(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0080272
(43) 공개일자 2020년07월06일

(51) 국제특허분류(Int. Cl.)
C12Q 1/6809 (2018.01) *C12Q 1/6883* (2018.01)
C12Q 1/6886 (2018.01) *G16B 20/10* (2019.01)
G16B 40/20 (2019.01)
(52) CPC특허분류
C12Q 1/6809 (2018.05)
C12Q 1/6883 (2018.05)
(21) 출원번호 10-2020-7014693
(22) 출원일자(국제) 2018년11월02일
심사청구일자 없음
(85) 번역문제출일자 2020년05월22일
(86) 국제출원번호 PCT/CN2018/113640
(87) 국제공개번호 WO 2019/085988
국제공개일자 2019년05월09일
(30) 우선권주장
62/580,906 2017년11월02일 미국(US)

(71) 출원인
더 차이니즈 유니버시티 오브 홍콩
중국 홍콩 엔. 티. 새틴 더 차이니즈 유니버시티 오브 홍콩
그레일, 인코포레이티드.
미국 캘리포니아 (우편번호: 94025) 멘로 파크 오브라이언 드라이브 1525
(72) 발명자
로 역밍 데니스
중국 홍콩 뉴 테리토리즈 샤틴 피 치우 빌딩 테크놀로지 라이센싱 오피스 룸 328 내
치우 로사 와이 쿤
중국 홍콩 뉴 테리토리즈 샤틴 피 치우 빌딩 테크놀로지 라이센싱 오피스 룸 328 내
(뒷면에 계속)
(74) 대리인
김진희, 김태홍

전체 청구항 수 : 총 43 항

(54) 발명의 명칭 **비침습적 산전 검사 및 암 검출을 위한 핵산 크기 범위의 용도**

(57) 요 약

염색체 영역이 복제수 이상(copy number aberration) 또는 후성유전체적 변경(epigenetic alteration)을 나타내는지를 결정하기 위해 크기-밴드 분석이 사용된다. 특이적인 크기에 초점을 맞추는 대신 다수의 크기 범위가 분석될 수 있다. 특이적인 크기 대신 다수의 크기 범위를 사용함으로써, 방법은 더 많은 서열 판독을 분석할 수 있고, 심지어 임상적-관련 DNA가 낮은 분율의 생물학적 시료일 수 있는 경우에도 염색체 영역이 복제수 이상을 나타내는지를 결정할 수 있다. 다수의 범위를 사용하는 것은, 게놈 영역에서 선택된 하위세트의 판독보다는 게놈 영역으로부터의 모든 서열 판독을 사용할 수 있게 한다. 분석의 정확도는 유사한 또는 더 높은 특이도에서 더 높은 민감도로 증가될 수 있다. 분석은 동일한 정확도를 달성하기 위해 더 적은 수의 시퀀싱 판독을 포함하여, 더 효율적인 과정을 초래할 수 있다.

(52) CPC특허분류
C12Q 1/6886 (2018.05)
G16B 20/10 (2019.02)
G16B 40/20 (2019.02)
C12Q 2537/16 (2013.01)
C12Q 2600/154 (2013.01)
C12Q 2600/156 (2013.01)
(72) 발명자
찬 콴 체
중국 홍콩 뉴 테리토리즈 샤틴 피 치우 빌딩 날리지 트랜스퍼 오피스 룸 301 내

지앙 페이영
중국 홍콩 뉴 테리토리즈 샤틴 피 치우 빌딩 날리지 트랜스퍼 오피스 룸 301 내

명 세 서

청구범위

청구항 1

염색체 영역이 대상체로부터의 생물학적 시료에서 복제수 이상(copy number aberration)을 나타내는지를 결정하는 방법으로서, 상기 생물학적 시료는 임상적-관련(clinically-relevant) DNA 분자 및 다른 DNA 분자를 포함하는 세포-무함유 DNA 분자의 혼합물을 포함하고, 상기 방법은 하기 단계를 포함하는, 방법:

복수의 크기 범위의 각각의 크기 범위에 대해:

크기 범위에 상응하는 생물학적 시료로부터 세포-무함유 DNA 분자의 제1 양을 측정하는 단계, 및

크기 범위에 상응하는 세포-무함유 DNA 분자의 상기 제1 양, 및 크기 범위 내의 것이 아닌 크기를 포함하는 제2 크기 범위 내의 DNA 분자의 제2 양을 사용하여 크기 비를 컴퓨터 시스템에 의해 계산하는 단계;

복수의 크기 범위에 대한 복수의 기준 크기 비를 포함하는 기준 크기 패턴을 수득하는 단계로서, 상기 기준 크기 패턴은 염색체 영역에 복제수 이상을 갖는 대상체로부터의 또는 복제수 이상을 갖지 않는 대상체로부터의 복수의 기준 시료로부터 결정되는 단계;

복수의 크기 비를 상기 기준 패턴과 비교하는 단계;

상기 비교에 기초하여 염색체 영역이 복제수 이상을 나타내는지를 결정하는 단계.

청구항 2

제1항에 있어서, 상기 임상적-관련 DNA 분자는 태아 DNA 또는 모체 DNA를 포함하는, 방법.

청구항 3

제1항에 있어서, 상기 임상적-관련 DNA 분자는 종양 DNA를 포함하고, 다른 DNA 분자는 비-종양 DNA를 포함하는, 방법.

청구항 4

제2항에 있어서, 상기 복제수 이상은 이수성(aneuploidy)인, 방법.

청구항 5

제3항에 있어서, 상기 복제수 이상은 암의 지표(indication)인, 방법.

청구항 6

제1항에 있어서, 복수의 크기 범위의 각각의 크기 범위는 밴드폭을 특징으로 하는, 방법.

청구항 7

제6항에 있어서, 상기 밴드폭은 50 bp 내지 200 bp 범위인, 방법.

청구항 8

제1항에 있어서, 상기 각각의 크기 범위는 복수의 크기 범위의 임의의 다른 크기 범위와 비-중첩되는, 방법.

청구항 9

제1항에 있어서, 상기 각각의 크기 범위는 복수의 크기 범위의 적어도 하나의 다른 크기 범위와 중첩되는, 방법.

청구항 10

제1항에 있어서, 상기 크기 비는 z-점수를 포함하는, 방법.

청구항 11

제1항에 있어서, 상기 제2 크기 범위는 복수의 크기 범위의 각각의 크기 범위보다 큰 범위인, 방법.

청구항 12

제1항에 있어서, 상기 제2 크기 범위는 생물학적 시료 내의 세포-무함유 DNA 분자의 모든 크기 또는 염색체 영역 내 세포-무함유 DNA 분자의 모든 크기를 포함하는, 방법.

청구항 13

제1항에 있어서, 상기 세포-무함유 DNA 분자는 게놈 영역으로부터의 것인, 방법.

청구항 14

제13항에 있어서, 상기 게놈 영역은 염색체인, 방법.

청구항 15

제13항에 있어서, 상기 게놈 영역은 염색체 아암(arm)인, 방법.

청구항 16

제1항에 있어서,

복수의 크기 비를 기준 크기 패턴과 비교하는 단계는 하기 단계를 포함하는, 방법:

복수의 크기 비의 각각의 크기 비를 상응하는 크기 범위에서 기준 크기 비와 비교하는 단계,

각각의 크기 비가 상응하는 크기 범위에서 기준 크기 비와 통계학적으로 유사한지 결정하는 단계.

청구항 17

제1항에 있어서,

복수의 크기 비를 기준 크기 패턴과 비교하는 단계는 하기 단계를 포함하는, 방법:

복수의 크기 범위에 대한 복수의 크기 비를 포함하는 크기 패턴을 결정하는 단계;

상기 크기 패턴을 기준 크기 패턴과 비교하는 단계,

크기 패턴이 기준 크기 패턴과 유사한 모양을 갖는지를 결정하는 단계.

청구항 18

제16항에 있어서,

상기 기준 크기 패턴은 복제수 이상을 갖는 대상체로부터의 복수의 기준 시료로부터 결정되고,

상기 방법은 추가로 하기 단계를 포함하는, 방법:

비교에 기초하여 염색체 영역이 복제수 이상을 나타내는지를 결정하는 단계.

청구항 19

제1항에 있어서,

기준 크기 패턴을 수득하고 복수의 크기 비를 상기 기준 크기 패턴과 비교하는 단계는 복수의 크기 비를 기계 학습(machine learning) 모델 내로 입력하는 단계를 포함하고,

상기 기계 학습 모델은 복수의 기준 시료로부터의 복수의 트레이닝(training) 크기 패턴을 사용하여 트레이닝된, 방법.

청구항 20

제1항에 있어서, 복수의 크기 비를 기준 크기 패턴과 비교하는 단계는, 상기 복수의 크기 비를, 복수의 기준 시

료로부터 결정되는 복수의 역치 값과 비교하는 단계를 포함하는, 방법.

청구항 21

대상체로부터의 생물학적 시료에서 암 분류를 결정하는 방법으로서, 상기 생물학적 시료는 종양 DNA 분자 및 비-종양 DNA 분자를 포함하는 세포-무함유 DNA 분자의 혼합물을 포함하고, 상기 방법은 하기 단계를 포함하는, 방법:

복수의 크기 범위의 각각의 크기 범위에 대해:

크기 범위에 상응하는 생물학적 시료로부터 메틸화된 세포-무함유 DNA 분자의 제1 양을 측정하는 단계, 및

크기 범위에 상응하는 메틸화된 세포-무함유 DNA 분자의 상기 제1 양, 및 크기 범위 내의 것이 아닌 크기를 포함하는 제2 크기 범위 내의 DNA 분자의 제2 양을 사용하여 메틸화 수준을 컴퓨터 시스템에 의해 계산하는 단계;

복수의 크기 범위에 대한 복수의 기준 메틸화 수준을 포함하는 기준 크기 패턴을 수득하는 단계로서, 상기 기준 크기 패턴은 암을 갖는 대상체로부터의 또는 암을 갖지 않는 대상체로부터의 복수의 기준 시료로부터 결정되는 단계;

복수의 메틸화 수준을 상기 기준 패턴과 비교하는 단계; 및

상기 비교에 기초하여 암의 수준을 결정하는 단계.

청구항 22

제21항에 있어서, 상기 제2 양은 메틸화된 세포-무함유 DNA 분자인, 방법.

청구항 23

제21항에 있어서, 상기 메틸화된 세포-무함유 DNA 분자는 염색체 아암으로부터의 것인, 방법.

청구항 24

제21항에 있어서,

복수의 메틸화 수준을 기준 크기 패턴과 비교하는 단계는 하기 단계를 포함하는, 방법:

복수의 크기 범위의 각각의 메틸화 수준을 상응하는 크기 범위에서 기준 메틸화 수준과 비교하는 단계,

각각의 메틸화 수준이 상응하는 크기 범위에서 기준 메틸화 수준과 통계학적으로 유사한지 결정하는 단계.

청구항 25

제21항에 있어서,

복수의 메틸화 수준을 기준 크기 패턴과 비교하는 단계는 하기 단계를 포함하는, 방법:

복수의 크기 범위에 대한 복수의 메틸화 수준을 포함하는 크기 패턴을 결정하는 단계;

상기 크기 패턴을 기준 크기 패턴과 비교하는 단계,

크기 패턴이 기준 크기 패턴과 유사한 모양을 갖는지를 결정하는 단계.

청구항 26

제24항에 있어서,

상기 기준 크기 패턴은 암을 갖는 대상체로부터의 복수의 기준 시료로부터 결정되고,

상기 방법은 추가로 하기 단계를 포함하는, 방법:

대상체가 암을 갖는지를 결정하는 단계.

청구항 27

제21항에 있어서, 메틸화된 세포-무함유 DNA 분자의 상기 제1 양은 게놈 영역으로부터의 것인, 방법.

청구항 28

제27항에 있어서, 상기 게놈 영역은 염색체 아암이고, 상기 염색체 아암은 1p, 1q, 8p, 8q, 13q 및 14q로 이루어진 군으로부터 선택되는, 방법.

청구항 29

제21항에 있어서, 복수의 메틸화 수준을 기준 크기 패턴과 비교하는 단계는, 복수의 메틸화 수준을 복수의 기준 시료로부터 결정되는 복수의 역치 값과 비교하는 단계를 포함하는, 방법.

청구항 30

제21항에 있어서,

복수의 크기 범위는 M개의 크기 범위를 포함하며,

메틸화된 세포-무함유 DNA 분자의 제1 양을 측정하는 단계는 크기 범위에 상응하고 N개의 게놈 영역에 대한 각각의 게놈 영역에 상응하는 메틸화된 세포-무함유 DNA 분자의 제1 양을 측정하는 단계를 포함하며,

크기 범위에 상응하고 게놈 영역에 상응하는 메틸화된 세포-무함유 DNA의 상기 제1 양 및 제2 양을 사용하여 메틸화 수준을 계산하는 것은 NXM개의 메틸화 수준의 측정 벡터를 발생시키며, N은 1 이상의 정수이고 M은 1 초과의 정수이고,

기준 크기 패턴은 N개의 게놈 영역 및 M개의 크기 범위에 대한 기준 메틸화 수준의 기준 벡터를 포함하고, 상기 기준 크기 패턴은 암을 갖는 대상체로부터의 또는 암을 갖지 않는 대상체로부터의 복수의 기준 시료로부터 결정되고,

복수의 메틸화 수준을 기준 크기 패턴과 비교하는 단계는 상기 측정 벡터를 상기 기준 벡터와 비교하는 단계를 포함하는, 방법.

청구항 31

대상체로부터의 생물학적 시료에서 암 분류를 결정하는 방법으로서, 상기 생물학적 시료는 종양 DNA 분자 및 비-종양 DNA 분자를 포함하는 세포-무함유 DNA 분자의 혼합물을 포함하고, 상기 방법은 하기 단계를 포함하는, 방법:

N개의 게놈 영역의 각각의 게놈 영역에 대해:

M개의 크기 범위의 각각의 크기 범위에 대해:

크기 범위에 상응하고 게놈 영역에 상응하는 생물학적 시료로부터 세포-무함유 DNA 분자의 제1 양을 측정하는 단계, 및

크기 범위에 상응하고 게놈 영역에 상응하는 세포-무함유 DNA 분자의 상기 제1 양, 및 크기 범위 내의 것이 아닌 크기를 포함하는 제2 크기 범위 내의 DNA 분자의 제2 양을 사용하여 크기 비를 컴퓨터 시스템에 의해 계산하여, NXM개의 크기 비의 측정 벡터를 발생시키는 단계로서, N은 1 이상의 정수이고 M은 1 초과의 정수인, 단계;

N개의 게놈 영역 및 M개의 크기 범위에 대한 기준 크기 비의 기준 벡터를 포함하는 기준 크기 패턴을 수득하는 단계로서, 상기 기준 크기 패턴은 암을 갖는 대상체로부터의 또는 암을 갖지 않는 대상체로부터의 복수의 기준 시료로부터 결정되는 단계;

측정 벡터를 상기 기준 패턴과 비교하는 단계; 및

상기 비교에 기초하여 암의 수준을 결정하는 단계.

청구항 32

제31항에 있어서, 각각의 게놈 영역은 염색체 아암인, 방법.

청구항 33

제31항에 있어서,

기준 크기 패턴은 기계 학습 모델을 사용하여 결정되고, 상기 기계 학습 모델은 서포트 벡터 머신(support vector machine), 결정 트리(decision tree), 나이브 베이즈 분류(naive Bayes classification), 로지스틱 회귀(logistic regression), 클러스터링 알고리즘(clustering algorithm), 주성분 분석(principal component analysis), 특이값 분해(singular value decomposition), t-분포 확률적 임베딩(t-distributed stochastic neighbor embedding), 및 인공 신경망(artificial neural network)으로 이루어진 군으로부터 선택되는 하나 이상을 포함하는, 방법.

청구항 34

제31항에 있어서, 측정 벡터를 기준 벡터와 비교하는 단계는, 암을 갖는 것으로 결정된 개체에 대한 그리고 암을 갖지 않는 것으로 결정된 개체에 대한 상이한 게놈 영역에 대한 크기 비를 포함하는 트레이닝 벡터의 트레이닝 세트로 트레이닝된 기계 학습 모델을 사용하는 단계를 포함하는, 방법.

청구항 35

제31항에 있어서, 상기 암은 간세포암종을 포함하는, 방법.

청구항 36

제31항에 있어서, 상기 암의 수준은 암의 확률을 포함하는, 방법.

청구항 37

제31항에 있어서,

기준 크기 패턴을 수득하고 측정 벡터를 기준 벡터와 비교하는 단계는 기계 학습 모델을 사용하는 단계를 포함하며,

상기 기계 학습 모델은 복수의 기준 크기 패턴을 사용하여 트레이닝되었으며,

측정 벡터를 기준 벡터와 비교하는 단계는 상기 기준 벡터에 대한 상기 측정 벡터의 유사성을 특징화하는 컷오프 값을 결정하는 단계를 포함하고,

암의 수준을 결정하는 단계는 상기 컷오프 값을 사용하는, 방법.

청구항 38

제31항에 있어서, 측정 벡터를 기준 벡터와 비교하는 단계는 NXM개의 크기 비를 복수의 기준 시료로부터 결정되는 복수의 역치 값과 비교하는 단계를 포함하는, 방법.

청구항 39

제1항의 동작을 수행하도록 컴퓨터 시스템을 제어하기 위한 복수의 명령을 저장하는 비-일시성(non-transitory) 컴퓨터 판독 가능 매체를 포함하는 컴퓨터 제품.

청구항 40

하기를 포함하는 시스템:

제39항의 컴퓨터 제품; 및

비-일시성 컴퓨터 판독 가능 매체 상에 저장된 명령을 실행하기 위한 하나 이상의 프로세서.

청구항 41

제1항 내지 제39항의 방법 중 임의의 방법을 수행하기 위한 수단을 포함하는 시스템.

청구항 42

제1항 내지 제39항의 방법 중 임의의 방법을 수행하도록 구성된 시스템.

청구항 43

제1항 내지 제39항의 방법 중 임의의 방법의 단계를 각각 수행하는 모듈을 포함하는 시스템.

발명의 설명

기 술 분 야

[0001] 관련 출원의 교차 참조

[0002] 본 출원은 2017년 11월 2일에 출원된 발명의 명칭이 "비침습적 산전 검사 및 암 검출을 위한 핵산 크기 범위의 용도(USING NUCLEIC ACID SIZE RANGE FOR NONINVASIVE PRENATAL TESTING AND CANCER DETECTION)"인 미국 가출원 62/580,906으로부터 우선권을 주장하며, 이의 전체 내용은 모든 목적을 위해 참조에 의해 본 명세서에 포함된다.

배 경 기 술

[0003] 임산부의 혈액 혈장 및 혈청에서 태아로부터 기원하는 순환형 세포-무함유 DNA(cfDNA; circulating cell-free DNA)의 존재의 실증(문헌[Lo *et al.*, Lancet 1997; 350:485-487])은 비침습적 산전 검사(NIPT; noninvasive prenatal testing)의 개발을 통해 산전 검사의 실시를 완전히 변형시켰다. NIPT는 예컨대 양수천자(amniocentesis) 및 융모막 융모 시료화(CVS; chorionic villus sampling)를 통한 침습적 조직 시료화와 연관된 위험을 피하는 데 있어서 이점을 가진다. 따라서 현재까지, NIPT는 태아 RhD 혈액 그룹 유전형 분석(genotyping)(문헌[Finning *et al. BMJ* 2008; 336:816-818; Lo *et al. N Engl J Med* 1998; 339:1734-1738]), 성-연관 장애에 대한 태아 성 결정(문헌[Costa *et al. N. Engl. J. Med.* 2002; 346:1502]), 염색체 이수성(chromosomal aneuploidy) 검출(문헌[Chiu *et al. Proc Natl Acad Sci U S A* 2008; 105:20458-20463; Fan *et al. Nature* 2012; 487:320-324; Chiu *et al. BMJ* 2011; 342:c7401; Bianchi *et al. N. Engl. J. Med.* 2014; 370:799-808; Yu *et al. Proc. Natl. Acad. Sci. U. S. A.* 2014; 111:8583-8; Norton *et al. N. Engl. J. Med.* 2015; 372:1589-1597]) 및 단일유전인자 장애(monogenic disorder)의 진단(문헌[Lam *et al. Clin. Chem.* 2012; 58:1467-75; Lo *et al. Sci. Transl. Med.* 2010; 2:61ra91-61ra91; Ma *et al. Gene* 2014; 544:252-258; New *et al. J. Clin. Endocrinol. Metab.* 2014; 99:E1022-E1030])에 사용되어 왔다. 특히, 모체(maternal) 혈장 DNA의 대규모 병렬(massively parallel) 시퀀싱을 사용하여, 보편적인 염색체 이수성에 대한 NIPT는 많은 국가들에서 임상 서비스를 위해 빠르게 채택되어 왔고, 매년 수백만명의 임산부에 의해 사용된다(문헌[Allyse *et al. Int. J. Womens. Health* 2015; 7:113-26; Chandrasekharan *et al. Sci Transl Med* 2014; 6:231fs15]).

[0004] 초기 확증 연구(문헌[Chiu *et al. BMJ* 2011; 342:c7401; Sparks *et al. Am. J. Obstet. Gynecol.* 2012; 206:319.e1-9])에서, NIPT는 이수성에 대해 높은-위험도에 있는 환자에서 수행되었고, 높은 양성예측치(PPV; positive predictive value)는 92%로부터 100%까지 달성되었다. 특정 모체 시료에서 태아 DNA의 상대 농도는 태아 DNA 분율(DNA fraction)로도 보편적으로 지칭되며, NIPT의 정확도의 중요한 결정인자이다(문헌[Chiu *et al. BMJ* 2011; 342:c7401; Jiang *et al. Bioinformatics* 2012; 28:2883-2890, *npj Genomic Med.* 2016; 1:16013]). 21번 삼염색체성(trisomy 21) 검출의 민감도(sensitivity)는 태아 DNA 분율의 감소와 함께 유의하게 저하될 것이다(문헌[Chiu *et al. BMJ* 2011; 342:c7401; Canick *et al. Prenat. Diagn.* 2013; 33:667-674]). 그러므로, 삼염색체성 검출에 대한 위음성 결과는 낮은 태아 DNA 분율을 갖는 임산부에서 발생할 것이다. 예를 들어, Canick 등은 다운 증후군을 갖는 212 사례 중에서, 위음성이 4 사례 있었고, 이들은 모두 4% 내지 7%의 태아 DNA 분율을 가졌음을 보고하였다(문헌[Canick *et al. Prenat. Diagn.* 2013; 33:667-674]).

[0005] NIPT를 수행하는 많은 실험실에서, 검사 실패 또는 노-콜(no-call) 결과는 분석에 비례하여 관찰될 것임을 주지하는 것이 중요하다. 일부 연구에서, 총 실험실 실패율은 8.8%만큼 높을 수 있을 것이다(문헌[Porreco *et al. Am. J. Obstet. Gynecol.* 2014; 211:365.e1-365.e12]). NIPT에서 결과를 수득하는 데 있어서 실패의 주요 원인들 중 하나는 일부 시료에서, 통상적으로 4% 미만인, 모체 혈장 DNA의 낮은 태아 DNA 분율이다(문헌[Gil *et al. Fetal Diagn. Ther.* 2014; 35:156-73]). 4% 미만의 태아 DNA 분율을 갖는 환자에서, 이수성의 유병률(prevalence)은 4.7%인 것으로 보고되었고, 이는 전체 코호트에서 0.4%의 유병률과 비교하여 유의하게 더 높은 것으로 실증되었다(문헌[Norton *et al. N. Engl. J. Med.* 2015; 372:1589-1597]). 따라서, 이러한 검사 실패는 궁극적으로, NIPT의 전체 성능에 유해한 영향을 줄 수 있다. 예를 들어, 더 높은 검사 실패율은 더 낮은 실제 PPV를 야기할 것으로 예시되었다(문헌[Yaron *Prenat. Diagn.* 2016; 36:391-6]). 이론적 추정(문헌[Yaron *Prenat. Diagn.* 2016; 36:391-6])에서, 실험실에서 0.1%의 실패율은 67%의 실제 PPV를 제공할 것이지만, 이수성의 높은 위험도와 연관된 것으로 보고된 검사 실패를 갖는 모든 이들 환자는, 태아가 미국 산부인과 학회

(ACOG; American Congress of Obstetricians and Gynecologists) 권고사항으로부터의 권고에 따라 사실상 이수성인지 확인하기 위해 침습적 검사를 받을 것임을 가정하면, 1%의 실패율은 16.7%의 실제 PPV를 야기할 것이다(문헌[Yaron *Prenat. Diagn.* 2016; 36:391-6]).

[0006] 대략 2%의 임신이 4% 미만의 태아 DNA 분율을 갖는 것으로 나타났다(문헌[Wang *et al. Prenat. Diagn.* 2013; 33:662-666]). 낮은 태아 DNA 분율을 보여주는 제1 혈액 시료를 갖는 환자에 대해 재채혈된 혈액은, 10주 내지 21주에서 태아 DNA의 증가가 매우 미묘하기 때문에(태아 DNA 분율에서 평균적으로 매주 대략 0.1% 증가함) 충분한 태아 DNA 분율을 보증할 것 같지 않다(문헌[Wang *et al. Prenat. Diagn.* 2013; 33:662-666]). 또한, 이러한 낮은 태아 DNA 분율은 높은 모체 체중을 갖는 여성에서 선호적으로 발생한다. 일부 연구에서, 4% 미만의 태아 DNA 분율로 인한 결과 보고의 실패율은 5.9%만큼 높을 수 있다(문헌[Hall *et al. PLoS One* 2014; 9:e96677]).

[0007] 따라서, 모체 혈장에서 낮은 태아 DNA 분율(예를 들어, 4% 미만)을 갖는 임산부에 대해 NIPT의 성능을 향상시키기 위한 접근법을 개발하는 것이 유용할 것이다. 이러한 향상은 보편적인 염색체 이수성(예를 들어, 21번 삼염색체성, 18번 삼염색체성, 13번 삼염색체성 및 성염색체 이수성), 뿐만 아니라 하위-염색체(sub-chromosomal) 이상(aberration)(예를 들어, 미세결실 및 미세중복(microduplication))에 대한 NIPT의 성능에 중요할 것이다. 또한, 복제수(copy number) 이상 및 암에 대한 검사의 정확도 및 효율을 향상시키는 것은 유사한 접근법으로 해결될 수 있다. 이들 필요성 및 다른 필요성은 하기에서 해결된다.

발명의 내용

[0008] 크기-밴드 분석은 염색체 영역이 복제수 이상을 나타내는지를 결정하는 데 사용되거나 암을 검출하는 데 사용된다. 특이적인 크기에 초점을 맞추는 대신 다수의 크기 범위가 분석될 수 있다. 특이적인 크기 대신 다수의 크기 범위를 사용함으로써, 방법은 임상적-관련 DNA가 낮은 분율의 생물학적 시료일 수 있는 경우에도 염색체 영역이 복제수 이상을 나타내는지를 결정할 수 있다. 다수의 범위를 사용하는 것은, 게놈 영역에서 선택된 하위세트의 판독보다는 게놈 영역으로부터의 모든 서열 판독을 사용할 수 있게 한다. 분석의 정확도는 유사한 또는 더 높은 특이도에서 더 높은 민감도와 함께 증가될 수 있다. 분석은 동일한 정확도를 달성하기 위해 더 적은 수의 시퀀싱 판독을 포함하여, 더 효율적인 과정을 초래할 수 있다. 분석이 더 낮은 분율의 임상적-관련 DNA로 수행될 수 있기 때문에, 분석은 보다 초기 단계의 임신 또는 암에서 수행될 수 있다.

[0009] 하기 상세한 설명 및 첨부된 도면을 참고하여 본 발명의 실시예의 속성 및 이점에 대해 더 잘 이해할 수 있다.

도면의 간단한 설명

[0010] 도 1은 본 발명의 구현예에 따른 혈장 DNA 크기-밴드 분석의 원리의 개략적인 예시를 보여준다.

도 2a는 본 발명의 구현예에 따른 혈장 DNA 단편의 크기에 대한 이수성 염색체에 대한 측정된 태아 DNA 분율을 보여준다.

도 2b는 본 발명의 구현예에 따른 정배수성(euploidy) 및 21번 삼염색체성 태아로부터의 DNA를 포함한 시료에 대한 크기 밴드에 대한 z-점수를 보여준다.

도 3은 본 발명의 구현예에 따라 4%의 태아 DNA 분율을 갖는 상이한 개별 임신에 걸쳐 이수성 염색체에 대한 측정된 게놈 표현(genomic representation)의 크기-밴드 기초의 변화 패턴을 보여준다.

도 4a는 본 발명의 구현예에 따라 정배수성 및 21번 삼염색체성 태아를 갖는 임신 사이에서 크기-밴드 기초의 변화 패턴의 히트맵 플롯을 보여준다.

도 4b는 본 발명의 구현예에 따라 정배수성 및 21번 삼염색체성 태아를 갖는 임신 사이에서 크기-밴드 기초의 변화 패턴의 t-SNE(t-분포 확률적 임베딩; t-distributed stochastic neighbor embedding) 플롯을 보여준다.

도 4c는 본 발명의 구현예에 따라 정배수성 및 21번 삼염색체성 태아를 갖는 임신 사이에서 종래의 z-점수 접근법을 사용한 z-점수 분포를 보여준다.

도 5a 및 도 5b는 본 발명의 구현예에 따라 상이한 크기 밴드 중에서 z-점수 패턴을 학습함으로써 신경망 기초 모델에 대한 성능 평가를 보여준다.

도 6은 염색체 영역이 본 발명의 구현예에 따라 대상체로부터의 생물학적 시료에서 복제수 이상을 나타내는지를 결정하는 방법을 보여준다.

도 7은 본 발명의 구현예에 따라 간세포암종(HCC) 환자의 혈장 DNA에서 측정된 메틸화의 크기-밴드 기초의 변화 패턴을 보여준다.

도 8은 본 발명의 구현예에 따라 대상체로부터 생물학적 시료에서 암 분류를 결정하는 방법을 보여준다.

도 9는 본 발명의 구현예에 따라 간세포암종(HCC) 환자의 혈장 DNA에서 측정된 복제수 이상의 크기-밴드 기초의 변화 패턴을 보여준다.

도 10은 본 발명의 구현예에 따라 암 검출을 위한 크기-밴드 게놈 표현(GR) 접근법에 대한 작업흐름(workflow)을 예시한다.

도 11a, 도 11b 및 도 11c는 본 발명의 구현예에 따라 크기-밴드 GR 접근법과 종래의 z-점수 접근법 사이의 비교를 보여준다.

도 12는 본 발명의 구현예에 따른 암 분류를 결정하는 방법을 보여준다.

도 13은 본 발명의 구현예에 따른 암 검출을 위한 크기-밴드 메틸화 밀도(MD) 접근법에 대한 작업흐름을 예시한다.

도 14a, 14b 및 14c는 본 발명의 구현예에 따라 크기-밴드 MD 접근법과 종래의 z-점수 접근법 사이의 비교를 보여준다.

도 15는 본 발명의 구현예에 따른 시스템을 예시한다.

도 16은 본 발명의 구현예에 따른 컴퓨터 시스템을 보여준다.

용어

용어 "시료", "생물학적 시료" 또는 "환자 시료"는 살아 있는 또는 죽은 대상체로부터 유래되는 임의의 조직 또는 물질을 포함하는 것으로 의미된다. 생물학적 시료는 세포-무함유 시료일 수 있으며, 이는 대상체로부터의 핵산 분자와 잠재적으로는 병원체, 예를 들어, 바이러스로부터의 핵산 분자의 혼합물을 포함할 수 있다. 생물학적 시료는 일반적으로, 핵산(예를 들어, DNA 또는 RNA) 또는 이의 단편을 포함한다. 용어 "핵산"은 일반적으로, 데옥시리보핵산(DNA), 리보핵산(RNA) 또는 이들의 임의의 하이브리드 또는 단편을 지칭한다. 시료 내의 핵산은 세포-무함유 핵산일 수 있다. 시료는 액체 시료 또는 고체 시료(예를 들어, 세포 또는 조직 시료)일 수 있다. 생물학적 시료는 체액, 예컨대 혈액, 혈장, 혈청, 소변, 질액, 수류(예를 들어, 고환의)로부터의 유체, 질 플러싱 유체(vaginal flushing fluid), 흉수(pleural fluid), 복수(ascitic fluid), 뇌척수액, 침, 땀, 눈물, 가래, 기관지폐포 세척액, 유두로부터의 배출액, 신체(예를 들어, 갑상선, 유방)의 상이한 부분들로부터의 흡인액 등일 수 있다. 대변 시료도 또한 사용될 수 있다. 다양한 구현예에서, 세포-무함유 DNA에 대해 농화되었던 생물학적 시료(예를 들어, 원심분리 프로토콜을 통해 수득된 혈장 시료) 내의 대부분의 DNA는 세포-무함유일 수 있다(예를 들어, 50%, 60%, 70%, 80%, 90%, 95% 또는 99% 초과의 DNA가 세포-무함유일 수 있음). 원심분리 프로토콜은 예를 들어, 3,000 g x 10분, 유체 부분을 수득하고 잔여 세포를 제거하기 위해 예를 들어, 30,000 g에서 또 다른 10분 동안 재-원심분리할 수 있다.

본원에 사용된 바와 같이, 용어 "*좌위*(*locus*)" 또는 이의 복수형 "*좌위들*"은 게놈에 걸쳐 변동을 갖는 임의의 길이의 뉴클레오타이드(또는 염기쌍)의 장소 또는 주소(address)이다. 용어 "*서열 판독*"은 핵산 분자의 모두 또는 일부, 예를 들어, DNA 단편으로부터 수득되는 서열을 지칭한다. 일 구현예에서, 단편의 단지 하나의 말단이 시퀀싱된다. 대안적으로, 단편의 두 말단 모두(예를 들어, 각각의 말단으로부터 약 30 bp)가 시퀀싱되어, 2개의 서열 판독을 발생시킬 수 있다. 그 후에, 짝형성된(paired) 서열 판독은 기준 게놈에 정렬될 수 있으며, 이는 단편의 길이를 제공할 수 있다. 보다 다른 구현예에서, 선형 DNA 단편은 예를 들어, 결찰에 의해 고리화될 수 있고, 결찰 부위를 포괄하는(spanning) 부분(part)이 시퀀싱될 수 있다.

본원에 사용된 바와 같이, 용어 "단편"(예를 들어, DNA 단편)은 적어도 3개의 연속 뉴클레오타이드를 포함하는 폴리뉴클레오타이드 또는 폴리펩타이드 서열의 일부를 지칭할 수 있다. 핵산 단편은 부모 폴리펩타이드의 생물학적 활성 및/또는 일부 특징을 보유할 수 있다. 핵산 단편은 이중 가닥 또는 단일 가닥이거나, 메틸화 또는 비메틸화되거나, 온전하거나 또는 닉킹되거나(nicked), 다른 거대분자, 예를 들어, 지질 입자, 단백질과 복합체화될 수 있거나 복합체화되지 않을 수 있다. 종양-유래 핵산은 종양 세포 내의 병원체로부터의 병원체 핵산을 포함하여 종양 세포로부터 방출되는 임의의 핵산을 지칭할 수 있다.

용어 "검정법"은 일반적으로, 핵산의 특성을 결정하는 기술을 지칭한다. 검정법(예를 들어, 제1 검정법 또는 제

2 검정법)은 일반적으로, 시료 내 핵산의 양, 시료 내 핵산의 유전적 동일성, 시료 내 핵산의 복제수 변동(복제수 변동), 시료 내 핵산의 메틸화 상태, 시료 내 핵산의 단편 크기 분포, 시료 내 핵산의 돌연변이 상태, 또는 시료 내 핵산의 단편화 패턴을 결정하는 기술을 지칭한다. 당업자에게 알려진 임의의 검정법이 본원에서 언급된 핵산의 임의의 특성을 검출하는 데 사용될 수 있다. 핵산의 특성은 핵산의 서열, 양, 유전적 동일성, 복제수, 하나 이상의 뉴클레오타이드 위치에서의 메틸화 상태, 크기, 하나 이상의 뉴클레오타이드 위치에서 핵산 내 돌연변이, 및 핵산의 단편화 패턴(예를 들어, 핵산이 단편화하는 뉴클레오타이드 위치(들))을 포함한다. 용어 "검정법"은 용어 "방법"과 상호교환적으로 사용될 수 있다. 검정법 또는 방법은 특정 민감도 및/또는 특이도를 가질 수 있고, 진단 툴로서의 이들의 상대적인 유용성은 ROC-AUC 통계를 사용하여 측정될 수 있다.

본원에 사용된 바와 같이, 용어 "무작위 시퀀싱"은 일반적으로, 시퀀싱된 핵산 단편이 시퀀싱 절차 전에 구체적으로 식별되거나 또는 사전-결정되지 않은 시퀀싱을 지칭한다. 특이적인 유전자 좌위를 표적화하는 서열-특이적인 프라이머가 요구되지 않는다. 일부 구현예에서, 어댑터(adapter)는 단편의 말단에 첨가되고, 시퀀싱을 위한 프라이머가 상기 어댑터에 부착된다. 따라서, 임의의 단편은 동일한 유니버셜 어댑터에 부착하는 동일한 프라이머를 이용하여 시퀀싱될 수 있고, 따라서 시퀀싱은 무작위일 수 있다. 대규모 병렬 시퀀싱은 무작위 시퀀싱을 사용하여 수행될 수 있다.

"*핵산*"은 데옥시리보뉴클레오타이드 또는 리보뉴클레오타이드 및 단일 가닥 또는 이중 가닥 형태의 이들의 중합체를 지칭할 수 있다. 용어는 합성, 천연 발생 및 비-천연 발생이며 기준 핵산과 유사한 결합 특성을 갖고 기준 뉴클레오타이드와 유사한 방식으로 대사되는 기지의(known) 뉴클레오타이드 유사체 또는 변형된 백본 잔기 또는 연결(linkage)을 함유하는 핵산을 포괄할 수 있다. 이러한 유사체의 예는 비제한적으로, 포스포로티오에이트, 포스포라미다이트, 메틸 포스포네이트, 카이랄-메틸 포스포네이트, 2-O-메틸 리보뉴클레오타이드, 펩타이드-핵산(PNA)을 포함할 수 있다.

다르게 나타내지 않는 한, 특정 핵산 서열은 또한, 이의 보존적으로 변형된 변이체(예를 들어, 축퇴성(degenerate) 코돈 치환) 및 상보적 서열, 뿐만 아니라 명쾌하게 나타낸 서열을 내재적으로 포괄한다. 구체적으로, 축퇴성 코돈 치환은, 하나 이상의 선택된(또는 모든) 코돈의 제3 위치가 혼합-염기 및/또는 데옥시이노신 잔기로 치환된 서열을 발생시킴으로써 달성될 수 있다(문헌[Batzer *et al.*, *Nucleic Acid Res.* 19:5081 (1991); Ohtsuka *et al.*, *J. Biol. Chem.* 260:2605-2608 (1985); Rossolini *et al.*, *Mol. Cell. Probes* 8:91-98 (1994)]). 용어 핵산은 유전자, cDNA, mRNA, 올리고뉴클레오타이드 및 폴리뉴클레오타이드와 상호교환적으로 사용된다.

용어 "*뉴클레오타이드*"는 천연 발생 리보뉴클레오타이드 또는 데옥시리보뉴클레오타이드 단량체를 지칭하는 것 외에도, 문맥상 분명하게 다르게 나타내지 않는 한, 뉴클레오타이드가 사용되는 특정 맥락(예를 들어, 상보적 염기에의 혼성화)에 관하여 기능적으로 동등한 이의 유도체 및 유사체를 포함하여 관련된 구조적 변이체를 지칭하는 것으로 이해될 수 있다.

"*서열 판독*"은 핵산 분자 중 임의의 부분 또는 모두로부터 시퀀싱된 뉴클레오타이드 열(string)을 지칭한다. 예를 들어, 서열 판독은 생물학적 시료에 존재하는 전체 핵산 단편일 수 있다. 또한 예로서, 서열 판독은 생물학적 시료에 존재하는 핵산 단편으로부터 시퀀싱된 뉴클레오타이드(예를 들어, 20 내지 150개 염기)의 짧은 열, 핵산 단편 중 하나의 말단 또는 두 말단 모두에서 뉴클레오타이드의 짧은 열, 또는 전체 핵산 단편의 시퀀싱일 수 있다. 서열 판독은 여러 가지 방식으로, 예를 들어, 혼성화 프로브 또는 포착 프로브에서 시퀀싱 기술을 사용하거나 프로브를 사용하여, 또는 증폭 기술, 예컨대 중합효소 연쇄 반응(PCR) 또는 단일 프라이머를 사용하는 선형 증폭 또는 등온 증폭에서, 또는 생물물리학적 측정, 예컨대 질량 분광분석법을 기초로 하여 수득될 수 있다. 서열 판독은 단일-분자 시퀀싱으로부터 수득될 수 있다. "*단일-분자 시퀀싱*"은 주형 DNA 분자의 클론 복사체로부터 염기 서열 정보를 해석할 필요 없이, 단일 주형 DNA 분자를 시퀀싱하여 서열 판독을 수득하는 것을 지칭한다. 단일-분자 시퀀싱은 전체 분자 또는 DNA 분자의 단지 일부를 시퀀싱할 수 있다. 대부분의 DNA 분자, 예를 들어, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95% 또는 99% 초과가 시퀀싱될 수 있다.

용어 "*유니버셜 시퀀싱*"은, 어댑터가 단편의 말단에 첨가되고, 시퀀싱용 프라이머가 상기 어댑터에 부착되는 시퀀싱을 지칭한다. 따라서, 임의의 단편이 동일한 프라이머로 시퀀싱될 수 있으므로, 시퀀싱은 무작위일 수 있다.

"임상적-관련(clinically-relevant)" DNA의 예는 모체 혈장 내 태아 DNA 및 환자의 혈장 내 종양 DNA를 포함한다. 또 다른 예는 이식 환자의 혈장 내 이식-연관 DNA의 양의 측정을 포함한다. 추가의 예는 대상체의 혈장 내 조혈모 DNA 및 비-조혈모 DNA의 상대량의 측정을 포함한다. 이러한 후자의 구현예는 조혈모 및 비-조혈모 조직

을 수반하는 병리학적 과정 또는 손상을 검출하거나 모니터링하거나 예후화하는 데 사용될 수 있다.

용어 "*암의 수준*"(또는 보다 일반적으로 "*질병의 수준*" 또는 "*질환의 수준*")은, 암이 존재하는지의 여부(즉, 존재 또는 부재), 암의 병기, 종양의 크기, 전이가 존재하는지의 여부, 신체의 총 종양 부담, 치료에 대한 암의 반응, 및/또는 암의 중증도의 다른 측정치(예를 들어, 암의 재발)를 지칭할 수 있다. 암의 수준은 수치(예를 들어, 확률) 또는 다른 지표, 예컨대 부호, 알파벳 글자 및 색상일 수 있다. 수준은 제로(0)일 수 있다. 암의 수준은 또한, 전악성(premalignant) 또는 전암성(precancerous) 질환(상태)을 포함한다. 암의 수준은 다양한 방식으로 사용될 수 있다. 예를 들어, 스크리닝은, 암을 갖고 있는 것으로 이전에는 알려지지 않은 개체에 암이 존재하는지 체크할 수 있다. 평가는 암을 진단받은 개체를 조사하여, 시간 경과에 따른 암의 진전을 모니터링하거나, 치료법의 효능을 연구하거나, 예후를 결정할 수 있다. 일 구현예에서, 예후는 환자가 암으로 사망할 가능성, 특정한 기간 또는 시간 후에 암이 진전되는 가능성, 또는 암이 전이될 가능성으로서 표현될 수 있다. 검출은 '스크리닝'을 의미할 수 있거나, 암의 제안적인 특색(예를 들어, 증상 또는 다른 양성 시험)을 갖는 개체가 암을 갖고 있는지 체크하는 것을 의미할 수 있다. "병상(pathology)의 수준"은 병원체와 연관된 병상의 수준을 지칭할 수 있으며, 이때 상기 수준은 암에 대해 상기 기재된 바와 같을 수 있다. 질병/질환의 수준 또한, 암에 대해 상기 기재된 바와 같을 수 있다. 암이 병원체와 연관되어 있는 경우, 암의 수준은 병상의 수준의 유형일 수 있다.

본원에 사용된 바와 같이, 용어 "*염색체 이수성*"은 2배체 게놈으로부터 염색체의 정량적 양의 변동을 의미한다. 상기 변동은 획득(gain) 또는 소실(loss)일 수 있다. 상기 변동은 하나의 염색체의 전체 또는 염색체의 영역을 수반할 수 있다.

본원에 사용된 바와 같이, 용어 "*서열 불균형*" 또는 "*이상*"은 기준 양으로부터 임상적으로 관련된 염색체 영역의 양에서 적어도 하나의 컷오프(cutoff) 값에 의해 정의되는 바와 같은 임의의 유의한 편차를 의미한다. 서열 불균형은 염색체 투여량 불균형, 대립유전자 불균형, 돌연변이 투약량 불균형, 복제수 불균형, 반수체 투약량 불균형, 및 다른 유사한 불균형을 포함할 수 있다. 일례로, 대립유전자 불균형은, 종양이 결실된 유전자의 하나의 대립유전자 또는 증폭된 유전자의 하나의 대립유전자, 또는 이의 게놈에서 2개 대립유전자의 차별적인 증폭을 가짐으로써 시료 내 특정 좌위에서 불균형을 생성할 때 발생할 수 있다. 또 다른 예로서, 환자는 종양 억제자 유전자에서 유전받은(inherited) 돌연변이를 가질 수 있을 것이다. 그 후에, 상기 환자는 종양 억제자 유전자의 비-돌연변이화된 대립유전자가 결실되는 종양을 발병시키게 될 수 있을 것이다. 따라서, 종양 내에서, 돌연변이 투여량 불균형이 존재한다. 종양이 이의 DNA를 환자의 혈장 내로 방출할 때, 종양 DNA는 혈장 내 환자의 (정상 세포로부터의) 구성적(constitutional) DNA와 혼합될 것이다. 본원에 기재된 방법의 사용을 통해, 혈장에서 이러한 DNA 혼합물의 돌연변이 투여량 불균형이 검출될 수 있다. 이상은 염색체 영역의 결실 또는 증폭을 포함할 수 있다.

포유류 게놈에서 "*DNA 메틸화*"는 전형적으로, CpG 디뉴클레오타이드 중에서 시토신 잔기의 5' 탄소에 메틸기의 첨가(즉, 5-메틸시토신)를 지칭한다. DNA 메틸화는 다른 맥락, 예를 들어, CHG 및 CHH의 시토신에서 발생할 수 있으며, 이때 H는 아데닌, 시토신 또는 티민이다. 시토신 메틸화는 또한, 5-하이드록시메틸시토신의 형태일 수 있다. 비-시토신 메틸화, 예컨대 N6-메틸아데닌이 또한, 보고되었다.

"*분류*"는 시료의 특정한 특성과 연관된 임의의 수치(들) 또는 다른 특징(들)을 지칭한다. 예를 들어, "+" 부호(또는 단어 "양성")는, 시료가 결실 또는 증폭을 갖고 있는 것으로 분류됨을 의미할 수 있을 것이다. 분류는 2진(binary)(예를 들어, 양성 또는 음성)일 수 있거나, 더 많은 수준의 분류(예를 들어, 1 내지 10, 또는 0 내지 1의 규모)를 가질 수 있다.

용어 "*컷오프*" 및 "*역치*"는 작동에 사용되는 사전-결정된 수치를 지칭할 수 있다. 역치 또는 기준 값은, 이 값의 초과 또는 미만에서 특정 분류, 예를 들어, 질환의 분류, 예컨대 대상체가 질환을 갖고 있는지의 여부 또는 질환의 중증도가 적용되는 값일 수 있다. 컷오프는 시료 또는 대상체의 특징을 참조하거나 참조하지 않으면서 사전-결정될 수 있다. 예를 들어, 컷오프는 시험되는 대상체의 연령 또는 성별에 기초하여 선택될 수 있다. 컷오프는 시험 데이터의 출력 후 그리고 이에 기초하여 선택될 수 있다. 예를 들어, 소정의 컷오프는, 시료의 시퀀싱이 소정의 깊이에 도달할 때 사용될 수 있다. 또 다른 예로서, 하나 이상의 질환의 기지의 분류 및 측정된 특징적인 값(예를 들어, 메틸화 수준, 통계학적 크기 값, 또는 카운트(count))을 갖는 기준 대상체는 상이한 질환들 및/또는 질환의 분류(예를 들어, 대상체가 질환을 갖고 있는지 여부)를 분간하기 위해 기준 수준을 결정하는 데 사용될 수 있다. 이들 용어 중 임의의 용어는 이들 맥락 중 임의의 맥락에서 사용될 수 있다. 당업자가 이해하는 바와 같이, 컷오프는 요망되는 민감도 및 특이도를 달성하도록 선택될 수 있다.

"*부위*"("*게놈 부위*"로도 지칭됨)는 단일 부위에 상응하며, 이는 단일 염기 위치 또는 상관(correlated) 염기 위치의 그룹, 예를 들어, 상관 염기 위치의 CpG 부위 또는 더 큰 그룹일 수 있다. "*좌위*"는 다수의 부위들을 포함하는 영역에 상응할 수 있다. 좌위는 단지 하나의 부위를 포함할 수 있으며, 이는 상기 좌위를 해당 맥락에서 부위에 동등하게 만들 것이다.

각각의 게놈 부위(예를 들어, CpG 부위)에 대한 "*메틸화 지수*"는 (예를 들어, 시퀀스 판독 또는 프로브로부터 결정된 바와 같은) DNA 단편의 비율을 지칭할 수 있으며, 부위를 망라하는 판독의 총 수에 걸쳐 해당 부위에서의 메틸화를 보여준다. "판독"은 DNA 단편으로부터 수득된 정보(예를 들어, 부위에서의 메틸화 상태)에 상응할 수 있다. 판독은 특정 메틸화 상태의 DNA 단편에 우선적으로 혼성화하는 시약(예를 들어, 프라이머 또는 프로브)을 사용하여 수득될 수 있다. 전형적으로, 이러한 시약은 DNA 분자의 메틸화 상태에 따라 이들 분자를 차별적으로 변형시키거나 차별적으로 인지하는 과정, 예를 들어, 비설파이트 전환, 또는 메틸화-민감성 제한 효소, 메틸화 결합 단백질, 또는 항-메틸시토신 항체로 처리한 후 적용된다. 또 다른 구현예에서, 메틸시토신 및 하이드록시메틸시토신을 인지하는 단일 분자 시퀀싱 기술은 메틸화 상태를 명시하고 메틸화 지수를 결정하는 데 사용될 수 있다.

영역의 "*메틸화 밀도*"는 영역 내의 부위를 망라하는 판독의 총 수로 나눈, 메틸화를 보여주는 영역 내의 부위에서의 판독의 수를 지칭할 수 있다. 상기 부위는 특이적인 특징, 예를 들어, CpG 부위라는 특징을 가질 수 있다. 따라서, 영역의 "CpG 메틸화 밀도"는 영역(예를 들어, 특정 CpG 부위, CpG 섬(island) 내의 CpG 부위, 또는 더 큰 영역) 내의 CpG 부위를 망라하는 판독의 총 수로 나눈, CpG 메틸화를 보여주는 판독의 수를 지칭할 수 있다. 예를 들어, 인간 게놈에서 각각의 100-kb 빈(bin)에 대한 메틸화 밀도는, 100-kb 영역으로 맵핑된(mapped) 시퀀스 판독에 의해 망라된 모든 CpG 부위의 비율로서 CpG 부위에서 비설파이트 처리(메틸화된 시토신에 상응함) 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 등에 대해서도 수행될 수 있다. 영역은 전체 게놈, 염색체 또는 염색체의 일부(예를 들어, 염색체 아암(arm))일 수 있을 것이다. CpG 부위의 메틸화 지수는, 영역이 해당 CpG 부위만 포함할 때, 상기 영역에 대한 메틸화 밀도와 동일하다. "메틸화된 시토신의 비율"은 영역 내의 분석된 시토신 잔기, 즉, CpG 맥락 외부의 시토신을 포함하여 이들의 총 수에 걸쳐, 메틸화된(예를 들어, 비설파이트 전환 후 전환되지 않는) 것으로 보이는 시토신 부위, "C"의 수를 지칭할 수 있다. 메틸화된 시토신의 메틸화 지수, 메틸화 밀도 및 비율은, "*메틸화 수준*"의 예이며, 이는 부위에서 메틸화된 판독의 카운트를 수반하는 다른 비를 포함할 수 있다. 비설파이트 전환 외에도, 비제한적으로 메틸화 상태에 민감한 효소(예를 들어, 메틸화-민감성 제한 효소), 메틸화 결합 단백질, 메틸화 상태에 민감한 플랫폼을 사용하는 단일 분자 시퀀싱(예를 들어, 나노포어 시퀀싱(문헌[Schreiber et al. Proc Natl Acad Sci 2013; 110: 18910-18915]) 및 Pacific Biosciences 단일 분자 실시간 분석(문헌[Flusberg et al. Nat Methods 2010; 7: 461-465]))을 포함하여 당업자에게 알려진 다른 과정이 DNA 분자의 메틸화 상태에 대한 정보를 얻는 데 사용될 수 있다.

"*메틸화-인식 시퀀싱*"은 비제한적으로 비설파이트 시퀀싱, 또는 메틸화-민감성 제한 효소 절단에 뒤이은 시퀀싱, 항-메틸시토신 항체 또는 메틸화 결합 단백질을 사용한 면역침전, 또는 메틸화 상태의 명시(elucidation)를 가능하게 하는 단일 분자 시퀀싱을 포함하여 당업자가 시퀀싱 과정 동안 DNA 분자의 메틸화 상태를 확인할 수 있게 하는 임의의 시퀀싱 방법을 지칭한다. "*메틸화-인식 검정법*" 또는 "*메틸화-민감성 검정법*"은 시퀀싱 기초 방법과 비-시퀀싱 기초 방법 둘 모두, 예컨대 MSP, 프로브 기초 조사(interrogation), 혼성화, 제한 효소 절단 및 뒤이은 밀도 측정, 항-메틸시토신 면역검정법, 메틸화된 시토신 또는 하이드록시메틸시토신의 비율의 질량 분광분석법 조사, 시퀀싱이 후속하지 않는 면역침전 등을 포함할 수 있다.

"분리 값"(또는, 상대 존재비(relative abundance))은 2개의 값, 예를 들어, DNA 분자의 2개 양, 2개의 분획 기여도(fractional contribution), 또는 2개의 메틸화 수준, 예컨대 시료(혼합물) 메틸화 수준 및 기준 메틸화 수준을 수반하는 차이 또는 비에 상응한다. 분리 값은 단순한 차이 또는 비일 수 있을 것이다. 예로서, x/y의 정비(direct ratio)는 분리 값, 뿐만 아니라 x/(x+y)이다. 분리 값은 다른 인자, 예를 들어, 곱셈 인자(multiplicative factor)를 포함할 수 있다. 다른 예로서, 값들의 함수의 차이 또는 비, 예를 들어, 2개 값의 자연 로그(ln)의 차이 또는 비가 사용될 수 있다. 분리 값은 차이 및/또는 비를 포함할 수 있다. 메틸화 수준은 예를 들어, (예를 들어, 특정 부위에서) 메틸화된 DNA 분자와 다른 DNA 분자(예를 들어, 특정 부위에서의 모든 다른 DNA 분자 또는 단지 비메틸화된 DNA 분자) 사이의 상대 존재비의 일례이다. 다른 DNA 분자의 양은 정규화 인자(normalization factor)로서 작용할 수 있다. 또 다른 예로서, 모든 또는 비메틸화된 DNA 분자의 강도에 대한 메틸화된 DNA 분자의 강도(예를 들어, 형광 또는 전기적 강도)가 결정될 수 있다. 상대 존재비는 또한, 1 부피 당 강도를 포함할 수 있다.

용어 "대조군", "대조군 시료", "기준", "기준 시료", "정상" 및 "정상 시료"는 일반적으로 특정 질환을 갖고 있지 않는, 또는 그렇지 않다면 건강한 시료를 설명하기 위해 상호교환적으로 사용될 수 있다. 일례에서, 본원에 개시된 바와 같은 방법은 종양을 갖고 있는 대상체 상에서 수행될 수 있으며, 이때, 기준 시료는 대상체의 건강한 조직으로부터 가져온 시료이다. 또 다른 예에서, 기준 시료는 질병, 예를 들어, 암 또는 특정 병기(stage)의 암을 갖는 대상체로부터 가져온 시료이다. 기준 시료는 대상체로부터, 또는 데이터베이스로부터 수득될 수 있다. 기준은 일반적으로, 대상체로부터 시료를 시퀀싱함으로써 수득되는 시퀀스 판독을 맵핑하는 데 사용되는 기준 게놈을 지칭한다. 기준 게놈은 일반적으로, 생물학적 시료 및 구성적(constitutional) 시료로부터의 시퀀스 판독이 정렬되고 비교될 수 있는 반수체 또는 2배체 게놈을 지칭한다. 반수체 게놈의 경우, 각각의 좌위에 오직 1개의 뉴클레오타이드가 존재한다. 2배체 게놈의 경우, 이형접합체성(heterozygous) 좌위가 식별될 수 있으며, 이때, 이러한 좌위는 2개의 대립유전자를 가지며, 여기서 대립유전자는 상기 좌위로의 정렬을 위한 매치를 가능하게 할 수 있다. 기준 게놈은 예를 들어, 하나 이상의 바이러스 게놈을 포함함으로써 바이러스에 상응할 수 있다.

본원에 사용된 바와 같이 어구 "건강한"은 일반적으로, 양호한 건강을 소유한 대상체를 지칭한다. 이러한 대상체는 임의의 악성 또는 비-악성 질병의 부재를 실증한다. "건강한 개체"는 검정되는 질환과 관련이 없는 다른 질병 또는 질환을 갖고 있을 수 있으며, 통상적으로 "건강한" 것으로 여겨지지 않을 수 있다.

용어 "암" 또는 "종양"은 상호교환적으로 사용될 수 있고, 일반적으로, 조직의 비정상적인 덩어리(mass)를 지칭하며, 상기 덩어리의 성장은 정상 조직의 성장을 능가하고 이와 조화되지 않는다. 암 또는 종양은 하기 특징에 따라 "양성" 또는 "악성"으로서 정의될 수 있다: 형태 및 기능성을 포함하여 세포 분화의 정도, 성장 속도, 국소 침습, 및 전이. "양성" 종양은 일반적으로 잘 분화되어 있으며, 악성 종양보다 특징적으로 더 느린 성장을 갖고, 기원 부위로 국소화된 채로 남아 있다. 또한, 양성 종양은 원위부로 침윤하거나, 침습하거나 전이하는 능력을 갖지 않는다. "악성" 종양은 일반적으로 불량하게 분화되어 있으며(퇴화(anaplasia)), 주변 조직의 점진적인 침윤, 침습 및 파괴가 동반되는 특징적으로 신속한 성장을 가진다. 더욱이, 악성 종양은 원위부로 전이하는 능력을 가진다. "병기"는, 악성 종양이 얼마나 진행되어 있는지 설명하는 데 사용될 수 있다. 초기 암 또는 악성물은 말기 악성물보다 신체에서 더 적은 종양 부담(burden), 일반적으로 더 적은 증상, 더 양호한 예후, 및 더 양호한 치료 결과와 연관이 있다. 말기 또는 진행 병기(advanced stage) 암 또는 악성물은 종종, 원위부 전이 및/또는 림프 확산(lymphatic spread)과 연관이 있다.

용어 "위양성"(FP)은 질환을 갖지 않는 대상체를 지칭할 수 있다. 위양성은 일반적으로, 종양, 암, 전암성 질환(예를 들어, 전암성 병변), 국소화된 또는 전이된 암, 비-악성 질병을 갖지 않거나 그렇지 않다면 건강한 대상체를 지칭한다. 용어 위양성은 일반적으로, 질환을 갖지 않지만 본 개시내용의 검정법 또는 방법에 의해 상기 질환을 갖는 것으로 식별되는 대상체를 지칭한다.

용어 "민감도" 또는 "진양성률"(TPR; true positive rate)은 진양성 및 위음성의 수의 합계로 나눈, 진양성의 수를 지칭할 수 있다. 민감도는 실제로 질환을 갖고 있는 집단의 비율을 올바르게 식별하는 검정법 또는 방법의 능력을 특징화할 수 있다. 예를 들어, 민감도는 암을 갖고 있는 집단 내의 대상체의 수를 올바르게 식별하는 방법의 능력을 특징화할 수 있다. 또 다른 예에서, 민감도는 암을 시사하는 하나 이상의 마커를 올바르게 식별하는 방법의 능력을 특징화할 수 있다.

용어 "특이도" 또는 "진음성률"(TNR; true negative rate)은 진음성 및 위양성의 수의 합계로 나눈 진음성의 수를 지칭할 수 있다. 특이도는 실제로 질환을 갖고 있지 않는 집단의 비율을 올바르게 식별하는 검정법 또는 방법의 능력을 특징화할 수 있다. 예를 들어, 특이도는 암을 갖고 있지 않는 집단 내의 대상체의 수를 올바르게 식별하는 방법의 능력을 특징화할 수 있다. 또 다른 예에서, 특이도는 암을 시사하는 하나 이상의 마커를 올바르게 식별하는 방법의 능력을 특징화할 수 있다.

용어 "ROC" 또는 "ROC 곡선"은 수신자 조작자 특성 곡선(receiver operator characteristic curve)을 지칭할 수 있다. ROC 곡선은 2진 분류기 시스템의 성능의 그래프 표현일 수 있다. 임의의 주어진 방법에 대해, ROC 곡선은 다양한 역치 설정에서 민감도를 특이도에 대해 도시함으로써 발생될 수 있다. 대상체에서 종양의 존재를 검출하는 방법의 민감도 및 특이도는 상기 대상체의 혈장 시료 내 다양한 농도의 종양-유래 핵산에서 결정될 수 있다. 더욱이, 3개 매개변수(예를 들어, 민감도, 특이도, 및 역치 설정) 중 적어도 하나를 고려하면, ROC 곡선은 임의의 미지의 매개변수에 대한 값 또는 예상 값을 결정할 수 있다. 미지의 매개변수는 ROC 곡선에 피팅된(fitted) 곡선을 사용하여 결정될 수 있다. 용어 "AUC" 또는 "ROC-AUC"는 일반적으로, 수신자 조작자 특성 곡선 아래 면적을 지칭한다. 이러한 계측(metric)은 방법의 민감도와 특이도 둘 모두를 고려하여, 상기 방법의 진단

유용성의 측정치를 제공할 수 있다. 일반적으로, ROC-AUC는 0.5 내지 1.0의 범위이며, 이때, 0.5에 더 근접한 값은 상기 방법이 제한된 진단 유용성(예를 들어, 더 낮은 민감도 및/또는 특이도)을 가짐을 시사하고, 1.0에 더 근접한 값은 더 큰 진단 유용성(예를 들어, 더 높은 민감도 및/또는 특이도)을 가짐을 시사한다. 예를 들어, 문헌[Pepe et al, " Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker," Am. J. Epidemiol 2004, 159 (9): 882-890]을 참조하고, 이는 참조에 의해 본 명세서에 포함된다. 확률 함수, 교차비(odds ratio), 정보 이론, 예측 값, 보정(적합도(goodness-of-fit)를 포함함), 및 재분류 측정을 사용하여 진단 유용성을 특징화하는 또 다른 접근법은 문헌[Cook, "Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction," Circulation 2007, 115: 928-935]에 따라 요약되어 있으며, 이는 참조에 의해 본 명세서에 포함된다.

용어 "약" 또는 "대략"은 당업자에 의해 결정된 바와 같은 특정 값에 대한 허용 가능한 오차 범위 내를 의미할 수 있으며, 이는 부분적으로는 값이 어떻게 측정되거나 결정되는가, 즉, 측정 시스템의 한계에 의존할 수 있다. 예를 들어, "약"은 당업의 관행에 따라, 1 이내 또는 1 초과의 표준 편차를 의미할 수 있다. 대안적으로, "약"은 주어진 값의 20% 이하, 10% 이하, 5% 이하, 또는 1% 이하의 범위를 의미할 수 있다. 대안적으로, 특히 생물학적 시스템 또는 과정에 관하여, 용어 "약" 또는 "대략"은 값의 승수(order of magnitude) 이내, 5-배 이내, 보다 바람직하게는 2-배 이내를 의미할 수 있다. 특정 값이 출원 및 청구항에 기재되어 있는 경우, 다르게 언급되지 않는 한, 특정 값에 대한 허용 가능한 오차 범위 내를 의미하는 용어 "약"이 추정되어야 한다. 용어 "약"은 당업자에 의해 보편적으로 이해되는 바와 같은 의미를 가질 수 있다. 용어 "약"은 ±10%를 지칭할 수 있다. 용어 "약"은 ±5%를 지칭할 수 있다.

본원에서 사용되는 용어는 특정 사례만 설명하기 위한 것이고, 제한하려는 것이 아니다. 본원에 사용된 바와 같은, 단수형("a", "an" 및 "the")은 맥락상 다르게 명확하게 시사하지 않는 한, 복수형도 포함시키려는 것이다. "또는"의 사용은 "포함하거나 또는"을 의미하고, 구체적으로 다르게 시사되지 않는 한 "배제하거나 또는"을 의미하도록 의도되지 않는다. 용어 "~에 기초하는"은 "적어도 부분적으로 ~에 기초하는"을 의미하도록 의도된다. 더욱이, 용어 "포함하는", "포함한다", "갖는", "가진다", "있는" 또는 이의 변형이 상세한 설명 및/또는 청구범위에서 사용되는 정도까지, 이러한 용어는 용어 "포함하는"과 유사한 방식으로 포함하도록 의도된다.

발명을 실시하기 위한 구체적인 내용

[0011] 세포-무함유 DNA의 크기-기초 분석은 염색체 이수성 및 암에 대해 생물학적 시료를 분석하는 데 사용되어 왔다. 그러나, 이전의 크기-기초 기법을 이용하여, 생물학적 시료가 낮은 백분율의 임상적-관련 DNA를 갖는 경우, 통계학적으로 유의한 결과를 수득하는 것은 어려울 수 있다. 임상적-관련 DNA의 분율이 낮은 경우, 이전의 크기-기초 분석은 단일 분석 기법으로서 의존한 것보다 또 다른 유형의 분석의 결과를 확증하는 데 사용될 수 있다. 본 발명의 구현예는, 더 많은 세포-무함유 DNA가 분석에 사용되게 할 수 있고 크기 패턴이 분석되게 할 수 있는 크기 밴드를 사용하는 단계를 수반한다. 그 결과, 크기-기초 분석은 훨씬 낮은 분율의 임상적-관련 DNA에서 정확하게 수행될 수 있다.

[0012] 이 연구에서, 본 발명자들은 NIPT에 필요한 태아 DNA 분율의 한계를 낮추기 위해 세포-무함유 DNA의 크기 분석을 적용하는 것을 목표로 하였다. 본 발명자들은 특이도에 유해한 영향을 주지 않으면서 NIPT의 민감도를 향상시키는 것을 목표로 한다. 유사한 기법은 암 분석에 적용될 수 있다. 특이적인 크기 대신에 다수의 크기 범위를 사용하는 것은 임상적-관련 DNA의 분율이 낮은 경우에도 생물학적 시료의 분석을 가능하게 하는 것으로 확인되었다. 구현예는 염색체 영역이 복제수 이상(CNA)을 나타내는지를 결정하기 위해 크기 밴드를 사용하는 단계를 포함할 수 있다. CNA는 이수성 또는 암과 관련이 있을 수 있다. 구현예는 또한, 암의 수준을 결정하기 위해 크기 밴드를 사용하는 단계를 포함할 수 있다.

[0013] **I. 크기-기초 분석**

[0014] 모체 혈장 내 태아-유래 분자는 모체 DNA 분자보다 짧은 것으로 실증되었다(문헌[Chan *et al. Clin Chem* 2004; 50:88-92; Lo *et al. Sci. Transl. Med.* 2010; 2:61ra91-61ra91]). 연구자들은 NIPT를 위해 모체 혈장 시료에서 태아 DNA를 농화시키기 위해 이러한 크기 차이를 이용하였다(문헌[Li *et al. Clin Chem* 2004; 50:1002-1011, *JAMA* 2005; 293:843-9; Lun *et al. Proc. Natl. Acad. Sci. U. S. A.* 2008; 105:19920-5]). Yu 등은, 짝형성된-말단 시퀀싱 데이터에서 이수성 염색체로부터 짧은 단편의 이탈적인 비율을 결정함으로써 태아 염색체 이수성이 검출될 수 있음을 예시하였었다(문헌[Yu *et al. Proc. Natl. Acad. Sci. U. S. A.* 2014; 111:8583-8]). 이러한 접근법은 모체 혈장 내 DNA 분자의 카운팅과 비교한 경우 양호한 NIPT 성능을 달성할 수 있다(문헌

[Yu *et al. Proc. Natl. Acad. Sci. U. S. A.* 2014; 111:8583-8]).

[0015] 낮은 태아 DNA 분율(예를 들어, 4% 미만)을 갖는 임산부에서 태아 염색체 비정상(abnormality)의 비침습적 검출의 정확도를 향상시키기 위해, 이전에 탐구되어 온 하나의 가능한 방법은 인 실리코(in silico) 크기 선별 또는 물리적 크기 선별을 통한 짧은 DNA 분자의 선별적 분석이다(예를 들어, 2008년 7월 23일에 출원된 WO 2009/013496으로서, 이는 모든 목적을 위해 참조에 의해 본 명세서에 포함됨). 이들 방법에서, 짧은 혈장 DNA 분자로부터의 데이터 또는 분자는 통계학적 분석, 질병 분류 및 사례 해석을 위한 기초를 형성한다. 태아-유래 DNA 분자가 모체-유래 DNA 분자와 비교하여 더 짧은 크기 분포를 가지므로, 짧은 DNA 단편의 선별적 분석은 태아-유래 DNA 분자를 우선적으로 농화시켜, 더 높은 태아 DNA 분율을 초래할 수 있을 것이다.

[0016] 태아 DNA 분율이 NIPT 성능을 지배하는 주된 인자이므로, 이는 NIPT의 정확도를 잠재적으로 향상시킬 수 있다. 그러나, 150 bp 초과의 길이를 갖는 시퀀싱된 판독의 인-실리코 선별은 효과적인 태아 DNA 분율을 증가시킬 수 있을 것이지만, 태아 DNA 분율과 카운팅되는 분자의 수 사이의 트레이드-오프(trade-off) 때문에 단일-분자 카운팅에 의해 이수성 검출의 민감도를 본질적으로 증가시키지 않을 것으로 보고되었다(문헌[Fan *et al. Clin. Chem.* 2010; 56:1279-1286]). 다시 말해, 표 1에 제시된 바와 같이, 짧은 DNA를 선별하는 이전의 접근법은 카운팅되는 혈장 DNA 단편의 수의 두드러진 감소때문에 시퀀싱 깊이를 증가시키지 않으면서 민감도를 향상시킬 수는 없었다. 분석되는 혈장 DNA 단편의 크기를 감소시키는 것은 분석되는 DNA 단편의 수를 감소시킨다. 예를 들어, 100 bp 미만의 길이만 분석된다면, DNA 단편은 48.5배 감소를 받는다. 이와 동시에, 더 작은 혈장 DNA 단편에 초점을 맞춤으로써, 태아 DNA 분율은 농화된다. 예를 들어, 100 bp 미만의 길이에 대해, 태아 DNA 분율은 1.78배 농화를 갖는다. 그러나, 1.78배 농화는, 분석되는 혈장 DNA 분자에서 48.5배 감소와 비교하여 작다.

표 1

[0017]

혈장 DNA 크기 (bp)	태아 DNA 분율에서의 배수(fold) 농화 (x)	분석되는 혈장 DNA 분자에서의 배수 농화 (x)
<150	1.93	4.67
<120	2.04	21.2
<110	1.91	32.3
<100	1.78	48.5

[0018] 한편, 본 발명자들은 이전에, 소정의 역치, 예를 들어, 150 bp 크기 미만의 DNA 분자를 이용함으로써 진단 특이도를 향상시키기 위해 또 다른 혈장 DNA 크기-기초 접근법(미국 특허 제8,620,593호)을 개발하였다. 이 방법에서, 잠재적인 이수성 염색체로부터 유래되는 혈장 DNA 분자의 평균 크기는 다른 염색체로부터 유래되는 혈장 DNA 분자의 평균 크기와 비교된다. 이 접근법은, 태아 염색체 이수성이 과다표현된(overrepresented) 염색체(예를 들어, 삼염색체성 염색체)로부터의 혈장 DNA 분자의 평균 크기의 단축 또는 과소표현된(underrepresented) 염색체(예를 들어, 일염색체성(monosomic) 염색체)에 대한 혈장 DNA 분자의 평균 크기의 연장을 초래할 것이기 때문에 염색체 이수성의 비-침습적 검출의 특이도를 향상시키는 것으로 나타났다. 그러나, 이러한 접근법은 카운팅되는 혈장 DNA 분자의 수의 감소때문에 민감도를 증강시키는 것으로 예상될 수 없었다.

[0019] 이수성 염색체의 복제수 변화를 정량화하기 위해 특정한 짧은 DNA 분자의 인 실리코 선별을 사용하려고 시도하는 일부 이전의 노력이 있었다(문헌[Fan *et al. Clin. Chem.* 2010; 56:1279-1286]). 그러나, 이러한 특이적인 크기 선별은 최종 임상 분류에 기여할 DNA 분자의 수를 감소시켜, 따라서 추계학적 변동을 증가시킬 것이다. 분석적으로, 추계학적 변동에서 이러한 증가는 변동 계수(CV) 또는 표준 편차(SD)의 증가로서 나타날 수 있다. 푸아송 분포(Poisson distribution)에 따르면, 분석되는 분자의 수에서 모든 4-배 감소에 대해, CV는 2-배 증가할 것이다. 한편, 순환형 태아 DNA의 분획 농도(fractional concentration)에서 모든 2-배 증가에 대해, 태아 염

색체 이수성의 올바른 진단에 도달하기 위해 당업자가 카운팅하는 데 필요할 분자의 수는 4-배만큼 저하될 것이다. 당업자가 150 bp 미만의 분자에 대해 크기 선별을 사용한다면, 태아 DNA 분율은 약 2-배 증가할 것이지만 혈장 DNA 분자의 수는 4.7-배 저하될 것이다. 따라서, 단순 크기 선별을 통한 태아 DNA 분율의 농화는 혈장 DNA 분자의 감소의 유해 효과를 효과적으로 보상할 수 없을 것이며, 이는 단순 인 실리코 크기 선별에 의해 NIPT에서 일관된 향상이 없었던 중요한 이유일 것이다(문헌[Fan *et al. Clin. Chem.* 2010; 56:1279-1286]).

[0020] **II. 크기 패턴**

[0021] 이 연구에서, 본 발명자들은 일련의 상이한 크기 범위에 걸쳐 분자 카운트의 상세한 변화 패턴을 이용함으로써 혈장 DNA 크기 정보를 통합하는 새로운 방식을 개발하였으며, 이는 경험적 데이터에 따르면 놀랍게도 검사 민감도에서 향상을 초래하였다. 이는, 혈장 DNA 분자를 하나 이상의 크기 밴드로 분획화하는 경우, 1개 크기 밴드당 훨씬 더 적은 시퀀싱된 DNA 분자가 있어야 하고, 각각의 밴드 내의 혈장 DNA 분자 단독으로는 민감도를 향상시킬 수 없었기 때문에 직관에 반대된다. 하나의 특정 밴드를 단독으로 사용하는 대신, 본 발명자들의 새로운 접근법은 성능을 향상시키기 위해 상이한 밴드에 걸쳐 관계를 사용하는 것이다.

[0022] 본 발명자들은, 이수성 염색체의 게놈 표현(GR)의 변화가 상이한 크기의 혈장 DNA 분자에 존재하는 측정된 태아 DNA 분율에 따라 변할 것으로 판단하였다. 본 발명자들은, 세포-무함유 태아 및 모체 DNA 크기가 2개의 별개의 단편화 패턴을 반영하기 때문에, 영향을 받는 염색체의 GR 변화들 사이의 관계가 비-무작위 방식에서 상이한 크기 범위(크기 밴드)에 연결될 것으로 가정하였다(문헌[Lo *et al. Sci. Transl. Med.* 2010; 2:61ra91-61ra91]). 따라서, 본 발명자들은 상이한 크기 밴드 중에서 이탈적인 염색체로부터 기원하는 GR 값의 상세한 변화 모양을 분석하기 위해 새로운 접근법을 개발하였다. 이 접근법의 개략적인 원리는 도 1에서 예시된다.

[0023] 도 1은 혈장 DNA 크기-밴드 분석의 원리의 개략적인 예시(100)를 보여준다. 모체 혈장은 각각 태아 세포 및 모체 세포로부터 기원하는 태아 DNA 분자(구획(104) 및 분자(106)에서 파형 적색 라인)와 모체 DNA 분자(구획(108) 및 분자(110)에서 파형 검정색 라인)의 혼합물을 포함한다. 태아 DNA 분자는 일반적으로, 모체 DNA 분자에 비해 좌측으로 시프트하고 있는 태아 DNA 크기 프로파일에 의해 입증된 바와 같이 모체 DNA 분자보다 짧다. 따라서, 측정된 태아 DNA 분율은, 일반적으로 더 짧은 크기 범위에서 농화하는 상이한 크기 밴드에 따라 변할 것이다. 따라서, 삼염색체성 태아를 갖는 임산부의 경우, 기준 그룹으로부터의 편차가 z-점수에 의해 측정될 수 있는 측정된 게놈 표현(GR)은 상이한 크기 밴드에 따라 변할 것으로 예상될 것이지만, 대조적으로 어떠한 특이적인 변화도 정배수성 태아를 갖는 임산부에서 발생하지 않을 것이다.

[0024] 도 1은 크기 밴드를 별개의 밴드로서 그리고 슬라이딩 윈도우로서 보여준다. 빈도 대 크기의 그래프에서, 상이한 유색 컬럼(예를 들어, 컬럼(112))은 크기 밴드를 별개의 크기 범위에 상응하는 것으로 보여준다. z-점수(chr21) 대 크기의 그래프(116 및 118)에서, 유색 컬럼(예를 들어, 컬럼(122) 및 컬럼(124))은 상이한 크기 밴드에 대한 z-점수를 보여준다. z-점수 대 크기의 그래프에서 라인(126 및 128)은 크기 밴드에 대한 결과를 슬라이딩 윈도우로서 보여준다. 이수성 태아를 갖는 임산부에서, 라인(128)은 특정 크기에 집중된 크기 밴드에 대한 z-점수를 나타낸다. 예를 들어, 라인(128) 상의 주어진 x-좌표 및 y-좌표를 갖는 데이터 포인트는 x-좌표에 의해 나타낸 크기 주변으로 집중된 크기 범위에 대한 y-좌표로 나타낸 z-점수를 가진다. 각각의 z-점수는 전체 크기 밴드에 대해 계산된 풀링된(pooled) z-점수이다. 그러므로, 정배수성 태아를 갖는 임산부의 그래프(116)에서, 라인(126)은 크기 밴드에 대한 결과를 슬라이딩 윈도우로서 보여준다. 이수성 태아를 갖는 임산부의 그래프(118)에서, 라인(128)은 크기 밴드에 대한 결과를 슬라이딩 윈도우로서 보여준다.

[0025] 크기 밴드가 별개의 또는 슬라이딩 윈도우에 기초하는지 상관없이, 크기 밴드의 z-점수의 모양 또는 패턴은 정배수성 태아를 갖는 임산부와 이수성 태아를 갖는 임산부 사이에서 명백하게 상이하다. 예를 들어, 그래프(116) 및 그래프(118)에서 제시된 바와 같이, 이수성 태아를 갖는 임산부는 정배수성 태아를 갖는 임산부에서의 더 원통형인 패턴과 비교하여 쌍봉(bimodal) 패턴을 보여준다.

[0026] 상이한 크기 밴드에 걸쳐 카운트의 패턴은 태아 DNA 분율, 종양 DNA 분율, 또는 다른 임상적-관련 DNA 분율과 관련이 있을 수 있다. 따라서, 상이한 크기 밴드에 걸친 일련의 분자 카운트 및 상이한 크기-밴드 기초 판독 사이의 관계를 동시에 정량화하는 이러한 새로운 접근법은, 단지 특이적인 크기의 DNA 분자를 사용하는 접근법과 비교하여 혈장 DNA 크기 특성을 통합하는 경우 혈장 DNA 분자를 상실하지 않을 것이다. 이러한 동시적인 정량화는 소정의 크기 컷오프 미만의 단지 단일 판독의 사용과 비교하여 정확도를 향상시킬 것이다. 혈장에서 복제수 변화의 크기-밴드 패턴은 비제한적으로 기계 학습 접근법, 예컨대 인공 신경망, k-최근접 이웃 알고리즘(k-nearest neighbors algorithm), 서포트 벡터 머신(support vector machine), 및 혼합 가우스 모델(mixture Gaussian 모델) 등의 사용으로 인지될 수 있다.

[0027] *A. 크기 패턴 데이터 분석의 확증*

[0028] 크기 패턴(즉, 특정 크기 밴드에서 세포-무함유 DNA의 양과 관련된 분율의 모양 또는 매개변수)은 세포-무함유 DNA의 특징에 의존할 수 있다. 예를 들어, 크기 패턴은, 도 1의 그래프(116 및 118)에서와 같이 생물학적 시료가 이수성 태아로부터의 세포-무함유 DNA를 포함하는지에 의존할 수 있다. 첫째, 상이한 크기의 DNA에 대한 태아 DNA 분율은, 소정의 크기의 세포-무함유 DNA가 모체 DNA와 비교하여 태아 DNA에 대해 농화되는지 보여주기 위해 분석된다. 둘째, 이수성 태아를 갖는 임산부로부터의 데이터는, 정배수성 태아를 갖는 임산부로부터의 데이터에 대한 크기 밴드를 사용하여 분석된다. 이들 분석은, 크기 패턴이 분석되어, CNA가 이수성 태아의 결과인 경우를 포함하여 CNA에서의 차이를 구별할 수 있음을 확증한다.

[0029] **1. 측정된 태아 DNA 분율은 상이한 크기 밴드에 따라 달라진다**

[0030] 태아 DNA 분율 변화가 비-무작위 방식에서 단편 크기에 따라 달라질 것이라는 가설을 확증하기 위해, 본 발명자들은 본 발명자들의 이전의 연구에서 기재된 데이터를 재분석하였다(문헌[Chan *et al. Proc. Natl. Acad. Sci.* 2016; 113:E8159-E8168]).

[0031] 도 2a는 50 내지 400 bp 범위의 혈장 DNA 단편 크기에 대한 이수성 염색체에 대해 측정된 태아 DNA 분율을 보여준다. x-축은 DNA 분자의 크기이고, y-축은 태아 DNA인 해당 크기에서의 DNA 분자의 분율이다. 예를 들어, 120 bp 크기에서 태아 DNA 분율은 70.5%이고, 이는 120 bp 크기를 갖는 DNA 분자 중에서, 70.5%는 태아로부터 유래되고 이들 중 29.5%는 임산부로부터 유래됨을 의미한다. 태아 DNA 분율은 남자 태아를 갖는 임산부로부터의 시료에 대한 염색체 Y 백분율로부터 결정되었다. 태아 DNA 분율은 각각 120 bp 및 280 bp 크기에서 농화되는 것으로 확인되었다. 최대 70.5%의 태아 DNA 분율은 120 bp 크기에서 확인되었으며, 이는 17.4%의 태아 DNA 분율을 갖는 200 bp 크기에서 최저값보다 4x 더 높다.

[0032] **2. 혈장 DNA에서의 CNA는 상이한 크기 밴드에 대해 달라진다**

[0033] 불균일한 패턴을 나타내는 태아 DNA 분율의 변화는 이수성 염색체로부터 기원하는 분자 카운트의 제시에 영향을 줄 것이다. 이수성 염색체는 비정상적인 수의 염색체를 가진다. 태아에서 비정상적인 수의 염색체는 모체 DNA와 비교하여 태아 DNA의 양에 영향을 줄 것이다. 예를 들어, 21번 삼염색체성은 21번 염색체를 단지 2개 대신에 3개 갖는다. 태아가 21번 삼염색체성을 갖는다면, 태아 DNA는 정상적인 정배수성 태아보다 높은 분율을 가진다. 태아 DNA가 종종 모체 DNA보다 짧으므로, 21번 삼염색체성을 갖는 태아를 갖는 임산부의 모체 시료는, 정배수성 태아를 갖는 임산부의 모체 시료와 비교하여 21번 염색체로부터 더 높은 농도의 짧은 DNA를 갖는 경향이 있을 것이다.

[0034] 도 2b는 21번 삼염색체성 태아를 갖는 임산부 및 정배수성 태아를 갖는 임산부에 대해 크기 밴드 슬라이딩 윈도우를 사용한 z-점수 결과를 보여준다. 크기 밴드 슬라이딩 윈도우의 밴드폭은 50 bp였다. 21번 삼염색체성 태아를 갖는 임산부는 4%의 태아 DNA 분율을 가졌다. 도 2b에 제시된 바와 같이, 21번 삼염색체성 태아에 대한 120-bp 위치는 분석된 모든 시료 중에서 최고 z-점수를 가졌고, 따라서, 측정된 복제수 이상의 최고 정도에 상응하였다. 상이한 크기 밴드는 120 bp 및 다른 크기에서 z-점수의 규모에 영향을 줄 것이다. 영향을 받는 염색체의 z-점수의 계산은 하기에 기재되어 있다.

[0035] 50-bp 밴드폭을 갖는 크기 밴드의 중간점이 길이 i에 위치한다는 것을 추정하면(예를 들어, 75 bp의 i에서 위치한 크기 밴드와 밴드의 중간점은 50 내지 100 bp일 것임), 표적화된 염색체(예를 들어, 염색체 21)로 맵핑하는 시퀀싱 판독의 백분율은 게놈 표현 i(즉, GR_i)로 표시되는 관심 특정 크기(예를 들어, 50 내지 100 bp) 범위 내에서 이러한 단편을 사용하여 계산될 수 있다. 길이 i에 대한 z-점수는 계산된다:

[0036] $$Z-점수_i = \frac{GR_i - M_i}{SD_i}$$

[0037] 여기서, M_i 및 SD_i는 길이 i에서 집중된 크기 밴드에 대한 표적화된 염색체의 게놈 표현의 평균 및 표준 편차를 나타내며, 이는 이 연구에서 정배수성 태아를 임신한 50명의 임산부로부터 추론되었다. 크기의 전스펙트럼(full spectrum)은 50 내지 400 bp 범위의 크기 프로파일에서 크기 밴드의 중간점의 장소를 동적으로 변화시킴으로써 조사될 것이다.

[0038] 도 2b에서, 본 발명자들은 21번 삼염색체성 태아를 갖는 임산부에 대한 크기-밴드 기초 z-점수 곡선(202)에서 규칙적인 파형(wave-like) 패턴을 관찰할 수 있다. 이러한 관찰은 상이한 크기 밴드에서 태아 DNA 분율의 변화

를 연상시켰다. 그러나, 정배수성 태아를 갖는 대조군에서는 이러한 패턴이 제시되지 않았다. 특정 크기 밴드에서 이러한 변화의 규모는 태아 DNA 분율의 변화와 상이한 것으로 보였다. 예를 들어, 120 bp에서 z-점수는 280 bp에서의 z-점수보다 훨씬 더 높았으나(도 2b), 태아 DNA 분율은 이들 2개 크기 사이에서 유사하였다(도 2a). 변동성은, 166 bp보다 짧은 길이와 비교하여 166 bp보다 긴 길이에서 더 빠르게 감소하는 분자 카운트의 결과일 수 있어서, 높은 시료화 변동은 긴 분자에 존재할 것이다.

[0039] 도 2b는 또한, x-축 상에서 "모두(All)"로 표지된 값에 상응하는 원형으로 예시된, 모든 크기에 대한 z-점수를 보여준다. 가장 높은 원형인 적색 원형(204)은 21번 삼염색체성에 상응한다. 적색 원형(204)은 3 미만의 z-점수를 가진다. 따라서, 당업자가 모든 단편을 사용하고 3의 z-점수를 컷오프로서 이용할 것이라면, 이러한 사례는 정배수성 태아로서 잘못 분류되어 위음성 결과를 초래할 것이다. 대조적으로, 당업자가 상이한 크기 밴드에 대해 다양해지는 z-점수에서 별개의 모양의 변화를 사용할 것이라면, 해당 사례는 대조군과 비교하여 21번 삼염색체성 사례로서 올바르게 식별될 수 있다.

[0040] *B. 크기 패턴 분석의 적용*

[0041] 정배수성 태아 또는 이수성 태아를 갖는 임산부에 대해 크기 패턴 데이터를 발생시켰다. 그 후에, 데이터는 기계 학습 모델을 포함한 상이한 기법에 의해 분석되어, 크기 패턴이 정배수성 태아를 갖는 임산부와 이수성 태아를 갖는 임산부 사이를 구별하는 데 사용될 수 있는지를 결정하였다.

[0042] **1. 혈장 내 CNA의 크기-밴드 모양은 낮은 태아 분율을 갖는 염색체 이수성을 알려준다**

[0043] 이러한 크기-밴드 기초 z-점수 패턴이 낮은 태아 DNA 분율을 갖는 다른 시료로 일반화될 수 있는지 평가하기 위해, 본 발명자들은 각각 21번 삼염색체성 태아를 갖는 48개 사례와 각각 정배수성 태아를 갖는 63개 사례를 포함하여 각각 남자 태아를 갖는 부가적인 111개의 모체 혈장 DNA 시료를 분석하였다. 태아 DNA 분율은 남자 태아로부터 유래된 Y 염색체 서열을 사용하여 추정되었다(문헌[Hudecova *et al. PLoS One* 2014; 9:e88484; Chiu *et al. BMJ* 2011; 342:c7401]). 4% 이하의 낮은 태아 DNA 분율을 갖는 충분한 사례를 갖기 위해, 삼염색체성 태아를 갖는 48명의 임산부에 대한 각각의 짝형성-말단 시퀀싱 데이터세트를 정배수성 태아를 갖는 사례로부터의 시퀀싱 데이터세트와 인 실리코 혼합하여, 4% 이하의 태아 DNA 분율 수준을 달성하였다.

[0044] 도 3은 4%의 태아 DNA 분율을 갖는 상이한 개별적인 임산분에 걸친 이수성 염색체에 대한 측정된 게놈 표현(GR)의 크기-밴드 기초의 변화 패턴을 보여준다. Y-축은 z-점수 값을 나타내었으며, 정배수성 태아를 갖는 임산부와 비교하여 이수성 태아를 갖는 여성 임산부에서 측정된 GR에 대한 파생도(degree of derivation)를 시사한다. X-축은 상이한 크기 밴드를 나타내었다. 적색 라인(또한 더 짙은 라인)은 삼염색체성 태아를 갖는 임산부를 나타내었고; 회색 라인은 정배수성 태아를 갖는 임산부를 나타내었다.

[0045] 도 3은 삼염색체성 태아를 갖는 대부분의 모든 사례가, 정배수성 태아를 갖는 사례와 비교하여 측정된 복제수 이상의 일관적으로 상이한 크기-밴드 기초 패턴을 나타내었음을 보여준다. 각각의 사례에서, 21번 삼염색체성 사례의 크기 패턴에 대한 라인은 정배수성 사례에 대한 패턴과 명백하게 상이하며, 이는 도 2b에 제시된 바와 같이 21번 삼염색체성이 모든 크기 단편에 대해 z-점수를 사용하는 것보다 더 쉽게 결정되게 할 수 있다.

[0046] 본 발명자들은 추가로, 삼염색체성을 임신한 임산부와 정배수성 사례 사이에서 데이터 구조를 시각화하기 위해 히트맵 및 t-SNE(t-분포 확률적 임베딩) 접근법을 사용하였다. 도 4a는 정배수성 태아를 갖는 임산부와 21번 삼염색체성 태아를 갖는 임산부 사이에서 크기-밴드 기초의 변화 패턴의 히트맵 플롯을 보여준다. 청색(예를 들어, 영역(402))은 정배수성을 나타내는 크기 밴드의 특징에 대한 것인 한편, 녹색(예를 들어, 영역(404))은 21번 삼염색체성을 나타내는 크기 밴드의 특징에 대한 것이다. 도 4a에서 대부분의 모든 사례(46/48, 96%)는 21번 삼염색체성 태아 사례를 함께 클러스터링하는 단계를 수반한다. 유사하게는, 도 4a에서 정배수성 태아를 수반하는 대부분의 모든 사례(62/63, 98%)는 함께 클러스터링되었다.

[0047] 도 4b는 정배수성 태아를 갖는 임산부와 21번 삼염색체성 태아를 갖는 임산부 사이에서 크기-밴드 기초의 변화 패턴의 t-SNE 플롯을 보여준다. t-SNE 플롯은 기계 학습으로부터 결정된 2가지 특징에 기초한다. t-SNE 플롯은, 21번 삼염색체성 사례를 갖는 임산부가 정배수성 사례를 갖는 임산부로부터 쉽게 구분될 수 있다는 일관된 결과를 제공하였으며(도 4b), 이는 혈장 DNA에서 측정된 복제수 이상의 크기-밴드 기초 모양이 낮은 태아 DNA 분율, 예컨대 4%를 갖는 사례에 대해 염색체 이수성을 알려줄 수 있었음을 시사한다.

[0048] 도 4c는 정배수성태아를 갖는 임산부와 21번 삼염색체성 태아를 갖는 임산부 사이에서 종래의 z-점수 접근법을 사용한 z-점수 분포를 보여준다. 파선은 3의 z-점수 역치를 나타낸다. 3의 z-점수 컷오프를 사용하는 경우, 21번 삼염색체성의 검출률은 단지 48%일 것이다. 다시 말해, 52%의 21번 삼염색체성은 위음성을 초래할 것이다.

또한, 도 4c는 1명의 정배수성 임산부가 21번 삼염색체성에 대해 위양성을 초래할 것임을 보여준다. 종래의 z-점수 접근법은, 임의의 위양성 또는 위음성을 발생시키지 않았던 도 4b에서의 t-SNE 접근법과 비교하여 더 낮은 민감도 및 특이도를 초래할 것이다.

[0049] **2. 낮은 태아 DNA 분율을 갖는 사례를 검출하기 위한 기계 학습 패턴 인지**

[0050] 본 발명자들은 태아 복제수 이상을 검출하기 위한 크기-밴드 기초 접근법의 사용을 추가로 실증하기 위해 신경망 모델을 이용하였다. 본 발명자들은 시료를 트레이닝 데이터세트 및 검사 데이터세트로 나누었다. 트레이닝 데이터세트는 21번 삼염색체성 태아를 갖는 33명의 임산부와 정배수성 태아를 갖는 63명의 사례를 포함하였고, 검사 데이터세트는 15명의 21번 삼염색체성 태아 사례와 50명의 정배수성 태아 사례를 함유하였다. 각각이 20개 뉴런을 포함하는 하나의 층으로 구축된 신경망을 사용하여, 크기 밴드에 감춰진 패턴을 포착하는 모델을 학습하였다. 이후, 본 발명자들은 이 모델을 검사 데이터세트에 적용하였다.

[0051] 도 5는 신경망 모델을 위한 트레이닝 데이터세트 및 검사 데이터세트를 보여준다. 21번 삼염색체성의 확률에 대해 0.7의 컷오프를 이용하여, 본 발명자들이 1%, 2%, 3% 및 4%의 태아 DNA 분율에 대해 98%의 특이도에서 각각 40%, 80%, 100%, 및 100% 민감도를 달성할 수 있음이 판명되었다. 심지어 1%의 낮은 태아 DNA 분율에서, 신경망 모델은 21번 삼염색체성에 대해 진양성을 식별하는 능력을 보여준다.

[0052] 신경망 모델 이외의 기계 학습 모델을 사용하여, 대상체에서 태아 이수성 또는 암의 확률을 결정할 수 있는 패턴 및 특징을 결정할 수 있다. 이들 기계 학습 모델의 트레이닝은 장애 또는 임상적-관련 특징에 의해 영향을 받는 대상체 및 그렇지 않은 대상체로부터의 시료를 포함하는 데이터세트를 사용할 수 있다. 트레이닝을 위해 고려될 수 있는 매개변수는 크기 밴드의 밴드폭, 크기 밴드의 중앙점, DNA 분자의 양, DNA 분자의 장소, 후성유전체적(epigenomic) 신호(예를 들어, 메틸화), 및 다른 변수를 포함한다.

[0053] **3. 복제수 이상을 검출하는 예시적인 방법**

[0054] 도 6은 염색체 영역이 대상체로부터의 생물학적 시료에서 복제수 이상을 나타내는지를 결정하는 방법(600)을 보여준다. 생물학적 시료는 임상적-관련 DNA 분자 및 다른 DNA 분자를 포함하는 세포-무함유 DNA 분자의 혼합물을 포함할 수 있다. 임상적-관련 DNA 분자는 태아 DNA 또는 모체 DNA를 포함할 수 있다. 임상적-관련 DNA 분자가 태아 DNA를 포함한다면, 다른 DNA는 모체 DNA를 포함할 수 있다. 임상적-관련 DNA 분자가 모체 DNA를 포함한다면, 다른 DNA는 태아 DNA를 포함할 수 있다. 임상적-관련 DNA는 종양 DNA를 포함할 수 있으며, 이때 다른 DNA 분자는 비-종양 DNA를 포함한다.

[0055] 블록(602)에서, 방법(600)은 복수의 크기 범위의 각각의 크기 범위에 대한 크기 범위에 상응하는 생물학적 시료로부터의 세포-무함유 DNA 분자의 제1 양을 측정하는 단계를 포함할 수 있다. 세포-무함유 DNA 분자는 특정 게놈 영역으로부터의 것일 수 있으며, 상기 게놈 영역은 염색체 또는 염색체의 일부일 수 있다. 예를 들어, 게놈 영역은 염색체 아암일 수 있다. 게놈 영역은 게놈으로부터의 임의의 영역일 수 있다. 일부 구현예에서, 세포-무함유 DNA 분자는 다수의 분리(disjoint) 또는 하나의 연속 게놈 영역으로부터의 것일 수 있다. 크기 범위는 본원에 기재된 크기 밴드일 수 있다.

[0056] 사용하려는 특정 크기 범위는 기계 학습 모델에 의해 결정될 수 있다. 기계 학습 모델은 데이터세트 상에서 트레이닝될 수 있고, 상기 모델은 복제수 이상 또는 임상 질환을 검출하기 위한 민감도 및 특이도를 최적화하기 위해 어떤 범위(예를 들어, 크기 범위의 중앙점 위치 및/또는 밴드폭)가 사용되는지에 따라 달라질 수 있다. 데이터세트는 복수의 기준 크기 패턴을 포함할 수 있다. 기계 학습 모델은, 크기 범위의 소정의 밴드폭이 유리함을 결정할 수 있다. 또한, 기계 학습 모델은, 소정의 크기 범위가 다른 것들보다 예측 결과에 더 중요할 수 있음을 결정할 수 있다. 예를 들어, 크기 범위는 100 bp 내지 150 bp의 임의의 크기 주변으로 집중된 슬라이딩 크기 범위인 것으로 결정될 수 있다. 다른 구현예에서, 기계 학습 모델은, 별개의, 비-중첩 크기 범위가 슬라이딩 크기 범위를 능가하여 향상된 결과를 제공할 수 있음을 결정할 수 있다. 트레이닝 세트에서의 민감도 및/또는 특이도 또는 다른 정확도에 관한 비용 함수는 기계 학습 모델에 대한 매개변수 및 특징 선별(예를 들어, 사용하는 크기 범위 및 특이적인 크기 비)을 업데이트하는 데 사용될 수 있다. 확증 데이터세트는 또한, 모델의 정확도를 확증하는 데 사용될 수 있다.

[0057] 블록(604)에서, 방법(600)은 복수의 크기 범위의 각각의 크기 범위에 대해, 크기 범위에 상응하는 세포-무함유 DNA 분자의 제1 양 및 크기 범위 내의 것이 아닌 크기를 포함하는 제2 크기 범위 내의 DNA 분자의 제2 양을 사용하여 크기 비를 컴퓨터 시스템에 의해 계산하는 단계를 포함할 수 있다. 크기 비는 세포-무함유 DNA 분자의 z-점수 또는 정규화된 양(예를 들어, 분율, 백분율 또는 상대 존재비)일 수 있다. 예를 들어, 크기 비는 게놈

표현(GR)일 수 있다. 다른 구현예에서, 크기 비는 GR로 계산된 z-점수일 수 있다(예를 들어, 도 2b에서 곡선(202) 상의 포인트에서의 z-점수 값).

[0058] 각각의 크기 범위는, 크기 범위에서 크기 범위의 수치를 기재하는 밴드폭을 가질 수 있다. 예를 들어, 밴드폭은 50 bp 내지 100 bp, 100 bp 내지 200 bp, 200 bp 내지 300 bp, 또는 300 bp 내지 400 bp의 범위일 수 있다. 100 bp에 집중된 50 bp의 밴드폭을 갖는 크기 범위는 75 bp 내지 125 bp를 경유할 것이다. 각각의 크기 범위는 복수의 크기 범위(예를 들어, 별개의 크기 밴드, 예컨대 도 1에서 컬럼(122) 및 컬럼(124))의 임의의 다른 크기 범위와 비-중첩할 수 있다. 다른 구현예에서, 각각의 크기 범위는 복수의 크기 범위의 적어도 하나의 다른 크기와 중첩할 수 있다. 이러한 방식으로, 크기 범위는 슬라이딩 윈도우로서 여겨질 수 있다. 따라서, 상기 슬라이딩 윈도우는 많은 크기(예를 들어, 도 1에서 라인(126) 또는 라인(128))에 걸쳐 연속적인 크기 비 값을 초래한다.

[0059] 제2 크기 범위는 복수의 크기 범위의 각각의 크기 범위보다 클 수 있다. 제2 크기 범위는 모든 크기의 세포-무함유 DNA 분자를 포함할 수 있거나, 측정된 세포-무함유 DNA 분자에 대한 게놈 영역에서 모든 크기의 세포-무함유 DNA 분자를 포함할 수 있다. 제2 크기 범위는 블록(602)에서 측정된 세포-무함유 DNA 분자와 동일한 게놈 영역(예를 들어, 동일한 염색체(들) 또는 염색체 아암(들))으로부터의 세포-무함유 DNA 분자를 포함할 수 있다. 제2 크기 범위는 또한, 블록(602)에서 측정된 세포-무함유 DNA 분자에 대한 게놈 영역 이외의 게놈 영역으로부터의 세포-무함유 DNA 분자를 포함할 수 있다. 예를 들어, 21번 삼염색체성의 경우, 블록(602)에서 측정된 세포-무함유 분자는 염색체 21로부터의 것일 수 있다. 이러한 사례에서, 제2 크기 범위는 다른 염색체(예를 들어, 기준으로서 역할을 하거나 전체 게놈에 걸친 상이한 염색체)로부터의 세포-무함유 DNA 분자를 포함할 수 있다. 그 후에, 방법(600)은 또한, 제2 크기 범위에 있는 세포-무함유 DNA 분자의 양을 측정하는 단계를 포함할 수 있다.

[0060] 블록(606)에서, 방법(600)은 복수의 크기 범위에 대한 복수의 기준 크기 비를 포함하는 기준 크기 패턴을 수득하는 단계를 포함할 수 있다. 기준 크기 패턴은 염색체 영역에 복제수 이상을 갖는 대상체로부터의 또는 복제수 이상을 갖지 않는 대상체로부터의 복수의 기준 시료로부터 결정될 수 있다. 예를 들어, 검사되는 복제수 이상이 태아 이수성과 관련이 있다면, 기준 시료는 정배수성 태아를 갖는 것으로 알려진 대상체로부터의 것일 수 있다. 다른 구현예에서, 기준 시료는 태아 이수성을 갖는 것으로 알려진 대상체로부터의 것일 수 있다. 복수의 크기 범위에 대한 각각의 기준 크기 비는, 생물학적 시료 대신에 기준 시료라는 점을 제외하고는, 블록(604)에서 계산된 크기 비와 동일한 방식으로 결정될 수 있다. 예를 들어, 도 2b에서, 기준 시료에 대한 크기 패턴은 곡선(202)을 제외하고는 도 2b의 곡선 중 임의의 하나일 수 있다. 기준 크기 패턴은 기준 시료에 대한 모든 크기 패턴의 통계학적 표현일 수 있다. 예를 들어, 기준 크기 패턴은 모든 크기 패턴의 평균(평균(mean), 중앙값 또는 모드(mode))일 수 있다. 예를 들어, 이러한 평균화된 기준 크기 패턴은 도 1에서 라인(126)일 수 있다.

[0061] 블록(608)에서, 방법(600)은 복수의 크기 비를 기준 크기 패턴과 비교하는 단계를 포함할 수 있다. 복수의 크기 비를 기준 크기 패턴과 비교하는 단계는 복수의 크기 비의 각각의 크기 비를 상응하는 크기 범위에서의 기준 크기 비와 비교하는 단계를 포함할 수 있다. 예를 들어, 복수의 크기 비는 도 1에서 라인(128)을 형성하는 포인트일 수 있다. 일부 사례에서, 복수의 크기 비는 라인(128)의 일부만 형성할 수 있다. 기준 크기 패턴이 도 1에서의 라인(126)이라고 가정하면, 복수의 크기 비를 기준 크기 패턴과 비교하는 단계는 라인(128)의 포인트와 라인(126)의 기준 포인트 사이의 통계학적 비교를 포함할 수 있다.

[0062] 각각의 크기 범위에 대한 각각의 크기 비는 상응하는 크기 범위에서의 기준 크기 비와 통계학적으로 유사한 것으로 결정될 수 있다. 통계학적 유사성은 역치를 사용하여 결정될 수 있다. 역치는, 크기 비가 기준 크기 비에 얼마나 근접해야 하는지 나타낼 수 있다. 역치는 기준 크기 비로부터 소정의 수의 표준 편차(예를 들어, 1, 2 또는 3)일 수 있다. 일부 구현예에서, 모든 크기 비가 기준 크기 비와 통계학적으로 유사할 필요는 없다. 대신에, 최소 수의 크기 비가 통계학적으로 유사할 수 있다. 예를 들어, 80%, 85%, 90% 또는 95%의 크기 비는 상응하는 기준 크기 비와 통계학적으로 유사할 수 있다.

[0063] 복수의 크기 비를 기준 크기 패턴과 비교하는 단계는, 복수의 크기 비를 복수의 기준 시료로부터 결정되는 복수의 역치 값과 비교하는 단계를 포함할 수 있다. 예를 들어, 각각의 크기 범위는 상이한 역치 값을 가질 수 있으며, 이러한 역치 값은 기준 시료에 대한 표준 편차에 기초할 수 있다. 단일 크기 범위는 또한, 상이한 역치 값을 가질 수 있으며, 이때 각각의 역치 값은 크기 비가 기준 시료로부터 상이한, 상이한 확실성(certainty) 수준과 연관이 있다. 비교는 초과된 역치 값의 수를 카운팅하는 단계, 및 그 수가 양 또는 분율(예를 들어, 0.5, 0.6, 0.7, 0.8 또는 0.9)을 초과하는지를 결정하는 단계를 포함할 수 있다. 그 수가 양을 초과한다면, 복제수

이상은 염색체 영역에 의해 나타나는 것으로 결정될 수 있다.

[0064] 일부 구현예에서, 복수의 크기 비를 기준 크기 패턴과 비교하는 단계는 복수의 크기 범위에 대한 복수의 크기 비를 포함하는 크기 패턴을 결정하는 단계를 포함할 수 있다. 크기 패턴은 크기 비를 크기 범위와 관련짓는 그래프일 수 있다. 예를 들어, 크기 패턴은 도 1에서 라인(128), 도 2b에서 곡선(202), 또는 도 3에서 임의의 21번 삼염색체성 라인일 수 있다. 크기 패턴은 기준 크기 패턴과 유사한 모양을 갖는 것으로 결정될 수 있다. 유사한 모양을 결정하는 단계는, 크기 패턴의 기울기(예를 들어, 1차 도함수) 및/또는 변곡점(예를 들어, 2차 도함수)이 기준 크기 패턴의 것들과 유사한지 결정하는 단계를 포함할 수 있다. 기울기 또는 변곡점의 유사성은 통계학적 유의성(예를 들어, 소정의 수의 표준 편차)을 나타낼 수 있는 역치를 사용하여 결정될 수 있다.

[0065] 일부 구현예에서, 복수의 크기 비를 기준 크기 패턴과 비교하는 단계는 신경망을 포함하는 기계 학습을 사용한 비교를 포함할 수 있다. 기계 학습 모델은, 크기 비를 어떻게 계산하는지, 크기 비를 기준 크기 패턴에 어떻게 비교하는지, 및/또는 크기 패턴이 기준 크기 패턴과 유사한지를 어떻게 결정하는지를 결정하는 데 사용될 수 있다. 크기 비를 어떻게 계산하는지는, 크기 범위의 밴드폭 및 제2 크기 범위의 크기와 밴드폭을 결정하는 단계를 포함할 수 있다. 크기 비를 기준 크기 패턴과 어떻게 비교하는지는, 상이한 크기 범위에 대한 가중치, 및 크기 패턴의 0차, 1차 또는 2차 도함수를 사용하는지의 여부를 결정하는 단계를 포함할 수 있다. 크기 패턴이 기준 패턴과 유사한지 어떻게 결정하는지는, 유사성에 대한 역치 값을 결정하는 단계를 포함할 수 있다.

[0066] 기준 크기 패턴을 수득하고 복수의 크기 비를 상기 기준 크기 패턴과 비교하는 단계는 복수의 크기 비를 기계 학습 모델에 입력하는 단계를 포함할 수 있다. 기계 학습 모델은 복수의 기준 시료로부터의 복수의 트레이닝 크기 패턴을 사용하여 트레이닝될 수 있다. 트레이닝된 기계 학습 모델(예를 들어, 신경망)은 염색체 영역에 이상을 갖는 시료의 확률을 출력할 수 있다.

[0067] 블록(610)에서, 방법(600)은 염색체 영역이 비교에 기초하여 복제수 이상을 나타내는지를 결정하는 단계를 포함할 수 있다. 복제수 이상은 21번 삼염색체성, 18번 삼염색체성, 13번 삼염색체성 및 성염색체 이수성을 포함한 이수성일 수 있다. 복제수 이상은 암의 지표일 수 있다. 방법(600)은 또한, 대상체의 암을 치료하는 단계 또는 이수성에 대한 플랜(plan)을 개발하는 단계를 포함할 수 있다.

[0068] 기준 크기 패턴이 복제수 이상을 갖는 대상체로부터의 복수의 기준 시료로부터 결정되고 비교가 크기 비 또는 크기 패턴이 기준 크기 패턴과 유사함을 보여준다면, 염색체 영역은 복제수 이상을 나타내는 것으로 결정될 수 있다. 또한, 비교가 크기 비 또는 크기 패턴과 기준 크기 패턴 사이에서 차이를 보여준다면, 염색체 영역은 복제수 이상을 나타내지 않는 것으로 결정될 수 있다. 일부 구현예에서, 복제수 이상을 나타내는 확률이 결정될 수 있다. 확률은, 크기 비 또는 크기 패턴이 기준 크기 패턴과 얼마나 유사한지 또는 다른지와 상관관계가 있을 수 있다. 확률은 신경망 또는 본원에 기재된 임의의 모델을 포함한 기계 학습 모델을 사용하여 결정될 수 있다.

[0069] 대안적으로, 기준 크기 패턴이 복제수 이상을 갖지 않는 대상체로부터의 복수의 기준 시료로부터 결정되고 비교가 크기 비 또는 크기 패턴이 기준 크기 패턴과 유사함을 보여준다면, 염색체 영역은 복제수 이상을 나타내지 않는 것으로 결정될 수 있다. 또한, 비교가 크기 비 또는 크기 패턴과 기준 크기 패턴 사이에서 차이를 보여준다면, 염색체 영역은 복제수 이상을 나타내는 것으로 결정될 수 있다.

[0070] *C. 낮은 태아 분율에서 향상된 정확도*

[0071] 혈장 DNA 내 측정된 복제수 이상의 크기-밴드 기초 패턴을 이용함으로써 접근법의 성능을 벤치마크하기 위해, 본 발명자들은 또한, 전형적인 z-점수(문헌[Chiu et al. Proc Natl Acad Sci U S A 2008; 105:20458-20463]) 및 크기 선별 방법을 사용하여 상이한 태아 DNA 분율, 예컨대 4%, 3%, 2% 및 1%에 걸쳐 특이도 및 민감도를 계산하였다. 태아 DNA가 모체 혈장 DNA에 존재하는 측정된 태아 DNA 분율의 최대를 120 bp에서 제공하였기 때문에(도 2a), 본 발명자들은 120 bp 주변의 크기 밴드가 모든 DNA 단편을 사용하는 것보다 양호한 성능을 제공할 것으로 가정하였다. 이를 위해, 본 발명자들은 105 내지 155 bp의 크기 밴드를 선택하였고, 상응하는 z-점수를 계산하였다.

[0072] 표 2는, 크기 선별과 함께 그리고 크기 선별 없이 종래의 카운팅-기초 방법과 비교된 크기-밴드 기초 패턴 인지의 성능을 보여준다. 혈장 DNA에서 측정된 복제수 이상의 크기-밴드 기초 패턴의 사용은 전형적인 z-점수 및 크기 선별 접근법과 비교하여 우수한 성능을 제공하였다. 예를 들어, 본 발명자들의 연구에서, 3%의 태아 DNA 분유에서, 측정된 복제수 이상의 크기-밴드 기초 패턴의 인지는 98%의 특이도와 함께 100% 민감도를 제공하였다. 비교로서, 종래의 카운팅 기초 접근법은 단지 10%의 민감도 및 98%의 특이도를 제공하였다. 150 bp 미만의 단편의 크기 선별을 사용하여, 민감도는 43%까지 향상되었다. 그러나, 120 bp까지 훨씬 더 짧은 크기의 단편의 선별

에서, 민감도는 20%까지 감소되었다. 이는, 본 발명에서 제안된 방법이 크기 선별을 사용하는 기존의 접근법을 능가하여 훨씬 더 양호한 분석 성능을 제공함을 나타낸다.

표 2

[0073]

태아 DNA 분율	종래의 카운팅-기초 접근법						측정된 복제수 이탈의 크기-밴드 기초 패턴(새로운 본 발명의 접근법)	
	크기 선별과 함께 (<120 bp)		크기 선별과 함께 (<150 bp)		크기 선별 없이			
	특이도	민감도	특이도	민감도	특이도	민감도	특이도	민감도
4%	96%	47%	100%	75%	98%	48%	98%	100%
3%	96%	20%	100%	43%	98%	10%	98%	100%
2%	96%	16%	100%	14%	98%	2%	98%	80%
1%	96%	6%	100%	6%	98%	2%	98%	40%

[0074] 증가된 정확도 외에도, 본 발명의 구현예는 감소된 양의 시퀀싱을 가능하게 할 수 있다. 크기 패턴 접근법은 소정의 크기의 서열 판독을 폐기하는 단계를 수반하지 않을 수 있고, 그 결과, 주어진 시퀀싱 깊이에서 더 많은 서열 판독이 분석에 사용된다. 따라서, 크기 패턴 접근법은 소정의 크기 범위에서 더 많은 판독을 제공하기 위해 부가적인 시퀀싱을 필요로 하지 않을 수 있다. 더욱이, 소정의 낮은 수준의 태아 분율에서 더 높은 시퀀싱 깊이를 이용하더라도, 크기 밴드 또는 크기 패턴을 사용하지 않는 접근법은 여전히, 21번 삼염색체성을 정확하게 결정하지 않을 수 있다. 낮은 태아 분율은, 크기 밴드 또는 크기 패턴이 분석되지 않는다면 21번 삼염색체성 사례와 정배수성 사례 사이에서 통계학적으로 유의한 크기 차이를 초래하지 않을 수 있다. 더욱이, 크기 밴드 또는 크기 패턴 없이 크기 선별을 사용하는 기존의 접근법이 다른 기법을 보충하는 데 사용될 수 있더라도, 크기 밴드 또는 크기 패턴을 사용하는 구현예는 독립적으로 사용되어 21번 삼염색체성 또는 복제수 이상을 결정할 수 있다.

[0075] 이 연구에서, 본 발명자들은 예를 들어 2%까지 연장하는 낮은 태아 DNA 분율을 갖는 임산부에 대해 NIPT가 수행되게 할 수 있는 신규 방법을 개발하였다. 신경망 모델 또는 다른 기계 학습 모델을 트레이닝하기 위해 더 많은 시료가 사용되므로, 본 발명자들은 검출 한계를 추가로 낮출 것으로 예상한다. 본 발명자들은, 모체 혈장 DNA에서 복제수 변화도가 삼염색체성 태아를 갖는 임산부와 정배수성 태아를 갖는 임산부 사이에서 상이한 크기 밴드에 관하여 별개의 패턴을 나타낼 것이라는 사실을 이용하였다. 이는, 2% 미만의 태아 DNA 분율까지 연장하는 태아 염색체 이수성의 비-침습적 검출의 한계를 낮춤으로써 넓은 집단 범위(broad population coverage)를 달성하기 위한 중요한 단계이다. 종래의 접근법을 사용하면, 4% 미만의 태아 DNA 분율을 수반하는 임산부는 NIPT에 적합하지 않았고, 일반적으로 보고 불가능한 결과 또는 검사 실패로 문제가 될 것이다.

[0076] 본 발명자들의 새로운 접근법은 더 낮은 검출 한계때문에 위음성률을 감소시킬 뿐만 아니라, 이수성을 가질 위험이 4% 미만의 태아 DNA 분율을 갖는 임산부에서 증가할 것임을 보여주는 많은 보고가 있었기 때문에 실제 PPV를 향상시키는 잠재성을 가진다(문헌[Norton *et al. N. Engl. J. Med.* 2015; 372:1589-1597]). 이전에, 일부 연구자는, 낮은 태아 DNA 분율을 갖는 임산부가 유전적 상담을 받고 이수성의 증가된 위험때문에 포괄적인 초음파 평가 및 진단 검사가 제공되어야 한다고 논쟁한다(문헌[Yaron *Prenat. Diagn.* 2016; 36:391-396]). 태아 DNA 분율이 일반적으로, 모체 체중과 반비례하기 때문에(문헌[Wang *et al. Prenat. Diagn.* 2013; 33:662-666;

Hudecova *et al. PLoS One* 2014; 9:e88484]), 높은 체질량 지수를 갖는 임산부는 특히, 낮은 태아 DNA 분율과 함께 시나리오를 민감하게 다루는 이러한 크기-밴드 기초 접근법의 능력으로부터 이익을 얻을 것이다. 본 발명자들의 새로운 접근법의 또 다른 사용은, 태아 DNA 분율이 일반적으로 더 낮을 때, NIPT가 임신 초기에(예를 들어, 임신 10주 전에) 수행되게 할 수 있을 것이다.

[0077] *D. 종양학에서 메틸화 수준 분석*

[0078] 복제수 이상(CNA)은 또한, 많은 암에서 존재한다. 그 결과, CNA는 대상체에서 암의 수준을 결정하는 데 사용될 수 있다. 또한, 암 환자는 종종, 소정의 게놈 영역에서 더 높은 수준의 메틸화를 나타낸다. 따라서, 메틸화 마커는 또한, 암의 수준을 결정하기 위해 크기 밴드 분석과 조합하여 사용될 수 있다.

[0079] **1. 메틸화를 이용한 크기 패턴 분석**

[0080] 본 발명자들은, 다른 유형의 암-연관 이상, 예컨대 메틸화는 또한, 비-암 대상체로부터 구별될 수 있을 특이적인 크기-밴드 기초 패턴을 작제하는 데 사용될 수 있을 것이다. 따라서, 본 발명자들은 또한, 상기 언급된 바와 같이 HCC 환자로부터의 4개의 혈장 DNA 시료를 추가로 분석하였다. 본 발명자들은 비제한적으로, 건강한 대상체의 기관에서 비메탈화되어야 하지만 암 환자에서는 메틸화되는 가능성을 훨씬 더 높게 갖는 영역에 대해 메틸화 수준을 정량화하기 위해, 표적화된 비설파이트(bisulfite) 시퀀싱을 사용하였다. 본 발명자들은 본원에 기재된 크기-밴드 기초 접근법을 적용하여, 건강한 대상체와 비교하여 메틸롬(methylomic) 이상의 측면에서 크기-밴드 연관된 패턴을 탐구하였다. 메틸화는 2013년 3월 15일에 출원된 미국 출원 13/842,209(2017년 8월 15일에 미국 특허 제9,732,390호로서 등록됨) 및 2015년 7월 20일에 출원된 미국 출원 14/803,692에 추가로 기재되어 있으며, 이들 둘 모두의 내용은 모든 목적을 위해 참조에 의해 본 명세서에 포함된다.

[0081] 도 7은 간세포암종(HCC) 환자의 혈장 DNA에서 측정된 메틸화의 크기-밴드 기초의 변화 패턴을 보여준다. z-점수는, HCC를 갖지 않는 것으로 알려진 건강한 대상체로부터의 기준 시료에 대한 평균적인(mean) 평균 메틸화 수준을 계산하고 평균 메틸화 수준과 연관된 표준 편차를 계산함으로써 계산된다. 각각의 크기 밴드에서 z-점수는 해당 크기 밴드에서의 메틸화 수준과 평균적인 평균 메틸화 수준 사이의 차이, 및 표준 편차로 나뉜 차이로서 계산된다. 도 7에서 파선은 +3 또는 -3의 z-점수를 나타내며, 이는 평균적인 평균 메틸화 수준으로부터의 통계학적 유의성을 나타내는 데 사용될 수 있다.

[0082] 적색 또는 더 짙은 라인(702, 704, 706 및 708)은 초기 HCC(eHCC)를 나타내었고, 회색 라인은 HCC가 없는 만성 B형 간염 바이러스(HBV) 보균자를 나타내었다. 도 7에서, 본 발명자들은 HCC 환자(라인(702, 704, 706 및 708))와 연관된 메틸롬 비정상의 별개의 크기-밴드 패턴을 확인시켜줄 수 있었고, 이는 HCC01, HCC02 및 HCC03에서 HBV 보균자(회색 라인)로부터 암 환자를 식별하게 할 수 있었다. 라인(702, 704 및 706)은, HBV 시료에 대한 회색 라인으로부터 상당히 더 높은 곳에서 보이는 적어도 2개의 피크를 갖는 패턴을 보여준다. 라인(708)은 회색 라인에 더 근접하지만 HBV 시료에 대한 회색 라인보다 높은 2개의 피크를 여전히 가진다. 각각의 그래프에서 "모두"로 표지된 최우측 데이터는 크기-밴드와 상관없이, 모든 데이터에 대한 풀링된(pooled) z-점수이다. HCC04에 대해, 비-무작위 크기-밴드 기초 곡선화 패턴은 모든 단편(원형(710)으로 표시됨)의 사용으로 이탈적인 메틸화의 전반적인 정도보다 더 유익한 것으로 판명되었다. 상이한 게놈 영역이 상이한 그래프에서 사용되었다. 염색체 아암 1q는 HCC01 및 HCC04에 사용되었으며, 10p는 HCC02에 사용되었고, 19q는 HCC03에 사용되었다. 다른 구현예에서, 예를 들어 비제한적으로, 저메틸화, 점 돌연변이, 하이드록시메틸화, 단편화 말단 등의 크기-밴드 기초의 변화 패턴이 또한, 암을 검출하는 데 사용될 수 있었다.

[0083] **2. 암의 수준을 결정하기 위한 예시적인 방법**

[0084] 도 8은 대상체로부터의 생물학적 시료에서 암의 수준을 결정하는 방법(800)을 보여준다. 생물학적 시료는 세포-무함유 DNA 분자의 혼합물을 포함할 수 있다. 세포-무함유 DNA 분자는 종양 DNA 분자 및 비-종양 DNA 분자를 포함할 수 있다.

[0085] 블록(802)에서, 방법(800)은 복수의 크기 범위의 각각의 크기 범위에 대한 크기 범위에 상응하는 생물학적 시료로부터의 메틸화된 세포-무함유 DNA 분자의 제1 양을 측정하는 단계를 포함할 수 있다. 메틸화된 세포-무함유 DNA 분자는 염색체 아암으로부터의 것일 수 있다. 크기 범위에 상응하는 메틸화된 세포-무함유 DNA 분자의 양을 측정하는 단계는, 세포-무함유 DNA 분자가 메틸화되는 점을 제외하고는, 방법(600) 또는 본원에 기재된 임의의 다른 방법에서 기재된 바와 같이 수행될 수 있다. 메틸화된 세포-무함유 DNA 분자의 제1 양은 하나 이상의 게놈 영역으로부터의 것일 수 있다. 게놈 영역은 염색체 아암, 예를 들어, 1p, 1q, 8p, 8q, 13q 또는 14p일 수 있다. 게놈 영역의 다양한 조합이 사용될 수 있다. 사용될 특정 영역은, 기지의 수준의 암을 갖는 시료의 트레이닝 세

트에서 암의 수준을 결정하기 위해 영역의 다양한 조합에 대한 정확도를 분석함으로써 결정될 수 있다.

[0086] 블록(804)에서, 방법(800)은 각각의 크기 범위에 대해, 크기 범위에 상응하는 메틸화된 세포-무함유 DNA 분자의 제1 양, 및 크기 범위 내의 것이 아닌 크기를 포함하는 제2 크기 범위 내의 DNA 분자의 제2 양을 사용하여 메틸화 수준을 컴퓨터 시스템에 의해 계산하는 단계를 포함할 수 있다. 제2 양은 메틸화된 세포-무함유 DNA 분자의 것일 수 있다. 이들 구현예 또는 다른 구현예에서, 제2 양은 비-메틸화된 세포-무함유 DNA 분자를 포함할 수 있다.

[0087] 메틸화 수준은 하나 이상의 부위에서 메틸화되거나 비메틸화된 DNA 분자의 DNA 분자의 z-점수 또는 정규화된 양(예를 들어, 분율, 백분율, 또는 상대 존재비)일 수 있다. 예를 들어, 메틸화 수준은 제2 양에 대한 제1 양의 비일 수 있다. 다른 구현예에서, 메틸화 수준은 z-점수일 수 있다. z-점수는 제2 양에 대한 크기 범위에 상응하는 세포-무함유 DNA 분자의 양의 비를 사용하여 계산될 수 있다. 따라서, 계산된 비와 평균적인 평균 비 사이의 차이를 표준 편차로 나누어서, z-점수를 결정한다. 평균적인 평균 비는 대조군(예를 들어, 비-암 환자, 기준 시료, 또는 암과 연관되지 않은 게놈 영역)에 대한 평균 메틸화 수준일 수 있다. 메틸화 수준이 z-점수라면, 크기 범위에 대한 메틸화 수준은 도 7에서 라인(702, 704, 706 및 708) 상의 임의의 포인트일 수 있다.

[0088] 블록(806)에서, 방법(800)은 복수의 크기 범위에 대한 복수의 기준 메틸화 수준을 포함하는 기준 크기 패턴을 수득하는 단계를 포함할 수 있다. 복수의 크기 범위는 기계 학습 알고리즘에 의해 결정될 수 있고, 방법(600)에 대해 기재된 것과 동일한 방식으로 결정될 수 있다. 기준 크기 패턴은 암을 갖는 대상체로부터의 또는 암을 갖지 않는 대상체로부터의 복수의 기준 시료로부터 결정될 수 있다. 예를 들어, 기준 시료는 HCC 또는 임의의 유형의 암을 갖지 않는 것으로 알려진 환자로부터의 것일 수 있다. 기준 크기 패턴은 HCC가 없는 만성 HBV 보균자로부터의 데이터에 기초할 수 있다. 예를 들어, 기준 크기 패턴은 도 7에서 HBV에 대한 임의의 회색 라인일 수 있다. 예를 들어, 기준 크기 패턴은 방법(600)으로 설명된 바와 같이, 기준 시료에 대한 모든 크기 패턴의 통계학적 표현일 수 있다.

[0089] 블록(808)에서, 방법(800)은 복수의 메틸화 수준을 기준 크기 패턴과 비교하는 단계를 포함할 수 있다. 방법(800)은 복수의 크기 비의 각각의 메틸화 수준을 상응하는 크기 범위에서의 기준 메틸화 수준과 비교하는 단계를 포함할 수 있다. 메틸화 수준을 기준 크기 패턴과 비교하는 단계는, 크기 대신에 메틸화 수준이 이용되는 점을 제외하고는, 크기 비가 방법(600)에서 기준 크기 패턴과 비교되는 방식으로 수행될 수 있다. 방법(800)은 각각의 메틸화 수준이 상응하는 크기 범위에서의 기준 메틸화 수준과 통계학적으로 유사한지 결정하는 단계를 포함할 수 있다. 일부 구현예에서, 방법(800)은 각각의 메틸화 수준 또는 일부 메틸화 수준이 상응하는 크기 범위에서의 기준 메틸화 수준과 통계학적으로 상이한지 결정하는 단계를 포함할 수 있다.

[0090] 일부 구현예에서, 복수의 메틸화 수준을 기준 크기 패턴과 비교하는 단계는 복수의 크기 범위에 대한 복수의 메틸화 수준을 포함하는 크기 패턴을 결정하는 단계를 포함할 수 있다. 크기 패턴은 기준 크기 패턴과 비교될 수 있다. 크기 패턴은 기준 크기 패턴과 유사한 모양을 갖는 것으로 결정될 수 있다. 방법(800)에서 기준 크기 패턴과의 비교는 방법(600)에서 기준 크기 패턴과의 비교와 유사할 수 있다.

[0091] 제1 양의 메틸화된 세포-무함유 DNA 분자가 1개 초과의 게놈 영역으로부터의 것이라면, 메틸화 수준은 게놈에서 이들의 장소에 기초하여 분석될 수 있다. 복수의 메틸화 수준은 다차원 벡터를 포함할 수 있다. 다차원 벡터는 N X M일 수 있으며, 이때 N은 크기 범위의 수이고 M은 게놈 영역의 수이다. 게놈 영역은 염색체, 염색체 아암, 또는 염색체 아암의 일부일 수 있다. 기준 크기 패턴은 유사하게는, 다차원 벡터(예를 들어, 크기 N X M)일 수 있다. 복수의 메틸화 수준은 기계 학습 모델 또는 다른 기법을 사용하여 기준 크기 패턴과 비교될 수 있다. 다차원 벡터 및 메틸화 수준의 사용은 하기에 기재된다(예를 들어, 도 13, 14a, 14b 및 14c).

[0092] 블록(810)에서, 방법(800)은 비교에 기초하여 암의 수준을 결정하는 단계를 포함할 수 있다. 암의 수준은 대상체가 암, 암의 가능성, 또는 종양 크기를 갖는지 또는 갖지 않는지를 포함할 수 있다.

[0093] 기준 크기 패턴이 암을 갖는 대상체로부터의 복수의 기준 시료로부터 결정되고 비교가 유사한 메틸화 수준 또는 유사한 모양의 결정을 포함한다면, 상기 대상체는 암을 갖는 것으로 결정될 수 있다. 이러한 기준 크기 패턴과 함께, 비교가 상이한 메틸화 수준 또는 상이한 모양의 결정을 포함한다면, 대상체는 암을 갖지 않는 것으로 결정될 수 있다. 기준 크기 패턴이 암을 갖지 않는 복수의 기준 시료로부터 결정되고 비교가 상이한 메틸화 수준 또는 모양의 결정을 포함한다면, 대상체는 암을 갖는 것으로 결정될 수 있다. 또한, 기준 크기 패턴이 암을 갖지 않는 복수의 기준 시료로부터 결정되고 비교가 유사한 메틸화 수준 또는 모양의 결정을 포함한다면, 대상체는 암을 갖지 않는 것으로 결정될 수 있다.

[0094] *E. 크기-밴드 매트릭스를 이용한 패턴 분석*

[0095] 다양한 암의 경우, 염색체 아암을 포함하는 소정의 게놈 영역은 복제수 이상을 가질 가능성이 더 클 수 있다. 따라서, 가능한 복제수 이상에 대해 염색체 아암에 의해 크기 범위를 분석하는 것이 암의 확률을 결정하거나 암을 검출하는 것을 돕는데 사용될 수 있다. 기계 학습 모델은 상이한 염색체 영역(예를 들어, 아암)에서의 크기 특징의 패턴에 기초하여 암 분류자를 결정하는 데 사용될 수 있다.

[0096] **1. 크기 패턴 분석**

[0097] 암 환자의 혈장에서 종양-유래 DNA의 크기 프로파일이 비(non)종양-유래 DNA 분자와 상이한 것으로 나타났으며, 상기 종양-유래 DNA의 크기 프로파일은 일반적으로 더 많은 짧은 DNA 분자를 포함하기 때문에(문헌[Jiang *et al. Proc. Natl. Acad. Sci.* 2015; 112:E1317-E1325]), 본 발명자들은 본 발명에서 기재된 크기-밴드 기초 접근법이 암-연관 이상, 예컨대 복제수 이상(CNA) 및 메틸롬 이상을 검출하는 데 유용할 것이라고 판단하였다. 일례로, 본 발명자들은 초기 간세포암종(HCC) 환자의 4개의 혈장 DNA 시료, 및 HCC 암을 갖지 않는 67명의 만성 B형 간염(HBV) 보균자(HBV 보균자)에게 크기-밴드 기초 패턴 인지를 적용하였다. 건강한 대조군의 30개의 혈장 DNA 시료를 사용하여, HCC 환자 및 HBV 보균자에서 CNA 및 메틸롬 이상을 지칭(call)하는 데 사용된 복제수 변화의 정상적인 기준 범위를 구축하였다.

[0098] 도 9는 간세포암종(HCC) 환자의 혈장 DNA 내의 측정된 복제수 이상의 크기-밴드 기초의 변화 패턴을 보여준다. 적색 라인은 초기 HCC(eHCC)를 나타내었고, 회색 라인은 HCC가 없는 만성 B형 간염 바이러스(HBV) 보균자를 나타내었다. 본 발명자들은, HCC 암 환자에서 측정된 CNA의 크기-밴드 패턴의 곡선(적색 또는 더 짙은 라인(902, 904, 906 및 908))이 HBV 보균자 환자에 대한 이들 곡선(회색 라인)과 별개였음을 관찰한다. 예를 들어, HCC01 및 HCC03 사례는 각각 13q 및 1p 염색체 아암 상에서 복제 획득(copy gain)을 가졌다.

[0099] HCC01 및 HCC03에서, 본 발명자들은 비-무작위 파형 크기-밴드 기초 패턴을 일관적으로 검출할 수 있었으며, 여기서, 210 bp에서 중간점을 갖는 크기 밴드는 복제수 변화를 보여주는 이의 좌측 및 우측에 비해 전환점인 경향이 있었고, 약 120 bp에서 크기-밴드 패턴은 "벨(bell) 곡선"의 경향을 보여주었다. 14q 결실을 받은 HCC02 사례에 대해, 반전된(inverted) "벨 곡선"이 존재하였다. HCC04 사례의 경우, 본 발명자들이 모든 단편에 대해 z-점수를 사용한다면, 본 발명자들은 3 미만 및 비-암 환자에 대한 z-점수의 범위 내에서 z-점수를 갖는 원형(910)에 의해 제시되는 바와 같이 암을 검출할 수 없었다. 그러나, 본 발명자들이 크기-밴드 기초 접근법을 이용한다면, 본 발명자들은 무작위 크기-밴드 기초 패턴(회색 라인)을 보여주는 비-암 환자로부터 HCC04를 구분할 수 있었다. 대조적으로, 이러한 비-무작위의 별개의 크기-밴드 기초 패턴은 대조군에 존재하지 않았다. 상이한 염색체 아암은 상이한 크기 패턴을 보여준다. 크기 패턴은 염색체 아암에 특이적인 크기 패턴을 참조할 필요가 있을 수 있다.

[0100] **2. 크기-밴드 GR 매트릭스를 이용한 암 분류자**

[0101] 암세포는 일반적으로, 임의의 염색체 아암에서 발생할 복제수 이상을 가지며, 이러한 복제수 이상은 종양 세포가 DNA를 암 환자의 혈액 순환 내로 방출된 경우 혈액 혈장에서 반영될 것이다. 종양-유래 세포-무함유 DNA 분자가 백그라운드 정상 세포-무함유 DNA와 비교하여 별개의 크기 특성을 갖는 것으로 제시되기 때문에(예를 들어, 종양 세포-무함유 DNA 분자는 정상 세포로부터 유래되는 백그라운드 세포-무함유 DNA보다 짧음), 상이한 크기 범위에 걸친 상대 종양 DNA 분율은 다양해질 것이다. 따라서, 암 환자의 혈장에 존재하는 상이한 크기 범위에 걸친 복제수 이상의 측정된 정도는 상이한 크기 범위에 걸친 상대 종양 DNA 분율의 함수일 것이다.

[0102] 본 발명자들은, 상이한 크기 범위에 걸친 측정된 복제수 이상의 상세한 패턴을 포착하는 것이 암 환자와 비-암 환자를 구분하는 데 있어서 성능을 향상시킬 것임을 제안하였다. 패턴은 다수의 영역을 또한 포함할 수 있다.

[0103] 도 10은 본 발명의 구현예에 따라 암 검출을 위한 크기-밴드 게놈 표현(GR) 접근법에 대한 작업흐름을 예시한다. 단계(1010)에서, 본 발명자들은 시퀀싱된 세포-무함유 DNA 단편을 기준 게놈에 맵핑하였다. 단계(1020)에서, 시퀀싱된 단편은 상이한 염색체 아암에 맵핑된다.

[0104] 단계(1030)에서, 시퀀싱된 단편은 상이한 크기 범위(크기 밴드)로 추가로 분류된다. 예를 들어, 크기 범위는 35~75 bp, 40~80 bp, 45~85 bp, 50~90 bp, 55~95 bp, 60~100 bp, 65~105 bp, 70~110 bp, 75~115 bp, 80~120 bp, 85~125 bp, 90~130 bp, 95~135 bp, 100~140 bp, 105~145 bp, 110~150 bp, 115~155 bp, 120~160 bp, 125~165 bp, 130~170 bp, 135~175 bp, 140~180 bp, 145~185 bp, 150~190 bp, 155~195 bp, 160~200 bp, 165~205 bp, 170~210 bp, 175~215 bp, 180~220 bp, 185~225 bp, 190~230 bp, 195~235 bp, 200~240 bp,

205~245 bp, 210~250 bp, 215~255 bp, 220~260 bp, 225~265 bp, 230~270 bp, 235~275 bp, 240~280 bp, 245~285 bp, 250~290 bp, 255~295 bp, 260~300 bp, 265~305 bp, 270~310 bp, 275~315 bp, 280~320 bp, 285~325 bp, 290~330 bp, 295~335 bp, 300~340 bp, 305~345 bp, 310~350 bp, 315~355 bp, 320~360 bp, 325~365 bp, 330~370 bp, 335~375 bp, 340~380 bp, 345~385 bp, 350~390 bp, 355~395 bp, 360~400 bp, 365~405 bp, 370~410 bp, 375~415 bp, 380~420 bp, 및 385~425 bp를 포함할 수 있으나, 이들로 한정되는 것은 아니다. 이러한 크기 범위는 모든 다른 구현예에도 사용될 수 있다.

[0105] 특정 크기 범위 내의 분자 그룹의 경우, 각각의 염색체 아암에 맵핑되는 시퀀싱된 단편의 비율이 계산될 것이며, 본원에서 게놈 표현(GR)으로 지칭된다. GR은 크기 범위 내의 특정 영역(또는 전체 게놈)에 상응하는 모든 DNA 단편의 비율이다. 단계(1030)는 상이한 크기 범위에 대한, 상이한 염색체 아암에 대한, 암을 갖고 있는 것으로 알려진 시료에 대한, 그리고 암을 갖지 않는 것으로 알려진 시료에 대한 GR을 보여준다.

[0106] 일례로서, 각각의 염색체 아암이 71개의 크기 범위를 포함하고 상염색체가 총 39개의 염색체 아암을 갖는다면, 크기 범위 및 염색체 아암은 2,769-차원 벡터를 초래한다. 단계(1040)는 가능한 다차원 벡터를 보여주는 표("크기-밴드 GR 매트릭스")를 보여준다. 제1 열(1042)은 암 시료 1에 상응하고, 71 X N 차원 벡터를 보여주며, 여기서, N은 염색체 아암의 수이다. 표는 암에 대한 M개의 시료 및 비-암에 대한 P개의 시료를 보여준다.

[0107] 단계(1050)에서, 다차원 벡터 및 상기 다차원 벡터로부터 형성된 크기-밴드 GR 매트릭스는 암 분류 모델을 트레이닝하는 데 사용될 수 있다. 기계 학습 알고리즘 또는 심층 학습(deep learning) 알고리즘은 서포트 벡터 머신(SVM), 결정 트리(decision tree), 나이브 베이즈 분류(naive Bayes classification), 로지스틱 회귀(logistic regression), 클러스터링 알고리즘(clustering algorithm), 주성분 분석(principal component analysis; PCA), 특이값 분해(singular value decomposition)(SVD), t-분포 확률적 임베딩(t-distributed stochastic neighbor embedding; tSNE), 및 인공 신경망(artificial neural network)을 비제한적으로 포함하는 암 분류자(classifier), 뿐만 아니라 분류자 세트를 작제한 다음 이들의 예측의 가중 보트(weighted vote)를 취함으로써 새로운 데이터 포인트를 분류하는 앙상블 방법을 트레이닝하는 데 사용될 수 있을 것이다. 일단 암 분류자가 트레이닝되면, 새로운 환자에 대한 암의 확률이 예측될 수 있다.

[0108] 트레이닝 데이터는 암 대상체 및 비-암 대상체를 포함할 수 있다. 세포-무함유 DNA 측정을 모델링하는 기계 학습 알고리즘(크기-밴드 GR, 메틸화 등)은, 암 대상체와 비-암 대상체 사이에서 최상의 분리를 제공하는 분류 경계(예를 들어, 선형 또는 비-선형 수식, 예컨대 로지스틱 회귀 수식에서 구조화된 계수 및 트레이닝된 가중의 세트를 사용함)를 구축하는 데 사용될 수 있다. 암-연관 데이터 포인트로의 최적의 분류화 경계로부터 세포-무함유 DNA 측정을 포함한 새로운 시료의 입력 벡터의 편차는 암일 가능성을 나타낼 것이다. 이러한 편차는 0 내지 1의 척도(scale) 내에서 암의 확률로 정규화되거나 변환될 수 있을 것이다. 확률이 높을수록, 암의 가능성이 더 높다. 소정의 역치(예를 들어, 0.6 초과) 초과의 암의 확률은 암을 갖는 양성 검사로서 여겨질 수 있다.

[0109] 간세포암종에 대해, 1p, 1q, 8p 및 8q는 복제수의 측면에서 보편적으로 이탈적인 것으로 보고되었다(문헌[Proc Natl Acad Sci USA. 2015 Mar 17;112(11):E1317-25]). 따라서, 크기-밴드 암 검출의 성능을 예시하기 위해, 본 발명자들은 많은 건강한 대조군(CTR), HBV 보균자(HBV), 간경변 대상체(간경변), 초기-병기(stage) HCC(eHCC), 중간-병기 HCC(iHCC) 및 진행-병기 HCC(aHCC)를 시퀀싱하기 위해 대규모 병렬 시퀀싱 플랫폼을 사용하였다. 트레이닝 데이터세트의 경우, 본 발명자들은 제한된 수의 진행-병기 HCC 환자를 시퀀싱한 다음, 진행-병기 HCC 환자의 시퀀싱 결과를 비-HCC 대상체의 시퀀싱 결과와 인위적으로 혼합하여, 종양 DNA 분율의 광범위한 범위가 0.01% 내지 50% 범위인 충분한 HCC 양성 환자 및 비-HCC 대상체를 함유하는 트레이닝 데이터세트를 형성하였다. 이를 위해, 사용되는 시퀀싱 판독의 비율을 다양하게 함으로써 34명의 HBV, 10명의 CTR 및 9명의 aHCC 대상체를 무작위로 반복적으로 혼합함으로써 401명의 HCC 환자가 생성되었고, 34명의 HBV, 15명의 간경변 및 10명의 CTR 대상체를 무작위로 반복적으로 혼합함으로써 175명의 비-HCC 환자가 생성되었다. SVM 알고리즘은 이러한 401명의 HCC 환자 및 175명의 비-HCC 환자를 사용하여 암 분류자를 트레이닝하는 데 사용되었다.

[0110] 단계(1060)에서, 트레이닝된 암 분류 모델은 새로운 시료가 암을 갖는지 또는 암을 갖지 않는지 예측하는 데 사용될 수 있다. 암의 확률은 모델에 의해 결정될 수 있으며, 이때 역치 초과의 확률은 암에 대한 양성 검사로서 여겨진다.

[0111] 암을 검출하기 위한 크기-밴드 접근법 및 종래의 z-점수 접근법이 30명의 CTR, 19명의 HBV, 14명의 간경변, 36명의 eHCC, 및 11명의 iHCC 대상체를 포함하는 검사 데이터세트에 적용되었다.

[0112] 도 11a는 암을 검출하기 위한 크기-밴드 접근법의 결과를 보여준다. SVM은 암 분류자를 트레이닝하는 데 사용되

었다. eHCC 대상체와 iHCC 대상체 둘 모두는 0.60 초과의 중앙값의 암 확률을 가졌으며, 이때 iHCC는 eHCC보다 높은 확률을 가졌다. CTR, HBV, 및 간경변 대상체는 0.20 미만의 중앙값 확률을 보여주었다. 암을 검출하기 위한 크기-밴드 접근법은 95%의 특이도에서 64% 민감도를 가졌다. 적색 점선은 95% 특이도에 상응한다.

[0113] 도 11b는 암을 검출하기 위한 종래의 z-점수 접근법의 결과를 보여준다. 적색 점선은 95% 특이도에 상응하고, 이는 약 4.2의 z-점수에 있었다. 염색체 아암 1p, 1q, 8p 및 8q가 예로서 사용되었다. 검사 시료의 각각의 아암에 대한 GR이 계산되었다. 상응하는 평균 및 표준 편차 또한, 계산되었다. 각각의 아암 z-점수는 (GR - 평균)/표준 편차로서 계산될 것이다. 절대 z-점수는 4개의 염색체 아암에 상응하는 4개의 절대 z-점수의 합계와 동일하였다. iHCC 대상체는 CTR, HBV, 간경변 및 eHCC 대상체보다 주목할 만하게 높은 암의 중앙값 절대 z-점수를 가졌다. iHCC에 대한 중앙값 절대 z-점수가 다른 대상체에 대한 절대 z-점수보다 높은 한편, 몇몇 iHCC 대상체의 z-점수는 다른 대상체와 꽤 유사하였다. 그러나, eHCC의 중앙값 절대 z-점수는 CTR, HBV 및 간경변 대상체보다 단지 약간 더 높았고, 3의 z-점수 역치 수준과 대략 동일하였다. 종래의 z-점수 접근법은 95%의 특이도에서 51% 민감도를 가졌다. 따라서, 크기-밴드 접근법은 종래의 z-점수 접근법을 능가하여 우수한 민감도를 보여준다.

[0114] 도 11c는 수신자 조작 특성 곡선(ROC; receiver operating characteristic curve) 분석으로 종래의 z-점수 접근법을 능가하는 크기-밴드 접근법의 우수성을 보여준다(0.84 대 0.82).

[0115] **3. 크기-밴드 게놈 표현(GR) 매트릭스를 이용한 예시적인 방법**

[0116] 도 12는 대상체로부터의 생물학적 시료에서 암 분류를 결정하는 방법(1200)의 일례를 보여준다. 생물학적 시료는 종양 DNA 분자 및 비-종양 DNA 분자를 포함하는 세포-무함유 DNA 분자의 혼합물을 포함할 수 있다.

[0117] 블록(1202)에서, 생물학적 시료로부터의 세포-무함유 DNA 분자의 제1 양이 측정될 수 있다. 세포-무함유 DNA 분자의 제1 양은 M개의 범위에 대한 각각의 크기 범위 및 N개의 게놈 영역에 대한 각각의 게놈 영역에 상응할 수 있다. 복수의 크기 범위는 방법(600) 또는 방법(800)과 함께 기재된 바와 같이 결정될 수 있다. 각각의 게놈 영역은 염색체 아암일 수 있다.

[0118] 블록(1204)에서, 크기 비는 세포-무함유 DNA 분자의 제1 양, 및 크기 범위 내의 것이 아닌 크기를 포함하는 제2 크기 범위 내의 세포-무함유 DNA 분자의 제2 양을 사용하여 계산될 수 있다. 크기 비는 방법(600)에서와 같이 계산될 수 있으나, 크기 비는 특정 게놈 영역(예를 들어, 염색체 아암)에 대한 것일 수 있다. 일례로서, 크기 비는 도 10에서 열(1004)에서 임의의 게놈 표현 GR1, GR2, GR3, ... GR 71일 수 있다. 크기 비의 계산은 NXM개의 크기 비의 측정 벡터를 발생시킬 수 있다. N은 1 초과의 정수일 수 있다. N 및 M은 2, 3, 4, 5 또는 6 초과를 포함하여 1 초과의 정수일 수 있다.

[0119] 블록(1206)에서, 기준 크기 패턴이 수득될 수 있다. 기준 크기 패턴은 N개의 게놈 영역 및 M개의 크기 범위에 대한 기준 크기 비의 기준 벡터를 포함할 수 있다. 기준 크기 패턴은 암을 갖는 대상체로부터의 또는 암을 갖지 않는 대상체로부터의 복수의 기준 시료로부터 결정될 수 있다. 기준 크기 패턴은 기계 학습 모델을 사용하여 결정될 수 있다.

[0120] 기계 학습 모델은 암을 갖는 개체로부터의 복수의 게놈 영역 중 각각의 게놈 영역에서의 크기 비를 포함한 크기 비의 트레이닝 세트를 사용하여 결정될 수 있다. 암 분류자는 기계 학습 알고리즘 또는 심층 학습 알고리즘을 사용하여 결정될 수 있다. 기계 학습 모델 또는 심층 학습 알고리즘은 서포트 벡터 머신(SVM), 결정 트리, 나이브 베이즈 분류, 로지스틱 회귀, 클러스터링 알고리즘, 주성분 분석(PCA), 특이값 분해(SVD), t-분포 확률적 임베딩(tSNE), 인공 신경망 또는 본원에 기재된 임의의 알고리즘을 포함할 수 있다. 트레이닝 세트는 암을 갖는 것으로 결정된 개체 및 암을 갖지 않는 것으로 결정된 개체에 대해 상이한 게놈 영역에서의 크기 비를 포함할 수 있다. 기계 학습 모델은 도 10에서의 암 분류자일 수 있다.

[0121] 블록(1208)에서, 측정 벡터는 기준 벡터와 비교될 수 있다. 비교는 기계 학습 모델을 사용하여 비교될 수 있다. 비교는 기준 벡터에 대한 측정 벡터의 유사성에 기초한 값을 초래할 수 있다.

[0122] 측정 벡터와 기준 벡터의 비교는 NXM개의 크기 비를 복수의 기준 시료로부터 결정된 복수의 역치 값과 비교하는 단계를 포함할 수 있다. 예를 들어, 각각의 크기 범위는 상이한 역치 값을 가질 수 있으며, 이러한 역치 값은 기준 시료에 대한 표준 편차에 기초할 수 있다. 이에, NXM개의 역치 값이 있을 수 있다. 단일 크기 범위는 또한, 상이한 역치 값을 가질 수 있으며, 이때, 각각의 역치 값은 크기 비가 기준 시료로부터 상이한, 상이한 확실성 수준과 연관이 있다. 비교는 초과된 역치 값의 수를 카운팅하는 단계 및 비교에 기초하여 암의 수준을 결정하는 단계를 포함할 수 있다. 초과된 역치 값의 더 높은 수준은 측정 벡터와 기준 벡터 사이에서 더 큰 차

이를 나타낼 수 있다.

[0123] 블록(1210)에서, 암의 수준은 비교에 기초하여 결정될 수 있다. 암은 간세포암종을 포함할 수 있다. 암은 결장직장암, 폐암, 비인두암, 난소암, 위암 및 혈액암을 포함할 수 있다. 방법(1200)은 암 대상체와 비-암 대상체 사이에서 구분을 가능하게 할 수 있다. 대상체는 기준 벡터에 대한 측정 벡터의 유사성에 기초한 값에 기초하여 암을 갖거나 암의 높은 가능성을 갖는 것으로 분류될 수 있다. 유사성에 기초한 값은 컷오프 값과 비교될 수 있다. 컷오프 값을 더 크게 초과하는 유사성에 기초한 값은 암의 더 높은 가능성 또는 중증도를 나타낼 수 있다. 상기 방법은 대상체가 암을 갖거나 암의 높은 가능성을 갖는 것으로 분류되는 경우 암을 치료하는 단계를 추가로 포함할 수 있다.

[0124] 방법(1200)은 암 대신에 자가면역 장애의 수준을 결정하도록 적응될 수 있다. 자가면역 장애는 전신 홍반성 루프스(SLE)를 포함할 수 있다. 크기 DNA 단편은 2014년 9월 19일에 출원된 미국 특허 공보 제2015/0087529 A1호에 기재된 바와 같이 SLE와 관련이 있는 것으로 확인되었고, 이의 내용은 모든 목적을 위해 참조에 의해 본 명세서에 포함된다. 자가면역 장애의 수준은 측정 벡터를 기준 벡터와 비교함으로써 결정될 수 있다. 기준 벡터는 기준 크기 패턴으로부터의 것일 수 있다. 기준 크기 패턴은 건강한 대상체 또는 기지의 수준의 자가면역 장애를 갖는 대상체로부터의 시료로부터 결정될 수 있다. 방법(1200)은 자가면역 장애를 갖는 대상체와 갖지 않는 대상체 사이에서 구분을 가능하게 할 수 있다.

[0125] **4. 크기-밴드 메틸화 밀도(MD) 매트릭스를 이용한 암 분류자**

[0126] 암세포는 일반적으로, 임의의 게놈 영역에서 발생할 특이적인 메틸화 패턴을 갖는다. 예를 들어, 암세포에서, Alu 반복 영역은 비-악성 세포와 비교하여 우선적으로 덜 메틸화될 수 있고, CpG 섬(island) 영역은 비-악성 세포와 비교하여 우선적으로 더 메틸화될 수 있다. 이러한 암-연관된 이탈적인 메틸화 신호는, 종양 세포가 DNA를 혈액 순환 내로 방출한 경우 암 환자의 혈액 혈장에서 반영될 수 있다. 상기 설명된 바와 같이, 상이한 크기 범위에 걸친 상대 종양 DNA 분율은 다양하다. 따라서, 암 환자의 혈장에 존재하는 상이한 크기 범위에 걸친 암-연관된 메틸화 수준의 측정된 정도는 상이한 크기 범위에 걸친 상대 종양 DNA 분율의 함수일 것이다.

[0127] 본 발명자들은, 상이한 크기 범위에 걸친 측정된 메틸화 이상의 상세한 패턴을 포착하는 것은 암 환자와 비-암 환자를 구분하는 데 있어서 성능을 향상시킬 것이라고 제안하였다.

[0128] 도 13은 본 발명의 구현예에 따른 암 검출을 위한 크기-밴드 메틸화 밀도(MD) 접근법에 대한 작업흐름을 예시한다. 단계(1310)에서, 본 발명자들은 Methy-Pipe(문헌[Jiang et al., PLoS One. 2014;9(6):e100360]) 또는 다른 메틸화-인지(aware) 정렬자(aligner)를 사용하여, 시퀀싱된 비설파이트-전환된 세포-무함유 DNA 단편을 기준 게놈에 맵핑하였다. 단계(1320)에서, 상이한 차별적으로 메틸화된 영역에 맵핑된 시퀀싱된 단편이 위치한다.

[0129] 단계(1330)에서, 시퀀싱된 단편은 상이한 크기 범위(크기 밴드)로 추가로 분류된다. 예를 들어, 크기 범위는 도 10에 대해 단계(1030)에 기재된 크기 범위를 포함하여 본원에 기재된 임의의 크기 범위를 포함할 수 있다.

[0130] 특정 크기 범위 내의 분자 그룹에 대해, 관심 영역(예를 들어, Alu 반복 또는 CpG 섬) 상의 시퀀싱된 CpG의 비율이 계산되어, 메틸화 수준을 반영하는 메틸화 밀도(MD)를 초래할 것이다. 영역은 간암 세포와 조혈모 세포(예를 들어, T 세포, B 세포, 호중구, 대식세포, 적혈아구 세포 등), 간세포 및 결장 세포를 포함한 다른 정상 세포 사이에서 상이한 메틸화 수준을 보여줄 수 있다. 단계(1330)는 상이한 크기 범위에 대한, 상이한 게놈 영역에 대한, 암을 갖고 있는 것으로 알려진 시료에 대한, 그리고 암을 갖지 않는 것으로 알려진 시료에 대한 MD를 보여준다.

[0131] 일례로서, 각각의 영역이 71개의 크기 범위를 포함하고 간암 세포와 다른 정상 세포사이에서 차별적인 메틸화를 보여주는 총 32,450개의 영역이 있다면, 크기 범위 및 게놈 영역은 2,303,950-차원 벡터를 초래한다. 단계(1340)는 가능한 다차원 벡터를 보여주는 표("크기-밴드 MD 매트릭스")를 보여준다. 표의 제1 열(1342)은 암 시료 1에 상응하고 71 X N 차원 벡터를 보여주며, 여기서, N은 게놈 영역의 수이다. 표는 암에 대한 M개의 시료 및 비-암에 대한 P개의 시료를 보여준다.

[0132] 단계(1350)에서, 다차원 벡터 및 상기 다차원 벡터로부터 형성된 크기-밴드 MD 매트릭스는 암 분류 모델을 트레이닝하는 데 사용될 수 있다. 트레이닝은 예를 들어, 도 10의 단계(1050)에 대한 것을 포함하여 본원에 기재된 바와 같이 분류를 수행하는 임의의 적합한 기계 학습 모델에 의한 것일 수 있다. 일단 암 분류자가 트레이닝되면, 새로운 환자에 대해 암을 나타내는 시료의 확률이 예측될 수 있다. 소정의 역치(예를 들어, 0.6 초과) 초과의 암의 확률은 암을 갖는 양성 검사로서 여겨질 수 있다.

[0133] 크기-밴드 메틸화 수준을 이용하여 암 검출의 성능을 예시하기 위해, 본 발명자들은 많은 건강한 대조군(CTR), HBV 보균자(HBV), 간경변 대상체(간경변), 초기-병기 HCC(eHCC), 중간-병기 HCC(iHCC) 및 진행-병기 HCC(aHCC)를 시퀀싱하기 위해 대규모 병렬 시퀀싱 플랫폼을 사용하였다. 트레이닝 데이터세트의 경우, 본 발명자들은 제한된 수의 진행-병기 HCC 환자를 시퀀싱한 다음, 진행-병기 HCC 환자의 시퀀싱 결과를 비-HCC 대상체의 시퀀싱 결과와 인위적으로 혼합하여, 종양 DNA 분율의 광범위한 범위가 0.01% 내지 50% 범위인 충분한 HCC 양성 환자 및 비-HCC 대상체를 함유하는 트레이닝 데이터세트를 형성하였다. 이를 위해, 사용되는 시퀀싱 판독의 비율을 다양하게 함으로써 27명의 HBV 및 7명의 aHCC 대상체를 무작위로 반복적으로 혼합함으로써 140명의 HCC 환자가 생성되었고, 7명의 HBV 및 20명의 CTR 대상체를 무작위로 반복적으로 혼합함으로써 140명의 비-HCC 환자가 생성되었다. SVM 알고리즘은 이러한 140명의 HCC 환자 및 140명의 비-HCC 환자를 사용하여 암 분류자를 트레이닝하는 데 사용되었다.

[0134] 단계(1360)에서, 트레이닝된 암 분류 모델은 새로운 시료가 암을 갖는지 또는 암을 갖지 않는지 예측하는 데 사용될 수 있다. 암의 확률은 모델에 의해 결정될 수 있으며, 이때 역치 초과의 확률은 암에 대한 양성 검사로서 여겨진다.

[0135] 도 14a, 14b 및 14c는 본 발명의 구현예에 따라 크기-밴드 MD와 종래의 z-점수 접근법 사이의 비교를 보여준다. 도 14a는 크기-밴드 MD 접근법에 대한 결과를 보여준다. 도 14b는 종래의 z-점수 접근법에 대한 결과를 보여준다.

[0136] 도 14a 및 14b는, 27명의 HBV, 36명의 eHCC 및 11명의 iHCC 대상체를 포함하는 검사 데이터세트에서, 암을 검출하기 위한 크기-밴드 메틸화 접근법이 종래의 z-점수 접근법보다 우수하였음을 보여준다. 종래의 z-점수 접근법은 하기 방식으로 수행되었다: (1) 모든 관심 영역으로부터 유래된 총 단편에 대한 풀링된 메틸화 수준("X"로 표시됨)이 계산되며; (2) 비-암 그룹에서 풀링된 메틸화 수준의 평균(M), 및 풀링된 메틸화 수준의 표준 편차(SD)가 계산되고; (3) 그 후에, 종래의 메틸화 z-점수가 계산된다: z-점수 = (X-M)/SD. SVM은 암 분류자를 트레이닝하는 데 사용되었다. 도 14a에서 크기-밴드 메틸화 접근법은 92.5%의 특이도에서 74.5% 민감도를 가졌다. 대조적으로, 도 14b에서 종래의 z-점수 접근법은 92.5%의 특이도에서 더 낮은 민감도, 65.9% 민감도를 가졌다. 증가된 민감도는 중요한 이익을 야기할 수 있다. 초기 암의 조기 검출은 일반적으로, 더 양호한 치료 결과와 연관이 있다. eHCC 그룹과 iHCC 그룹 둘 모두는 치료 가능한 병기인 것으로 여겨진다. 따라서, 치료 가능한 사례에서 민감도의 임의의 증가는 임상적 영향을 미치며, 환자에 대한 매우 상이한 생존율 프로파일로 해석될 수 있다.

[0137] 도 14c는 수신자 조작 특성 곡선(ROC) 분석에서 크기-밴드 메틸화 접근법의 우수성을 보여준다(SVM: 0.89 AUC 대 z-점수: 0.87 AUC).

[0138] 이에, 게놈 표현(GR)과 함께 다차원 벡터의 사용(예를 들어, 도 10 내지 도 12)은 GR 대신에 메틸화 밀도를 사용하는 분석을 위해 적응될 수 있다.

[0139] *F. 부가적인 크기 패턴 적용*

[0140] 크기-밴드 기초 패턴은 혈장 DNA에서 나타난 이상에 대한 기원을 알려줄 것이다. 일례로서, 임산부 맥락에서, 복제수 이상이 모체로부터 유래된다면, 모체 DNA 단편이 태아 DNA보다 길기 때문에 크기-밴드 패턴은 태아로부터 기원하는 것과 비교하여 역방향으로 발생할 것이다(문헌[Yu et al. Clin. Chem. 2017; 63:495-502]). 크기-밴드 기초 분자 진단은 또한, 점 돌연변이, 하위-염색체 이상 및 후성유전체적 비정상의 검출을 증강시키는 것을 포함하여, 다른 임상 질환, 예컨대 암(문헌[Jiang et al. Proc. Natl. Acad. Sci. 2015; 112:E1317-E1325])에서 세포-무함유 DNA의 분석에 적용될 수 있을 것이다. 임상 질환은 이식된 조직 또는 기관에 대한 면역-반응의 존재를 결정하는 단계를 포함할 수 있다.

[0141] 이외에도, 이는 SLE 환자(문헌[Chan et al. Proc. Natl. Acad. Sci. 2014; 111:E5302-E5311])의 혈장 DNA에 존재하는 분명한 복제수 변화가 참(true) 복제수 변화, 특히 세포보다 특정 DNA 서열에의 항-DNA 항체의 우선적인 결합으로 인한 것일 가능성이 있기 때문에, 전신 홍반성 루프스(SLE)와 같은 혈장 DNA에 존재하는 이상과 혼동하는(confounding) 혈장 DNA를 구분할 수 있게 할 것이다. 따라서, 크기-밴드 기초 분석은 SLE 환자의 혈장에 존재하는 측정된 복제수 이상에 대한 상이한 크기 밴드에 관하여 무작위한 모양 변화를 볼 것으로 예상될 것이다.

[0142] 구현예는 환자에서 질병 또는 질환의 수준 또는 확률을 결정한 후 상기 환자에서 질병 또는 질환을 치료하는 단계를 포함할 수 있다. 치료는 본원에서 언급된 참조문헌에 기재된 임의의 치료를 포함하여 임의의 적합한 치료

법, 약물 또는 수술을 포함할 수 있다. 참조문헌에서의 치료에 대한 정보는 참조에 의해 본 명세서에 포함된다.

[0143] **III. 재료 및 방법**

[0144] **시료 수합 및 처리**

[0145] 이러한 소급적 연구를 위해 분석된 익명의 데이터를 홍콩 중문 대학교(Chinese University of Hong Kong)의 병리학 서비스 대학교(UPS; University Pathology Service)에서 기존 환자 데이터로부터 수득하였다. 161개 시료로 구성된 환자 데이터를 UPS 실험실-개발 검사의 결과로서 발생시켰다. 종양 절개를 위해 홍콩 소재의 프린스 웨일즈 병원의 외과에 입원한 HCC를 갖는 익명의 환자를 모집하였다. 수술 전에 모든 혈액을 채혈하였다. 익명의 HBV 보균자 및 간경변 대상체를 홍콩 소재의 프린스 웨일즈 병원의 의학 및 치료과로부터 모집하였다. 혈액을 원심분리하여 혈장을 수득함으로써 시료를 수득하였다. 간략하게는, 말초 혈액 시료를 EDTA-함유 튜브에 수합하고, 상기 튜브를 후속해서 1,600 g, 4℃에서 10분 동안 원심분리하였다. 혈장 부분을 16,000 g, 4℃에서 10분 동안 재원심분리하여 세포-무함유 혈장을 수득하고, 추가의 분석 때까지 -80℃에서 보관하였다. QIAamp DSP DNA 혈액 미니 키트(Qiagen)를 사용하여 4 내지 10 mL의 혈장으로부터 DNA를 추출하였다. 혈장 DNA를 SpeedVac 농축기(Savant DNA120; Thermo Scientific)를 이용하여 1개 시료 당 75-μL 최종 부피로 농축시켰다.

[0146] **시퀀싱 라이브러리 제조 및 DNA 시퀀싱**

[0147] 추출된 혈장 DNA를 사용하여, 짝형성-말단 시퀀싱 시료 제조 키트를 제조업체의 설명에 따라 이용하여 인덱싱된 DNA 라이브러리를 작제하였다. 이 단계에서, 혈장 이중-가닥 DNA 분자는 말단이 다시 짝형성되어 평활 말단을 형성할 것이고, 동시에 잉여(extra) A 염기에 첨가되었다. PCR 증폭을 보조하며 플로우셀(flowcell)에 어닐링되고 시퀀싱을 용이하게 할 수 있는 어댑터를 A-태깅된 이중-가닥 혈장 DNA 분자에 결찰시켜, 시퀀싱 라이브러리를 형성하였다. 라이브러리는 이전에 기재된 바와 같이 각각의 말단에 대해 36, 50 또는 75 사이클의 사용으로 짝형성-말단 모드에서 시퀀싱될 수 있다(문헌[Yu *et al. Proc. Natl. Acad. Sci. U. S. A.* 2014; 111:8583-8]).

[0148] **서열 정렬**

[0149] 각각의 시료로부터의 서열을 이전에 기재된 바와 같이(문헌[Yu *et al. Proc. Natl. Acad. Sci. U. S. A.* 2014; 111:8583-8]) 짧은 올리고뉴클레오타이드 정렬 프로그램 2(SOAP2)(문헌[Li *et al. Bioinformatics* 2009; 25:1966-1967])를 사용하여 인간 기준 게놈(hg19)에 정렬하였다. 평균적으로, 각각의 시료는 1,200만개의 고유하게 맵핑된 짝형성-말단 판독(범위: 1,000만 내지 1,500만개)을 수득하였다.

[0150] **메틸화 수준**

[0151] 서열 판독의 부위의 메틸화 상태는 본원에 기재된 바와 같이 수득될 수 있다. 예를 들어, DNA 분자는 상기 DNA 분자의 시퀀스 판독을 사용하여 분석될 수 있으며, 이때 시퀀싱은 메틸화-인식이다. 예를 들어, 메틸화-인지 시퀀싱은 비제한적으로, 비설파이트 시퀀싱, 메틸화-민감성 제한 효소 분해에 뒤이은 시퀀싱, 항-메틸시토신 항체 또는 메틸화 결합 단백질을 사용한 면역침전, 또는 메틸화 상태의 명시(elucidation)를 가능하게 하는 단일 분자 시퀀싱을 포함할 수 있다. 다른 메틸화-인식 검정법이 또한 사용될 수 있다.

[0152] 시퀀스 판독은 각각 생물학적 시료로부터의 세포-무함유 DNA 분자의 메틸화 상태를 포함할 수 있다. 메틸화 상태는, 특정 시토신 잔기가 5-메틸시토신 또는 5-하이드록시메틸시토신인지의 여부를 포함할 수 있다. 시퀀스 판독은 다양한 방식으로, 각각 다양한 시퀀싱 기술, PCR 기술(예를 들어, 실시간 또는 디지털), 어레이, 및 단편의 서열을 식별하기 위한 다른 적합한 기술로 수득될 수 있다. 실시간 PCR은 DNA 그룹을 총괄적으로, 예를 들어, 부위에서 메틸화된 DNA의 수에 비례하는 강도 신호로서 분석하는 일례이다. 서열 판독은 2개의 부위의 근접성 및 상기 서열 판독의 길이에 따라 1개 초과의 부위를 망라할 수 있다.

[0153] 분석은, 메틸화-인식 시퀀싱으로부터 시퀀스 판독을 수신함으로써 수행될 수 있고, 따라서, 상기 분석은 DNA로부터 이전에 수득된 데이터 상에서 수행될 수 있다. 다른 구현예에서, 분석은 실제 시퀀싱, 또는 DNA 분자의 특성의 측정을 수행하는 다른 능동적인 단계를 포함할 수 있다. 시퀀싱은 여러 가지 방식으로, 예를 들어, 대규모 병렬 시퀀싱 또는 차세대 시퀀싱을 사용하여, 단일 분자 시퀀싱을 사용하여, 및/또는 이중 가닥 또는 단일 가닥 DNA 시퀀싱 라이브러리 제조 프로토콜 및 본원에 기재된 다른 기술을 사용하여 수행될 수 있다. 시퀀싱의 부분으로서, 시퀀스 판독 중 일부는 세포 핵산에 상응할 수 있는 것이 가능하다.

[0154] 시퀀싱은 예를 들어, 본원에 기재된 바와 같이 표적화된 시퀀싱일 수 있다. 예를 들어, 생물학적 시료는 바이러스로부터의 핵산 분자에 대해 농화될 수 있다. 바이러스로부터의 핵산 분자에 대한 생물학적 시료의 농화는, 상

기 바이러스의 일부 또는 전체 게놈에 결합하는 포착 프로브를 사용하는 단계를 포함할 수 있다. 다른 구현예는 바이러스의 특정 좌위에 특이적인 프라이머를 사용할 수 있다. 생물학적 시료는 인간 게놈의 일부, 예를 들어, 상염색체의 영역으로부터의 핵산 분자에 대해 농화될 수 있다. 도 1은 이러한 포착 프로브의 예를 제공한다. 다른 구현예에서, 시퀀싱은 무작위 시퀀싱을 포함할 수 있다.

[0155] 시퀀싱 장치에 의한 시퀀싱 후, 시퀀스 판독은 컴퓨터 시스템에 의해 수신될 수 있으며, 상기 시스템은 시퀀싱을 수행하는 시퀀싱 장치에 예를 들어, 유선 또는 무선 통신을 통해 또는 탈착 가능한 메모리 장치를 통해 통신 가능하게 커플링될 수 있다. 일부 구현예에서, 핵산 단편의 2개 말단 모두를 포함하는 하나 이상의 시퀀스 판독이 수신될 수 있다. DNA 분자의 장소는, DNA 분자의 하나 이상의 시퀀스 판독을 인간 게놈의 각각의 부분, 예를 들어, 특이적인 영역, 예컨대 차별적으로 메틸화된 영역(DMR)에 맵핑(정렬)함으로써 결정될 수 있다. 일 실시에서, 판독이 관심 영역에 맵핑되지 않는다면, 상기 판독은 무시될 수 있다. 다른 구현예에서, 특정 프로브(예를 들어, PCR 또는 다른 증폭 후)는 예컨대 특정 형광 색상을 통해 장소를 나타낼 수 있다. 식별은, 하나 이상의 부위에서 메틸화된 DNA의 양이 모두 필요하기 때문에 세포-무함유 DNA 분자가 하나 이상의 부위의 세트 중 하나에 상응하며, 즉, 특정 부위가 미지일 수 있다는 것일 수 있다.

[0156] 따라서, 시퀀싱 및 정렬 후, 개별적인 CpG 부위의 메틸화 상태는 CpG 맥락에서 시토신 잔기에서 메틸화된 서열 판독 "M"(메틸화된)의 카운트 및 비메틸화된 서열 판독 "U"(비메틸화된)의 카운트로부터 추론될 수 있을 것이다. 비설파이트 시퀀싱 데이터를 사용하여, 모체 혈액, 태반 및 모체 혈장의 전체 메틸롬을 작제하였다. 모체 혈장에서 특이적인 좌위의 평균 메틸화된 CpG 밀도(메틸화 밀도 *MD*라고도 함)는 방정식을 사용하여 계산될 수 있으며:

[0157] $$MD = \frac{M}{M+U}$$

[0158] 여기서, *M*은 유전자 좌위 내의 CpG 부위에서 메틸화된 판독의 카운트이고, U는 비메틸화된 판독의 카운트이다. 좌위 내에 1개 초과의 CpG 부위가 존재한다면, M 및 U는 상기 부위에 걸친 카운트에 상응한다.

[0159] 대안으로서, 메틸화 검정법은 Infinium HD 메틸화 검정법 프로토콜에 따라 비설파이트-전환된 게놈 DNA 상에서 수행될 수 있다. 혼성화된 비드칩(beadchip)은 Illumina iScan 장비 상에서 스캔될 수 있다. DNA 메틸화 데이터를 Genometudio(v2011.1) 메틸화 모듈(v1.9.0) 소프트웨어에 의해 분석하였으며, 이때, 내부 대조군 및 백그라운드 차감(subtraction)으로 정규화하였다. 개별적인 CpG 부위에 대한 메틸화 지수는 베타 값(β)으로 표시될 수 있으며, 이는 메틸화된 대립유전자와 비메틸화된 대립유전자 사이에서 형광 강도의 비를 사용하여 계산될 수 있다:

[0160] $$\beta = \frac{\text{메틸화된 대립유전자의 강도}}{\text{비메틸화된 대립유전자의 강도} + 100}$$

[0161] **태아 DNA 분율의 계산**

[0162] 남자 태아를 임신한 임산부에서, 모체 혈장 시료 내 태아 DNA 분율(*f*)은 염색체 Y에 정렬된 판독의 비율(%chrY)로부터 결정될 수 있다. 이전의 연구에서, 여자 태아를 임신한 임산부의 혈장 내 적은 수의 서열은 염색체 Y에 잘못 정렬된 것으로 제시되었다(문헌[Chiu *et al. Proc Natl Acad Sci U S A* 2008; 105:20458-20463]). 따라서, 남자 태아를 임신한 임산부의 혈장에서 %chrY는 남자 태아로부터 유래된 염색체 Y 판독과 염색체 Y에 잘못정렬된(misaligned) 모체 판독의 혼합이었다(문헌[Chiu *et al. BMJ* 2011; 342:c7401]). 남자 태아를 임신한 임산부에서 %chrY와 *f* 사이의 관계는 하기 방정식을 사용하여 표현될 수 있으며:

[0163] $$\%\text{chrY} = \%\text{chrY}_{\text{남자}} \times f - \%\text{chrY}_{\text{여자}} \times (1-f),$$

[0164] 여기서, $\%\text{chrY}_{\text{남자}}$는 100% 남자 DNA를 함유하는 혈장 시료 내 정렬된 염색체 Y 판독의 비율이고, $\%\text{chrY}_{\text{여자}}$는 100% 여자 DNA를 함유하는 혈장 시료 내 염색체 Y에 정렬된 판독의 비율이다.

[0165] 특정 구현예의 구체적인 세부사항은 본 발명의 구현예의 사상 및 범주 내에서 임의의 적합한 방식으로 조합될 수 있다. 그러나, 본 발명의 다른 구현예는 각각의 개별 양태에 관한 구체적인 구현예, 또는 이들 개별 양태의 구체적인 조합에 관한 것일 수 있다.

[0166] **IV. 예시적인 시스템**

[0167] 도 15는 본 발명의 구현예에 따른 시스템(1500)을 예시한다. 도시된 바와 같은 시스템은 시료(1505), 예컨대 시료 홀더(1510) 내의 세포-무함유 DNA 분자를 포함하며, 상기 시료(1505)는 검정물(1508)과 접촉되어, 물리적 특징(1515)의 신호를 제공할 수 있다. 시료 홀더의 일례는, 검정물의 프로브 및/또는 프라이머 또는 튜브를 포함하는 유동 세포일 수 있으며, 상기 튜브를 통해 액적이 이동한다(이때 상기 액적은 검정물을 포함함). 시료로부터의 물리적 특징(1515), 예컨대 형광 강도 값은 검출기(1520)에 의해 검출된다. 검출기는 간격(예를 들어, 주기적 간격)을 둔 측정을 행하여, 데이터 신호를 이루는 데이터 포인트를 수득할 수 있다. 일 구현예에서, 아날로그-디지털 전환기는 검출기로부터의 아날로그 신호를 복수의 시점에서 디지털 형태로 전환시킨다. 시료 홀더(1510) 및 검출기(1520)는 검정 장치, 예를 들어, 본원에 기재된 구현예에 따른, 시퀀싱을 수행하는 시퀀싱 장치를 형성할 수 있다. 디지털 신호(1525)는 검출기(1520)로부터 로직 시스템(1530)으로 전송된다. 디지털 신호(1525)는 로컬 메모리(1535), 외부 메모리(1540) 또는 저장 장치(1545)에 저장될 수 있다.

[0168] 로직 시스템(1530)은 컴퓨터 시스템, ASIC, 마이크로프로세서 장치 등일 수 있거나 이들을 포함할 수 있다. 상기 시스템은 또한, 디스플레이(예를 들어, 모니터, LED 디스플레이 등) 및 사용자 입력 장치(예를 들어, 마우스, 키보드, 버튼 등)를 포함하거나 이들과 커플링될 수 있다. 로직 시스템(1530) 및 다른 구성요소는 독립형 또는 네트워크 연결 컴퓨터 시스템의 일부일 수 있거나, 이들은 검출기(1520) 및/또는 시료 홀더(1510)를 포함하는 장치(예를 들어, 시퀀싱 장치)에 직접적으로 부착되거나 상기 장치에 통합될 수 있다. 로직 시스템(1530)은 또한, 프로세서(1550)에서 실행하는 소프트웨어를 포함할 수 있다. 로직 시스템(1530)은 임의의 본원에 기재된 방법을 수행하도록 시스템(1500)을 제어하는 명령을 저장하는 컴퓨터 판독 가능 매체를 포함할 수 있다. 예를 들어, 로직 시스템(1530)은, 시퀀싱 또는 다른 물리적 작동이 수행되도록, 시료 홀더(1510)를 포함하는 시스템에 명령을 제공할 수 있다. 이러한 물리적 작동은 특정 순서로 수행될 수 있으며, 예를 들어, 시약이 특정 순서로 첨가되고 제거된다. 이러한 물리적 작동은 시료를 수득하고 검정을 수행하는 데 사용될 수 있는 바와 같이 예를 들어, 로보틱 아암을 포함하는 로보틱스 시스템에 의해 수행될 수 있다.

[0169] 본 명세서에 언급된 컴퓨터 시스템 중 임의의 것이 임의의 적합한 수의 하위시스템을 이용할 수 있다. 이러한 하위시스템의 예는 도 16에서 컴퓨터 장치(1600)에서 도시된다. 일부 구현예에서, 컴퓨터 시스템은 단일 컴퓨터 장치를 포함하며, 여기서 하위시스템은 컴퓨터 장치의 구성요소일 수 있다. 다른 구현예에서, 컴퓨터 시스템은 내부 구성요소와 함께, 각각 하위시스템인, 다수의 컴퓨터 장치를 포함할 수 있다.

[0170] 도 16에 도시된 하위시스템은 시스템 버스(시스템 버스)(1675)를 통해서 서로 연결되어 있다. 추가의 하위시스템, 예컨대 프린터(1674), 키보드(1678), 고정 디스크(fixed disk)(1679), 디스플레이 어댑터(1682)에 커플링되는 모니터(1676), 및 다른 것들이 도시된다. I/O 컨트롤러(controller)(1671)에 커플링되는 주변 장치 및 입력/출력(I/O) 장치는 당업계에 알려진 임의의 수의 수단, 예컨대 시리얼 포트(serial port)(1677)에 의해 컴퓨터 시스템에 연결될 수 있다. 예를 들어, 시리얼 포트(1677) 또는 외부 인터페이스(external interface)(1681)(예를 들어, 이더넷(Ethernet), Wi-Fi 등)는 컴퓨터 장비(1600)를 광역 네트워크, 예컨대 인터넷, 마우스 입력 장치, 또는 스캐너에 연결하는 데 사용될 수 있다. 시스템 버스(1675)를 통한 상호연결은 중앙 프로세서(1673)가 각각의 하위시스템과 통신하고 시스템 메모리(1672) 또는 고정 디스크(1679)로부터의 명령의 실행, 뿐만 아니라 하위시스템 사이의 정보 교환을 제어할 수 있게 한다. 시스템 메모리(1672) 및/또는 고정 디스크(1679)는 컴퓨터 판독 가능 매체를 구현할 수 있다. 본 명세서에 언급된 값 중 임의의 것이 하나의 구성요소로부터 또 다른 구성요소로 출력될 수 있고, 사용자에게 출력될 수 있다.

[0171] 컴퓨터 시스템은 예를 들어, 외부 인터페이스(1681)에 의해 또는 내부 인터페이스에 의해 함께 연결된 복수의 동일한 구성요소 또는 하위시스템을 포함할 수 있다. 일부 구현예에서, 컴퓨터 시스템, 하위시스템 또는 장치는 네트워크를 통해 통신할 수 있다. 이러한 예에서, 하나의 컴퓨터는 클라이언트로 간주될 수 있고 또 다른 컴퓨터는 서버로 간주될 수 있으며, 여기서 각각은 동일한 컴퓨터 시스템의 부분일 수 있다. 클라이언트 및 서버는 각각 다중 시스템, 하위시스템 또는 구성요소를 포함할 수 있다.

[0172] 본 발명의 임의의 구현예는 소프트웨어(예를 들어, 주문형 반도체(application specific integrated circuit) 또는 필드 프로그래머블 게이트 어레이(field programmable gate array))를 사용하고/하거나 모듈러 또는 통합 방식으로 일반적으로 프로그래밍 가능한 프로세서와 함께 컴퓨터 소프트웨어를 사용하는 컨트롤 로직 형태로 실시될 수 있음을 이해해야 한다. 본 명세서에 제공된 개시내용 및 교시를 기초로, 당업자는 하드웨어 및 하드웨어와 소프트웨어의 조합을 사용하여 본 발명의 구현예를 구현하는 다른 방식 및/또는 방법을 알고 인지할 것이다.

[0173] 본 출원에 기재된 임의의 소프트웨어 구성요소 또는 기능은 예를 들어, 종래의 또는 개체-지향 기법을 사용하는

예를 들어, Java, C++, Python 또는 Perl과 같은 임의의 적합한 컴퓨터 언어를 사용하여 프로세서에 의해 실행될 소프트웨어 코드로서 실시될 수 있다. 소프트웨어 코드는 저장 및/또는 전송을 위해 컴퓨터 판독 가능 매체 상에 일련의 명령 또는 지령로서 저장될 수 있고, 적합한 매체는 랜덤 액세스 메모리(RAM), 판독 전용 메모리(ROM), 자기 매체, 예컨대 하드-드라이브 또는 플로피 디스크, 또는 광학 매체, 예컨대 컴팩트 디스크(CD) 또는 DVD(디지털 다목적 디스크), 플래시 메모리 등을 포함한다. 컴퓨터 판독 가능 매체는 이러한 저장 또는 전송 장치의 임의의 조합일 수 있다.

[0174] 이러한 프로그램은 또한 인터넷을 비롯한, 각종 프로토콜에 따른 유선, 광학 및/또는 무선 네트워크를 통한 전송을 위해 채택된 캐리어 신호를 사용하여 인코딩되고 전송될 수 있다. 이와 같이, 본 발명의 일 구현예에 따른 컴퓨터 판독 가능 매체는 이러한 프로그램으로 인코딩되는 데이터 신호를 사용하여 생성될 수 있다. 프로그램 코드로 인코딩된 컴퓨터 판독 가능 매체는 호환 장치와 함께 패키징될 수 있거나 (예를 들어, 인터넷 다운로드를 통해) 다른 장치와 별도로 제공될 수 있다. 임의의 이러한 컴퓨터 판독 가능 매체는 단일 컴퓨터 프로그램 제품(예를 들어, 하드 드라이브, CD, 또는 전체 컴퓨터 시스템) 상에 또는 그 내에 상주할 수 있고, 시스템 또는 네트워크 내의 상이한 컴퓨터 프로그램 제품 상에 또는 그 내에 존재할 수 있다. 컴퓨터 시스템은 모니터, 프린터, 또는 본 명세서에 언급된 결과 중 임의의 것을 사용자에게 제공하기에 적합한 다른 디스플레이를 포함할 수 있다.

[0175] 본 명세서에 기재된 방법 중 임의의 방법은, 단계를 수행하도록 구성될 수 있는, 하나 이상의 프로세서를 포함하는 컴퓨터 시스템에 의해 전체적으로 또는 부분적으로 수행될 수 있다. 따라서 실시예는, 잠재적으로는 각각의 단계 또는 각각의 단계의 그룹을 수행하는 상이한 구성요소와 함께, 본 명세서에 기재된 방법 중 임의의 방법의 단계를 수행하도록 구성된 컴퓨터 시스템에 관한 것일 수 있다. 넘버링된 단계로서 제시되더라도, 본원의 방법의 단계는 동시에 또는 상이한 순서로 수행될 수 있다. 또한, 이들 단계의 일부는 다른 방법으로부터의 다른 단계의 일부와 함께 사용될 수 있다. 또한, 단계의 전부 또는 일부는 선택적일 수 있다. 부가적으로, 임의의 방법의 임의의 단계는 이들 단계를 수행하기 위한 모듈, 회로 또는 다른 수단으로 수행될 수 있다.

[0176] 본 발명의 예시적인 구현예의 상기 설명은 예시 및 설명의 목적으로 제공되었다. 본 발명을 설명된 정확한 형태로 제한하거나 이에 철저하게 지키도록 하는 것을 의도하지는 않으며, 많은 수정 및 변경이 상기 교시를 감안하여 가능하다.

[0177] 전술한 상세한 설명에서, 설명을 위해, 많은 세부사항이 본 기술의 다양한 구현예의 이해를 제공하도록 제시되었다. 그러나, 당업자는 소정의 구현예가 이들 세부사항 중 일부 없이 또는 부가적인 세부사항과 함께 실행될 수 있음을 알 것이다.

[0178] 몇몇 구현예를 설명하였지만, 당업자는 다양한 변형, 대안적인 구성 및 균등물이 본 발명의 사상으로부터 벗어나지 않으면서 사용될 수 있음을 인지할 것이다. 부가적으로, 본 발명을 불필요하게 모호하게 하는 것을 피하기 위해 많은 잘 알려진 과정 및 요소는 기재되지 않았다. 부가적으로, 임의의 구체적인 구현예의 세부사항이 항상 해당 구현예의 변형으로 존재하지 않을 수 있거나, 다른 구현예에 부가될 수 있다.

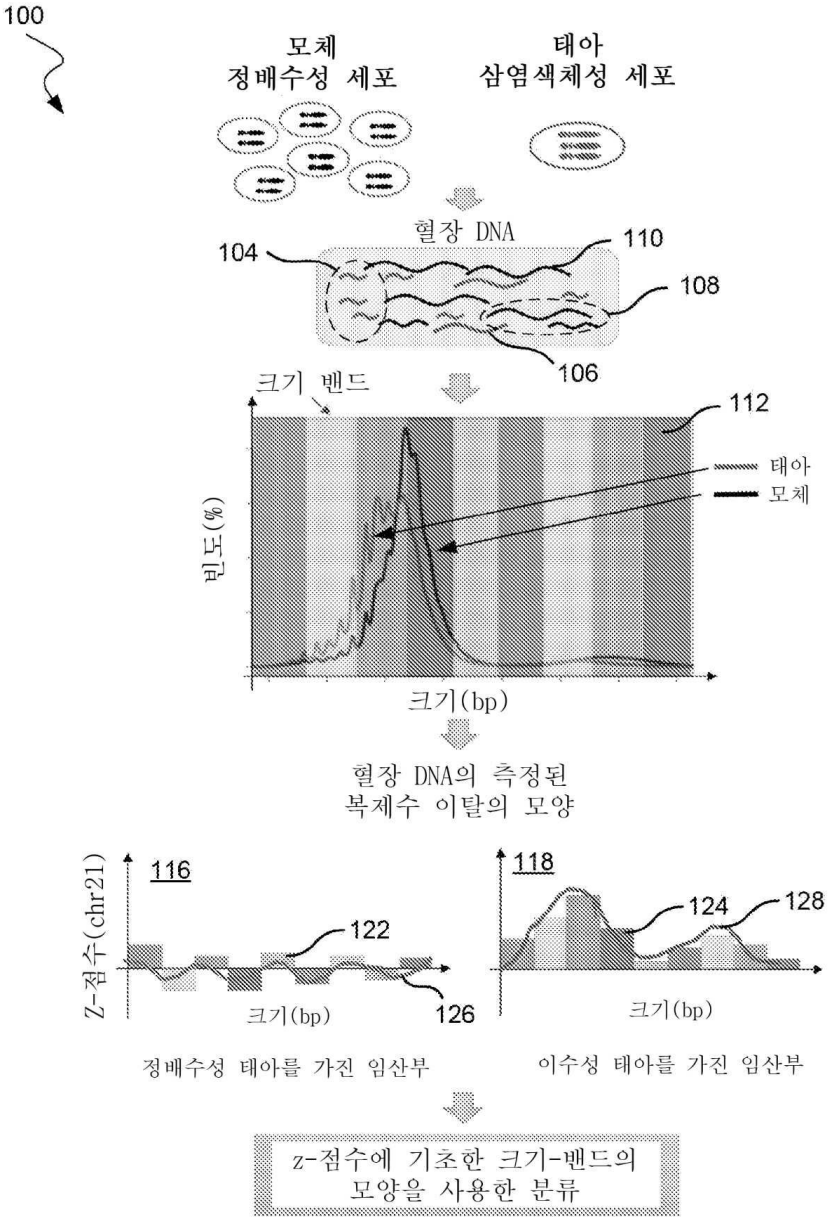
[0179] 값의 범위가 제공되는 경우, 문맥상 명백하게 다르게 나타내지 않는 한 해당 범위의 상한과 하한 사이에서 하한의 단위의 1/10까지 각각의 사이값(intervening value)이 또한 구체적으로 개시되는 것으로 이해된다. 언급된 범위 내의 임의의 언급된 값 또는 사이값과 해당 언급된 범위 내의 임의의 다른 언급된 값 또는 사이값 사이의 각각의 더 작은 범위가 포괄된다. 이들 더 작은 범위의 상한 및 하한은 독립적으로 상기 범위에서 포함되거나 배제될 수 있고, 두 한계 중 어느 하나 또는 둘 중 어느 것도 또는 둘 모두가 더 작은 범위에 포함되는 각각의 범위가 또한, 언급된 범위에서 임의의 구체적으로 배제된 한계를 받는 본 발명 내에 포괄된다. 언급된 범위가 한계 중 하나 또는 둘 모두를 포함하는 경우, 이들 포함된 한계 중 하나 또는 둘 모두를 배제하는 범위가 또한 포함된다.

[0180] 본원에서 그리고 첨부된 청구항에서 사용된 바와 같이, 단수형("a", "an", 및 "the")은 문맥상 명확하게 다르게 나타내지 않는 한 복수형을 포함한다. 따라서, 예를 들어, "일 방법"에 대한 지칭은 복수의 이러한 방법을 포함하고, "입자"에 대한 지칭은 하나 이상의 입자 및 당업자에게 알려진 이의 등가물에 대한 지칭 등을 포함한다. 지금까지, 본 발명은 명료성 및 이해를 위해 상세하게 기재되었다. 그러나, 소정의 변화 및 변형은 첨부된 청구항의 범위 내에서 실행될 수 있는 것으로 이해될 것이다.

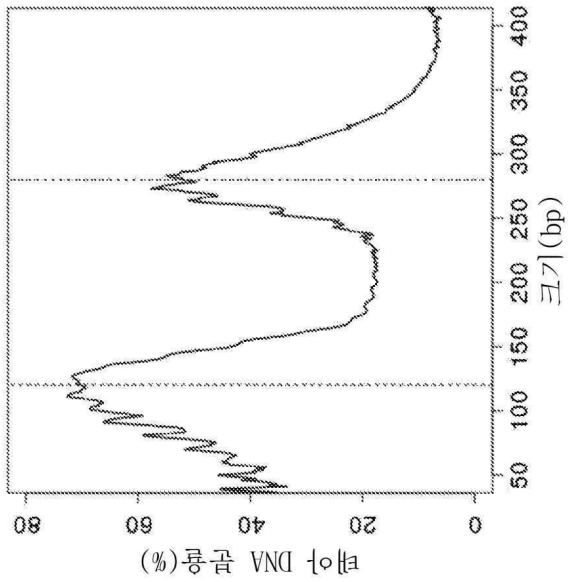
[0181] 본원에서 인용된 모든 공보, 특허 및 특허 출원은 이들 전문이 모든 목적을 위해 참조에 의해 본 명세서에 포함된다. 어느 것도 선행 기술로서 인정하는 것은 아니다.

도면

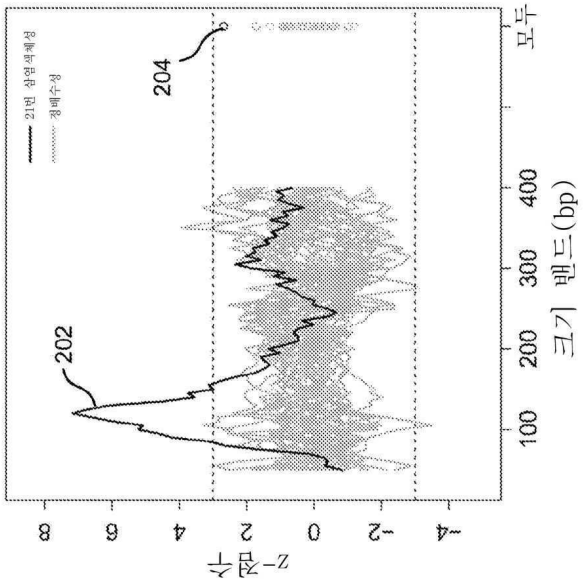
도면1


100
모체
정배수성 세포
태아
삼염색체성 세포
혈장 DNA
104
110
108
106
크기 밴드
112
태아
모체
빈도(%)
크기(bp)
혈장 DNA의 측정된
복제수 이탈의 모양
116
122
126
Z-점수(chr21)
크기(bp)
정배수성 태아를 가진 임산부
118
124
128
크기(bp)
이수성 태아를 가진 임산부
z-점수에 기초한 크기-밴드의
모양을 사용한 분류

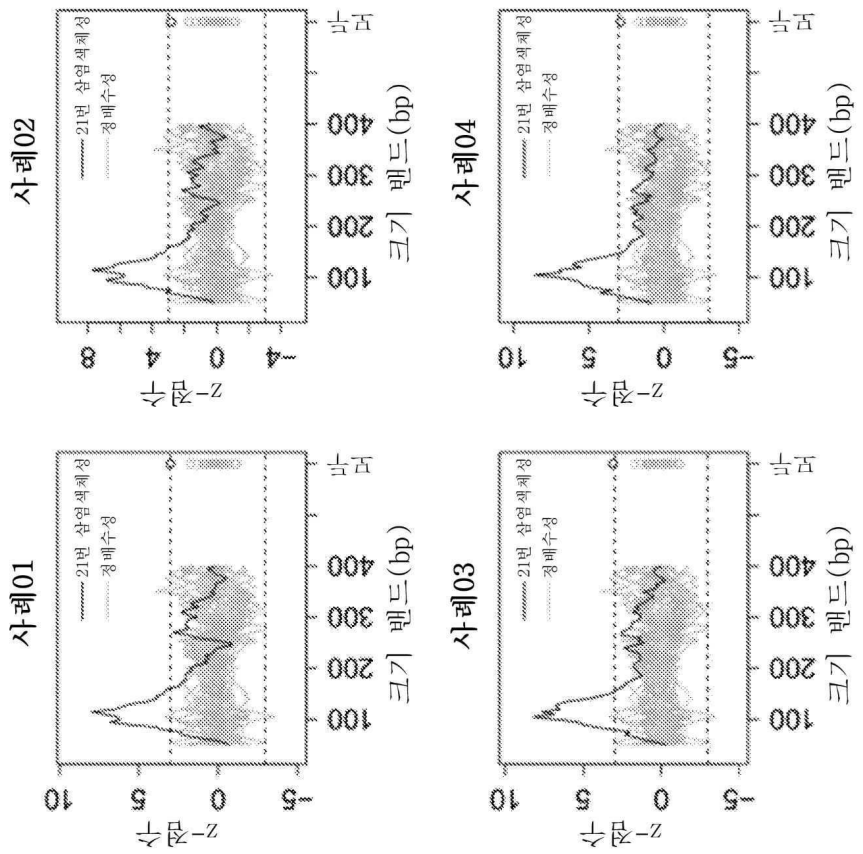
도면2a


테아 DNA 분율(%)
0
20
40
60
80
50
100
150
200
250
300
350
400
크기(bp)

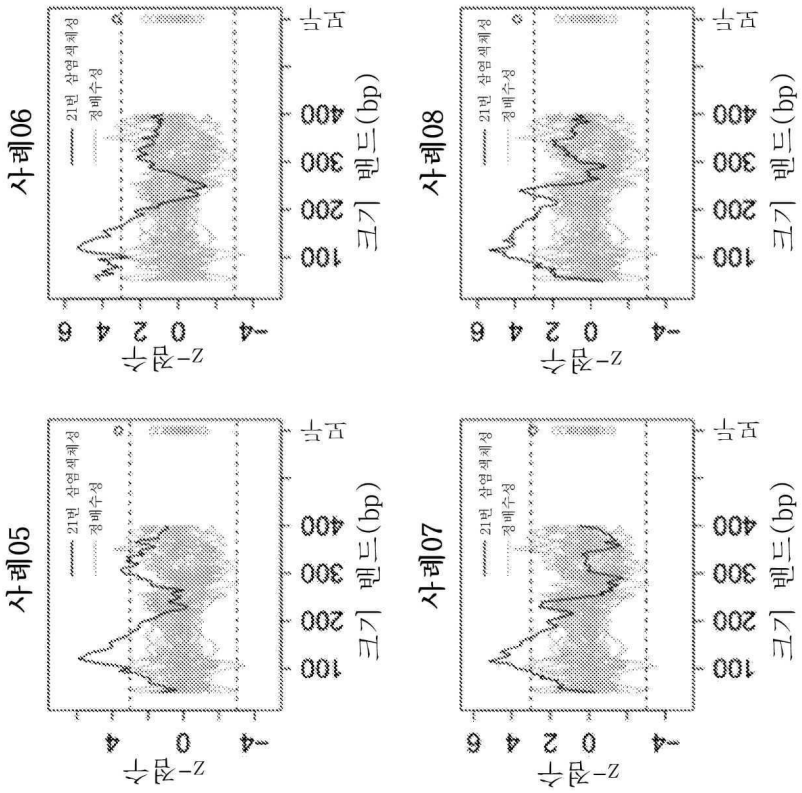
도면2b


Z-점수
-4
-2
0
2
4
6
8
100
200
300
400
모두
크기 밴드(bp)
202
204
21번 삼염색체성
정배수성

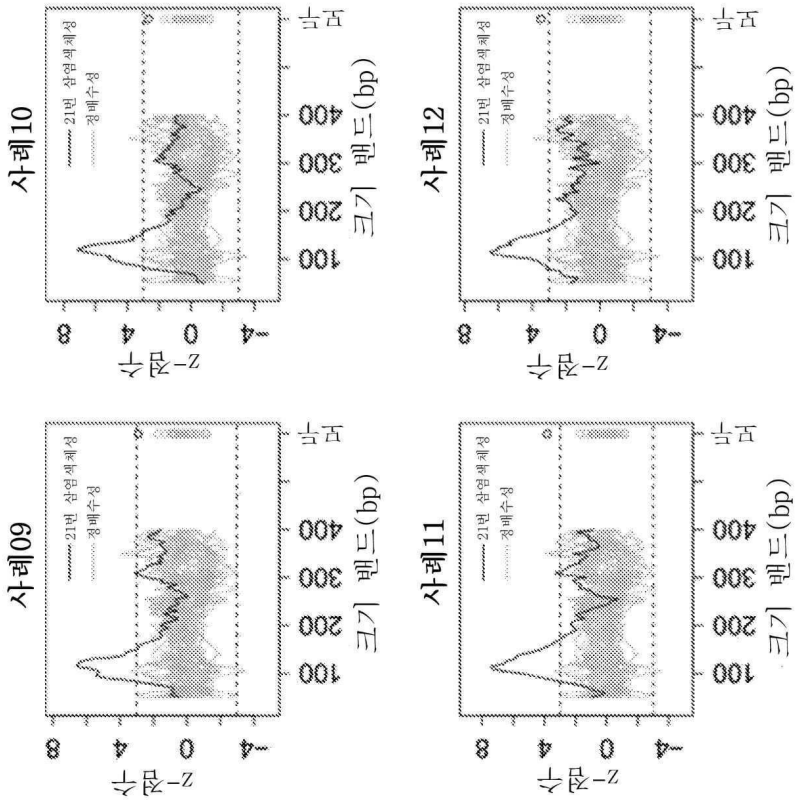
도면3a



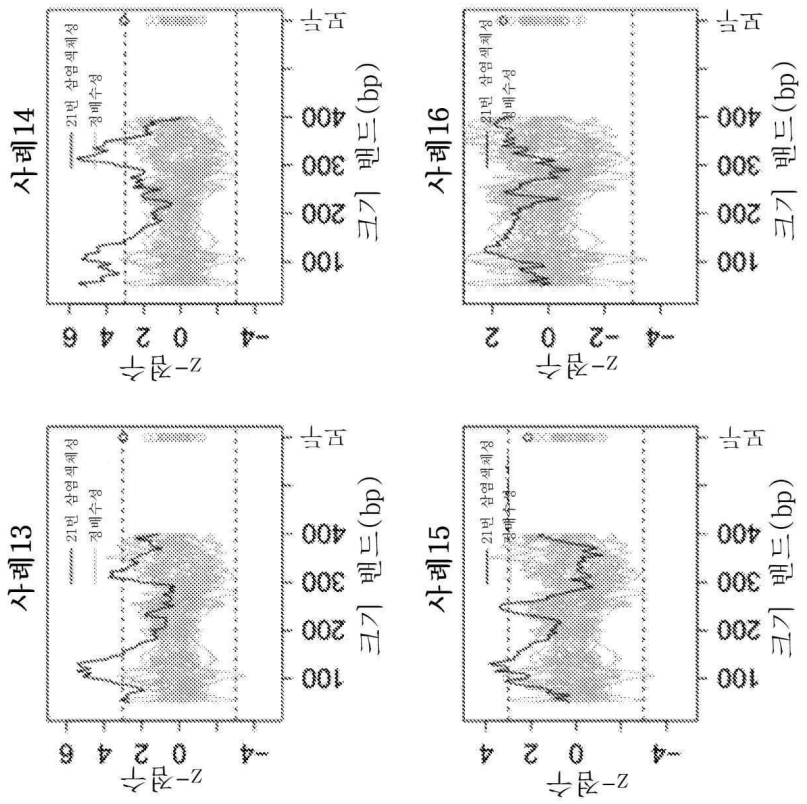
도면3b


사례06
사례08
사례05
사례07
21번 삼염색체성
정배수성
Z-점수
크기 밴드(bp)
모두

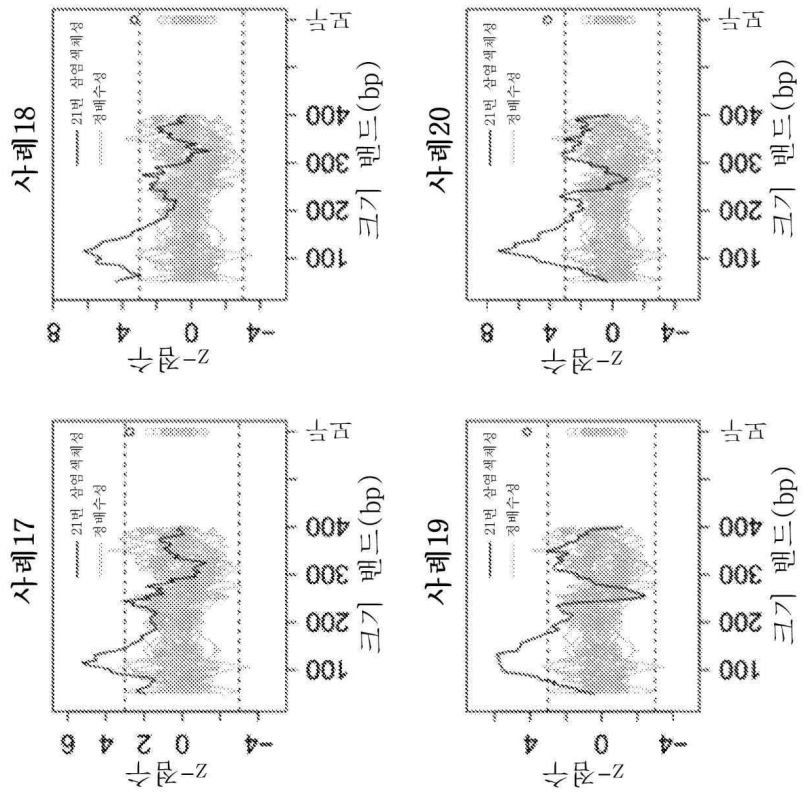
도면3c


사례10
사례12
사례09
사례11
21번 삼염색체성
정배수성
Z-점수
크기 밴드(bp)
모두

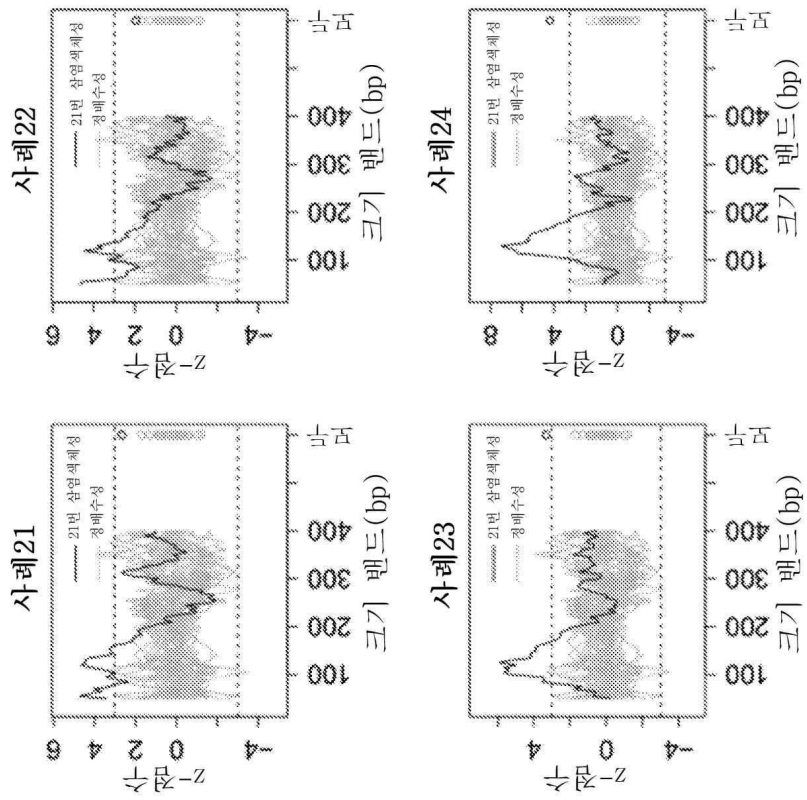
도면3d


사례14
사례16
사례13
사례15
21번 삼염색체성
정배수성
Z-점수
크기 밴드(bp)
모두

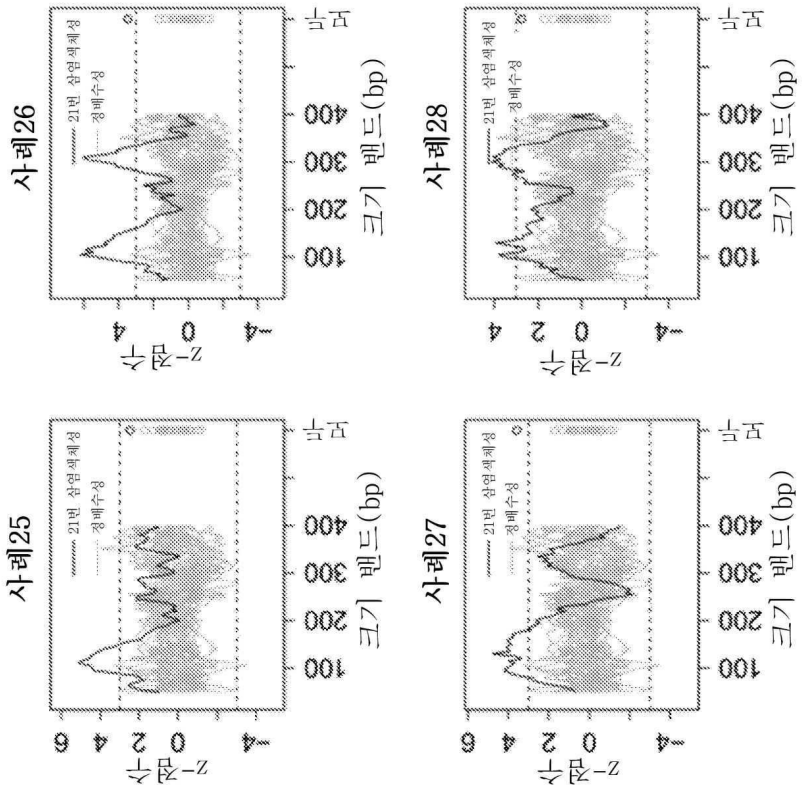
도면3e


사례18
사례20
사례17
사례19
21번 삼염색체성
정배수성
Z-점수
크기 밴드(bp)
모두

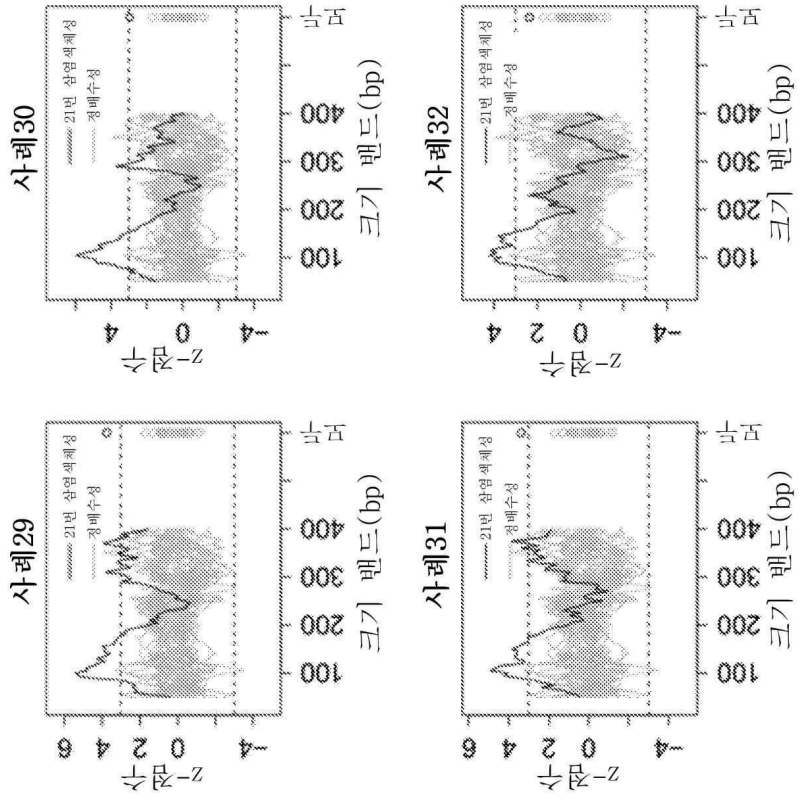
도면3f


사례22
사례24
사례21
사례23
21번 삼염색체성
정배수성
Z-점수
크기 밴드(bp)
모두

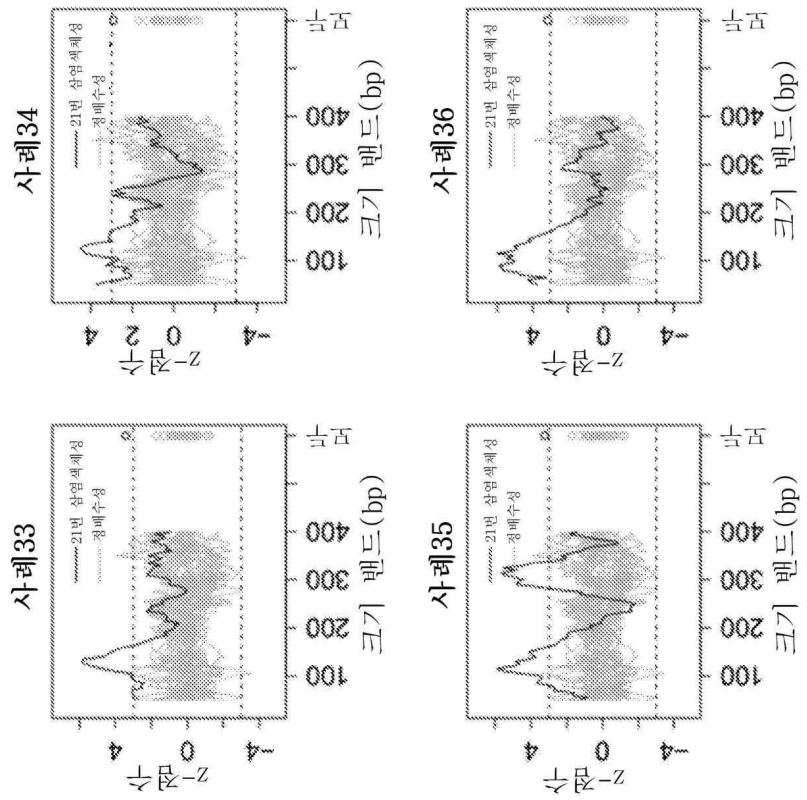
도면3g


사례26
사례28
사례25
사례27
21번 삼염색체성
정배수성
Z-점수
크기 밴드(bp)
모두

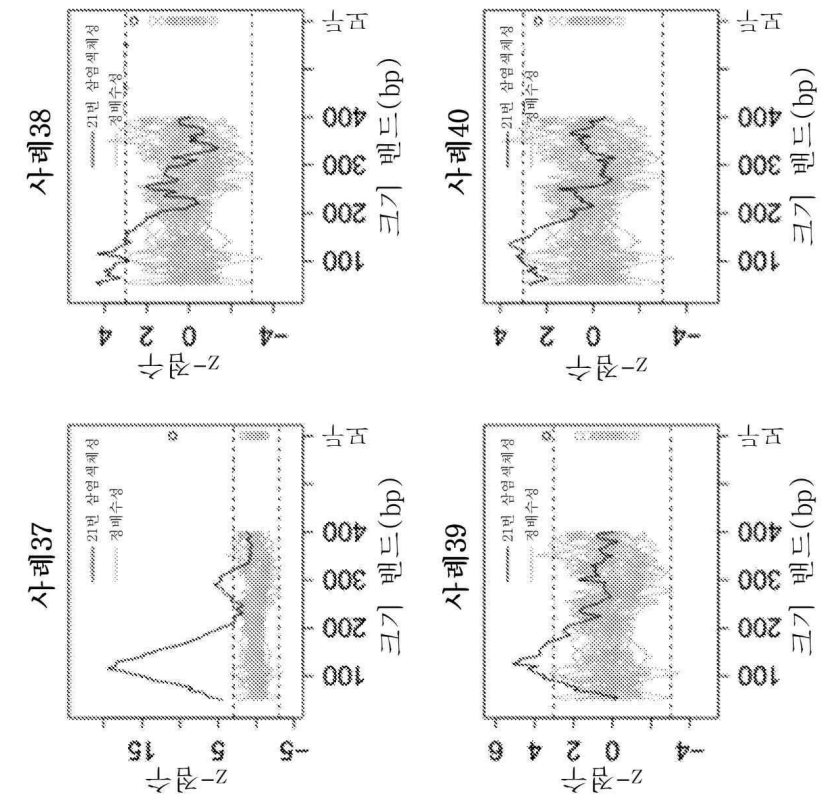
도면3h


사례30
사례29
사례32
사례31
21번 삼염색체성
정배수성
Z-점수
크기 밴드(bp)
모두

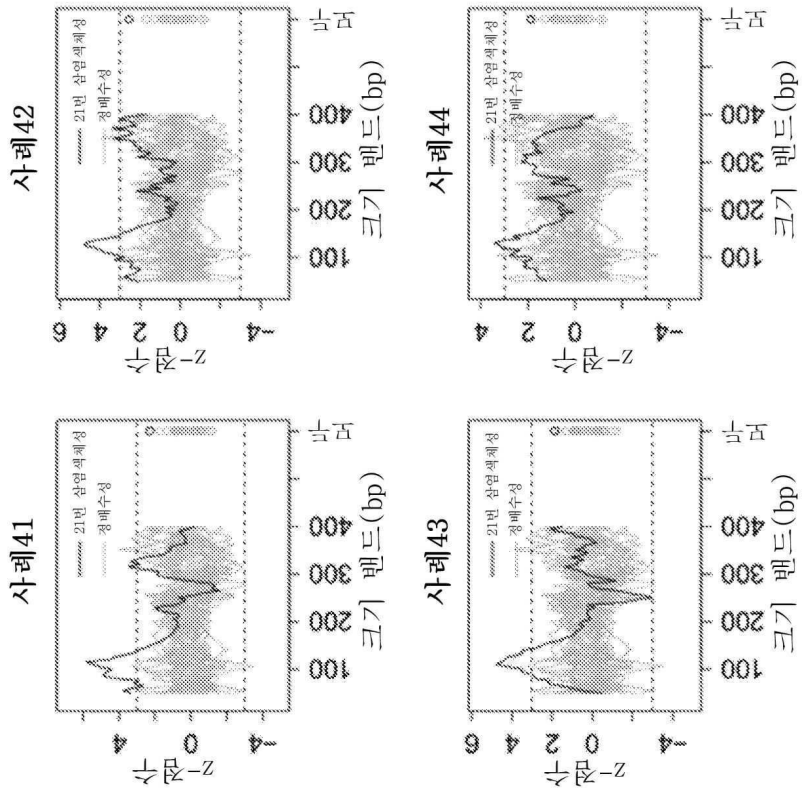
도면3i


사례34
사례33
사례36
사례35
21번 삼염색체성
정배수성
Z-점수
크기 밴드(bp)
모두

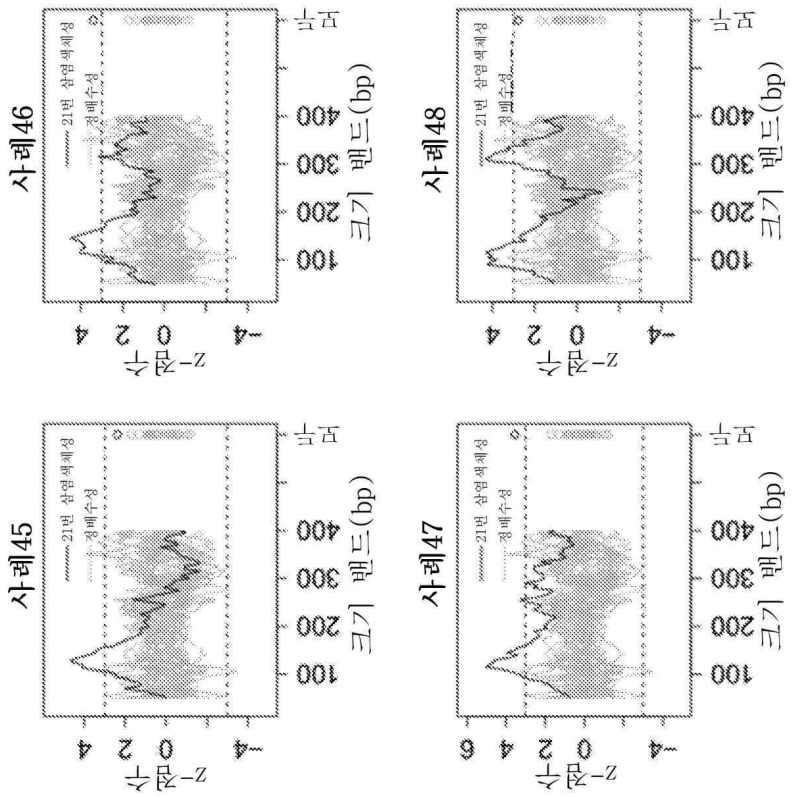
도면3j


사례38
사례37
사례40
사례39
21번 삼염색체성
정배수성
Z-점수
크기 밴드(bp)
모두

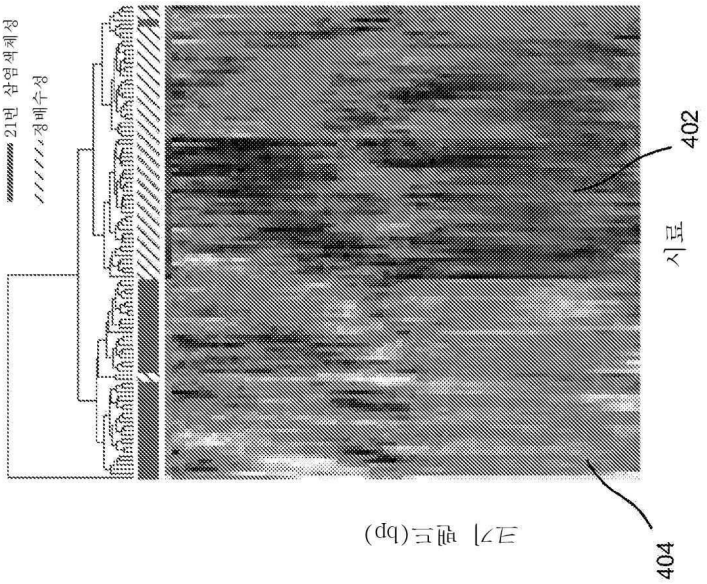
도면3k


사례42
사례41
사례44
사례43
21번 삼염색체성
정배수성
Z-점수
크기 밴드(bp)
모두

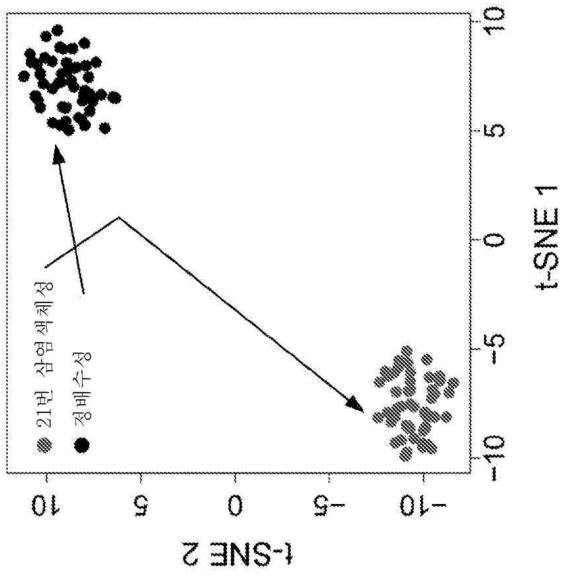
도면31


사례46
사례48
사례45
사례47
21번 삼염색체성
정배수성
Z-점수
크기 밴드(bp)
모두

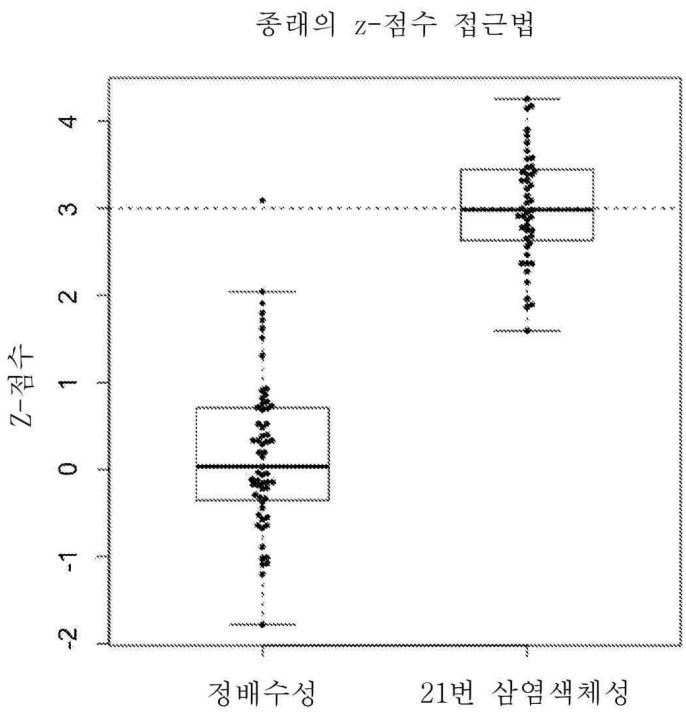
도면4a


21번 삼염색체성
정배수성
시료
크기 밴드(bp)
402
404

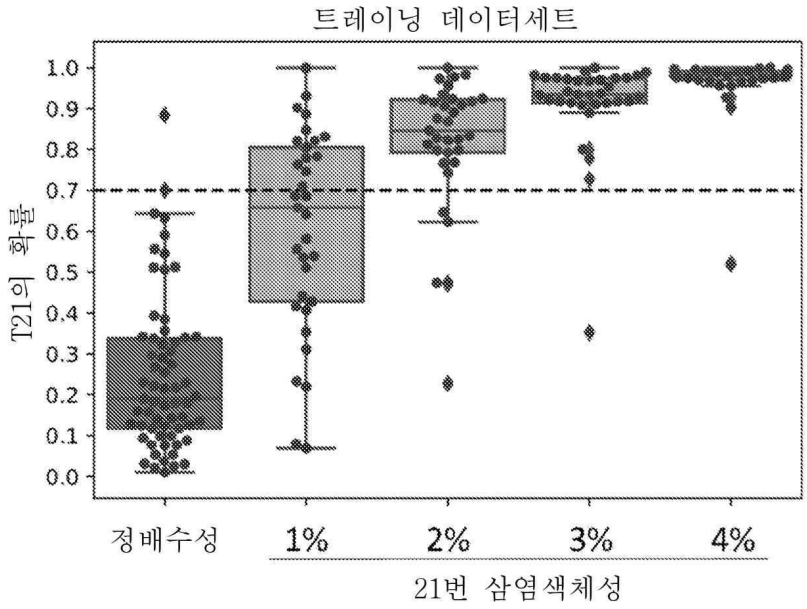
도면4b


21번 삼염색체성
정배수성
t-SNE 1
t-SNE 2
10
5
0
-5
-10

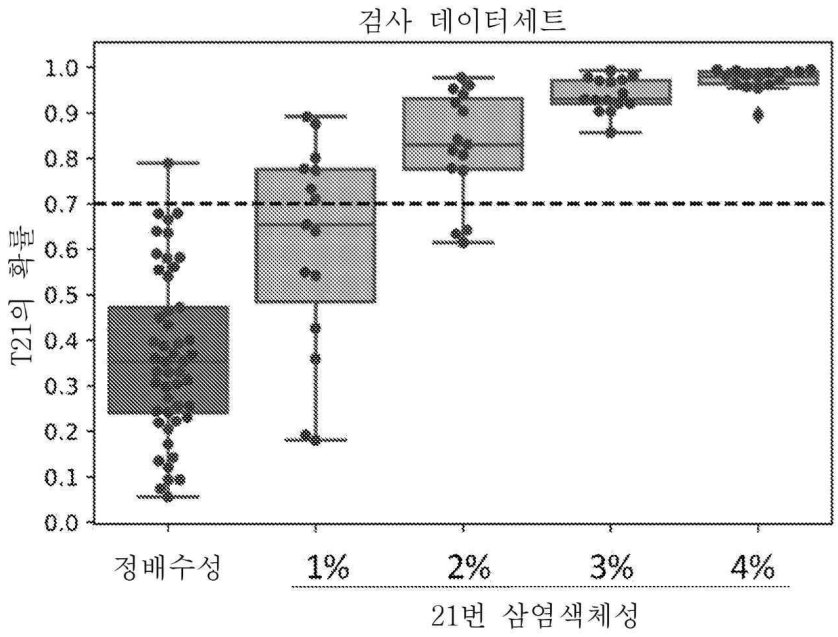
도면4c


종래의 z-점수 접근법
Z-점수
4
3
2
1
0
-1
-2
정배수성
21번 삼염색체성

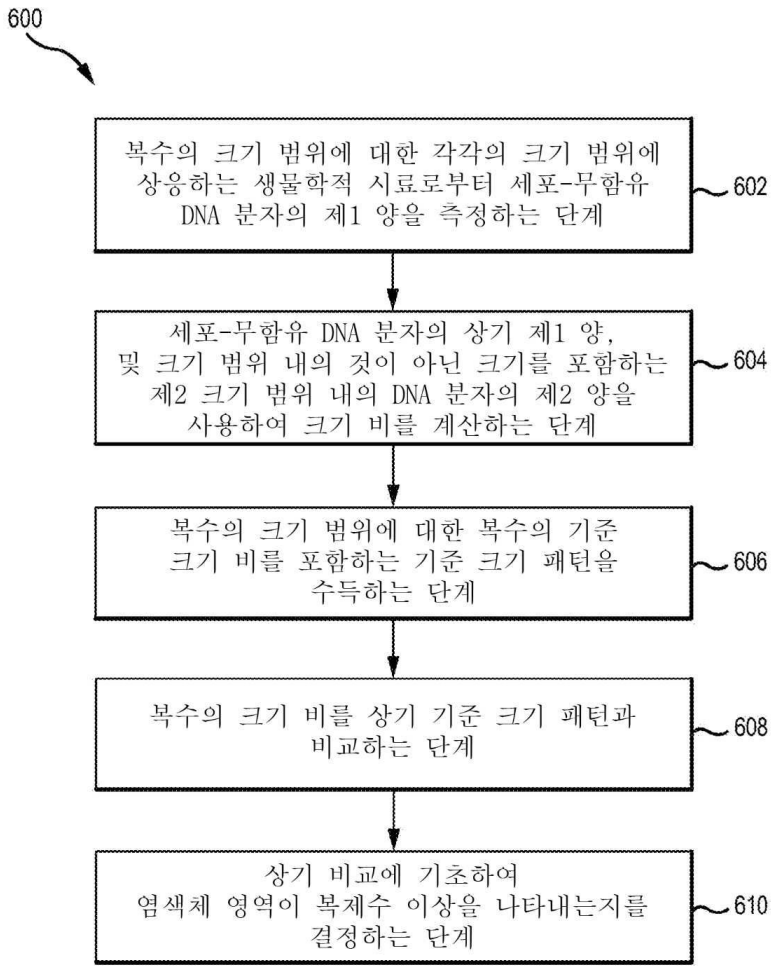
도면5a


트레이닝 데이터세트
T21의 확률
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
정배수성
1%
2%
3%
4%
21번 삼염색체성

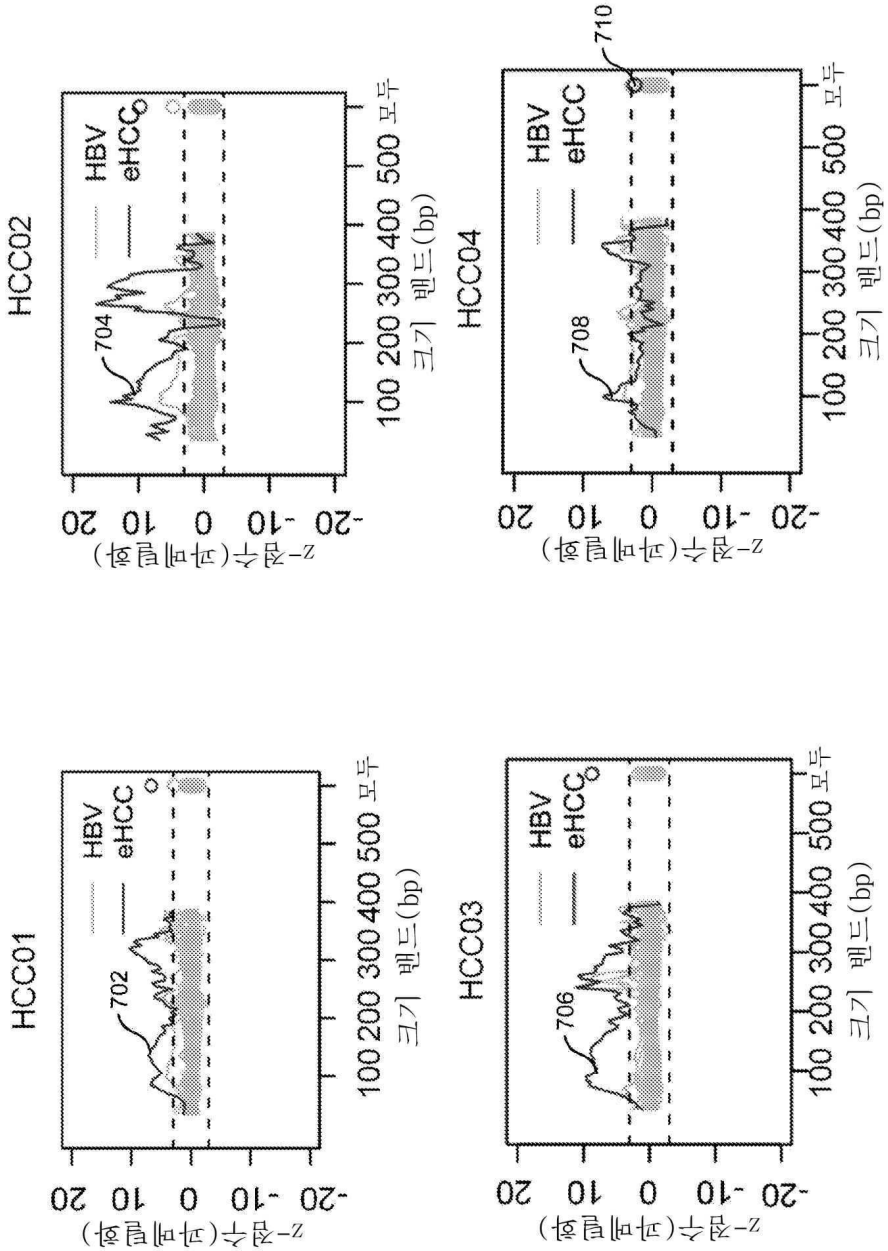
도면5b


검사 데이터세트
T21의 확률
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
정배수성
1%
2%
3%
4%
21번 삼염색체성

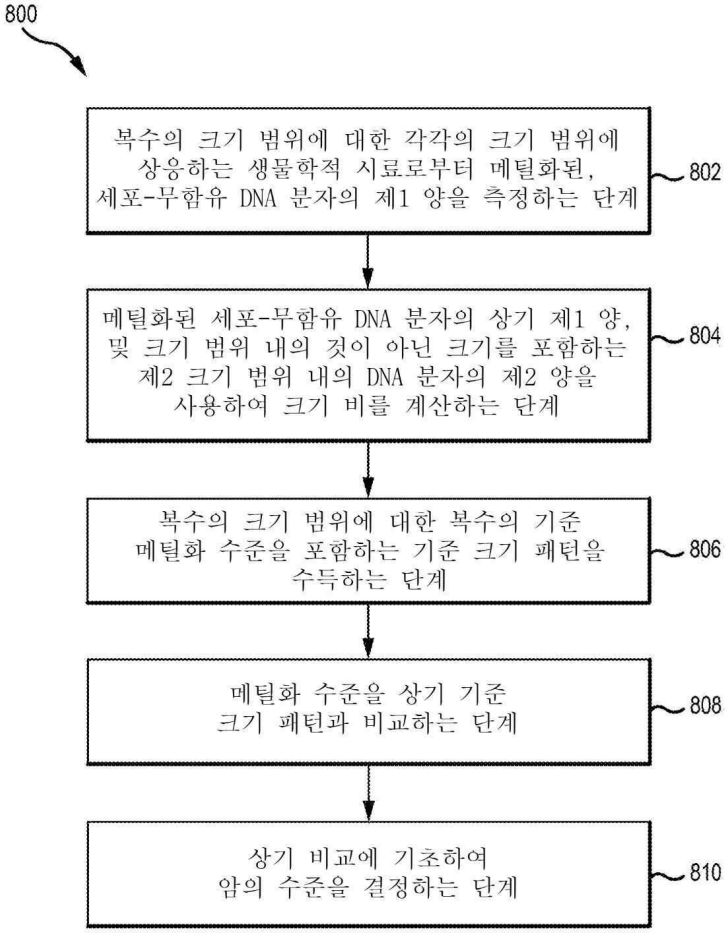
도면6


600
복수의 크기 범위에 대한 각각의 크기 범위에 상응하는 생물학적 시료로부터 세포-무함유 DNA 분자의 제1 양을 측정하는 단계
602
세포-무함유 DNA 분자의 상기 제1 양, 및 크기 범위 내의 것이 아닌 크기를 포함하는 제2 크기 범위 내의 DNA 분자의 제2 양을 사용하여 크기 비를 계산하는 단계
604
복수의 크기 범위에 대한 복수의 기준 크기 비를 포함하는 기준 크기 패턴을 수득하는 단계
606
복수의 크기 비를 상기 기준 크기 패턴과 비교하는 단계
608
상기 비교에 기초하여 염색체 영역이 복제수 이상을 나타내는지를 결정하는 단계
610

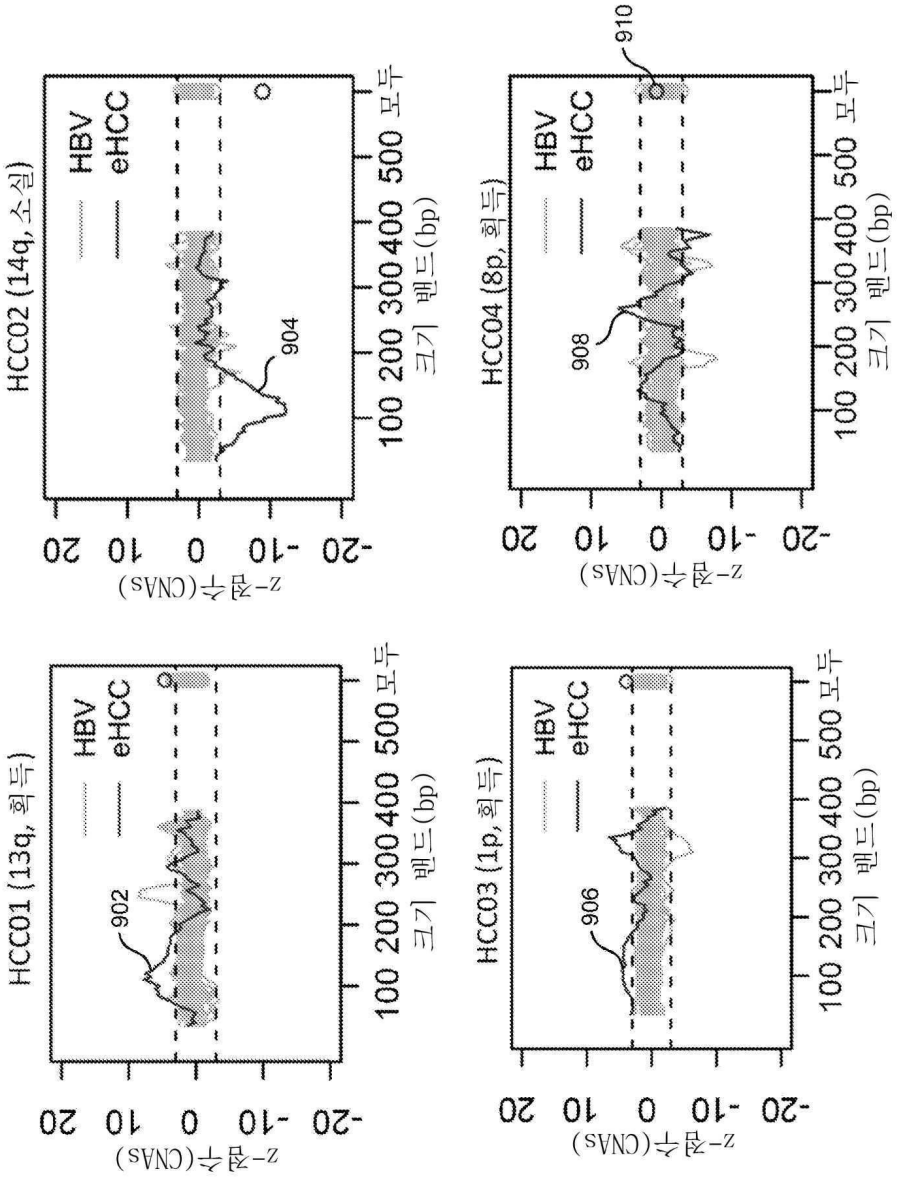
도면7


HCC01
HCC02
HCC03
HCC04
HBV
eHCC
702
704
706
708
710
Z-점수(과메틸화)
크기 밴드(bp)
100 200 300 400 500 모두
-20 -10 0 10 20

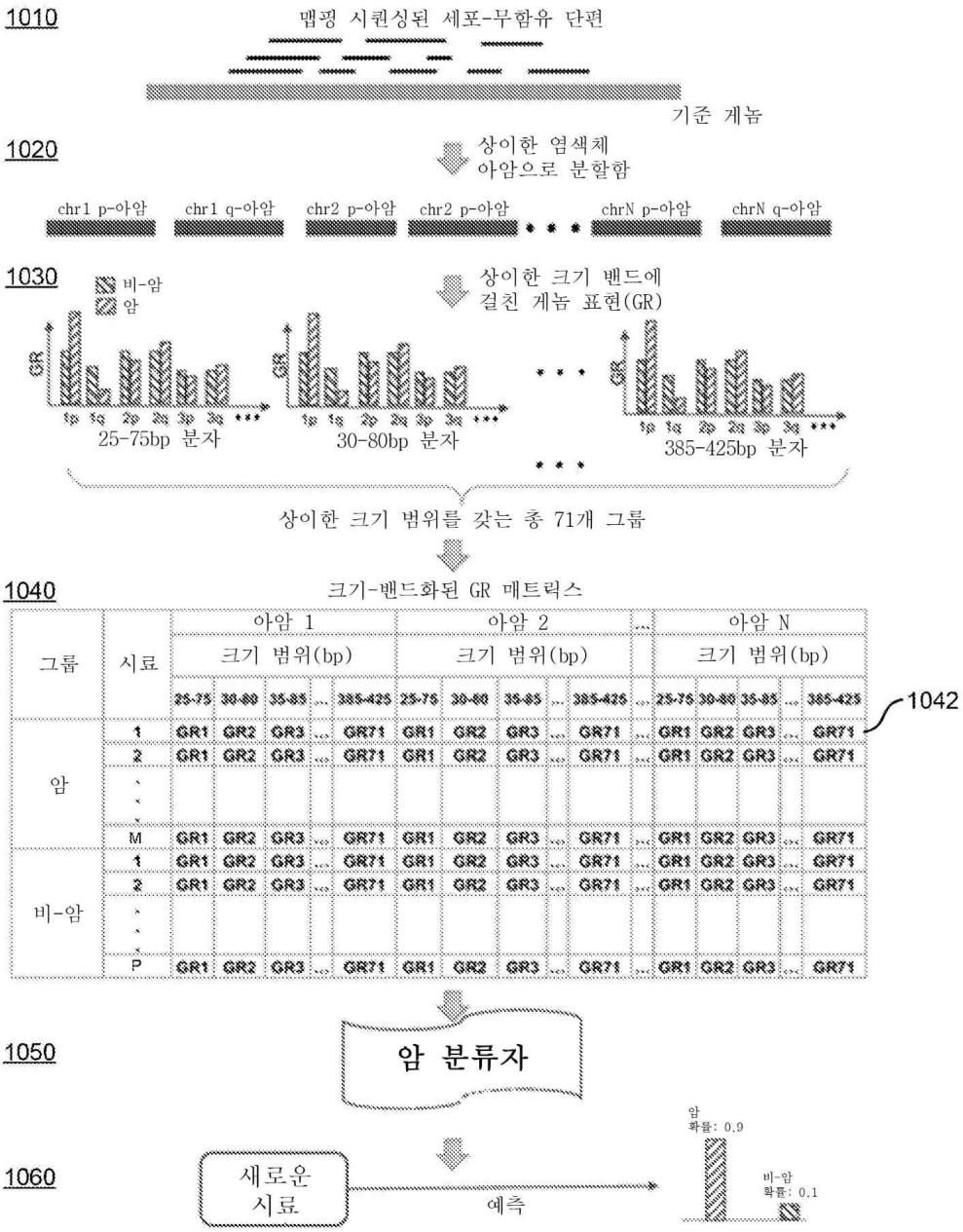
도면8


800
복수의 크기 범위에 대한 각각의 크기 범위에 상응하는 생물학적 시료로부터 메틸화된, 세포-무함유 DNA 분자의 제1 양을 측정하는 단계
802
메틸화된 세포-무함유 DNA 분자의 상기 제1 양, 및 크기 범위 내의 것이 아닌 크기를 포함하는 제2 크기 범위 내의 DNA 분자의 제2 양을 사용하여 크기 비를 계산하는 단계
804
복수의 크기 범위에 대한 복수의 기준 메틸화 수준을 포함하는 기준 크기 패턴을 수득하는 단계
806
메틸화 수준을 상기 기준 크기 패턴과 비교하는 단계
808
상기 비교에 기초하여 암의 수준을 결정하는 단계
810

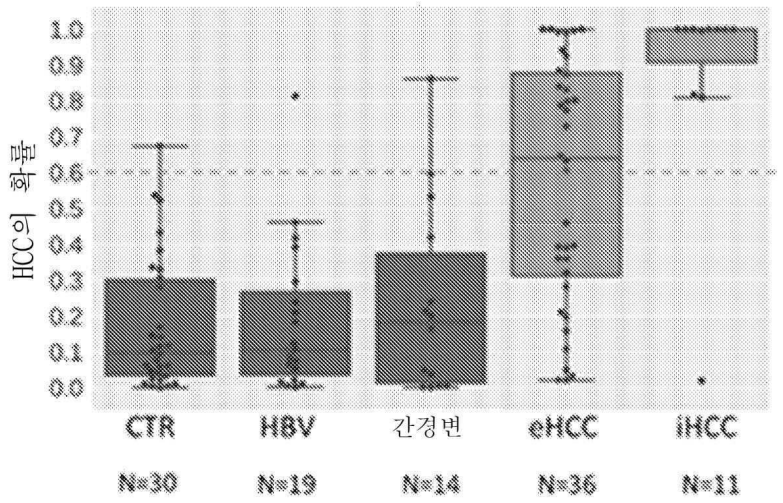
도면9


HCC02 (14q, 소실)
HBV
eHCC
904
z-점수(CNAs)
크기 밴드(bp)
HCC04 (8p, 획득)
908
910
HCC01 (13q, 획득)
902
HCC03 (1p, 획득)
906
100 200 300 400 500 모두
-20 -10 0 10 20

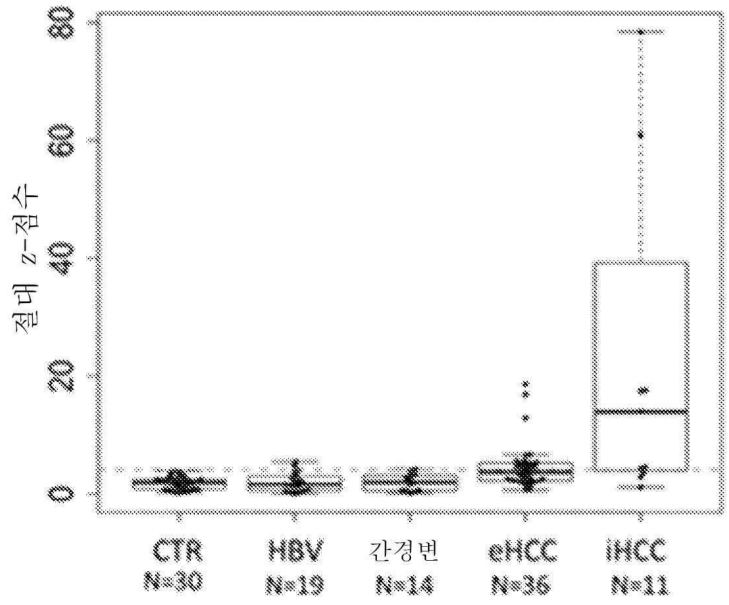
도면10


1010
맵핑 시퀀싱된 세포-무함유 단편
기준 게놈
1020
상이한 염색체
아암으로 분할함
chr1 p-아암
chr1 q-아암
chr2 p-아암
chr2 p-아암
chrN p-아암
chrN q-아암
1030
비-암
암
상이한 크기 밴드에
걸친 게놈 표현(GR)
GR
25-75bp 분자
30-80bp 분자
385-425bp 분자
상이한 크기 범위를 갖는 총 71개 그룹
1040
크기-밴드화된 GR 매트릭스
그룹
시료
아암 1
아암 2
아암 N
크기 범위(bp)
25-75
30-80
35-85
385-425
암
비-암
GR1
GR2
GR3
GR71
1042
1050
암 분류자
1060
새로운
시료
예측
암
확률: 0.9
비-암
확률: 0.1

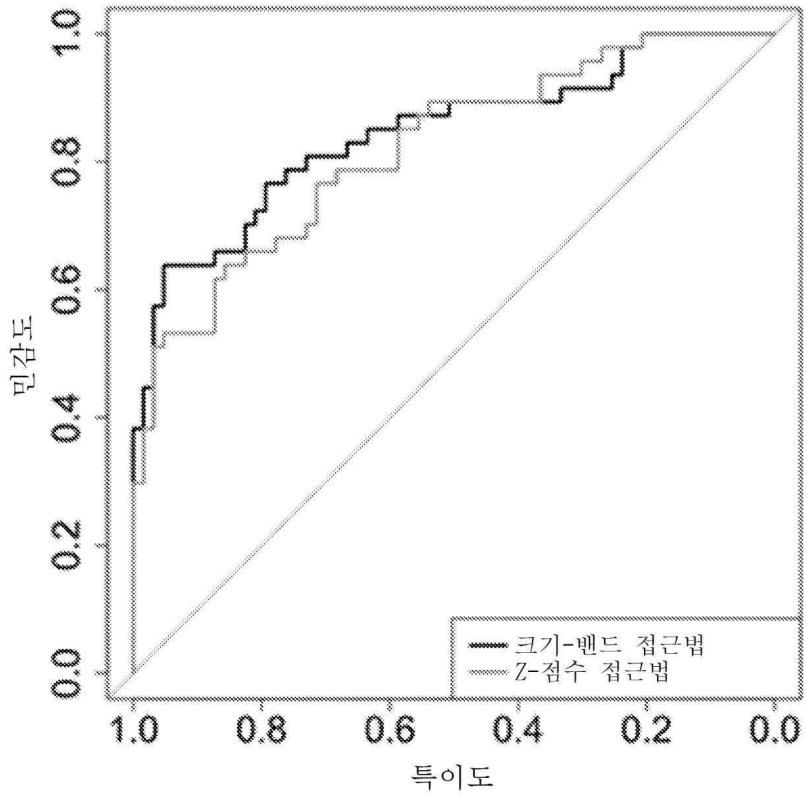
도면11a


HCC의 확률
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
CTR
N=30
HBV
N=19
간경변
N=14
eHCC
N=36
iHCC
N=11

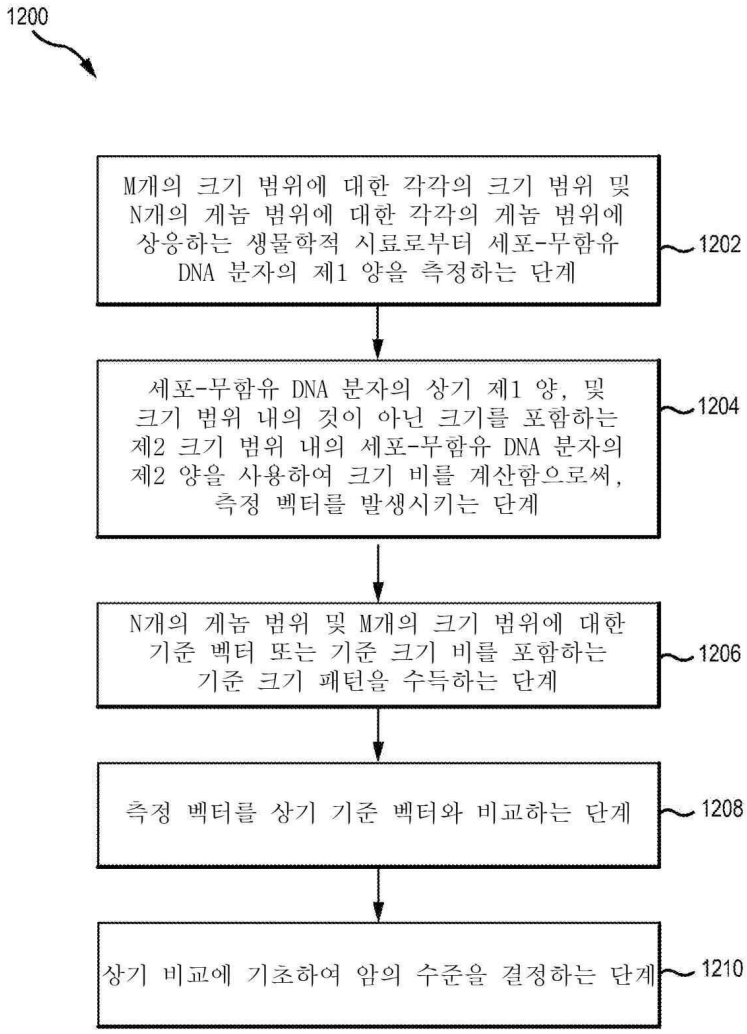
도면11b


절대 z-점수
80
60
40
20
0
CTR
N=30
HBV
N=19
간경변
N=14
eHCC
N=36
iHCC
N=11

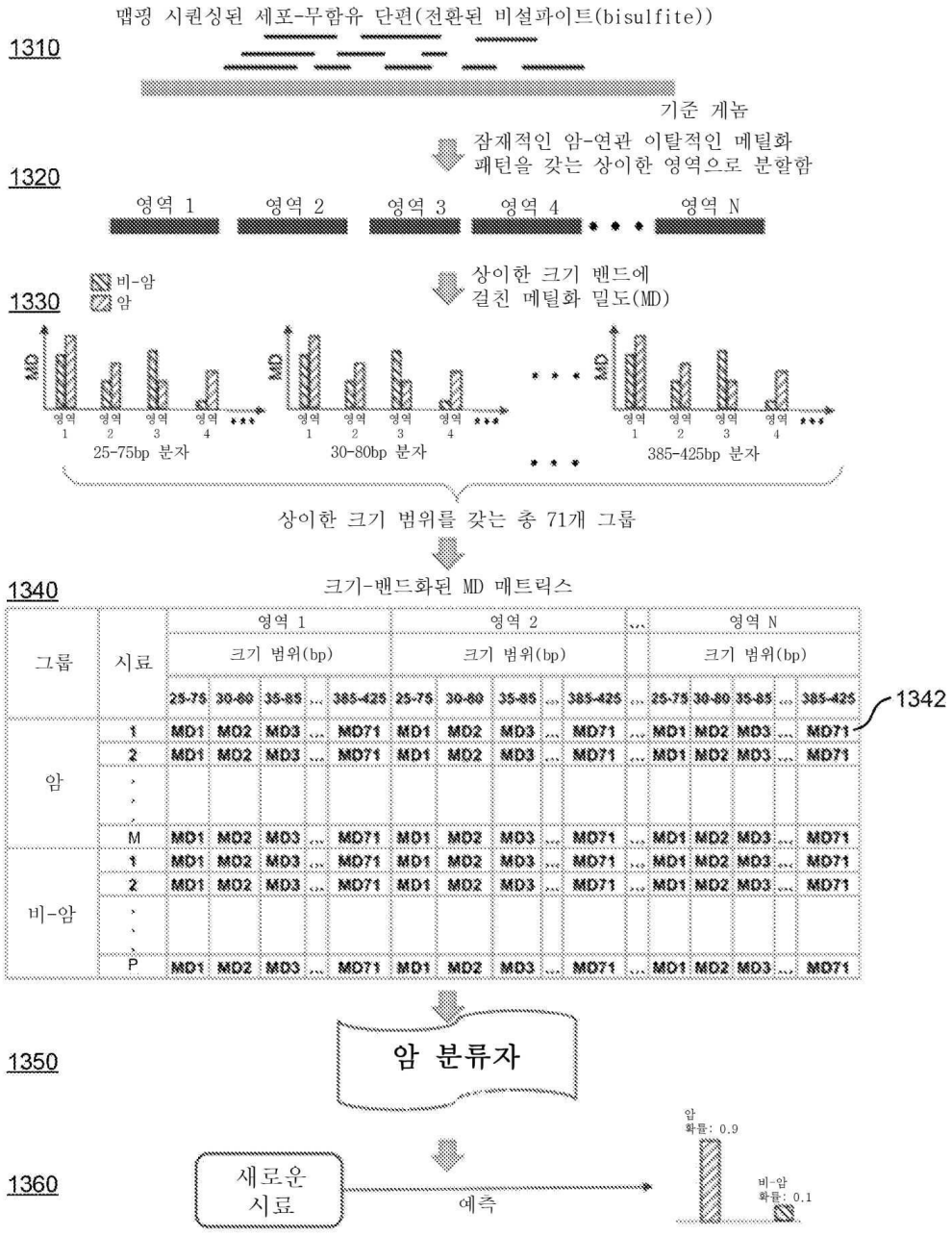
도면11c



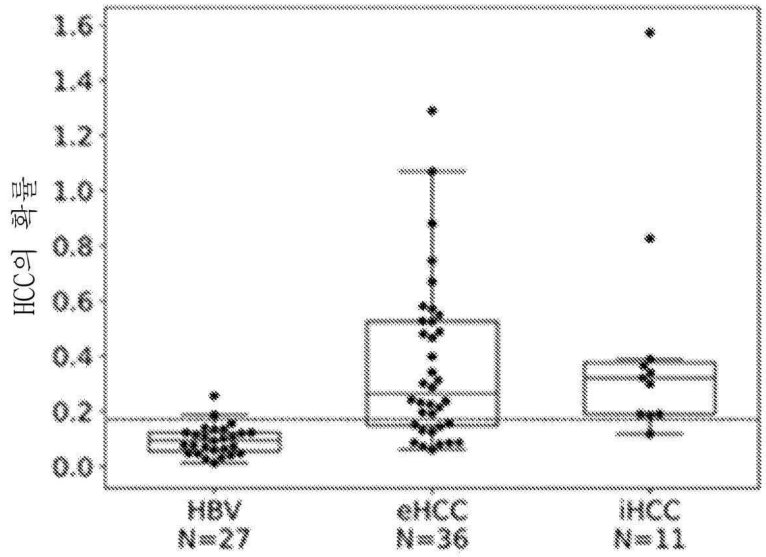
도면12


1200
M개의 크기 범위에 대한 각각의 크기 범위 및 N개의 게놈 범위에 대한 각각의 게놈 범위에 상응하는 생물학적 시료로부터 세포-무함유 DNA 분자의 제1 양을 측정하는 단계
1202
세포-무함유 DNA 분자의 상기 제1 양, 및 크기 범위 내의 것이 아닌 크기를 포함하는 제2 크기 범위 내의 세포-무함유 DNA 분자의 제2 양을 사용하여 크기 비를 계산함으로써, 측정 벡터를 발생시키는 단계
1204
N개의 게놈 범위 및 M개의 크기 범위에 대한 기준 벡터 또는 기준 크기 비를 포함하는 기준 크기 패턴을 수득하는 단계
1206
측정 벡터를 상기 기준 벡터와 비교하는 단계
1208
상기 비교에 기초하여 암의 수준을 결정하는 단계
1210

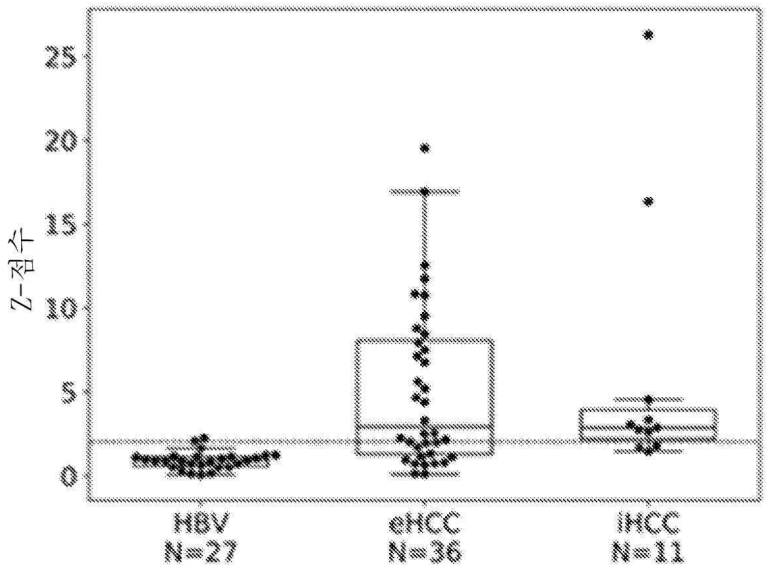
도면13


맵핑 시퀀싱된 세포-무함유 단편(전환된 비설파이트(bisulfite))
1310
기준 게놈
잠재적인 암-연관 이탈적인 메틸화 패턴을 갖는 상이한 영역으로 분할함
1320
영역 1
영역 2
영역 3
영역 4
영역 N
1330
비-암
암
상이한 크기 밴드에 걸친 메틸화 밀도(MD)
MD
영역 1
영역 2
영역 3
영역 4
25-75bp 분자
30-80bp 분자
385-425bp 분자
상이한 크기 범위를 갖는 총 71개 그룹
1340
크기-밴드화된 MD 매트릭스
그룹
시료
영역 1
영역 2
영역 N
크기 범위(bp)
25-75
30-80
35-85
385-425
암
비-암
MD1
MD2
MD3
MD71
1342
1350
암 분류자
1360
새로운 시료
예측
암
확률: 0.9
비-암
확률: 0.1

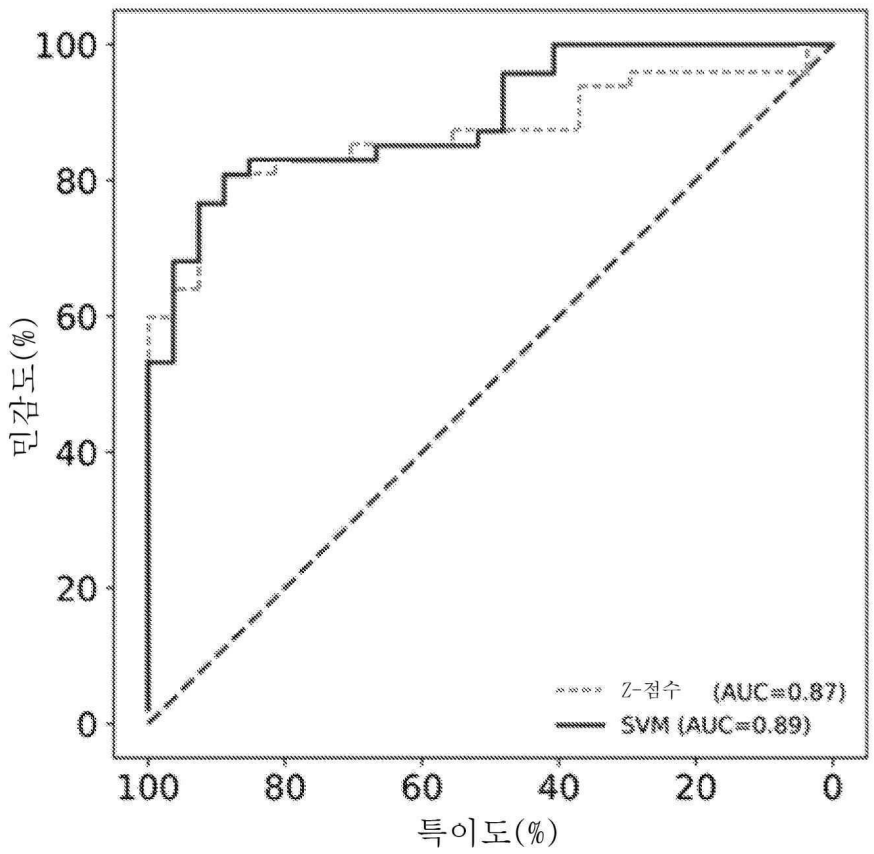
도면14a


HCC의 확률
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
HBV
N=27
eHCC
N=36
iHCC
N=11

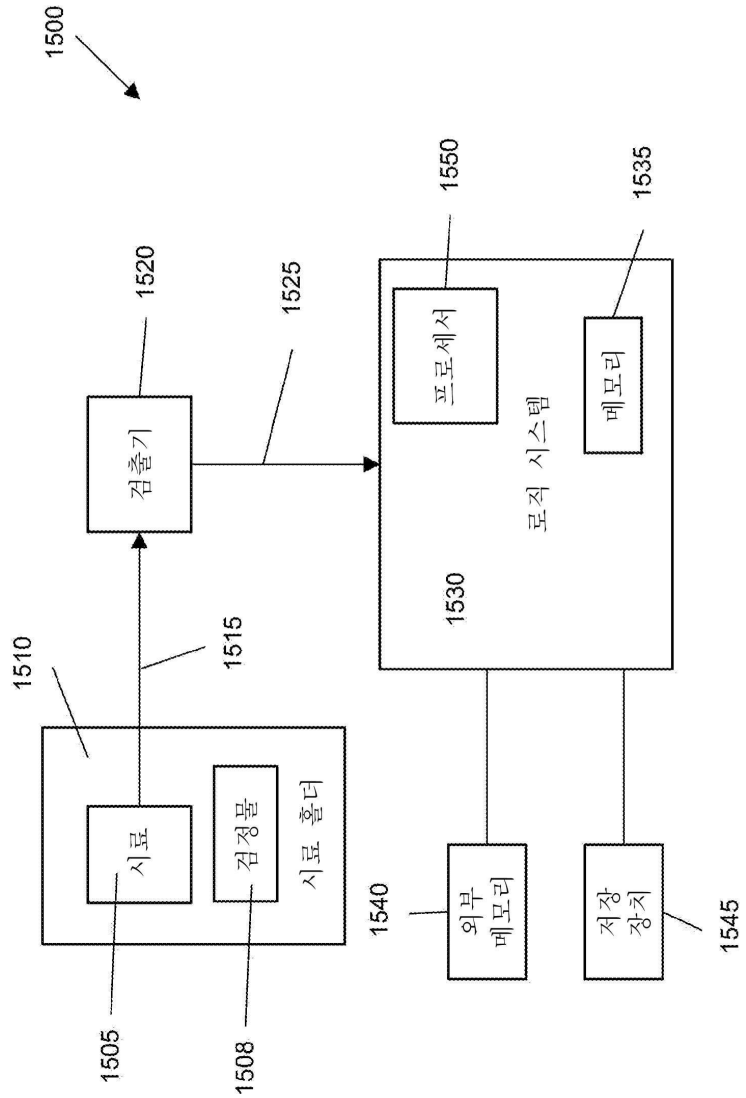
도면14b


Z-점수
25
20
15
10
5
0
HBV
N=27
eHCC
N=36
iHCC
N=11

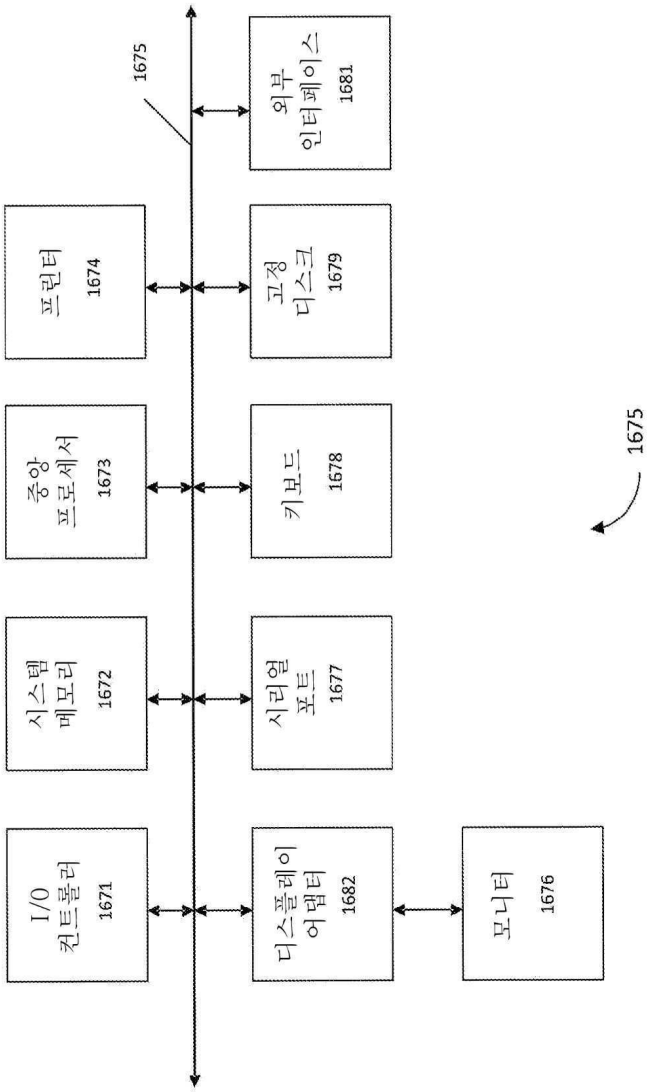
도면14c



도면15


1500
1520
1525
1550
1535
검출기
프로세서
메모리
로직 시스템
1530
1510
1515
시료
검정물
시료 홀더
1505
1508
1540
외부 메모리
저장 장치
1545

도면16


1675
외부 인터페이스 1681
프린터 1674
고정 디스크 1679
중앙 프로세서 1673
키보드 1678
시스템 메모리 1672
시리얼 포트 1677
I/O 컨트롤러 1671
디스플레이 어댑터 1682
모니터 1676
1675