(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년05월21일
(11) 등록번호 10-2665592
(24) 등록일자 2024년05월08일

(51) 국제특허분류(Int. Cl.)
G16B 20/00 (2019.01) *C12Q 1/6869* (2018.01)
C12Q 1/6883 (2018.01) *G16B 30/00* (2019.01)
G16B 40/00 (2019.01)
(52) CPC특허분류
G16B 20/00 (2019.02)
C12Q 1/6869 (2018.05)
(21) 출원번호 10-2023-7018438(분할)
(22) 출원일자(국제) 2014년05월23일
심사청구일자 2023년06월29일
(85) 번역문제출일자 2023년05월31일
(65) 공개번호 10-2023-0082691
(43) 공개일자 2023년06월08일
(62) 원출원 특허 10-2022-7011344
원출원일자(국제) 2014년05월23일
심사청구일자 2022년05월04일
(86) 국제출원번호 PCT/US2014/039389
(87) 국제공개번호 WO 2014/190286
국제공개일자 2014년11월27일
(30) 우선권주장
61/827,385 2013년05월24일 미국(US)

(56) 선행기술조사문헌
WO2013052913 A2
WO2013052913 A2
S. Chen 외, "A method for noninvasive detection of fetal large deletions/duplications by low coverage massively parallel sequencing", Prenatal Diagnosis 33, pp.584-590, 2013.5.17.
Y. Yang 외, "A novel k-mer mixture logistic regression for methylation susceptibility modeling of CpG dinucleotides in human gene promoters", BMC Bioinformatics 13(Suppl 3):S15, 2012.

(73) 특허권자
시쿼넘, 인코포레이티드
미국 캘리포니아주 92121 샌 디에고 존 홉킨스 코트 3595
(72) 발명자
자오 첸
미국 92122 캘리포니아주 샌디에이고 코스타 베르데 블러바드 8950 아파트먼트 4124
자쿨라 젤리코
미국 92129 캘리포니아주 샌디에이고 선댄스 에비뉴 12830
(뒷면에 계속)
(74) 대리인
김진희, 김태홍

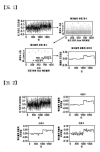
전체 청구항 수 : 총 19 항 심사관 : 성경아

(54) 발명의 명칭 **유전적 변이의 비침습 평가를 위한 방법 및 프로세스**

(57) 요 약

본원은 유전적 변이의 비침습 평가 방법을 제공한다. 이 방법은 맵핑된 서열 리드(mapped sequence reads)에 기초하고, 결정 분석을 이용한다. 결정 분석은 종종 분할(segmentation) 분석 및/또는 승산비(odds ratio) 분석을 포함한다.

대 표 도



(52) CPC특허분류
C12Q 1/6883 (2022.01)
G16B 30/00 (2019.02)
G16B 40/00 (2019.02)
C12Q 2535/122 (2019.08)
C12Q 2537/165 (2013.01)
C12Q 2600/156 (2013.01)

(72) 발명자
데시우 코스민
미국 92121 캘리포니아주 샌디에이고 씨 미스트 웨이 10545
김 성균
미국 91206 캘리포니아주 글렌데일 글렌모어 블러바드 662
마즐룸 아민
미국 92128 캘리포니아주 샌디에이고 케이프우드 레인 14013
핸넘 그레고리
미국 92129 캘리포니아주 샌디에이고 바바리언 드라이브 13125
에리히 마티아스
미국 92109 캘리포니아주 샌디에이고 다이아몬드 스트리트 1947

명 세 서

청구범위

청구항 1

불균질(heterogeneous) 샘플에서 미세중복 또는 미세결실의 존재 또는 부재를 결정하는 방법으로서,

(a) 기준 게놈의 부분들에 맵핑된 핵산 서열 리드의 카운트를 수득하는 단계로서, 이때 상기 서열 리드는 불균질 핵산을 포함하는 검사 샘플로부터의 순환 무세포 핵산의 리드인 단계;

(b) 각각의 부분들에 맵핑된 카운트를 표준화하여, 표준화된 카운트를 제공하는 단계;

(c) 상기 부분들의 표준화된 카운트 또는 상기 부분들의 서브세트의 표준화된 카운트를 분할하여, 하나 이상의 별개의 분절을 제공하는 단계;

(d) 하나 이상의 별개의 분절 중에서 후보 분절을 확인하는 단계;

(e) 후보 분절의 카운트 표시치를 생성하는 단계;

(f) 후보 분절에 대한 로그 승산비(LOR)를 결정하는 단계로서, 이때 LOR은 (i) (1) 미세중복 또는 미세결실을 가질 조건부 확률과 (2) 미세중복 또는 미세결실을 가질 선험적 확률의 제1 곱셈 결과, 및 (ii) (1) 미세중복 또는 미세결실을 갖지 않을 조건부 확률과 (2) 미세중복 또는 미세결실을 갖지 않을 선험적 확률의 제2 곱셈 결과의 몫의 로그인 것이고, 상기 미세중복 또는 미세결실을 가질 조건부 확률은 (e) 단계의 카운트 표시치에 따라 결정되는 것인 단계; 및

(g) 후보 분절에 대해 결정된 LOR 및 카운트 표시치에 따른 미세중복 또는 미세결실의 존재 또는 부재를 결정하는 단계를 포함하는 방법.

청구항 2

제1항에 있어서, (b) 단계의 상기 표준화가 구아닌 및 사이토신(GC) 편향(bias)의 LOESS 표준화(GC-LOESS 표준화), 주성분 표준화, 또는 GC-LOESS 표준화에 이은 주성분 표준화를 포함하는 방법.

청구항 3

제1항에 있어서, 상기 (c) 단계의 분할이 2종 이상의 상이한 분할 프로세스의 적용을 포함하는 것인 방법.

청구항 4

제3항에 있어서, 2종 이상의 상이한 분할 프로세스 각각이 Haar 웨이블렛 분할(Haar wavelet segmentation), 원형 이진 분할(circular binary segmentation), 최대 엔트로피 분할(maximum entropy segmentation), 에지 검출 커넬을 이용한 컨볼루션(convolution with edge detection kernel), 젠슨 샤논 발산(Jensen Shannon Divergence), 이진 반복 분할(Binary Recursive Segmentation) 및 푸리에 변환(Fourier transform)으로부터 독립적으로 선택되는 것인 방법.

청구항 5

제1항 내지 제4항 중 어느 한 항에 있어서, 상기 후보 분절은 5Mb미만의 길이인 방법.

청구항 6

제1항 내지 제4항 중 어느 한 항에 있어서, 후보 분절에 대한 카운트 표시치의 정량치를 생성하는 단계를 더 포함하는 방법.

청구항 7

제6항에 있어서, 상기 후보 분절에 대한 카운트 표시치의 정량치는 z-점수 정량치인 방법.

청구항 8

제7항에 있어서, 상기 z-점수가 (i) 검사 샘플에 대한 후보 분절에 대한 카운트 표시치로부터 (ii) 후보 분절에 대한 정배수체 카운트 표시치의 중간치를 뺀 뺄셈 결과를 (iii) 후보 분절에 대한 정배수체 카운트 표시치의 MAD로 나눈 값이고, 상기 검사 샘플에 대한 후보 분절에 대한 카운트 표시치는 후보 분절에 대한 총 표준화된 카운트를 검사 샘플에 대한 총 상염색체 카운트로 나눈 비이고, 정배수체 카운트 표시치의 중간치는 후보 분절에 대한 총 카운트를 상염색체 샘플에 대한 총 상염색체 카운트로 나눈 비의 중간치인 방법.

청구항 9

제8항에 있어서, 검사 샘플에 대하여, (i) 후보 분절에 대한 카운트 표시치의 z-점수 정량치가 3.95의 값보다 더 크거나 또는 동등하다는 결정, 및 (ii) LOR이 0보다 더 크다는 결정에 적어도 부분적으로 기초하여 미세중복 또는 미세결실의 존재를 결정하는 단계를 포함하는 방법.

청구항 10

제8항에 있어서, 검사 샘플에 대하여, (i) 후보 분절에 대한 카운트 표시치의 z-점수 정량치가 3.95의 값보다 작다는 결정, 또는 (ii) LOR이 0보다 더 작다는 결정; 또는 (i) 후보 분절에 대한 카운트 표시치의 z-점수 정량치가 3.95의 값보다 작다는 결정, 및 (ii) LOR이 0보다 더 작다는 결정에 적어도 부분적으로 기초하여 미세결실 또는 미세중복의 부재를 결정하는 단계를 포함하는 방법.

청구항 11

제8항에 있어서, 후보 분절이 위치하는 염색체의 염색체 표시치의 정량치를 생성하는 단계를 더 포함하는 방법.

청구항 12

제11항에 있어서, 염색체 표시치의 정량치가 z-점수 정량치인 방법.

청구항 13

제12항에 있어서, z-점수가 (i) 검사 샘플에 대한 염색체 카운트 표시치로부터 (ii) 정배수체 염색체 카운트 표시치의 중간치를 뺀 뺄셈 결과를 (iii) 정배수체 염색체 카운트 표시치의 MAD로 나눈 값이고, 상기 검사 샘플에 대한 염색체 카운트 표시치는 후보 분절이 위치하는 염색체의 총 카운트를 검사 샘플에 대한 총 상염색체 카운트로 나눈 비이고, 정배수체 염색체 카운트 표시치의 중간치는, 후보 분절이 위치하는 염색체의 총 카운트를 정배수체 샘플에 대한 총 상염색체 카운트로 나눈 비의 중간치인 방법.

청구항 14

제11항에 있어서, 후보 분절에 대한 카운트 표시치의 정량치를 염색체 표시치의 정량치와 비교하는 것인 방법.

청구항 15

제1항 내지 제4항 중 어느 한 항에 있어서, 검사 샘플에서 소수의 핵산 종의 양을 결정하는 단계를 더 포함하는 것인 방법.

청구항 16

제15항에 있어서, 상기 미세중복 또는 미세결실을 가질 조건부 확률은 소수의 핵산 종의 양 및 (e) 단계의 카운트 표시치에 따라 (f)단계에서 결정되는 것인 방법.

청구항 17

제1항 내지 제4항 중 어느 한 항에 있어서, 상기 검사 샘플은 태아를 가진 임신 대상체로부터이고, 미세중복 또는 미세결실을 가질 조건부 확률은 검사 샘플에 대해 결정된 태아 분획, 검사 샘플에 대한 후보 분절에 대한 카운트 표시치의 z-점수, 및 후보 분절에 대한 카운트 표시치에 대한 z-점수의 태아 분획-특이적 분포에 따라 결정되는 것인 방법.

청구항 18

제17항에 있어서, 미세중복 또는 미세결실을 가질 조건부 확률이 하기 방정식 23의 관계에 의해 결정되는 것인

방법:

방정식 23

$$Z \sim Normal\left(\frac{\mu_X}{\sigma_X}\frac{f}{2}, 1\right)$$

상기 식에서,

f는 태아 분획, X는 후보 분절에 대한 합산된 부분 카운트, $X \sim f(\mu_X, \sigma_X)$이고, 이때 μ_X 및 σ_X는 각각 X의 평균 및 표준 편차이고, $f(\cdot)$는 분포 함수이다.

청구항 19

제17항에 있어서, 상기 방법은

(i) 미세중복 또는 미세결실을 가질 조건부 확률이, 검사 샘플에 대한 후보 분절에 대한 카운트 표시치에 대한 z-점수와 후보 분절에 대한 카운트 표시치에 대한 z-점수의 태아 분획-특이적 분포 사이의 교차점이거나;

(ii) 미세중복 또는 미세결실을 갖지 않을 조건부 확률이 검사 샘플에 대한 후보 분절에 대한 카운트 표시치에 대한 z-점수와 정배수체 샘플에서 후보 분절에 대한 카운트 표시치에 대한 z-점수의 태아 분획-특이적 분포 사이의 교차점이거나;

(iii) 미세중복 또는 미세결실을 가질 선험적 확률 및 미세중복 또는 미세결실을 갖지 않을 선험적 확률이 검사 대상체를 포함하지 않는 다수의 샘플들로부터 결정되는 것이거나; 또는

(iv) (i), (ii) 및 (iii)의 하나 이상의 조합인 방법.

청구항 20

삭제

청구항 21

삭제

청구항 22

삭제

청구항 23

삭제

청구항 24

삭제

청구항 25

삭제

청구항 26

삭제

청구항 27

삭제

청구항 28

삭제

청구항 29

삭제

청구항 30

삭제

청구항 31

삭제

청구항 32

삭제

청구항 33

삭제

청구항 34

삭제

청구항 35

삭제

청구항 36

삭제

청구항 37

삭제

청구항 38

삭제

청구항 39

삭제

청구항 40

삭제

청구항 41

삭제

청구항 42

삭제

청구항 43

삭제

청구항 44

삭제

청구항 45

삭제

청구항 46

삭제

청구항 47

삭제

청구항 48

삭제

청구항 49

삭제

청구항 50

삭제

청구항 51

삭제

청구항 52

삭제

청구항 53

삭제

청구항 54

삭제

청구항 55

삭제

청구항 56

삭제

청구항 57

삭제

청구항 58

삭제

청구항 59

삭제

청구항 60

삭제

청구항 61

삭제

청구항 62

삭제

청구항 63

삭제

청구항 64

삭제

청구항 65

삭제

청구항 66

삭제

청구항 67

삭제

청구항 68

삭제

청구항 69

삭제

청구항 70

삭제

청구항 71

삭제

청구항 72

삭제

청구항 73

삭제

청구항 74

삭제

청구항 75

삭제

청구항 76

삭제

청구항 77

삭제

청구항 78

삭제

청구항 79

삭제

청구항 80

삭제

청구항 81

삭제

청구항 82

삭제

청구항 83

삭제

청구항 84

삭제

청구항 85

삭제

청구항 86

삭제

청구항 87

삭제

청구항 88

삭제

청구항 89

삭제

청구항 90

삭제

청구항 91

삭제

청구항 92

삭제

청구항 93

삭제

청구항 94

삭제

청구항 95

삭제

청구항 96

삭제

청구항 97

삭제

청구항 98

삭제

청구항 99

삭제

청구항 100

삭제

청구항 101

삭제

발명의 설명

기 술 분 야

[0001] 관련 특허 출원

[0002] 본 특허출원은 2013년 5월 24일자로 출원된 미국 가특허출원 제61/827,385호(발명의 명칭: METHODS AND PROCESSES FOR NON-INVASIVE ASSESSMENT OF GENETIC VARIATIONS; 발명자: Zeljko Dzakula et al.; 지정된 변리사 사건 번호 SEQ-6068-PV)의 이익을 주장한다. 본 특허출원은 (i) 2012년 10월 4일자로 출원된 미국 가특허출원 제61/709,899호(발명의 명칭: METHODS AND PROCESSES FOR NON-INVASIVE ASSESSMENT OF GENETIC VARIATIONS; 발명자: Cosmin Deciu, Zeljko Dzakula, Mathias Ehrich and Sung Kim; 지정된 변리사 사건 번호: SEQ-6034-PV3)의 이익을 주장하고; (ii) 2012년 6월 22일자로 출원된 미국 가특허출원 제61/663,477호(발명의 명칭: METHODS AND PROCESSES FOR NON-INVASIVE ASSESSMENT OF GENETIC VARIATIONS; 발명자: Zeljko Dzakula and Mathias Ehrich; 지정된 변리사 사건 번호: SEQ-6034-PV2)의 이익을 주장하고; (iii) 2011년 10월 6일자로 출원된 미국 가특허출원 제61/544,251호(발명의 명칭: METHODS AND PROCESSES FOR NON-INVASIVE ASSESSMENT OF GENETIC VARIATIONS; 발명자: Zeljko Dzakula and Mathias Ehrich; 지정된 변리사 사건 번호: SEQ-6034-PV)의 이익을 주장하는, 2012년 10월 5일자로 출원된 국제 PCT 출원 제PCT/US2012/059123호(발명의 명칭: METHODS AND PROCESSES FOR NON-INVASIVE ASSESSMENT OF GENETIC VARIATIONS; 발명자: Cosmin Deciu, Zeljko Dzakula, Mathias Ehrich and Sung Kim; 지정된 변리사 사건 번호: SEQ-6034-PC)의 계속출원인, 2012년 11월 5일자로 출원된 미국 특허출원 제13/669,136호(발명의 명칭: METHODS AND PROCESSES FOR NON-INVASIVE ASSESSMENT OF GENETIC VARIATIONS; 발명자: Cosmin Deciu, Zeljko Dzakula, Mathias Ehrich and Sung Kim; 지정된 변리사 사건 번호 SEQ-6034-CTt)와 관련되어 있다. 모든 본문, 표 및 도면을 포함하는 상기 출원들의 전체 내용은 본원

에 참고로 도입된다.

[0003] 분야

[0004] 본원에서 제공된 기술은 부분적으로 유전적 변이의 비침습 평가를 위한 방법, 프로세스 및 장치에 관한 것이다.

배 경 기 술

[0005] 살아있는 유기체(예를 들면, 동물, 식물 및 미생물)의 유전 정보 및 다른 형태의 복제 유전 정보(예를 들면, 바이러스)는 데옥시리보핵산(DNA) 또는 리보핵산(RNA)에 코딩되어 있다. 유전 정보는 화학적 또는 가상적 핵산의 일차 구조를 나타내는 뉴클레오티드 또는 변경된 뉴클레오티드의 연속물이다. 인간에서, 완전한 게놈은 이십사(24)개의 염색체 상에 위치하는 약 30,000개의 유전자를 함유한다(문헌(The Human Genome, T. Strachan, BIOS Scientific Publishers, 1992) 참조). 각각의 유전자는 전사 및 번역을 통한 발현 후 살아있는 세포 내에서 특정 생화학적 기능을 수행하는 특정 단백질을 코딩한다.

[0006] 많은 의학적 질환들이 하나 이상의 유전적 변이에 의해 야기된다. 일부 유전적 변이들은 예를 들면, 혈우병, 지중해빈혈, 뒤시엔느 근위축증(DMD), 헌팅톤병(HD), 알쯔하이머병 및 낭성 섬유증(CF)을 포함하는 의학적 질환을 야기한다(Human Genome Mutations, D. N. Cooper and M. Krawczak, BIOS Publishers, 1993). 이러한 유전 질환들은 특정 유전자의 DNA에서 단일 뉴클레오티드의 추가, 치환 또는 결실로부터 발생할 수 있다. 일부 출생 결함은 이배수성(aneuploidy)으로서도 지칭되는 염색체 이상, 예컨대, 삼염색체성 21(다운(Down) 증후군), 삼염색체성 13(파타우(Patau) 증후군), 삼염색체성 18(에드워드(Edward) 증후군), 일염색체성 X(터너(Turner) 증후군) 및 일부 성염색체 이배수성, 예컨대, 클라인펠터(Klinefelter) 증후군(XXY)에 의해 야기된다. 또 다른 유전적 변이는 종종 성염색체 X 및 Y를 기초로 확인될 수 있는 태아 성별이다. 일부 유전적 변이들은 개체가 다수의 질환들, 예를 들면, 당뇨병, 동맥경화증, 비만, 다양한 자가면역 질환들 및 암(예를 들면, 대장, 유방, 난소, 폐) 중 임의의 질환에 취약하게 만들 수 있거나 이러한 질환을 야기할 수 있다.

[0007] 하나 이상의 유전적 변이 또는 변경의 확인은 특정 의학적 질환의 진단 또는 특정 의학적 질환에 대한 소인의 확인으로 이어질 수 있다. 유전적 변이의 확인은 의학적 결정을 용이하게 할 수 있고/있거나 도움이 되는 의학적 절차를 이용하게 할 수 있다. 일부 실시양태에서, 하나 이상의 유전적 변이 또는 변경의 확인은 무세포 DNA의 분석을 포함한다. 무세포 DNA(CF-DNA)는 세포 사멸로부터 유래하고 말초 혈액에서 순환하는 DNA 단편으로 구성된다. 고농도의 CF-DNA는 일부 임상 질환, 예컨대, 암, 외상, 화상, 심근경색, 뇌졸중, 패혈증, 감염 및 다른 질환을 표시할 수 있다. 추가로, 무세포 태아 DNA(CFF-DNA)는 모체 혈류에서 검출될 수 있고 다양한 비침습 출생전 진단을 위해 사용될 수 있다.

발명의 내용

[0008] 일부 양태에서, 본원은 거짓(false) 음성 및 거짓 양성 확인을 감소시키면서 태아에서의 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 방법으로서, (a) 기준 게놈(reference genome)의 부분들에 맵핑된(mapped) 핵산 서열 리드(reads)의 카운트(counts)를 수득하는 단계로서, 이때 상기 서열 리드가 임신 여성 유래의 순환 무세포 핵산의 리드인 단계, (b) 각각의 부분들에 맵핑된 카운트를 정규화하여 계산된 게놈 구획(section) 수준을 제공하는 단계, (c) 계산된 게놈의 구획 수준에 따라 게놈 분절(segment)에 대한 프로파일을 생성하는 단계, (d) 상기 프로파일을 분할하여 2개 이상의 분해 랜더링(decomposition rendering)을 제공하는 단계, 및 (e) 상기 2개 이상의 분해 렌더링에 따라 거짓 음성 및 거짓 양성 확인을 감소시키면서 태아에서의 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 단계를 포함하는 방법을 제공한다.

[0009] 또한, 일부 양태에서, 본원은 거짓 음성 및 거짓 양성 확인을 감소시키면서 웨이블렛 이벤트(wavelet event)의 존재 또는 부재를 확인하는 방법으로서, (a) 기준 게놈의 부분들에 맵핑된 핵산 서열 리드의 카운트를 수득하는 단계로서, 이때 상기 서열 리드가 임신 여성으로부터의 순환 무세포 핵산의 리드인 단계, (b) 각각의 부분들에 맵핑된 카운트를 정규화하여 계산된 게놈 구획 수준을 제공하는 단계, (c) 부분 세트를 다수의 부분 서브세트로 분할하는 단계, (d) 계산된 게놈 구획 수준에 따라 각각의 서브세트에 대한 수준을 측정하는 단계, (e) 각각의 수준에 대한 유의 수준을 측정하는 단계, 및 (f) 각각의 수준에 대해 측정된 유의 수준에 따라 거짓 음성 및 거짓 양성 확인을 감소시키면서 웨이블렛 이벤트의 존재 또는 부재를 확인하는 단계를 포함하는 방법을 제공한다.

[0010] 또한, 일부 양태에서, 본 발명은 거짓 음성 및 거짓 양성 확인을 감소시키면서 태아에서의 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 방법으로서, (a) 기준 게놈의 부분들에 맵핑된 핵산 서열 리드의 카운트를 수득하는 단계로서, 이때 상기 서열 리드가 임신 여성으로부터의 순환 무세포 핵산의 리드인

단계, (b) 각각의 부분들에 맵핑된 카운트를 정규화하여 계산된 게놈 구획 수준을 제공하는 단계, (c) 게놈의 분절을 선택하여 부분 세트를 제공하는 단계, (d) 부분 세트를 반복적으로 분할하여 2개 이상의 부분 서브세트를 제공하는 단계, (e) 상기 2개 이상의 부분 서브세트 각각에 대한 수준을 측정하는 단계, 및 (f) (e)에서 측정된 수준에 따라 거짓 음성 및 거짓 양성 확인을 감소시키면서 샘플에 대해 태아에서의 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 단계를 포함하는 방법을 제공한다.

[0011] 또한, 본원은 하나 이상의 프로세서(processors) 및 메모리(memory)를 포함하는 시스템을 제공하는데, 이때 상기 메모리는 상기 하나 이상의 프로세서에 의해 실행가능한 지시를 포함하고, 상기 메모리는 기준 게놈의 부분들에 맵핑된 핵산 서열 리드의 카운트를 포함하고, 상기 서열 리드는 임신 여성으로부터의 순환 무세포 핵산의 리드이고, 상기 하나 이상의 프로세서에 의해 실행가능한 지시는 (a) 기준 게놈의 부분들에 맵핑된 핵산 서열 리드로서, 임신 여성으로부터의 순환 무세포 핵산의 리드인 핵산 서열 리드의 카운트를 수득하고, (b) 각각의 부분들에 맵핑된 카운트를 정규화하여 계산된 게놈 구획 수준을 제공하고, (c) 계산된 게놈 구획 수준에 따라 게놈의 분절에 대한 프로파일을 생성하고, (d) 상기 프로파일을 분할하여 2개 이상의 분해 렌더링을 제공하고, (e) 상기 2개 이상의 분해 렌더링에 따라 거짓 음성 및 거짓 양성 확인을 감소시키면서 태아에서의 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하도록 구성된다.

[0012] 기술의 일부 양태는 하기 설명, 실시예, 특허청구범위 및 도면에 더 기재되어 있다.

도면의 간단한 설명

[0013] 도면은 기술의 실시양태를 예시하고 한정하지 않는다. 명료함 및 용이한 예시를 위해, 도면은 비례척에 맞춰지지 않았고, 일부 경우 다양한 양태들이 특정 실시양태의 이해를 용이하게 하기 위해 과장되거나 확장된 상태로 표시될 수 있다.

도 1은 웨이블렛 방법의 도식을 보여준다. 정규화된 부분 카운트 데이터(상부 우측 패널)는 웨이블렛에 의해 평활화된 프로파일(하부 우측)을 생성하도록 웨이블렛에 의해 변환되어 있다. 불균일 이벤트는 웨이블렛 노이즈제거(denoising) 후에 명확히 보인다.

도 2는 역치화(thresholding) 없는 평준화의 효과를 보여준다. 최적 수준은 이벤트의 원하는 크기에 의해 결정될 수 있다.

도 3은 13번 염색체에 대한 샘플에 대한 불균일 프로파일(상부) 및 웨이블렛 변환된 프로파일(중간)을 포함하는 프로파일을 보여준다. 하부 패널은 13번 염색체에 대한 다수의 정배수체 기준 샘플들로부터 수득된 널 에지(null edge) 높이 분포를 보여준다. 중간 패널에서, 2개의 큰 차이(원형)는 불균일 이벤트의 경계에 상응한다.

도 4는 웨이블렛 또는 CBS 후 분절 병합의 일례를 보여준다. 3개의 처음 분할된 분절들(좌측 패널, 염색체의 우측 절반)은 단일 긴 스트레치(stretch)(우측 패널, 염색체의 우측 절반)로 병합되어 미세중복이 명확히 보이게 만들었다.

도 5의 A 내지 E는 웨이블렛에 의해 평활화된 염색체 프로파일(도 5의 B), CBS에 의해 평활화된 염색체 프로파일(도 5의 C) 및 분절 병합된 염색체 프로파일(도 5의 D 및 E)을 보여준다. 2종의 방법들로부터 가장 우수한 2개의 분절들을 서로 비교하고 "교차확인"하였다.

도 6a 및 도 6b는 결정 분석의 비한정적 예를 보여준다. 나타낸 순서도의 일부 요소들(예를 들면, 상자)은 임의적이다. 일부 실시양태에서, 추가 요소가 추가된다(예를 들면, 검증).

도 7은 650으로부터 확장된 비교의 비한정적 예를 보여준다.

도 8은 631 및 632로 표시된 2개의 웨이블렛 이벤트의 비교의 비한정적 예를 보여준다.

도 9는 웨이블렛에 의해 평활화되고 병합된 염색체 프로파일(B) 및 CBS에 의해 평활화되고 병합된 염색체 프로파일(C)인 염색체 프로파일(A)을 보여준다. 비교 후, 2종의 방법들로부터 가장 우수한 2개의 분절들은 서로 "교차거부"한다.

도 10은 디조지 증후군(DiGeorge syndrome)과 관련된 유전적 변이와 관련된 22번 염색체의 분절의 프로파일을 보여준다. 디조지 증후군과 관련된 유전 미세결실 및 미세중복은 이 영역으로 맵핑되어 있다. 좌측의 프로파일들(패널 A 내지 G)은 Haar 웨이블렛 및 CBS에 의해 분할되고 평활화되고 병합되고 비교되었다. 복합 프로파일들은 우측 패널(A'-G')에 나타나 있다. 유동 셀 당 샘플 적재량의 차이는 패널 A-A', 0.5-플렉스(plex); B-B', 1-플렉스; C-C', 2-플렉스, D-D', 3-플렉스; E-E', 4-플렉스; F-F', 5-플렉스 및 G-G', 6-플렉스에 나타나 있

다. 디조지 미세결실은 샘플 리드 커버리지(coverage)가 10배 감소하였을 때조차도 검출되었다(예를 들면, 패널 F' 참조).

도 11a는 1번 염색체의 프로파일에서 미세결실의 검출을 나타내는 복합 웨이블렛 이벤트를 보여준다. 도 11b는 2번 염색체의 프로파일에서 미세중복의 검출을 나타내는 복합 웨이블렛 이벤트를 보여준다.

도 12는 최대 엔트로피(entropy) 방법을 이용하여 12번 염색체에서 미세중복의 위치를 검출하는 것을 예시하는 대표적인 예를 보여준다.

도 13은 플레이트 상의 샘플의 위치를 표시하는 한 쌍의 숫자로 표지된 16개의 샘플들에 대한 디조지 영역의 확대도를 보여준다. 샘플 쌍 3_4(하부로부터 두 번째) 및 9_10(하부로부터 다섯 번째)은 치명적인 디조지 임신에 속한다. 모든 다른 샘플들은 정배수체로서 핵형분석되어 있다. 강조된 상자(회색 영역)는 디조지 영역과 PERUN 부분 선택(기준 게놈의 부분 chr22_368-chr22_451) 사이의 중첩을 나타낸다.

도 14는 디조지 영역에 대한 Z-점수를 보여준다. 각각의 데이터 점은 각각의 환자로부터의 2개의 별개의 혈장 분취물들로부터 수득된 2개의 프로파일들의 합계로부터 유도된다. Z-표준화는 영향을 받은 2개의 사례를 포함하는 모든 16명의 환자들을 기초로 수행되었다.

도 15 및 16은 각각 샘플 3_4(디조지) 및 1_2(정배수체)에 대한 대표적인 히스토그램을 보여준다. 각각의 히스토그램은 디조지 영역 내에 함유된 영역들의 15x15 격자 상에서 수득된 Z-점수의 분포를 보여준다. 상기 영역들은 디조지 영역의 좌측 및 우측 에지 둘 다를 외부 에지부터 출발하여 내부로 이동시키면서 한 부분씩 슬라이딩시킴으로써 선택되었다. 샘플 3_4 및 9_10에 대한 히스토그램(도시되어 있지 않음)은 일관되게 고갈을 보였는데, 이때 3_4에 대한 몇몇 Z-점수들만이 Z = -3을 초과하였다. 샘플 13_14에 대한 히스토그램(도시되어 있지 않음)은 일관되게 과다표시(overrepresentation)를 암시하였는데, 이때 몇몇 영역들만이 3 미만의 Z-점수를 제공하였다. 모든 다른 샘플들(예를 들면, 1_2)은 Z-점수의 [-3, 3] 분절 내에 국한되어 있는 상태로 유지되었다.

도 17은 16개 샘플들 각각에 대한 중간 Z-점수 및 이들의 ±3 MAD 신뢰 구간을 보여준다. 각각의 중간 Z-점수는 슬라이딩 에지로부터 수득된 영역들의 15x15 격자(225개의 영역)로부터 측정되었다. 공지된 디조지 샘플들(3_4 및 9_10)에 대한 Z-점수는 압도적인 대다수의 디조지 하위영역들에 대해 -3 미만으로 유지되었다. 샘플 13_14에서의 명백한 중복은 그의 Z-점수가 대체로 3을 초과한다는 사실에 의해 확인되었다. 모든 다른 샘플들의 Z-점수는 항상 [-3, 3] 분절 내에 국한되어 있었다.

도 18 및 19는 샘플 3_4(디조지) 및 1_2(정배수체) 각각에 대한 대표적인 히스토그램을 보여준다. 각각의 히스토그램은 디조지 영역에 대해 수득된 Z-점수의 분포를 보여준다. 각각의 Z-점수는 "리브 원 아웃(leave one out)" 방법을 이용하여 16개의 상이한 기준 샘플 세트들을 사용함으로써 계산되었다. 샘플 9_10에 대한 히스토그램(도시되어 있지 않음)은 고갈을 확인시켜주었다. 기준 세트에 따라 샘플 3_4는 명확히 고갈되었거나 경계선 Z-점수를 가졌다. 샘플 13_14에 대한 히스토그램(도시되어 있지 않음)은 몇몇 경계선 Z-점수와 함께 과다표시를 암시하였다. 1_2를 포함하는 모든 다른 샘플들은 Z-점수의 [-3, 3] 분절 내에 국한된 상태로 유지되었다.

도 20은 각각의 샘플에 대한 중간 Z-점수를 보여주는데, 이의 대표적인 예가 도 18 및 19에 나타나 있다. 중간 Z-점수 및 이들의 ±3 MAD 신뢰 구간은 "리브 원 아웃" 방법을 이용함으로써 확인된 16개의 상이한 기준 샘플 세트들로부터 계산되었다. 공지된 디조지 샘플(3_4 및 9_10)에 대한 Z-점수는 압도적인 대다수의 기준 샘플 서브세트에 대해 -3 미만으로 유지되었다. 샘플 13_14에서의 명백한 중복은 그의 Z-점수가 대체로 3을 초과한다는 사실에 의해 확인되었다. 모든 다른 샘플들의 Z-점수는 항상 [-3, 3] 분절 내에 국한되어 있었다.

도 21은 16개의 샘플들 각각에 대해 "리브 원 아웃" 기법에 의해 생성된 중간 Z-점수(y-축)에 비해 디조지 하위영역들의 15x15 격자를 사용함으로써 수득된 중간 Z-점수(x-축)를 보여준다. 대각선은 이상적인 일치(기울기=1, 절편=0)를 나타낸다.

도 22 및 23은 샘플 3_4(디조지) 및 1_2(정배수체) 각각에 대한 대표적인 히스토그램을 보여준다. 각각의 히스토그램은 16개의 상이한 기준 샘플 세트들을 사용함으로써 디조지 영역의 하위영역에 대해 수득된 Z-점수의 분포를 보여준다. 상기 하위영역은 225개 하위영역들의 15x15 격자로부터 무작위적으로 선택되었다. "리브 원 아웃" 분석은 샘플 3_4(도 37) 및 9_10(도시되어 있지 않음)에 대한 고갈을 확인시켜주었다. 샘플 13_14에 대한 히스토그램은 과다표시를 확인시켜주었다(도시되어 있지 않음). 1_2를 포함하는 모든 다른 샘플들은 Z-점수의 [-3, 3] 분절 내에 국한된 상태로 유지되었다.

도 24는 "리브 원 아웃" 방법을 이용하여 16개의 샘플들 각각에 대한 디조지 영역의 무작위적으로 선택된 하위영역에 대한 중간 Z-점수 및 이들의 ±3 MAD 신뢰 구간을 보여준다. 공지된 디조지 샘플들(3_4 및 9_10)에 대한 Z-점수는 대다수의 기준 샘플들에 대해 -3 미만으로 유지되었다. 샘플 13_14에서의 명백한 중복은 그의 Z-점수가 대체로 3을 초과한다는 사실에 의해 표시되었다. 샘플 17_18을 제외한 모든 다른 샘플들의 Z-점수는 항상 [-3, 3] 분절 내에 국한되어 있었다.

도 25 및 26은 16개의 상이한 기준 샘플 세트들을 사용함으로써 디조지 영역의 모든 225개의 하위영역들에 대해 수득된 Z-점수의 분포를 나타내는, 샘플 3_4(디조지) 및 1_2(정배수체) 각각에 대한 대표적인 히스토그램을 보여준다. 상기 225개의 하위영역들은 각각의 샘플에 대한 15x15 격자 상에서 슬라이딩 에지 방법을 이용함으로써 생성되었다. 슬라이딩 에지는 "리브 원 아웃" 분석과 병용되었다. 결과는 영향을 받은 샘플 3_4 및 9_10 둘 다에 대한 고갈을 확인시켜준다(도시되어 있지 않음). 샘플 13_14에 대한 히스토그램은 과다표시를 확인시켜주었다(도시되어 있지 않음). 17_18을 제외한, 1_2를 포함하는 모든 다른 샘플들이 Z-점수의 [-3, 3] 분절 내에 국한된 상태로 유지되었다(도시되어 있지 않음).

도 27은 15x15 격자 단독으로부터 유도된 중간 Z-점수에 비해 "리브 원 아웃" 기법과 함께 디조지 하위영역의 15x15 격자를 사용함으로써 수득된 중간 Z-점수를 보여준다. 대각선은 이상적인 일치(기울기=1, 절편=0)를 나타낸다.

도 28은 "리브 원 아웃" 기법과 함께 디조지 하위영역의 15x15 격자를 사용함으로써 수득된 Z-점수의 MAD가 15x15 격자 단독으로부터 유도된 Z-점수의 MAD와 비교되었다는 것을 보여준다. 대각선은 이상적인 일치(기울기=1, 절편=0)를 나타낸다.

도 29는 "리브 원 아웃" 방법과 함께 표준 디조지 영역의 하위영역의 완전한 15x15 격자를 사용함으로써 평가된 중간 Z-점수 및 이들의 ±3 MAD 신뢰 구간을 보여준다. 공지된 디조지 샘플들(3_4 및 9_10)에 대한 Z-점수는 대다수의 기준 샘플들에 대해 -3 미만으로 유지되었다. 샘플 13_14에서의 명맥한 중복은 그의 Z-점수가 대체로 3을 초과한다는 사실에 의해 표시되었다. 샘플 17_18을 제외한 모든 다른 샘플들의 Z-점수는 항상 [-3, 3] 분절 내에 국한되어 있었다.

도 30은 기술의 일부 실시양태가 실행될 수 있는 시스템의 예시적인 실시양태를 보여준다.

도 31은 로그 승산비(LOR) 방법을 이용한 LDTv2 남성 샘플에 대한 분류 결과를 보여준다.

도 32는 실시예 6에 기재된 방정식 23의 일부 양태를 그래프로 묘사한다.

도 33은 에파네크니코브 커넬(Epanechnikov kernel)(밴드폭=200 bp)에 의해 제공된 GC 밀도의 한 실시양태를 보여준다.

도 34는 HTRA1 유전자에 대한 GC 밀도(y-축)의 도표를 보여주는데, 이때 GC 밀도는 전체 게놈에 걸쳐 정규화되어 있다. 게놈 위치는 x-축 상에 나타나 있다.

도 35는 기준 게놈(직선) 및 샘플로부터 수득된 서열 리드(쇄선)에 대한 국소 게놈 편향(bias) 추정치(예를 들면, GC 밀도, x-축)의 분포를 보여준다. 편향 빈도(예를 들면, 밀도 빈도)는 y-축 상에 나타나 있다. GC 밀도 추정치는 전체 게놈에 걸쳐 정규화되어 있다. 이 예에서, 샘플은 기준물로부터 예상된 것보다 더 높은 GC 함량을 갖는 더 많은 리드를 갖는다.

도 36은 가중된 3차 다항식 피팅된 관계를 이용하여 기준 게놈에 대한 GC 밀도 추정치와 샘플에 대한 서열 리드의 GC 밀도 추정치의 분포를 비교하여 보여준다. GC 밀도 추정치(x-축)는 전체 게놈에 걸쳐 정규화되었다. GC 밀도 빈도는 샘플의 밀도 빈도로 나누어진 기준물의 밀도 빈도의 log2 비로서 y-축 상에 표시되어 있다.

도 37a는 게놈의 모든 부분들에 대한 중간 GC 밀도(x-축)의 분포를 보여준다. 도 37b는 다수의 샘플들에 대한 GC 밀도 분포에 따라 결정된 중간 절대 편차(MAD) 값(x-축)을 보여준다. GC 밀도 빈도는 y-축 상에 나타나 있다. 부분들은 다수의 기준 샘플들(예를 들면, 트레이닝 세트)에 대한 중간 GC 밀도 분포, 및 다수의 샘플들의 GC 밀도 분포에 따라 측정된 MAD 값에 따라 필터링되었다. 확립된 역치(예를 들면, MAD의 사분위수간 범위의 4배)를 벗어난 GC 밀도를 포함하는 부분들은 필터링 프로세스에 따라 고려사항으로부터 제거되었다.

도 38a는 게놈 내의 각각의 게놈 부분의 중간 리드 밀도(y-축, 예를 들면, 리드 밀도/부분) 및 상대적 위치(x-축, 부분 지수)를 포함하는 게놈에 대한 샘플의 리드 밀도 프로파일을 보여준다. 도 38b는 제1 주성분(PC1)을 보여주고, 도 38c는 500개 정배수체의 트레이닝 세트로부터 수득된 리드 밀도 프로파일의 주성분 분석으로부터

수득된 제2 주성분(PC2)을 보여준다.

도 39a 내지 39c는 (예를 들면, 2개의 수직선으로 묶인) 21번 염색체의 삼염색체성을 포함하는 게놈에 대한 샘플의 리드 밀도 프로파일의 일례를 보여준다. 각각의 게놈 부분의 상대적 위치는 x-축 상에 나타나 있다. 리드 밀도는 y-축 상에서 제공되어 있다. 도 39a는 미처리(raw)(예를 들면, 조절되지 않은) 리드 밀도 프로파일을 보여준다. 도 39b는 중간 프로파일의 빨셈을 포함하는 제1 조절을 포함하는 도 39a의 프로파일을 보여준다. 도 39c는 제2 조절을 포함하는 도 39b의 프로파일을 보여준다. 제2 조절은 이 샘플에서 발견된 그들의 표시치(representation)를 기초로 가중된 8x 주성분 프로파일들의 빨셈을 포함한다(예를 들면, 모델이 구축된다). 예를 들면, 샘플 프로파일 = $A^{*}PC1 + B^{*}PC2 + C^{*}PC3$ … 및 예를 들면, 도 39c에 나타낸 보정된 프로파일 = 샘플프로파일 - $A^{*}PC1 + B^{*}PC2 + C^{*}PC3$ …

도 40은 T21 검사를 위한 부트스트랩핑된(bootstrapped) 트레이닝 샘플로부터의 검사 p-값의 QQ-도표를 보여준다. QQ-도표는 일반적으로 2개의 분포를 비교한다. 도 40은 검사 샘플로부터의 ChAI 점수(y-축)와 균일한 분포(즉, p-값의 예상된 분포, x-축)의 비교를 보여준다. 각각의 점은 단일 검사 샘플의 log-p 값 점수를 나타낸다. 샘플이 분류되고 균일한 분포를 기초로 '예상된' 값(x-축)을 배정받는다. 하부 쇄선은 대각선을 나타내고 상부 선은 본페로니(Bonferroni) 역치를 나타낸다. 균일한 분포를 따르는 샘플은 하부 대각선(하부 쇄선) 상에 있을 것으로 예상될 것이다. 데이터 값은 부분들에서의 상관관계(예를 들면, 편향)로 인해 대각선을 벗어나 있는데, 이것은 예상된 것보다 더 높은 점수(낮은 p-값)를 갖는 샘플을 시사한다. 본원에 기재된 방법(예를 들면, ChAI, 예를 들면, 실시예 7 참조)은 이 관찰된 편향을 보정할 수 있다.

도 41a는 트레이닝 세트에서 남성 및 여성에 대한 PC2 계수의 차이를 보여주는 리드 밀도 도표를 보여준다. 도 41b는 PC2 계수를 사용한 성별 판정을 위한 수신자 조작 특성(ROC) 도표를 보여준다. 서열결정에 의해 수행된 성별 판정은 진정한 기준물을 위해 사용되었다.

도 42a 및 42b는 시스템의 한 실시양태를 보여준다.

발명을 실시하기 위한 구체적인 내용

[0014] 본원은 태아에서 태아 유전적 변이(예를 들면, 염색체 이배수성, 미세중복 또는 미세결실)의 존재 또는 부재를 확인하는 방법을 제공하는데, 이때 확인은 부분적으로 또는 전체적으로 핵산 서열에 따라 수행된다. 일부 실시양태에서, 핵산 서열은 임신 여성으로부터 수득된 샘플로부터(예를 들면, 임신 여성의 혈액으로부터) 수득된다. 또한, 본원은 개선된 데이터 조작 방법뿐만 아니라, 일부 실시양태에서 본원에 기재된 방법을 수행하는 시스템, 장치 및 모듈도 제공한다. 일부 실시양태에서, 본원에 기재된 방법에 의한 유전적 변이의 확인은 특정 의학적 질환의 진단으로 이어질 수 있거나 특정 의학적 질환에 대한 소인을 확인하게 할 수 있다. 유전적 변이의 확인은 의학적 결정을 용이하게 할 수 있고/있거나 도움이 되는 의학적 절차를 이용하게 할 수 있다.

[0015] 샘플

[0016] 본원은 핵산을 분석하는 방법 및 조성물을 제공한다. 일부 실시양태에서, 핵산 단편의 혼합물 중의 핵산 단편이 분석된다. 핵산의 혼합물은 상이한 뉴클레오티드 서열, 상이한 단편 길이, 상이한 기원(예를 들면, 게놈 기원, 태아 대 모체 기원, 세포 또는 조직 기원, 샘플 기원, 대상체 기원 등) 또는 이들의 조합을 갖는 2개 이상의 핵산 단편 종(species)을 포함할 수 있다.

[0017] 본원에 기재된 방법 및 장치에서 사용되는 핵산 또는 핵산 혼합물은 종종 대상체로부터 수득된 샘플로부터 단리된다. 대상체는 인간, 비-인간 동물, 식물, 세균, 진균 또는 원생생물을 포함하나 이들로 한정되지 않는 임의의 살아있거나 살아있지 않는 유기체일 수 있다. 포유동물, 파충류, 조류, 양서류, 어류, 유제류, 반추동물, 솟과 동물(예를 들면, 소), 말과 동물(예를 들면, 말), 염소과 동물 및 양과 동물(예를 들면, 양, 염소), 돼지과 동물(예를 들면, 돼지), 낙타과 동물(예를 들면, 낙타, 라마, 알파카), 원숭이, 유인원(예를 들면, 고릴라, 침팬지), 곰과 동물(예를 들면, 곰), 가금류, 개, 고양이, 마우스, 래트, 물고기, 돌고래, 고래 및 상어를 포함하나 이들로 한정되지 않는 임의의 인간 또는 비-인간 동물이 선택될 수 있다. 대상체는 남성 또는 여성(예를 들면, 여성, 임신 여성)일 수 있다. 대상체는 임의의 연령(예를 들면, 배아, 태아, 유아, 소아, 성체)일 수 있다.

[0018] 핵산은 임의의 종류의 적합한 생물학적 표본 또는 샘플(예를 들면, 검사 샘플)로부터 단리될 수 있다. 샘플 또는 검사 샘플은 대상체 또는 이의 부분(예를 들면, 인간 대상체, 임신 여성, 태아)으로부터 단리되거나 수득된 임의의 표본일 수 있다. 표본의 비한정적 예는 혈액 또는 혈액 생성물(예를 들면, 혈청, 혈장 등), 제대혈, 융모막융모, 양수, 뇌척수액, 척수액, 세척액(예를 들면, 기관지폐포, 위, 복막, 도관, 귀, 관절경), 생검 샘플

(예를 들면, 착상 전 배아로부터의 생검 샘플), 복강천자 샘플, 세포(혈액 세포, 태반 세포, 배아 또는 태아 세포, 태아 유핵 세포 또는 태아 세포 잔류물) 또는 이들의 부분(예를 들면, 미토콘드리아, 핵, 추출물 등), 여성 생식관의 세척물, 소변, 대변, 객담, 침, 코 점액, 전립선액, 세척물, 정액, 림프액, 담즙, 눈물, 땀, 모유, 가슴액 등 또는 이들의 조합물을 포함하나 이들로 한정되지 않는, 대상체로부터의 유체 또는 조직을 포함한다. 일부 실시양태에서, 생물학적 샘플은 대상체로부터의 자궁경부 면봉이다. 일부 실시양태에서, 생물학적 샘플은 혈액 및 종종 혈장 또는 혈청일 수 있다. 본원에서 사용된 용어 "혈액"은 임신 여성 또는 가능한 임신에 대해 검사받는 여성으로부터의 혈액 샘플 또는 제제를 지칭한다. 상기 용어는 통상적으로 정의된 바와 같은 전혈, 혈액 생성물 또는 혈액의 임의의 분획, 예컨대, 혈청, 혈장, 버피 코트 등을 포괄한다. 혈액 또는 이의 분획은 종종 뉴클레오좀(nucleosomes)(예를 들면, 모체 및/또는 태아 뉴클레오좀)을 포함한다. 뉴클레오좀은 핵산을 포함하고 종종 무세포 또는 세포내 뉴클레오좀이다. 혈액은 버피 코트도 포함한다. 버피 코트는 종종 피콜 구배를 이용함으로써 단리된다. 버피 코트는 백혈 세포(예를 들면, 백혈구, T-세포, B-세포, 혈소판 등)를 포함할 수 있다. 일부 실시양태에서, 버피 코트는 모체 및/또는 태아 핵산을 포함한다. 혈액 혈장은 항응고제로 처리된 혈액의 원심분리로부터 생성된 전혈의 분획을 지칭한다. 혈액 혈청은 혈액 샘플이 응고된 후 남아있는 유체의 묽은 부분을 지칭한다. 유체 또는 조직 샘플은 종종 병원 또는 진료소가 일반적으로 따르는 표준 프로토콜에 따라 채취된다. 혈액의 경우, 적절한 양(예를 들면, 3 내지 40 ㎖)의 말초 혈액이 종종 채취되고 준비 전에 또는 후에 표준 절차에 따라 저장될 수 있다. 핵산이 추출되는 유체 또는 조직 샘플은 무세포(예를 들면, 세포 무함유) 상태일 수 있다. 일부 실시양태에서, 유체 또는 조직 샘플은 세포 요소 또는 세포 잔류물을 함유할 수 있다. 일부 실시양태에서, 태아 세포 또는 암세포가 샘플에 포함될 수 있다.

[0019] 샘플은 종종 불균질한데, 이것은 1종 초과의 핵산 종이 샘플에 존재한다는 것을 의미한다. 예를 들면, 불균질한 핵산은 (i) 태아로부터 유래된 핵산 및 모체로부터 유래된 핵산, (ii) 암 핵산 및 비-암 핵산, (iii) 병원체 핵산 및 숙주 핵산, 및 보다 일반적으로 (iv) 돌연변이된 핵산 및 야생형 핵산을 포함할 수 있으나 이들로 한정되지 않는다. 샘플은 1종 초과의 세포, 예컨대, 태아 세포 및 모체 세포, 암세포 및 비-암세포, 또는 병원체 및 숙주 세포가 존재하기 때문에 불균질할 수 있다. 일부 실시양태에서, 소수의 핵산 종 및 다수의 핵산 종이 존재한다.

[0020] 본원에 기재된 기술의 출생전 적용을 위해, 유체 또는 조직 샘플은 검사에 적합한 임신 연령의 여성, 또는 가능한 임신에 대해 검사받는 여성으로부터 채취될 수 있다. 적합한 임신 연령은 수행되는 출생전 검사에 따라 달라질 수 있다. 일부 실시양태에서, 임신 여성 대상체는 종종 임신의 첫 번째 3개월, 종종 임신의 두 번째 3개월, 또는 종종 임신의 세 번째 3개월에 있다. 일부 실시양태에서, 유체 또는 조직은 태아 임신의 약 1주 내지 약 45주(예를 들면, 태아 임신의 1주 내지 4주, 4주 내지 8주, 8주 내지 12주, 12주 내지 16주, 16주 내지 20주, 20주 내지 24주, 24주 내지 28주, 28주 내지 32주, 32주 내지 36주, 36주 내지 40주, 또는 40주 내지 44주), 종종 태아 임신의 약 5주 내지 약 28주(예를 들면, 태아 임신의 6주, 7주, 8주, 9주, 10주, 11주, 12주, 13주, 14주, 15주, 16주, 17주, 18주, 19주, 20주, 21주, 22주, 23주, 24주, 25주, 26주 또는 27주)에서 임신 여성으로부터 채취된다. 일부 실시양태에서, 유체 또는 조직 샘플은 분만(예를 들면, 질 분만 또는 비-질 분만(예를 들면, 수술적 분만)) 동안 또는 직후에 임신 여성으로부터 채취된다.

[0021] 혈액 샘플의 획득 및 DNA의 추출

[0022] 본원의 방법은 종종 모체 및/또는 태아 유전적 변이의 존재 또는 부재를 검출하고/하거나 임신 동안 및 종종 후에 태아 및/또는 임신 여성의 건강을 모니터링하기 위해 비침습 수단으로서 모체 혈액에서 발견된 태아 DNA를 분리하고 농축하고 분석하는 단계를 포함한다. 따라서, 본원의 일부 방법들을 실시하는 제1 단계는 종종 임신 여성으로부터 혈액 샘플을 수득하고 샘플로부터 DNA를 추출하는 것을 포함한다.

[0023] 혈액 샘플의 획득

[0024] 혈액 샘플은 본 기술의 방법을 이용한 검사에 적합한 임신 연령의 임신 여성으로부터 수득될 수 있다. 적합한 임신 연령은 하기 논의된 바와 같이 검사되는 장애에 따라 달라질 수 있다. 여성으로부터의 혈액의 채취는 종종 병원 또는 진료소가 일반적으로 따르는 표준 프로토콜에 따라 수행된다. 적절한 양, 예를 들면, 전형적으로 5 내지 50 ㎖의 말초 혈액이 종종 채취되고 추가 준비 전에 표준 절차에 따라 저장될 수 있다. 혈액 샘플은 샘플에 존재하는 핵산의 분해 또는 질 손상을 최소화하는 방식으로 채취될 수 있거나, 저장될 수 있거나 수송될 수 있다.

[0025] 혈액 샘플의 준비

[0026] 모체 혈액에서 발견된 태아 DNA의 분석은 예를 들면, 전혈, 혈청 또는 혈장을 사용함으로써 수행될 수 있다. 모체 혈액으로부터 혈청 또는 혈장을 준비하는 방법은 공지되어 있다. 예를 들면, 임신 여성의 혈액은 혈액 응고를 방지하기 위해 EDTA를 함유하는 튜브 또는 특수 상업용 제품, 예컨대, 바큐테이너(Vacutainer) SST(벡톤 딕킨슨(Becton Dickinson), 미국 뉴저지주 프랭클린 레이크스 소재) 내에 놓여질 수 있고, 그 후 혈장이 원심분리를 통해 전혈로부터 수득될 수 있다. 혈청은 혈액 응고 후 원심분리를 이용하거나 이용하지 않고 수득될 수 있다. 원심분리가 이용되는 경우, 원심분리는 (배타적으로는 아니지만) 전형적으로 적절한 속도, 예를 들면, 1,500xg 내지 3,000xg에서 수행된다. 혈장 또는 혈청은 DNA 추출을 위해 새로운 튜브로 옮겨지기 전에 추가 원심분리 단계로 처리될 수 있다.

[0027] DNA는 전혈의 무세포 부분 이외에 여성으로부터의 전혈 샘플의 원심분리 및 혈장의 제거 후 수득될 수 있는, 버피 코트 부분이 풍부한 세포 분획으로부터도 회수될 수 있다.

[0028] DNA의 추출

[0029] 혈액을 포함하는 생물학적 샘플로부터 DNA를 추출하는 다수의 공지된 방법들이 존재한다. (예를 들면, 문헌(Sambrook and Russell, Molecular Cloning: A Laboratory Manual 3d ed., 2001)에 기재된) 일반적인 DNA 준비 방법을 따를 수 있고; 다양한 상업적으로 입수가능한 시약들 또는 키트들, 예컨대, 퀴아젠(Qiagen)의 QIAamp 순환 핵산 키트, QiaAmp DNA 미니 키트 또는 QiaAmp DNA 혈액 미니 키트(퀴아젠, 독일 힐덴 소재), 게노믹프렙(GenomicPrep)™ 혈액 DNA 단리 키트(프로메가(Promega), 미국 위스콘신주 매디슨 소재), 및 GFX™ 게놈 혈액 DNA 정제 키트(아머샴(Amersham), 미국 뉴저지주 피스카타웨이 소재)도 임신 여성으로부터의 혈액 샘플로부터 DNA를 수득하는 데에 사용될 수 있다. 이 방법들 중 1종 초과의 방법들이 병용될 수도 있다.

[0030] 일부 실시양태에서, 샘플은 먼저 1종 이상의 방법에 의해 태아 핵산에 대해 농축될 수 있거나 상대적으로 농축될 수 있다. 예를 들면, 본 기술의 조성물 및 프로세스를 단독으로 또는 다른 식별 인자들과 함께 사용하여 태아 DNA와 모체 DNA의 식별을 수행할 수 있다. 이 인자들의 예는 X 염색체와 Y 염색체 사이의 단일 뉴클레오티드 차이, Y 염색체 특이적 서열, 게놈의 다른 부분에 위치하는 다형성, 태아 DNA와 모체 DNA 사이의 크기 차이, 및 모체 조직과 태아 조직 사이의 메틸화 패턴의 차이를 포함하나 이들로 한정되지 않는다.

[0031] 특정 핵산 종에 대해 샘플을 농축하는 다른 방법은 2007년 5월 30일자로 출원된 PCT 특허출원 제PCT/US07/69991호, 2007년 6월 15일자로 출원된 PCT 특허출원 제PCT/US2007/071232호, 미국 가특허출원 제60/968,876호 및 미국 가특허출원 제60/968,878호(출원인에게 양도됨), 및 2005년 11월 28일자로 출원된 PCT 특허출원 제PCT/EP05/012707호(이들 모두 본원에 참고로 도입됨)에 기재되어 있다. 일부 실시양태에서, 모체 핵산은 샘플로부터 (부분적으로, 실질적으로, 거의 완전히 또는 완전히) 선택적으로 제거된다.

[0032] 용어 "핵산" 및 "핵산 분자"는 본 개시내용 전체에서 상호교환적으로 사용될 수 있다. 상기 용어들은 임의의 조성의 핵산, 예컨대, DNA(예를 들면, 상보적 DNA(cDNA), 게놈 DNA(gDNA) 등), RNA(예를 들면, 메신저 RNA(mRNA), 짧은 억제 RNA(siRNA), 리보좀 RNA(rRNA), tRNA, 마이크로RNA, 태아 또는 태반에 의해 고도로 발현된 RNA 등), 및/또는 DNA 또는 RNA 유사체(예를 들면, 염기 유사체, 당 유사체 및/또는 비천연 골격 등을 함유함), RNA/DNA 하이브리드 및 폴리아미드 핵산(PNA)(이들 모두 단일 또는 이중 가닥 형태로 존재할 수 있음)을 지칭하고, 달리 한정되어 있지 않은 한, 천연 발생 뉴클레오티드와 유사한 방식으로 작용할 수 있는 천연 뉴클레오티드의 공지된 유사체를 포괄할 수 있다. 일부 실시양태에서, 핵산은 시험관내 또는 숙주 세포, 세포, 세포핵 또는 세포의 세포질에서 복제할 수 있거나 복제될 수 있는 플라스미드, 파지, 자가 복제 서열(ARS), 센트로미어(centromere), 인공 염색체, 염색체 또는 다른 핵산일 수 있거나 이들로부터 유래될 수 있다. 일부 실시양태에서, 주형 핵산은 단일 염색체로부터 유래될 수 있다(예를 들면, 핵산 샘플은 이배수체 유기체로부터 수득된 샘플의 한 염색체로부터 유래될 수 있다). 달리 구체적으로 한정되어 있지 않은 한, 상기 용어는 기준 핵산과 유사한 결합 성질을 갖고 천연 발생 뉴클레오티드와 유사한 방식으로 대사되는 천연 뉴클레오티드의 공지된 유사체를 함유하는 핵산을 포괄한다. 달리 표시되어 있지 않은 한, 특정 핵산 서열은 명시적으로 표시된 서열뿐만 아니라 이의 보존적으로 변경된 변이체(예를 들면, 축퇴 코돈 치환), 대립형질, 오르톨로그(orthologs), 단일 뉴클레오티드 다형성(SNP) 및 상보적 서열도 함축적으로 포괄한다. 구체적으로, 축퇴 코돈 치환은 하나 이상의 선택된 (또는 모든) 코돈의 제3 위치가 혼합된 염기 및/또는 데옥시이노신 잔기로 치환되어 있는 서열을 생성함으로써 달성될 수 있다. 용어 핵산은 좌위, 유전자, cDNA, 및 유전자에 의해 코딩된 mRNA와 상호교환적으로 사용된다. 상기 용어는 뉴클레오티드 유사체로부터 합성된 RNA 또는 DNA의 등가물, 유도체, 변이체 및 유사체로서 단일 가닥("센스" 또는 "안티센스", "플러스" 가닥 또는 "마이너스" 가닥, "정방향" 리딩 프레임 또는 "역방향" 리딩 프레임) 및 이중 가닥 폴리뉴클레오티드도 포함할 수 있다. 용어 "유전자"는 폴리펩티드 쇄를 생성하는 데

에 관여하는 DNA의 분절을 의미하고; 상기 용어는 개별 코딩 분절(엑손) 사이의 개재 서열(인트론)뿐만 아니라 유전자 생성물의 전사/번역 및 전사/번역의 조절에 관여하는, 코딩 서열 앞 및 뒤의 영역(리더 및 트레일러)도 포함한다.

[0033] 데옥시리보뉴클레오티드는 데옥시아데노신, 데옥시사이티딘, 데옥시구아노신 및 데옥시타이미딘을 포함한다. RNA의 경우, 염기 사이토신은 우라실로 대체된다. 주형 핵산은 대상체로부터 수득된 핵산을 주형으로서 사용함으로써 제조될 수 있다.

[0034] 핵산 단리 및 프로세싱

[0035] 핵산은 당분야에서 공지된 방법에 의해 하나 이상의 공급원(예를 들면, 세포, 혈청, 혈장, 버피 코트, 림프액, 피부, 토양 등)으로부터 유래될 수 있다. 임의의 적합한 방법이 생물학적 샘플(예를 들면, 혈액 또는 혈액 생성물)로부터 DNA를 단리하고/하거나, 추출하고/하거나 정제하는 데에 사용될 수 있고, 이러한 방법의 비한정적 예는 (예를 들면, 문헌(Sambrook and Russell, Molecular Cloning: A Laboratory Manual 3d ed., 2001)에 기재된) DNA 준비 방법, 다양한 상업적으로 입수가능한 시약들 또는 키트들, 예컨대, 퀴아젠의 QIAamp 순환 핵산 키트, QIAamp DNA 미니 키트 또는 QIAamp DNA 혈액 미니 키트(퀴아젠, 독일 힐덴 소재), 게노믹프렙™ 혈액 DNA 단리 키트(프로메가, 미국 위스콘신주 매디슨 소재), 및 GFX™ 게놈 혈액 DNA 정제 키트(아머샴, 미국 뉴저지주 피스카타웨이 소재) 등 또는 이들의 조합을 포함한다.

[0036] 세포 용해 절차 및 시약은 당분야에서 공지되어 있고 일반적으로 화학적(예를 들면, 세제, 저장성 용액, 효소적 절차 등, 또는 이들의 조합), 물리적(예를 들면, 프렌치 프레스(French press), 초음파처리 등), 또는 전기용해적 용해 방법에 의해 수행될 수 있다. 임의의 적합한 용해 절차가 이용될 수 있다. 예를 들면, 화학적 방법은 일반적으로 용해제를 사용하여 세포를 파괴하고 세포로부터 핵산을 추출한 후 카오트로픽(chaotropic) 염으로 처리한다. 물리적 방법, 예컨대, 동결/해동 후 분쇄, 세포 프레스의 사용 등도 유용하다. 높은 염 용해 절차도 통상적으로 이용된다. 예를 들면, 알칼리성 용해 절차가 이용될 수 있다. 후자 절차는 전통적으로 페놀-클로로포름 용액의 사용을 도입하고, 3개의 용액을 사용하는 대안적 페놀-클로로포름 무함유 절차도 이용될 수 있다. 후자 절차에서, 한 용액은 15 mM 트리스(Tris)(pH 8.0), 10 mM EDTA 및 100 ㎍/㎖ Rnase A를 함유할 수 있고; 제2 용액은 0.2 N NaOH 및 1% SDS를 함유할 수 있고; 제3 용액은 3 M KOAc(pH 5.5)를 함유할 수 있다. 이 절차들은 본원에 전체적으로 도입된 문헌(Molecular Biology, John Wiley & Sons, N.Y., 6.3.1-6.3.6 (1989))에서 찾아볼 수 있다.

[0037] 핵산은 또 다른 핵산에 비해 상이한 시점에서 단리될 수 있고, 이때 각각의 샘플은 동일한 또는 상이한 공급원으로부터 유래된다. 핵산은 예를 들면, 핵산 라이브러리, 예컨대, cDNA 또는 RNA 라이브러리로부터 유래될 수 있다. 핵산은 샘플로부터의 핵산 분자의 핵산 정제 또는 단리 및/또는 증폭의 결과일 수 있다. 본원에 기재된 프로세스를 위해 제공된 핵산은 1개 샘플 또는 2개 이상의 샘플(예를 들면, 1개 이상, 2개 이상, 3개 이상, 4개 이상, 5개 이상, 6개 이상, 7개 이상, 8개 이상, 9개 이상, 10개 이상, 11개 이상, 12개 이상, 13개 이상, 14개 이상, 15개 이상, 16개 이상, 17개 이상, 18개 이상, 19개 이상 또는 20개 이상의 샘플)로부터의 핵산을 함유할 수 있다.

[0038] 일부 실시양태에서, 핵산은 세포외 핵산을 포함할 수 있다. 본원에서 사용된 용어 "세포외 핵산"은 세포를 실질적으로 갖지 않는 공급원으로부터 단리된 핵산을 지칭할 수 있고, "무세포" 핵산 및/또는 "무세포 순환" 핵산으로서도 지칭된다. 세포외 핵산은 혈액(예를 들면, 임신 여성의 혈액)에 존재할 수 있고 이 혈액으로부터 수득될 수 있다. 세포외 핵산은 종종 검출가능한 세포를 포함하지 않고 세포 요소 또는 세포 잔류물을 함유할 수 있다. 세포외 핵산을 위한 무세포 공급원의 비한정적 예는 혈액, 혈액 혈장, 혈액 혈청 및 소변이다. 본원에서 사용된 용어 "무세포 순환 샘플 핵산을 수득하는"은 샘플을 직접적으로 수득하는 것(예를 들면, 샘플, 예를 들면, 검사 샘플을 채취하는 것) 또는 샘플을 채취한 또 다른 대상체로부터 샘플을 수득하는 것을 포함한다. 이론에 의해 한정되지 않지만, 세포외 핵산은 종종 광범위한 일련의 길이를 갖는 세포외 핵산(예를 들면, "래더(ladder)")을 위한 기초를 제공하는, 세포 아폽토시스 및 세포 분해의 생성물일 수 있다.

[0039] 일부 실시양태에서, 세포외 핵산은 상이한 핵산 종들을 포함할 수 있으므로 본원에서 "불균질한" 세포외 핵산으로서 지칭된다. 예를 들면, 암을 갖는 사람으로부터의 혈액 혈청 또는 혈장은 암세포로부터의 핵산 및 비-암세포로부터의 핵산을 포함할 수 있다. 또 다른 예에서, 임신 여성으로부터의 혈액 혈청 또는 혈장은 모체 핵산 및 태아 핵산을 포함할 수 있다. 일부 경우, 태아 핵산은 종종 전체 핵산의 약 5% 내지 약 50%이다(예를 들면, 총 핵산의 약 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%,

43%, 44%, 45%, 46%, 47%, 48% 또는 49%가 태아 핵산이다). 일부 실시양태에서, 핵산에서 대다수의 태아 핵산은 약 500개 이하의 염기쌍, 약 250개 이하의 염기쌍, 약 200개 이하의 염기쌍, 약 150개 이하의 염기쌍, 약 100개 이하의 염기쌍, 약 50개 이하의 염기쌍 또는 약 25개 이하의 염기쌍의 길이를 갖는다.

[0040] 일부 실시양태에서, 핵산은 핵산을 함유하는 샘플(들)의 프로세싱 없이 본원에 기재된 방법을 수행하기 위해 제공될 수 있다. 일부 실시양태에서, 핵산은 핵산을 함유하는 샘플(들)의 프로세싱 후 본원에 기재된 방법을 수행하기 위해 제공된다. 예를 들면, 핵산은 샘플(들)로부터 추출될 수 있거나, 단리될 수 있거나, 정제될 수 있거나, 부분적으로 정제될 수 있거나 증폭될 수 있다. 본원에서 사용된 용어 "단리된"은 그의 최초 환경(예를 들면, 천연 발생 핵산인 경우 천연 환경 또는 외생적으로 발현된 경우 숙주 세포)으로부터 제거된 핵산을 지칭하므로, 그의 최초 환경으로부터 인간 중재에 의해(예를 들면, "인간의 손에 의해") 변경된다. 본원에서 사용된 용어 "단리된 핵산"은 대상체(예를 들면, 인간 대상체)로부터 제거된 핵산을 지칭할 수 있다. 단리된 핵산은 공급원 샘플에 존재하는 성분의 양보다 더 적은 비-핵산 성분(예를 들면, 단백질, 지질)을 구비할 수 있다. 단리된 핵산을 포함하는 조성물은 약 50% 내지 99% 초과의 비-핵산 성분을 함유하지 않을 수 있다. 단리된 핵산을 포함하는 조성물은 약 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% 또는 99% 초과의 비-핵산 성분을 함유하지 않을 수 있다. 본원에서 사용된 용어 "정제된"은 핵산을 정제 절차로 처리하기 전에 존재하는 비-핵산 성분의 양보다 더 적은 비-핵산 성분(예를 들면, 단백질, 지질, 탄수화물)을 함유하는 제공된 핵산을 지칭할 수 있다. 정제된 핵산을 포함하는 조성물은 약 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% 또는 99% 초과의 다른 비-핵산 성분을 함유하지 않을 수 있다. 본원에서 사용된 용어 "정제된"은 핵산이 유래된 샘플 공급원에 존재하는 핵산 종보다 더 적은 핵산 종을 함유하는 제공된 핵산을 지칭할 수 있다. 정제된 핵산을 포함하는 조성물은 약 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% 또는 99% 초과의 다른 핵산 종을 함유하지 않을 수 있다. 예를 들면, 태아 핵산은 모체 핵산 및 태아 핵산을 포함하는 혼합물로부터 정제될 수 있다. 일부 예에서, 태아 핵산의 작은 단편을 포함하는 뉴클레오좀은 모체 핵산의 보다 더 큰 단편을 포함하는 보다 더 큰 뉴클레오좀 복합체의 혼합물로부터 정제될 수 있다.

[0041] 일부 실시양태에서, 핵산은 본원에 기재된 방법 전, 동안 또는 후에 단편화되거나 절단된다. 단편화된 또는 절단된 핵산은 약 5개 내지 약 10,000개 염기쌍, 약 100개 내지 약 1,000개 염기쌍, 약 100개 내지 약 500개 염기쌍, 또는 약 10개, 15개, 20개, 25개, 30개, 35개, 40개, 45개, 50개, 55개, 60개, 65개, 70개, 75개, 80개, 85개, 90개, 95개, 100개, 200개, 300개, 400개, 500개, 600개, 700개, 800개, 900개, 1000개, 2000개, 3000개, 4000개, 5000개, 6000개, 7000개, 8000개 또는 9000개 염기쌍의 공칭, 평균치(average) 또는 평균(mean) 길이를 가질 수 있다. 단편은 당분야에서 공지된 적합한 방법에 의해 생성될 수 있고, 핵산 단편의 평균치, 평균 또는 공칭 길이는 적절한 단편 생성 절차를 선택함으로써 조절될 수 있다.

[0042] 핵산 단편은 중첩 뉴클레오티드 서열을 함유할 수 있고, 이러한 중첩 서열은 단편화되지 않은 대응물 핵산 또는 이의 분절의 뉴클레오티드 서열의 구축을 용이하게 할 수 있다. 예를 들면, 한 단편은 하위서열 x 및 y를 가질 수 있고, 또 다른 단편은 하위서열 y 및 z를 가질 수 있고, 이때 x, y 및 z는 5개 이상의 뉴클레오티드의 길이를 가질 수 있는 뉴클레오티드 서열이다. 일부 실시양태에서, 중첩 서열 y는 샘플로부터의 핵산에서 x-y-z 뉴클레오티드 서열의 구축을 용이하게 하는 데에 사용될 수 있다. 일부 실시양태에서, 핵산은 (예를 들면, 불완전한 또는 종결된 특이적 절단 반응으로부터) 부분적으로 단편화될 수 있거나 전체적으로 단편화될 수 있다.

[0043] 일부 실시양태에서, 핵산은 적합한 방법에 의해 단편화되거나 절단되고, 이러한 방법의 비한정적 예는 물리적 방법(예를 들면, 전단, 예를 들면, 초음파처리, 프렌치 프레스, 가열, UV 방사선조사 등), 효소적 방법(예를 들면, 효소적 절단제(예를 들면, 적합한 뉴클레아제(nuclease), 적합한 제한 효소, 적합한 메틸화 민감성 제한 효소)), 화학적 방법(예를 들면, 알킬화, DMS, 피페리딘, 산 가수분해, 염기 가수분해, 가열 등 또는 이들의 조합), 미국 특허출원 공보 제20050112590호에 기재된 공정 등 또는 이들의 조합을 포함한다.

[0044] 본원에서 사용된 바와 같이, "단편화" 또는 "절단"은 핵산 분자, 예컨대, 핵산 주형 유전자 분자 또는 이의 증폭된 생성물이 2개 이상의 보다 더 작은 핵산 분자들로 잘라질 수 있는 절차 또는 조건을 지칭한다. 이러한 단편화 또는 절단은 서열 특이적, 염기 특이적 또는 비특이적일 수 있고, 예를 들면, 화학적, 효소적 또는 물리적 단편화를 포함하는 임의의 다양한 방법, 시약 또는 조건에 의해 달성될 수 있다.

[0045] 본원에서 사용된 바와 같이, "단편", "절단 생성물", "절단된 생성물" 또는 이의 문법적 파생어는 핵산 주형 유전자 분자 또는 이의 증폭된 생성물의 단편화 또는 절단으로부터 생성된 핵산 분자를 지칭한다. 이러한 단편 또는 절단된 생성물은 절단 반응으로부터 생성된 모든 핵산 분자들을 지칭할 수 있지만, 전형적으로 이러한 단편 또는 절단된 생성물은 핵산 주형 유전자 분자, 또는 핵산 주형 유전자 분자의 상응하는 뉴클레오티드 서열을 함

유하는 이의 증폭된 생성물의 분절의 단편화 또는 절단으로부터 생성된 핵산 분자만을 지칭한다. 본원에서 사용된 용어 "증폭된"은 표적 핵산 또는 이의 분절과 동일한 또는 실질적으로 동일한 뉴클레오티드 서열을 갖는 앰플리콘 핵산을 선형적으로 또는 기하급수적으로 생성하는 공정으로 샘플 중의 표적 핵산을 처리하는 것을 지칭한다. 일부 실시양태에서, 용어 "증폭된"은 중합효소 연쇄 반응(PCR)을 포함하는 방법을 지칭한다. 예를 들면, 증폭된 생성물은 핵산 주형 서열의 증폭된 뉴클레오티드 영역보다 하나 이상 더 많은 뉴클레오티드를 함유할 수 있다(예를 들면, 프라이머는 핵산 주형 유전자 분자에 상보적인 뉴클레오티드들 이외에 "가외의" 뉴클레오티드, 예컨대, 전사 개시 서열을 함유하여, "가외의" 뉴클레오티드 또는 핵산 주형 유전자 분자의 증폭된 뉴클레오티드 영역에 상응하지 않는 뉴클레오티드를 함유하는 증폭된 생성물을 생성할 수 있다). 따라서, 단편은 대표적인 핵산 주형 분자로부터의 또는 이러한 분자에 기초한 뉴클레오티드 서열 정보를 적어도 부분적으로 함유하는 증폭된 핵산 분자의 분절 또는 부분으로부터 생성된 단편을 포함할 수 있다.

[0046] 본원에서 사용된 바와 같이, 용어 "상보적 절단 반응"은 동일한 표적 또는 기준 핵산 또는 단백질의 대안적 절단 패턴이 생성되도록 상이한 절단 시약을 사용하거나 동일한 절단 시약의 절단 특이성을 변경시킴으로써 동일한 핵산에 대해 수행되는 절단 반응을 지칭한다. 일부 실시양태에서, 핵산은 하나 이상의 반응 용기에서 하나 이상의 특이적 절단제(예를 들면, 1개, 2개, 3개, 4개, 5개, 6개, 7개, 8개, 9개 또는 10개 이상의 특이적 절단제)로 처리될 수 있다(예를 들면, 핵산은 별개의 용기에서 각각의 특이적 절단제로 처리된다). 본원에서 사용된 용어 "특이적 절단제"는 하나 이상의 특이적 부위에서 핵산을 절단할 수 있는 물질, 종종 화학물질 또는 효소를 지칭한다.

[0047] 핵산은 본원에 기재된 방법을 위해 핵산을 제공하기 전에 핵산에서 일부 뉴클레오티드들을 변경시키는 공정에 노출될 수 있다. 예를 들면, 핵산 내의 뉴클레오티드의 메틸화 상태를 기초로 핵산을 선택적으로 변경시키는 공정이 핵산에 적용될 수 있다. 추가로, 조건, 예컨대, 높은 온도, 자외선 방사선조사, x-방사선조사 등은 핵산 분자의 서열에서 변화를 유도할 수 있다. 핵산은 적합한 핵산 분석을 수행하기에 유용한 임의의 적합한 형태로 제공될 수 있다.

[0048] 핵산은 단일 가닥 또는 이중 가닥일 수 있다. 예를 들면, 단일 가닥 DNA는 예를 들면, 가열 또는 알칼리를 사용한 처리로 이중 가닥 DNA를 변성시킴으로써 생성될 수 있다. 일부 실시양태에서, 핵산은 올리고뉴클레오티드 또는 DNA 유사 분자, 예컨대, 펩티드 핵산(PNA)에 의한 이중체 DNA 분자의 가닥 침습에 의해 형성된 D 루프(loop) 구조로 존재한다. D 루프 형성은 예를 들면, 당분야에서 공지된 방법을 이용한 이. 콜라이(*E. Coli*) RecA 단백질의 첨가 및/또는 염 농도의 변경에 의해 용이해질 수 있다.

[0049] 태아 핵산 함량의 측정

[0050] 일부 실시양태에서, 핵산 중의 태아 핵산의 양(예를 들면, 농도, 상대적 양, 절대적 양, 카피 수 등)이 측정된다. 일부 실시양태에서, 샘플 중의 태아 핵산의 양은 "태아 분획(fetal fraction)"으로서 지칭된다. 일부 실시양태에서, "태아 분획"은 임신 여성으로부터 수득된 샘플(예를 들면, 혈액 샘플, 혈청 샘플, 혈장 샘플) 중의 순환 무세포 핵산 중의 태아 핵산의 분획을 지칭한다. 일부 실시양태에서, 태아 핵산의 양은 남성 태아에 대한 특이성을 갖는 마커(예를 들면, Y 염색체 STR 마커(예를 들면, DYS 19, DYS 385, DYS 392 마커); RhD 음성 여성에서 RhD 마커) 또는 다형성 서열의 대립형질 비에 따라, 또는 태아 핵산에 대한 특이성을 갖지만 모체 핵산에 대한 특이성을 갖지 않는 하나 이상의 마커(예를 들면, 모체와 태아 사이의 상이한 유전외적(epigenetic) 생체마커(예를 들면, 메틸화; 하기 더 상세히 기재됨), 또는 모체 혈액 혈장 중의 태아 RNA 마커에 따라 측정된다(예를 들면, 문헌(Lo, 2005, Journal of Histochemistry and Cytochemistry 53 (3): 293-296) 참조).

[0051] 태아 핵산 함량(예를 들면, 태아 분획)의 측정은 종종 예를 들면, 본원에 참고로 도입되는 미국 특허출원 공보 제2010/0105049호에 기재된 태아 정량자 분석(FQA)을 이용함으로써 수행된다. 이러한 종류의 분석은 샘플 중의 핵산의 메틸화 상태를 기초로 모체 샘플에서 태아 핵산을 검출하고 정량할 수 있게 한다. 일부 실시양태에서, 모체 샘플로부터의 태아 핵산의 양은 존재하는 핵산의 총량에 비해 상대적으로 측정되어 샘플 중의 태아 핵산의 백분율을 제공할 수 있다. 일부 실시양태에서, 태아 핵산의 카피 수가 모체 샘플에서 측정될 수 있다. 일부 실시양태에서, 태아 핵산의 양은 정확한 염색체 용량 분석을 가능하게 할 정도로(예를 들면, 태아 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 검출할 정도로) 서열 특이적(또는 부분 특이적) 방식으로 종종 충분한 민감도로 측정될 수 있다.

[0052] 태아 정량자 분석(FQA)은 본원에 기재된 방법들 중 임의의 방법과 함께 수행될 수 있다. 이러한 분석은 당분야에서 공지되어 있고/있거나 미국 특허출원 공보 제2010/0105049호에 기재되어 있는 임의의 방법, 예를 들면, 상이한 메틸화 상태를 기초로 모체 DNA와 태아 DNA를 식별할 수 있는 방법에 의해 수행될 수 있고 태아 DNA를 정

량(즉, 양을 측정)할 수 있다. 메틸화 상태를 기초로 핵산을 식별하는 방법은 예를 들면, MBD2의 메틸 결합 도메인이 항체의 Fc 단편에 융합되어 있는(MBD-FC) MBD2-Fc 단편(Gebhard et al. (2006) Cancer Res. 66(12):6118-28)을 사용한 메틸화 민감성 포획; 메틸화 특이적 항체; 중아황산 전환 방법, 예를 들면, MSP(메틸화 민감성 PCR), COBRA, 메틸화 민감성 단일 뉴클레오티드 프라이머 연장(Ms-SNuPE) 또는 시퀘놈 매스클레이브(Sequenom MassCLEAVE)™ 기술; 및 메틸화 민감성 제한 효소의 사용(예를 들면, 하나 이상의 메틸화 민감성 제한 효소를 사용하여 모체 샘플 중의 모체 DNA를 분해하여 태아 DNA를 농축하는 것)을 포함하나 이들로 한정되지 않는다. 메틸 민감성 효소도 메틸화 상태를 기초로 핵산을 식별하는 데에 사용될 수 있는데, 상기 효소는 예를 들면, 그의 DNA 인식 서열이 메틸화되어 있지 않은 경우 이 인식 서열에서 우선적으로 또는 실질적으로 절단할 수 있거나 분해할 수 있다. 따라서, 메틸화되지 않은 DNA 샘플은 메틸화된 DNA 샘플보다 더 작은 단편으로 절단될 것이고, 과다메틸화된 DNA 샘플은 절단되지 않을 것이다. 명확히 언급된 경우를 제외하고, 메틸화 상태를 기초로 핵산을 식별하는 임의의 방법이 본원의 기술의 조성물 및 방법과 함께 사용될 수 있다. 태아 DNA의 양은 예를 들면, 증폭 반응 동안 공지된 농도로 하나 이상의 경쟁자를 도입함으로써 측정될 수 있다. 태아 DNA의 양의 측정은 예를 들면, RT-PCR, 프라이머 연장, 서열결정 및/또는 카운팅에 의해 수행될 수도 있다. 일부 경우, 핵산의 양은 미국 특허출원 공보 제2007/0065823호에 기재된 BEAMing 기술을 이용함으로써 측정될 수 있다. 일부 실시양태에서, 제한 효율이 측정될 수 있고, 이 효율도 태아 DNA의 양을 측정하는 데에 사용될 수 있다.

[0053] 일부 실시양태에서, 태아 정량자 분석(FQA)은 예를 들면, 하기 방법으로 모체 샘플 중의 태아 DNA의 농도를 측정하는 데에 사용될 수 있다: a) 모체 샘플에 존재하는 DNA의 총량을 측정하고; b) 하나 이상의 메틸화 민감성 제한 효소를 사용하여 모체 샘플 중의 모체 DNA를 선택적으로 분해하여 태아 DNA를 농축하고; c) 단계 b)로부터의 태아 DNA의 양을 측정하고; d) 단계 c)로부터의 태아 DNA의 양을 단계 a)로부터의 DNA의 총량과 비교하여 모체 샘플에서 태아 DNA의 농도를 측정한다. 일부 실시양태에서, 예를 들면, 절대 카피 수 측정을 위한 경쟁적 PCR 방법을 이용하는 질량 분광측정 및/또는 시스템을 사용하여 모체 샘플 중의 태아 핵산의 절대 카피 수를 측정할 수 있다. 예를 들면, 문헌(Ding and Cantor (2003) Proc. Natl. Acad. Sci. USA 100:3059-3064) 및 미국 특허출원 공보 제2004/0081993호(이들 둘 다 본원에 참고로 도입됨)를 참조한다.

[0054] 일부 실시양태에서, 태아 분획은 예를 들면, 본원에 참고로 도입되는 미국 특허출원 공보 제2011/0224087호에 기재된 방법을 이용함으로써 다형성 서열(예를 들면, 단일 뉴클레오티드 다형성(SNP))의 대립형질 비를 기초로 측정될 수 있다. 이러한 방법에서, 모체 샘플에 대한 뉴클레오티드 서열 리드를 수득하고 기준 게놈 내의 정보 제공 다형성 부위(예를 들면, SNP)에서 제1 대립형질로 맵핑되는 뉴클레오티드 서열 리드의 총수를 제2 대립형질로 맵핑되는 뉴클레오티드 서열 리드의 총수와 비교함으로써 태아 분획을 측정한다. 일부 실시양태에서, 태아 대립형질은 예를 들면, 모체 핵산이 샘플 중의 태아 핵산과 모체 핵산의 혼합물에 크게 기여하는 것과 비교될 때 태아 대립형질이 상기 혼합물에 상대적으로 미미하게 기여한다는 사실에 의해 확인된다. 따라서, 모체 샘플 중의 태아 핵산의 상대적 존재도는 다형성 부위의 2개 대립형질들 각각에 대한 기준 게놈 상의 표적 핵산 서열로 맵핑된 유일한 서열 리드의 총수의 파라미터로서 측정될 수 있다.

[0055] 일부 실시양태에서, 태아 분획은 단편 길이 정보를 도입하는 방법(예를 들면, 본원에 참고로 도입되는 국제 특허출원 공보 제WO2013/177086호에 기재된 단편 길이 비(FLR) 분석 또는 태아 비 통계(FRS) 분석)을 이용함으로써 측정될 수 있다. 무세포 태아 핵산 단편은 일반적으로 모체로부터 유래된 핵산 단편보다 더 짧다(예를 들면, 문헌(Chan et al. (2004) Clin. Chem. 50:88-92); 및 문헌(Lo et al. (2010) Sci. Transl. Med. 2:61ra91) 참조). 따라서, 일부 실시양태에서 태아 분획은 특정 길이 역치 미만의 단편을 카운팅하고 카운트를 예를 들면, 특정 길이 역치 초과의 단편으로부터의 카운트 및/또는 샘플 중의 총 핵산의 양과 비교함으로써 측정될 수 있다. 특정 길이의 핵산 단편을 카운팅하는 방법은 국제 특허출원 공보 제WO2013/177086호에 더 상세히 기재되어 있다.

[0056] 일부 실시양태에서, 태아 분획은 부분 특이적 태아 분획 추정치에 따라 측정될 수 있다. 이론에 의해 한정되지 않지만, 태아 CCF 단편(예를 들면, 특정 길이 또는 길이 범위의 단편)으로부터의 일정량의 리드가 종종 (예를 들면, 동일한 서열결정 실시에서, 예를 들면, 동일한 샘플 내에서) 넓은 빈도로 부분들에 맵핑된다. 또한, 이론에 의해 한정되지 않지만, 다수의 샘플들 사이에 비교될 때 일부 부분들은 태아 CCF 단편(예를 들면, 특정 길이 또는 길이 범위의 단편)으로부터의 리드의 유사한 표시치를 갖는 경향을 나타내고, 그 표시치는 부위 특이적 태아 분획(예를 들면, 태아로부터 유래된 CCF 단편의 상대적 양, 백분율 또는 비)과 상관관계를 갖는다.

[0057] 일부 실시양태에서, 부분 특이적 태아 분획 추정치는 부분적으로 부분 특이적 파라미터들 및 이들과 태아 분획의 관계를 기초로 측정된다. 부분 특이적 파라미터는 부분에서 특정 크기(예를 들면, 크기 범위)의 CCF 단편 길이로부터의 리드의 양 또는 비율을 반영하는(예를 들면, 이러한 양 또는 비율과 상관관계를 갖는) 임의의 적합

한 파라미터일 수 있다. 부분 특이적 파라미터는 다수의 샘플들에 대해 측정된 부분 특이적 파라미터의 평균치, 평균 또는 중간치일 수 있다. 임의의 적합한 부분 특이적 파라미터가 사용될 수 있다. 부분 특이적 파라미터의 비한정적 예는 FLR(예를 들면, FRS), 선택된 단편 길이보다 더 짧은 길이를 갖는 리드의 양, 게놈 커버리지(즉, 커버리지), 맵핑가능성, 카운트(예를 들면, 부분들에 맵핑된 서열 리드의 카운트, 예를 들면, 정규화된 카운트, PERUN 정규화된 카운트, ChAI 정규화된 카운트), DNase I 민감성, 메틸화 상태, 아세틸화, 히스톤 분포, 구아닌-사이토신(GC) 함량, 염색질 구조 등 또는 이들의 조합을 포함한다. 부분 특이적 파라미터는 부분 특이적 방식으로 FLR 및/또는 FRS와 상관관계를 갖는 임의의 적합한 파라미터일 수 있다. 일부 실시양태에서, 일부 또는 모든 부분 특이적 파라미터들이 부분에 대한 FLR의 직접적인 또는 간접적인 표시치이다. 일부 실시양태에서, 부분 특이적 파라미터는 구아닌-사이토신(GC) 함량이 아니다.

[0058] 일부 실시양태에서, 부분 특이적 파라미터는 CCF 단편으로부터의 리드의 양을 나타내거나, 이러한 양과 상관관계를 갖거나 이러한 양에 비례하는 임의의 적합한 값이고, 이때 부분들에 맵핑된 리드는 선택된 단편 길이보다 더 짧은 길이를 갖는다. 일부 실시양태에서, 부분 특이적 파라미터는 부분들로 맵핑되는 상대적으로 짧은 CCF 단편(예를 들면, 약 200개 이하의 염기쌍)으로부터 유도된 리드의 양의 표시치이다. 선택된 단편 길이보다 더 짧은 길이를 갖는 CCF 단편은 종종 상대적으로 짧은 CCF 단편이고, 종종 선택된 단편 길이는 약 200개 이하의 염기쌍이다(예를 들면, 길이에서 약 190개, 180개, 170개, 160개, 150개, 140개, 130개, 120개, 110개, 100개, 90개 또는 80개 염기를 갖는 CCF 단편). CCF 단편, 또는 CCF 단편으로부터 유도된 리드의 길이는 임의의 적합한 방법(예를 들면, 서열결정 방법, 혼성화 방법)에 의해 측정될 수 있다(예를 들면, 추정될 수 있거나 유추될 수 있다). 일부 실시양태에서, CCF 단편의 길이는 페어링된-말단 서열결정 방법으로부터 수득된 리드에 의해 측정된다(예를 들면, 추정되거나 유추된다). 일부 실시양태에서, CCF 단편 주형의 길이는 CCF 단편으로부터 유도된 리드(예를 들면, 단일-말단 리드)의 길이로부터 직접적으로 측정된다.

[0059] 부분 특이적 파라미터는 하나 이상의 가중률에 의해 가중될 수 있거나 조절될 수 있다. 일부 실시양태에서, 가중된 또는 조절된 부분 특이적 파라미터는 샘플(예를 들면, 검사 샘플)에 대한 부분 특이적 태아 분획 추정치를 제공할 수 있다. 일부 실시양태에서, 가중 또는 조절은 일반적으로 부분(예를 들면, 부분들에 맵핑된 리드) 또는 또 다른 부분 특이적 파라미터의 카운트를 부분 특이적 태아 분획 추정치로 전환시키고, 이러한 전환은 종종 변환으로서 간주된다.

[0060] 일부 실시양태에서, 가중률은 부분적으로 태아 분획(예를 들면, 다수의 샘플들로부터 측정된 태아 분획)과 다수의 샘플들(예를 들면, 트레이닝 세트)에 대한 부분 특이적 파라미터 사이의 관계를 기술하고/하거나 정의하는 계수 또는 상수이다. 일부 실시양태에서, 가중률은 다수의 태아 분획 측정치와 다수의 부분 특이적 파라미터의 관계에 따라 측정된다. 관계는 하나 이상의 가중률에 의해 정의될 수 있고, 하나 이상의 가중률은 관계로부터 측정될 수 있다. 일부 실시양태에서, 가중률(예를 들면, 하나 이상의 가중률)은 (i) 다수의 샘플들 각각에 대해 측정된 태아 핵산의 분획, 및 (ii) 다수의 샘플들에 대한 부분 특이적 파라미터에 따라 부분에 대한 피팅된 관계로부터 측정된다.

[0061] 가중률은 적합한 관계(예를 들면, 적합한 수학적 관계, 대수적 관계, 피팅된 관계, 회귀, 회귀 분석, 회귀 모델)로부터 유도된 임의의 적합한 계수, 추정된 계수 또는 상수일 수 있다. 가중률은 적합한 관계에 따라 측정될 수 있거나, 적합한 관계로부터 유도될 수 있거나 적합한 관계로부터 추정될 수 있다. 일부 실시양태에서, 가중률은 피팅된 관계로부터 추정된 계수이다. 다수의 샘플들에 대한 관계의 피팅은 종종 트레이닝 모델로서 지칭된다. 관계를 피팅하는(예를 들면, 트레이닝 세트에 대한 모델을 트레이닝하는) 임의의 적합한 모델 및/또는 방법이 이용될 수 있다. 사용될 수 있는 적합한 모델의 비한정적 예는 회귀 모델, 선형 회귀 모델, 단순 회귀 모델, 통상의 최소 자승 회귀 모델, 다중 회귀 모델, 일반적인 다중 회귀 모델, 다항식 회귀 모델, 일반적인 선형 모델, 일반화된 선형 모델, 별개의 선택 회귀 모델, 로지스틱 회귀 모델, 다항 로짓(logit) 모델, 혼합된 로짓 모델, 프로빗(probit) 모델, 다항 프로빗 모델, 순서형 로짓 모델, 순서형 프로빗 모델, 푸아송(Poisson) 모델, 다변량 반응 회귀 모델, 다층 모델, 고정된 효과 모델, 무작위 효과 모델, 혼합된 모델, 비선형 회귀 모델, 비-파라미터 모델, 반파라미터 모델, 로부스트(robust) 모델, 변위치 모델, 등장성 모델, 주성분 모델, 최소 각도 모델, 국소 모델, 분할된 모델 및 변수 내 오차(errors-in-variables) 모델을 포함한다. 일부 실시양태에서, 고정된 관계는 회귀 모델이 아니다. 일부 실시양태에서, 고정된 관계는 결정 트리(tree) 모델, 지지체-백터 머신(support-vector machine) 모델 및 신경 네트워크(neural network) 모델로부터 선택된다. 모델(예를 들면, 회귀 모델, 관계)을 트레이닝한 결과는 종종 관계가 하나 이상의 계수(예를 들면, 가중률)를 포함하는 경우 수학적으로 기술될 수 있는 관계이다. 보다 더 복잡한 다변량 모델은 1개, 2개 또는 3개 이상의 가중률을 측정할 수 있다. 일부 실시양태에서, 모델은 다수의 샘플들로부터 수득된 태아 분획 및 2개 이상의 부분 특이적 파라미터

(예를 들면, 계수)(예를 들면, 매트릭스에 의해 다수의 샘플들에 피팅된 피팅된 관계)에 따라 트레이닝된다.

[0062] 가중률은 적합한 방법에 의해 적합한 관계(예를 들면, 적합한 수학적 관계, 대수적 관계, 피팅된 관계, 회귀, 회귀 분석, 회귀 모델)로부터 유도될 수 있다. 일부 실시양태에서, 피팅된 관계는 추정에 의해 피팅되고, 이의 비한정적 예는 최소 자승, 통상의 최소 자승, 선형, 부분적, 총, 일반화된, 가중된, 비선형, 반복적으로 재가중된, 리지(ridge) 회귀, 최소 절대 편차, 베이지안(Bayesian), 베이지안 다변량, 감소된-순위, LASSO, 가중된 순위 선택 기준(WRSC), 순위 선택 기준(RSC), 엘라스틱 넷 추정자(elastic net estimator)(예를 들면, 엘라스틱 넷 회귀) 및 이들의 조합을 포함한다.

[0063] 가중률은 게놈의 임의의 적합한 부분에 대해 측정될 수 있거나 이러한 임의의 적합한 부분과 관련될 수 있다. 가중률은 임의의 적합한 염색체의 임의의 적합한 부분에 대해 측정될 수 있거나 이러한 임의의 적합한 부분과 관련될 수 있다. 일부 실시양태에서, 가중률은 게놈 내의 일부 또는 모든 부분들에 대해 측정되거나 이들과 관련되어 있다. 일부 실시양태에서, 가중률은 게놈 내의 일부 또는 모든 염색체들에 대해 측정되거나 이들과 관련되어 있다. 가중률은 종종 선택된 염색체의 부분에 대해 측정되거나 이러한 부분과 관련되어 있다. 가중률은 하나 이상의 상염색체(autosome)의 부분에 대해 측정될 수 있거나 이러한 부분과 관련될 수 있다. 가중률은 상염색체 내의 부분들 또는 이들의 서브세트를 포함하는 복수의 부분들 중 부분에 대해 측정될 수 있거나 이러한 부분과 관련될 수 있다. 일부 실시양태에서, 가중률은 성염색체(예를 들면, ChrX 및/또는 ChrY)의 부분에 대해 측정되거나 이러한 부분과 관련되어 있다. 가중률은 하나 이상의 상염색체 및 하나 이상의 성염색체의 부분에 대해 측정될 수 있거나 이러한 부분과 관련될 수 있다. 일부 실시양태에서, 가중률은 모든 상염색체들 및 X 및 Y 염색체 내의 복수의 부분들 중 부분에 대해 측정되거나 이러한 부분과 관련되어 있다. 가중률은 X 및/또는 Y 염색체 내의 부분을 포함하지 않는 복수의 부분들 중 부분에 대해 측정될 수 있거나 이러한 부분과 관련될 수 있다. 일부 실시양태에서, 가중률은 염색체의 부분에 대해 측정되거나 이러한 부분과 관련되어 있고, 이때 염색체는 이배수성(예를 들면, 전체 염색체 이배수성)을 포함한다. 일부 실시양태에서, 가중률은 염색체의 부분에 대해서만 측정되거나 이러한 부분과만 관련되어 있고, 이때 염색체는 정배수체(예를 들면, 정배수체 염색체)가 아니다. 가중률은 13번, 18번 및/또는 21번 염색체 내의 부분을 포함하지 않는 복수의 부분들 중 부분에 대해 측정될 수 있거나 이러한 부분과 관련될 수 있다.

[0064] 일부 실시양태에서, 가중률은 하나 이상의 샘플(예를 들면, 샘플의 트레이닝 세트)에 따라 부분에 대해 측정된다. 가중률은 종종 부분에 대한 특이성을 갖는다. 일부 실시양태에서, 하나 이상의 가중률은 부분에 독립적으로 배정된다. 일부 실시양태에서, 가중률은 다수의 샘플들에 대한 태아 분획 측정치(예를 들면, 샘플 특이적 태아 분획 측정치)와 다수의 샘플에 따라 측정된 부분 특이적 파라미터의 관계에 따라 측정된다. 가중률은 종종 다수의 샘플들, 예를 들면, 약 20개 내지 약 100,000개 이상, 약 100개 내지 약 100,000개 이상, 약 500개 내지 약 100,000개 이상, 약 1,000개 내지 약 100,000개 이상, 또는 약 10,000개 내지 약 100,000개 이상의 샘플들로부터 측정된다. 가중률은 정배수체인 샘플(예를 들면, 정배수체 태아를 포함하는 대상체로부터의 샘플, 예를 들면, 이수체 염색체가 존재하지 않는 샘플)로부터 측정될 수 있다. 일부 실시양태에서, 가중률은 이수체 염색체를 포함하는 샘플(예를 들면, 정배수체 태아를 포함하는 대상체로부터의 샘플)로부터 수득된다. 일부 실시양태에서, 가중률은 정배수체 태아를 갖는 대상체 및 삼염색체성 태아를 갖는 대상체로부터의 다수의 샘플들로부터 측정된다. 가중률은 다수의 샘플들로부터 유도될 수 있고, 이때 상기 샘플들은 남성 태아 및/또는 여성 태아를 갖는 대상체로부터의 샘플들이다.

[0065] 태아 분획은 종종 가중률이 유도되는 트레이닝 세트의 하나 이상의 샘플에 대해 측정된다. 가중률이 측정되는 태아 분획은 종종 샘플 특이적 태아 분획 측정치이다. 가중률이 측정되는 태아 분획은 본원에 기재되어 있거나 당분야에서 공지되어 있는 임의의 적합한 방법에 의해 측정될 수 있다. 일부 실시양태에서, 태아 핵산 함량(예를 들면, 태아 분획)의 측정은 본원에 기재되어 있거나 당분야에서 공지되어 있는 적합한 태아 정량자 분석(FQA)을 이용함으로써 수행되고, 이러한 분석의 비한정적 예는 남성 태아에 대한 특이성을 갖는 마커에 따른 태아 분획 측정, 다형성 서열의 대립형질 비에 기초한 태아 분획 측정, 태아 핵산에 대한 특이성을 갖지만 모체 핵산에 대한 특이성을 갖지 않는 하나 이상의 마커에 따른 태아 분획 측정, 메틸화에 기초한 DNA 식별의 이용에 의한 태아 분획 측정(예를 들면, 문헌(A. Nygren, et al., (2010) Clinical Chemistry 56(10):1627-1635)), 경쟁 PCR 방법을 이용하는 질량 분광측정 방법 및/또는 시스템에 의한 태아 분획 측정, 본원에 참고로 도입되는 미국 특허출원 공보 제2010/0105049호에 기재된 방법에 의한 태아 분획 측정 등 또는 이들의 조합을 포함한다. 종종 태아 분획은 부분적으로 Y 염색체의 수준(예를 들면, 하나 이상의 게놈 구획 수준, 프로파일의 수준)에 따라 측정된다. 일부 실시양태에서, 태아 분획은 Y 염색체의 적합한 분석에 따라(예를 들면, 정량 실시간 PCR을 이용하여 태아 특이적 좌위(예컨대, 남아 임신에서 Y 염색체 상의 SRY 좌위)의 양을 모체 및 태아 둘 다에 공통

된 임의의 상염색체 상의 좌위의 양과 비교함으로써(예를 들면, Lo YM, et al.(1998) Am J Hum Genet 62:768-775)) 측정된다.

[0066] (예를 들면, 검사 샘플에 대한) 부분 특이적 파라미터는 하나 이상의 가중률(예를 들면, 트레이닝 세트로부터 유도된 가중률)에 의해 가중될 수 있거나 조절될 수 있다. 예를 들면, 가중률은 다수의 샘플들의 트레이닝 세트에 대한 부분 특이적 파라미터와 태아 분획 측정치의 관계에 따라 부분에 대해 유도될 수 있다. 그 다음, 검사 샘플의 부분 특이적 파라미터는 트레이닝 세트로부터 유도된 가중률에 따라 조절될 수 있고/있거나 가중될 수 있다. 일부 실시양태에서, 가중률이 유도된 부분 특이적 파라미터는 조절되거나 가중된 (예를 들면, 검사 샘플의) 부분 특이적 파라미터와 동일하다(예를 들면, 두 파라미터들이 FLR이다). 일부 실시양태에서, 가중률이 유도된 부분 특이적 파라미터는 조절되거나 가중된 (예를 들면, 검사 샘플의) 부분 특이적 파라미터와 상이하다. 예를 들면, 가중률은 샘플의 트레이닝 세트에 대한 커버리지(즉, 부분 특이적 파라미터)와 태아 분획 사이의 관계로부터 측정될 수 있고, 검사 샘플의 부분에 대한 FLR(즉, 또 다른 부분 특이적 파라미터)은 커버리지로부터 유도된 가중률에 따라 조절될 수 있다. 이론에 의해 한정되지 않지만, (예를 들면, 검사 샘플에 대한) 부분 특이적 파라미터는 종종 각각의 부분 특이적 파라미터와 공통된 부분 특이적 FLR 사이의 관계 및/또는 상관관계로 인해 (예를 들면, 트레이닝 세트의) 상이한 부분 특이적 파라미터로부터 유도된 가중률에 의해 조절될 수 있고/있거나 가중될 수 있다.

[0067] 부분 특이적 파라미터를 그 부분에 대해 측정된 가중률로 가중함으로써 샘플(예를 들면, 검사 샘플)에 대한 부분 특이적 태아 분획 추정치를 측정할 수 있다. 가중은 임의의 적합한 수학적 조작을 적용함으로써 가중률에 따라 부분 특이적 파라미터를 조절하고하거나 전환시키고/시키거나 변환시키는 단계를 포함할 수 있고, 상기 수학적 조작의 비한정적 예는 곱셈, 나눗셈, 덧셈, 뺄셈, 적분, 기호 연산, 대수적 연산, 알고리즘, 삼각법 또는 기하학 함수, 변환(예를 들면, 푸리에(Fourier) 변환) 등 또는 이들의 조합을 포함한다. 가중은 적합한 수학적 모델로 가중률에 따라 부분 특이적 파라미터를 조절하고/하거나, 전환시키고/시키거나 변환시키는 단계를 포함할 수 있다.

[0068] 일부 실시양태에서, 하나 이상의 부분 특이적 태아 분획 추정치에 따라 샘플에 대한 태아 분획이 측정된다. 일부 실시양태에서, 하나 이상의 부분에 대한 부분 특이적 파라미터의 가중 또는 조절에 따라 샘플(예를 들면, 검사 샘플)에 대한 태아 분획이 측정(예를 들면, 추정)된다. 일부 실시양태에서, 검사 샘플에 대한 태아 핵산의 분획은 조절된 카운트 또는 조절된 카운트 서브세트를 기초로 추정된다. 일부 실시양태에서, 검사 샘플에 대한 태아 핵산의 분획은 부분에 대한 조절된 FLR, 조절된 FRS, 조절된 커버리지 및/또는 조절된 맵핑가능성을 기초로 추정된다. 일부 실시양태에서, 약 1개 내지 약 500,000개, 약 100개 내지 약 300,000개, 약 500개 내지 약 200,000개, 약 1,000개 내지 약 200,000개, 약 1,500개 내지 약 200,000개, 또는 약 1,500개 내지 약 50,000개의 부분 특이적 파라미터들이 가중되거나 조절된다.

[0069] (예를 들면, 검사 샘플에 대한) 태아 분획은 임의의 적합한 방법에 의해 (예를 들면, 동일한 검사 샘플에 대한) 다수의 부분 특이적 태아 분획 추정치에 따라 측정될 수 있다. 일부 실시양태에서, 임신 여성으로부터의 검사 샘플에서 태아 핵산의 분획 추정의 정확성을 증가시키는 방법은 하나 이상의 부분 특이적 태아 분획 추정치를 측정하는 단계를 포함하고, 이때 샘플에 대한 태아 분획의 추정치는 하나 이상의 부분 특이적 태아 분획 추정치에 따라 측정된다. 일부 실시양태에서, 샘플(예를 들면, 검사 샘플)에 대한 태아 핵산의 분획의 추정 또는 측정은 하나 이상의 부분 특이적 태아 분획 추정치를 합산하는 단계를 포함한다. 합산은 다수의 부분 특이적 태아 분획 추정치에 따라 평균치, 평균, 중간치, AUC 또는 적분값을 측정하는 단계를 포함할 수 있다.

[0070] 일부 실시양태에서, 임신 여성으로부터의 검사 샘플에서 태아 핵산의 분획 추정의 정확도를 증가시키는 방법은 기준 게놈의 부분들에 맵핑된 서열 리드의 카운트를 수득하는 단계를 포함하고, 이때 상기 서열 리드는 임신 여성으로부터의 검사 샘플로부터의 순환 무세포 핵산의 리드이고, 이때 적어도 수득된 카운트의 서브세트는 게놈의 또 다른 영역의 총 카운트에 대한 상대적인 태아 핵산의 카운트보다 영역으로부터의 총 카운트에 대한 상대적인 태아 핵산으로부터 유도된 더 큰 수의 카운트를 기여하는 게놈의 영역으로부터 유도된다. 일부 실시양태에서, 태아 핵산의 분획의 추정치는 부분의 서브세트에 따라 측정되고, 이때 부분의 서브세트는 또 다른 부분의 태아 핵산의 카운트보다 태아 핵산으로부터 유도된 더 큰 수의 카운트가 맵핑되는 부분에 따라 선택된다. 일부 실시양태에서, 부분의 서브세트는 또 다른 부분의 비-태아 핵산에 대한 상대적인 태아 핵산의 카운트보다 비-태아 핵산에 대한 상대적인 태아 핵산으로부터 유도된 더 큰 수의 카운트가 맵핑되는 부분에 따라 선택된다. 부분의 전부 또는 서브세트로 맵핑된 카운트는 가중되어 가중된 카운트를 제공할 수 있다. 가중된 카운트는 태아 핵산의 분획을 추정하는 데에 사용될 수 있고, 카운트는 또 다른 부분의 태아 핵산의 카운트보다 태아 핵산으로부터 유도된 더 큰 수의 카운트가 맵핑되는 부분에 따라 가중될 수 있다. 일부 실시양태에서, 카운트는 또 다른

부분의 비-태아 핵산에 대한 상대적인 태아 핵산의 카운트보다 비-태아 핵산에 대한 상대적인 태아 핵산으로부터 유도된 더 큰 수의 카운트가 맵핑되는 부분에 따라 가중된다.

[0071] 샘플에 대한 다수의 부분 특이적 태아 분획 추정치에 따라 샘플(예를 들면, 검사 샘플)에 대한 태아 분획이 측정될 수 있고, 이때 부분 특이적 추정치는 게놈의 임의의 적합한 영역 또는 분절의 부분으로부터 유도된다. 적합한 염색체(예를 들면, 하나 이상의 선택된 염색체, 하나 이상의 상염색체, 성염색체(예를 들면, ChrX 및/또는 ChrY), 이배수체 염색체, 정배수체 염색체 등 또는 이들의 조합)의 하나 이상의 부분에 대한 부분 특이적 태아 분획 추정치가 측정될 수 있다.

[0072] 일부 실시양태에서, 태아 분획의 측정은 (a) 기준 게놈의 부분들에 맵핑된 서열 리드의 카운트를 수득하는 단계로서, 이때 상기 서열 리드가 임신 여성으로부터의 검사 샘플로부터의 순환 무세포 핵산의 리드인 단계; (b) 마이크로프로세서를 이용하여 각각의 부분과 독립적으로 관련된 가중률에 따라 (i) 각각의 부분들에 맵핑된 서열 리드의 카운트, 또는 (ii) 다른 부분 특이적 파라미터를 태아 핵산의 부분 특이적 분획에 가중함으로써 가중률에 따라 부분 특이적 태아 분획 추정치를 제공하는 단계로서, 이때 각각의 가중률이 (i) 다수의 샘플들 각각에 대한 태아 핵산의 분획과 (ii) 다수의 샘플들에 대한 각각의 부분들에 맵핑된 서열 리드의 카운트 또는 다른 부분 특이적 파라미터 사이에 각각의 부분들에 대한 피팅된 관계로부터 측정되는 것인 단계; 및 (c) 부분 특이적 태아 분획 추정치를 기초로 검사 샘플에 대한 태아 핵산의 분획을 추정하는 단계를 포함한다.

[0073] 세포외 핵산 중의 태아 핵산의 양이 정량될 수 있고 본원에서 제공된 방법과 함께 사용될 수 있다. 따라서, 일부 실시양태에서, 본원에 기재된 기술의 방법은 태아 핵산의 양을 측정하는 추가 단계를 포함한다. 태아 핵산의 양은 샘플 핵산을 준비하기 위한 프로세싱 전 또는 후에 대상체로부터의 핵산 샘플에서 측정될 수 있다. 일부 실시양태에서, 태아 핵산의 양은 샘플 핵산이 프로세싱되고 준비된 후에 샘플에서 측정되고, 이 양은 추가 평가를 위해 사용된다. 일부 실시양태에서, 결과는 샘플 핵산 중의 태아 핵산의 분획을 팩토링하는(factoring) 것(예를 들면, 카운트를 조절하거나, 샘플을 제거하거나, 판정하거나 판정하지 않는 것)을 포함한다.

[0074] 측정 단계는 본원에 기재된 방법에서 어느 한 시점 전, 이 시점 동안 또는 이 시점에서, 또는 본원에 기재된 일부(예를 들면, 이배수성 검출, 미세중복 또는 미세결실 검출, 태아 성별 결정) 방법 후에 수행될 수 있다. 예를 들면, 주어진 민감성 또는 특이성으로 태아 성별 또는 이배수성, 미세중복 또는 미세결실 확인 방법을 달성하기 위해, 태아 성별 또는 이배수성, 미세중복 또는 미세결실 확인 전, 동안 또는 후에 태아 핵산 정량 방법을 실시하여 약 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24% 또는 25% 이상의 태아 핵산을 갖는 샘플을 확인할 수 있다. 일부 실시양태에서, 일정한 역치 양의 태아 핵산(예를 들면, 약 15% 이상의 태아 핵산; 약 4% 이상의 태아 핵산)을 갖는 것으로 확인된 샘플은 예를 들면, 태아 성별 또는 이배수성, 미세중복 또는 미세결실 확인, 또는 이배수성 또는 유전적 변이의 존재 또는 부재에 대해 더 분석된다. 일부 실시양태에서, 예를 들면, 태아 성별의 확인, 또는 이배수성, 미세중복 또는 미세결실의 존재 또는 부재의 확인은 일정한 역치 양의 태아 핵산(예를 들면, 약 15% 이상의 태아 핵산; 약 4% 이상의 태아 핵산)을 갖는 샘플에 대해서만 선택된다(예를 들면, 선택되고 환자에게 통보된다).

[0075] 일부 실시양태에서, 태아 분획의 측정 또는 태아 핵산의 양의 측정은 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하기 위해 요구되지 않거나 필요하지 않다. 일부 실시양태에서, 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재의 확인은 태아 대 모체 DNA의 서열 식별을 요구하지 않는다. 일부 실시양태에서, 이것은 부모의 합산된 기여 때문이고, 특정 염색체, 염색체 부분 또는 이의 분절에서 태아 서열이 분석된다. 일부 실시양태에서, 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재의 확인은 모체 DNA로부터 태아 DNA를 식별시켜 줄 종래 서열 정보에 의존하지 않는다.

[0076] 핵산의 농축

[0077] 일부 실시양태에서, 핵산(예를 들면, 세포외 핵산)은 핵산의 하위집단 또는 종에 대해 농축되거나 상대적으로 농축된다. 핵산 하위집단은 예를 들면, 태아 핵산, 모체 핵산, 특정 길이 또는 길이 범위의 단편을 포함하는 핵산, 또는 특정 게놈 영역(예를 들면, 단일 염색체, 염색체의 세트 및/또는 일부 염색체 영역)으로부터의 핵산을 포함할 수 있다. 이러한 농축된 샘플은 본원에서 제공된 방법과 함께 사용될 수 있다. 따라서, 일부 실시양태에서, 본 기술의 방법은 샘플 중의 핵산, 예를 들면, 태아 핵산의 하위집단을 농축하는 추가 단계를 포함한다. 일부 실시양태에서, 전술된 태아 분획을 측정하는 방법도 태아 핵산을 농축하는 데에 이용될 수 있다. 일부 실시양태에서, 모체 핵산은 샘플로부터 (부분적으로, 실질적으로, 거의 완전히 또는 완전히) 선택적으로 제거된다. 일부 실시양태에서, 특정 낮은 카피 수 종 핵산(예를 들면, 태아 핵산)의 농축은 정량적 민감성을 개선할 수 있다. 핵산의 특정 종에 대해 샘플을 농축하는 방법은 예를 들면, 미국 특허 제6,927,028호, 국제 특허출원 공보

제WO2007/140417호, 국제 특허출원 공보 제WO2007/147063호, 국제 특허출원 공보 제WO2009/032779호, 국제 특허출원 공보 제WO2009/032781호, 국제 특허출원 공보 제WO2010/033639호, 국제 특허출원 공보 제WO2011/034631호, 국제 특허출원 공보 제WO2006/056480호 및 국제 특허출원 공보 제WO2011/143659호(이들 모두 본원에 참고로 도입됨)에 기재되어 있다.

[0078] 일부 실시양태에서, 핵산은 일부 표적 단편 종 및/또는 기준 단편 종에 대해 농축된다. 일부 실시양태에서, 핵산은 후술된 하나 이상의 길이 기초 분리 방법을 이용함으로써 특정 핵산 단편 길이 또는 단편 길이의 범위에 대해 농축된다. 일부 실시양태에서, 핵산은 본원에 기재되어 있고/있거나 당분야에서 공지되어 있는 하나 이상의 서열 기초 분리 방법을 이용함으로써 선택 게놈 영역(예를 들면, 염색체)으로부터의 단편에 대해 농축된다. 샘플에서 핵산 하위집단(예를 들면, 태아 핵산)을 농축하는 일부 방법들이 상세히 후술되어 있다.

[0079] 본원에 기재된 방법과 함께 이용될 수 있는 핵산 하위집단(예를 들면, 태아 핵산)을 농축하는 일부 방법들은 모체 핵산과 태아 핵산 사이의 유전외적 차이를 활용하는 방법을 포함한다. 예를 들면, 태아 핵산은 메틸화 차이를 기초로 모체 핵산으로부터 식별될 수 있고 분리될 수 있다. 메틸화 기초 태아 핵산 농축 방법은 본원에 참고로 도입되는 미국 특허출원 공보 제2010/0105049호에 기재되어 있다. 이러한 방법들은 종종 샘플 핵산을 메틸화 특이적 결합제(메틸-CpG 결합 단백질(MBD), 메틸화 특이적 항체 등)에 결합시키는 단계 및 상이한 메틸화 상태를 기초로 결합되지 않은 핵산으로부터 결합된 핵산을 분리하는 단계를 포함한다. 이러한 방법들은 하나 이상의 태아 핵산 영역에 대해 샘플을 농축하기 위해 모체 핵산을 선택적으로 및 완전히 또는 실질적으로 분해하는 효소를 사용하여 모체 샘플로부터의 핵산을 선택적으로 분해함으로써 모체 샘플에서 태아 핵산 영역을 농축할 수 있게 하는 메틸화 민감성 제한 효소(전술된 바와 같음; 예를 들면, HhaI 및 HpaII)의 사용을 포함할 수도 있다.

[0080] 본원에 기재된 방법과 함께 이용될 수 있는 핵산 하위집단(예를 들면, 태아 핵산)을 농축하는 또 다른 방법은 제한 엔도뉴클레아제(endonuclease) 향상 다형성 서열 방법, 예컨대, 본원에 참고로 도입되는 미국 특허출원 공보 제2009/0317818호에 기재된 방법이다. 이러한 방법은 비-표적 대립형질을 포함하는 핵산을 인식하나 표적 대립형질을 인식하지 않는 제한 엔도뉴클레아제를 사용하여 비-표적 대립형질을 포함하는 핵산을 절단하는 단계; 및 절단된 핵산을 증폭하지 않으면서 절단되지 않은 핵산을 증폭하는 단계로서, 이때 절단되지 않은 증폭된 핵산이 비-표적 핵산(예를 들면, 모체 핵산)에 비해 상대적으로 농축된 표적 핵산(예를 들면, 태아 핵산)을 나타내는 것인 단계를 포함한다. 일부 실시양태에서, 핵산은 예를 들면, 절단제에 의한 선택적 분해에 민감한 다형성 부위를 갖는 대립형질을 포함하도록 선택될 수 있다.

[0081] 본원에 기재된 방법과 함께 사용될 수 있는 핵산 하위집단(예를 들면, 태아 핵산)을 농축하는 일부 방법들은 선택적 효소 분해 방법을 포함한다. 이러한 방법은 엑소뉴클레아제(exonuclease) 분해로부터 표적 서열을 보호하여 샘플에서 원하지 않는 서열(예를 들면, 모체 DNA)의 제거를 용이하게 하는 단계를 포함한다. 예를 들면, 한 방법에서, 샘플 핵산을 변성시켜 단일 가닥 핵산을 생성하고, 단일 가닥 핵산을 적합한 어닐링 조건 하에서 하나 이상의 표적 특이적 프라이머 쌍과 접촉시키고, 어닐링된 프라이머를 뉴클레오티드 중합으로 연장하여 이중 가닥 표적 서열을 생성하고 단일 가닥(즉, 비-표적) 핵산을 분해하는 뉴클레아제를 사용하여 단일 가닥 핵산을 분해한다. 일부 실시양태에서, 상기 방법은 하나 이상의 추가 주기에 동안 반복될 수 있다. 일부 실시양태에서, 동일한 표적 특이적 프라이머 쌍을 사용하여 연장의 제1 주기 및 제2 주기 각각을 프라이밍하고, 일부 실시양태에서 제1 주기 및 제2 주기를 위해 상이한 표적 특이적 프라이머 쌍을 사용한다.

[0082] 본원에 기재된 방법과 함께 이용될 수 있는 핵산 하위집단(예를 들면, 태아 핵산)을 농축하는 일부 방법들은 대량 병렬 시그너처(signature) 서열결정(MPSS) 방법을 포함한다. MPSS는 전형적으로 어댑터(즉, 태그) 연결 단계에 이어서 어댑터 해독 단계, 및 약간씩 증가시키면서 핵산 서열을 판독하는 단계를 이용하는 고체상 방법이다. 태깅된 PCR 생성물은 전형적으로 각각의 핵산이 유일한 태그를 갖는 PCR 생성물을 생성하도록 증폭된다. 태그는 종종 PCR 생성물을 마이크로비드에 부착시키는 데에 사용된다. 연결 기초 서열 결정의 수회 라운드 후, 예를 들면, 서열 시그너처가 각각의 비드로부터 확인될 수 있다. MPSS 데이터세트에서 각각의 시그너처 서열(MPSS 태그)이 분석되고 모든 다른 시그너처와 비교되고, 모든 동일한 시그너처가 카운팅된다.

[0083] 일부 실시양태에서, 일부 농축 방법들(예를 들면, 일부 MPS 및/또는 MPSS 기초 농축 방법들)은 증폭(예를 들면, PCR) 기초 방법을 포함할 수 있다. 일부 실시양태에서, (예를 들면, 좌위 특이적 증폭 프라이머를 사용하는) 좌위 특이적 증폭 방법이 이용될 수 있다. 일부 실시양태에서, 다중체 SNP 대립형질 PCR 방법이 이용될 수 있다. 일부 실시양태에서, 다중체 SNP 대립형질 PCR 방법이 단일체 서열결정과 함께 이용될 수 있다. 예를 들면, 이러한 방법은 다중체 PCR(예를 들면, MASSARRAY 시스템)의 사용 및 앰플리콘(amplicons) 내로의 포획 프로브 서열의 도입에 이어서, 예를 들면, 일루미나(Illumina) MPSS 시스템을 사용한 서열결정을 포함할 수 있다. 일부 실

시양태에서, 다중체 SNP 대립형질 PCR 방법은 3-프라이머 시스템 및 인덱스화(indexed) 서열결정과 함께 이용될 수 있다. 예를 들면, 이러한 방법은 일부 좌위 특이적 정방향 PCR 프라이머 내로 도입된 제1 포획 프로브 및 좌위 특이적 역방향 PCR 프라이머 내로 도입된 어댑터 서열을 갖는 프라이머들과 함께 다중체 PCR(예를 들면, MASSARRAY 시스템)을 이용하여 앰플리콘을 생성한 후, 예를 들면, 일루미나 MPSS 시스템을 사용한 서열결정을 위해 이차 PCR을 이용하여 역방향 포획 서열 및 분자 인덱스 바코드를 도입하는 단계를 포함할 수 있다. 일부 실시양태에서, 다중체 SNP 대립형질 PCR 방법은 4-프라이머 시스템 및 인덱스화 서열결정과 함께 이용될 수 있다. 예를 들면, 이러한 방법은 좌위 특이적 정방향 PCR 프라이머 및 좌위 특이적 역방향 PCR 프라이머 둘 다 내로 도입된 어댑터 서열을 갖는 프라이머와 함께 다중체 PCR(예를 들면, MASSARRAY 시스템)을 이용한 후, 예를 들면, 일루미나 MPSS 시스템을 사용한 서열결정을 위해 이차 PCR을 이용하여 정방향 및 역방향 포획 서열들 및 분자 인덱스 바코드 둘 다를 도입하는 단계를 포함할 수 있다. 일부 실시양태에서, 마이크로플루이딕 방법이 이용될 수 있다. 일부 실시양태에서, 어레이 기초 마이크로플루이딕 방법이 이용될 수 있다. 예를 들면, 이러한 방법은 낮은 플렉스에서의 증폭을 위해 마이크로플루이딕 어레이(예를 들면, 플루이다임(Fluidigm))를 이용하는 단계, 및 인덱스 및 포획 프로브를 도입한 후 서열결정하는 단계를 포함할 수 있다. 일부 실시양태에서, 에멀젼 마이크로플루이딕 방법, 예를 들면, 디지털 소적 PCR이 이용될 수 있다.

[0084] 일부 실시양태에서, (예를 들면, 보편적인 또는 비-좌위 특이적 증폭 프라이머를 사용하는) 보편적인 증폭 방법이 이용될 수 있다. 일부 실시양태에서, 보편적인 증폭 방법이 풀-다운(pull-down) 방법과 함께 이용될 수 있다. 일부 실시양태에서, 방법은 보편적으로 증폭된 서열결정 라이브러리로부터의 바이오티닐화된 울트라머(ultramer) 풀-다운(예를 들면, 아질런트(Agilent) 또는 IDT로부터의 바이오티닐화된 풀-다운 분석)을 포함할 수 있다. 예를 들면, 이러한 방법은 표준 라이브러리를 제조하는 단계, 선택된 영역을 풀-다운 분석으로 농축하는 단계 및 이차 보편적인 증폭 단계를 포함할 수 있다. 일부 실시양태에서, 풀-다운 방법은 연결 기초 방법과 함께 이용될 수 있다. 일부 실시양태에서, 방법은 서열 특이적 어댑터 연결에 의한 바이오티닐화된 울트라머 풀-다운(예를 들면, HALOPLEX PCR, 할로 게노믹스(Halo Genomics))을 포함할 수 있다. 예를 들면, 이러한 방법은 선택자 프로브를 사용하여 제한 효소에 의해 분해된 단편을 포획한 후 포획된 생성물을 어댑터에 연결하는 단계 및 보편적인 증폭 후 서열결정을 수행하는 단계를 포함할 수 있다. 일부 실시양태에서, 풀-다운 방법은 연장 및 연결 기초 방법과 함께 이용될 수 있다. 일부 실시양태에서, 방법은 분자 역위 프로브(MIP) 연장 및 연결을 포함할 수 있다. 예를 들면, 이러한 방법은 서열 어댑터와 함께 분자 역위 프로브를 사용한 후 보편적인 증폭 및 서열결정을 수행하는 단계를 포함할 수 있다. 일부 실시양태에서, 상보적 DNA가 증폭 없이 합성되고 서열결정될 수 있다.

[0085] 일부 실시양태에서, 연장 및 연결 방법은 풀-다운 성분 없이 수행될 수 있다. 일부 실시양태에서, 방법은 좌위 특이적 정방향 및 역방향 프라이머 혼성화, 연장 및 연결을 포함할 수 있다. 이러한 방법은 보편적인 증폭 또는 증폭 없는 상보적 DNA 합성 후 서열결정을 수행하는 단계를 추가로 포함할 수 있다. 일부 실시양태에서, 이러한 방법은 분석 동안 배경 서열을 감소시킬 수 있거나 배제시킬 수 있다.

[0086] 일부 실시양태에서, 풀-다운 방법은 임의적인 증폭 성분과 함께 이용될 수 있거나 증폭 성분 없이 이용될 수 있다. 일부 실시양태에서, 방법은 보편적인 증폭 없이 포획 프로브를 완전히 도입하는 변경된 풀-다운 분석 및 연결을 포함할 수 있다. 예를 들면, 이러한 방법은 변경된 선택자 프로브를 사용하여 제한 효소에 의해 분해된 단편을 포획한 후 포획된 생성물을 어댑터에 연결하는 단계, 임의적인 증폭 단계 및 서열결정 단계를 포함할 수 있다. 일부 실시양태에서, 방법은 환형 단일 가닥 연결과 함께 어댑터 서열의 연장 및 연결을 이용하는 바이오티닐화된 풀-다운 분석을 포함할 수 있다. 예를 들면, 이러한 방법은 선택자 프로브를 사용하여 관심있는 영역(즉, 표적 서열)을 포획하는 단계, 프로브 연장 단계, 어댑터 연결 단계, 단일 가닥 환형 연결 단계, 임의적인 증폭 단계 및 서열결정 단계를 포함할 수 있다. 일부 실시양태에서, 서열결정 결과의 분석은 배경으로부터 표적 서열을 분리할 수 있다.

[0087] 일부 실시양태에서, 본원에 기재된 하나 이상의 서열 기초 분리 방법을 이용하여 선택 게놈 영역(예를 들면, 염색체)로부터의 단편에 대해 핵산을 농축한다. 서열 기초 분리는 일반적으로 관심있는 단편(예를 들면, 표적 및/또는 기준 단편)에 존재하고 샘플의 다른 단편에 실질적으로 존재하지 않거나 비-실질적인 양(예를 들면, 5% 이하)으로 상기 다른 단편에 존재하는 뉴클레오티드 서열에 기초한다. 일부 실시양태에서, 서열 기초 분리는 분리된 표적 단편 및/또는 분리된 기준 단편을 생성할 수 있다. 분리된 표적 단편 및/또는 분리된 기준 단편은 종종 핵산 샘플에 남아있는 단편들로부터 단리된다. 일부 실시양태에서, 분리된 표적 단편 및 분리된 기준 단편은 서로로부터 단리된다(예를 들면, 별도의 분석 구획에서 단리된다). 일부 실시양태에서, 분리된 표적 단편 및 분리된 기준 단편은 함께 단리된다(예를 들면, 동일한 분석 구획에서 단리된다). 일부 실시양태에서, 결합되지 않은

단편은 상이하게 제거될 수 있거나, 분해될 수 있거나 절단될 수 있다.

[0088] 일부 실시양태에서, 선택적 핵산 포획 공정을 이용하여 핵산 샘플로부터 표적 및/또는 기준 단편을 분리한다. 상업적으로 입수가능한 핵산 포획 시스템은 예를 들면, 님블레겐(Nimblegen) 서열 포획 시스템(로슈 님블레겐(Roche NimbleGen), 미국 위스콘신주 매디슨 소재); 일루미나 비드어레이(BEADARRAY) 플랫폼(일루미나, 미국 캘리포니아주 샌 디에고 소재); 아피메트릭스 진칩(Affymetrix GENECHIP) 플랫폼(아피메트릭스, 미국 캘리포니아주 산타 클라라 소재); 아질런트 슈어셀렉트(SureSelect) 표적 농축 시스템(아질런트 테크놀로지스(Agilent Technologies), 미국 캘리포니아주 산타 클라라 소재); 및 관련 플랫폼들을 포함한다. 이러한 방법은 전형적으로 포획 올리고뉴클레오티드를 표적 또는 기준 단편의 뉴클레오티드 서열의 분절 또는 전체에 혼성화시키는 단계를 포함하고, 고체상(예를 들면, 고체상 어레이) 및/또는 용액 기제 플랫폼의 사용을 포함할 수 있다. 포획 올리고뉴클레오티드(종종 "미끼(bait)"로서 지칭됨)는 선택된 게놈 영역 또는 좌위(예를 들면, 21번, 18번, 13번, X 및 Y 염색체 중 하나, 또는 기준 염색체)로부터의 핵산 단편에 우선적으로 혼성화하도록 선택될 수 있거나 디자인될 수 있다. 일부 실시양태에서, (예를 들면, 올리고뉴클레오티드 어레이를 사용하는) 혼성화 기초 방법을 이용하여 일부 염색체(예를 들면, 잠재적 이배수체 염색체, 기준 염색체 또는 관심있는 다른 염색체) 또는 이의 관심있는 분절로부터의 핵산 서열을 농축할 수 있다.

[0089] 일부 실시양태에서, 하나 이상의 길이 기초 분리 방법을 이용하여 특정 핵산 단편 길이, 길이 범위, 또는 특정 역치 또는 컷오프 미만 또는 초과의 길이에 대해 핵산을 농축한다. 핵산 단편 길이는 전형적으로 단편 내의 뉴클레오티드의 수를 지칭한다. 핵산 단편 길이는 종종 핵산 단편 크기로서도 지칭된다. 일부 실시양태에서, 개별 단편의 길이를 측정하지 않으면서 길이 기초 분리 방법을 수행한다. 일부 실시양태에서, 개별 단편의 길이를 측정하는 방법과 함께 길이 기초 분리 방법을 수행한다. 일부 실시양태에서, 길이 기초 분리는 크기 분별 절차를 지칭하고, 이때 분별된 풀의 전부 또는 일부가 단리될(예를 들면, 보유될) 수 있고/있거나 분석될 수 있다. 크기 분별 절차는 당분야에서 공지되어 있다(예를 들면, 어레이 상에서의 분리, 분자체에 의한 분리, 겔 전기영동에 의한 분리, 컬럼 크로마토그래피(예를 들면, 크기 배제 컬럼)에 의한 분리, 및 마이크로플루이딕 기초 방법). 일부 실시양태에서, 길이 기초 분리 방법은 예를 들면, 단편 환형화, 화학적 처리(예를 들면, 포름알데하이드, 폴리에틸렌 글리콜(PEG)), 질량 분광측정 및/또는 크기 특이적 핵산 증폭을 포함할 수 있다.

[0090] 본원에 기재된 방법과 함께 이용될 수 있는 일부 길이 기초 분리 방법들은 예를 들면, 선택적 서열 태깅 방법을 이용한다. 용어 "서열 태깅"은 인식가능하고 구별되는 서열을 핵산 또는 핵산 집단 내로 도입하는 것을 지칭한다. 본원에서 사용된 용어 "서열 태깅"은 본원에 나중에 기재된 용어 "서열 태그"와 상이한 의미를 갖는다. 이러한 서열 태깅 방법에서, 단편 크기 종(예를 들면, 짧은 단편) 핵산은 긴 핵산 및 짧은 핵산을 포함하는 샘플에서 선택적 서열 태깅으로 처리된다. 이러한 방법은 전형적으로 내부 프라이머 및 외부 프라이머를 포함하는 네스티드(nested) 프라이머 세트를 사용하여 핵산 증폭 반응을 수행하는 단계를 포함한다. 일부 실시양태에서, 내부 프라이머들 중 하나 또는 둘 다를 태깅하여 태그를 표적 증폭 생성물 내로 도입할 수 있다. 외부 프라이머는 일반적으로 (내부) 표적 서열을 보유하는 짧은 단편에 어닐링하지 않는다. 내부 프라이머는 짧은 단편에 어닐링할 수 있고 태그 및 표적 서열을 보유하는 증폭 생성물을 생성할 수 있다. 전형적으로, 긴 단편의 태깅은 예를 들면, 외부 프라이머의 사전 어닐링 및 연장에 의한 내부 프라이머의 차단된 연장을 포함하는 기작들의 조합을 통해 억제된다. 태깅된 단편의 농축은 예를 들면, 단일 가닥 핵산의 엑소뉴클레아제 분해 및 하나 이상의 태그에 대한 특이성을 갖는 증폭 프라이머를 사용한 태깅된 단편의 증폭을 포함하는 다양한 방법들 중 임의의 방법에 의해 달성될 수 있다.

[0091] 본원에 기재된 방법과 함께 이용될 수 있는 또 다른 길이 기초 분리 방법은 핵산 샘플을 폴리에틸렌 글리콜(PEG) 침전으로 처리하는 단계를 포함한다. 방법의 예는 국제 특허출원 공보 제WO2007/140417호 및 국제 특허출원 공보 제WO2010/115016호에 기재된 방법들을 포함한다. 이 방법은 일반적으로 작은(예를 들면, 300개 미만의 뉴클레오티드) 핵산을 실질적으로 침전시키지 않으면서 큰 핵산을 실질적으로 침전시키기에 충분한 조건 하에서 하나 이상의 1가 염의 존재 하에서 핵산 샘플을 PEG와 접촉시키는 단계를 수반한다.

[0092] 본원에 기재된 방법과 함께 이용될 수 있는 또 다른 크기 기초 농축 방법은 예를 들면, 서클리가제(circligase)를 사용한 연결에 의한 환형화를 포함한다. 짧은 핵산 단편은 전형적으로 긴 단편보다 더 높은 효율로 환형화될 수 있다. 환형화되지 않은 서열은 환형화된 서열로부터 분리될 수 있고, 농축된 짧은 단편은 추가 분석을 위해 사용될 수 있다.

[0093] 핵산 라이브러리

[0094] 일부 실시양태에서, 핵산 라이브러리는 특정 공정을 위해 제조되고/되거나, 조립되고/되거나 변경되는 복수의

폴리뉴클레오티드 분자들(예를 들면, 핵산의 샘플)이고, 상기 공정의 비한정적 예는 고체상(예를 들면, 고체 지지체, 예를 들면, 유동 셀, 비드) 상에의 고정, 농축, 증폭, 클로닝, 검출 및/또는 핵산 서열결정을 포함한다. 일부 실시양태에서, 핵산 라이브러리는 서열결정 공정 전 또는 동안에 제조된다. 핵산 라이브러리(예를 들면, 서열결정 라이브러리)는 당분야에서 공지된 적합한 방법에 의해 제조될 수 있다. 핵산 라이브러리는 표적화된 또는 표적화되지 않은 제조 공정에 의해 제조될 수 있다.

[0095] 일부 실시양태에서, 핵산의 라이브러리는 핵산을 고체 지지체에 고정시키도록 구성된 화학 모이어티(예를 들면, 작용기)를 포함하도록 변경된다. 일부 실시양태에서, 핵산의 라이브러리는 라이브러리를 고체 지지체에 고정시키도록 구성된 생체분자(예를 들면, 작용기) 및/또는 결합 쌍의 구성원을 포함하도록 변경되고, 상기 고체 지지체의 비한정적 예는 티록신 결합 글로불린, 스테로이드 결합 단백질, 항체, 항원, 햅텐, 효소, 렉틴, 핵산, 리프레서, 단백질 A, 단백질 G, 아비딘, 스트렙타비딘, 바이오틴, 보체 성분 C1q, 핵산 결합 단백질, 수용체, 탄수화물, 올리고뉴클레오티드, 폴리뉴클레오티드, 상보적 핵산 서열 등 및 이들의 조합물을 포함한다. 특정 결합 쌍의 일부 예는 하기 결합 쌍들을 포함하나 이들로 한정되지 않는다: 아비딘 모이어티 및 바이오틴 모이어티; 항원성 에피토프 및 항체 또는 이의 면역학적 반응성 단편; 항체 및 햅텐; 디곡시겐 모이어티 및 항-디곡시겐 항체; 플루오레세인 모이어티 및 항-플루오레세인 항체; 오퍼레이터 및 리프레서; 뉴클레아제 및 뉴클레오티드; 렉틴 및 폴리사카라이드; 스테로이드 및 스테로이드 결합 단백질; 활성 화합물 및 활성 화합물 수용체; 호르몬 및 호르몬 수용체; 효소 및 기질; 면역글로불린 및 단백질 A; 올리고뉴클레오티드 또는 폴리뉴클레오티드 및 이의 상응하는 상보체 등; 또는 이들의 조합물.

[0096] 일부 실시양태에서, 핵산의 라이브러리는 공지된 조성의 하나 이상의 폴리뉴클레오티드를 포함하도록 변경되고, 이때 상기 폴리뉴클레오티드의 비한정적 예는 식별자(예를 들면, 태그, 인덱스화 태그), 포획 서열, 표지, 어댑터, 제한 효소 부위, 프로모터, 인핸서, 복제기점, 줄기 루프, 상보적 서열(예를 들면, 프라이머 결합 부위, 어닐링 부위), 적합한 삽입 부위(예를 들면, 트랜스포존, 바이러스 삽입 부위), 변경된 뉴클레오티드 등 또는 이들의 조합물을 포함한다. 공지된 서열의 폴리뉴클레오티드는 적합한 위치, 예를 들면, 5' 말단, 3' 말단 또는 핵산 서열 내에서 추가될 수 있다. 공지된 서열의 폴리뉴클레오티드는 동일한 또는 상이한 서열일 수 있다. 일부 실시양태에서, 공지된 서열의 폴리뉴클레오티드는 표면(예를 들면, 유동 셀의 표면) 상에 고정된 하나 이상의 올리고뉴클레오티드에 혼성화하도록 구성된다. 예를 들면, 5' 공지된 서열을 포함하는 핵산 분자는 복수의 제1 올리고뉴클레오티드에 혼성화할 수 있는 반면, 3' 공지된 서열은 복수의 제2 올리고뉴클레오티드에 혼성화할 수 있다. 일부 실시양태에서, 핵산의 라이브러리는 염색체 특이적 태그, 포획 서열, 표지 및/또는 어댑터를 포함할 수 있다. 일부 실시양태에서, 핵산의 라이브러리는 하나 이상의 검출가능한 표지를 포함한다. 일부 실시양태에서, 하나 이상의 검출가능한 표지는 5' 말단, 3' 말단 및/또는 라이브러리의 핵산 내의 임의의 뉴클레오티드 위치에서 핵산 라이브러리 내로 도입될 수 있다. 일부 실시양태에서, 핵산의 라이브러리는 혼성화된 올리고뉴클레오티드를 포함한다. 일부 실시양태에서, 혼성화된 올리고뉴클레오티드는 표지된 프로브이다. 일부 실시양태에서, 핵산의 라이브러리는 고체상에의 고정 전에 혼성화된 올리고뉴클레오티드 프로브를 포함한다.

[0097] 일부 실시양태에서, 공지된 서열의 폴리뉴클레오티드는 보편적인 서열을 포함한다. 보편적인 서열은 2개 이상의 핵산 분자 또는 2개 이상의 핵산 분자 서브세트 내로 삽입되는 특정 뉴클레오티드 서열이고, 이때 보편적인 서열은 그 자신이 삽입되어 있는 모든 분자들 또는 분자 서브세트에 대해 동일하다. 보편적인 서열은 종종 보편적인 서열에 상보적인 단일 보편적인 프라이머를 사용하여 복수의 상이한 서열들에 혼성화하고/하거나 이러한 상이한 서열들을 증폭하도록 디자인된다. 일부 실시양태에서, 2개(예를 들면, 한 쌍) 이상의 보편적인 서열들 및/또는 보편적인 프라이머들이 사용된다. 보편적인 프라이머는 종종 보편적인 서열을 포함한다. 일부 실시양태에서, 어댑터(예를 들면, 보편적인 어댑터)는 보편적인 서열을 포함한다. 일부 실시양태에서, 하나 이상의 보편적인 서열을 사용하여 다수의 핵산 종 또는 서브세트를 포획하고/하거나, 확인하고/하거나 검출한다.

[0098] 핵산 라이브러리를 제조하는 일부 실시양태에서(예를 들면, 합성 절차에 의한 일부 서열결정에서), 핵산은 (예를 들면, 라이브러리 생성을 위한 제조에서) 수백 개 이하의 염기쌍 길이로 크기 선택되고/되거나 단편화된다. 일부 실시양태에서, 라이브러리 제조는 (예를 들면, ccfDNA를 사용할 때) 단편화 없이 수행된다.

[0099] 일부 실시양태에서, 연결 기초 라이브러리 제조 방법이 이용된다(예를 들면, 일루미나 TRUSEQ, 일루미나, 미국 캘리포니아주 샌 디에고 소재). 연결 기초 라이브러리 제조 방법은 초기 연결 단계에서 인덱스 서열을 도입할 수 있고 종종 단일-리드 서열결정, 페어링된-말단 서열결정 및 다중체화된 서열결정을 위한 샘플을 제조하는 데에 사용될 수 있는 어댑터(예를 들면, 메틸화된 어댑터) 디자인을 종종 이용한다. 예를 들면, 종종 핵산(예를 들면, 단편화된 핵산 또는 ccfDNA)은 필-인(fill-in) 반응, 엑소뉴클레아제 반응 또는 이들의 조합에 의해 말단 복구된다. 그 다음, 일부 실시양태에서, 생성된 블런트(blunt)-말단 복구된 핵산은 어댑터/프라이머의 3' 말단

상의 단일 뉴클레오티드 오버행(overhang)에 상보적인 단일 뉴클레오티드에 의해 연장될 수 있다. 임의의 뉴클레오티드가 연장/오버행 뉴클레오티드용으로 사용될 수 있다. 일부 실시양태에서, 핵산 라이브러리 제조는 어댑터 올리고뉴클레오티드를 연결하는 단계를 포함한다. 어댑터 올리고뉴클레오티드는 종종 유동 셀 고착제(anchors)에 상보적이고, 종종 핵산 라이브러리를 고체 지지체, 예컨대, 유동 셀의 내부 표면에 고정시키는 데에 사용된다. 일부 실시양태에서, 어댑터 올리고뉴클레오티드는 식별자, 하나 이상의 서열결정 프라이머 혼성화 부위(예를 들면, 보편적인 서열결정 프라이머, 단일 말단 서열결정 프라이머, 페어링된-말단 서열결정 프라이머, 다중체화된 서열결정 프라이머 등에 상보적인 서열) 또는 이들의 조합(예를 들면, 어댑터/서열결정, 어댑터/식별자, 어댑터/식별자/서열결정)을 포함한다.

[0100] 식별자는 식별자를 포함하는 핵산의 검출 및/또는 식별을 가능하게 하는, 핵산(예를 들면, 폴리뉴클레오티드) 내로 도입된 또는 이러한 핵산에 부착된 적합한 검출가능한 표지일 수 있다. 일부 실시양태에서, 식별자는 서열결정 방법 동안에 (예를 들면, 중합효소에 의해) 핵산 내로 도입되거나 핵산에 부착된다. 식별자의 비한정적 예는 핵산 태그, 핵산 인덱스 또는 바코드, 방사성표지(예를 들면, 동위원소), 금속성 표지, 형광 표지, 화학발광 표지, 인광 표지, 형광단 소광제(quencher), 염료, 단백질(예를 들면, 효소, 항체 또는 이의 부분, 링커, 결합쌍의 구성원) 등 또는 이들의 조합물을 포함한다. 일부 실시양태에서, 식별자(예를 들면, 핵산 인덱스 또는 바코드)는 뉴클레오티드 또는 뉴클레오티드 유사체의 유일한, 공지된 및/또는 식별가능한 서열이다. 일부 실시양태에서, 식별자는 6개 이상의 연속 뉴클레오티드이다. 다양한 상이한 여기 및 방사 스펙트럼을 갖는 다수의 형광단들이 사용될 수 있다. 임의의 적합한 종류 및/또는 수의 형광단들이 식별자로서 사용될 수 있다. 일부 실시양태에서, 1개 이상, 2개 이상, 3개 이상, 4개 이상, 5개 이상, 6개 이상, 7개 이상, 8개 이상, 9개 이상, 10개 이상, 20개 이상, 30개 이상 또는 50개 이상의 상이한 식별자들이 본원에 기재된 방법(예를 들면, 핵산 검출 및/또는 서열결정 방법)에서 사용된다. 일부 실시양태에서, 1종 또는 2종의 식별자(예를 들면, 형광 표지)가 라이브러리의 각각의 핵산에 연결된다. 식별자의 검출 및/또는 정량은 적합한 방법 또는 장치에 의해 수행될 수 있고, 이러한 방법 또는 장치의 비한정적 예는 유동 세포측정, 정량 중합효소 연쇄 반응(qPCR), 겔 전기영동, 광도계(luminometer), 형광계(fluorometer), 분광광도계, 적합한 유전자-칩 또는 마이크로어레이 분석, 웨스턴 블롯, 질량 분광측정, 크로마토그래피, 세포형광측정(cytofluorimetric) 분석, 형광 현미경관찰, 적합한 형광 또는 디지털 영상화 방법, 공초점 레이저 스캐닝 현미경관찰, 레이저 스캐닝 세포측정, 친화성 크로마토그래피, 수동 배치 모드(batch mode) 분리, 전기장 현탁, 적합한 핵산 서열결정 방법 및/또는 핵산 서열결정 장치 등 및 이들의 조합을 포함한다.

[0101] 일부 실시양태에서, 트랜스포존(transposon) 기초 라이브러리 제조 방법이 이용된다(예를 들면, EPICENTRE NEXTERA, 에피센터(Epicentre), 미국 위스콘신주 매디슨 소재). 트랜스포존 기초 방법은 전형적으로 (종종 플랫폼 특이적 태그 및 임의적인 바코드의 도입을 가능하게 하는) 단일 튜브 반응에서 시험관내 전위(transposition)를 이용하여 동시에 DNA를 단편화하고 태깅하고, 서열결정기-준비된(sequencer-ready) 라이브러리를 제조한다.

[0102] 일부 실시양태에서, 핵산 라이브러리 또는 이의 부분이 증폭된다(예를 들면, PCR 기초 방법에 의해 증폭된다). 일부 실시양태에서, 서열결정 방법은 핵산 라이브러리의 증폭을 포함한다. 핵산 라이브러리는 고체 지지체(예를 들면, 유동 셀 내의 고체 지지체) 상에의 고정 전 또는 후에 증폭될 수 있다. 핵산 증폭은 주형 및/또는 이의 상보체의 하나 이상의 카피를 생성함으로써 (예를 들면, 핵산 라이브러리에) 존재하는 핵산 주형 및/또는 이의 상보체의 수를 증폭하거나 증가시키는 공정을 포함한다. 증폭은 적합한 방법에 의해 수행될 수 있다. 핵산 라이브러리는 열순환(thermocycling) 방법 또는 등온 증폭 방법에 의해 증폭될 수 있다. 일부 실시양태에서, 롤링 서클 증폭 방법이 이용된다. 일부 실시양태에서, 증폭은 (예를 들면, 유동 셀 내의) 고체 지지체 상에서 일어나고, 이때 핵산 라이브러리 또는 이의 부분이 고정된다. 일부 서열결정 방법에서, 핵산 라이브러리는 유동 셀에 첨가되고, 적합한 조건 하에서 고착제와의 혼성화에 의해 고정된다. 이러한 종류의 핵산 증폭은 종종 고체상 증폭으로서 지칭된다. 고체상 증폭의 일부 실시양태에서, 증폭된 생성물의 전부 또는 부분은 고정된 프라이머로부터 시작되는 연장에 의해 합성된다. 고체상 증폭 방법은 하나 이상의 증폭 올리고뉴클레오티드(예를 들면, 프라이머)가 고체 지지체 상에 고정된다는 점을 제외하고 표준 용액상 증폭과 유사하다.

[0103] 일부 실시양태에서, 고체상 증폭은 표면에 고정된 1개의 올리고뉴클레오티드 프라이머 종만을 포함하는 핵산 증폭 반응을 포함한다. 일부 실시양태에서, 고체상 증폭은 복수의 상이한 고정된 올리고뉴클레오티드 프라이머 종들을 포함한다. 일부 실시양태에서, 고체상 증폭은 고체 표면 상에 고정된 1개의 올리고뉴클레오티드 프라이머 종, 및 용액 중의 상이한 제2 올리고뉴클레오티드 프라이머 종을 포함하는 핵산 증폭 반응을 포함할 수 있다. 다수의 상이한 고정된 또는 용액 기초 프라이머 종들이 사용될 수 있다. 고체상 핵산 증폭 반응의 비한정적 예

는 계면 중합, 가교 중합, 에멀젼 PCR, 와일드파이어(WildFire) 중합(예를 들면, 미국 특허출원 공보 제 20130012399호) 등 또는 이들의 조합을 포함한다.

[0104] 서열결정

[0105] 일부 실시양태에서, 핵산(예를 들면, 핵산 단편, 샘플 핵산, 무세포 핵산)이 서열결정된다. 일부 실시양태에서, 전체 또는 실질적으로 전체 서열이 수득되고 종종 부분 서열이 수득된다.

[0106] 일부 실시양태에서, 샘플 중의 일부 또는 모든 핵산들이 서열결정 전 또는 동안에 (예를 들면, 비특이적으로, 예를 들면, PCR 기초 방법에 의해) 농축되고/되거나 증폭된다. 일부 실시양태에서, 샘플 중의 특정 핵산 부분 또는 서브세트가 서열결정 전 또는 동안에 농축되고/되거나 증폭된다. 일부 실시양태에서, 핵산의 미리 선택된 풀의 부분 또는 서브세트가 무작위적으로 서열결정된다. 일부 실시양태에서, 샘플 중의 핵산은 서열결정 전 또는 동안에 농축되지 않고/않거나 증폭되지 않는다.

[0107] 본원에서 사용된 바와 같이, "리드"(즉, "리드", "서열 리드")는 본원에 기재되어 있거나 당분야에서 공지되어 있는 임의의 서열결정 공정에 의해 생성된 짧은 뉴클레오티드 서열이다. 리드는 핵산 단편의 한 말단으로부터 생성될 수 있고("단일-말단 리드"), 종종 핵산의 양 말단으로부터 생성된다(예를 들면, 페어링된-말단 리드, 이중-말단 리드).

[0108] 서열 리드의 길이는 종종 특정 서열결정 기술과 관련되어 있다. 예를 들면, 고속처리 방법은 크기 면에서 수십 개 내지 수백 개의 염기쌍(bp)만큼 상이할 수 있는 서열 리드를 제공한다. 예를 들면, 나노공극(Nanopore) 서열결정은 크기 면에서 수십 개 내지 수천 개의 염기쌍만큼 상이할 수 있는 서열 리드를 제공할 수 있다. 일부 실시양태에서, 서열 리드는 약 15 bp 내지 약 900 bp 길이의 평균, 중간, 평균치 또는 절대 길이를 갖는다. 일부 실시양태에서, 서열 리드는 약 100 bp 이상의 평균, 중간, 평균치 또는 절대 길이를 갖는다.

[0109] 일부 실시양태에서, 단일-말단 리드의 공칭, 평균치, 평균 또는 절대 길이는 종종 약 15개의 연속 뉴클레오티드 내지 약 50개 이상의 연속 뉴클레오티드, 약 15개의 연속 뉴클레오티드 내지 약 40개 이상의 연속 뉴클레오티드, 종종 약 15개의 연속 뉴클레오티드 또는 약 36개 이상의 연속 뉴클레오티드이다. 일부 실시양태에서, 단일-말단 리드의 공칭, 평균치, 평균 또는 절대 길이는 약 20개 내지 약 30개 염기, 또는 약 24개 내지 약 28개 염기이다. 일부 실시양태에서, 단일-말단 리드의 공칭, 평균치, 평균 또는 절대 길이는 약 1개, 2개, 3개, 4개, 5개, 6개, 7개, 8개, 9개, 10개, 11개, 12개, 13개, 14개, 15개, 16개, 17개, 18개, 19개, 21개, 22개, 23개, 24개, 25개, 26개, 27개, 28개 또는 약 29개 염기 이상이다.

[0110] 일부 실시양태에서, 페어링된-말단 리드의 공칭, 평균치, 평균 또는 절대 길이는 약 10개의 연속 뉴클레오티드 내지 약 25개의 연속 뉴클레오티드 이상(예를 들면, 길이에서 약 10개, 11개, 12개, 13개, 14개, 15개, 16개, 17개, 18개, 19개, 20개, 21개, 22개, 23개, 24개 또는 25개의 뉴클레오티드 이상), 또는 약 15개의 연속 뉴클레오티드 내지 약 20개의 연속 뉴클레오티드 이상이고, 종종 약 17개의 연속 뉴클레오티드 또는 약 18개의 연속 뉴클레오티드이다.

[0111] 리드는 일반적으로 물리적 핵산에서 뉴클레오티드 서열의 표시물이다. 예를 들면, ATGC로 표시된 서열을 함유하는 리드에서, "A"는 물리적 핵산에서 아데닌 뉴클레오티드를 표시하고, "T"는 물리적 핵산에서 타이민 뉴클레오티드를 표시하고, "G"는 물리적 핵산에서 구아닌 뉴클레오티드를 표시하고, "C"는 물리적 핵산에서 사이토신 뉴클레오티드를 표시한다. 임신 여성의 혈액으로부터 수득된 서열 리드는 태아 핵산과 모체 핵산의 혼합물로부터의 리드일 수 있다. 상대적으로 짧은 리드의 혼합물은 본원에 기재된 공정에 의해 임신 여성 및/또는 태아에 존재하는 게놈 핵산의 표시물로 변환될 수 있다. 예를 들면, 상대적으로 짧은 리드의 혼합물은 카피 수 변이(예를 들면, 모체 및/또는 태아 카피 수 변이), 유전적 변이 또는 이배수성, 미세중복 또는 미세결실의 표시물로 변환될 수 있다. 모체 핵산과 태아 핵산의 혼합물의 리드는 모체 염색체 및 태아 염색체 중 하나 또는 둘 다의 특징을 포함하는 복합 염색체 또는 이의 분절의 표시물로 변환될 수 있다. 일부 실시양태에서, 대상체로부터의 샘플의 핵산 서열 리드를 "수득하는 것" 및/또는 1명 이상의 기준 개체로부터의 생물학적 표본의 핵산 서열 리드를 "수득하는 것"은 핵산을 직접적으로 서열결정하여 서열 정보를 수득하는 것을 포함할 수 있다. 일부 실시양태에서, "수득하는 것"은 또 다른 개체에 의해 핵산으로부터 직접적으로 수득된 서열 정보를 제공받는 것을 포함할 수 있다.

[0112] 일부 실시양태에서, 게놈의 대표적인 부분이 서열결정되고 종종 "커버리지(coverage)" 또는 "배(fold) 커버리지"로서 지칭된다. 예를 들면, 1배 커버리지는 게놈의 뉴클레오티드 서열의 대략 100%가 리드로 표시된다는 것을 시사한다. 일부 실시양태에서, "배 커버리지"는 기준물로서 실시된 사전 서열결정을 지칭하는 상대적

용어이다. 예를 들면, 제2 서열결정 실시는 제1 서열결정 실시보다 2배 낮은 커버리지를 가질 수 있다. 일부 실시양태에서, 게놈은 중복적으로 서열결정되고, 이때 게놈의 주어진 영역은 2개 이상의 리드 또는 중첩 리드에 의해 커버될 수 있다(예를 들면, 1 초과의 "배 커버리지", 예를 들면, 2배 커버리지).

[0113] 일부 실시양태에서, 1명의 개체로부터의 1개 핵산 샘플이 서열결정된다. 일부 실시양태에서, 2개 이상의 샘플들 각각으로부터의 핵산이 서열결정되고, 이때 샘플들은 1명의 개체로부터 유래되거나 상이한 개체들로부터 유래된다. 일부 실시양태에서, 2개 이상의 생물학적 샘플들로부터의 핵산 샘플들이 풀링되고, 이때 각각의 생물학적 샘플이 1명의 개체 또는 2명 이상의 개체들로부터 유래되고, 풀이 서열결정된다. 후자 실시양태에서, 각각의 생물학적 샘플로부터의 핵산 샘플은 종종 하나 이상의 유일한 식별자에 의해 식별된다.

[0114] 일부 실시양태에서, 서열결정 방법은 서열결정 공정에서 서열 반응의 다중체화를 가능하게 하는 식별자를 사용한다. 유일한 식별자의 수가 클수록, 예를 들면, 서열결정 공정에서 다중체화될 수 있는 검출될 샘플 및/또는 염색체의 수가 커진다. 서열결정 공정은 임의의 적합한 수(예를 들면, 4개, 8개, 12개, 24개, 48개 또는 96개 이상)의 유일한 식별자를 사용함으로써 수행될 수 있다.

[0115] 서열결정 공정은 종종 고체상을 사용하고, 종종 고체상은 라이브러리로부터의 핵산이 부착될 수 있고 시약이 유동될 수 있고 부착된 핵산과 접촉될 수 있는 유동 셀을 포함한다. 유동 셀은 종종 유동 셀 레인을 포함하고, 식별자의 사용은 각각의 레인에서 다수의 샘플들을 분석하는 것을 용이하게 할 수 있다. 유동 셀은 종종 시약 용액이 결합된 분석물 상에서 차례로 통과하는 것을 유지하고/하거나 가능하게 하도록 구성될 수 있는 고체 지지체이다. 유동 셀은 종종 형태 면에서 평면이고, 광학적으로 투명하고, 일반적으로 밀리미터 또는 밀리미터 미만의 크기를 갖고, 종종 분석물/시약 상호작용이 일어나는 채널 또는 레인을 갖는다. 일부 실시양태에서, 주어진 유동 셀 레인에서 분석되는 샘플의 수는 단일 유동 셀 레인에서 라이브러리 제조 및/또는 프로브 디자인 동안에 사용된 유일한 식별자의 수에 의존한다. 예를 들면, 12개의 식별자를 사용하는 다중체화는 8개 레인 유동 셀에서 96개의 샘플들(예를 들면, 96웰 마이크로웰 플레이트 내의 웰의 수와 동등함)의 동시적인 분석을 가능하게 한다. 유사하게, 예를 들면, 48개의 식별자를 사용하는 다중체화는 8개 레인 유동 셀에서 384개의 샘플들(예를 들면, 384웰 마이크로웰 플레이트 내의 웰의 수와 동등함)의 동시적인 분석을 가능하게 한다. 상업적으로 입수 가능한 다중체 서열결정 키트의 비한정적 예는 일루미나의 다중체화 샘플 제조 올리고뉴클레오티드 키트 및 다중체화 서열결정 프라이머 및 PhiX 대조군 키트(예를 들면, 각각 일루미나의 카탈로그 번호 PE-400-1001 및 PE-400-1002)를 포함한다.

[0116] 임의의 적합한 핵산 서열결정 방법이 이용될 수 있고, 이러한 방법의 비한정적 예는 막심 & 길버트(Maxim & Gilbert), 쇄-종결 방법, 합성에 의한 서열결정, 연결에 의한 서열결정, 질량 분광측정에 의한 서열결정, 현미경관찰 기초 기법 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, 제1 세대 기술, 예를 들면, 마이크로플루이딕 생거(Sanger) 서열결정을 포함하는 자동화된 생거 서열결정 방법을 포함하는 생거 서열결정 방법이 본원에서 제공된 방법에서 이용될 수 있다. 일부 실시양태에서, 핵산 영상화 기술(예를 들면, 투과 전자 현미경관찰(TEM) 및 원자력 현미경관찰(AFM))의 사용을 포함하는 서열결정 기술이 이용될 수 있다. 일부 실시양태에서, 고속처리 서열결정 방법이 이용된다. 고속처리 서열결정 방법은 일반적으로 종종 유동 셀 내에서 대량 병렬 방식으로 서열결정되는 클론적으로 증폭된 DNA 주형 또는 단일 DNA 분자를 사용한다. 대량 병렬 방식으로 DNA를 서열결정할 수 있는 차세대(예를 들면, 2세대 및 3세대) 서열결정 기법이 본원에 기재된 방법을 위해 이용될 수 있고 본원에서 "대량 병렬 서열결정"(MPS)으로서 총칭된다. 일부 실시양태에서, MPS 서열결정 방법은 표적화된 방식을 이용하고, 이때 관심있는 특정 염색체, 유전자 또는 영역이 서열결정된다. 일부 실시양태에서, 비-표적화된 방식이 이용되고, 이때 샘플 중의 대다수 또는 모든 핵산들이 무작위적으로 서열결정되고/되거나, 증폭되고/되거나 포획된다.

[0117] 일부 실시양태에서, 표적화된 농축, 증폭 및/또는 서열결정 방식이 이용된다. 표적화된 방식은 종종 서열 특이적 올리고뉴클레오티드의 사용에 의한 추가 프로세싱을 위해 샘플 중의 핵산 서브세트를 단리하고/하거나, 선택하고/하거나 농축한다. 일부 실시양태에서, 서열 특이적 올리고뉴클레오티드의 라이브러리를 사용하여 샘플 중의 하나 이상의 핵산 세트를 표적화한다(예를 들면, 이 핵산 세트에 혼성화한다). 서열 특이적 올리고뉴클레오티드 및/또는 프라이머는 종종 관심있는 하나 이상의 염색체, 유전자, 엑손, 인트론 및/또는 조절 영역에 존재하는 특정 서열(예를 들면, 유일한 핵산 서열)에 대한 선택성을 갖는다. 임의의 적합한 방법 또는 방법의 조합이 하나 이상의 표적화된 핵산 서브세트의 농축, 증폭 및/또는 서열결정을 위해 이용될 수 있다. 일부 실시양태에서, 표적화된 서열은 하나 이상의 서열 특이적 고착제를 사용하여 고체상(예를 들면, 유동 셀, 비드)에 포획시킴으로써 단리되고/되거나 농축된다. 일부 실시양태에서, 표적화된 서열은 서열 특이적 프라이머 및/또는 프라이머 세트를 사용하는 중합효소 기초 방법(예를 들면, PCR 기초 방법, 임의의 적합한 중합효소 기초 연장)에

의해 농축되고/되거나 증폭된다. 서열 특이적 고착제는 종종 서열 특이적 프라이머로서 사용될 수 있다.

[0118] MPS 서열결정은 종종 합성 및 일부 영상화 공정에 의한 서열결정을 이용한다. 본원에 기재된 방법에서 이용될 수 있는 핵산 서열결정 기술은 합성에 의한 서열결정 및 가역적 종결요소(terminator) 기초 서열결정(예를 들면, 일루미나의 게놈 분석기; 게놈 분석기 II; HISEQ 2000; HISEQ 2500(일루미나, 미국 캘리포니아주 샌 디에고 소재)이다. 이 기술을 이용하여 수백만 개의 핵산(예를 들면, DNA) 단편들을 동시에 서열결정할 수 있다. 이러한 종류의 서열결정 기술의 일례에서, 그의 표면 상에 올리고뉴클레오티드 고착제(예를 들면, 어댑터 프라이머)가 결합되어 있는 8개의 개별 레인들을 갖는 광학적으로 투명한 슬라이드를 함유하는 유동 셀이 사용된다. 유동 셀은 종종 시약 용액이 결합된 분석물 상에서 차례로 통과하는 것을 유지하고/하거나 가능하게 하도록 구성될 수 있는 고체 지지체이다. 유동 셀은 종종 형태 면에서 평면이고, 광학적으로 투명하고, 일반적으로 밀리미터 또는 밀리미터 미만의 크기를 갖고, 종종 분석물/시약 상호작용이 일어나는 채널 또는 레인을 갖는다.

[0119] 일부 실시양태에서, 합성에 의한 서열결정은 주형-지정된 방식으로 뉴클레오티드를 프라이머 또는 기존 핵산 가닥에 (예를 들면, 공유 추가로) 반복적으로 추가하는 단계를 포함한다. 각각의 반복적 뉴클레오티드 추가가 검출되고, 핵산 가닥의 서열이 수득될 때까지 상기 공정이 다회 반복된다. 수득된 서열의 길이는 수행되는 추가 및 검출 단계의 수에 부분적으로 의존한다. 합성에 의한 서열결정의 일부 실시양태에서, 동일한 종류의 1개, 2개 또는 3개 이상의 뉴클레오티드(예를 들면, A, G, C 또는 T)가 뉴클레오티드 추가 라운드에서 추가되고 검출된다. 뉴클레오티드는 임의의 적합한 방법(예를 들면, 효소적 또는 화학적 방법)에 의해 추가될 수 있다. 예를 들면, 일부 실시양태에서, 중합효소 또는 연결효소(ligase)가 주형-지정된 방식으로 뉴클레오티드를 프라이머 또는 기존 핵산 가닥에 추가한다. 합성에 의한 서열결정의 일부 실시양태에서, 상이한 종류의 뉴클레오티드, 뉴클레오티드 유사체 및/또는 식별자가 사용된다. 일부 실시양태에서, 가역적 종결요소 및/또는 제거가능한(예를 들면, 절단가능한) 식별자가 사용된다. 일부 실시양태에서, 형광 표지된 뉴클레오티드 및/또는 뉴클레오티드 유사체가 사용된다. 일부 실시양태에서, 합성에 의한 서열결정은 절단(예를 들면, 식별자의 절단 및 제거) 및/또는 세척 단계를 포함한다. 일부 실시양태에서, 하나 이상의 뉴클레오티드의 추가는 본원에 기재되어 있거나 당분야에서 공지되어 있는 적합한 방법에 의해 검출되고, 이러한 방법의 비한정적 예는 임의의 적합한 영상화 장치, 적합한 카메라, 디지털 카메라, CCD(전하 커플 디바이스) 기초 영상화 장치(예를 들면, CCD 카메라), CMOS(상보적 금속 산화물 규소) 기초 영상화 장치(예를 들면, CMOS 카메라), 광 다이오드(예를 들면, 광전자증배관), 전자 현미경관찰, 필드-효과 트랜지스터(예를 들면, DNA 필드-효과 트랜지스터), ISFET 이온 센서(예를 들면, CHEMFET 센서) 등 또는 이들의 조합을 포함한다. 본원의 방법을 수행하기 위해 이용될 수 있는 다른 서열결정 방법은 디지털 PCR 및 혼성화에 의한 서열결정을 포함한다.

[0120] 본원의 방법을 수행하기 위해 이용될 수 있는 다른 서열결정 방법은 디지털 PCR 및 혼성화에 의한 서열결정을 포함한다. 디지털 중합효소 연쇄 반응(디지탈 PCR 또는 dPCR)은 샘플에서 핵산을 직접적으로 확인하고 정량하는 데에 이용될 수 있다. 일부 실시양태에서, 디지털 PCR은 에멀젼에서 수행될 수 있다. 예를 들면, 개별 핵산들이 예를 들면, 마이크로플루이딕 챔버 디바이스에서 분리되고, 각각의 핵산이 PCR에 의해 개별적으로 증폭된다. 핵산은 웰 당 1개 이하의 핵산이 존재하도록 분리될 수 있다. 일부 실시양태에서, 상이한 프로브들이 다양한 대립형질(예를 들면, 태아 대립형질 및 모체 대립형질)을 식별하는 데에 사용될 수 있다. 카피 수를 측정하기 위해 대립형질의 수를 셀 수 있다.

[0121] 일부 실시양태에서, 혼성화에 의한 서열결정이 이용될 수 있다. 이 방법은 복수의 폴리뉴클레오티드 서열들을 복수의 폴리뉴클레오티드 프로브들과 접촉시키는 단계를 포함하고, 이때 복수의 폴리뉴클레오티드 프로브들 각각은 임의적으로 기판에 고착될 수 있다. 일부 실시양태에서, 기판은 공지된 뉴클레오티드 서열의 어레이를 갖는 평평한 표면일 수 있다. 어레이에의 혼성화의 패턴은 샘플에 존재하는 폴리뉴클레오티드 서열을 확인하는 데에 이용될 수 있다. 일부 실시양태에서, 각각의 프로브는 비드, 예를 들면, 자성 비드 등에 고착된다. 비드에의 혼성화는 확인될 수 있고 샘플 내의 복수의 폴리뉴클레오티드 서열들을 확인하는 데에 이용될 수 있다.

[0122] 일부 실시양태에서, 나노공극 서열결정이 본원에 기재된 방법에서 이용될 수 있다. 나노공극 서열결정은 단일 분자 서열결정 기술이고, 이 기술에 의해 단일 핵산 분자(예를 들면, DNA)가 나노공극을 통과할 때 직접적으로 서열결정된다.

[0123] 본원에 기재된 방법을 수행하기 위한 적합한 MPS 방법, 시스템 또는 기술 플랫폼이 핵산 서열결정 리드를 수득하는 데에 이용될 수 있다. MPS 플랫폼의 비한정적 예는 일루미나/솔렉스(Solex)/HiSeq(예를 들면, 일루미나의 게놈 분석기; 게놈 분석기 II; HISEQ 2000; HISEQ), SOLiD, 로슈(Roche)/454, PACBIO 및/또는 SMRT, 헬리코스 트루(Helicos True) 단일 분자 서열결정, 이온 토렌트(Torrent) 및 이온 반도체 기초 서열결정(예를 들면, 라이

프 테크놀로지스(Life Technologies)에 의해 개발됨), 와일드파이어, 5500, 5500x1 W 및/또는 5500x1 W 유전 분석기 기초 기술(예를 들면, 라이프 테크놀로지스에 의해 개발되고 시판됨, 미국 특허출원 공보 제20130012399호); 폴로니(Polony) 서열결정, 피로서열결정(Pyrosequencing), 대량 병렬 시그너처 서열결정(MPSS), RNA 중합효소(RNAP) 서열결정, 레이저겐(LaserGen) 시스템 및 방법, 나노공극 기초 플랫폼, 화학적 민감성 필드-효과 트랜지스터(CHEMFET) 어레이, 전자 현미경관찰 기초 서열결정(예를 들면, 제트에스 제네틱스, 할사이온 몰레큘라(ZS Genetics, Halcyon Molecular)에 의해 개발됨), 및 나노볼(nanoball) 서열결정을 포함한다.

[0124] 일부 실시양태에서, 염색체 특이적 서열결정이 수행된다. 일부 실시양태에서, DANSR(선택된 영역의 디지털 분석)을 이용하는 염색체 특이적 서열결정이 수행된다. 선택된 영역의 디지털 분석은 개재 '가교' 올리고뉴클레오티드를 통해 2개의 좌위 특이적 올리고뉴클레오티드들을 cfDNA 의존적으로 연쇄적으로 연결하여 PCR 주형을 형성함으로써 수백 개의 좌위들의 동시적 정량을 가능하게 한다. 일부 실시양태에서, 염색체 특이적 서열결정은 염색체 특이적 서열이 풍부한 라이브러리를 생성함으로써 수행된다. 일부 실시양태에서, 서열 리드는 선택된 염색체 세트에 대해서만 수득된다. 일부 실시양태에서, 서열 리드는 21번 염색체, 18번 염색체 및 13번 염색체에 대해서만 수득된다.

[0125] 맵핑 리드

[0126] 서열 리드가 맵핑될 수 있고, 특정된 핵산 영역(예를 들면, 염색체, 또는 이의 부분 또는 분절)로 맵핑되는 리드의 수는 카운트로서 지칭된다. 임의의 적합한 맵핑 방법(예를 들면, 프로세스, 알고리즘, 프로그램, 소프트웨어, 모듈 등 또는 이들의 조합)이 이용될 수 있다. 맵핑 프로세스의 일부 양태는 이하에 기재되어 있다.

[0127] 뉴클레오티드 서열 리드(즉, 물리적 게놈 위치가 공지되어 있지 않은 단편으로부터의 서열 정보)의 맵핑은 다수의 방식으로 수행될 수 있고, 종종 수득된 서열 리드를 기준 게놈 내의 일치 서열과 정렬하는 단계를 포함한다. 이러한 정렬에서, 서열 리드는 일반적으로 기준 서열과 정렬되고, 정렬되는 서열 리드는 "맵핑된" 것, "맵핑된 서열 리드" 또는 "맵핑된 리드"로서 표기된다. 일부 실시양태에서, 맵핑된 서열 리드는 "히트(hit)" 또는 "카운트"로서 지칭된다. 일부 실시양태에서, 맵핑된 서열 리드는 다양한 파라미터에 따라 함께 분류되고, 하기에 더 상세히 논의되어 있는 특정 부분으로 배정된다.

[0128] 본원에서 사용된 바와 같이, 용어 "정렬된", "정렬" 또는 "정렬하는"은 일치(예를 들면, 100% 동일성) 또는 부분적 일치로서 확인될 수 있는 2개 이상의 핵산 서열을 지칭한다. 정렬은 수동으로 수행될 수 있거나 컴퓨터(예를 들면, 소프트웨어, 프로그램, 모듈 또는 알고리즘)에 의해 수행될 수 있고, 이러한 컴퓨터 프로그램의 비한정적 예는 일루미나 게놈 분석 파이프라인의 부분으로서 배포된 뉴클레오티드 데이터의 효율적 국소 정렬(ELAND) 컴퓨터 프로그램을 포함하다. 서열 리드의 정렬은 100% 서열 일치일 수 있다. 일부 경우, 정렬은 100% 미만의 서열 일치(즉, 완벽하지 않은 일치, 부분적 일치, 부분적 정렬)이다. 일부 실시양태에서, 정렬은 약 99%, 98%, 97%, 96%, 95%, 94%, 93%, 92%, 91%, 90%, 89%, 88%, 87%, 86%, 85%, 84%, 83%, 82%, 81%, 80%, 79%, 78%, 77%, 76% 또는 75% 일치이다. 일부 실시양태에서, 정렬은 불일치를 포함한다. 일부 실시양태에서, 정렬은 1개, 2개, 3개, 4개 또는 5개의 불일치를 포함한다. 어느 한 가닥을 사용하여 2개 이상의 서열을 정렬할 수 있다. 일부 실시양태에서, 핵산 서열은 또 다른 핵산 서열의 역 상보체와 정렬된다.

[0129] 컴퓨터를 이용한 다양한 방법들이 각각의 서열 리드를 부분들로 맵핑하는 데에 이용될 수 있다. 서열을 정렬하는 데에 사용될 수 있는 컴퓨터 알고리즘의 비한정적 예는 BLAST, BLITZ, FASTA, BOWTIE 1, BOWTIE 2, ELAND, MAQ, PROBEMATCH, SOAP 또는 SEQMAP, 또는 이들의 변경물 또는 이들의 조합물을 포함하나 이들로 한정되지 않는다. 일부 실시양태에서, 서열 리드는 기준 게놈 내의 서열과 정렬될 수 있다. 일부 실시양태에서, 서열 리드는 예를 들면, 진뱅크(GenBank), dbEST, dbSTS, EMBL(유럽 분자생물학 실험실(European Molecular Biology Laboratory)) 및 DDBJ(일본의 DNA 데이터뱅크(DNA Databank of Japan))를 포함하는, 당분야에서 공지된 핵산 데이터베이스에서 발견될 수 있고/있거나 이러한 데이터베이스 내의 서열과 정렬될 수 있다. BLAST 또는 유사한 수단이 확인된 서열을 서열 데이터베이스에서 검색하는 데에 사용될 수 있다. 그 다음, 예를 들면, 검색 히트가 확인된 서열을 적절한 부분(이하에 기재됨)으로 분류하는 데에 사용될 수 있다.

[0130] 일부 실시양태에서, 맵핑된 서열 리드 및/또는 맵핑된 서열 리드와 관련된 정보가 적합한 컴퓨터 판독가능한 포맷으로 비-일시적 컴퓨터 판독가능한 저장 매체 상에 저장되고/되거나 이러한 저장 매체로부터 접근된다. "컴퓨터 판독가능한 포맷"은 종종 일반적으로 본원에서 포맷으로서 지칭된다. 일부 실시양태에서, 맵핑된 서열 리드는 적합한 이진(binary) 포맷, 텍스트 포맷 등 또는 이들의 조합물로 저장되고/되거나 접근된다. 이진 포맷은 종종 BAM 포맷이다. 텍스트 포맷은 종종 서열 정렬/맵(SAM) 포맷이다. 이진 및/또는 텍스트 포맷의 비한정적 예는 BAM, SAM, SRF, FASTQ, Gzip 등 또는 이들의 조합물을 포함한다. 일부 실시양태에서, 맵핑된 서열 리드는 전

통적인 포맷(예를 들면, SAM 포맷 또는 BAM 포맷)보다 더 작은 저장 공간(예를 들면, 더 작은 바이트)을 요구하는 포맷으로 저장되고/되거나 전환된다. 일부 실시양태에서, 제1 포맷의 맵핑된 서열 리드는 제1 포맷보다 더 작은 저장 공간을 요구하는 제2 포맷으로 압축된다. 본원에서 사용된 용어 "압축된"은 데이터 압축, 소스 암호화 및/또는 비트-속도 감소의 프로세스를 지칭하고, 이때 컴퓨터 판독가능한 데이터 파일이 크기 면에서 감소된다. 일부 실시양태에서, 맵핑된 서열 리드는 SAM 포맷으로부터 이진 포맷으로 압축된다. 일부 데이터는 종종 파일이 압축된 후 손실된다. 종종 데이터는 압축 프로세스에서 전혀 손실되지 않는다. 일부 파일 압축 실시양태에서, 일부 데이터는 맵핑된 서열 리드에 대한 정보를 포함하는 또 다른 데이터 파일에 대한 인덱스 및/또는 기준물로 대체된다. 일부 실시양태에서, 맵핑된 서열 리드는 리드 카운트, (예를 들면, 리드가 맵핑되는 염색체를 식별시켜주는) 염색체 식별자 및 (예를 들면, 리드가 맵핑되는 염색체 상의 위치를 식별시켜주는) 염색체 위치 식별자를 포함하거나 이들로 구성된 이진 포맷으로 저장된다. 일부 실시양태에서, 이진 포맷은 20 바이트 어레이, 16 바이트 어레이, 8 바이트 어레이, 4 바이트 어레이 또는 2 바이트 어레이를 포함한다. 일부 실시양태에서, 맵핑된 리드 정보는 10 바이트 포맷, 9 바이트 포맷, 8 바이트 포맷, 7 바이트 포맷, 6 바이트 포맷, 5 바이트 포맷, 4 바이트 포맷, 3 바이트 포맷 또는 2 바이트 포맷의 어레이로 저장된다. 종종 맵핑된 리드 데이터는 5 바이트 포맷을 포함하는 4 바이트 어레이로 저장된다. 일부 실시양태에서, 이진 포맷은 1 바이트 염색체 서수 및 4 바이트 염색체 위치를 포함하는 5 바이트 포맷을 포함한다. 일부 실시양태에서, 맵핑된 리드는 서열 정렬/맵(SAM) 포맷보다 약 100배, 약 90배, 약 80배, 약 70배, 약 60배, 약 55배, 약 50배, 약 45배, 약 40배 또는 약 30배 더 작은 압축된 이진 포맷으로 저장된다. 일부 실시양태에서, 맵핑된 리드는 GZip 포맷보다 약 2배 내지 약 50배 더 작은(예를 들면, 약 30배, 25배, 20배, 19배, 18배, 17배, 16배, 15배, 14배, 13배, 12배, 11배, 10배, 9배, 8배, 7배, 6배 또는 약 5배 더 작은) 압축 이진 포맷으로 저장된다.

[0131] 일부 실시양태에서, 시스템은 압축 모듈(예를 들면, 도 42a의 4)을 포함한다. 일부 실시양태에서, 비-일시적 컴퓨터 판독가능한 저장 매체 상에서 컴퓨터 판독가능한 포맷으로 저장된 맵핑된 서열 리드 정보는 압축 모듈에 의해 압축된다. 압축 모듈은 종종 맵핑된 서열 리드를 적합한 포맷으로 전환시키고 맵핑된 서열 리드를 적합한 포맷으로부터 전환시킨다. 일부 실시양태에서, 압축 모듈은 맵핑된 서열 리드를 제1 포맷으로 수용할 수 있고(예를 들면, 도 42a의 1), 이를 압축된 포맷(예를 들면, 이진 포맷, 5)으로 전환시킬 수 있고, 압축된 리드를 또 다른 모듈(예를 들면, 편향 밀도 모듈, 6)에게 전달한다. 압축 모듈은 종종 이진 포맷(5)(예를 들면, BReads 포맷)으로 서열 리드를 제공한다. 압축 모듈의 비한정적 예는 GZIP, BGZF 및 BAM 등 또는 이들의 변경물을 포함한다.

[0132] 하기 예는 자바(java)를 사용하여 정수를 4 바이트 어레이로 전환시키는 일례를 제공한다:

```
public static final byte[ ]
convertToByteArray(int value)
{
return new byte[ ] {
(byte)(value >>> 24),
(byte)(value >>> 16),
(byte)(value >>> 8),
(byte)value};
}
```

[0133]

[0134] 일부 실시양태에서, 리드는 기준 게놈 내의 부분으로 유일하게 또는 유일하지 않게 맵핑될 수 있다. 리드는 기준 게놈 내의 단일 서열과 정렬되는 경우 "유일하게 맵핑된" 것으로서 간주된다. 리드는 기준 게놈 내의 2개 이상의 서열들과 정렬되는 경우 "유일하지 않게 맵핑된" 것으로서 간주된다. 일부 실시양태에서, 유일하지 않게 맵핑된 리드는 추가 분석(예를 들면, 정량)으로부터 제거된다. 일부 실시양태에서, 기준 게놈과 맵핑되는 개별 샘플로부터의 리드 사이에 존재할 수 있는 단일 뉴클레오티드 다형성을 설명하기 위해 미미한 정도(0개 또는 1개)의 불일치가 허용될 수 있다. 일부 실시양태에서, 기준 서열로 맵핑되는 리드에 대해 어느 정도의 불일치도 허용되지 않는다.

[0135] 본원에서 사용된 바와 같이, 용어 "기준 게놈"은 부분적 게놈이든 아니면 완전한 게놈이든 관계없이 대상체로부터의 확인된 서열의 기준물로서 사용될 수 있는 임의의 유기체 또는 바이러스의 임의의 특정 공지된, 서열결정

된 또는 특징규명된 게놈을 지칭할 수 있다. 예를 들면, 인간 대상체뿐만 아니라 많은 다른 유기체를 위해서도 사용되는 기준 게놈은 월드 와이드 웹 URL ncbi.nlm.nih.gov의 국립 생물공학 정보 센터에서 발견될 수 있다. "게놈"은 핵산 서열로 발현된 유기체 또는 바이러스의 완전한 유전 정보를 지칭한다. 본원에서 사용된 바와 같이, 기준 서열 또는 기준 게놈은 종종 한 개체 또는 다수의 개체들로부터의 조립된 또는 부분적으로 조립된 게놈 서열이다. 일부 실시양태에서, 기준 게놈은 1명 이상의 인간 개체로부터의 조립된 또는 부분적으로 조립된 게놈 서열이다. 일부 실시양태에서, 기준 게놈은 염색체로 배정된 서열을 포함한다.

[0136] 일부 실시양태에서, 샘플 핵산이 임신 여성으로부터 유래된 경우, 기준 서열은 종종 태아, 태아의 모 또는 태아의 부로부터 유래되지 않고, 본원에서 "외부 기준물"로서 지칭된다. 일부 실시양태에서, 모체 기준물이 제조되고 사용될 수 있다. 외부 기준물을 기초로 임신 여성으로부터의 기준물이 제조된 경우("모체 기준 서열"), 태아 DNA를 실질적으로 함유하지 않는 임신 여성의 DNA로부터의 리드가 종종 외부 기준 서열로 맵핑되고 조립된다. 일부 실시양태에서, 외부 기준물은 임신 여성과 실질적으로 동일한 인종을 갖는 개체의 DNA로부터 유래된다. 모체 기준 서열은 모체 게놈 DNA를 완전히 커버하지 않을 수 있고(예를 들면, 이 기준 서열은 모체 게놈 DNA의 약 50%, 60%, 70%, 80% 또는 90% 이상을 커버할 수 있고), 모체 기준물은 모체 게놈 DNA 서열과 완벽히 일치하지 않을 수 있다(예를 들면, 모체 기준 서열은 다수의 불일치를 포함할 수 있다).

[0137] 일부 실시양태에서, 게놈 영역(예를 들면, 부분, 게놈 부분, 부분)에 대한 맵핑가능성이 평가된다. 맵핑가능성은 뉴클레오티드 서열 리드를, 전형적으로 특정된 수의 불일치(예를 들면, 0개, 1개 또는 2개 이상의 불일치를 포함함)를 갖는 기준 게놈의 부분과 명확히 정렬하는 능력이다. 주어진 게놈 영역에 대해, 미리 설정된 리드 길이의 슬라이딩 윈도우(window) 방법을 이용하고 수득된 리드-수준 맵핑가능성 값의 평균을 산출함으로써 예상된 맵핑가능성을 추정할 수 있다. 유일한 뉴클레오티드 서열의 스트레치를 포함하는 게놈 영역은 종종 높은 맵핑가능성 값을 갖는다.

[0138] 부분

[0139] 일부 실시양태에서, 맵핑된 서열 리드(즉, 서열 태그)는 다양한 파라미터에 따라 함께 분류되고 특정 부분(예를 들면, 기준 게놈의 부분)으로 배정된다. 종종, 개별 맵핑된 서열 리드가 샘플에 존재하는 부분(예를 들면, 부분의 존재, 부재 또는 양)을 확인하는 데에 사용될 수 있다. 일부 실시양태에서, 부분의 양은 샘플 중의 보다 더 큰 서열(예를 들면, 염색체)의 양을 표시한다. 용어 "부분"은 본원에서 "게놈 구획", "빈(bin)", "영역", "구획", "기준 게놈의 부분", "염색체의 부분" 또는 "게놈 부분"으로서도 지칭될 수 있다. 일부 실시양태에서, 부분은 전체 염색체, 염색체의 분절, 기준 게놈의 분절, 다수의 염색체에 걸쳐있는 분절, 다수의 염색체 분절 및/또는 이들의 조합물이다. 일부 실시양태에서, 부분은 특정 파라미터를 기초로 미리 정해진다. 일부 실시양태에서, 부분은 게놈의 분할을 기초로 임의로 정해진다(예를 들면, 크기, GC 함량, 연속 영역, 임의로 정해진 크기의 연속 영역 등에 의해 분할된다).

[0140] 일부 실시양태에서, 부분은 예를 들면, 서열의 길이 또는 특정 특징 또는 특징들을 포함하는 하나 이상의 파라미터를 기초로 기술된다. 부분은 당분야에서 공지되어 있거나 본원에 기재되어 있는 임의의 적합한 기준을 사용함으로써 고려사항으로부터 선택될 수 있고/있거나, 필터링될 수 있고/있거나 제거될 수 있다. 일부 실시양태에서, 부분은 게놈 서열의 특정 길이에 기초한다. 일부 실시양태에서, 방법은 복수의 부분들에 대한 다수의 맵핑된 서열 리드들의 분석을 포함할 수 있다. 부분은 거의 동일한 길이를 가질 수 있거나, 부분은 상이한 길이를 가질 수 있다. 일부 실시양태에서, 부분은 거의 동등한 길이를 갖는다. 일부 실시양태에서, 상이한 길이의 부분들은 조절되거나 가중된다. 일부 실시양태에서, 부분은 약 10 킬로염기(kb) 내지 약 100 kb, 약 20 kb 내지 약 80 kb, 약 30 kb 내지 약 70 kb, 약 40 kb 내지 약 60 kb, 또는 종종 약 50 kb이다. 일부 실시양태에서, 부분은 약 10 kb 내지 약 20 kb이다. 부분은 서열의 연속물로 한정되지 않는다. 따라서, 부분은 연속 및/또는 비-연속 서열로 구성될 수 있다. 부분은 단일 염색체로 한정되지 않는다. 일부 실시양태에서, 부분은 한 염색체의 전부 또는 일부, 또는 2개 이상의 염색체의 전부 또는 일부를 포함한다. 일부 실시양태에서, 부분은 1개 또는 2개 이상의 전체 염색체에 걸쳐 있을 수 있다. 추가로, 부분은 다수의 염색체들의 연결된 또는 연결되지 않은 영역들에 걸쳐 있을 수 있다.

[0141] 일부 실시양태에서, 부분은 관심있는 염색체, 예를 들면, 유전적 변이(예를 들면, 13번 염색체, 18번 염색체 및/또는 21번 염색체 또는 성염색체의 이배수성)가 평가되는 염색체 내의 특정 염색체 분절일 수 있다. 부분은 병원체 게놈(예를 들면, 세균, 진균 또는 바이러스) 또는 이의 단편일 수도 있다. 부분은 유전자, 유전자 단편, 조절 서열, 인트론, 엑손 등일 수 있다.

[0142] 일부 실시양태에서, 게놈(예를 들면, 인간 게놈)은 특정 영역의 정보 내용을 기초로 부분으로 분할된다. 일부

실시양태에서, 게놈의 분할은 게놈에 걸쳐 유사한 영역들(예를 들면, 동일한 또는 상동성 영역들 또는 서열들)을 제거할 수 있고 유일한 영역들만을 남길 수 있다. 분할 동안 제거된 영역은 단일 염색체 내에 존재할 수 있거나 다수의 염색체들에 걸쳐 있을 수 있다. 일부 실시양태에서, 분할된 게놈은 보다 더 빠른 정렬을 위해 줄어들고 최적화되어, 종종 유일하게 식별가능한 서열에 초점을 맞출 수 있게 한다.

[0143] 일부 실시양태에서, 분할은 유사한 영역들을 하향 가중할 수 있다. 부분을 하향 가중하는 프로세스는 하기 더 상세히 논의되어 있다.

[0144] 일부 실시양태에서, 염색체를 능가하는 영역으로의 게놈의 분할은 분류 면에서 생성된 정보 획득에 기초할 수 있다. 예를 들면, 확인된 정상 대상체 군과 비정상 대상체 군(예를 들면, 각각 정배수체 대상체와 삼염색체성 대상체)을 식별하기 위해 특정 게놈 위치의 유의성을 측정하는 p-값 프로파일을 사용하여 정보 내용을 정량할 수 있다. 일부 실시양태에서, 염색체를 능가하는 영역으로의 게놈의 분할은 임의의 다른 기준, 예컨대, 태그를 정렬하는 동안의 속도/편의성, GC 함량(예를 들면, 높은 또는 낮은 GC 함량), GC 함량의 균일성, 서열 함량의 다른 척도(예를 들면, 개별 뉴클레오티드의 분율, 피리미딘 또는 푸린의 분율, 천연 대 비천연 핵산의 분율, 메틸화된 뉴클레오티드의 분율 및 CpG 함량), 메틸화 상태, 이중체 용융 온도, 서열결정 또는 PCR로의 처리가능성, 기준 게놈의 개별 부분으로 배정된 불확실성 값, 및/또는 특정 특징에 대한 표적화된 검색에 기초할 수 있다.

[0145] 염색체의 "분절"은 일반적으로 염색체의 일부이고, 전형적으로 부분과 상이한 염색체의 일부이다. 염색체의 분절은 종종 부분과 상이한 염색체의 영역에 존재하고, 종종 부분과 폴리뉴클레오티드를 공유하지 않고, 종종 부분에 존재하는 폴리뉴클레오티드를 포함한다. 염색체의 분절은 종종 부분보다 더 많은 수의 뉴클레오티드를 함유하고(예를 들면, 분절은 종종 부분을 포함하고), 종종 염색체의 분절은 부분보다 더 적은 수의 뉴클레오티드를 함유한다(예를 들면, 분절은 종종 부분 내에 존재한다).

[0146] 카운트

[0147] 일부 실시양태에서, 하나 이상의 부분들(예를 들면, 기준 게놈의 부분)에 맵핑된 리드의 수를 측정하기 위해 선택된 특징 또는 변수를 기초로 맵핑되거나 분할된 서열 리드를 정량할 수 있다. 일부 실시양태에서, 부분들에 맵핑된 서열 리드의 양은 카운트(예를 들면, 카운트 값)로서 지칭된다. 종종 카운트는 부분과 관련되어 있다. 일부 실시양태에서, 2개 이상의 부분(예를 들면, 부분 세트)에 대한 카운트는 수학적으로 조작된다(예를 들면, 평균 산출, 덧셈, 정규화 등 또는 이들의 조합). 일부 실시양태에서, 카운트는 부분들에 맵핑된(즉, 부분과 관련된) 서열 리드의 일부 또는 전부로부터 측정된다. 일부 실시양태에서, 카운트는 미리 정해진 맵핑된 서열 리드 서브세트로부터 측정된다. 미리 정해진 맵핑된 서열 리드 서브세트는 임의의 적합한 특징 또는 변수를 사용함으로써 정해질 수 있거나 선택될 수 있다. 일부 실시양태에서, 미리 정해진 맵핑된 서열 리드 서브세트는 1개 내지 n개의 서열 리드를 포함할 수 있고, 이때 n은 검사 대상체 또는 기준 대상체 샘플로부터 생성된 모든 서열 리드들의 합계와 동등한 수를 나타낸다.

[0148] 일부 실시양태에서, 카운트는 당분야에서 공지된 적합한 방법, 연산 또는 수학적 프로세스에 의해 프로세싱되거나 조작된 서열 리드로부터 유도된다. 카운트(예를 들면, 카운트 값)는 적합한 방법, 연산 또는 수학적 프로세스에 의해 측정될 수 있다. 일부 실시양태에서, 카운트는 부분과 관련된 서열 리드로부터 유도되고, 이때 서열 리드의 일부 또는 전부가 가중되거나, 제거되거나, 필터링되거나, 정규화되거나, 조절되거나, 평균으로서 산출되거나, 평균으로서 유도되거나, 더해지거나, 빼지거나 이들의 조합에 의해 프로세싱된다. 일부 실시양태에서, 카운트는 미처리 서열 리드 및/또는 필터링된 서열 리드로부터 유도된다. 일부 실시양태에서, 카운트 값은 수학적 프로세스에 의해 측정된다. 일부 실시양태에서, 카운트 값은 부분들에 맵핑된 서열 리드의 평균치, 평균 또는 합계이다. 종종, 카운트는 카운트의 평균 수이다. 일부 실시양태에서, 카운트는 불확실성 값과 관련되어 있다.

[0149] 일부 실시양태에서, 카운트는 조작될 수 있거나 변환될 수 있다(예를 들면, 정규화될 수 있거나, 조합될 수 있거나, 더해질 수 있거나, 필터링될 수 있거나, 선택될 수 있거나, 평균으로서 산출될 수 있거나, 평균으로서 유도될 수 있거나, 이들의 조합에 의해 프로세싱될 수 있다). 일부 실시양태에서, 카운트는 정규화된 카운트를 생성하도록 변환될 수 있다. 카운트는 당분야에서 공지되어 있고/있거나 본원에 기재되어 있는 방법(예를 들면, 부분별 정규화, GC 함량에 의한 정규화, 선형 및 비선형 최소 자승 회귀, GC LOESS, LOWESS, PERUN, ChAI, RM, GCRM, cQn 및/또는 이들의 조합)에 의해 프로세싱(예를 들면, 정규화)될 수 있다.

[0150] 카운트(예를 들면, 미처리, 필터링된 및/또는 정규화된 카운트)는 프로세싱될 수 있고 하나 이상의 수준으로 정

규화될 수 있다. 수준 및 프로파일은 이하에 더 상세히 기재되어 있다. 일부 실시양태에서, 카운트는 프로세싱될 수 있고/있거나 기준 수준으로 정규화될 수 있다. 기준 수준은 본원에서 나중에 다루어진다. 수준에 따라 프로세싱된 카운트(예를 들면, 프로세싱된 카운트)는 불확실성 값(예를 들면, 계산된 분산, 오차, 표준 편차, Z-점수, p-값, 평균 절대 편차 등)과 관련될 수 있다. 일부 실시양태에서, 불확실성 값은 수준 초과 및 미만의 범위를 한정한다. 편차에 대한 값은 불확실성 값 대신에 사용될 수 있고, 편차의 척도의 비한정적 예는 표준 편차, 평균 절대 편차, 중간 절대 편차, 표준 점수(예를 들면, Z-점수, 표준 점수, 표준화된 변수) 등을 포함한다.

[0151] 카운트는 종종 태아를 가진 임신 여성으로부터의 핵산 샘플로부터 수득된다. 하나 이상의 부분들에 맵핑된 핵산 서열 리드의 카운트는 태아 및 태아의 모체(예를 들면, 임신 여성 대상체) 둘 다를 표시하는 카운트이다. 일부 실시양태에서, 부분들에 맵핑된 카운트 중 일부는 태아 게놈으로부터의 카운트이고, 동일한 부분들에 맵핑된 카운트 중 일부는 모체 게놈으로부터의 카운트이다.

[0152] 데이터 프로세싱 및 정규화

[0153] 카운팅되어 있는 맵핑된 서열 리드는 본원에서 미처리 데이터로서 지칭되는데, 이는 상기 데이터가 조작되지 않은 카운트(예를 들면, 미처리 카운트)를 나타내기 때문이다. 일부 실시양태에서, 데이터 세트에서 서열 리드 데이터는 결과 제공을 용이하게 하기 위해 더 프로세싱될 수 있고(예를 들면, 수학적으로 및/또는 통계학적으로 조작될 수 있고)/있거나 디스플레이될 수 있다. 일부 실시양태에서, 보다 더 큰 데이터 세트를 포함하는 데이터 세트는 추가 분석을 용이하게 하기 위한 사전프로세싱으로부터 이익을 얻을 수 있다. 데이터 세트의 사전프로세싱은 종종 중복되고/되거나 비-정보제공 부분 또는 기준 게놈의 부분(예를 들면, 비-정보제공 데이터를 갖는 기준 게놈의 부분, 중복 맵핑된 리드, 0의 중간치 카운트를 갖는 부분, 과다표시된 또는 과소표시된 서열)의 제거를 포함한다. 이론에 의해 한정되지 않지만, 데이터 프로세싱 및/또는 사전프로세싱은 (i) 무의미한(noisy) 데이터를 제거할 수 있고/있거나, (ii) 비-정보제공 데이터를 제거할 수 있고/있거나, (iii) 중복 데이터를 제거할 수 있고/있거나, (iv) 보다 더 큰 데이터 세트의 복잡성을 감소시킬 수 있고/있거나, (v) 데이터를 한 형태로부터 하나 이상의 다른 형태로 변환하는 것을 용이하게 할 수 있다. 데이터 또는 데이터 세트와 관련하여 사용될 때 용어 "사전프로세싱" 및 "프로세싱"은 본원에서 "프로세싱"으로서 총칭된다. 일부 실시양태에서, 프로세싱은 데이터가 더 잘 추가 분석될 수 있게 만들 수 있고 결과를 생성할 수 있다. 일부 실시양태에서, 하나 이상의 또는 모든 프로세싱 방법(예를 들면, 정규화 방법, 부분 필터링, 맵핑, 검증 등 또는 이들의 조합)이 프로세서에 의해, 마이크로프로세서에 의해, 컴퓨터에 의해, 메모리와 함께 및/또는 마이크로프로세서 제어 장치에 의해 수행된다.

[0154] 본원에서 사용된 용어 "무의미한 데이터"는 (a) 분석되거나 작도될 때 데이터 점들 사이에 유의한 분산을 갖는 데이터, (b) 유의한 표준 편차(예를 들면, 3 초과의 표준 편차)를 갖는 데이터, (c) 평균의 유의한 표준 오차를 갖는 데이터 등 및 이들의 조합물을 지칭한다. 무의미한 데이터는 종종 출발 물질(예를 들면, 핵산 샘플)의 양 및 질로 인해 생성되고, 종종 서열 리드를 생성하는 데에 사용된 DNA를 제조하거나 복제하는 프로세스의 일부로서 생성된다. 일부 실시양태에서, 무의미한 데이터는 PCR 기초 방법을 이용함으로써 제조되었을 때 과다표시되는 일부 서열들로부터 생성된다. 본원에 기재된 방법은 무의미한 데이터의 기여를 감소시킬 수 있거나 제거할 수 있으므로, 제공된 결과에 대한 무의미한 데이터의 영향을 감소시킬 수 있다.

[0155] 본원에서 사용된 용어 "비-정보제공 데이터", "기준 게놈의 비-정보제공 부분" 및 "비-정보제공 부분"은 예정된 역치 값과 유의하게 상이하거나 예정된 컷오프 값 범위를 벗어나는 수치 값을 갖는 부분 또는 이로부터 유도된 데이터를 지칭한다. 본원에서 용어 "역치" 및 "역치 값"은 자격을 갖춘 데이터 세트를 사용함으로써 계산되고 유전적 변이(예를 들면, 카피 수 변이, 이배수성, 미세중복, 미세결실, 염색체 이상 등)의 진단의 한계로서 작용하는 임의의 수를 지칭한다. 일부 실시양태에서, 본원에 기재된 방법에 의해 수득된 결과는 역치를 초과하고, 대상체는 유전적 변이(예를 들면, 삼염색체성 21)를 갖는 것으로서 진단된다. 일부 실시양태에서, 역치 값 또는 값 범위는 종종 (예를 들면, 기준물 및/또는 대상체로부터의) 서열 리드 데이터를 수학적으로 및/또는 통계학적으로 조작함으로써 계산되고, 일부 실시양태에서 역치 값 또는 값 범위를 생성하도록 조작된 서열 리드 데이터는 (예를 들면, 기준물 및/또는 대상체로부터의) 서열 리드 데이터이다. 일부 실시양태에서, 불확실성 값이 측정된다. 불확실성 값은 일반적으로 분산 또는 오차의 척도이고, 분산 또는 오차의 임의의 적합한 척도일 수 있다. 일부 실시양태에서, 불확실성 값은 표준 편차, 표준 오차, 계산된 분산, p-값 또는 평균 절대 편차(MAD)이다. 일부 실시양태에서, 불확실성 값은 실시예 4의 식에 따라 계산될 수 있다.

[0156] 임의의 적합한 절차가 본원에 기재된 데이터 세트의 프로세싱을 위해 사용될 수 있다. 데이터 세트의 프로세싱

을 위해 사용되기에 적합한 절차의 비한정적 예는 필터링, 정규화, 가중, 피크 높이의 모니터링, 피크 면적의 모니터링, 피크 에지의 모니터링, 면적 비의 측정, 데이터의 수학적 프로세싱, 데이터의 통계학적 프로세싱, 통계학적 알고리즘의 적용, 고정된 변수를 사용한 분석, 최적화된 변수를 사용한 분석, 추가 프로세싱을 위한 패턴 또는 경향을 확인하기 위한 데이터의 작도 등 및 이들의 조합을 포함한다. 일부 실시양태에서, 데이터 세트는 다양한 특징들(예를 들면, GC 함량, 중복 맵핑된 리드, 센트로미어 영역, 텔로미어 영역 등 및 이들의 조합) 및/또는 변수(예를 들면, 태아 성별, 모체 연령, 모체 배수성, 태아 핵산의 % 기여 등 또는 이들의 조합)을 기초로 프로세싱된다. 일부 실시양태에서, 본원에 기재된 데이터 세트의 프로세싱은 크고/크거나 복잡한 데이터 세트의 복잡성 및/또는 차원수를 감소시킬 수 있다. 복잡한 데이터 세트의 비한정적 예는 상이한 연령 및 인종 배경을 갖는 하나 이상의 검사 대상체 및 복수의 기준 대상체로부터 생성된 서열 리드 데이터를 포함한다. 일부 실시양태에서, 데이터 세트는 각각의 검사 및/또는 기준 대상체에 대한 수천 개 내지 수백만 개의 서열 리드를 포함할 수 있다.

[0157] 일부 실시양태에서, 데이터 프로세싱은 임의의 수의 단계로 수행될 수 있다. 예를 들면, 일부 실시양태에서 데이터는 단일 프로세싱 절차만을 이용함으로써 프로세싱될 수 있고, 일부 실시양태에서 데이터는 1개 이상, 5개 이상, 10개 이상 또는 20개 이상의 프로세싱 단계(예를 들면, 1개 이상의 프로세싱 단계, 2개 이상의 프로세싱 단계, 3개 이상의 프로세싱 단계, 4개 이상의 프로세싱 단계, 5개 이상의 프로세싱 단계, 6개 이상의 프로세싱 단계, 7개 이상의 프로세싱 단계, 8개 이상의 프로세싱 단계, 9개 이상의 프로세싱 단계, 10개 이상의 프로세싱 단계, 11개 이상의 프로세싱 단계, 12개 이상의 프로세싱 단계, 13개 이상의 프로세싱 단계, 14개 이상의 프로세싱 단계, 15개 이상의 프로세싱 단계, 16개 이상의 프로세싱 단계, 17개 이상의 프로세싱 단계, 18개 이상의 프로세싱 단계, 19개 이상의 프로세싱 단계, 또는 20개 이상의 프로세싱 단계)를 이용함으로써 프로세싱될 수 있다. 일부 실시양태에서, 프로세싱 단계는 2회 이상 반복된 동일한 단계(예를 들면, 2회 이상의 필터링, 2회 이상의 정규화)일 수 있고, 일부 실시양태에서 프로세싱 단계는 동시에 또는 순차적으로 수행된 2회 이상의 상이한 프로세싱 단계(예를 들면, 필터링, 정규화; 정규화, 피크 높이 및 에지의 모니터링; 필터링, 정규화, 기준물로의 정규화, p-값을 측정하기 위한 통계학적 조작 등)일 수 있다. 일부 실시양태에서, 임의의 적합한 수 및/또는 조합의 동일한 또는 상이한 프로세싱 단계들이 결과의 제공을 용이하게 하기 위해 서열 리드 데이터를 프로세싱하는 데에 사용될 수 있다. 일부 실시양태에서, 본원에 기재된 기준에 의한 데이터 세트의 프로세싱은 데이터 세트의 복잡성 및/또는 차원수를 감소시킬 수 있다.

[0158] 일부 실시양태에서, 하나 이상의 프로세싱 단계는 하나 이상의 필터링 단계를 포함할 수 있다. 본원에서 사용된 "필터링"은 고려사항으로부터 부분 또는 기준 게놈의 부분을 제거하는 것을 지칭한다. 기준 게놈의 부분은 중복 데이터(예를 들면, 중복 또는 중첩 맵핑된 리드), 비-정보제공 데이터(예를 들면, 0의 중간치 카운트를 갖는 기준 게놈의 부분), 과다표시된 또는 과소표시된 서열을 갖는 기준 게놈의 부분, 무의미한 데이터 또는 이들의 조합물을 포함하나 이들로 한정되지 않는 임의의 적합한 기준을 기초로 제거를 위해 선택될 수 있다. 필터링 프로세스는 종종 고려사항으로부터 기준 게놈의 하나 이상의 부분을 제거하는 단계 및 기준 게놈, 염색체 또는 염색체들, 또는 고려되는 게놈의 부분에 대한 카운팅된 또는 합산된 카운트로부터 제거를 위해 선택된 기준 게놈의 하나 이상의 부분에서의 카운트를 빼는 단계를 포함한다. 일부 실시양태에서, 기준 게놈의 부분은 연속적으로(예를 들면, 각각의 개별 부분의 제거 효과의 평가를 가능하게 하기 위해 한번에 하나씩) 제거될 수 있고, 일부 실시양태에서 제거를 위해 표시된 기준 게놈의 모든 부분들이 동시에 제거될 수 있다. 일부 실시양태에서, 일정 수준 초과 또는 미만의 분산을 특징으로 하는 기준 게놈의 부분이 제거되는데, 이것은 종종 본원에서 기준 게놈의 "무의미한" 부분의 필터링으로서 지칭된다. 일부 실시양태에서, 필터링 프로세스는 부분, 염색체 또는 염색체의 분절의 평균 프로파일 수준으로부터 프로파일 분산의 예정된 배수만큼 벗어나는 데이터 점을 데이터 세트로부터 수득하는 단계를 포함하고, 일부 실시양태에서 필터링 프로세스는 부분, 염색체 또는 염색체의 분절의 평균 프로파일 수준으로부터 프로파일 분산의 예정된 배수만큼 벗어나지 않는 데이터 점을 데이터 세트로부터 제거하는 단계를 포함한다. 일부 실시양태에서, 필터링 프로세스는 유전적 변이의 존재 또는 부재에 대해 분석되는 기준 게놈의 후보 부분의 수를 감소시키는 데에 사용된다. 유전적 변이(예를 들면, 미세결실, 미세중복)의 존재 또는 부재에 대해 분석되는 기준 게놈의 후보 부분의 수의 감소는 종종 데이터 세트의 복잡성 및/또는 차원수를 감소시키고, 종종 유전적 변이 및/또는 유전 이상을 검색하고/하거나 확인하는 속도를 두 자릿수 이상의 크기만큼 증가시킨다.

[0159] 일부 실시양태에서, 하나 이상의 프로세싱 단계는 하나 이상의 정규화 단계를 포함할 수 있다. 정규화는 본원에 기재되어 있거나 당분야에서 공지되어 있는 적합한 방법에 의해 수행될 수 있다. 일부 실시양태에서, 정규화는 상이한 스케일로 측정된 값을 개념적으로 공통된 스케일로 조절하는 단계를 포함한다. 일부 실시양태에서, 정규화는 조절된 값의 확률 분포를 정렬하는 정교한 수학적 조절을 포함한다. 일부 실시양태에서, 정규화는 분포를

표준 분포로 정렬하는 단계를 포함한다. 일부 실시양태에서, 정규화는 일부 중대한 결과(예를 들면, 오차 및 예외)의 영향을 제거하는 방식으로 상이한 데이터 세트에 대한 상응하는 정규화된 값의 비교를 가능하게 하는 수학적 조절을 포함한다. 일부 실시양태에서, 정규화는 크기 조정을 포함한다. 정규화는 종종 하나 이상의 데이터 세트를 예정된 변수 또는 식으로 나누는 단계를 포함한다. 정규화 방법의 비한정적 예는 부분별 정규화, GC 함량에 의한 정규화, 선형 및 비선형 최소 자승 회귀, LOESS, GC LOESS, LOWESS(국소적으로 가중된 산란도 평활화), PERUN, ChAI, 반복부 마스킹(RM), GC-정규화 및 반복부 마스킹(GCRM), cQn 및/또는 이들의 조합을 포함한다. 일부 실시양태에서, 유전적 변이(예를 들면, 이배수성, 미세중복, 미세결실)의 존재 또는 부재의 확인은 정규화 방법(예를 들면, 부분별 정규화, GC 함량에 의한 정규화, 선형 및 비선형 최소 자승 회귀, LOESS, GC LOESS, LOWESS(국소적으로 가중된 산란도 평활화), PERUN, ChAI, 반복부 마스킹(RM), GC-정규화 및 반복부 마스킹(GCRM), cQn, 당분야에서 공지된 정규화 방법 및/또는 이들의 조합)을 이용한다.

[0160] 임의의 적합한 수의 정규화가 이용될 수 있다. 일부 실시양태에서, 데이터 세트는 1회 이상, 5회 이상, 10회 이상 또는 심지어 20회 이상 정규화될 수 있다. 데이터 세트는 임의의 적합한 특징 또는 변수(예를 들면, 샘플 데이터, 기준 데이터, 또는 이들 둘 다)를 표시하는 값(예를 들면, 정규화 값)으로 정규화될 수 있다. 사용될 수 있는 데이터 정규화의 종류의 비한정적 예는 하나 이상의 선택된 검사 또는 기준 부분에 대한 미처리 카운트 데이터를 선택된 부분 또는 구획이 맵핑되는 염색체 또는 전체 게놈에 맵핑된 카운트의 총수로 정규화하는 것; 하나 이상의 선택된 부분에 대한 미처리 카운트 데이터를 선택된 부분 또는 분절이 맵핑되는 하나 이상의 부분 또는 염색체에 대한 중간 기준 카운트로 정규화하는 것; 미처리 카운트 데이터를 미리 정규화된 데이터 또는 이의 유도체로 정규화하는 것; 및 미리 정규화된 데이터를 하나 이상의 다른 예정된 정규화 변수로 정규화하는 것을 포함한다. 데이터 세트의 정규화는 종종 예정된 정규화 변수로서 선택된 특징 또는 성질에 따라 통계학적 오차를 단리하는 효과를 갖는다. 또한, 데이터 세트의 정규화는 종종 데이터가 공통된 스케일(예를 들면, 예정된 정규화 변수)을 갖게 함으로써 상이한 스케일을 갖는 데이터의 데이터 특성의 비교를 가능하게 한다. 일부 실시양태에서, 통계학적으로 유도된 값으로의 1회 이상의 정규화는 데이터 차이를 최소화하고 이상치 데이터의 중요성을 감소시키는 데에 이용될 수 있다. 정규화 값과 관련하여 부분 또는 기준 게놈의 부분의 정규화는 종종 "부분별 정규화"로서 지칭된다.

[0161] 일부 실시양태에서, 정규화를 포함하는 프로세싱 단계는 정적 윈도우(window)로의 정규화를 포함하고, 일부 실시양태에서 정규화를 포함하는 프로세싱 단계는 이동 또는 슬라이딩 윈도우로의 정규화를 포함한다. 본원에서 사용된 용어 "윈도우"는 분석을 위해 선택된 하나 이상의 부분을 지칭하고, 종종 비교를 위한 기준물로서 사용된다(예를 들면, 정규화 및/또는 다른 수학적 또는 통계학적 조작을 위해 사용된다). 본원에서 사용된 용어 "정적 윈도우로의 정규화"는 검사 대상체 데이터 세트와 기준 대상체 데이터 세트 사이의 비교를 위해 선택된 하나 이상의 부분을 사용하는 정규화 프로세스를 지칭한다. 일부 실시양태에서, 선택된 부분은 프로파일을 생성하는 데에 사용된다. 정적 윈도우는 일반적으로 조작 및/또는 분석 동안 변화지 않는 예정된 부분 세트를 포함한다. 본원에서 사용된 용어 "이동 윈도우로의 정규화" 및 "슬라이딩 윈도우로의 정규화"는 선택된 검사 부분의 게놈 영역(예를 들면, 바로 옆의 유전적 주변, 인접한 부분 또는 구획 등)에 위치하는 부분에 대해 수행된 정규화를 지칭하고, 이때 하나 이상의 선택된 검사 부분이 선택된 검사 부분을 직접적으로 둘러싸는 부분으로 정규화된다. 일부 실시양태에서, 선택된 부분은 프로파일을 생성하는 데에 사용된다. 슬라이딩 또는 이동 윈도우 정규화는 종종 인접한 검사 부분으로의 반복적 이동 또는 슬라이딩, 및 새로 선택된 검사 부분을 직접적으로 둘러싸거나 인접하는 부분으로의 새로 선택된 검사 부분의 정규화를 포함하고, 이때 인접한 윈도우는 하나 이상의 부분을 공유한다. 일부 실시양태에서, 복수의 선택된 검사 부분들 및/또는 염색체들이 슬라이딩 윈도우 프로세스에 의해 분석될 수 있다.

[0162] 일부 실시양태에서, 슬라이딩 또는 이동 윈도우로의 정규화는 하나 이상의 값을 생성할 수 있고, 이때 각각의 값은 게놈(예를 들면, 염색체)의 상이한 영역들로부터 선택된 상이한 기준 부분 세트로의 정규화를 표시한다. 일부 실시양태에서, 생성된 하나 이상의 값은 누적 합산(예를 들면, 선택된 부분, 도메인(예를 들면, 염색체의 일부) 또는 염색체에 대한 정규화된 카운트 프로파일의 적분의 수치적 추정치)이다. 슬라이딩 또는 이동 윈도우 프로세스에 의해 생성된 값은 프로파일을 생성하고 결과에 도달하는 것을 용이하게 하는 데에 사용될 수 있다. 일부 실시양태에서, 하나 이상의 부분의 누적 합산은 게놈 위치의 함수로서 디스플레이될 수 있다. 이동 또는 슬라이딩 윈도우 분석은 종종 미세결실 및/또는 미세삽입의 존재 또는 부재에 대해 게놈을 분석하는 데에 이용된다. 일부 실시양태에서, 하나 이상의 부분의 누적 합계의 디스플레이는 유전적 변이(예를 들면, 미세결실, 미세중복)의 영역의 존재 또는 부재를 확인하는 데에 이용된다. 일부 실시양태에서, 이동 또는 슬라이딩 윈도우 분석은 미세결실을 함유하는 게놈 영역을 확인하는 데에 이용되고, 일부 실시양태에서 이동 또는 슬라이딩 윈도

우 분석은 미세중복을 함유하는 게놈 영역을 확인하는 데에 이용된다.

[0163] 예를 들면, 이용될 수 있는 정규화 프로세스, 예컨대, LOESS, PERUN, ChAI 및 주성분 정규화 방법의 일부 예들은 이하에 더 상세히 기재되어 있다.

[0164] 일부 실시양태에서, 프로세싱 단계는 가중을 포함한다. 본원에서 사용된 용어 "가중된", "가중" 또는 "가중 함수" 또는 이의 문법적 파생어 또는 등가어는 일부 데이터 세트 특징 또는 변수의 영향을 다른 데이터 세트 특징 또는 변수에 비해 변경시키는 데에(예를 들면, 기준 게놈의 선택된 부분 또는 부분들에서의 데이터의 질 또는 유용성을 기초로 하나 이상의 부분 또는 기준 게놈의 부분에 함유된 데이터의 유의성 및/또는 기여를 증가시키거나 감소시키는 데에) 종종 이용되는, 데이터 세트의 일부 또는 전부의 수학적 조작을 지칭한다. 일부 실시양태에서, 가중 함수는 상대적으로 작은 측정 분산을 갖는 데이터의 영향을 증가시키고/시키거나 상대적으로 큰 측정 분산을 갖는 데이터의 영향을 감소시키는 데에 사용될 수 있다. 예를 들면, 과소표시된 또는 낮은 질의 서열 데이터를 갖는 기준 게놈의 부분은 데이터 세트에 대한 영향을 최소화하기 위해 "하향 가중될" 수 있는 반면, 기준 게놈의 선택된 부분은 데이터 세트에 대한 영향을 증가시키기 위해 "상향 가중될" 수 있다. 가중 함수의 비한정적 예는 $[1/(\text{표준 편차})^2]$이다. 가중 단계는 종종 정규화 단계와 실질적으로 유사한 방식으로 수행된다. 일부 실시양태에서, 데이터 세트는 예정된 변수(예를 들면, 가중 변수)로 나누어진다. 예정된 변수(예를 들면, 최소화된 표적 함수, Phi)는 종종 데이터 세트의 상이한 부분들을 상이하게 가중하도록(예를 들면, 다른 종류의 데이터의 영향을 감소시키면서 일부 종류의 데이터의 영향을 증가시키도록) 선택된다.

[0165] 일부 실시양태에서, 프로세싱 단계는 하나 이상의 수학적 및/또는 통계학적 조작을 포함할 수 있다. 임의의 적합한 수학적 및/또는 통계학적 조작이 본원에 기재된 데이터 세트를 분석하고/하거나 조작하는 데에 단독으로 또는 조합으로 이용될 수 있다. 임의의 적합한 수의 수학적 및/또는 통계학적 조작이 이용될 수 있다. 일부 실시양태에서, 데이터 세트는 1회 이상, 5회 이상, 10회 이상 또는 20회 이상 수학적으로 및/또는 통계학적으로 조작될 수 있다. 이용될 수 있는 수학적 및 통계학적 조작의 비한정적 예는 덧셈, 뺄셈, 곱셈, 나눗셈, 대수적 함수, 최소 자승 추정자, 곡선 피팅, 미분 방정식, 유리 다항식, 이중 다항식, 직교 다항식, z-점수, p-값, chi 값, phi 값, 피크 수준의 분석, 피크 에지 위치의 확인, 피크 면적 비의 계산, 중간 염색체 수준의 분석, 평균 절대 편차의 계산, 제곱 잔차의 합계, 평균, 표준 편차, 표준 오차 등 또는 이들의 조합을 포함한다. 수학적 및/또는 통계학적 조작은 서열 리드 데이터 또는 이의 프로세싱된 생성물의 전부 또는 일부에 대해 수행될 수 있다. 통계학적으로 조작될 수 있는 데이터 세트 변수 또는 특징의 비한정적 예는 미처리 카운트, 필터링된 카운트, 정규화된 카운트, 피크 높이, 피크 폭, 피크 면적, 피크 에지, 측면 허용오차(lateral tolerance), P-값, 중간 수준, 평균 수준, 게놈 영역 내의 카운트 분포, 핵산 종의 상대적 표시치 등 또는 이들의 조합을 포함한다.

[0166] 일부 실시양태에서, 프로세싱 단계는 하나 이상의 통계학적 알고리즘의 사용을 포함할 수 있다. 임의의 적합한 통계학적 알고리즘이 본원에 기재된 데이터 세트를 분석하고/하거나 조작하는 데에 단독으로 또는 조합으로 사용될 수 있다. 임의의 적합한 수의 통계학적 알고리즘이 사용될 수 있다. 일부 실시양태에서, 데이터 세트는 1개 이상, 5개 이상, 10개 이상 또는 20개 이상의 통계학적 알고리즘을 사용함으로써 분석될 수 있다. 본원에 기재된 방법과 함께 사용되기에 적합한 통계학적 알고리즘의 비한정적 예는 결정 트리, 카운터널(counternulls), 다중 비교, 옴니버스 검정, 베렌스-피셔 문제(Behrens-Fisher problem), 부트스트랩핑(bootstrapping), 독립적인 유의성 검정을 조합하기 위한 피셔의 방법, 널(null) 가설, I형 오차, II형 오차, 정확 검정, 1-샘플 Z 검정, 2-샘플 Z 검정, 1-샘플 t-검정, 페어링된 t-검정, 등분산을 갖는 2-샘플 풀링된 t-검정, 이분산을 갖는 2-샘플 비-풀링된 t-검정, 1-비율 z-검정, 풀링된 2-비율 z-검정, 비-풀링된 2-비율 z-검정, 1-샘플 chi-제곱 검정, 분산의 등가성에 대한 2-샘플 F 검정, 신뢰 구간, 신뢰가능한 구간, 유의성, 메타 분석, 단순 선형 회귀, 로부스트 선형 회귀 등 또는 이들의 조합을 포함한다. 통계학적 알고리즘을 사용함으로써 분석될 수 있는 데이터 세트 변수 또는 특징의 비한정적 예는 미처리 카운트, 필터링된 카운트, 정규화된 카운트, 피크 높이, 피크 폭, 피크 에지, 측면 허용오차, P-값, 중간 수준, 평균 수준, 게놈 영역 내의 카운트 분포, 핵산 종의 상대적 표시치 등 또는 이들의 조합을 포함한다.

[0167] 일부 실시양태에서, 데이터 세트는 다수의(예를 들면, 2개 이상의) 통계학적 알고리즘(예를 들면, 최소 자승 회귀, 주성분 분석, 선형 판별식 분석, 이차 판별식 분석, 배깅(bagging), 신경 네트워크, 지지체 벡터 머신 모델, 무작위 포레스트(random forests), 분류 트리 모델, K-최근접 이웃(K-nearest neighbors), 로지스틱 회귀 및/또는 손실 평활화(loss smoothing)), 및/또는 수학적 및/또는 통계학적 조작(예를 들면, 본원에서 조작으로서 지칭됨)을 이용함으로써 분석될 수 있다. 일부 실시양태에서, 다수의 조작의 이용은 결과를 제공하는 데에

사용될 수 있는 N-차원 공간을 생성할 수 있다. 일부 실시양태에서, 다수의 조작을 이용한 데이터 세트의 분석은 데이터 세트의 복잡성 및/또는 차원수를 감소시킬 수 있다. 예를 들면, 기준 데이터 세트에 대한 다수의 조작의 이용은 기준 샘플의 유전 상태(예를 들면, 선택된 유전적 변이에 대한 양성 또는 음성)에 따라 유전적 변이의 존재 또는 부재를 표시하는 데에 사용될 수 있는 N-차원 공간(예를 들면, 확률도)를 생성할 수 있다. 실질적으로 유사한 조작 세트를 사용한 검사 샘플의 분석은 검사 샘플 각각에 대한 N-차원 점을 생성하는 데에 사용될 수 있다. 검사 대상체 데이터 세트의 복잡성 및/또는 차원수는 종종 기준 데이터로부터 생성된 N-차원 공간과 용이하게 비교될 수 있는 단일 값 또는 N-차원 점까지 감소된다. 기준 대상체 데이터가 차지하는 N-차원 공간 내에 속하는 검사 샘플 데이터는 기준 대상체의 유전 상태와 실질적으로 유사한 유전 상태를 표시한다. 기준 대상체 데이터가 차지하는 N-차원 공간을 벗어나는 검사 샘플 데이터는 기준 대상체의 유전 상태와 실질적으로 유사하지 않은 유전 상태를 표시한다. 일부 실시양태에서, 기준물은 정배수체이거나 유전적 변이 또는 의학적 질환을 갖지 않는다.

[0168] 일부 실시양태에서, 데이터 세트가 카운팅되고 임의적으로 필터링되고 정규화된 후, 프로세싱된 데이터 세트는 하나 이상의 필터링 및/또는 정규화 절차에 의해 더 조작될 수 있다. 일부 실시양태에서, 하나 이상의 필터링 및/또는 정규화 절차에 의해 더 조작된 데이터 세트는 프로파일을 생성하는 데에 사용될 수 있다. 일부 실시양태에서, 하나 이상의 필터링 및/또는 정규화 절차는 종종 데이터 세트 복잡성 및/또는 차원수를 감소시킬 수 있다. 결과는 감소된 복잡성 및/또는 차원수의 데이터 세트를 기초로 제공될 수 있다.

[0169] 일부 실시양태에서, 부분은 오차의 척도(예를 들면, 표준 편차, 표준 오차, 계산된 분산, p-값, 평균 절대 오차(MAE), 평균치 절대 편차 및/또는 평균 절대 편차(MAD)에 따라 필터링될 수 있다. 일부 실시양태에서, 오차의 척도는 카운트 가변성을 지칭한다. 일부 실시양태에서, 부분은 카운트 가변성에 따라 필터링될 수 있다. 일부 실시양태에서, 카운트 가변성은 다수의 샘플들(예를 들면, 다수의 대상체들, 예를 들면, 50명 이상, 100명 이상, 500명 이상, 1,000명 이상, 5,000명 이상 또는 10,000명 이상의 대상체들로부터 수득된 다수의 샘플들)에 대한 기준 게놈의 부분(즉, 부분)에 맵핑된 카운트에 대해 측정된 오차의 척도이다. 일부 실시양태에서, 예정된 상한 범위 초과의 카운트 가변성을 갖는 부분이 필터링된다(예를 들면, 고려사항으로부터 배제된다). 일부 실시양태에서, 예정된 상한 범위는 약 50, 약 52, 약 54, 약 56, 약 58, 약 60, 약 62, 약 64, 약 66, 약 68, 약 70, 약 72, 약 74 또는 약 76 이상의 MAD 값이다. 일부 실시양태에서, 예정된 하한 범위 미만의 카운트 가변성을 갖는 부분이 필터링된다(예를 들면, 고려사항으로부터 배제된다). 일부 실시양태에서, 예정된 하한 범위는 약 40, 약 35, 약 30, 약 25, 약 20, 약 15, 약 10, 약 5, 약 1 또는 약 0 이하의 MAD 값이다. 일부 실시양태에서, 예정된 범위를 벗어난 카운트 가변성을 갖는 부분이 필터링된다(예를 들면, 고려사항으로부터 배제된다). 일부 실시양태에서, 예정된 범위는 0 초과 내지 약 76 미만, 약 74 미만, 약 73 미만, 약 72 미만, 약 71 미만, 약 70 미만, 약 69 미만, 약 68 미만, 약 67 미만, 약 66 미만, 약 65 미만, 약 64 미만, 약 62 미만, 약 60 미만, 약 58 미만, 약 56 미만, 약 54 미만, 약 52 미만 또는 약 50 미만의 MAD 값이다. 일부 실시양태에서, 예정된 범위는 0 초과 내지 약 67.7 미만의 MAD 값이다. 일부 실시양태에서, 예정된 범위 내의 카운트 가변성을 갖는 부분이 선택된다(예를 들면, 유전적 변이의 존재 또는 부재를 확인하기 위해 사용된다).

[0170] 일부 실시양태에서, 부분의 카운트 가변성은 분포(예를 들면, 정규 분포)를 나타낸다. 일부 실시양태에서, 부분은 분포의 변위치 내에서 선택된다. 일부 실시양태에서, 분포에 대한 약 99.9%, 99.8%, 99.7%, 99.6%, 99.5%, 99.4%, 99.3%, 99.2%, 99.1%, 99.0%, 98.9%, 98.8%, 98.7%, 98.6%, 98.5%, 98.4%, 98.3%, 98.2%, 98.1%, 98.0%, 97%, 96%, 95%, 94%, 93%, 92%, 91%, 90%, 85% 또는 80% 이하의 변위치, 또는 약 75% 이하의 변위치를 갖는 부분이 선택된다. 일부 실시양태에서, 카운트 가변성의 분포의 99% 변위치 내의 부분이 선택된다. 일부 실시양태에서, 99% 변위치 내에서 0 초과 내지 67.725 미만의 MAD를 갖는 부분이 선택되어, 기준 게놈의 안정한 부분 세트를 확인시켜 준다.

[0171] 예를 들면, PERUN과 관련하여 부분 필터링의 비한정적 예가 본원 및 국제 특허출원 제PCT/US12/59123호(WO2013/052913)(모든 본문, 표, 방정식 및 도면을 포함하는 전체 내용이 본원에 참고로 도입됨)에서 제공되어 있다. 부분은 오차의 척도를 기초로 또는 부분적으로 오차의 척도를 기초로 필터링될 수 있다. 일부 실시양태에서, 편차의 절대 값, 예컨대, R-계수를 포함하는 오차의 척도가 부분 제거 또는 가중을 위해 사용될 수 있다. 일부 실시양태에서, R-계수는 실제 측정으로부터의 예상된 카운트 값으로 나누어진 실제 측정으로부터의 예상된 카운트 값의 절대 편차의 합계로서 정의된다(예를 들면, 본원의 방정식 B). 편차의 절대 값을 포함하는 오차의 척도가 사용될 수 있지만, 오차의 적합한 척도가 대안적으로 사용될 수 있다. 일부 실시양태에서, 편차의 절대 값을 포함하지 않는 오차의 척도, 예컨대, 제곱에 기초한 분산이 사용될 수 있다. 일부 실시양태에서, 부분은 맵핑가능성의 척도(예를 들면, 맵핑가능성 점수)에 따라 필터링되거나 가중된다. 부분은 종종 부분들에 맵핑된

상대적으로 낮은 수의 서열 리드(예를 들면, 부분들에 맵핑된 0개, 1개, 2개, 3개, 4개 또는 5개의 리드)에 따라 필터링되거나 가중된다. 부분은 수행되는 분석의 종류에 따라 필터링될 수 있거나 가중될 수 있다. 예를 들면, 13번, 18번 및/또는 21번 염색체 이배수성 분석을 위해, 성염색체가 필터링될 수 있고, 상염색체 또는 상염색체 서브세트만이 분석될 수 있다.

[0172] 특정 실시양태에서, 하기 필터링 프로세스가 이용될 수 있다. 주어진 염색체(예를 들면, 21번 염색체) 내의 동일한 부분(예를 들면, 기준 게놈의 부분) 세트가 선택되고, 영향을 받은 샘플 및 영향을 받지 않은 샘플에서 리드의 수가 비교된다. 갭(gap)은 삼염색체성 21 및 정배수체 샘플과 관련되고, 21번 염색체의 대부분을 커버하는 부분 세트를 포함한다. 상기 부분 세트는 정배수체 샘플과 T21 샘플 사이에 동일하다. 부분이 한정될 수 있기 때문에 부분 세트와 단일 구획 사이의 식별은 중요하지 않다. 동일한 게놈 영역이 상이한 환자들에서 비교된다. 이 프로세스는 T21 이외에 또는 대신에 삼염색체성 분석, 예컨대, T13 또는 T18에 대한 삼염색체성 분석을 위해 이용될 수 있다.

[0173] 일부 실시양태에서, 데이터 세트가 카운팅되고 임의적으로 필터링되고 정규화된 후, 프로세싱된 데이터 세트는 가중에 의해 조작될 수 있다. 일부 실시양태에서, 하나 이상의 부분이 선택된 부분에 함유된 데이터(예를 들면, 무의미한 데이터, 비-정보제공 데이터)의 영향을 감소시키기 위한 가중을 위해 선택될 수 있고, 일부 실시양태에서 하나 이상의 부분이 선택된 부분에 함유된 데이터(예를 들면, 작은 측정된 분산을 갖는 데이터)의 영향을 향상시키거나 증가시키기 위한 가중을 위해 선택될 수 있다. 일부 실시양태에서, 데이터 세트는 큰 분산을 갖는 데이터의 영향을 감소시키고 작은 분산을 갖는 데이터의 영향을 증가시키는 단일 가중 함수를 사용함으로써 가중된다. 가중 함수는 종종 큰 분산을 갖는 데이터의 영향을 감소시키고 작은 분산을 갖는 데이터의 영향을 증가시키는 데에 사용된다(예를 들면, $[1/(\text{표준 편차})^2]$). 일부 실시양태에서, 가중에 의해 더 조작된 프로세싱된 데이터의 프로파일 도표가 분류 및/또는 결과의 제공을 용이하게 하기 위해 생성된다. 결과는 가중된 데이터의 프로파일 도표를 기초로 제공될 수 있다.

[0174] 부분의 필터링 또는 가중은 분석에서 하나 이상의 적합한 점에서 수행될 수 있다. 예를 들면, 부분은 서열 리드가 기준 게놈의 부분들로 맵핑되기 전 또는 후에 필터링될 수 있거나 가중될 수 있다. 일부 실시양태에서, 부분은 개별 게놈 부분에 대한 실험적 편향이 측정되기 전 또는 후에 필터링될 수 있거나 가중될 수 있다. 일부 실시양태에서, 부분은 게놈 구획 수준이 계산되기 전 또는 후에 필터링될 수 있거나 가중될 수 있다.

[0175] 일부 실시양태에서, 데이터 세트가 카운팅되고 임의적으로 필터링되고 정규화되고 임의적으로 가중된 후, 프로세싱된 데이터 세트는 하나 이상의 수학적 및/또는 통계학적(예를 들면, 통계학적 함수 또는 통계학적 알고리즘) 조작에 의해 조작될 수 있다. 일부 실시양태에서, 프로세싱된 데이터 세트는 하나 이상의 선택된 부분, 염색체 또는 염색체의 부분에 대한 Z-점수를 계산함으로써 더 조작될 수 있다. 일부 실시양태에서, 프로세싱된 데이터 세트는 P-값을 계산함으로써 더 조작될 수 있다. Z-점수 및 p-값을 계산하기 위한 방정식의 한 실시양태는 방정식 1(실시예 2)에 제시되어 있다. 일부 실시양태에서, 수학적 및/또는 통계학적 조작은 배수성 및/또는 태아 분획에 관한 하나 이상의 가정을 포함한다. 일부 실시양태에서, 하나 이상의 통계학적 및/또는 수학적 조작에 의해 더 조작된 프로세싱된 데이터의 프로파일 도표는 분류 및/또는 결과의 제공을 용이하게 하기 위해 생성된다. 결과는 통계학적으로 및/또는 수학적으로 조작된 데이터의 프로파일 도표를 기초로 제공될 수 있다. 통계학적으로 및/또는 수학적으로 조작된 데이터의 프로파일 도표를 기초로 제공된 결과는 배수성 및/또는 태아 분획에 관한 하나 이상의 가정을 포함한다.

[0176] 일부 실시양태에서, 데이터 세트가 카운팅되고 임의적으로 필터링되고 정규화된 후, N-차원 공간 및/또는 N-차원 점을 생성하기 위해 프로세싱된 데이터 세트에 대한 다수의 조작이 수행된다. 결과는 N-차원에서 분석된 데이터 세트의 프로파일 도표를 기초로 제공될 수 있다.

[0177] 일부 실시양태에서, 데이터 세트의 일부로서 또는 데이터 세트가 프로세싱되고/되거나 조작된 후에 하나 이상의 피크 수준 분석, 피크 폭 분석, 피크 에지 위치 분석, 피크 측면 허용오차 등, 이들로부터 유도된 분석, 또는 이들의 조합을 이용하여 데이터 세트를 프로세싱한다. 일부 실시양태에서, 하나 이상의 피크 수준 분석, 피크 폭 분석, 피크 에지 위치 분석, 피크 측면 허용오차 등, 이들로부터 유도된 분석, 또는 이들의 조합을 이용함으로써 프로세싱된 데이터의 프로파일 도표는 분류 및/또는 결과의 제공을 용이하게 하기 위해 생성된다. 결과는 하나 이상의 피크 수준 분석, 피크 폭 분석, 피크 에지 위치 분석, 피크 측면 허용오차 등, 이들로부터 유도된 분석, 또는 이들의 조합을 이용함으로써 프로세싱된 데이터의 프로파일 도표를 기초로 제공될 수 있다.

[0178] 일부 실시양태에서, 해당 유전적 변이를 실질적으로 갖지 않는 하나 이상의 기준 샘플은 기준 중간 카운트 프로파일을 생성하는 데에 사용될 수 있고, 이 기준 중간 카운트 프로파일은 유전적 변이의 부재를 표시하는 예정된

값을 생성할 수 있고, 검사 대상체가 유전적 변이를 보유한 경우 유전적 변이가 검사 대상체에서 위치하는 게놈 위치에 상응하는 영역에서 예정된 값으로부터 종종 벗어난다. 유전적 변이와 관련된 의학적 질환에 대한 위험을 갖거나 이러한 의학적 질환을 앓고 있는 검사 대상체에서, 선택된 부분 또는 구획에 대한 수치 값은 영향을 받지 않은 게놈 위치에 대한 예정된 값과 유의하게 상이할 것으로 예상된다. 일부 실시양태에서, 해당 유전적 변이를 보유하는 것으로 공지된 하나 이상의 기준 샘플은 기준 중간 카운트 프로파일을 생성하는 데에 사용될 수 있고, 이 기준 중간 카운트 프로파일은 유전적 변이의 존재를 표시하는 예정된 값을 생성할 수 있고, 검사 대상체가 유전적 변이를 보유하지 않는 게놈 위치에 상응하는 영역에서 예정된 값으로부터 종종 벗어난다. 유전적 변이와 관련된 의학적 질환에 대한 위험을 갖지 않거나 이러한 의학적 질환을 앓고 있지 않은 검사 대상체에서, 선택된 부분 또는 구획에 대한 수치 값은 영향을 받은 게놈 위치에 대한 예정된 값과 유의하게 상이할 것으로 예상된다.

[0179] 일부 실시양태에서, 데이터의 분석 및 프로세싱은 하나 이상의 가정의 사용을 포함할 수 있다. 적합한 수 또는 종류의 가정이 데이터 세트를 분석하거나 프로세싱하는 데에 사용될 수 있다. 데이터 프로세싱 및/또는 분석을 위해 사용될 수 있는 가정의 비한정적 예는 모체 배수성, 태아 기여, 기준 집단에서의 일부 서열들의 우세, 인종적 배경, 관련된 가족 구성원에서 선택된 의학적 질환의 우세, GC 정규화 및 반복부 마스킹(예를 들면, GCRM) 후 상이한 환자들 및/또는 실시로부터의 미처리 카운트 프로파일들 사이의 유사성, 동일한 일치가 PCR 인공물을 나타낸다는 가정(예를 들면, 동일한 염기 위치), 태아 정량자 분석(예를 들면, FQA)에 내재하는 가정, 쌍생아에 대한 가정(예를 들면, 2명의 쌍둥이 중 1명만이 영향을 받은 경우 유효 태아 분획은 총 측정된 태아 분획의 단지 50%임(세쌍둥이, 네쌍둥이 등의 경우에도 유사함)), 태아 무세포 DNA(예를 들면, cfDNA)가 전체 게놈을 균일하게 커버한다는 가정 등 및 이들의 조합을 포함한다.

[0180] 맵핑된 서열 리드의 질 및/또는 깊이가 원하는 신뢰 수준(예를 들면, 95% 이상의 신뢰 수준)에서 유전적 변이의 존재 또는 부재의 결과 예측을 허용하지 않는 경우, 정규화된 카운트 프로파일을 기초로 하나 이상의 추가 수학적 조작 알고리즘 및/또는 통계학적 예측 알고리즘이 데이터 분석 및/또는 결과의 제공에 유용한 추가 수치 값을 생성하는 데에 사용될 수 있다. 본원에서 사용된 용어 "정규화된 카운트 프로파일"은 정규화된 카운트를 사용함으로써 생성된 프로파일을 지칭한다. 정규화된 카운트 및 정규화된 카운트 프로파일을 생성하는 데에 이용될 수 있는 방법의 예는 본원에 기재되어 있다. 인지된 바와 같이, 카운팅되어 있는 맵핑된 서열 리드는 검사 샘플 카운트 또는 기준 샘플 카운트에 대하여 정규화될 수 있다. 일부 실시양태에서, 정규화된 카운트 프로파일은 도표로서 제시될 수 있다.

[0181] LOESS 정규화

[0182] LOESS는 k-최근접 이웃 기초 메타-모델에서 다수의 회귀 모델들을 병용하는, 당분야에서 공지된 회귀 모델링 방법이다. LOESS는 종종 국소적으로 가중된 다항식 회귀로서 지칭된다. 일부 실시양태에서, GC LOESS는 LOESS 모델을 기준 게놈의 부분에 대한 단편 카운트(예를 들면, 서열 리드, 카운트)와 GC 조성 사이의 관계에 적용한다. LOESS를 이용하여 데이터 점 세트를 통해 평활 곡선을 작도하는 것은 특히 각각의 평활화된 값이 y-축 산점도 기준 변수의 값의 범위에 걸쳐 가중된 이차 최소 자승 회귀에 의해 제공될 때 종종 LOESS 곡선으로서 지칭된다. 데이터 세트에서 각각의 점에 대해, LOESS 방법은 반응이 평가되는 점 근처의 설명 변수 값으로 저차(low-degree) 다항식을 데이터 서브세트에 피팅한다. 상기 다항식은 반응이 평가되는 점 근처의 점에게 더 많은 가중을 주고 더 멀리 떨어져 있는 점에게 더 적은 가중을 주는 가중된 최소 자승을 사용함으로써 피팅된다. 그 다음, 점에 대한 회귀 함수의 값이 그 데이터 점에 대한 설명 변수 값을 사용하여 국소 다항식을 평가함으로써 수득된다. LOESS 피트는 회귀 함수 값이 각각의 데이터 점에 대해 계산된 후 종종 완결된 것으로서 간주된다. 이 방법의 세부사항들 중 대부분, 예컨대, 다항식 모델의 차수 및 가중은 변화가능하다.

[0183] PERUN 정규화

[0184] 핵산 표시자와 관련된 오차를 감소시키는 정규화 방법은 본원 및 국제 특허출원 제PCT/US12/59123호(WO2013/052913)(모든 본문, 표, 방정식 및 도면을 포함하는 전체 내용이 본원에 참고로 도입됨)에 기재된 파라미터화된 오차 제거 및 비-편향된 정규화(PERUN)로서 본원에서 지칭된다. PERUN 방법은 다양한 핵산 표시자(예를 들면, 핵산 서열 리드)에 기초한 예측을 혼동시키는 오차의 영향을 감소시킬 목적으로 이러한 핵산 표시자에 적용될 수 있다.

[0185] 예를 들면, PERUN 방법은 샘플로부터의 핵산 서열 리드에 적용될 수 있고 게놈 구획 수준 측정을 방해할 수 있는 오차의 영향을 감소시킬 수 있다. 이러한 적용은 핵산 서열 리드를 사용하여 뉴클레오티드 서열의 상이한 수준(예를 들면, 부분, 게놈 구획 수준)으로서 나타나는 유전적 변이의 존재 또는 부재를 대상체에서 확인하는 데

에 유용하다. 부분에서의 변이의 비한정적 예는 염색체 이배수성(예를 들면, 삼염색체성 21, 삼염색체성 18, 삼염색체성 13) 및 성염색체(예를 들면, 여성에서의 XX 대 남성에서의 XY)의 존재 또는 부재이다. 상염색체(예를 들면, 성염색체 이외의 염색체)의 삼염색체성은 영향을 받은 상염색체로서 지칭될 수 있다. 게놈 구획 수준에서의 변이의 다른 비한정적 예는 미세결실, 미세삽입, 중복 및 모자이크 현상을 포함한다.

[0186] 일부 적용에서, PERUN 방법은 기준 게놈의 특정 부분들에 맵핑된 핵산 리드를 정규화함으로써 실험적 편향을 감소시킬 수 있고, 이때 상기 부분은 부분으로서 지칭되고 종종 기준 게놈의 부분으로서 지칭된다. 이러한 적용에서, PERUN 방법은 일반적으로 3차원에서 다수의 샘플들에 걸쳐 기준 게놈의 특정 부분에서 핵산 리드의 카운트를 정규화한다. PERUN 및 이의 적용의 상세한 설명은 본원의 실시예 부분, 국제 특허출원 제PCT/US12/59123호(WO2013/052913) 및 미국 특허출원 공보 제20130085681호(모든 본문, 표, 방정식 및 도면을 포함하는 전체 내용이 본원에 참고로 도입됨)에서 제공되어 있다.

[0187] 일부 실시양태에서, PERUN 방법은 (a) 검사 샘플에 대한 기준 게놈의 부분들에 맵핑된 서열 리드 카운트, (b) 검사 샘플에 대한 실험적 편향(예를 들면, GC 편향), 및 (c) (i) 서열 리드가 맵핑되는 기준 게놈의 부분에 대한 실험적 편향과 (ii) 상기 부분들에 맵핑된 서열 리드의 카운트 사이의 피팅된 관계에 대한 하나 이상의 피트 파라미터(예를 들면, 피트의 추정치)로부터 기준 게놈의 부분에 대한 게놈 구획 수준을 계산하는 단계를 포함한다. 기준 게놈의 각각의 부분들에 대한 실험적 편향은 (i) 기준 게놈의 각각의 부분들에 맵핑된 서열 리드의 카운트와 (ii) 기준 게놈의 각각의 부분들에 대한 맵핑 특징 사이의 각각 샘플에 대한 피팅된 관계에 따라 다수의 샘플들에 걸쳐 측정될 수 있다. 각각의 샘플에 대한 이 피팅된 관계는 3차원에서 다수의 샘플들에 대해 조립될 수 있다. PERUN 방법이 실험적 편향에 따라 조립의 순서를 정하지 않고 실시될 수 있지만, 일부 실시양태에서 조립은 실험적 편향에 따라 순서가 정해질 수 있다. 각각의 샘플에 대한 피팅된 관계 및 기준 게놈의 각각의 부분들에 대한 피팅된 관계는 당분야에서 공지된 적합한 피팅 프로세스에 의해 선형 함수 또는 비선형 함수에 독립적으로 피팅될 수 있다.

[0188] 일부 실시양태에서, 관계는 기하학적 및/또는 도식적 관계이다. 일부 실시양태에서, 관계는 수학적 관계이다. 일부 실시양태에서, 관계는 작도된다. 일부 실시양태에서, 관계는 선형 관계이다. 일부 실시양태에서, 관계는 비선형 관계이다. 일부 실시양태에서, 관계는 회귀(예를 들면, 회귀 선)이다. 회귀는 선형 회귀 또는 비선형 회귀일 수 있다. 관계는 수학적 방정식으로 표현될 수 있다. 종종, 관계는 부분적으로 하나 이상의 상수에 의해 정의된다. 관계는 당분야에서 공지된 방법에 의해 생성될 수 있다. 일부 실시양태에서, 하나 이상의 샘플에 대한 2차원 관계가 생성될 수 있고, 오차를 증명하는 또는 오차를 증명할 수 있는 변수가 하나 이상의 차원에 대해 선택될 수 있다. 관계는 예를 들면, 사용자에 의해 제공된 2개 이상의 변수의 값을 사용하여 그래프를 작도하는, 당분야에서 공지된 도식화 소프트웨어를 사용함으로써 생성될 수 있다. 관계는 당분야에서 공지된 방법(예를 들면, 도식화 소프트웨어)을 사용함으로써 피팅될 수 있다. 일부 관계는 선형 회귀에 의해 피팅될 수 있고, 선형 회귀는 기울기 값 및 절편 값을 생성할 수 있다. 예를 들면, 일부 관계는 종종 선형이 아니고 비선형 함수, 예컨대, 포물선, 쌍곡선 또는 지수 함수(예를 들면, 이차 함수)에 의해 피팅될 수 있다.

[0189] PERUN 방법에서, 하나 이상의 피팅된 관계는 선형일 수 있다. 임신 여성으로부터의 무세포 순환 핵산을 분석하는 경우, 실험적 편향이 GC 편향이고 맵핑 특징이 GC 함량일 때, (i) 각각의 부분들에 맵핑된 서열 리드의 카운트와 (ii) 기준 게놈의 각각의 부분들에 대한 GC 함량 사이의 샘플에 대한 피팅된 관계는 선형일 수 있다. 후자 피팅된 관계의 경우, 기울기는 GC 편향에 관한 것이고, GC 편향 계수는 피팅된 관계가 다수의 샘플들에 걸쳐 조립될 때 각각의 샘플에 대해 측정될 수 있다. 이러한 실시양태에서, (i) 부분에 대한 GC 편향 계수와 (ii) 부분들에 맵핑된 서열 리드의 카운트 사이의 다수의 샘플들 및 부분에 대한 피팅된 관계도 선형일 수 있다. 절편 및 기울기는 후자 피팅된 관계로부터 수득될 수 있다. 이러한 적용에서, 기울기는 GC 함량을 기초로 샘플 특이적 편향을 다루고, 절편은 모든 샘플들에 공통된 부분 특이적 약화 패턴을 다룬다. PERUN 방법은 결과(예를 들면, 유전적 변이의 존재 또는 부재; 태아 성별의 확인)를 제공하기 위해 게놈 구획 수준을 계산할 때 이러한 샘플 특이적 편향 및 부분 특이적 약화를 유의하게 감소시킬 수 있다.

[0190] 일부 실시양태에서, PERUN 정규화는 선형 함수에의 피팅을 이용하고 방정식 A, 방정식 B 또는 이들로부터 유도된 식에 의해 기재된다.

[0191] 방정식 A:

[0192] $M = LI + GS$

[0193] 방정식 B:

[0194] $L = (M - GS)/I$

[0195] 일부 실시양태에서, L은 PERUN 정규화된 수준 또는 프로파일이다. 일부 실시양태에서, L은 PERUN 정규화 절차로부터의 원하는 결과이다. 일부 실시양태에서, L은 부분 특이적이다. 일부 실시양태에서, L은 기준 게놈의 다수의 부분들에 따라 측정되고 게놈, 염색체, 또는 이들의 부분 또는 분절의 PERUN 정규화된 수준을 나타낸다. 수준 L은 종종 추가 분석을 위해(예를 들면, Z-값, 모체 결실/중복, 태아 미세결실/미세중복, 태아 성별, 성 이배수성 등을 확인하기 위해) 사용된다. 방정식 B에 따라 정규화하는 방법은 파라미터화된 오차 제거 및 비-편향된 정규화(PERUN)로서 명명된다.

[0196] 일부 실시양태에서, G는 선형 모델, LOESS 또는 임의의 등가 방법을 이용함으로써 측정된 GC 편향 계수이다. 일부 실시양태에서, G는 기울기이다. 일부 실시양태에서, GC 편향 계수 G는 기준 게놈으로부터 측정된 부분 i에 대한 카운트 M(예를 들면, 미처리 카운트) 및 부분 i의 GC 함량에 대한 회귀의 기울기로서 평가된다. 일부 실시양태에서, G는 M으로부터 추출되고 관계에 따라 측정된 이차 정보를 나타낸다. 일부 실시양태에서, G는 부분 특이적 카운트의 세트와 샘플(예를 들면, 검사 샘플)에 대한 부분 특이적 GC 함량 값의 세트의 관계를 나타낸다. 일부 실시양태에서, 부분 특이적 GC 함량은 기준 게놈으로부터 유도된다. 일부 실시양태에서, 부분 특이적 GC 함량은 관찰된 또는 측정된(예를 들면, 샘플로부터 측정된) GC 함량으로부터 유도된다. GC 편향 계수는 종종 샘플 군 내의 각각의 샘플에 대해 측정되고, 일반적으로 시험 샘플에 대해 측정된다. GC 편향 계수는 종종 샘플 특이적이다. 일부 실시양태에서, GC 편향 계수는 상수이다. 일부 실시양태에서, 일단 샘플에 대해 유도된 GC 편향 계수는 변하지 않는다.

[0197] 일부 실시양태에서, I는 절편이고, S는 선형 관계로부터 유도된 기울기이다. 일부 실시양태에서, I 및 S가 유도되는 관계는 G가 유도되는 관계와 상이하다. 일부 실시양태에서, I 및 S가 유도되는 관계는 주어진 실험 셋업에 대해 고정된다. 일부 실시양태에서, I 및 S는 다수의 샘플들에 따른 카운트(예를 들면, 미처리 카운트) 및 GC 편향 계수에 따라 선형 관계로부터 유도된다. 일부 실시양태에서, I 및 S는 시험 샘플과 관계없이 유도된다. 일부 실시양태에서, I 및 S는 다수의 샘플들로부터 유도된다. I 및 S는 종종 부분 특이적이다. 일부 실시양태에서, I 및 S는 정배수체 샘플에서 기준 게놈의 모든 부분들에 대해 L이 1이라는 가정 하에 측정된다. 일부 실시양태에서, 선형 관계는 정배수체 샘플에 대해 결정되고, 선택된 부분(L이 1이라고 가정함)에 대한 특이성을 갖는 I 및 S 값이 측정된다. 일부 실시양태에서, 동일한 절차가 인간 게놈 내의 기준 게놈의 모든 부분들에 적용되고, 절편 I 및 기울기 S의 세트가 모든 부분에 대해 측정된다.

[0198] 일부 실시양태에서, 교차검증 방법이 적용된다. 교차검증은 종종 회전 평가로서 지칭된다. 일부 실시양태에서, 교차검증 방법은 예측 모델(예를 들면, PERUN)이 검사 샘플을 사용하여 실제로 얼마나 정확히 수행할 것인지를 평가하기 위해 적용된다. 일부 실시양태에서, 한 라운드의 교차검증은 데이터의 샘플을 상보적 서브세트로 분할하는 단계, 한 서브세트(예를 들면, 종종 트레이닝 세트로서 지칭됨)에 대한 교차검증 분석을 수행하는 단계, 및 또 다른 서브세트(예를 들면, 종종 검증 세트 또는 검사 세트로서 지칭됨)를 사용하여 분석을 검증하는 단계를 포함한다. 일부 실시양태에서, 상이한 구획들 및/또는 상이한 서브세트를 사용하여 다수의 라운드의 교차검증을 수행한다. 교차검증 방법의 비한정적 예는 리브-원-아웃, 슬라이딩 에지, K-폴드(fold), 2-폴드, 반복부무작위 서브샘플링 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, 교차검증은 공지된 정배수체 태아를 포함하는 90%의 샘플 세트를 함유하는 작업 세트를 무작위적으로 선택하고 그 서브세트를 사용하여 모델을 트레이닝한다. 일부 실시양태에서, 무작위 선택을 100회 반복하여 모든 부분들에 대한 100개 기울기 및 100개 절편의 세트를 생성한다.

[0199] 일부 실시양태에서, M의 값은 검사 샘플로부터 유도된 측정된 값이다. 일부 실시양태에서, M은 부분에 대해 측정된 미처리 카운트이다. 일부 실시양태에서, 값 I 및 S가 부분에 대해 사용가능한 경우, 측정치 M은 검사 샘플로부터 측정되고 방정식 B에 따라 게놈, 염색체, 또는 이들의 분절 또는 부분에 대한 PERUN 정규화된 수준 L을 측정하는 데에 사용된다.

[0200] 따라서, PERUN 방법을 다수의 샘플들에 걸쳐 서열 리드에 동시에 적용하는 것은 (i) 샘플 특이적 실험적 편향(예를 들면, GC 편향) 및 (ii) 샘플들에 공통된 부분 특이적 약화에 의해 약기된 오차를 유의하게 감소시킬 수 있다. 오차의 이들 2개 공급원들 각각이 별도로 또는 순차적으로 다루어지는 다른 방법은 종종 PERUN 방법만큼 효과적으로 이들을 감소시킬 수 없다. 이론에 의해 한정되지 않지만, PERUN 방법은 부분적으로 그의 일반적인 덧셈 프로세스가 다른 정규화 방법(예를 들면, GC-LOESS)에서 이용된 일반적인 곱셈 프로세스만큼 많이 넓게 확대하지 않기 때문에 오차를 더 효과적으로 감소시킨다고 생각된다.

[0201] 추가 정규화 및 통계학적 기법이 PERUN 방법과 함께 이용될 수 있다. 추가 프로세스는 PERUN 방법의 이용 전,

후 및/또는 동안에 적용될 수 있다. PERUN 방법과 함께 이용될 수 있는 프로세스의 비한정적 예는 이하에 기재되어 있다.

[0202] 일부 실시양태에서, GC 함량에 대한 게놈 구획 수준의 이차 정규화 또는 조절이 PERUN 방법과 함께 이용될 수 있다. 적합한 GC 함량 조절 또는 정규화 절차(예를 들면, GC-LOESS, GCRM)가 이용될 수 있다. 일부 실시양태에서, 특정 샘플이 추가 GC 정규화 프로세스의 적용을 위해 확인될 수 있다. 예를 들면, PERUN 방법의 적용은 각각의 샘플에 대한 GC 편향을 측정할 수 있고, 특정 역치 초과의 GC 편향과 관련된 샘플이 추가 GC 정규화 프로세스를 위해 선택될 수 있다. 이러한 실시양태에서, 예정된 역치 수준이 추가 GC 정규화를 위해 이러한 샘플을 선택하는 데에 사용될 수 있다.

[0203] 일부 실시양태에서, 부분 필터링 또는 가중 프로세스가 PERUN 방법과 함께 이용될 수 있다. 적합한 부분 필터링 또는 가중 프로세스가 이용될 수 있고, 비한정적 예가 본원, 국제 특허출원 제PCT/US12/59123호(WO2013/052913) 및 미국 특허출원 공보 제20130085681호(모든 본문, 표, 방정식 및 도면을 포함하는 전체 내용이 본원에 참고로 도입됨)에 기재되어 있다. 일부 실시양태에서, 모체 삽입, 중복 및/또는 결실(예를 들면, 모체 및/또는 태아 카피 수 변이)과 관련된 오차를 감소시키는 정규화 기법이 PERUN 방법과 함께 이용된다.

[0204] PERUN 방법에 의해 계산된 게놈 구획 수준은 결과를 제공하는 데에 직접적으로 사용될 수 있다. 일부 실시양태에서, 게놈 구획 수준은 태아 분획이 약 2% 내지 약 6% 이상(예를 들면, 약 4% 이상의 태아 분획)인 샘플에 대한 결과를 제공하는 데에 직접적으로 사용될 수 있다. PERUN 방법에 의해 계산된 게놈 구획 수준은 종종 결과의 제공을 위해 더 프로세싱된다. 일부 실시양태에서, 계산된 게놈 구획 수준은 표준화된다. 일부 실시양태에서, 검사 부분(예를 들면, 21번 염색체)에 대한 계산된 게놈 구획 수준의 합계, 평균 또는 중간치를 검사 부분 이외의 부분(예를 들면, 21번 염색체 이외의 상염색체)에 대한 계산된 게놈 구획 수준의 합계, 평균 또는 중간치로 나누어 실험적 게놈 구획 수준을 생성할 수 있다. 실험적 게놈 구획 수준 또는 미처리 게놈 구획 수준은 표준화 분석, 예컨대, Z-점수의 계산의 일부로서 사용될 수 있다. Z-점수는 예상된 게놈 구획 수준을 실험적 게놈 구획 수준 또는 미처리 게놈 구획 수준으로부터 뺌으로써 샘플에 대해 생성될 수 있고, 생성된 값은 샘플에 대한 표준 편차로 나누어질 수 있다. 일부 실시양태에서, 생성된 Z-점수는 상이한 샘플들에 대해 분포되고 분석되거나, 다른 변수, 예컨대, 태아 분획 등과 관련되고 분석되어 결과를 제공할 수 있다.

[0205] 상기 인지된 바와 같이, PERUN 방법은 GC 편향 및 GC 함량에 따른 정규화 그 자체로 한정되지 않고, 다른 오차 공급원과 관련된 오차를 감소시키는 데에 사용될 수 있다. 비-GC 함량 편향의 공급원의 비한정적 예는 맵핑가능성이다. GC 편향 및 함량 이외의 정규화 파라미터가 다루어질 때, 하나 이상의 피팅된 관계는 비선형(예를 들면, 쌍곡선, 지수)일 수 있다. 일부 실시양태에서, 실험적 편향이 비선형 관계로부터 측정되는 경우, 예를 들면, 실험적 편향 곡률 추정치가 분석될 수 있다.

[0206] PERUN 방법은 다양한 핵산 표시자들에게 적용될 수 있다. 핵산 표시자의 비한정적 예는 마이크로어레이 상의 특정 위치에서의 핵산 서열 리드 및 핵산 수준이다. 서열 리드의 비한정적 예는 무세포 순환 DNA, 무세포 순환 RNA, 세포 DNA 및 세포 RNA로부터 수득된 서열 리드를 포함한다. PERUN 방법은 적합한 기준 서열, 예컨대, 게놈 기준 DNA, 세포 기준 RNA(예를 들면, 전사체(transcriptome)), 및 이들의 부분(예를 들면, DNA 또는 RNA 전사체의 게놈 상보체의 일부(들), 염색체의 일부(들))에 맵핑된 서열 리드에 적용될 수 있다.

[0207] 따라서, 일부 실시양태에서, 세포 핵산(예를 들면, DNA 또는 RNA)은 핵산 표시자로서 사용될 수 있다. PERUN 방법을 이용하여 기준 게놈 부분들에 맵핑된 세포 핵산 리드를 정규화할 수 있다. 특정 단백질에 결합된 세포 핵산은 종종 염색질 면역침전(ChIP) 공정으로서 지칭된다. ChIP에 의해 농축된 핵산은 세포 단백질과 회합된 핵산, 예를 들면, DNA 또는 RNA이다. 당분야에서 공지된 기술을 이용하여 ChIP에 의해 농축된 핵산의 리드를 수득할 수 있다. ChIP에 의해 농축된 핵산의 리드를 기준 게놈의 하나 이상의 부분들로 맵핑할 수 있고, PERUN 방법을 이용하여 맵핑 결과를 정규화하여 결과를 제공할 수 있다.

[0208] 일부 실시양태에서, 세포 RNA가 핵산 표시자로서 사용될 수 있다. 세포 RNA 리드를 기준 RNA 부분들로 맵핑할 수 있고 PERUN 방법을 이용하여 정규화하여 결과를 제공할 수 있다. 전사체로서 지칭되는 세포 RNA 또는 이의 분절에 대한 공지된 서열은 샘플로부터의 RNA 리드가 맵핑될 수 있는 기준물로서 사용될 수 있다. 당분야에서 공지된 기술을 이용하여 샘플 RNA의 리드를 수득할 수 있다. PERUN 방법을 이용하여 기준물로 맵핑된 RNA 리드의 결과를 정규화하여 결과를 제공할 수 있다.

[0209] 일부 실시양태에서, 마이크로어레이 핵산 수준이 핵산 표시자로서 사용될 수 있다. PERUN 방법을 이용하여 어레이 상에서 특정 어드레스(address) 또는 혼성화 핵산에 대해 샘플들 전체에 걸쳐 수득된 핵산 수준을 분석함으

로써 마이크로어레이 분석에 의해 제공된 핵산 표시자를 정규화할 수 있다. 이 방식에서, 마이크로어레이 상의 특정 어드레스 또는 혼성화 핵산은 맵핑된 핵산 서열 리드에 대한 부분과 유사하고, PERUN 방법은 마이크로어레이 데이터를 정규화하여 개선된 결과를 제공하는 데에 이용될 수 있다.

[0210] ChAI 정규화

[0211] 핵산 표시자와 관련된 오차를 감소시키는 데에 이용될 수 있는 또 다른 정규화 방법은 본원에서 ChAI로서 지칭되고 종종 주성분 분석을 이용한다. 일부 실시양태에서, 주성분 분석은 (a) 기준 게놈의 부분을 리드 밀도 분포에 따라 필터링하여 필터링된 부분의 리드 밀도를 포함하는 검사 샘플에 대한 리드 밀도 프로파일을 제공하는 단계로서, 이때 리드 밀도가 임신 여성으로부터의 검사 샘플로부터의 순환 무세포 핵산의 서열 리드를 포함하고, 리드 밀도 분포가 다수의 샘플들에 대한 부분의 리드 밀도에 대해 측정되는 것인 단계, (b) 주성분 분석에 의해 공지된 정배수체 샘플 세트로부터 수득된 하나 이상의 주성분에 따라 검사 샘플에 대한 리드 밀도 프로파일을 조절하여 조절된 리드 밀도를 포함하는 검사 샘플 프로파일을 제공하는 단계, 및 (c) 검사 샘플 프로파일을 기준 프로파일과 비교하여 비교를 제공하는 단계를 포함한다. 일부 실시양태에서, 주성분 분석은 (d) 상기 비교에 따라 검사 샘플에 대한 유전적 변이의 존재 또는 부재를 확인하는 단계를 포함한다.

[0212] 부분 필터링

[0213] 일부 실시양태에서, 하나 이상의 부분(예를 들면, 게놈의 부분)은 필터링 프로세스에 의해 고려사항으로부터 제거된다. 일부 실시양태에서, 하나 이상의 부분이 필터링되어(예를 들면, 필터링 프로세스로 처리되어) 필터링된 부분을 제공한다. 일부 실시양태에서, 필터링 프로세스는 일부 부분들을 제거하고 부분(예를 들면, 부분의 서브세트)을 보유한다. 필터링 프로세스 후, 보유된 부분은 종종 본원에서 필터링된 부분으로서 지칭된다. 일부 실시양태에서, 기준 게놈의 부분이 필터링된다. 일부 실시양태에서, 필터링 프로세스에 의해 제거되는 기준 게놈의 부분은 유전적 변이(예를 들면, 염색체 이배수성, 미세중복, 미세결실)의 존재 또는 부재의 확인에 포함되지 않는다. 일부 실시양태에서, 리드 밀도와 관련된 부분(예를 들면, 리드 밀도가 부분에 대한 것인 경우)은 필터링 프로세스에 의해 제거되고, 제거된 부분과 관련된 리드 밀도는 유전적 변이(예를 들면, 염색체 이배수성, 미세중복, 미세결실)의 존재 또는 부재의 확인에 포함되지 않는다. 일부 실시양태에서, 리드 밀도 프로파일은 필터링된 부분의 리드 밀도를 포함하고/하거나 이러한 리드 밀도로 구성된다. 당분야에서 공지되어 있거나 본원에 기재되어 있는 임의의 적합한 기준 및/또는 방법을 이용하여 부분을 고려사항으로부터 선택할 수 있고/있거나, 필터링할 수 있고/있거나 제거할 수 있다. 부분의 필터링을 위해 사용되는 기준의 비한정적 예는 중복 데이터(예를 들면, 중복 또는 중첩 맵핑된 리드), 비-정보제공 데이터(예를 들면, 0의 맵핑된 카운트를 갖는 기준 게놈의 부분), 과다표시된 또는 과소표시된 서열을 갖는 기준 게놈의 부분, GC 함량, 무의미한 데이터, 맵핑가능성, 카운트, 카운트 가변성, 리드 밀도, 리드 밀도의 가변성, 불확실성의 척도, 반복가능성 측정치 등 또는 이들의 조합을 포함한다. 부분은 종종 카운트의 분포 및/또는 리드 밀도의 분포에 따라 필터링된다. 일부 실시양태에서, 부분은 카운트 및/또는 리드 밀도의 분포에 따라 필터링되고, 이때 카운트 및/또는 리드 밀도는 하나 이상의 기준 샘플로부터 수득된다. 하나 이상의 기준 샘플은 종종 본원에서 트레이닝 세트로서 지칭된다. 일부 실시양태에서, 부분은 카운트 및/또는 리드 밀도의 분포에 따라 필터링되고, 이때 카운트 및/또는 리드 밀도는 하나 이상의 검사 샘플로부터 수득된다. 일부 실시양태에서, 부분은 리드 밀도 분포에 대한 불확실성의 척도에 따라 필터링된다. 일부 실시양태에서, 리드 밀도에서의 큰 편차를 나타내는 부분은 필터링 프로세스에 의해 제거된다. 예를 들면, 리드 밀도의 분포(예를 들면, 평균치, 평균 또는 중간 리드 밀도의 분포, 예를 들면, 도 37a)가 측정될 수 있고, 이때 분포에서 각각의 리드 밀도는 동일한 부분들에 맵핑된다. 불확실성의 척도(예를 들면, MAD)는 다수의 샘플들에 대한 리드 밀도의 분포를 비교함으로써 측정될 수 있고, 이때 게놈의 각각의 부분은 불확실성의 척도와 관련되어 있다. 상기 예에 따라, 부분은 각각의 부분과 관련된 불확실성의 척도(예를 들면, 표준 편차(SD), MAD) 및 예정된 역치에 따라 필터링될 수 있다. 도 37b는 다수의 샘플들에 대한 리드 밀도 분포에 따라 측정된, 부분에 대한 MAD 값의 분포를 보여준다. 예정된 역치는 허용가능한 MAD 값의 범위를 둘러싸는 수직 쇄선에 의해 표시된다. 도 37b의 예에서, 허용가능한 범위 내의 MAD 값을 포함하는 부분은 보유되고, 허용가능한 범위 밖의 MAD 값을 포함하는 부분은 필터링 프로세스에 의해 고려사항으로부터 제거된다. 일부 실시양태에서, 상기 예에 따라, 불확실성의 예정된 측정치를 벗어나는 리드 밀도 값(예를 들면, 중간, 평균치 또는 평균 리드 밀도)을 포함하는 부분은 종종 필터링 프로세스에 의해 고려사항으로부터 제거된다. 일부 실시양태에서, 분포의 사분위수간 범위를 벗어나는 리드 밀도 값(예를 들면, 중간, 평균치 또는 평균 리드 밀도)을 포함하는 부분은 필터링 프로세스에 의해 고려사항으로부터 제거된다. 일부 실시양태에서, 분포의 사분위수간 범위를 2배, 3배, 4배 또는 5배 초과의 수준으로 벗어나는 리드 밀도 값을 포함하는 부분은 필터링 프로세스에 의해 고려사항으로부터 제거된다. 일부 실시양태에서, 2 시그마, 3 시그마, 4 시그마, 5 시그마, 6 시그마, 7

시그마 또는 8 시그마(예를 들면, 시그마가 표준 편차에 의해 한정된 범위인 경우) 초과의 수준으로 벗어나는 리드 밀도 값을 포함하는 부분은 필터링 프로세스에 의해 고려사항으로부터 제거된다.

[0214] 일부 실시양태에서, 시스템은 필터링 모듈(18, 도 42a)을 포함한다. 필터링 모듈은 종종 부분(예를 들면, 예정된 크기 및/또는 중첩의 부분, 기준 게놈 내에 위치하는 부분), 및 부분과 관련된 리드 밀도를 종종 또 다른 적합한 모듈(예를 들면, 분포 모듈(12), 도 42a)로부터 수용하고/하거나, 회수하고/하거나 저장한다. 일부 실시양태에서, 선택된 부분(예를 들면, 20(도 42a), 예를 들면, 필터링된 부분)은 필터링 모듈에 의해 제공된다. 일부 실시양태에서, 필터링 모듈은 필터링된 부분을 제공하고/하거나 고려사항으로부터 부분을 제거하는 데에 요구된다. 일부 실시양태에서, 필터링 모듈은 고려사항으로부터 리드 밀도를 제거하고, 이때 리드 밀도는 제거된 부분과 관련되어 있다. 필터링 모듈은 종종 선택된 부분(예를 들면, 필터링된 부분)을 또 다른 적합한 모듈(예를 들면, 분포 모듈(12), 도 42a)에게 제공한다. 필터링 모듈의 비한정적 예는 실시예 7에서 제공되어 있다.

[0215] 편향 추정치

[0216] 서열결정 기술은 다수의 편향 공급원들에 취약할 수 있다. 종종 서열결정 편향은 국소 편향(예를 들면, 국소 게놈 편향)이다. 국소 편향은 종종 서열 리드의 수준에서 나타난다. 국소 게놈 편향은 임의의 적합한 국소 편향일 수 있다. 국소 편향의 비한정적 예는 서열 편향(예를 들면, GC 편향, AT 편향 등), DNase I 민감성과 관련된 편향, 엔트로피, 반복 서열 편향, 염색질 구조 편향, 중합효소 오류 속도 편향, 팔린드롬(palindrome) 편향, 도립된 반복부 편향, PCR 관련된 편향 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, 국소 편향의 공급원은 확인되어 있지 않거나 공지되어 있지 않다.

[0217] 일부 실시양태에서, 국소 게놈 편향 추정치가 측정된다. 국소 게놈 편향 추정치는 종종 본원에서 국소 게놈 편향 평가치로서 지칭된다. 국소 게놈 편향 추정치는 기준 게놈, 또는 이의 분절 또는 부분에 대해 측정될 수 있다. 일부 실시양태에서, 국소 게놈 편향 추정치는 하나 이상의 서열 리드(예를 들면, 샘플의 일부 또는 모든 서열 리드)에 대해 측정된다. 국소 게놈 편향 추정치는 종종 기준물(예를 들면, 기준 게놈)의 상응하는 부위 및/또는 위치에 대한 국소 게놈 편향 추정에 따라 서열 리드에 대해 측정된다. 일부 실시양태에서, 국소 게놈 편향 추정치는 서열(예를 들면, 서열 리드, 기준 게놈의 서열)의 편향의 정량적 척도를 포함한다. 국소 게놈 편향 추정치는 적합한 방법 또는 수학적 프로세스에 의해 측정될 수 있다. 일부 실시양태에서, 국소 게놈 편향 추정치는 적합한 분포 및/또는 적합한 분포 함수(예를 들면, PDF)에 의해 측정된다. 일부 실시양태에서, 국소 게놈 편향 추정치는 PDF의 정량적 표시치를 포함한다. 일부 실시양태에서, 국소 게놈 편향 추정치(예를 들면, 확률 밀도 추정치(PDE), 커넬(kernel) 밀도 추정치)는 국소 편향 함량의 확률 밀도 함수(예를 들면, PDF, 예를 들면, 커넬 밀도 함수)에 의해 측정된다. 일부 실시양태에서, 밀도 추정치는 커넬 밀도 추정치를 포함한다. 국소 게놈 편향 추정치는 종종 분포의 평균치, 평균 또는 중간치로서 표현된다. 종종, 국소 게놈 편향 추정치는 적합한 분포의 합계 또는 적분(예를 들면, 곡선하면적(AUC))으로서 표현된다.

[0218] PDF(예를 들면, 커넬 밀도 함수, 예를 들면, 에파네크니코브 커넬 밀도 함수)는 종종 밴드폭 변수(예를 들면, 밴드폭)를 포함한다. 밴드폭 변수는 종종 PDF를 사용할 때 확률 밀도 추정치(PDE)가 유도되는 윈도우의 크기 및/또는 길이를 한정한다. PDE가 유도되는 윈도우는 종종 한정된 길이의 폴리뉴클레오티드를 포함한다. 일부 실시양태에서, PDE가 유도되는 윈도우는 부분이다. 부분(예를 들면, 부분 크기, 부분 길이)은 종종 밴드폭 변수에 따라 결정된다. 밴드폭 변수는 국소 게놈 편향 추정치를 측정하는 데에 사용되는 윈도우의 길이 또는 크기, 즉 국소 게놈 편향 추정치가 측정되는 폴리뉴클레오티드 분절(예를 들면, 뉴클레오티드 염기의 연속 분절)의 길이를 결정한다. 임의의 적합한 밴드폭을 사용하여 PDE(예를 들면, 리드 밀도, 국소 게놈 편향 추정치(예를 들면, GC 밀도))를 측정할 수 있고, 이러한 밴드폭의 비한정적 예는 약 5개 염기 내지 약 100,000개 염기, 약 5개 염기 내지 약 50,000개 염기, 약 5개 염기 내지 약 25,000개 염기, 약 5개 염기 내지 약 10,000개 염기, 약 5개 염기 내지 약 5,000개 염기, 약 5개 염기 내지 약 2,500개 염기, 약 5개 염기 내지 약 1,000개 염기, 약 5개 염기 내지 약 500개 염기, 약 5개 염기 내지 약 250개 염기, 약 20개 염기 내지 약 250개 염기 등의 밴드폭을 포함한다. 일부 실시양태에서, 약 400개 염기 이하, 약 350개 염기 이하, 약 300개 염기 이하, 약 250개 염기 이하, 약 225개 염기 이하, 약 200개 염기 이하, 약 175개 염기 이하, 약 150개 염기 이하, 약 125개 염기 이하, 약 100개 염기 이하, 약 75개 염기 이하, 약 50개 염기 이하 또는 약 25개 염기 이하의 밴드폭을 사용하여 국소 게놈 편향 추정치(예를 들면, GC 밀도)를 측정한다. 일부 실시양태에서, 주어진 대상체 및/또는 샘플에 대해 수득된 서열 리드의 평균치, 평균, 중간 또는 최대 리드 길이에 따라 결정된 밴드폭을 사용하여 국소 게놈 편향 추정치(예를 들면, GC 밀도)를 측정한다. 종종, 주어진 대상체 및/또는 샘플에 대해 수득된 서열 리드의 평균치, 평균, 중간 또는 최대 리드 길이와 거의 동등한 밴드폭을 사용하여 국소 게놈 편향 추정치(예를 들면, GC 밀도)를 측정한다. 일부 실시양태에서, 약 250개, 240개, 230개, 220개, 210개, 200개, 190개, 180개, 160

개, 150개, 140개, 130개, 120개, 110개, 100개, 90개, 80개, 70개, 60개, 50개, 40개, 30개, 20개 또는 약 10개 염기의 밴드폭을 사용하여 국소 게놈 편향 추정치(예를 들면, GC 밀도)를 측정한다.

[0219] 국소 게놈 편향 추정치(예를 들면, 국소 GC 함량)가 보다 더 낮은 해상도에서 측정될 수 있지만, 국소 게놈 편향 추정치는 단일 염기 해상도에서 측정될 수 있다. 일부 실시양태에서, 국소 게놈 편향 추정치는 국소 편향 함량에 대해 측정된다. (예를 들면, PDF를 사용함으로써 측정된) 국소 게놈 편향 추정치는 종종 윈도우를 사용함으로써 측정된다. 일부 실시양태에서, 국소 게놈 편향 추정치는 미리 선택된 수의 염기를 포함하는 윈도우의 사용을 포함한다. 종종, 윈도우는 연속 염기의 분절을 포함한다. 종종, 윈도우는 비-연속 염기의 하나 이상의 부분을 포함한다. 종종, 윈도우는 하나 이상의 부분(예를 들면, 게놈의 부분)을 포함한다. 윈도우 크기 또는 길이는 종종 밴드폭 및 PDF에 따라 측정된다. 일부 실시양태에서, 윈도우는 밴드폭의 약 10배 이상, 8배 이상, 7배 이상, 6배 이상, 5배 이상, 4배 이상, 3배 이상 또는 2배 이상의 길이를 갖는다. PDF(예를 들면, 커넬 밀도 함수)가 밀도 추정치를 측정하는 데에 사용될 때 윈도우는 종종 선택된 밴드폭의 2배 길이를 갖는다. 윈도우는 임의의 적합한 수의 염기를 포함할 수 있다. 일부 실시양태에서, 윈도우는 약 5개 염기 내지 약 100,000개 염기, 약 5개 염기 내지 약 50,000개 염기, 약 5개 염기 내지 약 25,000개 염기, 약 5개 염기 내지 약 10,000개 염기, 약 5개 염기 내지 약 5,000개 염기, 약 5개 염기 내지 약 2,500개 염기, 약 5개 염기 내지 약 1,000개 염기, 약 5개 염기 내지 약 500개 염기, 약 5개 염기 내지 약 250개 염기, 또는 약 20개 염기 내지 약 250개 염기를 포함한다. 일부 실시양태에서, 게놈 또는 이의 분절은 복수의 윈도우로 분할된다. 게놈의 영역을 포괄하는 윈도우는 중첩될 수 있거나 중첩되지 않을 수 있다. 일부 실시양태에서, 윈도우는 서로로부터 동등한 거리를 두고 위치한다. 일부 실시양태에서, 윈도우는 서로로부터 상이한 거리를 두고 위치한다. 일부 실시양태에서, 게놈 또는 이의 분절은 복수의 슬라이딩 윈도우로 분할되고, 이때 윈도우는 게놈 또는 이의 분절 전체에 걸쳐 증가하면서 슬라이딩되고, 이때 증가할 때마다 각각의 윈도우가 국소 게놈 편향 추정치(예를 들면, 국소 GC 밀도)를 포함한다. 윈도우는 임의의 수치적 패턴에 따라 또는 임의의 비-주제적(athematic) 정해진 서열에 따라 게놈 전체에 걸쳐 적절하게 증가하면서 슬라이딩될 수 있다. 일부 실시양태에서, 국소 게놈 편향 추정치 측정을 위해, 윈도우는 약 10,000 bp 이상, 약 5,000 bp 이상, 약 2,500 bp 이상, 약 1,000 bp 이상, 약 750 bp 이상, 약 500 bp 이상, 약 400개 염기 이상, 약 250 bp 이상, 약 100 bp 이상, 약 50 bp 이상 또는 약 25 bp 이상씩 증가하면서 게놈 또는 그의 분절 전체에 걸쳐 슬리이딩된다. 일부 실시양태에서, 국소 게놈 편향 추정치 측정을 위해, 윈도우는 약 25 bp, 24 bp, 23 bp, 22 bp, 21 bp, 20 bp, 19 bp, 18 bp, 17 bp, 16 bp, 15 bp, 14 bp, 13 bp, 12 bp, 11 bp, 10 bp, 9 bp, 8 bp, 7 bp, 6 bp, 5 bp, 4 bp, 3 bp, 2 bp 또는 약 1 bp씩 증가하면서 게놈 또는 그의 분절 전체에 걸쳐 슬리이딩된다. 예를 들면, 국소 게놈 편향 추정치 측정을 위해, 윈도우는 약 400 bp(예를 들면, 200 bp의 밴드폭)를 포함할 수 있고 1 bp씩 증가하면서 게놈 전체에 걸쳐 슬라이딩될 수 있다. 일부 실시양태에서, 국소 게놈 편향 추정치는 커넬 밀도 함수 및 약 200 bp의 밴드폭을 사용함으로써 게놈 또는 그의 분절에서 각각의 염기에 대해 측정된다.

[0220] 일부 실시양태에서, 국소 게놈 편향 추정치는 국소 GC 함량 및/또는 국소 GC 함량의 표시치이다. 본원에서 사용된(예를 들면, 국소 편향, 국소 편향 추정치, 국소 편향 함량, 국소 게놈 편향, 국소 GC 함량 등을 기술하는 데에 사용된) 용어 "국소"는 10,000 bp 이하의 폴리뉴클레오티드 분절을 지칭한다. 일부 실시양태에서, 용어 "국소"는 5000 bp 이하, 4000 bp 이하, 3000 bp 이하, 2000 bp 이하, 1000 bp 이하, 500 bp 이하, 250 bp 이하, 200 bp 이하, 175 bp 이하, 150 bp 이하, 100 bp 이하, 75 bp 이하 또는 50 bp 이하의 폴리뉴클레오티드 분절을 지칭한다. 국소 GC 함량은 종종 게놈의 국소 분절, 서열 리드 또는 서열 리드 조립체(예를 들면, 콘티그(contig), 프로파일 등)에 대한 GC 함량의 표시치(예를 들면, 수학적 또는 정량적 표시치)이다. 예를 들면, 국소 GC 함량은 국소 GC 편향 추정치 또는 GC 밀도일 수 있다.

[0221] 하나 이상의 GC 밀도는 종종 기준물 또는 샘플(예를 들면, 검사 샘플)의 폴리뉴클레오티드에 대해 측정된다. 일부 실시양태에서, GC 밀도는 (예를 들면, 5,000 bp 이하의 폴리뉴클레오티드 분절에 대한) 국소 GC 함량의 표시치(예를 들면, 수학적 또는 정량적 표시치)이다. 일부 실시양태에서, GC 밀도는 국소 게놈 편향 추정치이다. 본원에 기재되어 있고/있거나 당분야에서 공지되어 있는 적합한 프로세스를 이용하여 GC 밀도를 측정할 수 있다. 적합한 PDF(예를 들면, 커넬 밀도 함수(예를 들면, 에파네크니코브 커넬 밀도 함수, 예를 들면, 도 33 참조))를 사용하여 GC 밀도를 측정할 수 있다. 일부 실시양태에서, GC 밀도는 PDE(예를 들면, 커넬 밀도 추정치)이다. 일부 실시양태에서, GC 밀도는 하나 이상의 구아닌(G) 및/또는 사이토신(C) 뉴클레오티드의 존재 또는 부재에 의해 정해진다. 역으로, 일부 실시양태에서, GC 밀도는 하나 이상의 아데닌(A) 및/또는 타이미딘(T) 뉴클레오티드의 존재 또는 부재에 의해 정해질 수 있다. 일부 실시양태에서, 국소 GC 함량에 대한 GC 밀도는 전체 게놈 또는 이의 분절(예를 들면, 상염색체, 염색체 세트, 단일 염색체, 유전자, 예를 들면, 도 34 참조)에 대해 측정된 GC 밀도에 따라 정규화된다. 하나 이상의 GC 밀도는 샘플(예를 들면, 검사 샘플) 또는 기준 샘플의 폴리뉴클레오티

드에 대해 측정될 수 있다. GC 밀도는 종종 기준 게놈에 대해 측정된다. 일부 실시양태에서, GC 밀도는 기준 게놈에 따른 서열 리드에 대해 측정된다. 리드의 GC 밀도는 종종 리드가 맵핑되는 기준 게놈의 상응하는 부위 및/또는 위치에 대해 측정된 GC 밀도에 따라 측정된다. 일부 실시양태에서, 기준 게놈 상의 위치에 대해 측정된 GC 밀도는 리드에 대해 배정되고/되거나 제공되고, 이때 리드 또는 이의 분절은 기준 게놈 상의 동일한 위치로 맵핑된다. 임의의 적합한 방법이 리드에 대한 GC 밀도를 생성할 목적으로 기준 게놈 상의 맵핑된 리드의 위치를 결정하는 데에 사용될 수 있다. 일부 실시양태에서, 맵핑된 리드의 중간 위치는 리드에 대한 GC 밀도가 측정되는 기준 게놈 상의 위치를 결정한다. 예를 들면, 리드의 중간 위치가 기준 게놈의 염기 번호 x에서 12번 염색체로 맵핑되는 경우, 상기 리드의 GC 밀도는 종종 기준 게놈의 염기 번호 x에서 또는 x 근처에서 12번 염색체 상에 위치하는 위치에 대한 커넬 밀도 추정치에 의해 결정된 GC 밀도로서 제공된다. 일부 실시양태에서, GC 밀도는 기준 게놈에 따라 리드의 일부 또는 전체 염기 위치에 대해 측정된다. 종종, 리드의 GC 밀도는 기준 게놈 상의 복수의 염기 위치에 대해 측정된 2개 이상의 GC 밀도의 평균, 합계, 중간치 또는 적분을 포함한다.

[0222] 일부 실시양태에서, 국소 게놈 편향 추정치(예를 들면, GC 밀도)는 정량되고/되거나 값으로서 제공된다. 국소 게놈 편향 추정치(예를 들면, GC 밀도)는 종종 평균치, 평균 및/또는 중간치로서 표현된다. 국소 게놈 편향 추정치(예를 들면, GC 밀도)는 종종 PDE의 최대 피크 높이로서 표현된다. 종종 국소 게놈 편향 추정치(예를 들면, GC 밀도)는 적합한 PDE의 합계 또는 적분(예를 들면, 곡선하면적(AUC))으로서 표현된다. 일부 실시양태에서, GC 밀도는 커넬 중량을 포함한다. 일부 실시양태에서, 리드의 GC 밀도는 커넬 중량의 평균치, 평균, 합계, 중간치, 최대 피크 높이 또는 적분과 거의 동등한 값을 포함한다.

[0223] 편향 빈도

[0224] 편향 빈도는 종종 하나 이상의 국소 게놈 편향 추정치(예를 들면, GC 밀도)에 따라 결정된다. 편향 빈도는 종종 샘플, 기준물(예를 들면, 기준 게놈, 기준 서열) 또는 이들의 부분에 대한 국소 게놈 편향 추정치의 발생 횟수의 카운트 또는 합계이다. 편향 빈도는 종종 샘플, 기준물 또는 이들의 부분에 대한 국소 게놈 편향 추정치(예를 들면, 각각의 국소 게놈 편향 추정치)의 발생 횟수의 카운트 또는 합계이다. 일부 실시양태에서, 편향 빈도는 GC 밀도 빈도이다. GC 밀도 빈도는 종종 하나 이상의 GC 밀도에 따라 결정된다. 예를 들면, GC 밀도 빈도는 값 x의 GC 밀도가 전체 게놈 또는 이의 분절 전체에 걸쳐 표시되는 횟수를 나타낼 수 있다. 편향 빈도는 종종 국소 게놈 편향 추정치의 분포이고, 이때 각각의 국소 게놈 편향 추정치의 발생 횟수는 편향 빈도로서 표시된다(예를 들면, 도 35 참조). 편향 빈도는 종종 수학적으로 조작되고/되거나 정규화된다. 편향 빈도는 적합한 방법에 의해 수학적으로 조작될 수 있고/있거나 정규화될 수 있다. 일부 실시양태에서, 편향 빈도는 샘플, 기준물 또는 이들의 부분(예를 들면, 상염색체, 염색체의 서브세트, 단일 염색체, 또는 이들의 리드)에 대한 각각의 국소 게놈 편향 추정치의 표시치(예를 들면, 분율, 백분율)에 따라 정규화된다. 편향 빈도는 샘플 또는 기준물의 일부 또는 모든 국소 게놈 편향 추정치에 대해 측정될 수 있다. 일부 실시양태에서, 편향 빈도는 검사 샘플의 일부 또는 모든 서열 리드에 대한 국소 게놈 편향 추정치에 대해 측정될 수 있다.

[0225] 일부 실시양태에서, 시스템은 편향 밀도 모듈(6)을 포함한다. 편향 밀도 모듈은 맵핑된 서열 리드(5) 및 기준 서열(2)을 임의의 적합한 포맷으로 수용할 수 있고/있거나, 회수할 수 있고/있거나 저장할 수 있고, 국소 게놈 편향 추정치, 국소 게놈 편향 분포, 편향 빈도, GC 밀도, GC 밀도 분포 및/또는 GC 밀도 빈도(총괄적으로 상자(7)로 표시됨)를 생성한다. 일부 실시양태에서, 편향 밀도 모듈은 데이터 및/또는 정보(예를 들면, 7)를 또 다른 적합한 모듈(예를 들면, 관계 모듈(8))에게 전달한다.

[0226] 관계

[0227] 일부 실시양태에서, 국소 게놈 편향 추정치와 편향 빈도 사이에 하나 이상의 관계가 생성된다. 본원에서 사용된 용어 "관계"는 2개 이상의 변수들 또는 값들 사이의 수학적 및/또는 도식적 관계를 지칭한다. 관계는 적합한 수학적 및/또는 도식적 프로세스에 의해 생성될 수 있다. 관계의 비한정적 예는 함수, 상관관계, 분포, 선형 또는 비선형 방정식, 선, 회귀, 피팅된 회귀 등 또는 이들의 조합의 수학적 및/또는 도식적 표시를 포함한다. 종종, 관계는 피팅된 관계를 포함한다. 일부 실시양태에서, 피팅된 관계는 피팅된 회귀를 포함한다. 종종, 관계는 가중된 2개 이상의 변수들 또는 값들을 포함한다. 일부 실시양태에서, 관계는 피팅된 회귀를 포함하고, 이때 관계의 하나 이상의 변수 또는 값이 가중된다. 종종, 회귀는 가중된 방식으로 피팅된다. 종종, 회귀는 가중 없이 피팅된다. 일부 실시양태에서, 관계의 생성은 작도 또는 도식화를 포함한다.

[0228] 일부 실시양태에서, 국소 게놈 편향 추정치와 편향 빈도 사이에 적합한 관계가 결정된다. 일부 실시양태에서, 샘플에 대한 (i) 국소 게놈 편향 추정치와 (ii) 편향 빈도 사이의 관계 생성은 샘플 편향 관계를 제공한다. 일부 실시양태에서, 기준물에 대한 (i) 국소 게놈 편향 추정치와 (ii) 편향 빈도 사이의 관계 생성은 기준 편향

관계를 제공한다. 일부 실시양태에서, GC 밀도와 GC 밀도 빈도 사이에 관계가 생성된다. 일부 실시양태에서, 샘플에 대한 (i) GC 밀도와 (ii) GC 밀도 빈도 사이의 관계 생성은 샘플 GC 밀도 관계를 제공한다. 일부 실시양태에서, 기준물에 대한 (i) GC 밀도와 (ii) GC 밀도 빈도 사이의 관계 생성은 기준 GC 밀도 관계를 제공한다. 일부 실시양태에서, 국소 게놈 편향 추정치가 GC 밀도인 경우, 샘플 편향 관계는 샘플 GC 밀도 관계이고, 기준 편향 관계는 기준 GC 밀도 관계이다. 기준 GC 밀도 관계 및/또는 샘플 GC 밀도 관계의 GC 밀도는 종종 국소 GC 함량의 표시치(예를 들면, 수학적 또는 정량적 표시치)이다. 일부 실시양태에서, 국소 게놈 편향 추정치와 편향 빈도 사이의 관계는 분포를 포함한다. 일부 실시양태에서, 국소 게놈 편향 추정치와 편향 빈도 사이의 관계는 피팅된 관계(예를 들면, 피팅된 회귀)를 포함한다. 일부 실시양태에서, 국소 게놈 편향 추정치와 편향 빈도 사이의 관계는 피팅된 선형 또는 비선형 회귀(예를 들면, 다항식 회귀)를 포함한다. 일부 실시양태에서, 국소 게놈 편향 추정치와 편향 빈도 사이의 관계는 가중된 관계를 포함하고, 이때 국소 게놈 편향 추정치 및/또는 편향 빈도는 적합한 프로세스에 의해 가중된다. 일부 실시양태에서, 가중된 피팅된 관계(예를 들면, 가중된 피팅)는 변위치 회귀, 파라미터화된 분포, 또는 내삽을 갖는 실험적 분포를 포함하는 프로세스에 의해 수득될 수 있다. 일부 실시양태에서, 검사 샘플, 기준물 또는 이들의 부분에 대한 국소 게놈 편향 추정치와 편향 빈도 사이의 관계는 다항식 회귀를 포함하고, 이때 국소 게놈 편향 추정치가 가중된다. 일부 실시양태에서, 가중된 피팅된 모델은 분포의 가중 값을 포함한다. 분포의 값은 적합한 프로세스에 의해 가중될 수 있다. 일부 실시양태에서, 분포의 중간치에 더 가까운 값보다 더 작은 가중을 갖는, 분포의 꼬리 근처에 위치하는 값이 제공된다. 예를 들면, 국소 게놈 편향 추정치(예를 들면, GC 밀도)와 편향 빈도(예를 들면, GC 밀도 빈도) 사이의 분포의 경우, 가중은 주어진 국소 게놈 편향 추정치에 대한 편향 빈도에 따라 결정되고, 이때 평균으로부터 더 먼 편향 빈도를 포함하는 국소 게놈 편향 추정치보다 더 큰 가중을 갖는, 분포의 평균에 더 가까운 편향 빈도를 포함하는 국소 게놈 편향 추정치가 제공된다.

[0229] 일부 실시양태에서, 시스템은 관계 모듈(8)을 포함한다. 관계 모듈은 관계뿐만 아니라 관계를 정의하는 함수, 계수, 상수 및 변수도 생성할 수 있다. 관계 모듈은 적합한 모듈(예를 들면, 편향 밀도 모듈(6))로부터 데이터 및/또는 정보(예를 들면, 7)를 수용할 수 있고/있거나, 저장할 수 있고/있거나 회수할 수 있고 관계를 생성할 수 있다. 관계 모듈은 종종 국소 게놈 편향 추정치의 분포를 생성하고 비교한다. 관계 모듈은 데이터 세트를 비교할 수 있고 종종 회귀 및/또는 피팅된 관계를 생성할 수 있다. 일부 실시양태에서, 관계 모듈은 하나 이상의 분포(예를 들면, 샘플 및/또는 기준물의 국소 게놈 편향 추정치의 분포)를 비교하고 서열 리드의 카운트에 대한 가중률 및/또는 가중 배정(9)을 또 다른 적합한 모듈(예를 들면, 편향 보정 모듈)에게 제공한다. 종종, 관계 모듈은 서열 리드의 정규화된 카운트를 분포 모듈(21)에게 직접적으로 제공하고, 이때 카운트는 관계 및/또는 비교에 따라 정규화된다.

[0230] 비교의 생성 및 이의 사용

[0231] 일부 실시양태에서, 서열 리드에서 국소 편향을 감소시키는 프로세스는 서열 리드의 카운트를 정규화하는 단계를 포함한다. 서열 리드의 카운트는 종종 시험 샘플과 기준물의 비교에 따라 정규화된다. 예를 들면, 종종 서열 리드의 카운트는 검사 샘플의 서열 리드의 국소 게놈 편향 추정치를 기준물(예를 들면, 기준 게놈 또는 이의 부분)의 국소 게놈 편향 추정치와 비교함으로써 정규화된다. 일부 실시양태에서, 서열 리드의 카운트는 검사 샘플의 국소 게놈 편향 추정치의 편향 빈도를 기준물의 국소 게놈 편향 추정치의 편향 빈도와 비교함으로써 정규화된다. 일부 실시양태에서, 서열 리드의 카운트는 샘플 편향 관계와 기준 편향 관계와 비교하여 비교를 생성함으로써 정규화된다.

[0232] 서열 리드의 카운트는 종종 2개 이상의 관계의 비교에 따라 정규화된다. 일부 실시양태에서, 2개 이상의 관계를 비교하여 서열 리드에서 국소 편향을 감소시키는 데에(예를 들면, 카운트를 정규화하는 데에) 사용되는 비교를 제공한다. 2개 이상의 관계는 적합한 방법에 의해 비교될 수 있다. 일부 실시양태에서, 비교는 제1 관계와 제2 관계의 덧셈, 뺄셈, 곱셈 및/또는 나눗셈을 포함한다. 일부 실시양태에서, 2개 이상의 관계의 비교는 적합한 선형 회귀 및/또는 비선형 회귀의 이용을 포함한다. 일부 실시양태에서, 2개 이상의 관계의 비교는 적합한 다항식 회귀(예를 들면, 3차 다항식 회귀)를 포함한다. 일부 실시양태에서, 비교는 제1 회귀와 제2 회귀의 덧셈, 뺄셈, 곱셈 및/또는 나눗셈을 포함한다. 일부 실시양태에서, 2개 이상의 관계는 다수의 회귀의 추론 골격을 포함하는 프로세스에 의해 비교된다. 일부 실시양태에서, 2개 이상의 관계는 적합한 다변량 분석을 포함하는 프로세스에 의해 비교된다. 일부 실시양태에서, 2개 이상의 관계는 기저 함수(예를 들면, 블렌딩 함수, 예를 들면, 다항식 기저, 푸리에 기저 등), 스플라인(spline), 방사형 기저(radial basis) 함수 및/또는 웨이블렛을 포함하는 프로세스에 의해 비교된다.

[0233] 일부 실시양태에서, 검사 샘플 및 기준물에 대한 편향 빈도를 포함하는 국소 게놈 편향 추정치의 분포는 다항식

회귀를 포함하는 프로세스에 의해 비교되고, 이때 국소 게놈 편향 추정치는 가중된다. 일부 실시양태에서, (i) 비(각각의 비는 기준물의 국소 게놈 편향 추정치의 편향 빈도 및 샘플의 국소 게놈 편향 추정치의 편향 빈도를 포함함)와 (ii) 국소 게놈 편향 추정치 사이에 다항식 회귀가 생성된다. 일부 실시양태에서, (i) 기준물의 국소 게놈 편향 추정치의 편향 빈도 대 샘플의 국소 게놈 편향 추정치의 편향 빈도의 비, 및 (ii) 국소 게놈 편향 추정치 사이에 다항식 회귀가 생성된다. 일부 실시양태에서, 검사 샘플 및 기준물의 리드에 대한 국소 게놈 편향 추정치의 분포의 비교는 기준물 및 샘플에 대한 국소 게놈 편향 추정치의 편향 빈도의 로그 비(예를 들면, log2 비)를 측정하는 단계를 포함한다. 일부 실시양태에서, 국소 게놈 편향 추정치의 분포의 비교는 기준물에 대한 국소 게놈 편향 추정치의 편향 빈도의 로그 비(예를 들면, log2 비)를 샘플에 대한 국소 게놈 편향 추정치의 편향 빈도의 로그 비(예를 들면, log2 비)로 나누는 단계를 포함한다(예를 들면, 실시예 7 및 도 36 참조).

[0234] 비교에 따른 카운트의 정규화는 전형적으로 일부 카운트를 조절하고 다른 카운트를 조절하지 않는다. 카운트의 정규화는 종종 모든 카운트를 조절하고 종종 서열 리드의 임의의 카운트를 조절하지 않는다. 서열 리드에 대한 카운트는 종종 가중률을 측정하는 단계를 포함하는 프로세스에 의해 정규화되고 종종 상기 프로세스는 가중률을 직접적으로 생성하고 사용하는 단계를 포함하지 않는다. 비교에 따른 카운트의 정규화는 종종 서열 리드의 각각의 카운트에 대한 가중률을 측정하는 단계를 포함한다. 가중률은 종종 서열 리드에 대한 특이성을 갖고 특이적 서열 리드의 카운트에 적용된다. 가중률은 종종 2개 이상의 편향 관계의 비교(예를 들면, 기준 편향 관계와 비교된 샘플 편향 관계)에 따라 측정된다. 정규화된 카운트는 종종 가중률에 따라 카운트 값을 조절함으로써 측정된다. 가중률에 따른 카운트의 조절은 종종 서열 리드에 대한 카운트와 가중률의 덧셈, 뺄셈, 곱셈 및/또는 나눗셈을 포함한다. 가중률 및/또는 정규화된 카운트는 종종 회귀(예를 들면, 회귀 선)로부터 측정된다. 정규화된 카운트는 종종 기준물(예를 들면, 기준 게놈) 및 검사 샘플의 국소 게놈 편향 추정치의 편향 빈도들 사이의 비교로부터 생성된 회귀 선(예를 들면, 피팅된 회귀 선)으로부터 직접적으로 수득된다. 일부 실시양태에서, 샘플의 리드의 각각의 카운트는 (ii) 기준물의 국소 게놈 편향 추정치의 편향 빈도와 비교된 (i) 리드의 국소 게놈 편향 추정치의 편향 빈도의 비교에 따라 정규화된 카운트 값으로서 제공된다. 일부 실시양태에서, 샘플에 대해 수득된 서열 리드의 카운트는 정규화되고, 서열 리드에서의 편향은 감소된다.

[0235] 종종, 시스템은 편향 보정 모듈(10)을 포함한다. 일부 실시양태에서, 편향 보정 모듈의 함수는 관계 모델링 모듈(8)에 의해 수행된다. 편향 보정 모듈은 적합한 모듈(예를 들면, 관계 모듈(8), 압축 모듈(4))로부터 맵핑된 리드 서열 및 가중률(예를 들면, 9)을 수용할 수 있고/있거나, 회수할 수 있고/있거나 저장할 수 있다. 일부 실시양태에서, 편향 보정 모듈은 맵핑된 리드에 대한 카운트를 제공한다. 일부 실시양태에서, 편향 보정 모듈은 가중 배정 및/또는 편향 보정 계수를 서열 리드의 카운트에 적용하여 정규화된 및/또는 조절된 카운트를 제공한다. 편향 보정 모듈은 종종 정규화된 카운트를 또 다른 적합한 모듈(예를 들면, 분포 모듈(21))에게 제공한다.

[0236] 일부 실시양태에서, 카운트의 정규화는 GC 밀도 이외의 하나 이상의 특징을 팩토링하는 단계, 및 서열 리드의 카운트를 정규화하는 단계를 포함한다. 일부 실시양태에서, 카운트의 정규화는 하나 이상의 상이한 국소 게놈 편향 추정치를 팩토링하는 단계 및 서열 리드의 카운트를 정규화하는 단계를 포함한다. 일부 실시양태에서, 서열 리드의 카운트는 하나 이상의 특징(예를 들면, 하나 이상의 편향)에 따라 결정된 가중에 따라 가중된다. 일부 실시양태에서, 카운트는 하나 이상의 조합된 가중에 따라 정규화된다. 종종, 하나 이상의 조합된 가중에 따라 하나 이상의 특징을 팩토링하고/하거나 카운트를 정규화하는 것은 다변량 모델의 사용을 포함하는 프로세스에 의해 수행된다. 임의의 적합한 다변량 모델이 카운트를 정규화하는 데에 사용될 수 있다. 다변령 모델의 비한정적 예는 다변량 선형 회귀, 다변량 변위치 회귀, 실험적 데이터의 다변량 내삽, 비선형 다변량 모델 등 또는 이들의 조합을 포함한다.

[0237] 일부 실시양태에서, 시스템은 다변량 보정 모듈(13)을 포함한다. 다변량 보정 모듈은 편향 밀도 모듈(6), 관계 모듈(8) 및/또는 편향 보정 모율(10)의 기능을 다회 수행하여 다수의 편향에 대해 카운트를 조절할 수 있다. 일부 실시양태에서, 다변량 보정 모듈은 하나 이상의 편향 밀도 모듈(6), 관계 모듈(8) 및/또는 편향 보정 모듈(10)을 포함한다. 종종, 다변량 보정 모듈은 정규화된 카운트(11)를 또 다른 적합한 모듈(예를 들면, 분포 모듈(21))에게 제공한다.

[0238] 가중된 부분

[0239] 일부 실시양태에서, 부분은 가중된다. 일부 실시양태에서, 하나 이상의 부분은 가중되어 가중된 부분을 제공한다. 부분의 가중은 종종 부분 의존성을 제거한다. 부분은 적합한 프로세스에 의해 가중될 수 있다. 일부 실시양태에서, 하나 이상의 부분은 고유 함수(예를 들면, 고유함수)에 의해 가중된다. 일부 실시양태에서, 고유 함수는 부분을 직교 고유 부분으로 대체하는 단계를 포함한다. 일부 실시양태에서, 시스템은 부분 가중 모듈(42)을

포함한다. 일부 실시양태에서, 가중 모듈은 리드 밀도, 리드 밀도 프로파일 및/또는 조절된 리드 밀도 프로파일을 수용하고/하거나, 회수하고/하거나 저장한다. 일부 실시양태에서, 가중된 부분은 부분 가중 모듈에 의해 제공된다. 일부 실시양태에서, 가중 모듈은 부분을 가중하는 데에 요구된다. 가중 모듈은 당분야에서 공지되어 있거나 본원에 기재되어 있는 하나 이상의 가중 방법으로 부분을 가중할 수 있다. 가중 모듈은 종종 가중된 부분을 또 다른 적합한 모듈(예를 들면, 점수화 모듈(46), PCA 통계 모듈(33), 프로파일 생성 모듈(26) 등)에게 제공한다.

[0240] 주성분 분석

[0241] 일부 실시양태에서, 리드 밀도 프로파일(예를 들면, 검사 샘플의 리드 밀도 프로파일(도 39a))은 주성분 분석(PCA)에 따라 조절된다. 하나 이상의 기준 샘플의 리드 밀도 프로파일 및/또는 검사 대상체의 리드 밀도 프로파일은 PCA에 따라 조절될 수 있다. PCA 관련된 프로세스로 리드 밀도 프로파일로부터 편향을 제거하는 것은 종종 본원에서 프로파일을 조절하는 것으로서 지칭된다. PCA는 적합한 PCA 방법 또는 이의 변법에 의해 수행될 수 있다. PCA 방법의 비한정적 예는 표준 상관관계 분석(CCA), 카루넨-루베(Karhunen-Loeve) 변환(KLT), 호텔링(Hotelling) 변환, 적절한 직교 분해(POD), X의 단일 값 분해(SVD), XTX의 고유값 분해(EVD), 인자 분석, 엑카르트-영(Eckart-Young) 정리, 슈미츠-미르스키(Schmidt-Mirsky) 정리, 실험적 직교 함수(EOF), 실험적 고유함수 분해, 실험적 성분 분석, 준고조차(quasiharmonic) 모드, 스펙트럼 분해, 실험적 모달 분석 등, 또는 이들의 변법 또는 조합을 포함한다. PCA는 종종 리드 밀도 프로파일에서 하나 이상의 편향을 확인한다. PCA에 의해 확인된 편향은 종종 본원에서 주성분으로서 지칭된다. 일부 실시양태에서, 하나 이상의 편향은 적합한 방법을 이용하여 리드 밀도 프로파일을 하나 이상의 주성분에 따라 조절함으로써 제거될 수 있다. 리드 밀도 프로파일은 하나 이상의 주성분과 리드 밀도 프로파일의 덧셈, 뺄셈, 곱셈 및/또는 나눗셈에 의해 조절될 수 있다. 일부 실시양태에서, 하나 이상의 편향은 하나 이상의 주성분을 리드 밀도 프로파일로부터 뺌으로써 리드 밀도 프로파일로부터 제거될 수 있다. 리드 밀도 프로파일에서의 편향이 종종 프로파일의 PCA에 의해 확인되고/되거나 정량되지만, 주성분을 종종 리드 밀도의 수준에서 프로파일로부터 뺀다. PCA는 종종 하나 이상의 주성분을 확인한다. 일부 실시양태에서, PCA는 제1, 제2, 제3, 제4, 제5, 제6, 제7, 제8, 제9 및 제10 이상의 주성분을 확인한다. 일부 실시양태에서, 1개, 2개, 3개, 4개, 5개, 6개, 7개, 8개, 9개 또는 10개 이상의 주성분이 프로파일을 조절하는 데에 사용된다. 종종, 주성분은 PCA에서의 출현 순서로 프로파일을 조절하는 데에 사용된다. 예를 들면, 3개의 주성분들을 리드 밀도 프로파일로부터 빼는 경우, 제1, 제2 및 제3 주성분이 사용된다. 종종, 주성분에 의해 확인된 편향은 프로파일을 조절하는 데에 사용되지 않는 프로파일의 특징을 포함한다. 예를 들면, PCA는 주성분으로서 유전적 변이(예를 들면, 이배수성, 미세중복, 미세결실, 결실, 전위, 삽입) 및/또는 성별 차이를 확인할 수 있다(예를 들면, 도 38c에서 관찰된 바와 같음). 따라서, 일부 실시양태에서, 하나 이상의 주성분은 프로파일을 조절하는 데에 사용되지 않는다. 예를 들면, 종종 제1, 제2 및 제4 주성분들이 프로파일을 조절하는 데에 사용되고, 이때 제3 주성분은 프로파일을 조절하는 데에 사용되지 않는다. 주성분은 임의의 적합한 샘플 또는 기준물을 사용함으로써 PCA로부터 수득될 수 있다. 일부 실시양태에서, 주성분은 검사 샘플(예를 들면, 검사 대상체)로부터 수득된다. 일부 실시양태에서, 주성분은 하나 이상의 기준물(예를 들면, 기준 샘플, 기준 서열, 기준 세트)로부터 수득된다. 예를 들면, 도 38a 내지 38c에 나타낸 바와 같이, PCA는 다수의 샘플들을 포함하는 트레이닝 세트(도 38a)로부터 수득된 중간 리드 밀도 프로파일에 대해 수행되어 제1 주성분(도 38b) 및 제2 주성분(도 38c)을 확인한다. 일부 실시양태에서, 주성분은 해당 유전적 변이를 갖지 않는 것으로 공지된 대상체의 세트로부터 수득된다. 일부 실시양태에서, 주성분은 공지된 정배수체의 세트로부터 수득된다. 주성분은 종종 기준물(예를 들면, 트레이닝 세트)의 하나 이상의 리드 밀도 프로파일을 사용함으로써 수행된 PCA에 따라 확인된다. 기준물로부터 수득된 하나 이상의 주성분을 종종 검사 대상체의 리드 밀도 프로파일(예를 들면, 도 39b)로부터 빼어 조절된 프로파일(예를 들면, 도 39c)을 제공한다.

[0242] 일부 실시양태에서, 시스템은 PCA 통계 모델(33)을 포함한다. PCA 통계 모듈은 또 다른 적합한 모듈(예를 들면, 프로파일 생성 모듈(26))로부터 리드 밀도 프로파일을 수용할 수 있고/있거나 회수할 수 있다. PCA는 종종 PCA 통계 모듈에 의해 수행된다. PCA 통계 모듈은 종종 리드 밀도 프로파일을 수용하고/하거나, 회수하고/하거나 저장하고 기준 세트(32), 트레이닝 세트(30) 및/또는 하나 이상의 검사 대상체(28)로부터의 리드 밀도 프로파일을 프로세싱한다. PCA 통계 모듈은 주성분을 생성할 수 있고/있거나 제공할 수 있고/있거나, 하나 이상의 주성분에 따라 리드 밀도 프로파일을 조절할 수 있다. 조절된 리드 밀도 프로파일(예를 들면, 40, 38)은 종종 PCA 통계 모듈에 의해 제공된다. PCA 통계 모듈은 조절된 리드 밀도 프로파일(예를 들면, 38, 40)을 또 다른 적합한 모듈(예를 들면, 부분 가중 모듈(42), 점수화 모듈(46))에게 제공할 수 있고/있거나 전달할 수 있다. 일부 실시양태에서, PCA 통계 모듈은 성별 판정(36)을 제공할 수 있다. 성별 판정은 종종 PCA 및/또는 하나 이상의 주성분에 따라 결정된 태아 성별의 확인이다. 일부 실시양태에서, PCA 통계 모듈은 하기 표시된 R 코드의 일부, 전부 또

는 변경물을 포함한다. 주성분을 산출하기 위한 R 코드는 일반적으로 데이터를 지우는(예를 들면, 중간치를 빼고 부분을 필터링하고 극한 값을 트리밍하는) 것으로 시작한다:

```
#Clean the data outliers for PCA
dclean <- (dat - m)[mask,]

for (j in 1:ncol(dclean))
{
q <- quantile(dclean[,j],c(.25,.75))
qmin <- q[1] - 4*(q[2]-q[1])
qmax <- q[2] + 4*(q[2]-q[1])
dclean[dclean[,j] < qmin,j] <- qmin
dclean[dclean[,j] > qmax,j] <- qmax
}
```

[0243]

[0244] 그 다음, 주성분이 산출된다:

```
#Compute principal components
pc <- prcomp(dclean)$x
```

[0245]

[0246] 마지막으로, 각각의 샘플의 PCA-조절된 프로파일이 다음과 같이 산출될 수 있다:

```
#Compute residuals
mm <- model.matrix(~pc[,1:numpc])
for (j in 1:ncol(dclean))
dclean[,j] <- dclean[,j] - predict(lm(dclean[,j]~mm))
```

[0247]

[0248] 프로파일의 비교

[0249] 일부 실시양태에서, 결과의 확인은 비교를 포함한다. 일부 실시양태에서, 리드 밀도 프로파일 또는 이의 부분은 결과를 제공하는 데에 사용된다. 일부 실시양태에서, 결과의 확인(예를 들면, 유전적 변이의 존재 또는 부재의 확인)은 2개 이상의 리드 밀도 프로파일들의 비교를 포함한다. 리드 밀도 프로파일들의 비교는 종종 게놈의 선택된 분절에 대해 생성된 리드 밀도 프로파일들의 비교를 포함한다. 예를 들면, 검사 프로파일은 기준 프로파일과 비교되고, 이때 종종 검사 프로파일 및 기준 프로파일은 실질적으로 동일한 분절인 게놈(예를 들면, 기준 게놈)의 분절에 대해 결정되었다. 리드 밀도 프로파일들의 비교는 종종 리드 밀도 프로파일의 2개 이상의 부분 서브세트들의 비교를 포함한다. 리드 밀도 프로파일의 부분 서브세트는 게놈의 분절(예를 들면, 염색체 또는 이의 분절)을 표시할 수 있다. 리드 밀도 프로파일은 임의의 양의 부분 서브세트를 포함할 수 있다. 종종, 리드 밀도 프로파일은 2개 이상, 3개 이상, 4개 이상 또는 5개 이상의 서브세트를 포함한다. 일부 실시양태에서, 리드 밀도 프로파일은 2개의 부분 서브세트를 포함하고, 이때 각각의 부분은 인접한 기준 게놈의 분절들을 표시한다. 일부 실시양태에서, 검사 프로파일은 기준 프로파일과 비교될 수 있고, 이때 검사 프로파일 및 기준 프로파일 둘 다가 제1 부분 서브세트 및 제2 부분 서브세트를 포함하고, 이때 제1 서브세트 및 제2 서브세트는 게놈의 상이한 분절들을 표시한다. 리드 밀도 프로파일의 일부 부분 서브세트들은 유전적 변이를 포함할 수 있고, 다른 부분 서브세트는 종종 유전적 변이를 실질적으로 갖지 않는다. 종종, 프로파일(예를 들면, 검사 프로파일)의 모든 부분 서브세트들은 유전적 변이를 실질적으로 갖지 않는다. 종종, 프로파일(예를 들면, 검사 프로파일)의 모든 부분 서브세트들은 유전적 변이를 포함한다. 일부 실시양태에서, 검사 프로파일은 유전적 변이를 포함하는 제1 부분 서브세트 및 유전적 변이를 실질적으로 갖지 않는 제2 부분 서브세트를 포함할 수 있다.

[0250] 일부 실시양태에서, 본원에 기재된 방법은 비교(예를 들면, 검사 프로파일과 기준 프로파일의 비교)를 수행하는 단계를 포함한다. 2개 이상의 데이터 세트, 2개 이상의 관계 및/또는 2개 이상의 프로파일이 적합한 방법에 의해 비교될 수 있다. 데이터 세트, 관계 및/또는 프로파일의 비교에 적합한 통계학적 방법의 비한정적 예는 베렌스-피셔 방법, 부트스트랩핑, 독립적인 유의성 검정을 조합하기 위한 피셔의 방법, 네이만-피어슨(Neyman-Pearson) 검정, 확인 데이터 분석, 탐구 데이터 분석, 정확 검정, F-검정, Z-검정, T-검정, 불확실성의 척도의 계산 및/또는 비교, 널 가설, 카운터널 등, chi-제곱 검정, 옴니버스 검정, 유의 수준(예를 들면, 통계학적 유의성)의 계산 및/또는 비교, 메타 분석, 다변량 분석, 회귀, 단순 선형 회귀, 로부스트 선형 회귀 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, 2개 이상의 데이터 세트, 관계 및/또는 프로파일의 비교는 불확실성의 척도를 측정하고/하거나 비교하는 단계를 포함한다. 본원에서 사용된 "불확실성의 척도"는 유의성(예를 들면,

통계학적 유의성)의 척도, 오차의 척도, 분산의 척도, 신뢰도의 척도 등 또는 이들의 조합을 지칭한다. 불확실성의 척도는 값(예를 들면, 역치) 또는 값의 범위(예를 들면, 구간, 신뢰 구간, 베이지안 신뢰 구간, 역치 범위)일 수 있다. 불확실성의 척도의 비한정적 예는 p-값, 편차(예를 들면, 표준 편차, 시그마, 절대 편차, 평균 절대 편차 등)의 적합한 척도, 오차(예를 들면, 표준 오차, 평균 제곱 오차, 평균 제곱근 오차 등)의 적합한 척도, 분산의 적합한 척도, 적합한 표준 점수(예를 들면, 표준 편차, 누적 백분율, 백분위수 당량, Z-점수, T-점수, R-점수, 표준 9등급(nine)(구간척도(stanine)), 구간척도에서의 % 등) 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, 유의 수준의 측정은 불확실성의 척도(예를 들면, p-값)의 측정을 포함한다. 일부 실시양태에서, 2개 이상의 데이터 세트, 관계 및/또는 프로파일이 다수(예를 들면, 2종 이상)의 통계학적 방법(예를 들면, 최소 자승 회귀, 주성분 분석, 선형 판별식 분석, 이차 판별식 분석, 배깅, 신경 네트워크, 지지체 벡터 머신 모델, 무작위 포레스트, 분류 트리 모델, K-최근접 이웃, 로지스틱 회귀 및/또는 손실 평활화), 및/또는 임의의 적합한 수학적 및/또는 통계학적 조작(예를 들면, 본원에서 조작으로서 지칭됨)을 이용함으로써 분석될 수 있고/있거나 비교될 수 있다.

[0251] 일부 실시양태에서, 2개 이상의 리드 밀도 프로파일들의 비교는 2개 이상의 리드 밀도 프로파일들에 대한 불확실성의 척도의 측정 및/또는 비교를 포함한다. 리드 밀도 프로파일 및/또는 관련된 불확실성의 척도는 종종 데이터 세트의 수학적 및/또는 통계학적 조작의 해석을 용이하게 하고/하거나 결과를 제공하기 위해 비교된다. 검사 대상체에 대해 생성된 리드 밀도 프로파일은 하나 이상의 기준물(예를 들면, 기준 샘플, 기준 대상체 등)에 대해 생성된 리드 밀도 프로파일과 비교된다. 일부 실시양태에서, 결과는 검사 대상체로부터의 리드 밀도 프로파일을 염색체, 또는 이의 부분 또는 분절에 대한 기준물로부터의 리드 밀도 프로파일과 비교함으로써 제공되고, 이때 기준 리드 밀도 프로파일은 유전적 변이를 보유하지 않는 것으로 공지된 기준 대상체(예를 들면, 기준물)의 세트로부터 수득된다. 일부 실시양태에서, 결과는 검사 대상체로부터의 리드 밀도 프로파일을 염색체, 또는 이의 부분 또는 분절에 대한 기준물로부터의 리드 밀도 프로파일과 비교함으로써 제공되고, 이때 기준 리드 밀도 프로파일은 특정 유전적 변이(예를 들면, 염색체 이배수성, 삼염색체성, 미세중복, 미세결실)를 보유하는 것으로 공지된 기준 대상체의 세트로부터 수득된다.

[0252] 일부 실시양태에서, 검사 대상체의 리드 밀도 프로파일은 유전적 변이의 부재를 표시하는 예정된 값과 비교되고, 종종 유전적 변이가 위치하는 게놈 위치에 상응하는 하나 이상의 게놈 위치(예를 들면, 부분)에서 예정된 값으로부터 벗어난다. 예를 들면, 검사 대상체(예를 들면, 유전적 변이와 관련된 의학적 질환에 대한 위험을 갖거나 이러한 의학적 질환을 앓고 있는 대상체)에서, 검사 대상체가 해당 유전적 변이를 포함할 때 리드 밀도 프로파일은 선택된 부분에 대해 기준물(예를 들면, 기준 서열, 기준 대상체, 기준 세트)의 리드 밀도 프로파일과 유의하게 상이할 것으로 예상된다. 검사 대상체가 해당 유전적 변이를 포함하지 않을 때 검사 대상체의 리드 밀도 프로파일은 종종 선택된 부분에 대해 기준물(예를 들면, 기준 서열, 기준 대상체, 기준 세트)의 리드 밀도 프로파일과 실질적으로 동일하다. 리드 밀도 프로파일은 종종 예정된 역치 및/또는 역치 범위와 비교된다(예를 들면, 도 40 참조). 본원에서 사용된 용어 "역치"는 자격을 갖춘 데이터 세트를 사용함으로써 계산되고 유전적 변이(예를 들면, 카피 수 변이, 이배수성, 염색체 이상, 미세중복, 미세결실 등)의 진단의 한계로서 작용하는 임의의 수를 지칭한다. 일부 실시양태에서, 본원에 기재된 방법에 의해 수득된 결과는 역치를 초과하고, 대상체는 유전적 변이(예를 들면, 삼염색체성)를 갖는 것으로서 진단된다. 일부 실시양태에서, 역치 값 또는 값 범위는 종종 (예를 들면, 기준물 및/또는 대상체로부터의) 서열 리드 데이터를 수학적으로 및/또는 통계학적으로 조작함으로써 계산된다. 유전적 변이의 존재 또는 부재를 표시하는 값의 예정된 역치 또는 역치 범위는 유전적 변이의 존재 또는 부재를 확인하는 데에 유용한 결과를 여전히 제공하면서 달라질 수 있다. 일부 실시양태에서, 정규화된 리드 밀도 및/또는 정규화된 카운트를 포함하는 리드 밀도 프로파일은 분류 및/또는 결과의 제공을 용이하게 하기 위해 생성된다. 결과는 정규화된 카운트를 포함하는 리드 밀도 프로파일의 도표를 기초로 (예를 들면, 이러한 리드 밀도 프로파일의 도표를 사용함으로써) 제공될 수 있다.

[0253] 일부 실시양태에서, 시스템은 점수화 모듈(46)을 포함한다. 점수화 모듈은 또 다른 적합한 모듈(예를 들면, 프로파일 생성 모듈(26), PCA 통계 모듈(33), 부분 가중 모듈(42) 등)로부터 리드 밀도 프로파일(예를 들면, 조절되고 정규화된 리드 밀도 프로파일)을 수용할 수 있고/있거나, 회수할 수 있고/있거나 저장할 수 있다. 점수화 모듈은 2개 이상의 리드 밀도 프로파일(예를 들면, 검사 프로파일, 기준 프로파일, 트레이닝 세트, 검사 대상체)을 수용할 수 있고/있거나, 회수할 수 있고/있거나, 저장할 수 있고/있거나 비교할 수 있다. 점수화 모듈은 종종 점수(예를 들면, 도표, 프로파일 통계치, 비교(예를 들면, 2개 이상의 프로파일들 사이의 차이), Z-점수, 불확실성의 척도, 판정 대역, 샘플 판정(50)(예를 들면, 유전적 변이의 존재 또는 부재의 확인) 및/또는 결과)를 제공할 수 있다. 점수화 모듈은 점수를 최종 사용자 및/또는 또 다른 적합한 모듈(예를 들면, 디스플레이, 프린터 등)에게 제공할 수 있다. 일부 실시양태에서, 점수화 모듈은 특정 검정(예를 들면, High-chr21 카운

트)을 위해 Chi-제곱 통계치를 산출하기 위한 R 함수를 포함하는 하기 표시된 R 코드의 일부, 전부 또는 변경물을 포함한다.

[0254] 3개의 파라미터들은 다음과 같다:

[0255] x = 샘플 리드 데이터(부분 x 샘플)

[0256] m = 부분에 대한 중간 값

[0257] y = 검정 벡터(예를 들면, chr21에 대한 진실을 제외한 모든 부분들에 대한 거짓)

```
getChisqP <- function(x,m,y)
{
ahigh <- apply(x[!y,],2,function(x) sum((x>m[!y])))
alow <- sum((!y))-ahigh
bhigh <- apply(x[y,],2,function(x) sum((x>m[y])))
blow <- sum(y)-bhigh
p <- sapply(1:length(ahigh), function(i) {
p <- chisq.test(matrix(c(ahigh[i],alow[i],bhigh[i],blow[i]),2))$p.value/2
if (ahigh[i]/alow[i] > bhigh[i]/blow[i]) p <- max(p,1-p)
else p <- min(p,1-p); p})
```

[0258] return(p)

[0259] 하이브리드 회귀 정규화

[0260] 일부 실시양태에서, 하이브리드 정규화 방법이 이용된다. 일부 실시양태에서, 하이브리드 정규화 방법은 편향(예를 들면, GC 편향)을 감소시킨다. 일부 실시양태에서, 하이브리드 정규화는 (i) 2개의 변수(예를 들면, 카운트 및 GC 함량)의 관계의 분석, 및 (ii) 이 분석에 따른 정규화 방법의 선택 및 적용을 포함한다. 일부 실시양태에서, 하이브리드 정규화는 (i) 회귀(예를 들면, 회귀 분석) 및 (ii) 이 회귀에 따른 정규화 방법의 선택 및 적용을 포함한다. 일부 실시양태에서, 제1 샘플(예를 들면, 제1 샘플 세트)에 대해 수득된 카운트는 또 다른 샘플(예를 들면, 제2 샘플 세트)로부터 수득된 카운트와 상이한 방법에 의해 정규화된다. 일부 실시양태에서, 제1 샘플(예를 들면, 제1 샘플 세트)에 대해 수득된 카운트는 제1 정규화 방법에 의해 정규화되고, 제2 샘플(예를 들면, 제2 샘플 세트)로부터 수득된 카운트는 제2 정규화 방법에 의해 정규화된다. 예를 들면, 일부 실시양태에서, 제1 정규화 방법은 선형 회귀의 사용을 포함하고, 제2 정규화 방법은 비선형 회귀(예를 들면, LOESS, GC-LOESS, LOWESS 회귀, LOESS 평활화)의 사용을 포함한다.

[0261] 일부 실시양태에서, 하이브리드 정규화 방법은 게놈 또는 염색체의 부분들에 맵핑된 서열 리드(예를 들면, 카운트, 맵핑된 카운트, 맵핑된 리드)를 정규화하는 데에 사용된다. 일부 실시양태에서, 미처리 카운트가 정규화되고, 일부 실시양태에서 조절된, 가중된, 필터링된 또는 미리 정규화된 카운트는 하이브리드 정규화 방법에 의해 정규화된다. 일부 실시양태에서, 게놈 구획 수준 또는 Z-점수가 정규화된다. 일부 실시양태에서, 게놈 또는 염색체의 선택된 부분들에 맵핑된 카운트는 하이브리드 정규화 방법에 의해 정규화된다. 카운트는 게놈의 부분들에 맵핑된 서열 리드의 적합한 척도를 지칭할 수 있고, 이 카운트의 비한정적 예는 미처리 카운트(예를 들면, 프로세싱되지 않은 카운트), 정규화된 카운트(예를 들면, PERUN, ChAI 또는 적합한 방법에 의해 정규화됨), 부분 수준(예를 들면, 평균치 수준, 평균 수준, 중간 수준 등), Z-점수 등 또는 이들의 조합을 포함한다. 카운트는 하나 이상의 샘플(예를 들면, 검사 샘플, 임신 여성으로부터의 샘플)로부터의 미처리 카운트 또는 프로세싱된 카운트일 수 있다. 일부 실시양태에서, 카운트는 하나 이상의 대상체로부터 수득된 하나 이상의 샘플로부터 수득된다.

[0262] 일부 실시양태에서, 정규화 방법(예를 들면, 정규화 방법의 종류)은 회귀(예를 들면, 회귀 분석) 및/또는 상관계수에 따라 선택된다. 회귀 분석은 변수들(예를 들면, 카운트 및 GC 함량) 사이의 관계를 평가하는 통계학적 기법을 지칭한다. 일부 실시양태에서, 회귀는 기준 게놈의 다수의 부분들의 각각의 부분들에 대한 카운트 및 GC 함량의 척도에 따라 생성된다. GC 함량의 적합한 척도가 사용될 수 있고, 이의 비한정적 예는 구아닌, 사이토신, 아데닌, 타이민, 푸린(GC), 또는 피리미딘(AT 또는 ATU) 함량의 척도, 용융 온도(T_m)(예를 들면, 변성 온도, 어닐링 온도, 혼성화 온도), 자유 에너지의 척도 등 또는 이들의 조합을 포함한다. 구아닌(G), 사이토신(C), 아데닌(A), 타이민(T), 푸린(GC) 또는 피리미딘(AT 또는 ATU) 함량의 척도는 비 또는 백분율로서 표현될

수 있다. 일부 실시양태에서, 임의의 적합한 비 또는 백분율이 사용되고, 이의 비한정적 예는 GC/AT, GC/총 뉴클레오티드, GC/A, GC/T, AT/총 뉴클레오티드, AT/GC, AT/G, AT/C, G/A, C/A, G/T, G/A, G/AT, C/T 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, GC 함량의 척도는 GC 대 총 뉴클레오티드 함량의 비 또는 백분율이다. 일부 실시양태에서, GC 함량의 척도는 기준 게놈의 부분들에 맵핑된 서열 리드에 대한 GC 대 총 뉴클레오티드 함량의 비 또는 백분율이다. 일부 실시양태에서, GC 함량은 기준 게놈의 각각의 부분들에 맵핑된 서열 리드에 따라 및/또는 이 서열 리드로부터 결정되고, 상기 서열 리드는 샘플(예를 들면, 임신 여성으로부터 수득된 샘플)로부터 수득된다. 일부 실시양태에서, GC 함량의 척도는 서열 리드에 따라 및/또는 서열 리드로부터 결정되지 않는다. 일부 실시양태에서, GC 함량의 척도는 하나 이상의 대상체로부터 수득된 하나 이상의 샘플에 대해 결정된다.

[0263] 일부 실시양태에서, 회귀의 생성은 회귀 분석 또는 상관관계 분석의 생성을 포함한다. 적합한 회귀가 사용될 수 있고, 이의 비한정적 예는 회귀 분석(예를 들면, 선형 회귀 분석), 적합도(goodness of fit) 분석, 피어슨의 상관관계 분석, 순위 상관관계, 설명되지 않는 분산의 분율, 나쉬-서트클리프(Nash-Sutcliffe) 모델 효율 분석, 회귀 모델 검증, 상실에서의 비례적 감소, 평균 제곱근 편차 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, 회귀 선이 생성된다. 일부 실시양태에서, 회귀의 생성은 선형 회귀의 생성을 포함한다. 일부 실시양태에서, 회귀의 생성은 비선형 회귀(예를 들면, LOESS 회귀, LOWESS 회귀)의 생성을 포함한다.

[0264] 일부 실시양태에서, 회귀는 예를 들면, 카운트와 GC 함량의 척도 사이의 상관관계(예를 들면, 선형 상관관계)의 존재 또는 부재를 확인한다. 일부 실시양태에서, 회귀(예를 들면, 선형 회귀)가 생성되고, 상관 계수가 결정된다. 일부 실시양태에서, 적합한 상관 계수가 결정되고, 이의 비한정적 예는 측정의 계수, R^2 값, 피어슨의 상관계수 등을 포함한다.

[0265] 일부 실시양태에서, 회귀(예를 들면, 회귀 분석, 선형 회귀)에 대한 적합도가 측정된다. 적합도는 종종 시각적 또는 수학적 분석에 의해 측정된다. 평가는 종종 적합도가 비선형 회귀의 경우 더 큰지 아니면 선형 회귀의 경우 더 큰지를 결정하는 것을 포함한다. 일부 실시양태에서, 상관 계수는 적합도의 척도이다. 일부 실시양태에서, 회귀에 대한 적합도의 평가는 상관 계수 및/또는 상관 계수 컷오프 값에 따라 결정된다. 일부 실시양태에서, 적합도의 평가는 상관 계수를 상관 계수 컷오프 값과 비교하는 단계를 포함한다. 일부 실시양태에서, 회귀에 대한 적합도의 평가는 선형 회귀를 표시한다. 예를 들면, 일부 실시양태에서, 적합도는 비선형 회귀보다 선형 회귀의 경우 더 크고, 적합도의 평가는 선형 회귀를 표시한다. 일부 실시양태에서, 평가는 선형 회귀를 표시하고, 선형 회귀는 카운트를 정규화하는 데에 사용된다. 일부 실시양태에서, 회귀에 대한 적합도의 평가는 비선형 회귀를 표시한다. 예를 들면, 일부 실시양태에서, 적합도는 선형 회귀보다 비선형 회귀의 경우 더 크고, 적합도의 평가는 비선형 회귀를 표시한다. 일부 실시양태에서, 평가는 비선형 회귀를 표시하고, 비선형 회귀는 카운트를 정규화하는 데에 사용된다.

[0266] 일부 실시양태에서, 상관 계수가 상관 계수 컷오프 이상일 때 적합도의 평가는 선형 회귀를 표시한다. 일부 실시양태에서, 상관 계수가 상관 계수 컷오프 미만일 때 적합도의 평가는 비선형 회귀를 표시한다. 일부 실시양태에서, 상관 계수 컷오프는 예정된다. 일부 실시양태에서, 상관 계수 컷오프는 약 0.5 이상, 약 0.55 이상, 약 0.6 이상, 약 0.65 이상, 약 0.7 이상, 약 0.75 이상, 약 0.8 이상 또는 약 0.85 이상이다.

[0267] 예를 들면, 일부 실시양태에서, 상관 계수가 약 0.6 이상일 때 선형 회귀를 포함하는 정규화 방법이 이용된다. 일부 실시양태에서, 상관 계수가 0.6의 상관 계수 컷오프 이상일 때 샘플의 카운트(예를 들면, 기준 게놈의 부분 당 카운트, 부분 당 카운트)는 선형 회귀에 따라 정규화되고, 그렇지 않으면 (예를 들면, 상기 계수가 0.6의 상관 계수 컷오프 미만일 때) 카운트는 비선형 회귀에 따라 정규화된다. 일부 실시양태에서, 정규화 프로세스는 기준 게놈의 다수의 부분들의 각각의 부분들에 대한 (i) 카운트 및 (ii) GC 함량에 대한 선형 회귀 또는 비선형 회귀를 생성하는 단계를 포함한다. 일부 실시양태에서, 상관 계수가 0.6의 상관 계수 컷오프 미만일 때 비선형 회귀(예를 들면, LOWESS, LOESS)를 포함하는 정규화 방법이 이용된다. 일부 실시양태에서, 상관 계수(예를 들면, 상관 계수)가 약 0.7 미만, 약 0.65 미만, 약 0.6 미만, 약 0.55 미만 또는 약 0.5 미만의 상관 계수 컷오프일 때 비선형 회귀(예를 들면, LOWESS)를 포함하는 정규화 방법이 이용된다. 예를 들면, 일부 실시양태에서, 상관 계수가 약 0.6의 상관 계수 컷오프 미만일 때 비선형 회귀(예를 들면, LOWESS, LOESS)를 포함하는 정규화 방법이 이용된다.

[0268] 일부 실시양태에서, 특정 종류의 회귀(예를 들면, 선형 또는 비선형 회귀)가 선택되고, 회귀가 생성된 후, 카운트는 회귀를 카운트로부터 뺌으로써 정규화된다. 일부 실시양태에서, 회귀를 카운트로부터 빼는 것은 감소된 편향(예를 들면, GC 편향)을 갖는 정규화된 카운트를 제공한다. 일부 실시양태에서, 선형 회귀를 카운트로부터 뺀

다. 일부 실시양태에서, 비선형 회귀(예를 들면, LOESS, GC-LOESS, LOWESS 회귀)를 카운트로부터 뺀다. 임의의 적합 방법이 회귀 선을 카운트로부터 빼는 데에 이용될 수 있다. 예를 들면, 카운트 x가 0.5의 GC 함량을 포함하는 부분 *i*(예를 들면, 부분 *i*)로부터 유도되고 회귀 선이 0.5의 GC 함량에서 카운트 y를 결정하는 경우, x-y는 부분 *i*에 대한 정규화된 카운트이다. 일부 실시양태에서, 카운트는 회귀를 빼기 전 및/또는 후에 정규화된다. 일부 실시양태에서, 하이브리드 정규화 방법에 의해 정규화된 카운트는 게놈 또는 이의 분절의 게놈 구획 수준, Z-점수, 수준 및/또는 프로파일을 생성하는 데에 사용된다. 일부 실시양태에서, 하이브리드 정규화 방법에 의해 정규화된 카운트는 (예를 들면, 태아에서) 유전적 변이의 존재 또는 부재를 확인하기 위해 본원에 기재된 방법에 의해 분석된다.

[0269] 일부 실시양태에서, 하이브리드 정규화 방법은 정규화 전 또는 후에 하나 이상의 부분을 필터링하거나 가중하는 단계를 포함한다. 본원에 기재된 부분(예를 들면, 기준 게놈의 부분)을 필터링하는 방법을 포함하는, 부분을 필터링하는 적합한 방법이 이용될 수 있다. 일부 실시양태에서, 부분(예를 들면, 기준 게놈의 부분)은 하이브리드 정규화 방법을 적용하기 전에 필터링된다. 일부 실시양태에서, 선택된 부분(예를 들면, 카운트 가변성에 따라 선택된 부분)에 맵핑된 서열결정 리드의 카운트만이 하이브리드 정규화에 의해 정규화된다. 일부 실시양태에서, 기준 게놈의 필터링된 부분(예를 들면, 카운트 가변성에 따라 필터링된 부분)에 맵핑된 서열결정 리드의 카운트는 하이브리드 정규화 방법을 이용하기 전에 제거된다. 일부 실시양태에서, 하이브리드 정규화 방법은 적합한 방법(예를 들면, 본원에 기재된 방법)에 따라 부분(예를 들면, 기준 게놈의 부분)을 선택하거나 필터링하는 단계를 포함한다. 일부 실시양태에서, 하이브리드 정규화 방법은 다수의 검사 샘플들의 각각의 부분들에 맵핑된 카운트에 대한 불확실성 값에 따라 부분(예를 들면, 기준 게놈의 부분)을 선택하거나 필터링하는 단계를 포함한다. 일부 실시양태에서, 하이브리드 정규화 방법은 카운트 가변성에 따라 부분(예를 들면, 기준 게놈의 부분)을 선택하거나 필터링하는 단계를 포함한다. 일부 실시양태에서, 하이브리드 정규화 방법은 GC 함량, 반복 요소, 반복 서열, 인트론, 엑손 등 또는 이들의 조합에 따라 부분(예를 들면, 기준 게놈의 부분)을 선택하거나 필터링하는 단계를 포함한다.

[0270] 예를 들면, 일부 실시양태에서, 다수의 임신 여성 대상체들로부터의 다수의 샘플들이 분석되고, 부분(예를 들면, 기준 게놈의 부분) 서브세트가 카운트 가변성에 따라 선택된다. 일부 실시양태에서, 선형 회귀가 임신 여성 대상체로부터 수득된 샘플에 대한 각각의 선택된 부분에 대한 (i) 카운트 및 (ii) GC 함량에 대한 상관 계수를 결정하는 데에 사용된다. 일부 실시양태에서, 예정된 상관 계수 컷오프 값(예를 들면, 약 0.6)보다 더 큰 상관 계수가 결정되고, 적합도의 평가가 선형 회귀를 표시하고, 카운트가 선형 회귀를 카운트로부터 뺌으로써 정규화된다. 일부 실시양태에서, 예정된 상관 계수 컷오프 값(예를 들면, 약 0.6)보다 더 작은 상관 계수가 결정되고, 적합도의 평가가 비선형 회귀를 표시하고, LOESS 회귀가 생성되고, 카운트가 LOESS 회귀를 카운트로부터 뺌으로써 정규화된다.

[0271] 프로파일

[0272] 일부 실시양태에서, 프로세싱 단계는 데이터 세트 또는 이의 유도체(예를 들면, 당분야에서 공지되어 있고/있거나 본원에 기재되어 있는 하나 이상의 수학적 및/또는 통계학적 데이터 프로세싱 단계의 생성물)의 다양한 양태들로부터 하나 이상의 프로파일(예를 들면, 프로파일 도표)을 생성하는 단계를 포함할 수 있다. 본원에서 사용된 용어 "프로파일"은 다량의 데이터에서 패턴 및/또는 상관관계의 확인을 용이하게 할 수 있는 데이터의 수학적 및/또는 통계학적 조작의 생성물을 지칭한다. "프로파일"은 종종 하나 이상의 기준을 기초로 데이터 또는 데이터 세트의 하나 이상의 조작으로부터 생성된 값을 포함한다. 프로파일은 종종 다수의 데이터 점들을 포함한다. 임의의 적합한 수의 데이터 점이 데이터 세트의 성질 및/또는 복잡성에 따라 프로파일에 포함될 수 있다. 일부 실시양태에서, 프로파일은 2개 이상의 데이터 점, 3개 이상의 데이터 점, 5개 이상의 데이터 점, 10개 이상의 데이터 점, 24개 이상의 데이터 점, 25개 이상의 데이터 점, 50개 이상의 데이터 점, 100개 이상의 데이터 점, 500개 이상의 데이터 점, 1000개 이상의 데이터 점, 5000개 이상의 데이터 점, 10,000개 이상의 데이터 점, 또는 100,000개 이상의 데이터 점을 포함할 수 있다.

[0273] 일부 실시양태에서, 프로파일은 데이터 세트의 전체를 표시하고, 일부 실시양태에서 프로파일은 데이터 세트의 부분 또는 서브세트를 표시한다. 즉, 프로파일은 종종 임의의 데이터를 제거하도록 필터링되지 않은 데이터를 표시하는 데이터 점을 포함하거나 이러한 데이터 점으로부터 생성되고, 종종 프로파일은 원하지 않는 데이터를 제거하도록 필터링된 데이터를 표시하는 데이터 점을 포함하거나 이러한 데이터 점으로부터 생성된다. 일부 실시양태에서, 프로파일에서의 데이터 점은 부분에 대한 데이터 조작의 결과를 표시한다. 일부 실시양태에서, 프로파일에서의 데이터 점은 부분의 군에 대한 데이터 조작의 결과를 포함한다. 일부 실시양태에서, 부분의 군은 서로 인접할 수 있고, 일부 실시양태에서 부분의 군은 염색체 또는 게놈의 상이한 부분들로부터의 군일 수

있다.

[0274] 데이터 세트로부터 유도된 프로파일에서의 데이터 점은 임의의 적합한 데이터 분류를 표시할 수 있다. 데이터가 분류되어 프로파일 데이터 점을 생성할 수 있는 카테고리의 비한정적 예는 크기에 기초한 부분, 서열 특징(예를 들면, GC 함량, AT 함량, 염색체 상의 위치(예를 들면, 짧은 아암(arm), 긴 아암, 센트로미어, 텔로미어) 등)에 기초한 부분, 발현 수준, 염색체 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, 프로파일은 또 다른 프로파일로부터 수득된 데이터 점으로부터 생성될 수 있다(예를 들면, 재정규화된 데이터 프로파일을 생성하도록 상이한 정규화 값으로 재정규화된 정규화된 데이터 프로파일). 일부 실시양태에서, 또 다른 프로파일로부터 수득된 데이터 점으로부터 생성된 프로파일은 데이터 점의 수 및/또는 데이터 세트의 복잡성을 감소시킨다. 데이터 점의 수 및/또는 데이터 세트의 복잡성의 감소는 종종 데이터의 해석을 용이하게 하고/하거나 결과의 제공을 용이하게 한다.

[0275] 프로파일(예를 들면, 게놈 프로파일, 염색체 프로파일, 염색체의 분절의 프로파일)은 종종 2개 이상의 부분들에 대한 정규화된 또는 비-정규화된 카운트의 집합체이다. 프로파일은 종종 하나 이상의 수준(예를 들면, 게놈 구획 수준)을 포함하고, 종종 2개 이상의 수준을 포함한다(예를 들면, 프로파일은 종종 다수의 수준을 갖는다). 수준은 일반적으로 거의 동일한 카운트 또는 정규화된 카운트를 갖는 부분 세트에 대한 수준이다. 일부 실시양태에서, 프로파일은 가중될 수 있거나, 제거될 수 있거나, 필터링될 수 있거나, 정규화될 수 있거나, 조절될 수 있거나, 평균으로서 산출될 수 있거나, 평균으로서 유도될 수 있거나, 더해질 수 있거나, 빼질 수 있거나, 프로세싱될 수 있거나 이들의 임의의 조합에 의해 변환될 수 있는 하나 이상의 부분을 포함한다. 프로파일은 종종 2개 이상의 수준을 정의하는 부분들에 맵핑된 정규화된 카운트를 포함하고, 이때 상기 카운트는 적합한 방법에 의해 상기 수준 중 하나에 따라 더 정규화된다. 종종, 프로파일의 카운트(예를 들면, 프로파일 수준)는 불확실성 값과 관련되어 있다.

[0276] 하나 이상의 수준을 포함하는 프로파일은 종종 패딩된다(예를 들면, 홀 패딩(hole padding)). 패딩(예를 들면, 홀 패딩)은 모체 미세결실 또는 모체 중복(예를 들면, 카피 수 변이)에 기인하는 프로파일 내의 수준을 확인하고 조절하는 프로세스를 지칭한다. 일부 실시양태에서, 태아 미세중복 또는 태아 미세결실에 기인하는 수준이 패딩된다. 일부 실시양태에서, 프로파일 내의 미세중복 또는 미세결실은 프로파일(예를 들면, 염색체의 프로파일)의 전체 수준을 인위적으로 상승시키거나 하강시켜 염색체 이배수성(예를 들면, 삼염색체성)의 거짓 양성 또는 거짓 음성 확인을 유발할 수 있다. 일부 실시양태에서, 미세중복 및/또는 결실에 기인하는 프로파일 내의 수준은 종종 패딩 또는 홀 패딩으로서 지칭되는 프로세스에 의해 확인되고 조절된다(예를 들면, 패딩되고/되거나 제거된다). 일부 실시양태에서, 프로파일은 프로파일 내의 제2 수준과 유의하게 상이한 하나 이상의 제1 수준을 포함하고, 하나 이상의 제1 수준 각각은 모체 카피 수 변이, 태아 카피 수 변이, 또는 모체 카피 수 변이 및 태아 카피 수 변이를 포함하고, 하나 이상의 제1 수준은 조절된다.

[0277] 하나 이상의 수준을 포함하는 프로파일은 제1 수준 및 제2 수준을 포함할 수 있다. 일부 실시양태에서, 제1 수준은 제2 수준과 상이(예를 들면, 유의하게 상이)하다. 일부 실시양태에서, 제1 수준은 제1 부분 세트를 포함하고, 제2 수준은 제2 부분 세트를 포함하고, 제1 부분 세트는 제2 부분 세트의 서브세트가 아니다. 일부 실시양태에서, 제1 부분 세트는 제2 부분 세트와 상이하고, 이들로부터 제1 수준 및 제2 수준이 측정된다. 일부 실시양태에서, 프로파일은 프로파일 내의 제2 수준과 상이한(예를 들면, 유의하게 상이한, 예를 들면, 유의하게 상이한 값을 갖는) 다수의 제1 수준을 가질 수 있다. 일부 실시양태에서, 프로파일은 프로파일 내의 제2 수준과 유의하게 상이한 하나 이상의 제1 수준을 포함하고, 하나 이상의 제1 수준은 조절된다. 일부 실시양태에서, 프로파일은 프로파일 내의 제2 수준과 유의하게 상이한 하나 이상의 제1 수준을 포함하고, 하나 이상의 제1 수준 각각은 모체 카피 수 변이, 태아 카피 수 변이, 또는 모체 카피 수 변이 및 태아 카피 수 변이를 포함하고, 하나 이상의 제1 수준은 조절된다. 일부 실시양태에서, 프로파일 내의 제1 수준은 프로파일로부터 제거되거나 조절된다(예를 들면, 패딩된다). 프로파일은 하나 이상의 제2 수준과 유의하게 상이한 하나 이상의 제1 수준을 포함하는 다수의 수준을 포함할 수 있고, 종종 프로파일 내의 대다수의 수준이 제2 수준이고, 이 제2 수준은 서로 거의 동등하다. 일부 실시양태에서, 프로파일 내의 수준 중 50% 초과, 60% 초과, 70% 초과, 80% 초과, 90% 초과 또는 95% 초과의 수준이 제2 수준이다.

[0278] 프로파일은 종종 도표로서 디스플레이된다. 예를 들면, 부분의 카운트(예를 들면, 정규화된 카운트)를 표시하는 하나 이상의 수준이 작도될 수 있고 가시화될 수 있다. 생성될 수 있는 프로파일 도표의 비한정적 예는 미처리 카운트(예를 들면, 미처리 카운트 프로파일 또는 미처리 프로파일), 정규화된 카운트, 가중된 부분, z-점수, p-값, 면적 비 대 피팅된 배수성, 중간 수준 대 피팅된 태아 분획과 측정된 태아 분획 사이의 비, 주성분 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, 프로파일 도표는 조작된 데이터의 가시화를 가능하게 한다. 일부

실시양태에서, 프로파일 도표는 결과(예를 들면, 면적 비 대 피팅된 배수성, 중간 수준 대 피팅된 태아 분획과 측정된 태아 분획 사이의 비, 주성분)를 제공하는 데에 이용될 수 있다. 본원에서 사용된 용어 "미처리 카운트 프로파일 도표" 또는 "미처리 프로파일 도표"는 영역(예를 들면, 게놈, 부분, 염색체, 기준 게놈의 염색체 부분 또는 염색체의 분절)에서의 총 카운트로 정규화된 영역 내의 각각의 부분에서의 카운트의 도표를 지칭한다. 일부 실시양태에서, 프로파일은 정적 윈도우 프로세스를 이용함으로써 생성될 수 있고, 일부 실시양태에서 프로파일은 슬라이딩 윈도우 프로세스를 이용함으로써 생성될 수 있다.

[0279] 검사 대상체에 대해 생성된 프로파일은 종종 데이터 세트의 수학적 및/또는 통계학적 조작의 해석을 용이하게 하고/하거나 결과를 제공하기 위해 하나 이상의 기준 대상체에 대해 생성된 프로파일과 비교된다. 일부 실시양태에서, 프로파일은 하나 이상의 출발 가정(예를 들면, 핵산의 모체 기여(예를 들면, 모체 분획), 핵산의 태아 기여(예를 들면, 태아 분획), 기준 샘플의 배수성 등 또는 이들의 조합)을 기초로 생성된다. 일부 실시양태에서, 검사 프로파일은 종종 유전적 변이의 부재를 표시하는 예정된 값 주변에 모이고, 검사 대상체가 유전적 변이를 보유하는 경우 종종 유전적 변이가 검사 대상체에서 위치하는 게놈 위치에 상응하는 영역에서 예정된 값으로부터 벗어난다. 유전적 변이와 관련된 의학적 질환에 대한 위험을 갖거나 이러한 의학적 질환을 앓고 있는 검사 대상체에서, 선택된 부분에 대한 수치 값은 영향을 받지 않은 게놈 위치에 대한 예정된 값과 유의하게 상이할 것으로 예상된다. 출발 가정(예를 들면, 고정된 배수성 또는 최적화된 배수성, 고정된 태아 분획 또는 최적화된 태아 분획, 또는 이들의 조합)에 따라, 유전적 변이의 존재 또는 부재를 확인하는 데에 유용한 결과를 여전히 제공하면서 유전적 변이의 존재 또는 부재를 표시하는 예정된 역치 또는 컷오프 값, 또는 역치 값 범위가 달라질 수 있다. 일부 실시양태에서, 프로파일은 표현형을 표시하고/하거나 나타낸다.

[0280] 비한정적 예로서, 정규화된 샘플 및/또는 기준 카운트 프로파일은 (a) 유전적 변이를 보유하지 않는 것으로 공지된 기준 세트로부터 선택된 염색체, 또는 이의 부분 또는 분절에 대한 기준 중간 카운트를 계산하고; (b) 비-정보제공 부분을 기준 샘플 미처리 카운트로부터 제거하고(예를 들면, 필터링); (c) 기준 게놈의 모든 남은 부분들에 대한 기준 카운트를 선택된 염색체 또는 선택된 게놈 위치에서 기준 샘플에 대한 카운트의 총 잔류 수(예를 들면, 기준 게놈의 비-정보제공 부분의 제거 후 남은 카운트의 합계)로 정규화하여 정규화된 기준 대상체 프로파일을 생성하고; (d) 상응하는 부분을 검사 대상체 샘플로부터 제거하고; (e) 하나 이상의 선택된 게놈 위치에 대한 남은 검사 대상체 카운트를, 선택된 게놈 위치를 함유하는 염색체 또는 염색체들에 대한 잔류 기준 중간 카운트의 합계로 정규화하여 정규화된 검사 대상체 프로파일을 생성함으로써 미처리 서열 리드 데이터로부터 수득될 수 있다. 일부 실시양태에서, (b)에서 필터링된 부분에 의해 감소된, 전체 게놈에 대한 추가 정규화 단계는 (c)와 (d) 사이에 포함될 수 있다.

[0281] 데이터 세트 프로파일은 카운팅된 맵핑된 서열 리드 데이터의 하나 이상의 조작에 의해 생성될 수 있다. 일부 실시양태는 하기 사항들을 포함한다. 서열 리드가 맵핑되고, 각각의 게놈 부분들로 맵핑되는 서열 태그의 수가 측정된다(예를 들면, 카운팅된다). 미처리 카운트 프로파일은 카운팅되는 맵핑된 서열 리드로부터 생성된다. 일부 실시양태에서, 결과는 검사 대상체로부터의 미처리 카운트 프로파일을, 유전적 변이를 보유하지 않는 것으로 공지된 기준 대상체 세트로부터의 염색체, 또는 이의 부분 또는 분절에 대한 기준 중간 카운트 프로파일과 비교함으로써 제공된다.

[0282] 일부 실시양태에서, 서열 리드 데이터는 임의적으로 무의미한 데이터 또는 비-정보제공 부분을 제거하도록 필터링된다. 필터링 후, 남은 카운트는 전형적으로 필터링된 데이터 세트를 생성하도록 합산된다. 일부 실시양태에서, 필터링된 카운트 프로파일은 필터링된 데이터 세트로부터 생성된다.

[0283] 서열 리드 데이터가 카운팅되고 임의적으로 필터링된 후, 데이터 세트는 수준 또는 프로파일을 생성하도록 정규화될 수 있다. 데이터 세트는 하나 이상의 선택된 부분을 적합한 정규화 기준 값으로 정규화함으로써 정규화될 수 있다. 일부 실시양태에서, 정규화 기준 값은 부분이 선택되는 염색체 또는 염색체들에 대한 총 카운트를 표시한다. 일부 실시양태에서, 정규화 기준 값은 유전적 변이를 보유하지 않는 것으로 공지된 기준 대상체 세트로부터 준비된 기준 데이터 세트로부터 하나 이상의 상응하는 부분, 염색체의 부분 또는 염색체를 표시한다. 일부 실시양태에서, 정규화 기준 값은 유전적 변이의 존재 또는 부재에 대해 분석되는 검사 대상체로부터 준비된 검사 대상체 데이터 세트로부터 하나 이상의 상응하는 부분, 염색체의 부분 또는 염색체를 표시한다. 일부 실시양태에서, 정규화 프로세스는 정적 윈도우 방법을 이용함으로써 수행되고, 일부 실시양태에서 정규화 프로세스는 이동 또는 슬라이딩 윈도우 방법을 이용함으로써 수행된다. 일부 실시양태에서, 정규화된 카운트를 포함하는 프로파일은 분류 및/또는 결과의 제공을 용이하게 하기 위해 생성된다. 결과는 정규화된 카운트를 포함하는 프로파일의 도표를 기초로(예를 들면, 이러한 프로파일의 도표를 사용함으로써) 제공될 수 있다.

[0284] 수준

[0285] 일부 실시양태에서, 값(예를 들면, 수, 정량적 값)은 수준에 기인한다. 수준은 적합한 방법, 연산 또는 수학적 프로세스에 의해 측정될 수 있다(예를 들면, 프로세싱된 수준). 수준은 종종 부분 세트에 대한 카운트(예를 들면, 정규화된 카운트)이거나 이러한 카운트로부터 유도된다. 일부 실시양태에서, 부분의 수준은 부분들에 맵핑된 카운트(예를 들면, 카운트, 정규화된 카운트)의 총수와 실질적으로 동등하다. 종종, 수준은 당분야에서 공지된 적합한 방법, 연산 또는 수학적 프로세스에 의해 프로세싱되거나, 변환되거나 조작된 카운트로부터 결정된다. 일부 실시양태에서, 수준은 프로세싱된 카운트로부터 유도되고, 프로세싱된 카운트의 비한정적 예는 가중된, 제거된, 필터링된, 정규화된, 조절된, 평균으로서 산출된, 평균(예를 들면, 평균 수준)으로서 유도된, 더해진, 빼진 또는 변환된 카운트 또는 이들의 조합을 포함한다. 일부 실시양태에서, 수준은 정규화된 카운트(예를 들면, 부분의 정규화된 카운트)를 포함한다. 수준은 적합한 프로세스에 의해 정규화된 카운트에 대한 수준일 수 있고, 이 프로세스의 비한정적 예는 부분별 정규화, GC 함량에 의한 정규화, 선형 및 비선형 최소 자승 회귀, GC LOESS, LOWESS, PERUN, ChAI, RM, GCRM, cQn 등 및/또는 이들의 조합을 포함한다. 수준은 정규화된 카운트 또는 상대적 양의 카운트를 포함할 수 있다. 일부 실시양태에서, 수준은 평균으로서 산출된 2개 이상의 부분에 대한 카운트 또는 정규화된 카운트에 대한 수준이고, 상기 수준은 평균 수준으로서 지칭된다. 일부 실시양태에서, 수준은 평균 수준으로서 지칭되는 평균 카운트 또는 정규화된 카운트의 평균을 갖는 부분의 세트에 대한 수준이다. 일부 실시양태에서, 수준은 미처리 카운트 및/또는 필터링된 카운트를 포함하는 부분에 대해 유도된다. 일부 실시양태에서, 수준은 미처리 상태의 카운트에 기초한다. 일부 실시양태에서, 수준은 불확실성 값(예를 들면, 표준 편차, MAD)과 관련되어 있다. 일부 실시양태에서, 수준은 Z-점수 또는 p-값에 의해 표시된다. 하나 이상의 부분에 대한 수준은 본원에서 "게놈 구획 수준"과 동의어이다.

[0286] 2개 이상의 수준(예를 들면, 프로파일 내의 2개 이상의 수준)에 대한 정규화된 또는 비-정규화된 카운트는 종종 수준에 따라 수학적으로 조작된다(예를 들면, 덧셈, 곱셈, 평균 산출, 정규화 등 또는 이들의 조합에 의해 조작된다). 예를 들면, 2개 이상의 수준에 대한 정규화된 또는 비-정규화된 카운트는 프로파일 내의 수준 중 하나, 일부 또는 전부에 따라 정규화될 수 있다. 일부 실시양태에서, 프로파일 내의 모든 수준의 정규화된 또는 비-정규화된 카운트는 프로파일 내의 1개 수준에 따라 정규화된다. 일부 실시양태에서, 프로파일 내의 제1 수준의 정규화된 또는 비-정규화된 카운트는 프로파일 내의 제2 수준의 정규화된 또는 비-정규화된 카운트에 따라 정규화된다.

[0287] 수준(예를 들면, 제1 수준, 제2 수준)의 비한정적 예는 프로세싱된 카운트를 포함하는 부분 세트에 대한 수준, 카운트의 평균, 중간치 또는 평균치를 포함하는 부분 세트에 대한 수준, 정규화된 카운트를 포함하는 부분 세트에 대한 수준 등 또는 이들의 임의의 조합이다. 일부 실시양태에서, 프로파일 내의 제1 수준 및 제2 수준은 동일한 염색체들로 맵핑된 부분의 카운트로부터 유도된다. 일부 실시양태에서, 프로파일 내의 제1 수준 및 제2 수준은 상이한 염색체들로 맵핑된 부분의 카운트로부터 유도된다.

[0288] 일부 실시양태에서, 수준은 하나 이상의 부분들에 맵핑된 정규화된 또는 비-정규화된 카운트로부터 결정된다. 일부 실시양태에서, 수준은 2개 이상의 부분들로 맵핑된 정규화된 또는 비-정규화된 카운트로부터 결정되고, 이 때 각각의 부분들에 대한 정규화된 카운트는 종종 거의 동일하다. 수준에 대한 부분 세트에서의 카운트(예를 들면, 정규화된 카운트)의 변경이 있을 수 있다. 수준에 대한 부분 세트에서, 이 세트의 다른 부분(예를 들면, 피크 및/또는 고랑(dips))에서의 카운트와 유의하게 상이한 카운트를 갖는 하나 이상의 부분이 존재할 수 있다. 임의의 적합한 수의 부분과 관련된 임의의 적합한 수의 정규화된 또는 비-정규화된 카운트는 수준을 정의할 수 있다.

[0289] 일부 실시양태에서, 하나 이상의 수준은 게놈의 부분들의 전부 또는 일부의 정규화된 또는 비-정규화된 카운트로부터 결정될 수 있다. 종종, 수준은 염색체 또는 이의 분절의 정규화된 또는 비-정규화된 카운트의 전부 또는 일부로부터 결정될 수 있다. 일부 실시양태에서, 2개 이상의 부분들(예를 들면, 부분 세트)로부터 유도된 2개 이상의 카운트가 수준을 결정한다. 일부 실시양태에서, 2개 이상의 카운트(예를 들면, 2개 이상의 부분으로부터의 카운트)가 수준을 결정한다. 일부 실시양태에서, 2개 내지 약 100,000개의 부분으로부터의 카운트가 수준을 결정한다. 일부 실시양태에서, 2개 내지 약 50,000개, 2개 내지 약 40,000개, 2개 내지 약 30,000개, 2개 내지 약 20,000개, 2개 내지 약 10,000개, 2개 내지 약 5,000개, 2개 내지 약 2,500개, 2개 내지 약 1,250개, 2개 내지 약 1,000개, 2개 내지 약 500개, 2개 내지 약 250개, 2개 내지 약 100개, 또는 2개 내지 약 60개의 부분으로부터의 카운트가 수준을 결정한다. 일부 실시양태에서, 약 10개 내지 약 50개의 부분으로부터의 카운트가 수준을 결정한다. 일부 실시양태에서, 약 20개 내지 약 40개 이상의 부분으로부터의 카운트가 수준을 결정한다. 일부 실시양태에서, 수준은 약 2개, 3개, 4개, 5개, 6개, 7개, 8개, 9개, 10개, 11개, 12개, 13개, 14개,

15개, 16개, 17개, 18개, 19개, 20개, 21개, 22개, 23개, 24개, 25개, 26개, 27개, 28개, 29개, 30개, 31개, 32개, 33개, 34개, 35개, 36개, 37개, 38개, 39개, 40개, 45개, 50개, 55개 또는 60개 이상의 부분으로부터의 카운트를 포함한다. 일부 실시양태에서, 수준은 부분 세트(예를 들면, 기준 게놈의 부분 세트, 염색체의 부분 세트 또는 염색체 분절의 부분 세트)에 상응한다.

[0290] 일부 실시양태에서, 수준은 인접한 부분들의 정규화된 또는 비-정규화된 카운트에 대해 결정된다. 일부 실시양태에서, 인접한 부분들(예를 들면, 부분 세트)은 게놈의 인접한 분절, 또는 염색체 또는 유전자의 인접한 분절을 표시한다. 예를 들면, 2개 이상의 인접한 부분들은 이들을 한 줄로 병합함으로써 정렬될 때 각각의 부분보다 더 긴 DNA 서열의 서열 조립체를 표시할 수 있다. 예를 들면, 2개 이상의 인접한 부분들은 온전한 게놈, 염색체, 유전자, 인트론, 엑손 또는 이들의 분절을 표시할 수 있다. 일부 실시양태에서, 수준은 인접한 부분들 및/또는 비-인접한 부분들의 집합체(예를 들면, 세트)로부터 측정된다.

[0291] 상이한 수준

[0292] 일부 실시양태에서, 정규화된 카운트의 프로파일은 프로파일 내의 또 다른 수준(예를 들면, 제2 수준)과 유의하게 상이한 수준(예를 들면, 제1 수준)을 포함한다. 제1 수준은 제2 수준보다 더 높을 수 있거나 낮을 수 있다. 일부 실시양태에서, 제1 수준은 카피 수 변이(예를 들면, 모체 카피 수 변이, 태아 카피 수 변이, 또는 모체 카피 수 변이 및 태아 카피 수 변이)를 포함하는 하나 이상의 리드를 포함하는 부분 세트에 대한 수준이고, 제2 수준은 카피 수 변이를 실질적으로 갖지 않는 리드를 포함하는 부분 세트에 대한 수준이다. 일부 실시양태에서, "유의하게 상이한"은 관찰가능한 차이를 지칭한다. 일부 실시양태에서, "유의하게 상이한"은 통계학적으로 상이한 또는 통계학적으로 유의한 차이를 지칭한다. 통계학적으로 유의한 차이는 종종 관찰된 차이의 통계학적 평가이다. 통계학적으로 유의한 차이는 당분야의 적합한 방법에 의해 평가될 수 있다. 임의의 적합한 역치 또는 범위를 사용하여 2개의 수준이 유의하게 상이한지를 확인할 수 있다. 일부 실시양태에서, 약 0.01% 이상(예를 들면, 수준 값들 중 하나 또는 어느 하나의 0.01%)만큼 상이한 2개의 수준(예를 들면, 평균 수준)은 유의하게 상이하다. 일부 실시양태에서, 약 0.1% 이상만큼 상이한 2개의 수준(예를 들면, 평균 수준)은 유의하게 상이하다. 일부 실시양태에서, 약 0.5% 이상만큼 상이한 2개의 수준(예를 들면, 평균 수준)은 유의하게 상이하다. 일부 실시양태에서, 약 0.5%, 0.75%, 1%, 1.5%, 2%, 2.5%, 3%, 3.5%, 4%, 4.5%, 5%, 5.5%, 6%, 6.5%, 7%, 7.5%, 8%, 8.5%, 9%, 9.5% 또는 약 10% 이상만큼 상이한 2개의 수준(예를 들면, 평균 수준)은 유의하게 상이하다. 일부 실시양태에서, 2개의 수준(예를 들면, 평균 수준)은 유의하게 상이하고, 어느 한 수준에서 중첩되지 않고/않거나 1개 또는 2개의 수준에 대해 계산된 불확실성 값에 의해 정의된 범위에서 중첩되지 않는다. 일부 실시양태에서, 불확실성 값은 시그마로서 표현된 표준 편차이다. 일부 실시양태에서, 2개의 수준(예를 들면, 평균 수준)은 유의하게 상이하고 불확실성 값의 약 1배 이상(예를 들면, 1 시그마)만큼 상이하다. 일부 실시양태에서, 2개의 수준(예를 들면, 평균 수준)은 유의하게 상이하고 불확실성 값의 약 2배 이상(예를 들면, 2 시그마), 불확실성 값의 약 3배 이상, 약 4배 이상, 약 5배 이상, 약 6배 이상, 약 7배 이상, 약 8배 이상, 약 9배 이상 또는 약 10배 이상만큼 상이하다. 일부 실시양태에서, 2개의 수준(예를 들면, 평균 수준)은 불확실성 값의 약 1.1배, 1.2배, 1.3배, 1.4배, 1.5배, 1.6배, 1.7배, 1.8배, 1.9배, 2.0배, 2.1배, 2.2배, 2.3배, 2.4배, 2.5배, 2.6배, 2.7배, 2.8배, 2.9배, 3.0배, 3.1배, 3.2배, 3.3배, 3.4배, 3.5배, 3.6배, 3.7배, 3.8배, 3.9배 또는 4.0배 이상만큼 상이할 때 유의하게 상이하다. 일부 실시양태에서, 2개의 수준 사이의 차이가 증가할 때 신뢰 수준이 증가한다. 일부 실시양태에서, 2개의 수준 사이의 차이가 감소할 때 및/또는 불확실성 값이 증가할 때 신뢰 수준이 감소한다. 예를 들면, 종종 신뢰 수준은 수준과 표준 편차(예를 들면, MAD) 사이의 차이의 비와 함께 증가한다.

[0293] 하나 이상의 예측 알고리즘이 서로 독립적으로 또는 의존적으로 가중될 수 있는 변수 조건 하에서 수집된 검출 데이터에 대한 유의성을 확인하거나 의미를 부여하는 데에 사용될 수 있다. 본원에서 사용된 용어 "변수"는 값 또는 값의 세트를 갖는 알고리즘의 인자, 양 또는 함수를 지칭한다.

[0294] 일부 실시양태에서, 제1 부분 세트는 종종 제2 부분 세트와 상이한(예를 들면, 중첩되지 않는) 부분을 포함한다. 예를 들면, 종종 정규화된 카운트의 제1 수준은 프로파일 내의 정규화된 카운트의 제2 수준과 유의하게 상이하고, 제1 수준은 제1 부분 세트에 대한 수준이고, 제2 수준은 제2 부분 세트에 대한 수준이고, 부분은 제1 부분 세트 및 제2 부분 세트에서 중첩되지 않는다. 일부 실시양태에서, 제1 부분 세트는 제2 부분 세트의 서브세트가 아니고, 이 세트들로부터 각각 제1 수준 및 제2 수준이 측정된다. 일부 실시양태에서, 제1 부분 세트는 제2 부분 세트와 상이하고/하거나 구별되고, 이 세트들로부터 각각 제1 수준 및 제2 수준이 측정된다.

[0295] 일부 실시양태에서, 제1 부분 세트는 프로파일 내의 제2 부분 세트의 서브세트이다. 예를 들면, 종종 프로파일

내의 제2 부분 세트에 대한 정규화된 카운트의 제2 수준은 프로파일 내의 제1 수준에 대한 제1 부분 세트의 정규화된 카운트를 포함하고, 제1 부분 세트는 프로파일 내의 제2 부분 세트의 서브세트이다. 일부 실시양태에서, 평균치, 평균 또는 중간 수준은 제2 수준으로부터 유도되고, 이때 제2 수준은 제1 수준을 포함한다. 일부 실시양태에서, 제2 수준은 전체 염색체를 표시하는 제2 부분 세트를 포함하고, 제1 수준은 제1 부분 세트를 포함하고, 이때 제1 세트는 제2 부분 세트의 서브세트이고, 제1 수준은 염색체에 존재하는 모체 카피 수 변이, 태아 카피 수 변이, 또는 모체 카피 수 변이 및 태아 카피 수 변이를 표시한다.

[0296] 일부 실시양태에서, 제2 수준의 값은 제1 수준보다 염색체 또는 이의 분절에 대한 카운트 프로파일의 평균, 평균치 또는 중간 수준에 더 가깝다. 일부 실시양태에서, 제2 수준은 염색체, 염색체의 부분 또는 이의 분절의 평균 수준이다. 일부 실시양태에서, 제1 수준은 염색체 또는 이의 분절을 표시하는 우세한 수준(예를 들면, 제2 수준)과 유의하게 상이하다. 프로파일은 제2 수준과 유의하게 상이한 다수의 제1 수준을 포함할 수 있고, 각각의 제1 수준은 독립적으로 제2 수준보다 더 높을 수 있거나 낮을 수 있다. 일부 실시양태에서, 제1 수준 및 제2 수준은 동일한 염색체로부터 유도되고, 제1 수준은 제2 수준보다 더 높거나 더 낮고, 제2 수준은 염색체의 우세한 수준이다. 일부 실시양태에서, 제1 수준 및 제2 수준은 동일한 염색체로부터 유도되고, 제1 수준은 카피 수 변이(예를 들면, 모체 및/또는 태아 카피 수 변이, 결실, 삽입, 중복)를 표시하고, 제2 수준은 염색체 또는 이의 분절에 대한 부분의 평균 수준 또는 우세한 수준이다.

[0297] 일부 실시양태에서, 제2 수준에 대한 제2 부분 세트에서의 리드는 유전적 변이(예를 들면, 카피 수 변이, 모체 및/또는 태아 카피 수 변이)를 실질적으로 포함하지 않는다. 종종, 제2 수준에 대한 제2 부분 세트는 일부 가변성(예를 들면, 수준의 가변성, 부분에 대한 카운트의 가변성)을 포함한다. 일부 실시양태에서, 실질적으로 존재하지 않는 카피 수 변이와 관련된 수준에 대한 부분 세트 내의 하나 이상의 부분은 모체 및/또는 태아 게놈에 존재하는 카피 수 변이를 갖는 하나 이상의 리드를 포함한다. 예를 들면, 종종 부분 세트는 염색체의 작은 분절(예를 들면, 10개 미만의 부분)에 존재하는 카피 수 변이를 포함하고, 상기 부분 세트는 실질적으로 존재하지 않는 카피 수 변이와 관련된 수준에 대한 부분 세트이다. 따라서, 카피 수 변이를 실질적으로 포함하지 않는 부분 세트는 수준의 약 10개, 9개, 8개, 7개, 6개, 5개, 4개, 3개, 2개 또는 1개 미만의 부분에 존재하는 카피 수 변이를 여전히 포함할 수 있다.

[0298] 일부 실시양태에서, 제1 수준은 제1 부분 세트에 대한 수준이고, 제2 수준은 제2 부분 세트에 대한 수준이고, 제1 부분 세트와 제2 부분 세트는 인접한다(예를 들면, 염색체 또는 이의 분절의 핵산 서열 면에서 인접한다). 일부 실시양태에서, 제1 부분 세트와 제2 부분 세트는 인접하지 않는다.

[0299] 태아 핵산과 모체 핵산의 혼합물로부터의 상대적으로 짧은 서열 리드를 사용하여 수준 및/또는 프로파일로 변환될 수 있는 카운트를 제공할 수 있다. 카운트, 수준 및 프로파일은 전자적 또는 유형의 형태로 도식화될 수 있고 가시화될 수 있다. (예를 들면, 수준 및/또는 프로파일로서 표시되는) 부분들에 맵핑된 카운트는 태아 및/또는 임신 여성에 존재하는 태아 및/또는 모체 게놈, 염색체, 또는 염색체의 부분 또는 분절의 가시적 표시치를 제공할 수 있다.

[0300] 기준 수준 및 정규화된 기준 값

[0301] 일부 실시양태에서, 프로파일은 기준 수준(예를 들면, 기준물로서 사용된 수준)을 포함한다. 종종, 정규화된 카운트의 프로파일은 기준 수준을 제공하고, 이 기준 수준으로부터 기대 수준 및 기대 범위가 결정된다(기대 수준 및 범위에 대해서는 하기 논의를 참조한다). 기준 수준은 종종 모체 및 태아 둘 다로부터의 맵핑된 리드를 포함하는 부분의 정규화된 카운트에 대한 수준이다. 기준 수준은 종종 태아 및 모체(예를 들면, 임신 여성)로부터의 맵핑된 리드의 정규화된 카운트의 합계이다. 일부 실시양태에서, 기준 수준은 정배수체 모체 및/또는 정배수체 태아로부터의 맵핑된 리드를 포함하는 부분에 대한 수준이다. 일부 실시양태에서, 기준 수준은 태아 및/또는 모체 유전적 변이(예를 들면, 이배수성(예를 들면, 삼염색체성), 카피 수 변이, 미세중복, 미세결실, 삽입)를 갖는 맵핑된 리드를 포함하는 부분에 대한 수준이다. 일부 실시양태에서, 기준 수준은 모체 및/또는 태아 유전적 변이(예를 들면, 이배수성(예를 들면, 삼염색체성), 카피 수 변이, 미세중복, 미세결실, 삽입)를 실질적으로 포함하지 않는 부분에 대한 수준이다. 일부 실시양태에서, 제2 수준은 기준 수준으로서 사용된다. 일부 실시양태에서, 프로파일은 정규화된 카운트의 제1 수준 및 정규화된 카운트의 제2 수준을 포함하고, 제1 수준은 제2 수준과 유의하게 상이하고, 제2 수준은 기준 수준이다. 일부 실시양태에서, 프로파일은 제1 부분 세트에 대한 정규화된 카운트의 제1 수준 및 제2 부분 세트에 대한 정규화된 카운트의 제2 수준을 포함하고, 제1 부분 세트는 모체 및/또는 태아 카피 수 변이를 갖는 맵핑된 리드를 포함하고, 제2 부분 세트는 모체 카피 수 변이 및/또는 태아 카피 수 변이를 실질적으로 갖지 않는 맵핑된 리드를 포함하고, 제2 수준은 기준 수준이다.

[0302] 일부 실시양태에서, 프로파일의 하나 이상의 수준에 대한 부분들에 맵핑된 카운트는 기준 수준의 카운트에 따라 정규화된다. 일부 실시양태에서, 기준 수준의 카운트에 따라 수준의 카운트를 정규화하는 것은 수준의 카운트를 기준 수준의 카운트, 또는 이의 배수 또는 분율로 나누는 것을 포함한다. 기준 수준의 카운트에 따라 정규화된 카운트는 종종 또 다른 프로세스(예를 들면, PERUN, ChAI)에 따라 정규화되고, 기준 수준의 카운트도 종종 (예를 들면, PERUN, ChAI에 의해) 정규화된다. 일부 실시양태에서, 수준의 카운트는 기준 수준의 카운트에 따라 정규화되고, 기준 수준의 카운트는 정규화 전 또는 후에 적합한 값까지 크기 조정될 수 있다. 기준 수준의 카운트를 크기 조정하는 프로세스는 임의의 적합한 상수(즉, 수)를 포함할 수 있고, 임의의 적합한 수학적 조작이 기준 수준의 카운트에 적용될 수 있다.

[0303] 정규화된 기준 값(NRV)은 종종 기준 수준의 정규화된 카운트에 따라 측정된다. NRV의 측정은 기준 수준의 카운트에 적용된 임의의 적합한 정규화 프로세스(예를 들면, 수학적 조작)을 포함할 수 있고, 이때 동일한 정규화 프로세스가 동일한 프로파일 내의 다른 수준의 카운트를 정규화하는 데에 사용된다. NRV의 측정은 종종 기준 수준을 그 자체로 나누는 것을 포함한다. NRV의 측정은 종종 기준 수준을 그 자체의 배수로 나누는 것을 포함한다. NRV의 측정은 종종 기준 수준을 기준 수준 및 상수(예를 들면, 임의의 수)의 합계 또는 차이로 나누는 것을 포함한다.

[0304] NRV는 종종 널(null) 값으로서 지칭된다. NRV는 임의의 적합한 값일 수 있다. 일부 실시양태에서, NRV는 0 이외의 임의의 값이다. 일부 실시양태에서, NRV는 정수이다. 일부 실시양태에서, NRV는 양의 정수이다. 일부 실시양태에서, NRV는 1, 10, 100 또는 1000이다. 종종, NRV는 1과 동등하다. 일부 실시양태에서, NRV는 0과 동등하다. 기준 수준의 카운트는 임의의 적합한 NRV로 정규화될 수 있다. 일부 실시양태에서, 기준 수준의 카운트는 0의 NRV로 정규화된다. 종종, 기준 수준의 카운트는 1의 NRV로 정규화된다.

[0305] 기대 수준

[0306] 기대 수준은 종종 미리 정해진 수준(예를 들면, 이론적 수준, 예측된 수준)이다. "기대 수준"은 종종 본원에서 "예정된 수준 값"으로서 지칭된다. 일부 실시양태에서, 기대 수준은 카피 수 변이를 포함하는 부분 세트에 대한 정규화된 카운트의 수준에 대한 예측된 값이다. 일부 실시양태에서, 기대 수준은 카피 수 변이를 실질적으로 포함하지 않는 부분 세트에 대해 결정된다. 기대 수준은 염색체 배수성(예를 들면, 0개, 1개, 2개(즉, 이배수체), 3개 또는 4개 염색체) 또는 미세배수성(동형접합 또는 이형접합 결실, 중복, 삽입 또는 이들의 부재)에 대해 결정될 수 있다. 종종, 기대 수준은 모체 미세배수성(예를 들면, 모체 및/또는 태아 카피 수 변이)에 대해 결정된다.

[0307] 유전적 변이 또는 카피 수 변이에 대한 기대 수준은 임의의 적합한 방법에 의해 결정될 수 있다. 종종, 기대 수준은 수준(예를 들면, 수준에 대한 부분 세트로 맵핑된 카운트)의 적합한 수학적 조작에 의해 결정된다. 일부 실시양태에서, 기대 수준은 종종 기대 수준 상수로서 지칭되는 상수를 사용함으로써 결정된다. 카피 수 변이에 대한 기대 수준은 종종 기준 수준, 기준 수준의 정규화된 카운트 또는 NRV를 기대 수준 상수와 곱하거나, 기대 수준 상수를 더하거나, 기대 수준 상수를 빼거나 기대 수준 상수로 나눔으로써, 또는 이들의 조합된 조작에 의해 계산된다. 종종, 동일한 대상체, 샘플 또는 검사 군에 대해 측정된 기대 수준(예를 들면, 모체 및/또는 태아 카피 수 변이의 기대 수준)은 동일한 기준 수준 또는 NRV에 따라 결정된다.

[0308] 종종, 기대 수준은 기준 수준, 기준 수준의 정규화된 카운트 또는 NRV를 기대 수준 상수와 곱함으로써 결정되고, 이때 기준 수준, 기준 수준의 정규화된 카운트 또는 NRV는 0과 동등하지 않다. 일부 실시양태에서, 기대 수준은 기대 수준 상수를 0과 동등한 기준 수준, 기준 수준의 정규화된 카운트 또는 NRV에 더함으로써 결정된다. 일부 실시양태에서, 기대 수준, 기준 수준의 정규화된 카운트, NRV 및 기대 수준 상수는 크기 조정가능하다. 크기조정하는 프로세스는 임의의 적합한 상수(즉, 수) 및 임의의 적합한 수학적 조작을 포함할 수 있고, 이때 동일한 크기조정 프로세스가 고려되는 모든 값들에게 적용된다.

[0309] 기대 수준 상수

[0310] 기대 수준 상수는 적합한 방법에 의해 결정될 수 있다. 일부 실시양태에서, 기대 수준 상수는 임의로 결정된다. 종종, 기대 수준 상수는 실험적으로 결정된다. 일부 실시양태에서, 기대 수준 상수는 수학적 조작에 따라 결정된다. 일부 실시양태에서, 기대 수준 상수는 기준물(예를 들면, 기준 게놈, 기준 샘플, 기준 검사 데이터)에 따라 결정된다. 일부 실시양태에서, 기대 수준 상수는 유전적 변이 또는 카피 수 변이(예를 들면, 중복, 삽입 또는 결실)의 존재 또는 부재를 표시하는 수준에 대해 예정된다. 일부 실시양태에서, 기대 수준 상수는 모체 카피 수 변이, 태아 카피 수 변이, 또는 모체 카피 수 변이 및 태아 카피 수 변이의 존재 또는 부재를 표시하는 수준

에 대해 예정된다. 카피 수 변이에 대한 기대 수준 상수는 임의의 적합한 상수 또는 상수의 세트일 수 있다.

[0311] 일부 실시양태에서, 동형접합 중복(예를 들면, 동형접합 중복)에 대한 기대 수준 상수는 약 1.6 내지 약 2.4, 약 1.7 내지 약 2.3, 약 1.8 내지 약 2.2, 또는 약 1.9 내지 약 2.1일 수 있다. 일부 실시양태에서, 동형접합 중복에 대한 기대 수준 상수는 약 1.6, 1.7, 1.8, 1.9, 2.0, 2.1, 2.2, 2.3 또는 약 2.4이다. 종종, 동형접합 중복에 대한 기대 수준 상수는 약 1.90, 1.92, 1.94, 1.96, 1.98, 2.0, 2.02, 2.04, 2.06, 2.08 또는 약 2.10이다. 종종, 동형접합 중복에 대한 기대 수준 상수는 약 2이다.

[0312] 일부 실시양태에서, 이형접합 중복(예를 들면, 동형접합 중복)에 대한 기대 수준 상수는 약 1.2 내지 약 1.8, 약 1.3 내지 약 1.7, 또는 약 1.4 내지 약 1.6이다. 일부 실시양태에서, 이형접합 중복에 대한 기대 수준 상수는 약 1.2, 1.3, 1.4, 1.5, 1.6, 1.7 또는 약 1.8이다. 종종, 이형접합 중복에 대한 기대 수준 상수는 약 1.40, 1.42, 1.44, 1.46, 1.48, 1.5, 1.52, 1.54, 1.56, 1.58 또는 약 1.60이다. 일부 실시양태에서, 이형접합 중복에 대한 기대 수준 상수는 약 1.5이다.

[0313] 일부 실시양태에서, 카피 수 변이의 부재(예를 들면, 모체 카피 수 변이 및/또는 태아 카피 수 변이의 부재)에 대한 기대 수준 상수는 약 1.3 내지 약 0.7, 약 1.2 내지 약 0.8, 또는 약 1.1 내지 약 0.9이다. 일부 실시양태에서, 카피 수 변이의 부재에 대한 기대 수준 상수는 약 1.3, 1.2, 1.1, 1.0, 0.9, 0.8 또는 약 0.7이다. 종종, 카피 수 변이의 부재에 대한 기대 수준 상수는 약 1.09, 1.08, 1.06, 1.04, 1.02, 1.0, 0.98, 0.96, 0.94 또는 약 0.92이다. 일부 실시양태에서, 카피 수 변이의 부재에 대한 기대 수준 상수는 약 1이다.

[0314] 일부 실시양태에서, 이형접합 결실(예를 들면, 모체, 태아, 또는 모체 및 태아 이형접합 결실)에 대한 기대 수준 상수는 약 0.2 내지 약 0.8, 약 0.3 내지 약 0.7, 또는 약 0.4 내지 약 0.6이다. 일부 실시양태에서, 이형접합 결실에 대한 기대 수준 상수는 약 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 또는 약 0.8이다. 종종 이형접합 결실에 대한 기대 수준 상수는 약 0.40, 0.42, 0.44, 0.46, 0.48, 0.5, 0.52, 0.54, 0.56, 0.58 또는 약 0.60이다. 일부 실시양태에서, 이형접합 결실에 대한 기대 수준 상수는 약 0.5이다.

[0315] 일부 실시양태에서, 동형접합 결실(예를 들면, 동형접합 결실)에 대한 기대 수준 상수는 약 -0.4 내지 약 0.4, 약 -0.3 내지 약 0.3, 약 -0.2 내지 약 0.2, 또는 약 -0.1 내지 약 0.1일 수 있다. 일부 실시양태에서, 동형접합 결실에 대한 기대 수준 상수는 약 -0.4, -0.3, -0.2, -0.1, 0.0, 0.1, 0.2, 0.3 또는 약 0.4이다. 종종, 동형접합 결실에 대한 기대 수준 상수는 약 -0.1, -0.08, -0.06, -0.04, -0.02, 0.0, 0.02, 0.04, 0.06, 0.08 또는 약 0.10이다. 종종, 동형접합 결실에 대한 기대 수준 상수는 약 0이다.

[0316] 기대 수준 범위

[0317] 일부 실시양태에서, 유전적 변이 또는 카피 수 변이(예를 들면, 모체 카피 수 변이, 태아 카피 수 변이, 또는 모체 카피 수 변이 및 태아 카피 수 변이)의 존재 또는 부재는 기대 수준 범위 내에 또는 밖에 속하는 수준에 의해 확인된다. 기대 수준 범위는 종종 기대 수준에 따라 결정된다. 일부 실시양태에서, 기대 수준 범위는 유전적 변이를 실질적으로 포함하지 않거나 카피 수 변이를 실질적으로 포함하지 않는 수준에 대해 결정된다. 적합한 방법이 기대 수준 범위를 결정하는 데에 이용될 수 있다.

[0318] 일부 실시양태에서, 기대 수준 범위는 수준에 대해 계산된 적합한 불확실성 값에 따라 정의된다. 불확실성 값의 비한정적 예는 표준 편차, 표준 오차, 계산된 분산, p-값 및 평균 절대 편차(MAD)이다. 일부 실시양태에서, 유전적 변이 또는 카피 수 변이에 대한 기대 수준 범위는 부분적으로 수준(예를 들면, 제1 수준, 제2 수준, 제1 수준 및 제2 수준)에 대한 불확실성 값을 계산함으로써 결정된다. 일부 실시양태에서, 기대 수준 범위는 프로파일(예를 들면, 염색체 또는 이의 분절에 대한 정규화된 카운트의 프로파일)에 대해 계산된 불확실성 값에 따라 한정된다. 일부 실시양태에서, 불확실성 값은 유전적 변이를 실질적으로 포함하지 않거나 카피 수 변이를 실질적으로 포함하지 않는 수준에 대해 계산된다. 일부 실시양태에서, 불확실성 값은 제1 수준, 제2 수준, 또는 제1 수준 및 제2 수준에 대해 계산된다. 일부 실시양태에서, 불확실성 값은 제1 수준, 제2 수준, 또는 제1 수준을 포함하는 제2 수준에 대해 결정된다.

[0319] 기대 수준 범위는 종종 부분적으로 불확실성 값을 상수(예를 들면, 예정된 상수) n과 곱하거나, n에 더하거나, n으로부터 빼거나 n으로 나눔으로써 계산된다. 적합한 수학적 절차 또는 절차들의 조합이 이용될 수 있다. 상수 n(예를 들면, 예정된 상수 n)은 종종 신뢰 구간으로서 지칭된다. 선택된 신뢰 구간은 선택된 상수 n에 따라 결정된다. 상수 n(예를 들면, 예정된 상수 n, 신뢰 구간)은 적합한 방식에 의해 결정될 수 있다. 상수 n은 0보다 더 큰 수 또는 0보다 더 큰 수의 분율일 수 있다. 상수 n은 정수일 수 있다. 종종, 상수 n은 10 미만의 수이다. 일부 실시양태에서, 상수 n은 약 10 미만, 약 9 미만, 약 8 미만, 약 7 미만, 약 6 미만, 약 5 미만, 약 4

미만, 약 3 미만 또는 약 2 미만의 수이다. 일부 실시양태에서, 상수 *n*은 약 10, 9.5, 9, 8.5, 8, 7.5, 7, 6.5, 6, 5.5, 5, 4.5, 4, 3.5, 3, 2.5, 2 또는 1이다. 상수 *n*은 공지된 유전 소인을 갖는 대상체(임신 여성 및/또는 태아)로부터 유도된 데이터로부터 실험적으로 결정될 수 있다.

[0320] 종종, 불확실성 값 및 상수 *n*은 범위(예를 들면, 불확실성 컷오프)를 한정한다. 예를 들면, 종종 불확실성 값은 표준 편차(예를 들면, +/- 5)이고 상수 *n*(예를 들면, 신뢰 구간)을 곱하여 범위 또는 불확실성 컷오프(예를 들면, 5*n* 내지 -5*n*)를 한정한다.

[0321] 일부 실시양태에서, 유전적 변이(예를 들면, 모체 카피 수 변이, 태아 카피 수 변이, 또는 모체 카피 수 변이 및 태아 카피 수 변이)에 대한 기대 수준 범위는 기대 수준 + 상수 *n* x 불확실성(예를 들면, *n* x 시그마(예를 들면, 6 시그마))의 합계이다. 일부 실시양태에서, *k*로 표시되는 유전적 변이 또는 카피 수 변이에 대한 기대 수준 범위는 하기 식에 의해 한정될 수 있다:

[0322] 식 R: (기대 수준 범위)$_k$ = (기대 수준)$_k$ + $n\sigma$

[0323] 상기 식에서, σ는 불확실성 값이고, *n*은 상수(예를 들면, 예정된 상수)이고, 기대 수준 범위 및 기대 수준은 유전적 변이 *k*(예를 들면, *k* = 이형접합 결실, 예를 들면, *k* = 유전적 변이의 부재)에 대한 것이다. 예를 들면, 기대 수준이 1과 동등하고(예를 들면, 카피 수 변이의 부재) 불확실성 값(즉, σ)이 +/- 0.05와 동등하고 *n*이 3인 경우, 기대 수준 범위는 1.15 내지 0.85로서 한정된다. 일부 실시양태에서, 이형접합 중복에 대한 기대 수준 범위는 이형접합 중복에 대한 기대 수준이 1.5이고 *n*이 3이고 불확실성 값 σ가 +/- 0.05일 때 1.65 내지 1.35로서 결정된다. 일부 실시양태에서, 이형접합 결실에 대한 기대 수준 범위는 이형접합 중복에 대한 기대 수준이 0.5이고 *n*이 3이고 불확실성 값 σ가 +/- 0.05일 때 0.65 내지 0.35로서 결정된다. 일부 실시양태에서, 동형접합 중복에 대한 기대 수준 범위는 이형접합 중복에 대한 기대 수준이 2.0이고 *n*이 3이고 불확실성 값 σ가 +/- 0.05일 때 2.15 내지 1.85로서 결정된다. 일부 실시양태에서, 동형접합 결실에 대한 기대 수준 범위는 이형접합 중복에 대한 기대 수준이 0.0이고 *n*이 3이고 불확실성 값 σ가 +/- 0.05일 때 0.15 내지 -0.15로서 결정된다.

[0324] 일부 실시양태에서, 동형접합 카피 수 변이(예를 들면, 모체, 태아, 또는 모체 및 태아 동형접합 카피 수 변이)에 대한 기대 수준 범위는 부분적으로 상응하는 이형접합 카피 수 변이에 대한 기대 수준 범위에 따라 결정된다. 예를 들면, 종종 동형접합 중복에 대한 기대 수준 범위는 이형접합 중복에 대한 기대 수준 범위의 상한보다 더 큰 모든 값들을 포함한다. 일부 실시양태에서, 동형접합 중복에 대한 기대 수준 범위는 이형접합 중복에 대한 기대 수준 범위의 상한 이상의 모든 값들을 포함한다. 일부 실시양태에서, 동형접합 중복에 대한 기대 수준 범위는 이형접합 중복에 대한 기대 수준 범위의 상한보다 더 크고 식 R에 의해 한정된 상한보다 더 작은 모든 값들을 포함하고, 이때 σ는 불확실성 값이고 양의 값이며, *n*은 상수이고, *k*는 동형접합 중복이다. 일부 실시양태에서, 동형접합 중복에 대한 기대 수준 범위는 이형접합 중복에 대한 기대 수준 범위의 상한 이상이고 식 R에 의해 한정된 상한 이하인 모든 값들을 포함하고, 이때 σ는 불확실성 값이고 σ는 양의 값이며, *n*은 상수이고, *k*는 동형접합 중복이다.

[0325] 일부 실시양태에서, 동형접합 결실에 대한 기대 수준 범위는 이형접합 결실에 대한 기대 수준 범위의 하한보다 더 작은 모든 값들을 포함한다. 일부 실시양태에서, 동형접합 결실에 대한 기대 수준 범위는 이형접합 결실에 대한 기대 수준 범위의 하한 이하인 모든 값들을 포함한다. 일부 실시양태에서, 동형접합 결실에 대한 기대 수준 범위는 이형접합 결실에 대한 기대 수준 범위의 하한보다 더 작고 식 R에 의해 한정된 하한보다 더 큰 모든 값들을 포함하고, 이때 σ는 불확실성 값이고 σ는 음의 값이며, *n*은 상수이고, *k*는 동형접합 결실이다. 일부 실시양태에서, 동형접합 결실에 대한 기대 수준 범위는 이형접합 결실에 대한 기대 수준 범위의 하한 이하이고 식 R에 의해 한정된 하한 이상인 모든 값들을 포함하고, 이때 σ는 불확실성 값이고 σ는 음의 값이며, *n*은 상수이고, *k*는 동형접합 결실이다.

[0326] 불확실성 값은 역치 값을 결정하는 데에 사용될 수 있다. 일부 실시양태에서, 범위(예를 들면, 역치 범위)는 미처리 카운트, 필터링된 카운트 및/또는 정규화된 카운트로부터 결정된 불확실성 값을 계산함으로써 수득된다. 일부 실시양태에서, 범위는 수준(예를 들면, 수준의 정규화된 카운트)에 대한 불확실성 값을, 컷오프 역치로서 선택된 불확실성의 배수(예를 들면, 표준 편차의 수)를 나타내는 예정된 상수(예를 들면, 1, 2, 3, 4, 5, 6 등)와 곱하여(예를 들면, 3 표준 편차의 경우 3과 곱하여) 범위를 생성함으로써 결정될 수 있다. 일부 실시양태에서, 범위는 값(예를 들면, 예정된 값, 불확실성 값, 예정된 상수와 곱해진 불확실성 값)을 수준에 더하고/더하거나 수준으로부터 빼어 범위를 생성함으로써 결정될 수 있다. 예를 들면, 수준이 1과 동등하고 표준 편차가 +/- 0.2이고 예정된 상수가 3인 경우, 범위는 (1 + 3(0.2)) 내지 (1 + 3(-0.2)), 또는 1.6 내지 0.4로서 계산될 수 있다. 범위는 종종 카피 수 변이에 대한 기대 범위 또는 기대 수준 범위를 한정할 수 있다. 일부 실시양

태에서, 역치 값을 초과하거나, 범위 밖에 속하거나 값의 범위 내에 속하는 부분들의 일부 또는 전부가 정규화 프로세스의 일부로서, 또는 정규화 프로세스의 전 또는 후에 제거된다. 일부 실시양태에서, 계산된 역치 값을 초과하거나, 범위 밖에 또는 범위 내에 속하는 부분들의 일부 또는 전부가 정규화 또는 분류 프로세스의 일부로서, 또는 정규화 또는 분류 프로세스 전에 가중되거나 조절된다. 가중의 예는 본원에 기재되어 있다. 본원에서 사용된 용어 "중복 데이터" 및 "중복 맵핑된 리드"는 게놈 위치(예를 들면, 염기 위치)로 이미 배정되어 있고/있거나 부분에 대해 카운팅되어 있는 것으로 확인되는, 샘플로부터 유도된 서열 리드를 지칭한다.

[0327] 일부 실시양태에서, 불확실성 값은 하기 식에 따라 결정된다:

[0328] $$Z = \frac{L_A - L_o}{\sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_o^2}{N_o}}}$$

[0329] 상기 식에서, Z는 2개의 수준 사이의 표준화된 편차를 표시하고, L은 평균(중간) 수준이고, 시그마는 표준 편차(또는 MAD)이다. 아래첨자 O은 프로파일의 분절(예를 들면, 제2 수준, 염색체, NRV, "정배수체 수준", 카피 수 변이 부재를 표시하는 수준)를 표시하고, A는 프로파일의 또 다른 분절(예를 들면, 제1 수준, 카피 수 변이를 표시하는 수준, 이배수성(예를 들면, 삼염색체성)을 표시하는 수준)을 표시한다. 변수 N_O는 아래첨자 O으로 표시된 프로파일의 분절 내의 부분의 총수를 표시한다. N_A는 아래첨자 A로 표시된 프로파일의 분절 내의 부분의 총수를 표시한다.

[0330] 카피 수 변이의 분류

[0331] 또 다른 수준(예를 들면, 제2 수준)과 유의하게 상이한 수준(예를 들면, 제1 수준)은 종종 기대 수준 범위에 따라 카피 수 변이(예를 들면, 모체 및/또는 태아 카피 수 변이, 태아 카피 수 변이, 결실, 중복, 삽입)로서 분류될 수 있다. 일부 실시양태에서, 카피 수 변이의 존재는 제1 수준이 제2 수준과 유의하게 상이하고 제1 수준이 카피 수 변이에 대한 기대 수준 범위 내에 속할 때 분류된다. 예를 들면, 카피 수 변이(예를 들면, 모체 및/또는 태아 카피 수 변이, 태아 카피 수 변이)는 제1 수준이 제2 수준과 유의하게 상이하고 제1 수준이 카피 수 변이에 대한 기대 수준 범위 내에 속할 때 분류될 수 있다. 일부 실시양태에서, 이형접합 중복(예를 들면, 모체 또는 태아, 또는 모체 및 태아 이형접합 중복) 또는 이형접합 결실(예를 들면, 모체 또는 태아, 또는 모체 및 태아 이형접합 결실)은 제1 수준이 제2 수준과 유의하게 상이하고 제1 수준이 각각 이형접합 중복 또는 이형접합 결실에 대한 기대 수준 범위 내에 속할 때 분류된다. 일부 실시양태에서, 동형접합 중복 또는 동형접합 결실은 제1 수준이 제2 수준과 유의하게 상이하고 제1 수준이 각각 동형접합 중복 또는 동형접합 결실에 대한 기대 수준 범위 내에 속할 때 분류된다.

[0332] 수준 조절

[0333] 일부 실시양태에서, 하나 이상의 수준은 조절된다. 수준을 조절하는 프로세스는 종종 패딩으로서 지칭된다. 일부 실시양태에서, 프로파일(예를 들면, 게놈의 프로파일, 염색체 프로파일, 염색체의 부분 또는 분절의 프로파일) 내의 다수의 수준은 조절된다. 일부 실시양태에서, 프로파일 내의 약 1개 내지 약 10,000개 이상의 수준은 조절된다. 일부 실시양태에서, 프로파일 내의 약 1개 내지 약 1,000개, 1개 내지 약 900개, 1개 내지 약 800개, 1개 내지 약 700개, 1개 내지 약 600개, 1개 내지 약 500개, 1개 내지 약 400개, 1개 내지 약 300개, 1개 내지 약 200개, 1개 내지 약 100개, 1개 내지 약 50개, 1개 내지 약 25개, 1개 내지 약 20개, 1개 내지 약 15개, 1개 내지 약 10개, 또는 1개 내지 약 5개의 수준은 조절된다. 일부 실시양태에서, 1개의 수준은 조절된다. 일부 실시양태에서, 제2 수준과 유의하게 상이한 수준(예를 들면, 정규화된 카운트 프로파일의 제1 수준)은 조절된다. 일부 실시양태에서, 카피 수 변이로서 분류된 수준은 조절된다. 일부 실시양태에서, 제2 수준과 유의하게 상이한 수준(예를 들면, 정규화된 카운트 프로파일의 제1 수준)은 카피 수 변이(예를 들면, 카피 수 변이, 예를 들면, 모체 카피 수 변이)로서 분류되고 조절된다. 일부 실시양태에서, 수준(예를 들면, 제1 수준)은 모체 카피 수 변이, 태아 카피 수 변이, 또는 모체 카피 수 변이 및 태아 카피 수 변이에 대한 기대 수준 범위 내에 있고, 상기 수준은 조절된다. 일부 실시양태에서, 하나 이상의 수준(예를 들면, 프로파일 내의 수준)은 조절되지 않는다. 일부 실시양태에서, 수준(예를 들면, 제1 수준)은 카피 수 변이에 대한 기대 수준 범위 밖에 있고, 상기 수준은 조절되지 않는다. 종종, 카피 수 변이의 부재에 대한 기대 수준 범위 내의 수준은 조절되지 않는다. 프로파일 내의 하나 이상의 수준에 대한 임의의 적합한 수의 조절을 수행할 수 있다. 일부 실시양태에서, 하나 이상의 수준은 조절된다. 일부 실시양태에서, 2개 이상, 3개 이상, 5개 이상, 6개 이상, 7개 이상, 8

개 이상, 9개 이상 또는 종종 10개 이상의 수준은 조절된다.

[0334] 일부 실시양태에서, 제1 수준의 값은 제2 수준의 값에 따라 조절된다. 일부 실시양태에서, 카피 수 변이를 표시하는 것으로서 확인된 제1 수준은 제2 수준의 값으로 조절되고, 이때 제2 수준은 종종 카피 수 변이의 부재와 관련되어 있다. 일부 실시양태에서, 카피 수 변이를 표시하는 것으로서 확인된 제1 수준의 값은 제1 수준의 값이 제2 수준의 값과 거의 동등하도록 조절된다.

[0335] 조절은 적합한 수학적 연산을 포함할 수 있다. 일부 실시양태에서, 조절은 하나 이상의 수학적 연산을 포함한다. 일부 실시양태에서, 수준은 정규화, 필터링, 평균 산출, 곱셈, 나눗셈, 덧셈 또는 뺄셈, 또는 이들의 조합에 의해 조절된다. 일부 실시양태에서, 수준은 예정된 값 또는 상수에 의해 조절된다. 일부 실시양태에서, 수준은 수준의 값을 또 다른 수준의 값으로 변경시킴으로써 조절된다. 예를 들면, 제1 수준은 그의 값을 제2 수준의 값으로 변경시킴으로써 조절될 수 있다. 이러한 경우 값은 프로세싱된 값(예를 들면, 평균, 정규화된 값 등)일 수 있다.

[0336] 일부 실시양태에서, 수준은 카피 수 변이(예를 들면, 모체 카피 수 변이)로서 분류되고 본원에서 예정된 조절 값(PAV)으로서 지칭되는 예정된 값에 따라 조절된다. 종종, PAV는 특정 카피 수 변이에 대해 결정된다. 종종, 특정 카피 수 변이(예를 들면, 동형접합 중복, 동형접합 결실, 이형접합 중복, 이형접합 결실)에 대해 결정된 PAV는 특정 카피 수 변이(예를 들면, 동형접합 중복, 동형접합 결실, 이형접합 중복, 이형접합 결실)로서 분류된 수준을 조절하는 데에 사용된다. 일부 실시양태에서, 수준은 카피 수 변이로서 분류된 후 분류된 카피 수 변이의 종류에 대한 특이성을 갖는 PAV에 따라 조절된다. 일부 실시양태에서, 수준(예를 들면, 제1 수준)은 모체 카피 수 변이, 태아 카피 수 변이, 또는 모체 카피 수 변이 및 태아 카피 수 변이로서 분류되고 PAV를 상기 수준에 더하거나 상기 수준으로부터 뺌으로써 조절된다. 종종, 수준(예를 들면, 제1 수준)은 모체 카피 수 변이로서 분류되고 PAV를 상기 수준에 더함으로써 조절된다. 예를 들면, 중복(예를 들면, 모체, 태아, 또는 모체 및 태아 동형접합 중복)으로서 분류된 수준은 특정 중복(예를 들면, 동형접합 중복)에 대해 결정된 PAV를 더하여 조절된 수준을 제공함으로써 조절될 수 있다. 종종, 카피 수 중복에 대해 결정된 PAV는 음의 값이다. 일부 실시양태에서, 중복에 대해 결정된 PAV를 사용함으로써 중복을 표시하는 수준으로의 조절을 제공하는 것은 상기 수준의 값을 감소시킨다. 일부 실시양태에서, 제2 수준과 유의하게 상이한 수준(예를 들면, 제1 수준)은 카피 수 결실(예를 들면, 동형접합 결실, 이형접합 결실, 동형접합 중복, 동형접합 중복)로서 분류되고, 제1 수준은 카피 수 결실에 대해 결정된 PAV를 더함으로써 조절된다. 종종, 카피 수 결실에 대해 결정된 PAV는 양의 값이다. 일부 실시양태에서, 결실에 대해 결정된 PAV를 사용함으로써 결실을 표시하는 수준으로의 조절을 제공하는 것은 상기 수준의 값을 증가시킨다.

[0337] PAV는 임의의 적합한 값일 수 있다. 종종, PAV는 카피 수 변이(예를 들면, 분류된 카피 수 변이)에 따라 결정되고 이러한 카피 수 변이에 대한 특이성을 갖는다. 일부 실시양태에서, PAV는 카피 수 변이(예를 들면, 분류된 카피 수 변이)에 대한 기대 수준 및/또는 PAV 계수에 따라 결정된다. PAV는 종종 기대 수준을 PAV 계수와 곱함으로써 결정된다. 예를 들면, 카피 수 변이에 대한 PAV는 카피 수 변이(예를 들면, 이형접합 결실)에 대해 결정된 기대 수준을 동일한 카피 수 변이(예를 들면, 이형접합 결실)에 대해 결정된 PAV 계수와 곱함으로써 결정될 수 있다. 예를 들면, 카피 수 변이 k(예를 들면, k = 이형접합 결실)에 대한 PAV는 하기 식에 의해 결정될 수 있다:

[0338] $PAV_k = (\text{기대 수준})_k \times (\text{PAV 계수})_k$

[0339] PAV 계수는 임의의 적합한 값일 수 있다. 일부 실시양태에서, 동형접합 중복에 대한 PAV 계수는 약 -0.6 내지 약 -0.4이다. 일부 실시양태에서, 동형접합 중복에 대한 PAV 계수는 약 -0.60, -0.59, -0.58, -0.57, -0.56, -0.55, -0.54, -0.53, -0.52, -0.51, -0.50, -0.49, -0.48, -0.47, -0.46, -0.45, -0.44, -0.43, -0.42, -0.41 또는 -0.40이다. 종종, 동형접합 중복에 대한 PAV 계수는 약 -0.5이다.

[0340] 예를 들면, NRV가 약 1이고 동형접합 중복에 대한 기대 수준이 약 2와 동등한 경우, 동형접합 중복에 대한 PAV는 상기 식에 따라 약 -1로서 결정된다. 이 경우, 동형접합 중복으로서 분류된 제1 수준은 예를 들면, 약 -1을 제1 수준의 값에 더함으로써 조절된다.

[0341] 일부 실시양태에서, 이형접합 중복에 대한 PAV 계수는 약 -0.4 내지 약 -0.2이다. 일부 실시양태에서, 이형접합 중복에 대한 PAV 계수는 약 -0.40, -0.39, -0.38, -0.37, -0.36, -0.35, -0.34, -0.33, -0.32, -0.31, -0.30, -0.29, -0.28, -0.27, -0.26, -0.25, -0.24, -0.23, -0.22, -0.21 또는 -0.20이다. 종종, 이형접합 중복에 대한 PAV 계수는 약 -0.33이다.

[0342] 예를 들면, NRV가 약 1이고 이형접합 중복의 기대 수준이 약 1.5와 동등한 경우, 이형접합 중복에 대한 PAV는 상기 식에 따라 약 -0.495로서 결정된다. 이 경우, 이형접합 중복으로서 분류된 제1 수준은 예를 들면, 약 -0.495를 제1 수준의 값에 더함으로써 조절된다.

[0343] 일부 실시양태에서, 이형접합 결실에 대한 PAV 계수는 약 0.4 내지 약 0.2이다. 일부 실시양태에서, 이형접합 결실에 대한 PAV 계수는 약 0.40, 0.39, 0.38, 0.37, 0.36, 0.35, 0.34, 0.33, 0.32, 0.31, 0.30, 0.29, 0.28, 0.27, 0.26, 0.25, 0.24, 0.23, 0.22, 0.21 또는 0.20이다. 종종, 이형접합 결실에 대한 PAV 계수는 약 0.33이다.

[0344] 예를 들면, NRV가 약 1이고 이형접합 결실에 대한 기대 수준이 약 0.5와 동등한 경우, 이형접합 결실에 대한 PAV는 상기 식에 따라 약 -0.495로서 결정된다. 이 경우, 이형접합 결실로서 분류된 제1 수준은 예를 들면, 약 -0.495를 제1 수준의 값에 더함으로써 조절된다.

[0345] 일부 실시양태에서, 동형접합 결실에 대한 PAV 계수는 약 0.6 내지 약 0.4이다. 일부 실시양태에서, 동형접합 결실에 대한 PAV 계수는 약 0.60, 0.59, 0.58, 0.57, 0.56, 0.55, 0.54, 0.53, 0.52, 0.51, 0.50, 0.49, 0.48, 0.47, 0.46, 0.45, 0.44, 0.43, 0.42, 0.41 또는 0.40이다. 종종, 동형접합 결실에 대한 PAV 계수는 약 0.5이다.

[0346] 예를 들면, NRV가 약 1이고 동형접합 결실의 기대 수준이 약 0과 동등한 경우, 동형접합 결실에 대한 PAV는 상기 식에 따라 약 1로서 측정된다. 이 경우, 동형접합 결실로서 분류된 제1 수준은 예를 들면, 약 1을 제1 수준의 값에 더함으로써 조절된다.

[0347] 일부 실시양태에서, PAV는 카피 수 변이에 대한 기대 수준(예를 들면, 카피 수 변이의 기대 수준)과 동등하거나 거의 동등하다.

[0348] 일부 실시양태에서, 수준의 카운트는 조절을 수행하기 전에 정규화된다. 일부 실시양태에서, 프로파일 내의 일부 또는 모든 수준의 카운트는 조절을 수행하기 전에 정규화된다. 예를 들면, 수준의 카운트는 기준 수준의 카운트 또는 NRV에 따라 정규화될 수 있다. 일부 실시양태에서, 수준(예를 들면, 제2 수준)의 카운트는 기준 수준의 카운트 또는 NRV에 따라 정규화되고, 프로파일 내의 모든 다른 수준(예를 들면, 제1 수준)의 카운트는 조절을 수행하기 전에 동일한 기준 수준의 카운트 또는 NRV에 비해 상대적으로 정규화된다.

[0349] 일부 실시양태에서, 프로파일의 수준은 1회 이상의 조절로부터 생성된다. 일부 실시양태에서, 프로파일의 수준은 프로파일 내의 하나 이상의 수준이 조절된 후 측정된다. 일부 실시양태에서, 프로파일의 수준은 1회 이상의 조절이 수행된 후 다시 계산된다.

[0350] 일부 실시양태에서, 카피 수 변이(예를 들면, 모체 카피 수 변이, 태아 카피 수 변이, 또는 모체 카피 수 변이 및 태아 카피 수 변이)는 조절로부터 확인된다(예를 들면, 직접적으로 또는 간접적으로 확인된다). 예를 들면, 조절된 프로파일 내의 수준(예를 들면, 조절된 제1 수준)은 모체 카피 수 변이로서 확인될 수 있다. 일부 실시양태에서, 조절의 규모는 카피 수 변이(예를 들면, 이형접합 결실, 동형접합 중복 등)의 종류를 표시한다. 일부 실시양태에서, 프로파일 내의 조절된 수준은 카피 수 변이에 대한 PAV의 값에 따라 카피 수 변이를 표시하는 것으로서 확인될 수 있다. 예를 들면, 주어진 프로파일에 대한 PAV는 동형접합 중복의 경우 약 -1이고 이형접합 중복의 경우 약 -0.5이고 이형접합 결실의 경우 약 0.5이고 동형접합 결실의 경우 약 1이다. 상기 예에서, 예를 들면, 약 -1까지 조절된 수준은 동형접합 중복으로서 확인될 수 있다. 일부 실시양태에서, 하나 이상의 카피 수 변이는 1회 이상의 조절을 포함하는 프로파일 또는 수준으로부터 확인될 수 있다.

[0351] 일부 실시양태에서, 프로파일 내의 조절된 수준은 비교된다. 일부 실시양태에서, 예외 및 오류는 조절된 수준을 비교함으로써 확인된다. 예를 들면, 종종 프로파일 내의 하나 이상의 조절된 수준이 비교되고, 특정 수준이 예외 또는 오류로서 확인될 수 있다. 일부 실시양태에서, 예외 또는 오류는 수준을 구성하는 하나 이상의 부분 내에서 확인된다. 예외 또는 오류는 (예를 들면, 프로파일 내의) 동일한 수준, 또는 인접하는, 근접하는, 서로 접하는 또는 이웃하는 부분들을 표시하는 하나 이상의 수준 내에서 확인될 수 있다. 일부 실시양태에서, 하나 이상의 조절된 수준은 인접하는, 근접하는, 서로 접하는 또는 이웃하는 부분들의 수준이고, 이때 하나 이상의 조절된 수준이 비교되고, 예외 또는 오류가 확인된다. 예외 또는 오류는 프로파일 또는 수준에서의 피크 또는 고랑일 수 있고, 이때 피크 또는 고랑의 원인은 공지되어 있거나 공지되어 있지 않다. 일부 실시양태에서, 조절된 수준이 비교되고 예외 또는 오류가 확인되고, 이때 예외 또는 오류는 확률적, 시스템적, 무작위적 또는 사용자 오류에 기인한다. 일부 실시양태에서, 조절된 수준이 비교되고, 예외 또는 오류가 프로파일로부터 제거된다. 일부 실시양태에서, 조절된 수준이 비교되고, 예외 또는 오류가 조절된다.

[0352] 수준에 기초한 태아 분획 측정

[0353] 일부 실시양태에서, 태아 분획은 모체 및/또는 태아 카피 수 변이를 표시하는 것으로서 분류된 수준에 따라 측정된다. 예를 들면, 태아 분획의 측정은 종종 태아 분획의 측정에 사용된 모체 및/또는 태아 카피 수 변이에 대한 기대 수준을 평가하는 단계를 포함한다. 일부 실시양태에서, 태아 분획은 동일한 종류의 카피 수 변이에 대해 결정된 기대 수준 범위에 따라 카피 수 변이를 표시하는 것으로서 분류된 수준(예를 들면, 제1 수준)에 대해 측정된다. 종종, 태아 분획은 기대 수준 범위 내에 속하는 관찰된 수준에 따라 측정됨으로써 모체 및/또는 태아 카피 수 변이로서 분류된다. 일부 실시양태에서, 태아 분획은 모체 및/또는 태아 카피 수 변이로서 분류된 관찰된 수준(예를 들면, 제1 수준)이 동일한 모체 및/또는 태아 카피 수 변이에 대해 결정된 기대 수준과 상이할 때 측정된다.

[0354] 일부 실시양태에서, 수준(예를 들면, 제1 수준, 관찰된 수준)은 제2 수준과 유의하게 상이하고, 제1 수준은 모체 및/또는 태아 카피 수 변이로서 분류되고, 태아 분획은 제1 수준에 따라 측정된다. 일부 실시양태에서, 제1 수준은 프로파일 내의 제2 수준과 유의하게 상이한 관찰된 및/또는 실험적으로 수득된 수준이고, 태아 분획은 제1 수준에 따라 측정된다. 일부 실시양태에서, 제1 수준은 평균치, 평균 또는 합산된 수준이고, 태아 분획은 제1 수준에 따라 측정된다. 일부 실시양태에서, 제1 수준 및 제2 수준은 관찰된 및/또는 실험적으로 수득된 수준이고, 태아 분획은 제1 수준에 따라 측정된다. 일부 경우, 제1 수준은 제1 부분 세트에 대한 정규화된 카운트를 포함하고, 제2 수준은 제2 부분 세트에 대한 정규화된 카운트를 포함하고, 태아 분획은 제1 수준에 따라 측정된다. 일부 실시양태에서, 제1 수준의 제1 부분 세트는 카피 수 변이를 포함하고(예를 들면, 제1 수준은 카피 수 변이를 표시함), 태아 분획은 제1 수준에 따라 측정된다. 일부 실시양태에서, 제1 수준의 제1 부분 세트는 동형접합 또는 이형접합 모체 카피 수 변이를 포함하고, 태아 분획은 제1 수준에 따라 측정된다. 일부 실시양태에서, 프로파일은 제1 부분 세트에 대한 제1 수준 및 제2 부분 세트에 대한 제2 수분을 포함하고, 제2 부분 세트는 카피 수 변이(예를 들면, 모체 카피 수 변이, 태아 카피 수 변이, 또는 모체 카피 수 변이 및 태아 카피 수 변이)를 실질적으로 포함하지 않고, 태아 분획은 제1 수준에 따라 측정된다.

[0355] 일부 실시양태에서, 수준(예를 들면, 제1 수준, 관찰된 수준)은 제2 수준과 유의하게 상이하고, 제1 수준은 모체 및/또는 태아 카피 수 변이에 대한 것으로서 분류되고, 태아 분획은 제1 수준 및/또는 카피 수 변이의 기대 수준에 따라 측정된다. 일부 실시양태에서, 제1 수준은 카피 수 변이에 대한 기대 수준에 따라 카피 수 변이에 대한 것으로서 분류되고, 태아 분획은 제1 수준과 기대 수준 사이의 차이에 따라 측정된다. 일부 실시양태에서, 수준(예를 들면, 제1 수준, 관찰된 수준)은 모체 및/또는 태아 카피 수 변이로서 분류되고, 태아 분획은 제1 수준과 카피 수 변이의 기대 수준 사이의 차이의 2배로서 측정된다. 일부 실시양태에서, 수준(예를 들면, 제1 수준, 관찰된 수준)은 모체 및/또는 태아 카피 수 변이로서 분류되고, 제1 수준을 기대 수준으로부터 뺌으로써 차이를 제공하고, 태아 분획은 상기 차이의 2배로서 측정된다. 일부 실시양태에서, 수준(예를 들면, 제1 수준, 관찰된 수준)은 모체 및/또는 태아 카피 수 변이로서 분류되고, 기대 수준을 제1 수준으로부터 뺌으로써 차이를 제공하고, 태아 분획은 상기 차이의 2배로서 측정된다.

[0356] 종종, 태아 분획은 %로서 제공된다. 예를 들면, 태아 분획을 100으로 나누어 % 값을 제공할 수 있다. 예를 들면, 모체 동형접합 중복을 표시하고 155의 수준을 갖는 제1 수준, 및 150의 수준을 갖는 모체 동형접합 중복에 대한 기대 수준의 경우, 태아 분획은 10%로서 측정될 수 있다(예를 들면, (태아 분획 = 2 x (155 - 150)).

[0357] 일부 실시양태에서, 태아 분획은 카피 수 변이로서 분류되는 프로파일 내의 2개 이상의 수준으로부터 측정된다. 예를 들면, 종종 프로파일 내의 2개 이상의 수준(예를 들면, 2개 이상의 제1 수준)은 기준 수준(예를 들면, 제2 수준, 카피 수 변이를 실질적으로 포함하지 않는 수준)과 유의하게 상이한 것으로서 확인되고, 상기 2개 이상의 수준은 모체 및/또는 태아 카피 수 변이를 표시하는 것으로서 분류되고, 태아 분획은 상기 2개 이상의 수준 각각으로부터 측정된다. 일부 실시양태에서, 태아 분획은 프로파일 내에서의 약 3회 이상, 약 4회 이상, 약 5회 이상, 약 6회 이상, 약 7회 이상, 약 8회 이상 또는 약 9회 이상의 태아 분획 측정으로부터 측정된다. 일부 실시양태에서, 태아 분획은 프로파일 내에서의 약 10회 이상, 약 20회 이상, 약 30회 이상, 약 40회 이상, 약 50회 이상, 약 60회 이상, 약 70회 이상, 약 80회 이상 또는 약 90회 이상의 태아 분획 측정으로부터 측정된다. 일부 실시양태에서, 태아 분획은 프로파일 내에서의 약 100회 이상, 약 200회 이상, 약 300회 이상, 약 400회 이상, 약 500회 이상, 약 600회 이상, 약 700회 이상, 약 800회 이상, 약 900회 이상 또는 약 1000회 이상의 태아 분획 측정으로부터 측정된다. 일부 실시양태에서, 태아 분획은 프로파일 내에서의 약 10회 내지 약 1000회, 약 20회 내지 약 900회, 약 30회 내지 약 700회, 약 40회 내지 약 600회, 약 50회 내지 약 500회, 약 50회 내지 약 400회, 약 50회 내지 약 300회, 약 50회 내지 약 200회, 또는 약 50회 내지 약 100회의 태아 분획 측정으

로부터 측정된다.

[0358] 일부 실시양태에서, 태아 분획은 프로파일 내에서의 다회 태아 분획 측정의 평균치 또는 평균으로서 측정된다. 일부 실시양태에서, 다회 태아 분획 측정으로부터 측정된 태아 분획은 다회 태아 분획 측정의 평균(예를 들면, 평균치, 평균, 표준 평균치, 중간치 등)이다. 종종, 다회 태아 분획 측정으로부터 측정된 태아 분획은 당분야에서 공지되어 있거나 본원에 기재되어 있는 적합한 방법에 의해 측정된 평균 값이다. 일부 실시양태에서, 태아 분획 측정의 평균 값은 가중된 평균이다. 일부 실시양태에서, 태아 분획 측정의 평균 값은 비가중된 평균이다. 다회 태아 분획 측정으로부터 수득된 평균, 중간 또는 평균치 태아 분획 측정치(즉, 평균, 중간 또는 평균치 태아 분획 측정 값)는 종종 불확실성 값(예를 들면, 분산, 표준 편차, MAD 등)과 관련되어 있다. 일부 실시양태에서, 다회 측정으로부터 평균, 중간 또는 평균치 태아 분획 값을 측정하기 전에, 하나 이상의 벗어난 측정치는 제거된다(본원에 더 상세히 기재됨).

[0359] 프로파일 내의 일부 태아 분획 측정치는 종종 태아 분획의 전체 측정치(예를 들면, 평균 또는 평균치 태아 분획 측정치)에 포함되지 않는다. 일부 실시양태에서, 태아 분획 측정치는 프로파일 내의 제1 수준(예를 들면, 제2 수준과 유의하게 상이한 제1 수준)으로부터 유도되고, 제1 수준은 유전적 변이를 표시하지 않는다. 예를 들면, 프로파일 내의 일부 제1 수준(예를 들면, 봉우리(spike) 또는 고랑)은 예외 또는 공지되지 않은 원인으로부터 생성된다. 이러한 값은 종종 진짜 카피 수 변이로부터 수득된 다른 태아 분획 측정치와 유의하게 상이한 태아 분획 측정치를 생성한다. 일부 실시양태에서, 프로파일 내의 다른 태아 분획 측정치와 유의하게 상이한 태아 분획 측정치는 확인되고 태아 분획 측정치로부터 제거된다. 예를 들면, 예외적인 봉우리 및 고랑으로부터 수득된 일부 태아 분획 측정치는 이를 프로파일 내의 다른 태아 분획 측정치와 비교함으로써 확인되고 태아 분획의 전체 측정치로부터 배제된다.

[0360] 일부 실시양태에서, 평균, 중간 또는 평균치 태아 분획 측정치와 유의하게 상이한 독립적인 태아 분획 측정치는 확인되고/되거나, 인식되고/되거나 관찰가능한 차이이다. 일부 실시양태에서, 용어 "유의하게 상이한"은 통계학적으로 상이하고/하거나 통계학적으로 유의한 차이를 의미할 수 있다. "독립적인" 태아 분획 측정치는 카피 수 변이로서 분류된 특정 수준으로부터 측정된 태아 분획(예를 들면, 일부 실시양태에서 단일 측정치)일 수 있다. 임의의 적합한 역치 또는 범위를 사용하여 태아 분획 측정치가 평균, 중간 또는 평균치 태아 분획 측정치와 유의하게 상이한지를 확인할 수 있다. 일부 실시양태에서, 태아 분획 측정치는 평균, 중간 또는 평균치 태아 분획 측정치와 유의하게 상이하고, 측정치는 평균치 또는 평균 값으로부터 % 이탈로서 표현될 수 있다. 일부 실시양태에서, 평균, 중간 또는 평균치 태아 분획 측정치와 유의하게 상이한 태아 분획 측정치는 약 10% 이상만큼 상이하다. 일부 실시양태에서, 평균, 중간 또는 평균치 태아 분획 측정치와 유의하게 상이한 태아 분획 측정치는 약 15% 이상만큼 상이하다. 일부 실시양태에서, 평균, 중간 또는 평균치 태아 분획 측정치와 유의하게 상이한 태아 분획 측정치는 약 15% 내지 약 100% 이상만큼 상이하다.

[0361] 일부 실시양태에서, 태아 분획 측정치는 평균 또는 평균치 태아 분획 측정치와 관련된 불확실성 값의 배수에 따라 평균, 중간 또는 평균치 태아 분획 측정치와 유의하게 상이하다. 종종, 불확실성 값 및 상수 n(예를 들면, 신뢰 구간)은 범위(예를 들면, 불확실성 컷오프)를 한정한다. 예를 들면, 종종 불확실성 값은 태아 분획 측정치에 대한 표준 편차(예를 들면, +/- 5)이고 상수 n(예를 들면, 신뢰 구간)과 곱해져 범위 또는 불확실성 컷오프(예를 들면, 종종 5 시그마로서 지칭되는 $5n$ 내지 $-5n$)를 한정한다. 일부 실시양태에서, 독립적인 태아 분획 측정치는 불확실성 컷오프에 의해 한정된 범위를 벗어나고 평균, 중간 또는 평균치 태아 분획 측정치와 유의하게 상이한 것으로서 간주된다. 예를 들면, 평균 값이 10이고 불확실성 컷오프가 3인 경우, 13 초과 또는 7 미만의 독립적인 태아 분획이 유의하게 상이하다. 일부 실시양태에서, 평균, 중간 또는 평균치 태아 분획 측정치와 유의하게 상이한 태아 분획 측정치는 불확실성 값의 n배(예를 들면, n x 시그마) 이상만큼 상이하고, 이때 n은 약 1, 2, 3, 4, 5, 6, 7, 8, 9 또는 10 이상이다. 일부 실시양태에서, 평균, 중간 또는 평균치 태아 분획 측정치와 유의하게 상이한 태아 분획 측정치는 불확실성 값의 n배(예를 들면, n x 시그마) 이상만큼 상이하고, 이때 n은 약 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9, 3.0, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9 또는 4.0 이상이다.

[0362] 일부 실시양태에서, 수준은 태아 및/또는 모체 미세배수성을 표시한다. 일부 실시양태에서, 수준(예를 들면, 제1 수준, 관찰된 수준)은 제2 수준과 유의하게 상이하고, 제1 수준은 모체 및/또는 태아 카피 수 변이로서 분류되고, 제1 수준 및/또는 제2 수준은 태아 미세배수성 및/또는 모체 미세배수성을 표시한다. 일부 실시양태에서, 제1 수준은 태아 미세배수성을 표시한다. 일부 실시양태에서 제1 수준은 모체 미세배수성을 표시한다. 종종, 제1 수준은 태아 미세배수성 및 모체 미세배수성을 표시한다. 일부 실시양태에서, 수준(예를 들면, 제1 수준, 관찰된 수준)은 제2 수준과 유의하게 상이하고, 제1 수준은 모체 및/또는 태아 카피 수 변이로서 분류되고, 제1

수준은 태아 및/또는 모체 미세배수성을 표시하고, 태아 분획은 태아 및/또는 모체 미세배수성에 따라 측정된다. 일부 경우, 제1 수준은 모체 및/또는 태아 카피 수 변이로서 분류되고, 제1 수준은 태아 미세배수성을 표시하고, 태아 분획은 태아 미세배수성에 따라 측정된다. 일부 실시양태에서, 제1 수준은 모체 및/또는 태아 카피 수 변이로서 분류되고, 제1 수준은 모체 미세배수성을 표시하고, 태아 분획은 모체 미세배수성에 따라 측정된다. 일부 실시양태에서, 제1 수준은 모체 및/또는 태아 카피 수 변이로서 분류되고, 제1 수준은 모체 및 태아 미세배수성을 표시하고, 태아 분획은 모체 및 태아 미세배수성에 따라 측정된다.

[0363] 일부 실시양태에서, 태아 분획의 측정은 태아 및/또는 모체 미세배수성의 측정을 포함한다. 일부 실시양태에서, 수준(예를 들면, 제1 수준, 관찰된 수준)은 제2 수준과 유의하게 상이하고, 제1 수준은 모체 및/또는 태아 카피 수 변이로서 분류되고, 태아 및/또는 모체 미세배수성은 제1 수준 및/또는 제2 수준에 따라 측정되고, 태아 분획은 측정된다. 일부 실시양태에서, 제1 수준은 모체 및/또는 태아 카피 수 변이로서 분류되고, 태아 미세배수성은 제1 수준 및/또는 제2 수준에 따라 측정되고, 태아 분획은 태아 미세배수성에 따라 측정된다. 일부 실시양태에서, 제1 수준은 모체 및/또는 태아 카피 수 변이로서 분류되고, 모체 미세배수성은 제1 수준 및/또는 제2 수준에 따라 측정되고, 태아 분획은 모체 미세배수성에 따라 측정된다. 일부 실시양태에서, 제1 수준은 모체 및/또는 태아 카피 수 변이로서 분류되고, 모체 및 태아 미세배수성은 제1 수준 및/또는 제2 수준에 따라 측정되고, 태아 분획은 모체 및 태아 미세배수성에 따라 측정된다.

[0364] 태아 분획은 종종 모체의 미세배수성이 주어진 수준 또는 카피 수 변이로서 분류된 수준에 대해 태아의 미세배수성과 상이할(예를 들면, 동일하지 않을) 때 측정된다. 일부 실시양태에서, 태아 분획은 모체가 중복에 대한 동형접합성(예를 들면, 2의 미세배수성)을 갖고 태아가 동일한 중복에 대한 이형접합성(예를 들면, 1.5의 미세배수성)을 가질 때 측정된다. 일부 실시양태에서, 태아 분획은 모체가 중복에 대한 이형접합성(예를 들면, 1.5의 미세배수성)을 갖고 태아가 동일한 중복에 대한 동형접합성(예를 들면, 2의 미세배수성)을 갖거나 중복이 태아에 부재할(예를 들면, 1의 미세배수성) 때 측정된다. 일부 실시양태에서, 태아 분획은 모체가 결실에 대한 동형접합성(예를 들면, 0의 미세배수성)을 갖고 태아가 동일한 결실에 대한 이형접합성(예를 들면, 0.5의 미세배수성)을 가질 때 측정된다. 일부 실시양태에서, 태아 분획은 모체가 결실에 대한 이형접합성(예를 들면, 0.5의 미세배수성)을 갖고 태아가 동일한 결실에 대한 동형접합성(예를 들면, 0의 미세배수성)을 갖거나 결실이 태아에 부재할(예를 들면, 1의 미세배수성) 때 측정된다.

[0365] 일부 실시양태에서, 태아 분획은 모체의 미세배수성이 카피 수 변이로서 확인된 주어진 수준에 대해 태아의 미세배수성과 동일할(예를 들면, 동일한 것으로서 확인될) 때 측정될 수 없다. 예를 들면, 일부 실시양태에서, 모체 및 태아 둘 다가 동일한 수의 카피의 카피 수 변이를 보유하는 경우 주어진 수준에 대한 태아 분획은 측정되지 않는다. 예를 들면, 태아 분획은 모체 및 태아 둘 다가 동일한 결실에 대한 동형접합성 또는 동일한 중복에 대한 동형접합성을 가질 때 카피 수 변이로서 분류된 수준에 대해 측정될 수 없다. 일부 실시양태에서, 태아 분획은 모체 및 태아가 동일한 결실에 대한 이형접합성 또는 동일한 중복에 대한 이형접합성을 가질 때 카피 수 변이로서 분류된 수준에 대해 측정될 수 없다. 샘플에 대한 다회 태아 분획 측정이 수행되는 실시양태에서, 평균, 중간 또는 평균치 값으로부터 유의하게 벗어나는 측정치는 모체 배수성이 태아 배수성과 동등한 카피 수 변이로부터 비롯될 수 있고, 이러한 측정치는 고려사항으로부터 제거될 수 있다.

[0366] 일부 실시양태에서, 모체 카피 수 변이 및 태아 카피 수 변이의 미세배수성은 공지되어 있지 않다. 일부 실시양태에서, 카피 수 변이에 대한 태아 및/또는 모체 미세배수성의 측정이 없는 경우, 태아 분획이 생성되고 평균, 중간 또는 평균치 태아 분획 측정치와 비교된다. 평균, 중간 또는 평균치 태아 분획 측정치와 유의하게 상이한 카피 수 변이에 대한 태아 분획 측정치는 종종 모체 및 태아의 미세배수성이 카피 수 변이에 대해 동일하기 때문이다. 평균, 중간 또는 평균치 태아 분획 측정치와 유의하게 상이한 태아 분획 측정치는 종종 차이의 원천 또는 원인과 관계없이 전체 태아 분획 측정치로부터 배제된다. 일부 실시양태에서, 모체 및/또는 태아의 미세배수성은 당분야에서 공지된 방법(예를 들면, 표적화된 서열결정 방법)에 의해 측정되고/되거나 검증된다.

[0367] 태아 배수성

[0368] 일부 실시양태에서, 부분적으로 유전적 변이(예를 들면, 염색체 이배수성, 삼염색체성)의 존재 또는 부재를 확인하기 위해 태아 배수성 측정이 이용된다. 태아 배수성은 부분적으로 본원에 기재된 방법을 포함하는, 적합한 태아 분획 측정 방법에 의해 측정된 태아 분획의 측정치로부터 측정될 수 있다. 일부 실시양태에서, 태아 배수성은 태아 분획 측정 및 방정식 8, 20 또는 21, 또는 이들의 변경된 식 또는 유도된 식에 따라 측정된다(실시예 2). 일부 실시양태에서, 태아 배수성은 하기 방법에 의해 측정된다. 일부 실시양태에서, 하기 각각의 방법은 다수의 샘플들에 대한 게놈의 부분(즉, 부분 i)에 대해 측정된 계산된 기준 카운트 F_i(종종 f_i로서 표시됨)를 요구

하고, 이때 상기 게놈의 부분 i에 대한 태아의 배수성은 정배수체이다. 일부 실시양태에서, 불확실성 값(예를 들면, 표준 편차, σ)은 기준 카운트 f_i에 대해 측정된다. 일부 실시양태에서, 기준 카운트 f_i, 불확실성 값, 검사 샘플 카운트 및/또는 측정된 태아 분획(F)이 하기 방법에 따라 태아 배수성을 측정하는 데에 사용된다. 일부 실시양태에서, 기준 카운트(예를 들면, 평균치, 평균 또는 중간 기준 카운트)는 본원에 기재된 방법(예를 들면, 부분별 정규화, GC 함량에 의한 정규화, 선형 및 비선형 최소 자승 회귀, LOESS, GC LOESS, LOWESS, PERUN, ChAI, RM, GCRM 및/또는 이들의 조합)에 의해 정규화된다. 일부 실시양태에서, 정배수체인 게놈의 분절의 기준 카운트는 기준 카운트가 PERUN에 의해 정규화될 때 1과 동등하다. 일부 실시양태에서, (예를 들면, 정배수체인 것으로 공지된 태아에 대한) 기준 카운트 및 게놈의 부분 또는 분절에 대한 검사 샘플의 카운트 둘 다가 PERUN에 의해 정규화되고, 기준 카운트는 1과 동등하다. 마찬가지로, 일부 실시양태에서, 정배수체인 게놈의 부분 또는 분절의 기준 카운트는 이 카운트가 기준 카운트의 중간치에 의해 정규화될(즉, 나누어질) 때 1과 동등하다. 예를 들면, 일부 실시양태에서, (예를 들면, 정배수체인 태아에 대한) 기준 카운트 및 게놈의 부분 또는 분절에 대한 검사 샘플의 카운트 둘 다가 중간 기준 카운트에 의해 정규화되고, 정규화된 기준 카운트는 1과 동등하고, 검사 샘플 카운트는 중간 기준 카운트에 의해 정규화된다(예를 들면, 나누어진다). 일부 실시양태에서, (예를 들면, 정배수체인 태아에 대한) 기준 카운트 및 게놈의 부분 또는 분절에 대한 검사 샘플의 카운트 둘 다가 GCRM, GC, RM 또는 적합한 방법에 의해 정규화된다. 일부 실시양태에서, 기준 카운트는 평균치, 평균 또는 중간 기준 카운트이다. 기준 카운트는 종종 부분에 대한 정규화된 카운트(예를 들면, 정규화된 게놈 구획 수준)이다. 일부 실시양태에서, 기준 카운트 및 검사 샘플에 대한 카운트는 미처리 카운트이다. 일부 실시양태에서, 기준 카운트는 평균치, 평균 또는 중간 카운트 프로파일로부터 측정된다. 일부 실시양태에서, 기준 카운트는 계산된 게놈 구획 수준이다. 일부 실시양태에서, 기준 샘플의 기준 카운트 및 검사 샘플(예를 들면, 환자 샘플, 예를 들면, y_i)의 카운트는 동일한 방법 또는 프로세스에 의해 정규화된다.

[0369] 일부 실시양태에서, 태아 분획(F)의 측정치가 측정된다. 그 다음, 이 태아 분획 값은 방정식 8, 이의 유도된 식 또는 변경된 식에 따라 태아 배수성을 측정하는 데에 사용된다. 일부 실시양태에서, 태아가 정배수체인 경우 음의 값이 수득되고, 태아가 정배수체가 아닌 경우 양의 값이 수득된다. 일부 실시양태에서, 음의 값은 태아가 고려되는 게놈의 분절에 대한 정배수체라는 것을 표시한다. 일부 실시양태에서, 음의 값이 아닌 값은 태아가 이배수성(예를 들면, 중복)을 포함한다는 것을 표시한다. 일부 실시양태에서, 음의 값이 아닌 값은 태아가 삼염색체성을 포함한다는 것을 표시한다. 일부 실시양태에서, 임의의 양의 값은 태아가 이배수성(예를 들면, 삼염색체성, 중복)을 포함한다는 것을 표시한다.

[0370] 일부 실시양태에서, 제곱 잔차의 합계가 측정된다. 예를 들면, 방정식 8로부터 유도된 제곱 잔차의 합계를 나타내는 방정식은 방정식 18로 예시된다. 일부 실시양태에서, 제곱 잔차의 합계는 1의 값으로 설정된 배수성 값 X(방정식 9 참조) 및 3/2의 값으로 설정된 배수성 값(방정식 13 참조)에 대해 방정식 8로부터 측정된다. 일부 실시양태에서, 제곱 잔차의 합계(방정식 9 및 13)는 게놈 또는 염색체의 분절(예를 들면, 게놈의 분절 내의 기준 게놈 i의 모든 부분들)에 대해 측정된다. 예를 들면, 제곱 잔차의 합계(예를 들면, 방정식 9 및 13)는 21번 염색체, 13번 염색체, 18번 염색체 또는 이들의 부분에 대해 측정될 수 있다. 일부 실시양태에서, 태아의 배수성 상태를 확인하기 위해, 방정식 13의 결과를 방정식 9로부터 빼어 값 phi에 도달한다(예를 들면, 방정식 14 참조). 일부 실시양태에서, 값 phi의 부호(즉, 양 또는 음)는 태아 이배수성의 존재 또는 부재를 결정한다. 일부 실시양태에서, 음의 값인 phi 값(예를 들면, 방정식 14로부터 유도됨)은 이배수성의 부재(예를 들면, 태아가 기준 게놈 i의 부분에 대한 정배수체라는 것)를 표시하고, 음의 값이 아닌 phi 값은 이배수성(예를 들면, 삼염색체성)의 존재를 표시한다.

[0371] 일부 실시양태에서, 기준 카운트 f_i, 기준 카운트에 대한 불확실성 값 σ 및/또는 측정된 태아 분획(F)은 기준 게놈 i의 모든 부분들의 합계에 대한 제곱 잔차의 합계를 측정하기 위해 방정식 9 및 13에서 사용된다. 일부 실시양태에서, 기준 카운트 f_i, 기준 카운트에 대한 불확실성 값 σ 및/또는 측정된 태아 분획(F)은 태아 배수성을 측정하기 위해 방정식 9 및 13에서 사용된다. 일부 실시양태에서, 검사 샘플에 대한 카운트(예를 들면, 정규화된 카운트, 예를 들면, 계산된 게놈 구획 수준)(부분 i에 대해 y_i로 표시됨)는 부분 i에 대한 태아의 배수성 상태를 확인하는 데에 사용된다. 예를 들면, 일부 실시양태에서 게놈의 분절에 대한 배수성 상태는 기준 카운트 f_i, (예를 들면, 기준 카운트로부터의) 불확실성 값, 검사 샘플에 대해 측정된 태아 분획(F) 및 검사 샘플에 대해 측정된 카운트 y_i에 따라 확인되고, 이때 배수성 상태는 방정식 14, 또는 이의 유도된 식 또는 변경된 식에 따라 확인된다. 일부 실시양태에서, 카운트 y_i 및/또는 기준 카운트는 본원에 기재된 방법(예를 들면, 부분별 정

규화, GC 함량에 의한 정규화, 선형 및 비선형 최소 자승 회귀, LOESS, GC LOESS, LOWESS, PERUN, ChAI, RM, GCRM 및 이들의 조합)에 의해 정규화된다. 일부 실시양태에서, 게놈 또는 염색체의 부분 또는 분절에 대한 태아 배수성 상태(예를 들면, 정배수체, 이수체, 삼염색체성)는 상기 단락들 및 실시예 단락에 기재된 비한정적 예에 의해 확인된다.

[0372] 일부 실시양태에서, 태아 분획이 검사 샘플로부터 측정되고, 카운트 y가 검사 샘플에 대해 측정되고, 이들 둘 다가 검사 샘플로부터 태아에 대한 배수성을 측정하는 데에 사용된다. 본원에 기재된 방법의 일부 실시양태에서, X로 표시된 태아 배수성의 값은 고정되지 않거나 추정되지 않는다. 본원에 기재된 방법의 일부 실시양태에서, 태아 분획 F는 고정된다. 일부 실시양태에서, 배수성(예를 들면, 배수성 값)은 방정식 20 또는 21에 따라 게놈의 부분 또는 분절에 대해 측정된다(실시예 2). 이 방법의 일부 실시양태에서, 배수성 값이 측정되고, 이때 이 값은 1, 3/2 또는 5/4에 가깝다. 일부 실시양태에서, 약 1의 배수성 값은 정배수체 태아를 표시하고, 약 3/2의 값은 태아 삼염색체성을 표시하고, 쌍둥이의 경우 약 5/4의 값은 1명의 태아가 삼염색체성을 포함하고 다른 태아가 고려되는 게놈의 부분 또는 분절에 대한 정배수체라는 것을 표시한다. 태아 배수성 측정으로부터 태아 이배수성의 존재 또는 부재를 확인하는 것에 대한 추가 정보는 하기 또 다른 단락에서 논의되어 있다.

[0373] 일부 실시양태에서, 태아 분획이 측정되고 그의 측정된 값에서 고정되고 태아 배수성이 회귀로부터 측정된다. 임의의 적합한 회귀가 이용될 수 있고, 이의 비한정적 예는 선형 회귀, 비선형 회귀(예를 들면, 다항식 회귀) 등을 포함한다. 일부 실시양태에서, 선형 회귀는 방정식 8, 20 또는 21, 및/또는 이들의 유도된 식 또는 변경된 식에 따라 이용된다. 일부 실시양태에서, 이용된 선형 회귀는 방정식 8, 20 또는 21, 및/또는 이들의 유도된 식 또는 변경된 식으로부터 유도된 제곱 잔차의 합계에 따른다. 일부 실시양태에서, 태아 배수성은 방정식 8, 20 또는 21, 및/또는 이들의 유도된 식 또는 변경된 식에 따라 측정되고, 회귀는 이용되지 않는다. 일부 실시양태에서, 태아 배수성은 기준 게놈 i의 다수의 부분들에 대해 방정식 8, 20 또는 21, 및/또는 이들의 유도된 식 또는 변경된 식으로부터 유도된 제곱 잔차의 합계에 따라 측정되고, 회귀는 이용되지 않는다. 방정식의 유도된 식은 방정식의 수학적 증명으로부터 수득된 방정식의 임의의 변경된 식이다.

[0374] 일부 실시양태에서, (앞서 본원에 기재된) 기준 카운트 f_i, 불확실성 값 σ 및/또는 측정된 태아 분획(F)은 태아 배수성을 측정하기 위해 방정식 20 및 21에서 사용된다. 일부 실시양태에서, 기준 카운트 f_i, 불확실성 값 σ 및/또는 측정된 태아 분획(F)은 부분 i에 대한 또는 기준 게놈 i의 다수의 부분들 전체(예를 들면, 염색체 또는 이의 분절에 대한 기준 게놈 i의 모든 부분들 전체)에 대한 태아 배수성 X를 측정하기 위해 방정식 20 또는 21에서 사용된다. 일부 실시양태에서, 검사 샘플에 대한 카운트(예를 들면, 정규화된 카운트, 계산된 게놈 구획 수준)(부분 i에 대해 y_i로 표시됨)는 기준 게놈 i의 다수의 부분들로 표시된 게놈의 분절에 대한 태아의 배수성을 측정하는 데에 사용된다. 예를 들면, 일부 실시양태에서, 게놈의 분절에 대한 배수성 X는 기준 카운트 f_i, 불확실성 값, 검사 샘플에 대해 측정된 태아 분획(F) 및 검사 샘플에 대해 측정된 카운트 y_i에 따라 측정되고, 이때 배수성은 방정식 20 또는 21, 또는 이들의 유도된 식 또는 변경된 식에 따라 측정된다. 일부 실시양태에서, 카운트 y_i 및/또는 기준 카운트는 본원에 기재된 방법(예를 들면, 부분별 정규화, GC 함량에 의한 정규화, 선형 및 비선형 최소 자승 회귀, LOESS, GC LOESS, LOWESS, PERUN, ChAI, RM, GCRM 및 이들의 조합)에 의해 정규화된다. 일부 실시양태에서, 카운트 y_i 및/또는 기준 카운트는 본원에 기재된 방법(예를 들면, 부분별 정규화, GC 함량에 의한 정규화, 선형 및 비선형 최소 자승 회귀, LOESS, GC LOESS, LOWESS, PERUN, ChAI, RM, GCRM 및 이들의 조합)에 의해 정규화되고/되거나 프로세싱된다. 일부 실시양태에서, 카운트 y_i 및 f_i는 게놈 또는 염색체의 동일한 부분 또는 분절로 맵핑된 카운트이다.

[0375] 불확실성 값 σ는 오차의 적합한 측정치일 수 있고, 이러한 측정치의 비한정적 예는 표준 편차, 표준 오차, 계산된 분산, p-값 및/또는 평균 절대 편차(MAD)를 포함한다. 불확실성 값 σ는 임의의 적합한 측정치에 대해 측정될 수 있고, 이러한 측정치의 비한정적 예는 Z-점수, Z-값, t-값, p-값, 교차검증 오차, 게놈 구획 수준, 계산된 게놈 구획 수준, 수준, 카운트 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, σ는 1의 값으로 설정된다. 일부 실시양태에서, σ는 1의 값으로 설정되지 않는다. 일부 실시양태에서, σ의 값은 추정되고 종종 측정되고/되거나 계산된다.

[0376] 일부 실시양태에서, M_i은 게놈 i의 부분에 대한 모체의 배수성(즉, 모체 배수성)이다. 일부 실시양태에서, M_i은 y_i가 측정되는 환자와 동일한 환자(예를 들면, 동일한 검사 샘플)에 대해 측정된다. 일부 실시양태에서, 모체 배

수성 M_i은 공지되어 있거나 본원에 기재된 방법에 따라 측정된다. 일부 실시양태에서, 모체 배수성은 패딩 전 또는 후에(예를 들면, 수준을 조절한 후에) 측정된다. 일부 실시양태에서, M_i은 프로파일의 가시화로부터 추정되거나 측정된다. 일부 실시양태에서, 모체 배수성 M_i는 공지되어 있지 않다. 일부 실시양태에서, 모체 배수성 M_i은 추정된다. 예를 들면, 일부 실시양태에서, 모체가 평가되는 게놈의 분절에서 결실 및/또는 중복을 갖지 않는다는 것이 추정되거나 공지되어 있다. 일부 실시양태에서, 모체 배수성이 1인 것이 추정되거나 공지되어 있다. 일부 실시양태에서, 모체 배수성은 패딩 후(예를 들면, 수준을 조절한 후)에 1의 값으로 설정된다. 일부 실시양태에서, 모체 배수성은 무시되고 1의 값으로 설정된다. 일부 실시양태에서, 방정식 21은 모체가 평가되는 게놈의 분절에서 결실 및/또는 중복을 갖지 않는다는 가정 하에 방정식 20으로부터 유도된다.

[0377] 일부 실시양태에서, 태아 배수성을 측정하는 방법은 임신 여성으로부터 수득된 검사 샘플에 대한 핵산 서열 리드에 따른다. 일부 실시양태에서, 서열 리드는 샘플(예를 들면, 검사 샘플)로부터의 순환 무세포 핵산의 리드이다. 일부 실시양태에서, 태아 배수성을 측정하는 방법은 기준 게놈의 부분들에 맵핑된 서열 리드의 카운트를 수득하는 단계를 포함한다. 일부 실시양태에서, 서열 리드는 기준 게놈의 부분 서브세트로 맵핑된다. 일부 실시양태에서, 태아 배수성의 측정은 태아 분획의 측정을 포함한다. 일부 실시양태에서, 태아 배수성의 측정은 게놈 구획 수준의 계산 또는 측정을 포함한다. 일부 실시양태에서, 태아 배수성의 측정은 태아 분획의 측정 및 게놈 구획 수준의 계산 또는 측정을 포함한다. 일부 실시양태에서, 태아 분획 및 계산된 게놈 구획 수준은 동일한 검사 샘플(예를 들면, 검사 샘플의 동일한 부분)로부터 측정된다. 일부 실시양태에서, 태아 분획 및 계산된 게놈 구획 수준은 동일한 검사 샘플(예를 들면, 검사 샘플의 동일한 부분)로부터 수득된 동일한 리드로부터 측정된다. 일부 실시양태에서, 태아 분획 및 계산된 게놈 구획 수준은 동일한 서열결정 실시 및/또는 동일한 유동 셀로부터 수득된 동일한 리드로부터 측정된다. 일부 실시양태에서, 태아 분획 및 계산된 게놈 구획 수준은 동일한 장치 및/또는 기계(예를 들면, 서열결정 장치, 유동 셀 등)로부터 측정된다.

[0378] 일부 실시양태에서, 태아 배수성을 측정하는 방법은 태아 분획 측정치 및 정규화된 카운트(예를 들면, 계산된 게놈 구획 수준)에 따라 결정되고, 이때 태아 분획 측정치 및 정규화된 카운트(예를 들면, 계산된 게놈 구획 수준)는 검사 샘플의 상이한 부분들(예를 들면, 동일한 대상체 또는 환자로부터 거의 동일한 시간에 채취된 상이한 분취물들 또는 예를 들면, 상이한 검사 샘플들)로부터 측정된다. 예를 들면, 종종 태아 분획은 검사 샘플의 제1 부분으로부터 측정되고, 정규화된 카운트 및/또는 게놈 구획 수준은 검사 샘플의 제2 부분으로부터 측정된다. 일부 실시양태에서, 태아 분획 및 계산된 게놈 구획 수준은 동일한 대상체(예를 들면, 환자)로부터 채취된 상이한 검사 샘플들(예를 들면, 검사 샘플의 상이한 부분들)로부터 측정된다. 일부 실시양태에서, 태아 분획 및 계산된 게놈 구획 수준은 상이한 시간에 수득된 리드로부터 측정된다. 일부 실시양태에서, 태아 분획 측정치 및 정규화된 카운트(예를 들면, 계산된 게놈 구획 수준)는 상이한 장치 및/또는 상이한 기계(예를 들면, 서열결정 장치, 유동 셀 등)로부터 측정된다.

[0379] 결정 분석 특징

[0380] 일부 실시양태에서, 결과의 확인(예를 들면, 판정), 또는 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재의 확인은 결정 분석에 따라 수행된다. 예를 들면, 결정 분석은 종종 하나 이상의 결과를 생성하는 하나 이상의 방법의 적용; 결과의 평가; 및 결과, 평가 및/또는 결정의 가능한 결과에 기초하고 최종 결정이 내려지는 프로세스의 일부 시점에서 종결되는 일련의 결정을 포함한다. 일부 실시양태에서, 결정 분석은 결정 트리이다. 일부 실시양태에서, 결정 분석은 하나 이상의 프로세스(예를 들면, 프로세스 단계, 예를 들면, 알고리즘)의 조화된 사용을 포함한다. 결정 분석은 사람, 시스템, 장치, 소프트웨어(예를 들면, 모듈), 컴퓨터, 프로세서(예를 들면, 마이크로프로세서) 등 또는 이들의 조합에 의해 수행될 수 있다. 일부 실시양태에서, 결정 분석은 결정 분석이 이용되지 않는 경우(예를 들면, 정규화된 카운트로부터 직접적으로 확인되는 경우)에 비해 거짓 음성 및 거짓 양성 확인을 감소시키면서 태아에서의 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 방법을 포함한다. 일부 실시양태에서, 결정 분석은 하나 이상의 미세중복 또는 미세결실과 관련된 상태의 존재 또는 부재를 확인하는 단계를 포함한다. 예를 들면, 일부 실시양태에서, 결정 분석은 대상체로부터의 검사 샘플에 대해 디조지 증후군과 관련된 하나 이상의 유전적 변이의 존재 또는 부재를 확인하는 단계를 포함한다. 일부 실시양태에서, 결정 분석은 대상체로부터의 검사 샘플에 대해 디조지 증후군의 존재 또는 부재를 확인하는 단계를 포함한다.

[0381] 일부 실시양태에서, 결정 분석은 게놈 또는 게놈의 분절(예를 들면, 염색체 또는 이의 부분)에 대한 프로파일을 생성하는 단계를 포함한다. 프로파일은 공지되어 있거나 본원에 기재되어 있는 임의의 적합한 방법에 의해 생성될 수 있고, 종종 기준 게놈의 부분들에 맵핑된 서열 리드의 카운트의 수득, 카운트의 정규화, 수준의 정규화,

패딩 등 또는 이들의 조합을 포함한다. 기준 게놈에 맵핑된 서열 리드의 카운트의 수득은 (예를 들면, 임신 여성 대상체로부터의) 샘플의 수득, 샘플로부터의 핵산(예를 들면, 순환 무세포 핵산)의 서열결정, 서열 리드의 수득, 기준 게놈의 부분으로의 서열 리드의 맵핑 등 또는 이들의 조합을 포함할 수 있다. 일부 실시양태에서, 프로파일의 생성은 기준 게놈의 부분들에 맵핑된 카운트를 정규화하여 계산된 게놈 구획 수준을 제공하는 것을 포함한다.

[0382] 일부 실시양태에서, 결정 분석은 분할을 포함한다. 일부 실시양태에서, 분할은 프로파일을 변경시키고/시키거나 변환시켜 프로파일의 하나 이상의 분해 렌더링을 제공한다. 분할 프로세스로 처리된 프로파일은 종종 기준 게놈 또는 이의 일부(예를 들면, 상염색체 및 성염색체) 내의 부분(예를 들면, 빈(bin))으로 맵핑된 정규화된 카운트의 프로파일이다. 본원에서 다루어진 바와 같이, 부분들에 맵핑된 미처리 카운트는 하나 이상의 적합한 정규화 프로세스(예를 들면, PERUN, LOESS, GC-LOESS, 주성분 정규화(ChAI) 또는 이들의 조합)에 의해 정규화되어 결정 분석의 일부로서 분할되는 프로파일을 생성할 수 있다. 프로파일의 분해 렌더링은 종종 프로파일의 변환이다. 프로파일의 분해 렌더링은 종종 게놈 또는 염색체, 또는 이들의 분절의 표시치로의 프로파일의 변환이다.

[0383] 일부 실시양태에서, 분할을 위해 이용된 분할 프로세스는 프로파일 내의 하나 이상의 다른 수준과 상이한(예를 들면, 실질적으로 또는 유의하게 상이한) 프로파일 내의 하나 이상의 수준의 위치를 찾아내고 이 수준을 확인한다. 프로파일 내의 또 다른 수준과 상이하고 프로파일 내의 또 다른 수준과 상이한 에지를 갖는, 분할 프로세스에 따라 프로파일에서 확인된 수준은 본원에서 웨이블렛으로서 지칭되고, 보다 일반적으로 별개의 분절에 대한 수준으로 지칭된다. 분할 프로세스는 하나 이상의 별개의 분절 또는 웨이블렛이 확인될 수 있는 분해 렌더링을 정규화된 카운트 또는 수준으로부터 생성할 수 있다. 별개의 분절은 일반적으로 분할되는 것(예를 들면, 염색체, 염색체들, 상염색체)보다 더 작은 부분(예를 들면, 빈)을 커버한다.

[0384] 일부 실시양태에서, 분할은 프로파일 내의 별개의 분절 및 웨이블렛의 에지의 위치를 찾아내고 이 에지를 확인한다. 일부 실시양태에서, 하나 이상의 별개의 분절 및 웨이블렛의 한쪽 또는 양쪽 에지가 확인된다. 예를 들면, 분할 프로세스는 프로파일 내의 별개의 분절 또는 웨이블렛의 우측 및/또는 좌측 에지의 위치(예를 들면, 게놈 좌표, 예를 들면, 부분 위치)를 확인할 수 있다. 별개의 분절 또는 웨이블렛은 종종 2개의 에지를 포함한다. 예를 들면, 별개의 분절 또는 웨이블렛은 좌측 에지 및 우측 에지를 포함할 수 있다. 일부 실시양태에서, 표시 또는 시야에 따라, 좌측 에지는 5'-에지일 수 있고 우측 에지는 프로파일 내의 핵산 분절의 3'-에지일 수 있다. 일부 실시양태에서, 좌측 에지는 3'-에지일 수 있고, 우측 에지는 프로파일 내의 핵산 분절의 5'-에지일 수 있다. 종종, 프로파일의 에지는 분할 전에 공지되어 있으므로, 일부 실시양태에서 프로파일의 에지는 수준의 어떤 에지가 5'-에지이고 어떤 에지가 3'-에지인지를 결정한다. 일부 실시양태에서, 프로파일 및/또는 별개의 분절(예를 들면, 웨이블렛)의 한쪽 또는 양쪽 에지는 염색체의 에지이다.

[0385] 일부 실시양태에서, 별개의 분절 또는 웨이블렛의 에지는 기준 샘플(예를 들면, 기준 프로파일)에 대해 생성된 분해 렌더링에 따라 결정된다. 일부 실시양태에서, 널 에지 높이 분포는 기준 프로파일(예를 들면, 염색체 또는 이의 분절의 프로파일)의 분해 렌더링에 따라 결정된다(예를 들면, 도 3 참조). 일부 실시양태에서, 프로파일 내의 별개의 분절 또는 웨이블렛의 에지는 별개의 분절 또는 웨이블렛의 수준이 널 에지 높이 분포를 벗어날 때 확인된다. 일부 실시양태에서, 프로파일 내의 별개의 분절 또는 웨이블렛의 에지는 기준 프로파일에 대한 분해 렌더링에 따라 계산된 Z-점수에 따라 확인된다.

[0386] 종종, 분할은 프로파일에서 2개 이상의 별개의 분절 또는 웨이블렛(예를 들면, 2개 이상의 단편화된 수준, 2개 이상의 단편화된 분절)을 생성한다. 일부 실시양태에서, 분할 프로세스로부터 유도된 분해 렌더링은 과다분할되거나 단편화되고, 다수의 별개의 분절들 또는 웨이블렛들을 포함한다. 종종, 분할에 의해 생성된 별개의 분절 또는 웨이블렛은 실질적으로 상이하고, 종종 분할에 의해 생성된 별개의 분절 또는 웨이블렛은 실질적으로 유사하다. 실질적으로 유사한 별개의 분절 또는 웨이블렛(예를 들면, 실질적으로 유사한 수준)은 종종 불확실성의 예정된 수준 미만의 정도로 상이한 게놈 구획 수준(예를 들면, 수준)을 각각 갖는 분할된 프로파일 내의 2개 이상의 인접한 별개의 분절 또는 웨이블렛을 지칭한다. 일부 실시양태에서, 실질적으로 유사한 별개의 분절 또는 웨이블렛은 서로 인접하고 개재 분절 또는 웨이블렛에 의해 분리되지 않는다. 일부 실시양태에서, 실질적으로 유사한 별개의 분절 또는 웨이블렛은 하나 이상의 보다 더 작은 분절 또는 웨이블렛에 의해 분리된다. 일부 실시양태에서, 실질적으로 유사한 별개의 분절 또는 웨이블렛은 약 1개 내지 약 20개, 약 1개 내지 약 15개, 약 1개 내지 약 10개, 또는 약 1개 내지 약 5개의 부분(예를 들면, 빈)에 의해 분리되고, 이때 하나 이상의 개재 부분이 실질적으로 유사한 별개의 분절 또는 웨이블렛 각각의 수준과 유의하게 상이한 수준을 갖는다. 일부 실시양태에서, 실질적으로 유사한 별개의 분절 또는 웨이블렛의 수준은 불확실성 수준의 약 3배 미만, 약 2배 미만, 약 1배 미만 또는 약 0.5배 미만의 정도로 상이하다. 일부 실시양태에서, 실질적으로 유사한 별개의 분절 또는

웨이블렛은 3 MAD 미만(예를 들면, 3 시그마 미만), 2 MAD 미만, 1 MD 미만 또는 약 0.5 MAD 미만의 정도로 상이한 중간 게놈 수준을 포함하고, 이때 MAD는 분절 또는 웨이블렛 각각의 중간 게놈 구획 수준으로부터 계산된다. 일부 실시양태에서, 실질적으로 상이한 별개의 분절 또는 웨이블렛은 인접하지 않거나 10개 이상, 15개 이상 또는 20개 이상의 부분에 의해 분리된다. 실질적으로 상이한 별개의 분절 또는 웨이블렛은 일반적으로 실질적으로 상이한 수준을 갖는다. 일부 실시양태에서, 실질적으로 상이한 별개의 분절 또는 웨이블렛은 불확실성 수준의 약 2.5배 초과, 약 3배 초과, 약 4배 초과, 약 5배 초과 또는 약 6배 초과의 정도로 상이한 수준을 포함한다. 일부 실시양태에서, 실질적으로 상이한 별개의 분절 또는 웨이블렛은 2.5 MAD 초과(예를 들면, 2.5 시그마 초과), 3 MAD 초과, 4 MAD 초과, 약 5 MAD 초과 또는 약 6 MAD 초과의 정도로 상이한 중간 게놈 구획 수준을 포함하고, 이때 MAD는 별개의 분절 또는 웨이블렛 각각의 중간 게놈 구획 수준으로부터 계산된다.

[0387] 일부 실시양태에서, 분할 프로세스는 프로파일 또는 이의 분절 내의 하나 이상의 별개의 분절 또는 웨이블렛(예를 들면, 수준)에 대한 수준(예를 들면, 정량적 값, 예를 들면, 평균 또는 중간 수준), 불확실성 수준(예를 들면, 불확실성 값), Z-점수, Z-값, p-값 등 또는 이들의 조합을 측정하는 단계를 포함한다. 일부 실시양태에서, 수준(예를 들면, 정량적 값, 예를 들면, 평균 또는 중간 수준), 불확실성 수준(예를 들면, 불확실성 값), Z-점수, Z-값, p-값 등 또는 이들의 조합은 별개의 분절 또는 웨이블렛에 대해 측정된다(예를 들면, 계산된다).

[0388] 일부 실시양태에서, 분할은 하나의 프로세스 또는 다수의 하위프로세스를 포함하는 프로세스에 의해 달성되고, 이 프로세스의 비한정적 예는 분해 발생 프로세스(예를 들면, 웨이블렛 분해 발생 프로세스), 역치화, 평준화, 평활화 등 또는 이들의 조합을 포함한다. 역치화, 평준화, 평활화 등은 분해 발생 프로세스와 함께 수행될 수 있고, 웨이블렛 분해 렌더링 프로세스와 관련하여 이하에 기재되어 있다.

[0389] 웨이블렛 분할 프로세스

[0390] 일부 실시양태에서, 분할은 웨이블렛 분해 발생 프로세스에 따라 수행된다. 일부 실시양태에서, 분할은 2개 이상의 웨이블렛 분해 발생 프로세스에 따라 수행된다. 일부 실시양태에서, 웨이블렛 분해 발생 프로세스는 프로파일에서 하나 이상의 웨이블렛을 확인하고 프로파일의 분해 렌더링을 제공한다.

[0391] 전체적으로 또는 부분적으로 분할은 본원에 기재되어 있거나 당분야에서 공지되어 있는 임의의 적합한 웨이블렛 분해 발생 프로세스에 의해 수행될 수 있다. 웨이블렛 분해 발생 프로세스의 비한정적 예는 Haar 웨이블렛 분할(Haar, Alfred (1910) "Zur Theorie der orthogonalen Funktionensysteme", Mathematische Annalen 69 (3): 331-371; Nason, G.P. (2008) "Wavelet methods in Statistics", R. Springer, New York)(예를 들면, 웨이브쓰레쉬(WaveThresh)), 웨이브쓰레쉬, 적합한 반복 이진 분할 프로세스, 서큘러 이진 분할 프로세스(CBS; circular binary segmentation)(Olshen, AB, Venkatraman, ES, Lucito, R, Wigler, M (2004) "Circular binary segmentation for the analysis of array-based DNA copy number data", Biostatistics, 5, 4:557-72; Venkatraman, ES, Olshen, AB (2007) "A faster circular binary segmentation algorithm for the analysis of array CGH data", Bioinformatics, 23, 6:657-63), 최대 중첩 별개의 웨이블렛 변환(MODWT)(L. Hsu, S. Self, D. Grove, T. Randolph, K. Wang, J. Delrow, L. Loo, and P. Porter, "Denoising array-based comparative genomic hybridization data using wavelets", Biostatistics (Oxford, England), vol. 6, no. 2, pp. 211-226, 2005), 정지 웨이블렛(SWT)(Y. Wang and S. Wang, "A novel stationary wavelet denoising algorithm for array-based DNA copy number data", International Journal of Bioinformatics Research and Applications, vol. 3, no. 2, pp. 206-222, 2007), 이중-트리 복합 웨이블렛 변환(DTCWT)(Nha, N., H. Heng, S. Oraintara and W. Yuhang (2007) "Denoising of Array-Based DNA Copy Number Data Using The Dual-tree Complex Wavelet Transform." 137-144), 최대 엔트로피 분할(maximum entropy segmentation), 에지 검출 커넬을 이용한 컨볼루션(convolution with edge detection kernel), 젠슨 샤논 발산(Jensen Shannon Divergence), 쿨백-레이블러 발산(Kullback-Leibler divergence), 이진 반복 분할(Binary Recursive Segmentation), 푸리에 변환(Fourier transform) 등 또는 이들의 조합을 포함한다.

[0392] 웨이블렛 분해 발생 프로세스는 적합한 언어(예를 들면, 당분야에서 공지된 컴퓨터 프로그래밍 언어)로 작성된 적합한 소프트웨어, 모듈 및/또는 코드, 및/또는 연산 시스템에 의해 표시될 수 있거나 수행될 수 있고, 이들의 비한정적 예는 UNIX, 리눅스(Linux), 오라클(oracle), 윈도우(windows), 우분투(Ubuntu), 액션스크립(ActionScript), C, C++, C#, 하스켈(Haskell), 자바(Java), 자바스크립(JavaScript), 오브젝티브(Objective)-C, 펄(Perl), 파이톤(Python), 루비(Ruby), 스몰토크(Smalltalk), SQL, 비주얼 베이직(Visual Basic), COBOL, 포트란(Fortran), UML, HTML(예를 들면, PHP를 가짐), PGP, G, R, S 등 또는 이들의 조합물을 포함한다. 일부 실시양태에서, 적합한 웨이블렛 분해 발생 프로세스는 S 또는 R 코드, 또는 팩키지(예를 들면, R 팩키지)로 표

시된다. 웨이블렛 분해 발생 프로세스를 위한 R, R 소스 코드, R 프로그램, R 팩키지 및 R 문서화가 CRAN 또는 CRAN 미러 사이트(예를 들면, The Comprehensive R Archive Network(CRAN); 월드 와이드 웹 URL cran.us.r-project.org)로부터 다운로드로 입수될 수 있다. CRAN은 R에 대한 코드 및 문서화의 동일한 최신 버전을 저장하는 전세계 ftp 및 웹 서버의 네트워크이다. 예를 들면, 웨이브쓰레쉬(웨이브쓰레쉬: 웨이블렛 통계 및 변환; 월드 와이드 웹 URL cran.r-project.org/web/packages/wavethresh/index.html) 및 웨이브쓰레쉬의 상세한 설명(팩키지 '웨이브쓰레쉬'; 월드 와이드 웹 RL cran.r-project.org/web/packages/wavethresh/wavethresh.pdf)이 다운로드로 입수될 수 있다. 일부 실시양태에서, 웨이블렛 분해 발생 프로세스(예를 들면, 최대 엔트로피 분할)를 위한 R 코드는 실시예 4에 기재되어 있다. CBS 방법을 위한 R 코드의 일례는 다운로드될 수 있다(예를 들면, DNAcopy; 월드 와이드 웹 URL bioconductor.org/packages/2.12/bioc/html/DNAcopy.html 또는 팩키지 'DNAcopy'; 월드 와이드 웹 URL bioconductor.org/packages/release/bioc/manuals/DNAcopy/man/DNAcopy.pdf).

[0393] 일부 실시양태에서, 웨이블렛 분해 발생 프로세스(예를 들면, Haar 웨이블렛 분할, 예를 들면, 웨이브쓰레쉬)는 역치화를 포함한다. 일부 실시양태에서, 역치화는 노이즈(noise)로부터 신호를 식별한다. 일부 실시양태에서, 역치화는 어떤 웨이블렛 계수(예를 들면, 접속점(node))가 신호를 표시하고 보유되어야 하는지 및 어떤 웨이블렛 계수가 노이즈의 반영을 표시하고 제거되어야 하는지를 결정한다. 일부 실시양태에서, 역치화는 하나 이상의 가변 파라미터를 포함하고, 이때 사용자는 상기 파라미터의 값을 설정한다. 일부 실시양태에서, 역치화 파라미터(예를 들면, 역치화 파라미터, 폴리시(policy) 파라미터)는 웨이블렛 분해 발생 프로세스에서 이용된 분할의 양을 기술할 수 있거나 정의할 수 있다. 임의의 적합한 파라미터 값이 사용될 수 있다. 일부 실시양태에서, 역치화 파라미터가 사용된다. 일부 실시양태에서, 역치화 파라미터 값은 소프트(soft) 역치화이다. 일부 실시양태에서, 소프트 역치화는 작은 계수 및 무의미한 계수를 제거하는 데에 이용된다. 일부 실시양태에서, 하드(hard) 역치화가 이용된다. 일부 실시양태에서, 역치화는 폴리시 파라미터를 포함한다. 임의의 적합한 폴리시 값이 사용될 수 있다. 일부 실시양태에서, 사용된 폴리시는 "보편적"이고, 일부 실시양태에서 사용된 폴리시는 "확실"하다.

[0394] 일부 실시양태에서, 웨이블렛 분해 발생 프로세스(예를 들면, Haar 웨이블렛 분할, 예를 들면, 웨이브쓰레쉬)는 평준화를 포함한다. 일부 실시양태에서, 역치화 후 일부 고수준 계수가 남아있다. 이 계수는 원래의 신호에서 급격한 변화 또는 큰 급증을 표시하고, 일부 실시양태에서 평준화에 의해 제거된다. 일부 실시양태에서, 평준화는 값을 분해 수준 c로서 공지된 파라미터에 배정하는 단계를 포함한다. 일부 실시양태에서, 최적 분해 수준은 하나 이상의 측정된 값, 예컨대, 염색체의 길이(예를 들면, 프로파일의 길이), 검출할 원하는 웨이블렛 길이, 태아 분획, 서열 커버리지(예를 들면, 플레스 수준) 및 정규화된 프로파일의 노이즈 수준에 따라 결정된다. 게놈, 염색체 또는 프로파일의 분절의 길이(N_{chr})가 주어진 경우, 웨이블렛 분해 수준 c는 종종 방정식 $N_{micro} = N_{chr}/2^{c+1}$에 따라 최소 웨이블렛 길이 N_{micro}와 관련되어 있다. 일부 실시양태에서, N_{micro} 이상의 크기의 미세결실을 검출하기 위해, 원하는 분해 수준 c를 하기 방정식에 따라 측정한다: $c = \log2(N_{chr}/N_{micro}) - 1$. 예를 들면, N_{chr}가 기준 게놈의 4096개 부분이고 N_{micro}가 기준 게놈의 128개 부분인 경우, 분해 수준 c는 4이고 $c \pm 1$ 수준은 일부 경우 사용될 수 있다(즉, 약 3 내지 약 5). 일부 실시양태에서, 분해 수준 c는 약 1, 2, 3, 4, 5, 6, 7, 8, 9 또는 10이다. 일부 실시양태에서, 검출할 최소 원하는 웨이블렛 길이인 N_{micro}는 약 1 Mb, 2 Mb, 3 Mb, 4 Mb, 5 Mb, 6 Mb, 7 Mb, 8 Mb, 10 Mb 또는 15Mb이거나 약 20 Mb보다 더 크다. 일부 실시양태에서, N_{micro}는 예정된다. 일부 실시양태에서, 서열 커버리지(예를 들면, 플렉스 수준) 및 태아 분획의 양은 N_{micro}와 반비례한다. 예를 들면, 검출할 최소 원하는 웨이블렛 길이는 샘플 중의 태아 분획의 양이 증가할수록 감소한다(즉, 해상도가 증가한다). 일부 실시양태에서, 검출할 최소 원하는 웨이블렛 길이는 커버리지 증가할수록(즉, 플렉스 수준이 감소할수록) 감소한다(즉, 해상도가 증가한다). 예를 들면, 약 10%의 태아 분획을 포함하는 샘플의 경우, 4-플렉스는 약 1 Mb 이상의 N_{micro}를 생성하고, 12-플렉스는 약 3 Mb 이상의 N_{micro}를 생성한다. 일부 실시양태에서, 역치화는 평준화 전에 수행되고, 종종 역치화는 평준화 후에 수행된다.

[0395] 최대 엔트로피 분할 프로세스

[0396] 일부 실시양태에서, 적합한 분해 발생 프로세스는 최대 엔트로피 분할 프로세스를 포함한다. 일부 실시양태에서, 최대 엔트로피 분할은 분해 렌더링을 확인하는 단계를 포함한다. 일부 실시양태에서, 최대 엔트로피 분할은 서브-염색체 이상(예를 들면, 미세중복, 미세결실)의 존재 또는 부재를 확인하는 단계를 포함한다.

[0397] 일부 실시양태에서, 최대 엔트로피 분할은 게놈의 분절(예를 들면, 부분 세트, 프로파일)을 반복적으로 분할하

는 단계를 포함한다. 일부 실시양태에서, 최대 엔트로피 분할 프로세스는 수준(예를 들면, 게놈 구획 수준)에 따라 게놈의 분절을 분할한다. 일부 실시양태에서, 최대 엔트로피 분할은 프로파일의 분할된 부분에 대한 수준을 측정하는 단계를 포함한다. 일부 실시양태에서, 최대 엔트로피 분할은 게놈의 분절을 2개의 분절들(예를 들면, 2개의 부분 세트들)로 나누고 2개의 분절들에 대한 수준을 계산한다. 일부 실시양태에서, 2개의 분절들에 대한 수준은 분할(예를 들면, 분할)이 수행되기 전 또는 후에 계산된다. 일부 실시양태에서, 분할 부위(예를 들면, 분할 부위, 나누는 부위)는 2개의 생성된 분절들의 수준 사이의 차이를 최대화하도록 선택된다. 일부 실시양태에서, 최대 엔트로피 분할은 프로파일(예를 들면, 분절) 내의 모든 가능한 분할 부위에 대한 가상 분할 이벤트로부터 생성될 2개의 가상 분절들 사이의 수준 차이를 측정하고, 수준의 최대 차이가 예측되는 부위를 선택한 후 프로파일을 2개의 분절들로 나눈다(예를 들면, 분할한다). 일부 실시양태에서, 최근에 나누어진 2개의 인접한 분절들은 적합한 통계학적 방법에 의해 유의하게 상이하거나 유의하게 상이하지 않은 것으로서 확인되고, 이러한 방법의 비한정적 예는 t-검정, t 기초 기준 등을 포함한다. 일부 실시양태에서, 최대 엔트로피 분할은 제1 부분 서브세트의 수준이 제2 부분 서브세트의 수준과 유의하게 상이할 때 제1 부분 서브세트와 제2 부분 서브세트를 분할하는 단계를 포함한다. 일부 실시양태에서, 제1 부분 서브세트와 제2 부분 서브세트는 서로 인접한다.

[0398] 일부 실시양태에서, 최근에 나누어진 2개의 인접한 분절들은 유의하게 상이한 것으로 확인되고, 각각의 분절은 최대 엔트로피 분할에 따라(예를 들면, 수준의 최대 차이를 발생시키는 분할 부위에 따라) 다시 분할된다. 일부 실시양태에서, 최대 엔트로피 분할은 부분(예를 들면, 프로파일) 세트를 반복적으로 분할하여 2개 이상의 부분 서브세트들을 제공하고, 이때 각각의 생성된 서브세트는 인접한 부분 서브세트의 수준과 유의하게 상이한 수준을 포함한다.

[0399] 일부 실시양태에서, 최대 엔트로피 분할은 하나 이상의 별개의 분절을 확인하는 단계를 포함한다. 일부 실시양태에서, 최대 엔트로피 분할은 제2 수준과 유의하게 상이한 제1 수준을 확인하는 단계를 포함한다. 별개의 분절은 종종 프로파일 내의 분절의 제2 수준(예를 들면, 기준 수준)과 유의하게 상이한 제1 수준을 갖는다. 일부 실시양태에서, 별개의 분절은 기준 수준(예를 들면, 널 수준, 널 프로파일)에 따라 확인된다. 일부 실시양태에서, 기준 수준은 전체 프로파일 또는 이의 부분의 수준이다. 일부 실시양태에서, 기준 수준은 정배수체인 것으로서 공지되어 있거나 카피 수 변이(예를 들면, 미세중복 또는 미세결실)를 결여하는 것으로서 공지되어 있는 기준 프로파일 또는 기준 프로파일의 부분(예를 들면, 분절)이다. 일부 실시양태에서, 별개의 분절은 제2 수준(예를 들면, 기준 수준)과 유의하게 상이한 제1 수준(예를 들면, 웨이블렛)을 갖고, 제2 수준은 기준 수준이다. 일부 실시양태에서, 최대 엔트로피 분할은 확인된 별개의 분절에 따라 및/또는 제2 수준과 유의하게 상이한 제1 수준에 따라 거짓 음성 및 거짓 양성 확인을 감소시키면서 샘플에 대해 태아에서의 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 단계를 포함한다.

[0400] 일부 실시양태에서, 최대 엔트로피 분할은 분할된(예를 들면, 나누어진) 2개의 부분 서브세트들을 재연결하는 단계를 포함한다. 일부 실시양태에서, 나누어진 2개의 분절들은 유의하게 상이하지 않고 상기 2개의 분절들은 재연결된다. 일부 실시양태에서, 분할된 2개의 부분 서브세트들 각각의 수준은 (예를 들면, 미리 정해진 역치, 예를 들면, Z-점수 및/또는 불확실성 수준, 예를 들면, MAD에 따라) 유의하게 상이하지 않고, 상기 서브세트들은 재연결된다. 일부 실시양태에서, 재연결된 분절들은 다시 분할되지 않는다.

[0401] 일부 실시양태에서, 결정 분석은 2개 이상의 분해 렌더링을 생성하는 2개 이상의 분할 프로세스를 포함한다. 일부 실시양태에서, 결정 분석은 분해 렌더링을 독립적으로 생성하는 2개 이상의 상이한 분할 프로세스(예를 들면, 분해 발생 프로세스)를 이용하는 단계를 포함한다. 일부 실시양태에서, 결정 분석은 제1 분할 프로세스 및 제2 분할 프로세스를 포함하고, 제1 분할 프로세스 및 제2 분할 프로세스는 동시에 수행된다. 일부 실시양태에서, 제1 분할 프로세스 및 제2 분할 프로세스는 연속적으로 수행된다. 일부 실시양태에서, 결정 분석은 분석된 샘플 및 이용된 분할 프로세스의 종류에 따라 실질적으로 동일하거나 상이한 분해 렌더링을 독립적으로 생성하는 2개 이상의 상이한 분할 프로세스를 포함한다. 일부 실시양태에서, 제1 분할 프로세스는 웨이블렛 분할 프로세스(예를 들면, Haar 웨이블렛 프로세스)를 포함하고, 제2 분할 프로세스는 서큘러 이진 분할 프로세스를 포함한다.

[0402] 폴리싱(Polishing)

[0403] 일부 실시양태에서, 분해 렌더링을 폴리싱하여 폴리싱된 분해 렌더링을 제공한다. 일부 실시양태에서, 분해 렌더링은 2회 이상 폴리싱된다. 일부 실시양태에서, 분해 렌더링은 분할 프로세스의 하나 이상의 단계 전 및/또는 후에 폴리싱된다. 일부 실시양태에서, 결정 분석은 2개 이상의 분할 프로세스를 포함하고, 각각의 분할 프로세

스는 하나 이상의 폴리싱 프로세스를 포함한다. 분해 렌더링은 폴리싱된 분해 렌더링 또는 폴리싱되지 않은 분해 렌더링을 지칭할 수 있다.

[0404] 따라서, 일부 실시양태에서, 분할 프로세스는 폴리싱을 포함한다. 일부 실시양태에서, 폴리싱 프로세스는 (예를 들면, 분해 렌더링에서) 2개 이상의 실질적으로 유사한 별개의 분절들 또는 웨이블렛들을 확인하고 이들을 단일 별개의 분절 또는 웨이블렛으로 병합한다(예를 들면, 도 4). 일부 실시양태에서, 폴리싱 프로세스는 실질적으로 유사한 2개 이상의 인접한 분절들 또는 웨이블렛들을 확인하고 이들을 단일 수준, 분절 또는 웨이블렛으로 병합한다. 따라서, 일부 실시양태에서, 폴리싱 프로세스는 병합 프로세스를 포함한다. 일부 실시양태에서, 인접한 단편화된 별개의 분절들 또는 웨이블렛들은 그들의 게놈 구획 수준에 따라 병합된다. 일부 실시양태에서, 2개 이상의 인접한 별개의 분절들 또는 웨이블렛들의 병합은 궁극적으로 병합되는 2개 이상의 인접한 별개의 분절들 또는 웨이블렛들에 대한 중간 수준을 계산하는 단계를 포함한다. 일부 실시양태에서, 실질적으로 유사한 2개 이상의 인접한 별개의 분절들 또는 웨이블렛들은 병합되고 폴리싱되어 단일 분절, 웨이블렛 또는 수준을 생성한다. 일부 실시양태에서, 2개 이상의 인접한 별개의 분절들 또는 웨이블렛들은 문헌(Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics (2005) Nov 15;21(22):4084-91)에 기재된 프로세스에 의해 병합된다. 일부 실시양태에서, 2개 이상의 인접한 별개의 분절들 또는 웨이블렛들은 GLAD로서 공지되어 있고 문헌(Hupe,P. et al. (2004) "Analysis of array CGH data: from signal ratio to gain and loss of DNA regions", Bioinformatics, 20, 3413-3422)에 기재되어 있는 프로세스에 의해 병합된다.

[0405] 후보 분절 또는 웨이블렛 이벤트의 확인

[0406] 일부 실시양태에서, 결정 분석은 분해 렌더링에서 후보 분절 또는 웨이블렛 이벤트를 확인하는 단계를 포함한다. 후보 분절은 분해 렌더링에서 가장 유의한 별개의 분절인 것으로서 확인되고, 웨이블렛 이벤트는 웨이블렛 분해 렌더링에서 확인된 가장 유의한 웨이블렛인 것으로서 확인된다. "후보 분절"은 또한 임의의 종류의 분할 프로세스 및 분해 렌더링을 이용하는 분할로부터 생성된 분해 렌더링 내의 가장 유의한 별개의 분절이다. 후보 분절은 웨이블렛 분할 프로세스가 이용될 때 "웨이블렛 이벤트"와 동의어이다. 후보 분절은 일반적으로 분해 렌더링에서 가장 유의한 별개의 분절이고, 종종 분절에 의해 커버된 부분(예를 들면, 빈)의 수 및/또는 분절에 대한 정규화된 카운트의 수준의 절대 값의 관점에서 가장 유의하다. 후보 분절은 종종 분해 렌더링에서 다른 별개의 분절보다 더 크고 종종 실질적으로 더 크다. 일부 실시양태에서, 1개의 후보 분절만이 분해 렌더링에서 확인된다. 일부 실시양태에서, 하나 이상의 별개의 분절이 분해 렌더링에서 확인되고, 하나 이상의 별개의 분절 중 하나는 후보 분절로서 확인된다. 일부 실시양태에서, 후보 분절은 제2 별개의 분절의 수준보다 실질적으로 더 큰 수준을 갖는 제1 별개의 분절이고, 이때 제1 별개의 수준이 분해 렌더링에서 가장 큰 수준이다. 후보 분절은 적합한 방법에 의해 확인될 수 있다. 일부 실시양태에서, 후보 분절은 곡선하면적(AUC) 분석에 의해 확인된다. 일부 실시양태에서, 결정 분석은 AUC 분석을 포함한다. 일부 실시양태에서, 제1 별개의 분절이 수준을 갖고/갖거나 분해 렌더링에서 또 다른 별개의 분절에 대한 부분보다 실질적으로 더 큰 다수의 부분들을 커버하는 일부 실시양태에서, 제1 분절은 더 큰 AUC를 포함한다. 수준이 AUC에 대해 분석되는 경우, 수준의 절대 값이 종종 사용된다(예를 들면, 정규화된 카운트에 상응하는 수준은 결실에 대해 음의 값을 가질 수 있고 중복에 대해 양의 값을 가질 수 있다). 일부 실시양태에서, AUC는 계산된 AUC의 절대 값(예를 들면, 수득된 양의 값)으로서 측정된다. 일부 실시양태에서, 후보 분절은 (예를 들면, AUC 분석 또는 적합한 방법에 의해) 일단 확인되면 임의적으로 검증된 후 후보 분절이 유전적 변이(예를 들면, 이배수성, 미세결실 또는 미세중복)를 표시하는지를 확인하기 위한 z-점수 계산 등을 위해 선택된다.

[0407] 로그 승산비 분석

[0408] 승산비 또는 로그 승산비(LOR)는 종종 비교에서의 사용 및/또는 샘플에 대한 결정(예를 들면, 유전적 변이의 존재 또는 부재의 결정)에서의 사용을 위해 계산된다. LOR은 종종 (A) 및 (B)의 몫(quotient)의 로그로서 계산되고, 이때 (A)는 (1) 유전적 변이를 가질 조건부 확률과 (2) 유전적 변이를 가질 선험적 확률의 제1 곱셈 결과이고, (b)는 (1) 유전적 변이를 갖지 않을 조건부 확률과 (2) 유전적 변이를 갖지 않을 선험적 확률의 제2 곱셈 결과이다. 유전적 변이는 종종 염색체 이배수성(예를 들면, 1개, 3개 또는 4개 카피의 전체 염색체), 미세결실 또는 미세삽입이다.

[0409] LOR 계산은 종종 검사 샘플에 대해 측정된 태아 분획을 적용하는 단계를 포함하고, 종종 검사 샘플에 대해 확인된 염색체 또는 후보 분절에 대한 카운트 표시치(count representation)를 적용하는 단계를 포함한다. 일부 실시양태에서, 염색체 이배수성을 가질 조건부 확률은 태아 분획 및 카운트 표시치에 따라 결정된다. 따라서, 일

부 실시양태에서, 일부 방법은 기준 게놈의 부분들에 맵핑된 핵산 서열 리드의 카운트에 따라 염색체 카운트 표시치 및/또는 후보 분절 카운트 표시치를 측정하는 단계를 포함하고, 이때 서열 리드는 종종 태아를 가진 임신 여성으로부터의 검사 샘플에 대한 순환 무세포 핵산의 리드이다. 후보 분절은 종종 검증된 후보 분절이다(본원에 기재됨).

[0410] 염색체 카운트 표시치는 종종 염색체(예를 들면, 모든 상염색체)보다 더 큰 게놈 또는 이의 서브세트의 부분에서의 카운트로 나누어진, 염색체 내의 부분(예를 들면, 빈)으로 맵핑된 카운트이다. 염색체 카운트 표시치는 종종 정량되고, 임의의 적합한 정량이 이용될 수 있다(예를 들면, z-점수). z-점수가 염색체 카운트 표시치를 정량하는 실시양태의 경우, z-점수는 종종 값(B)로 나누어진 뺄셈 결과(A)이다. 뺄셈 결과(A)는 종종 (i) 검사 샘플 염색체 카운트 표시치로부터 (ii) 정배수체 염색체 카운트 표시치의 중간치를 뺀 결과이다. 값(B)는 종종 정배수체 염색체 카운트 표시치의 MAD이다. 검사 샘플 염색체 카운트 표시치는 종종 검사 샘플에 대한 (a) 상기 염색체 내의 부분에서의 카운트 대 (b) 상염색체 내의 부분에서의 카운트의 비이다. 정배수체 염색체 카운트 표시치의 중간치는 종종 정배수체에 대한 (a) 상기 염색체 내의 부분에서의 카운트 대 (b) 상염색체 내의 부분에서의 카운트의 비의 중간치이다. 카운트는 종종 정규화된 카운트이고, 이때 게놈 부분들에 맵핑된 카운트는 하나 이상의 적합한 정규화 프로세스에 의해 정규화될 수 있다. 이용될 수 있는 정규화 프로세스의 비한정적 예는 당분야에서 공지되어 있고 본원에 기재되어 있다(예를 들면, LOESS, GC-LOESS, PERUN, ChAI, 주성분 정규화 프로세스).

[0411] 후보 분절 카운트 표시치는 종종 후보 분절(예를 들면, 모든 상염색체)보다 더 큰 게놈의 부분 또는 이의 서브세트에서의 카운트로 나누어진, 후보 분절 내에 있거나 후보 분절에 의해 커버된 부분(예를 들면, 빈)에 맵핑된 카운트이다. 후보 분절 카운트 표시치는 종종 정량되고, 임의의 적합한 정량이 이용될 수 있다(예를 들면, z-점수). z-점수가 후보 분절 카운트 표시치를 정량하는 실시양태의 경우, z-점수는 종종 값(B)로 나누어진 뺄셈 결과(A)이다. 뺄셈 결과(A)는 종종 (i) 검사 샘플 후보 분절 카운트 표시치로부터 (ii) 정배수체 후보 분절 카운트 표시치의 중간치를 뺀 결과이다. 값(B)는 종종 정배수체 후보 분절 카운트 표시치의 MAD이다. 검사 샘플 후보 분절 카운트 표시치는 종종 검사 샘플에 대한 (a) 후보 분절 내의 부분에서의 카운트 대 (b) 상염색체 내의 부분에서의 카운트의 비이다. 정배수체 후보 분절 카운트 표시치의 중간치는 종종 정배수체에 대한 (a) 후보 분절 내의 부분에서의 카운트 대 (b) 상염색체 내의 부분에서의 카운트의 비의 중간치이다. 카운트는 종종 정규화된 카운트이고, 이때 게놈 부분들에 맵핑된 카운트는 하나 이상의 적합한 정규화 프로세스에 의해 정규화될 수 있다. 이용될 수 있는 정규화 프로세스의 비한정적 예는 당분야에서 공지되어 있고 본원에 기재되어 있다(예를 들면, LOESS, GC-LOESS, PERUN, ChAI, 주성분 정규화 프로세스).

[0412] LOR 계산을 포함하는 방법은 종종 검사 샘플에 대한 태아 분획을 측정하는 단계를 포함한다. 태아 분획은 당분야에서 공지된 임의의 적합한 방법을 이용함으로써 측정될 수 있고, 이 방법의 비한정적 예는 본원에 기재되어 있다(예를 들면, Y 염색체 좌위(예를 들면, SRY 좌위) 정량, FRS 정량).

[0413] 일부 LOR 계산 실시양태에서, 유전적 변이를 가질 조건부 확률은 검사 샘플에 대해 측정된 태아 분획, 검사 샘플에 대한 염색체 카운트 표시치 또는 후보 분절 카운트 표시치에 대한 z-점수, 및 염색체 카운트 표시치 또는 후보 분절 카운트 표시치에 대한 z-점수의 태아 분획 특이적 분포에 따라 평가된다. 일부 실시양태에서, 유전적 변이를 가질 조건부 확률은 하기 실시예 6에 기재된 방정식 23에서의 관계에 의해 측정되고, 이때 f는 태아 분획, X는 염색체 또는 후보 분절에 대한 합산된 부분, $X \sim f(\mu_X, \sigma_X)$이고, 이때 μ_X 및 σ_X는 각각 X의 평균 및 표준 편차이고, $f(\cdot)$는 분포 함수이다. 유전적 변이를 가질 조건부 확률은 종종 검사 샘플 염색체 카운트 표시치 또는 후보 분절 카운트 표시치에 대한 z-점수와 염색체 카운트 표시치 또는 후보 분절 카운트 표시치에 대한 z-점수의 태아 분획 특이적 분포 사이의 교차점이다(예를 들면, T21 예에 대한 도 32 참조). 실시예 6은 도 32와 관련하여 미세중복 이벤트 또는 미세결실 이벤트의 존재 또는 부재가 확인되는 상황에서 정배수체 분포에 대한 분포 변동을 기술한다.

[0414] 염색체 이배수성을 갖지 않을 조건부 확률은 염색체 카운트 표시치 또는 후보 분절 카운트 표시치, 및 정배수체에 대한 카운트 표시치에 따라 결정된다. 유전적 변이를 갖지 않을 조건부 확률은 종종 정배수체에서 염색체 카운트 표시치의 z-점수와 염색체 카운트 표시치에 대한 z-점수의 분포 사이의 교차점이다(예를 들면, T21 예에 대한 도 32 참조).

[0415] 유전적 변이를 가질 선험적 확률 및 유전적 변이를 갖지 않을 선험적 확률은 종종 예를 들면, 당분야에서 공지된 통계 데이터를 사용함으로써 하나 이상의 환자 집단에 대해 측정된다. 예를 들면, T21이 발생할 확률 및 T21이 발생하지 않을 확률은 특정 지리학적 지역 내의 집단에 대해 용이하게 측정될 수 있다. 선험적 확률은 종종

검사 대상체를 포함하지 않는 다수의 샘플들로부터 측정된다.

[0416] 비교 및 결정 분석

[0417] 일부 실시양태에서, 결정 분석은 비교를 포함한다. 일부 실시양태에서, 비교는 2개 이상의 분해 렌더링들을 비교하는 단계를 포함한다. 일부 실시양태에서, 비교는 2개 이상의 후보 분절들을 비교하는 단계를 포함한다. 일부 실시양태에서, 2개 이상의 후보 분절들 각각은 상이한 분해 렌더링으로부터의 후보 분절들이다. 예를 들면, 제1 후보 분절은 제1 분해 렌더링으로부터의 후보 분절일 수 있고, 제2 후보 분절은 제2 분해 렌더링으로부터의 후보 분절일 수 있다. 일부 실시양태에서, 비교는 2개 이상의 분해 렌더링들이 실질적으로 동일한지 아니면 상이한지를 확인하는 단계를 포함한다. 일부 실시양태에서, 비교는 2개 이상의 후보 분절들이 실질적으로 동일한지 아니면 상이한지를 확인하는 단계를 포함한다.

[0418] 일부 실시양태에서, 2개의 분해 렌더링들은 각각의 렌더링이 후보 분절을 포함하고 각각의 분해 렌더링으로부터의 후보 분절이 실질적으로 동일한 것으로서 확인될 때 실질적으로 동일하다. 2개의 후보 분절들은 적합한 비교 방법에 의해 실질적으로 동일하거나 상이한 것으로 확인될 수 있고, 이러한 방법의 비한정적 예는 시각적 검사, 2개의 후보 분절들의 수준 또는 Z-점수의 비교, 2개의 후보 분절들의 에지의 비교, 2개의 후보 분절들 또는 이들의 상응하는 분해 렌더링들의 중첩 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, 2개의 후보 분절들의 에지는 실질적으로 동일하고, 2개의 후보 분절들은 실질적으로 동일하다. 일부 실시양태에서, 후보 분절의 에지는 또 다른 후보 분절의 에지와 실질적으로 동일하고, 2개의 에지는 10개 미만, 9개 미만, 8개 미만, 7개 미만, 6개 미만, 5개 미만, 4개 미만, 3개 미만, 2개 미만 또는 1개 미만의 부분(예를 들면, 빈)에 의해 분리된다. 일부 실시양태에서, 2개의 에지는 실질적으로 동일하고 동일한 위치(예를 들면, 동일한 부분)에 존재한다. 일부 실시양태에서, 실질적으로 동일한 2개의 후보 분절들은 (예를 들면, 불확실성 수준 내에서, 예를 들면, 불확실성 수준의 약 3배, 2배 또는 1배 이하에서) 실질적으로 동일한 수준, Z-점수 등을 포함한다. 일부 실시양태에서, 2개의 후보 분절들은 실질적으로 상이한 에지 및/또는 실질적으로 상이한 수준을 포함하고 비교에 따라 실질적으로 동일하지 않은(예를 들면, 상이한) 것으로서 확인된다.

[0419] 일부 실시양태에서, 비교는 하나 이상의 복합 후보 분절을 생성하는 단계 및 하나 이상의 복합 후보 분절의 비교를 포함하는 비교를 기초로(예를 들면, 부분적으로 또는 전적으로 기초로) 이배수성, 미세결실 또는 미세중복의 존재 또는 부재를 확인하는 단계를 포함한다. 복합 후보 분절은 임의의 적합한 방법에 의해 생성될 수 있다. 일부 실시양태에서, 복합 후보 분절은 2개 이상의 후보 분절(예를 들면, 수준, AUC 및/또는 에지)의 평균을 산출함으로써 생성된다. 일부 실시양태에서, 복합 후보 분절은 2개 이상의 후보 분절들을 중첩시킴으로써 생성된다. 일부 실시양태에서, 2개 이상의 후보 분절들은 실질적으로 동일하고 복합 후보 분절은 생성된다(예를 들면, 도 11).

[0420] 비교는 종종 이하에 기재된 바와 같이 2개 이상의 분해 렌더링으로부터 유도된 후보 분절(예를 들면, 웨이블렛 이벤트)을 정량하는 단계 및 비교를 이용하여 샘플에서 유전적 변이(예를 들면, 염색체 이배수성, 미세중복 또는 미세결실)의 존재 또는 부재를 확인하는 단계를 포함한다.

[0421] 일부 실시양태에서, 비교는 2개 이상의 분해 렌더링에서 확인된 후보 분절(예를 들면, 웨이블렛 이벤트)로부터 복합 후보 분절(예를 들면, 복합 웨이블렛 이벤트)의 존재 또는 부재를 확인하는 단계를 포함한다. 일부 실시양태에서, 2개 이상의 후보 분절들(예를 들면, 2개 이상의 분해 렌더링들로부터 유도된 웨이블렛 이벤트들)은 중첩되거나 실질적으로 동일하고 복합 후보 분절(예를 들면, 복합 웨이블렛 이벤트)의 존재는 확인된다(도 11). 복합 웨이블렛 이벤트의 존재 또는 부재는 임의의 적합한 방법에 의해 확인될 수 있다. 일부 실시양태에서, 복합 후보 분절(예를 들면, 복합 웨이블렛 이벤트)의 존재 또는 부재는 2개 이상의 후보 분절(예를 들면, 복합 웨이블렛 이벤트, 예를 들면, 수준, AUC 및/또는 에지)의 평균을 산출함으로써 확인된다. 일부 실시양태에서, 복합 후보 분절(예를 들면, 복합 웨이블렛 이벤트)의 존재 또는 부재는 2개 이상의 후보 분절들(예를 들면, 웨이블렛 이벤트들)을 중첩시킴으로써 확인된다. 일부 실시양태에서, 복합 후보 분절(예를 들면, 복합 웨이블렛 이벤트)의 존재는 2개 이상의 후보 분절들(예를 들면, 웨이블렛 이벤트들)이 중첩되거나 실질적으로 동일할 때 확인된다.

[0422] 일부 실시양태에서, 2개 이상의 후보 분절들(예를 들면, 2개 이상의 분해 렌더링들로부터 유도된, 예를 들면, 복합 웨이블렛 이벤트들)은 중첩되지 않거나 상이하고(예를 들면, 실질적으로 상이하고) 복합 후보 분절의 부재(예를 들면, 복합 웨이블렛 이벤트의 부재)는 확인된다. 일부 실시양태에서, 복합 후보 분절(예를 들면, 복합 웨이블렛 이벤트)의 부재는 염색체 이배수성, 미세중복 또는 미세결실의 부재를 표시한다.

[0423] 일부 실시양태에서, 결정 분석은 결과를 확인하는(예를 들면, 태아에서의 유전적 변이의 존재 또는 부재를 확인하는) 단계를 포함한다. 일부 실시양태에서, 결정 분석은 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 방법을 포함한다. 일부 실시양태에서, 결정 분석은 본원에 기재된 결정 분석을 이용하지 않고(예를 들면, 분할, 하나 이상의 후보 분절의 존재 또는 부재의 확인 및/또는 하나 이상의 후보 분절의 정량 없이) 유전적 변이의 존재 또는 부재를 확인할 때에 비해 거짓 음성 및 거짓 양성 확인을 감소시키면서 (예를 들면, 태아에서) 유전적 변이의 존재 또는 부재를 확인하는 방법을 포함한다. 일부 실시양태에서, 결정 분석은 일련의 방법 또는 방법 단계들을 포함한다. 결정 분석의 비한정적 예는 도 6 내지 8에 제시되어 있고 본원에 기재되어 있다. 일부 실시양태에서, 결정 분석은 카운트를 수득하는 단계 및 프로파일을 생성하고/하거나 수득하는 단계를 포함한다. 일부 실시양태에서, 결정 분석은 프로파일을 분할하는 단계 및 분해 렌더링을 생성하는 단계를 포함한다. 일부 실시양태에서, 분해 렌더링 또는 이의 분절(예를 들면, 염색체를 표시하는 분절, 수준, 별개의 분절 또는 웨이블렛, 후보 분절 또는 웨이블렛 이벤트, 복합 분절 또는 복합 웨이블렛)은 적합한 방법에 의해 정량된다. 적합한 정량 방법의 비한정적 예는 당분야에서 공지되어 있고 부분적으로 본원에 기재되어 있고 예를 들면, Z-점수, p-값, t-값, 수준 또는 수준, AUC, 배수성, 불확실성 수준 등 또는 이들의 조합을 측정하는 방법을 포함한다.

[0424] 일부 실시양태에서, 결정 분석은 2개 이상의 분할 방법으로 프로파일을 분할하는 단계를 포함한다. 일부 실시양태에서, 결정 분석은 50개 이상의 분할 방법을 포함한다. 일부 실시양태에서, 결정 분석은 50개 미만, 40개 미만, 30개 미만, 20개 미만, 10개 미만 또는 약 5개 미만의 분할 방법을 포함한다. 일부 실시양태에서, 결정 분석은 약 10개, 9개, 8개, 7개, 6개, 5개, 4개, 3개 또는 2개의 분할 방법을 포함한다. 일부 실시양태에서, 각각의 분할 방법(예를 들면, 2개의 방법이 이용되는 경우, 예를 들면, 도 6a, 611 및 612)은 프로파일의 분해 렌더링을 제공한다. 일부 실시양태에서, 2개 이상의 분할 방법에 의해 제공된 분해 렌더링은 동일하거나, 실질적으로 동일하거나 상이하다.

[0425] 일부 실시양태에서, 분할 후 폴리싱(예를 들면, 도 6a, 621 및 622; 도 6b, 623)이 수행된다. 일부 실시양태에서, 하나 이상의 분할 프로세스로부터 유도된 하나 이상의 분해 렌더링은 종종 동일한 폴리싱 방법에 의해 폴리싱된다. 일부 실시양태에서, 하나 이상의 분할 단계로부터 유도된 하나 이상의 분해 렌더링은 상이한 폴리싱 방법에 의해 폴리싱된다. 일부 실시양태에서, 분해 렌더링은 1개, 2개 또는 3개 이상의 폴리싱 방법에 의해 폴리싱된다. 일부 실시양태에서, 각각의 분해 렌더링은 한 방법에 의해 폴리싱되고 이 방법은 각각의 분해 렌더링에 대해 동일하다.

[0426] 일부 실시양태에서, 후보 분절(예를 들면, 웨이블렛 이벤트)의 존재 또는 부재는 분할 및 임의적으로 폴리싱 후 확인된다(예를 들면, 도 6a, 631 및 632; 도 6b, 623). 일부 실시양태에서, 폴리싱 프로세스는 생략되고 후보 분절(예를 들면, 웨이블렛 이벤트)은 분할로부터 유도된 분해 렌더링으로부터 직접적으로 확인된다. 일부 실시양태에서, 후보 분절(예를 들면, 웨이블렛 이벤트)은 폴리싱된 분해 렌더링에서 및/또는 폴리싱된 분해 렌더링으로부터 확인된다. 일부 실시양태에서, 후보 분절(예를 들면, 웨이블렛 이벤트)은 하나 이상의 분해 렌더링에서 확인되지 않고, 유전적 변이의 부재가 확인된다. 일부 실시양태에서, 후보 분절(예를 들면, 웨이블렛 이벤트)이 하나 이상의 분해 렌더링(예를 들면, 폴리싱된 분해 렌더링) 중 하나에서 확인되지 않는 경우, 결정 분석은 종결된다.

[0427] 일부 실시양태에서, 후보 분절(예를 들면, 웨이블렛 이벤트)은 일단 확인되면 정량된다(예를 들면, 도 6a, 641 및 642; 도 6b, 644(예를 들면, z-점수 또는 LOR 정량)). 후보 분절(예를 들면, 웨이블렛 이벤트)은 적합한 방법에 의해 정량될 수 있고, 이러한 방법의 비한정적 예는 Z-점수의 계산, p-값의 계산, t-값의 측정, 수준의 측정, 배수성의 측정, 계산된 불확실성 수준 등 또는 이들의 조합을 포함한다.

[0428] 일부 실시양태에서, 결정 분석은 비교를 포함한다(예를 들면, 도 6a, 6b 및 8에서 650, 651, 810). 일부 실시양태에서, 정량 후 비교가 수행된다(예를 들면, 도 6a, 641, 642 및 643; 도 6b, 651). 일부 실시양태에서, 웨이블렛 또는 후보 분절 확인 후 비교가 수행된다(예를 들면, 도 6a, 631 및 632; 도 6b, 633). 종종, 염색체 정량 후 비교가 수행된다(예를 들면, 도 6a, 643; 도 6b, 645(예를 들면, z-점수 또는 LOR 정량)). 일부 실시양태에서, 비교 후 결정이 수행된다(예를 들면, 도 6a, 660; 도 6b, 661).

[0429] 일부 실시양태에서, 검증된 후보 분절을 포함하는 후보 분절("후보 분절"로서 총칭됨)이 정량된다. 본원에 기재된 바와 같이, 후보 분절은 종종 후보 분절 카운트 표시치로서 정량되고, 후보 분절 카운트 표시치는 종종 z-점수에 의해 정량된다. 염색체 카운트 표시치는 종종 후보 분절이 위치하는 염색체에 대해 생성되고 정량된다. 염색체 카운트 표시치는 본원에 기재되어 있고, 마찬가지로 본원에 기재되어 있는 z-점수에 의해 정량될 수 있다.

후보 분절 카운트 표시치 및/또는 염색체 카운트 표시치에 대한 카운트는 본원에 기재된 바와 같이 정규화된 카운트이다.

[0430] 일부 실시양태에서, 제1 후보 분절 카운트 표시치의 z-점수 정량치가 생성되고, 제2 후보 분절 카운트 표시치의 z-점수 정량치가 생성되고, 이때 제1 후보 분절 및 제2 후보 분절은 상이한 2종의 분할로부터 확인된다. 일부 실시양태는 (i) 1 미만의 계수(예를 들면, 약 0.6 내지 약 0.8)와 곱해진 제1 후보 분절 카운트 표시치의 z-점수 정량치 및 (ii) 상기 계수와 곱해진 제2 후보 분절 카운트 표시치의 z-점수 정량치의 최소치를 측정하는 것을 포함한다.

[0431] 일부 실시양태에서, 후보 분절 카운트 표시치의 정량치는 후보 분절이 위치하는 염색체에 대한 염색체 카운트 표시치의 정량치와 비교된다. 일부 실시양태는 염색체 표시치의 z-점수 정량치가 이전 단락에서 언급된 최소치보다 더 작은지, 이 최소치보다 더 큰지 아니면 이 최소치와 동등한지를 확인하는 것을 포함한다. 일부 실시양태는 염색체 카운트 표시치의 z-점수 정량치가 역치 z-점수 값(예를 들면, 약 3.95(예를 들면, 약 3.5 내지 약 4.5)의 값)보다 더 작은지, 이 값보다 더 큰지 아니면 이 값과 동등한지를 확인하는 것을 포함한다.

[0432] 일부 실시양태는 검사 샘플에 대해 (i) 염색체 카운트 표시치의 z-점수 정량치가 역치 z-점수 값(예를 들면, 약 3.95(예를 들면, 약 3.5 내지 약 4.5)의 값) 이상이고 (ii) 염색체 카운트 표시치의 z-점수 정량치가 이전 단락에서 언급된 최소치 이상인 경우 염색체 이배수성의 존재를 확인하는 것을 포함한다. 일부 실시양태는 검사 샘플에 대해 (i) 염색체 카운트 표시치의 z-점수 정량치가 역치 z-점수 값(예를 들면, 약 3.95(예를 들면, 약 3.5 내지 약 4.5)의 값)보다 더 작고/작거나, (ii) 염색체 카운트 표시치의 z-점수 정량치가 상기 최소치보다 더 작은 경우 염색체 이배수성의 부재를 확인하는 것을 포함한다. 염색체 이배수성은 종종 삼염색체성 또는 일염색체성이고, 종종 1개, 3개 또는 4개 염색체의 존재이다.

[0433] 일부 실시양태는 제1 후보 분절 카운트 표시치의 z-점수 정량치가 역치 z-점수 값(예를 들면, 약 3.95(예를 들면, 약 3.5 내지 약 4.5)의 값)보다 더 작은지, 이 값보다 더 큰지 아니면 이 값과 동등한지를 확인하는 것 및 제2 후보 분절 카운트 표시치의 z-점수 정량치가 역치 z-점수 값(예를 들면, 약 3.95(예를 들면, 약 3.5 내지 약 4.5)의 값)보다 더 작은지, 이 값보다 더 큰지 아니면 이 값과 동등한지를 확인하는 것을 포함한다. 일부 실시양태는 제1 후보 분절과 제2 후보 분절이 실질적으로 동일한지 아니면 중첩되는지를 확인하는 것을 포함한다.

[0434] 일부 실시양태는 검사 샘플에 대해 (i) 제1 후보 분절 카운트 표시치의 z-점수 정량치가 역치 z-점수 값(예를 들면, 약 3.95(예를 들면, 약 3.5 내지 약 4.5)의 값) 이상이고 제2 후보 분절 카운트 표시치의 z-점수 정량치가 역치 z-점수 값(예를 들면, 약 3.95(예를 들면, 약 3.5 내지 약 4.5)의 값) 이상이고, (ii) 제1 후보 분절과 제2 후보 분절이 실질적으로 동일하거나 중첩되는 경우 미세결실 또는 미세삽입의 존재를 확인하는 것을 포함한다. 일부 실시양태는 검사 샘플에 대해 (i) 제1 후보 분절 카운트 표시치의 z-점수 정량치가 역치 z-점수 값(예를 들면, 약 3.95(예를 들면, 약 3.5 내지 약 4.5)의 값)보다 더 작고/작거나 제2 후보 분절 카운트 표시치의 z-점수 정량치가 역치 z-점수 값(예를 들면, 약 3.95(예를 들면, 약 3.5 내지 약 4.5)의 값)보다 더 작고/작거나, (ii) 제1 후보 분절과 제2 후보 분절이 실질적으로 동일하지 않거나 중첩되지 않는 경우 미세결실 또는 미세삽입의 부재를 확인하는 것을 포함한다.

[0435] 일부 실시양태에서, 비교는 2개 이상의 값(예를 들면, 정량, 예를 들면, 프로파일의 정량 및/또는 후보 분절(예를 들면, 웨이블렛 이벤트)의 정량으로부터 유도된 값)을 비교한다. 일부 실시양태에서, 비교는 후보 분절(예를 들면, 웨이블렛 이벤트) 또는 프로파일의 정량치를 예정된 값 또는 역치와 비교한다. 비교의 비한정적 예는 도 7에 제시되어 있다. 일부 실시양태에서, 비교는 Z-점수의 비교를 포함한다. 일부 실시양태에서, 비교는 전체 염색체에 대한 전체 염색체 표시치(염색체의 프로파일)에 대한 Z-점수의 절대 값(즉, $|Z_{chr}|$)의 비교를 포함한다. 값 $|Z_{chr}|$는 종종 예정된 값, 역치 또는 비교 특징(예를 들면, 도 7의 710에서 역치 3.95)과 비교된다. 일부 실시양태에서, Z-점수의 비교를 위해 사용된 예정된 값 또는 비교 특징은 약 2.8, 2.9, 3.0, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.75, 3.8, 3.85, 3.9, 3.95, 4.0, 4.05, 4.1, 4.15, 4.2, 4.3, 4.4 또는 약 4.5이다. 값 $|Z_{chr}|$는 종종 분해 렌더링으로부터의 후보 분절 및 그의 부분 카운트 표시치에 대한 Z-점수의 절대 값(예를 들면, 실시예 3에서 $|Z_{wave}|$ 및 $|Z_{cbs}|$, 및 도 7에서 $|Z_{A4}|$ 및 $|Z_{B4}|$)과 비교된다.

[0436] 일부 실시양태에서, 비교의 결과는 결과의 또 다른 비교 또는 결정을 위한 결정이다. 일부 실시양태에서, 제1 비교의 결과(예를 들면, 도 7, 710)는 일련의 비교에서 다음 비교를 결정하는 결정이다. 예를 들면, 제1 비교(예를 들면, 도 7, 710)는 $|Z_{chr}|$이 예정된 값 이상이라고 결정할 수 있고, 제2 비교(예를 들면, 도 7, 721)는

$|Z_{chr}|$을 $|Z_{A4}|$ 및/또는 $|Z_{B4}|$와 비교한다. 대안적으로, 제1 비교(예를 들면, 도 7, 710)는 $|Z_{chr}|$이 예정된 값보다 더 작다고 결정할 수 있고, 제2 비교(예를 들면, 도 7, 722)는 결정 분석(예를 들면, 도 6a, 631 및 632)에서 이미 확인된 후보 분절(예를 들면, 웨이블렛 이벤트)이 실질적으로 동일한지 아니면 상이한지를 결정한다.

[0437] 일부 실시양태에서, 제1 비교(예를 들면, 도 7, 710)의 결과는 제2 비교를 연속적으로 결정하는 결정이고, 제2 비교로부터 유도된 결정은 제3 비교 등을 결정한다. 일부 실시양태에서, 제1 비교는 $|Z_{chr}|$이 예정된 값 이상이라고 결정할 수 있고, 제2 비교(예를 들면, 도 7, 721)는 $|Z_{chr}|$이 $|Z_{A4}|$ 및/또는 $|Z_{B4}|$ 또는 이의 분율(예를 들면, 예정된 값 α와 곱해진 $|Z_{A4}|$ 및/또는 $|Z_{B4}|$)보다 더 크다고 결정할 수 있고, 전체 염색체 이배수성의 존재가 확인된다. 삼염색체성과 일염색체성은 적합한 방법에 의해 식별될 수 있다.

[0438] 일부 실시양태에서, 제1 비교는 $|Z_{chr}|$이 예정된 값 이상이라고 결정할 수 있고, 제2 비교(예를 들면, 도 7, 721)는 $|Z_{chr}|$이 $|Z_{A4}|$ 및/또는 $|Z_{B4}|$ 또는 이의 분율(예를 들면, 예정된 값 α와 곱해진 $|Z_{A4}|$ 및/또는 $|Z_{B4}|$)보다 더 작다고 결정할 수 있고, 제3 비교가 수행된다. 일부 실시양태에서, 제1 비교는 $|Z_{chr}|$이 예정된 값보다 더 작다고 결정할 수 있고, 제2 비교는 확인된 후보 분절(예를 들면, 웨이블렛 이벤트)이 중첩되거나 실질적으로 동일하다고(복합 후보 분절) 결정하고, 제3 비교는 $|Z_{A4}|$ 및 $|Z_{B4}|$이 예정된 값(예를 들면, 3.95) 이상이라고 결정하고, 미세중복 및/또는 미세결실의 존재가 확인된다. 미세중복과 미세결실은 적합한 방법에 의해 식별될 수 있다. 예를 들면, 미세중복은 양의 Z-점수를 가질 수 있고, 미세결실은 음의 Z-점수를 가질 수 있다.

[0439] 일부 실시양태에서, 비교는 2개 이상의 후보 분절들(예를 들면, 웨이블렛 이벤트들)이 중첩되지 않거나 실질적으로 동일하지 않고(예를 들면, 실질적으로 상이하고, 예를 들면, 도 8, 822) 유전적 변이가 프로파일에 존재하지 않는다고 결정할 수 있다. 일부 실시양태에서, 비교는 2개 이상의 후보 분절들(예를 들면, 웨이블렛 이벤트들, 예를 들면, 하나 이상의 분해 렌더링에서 확인된 모든 후보 분절들(예를 들면, 웨이블렛 이벤트들))이 중첩되거나 실질적으로 동일하다고(예를 들면, 도 8, 821) 결정할 수 있고, 미세중복 또는 미세결실의 존재 또는 부재가 확인된다. 일부 실시양태에서, 미세중복 또는 미세결실의 존재 또는 부재는 복합 후보 분절(예를 들면, 복합 웨이블렛 이벤트)의 정량에 따라 확인된다.

[0440] 일부 실시양태에서, 결정 분석은 2회 이상의 분할, 폴리싱 및 후보 분절(예를 들면, 웨이블렛 이벤트)의 확인을 포함한다. 일부 실시양태에서, 결정 분석은 2개 이상의 후보 분절들(예를 들면, 웨이블렛 이벤트들)의 정량을 포함할 수 있다. 일부 실시양태에서, 결정 분석은 염색체의 프로파일의 정량을 포함할 수 있다. 일부 실시양태에서, 결정 분석은 1회 이상의 비교를 포함한다. 일부 실시양태에서, 결정 분석은 유전적 변이의 존재 또는 부재의 확인을 포함한다.

[0441] 일부 실시양태에서, 결정 분석은 분할, 폴리싱, 후보 분절(예를 들면, 웨이블렛 이벤트)의 확인, 1회 이상의 비교 및 유전적 변이의 존재 또는 부재의 확인을 포함하고/하거나 이들로 구성된다. 일부 실시양태에서, 결정 분석은 분할, 폴리싱, 후보 분절(예를 들면, 웨이블렛 이벤트)의 확인, 정량, 1회 이상의 비교 및 유전적 변이의 존재 또는 부재의 확인을 포함하고/하거나 이들로 구성된다. 일부 실시양태에서, 결정 분석은 분할, 폴리싱, 후보 분절(예를 들면, 웨이블렛 이벤트)의 확인, 복합 후보 분절(예를 들면, 웨이블렛 이벤트)의 존재 또는 부재의 확인, 복합 후보 분절(예를 들면, 복합 웨이블렛 이벤트)의 정량, 1회 이상의 비교 및 유전적 변이의 존재 또는 부재의 확인을 포함하고/하거나 이들로 구성된다. 일부 실시양태에서, 결정 분석은 분할, 폴리싱, 후보 분절(예를 들면, 웨이블렛 이벤트)의 확인, 후보 분절(예를 들면, 웨이블렛 이벤트)의 정량, 염색체의 프로파일의 정량, 비교 및 유전적 변이의 존재 또는 부재의 확인을 포함하고/하거나 이들로 구성된다. 일부 실시양태에서, 결정 분석은 검증을 포함한다.

[0442] 일부 실시양태에서, 비교 또는 결정 분석은 승산비 또는 로그 승산비(LOR)와의 비교를 포함한다. 일부 실시양태에서 비교 또는 결정은 계산된 LOR이 0보다 더 큰지 아니면 더 작은지를 확인하는 단계를 포함한다.

[0443] 일부 실시양태에서, 비교 또는 결정은 염색체 카운트 표시치의 z-점수 정량치를 생성하는 단계 및 염색체 카운트 표시치가 값(예를 들면, 약 3.95(예를 들면, 약 3.5 내지 약 4.5)의 z-점수 값)보다 더 작은지, 이 값보다 더 큰지 아니면 이 값과 동등한지를 확인하는 단계를 포함한다. 일부 실시양태에서, 결정은 검사 샘플에 대해 (i) 염색체 카운트 표시치의 z-점수 정량치가 값(예를 들면, 약 3.95의 값) 이상이고 (ii) LOR이 0보다 더 큰 경우 염색체 이배수성의 존재를 결정하는(확인하는) 단계를 포함한다. 일부 실시양태에서, 결정은 검사 샘플에 대해 (i) 염색체 카운트 표시치의 z-점수 정량치가 값(예를 들면, 약 3.95의 값)보다 더 작고/작거나, (ii) LOR이 0보다 더 작은 경우 염색체 이배수성의 부재를 결정하는(확인하는) 단계를 포함한다. 염색체 이배수성은 종

종 삼염색체성 또는 일염색체성, 또는 1개, 3개 또는 4개 카피의 염색체이다.

[0444] 일부 실시양태에서, 비교 또는 결정은 후보 분절 카운트 표시치의 z-점수 정량치를 생성하는 단계 및 후보 분절 카운트 표시치가 값(예를 들면, 약 3.95(예를 들면, 약 3.5 내지 약 4.5)의 z-점수 값)보다 더 작은지, 이 값보다 더 큰지 아니면 이 값과 동등한지를 확인하는 단계를 포함한다. 일부 실시양태에서, 결정은 검사 샘플에 대해 (i) 후보 분절 카운트 표시치의 z-점수 정량치가 값(예를 들면, 약 3.95의 값) 이상이고 (ii) LOR이 0보다 더 큰 경우 미세결실 또는 미세삽입 이벤트의 존재를 결정하는(확인하는) 단계를 포함한다. 일부 실시양태에서, 결정은 검사 샘플에 대해 (i) 후보 분절 카운트 표시치의 z-점수 정량치가 값(예를 들면, 약 3.95의 값)보다 더 작고/작거나, (ii) LOR이 0보다 더 작은 경우 미세결실 또는 미세삽입 이벤트의 부재를 결정하는(확인하는) 단계를 포함한다. 미세결실 이벤트는 종종 디조지 증후군과 관련된 이벤트이다.

[0445] 결과

[0446] 본원에 기재된 방법은 샘플에 대한 유전적 변이(예를 들면, 태아 이배수성)의 존재 또는 부재의 확인을 제공하여 결과를 제공(예를 들면, 유전적 변이(예를 들면, 태아 이배수성)의 존재 또는 부재를 확인시켜주는 결과를 제공)할 수 있다. 유전적 변이는 종종 기준물에 비해 검사 대상체의 게놈 또는 유전 정보에서 검출가능한 변화를 초래하는, 유전 정보(예를 들면, 염색체, 염색체의 분절, 다형성 영역, 전위된 영역, 변경된 뉴클레오티드 서열 등 또는 이들의 조합)의 획득, 상실 및/또는 변경(예를 들면, 중복, 결실, 융합, 삽입, 돌연변이, 재조직화, 치환 또는 비정상적인 메틸화)을 포함한다. 유전적 변이의 존재 또는 부재는 부분들에 맵핑된 서열 리드(예를 들면, 카운트, 기준 게놈의 게놈 부분의 카운트)를 변환시키고/시키거나, 분석하고/하거나 조작함으로써 확인될 수 있다. 일부 실시양태에서, 결과의 확인은 임신 여성으로부터의 핵산을 분석하는 단계를 포함한다. 일부 실시양태에서, 결과는 임신 여성으로부터 수득된 카운트(예를 들면, 정규화된 카운트)에 따라 확인되고, 이때 카운트는 임신 여성으로부터 수득된 핵산으로부터의 카운트이다.

[0447] 본원에 기재된 방법은 종종 태아를 가진 임신 여성으로부터의 검사 샘플에 대한 태아 이배수성(예를 들면, 전체적인 염색체 이배수성, 부분적인 염색체 이배수성 또는 분절 염색체 이상(예를 들면, 모자이크 현상, 결실 및/또는 삽입))의 존재 또는 부재를 확인한다. 일부 실시양태에서, 본원에 기재된 방법은 태아를 가진 임신 여성으로부터의 샘플에 대한 정배수성 또는 정배수성의 결여(비-정배수성)를 검출한다. 본원에 기재된 방법은 하나 이상의 염색체(예를 들면, 13번 염색체, 18번 염색체, 21번 염색체 또는 이들의 조합) 또는 이의 분절에 대한 삼염색체성을 검출한다.

[0448] 일부 실시양태에서, 유전적 변이(예를 들면, 태아 이배수성)의 존재 또는 부재는 본원에 기재된 방법, 당분야에서 공지된 방법, 또는 이들의 조합에 의해 확인된다. 유전적 변이의 존재 또는 부재는 일반적으로 기준 게놈의 부분들에 맵핑된 서열 리드의 카운트로부터 확인된다. 유전적 변이의 존재 또는 부재를 확인하는 데에 사용된 서열 리드의 카운트는 종종 미처리 카운트 및/또는 필터링된 카운트이고, 종종 정규화된 카운트이다. 적합한 정규화 프로세스 또는 프로세스들을 이용하여 정규화된 카운트를 생성할 수 있고, 이러한 프로세스의 비한정적 예는 부분별 정규화, GC 함량에 의한 정규화, 선형 및 비선형 최소 자승 회귀, LOESS, GC LOESS, LOWESS, PERUN, ChAI, RM, GCRM 및 이들의 조합을 포함한다. 정규화된 카운트는 종종 특정 부분 세트 또는 세트들에 대한 프로파일 내의 하나 이상의 수준 또는 수준들로서 표현된다. 정규화된 카운트는 종종 유전적 변이의 존재 또는 부재의 확인 전에 조절되거나 패딩된다.

[0449] 일부 실시양태에서, 결과는 하나 이상의 수준에 따라 확인된다. 일부 실시양태에서, 유전적 변이(예를 들면, 염색체 이배수성)의 존재 또는 부재는 하나 이상의 조절된 수준에 따라 확인된다. 일부 실시양태에서, 유전적 변이(예를 들면, 염색체 이배수성)의 존재 또는 부재는 1개 내지 약 10,000개의 조절된 수준을 포함하는 프로파일에 따라 확인된다. 종종, 유전적 변이(예를 들면, 염색체 이배수성)의 존재 또는 부재의 확인은 약 1회 내지 약 1,000회, 1회 내지 약 900회, 1회 내지 약 800회, 1회 내지 약 700회, 1회 내지 약 600회, 1회 내지 약 500회, 1회 내지 약 400회, 1회 내지 약 300회, 1회 내지 약 200회, 1회 내지 약 100회, 1회 내지 약 50회, 1회 내지 약 25회, 1회 내지 약 20회, 1회 내지 약 15회, 1회 내지 약 10회, 또는 1회 내지 약 5회의 조절을 포함하는 프로파일에 따라 확인된다. 일부 실시양태에서, 유전적 변이(예를 들면, 염색체 이배수성)의 존재 또는 부재는 약 1회의 조절(예를 들면, 1개의 조절된 수준)을 포함하는 프로파일에 따라 확인된다. 일부 실시양태에서, 결과는 1회 이상, 2회 이상, 3회 이상, 5회 이상, 6회 이상, 7회 이상, 8회 이상, 9회 이상 또는 종종 10회 이상의 조절을 포함하는 하나 이상의 프로파일(예를 들면, 염색체 또는 이의 분절의 프로파일)에 따라 확인된다. 일부 실시양태에서, 유전적 변이(예를 들면, 염색체 이배수성)의 존재 또는 부재는 프로파일 내의 일부 수준들이 조절되어 있지 않은 프로파일에 따라 확인된다. 일부 실시양태에서, 유전적 변이(예를 들면, 염색체 이배수성)의 존

재 또는 부재는 조절되지 않은 프로파일에 따라 확인된다.

[0450] 일부 실시양태에서, 프로파일 내의 수준(예를 들면, 제1 수준)의 조절은 거짓 확인 또는 거짓 결과를 감소시킨다. 일부 실시양태에서, 프로파일 내의 수준(예를 들면, 제1 수준)의 조절은 거짓 확인 또는 거짓 결과의 빈도 및/또는 확률(예를 들면, 통계학적 확률, 가능성)을 감소시킨다. 거짓 확인 또는 결과는 정확하지 않은 확인 또는 결과일 수 있다. 거짓 확인 또는 결과는 대상체(예를 들면, 임신 여성, 태아 및/또는 이들의 조합)의 실제 또는 진짜 유전 구성, 또는 실제 또는 진짜 유전 소인(예를 들면, 유전적 변이의 존재 또는 부재)을 반영하지 않는 확인 또는 결과일 수 있다. 일부 실시양태에서, 거짓 확인 또는 결과는 거짓 음성 확인이다. 일부 실시양태에서, 음성 확인 또는 음성 결과는 유전적 변이(예를 들면, 이배수성, 카피 수 변이)의 부재이다. 일부 실시양태에서, 거짓 확인 또는 거짓 결과는 거짓 양성 확인 또는 거짓 양성 결과이다. 일부 실시양태에서, 양성 확인 또는 양성 결과는 유전적 변이(예를 들면, 이배수성, 카피 수 변이)의 존재이다. 일부 실시양태에서, 확인 또는 결과는 진단에서 사용된다. 일부 실시양태에서, 확인 또는 결과는 태아에 대한 확인 또는 결과이다.

[0451] 유전적 변이(예를 들면, 태아 이배수성)의 존재 또는 부재는 종종 부분 세트에 대한 카운트를 기준물과 비교하지 않고 확인된다. 검사 샘플에 대해 측정되고 검사 영역(예를 들면, 관심있는 부분 세트) 내에 있는 카운트는 본원에서 "검사 카운트"로서 지칭된다. 검사 카운트는 종종 본원에 기재된 프로세싱된 카운트, 평균으로서 산출된 또는 합산된 카운트, 표시치, 정규화된 카운트, 또는 하나 이상의 수준 또는 수준들이다. 일부 실시양태에서, 검사 카운트는 부분 세트에 대해 평균으로서 산출되거나 합산되고(예를 들면, 평균치, 평균, 중간치, 모드 또는 합계가 계산되고), 평균으로서 산출되거나 합산된 카운트는 역치 또는 범위와 비교된다. 검사 카운트는 종종 제1 부분 세트에 대한 카운트 대 제2 부분 세트에 대한 카운트의 비 또는 백분율로서 표현될 수 있는 표시치로서 표현될 수 있다. 일부 실시양태에서, 제1 부분 세트는 하나 이상의 검사 염색체(예를 들면, 13번 염색체, 18번 염색체, 21번 염색체, 또는 이들의 조합)에 대한 부분 세트이고, 종종 제2 부분 세트는 게놈 또는 게놈의 일부(예를 들면, 상염색체, 또는 상염색체 및 성염색체)에 대한 부분 세트이다. 일부 실시양태에서, 표시치는 역치 또는 범위와 비교된다. 일부 실시양태에서, 검사 카운트는 부분 세트에 대한 정규화된 카운트에 대한 하나 이상의 수준 또는 수준들로서 표현되고, 하나 이상의 수준 또는 수준들은 역치 또는 범위와 비교된다. 특정 역치 초과 또는 미만의, 특정 범위 내의 또는 특정 범위 밖의 검사 카운트(예를 들면, 평균으로서 산출되거나 합산된 카운트, 표시치, 정규화된 카운트, 하나 이상의 수준 또는 수준들)는 종종 유전적 변이의 존재 또는 정배수성의 결여(예를 들면, 비-정배수성)를 확인시켜준다. 특정 역치 초과 또는 미만의, 특정 범위 내의 또는 특정 범위 밖의 검사 카운트(예를 들면, 평균으로서 산출되거나 합산된 카운트, 표시치, 정규화된 카운트, 하나 이상의 수준 또는 수준들)는 종종 유전적 변이의 부재 또는 정배수성을 확인시켜준다.

[0452] 유전적 변이(예를 들면, 태아 이배수성)의 존재 또는 부재는 종종 카운트의 비교에 의해 확인되고, 이러한 카운트의 비한정적 예는 검사 카운트, 기준 카운트, 미처리 카운트, 필터링된 카운트, 평균으로서 산출되거나 합산된 카운트, 표시치(예를 들면, 염색체 표시치), 정규화된 카운트, (예를 들면, 부분 세트, 예를 들면, 게놈 구획 수준, 프로파일에 대한) 하나 이상의 수준 또는 수준들, Z-점수 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, 검사 카운트는 기준물(예를 들면, 기준 카운트)과 비교된다. 기준물(예를 들면, 기준 카운트)은 카운트의 적합한 측정치일 수 있고, 이러한 카운트의 비한정적 예는 미처리 카운트, 필터링된 카운트, 평균으로서 산출되거나 합산된 카운트, 표시치(예를 들면, 염색체 표시치), 정규화된 카운트, (예를 들면, 부분 세트, 예를 들면, 게놈 구획 수준, 프로파일에 대한) 하나 이상의 수준 또는 수준들, Z-점수 등 또는 이들의 조합을 포함한다. 기준 카운트는 종종 정배수체 검사 영역에 대한 카운트, 또는 정배수체인 게놈 또는 염색체의 분절로부터의 카운트이다. 일부 실시양태에서, 기준 카운트 및 검사 카운트는 동일한 샘플 및/또는 동일한 대상체로부터 수득된다. 일부 실시양태에서, 기준 카운트는 상이한 샘플 및/또는 상이한 대상체로부터 수득된다. 일부 실시양태에서, 기준 카운트는 검사 카운트가 유도되고/되거나 측정되는 게놈의 상응하는 분절로부터 측정되고/되거나 이 분절과 비교된다. 상응하는 분절은 기준 게놈의 동일한 위치로 맵핑되는 분절, 부분 또는 부분 세트를 지칭한다. 일부 실시양태에서, 기준 카운트는 시험 카운트가 유도되고/되거나 측정되는 게놈의 상이한 분절로부터 측정되고/되거나 이 분절과 비교된다.

[0453] 일부 실시양태에서, 검사 카운트는 종종 제1 부분 세트에 대한 카운트이고, 기준물은 제1 부분 세트와 상이한 제2 부분 세트에 대한 카운트를 포함한다. 기준 카운트는 종종 검사 샘플이 수득된 임신 여성과 동일한 임신 여성으로부터의 핵산 샘플에 대한 카운트이다. 일부 실시양태에서, 기준 카운트는 검사 샘플이 수득된 여성과 상이한 1명 이상의 임신 여성으로부터의 핵산 샘플에 대한 카운트이다. 일부 실시양태에서, 제1 부분 세트는 13번 염색체, 18번 염색체, 21번 염색체, 이들의 분절 또는 이들의 조합에 존재하고, 제2 부분 세트는 또 다른 염색체 또는 염색체들, 또는 이들의 분절에 존재한다. 비한정적 예에서, 제1 부분 세트가 21번 염색체 또는 이의 분

절에 존재하는 경우, 제2 부분 세트는 종종 또 다른 염색체(예를 들면, 1번 염색체, 13번 염색체, 14번 염색체, 18번 염색체, 19번 염색체, 이들의 분절 또는 이들의 조합)에 존재한다. 기준물은 종종 전형적으로 정배수체인 염색체 또는 이의 분절에 위치한다. 예를 들면, 1번 염색체 및 19번 염색체는 종종 1번 염색체 및 19번 염색체 이배수성과 관련된 높은 초기 태아 사망률로 인해 태아에서 정배수체이다. 검사 카운트와 기준 카운트 사이의 편차 척도가 생성될 수 있다.

[0454] 일부 실시양태에서, 기준물은 검사 카운트의 경우와 동일한 부분 세트에 대한 카운트를 포함하고, 이때 기준물에 대한 카운트는 하나 이상의 기준 샘플(예를 들면, 다수의 기준 대상체들로부터의 다수의 기준 샘플들)로부터의 카운트이다. 기준 샘플은 종종 검사 샘플이 수득된 여성과 상이한 1명 이상의 임신 여성으로부터의 샘플이다. 검사 카운트와 기준 카운트 사이의 편차 척도(예를 들면, 불확실성의 척도, 불확실성 값)가 생성될 수 있다. 일부 실시양태에서, 편차 척도는 검사 카운트로부터 결정된다. 일부 실시양태에서, 편차 척도는 기준 카운트로부터 결정된다. 일부 실시양태에서, 편차 척도는 전체 프로파일 또는 프로파일 내의 부분 서브세트로부터 결정된다.

[0455] 적합한 편차 척도가 선택될 수 있고, 이의 비한정적 예는 표준 편차, 평균 절대 편차, 중간 절대 편차, 최대 절대 편차, 표준 점수(예를 들면, z-값, z-점수, 표준 점수, 표준화된 변수) 등을 포함한다. 일부 실시양태에서, 기준 샘플은 검사 영역에 대한 정배수체이고, 검사 카운트와 기준 카운트 사이의 편차가 평가된다. 일부 실시양태에서, 유전적 변이의 존재 또는 부재는 게놈 또는 염색체의 분절 또는 부분에 대한 검사 카운트와 기준 카운트 사이의 편차의 수(예를 들면, 편차의 척도, MAD)에 따라 확인된다. 일부 실시양태에서, 유전적 변이의 존재는 검사 카운트와 기준 카운트 사이의 편차의 수가 약 1보다 더 크거나, 약 1,5보다 더 크거나, 약 2보다 더 크거나, 약 2.5보다 더 크거나, 약 2.6보다 더 크거나, 약 2.7보다 더 크거나, 약 2.8보다 더 크거나, 약 2.9보다 더 크거나, 약 3보다 더 크거나, 약 3.1보다 더 크거나, 약 3.2보다 더 크거나, 약 3.3보다 더 크거나, 약 3.4보다 더 크거나, 약 3.5보다 더 크거나, 약 4보다 더 크거나, 약 5보다 더 크거나 약 6보다 더 클 때 확인된다. 예를 들면, 종종 검사 카운트는 3의 편차 척도(예를 들면, 3 시그마, 3 MAD) 초과의 수준만큼 기준 카운트와 상이하고, 유전적 변이의 존재가 확인된다. 일부 실시양태에서, 임신 여성으로부터 수득된 검사 카운트는 3의 편차 척도(예를 들면, 3 시그마, 3 MAD) 초과의 수준만큼 기준 카운트보다 더 크고, 태아 염색체 이배수성(예를 들면, 태아 삼염색체성)의 존재가 확인된다. 검사 카운트와 기준 카운트 사이의 3 초과의 편차는 종종 비-정배수체 검사 영역(예를 들면, 유전적 변이의 존재)을 표시한다. 정배수성을 표시하는 기준 카운트보다 유의하게 더 큰 검사 카운트는 종종 삼염색체성을 확인시켜준다. 일부 실시양태에서, 임신 여성으로부터 수득된 검사 카운트는 3의 편차 척도(예를 들면, 3 시그마, 3 MAD) 초과의 수준만큼 기준 카운트보다 더 작고, 태아 염색체 이배수성(예를 들면, 태아 일염색체성)의 존재가 확인된다. 정배수성을 표시하는 기준 카운트보다 유의하게 더 작은 검사 카운트는 종종 일염색체성을 확인시켜준다.

[0456] 일부 실시양태에서, 유전적 변이의 부재는 검사 카운트와 기준 카운트 사이의 편차의 수가 약 3.5 미만, 약 3.4 미만, 약 3.3 미만, 약 3.2 미만, 약 3.1 미만, 약 3.0 미만, 약 2.9 미만, 약 2.8 미만, 약 2.7 미만, 약 2.6 미만, 약 2.5 미만, 약 2.0 미만, 약 1.5 미만 또는 약 1.0 미만일 때 확인된다. 예를 들면, 종종 검사 카운트는 3의 편차 척도(예를 들면, 3 시그마, 3 MAD) 미만의 수준만큼 기준 카운트와 상이하고, 유전적 변이의 부재가 확인된다. 일부 실시양태에서, 임신 여성으로부터 수득된 검사 카운트는 3의 편차 척도(예를 들면, 3 시그마, 3 MAD) 미만의 수준만큼 기준 카운트와 상이하고, 태아 염색체 이배수성(예를 들면, 태아 정배수체)의 부재가 확인된다. 일부 실시양태에서, 예를 들면, 검사 카운트와 기준 카운트 사이의 3 미만의 편차(예를 들면, 표준 편차에 대한 3-시그마)는 종종 정배수체 검사 영역(예를 들면, 유전적 변이의 부재)을 표시한다. 검사 샘플에 대한 검사 카운트와 하나 이상의 기준 대상체에 대한 기준 카운트 사이의 편차 척도는 작도될 수 있고 가시화될 수 있다(예를 들면, z-점수 도표).

[0457] 임의의 다른 적합한 기준물이 검사 샘플의 검사 영역에 대한 유전적 변이의 존재 또는 부재의 확인(또는 정배수체 또는 비-정배수체의 확인)을 위해 검사 카운트로 팩토링될 수 있다. 예를 들면, 태아 분획 측정치는 유전적 변이의 존재 또는 부재를 확인하기 위해 검사 카운트로 팩토링될 수 있다. 태아 분획을 정량하는 적합한 프로세스가 이용될 수 있고, 이러한 프로세스의 비한정적 예는 질량 분광측정 프로세스, 서열결정 프로세스 또는 이들의 조합을 포함한다.

[0458] 일부 실시양태에서, 태아 염색체 이배수성(예를 들면, 삼염색체성)의 존재 또는 부재는 부분적으로 태아 배수성 측정치로부터 확인된다. 일부 실시양태에서, 태아 배수성은 본원에 기재된 적합한 방법에 의해 측정된다. 일부 실시양태에서, 약 1.20 이상, 1.25 이상, 1.30 이상, 약 1.35 이상, 약 1.4 이상 또는 약 1.45 이상의 태아 배수성 측정치는 태아 염색체 이배수성의 존재(예를 들면, 태아 삼염색체성의 존재)를 표시한다. 일부 실시양태에

서, 약 1.20 내지 약 2.0, 약 1.20 내지 약 1.9, 약 1.20 내지 약 1.85, 약 1.20 내지 약 1.8, 약 1.25 내지 약 2.0, 약 1.25 내지 약 1.9, 약 1.25 내지 약 1.85, 약 1.25 내지 약 1.8, 약 1.3 내지 약 2.0, 약 1.3 내지 약 1.9, 약 1.3 내지 약 1.85, 약 1.3 내지 약 1.8, 약 1.35 내지 약 2.0, 약 1.35 내지 약 1.9, 약 1.35 내지 약 1.8, 약 1.4 내지 약 2.0, 약 1.4 내지 약 1.85, 또는 약 1.4 내지 약 1.8의 태아 배수성 측정치는 태아 염색체 이배수성의 존재(예를 들면, 태아 삼염색체성의 존재)를 표시한다. 일부 실시양태에서, 태아 이배수성은 삼염색체성이다. 일부 실시양태에서, 태아 이배수성은 13번 염색체, 18번 염색체 및/또는 21번 염색체의 삼염색체성이다.

[0459] 일부 실시양태에서, 약 1.35 미만, 약 1.30 미만, 약 1.25 미만, 약 1.20 미만 또는 약 1.15 미만의 태아 배수성은 태아 이배수성의 부재(예를 들면, 태아 삼염색체성의 부재, 예를 들면, 정배수체)를 표시한다. 일부 실시양태에서, 약 0.7 내지 약 1.35, 약 0.7 내지 약 1.30, 약 0.7 내지 약 1.25, 약 0.7 내지 약 1.20, 약 0.7 내지 약 1.15, 약 0.75 내지 약 1.35, 약 0.75 내지 약 1.30, 약 0.75 내지 약 1.25, 약 0.75 내지 약 1.20, 약 0.75 내지 약 1.15, 약 0.8 내지 약 1.35, 약 0.8 내지 약 1.30, 약 0.8 내지 약 1.25, 약 0.8 내지 약 1.20, 또는 약 0.8 내지 약 1.15의 태아 배수성 측정치는 태아 염색체 이배수성의 부재(예를 들면, 태아 삼염색체성의 부재, 예를 들면, 정배수체)를 표시한다.

[0460] 일부 실시양태에서, 약 0.8 미만, 약 0.75 미만, 약 0.70 미만 또는 약 0.6 미만의 태아 배수성은 태아 이배수성의 존재(예를 들면, 염색체 결실의 존재)를 표시한다. 일부 실시양태에서, 약 0 내지 약 0.8, 약 0 내지 약 0.75, 약 0 내지 약 0.70, 약 0 내지 약 0.65, 약 0 내지 약 0.60, 약 0.1 내지 약 0.8, 약 0.1 내지 약 0.75, 약 0.1 내지 약 0.70, 약 0.1 내지 약 0.65, 약 0.1 내지 약 0.60, 약 0.2 내지 약 0.8, 약 0.2 내지 약 0.75, 약 0.2 내지 약 0.70, 약 0.2 내지 약 0.65, 약 0.2 내지 약 0.60, 약 0.25 내지 약 0.8, 약 0.25 내지 약 0.75, 약 0.25 내지 약 0.70, 약 0.25 내지 약 0.65, 약 0.25 내지 약 0.60, 약 0.3 내지 약 0.8, 약 0.3 내지 약 0.75, 약 0.3 내지 약 0.70, 약 0.3 내지 약 0.65, 또는 약 0.3 내지 약 0.60의 태아 배수성 측정치는 태아 염색체 이배수성의 존재(예를 들면, 염색체 결실의 존재)를 표시한다. 일부 실시양태에서, 측정된 태아 이배수성은 전체 염색체 결실이다.

[0461] 일부 실시양태에서, (예를 들면, 상기 배수성 측정치의 범위들 중 하나 이상의 범위에 따른) 태아 이배수성의 존재 또는 부재의 확인은 판정 대역에 따라 확인된다. 일부 실시양태에서, 값(예를 들면, 배수성 값, 태아 분획 값, 불확실성 수준) 또는 일군의 값들이 미리 한정된 범위(예를 들면, 대역, 판정 대역) 내에 속할 때 판정(예를 들면, 유전적 변이의 존재 또는 부재, 예를 들면, 결과를 확인하는 판정)이 수행된다. 일부 실시양태에서, 판정 대역은 동일한 환자 샘플로부터 수득된 일군의 값들에 따라 한정된다. 일부 실시양태에서, 판정 대역은 동일한 염색체 또는 이의 분절로부터 유도된 일군의 값들에 따라 한정된다. 일부 실시양태에서, 배수성 측정치에 기초한 판정 대역은 신뢰 수준(예를 들면, 높은 신뢰 수준, 예를 들면, 낮은 불확실성 수준) 및/또는 태아 분획에 따라 한정된다. 일부 실시양태에서, 판정 대역은 배수성 측정치 및 약 2.0% 이상, 약 2.5% 이상, 약 3% 이상, 약 3.25% 이상, 약 3.5% 이상, 약 3.75% 이상 또는 약 4.0% 이상의 태아 분획에 따라 한정된다. 예를 들면, 일부 실시양태에서, 태아를 가진 임신 여성으로부터 수득된 샘플에 대한 2% 이상 또는 4% 이상의 태아 분획 측정치와 함께 1.25 초과의 배수성 측정치를 기초로 태아가 삼염색체성 21을 포함한다고 판정한다. 일부 실시양태에서, 예를 들면, 태아를 가진 임신 여성으로부터 수득된 샘플에 대한 2% 이상 또는 4% 이상의 태아 분획 측정치와 함께 1.25 미만의 배수성 측정치를 기초로 태아가 정배수체라고 판정한다. 일부 실시양태에서, 판정 대역은 약 99% 이상, 약 99.1% 이상, 약 99.2% 이상, 약 99.3% 이상, 약 99.4% 이상, 약 99.5% 이상, 약 99.6% 이상, 약 99.7% 이상, 약 99.8% 이상 또는 약 99.9% 이상의 신뢰 수준에 의해 한정된다. 일부 실시양태에서, 판정 대역을 사용하지 않고 판정한다. 일부 실시양태에서, 판정 대역 및 추가 데이터 또는 정보를 사용하여 판정한다. 일부 실시양태에서, 판정 대역을 사용하지 않고 배수성 값을 기초로 판정한다. 일부 실시양태에서, 배수성 값을 계산하지 않고 판정한다. 일부 실시양태에서, 프로파일의 시각적 검사(예를 들면, 게놈 구획 수준의 시각적 검사)를 기초로 판정한다. 전체적으로 또는 부분적으로 본원에 기재된 방법에 의해 수득된 측정치, 값 및/또는 데이터를 기초로 임의의 적합한 방법으로 판정할 수 있고, 이러한 측정치, 값 및/또는 데이터의 비한정적 예는 태아 배수성 측정치, 태아 분획 측정치, 모체 배수성, 불확실성 및/또는 신뢰도 측정치, 부분 수준, 수준, 프로파일, z-점수, 기대된 염색체 표시치, 측정된 염색체 표시치, 카운트(예를 들면, 정규화된 카운트, 미처리 카운트), 태아 또는 모체 카피 수 변이(예를 들면, 분류된 카피 수 변이), 유의하게 상이한 수준, 조절된 수준(예를 들면, 패딩) 등 또는 이들의 조합을 포함한다.

[0462] 일부 실시양태에서, 비-판정 대역은 판정하지 않는 대역이다. 일부 실시양태에서, 비-판정 대역은 낮은 정확성, 높은 위험, 높은 오차, 낮은 신뢰 수준, 높은 불확실성 수준 등 또는 이들의 조합을 표시하는 값 또는 일군의

값들에 의해 한정된다. 일부 실시양태에서, 비-판정 대역은 부분적으로 약 5% 이하, 약 4% 이하, 약 3% 이하, 약 2.5% 이하, 약 2.0% 이하, 약 1.5% 이하 또는 약 1.0% 이하의 태아 분획에 의해 한정된다.

[0463] 유전적 변이는 종종 의학적 질환과 관련되어 있다. 유전적 변이를 확인시켜주는 결과는 종종 질환(예를 들면, 의학적 질환), 질병, 증후군 또는 기형의 존재 또는 부재를 확인시켜주는 결과이거나, 질환, 질병, 증후군 또는 기형(예를 들면, 표 1에 나열된 비한정적 예)의 검출을 포함한다. 일부 실시양태에서, 진단은 결과의 평가를 포함한다. 본원에 기재된 방법으로 질환(예를 들면, 의학적 질환), 질병, 증후군 또는 기형의 존재 또는 부재를 확인시켜주는 결과는 종종 추가 검사(예를 들면, 핵형분석 및/또는 양수천자)에 의해 독립적으로 검증될 수 있다. 데이터의 분석 및 프로세싱은 하나 이상의 결과를 제공할 수 있다. 본원에서 사용된 용어 "결과"는 유전적 변이(예를 들면, 이배수성, 카피 수 변이)의 존재 또는 부재의 확인을 용이하게 하는 데이터 프로세싱의 결과를 지칭할 수 있다. 일부 실시양태에서, 본원에서 사용된 용어 "결과"는 유전적 변이(예를 들면, 이배수성, 카피 수 변이)의 존재 또는 부재를 예측하고/하거나 확인하는 결론을 지칭한다. 일부 실시양태에서, 본원에서 사용된 용어 "결과"는 대상체(예를 들면, 태아)에서 유전적 변이(예를 들면, 이배수성, 카피 수 변이)의 존재 또는 부재의 위험 또는 확률을 예측하고/하거나 확인하는 결론을 지칭한다. 진단은 종종 결과의 사용을 포함한다. 예를 들면, 건강 전문가는 결과를 분석할 수 있고 결과를 기초로 또는 부분적으로 결과를 기초로 진단을 제공할 수 있다. 일부 실시양태에서, 질환, 증후군 또는 기형(예를 들면, 표 1에 나열됨)의 확인, 검출 또는 진단은 유전적 변이의 존재 또는 부재를 확인시켜주는 결과의 사용을 포함한다. 일부 실시양태에서, 카운팅된 맵핑된 서열 리드 또는 이의 변환물에 기초한 결과는 유전적 변이의 존재 또는 부재를 확인시켜 준다. 일부 실시양태에서, 본원에 기재된 하나 이상의 방법(예를 들면, 데이터 프로세싱 방법)을 이용함으로써 생성된 결과는 표 1에 나열된 하나 이상의 질환, 증후군 또는 기형의 존재 또는 부재를 확인시켜준다. 일부 실시양태에서, 진단은 질환, 증후군 또는 기형의 존재 또는 부재의 확인을 포함한다. 종종, 진단은 질환, 증후군 또는 기형의 성질 및/또는 원인으로서 유전적 변이의 확인을 포함한다. 일부 실시양태에서, 결과는 진단이 아니다. 결과는 종종 하나 이상의 확률의 고려와 관련하여 본원에 기재된 프로세싱 방법을 이용함으로써 생성된 하나 이상의 수치 값을 포함한다. 위험 또는 확률의 고려는 불확실성 값, 가변성의 측정치, 신뢰 수준, 민감성, 특이성, 표준 편차, 변동 계수(CV) 및/또는 신뢰 수준, Z-점수, Chi 값, Phi 값, 배수성 값, 피팅된 태아 분획, 면적 비, 중간 수준 등 또는 이들의 조합을 포함할 수 있으나 이들로 한정되지 않는다. 확률의 고려는 대상체가 유전적 변이를 가질 위험에 있는지 아니면 유전적 변이를 갖는지를 확인하는 것을 용이하게 할 수 있고, 유전 장애의 존재 또는 부재를 확인시켜주는 결과는 종종 이러한 고려를 포함한다.

[0464] 결과는 종종 표현형이다. 결과는 종종 관련된 신뢰 수준을 갖는 표현형이다(예를 들면, 불확실성 값, 예를 들면, 태아는 99%의 신뢰 수준으로 삼염색체성 21에 대한 양성을 나타내고, 검사 대상체는 95%의 신뢰 수준에서 유전적 변이와 관련된 암에 대한 음성을 나타낸다). 결과 값을 생성하는 상이한 방법들은 종종 상이한 종류의 결과를 생성할 수 있다. 일반적으로, 본원에 기재된 방법을 이용함으로써 생성된 결과 값을 기초로 만들어질 수 있는 4종의 가능한 점수 또는 판정이 존재한다: 진짜 양성, 거짓 양성, 진짜 음성 및 거짓 음성. 본원에서 사용된 용어 "점수", "점수들", "판정" 및 "판정들"은 특정 유전적 변이가 대상체/샘플에 존재하거나 부재할 확률의 계산을 지칭한다. 점수의 값은 예를 들면, 변이, 차이, 또는 유전적 변이에 상응할 수 있는 맵핑된 서열 리드의 비를 측정하는 데에 사용될 수 있다. 예를 들면, 선택된 유전적 변이 또는 부분에 대한 양의 점수를 기준 게놈에 대해 데이터 세트로부터 계산하여 종종 의학적 질환(예를 들면, 암, 임신중독증, 삼염색체성, 일염색체성 등)와 관련된 유전적 변이의 존재 또는 부재를 확인할 수 있다. 일부 실시양태에서, 결과는 수준, 프로파일 및/또는 도표(예를 들면, 프로파일 도표)를 포함한다. 결과가 프로파일을 포함하는 실시양태에서, 적합한 프로파일 또는 프로파일의 조합이 결과를 위해 사용될 수 있다. 결과를 위해 사용될 수 있는 프로파일의 비한정적 예는 z-점수 프로파일, p-점수 프로파일, chi 값 프로파일, phi 값 프로파일 등 또는 이들의 조합을 포함한다.

[0465] 유전적 변이의 존재 또는 부재를 확인하기 위해 생성된 결과는 종종 널 결과(예를 들면, 2개의 클러스터들 사이의 데이터 점, 유전적 변이의 존재 및 부재 둘 다에 대한 값을 포괄하는 표준 편차를 갖는 수치 값, 조사되는 유전적 변이를 갖거나 갖지 않는 대상체에 대한 프로파일 도표와 유사하지 않은 프로파일 도표를 갖는 데이터 세트)를 포함한다. 일부 실시양태에서, 널 결과를 표시하는 결과는 여전히 확정적인 결과이고, 확인은 유전적 변이의 존재 또는 부재의 확인을 위해 추가 정보 및/또는 데이터 생성 및/또는 분석의 반복을 필요로 할 수 있다.

[0466] 일부 실시양태에서, 결과는 본원에 기재된 하나 이상의 프로세싱 단계를 수행한 후에 생성될 수 있다. 일부 실시양태에서, 결과는 본원에 기재된 프로세싱 단계들 중 하나의 결과로서 생성되고, 일부 실시양태에서 결과는 데이터 세트의 각각의 통계학적 및/또는 수학적 조작이 수행된 후 생성될 수 있다. 유전적 변이의 존재 또는 부

재의 확인에 대한 결과는 대상체 또는 샘플에 대한 유전적 변이의 존재 또는 부재와 관련된 적합한 형태로 표현될 수 있고, 이 형태는 확률(예를 들면, 승산비, p-값), 가능성, 클러스터 내의 또는 밖의 값, 역치 값 초과 또는 미만의 값, 범위(예를 들면, 역치 범위) 내의 값, 분산 또는 신뢰의 측정치를 갖는 값 또는 위험 계수를 포함하나 이들로 한정되지 않는다. 일부 실시양태에서, 샘플들 사이의 비교는 샘플 동일성의 확인을 가능하게 한다(예를 들면, 반복된 샘플 및/또는 혼합된(예를 들면, 잘못 표지된 또는 조합된) 샘플 등의 확인을 가능하게 한다).

[0467] 일부 실시양태에서, 결과는 예정된 역치 또는 컷오프 값 초과 또는 미만(예를 들면, 1 초과 또는 1 미만)의 값, 및 이 값과 관련된 불확실성 또는 신뢰 수준을 포함한다. 일부 실시양태에서, 예정된 역치 또는 컷오프 값은 기대 수준 또는 기대 수준 범위이다. 결과는 데이터 프로세싱에서 사용된 가정을 기술할 수도 있다. 일부 실시양태에서, 결과는 예정된 값 범위(예를 들면, 역치 범위) 내 또는 밖에 속하는 값, 및 상기 범위 내 또는 밖에 있는 그 값에 대한 관련된 불확실성 또는 신뢰 수준을 포함한다. 일부 실시양태에서, 결과는 예정된 값과 동등한(예를 들면, 1과 동등하거나 0과 동등한) 값 또는 예정된 값 범위 내의 값과 동등한 값, 및 동등하거나 범위 내에 또는 밖에 있는 그 값에 대한 그의 관련된 불확실성 또는 신뢰 수준을 포함한다. 결과는 종종 도표(예를 들면, 프로파일 도표)로서 도식적으로 표시된다.

[0468] 상기 인지된 바와 같이, 결과는 진짜 양성, 진짜 음성, 거짓 양성 또는 거짓 음성으로서 특징규명될 수 있다. 본원에서 사용된 용어 "진짜 양성"은 유전적 변이를 갖는 것으로 정확히 진단된 대상체를 지칭한다. 본원에서 사용된 용어 "거짓 양성"은 유전적 변이를 갖는 것으로서 잘못 확인된 대상체를 지칭한다. 본원에서 사용된 용어 "진짜 음성"은 유전적 변이를 갖지 않는 것으로서 정확히 확인된 대상체를 지칭한다. 본원에서 사용된 용어 "거짓 음성"은 유전적 변이를 갖지 않는 것으로서 잘못 확인된 대상체를 지칭한다. 임의의 주어진 방법에 대한 수행의 2개 측정치가 이들 발생의 비를 기초로 계산될 수 있다: (i) 일반적으로 양성으로서 정확히 확인되는 예측된 양성의 분율인 민감성 값; 및 (ii) 일반적으로 음성으로서 정확히 확인되는 예측된 음성의 분율인 특이성 값.

[0469] 일부 실시양태에서, 민감성, 특이성 및/또는 신뢰 수준 중 하나 이상은 백분율로서 표현된다. 일부 실시양태에서, 독립적으로 각각의 변수에 대한 백분율은 약 90%보다 더 크다(예를 들면, 약 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% 또는 99%, 또는 99% 초과(예를 들면, 약 99.5% 이상, 약 99.9% 이상, 약 99.95% 이상, 약 99.99% 이상)). 일부 실시양태에서, 변동 계수(CV)는 백분율로서 표현되고, 종종 이 백분율은 약 10% 이하(예를 들면, 약 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2% 또는 1%, 또는 1% 미만(예를 들면, 약 0.5% 이하, 약 0.1% 이하, 약 0.05% 이하, 약 0.01% 이하))이다. 일부 실시양태에서, 확률(예를 들면, 특정 결과가 우연에 기인하지 않는 것)은 Z-점수, p-값, 또는 t-검정의 결과로서 표현된다. 일부 실시양태에서, 본원에 기재된 하나 이상의 데이터 프로세싱 조작을 이용하여 결과에 대한 측정된 분산, 신뢰 수준, 민감성, 특이성 등(예를 들면, 신뢰 파라미터로서 총칭됨)을 생성할 수 있다. 결과 생성 및 관련된 신뢰 수준의 구체적인 예는 실시예 단락 및 국제특허출원 제PCT/US12/59123호(WO2013/052913)(모든 본문, 표, 방정식 및 도면을 포함하는 전체 내용이 본원에 참고로 도입됨)에 기재되어 있다.

[0470] 본원에서 사용된 용어 "민감성"은 진짜 양성의 수 + 거짓 음성의 수로 나누어진 진짜 양성의 수를 지칭하고, 이때 민감성(sens)은 $0 \leq sens \leq 1$의 범위 내에 있을 수 있다. 본원에서 사용된 용어 "특이성"은 진짜 음성의 수 + 거짓 양성의 수로 나누어진 진짜 음성의 수를 지칭하고, 이때 특이성(spec)은 $0 \leq spec \leq 1$의 범위 내에 있을 수 있다. 일부 실시양태에서, 1, 100%, 또는 거의 1(예를 들면, 약 90% 내지 약 99%)과 동등한 민감성 및 특이성을 갖는 방법이 종종 선택된다. 일부 실시양태에서, 1 또는 100%와 동등한 민감성을 갖는 방법이 선택되고, 일부 실시양태에서 거의 1의 민감성(예를 들면, 약 90%의 민감성, 약 91%의 민감성, 약 92%의 민감성, 약 93%의 민감성, 약 94%의 민감성, 약 95%의 민감성, 약 96%의 민감성, 약 97%의 민감성, 약 98%의 민감성 또는 약 99%의 민감성)을 갖는 방법이 선택된다. 일부 실시양태에서, 1 또는 100%의 특이성을 갖는 방법이 선택되고, 일부 실시양태에서 거의 1의 특이성(예를 들면, 약 90%의 특이성, 약 91%의 특이성, 약 92%의 특이성, 약 93%의 특이성, 약 94%의 특이성, 약 95%의 특이성, 약 96%의 특이성, 약 97%의 특이성, 약 98%의 특이성 또는 약 99%의 특이성)을 갖는 방법이 선택된다.

[0471] 일부 실시양태에서, 태아에 대한 유전적 변이(예를 들면, 염색체 이배수성)의 존재 또는 부재가 확인된다. 이러한 실시양태에서, 태아 유전적 변이(예를 들면, 태아 염색체 이배수성)의 존재 또는 부재가 확인된다.

[0472] 일부 실시양태에서, 샘플에 대한 유전적 변이(예를 들면, 염색체 이배수성)의 존재 또는 부재가 확인된다. 이러한 실시양태에서, 샘플 핵산에서의 유전적 변이(예를 들면, 염색체 이배수성)의 존재 또는 부재가 확인된다. 일

부 실시양태에서, 검출되거나 검출되지 않는 변이가 한 공급원으로부터의 샘플 핵산에 존재하지만 또 다른 공급원으로부터의 샘플 핵산에는 존재하지 않는다. 공급원의 비한정적 예는 태반 핵산, 태아 핵산, 모체 핵산, 암세포 핵산, 비-암세포 핵산 등 및 이들의 조합물을 포함한다. 비한정적 예에서, 검출되거나 검출되지 않는 특정 유전적 변이는 (i) 태반 핵산에는 존재하지만 태아 핵산 및 모체 핵산에는 존재하지 않거나; (ii) 태아 핵산에는 존재하지만 모체 핵산에는 존재하지 않거나; (iii) 모체 핵산에는 존재하지만 태아 핵산에는 존재하지 않는다.

[0473] 하나 이상의 결과가 생성된 후, 결과는 종종 유전적 변이 및/또는 관련된 의학적 질환의 존재 또는 부재의 확인을 제공하는 데에 사용된다. 결과는 전형적으로 건광 관리 전문가(예를 들면, 실험실 기술자 또는 관리자; 의사 또는 보조자)에게 제공된다. 종종, 결과는 결과 모듈에 의해 제공된다. 일부 실시양태에서, 결과는 작도 모듈에 의해 제공된다. 일부 실시양태에서, 결과는 장치의 주변장치 또는 부품 상에서 제공된다. 예를 들면, 종종 결과는 프린터 또는 디스플레이에 의해 제공된다. 일부 실시양태에서, 유전적 변이의 존재 또는 부재를 확인시켜주는 결과는 보고서의 형태로 건강관리 전문가에게 제공되고, 일부 실시양태에서 보고서는 결과 값 및 관련된 신뢰 파라미터의 디스플레이를 포함한다. 일반적으로, 결과는 유전적 변이 및/또는 의학적 질환의 존재 또는 부재의 확인을 용이하게 하는 적합한 포맷으로 디스플레이될 수 있다. 데이터 세트를 보고하고/하거나 디스플레이하거나 결과를 보고하는 데에 사용되기에 적합한 포맷의 비한정적 예는 디지털 데이터, 그래프, 2D 그래프, 3D 그래프 및 4D 그래프, 사진, 상형문자, 차트, 막대 그래프, 파이 그래프, 도표, 순서도, 산점도, 지도, 히스토그램, 밀도 차트, 함수 그래프, 회로도, 블록도, 버블 지도, 성상도, 이체선도, 통계지도, 방사도(spider chart), 벤(Venn) 도표, 계산 도표 등 및 이들의 조합을 포함한다. 결과 표시의 다양한 예들이 도면에 제시되어 있고 실시예에 기재되어 있다.

[0474] 일부 실시양태에서, 결과를 생성하는 것은 핵산 서열 리드 데이터 등을 대상체의 세포 핵산의 표시치로 변환하는 것으로서 간주될 수 있다. 예를 들면, 대상체로부터의 핵산의 서열 리드를 분석하고 염색체 프로파일 및/또는 결과를 생성하는 것은 상대적으로 작은 서열 리드 단편을 상대적으로 큰 염색체 구조물의 표시치로 변환하는 것으로서 간주될 수 있다. 일부 실시양태에서, 결과는 대상체(예를 들면, 임신 여성)로부터의 서열 리드를 상기 대상체(예를 들면, 모체 및/또는 태아 핵산)에 존재하는 기존 구조물(예를 들면, 게놈, 염색체 또는 이들의 분절)의 표시치로 변환하는 것으로부터 생성된다. 일부 실시양태에서, 결과는 제1 대상체(예를 들면, 임신 여성)로부터의 서열 리드를 구조물(예를 들면, 게놈, 염색체 또는 이들의 분절)의 복합 표시치로 변환하는 제1 변환, 및 제1 대상체(예를 들면, 임신 여성) 및/또는 제2 대상체(예를 들면, 태아)에 존재하는 구조물의 표시치를 생성하는 상기 복합 표시치의 제2 변환을 포함한다.

[0475] 일부 실시양태에서, 결과는 하나 이상의 후보 분절의 분석에 따라 생성될 수 있다. 일부 실시양태에서, 유전적 변이의 존재 또는 부재는 별개의 분절, 후보 분절 또는 복합 후보 분절(예를 들면, 별개의 분절, 후보 분절 또는 복합 후보 분절의 존재 또는 부재)에 따라 확인된다. 일부 실시양태에서, 동일한 프로파일의 2개의 분해 렌더링으로부터 유도된 2개의 후보 분절들은 (예를 들면, 비교에 따를 때) 실질적으로 동일하고, 염색체 이배수성, 미세중복 또는 미세결실의 존재가 확인된다. 일부 실시양태에서, 복합 후보 분절의 존재는 염색체 이배수성, 미세중복 또는 미세결실의 존재를 표시한다. 일부 실시양태에서, 전체 염색체 이배수성의 존재는 프로파일 내의 별개의 분절, 후보 분절 또는 복합 후보 분절의 존재에 따라 확인되고, 상기 프로파일은 게놈의 분절(예를 들면, 염색체보다 더 큰 분절, 예를 들면, 2개 이상의 염색체를 표시하는 분절, 전체 게놈을 표시하는 분절)이다. 일부 실시양태에서, 전체 염색체 이배수성의 존재는 프로파일 내의 별개의 분절, 후보 분절 또는 복합 후보 분절의 존재에 따라 확인되고, 별개의 분절 에지들은 염색체의 에지와 실질적으로 동일하다. 일부 실시양태에서, 프로파일 내의 별개의 분절, 후보 분절 또는 복합 후보 분절의 하나 이상의 에지가 염색체의 에지와 상이하고/하거나 별개의 분절이 염색체 내에 있을 때 미세중복 또는 미세결실의 존재가 확인된다. 일부 실시양태에서, 미세중복의 존재가 확인되고, 별개의 분절, 후보 분절 또는 복합 후보 분절에 대한 수준 또는 AUC는 기준 수준(예를 들면, 정배수체 영역)보다 실질적으로 더 크다. 일부 실시양태에서, 미세결실의 존재가 확인되고, 별개의 분절, 후보 분절 또는 복합 후보 분절에 대한 수준 또는 AUC는 기준 수준(예를 들면, 정배수체 영역)보다 실질적으로 더 작다. 일부 실시양태에서, 2개 이상의 상이한 분해 렌더링에서 확인된 후보 분절들은 실질적으로 동일하지 않고(예를 들면, 상이하고), 염색체 이배수성, 미세중복 및/또는 미세결실의 부재가 확인된다. 일부 실시양태에서, 프로파일 또는 프로파일의 분해 렌더링 내의 별개의 분절, 후보 분절 또는 복합 후보 분절의 부재는 염색체 이배수성, 미세중복 또는 미세결실의 부재를 표시한다.

[0476] 검증

[0477] 일부 실시양태에서, 본원에 기재된 방법은 검증을 포함한다. 일부 실시양태에서, 결정 분석(예를 들면, 결정 트

리), 유전적 변이(예를 들면, 카피 수 변이, 미세중복, 미세결실, 이배수성)의 존재 또는 부재의 확인, 판정의 수행 및/또는 결과의 확인은 검증을 포함한다. 임의의 적합한 검증 프로세스가 본원에 기재된 방법, 판정 또는 결과를 검증하는 데에 이용될 수 있다.

[0478] 일부 실시양태에서, 검증은 분해 렌더링에서 확인된 후보 분절의 검증 또는 비검증을 포함한다. 검증된 후보 분절은 후보 분절의 존재를 확인시켜준다. 검증되지 않은 후보 분절은 후보 분절의 존재를 표시하는 판정을 후보 분절의 부재로 바꾼다. 예를 들면, 일부 실시양태에서, 분할 프로세스에 의한 후보 분절의 확인 후, 검증이 수행될 수 있고, 이때 후보 분절이 검증되거나 검증되지 않는다. 검증되지 않은 후보 분절은 프로파일 내의 염색체 이배수성, 미세중복 또는 미세결실의 부재를 표시한다. 일부 실시양태에서, 검증은 거짓 음성 및/또는 거짓 양성 확인을 감소시키면서 후보 분절의 존재 또는 부재를 확인하는 단계를 포함한다. 후보 분절은 적합한 방법에 의해 검증될 수 있고, 이러한 방법의 비한정적 예는 "슬라이딩 에지" 프로세스, "리브 원 아웃" 프로세스 등 또는 이들의 조합을 포함한다.

[0479] 일부 실시양태에서, 검증은 후보 분절 또는 복합 후보 분절에 대한 유의 수준을 생성하는 단계를 포함한다. 일부 실시양태에서, 유의 수준은 Z-점수, z-값, p-값 등이다. 일부 실시양태에서, 검증은 불확실성 수준을 생성하는 단계를 포함한다. 일부 실시양태에서, 불확실성 수준은 유의 수준과 관련되어 있다. 예를 들면, 종종 평균치, 평균 또는 중간 유의 수준이 측정되고, 평균치, 평균 또는 중간 유의 수준에 대한 불확실성 수준이 측정된다.

[0480] 일부 실시양태에서, 후보 분절은 유의 수준 및/또는 불확실성 값에 따라 검증되거나 검증되지 않는다. 검증된 또는 검증되지 않은 별개의 분절은 검증된 또는 검증되지 않은 복합 후보 분절일 수 있다. 일부 실시양태에서, 검증된 후보 분절의 존재 또는 부재는 후보 분절에 대한 유의 수준 및/또는 불확실성 수준에 따라 확인된다. 일부 실시양태에서, 검증된 후보 분절의 부재는 염색체 이배수성, 미세중복 또는 미세결실의 부재를 표시한다. 일부 실시양태에서, 검증된 후보 분절의 존재는 후보 분절의 존재를 확인시켜준다. 일부 실시양태에서, 2개 이상의 검증된 후보 분절들의 존재는 복합 후보 분절의 확인 또는 생성으로 이어진다. 일부 실시양태에서, 부분적으로 하나 이상의 검증된 후보 분절의 존재는 증가된 신뢰 수준으로 염색체 이배수성, 미세중복 또는 미세결실의 존재를 확인시켜준다. 일부 실시양태에서, 부분적으로 후보 분절의 존재는 디조지 증후군의 존재를 표시한다. 일부 실시양태에서, 검증된 후보 분절의 부재는 염색체 이배수성, 미세중복 또는 미세결실의 부재를 표시한다.

[0481] 슬라이딩 에지 검증

[0482] 일부 실시양태에서, 검증은 "슬라이딩 에지" 프로세스를 포함한다. 적합한 "슬라이딩 에지" 프로세스는 직접적으로 이용될 수 있거나 분해 렌더링에서 분절을 검증하기 위해 채택될 수 있다. 일부 실시양태에서, "슬라이딩 에지" 프로세스는 후보 분절(예를 들면, 부분 세트로 표시된 후보 분절) 또는 후보 분절을 포함하거나 후보 분절인 것으로 의심되는 분절을 다수의 부분 서브세트로 분할하는 단계를 포함한다. 일부 실시양태에서, 후보 분절은 전체 염색체 또는 염색체의 분절에 대한 부분 세트이다. 일부 실시양태에서, 후보 분절은 공지된 유전적 변이 또는 공지된 유전 장애와 관련된 영역을 포함하는 부분 세트이다. 일부 실시양태에서, 후보 분절은 디조지 영역을 포함한다.

[0483] 일부 실시양태에서, "슬라이딩 에지" 프로세스는 확인된 후보 분절(부분 세트)을 다수의 부분 서브세트로 분할하는 단계를 포함하고, 이때 각각의 부분 서브세트는 유사하지만 상이한 에지를 갖는 후보 분절을 표시한다. 일부 실시양태에서, 최초로 확인된 후보 분절이 분석에 포함된다. 예를 들면, 최초로 확인된 후보 분절은 다수의 부분 서브세트들 중 하나로서 포함된다. 부분 서브세트는 최초로 확인된 별개의 분절의 한쪽 또는 양쪽 에지를 임의의 적합한 방법으로 변화시킴으로써 확인될 수 있다. 일부 실시양태에서, 좌측 에지를 변화시켜 상이한 좌측 에지를 갖는 별개의 분절을 생성할 수 있다. 일부 실시양태에서, 우측 에지를 변화시켜 상이한 우측 에지를 갖는 별개의 분절을 생성할 수 있다. 일부 실시양태에서, 우측 에지 및 좌측 에지 둘 다가 변화될 수 있다. 일부 실시양태에서, 상기 에지들은 이 에지들을 기준 게놈의 하나 이상의 인접한 부분만큼 원래의 에지의 좌측 또는 우측으로 이동시킴으로써 변화된다.

[0484] 실시예 5에 기재된 슬라이딩 에지 방법의 한 실시양태에서, 최초 별개의 분절은 양쪽 에지를 기준 게놈의 15개 부분만큼 이동시켜 별개의 분절의 15x15 격자(예를 들면, 225개의 상이한 부분 서브세트)를 생성함으로써 변화된다. 예를 들면, 우측 에지를 안정한 상태로 유지하면서, 좌측 에지를 기준 게놈의 7개 부분만큼 우측으로 이동시킨 후 기준 게놈의 7개 부분만큼 좌측으로 이동시켜 15개의 가능한 좌측 에지를 생성할 수 있다. 15개 좌측 에지 각각을 안정한 상태로 유지하면서, 우측 에지를 기준 게놈의 7개 부분만큼 우측으로 이동시키고 기준 게놈의 7개 부분만큼 좌측으로 이동시켜 15개의 가능한 우측 에지를 생성할 수 있다. 생성된 서브세트는 225개의 상

이한 별개의 분절들(예를 들면, 기준 게놈의 부분 서브세트)을 포함한다.

[0485] 일부 실시양태에서, 에지들 중 어느 하나 또는 둘 다가 기준 게놈의 5개 내지 30개 부분만큼 변화된다. 일부 실시양태에서, 에지는 기준 게놈의 1개, 2개, 3개, 4개, 5개, 6개, 7개, 8개, 9개, 10개, 11개, 12개, 13개, 14개, 15개, 16개, 17개, 18개, 19개, 20개, 21개, 22개, 23개, 24개, 25개, 26개, 27개, 28개, 29개 또는 30개 부분만큼 어느 한 방향으로 이동된다. 일부 실시양태에서, 부분 크기와 관계없이 에지는 어느 한 에지 또는 양쪽 에지에 대해 약 100,000개 내지 약 2,000,000개 염기쌍, 250,000개 내지 약 1,500,000개 염기쌍, 또는 약 500,000개 내지 약 1,000,000개 염기쌍의 에지 범위를 생성하도록 변화된다. 일부 실시양태에서, 부분 크기와 관계없이 에지는 어느 한 에지 또는 양쪽 에지에 대해 약 500,000개, 600,000개, 700,000개, 750,000개, 800,000개, 900,000개 또는 약 1,000,000개 염기쌍의 에지 범위를 생성하도록 변화된다.

[0486] 일부 실시양태에서, 확인된 별개의 분절은 제1 말단 및 제2 말단을 포함하고, 분할은 (i) 반복적 제거로 하나 이상의 부분을 부분 세트의 제1 말단으로부터 제거함으로써 반복적 제거마다 부분 서브세트를 제공하는 단계, (ii) n회의 반복 후 (i)에서의 반복적 제거를 종결하여 n + 1개의 부분 서브세트를 제공하는 단계로서, 이때 상기 부분 세트가 서브세트이고 각각의 서브세트가 상이한 수의 부분, 제1 서브세트 말단 및 제2 서브세트 말단을 포함하는 것인 단계, (iii) 반복적 제거로 하나 이상의 부분을 (ii)에서 제공된 n + 1개의 부분 서브세트 각각의 제2 서브세트 말단으로부터 제거하는 단계, 및 (iv) n회의 반복 후 (iii)에서의 반복적 제거를 종결하여 다수의 부분 서브세트를 제공하는 단계를 포함한다. 일부 실시양태에서, 다수의 서브세트는 (n + 1)2개의 서브세트와 동등하다. 일부 실시양태에서, n은 5 내지 30의 정수와 동등하다. 일부 실시양태에서, n은 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 또는 30과 동등하다.

[0487] 슬라이딩 에지 방법의 일부 실시양태에서, 기준 게놈의 부분 서브세트 각각에 대한 유의 수준(예를 들면, Z-점수, p-값)이 측정되고, 모든 서브세트들에 대해 측정된 유의 수준에 따라 평균치, 평균 또는 중간 유의 수준이 결정된다.

[0488] 일부 실시양태에서, 유의 수준은 Z-점수 또는 p-값이다. 일부 실시양태에서, Z-점수는 하기 식에 따라 계산된다:

[0489] $Z_i = (E_i - \text{Med.}E_{(n)})/\text{MAD}$

[0490] 상기 식에서, E_i는 별개의 분절 i의 수준의 정량적 측정치이고, $\text{Med.}E_{(n)}$은 슬라이딩 에지 프로세스에 의해 생성된 모든 별개의 분절들에 대한 중간 수준이고, MAD는 $\text{Med.}E_{(n)}$에 대한 중간 절대 편차이고, Z_i는 별개의 분절 i에 대한 수득된 Z-점수이다. 일부 실시양태에서, MAD는 불확실성의 임의의 적합한 측정치로 대체될 수 있다. 일부 실시양태에서, E_i는 수준의 임의의 적합한 측정치이고, 이의 비한정적 예는 중간 수준, 평균치 수준, 평균 수준, 부분에 대한 카운트의 합계 등을 포함한다.

[0491] 일부 실시양태에서, 슬라이딩 에지 프로세스에 의해 생성된 모든 별개의 분절들에 대한 중간, 평균 또는 평균치 Z-점수가 측정되고, 상기 중간, 평균 또는 평균치 Z-점수로부터 불확실성 수준(예를 들면, MAD)이 생성된다. 일부 실시양태에서, 별개의 분절(예를 들면, 확인된 최초 별개의 분절)은 슬라이딩 에지 프로세스에 의해 생성된 모든 별개의 분절들에 대해 측정된 중간, 평균 또는 평균치 Z-점수, 및 중간, 평균 또는 평균치 Z-점수에 대한 불확실성 수준에 따라 검증되거나 검증되지 않는다. 일부 실시양태에서, 유의 수준(예를 들면, Z-점수)에 대한 예정된 범위(예를 들면, 역치 범위)가 예정된다. 일부 실시양태에서, 후보 분절의 부재에 대한 Z-점수의 예정된 범위는 약 3.5 내지 약 -3.5, 약 3.25 내지 약 -3.25, 약 3.0 내지 약 -3.0, 약 2.75 내지 약 -2.75, 또는 약 2.5 내지 약 -2.5이다. 일부 실시양태에서, 예정된 범위 밖의 값을 갖는 중간, 평균 또는 평균치 Z-점수는 "슬라이딩 에지" 방법에 따라 검증된 별개의 분절의 존재를 확인시켜준다. 일부 실시양태에서, 예정된 범위 내의 값을 갖는 중간, 평균 또는 평균치 Z-점수는 "슬라이딩 에지" 방법에 따라 후보 분절을 검증하지 않고/않거나 후보 분절의 부재(예를 들면, 검증된 후보 분절의 부재)를 확인시켜준다. 일부 실시양태에서, 약 2, 2.25, 2.5, 2.75, 3.0, 3.25 또는 3.5 초과의 절대 값을 갖는 중간, 평균 또는 평균치 Z-점수는 "슬라이딩 에지" 방법에 따라 별개의 분절의 존재를 확인시켜주고/주거나 별개의 분절을 검증한다. 일부 실시양태에서, 약 2, 2.25, 2.5, 2.75, 3.0, 3.25 또는 3.5 미만의 절대 값을 갖는 중간, 평균 또는 평균치 Z-점수는 "슬라이딩 에지" 방법에 따라 후보 분절의 부재를 확인시켜주고/주거나 후보 분절을 검증하지 않는다. 일부 실시양태에서, 중간 Z-점수와 관련된 불확실성 값은 부분적으로 별개의 분절이 검증되는지 아니면 검증되지 않는지를 확인시켜준다. 일부 실시양태에서, 후보 분절은 중간, 평균 또는 평균치 Z-점수가 역치 범위 밖에 있고 불확실성 값(예를 들면, MAD)

이 불확실성 값의 0%(예를 들면, 중첩되지 않음), 5%, 10%, 20%, 25%, 30%, 35% 또는 40% 미만의 수준으로 역치 범위와 중첩되는 경우 검증된다. 일부 실시양태에서, 후보 분절은 중간, 평균 또는 평균치 Z-점수가 역치 범위 밖에 있고 불확실성 값(예를 들면, MAD)이 불확실성 값의 약 25%, 30%, 40%, 50%, 60% 또는 약 70% 초과의 수준으로 역치 범위와 중첩되는 경우 검증되지 않는다.

[0492] 일부 실시양태에서, 분포는 슬라이딩 에지 방법에 의해 생성된 모든 별개의 분절들에 대해 측정된 유의 수준(예를 들면, Z-점수)에 대해 생성된다(예를 들면, 도 13 및 14 참조). 일부 실시양태에서, 별개의 분절은 중간, 평균 또는 평균치 유의 수준 및/또는 유의 수준의 분포에 따라 검증되거나 검증되지 않는다. 일부 실시양태에서, 분포의 약 50%, 60%, 70%, 75%, 80%, 85%, 90% 또는 약 95% 이상이 유의 수준에 대한 예정된 범위 밖에 있고, 별개의 분절이 검증된다. 예를 들면, 3.0 내지 -3.0의 예정된 Z-점수 범위의 경우, 검증된 후보 분절은 중간 Z-점수, 및 3.0 초과의 절대 값을 갖는 Z-점수의 분포의 70% 이상을 가질 수 있다.

[0493] 리브 원 아웃 검증

[0494] 일부 실시양태에서, 검증은 "리브 원 아웃" 프로세스를 포함한다. 적합한 "리브 원 아웃" 프로세스가 이용될 수 있다. 일부 실시양태에서, "리브 원 아웃" 프로세스는 선택된 기준 샘플 세트와 관련된 신뢰 수준을 제공한다. 일부 실시양태에서, "리브 원 아웃" 프로세스는 선택된 기준 샘플 세트와 관련된 불확실성 수준을 제공한다. 일부 실시양태에서, "리브 원 아웃" 프로세스는 선택된 기준 샘플 세트에 따라 측정된 신뢰 수준 및/또는 불확실성 수준에 따라 후보 분절을 검증하거나 검증하지 않는다.

[0495] 일부 실시양태에서, "리브 원 아웃" 프로세스는 검사 샘플 및 2개 이상의 기준 샘플(예를 들면, 본원에서 종종 최초 세트로서 지칭되는 기준 샘플 세트)에 대해 수행된다. 일부 실시양태에서, 검사 샘플은 2개 이상의 기준 샘플 중 하나로서 포함된다. 일부 실시양태에서, 검사 샘플은 2개 이상의 기준 샘플 중 하나로서 포함되지 않는다. 일부 실시양태에서, "리브 원 아웃" 프로세스는 최초 샘플 세트로부터 2개 이상의 기준 샘플 중 하나를 제거하여 기준 샘플 서브세트를 제공하는 단계를 포함한다. 일부 실시양태에서, 최초 세트로부터 기준 샘플을 제거하는 프로세스는 상기 세트 내의 각각의 기준 샘플에 대해 반복된다. 종종, 기준 샘플이 최초 세트로부터 제거될 때, 존재하는 경우 먼저 제거된 기준 샘플이 최초 세트로 복귀된다. 일부 실시양태에서, 1개의 기준 샘플만이 임의의 한 서브세트로부터 제거된다. 결과는 종종 다수의 기준 샘플 서브세트들(본원에서 종종 다수의 샘플 서브세트들로서 지칭됨)이고, 이때 각각의 서브세트는 최초 세트로부터 기준 샘플들 중 하나를 결여한다.

[0496] 일부 실시양태에서, "리브 원 아웃" 프로세스는 기준 샘플 서브세트의 각각의 서브세트에 따라 유의 수준을 측정하는 단계를 포함한다. 일부 실시양태에서, 평균, 평균치 또는 중간 유의 수준은 모든 서브세트들에 대해 측정된 유의 수준 값으로부터 계산된다. 일부 실시양태에서, 불확실성 수준(예를 들면, MAD)은 평균, 평균치 또는 중간 유의 수준에 따라 계산된다. 일부 실시양태에서, 별개의 분절은 "리브 원 아웃" 프로세스에 따라 생성된 중간, 평균 또는 평균치 유의 수준 및/또는 불확실성 수준에 따라 검증되거나 검증되지 않는다.

[0497] "리브 원 아웃" 프로세스의 일부 실시양태에서, 유의 수준은 Z-점수 또는 p-값이다. 일부 실시양태에서, "리브 원 아웃" 프로세스에 대한 Z-점수는 하기 식에 따라 계산된다:

[0498] $Z_i = (E_i - \mathrm{Med}.E_{(n)})/\mathrm{MAD}$

[0499] 상기 식에서, E_i는 분절 i의 수준의 정량적 측정치이고, $\mathrm{Med}.E_{(n)}$은 기준 샘플 서브세트에 대한 분절 i에 대한 중간 수준이고, MAD는 $\mathrm{Med}.E_{(n)}$에 대한 중간 절대 편차이고, Z_i는 분절 i에 대한 수득된 Z-점수이다. 일부 실시양태에서, MAD는 불확실성의 임의의 적합한 측정치로 대체될 수 있다. 일부 실시양태에서, E_i는 수준의 임의의 적합한 측정치이고, 이의 비한정적 예는 중간 수준, 평균치 수준, 평균 수준, 부분에 대한 카운트의 합계 등을 포함한다.

[0500] 일부 실시양태에서, 검증은 "슬라이딩 에지" 프로세스 및 "리브 원 아웃" 프로세스를 포함한다. 예를 들면, 일부 실시양태에서, (예를 들면, "리브 원 아웃" 프로세스로부터 생성된) 기준 샘플 서브세트는 "슬라이딩 에지 프로세스"에 의해 생성된 기준 샘플 세트로부터 생성된다. 예를 들면, 주어진 검사 샘플에 대해, "슬라이딩 에지" 프로세스는 분할 프로세스로부터 확인된 별개의 분절에 대해 225개의 분절을 생성할 수 있고, 그 후 "리브 원 아웃" 프로세스는 10개 기준 샘플의 세트를 사용함으로써 수행된다. 상기 예에서, 복합 중간, 평균 또는 평균치 유의 수준(예를 들면, 복합 중간 Z-점수) 및 복합 불확실성 수준(예를 들면, 복합 MAD)이 수득된 2250개의 Z-점수로부터 계산된다. 일부 실시양태에서, 분할 프로세스에 의해 확인된 별개의 분절은 복합 중간 유의 수준(예를 들면, 복합 중간 Z-점수) 및/또는 복합 불확실성 수준(예를 들면, 복합 MAD)에 따라 검증되거나 검증되지

않는다.

[0501] 일부 실시양태에서, 결정 분석은 후보 분절(예를 들면, 복합 후보 분절)에 대한 Z-점수 또는 복합 Z-점수에 따라 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 단계를 포함한다. 일부 실시양태에서, 후보 분절은 삼염색체성을 표시하고, 후보 분절은 전체 염색체를 표시하는 부분 세트에 대한 후보 분절이다. 일부 실시양태에서, 후보 분절은 전체 염색체를 표시하는 부분 세트에 대한 절대 Z-점수가 예정된 값 또는 역치 이상일 때 전체 염색체 이배수성을 표시한다(예를 들면, 도 7 참조). 일부 실시양태에서, 후보 분절은 전체 염색체를 표시하는 부분 세트에 대한 절대 Z-점수가 약 2.8, 2.9, 3.0, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.75, 3.8, 3.85, 3.9, 3.95, 4.0, 4.05, 4.1, 4.15, 4.2, 4.3, 4.4 또는 약 4.5의 예정된 값 이상일 때 전체 염색체 이배수성을 표시한다. 일부 실시양태에서, 후보 분절은 전체 염색체를 표시하는 부분 세트에 대한 절대 Z-점수가 3.95 이상일 때 삼염색체성을 표시한다. 일부 실시양태에서, 후보 분절은 전체 염색체를 표시하는 부분 세트에 대한 절대 Z-점수가 (i) Haar 웨이블렛 분해 프로세스에 따라 확인된 별개의 분절 또는 (ii) CBS 프로세스에 따라 확인된 별개의 분절에 대해 측정된 Z-점수의 절대 값 이상일 때 삼염색체성을 표시한다. 일부 실시양태에서, 후보 분절은 전체 염색체를 표시하는 부분 세트에 대한 절대 Z-점수가 (i) Haar 웨이블렛 분해 프로세스에 따라 확인된 별개의 분절 또는 (ii) CBS 프로세스에 따라 확인된 별개의 분절에 대해 측정된 Z-점수의 절대 값의 배수 이상일 때 삼염색체성을 표시한다. 일부 실시양태에서, Z-점수의 절대 값의 배수는 약 0.4, 0.5, 0.6, 0.7, 0.8 또는 약 0.9와 곱해진 Z-점수의 절대 값이다.

[0502] 일부 실시양태에서, 후보 분절(예를 들면, 유의한 후보 분절)은 전체 염색체를 표시하는 부분 세트에 대한 절대 Z-점수가 3.95 이상이고 (i) Haar 웨이블렛 분해 프로세스에 따라 확인된 별개의 분절 또는 (ii) CBS 프로세스에 따라 확인된 별개의 분절에 대해 측정된 Z-점수의 절대 값 이상일 때 삼염색체성을 표시한다. 일부 실시양태에서, 후보 분절은 전체 염색체를 표시하는 부분 세트에 대한 절대 Z-점수가 3.95 이상이고 (i) Haar 웨이블렛 분해 프로세스에 따라 확인된 별개의 분절 또는 (ii) CBS 프로세스에 따라 확인된 별개의 분절에 대해 측정된 Z-점수의 절대 값의 배수 이상일 때 삼염색체성을 표시한다. 일부 실시양태에서, Z-점수의 절대 값의 배수는 약 0.4, 0.5, 0.6, 0.7, 0.8 또는 약 0.9와 곱해진 Z-점수의 절대 값이다.

[0503] 일부 실시양태에서, 후보 분절은 삼염색체성을 표시하지 않고, 미세결실 또는 미세중복의 존재는 (i) Haar 웨이블렛 분해 프로세스에 따라 확인된 별개의 분절 또는 (ii) CBS 프로세스에 따라 확인된 별개의 분절에 대해 측정된 Z-점수의 절대 값이 약 2.8, 2.9, 3.0, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.75, 3.8, 3.85, 3.9, 3.95, 4.0, 4.05, 4.1, 4.15, 4.2, 4.3, 4.4 또는 약 4.5 이상일 때 확인된다. 일부 실시양태에서, 후보 분절은 삼염색체성을 표시하지 않고, 미세결실 또는 미세중복의 존재는 확인된다. 일부 실시양태에서, 후보 분절은 삼염색체성을 표시하지 않고, 미세결실 또는 미세중복의 존재는 (i) Haar 웨이블렛 분해 프로세스에 따라 확인된 별개의 분절 또는 (ii) CBS 프로세스에 따라 확인된 별개의 분절에 대해 측정된 Z-점수의 절대 값이 3.95 이상일 때 확인된다. 일부 실시양태에서, 후보 분절은 삼염색체성을 표시하지 않고, 미세결실 또는 미세중복의 존재는 확인되고, Haar 웨이블렛 분해 프로세스에 따라 확인된 별개의 분절은 CBS 프로세스에 따라 확인된 별개의 분절과 실질적으로 동일하다.

[0504] 일부 실시양태에서, 결과의 확인(예를 들면, 유전적 변이, 예를 들면, 태아에서의 유전적 변이의 존재 또는 부재의 확인)은 결정 분석을 포함한다. 일부 실시양태에서, 거짓 음성 및 거짓 양성 확인을 감소시키면서 태아에서의 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 방법은 결정 분석을 포함한다. 일부 실시양태에서, 결정 분석은 일련의 방법들 또는 방법 단계들을 포함한다. 결정 분석의 비한정적 예는 도 6 내지 8에 제시되어 있고 본원에 기재되어 있다.

[0505] 결과의 용도

[0506] 유전적 변이의 존재 또는 부재를 확인시켜주는 하나 이상의 결과를 포함하는 보고서를 제공받는 건강관리 전문가 또는 다른 자격을 갖춘 개체는 상기 보고서에서 검사 대상체 또는 환자의 상태에 대해 판정하는 디스플레이된 데이터를 사용할 수 있다. 일부 실시양태에서, 건강관리 전문가는 제공된 결과를 기초로 권고할 수 있다. 일부 실시양태에서, 건강관리 전문가 또는 자격을 갖춘 개체는 보고서에서 제공된 결과 값 또는 값들 및 관련된 신뢰 파라미터를 기초로 유전적 변이의 존재 또는 부재에 대한 판정 또는 점수를 검사 대상체 또는 환자에게 제공할 수 있다. 일부 실시양태에서, 건강관리 전문가 또는 자격을 갖춘 개체는 제공된 보고서의 시각적 관찰을 이용하여 점수 또는 판정을 수동으로 만든다. 일부 실시양태에서, 점수 또는 판정은 종종 소프트웨어로 구현된 자동화된 루틴(routine)에 의해 만들어지고 정보를 검사 대상체 또는 환자에게 제공하기 전에 정확성에 대해 건강관리 전문가 또는 자격을 갖춘 개체에 의해 검토된다. 본원에서 사용된 용어 "보고서를 제공받는"은 검토 시

건강관리 전문가 또는 다른 자격을 갖춘 개체가 검사 대상체 또는 환자에서의 유전적 변이의 존재 또는 부재에 대해 확인할 수 있게 하는, 결과를 포함하는 작성된 및/또는 도식적 표시치를 통신 수단으로 수득하는 것을 지칭한다. 상기 보고서는 컴퓨터 또는 인간 데이터 도입에 의해 생성될 수 있고 전자적 수단의 이용을 통해(예를 들면, 동일한 또는 상이한 물리적 장소에서 인터넷 상에서, 컴퓨터를 통해, 팩스를 통해, 또는 한 네트워크 위치로부터 또 다른 위치로), 또는 데이터를 보내거나 받는 또 다른 방법(예를 들면, 우편 서비스, 택배 서비스 등)에 의해 전달될 수 있다. 일부 실시양태에서, 결과는 언어, 문서 또는 다른 파일 형태를 포함하나 이들로 한정되지 않는 적합한 매체로 건강관리 전문가에게 전달된다. 상기 파일은 예를 들면, 음향 파일, 컴퓨터 판독가능한 파일, 종이 파일, 실험실 파일 또는 의료 기록 파일일 수 있으나 이들로 한정되지 않는다.

[0507] 본원에서 사용된 용어 "결과를 제공하는" 및 이의 문법적 등가어는 실험실로부터 정보(예를 들면, 실험실 파일)를 수득하는 것을 포함하나 이것으로 한정되지 않는, 이러한 정보를 수득하는 방법도 지칭할 수 있다. 실험실 파일은 의학적 질환의 존재 또는 부재를 확인하기 위해 하나 이상의 분석 또는 하나 이상의 데이터 프로세싱 단계를 수행한 실험실에 의해 생성될 수 있다. 실험실은 실험실 파일로부터 의학적 질환의 존재 또는 부재를 확인하는 사람과 동일한 위치 또는 상이한 위치(예를 들면, 또 다른 국가)에 존재할 수 있다. 예를 들면, 실험실 파일은 한 위치에서 생성되어, 그 내의 정보가 임신 여성 대상체에게 전달될 또 다른 위치로 전달될 수 있다. 일부 실시양태에서, 실험실 파일은 유형 형태 또는 전자적 형태(예를 들면, 컴퓨터 판독가능한 형태)로 존재할 수 있다.

[0508] 일부 실시양태에서, 결과는 실험실로부터 건강관리 전문가, 의사 또는 자격을 갖춘 개체에게 제공될 수 있고, 건강관리 전문가, 의사 또는 자격을 갖춘 개체는 결과를 기초로 진단할 수 있다. 일부 실시양태에서, 결과는 실험실로부터 건강관리 전문가, 의사 또는 자격을 갖춘 개체에게 제공될 수 있고, 건강관리 전문가, 의사 또는 자격을 갖춘 개체는 부분적으로 추가 데이터 및/또는 정보 및 다른 결과와 함께 결과를 기초로 진단할 수 있다.

[0509] 건강관리 전문가 또는 자격을 갖춘 개체는 보고서에서 제공된 결과 또는 결과들을 기초로 적합한 권고를 제공할 수 있다. 제공된 결과 보고서를 기초로 제공될 수 있는 권고의 비한정적 예는 수술, 방사선요법, 화학요법, 유전 상담, 출생 후 치료 해법(예를 들면, 생활 설계, 장기간 보조 관리, 약제, 증상 치료), 임신 종결, 장기 이식, 수혈 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, 권고는 제공된 결과에 기초한 분류(예를 들면, 다운 증후군, 터너 증후군, T13에서의 유전적 변이와 관련된 의학적 질환, T18에서의 유전적 변이와 관련된 의학적 질환)에 의존한다.

[0510] 실험실 직원(예를 들면, 실험실 관리자)은 유전적 변이의 존재 또는 부재의 확인(또는 검사 영역에 대한 정배수체 또는 비-정배수체의 확인)의 근거가 되는 값(예를 들면, 검사 카운트, 기준 카운트, 편차 수준)을 분석할 수 있다. 아슬아슬한 또는 의문스러운 유전적 변이의 존재 또는 부재에 대한 판정의 경우, 실험실 직원은 검사 대상체로부터의 동일한 또는 상이한 샘플 핵산을 사용하는 동일한 검사를 다시 지시할 수 있고/있거나 상이한 검사(예를 들면, 태아 이배수성 확인의 경우 핵형분석 및/또는 양수천자)를 지시할 수 있다.

[0511] 유전적 변이 및 의학적 질환

[0512] 유전적 변이의 존재 또는 부재는 본원에 기재된 방법 또는 장치를 이용함으로써 확인될 수 있다. 일부 실시양태에서, 하나 이상의 유전적 변이의 존재 또는 부재는 본원에 기재된 방법 및 장치에 의해 제공된 결과에 따라 확인된다. 유전적 변이는 일반적으로 일부 개체들에 존재하는 특정 유전 표현형이고, 종종 유전적 변이는 통계학적으로 유의한 개체 하위집단에 존재한다. 일부 실시양태에서, 유전적 변이는 염색체 이상(예를 들면, 이배수성), 부분적 염색체 이상 또는 모자이크 현상이고, 이들 각각은 본원에 더 상세히 기재되어 있다. 유전적 변이의 비한정적 예는 하나 이상의 결실(예를 들면, 미세결실), 중복(예를 들면, 미세중복), 삽입, 돌연변이, 다형성(예를 들면, 단일 뉴클레오티드 다형성), 융합, 반복부(예를 들면, 짧은 직렬 반복부), 상이한 메틸화 부위, 상이한 메틸화 패턴 등 또는 이들의 조합을 포함한다. 삽입, 반복부, 결실, 중복, 돌연변이 또는 다형성은 임의의 길이를 가질 수 있고, 일부 실시양태에서 약 1 염기 또는 염기쌍(bp) 내지 약 250 메가염기(Mb)의 길이를 갖는다. 일부 실시양태에서, 삽입, 반복부, 결실, 중복, 돌연변이 또는 다형성은 약 1 염기 또는 염기쌍(bp) 내지 약 1,000 킬로염기(kb)의 길이(예를 들면, 약 10 bp, 50 bp, 100 bp, 500 bp, 1 kb, 5 kb, 10 kb, 50 kb, 100 kb, 500 kb 또는 1000 kb의 길이)를 갖는다.

[0513] 유전적 변이는 종종 결실이다. 일부 실시양태에서, 결실은 염색체 또는 DNA 서열의 일부가 상실되는 돌연변이(예를 들면, 유전 이상)이다. 결실은 종종 유전 물질의 상실이다. 임의의 수의 뉴클레오티드가 결실될 수 있다. 결실은 하나 이상의 전체 염색체, 염색체의 분절, 대립형질, 유전자, 인트론, 엑손, 임의의 비-코딩 영역, 임의의 코딩 영역, 이들의 분절 또는 이들의 조합물의 결실을 포함할 수 있다. 결실은 미세결실을 포함할 수 있다.

결실은 단일 염기의 결실을 포함할 수 있다.

[0514] 유전적 변이는 종종 유전자 중복이다. 일부 실시양태에서, 중복은 염색체 또는 DNA 서열의 일부가 카피되고 게놈 내로 다시 삽입되는 돌연변이(예를 들면, 유전 이상)이다. 일부 실시양태에서, 유전자 중복(즉, 중복)은 DNA의 영역의 임의의 중복이다. 일부 실시양태에서, 중복은 게놈 또는 염색체 내에서 종종 직렬로 반복되는 핵산 서열이다. 일부 실시양태에서, 중복은 하나 이상의 전체 염색체, 염색체의 분절, 대립형질, 유전자, 인트론, 엑손, 임의의 비-코딩 영역, 임의의 코딩 영역, 이들의 분절 또는 이들의 조합물의 카피를 포함할 수 있다. 중복은 미세중복을 포함할 수 있다. 중복은 종종 중복된 핵산의 하나 이상의 카피를 포함한다. 중복은 종종 1회 이상 반복된(예를 들면, 1회, 2회, 3회, 4회, 5회, 6회, 7회, 8회, 9회 또는 10회 반복된) 유전 영역으로서 특징규명된다. 중복의 범위는 일부 경우 작은 영역(수천 개의 염기쌍)부터 전체 염색체까지 이를 수 있다. 중복은 종종 상동성 재조합에서의 오류의 결과로서 일어나거나 레트로트랜스포존(retrotransposon) 이벤트로 인해 일어난다. 중복은 특정 종류의 증식 질환들과 관련되어 있다. 중복은 게놈 마이크로어레이 또는 비교 유전 혼성화(CGH)를 이용함으로써 특징규명될 수 있다.

[0515] 유전적 변이는 종종 삽입이다. 삽입은 종종 하나 이상의 뉴클레오티드 염기쌍을 핵산 서열 내로 추가하는 것이다. 삽입은 종종 미세삽입이다. 일부 실시양태에서, 삽입은 염색체의 분절을 게놈, 염색체 또는 이들의 분절 내로 추가하는 것을 포함한다. 일부 실시양태에서, 삽입은 대립형질, 유전자, 인트론, 엑손, 임의의 비-코딩 영역, 임의의 코딩 영역, 이들의 분절 또는 이들의 조합물을 게놈 또는 이의 분절 내로 추가하는 것을 포함한다. 일부 실시양태에서, 삽입은 기원이 공지되어 있지 않은 핵산을 게놈, 염색체 또는 이들의 분절 내로 추가(즉, 삽입)하는 것을 포함한다. 일부 실시양태에서, 삽입은 단일 염기의 추가(즉, 삽입)를 포함한다.

[0516] 본원에서 사용된 바와 같이, "카피 수 변이"는 일반적으로 유전적 변이 또는 염색체 이상의 한 클래스 또는 종류이다. 카피 수 변이는 결실(예를 들면, 미세결실), 중복(예를 들면, 미세중복) 또는 삽입(예를 들면, 미세삽입)일 수 있다. 종종, 본원에서 사용된 접두사 "미세"는 5 Mb 미만의 길이를 갖는 핵산의 분절이다. 카피 수 변이는 염색체의 분절의 하나 이상의 결실(예를 들면, 미세결실), 중복 및/또는 삽입(예를 들면, 미세중복, 미세삽입)을 포함할 수 있다. 일부 실시양태에서, 중복은 삽입을 포함한다. 일부 실시양태에서, 삽입은 중복이다. 일부 실시양태에서, 삽입은 중복이 아니다. 예를 들면, 종종 부분 내의 서열의 중복은 중복이 발견되는 부분에 대한 카운트를 증가시킨다. 종종, 부분 내의 서열의 중복은 수준을 증가시킨다. 일부 실시양태에서, 제1 수준을 구성하는 부분에 존재하는 중복은 중복이 부재하는 제2 수준에 비해 상대적으로 수준을 증가시킨다. 일부 실시양태에서, 삽입은 부분의 카운트를 증가시키고, 삽입을 나타내는 서열은 동일한 부분 내의 또 다른 위치에서 존재한다(즉, 중복된다). 일부 실시양태에서, 삽입은 부분의 카운트 또는 수준을 유의하게 증가시키지 않고, 삽입되는 서열은 동일한 부분 내의 서열의 중복이 아니다. 일부 실시양태에서, 삽입은 중복으로서 검출되지 않거나 표시되지 않고, 삽입을 표시하는 중복 서열은 동일한 부분에 존재하지 않는다.

[0517] 일부 실시양태에서, 카피 수 변이는 태아 카피 수 변이이다. 종종, 태아 카피 수 변이는 태아의 게놈 내의 카피 수 변이이다. 일부 실시양태에서, 카피 수 변이는 모체 및/또는 태아 카피 수 변이이다. 일부 실시양태에서, 모체 및/또는 태아 카피 수 변이는 임신 여성(예를 들면, 태아를 가진 여성 대상체), 출산한 여성 대상체 또는 태아를 보유할 수 있는 여성의 게놈 내의 카피 수 변이이다. 카피 수 변이는 이종 카피 수 변이일 수 있고, 이때 변이(예를 들면, 중복 또는 결실)는 게놈의 한 대립형질 상에 존재한다. 카피 수 변이는 동종 카피 수 변이일 수 있고, 이때 변이는 게놈의 두 대립형질 상에 존재한다. 일부 실시양태에서, 카피 수 변이는 이종 또는 동종 태아 카피 수 변이이다. 일부 실시양태에서, 카피 수 변이는 이종 또는 동종 모체 및/또는 태아 카피 수 변이이다. 카피 수 변이는 종종 모체 게놈 및 태아 게놈에 존재하거나, 모체 게놈에 존재하고 태아 게놈에 존재하지 않거나, 태아 게놈에 존재하고 모체 게놈에 존재하지 않는다.

[0518] "배수성"은 태아 또는 모체에 존재하는 염색체의 수를 지칭한다. 일부 실시양태에서, "배수성"은 "염색체 배수성"과 동일하다. 예를 들면, 인간에서 상염색체성 염색체는 종종 쌍으로 존재한다. 예를 들면, 유전적 변이의 부재 하에서 대다수의 인간들은 2개의 각각의 상염색체성 염색체(예를 들면, 1번 염색체 내지 22번 염색체)를 갖는다. 인간에서 2개의 상염색체성 염색체의 정상 상보체의 존재는 종종 정배수체로서 지칭된다. "미세배수성"은 의미에서 배수성과 유사하다. "미세배수성"은 종종 염색체의 분절의 배수성을 지칭한다. 용어 "미세배수성"은 종종 염색체 내의 카피 수 변이(예를 들면, 결실, 중복 및/또는 삽입)의 준재 또는 부재(예를 들면, 동형접합 또는 이형접합 결실, 중복 또는 삽입 등 또는 이들의 부재)를 지칭한다. "배수성" 및 "미세배수성"은 종종 프로파일 내의 수준의 카운트의 정규화 후 측정된다. 따라서, 상염색체성 염색체 쌍(예를 들면, 정배수체)을 표시하는 수준은 종종 1의 배수성으로 정규화된다. 유사하게, 중복, 결실 또는 삽입의 부재를 표시하는 염색체의 분절 내의 수준은 종종 1의 미세배수성으로 정규화된다. 배수성 및 미세배수성은 종종 부분-특이적(예를 들면,

부분 특이적) 및 샘플-특이적이다. 배수성은 종종 각각 정배수체(예를 들면, 2개 염색체), 존재하는 1개 염색체(예를 들면, 염색체 결실), 염색체 부재, 3개 염색체(예를 들면, 삼염색체성) 및 4개 염색체를 표시하는 1, ½, 0, 3/2 및 2의 값과 ½의 정배수로서 정의된다. 마찬가지로, 미세배수성은 각각 정배수체(예를 들면, 카피 수 변이 없음), 이형접합 결실, 동형접합 결실, 이형접합 중복 및 동형접합 중복을 표시하는 1, ½, 0, 3/2 및 2의 값과 ½의 정배수로서 정의된다. 태아에 대한 배수성 값의 일부 예들이 표 2에서 제공되어 있다.

[0519] 일부 실시양태에서, 태아의 미세배수성은 태아의 모체(즉, 임신 여성 대상체)의 미세배수성과 일치한다. 일부 실시양태에서, 태아의 미세배수성은 태아의 모체의 미세배수성과 일치하고, 모체 및 태아 둘 다가 동일한 이형접합 카피 수 변이 또는 동형접합 카피 수 변이를 보유하거나, 둘 다가 정배수체이다. 일부 실시양태에서, 태아의 미세배수성은 태아의 모체의 미세배수성과 상이하다. 예를 들면, 종종 태아의 미세배수성은 카피 수 변이에 대한 이형접합성을 갖고, 모체는 카피 수 변이에 대한 동형접합성을 갖고, 태아의 미세배수성은 특정된 카피 수 변이에 대해 모체의 미세배수성과 일치하지 않는다(예를 들면, 동등하지 않다).

[0520] 미세배수성은 종종 기대 수준과 관련되어 있다. 예를 들면, 종종 수준(예를 들면, 프로파일 내의 수준, 종종 카피 수 변이를 실질적으로 포함하지 않는 수준)은 1의 값(예를 들면, 1의 배수성, 1의 미세배수성)으로 정규화되고, 동형접합 중복의 미세배수성은 2이고, 이형접합 중복은 1.5이고, 이형접합 결실은 0.5이고, 동형접합 결실은 0이다.

[0521] 일부 실시양태에서, 존재 또는 부재가 대상체에 대해 확인되는 유전적 변이는 의학적 질환과 관련되어 있다. 따라서, 본원에 기재된 기술은 의학적 질환 또는 의학적 상태와 관련되어 있는 하나 이상의 유전적 변이의 존재 또는 부재를 확인하는 데에 이용될 수 있다. 의학적 질환의 비한정적 예는 지적 장애(예를 들면, 다운 증후군), 비정상적인 세포 증식(예를 들면, 암), 미생물 핵산의 존재(예를 들면, 바이러스, 세균, 진균, 효모) 및 임신중독증과 관련된 의학적 질환을 포함한다.

[0522] 유전적 변이, 의학적 질환 및 상태의 비한정적 예는 이하에 기재되어 있다.

[0523] 태아 성별

[0524] 일부 실시양태에서, 태아 성별 또는 성별 관련 장애(예를 들면, 성염색체 이배수성)의 예측은 본원에 기재된 방법 또는 장치에 의해 결정될 수 있다. 성별 확인은 일반적으로 성염색체에 기초한다. 인간에서, 2개의 성염색체들, 즉 X 염색체 및 Y 염색체가 존재한다. Y 염색체는 남성으로서의 배아 발생을 유발하는 유전자인 SRY를 함유한다. 인간 및 다른 포유동물의 Y 염색체는 정상적인 정자 생성을 위해 필요한 다른 유전자도 함유한다. XX를 갖는 개체는 여성이고 XY는 남성이고, 종종 성염색체 이배수성으로서 지칭되는 비한정적 변이는 XO, XYY, XXX 및 XXY를 포함한다. 일부 실시양태에서, 남성은 2개의 X 염색체 및 1개의 Y 염색체(XXY; 클라인펠터 증후군), 또는 1개의 X 염색체 및 2개의 Y 염색체(XYY 증후군; 자콥 증후군)를 갖고, 일부 여성들은 3개의 X 염색체(XXX; 삼중 X 증후군) 또는 2개 대신에 1개의 X 염색체(XO; 터너 증후군)를 갖는다. 일부 실시양태에서, 개체 내의 일부 세포들만이 성염색체 이배수성에 의해 영향을 받고, 이것은 모자이크 현상(예를 들면, 터너 모자이크 현상)으로서 지칭될 수 있다. 다른 경우는 SRY가 손상되거나(XY 여성을 유발함) X로 카피되는(XX 남성을 유발함) 경우를 포함한다.

[0525] 일부 경우, 자궁에서 태아의 성별을 확인하는 것이 유리할 수 있다. 예를 들면, 하나 이상의 성-연관 장애의 가족력을 갖는 환자(예를 들면, 임신 여성)는 태아가 이러한 장애를 유전받을 위험을 평가하는 것을 돕기 위해 그녀가 보유하는 태아의 성별을 확인하는 것을 원할 수 있다. 성-연관 장애는 X-연관 장애 및 Y-연관 장애를 포함하나 이들로 한정되지 않는다. X-연관 장애는 X-연관 열성 장애 및 X-연관 우성 장애를 포함한다. X-연관 열성 장애의 예는 면역 장애(예를 들면, 만성 육아종 질환(CYBB), 비스콧-알드리치(Wiskott-Aldrich) 증후군, X-연관 중증 복합 면역결핍, X-연관 무감마글로불린혈증, 과다-IgM 증후군 1형, IPEX, X-연관 림프증식 질환, 프로퍼딘(Properdin) 결핍), 혈액 장애(예를 들면, 혈우병 A, 혈우병 B, X-연관 철적혈모구빈혈), 내분비 장애(예를 들면, 안드로겐 불감성 증후군/케네디병(Kennedy disease), KAL1 칼만(Kallmann) 증후군, X-연관 선천성 부신 발육부전), 대사 장애(예를 들면, 오르니틴 트랜스카브아밀라제(transcarbamylase) 결핍, 안뇌신 증후군, 부신백색질형성장애, 글루코스-6-포스페이트 데하이드로게나제(dehydrogenase) 결핍, 피루베이트 데하이드로게나제 결핍, 다논병(Danon disease)/글리코겐 저장 질환 IIb형, 파브리병(Fabry's disease), 헌터(Hunter) 증후군, 레슈-니한(Lesch-Nyhan) 증후군, 멘케스병(Menkes disease)/후두각 증후군), 신경계 장애(예를 들면, 코핀-로우리(Coffin-Lowry) 증후군, MASA 증후군, X-연관 알파 지중해빈혈, 정신지체 증후군, 시더리우스(Siderius) X-연관 정신지체 증후군, 색맹, 눈 백색증, 노리에병(Norrie disease), 범맥락막위축, 샤르코-마리-투쓰병(Charcot-Marie-Tooth disease)(CMTX2-3), 펠리체우스-메르츠바허병(Pelizaeus-Merzbacher disease), SMAX2), 피부 및

관련 조직 장애(예를 들면, 선천성 이상각화증, 땀저하외배엽형성이상증(EDA), X-연관 비늘증, X-연관 내피 각막 퇴행위축), 신경근 장애(예를 들면, 베커 근육 퇴행위축/뒤시엔느(Duchenne), 중심핵근육병증(MTM1), 콘라디-후너만(Conradi-Hunermann) 증후군, 에머리-드레이푸스(Emery-Dreifuss) 근육 퇴행위축 1), 비뇨 장애(예를 들면, 알포트(Alport) 증후군, 덴트병(Dent's disease), X-연관 신장기원 요붕증), 골/치아 장애(예를 들면, AMELX 불완전사기질형성증), 및 다른 장애(예를 들면, 바쓰(Barth) 증후군, 맥레오드(McLeod) 증후군, 스미쓰-피네만-메이어스(Smith-Fineman-Myers) 증후군, 심슨-골라비-베멜(Simpson-Golabi-Behmel) 증후군, 모어 트라네브제르그(Mohr-Tranebjærg) 증후군, 나소디지토어쿠스틱(Nasodigitoacoustic) 증후군)을 포함하나 이들로 한정되지 않는다. X-연관 우성 장애의 예는 X-연관 저인산혈증, 국소 피부 형성저하증, 프래질(Fragile) X 증후군, 에카르디(Aicardi) 증후군, 색소실조증(Incontinentia pigmenti), 레트(Rett) 증후군, CHILD 증후군, 루잔-프린스(Lujan-Fryns) 증후군 및 구강안면손가락(Orofaciodigital) 증후군 1을 포함하나 이들로 한정되지 않는다. Y-연관 장애의 예는 남성 불임, 색소성 망막염 및 무정자증을 포함하나 이들로 한정되지 않는다.

[0526] 염색체 이상

[0527] 일부 실시양태에서, 태아 염색체 이상의 존재 또는 부재는 본원에 기재된 방법 또는 장치를 이용함으로써 확인될 수 있다. 염색체 이상은 전체 염색체 또는 하나 이상의 유전자를 포함하는 염색체 영역의 획득 또는 상실을 포함하나 이들로 한정되지 않는다. 염색체 이상은 일염색체성, 삼염색체성, 다염색체성, 이형접합성의 상실, 하나 이상의 뉴클레오티드 서열(예를 들면, 하나 이상의 유전자)의 전위, 결실 및/또는 중복(불균형 전위에 의해 야기된 결실 및 중복을 포함함)을 포함한다. 본원에서 사용된 용어 "염색체 이상" 또는 "이배수성"은 대상체 염색체의 구조와 정상 상동성 염색체의 구조 사이의 편차를 지칭한다. 용어 "정상"은 특정 종의 건강한 개체, 예를 들면, 정배수체 게놈에서 발견된 우세한 핵형 또는 밴딩 패턴(인간에서 46, XX 또는 46, XY)을 지칭한다. 상이한 유기체가 광범위하게 상이한 염색체 상보체를 갖기 때문에, 용어 "이배수성"은 염색체의 구체적인 수를 지칭하는 것이 아니라, 유기체의 주어진 세포 또는 세포들 내의 염색체 함량이 비정상적인 상황을 지칭한다. 일부 실시양태에서, 본원에서 용어 "이배수성"은 전체 염색체 또는 염색체의 일부의 상실 또는 획득에 의해 야기된 유전 물질의 불균형을 지칭한다. "이배수성"은 염색체의 분절의 하나 이상의 결실 및/또는 삽입을 지칭할 수 있다. 일부 실시양태에서, 용어 "정배수체"는 염색체의 정상 상보체를 지칭한다.

[0528] 본원에서 사용된 용어 "일염색체성"은 정상 상보체의 1개 염색체의 결여를 지칭한다. 부분적 일염색체성은 불균형 전위 또는 결실에서 발생할 수 있고, 이때 염색체의 분절만이 단일 카피로 존재한다. 예를 들면, 성염색체의 일염색체성(45, X)은 터너 증후군을 야기한다. 용어 "이염색체성"은 염색체의 2개 카피의 존재를 지칭한다. 각각의 염색체의 2개 카피를 갖는 유기체, 예컨대, 인간(이배수체 또는 "정배수체"인 유기체)의 경우 이염색체성은 정상 상태이다. 정상적으로 각각의 염색체의 3개 이상의 카피를 갖는 유기체(삼배수체 이상인 유기체)의 경우, 이염색체성은 이배수체 염색체 상태이다. 단위생식 이염색체성에서, 염색체의 카피 둘 다가 동일한 부모로부터 나온다(다른 부모로부터의 기여가 없다).

[0529] 본원에서 사용된 용어 "삼염색체성"은 특정 염색체의 2개 카피 대신에 3개 카피의 존재를 지칭한다. 인간 다운 증후군에서 발견되는 추가 21번 염색체의 존재는 "삼염색체성 21"로서 지칭된다. 삼염색체성 18 및 삼염색체성 13은 2종의 다른 인간 상염색체 삼염색체성이다. 성염색체의 삼염색체성은 여성(예를 들면, 삼중 X 증후군에서 47, XXX) 또는 남성(예를 들면, 클라인펠터 증후군에서 47, XXY; 또는 자콥 증후군에서 47, XYY)에서 관찰될 수 있다. 일부 실시양태에서, 삼염색체성은 대부분의 또는 모든 상염색체의 중복이다. 일부 실시양태에서, 삼염색체성은 (예를 들면, 정배수체에 대한 특정 종류의 염색체의 2개 존재(즉, 한 쌍) 대신에) 특정 종류의 염색체의 3개 존재(예를 들면, 3개 카피)를 초래하는 전체 염색체 이배수성이다.

[0530] 본원에서 사용된 용어 "사염색체성" 및 "오염색체성"은 각각 염색체의 4개 또는 5개 카피의 존재를 지칭한다. 상염색체에서 거의 관찰되지 않지만, XXXX, XXXY, XXYY, XYYY, XXXXX, XXXXY, XXXYY, XXYYY 및 XYYYY를 포함하는 성염색체 사염색체성 및 오염색체성이 인간에서 보고되었다.

[0531] 염색체 이상은 다양한 기작들에 의해 야기될 수 있다. 기작은 (i) 약화된 유사분열 체크포인트의 결과로서 발생하는 비분리, (ii) 다수의 염색체들에서 비분리를 야기하는 불활성 유사분열 체크포인트, (iii) 1개의 동원체가 양쪽 유사분열 방추극에 부착될 때 발생하는 메로텔릭(merotelic) 부착, (iv) 2개 초과의 방추극이 형성될 때 형성되는 다중극 방추, (v) 1개의 방추극만이 형성될 때 형성되는 단일극 방추, 및 (vi) 단일극 방추 기작의 최종 결과로서 발생하는 사배수체 중간체를 포함하나 이들로 한정되지 않는다.

[0532] 본원에서 사용된 용어 "부분적 일염색체성" 및 "부분적 삼염색체성"은 염색체의 일부의 상실 또는 획득에 의해 야기된 유전 물질의 불균형을 지칭한다. 부분적 일염색체성 또는 부분적 삼염색체성은 불균형 전위로부터 발생

할 수 있고, 이때 개체는 2개의 상이한 염색체들의 절단 및 융합으로부터 형성된 유도체 염색체를 보유한다. 이 상황에서, 개체는 한 염색체의 일부의 3개 카피(2개의 정상 카피, 및 유도체 염색체 상에 존재하는 분절), 및 유도체 염색체에 포함된 다른 염색체의 일부의 단지 1개 카피를 가질 것이다.

[0533] 본원에서 사용된 용어 "모자이크 현상"은 유기체의 모든 세포들이 아니라 일부 세포들에서의 이배수성을 지칭한다. 일부 염색체 이상은 모자이크 염색체 이상 및 비-모자이크 염색체 이상으로서 존재할 수 있다. 예를 들면, 일부 삼염색체성 21 개체들은 모자이크 다운 증후군을 갖고, 일부는 비-모자이크 다운 증후군을 갖는다. 상이한 기작들이 모자이크 현상을 유발할 수 있다. 예를 들면, (i) 초기 접합체는 정상적으로 단순 삼염색체성 21을 초래할 3개의 21번째 염색체들을 가질 수 있으나, 세포 분열의 과정 동안 하나 이상의 세포주가 21번째 염색체들 중 하나를 상실하고; (ii) 초기 접합체는 2개의 21번째 염색체들을 가질 수 있으나, 세포 분열의 과정 동안 21번째 염색체들 중 하나가 중복되었다. 체세포 모자이크 현상은 완전한 또는 모자이크 이배수성을 수반하는 유전 증후군과 전형적으로 관련된 기작과 상이한 기작을 통해 발생할 가능성이 있다. 예를 들면, 체세포 모자이크 현상은 일부 종류의 암 및 신경에서 확인되었다. 일부 경우, 삼염색체성 12는 만성 림프구성 백혈병(CLL)에서 확인되었고, 삼염색체성 8은 급성 골수성 백혈병(AML)에서 확인되었다. 또한, 개체가 염색체의 절단에 취약한 유전 증후군(염색체 불안정성 증후군)은 종종 다양한 종류의 암에 대한 증가된 위험과 관련되어 있으므로, 암발생에서 체세포 이배수성의 역할을 강조한다. 본원에 기재된 방법 및 프로토콜은 비-모자이크 염색체 이상 및 모자이크 염색체 이상의 존재 또는 부재를 확인할 수 있다.

[0534] 표 1A 및 1B는 본원에 기재된 방법 및 장치에 의해 잠재적으로 확인될 수 있는 염색체 상태, 증후군 및/또는 이상의 비한정적 목록을 제공한다. 표 1B는 2011년 10월 6일자 DECIPHER 데이터베이스(예를 들면, GRCh37로 맵핑된 위치들에 기초한 버전 5.1; 균일 자원 위치주소(URL) dechipher.sanger.ac.uk에서 입수가능함)로부터의 목록이다.

[0535] [표 1A]

염색체	이상	질환 관련
X	XO	터너 증후군
Y	XXY	클라인펠터 증후군
Y	XYY	이중 Y 증후군
Y	XXX	삼염색체성 X 증후군
Y	XXXX	4 X 증후군
Y	Xp21 결실	뒤시엔느/벡커 증후군, 선천성 부신 발육부전, 만성 육아종 질환
Y	Xp22 결실	스테로이드 설파타제 결핍
Y	Xq26 결실	X-연관 림프증식 질환
1	1p(체세포성) 일염색체성 삼염색체성	신경모세포종
2	일염색체성 삼염색체성 2q	성장 지체, 발달 및 정신 지연, 및 경미한 신체적 기형
3	일염색체성 삼염색체성 (체세포성)	비-호지킨 림프종
4	일염색체성 삼염색체성 (체세포성)	급성 비-림프구성 백혈병(ANLL)
5	5p	묘성; 르준(Lejeune) 증후군
5	5q (체세포성) 일염색체성 삼염색체성	골수형성이상 증후군
6	일염색체성 삼염색체성 (체세포성)	투명 세포 육종
7	7q11.23 결실	윌리암 증후군

[0536]

염색체	이상	질환 관련
7	일염색체성 삼염색체성	소아기의 일염색체성 7 증후군; 체세포성; 신장 피질 선종; 골수형성이상 증후군
8	8q24.1 결실	랑거-귀돈(Langer-Giedon) 증후군
8	일염색체성 삼염색체성	골수형성이상 증후군; 와카니(Warkany) 증후군; 체세포성: 만성 골수성 백혈병
9	일염색체성 9p	알피(Alfi's) 증후군
9	일염색체성 9p 부분적 삼염색체성	레토어(Rethore) 증후군
9	삼염색체성	완전한 삼염색체성 9 증후군; 모자이크성 삼염색체성 9 증후군
10	일염색체성 삼염색체성 (체세포성)	ALL 또는 ANLL
11	11p-	아니리디아(Aniridia); 윌름스(Wilms) 종양
11	11q-	자콥센(Jacobsen) 증후군
11	일염색체성 (체세포성) 삼염색체성	영향을 받은 골수 계통(ANLL, MDS)
12	일염색체성 삼염색체성 (체세포성)	CLL, 청소년 과립층 세포 종양(JGCT)
13	13q-	13q-증후군; 오벨리(Orbeli) 증후군
13	13q14 결실	망막모세포종
13	일염색체성 삼염색체성	파타우(Patau) 증후군
14	일염색체성 삼염색체성 (체세포성)	골수 장애(MDS, ANLL, 비정형 CML)
15	15q11-q13 결실 일염색체성	프라더-윌리(Prader-Willi), 앙겔만(Angelman) 증후군

[0537]

염색체	이상	질환 관련
15	삼염색체성 (체세포성)	영향을 받은 골수 및 림프 계통(예를 들면, MDS, ANLL, ALL, CLL)
16	16q13.3 결실	루벤스테인-테이비(Rubenstein-Taybi)
3	일염색체성 삼염색체성 (체세포성)	유두상 신장 세포 암종(악성)
17	17p-(체세포성)	골수 악성종양에서의 17p 증후군
17	17q11.2 결실	스미쓰-마게니스(Smith-Magenis)
17	17q13.3	밀러-디에커(Miller-Dieker)
17	일염색체성 삼염색체성 (체세포성)	신장 피질 선종
17	17p11.2-12 삼염색체성	샤르코-마리-투쓰(Charcot-Marie Tooth) 증후군 1 형; HNPP
18	18p-	18p 부분적 일염색체성 증후군 또는 그루시 라미 티에프리(Grouchy Lamy Thieffry) 증후군
18	18q-	그루시 라미 살몬 랜드리(Grouchy Lamy Salmon Landry) 증후군
18	일염색체성 삼염색체성	에드워즈(Edwards) 증후군
19	일염색체성 삼염색체성	
20	20p-	삼염색체성 20p 증후군
20	20p11.2-12 결실	알라질(Alagille)
20	20q-	체세포성: MDS, ANLL, 진성적혈구증가증, 만성 호중구 백혈병
20	일염색체성 삼염색체성 (체세포성)	유두상 신장 세포 암종(악성)
22	22q11.2 결실	디조지 증후군, 구개심장안면 증후군, 대동맥궁 기형 안면 증후군, 상염색체 우성 오피츠(Opitz) G/BBB 증후군, 카일러(Caylor) 심장안면 증후군
22	일염색체성 삼염색체성	완전한 삼염색체성 22 증후군

[0538]

[0539] [표 1B]

증후군	염색체	시작	말단	간격(Mb)	등급
12q14 미세결실 증후군	12	65,071,919	68,645,525	3.57	
15q13.3 미세결실 증후군	15	30,769,995	32,701,482	1.93	
15q24 반복성 미세결실 증후군	15	74,377,174	76,162,277	1.79	
15q26 과다성장 증후군	15	99,357,970	102,521,392	3.16	
16p11.2 미세중복 증후군	16	29,501,198	30,202,572	0.70	
16p11.2-p12.2 미세결실 증후군	16	21,613,956	29,042,192	7.43	
16p13.11 반복성 미세결실(신경인지 장애 민감성 좌위)	16	15,504,454	16,284,248	0.78	
16p13.11 반복성 미세중복(신경인지 장애 민감성 좌위)	16	15,504,454	16,284,248	0.78	
17q21.3 반복성 미세결실 증후군	17	43,632,466	44,210,205	0.58	1
1p36 미세결실 증후군	1	10,001	5,408,761	5.40	1
1q21.1 반복성 미세결실(신경발달 장애에 대한 민감성 좌위)	1	146,512,930	147,737,500	1.22	3
1q21.1 반복성 미세중복(신경발달 장애에 대한 가능한 민감성 좌위)	1	146,512,930	147,737,500	1.22	3
1q21.1 저혈소판 요골 무형성(TAR) 증후군에 대한 민감성 좌위	1	145,401,253	145,928,123	0.53	3
22q11 결실 증후군 (구개심장안면/디조지 증후군)	22	18,546,349	22,336,469	3.79	1
22q11 중복 증후군	22	18,546,349	22,336,469	3.79	3
22q11.2 원위 결실 증후군	22	22,115,848	23,696,229	1.58	
22q13 결실 증후군(펠란 맥더미드(Phelan-Mcdermid) 증후군)	22	51,045,516	51,187,844	0.14	1
2p15-16.1 미세결실 증후군	2	57,741,796	61,738,334	4.00	
2q33.1 결실 증후군	2	196,925,089	205,206,940	8.28	1
2q37 일염색체성	2	239,954,693	243,102,476	3.15	1

[0540]

증후군	염색체	시작	말단	간격(Mb)	등급
3q29 미세결실 증후군	3	195,672,229	197,497,869	1.83	
3q29 미세중복 증후군	3	195,672,229	197,497,869	1.83	
7q11.23 중복 증후군	7	72,332,743	74,616,901	2.28	
8p23.1 결실 증후군	8	8,119,295	11,765,719	3.65	
9q 서브텔로머성(subtelomeric) 결실 증후군	9	140,403,363	141,153,431	0.75	1
성인 발병 상염색체 우성 백색질형성장애(ADLD)	5	126,063,045	126,204,952	0.14	
앙겔만 증후군(1 형)	15	22,876,632	28,557,186	5.68	1
앙겔만 증후군(2 형)	15	23,758,390	28,557,186	4.80	1
ATR-16 증후군	16	60,001	834,372	0.77	1
AZFa	Y	14,352,761	15,154,862	0.80	
AZFb	Y	20,118,045	26,065,197	5.95	
AZFb+ AZFc	Y	19,964,826	27,793,830	7.83	
AZFc	Y	24,977,425	28,033,929	3.06	
고양이-눈 증후군(I 형)	22	1	16,971,860	16.97	
샤르코-마리-투쓰 증후군 1A 형(CMT1A)	17	13,968,607	15,434,038	1.47	1
묘성 증후군(5p 결실)	5	10,001	11,723,854	11.71	1
뇌 아밀로이드 혈관병증을 갖는 초기 발병 알쯔하이머병	21	27,037,956	27,548,479	0.51	
가족성 선종 폴립증	5	112,101,596	112,221,377	0.12	
압박 마비 유전 체질(HNPP)	17	13,968,607	15,434,038	1.47	1
레리-웨일(Leri-Weill) 연골뼈형성이상증(LWD) - SHOX 결실	X	751,878	867,875	0.12	
레리-웨일 연골뼈형성이상증(LWD) - SHOX 결실	X	460,558	753,877	0.29	
밀러-디에커(Miller-Dieker) 증후군(MDS)	17	1	2,545,429	2.55	1
NF1-미세결실 증후군	17	29,162,822	30,218,667	1.06	1
펠리체우스-메르츠바허병(Pelizaeus-Merzbacher disease)	X	102,642,051	103,131,767	0.49	
포토키-룹스키(Potocki-	17	16,706,021	20,482,061	3.78	

[0541]

증후군	염색체	시작	말단	간격(Mb)	등급
Lupski) 증후군 (17p11.2 중복 증후군)					
포토키 샤퍼(Potocki-Shaffer) 증후군	11	43,985,277	46,064,560	2.08	1
프라더-윌리 증후군(1 형)	15	22,876,632	28,557,186	5.68	1
프라더-윌리 증후군(2 형)	15	23,758,390	28,557,186	4.80	1
RCAD(신장낭 및 당뇨병)	17	34,907,366	36,076,803	1.17	
루빈스타인-테이비 증후군	16	3,781,464	3,861,246	0.08	1
스미쓰-마게니스 증후군	17	16,706,021	20,482,061	3.78	1
소토스(Sotos) 증후군	5	175,130,402	177,456,545	2.33	1
갈림손/발 기형 1(SHFM1)	7	95,533,860	96,779,486	1.25	
스테로이드 설파타제 결핍(STS)	X	6,441,957	8,167,697	1.73	
WAGR 11p13 결실 증후군	11	31,803,509	32,510,988	0.71	
윌리암스-보이렌(Williams-Beuren) 증후군(WBS)	7	72,332,743	74,616,901	2.28	1
울프-허쉬호른(Wolf-Hirschhorn) 증후군	4	10,001	2,073,670	2.06	1
Xq28(MECP2) 중복	X	152,749,900	153,390,999	0.64	

[0542]

[0543] 1 등급 상태는 종종 하기 특성들 중 하나 이상을 갖는다: 병원성 비정상; 유전학자들 사이의 강한 동의; 고도 침투성; 여전히 가변성 표현형을 가질 수 있으나 일부 공통된 특성을 가짐; 문헌의 모든 사례들이 임상 표현형을 가짐; 비정상을 갖는 건강한 개체의 사례 부재; DVG 데이터베이스 상에서 보고되어 있지 않거나 건강한 집단에서 발견되지 않음; 단일 유전자 또는 다중 유전자 용량 효과를 확인시켜주는 기능성 데이터; 확인된 또는 강한 후보 유전자; 정의된 임상 관리 영향; 감시할 가치가 있는 공지된 암 위험; 다수의 정보 공급원(OMIM, 진리뷰스(Genereviews), 오파넷(Orphanet), 유니크(Unique), 위키페디아(Wikipedia)); 및/또는 진단 용도(생식 상담)를 위한 사용가능성.

[0544] 2 등급 상태는 종종 하기 특성들 중 하나 이상을 갖는다: 가능성이 있는 병원성 비정상; 고도 침투성; DD 이외의 일관된 특징을 갖지 않는 가변성 표현형; 문헌에서의 작은 수의 사례/보고; 모든 보고된 사례가 임상 표현형을 가짐; 기능성 데이터 또는 확인된 병원성 유전자의 부재; 다수의 정보 공급원(OMIM, 진리뷰스, 오파넷, 유니크, 위키페디아); 및/또는 진단 목적 및 생식 상담을 위해 사용될 수 있음.

[0545] 3 등급 상태는 종종 하기 특성들 중 하나 이상을 갖는다: 민감성 좌위; 기재된 발단자(proband)의 건강한 개체 또는 영향을 받지 않은 부모; 대조군 집단에의 존재; 비-침투성; 경미하고 특이적이지 않은 표현형; 덜 일관된 특징; 기능성 데이터 또는 확인된 병원성 유전자의 부재; 보다 더 한정된 데이터 공급원; 제2 진단의 가능성이 대다수로부터 벗어나는 사례 또는 신규 임상 발견이 존재하는 경우에 대한 가능성을 남김; 및/또는 진단 목적으로 사용될 때 주의 및 생식 상담에 대한 신중한 충고.

[0546] 임신중독증

[0547] 일부 실시양태에서, 임신중독증의 존재 또는 부재는 본원에 기재된 방법 또는 장치를 이용함으로써 확인된다. 임신중독증은 임신 동안에 고혈압이 발생하는 상태(즉, 임신에 의해 유도된 고혈압)이고 소변 중의 상당한 양의 단백질과 관련되어 있다. 일부 실시양태에서, 임신중독증은 상승된 수준의 세포외 핵산 및/또는 메틸화 패턴의 변경과도 관련되어 있다. 예를 들면, 세포외 태아로부터 유래된 과다메틸화된 RASSF1A 수준과 임신중독증의 중증도 사이에 양의 상관관계가 관찰되었다. 일부 예에서, 증가된 DNA 메틸화는 정상 대조군에 비해 임신중독증성 태반에서 H19 유전자에 대해 관찰된다.

[0548] 임신중독증은 전세계적으로 모체 및 태아/신생아 사망률 및 이환율의 주원인들 중 하나이다. 혈장 및 혈청 중의 순환 무세포 핵산은 출생전 진단을 포함하는 상이한 의학 분야들에서 기대되는 임상 적용을 갖는 신규 생체마커이다. 임박한 임신중독증에 대한 표시자로서 모체 혈장 중의 무세포 태아 (cff)DNA의 정량적 변화가 예를 들면, 남성 특이적 SRY 또는 DYS 14 좌위에 대한 실시간 정량 PCR을 이용한 상이한 연구들에서 보고되었다. 초기 발병 임신중독증의 경우, 상승된 수준이 처음 3개월에서 관찰될 수 있다. 증상의 발생 전 증가된 수준의 cffDNA는 조직 산화 스트레스 및 증가된 태반 아폽토시스 및 괴사를 유발하는, 융모사이 공간 내의 저산소증/재산소화에 기인할 수 있다. 모체 순환계 내로의 cffDNA의 증가된 배출에 대한 증거 이외에 임신중독증에서 cffDNA의 감소된 신장 제거에 대한 증거도 존재한다. 태아 DNA의 양이 현재 Y 염색체 특이적 서열을 정량함으로써 측정되기 때문

에, 대안적 방법, 예컨대, 총 무세포 DNA의 측정 또는 성별 독립적 태아 유전외적 마커, 예컨대, DNA 메틸화의 사용이 대안을 제공한다. 태반 유래의 무세포 RNA는 임상 실시에서 임신중독증을 스크리닝하고 진단하는 데에 사용될 수 있는 또 다른 대안적 생체마커이다. 태아 RNA는 이를 분해로부터 보호하는, 세포보다 더 작은 태반 입자와 회합되어 있다. 태아 RNA 수준은 종종 대조군에 비해 임신중독증을 갖는 임신 여성에서 10배 더 높으므로, 임상 실시에서 임신중독증을 스크리닝하고 진단하는 데에 사용될 수 있는 대안적 생체마커이다.

[0549] 병원체

[0550] 일부 실시양태에서, 병원성 상태의 존재 또는 부재는 본원에 기재된 방법 또는 장치에 의해 확인된다. 병원성 상태는 세균, 바이러스 또는 진균을 포함하나 이들로 한정되지 않는 병원체에 의한 숙주의 감염에 의해 야기될 수 있다. 병원체는 전형적으로 숙주 핵산과 식별될 수 있는 핵산(예를 들면, 게놈 DNA, 게놈 RNA, mRNA)을 보유하기 때문에, 본원에서 제공된 방법 및 장치는 병원체의 존재 또는 부재를 확인하는 데에 이용될 수 있다. 종종, 병원체는 특정 병원체 특유의 특성, 예를 들면, 유전외적 상태 및/또는 하나 이상의 서열 변경, 중복 및/또는 결실을 갖는 핵산을 보유한다. 따라서, 본원에서 제공된 방법은 특정 병원체 또는 병원체 변이체(예를 들면, 균주)를 확인하는 데에 사용될 수 있다.

[0551] 암

[0552] 일부 실시양태에서, 세포 증식 장애(예를 들면, 암)의 존재 또는 부재는 본원에 기재된 방법 또는 장치를 이용함으로써 확인된다. 예를 들면, 혈청 중의 무세포 핵산의 수준은 건강한 환자에 비해 다양한 종류의 암을 갖는 환자들에서 상승될 수 있다. 예를 들면, 전이성 질환을 갖는 환자는 종종 비-전이성 환자보다 대략 2배 더 높은 혈청 DNA 수준을 갖는다. 전이성 질환을 갖는 환자는 예를 들면, 암 특이적 마커 및/또는 일부 단일 뉴클레오티드 다형성 또는 짧은 직렬 반복부에 의해 확인될 수도 있다. 상승된 수준의 순환 DNA와 양의 상관관계를 가질 수 있는 암 종류의 비한정적 예는 유방암, 대장암, 위장암, 간세포암, 폐암, 흑색종, 비-호지킨 림프종, 백혈병, 다발성 골수종, 방광암, 간암, 자궁경부암, 식도암, 췌장암 및 전립선암을 포함한다. 다양한 암들이 비-암성 건강한 세포로부터의 핵산과 식별될 수 있는 특성, 예를 들면, 유전외적 상태 및/또는 서열 변경, 중복 및/또는 결실을 갖는 핵산을 보유할 수 있고 종종 이러한 핵산을 혈류 내로 방출할 수 있다. 이러한 특성은 예를 들면, 특정 종류의 암에 특이적일 수 있다. 따라서, 본원에서 제공된 방법이 특정 종류의 암을 확인하는 데에 이용될 수 있다는 것도 고려된다.

[0553] 소프트웨어는 이하에 더 상세히 기재된 바와 같이 본원에 기재된 프로세스들에서 카운팅, 데이터 프로세싱, 결과의 생성, 및/또는 생성된 결과에 기초한 하나 이상의 권고의 제공을 포함하나 이들로 한정되지 않는 하나 이상의 단계를 수행하는 데에 사용될 수 있다.

[0554] 기계, 소프트웨어 및 인터페이스

[0555] 본원에 기재된 일부 프로세스들 및 방법들(예를 들면, 서열 리드, 카운트, 수준(예를 들면, 수준들) 및/또는 프로파일의 정량, 맵핑, 정규화, 범위 설정, 조절, 분류, 카운팅 및/또는 측정)은 종종 컴퓨터, 프로세서, 소프트웨어, 모듈 또는 다른 장치 없이 수행될 수 없다. 본원에 기재된 방법은 전형적으로 컴퓨터에 의해 실행되는 방법이고, 방법의 하나 이상의 부분은 종종 하나 이상의 프로세서(예를 들면, 마이크로프로세서), 컴퓨터, 또는 마이크로프로세서 제어 장치에 의해 수행된다. 본 명세서에 기재된 방법에 대한 실시양태들은 일반적으로 본원에 기재된 시스템, 장치 및 컴퓨터 프로그램 제품에서 지시에 의해 실행되는 동일한 또는 관련된 프로세스들에 적용될 수 있다. 일부 실시양태에서, 본원에 기재된 프로세스 및 방법(예를 들면, 서열 리드, 카운트, 수준 및/또는 프로파일의 정량, 카운팅 및/또는 측정)은 자동화된 방법에 의해 수행된다. 일부 실시양태에서, 본원에 기재된 하나 이상의 단계 및 방법은 프로세서 및/또는 컴퓨터에 의해 수행되고/되거나, 메모리와 함께 수행된다. 일부 실시양태에서, 서열 리드, 카운트, 맵핑, 맵핑된 서열 태그, 수준 또는 프로파일의 측정, 정규화, 비교, 범위 설정, 분류, 조절, 작도, 결과, 변환 및 확인을 포함하는 자동화된 방법은 소프트웨어, 모듈, 프로세서, 주변장치 및/또는 장치에서 구현된다. 본원에서 사용된 바와 같이, 소프트웨어는 본원에 기재된 바와 같이 프로세서에 의해 실행될 때 컴퓨터 작동을 수행하는 컴퓨터 판독가능한 프로그램 지시를 지칭한다.

[0556] 검사 대상체(예를 들면, 환자, 임신 여성) 및/또는 기준 대상체로부터 유래된 서열 리드, 카운트, 수준 및 프로파일은 유전적 변이의 존재 또는 부재를 확인하기 위해 더 분석되고 프로세싱될 수 있다. 서열 리드, 카운트, 수준 및/또는 프로파일은 종종 "데이터" 또는 "데이터 세트"로서 지칭된다. 일부 실시양태에서, 데이터 또는 데이터 세트는 하나 이상의 특징 또는 변수(예를 들면, 서열에 기초한 특징 또는 변수[예를 들면, GC 함량, 특정 뉴클레오티드 서열 등], 기능 특이적 특징 또는 변수[예를 들면, 발현된 유전자, 암 유전자 등], 위치에 기초한

특징 또는 변수[게놈 특이적, 염색체 특이적, 부분 또는 부분 특이적] 등 및 이들의 조합)를 특징으로 할 수 있다. 일부 실시양태에서, 데이터 또는 데이터 세트는 하나 이상의 특징 또는 변수를 기초로 2개 이상의 차원을 갖는 매트릭스로 조직화될 수 있다. 매트릭스로 조직화된 데이터는 임의의 적합한 특징 또는 변수를 사용함으로써 조직화될 수 있다. 매트릭스 내의 데이터의 비한정적 예는 모체 연령, 모체 배수성 및 태아 기여에 의해 조직화되는 데이터를 포함한다. 일부 실시양태에서, 하나 이상의 특징 또는 변수를 특징으로 하는 데이터 세트는 종종 카운팅 후 프로세싱된다.

[0557] 장치, 소프트웨어 및 인터페이스를 이용하여 본원에 기재된 방법을 수행할 수 있다. 예를 들면, 사용자는 장치, 소프트웨어 및 인터페이스를 사용함으로써 통계학적 분석 알고리즘, 통계학적 유의성 알고리즘, 통계학적 알고리즘, 반복적 단계, 검증 알고리즘 및 도식적 표시를 실행하는 것을 포함할 수 있는 특정 정보, 프로그램 또는 프로세스(예를 들면, 서열 리드의 맵핑, 맵핑된 데이터의 프로세싱 및/또는 결과의 제공)의 사용을 위한 옵션을 도입할 수 있거나 요청할 수 있거나, 질의할 수 있거나 결정할 수 있다. 일부 실시양태에서, 데이터 세트는 입력 정보로서 사용자에 의해 도입될 수 있고, 사용자는 적합한 하드웨어 매체(예를 들면, 플래쉬 드라이브)로 하나 이상의 데이터 세트를 다운로딩할 수 있고/있거나, 사용자는 결과의 후속 프로세싱 및/또는 제공을 위해 데이터 세트를 한 시스템으로부터 또 다른 시스템으로 보낼 수 있다(예를 들면, 서열 리드 데이터를 서열결정기로부터 서열 리드 맵핑용 컴퓨터 시스템으로 보내고; 맵핑된 서열 리드 데이터를 결과 및/또는 보고의 프로세싱 및 생성용 컴퓨터 시스템으로 보낸다).

[0558] 시스템은 전형적으로 하나 이상의 장치를 포함한다. 각각의 장치는 하나 이상의 메모리, 하나 이상의 프로세서 및 지시를 포함한다. 시스템이 2개 이상의 장치를 포함하는 경우, 장치들 중 일부 또는 전부가 동일한 위치에 위치할 수 있고/있거나, 장치들 중 일부 또는 전부가 상이한 위치에 위치할 수 있고/있거나, 장치들 전부가 한 위치에 위치할 수 있고/있거나, 장치들 전부가 상이한 위치에 위치할 수 있다. 시스템이 2개 이상의 장치를 포함하는 경우, 장치들 중 일부 또는 전부가 사용자와 동일한 위치에 위치할 수 있고/있거나, 장치들 중 일부 또는 전부가 사용자와 상이한 위치에 위치할 수 있고/있거나, 장치들 전부가 사용자와 동일한 위치에 위치할 수 있고/있거나, 장치들 전부가 사용자와 상이한 하나 이상의 위치에 위치할 수 있다.

[0559] 시스템은 종종 연산 장치 및 서열결정 장치를 포함하고, 이때 서열결정 장치는 물리적 핵산을 수용하고 서열 리드를 생성하도록 구성되고, 연산 장치는 서열결정 장치로부터의 리드를 프로세싱하도록 구성된다. 연산 장치는 종종 서열 리드로부터 유전적 변이(예를 들면, 카피 수 변이; 태아 염색체 이배수성)의 존재 또는 부재를 확인하도록 구성된다.

[0560] 예를 들면, 사용자는 인터넷 접근을 통해 데이터 세트를 획득할 수 있는 소프트웨어에게 질의를 할 수 있고, 일부 실시양태에서 프로그래밍가능한 프로세서가 주어진 파라미터를 기초로 적합한 데이터 세트를 획득하도록 촉구될 수 있다. 프로그래밍가능한 프로세서는 사용자가 주어진 파라미터를 기초로 프로세서에 의해 선택된 하나 이상의 데이터 세트 옵션을 선택하도록 촉구할 수도 있다. 프로그래밍가능한 프로세서는 사용자가 인터넷을 통해 발견된 정보, 다른 내부 또는 외부 정보 등을 기초로 프로세서에 의해 선택된 하나 이상의 데이터 세트 옵션을 선택하도록 촉구할 수 있다. 옵션은 하나 이상의 데이터 특징 선택, 하나 이상의 통계학적 알고리즘, 하나 이상의 통계학적 분석 알고리즘, 하나 이상의 통계학적 유의성 알고리즘, 반복적 단계, 하나 이상의 검증 알고리즘, 및 방법, 장치 또는 컴퓨터 프로그램의 하나 이상의 도식적 표시를 선택하기 위해 선택될 수 있다.

[0561] 본원에서 다루어진 시스템은 컴퓨터 시스템의 일반적인 구성요소, 예컨대, 네트워크 서버, 휴대용(laptop) 시스템, 데스크탑 시스템, 초소형(handheld) 시스템, 개인 디지털 보조장치, 연산 키오스크(kiosk) 등을 포함할 수 있다. 컴퓨터 시스템은 하나 이상의 입력 수단, 예컨대, 키보드, 터치 스크린, 마우스, 음성 인식, 또는 사용자가 데이터를 시스템 내로 도입할 수 있게 하는 다른 수단을 포함할 수 있다. 시스템은 디스플레이 스크린(예를 들면, CRT 또는 LCD), 스피커, FAC 기기, 프린터(예를 들면, 레이저, 잉크 젯, 임팩트, 흑백 또는 칼라 프린터), 또는 정보(예를 들면, 결과 및/또는 보고)의 시각적, 청각적 및/또는 하드카피 출력물을 제공하는 데에 유용한 다른 출력 수단을 포함하나 이들로 한정되지 않는 하나 이상의 출력 수단을 추가로 포함할 수 있다.

[0562] 시스템에서, 입력 수단 및 출력 수단은 다른 구성요소들 중에서 프로그램 지시를 실행하기 위한 마이크로프로세서, 및 프로그램 코드 및 데이터를 저장하기 위한 메모리를 포함할 수 있는 중심 프로세싱 유닛에 연결될 수 있다. 일부 실시양태에서, 프로세스는 단일 지리학적 장소에 위치하는 단일 사용자 시스템으로서 실행될 수 있다. 일부 실시양태에서, 프로세스는 다중사용자 시스템으로서 실행될 수 있다. 다중사용자 실행의 경우, 다수의 중심 프로세싱 유닛들이 네트워크에 의해 연결될 수 있다. 네트워크는 건축물의 한 부분 내의 단일 부서 또는 전체 건축물을 포괄하거나, 다수의 건축물들에 걸쳐 있거나, 한 지역에 걸쳐 있거나 전체 국가에 걸쳐 있는 지역

네트워크 또는 전세계 네트워크일 수 있다. 네트워크는 공급자에 의해 소유되고 제어되는 사유 네트워크일 수 있거나, 사용자가 웹 페이지에 접근하여 정보를 도입하고 검색하는 인터넷 기초 서비스로서 실행될 수 있다. 따라서, 일부 실시양태에서, 시스템은 사용자의 관점에서 근거리 또는 원거리에 위치할 수 있는 하나 이상의 기계를 포함한다. 한 위치 또는 다수의 위치들에서 하나 초과의 기계가 사용자에 의해 접근될 수 있고, 데이터가 연속적으로 및/또는 동시에 맵핑될 수 있고/있거나 프로세싱될 수 있다. 따라서, 예컨대, 적합한 배열 및 제어가 예컨대, 지역 네트워크, 원격 네트워크 및/또는 "클라우드(cloud)" 연산 플랫폼에서 다수의 기계들을 이용하여 데이터를 맵핑하고/하거나 프로세싱하는 데에 이용될 수 있다.

[0563] 일부 실시양태에서, 시스템은 통신 인터페이스를 포함할 수 있다. 통신 인터페이스는 컴퓨터 시스템과 하나 이상의 외부 디바이스 사이의 소프트웨어 및 데이터의 전달을 가능하게 한다. 통신 인터페이스의 비한정적 예는 모뎀, 네트워크 인터페이스(예컨대, 에터넷(Ethernet) 카드), 통신 포트, PCMCIA 슬롯 및 카드 등을 포함한다. 통신 인터페이스를 통해 전달된 소프트웨어 및 데이터는 일반적으로 통신 인터페이스에 의해 제공받을 수 있는 전자, 전자기, 광학 및/또는 다른 신호일 수 있는 신호의 형태로 존재한다. 신호는 종종 채널을 통해 통신 인터페이스에게 제공된다. 채널은 종종 신호를 보유하고 와이어 또는 케이블, 광섬유, 전화선, 휴대전화 링크, RF 링크 및/또는 다른 통신 채널을 이용함으로써 실행될 수 있다. 따라서, 예를 들면, 통신 인터페이스는 신호 검출 모듈에 의해 검출될 수 있는 신호 정보를 제공받는 데에 사용될 수 있다.

[0564] 데이터는 수동 입력 디바이스 또는 직접 데이터 도입 디바이스(DDE)를 포함하나 이들로 한정되지 않는 적합한 디바이스 및/또는 방법에 의해 입력될 수 있다. 수동 디바이스의 비한정적 예는 키보드, 컨셉(concept) 키보드, 터치 민감성 스크린, 라이트 펜, 마우스, 트랙커 볼(tracker ball), 조이스틱(joystick), 그래픽 태블릿(graphic tablet), 스캐너, 디지털 카메라, 비디오 디지털화기 및 음성 인식 디바이스를 포함한다. DDE의 비한정적 예는 바 코드 판독기, 자기 스트립 코드, 스마트 카드, 자기 잉크 캐릭터 인식, 광학 캐릭터 인식, 광학 마크 인식 및 반환 문서를 포함한다.

[0565] 일부 실시양태에서, 서열결정 장치로부터의 출력물은 입력 디바이스를 통해 입력될 수 있는 데이터로서 사용될 수 있다. 일부 실시양태에서, 맵핑된 서열 리드는 입력 디바이스를 통해 입력될 수 있는 데이터로서 사용될 수 있다. 일부 실시양태에서, 시뮬레이팅된 데이터는 인 실리코(in silico) 프로세스에 의해 생성되고, 시뮬레이팅된 데이터는 입력 디바이스를 통해 입력될 수 있는 데이터로서 사용된다. 용어 "인 실리코"는 컴퓨터를 이용함으로써 수행된 연구 및 실험을 지칭한다. 인 실리코 프로세스는 본원에 기재된 프로세스에 따라 서열 리드를 맵핑하고 맵핑된 서열 리드를 프로세싱하는 것을 포함하나 이들로 한정되지 않는다.

[0566] 시스템은 본원에 기재된 프로세스를 수행하는 데에 유용한 소프트웨어를 포함할 수 있고, 소프트웨어는 이러한 프로세스를 수행하기 위한 하나 이상의 모듈(예를 들면, 서열결정 모듈, 논리 프로세싱 모듈, 데이터 디스플레이 조직화 모듈)을 포함할 수 있다. 용어 "소프트웨어"는 컴퓨터에 의해 실행될 때 컴퓨터 작동을 수행하는 컴퓨터 판독가능한 프로그램 지시를 지칭한다. 하나 이상의 프로세서에 의해 실행될 수 있는 지시는 종종 실행될 때 하나 이상의 프로세서가 본원에 기재된 방법을 실행하게 할 수 있는 실행가능한 코드로서 제공된다. 본원에 기재된 모듈은 소프트웨어로서 존재할 수 있고, 소프트웨어에서 구현된 지시(예를 들면, 프로세스, 루틴, 서브루틴)는 프로세서에 의해 실행될 수 있거나 수행될 수 있다. 예를 들면, 모듈(예를 들면, 소프트웨어 모듈)은 특정 프로세스 또는 과제를 수행하는 프로그램의 일부일 수 있다. 용어 "모듈"은 보다 더 큰 장치 또는 소프트웨어 시스템에서 사용될 수 있는 자가-함유된 기능성 유닛을 지칭한다. 모듈은 모듈의 기능을 수행하기 위한 지시 세트를 포함할 수 있다. 모듈은 데이터 및/또는 정보를 변환시킬 수 있다. 데이터 및/또는 정보는 적합한 형태로 존재할 수 있다. 예를 들면, 데이터 및/또는 정보는 디지털 또는 아날로그일 수 있다. 일부 실시양태에서, 데이터 및/또는 정보는 팩킷(packet), 바이트(byte), 캐릭터(character) 또는 비트(bit)일 수 있다. 일부 실시양태에서, 데이터 및/또는 정보는 임의의 모아진, 조립된 또는 사용가능한 데이터 또는 정보일 수 있다. 데이터 및/또는 정보의 비한정적 예는 적합한 매체, 사진, 비디오, 음향(예를 들면, 들리거나 들리지 않는 주파수), 숫자, 상수, 값, 물체, 시간, 함수, 지시, 지도, 참고자료, 서열, 리드, 맵핑된 리드, 수준, 범위, 역치, 신호, 디스플레이, 표시치 또는 이의 변환물을 포함한다. 모듈은 데이터 및/또는 정보를 수용할 수 있거나 제공받을 수 있고, 데이터 및/또는 정보를 제2 형태로 변환시킬 수 있고, 상기 제2 형태를 장치, 주변장치, 구성요소 또는 또 다른 모듈에게 제공할 수 있거나 전달할 수 있다. 예를 들면, 모듈은 하기 비한정적 기능들 중 하나 이상을 수행할 수 있다: 서열 리드의 맵핑, 카운트의 제공, 부분의 조립, 수준의 제공 또는 측정, 카운트 프로파일의 제공, 정규화(예를 들면, 리드의 정규화, 카운트의 정규화 등), 정규화된 카운트 프로파일 또는 정규화된 카운트의 수준의 제공, 2개 이상의 수준의 비교, 불확실성 값의 제공, 기대 수준 및 기대 범위(예를 들면, 기대 수준 범위, 역치 범위 및 역치 값)의 제공 또는 측정, 수준의 조절(예를 들면, 제1 수준의 조절, 제2 수준의 조

절, 염색체 또는 이의 분절의 프로파일의 조절 및/또는 패딩)의 제공, 확인(예를 들면, 카피 수 변이, 유전적 변이 또는 이배수성의 확인)의 제공, 분류, 작도 및/또는 결과의 확인. 일부 실시양태에서, 프로세서는 모듈에서 지시를 수행할 수 있다. 일부 실시양태에서, 하나 이상의 프로세서가 모듈 또는 모듈 군에서 지시를 수행하는 데에 요구된다. 모듈은 데이터 및/또는 정보를 또 다른 모듈, 장치 또는 공급원에게 제공할 수 있고 또 다른 모듈, 장치 또는 공급원으로부터 데이터 및/또는 정보를 제공받을 수 있다.

[0567] 컴퓨터 프로그램 제품은 종종 유형의 컴퓨터 판독가능한 매체 상에서 구현되고, 종종 비-일시적 컴퓨터 판독가능한 매체 상에서 유형적으로 구현된다. 모듈은 종종 컴퓨터 판독가능한 매체(예를 들면, 디스크, 드라이브) 상에 또는 메모리(예를 들면, 무작위 접근 메모리) 내에 저장된다. 모듈로부터의 지시를 실행할 수 있는 모듈 및 프로세서는 한 장치 또는 상이한 장치에 위치할 수 있다. 모듈을 위한 지시를 실행할 수 있는 모듈 및/또는 프로세서는 사용자와 동일한 위치(예를 들면, 지역 네트워크) 또는 사용자와 상이한 위치(예를 들면, 원격 네트워크, 클라우드 시스템)에 위치할 수 있다. 방법이 2개 이상의 모듈들과 함께 수행되는 실시양태에서, 모듈들은 동일한 장치에 위치할 수 있고, 하나 이상의 모듈은 동일한 물리적 위치에서 상이한 장치에 위치할 수 있고, 하나 이상의 모듈은 상이한 물리적 위치에서 상이한 장치에 위치할 수 있다.

[0568] 일부 실시양태에서, 장치는 모듈에서 지시를 수행하기 위한 하나 이상의 프로세서를 포함한다. 기준 게놈의 부분들에 맵핑된 서열 리드의 카운트는 종종 본원에 기재된 방법을 수행하도록 구성된 지시를 실행하는 프로세서에 의해 접근된다. 프로세서에 의해 접근되는 카운트는 시스템의 메모리 내에 존재할 수 있고, 상기 카운트는 수득된 후 접근될 수 있고 시스템의 메모리 내에 놓일 수 있다. 일부 실시양태에서, 장치는 모듈로부터의 하나 이상의 지시(예를 들면, 프로세스, 루틴 및/또는 서브루틴)를 수행할 수 있고/있거나 실행할 수 있는 프로세서(예를 들면, 하나 이상의 프로세서)를 포함한다. 일부 실시양태에서, 장치는 다수의 프로세서들, 예컨대, 동시에 작동하는 통합된 프로세서들을 포함한다. 일부 실시양태에서, 장치는 하나 이상의 외부 프로세서(예를 들면, 내부 또는 외부 네트워크, 서버, 저장 디바이스 및/또는 저장 네트워크(예를 들면, 클라우드))와 함께 작동한다. 일부 실시양태에서, 장치는 모듈을 포함한다. 일부 실시양태에서, 장치는 하나 이상의 모듈을 포함한다. 모듈을 포함하는 장치는 종종 하나 이상의 데이터 및/또는 정보를 다른 모듈로부터 제공받을 수 있고 다른 모듈에게 전달할 수 있다. 일부 실시양태에서, 장치는 주변장치 및/또는 구성요소를 포함한다. 일부 실시양태에서, 장치는 데이터 및/또는 정보를 다른 모듈, 주변장치 및/또는 구성요소에게 전달할 수 있고 다른 모듈, 주변장치 및/또는 구성요소로부터 전달받을 수 있는 하나 이상의 주변장치 또는 구성요소를 포함할 수 있다. 일부 실시양태에서, 장치는 데이터 및/또는 정보를 제공하는 주변장치 및/또는 구성요소와 상호작용한다. 일부 실시양태에서, 주변장치 및 구성요소는 기능을 수행하는 데에 있어서 장치를 보조하거나 모듈과 직접적으로 상호작용한다. 주변장치 및/또는 구성요소의 비한정적 예는 스캐너, 프린터, 디스플레이(예를 들면, 모니터, LED, LCT 또는 CRT), 카메라, 마이크로폰, 패드(예를 들면, 아이패드, 태블릿), 터치 스크린, 스마트 전화, 이동 전화, USB I/O 디바이스, USB 대량 저장 디바이스, 키보드, 컴퓨터 마우스, 디지털 펜, 모뎀, 하드 드라이브, 점프 드라이브, 플래쉬 드라이브, 프로세서, 서버, CD, DVD, 그래픽 카드, 전문화된 I/O 디바이스(예를 들면, 서열결정기, 포토 셀, 광전자증배관, 광학 판독기, 센서 등), 하나 이상의 유동 셀, 유체 취급 구성요소, 네트워크 인터페이스 제어기, ROM, RAM, 무선 전달 방법 및 디바이스(블루투쓰(Bluetooth), 와이파이(WiFi) 등), 월드 와이드 웹(www), 인터넷, 컴퓨터 및/또 다른 모듈을 포함하나 이들로 한정되지 않는 적합한 컴퓨터 주변장치, I/O 또는 저장 방법 또는 디바이스를 포함한다.

[0569] 종종, 소프트웨어는 플로피 디스크, 하드 디스크 및 자기 테이프를 포함하는 자기 매체; 및 CD-ROM 디스크, DVD 디스크, 자기광학 디스크, 플래쉬 드라이브, RAM, 플로피 디스크 등을 포함하는 광학 매체, 및 프로그램 지시가 기록될 수 있는 다른 이러한 매체를 포함하나 이들로 한정되지 않는 컴퓨터 판독가능한 매체 상에 기록된 프로그램 지시를 함유하는 프로그램 제품 상에 제공된다. 온라인 실행에서, 단체에 의해 유지되는 서버 및 웹 사이트가 소프트웨어 다운로드를 원거리 사용자에게 제공하도록 구성될 수 있거나, 원거리 사용자가 단체에 의해 유지된 원격 시스템에 접근하여 소프트웨어를 원격으로 접근할 수 있다. 소프트웨어는 입력 정보를 수득할 수 있거나 수용할 수 있다. 소프트웨어는 구체적으로 데이터를 수득하거나 수용하는 모듈(예를 들면, 서열 리드 데이터 및/또는 맵핑된 리드 데이터를 수용하는 데이터 수용 모듈)을 포함할 수 있고 구체적으로 데이터를 프로세싱하는 모듈(예를 들면, 수용된 데이터를 프로세싱하는(예를 들면, 필터링하고/하거나, 정규화하고/하거나, 결과를 제공하고/하거나 보고하는) 프로세싱 모듈)을 포함할 수 있다. 용어 입력 정보를 "수득하는" 및 "수용하는"은 근거리 또는 원거리 장소로부터의 컴퓨터 통신 수단, 인간 데이터 도입, 또는 임의의 다른 데이터 수용 방법으로 데이터(예를 들면, 서열 리드, 맵핑된 리드)를 수용하는 것을 지칭한다. 입력 정보는 이것이 수용되는 위치와 동일한 위치에서 생성될 수 있거나, 상이한 위치에서 생성되어 수용 위치로 전달될 수 있다. 일부 실시양태에서, 입력 정보는 프로세싱되기(예를 들면, 프로세싱될 수 있는 포맷으로 놓이기(예를 들면, 표로 작성되

기)) 전에 변형된다.

[0570] 일부 실시양태에서, 컴퓨터 프로그램 제품, 예를 들면, 그 내부에 구현된 컴퓨터 판독가능한 프로그램 코드를 갖는 컴퓨터 사용가능한 매체를 포함하는 컴퓨터 프로그램 제품이 제공되고, 이때 상기 컴퓨터 판독가능한 프로그램 코드는 실행되어 하기 단계들을 포함하는 방법을 수행하도록 개조되어 있다: (a) 검사 대상체로부터 샘플 핵산의 서열 리드를 수득하는 단계; (b) (a)에서 수득된 서열 리드를 부분으로 나누어진 공지된 게놈으로 맵핑하는 단계; (c) 상기 부분 내의 맵핑된 서열 리드를 카운팅하는 단계; (d) (c)에서 수득된 부분에 대한 카운트를 정규화함으로써 샘플 정규화된 카운트 프로파일을 생성하는 단계; 및 (e) (d)에서 수득된 샘플 정규화된 카운트 프로파일로부터 유전적 변이의 존재 또는 부재를 확인하는 단계.

[0571] 일부 실시양태에서, 소프트웨어는 하나 이상의 알고리즘을 포함할 수 있다. 알고리즘은 유한 수열의 지시에 따라 데이터를 프로세싱하고/하거나 결과 또는 보고를 제공하는 데에 사용될 수 있다. 알고리즘은 종종 과제를 완료하기 위한 정해진 지시의 목록이다. 지시는 초기 상태부터 출발하여 정해진 일련의 연속적인 상태를 통해 진행하고 궁극적으로 최종 종결 상태로 종결되는 연산을 기술할 수 있다. 한 상태로부터 다음 상태로의 전이는 반드시 확정적인 것은 아니다(예를 들면, 일부 알고리즘들은 무작위성을 도입한다). 예를 들면(그러나, 한정되지 않음), 알고리즘은 검색 알고리즘, 분류 알고리즘, 병합 알고리즘, 수치 알고리즘, 그래프 알고리즘, 스트링 알고리즘, 모델링 알고리즘, 연산 게노메트릭(genometric) 알고리즘, 조합 알고리즘, 기계 학습 알고리즘, 암호작성 알고리즘, 데이터 압축 알고리즘, 파싱(parsing) 알고리즘 등일 수 있다. 알고리즘은 1개의 알고리즘, 또는 함께 작동하는 2개 이상의 알고리즘을 포함할 수 있다. 알고리즘은 임의의 적합한 복잡성 클래스 및/또는 파라미터화된 복잡성의 알고리즘일 수 있다. 알고리즘은 계산 및/또는 데이터 프로세싱을 위해 사용될 수 있고, 일부 실시양태에서 확정적 또는 확률적/예측 방법에서 사용될 수 있다. 알고리즘은 적합한 프로그래밍 언어의 사용에 의해 연산 환경에서 실행될 수 있고, 이러한 언어의 비한정적 예는 C, C++, 자바(Java), 펄(Perl), 파이톤(Python), 포트란(Fortran) 등이다. 일부 실시양태에서, 알고리즘은 오차의 한계, 통계학적 분석, 통계학적 유의성, 및/또는 다른 정보 또는 데이터 세트와의 비교(예를 들면, 신경 네트 또는 클러스터링 알고리즘을 사용할 때 적용가능함)를 포함하도록 구성될 수 있거나 변형될 수 있다.

[0572] 일부 실시양태에서, 여러 알고리즘들이 소프트웨어에서 사용되도록 실행될 수 있다. 일부 실시양태에서, 이 알고리즘들은 미처리 데이터로 트레이닝될 수 있다. 각각의 새로운 미처리 데이터 샘플에 대해, 트레이닝된 알고리즘은 대표적인 프로세싱된 데이터 세트 또는 결과를 생성할 수 있다. 프로세싱된 데이터 세트는 종종 프로세싱된 모 데이터 세트에 비해 감소된 복잡성을 갖는다. 일부 실시양태에서, 프로세싱된 세트를 기초로, 트레이닝된 알고리즘의 성능은 민감성 및 특이성을 기초로 평가될 수 있다. 일부 실시양태에서, 가장 높은 민감성 및/또는 특이성을 갖는 알고리즘이 확인될 수 있고 사용될 수 있다.

[0573] 일부 실시양태에서, 시뮬레이팅된(또는 시뮬레이션) 데이터는 예를 들면, 알고리즘을 트레이닝하거나 알고리즘을 시험함으로써 데이터 프로세싱을 도울 수 있다. 일부 실시양태에서, 시뮬레이팅된 데이터는 서열 리드의 상이한 군들의 다양한 가상 샘플링을 포함한다. 시뮬레이팅된 데이터는 실제 집단으로부터 예상된 데이터에 기초할 수 있거나, 알고리즘을 시험하고/하거나 정확한 분류를 배정하도록 왜곡될 수 있다. 시뮬레이팅된 데이터는 본원에서 "가상" 데이터로서도 지칭된다. 일부 실시양태에서, 시뮬레이션은 컴퓨터 프로그램에 의해 수행될 수 있다. 시뮬레이팅된 데이터 세트의 사용에서 한 가능한 단계는 확인된 결과의 신뢰도, 예를 들면, 무작위 샘플링이 원래의 데이터와 얼마나 잘 일치하는지 또는 원래의 데이터를 얼마나 가장 잘 표시하는지를 평가하는 것이다. 한 방법은 무작위 샘플이 선택된 샘플보다 더 우수한 점수를 가질 확률을 추정하는 확률 값(p-값)을 계산하는 것이다. 일부 실시양태에서, 실험 모델이 평가될 수 있고, 이때 하나 이상의 샘플이 (해상 변경을 이용하거나 이용하지 않을 때) 기준 샘플과 일치한다고 가정된다. 일부 실시양태에서, 또 다른 분포, 예컨대, 포아송(Poisson) 분포가 확률 분포를 정의하는 데에 사용될 수 있다.

[0574] 일부 실시양태에서, 시스템은 하나 이상의 프로세서를 포함할 수 있다. 프로세서는 통신 버스(bus)에 연결될 수 있다. 컴퓨터 시스템은 주 메모리, 종종 무작위 접근 메모리(RAM)를 포함할 수 있고, 이차 메모리도 포함할 수 있다. 일부 실시양태에서, 메모리는 비-일시적 컴퓨터 판독가능한 저장 매체를 포함한다. 이차 메모리는 예를 들면, 하드 디스크 드라이브 및/또는 제거가능한 저장 드라이브(플로피 디스크 드라이브, 자기 테이프 드라이브, 광학 디스크 드라이브, 메모리 카드 등을 나타냄)를 포함할 수 있다. 제거가능한 저장 드라이브는 종종 제거가능한 저장 유닛으로부터 읽고/읽거나 제거가능한 저장 유닛에 쓴다. 제거가능한 저장 유닛의 비한정적 예는 예를 들면, 제거가능한 저장 드라이브에 의해 읽혀질 수 있고 쓰여질 수 있는 플로피 디스크, 자기 테이프, 광학 디스크 등을 포함한다. 제거가능한 저장 유닛은 그 내부에 저장된 컴퓨터 소프트웨어 및/또는 데

이터를 갖는 컴퓨터 사용가능한 저장 매체를 포함할 수 있다.

[0575] 프로세서는 시스템에서 소프트웨어를 실행할 수 있다. 일부 실시양태에서, 프로세서는 사용자가 수행할 수 있는 본원에 기재된 과제를 자동적으로 수행하도록 프로그래밍될 수 있다. 따라서, 프로세서, 또는 이러한 프로세서에 의해 수행된 알고리즘은 사용자로부터의 감시 또는 입력을 거의 또는 전혀 요구하지 않을 수 있다(예를 들면, 소프트웨어는 자동적으로 기능을 실행하도록 프로그래밍될 수 있다). 일부 실시양태에서, 프로세스의 복잡성은 1명의 사람 또는 일군의 사람들이 유전적 변이의 존재 또는 부재를 확인하기 위해 충분히 짧은 시간 내에 프로세스를 수행할 수 없을 정도로 크다.

[0576] 일부 실시양태에서, 이차 메모리는 컴퓨터 프로그램 또는 다른 지시가 컴퓨터 시스템 내로 적재될 수 있게 하는 다른 유사한 수단을 포함할 수 있다. 예를 들면, 시스템은 제거가능한 저장 유닛 및 인터페이스 디바이스를 포함할 수 있다. 이러한 시스템의 비한정적 예는 프로그램 카트리지 및 카트리지 인터페이스(예컨대, 비디오 게임 디바이스에서 발견된 것), 제거가능한 메모리 칩(예컨대, EPROM 또는 PROM) 및 관련 소켓, 및 소프트웨어 및 데이터가 제거가능한 저장 유닛으로부터 컴퓨터 시스템으로 전달될 수 있게 하는 다른 제거가능한 저장 유닛 및 인터페이스를 포함한다.

[0577] 일부 실시양태에서, 한 엔터티(entity)가 본원에 기재된 방법, 시스템, 장치 또는 컴퓨터 프로그램 제품에서 서열 리드의 카운트를 생성할 수 있고, 서열 리드를 부분들로 맵핑할 수 있고, 맵핑된 리드를 카운팅할 수 있고, 카운팅된 맵핑된 리드를 사용할 수 있다. 일부 실시양태에서, 부분들에 맵핑된 서열 리드의 카운트는 종종 본원에 기재된 방법, 시스템, 장치 또는 컴퓨터 프로그램 제품에서 제2 엔터티에 의한 사용을 위해 한 엔터티에 의해 제2 엔터티에게 전달된다.

[0578] 일부 실시양태에서, 한 엔터티는 서열 리드를 생성하고 제2 엔터티는 서열 리드를 기준 게놈 내의 부분들로 맵핑한다. 제2 엔터티는 종종 맵핑된 리드를 카운팅하고 카운팅된 맵핑된 리드를 본원에 기재된 방법, 시스템, 장치 또는 컴퓨터 프로그램 제품에서 사용한다. 일부 실시양태에서, 제2 엔터티는 맵핑된 리드를 제3 엔터티에게 전달하고, 제3 엔터티는 맵핑된 리드를 카운팅하고 맵핑된 리드를 본원에 기재된 방법, 시스템, 장치 또는 컴퓨터 프로그램 제품에서 사용한다. 일부 실시양태에서, 제2 엔터티는 맵핑된 리드를 카운팅하고 카운팅된 맵핑된 리드를 제3 엔터티에게 전달하고, 제3 엔터티는 카운팅된 맵핑된 리드를 본원에 기재된 방법, 시스템, 장치 또는 컴퓨터 프로그램에서 사용한다. 제3 엔터티를 포함하는 실시양태에서, 제3 엔터티는 종종 제1 엔터티와 동일하다. 즉, 제1 엔터티는 종종 서열 리드를 제2 엔터티에게 전달하고, 상기 제2 엔터티는 서열 리드를 기준 게놈 내의 부분들로 맵핑할 수 있고/있거나 맵핑된 리드를 카운팅할 수 있고, 제2 엔터티는 맵핑되고/되거나 카운팅된 리드를 제3 엔터티에게 전달할 수 있다. 제3 엔터티는 맵핑되고/되거나 카운팅된 리드를 종종 본원에 기재된 방법, 시스템, 장치 또는 컴퓨터 프로그램 제품에서 사용할 수 있고, 이때 제3 엔터티는 종종 제1 엔터티와 동일하고, 종종 제3 엔터티는 제1 또는 제2 엔터티와 상이하다.

[0579] 일부 실시양태에서, 한 엔터티는 임신 여성으로부터 혈액을 수득하고 임의적으로 상기 혈액(예를 들면, 혈장 또는 혈청)으로부터 핵산을 단리하고 상기 혈액 또는 핵산을, 이 핵산으로부터 서열 리드를 생성하는 제2 엔터티에게 전달한다.

[0580] 도 30은 본원에 기재된 다양한 시스템, 방법, 알고리즘 및 데이터 구조물이 실행될 수 있는 연산 환경(510)의 비한정적 예를 보여준다. 연산 환경(510)은 적합한 연산 환경의 일례일 뿐이고, 본원에 기재된 시스템, 방법 및 데이터 구조물의 사용 또는 기능성의 범위에 대한 임의의 한정을 암시하기 위한 것이 아니다. 연산 환경(510)은 연산 환경(510)에 예시된 구성요소들 중 어느 하나 또는 조합에 관한 임의의 의존 또는 요구를 갖는 것으로서 해석되어서도 안 된다. 일부 실시양태에서, 도 30에 나타낸 시스템, 방법 및 데이터 구조물의 서브세트가 사용될 수 있다. 본원에 기재된 시스템, 방법 및 데이터 구조물은 다수의 다른 일반 목적 또는 특수 목적의 연산 시스템 환경 또는 구성과 함께 작동할 수 있다. 적합할 수 있는 공지된 연산 시스템, 환경 및/또는 구성의 예는 개인용 컴퓨터, 서버 컴퓨터, 씬 클라이언트(thin clients), 씩 클라이언트(thick client), 초소형 또는 휴대용 디바이스, 멀티프로세서 시스템, 마이크로프로세서 기초 시스템, 셋 탑 박스, 프로그래밍가능한 가전제품, 네트워크 PC, 미니컴퓨터, 메인프레임 컴퓨터, 상기 시스템들 또는 디바이스들 중 임의의 시스템 또는 디바이스를 포함하는 분산된 연산 환경 등을 포함할 수 있으나 이들로 한정되지 않는다.

[0581] 도 30의 연산 환경(510)은 프로세싱 유닛(521), 시스템 메모리(522), 및 시스템 메모리(522)를 포함하는 다양한 시스템 구성요소들을 프로세싱 유닛(521)에 작동가능하게 커플링시키는 시스템 버스(523)를 포함하는, 컴퓨터(520) 형태의 일반 목적 연산 디바이스를 포함한다. 컴퓨터(520)의 프로세서가 단일 중앙-프로세싱 유닛(CPU), 또는 통상적으로 병렬 프로세싱 환경으로서 지칭되는 복수의 프로세싱 유닛들을 포함하도록 단지 1개의 프로세

싱 유닛(521)이 존재할 수 있거나 하나 초과의 프로세싱 유닛(521)이 존재할 수 있다. 컴퓨터(520)는 보편적인 컴퓨터, 분산 컴퓨터, 또는 임의의 다른 종류의 컴퓨터일 수 있다.

[0582] 시스템 버스(523)는 메모리 버스 또는 메모리 제어기, 주변 버스, 및 다양한 버스 구조물들 중 임의의 버스 구조물을 사용하는 지역 버스를 포함하는 여러 종류의 버스 구조물들 중 임의의 버스 구조물일 수 있다. 시스템 메모리는 단순히 메모리로서 지칭될 수도 있고 읽기 전용 메모리(ROM) 및 무작위 접근 메모리(RAM)를 포함한다. 예컨대, 스타트-업(start-up) 동안 컴퓨터(520) 내의 요소들 사이에 정보를 전달하는 데에 도움을 주는 기본 루틴을 함유하는 기본 입력/출력 시스템(BIOS)(526)이 ROM(524)에 저장된다. 컴퓨터(520)는 하드 디스크(도시되어 있지 않음)로부터 읽고 쓰기 위한 하드 디스크 드라이브 인터페이스(527), 제거가능한 자기 디스크(529)로부터 읽거나 쓰기 위한 자기 디스크 드라이브(528), 및 제거가능한 광학 디스크(531), 예컨대, CD ROM 또는 다른 광학 매체로부터 읽거나 쓰기 위한 광학 디스크 드라이브(530)를 추가로 포함할 수 있다.

[0583] 하드 디스크 드라이브(527), 자기 디스크 드라이브(528) 및 광학 디스크 드라이브(530)는 각각 하드 디스크 드라이브 인터페이스(532), 자기 디스크 드라이브 인터페이스(533) 및 광학 디스크 드라이브 인터페이스(534)에 의해 시스템 버스(523)에 연결된다. 상기 드라이브들 및 이들의 관련된 컴퓨터 판독가능한 매체는 컴퓨터 판독가능한 지시, 데이터 구조물, 프로그램 모듈 및 다른 데이터의 지구성 저장(nonvolatile storage)을 컴퓨터(520)에게 제공한다. 컴퓨터에 의해 접근될 수 있는 데이터를 저장할 수 있는 임의의 종류의 컴퓨터 판독가능한 매체, 예컨대, 자기 카세트, 플래쉬 메모리 카드, 디지털 비디오 디스크, 베르누이(Bernoulli) 카트리지, 무작위 접근 메모리(RAM), 읽기 전용 메모리(ROM) 등이 운영 환경에서 사용될 수 있다.

[0584] 운영 시스템(535), 하나 이상의 응용 프로그램(536), 다른 프로그램 모듈(537) 및 프로그램 데이터(538)를 포함하는 다수의 프로그램 모듈들이 하드 디스크, 자기 디스크(529), 광학 디스크(531), ROM(524) 또는 RAM 상에 저장될 수 있다. 사용자는 입력 디바이스, 예컨대, 키보드(540) 및 포인팅 디바이스(542)를 통해 명령 및 지시를 개인용 컴퓨터(520)에 도입할 수 있다. 다른 입력 디바이스(도시되어 있지 않음)는 마이크로폰, 조이스틱, 게임 패드, 위성방송 수신 안테나, 스캐너 등을 포함할 수 있다. 이 입력 디바이스들 및 다른 입력 디바이스들은 종종 시스템 버스에 커플링되는 직렬 포트 인터페이스(546)를 통해 프로세싱 유닛(521)에 연결되지만, 다른 인터페이스, 예컨대, 병렬 포트, 게임 포트 또는 보편적인 직렬 버스(USB)에 의해 연결될 수 있다. 모니터(547) 또는 다른 종류의 디스플레이 디바이스도 인터페이스, 예컨대, 비디오 어댑터(548)를 통해 시스템 버스(523)에 연결된다. 컴퓨터는 모니터 이외에 전형적으로 다른 주변 출력 디바이스(도시되어 있지 않음), 예컨대, 스피커 및 프린터를 포함한다.

[0585] 컴퓨터(520)는 하나 이상의 원격 컴퓨터, 예컨대, 원격 컴퓨터(549)에의 논리적 연결을 이용하여 네트워킹 환경에서 작동할 수 있다. 이 논리적 연결은 컴퓨터(520) 또는 이 컴퓨터의 일부에 커플링된 통신 디바이스에 의해 또는 다른 방식으로 달성될 수 있다. 원격 컴퓨터(549)는 또 다른 컴퓨터, 서버, 루터(router), 네트워크 PC, 클라이언트, 피어(peer) 디바이스 또는 다른 통상의 네트워크 노드일 수 있고, 메모리 저장 디바이스(550)만이도 30에 예시되어 있지만 전형적으로 컴퓨터(520)와 관련하여 전술된 요소들 중 대다수 또는 전부를 포함한다. 도 30에 도시된 논리적 연결은 지역 네트워크(LAN)(551) 및 광역 네트워크(WAN)(552)를 포함한다. 이러한 네트워킹 환경은 사무실 네트워크, 전사적 컴퓨터 네트워크, 인트라넷 및 인터넷(모드 네트워크의 종류임)에서 매우 흔하다.

[0586] LAN-네트워킹 환경에서 사용될 때, 컴퓨터(520)는 통신 디바이스의 일종인 네트워크 인터페이스 또는 어댑터(553)를 통해 지역 네트워크(551)에 연결된다. WAN-네트워킹 환경에서 사용될 때, 컴퓨터(520)는 종종 통신 디바이스의 일종인 모뎀(554), 또는 광역 네트워크(552) 상에서의 통신을 확립하기 위한 임의의 다른 종류의 통신 디바이스를 포함한다. 내부 또는 외부에 존재할 수 있는 모뎀(554)은 직렬 포트 인터페이스(546)를 통해 시스템 버스(523)에 연결된다. 네트워킹 환경에서, 개인용 컴퓨터(520)와 관련하여 도시된 프로그램 모듈 또는 이의 부분은 원격 메모리 저장 디바이스에 저장될 수 있다. 도시된 네트워크 연결이 비한정적 예이고 컴퓨터들 사이의 통신 연결을 확립하기 위한 다른 통신 디바이스가 사용될 수 있다는 것이 인식된다.

[0587] 모듈

[0588] 하나 이상의 모듈이 본원에 기재된 방법에서 사용될 수 있고, 이의 비한정적 예는 논리 프로세싱 모듈, 서열결정 모듈, 맵핑 모듈, 카운팅 모듈, 필터링 모듈, 가중 모듈, 정규화 모듈, GC 편향 모듈, 수준 모듈, 비교 모듈, 범위 설정 모듈, 분류 모듈, 작도 모듈, 표시 모듈, 관계 모듈, 결과 모듈 및/또는 데이터 디스플레이 조직화 모듈 등 또는 이들의 조합을 포함한다. 모듈은 종종 마이크로프로세서에 의해 제어된다. 일부 실시양태에서, 모듈, 또는 하나 이상의 모듈을 포함하는 장치는 데이터 및/또는 정보를 또 다른 모듈, 장치, 구성요소, 주

변장치 또는 장치의 작동자에게 또는 이로부터 모으고/으거나, 조립하고/하거나, 수용하고/하거나, 수득하고/하거나, 접근하고/하거나, 회수하고/하거나, 제공하고/하거나 전달한다. 일부 실시양태에서, 데이터 및/또는 정보(예를 들면, 서열결정 리드)는 하나 이상의 유동 셀, 카메라, 검출기(예를 들면, 광 검출기, 광 셀, 전기 검출기(예를 들면, 진폭 조절 검출기, 주파수 및 위상 조절 검출기, 위상-잠겨진 루프 검출기), 카운터, 센서(예를 들면, 압력, 온도, 부피, 유동 또는 중량의 센서), 유체 취급 디바이스, 프린터, 디스플레이(예를 들면, LED, LCT 또는 CRT) 등 및 이들의 조합 중 하나 이상을 포함하는 장치에 의해 모듈에게 제공된다. 예를 들면, 종종 장치의 작동자는 상수, 역치 값, 식 또는 예정된 값을 모듈에게 제공한다. 모듈은 종종 데이터 및/또는 정보를 또 다른 모듈 또는 장치에게 전달하거나 또 다른 모듈 또는 장치로부터 전달받도록 구성된다. 모듈은 또 다른 모듈로부터 데이터 및/또는 정보를 수용할 수 있고, 이러한 또 다른 모듈의 비한정적 예는 논리 프로세싱 모듈, 서열결정 모듈, 맵핑 모듈, 카운팅 모듈, 필터링 모듈, 가중 모듈, 정규화 모듈, GC 편향 모듈, 수준 모듈, 비교 모듈, 범위 설정 모듈, 분류 모듈, 작도 모듈, 표시 모듈, 관계 모듈, 결과 모듈 및/또는 데이터 디스플레이 조직화 모듈 등 또는 이들의 조합을 포함한다. 모듈은 데이터 및/또는 정보를 조작할 수 있고/있거나 변환시킬 수 있다. 모듈로부터 유도되거나 모듈에 의해 변환된 데이터 및/또는 정보는 또 다른 적합한 장치 및/또는 모듈에게 전달될 수 있고, 이러한 장치 및/또는 모듈의 비한정적 예는 논리 프로세싱 모듈, 서열결정 모듈, 맵핑 모듈, 카운팅 모듈, 필터링 모듈, 가중 모듈, 정규화 모듈, GC 편향 모듈, 수준 모듈, 비교 모듈, 범위 설정 모듈, 분류 모듈, 작도 모듈, 표시 모듈, 관계 모듈, 결과 모듈 및/또는 데이터 디스플레이 조직화 모듈 등 또는 이들의 조합을 포함한다. 모듈을 포함하는 장치는 하나 이상의 프로세서를 포함할 수 있다. 일부 실시양태에서, 데이터 및/또는 정보는 모듈을 포함하는 장치에 의해 수용되고/되거나 제공된다. 모듈을 포함하는 장치는 모듈의 하나 이상의 지시(예를 들면, 프로세스, 루틴 및/또는 서브루틴)를 수행할 수 있고/있거나 실행할 수 있는 프로세서(예를 들면, 하나 이상의 프로세서)를 포함할 수 있다. 일부 실시양태에서, 모듈은 하나 이상의 외부 프로세서(예를 들면, 내부 또는 외부 네트워크, 서버, 저장 장치 및/또는 저장 네트워크(예를 들면, 클라우드))로 작동한다.

[0589] 논리 프로세싱 모듈

[0590] 일부 실시양태에서, 논리 프로세싱 모듈은 데이터 및/또는 정보를 편성하고/하거나, 제어하고/하거나, 한정하고/하거나, 조직화하고/하거나, 정돈하고/하거나, 분산시키고/시키거나, 분할하고/하거나, 변환시키고/시키거나 조절하거나, 데이터 및/또는 정보를 하나 이상의 다른 모듈, 주변장치 또는 디바이스에게 전달하고 하나 이상의 다른 모듈, 주변장치 또는 디바이스로부터 전달받는다.

[0591] 데이터 디스플레이 조직화 모듈

[0592] 일부 실시양태에서, 데이터 디스플레이 조직화 모듈은 데이터 및/또는 정보를 적합한 시각적 매체로 프로세싱하고/하거나 변환시키고, 이러한 시각적 매체의 비한정적 예는 영상, 비디오 및/또는 문자(예를 들면, 숫자, 글자 및 부호)를 포함한다. 일부 실시양태에서, 데이터 디스플레이 조직화 모듈은 적합한 디스플레이(예를 들면, 모니터, LED, LCD, CRT 등 또는 이들의 조합), 프린터, 적합한 주변장치 또는 디바이스 상에서의 제시를 위해 데이터 및/또는 정보를 프로세싱하고/하거나, 변환시키고/시키거나 전달한다. 일부 실시양태에서, 데이터 디스플레이 조직화 모듈은 데이터 및/또는 정보를 태아 또는 모체 게놈, 염색체 또는 이의 부분의 시각적 표시치로 프로세싱하거나 변환시킨다.

[0593] 서열결정 모듈

[0594] 일부 실시양태에서, 서열결정 모듈은 서열 리드를 수득하고/하거나, 생성하고/하거나, 모으고/으거나, 조립하고/하거나, 조작하고/하거나, 변환시키고/시키거나, 프로세싱하고/하거나, 변환시키고/시키거나 전달한다. 본원에서 사용된 "서열 수용 모듈"은 "서열결정 모듈"과 동일하다. 서열결정 모듈을 포함하는 장치는 당분야에서 공지된 서열결정 기술을 이용하여 핵산의 서열을 결정하는 임의의 장치일 수 있다. 일부 실시양태에서, 서열결정 모듈은 서열 리드를 정렬할 수 있거나, 조립할 수 있거나, 단편화할 수 있거나, 보완할 수 있거나, 역보완할 수 있거나, 오류 점검할 수 있거나 오류 보정할 수 있다.

[0595] 맵핑 모듈

[0596] 서열 리드는 맵핑 모듈, 또는 맵핑 모듈을 포함하는 장치에 의해 맵핑될 수 있고, 이 맵핑 모듈은 일반적으로 리드를 기준 게놈 또는 이의 분절로 맵핑한다. 맵핑 모듈은 당분야에서 공지된 적합한 방법으로 서열결정 리드를 맵핑할 수 있다. 일부 실시양태에서, 맵핑 모듈, 또는 맵핑 모듈을 포함하는 장치는 맵핑된 서열 리드를 제공하는 데에 요구된다.

[0597] 카운팅 모듈

[0598] 카운트는 카운팅 모듈, 또는 카운팅 모듈을 포함하는 장치에 의해 제공될 수 있다. 일부 실시양태에서, 카운팅 모듈은 기준 게놈으로 맵핑된 서열 리드를 카운팅한다. 일부 실시양태에서, 카운팅 모듈은 당분야에서 공지된 카운팅 방법에 따라 카운트를 생성하고/하거나, 조립하고/하거나 제공한다. 일부 실시양태에서, 카운팅 모듈, 또는 카운팅 모듈을 포함하는 장치는 카운트를 제공하는 데에 요구된다.

[0599] 필터링 모듈

[0600] 부분(예를 들면, 기준 게놈의 부분)의 필터링은 필터링 모듈(예를 들면, 필터링 모듈을 포함하는 장치)에 의해 제공될 수 있다. 일부 실시양태에서, 필터링 모듈은 필터링된 부분 데이터(예를 들면, 필터링된 부분)를 제공하고/하거나 부분을 고려사항으로부터 제거하는 데에 요구된다. 일부 실시양태에서, 필터링 모듈은 부분들에 맵핑된 카운트를 고려사항으로부터 제거한다. 일부 실시양태에서, 필터링 모듈은 부분들에 맵핑된 카운트를 수준 또는 프로파일의 측정으로부터 제거한다. 필터링 모듈은 당분야에서 공지되어 있거나 본원에 기재되어 있는 하나 이상의 필터링 방법으로 데이터(예를 들면, 카운트, 부분들에 맵핑된 카운트, 부분, 부분 수준, 정규화된 카운트, 미처리 카운트 등)를 필터링할 수 있다.

[0601] 가중 모듈

[0602] 부분(예를 들면, 기준 게놈의 부분)의 가중은 가중 모듈(예를 들면, 가중 모듈을 포함하는 장치)에 의해 제공될 수 있다. 일부 실시양태에서, 가중 모듈은 게놈 구획을 가중하고/하거나 가중된 부분 값을 제공하는 데에 요구된다. 가중 모듈은 당분야에서 공지되어 있거나 본원에 기재되어 있는 하나 이상의 가중 방법으로 부분을 가중할 수 있다.

[0603] 정규화 모듈

[0604] 정규화된 데이터(예를 들면, 정규화된 카운트)는 정규화 모듈(예를 들면, 정규화 모듈을 포함하는 장치)에 의해 제공될 수 있다. 일부 실시양태에서, 정규화 모듈은 서열결정 리드로부터 수득된 정규화된 데이터(예를 들면, 정규화된 카운트)를 제공하는 데에 요구된다. 정규화 모듈은 본원에 기재되어 있거나(예를 들면, PERUN, ChAI, 하이브리드 정규화 등 또는 이들의 조합) 또는 당분야에서 공지되어 있는 하나 이상의 정규화 방법으로 데이터(예를 들면, 카운트, 필터링된 카운트, 미처리 카운트)를 정규화할 수 있다.

[0605] GC 편향 모듈

[0606] GC 편향의 측정(예를 들면, 기준 게놈의 각각의 부분(예를 들면, 부분, 기준 게놈의 부분)에 대한 GC 편향의 측정)은 GC 편향 모듈(예를 들면, GC 편향 모듈을 포함하는 장치)에 의해 제공될 수 있다. 일부 실시양태에서, GC 편향 모듈은 GC 편향의 측정을 제공하는 데에 요구된다. 일부 실시양태에서, GC 편향 모듈은 기준 게놈의 각각의 부분들에 맵핑된 서열 리드의 카운트와 각각의 부분의 GC 함량 사이의 피팅된 관계(예를 들면, 피팅된 선형 관계)로부터 GC 편향의 측정을 제공한다. GC 편향 모듈은 종종 정규화 모듈(예를 들면, PERUN, ChAI 정규화 모듈)의 일부이다.

[0607] 수준 모듈

[0608] 수준(예를 들면, 수준들)의 측정 및/또는 기준 게놈의 부분에 대한 게놈 구획 수준의 계산은 수준 모듈(예를 들면, 수준 모듈을 포함하는 장치)에 의해 제공될 수 있다. 일부 실시양태에서, 수준 모듈은 (예를 들면, 방정식 A, B, L, M, N, O 및/또는 Q에 따라) 수준 또는 계산된 게놈 구획 수준을 제공하는 데에 요구된다. 일부 실시양태에서, 수준 모듈은 기준 게놈의 각각의 부분들에 맵핑된 서열 리드의 GC 편향과 카운트 사이의 피팅된 관계(예를 들면, 피팅된 선형 관계)로부터 수준을 제공한다. 일부 실시양태에서, 수준 모듈은 PERUN의 일부로서 게놈 구획 수준을 계산한다. 일부 실시양태에서, 수준 모듈은 방정식 $L_i = (m_i - G_iS)I^{-1}$에 따라 게놈 구획 수준(즉, L_i)을 제공하고, 이때 G_i는 GC 편향이고, m_i는 기준 게놈의 각각의 부분들에 맵핑된 측정된 카운트이고, i는 샘플이고, I는 절편이고, S는 기준 게놈의 각각의 부분들에 맵핑된 서열 리드의 GC 편향과 카운트 사이의 피팅된 관계(예를 들면, 피팅된 선형 관계)의 기울기이다.

[0609] 비교 모듈

[0610] 제1 수준은 비교 모듈, 또는 비교 모듈을 포함하는 장치에 의해 제2 수준과 유의하게 상이한 것으로서 확인될 수 있다. 일부 실시양태에서, 비교 모듈, 또는 비교 모듈을 포함하는 장치는 2개의 수준들 사이의 비교를 제공

하는 데에 요구된다.

[0611] 범위 설정 모듈

[0612] 다양한 카피 수 변이(예를 들면, 중복, 삽입 및/또는 결실)에 대한 기대 범위(예를 들면, 기대 수준 범위) 또는 카피 수 변이의 부재에 대한 범위는 범위 설정 모듈, 또는 범위 설정 모듈을 포함하는 장치에 의해 제공될 수 있다. 일부 실시양태에서, 기대 수준은 범위 설정 모듈, 또는 범위 설정 모듈을 포함하는 장치에 의해 제공된다. 일부 실시양태에서, 범위 설정 모듈, 또는 범위 설정 모듈을 포함하는 장치는 기대 수준 및/또는 범위를 제공하는 데에 요구된다.

[0613] 분류 모듈

[0614] 카피 수 변이(예를 들면, 모체 및/또는 태아 카피 수 변이, 태아 카피 수 변이, 중복, 삽입, 결실)는 분류 모듈, 또는 분류 모듈을 포함하는 장치에 의해 분류될 수 있다. 일부 실시양태에서, 카피 수 변이(예를 들면, 모체 및/또는 태아 카피 수 변이)는 분류 모듈에 의해 분류된다. 일부 실시양태에서, 또 다른 수준(예를 들면, 제2 수준)과 유의하게 상이한 것으로서 확인된 수준(예를 들면, 제1 수준)은 분류 모듈에 의해 카피 수 변이의 표시치로서 확인된다. 일부 실시양태에서, 카피 수 변이의 부재는 분류 모듈에 의해 확인된다. 일부 실시양태에서, 카피 수 변이의 확인은 분류 모듈을 포함하는 장치에 의해 확인될 수 있다. 분류 모듈은 모체 및/또는 태아 카피 수 변이, 태아 카피 수 변이, 중복, 결실 또는 삽입, 또는 이들의 결여 또는 조합을 분류하도록 전문화될 수 있다. 예를 들면, 모체 결실을 확인하는 분류 모듈은 태아 중복을 확인하는 분류 모듈과 상이할 수 있고/있거나 구별될 수 있다. 일부 실시양태에서, 분류 모듈, 또는 분류 모듈을 포함하는 장치는 카피 수 변이, 또는 카피 수 변이를 확인시켜주는 결과를 확인하는 데에 요구된다.

[0615] 작도 모듈

[0616] 일부 실시양태에서, 작도 모듈은 데이터 및/또는 정보를 적합한 시각적 매체로 프로세싱하고/하거나 변환시키고, 이러한 시각적 매체의 비한정적 예는 차트, 도표, 그래프 등 또는 이들의 조합을 포함한다. 일부 실시양태에서, 작도 모듈은 적합한 디스플레이(예를 들면, 모니터, LED, LCD, CRT 등 또는 이들의 조합), 프린터, 적합한 주변장치 또는 디바이스 상에서의 제시를 위해 데이터 및/또는 정보를 프로세싱하고/하거나, 변환시키고/시키거나 전달한다. 일부 실시양태에서, 작도 모듈은 카운트, 수준 및/또는 프로파일의 시각적 디스플레이를 제공한다. 일부 실시양태에서, 데이터 디스플레이 조직화 모듈은 데이터 및/또는 정보를 태아 또는 모체 게놈, 염색체 또는 이들의 일부의 시각적 표시치로 프로세싱하거나 변환시킨다. 일부 실시양태에서, 작도 모듈, 또는 작도 모듈을 포함하는 장치는 카운트, 수준 또는 프로파일을 작도하는 데에 요구된다.

[0617] 관계 모듈

[0618] 일부 실시양태에서, 관계 모듈은 데이터 및/또는 정보를 관계로 프로세싱하고/하거나 변환시킨다. 일부 실시양태에서, 관계는 관계 모듈에 의해 생성되고/되거나 관계 모듈로부터 전달된다.

[0619] 결과 모듈

[0620] 일부 실시양태에서, 유전적 변이(이배수성, 태아 이배수성, 카피 수 변이)의 존재 또는 부재는 결과 모듈, 또는 결과 모듈을 포함하는 장치에 의해 확인된다. 일부 실시양태에서, 유전적 변이는 결과 모듈에 의해 확인된다. 종종 이배수성의 존재 또는 부재는 결과 모듈에 의해 확인된다. 일부 실시양태에서, 유전적 변이(이배수성, 카피 수 변이)를 확인시켜주는 결과는 결과 모듈, 또는 결과 모듈을 포함하는 장치에 의해 확인될 수 있다. 결과 모듈은 특정 유전적 변이(예를 들면, 삼염색체성, 삼염색체성 21, 삼염색체성 18)를 확인하도록 전문화될 수 있다. 예를 들면, 삼염색체성 21을 확인하는 결과 모듈은 삼염색체성 18을 확인하는 결과 모듈과 상이할 수 있고/있거나 구별될 수 있다. 일부 실시양태에서, 결과 모듈, 또는 결과 모듈을 포함하는 장치는 유전적 변이를 확인하거나 유전적 변이(예를 들면, 이배수성, 카피 수 변이)를 확인시켜주는 결과를 확인하는 데에 요구된다. 본원에 기재된 방법에 의해 확인된 유전적 변이, 또는 유전적 변이를 확인시켜주는 결과는 추가 검사(예를 들면, 모체 및/또는 태아 핵산의 표적화된 서열결정)에 의해 독립적으로 검증될 수 있다.

[0621] 변환

[0622] 상기 인지된 바와 같이, 데이터는 종종 한 형태로부터 또 다른 형태로 변환된다. 본원에서 사용된 용어 "변환된", "변환" 및 이의 문법적 파생어 또는 등가어는 물리적 출발 물질(예를 들면, 검사 대상체 및/또는 기준 대상체 샘플 핵산)로부터의 데이터를 물리적 출발 물질의 디지털 표시치(예를 들면, 서열 리드 데이터)로 변경시키는 것을 지칭하고, 일부 실시양태에서 결과를 제공하는 데에 사용될 수 있는 디지털 표시치의 하나 이상

의 수치 값 또는 도식적 표시치로의 추가 변환을 포함한다. 일부 실시양태에서, 디지털적으로 표시된 데이터의 하나 이상의 수치 값 및/또는 도식적 표시치는 검사 대상체의 물리적 게놈의 외관을 표시하는 데(예를 들면, 게놈 삽입, 중복 또는 결실의 존재 또는 부재를 사실적으로 표시하거나 시각적으로 표시하거나; 의학적 질환과 관련된 서열의 물리적 양의 변경의 존재 또는 부재를 표시하는 데)에 사용될 수 있다. 사실적 표시치는 종종 출발 물질의 디지털 표시치의 하나 이상의 수치 값 또는 도식적 표시치로 더 변환된다. 이 방법들은 물리적 출발 물질을 수치 값 또는 도식적 표시치, 또는 검사 대상체의 게놈의 물리적 외관의 표시치로 변환시킬 수 있다.

[0623] 일부 실시양태에서, 데이터 세트의 변환은 데이터 복잡성 및/또는 데이터 차원수를 감소시킴으로써 결과의 제공을 용이하게 한다. 데이터 세트 복잡성은 종종 물리적 출발 물질을 이 출발 물질의 사실적 표시치(예를 들면, 물리적 출발 물질을 표시하는 서열 리드)로 변환시키는 프로세스 동안 감소된다. 적합한 특징 또는 변수가 데이터 세트 복잡성 및/또는 차원수를 감소시키는 데에 사용될 수 있다. 데이터 프로세싱을 위한 표적 특징으로서 사용되기 위해 선택될 수 있는 특징의 비한정적 예는 GC 함량, 태아 성별 예측, 염색체 이배수성의 확인, 특정 유전자 또는 단백질의 확인, 암의 확인, 질환, 유전된 유전자/형질, 염색체 이상, 생물학적 카테고리, 화학적 카테고리, 생화학적 카테고리, 유전자 또는 단백질의 카테고리, 유전자 존재론(ontology), 단백질 존재론, 함께 조절되는 유전자, 세포 신호전달 유전자, 세포 주기 유전자, 상기 유전자들과 관련된 단백질, 유전자 변이체, 단백질 변이체, 함께 조절되는 유전자, 함께 조절되는 단백질, 아미노산 서열, 뉴클레오티드 서열, 단백질 구조 데이터 등 및 이들의 조합을 포함한다. 데이터 세트 복잡성 및/또는 차원수 감소의 비한정적 예는 복수의 서열 리드를 프로파일 도표로 단순화하는 것, 복수의 서열 리드를 수치 값(예를 들면, 정규화된 값, Z-점수, p-값)으로 단순화하는 것, 다수의 분석 방법들을 확률 도표 또는 단일 점으로 단순화하는 것, 유도된 양의 주성분 분석 등 또는 이들의 조합을 포함한다.

[0624] 실시예

[0625] 하기 실시예들은 단지 예시로서 제공되고 한정으로서 제공되지 않는다. 따라서, 하기 실시예들은 일부 실시양태를 예시하고 기술을 한정하지 않는다. 당업자는 본질적으로 동일한 또는 유사한 결과를 생성하도록 변경될 수 있거나 변형될 수 있는 다양한 비-임계적 파라미터들을 용이하게 인식할 것이다.

[0626] 실시예 1: 유전적 변이와 관련된 상태를 검출하는 PERUN 및 일반적인 방법.

[0627] 본원에 기재된 방법 및 기저 이론은 유전적 변이와 관련된 다양한 상태를 검출하고 유전적 변이의 존재 또는 부재를 확인시켜주는 결과를 제공하거나 유전적 변이의 존재 또는 부재를 확인하는 데에 이용될 수 있다.

[0628] 기준 게놈의 비-정보제공 부분의 제거

[0629] 기준 게놈의 비-정보제공 부분을 제거하기 위한 다수의 시도들은 부분 선택이 분류를 개선하는 잠재력을 갖는다는 것을 시사하였다.

[0630] 방정식 A:

[0631] $M = LI + GS$

[0632] 방정식 A에서 다양한 용어들은 하기 의미를 갖는다:

[0633] - M: 원하지 않는 변이에 의해 오염된 일차 정보를 표시하는 측정된 카운트.

[0634] - L: 염색체 수준 - 이것은 데이터 프로세싱 절차로부터의 원하는 출력물이다. L은 정배수체로부터의 태아 및/또는 모체 이상을 표시한다. 이것은 확률적 오차 및 시스템적 편향 둘 다에 의해 감춰지는 양이다. 염색체 수준 L은 샘플 특이적 및 부분 특이적이다.

[0635] - G: 선형 모델, LOESS 또는 임의의 등가 방법을 이용함으로써 측정된 GC 편향 계수. G는 통상적으로 기준 게놈으로부터 유도된, M 및 부분 특이적 GC 함량 값의 세트로부터 추출된 이차 정보를 표시한다(그러나, 실제 관찰된 GC 함량으로부터도 유도될 수 있다). G는 샘플 특이적이고 게놈 위치에 따라 변경되지 않는다. 이것은 원하지 않는 변이의 부분을 요약한다.

[0636] - I: 선형 모델의 절편. 이 모델 파라미터는 주어진 실험 셋업에 대해 고정되고 샘플에 의존하지 않고 부분 특이적이다.

[0637] - S: 선형 모델의 기울기. 이 모델 파라미터는 주어진 실험 셋업에 대해 고정되고 샘플에 의존하지 않고 부분

특이적이다.

[0638] 양 *M* 및 *G*가 측정된다. 처음에, 부분 특이적 값 *I* 및 *S*는 공지되어 있지 않다. 공지되어 있지 않은 *I* 및 *S*를 평가하기 위해, 본 발명자들은 정배수체 샘플에서 기준 게놈의 모든 부분들에 대해 *L*이 1이라고 가정해야 한다. 이 가정은 항상 옳지는 않지만, 정상 염색체 수준을 갖는 샘플이 결실/중복을 갖는 임의의 샘플을 압도할 것이라고 합리적으로 예측할 수 있다. 정배수체 샘플에 적용된 선형 모델은 선택된 부분에 대한 특이성을 갖는 *I* 및 *S* 파라미터 값을 추출한다(*L*이 1이라고 가정함). 동일한 절차를 인간 게놈 내의 기준 게놈의 모든 부분들에 적용하여 모든 게놈 위치에 대한 절편 *I* 및 기울기 *S*의 세트를 생성한다. 교차검증은 모든 LDTv2CE 정배수체의 90%를 함유하는 작업 세트를 무작위적으로 선택하고 그 서브세트를 사용하여 모델을 트레이닝한다. 무작위 선택을 100회 반복하여 모든 부분들에 대한 100개 기울기 및 100개 절편의 세트를 생성한다.

[0639] 측정된 카운트로부터의 염색체 수준의 추출

[0640] 모델 파라미터 값 *I* 및 *S*가 모든 부분에 대해 이용될 수 있다고 가정하면서, 새로운 검사 샘플에 대해 수집된 측정치 *M*을 사용하여 하기 방정식 B에 따라 염색체 수준을 평가한다.

[0641] $L = (M - GS)/I$

[0642] 방정식 A에서처럼, GC 편향 계수 *G*는 부분별로 측정된 미처리 카운트 *M*과 기준 게놈의 GC 함량 사이의 회귀의 기울기로서 평가된다. 그 다음, 염색체 수준 *L*은 추가 분석(Z-값, 모체 결실/중복, 태아 미세결실/미세중복, 태아 성별, 성 이배수성 등)을 위해 사용된다. 방정식 B에 의해 요약된 절차는 파라미터화된 오차 제거 및 비-편향된 정규화(PERUN)로서 명명된다.

[0643] 실시예 2: 식의 예

[0644] 본원에 기재된 방법에서 이용될 수 있는 수학적 및/또는 통계학적 식의 비한정적 예가 하기 제공된다.

[0645] 평균 수준에서의 불확실성에 대한 추정치를 고려하여 1의 기대 수준으로부터의 편차와 관련된 Z-점수, 및 Z-점수로부터 계산된 p-값을 평가할 수 있다. p-값은 t-분포에 기초하고, 이 분포의 차수는 피크에서 기준 게놈의 부분의 수에 의해 결정된다. 원하는 신뢰 수준에 따라 컷오프는 노이즈를 억제할 수 있고 실제 신호의 분명한 검출을 가능하게 할 수 있다.

[0646] 방정식 1:

[0647] $$Z = \frac{\Delta_1 - \Delta_2}{\sqrt{\sigma_1^2\left(\frac{1}{N_1}+\frac{1}{n_1}\right)+\sigma_2^2\left(\frac{1}{N_2}+\frac{1}{n_2}\right)}}$$

[0648] 방정식 1은 2개의 상이한 샘플들로부터의 피크 수준을 직접적으로 비교하는 데에 이용될 수 있고, 이때 N 및 n은 각각 전체 염색체 및 비정상적인 염색체 내의 기준 게놈의 부분의 수를 지칭한다. 2개의 샘플들 사이의 유사성을 측정하는 p-값을 제공할 t-검정의 차수는 2개의 비정상적인 스트레치들 중 보다 더 짧은 스트레치 내의 기준 게놈의 부분의 수에 의해 결정된다.

[0649] 방정식 8은 태아 이배수성에 대한 유전적 변이의 존재 또는 부재를 확인하기 위해 태아 분획, 모체 배수성 및 중간 기준 카운트를 분류 체계 내로 도입하는 데에 이용될 수 있다.

[0650] 방정식 8:

[0651] $$y_i = (1 - F)M_i f_i + FXf_i$$

[0652] 상기 식에서,

[0653] Y_i는 중간 카운트 프로파일 내의 부분에 상응하는 검사 샘플의 부분에 대한 측정된 카운트를 표시하고, *F*는 태아 분획을 표시하고, *X*는 태아 배수성을 표시하고, M_i는 각각의 부분으로 배정된 모체 배수성을 표시한다. 방정식 8에서 *X*를 위해 사용된 가능한 값은 다음과 같다: 태아가 정배수체인 경우 1; 태아가 삼배수체인 경우 3/2; 및 쌍둥이 태아가 존재하고 1명이 영향을 받고 1명이 영향을 받지 않는 경우 5/4. 방정식 8에서 용어 *F*가 총 태아 DNA를 표시하므로, 모든 태아 DNA가 고려되어야 하기 때문에, 5/4는 1명의 태아가 영향을 받고 나머지 1명의 태아가 영향을 받지 않는 쌍둥이의 경우 사용된다. 일부 실시양태에서, 모체 게놈 내의 큰 결실 및/또는 중복은 모체 배수성 M_i를 각각의 부분 또는 부분으로 배정함으로써 설명될 수 있다. 일부 실시양태에서, 모체 배수성은

종종 1/2의 배수로서 배정되고, 부분별 정규화를 이용함으로써 평가될 수 있다. 모체 배수성이 종종 1/2의 배수이기 때문에, 모체 배수성은 용이하게 설명될 수 있으므로, 유도를 단순화하기 위한 추가 방정식에 포함되지 않을 것이다.

[0654] X = 1(예를 들면, 정배수체 가정)에서 방정식 8을 평가할 때, 태아 분획이 삭제되고, 하기 방정식이 제곱 잔차의 합계를 제공한다.

[0655] 방정식 9:

[0656] $$\varphi_E = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}(y_i - f_i)^2 = \sum_{i=1}^{N} \frac{y_i^2}{\sigma_i^2} - 2\sum_{i=1}^{N} \frac{y_i f_i}{\sigma_i^2} + \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2} = \Xi_{yy} - 2\Xi_{fy} + \Xi_{ff}$$

[0657] 방정식 9 및 후속 계산을 단순화하기 위해, 하기 방정식들을 이용한다.

[0658] 방정식 10:

[0659] $$\Xi_{yy} = \sum_{i=1}^{N} \frac{y_i^2}{\sigma_i^2}$$

[0660] 방정식 11:

[0661] $$\Xi_{ff} = \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2}$$

[0662] 방정식 12:

[0663] $$\Xi_{fy} = \sum_{i=1}^{N} \frac{y_i f_i}{\sigma_i^2}$$

[0664] X = 3/2(예를 들면, 삼배수체 가정)에서 방정식 8을 평가할 때, 하기 방정식이 제곱 잔차의 합계를 제공한다.

[0665] 방정식 13:

[0666] $$\varphi_T = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left(y_i - f_i - \frac{1}{2}Ff_i\right)^2 = \Xi_{yy} - 2\Xi_{fy} + \Xi_{ff} + F(\Xi_{ff} - \Xi_{fy}) + \frac{1}{4}F^2\Xi_{ff}$$

[0667] 방정식 9와 방정식 13 사이의 차이는 대안적 가설(예를 들면, 삼염색체성 외둥이, X = 3/2)에 대한 널 가설(예를 들면, 정배수체, X = 1)을 검정하는 데에 사용될 수 있는 함수 결과(예를 들면, phi)를 형성한다.

[0668] 방정식 14:

[0669] $$\varphi = \varphi_E - \varphi_T = F(\Xi_{fy} - \Xi_{ff}) - \frac{1}{4}F^2\Xi_{ff}$$

[0670] 방정식 18:

[0671] $$\varphi = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}[y_i - (1-F)M_i f_i - FXf_i]^2$$

$$= \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left[y_i^2 - 2(1-F)M_i f_i y_i - 2FXf_i y_i + (1-F)^2 M_i^2 f_i^2 + 2F(1-F)XM_i f_i^2 + F^2X^2 f_i^2\right]$$

[0672] 최적 배수성 값은 종종 방정식 20에 의해 제공된다.

[0673] 방정식 20:

[0674] $$X = \frac{\sum_{i=1}^{N} \frac{f_i y_i}{\sigma_i^2} - (1-F)\sum_{i=1}^{N} \frac{M_i f_i^2}{\sigma_i^2}}{F\sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2}}$$

[0675] 모체 배수성 M_i에 대한 용어는 일부 수학적 유도식으로부터 생략될 수 있다. 그 결과 X에 대한 표현은 상대적으로 단순하고 종종 가장 빈번하게 발생하는 특별한 경우인 모체가 평가되는 염색체 또는 염색체들에서 결실 또는 중복을 갖지 않는 때에 상응한다.

[0676] 방정식 21:

[0677] $$X = \frac{\Xi_{fy} - (1-F)\,\Xi_{ff}}{F\,\Xi_{ff}} = \frac{\Xi_{fy}}{F\,\Xi_{ff}} - \frac{1-F}{F} = 1 + \frac{1}{F}\left(\frac{\Xi_{fy}}{\Xi_{ff}} - 1\right)$$

[0678] Xi_{ff} 및 Xi_{fy}는 각각 방정식 11 및 12에 의해 제공된다. 모든 실험 오차가 무시할만한 실시양태에서, 방정식 21의 풀이는 정배수체에 대한 1의 값을 생성하고, 이때 $Xi_{ff} = Xi_{fy}$이다. 모든 실험 오차가 무시할만한 일부 실시양태에서, 방정식 21의 풀이는 삼배수체에 대한 3/2의 값을 생성한다(Xi_{ff}와 Xi_{fy} 사이의 삼배수체 관계에 대해서는 방정식 15를 참조한다).

[0679] [표 2]

[0680]

임신 상태	태아 21 번 염색체	태아 18 번 염색체	태아 13 번 염색체	태아 X 염색체	태아 Y 염색체
여성 T21	$P_{ij}^{F}=3/2$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=0$
여성 T18	$P_{ij}^{F}=1$	$P_{ij}^{F}=3/2$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=0$
여성 T13	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=3/2$	$P_{ij}^{F}=1$	$P_{ij}^{F}=0$
남성 T21	$P_{ij}^{F}=3/2$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1/2$	$P_{ij}^{F}=1/2$
남성 T18	$P_{ij}^{F}=1$	$P_{ij}^{F}=3/2$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1/2$	$P_{ij}^{F}=1/2$
남성 T13	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=3/2$	$P_{ij}^{F}=1/2$	$P_{ij}^{F}=1/2$
남성 정배수체	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1/2$	$P_{ij}^{F}=1/2$
터너	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1/2$	$P_{ij}^{F}=0$
자콥스	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1/2$	$P_{ij}^{F}=1$
클라인펠터	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1/2$
삼중 X	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=1$	$P_{ij}^{F}=3/2$	$P_{ij}^{F}=0$

[0681] 실시예 3: 결정 트리 분석

[0682] 임의의 염색체 상에서 태아 이배수성(보고되어 있지 않거나 이전에 공지되어 있지 않은 이배수성을 포함함)을 검출할 수 있는 결정 트리 방법을 개발하였다. 추가로, 정규화 절차(예를 들면, PERUN)를 적용한 후 게놈에서의 불균일 커버리지 이벤트를 검출할 수 있다. 도 1(상부 좌측 패널)에 나타낸 불균일 이벤트들은 미세결실/중복을 표시할 수 있다. 이 이벤트들은 독립적으로 검출될 수 있다.

[0683] 방법

[0684] 미리 정해진 게놈 좌표를 갖는 T21, T18 및 T13을 검출하는 경우와 달리, 불균일 커버리지 이벤트는 게놈 내의 임의의 위치에서 일어날 수 있다. 미리 정해진 위치에서의 유전적 변이의 검출은 표시된 위치에서 유의 수준을 측정하는 것만을 요구한다. 본 실시예에 기재된 알고리즘은 일관적으로 상승된 또는 감소된 부분 카운트/커버리지를 갖는 영역을 검색하고 이러한 이벤트의 경계를 정확히 결정한다. 2종의 오르토고날 방법들의 힘을 이용하는 방법이 본 실시예에 기재되어 있다.

[0685] 웨이블렛 분해

[0686] 제1 알고리즘은 웨이블렛 변환을 이용한다. 웨이블렛 변환은 신호 프로세싱 목적에 특히 유용한 수학적 수단이다. 이 변형된 적용에서, 먼저 전체 게놈 서열결정 데이터를 정렬하고 부분으로 나누고 정규화하여 GC 편향을 제거하였다. PERUN 정규화(본원에 기재됨)가 GC 편향을 감소시키는 데에 이용되었고 다른 GC 편향 감소 프로세스(예를 들면, 본원에 기재된 ChAI 프로세스)가 이용될 수 있다. 그 후, 웨이블렛 평활화 방법을 정규화된 프로파일에 적용하여 데이터에서 노이즈를 감소시킴으로써 미세결실/미세중복 이벤트가 명확히 보이게 하였다. 웨이블렛 방법에 대한 도식적 도표는 도 1에 제시되어 있다.

[0687] 웨이블렛 분해를 위해 표준 Haar 웨이블렛을 이용하였다. 원칙적으로, 보다 더 복잡한 웨이블렛 변환이 평활화 후 이벤트의 위치를 확인하는 데에 이용될 수 있다. 노이즈로부터 신호를 식별하기 위해, 어떤 웨이블렛 계수가 신호를 표시하고 보유되어야 하는지, 및 어떤 웨이블렛 계수가 노이즈를 반영할 가능성이 있고 제거되어야 하는지를 확인할 필요가 있다. 이 단계는 역치화로서 지칭된다. 공지된 바와 같이, 큰 크기 및 낮은 수준의 계수는

신호의 경향을 보존하는 반면, 작은 크기 및 높은 수준의 계수는 신호의 세부사항을 보존한다. "소프트" 역치화 방법을 이용하여 작고 무의미한 계수를 제거하였다[Donoho and Johnstone, (1995) WaveLab and Reproducible Research]. 역치화 후, 일부 높은 수준의 계수들은 남아있을 수 있다. 이 계수들은 최초 신호에서 급격한 변화 또는 큰 급등을 나타내고 제거된다. 이 단계는 "평준화"로서 지칭되고, 도 2는 역치화 없는 평준화의 효과를 예시한다(예를 들면, 보다 더 많은 세부사항이 증가된 수준으로 보존된다). 평준화의 최적 선택은 많은 인자들, 예컨대, 염색체의 길이, 검출할 원하는 이벤트 길이 및 정규화된 프로파일의 노이즈 수준에 의존할 수 있다. 염색체 길이 N_{chr}(2의 거듭제곱의 가장 가까운 수까지 연장됨) 및 웨이블렛 분해 수준 c가 주어질 때, 웨이블렛 프로파일의 최소 분절 길이는 $L = N_{chr}/2^{c+1}$이다. 따라서, 크기 N_{micro}의 미세결실을 검출하기 위해, 원하는 분해 수준은 $c = \log_2(N_{chr}/N_{micro})-1$이다. 예를 들면, N_{chr}이 기준 게놈의 4096개 부분이고 미세결실이 기준 게놈의 N_{micro} = 128개 부분의 크기를 갖는 경우, 분해 수준은 $c = 4$이어야 한다. $c \pm 1$의 분해 수준도 이용될 수 있다.

[0688] 서큘러 이진 분할(CBS) 방법

[0689] 웨이블렛 방법은 잠재적인 미세결실/미세중복의 위치를 확인할 수 있다. 그러나, 상기 방법은 단독으로 진짜 이벤트의 존재를 보장하지 않는다. 프로파일이 잘 정규화되지 않은 경우, 웨이블렛 알고리즘은 GC 잔차에 의해 야기된 국소 변동에 의해 오도될 수 있다. 더욱이, 검출된 에지의 정확성은 웨이블렛 계수 절단 순서에 의해 한정된다. 거짓 양성을 감소시키기 위해, 독립적인 방법을 이용하여 웨이블렛 결과를 검증하였다. 서큘러 이진 분할(CBS)은 어레이 CGH 데이터를 사용하여 카피 수 변이(CNV)를 검출하기 위해 최초로 제안된 방법이다. 이 방법은 변화 점을 정확히 찾아낼 수 있다. CBS는 확률 비 통계를 이용하여 염색체를 동등한 카피 수 영역들로 반복적으로 분할함으로써 작동한다[Olshen AB, Venkatraman ES, Lucito R, Wigler M. Biostatistics (2004) Oct;5(4):557-72]. 이것은 일반적으로 잘 작동하지만, 노이즈가 신호에서 높을 때 게놈을 과다분할하는 경향을 갖는다[Lai, WR, Johnson, MD, Kucherlapati, R, Park, PJ Bioinformatics (2005) 21, 19:3763-70]. 본 실시예에서, 상기 방법은 PERUN에 의해 정규화된 부분 카운트 데이터를 사용하여 작동하도록 개조되어 있고 웨이블렛 결과를 검증하기 위한 독립적인 방법으로서 이용된다. 도 5는 CBS 알고리즘을 예시한다.

[0690] 웨이블렛 또는 CBS에 의해 평활화된 프로파일에 대한 분절의 병합

[0691] 웨이블렛 또는 CBS는 관심있는 염색체를 동등한 카피 수의 영역들/분절들로 분할한다. 각각의 분절은 CNV에 대한 잠재적인 후보를 나타낸다. 이전 단락들에서 논의된 바와 같이, CBS는 염색체를 과다분할하는 경향을 가지므로, 광범위한 CNV 영역이 보다 더 작은 여러 조각들로 나누어질 수 있었다. 웨이블렛 방법의 경우 유사한 상황이 발생할 수 있다. 도 4는 이것을 예시하는데, 이때 큰 중복이 처음에 CBS 방법에 의해 3개의 조각으로 분할되어 CNV 폭의 과소평가를 초래하였다(도 4, 하부 좌측 패널). 이 결점을 극복하기 위해, 문헌(Willenbrock H, Fridlyand J., Bioinformatics (2005) Nov 15;21(22):4084-91)에 의해 제안된 알고리즘을 적용하여 웨이블렛 또는 CBS에 의해 평활화된 프로파일들을 동등한 카피 수 영역의 보다 더 긴 스트레치로 더 병합하였다. 2개의 분절들은 이 2개의 분절들로 맵핑된 부분 카운트들이 유의하게 상이하지 않거나 예측된 분절 값이 동적으로 측정된 역치보다 더 가까운 경우 병합된다[Willenbrock and Fridlyand 2005]. 도 4의 하부 우측 패널은 분절 병합의 효과를 예시하는데, 이때 명확한 미세중복이 분절의 병합 후 가시화되었다.

[0692] 웨이블렛 및 CBS 알고리즘으로부터 유도된 통계치

[0693] 각각의 염색체에 대해, 웨이블렛/CBS에 의해 평활화되고 분절 병합된 프로파일로부터 3개의 핵심 z-점수 통계치를 유추할 수 있다.

[0694] (1) 웨이블렛에 의해 평활화된 프로파일 및 이의 부분 카운트 표시치 z-점수(Z_{wave})로부터 가장 우수한 분절(후보 분절). 후보 분절의 샘플 카운트 표시치는 검사 샘플에 대한 총 정규화된 상염색체 카운트로 나누어진 분절에서의 총 정규화된 카운트이다. 후보 분절에 대한 중간 카운트 표시치를 정배수체 샘플 세트에 대해 생성하고, 후보 분절에 대한 정배수체 카운트 표시치에 대한 MAD를 측정한다. 상기 분절에 대한 Z_{wave} 통계치는 검사 샘플 카운트 표시치 - 정배수체 중간 카운트 표시치이고, 뺄셈 결과를 MAD로 나눈다.

[0695] (2) CBS에 의해 평활화된 프로파일 및 이의 부분 카운트 표시치 z-점수(Z_{cbs})로부터 가장 우수한 분절. 후보 분절의 샘플 카운트 표시치는 검사 샘플에 대한 총 정규화된 상염색체 카운트로 나누어진 분절에서의 총 정규화된 카운트이다. 후보 분절에 대한 중간 카운트 표시치를 정배수체 샘플 세트에 대해 생성하고, 후보 분절에 대한 정배수체 카운트 표시치에 대한 MAD를 측정한다. 분절에 대한 Z_{cbs} 통계치는 검사 샘플 카운트 표시치 - 정배수

체 중간 카운트 표시치이고, 뺄셈 결과를 MAD로 나눈다.

[0696] (3) 전체 염색체에 대한 총 염색체 표시치(Z_{chr}). 샘플 카운트 표시치는 검사 샘플에 대한 총 정규화된 상염색체 카운트로 나누어진, 후보 분절이 존재하는 염색체에서의 총 정규화된 카운트이다. 염색체에 대한 중간 카운트 표시치를 정배수체 샘플 세트에 대해 생성하고, 정배수체 카운트 표시치에 대한 MAD를 측정한다. 염색체에 대한 Z_{chr} 통계치는 검사 샘플 카운트 표시치 - 정배수체 중간 카운트 표시치이고, 뺄셈 결과를 MAD로 나눈다.

[0697] 가장 우수한 분절은 종종 그 염색체 상의 모든 분절들 중에서 가장 큰 곡선하면적(AUC)을 갖는 분절이다. 이러한 분절은 관심있는 염색체에 대한 가장 의미 있는 발견을 표시한다. 예를 들면, 도 4의 하부 우측 패널은 2개의 분절들을 갖고, 이들 중 제2 분절은 가장 큰 AUC를 갖는다. 도 5는 각각의 염색체에 대한 웨이블렛 및 CBS 평활화, 및 유도된 통계치를 요약한다.

[0698] CNV 검출을 위한 결정 트리

[0699] 일단 각각의 염색체에 대한 3개의 핵심 통계치가 계산되면, 이들을 사용하여 주어진 샘플에서 삼염색체성, 미세결실 또는 미세중복의 존재를 확인한다. 결정 트리는 하기 제시되어 있다:

[0700] 1. a. $|Z_{chr}| \geq 3.95$이고, b. $|Z_{chr}| \geq \min(\alpha|Z_{wave}|, \alpha|Z_{cbs}|)$인 경우, 염색체는 삼염색체성 또는 일염색체성으로서 분류된다.

[0701] 2. a. 삼염색체성 또는 일염색체성이 아니고, b. $|Z_{wave}| \geq 3.95$이고 $|Z_{cbs}| \geq 3.95$이고, c. 웨이블렛 및 CBS의 가장 우수한 분절이 중첩되는 경우, 염색체는 미세결실/미세중복을 갖는 것으로서 분류된다.

[0702] 조건 1은 본질적으로 총 염색체 Z-점수가 유의할 것을 요구하고 그의 크기는 웨이블렛 또는 CBS의 가장 우수한 분절에 필적할만해야 한다는 것을 주목한다. 조건 2는 웨이블렛 및 CBS의 가장 우수한 분절이 유의할 것을 요구하고, 이들이 서로 중첩되어 결과를 교차확인시켜줄 것을 요구한다(도 5). 일부 경우, CBS 및 웨이블렛 방법에 의해 확인된 웨이블렛 이벤트는 중첩되지 않는데(도 9), 이것은 미세중복 또는 미세결실의 부재를 시사한다. 대다수의 적용에서, 3.95의 Z-점수 컷오프(예를 들면, 예정된 역치)는 원하는 민감성 및 특이성을 달성하기 위해 약간 증가될 수 있거나 감소될 수 있다. 또한, 예정된 값인 α는 종종 본원에서 제시된 대다수의 적용에서 0.6 내지 0.8로 설정된다.

[0703] 결과

[0704] 검출 방법을 22번 염색체 상에서 3 MB 미만의 크기로 확장되어 있는 미세결실 22q11의 2개 사례에 적용하였다(도 10). 3 MB 해상을 달성하기 위해, 샘플을 처음에 0.5-플렉스에서 서열결정하였다. 22q11 미세결실은 커버리지가 10배 감소하였을 때조차도 검출되었다(도 10, F'). 한 샘플의 검출 결과는 도 10에 제시되어 있고, 이때 강조된 영역은 미세결실 이벤트(약 2.5 MB)를 표시한다.

[0705] 알고리즘을 상이한 연구로부터의 샘플에도 적용하였고, 19개의 추정되는 미세결실/미세중복 사례를 검출하였다. 검출된 사례들 중 2개는 도 11a 및 11b 상에 제시되어 있다.

[0706] 실시예 4: 최대 엔트로피

[0707] 최대 엔트로피는 게놈을 균일한 수준의 분절로 분할하는 본원에 기재된 자동화된 알고리즘이다. 이 알고리즘은 인간 게놈의 GC 함량 분포를 특징규명하기 위해 문헌(Cohen N, Dagan T, Stone L, Graur D., (2005) Mol. Biol. Evol., May;22(5):1260-72)의 저자들에 의해 이용된 절차에 기초한다. 미세결실/중복(예를 들면, 미세결실/미세중복 에지)의 위치를 검출할 목적으로 코헨(Cohen)의 방법을 채택하였다. 변경은 분할을 최소 길이까지 제한하는 능력 및 분할 를 종결하는 데에 사용된 t-값 기초 기준을 포함하였다. t-값을 사용하여 염색체(예를 들면, 분절, 구획 또는 영역)에서 확인된 모든 분절들이 균질한지를 확인하였다. 새로 확인된 분절들이 t-값에 기초한 유의성 검정을 통과하지 못한 경우, 이들을 다시 함께 병합하고 분할을 중단하였다. 이 작업을 수행한 코드는 하기 포함되어 있다:

```
# PATH: C:\SEQUENOM_PROJECTS\2013\2013_01_10_PARTITIONING_MAX_ENTROPY\
# FILE: PARTITIONING_MAX_ENTROPY.r
# AUTHOR: Zeljko Jovan Dzakula
# See: Cohen et al. (2005): GC Composition of the Human Genome: In Search
of Isochores. Mole Biol. Evol. 22(5):1260-1272.
# DATE: 1/10/2013

rm( list=ls() );

shannonEntropy <- function( values ) {
    meanValue <- mean( values, na.rm=TRUE );
    entropy <- -meanValue * log2( meanValue ) -
           ( 1 - meanValue ) * log2( 1 - meanValue );
    return( entropy );
} # shannonEntropy

partitionShannonEntropy <- function( values, edge ) {
    nValues <- length( values );
    if ( edge < 1 | edge >= nValues ) {
        return( NA );
    } # if edge
    overallEntropy <- shannonEntropy( values );
    leftEntropy <- shannonEntropy( values[ 1:edge ] );
    rightEntropy <- shannonEntropy( values[ ( edge + 1 ):nValues ] );
```

[0708]

```
    return( overallEntropy - ( leftEntropy * edge + rightEntropy *
( nValues - edge ) ) / nValues );
} # partitionShannonEntropy

findBestEdge <- function( values ) {
    epsilon <- 1e-6;
    bestEdge <- 1;
    bestEntropy <- -1e7;
    values <- values[ !is.na( values ) ];
    minValue <- min( values, na.rm=TRUE ) - epsilon;
    maxValue <- max( values, na.rm=TRUE ) + epsilon;
    values <- ( values - minValue ) / ( maxValue - minValue );
    nValues <- length( values );
    for ( candidateEdge in 1:( nValues - 1 ) ) {
        candidateEntropy <- partitionShannonEntropy( values,
candidateEdge );
        #print( paste( "", candidateEdge, candidateEntropy ) );
        if ( candidateEntropy > bestEntropy ) {
            bestEdge <- candidateEdge;
            bestEntropy <- candidateEntropy;
        } # if candidateEntropy
    } # for candidateEdge
    return( bestEdge );
} # findBestEdge

isPartitionHomogeneous <- function( values, edge, cutoff=10 ) {
    nValues <- length( values );
    if ( edge < 1 | edge >= nValues ) {
        print( paste( "Illegal argument: edge=", edge, sep="" ) );
        return( TRUE );
    } # if edge
    leftMean <- median( values[ 1:edge ], na.rm=TRUE );
    leftMAD <- mad( values[ 1:edge ], na.rm=TRUE );
    rightMean <- median( values[ ( edge + 1 ):nValues ], na.rm=TRUE );
    rightMAD <- mad( values[ ( edge + 1 ):nValues ], na.rm=TRUE );
    tValue <- abs( leftMean - rightMean ) / sqrt(
        leftMAD * leftMAD / edge + rightMAD * rightMAD / ( nValues -
edge ) );
    print( tValue );
    return( tValue < cutoff );
} # isPartitionHomogeneous

partitionRegions <- function( values, regions, minLength, cutoff=10 ) {
    nValues <- length( values );
    rownames( regions ) <- c( "Start", "End", "Homogeneous",
"MedianLevel", "MAD" );
    finished <- FALSE;
    while ( !finished ) {
        nRegions <- ncol( regions );
        for ( region in 1:nRegions ) {
            if ( regions[ "Homogeneous", region ] == FALSE ) {
                startPoint <- regions[ "Start", region ];
```

[0709]

```
                endPoint <- regions[ "End", region ];
                relativeEdge <-
findBestEdge( values[ startPoint:endPoint ] );
                absoluteEdge <- startPoint + relativeEdge - 1;
                homogeneous <- isPartitionHomogeneous(
                    values[ startPoint:endPoint ], relativeEdge,
cutoff=cutoff );
                if ( ( homogeneous == TRUE ) |
                     ( edge == endPoint ) |
                     ( relativeEdge < minLength ) |
                     ( endPoint - absoluteEdge ) < minLength ) {
                    print( paste( "Homogeneous: ", startPoint, "-",
absoluteEdge, "-", endPoint, sep="" ) );
                    regions[ "Homogeneous", region ] <- TRUE;
                } else {
                    print( paste( "Inhomogeneous: ", startPoint, "-",
absoluteEdge, "-", endPoint, sep="" ) );
                    regions[ , region ] <- c( startPoint, absoluteEdge,
FALSE, NA, NA );
                    regions <- cbind( regions,
                        c( absoluteEdge + 1, endPoint, FALSE, NA, NA ) );
                } # if homogeneous else
            } # if !homogeneous
        } # for region
        finished <- TRUE;
        nRegions <- ncol( regions );
        for ( region in 1:nRegions ) {
            finished <- finished & regions[ "Homogeneous", region ];
        } # for region
        if ( finished ) {
            regions <- t( regions );
            nRegions <- nrow( regions );
            rownames( regions ) <- 1:nRegions;
            for ( region in 1:nRegions ) {
                startPoint <- regions[ region, "Start" ];
                endPoint <- regions[ region, "End" ];
                regions[ region, "MedianLevel" ] <-
                    median( values[ startPoint:endPoint ], na.rm=TRUE );
                regions[ region, "MAD" ] <-
                    mad( values[ startPoint:endPoint ], na.rm=TRUE );
            } # for region
            return( regions );
        } # if done
    } # while TRUE
} # partition

partition <- function( values, minLength, cutoff=10 ) {
    nValues <- length( values );
    regions <- data.frame( "InnitialEdges"=c( 1, nValues, FALSE, NA,
NA ) );
    #regions <- data.frame( "FirstHalf"=c( 1, floor( nValues / 2 ), FALSE,
NA, NA ) );
```

[0710]

```
    #regions <- cbind( regions, c( floor( nValues / 2 ), nValues, FALSE,
NA, NA ) );
    rownames( regions ) <- c( "Start", "End", "Homogeneous",
"MedianLevel", "MAD" );
    return( partitionRegions( values, regions, minLength, cutoff ) );
} # partition

#########################################################################
#######
#########################################################################
#######
#########################################################################
#######
#########################################################################
#######
plotRegions <- function( chrProfile, regions, chr="", srID="" ) {
    plot( chrProfile, type="l" );
    for ( region in 1:nrow( regions ) ) {
        abline( v=regions[ region, "Start" ], col="gray" );
        abline( v=regions[ region, "End" ], col="gray" );
        polygon( x=c( regions[ region, "Start" ], regions[ region,
"Start" ],
                      regions[ region, "End" ], regions[ region,
"End" ] ),
                 y=c( regions[ region, "MedianLevel" ] - regions[ region,
"MAD" ],
                      regions[ region, "MedianLevel" ] + regions[ region,
"MAD" ],
                      regions[ region, "MedianLevel" ] + regions[ region,
"MAD" ],
                      regions[ region, "MedianLevel" ] - regions[ region,
"MAD" ] ),
                 col="yellow" );
        lines( x=c( regions[ region, "Start" ], regions[ region,
"End" ] ),
                 y=c( regions[ region, "MedianLevel" ], regions[ region,
"MedianLevel" ] ),
                 lwd=3 );
    } # for region
    par( new=TRUE );
    plot( chrProfile, type="l", lwd=1.5, main=paste( srID, ", Chr", chr,
sep="" ) );
} # plotRegions
```

[0711]

[0712] 상기 R-스크립트를 사용하여 OBX 연구로부터 샘플에서 다수의 추정되는 미세결실 및 미세중복을 검출하였다(표 3 및 도 12).

[0713] [표 3]

최대 엔트로피 방법으로 검출된 미세중복 및 미세결실 이벤트

검출된 이벤트	염색체
미세결실	5
미세중복	22
미세결실	22
미세중복	10
미세중복	21
미세중복	12
미세결실	15
미세결실	1
미세중복	19
미세중복	2
미세결실	7
미세결실	6
미세결실	4
미세중복	6
미세중복	4

[0714]

[0715] 실시예 5: 검증 방법

[0716] "리브-원-아웃" 및 "슬라이딩 에지"로서 지칭되는 2종의 검증 방법들을 PERUN과 커플링시켜 22번 염색체 상에서 구개심장안면 증후군(디조지 증후군, 22q11)으로서 공지된 서브-염색체 이상을 확인하고 22q11 영역에서 정배수체(즉, "정상") 염색체 수준으로부터의 관찰된 편차의 통계학적 유의성을 평가하였다. 상기 방법은 결실/중복의 표적화된 검출 및 비-표적화된 검출 둘 다에 적용될 수 있었다.

[0717] 16개의 임신 여성 혈장 샘플들을 종래 기재된 바와 같이(Jensen TJ, Dzakula Z, Deciu C, van den Boom D, Ehrich M. (2012): Detection of microdeletion 22q11.2 in a fetus by next-generation sequencing of maternal plasma. Clin. Chem. 58(7):1148-51) 채취하고 프로세싱하고 서열결정하였다. 2개의 샘플들은 핵형분석에 의해 확인되었을 때 디조지 태아를 가진 모체에 속한다. 남은 14개의 샘플들은 정배수체 임신(즉, "정상 태아")에 상응하고 22q11의 검출 및 특징규명을 위한 기준물로서 사용된다.

[0718] ELAND 정렬을 위해 최적화된 PERUN 부분 파라미터를 사용하여 모든 16개의 샘플들에 대한 최초 미처리 카운트(샘플당 2개의 데이터 세트)를 재프로세싱하였다. 종래 기재된 바와 같이, 상기 파라미터를 LDT2CE 데이터에 대해 트레이닝하고 교차검증을 기초로 부분 선택을 수행하였다. 맵핑가능성 기초 필터링을 적용하지 않았다. 이차 LOESS 정규화는 PERUN 정규화가 적용된 후 남아있는 임의의 GC 편향을 제거하였다. 각각의 샘플을 2회 측정하여 총 32개의 프로파일을 생성하였기 때문에, 페어링된 프로파일들을 단일 PERUN 프로파일로 조합하여 16개의 프로파일들(샘플당 1개)을 생성하였다. 프로파일들을 조합하기 전에, 그들의 표준 편차는 0.020 내지 0.030이었다. 일치된 프로파일의 추가는 1.414(2의 제곱근)의 예상된 개선보다 약간 더 작은 약 1.2의 계수(1.14 내지 1.27)만큼 가변성을 감소시켰다. 정규화된 카운트 프로파일은 미처리 프로파일 및 GCRM 프로파일뿐만 아니라 14개의 정배수체 샘플들에 기초한 중간 기준 프로파일에 대한 정규화의 결과에 비해 유의하게 개선되었다. 개선은 게놈 전체에 걸쳐 프로파일의 감소된 표준 편차 및 고도의 균일성으로부터 분명히 드러난다. 16개의 미처리 프로파일들의 표준 편차(총 카운트에 대하여 크기조정되고 기준 게놈의 부분의 수와 곱해짐)는 0.55 내지 0.64이지만, 상응하는 16개의 PERUN 프로파일들의 표준 편차는 0.016 내지 0.026이다. 도 13은 모든 16개의 프로파일들을 보여주는데, 이때 (chr22_368부터 chr22_451에 이르는) 디조지 영역은 배경에서 회색 리본으로 묘사되어 있다. 교차검증에 기초한 부분 필터링은 기준 게놈의 하기 부분 세트만을 남기면서 22q11 미세결실로부터 기준 게놈의 많은 부분들을 제거한다: chr22_371, chr22_372, chr22_380, chr22_381, chr22_382, chr22_383, chr22_384, chr22_385, chr22_386, chr22_387, chr22_388, chr22_389, chr22_390, chr22_391, chr22_392, chr22_393, chr22_394, chr22_395, chr22_396, chr22_397, chr22_398, chr22_399, chr22_400, chr22_401, chr22_402, chr22_403, chr22_404, chr22_415, chr22_416, chr22_417, chr22_418, chr22_419, chr22_422, chr22_423, chr22_424, chr22_426, chr22_427, chr22_428, chr22_439, chr22_440, chr22_441, chr22_442, chr22_443, chr22_444, chr22_445 및 chr22_446.

[0719] 도 13은 상세한 시각적 검사를 가능하게 하기 위해 디조지 영역에서 확장된 PERUN 프로파일을 보여준다. 22q11 결실은 영향을 받은 사례(3_4 및 9_10)에서 분명히 드러난다.

[0720] 22q11 결실의 존재 또는 부재에 대한 판정에서 신뢰도를 정량하기 위해, 22번 염색체의 위치 18,546,349-22,336,469를 커버하는 표준 디조지 영역에 대한 Z-점수를 평가하였다. 이 영역 내의 기준 게놈의 부분에 대한 PERUN 수준을 각각의 샘플에 대해 별도로 합산하였다. 모든 샘플들에 대한 기술적 복제물이 측정되었기 때문에, 2개의 합산의 평균치를 디조지 영역 내의 염색체 물질의 표시치로서 사용하였다. (정배수체 및 영향을 받은 샘플 둘 다를 포함하는) 모든 표시치의 중간치를 개별 표시치로부터 빼고, 차이를 모든 표시치의 MAD로 나누어 Z-점수를 생성하였다. 결과는 도 14에 제시되어 있다. 영향을 받은 2개의 사례 3_4 및 9_10은 상기 2개의 샘플들에서 결실의 존재를 확인시켜주는 -3 미만의 Z-점수를 갖는다. 또한, 한 샘플(13_14)에서 높은 양의 Z-점수는 관심있는 영역에서의 가능한 중복을 표시하였다. 도 13의 가시적 검사는 13_14 PERUN 프로파일의 중간 부분이 디조지 영역 내의 과다표시된 구획을 함유하였다는 것을 확인시켜주었다.

[0721] 도 13에 나타낸 프로파일은 영향을 받은 1개의 사례(3_4)에서 관찰된 결실이 표준 디조지 결실과 상이하다는 것을 암시하였다. 3_4 프로파일은 비정상의 예상된 에지의 좌측으로 1 Mbp(기준 게놈의 20개 부분) 초과의 길이까지 단지 부분적으로 대폭 감소되었는데, 이때 결실의 우측 에지는 부분 chr22_426에 가까이 위치하였다. 3_4에서 결실의 좌측 에지도 22q11의 예상된 좌측 에지로부터 좌측으로 이동된 것으로서 나타났다. 실제 결실 정도는 Z 값에 분명히 영향을 미쳤다. 본 연구의 첫 번째 목적은 출발 및 종결 결실/중복 에지의 위치가 주어졌을 때 결실/중복 판정에서의 신뢰도를 평가하는 것이었다.

[0722] 결실 에지의 위치의 영향을 평가하고 신뢰 구간을 Z-점수와 연관시키기 위해, 표준 디조지 영역과 부분적으로 중첩되는(또는 표준 디조지 영역 내에 함유되어 있는) 225개의 상이한 영역들의 표시치를 기초로 Z-점수의 평가를 반복하였다. 상기 영역들 중 가장 큰 영역은 chr22_371에서 시작되어 chr22_446에서 종결된다. 결실에서 마지막 부분인 chr2_447은 PERUN 교차검증에 의해 필터링되는 것으로서 계산에 포함되지 않았다. 남은 영역들은 표준 디조지 영역의 좌측 에지로부터 첫 번째 부분과 열다섯 번째 부분 사이의 임의의 위치에서 시작된다. 나아가, 상기 영역들은 표준 디조지 영역의 우측 에지와 이보다 앞서는 열다섯 번째 부분 사이의 임의의 위치에서 종결된다. 이것은 영역 출발/종결 점의 15x15 격자를 생성하였다. 모든 샘플들에 대한 표시치를 이 격자 상에서 평가하였고, 모든 격자 점에 대한 Z-값을 수득하였다. 모든 샘플들은 Z-표준화를 위해 사용된 중간 표시치 및 MAD 표시치에 기여하였다. (샘플 3_4 및 1_2에 대한) 모든 225개의 가능한 영역들에 대한 수득된 Z-값의 대표적인 히스토그램이 도 15 및 16에 제시되어 있다.

[0723] 도 17의 산점도는 모든 샘플들에 대한 모든 225개의 가능한 영역들에 대한 수득된 Z-값의 히스토그램을 요약한다. 도 17은 샘플당 중간 Z-값, 및 중간 Z-값 주변의 3개의 MAD 신뢰 구간만을 보여준다. 결실 에지의 선택에 따라, 샘플 13_14는 대체로 과다표시되는 듯하였다(15x15 격자에서 대다수의 영역들에 대한 Z > 3). 영향을 받은 2개의 샘플들(3_4 및 9_10)에 대한 3개의 MAD 신뢰 구간은 "정상" 영역과 부분적으로 중첩되었지만(-3부터 3까지), 상기 두 샘플들은 대체로 -3 미만으로 유지되었다.

[0724] 도 20은 미세결실 에지의 선택으로부터 도출되는 가변성을 조사하였다. 가변성의 또 다른 가능한 공급원은 Z-점수 표준화를 위한 기준 샘플의 선택일 수 있다. 기준 샘플의 선택이 결실/중복 판정에서 Z-점수 및 수득된 신뢰도(또는 신뢰도의 결여)에서의 가변성에 기여하는 정도를 평가하기 위해, 단일 영역뿐만 아니라 영역의 15x15 격자에 대해서도 "리브 원 아웃" 분석을 수행하였다. 표준 디조지 영역에만 적용되었을 때, "리브 원 아웃" 분석은 기준물의 선택이 Z-점수에서의 가변성에 유의하게 기여한다는 것을 표시하지 못하였다(도 18 내지 20). 그러나, 영역의 15x15 격자에 대해 수행된 보다 더 포괄적인 "리브 원 아웃" 분석은 기준 샘플의 선택이 Z-점수에서의 가변성에 유의한 영향을 미쳤다는 것을 확인시켜주었다(도 21 내지 23).

[0725] 도 21은 디조지 하위영역의 15x15 격자를 사용함으로써 수득된 중간 Z-값과 "리브 원 아웃" 기법에 의해 생성된 중간 Z-값 사이의 일치를 입증하였다.

[0726] 도 22 내지 24는 하위영역의 15x15 격자로부터 무작위적으로 선택된 디조지 하위영역에 대한 "리브 원 아웃" 기법의 이용을 입증하였다. 표준 디조지 영역의 경우와 동일한 결론이 도출될 수 있다.

[0727] 도 25 내지 29는 "리브 원 아웃" 기법과 영역 에지의 슬라이딩을 병용한 결과를 보여준다. 15x15 하위영역 각각에 대해, 16개의 기준 세트들은 샘플당 3,600개의 Z-점수의 세트로 조합된 Z-점수를 생성하였다. 대표적인 도 25 및 26에 나타낸 분포는 추가 세부사항을 제공하면서 도 14 내지 24로부터 도출된 결론과 일반적으로 일치하였다. 일부 히스토그램들에서 관찰된 날카로운 피크는 샘플 서브세트가 종종 Z-점수 표준화를 위한 중간 값을

제공한다는 사실에 의해 설명될 수 있다.

[0728] "슬라이딩 에지" 및 "리브 원 아웃" 방법은 판정을 검증하는 데에 함께 또는 독립적으로 이용될 수 있다. 또한, 상기 2종의 별도의 절차들은 병용되어 유도된 Z-점수에서의 불확실성의 보다 더 포괄적인 사진을 생성한다. 상기 2종의 기법들은 표적화된 결실/중복 검출과 관련하여 본원에 적용되었지만, 원칙적으로 게놈 전체에 걸쳐 이전에 공지되어 있지 않은 결실/중복에 대한 비-표적화된 "헌트(hunt)"까지 확장될 수 있다. 일단 영향을 받은 영역이 웨이블렛, 최대 엔트로피, 서큘러 이진 분할, 에지 검출 커넬을 이용한 컨볼루션 또는 일부 다른 적절한 방법을 이용함으로써 대략적으로 파악되면, 슬라이딩 에지 기법 및 리브 원 아웃 기법을 적용하여 새로 검출된 결실/중복의 정도 및 신뢰도를 알아낼 수 있다.

[0729] 슬라이딩 에지 분석에 대한 동기부여는 환자 3_4에서 관찰된 22q11 결실의 정도와 표준 디조지 결실 사이에 관찰된 불일치로부터 나온다. 밝혀진 바와 같이, 3_4 샘플에서 관찰된 2.5 Mb 결실은 이 계산이 수행된 시점에서 공지되어 있는 것보다 더 잘 표시되었다. 이들 두 공급원에 따를 때, 전형적인 결실은 표준 8 Mb라기보다는 약 3 Mb이다: (C. Carlson, et al., (1997) The American Journal of Human Genetics, Volume 61, Issue 3, 620-629); (Schwinger E, Devriendt K, Rauch A, Philip N. Clinical utility gene card for: DiGeorge syndrome, velocardiofacial syndrome, Shprintzen syndrome, chromosome 22q11.2 deletion syndrome (22q11.2, TBX1). Eur J Hum Genet. 2010 Sep;18(9). doi: 10.1038/ejhg.2010.5. Epub 2010 Feb 3. PubMed PMID: 20125192; PubMed Central PMCID: PMC2987430).

[0730] 디조지 결실은 일부 7% 내지 8%의 사례에서 1.5 Mb만큼 짧은 것으로서 보고되어 있다. (높은) 임상적 관련성을 갖는 염색체 이상의 예상된 크기와 실제 크기 사이의 이러한 차이는 본 데이터가 암시하는 것보다 더 흔할 수 있다. 이러한 이유로, 특정 이상에 맞게 조정된 표적화된 방법, 및 먼저 이상을 발견한 후 임상적 해석의 데이터베이스를 질의하는 보다 더 일반적인 비-표적화된 방법 둘 다가 슬라이딩 에지 및 리브 원 아웃 분석으로부터 이익을 얻을 것이다.

[0731] 실시예 6: 유전적 변이의 로그 승산비 검출

[0732] 태아 분획은 유전적 변이에 대한 비침습 출생전 검사(NIPT)에서 일정한 역할을 수행한다. 높은 태아 분획(예를 들면, 24%) 및 약간 상승된 z-점수(예를 들면, z = 3.2)를 갖는 샘플이 거짓 양성 분류를 초래할 수 있다는 것이 관찰되었다. 이러한 샘플에 대한 z-점수는 예를 들면, 진짜 삼염색체성인 경우 3보다 훨씬 더 높아야 한다. 로그 승산비(LOR) 프로세스는 이 문제를 해결하고 거짓 양성 판정의 가능성을 감소시키기 위해 개발되었다.

[0733] 검사 샘플에 대한 LOR은 그의 관찰된 z-점수 및 태아 분획이 주어졌을 때 삼염색체성 21(T21) 대 비-T21의 확률에 따라 방정식 22에 의해 계산될 수 있다:

[0734] 방정식 22

[0735] $$log\frac{P(T21|Z,\hat{f})}{P(\overline{T21}|Z,\hat{f})} = log\frac{P(Z|T21,\hat{f})P(T21,\hat{f})}{P(Z|,\hat{f})P(\overline{T21},\hat{f})} = log\frac{P(Z|T21,\hat{f})P(T21)}{P(Z|\overline{T21},\hat{f})P(\overline{T21})}$$

[0736] 상기 식에서,

[0737] 본원에서 *f*^로서도 지칭되는 $\hat{f}$는 (예를 들면, 남성 샘플에 대한 chrY 또는 공지된 다른 태아 분획 측정 기법에 의해) 측정된 태아 분획이고,

[0738] $P(T21|Z,\hat{f})$ 및 $P(\overline{T21}|Z,\hat{f})$는 Z 및 *f*^가 하기 기재된 바와 같이 주어졌을 때 각각 $T21$(T21을 가짐) 및 $\overline{T21}$ (비-T21)의 사후 확률이고,

[0739] 이때, $P(T21)$ 및 $P(\overline{T21})$는 각각 T21 및 비-T21에 대한 사후 확률이고,

[0740] $P(Z|T21,\hat{f})$ 및 $P(Z|\overline{T21},\hat{f})$는 본원에서 유도된, 각각 T21 및 비-T21에 대한 조건부 확률이다.

[0741] 조건부 확률은 검사 샘플에 대한 z-점수(Z) 및 계산된 태아 분획(*f*^)에 따라 결정될 수 있다(방정식 22의 우측 부분 참조). 영향을 받지 않은 정배수체 샘플의 경우, X는 이벤트 영역에 대한 합산된 빈(bin) 카운트를 나타낸다. 서열결정에서 고유 무작위성으로 인해, X는 $X \sim f(\mu_X, \sigma_X)$를 갖는 무작위 변수이고, 이때 μ_X 및 σ_X는 각

각 평균 및 표준 편차이고, f(·)는 분포 함수이다. 유사하게, 영향을 받은 삼염색체성 샘플의 경우, 영향을 받은 영역에 대한 빈 카운트는 Y ~ $f(\mu_Y, \sigma_Y)$이고, 이때 μ_Y는 $\mu_X(1 + f/2)$이고 f는 태아 분획이다. σ_Y가 σ_X와 거의 동등하다고 가정할 때, z-점수 분포는 방정식 23에서와 같이 기재될 수 있다:

[0742] 방정식 23

[0743] $$Z \sim Normal\left(\frac{\mu_X}{\sigma_X}\frac{f}{2}, 1\right)$$

[0744] 상기 식에서,

[0745] μ_X 및 σ_X는 정배수체 샘플의 큰 풀로부터 실험적으로 평가될 수 있다.

[0746] 정배수체 샘플의 경우, 그의 z-점수는 태아 분획과 무관하고 표준 정규 분포를 따른다.

[0747] 방정식 23은 도 32에 도식적으로 표시되어 있다. 정배수체 대상체에 대한 z-점수 분포는 0에 집중되고, 분포의 중심은 태아 분획에 민감하지 않다. 방정식 23의 도식적 표시로서, T21 각각에 대한 별개의 z-점수 분포는 상이한 z-점수에 집중되고, 이때 각각의 별개의 분포는 상이한 태아 분획에 대한 분포이다. 검사 샘플에 대해 측정된 태아 분획의 적용은 T21에 대한 어떤 z-점수 분포가 평가될지를 결정한다. 검사 샘플에 대해 측정된 z-점수의 적용은 T21에 대한 조건부 확률을 확인한다. T21에 대한 조건부 확률은 검사 샘플에 대해 측정된 z-점수와 그 검사 샘플에 대해 측정된 태아 분획에 따라 선택된 T21에 대한 z-점수 분포의 교차점에 존재한다. 검사 샘플에 대해 측정된 z-점수의 적용은 비-T21에 대한 조건부 확률도 확인한다. 비-T21에 대한 조건부 확률은 검사 샘플에 대해 측정된 z-점수와 비-T21에 대한 z-점수 분포의 교차점에 존재한다.

[0748] T21 및 비-T21에 대한 조건부 확률을 방정식 22에 적용하여 검사 샘플에 대한 LOR 계산을 제공한다. 3.95 초과의 z-점수 및 0 초과의 LOR을 갖는 검사 샘플은 유전적 변이의 존재(예를 들면, T21의 존재)를 갖는 것으로서 분류된다. 3.95 미만의 z-점수 및/또는 0 미만의 LOR을 갖는 검사 샘플은 유전적 변이의 부재를 갖는 것으로서 분류된다.

[0749] 도 31은 LOR 방법을 이용하여 LDTv2 남성 샘플에 대한 분류 결과를 보여준다. 알 수 있는 바와 같이, LOR은 태아 분획이 증가함에 따라 급속히 발산된다. 따라서, 높은 태아 분획을 갖는 경계선 샘플이 존재할 수 있고 신뢰 가능하게 분류되었다. 구체적으로, 0 초과의 LOR을 갖는 개방된 원으로 표시된 샘플은 T21로서 정확히 분류되었고, 0 미만의 LOR을 갖는 개방된 원으로 표시된 샘플은 비-T21로서 정확히 분류되었다.

[0750] 방정식 22는 측정된 태아 분획이 진짜 태아 분획과 동일하다고 가정하는데, 이것은 종종 실제로는 그렇지 않다. 측정 불확실성을 보상하기 위해, 진보된 LOR 방법을 방정식 24에 따라 개발하였다:

[0751] 방정식 24

[0752] $$\frac{P(T21|Z,\hat{f})}{P(\overline{T21}|Z,\hat{f})} = \frac{P(T21,Z,\hat{f})}{P(\overline{T21},Z,\hat{f})} = \frac{\sum_f P(T21,Z,f,\hat{f})}{\sum_f P(\overline{T21},Z,f,\hat{f})}$$

$$= \frac{\sum_f P(Z|T21,f,\hat{f})P(T21|f,\hat{f})P(f|\hat{f})P(\hat{f})}{\sum_f P(Z|\overline{T21},f,\hat{f})P(\overline{T21}|f,\hat{f})P(f|\hat{f})P(\hat{f})}$$

$$= \frac{\sum_f P(Z|T21,f,\hat{f})P(f|\hat{f})}{\sum_f P(Z|\overline{T21},f,\hat{f})P(f|\hat{f})}\frac{P(T21)}{P(\overline{T21})} = \frac{\sum_f P(Z|T21,f)P(f|\hat{f})}{\sum_f P(Z|\overline{T21},f)P(f|\hat{f})}\frac{P(T21)}{P(\overline{T21})}$$

[0753] 상기 식에서,

[0754] f는 진짜 태아 분획이고 $\hat{f}$는 측정된 태아 분획이다. 방정식 24는 방정식 22에 비해 $\hat{f}$가 주어졌을 때 f의 조건부 확률에 대한 평균을 산출한다. 따라서, 진보된 LOR은 조건부 확률의 가중된 평균을 사용하고, 측정된 태아 분획 값에 더 가까운 가능한 태아 분획 값에게 보다 더 큰 가중을 제공한다. 따라서, 방정식 24를 기초로, 진보된 LOR은 진짜 양성 샘플을 회수하였다. 따라서, 진보된 LOR은 검사 샘플에서 유전적 변이의 존재 또는 부재를 확인하는 데에 이용될 수 있다.

[0755] LOR 방법은 (예를 들면, T21 이외의) 다양한 종류의 염색체 이배수성의 존재 또는 부재를 확인하는 데에 적용될

수 있고, 다양한 종류의 다른 유전적 변이(예를 들면, 21번 염색체 이외의 염색체의 이배수성, 미세중복, 미세결실)의 존재 또는 부재를 확인하는 데에 적용될 수 있다. 양성 이벤트(예를 들면, 염색체 이배수성(예를 들면, 일염색체성, 삼염색체성)의 존재); 미세중복의 존재; 미세결실의 존재)는 z-점수가 3.95 이상이고 LOR이 0보다 더 클 때 확인된다. 미세중복 또는 미세결실의 존재 또는 부재를 확인하기 위해, 예를 들면, 각각 웨이블렛 평활화된 분해 렌더링 또는 CBS 평활화된 분해 렌더링으로부터 생성된 Z_{wave} 또는 Z_{cbs} z-점수를 사용할 수 있다(예를 들면, 실시예 3 참조). 도 32에 나타낸 관계는 미세결실 또는 미세중복 이벤트에 대해 유사하지만 상이하다. 미세중복 이벤트의 경우, 중복 분절에 대한 μ_X 및 σ_X는 염색체 삼염색체성에 대한 값보다 더 작은 값을 갖고, 미세중복 z-점수 분포는 도 32에서 정배수체 z-점수 분포에 더 가깝게 좌측으로 이동한다. 미세중복 이벤트에 대한 z-점수 분포는 도 32에서 정배수체 z-점수 분포의 우측 상에서 유지될 것이다. 그러나, 미세결실 이벤트에 대한 z-점수 분포는 도 32에서 정배수체 z-점수 분포의 좌측으로 이동할 것이다. (예를 들면, 21번 염색체 이외의 염색체에 대한) 염색체 이배수성의 존재 또는 부재를 확인하기 위해, 예를 들면, Z_{chr} z-점수 값을 사용할 수 있다(예를 들면, 실시예 3 참조).

[0756] 실시예 7: ChAI 정규화 프로세스

[0757] ChAI는 검사 대상체(예를 들면, 임신 여성)로부터 수득된 서열 리드로부터 태아에서의 유전적 변이(예를 들면, 이배수성, 미세중복, 미세결실)의 존재 또는 부재를 확인하는 데에 사용될 수 있는 시스템이다. ChAI에 대한 시스템 순서도의 예는 도 42a 및 42b에 제시되어 있다. 서열 리드를 본원에서 종종 트레이닝 세트로서 지칭되는 임신 여성 검사 대상체 및 하나 이상의 기준 대상체로부터 수득하였다. 트레이닝 세트의 임신 여성 대상체는 다른 검사 방법에 의해 확인되었을 때 정배수체인 태아를 가졌다.

[0758] 서열 리드를 먼저 SAM 또는 BAM 포맷으로부터 이진 리드 포맷(BReads 포맷)(ChAI가 훨씬 더 신속하게 실시되는 것을 가능하게 함)으로 압축하였다. BReads 포맷은 기준 게놈에 따라 측정된 염색체 및 염기쌍 위치를 포함하는 각각의 리드의 게놈 위치를 저장하고 다른 정보를 버린다. BReads 파일은 함유된 리드의 카운트로 시작된다. 이것은 메모리 재할당에 대한 필요성을 제거함으로써 로딩 횟수를 개선한다. 값을 4 바이트 어레이로서 디스크 상에 저장하였다. 그 다음, 5 바이트 포맷(염색체 서수를 위한 1 바이트(1-22, X, Y, M의 제로-인덱스) 및 염색체 위치를 위한 4 바이트)을 사용하여 리드를 저장하였다. 처음 4 바이트로부터 서열 리드 카운트를 먼저 읽음으로써 BReads 파일을 로딩하였다. 그 다음, 각각의 서열 리드는 한번에 5 바이트씩 로딩되는데, 이때 처음 바이트는 염색체 서수를 표시하고 다음 4 바이트는 정수 위치로 전환된다. 특정 리드 인덱스에 대한 디스크-스킵(disk-skip) 명령어를 사용함으로써 리드의 무작위 샘플링을 신속히 수행할 수 있다.

[0759] 일례로서, 17,673,732개의 맵핑된 리드들에 대해 상이한 포맷들의 디스크 용량이 표 4에서 Breads 포맷의 디스크 용량과 비교되어 있다.

[0760] [표 4]

17,673,732 개의 리드를 갖는 샘플에 기초한 상이한 포맷에 대한 디스크 용량

포맷	*공간 용량*
SAM	4.0 GB
맵핑된 리드 위치	247 MB
GZip 리드 위치	97 MB
BReads	85 MB

[0761]

[0762] BReads 포맷은 원래의 SAM 파일보다 대략 50배 더 작았고 GZip 포맷보다 약 12% 더 작은 공간을 사용하였다. BReads는 1회 메모리 할당을 위해 헤드에서 리드의 수를 저장한다는 장점도 가졌고, 리드들이 순서대로 읽혀질 필요가 없기 때문에 신속히 샘플링될 수 있다. 이 특징들은 다른 포맷을 사용하였을 때에는 가능하지 않았다.

[0763] GC 편향의 모델링

[0764] 그 다음, 각각의 샘플에 대한 GC 편향 모델을 기억시켰다. 트레이닝용으로 지정된 샘플을 사용하여 부분적으로 부분 필터를 생성하고 GC 편향만으로 잘 설명되지 않는 다른 게놈 편향을 기억시켰다. 마지막으로, 트레이닝 통계치를 사용하여 검사 샘플을 필터링하고 점수화하였다.

[0765] ChAI는 국소 GC 함량의 밀도 추정치를 사용하여 GC 편향을 모델링하였다. 커널 함수, 예컨대, 에파네크니코브 커널을 사용하여 기준 게놈으로부터 GC 밀도를 추정하였다(도 33). 가우스(Gaussian) 또는 트리웨이트(triweight) 커널을 포함하는 다른 커널도 적절하다. 밴드폭은 200 bp로서 선택되었으나, 밴드폭 파라미터는 변

경가능하다.

[0766] 커넬을 사용하여 기준 게놈에 대한 GC 밀도를 염기쌍 해상도에서 추정하였다(예를 들면, 도 34에 나타낸 바와 같음). 기준물의 GC 밀도 추정치를 사용하여 샘플로부터의 각각의 리드의 국소 GC 함량을 측정하였다. 그 다음, 샘플에 대한 GC 밀도 추정치의 분포를 전체 기준 게놈에 대한 분포와 비교하여 GC 편향을 측정하였다(도 35). AT 풍부 영역(GC 밀도 = 0)으로 맵핑되는 리드 및 기준 값을 버렸다. 샘플 분포의 밀도로 나누어진 기준 분포의 밀도의 로그 비에 대해 피팅된 다항식을 이용하여 샘플의 GC 밀도 분포와 기준물 사이의 차이를 모델링하였다(도 36). 모델을 가중된 방식으로 피팅하였는데, 이때 각각의 가중은 주어진 GC 밀도 값에 대한 샘플의 분포 밀도 값으로서 수득되었다. 이것은 분포의 꼬리가 피트를 과도하게 유도하지 않았다는 것을 보장하였다. 다른 피팅 모델, 예컨대, 변위치 회귀 모델 또는 파라미터화된 분포가 적절한 경우 편향의 분포에 대해 이용될 수 있다.

[0767] 피트 GC 모델을 사용하여 샘플에 대한 서열 리드의 각각의 카운트를 가중하여 기준물에 비해 그의 과다표시 또는 과소표시에 대해 조절하였다. ChAI 알고리즘은 이 가중을 리드 밀도의 추정에 도입하여 GC 편향을 보정할 수 있었다.

[0768] 다차원적 편향 보정

[0769] GC 편향은 게놈에서 리드 패턴에 영향을 미치는 여러 편향들 중 단지 하나이었다. 일반화된 다변량 모델을 사용하여 리드 가중을 추정하기 위해 종종 추가 편향을 모델링하고 보정하였다. 이 보정을 다음과 같이 수행하였다:

[0770] 1. 게놈 위치의 각각 서브세트에서 시험 샘플 및 기준 게놈에 대한 N 편향 값을 추정하였다.

[0771] 2. N-차원 평활화 커넬 또는 적절한 파라미터 함수를 사용하여 편향 값의 밀도를 모델링하였다.

[0772] 3. 기준물로부터 수득된 밀도 값의 세트 및 검사 밀도에 대한 로그 비를 계산하였다.

[0773] 4. 선택된 점을 다변량 모델(예를 들면, 각각의 차원에 대한 가중된 3차 다항식)과 함께 사용하여 밀도의 로그 비를 모델링하였다.

[0774] 5. 모델을 사용하여 기준물에 비해 주어진 리드의 빈도의 비를 추정하였고, 적절한 가중을 배정하였다.

[0775] 부분 필터링

[0776] 게놈에 대한 서열 리드(예를 들면, 카운트)의 표시치를 기초로 염색체 이상에 대해 샘플을 점수화하였다. 이 표시치는 국소 GC 추정을 위해 사용된 밀도 함수와 유사한 밀도 함수를 사용함으로써 측정되었다. 리드 밀도 커넬은 일반적으로 훨씬 더 큰 밴드폭을 갖고, 이때 디폴트는 50,000 bp이다, 리드의 각각의 카운트는 GC 편향 모델로부터의 그의 가중과 동등한 값으로 밀도에 기여한다. 리드 밀도는 임의의 또는 모든 염기쌍에서 평가될 수 있으나, 연산 수행의 경우 일부 위치들만을 사용하였다. 이 위치들은 "부분"으로서 지칭되었다. 부분은 리드 밀도를 추정하는 것이 가장 중요한 곳마다 위치될 수 있다. 염색체 이배수성의 분류를 위해 처음에(예를 들면, 필터링 전에) 부분을 게놈 전체에 균일하게 공간적으로 분리시켰다. 각각의 부분은 50,000 bp 윈도우로 구성되었고, 필터링 전에 25,000 bp만큼 다음 인접 부분과 중첩되었다.

[0777] 일부 부분들은 잘 맵핑되지 않은 게놈 영역을 포함하고, 이것은 샘플 사이에 리드 밀도에서의 극단의 변동을 유발하였다. ChAI는 트레이닝 세트를 사용하는 필터링 프로세스로 이 부분들을 확인하고 제거하였다. 중간 값(예를 들면, 도 37a) 및/또는 MAD 값(예를 들면, 도 37b)에서 큰 편차를 보여준 부분을 고려사항으로부터 제거하였다. 이 편차의 역치는 트레이닝 집단 사분위수로부터 사분위수간 범위의 4배 초과의 수준만큼 벗어나는 임의의 값으로서 수득되었다(도 37a 및 37b). 이 역치를 미세하게 조정하여 ChAI 파라미터의 특정 세트에 대한 검사 성능을 최대화할 수 있다.

[0778] 트레이닝 및 점수화

[0779] 필터링된 부분들에 맵핑된 리드만을 사용하여 각각의 샘플의 게놈 리드 밀도 프로파일을 계산하였다. 그 다음, 트레이닝 세트의 일부인 샘플을 사용하여 트레이닝 통계치를 추정하였고, 이 통계치를 사용하여 검사 세트를 점수화하였다. 이 통계치는 점수화 검사 통계치에 대한 부분 중간치, 주성분 및 널 분포로 구성되었다. 임의의 수의 생물학적 인공물(artifact) 및 기술적 인공물로부터 존재할 수 있는 게놈-범위 리드 편향을 모델링하기 위해 부분 중간치 및 주성분을 사용하였다(도 38a 내지 38c). 나머지 샘플에 대한 극단의 부분 값의 영향을 최소화하기 위해, 샘플에서 다른 부분에 걸쳐 4xIQR을 벗어나는 각각의 값을 4xIQR로 트리밍하였다.

[0780] 먼저 검사 부분 값으로부터 트레이닝된 중간 값을 뺌으로써 감춰진 편향에 대해 검사 샘플을 보정하였다. 상위 트레이닝된 주성분들과 상관관계를 갖는 샘플 값의 성분도 제거하였다. 이것은 주성분 항을 기초로 다변량 선형 회귀를 이용하여 부분 값을 모델링함으로써 수행되었다(도 39a 내지 39c). 모델에 의해 예측된 값을 샘플 값으로부터 빼어 비-편향된 잔차만을 남겼다. 사용된 주성분의 수는 임의적이고, 이때 디폴트는 8이다.

[0781] 보정 후, 피셔 정확 검정을 이용하여 샘플을 점수화하였다. 이 검정은 관심있는 염색체 영역에서 트레이닝된 중간치보다 더 크거나 더 작은 값을 갖는 부분의 수를 비교한다. 이 카운트를 게놈 내의 나머지 부분에 대해 평가하였다. 점수화 통계치를 음의 log10 p-값으로서 수득하였다. 다른 점수화 통계치, 예컨대, 윌콕손 부호-순위 검정(Wilcoxon signed-rank test) 또는 F-검정을 이 단계에서 이용할 수 있다.

[0782] 부분들 사이의 잔류 상관관계로 인해, 검정 통계치는 트레이닝 샘플 및 검사 샘플 둘 다에서 상승되었다. 이 상승은 트레이닝 세트의 부트스트랩으로부터 추정되었다(도 40).

[0783] 이 널 분포를 실험 배경으로서 사용하여 검사 샘플에 대한 점수를 보정하였다. 널 분포의 꼬리의 파레토(Pareto) 외삽을 이용하여 실험 분포에서의 점수보다 훨씬 더 큰 점수를 보정하였다.

[0784] 성별 판정

[0785] 샘플의 주성분 프로파일로부터 성별을 확인하였다. 트레이닝 데이터 세트에서, 이차 주성분(예를 들면, PC2)은 성별과 높은 상관관계를 가졌다. 이 성분의 회귀 계수를 검정 통계치로서 사용하는 것은 성별의 고도로 정확한 검사이었다(도 41a 및 41b).

[0786] 부분 의존성의 제거

[0787] 방법의 예측력을 개선하기 위해 ChAI 실시 동안에 추가 단계를 수행하였다. 이것은 부분-샘플 매트릭스에서 상관관계 구조의 양을 감소시키는 것을 포함하고, 이러한 감소는 변수 독립성의 검정 가정을 더 잘 뒷받침하고 널 순열에서 유의한 점수의 빈도를 감소시켰다. 상기 방법은 부분을, 거의 모든 동일한 정보를 함유하되 상관관계 구조를 갖지 않는 오르토고날 고유 부분으로 대체하는 단계를 포함하였다.

[0788] 제1 단계는 트레이닝 부분 세트 M에 대한 변환 매트릭스 Meig를 기억시키는 것이었다:

[0789] 1. SVD 분해: $M = U * D * VT$

[0790] 2. (예를 들면, D의 N 대각선 요소의 누적 분율이 95%보다 크도록) 독립적인 고유 부분 N의 수를 선택한다.

[0791] 3. 슈도인버스(pseudoinverse)를 계산한다: $Meig = pinv(U[\cdots,1:N] * D[1:N, 1:N])$

[0792] 부분 매트릭스 M의 임의의 서브세트를 그의 상응하는 Meig와 좌측-곱셈하여 그 서브세트의 차원-감소된 상관관계 부재 표시치를 생성하였다. 이 방식으로 트레이닝 데이터 세트에 대한 Meig를 유도하였고 추가 변경 없이 검사 샘플에 적용하였다.

[0793] Meig는 검정 변수를 변환시키는 데에도 사용되었다. 검정 변수는 예상된 편차의 위치(예를 들면, 21번 염색체 부분)에서 하나씩 존재하는 모든 0으로 구성된 벡터로서 표시되었다. 좌측-곱셈을 통해 Meig로 이 벡터를 변환시켜 변환된 부분 데이터를 적절하게 일치시켰다.

[0794] 이 방법만이 트레이닝 세트에 존재하는 샘플만큼 많은 독립적인 고유 부분들을 생성할 수 있다. 예를 들면, 50,000개 부분 및 1,000개 샘플의 트레이닝 세트의 경우, 변환된 데이터는 1,000개 이하의 부분을 함유할 것이다. 이것은 부분의 수를 현저히 감소시키는 과다보정일 가능성이 있었다. 상기 방법은 부분 데이터의 보다 더 작은 서브세트에 대해 별개의 Meig 변환을 연산하고 이들을 별도로 적용함으로써 더 느슨하게 수행될 수 있다. 이것은 인접한 부분들로부터 국소 상관관계 구조를 제거하는 데에 특히 유용하였다. 다른 방법을 이용하여 부분 상관관계 구조를 감소시킬 수도 있다. 예를 들면, 많은 클러스터링 방법들을 이용하여 (예를 들면, 군 평균 또는 중심을 기초로) 부분들을 나누고 이들을 보다 더 작은 밀집된 부분 세트로 대체할 수 있다.

[0795] 분포/프로파일 생성 모듈

[0796] 서열 리드 데이터(예를 들면, BReads)로부터 리드 밀도 프로파일을 생성하기 위해 스크립트(script)를 자바로 작성하였다. 각각의 서열 리드에 대한 리드 데이터를 수집하고, 적절한 리드 밀도 윈도우(예를 들면, 부분에 대한 개별 리드 밀도)에서 샘플의 GC 편향 보정에 따라 부분의 중앙 또는 중간 점으로부터 리드까지의 거리만큼 가중된 밀도 프로파일을 업데이트하기 위해 하기 코드를 디자인하였다. 하기 스크립트는 관계 모듈 또는 편향

보정 모듈로부터 생성된 가중되고/되거나 정규화된 카운트를 판정할 수 있거나 사용할 수 있다. 일부 실시양태에서, 분포 모듈은 하기 나타낸 자바 스크립트의 일부 또는 전부, 또는 이의 변경물을 포함할 수 있다. 일부 실시양태에서, 프로파일 생성 모듈은 하기 나타낸 자바 스크립트의 일부 또는 전부, 또는 이의 변경물을 포함할 수 있다:

```
package utilities.genome;
import java.util.Iterator;

import utilities.data.VectorUtil;
import utilities.text.DataFormatter;

public class ChromDensScaleRunnable implements Runnable{

	private GenomeScaleBoolean mask;
	private GenomeScaleFloat density;
	private final String modelPath;
	private final String brPath;
	private final int bandwidth;
	private final GenomeFloat gcdens;
	private final int stepSize;
	private final int sampleSize;
	private final int shift;
	private final String report;

	public ChromDensScaleRunnable(String modelPath, String brPath, int bandwidth,
GenomeFloat gcdens, int stepSize, GenomeScaleBoolean mask, String report, int sampleSize, int
shift)
	{
		this.modelPath = modelPath;
		this.brPath = brPath;
		this.bandwidth = bandwidth;
		this.gcdens = gcdens;
		this.stepSize = stepSize;
		this.mask = mask;
		this.report = report;
		this.sampleSize = sampleSize;
		this.shift = shift;
```

[0797]

```
	}
	public void run()
	{
		double[] mdat = (gcdens==null) ? null : VectorUtil.loadDoubleFromFile(modelPath,
6);

		//Build density
		density = new GenomeScaleFloat(stepSize);

		double correction = 0;

		try
		{
		Iterator<GenomicPosition> readIterator = (sampleSize==-1) ?
GenomeIO.scanBReads(brPath) : new BReadsSampler(brPath, sampleSize,true);

		while (readIterator.hasNext())
		{
			GenomicPosition gp = readIterator.next().shift(shift);
			int pos = gp.pos;
			int start = Math.max(0, pos-bandwidth);
			int end = Math.min(pos+bandwidth, GenomeUtil.chromosomeSize(gp.chr)-
1);

			int cindex = gp.chr.ordinal();

			double weight;
			if (gcdens!=null)
			{
				float gc = gcdens.values[cindex][pos-1];
				if (gc==0) continue;

				weight = modelWeight(mdat, gc);
			}else weight = 1;


			int[][] gpoints = density.getScalePoints(cindex, start, end, mask);

			if (gpoints[0].length==0 || Double.isNaN(weight)) continue;
			if (weight>2) weight = 2;
			if (weight<.5) weight = .5;

			correction += weight;

			for (int i=0;i<gpoints[0].length;i++)
				density.values[cindex][gpoints[0][i]] += kernel((gpoints[1][i]-
pos)/(double)bandwidth) * weight;
		}
```

[0798]

```
                }catch (Exception e)
                {
                        System.out.println("THROW!");
                        e.printStackTrace();
                        System.exit(0);
                }

                //System.out.println(GenomeIO.countReadsFromBReads(brPath));
                //System.out.println(correction);

                //Normalize intensity
                for (int i=0;i<density.values.length;i++)
                        for (int j=0;j<density.values[i].length;j++)
                        {
                                float blah = density.values[i][j];

                                density.values[i][j] /= correction;


                                if (Double.isNaN(density.values[i][j]) ||
Double.isInfinite(density.values[i][j]))
                                {
                                        System.out.println("NA val2: "+modelPath+",
"+density.values[i][j]+",  "+blah+",  "+correction);
                                        System.exit(0);
                                }
                        }


                if (report!=null) System.out.println(report);
        }

        public GenomeScaleFloat density()
        {
                return density;
        }

        private static double kernel(double x)
        {
                return .75 * (1.0 - x*x);
        }

        public static double modelWeight(double[] mdat, double gcdens)
        {
                if (mdat[5]==1) gcdens = Math.log(gcdens);

                double x2 = gcdens * gcdens;
```

[0799]

```
                return Math.pow(2, mdat[0] + mdat[1] * gcdens + mdat[2] * x2 + mdat[3] * x2 *
gcdens);
        }

}
```

[0800]

[0801] 필터링 모듈

[0802] 리드 밀도 프로파일의 부분을 필터링하기 위해 스크립트를 R로 작성하였다. 이 코드는 모든 샘플들에 걸쳐 리드 밀도 프로파일을 조사하고, 사분위수간 범위를 기초로 보유되는 부분 및/또는 버려지는(예를 들면, 분석으로부터 제거되는) 부분을 확인한다. 일부 실시양태에서, 필터링 모듈은 하기 나타낸 자바 스크립트의 일부 또는 전부, 또는 이의 변경물을 포함한다:

```
rcodepath <- "I:/ghannum/Projects/Binless/RCode"
mdistpath <-
"I:/ghannum/Projects/Binless/Reference/MarkerDistribution_LDTv2_200_50000_50000.txt"
outpath <-
"I:/ghannum/Projects/Binless/Reference/LDTv2_200_50000_50000_MarkerMask.txt"

args <- commandArgs(trailingOnly = TRUE)
rcodepath <- args[1]
mdistpath <- args[2]
outpath <- args[3]

source(paste(rcodepath,"/src/utilities/scanmatrix.R",sep=""))

dat <- scanMatrix(mdistpath,rownames=FALSE,colnames=TRUE)

m <- apply(dat,1,median)
v <- apply(dat,1,mad)

qm <- quantile(m,c(.25,.75))
qv <- quantile(v,c(.25,.75))

scalem <- qm[2]-qm[1]
scalev <- qv[2]-qv[1]

ok <- m > qm[1]-4*scalem & m < qm[2]+4*scalem & v > qv[1]-4*scalev & v < qv[2]+4*scalev

write.table(matrix(as.integer(ok),1),row.names=F,col.names=F,quote=F,file=outpath,sep="")
```

[0803]

[0804] 편향 밀도 모듈, 관계 모듈, 편향 보정 모듈 및 작도 모듈

[0805] 편향 밀도를 생성하고 관계를 생성하고 비교하고 서열 리드에서 편향을 보정하기 위해 스크립트를 R로 작성하였다. 이 코드는 일반적으로 마이크로프로세서가 하나 이상의 샘플을 분석하고 각각의 샘플 및 기준물에 대한 국소 게놈 편향 추정치(예를 들면, GC 밀도)를 기초로 편향 모델(예를 들면, 관계 및/또는 관계의 비교)을 구축하도록 지시한다. 하기 스크립트는 부분적으로 하나 이상의 프로세서가 (a) 검사 샘플의 서열 리드에 대한 (i) 구아닌 및 사이토신(GC) 밀도와 (ii) GC 밀도 빈도 사이의 관계를 생성하여 샘플 GC 밀도 관계를 생성하고, (b) 샘플 GC 밀도 관계를 기준 GC 밀도 관계와 비교함으로써 스크립트가 적절히 변경된 비교를 생성하고(이때, 기준 GC 밀도 관계는 기준물에 대한 (i) GC 밀도와 (ii) GC 밀도 빈도 사이의 관계임), (c) (b)에서 확인된 비교에 따라 샘플에 대한 서열 리드의 카운트를 정규화하도록(이때, 샘플에 대한 서열 리드에서 편향이 감소됨) 지시한다. 일부 실시양태에서, 편향 밀도 모듈, 관계 모듈, 편향 보정 모듈 및/또는 작도 모듈은 하기 나타낸 스크립트의 일부, 전부, 또는 일부 또는 전부의 변경물을 포함한다:

```
gcpath <- "I:/ghannum/Projects/Binless/Reference/BiasMaps/DnaseDensity_200_dist.txt"
inpath <- "I:/ghannum/Projects/Binless/Models/LDTv2_DNase-200"
outpath <- "I:/ghannum/Projects/Binless/Models/LDTv2_DNase-200"
makePlots <- TRUE
logTransform <- TRUE

args <- commandArgs(trailingOnly = TRUE)
gcpath <- args[1]
inpath <- args[2]
outpath <- args[3]
makePlots <- args[4]
logTransform <- as.logical(args[5])
paths <- dir(inpath)
paths <- paths[grep("_BiasDistr.txt$",paths)]

gcref <- scan(gcpath,0)
gcref <- gcref[gcref!=0]

if (logTransform) gcref <- log(gcref)
```

[0806]

```
from <- quantile(gcref,.005)
to <- quantile(gcref,.995)

x <- seq(from,to,length.out=100);

d1y <- predict(smooth.spline(density(gcref,from=from,to=to)),x)$y
if (!logTransform) d1y <- sapply(d1y,function(x){max(x,0)})

print(paste("Processing",length(paths),"models."))

for (f in paths)
{
        distr <- scan(paste(inpath,"/",f,sep=""),0)
        distr <- distr[distr!=0]
        if (logTransform) distr <- log(distr)
        d2y <- predict(smooth.spline(density(distr,from=from,to=to)),x)$y
        if (!logTransform) d2y <- sapply(d2y,function(x){max(x,0)})
        pp <- log2(d1y / d2y)
        pp[pp > 2] <- 2; pp[pp < -2] <- -2
        ok <- !is.na(pp)
        mod <- lm(pp[ok]~x+I(x^2)+I(x^3), data=list(x=x[ok]), w=d2y[ok])
        w <- 2^predict(mod,list(x=distr))
        fname <- substr(f,1,nchar(f)-14)
        out <- c(mod$coefficients,mean(w))
        out[out==Inf] <- "Infinity"
        out[out==-Inf] <- "-Infinity"
        write.table(matrix(c(out,as.integer(logTransform)),ncol=1),file=paste(outpath,"/",fnam
e,"_BiasMod.txt",sep=""),row.names=F,col.names=F,quote=F)

        if (makePlots)
        {

        png(units="in",height=4,width=4,res=300,file=paste(outpath,"/",fname,"_BiasMod.pn
g",sep=""))
                if (logTransform)
                {
```

[0807]

```
                plot(x[ok],pp[ok],ylim=c(-4,4),xlab="Bias Density",ylab="Log2 Ratio
(Reference / Sample)")
            }else plot(x[ok],pp[ok],ylim=c(-4,4),xlab="Log-Bias Density",ylab="Log2
Ratio (Reference / Sample)")
            abline(h=0,lty=2)
            lines(x[ok],predict(mod),col=3)
            dev.off()
        }
}

##Demo transformation
#load("I:/ghannum/Projects/Binless/2012_11_13_cewi_PERUN_19FCs_AltGCbias_chrFracti
ons.RData")
#
#d <- dir("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/")
#d <- d[grep("_GCDistr.txt",d)]
#v <- as.numeric(df.cewi.GCbiasTable[,"gcBiasRobust"])[1:length(d)]
#
#a <-
scan(paste("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/",d[which.min(v)],sep=""),0);
a <- sort(a)
#b <-
scan(paste("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/",d[which.max(v)],sep=""),0);
d <- sort(d)
#
#r <- scan("I:/ghannum/Projects/Binless/Reference/GCDensity_200_density.txt",0)
#
#plot(density(r),ylim=c(0,1e10),xlab="GC Density"); lines(density(a),col=3);
lines(density(b),col=2)
#
#a <- a[a!=0]
#b <- b[b!=0]
#r <- r[r!=0]
#
#plot(density(r),ylim=c(0,1e10),xlab="GC Density"); lines(density(a),col=3);
lines(density(b),col=2)
#
```

[0808]

```
#modA <-
as.numeric(scan(paste("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/",substr(d[which.
min(v)],1,nchar(d[which.min(v)])-12),"_GCMod.txt",sep=""),""))
#modB <-
as.numeric(scan(paste("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/",substr(d[which.
max(v)],1,nchar(d[which.max(v)])-12),"_GCMod.txt",sep=""),""))
#
#wa <- sapply(a,function(x){2^sum(c(1,x,x^2,x^3)*modA[1:4])})
#wb <- sapply(b,function(x){2^sum(c(1,x,x^2,x^3)*modB[1:4])})
#
#wa <- wa/(length(wa)*modA[5])
#wb <- wb/(length(wb)*modB[5])
#
#plot(density(r),ylim=c(0,1e10),xlab="GC Density"); lines(density(a,weights=wa),col=3);
lines(density(b,weights=wb),col=2)
```

[0809]

[0810] 실시예 8: 실시양태의 예

[0811] 하기 실시예는 일부 실시양태들을 예시하고 기술을 한정하지 않는다.

[0812] A1. 거짓 음성 및 거짓 양성 확인을 감소시키면서 태아에서의 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 방법으로서,

[0813] (a) 기준 게놈의 부분들에 맵핑된 핵산 서열 리드의 카운트를 수득하는 단계로서, 이때 상기 서열 리드가 태아를 가진 임신 여성으로부터의 순환 무세포 핵산의 리드인 단계;

[0814] (b) 각각의 부분들에 맵핑된 카운트를 정규화하여 계산된 게놈 구획 수준을 제공하는 단계;

[0815] (c) 계산된 게놈 구획 수준에 따라 게놈의 분절에 대한 프로파일을 생성하는 단계;

[0816] (d) 상기 프로파일을 분할하여 2개 이상의 분해 렌더링을 제공하는 단계; 및

[0817] (e) 상기 2개 이상의 분해 렌더링에 따라 거짓 음성 및 거짓 양성 확인을 감소시키면서 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 단계

[0818] 를 포함하는 방법.

[0819] A2. 실시양태 A1에 있어서, 분할이 역치화를 포함하는 것인 방법.

[0820] A3. 실시양태 A2에서 있어서, 역치화가 소프트 역치화를 포함하는 것인 방법.

[0821] A4. 실시양태 A2 또는 A3에 있어서, 역치화가 폴리시(policy)를 포함하는 것인 방법.

[0822] A5. 실시양태 A4에 있어서, 폴리시가 보편적인 것인 방법.

[0823] A6. 실시양태 A4에 있어서, 폴리시가 확실한 것인 방법.

[0824] A7. 실시양태 A2 내지 A6 중 어느 한 실시양태에 있어서, 역치화가 웨이브쓰레쉬에 의해 수행되는 것인 방법.

[0825] A8. 실시양태 A1 내지 A7 중 어느 한 실시양태에 있어서, 분할이 평준화를 포함하는 것인 방법.

[0826] A9. 실시양태 A8에 있어서, 평준화가 태아 분획에 따라 수행되는 것인 방법.

[0827] A10. 실시양태 A8 또는 A9에 있어서, 평준화가 커버리지에 따라 수행되는 것인 방법.

[0828] A11. 실시양태 A8 내지 A10 중 어느 한 실시양태에 있어서, 평준화가 검출할 최소 분절 길이에 따라 수행되는 것인 방법.

[0829] A12. 실시양태 A8 내지 A11 중 어느 한 실시양태에 있어서, 역치화 및 평준화가 수행되고 역치화가 평준화 전에 수행되는 것인 방법.

[0830] A13. 실시양태 A1 내지 A12 중 어느 한 실시양태에 있어서, (d)에서의 분할이 2종 이상의 상이한 분해 발생 프로세스에 따라 수행되는 것인 방법.

[0831] A14. 실시양태 A13에 있어서, 2종 이상의 상이한 분해 발생 프로세스 각각이 Haar 웨이블렛 분할, 서큘러 이진 분할, 최대 엔트로피 분할, 에지 검출 커넬을 이용한 컨볼루션, 젠슨 샤논 발산, 이진 반복 분할 및 푸리에 변환으로부터 독립적으로 선택되는 것인 방법.

[0832] A15. 실시양태 A13 또는 A14에 있어서, 2종 이상의 상이한 분해 발생 프로세스 중 하나가 서큘러 이진 분할인 방법.

[0833] A16. 실시양태 A13 내지 A15 중 어느 한 실시양태에 있어서, 2종 이상의 상이한 분해 발생 프로세스 중 하나가 Haar 웨이블렛인 방법.

[0834] A17. 실시양태 A13 내지 A16 중 어느 한 실시양태에 있어서, (d)에서의 분할이 Haar 웨이블렛 및 서큘러 이진 분할을 포함하는 것인 방법.

[0835] A18. 실시양태 A13 내지 A17 중 어느 한 실시양태에 있어서, 2종 이상의 분해 발생 프로세스가 동시에 적용되는 것인 방법.

[0836] A19. 실시양태 A13 내지 A17 중 어느 한 실시양태에 있어서, 2종 이상의 분해 발생 프로세스가 연속적으로 적용되는 것인 방법.

[0837] A20. 실시양태 A1 내지 A19 중 어느 한 실시양태에 있어서, 2개 이상의 분해 렌더링 중 하나 이상을 폴리싱하여

하나 이상의 폴리싱된 분해 렌더링을 제공하는 단계를 포함하는 방법.

[0838] A21. 실시양태 A20에 있어서, 폴리싱이 분해 렌더링에서 인접한 단편화된 수준들을 병합하는 단계를 포함하는 것인 방법.

[0839] A22. 실시양태 A20 또는 A21에 있어서, 인접한 단편화된 수준들이 그들의 게놈 구획 수준들에 따라 병합되는 것인 방법.

[0840] A23. 실시양태 A1 내지 A22 중 어느 한 실시양태에 있어서, 2개 이상의 분해 렌더링 중 하나 이상에서 후보 분절을 확인하는 단계를 포함하는 방법.

[0841] A23.1. 실시양태 A23에 있어서, 후보 분절이 하나 이상의 폴리싱된 분해 렌더링에서 확인되는 것인 방법.

[0842] A24. 실시양태 A23 또는 A23.1에 있어서, 후보 분절의 에지를 확인하는 단계를 포함하는 방법.

[0843] A25. 실시양태 A23 또는 A24에 있어서, 후보 분절의 수준을 측정하는 단계를 포함하는 방법.

[0844] A26. 실시양태 A23 또는 A25에 있어서, 후보 분절이 널(null) 프로파일에 따라 확인되는 것인 방법.

[0845] A27. 실시양태 A1 내지 A26 중 어느 한 실시양태에 있어서, (a)에서의 카운트가 임신 여성으로부터 수득된 샘플로부터 수득되는 것인 방법.

[0846] A28. 실시양태 A27에 있어서, 널 프로파일이 샘플로부터 생성되는 것인 방법.

[0847] A29. 실시양태 A26 또는 A27에 있어서, 널 프로파일이 기준 샘플로부터 생성되는 것인 방법.

[0848] A30. 실시양태 A23 내지 A29 중 어느 한 실시양태에 있어서, 후보 분절이 곡선하면적(AUC) 분석에 따라 확인되는 것인 방법.

[0849] A31. 실시양태 A23 내지 A30 중 어느 한 실시양태에 있어서, 2개 이상의 후보 분절들을 비교하여 비교를 제공하는 단계를 포함하는 방법.

[0850] A32. 실시양태 A31에 있어서, 제1 후보 분절이 제1 분해 렌더링으로부터의 후보 분절이고 제2 후보 분절이 제2 분해 렌더링으로부터의 후보 분절인 방법.

[0851] A33. 실시양태 A31 또는 A32에 있어서, 2개 이상의 후보 분절들이 비교에 따라 실질적으로 동일한 것으로서 확인되는 것인 방법.

[0852] A33.1. 실시양태 A31 또는 A32에 있어서, 2개 이상의 후보 분절들이 비교에 따라 상이한 것으로서 확인되는 것인 방법.

[0853] A33.2. 실시양태 A31 내지 A33.1 중 어느 한 실시양태에 있어서, 염색체 이배수성의 존재 또는 부재가 비교에 따라 확인되는 것인 방법.

[0854] A34. 실시양태 A31 내지 A33.2 중 어느 한 실시양태에 있어서, 비교가 2개 이상의 후보 분절들을 중첩시키는 단계를 포함하는 것인 방법.

[0855] A34.1. 실시양태 A31 또는 A34에 있어서, 비교에 따라 복합 후보 분절의 존재 또는 부재를 확인하는 단계를 포함하는 방법.

[0856] A35. 실시양태 34 또는 34.1에 있어서, 제1 후보 분절이 제2 후보 분절과 실질적으로 중첩되고 복합 후보 분절의 존재가 확인되는 것인 방법.

[0857] A35.1. 실시양태 34 또는 34.1에 있어서, 제1 후보 분절이 제2 후보 분절과 실질적으로 중첩되지 않고 복합 후보 분절의 부재가 확인되는 것인 방법.

[0858] A36. 실시양태 A34.1 내지 A35.1 중 어느 한 실시양태에 있어서, 염색체 이배수성의 존재 또는 부재가 복합 후보 분절의 존재 또는 부재에 따라 (e)에서 확인되는 것인 방법.

[0859] A37. 실시양태 A23 내지 A36 중 어느 한 실시양태에 있어서, 분해 렌더링에서 확인된 후보 분절을 검증하여 검증된 후보 분절을 제공하는 단계를 포함하는 방법.

[0860] A38. 실시양태 A37에 있어서, 검증이 슬라이딩 에지 프로세스를 수행하는 단계를 포함하는 것인 방법.

[0861] A39. 실시양태 A37 또는 A38에 있어서, 검증이 리브 원 아웃 프로세스를 수행하는 단계를 포함하는 것인 방법.

[0862] A40. 실시양태 A39에 있어서, 검증이 슬라이딩 에지 프로세스 및 리브 원 아웃 프로세스를 수행하는 단계를 포함하는 것인 방법.

[0863] A41. 실시양태 A37 내지 A40 중 어느 한 실시양태에 있어서, 검증이 후보 분절에 대한 유의 수준을 생성하는 단계를 포함하는 것인 방법.

[0864] A42. 실시양태 A37 내지 A41 중 어느 한 실시양태에 있어서, 검증이 복합 후보 분절에 대한 유의 수준을 생성하는 단계를 포함하는 것인 방법.

[0865] A43. 실시양태 A41 또는 A42에 있어서, 유의 수준이 Z-점수인 방법.

[0866] A44. 실시양태 A41 내지 A43 중 어느 한 실시양태에 있어서, 불확실성 수준이 유의 수준과 관련되어 있는 것인 방법.

[0867] A45. 실시양태 A44에 있어서, 검증된 후보 분절의 존재 또는 부재가 후보 분절의 유의 수준 및 불확실성 수준에 따라 확인되는 것인 방법.

[0868] A46. 실시양태 A44 또는 A45에 있어서, 염색체 이배수성, 미세중복 또는 미세결실의 존재가 유의 수준 및 불확실성 수준에 따라 확인되고, 이때 유의 수준 및 불확실성 수준 둘 다가 복합 후보 분절에 대해 생성되는 것인 방법.

[0869] A47. 실시양태 A46에 있어서, 염색체 이배수성, 미세중복 또는 미세결실의 존재가 Z-점수, 및 Z-점수와 관련된 불확실성 수준에 따라 확인되고, 이때 Z-점수 및 불확실성 수준 둘 다가 복합 후보 분절에 대해 생성되는 것인 방법.

[0870] A47.1. 실시양태 A47에 있어서, Z-점수가 약 3.95 이상의 절대 값을 갖는 것인 방법.

[0871] A48. 실시양태 A1 내지 A47.1 중 어느 한 실시양태에 있어서, 염색체 이배수성의 존재 또는 부재가 확인되는 것인 방법.

[0872] A48.1. 실시양태 A1 내지 A48 중 어느 한 실시양태에 있어서, 염색체 이배수성이 삼염색체성인 방법.

[0873] A48.2. 실시양태 A1 내지 A48.1 중 어느 한 실시양태에 있어서, 염색체 이배수성이 일염색체성인 방법.

[0874] A49. 실시양태 A1 내지 A48.3 중 어느 한 실시양태에 있어서, 미세중복의 존재 또는 부재가 확인되는 것인 방법.

[0875] A50. 실시양태 A1 내지 A48.3 중 어느 한 실시양태에 있어서, 미세결실의 존재 또는 부재가 확인되는 것인 방법.

[0876] A51. 실시양태 A1 내지 A50 중 어느 한 실시양태에 있어서, 디조지 증후군을 표시하는 미세결실의 존재 또는 부재가 확인되는 것인 방법.

[0877] A52. 실시양태 A1 내지 A51 중 어느 한 실시양태에 있어서, (a), (b), (c), (d) 및 (e) 중 하나 이상 또는 전부가 프로세서에 의해 수행되는 것인 방법.

[0878] A53. 실시양태 A52에 있어서, 프로세서가 마이크로프로세서인 방법.

[0879] A54. 실시양태 A1 내지 A53 중 어느 한 실시양태에 있어서, (a), (b), (c), (d) 및 (e) 중 하나 이상 또는 전부가 컴퓨터에 의해 수행되는 것인 방법.

[0880] A55. 실시양태 A1 내지 A54 중 어느 한 실시양태에 있어서, (a), (b), (c), (d) 및 (e) 중 하나 이상 또는 전부가 메모리(memory)와 함께 수행되는 것인 방법.

[0881] A56. 실시양태 A1 내지 A55 중 어느 한 실시양태에 있어서, (a), (b), (c), (d) 및 (e) 중 하나 이상 또는 전부가 마이크로프로세서 제어 장치에 의해 수행되는 것인 방법.

[0882] A57. 실시양태 A1 내지 A56 중 어느 한 실시양태에 있어서, (a) 전에, 임신 여성으로부터 수득된 샘플 중의 핵산을 서열결정하여 핵산 서열결정 리드를 제공하는 단계를 포함하는 방법.

[0883] A58. 실시양태 A1 내지 A57 중 어느 한 실시양태에 있어서, (a) 전에 핵산 서열 리드를 기준 게놈의 부분 또는

전체 기준 게놈으로 맵핑하는 단계를 포함하는 방법.

[0884] B1. 거짓 음성 및 거짓 양성 확인을 감소시키면서 후보 분절의 존재 또는 부재를 확인하는 방법으로서,

[0885] (a) 기준 게놈의 부분들에 맵핑된 핵산 서열 리드의 카운트를 수득하는 단계로서, 이때 서열 리드가 임신 여성으로부터의 순환 무세포 핵산의 리드인 단계;

[0886] (b) 각각의 부분들에 맵핑된 카운트를 정규화하여 계산된 게놈 구획 수준을 제공하는 단계;

[0887] (c) 부분 세트를 다수의 부분 서브세트로 분할하는 단계;

[0888] (d) 계산된 게놈 구획 수준에 따라 각각의 서브세트에 대한 수준을 측정하는 단계;

[0889] (e) 각각의 수준에 대한 유의 수준을 측정하는 단계; 및

[0890] (f) 각각의 수준에 대해 측정된 유의 수준에 따라 거짓 음성 및 거짓 양성 확인을 감소시키면서 후보 분절의 존재 또는 부재를 확인하는 단계

[0891] 를 포함하는 방법.

[0892] B1.1 실시양태 B1에 있어서, 부분 세트가 후보 분절을 포함하는 것으로 의심되는 것인 방법.

[0893] B2. 실시양태 B1 또는 B1.1에 있어서, 모든 부분 서브세트들에 대한 각각의 수준에 대해 측정된 유의 수준에 따라 중간 유의 수준을 생성하는 단계를 포함하는 방법.

[0894] B3. 실시양태 B1 또는 B2에 있어서, 모든 부분 서브세트들에 대한 각각의 수준에 대해 측정된 유의 수준의 분포를 생성하는 단계를 포함하는 방법.

[0895] B4. 실시양태 B1 내지 B3 중 어느 한 실시양태에 있어서, 모든 부분 서브세트들에 대한 모든 수준에 대해 측정된 유의 수준에 따라 불확실성 값을 생성하는 단계를 포함하는 방법.

[0896] B5. 실시양태 B4에 있어서, (f)에서의 확인이 중간 유의 수준 및 불확실성 값에 따른 것인 방법.

[0897] B6. 실시양태 B3 또는 B4에 있어서, (f)에서의 확인이 중간 유의 수준 및 유의 수준의 분포에 따른 것인 방법.

[0898] B6.1. 실시양태 B1 내지 B6 중 어느 한 실시양태에 있어서, (f)에서의 확인이 유의 수준에 대해 측정된 예정된 범위에 따른 것인 방법.

[0899] B6.2. 실시양태 B6.1에 있어서, 실시양태 B3에서 생성된 유의 수준의 분포의 75% 이상이 유의 수준에 대한 예정된 범위를 벗어날 때 후보 분절의 존재가 확인되는 것인 방법.

[0900] B6.3. 실시양태 B6.1 또는 B6.2에 있어서, 실시양태 B4에서 생성된 불확실성 값의 75% 이상이 유의 수준에 대한 예정된 범위를 벗어날 때 후보 분절의 존재가 확인되는 것인 방법.

[0901] B7. 실시양태 B1 내지 B6.3 중 어느 한 실시양태에 있어서, 유의 수준이 Z-점수인 방법.

[0902] B7.1. 실시양태 B7에 있어서, 예정된 범위가 약 3 내지 약 -3의 Z-점수인 방법.

[0903] B8. 실시양태 B4 내지 B7.1 중 어느 한 실시양태에 있어서, 불확실성 값이 중간 절대 편차인 방법.

[0904] B9. 실시양태 B1 내지 B8 중 어느 한 실시양태에 있어서, 부분 세트가 제1 말단 및 제2 말단을 포함하고, (c)에서의 분할이

[0905] (i) 반복적 제거로 하나 이상의 부분을 부분 세트의 제1 말단으로부터 제거하여 반복적 제거마다 부분 서브세트를 제공하는 단계;

[0906] (ii) n회의 반복 후 (i)에서의 반복적 제거를 종결하여 n + 1개의 부분 서브세트를 제공하는 단계로서, 이때 상기 부분 세트가 서브세트이고 각각의 서브세트가 상이한 수의 부분, 제1 서브세트 말단 및 제2 서브세트 말단을 포함하는 것인 단계,

[0907] (iii) 반복적 제거로 하나 이상의 부분을 (ii)에서 제공된 n + 1개의 부분 서브세트 각각의 제2 서브세트 말단으로부터 제거하는 단계; 및

[0908] (iv) n회의 반복 후 (iii)에서의 반복적 제거를 종결하여 다수의 부분 서브세트를 제공하는 단계

[0909] 를 포함하는 것인 방법.

[0910] B10. 실시양태 B9에 있어서, 다수의 서브세트가 (n + 1)2개의 서브세트와 동등한 것인 방법.

[0911] B11. 실시양태 B9 또는 B10에 있어서, n이 5 내지 30의 정수와 동등한 것인 방법.

[0912] B12. 실시양태 B9 내지 B11 중 어느 한 실시양태에 있어서, n이 15와 동등한 것인 방법.

[0913] B13. 실시양태 B1 내지 B12 중 어느 한 실시양태에 있어서, 부분 세트가 염색체에 존재하는 것인 방법.

[0914] B14. 실시양태 B13에 있어서, 부분 세트가 공지된 유전적 변이 또는 공지된 유전 장애와 관련된 영역을 포함하는 것인 방법.

[0915] B14. 실시양태 B13 또는 B14에 있어서, 부분 세트가 디조지 영역을 포함하는 것인 방법.

[0916] B15. 실시양태 B1 내지 B14 중 어느 한 실시양태에 있어서, (a) 내지 (e)가 검사 샘플 및 2개 이상의 기준 샘플들에 대해 수행되는 것인 방법.

[0917] B16. 실시양태 B15에 있어서,

[0918] (i) (a) 전에 2개 이상의 기준 샘플들 중 하나를 제거하여 기준 샘플 서브세트를 제공하는 단계;

[0919] (ii) 기준 샘플 서브세트 각각에 대해 (a) 내지 (e)를 수행하는 단계;

[0920] (iii) 기준 샘플 서브세트 각각에 대한 중간 유의 수준을 실시양태 B2에 따라 생성하는 단계;

[0921] (iv) (iii)에서 생성된 중간치에 따라 복합 중간 유의 수준을 생성하는 단계; 및

[0922] (v) (iv)에서의 복합 중간 유의 수준에 대한 복합 불확실성 수준을 생성하는 단계

[0923] 를 포함하고, 이때 (f)에서의 확인이 복합 중간 유의 수준 및 복합 불확실성 수준에 따른 것인 방법.

[0924] B17. 실시양태 B16에 있어서, 기준 샘플 서브세트 각각이 상이한 기준 샘플 세트를 포함하는 것인 방법.

[0925] B18. 실시양태 B16 또는 B17에 있어서, 제거된 2개 이상의 기준 샘플들 중 각각 하나가 서브세트들 중 하나로부터만 제거되는 것인 방법.

[0926] B19. 실시양태 B1 내지 B18 중 어느 한 실시양태에 있어서, (a), (b), (c), (d), (e) 및 (f) 중 하나 이상 또는 전부가 프로세서에 의해 수행되는 것인 방법.

[0927] B20. 실시양태 B19에 있어서, 프로세서가 마이크로프로세서인 방법.

[0928] B21. 실시양태 B1 내지 B20 중 어느 한 실시양태에 있어서, (a), (b), (c), (d), (e) 및 (f) 중 하나 이상 또는 전부가 컴퓨터에 의해 수행되는 것인 방법.

[0929] B22. 실시양태 B1 내지 B21 중 어느 한 실시양태에 있어서, (a), (b), (c), (d), (e) 및 (f) 중 하나 이상 또는 전부가 메모리와 함께 수행되는 것인 방법.

[0930] B23. 실시양태 B1 내지 B22 중 어느 한 실시양태에 있어서, (a), (b), (c), (d), (e) 및 (f) 중 하나 이상 또는 전부가 마이크로프로세서 제어 장치에 의해 수행되는 것인 방법.

[0931] B24. 실시양태 B1 내지 B23 중 어느 한 실시양태에 있어서, (a) 전에, 임신 여성으로부터 수득된 샘플에서 핵산을 서열결정하여 핵산 서열결정 리드를 제공하는 단계를 포함하는 방법.

[0932] B25. 실시양태 B1 내지 B24 중 어느 한 실시양태에 있어서, (a) 전에 핵산 서열 리드를 기준 게놈의 부분 또는 전체 기준 게놈으로 맵핑하는 단계를 포함하는 방법.

[0933] C1. 거짓 음성 및 거짓 양성 확인을 감소시키면서 태아에서의 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 방법으로서,

[0934] (a) 기준 게놈의 부분들에 맵핑된 핵산 서열 리드의 카운트를 수득하는 단계로서, 이때 상기 서열 리드가 임신 여성으로부터의 순환 무세포 핵산의 리드인 단계;

[0935] (b) 각각의 부분들에 맵핑된 카운트를 정규화하여 계산된 게놈 구획 수준을 제공하는 단계;

[0936] (c) 게놈의 분절을 선택하여 부분 세트를 제공하는 단계;

[0937] (d) 부분 세트를 반복적으로 분할하여 2개 이상의 부분 서브세트를 제공하는 단계;

[0938] (e) 2개 이상의 부분 서브세트 각각에 대한 수준을 측정하는 단계; 및

[0939] (f) (e)에서 측정된 수준에 따라 거짓 음성 및 거짓 양성 확인을 감소시키면서 샘플에 대해 태아에서의 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 단계

[0940] 를 포함하는 방법.

[0941] C2. 실시양태 C1에 있어서, (e)에서 확인된 2개 이상의 부분 서브세트 각각에 대한 수준이 유의하게 상이한지를 확인하는 단계를 포함하는 방법.

[0942] C3. 실시양태 C1 또는 C2에 있어서, (d)에서의 분할이 반복적 분할을 포함하는 것인 방법.

[0943] C4. 실시양태 C3에 있어서, 반복적 분할이 이진 반복적 분할을 포함하는 것인 방법.

[0944] C5. 실시양태 C3에 있어서, 반복적 분할이 최대 엔트로피-기초 분할을 포함하는 것인 방법.

[0945] C6. 실시양태 C2 내지 C5 중 어느 한 실시양태에 있어서, 제1 부분 서브세트의 수준이 제2 부분 서브세트의 수준과 유의하게 상이하게 상이할 때 제1 부분 서브세트 및 제2 부분 서브세트를 분할하는 단계를 포함하고, 이때 제1 부분 서브세트와 제2 부분 서브세트가 서로 인접하는 것인 방법.

[0946] C7. 실시양태 C2 내지 C6 중 어느 한 실시양태에 있어서, 제3 부분 서브세트 및 제4 부분 서브세트를 재연결하여 재연결된 부분 서브세트를 제공하는 단계를 포함하고, 이때 제3 부분 서브세트의 수준과 제4 부분 서브세트의 수준이 유의하게 상이하지 않고, 제3 부분 서브세트와 제4 부분 서브세트가 서로 인접하고, 재연결된 부분이 다시 분할되지 않는 것인 방법.

[0947] C8. 실시양태 C1 내지 C7 중 어느 한 실시양태에 있어서, (a), (b), (c), (d), (e) 및 (f) 중 하나 이상 또는 전부가 프로세서에 의해 수행되는 것인 방법.

[0948] C9. 실시양태 C8에 있어서, 프로세서가 마이크로프로세서인 방법.

[0949] C10. 실시양태 C1 내지 C9 중 어느 한 실시양태에 있어서, (a), (b), (c), (d), (e) 및 (f) 중 하나 이상 또는 전부가 컴퓨터에 의해 수행되는 것인 방법.

[0950] C11. 실시양태 C1 내지 C10 중 어느 한 실시양태에 있어서, (a), (b), (c), (d), (e) 및 (f) 중 하나 이상 또는 전부가 메모리와 함께 수행되는 것인 방법.

[0951] C12. 실시양태 C1 내지 C11 중 어느 한 실시양태에 있어서, (a), (b), (c), (d), (e) 및 (f) 중 하나 이상 또는 전부가 마이크로프로세서 제어 장치에 의해 수행되는 것인 방법.

[0952] C13. 실시양태 C1 내지 C12 중 어느 한 실시양태에 있어서, (a) 전에, 임신 여성으로부터 수득된 샘플 중의 핵산을 서열결정하여 핵산 서열결정 리드를 제공하는 단계를 포함하는 방법.

[0953] C14. 실시양태 C1 내지 C13 중 어느 한 실시양태에 있어서, (a) 전에 핵산 서열 리드를 기준 게놈의 부분 또는 전체 기준 게놈으로 맵핑하는 단계를 포함하는 방법.

[0954] D1. 태아에서의 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 방법으로서,

[0955] (a) 기준 게놈의 부분들에 맵핑된 핵산 서열 리드의 카운트를 정규화하여 정규화된 카운트를 제공하는 단계로서, 이때 상기 서열 리드가 태아를 가진 임신 여성으로부터의 검사 샘플로부터의 순환 무세포 핵산의 리드인 단계;

[0956] (b) 부분 서브세트에서 부분의 정규화된 카운트 또는 정규화된 카운트를 분할하여 하나 이상의 별개의 분절을 제공하는 단계;

[0957] (c) 하나 이상의 별개의 분절들 중에서 후보 분절을 확인하는 단계; 및

[0958] (d) 후보 분절에 따라 염색체 이배수성, 미세중복 또는 미세결실의 존재 또는 부재를 확인하는 단계

[0959] 를 포함하는 방법.

[0960] D2. 실시양태 D1에 있어서, 분할이 역치화를 포함하는 것인 방법.

[0961] D3. 실시양태 D1 또는 D2에 있어서, 분할이 평준화를 포함하는 것인 방법.

[0962] D4. 실시양태 D3에 있어서, 평준화가 태아 분획, 커버리지, 최소 분절 길이 또는 이들의 조합에 따라 수행되는

것인 방법.

[0963] D5. 실시양태 D1 내지 D4 중 어느 한 실시양태에 있어서, 역치화 및 평준화가 수행되고 역치화가 평준화 전에 수행되는 것인 방법.

[0964] D5.1. 실시양태 D1 내지 D5 중 어느 한 실시양태에 있어서, (b)에서의 분할이 부분의 정규화된 카운트에 대해 수행되는 것인 방법.

[0965] D5.2. 실시양태 D1 내지 D5 중 어느 한 실시양태에 있어서, (b)에서의 분할이 부분 서브세트에서의 정규화된 카운트에 대해 수행되는 것인 방법.

[0966] D5.3. 실시양태 D5.2에 있어서, 부분 서브세트가 염색체의 모든 부분 또는 염색체의 모든 부분 서브세트인 방법.

[0967] D5.4. 실시양태 D1 내지 D5.3 중 어느 한 실시양태에 있어서, 정규화된 카운트가 수준을 갖는 프로파일에서의 정규화된 카운트이고 상기 프로파일이 (b)에서 분할되는 것인 방법.

[0968] D5.5. 실시양태 D1 내지 D5.4 중 어느 한 실시양태에 있어서, 분할이 별개의 분절을 포함하는 분해 렌더링을 생성하는 것인 방법.

[0969] D5.6. 실시양태 D1 내지 D5.5 중 어느 한 실시양태에 있어서, (a)에서의 정규화가 구아닌 및 사이토신(GC) 편향의 LOESS 정규화(GC-LOESS 정규화)를 포함하는 것인 방법.

[0970] D5.7. 실시양태 D1 내지 D5.6 중 어느 한 실시양태에 있어서, (a)에서의 정규화가 주성분 정규화를 포함하는 것인 방법.

[0971] D5.8. 실시양태 D1 내지 D5.7 중 어느 한 실시양태에 있어서, (a)에서의 정규화가 GC-LOESS 정규화에 이어서 주성분 정규화를 포함하는 것인 방법.

[0972] D5.9. 실시양태 D1 내지 D5.8 중 어느 한 실시양태에 있어서, (a)에서의 정규화가

[0973] (1) (i) 각각의 부분들에 맵핑된 서열 리드의 카운트와 (ii) 각각의 부분들에 대한 GC 함량 사이의 피팅된 관계를 기초로 검사 샘플에 대한 구아닌 및 사이토신(GC) 편향 계수를 측정하는 단계로서, 이때 GC 편향 계수가 선형의 피팅된 관계에 대한 기울기 또는 비선형의 피팅된 관계에 대한 곡률 추정치인 단계; 및

[0974] (2) 각각의 부분들에 대한 (a)의 카운트, (b)의 GC 편향 계수, 및 (i) 다수의 샘플들 각각에 대한 GC 편향 계수와 (ii) 다수의 샘플들에 대한 각각의 부분들에 맵핑된 서열 리드의 카운트 사이의 피팅된 관계를 기초로 마이크로프로세서를 이용하여 각각의 부분들에 대한 게놈 구획 수준을 계산하여 계산된 게놈 구획 수준을 제공하는 단계

[0975] 를 포함하는 것인 방법.

[0976] D6. 실시양태 D1 내지 D5.9 중 어느 한 실시양태에 있어서, (b)에서의 분할이 2종 이상의 상이한 분할 프로세스의 적용을 포함하는 것인 방법.

[0977] D7. 실시양태 D6에 있어서, 2종 이상의 상이한 분할 프로세스 각각이 Haar 웨이블렛 분할, 서큘러 이진 분할, 최대 엔트로피 분할, 에지 검출 커넬을 이용한 컨볼루션, 젠슨 샤논 발산, 이진 반복 분할 및 푸리에 변환으로부터 독립적으로 선택되는 것인 방법.

[0978] D8. 실시양태 D6 또는 D7에 있어서, 2종 이상의 상이한 분할 프로세스 중 하나가 서큘러 이진 분할인 방법.

[0979] D9. 실시양태 D6 내지 D8 중 어느 한 실시양태에 있어서, 2종 이상의 상이한 분할 프로세스 중 하나가 Haar 웨이블렛 분할인 방법.

[0980] D10. 실시양태 D6 내지 D9 중 어느 한 실시양태에 있어서, (b)에서의 분할이 Haar 웨이블렛 분할 프로세스 및 서큘러 이진 분할 프로세스를 포함하는 것인 방법.

[0981] D11. 실시양태 D6 내지 D10 중 어느 한 실시양태에 있어서, 2종 이상의 분할 프로세스가 동시에 수행되는 것인 방법.

[0982] D12. 실시양태 D1 내지 D11 중 어느 한 실시양태에 있어서, 분할이 분해 렌더링에서 인접한 단편화된 수준들을 병합하는 단계를 포함하는 폴리싱 프로세스를 포함하는 것인 방법.

[0983] D13. 실시양태 D1 내지 D12 중 어느 한 실시양태에 있어서, 후보 분절의 하나 이상의 에지를 확인하는 단계를 포함하는 방법.

[0984] D14. 실시양태 D1 내지 D13 중 어느 한 실시양태에 있어서, 후보 분절에 의해 커버되는 부분들의 수를 측정하는 단계를 포함하는 방법.

[0985] D15. 실시양태 D1 내지 D14 중 어느 한 실시양태에 있어서, 후보 분절의 수준을 측정하는 단계를 포함하는 방법.

[0986] D15.1. 실시양태 D1 내지 D15 중 어느 한 실시양태에 있어서, 후보 분절이 곡선하면적(AUC) 분석에 따라 확인되는 것인 방법.

[0987] D16. 실시양태 D15.1에 있어서, AUC 분석이 후보 분절에 의해 커버되는 부분들의 수 및/또는 후보 분절에 대한 수준의 AUC 분석인 방법.

[0988] D16.1. 실시양태 D1 내지 D16 중 어느 한 실시양태에 있어서, 후보 분절을 검증하여 검증된 후보 분절을 제공하는 단계를 포함하는 방법.

[0989] D16.2. 실시양태 D16.1에 있어서, 검증이 슬라이딩 에지 프로세스를 수행하는 단계를 포함하는 것인 방법.

[0990] D16.3. 실시양태 D16.1 또는 D16.2에 있어서, 검증이 리브 원 아웃 프로세스를 수행하는 단계를 포함하는 것인 방법.

[0991] D16.4. 실시양태 D16.3에 있어서, 검증이 슬라이딩 에지 프로세스 및 리브 원 아웃 프로세스를 수행하는 단계를 포함하는 것인 방법.

[0992] D16.5. 실시양태 D16.1 내지 D16.4 중 어느 한 실시양태에 있어서, 검증이 후보 분절에 대한 유의 수준을 생성하는 단계를 포함하는 것인 방법.

[0993] D16.6. 실시양태 D16.1 내지 D16.5 중 어느 한 실시양태에 있어서, 검증이 복합 후보 분절에 대한 유의 수준을 생성하는 단계를 포함하는 것인 방법.

[0994] D16.7. 실시양태 D1 내지 D16.6 중 어느 한 실시양태에 있어서, 제1 분할로부터 제1 후보 분절을 확인하고 제1 분할과 상이한 제2 분할로부터 제2 후보 분절을 확인하는 단계를 포함하는 방법.

[0995] D16.8. 실시양태 D16.7에 있어서, 제1 후보 분절과 제2 후보 분절이 실질적으로 동일한지 아니면 실질적으로 상이한지를 확인하는 단계를 포함하는 방법.

[0996] D16.9. 실시양태 D16.7 또는 D16.8에 있어서, 제1 후보 분절과 제2 후보 분절이 실질적으로 상이할 때 미세결실 또는 미세중복의 부재를 확인하는 단계를 포함하는 방법.

[0997] D17. 실시양태 D1 내지 D16.9 중 어느 한 실시양태에 있어서, 후보 분절 또는 검증된 후보 분절의 정량치를 생성하는 단계를 포함하는 방법.

[0998] D18. 실시양태 D17에 있어서, 정량치가 후보 분절 또는 검증된 후보 분절에 대한 카운트 표시치인 방법.

[0999] D19. 실시양태 D18에 있어서, 정량치가 후보 분절 또는 검증된 후보 분절에 대한 카운트 표시치의 z-점수 정량치인 방법.

[1000] D20. 실시양태 D19에 있어서, 후보 분절 또는 검증된 후보 분절에 대한 z-점수가 (i) 검사 샘플 카운트 표시치로부터 (ii) 정배수체 카운트 표시치의 중간치를 뺀 뺄셈 결과를 (iii) 정배수체 카운트 표시치의 MAD로 나눈 값이고, 이때 (i) 검사 샘플 카운트 표시치가 검사 샘플에 대한 총 카운트를 총 상염색체 카운트로 나눈 비이고, (ii) 정배수체 중간 카운트 표시치가 정배수체 샘플에 대한 총 카운트를 총 상염색체 카운트로 나눈 비의 중간치인 방법.

[1001] D21. 실시양태 D17 내지 D20 중 어느 한 실시양태에 있어서, 후보 분절 또는 검증된 후보 분절이 위치하는 염색체의 염색체 표시치의 정량치를 생성하는 단계를 포함하는 방법.

[1002] D22. 실시양태 D21에 있어서, 염색체 표시치의 정량치가 z-점수 정량치인 방법.

[1003] D23. 실시양태 D22에 있어서, 염색체에 대한 z-점수가 (i) 검사 샘플 카운트 표시치로부터 (ii) 정배수체 카운트 표시치의 중간치를 뺀 뺄셈 결과를 (iii) 정배수체 카운트 표시치의 MAD로 나눈 값이고, 이때 (i) 검사 샘플

카운트 표시치는, 검사 샘플에 대한, 후보 분절이 위치하는 염색체의 총 카운트를 총 상염색체 카운트로 나눈 비이고, (ii) 정배수체 카운트 표시치의 중간치는, 정배수체 샘플에 대한, 후보 분절이 위치하는 염색체의 총 카운트를 총 상염색체 카운트로 나눈 비의 중간치인 방법.

[1004] D24. 실시양태 D17 내지 D23 중 어느 한 실시양태에 있어서, 후보 분절 또는 검증된 후보 분절의 정량치를 염색체 표시치의 정량치와 비교하는 것인 방법.

[1005] D25. 실시양태 D24에 있어서, 제1 후보 분절 또는 검증된 제1 후보 분절의 z-점수 정량치를 생성하고, 제2 후보 분절 또는 검증된 제2 후보 분절의 z-점수 정량치를 생성하고, 제1 후보 분절 및 제2 후보 분절을 2종의 상이한 분할로부터 확인하는 것인 방법.

[1006] D26. 실시양태 D25에 있어서, (i) 제1 후보 분절 또는 검증된 제1 후보 분절의 z-점수 정량치를 1 미만의 계수(factor)와 곱한 값 및 (ii) 제2 후보 분절 또는 검증된 제2 후보 분절의 z-점수 정량치를 상기 계수와 곱한 값의 최소치를 측정하는 단계를 포함하는 방법.

[1007] D27. 실시양태 D26에 있어서, 염색체 표시치의 z-점수 정량치가 최소치보다 더 작은지, 최소치보다 더 큰지 아니면 최소치와 동등한지를 확인하는 단계를 포함하는 방법.

[1008] D28. 실시양태 D25에 있어서, 염색체 표시치의 z-점수 정량치가 3.95의 값보다 더 작은지, 이 값보다 더 큰지 아니면 이 값과 동등한지를 확인하는 단계를 포함하는 방법.

[1009] D29. 실시양태 D28에 있어서, 검사 샘플에 대하여, (i) 염색체 표시치의 z-점수 정량치가 3.95의 값 이상이고 (ii) 염색체 표시치의 z-점수 정량치가 최소치 이상인 경우, 염색체 이배수성의 존재를 확인하는 단계를 포함하는 방법.

[1010] D30. 실시양태 28에 있어서, 검사 샘플에 대하여, (i) 염색체 표시치의 z-점수 정량치가 3.95의 값보다 더 작고/작거나 (ii) 염색체 표시치의 z-점수 정량치가 최소치보다 더 작은 경우, 염색체 이배수성의 부재를 확인하는 단계를 포함하는 방법.

[1011] D31. 실시양태 D29 또는 D30에 있어서, 염색체 이배수성이 삼염색체성 또는 일염색체성인 방법.

[1012] D32. 실시양태 D30에 있어서, 제1 후보 분절 또는 검증된 제1 후보 분절의 z-점수 정량치가 3.95의 값보다 더 작은지, 이 값보다 더 큰지 아니면 이 값과 동등한지를 확인하는 단계, 및 제2 후보 분절 또는 검증된 제2 후보 분절의 z-점수 정량치가 3.95의 값보다 더 작은지, 이 값보다 더 큰지 아니면 이 값과 동등한지를 확인하는 단계를 포함하는 방법.

[1013] D34. 실시양태 D32에 있어서, 제1 후보 분절 및 제2 후보 분절, 또는 이들의 검증된 분절이 실질적으로 동일한지를 확인하는 단계를 포함하는 방법.

[1014] D35. 실시양태 D34에 있어서, 검사 샘플에 대하여, (i) 제1 후보 분절 또는 검증된 제1 후보 분절의 z-점수 정량치가 3.95의 값 이상이고 제2 후보 분절 또는 검증된 제2 후보 분절의 z-점수 정량치가 3.95의 값 이상이고 (ii) 제1 후보 분절 및 제2 후보 분절, 또는 이들의 검증된 분절이 실질적으로 동일한 경우, 미세결실 또는 미세삽입의 존재를 확인하는 단계를 포함하는 방법.

[1015] D36. 실시양태 D34에 있어서, 검사 샘플에 대하여, (i) 제1 후보 분절 또는 검증된 제1 후보 분절의 z-점수 정량치가 3.95의 값보다 더 작고/작거나 제2 후보 분절 또는 검증된 제2 후보 분절의 z-점수 정량치가 3.95의 값보다 더 작고/작거나, (ii) 제1 후보 분절 및 제2 후보 분절, 또는 이들의 검증된 분절이 실질적으로 동일하지 않은 경우, 미세결실 또는 미세삽입의 부재를 확인하는 단계를 포함하는 방법.

[1016] D37. 실시양태 D17 내지 D23 중 어느 한 실시양태에 있어서, 후보 분절 또는 검증된 후보 분절에 대한 카운트 표시치의 z-점수 정량치를 측정하는 단계 및 이 정량치가 3.95의 값보다 더 작은지, 이 값보다 더 큰지 아니면 이 값과 동등한지를 확인하는 단계를 포함하는 방법.

[1017] D37.1. 실시양태 D17 내지 D23 중 어느 한 실시양태에 있어서, 염색체 표시치의 z-점수 정량치를 측정하는 단계 및 이 정량치가 3.95의 값보다 더 작은지, 이 값보다 더 큰지 아니면 이 값과 동등한지를 확인하는 단계를 포함하는 방법.

[1018] D38. 실시양태 D37 및/또는 D37.1에 있어서, 로그 승산비(LOR)를 계산하는 단계를 포함하며, 상기 LOR은 (i) (1) 유전적 변이를 가질 조건부 확률과 (2) 유전적 변이를 가질 선험적 확률의 제1 곱셈 결과를 (ii) (1) 유전

적 변이를 갖지 않을 조건부 확률과 (2) 유전적 변이를 갖지 않을 선험적 확률의 제2 곱셈 결과로 나눈 몫의 로그인 방법.

[1019] D39. 실시양태 D38에 있어서, 유전적 변이를 가질 조건부 확률이 검사 샘플에 대해 측정된 태아 분획, 검사 샘플에 대해 측정된 분절에 대한 카운트 표시치의 z-점수 및 분절에 대한 카운트 표시치에 대한 z-점수의 태아 분획에 대한 분포에 따라 결정되는 것인 방법.

[1020] D39.1. 실시양태 D39에 있어서, 유전적 변이를 가질 조건부 확률이 하기 방정식 23의 관계에 의해 결정되는 것인 방법:

[1021] 방정식 23

[1022] $$Z \sim Normal\left(\frac{\mu_X}{\sigma_X}\frac{f}{2}, 1\right)$$

[1023] 상기 식에서,

[1024] f는 태아 분획, X는 유전적 변이를 커버하는 분절에 대한 합산된 부분 카운트, $X \sim f(\mu_X, \sigma_X)$이고, 이때 μ_X 및 σ_X는 각각 X의 평균 및 표준 편차이고, $f(\cdot)$는 분포 함수이다.

[1025] D40. 실시양태 D39 또는 D39.1에 있어서, 유전적 변이를 가질 조건부 확률이 분절에 대한 카운트 표시치의 검사 샘플에 대한 Z-점수와 분절에 대한 카운트 표시치에 대한 z-점수의 태아 분획에 대한 분포 사이의 교차점인 방법.

[1026] D41. 실시양태 D38에 있어서, 유전적 변이를 갖지 않을 조건부 확률이 검사 샘플에 대해 측정된 분절에 대한 카운트 표시치의 z-점수와 정배수체에서의 분절에 대한 카운트 표시치에 대한 z-점수의 분포 사이의 교차점인 방법.

[1027] D42. 실시양태 D38 내지 D41 중 어느 한 실시양태에 있어서, 유전적 변이를 가질 선험적 확률 및 유전적 변이를 갖지 않을 선험적 확률이 검사 대상체를 포함하지 않는 다수의 샘플들로부터 측정되는 것인 방법.

[1028] D43. 실시양태 D38 내지 D42 중 어느 한 실시양태에 있어서, LOR이 0보다 더 큰지 아니면 0보다 더 작은지를 확인하는 단계를 포함하는 방법.

[1029] D44. 실시양태 D37 내지 D43 중 어느 한 실시양태에 있어서, 검사 샘플에 대하여, (i) 염색체 표시치의 z-점수 정량치가 3.95의 값 이상이고 (ii) LOR이 0보다 더 큰 경우, 염색체 이배수성의 존재를 확인하는 단계를 포함하는 방법.

[1030] D45. 실시양태 D37 내지 D43 중 어느 한 실시양태에 있어서, 검사 샘플에 대하여, (i) 염색체 표시치의 z-점수 정량치가 3.95의 값보다 더 작고/작거나, (ii) LOR이 0보다 더 작은 경우, 염색체 이배수성의 부재를 확인하는 단계를 포함하는 방법.

[1031] D46. 실시양태 D44 또는 D45에 있어서, 염색체 이배수성이 삼염색체성 또는 일염색체성인 방법.

[1032] D47. 실시양태 D37 내지 D43 중 어느 한 실시양태에 있어서, 검사 샘플에 대하여, (i) 후보 분절 또는 검증된 후보 분절에 대한 카운트 표시치의 z-점수 정량치가 3.95의 값 이상이고 (ii) LOR이 0보다 더 큰 경우, 미세결실 또는 미세중복의 존재를 확인하는 단계를 포함하는 방법.

[1033] D48. 실시양태 D37 내지 D43 중 어느 한 실시양태에 있어서, 검사 샘플에 대하여, (i) 후보 분절 또는 검증된 후보 분절에 대한 카운트 표시치의 z-점수 정량치가 3.95의 값보다 더 작고/작거나, (ii) LOR이 0보다 더 작은 경우, 미세결실 또는 미세중복의 부재를 확인하는 단계를 포함하는 방법.

[1034] D49. 실시양태 D47 또는 D48에 있어서, 미세결실이 디조지 증후군과 관련되어 있는 것인 방법.

[1035] D49.1. 실시양태 D1 내지 D49 중 어느 한 실시양태에 있어서, 카운트 표시치가 정규화된 카운트 표시치인 방법.

[1036] D50. 실시양태 D1 내지 D49.1 중 어느 한 실시양태에 있어서, (a), (b), (c) 및 (d) 중 하나 이상 또는 전부가 시스템 내의 마이크로프로세서에 의해 수행되는 것인 방법.

[1037] D51. 실시양태 D1 내지 D50 중 어느 한 실시양태에 있어서, (a), (b), (c) 및 (d) 중 하나 이상 또는 전부가 컴

퓨터에 의해 수행되는 것인 방법.

[1038] D52. 실시양태 D1 내지 D51 중 어느 한 실시양태에 있어서, (a), (b), (c) 및 (d) 중 하나 이상 또는 전부가 메모리와 함께 수행되는 것인 방법.

[1039] D53. 실시양태 D1 내지 D52 중 어느 한 실시양태에 있어서, (a) 전에, 임신 여성으로부터 수득된 샘플 중의 핵산을 서열결정하여 핵산 서열 리드를 제공하는 단계를 포함하는 방법.

[1040] D54. 실시양태 D1 내지 D53 중 어느 한 실시양태에 있어서, (a) 전에 핵산 서열 리드를 기준 게놈의 부분들로 맵핑하는 단계를 포함하는 방법.

[1041] E1. 태아에서의 염색체 이배수성의 존재 또는 부재를 확인하는 방법으로서,

[1042] (a) 기준 게놈의 부분들에 맵핑된 핵산 서열 리드의 카운트에 따라 염색체 카운트 표시치를 측정하는 단계로서, 이때 상기 서열 리드가 태아를 가진 임신 여성으로부터의 검사 샘플에 대한 순환 무세포 핵산의 리드인 단계;

[1043] (b) 검사 샘플에 대한 태아 분획을 측정하는 단계;

[1044] (c) 로그 승산비(LOR)를 계산하는 단계로서, 상기 LOR은 (i) (1) 염색체 이배수성을 가질 조건부 확률과 (2) 염색체 이배수성을 가질 선험적 확률의 제1 곱셈 결과를 (ii) (1) 염색체 이배수성을 갖지 않을 조건부 확률과 (2) 염색체 이배수성을 갖지 않을 선험적 확률의 제2 곱셈 결과로 나눈 몫의 로그이고, 이때 염색체 이배수성을 가질 조건부 확률이 (b)의 태아 분획 및 (a)의 카운트 표시치에 따라 결정되는 것인 단계; 및

[1045] (d) LOR 및 염색체 카운트 표시치에 따라 염색체 이배수성의 존재 또는 부재를 확인하는 단계

[1046] 를 포함하는 방법.

[1047] E1.1. 실시양태 E1에 있어서, 염색체 카운트 표시치가 염색체 내의 모든 부분들에 대한 카운트를 상염색체 내의 모든 부분들에 대한 카운트로 나눈 값인 방법.

[1048] E2. 실시양태 E1 또는 E1.1에 있어서, 염색체 카운트 표시치의 z-점수 정량치를 제공하는 단계를 포함하는 방법.

[1049] E3. 실시양태 E2에 있어서, z-점수가 (i) 검사 샘플 염색체 카운트 표시치로부터 (ii) 정배수체 카운트 표시치의 중간치를 뺀 뺄셈 결과를 (iii) 정배수체 카운트 표시치의 MAD로 나눈 값이고, 이때 (i) 검사 샘플 염색체 카운트 표시치가 염색체 내의 부분들의 카운트를 상염색체 내의 부분들의 카운트로 나눈 비이고, (ii) 정배수체 카운트 표시치의 중간치가 정배수체에 대한 염색체 내의 부분들의 카운트를 상염색체 내의 부분들의 카운트로 나눈 비의 중간치인 방법.

[1050] E4. 실시양태 E1 내지 E3 중 어느 한 실시양태에 있어서, 유전적 변이를 가질 조건부 확률이 (b)에서 검사 샘플에 대해 측정된 태아 분획, (a)에서 검사 샘플에 대한 염색체 카운트 표시치에 대한 z-점수 및 염색체 카운트 표시치에 대한 z-점수의 태아 분획 특이적 분포에 따라 결정되는 것인 방법.

[1051] E5. 실시양태 E4에 있어서, 유전적 변이를 가질 조건부 확률이 하기 방정식 23의 관계에 의해 결정되는 것인 방법:

[1052] 방정식 23

[1053] $$Z \sim Normal\left(\frac{\mu_X}{\sigma_X}\frac{f}{2}, 1\right)$$

[1054] 상기 식에서,

[1055] f는 태아 분획, X는 염색체에 대한 합산된 부분들, $X \sim f(\mu_X, \sigma_X)$이고, 이때 μ_X 및 σ_X는 각각 X의 평균 및 표준 편차이고, $f(\cdot)$는 분포 함수이다.

[1056] E6. 실시양태 E4 또는 E5에 있어서, 유전적 변이를 가질 조건부 확률이 (a)의 검사 샘플 염색체 카운트 표시치에 대한 z-점수와 염색체 카운트 표시치에 대한 z-점수의 태아 분획 특이적 분포 사이의 교차점인 방법.

[1057] E7. 실시양태 E1 내지 E6 중 어느 한 실시양태에 있어서, 염색체 이배수성을 갖지 않을 조건부 확률이 (a)의 염색체 카운트 표시치 및 정배수체에 대한 카운트 표시치에 따라 결정되는 것인 방법.

[1058] E8. 실시양태 E7에 있어서, 유전적 변이를 갖지 않을 조건부 확률이 염색체 카운트 표시치의 z-점수와 정배수체에서의 염색체 카운트 표시치에 대한 z-점수의 분포 사이의 교차점인 방법.

[1059] E9. 실시양태 E1 내지 E8 중 어느 한 실시양태에 있어서, 유전적 변이를 가질 선험적 확률 및 유전적 변이를 갖지 않을 선험적 확률이 검사 대상체를 포함하지 않는 다수의 샘플들로부터 측정되는 것인 방법.

[1060] E10. 실시양태 E1 내지 E9 중 어느 한 실시양태에 있어서, LOR이 0보다 더 큰지 아니면 0보다 더 작은지를 확인하는 단계를 포함하는 방법.

[1061] E11. 실시양태 E1 내지 E10 중 어느 한 실시양태에 있어서, 기준 게놈의 부분들에 맵핑된 핵산 서열 리드의 카운트가 정규화된 카운트인 방법.

[1062] E12. 실시양태 E11에 있어서, 카운트가 GC-LOESS 정규화를 포함하는 정규화에 의해 정규화되는 것인 방법.

[1063] E13. 실시양태 E11 또는 E12에 있어서, 카운트가 주성분 정규화를 포함하는 정규화에 의해 정규화되는 것인 방법.

[1064] E14. 실시양태 E11 내지 E13 중 어느 한 실시양태에 있어서, 카운트가 GC-LOESS 정규화에 이어서 주성분 정규화를 포함하는 정규화에 의해 정규화되는 것인 방법.

[1065] E14.1. 실시양태 E11 내지 E14 중 어느 한 실시양태에 있어서, 카운트가

[1066] (1) (i) 각각의 부분들에 맵핑된 서열 리드의 카운트와 (ii) 각각의 부분들에 대한 GC 함량 사이의 피팅된 관계를 기초로 검사 샘플에 대한 구아닌 및 사이토신(GC) 편향 계수를 측정하는 단계로서, 이때 GC 편향 계수가 선형의 피팅된 관계에 대한 기울기 또는 비선형의 피팅된 관계에 대한 곡률 추정치인 단계; 및

[1067] (2) 각각의 부분들에 대한 (a)의 카운트, (b)의 GC 편향 계수, 및 (i) 다수의 샘플들 각각에 대한 GC 편향 계수와 (ii) 다수의 샘플들에 대한 각각의 부분들에 맵핑된 서열 리드의 카운트 사이의 피팅된 관계를 기초로 마이크로프로세서를 이용하여 각각의 부분들에 대한 게놈 구획 수준을 계산하여 계산된 게놈 구획 수준을 제공하는 단계

[1068] 를 포함하는 정규화에 의해 정규화되는 것인 방법.

[1069] E15. 실시양태 E1 내지 E14.1 중 어느 한 실시양태에 있어서, 염색체 카운트 표시치의 z-점수 정량치를 측정하는 단계 및 상기 정량치가 3.95의 값보다 더 작은지, 이 값보다 더 큰지 아니면 이 값과 동등한지를 확인하는 단계를 포함하는 방법.

[1070] E16. 실시양태 E15에 있어서, 검사 샘플에 대하여, (i) 염색체 카운트 표시치의 z-점수 정량치가 3.95의 값 이상이고 (ii) LOR이 0보다 더 큰 경우, 염색체 이배수성의 존재를 확인하는 단계를 포함하는 방법.

[1071] E17. 실시양태 E15에 있어서, 검사 샘플에 대하여, (i) 염색체 표시치의 z-점수 정량치가 3.95의 값보다 더 작고/작거나, (ii) LOR이 0보다 더 작은 경우, 염색체 이배수성의 부재를 확인하는 단계를 포함하는 방법.

[1072] E18. 실시양태 E16 또는 E17에 있어서, 염색체 이배수성이 삼염색체성 또는 일염색체성인 방법.

[1073] E18.1. 실시양태 E1 내지 E18.1 중 어느 한 실시양태에 있어서, 카운트 표시치가 정규화된 카운트 표시치인 방법.

[1074] E19. 실시양태 E1 내지 E18.1 중 어느 한 실시양태에 있어서, (a), (b), (c) 및 (d) 중 하나 이상 또는 전부가 시스템 내의 마이크로프로세서에 의해 수행되는 것인 방법.

[1075] E20. 실시양태 E1 내지 E19 중 어느 한 실시양태에 있어서, (a), (b), (c) 및 (d) 중 하나 이상 또는 전부가 컴퓨터에 의해 수행되는 것인 방법.

[1076] E21. 실시양태 E1 내지 E20 중 어느 한 실시양태에 있어서, (a), (b), (c) 및 (d) 중 하나 이상 또는 전부가 메모리와 함께 수행되는 것인 방법.

[1077] E22. 실시양태 E1 내지 E21 중 어느 한 실시양태에 있어서, (a) 전에, 임신 여성으로부터 수득된 샘플 중의 핵산을 서열결정하여 핵산 서열 리드를 제공하는 단계를 포함하는 방법.

[1078] E23. 실시양태 E1 내지 E22 중 어느 한 실시양태에 있어서, (a) 전에 핵산 서열 리드를 기준 게놈의 부분들로 맵핑하는 단계를 포함하는 방법.

[1079] 본원에서 참조된 각각의 특허, 특허출원, 공개문헌 및 문헌 전체가 참고로 도입된다. 상기 특허, 특허출원, 공개문헌 및 문헌의 인용은 이들 중 임의의 참고문헌이 적절한 선행기술임을 인정하는 것이 아니고 이 공개문헌들 또는 문헌들의 내용 또는 날짜에 대한 임의의 인정을 구성하는 것도 아니다.

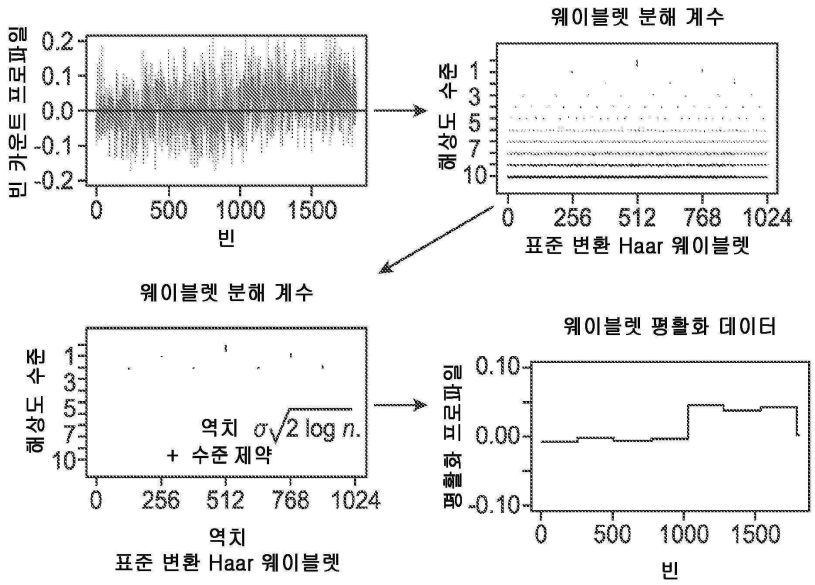
[1080] 본 기술의 기본 양태로부터 벗어나지 않으면서 전술된 기술을 변경시킬 수 있다. 기술이 하나 이상의 특정 실시양태와 관련하여 실질적으로 상세히 기재되어 있지만, 당업자는 본원에 구체적으로 개시된 실시양태를 변경시킬 수 있고 이 변경 및 개선이 본 기술의 범위 및 사상 내에 있다는 것을 인식할 것이다.

[1081] 본원에 예시적으로 기재된 기술은 본원에 구체적으로 개시되지 않은 임의의 요소(들)의 부재 하에서 실시될 수 있다. 따라서, 예를 들면, 각각의 경우 본원에서 용어 "포함하는", "본질적으로 구성된" 및 "구성된" 중 임의의 용어는 다른 2개의 용어들 중 어느 하나로 대체될 수 있다. 사용된 용어 및 표현은 기술하기 위한 용어로서 사용되고 한정하는 용어로서 사용되지 않고, 이러한 용어 및 표현의 사용은 제시되어 있고 기재되어 있는 특징 또는 이의 분절의 임의의 등가물을 배제하지 않고, 다양한 변경이 특허청구된 기술의 범위 내에서 가능하다. 하나의 요소 또는 하나 초과의 요소가 기재된다는 것이 문맥상 명확하지 않은 한, 단수형 용어는 그가 수식하는 하나의 또는 복수의 요소를 지칭할 수 있다(예를 들면, "시약"은 하나 이상의 시약을 의미할 수 있다). 본원에서 사용된 용어 "약"은 기저 파라미터의 10% 이내(즉, + 또는 - 10%)의 값을 지칭하고, 일련의 값들의 시작부에서 용어 "약"의 사용은 각각의 값을 수식한다(즉, "약 1, 2 및 3"은 약 1, 약 2 및 약 3을 지칭한다). 예를 들면, "약 100 그램"의 중량은 90 그램 내지 110 그램의 중량을 포함할 수 있다. 또한, 값의 목록(예를 들면, 약 50%, 60%, 70%, 80%, 85% 또는 86%)이 본원에 기재되어 있을 때, 상기 목록은 이의 모든 중간 값 및 분수 값(예를 들면, 54%, 85.4%)을 포함한다. 따라서, 본 기술이 대표적인 실시양태 및 임의적인 특징에 의해 구체적으로 개시되어 있지만, 당업자는 본원의 사상을 변형시킬 수 있거나 변경시킬 수 있고, 이러한 변형 및 변경이 본 기술의 범위 내에 있는 것으로서 간주된다는 것을 이해해야 한다.

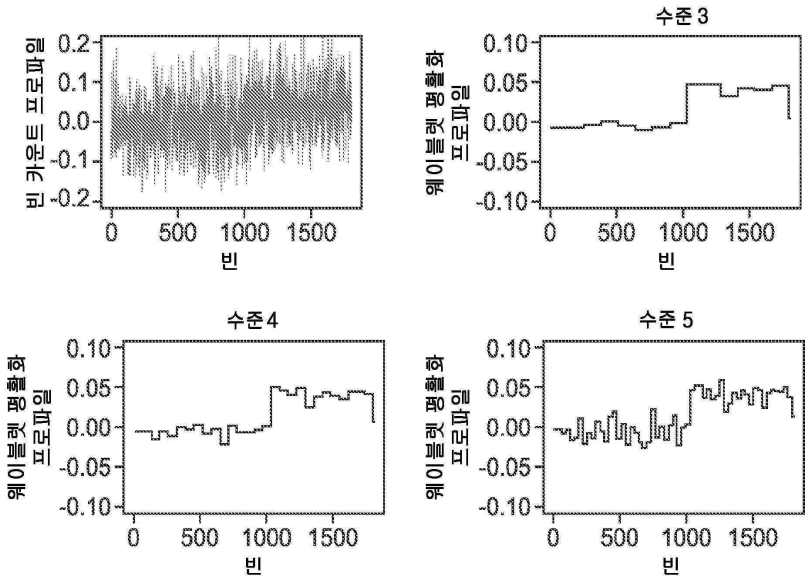
[1082] 본 기술의 일부 실시양태들은 하기 특허청구범위에 기재되어 있다.

도면

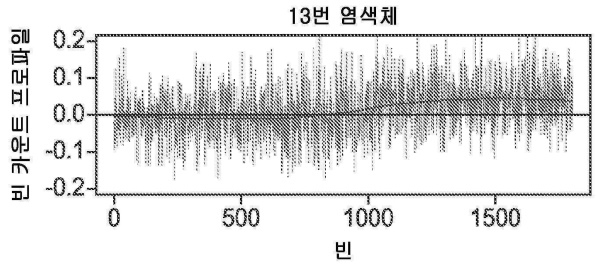
도면1

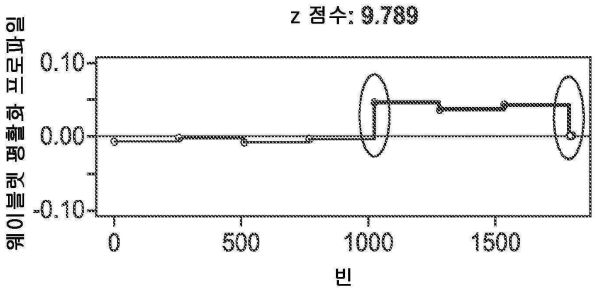


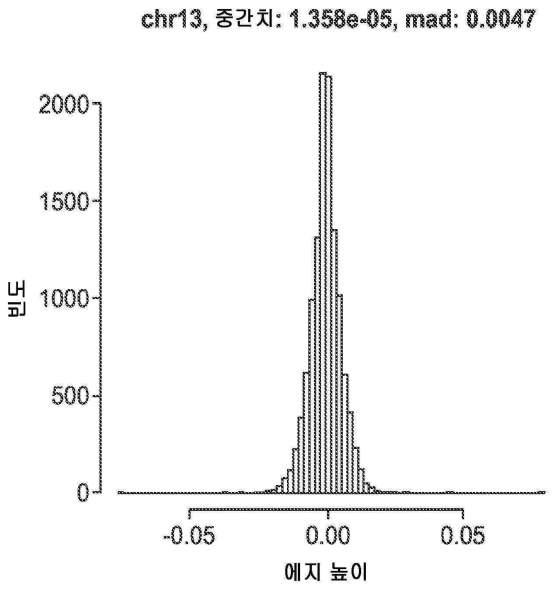
도면2


빈 카운트 프로파일
0.2
0.1
0.0
-0.1
-0.2
0
500
1000
1500
빈
수준 3
웨이블렛 평활화 프로파일
0.10
0.05
0.00
-0.05
-0.10
0
500
1000
1500
빈
수준 4
웨이블렛 평활화 프로파일
0.10
0.05
0.00
-0.05
-0.10
0
500
1000
1500
빈
수준 5
웨이블렛 평활화 프로파일
0.10
0.05
0.00
-0.05
-0.10
0
500
1000
1500
빈

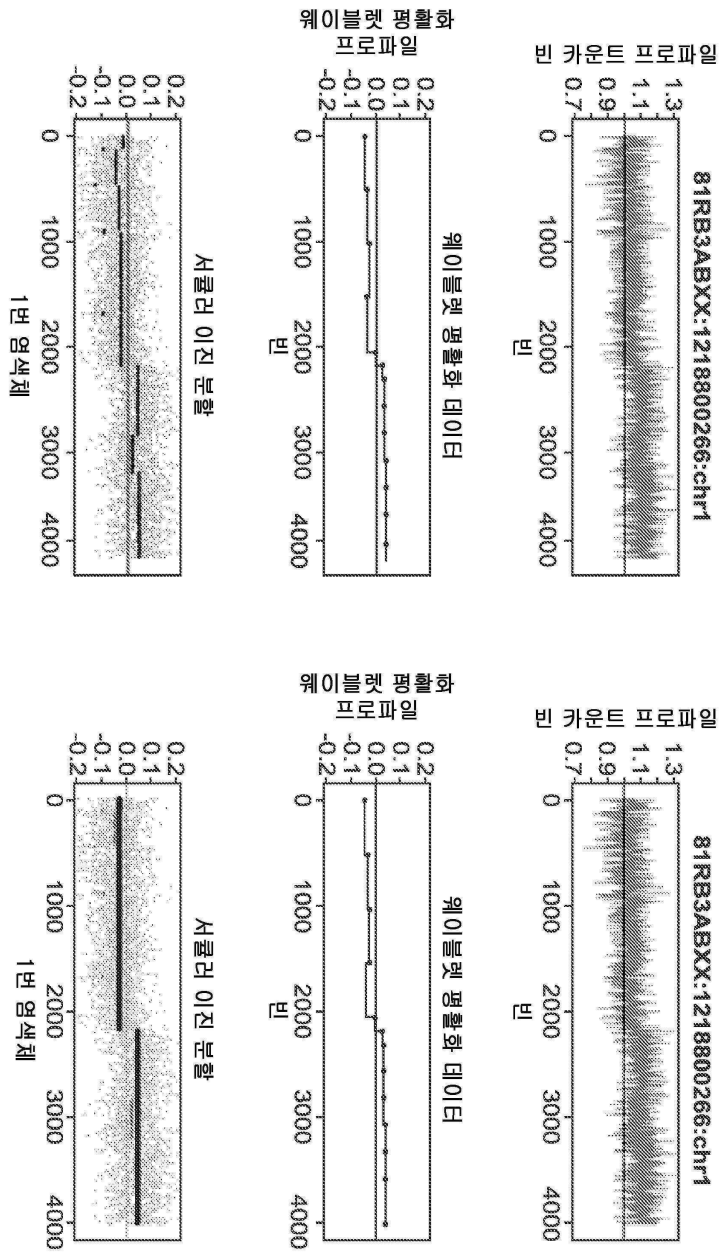
도면3


13번 염색체
빈 카운트 프로파일
0.2
0.1
0.0
-0.1
-0.2
0
500
1000
1500
빈

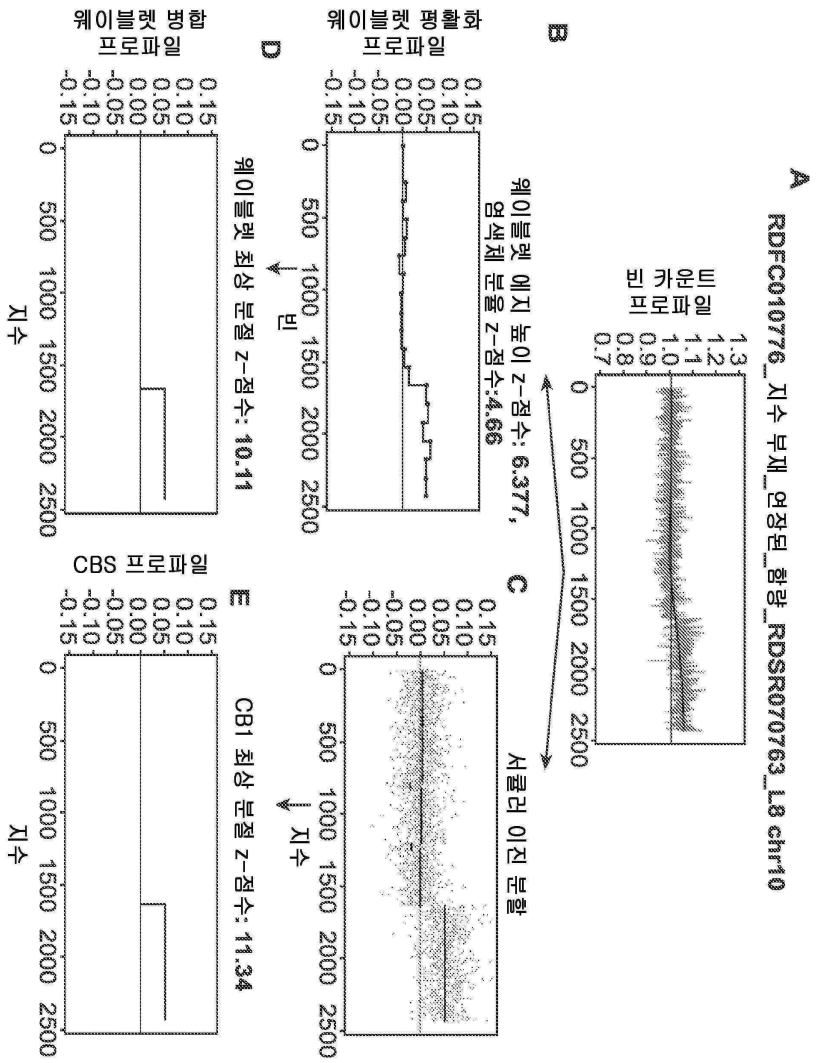

z 점수: 9.789
웨이블렛 평활화 프로파일
0.10
0.00
-0.10
0
500
1000
1500
빈


chr13, 중간치: 1.358e-05, mad: 0.0047
빈도
2000
1500
1000
500
0
-0.05
0.00
0.05
에지 높이

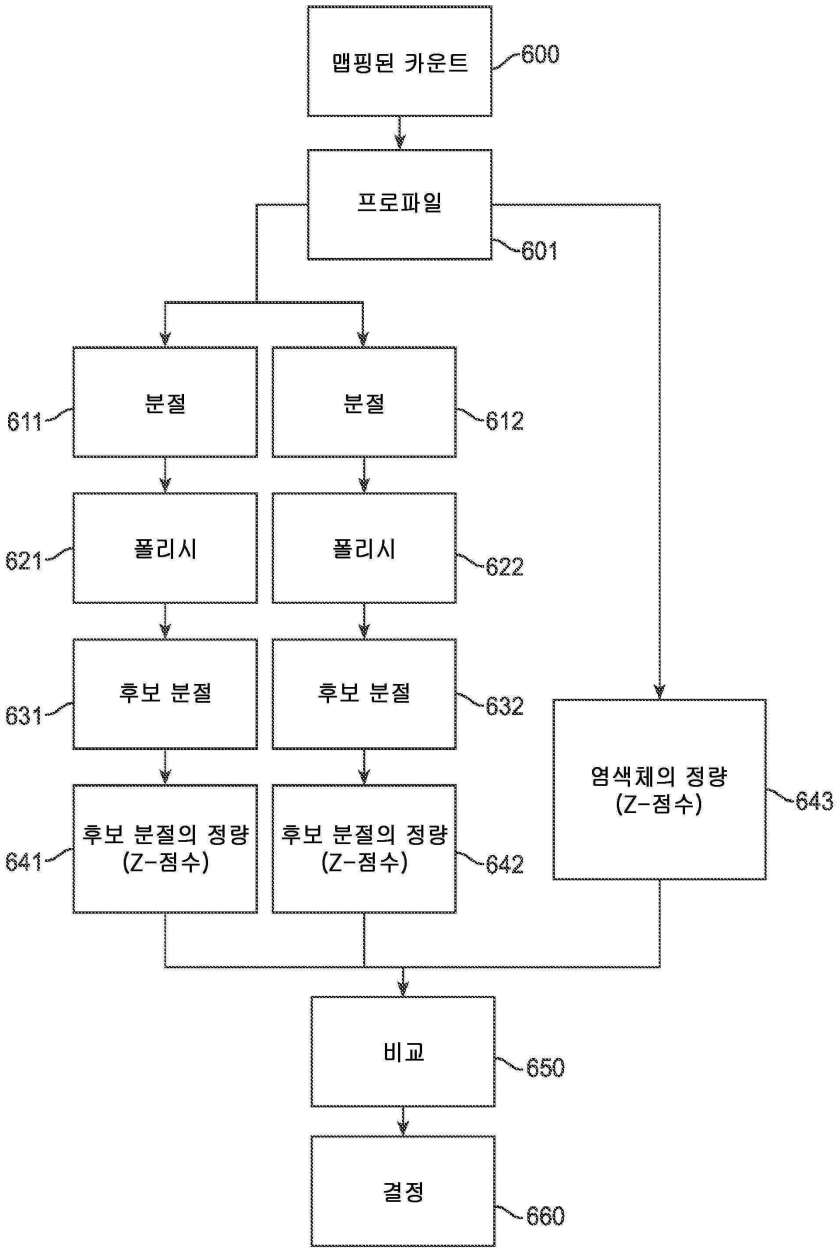
도면4


81RB3ABXX:1218800266:chr1
빈 카운트 프로파일
웨이블렛 평활화 프로파일
웨이블렛 평활화 데이터
서큘러 이진 분할
1번 염색체
빈

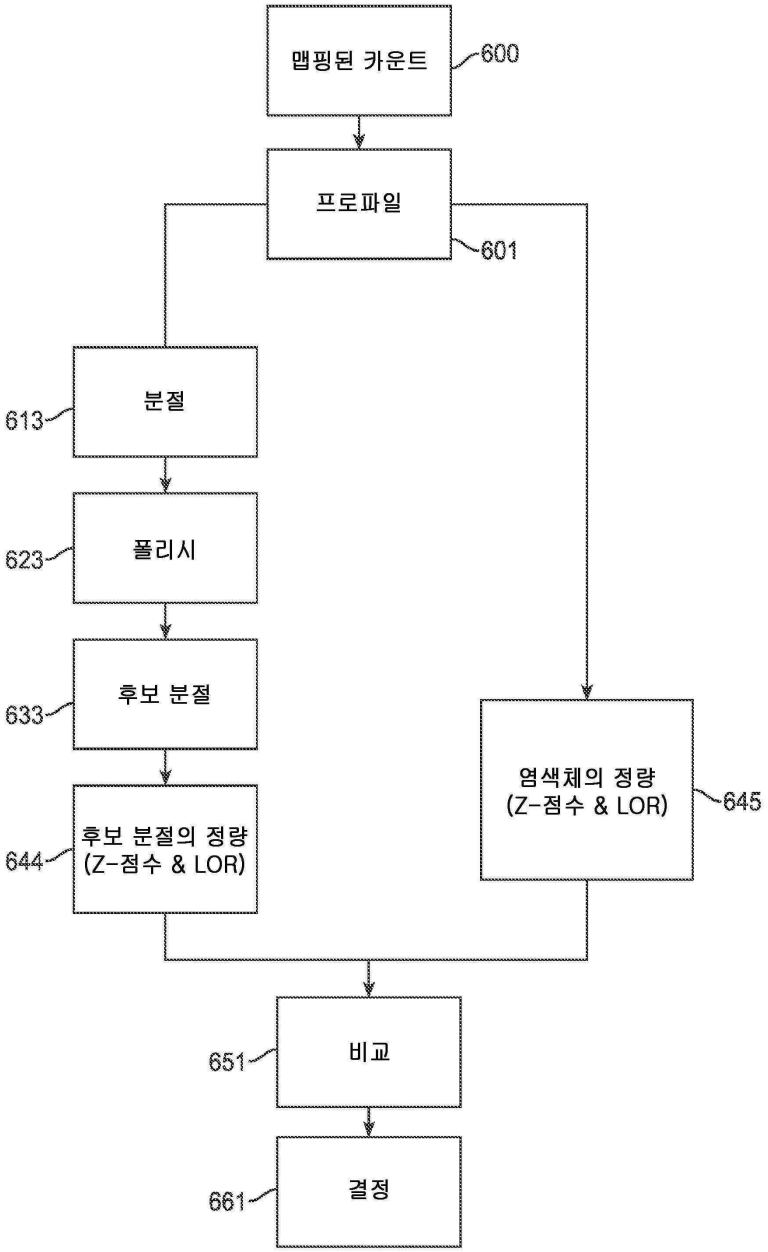
도면5


A
RDFC010776_지수 부재_연장된_항량_RDSR07063_L8 chr10
빈 카운트 프로파일
B
웨이블렛 에지 높이 z-점수: 6.377,
염색체 분할 z-점수:4.66
웨이블렛 평활화 프로파일
C
서큘러 이진 분할
빈
지수
D
웨이블렛 최상 분절 z-점수: 10.11
웨이블렛 병합 프로파일
E
CB1 최상 분절 z-점수: 11.34
CBS 프로파일

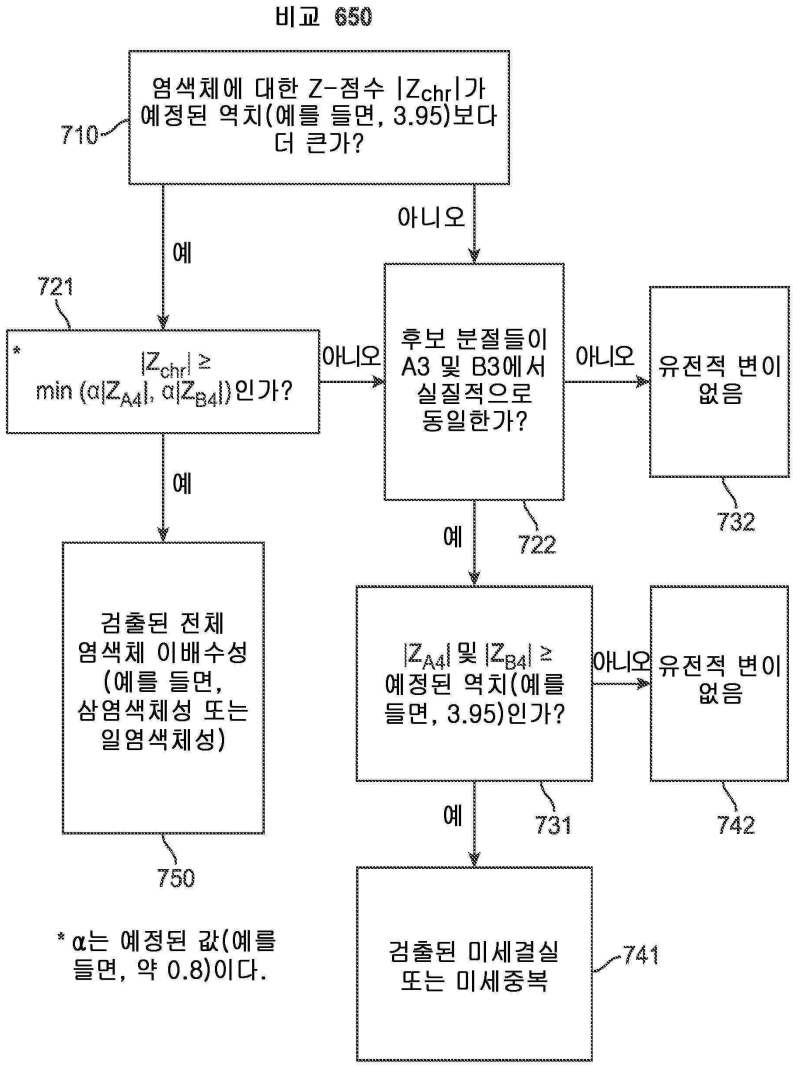
도면6a


맵핑된 카운트
600
프로파일
601
분절
611
분절
612
폴리시
621
폴리시
622
후보 분절
631
후보 분절
632
후보 분절의 정량 (Z-점수)
641
후보 분절의 정량 (Z-점수)
642
염색체의 정량 (Z-점수)
643
비교
650
결정
660

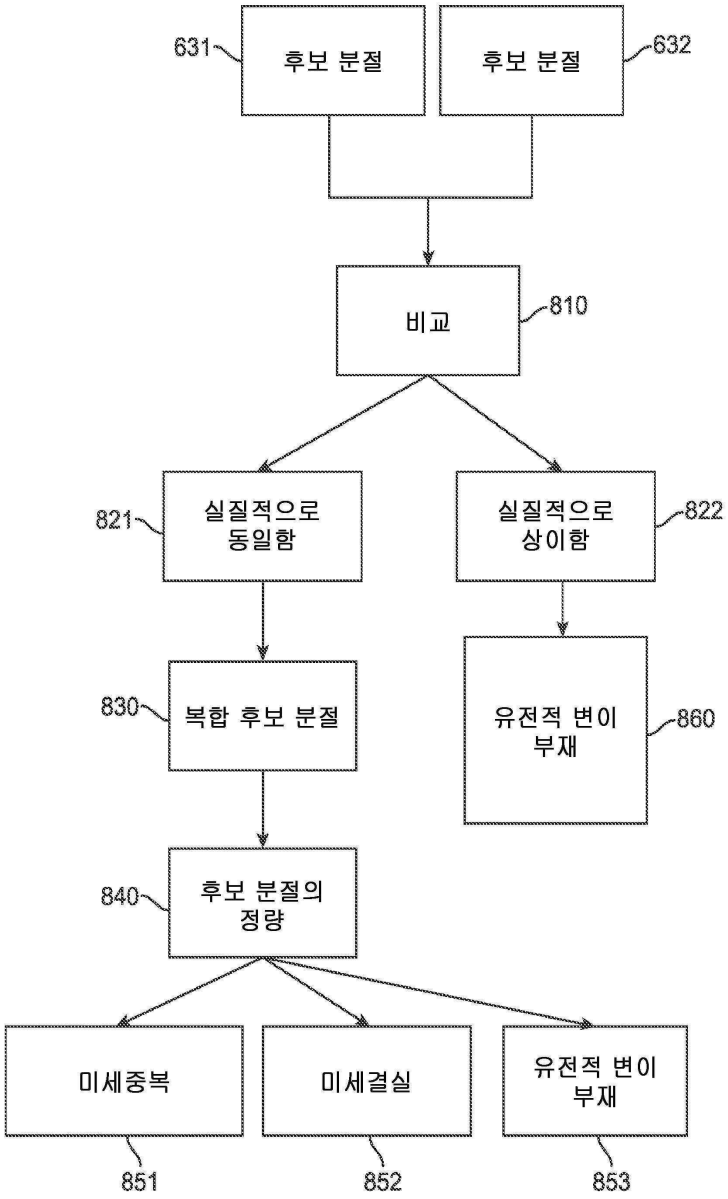
도면6b


맵핑된 카운트
600
프로파일
601
분절
613
폴리시
623
후보 분절
633
후보 분절의 정량
(Z-점수 & LOR)
644
염색체의 정량
(Z-점수 & LOR)
645
비교
651
결정
661

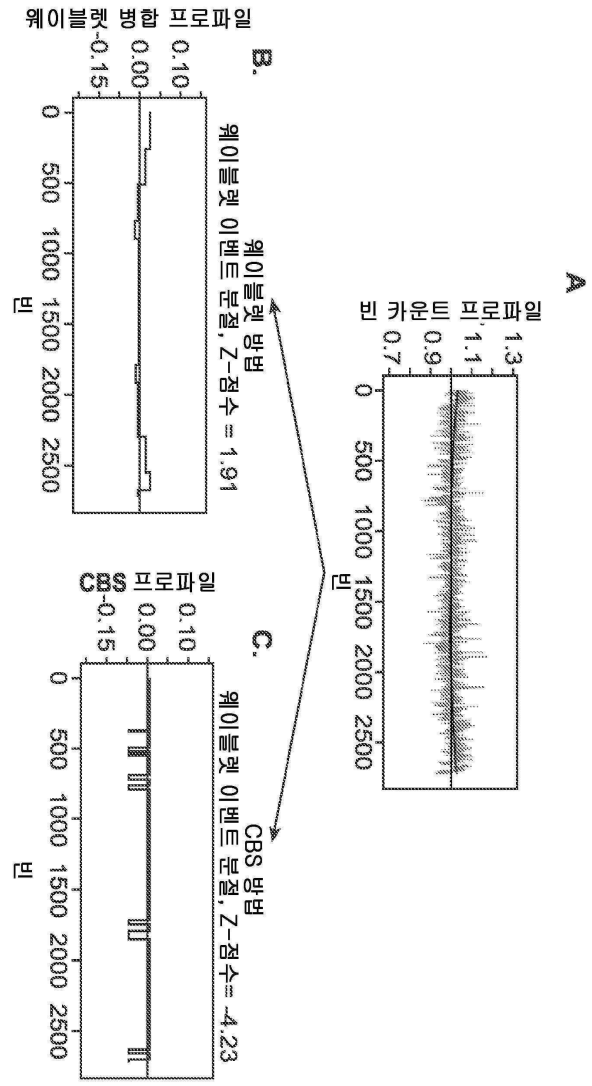
도면7


비교 650
염색체에 대한 Z-점수 $|Z_{chr}|$가 예정된 역치(예를 들면, 3.95)보다 더 큰가?
710
예
아니오
721
* $|Z_{chr}| \geq \min(\alpha|Z_{A4}|, \alpha|Z_{B4}|)$인가?
아니오
후보 분절들이 A3 및 B3에서 실질적으로 동일한가?
아니오
유전적 변이 없음
732
예
722
검출된 전체 염색체 이배수성 (예를 들면, 삼염색체성 또는 일염색체성)
750
$|Z_{A4}|$ 및 $|Z_{B4}| \geq$ 예정된 역치(예를 들면, 3.95)인가?
아니오
유전적 변이 없음
742
예
731
검출된 미세결실 또는 미세중복
741
* α는 예정된 값(예를 들면, 약 0.8)이다.

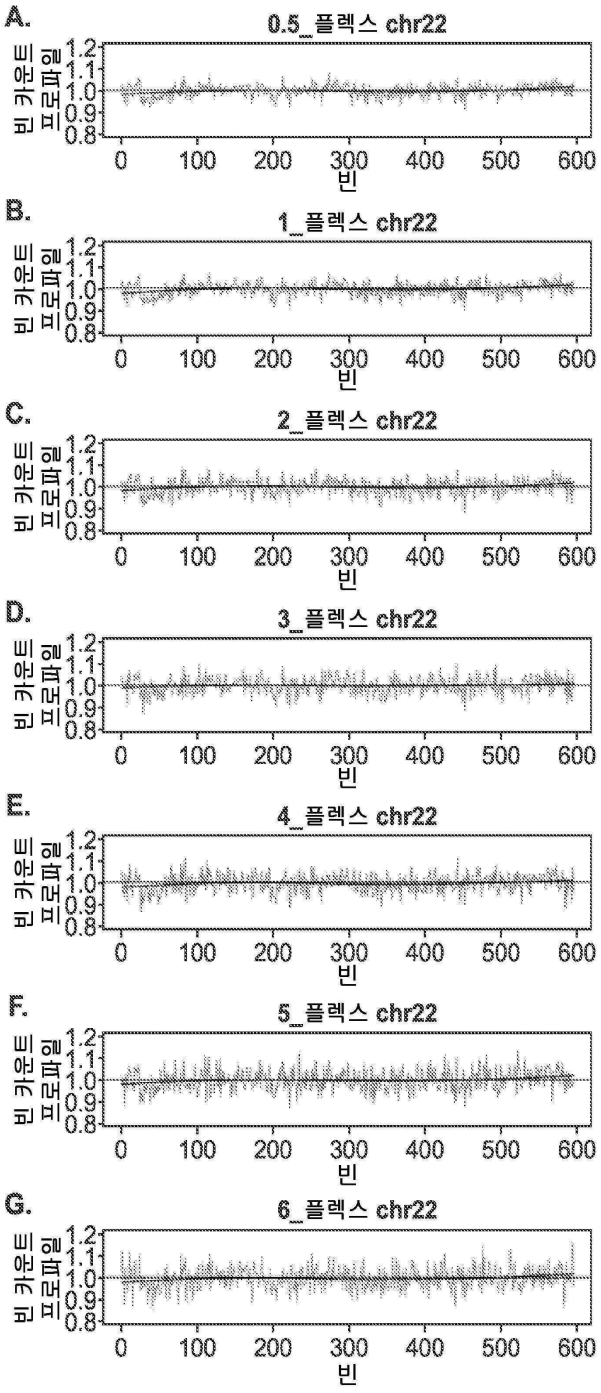
도면8


631
후보 분절
후보 분절
632
비교
810
821
실질적으로 동일함
실질적으로 상이함
822
830
복합 후보 분절
유전적 변이 부재
860
840
후보 분절의 정량
미세중복
미세결실
유전적 변이 부재
851
852
853

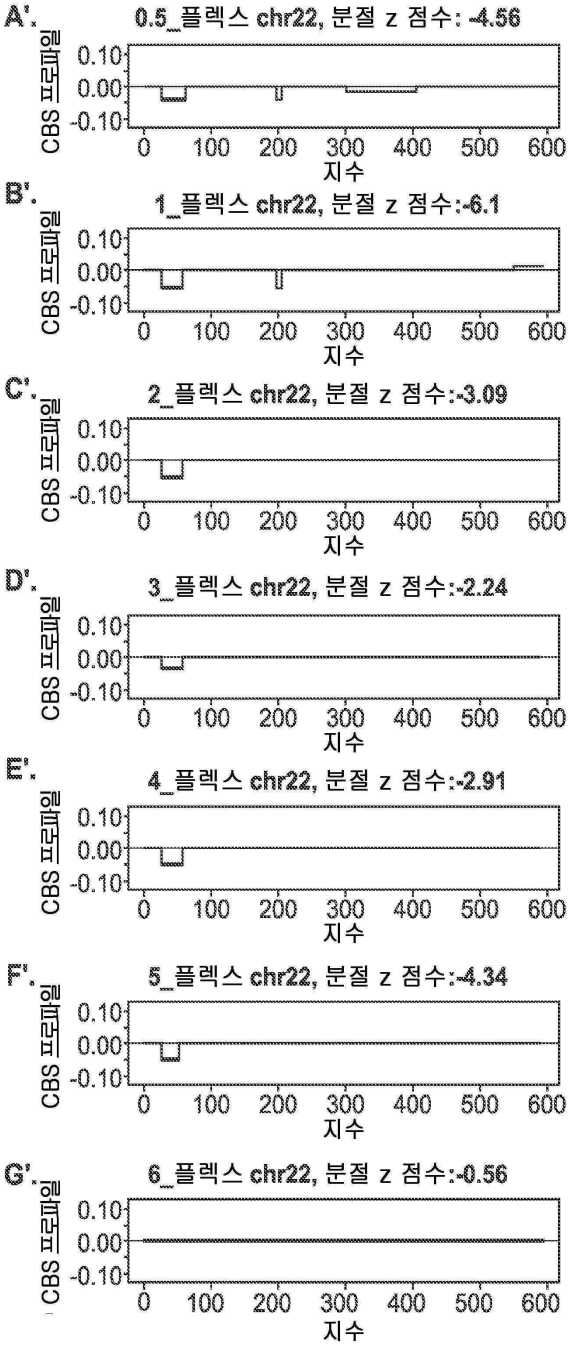
도면9


A
빈 카운트 프로파일
0.7
0.9
1.1
1.3
0
500
1000
1500
2000
2500
피
웨이블렛 방법
웨이블렛 이벤트 분절, Z-점수 = 1.91
B.
웨이블렛 병합 프로파일
-0.15
0.00
0.10
CBS 방법
웨이블렛 이벤트 분절, Z-점수= -4.23
C.
CBS 프로파일

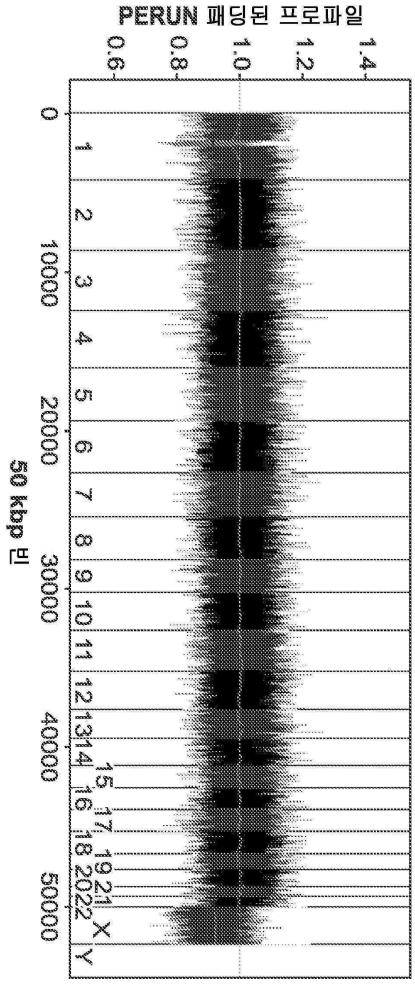
도면10a


A.
0.5_플렉스 chr22
B.
1_플렉스 chr22
C.
2_플렉스 chr22
D.
3_플렉스 chr22
E.
4_플렉스 chr22
F.
5_플렉스 chr22
G.
6_플렉스 chr22
빈 카운트 프로파일
빈
0
100
200
300
400
500
600
0.8
0.9
1.0
1.1
1.2

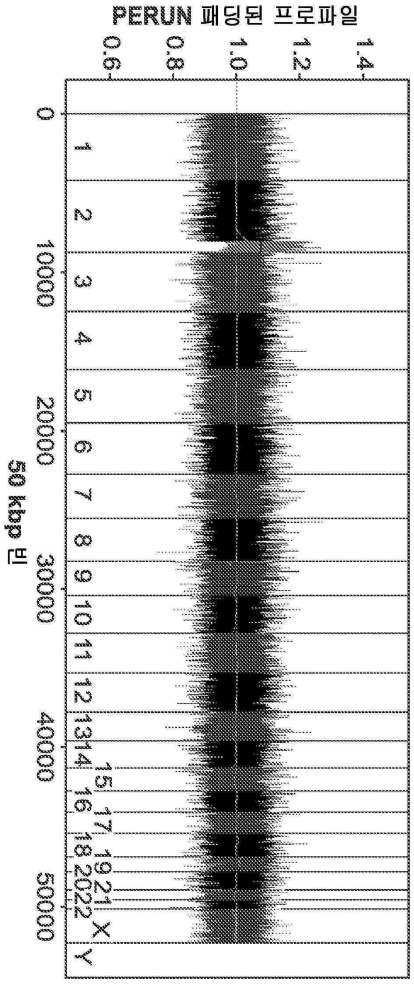
도면10b


A'.
0.5_플렉스 chr22, 분절 z 점수: -4.56
B'.
1_플렉스 chr22, 분절 z 점수:-6.1
C'.
2_플렉스 chr22, 분절 z 점수:-3.09
D'.
3_플렉스 chr22, 분절 z 점수:-2.24
E'.
4_플렉스 chr22, 분절 z 점수:-2.91
F'.
5_플렉스 chr22, 분절 z 점수:-4.34
G'.
6_플렉스 chr22, 분절 z 점수:-0.56
CBS 프로파일
0.10
0.00
-0.10
0
100
200
300
400
500
600
지수

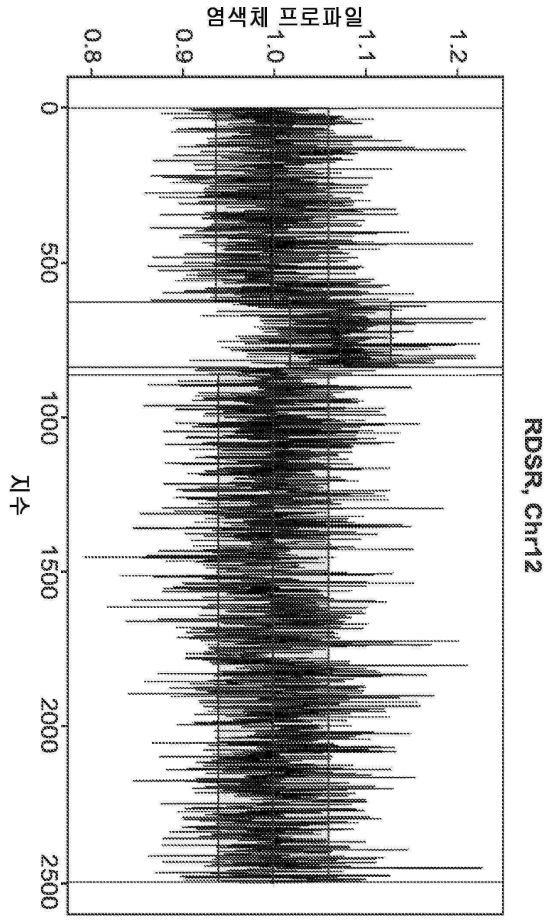
도면11a


PERUN 패딩된 프로파일
0.6
0.8
1.0
1.2
1.4
0
10000
20000
30000
40000
50000
50 kbp
피
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

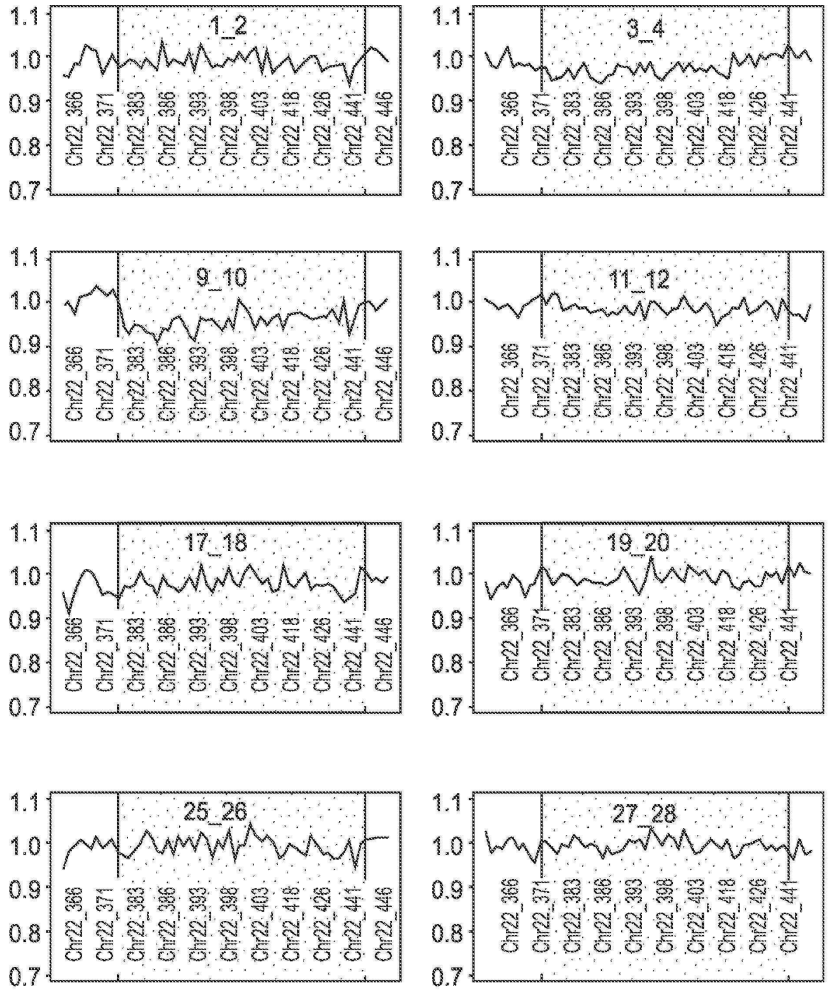
도면11b


PERUN 패딩된 프로파일
0.6
0.8
1.0
1.2
1.4
0
10000
20000
30000
40000
50000
50 kbp 단
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
21
22
X
Y

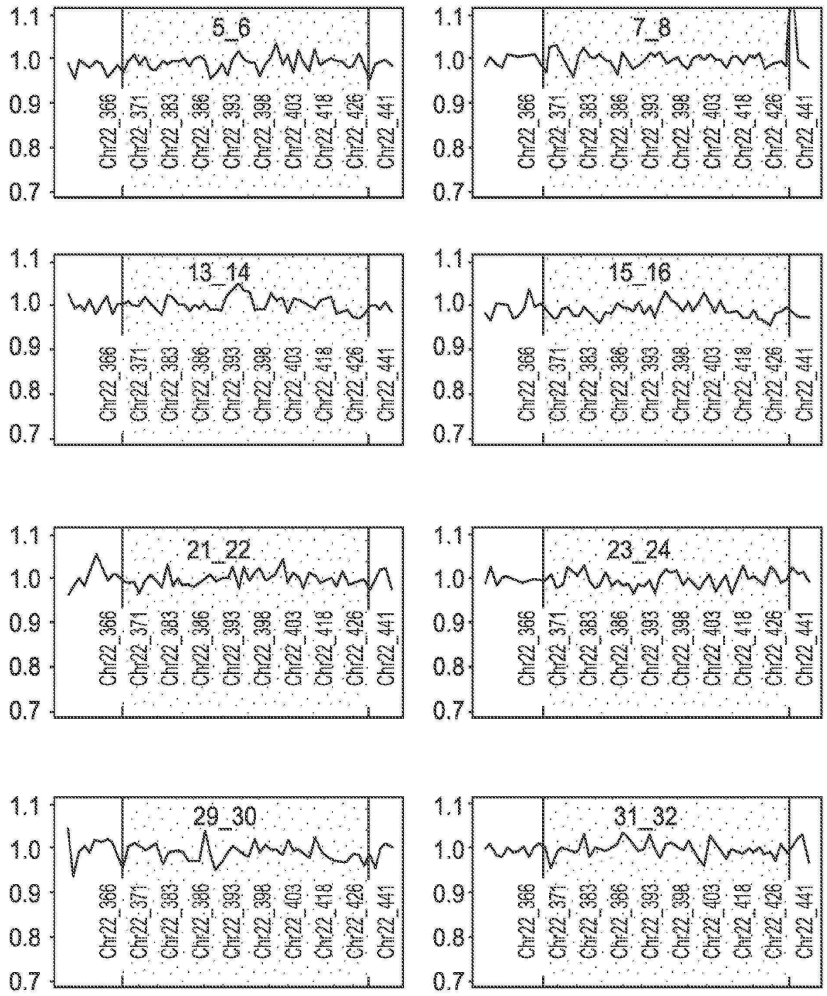
도면12



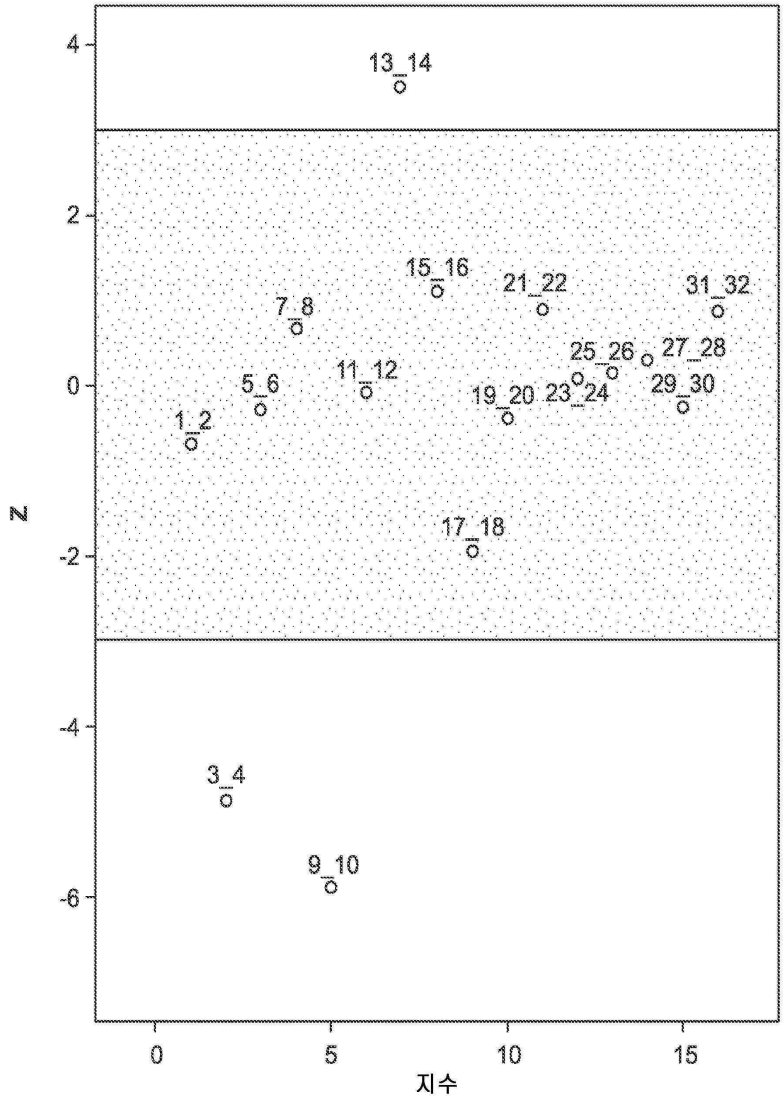
도면13a


1_2
3_4
9_10
11_12
17_18
19_20
25_26
27_28
1.1
1.0
0.9
0.8
0.7
Chr22_366
Chr22_371
Chr22_383
Chr22_386
Chr22_393
Chr22_398
Chr22_403
Chr22_418
Chr22_426
Chr22_441
Chr22_446

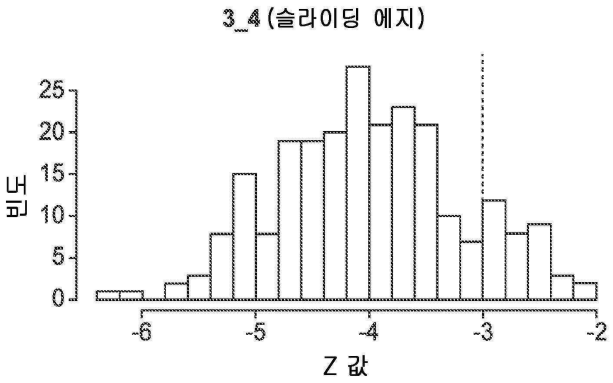
도면13b


5_6
7_8
13_14
15_16
21_22
23_24
29_30
31_32
1.1
1.0
0.9
0.8
0.7
Chr22_366
Chr22_371
Chr22_383
Chr22_386
Chr22_393
Chr22_398
Chr22_403
Chr22_418
Chr22_426
Chr22_441

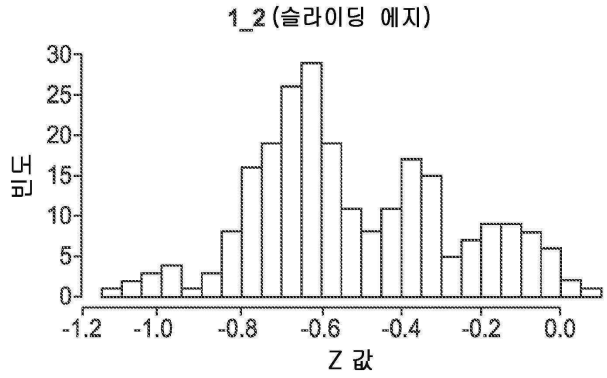
도면14


13_14
15_16
21_22
31_32
7_8
25_26
27_28
11_12
5_6
19_20
23_24
29_30
1_2
17_18
3_4
9_10
Z
지수
4
2
0
-2
-4
-6
0
5
10
15

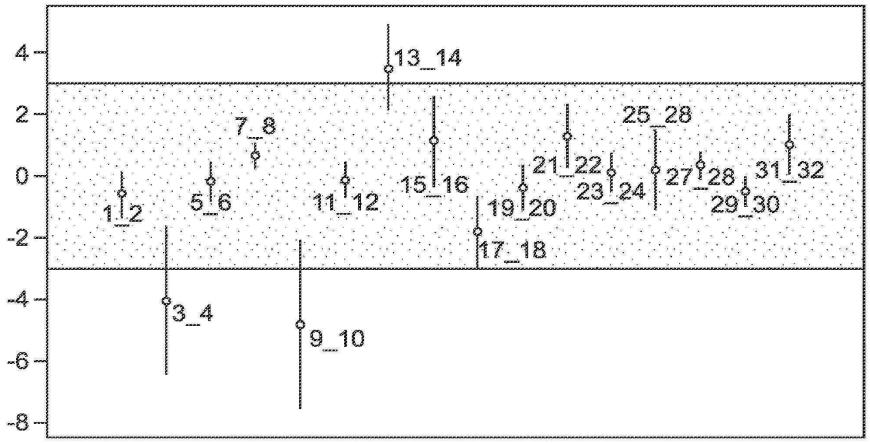
도면15


3_4 (슬라이딩 에지)
빈도
Z 값
25
20
15
10
5
0
-6
-5
-4
-3
-2

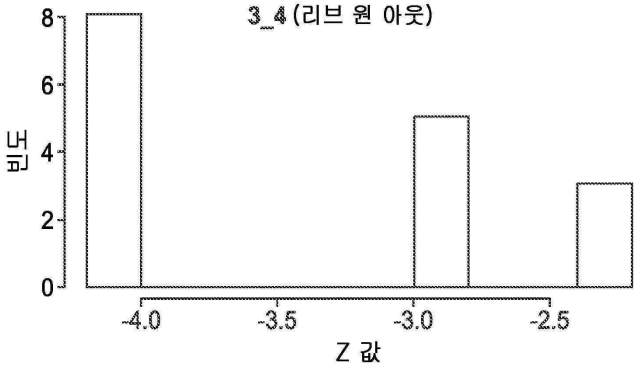
도면16



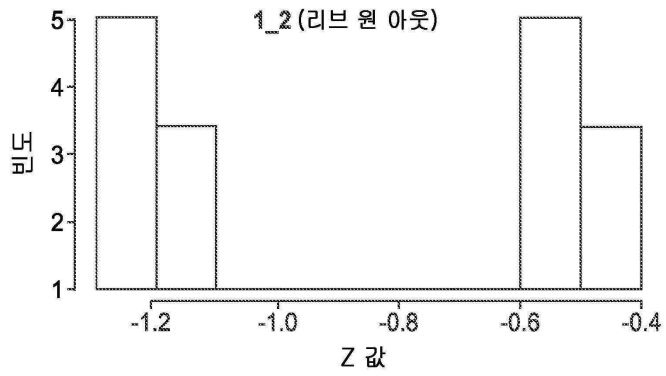
도면17



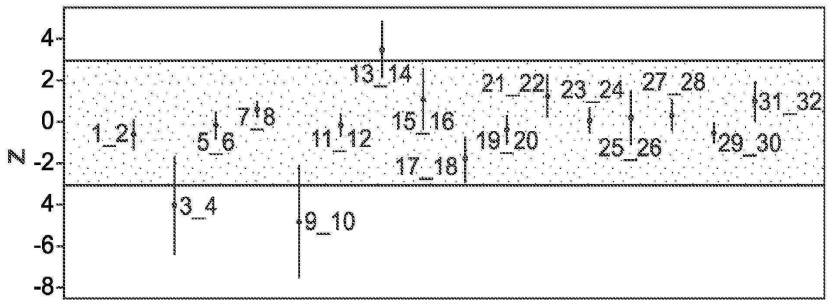
도면18



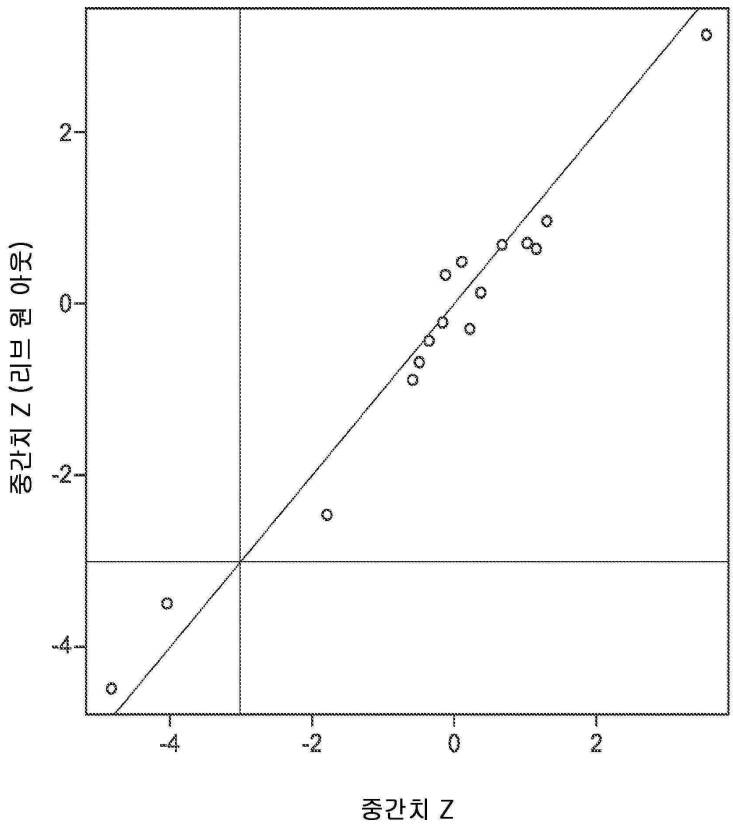
도면19



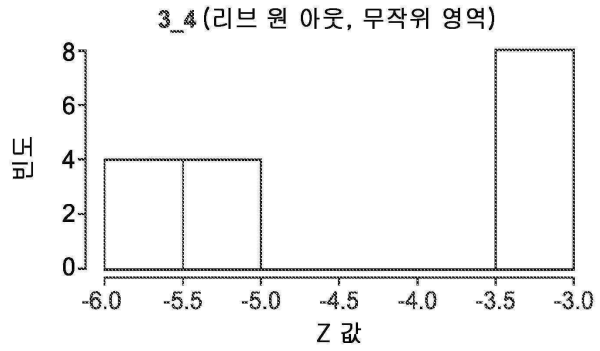
도면20



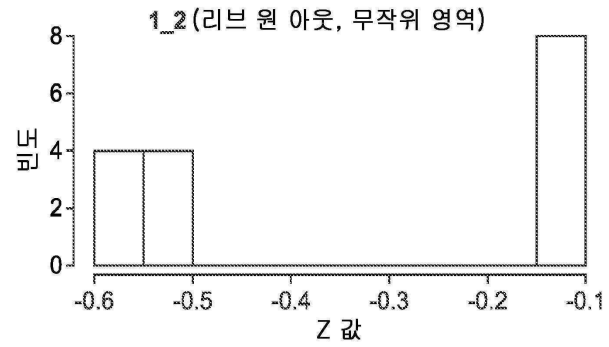
도면21



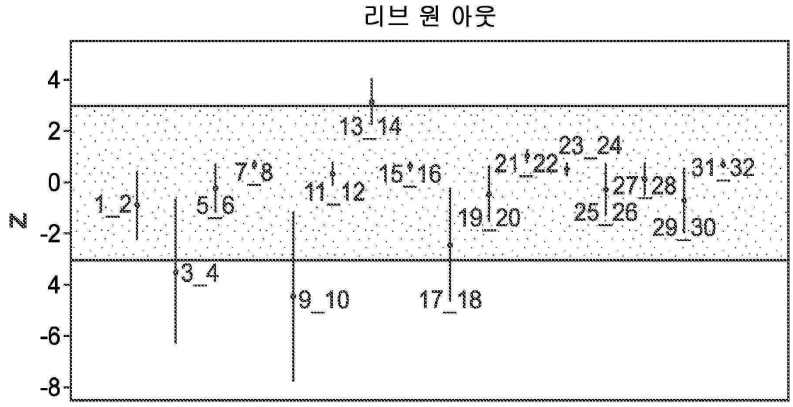
도면22



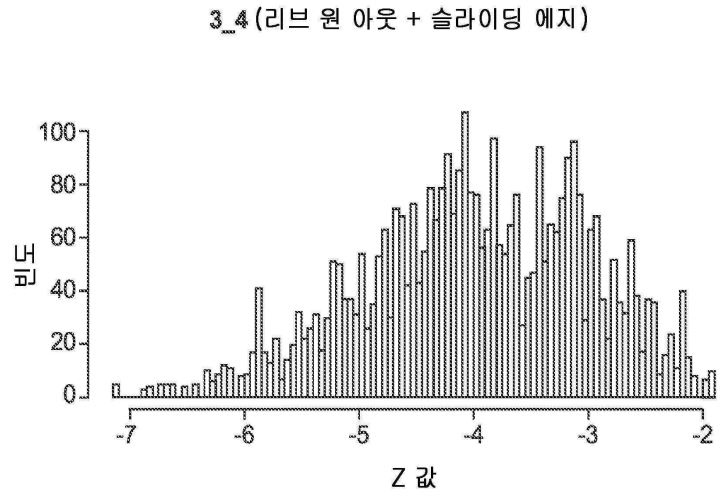
도면23



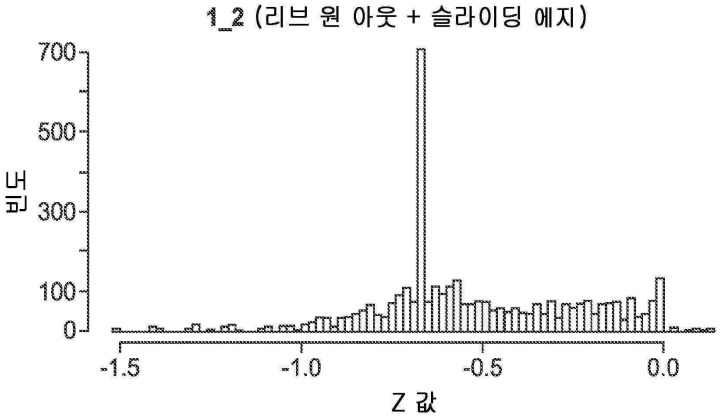
도면24



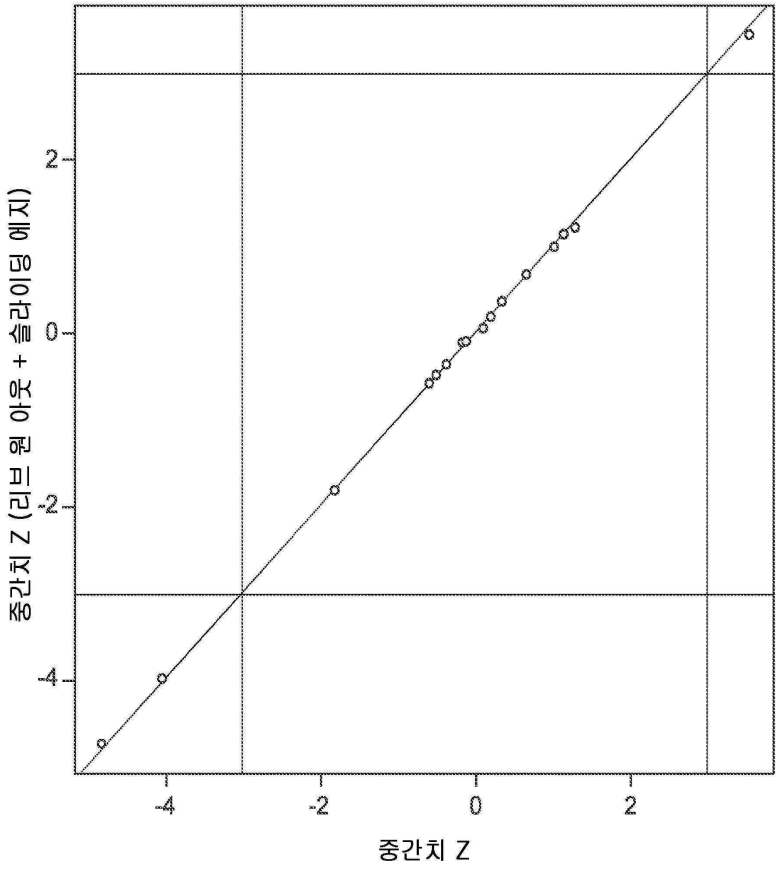
도면25


3_4 (리브 원 아웃 + 슬라이딩 에지)
100
80
60
40
20
0
-7
-6
-5
-4
-3
-2
Z 값

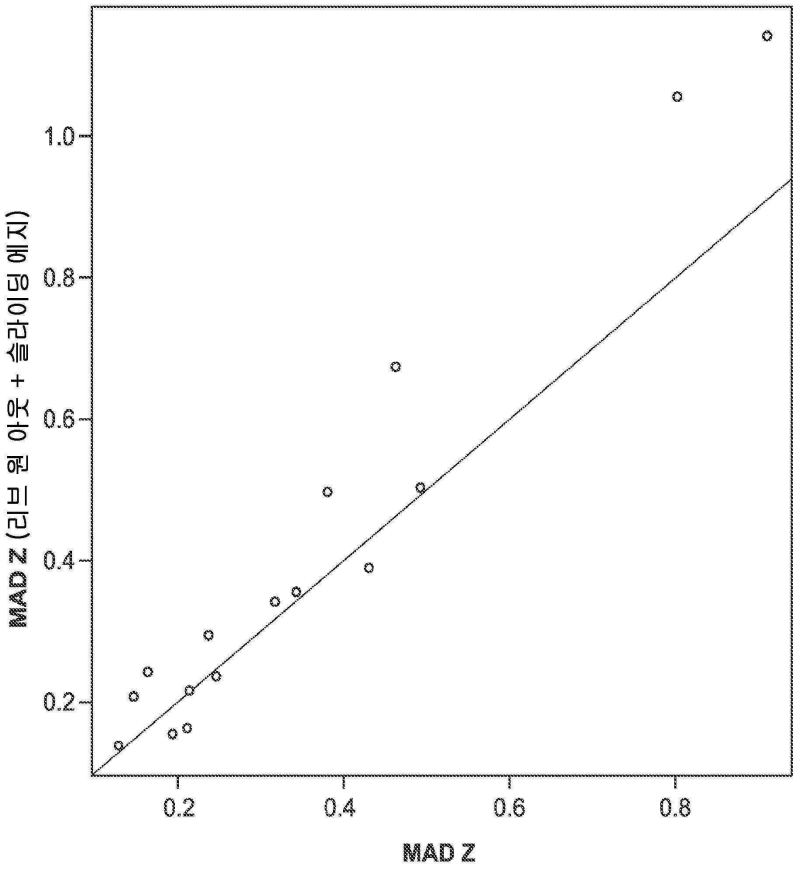
도면26


1_2 (리브 원 아웃 + 슬라이딩 에지)
700
500
300
100
0
-1.5
-1.0
-0.5
0.0
Z 값

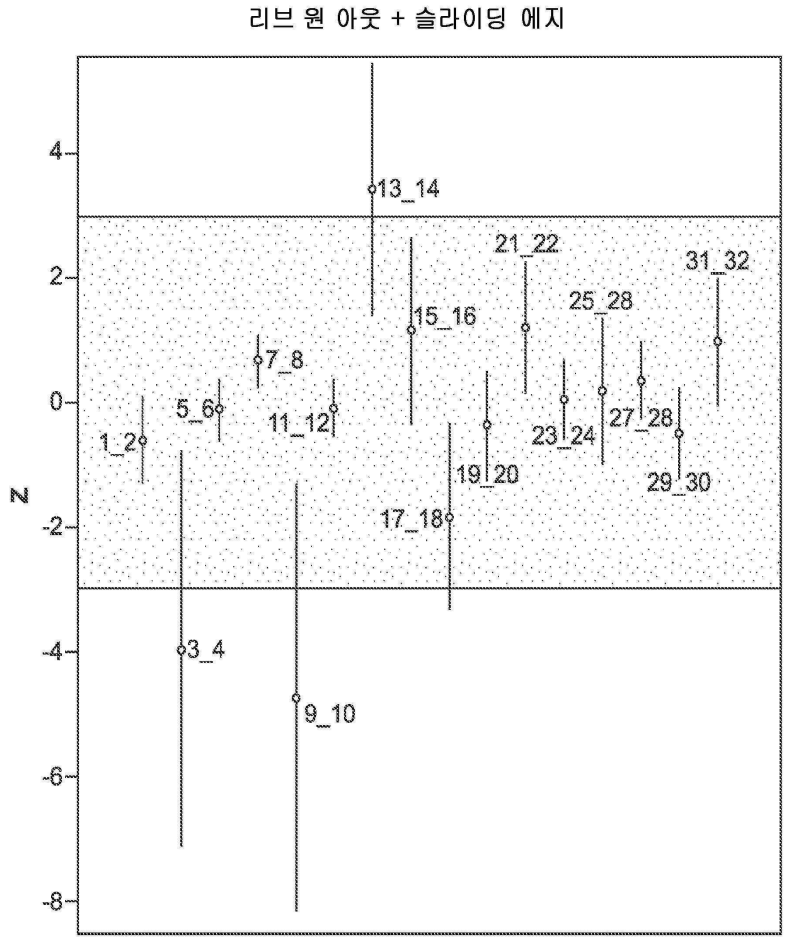
도면27



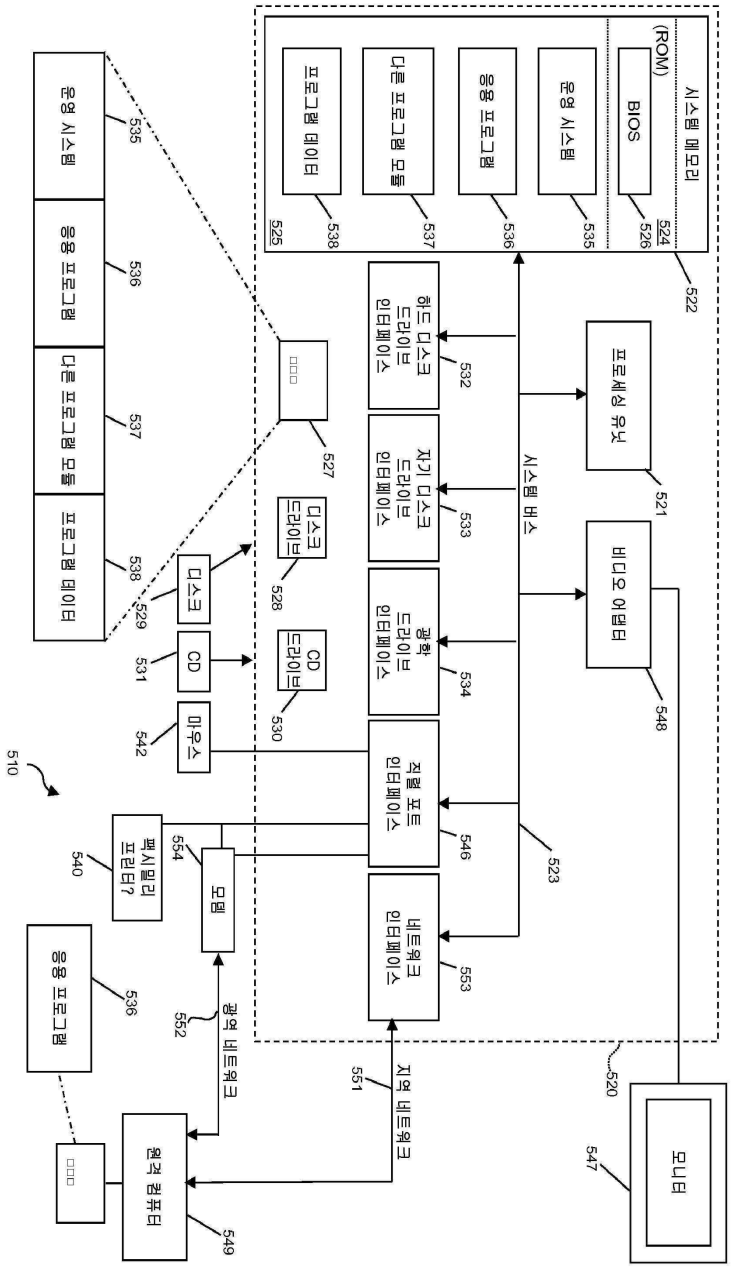
도면28



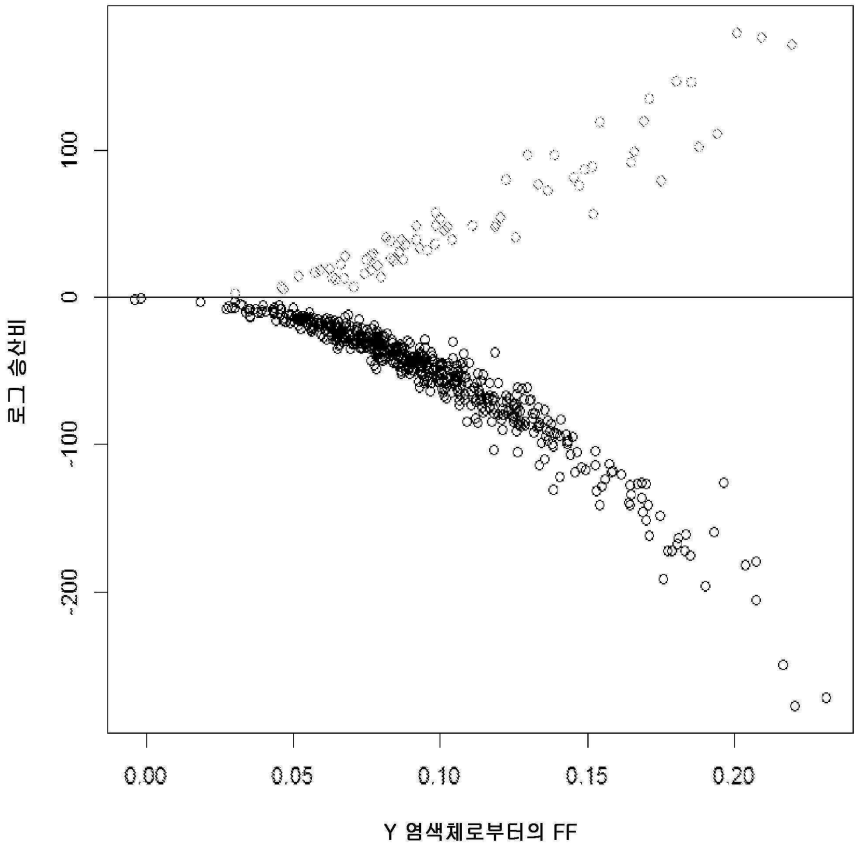
도면29


리브 원 아웃 + 슬라이딩 에지
4
2
0
-2
-4
-6
-8
N
13_14
21_22
31_32
25_28
15_16
7_8
5_6
11_12
1_2
23_24
27_28
19_20
29_30
17_18
3_4
9_10

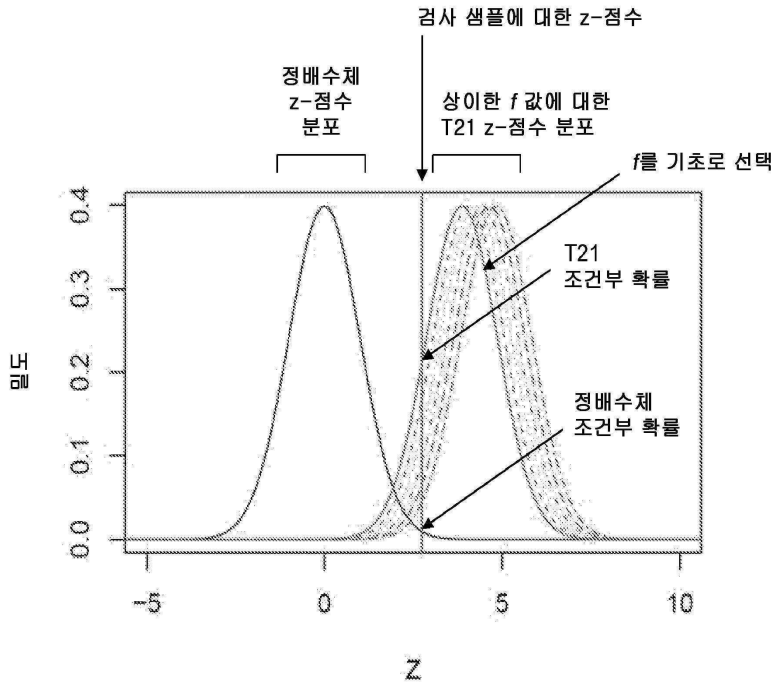
도면30


시스템 메모리
(ROM)
BIOS
운영 시스템
응용 프로그램
다른 프로그램 모듈
프로그램 데이터
524
526
535
536
537
538
525
522
하드 디스크 드라이브 인터페이스
자기 디스크 드라이브 인터페이스
광학 드라이브 인터페이스
직렬 포트 인터페이스
네트워크 인터페이스
시스템 버스
523
532
533
534
546
553
프로세싱 유닛
비디오 어댑터
521
548
520
모니터
547
551
지역 네트워크
552
광역 네트워크
549
원격 컴퓨터
554
모뎀
540
펙시밀리 프린터?
542
마우스
531
CD
530
CD 드라이브
529
디스크
528
디스크 드라이브
527
510

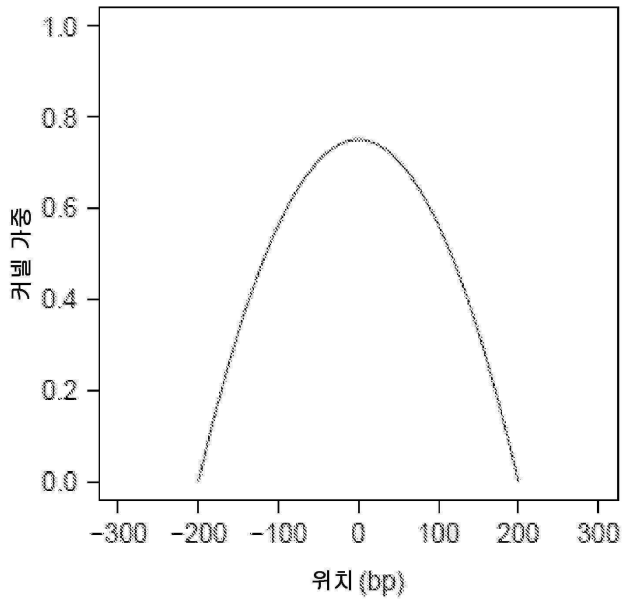
도면31


로그 승산비
100
0
-100
-200
0.00
0.05
0.10
0.15
0.20
Y 염색체로부터의 FF

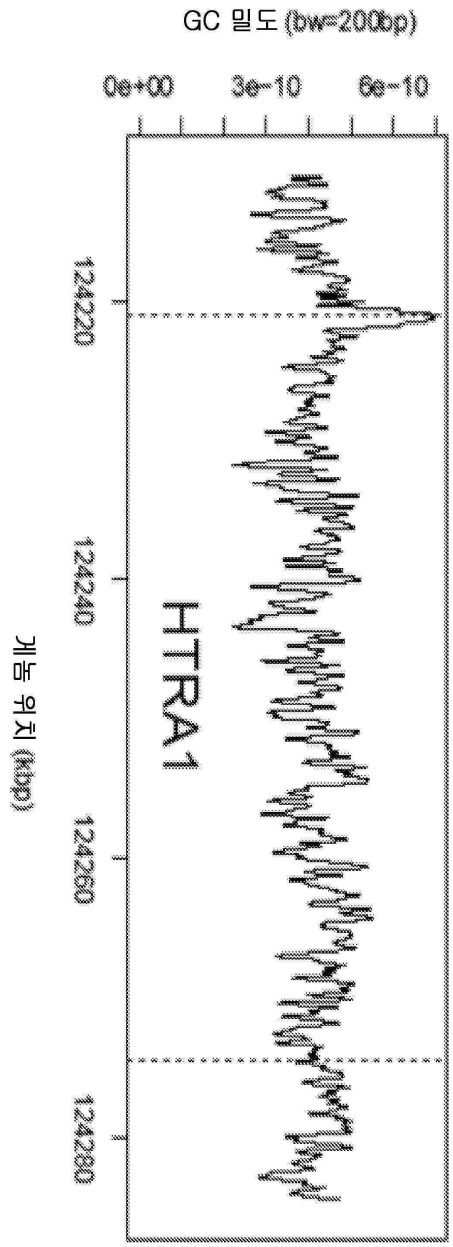
도면32


검사 샘플에 대한 z-점수
정배수체
z-점수
분포
상이한 f 값에 대한
T21 z-점수 분포
f를 기초로 선택
T21
조건부 확률
정배수체
조건부 확률
밀도
0.0
0.1
0.2
0.3
0.4
-5
0
5
10
Z

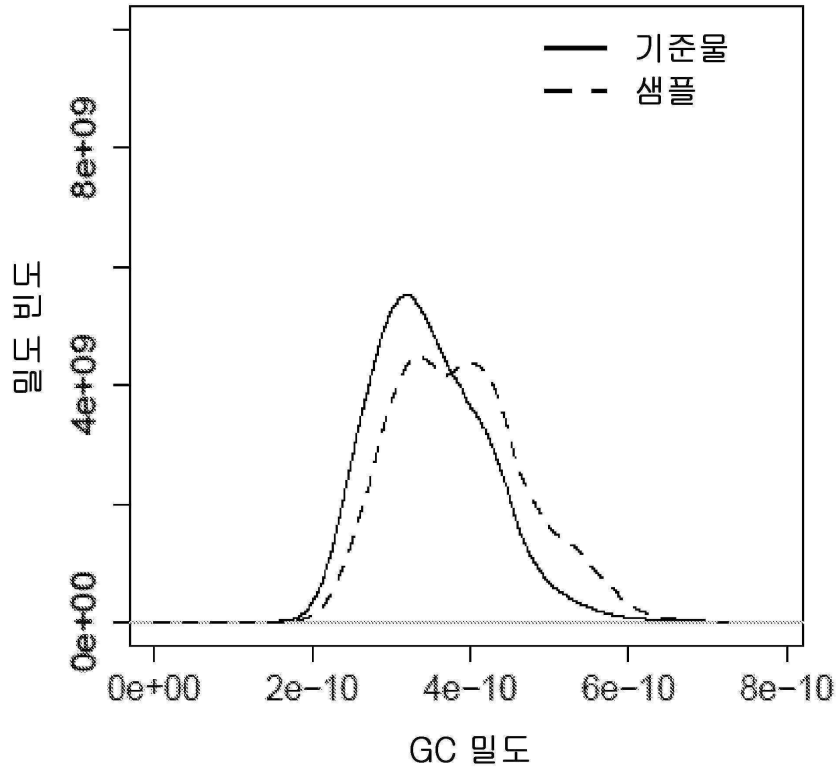
도면33



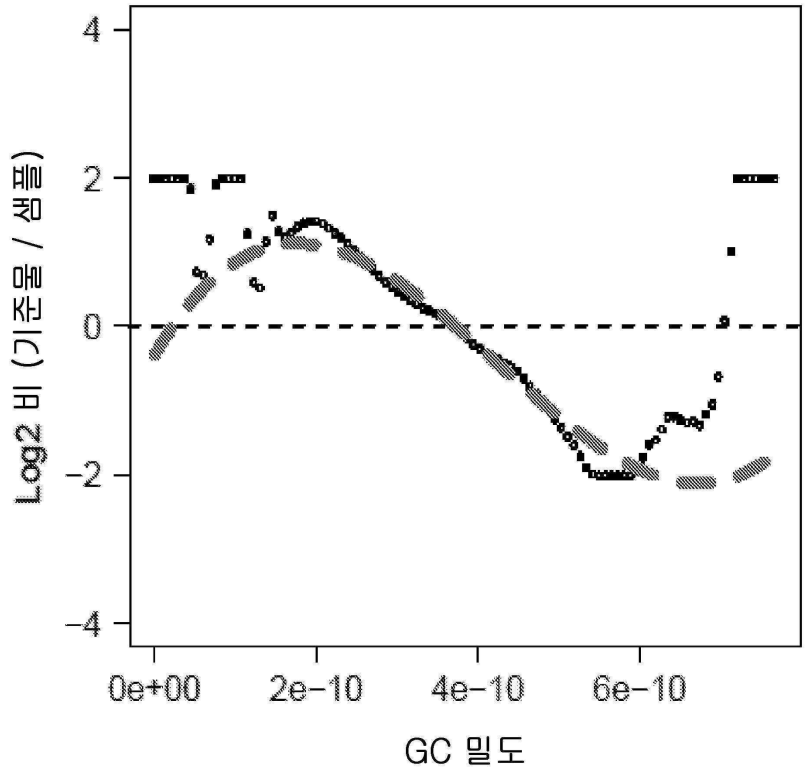
도면34


GC 밀도 (bw=200bp)
0e+00
3e-10
6e-10
124220
124240
124260
124280
HTRA1
게놈 위치 (kbp)

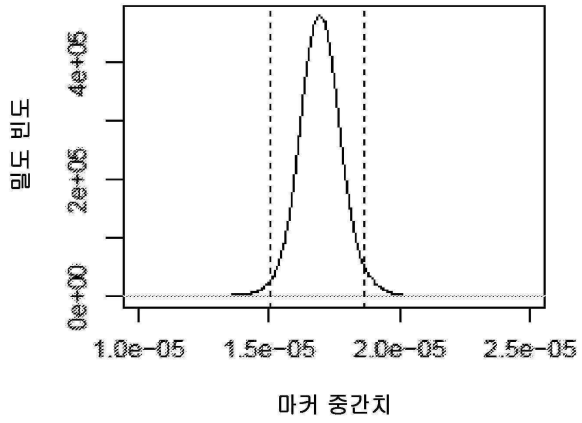
도면35


기준물
샘플
밀도 빈도
0e+00
4e+09
8e+09
0e+00
2e-10
4e-10
6e-10
8e-10
GC 밀도

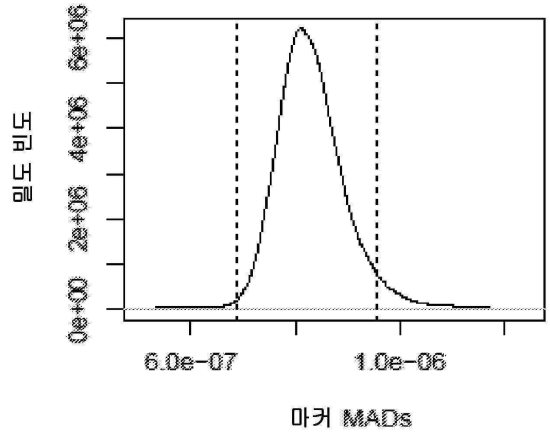
도면36


Log2 비 (기준물 / 샘플)
4
2
0
-2
-4
0e+00
2e-10
4e-10
6e-10
GC 밀도

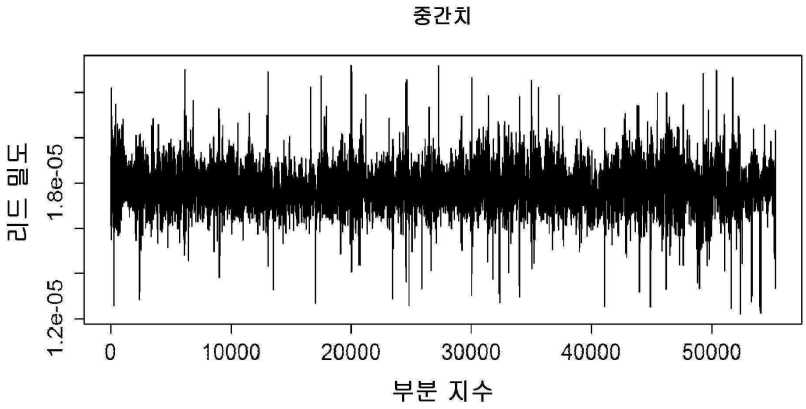
도면37a



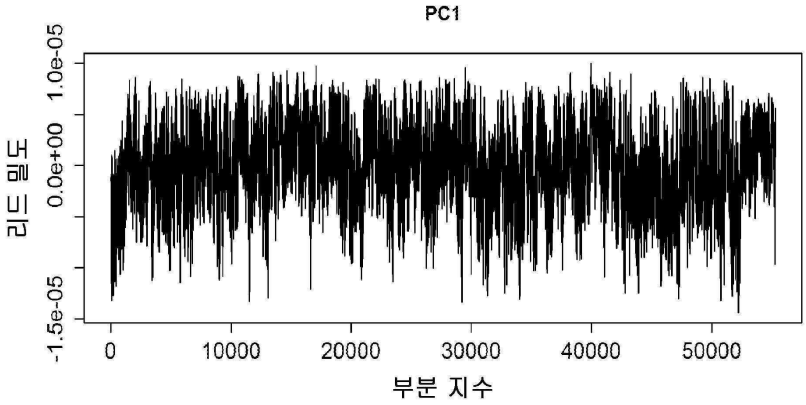
도면37b



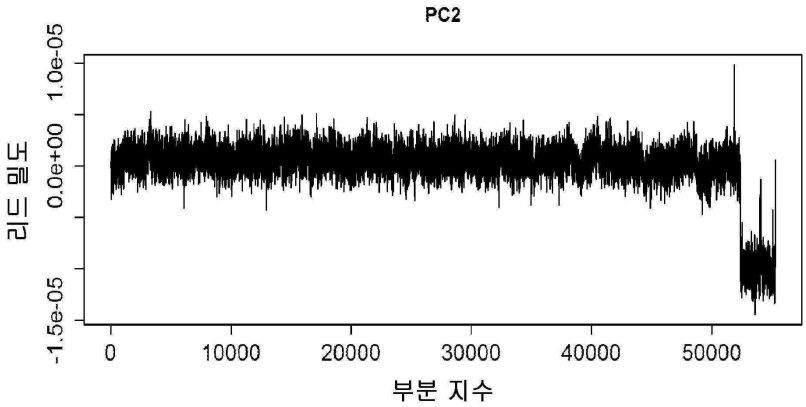
도면38a



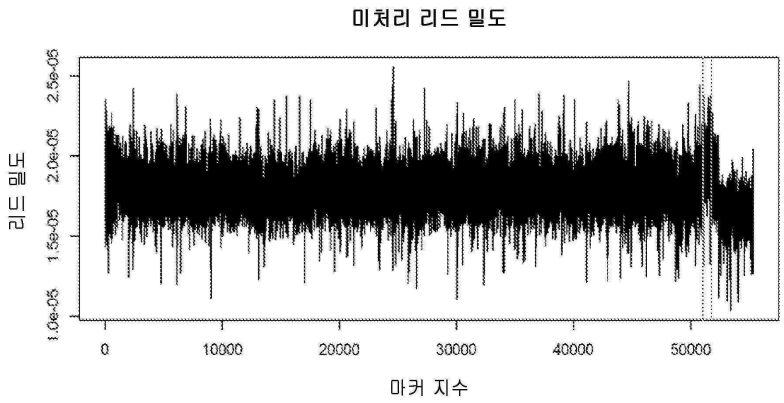
도면38b



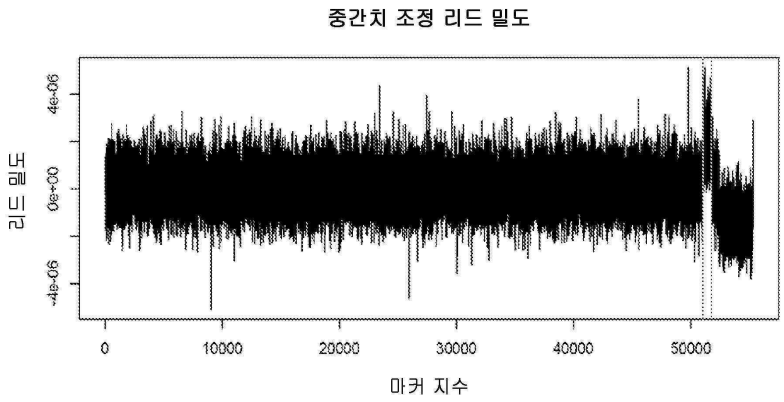
도면38c



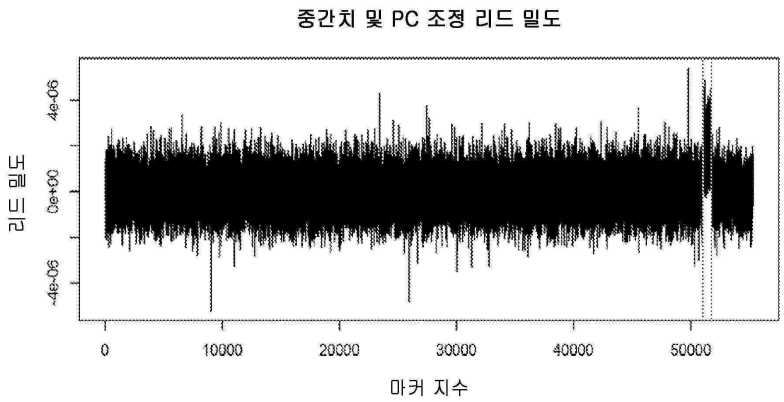
도면39a



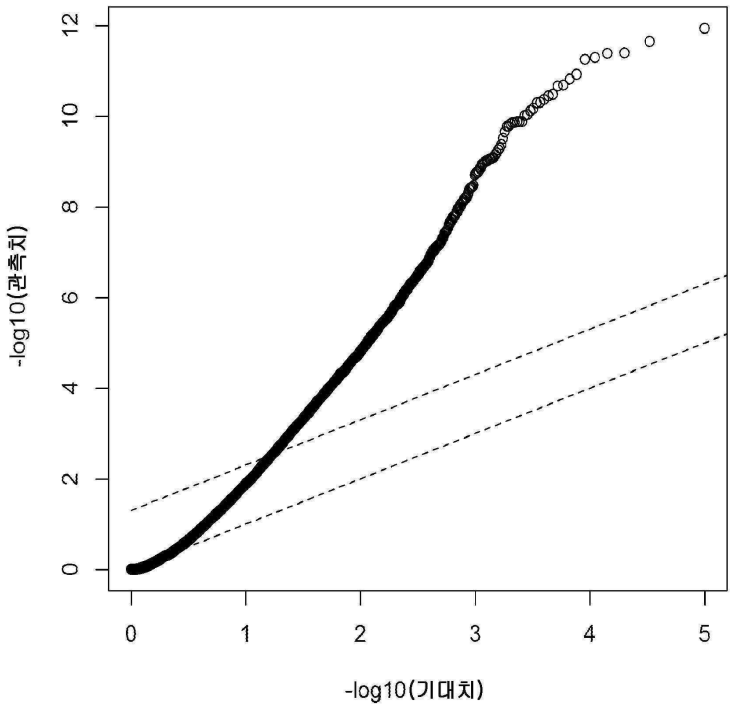
도면39b



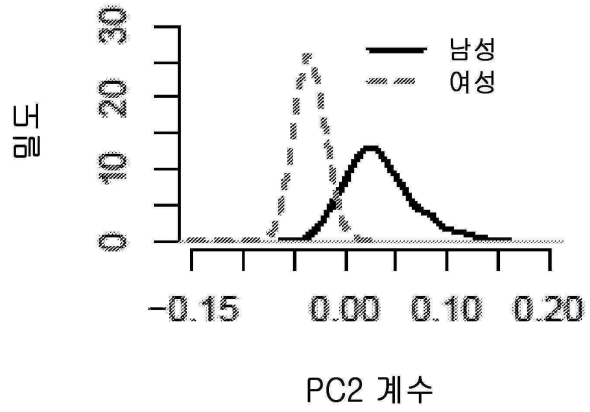
도면39c



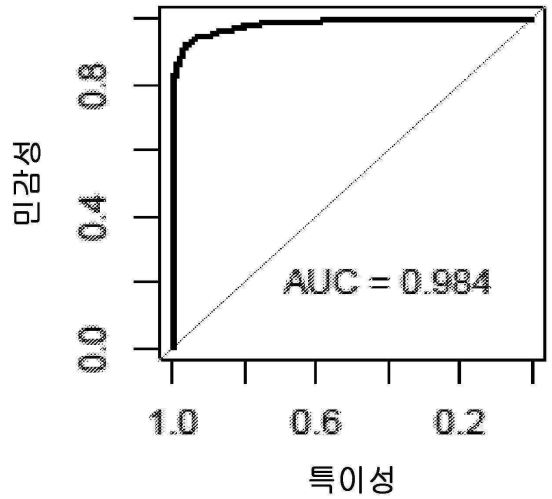
도면40



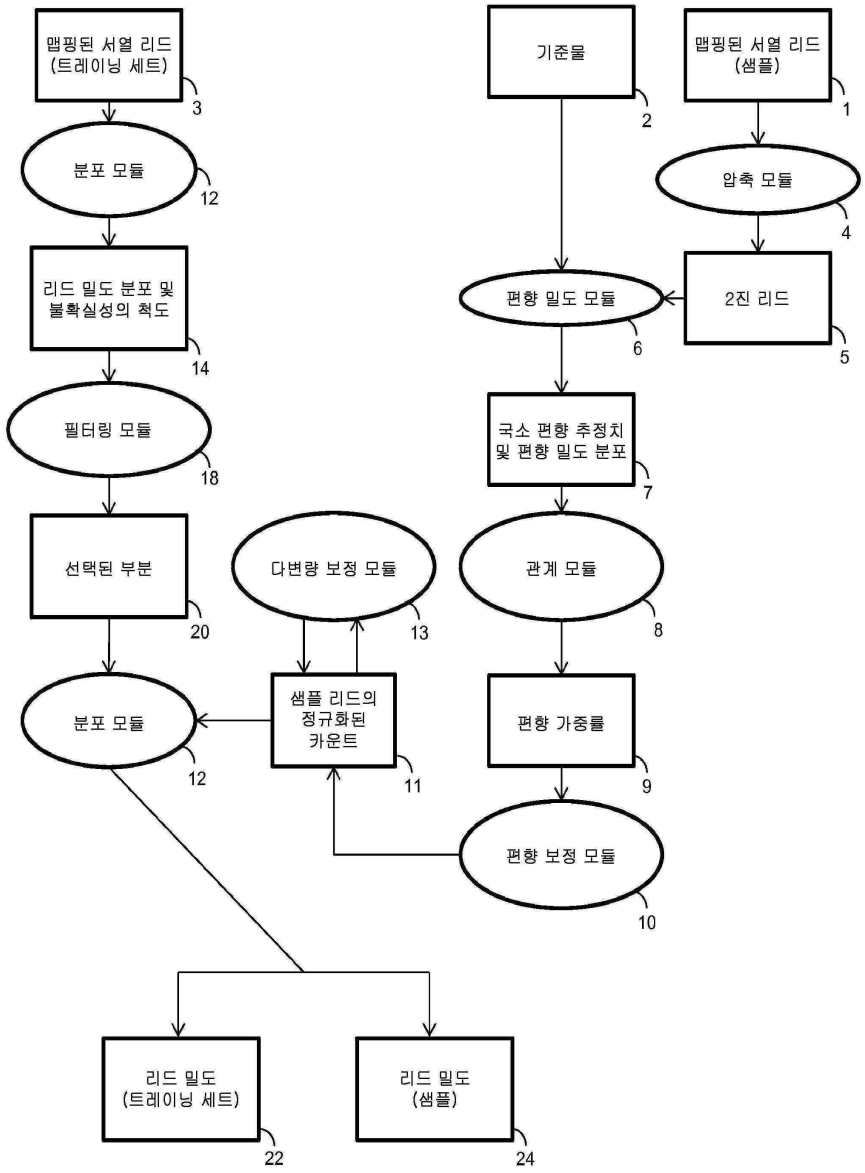
도면41a


밀도
0
10
20
30
남성
여성
−0.15
0.00
0.10
0.20
PC2 계수

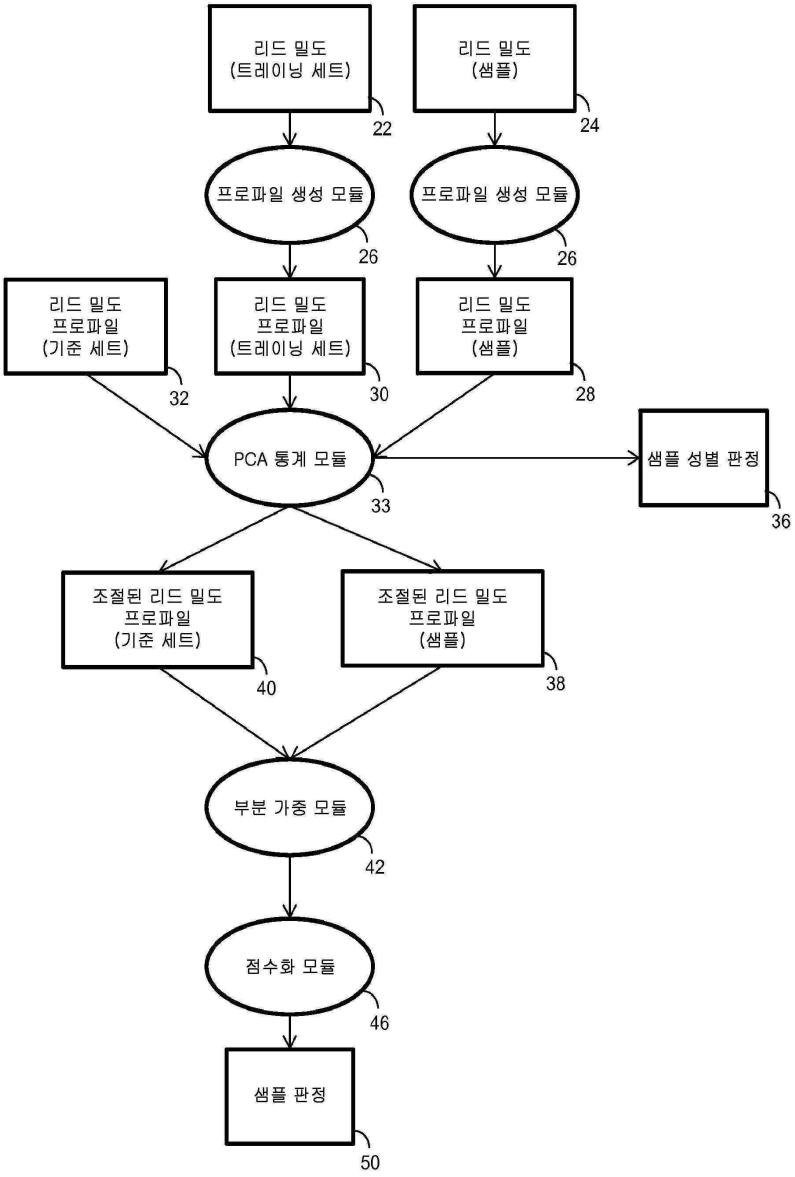
도면41b


민감성
0.0
0.4
0.8
AUC = 0.984
1.0
0.6
0.2
특이성

도면42a


맵핑된 서열 리드
(트레이닝 세트)
3
분포 모듈
12
리드 밀도 분포 및
불확실성의 척도
14
필터링 모듈
18
선택된 부분
20
분포 모듈
12
기준물
2
맵핑된 서열 리드
(샘플)
1
압축 모듈
4
2진 리드
5
편향 밀도 모듈
6
국소 편향 추정치
및 편향 밀도 분포
7
관계 모듈
8
편향 가중률
9
편향 보정 모듈
10
다변량 보정 모듈
13
샘플 리드의
정규화된
카운트
11
리드 밀도
(트레이닝 세트)
22
리드 밀도
(샘플)
24

도면42b


리드 밀도
(트레이닝 세트)
22
리드 밀도
(샘플)
24
프로파일 생성 모듈
26
프로파일 생성 모듈
26
리드 밀도
프로파일
(기준 세트)
32
리드 밀도
프로파일
(트레이닝 세트)
30
리드 밀도
프로파일
(샘플)
28
PCA 통계 모듈
33
샘플 성별 판정
36
조절된 리드 밀도
프로파일
(기준 세트)
40
조절된 리드 밀도
프로파일
(샘플)
38
부분 가중 모듈
42
점수화 모듈
46
샘플 판정
50