(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0007159
(43) 공개일자 2024년01월16일

(51) 국제특허분류(Int. Cl.)
G16B 40/20 (2019.01) *C12Q 1/6869* (2018.01)
G16B 20/20 (2019.01) *G16B 30/10* (2019.01)
G16B 40/10 (2019.01)
(52) CPC특허분류
G16B 40/20 (2019.02)
C12Q 1/6869 (2018.05)
(21) 출원번호 10-2023-7038839
(22) 출원일자(국제) 2022년04월12일
심사청구일자 없음
(85) 번역문제출일자 2023년11월10일
(86) 국제출원번호 PCT/CN2022/086260
(87) 국제공개번호 WO 2022/218290
국제공개일자 2022년10월20일
(30) 우선권주장
63/173,728 2021년04월12일 미국(US)

(71) 출원인
더 차이니즈 유니버시티 오브 홍콩
중국 뉴 테리토리즈 홍콩 새틴 더 차이니즈 유니버시티 오브 홍콩 피 치우 빌딩 오피스 오브 리서치 앤 놀리지 트랜스퍼 서비시즈 룸 301
(72) 발명자
로 육밍 데니스
중국 홍콩 뉴 테리토리즈 샤틴 피 치우 빌딩 룸 301 오알케이티에스 더 차이니즈 유니버시티 오브 홍콩 내
치우 로싸 웨이 쿤
중국 홍콩 뉴 테리토리즈 샤틴 피 치우 빌딩 룸 301 오알케이티에스 더 차이니즈 유니버시티 오브 홍콩 내
(뒷면에 계속)
(74) 대리인
김진희, 김태홍

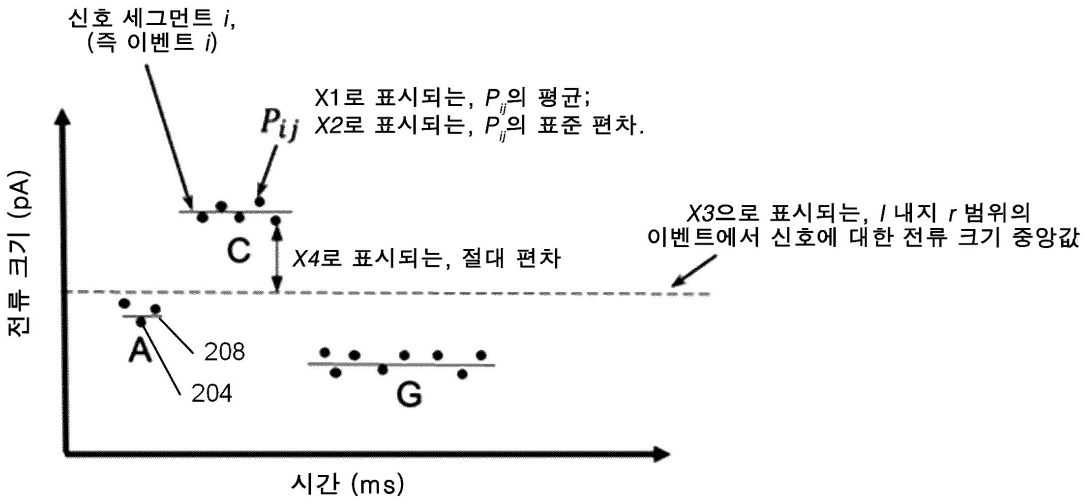
전체 청구항 수 : 총 50 항

(54) 발명의 명칭 **전기 신호를 이용한 염기 변형 분석**

(57) 요 약

전기 신호 및 다른 데이터를 사용하여 염기 변형을 결정하는 시스템 및 방법이 본원에 기재된다. 구현예는 다양한 염기 변형에 의해 영향을 받는 나노포어(116)를 사용하여 획득된 것들과 같은 시퀀싱과 관련된 전기 신호에서 유도된 특징뿐만 아니라 메틸화 상태가 결정되는 표적 위치 주변의 윈도우에서 뉴클레오티드의 정체를 활용할 수 있다. 다른 특징은 뉴클레오티드에 대응하는 전기 신호 세그먼트의 통계값과 핵산 분자 영역의 윈도우에서 전기 신호의 통계값의 벡터를 포함할 수 있다. 검출된 염기 변형은 생물학적 샘플의 추가 분석을 위해 사용될 수 있다.

대 표 도



(52) CPC특허분류
G16B 20/20 (2019.02)
G16B 30/10 (2019.02)
G16B 40/10 (2019.02)
C12Q 2563/116 (2013.01)
C12Q 2600/154 (2013.01)
(72) 발명자
찬 콴 치
중국 홍콩 뉴 테리토리즈 샤틴 피 치우 빌딩 룸 301 오알케이티에스 더 차이니즈 유니버시티 오브 홍콩 내
지앙 페이용
중국 홍콩 뉴 테리토리즈 샤틴 피 치우 빌딩 룸 301 오알케이티에스 더 차이니즈 유니버시티 오브 홍콩 내
청 숙 항
중국 홍콩 뉴 테리토리즈 샤틴 피 치우 빌딩 룸 301 오알케이티에스 더 차이니즈 유니버시티 오브 홍콩 내
덩 지아엔
중국 홍콩 뉴 테리토리즈 샤틴 피 치우 빌딩 룸 301 오알케이티에스 더 차이니즈 유니버시티 오브 홍콩 내

명 세 서

청구범위

청구항 1

하기를 포함하는, 핵산 분자에서 뉴클레오티드의 변형을 검출하는 방법:

입력 데이터 구조를 수신하는 것으로서, 입력 데이터 구조는 샘플 핵산 분자에서 시퀀싱된 뉴클레오티드의 윈도우에 대응하고, 샘플 핵산 분자는 뉴클레오티드에 대응하는 전기 신호를 측정하여 시퀀싱되고, 입력 데이터 구조는 하기 특성에 대한 값을 포함하는 것:

윈도우 내 각 뉴클레오티드에 대해:

뉴클레오티드의 정체,

각 윈도우 내 표적 위치에 대한 뉴클레오티드의 위치, 및

뉴클레오티드에 대응하는 전기 신호 세그먼트의 제1 세그먼트 통계값을 포함하는 벡터;

입력 데이터 구조를 모델에 입력하는 것으로서, 모델은

제1 복수의 제1 데이터 구조를 수신하되, 제1 복수의 제1 데이터 구조의 각 제1 데이터 구조는 복수의 제1 핵산 분자의 각 핵산 분자에서 시퀀싱된 뉴클레오티드의 각 윈도우에 대응하고, 각 제1 핵산 분자는 뉴클레오티드에 대응하는 전기 신호를 측정하여 시퀀싱되며, 변형은 각 제1 핵산 분자의 각 윈도우의 표적 위치에서 뉴클레오티드의 알려진 제1 상태를 가지며, 각 제1 데이터 구조는 입력 데이터 구조와 동일한 특성에 대한 값을 포함하는 것,

복수의 제1 훈련 샘플을 저장하되, 각각 제1 복수의 제1 데이터 구조 중 하나 및 표적 위치에서 뉴클레오티드의 제1 상태를 표시하는 제1 표지를 포함하는 것,

복수의 제1 훈련 샘플을 사용하되, 제1 복수의 제1 데이터 구조가 모델에 입력될 때 제1 표지의 대응하는 표지와 일치하거나 일치하지 않는 모델의 출력에 기반하여 모델의 매개변수를 최적화하는 것으로서, 모델의 출력은 각 윈도우의 표적 위치에서 뉴클레오티드에 변형을 갖는지 여부를 특정하는 것에 의해 훈련되는 것,

모델을 사용하여 변형이 입력 데이터 구조의 윈도우 내 표적 위치에서 뉴클레오티드에 존재하는지 여부를 결정하는 것.

청구항 2

제1항에 있어서, 제1 세그먼트 통계값이 뉴클레오티드에 대응하는 전기 신호 세그먼트의 평균을 나타내는, 방법.

청구항 3

제1항에 있어서, 제1 세그먼트 통계값이 뉴클레오티드에 대응하는 전기 신호 세그먼트의 전기 신호의 변동을 나타내는, 방법.

청구항 4

제1항에 있어서, 제1 세그먼트 통계값이 뉴클레오티드에 대응하는 전기 신호의 세그먼트 평균의 정규화된 값을 나타내는, 방법.

청구항 5

제1항, 제2항 또는 제4항 중 어느 한 항에 있어서, 벡터가 뉴클레오티드에 대응하는 전기 신호 세그먼트의 변동을 나타내는 제2 세그먼트 통계값을 포함하는, 방법.

청구항 6

제1항, 제2항 또는 제3항 중 어느 한 항에 있어서, 벡터가 뉴클레오티드에 대응하는 전기 신호의 세그먼트 평균의 정규화된 값을 나타내는 제2 세그먼트 통계값을 포함하는, 방법.

청구항 7

제2항에 있어서,

벡터가 뉴클레오티드에 대응하는 전기 신호 세그먼트의 변동을 나타내는 제2 세그먼트 통계값을 포함하고,

벡터가 제1 세그먼트 통계값의 정규화된 값을 나타내는 제3 세그먼트 통계값을 포함하는, 방법.

청구항 8

제1항 내지 제7항 중 어느 한 항에 있어서, 입력 데이터 구조가 윈도우보다 크거나 같은 핵산 분자 영역에서 전기 신호의 제1 영역 통계값에 대한 값을 포함하는, 방법.

청구항 9

제8항에 있어서, 제1 영역 통계값이 상기 영역에서의 전기 신호의 평균 또는 중앙값을 나타내는, 방법.

청구항 10

제8항에 있어서, 제1 영역 통계값은 상기 영역에서의 전기 신호의 평균 또는 중앙값으로부터의 전기 신호의 변동 절대값의 중앙값 또는 평균을 나타내는, 방법.

청구항 11

제9항에 있어서, 입력 데이터 구조가 상기 영역에서의 전기 신호의 평균 또는 중앙값으로부터의 전기 신호의 변동 절대값의 중앙값 또는 평균을 나타내는 제2 영역 통계값을 추가로 포함하는, 방법.

청구항 12

제8항 내지 제11항 중 어느 한 항에 있어서, 상기 영역이 샘플 핵산 분자의 한 가닥 상에 있는, 방법.

청구항 13

제8항 내지 제12항 중 어느 한 항에 있어서, 상기 영역이 샘플 핵산 분자이거나 적어도 5, 10, 15, 20, 25, 30, 50, 100, 200, 300, 400, 500, 또는 1000, 5000, 10000, 50000 또는 100만개의 뉴클레오티드를 포함하는, 방법.

청구항 14

제8항 내지 13항 중 어느 한 항에 있어서, 상기 영역은 뉴클레오티드 주위가 중앙에 있는, 방법.

청구항 15

제1항 내지 제14항 중 어느 한 항에 있어서, 윈도우가 샘플 핵산 분자의 두 가닥 상의 뉴클레오티드를 포함하는, 방법.

청구항 16

제1항 내지 제15항 중 어느 한 항에 있어서, 변형이 메틸화 또는 산화인, 방법.

청구항 17

제1항 내지 제16항 중 어느 한 항에 있어서, 전기 신호가 전류, 전압, 저항, 인덕턴스, 정전용량 또는 임피던스인, 방법.

청구항 18

제1항 내지 제17항 중 어느 한 항에 있어서, 나노포어를 사용하여 샘플 핵산 분자를 시퀀싱하는 것을 추가로 포함하는 방법.

청구항 19

제1항에 있어서,

변형이 메틸화이고,

샘플 핵산 분자가 무세포이고 태아를 임신한 여성 대상체의 생물학적 샘플로부터 얻어지며,

표적 위치에서 뉴클레오티드의 변형 상태를 사용하여 샘플 핵산 분자가 태아 또는 모체 기원인지 여부를 결정하는 것으로서, 변형 상태는 변형이 존재하는지 여부, 및 선택적으로 샘플 핵산 분자의 하나 이상의 다른 뉴클레오티드의 변형 상태인 것을 추가로 포함하는, 방법.

청구항 20

제19항에 있어서, 샘플 핵산 분자가 태아 또는 모체 기원인지 여부를 결정하는 것이 다음을 포함하는, 방법:

하나 이상의 뉴클레오티드의 변형 상태를 사용하여 샘플 핵산 분자의 메틸화 수준을 결정하는 것; 및

샘플 핵산 분자의 메틸화 수준을 참조값과 비교하는 것.

청구항 21

제20항에 있어서, 참조값이 하나 이상의 모체 핵산 분자의 메틸화 수준으로부터 결정되는, 방법.

청구항 22

제20항에 있어서,

샘플 핵산 분자의 메틸화 수준을 참조값과 비교하는 것이 샘플 핵산 분자의 메틸화 수준이 참조값보다 낮음을 결정하는 것을 포함하고,

샘플 핵산 분자가 태아 또는 모체 기원인지 여부를 결정하는 것이 비교를 사용하여 샘플 핵산 분자가 태아 기원임을 결정하는 것을 포함하는, 방법.

청구항 23

제19항에 있어서,

샘플 핵산 분자를 사전 결정된 게놈 영역에 대해 정렬함으로써 식별하는 것을 추가로 포함하는 방법.

청구항 24

제19항에 있어서,

샘플 핵산 분자가 복수의 샘플 핵산 분자 중 하나의 샘플 핵산 분자이고,

다음 단계를 추가로 포함하는, 방법:

변형 상태를 사용하여 각 복수의 샘플 핵산 분자가 태아 또는 모체 기원인지 여부를 결정하는 것, 및

복수의 샘플 핵산 분자의 태아 또는 모체 기원의 결정을 사용하여 태아 분율을 결정하는 것.

청구항 25

제1항에 있어서,

변형이 메틸화이고,

샘플 핵산 분자가 무세포이고 태아를 임신한 여성 대상체의 생물학적 샘플로부터 얻어지며,

샘플 핵산 분자가 복수의 샘플 핵산 분자 중 하나의 샘플 핵산 분자이고,

다음 단계를 추가로 포함하는, 방법:

복수의 샘플 핵산 분자를 태아 게놈의 영역에 대해 정렬함으로써 식별하는 것,

복수의 샘플 핵산 분자 중 각 샘플 핵산 분자의 하나 이상의 뉴클레오티드의 변형 상태를 결정하는 것,

복수의 샘플 핵산 분자 중 각 샘플 핵산 분자에 대한 하나 이상의 뉴클레오티드의 변형 상태를 사용하여 영역의 메틸화 수준을 결정하는 것, 및

메틸화 수준을 사용하여 카피수 이상이 태아 게놈의 영역에 존재하는지 여부를 결정하는 것.

청구항 26

하기를 포함하는, 핵산 분자에서 뉴클레오티드의 변형을 검출하는 방법:

제1 복수의 제1 데이터 구조를 수신하는 것으로서, 제1 복수의 제1 데이터 구조의 각 제1 데이터 구조는 복수의 제1 핵산 분자의 각 핵산 분자에서 시퀀싱된 뉴클레오티드의 각 윈도우에 대응하고, 각 제1 핵산 분자는 뉴클레오티드에 대응하는 전기 신호를 측정하여 시퀀싱되며, 변형은 각 제1 핵산 분자의 각 윈도우의 표적 위치에서 뉴클레오티드의 알려진 제1 상태를 가지며, 각 제1 데이터 구조는 다음 특성에 대한 값을 포함하는 것:

윈도우 내 각 뉴클레오티드에 대해:

뉴클레오티드의 정체,

각 윈도우 내 표적 위치에 대한 뉴클레오티드의 위치, 및

뉴클레오티드에 대응하는 전기 신호 세그먼트의 제1 세그먼트 통계값을 포함하는 벡터;

복수의 제1 훈련 샘플을 저장하는 것으로서, 각각 제1 복수의 제1 데이터 구조 중 하나 및 표적 위치에서 뉴클레오티드의 변형에 대한 제1 상태를 표시하는 제1 표지를 포함하는 것; 및

제1 복수의 제1 데이터 구조가 모델에 입력될 때 제1 표지의 대응하는 표지와 일치하거나 일치하지 않는 모델의 출력에 기반하여 모델의 매개변수를 최적화하여 복수의 제1 훈련 샘플을 사용하여 모델을 훈련하는 것으로서, 모델의 출력은 각 윈도우의 표적 위치에서 뉴클레오티드가 변형을 갖는지 여부를 특정하는 것.

청구항 27

제26항에 있어서,

제2 복수의 제2 데이터 구조를 수신하는 것으로서, 제2 복수의 제2 데이터 구조의 각 제2 데이터 구조는 복수의 제2 핵산 분자의 각 핵산 분자에서 시퀀싱된 뉴클레오티드의 각 윈도우에 대응하며, 변형은 각 제2 핵산 분자의 각 윈도우 내 표적 위치에서 뉴클레오티드의 알려진 제2 상태를 갖고, 각 제2 데이터 구조는 제1 복수의 제1 데이터 구조와 동일한 특성에 대한 값을 포함하는 것;

복수의 제2 훈련 샘플을 저장하는 것으로서, 각각 제2 복수의 제2 데이터 구조 중 하나 및 표적 위치에서 뉴클레오티드의 제2 상태를 표시하는 제2 표지를 포함하는 것을 추가로 포함하고;

훈련은 제1 상태 또는 제2 상태는 변형이 존재하는 것이고, 다른 상태는 변형이 부재하는 것이며,

모델은 제2 복수의 제2 데이터 구조가 모델에 입력될 때 제2 표지의 대응하는 표지와 일치하거나 일치하지 않는 모델의 출력에 기반하여 모델의 매개변수를 최적화하여 복수의 제2 훈련 샘플을 사용하는 것을 추가로 포함하는, 방법.

청구항 28

제27항에 있어서, 복수의 제1 핵산 분자가 복수의 제2 핵산 분자와 동일한, 방법.

청구항 29

제26항에 있어서,

제1 복수의 제1 데이터 구조와 관련된 각 윈도우가 제1 핵산 분자의 제1 가닥의 뉴클레오티드 및 제1 핵산 분자의 제2 가닥의 뉴클레오티드를 포함하고,

각 제1 데이터 구조가 윈도우 내 각 뉴클레오티드에 대해 가닥 특성의 값을 추가로 포함하며, 가닥 특성은 뉴클레오티드가 제1 가닥 또는 제2 가닥에 존재함을 표시하는, 방법.

청구항 30

제26항에 있어서, 변형이 표적 위치에서 뉴클레오티드의 메틸화를 포함하는, 방법.

청구항 31

제30항에 있어서, 알려진 제1 상태가 제1 데이터 구조의 제1 부분에 대한 메틸화 상태 및 제1 데이터 구조의 제2 부분에 대한 비메틸화 상태를 포함하는, 방법.

청구항 32

제26항에 있어서, 제1 세그먼트 통계값이 뉴클레오티드에 대응하는 전기 신호 세그먼트의 평균을 나타내는, 방법.

청구항 33

제26항에 있어서, 제1 세그먼트 통계값이 뉴클레오티드에 대응하는 전기 신호 세그먼트의 전기 신호의 변동을 나타내는, 방법.

청구항 34

제26항에 있어서, 제1 세그먼트 통계값이 뉴클레오티드에 대응하는 전기 신호의 세그먼트 평균의 정규화된 값을 나타내는, 방법.

청구항 35

제26항, 제32항 또는 제34항 중 어느 한 항에 있어서, 벡터가 뉴클레오티드에 대응하는 전기 신호 세그먼트의 변동을 나타내는 제2 세그먼트 통계값을 포함하는, 방법.

청구항 36

제26항, 제32항 또는 제33항 중 어느 한 항에 있어서, 벡터가 뉴클레오티드에 대응하는 전기 신호 세그먼트의 평균의 정규화된 값을 나타내는 제2 세그먼트 통계값을 포함하는, 방법.

청구항 37

제32항에 있어서,

벡터가 뉴클레오티드에 대응하는 전기 신호 세그먼트의 변동을 나타내는 제2 세그먼트 통계값을 포함하고,

벡터가 제1 세그먼트 통계값의 정규화된 값을 나타내는 제3 세그먼트 통계값을 포함하는, 방법.

청구항 38

제26항 내지 제37항 중 어느 한 항에 있어서, 각 제1 데이터 구조가 윈도우보다 크거나 같은 각 핵산 분자 영역에서 전기 신호의 제1 영역 통계값에 대한 값을 포함하는, 방법.

청구항 39

제38항에 있어서, 제1 영역 통계값이 상기 영역에서 전기 신호의 평균 또는 중앙값을 나타내는, 방법.

청구항 40

제38항에 있어서, 제1 영역 통계값이 상기 영역에서 전기 신호의 평균 또는 중앙값으로부터의 전기 신호의 변동 절대값의 중앙값 또는 평균을 나타내는, 방법.

청구항 41

제39항에 있어서, 제1 데이터 구조가 상기 영역에서 전기 신호의 평균 또는 중앙값으로부터의 전기 신호의 변동 절대값의 중앙값 또는 평균을 나타내는 제2 영역 통계값을 추가로 포함하는, 방법.

청구항 42

제38항 내지 제41항 중 어느 한 항에 있어서, 상기 영역이 각 핵산 분자의 한 가닥 상에 있는, 방법.

청구항 43

제38항 내지 제45항 중 어느 한 항에 있어서, 상기 영역이 각 핵산 분자이거나 적어도 5, 10, 15, 20 25, 30, 50, 100, 200, 300, 400, 500, 또는 1000, 5000, 10000, 50000 또는 100만개의 뉴클레오티드를 포함하는, 방법.

청구항 44

제38항 내지 제43항 중 어느 한 항에 있어서, 상기 영역은 뉴클레오티드 주위가 중앙에 있는, 방법.

청구항 45

제26항 내지 제44항 중 어느 한 항에 있어서, 윈도우가 각 핵산 분자의 두 가닥 상에서 뉴클레오티드를 포함하는, 방법.

청구항 46

실행될 때 제1항 내지 제45항 중 어느 한 항의 방법을 수행하도록 컴퓨터 시스템을 제어하는 복수의 명령을 저장하는 비일시적 컴퓨터 판독 가능 매체를 포함하는 컴퓨터 제품.

청구항 47

다음을 포함하는 시스템:

제46항의 컴퓨터 제품; 및

컴퓨터 판독 가능 매체에 저장된 명령을 실행하기 위한 하나 이상의 처리장치.

청구항 48

상기 방법 중 어느 하나를 수행하기 위한 수단을 포함하는 시스템.

청구항 49

상기 방법 중 어느 하나를 수행하도록 구성된 하나 이상의 처리장치를 포함하는 시스템.

청구항 50

상기 방법 중 어느 하나의 단계를 각각 수행하는 모듈을 포함하는 시스템.

발명의 설명

기 술 분 야

[0001] **관련 출원에 대한 상호 참조**

[0002] 본 출원은 그 전체가 모든 목적을 위해 본원에 참조로 포함된, 2021년 4월 12일에 출원된 미국 특허 가출원 63/173,728에 대한 우선권을 주장한다.

배 경 기 술

[0003] 핵산의 염기 변형의 존재는 바이러스, 박테리아, 식물, 진균, 선충류, 곤충 및 척추동물(예를 들어, 인간) 등을 포함한 상이한 유기체에 걸쳐 다양하다. 가장 일반적인 염기 변형은 상이한 위치에서 상이한 DNA 염기에 대한 메틸기의 부가, 소위 메틸화이다. 메틸화는 시토신, 아데닌, 티민 및 구아닌에서, 예컨대 5mC(5-메틸시토신), 4mC(N4-메틸시토신), 5hmC(5-하이드록시메틸시토신), 5fC(5-포르밀시토신), 5caC(5-카복실시토신), 1mA(N1-메틸아데닌), 3mA(N3-메틸아데닌), N6-메틸아데닌(6mA), 7mA(N7-메틸아데닌), 3mC(N3-메틸시토신), 2mG(N2-메틸구아닌), 6mG(O6-메틸구아닌), 7mG(N7-메틸구아닌), 3mT(N3-메틸티민) 및 4mT(O4-메틸티민)가 발견되었다. 척추동물 게놈에서 5mC는 가장 일반적인 유형의 염기 메틸화이며 구아닌이(즉, CpG 맥락에서) 그 뒤를 따른다.

[0004] DNA 메틸화는 포유동물 발달에 필수적이며 유전자 발현 및 침묵화, 배아 발달, 전사, 염색질 구조, X 염색체 불

활성화, 반복 요소의 활성에 대한 보호, 유사분열 동안 게놈 안정성 유지 및 기원-부모 게놈 각인의 조절에서 주목할만한 역할을 한다.

[0005] DNA 메틸화는 조율되는 방식으로 프로모터 및 인핸서의 침묵화에 많은 중요한 역할을 한다(Robertson, 2005; Smith and Meissner, 2013). 각인 장애(예를 들어, 베크위드-위드만 증후군 및 프라더-윌리 증후군), 반복 불안정성 질환(예를 들어, 취약 X 증후군), 자가면역 질환(예를 들어, 전신 홍반성 루푸스), 대사 장애(예를 들어, 제1형 및 제2형 당뇨병), 신경 장애, 노화 등을 포함하지만 이에 제한되지 않는 여러 인간 질환이 DNA 메틸화 이상과 관련된 것으로 발견되었다.

[0006] DNA 분자에 대한 메틸로믹 변형의 정확한 측정은 수많은 임상적 의미를 가질 것이다. DNA 메틸화를 측정하기 위해 널리 사용되는 방법 중 하나는 바이설파이트 시퀀싱(BS-seq)의 사용을 통하는 것이다(Lister 등, 2009; Frommer 등, 1992). 이 접근에서 DNA 샘플은 먼저 비메틸화 시토신(예를 들어, C)을 우라실로 전환하는 바이설파이트로 처리된다. 대조적으로, 메틸화 시토신은 변하지 않은 채로 남아 있다. 그런 다음 바이설파이트 변형 DNA가 DNA 시퀀싱으로 분석된다. 또 다른 접근에서, 바이설파이트 전환 후, 변형된 DNA는 상이한 메틸화 프로파일의 바이설파이트 전환된 DNA를 구별할 수 있는 프라이머를 사용하여 중합효소 연쇄 반응(PCR) 증폭을 거친다(Herman 등, 1996). 후자의 접근을 메틸화 특이적 PCR이라고 한다.

[0007] 이러한 바이설파이트 기반 접근의 한 가지 단점은 바이설파이트 전환 단계가 처리된 DNA의 대부분을 유의하게 분해하는 것으로 보고되었다는 것이다(Grunau, 2001). 또 다른 단점은 바이설파이트 전환 단계가 강한 CG 편향을 생성하여(Olova 등, 2018), 전형적으로 이종 메틸화 상태를 갖는 DNA 혼합물의 신호 대 노이즈 비의 감소를 초래할 것이라는 점이다. 더욱이, 바이설파이트 시퀀싱은 바이설파이트 처리 동안 DNA 분해로 인해 긴 DNA 분자를 시퀀싱하는 이상적인 방법이 아니다.

[0008] 핵산 염기 변형의 바이설파이트 비사용 결정을 달성하기 위한 많은 지속적인 노력이 있어왔다. 그러나 바이설파이트 시퀀싱에 필적하는 민감도 및 특이성 수준을 달성한 상업적으로 실행 가능한 도구가 부족하다. 나노포어(Nanopore) 시퀀싱은 샘플의 화학적 표지화를 필요로 하지 않는 매력적인 시퀀싱 유형이다. 나노포어 시퀀싱을 사용한 염기 변형의 검출은 상대적으로 비용이 저렴하고 효율적일 수 있다.

[0009] 따라서, 나노포어 시퀀싱을 사용하여 염기 변형을 결정할 필요가 있다. 본 개시내용에서, 본 발명자들은 염기 변형 결정에 대한 높은 민감도 및 특이성으로 나노포어 시퀀싱에 의해 생성된 전류 신호를 처리하는 새로운 방법 및 시스템을 기재한다.

발명의 내용

[0010] 기재된 구현예는 주형 DNA 전처리, 예컨대 효소적 및/또는 화학적 전환, 또는 단백질 및/또는 항체 결합 없이 염기 변형, 예컨대 핵산에서 5mC의 결정을 허용한다. 본 개시내용에 제시된 구현예는 예를 들어 4mC, 5hmC, 5fC, 5caC, 1mA, 3mA, 6mA, 7mA, 3mC, 2mG, 6mG, 7mG, 3mT, 4mT 등을 포함하지만 이에 제한되지 않는 상이한 유형의 염기 변형을 검출하기 위해 사용될 수 있다. 이러한 구현예는 다양한 염기 변형에 의해 영향을 받는 시퀀싱과 관련된 전기 신호에서 유도된 특징, 예컨대 나노포어를 사용하여 얻은 것들뿐만 아니라 메틸화 상태가 결정되는 표적 위치 주위의 윈도우에서 뉴클레오티드의 정체를 활용할 수 있다. 뉴클레오티드에 대한 미가공 전기 신호는 또한 뉴클레오티드의 상류 또는 하류의 뉴클레오티드와 관련될 수 있다. 미가공 전기 신호는 적합한 기술을 사용하여 상이한 뉴클레오티드에 할당될 수 있다.

[0011] 본 발명의 구현예는 나노포어 시퀀싱과 함께 사용될 수 있다. 나노포어 시퀀싱 시스템의 한 예는 Oxford Nanopore Technologies에 의해 상용화된 시스템이다. 방법은 나노포어를 사용하여 측정된 전기신호를 사용할 수 있다. 방법은 뉴클레오티드의 정체, 표적 위치에 대한 뉴클레오티드의 위치, 뉴클레오티드에 대응하는 전기 신호 세그먼트의 통계값을 포함하는 벡터, 및 핵산 분자 영역의 윈도우에서 전기 신호의 통계값을 사용할 수 있다.

[0012] 본 발명자들이 개발한 방법은 연구 및 진단 목적을 포함하지만 이에 제한되지 않는 다양한 목적으로 샘플의 메틸화 프로파일을 평가하기 위해 생물학적 샘플에서 염기 변형을 검출하는 도구로서 역할을 할 수 있다. 검출된 메틸화 프로파일은 상이한 분석을 위해 사용될 수 있다. 메틸화 프로파일은 DNA의 기원(예를 들어, 모체 또는 태아, 조직, 박테리아)을 검출하기 위해 사용될 수 있다. 조직의 비정상적인 메틸화 프로파일의 검출은 개인의 발달 및 다른 장애의 식별을 돕는다.

[0013] 본 발명의 구현예의 성질 및 장점에 대한 더 우수한 이해는 하기 상세한 설명 및 첨부 도면을 참조하여 얻을 수

있다.

도면의 간단한 설명

[0014] **도 1**은 나노포어 시퀀싱을 예시한다.

도 2는 본 발명의 구현예에 따른 상이한 신호 특징을 예시한다.

도 3은 본 발명의 구현예에 따른 전류 신호 분할 및 신호 특징 벡터의 작제을 예시한다.

도 4는 본 발명의 구현예에 따른 나노포어를 통과하는 각 뉴클레오티드에 대한 이벤트의 길이(즉, 기간) 분포의 그래프이다.

도 5는 본 발명의 구현예에 따른 전류 패턴, 시퀀싱 위치 및 시퀀싱 맥락을 포함하는 통합 제시 행렬을 사용하여 5mC 검출을 위한 원리를 예시한다.

도 6은 본 발명의 구현예에 따른 이중 가닥 DNA의 두 가닥에 기반하여 전류 패턴, 시퀀싱 위치 및 시퀀싱 맥락을 포함하는 통합 제시 행렬을 사용하여 염기 변형 검출을 위한 원리를 예시한다.

도 7은 본 발명의 구현예에 따른 염기 변형 분석의 성능에 대한 커널 크기의 영향을 나타낸다.

도 8은 본 발명의 구현예에 따른 메틸화 검출 측면에서 훈련 및 시험을 위해 사용된 시퀀싱 분자의 수를 나타낸다.

도 9a-9d는 본 발명의 구현예에 따른 IPM-CNN 및 IPM-RNN 접근을 사용하여 WGA DNA 및 M.SssI-처리 DNA 데이터세트 간 CpG가 메틸화될 확률의 상자그래프이다.

도 10a 및 **도 10b**는 본 발명의 구현예에 따른 훈련 데이터세트 및 시험 데이터세트에 대한 ROC(수신자 조작 특성(ROC)) 곡선을 나타낸다.

도 11은 본 발명의 구현예에 따른 메틸화 분석을 위한 상이한 도구의 성능의 표이다.

도 12는 본 발명의 구현예에 따른 핵산 분자 내 뉴클레오티드의 변형을 검출하는 과정의 흐름도이다.

도 13은 본 발명의 구현예에 따른 핵산 분자 내 뉴클레오티드의 변형을 검출하는 과정의 흐름도이다.

도 14는 본 발명의 구현예에 따른 측정 시스템을 예시한다.

도 15는 본 발명의 구현예에 따른 시스템 및 방법과 함께 사용 가능한 예시적 컴퓨터 시스템의 블록 다이어그램을 나타낸다.

도 16은 본 발명의 구현예에 따른 ROC 곡선 아래 면적(AUC)에 대한 상이한 매개변수 조합의 효과의 그래프를 나타낸다.

도 17은 본 발명의 구현예에 따른 AUC에 대한 윈도우 크기의 효과의 그래프를 나타낸다.

도 18은 본 발명의 구현예에 따른 전류 패턴, 시퀀싱 위치 및 시퀀싱 맥락을 포함하는 통합 제시 행렬을 사용하여 6mA 검출을 위한 원리를 예시한다.

도 19는 본 발명의 구현예에 따른 6mA 검출의 AUC 그래프를 나타낸다.

도 20은 본 발명의 구현예에 따른 백혈구 연층 및 NPC 종양 샘플로부터 유래된 DNA에 대한 IPM-RNN 모델에 의해 결정된 단일 분자 메틸화 수준의 비교이다.

도 21은 본 발명의 구현예에 따른 단일 분자 메틸화 패턴의 예를 나타낸다.

도 22는 본 발명의 구현예에 따른 모체 특이적 및 태아 특이적 무세포 DNA 분자의 단일 분자 메틸화 수준의 그래프이다.

도 23은 본 발명의 구현예에 따른 IPM-CNN 모델에 의해 결정된 메틸화 패턴을 사용하여 무세포 DNA 분자의 태아 및 모체 기원을 결정하기 위한 ROC 곡선이다.

발명을 실시하기 위한 구체적인 내용

[0015] 용어

[0016] *"조직"*은 기능 단위로 함께 그룹화되는 세포 그룹에 대응한다. 단일 조직에서 하나를 초과하는 유형의 세포가 발견될 수 있다. 상이한 유형의 조직은 상이한 유형의 세포(예를 들어, 간세포, 폐포 세포 또는 혈액 세포)로 구성될 수 있지만, 또한 상이한 유기체로부터의 조직(모체 대 태아; 이식을 받은 대상체의 조직; 미생물 또는 바이러스에 의해 감염된 유기체의 조직) 또는 건강한 세포 대 종양 세포에 대응할 수 있다. "참조 조직"은 조직 특이적 메틸화 수준을 결정하기 위해 사용된 조직에 대응할 수 있다. 상이한 개체로부터의 동일한 조직 유형의 다중 샘플이 해당 조직 유형에 대한 조직 특이적 메틸화 수준을 결정하기 위해 사용될 수 있다.

[0017] "*생물학적 샘플*"은 인간 대상체로부터 채취된 임의의 세포 샘플을 지칭한다. 생물학적 샘플은 조직 생검, 미세바늘 흡인물 또는 혈액 세포일 수 있다. 샘플은 또한 임산부로부터 채취된 무세포 샘플, 예를 들어 혈장, 혈청 또는 소변일 수 있다. 다양한 구현예에서, 무세포 DNA가 농축된 임산부로부터의 생물학적 샘플(예를 들어, 원심분리 프로토콜을 통해 얻은 혈장 샘플) 내 DNA의 대부분은 무세포일 수 있으며, 예를 들어 DNA의 50%, 60%, 70%, 80%, 90%, 95% 또는 99% 초과는 무세포일 수 있다. 원심 분리 프로토콜은 예를 들어 3,000 g x 10분, 유체 부분 수득, 및 예를 들어 30,000 g에서 추가 10분 동안 재원심분리하여 잔류 세포를 제거하는 것을 포함할 수 있다. 특정 구현예에서, 3,000 g 원심분리 단계 후에, 유체 부분의 여과가 후속 조치될 수 있다(예를 들어 직경이 5 μm 이하인 포어 크기의 필터를 사용함).

[0018] "*서열 판독*"은 핵산 분자의 임의의 부분 또는 전부로부터 시퀀싱된 뉴클레오티드 문자열을 지칭한다. 예를 들어, 서열 판독은 핵산 단편으로부터 시퀀싱된 뉴클레오티드의 짧은 문자열(예를 들어, 20-150), 핵산 단편의 한쪽 또는 양쪽 말단에서의 뉴클레오티드의 짧은 문자열, 또는 생물학적 샘플에 존재하는 전체 핵산 단편의 시퀀싱일 수 있다. 서열 판독은 다양한 방식으로, 예를 들어 시퀀싱 기술을 사용하거나 탐침을 사용하여, 예를 들어 혼성화 어레이 또는 포획 탐침에서, 또는 증폭 기술, 예컨대 중합효소 연쇄 반응(PCR) 또는 단일 프라이머를 사용한 선형 증폭 또는 등온 증폭으로 얻을 수 있다.

[0019] "*부위*"("*게놈 부위*"라고도 함)는 단일 부위에 대응하며, 이는 단일 염기 위치 또는 연관된 염기 위치의 그룹, 예를 들어 CpG 부위 또는 연관된 염기 위치의 더 큰 그룹일 수 있다. "유전자좌"는 다수의 부위를 포함하는 영역에 대응할 수 있다. 유전자좌는 하나의 부위만 포함할 수 있으며, 이는 유전자좌를 해당 맥락에서 부위와 동일하게 만든다.

[0020] "*메틸화 상태*"는 주어진 부위에서의 메틸화 상태를 지칭한다. 예를 들어, 부위는 메틸화되었거나 메틸화되지 않았거나 일부 경우에는 결정되지 않았을 수 있다.

[0021] "*메틸화 지수*"는 그 부위를 포괄하는 전체 판독 수 대비 해당 부위에서 메틸화를 나타내는 DNA 단편의 비율(예를 들어, 서열 판독 또는 탐침으로부터 결정됨)을 지칭할 수 있다. "판독"은 DNA 단편에서 얻은 정보(예를 들어, 특정 부위에서의 메틸화 상태)에 대응할 수 있다. 판독은 하나 이상의 부위에서 특정 메틸화 상태의 DNA 단편에 우선적으로 혼성화하는 시약(예를 들어, 프라이머 또는 탐침)을 사용하여 얻을 수 있다. 전형적으로 이러한 시약은 이의 메틸화 상태에 따라 DNA 분자를 차별적으로 변형하거나 차별적으로 인식하는 공정, 예를 들어 바이설파이트 전환, 메틸화 민감성 제한 효소, 또는 메틸화 결합 단백질, 또는 메틸시토신 및 하이드록시메틸시토신을 인식하는 항-메틸시토신 항체 또는 단일 분자 시퀀싱 기술(예를 들어, 단일 분자, 실시간 시퀀싱(예를 들어, Pacific Biosciences) 및 나노포어 시퀀싱(예를 들어, Oxford Nanopore Technologies))으로의 처리 후에 적용된다.

[0022] "*메틸화 밀도*"는 메틸화를 나타내는 영역 내의 위치에서의 판독 수를 해당 영역의 부위를 포괄하는 총 판독 수로 나눈 것을 지칭할 수 있다. 부위는 특징적 특성, 예컨대 CpG 부위를 가질 수 있다. 따라서, 영역의 "CpG 메틸화 밀도"는 CpG 메틸화를 나타내는 판독 수를 해당 영역의 CpG 부위(예를 들어, 특정 CpG 부위, CpG 섬 내의 CpG 부위, 또는 더 넓은 영역)를 포괄하는 총 판독 수로 나눈 것을 지칭할 수 있다. 예를 들어, 인간 게놈의 각 100 kb 빈(bin)에 대한 메틸화 밀도는 100 kb 영역에 매핑된 서열 판독에 포괄된 모든 CpG 부위의 비율로 CpG 부위에서 바이설파이트 처리(메틸화 시토신에 대응함) 후 전환되지 않은 시토신의 총 수로부터 결정될 수 있다. 이 분석은 다른 빈 크기(예를 들어, 500 bp, 5 kb, 10 kb, 50 kb 또는 1 Mb 등)에 대해서도 수행될 수 있다. 영역은 전체 게놈, 염색체 또는 염색체의 일부(예를 들어, 염색체 아암)일 수 있다. 대안적으로, 메틸화 밀도는 본 개시내용에 기재된 구현예를 사용하는 나노포어 시퀀싱을 사용하여 바이설파이트 전환 없이 결정될 수 있다. CpG 부위의 메틸화 지수는 영역이 해당 CpG 부위만 포함할 때 이 영역의 메틸화 밀도와 동일하다. "메틸화 시토신의 비율"은 분석된 총 시토신 잔기 수 대비, 즉 이 영역에서 CpG 맥락 외부의 시토신을 포함하여, 메틸화된(예를 들어 바이설파이트 전환 후 전환되지 않은) 것으로 나타나는 시토신 부위의 수("C")를 지칭할 수 있다. 메틸화 지수, 메틸화 밀도, 하나 이상의 부위에서 메틸화 분자 수, 및 하나 이상의 부위에서 메틸화 분자(예를

들어, 시토신)의 비율이 "*메틸화 수준*"의 예이다. 바이설파이트 전환과는 별개로, 메틸화 상태에 민감한 효소(예를 들어, 메틸화 민감성 제한 효소), 메틸화 결합 단백질, 메틸화 상태에 민감한 플랫폼을 사용한 단일 분자 시퀀싱(예를 들어, 나노포어 시퀀싱(Schreiber 등 Proc Natl Acad Sci 2013; 110: 18910-18915) 및 단일 분자에 의한, 실시간 시퀀싱(예를 들어, Pacific Biosciences)(Flusberg 등Nat Methods Flusberg 등 Nat Methods 2010; 7: 461-465))을 포함하지만 이에 제한되지 않는, 당업자에게 알려진 다른 공정이 DNA 분자의 메틸화 상태를 조사하기 위해 사용될 수 있다.

[0023] "*메틸롬(methylome)*"은 게놈 내 복수의 부위 또는 유전자좌에서의 DNA 메틸화 양의 측정을 제공한다. 메틸롬은 게놈 전체, 게놈의 상당 부분, 또는 게놈의 상대적으로 작은 부분(들)에 대응할 수 있다.

[0024] "*임신 혈장 메틸롬*"은 임신한 동물(예를 들어, 인간)의 혈장 또는 혈청으로부터 결정된 메틸롬이다. 혈장 및 혈청은 무세포 DNA를 포함하므로 임신 혈장 메틸롬은 무세포 메틸롬의 예이다. 임신 혈장 메틸롬은 신체 내 상이한 기관 또는 조직 또는 세포로부터의 DNA 혼합물이므로 혼합 메틸롬의 예이기도 한다. 한 구현예에서, 이러한 세포는 적혈(즉, 적혈구) 계통, 골수 계통(예를 들어, 호중구 및 이의 전구체) 및 거핵구 계통의 세포를 포함하지만 이에 제한되지 않는 조혈 세포이다. 임신 중 혈장 메틸롬은 태아 및 모체의 메틸롬 정보를 함유할 수 있다. "세포성 메틸롬"은 환자의 세포(예를 들어, 혈액 세포)에서 결정된 메틸롬에 대응한다. 혈액 세포의 메틸롬을 혈액 세포 메틸롬이라고 한다.

[0025] "*메틸화 프로파일*"은 여러 부위 또는 영역에 있어서 DNA 또는 RNA 메틸화와 관련된 정보를 포함한다. DNA 메틸화와 관련된 정보는 CpG 부위의 메틸화 지수, 영역에서 CpG 부위의 메틸화 밀도(약어로 MD), 인접 영역에 대한 CpG 부위의 분포, 하나 초과의 CpG 부위를 함유하는 영역 내의 각 개별 CpG 부위에 대한 메틸화 패턴 또는 수준 및 비-CpG 메틸화를 포함할 수 있지만 이에 제한되지 않는다. 한 구현예에서, 메틸화 프로파일은 하나 초과의 유형의 염기(예를 들어 시토신 또는 아데닌)의 메틸화 또는 비메틸화 패턴을 포함할 수 있다. 게놈의 상당 부분의 메틸화 프로파일은 메틸롬과 동등한 것으로 간주될 수 있다. 포유동물 게놈의 "DNA 메틸화"는 전형적으로 CpG 디뉴클레오티드 중 시토신 잔기의 5' 탄소에 대한 메틸기 부가(즉, 5-메틸시토신)를 지칭한다. DNA 메틸화는 다른 맥락의 시토신, 예를 들어 CHG 및 CHH에서 발생할 수 있고, H는 아데닌, 시토신 또는 티민이다. 시토신 메틸화는 5-하이드록시메틸시토신 형태일 수도 있다. 비시토신 메틸화, 예컨대 N^6-메틸 아데닌도 보고되었다.

[0026] "*메틸화 패턴*"은 메틸화 및 비메틸화 염기의 순서를 지칭한다. 예를 들어, 메틸화 패턴은 단일 DNA 가닥, 단일 이중 가닥 DNA 분자 또는 또 다른 유형의 핵산 분자에서 메틸화 염기의 순서일 수 있다. 예를 들어, 3개의 연속 CpG 부위는 하기 메틸화 패턴: UUU, MMM, UMM, UMU, UUM, MUM, MUU 또는 MMU 중 임의의 것을 가질 수 있고, "U"는 비메틸화 부위를 표시하고 "M"은 메틸화 부위를 표시한다. 이 개념을 메틸화를 포함하지만 이에 제한되지 않는 염기 변형으로 확장하면, 변형된 및 변형되지 않은 염기의 순서를 지칭하는 용어 "*변형 패턴*"을 사용하게 된다. 예를 들어, 변형 패턴은 단일 DNA 가닥, 단일 이중 가닥 DNA 분자, 또는 또 다른 유형의 핵산 분자에서 변형된 염기의 순서일 수 있다. 예를 들어, 잠재적으로 변형 가능한 3개의 연속 부위는 하기 변형 패턴: UUU, MMM, UMM, UMU, UUM, MUM, MUU 또는 MMU 중 임의의 것을 가질 수 있고, "U"는 변형되지 않은 부위를 표시하고 "M"은 변형된 부위를 표시한다. 메틸화에 기반하지 않은 염기 변형의 한 예는 산화 변화, 예컨대 8-옥소-구아닌이다.

[0027] 용어 "*과메틸화*" 및 "*저메틸화*"는 그 단일 분자 메틸화 수준에 의해 측정된 단일 DNA 분자의 메틸화 밀도, 예를 들어 분자 내의 메틸화 염기 또는 뉴클레오티드의 수를 해당 분자 내의 메틸화 가능한 염기 또는 뉴클레오티드 총 수로 나눈 것을 지칭할 수 있다. 과메틸화 분자는 단일 분자 메틸화 수준이 적용마다 정의될 수 있는 역치 이상인 분자이다. 역치는 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% 또는 95%일 수 있다. 저메틸화 분자는 단일 분자 메틸화 수준이 적용마다 정의될 수 있고 적용마다 변화할 수 있는 역치 이하인 분자이다. 역치는 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% 또는 95%일 수 있다.

[0028] 용어 "과메틸화" 및 "저메틸화"는 또한 이들 분자의 다중 분자 메틸화 수준에 의해 측정된 바와 같은 DNA 분자 집단의 메틸화 수준을 지칭할 수 있다. 과메틸화 분자 집단은 다중 분자 메틸화 수준이 적용마다 정의될 수 있고 적용마다 변화할 수 있는 역치 이상인 집단이다. 역치는 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% 또는 95%일 수 있다. 저메틸화 분자 집단은 다중 분자 메틸화 수준이 적용마다 정의될 수 있는 역치 이하인 집단이다. 역치는 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% 및 95%일 수 있다. 한 구현예에서, 분자 집단은 하나 이상의 선택된 게놈 영역에 대해 정렬될 수 있다. 한 구현예에서, 선택된 게놈 영역(들)은 질환, 예컨대 유전 장애, 각인 장애, 후성적 장애, 대사 장애 또는 신경학적 장애와 관련될 수 있다. 선택된 게놈 영역(들)은 50개 뉴클레오티드(nt), 100 nt, 200 nt, 300 nt, 500 nt, 1000 nt, 2 knt, 5 knt, 10 knt, 20 knt,

30 knt, 40 knt, 50 knt, 60 knt, 70 knt, 80 knt, 90 knt, 100 knt, 200 knt, 300 knt, 400 knt, 500 knt 또는 1 Mnt의 길이를 가질 수 있다.

[0029] 본원에 사용된 용어 "*분류*"는 샘플의 특정 특성과 관련된 임의의 수치(들) 또는 다른 문자(들)를 지칭한다. 예를 들어, "+" 기호(또는 단어 "양성")는 샘플이 삭제 또는 증폭을 갖는 것으로 분류됨을 의미할 수 있다. 분류는 이진형(예를 들어, 양성 또는 음성)일 수 있거나 더 많은 분류 수준(예를 들어, 1에서 10 또는 0에서 1까지의 척도)을 가질 수 있다.

[0030] 용어 "*컷오프*" 및 "*역치*"는 연산에서 사용되는 미리 정해진 수치를 지칭한다. 예를 들어 컷오프 크기는 그 초과에서 단편이 제외되는 크기를 지칭할 수 있다. 역치는 특정 분류가 적용되는 값보다 높거나 낮을 수 있다. 이러한 용어 중 하나가 이러한 맥락 중 하나에서 사용될 수 있다. 컷오프 또는 역치는 "참조값"이거나 특정 분류를 나타내거나 2개 이상의 분류를 구별하는 참조값에서 유도될 수 있다. 이러한 참조값은 당업자에게 이해될 바와 같이 다양한 방식으로 결정될 수 있다. 예를 들어, 계측은 상이한 알려진 분류를 가진 대상체의 상이한 두 코호트에 대해 결정될 수 있으며 참조값은 하나의 분류(예를 들어, 평균)의 대표값 또는 계측의 두 클러스터 간(예를 들어, 원하는 민감도 및 특이성을 얻기 위해 선택된) 값으로서 선택될 수 있다. 또 다른 예로, 참조값은 샘플의 통계적 분석 또는 시뮬레이션에 기반하여 결정될 수 있다.

[0031] "*병리 수준*"(또는 장애 수준)은 유기체의 세포 분석을 통해 측정될 수 있는 유기체와 관련된 병리의 양, 정도 또는 중등도를 지칭할 수 있다. 병리학의 또 다른 예는 이식된 장기의 거부이다. 다른 예시적 병리는 게놈 각인 장애, 자가면역 공격(예를 들어, 신장을 손상시키는 루푸스 신염 또는 신경계를 손상시키는 다발성 경화증), 염증성 질환(예를 들어, 간염), 섬유증 과정(예를 들어, 간경화), 지방 침윤(예를 들어, 지방간 질환), 퇴행성 과정(예를 들어, 알츠하이머병) 및 허혈성 조직 손상(예를 들어, 심근경색 또는 뇌졸중)을 포함할 수 있다. 대상체의 건강한 상태는 병리가 없는 분류로 간주될 수 있다.

[0032] "*임신 관련 장애*"는 모체 및/또는 태아 조직에서 유전자의 비정상적인 상대 발현 수준을 특징으로 하는 임의의 장애를 포함한다. 이러한 장애는 자간전증, 자궁내 성장 제한, 침습성 태반, 조산, 신생아 용혈성 질환, 태반 부전, 태아 수종, 태아 기형, HELLP(용혈, 상승된 간 효소 및 낮은 혈소판 수) 증후군, 전신 홍반성 루푸스(SLE), 및 모체의 다른 면역학적 질환을 포함하지만 이에 제한되지 않는다. 일부 구현예에서, 임신 관련 장애는 임신 기간 동안의 생리학적 또는 형태학적 이상과 관련된 임의의 병태이다.

[0033] 약어 "*bp*"는 염기쌍을 지칭한다. 일부 경우에 DNA 단편이 단일 가닥이고 염기쌍을 포함하지 않더라도 "bp"는 DNA 단편의 길이를 표시하기 위해 사용될 수 있다. 단일 가닥 DNA의 맥락에서 "bp"는 뉴클레오티드의 길이를 제공하는 것으로 해석될 수 있다.

[0034] 약어 "*nt*"는 뉴클레오티드를 지칭한다. 일부 경우, "nt"는 염기 단위 내 단일 가닥 DNA의 길이를 표시하기 위해 사용될 수 있다. 또한, "nt"는 상대 위치, 예컨대 분석되는 유전자좌의 상류 또는 하류를 표시하기 위해 사용될 수 있다. 기술적 개념화, 데이터 제시, 처리 및 분석과 관련된 일부 맥락에서, "nt" 및 "bp"는 상호교환적으로 사용될 수 있다.

[0035] 용어 "*서열 맥락*"은 DNA 스트레치의 염기 조성(A, C, G 또는 T) 및 염기 순서를 지칭할 수 있다. 이러한 DNA 스트레치는 염기 변형 분석의 대상 또는 표적인 염기를 둘러쌀 수 있다. 예를 들어, 서열 맥락은 염기 변형 분석의 대상인 염기의 상류 및/또는 하류의 염기를 지칭할 수 있다.

[0036] 용어 "*기계 학습 모델*"은 시험 데이터에 대한 예측을 하기 위해 샘플 데이터(예를 들어, 훈련 데이터)를 사용하는 것에 기반한 모델을 포함할 수 있고, 따라서 지도 학습을 포함할 수 있다. 기계 학습 모델은 종종 컴퓨터 또는 처리장치를 사용하여 개발된다. 기계 학습 모델은 통계 모델을 포함할 수 있다.

[0037] 용어 "*데이터 분석 프레임워크*"는 데이터를 입력으로 취하고 예측된 결과를 출력할 수 있는 알고리즘 및/또는 모델을 포함할 수 있다. "데이터 분석 프레임워크"의 예는 통계 모델, 수학적 모델, 기계 학습 모델, 다른 인공지능 모델 및 이의 조합을 포함한다.

[0038] 용어 "*실시간 시퀀싱*"은 시퀀싱에 관여된 공정 동안 데이터 수집 또는 모니터링이 관여하는 기술을 지칭할 수 있다. 예를 들어, 실시간 시퀀싱은 뉴클레오티드 가닥이 나노포어를 전위할 때 나노포어를 통한 이온 전류의 전기 신호 모니터링이 관여할 수 있다.

[0039] 용어 "*전기 신호*"는 정보를 전달하는 전압 또는 전류를 지칭할 수 있다. 전기 신호는 다양한 규칙적 및/또는 불규칙적 신호 파형 유형 및/또는 형태, 예컨대 사각파, 직사각파, 삼각파, 톱니 파형 또는 다양한 펄스 및 스파

이크로 표현될 수 있다. 전기 신호는 경시적인 전압 또는 전류 변동의 시각적 표현을 포함할 수 있다. 전기 신호의 측정은 특정 시간(예를 들어, 밀리초)에 샘플링될 수 있다. 예를 들어, 전류는 1 kHz, 2 kHz, 3 kHz, 4 kHz, 5 kHz, 10 kHz, 20 kHz, 30 kHz, 40 kHz, 50 kHz, 100 kHz 등의 주파수로 샘플링된다.

[0040] 용어 "*신호 세그먼트*" 또는 "*세그먼트*"는 특정 뉴클레오티드 시퀀싱과 관련된 전기 신호의 추적기록 부분을 지칭할 수 있다. 세그먼트는 나노포어 시퀀싱의 염기 호출로부터 결정된 뉴클레오티드에 대응할 수 있다. 세그먼트는 특정 추적 기간을 포괄할 수 있다. 상이한 세그먼트는 상이한 기간을 가질 수 있다. 세그먼트는 중첩되지 않을 수 있다. 일부 구현예에서 전기 신호 크기는 세그먼트에서 특정 변동을 가질 수 있다. 예를 들어, 전기 신호 크기는 세그먼트의 전기 신호 크기의 평균 또는 중앙값의 5%, 10%, 20%, 30% 또는 40% 이내일 수 있다.

[0041] 용어 "*약*" 또는 "*대략*"은 당업자에 의해 결정된 바와 같은 특정 값에 대해 허용 가능한 오차 범위 이내를 의미할 수 있으며, 이는 부분적으로 값이 측정되거나 결정될 방법, 즉 측정 시스템의 한계에 따라 의존할 것이다. 예를 들어, "약"은 당분야의 실시에 따라 1 또는 1 초과의 표준 편차 이내를 의미할 수 있다. 대안적으로, "약"은 주어진 값의 최대 20%, 최대 10%, 최대 5% 또는 최대 1%의 범위를 의미할 수 있다. 대안적으로, 특히 생물학적 시스템 또는 과정과 관련하여, 용어 "약" 또는 "대략"은 값의 10배 이내, 5배 이내, 더욱 바람직하게는 2배 이내를 의미할 수 있다. 출원 및 청구범위에 특정 값이 기재되는 경우, 달리 명시되지 않는 한 용어 "약"은 특정 값에 대해 허용 가능한 오차 범위 내임을 의미하는 것으로 가정되어야 한다. 용어 "약"은 당업자에 의해 일반적으로 이해되는 의미를 가질 수 있다. 용어 "약"은 ±10%를 지칭할 수 있다. 용어 "약"은 ±5%를 지칭할 수 있다.

[0042] 상세한 설명

[0043] 나노포어 시퀀싱을 사용하여 염기 변형(예를 들어, 메틸화)을 검출하는 정확하고 효율적인 방법이 요망된다. 연구 조사에서는 DNA 메틸화를 분석하기 위해 나노포어 시퀀싱에 의해 생성된 전기 신호를 사용하는 타당성을 연구했다(Simpson 등 Nat Methods. 2017;14:407-410; Liu 등 Nat Commun. 2019;10:2449; Ni 등 Bioinformatics. 2019;35:4586-4595). 5-메틸시토신(5mC) 검출에 대해 보고된 성능은 많은 검증 연구에서 최적 미만이었다. 예를 들어, 샘플 NA12878에 기반한 인간(*H. sapiens*) R9.4 1D 데이터를 분석할 때 DeepSignal이라는 연산 도구를 사용한 5mC 검출의 민감도는 79%, 특이성은 88%로 보고되었다(Ni 등 Bioinformatics. 2019;35:4586-4595). 더 높은 특이성(예를 들어, >95%)을 달성하려는 경우 민감도는 더욱 악화될 것으로 예상된다. nanopolish(Liu 등 Nat Commun. 2019;10:2449)라는 또 다른 도구의 경우, 동일한 데이터세트를 분석할 때 민감도는 0.61에 불과하고 특이성은 0.46이었다. nanopolish 소프트웨어는 하기 가정으로 은닉 마르코프(Markov) 모델에 기반했다: (1) DNA 서열에서 6-뉴클레오티드 올리고머(즉, 6-머)의 전기 신호는 가우스 분포를 따랐으며; (2) 특정 염기에 대한 메틸화 상태(메틸화 또는 비메틸화)의 확률은 이전 염기의 메틸화 상태에만 의존하고; (3) 특정 전류 수준의 출력 확률은 전류 신호를 생성하는 메틸화 상태에만 의존하고 임의의 다른 메틸화 상태나 다른 전류 신호에는 의존하지 않는다. 이러한 가정은 나노포어 시퀀싱 동안 생성된 실제 전류 신호에서 부정확할 수 있으므로 더 낮은 민감도 및 특이성을 야기할 수 있다.

[0044] Oxford Nanopore 시퀀싱에 기반하는 DNA 메틸화 분석을 위한 DeepMod라는 최근 연산 도구에서는 양방향 순환 신경망(RNN)을 사용하려고 시도했다. 그러나 이러한 접근의 설계는 시퀀싱 판독의 예측 결과를 전기 신호로 집계하여 게놈 위치에서 메틸화 수준을 측정하는 것을 목표로 하였으므로 단일 분자 수준에서 메틸화 패턴을 분석하는 능력이 부족하였다. 또한 대장균(*Escherichia coli*), 클라미도모나스 레인하르드티이(*Chlamydomonas reinhardtii*) 및 인간(*Homo sapiens*)을 포함한 데이터세트에 걸친 시퀀싱 깊이 중앙값은 대략 33x였다. 많은 상업적 적용에서, 경제적 비용 및 분석 시간을 절약하기 위해 더 낮은 시퀀싱 깊이가 요망될 것이다. DeepMod 소프트웨어가 단일 분자 수준에서 실질적으로 의미 있는 정확성으로 메틸화 패턴을 분석할 수 있었는지 여부는 알려져 있지 않다.

[0045] 한 연구에서 Yuen 등은 나노포어 시퀀싱으로부터 CpG 메틸화 검출을 위한 도구를 체계적으로 벤치마킹했으며, 대부분의 도구가 CpG 부위당 예상되는 메틸화 백분율과 높은 분산 및 낮은 일치도를 나타낸다고 결론지었다(Yuen 등 bioRxiv. 2020;doi: doi.org/10.1101/2020.10.14.340315).

[0046] Tse 등은 Pacific Biosciences(PacBio)의 단일 분자 실시간 시퀀싱(SMRT-seq)을 사용하여, 광학 신호, 예컨대 DNA 중합 동안 형광단-표지 뉴클레오티드의 혼입에 의해 생성된 인터펄스 지속기간(IPD) 및 펄스 폭(PW)을 포함하는 DNA 중합효소의 동역학적 특징이 컨볼루션 신경망을 사용하여 하나 초과의 염기로 구성된 분석 측정 윈도우에 기반하여, 메틸화 및 비메틸화 CpG 부위를 구별하는 데 사용될 수 있음을 보고하였다(Tse 등 Proc Natl Acad Sci USA. 2021;118: e2019768118; U.S. 특허 번호 11,091,794). 이러한 측정 윈도우는 IPD 및 PW를 상이

한 시퀀싱 맥락 및 시퀀싱 위치로 조직한다. 그러나 나노포어 시퀀싱은 나노포어를 통과하는 이중 가닥 DNA 가닥에 의해 유발되는 전류 신호에 따라 완전히 상이한 시퀀싱 메커니즘을 사용했다. 이러한 미가공 전기 신호는 나노포어를 통과하는 상이한 뉴클레오티드에 따라 달라지며, 특정 뉴클레오티드의 전기 신호는 해당 뉴클레오티드 근처의 상류 및 하류 뉴클레오티드에 의해 영향을 받을 것이다. 따라서, 상이한 뉴클레오티드는 검출된 상이한 길이의 전기 신호 추적기록을 가질 것이며, 심지어 동일한 뉴클레오티드라도 상이한 길이의 전기 신호 추적기록을 가질 것이다. 특정 뉴클레오티드 또는 나노포어를 통과하는 하나 초과의 뉴클레오티드와 관련된 전기 신호를 분석할 때, 각 염기에서 검출된 전기 신호 추적기록의 길이는 시간이 지나도 고정되지 않는다. 대조적으로, PacBio SMRT-seq를 사용한 5mC 검출에 대한 이전 연구는 각 뉴클레오티드에 대한 광학 신호와 관련된 2개의 고정된 측정, 즉 IPD 및 PW에 기반하였다(Tse 등 Proc Natl Acad Sci USA. 2021;118: e2019768118). 따라서 Tse 등의 연구(Tse 등 Proc Natl Acad Sci USA. 2021;118: e2019768118)에 제시된 훈련된 모델은 나노포어 시퀀싱에 의해 생성된 이러한 전기 신호에 적용할 수 없다.

[0047] 본원에 기재된 구현예는 나노포어 시퀀싱으로부터 얻은 전기 신호를 사용하여 뉴클레오티드 변형을 검출한다. 뉴클레오티드 변형은 본원에 기재된 임의의 메틸화를 포함할 수 있다. 나노포어 시퀀싱으로 얻은 정보는 뉴클레오티드의 정체, 표적 위치에 대한 뉴클레오티드의 위치, 뉴클레오티드에 대응하는 전기 신호의 세그먼트에 대한 통계값, 및 핵산 분자 영역의 윈도우에서 전기 신호의 통계값을 포함하는 벡터를 포함할 수 있다.

[0048] 본 개시내용에 제시된 구현예는 유기체로부터 얻은 세포 샘플(예를 들어, 세포주, 고형 장기, 고형 조직, 내시경을 통해 얻은 샘플, 융모막 융모 샘플)로부터 얻은 DNA에 대해 사용될 수 있다. 본 개시내용의 구현예는 환경(예를 들어, 박테리아, 세포 오염물질), 식품(예를 들어, 고기)에서 얻은 세포성 샘플에 대해서도 사용될 수 있다. 본 개시내용에 제시된 구현예는 임산부에서 얻은 혈장 또는 혈청에 대해서도 사용될 수 있다. 일부 구현예에서, 본 개시내용에 제시된 방법은 또한 예를 들어 혼성화 탐침(Albert 등, 2007; Okou 등, 2007; Lee 등, 2011), 또는 물리적 분리에 기반한(예를 들어, 크기 등에 기반한) 또는 제한 효소 분해(예를 들어, Msp I) 또는 Cas9 기반 농축(Watson 등, 2019) 후 접근을 사용하여 게놈의 일부가 먼저 농축되는 단계 후에 적용될 수 있다. 본 발명은 작동하기 위해 효소적 또는 화학적 전환을 필요로 하지 않지만, 특정 구현예에서는 이러한 전환 단계가 본 발명의 성능을 추가 향상시키기 위해 포함될 수 있다.

[0049] 본 개시내용의 구현예는 변형된 염기를 정확하고 효율적으로 검출할 수 있도록 나노포어 시퀀싱을 개선한다. 염기 변형은 직접적으로 검출될 수 있다. 구현예는 검출을 위한 모든 변형 정보를 보존하지 못할 수 있는, 효소적 또는 화학적 전환을 피할 수 있다. 또한 특정 효소적 또는 화학적 전환은 특정 유형의 변형과 호환되지 않을 수 있다. 본 개시내용의 구현예는 또한 염기 변형 정보를 PCR 생성물에 전달하지 않을 수 있는, PCR에 의한 증폭을 피할 수 있다. 추가로, DNA의 두 가닥 모두 함께 시퀀싱될 수 있으며, 이로써 다른 가닥에 대한 상보적 서열을 갖는 한 가닥의 서열이 쌍을 이룰 수 있게 한다. 대조적으로, PCR 증폭은 이중 가닥 DNA의 두 가닥을 분할하므로 두 구성 가닥으로부터의 서열을 조합 분석하는 것이 어렵다.

[0050] 또한, 나노포어 시퀀싱은 다른 시퀀싱 기술보다 더 비용 효과적이고 휴대성이 뛰어나다. 예를 들어, 나노포어 시퀀싱 시스템인 Oxford Nanopore Technologies MinION™은 대략 5,000 USD인 반면, 광학 신호 기반 시퀀싱 시스템인 PacBio SMRT™ Sequel II 시스템은 500,000 내지 700,000 USD 수준이다. 나노포어 시퀀싱 속도는 초당 약 450개 뉴클레오티드인 반면, PacBio SMRT™ 시퀀싱은 초당 약 5개 뉴클레오티드이다. 따라서 동일한 시기 내에 나노포어 시퀀싱은 광학 신호 기반 시퀀싱 시스템보다 많은 데이터를 얻을 수 있다.

[0051] 효소적 또는 화학적 전환을 갖거나 갖지 않고 결정된 메틸화 프로파일은 생물학적 샘플을 분석하기 위해 사용될 수 있다. 한 구현예에서, 메틸화 프로파일은 세포성 DNA(예를 들어, 모체 또는 태아, 조직 또는 바이러스)의 기원을 검출하기 위해 사용될 수 있다. 조직의 비정상적인 메틸화 프로파일의 검출은 개인의 발달 장애 식별을 돕는다. 단일 분자의 메틸화 패턴은 키메라(예를 들어, 바이러스와 인간 사이) 및 하이브리드 DNA(예를 들어, 천연 게놈에서 일반적으로 융합되지 않은 두 유전자 사이); 또는 두 종 사이(예를 들어, 유전적 또는 게놈 조작을 통해)를 식별할 수 있다.

[0052] I. 나노포어 시퀀싱 원리

[0053] 단일 분자 시퀀싱 기술의 예는 나노포어 시퀀싱(Oxford Nanopore Technologies)이다. **도 1**은 DNA 분자(예를 들어, DNA 분자(104))의 나노포어 시퀀싱에 대한 원리를 나타낸다. 단일 DNA 분자가 나노미터 크기의 포어를 통과함에 따라, 막을 통한 이온 전류 흐름에 의해 유발되는 전기 신호 패턴이 핵산의 서열을 결정하기 위해 사용되었다. 이러한 포어는, 예를 들어 단백질(예를 들어, 알파-헤모라이신, 에어로라이신 및 마이코박테리움 스메그마티스(*Mycobacterium smegmatis*) 포린 A(MspA)) 또는 합성 물질, 예컨대 실리콘 또는 그래핀(Magi 등, Brief

Bioinform. 2018;19:1256-1272)에 의해 작제될 수 있지만 이에 제한되지 않는다.

[0054] 한 구현예에서, 이중 가닥 DNA 분자는 말단 복구 과정을 거쳤다. 이러한 과정은 DNA를 무딘 말단 DNA로 전환한 다음, 시퀀싱 어댑터 결찰을 촉진하는 A 꼬리를 부가할 것이다. 각각 모터 단백질(즉, 모터 어댑터)(예를 들어, 모터 단백질(108))을 운반하는 시퀀싱 어댑터는 DNA 분자의 양쪽 말단에 결찰된다. 시퀀싱 과정은 모터 단백질(예를 들어, 모터 단백질(112))이 이중 가닥 DNA를 풀어 제1 가닥이 나노포어를 통과할 수 있도록 하면서 시작된다. DNA 가닥이 나노포어(116)를 통과할 때 센서(예를 들어, 전극)는 서열 맥락뿐만 아니라 관련 염기 변형(1-차원(1D) 판독이라고 함)에 따라 경시적인(밀리초, ms) 피코암페어(pA)의 이온 전류 변화를 측정한다. 그래프(120)는 예시적 전류 신호 대 시간을 나타낸다. 또 다른 구현예에서, 헤어핀 서열 어댑터는 이중 가닥 DNA 분자에 대해 제1 가닥 및 그 상보적 가닥을 함께 공유적으로 묶기 위해 사용될 것이다. 따라서 시퀀싱 동안 이중 가닥 DNA 분자의 한 가닥이 시퀀싱된 다음 상보적 가닥이 시퀀싱되며($1D^2$ 또는 2차원(2D) 판독이라고 함), 이는 시퀀싱 정확성을 잠재적으로 개선할 수 있었다. 또 다른 구현예에서, 단백질에 의해 묶인 이중 가닥 DNA 분자의 한쪽 말단은 동일한 분자의 제1 가닥의 시퀀싱이 완료된 후 상보적 가닥의 시퀀싱 가능성을 증가시켜 $1D^2$ 판독을 생성할 것이다.

[0055] 미가공 신호(예를 들어, 그래프(120)의 전류)가 염기 호출 및 염기 변형 분석을 위해 사용된다. 일부 구현예에서, 염기 호출 및 염기 변형 분석은 기계 학습 접근, 예를 들어, 그러나 비제한적으로 순환 신경망(RNN), 컨볼루션 신경망(CNN), 은닉 마르코프 모델(HMM), 또는 이의 하나 이상의 조합에 의해 수행된다.

[0056] 한 구현예에서, 본 발명자들은 나노포어 시퀀싱에 의해 생성된 전류 신호를 처리하는 새로운 방법을 개발했으며, 처리된 신호는 컨볼루션 신경망(CNN) 또는 순환 신경망(RNN)에 기반하여 단일 분자 수준으로 DNA 메틸화의 결정을 위해 분석되었다.

[0057] II. 전류 신호 분석

[0058] 나노포어 시퀀싱으로부터의 전류 신호가 분석되어 염기 변형을 식별할 수 있다. 그러나, 도 1에 기재된 기계 학습 접근은 나노포어를 이용하여 얻은 미가공 전류의 입력만을 사용하지는 않는다. 본원에 기재된 구현예는 전류 부분의 하나 이상의 통계값을 사용한다. 이러한 하나 이상의 통계값의 벡터는 뉴클레오티드의 정체 및 뉴클레오티드의 위치를 포함하는, 뉴클레오티드의 윈도우에 대응하는 다른 정보와 조합될 수 있다. 뉴클레오티드의 위치는 윈도우 내의 표적 위치에 관한 것일 수 있으며, 표적 위치는 변형 또는 이의 부재가 검출되는 위치이다. 뉴클레오티드 윈도우에 대한 정보는 핵산 분자 영역의 전기 신호에 대한 통계값과 함께 포함되어 입력 데이터 구조를 형성할 수 있다. 이러한 입력 데이터 구조에 대해 훈련된 모델이 염기 변형을 검출하기 위해 사용될 수 있다.

[0059] A. 전류 벡터 매개변수

[0060] 나노포어를 통과하는 뉴클레오티드 가닥의 경우, *N*개의 이벤트(즉, 식별된 상이한 뉴클레오티드와 관련된 신호 세그먼트)를 검출할 것 이다. 한 구현예에서, 하나의 이벤트는 특정 시간 단위(예를 들어, 밀리초)에 샘플링된 일련의 전기 신호를 사용하여 염기 호출 동안 식별된 하나의 뉴클레오티드에 대응한다. 한 예에서, 전류는 4 kHz의 주파수에서 샘플링되었다(Rang 등 Genome Biol. 2018;19:90). 또 다른 구현예에서, 하나의 이벤트는 특정 시간 속도로 샘플링된 일련의 전기 신호를 사용하여 염기 호출 동안 식별된 하나 초과의 뉴클레오티드에 대응한다.

[0061] **도 2**는 전류 신호의 그래프를 나타낸다. 전류 크기는 y축의 피코암페어이다. 밀리초 단위의 시간은 x축에 있다. 점(예를 들어, 점(204))은 개별 신호 측정을 나타낸다. 인접한 점을 통과하는 선(예를 들어, 선(208))은 뉴클레오티드(예를 들어, 선(208)의 A)와 관련된 신호 측정의 신호 세그먼트를 나타낸다. 이벤트 i에 대해 m_i개의 전류 신호가 있다고 가정하면, 이벤트 i에 대한 전류 신호 j의 크기는 P_{ij}로 표시되었다. 한 구현예에서, 뉴클레오티드의 경우, ***X1, X2, X3, X4*** 및 ***X5***를 포함하는 신호 특징 벡터가 해당 뉴클레오티드와 관련된 전기 신호의 패턴을 특성규명하기 위해 사용된다. ***X1, X2*** 및 ***X3***에 대한 정의가 도 2에 예시된다. ***X1***은 P_{ij}의 평균이다. ***X2***는 P_{ij}의 표준 편차이다. ***X3***은 P_{ij}의 중앙값이다. ***X4***는 ***X3***으로부터의 전류의 절대 편차의 중앙값이다(도 2에는 하나의 절대 편차만 표지됨). ***X5***는 표준 편차로 나눈 전류 신호의 평균으로부터의 ***X1***의 차이이다. ***X5***는 세그먼트의 전류 신호의 z-점수로 간주될 수 있다.

[0062] 한 구현예에서, P_{ij}는 정규화된 신호일 수 있다. 정규화에는 부분 또는 전체 뉴클레오티드 가닥에 관한 최소 및

최대 값을 사용하여 정규화된 신호값이 0과 1 범위 내에 있도록 원래 범위에서 전류 신호를 재조정하는 것이 관여할 수 있다. 정규화에는 정규화된 신호값의 평균이 0이고 표준 편차가 1이도록 전류 신호를 재조정하는 것이 관여할 수 있다. 정규화에는 부분 또는 전체 뉴클레오티드 가닥과 관련된 중앙값 및 편차를 사용하여 전류 신호를 재조정하는 것이 관여할 수 있다.

[0063] *X1* 및 *X2*는 이벤트 *i*와 관련된 P_{ij}의 평균 및 표준 편차를 나타낸다.

[0064] *X1*은 다음에 의해 정의된다:

[0065] $$X1 = \frac{\sum_{j=1}^{j=m_i} P_{ij}}{m_i}$$

[0066] *X2*는 다음에 의해 정의된다:

[0067] $$X2 = \sqrt{\frac{\sum_{j=1}^{j=m_i} (P_{ij} - X1)^2}{m_i - 1}}$$

[0068] *X3*은 다음에 의해 정의된다:

[0069] ***X3 = 중앙값(P_{ij}),***

[0070] 식 중, *i*는 염기 변형 분석을 위한 조사 염기 주위의 이벤트(예를 들어, CpG 부위의 메틸화)를 포함하여 *l*내지 *r*의 범위이다. 변수 *l* 및 *r*은 일련의 이벤트 윈도우(뉴클레오티드 서열에 대응)의 왼쪽 및 오른쪽을 나타낸다. *l*과 *r* 간 뉴클레오티드 서열은 일반적으로 아래에 논의된 전류 신호 패턴의 통합 제시 행렬(IPM으로 지칭됨)보다 길어야 한다. 주어진 이벤트 *i*에 대해 *j*는 1 내지 m_i의 범위이다. *X3*은 모든 세그먼트를 결정하는 데 사용되는 전류 신호 중앙값일 수 있다. *X3*은 단 하나의 세그먼트 초과의 전류를 사용하여 결정되므로 *X3*은 모든 세그먼트에 대해 동일한 값일 수 있다. 일부 구현예에서, *X3*은 특정 윈도우에 대한 것일 수 있다. 다른 구현예에서, *X3*은 여러 윈도우에 걸친 중앙값일 수 있다.

[0071] *X4*는 다음에 의해 정의된다:

[0072] ***X4* = 중앙값(| P_{ij} - *X3* |)**

[0073] 식 중, | · |는 절대값을 나타내고, *i*는 염기 변형 분석을 위한 조사 염기 주위의 이벤트(예를 들어, CpG 부위의 메틸화)를 포함하여 *l* 내지 *r*의 범위이다. 주어진 *i*에 대해 *j*는 1 내지 m_i의 범위이다. *X4*는 모든 세그먼트를 결정하는 데 사용되는 전류 신호의 절대 편차의 중앙값일 수 있다. *X4*는 단 하나의 세그먼트 초과의 전류를 사용하여(예를 들어, 샘플링된 모든 전류값을 사용하여) 계산될 수 있으므로 모든 세그먼트에 대해 동일한 값일 수 있다.

[0074] *X5*는 다음에 의해 정의된다:

[0075] $$X5 = \frac{X1-\mu}{\sigma},$$

[0076] 식 중, $\mu = \frac{\sum_{i=l}^{i=r}\sum_{j=1}^{j=m_r} P_{ij}}{M-1}$ 이며, $\sigma = \sqrt{\frac{\sum_{i=l}^{i=r}\sum_{j=1}^{j=m_r}(P_{ij}-\mu)^2}{M-1}}$ 이고,

[0077] i는 염기 변형 분석을 위한 조사 염기 주위의 이벤트(예를 들어, CpG 부위의 메틸화)를 포함하여 l 내지 r의 범위이다. 주어진 i에 대해 j는 1 내지 m_i의 범위이다. M은 l 내지 r 범위의 이벤트에 대해 샘플링된 전류 신호의 총 수이다. 복수의 전류 신호와 관련되고 X3을 결정하는 데 사용되는 영역의 크기는 DNA 단편의 크기일 수 있다. 예를 들어, DNA 단편이 500 bp인 경우, 영역의 크기는 500이다. 단편이 300 bp인 경우, 영역의 크기는 300이다. 일부 구현예에서 DNA 단편을 X3을 결정하기 위한 더 작은 하위단편으로 추가 분할하는 것이 유용할 수 있다. X3을 결정하는 데 사용되는 영역의 크기는 5 nt, 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, 60 nt, 70 nt, 90 nt, 100 nt, 200 nt, 300 nt, 400 nt, 500 nt, 600 nt, 800 nt, 900 nt, 1 kb, 2 kb, 3 kb, 4 kb, 5 kb, 10

kb, 50 kb 등일 수 있다.

[0078] *X1* 및 *X2*는 이벤트 *i* 내의 신호 변화를 반영하기 위해 사용되어, 각 뉴클레오티드에 대한 전기 신호의 국소 패턴을 나타낼 수 있다. *X3, X4* 및 *X5*는 *l* 내지 *r* 범위의 다른 주위 이벤트와 관련된 이벤트 *i*의 신호 변화를 반영하기 위해 사용될 수 있다. 일부 구현예에서, 주위 이벤트는 염기 변형 분석을 위한 조사 염기의 X-nt 상류 및 Y-nt 하류일 수 있다. X는 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000 및 10000을 포함할 수 있지만 이에 제한되지 않으며; Y는 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000 및 100000을 포함할 수 있지만 이에 제한되지 않는다. 한 구현예에서, 주위 이벤트는 나노포어를 통과하는 전체 뉴클레오티드 가닥일 수 있다.

[0079] B. 단일 가닥 분석

[0080] **도 3**은 전류 신호의 그래프를 나타낸다. 전류 크기는 y축의 피코암페어이다. 밀리초 단위의 시간은 x축에 있다. 추적기록(304)은 경시적인 전류 크기가다. 신호 세그먼트(예를 들어, 세그먼트(308))는 뉴클레오티드와 관련된 추적기록(304)의 일부이다. 전류 변화는 나노포어를 통과하는 상이한 뉴클레오티드에 따라 달라질 것이다. 나노포어 시퀀싱의 염기 호출은 일반적으로 전류 신호를 상이한 국소 고정 상태(즉, 이벤트)로 변환하는 데 의존한다. 전류 신호를 상이한 이벤트로 변환하는 과정을 전기 신호 분할이라고 한다. 이온 전류 변화는 신호 세그먼트의 하나 이상의 뉴클레오티드에 대응하는 이벤트의 크기(예를 들어, 피코암페어, pA로 측정됨), 이온 전류의 방향, 신호 세그먼트의 하나 이상의 뉴클레오티드에 대응하는 전류 이벤트의 기간, 이온 전류의 변화 속도, 및 상이한 신호 세그먼트에 걸친 상대 크기를 포함하지만 이에 제한되지 않는다. 크기는 전류의 강도 또는 세기를 지칭할 수 있으며 교류를 의미할 필요는 없다. 이러한 전류 이벤트는 예를 들어 Tombo(Stoiber 등 bioRxiv. 2016; doi.org/10.1101/094672)라는 소프트웨어를 사용하여 상이한 염기에 할당된다. 하나의 뉴클레오티드는 상이한 크기를 갖는 일련의 이벤트와 관련될 것이다. 이러한 도구(Tombo)는 이러한 염기가 만 휘트니(Mann-Whitney) U-시험에 기반하여 변형되었는지 여부를 추정하기 위해 2개 샘플 간 게놈 염기에 할당된 나노포어 신호의 차이를 시험하고자 시도하였다(Stoiber 등 bioRxiv. 2016; doi.org/10.1101/094672). 이 도구(Tombo)는 상류 및 하류 신호 그리고 서열 맥락을 고려하지 않았으며, 상이한 서열 판독으로부터의 모든 신호가 게놈 염기로 집계되었으므로 단일 분자 수준에서 메틸화 패턴을 분석할 수 없었다. Tombo의 성능은 다른 도구, 예컨대 Nanopolish 및 DeepSginal의 성능과 비교되었다(Yuen 등 bioRxiv. 2020;doi: doi.org/10.1101/2020.10.14.340315).

[0081] 한 구현예에서, 뉴클레오티드와 관련된 신호 세그먼트 내의 전류 패턴을 특성규명하기 위해, 해당 신호 세그먼트 내 이벤트의 전류 크기의 평균(X1) 및 표준 편차(X2)가 계산된다. 전체 분자와 관련된 이벤트의 전류 크기의 중앙값(X3) 및 전체 분자와 관련된 이벤트의 전류 크기의 절대 편차 중앙값(X4)이 결정된다. 신호 세그먼트에 대해 정규화된 신호(X5)는 아래 공식에 의해 결정된다:

[0082] $X5 = \frac{X1-\mu}{\sigma}$,

[0083] 식 중, *X*1은 해당 뉴클레오티드와 관련된 해당 신호 세그먼트 내 이벤트의 전류 크기의 평균이고; μ는 조사 중인 전체 분자 내 이벤트의 전류 크기의 평균이고; σ는 조사 중인 전체 분자 내 이벤트의 전류 크기의 표준 편차이다. 한 구현예에서, 평균 및 표준 편차는 가장 큰 값 및 가장 작은 값의 소규모 지정 백분율의 제거 후 유도될 수 있다.

[0084] 뉴클레오티드의 경우, X1, X2, X3, X4 및 X5를 포함하는 신호 특징 벡터가 해당 뉴클레오티드와 관련된 전기 신호의 패턴을 반영하기 위해 사용된다. 예를 들어, 세그먼트(308)는 [X1, X2, X3, X4, X5]의 신호 특징 벡터를 가질 수 있다.

[0085] X1 및 X2는 신호 세그먼트 *i* 내 이벤트의 전류 크기의 평균 및 표준 편차를 나타낸다. X3은 전체 분자와 관련된 이벤트의 전류 크기의 중앙값을 나타낸다. X4는 전체 분자와 관련된 이벤트의 전류 크기의 절대 편차 중앙값을 나타낸다. X5는 신호 세그먼트 *i*에 대한 정규화된 신호를 나타낸다.

[0086] **도 4**는 신호 세그먼트 길이의 주파수의 그래프이다. 뉴클레오티드와 관련된 전류 이벤트의 길이(즉, 밀리초 단위의 기간)는 x축에 있다. 길이의 주파수는 y축에 나타낸다. 도 4는 뉴클레오티드와 관련된 각 신호 세그먼트의

길이가 중앙값 9(범위: 1 - 3540)로 가변적이었음을 나타낸다.

[0087] 염기 변형은 그 상류 및 하류 뉴클레오티드와 관련된 전기 신호에 영향을 미칠 것이다. 본 개시내용에서, 본 발명자들은 성능을 개선하기 위해 염기 변형 분석을 위한 뉴클레오티드와 관련된 전류 신호, 관심 뉴클레오티드 근처의 뉴클레오티드와 관련된 전류 신호뿐만 아니라 시퀀싱 맥락을 종합적으로 활용했다. CpG 부위의 DNA 메틸화(즉, 시토신의 5번째 탄소에서의 메틸화)는 척추동물 게놈에서 가장 일반적인 유형의 염기 메틸화이다. CpG 부위에서의 DNA 메틸화 분석은 본 개시내용에 대한 예시적 예로 사용되었다.

[0088] **도 5**는 나노포어 시퀀싱을 통해 한 가닥으로부터의 전류 신호를 사용하여 메틸화를 결정하는 과정을 나타낸다. 블록(504)에서 이중 가닥 DNA 분자가 제공된다. 블록(508)에서, 이중 가닥 DNA 분자는 나노포어 시퀀싱에 적합한, 시퀀싱 어댑터와 결찰된다. 블록(512)에서, 나노포어 시퀀싱이 수행된다. 단일 이중 가닥 분자의 한 가닥이 막에 포매된 포어를 통해 이동하여 나노포어를 통해 흐르는 이온 전류 신호를 변경했다. 블록(516)에서, 전류 신호가 얻어진다. 이온 전류 신호는 예를 들어 트랜스 전극에 의해 측정될 수 있다.

[0089] 전류 신호는, 예를 들어 Tombo(Stoiber 등 bioRxiv. 2016; doi.org/10.1101/094672)를 사용하여 분할 단계에 의해 처리될 것이다. 이러한 분할된 전기적 이벤트는 상이한 뉴클레오티드에 할당될 것이다. 블록(520)에서 통합 제시 행렬(IPM)이 작제된다. IPM은 각 염기에 대한 전류 신호, 시퀀싱 맥락, 및 염기 변형 분석을 위한 유전자좌 근처 또는 주위의 일련의 뉴클레오티드에 걸친 시퀀싱 위치 정보를 포함하는 전류 신호 패턴의 행렬이다. 한 구현예에서, 뉴클레오티드와 관련된 분할된 전기적 이벤트는 신호 특징 벡터, 즉 [X1, X2, X3, X4, X5]에 의해 기재되었다. CpG 부위 내의 시토신 및 예를 들어, 다수의 신호 특징 벡터와 함께 해당 시토신의 10-nt 상류 및 하류(즉, 총 21 nt)가 전류 신호 패턴의 IPM을 형성하기 위해 사용되었다. 예시적 목적을 위해 5'-T[CCATGC]CATCGTC[GATGCA]G-3'의 21-nt 서열을 예로 사용하여 IPM 524를 생성하였다. 괄호 안의 염기는 단순화를 위해 생략되었다("…"로 표시). 아데닌 염기("A")에 대응하는 -2 위치의 경우 "A"와 관련된 신호 특징 벡터 [X1 =1.7, X2 =0.29, X3 =24.2, X4 =436, X5 =-0.3]는 "-2" 열과 "A" 행 사이의 해당 셀에 채워졌다. 같은 열의 다른 셀은 "0"으로 채워졌다. 21-nt 서열 맥락과 관련된 각 뉴클레오티드에 대한 나머지 신호 특징 벡터는 동일한 규칙을 사용하여 채워져 21-nt IPM을 형성했다. 따라서 이러한 IPM은 전류 신호 패턴, 시퀀싱 맥락, 시퀀싱 위치뿐만 아니라 경시적으로 변하는 패턴을 동시에 인코딩할 것이다. 메틸화 및 비메틸화 DNA 데이터세트에서 유래한 다수의 IPM이 CNN 또는 RNN 모델을 훈련하는 데 사용되었으며 이후 시험 샘플의 CpG 부위에서 메틸화 상태를 결정하기 위해 사용되었다.

[0090] 블록(528)은 CNN 분석을 나타낸다. CNN 분석을 위해 IPM이 입력 레이어에 공급되고, 컨볼루션 레이어 및 출력 레이어의 공정이 이어졌다. CpG에 대한 메틸화 확률(즉, 출력 메틸화 점수, 0 내지 1 범위)은 출력 레이어에서 시그모이드 함수에 기반하여 결정되었다. 이 접근은 IPM-CNN으로 지칭된다. 한 구현예에서, 메틸화 CpG 부위(M.SssI-처리 DNA) 및 비메틸화 CpG 부위(WGA(전체 게놈 증폭(WGA) DNA)에 대한 IPM이 CNN 모델을 훈련하는 데 사용되었다. M.Sss 처리 DNA에서 유도된 데이터세트에서 CpG 부위에 대한 메틸화의 표적값은 "1"로 정의된 반면, WGA DNA에서 유도된 데이터세트에서 CpG 부위에 대한 메틸화의 표적값은 "0"으로 정의되었다. IPM-CNN의 최적 매개변수는 모델 매개변수를 반복적으로 업데이트하여 시그모이드 함수에 의해 계산된 출력 점수와 원하는 표적 출력(이진값: 0 또는 1) 간 전체 예측 오차를 최소화함으로써 얻어졌다. 전체 예측 오차는 딥 러닝 알고리즘(keras.io/)의 시그모이드 교차 엔트로피 손실 함수에 의해 결정되었다. 훈련 데이터세트에서 학습된 모델 매개변수는 시험 데이터세트의 메틸화 상태를 분석하여 CpG 부위가 메틸화될 가능성을 제시하는 확률 점수(즉, 메틸화 확률)를 출력하는 데 사용되었다. 한 구현예에서 CNN 모델은 커널 크기가 25인 32, 64, 128, 256개의 필터를 각각 갖는 4개의 2차원(2D) 컨벌루션 레이어를 활용했다. 정류 선형 장치(ReLU)의 활성화 함수가 이러한 컨볼루션 레이어에 사용되었다. 이후에 배치 정규화 레이어가 적용되었다. 평탄화된 레이어가 더 추가되고, 드롭아웃 비율이 0.5인 드롭아웃 레이어에 이어, ReLU 활성화 함수를 사용하여 200개의 뉴런을 포함하는 완전 연결 레이어가 추가되었다. 시그모이드 활성화 함수로 하나의 뉴런을 갖는 출력 레이어가 최종적으로 적용되어 메틸화되는 CpG 부위에 대한 확률 점수(즉, 메틸화 확률)를 산출하였다. CNN 모델을 위한 프로그램은 Keras 딥 러닝 프레임워크(https://keras.io/)에 기반하여 구현되었다.

[0091] 블록(532)은 RNN 분석을 나타낸다. RNN 분석을 위해 IPM이 입력 레이어에 공급되고 장단기 메모리(LSTM) 레이어 및 출력 레이어의 공정이 이어졌다. CpG에 대한 메틸화 확률(0 내지 1 범위)은 출력 레이어의 시그모이드 함수에 기반하여 결정되었다. 이 접근은 IPM-RNN으로 지칭된다. IPM-RNN에서 사용된 것과 유사한 훈련 절차를 사용하여, 모델 매개변수를 반복적으로 업데이트하여 시그모이드 함수에 의해 계산된 출력 점수와 원하는 표적 출력(이진값: 0 또는 1) 간 전체 예측 오차를 최소화함으로써 IPM-RNN의 최적 매개변수를 얻었다. 훈련 데이터세트에서 학습된 모델 매개변수가 시험 데이터세트의 메틸화 상태를 분석하여 CpG 부위가 메틸화될 가능성을 제시하

는 확률 점수(즉, 메틸화 확률)를 출력하는 데 사용되었다. 한 구현예에서, LSTM 장치를 갖는 RNN 모델이 각각 256개의 은닉 노드를 갖는 2개의 완전 연결 은닉 레이어와 함께 사용되었다. 마지막 레이어에는 드롭아웃 비율이 0.2인 드롭아웃 레이어가 뒤따랐다. 시그모이드 활성화 함수로 하나의 뉴런을 갖는 출력 레이어가 최종적으로 적용되어 CpG 부위가 메틸화되는 확률 점수(즉, 메틸화 확률)를 산출하였다. CNN 모델을 위한 프로그램은 Keras 딥 러닝 프레임워크(keras.io/)에 기반하여 구현되었다.

[0092] C. 이중 가닥 분석

[0093] **도 6**은 나노포어 시퀀싱을 통해 두 DNA 가닥 모두의 전류 신호를 사용하여 메틸화를 결정하는 과정을 나타낸다. 한 구현예에서, 제2 뉴클레오티드 가닥(상보적 가닥, 또는 크릭 가닥으로 지칭됨)이 제1 뉴클레오티드 가닥(왓슨 가닥으로 지칭됨)의 동일한 나노포어 통과 완료에 바로 뒤따르는 방식으로 이중 가닥 DNA 분자가 시퀀싱될 때, 이중 가닥 DNA의 두 뉴클레오티드 가닥 모두로부터 전류 신호가 얻어질 수 있다. 동일한 나노포어에서 이중 가닥 DNA의 두 뉴클레오티드 가닥을 모두 순차적으로 시퀀싱하는 기술은 $1D^2$ 또는 2D 시퀀싱으로 지칭된다. 블록(604)에서 이중 가닥 DNA 분자가 제공된다. 블록(608)에서, 이중 가닥 DNA 분자는 나노포어 시퀀싱에 적합한, 시퀀싱 어댑터와 결찰된다. 블록(612)에서, 단일 이중 가닥 분자의 한 가닥이 막에 포매된 포어를 통해 이동한 다음, 상보적 가닥이 이동했다. 블록(616)에서, 각 이중 가닥 DNA 분자의 두 가닥 모두에 대해 전류 신호가 얻어진다. 이온 전류 신호는 트랜스 전극에 의해 측정될 수 있다. 얻어진 전류 신호가 Guppy(Oxford Nanopore Technologies Ltd)를 사용하여 시퀀싱(즉, 염기 호출)된 DNA 분자의 뉴클레오티드 정보를 추론하기 위해 사용되었다. 일부 구현예에서, Albacore(nanoporetech.com/), WaveNano(Wang 등 Quantitative Biology. 2018;6:359-368), Chiron(Teng 등 GigaScience. 2018;7:giy037), Flappie(github.com/nanoporetech/flappie), Scrappie(github.com/nanoporetech/scrappie) 등을 포함하지만 이에 제한되지 않는 다른 염기 호출 도구가 사용될 수 있었다.

[0094] 특정 시간 속도(예를 들어, 밀리초)로 샘플링된 전류 신호는 염기 변형 분석을 위해 상이한 검출된 뉴클레오티드에 할당될 것이다. 전류 신호는 예를 들어 Tombo (Stoiber 등 bioRxiv. 2016; doi.org/10.1101/094672)를 사용하여 분할 단계에 의해 처리될 것이다. 이러한 분할된 전기적 이벤트는 상이한 뉴클레오티드에 할당될 것이다. 블록(620)에서, 각 이중 가닥 DNA 분자로부터의 두 가닥을 모두 포함하는 통합 제시 행렬(IPM)이 작제된다. 한 구현예에서, 뉴클레오티드와 관련된 분할된 전기적 이벤트가 신호 특징 벡터, 즉 [*X1, X2, X3, X4, X5*]에 의해 기재되었다. 상보적 가닥의 대응하는 염기로부터의 신호 특징 벡터, 즉 [*X1', X2', X3', X4', X5'*]가 얻어졌다. CpG 부위 내의 시토신 및 예를 들어, 다수의 신호 특징 벡터와 함께 해당 시토신의 10-nt 상류 및 하류(즉, 총 21 nt)가 전류 신호 패턴의 IPM을 형성하기 위해 사용되었다. 동일한 이중 가닥 DNA 분자의 상보적 가닥에서 대응하는 염기로부터의 IPM이 얻어졌다. 왓슨 및 크릭 가닥에서 유도된 IPM이 조합되어 염기 변형 분석을 위한 더 고차원의 새로운 IPM 행렬을 형성했다.

[0095] 일부 구현예에서, NanoMod(Liu 등 BMC Genomics. 2019;20:78), Albacore(nanoporetech.com/), Chiron(Teng 등 GigaScience. 2018;7:giy037), Nanopolish(Simpson 등 Nat Methods. 2017;13:407-410), Scrappie(https://github.com/nanoporetech/scrappie), UNCALLED(Kovaka 등 Nat Biotechnol. 2020; doi:10.1038/s41587-020-0731-9) 등을 포함하는 다른 연산 도구가 상이한 뉴클레오티드에 전류 신호를 할당하는 데 사용될 수 있었다. 이중 가닥 분석을 위해 기재된 이러한 연산 도구 및 다른 기술이 단일 가닥 분석을 위해 사용될 수 있다.

[0096] 예시 목적으로, 5'-T[CCATGC]CATCGTC[GATGCA]G-3'의 21-nt 서열이 IPM 624에 대한 기준으로서 일례로 사용되었다. IPM 624는 IPM 524와 유사할 수 있지만 왓슨 및 크릭 가닥을 둘 모두 포함한다. 단순화를 위해 괄호 안의 염기는 생략되었다("..."로 표시됨). 왓슨 가닥의 아데닌 염기("A")에 대응하는 -2 위치의 경우 "A"와 관련된 신호 특징 벡터, 즉 [X1 = 1.7, X2 = 0.29, X3 = 436, X4 = 24.2, X5 = -0.3]은 "왓슨 가닥"으로 표시된 영역에서 "-2" 열과 "A" 행 사이의 해당 셀에 채워졌다. 상보적 가닥(즉, 크릭 가닥)의 그 대응하는 염기 "T"의 경우, "T"와 관련된 신호 특징 벡터, [X1' = -1.9, X2' = 0.23, X3' = 24.2, X4' = 436, X5' = -1.4]는 "크릭 가닥"으로 표시된 영역에서 "-2" 열과 "T" 행 사이의 해당 셀에 채워졌다. 같은 열의 다른 셀은 "0"으로 채워졌다. 일부 구현예에서, 신호 특징 벡터의 요소 순서는 변화될 수 있다. 예를 들어, [X2, X1, X3, X4, X5], [X2, X3, X4, X5, X1], [X1, X3, X5, X4, X2] 또는 다른 조합이 사용될 수 있다. 일부 구현예에서, 신호 특징 벡터의 크기는 5로 제한되지 않을 수 있다. 예를 들어, 신호 특징 벡터의 크기는 더 많은 처리된 전기 신호 특징 또는 미가공 전기 신호를 추가하여 6, 7, 8, 9, 10, 15, 20, 30, 40, 50, 100 등을 포함할 수 있지만 이에 제한되지 않는다. 신호 특징 벡터의 크기는 신호 특징 벡터의 일부 특징을 편집하거나 삭제함으로써 1, 2, 3, 4

를 포함할 수 있으나 이에 제한되지 않는다.

[0097] 21-nt 서열 맥락과 관련된 각 뉴클레오티드에 대한 나머지 신호 특징 벡터가 동일한 규칙을 사용하여 채워져서 21-nt IPM을 형성했다. 따라서 이러한 IPM은 전류 신호 패턴, 시퀀싱 맥락, 시퀀싱 위치뿐만 아니라 경시적으로 변하는 패턴을 동시에 인코딩할 것이다. 메틸화 및 비메틸화 DNA 데이터세트에서 유래한 다수의 IPM이 CNN 또는 RNN 모델을 훈련하는 데 사용되었으며, 이후 시험 샘플의 CpG 부위에서 메틸화 상태를 결정하기 위해 사용되었다.

[0098] 블록(628)은 CNN 분석을 나타낸다. 구현예에서, CNN 모델은 각각 커널 크기가 1x25인 32, 64, 128, 256개의 필터를 갖는, 4개의 2차원(2D) 콘볼루션 레이어를 활용했다. 정류 선형 장치(ReLU)의 활성화 함수가 이러한 컨볼루션 레이어에 사용되었다. 이후에 배치 정규화 레이어가 적용되었다. 평탄화된 레이어가 더 추가되고, 드롭아웃 비율이 0.5인 드롭아웃 레이어에 이어, ReLU 활성화 함수를 사용하여 200개의 뉴런을 포함하는 완전 연결 레이어가 추가되었다. 시그모이드 활성화 함수로 하나의 뉴런을 갖는 출력 레이어가 최종적으로 적용되어 CpG 부위가 메틸화되는 확률 점수(즉, 메틸화 확률)를 산출하였다. CNN 모델을 위한 프로그램은 Keras 딥 러닝 프레임워크(keras.io/)에 기반하여 구현되었다. 일부 구현예에서, 커널 크기 ***n*** x ***m***을 변경할 수 있으며, "***n***"은 1, 2, 3, 4, 5, 10, 15, 20, 30, 35, 40, 45, 50, 100 등을 포함할 수 있지만 이에 제한되지 않고, "***m***"은 1, 2, 3, 4, 5, 10, 15, 20, 30, 35, 40, 45, 50, 100 등을 포함할 수 있지만 이에 제한되지 않는다.

[0099] **도 7**은 염기 변형 분석의 성능에 대한 커널 크기의 영향의 표이다. 첫 번째 열은 상이한 커널 크기를 나타낸다. 두 번째 열은 훈련 데이터세트로부터의 AUC(ROC[수신자 조작 특성] 곡선 아래 면적)를 나타낸다. 세 번째 열은 시험 데이터세트로부터의 AUC를 나타낸다. 도 7은 1x5, 1x10, 1x15, 1x20 및 1x25와 같은 다양한 커널 크기가 각각 0.96, 0.96, 0.97, 0.96, 및 0.96의 AUC로 표시된 바와 같이, 메틸화 CpG 부위와 비메틸화 CpG 부위를 구별하는 데 있어서 필적하는 성능을 제공할 것임을 나타낸다.

[0100] 블록(632)은 RNN 분석을 나타낸다. 구현예에서, LSTM 장치를 갖는 RNN 모델이 각각 256개의 은닉 노드를 갖는, 2개의 완전 연결 은닉 레이어와 함께 사용되었다. LSTM 은닉 장치의 전류 출력은 전류 입력 및 LSTM 셀에 저장된 이전 정보에 의해 결정된다. 일례로, 21-nt IPM의 첫 번째 행을 표시하는 위치와 관련된 신호 특징 벡터 [***X1, X2, X3, X4, X5***]는 특정 시간 단계에서 LSTM 장치에 대한 입력 X_t로 간주되었다. 순방향 LSTM RNN은 하기와 같은 작업에 기반하여 시간 단계에 따라 은닉 레이어 ***H***를 재귀적으로 계산할 것이다(Gers 등 IEEE Transactions on Neural Networks. 2001;12:1333-1340).

[0101] $$A_{t,F} = sigmoid\left(W_{xa,F}X_{t,F} + W_{ha,F}H_{t-1,F} + W_{ca,F}\odot C_{t-1,F} + b_{a,F}\right),$$

[0102] $$F_{t,F} = sigmoid\left(W_{xf,F}X_{t,F} + W_{hf,F}H_{t-1,F} + W_{cf,F}\odot C_{t-1,F} + b_{f,F}\right),$$

[0103] $$C_{t,F} = F_{t,F}\odot C_{t-1,F} + A_{t,F}\odot tanh\left(W_{xc,F}X_{t,F} + W_{hc,F}\odot C_{t,F} + b_{c,F}\right),$$

[0104] $$O_{t,F} = sigmoid\left(W_{xo,F}X_{t,F} + W_{ho,F}H_{t-1,F} + W_{co,F}\odot C_{t,F} + b_{o,F}\right),$$

[0105] $$H_{t,F} = O_{t,F}\odot tanh(C_{t,F}).$$

[0106] 역방향 LSTM RNN은 하기와 같은 작업에 기반하여 시간 단계에 따라 은닉 레이어 ***H***를 재귀적으로 계산할 것이다(Gers 등 IEEE Transactions on Neural Networks. 2001;12:1333-1340).

[0107] $$A_{t,B} = sigmoid\left(W_{xa,B}X_{t,B} + W_{ha,B}H_{t-1,B} + W_{ca,B}\odot C_{t-1,B} + b_{a,B}\right),$$

[0108] $$F_{t,B} = sigmoid\left(W_{xf,B}X_{t,B} + W_{hf,B}H_{t-1,B} + W_{cf,B}\odot C_{t-1,B} + b_{f,B}\right),$$

[0109] $$C_{t,B} = F_{t,F}\odot C_{t-1,B} + A_{t,B}\odot tanh\left(W_{xc,F}x_{t,B} + W_{hc,B}\odot C_{t,B} + b_{c,B}\right),$$

[0110] $$O_{t,B} = sigmoid\left(W_{xo,B}X_{t,B} + W_{ho,B}H_{t-1,B} + W_{co,B}\odot C_{t,B} + b_{o,B}\right),$$

[0111] $$H_{t,B} = O_{t,B}\odot tanh(C_{t,B}).$$

[0112] 식 중, W 및 b는 가중치 및 편향이고; X는 입력 벡터이고; A는 입력 게이트의 활성화 벡터이고; F는 망각 게이트의 시그모이드 함수이고; C는 셀 상태이고; O는 출력 게이트의 시그모이드 함수이고 H는 LSTM 은닉 장치의 출력이다.

[0113] 순방향 및 역방향 LSTM RNN 장치의 출력이 조합된다.

[0114] $Z_t = H_{t,F} \oplus H_{t,B}$.

[0115] LSTM RNN 출력의 마지막 레이어에 드롭아웃 비율이 0.2인 드롭아웃 레이어가 뒤따랐다. 시그모이드 활성화 함수로 하나의 뉴런을 갖는 출력 레이어가 최종적으로 적용되어 CpG 부위가 메틸화되는 확률 점수(즉, 메틸화 확률)를 산출하였다. CNN 모델을 위한 프로그램은 Keras 딥 러닝 프레임워크(keras.io/)에 기반하여 구현되었다.

[0116] D. 매개변수 분석

[0117] AUC(ROC[수신자 조작 특성] 곡선 아래 면적)에 대한 상이한 전류 벡터 매개변수 및 상이한 윈도우 크기의 효과가 분석된다. 본 발명자들은 본 개시내용에 제시된 구현예에 따라 IPM-CNN 모델에 기반하여 IPM에서 상이한 매개변수를 사용하여 구별 능력을 분석했다. 이를 위해 WGA DNA 및 M.SssI 처리 DNA 데이터세트로부터 각각 8,282개 분자(38,238개 CpG 부위) 및 8,247개 분자(39,708개 CpG 부위)가 분석되었다.

[0118] **도 16**은 AUC에 대한 상이한 매개변수 조합의 효과의 그래프를 나타낸다. 전류 벡터 매개변수의 상이한 조합은 x축에 있고 AUC는 y축에 있다. 도 16은 IPM에서 X1, X2, X3, X4 및 X5의, 그러나 이에 제한되지는 않는 상이한 매개변수 조합의 사용이 CpG 메틸화 분석의 상이한 성능을 야기하였음을 나타낸다. 예를 들어, IPM에서 X1의 사용은 0.954의 AUC를 초래한 반면, IPM에서 X1 및 X2의 조합은 0.893의 AUC를 제공했다. IPM에서 X1, X2, 및 X3의 조합은 AUC를 0.963으로 높였다. IPM에서 X1, X2, X3 및 X4의 조합은 AUC를 0.978로 더 높였으며, 이 예에서 X1, X2, X3, X4 및 X5의 사용으로 0.977의 AUC에서 성능 정체가 이어졌다. 따라서, 일부 구현예에서, IPM에서 매개변수의 상이한 조합은 메틸화 및 비메틸화 CpG 부위 간 구별에서 원하는 성능을 결정할 수 있도록 할 것이다.

[0119] 조합이 아닌 X1, X2, X3, X4 및 X5의 개별적인 사용이 시험되었다. X1, X2, X3, X4 및 X5를 개별적으로 사용한 결과 각각 0.95, 0.92, 0.98, 0.88 및 0.95의 AUC를 초래하였다. X3(즉, 영역의 P_{ij}의 중앙값)은 0.98의 높은 AUC를 초래했다. 높은 AUC는 적어도 부분적으로 전체 단편 수준에서 메틸화 차이의 결과일 수 있다. 사용된 데이터세트에는 WGA(완전 비메틸화) 및 M.SssI(완전 메틸화)이 관여하였다. 그러나 실제로는 단편이 완전 메틸화되지 않거나 완전히 비메틸화되지 않을 것이다. 완전 메틸화되지 않거나 완전 비메틸화되지 않은 샘플에 대해 X3 자체의 사용은 이렇게 높지 않은 AUC를 초래할 수 있다.

[0120] **도 17**은 AUC에 대한 윈도우 크기의 효과의 그래프를 나타낸다. x축은 뉴클레오티드 단위의 윈도우 크기를 나타낸다. y축은 AUC를 나타낸다. IPM에서 사용된 뉴클레오티드의 수(윈도우 크기로도 지칭됨)는 나노포어 시퀀싱 동안 생성된 전류 신호의 상이한 정보 내용을 포착할 것이고 메틸화 분석의 성능에 영향을 미칠 수 있다. 도 17은 IPM-CNN 모델을 사용하는 메틸화 및 비메틸화 CpG 부위 간 구별에서의 성능이 IPM에 사용된 뉴클레오티드 수가 1에서 10 nt로 증가함에 따라 AUC를 0.715에서 0.969로 점차 증가시키는 것으로 나타났음을 보여준다. 이 예에서 성능 정체는 7 nt의 윈도우 크기에서 도달했다. 따라서, 일부 구현예에서, IPM의 윈도우 크기 조정은 메틸화 및 비메틸화 CpG 부위 간 구별에서 원하는 성능을 결정할 수 있도록 할 것이다.

[0121] 구현예는 가장 높은 AUC를 야기하는 전류 벡터 매개변수 또는 윈도우 크기의 조합을 사용할 것을 필요로 하지 않을 수 있다. 특정 용도에는 더 낮은 AUC가 충분할 수도 있고, 또는 더 높은 AUC가 추가 매개변수와 관련된 추가 연산 및 저장 비용을 감당할 가치가 없을 수도 있다. 또한, 상이한 매개변수가 조정되어 원하는 AUC, 특이성 및/또는 민감도를 달성할 수 있다. 예를 들어, 더 큰 윈도우 크기가 X1, X2, X3, X4 및 X5 중 더 적은 수의 매개변수를 사용하는 것을 보상하기 위해 사용될 수 있다.

[0122] E. 6mA 변형의 검출

[0123] 5mC 이외의 변형에 대한 전류 신호 분석의 적용 가능성을 결정하기 위해, 전류 신호 분석이 N6-메틸아데닌(6mA)을 검출하기 위해 사용되었다.

[0124] **도 18**은 나노포어 시퀀싱을 통해 한 가닥으로부터의 전류 신호를 사용하여 6mA 메틸화를 결정하는 과정을 나타낸다. 도 18은 5mC 메틸화를 결정하는 과정을 나타낸 도 5와 유사하다. 블록(1804)에서 이중 가닥 DNA 분자가 제공된다. 블록(1808)에서, 이중 가닥 DNA 분자는 나노포어 시퀀싱에 적합한, 시퀀싱 어댑터와 결찰된다. 블록

(1812)에서, 나노포어 시퀀싱이 수행된다. 블록(1816)에서 전류 신호가 얻어진다. 블록(1820)에서 통합 제시 행렬(IPM)이 작제된다. 블록(1804-1820)은 블록(504-520)과 동일할 수 있다.

[0125] 6mA 메틸화를 결정하기 위한 예시 목적으로, 5'-G[TACCCG]GGTACTG[TCTAGA]G-3'의 21-nt 서열이 메틸화 분석의 대상인 뉴클레오티드 A(예를 들어 0의 위치에 대응함)가 중앙에 IPM의 기반으로서 일례로 사용되었다. IPM(1824)은 21-nt 서열을 사용한 결과를 나타낸다. 단순화를 위해 괄호 안의 염기는 생략되었다("..."로 표시됨). 한 가닥의 아데닌 염기("A")에 대응하는 위치 0의 경우 "A"와 관련된 신호 특징 벡터(즉, [X1 = 0.39, X2 = 0.04, X3 = 389, X4 = 46.3, X5 = 0.32])가 행렬의 "0" 열과 "A" 행 사이의 해당 셀에 채워졌다. 같은 열의 다른 셀은 "0"으로 채워졌다. 일부 구현예에서, 신호 특징 벡터의 요소 순서는 변화될 수 있다. 예를 들어, [X2, X1, X3, X4, X5], [X2, X3, X4, X5, X1], [X1, X3, X5, X4, X2] 또는 다른 조합이 사용될 수 있다. 일부 구현예에서, 신호 특징 벡터의 크기는 5만이 아닐 수도 있다. 예를 들어, 신호 특징 벡터의 크기는 더 많은 처리된 전기 신호 특징 또는 미가공 전기 신호를 추가하여 6, 7, 8, 9, 10, 15, 20, 30, 40, 50, 100 등을 포함할 수 있으나 이에 제한되지 않는다. 신호 특징 벡터의 크기는 신호 특징 벡터의 일부 특징을 편집하거나 삭제함으로써 1, 2, 3, 또는 4를 포함할 수 있으나 이에 제한되지 않는다.

[0126] 21-nt 서열 맥락과 관련된 각 뉴클레오티드에 대한 나머지 신호 특징 벡터는 동일한 규칙을 사용하여 채워져서 21-nt IPM을 형성했다. 따라서 이러한 IPM은 전류 신호 패턴, 시퀀싱 맥락, 시퀀싱 위치뿐만 아니라 경시적으로 변화하는 패턴을 동시에 인코딩할 것이다. 뉴클레오티드 A와 관련하여 메틸화 및 비메틸화 DNA 데이터세트에서 유래한 다수의 IPM이 CNN 또는 RNN 모델을 훈련하는 데 사용되었으며, 이후 시험 샘플의 A 부위에서 메틸화 상태를 결정하기 위해 사용되었다. 블록(1828)은 CNN 분석을 나타내고, 블록(1832)은 RNN 분석을 나타낸다. 이들 블록은 블록(528 및 532)과 동일할 수 있다.

[0127] 상기 예시된 본 발명의 접근(IPM-CNN 또는 IPM-RNN)이 아데닌 메틸화(6mA)를 결정할 수 있는지 여부를 시험하기 위해, 본 발명자들은 이전 연구(Rand 등 *Nat Methods* 2017;14:411-413)로부터 pUC19 플라스미드 DNA의 나노포어 시퀀싱 결과를 포함하는 2개의 공개 데이터세트를 다운로드했다. 제1 데이터세트(6mA 데이터세트)는 *dam* 및 *dcm* 메틸트랜스퍼라제를 둘 모두 함유하는 대장균에서 성장시킨 pUC19 플라스미드 DNA로부터 생성되었고, 모든 GATC 모티프가 A 부위에서 메틸화된 것으로 예상되었다. 제2 데이터세트(uA 데이터세트)는 변형되지 않은 뉴클레오티드를 사용하여 PCR 증폭을 거친 DNA로부터 생성되었고, 모든 A 부위가 비메틸화된 것으로 제시되었다. 훈련 과정에서 본 발명자들은 IPM-CNN 모델을 사용하여 6mA 데이터세트로부터의 GATC 모티프를 포함하는 2052개 분자 및 uA 데이터세트로부터의 2081개 분자를 분석했다.

[0128] **도 19**는 IPM-CNN 모델을 사용하여 생성된 AUC를 나타낸다. x축은 특이성을 나타낸다. y축은 민감도를 나타낸다. 선(1904)은 훈련 데이터세트로부터의 결과를 나타낸다. 훈련 데이터세트를 사용한 AUC는 0.94이다. 훈련 과정에서 본 발명자들은 훈련된 IPM-CNN 모델을 6mA 데이터세트로부터의 GATC 모티프를 함유하는 522개 분자 및 uA 데이터세트로부터의 481개 분자에 적용했다. 시험 데이터세트를 사용한 AUC는 0.92이다. 또한 IPM-RNN 모델을 사용할 때 훈련 및 시험 데이터세트 둘 모두에 대해 0.89의 AUC가 달성되었다. 이러한 데이터는 IPM-CNN 및 IPM-RNN이 6mA 부위를 비메틸화 A 부위와 구별할 수 있도록 함을 제시했다.

[0129] 구현예에서, 인간 또는 비인간 DNA에 대한 6mA 결정을 위한 훈련 데이터세트는 각각 6mA 뉴클레오티드 및 비메틸화 A 뉴클레오티드를 사용하는 PCR 증폭에 기반하여 작제될 수 있다. 몇 번의 PCR 주기 후에 대부분의 DNA 분자는 6mA 뉴클레오티드로 증폭된 DNA로부터 생성된 데이터세트에 대해 6mA 뉴클레오티드를 운반할 것인 반면, 대부분의 DNA 분자는 비메틸화 A 뉴클레오티드로 증폭된 DNA로부터 생성된 데이터세트에 대해 비메틸화 A 뉴클레오티드를 운반할 것이다. 이 두 가지 유형의 데이터세트는 시험 샘플에서 A 뉴클레오티드의 메틸화 상태를 결정하기 위해 CNN 및/또는 RNN 모델을 훈련하는 데 사용될 수 있다.

[0130] 5mC 외에 6mA를 검출하기 위한 전류 신호 분석의 사용은 이러한 분석의 다른 메틸화 유형에 대한 적용 가능성을 실증한다. 따라서, 이들 방법은 본원에 기재된 다른 메틸화를 정확하게 검출할 것이다.

[0131] F. 인간 대상체의 비종양 및 종양 조직 간 CpG 메틸화 분석

[0132] 본원에 기재된 구현예를 사용하여 결정된 부위의 메틸화는 상이한 유형의 조직을 구별하기 위해 사용될 수 있다. 본 개시내용의 구현예에 따라 IPM-RNN 모델을 사용하여 본 발명자들은 비인두암종(NPC) 종양 및 백혈구 연층 샘플에서 유래하는 세포성 DNA 분자의 메틸화 패턴을 분석하였다. 이를 위해 본 발명자들은 NPC 종양으로부터의 147개 분자를 사용했으며, 크기 중앙값은 4,406 bp(사분위간 범위(IQR): 1,962 - 8,128 bp)였고 분자당 중앙값은 32 CpG(IQR: 13 - 61)였다. 본 발명자들은 백혈구 연층으로부터 또 다른 147개 분자를 분석했으며, 크

기 중앙값은 6,823 bp(사분위간 범위(IQR): 2,515 내지 9,304 bp)였고 분자당 중앙값은 49 CpG(IQR: 23 - 118)였다.

[0133] **도 20**은 백혈구 연층 샘플로부터 및 NPC 종양 조직 샘플로부터의 DNA 분자의 비교 그래프를 나타낸다. x축은 조직 유형을 나타낸다. y축은 메틸화 수준을 %로 나타낸다. 백혈구 연층의 단일 분자 메틸화 수준(즉, 메틸화된 것으로 결정된 분자 내 CpG 부위의 백분율)(중앙값: 74.8%, IQR: 71.1% 내지 80.1%)은 NPC 종양(중앙값: 50; IQR: 45.7 내지 53.1)에서보다 유의하게 더 높은 것으로 나타났다(*P* 값 <0.0001, 윌콕슨(Wilcoxon) 순위 합산 시험). 종양 조직에서 유래된 DNA 분자는 저메틸화된 것으로 나타났으며, 이는 단기 판독 바이설파이트 시퀀싱에 기반한 이전 결론과 일치하였다(Chan 등 Proc Natl Acad Sci USA 2013;110:18761-8). 그러나 본원에 기재된 새로운 나노포어 시퀀싱 기술은 거의 전체 길이의 DNA 분자의 시퀀싱 및 DNA 분자에 대한 메틸화 패턴 분석을 허용한다. 예를 들어, 나노포어 시퀀싱은 짧은 판독 시퀀싱 플랫폼(예를 들어, Illumina)에 의해 조사될 수 없던, 600 bp 초과 크기의 DNA 분자를 분석할 수 있다.

[0134] **도 21**은 종양 DNA 분자 및 백혈구 연층 DNA 분자의 메틸화 패턴을 예시한다. 검은색 원(예를 들어, 원(2104))은 메틸화 CpG 부위를 표시한다. 채워지지 않은 원(예를 들어, 원(2108))은 비메틸화 CpG 부위를 표시한다. 원은 분석 중인 DNA 분자의 5' 말단에 대한 CpG 부위의 상대 위치를 나타낸다(즉, 도면에서 DNA 분자의 왼쪽이 5' 말단에 더 가까움). 도 21에 나타난 바와 같이, 종양 조직에서 유래된 DNA 분자는 백혈구 연층 샘플에서 유래된 분자와 비교하여 분자 내에서 더 많은 비메틸화 CpG 부위를 운반하는 경향이 있었다. 백혈구 연층 샘플로부터의 분자 중 5.4%만이 50% 미만의 단일 분자 메틸화 수준을 가졌고 길이 중앙값은 2,091 bp였다. 대조적으로 NPC 종양 조직으로부터의 분자 중 39.5%가 50% 미만의 단일 분자 메틸화 수준을 가졌고 길이 중앙값은 2,924 bp였다. DNA 분자의 길이는 897 bp 내지 10,424 bp 범위였다.

[0135] 이들 데이터는 본원에 기재된 메틸화를 검출하기 위한 나노포어 시퀀싱 기술이 각 DNA 분자(예를 들어, 비종양 DNA 대 종양 DNA 분자)의 기원 조직을 조직 생검 샘플과 구별하기 위해 단일 분자 메틸화 패턴 분석에서 사용될 수 있음을 나타낸다. 조직 생검으로부터의 단일 분자 메틸화 패턴 분석은 종양 등급 또는 하위유형 검사, 암 또는 다른 질환의 치료 모니터링, 장기 이상(예를 들어, 신장 부전) 평가 등을 허용할 것이다.

[0136] G. 태아 및 모체 DNA 분자 간 분석

[0137] 본원에 기재된 구현예를 사용하여 결정된 부위의 메틸화는 태아 및 모체 DNA 분자 간을 구별하기 위해 사용될 수 있다. IPM-CNN 모델에 따르면, 본 발명자들은 모체 백혈구 연층과 태반 조직 간 SNP 정보를 활용하여, 임신 3기 임신부로부터 얻은 1,262개의 태아 특이적 무세포 DNA 분자(크기 중앙값: 530 bp, IQR: 361 - 779 bp) 및 6,108개의 모체 DNA 특이적 무세포 DNA 분자(크기 중앙값: 668 bp, IQR: 448 - 1,089 bp)에 대해 적어도 5개의 CpG 부위를 사용하여 단일 분자 메틸화 패턴을 결정했다. 이러한 임산부의 혈장 DNA 중 태아 DNA 분율은 26.0%였다.

[0138] **도 22**는 모체 특이적 및 태아 특이적 DNA 분자 간 단일 분자 메틸화 수준을 나타낸다. x축은 무세포 DNA 분자의 범주: 모체 특이적 또는 태아 특이적을 나타낸다. y축은 단일 분자 메틸화 수준을 %로 나타낸다. 단일 혈장 DNA 분자의 메틸화 수준 중앙값(즉, 메틸화된 것으로 결정된 분자 내 CpG 부위의 백분율)은 태아 특이적 무세포 DNA 분자의 경우 66.6%(IQR: 28.5 - 86.6%)였으며, 이는 모체 특이적 무세포 DNA 분자에 대한 수준(중앙값: 78.5%, IQR: 50 - 93.7%)보다 유의하게 더 낮았다(*P* 값: < 0.0001, 만-휘트니 *U* 시험). 결과는 무세포 DNA 분자의 메틸화 정보의 사용이 각 혈장 DNA 분자의 모체 및 태아 기원을 구별할 수 있도록 함을 제시했다.

[0139] 추가로, 2021년 2월 5일에 출원된 미국 특허 출원 번호 17/168,950에 기재된 바와 같이 IPM-CNN 모델에 의해 결정된 메틸화 패턴을 백혈구 연층 및 태반 조직의 각 참조 메틸화 패턴과 비교하여, 임산부에서 태아 및 모체 기원의 혈장 DNA 분자 간 구별에 대해 0.87의 AUC를 달성할 수 있었다.

[0140] **도 23은** IPM-CNN 모델에 의해 결정된 메틸화 패턴에 기반하여 임산부에서 무세포 DNA 분자의 태아 및 모체 기원 분석에 대한 ROC 곡선을 나타낸다. x축은 특이성이고, y축은 민감도이다.

[0141] III. IPM 기반 메틸화 결정 평가용 데이터세트

[0142] 비메틸화 데이터세트는 전체 게놈 증폭(WGA)을 통해 제조된 증폭된 DNA(WGA DNA 데이터세트로 표시됨)로부터의 시퀀싱 결과를 함유했다. WGA에서 변형되지 않은 뉴클레오티드의 사용은 염기 변형을 거의 함유하지 않는 증폭된 DNA를 초래하였다(소량의 입력 게놈 DNA 제외). 메틸화 데이터세트는 시퀀싱 전 M.SssI(스피로플라즈마(*Spiroplasma*) *sp.* 균주 MQ1로부터의 메틸트랜스퍼라제 유전자를 함유하는 대장균 균주로부터 단리된 CpG 메틸트랜스퍼라제는 이중 가닥 DNA의 모든 CpG 부위를 메틸화할 것임)에 의해 처리된 DNA(M.SssI-처리 DNA 데이터세

트로 표시됨)로부터의 시퀀싱 결과를 함유하였다. M.SssI 메틸트랜스퍼라제는 CpG 부위를 메틸화했다.

[0143] WGA DNA 데이터세트를 제조하기 위해 반응 혼합물(phi29 반응 완충액 및 dNTP 함유)을 열 블록에서 95℃에서 5분 동안 인큐베이션한 후 4℃로 냉각하여 엑소뉴클레아제 내성 무작위 프라이머를 1 ng의 DNA 주형에 사전 어닐링한다. 이어서, phi29 폴리머라제를 반응 혼합물에 첨가하고 30℃에서 4시간 동안 인큐베이션하였다. DNA를 Ampure XP 비드로 정제 하고 Qubit 형광계로 정량했다. 전형적으로 200 ng의 DNA를 20 μl 반응으로부터 얻을 수 있었다.

[0144] M.SssI-처리 DNA 데이터세트를 제조하기 위해 WGA 후 DNA의 절반을 M.SssI 효소로 처리하였다. 메틸트랜스퍼라제 반응 완충액, S-아데노실메티오닌(SAM) 및 M.SssI를 DNA와 혼합하고 37℃에서 2시간 동안 인큐베이션하였다. 65℃에서 20분 동안 가열하여 반응을 중단시켰다. 결찰 시퀀싱 키트(SQK-LSK109)(Oxford Nanopore)를 라이브러리 제조에 사용했다. DNA를 NEBNext Ultra II 말단 복구/dA 테일링 모듈과 함께 FFPE DNA 복구 혼합물로 처리하였다. Ampure XP 비드 세척 후 어댑터 혼합물, 결찰 완충액 및 NEBNext Quick T4 DNA 리가제를 첨가하여 시퀀싱 어댑터를 복구된 DNA에 결찰했다. 결찰된 DNA를 Ampure XP 비드로 세척하고 짧은 단편 완충액으로 세척했다. 라이브러리를 용출 완충액에 재현탁했다. R9.4.1 플로우셀을 WGA(샘플_01) 및 M.SssI 처리(샘플_02) 라이브러리 각각의 시퀀싱에 사용하였다. 플로우셀을 먼저 플러시 테터(Flush Tether) 및 플러시 완충액을 함유하는 플로우셀 프라이밍 혼합물로 프라이밍하였다. 그런 다음 시퀀싱 완충액, 로딩 비드 및 DNA 라이브러리를 혼합하여 라이브러리 로딩 혼합물을 제조했다. 라이브러리 로딩 혼합물을 플로우 셀 샘플 포트에 한 방울씩 첨가했다. 로딩된 플로우 셀을 PromethION의 슬롯에 연결하고 기본 매개변수를 사용하여 64시간 동안 시퀀싱했다.

[0145] 본 발명자들은 각각 샘플_01 및 샘플_02에 대해 1560만개 및 1530만개 나노포어 시퀀싱 판독을 얻었으며, 그 중 1380만개(88.7%) 및 1380만개(90.7%) 판독을 Minimap2(Li H, Bioinformatics. 2018;34(18):3094-3100)를 사용하여 인간 참조 게놈(UCSC hg19)에 대해 정렬될 수 있었다. 판독 길이 중앙값은 샘플_01 및 샘플_02에 대해 각각 510 nt(사분위간 범위(IQR): 333 - 778 nt) 및 606 nt(IQR: 382 - 911 nt) 였다. 일부 구현예에서, BLASR(Mark J Chaisson 등, BMC Bioinformatics. 2012; 13: 238), BLAST(Altschul SF 등, J Mol Biol. 1990;215(3):403-410), BLAT(Kent WJ, Genome Res. 2002;12(4):656-664), BWA(Li H 등, Bioinformatics. 2010;26(5):589-595), NGMLR(Sedlazeck FJ 등, Nat Methods. 2018;15(6):461-468), 및 LAST(Kielbasa SM 등, Genome Res. 2011;21(3):487-493)를 시퀀싱된 판독을 참조 게놈에 대해 정렬하는 데 사용할 수 있었다.

[0146] **도 8**은 IPM에 기반하여 CNN 및 RNN 모델을 훈련하고 시험하는 데 사용된 시퀀싱 분자의 수를 나타내는 표이다. 첫 번째 열은 데이터세트이다. M.SssI-처리 DNA는 메틸화 DNA 데이터세트이고, WGA DNA는 비메틸화 DNA 데이터세트이다. 두 번째 열은 훈련에 사용된 분자 수 및 CpG 부위 수이다. 세 번째 열은 시험에 사용된 분자 수 및 CpG 부위 수이다. 훈련 데이터세트의 경우 본 발명자들은 M.SssI-처리 DNA(메틸화 DNA) 및 WGA DNA(비메틸화 DNA)로부터 각각 무작위로 7,989개 및 8,052개의 시퀀싱 분자를 사용했다. 이러한 훈련 데이터세트는 38,470개의 메틸화 CpG 부위 및 37,150개의 비메틸화 CpG 부위로 이루어졌다. 시험 데이터세트의 경우 본 발명자들은 M.SssI-처리 DNA(메틸화 DNA) 및 WGA DNA(비메틸화 DNA)로부터 각각 무작위로 4,826개 및 5,041개의 시퀀싱 분자를 사용했다. 이러한 훈련 데이터세트는 9,716개의 메틸화 CpG 부위 및 11,444개의 비메틸화 CpG 부위로 이루어졌다.

[0147] **도 9a-9d**는 IPM-CNN 및 IPM-RNN 접근을 사용하는 WGA DNA 및 M.SssI-처리 DNA 데이터세트 간 CpG에 있어서 메틸화될 확률의 상자그래프이다. 그래프의 x축에는 데이터세트가 있다. 메틸화 확률은 y축에 있다. 도 9a 및 9b는 IPM-CNN 분석을 사용한 결과를 나타낸다. 도 9a는 훈련 데이터세트의 IPM-CNN 분석을 나타내며, M.SssI-처리 DNA 데이터세트에서 CpG에 대한 메틸화 확률(중앙값: 0.99; IQR: 0.987 - 0.999)은 WGA DNA 데이터세트에서의 확률(중앙값: 0.03, IQR: 0.001 - 0.15)보다 유의하게 더 높았다(*P* 값 < 0.0001, 만-휘트니 U 시험). 도 9b는 시험 데이터세트의 IPM-CNN 분석을 나타내며, 이는 또한 WGA(중앙값: 0.4; IQR: 0.002 - 0.18)와 M.SssI-처리 DNA 데이터세트(중앙값: 0.99, IQR: 0.980 - 0.999) 간 CpG에 있어서 메틸화될 확률의 유의차를 나타내었다(*P* 값 < 0.0001, 만-휘트니 U 시험).

[0148] 도 9c 및 9d는 IPM-RNN 분석을 사용한 결과를 나타낸다. 도 9c는 훈련 데이터세트의 IPM-RNN 분석을 나타내며, M.SssI-처리 DNA 데이터세트에서 CpG에 대해 메틸화될 확률(중앙값: 0.994; IQR: 0.92 - 0.99)은 WGA DNA 데이터세트에서의 확률(중앙값: 0.079; IQR: 0.059 - 0.118)보다 유의하게 더 높았다(*P* 값 < 0.0001, 만-휘트니 U 시험). 도 9d는 시험 데이터세트의 IPM-RNN 분석을 나타내며, 이는 또한 WGA(중앙값: 0.077; IQR: 0.057 - 0.115)와 M.SssI-처리 DNA 데이터세트(중앙값: 0.994, IQR: 0.919 - 0.999) 간 CpG에 있어서 메틸화될 확률의 유의차를 나타내었다(*P* 값 < 0.0001, 만-휘트니 U 시험). 이러한 결과는 본 개시내용에 제시된 구현예에 따라

CpG 부위의 메틸화 상태를 결정하기 위해 나노포어 시퀀싱에 의해 생성된 전기 신호를 사용하는 것이 타당하였음을 표시한다. 한 구현예에서, 0.5의 메틸화 확률 컷오프가 CpG 부위에서 메틸화 상태를 결정하는 데 사용될 수 있다. 이 컷오프를 사용하면 IPM-CNN 분석에서 DNA 메틸화 검출에 대한 특이성 및 민감도가 훈련 데이터세트의 경우 각각 96% 및 91%, 그리고 시험 데이터세트의 경우 각각 93% 및 88%였다. IPM-RNN 분석의 경우, DNA 메틸화 검출의 특이성 및 민감도는 훈련 및 시험 데이터세트 둘 모두에서 각각 97% 및 88%였다. 일부 구현예에서, 메틸화 확률에 대한 컷오프는 다양한 적용에 따라 조정될 수 있다.

[0149] **도 10a** 및 **10b**는 수신자 조작 특성(ROC) 곡선 분석을 나타낸다. 특이성은 x축에 나타낸다. 민감도는 y축에 나타낸다. 도 10a는 훈련 데이터세트에 대한 결과를 나타낸다. 도 10b는 시험 데이터세트에 대한 결과를 나타낸다. IPM-CNN 결과는 선(1004 및 1008)으로 나타낸다. IPM-RNN 결과는 선(1012 및 1016)으로 나타낸다. DeepMod(Liu 등 Nat Commun. 2019;10:2449) 결과는 선(1020 및 1024)으로 나타낸다. Nanopolish(Liu 등 Nat Commun. 2019;10:2449) 결과는 선(1028 및 1032)으로 나타낸다. IPM 기반 CNN 및 RNN 분석은 0.95 이상의 ROC 곡선 아래 면적(AUC)과 함께 훈련 및 시험 데이터세트 둘 모두에 대해 우수한 성능을 제공했다. IPM 기반 CNN 및 RNN 모델은 DeepMod(0.83) 및 nanopolish(0.91)와 비교하여, 시험 데이터세트에서 ROC 곡선 아래 면적(AUC) 0.95 및 0.97로 더 우수한 성능을 야기하였다. IPM 기반 RNN 또는 CNN 대 DeepMod 및 nanopolish를 포함하는 다른 도구의 모든 비교에 대한 P 값(DeLong 시험)은 < 0.0001로 나타났다. 이러한 결과는 IPM-CNN 및 IPM-RNN이 DNA 메틸화 분석에 있어서 다른 도구보다 더 우수함을 표시하였다.

[0150] **도 11**은 상이한 분석에 있어서 주어진 특이성에 대한 민감도의 표이다. 첫 번째 열은 분석 유형을 나타낸다. 두 번째 열은 민감도를 나타낸다. 세 번째 열은 특이성을 나타낸다. 도 11은 주어진 특이성으로 IPM-CNN 및 IPM-RNN 분석이 훨씬 더 높은 민감도를 달성했음을 나타낸다. 예를 들어, 특이성 90%로, IPM-CNN 및 IPM-RNN 분석은 각각 90% 및 93%의 민감도를 달성한 반면 DeepMod 및 nanopolish 접근은 각각 53% 및 74%의 민감도만을 달성했다. 95%의 특이성으로, IPM-CNN 및 IPM-RNN 분석은 각각 86% 및 90%의 민감도를 달성한 반면 DeepMod 및 nanopolish 접근은 각각 38% 및 55%의 민감도만을 달성했다. 99%의 특이성으로, IPM-CNN 및 IPM-RNN 분석은 각각 70% 및 83%의 민감도를 달성한 반면 DeepMod 및 nanopolish는 각각 13% 및 16%의 민감도만을 달성했다. 이러한 결과는 서열 세그먼트에 대한 전류 신호 패턴의 통합 제시 행렬이 DNA 메틸화 결정의 정확성을 크게 개선할 것임을 추가로 실증했다. 특히 IPM-RNN은 이러한 접근 중에서 가장 우수한 성능을 야기했다.

[0151] 일부 구현예에서, IPM의 경우, 염기 변형 분석을 거친 염기 주위의 DNA 스트레치의 길이는 대칭 또는 비대칭일 수 있다. 예를 들어, 해당 염기의 X-nt 상류 및 Y-nt 하류가 염기 변형 분석에 사용될 수 있다. X는 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, 및 10000을 포함할 수 있지만 이에 제한되지 않으며; Y는 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, 및 10000을 포함할 수 있지만 이에 제한되지 않는다. X 및 Y는 동일하거나 상이할 수 있다.

[0152] 일부 구현예에서, 핵산의 염기 변형은 바이러스, 박테리아, 식물, 진균, 선충류, 곤충 및 척추동물(예를 들어, 인간) 등을 포함하는 상이한 유기체에 걸쳐 본 개시내용의 구현예에 따라 분석될 것이다. 가장 일반적인 염기 변형은 메틸화로 불리는 상이한 위치의 상이한 DNA 염기에 대한 메틸기의 부가이다. 메틸화는 시토신, 아데닌, 티민 및 구아닌에서, 예컨대 5mC(5-메틸시토신), 4mC(N4-메틸시토신), 5hmC(5-하이드록시메틸시토신), 5fC(5-포르밀시토신), 5caC(5-카복실시토신), 1mA(N1-메틸아데닌), 3mA(N3-메틸아데닌), 6mA(N6-메틸아데닌), 7mA(N7-메틸아데닌), 3mC(N3-메틸시토신), 2mG(N2-메틸구아닌), 6mG(O6-메틸구아닌), 7mG(N7-메틸구아닌), 3mT(N3-메틸티민) 및 4mT(O4-메틸티민)이 발견되었다.

[0153] 일부 구현예에서, 전류 신호 패턴의 통합 제시 행렬은 선형 회귀, 로지스틱 회귀, 심층 순환 신경망(예를 들어, 장단기 메모리, LSTM), 베이즈 분류기, 은닉 마르코프 모델(HMM), 선형 판별 분석(LDA), k-평균 클러스터링, 노이즈를 포함하는 적용의 밀도 기반 공간 클러스터링(DBSCAN), 랜덤 포레스트 알고리즘 및 지원 벡터 머신(SVM)을 포함하지만 이에 제한되지 않는 상이한 통계적 및/또는 수학적 모델에 의해 분석될 수 있다. 또 다른 구현예에서, 자연어 처리가 염기 변형 분석을 위해 전기 신호 분석에 적용될 수 있다.

[0154] 일부 구현예에서, 생물학적 나노포어, 예컨대 단백질 α-헤모라이신 및 단백질 조작 기술에 의한 그 변형, 프로그래밍된 박테리아에 의해 생성된 포어 단백질, 합성 물질로부터 제작된 고체 상태 나노포어, 그래핀 등을 포함

하지만 이에 제한되지 않는 상이한 유형의 나노포어가 사용될 수 있다.

[0155] 구현예에서, 이들 방법은 참조 게놈, 예컨대 인간 참조 게놈(hg19), 예를 들어 긴 산재 핵 요소(LINE) 반복부를 참조하여 가이드 RNA를 설계함으로써 상동 서열을 공유하는 다수의 긴 DNA 분자를 표적화하기 위해 사용될 수 있다. 한 예에서, 이러한 분석은 태아 이수성 검출을 위한 임산부의 모체 혈장에서의 순환 무세포 DNA 분석에 사용될 수 있다(Kinde 등 PLOS One 2012;7(7):e41162). 구현예에서, 탈활성화 또는 '사멸' Cas9(dCas9) 및 그 관련된 단일 가이드 RNA(sgRNA)가 이중 가닥 DNA 분자를 절단하지 않고 표적화된 긴 DNA를 농축하는 데 사용될 수 있다. 예를 들어, sgRNA의 3' 말단은 여분의 범용 단형 서열을 보유하도록 설계될 수 있다. dCas9에 의해 결합된 표적 긴 DNA 분자를 포착하기 위해 범용 단형 서열에 상보적인 비오틴화 단일 가닥 올리고뉴클레오티드를 사용할 수 있다. 또 다른 구현예에서, 농축을 용이하게 하기 위해 비오틴화 dCas9 단백질 또는 sgRNA, 또는 둘 모두를 사용할 수 있다.

[0156] 구현예에서, 화학적, 물리적, 효소적, 겔 기반 및 자기 비드 기반 방법 또는 이러한 접근보다 많은 것을 조합하는 방법을 포함하지만 이에 제한되지 않는 접근을 사용하여 하나 이상의 특정 관심 게놈 영역에 제한하지 않고 긴 DNA 단편을 농축하기 위해 크기 선택을 수행할 수 있다.

[0157] IV. 예시적 방법

[0158] 이 섹션에서는 염기 변형을 검출하기 위해 기계 학습 모델을 사용하고 염기 변형을 검출하기 위해 기계 학습 모델을 훈련하는 예시적 방법을 나타낸다.

[0159] A. 변형의 검출

[0160] **도 12**는 핵산 분자에서 뉴클레오티드의 변형을 검출하는 것과 관련된 예시적 공정(1200)의 흐름도이다. 변형은 본원에 기재된 임의의 메틸화 또는 임의의 산화를 포함할 수 있다. 산화는 8-옥소-구아닌일 수 있다. 일부 구현에서, 도 12의 하나 이상의 공정 블록은 시스템(예를 들어, 측정 시스템(1400))에 의해 수행될 수 있다. 일부 구현에서, 도 12의 하나 이상의 공정 블록은 시스템과 별개이거나 시스템을 포함하는 또 다른 장치 또는 장치 그룹에 의해 수행될 수 있다. 추가적으로 또는 대안적으로, 도 12의 하나 이상의 공정 블록은 측정 시스템(1400)의 하나 이상의 구성요소, 예컨대 검출기(1420), 로직 시스템(1430), 로컬 메모리(1435), 외부 메모리(1440), 저장 장치(1445) 및/또는 처리장치(1450)에 의해 수행될 수 있다.

[0161] 블록(1210)에서, 입력 데이터 구조가 수신된다. 입력 데이터 구조는 샘플 핵산 분자에서 시퀀싱된 뉴클레오티드의 윈도우에 대응할 수 있다. 샘플 핵산 분자는 뉴클레오티드에 대응하는 전기 신호를 측정하여 시퀀싱된다. 전기 신호는 전류, 전압, 저항, 인덕턴스, 정전용량 또는 임피던스일 수 있다. 시퀀싱은 나노포어를 사용하여 이루어질 수 있다. 공정(1200)은 나노포어를 사용하는 샘플 핵산의 시퀀싱을 추가로 포함할 수 있다. 나노포어는 본원에 기재된 임의의 나노포어일 수 있다.

[0162] 입력 데이터 구조는 몇몇 특성에 대한 값을 포함할 수 있다. 특성은 윈도우 내 각 뉴클레오티드에 있어서, 뉴클레오티드의 정체, 각 윈도우 내 표적 위치에 대한 뉴클레오티드의 위치, 및 뉴클레오티드에 대응하는 전기 신호 세그먼트의 제1 세그먼트 통계값을 포함하는 벡터를 포함할 수 있다. 특성은 윈도우보다 크거나 같은 핵산 분자 영역에서 전기 신호의 제1 영역 통계값을 포함할 수 있다. 예를 들어, 입력 데이터 구조는 통합 제시 행렬[IPM]을 포함할 수 있다.

[0163] 뉴클레오티드의 정체는 염기(예를 들어, A, T, C 또는 G)일 수 있다. 염기는 나노포어 시퀀싱을 사용한 염기 호출 기술을 통해 결정될 수 있다. 염기 호출 기술은 전기 신호의 세그먼트를 뉴클레오티드와 관련시킬 수 있다. 뉴클레오티드의 위치는 표적 위치 대비 뉴클레오티드 거리일 수 있다. 예를 들어, 뉴클레오티드가 표적 위치로부터 한 방향으로 1개의 뉴클레오티드만큼 떨어져 있을 때 위치는 +1일 수 있고, 뉴클레오티드가 표적 위치로부터 반대 방향으로 1개의 뉴클레오티드만큼 떨어져 있을 때 위치는 -1일 수 있다.

[0164] 제1 세그먼트 통계값은 뉴클레오티드에 대응하는 전기 신호 세그먼트의 평균을 나타낼 수 있다. 일부 구현예에서, 제1 세그먼트 통계값은 뉴클레오티드에 대응하는 전기 신호 세그먼트의 전기 신호의 변동(예를 들어, 표준 편차)을 나타낼 수 있다. 구현예에서, 제1 세그먼트 통계값은 뉴클레오티드에 대응하는 전기 신호의 세그먼트 평균의 정규화된 값을 나타낼 수 있다. 정규화는 제1 세그먼트 통계값이 특정 범위(예를 들어, 0 내지 1의 범위)에 있도록 재조정하는 것을 포함할 수 있다. 정규화는 뉴클레오티드 가닥의 일부 또는 전부에 대한 중앙값, 평균값 및/또는 편차를 사용하는 것을 포함할 수 있다. 정규화는 z-점수(예를 들어, X5)를 포함하여 본원에 기재된 임의의 것일 수 있다.

[0165] 벡터는 뉴클레오티드에 대응하는 전기 신호 세그먼트의 변동을 나타내는 제2 세그먼트 통계값을 포함할 수 있다. 벡터는 제1 세그먼트 통계값의 정규화된 값을 나타내는 제3 세그먼트 통계값을 포함할 수 있다. 벡터는 본원에 기재된 변수 X1, X2 및 X5의 임의의 조합을 포함할 수 있다.

[0166] 제1 영역 통계값은 영역에서의 전기 신호의 평균 또는 중앙값을 나타낼 수 있다. 예를 들어, 제1 영역 통계값은 X3일 수 있다. 구현예에서, 제1 영역 통계값은 영역의 전기 신호의 평균 또는 중앙값으로부터 전기 신호의 변동 절대값의 중앙값 또는 평균을 나타낼 수 있다. 변동은 표준 편차일 수 있다. 예를 들어, 제1 영역 통계값은 X4일 수 있다. 일부 구현예에서, 제1 영역 통계값은 선택적일 수 있다.

[0167] 입력 데이터 구조는 영역에서 전기 신호의 평균 또는 중앙값으로부터 전기 신호의 변동 절대값의 중앙값 또는 평균을 나타내는 제2 영역 통계값을 추가로 포함할 수 있다. 예를 들어, 제2 영역 통계값은 X4일 수 있다.

[0168] 제1 영역 통계값은 윈도우 내의 상이한 뉴클레오티드에 대해 동일한 값일 수 있다. 제2 영역 통계값은 윈도우 내의 상이한 뉴클레오티드에 대해 동일한 값일 수 있다. 결과적으로, 제1 영역 통계값 및 제2 영역 통계값은 제1 세그먼트 통계값 및/또는 제2 세그먼트 통계값을 갖는 벡터와 별개로 간주될 수 있다. 대안적으로, 값이 뉴클레오티드 전체에 걸쳐 동일하더라도 각 뉴클레오티드에 있어서 벡터는 또한 제1 영역 통계값을 포함할 수 있고/있거나 제2 영역 통계값이 벡터에 포함될 수 있다. 영역 통계값을 반복하는 접근이 IPM 524 및 IPM 624에 예시되었다.

[0169] 영역은 샘플 핵산 분자의 한 가닥 상에 있을 수 있다. 일부 구현예에서, 영역은 샘플 핵산 분자의 두 가닥 상에 있을 수 있다. 윈도우는 샘플 핵산 분자의 두 가닥의 뉴클레오티드를 포함할 수 있다. 영역은 샘플 핵산 분자일 수 있다. 영역은 적어도 5, 10, 15, 20 25, 30, 50, 100, 200, 300, 400, 500, 1000, 5000, 10000, 50000 또는 100만개의 뉴클레오티드를 포함할 수 있다. 일부 구현예에서, 영역은 50, 100, 200, 300, 400, 500, 1000, 5000, 10000, 50000 또는 100만개 미만의 뉴클레오티드일 수 있다. 영역은 표적 위치에서 뉴클레오티드 주위가 중앙에 있을 수 있다.

[0170] 뉴클레오티드 윈도우는 표적 위치에서 뉴클레오티드 주위가 중앙에 있을 수 있다. 일부 구현예에서, 윈도우는 표적 위치에서 뉴클레오티드 주위가 중앙에 있지 않을 수 있다. 윈도우는 표적 위치에서 뉴클레오티드로부터 X-nt 상류 및 Y-nt 하류를 포함할 수 있다. X는 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, 및 10000을 포함할 수 있지만 이에 제한되지 않으며; Y는 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000 및 10000을 포함할 수 있지만 이에 제한되지 않는다. 윈도우의 최소 뉴클레오티드 수는 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 100, 200개, 또는 표적 위치의 상류 및 하류의 뉴클레오티드 수 중 임의의 것의 합보다 1이 더 클 수 있다. 윈도우는 도 5에 나타내고 기재된 윈도우과 유사할 수 있다.

[0171] 윈도우는 도 6에 기재된 기술과 유사하게 핵산 분자의 두 가닥을 포함할 수 있다.

[0172] 블록(1220)에서, 입력 데이터 구조가 모델에 입력된다. 모델은 제1 복수의 제1 데이터 구조를 수신하여 훈련된다. 제1 복수의 데이터 구조의 각 제1 데이터 구조는 복수의 제1 핵산 분자의 각 핵산 분자에서 시퀀싱된 뉴클레오티드의 각 윈도우에 대응한다. 각 제1 핵산 분자는 뉴클레오티드에 대응하는 전기 신호를 측정하여 시퀀싱된다. 변형은 각 제1 핵산 분자의 각 윈도우의 표적 위치에서 뉴클레오티드의 알려진 제1 상태를 갖는다. 각 제1 데이터 구조는 입력 데이터 구조와 동일한 특성에 대한 값을 포함한다. 모델은 본원에 기재된 임의의 기계 학습 모델일 수 있다.

[0173] 모델은 복수의 제1 훈련 샘플을 저장하여 추가로 훈련된다. 각 제1 훈련 샘플은 제1 복수의 제1 데이터 구조 중 하나 및 표적 위치에서 뉴클레오티드의 제1 상태를 표시하는 제1 표지를 포함한다. 또한, 제1 복수의 제1 데이터 구조가 모델에 입력될 때, 모델은 복수의 제1 훈련 샘플을 사용하여, 제1 표지의 대응하는 표지와 일치하거나 일치하지 않는 모델의 출력에 기반하여 모델의 매개변수를 최적화함으로써 모델을 훈련한다. 모델의 출력은 각 윈도우의 표적 위치에서 뉴클레오티드가 변형을 갖는지 여부를 특정한다. 훈련은 도 13에서 후술된 바와 같을 수 있다.

[0174] 블록(1230)에서, 모델을 사용하여 변형이 입력 데이터 구조의 윈도우 내 표적 위치에서 뉴클레오티드에 존재하는지 여부가 결정된다.

[0175] 변형 상태가 추가 분석에서 사용될 수 있다. 임산부로부터 얻은 샘플에서, 본 개시내용의 구현예는 메틸화 상태에 기반하여 혈장 DNA 분자의 태아 또는 모체 기원을 결정하기 위해 사용될 수 있다. 모체 또는 태아 기원은 참조값보다 높거나 낮은 메틸화 수준을 갖는 게놈 영역에 의해 결정될 수 있다. 구현예에서, 임산부로부터 얻은 샘플은 무세포, 예를 들어, 혈장 또는 혈청일 수 있다. 일부 구현예에서, 샘플 핵산 분자는 사전 결정된 게놈 영역에 대해 정렬함으로써 식별될 수 있다. 사전 결정된 게놈 영역은 태아 또는 모체 게놈에서 과메틸화 또는 저메틸화된 것이 알려져 있을 수 있다. 방법은 표적 위치에서 뉴클레오티드의 변형 상태 및 선택적으로 샘플 핵산 분자의 하나 이상의 다른 뉴클레오티드의 변형 상태를 사용하여 샘플 핵산이 태아 또는 모체 기원임을 결정하는 단계를 포함할 수 있다.

[0176] 샘플 핵산 분자가 태아 또는 모체 기원인지 여부를 결정하는 단계는 하나 이상의 뉴클레오티드의 메틸화 상태를 사용하여 샘플 핵산 분자의 메틸화 수준을 결정하는 것을 포함할 수 있다. 샘플 핵산 분자의 메틸화 수준은 참조값과 비교될 수 있다. 참조값은 하나 이상의 모체 핵산 분자의 메틸화 수준으로부터 결정될 수 있다. 샘플 핵산 분자의 메틸화 수준을 참조값과 비교하는 것은 샘플 핵산 분자의 메틸화 수준이 참조값보다 낮음을 결정하는 것을 포함할 수 있다. 샘플 핵산 분자가 태아 또는 모체 기원인지 여부를 결정하는 단계는 비교를 사용하여 샘플 핵산 분자가 태아 기원임을 결정하는 것을 포함할 수 있다.

[0177] 일부 구현예에서, 샘플 핵산 분자는 복수의 샘플 핵산 분자 중 하나의 샘플 핵산 분자일 수 있다. 방법은 메틸화 상태를 사용하여 복수의 샘플 핵산 분자 각각이 태아 또는 모체 기원인지 여부를 결정하는 단계를 추가로 포함할 수 있다. 태아 분율은 복수의 샘플 핵산 분자의 태아 또는 모체 기원의 결정을 사용하여 결정될 수 있다.

[0178] 일부 구현예에서, 변형 상태는 복사수 이상이 영역에 존재하는지 여부를 결정하기 위해 사용될 수 있다. 변형은 메틸화일 수 있다. 샘플 핵산 분자는 무세포 일 수 있고, 태아를 임신한 여성 대상체의 생물학적 샘플로부터 얻어질 수 있다. 샘플 핵산 분자는 복수의 샘플 핵산 분자 중 하나의 샘플 핵산 분자일 수 있다. 방법은 복수의 샘플 핵산 분자를 태아 게놈의 영역에 대해 정렬함으로써 식별하는 단계를 추가로 포함할 수 있다. 복수의 샘플 핵산 분자 중 각 샘플 핵산 분자의 하나 이상의 뉴클레오티드의 변형 상태가 결정될 수 있다. 영역의 메틸화 수준은 복수의 샘플 핵산 분자 중 각 샘플 핵산 분자에 대한 하나 이상의 뉴클레오티드의 메틸화 상태를 사용하여 결정될 수 있다. 방법은 메틸화 수준을 사용하여 태아 게놈의 영역에서 카피수 이상이 존재하는지 여부를 결정하는 단계를 추가로 포함할 수 있다. 영역은 염색체일 수 있고, 방법은 카피수 이상이 존재하는지 결정하는 단계 및 태아가 염색체 이수성을 갖는지 결정하는 단계를 추가로 포함할 수 있다.

[0179] 변형은 하나 이상의 뉴클레오티드에 존재하는 것으로 결정될 수 있다. 장애의 분류는 하나 이상의 뉴클레오티드에서 변형의 존재를 사용하여 결정될 수 있다. 장애의 분류는 변형의 수를 사용하는 것을 포함할 수 있다. 변형의 수는 역치과 비교될 수 있다. 대안적으로 또는 추가적으로, 분류는 하나 이상의 변형의 위치를 포함할 수 있다. 하나 이상의 변형의 위치는 핵산 분자의 서열 판독을 참조 게놈에 대해 정렬함으로써 결정될 수 있다. 장애와 연관된 것으로 알려진 특정 위치가 변형을 갖는 것으로 나타나면 장애가 결정될 수 있다. 예를 들어, 메틸화 부위의 패턴이 장애에 대한 참조 패턴과 비교될 수 있으며, 장애의 결정은 비교에 기반할 수 있다. 참조 패턴과의 일치 또는 참조 패턴과의 상당한 일치(예를 들어, 80%, 90% 또는 95% 이상)는 장애 또는 장애의 높은 가능성을 표시할 수 있다. 장애는 임의의 임신 관련 장애(예를 들어, 자간전증, 자궁내 성장 제한, 침습성 태반 및 조산)일 수 있다.

[0180] 통계적으로 유의한 수의 핵산 분자가 하나 이상의 임신 대상체에서 장애, 조직 기원, 또는 임상적으로 관련된 DNA 분율에 대한 정확한 결정을 제공하기 위해 분석될 수 있다. 일부 구현예에서, 적어도 1,000개의 핵산 분자가 분석된다. 다른 구현예에서, 적어도 10,000, 또는 50,000, 또는 100,000, 또는 500,000, 또는 1,000,000 또는 5,000,000개 이상의 핵산 분자가 분석될 수 있다. 추가 예로서, 적어도 10,000 또는 50,000 또는 100,000 또는 500,000 또는 1,000,000 또는 5,000,000개의 서열 판독이 생성될 수 있다.

[0181] 방법은 장애의 분류가 대상체가 장애를 갖는지를 결정하는 단계를 포함할 수 있다. 분류는 변형의 수 및/또는 변형 부위를 사용하여, 장애의 수준을 포함할 수 있다.

[0182] 태아 DNA 분율, 태아 메틸화 프로파일, 모체 메틸화 프로파일, 각인 유전자 영역의 존재가 하나 이상의 뉴클레오티드에서 변형의 존재를 사용하여 결정될 수 있다.

[0183] 공정(1200)은 추가적인 구현, 예컨대 아래에 기재된 및/또는 본원의 다른 곳에 기재된 하나 이상의 다른 공정과 관련하여 임의의 단일 구현 또는 임의의 구현의 조합을 포함할 수 있다.

[0184] 도 12는 공정(1200)의 예시적 블록을 나타내지만, 일부 구현에서 공정(1200)은 도 12에 도시된 것에 비해 추가

블록, 더 적은 수의 블록, 상이한 블록, 또는 상이하게 배열된 블록을 포함할 수 있다. 추가적으로 또는 대안적으로, 공정(1200)의 블록 중 2개 이상이 병렬로 수행될 수 있다.

[0185] B. 모델 훈련

[0186] **도 13**은 핵산 분자에서 뉴클레오티드의 변형을 검출하는 예시적 방법(1300)을 나타낸다. 예시적 방법(1300)은 변형을 검출하기 위해 모델을 훈련하는 방법일 수 있다. 변형은 메틸화를 포함할 수 있다. 메틸화는 본원에 기재된 임의의 메틸화를 포함할 수 있다. 변형은 개별 상태를 가질 수 있고, 예컨대 메틸화 및 비메틸화될 수 있고, 잠재적으로 메틸화 유형을 특정할 수 있다. 따라서 뉴클레오티드의 2개 초과의 상태(분류)가 있을 수 있다. 도 13의 훈련은 도 12의 방법(1200)과 함께 사용될 수 있다.

[0187] 블록(1310)에서, 복수의 제1 데이터 구조가 수신된다. 데이터 구조의 다양한 예가 본원에, 예를 들어 도 5 및 도 6에 기재되어 있다. 제1 복수의 제1 데이터 구조 중 각 제1 데이터 구조는 복수의 제1 핵산 분자 중 각 핵산 분자에서 시퀀싱된 뉴클레오티드의 각 윈도우에 대응할 수 있다. 제1 복수의 데이터 구조와 관련된 각 윈도우는 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21개 이상의 연속 뉴클레오티드를 포함하는 4개 이상의 연속 뉴클레오티드를 포함할 수 있다. 각 윈도우는 동일한 수의 연속 뉴클레오티드를 가질 수 있다. 윈도우는 중첩될 수 있다. 각 윈도우는 제1 핵산 분자의 제1 가닥의 뉴클레오티드 및 제1 핵산 분자의 제2 가닥의 뉴클레오티드를 포함할 수 있다. 제1 데이터 구조는 또한 윈도우 내 각 뉴클레오티드에 대해 가닥 특성값을 포함할 수 있다. 가닥 특성은 뉴클레오티드의 존재 또는 제1 가닥 또는 제2 가닥 중 하나를 표시할 수 있다. 윈도우는 제1 가닥의 대응하는 위치의 뉴클레오티드와 상보적이지 않은 제2 가닥의 뉴클레오티드를 포함할 수 있다. 일부 구현예에서, 제2 가닥의 모든 뉴클레오티드는 제1 가닥의 뉴클레오티드와 상보적이다. 일부 구현예에서, 각 윈도우는 제1 핵산 분자의 한 가닥에만 뉴클레오티드를 포함할 수 있다.

[0188] 제1 복수의 제1 데이터 구조는 5,000 내지 10,000, 10,000 내지 50,000, 50,000 내지 100,000, 100,000 내지 200,000, 200,000 내지 500,000, 500,000 내지 1,000,000 또는 1,000,000개 이상의 제1 데이터 구조를 포함할 수 있다. 복수의 제1 핵산 분자는 적어도 1,000, 10,000, 50,000, 100,000, 500,000, 1,000,000, 5,000,000개 이상의 핵산 분자를 포함할 수 있다. 추가 예로서, 적어도 10,000 또는 50,000 또는 100,000 또는 500,000 또는 1,000,000 또는 5,000,000개의 서열 판독이 생성될 수 있다.

[0189] 각각의 제1 핵산 분자는 뉴클레오티드에 대응하는 전기 신호를 측정하여 시퀀싱된다. 전기 신호는 나노포어 시퀀싱으로부터 나올 수 있다.

[0190] 변형은 각 제1 핵산 분자의 각 윈도우의 표적 위치에서 뉴클레오티드의 알려진 제1 상태를 갖는다. 제1 상태는 변형이 뉴클레오티드에 부재하는 것일 수 있거나 변형이 뉴클레오티드에 존재하는 것일 수 있다. 변형은 제1 핵산 분자에 부재하는 것으로 알려져 있을 수 있거나, 제1 핵산 분자는 변형이 부재하도록 처리를 거칠 수 있다. 변형은 제1 핵산 분자에 존재하는 것으로 알려져 있을 수 있거나, 제1 핵산 분자는 변형이 존재하도록 처리를 거칠 수 있다. 제1 상태가 변형이 부재하는 것인 경우, 변형은 각 제1 핵산 분자의 각 윈도우에 부재할 수 있고 표적 위치에만 부재하는 것이 아닐 수 있다. 알려진 제1 상태는 제1 데이터 구조의 제1 부분에 대한 메틸화 상태 및 제1 데이터 구조의 제2 부분에 대한 비메틸화 상태를 포함할 수 있다. 메틸화에 대해 알려진 제1 상태는 바이설파이트 시퀀싱을 사용하거나 단일 분자 실시간 시퀀싱으로부터의 광학 신호를 사용하는 기술을 통해 결정될 수 있다.

[0191] 표적 위치가 각 윈도우의 중앙에 있을 수 있다. 짝수의 뉴클레오티드에 걸친 윈도우의 경우, 표적 위치는 윈도우 중앙의 바로 상류 또는 바로 하류의 위치일 수 있다. 일부 구현예에서, 표적 위치는 첫 번째 위치 또는 마지막 위치를 포함하여 각 윈도우의 임의의 다른 위치일 수 있다. 예를 들어, 윈도우가 첫 번째 위치부터 n번째 위치(상류 또는 하류)까지 한 가닥의 n개 뉴클레오티드에 걸친 경우, 표적 위치는 첫 번째 위치부터 n번째 위치까지 중 임의의 것일 수 있다.

[0192] 각 제1 데이터 구조는 윈도우 내의 특성에 대한 값을 포함한다. 특성은 블록(1210)에 기재된 특성 중 임의의 것일 수 있다.

[0193] 블록(1320)에서, 복수의 제1 훈련 샘플이 저장된다. 각 제1 훈련 샘플은 제1 복수의 제1 데이터 구조 중 하나 및 표적 위치에서 뉴클레오티드의 변형에 대한 제1 상태를 표시하는 제1 표지를 포함한다.

[0194] 블록(1330)에서, 제2 복수의 제2 데이터 구조가 수신된다. 블록(1330)은 선택적이다. 제2 복수의 제2 데이터 구조의 각 제2 데이터 구조는 복수의 제2 핵산 분자의 각 핵산 분자에서 시퀀싱된 뉴클레오티드의 각 윈도우에 대응한다. 제2 복수의 핵산 분자는 복수의 제1 핵산 분자와 동일하거나 상이할 수 있다. 변형은 각 제2 핵산 분자

의 각 윈도우 내 표적 위치에서 뉴클레오티드의 알려진 제2 상태를 갖는다. 제2 상태는 제1 상태와 상이한 상태이다. 예를 들어 제1 상태가 변형이 존재하는 것인 경우, 제2 상태는 변형이 부재하는 것이며 그 반대의 경우도 마찬가지이다. 각 제2 데이터 구조는 제1 복수의 제1 데이터 구조와 동일한 특성에 대한 값을 포함한다.

[0195] 블록(1340)에서, 복수의 제2 훈련 샘플이 저장된다. 블록(1340)은 선택적이다. 각 제2 훈련 샘플은 제2 복수의 제2 데이터 구조 중 하나 및 표적 위치에서 뉴클레오티드의 변형에 대한 제2 상태를 표시하는 제2 표지를 포함한다.

[0196] 블록(1350)에서, 모델은 복수의 제1 훈련 샘플 및 선택적으로 복수의 제2 훈련 샘플을 사용하여 훈련된다. 훈련은 제1 복수의 제1 데이터 구조 및 선택적으로 제2 복수의 제2 데이터 구조가 모델에 입력될 때 제1 표지 및 선택적으로 제2 표지의 대응하는 표지와 일치하거나 일치하지 않는 모델의 출력에 기반하여 모델의 매개변수를 최적화하여 수행된다. 모델의 출력은 각 윈도우의 표적 위치에서 뉴클레오티드가 변형을 갖는지 여부를 특정한다. 모델이 이상값을 제1 상태와 상이한 상태인 것으로 식별할 수 있으므로 방법은 복수의 제1 훈련 샘플만을 포함할 수 있다. 모델은 기계 학습 모델로도 지칭되는 통계 모델일 수 있다.

[0197] 일부 구현예에서, 모델의 출력은 각 복수의 상태에 있을 확률을 포함할 수 있다. 확률이 가장 높은 상태가 상태로 간주될 수 있다.

[0198] 모델은 컨볼루션 신경망(CNN)을 포함할 수 있다. CNN은 제1 복수의 데이터 구조 및 선택적으로 제2 복수의 데이터 구조를 필터링하도록 구성된 컨볼루션 필터 세트를 포함할 수 있다. 필터는 본원에 기재된 임의의 필터일 수 있다. 각 레이어에 대한 필터의 수는 10 내지 20, 20 내지 30, 30 내지 40, 40 내지 50, 50 내지 60, 60 내지 70, 70 내지 80, 80 내지 90, 90 내지 100, 100 내지 150, 150 내지 200개 이상일 수 있다. 필터의 커널 크기는 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 15 내지 20, 20 내지 30, 30 내지 40 이상일 수 있다. CNN은 필터링된 제1 복수의 데이터 구조 및 선택적으로 필터링된 제2 복수의 데이터 구조를 수신하도록 구성된 입력 레이어을 포함할 수 있다. CNN은 복수의 노드를 포함하는 복수의 은닉 레이어를 포함할 수도 있다. 복수의 은닉 레이어 중 제1 레이어는 입력 레이어에 커플링된다. CNN은 복수의 은닉 레이어 중 마지막 레이어에 커플링되고 출력 데이터 구조를 출력하도록 구성된 출력 레이어를 추가로 포함할 수 있다. 출력 데이터 구조는 특성을 포함할 수 있다.

[0199] 모델은 순환 신경망(RNN)을 포함할 수 있다. RNN 모델은 측정 윈도우에서 복수의 뉴클레오티드와 관련된 다수의 장단기 메모리(LSTM) 장치를 포함한다. LSTM 장치의 수는 측정 윈도우의 뉴클레오티드 수와 동일할 수 있다. 일부 구현예에서, LSTM 장치의 수는 측정 윈도우의 뉴클레오티드 수보다 적을 수 있다. LSTM 장치의 수는 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 40, 50, 100, 200, 300, 400, 500, 1,000, 2,000, 3,000, 4,000, 5,000, 10,000, 50,000개 등일 수 있지만 이에 제한되지 않는다. 하나의 LSTM 장치는 여러 회의 선형 또는 비선형 변환을 거칠, 전류 신호 특징과 관련된 정보를 다음 LSTM 장치로 전송할 수 있었다. LSTM 장치를 통한 이러한 정보 전송은 일반적으로 순차적 방식으로(예를 들어, 시간 단계에 따라) 조직된다. LSTM 장치를 통한 이러한 정보 전송은 양방향일 수 있다(즉, 시간적 순서 및 예약된 시간적 순서를 포함함). 각 LSTM 장치는 프로그래밍 가능 작업, 예컨대 망각 게이트, 입력 게이트, 셀 상태 및 출력 게이트를 포함한다. 이러한 연산을 통해 하나의 LSTM은 이전 시간 단계에서 오는 전류 신호 정보를 기억할지, 아니면 관련이 없고 망각할 수 있는지 여부를 결정할 수 있다(망각 게이트). 하나의 LSTM 장치는 이러한 장치에 대한 입력으로부터 새로운 정보를 학습하려고 시도한다(입력 게이트). 장치는 업데이트된 정보를 현재 시간 단계에서 다음 시간 단계로 전달한다(출력 게이트). 본원에서 셀 상태는 모든 시간 단계와 더불어 정보를 운반한다. LSTM 장치의 여러 레이어가 사용될 수 있다. LSTM 레이어의 수는 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30개 등일 수 있다. 레이어 간 전체 연결이 사용될 수 있다. 시그모이드 함수가 일반적으로 입력, 출력 및 망각 게이트에 대한 게이팅 함수로 사용된다. 시그모이드 함수의 출력값은 0 내지 1일 수 있으며, 게이트에 걸쳐 정보의 흐름 없음 또는 완전한 흐름을 결정한다. 쌍곡선 탄젠트 활성화 함수(Tanh로도 지칭됨)가 출력 게이트로부터의 정보값을 처리하여 -1 내지 1의 값을 갖는 새로운 정보를 형성하는 출력 활성화 함수로 사용될 수 있고, 이는 다음 LSTM 장치로 전달될 수 있다. 일부 구현예에서, 이진 단계 함수, 선형 활성화 함수, 시그모이드 함수, 정류 선형 장치 등을 포함하지만 이에 제한되지 않는 다른 활성화 함수를 사용할 수 있다. LSTM의 마지막 레이어에 의해 생성된 값은 각 뉴런이 완전 연결되는 출력 레이어(즉, 특정 수의 뉴런을 갖는, 조밀한 레이어)로 전달될 수 있다. 조밀한 레이어의 뉴런 수는 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 1000, 2000개 등일 수 있지만 이에 제한되지 않는다. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 100, 5000, 1000 등을 포함하지만 이에 제한되지 않는 수의 조밀한 레이어를 사용할 수 있다. 출력 레이어는 예를 들어 메틸화 상태를 분류하기 위해 사용될 수 있는, 시그모이드 활성화 함수 또는 SoftMax

활성화 함수에 기반하여 메틸화 점수를 출력할 수 있다. 예를 들어, 메틸화 점수가 0.5보다 크면 염기가 메틸화된 것으로 결정된다. 그렇지 않으면 염기가 비메틸화된 것으로 결정된다. 일부 구현예에서, 메틸화 상태를 분류하기 위해 사용되는 역치는 적어도 0.1, 0.2, 0.3, 0.4, 0.6, 0.7, 0.8, 0.9 등일 수 있으나 이에 제한되지 않는다. 일부 구현예에서, 모델에서 일부 뉴런은 과적합 문제를 최소화하기 위해 탈락될 수 있다. 탈락된 뉴런의 백분율은 1%, 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70% 등일 수 있지만 이에 제한되지 않으며, 이는 상이한 레이어에 따라 상이할 수 있다.

[0200] 모델은 지도 학습 모델을 포함할 수 있다. 지도 학습 모델은 분석 학습, 인공 신경망, 역전파, 부스팅(메타 알고리즘), 베이지안 통계, 사례 기반 추론, 의사결정 트리 학습, 귀납적 로직 프로그래밍, 가우스 공정 회귀, 유전적 프로그래밍, 데이터 처리의 그룹 방법, 커널 추정기, 학습 오토마타, 학습 분류기 시스템, 최소 메시지 길이(의사결정 트리, 의사결정 그래프 등), 다중선형 부분공간 학습, 나이브 베이즈 분류기, 최대 엔트로피 분류기, 조건부 랜덤 필드, 최근접 이웃 알고리즘, PAC(얼추 거의 맞춤(probably approximately correct)) 학습, 리플다운(ripple down) 규칙, 지식 획득 방법론, 기호 기계 학습 알고리즘, 하위기호 기계 학습 알고리즘, 지원 벡터 기계, 최소 복잡도 기계(MCM), 랜덤 포레스트, 분류기 앙상블, 순서 분류, 데이터 사전-처리, 불균형 데이터세트 취급, 통계적 관계 학습 또는 다참조 분류 알고리즘인 Proaftn을 포함하는 상이한 접근 및 알고리즘을 포함할 수 있다. 모델은 선형 회귀, 로지스틱 회귀, 심층 순환 신경망(예를 들어, 장단기 메모리, LSTM), 베이즈 분류기, 은닉 마르코프 모델(HMM), 선형 판별 분석(LDA), k-평균 클러스터링, 노이즈 포함 적용의 밀도 기반 공간적 클러스터링(DBSCAN), 랜덤 포레스트 알고리즘, 지원 벡터 머신(SVM) 또는 본원에 기재된 임의의 모델일 수 있다.

[0201] 기계 학습 모델을 훈련하는 것의 일부로서, 기계 학습 모델의 매개변수(예를 들어 신경망의 활성화 함수에 사용될 수 있는, 예컨대 가중치, 역치 등)는 훈련 샘플(훈련 세트)에 기반하여 최적화되어 표적 위치에서 뉴클레오티드의 변형을 분류하는 데 최적화된 정확성을 제공할 수 있다. 다양한 형태의 최적화, 예컨대 역전파, 경험적 위험 최소화, 구조적 위험 최소화가 수행될 수 있다. 샘플의 검증 세트(데이터 구조 및 표지)가 모델의 정확성을 검증하기 위해 사용될 수 있다. 교차 검증이 훈련 및 검증을 위해 훈련 세트의 다양한 부분을 사용하여 수행될 수 있다. 모델은 복수의 하위모델을 포함하여 앙상블 모델을 제공할 수 있다. 하위모델은 일단 조합되면 더 정확한 최종 모델을 제공하는 더 약한 모델일 수 있다.

[0202] V. 예시적 시스템

[0203] **도 14**는 본 발명의 구현예에 따른 측정 시스템(1400)을 예시한다. 나타낸 시스템은 샘플(1405), 예컨대 샘플 홀더(1410) 내 DNA 분자를 포함하며, 샘플(1405)은 검정(1408)과 접촉되어 물리적 특성(1415)의 신호를 제공할 수 있다. 샘플 홀더의 예는 검정의 탐침 및/또는 프라이머 또는 액적이 이동하는 튜브(검정을 포함하는 액적과 함께)를 포함하는 플로우 셀일 수 있다. 샘플로부터의 물리적 특성(1415)(예를 들어, 형광 강도, 전압 또는 전류)은 검출기(1420)에 의해 검출된다. 검출기(1420)는 데이터 신호를 구성하는 데이터 포인트를 얻기 위해 간격(예를 들어, 주기적인 간격)을 두고 측정을 수행할 수 있다. 한 구현예에서, 아날로그에서-디지털로의 전환기는 검출기로부터의 아날로그 신호를 복수 회 디지털 형태로 전환한다. 샘플 홀더(1410) 및 검출기(1420)는 검정 장치, 예를 들어 본원에 기재된 구현예에 따라 시퀀싱을 수행하는 시퀀싱 장치를 형성할 수 있다. 데이터 신호(1425)는 검출기(1420)로부터 로직 시스템(1430)으로 전송된다. 데이터 신호(1425)는 로컬 메모리(1435), 외부 메모리(1440) 또는 저장 장치(1445)에 저장될 수 있다.

[0204] 로직 시스템(1430)은 컴퓨터 시스템, ASIC, 마이크로프로세서 등일 수 있거나 이를 포함할 수 있다. 이는 또한 디스플레이(예를 들어, 모니터, LED 디스플레이 등) 및 사용자 입력 장치(예를 들어, 마우스, 키보드, 버튼 등)를 포함하거나 이와 커플링될 수 있다. 로직 시스템(1430) 및 다른 구성요소는 독립형 또는 네트워크 연결 컴퓨터 시스템의 일부일 수 있거나 검출기(1420) 및/또는 샘플 홀더(1410)를 포함하는 장치(예를 들어, 시퀀싱 장치)에 직접 부착되거나 통합될 수 있다. 로직 시스템(1430)은 또한 처리장치(1450)에서 실행하는 소프트웨어를 포함할 수 있다. 로직 시스템(1430)은 본원에 기재된 방법 중 임의의 것을 수행하기 위해 시스템(1400)을 제어하기 위한 명령을 저장하는 컴퓨터 판독 가능 매체를 포함할 수 있다. 예를 들어, 로직 시스템(1430)은 시퀀싱 또는 다른 물리적 작업이 수행되도록 샘플 홀더(1410)를 포함하는 시스템에 명령어를 제공할 수 있다. 이러한 물리적 작업은 특정 순서로 수행될 수 있다, 예를 들어, 시약은 특정 순서로 첨가 및 제거된다. 이러한 물리적 작업은 샘플을 얻고 검정을 수행하기 위해 사용될 수 있는, 로봇 팔을 포함하는 로봇 시스템에 의해 수행될 수 있다.

[0205] 본원에 언급된 임의의 컴퓨터 시스템은 임의의 적합한 수의 하위시스템을 활용할 수 있다. 이러한 하위시스템의

예를 컴퓨터 시스템(10)에서 **도 15**에 나타낸다. 일부 구현예에서, 컴퓨터 시스템은 단일 컴퓨터 장치를 포함하며, 하위시스템은 컴퓨터 장치의 구성요소일 수 있다. 다른 구현예에서, 컴퓨터 시스템은 각각 내부 구성요소를 갖는 하위시스템인 다수의 컴퓨터 장치를 포함할 수 있다. 컴퓨터 시스템은 데스크톱 및 노트북 컴퓨터, 태블릿, 휴대폰, 다른 모바일 장치 및 클라우드 기반 시스템을 포함할 수 있다.

[0206] 도 15에 나타낸 하위시스템은 시스템 버스(75)를 통해 상호연결된다. 디스플레이 어댑터(82) 등에 커플링된 추가적인 하위시스템, 예컨대 프린터(74), 키보드(78), 저장 장치(들)(79), 모니터(76)(예를 들어, 디스플레이 스크린, 예컨대 LED)를 나타낸다. I/O 컨트롤러(71)에 커플링되는 주변장치 및 입/출력(I/O) 장치는 당분야에 알려진 임의의 수의 수단, 예컨대 입/출력(I/O) 포트(77)(예를 들어, USB, Lightning, Thunderbolt™)에 의해 컴퓨터 시스템에 연결될 수 있다. 예를 들어, I/O 포트(77) 또는 외부 인터페이스(81)(예를 들어, 이더넷, 와이파이 등)가 컴퓨터 시스템(10)을 광역 네트워크, 예컨대 인터넷, 마우스 입력 장치 또는 스캐너에 연결하기 위해 사용될 수 있다. 시스템 버스(75)를 통한 상호연결은 중앙 처리장치(73)가 각 하위시스템과 통신하고 시스템 메모리(72) 또는 저장 장치(들)(79)(예를 들어, 고정 디스크, 예컨대 하드 드라이브 또는 광 디스크)로부터의 복수 명령의 실행뿐만 아니라 하위시스템 간의 정보 교환을 제어할 수 있도록 한다. 시스템 메모리(72) 및/또는 저장 장치(들)(79)는 컴퓨터 판독 가능 매체를 구현할 수 있다. 또 다른 하위시스템은 데이터 수집 장치(85), 예컨대 카메라, 마이크, 가속도계 등이다. 본원에 언급된 임의의 데이터가 하나의 구성요소에서 또 다른 구성요소로 출력될 수 있으며 사용자에게 출력될 수 있다.

[0207] 컴퓨터 시스템은, 예를 들어 외부 인터페이스(81)에 의해, 내부 인터페이스에 의해, 또는 하나의 구성요소에서 또 다른 구성요소로 연결 및 제거될 수 있는 제거 가능 저장 장치를 통해 함께 연결된, 복수의 동일한 구성요소 또는 하위시스템을 포함할 수 있다. 일부 구현예에서, 컴퓨터 시스템, 하위시스템, 또는 장치는 네트워크를 통해 통신할 수 있다. 이러한 경우, 하나의 컴퓨터는 클라이언트로 간주되고 또 다른 컴퓨터는 서버로 간주될 수 있으며, 각각 동일한 컴퓨터 시스템의 일부일 수 있다. 클라이언트 및 서버는 각각 다수의 시스템, 하위시스템 또는 구성요소를 포함할 수 있다.

[0208] 구현예의 측면은 하드웨어 회로(예를 들어, 애플리케이션 특이적 집적 회로 또는 현장 프로그래밍 가능 게이트 어레이)를 사용하고/하거나 모듈식 또는 통합 방식으로 일반적으로 프로그래밍 가능 처리장치를 갖는 컴퓨터 소프트웨어를 사용하여 제어 로직의 형태로 구현될 수 있다. 본원에 사용된 바와 같이, 처리장치는 단일 코어 처리장치, 동일한 통합 칩 상의 멀티 코어 처리장치, 또는 단일 회로 기판 또는 네트워크상의 하드웨어뿐만 아니라 전용 하드웨어 상의 다중 처리 장치를 포함할 수 있다. 본원에 제공된 개시 및 교시에 기반하여, 당업자는 하드웨어 그리고 하드웨어와 소프트웨어의 조합을 사용하여 본 발명의 구현예를 구현하는 다른 방식 및/또는 방법을 알고 인식할 것이다.

[0209] 본 출원에 기재된 임의의 소프트웨어 구성요소 또는 기능은 예를 들어 Java, C, C++, C#, Objective-C, Swift와 같은 임의의 적합한 컴퓨터 언어, 또는 예를 들어 기존 또는 객체 지향 기술을 사용하는 Perl 또는 Python과 같은 스크립팅 언어를 사용하여 처리장치에 의해 실행될 소프트웨어 코드로 구현될 수 있다. 소프트웨어 코드는 저장 및/또는 전송을 위해 컴퓨터 판독 가능 매체에 일련의 명령 또는 명령어로 저장될 수 있다. 적합한 비일시적 컴퓨터 판독 가능 매체는 랜덤 액세스 메모리(RAM), 읽기 전용 메모리(ROM), 자기 매체, 예컨대 하드 드라이브 또는 플로피 디스크, 또는 광학 매체, 예컨대 컴팩트 디스크(CD) 또는 DVD(디지털 다용도 디스크) 또는 블루레이 디스크, 플래시 메모리 등을 포함할 수 있다. 컴퓨터 판독 가능 매체는 이러한 저장 또는 전송 장치의 임의의 조합일 수 있다.

[0210] 이러한 프로그램은 또한 인터넷을 포함하는 다양한 프로토콜을 따르는 유선, 광학 및/또는 무선 네트워크를 통한 전송에 적합한 반송파 신호를 사용하여 인코딩되고 전송될 수 있다. 따라서, 컴퓨터 판독 가능 매체는 이러한 프로그램으로 인코딩된 데이터 신호를 사용하여 생성될 수 있다. 프로그램 코드로 인코딩된 컴퓨터 판독 가능 매체는 호환 장치와 함께 패키지로 제공되거나 다른 장치와 별도로(예를 들어, 인터넷 다운로드를 통해) 제공될 수 있다. 임의의 이러한 컴퓨터 판독 가능 매체는 단일 컴퓨터 제품(예를 들어, 하드 드라이브, CD 또는 전체 컴퓨터 시스템) 상에 또는 내에 있을 수 있으며 시스템 또는 네트워크 내의 상이한 컴퓨터 제품 상에 또는 내에 있을 수 있다. 컴퓨터 시스템은 사용자에게 본원에 언급된 결과 중 임의의 것을 제공하기 위한 모니터, 프린터 또는 다른 적절한 디스플레이를 포함할 수 있다.

[0211] 본원에 기재된 임의의 방법은 단계를 수행하도록 구성될 수 있는, 하나 이상의 처리장치를 포함하는 컴퓨터 시스템으로 전체적으로 또는 부분적으로 수행될 수 있다. 따라서, 구현예는 잠재적으로 각 단계 또는 각 단계 그룹을 수행하는 상이한 구성요소를 사용하여, 본원에 기재된 임의의 방법의 단계를 수행하도록 구성된 컴퓨터 시

스템에 관한 것일 수 있다. 번호가 매겨진 단계로 제시되지만, 본원의 방법 단계는 동시에 또는 상이한 시간에 또는 상이한 순서로 수행될 수 있다. 추가적으로, 이러한 단계의 일부는 다른 방법으로부터의 다른 단계의 일부와 함께 사용될 수 있다. 또한 단계의 전부 또는 일부는 선택적일 수 있다. 추가적으로, 임의의 방법의 임의의 단계는 이들 단계를 수행하기 위한 모듈, 장치, 회로, 또는 시스템의 다른 수단을 사용하여 수행될 수 있다.

[0212] 특정 구현예의 특정 세부사항이 본 발명의 구현예의 정신 및 범위를 벗어나지 않고 임의의 적합한 방식으로 조합될 수 있다. 그러나, 본 발명의 다른 구현예는 각 개별적 측면, 또는 이들 개별적 측면의 특정 조합과 관련된 특정 구현예에 관한 것일 수 있다.

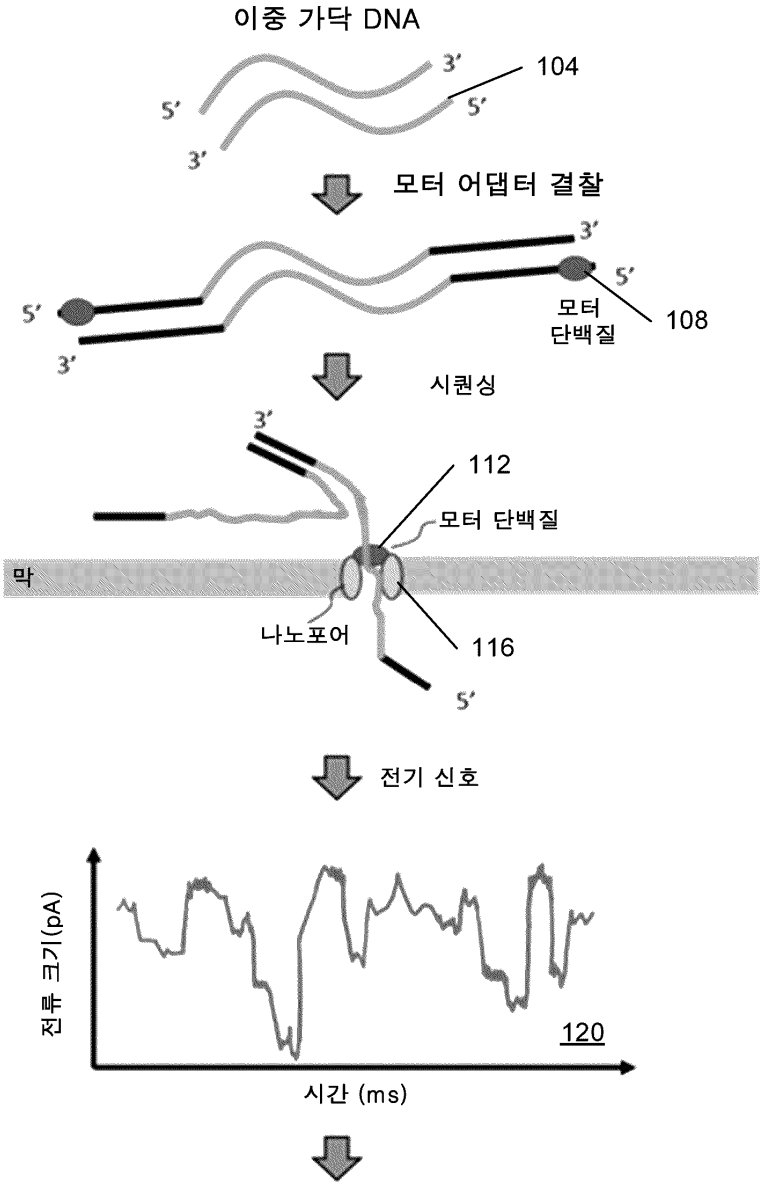
[0213] 본 개시내용의 예시적 구현예에 대한 상기 설명은 예시 및 설명의 목적으로 제시되었다. 이는 본 개시를 기재된 정확한 형태로 배타적이도록 하거나 제한하려는 의도가 아니며, 상기 교시에 비추어 많은 변형 및 변동이 가능하다.

[0214] 관사("a", "an" 또는 "the")의 표현은 달리 구체적으로 표시되지 않는 한 "하나 이상"을 의미하도록 의도된다. "또는"의 사용은 "포괄적 또는"을 의미하도록 의도되며, 달리 구체적으로 표시되지 않는 한 "배타적 또는"을 의미하도록 의도되지 않는다. "제1" 구성요소에 대한 지칭이 반드시 제2 구성요소가 제공되는 것을 필요로 하지는 않는다. 또한, "제1" 또는 "제2" 구성요소에 대한 지칭은 명시적으로 언급되지 않는 한 참조된 구성요소를 특정 위치로 제한하지 않는다. 용어 "~에 기반하는"은 "적어도 부분적으로 ~에 기반하는"을 의미하도록 의도된다.

[0215] 본원에 언급된 모든 특허, 특허 출원, 간행물 및 설명은 모든 목적을 위해 이의 전체가 참조로 포함된다. 어느 것도 선행 기술로 인정되는 것은 아니다.

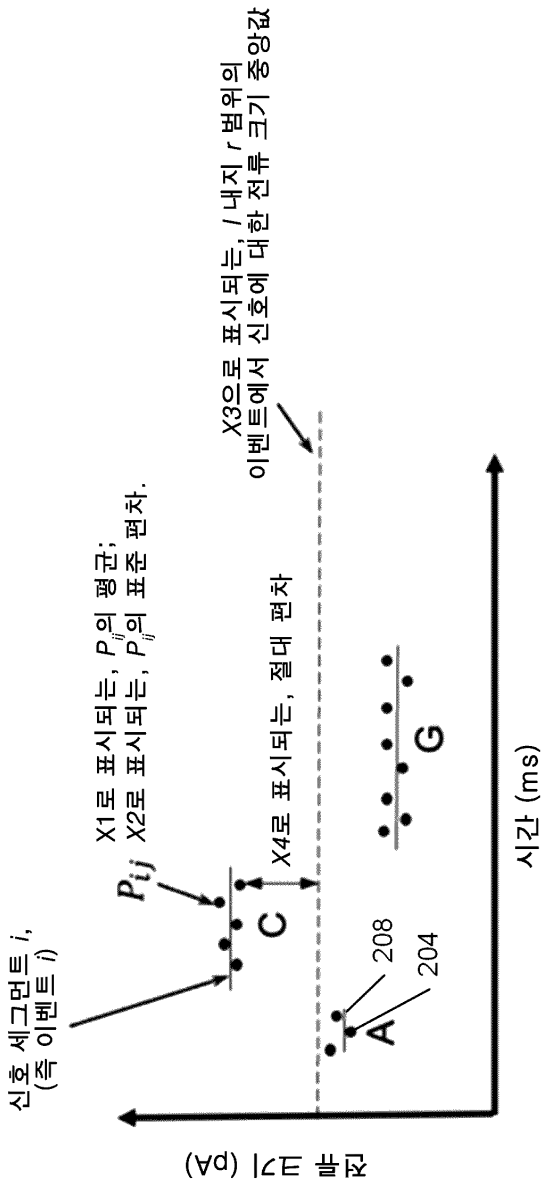
도면

도면1

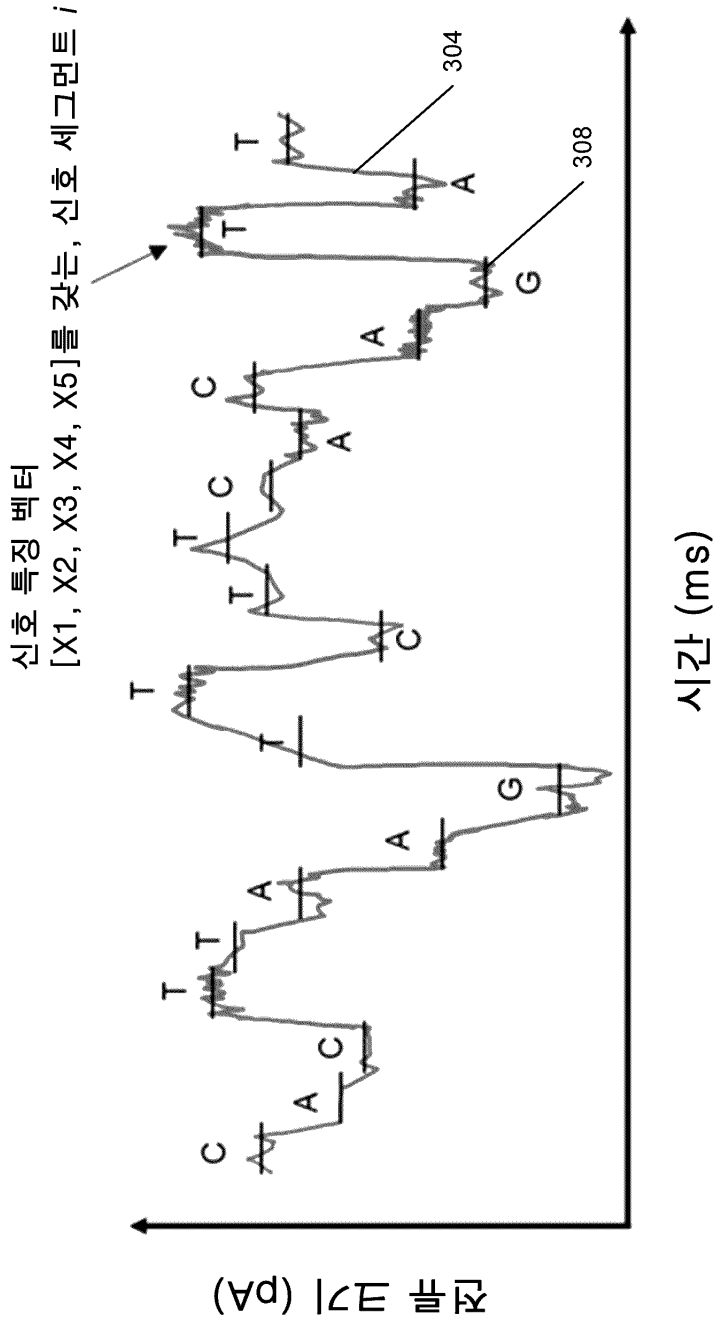


1) 염기 호출
2) 염기 변형 분석

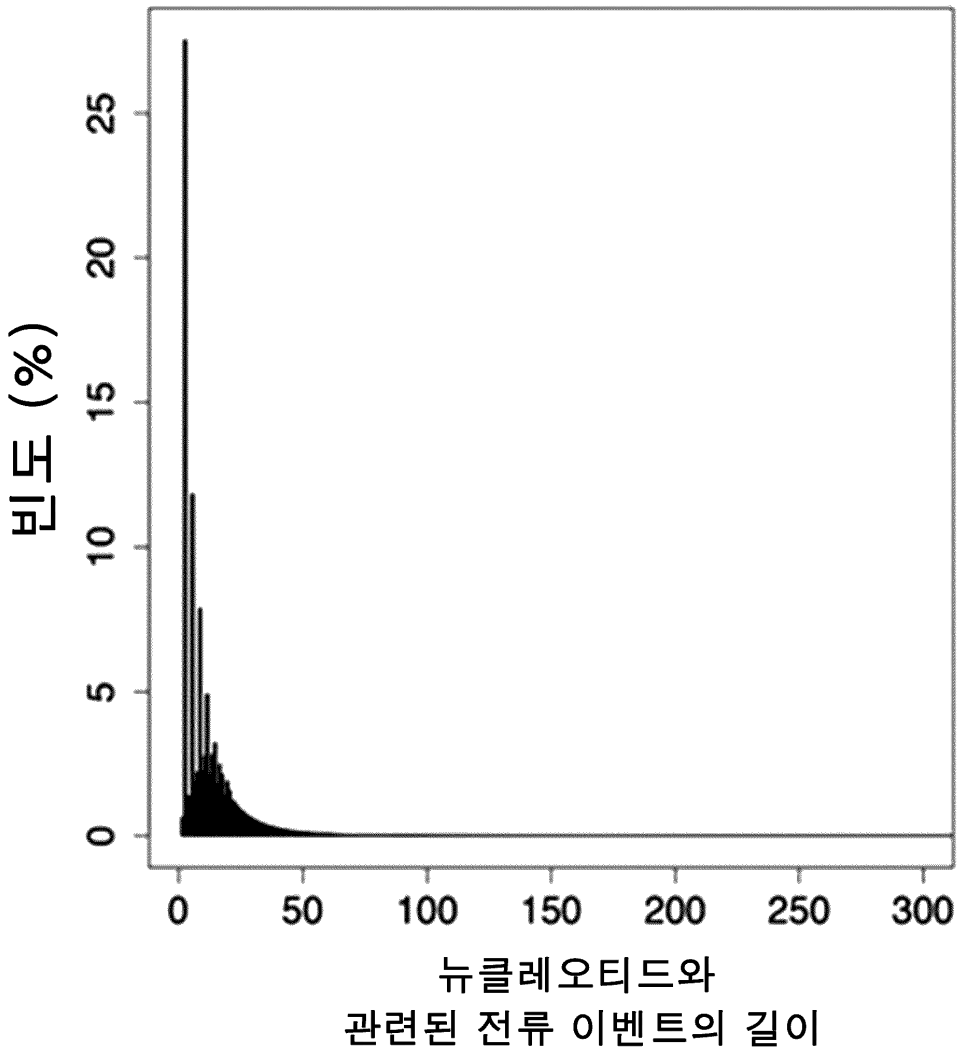
도면2


신호 세그먼트 i,
(즉 이벤트 i)
P_{ij}
X1로 표시되는, P_{ij}의 평균;
X2로 표시되는, P_{ij}의 표준 편차.
X3으로 표시되는, I 내지 r 범위의
이벤트에서 신호에 대한 전류 크기 중앙값
C
X4로 표시되는, 절대 편차
208
204
A
G
시간 (ms)
전류 크기 (pA)

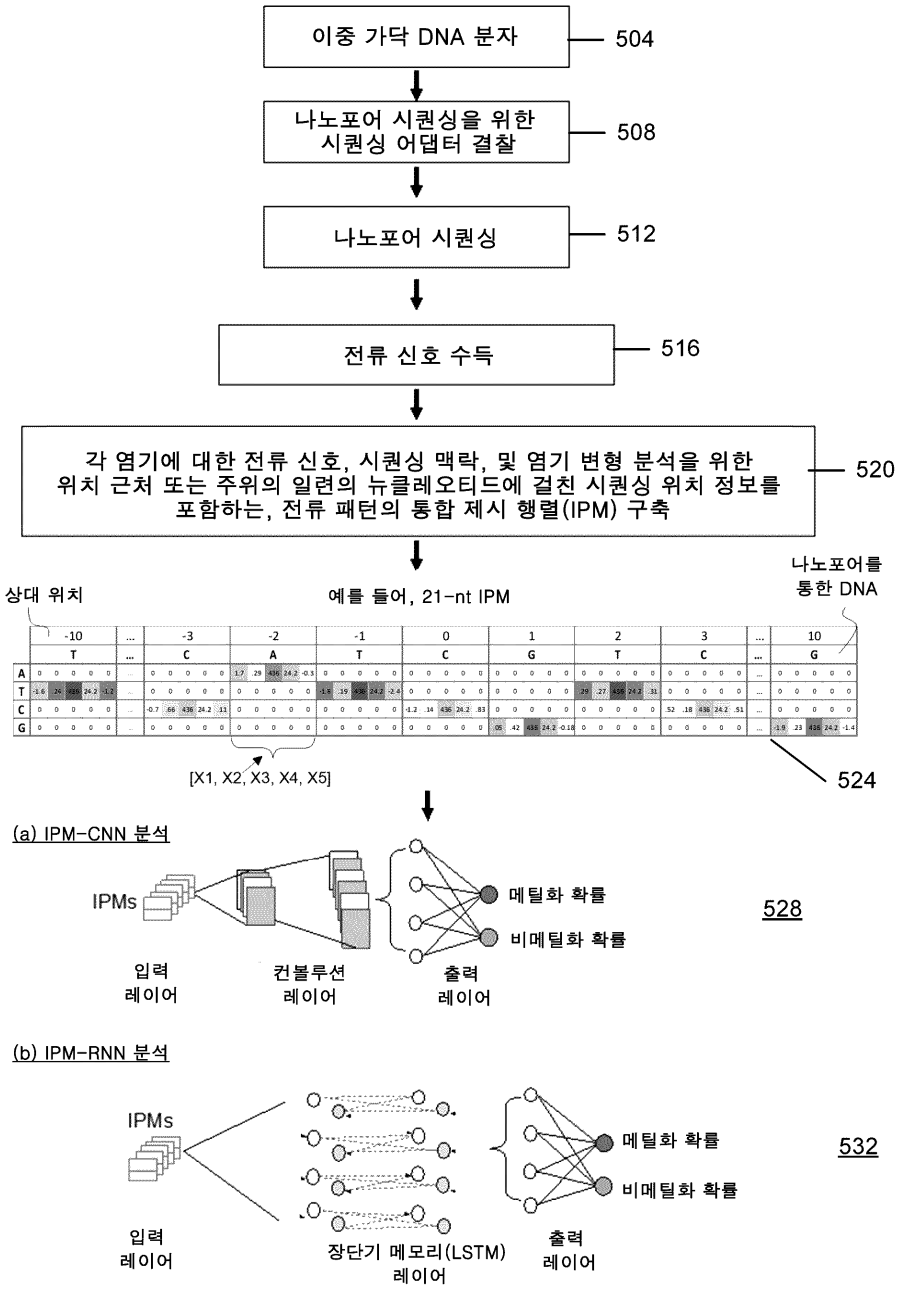
도면3



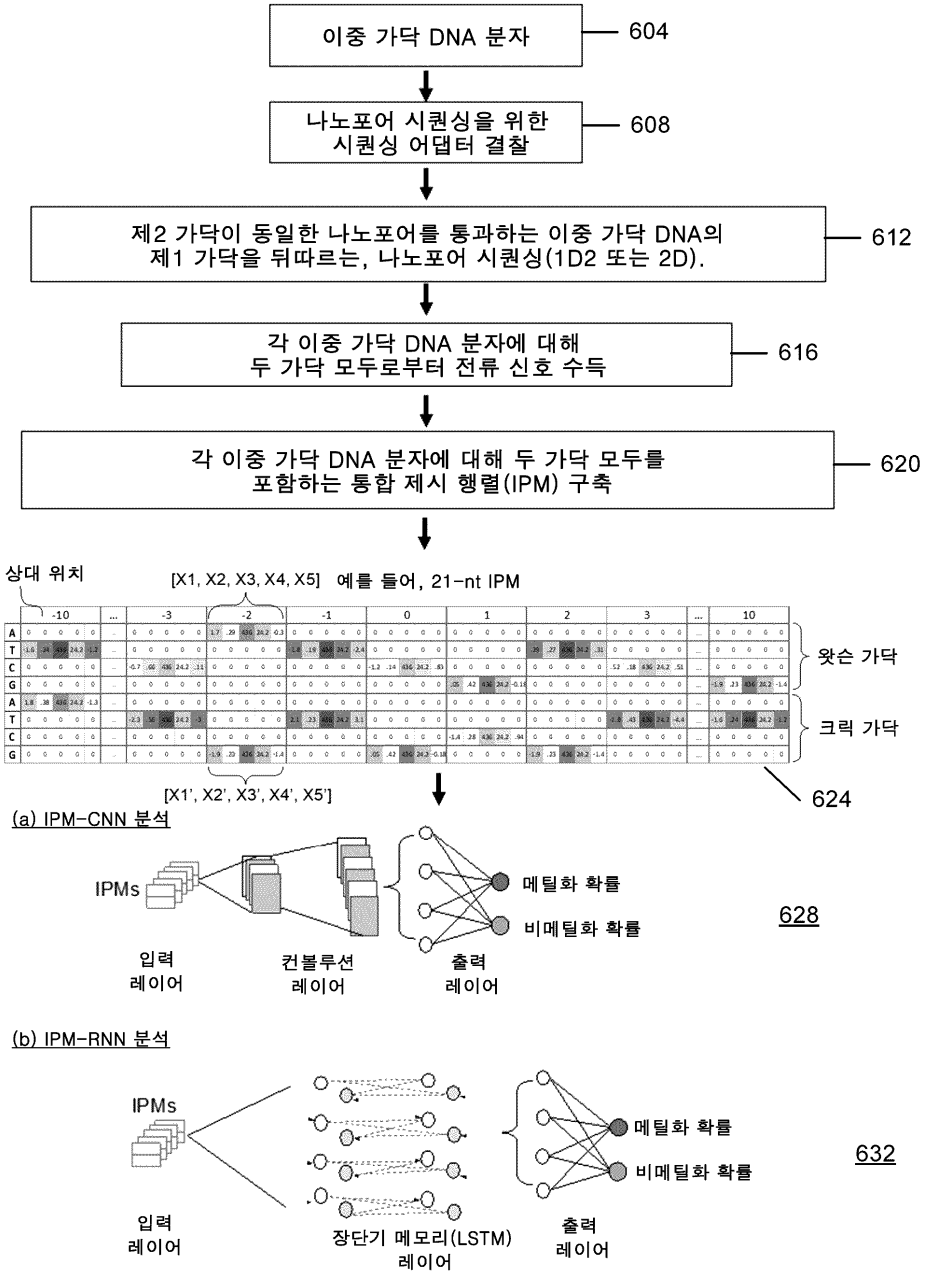
도면4



도면5


이중 가닥 DNA 분자
504
나노포어 시퀀싱을 위한 시퀀싱 어댑터 결찰
508
나노포어 시퀀싱
512
전류 신호 수득
516
각 염기에 대한 전류 신호, 시퀀싱 맥락, 및 염기 변형 분석을 위한 위치 근처 또는 주위의 일련의 뉴클레오티드에 걸친 시퀀싱 위치 정보를 포함하는, 전류 패턴의 통합 제시 행렬(IPM) 구축
520
상대 위치
예를 들어, 21-nt IPM
나노포어를 통한 DNA
[X1, X2, X3, X4, X5]
524
(a) IPM-CNN 분석
IPMs
메틸화 확률
비메틸화 확률
528
입력 레이어
컨볼루션 레이어
출력 레이어
(b) IPM-RNN 분석
IPMs
메틸화 확률
비메틸화 확률
532
입력 레이어
장단기 메모리(LSTM) 레이어
출력 레이어

도면6


이중 가닥 DNA 분자
604
나노포어 시퀀싱을 위한 시퀀싱 어댑터 결찰
608
제2 가닥이 동일한 나노포어를 통과하는 이중 가닥 DNA의 제1 가닥을 뒤따르는, 나노포어 시퀀싱(1D2 또는 2D).
612
각 이중 가닥 DNA 분자에 대해 두 가닥 모두로부터 전류 신호 수득
616
각 이중 가닥 DNA 분자에 대해 두 가닥 모두를 포함하는 통합 제시 행렬(IPM) 구축
620
상대 위치
[X1, X2, X3, X4, X5] 예를 들어, 21-nt IPM
왓슨 가닥
크릭 가닥
[X1', X2', X3', X4', X5']
624
(a) IPM-CNN 분석
IPMs
메틸화 확률
비메틸화 확률
628
입력 레이어
컨볼루션 레이어
출력 레이어
(b) IPM-RNN 분석
IPMs
메틸화 확률
비메틸화 확률
632
입력 레이어
장단기 메모리(LSTM) 레이어
출력 레이어

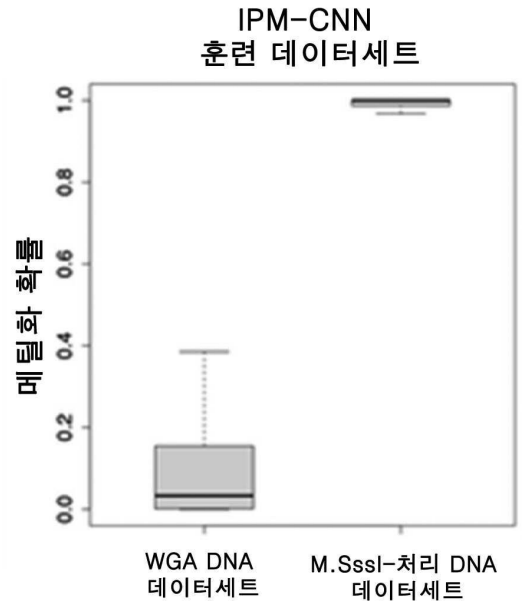
도면7

커널 크기	AUC	
	훈련 데이터세트	시험 데이터세트
1x5	0.98	0.96
1x10	0.98	0.96
1x15	0.97	0.97
1x20	0.98	0.96
1x25	0.98	0.96
1x30	0.97	0.94

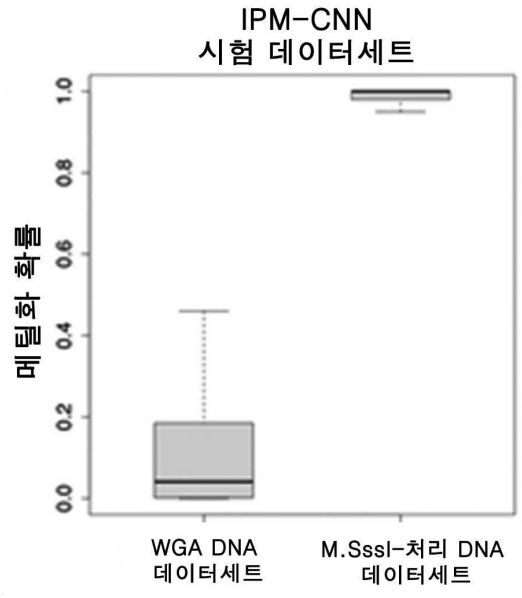
도면8

데이터세트	분자의 수 (CpG 부위의 수)	
	훈련	시험
M.SssI-처리 DNA	7,989 (38,470)	4,826 (9,716)
WGA DNA	8,052 (37,150)	5,041 (11,444)

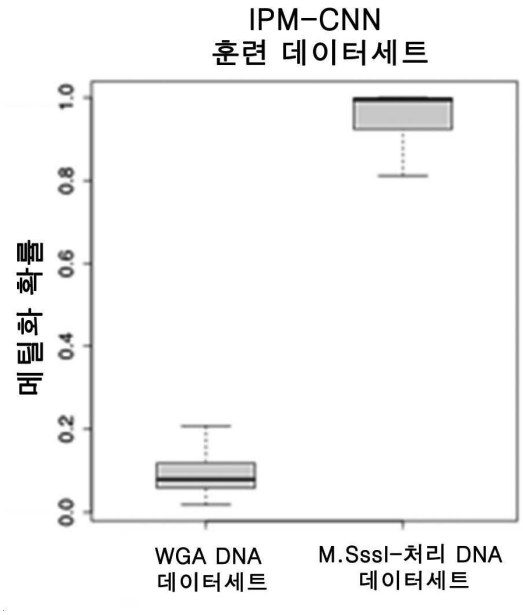
도면9a



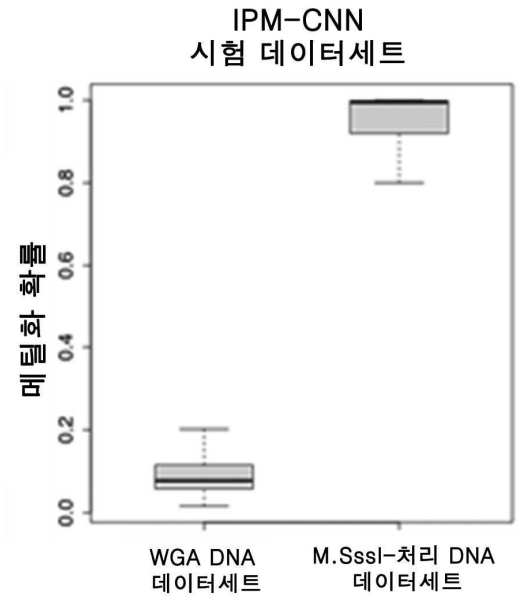
도면9b


IPM-CNN
시험 데이터세트
메틸화 확률
0.0
0.2
0.4
0.6
0.8
1.0
WGA DNA
데이터세트
M.SssI-처리 DNA
데이터세트

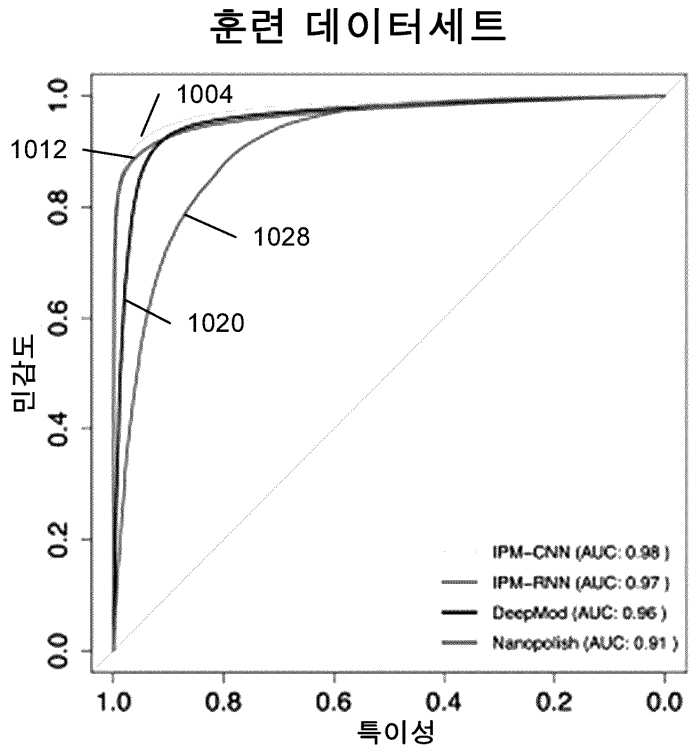
도면9c


IPM-CNN
훈련 데이터세트
메틸화 확률
0.0
0.2
0.4
0.6
0.8
1.0
WGA DNA
데이터세트
M.SssI-처리 DNA
데이터세트

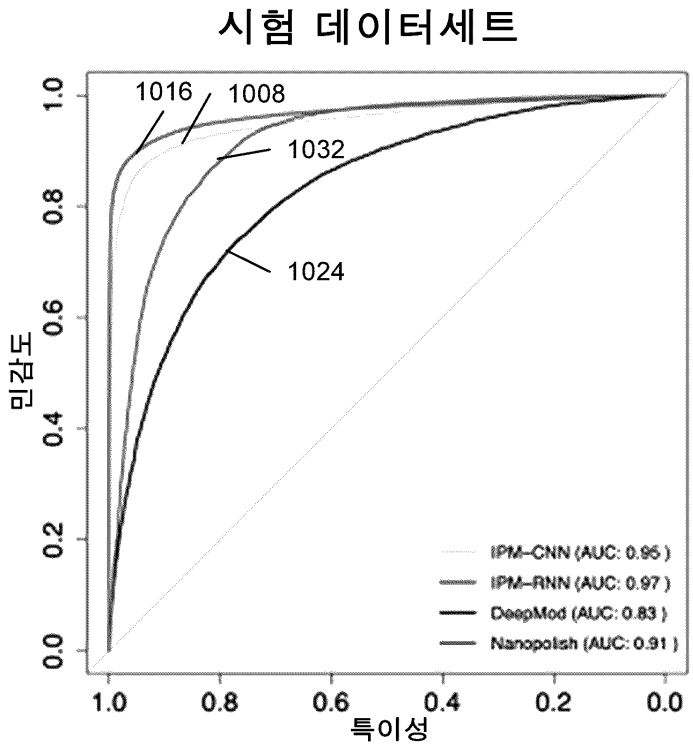
도면9d


IPM-CNN
시험 데이터세트
메틸화 확률
0.0
0.2
0.4
0.6
0.8
1.0
WGA DNA
데이터세트
M.SssI-처리 DNA
데이터세트

도면10a


훈련 데이터세트
1004
1012
1028
1020
민감도
특이성
1.0
0.8
0.6
0.4
0.2
0.0
IPM-CNN (AUC: 0.98)
IPM-RNN (AUC: 0.97)
DeepMod (AUC: 0.96)
Nanopolish (AUC: 0.91)

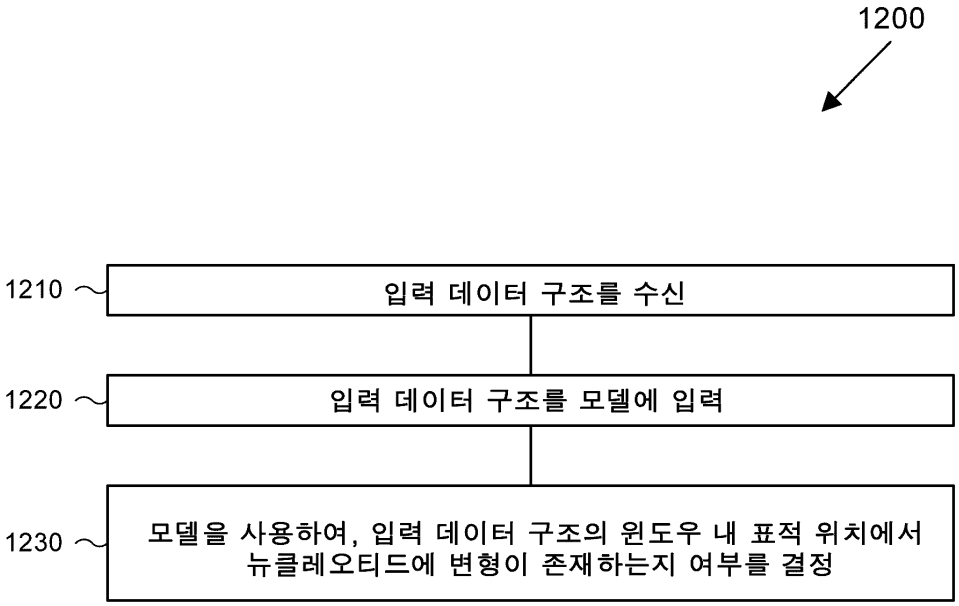
도면10b



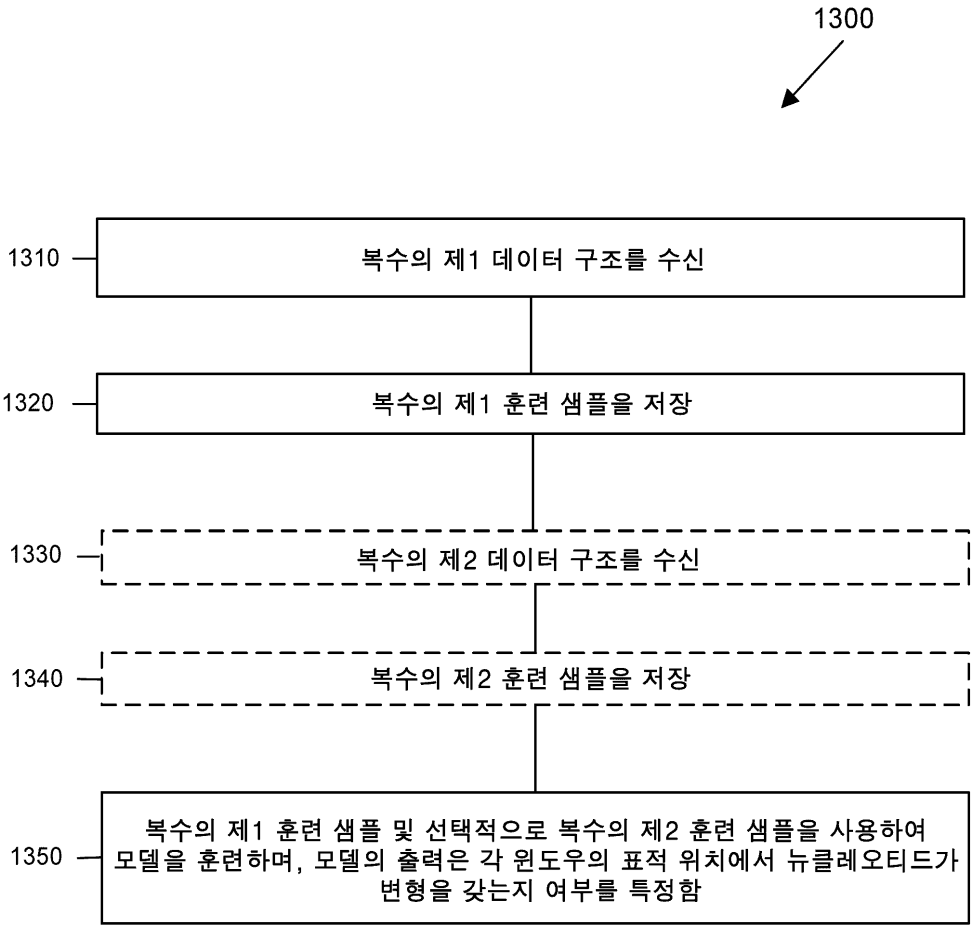
도면11

	민감도	특이성
IPM-CNN	**90%**	90%
IPM-RNN	**93%**	
DeepMod	53%	
nanopolish	74%	
IPM-CNN	**86%**	95%
IPM-RNN	**90%**	
DeepMod	38%	
nanopolish	55%	
IPM-CNN	**70%**	99%
IPM-RNN	**83%**	
DeepMod	13%	
nanopolish	16%	

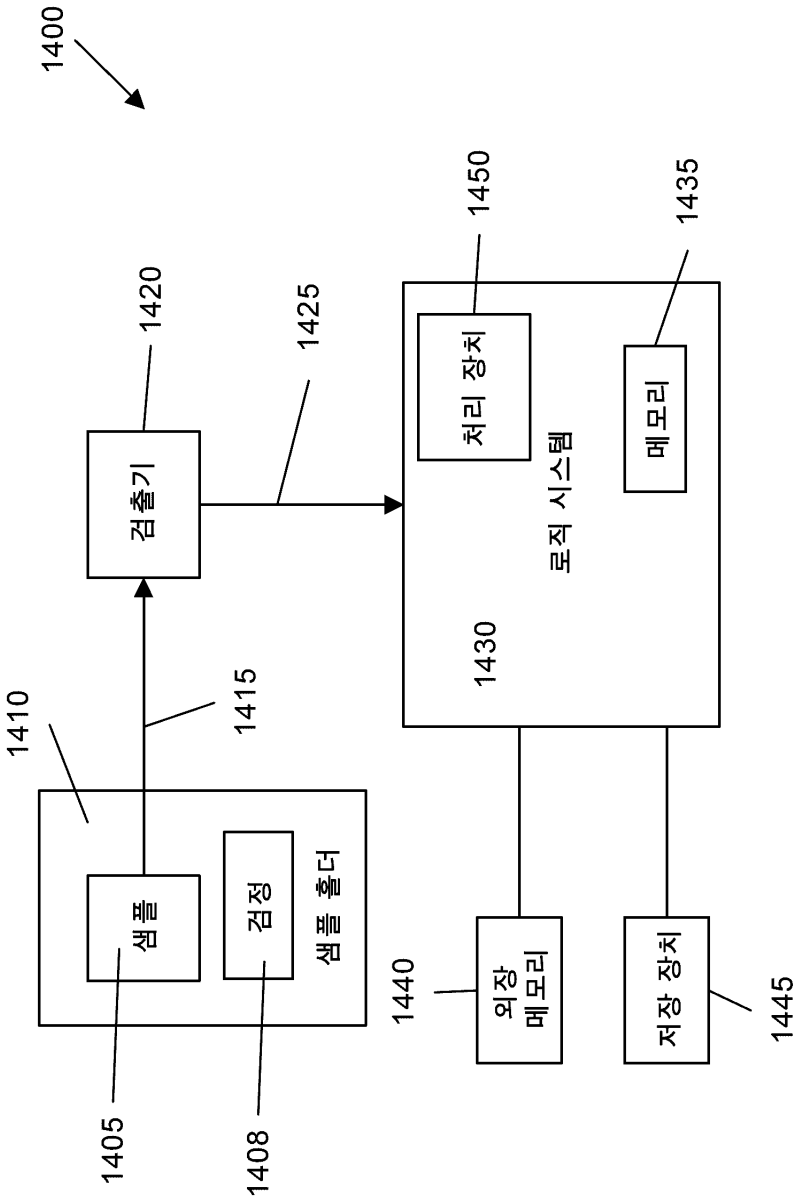
도면12


1200
1210
입력 데이터 구조를 수신
1220
입력 데이터 구조를 모델에 입력
1230
모델을 사용하여, 입력 데이터 구조의 윈도우 내 표적 위치에서 뉴클레오티드에 변형이 존재하는지 여부를 결정

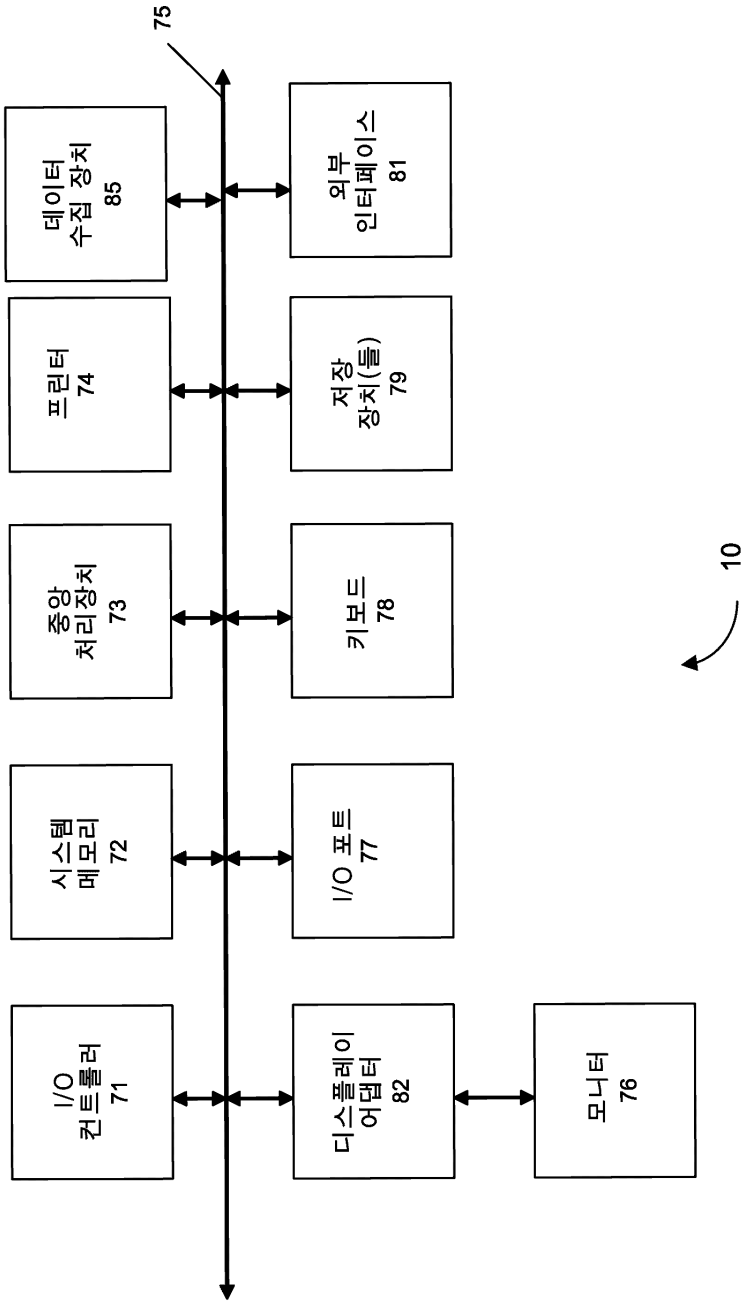
도면13


1300
1310
복수의 제1 데이터 구조를 수신
1320
복수의 제1 훈련 샘플을 저장
1330
복수의 제2 데이터 구조를 수신
1340
복수의 제2 훈련 샘플을 저장
1350
복수의 제1 훈련 샘플 및 선택적으로 복수의 제2 훈련 샘플을 사용하여 모델을 훈련하며, 모델의 출력은 각 윈도우의 표적 위치에서 뉴클레오티드가 변형을 갖는지 여부를 특정함

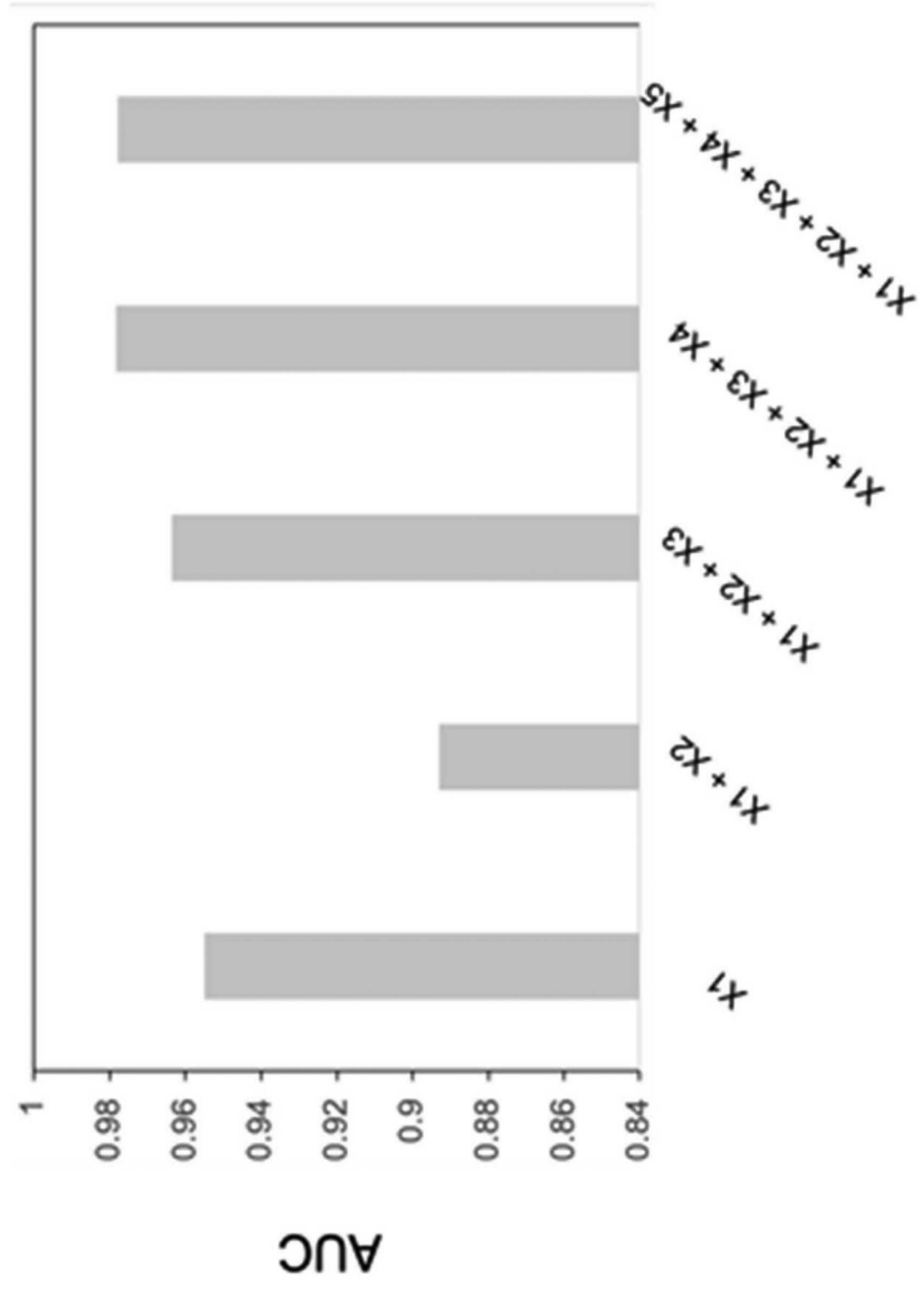
도면14


1400
1420
1425
1450
1435
검출기
처리 장치
메모리
로직 시스템
1430
1410
1415
샘플
검정
샘플 홀더
1405
1408
1440
외장 메모리
저장 장치
1445

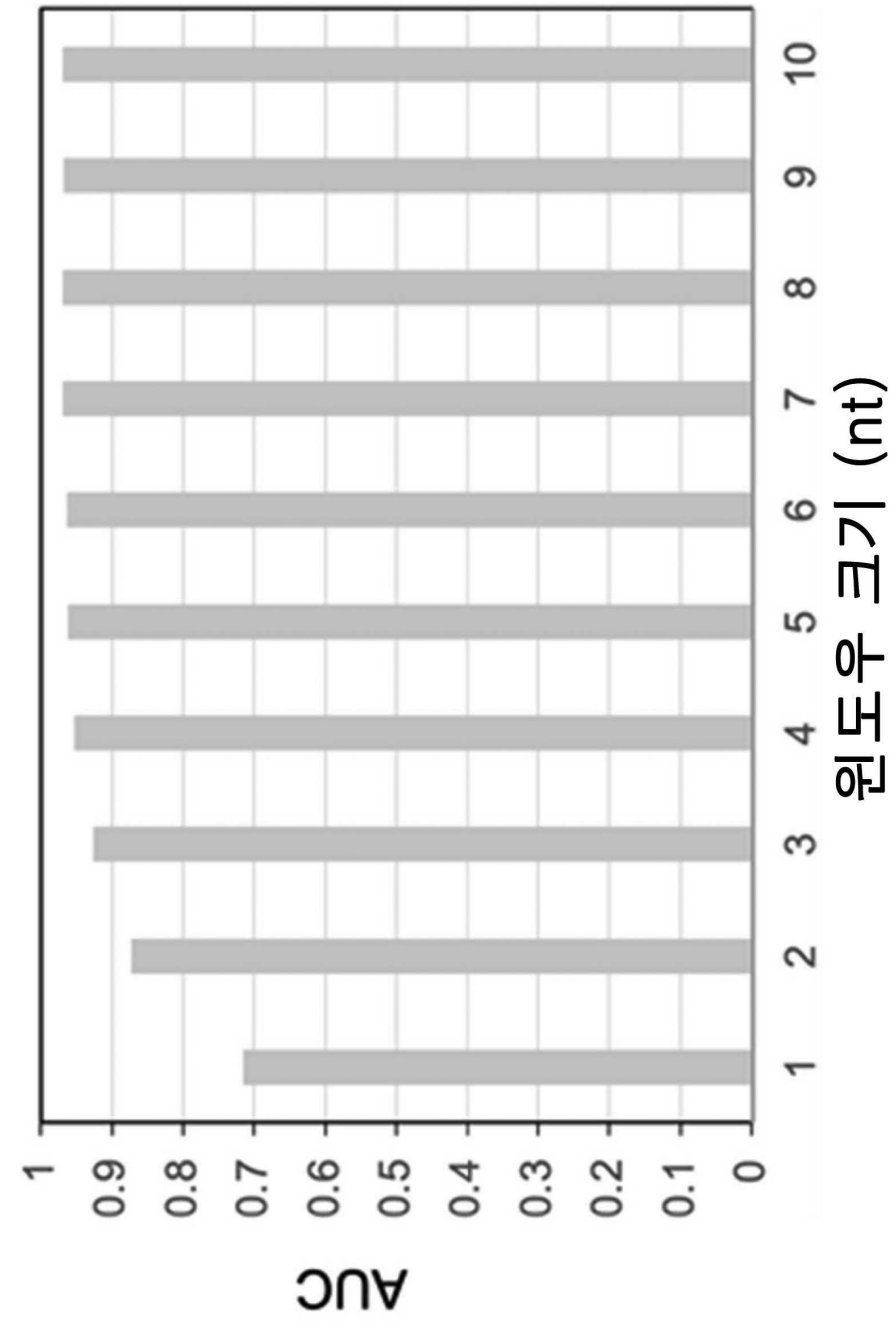
도면15


75
데이터 수집 장치 85
외부 인터페이스 81
프린터 74
저장 장치(들) 79
중앙 처리장치 73
키보드 78
10
시스템 메모리 72
I/O 포트 77
I/O 컨트롤러 71
디스플레이 어댑터 82
모니터 76

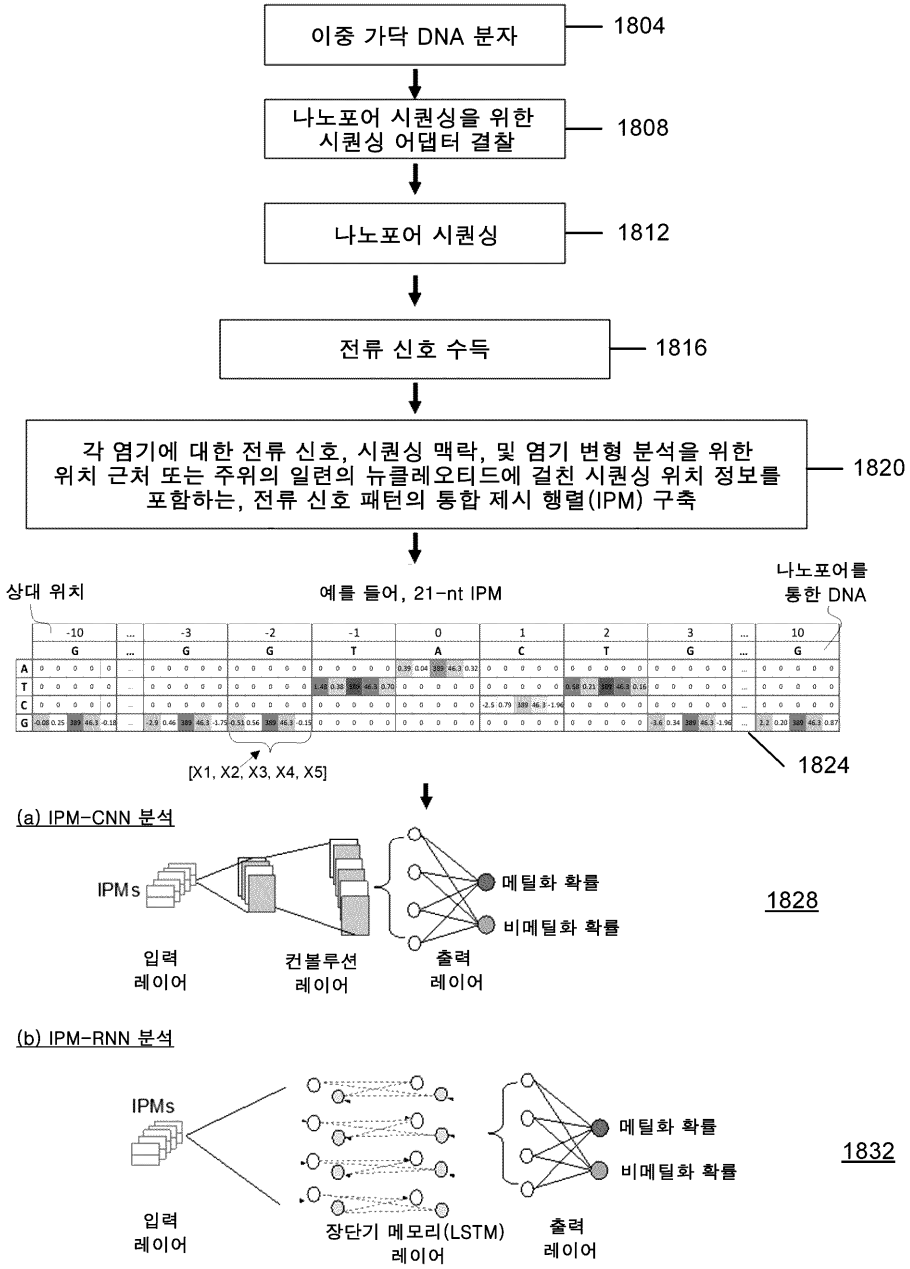
도면16


AUC
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
x7
x7×x2
x7×x2×x3
x7×x2×x3×x4
x7×x2×x3×x4×x5

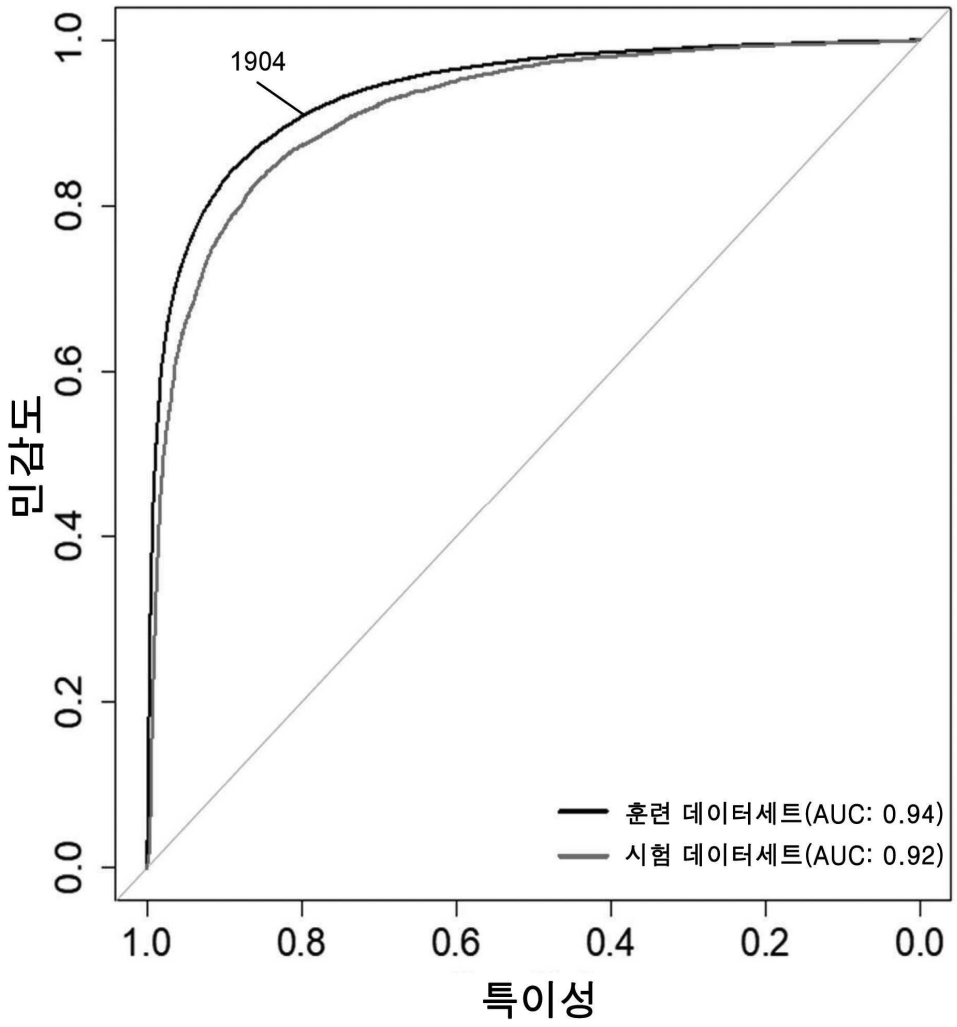
도면17



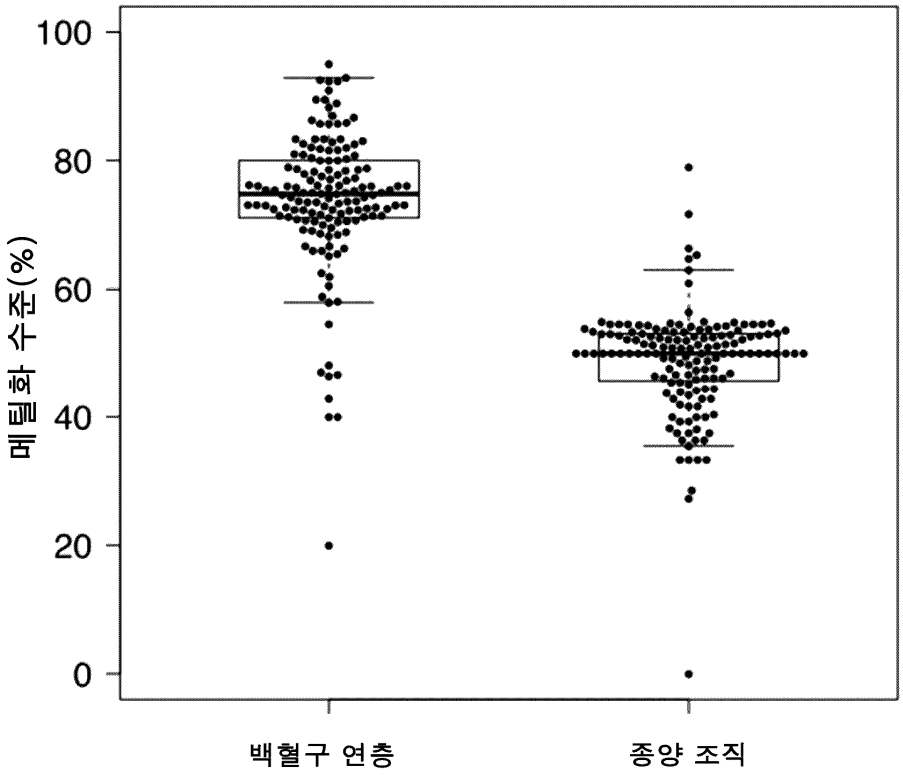
도면18


이중 가닥 DNA 분자
1804
나노포어 시퀀싱을 위한 시퀀싱 어댑터 결찰
1808
나노포어 시퀀싱
1812
전류 신호 수득
1816
각 염기에 대한 전류 신호, 시퀀싱 맥락, 및 염기 변형 분석을 위한 위치 근처 또는 주위의 일련의 뉴클레오티드에 걸친 시퀀싱 위치 정보를 포함하는, 전류 신호 패턴의 통합 제시 행렬(IPM) 구축
1820
상대 위치
예를 들어, 21-nt IPM
나노포어를 통한 DNA
[X1, X2, X3, X4, X5]
1824
(a) IPM-CNN 분석
IPMs
메틸화 확률
비메틸화 확률
입력 레이어
컨볼루션 레이어
출력 레이어
1828
(b) IPM-RNN 분석
IPMs
메틸화 확률
비메틸화 확률
입력 레이어
장단기 메모리(LSTM) 레이어
출력 레이어
1832

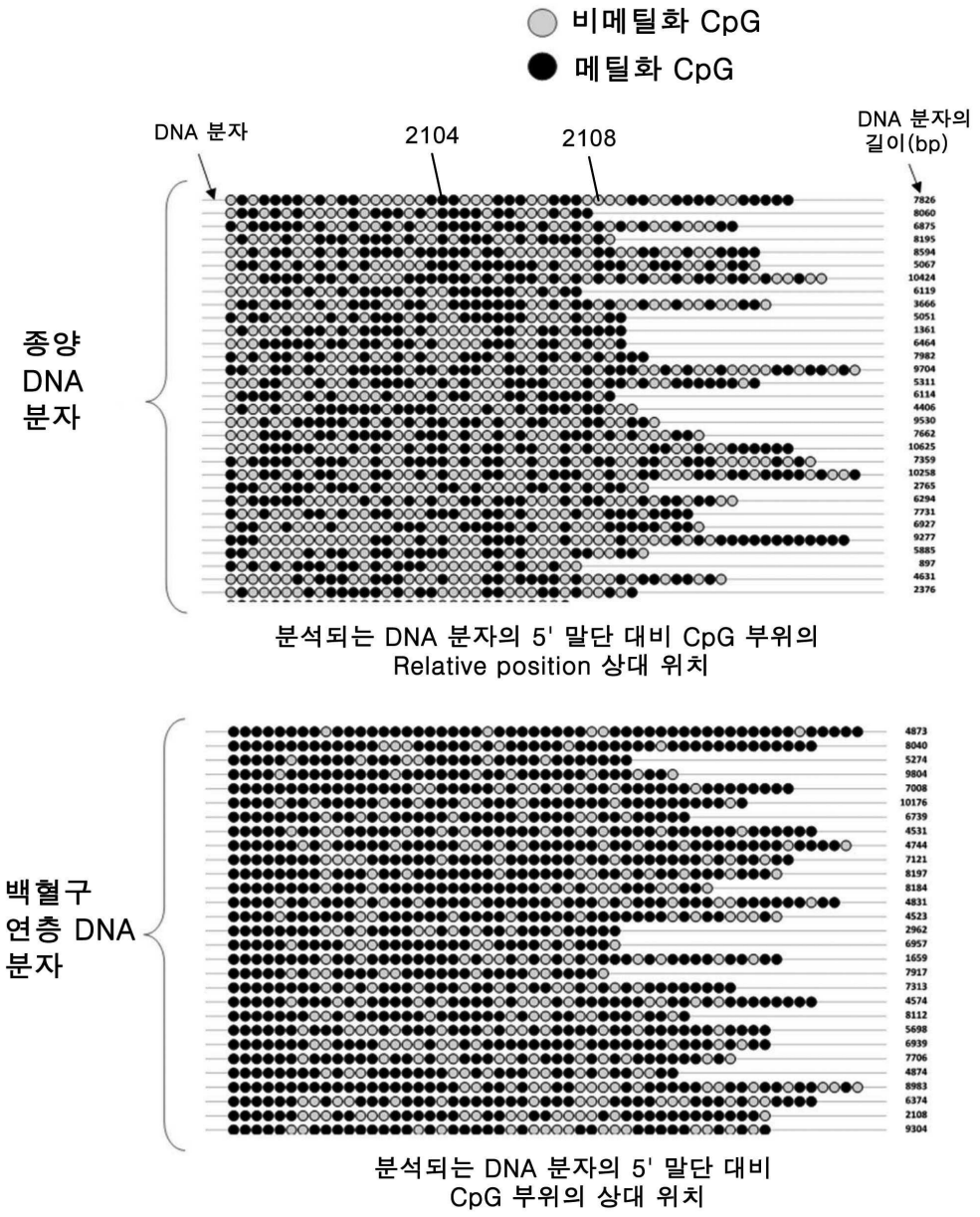
도면19


1904
민감도
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.8
0.6
0.4
0.2
0.0
특이성
훈련 데이터세트(AUC: 0.94)
시험 데이터세트(AUC: 0.92)

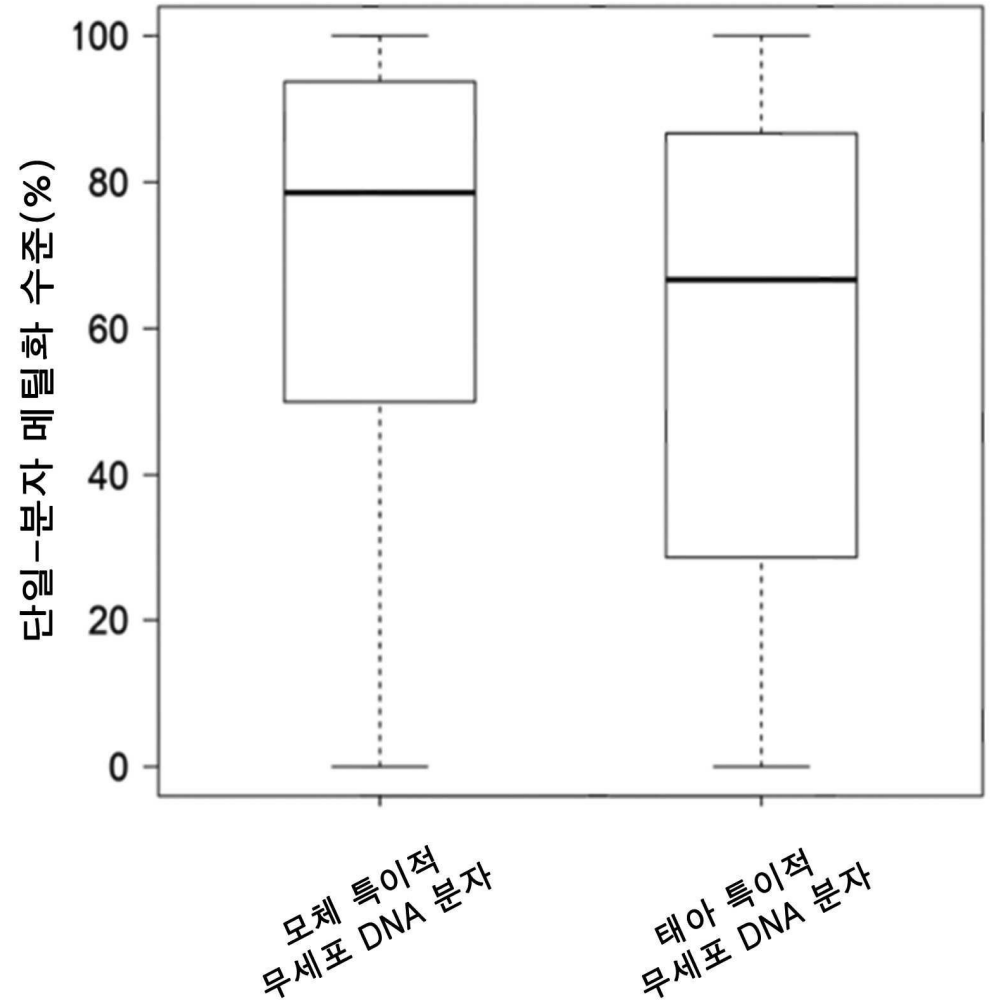
도면20



도면21


비메틸화 CpG
메틸화 CpG
DNA 분자
2104
2108
DNA 분자의 길이(bp)
종양 DNA 분자
분석되는 DNA 분자의 5' 말단 대비 CpG 부위의 Relative position 상대 위치
백혈구 연층 DNA 분자
분석되는 DNA 분자의 5' 말단 대비 CpG 부위의 상대 위치

도면22



도면23