(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년03월11일
(11) 등록번호 10-2373647
(24) 등록일자 2022년03월08일

(51) 국제특허분류(Int. Cl.)
G16B 25/00 (2019.01) *G16B 30/00* (2019.01)
G16B 40/00 (2019.01)
(52) CPC특허분류
G16B 25/00 (2019.02)
G16B 30/00 (2019.02)
(21) 출원번호 10-2016-7013043
(22) 출원일자(국제) 2014년10월21일
심사청구일자 2019년09월23일
(85) 번역문제출일자 2016년05월18일
(65) 공개번호 10-2016-0073405
(43) 공개일자 2016년06월24일
(86) 국제출원번호 PCT/US2014/061635
(87) 국제공개번호 WO 2015/061359
국제공개일자 2015년04월30일
(30) 우선권주장
61/893,830 2013년10월21일 미국(US)
(56) 선행기술조사문헌
WO2013109981 A1*
H. Christina Fan 외, Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood, PNAS, 2008.10.21., Vol.105, No.42, pp.16266-16271.
Alberto Magi 외, Read count approach for DNA copy number variants detection, Bioinformatics, 2012.02.15., Vol.28, Issue.4, pp.470-478.
Eric Z. Chen 외, Noninvasive Prenatal Diagnosis of Fetal Trisomy 18 and Trisomy 13 by Maternal Plasma DNA Sequencing, PLOS ONE, 2011.07.06., Vol.6, Issue.7, pp.1-7.
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
베리나타 헬스, 인코포레이티드
미국, 캘리포니아 92122, 샌 디에고, 일루미나 웨이 5200
(72) 발명자
츄도바, 다르야 아이.
미국, 캘리포니아 95123, 산 호세, 타오르미노 애버뉴 5931
아브듀에바, 다이아나
미국, 캘리포니아 94563, 오린다, 오차드 알디. 227
라바, 리차드 피.
미국, 캘리포니아 94062, 레드우드 시티, 에지우드 알이. 711
(74) 대리인
강명구, 이경민

전체 청구항 수 : 총 62 항 심사관 : 정태수

(54) 발명의 명칭 **사본수 변동을 결정함에 있어서 검출의 감수성을 향상시키기 위한 방법**

(57) 요 약

다양한 의학적 상태와 연관되는 것으로 공지된 또는 의심되는 사본수 변동 (CNV)을 결정하기 위한 방법이 개시된다. 일부 구체예에서, 모계와 태아 무세포 DNA를 포함하는 모계 표본을 이용하여 태아의 사본수 변동 (CNV)을 결정하기 위한 방법이 제공된다. 일부 구체예에서, 다양한 의학적 상태와 연관되는 것으로 공지된 또는 의심되는

(뒷면에 계속)

대 표 도

CNV를 결정하기 위한 방법이 제공된다. 본원에서 개시된 일부 구체예는 표본내 GC-함량 바이어스를 제거함으로써, 서열 데이터 분석의 감수성 및/또는 특이성을 향상시키는 방법을 제공한다. 일부 구체예에서, 표본내 GC-함량 바이어스의 제거는 영향을 받지 않은 훈련 표본 전체에 대하여 통상적인 체계적 변동에 대해 교정된 서열 데이터에 근거된다. 또한, 관심되는 서열의 CNV의 평가를 위한 시스템 및 컴퓨터 프로그램 제품이 개시된다.

(52) CPC특허분류
G16B 40/00 (2019.02)

명 세 서

청구범위

청구항 1

시험 표본에서 관심되는 핵산 서열의 사본수를 평가하기 위해, 하나 또는 그 이상의 프로세서와 시스템 메모리를 포함하는 컴퓨터 시스템에서 실행된 방법에 있어서, 상기 방법은 다음 단계를 포함하는 것을 특징으로 하는 방법:

(a) 컴퓨터 시스템에서, 시험 표본으로부터 핵산 서열분석기에 의해 획득된 서열 리드를 제공하되, 상기 시험 표본이 하나 또는 그 이상의 유전체로부터 획득되는 핵산 분자를 포함하는 단계;

(b) 컴퓨터 시스템에 의해, 시험 표본의 서열 리드를 관심되는 핵산 서열을 포함하는 참조 유전체에 맞춰 정렬하여, 시험 서열 태그를 제공하는 단계;

(c) 컴퓨터 시스템에 의해, 각 빈에서 위치된 시험 서열 태그의 리드 커버리지를 결정하되, 참조 유전체가 복수의 빈으로 분할되고, 그리고 리드 커버리지가 빈에서 서열 태그의 존재비를 지시하는 단계;

(d) 컴퓨터 시스템에 의해, 관심되는 핵산 서열에 대한 전역 프로필을 제공하되, 전역 프로필이 각 빈에서 예상된 리드 커버리지를 포함하고, 그리고 예상된 리드 커버리지가 시험 표본과 실제적으로 동일한 방식으로 염기서열결정되고 정렬된 핵산 분자를 포함하는 영향을 받지 않은 훈련 표본의 훈련 세트로부터 획득되고, 예상된 리드 커버리지가 각 빈마다 변동을 표시하는 단계;

(e) 컴퓨터 시스템에 의해, 최소한 관심되는 핵산 서열의 각 빈에서 예상된 리드 커버리지를 이용하여 시험 서열 태그의 리드 커버리지를 조정하여, 관심되는 핵산 서열에 대한 전역-프로필-교정된 리드 커버리지를 획득하는 단계;

(f) 컴퓨터 시스템에 의해, 시험 표본의 GC 함량 수준 및 상기 시험 표본의 전역-프로필-교정된 리드 커버리지 사이의 관계에 근거하여 전역-프로필-교정된 리드 커버리지를 조정하여, 관심되는 핵산 서열에 대한 표본-GC-교정된 리드 커버리지를 획득하는 단계; 그리고

(g) 컴퓨터 시스템에 의해, 표본-GC-교정된 리드 커버리지에 근거하여 시험 표본에서 관심되는 핵산 서열의 사본수를 평가하되, 표본-GC-교정된 리드 커버리지가 신호 수준을 향상시키거나, 관심되는 핵산 서열의 사본수를 결정하기 위한 잡음 수준을 감소시키거나, 또는 둘 모두를 유발하는 단계.

청구항 2

청구항 1에 있어서, 서열 리드를 제공하기 전, 서열분석기를 이용하여 시험 표본으로부터 획득되는 핵산을 염기서열결정하여, 서열 리드를 산출하는 단계를 더욱 포함하는 것을 특징으로 하는 방법.

청구항 3

청구항 2에 있어서, 마커 핵산을 염기서열결정하기 전, 이들 핵산을 시험 표본과 결합하는 단계를 더욱 포함하는 것을 특징으로 하는 방법.

청구항 4

청구항 3에 있어서, 마커 핵산은 자연발생 데옥시리보핵산, 자연발생 리보핵산, 펩티드 핵산 (PNA), 모르폴리노 핵산, 잠금된 핵산, 글리콜 핵산, 트레오스 핵산, 그리고 이들의 임의의 조합으로 구성된 군에서 선택되는 것을 특징으로 하는 방법.

청구항 5

청구항 1에 있어서, 서열 리드는 임신 여성의 무세포 DNA 및 임신 여성에 의해 잉태된 태아의 무세포 DNA의 서열로부터 획득된 것을 특징으로 하는 방법.

청구항 6

청구항 1에 있어서, 마스크가 적용된 빈에서 리드 커버리지를 고려 사항으로부터 배제하는 서열 마스크를 적용하는 단계를 더욱 포함하는 것을 특징으로 하는 방법.

청구항 7

청구항 6에 있어서, 서열 마스크는 다음 단계를 포함하는 방법에 의해 획득되는 것을 특징으로 하는 방법:

컴퓨터 시스템에서, 복수의 영향을 받지 않은 훈련 표본으로부터 획득되는 서열 리드를 포함하는 훈련 세트를 제공하는 단계;

컴퓨터 시스템에 의해, 훈련 세트의 서열 리드를 참조 유전체에 맞춰 정렬하여, 훈련 표본에 대한 훈련 서열 태그를 제공하는 단계;

컴퓨터 시스템에 의해, 참조 유전체를 복수의 빈으로 분할하는 단계;

컴퓨터 시스템에 의해, 각 훈련 표본을 위한 각 빈에서 훈련 서열 태그의 리드 커버리지를 결정하는 단계; 그리고

컴퓨터 시스템에 의해, 마스크가 적용되지 않은 빈 및 마스크가 적용된 빈을 포함하는 서열 마스크를 창출하되, 마스크가 적용된 빈 각각이 마스크 적용 역치를 초과하는 분포 지수를 갖고, 상기 분포 지수가 훈련 표본의 리드 커버리지의 분포에 관련되는 단계.

청구항 8

청구항 7에 있어서, 서열 마스크를 창출하기 전, 각 빈에서 예상된 리드 커버리지에 따라 훈련 서열 태그의 리드 커버리지를 조정하여, 이들 빈에서 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 획득하는 단계를 더욱 포함하고, 이들 리드 커버리지는 이후, 서열 마스크를 창출하는데 이용되는 것을 특징으로 하는 방법.

청구항 9

청구항 7에 있어서, 분포 지수는 훈련 표본의 리드 커버리지의 분산에 수학적으로 관련된 것을 특징으로 하는 방법.

청구항 10

청구항 9에 있어서, 분포 지수는 변동 계수인 것을 특징으로 하는 방법.

청구항 11

청구항 6에 있어서, 관심되는 핵산 서열에서 사용된 마스크가 적용된 빈은 첫 번째 마스크 적용 역치를 갖고, 그리고 정규화 서열에서 사용된 마스크가 적용된 빈은 두 번째 마스크 적용 역치를 갖는 것을 특징으로 하는 방법.

청구항 12

청구항 11에 있어서, 첫 번째 마스크 적용 역치와 두 번째 마스크 적용 역치의 조합은 다른 역치를 이용하여 획득된 서열 마스크보다, 영향을 받지 않은 표본에서 관심되는 서열을 포함하는 영역에 걸쳐 리드 커버리지의 더욱 낮은 변동을 유발하는 서열 마스크를 제공하는 것을 특징으로 하는 방법.

청구항 13

청구항 6에 있어서, 서열 마스크는 빈 내에 훈련 표본 전체에 대하여 매핑 품질평가점수의 분포에 의해 규정된 마스크가 적용된 빈 및 마스크가 적용되지 않은 빈을 포함하고, 매핑 품질평가점수는 복수의 영향을 받지 않은 훈련 표본의 서열 리드를 참조 유전체에 맞춰 정렬하는 것으로부터 도출된 것을 특징으로 하는 방법.

청구항 14

청구항 1 내지 13 중 어느 한 항에 있어서, 작업 (g)에서 시험 표본에서 관심되는 핵산 서열의 사본수를 평가하는 것은 정규화 서열의 리드 커버리지 정보를 이용하여, 시험 표본에 대한 관심되는 핵산 서열의 서열 도스를 계산하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 15

청구항 14에 있어서, 서열 도스를 계산하는 것은 관심되는 핵산 서열에서 시험 서열 태그의 표본-GC-교정된 리드 커버리지를 정규화 서열에서 시험 서열 태그의 표본-GC-교정된 리드 커버리지로 나눗셈하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 16

청구항 15에 있어서, 정규화 서열은 하나 또는 그 이상의 로버스트 상염색체 서열 또는 이들의 분절을 포함하는 것을 특징으로 하는 방법.

청구항 17

청구항 1 내지 13 중 어느 한 항에 있어서, 작업 (g)에서 시험 표본에서 관심되는 핵산 서열의 사본수를 평가하는 것은 정규화 서열의 리드 커버리지 정보를 이용하여, 정규화된 염색체 값 또는 시험 표본에 대한 관심되는 핵산 서열의 정규화된 분절 값을 계산하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 18

청구항 1 내지 13 중 어느 한 항에 있어서, 시험 표본은 2개의 상이한 유전체로부터 획득되는 핵산의 혼합물을 포함하는 것을 특징으로 하는 방법.

청구항 19

청구항 18에 있어서, 상기 핵산은 무세포 DNA 분자를 포함하는 것을 특징으로 하는 방법.

청구항 20

청구항 1 내지 13 중 어느 한 항에 있어서, 시험 표본은 태아와 모계 무세포 핵산을 포함하는 것을 특징으로 하는 방법.

청구항 21

청구항 1 내지 13 중 어느 한 항에 있어서, 시험 표본은 2개 또는 그 이상의 태아로부터 획득되는 태아 무세포 핵산을 포함하는 것을 특징으로 하는 방법.

청구항 22

청구항 1 내지 13 중 어느 한 항에 있어서, 시험 표본은 동일한 개체로부터 유래된 암성 세포 및 영향을 받지 않은 세포로부터 획득되는 핵산을 포함하는 것을 특징으로 하는 방법.

청구항 23

청구항 1 내지 13 중 어느 한 항에 있어서, 시험 표본에서 관심되는 핵산 서열의 사본수를 평가하는 것은 완전한 또는 부분적인 태아 이수성의 존재 또는 부재를 결정하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 24

청구항 1 내지 13 중 어느 한 항에 있어서, 작업 (f) 후, 사본수 변동의 평가에서 고려 사항으로부터 표본-GC-교정된 리드 커버리지의 이상점 빈을 제거하는 단계를 더욱 포함하는 것을 특징으로 하는 방법.

청구항 25

청구항 24에 있어서, 이상점 빈은 중앙 표본-GC-교정된 리드 커버리지가 모든 빈의 중앙값으로부터 약 1 중위 절대 편차보다 큰 빈을 포함하는 것을 특징으로 하는 방법.

청구항 26

청구항 1 내지 13 중 어느 한 항에 있어서, 각 빈에서 예상된 리드 커버리지는 훈련 표본의 리드 커버리지의 중앙값 또는 평균을 포함하고, 그리고 여기서 작업 (e)에서 시험 서열 태그의 리드 커버리지를 조정하는 것은 각 빈에 대한 시험 서열 태그의 리드 커버리지를, 상기 빈으로부터 훈련 표본의 리드 커버리지의 중앙값 또는 평균

으로 나눗셈하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 27

청구항 1 내지 13 중 어느 한 항에 있어서, 작업 (e)에서 시험 서열 태그의 리드 커버리지를 조정하는 것은 (i) 하나 또는 그 이상의 로버스트 염색체 또는 영역 내에 복수의 빈에서 예상된 리드 커버리지와 시험 서열 태그의 리드 커버리지 사이에 수학적 관계를 획득하고, 그리고 (ii) 수학적 관계를 관심되는 서열 내에 빈에 적용하여 전역-프로필-교정된 리드 커버리지를 획득하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 28

청구항 27에 있어서,

(i)에서 관계는 선형 회귀에 의해 획득되고:

y_a = *인터셉트* + *기울기* * gwp_a

여기서 *ya*는 하나 또는 그 이상의 로버스트 염색체 또는 영역에서 시험 표본을 위한 빈 *a*의 리드 커버리지이고, 그리고 *gwpa*는 영향을 받지 않은 훈련 표본을 위한 빈 *a*에 대한 전역 프로필이고; 그리고

(ii)에서 전역-프로필-교정된 리드 커버리지를 획득하는 것은 아래와 같은 전역-프로필-교정된 리드 커버리지 z_b를 획득하는 것을 포함하고:

$z_b = y_b / (\text{인터셉트} + \text{기울기} * gwp_b) - 1$

여기서 *yb*는 관심되는 서열에서 시험 표본을 위한 빈 b의 관찰된 리드 커버리지이고, 그리고 *gwpb*는 영향을 받지 않은 훈련 표본을 위한 빈 b에 대한 전역 프로필인 것을 특징으로 하는 방법.

청구항 29

청구항 1 내지 13 중 어느 한 항에 있어서, (e)로부터 시험 서열 태그의 전역-프로필-교정된 리드 커버리지는 관심되는 핵산 서열에서 빈의 전역-프로필-교정된 리드 커버리지 및 정규화 서열에서 빈의 전역-프로필-교정된 리드 커버리지를 포함하는 것을 특징으로 하는 방법.

청구항 30

청구항 1 내지 13 중 어느 한 항에 있어서, 작업 (f)에서 전역-프로필-교정된 리드 커버리지를 조정하는 것은

참조 유전체에서 빈을 복수의 GC 군으로 그룹화하고, 각 GC 군은 복수 빈을 포함하고, 여기서 이들 복수 빈은 시험 서열 태그를 내포하고 유사한 GC 함량을 갖고;

복수의 로버스트 상염색체에 대한 각 GC 군에 대한 전역-프로필-교정된 리드 커버리지의 예상된 리드 커버리지를 결정하고; 그리고

동일한 GC 군의 결정된 예상된 리드 커버리지에 근거하여 각 GC 군에 대한 시험 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하여, 관심되는 핵산 서열에서 시험 서열 태그의 표본-GC-교정된 리드 커버리지를 획득하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 31

청구항 30에 있어서, 전역-프로필-교정된 리드 커버리지의 예상된 리드 커버리지는 복수의 로버스트 상염색체의 GC 군에 대한 리드 커버리지의 평균 또는 중앙값인 것을 특징으로 하는 방법.

청구항 32

청구항 30에 있어서, 시험 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하는 것은 예상된 리드 커버리지를 전역-프로필-교정된 리드 커버리지로부터 감산하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 33

청구항 1 내지 13 중 어느 한 항에 있어서, 작업 (f)에서 전역-프로필-교정된 리드 커버리지를 조정하는 것은

선형 또는 비선형 수학 함수를 복수의 로버스트 상염색체로부터 획득되는 데이터 포인트에 적합시키고, 여기서 각 데이터 포인트는 리드 커버리지 값을 GC 함량 값에 관련시키고;

고려 중인 빈의 GC 함량 값에서 수학 함수의 리드 커버리지 값에 동등한, 각 빈에 대한 예상된 리드 커버리지에 근거하여 각 빈에서 시험 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 34

청구항 33에 있어서, 시험 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하는 것은 예상된 리드 커버리지를 전역-프로필-교정된 리드 커버리지로부터 감산하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 35

청구항 30에 있어서, 로버스트 상염색체는 관심되는 염색체(들)를 제외한 모든 상염색체를 포함하는 것을 특징으로 하는 방법.

청구항 36

청구항 30에 있어서, 로버스트 상염색체는 chr X, Y, 13, 18, 그리고 21을 제외한 모든 상염색체를 포함하는 것을 특징으로 하는 방법.

청구항 37

청구항 30에 있어서, 로버스트 상염색체는 정상적인 이배체 상태로부터 일탈하는 것으로 시험 표본으로부터 결정된 것들을 제외한 모든 상염색체를 포함하는 것을 특징으로 하는 방법.

청구항 38

청구항 1 내지 13 중 어느 한 항에 있어서, 복수의 영향을 받지 않은 개체, 시험 표본, 또는 둘 모두로부터 무세포 DNA를 추출하는 단계를 더욱 포함하는 것을 특징으로 하는 방법.

청구항 39

청구항 1 내지 13 중 어느 한 항에 있어서, 서열 리드는 개체의 전체 유전체 내에서 획득된 약 20 내지 50-bp의 서열을 포함하는 것을 특징으로 하는 방법.

청구항 40

청구항 1 내지 13 중 어느 한 항에 있어서, (a)의 서열 리드는 바코드화된 25-mer을 포함하는 것을 특징으로 하는 방법.

청구항 41

청구항 1 내지 13 중 어느 한 항에 있어서, 시험 서열 태그와 훈련 서열 태그의 리드 커버리지는 마스크가 적용되지 않은 부위 수치 (NES 수치)에 근거되고, 여기서 NES 수치는 마스크가 적용되지 않은 부위에 매핑된 비다중 서열 태그의 개수인 것을 특징으로 하는 방법.

청구항 42

청구항 41에 있어서, NES 수치는 마스크가 적용되지 않은 부위에 매핑된 독특하게 정렬된, 비다중 서열 태그의 개수인 것을 특징으로 하는 방법.

청구항 43

청구항 1 내지 13 중 어느 한 항에 있어서, 빈 크기는 약 1000 bp와 1,000,000 bp 사이인 것을 특징으로 하는 방법.

청구항 44

청구항 1 내지 13 중 어느 한 항에 있어서, 빈 크기는 약 100,000 bp인 것을 특징으로 하는 방법.

청구항 45

청구항 1 내지 13 중 어느 한 항에 있어서, 시험 표본의 서열 리드의 개수를 이용한 계산에 의해 빈 크기를 결정하는 단계를 더욱 포함하는 것을 특징으로 하는 방법.

청구항 46

청구항 45에 있어서, 각 빈에서 서열 태그의 개수는 최소한 약 1000 bp인 것을 특징으로 하는 방법.

청구항 47

관심되는 핵산 서열의 사본수의 평가를 위한 서열 마스크를 창출하기 위한, 하나 또는 그 이상의 프로세서와 시스템 메모리를 포함하는 컴퓨터 시스템에서 실행된 방법에 있어서, 상기 방법은 다음 단계를 포함하는 것을 특징으로 하는 방법:

(a) 컴퓨터 시스템에서, 복수의 영향을 받지 않은 훈련 표본으로부터 획득되는 서열 리드를 포함하는 훈련 세트를 제공하는 단계;

(b) 컴퓨터 시스템에 의해, 훈련 세트의 서열 리드를 관심되는 핵산 서열을 포함하는 참조 유전체에 맞춰 정렬하여, 훈련 표본에 대한 훈련 서열 태그를 제공하는 단계;

(c) 컴퓨터 시스템에 의해, 참조 유전체를 복수의 빈으로 분할하는 단계;

(d) 컴퓨터 시스템에 의해, 각 훈련 표본을 위한 각 빈에서 훈련 서열 태그의 리드 커버리지를 영향을 받지 않은 훈련 표본 각각에 대해 결정하되, 리드 커버리지가 빈에서 서열 태그의 존재비를 지시하는 단계;

(e) 모든 훈련 표본 전체에 대하여 훈련 서열 태그의 예상된 리드 커버리지를 각 빈에 대해 결정하는 단계;

(f) 컴퓨터 시스템에 의해, 각 빈에서 예상된 리드 커버리지에 따라 각 훈련 표본을 위한 각 빈에서 훈련 서열 태그의 리드 커버리지를 조정하여, 각 훈련 표본을 위한 빈에서 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 획득하는 단계;

(g) 컴퓨터 시스템에 의해, 각 훈련 표본의 GC 함량 수준 및 상기 훈련 표본의 전역-프로필-교정된 리드 커버리지 사이의 관계에 근거하여 각 훈련 표본에 대한 전역-프로필-교정된 리드 커버리지를 조정하여, 표본-GC-교정된 리드 커버리지를 획득하는 단계; 그리고

(h) 컴퓨터 시스템에 의해, 각 빈에서 훈련 표본 전체에 대하여 표본-GC-교정된 리드 커버리지의 변동에 근거하여 참조 유전체 전체에 대하여 마스크가 적용되지 않은 빈 및 마스크가 적용된 빈을 포함하는 서열 마스크를 창출하는 단계.

청구항 48

청구항 47에 있어서, 각 빈에 대해 (e)에서 결정된 예상된 리드 커버리지는 훈련 표본의 리드 커버리지의 중앙값 또는 평균을 포함하는 것을 특징으로 하는 방법.

청구항 49

청구항 48에 있어서, 작업 (f)에서 훈련 서열 태그의 리드 커버리지를 조정하는 것은 각 훈련 표본의 각 빈에 대한 훈련 서열 태그의 리드 커버리지로부터 중앙값 또는 평균을 감산하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 50

청구항 48에 있어서, 작업 (f)에서 훈련 서열 태그의 리드 커버리지를 조정하는 것은 각 훈련 표본의 각 빈에 대한 훈련 서열 태그의 리드 커버리지를 중앙값 또는 평균으로 나눗셈하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 51

청구항 47에 있어서, 관심되는 핵산 서열에서 사용된 마스크가 적용된 빈은 첫 번째 마스크 적용 역치를 갖고, 그리고 정규화 서열에서 사용된 마스크가 적용된 빈은 두 번째 마스크 적용 역치를 갖는 것을 특징으로 하는 방법.

청구항 52

청구항 51에 있어서, 첫 번째 마스크 적용 역치와 두 번째 마스크 적용 역치의 조합은 다른 역치를 이용하여 획득된 서열 마스크보다, 영향을 받지 않은 표본에서 관심되는 서열을 포함하는 영역에 걸쳐 리드 커버리지의 더욱 낮은 변동을 유발하는 서열 마스크를 제공하는 것을 특징으로 하는 방법.

청구항 53

삭제

청구항 54

청구항 47에 있어서, 각 훈련 표본에 대한 전역-프로필-교정된 리드 커버리지를 조정하는 것은

참조 유전체에서 모든 빈을 복수의 GC 군으로 그룹화하고, 각 GC 군은 유사한 GC 함량을 갖는 복수 빈을 포함하고;

복수의 로버스트 상염색체에 대한 각 GC 군에 대한 전역-프로필-교정된 리드 커버리지의 예상된 리드 커버리지를 결정하고; 그리고

동일한 GC 군의 결정된 예상된 리드 커버리지에 근거하여 각 GC 군에 대한 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하여, 관심되는 핵산 서열에서 훈련 서열 태그의 표본-GC-교정된 리드 커버리지를 획득하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 55

청구항 54에 있어서, 전역-프로필-교정된 리드 커버리지의 예상된 리드 커버리지는 복수의 로버스트 상염색체의 GC 군에 대한 리드 커버리지의 평균 또는 중앙값인 것을 특징으로 하는 방법.

청구항 56

청구항 54에 있어서, 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하는 것은 예상된 리드 커버리지를 전역-프로필-교정된 리드 커버리지로부터 감산하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 57

청구항 47에 있어서, 각 훈련 표본에 대한 전역-프로필-교정된 리드 커버리지를 조정하는 것은

선형 또는 비선형 수학 함수를 복수의 로버스트 상염색체로부터 획득되는 데이터 포인트에 적합시키고, 여기서 각 데이터 포인트는 리드 커버리지 값을 GC 함량 값에 관련시키고;

빈의 GC 함량 값에서 수학 함수의 리드 커버리지 값에 동등한, 각 빈에 대한 예상된 리드 커버리지에 근거하여 각 빈에서 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 58

청구항 57에 있어서, 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하는 것은 예상된 리드 커버리지를 전역-프로필-교정된 리드 커버리지로부터 감산하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 59

청구항 1 또는 47에 있어서, 표본은 혈액 표본, 소변 표본, 또는 타액 표본인 것을 특징으로 하는 방법.

청구항 60

청구항 1 또는 47에 있어서, 표본은 혈장 표본인 것을 특징으로 하는 방법.

청구항 61

청구항 1 또는 47에 있어서, 각각의 관심되는 핵산 서열에 대해 확인된 서열 태그의 개수는 최소한 약 10,000개인 것을 특징으로 하는 방법.

청구항 62

시험 표본에서 관심되는 핵산 서열의 사본수의 평가를 위한 컴퓨터 시스템에 있어서, 상기 시스템은

프로세서; 그리고

다음 단계를 포함하는 방법을 이용하여 시험 표본에서 사본수를 평가하기 위해 상기 프로세서에서 이행을 위한 명령이 그 안에 저장된 하나 또는 그 이상의 컴퓨터-판독가능 저장 매체를 포함하는 것을 특징으로 하는 시스템:

(a) 컴퓨터 시스템에서, 시험 표본의 서열 리드를 제공하는 단계;

(b) 컴퓨터 시스템에 의해, 시험 표본의 서열 리드를 관심되는 핵산 서열을 포함하는 참조 유전체에 맞춰 정렬하여, 시험 서열 태그를 제공하는 단계;

(c) 컴퓨터 시스템에 의해, 각 빈에서 위치된 시험 서열 태그의 리드 커버리지를 결정하되, 참조 유전체가 복수의 빈으로 분할되는 단계;

(d) 컴퓨터 시스템에 의해, 관심되는 핵산 서열에 대한 전역 프로필을 제공하되, 전역 프로필이 각 빈에서 예상된 리드 커버리지를 포함하고, 그리고 예상된 리드 커버리지가 시험 표본과 실제적으로 동일한 방식으로 염기서열결정되고 정렬된 영향을 받지 않은 훈련 표본의 훈련 세트로부터 획득되고, 예상된 리드 커버리지가 각 빈마다 변동을 표시하는 단계;

(e) 컴퓨터 시스템에 의해, 각 빈에서 예상된 리드 커버리지에 따라 시험 서열 태그의 리드 커버리지를 조정하여, 시험 서열 태그의 각 빈에서 전역-프로필-교정된 리드 커버리지를 획득하는 단계;

(f) 컴퓨터 시스템에 의해, 시험 표본의 GC 함량 수준 및 시험 서열 태그의 빈에 대한 상기 시험 표본의 전역-프로필-교정된 리드 커버리지 사이의 관계에 근거하여 전역-프로필-교정된 리드 커버리지를 조정하여, 관심되는 핵산 서열에서 시험 서열 태그의 표본-GC-교정된 리드 커버리지를 획득하는 단계; 그리고

(g) 컴퓨터 시스템에 의해, 표본-GC-교정된 리드 커버리지에 근거하여 시험 표본에서 관심되는 핵산 서열의 사본수를 평가하는 단계.

청구항 63

컴퓨터 시스템의 하나 또는 그 이상의 프로세서에 의해 실행될 때, 컴퓨터 시스템이 태아와 모계 무세포 핵산을 포함하는 시험 표본에서 관심되는 염색체 또는 핵산 서열의 사본수의 평가를 위한 방법을 실행하도록 유발하는 컴퓨터-실행가능 명령이 그 안에 저장된 컴퓨터-판독가능 비-일시적 저장 매체에 있어서, 상기 방법은 다음 단계를 포함하는 것을 특징으로 하는 비-일시적 저장 매체:

(a) 컴퓨터 시스템에서, 시험 표본으로부터 핵산 서열분석기에 의해 획득된 서열 리드를 제공하되, 상기 시험 표본이 하나 또는 그 이상의 유전체로부터 획득되는 핵산 분자를 포함하는 단계;

(b) 컴퓨터 시스템에 의해, 시험 표본의 서열 리드를 관심되는 핵산 서열을 포함하는 참조 유전체에 맞춰 정렬하여, 시험 서열 태그를 제공하는 단계;

(c) 컴퓨터 시스템에 의해, 각 빈에서 위치된 시험 서열 태그의 리드 커버리지를 결정하되, 참조 유전체가 복수의 빈으로 분할되고, 그리고 리드 커버리지가 빈에서 서열 태그의 존재비를 지시하는 단계;

(d) 컴퓨터 시스템에 의해, 관심되는 핵산 서열에 대한 전역 프로필을 제공하되, 전역 프로필이 각 빈에서 예상된 리드 커버리지를 포함하고, 그리고 예상된 리드 커버리지가 시험 표본과 실제적으로 동일한 방식으로 염기서열결정되고 정렬된 핵산 분자를 포함하는 영향을 받지 않은 훈련 표본의 훈련 세트로부터 획득되고, 예상된 리드 커버리지가 각 빈마다 변동을 표시하는 단계;

(e) 컴퓨터 시스템에 의해, 최소한 관심되는 핵산 서열의 각 빈에서 예상된 리드 커버리지를 이용하여 시험 서열 태그의 리드 커버리지를 조정하여, 관심되는 핵산 서열에 대한 전역-프로필-교정된 리드 커버리지를 획득하

는 단계;

(f) 컴퓨터 시스템에 의해, 시험 표본의 GC 함량 수준 및 상기 시험 표본의 전역-프로필-교정된 리드 커버리지 사이의 관계에 근거하여 전역-프로필-교정된 리드 커버리지를 조정하여, 관심되는 핵산 서열에 대한 표본-GC-교정된 리드 커버리지를 획득하는 단계; 그리고

(g) 컴퓨터 시스템에 의해, 표본-GC-교정된 리드 커버리지에 근거하여 시험 표본에서 관심되는 핵산 서열의 사본수를 평가하되, 표본-GC-교정된 리드 커버리지가 신호 수준을 향상시키거나, 관심되는 핵산 서열의 사본수를 결정하기 위한 잡음 수준을 감소시키거나, 또는 둘 모두를 유발하는 단계.

발명의 설명

기 술 분 야

[0001] **관련된 출원에 대한 교차 참조**

[0002] 본 출원은 35 U.S.C. § 119(e) 하에, 2013년 10월 21자 제출된, U.S. 특허가출원 번호 61/893,830, 발명의 명칭: 사본수 변동을 결정함에 있어서 검출의 감수성을 향상시키기 위한 방법에 우선권을 주장하고, 이것은 모든 점에서 전체적으로 본원에 참조로서 편입된다.

배 경 기 술

[0003] **배경**

[0004] 인간 의학적 연구에서 결정적인 노력 중에서 한 가지는 불리한 건강 결과를 발생시키는 유전자 비정상의 발견이다. 많은 경우에, 특이적 유전자 및/또는 결정적인 진단적 마커가 비정상적인 사본수에서 존재하는 유전체의 부분에서 확인되었다. 가령, 출생전 진단에서, 전체 염색체의 추가 사본 또는 사본 결여가 빈번하게 발생하는 유전자 병변이다. 암에서, 전체 염색체 또는 염색체 분절의 사본의 결실 또는 증가, 그리고 유전체의 특정한 영역의 더욱 높은 수준 증폭은 통상적인 일이다.

[0005] 사본수 변동 (CNV)에 관한 대부분의 정보는 구조적 비정상의 인식을 허용하는 세포유전학적 분해능에 의해 제공되었다. 유전 선별검사 및 생물학적 선량측정을 위한 전통적인 절차는 핵형의 분석을 위한 세포를 획득하기 위해, 침습성 시술, 예를 들면, 양수천자, 탯줄천자, 또는 융모막 융모 표본추출 (CVS)을 활용하였다. 세포 배양을 필요로 하지 않는 더욱 신속한 시험 방법에 대한 필요를 인식하여, 형광 제자리 혼성화 (FISH), 정량적 형광 PCR (QF-PCR) 및 어레이-비교 유전체 혼성화 (어레이-CGH)가 사본수 변동의 분석을 위한 분자-세포유전학적 방법으로서 개발되었다.

[0006] 인간 의학적 연구에서 결정적인 노력 중에서 한 가지는 불리한 건강 결과를 발생시키는 유전자 비정상의 발견이다. 많은 경우에, 특이적 유전자 및/또는 결정적인 진단적 마커가 비정상적인 사본수에서 존재하는 유전체의 부분에서 확인되었다. 가령, 출생전 진단에서, 전체 염색체의 추가 사본 또는 사본 결여가 빈번하게 발생하는 유전자 병변이다. 암에서, 전체 염색체 또는 염색체 분절의 사본의 결실 또는 증가, 그리고 유전체의 특정한 영역의 더욱 높은 수준 증폭은 통상적인 일이다.

[0007] 사본수 변동 (CNV)에 관한 대부분의 정보는 구조적 비정상의 인식을 허용하는 세포유전학적 분해능에 의해 제공되었다. 유전 선별검사 및 생물학적 선량측정을 위한 전통적인 절차는 핵형의 분석을 위한 세포를 획득하기 위해, 침습성 시술, 예를 들면, 양수천자, 탯줄천자, 또는 융모막 융모 표본추출 (CVS)을 활용하였다. 세포 배양을 필요로 하지 않는 더욱 신속한 시험 방법에 대한 필요를 인식하여, 형광 제자리 혼성화 (FISH), 정량적 형광 PCR (QF-PCR) 및 어레이-비교 유전체 혼성화 (어레이-CGH)가 사본수 변동의 분석을 위한 분자-세포유전학적 방법으로서 개발되었다.

[0008] 상대적으로 짧은 시간 내에 전체 유전체를 염기서열결정하는 것을 허용하는 기술의 출현, 그리고 순환하는 무세포 DNA (cfDNA)의 발견은 침습성 표본추출 방법과 연관된 위험 없이, 비교되는 한 염색체로부터 유래하는 유전 물질을 다른 것의 유전 물질과 비교하는 기회를 제공하였고, 이것은 관심되는 유전자 서열의 다양한 종류의 사본수 변동을 진단하는 도구를 제공한다.

[0009] 일부 적용에서 사본수 변동 (CNV)의 진단은 고조된 기술적인 과제를 수반한다. 가령, 이란성 쌍둥이 다태 (또는

다중접합자) 임신에 대한 CNV의 비침습성 출생전 진단 (NIPD)은 단일 임신보다 어려운데, 그 이유는 태아 cfDNA의 전체 분율이 태아의 숫자에 비례하여 변하지 않고, cfDNA의 태아 분율을 태아의 숫자의 차수에 의해 낮추는데, 이것은 차례로, 분석 동안 신호 대 잡음 비율을 감소시킨다. 추가적으로, Y 염색체 기초된 진단, 예를 들면, 성별 확인은 Y 염색체에 관련된 제약에 의해 영향을 받는다. 구체적으로, Y 염색체의 리드 커버리지는 상염색체의 리드 커버리지보다 낮고, 그리고 Y 염색체 상에서 반복된 서열은 그들의 정확한 위치까지 리드 (read)의 매핑을 복잡하게 만든다. 게다가, 일부 현재 염기서열결정 프로토콜은 극초단 리드, 예를 들면, 25mer 리드와 태그를 활용하고, 또 다른 정렬 과제를 나타내는데, 그 이유는 25mer 태그가 대부분의 편재성 반복적 요소의 전형적인 크기보다 짧기 때문이다. 본원에서 개시된 일부 구체예는 CNV의 평가를 위해 서열 데이터를 분석함에 있어서 감수성 및/또는 특이성을 향상시키는 방법을 제공한다.

[0010] 제한된 수준의 cfDNA에 기인하는 불충분한 감수성, 그리고 유전체 정보의 내재하는 성격에 기인하는 기술의 염기서열결정 바이어스를 비롯하여, 비침습성 출생전 진단학에서 현재 방법의 한계는 다양한 임상적 세팅에서 사본수 변화를 확실하게 진단하기 위해 특이성, 감수성, 그리고 적용가능성 중에서 한 가지 또는 모두를 제공할 비침습성 방법에 대한 지속적인 요구의 기초가 된다. 본원에서 개시된 구체예는 상기 요구 중에서 일부를 완수하고, 그리고 특히, 비침습성 출생전 진단학의 실시에 적용가능한 신뢰할 만한 방법을 제공한다.

발명의 내용

해결하려는 과제

[0011] **요약**

[0012] 일부 구체예에서, 임의의 태아 이수성의 사본수 변동 (CNV), 그리고 다양한 의학적 상태와 연관되는 것으로 공지된 또는 의심되는 CNV를 결정하기 위한 방법이 제공된다. 이들 방법은 유전체 서열의 GC 변동에 관련된 잡음과 오차를 감소시키기 위한 메커니즘을 포함한다. 본 발명 방법에 따라 결정될 수 있는 CNV는 염색체 1-22, X와 Y 중에서 임의의 한 가지 또는 그 이상의 삼염색체성과 일염색체성, 다른 염색체 다염색체성, 그리고 이들 염색체 중에서 임의의 한 가지 또는 그 이상의 분절의 결실 및/또는 중복을 포함한다.

[0013] 다른 구체예는 시험 표본에서 관심되는 핵산 서열, 예를 들면, 임상적으로 유관한 서열의 사본수 변동 (CNV)을 확인하기 위한 방법을 제공한다. 상기 방법은 완전한 염색체 또는 염색체의 분절 대신에, 관심되는 서열의 사본수 변동을 사정한다.

[0014] 일부 구체예에서, 상기 방법은 하나 또는 그 이상의 유전체의 핵산을 포함하는 시험 표본에서 관심되는 핵산 서열의 사본수를 평가하기 위해, 하나 또는 그 이상의 프로세서와 시스템 메모리를 포함하는 컴퓨터 시스템에서 실행된다. 상기 방법은 다음을 포함한다: (a) 시험 표본으로부터 핵산 서열분석기에 의해 획득된 서열 리드를 제공하고; (b) 시험 표본의 서열 리드를 관심되는 핵산 서열을 포함하는 참조 유전체에 맞춰 정렬하고, 따라서 시험 서열 태그를 제공하고; (c) 각 빈 (bin)에 위치된 시험 서열 태그의 리드 커버리지를 결정하고, 여기서 참조 유전체는 복수의 빈으로 분할되고; (d) 관심되는 핵산 서열에 대한 전역 프로필을 제공하고, 여기서 전역 프로필은 각 빈에서 예상된 리드 커버리지를 포함하고, 그리고 여기서 예상된 리드 커버리지는 시험 표본과 실제적으로 동일한 방식으로 염기서열결정되고 정렬되는 영향을 받지 않은 (가령, 이배체) 훈련 표본의 훈련 세트로부터 획득되고, 예상된 리드 커버리지는 각 빈마다 변동을 표시하고; (e) 최소한 관심되는 핵산 서열의 각 빈에서 예상된 리드 커버리지를 이용하여 시험 서열 태그의 리드 커버리지를 조정하고, 따라서 관심되는 핵산 서열에 대한 전역-프로필-교정된 리드 커버리지를 획득하고; (f) GC 함량 수준 및 전역-프로필-교정된 리드 커버리지 사이의 관계에 근거하여 전역-프로필-교정된 리드 커버리지를 조정하고, 따라서 관심되는 핵산 서열에 대한 표본-GC-교정된 리드 커버리지를 획득하고; 그리고 (g) 표본-GC-교정된 리드 커버리지에 근거하여 시험 표본에서 관심되는 핵산 서열의 사본수를 평가한다. 일부 구체예에서, 단계 (c)에서 결정된 리드 커버리지는 라이브러리 깊이 차이에 대한 정규화 후 획득된다. 라이브러리 정규화는 리드 커버리지를 본원에서 설명된 바와 같이, 이배체인 것으로 예상되는 로버스트 염색체에 매핑하는 리드의 총수에 의해 나눗셈하는 것을 수반할 수 있다. 대안으로, 라이브러리 깊이 정규화는 리드 커버리지를 전체 유전체에 매핑하는 리드의 개수로 나눗셈하고, 따라서 서열 태그 밀도 비율을 생산하는 것을 수반할 수 있다. 일부 구체예에서, 표본 그 자체에 대한 염기서열결정 데이터가 이배체 리드 커버리지를 갖는 것으로 추정되는 유전체 영역을 도출하고, 그리고 이들 영역을 라이브러리 정규화에서 이용하는데 이용될 수 있다. 라이브러리 깊이 정규화는 전형적으로 (c) 이후에 수행된 다른 형태의 정규화, 예를 들면, (f)에서 획득된 전역-프로필-교정된 리드 커버리지의 정규화와 별개로 수행된다. 다른 형태의 "정규화"는 본원에서 설명된 바와 같이 "서열 도스 (sequence dose)"를 생산한다.

[0015] 일부 구체예에서, 상기 방법은 빈의 리드 커버리지를 결정하는 작업 (c) 전에, 마스크가 적용된 빈에서 고려 범위에서 배제하는 서열 마스크를 적용하는 것을 더욱 수반한다. 일부 구체예에서, 서열 마스크는 복수의 영향을 받지 않은 훈련 표본의 서열 리드로부터 획득된다. 서열 마스크는 훈련 세트의 서열 리드를 참조 유전체에 맞춰 정렬하고, 따라서 훈련 표본에 대한 훈련 서열 태그를 제공함으로써 획득된다. 상기 방법은 또한, 참조 유전체를 복수의 빈으로 분할하고, 그리고 각 훈련 표본에 대해 각 빈에서 훈련 서열 태그의 리드 커버리지를 결정하는 것을 수반한다. 상기 방법은 마스크가 적용되지 않은 빈 및 마스크가 적용된 빈을 포함하는 서열 마스크를 창출하는 것을 더욱 수반한다. 마스크가 적용된 빈 각각은 마스크 적용 역치를 초과하는 분포 지수를 갖고, 분포 지수는 훈련 표본의 리드 커버리지의 분포에 관련된다. 일부 구체예에서, 마스크가 적용된 빈 및 마스크가 적용되지 않은 빈을 결정하는데 이용된 분포 지수는 훈련 표본의 리드 커버리지의 분산, 예를 들면, 변동 계수에 수학적으로 관련된다. 분포 지수는 빈에 마스크를 적용하기 위한 규준으로서 실행되는데, 그 이유는 훈련 표본 사이에서 큰 가변성 또는 분산을 표시하는 빈이 높은 분포 지수를 갖고, 그리고 이런 이유로, 사본수를 특징짓는데 이용하기에는 신뢰성이 없기 때문이다.

[0016] 일부 구체예에서, 상기 방법은 먼저, 서열 마스크를 창출하거나 또는 적용하기 전에, 영향을 받지 않은 훈련 표본 (또는 전역 프로필)에서 통상적인 체계적 변동을 제거한다. 이것은 각 빈에서 예상된 리드 커버리지에 따라 훈련 서열 태그의 리드 커버리지를 조정하고, 따라서 빈에서 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 획득함으로써 행위될 수 있고, 이들 리드 커버리지는 이후, 서열 마스크를 창출하는데 이용된다. 일부 구체예에서, 정규화된 리드 커버리지 양이 마스크를 계산하는데 이용된다. 정규화된 리드 커버리지 양은 정규화 서열의 리드 커버리지에 비하여 관심되는 핵산 서열의 리드 커버리지의 비율이다. 일부 구체예에서, 관심되는 핵산 서열 상에서 마스크가 적용된 빈은 첫 번째 마스크 적용 역치를 갖고, 그리고 정규화 서열 상에서 마스크가 적용된 빈은 두 번째 마스크 적용 역치를 갖는다. 일부 구체예에서, 첫 번째 마스크 적용 역치와 두 번째 마스크 적용 역치의 조합은 다른 역치를 이용하여 획득된 마스크보다, 영향을 받지 않은 표본에서 관심되는 서열을 포함하는 영역에 걸쳐 리드 커버리지의 더욱 낮은 변동을 유발하는 서열 마스크를 제공한다. 리드 커버리지의 변동은 표본 및 실행 전체에 대하여 분산을 제어하는 서열 마스크의 능력을 반영하고, 그리고 따라서, 더욱 낮은 변동은 영향을 받은 표본 및 영향을 받지 않은 표본 사이에 분리를 증가시킨다. 일부 구체예에서, 마스크 적용 역치는 검증 표본에서 리드 커버리지의 작은 분산 계수 및/또는 ROC 분석에서 큰 d' 값을 유발한다.

[0017] 일부 구체예에서, 서열 마스크는 빈 내에 훈련 표본 전체에 대하여 매핑 품질평가점수의 분포에 의해 규정된 마스크가 적용된 빈 및 마스크가 적용되지 않은 빈을 포함한다. 매핑 품질평가점수는 복수의 영향을 받지 않은 훈련 표본의 서열 리드를 참조 유전체에 맞춰 정렬하는 것으로부터 유래된다.

[0018] 일부 구체예에서, 시험 표본에서 관심되는 핵산 서열의 사본수를 평가하는 것은 정규화 서열의 리드 커버리지 정보를 이용하여, 시험 표본에 대한 관심되는 핵산 서열의 서열 도스를 계산하는 것을 포함한다. 일부 구체예에서, 서열 도스를 계산하는 것은 관심되는 핵산 서열에서 시험 서열 태그의 리드 커버리지 (가령, 표본-GC-교정된 리드 커버리지)를 정규화 서열에서 시험 서열 태그의 리드 커버리지로 나눗셈하는 것을 포함한다. 다른 방법, 예를 들면, 유전체의 다른 정규화 영역의 정규화된 리드 커버리지로부터 관심되는 서열의 정규화된 리드 커버리지를 모형화하는 선형 회귀 또는 로버스트 선형 회귀를 이용하는 방법이 서열 도스를 계산하는데 이용될 수 있다.

[0019] 일부 구체예에서, 정규화 서열은 하나 또는 그 이상의 로버스트 상염색체 서열 또는 이들의 분절을 포함한다. 일부 구체예에서, 로버스트 상염색체는 관심되는 염색체(들)를 제외한 모든 상염색체를 포함한다. 일부 구체예에서, 로버스트 상염색체는 chr X, Y, 13, 18, 그리고 21을 제외한 모든 상염색체를 포함한다. 일부 구체예에서, 로버스트 상염색체는 정상적인 이배체 상태로부터 일탈하는 표본으로부터 결정된 것들을 제외한 모든 상염색체를 포함한다.

[0020] 일부 구체예에서, 사본수를 평가하는 것은 정규화 서열의 리드 커버리지 정보를 이용하여, 시험 표본에 대한 관심되는 핵산 서열의 정규화된 염색체 값 또는 정규화된 분절 값을 계산하는 것을 더욱 포함한다.

[0021] 일부 구체예에서, 시험 표본은 2개의 상이한 유전체로부터 핵산의 혼합물을 포함한다. 일부 구체예에서, 시험 표본은 cfDNA 분자를 포함한다. 일부 구체예에서, 시험 표본은 태아와 모계 무세포 핵산을 포함한다. 일부 구체예에서, 시험 표본은 2개 또는 그 이상의 태아로부터 태아 무세포 핵산을 포함한다. 일부 구체예에서, 시험 표본은 동일한 개체로부터 암성 세포 및 영향을 받지 않은 세포로부터 핵산 (세포 유전체 DNA 및/또는 cfDNA)을 포함한다.

[0022] 일부 구체예에서, 시험 표본에서 관심되는 핵산 서열의 사본수를 평가하는 것은 완전한 또는 부분적인 태아 이

수성의 존재 또는 부재를 결정하는 것을 수반한다.

[0023] 일부 구체예에서, 표본-GC-교정된 리드 커버리지를 획득하는 작업 (f) 후, 상기 방법은 CNV의 평가에서 고려 사항으로부터 표본-GC-교정된 리드 커버리지의 이상점 빈을 제거하는 것을 더욱 수반한다. 일부 구체예에서, 이상점 빈은 중앙 표본-GC-교정된 리드 커버리지가 각 염색체에서 모든 빈의 중앙값으로부터 약 3 중위 절대 편차보다 큰 빈이다.

[0024] 일부 구체예에서, 각 빈에서 예상된 리드 커버리지는 훈련 표본 전체에 대하여 중앙값 또는 평균이다. 일부 구체예에서, 훈련 표본에서 리드 커버리지는 전역 프로필을 중앙값 또는 평균 정규화된 리드 커버리지로서 연산하기에 앞서, GC 함량 변동에 대해 교정된다.

[0025] 일부 실행에서, 시험 서열 태그의 리드 커버리지는 (i) 하나 또는 그 이상의 로버스트 염색체 또는 영역 내에 복수의 빈에서 예상된 리드 커버리지와 대비하여 시험 서열 태그의 리드 커버리지 사이에 수학적 관계를 획득하고, 그리고 (ii) 수학적 관계를 관심되는 서열에서 빈에 적용함으로써 조정된다. 일부 실행에서, 시험 표본에서 리드 커버리지는 영향을 받지 않은 훈련 표본으로부터 예상된 리드 커버리지 값 및 유전체의 로버스트 염색체 또는 다른 로버스트 영역에서 시험 표본에 대한 리드 커버리지 값 사이에 선형 관계를 이용하여 변동에 대해 교정된다. 조정은 전역-프로필-교정된 리드 커버리지를 유발한다. 일부 경우에, 조정은 아래와 같이, 로버스트 염색체 또는 영역에서 빈의 부분집합에 대한 시험 표본에 대한 리드 커버리지를 획득하는 것을 수반한다:

[0026] $y_a = 인터셉트 + 기울기 * gwp_a$

[0027] 여기서 ya는 하나 또는 그 이상의 로버스트 염색체 또는 영역에서 시험 표본을 위한 빈 a의 리드 커버리지이고, 그리고 gwpa는 영향을 받지 않은 훈련 표본을 위한 빈 a에 대한 전역 프로필이다. 이러한 과정은 이후, 아래와 같이, 서열 또는 관심 영역에 대한 전역-프로필-교정된 리드 커버리지 zb를 연산한다:

[0028] $z_b = y_b / (인터셉트 + 기울기 * gwp_b) - 1$

[0029] 여기서 yb는 관심되는 서열 (이것은 로버스트 염색체 또는 영역 외부에 체류할 수 있다)에서 시험 표본을 위한 빈 b의 관찰된 리드 커버리지이고, 그리고 gwpb는 영향을 받지 않은 훈련 표본을 위한 빈 b에 대한 전역 프로필이다. 분모 (인터셉트 + 기울기 * gwpb)는 영향을 받지 않은 시험 표본에서 관찰되는 빈 b에 대한 리드 커버리지이다. 사본수 변동을 품는 관심되는 서열의 경우에, 빈 b에 대한 관찰된 리드 커버리지 및 따라서, 전역-프로필-교정된 리드 커버리지 값은 영향을 받지 않은 표본의 리드 커버리지로부터 유의미하게 일탈할 것이다. 가령, 교정된 리드 커버리지 zb는 영향을 받은 염색체에서 빈에 대한 삼염색체성 표본의 경우에 태아 분율에 비례할 것이다. 이러한 과정은 로버스트 염색체에서 인터셉트와 기울기를 연산함으로써 표본 내에서 정규화하고, 그리고 이후, 표적 염색체 (또는 관심되는 다른 서열)가 동일한 표본 내에서 로버스트 염색체에 대해 유지하는 관계 (기울기 및 인터셉트에 의해 설명된 바와 같이)로부터 어떻게 일탈하는 지를 평가한다.

[0030] 일부 구체예에서, (e)로부터 시험 서열 태그의 전역-프로필-교정된 리드 커버리지는 관심되는 핵산 서열에서 빈의 전역-프로필-교정된 리드 커버리지 및 정규화 서열에서 빈의 전역-프로필-교정된 리드 커버리지를 포함한다.

[0031] 일부 구체예에서, 작업 (f)에서 전역-프로필-교정된 리드 커버리지를 조정하는 것은 참조 유전체에서 빈을 복수의 GC 군으로 그룹화하고, 각 GC 군은 복수 빈을 포함하고, 여기서 복수 빈은 시험 서열 태그를 내포하고 유사한 GC 함량을 갖고; 복수의 로버스트 상염색체에 대한 각 GC 군에 대한 전역-프로필-교정된 리드 커버리지의 기댓값을 결정하고; 그리고 동일한 GC 군의 결정된 기대값에 근거하여 각 GC 군에 대한 시험 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하고, 따라서 관심되는 핵산 서열에서 시험 서열 태그의 표본-GC-교정된 리드 커버리지를 획득하는 것을 포함한다.

[0032] 일부 구체예에서, 전역-프로필-교정된 리드 커버리지의 기댓값은 복수의 로버스트 상염색체의 GC 군에 대한 리드 커버리지의 평균 또는 중앙값이다. 일부 구체예에서, 시험 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하는 것은 기댓값을 전역-프로필-교정된 리드 커버리지로부터 감산함으로써 달성된다.

[0033] 일부 구체예에서, 작업 (f)에서 전역-프로필-교정된 리드 커버리지를 조정하는 것은 선형 또는 비선형 수학 함수를 복수의 로버스트 상염색체로부터 데이터 포인트에 적합시키는 것을 수반하고, 여기서 각 데이터 포인트는 리드 커버리지 값을 GC 함량 값에 관련시킨다. 상기 방법은 이후, 고려 중인 빈의 GC 함량 값에서 수학 함수의 리드 커버리지 값에 동등한 값으로 리드 커버리지를 조정한다. 일부 구체예에서, 상기 방법은 기댓값을 전역-프로필-교정된 리드 커버리지로부터 감산한다. 다른 구체예에서, 상기 방법은 리드 커버리지 양을 기댓값을 나눗

셈한다.

[0034] 일부 구체예에서, CNV를 평가하기 위한 방법은 또한, 복수의 영향을 받지 않은 개체 및/또는 시험 표본으로부터 무세포 DNA를 추출하는 것을 수반한다. 일부 구체예에서, 이들 방법은 또한, 서열분석기를 이용하여 시험 표본으로부터 핵산을 염기서열결정하고, 따라서 시험 표본의 서열 리드를 산출하는 것을 수반한다. 일부 구체예에서, 서열 리드는 개체의 전체 유전체 내에 어디에서든 약 20 내지 50-bp의 서열을 포함한다. 일부 구체예에서, 서열 리드는 바코드화된 25-mer을 포함한다.

[0035] 일부 구체예에서, 시험 서열 태그와 훈련 서열 태그의 리드 커버리지는 비-배제된 부위 수치 (NES 수치)에 근거되고, 여기서 NES 수치는 비-배제된 부위에 매핑된 비다중 및/또는 독특하게 정렬된 서열 태그의 개수이다.

[0036] 일부 구체예에서, 관심되는 핵산 서열은 약 1000 bp 및 1,000,000 bp 사이의 빈으로 분할된다. 일부 구체예에서, 빈 크기는 약 100,000 bp이다. 일부 구체예에서, 빈 크기는 시험 표본의 서열 리드의 개수에 관하여 계산된다. 일부 구체예에서, 각 빈에서 서열 태그의 개수는 최소한 약 1000 bp이다.

[0037] 본원에서 개시된 일부 구체예는 관심되는 핵산 서열의 사본수의 평가를 위한 서열 마스크를 창출하기 위한 방법을 제공한다. 상기 방법은 다음을 포함한다: (a) 컴퓨터 시스템에서, 복수의 영향을 받지 않은 훈련 표본으로부터 서열 리드를 포함하는 훈련 세트를 제공하고; (b) 훈련 세트의 서열 리드를 관심되는 핵산 서열을 포함하는 참조 유전체에 맞춰 정렬하고, 따라서 훈련 표본에 대한 훈련 서열 태그를 제공하고; (c) 참조 유전체를 복수의 빈으로 분할하고; (d) 각 훈련 표본을 위한 각 빈에서 훈련 서열 태그의 리드 커버리지를 각 영향을 받지 않은 훈련 표본에 대해 결정하고; (e) 모든 훈련 표본 전체에 대하여 훈련 서열 태그의 예상된 리드 커버리지를 각 빈에 대해 결정하고; (f) 각 빈에서 예상된 리드 커버리지에 따라 각 훈련 표본을 위한 각 빈에서 훈련 서열 태그의 리드 커버리지를 조정하고, 따라서 각 훈련 표본을 위한 빈에서 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 획득하고; 그리고 (g) 참조 유전체 전체에 대하여 마스크가 적용되지 않은 빈 및 마스크가 적용된 빈을 포함하는 서열 마스크를 창출하고, 여기서 마스크가 적용된 빈 각각은 마스크 적용 역치를 초과하는 분포 특징을 갖고, 그리고 분포 특징은 훈련 표본 전체에 대하여 빈에서 훈련 서열 태그의 조정된 리드 커버리지에 대해 제공된다.

[0038] 일부 구체예에서, 각 빈에 대해 (e)에서 결정된 예상된 리드 커버리지는 훈련 표본의 리드 커버리지의 중앙값 또는 평균을 포함한다. 일부 구체예에서, 작업 (f)에서 훈련 서열 태그의 리드 커버리지를 조정하는 것은 각 훈련 표본의 각 빈에 대한 훈련 서열 태그의 리드 커버리지로부터 중앙값 또는 평균을 감산하는 것을 포함한다. 일부 구체예에서, 조정은 각 훈련 표본의 각 빈에 대한 훈련 서열 태그의 리드 커버리지를 중앙값 또는 평균으로 나눗셈함으로써 행위된다.

[0039] 일부 구체예에서, 관심되는 핵산 서열에서 마스크가 적용된 빈은 첫 번째 마스크 적용 역치를 갖고, 그리고 정규화 서열에서 마스크가 적용된 빈은 두 번째 마스크 적용 역치를 갖는다. 일부 구체예에서, 첫 번째 마스크 적용 역치와 두 번째 마스크 적용 역치의 조합은 다른 역치를 이용하여 획득된 마스크보다, 영향을 받지 않은 표본에서 관심되는 서열을 포함하는 영역에 걸쳐 리드 커버리지의 더욱 낮은 변동을 유발하는 서열 마스크를 제공한다.

[0040] 일부 구체예에서, 서열 마스크를 창출하기 위한 방법은 (f) 이후 및 (g) 이전에, 각 훈련 표본 내에 현존하는 GC 함량 수준 및 전역-프로필-교정된 리드 커버리지 사이의 관계에 근거하여 각 훈련 표본의 빈에 대한 전역-프로필-교정된 리드 커버리지를 조정하고, 따라서 각 훈련 표본에 대한 훈련 서열 태그의 표본-GC-교정된 리드 커버리지를 획득하는 것을 더욱 수반한다.

[0041] 일부 구체예에서, 각 훈련 표본에 대한 전역-프로필-교정된 리드 커버리지의 조정은 참조 유전체에서 모든 빈을 복수의 GC 군으로 그룹화하고, 각 GC 군은 유사한 GC 함량을 갖는 복수 빈을 포함하고; 복수의 로버스트 상염색체에 대한 각 GC 군에 대한 전역-프로필-교정된 리드 커버리지의 기댓값을 결정하고; 그리고 동일한 GC 군의 결정된 기대값에 근거하여 각 GC 군에 대한 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하고, 따라서 관심되는 핵산 서열에서 훈련 서열 태그의 표본-GC-교정된 리드 커버리지를 획득하는 것을 수반한다.

[0042] 일부 구체예에서, 전역-프로필-교정된 리드 커버리지의 기댓값은 복수의 로버스트 상염색체의 GC 군에 대한 리드 커버리지의 평균 또는 중앙값이다. 일부 구체예에서, 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하는 것은 기댓값을 전역-프로필-교정된 리드 커버리지로부터 감산하는 것을 수반한다.

[0043] 일부 구체예에서, 각 훈련 표본에 대한 전역-프로필-교정된 리드 커버리지를 조정하는 것은 선형 또는 비선형 수학 함수를 복수의 로버스트 상염색체로부터 데이터 포인트에 적합시키는 것을 수반하고, 여기서 각 데이터 포

인트는 리드 커버리지 값을 GC 함량 값에 관련시킨다. 상기 방법은 이후, 빈의 GC 함량 값에서 수학 함수의 리드 커버리지 값에 동등한, 각 빈에 대한 리드 커버리지의 기댓값에 근거하여 각 빈에서 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정한다.

[0044] 일부 구체예에서, 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 조정하는 것은 기댓값을 전역-프로필-교정된 리드 커버리지로부터 감산하는 것을 포함한다. 다른 구체예에서, 리드 커버리지는 기대값에 의해 나눗셈된다.

[0045] 일부 구체예에서, 시험 표본은 혈액, 혈장, 혈청, 소변과 타액 표본에서 선택되는 모계 표본일 수 있다. 구체예 중에서 임의의 한 가지에서, 시험 표본은 혈장 표본일 수 있다. 모계 표본의 핵산 분자는 태아와 모계 무세포 DNA 분자의 혼합물이다. 핵산의 염기서열결정은 차세대 염기서열결정 (NGS)을 이용하여 수행될 수 있다. 일부 구체예에서, 염기서열결정은 가역성 염료 종결인자로 합성에 의한 염기서열결정을 이용한 대량 병렬 염기서열결정이다. 다른 구체예에서, 염기서열결정은 결찰에 의한 염기서열결정이다. 또 다른 구체예에서, 염기서열결정은 단일 분자 염기서열결정이다. 임의선택적으로, 증폭 단계는 염기서열결정에 앞서 수행된다.

[0046] 다른 구체예는 시험 표본에서 관심되는 핵산 서열, 예를 들면, 임상적으로 유관한 서열의 사본수 변동 (CNV)을 확인하기 위한 방법을 제공한다. 상기 방법은 완전한 염색체 또는 염색체의 분절 대신에, 관심되는 서열의 사본수 변동을 사정한다.

[0047] 컴퓨터 시스템 상에서 구현된 일정한 구체예에서, 관심되는 하나 또는 그 이상의 염색체 또는 관심되는 염색체 분절 각각에 대해 확인된 서열 태그의 개수는 최소한 약 10,000, 또는 최소한 약 100,000이다.

[0048] 개시된 구체예는 또한, 언급된 작업 및 본원에서 설명된 다른 연산적 작업을 수행하기 위한 프로그램 명령이 제공되는 비-일시적 컴퓨터 판독가능 매체를 포함하는 컴퓨터 프로그램 제품을 제공한다.

[0049] 일부 구체예는 시험 표본에서 관심되는 핵산 서열의 사본수의 평가를 위한 시스템을 제공한다. 상기 시스템은 표본으로부터 핵산 서열 정보를 제공하는 시험 표본으로부터 핵산을 제공받기 위한 서열분석기, 프로세서; 그리고 본원에서 언급된 방법을 이용하여 시험 표본에서 사본수를 평가하기 위해 상기 프로세서에서 이행을 위한 명령이 그 안에 저장된 하나 또는 그 이상의 컴퓨터-판독가능 저장 매체를 포함한다.

[0050] 일부 구체예에서, 방법은 부가적으로, 상기 시험 표본의 상기 태아와 모계 핵산 분자에 대한 상기 서열 정보를 획득하기 위해, 상기 시험 표본의 상기 핵산 분자 중에서 최소한 일부를 염기서열결정하는 것을 포함한다. 염기서열결정은 모계 시험 표본으로부터 모계와 태아 핵산에서 대량으로 병렬 염기서열결정하여 서열 리드를 생산하는 것을 수반할 수 있다.

[0051] 비록 본원에서 실례가 인간에 관계하고, 그리고 언어가 인간 관심사에 일차적으로 관계하지만, 본원에서 설명된 개념은 임의의 식물 또는 동물로부터 유전체에 적용가능하다. 본 발명의 이런 저런 목적과 특질은 다음 설명과 첨부된 청구항으로부터 더욱 충분히 명백해질 것이고, 또는 본원에서 진술된 바와 같은 발명의 실시에 의해 학습될 수 있다.

[0052] **참조로서 편입**

[0053] 본원에서 언급된 모든 특허, 특허 출원, 그리고 다른 간행물뿐만 아니라 이들 참고문헌 내에 개시된 모든 서열은 마치 각 개별 간행물, 특허 또는 특허 출원이 구체적으로 및 개별적으로, 참조로서 편입되는 것으로 지시되는 것과 동일한 정도로, 명시적으로 본원에 참조로서 편입된다. 인용된 모든 문서는 유관한 부분에서, 본원에서 그들의 인용의 문맥에 의해 지시된 목적을 위해 전체적으로 본원에 참조로서 편입된다. 하지만, 임의의 문서의 인용은 이것이 본 발명에 대하여 선행 기술이라는 시인으로서 해석되지 않는다.

도면의 간단한 설명

[0054] **도면의 간단한 설명**

도면 1은 핵산의 혼합물을 포함하는 시험 표본에서 사본수 변동의 존재 또는 부재를 결정하기 위한 방법 100의 흐름도이다.

도면 2는 사본수의 평가를 위해 이용된 관심되는 핵산 서열의 리드 커버리지를 결정하기 위한 과정의 흐름도를 묘사한다.

도면 3a는 시험 표본으로부터 서열 데이터에서 잡음을 감소시키기 위한 과정의 실례의 흐름도를 보여준다.

도면 3b-3k는 도면 3a에서 묘사된 과정의 다양한 시기에서 획득된 데이터의 분석을 나타낸다.

도면 4a는 서열 데이터에서 잡음을 감소시키기 위한 서열 마스크를 창출하기 위한 과정의 흐름도를 보여준다.

도면 4b는 MapQ 점수가 정규화된 리드 커버리지 양의 CV와 강한 변함없는 상관을 갖는다는 것을 보여준다.

도면 5는 시험 표본을 처리하고 궁극적으로 진단을 내리기 위한 분산된 시스템의 블록 선도이다.

도면 6은 시험 표본을 처리함에 있어서 상이한 작업이 어떻게 그룹화되어 시스템의 상이한 요소에 의해 취급될 수 있는 지를 개략적으로 도해한다.

도면 7a와 7b는 실시예 1a에서 설명되는 단축된 프로토콜 (도면 7a), 그리고 실시예 1b에서 설명된 프로토콜 (도면 7b)에 따라 제조된 cfDNA 염기서열결정 라이브러리의 전기영동도를 보여준다.

도면 8은 118건의 쌍둥이 임신으로부터 모계 혈장 표본에 대한 정규화된 염색체 값 (NCV) 분포를 보여준다. (A) 염색체 21과 18에 대한 NCV 분포; 3개 표본은 T21 영향을 받은 것으로 분류되었고 (T21에 대해 모자이크인 태아 포함), 그리고 1개 표본은 T18 영향을 받은 것으로 분류되었다. (B) 염색체 Y에 대한 NCV 분포. 코호트는 암컷/암컷으로서 임상적으로 분류된 표본 또는 최소한 하나의 수컷 태아를 내포하는 표본 (수컷/암컷 및 수컷/수컷)으로 분할되었고, 그리고 Y 염색체의 존재는 염색체 Y에 대한 NCV를 이용하여 결정되었다.

도면 9는 NIPT 연구에서 분석된 쌍둥이 표본을 보여준다. 상업적으로 가용한 NIPT 시험의 성과를 사정하기 위한 다양한 연구에서 이용된 쌍둥이 표본의 숫자.

발명을 실시하기 위한 구체적인 내용

[0055] **상세한 설명**

[0056] 개시된 구체예는 태아와 모계 무세포 핵산을 포함하는 시험 표본에서 Y 염색체의 사본수의 평가를 위한 방법, 기구, 그리고 시스템에 관계한다. 일부 구체예에서, 관심되는 서열은 가령, 유전자 또는 질환 상태와 연관되는 것으로 알려져 있거나 또는 의심되는 킬로베이스 (kb) 내지 메가베이스 (Mb)에서부터 전체 염색체까지의 범위에서 변하는 유전체 분절 서열을 포함한다. 일부 구체예에서, Y 염색체의 사본수가 태아 성별을 결정하는데 이용된다. 일부 구체예에서, 본 발명 방법에 따라 결정될 수 있는 CNV는 성염색체 Y의 일염색체성과 삼염색체성 (가령, 47,XXY 및 47,XYY), 성염색체의 다른 다염색체성, 예를 들면, 사염색체성과 오염색체성 (가령, XXXXY 및 XYYYY), 그리고 성염색체 중에서 임의의 한 가지 또는 그 이상의 분절의 결실 및/또는 중복을 포함한다. 관심되는 서열의 다른 실례는 널리 공지된 이수체와 연관된 염색체, 예를 들면, 삼염색체성 XXX, 삼염색체성 21, 그리고 질환, 예를 들면, 암에서 크게 증가되는 염색체의 분절, 예를 들면, 급성 골수성 백혈병에서 부분적인 삼염색체성 8을 포함한다.

[0057] 달리 지시되지 않으면, 본원에서 개시된 방법과 시스템의 실시는 당해 분야의 기술 범위 안에 있는, 분자생물학, 미생물학, 단백질 정제, 단백질 가공, 단백질과 DNA 염기서열결정, 그리고 재조합 DNA 분야에서 통상적으로 이용되는 전통적인 기술과 기구를 수반한다. 이런 기술과 기구는 당업자에게 공지되어 있고 다양한 문서와 참고서에서 설명된다 (가령, Sambrook et al., "Molecular Cloning: A Laboratory Manual," Third Edition (Cold Spring Harbor), [2001]); 그리고 Ausubel et al., "Current Protocols in Molecular Biology" [1987]를 참조한다.).

[0058] 수치 범위는 범위를 규정하는 숫자를 포괄한다. 명세서 전반에서 제공된 모든 최고 수치 한계는 마치 더욱 낮은 수치 한계가 본원에서 명시적으로 기재된 것처럼, 모든 더욱 낮은 수치 한계를 포함하는 것으로 의도된다. 명세서 전반에서 제공된 모든 최소 수치 한계는 마치 더욱 높은 수치 한계가 본원에서 명시적으로 기재된 것처럼, 모든 더욱 높은 수치 한계를 포함할 것이다. 명세서 전반에서 제공된 모든 수치 범위는 마치 더욱 좁은 수치 범위가 본원에서 충분히 명시적으로 기재된 것처럼, 더욱 넓은 수치 범위에 들어가는 모든 더욱 좁은 수치 범위를 포함할 것이다. 본원에서 제공된 표제는 본 발명을 제한하는 것으로 의도되지 않는다.

[0059] 본원에서 달리 정의되지 않으면, 본원에서 이용된 모든 기술 용어와 과학 용어는 당업자에 의해 통상적으로 이해되는 바와 동일한 의미를 갖는다. 본원에서 포함된 용어를 포함하는 다양한 과학 사전은 널리 공지되어 있고 당업자에게 가용하다. 비록 본원에서 설명된 것들과 유사하거나 또는 동등한 임의의 방법과 물질이 본원에서 개시된 구체예의 실시 또는 시험에서 용도를 발견하지만, 일부 방법과 물질이 설명된다.

[0060] 바로 아래에 규정된 용어는 본 명세서를 전체로서 참조함으로써 더욱 충분히 설명된다. 본 발명은 설명된 특정

방법, 프로토콜, 그리고 시약에 한정되지 않는 것으로 이해되는데, 그 이유는 이들이 당업자에 의해 이용되는 상황에 따라 변할 수 있기 때문이다.

[0061] **정의**

[0062] 본원에서 이용된 바와 같이, 단수 용어 ("a," "an," 그리고 "the")는 문맥에서 명확하게 달리 지시되지 않으면, 복수 참조를 포함한다.

[0063] 달리 지시되지 않으면, 핵산은 5'에서 3' 방향으로 왼쪽에서 오른쪽으로 기재되고, 그리고 아미노산 서열은 아미노에서 카르복시 방향으로 왼쪽에서 오른쪽으로 기재된다.

[0064] 용어 "사정하는"은 CNV에 대한 핵산 표본을 분석하는 문맥에서 본원에서 이용될 때, 호출의 3가지 유형 중에서 한 가지에 의해 염색체 또는 분절 이수성의 상태를 특징짓는 것을 지칭한다: "정상적임" 또는 "영향을 받지 않음", "영향을 받음" 및 "호출 없음." 정상적임 및 영향을 받음을 호출하기 위한 역치는 전형적으로 세팅된다. 이수성 또는 다른 사본수 변동에 관련된 파라미터가 표본에서 계측되고, 그리고 계측된 값이 역치와 비교된다. 중복 유형 이수체의 경우에, 영향을 받음의 호출은 염색체 또는 분절 도스 (또는 다른 계측된 값 서열 함량)이 영향을 받은 표본에 대한 규정된 역치 세트를 초과하면 만들어진다. 이런 이수성의 경우에, 정상적임의 호출은 염색체 또는 분절 도스가 정상적인 표본에 대한 역치 세트 미만이면 만들어진다. 대조적으로, 결실 유형 이수성의 경우에, 영향을 받음의 호출은 염색체 또는 분절 도스가 영향을 받은 표본에 대한 규정된 역치 미만이면 만들어지고, 그리고 정상적임의 호출은 염색체 또는 분절 도스가 정상적인 표본에 대한 역치 세트를 초과하면 만들어진다. 가령, 삼염색체성의 존재에서, "정상적인" 호출은 파라미터, 예를 들면, 신뢰도의 사용자-규정된 역치 미만인 시험 염색체 도스의 값에 의해 결정되고, 그리고 "영향을 받은" 호출은 파라미터, 예를 들면, 신뢰도의 사용자-규정된 역치 초과인 시험 염색체 도스에 의해 결정된다. "호출 없음" 결과는 파라미터, 예를 들면, "정상적인" 또는 "영향을 받은" 호출을 만들기 위한 역치 사이에 있는 시험 염색체 도스에 의해 결정된다. 용어 "호출 없음"은 "미분류"와 교체가능하게 이용된다.

[0065] 본원에서 용어 "사본수 변동"은 참조 표본 내에 존재하는 핵산 서열의 사본수와 비교하여, 시험 표본 내에 존재하는 핵산 서열의 사본의 숫자에서 변이를 지칭한다. 일정한 구체예에서, 핵산 서열은 1 kb 또는 그 이상이다. 일부 경우에, 핵산 서열은 전체 염색체 또는 이의 유의미한 부분이다. "사본수 변동체"는 사본수 차이가 시험 표본 내에 관심되는 핵산 서열 및 관심되는 핵산 서열의 예상된 수준의 비교에 의해 발견되는 핵산의 서열을 지칭한다. 가령, 시험 표본 내에 관심되는 핵산 서열의 수준은 유자격 표본 내에 존재하는 것과 비교된다. 사본수 변동체/변동은 미세결실을 비롯한 결실, 미세삽입을 비롯한 삽입, 중복, 증식, 그리고 전위를 포함한다. CNV는 염색체 이수성 및 부분적인 이수성을 포괄한다.

[0066] 본원에서 용어 "이수성"은 전체 염색체, 또는 염색체의 부분의 상실 또는 획득에 의해 유발된 유전 물질의 불균형을 지칭한다.

[0067] 본원에서 용어 "염색체 이수성" 및 "완전한 염색체 이수성"은 전체 염색체의 상실 또는 획득에 의해 유발된 유전 물질의 불균형을 지칭하고, 그리고 생식계열 이수성 및 모자이크 이수성을 포함한다.

[0068] 본원에서 용어 "부분적인 이수성" 및 "부분적인 염색체 이수성"은 염색체의 부분의 상실 또는 획득에 의해 유발된 유전 물질의 불균형, 예를 들면, 부분적인 일염색체성 및 부분적인 삼염색체성을 지칭하고, 그리고 전위, 결실 및 삽입으로부터 발생하는 불균형을 포괄한다.

[0069] 용어 "복수"는 하나 이상의 요소를 지칭한다. 가령, 상기 용어는 본원에서, 본원에서 개시된 방법을 이용하여 시험 표본 및 유자격 표본에서 사본수 변동의 유의미한 차이를 확인하는데 충분한 다수의 핵산 분자 또는 서열 태그에 관하여 이용된다. 일부 구체예에서, 약 20과 40bp 사이의 최소한 약 3 x 10^6 서열 태그가 각 시험 표본에 대해 획득된다. 일부 구체예에서, 각 시험 표본은 최소한 약 5 x 10^6, 8 x 10^6, 10 x 10^6, 15 x 10^6, 20 x 10^6, 30 x 10^6, 40 x 10^6, 또는 50 x 10^6 서열 태그에 대한 데이터를 제공하고, 각 서열 태그는 약 20 내지 40bp를 포함한다.

[0070] 용어 "폴리뉴클레오티드," "핵산" 및 "핵산 분자"는 교체가능하게 이용되고, 그리고 한 뉴클레오티드의 펜토오스의 3' 위치가 포스포디에스테르 기에 의해, 그 다음 뉴클레오티드의 펜토오스의 5' 위치에 연결되는 뉴클레오티드의 공유 연결된 서열 (즉, RNA의 경우에 리보뉴클레오티드 및 DNA의 경우에 데옥시리보뉴클레오티드)을 지칭한다. 뉴클레오티드는 RNA 및 DNA 분자, 예를 들면, cfDNA 분자를 포함하지만 이들에 한정되지 않는, 임의의 형태의 핵산의 서열을 포함한다. 용어 "폴리뉴클레오티드"는 제한 없이, 단일- 및 이중 가닥 폴리뉴클레오티드

를 포함한다.

[0071] 본원에서 용어 "부분"은 총합으로 1 인간 유전체의 서열 정보보다 적은 양에 이르는 생물학적 표본에서 태아와 모계 핵산 분자의 서열 정보의 양에 관하여 이용된다.

[0072] 본원에서 용어 "시험 표본"은 전형적으로, 사본수 변동에 대해 선별검사되는 최소한 하나의 핵산 서열을 포함하는 핵산 또는 핵산의 혼합물을 포함하는, 생물학적 유체, 세포, 조직, 장기, 또는 생물체로부터 유래된 표본을 지칭한다. 일정한 구체예에서, 표본은 사본수가 변이를 겪은 것으로 의심되는 최소한 하나의 핵산 서열을 포함한다. 이런 표본은 객담/경구 유체, 양수, 혈액, 혈액 분획물, 또는 미세 바늘 생검 표본 (가령, 외과적 생검, 미세 바늘 생검 등), 소변, 복막액, 흉수, 기타 등등을 포함하지만 이들에 한정되지 않는다. 비록 표본이 종종 인간 개체 (가령, 환자)로부터 채취되긴 하지만, 이들 검정은 개, 고양이, 말, 염소, 양, 소, 돼지 등을 포함하지만 이들에 한정되지 않는 임의의 포유동물로부터 표본에서 사본수 변동 (CNVs)에 이용될 수 있다. 표본은 생물학적 공급원으로부터 획득된 그대로 직접적으로 이용되거나, 또는 표본의 형질을 변형하는 선처리 이후에 이용될 수 있다. 가령, 이런 선처리는 혈액으로부터 혈장을 준비하고, 점성 유체를 희석하고, 기타 등등을 포함할 수 있다. 선처리의 방법은 또한, 여과, 침전, 희석, 증류, 혼합, 원심분리, 동결, 동결건조, 농축, 증폭, 핵산 단편화, 간섭 성분의 비활성화, 시약의 첨가, 용해 등을 수반할 수 있지만, 이들에 한정되지 않는다. 이런 선처리 방법이 표본에 대하여 이용되면, 이런 선처리 방법은 전형적으로, 관심되는 핵산(들)이 때때로 처리되지 않은 시험 표본 (가령, 다시 말하면, 임의의 이런 선처리 방법에 종속되지 않는 표본))에서 농도에 비례하는 농도로 시험 표본 내에 남아있도록 하는 것이다. 이런 "처리된" 또는 "가공된" 표본은 여전히, 본원에서 설명된 방법에 대하여 생물학적 "시험" 표본인 것으로 고려된다.

[0073] 본원에서 용어 "유자격 표본" 또는 "영향을 받지 않은 표본"은 시험 표본 내에 핵산이 비교되는 공지된 사본수에서 존재하는 핵산의 혼합물을 포함하는 표본을 지칭하고, 그리고 이것은 관심되는 핵산 서열에 대해 정상적인, 다시 말하면, 이수체가 아닌 표본이다. 일부 구체예에서, 유자격 표본은 서열 마스크 또는 서열 프로필을 도출하기 위한 훈련 세트의 영향을 받지 않은 훈련 표본으로서 이용된다. 일정한 구체예에서, 유자격 표본은 하나 또는 그 이상의 정규화 염색체 또는 고려 중인 염색체에 대한 분절을 확인하는데 이용된다. 가령, 유자격 표본은 염색체 21에 대한 정규화 염색체를 확인하는데 이용될 수 있다. 이런 경우에, 유자격 표본은 삼염색체성 21 표본이 아닌 표본이다. 다른 실례는 염색체 X에 대한 유자격 표본으로서 단지 암컷만을 이용하는 것을 수반한다. 유자격 표본은 또한, 다른 목적, 예를 들면, 영향을 받은 표본을 호출하기 위한 역치를 결정하고, 참조 서열 상에서 마스크 영역을 규정하기 위한 역치를 확인하고, 유전체의 상이한 영역에 대한 예상된 리드 커버리지 양을 결정하고, 기타 등등에 이용될 수 있다.

[0074] 본원에서 용어 "훈련 세트"는 영향을 받은 및/또는 영향을 받지 않은 표본을 포함할 수 있고, 그리고 시험 표본을 분석하기 위한 모형을 개발하는데 이용되는 한 세트의 훈련 표본을 지칭한다. 일부 구체예에서, 훈련 세트는 영향을 받지 않은 표본을 포함한다. 이들 구체예에서, CNV를 결정하기 위한 역치는 관심되는 사본수 변동에 대해 영향을 받지 않은 표본의 훈련 세트를 이용하여 확립된다. 훈련 세트에서 영향을 받지 않은 표본은 정규화 서열, 예를 들면, 정규화 염색체를 확인하기 위한 유자격 표본으로서 이용될 수 있고, 그리고 영향을 받지 않은 표본의 염색체 도스는 관심되는 각 서열, 예를 들면, 염색체에 대한 역치를 세팅하는데 이용된다. 일부 구체예에서, 훈련 세트는 영향을 받은 표본을 포함한다. 훈련 세트에서 영향을 받은 표본은 영향을 받은 시험 표본이 영향을 받지 않은 표본과 쉽게 구별될 수 있다는 것을 실증하는데 이용될 수 있다.

[0075] 본원에서 "훈련 세트"는 또한, 관심되는 개체군의 통계학적 표본의 한 세트의 개체에 관하여 이용되고, 이들 개체의 데이터는 개체군에 일반화가능한 관심되는 하나 또는 그 이상의 정량적 값을 결정하는데 이용된다. 통계학적 표본은 관심되는 개체군에서 개체의 부분집합이다. 개체는 인간, 동물, 조직, 세포, 다른 생물학적 표본 (즉, 통계학적 표본은 복수 생물학적 표본을 포함할 수 있다), 그리고 통계학적 분석을 위한 데이터 포인트를 제공하는 다른 개별 실체일 수 있다.

[0076] 통상적으로, 훈련 세트는 검증 세트와 함께 이용된다. 본원에서 용어 "검증 세트"는 통계학적 표본에서 한 세트의 개체에 관하여 이용되고, 이들 개체의 데이터는 훈련 세트를 이용하여 결정된 관심되는 정량적 값을 검증하거나 또는 평가하는데 이용된다. 일부 구체예에서, 예로서, 훈련 세트는 참조 서열에 대한 마스크를 계산하기 위한 데이터를 제공한다; 검증 세트는 마스크를 검증하거나 또는 평가하기 위한 데이터를 제공한다.

[0077] 본원에서 "사본수의 평가"는 서열의 사본수에 관련된 유전자 서열의 상태의 통계학적 평가에 관하여 이용된다. 가령, 일부 구체예에서, 평가는 유전자 서열의 존재 또는 부재의 결정을 포함한다. 일부 구체예에서, 평가는 유전자 서열의 부분적인 또는 완전한 이수성의 결정을 포함한다. 다른 구체예에서, 평가는 유전자 서열의 사본수

에 근거하여 2개 또는 그 이상의 표본 사이에 구별을 포함한다. 일부 구체예에서, 평가는 유전자 서열의 사본수에 근거된 통계학적 분석, 예를 들면, 정규화 및 비교를 포함한다.

[0078] 용어 "유자격 핵산"은 "유자격 서열"과 교체가능하게 이용되는데, 이것은 시험 서열 또는 시험 핵산의 양이 비교되는 서열이다. 유자격 서열은 바람직하게는, 공지된 표현에서 생물학적 표본 내에 존재하는 것이다, 다시 말하면, 유자격 서열의 양이 알려져 있다. 일반적으로, 유자격 서열은 "유자격 표본" 내에 존재하는 서열이다. "관심되는 유자격 서열"은 유자격 표본 내에서 양이 알려져 있는 유자격 서열이고, 그리고 의학적 상태를 앓는 개체에서 서열 표현에서 차이와 연관되는 서열이다.

[0079] 본원에서 용어 "관심되는 서열" 또는 "관심되는 핵산 서열"은 병든 개체와 대비하여 건강한 개체에서 서열 표현에서 차이와 연관되는 핵산 서열을 지칭한다. 관심되는 서열은 질환 또는 유전적 장애에서 잘못 표현되는, 다시 말하면, 지나치게 많거나 또는 불충분하게 표현되는 염색체 상에서 서열일 수 있다. 관심되는 서열은 염색체의 부분, 다시 말하면, 염색체 분절, 또는 전체 염색체일 수 있다. 가령, 관심되는 서열은 이수성 상태에서 지나치게 많은 염색체, 또는 암에서 불충분하게 표현되는 종양-억제인자를 인코딩하는 유전자이다. 관심되는 서열은 개체의 세포의 전체 개체군 또는 아개체군에서 지나치게 많거나 또는 불충분하게 표현되는 서열을 포함한다. "관심되는 유자격 서열"은 유자격 표본 내에 관심되는 서열이다. "관심되는 시험 서열"은 시험 표본 내에 관심되는 서열이다.

[0080] 본원에서 용어 "정규화 서열"은 정규화 서열과 연관된 관심되는 서열에 매핑된 서열 태그의 개수를 정규화하는데 이용되는 서열을 지칭한다. 일부 구체예에서, 정규화 서열은 로버스트 염색체를 포함한다. "로버스트 염색체"는 이수체일 개연성이 낮은 것이다. 인간 염색체를 수반하는 일부 경우에, 로버스트 염색체는 X 염색체, Y 염색체, 염색체 13, 염색체 18, 그리고 염색체 21 이외에 임의의 염색체이다. 일부 구체예에서, 정규화 서열은 이것이 정규화 파라미터로서 이용되는 관심되는 서열의 가변성을 근사하는 표본 및 염기서열결정 실행 사이에서, 이것에 매핑되는 서열 태그의 개수에서 가변성을 표시한다. 정규화 서열은 영향을 받은 표본을 하나 또는 그 이상의 영향을 받지 않은 표본과 구별할 수 있다. 일부 실행에서, 정규화 서열은 다른 잠재적 정규화 서열, 예를 들면, 다른 염색체와 비교할 때, 영향을 받은 표본을 하나 또는 그 이상의 영향을 받지 않은 표본으로부터 최고로 또는 효과적으로 구별한다. 일부 구체예에서, 정규화 서열의 가변성은 표본 및 염기서열결정 실행 전체에 대하여 관심되는 서열에 대한 염색체 도스에서 가변성으로서 계산된다. 일부 구체예에서, 정규화 서열은 한 세트의 영향을 받지 않은 표본에서 확인된다.

[0081] "정규화 염색체," "정규화 분모 염색체," 또는 "정규화 염색체 서열"은 "정규화 서열"의 실례이다. "정규화 염색체 서열"은 단일 염색체 또는 일군의 염색체로 구성될 수 있다. 일부 구체예에서, 정규화 서열은 2개 또는 그 이상의 로버스트 염색체를 포함한다. 일정한 구체예에서, 로버스트 염색체는 모두 염색체, X, Y, 13, 18, 그리고 21 이외에 상염색체이다. "정규화 분절"은"정규화 서열"의 다른 실례이다. "정규화 분절 서열"은 한 염색체의 단일 분절로 구성될 수 있거나 또는 이것은 동일한 또는 상이한 염색체의 2개 또는 그 이상의 분절로 구성될 수 있다. 일정한 구체예에서, 정규화 서열은 가변성, 예를 들면, 과정-관련된, 염색체간 (실행내), 그리고 염기서열결정간 (실행간) 가변성을 정규화하는 것으로 의도된다.

[0082] 본원에서 용어 "구별가능성"은 하나 또는 그 이상의 영향을 받지 않은, 다시 말하면, 정상적인 표본을 하나 또는 그 이상의 영향을 받은, 다시 말하면, 이수체 표본과 구별할 수 있게 하는 정규화 염색체의 특징을 지칭한다. 최대 "구별가능성"을 표시하는 정규화 염색체는 한 세트의 유자격 표본 내에 관심되는 염색체에 대한 염색체 도스 및 하나 또는 그 이상의 영향을 받은 표본 내에 상응하는 염색체에서 관심되는 동일한 염색체에 대한 염색체 도스의 분포 사이에 최대 통계학적 차이를 제공하는 염색체 또는 일군의 염색체이다.

[0083] 본원에서 용어 "가변성"은 하나 또는 그 이상의 영향을 받지 않은, 다시 말하면, 정상적인 표본을 하나 또는 그 이상의 영향을 받은, 다시 말하면, 이수체 표본과 구별할 수 있게 하는 정규화 염색체의 다른 특징을 지칭한다. 한 세트의 유자격 표본에서 계측되는 정규화 염색체의 가변성은 이것이 정규화 파라미터로서 역할을 하는 관심되는 염색체에 매핑되는 서열 태그의 개수에서 가변성을 근사하는, 이것에 매핑되는 서열 태그의 개수에서 가변성을 지칭한다.

[0084] 본원에서 용어 "서열 태그 밀도"는 참조 유전체 서열에 매핑되는 서열 리드의 개수를 지칭한다, 예를 들면, 염색체 21에 대한 서열 태그 밀도는 참조 유전체의 염색체 21에 매핑되는, 염기서열결정 방법에 의해 산출된 서열 리드의 개수이다.

[0085] 본원에서 용어 "서열 태그 밀도 비율"은 참조 유전체의 염색체, 예를 들면, 염색체 21에 매핑되는 서열 태그의

개수 대 참조 유전체 염색체의 길이의 비율을 지칭한다.

[0086] 본원에서 용어 "서열 도스 (sequence dose)"는 관심되는 서열에 대해 확인된 서열 태그의 개수를 정규화 서열에 대해 확인된 서열 태그의 개수에 관련시키는 파라미터를 지칭한다. 일부 경우에, 서열 도스는 관심되는 서열에 대한 서열 태그 리드 커버리지 대 정규화 서열에 대한 서열 태그 리드 커버리지의 비율이다. 일부 경우에, 서열 도스는 관심되는 서열의 서열 태그 밀도를 정규화 서열의 서열 태그 밀도에 관련시키는 파라미터를 지칭한다. "시험 서열 도스"는 관심되는 서열, 예를 들면, 염색체 21의 서열 태그 밀도를 시험 표본에서 결정된 정규화 서열, 예를 들면, 염색체 9의 서열 태그 밀도에 관련시키는 파라미터이다. 유사하게, "유자격 서열 도스"는 관심되는 서열의 서열 태그 밀도를 유자격 표본에서 결정된 정규화 서열의 서열 태그 밀도에 관련시키는 파라미터이다.

[0087] 용어 "리드 커버리지 (coverage)"는 규정된 서열에 매핑된 서열 태그의 존재비를 지칭한다. 리드 커버리지는 서열 태그 밀도 (또는 서열 태그의 수치), 서열 태그 밀도 비율, 정규화된 리드 커버리지 양, 조정된 리드 커버리지 값 등에 의해 정량적으로 표시될 수 있다.

[0088] 용어 "리드 커버리지 양"은 미가공 리드 커버리지의 변형이고, 그리고 종종, 유전체의 영역, 예를 들면, 빈 내에 서열 태그의 상대적 양 (때때로, 수치로 불린다)을 나타낸다. 리드 커버리지 양은 유전체의 영역에 대한 미가공 리드 커버리지 또는 수치를 정규화하고, 조정하고 및/또는 교정함으로써 획득될 수 있다. 가령, 영역에 대한 정규화된 리드 커버리지 양은 영역에 매핑된 서열 태그 수치를 전체 유전체에 매핑된 총수 서열 태그로 나눗셈함으로써 획득될 수 있다. 정규화된 리드 커버리지 양은 상이한 표본 전체에 대하여 빈의 리드 커버리지의 비교를 허용하는데, 이것은 상이한 깊이의 염기서열결정을 가질 수 있다. 이것은 서열 도스와 상이한데, 후자가 전형적으로, 전체 유전체의 부분집합에 매핑된 태그 수치로 나눗셈함으로써 획득된다는 점에서 그러하다. 부분집합은 정규화 분절 또는 염색체이다. 리드 커버리지 양은 정규화되는 지의 여부에 상관없이, 유전체 상에서 영역별로 전역 프로필 변동, G-C 분율 변동, 로버스트 염색체에서 이상점 등에 대해 교정될 수 있다.

[0089] 본원에서 용어 "차세대 염기서열결정 (NGS)"은 클론에 의해 증폭된 분자 및 단일 핵산 분자의 대량으로 병렬 염기서열결정을 허용하는 염기서열결정 방법을 지칭한다. NGS의 무제한적 실례는 가역성 염료 종결인자를 이용한 합성에 의한 염기서열결정, 그리고 결찰에 의한 염기서열결정을 포함한다.

[0090] 본원에서 용어 "파라미터"는 물리적 성질을 특징짓는 수치 값을 지칭한다. 빈번하게, 파라미터는 정량적 데이터 세트 및/또는 정량적 데이터 세트 사이에 수치 관계를 수치적으로 특징짓는다. 가령, 염색체에 매핑된 서열 태그의 개수 및 이들 태그가 매핑되는 염색체의 길이 사이에 비율 (또는 비율의 함수)이 파라미터이다.

[0091] 본원에서 용어 "역치값" 및 "유자격 역치값"은 표본, 예를 들면, 의학적 상태를 앓는 것으로 의심되는 생물체로부터 핵산을 내포하는 시험 표본을 특징짓는 컷오프로서 이용되는 임의의 숫자를 지칭한다. 역치는 이런 파라미터 값을 발생시키는 표본이 생물체가 의학적 상태를 앓는다는 것을 암시하는 지를 결정하기 위해 파라미터 값과 비교될 수 있다. 일정한 구체예에서, 유자격 역치값은 유자격 데이터 세트를 이용하여 계산되고, 그리고 생물체에서 사본수 변동, 예를 들면, 이수성의 진단의 한계로서 역할을 한다. 역치가 본원에서 개시된 방법으로부터 획득된 결과에 의해 능가되면, 개체는 사본수 변동, 예를 들면, 삼염색체성 21로 진단될 수 있다. 본원에서 설명된 방법에 대한 적절한 역치값은 표본의 훈련 세트에 대해 계산된 정규화된 값 (가령, 염색체 도스, NCVs 또는 NSVs)을 분석함으로써 확인될 수 있다. 역치값은 유자격 (즉, 영향을 받지 않은) 표본 및 영향을 받은 표본 둘 모두를 포함하는 훈련 세트에서 유자격 (즉, 영향을 받지 않은) 표본을 이용하여 확인될 수 있다. 염색체 이수성을 갖는 것으로 알려진 훈련 세트에서 표본 (즉, 영향을 받은 표본)은 선택된 역치가 시험 세트에서 영향을 받은 표본을 영향을 받지 않은 표본과 구별하는데 유용하다는 것을 확증하는데 이용될 수 있다 (본원에서 실시예 참조). 역치의 선택은 사용자가 분류를 하고자 하는 신뢰의 수준에 의존한다. 일부 구체예에서, 적절한 역치값을 확인하는데 이용된 훈련 세트는 최소한 10, 최소한 20, 최소한 30, 최소한 40, 최소한 50, 최소한 60, 최소한 70, 최소한 80, 최소한 90, 최소한 100, 최소한 200, 최소한 300, 최소한 400, 최소한 500, 최소한 600, 최소한 700, 최소한 800, 최소한 900, 최소한 1000, 최소한 2000, 최소한 3000, 최소한 4000개, 또는 그 이상의 유자격 표본을 포함한다. 역치값의 진단적 유용성을 향상시키기 위해 유자격 표본의 더욱 큰 세트를 이용하는 것이 유리할 수 있다.

[0092] 용어 "빈 (bin)"은 서열의 분절 또는 유전체의 분절을 지칭한다. 일부 구체예에서, 빈은 서로 인접하고 유전체 또는 염색체 내에서 위치에 의해 분리된다. 각 빈은 참조 유전체에서 뉴클레오티드의 서열을 규정할 수 있다. 빈의 크기는 특정 적용 및 서열 태그 밀도에 의해 요구되는 분석에 따라 1 kb, 100 kb, 1Mb 등일 수 있다. 참조 서열 내에서 그들의 위치 이외에, 빈은 다른 특징, 예를 들면, 표본 리드 커버리지 및 서열 구조 특징, 예를 들

면, G-C 함량을 가질 수 있다.

[0093] 본원에서 용어 "마스크 적용 역치"는 서열 빈 내에 서열 태그의 개수에 근거된 값이 비교되는 양을 지칭하는데 이용되고, 여기서 마스크 적용 역치를 초과하는 값을 갖는 빈은 마스크가 적용된다. 일부 구체예에서, 마스크 적용 역치는 백분위수 순위, 절대적 수치, 매핑 품질평가점수, 또는 다른 적합한 값일 수 있다. 일부 구체예에서, 마스크 적용 역치는 복수의 영향을 받지 않은 표본 전체에 대하여 변동 계수의 백분위수 순위로서 규정될 수 있다. 다른 구체예에서, 마스크 적용 역치는 매핑 품질평가점수, 예를 들면, MapQ 점수로서 규정될 수 있는데, 이것은 서열 리드를 참조 유전체에 맞춰 정렬하는 신뢰도에 관계한다. 주목할 점은 마스크 적용 역치값이 사본수 변동 (CNV) 역치값과 상이하고, 후자가 CNV에 관련된 의학적 상태를 앓는 것으로 의심되는 생물체로부터 핵산을 내포하는 표본을 특징짓는 컷오프라는 것이다. 일부 구체예에서, CNV 역치값은 본원의 다른 곳에서 설명된 정규화된 염색체 값 (NCV) 또는 정규화된 분절 값 (NSV)에 상대적으로 규정된다.

[0094] 본원에서 용어 "정규화된 값"은 관심되는 서열 (가령, 염색체 또는 염색체 분절)에 대해 확인된 서열 태그의 개수를 정규화 서열 (가령, 정규화 염색체 또는 정규화 염색체 분절)에 대해 확인된 서열 태그의 개수에 관련시키는 수치 값을 지칭한다. 가령, "정규화된 값"은 본원의 다른 곳에서 설명된 바와 같은 염색체 도스일 수 있고, 또는 이것은 NCV일 수 있고, 또는 이것은 본원의 다른 곳에서 설명된 바와 같은 NSV일 수 있다.

[0095] 용어 "리드"는 핵산 표본의 부분으로부터 서열 리드를 지칭한다. 전형적으로, 리드는 표본 내에 인접한 염기쌍의 짧은 서열을 나타내지만, 반드시 그러한 것은 아니다. 리드는 표본 부분의 염기쌍 서열 (ATCG에서)에 의해 상징적으로 표현될 수 있다. 이것은 기억 장치에서 저장되고, 그리고 이것이 참조 서열에 정합하거나 또는 다른 규준에 부합하는 지를 결정하기 위해 적절히 처리될 수 있다. 리드는 염기서열결정 기구로부터 직접적으로 또는 표본에 관한 저장된 서열 정보로부터 간접적으로 획득될 수 있다. 일부 경우에, 리드는 더욱 큰 서열 또는 영역을 확인하는데 이용될 수 있는, 예를 들면, 염색체 또는 유전체 영역 또는 유전자에 정렬되고 특이적으로 배정될 수 있는 충분한 길이 (가령, 최소한 약 25 bp)의 DNA 서열이다.

[0096] 용어 "유전체 리드"는 개체의 전체 유전체 내에 임의의 분절의 리드에 관하여 이용된다.

[0097] 본원에서 용어 "서열 태그"는 용어 "매핑된 서열 태그"와 교체가능하게 이용되고 정렬에 의해 더욱 큰 서열, 예를 들면, 참조 유전체에 특이적으로 배정된, 다시 말하면, 매핑된 서열 리드를 지칭한다. 매핑된 서열 태그는 참조 유전체에 독특하게 매핑된다, 다시 말하면, 이들은 참조 유전체의 단일 위치에 배정된다. 달리 명시되지 않으면, 참조 서열 상에서 동일한 서열에 매핑하는 태그는 1회 계수된다. 태그는 데이터 구조 또는 데이터의 다른 기계조립으로서 제공될 수 있다. 일정한 구체예에서, 태그는 리드 서열 및 상기 리드에 대한 연관된 정보, 예를 들면, 유전체 내에 서열의 위치, 예를 들면, 염색체 상에서 위치를 내포한다. 일정한 구체예에서, 위치는 양성 가닥 방향에 대해 특정된다. 태그는 참조 유전체에 맞춰 정렬함에 있어서 부정합의 한계 양을 제공하도록 규정될 수 있다. 일부 구체예에서, 참조 유전체 상에서 하나 이상의 위치에 매핑될 수 있는 태그, 다시 말하면, 독특하게 매핑하지 않는 태그는 분석에 포함될 수 없다.

[0098] 용어 "비다중 서열 태그"는 동일한 부위에 매핑하지 않는 서열 태그를 지칭하고, 이것은 일부 구체예에서 정규화된 염색체 값 (NCVs)을 결정하는 목적으로 계수된다. 때때로, 복수 서열 리드가 참조 유전체 상에서 동일한 위치에 정렬되고, 잉여 또는 중복 서열 태그를 산출한다. 일부 구체예에서, 동일한 위치에 매핑하는 중복 서열 태그는 NCV를 결정하는 목적으로, 제외되거나 또는 하나의 "비다중 서열 태그"로서 계수된다. 일부 구체예에서, 비-배제된 부위에 정렬된 비다중 서열 태그가 NCV를 결정하기 위한 "비-배제된-부위 수치" (NES 수치)를 산출하기 위해 계수된다.

[0099] 용어 "부위"는 참조 유전체 상에서 독특한 위치 (즉, 염색체 ID, 염색체 위치 및 방향)를 지칭한다. 일부 구체예에서, 부위는 서열 상에서 잔기, 서열 태그, 또는 분절의 위치일 수 있다.

[0100] "배제된 부위"는 서열 태그를 계수하는 목적으로 배제된, 참조 유전체의 영역에서 발견되는 부위이다. 일부 구체예에서, 배제된 부위는 반복 서열을 내포하는 염색체의 영역, 예를 들면, 동원체와 말단소립, 그리고 하나 이상의 염색체에 공통적인 염색체의 영역, 예를 들면, Y 염색체 상에 존재하고, 또한 X 염색체 상에 존재하는 영역에서 발견된다.

[0101] "비-배제된 부위" (NESs)는 서열 태그를 계수하는 목적으로 참조 유전체에서 배제되지 않는 부위이다.

[0102] "비-배제된-부위 수치" (NES 수치)는 참조 유전체 상에서 NES에 매핑되는 서열 태그의 개수이다. 일부 구체예에서, NES 수치는 NES에 매핑되는 비다중 서열 태그의 개수이다. 일부 구체예에서, 리드 커버리지 및 관련된 파라미터, 예를 들면, 정규화된 리드 커버리지 양, 전역 프로필 이전된 리드 커버리지 양, 그리고 염색체 도스가

NES 수치에 근거된다. 한 가지 실례에서, 염색체 도스는 관심되는 염색체에 대한 NES 수치의 숫자 대 정규화 염색체에 대한 NES 수치의 숫자의 비율로서 계산된다.

[0103] 정규화된 염색체 값 (NCV)은 시험 표본의 리드 커버리지를 한 세트의 훈련/유자격 표본의 리드 커버리지에 관련시킨다. 일부 구체예에서, NCV는 염색체 도스에 근거된다. 일부 구체예에서, NCV는 시험 표본 내에 관심되는 염색체의 염색체 도스 및 한 세트의 유자격 표본 내에 상응하는 염색체 도스의 평균 사이의 차이에 관계하고, 그리고 다음과 같이 계산될 수 있다:

[0104] $$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

[0105] 여기서 $\hat{\mu}_j$ 및 $\hat{\sigma}_j$ 는 각각, 한 세트의 유자격 표본에서 j-번째 염색체 도스에 대한 추정된 평균과 표준 편차이고, 그리고 x_{ij} 는 시험 표본 i에 대한 관찰된 j-번째 염색체 비율 (도스)이다.

[0106] 일부 구체예에서, NCV는 아래와 같이, 시험 표본 내에 관심되는 염색체의 염색체 도스를 동일한 흐름 셀 상에서 염기서열결정된 다중화된 표본 내에 상응하는 염색체 도스의 중앙값에 관련시킴으로써 "온 더 플라이" 계산될 수 있다:

[0107] $$NCV_{ij} = \frac{x_{ij} - M_j}{\hat{\sigma}_j}$$

[0108] 여기서 M_j 는 동일한 흐름 셀 상에서 염기서열결정된 한 세트의 다중화된 표본에서 j-번째 염색체 도스에 대한 추정된 중앙값이고; $\hat{\sigma}_j$ 는 하나 또는 그 이상의 흐름 셀 상에서 염기서열결정된 하나 또는 그 이상의 세트의 다중화된 표본에서 j-번째 염색체 도스에 대한 표준 편차이고, 그리고 x_i 는 시험 표본 i에 대한 관찰된 j-번째 염색체 도스이다. 이러한 구체예에서, 시험 표본 i는 M_j 가 결정되는 동일한 흐름 셀 상에서 염기서열결정된 다중화된 표본 중에서 하나이다.

[0109] 가령, 하나의 흐름 셀 상에서 64개 다중화된 표본 중에서 한 가지로서 염기서열결정되는, 시험 표본 A에서 관심되는 염색체 21의 경우에, 시험 표본 A에서 염색체 21에 대한 NCV는 흐름 셀 1 상에서, 또는 추가 흐름 셀, 예를 들면, 20의 64개 다중화된 표본에 대해 결정된 염색체 21에 대한 도스의 표준 편차에 의해 나눗셈된, 표본 A에서 염색체 21의 도스 - (마이너스) 64개 다중화된 표본에서 결정된 염색체 21에 대한 도스의 중앙값으로서 계산된다.

[0110] 본원에서 이용된 바와 같이, 용어 "정렬된," "정렬," 또는 "정렬하는"은 리드 또는 태그를 참조 서열에 비교하고, 그리고 따라서, 참조 서열이 리드 서열을 내포하는 지를 결정하는 과정을 지칭한다. 참조 서열이 리드를 내포하면, 리드는 참조 서열, 또는 일정한 구체예에서, 참조 서열에서 특정 위치에 매핑될 수 있다. 일부 경우에, 정렬은 리드가 특정 참조 서열의 구성원인 지의 여부 (즉, 리드가 참조 서열에서 존재하거나 또는 부재하는 지의 여부)를 단순히 말한다. 가령, 인간 염색체 13에 대한 참조 서열에 리드의 정렬은 리드가 염색체 13에 대한 참조 서열 내에 존재하는 지의 여부를 말할 것이다. 이러한 정보를 제공하는 도구는 세트 멤버십 검사기로 불릴 수 있다. 일부 경우에, 정렬은 리드 또는 태그가 매핑하는, 참조 서열 내에 위치를 부가적으로 지시한다. 가령, 참조 서열이 전체 인간 유전체 서열이면, 정렬은 리드가 염색체 13 상에 존재한다는 것을 지시할 수 있고, 그리고 리드가 염색체 13의 특정 가닥 및/또는 부위 상에 있다는 것을 더욱 지시할 수 있다.

[0111] 정렬된 리드 또는 태그는 참조 유전체로부터 공지된 서열에 그들의 핵산 분자의 순서의 면에서 정합으로서 확인되는 하나 또는 그 이상의 서열이다. 정렬은 비록 전형적으로 컴퓨터 알고리즘에 의해 실행되지만, 수동으로 행위될 수 있는데, 그 이유는 본원에서 개시된 방법을 실행하기 위한 합리적인 기간 내에 리드를 정렬하는 것이 불가능할 것이기 때문이다. 서열을 정렬하는 것으로부터 알고리즘의 한 가지 실례는 Illumina 유전체 분석 파이프라인의 일부로서 배포된 뉴클레오티드 데이터의 효율적인 국부 정렬 (ELAND) 컴퓨터 프로그램이다. 대안으로, Bloom 필터 또는 유사한 세트 멤버십 검사기가 리드를 참조 유전체에 맞춰 정렬하는데 이용될 수 있다. 전체적

으로 본원에 참조로서 편입되는, 2011년 10월 27일자 제출된 US 특허 출원 번호 61/552,374를 참조한다. 정렬에 있어서 서열 리드의 정합은 100% 서열 정합 또는 100%보다 적은 서열 정합 (비-완벽한 정합)일 수 있다.

[0112] 용어 "정렬 프로필"은 관심되는 참조 서열에서 염기쌍 빈으로서 확인될 수 있는 위치에 정렬된 서열 태그의 분포에 관하여 이용된다.

[0113] 본원에서 이용된 용어 "매핑"은 서열 리드를 더욱 큰 서열, 예를 들면, 참조 유전체에 정렬에 의해 특이적으로 배정하는 것을 지칭한다.

[0114] 본원에서 이용된 바와 같이, 용어 "참조 유전체" 또는 "참조 서열"은 개체로부터 확인된 서열을 참조하는데 이용될 수 있는, 임의의 생물체 또는 바이러스의 임의의 특정 공지된 유전체 서열 (부분적인 또는 완전한 지에 상관없이)을 지칭한다. 가령, 인간 개체뿐만 아니라 많은 다른 생물체에 대해 이용된 참조 유전체는 ncbi.nlm.nih.gov에서 the National Center for Biotechnology Information에서 발견된다. "유전체"는 핵산 서열에서 발현되는, 생물체 또는 바이러스의 완전한 유전 정보를 지칭한다.

[0115] 다양한 구체예에서, 참조 서열은 이것에 정렬되는 리드보다 훨씬 크다. 가령, 이것은 최소한 약 100 배, 또는 최소한 약 1000 배, 또는 최소한 약 10,000 배, 또는 최소한 약 10^5 배, 또는 최소한 약 10^6 배, 또는 최소한 약 10^7 배 클 수 있다.

[0116] 한 가지 실례에서, 참조 서열은 전장 인간 유전체의 서열이다. 이런 서열은 유전체 참조 서열로서 지칭될 수 있다. 다른 실례에서, 참조 서열은 특정한 인간 염색체, 예를 들면, 염색체 13에 한정된다. 일부 구체예에서, 참조 Y 염색체는 인간 유전체 이형 hg19로부터 Y 염색체 서열이다. 이런 서열은 염색체 참조 서열로서 지칭될 수 있다. 참조 서열의 다른 실례는 다른 종의 유전체뿐만 아니라 임의의 종의 염색체, 아염색체 영역 (가령, 가닥) 등을 포함한다.

[0117] 다양한 구체예에서, 참조 서열은 복수 개체로부터 유래된 공통 서열 또는 다른 조합이다. 하지만, 일정한 적용에서, 참조 서열은 특정 개체로부터 취해질 수 있다.

[0118] 본원에서 용어 "임상적으로-유관한 서열"은 유전자 또는 질환 상태와 연관되거나 또는 연루되는 것으로 알려져 있거나 또는 의심되는 핵산 서열을 지칭한다. 임상적으로-유관한 서열의 부재 또는 존재를 결정하는 것은 의학적 상태의 진단을 결정하거나 또는 질병의 진단을 확증하고, 또는 질환의 발달에 대한 예후를 제공하는데 유용할 수 있다.

[0119] 본원에서 핵산 또는 핵산의 혼합물의 문맥에서 이용될 때 용어 "유래된"은 핵산(들)이 그들이 기원하는 공급원으로부터 획득되는 수단을 지칭한다. 가령, 한 구체예에서, 2개의 상이한 유전체로부터 유래되는 핵산의 혼합물은 이들 핵산, 예를 들면, cfDNA가 자연발생 과정, 예를 들면, 괴사 또는 아폽토시스를 통해 세포에 의해 자연적으로 방출되었다는 것을 의미한다. 다른 구체예에서, 2개의 상이한 유전체로부터 유래되는 핵산의 혼합물은 이들 핵산이 개체로부터 2가지 상이한 유형의 세포로부터 추출되었다는 것을 의미한다.

[0120] 본원에서 특정한 정량적 값을 획득하는 문맥에서 이용될 때 용어 "에 근거된"은 특정한 정량적 값을 출력으로서 계산하기 위해 다른 양을 입력으로서 이용하는 것을 지칭한다.

[0121] 본원에서 용어 "환자 표본"은 환자, 다시 말하면, 의학적 관심, 간호 또는 치료의 수용자로부터 획득된 생물학적 표본을 지칭한다. 환자 표본은 본원에서 설명된 표본 중에서 한 가지일 수 있다. 일정한 구체예에서, 환자 표본은 비침습성 시술에 의해 획득된다, 예를 들면, 말초혈 표본 또는 대변 표본. 본원에서 설명된 방법은 인간에 한정될 필요가 없다. 따라서, 다양한 수의학적 적용이 예기되는데, 이러한 사례에서 환자 표본은 비인간 포유동물 (가령, 고양이, 돼지, 말, 소 등)로부터 표본일 수 있다.

[0122] 본원에서 용어 "혼합된 표본"은 상이한 유전체로부터 유래되는 핵산의 혼합물을 내포하는 표본을 지칭한다.

[0123] 본원에서 용어 "모계 표본"은 임신 개체, 예를 들면, 여성으로부터 획득된 생물학적 표본을 지칭한다.

[0124] 본원에서 용어 "생물학적 유체"는 생물학적 공급원으로부터 채취된 액체를 지칭하고, 그리고 예로서, 혈액, 혈청, 혈장, 객담, 세척액, 뇌척수액, 소변, 정액, 땀, 눈물, 타액 등을 포함한다. 본원에서 이용된 바와 같이, 용어 "혈액," "혈장" 및 "혈청"은 분획물 또는 이들의 처리된 부분을 명시적으로 포괄한다. 유사하게, 표본이 생검, 면봉, 도말 등으로부터 채취되는 경우에, "표본"은 생검, 면봉, 도말 등으로부터 유래된 처리된 분획물 또는 부분을 명시적으로 포괄한다.

[0125] 본원에서 용어 "모계 핵산" 및 "태아 핵산"은 각각, 임신 암컷 개체의 핵산 및 임신 암컷에 의해 잉태되는 태아의 핵산을 지칭한다.

[0126] 본원에서 이용된 바와 같이, 용어 "에 상응하는"은 때때로, 상이한 개체의 유전체 내에 존재하고, 그리고 모든 유전체에서 반드시 동일한 서열을 갖는 것은 아니지만, 관심되는 서열, 예를 들면, 유전자 또는 염색체의 유전 정보보다는 동일성을 제공하는데 역할을 하는 핵산 서열, 예를 들면, 유전자 또는 염색체를 지칭한다.

[0127] 본원에서 이용된 바와 같이, 원하는 표본과 관련하여 이용된 용어 "실제적으로 무세포"는 표본과 정상적으로 연관된 세포 성분이 제거되는 원하는 표본의 제조물을 포괄한다. 가령, 혈장 표본은 이것과 정상적으로 연관되는 혈액 세포, 예를 들면, 적혈구를 제거함으로써, 실제적으로 무세포가 되도록 만들어진다. 일부 구체예에서, 실제적으로 무세포 표본은 만약 그렇지 않으면, CNV에 대해 시험되는 원하는 유전 물질에 기여할 세포가 제거되도록 처리된다.

[0128] 본원에서 이용된 바와 같이, 용어 "태아 분획물"은 태아와 모계 핵산을 포함하는 표본 내에 존재하는 태아 핵산의 분획물을 지칭한다. 태아 분획물은 종종, 모체의 혈액에서 cfDNA를 특징짓는데 이용된다.

[0129] 본원에서 이용된 바와 같이 용어 "염색체"는 생존 세포의 유전-보유 유전자 운반체를 지칭하고, 이것은 DNA 및 단백질 성분 (특히, 히스톤)을 포함하는 염색질 가닥으로부터 유래된다. 전통적인 국제적으로 인식된 개별 인간 유전체 염색체 넘버링 시스템이 본원에서 이용된다.

[0130] 본원에서 이용된 바와 같이, 용어 "폴리뉴클레오티드 길이"는 참조 유전체의 서열 또는 영역에서 핵산 분자 (뉴클레오티드)의 절대적 숫자를 지칭한다. 용어 "염색체 길이"는 염기쌍에서 제공된, 예를 들면, 월드와이드웹 상에서 |genome|.|ucsc|.|edu/cgi-bin/hgTracks?hgsid=167155613&chromInfoPage=에서 발견되는 인간 염색체의 NCBI36/hg18 어셈블리에서 제공된 염색체의 공지된 길이를 지칭한다.

[0131] 본원에서 용어 "개체"는 인간 개체뿐만 아니라 비인간 개체, 예를 들면, 포유동물, 무척추동물, 척추동물, 균류, 효모, 세균, 그리고 바이러스를 지칭한다. 비록 본원에서 실례가 인간에 관계하고, 그리고 상기 용어가 인간 우려에 일차적으로 관계하지만, 본원에서 개시된 개념은 임의의 식물 또는 동물로부터 유전체에 적용가능하고, 그리고 수의학, 동물 과학, 연구 실험실, 기타 등등의 분야에서 유용하다.

[0132] 본원에서 용어 "상태"는 모든 질환 및 장애를 포함하는 광범위한 용어로서 "의학적 상태"를 지칭하지만, 개인의 건강에 영향을 주거나, 의학적 도움으로부터 이익을 얻거나, 또는 의학적 치료에 대한 함의를 가질지도 모르는 [손상] 및 정상적인 건강 환경, 예를 들면, 임신을 포함할 수 있다.

[0133] 본원에서 염색체 이수성에 관하여 이용될 때 용어 "완전한"은 전체 염색체의 획득 또는 상실을 지칭한다.

[0134] 본원에서 염색체 이수성에 관하여 이용될 때 용어 "부분적인"은 염색체의 부분, 다시 말하면, 분절의 획득 또는 상실을 지칭한다.

[0135] 본원에서 용어 "모자이크"는 단일 수정란으로부터 발달한 한 개체에서 상이한 핵형을 갖는 두 세포 개체군의 존재를 표시하는 것을 지칭한다. 모자이크현상은 단지 성체 세포의 부분집합에만 전파되는, 발달 동안 돌연변이로부터 발생할 수 있다.

[0136] 본원에서 용어 "비-모자이크"는 한 핵형의 세포로 구성된 생물체, 예를 들면, 인간 태아를 지칭한다.

[0137] 본원에서 염색체 도스를 결정하는 것에 관하여 이용될 때 용어 "염색체를 이용하는"은 염색체에 대해 획득된 서열 정보, 다시 말하면, 염색체에 대해 획득된 서열 태그의 개수를 이용하는 것을 지칭한다.

[0138] 본원에서 이용된 바와 같이, 용어 "감수성"은 진양성 및 가음성의 합계에 의해 나눗셈된 진양성의 숫자에 필적한다.

[0139] 본원에서 이용된 바와 같이, 용어 "특이성"은 진음성 및 가양성의 합계에 의해 나눗셈된 진음성의 숫자에 필적한다.

[0140] 본원에서 용어 "농축한다"는 모계 표본의 부분에서 내포된 다형성 표적 핵산을 증폭하고, 그리고 증폭된 산물을, 상기 부분이 제거된 모계 표본의 나머지와 결합하는 과정을 지칭한다. 가령, 모계 표본의 나머지는 본래 모계 표본일 수 있다.

[0141] 본원에서 용어 "본래 모계 표본"은 임신 개체, 예를 들면, 여성으로부터 획득된 비-농축된 생물학적 표본을 지칭하고, 상기 개체는 다형성 표적 핵산을 증폭하기 위해 일부분이 제거된 공급원으로서 역할을 한다. "본래 표

본"은 임신 개체로부터 획득된 임의의 표본, 그리고 이의 처리된 분획물, 예를 들면, 모계 혈장 표본으로부터 추출된 정제된 cfDNA 표본일 수 있다.

[0142] 본원에서 이용된 바와 같이, 용어 "프라이머"는 연장 산물의 합성을 유도하는 조건 (가령, 이들 조건은 뉴클레오티드, 유도제, 예를 들면, DNA 중합효소, 그리고 적합한 온도와 pH를 포함한다) 하에 배치될 때 합성의 개시의 포인트로서 행동할 수 있는 단리된 올리고뉴클레오티드를 지칭한다. 프라이머는 바람직하게는, 증폭에서 최고 효율을 위해 단일 가닥이지만, 대안으로 이중 가닥일 수도 있다. 이중 가닥이면, 프라이머는 먼저 연장 산물을 제조하는데 이용되기 전에, 이의 가닥이 분리되도록 처리된다. 바람직하게는, 프라이머는 올리고데옥시리보뉴클레오티드이다. 프라이머는 유도제의 존재에서 연장 산물의 합성을 기폭할 만큼 충분히 길어야 한다. 프라이머의 정확한 길이는 온도, 프라이머의 공급원, 방법의 용도, 그리고 프라이머 설계에 이용된 파라미터를 비롯한 많은 인자에 의존할 것이다.

[0143] 관용구 "투여되도록 유발한다"는 개체에 문제의 작용제(들)/화합물(들)의 투여를 제어하고 및/또는 허용하는 의학적 전문가 (가령, 의사), 또는 개체의 의료를 제어하거나 또는 주도하는 개인에 의해 취해진 행위를 지칭한다. 투여되도록 유발하는 것은 적절한 치료적 또는 예방적 섭생을 진단하고 및/또는 결정하고, 및/또는 개체에 대한 특정 작용제(들)/화합물을 처방하는 것을 수반할 수 있다. 이런 처방은 예로서, 처방전 양식의 초안을 작성하고, 병력을 주해하고, 기타 등등을 포함할 수 있다. 유사하게, 예로서 진단적 절차에 대해 "수행되도록 유발한다"는 개체에 또는 개체에서 한 가지 또는 그 이상의 진단적 프로토콜의 성과를 제어하고 및/또는 허용하는 의학적 전문가 (가령, 의사), 또는 개체의 의료를 제어하거나 또는 주도하는 개인에 의해 취해진 행위를 지칭한다.

[0144] **도입**

[0145] 2개 또는 그 이상의 상이한 유전체로부터 유래된 핵산의 혼합물을 포함하고, 그리고 관심되는 하나 또는 그 이상의 서열의 양에서 다른 것으로 알려져 있거나 또는 의심되는 시험 표본 내에 상이한 관심되는 서열의 사본수 및 사본수 변동 (CNV)을 결정하기 위한 방법, 기구, 그리고 시스템이 본원에서 개시된다. 본원에서 개시된 방법과 기구에 의해 결정된 사본수 변동은 전체 염색체의 획득 또는 상실, 현미경적으로 가시적인 매우 큰 염색체 분절을 수반하는 변경, 그리고 크기에서 단일 뉴클레오티드에서부터 킬로베이스 (kb) 내지 메가베이스 (Mb)까지의 범위에서 변하는 DNA 분절의 초현미경적 사본수 변동의 존재비를 포함한다.

[0146] 일부 구체예에서, 모계와 태아 무세포 DNA를 내포하는 모계 표본을 이용하여 태아의 사본수 변동 (CNV)을 결정하기 위한 방법이 제공된다. 본원에서 개시된 일부 구체예는 표본내 GC-함량 바이어스를 제거함으로써, 서열 데이터 분석의 감수성 및/또는 특이성을 향상시키는 방법을 제공한다. 일부 구체예에서, 표본내 GC-함량 바이어스의 제거는 영향을 받지 않은 훈련 표본 전체에 대하여 공통적인 체계적 변동에 대해 교정된 서열 데이터에 근거된다.

[0147] 개시된 일부 구체예는 낮은 잡음 및 높은 신호에서 서열 리드 커버리지 양을 결정하는 방법을 제공하고, 전통적인 방법에 의해 획득된 서열 리드 커버리지 양에 비하여 향상된 감수성, 선택성, 및/또는 효율로 사본수와 CNV에 관련된 다양한 유전적 장애를 결정하기 위한 데이터를 제공한다. 묘사된 과정은 고려 중인 유전체 (가령, 태아의 유전체)로부터 DNA의 상대적으로 낮은 분율을 갖는 표본에서 신호를 향상시키는데 특히 효과적인 것으로 밝혀졌다. 이런 표본의 실례는 형제간 쌍둥이, 삼둥이 등을 임신한 개체로부터 모체 혈액 표본이고, 여기서 상기 과정은 이들 태아 중에서 하나의 유전체에서 사본수 변동을 사정한다.

[0148] 이들 방법은 임의의 태아 이수성의 CNV, 그리고 다양한 의학적 상태와 연관되는 것으로 공지된 또는 의심되는 CNV를 결정하는데 적용가능하다. 인간 개체를 수반하는 일부 구체예에서, 본 발명 방법에 따라 결정될 수 있는 CNV는 염색체 1-22, X와 Y 중에서 임의의 한 가지 또는 그 이상의 삼염색체성 및 일염색체성, 다른 염색체 다염색체성, 그리고 이들 염색체 중에서 임의의 한 가지 또는 그 이상의 분절의 결실 및/또는 중복을 포함하는데, 이들은 시험 표본의 핵산을 단지 1회 염기서열결정함으로써 검출될 수 있다. 임의의 이수성은 시험 표본의 핵산을 단지 1회 염기서열결정함으로써 획득되는 염기서열결정 정보로부터 결정될 수 있다.

[0149] 인간 유전체에서 CNV는 인간 다양성 및 질환에 대한 소인에 유의미하게 영향을 준다 (Redon et al., Nature 23:444-454 [2006], Shaikh et al. Genome Res 19:1682-1690 [2009]). CNV는 상이한 기전을 통해 유전성 질환에 기여하고, 많은 경우에 유전자량의 불균형 또는 유전자 붕괴를 유발하는 것으로 알려져 있었다. 유전 질환과의 직접적인 상관에 더하여, CNV는 유해할 수 있는 표현형 변화를 매개하는 것으로 알려져 있다. 최근에, 여러 연구가 정상적인 대조와 비교하여 복합적인 장애, 예를 들면, 자폐증, ADHD, 그리고 정신분열병에서 희귀한 또

는 데노보 CNV의 증가된 부담을 보고하였는데, 이것은 희귀한 또는 독특한 CNV의 잠재적 병원성을 강조한다 (Sebat et al., 316:445 - 449 [2007]; Walsh et al., Science 320:539 - 543 [2008]). CNV는 일차적으로, 결실, 중복, 삽입, 그리고 불균형 전위 이벤트에 기인한 유전체 재배열로부터 발생한다.

[0150] 본원에서 설명된 방법과 기구는 차세대 염기서열결정 기술 (NGS)을 이용할 수 있는데, 이것은 대량으로 병렬 염기서열결정이다. 일정한 구체예에서, 클론에 의해 증폭된 DNA 주형 또는 단일 DNA 분자가 흐름 셀 내에서 대량으로 병렬 방식으로 염기서열결정된다 (가령, Volkerding et al. Clin Chem 55:641-658 [2009]; Metzker M Nature Rev 11:31-46 [2010]에서 설명된 바와 같이). 고처리량 서열 정보에 더하여, NGS는 각 서열 리드가 개별 클론 DNA 주형 또는 단일 DNA 분자를 나타내는 계수가능한 "서열 태그"라는 점에서, 정량적 정보를 제공한다. NGS의 염기서열결정 기술은 파이로시퀀싱, 가역성 염료 종결인자로 합성에 의한 염기서열결정, 올리고뉴클레오티드 프로브 결찰에 의한 염기서열결정 및 이온 반도체 염기서열결정을 포함한다. 개체 표본으로부터 DNA는 개별적으로 염기서열결정 (즉, 단일플렉스 염기서열결정)될 수 있고, 또는 복수 표본으로부터 DNA는 DNA 서열의 수억 개 리드를 산출하기 위해, 단일 염기서열결정 실행에서 색인된 유전체 분자로서 모아지고 염기서열결정 (즉, 멀티플렉스 염기서열결정)될 수 있다. 본 발명 방법에 따라 서열 정보를 획득하는데 이용될 수 있는 염기서열결정 기술의 실례가 아래에 설명된다.

[0151] DNA 표본을 이용하는 다양한 CNV 분석은 서열분석기로부터 서열 리드를 참조 서열에 정렬하거나 또는 매핑하는 것을 수반한다. 참조 서열은 전체 유전체의 서열, 염색체의 서열, 아염색체 영역의 서열 등일 수 있다. 참조 서열의 이들 특징으로 인해, Y 염색체의 CNV의 진단은 상염색체와 비교하여 고조된 기술적인 과제를 수반하는데, 그 이유는 Y 염색체의 리드 커버리지가 상염색체의 리드 커버리지 보다 낮고, 그리고 Y 염색체 상에서 반복된 서열이 리드의 정확한 위치로의 매핑을 복잡하게 만들기 때문이다. 약 10 Mb의 독특한 Y 서열이 현재 NGS 기술에 의해 접근가능하지만, 성별 검출은 모계 표본 내에 태아 cfDNA의 양이 모계 DNA의 것보다 최소한 1 크기 자릿수 낮은 태아 진단 권역에서 여전히 도전적인 과제로서 남아있는데, 이것은 비특이적 매핑의 문제를 강조한다.

[0152] 추가적으로, 일부 현재 염기서열결정 프로토콜은 극초단 리드, 예를 들면, 25mer 리드 및 태그를 활용한다. 염기서열결정 프로토콜의 과정에서 활용된 극초단 염기서열결정은 서열 정렬에 대한 기술적인 과제를 제공하였던 짧은 리드 길이를 산출하는데, 그 이유는 인간 유전체의 거의 절반이 반복에 의해 커버되기 때문인데, 이들 중에서 다수가 약 수십 년 동안 알려져 있다. 연산적 관점으로부터, 반복은 정렬에서 불명료를 창출하고, 이것은 차례로, 심지어 전체 염색체 계수 수준에서 바이어스와 오차를 발생시킬 수 있다.

[0153] **CNV 평가**

[0154] *CNV의 결절을 위한 방법*

[0155] 본원에서 개시된 방법에 의해 제공된 서열 리드 커버리지 값을 이용하여, 전통적인 방법에 의해 획득된 서열 리드 커버리지 값을 이용하는 것에 비하여 향상된 감수성, 선택성, 및/또는 효율로 서열, 염색체, 또는 염색체 분절의 사본수와 CNV에 관련된 다양한 유전적 장애를 결정할 수 있다. 가령, 일부 구체예에서, 마스크가 적용된 참조 서열은 태아와 모계 핵산 분자를 포함하는 모계 시험 표본 내에 임의의 2가지 또는 그 이상의 상이한 완전한 태아 염색체 이수성의 존재 또는 부재를 결정하는데 이용된다. 아래에 제공된 예시적인 방법은 리드를 참조 서열 (참조 유전체 포함)에 맞춰 정렬한다. 정렬은 마스크가 적용되지 않은 또는 마스크가 적용된 참조 서열 상에서 수행될 수 있고, 따라서 참조 서열에 매핑된 서열 태그를 산출한다. 일부 구체예에서, 참조 서열의 마스크가 적용되지 않은 분절에 속하는 서열 태그만 사본수 변동을 결정하는데 고려된다.

[0156] 일부 구체예에서, 모계 시험 표본 내에 임의의 완전한 태아 염색체 이수성의 존재 또는 부재를 결정하기 위한 방법은 (a) 모계 시험 표본 내에 태아와 모계 핵산에 대한 서열 정보를 획득하고; (b) 서열 정보 및 앞서 설명된 방법을 이용하여, 염색체 1-22, X와 Y에서 선택되는 관심되는 염색체 각각에 대한 서열 태그의 개수 또는 그것들로부터 유래된 서열 리드 커버리지 양을 확인하고, 그리고 하나 또는 그 이상의 정규화 염색체 서열에 대한 서열 태그의 개수를 확인하고; (c) 관심되는 염색체 각각에 대해 확인된 서열 태그의 개수 및 정규화 염색체 각각에 대해 확인된 서열 태그의 개수를 이용하여, 관심되는 염색체 각각에 대한 단일 염색체 도스를 계산하고; 그리고 (d) 각 염색체 도스를 역치값에 비교하고, 그리고 따라서, 모계 시험 표본 내에 임의의 완전한 태아 염색체 이수성의 존재 또는 부재를 결정하는 것을 포함한다.

[0157] 일부 구체예에서, 앞서 설명된 단계 (a)는 시험 표본의 태아와 모계 핵산 분자에 대한 상기 서열 정보를 획득하기 위해, 시험 표본의 핵산 분자 중에서 최소한 일부를 염기서열결정하는 것을 포함할 수 있다. 일부 구체예에

서, 단계 (c)는 관심되는 염색체 각각에 대한 단일 염색체 도스를, 관심되는 염색체 각각에 대해 확인된 서열 태그의 개수 및 정규화 염색체 서열(들)에 대해 확인된 서열 태그의 개수의 비율로서 계산하는 것을 포함한다. 일부 다른 구체예에서, 염색체 도스는 서열 태그의 개수로부터 유래된 처리된 서열 리드 커버리지 양에 근거된다. 일부 구체예에서, 단지 독특한, 비다중 서열 태그만 처리된 서열 리드 커버리지 양을 계산하는데 이용된다. 일부 구체예에서, 처리된 서열 리드 커버리지 양은 서열 태그 밀도 비율인데, 이것은 서열 길이에 의해 표준화된 서열 태그의 개수이다. 일부 구체예에서, 처리된 서열 리드 커버리지 양은 정규화된 서열 태그이고, 이것은 유전체의 전부 또는 실제적인 부분에 의해 나눗셈된 관심되는 서열의 서열 태그의 개수이다. 일부 구체예에서, 처리된 서열 리드 커버리지 양은 관심되는 서열의 전역 프로필에 따라 조정된다. 일부 구체예에서, 처리된 서열 리드 커버리지 양은 GC 함량 및 시험되는 표본에 대한 서열 리드 커버리지 사이에 표본내 상관에 따라 조정된다. 일부 구체예에서, 처리된 서열 리드 커버리지 양은 이들 과정의 조합으로부터 발생하는데, 이들은 본원의 다른 곳에서 더욱 설명된다.

[0158] 일부 구체예에서, 염색체 도스가 관심되는 염색체 각각에 대한 처리된 서열 리드 커버리지 양 및 정규화 염색체 서열(들)에 대한 처리된 서열 리드 커버리지 양의 비율로서 계산된다.

[0159] 상기 구체예 중에서 임의의 한 가지에서, 완전한 염색체 이수성은 완전한 염색체 삼염색체성, 완전한 염색체 일염색체성 및 완전한 염색체 다염색체성에서 선택된다. 완전한 염색체 이수성은 염색체 1-22, X와 Y 중에서 임의의 한 가지의 완전한 이수성에서 선택된다. 가령, 상기 상이한 완전한 태아 염색체 이수성은 삼염색체성 2, 삼염색체성 8, 삼염색체성 9, 삼염색체성 20, 삼염색체성 21, 삼염색체성 13, 삼염색체성 16, 삼염색체성 18, 삼염색체성 22, 47,XXX, 47,XYY, 그리고 일염색체성 X에서 선택된다.

[0160] 상기 구체예 중에서 임의의 한 가지에서, 단계 (a)-(d)는 상이한 모계 개체로부터 시험 표본에 대해 반복되고, 그리고 상기 방법은 각 시험 표본 내에 임의의 2가지 또는 그 이상의 상이한 완전한 태아 염색체 이수성의 존재 또는 부재를 결정하는 것을 포함한다.

[0161] 상기 구체예 중에서 임의의 한 가지에서, 상기 방법은 정규화된 염색체 값 (NCV)을 계산하는 것을 더욱 포함할 수 있고, 여기서 NCV는 아래와 같이, 염색체 도스를 한 세트의 유자격 표본 내에 상응하는 염색체 도스의 평균에 관련시킨다:

$$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

[0162]

[0163] 여기서 $\hat{\mu}_j$ 및 $\hat{\sigma}_j$ 는 각각, 한 세트의 유자격 표본에서 *j*-번째 염색체 도스에 대한 추정된 평균과 표준 편차이고, 그리고 x_{ij} 는 시험 표본 *i*에 대한 관찰된 *j*-번째 염색체 도스이다.

[0164] 일부 구체예에서, NCV는 아래와 같이, 시험 표본 내에 관심되는 염색체의 염색체 도스를 동일한 흐름 셀 상에서 염기서열결정된 다중화된 표본 내에 상응하는 염색체 도스의 중앙값에 관련시킴으로써 "온 더 플라이" 계산될 수 있다:

$$NCV_{ij} = \frac{x_{ij} - M_j}{\hat{\sigma}_j}$$

[0165]

[0166] 여기서 M_j 는 동일한 흐름 셀 상에서 염기서열결정된 한 세트의 다중화된 표본에서 j-번째 염색체 도스에 대한 추정된 중앙값이고; $\hat{\sigma}_j$ 는 하나 또는 그 이상의 흐름 셀 상에서 염기서열결정된 하나 또는 그 이상의 세트의 다중화된 표본에서 j-번째 염색체 도스에 대한 표준 편차이고, 그리고 x_i 는 시험 표본 i에 대한 관찰된 j-번째 염색체 도스이다. 이러한 구체예에서, 시험 표본 i는 M_j 가 결정되는 동일한 흐름 셀 상에서 염기서열결정된 다중화된 표본 중에서 하나이다.

[0167] 일부 구체예에서, 태아와 모계 핵산을 포함하는 모계 시험 표본 내에 상이한 부분적인 태아 염색체 이수성의 존

재 또는 부재를 결정하기 위한 방법이 제공된다. 상기 방법은 상기 개설된 바와 같이, 완전한 이수성을 검출하기 위한 방법과 유사한 절차를 수반한다. 하지만, 완전한 염색체를 분석하는 대신에, 염색체의 분절이 분석된다. US 특허 출원 공개 번호 2013/0029852를 참조하고, 이것은 참조로서 통합된다.

[0168] 도면 1은 일부 구체예에 따라서 사본수 변동의 존재를 결정하기 위한 방법을 보여준다. 작업 130과 135에서, 유자격 서열 태그 리드 커버리지 및 시험 서열 태그 리드 커버리지가 결정된다. 본 발명은 전통적인 방법에 비하여 향상된 감수성과 선택성을 제공하는 리드 커버리지 양을 결정하는 과정을 제공한다. 작업 130과 135는 별표에 의해 표시되고, 그리고 이들 작업이 선행 기술에 비하여 향상에 기여한다는 것을 지시하기 위해 굵은 선의 상자에 의해 강조된다. 일부 구체예에서, 서열 태그 리드 커버리지 양은 분석의 감수성과 선택성을 향상시키기 위해 정규화되고, 조정되고, 손질되고, 그리고 만약 그렇지 않으면 처리된다. 이들 과정은 본원의 다른 곳에서 더욱 설명된다.

[0169] 개요 관점으로부터, 상기 방법은 시험 표본의 CNV의 결정에서 유자격 훈련 표본의 정규화 서열을 이용한다. 일부 구체예에서, 유자격 훈련 표본은 영향을 받지 않고 정상적인 사본수를 갖는다. 정규화 서열은 실행내 및 실행간 가변성에 대한 치수를 정규화하기 위한 메커니즘을 제공한다. 정규화 서열은 관심되는 임의의 한 가지 서열, 예를 들면, 염색체 또는 이의 분절에 대한 정상적인 사본수를 갖는 세포를 포함하는 것으로 알려진 개체로부터 획득된 한 세트의 유자격 표본으로부터 서열 정보를 이용하여 확인된다. 정규화 서열의 결정은 도면 1에서 묘사된 방법의 구체예의 단계 110, 120, 130, 145와 146에서 개설된다. 일부 구체예에서, 정규화 서열은 시험 서열에 대한 서열 도스를 계산하는데 이용된다. 단계 150을 참조한다. 일부 구체예에서, 정규화 서열은 또한, 시험 서열의 서열 도스가 비교되는 역치를 계산하는데 이용된다. 단계 150을 참조한다. 정규화 서열 및 시험 서열로부터 획득된 서열 정보는 시험 표본에서 염색체 이수성의 통계학적으로 의미있는 확인을 결정하는데 이용된다 (단계 160).

[0170] 일부 구체예에 따라 사본수 변동의 존재를 결정하기 위한 방법의 상세로 돌아가서, 도면 1은 생물학적 표본 내에 관심되는 서열, 예를 들면, 염색체 또는 이의 분절의 CNV를 결정하기 위한 구체예의 흐름 도표 100를 제공한다. 일부 구체예에서, 생물학적 표본은 개체로부터 획득되고, 그리고 상이한 유전체에 의해 기여된 핵산의 혼합물을 포함한다. 상이한 유전체는 두 개체에 의해 표본에 기여될 수 있다, 예를 들면, 상이한 유전체는 태아 및 태아를 잉태하는 모체에 의해 기여된다. 또한, 상이한 유전체는 셋 또는 그 이상 개체에 의해 표본에 기여될 수 있다, 예를 들면, 상이한 유전체는 2개 또는 그 이상의 태아 및 이들 태아를 잉태하는 모체에 의해 기여된다. 대안으로, 이들 유전체는 동일한 개체, 예를 들면, 암 환자로부터 혈장 표본으로부터 이수체 암성 세포 및 정상적인 정배수 세포에 의해 표본에 기여된다.

[0171] 환자의 시험 표본을 분석하는 것과 별개로, 하나 또는 그 이상의 정규화 염색체 또는 하나 또는 그 이상의 정규화 염색체 분절이 관심되는 각 가능한 염색체에 대해 선별된다. 정규화 염색체 또는 분절은 환자 표본의 정상적인 시험으로부터 비동기적으로 확인되고, 이것은 임상적 세팅에서 발생할 수 있다. 다시 말하면, 정규화 염색체 또는 분절은 환자 표본을 시험하기에 앞서 확인된다. 정규화 염색체 또는 분절 및 관심되는 염색체 또는 분절 사이에 연관은 시험 동안 이용을 위해 저장된다. 아래에 설명된 바와 같이, 이런 연관은 전형적으로, 많은 표본의 시험에 걸쳐 있는 기간 동안 유지된다. 다음의 논의는 관심되는 개별 염색체 또는 분절에 대한 정규화 염색체 또는 염색체 분절을 선별하기 위한 구체예에 관계한다.

[0172] 유자격 정규화 서열을 확인하고, 그리고 시험 표본 내에 CNV의 통계학적으로 의미있는 확인을 결정하는데 이용을 위한 분산 값을 제공하기 위해, 한 세트의 유자격 표본이 획득된다. 단계 110에서, 복수의 생물학적 유자격 표본이 관심되는 임의의 한 가지 서열에 대한 정상적인 사본수를 갖는 세포를 포함하는 것으로 알려진 복수의 개체로부터 획득된다. 한 구체예에서, 유자격 표본은 세포유전학적 수단을 이용하여, 염색체의 정상적인 사본수를 갖는 것으로 확증되었던 태아를 임신한 모체로부터 획득된다. 생물학적 유자격 표본은 생물학적 유체, 예를 들면, 혈장, 또는 아래에 설명된 바와 같은 임의의 적절한 표본일 수 있다. 일부 구체예에서, 유자격 표본은 핵산 분자, 예를 들면, cfDNA 분자의 혼합물을 내포한다. 일부 구체예에서, 유자격 표본은 태아와 모계 cfDNA 분자의 혼합물을 내포하는 모계 혈장 표본이다. 정규화 염색체 및/또는 이들의 분절에 대한 서열 정보는 임의의 공지된 염기서열결정 방법을 이용하여, 핵산, 예를 들면, 태아와 모계 핵산 중에서 최소한 일부를 염기서열결정함으로써 획득된다. 바람직하게는, 본원의 다른 곳에서 설명된 차세대 염기서열결정 (NGS) 방법 중에서 임의의 한 가지가 태아와 모계 핵산을 단일 또는 클론에 의해 증폭된 분자로서 염기서열결정하는데 이용된다. 다양한 구체예에서, 유자격 표본은 염기서열결정에 앞서 및 염기서열결정 동안 아래에 개시된 바와 같이 처리된다. 이들은 본원에서 개시된 바와 같은 기구, 시스템, 그리고 키트를 이용하여 처리될 수 있다.

[0173] 단계 120에서, 유자격 표본에서 내포된 모든 유자격 핵산 각각의 최소한 일부가 염기서열결정되어 수백만 개의 서열 리드, 예를 들면, 36bp 리드가 산출되고, 이들은 참조 유전체, 예를 들면, hg18에 정렬된다. 일부 구체예에서, 서열 리드는 약 20bp, 약 25bp, 약 30bp, 약 35bp, 약 40bp, 약 45bp, 약 50bp, 약 55bp, 약 60bp, 약 65bp, 약 70bp, 약 75bp, 약 80bp, 약 85bp, 약 90bp, 약 95bp, 약 100bp, 약 110bp, 약 120bp, 약 130, 약 140bp, 약 150bp, 약 200bp, 약 250bp, 약 300bp, 약 350bp, 약 400bp, 약 450bp, 또는 약 500bp를 포함한다. 기술적인 진전은 페어드 엔드 리드가 산출될 때, 약 1000bp보다 큰 리드를 실시가능하게 하는 500bp보다 큰 싱글 엔드 리드를 가능하게 할 것으로 예상된다. 한 구체예에서, 매핑된 서열 리드는 36bp를 포함한다. 다른 구체예에서, 매핑된 서열 리드는 25bp를 포함한다.

[0174] 서열 리드는 참조 유전체에 정렬되고, 그리고 참조 유전체에 독특하게 매핑되는 리드는 서열 태그로서 알려져 있다. 마스크가 적용된 참조 서열의 마스크가 적용된 분절에 속하는 서열 태그는 CNV의 분석을 위해 계수되지 않는다.

[0175] 한 구체예에서, 20 내지 40bp 리드를 포함하는 최소한 약 3×10^6 유자격 서열 태그, 최소한 약 5×10^6 유자격 서열 태그, 최소한 약 8×10^6 유자격 서열 태그, 최소한 약 10×10^6 유자격 서열 태그, 최소한 약 15×10^6 유자격 서열 태그, 최소한 약 20×10^6 유자격 서열 태그, 최소한 약 30×10^6 유자격 서열 태그, 최소한 약 40×10^6 유자격 서열 태그, 또는 최소한 약 50×10^6 유자격 서열 태그가 참조 유전체에 독특하게 매핑하는 리드로부터 획득된다.

[0176] 단계 130에서, 유자격 표본에서 핵산을 염기서열결정하는 것으로부터 획득된 모든 태그는 유자격 서열 태그 리드 커버리지를 획득하기 위해 계수된다. 유사하게, 작업 135에서, 시험 표본으로부터 획득된 모든 태그는 시험 서열 태그 리드 커버리지를 획득하기 위해 계수된다. 본 발명은 전통적인 방법에 비하여 향상된 감수성과 선택성을 제공하는 리드 커버리지 양을 결정하는 과정을 제공한다. 작업 130과 135는 별표에 의해 표시되고, 그리고 이들 작업이 선행 기술에 비하여 향상에 기여한다는 것을 지시하기 위해 굵은 선의 상자에 의해 강조된다. 일부 구체예에서, 서열 태그 리드 커버리지 양은 분석의 감수성과 선택성을 향상시키기 위해 정규화되고, 조정되고, 손질되고, 그리고 만약 그렇지 않으면 처리된다. 이들 과정은 본원의 다른 곳에서 더욱 설명된다.

[0177] 모든 유자격 서열 태그가 유자격 표본 각각에서 매핑되고 계수됨에 따라서, 유자격 표본 내에 관심되는 서열, 예를 들면, 임상적으로-유관한 서열에 대한 서열 태그 리드 커버리지가 결정되고, 정규화 서열이 차후에 확인되는 추가 서열에 대한 서열 태그 리드 커버리지 역시 결정된다.

[0178] 일부 구체예에서, 관심되는 서열은 완전한 염색체 이수성과 연관되는 염색체, 예를 들면, 염색체 21이고, 그리고 유자격 정규화 서열은 염색체 이수성과 연관되지 않고 서열 태그 리드 커버리지에서 변동이 관심되는 서열(즉, 염색체), 예를 들면, 염색체 21의 것을 근사하는 완전한 염색체이다. 선별된 정규화 염색체(들)은 관심되는 서열의 서열 태그 리드 커버리지에서 변동을 최고로 근사하는 염색체 또는 일군의 염색체일 수 있다. 염색체 1-22, X와 Y 중에서 임의의 한 가지 또는 그 이상이 관심되는 서열일 수 있고, 그리고 하나 또는 그 이상의 염색체가 유자격 표본 내에 임의의 한 가지 염색체 1-22, X와 Y 각각에 대한 정규화 서열로서 확인될 수 있다. 정규화 염색체는 개별 염색체일 수 있고, 또는 본원의 다른 곳에서 설명된 바와 같은 일군의 염색체일 수 있다.

[0179] 다른 구체예에서, 관심되는 서열은 부분적인 이수성, 예를 들면, 염색체 결실 또는 삽입, 또는 불균형 염색체 전위와 연관된 염색체의 분절이고, 그리고 정규화 서열은 부분적인 이수성과 연관되지 않고 서열 태그 리드 커버리지에서 변동이 부분적인 이수성과 연관된 염색체 분절의 것을 근사하는 염색체 분절 (또는 분절의 군)이다. 선별된 정규화 염색체 분절(들)은 관심되는 서열의 서열 태그 리드 커버리지에서 변동을 최고로 근사하는 하나 또는 그 이상의 것일 수 있다. 임의의 한 가지 또는 그 이상의 염색체 1-22, X와 Y의 임의의 한 가지 또는 그 이상 분절이 관심되는 서열일 수 있다.

[0180] 다른 구체예에서, 관심되는 서열은 부분적인 이수성과 연관된 염색체의 분절이고, 그리고 정규화 서열은 전체 염색체 또는 염색체들이다. 또 다른 구체예에서, 관심되는 서열은 이수성과 연관된 전체 염색체이고, 그리고 정규화 서열은 이수성과 연관되지 않는 염색체 분절 또는 분절들이다.

[0181] 단일 서열 또는 일군의 서열이 임의의 한 가지 또는 그 이상의 관심되는 서열에 대한 정규화 서열(들)로서 유자격 표본에서 확인되는 지에 상관없이, 유자격 정규화 서열은 유자격 표본에서 결정된 바와 같은 관심되는 서열의 것을 최고로 또는 효과적으로 근사하는 서열 태그 리드 커버리지에서 변동을 갖도록 선택될 수 있다. 가령, 유자격 정규화 서열은 관심되는 서열을 정규화하는데 이용될 때 유자격 표본 전체에 대하여 가장 작은 가변성을

발생시키는 서열이다, 다시 말하면, 정규화 서열의 가변성은 유자격 표본에서 결정된 관심되는 서열의 것에 가장 가깝다. 달리 말하면, 유자격 정규화 서열은 유자격 표본 전체에 대하여 서열 도스 (관심되는 서열에 대한)에서 최소 변동을 발생시키도록 선별된 서열이다. 따라서, 상기 과정은 정규화 염색체로서 이용될 때, 관심되는 서열에 대한 실행간 염색체 도스에서 가장 작은 가변성을 발생시킬 것으로 예상되는 서열을 선별한다.

[0182] 임의의 한 가지 또는 그 이상의 관심되는 서열에 대해 유자격 표본에서 확인된 정규화 서열은 염기서열결정 라이브러리를 산출하고, 그리고 표본을 염기서열결정하는데 필요한 절차가 시간의 추이에서 본질적으로 변경되지 않는다면, 수일, 수주, 수개월, 그리고 아마도 수년에 걸쳐 시험 표본에서 이수성의 존재 또는 부재를 결정하기 위해 선택되는 정규화 서열로서 남아있다. 앞서 설명된 바와 같이, 이수성의 존재를 결정하기 위한 정규화 서열은 관심되는 서열에 대한 정규화 파라미터로서 이용되는 관심되는 서열의 가변성을 최고로 근사하는, 표본, 예를 들면, 상이한 표본 및 염기서열결정 실행, 예를 들면, 동일자 및/또는 상이한 일자에서 발생하는 염기서열결정 실행 사이에서 이것에 매핑되는 서열 태그의 개수에서 가변성 (아마도 다른 이유 중에서 특히)에 대해 선택된다. 이들 절차에서 실제적인 변경은 모든 서열에 매핑되는 태그의 개수에 영향을 줄 것이고, 이것은 차례로, 어떤 서열 또는 어떤 군의 서열이 관심되는 서열(들)의 것을 가장 가깝게 근사하는, 동일한 및/또는 상이한 염기서열결정 실행에서 동일자 또는 상이한 일자에서 표본 전체에 대하여 가변성을 가질 것인 지를 결정할 것이고, 이것은 정규화 서열의 세트가 다시 결정되도록 하는 것을 필요로 할 것이다. 절차에서 실제적인 변경은 염기서열결정 라이브러리를 제조하는데 이용된 실험실 프로토콜에서 변화 (이들은 단일플렉스 염기서열결정 대신에 멀티플렉스 염기서열결정을 위한 표본을 준비하는 것에 관련된 변화를 포함한다), 그리고 염기서열결정 플랫폼에서 변화 (이들은 염기서열결정하는데 이용된 화학에서 변화를 포함한다)를 포함한다.

[0183] 일부 구체예에서, 관심되는 특정 서열을 정규화하는데 선택되는 정규화 서열은 하나 또는 그 이상의 유자격, 표본을 하나 또는 그 이상의 영향을 받은 표본과 최고로 구별하는 서열인데, 이것은 정규화 서열이 최대 구별가능성을 갖는 서열이라는 것을 암시한다, 다시 말하면, 정규화 서열의 구별가능성은 영향을 받은 시험 표본이 다른 영향을 받지 않은 표본과 쉽게 구별되도록, 영향을 받은 시험 표본 내에 관심되는 서열에 최적 구별을 제공할 정도이다. 다른 구체예에서, 정규화 서열은 가장 작은 가변성 및 최대 구별가능성의 조합을 갖는 서열이다.

[0184] 구별가능성의 수준은 아래에 설명되고 실시예에서 도시된 바와 같이, 유자격 표본의 개체군에서 서열 도스, 예를 들면, 염색체 도스 또는 분절 도스 및 하나 또는 그 이상의 시험 표본에서 염색체 도스(들) 사이에 통계학적 차이로서 결정될 수 있다. 가령, 구별가능성은 t 검증 값으로서 수치적으로 표현될 수 있는데, 이것은 유자격 표본의 개체군에서 염색체 도스 및 하나 또는 그 이상의 시험 표본에서 염색체 도스(들) 사이에 통계학적 차이를 나타낸다. 유사하게, 구별가능성은 염색체 도스 대신에 분절 도스에 근거될 수 있다. 대안으로, 구별가능성은 정규화된 염색체 값 (NCV)으로 수치적으로 표현될 수 있는데, 이것은 NCV에 대한 분포가 정상적이기만 하면, 염색체 도스에 대한 z-점수이다. 유사하게, 염색체 분절이 관심되는 서열인 경우에, 분절 도스의 구별가능성은 정규화된 분절 값 (NSV)으로서 수치적으로 표현될 수 있는데, 이것은 NSV에 대한 분포가 정상적이기만 하면, 염색체 분절 도스에 대한 z-점수이다. z-점수를 결정함에 있어서, 한 세트의 유자격 표본에서 염색체 또는 분절 도스의 평균 및 표준 편차가 이용될 수 있다. 대안으로, 유자격 표본 및 영향을 받은 표본을 포함하는 훈련 세트에서 염색체 또는 분절 도스의 평균 및 표준 편차가 이용될 수 있다. 다른 구체예에서, 정규화 서열은 가장 작은 가변성 및 최대 구별가능성 또는 작은 가변성 및 큰 구별가능성의 최적 조합을 갖는 서열이다.

[0185] 상기 방법은 내재적으로 유사한 특징을 갖고, 표본 및 염기서열결정 실행 사이에서 유사하게 변하기 쉽고, 그리고 시험 표본에서 서열 도스를 결정하는데 유용한 서열을 확인한다.

[0186] *서열 도스의 결정*

[0187] 일부 구체예에서, 관심되는 하나 또는 그 이상의 염색체 또는 분절에 대한 염색체 또는 분절 도스는 도면 1에서 도시된 단계 146에서 설명된 바와 같이 모든 유자격 표본에서 결정되고, 그리고 정규화 염색체 또는 분절 서열은 단계 145에서 확인된다. 일부 정규화 서열은 서열 도스가 계산되기 전에 제공된다. 이후, 하나 또는 그 이상의 정규화 서열이 아래에 더욱 설명된 바와 같이 다양한 규준에 따라 확인된다. 단계 145를 참조한다. 일부 구체예에서, 가령, 확인된 정규화 서열은 모든 유자격 표본 전체에 대하여 관심되는 서열에 대한 서열 도스에서 가장 작은 가변성을 유발한다.

[0188] 단계 146에서, 계산된 유자격 태그 밀도에 근거하여, 관심되는 서열에 대한 유자격 서열 도스, 다시 말하면, 염색체 도스 또는 분절 도스가 관심되는 서열에 대한 서열 태그 리드 커버리지 및 정규화 서열이 단계 145에서 차후에 확인되는 추가 서열에 대한 유자격 서열 태그 리드 커버리지의 비율로서 결정된다. 확인된 정규화 서열은 시험 표본에서 서열 도스를 결정하기 위해 차후에 이용된다.

[0189] 한 구체예에서, 유자격 표본에서 서열 도스는 관심되는 염색체에 대한 서열 태그의 개수 및 유자격 표본에서 정규화 염색체 서열에 대한 서열 태그의 개수의 비율로서 계산되는 염색체 도스이다. 정규화 염색체 서열은 단일 염색체, 일군의 염색체, 한 염색체의 분절, 또는 상이한 염색체로부터 일군의 분절일 수 있다. 따라서, 관심되는 염색체에 대한 염색체 도스는 유자격 표본에서, 관심되는 염색체에 대한 태그의 개수 및 (i) 단일 염색체로 구성된 정규화 염색체 서열, (ii) 2개 또는 그 이상의 염색체로 구성된 정규화 염색체 서열, (iii) 한 염색체의 단일 분절로 구성된 정규화 분절 서열, (iv) 한 염색체로부터 2개 또는 그 이상의 분절로 구성된 정규화 분절 서열, 또는 (v) 2개 또는 그 이상 염색체의 2개 또는 그 이상 분절로 구성된 정규화 분절 서열에 대한 태그의 개수의 비율로서 결정된다. (i)-(v)에 따라 관심되는 염색체 21에 대한 염색체 도스를 결정하기 위한 실례는 다음과 같다: 관심되는 염색체, 예를 들면, 염색체 21에 대한 염색체 도스는 염색체 21의 서열 태그 리드 커버리지 및 다음의 서열 태그 리드 커버리지 중에서 한 가지: (i) 모든 나머지 염색체 각각, 다시 말하면, 염색체 1-20, 염색체 22, 염색체 X, 그리고 염색체 Y; (ii) 2개 또는 그 이상 나머지 염색체의 모든 가능한 조합; (iii) 다른 염색체, 예를 들면, 염색체 9의 분절; (iv) 한 가지 다른 염색체의 2개 분절, 예를 들면, 염색체 9의 2개 분절; (v) 2개의 상이한 염색체의 2개 분절, 예를 들면, 염색체 9의 분절 및 염색체 14의 분절의 비율로서 결정된다.

[0190] 다른 구체예에서, 유자격 표본에서 서열 도스는 염색체 도스와 대조되는 분절 도스인데, 상기 분절 도스는 전체 염색체가 아닌 관심되는 분절에 대한 서열 태그의 개수 및 유자격 표본 내에 정규화 분절 서열에 대한 서열 태그의 개수의 비율로서 계산된다. 정규화 분절 서열은 상기 논의된 정규화 염색체 또는 분절 서열 중에서 한 가지일 수 있다.

[0191] *정규화 서열의 확인*

[0192] 단계 145에서, 정규화 서열이 관심되는 서열에 대해 확인된다. 일부 구체예에서, 가령, 정규화 서열은 예로서, 모든 유자격 훈련 표본 전체에 대하여 관심되는 서열에 대한 서열 도스에서 가장 작은 가변성을 유발하는 계산된 서열 도스에 근거된 서열이다. 상기 방법은 유사한 특징을 내재적으로 갖고, 표본 및 염기서열결정 실행 사이에서 유사하게 변하기 쉽고, 그리고 시험 표본에서 서열 도스를 결정하는데 유용한 서열을 확인한다.

[0193] 하나 또는 그 이상의 관심되는 서열에 대한 정규화 서열은 한 세트의 유자격 표본에서 확인될 수 있고, 그리고 유자격 표본에서 확인된 서열은 각 시험 표본에서 이수성의 존재 또는 부재를 결정하기 위해, 각 시험 표본 (단계 150) 내에 하나 또는 그 이상의 관심되는 서열에 대한 서열 도스를 계산하는데 차후 이용된다. 관심되는 염색체 또는 분절에 대해 확인된 정규화 서열은 상이한 염기서열결정 플랫폼이 이용될 때 및/또는 염기서열결정되는 핵산의 정제 및/또는 염기서열결정 라이브러리의 제조에서 차이가 존재할 때, 상이할 수 있다. 본원에서 설명된 방법에 따른 정규화 서열의 이용은 이용되는 표본 준비 및/또는 염기서열결정 플랫폼과 상관없이, 염색체 또는 이의 분절의 사본수에서 변이의 특정하고 민감한 척도를 제공한다.

[0194] 일부 구체예에서, 하나 이상의 정규화 서열이 확인된다, 다시 말하면, 상이한 정규화 서열이 관심되는 한 서열에 대해 결정될 수 있고, 그리고 복수 서열 도스가 관심되는 한 서열에 대해 결정될 수 있다. 가령, 관심되는 염색체 21에 대한 염색체 도스에서 변동, 예를 들면, 변동 계수 (CV= 표준 편차/평균)는 염색체 14의 서열 태그 리드 커버리지가 이용될 때 최소이다. 하지만, 2, 3, 4, 5, 6, 7, 8개 또는 그 이상의 정규화 서열이 시험 표본 내에 관심되는 서열에 대한 서열 도스를 결정하는데 이용을 위해 확인될 수 있다. 실례로서, 임의의 한 가지 시험 표본 내에 염색체 21에 대한 두 번째 도스는 염색체 7, 염색체 9, 염색체 11 또는 염색체 12를 정규화 염색체 서열로서 이용하여 결정될 수 있는데, 그 이유는 이들 염색체 모두 염색체 14에 대한 CV에 가까운 CV를 갖기 때문이다.

[0195] 일부 구체예에서, 단일 염색체가 관심되는 염색체에 대한 정규화 염색체 서열로서 선택될 때, 정규화 염색체 서열은 시험된 모든 표본, 예를 들면, 유자격 표본 전체에 대하여 가장 작은 가변성을 갖는 관심되는 염색체에 대한 염색체 도스를 유발하는 염색체일 것이다. 일부 경우에, 최고 정규화 염색체는 최소 변동을 가질 수 없을 수도 있지만, 시험 표본 또는 표본들을 유자격 표본과 최고로 구별하는 유자격 도스의 분포를 가질 수 있다, 다시 말하면, 최고 정규화 염색체는 가장 낮은 변동을 가질 수 없을 수도 있지만, 최대 구별가능성을 가질 수 있다.

[0196] 일부 구체예에서, 정규화 서열은 하나 또는 그 이상의 로버스트 상염색체 서열 또는 이의 분절을 포함한다. 일부 구체예에서, 로버스트 상염색체는 관심되는 염색체(들)를 제외한 모든 상염색체를 포함한다. 일부 구체예에서, 로버스트 상염색체는 chr X, Y, 13, 18과 21을 제외한 모든 상염색체를 포함한다. 일부 구체예에서, 로버스트 상염색체는 정상적인 이배체 상태로부터 일탈하는 것으로 표본으로부터 결정된 것들을 제외한 모든 상염색체를 포함하는데, 이것은 정상적인 이배체 유전체에 비하여 비정상적인 사본수를 갖는 암 유전체를 결정하는데 유

용할 수 있다,

[0197] *시험 표본에서 이수성의 결정*

[0198] 유자격 표본 내에 정규화 서열(들)의 확인에 근거하여, 서열 도스가 하나 또는 그 이상의 관심되는 서열에서 다른 유전체로부터 유래된 핵산의 혼합물을 포함하는 시험 표본 내에 관심되는 서열에 대해 결정된다.

[0199] 단계 115에서, 시험 표본은 관심되는 서열의 임상적으로-유관한 CNV를 보유하는 것으로 의심되거나 또는 공지된 개체로부터 획득된다. 시험 표본은 생물학적 유체, 예를 들면, 혈장, 또는 아래에 설명된 바와 같은 임의의 적절한 표본일 수 있다. 설명된 바와 같이, 표본은 비침습성 시술, 예를 들면, 단순한 채혈을 이용하여 획득될 수 있다. 일부 구체예에서, 시험 표본은 핵산 분자, 예를 들면, cfDNA 분자의 혼합물을 내포한다. 일부 구체예에서, 시험 표본은 태아와 모계 cfDNA 분자의 혼합물을 내포하는 모계 혈장 표본이다.

[0200] 단계 125에서, 시험 표본 내에 시험 핵산 중에서 최소한 일부가 수백만 개의 서열 리드, 예를 들면, 36bp 리드를 산출하기 위해, 유자격 표본에 대해 설명된 바와 같이 염기서열결정된다. 단계 120에서처럼, 시험 표본 내에 핵산을 염기서열결정하는 것으로부터 산출된 리드는 참조 유전체에 독특하게 매핑되거나 또는 정렬되어 태그가 생산된다. 단계 120에서 설명된 바와 같이, 20 내지 40bp 리드를 포함하는 최소한 약 3×10^6 유자격 서열 태그, 최소한 약 5×10^6 유자격 서열 태그, 최소한 약 8×10^6 유자격 서열 태그, 최소한 약 10×10^6 유자격 서열 태그, 최소한 약 15×10^6 유자격 서열 태그, 최소한 약 20×10^6 유자격 서열 태그, 최소한 약 30×10^6 유자격 서열 태그, 최소한 약 40×10^6 유자격 서열 태그, 또는 최소한 약 50×10^6 유자격 서열 태그가 참조 유전체에 독특하게 매핑하는 리드로부터 획득된다. 일정한 구체예에서, 염기서열결정 기구에 의해 생산된 리드는 전자 형식으로 제공된다. 정렬은 아래에 논의된 바와 같이 연산 기구를 이용하여 달성된다. 개별 리드는 이들 리드가 참조 유전체와 독특하게 부합하는 부위를 확인하기 위해, 종종 방대한 (수백만 개의 염기쌍) 참조 유전체에 대하여 비교된다. 일부 구체예에서, 정렬 절차는 리드 및 참조 유전체 사이에 제한된 부정합을 허용한다. 일부 경우에, 리드 내에 1, 2, 또는 3개 염기쌍이 참조 유전체 내에 상응하는 염기쌍과 부정합하도록 허용되고, 그리고 그럼에도 불구하고, 매핑이 여전히 이루어진다.

[0201] 단계 135에서, 시험 표본 내에 핵산을 염기서열결정하는 것으로부터 획득된 모든 또는 대부분의 태그는 아래에 설명된 바와 같이 연산 기구를 이용하여 시험 서열 태그 리드 커버리지를 결정하기 위해 계수된다. 일부 구체예에서, 각 리드는 참조 유전체의 특정 영역 (많은 경우에 염색체 또는 분절)에 정렬되고, 그리고 리드는 부위 정보를 리드에 첨부함으로써 태그로 전환된다. 이러한 과정이 전개됨에 따라서, 연산 기구는 참조 유전체의 각 영역 (많은 경우에 염색체 또는 분절)에 매핑하는 태그/리드의 개수의 작동 수치를 유지할 수 있다. 수치는 관심되는 각 염색체 또는 분절 및 각 상응하는 정규화 염색체 또는 분절에 대해 저장된다.

[0202] 일정한 구체예에서, 참조 유전체는 진정한 생물학적 유전체의 일부이지만 참조 유전체 내에 포함되지 않는 하나 또는 그 이상의 배제된 영역을 갖는다. 이들 배제된 영역에 잠재적으로 정렬하는 리드는 계수되지 않는다. 배제된 영역의 실례는 긴 반복된 서열의 영역, X와 Y 염색체 사이에 유사성의 영역 등을 포함한다. 앞서 설명된 마스크 적용 기술에 의해 획득된 마스크가 적용된 참조 서열을 이용하여, 참조 서열의 마스크가 적용되지 않은 분절 상에서 태그만 CNV의 분석에 고려된다.

[0203] 일부 구체예에서, 상기 방법은 복수 리드가 참조 유전체 또는 서열 상에서 동일한 부위에 정렬할 때, 태그를 1회 이상 계수할 지를 결정한다. 2개의 태그가 동일한 서열을 갖고, 이런 이유로 참조 서열 상에서 동일한 부위에 정렬하는 경우가 있을 수 있다. 태그를 계수하는데 이용된 방법은 일정한 상황 하에, 동일한 염기서열결정된 표본으로부터 유래되는 동일한 태그를 수치로부터 배제할 수 있다. 불균형한 숫자의 태그가 소정의 표본에서 동일하면, 이것은 상기 절차에서 강한 바이어스 또는 다른 결함이 있다는 것을 암시한다. 이런 이유로, 일정한 구체예에 따라서, 계수 방법은 이전에 계수되었던 표본으로부터 태그와 동일한, 소정의 표본으로부터 태그를 계수하지 않는다.

[0204] 언제 단일 표본으로부터 동일한 태그를 무시할 지를 선택하기 위한 다양한 규준이 세팅될 수 있다. 일정한 구체예에서, 계수되는 태그의 규정된 백분율은 독특해야 한다. 이러한 역치보다 많은 태그가 독특하지 않으면, 이들은 무시된다. 가령, 규정된 백분율이 최소한 50%가 독특할 것을 요구하면, 동일한 태그는 독특한 태그의 백분율이 표본에 대해 50%를 초과할 때까지 계수되지 않는다. 다른 구체예에서, 독특한 태그의 역치 숫자는 최소한 약 60%이다. 다른 구체예에서, 독특한 태그의 역치 백분율은 최소한 약 75%, 또는 최소한 약 90%, 또는 최소한 약 95%, 또는 최소한 약 98%, 또는 최소한 약 99%이다. 역치는 염색체 21의 경우에 90%에서 세팅될 수 있다. 30M

태그가 염색체 21에 정렬되면, 이들 중에서 최소한 27M이 독특해야 한다. 3M 계수된 태그가 독특하지 않고, 그리고 30 백만 및 첫 번째 태그가 독특하지 않으면, 이것은 계수되지 않는다. 언제 추가의 동일한 태그를 계수하지 않을 지를 결정하는데 이용되는 특정 역치 또는 다른 규준의 선택은 적절한 통계학적 분석을 이용하여 선별될 수 있다. 이러한 역치 또는 다른 규준에 영향을 주는 한 가지 인자는 태그가 정렬될 수 있는 유전체의 크기에 대한 염기서열결정된 표본의 상대적 양이다. 다른 인자는 리드의 크기 및 유사한 고려 사항을 포함한다.

[0205] 한 구체예에서, 관심되는 서열에 매핑된 시험 서열 태그의 개수는 이들이 매핑되는 관심되는 서열의 공지된 길이에 정규화되어, 시험 서열 태그 밀도 비율이 제공된다. 유자격 표본에 대해 설명된 바와 같이, 관심되는 서열의 공지된 길이에 정규화가 필요하지 않고, 그리고 인간 해석을 위해 이를 단순화하기 위해 숫자에서 자리수를 감소시키는 단계로서 포함될 수 있다. 모든 매핑된 시험 서열 태그가 시험 표본에서 계수됨에 따라서, 관심되는 서열, 예를 들면, 임상적으로-유관한 서열에 대한 서열 태그 리드 커버리지가 시험 표본에서 결정되고, 유자격 표본에서 확인된 최소한 하나의 정규화 서열에 상응하는 추가 서열에 대한 서열 태그 리드 커버리지 역시 결정된다.

[0206] 단계 150에서, 유자격 표본 내에 최소한 하나의 정규화 서열의 동일성에 근거하여, 시험 서열 도스가 시험 표본 내에 관심되는 서열에 대해 결정된다. 다양한 구체예에서, 시험 서열 도스가 본원에서 설명된 바와 같이, 관심되는 서열 및 상응하는 정규화 서열의 서열 태그 리드 커버리지를 이용하여 연산적으로 결정된다. 이러한 착수를 책임지는 연산 기구는 관심되는 서열 및 이의 연관된 정규화 서열 사이에 연관에 전자적으로 접근할 것이고, 이것은 데이터베이스, 표, 그래프에 저장되거나, 또는 프로그램 명령에서 코드로서 포함될 수 있다.

[0207] 본원의 다른 곳에서 설명된 바와 같이, 최소한 하나의 정규화 서열은 단일 서열 또는 일군의 서열일 수 있다. 시험 표본 내에 관심되는 서열에 대한 서열 도스는 시험 표본 내에 관심되는 서열에 대해 결정된 서열 태그 리드 커버리지 및 시험 표본에서 결정된 최소한 하나의 정규화 서열의 서열 태그 리드 커버리지의 비율이고, 여기서 시험 표본 내에 정규화 서열은 관심되는 특정 서열에 대해 유자격 표본에서 확인된 정규화 서열에 상응한다. 가령, 유자격 표본 내에 염색체 21에 대해 확인된 정규화 서열이 염색체, 예를 들면, 염색체 14인 것으로 결정되면, 염색체 21 (관심되는 서열)에 대한 시험 서열 도스는 시험 표본에서 각각 결정된 염색체 21에 대한 서열 태그 리드 커버리지 및 염색체 14에 대한 서열 태그 리드 커버리지의 비율로서 결정된다. 유사하게, 염색체 이수성과 연관된 염색체 13, 18, X, Y, 그리고 다른 염색체에 대한 염색체 도스가 결정된다. 관심되는 염색체에 대한 정규화 서열은 하나 또는 일군의 염색체, 또는 하나 또는 일군의 염색체 분절일 수 있다. 앞서 설명된 바와 같이, 관심되는 서열은 염색체의 부분, 예를 들면, 염색체 분절일 수 있다. 따라서, 염색체 분절에 대한 도스가 시험 표본 내에 분절에 대해 결정된 서열 태그 리드 커버리지 및 시험 표본 내에 정규화 염색체 분절에 대한 서열 태그 리드 커버리지의 비율로서 결정될 수 있고, 여기서 시험 표본 내에 정규화 분절은 관심되는 특정 분절에 대해 유자격 표본에서 확인된 정규화 분절 (단일 또는 일군의 분절)에 상응한다. 염색체 분절은 크기에서 킬로베이스 (kb) 내지 메가베이스 (Mb) (가령, 약 1kb 내지 10 kb, 또는 약 10 kb 내지 100 kb, 또는 약 100kb 내지 1 Mb)의 범위에서 변할 수 있다.

[0208] 단계 155에서, 역치값은 복수의 유자격 표본에서 결정된 유자격 서열 도스 및 관심되는 서열에 대해 이수체인 것으로 알려진 표본에 대해 결정된 서열 도스에 대해 확립된 표준 편차 값으로부터 도출된다. 주목할 점은 이러한 작업이 전형적으로, 환자 시험 표본의 분석과 비동기적으로 수행된다는 것이다. 이것은 예로서, 유자격 표본으로부터 정규화 서열의 선별과 동시에 수행될 수 있다. 정확한 분류는 상이한 부류, 다시 말하면, 이수성의 유형에 대한 확률 분포 사이의 차이에 의존한다. 일부 실례에서, 역치는 각 유형의 이수성, 예를 들면, 삼염색체성 21에 대한 경험적 분포에서 선택된다. 가능한 역치값은 태아와 모계 핵산의 혼합물을 포함하는 모계 표본으로부터 추출된 cfDNA를 염기서열결정함으로써 염색체 이수성을 결정하기 위한 방법의 이용을 설명하는 실시예에서 설명된 바와 같이, 삼염색체성 13, 삼염색체성 18, 삼염색체성 21, 그리고 일염색체성 X 이수성을 분류하기 위해 확립되었다. 염색체의 이수성에 대해 영향을 받은 표본을 구별하기 위해 결정되는 역치값은 상이한 이수성에 대한 역치와 동일하거나 상이할 수 있다. 실시예에서 보여지는 바와 같이, 관심되는 각 염색체에 대한 역치값은 표본 및 염기서열결정 실행 전체에 대하여 관심되는 염색체의 도스에서 가변성으로부터 결정된다. 관심되는 임의의 염색체에 대한 염색체 도스가 덜 가변적일수록, 모든 영향을 받지 않은 표본 전체에 대하여 관심되는 염색체에 대한 도스에서 확산이 더욱 좁은데, 이들 표본은 상이한 이수성을 결정하기 위한 역치를 세팅하는데 이용된다.

[0209] 환자 시험 표본을 분류하는 것과 연관된 과정 흐름으로 돌아가서, 단계 160에서, 관심되는 서열의 사본수 변동이 관심되는 서열에 대한 시험 서열 도스를 유자격 서열 도스로부터 확립된 최소한 하나의 역치값에 비교함으로써 시험 표본에서 결정된다. 이러한 작업은 서열 태그 리드 커버리지를 계측하고 및/또는 분절 도스를 계산하는

데 이용된 동일한 연산 기구에 의해 수행될 수 있다.

[0210] 단계 160에서, 관심되는 시험 서열에 대한 계산된 도스는 표본을 "정상적임," "영향을 받음" 또는 "호출 없음"으로서 분류하는 사용자-규정된 "신뢰도의 역치"에 따라 선택되는 역치값으로서 세팅된 것과 비교된다. "호출 없음" 표본은 확정적인 진단이 신뢰도 있게 내려질 수 없는 표본이다. 각 유형의 영향을 받은 표본 (가령, 삼염색체성 21, 부분적인 삼염색체성 21, 일염색체성 X)은 자체 역치를 갖는데, 하나는 정상적인 (영향을 받지 않은) 표본을 호출하기 위한 것이고, 다른 하나는 영향을 받은 표본을 호출하기 위한 것이다 (비록 일부 경우에 이들 두 역치가 일치하긴 하지만). 본원의 다른 곳에서 설명된 바와 같이, 일부 상황 하에 호출 없음은 시험 표본 내에 핵산의 태아 분율이 충분히 높으면, 호출 (영향을 받음 또는 정상적임)로 전환될 수 있다. 시험 서열의 분류는 이러한 과정 흐름의 다른 작업에서 이용된 연산 기구에 의해 보고될 수 있다. 일부 경우에, 분류는 전자 형식으로 보고되고, 그리고 관심 있는 개인에게 표시되거나, 이메일로 보내지거나, 문자로 보내지거나, 기타 등등일 수 있다.

[0211] 일부 구체예에서, CNV의 결정은 앞서 설명된 바와 같이, 염색체 또는 분절 도스를 한 세트의 유자격 표본에서 상응하는 염색체 또는 분절 도스의 평균에 관련시키는 NCV 또는 NSV를 계산하는 것을 포함한다. 이후, CNV는 NCV/NSV를 미리 결정된 사본수 평가 역치값에 비교함으로써 결정될 수 있다.

[0212] 사본수 평가 역치는 가양성 및 가음성의 비율이 최적화되도록 선택될 수 있다. 사본수 평가 역치가 더욱 높을수록, 가양성이 발생할 개연성이 덜하다. 유사하게, 역치가 더욱 낮을수록, 가음성이 발생할 개연성이 덜하다. 따라서, 첫 번째 이상적인 역치 (그 초과에서 단지 진양성만 분류된다), 그리고 두 번째 이상적인 역치 (그 미만에서 단지 진음성만 분류된다) 사이에 교환이 존재한다.

[0213] 역치는 주로, 한 세트의 영향을 받지 않은 표본에서 결정된 바와 같이, 관심되는 특정 염색체에 대한 염색체 도스에서 가변성에 따라 세팅된다. 가변성은 표본 내에 존재하는 태아 cDNA의 분율을 비롯한 다수의 인자에 의존한다. 가변성 (CV)은 영향을 받지 않은 표본의 개체군 전체에 대하여 염색체 도스에 대한 평균 또는 중앙값 및 표준 편차에 의해 결정된다. 따라서, 이수성을 분류하기 위한 역치(들)는 아래에 따라, NCV를 이용한다:

$$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

[0214]

[0215] (여기서 $\hat{\mu}_j$ 및 $\hat{\sigma}_j$ 는 각각, 한 세트의 유자격 표본에서 j-번째 염색체 도스에 대한 추정된 평균과 표준 편차이고, 그리고 x_{ij} 는 시험 표본 i에 대한 관찰된 j-번째 염색체 도스이다.)

[0216] 연관된 태아 분율은 아래와 같다:

$$FF_{ij} = 2 \times \left| \frac{NCV_{ij} \times \hat{\sigma}_j}{\hat{\mu}_j} \right| = 2 \times NCV \times CV$$

[0217]

[0218] 따라서, 관심되는 염색체의 모든 NCV에 대해, 소정의 NCV 값과 연관된 예상된 태아 분율은 영향을 받지 않은 표본의 개체군 전체에 대하여 관심되는 염색체에 대한 염색체 비율의 평균 및 표준 편차에 근거하여 CV로부터 계산될 수 있다.

[0219] 차후에, 태아 분율 및 NCV 값 사이에 관계에 근거하여, 결정 경계가 선택될 수 있는데, 그 초과에서 표본은 정규 분포 사분위수에 근거하여 양성 (영향을 받음)인 것으로 결정된다. 앞서 설명된 바와 같이, 역치는 진양성의 검출 및 가음성 결과의 비율 사이에 최적 교환을 위해 세팅된다. 따라서, 세팅되는 역치는 가양성 및 가음성을 최적화하도록 선택된다.

[0220] 일정한 구체예는 태아와 모계 핵산 분자를 포함하는 생물학적 표본에서 태아 염색체 이수성의 출생전 진단을 제공하기 위한 방법을 제공한다. 진단은 생물학적 시험 표본, 예를 들면, 모계 혈장 표본으로부터 유래된 태아와 모계 핵산 분자의 혼합물의 최소한 일부로부터 서열 정보를 획득하고, 염기서열결정 데이터로부터 관심되는 하나 또는 그 이상의 염색체에 대한 정규화 염색체 도스, 및/또는 관심되는 하나 또는 그 이상의 분절에 대한 정규화 분절 도스를 연산하고, 그리고 각각, 시험 표본에서 관심되는 염색체에 대한 염색체 도스 및/또는 관심되

는 분절에 대한 분절 도스, 그리고 복수의 유자격 (정상적인) 표본에서 확립된 역치값 사이에 통계학적으로 유의한 차이를 결정하고, 그리고 통계학적 차이에 근거하여 출생전 진단을 제공하는 것에 근거하여 만들어진다. 상기 방법의 단계 160에서 설명된 바와 같이, 정상적임 또는 영향을 받음의 진단이 이루어진다. "호출 없음"은 정상적임 또는 영향을 받음에 대한 진단이 신뢰성 있게 이루어질 수 없는 경우에 제공된다.

[0221] 일부 구체예에서, 2가지 역치가 선택될 수 있다. 첫 번째 역치는 가양성 비율을 최소화하도록 선택되고, 그 초과에서 표본은 "영향을 받음"으로서 분류될 것이고, 그리고 두 번째 역치는 가음성 비율을 최소화하도록 선택되고, 그 미만에서 표본은 "영향을 받지 않음"으로서 분류될 것이다. 두 번째 역치 초과, 하지만 첫 번째 역치 미만의 NCV를 갖는 표본은 "이수성 의심됨" 또는 "호출 없음" 표본으로서 분류될 수 있는데, 이들에 대한 이수성의 존재 또는 부재는 독립된 수단에 의해 확증될 수 있다. 첫 번째와 두 번째 역치 사이에 영역은 "호출 없음" 영역으로서 지칭될 수 있다.

[0222] 일부 구체예에서, 의심되는 및 호출 없음 역치가 표 2에서 도시된다. 목격될 수 있는 바와 같이, NCV의 역치는 상이한 염색체 전체에 대하여 변한다. 일부 구체예에서, 역치는 앞서 설명된 바와 같이 표본에 대한 FF에 따라 변한다. 여기에서 적용된 역치 기술은 일부 구체예에서, 향상된 감수성과 선택성에 기여한다.

표 2

[0223] **의심되는 및 영향을 받은 NCV 역치 브래킷 호출 없음 범위**

	의심됨	영향을 받음
Chr 13	3.5	4.0
Chr 18	3.5	4.5
Chr 21	3.5	4.0
Chr X (XO, XXX)	4.0	4.0
Chr Y (XX 대 XY)	6.0	6.0

[0224] **서열 리드 커버리지 결정**

[0225] *서열 리드 커버리지를 결정하기 위한 일반적인 과정*

[0226] 개시된 일부 구체예는 낮은 잡음 및 높은 신호에서 서열 리드 커버리지 양을 결정하는 방법을 제공하고, 전통적인 방법에 의해 획득된 서열 리드 커버리지 양에 비하여 향상된 감수성, 선택성, 및/또는 효율로 사본수와 CNV에 관련된 다양한 유전적 장애를 결정하기 위한 데이터를 제공한다. 일정한 구체예에서, 시험 표본으로부터 서열이 서열 리드 커버리지 양을 획득하기 위해 처리된다.

[0227] 이러한 과정은 다른 공급원으로부터 가용한 일정한 정보를 이용한다. 일부 실행에서, 이러한 정보 모두 영향을 받지 않는 (가령, 이수체가 아닌) 것으로 공지된 표본의 훈련 세트로부터 획득된다. 다른 구체예에서, 정보의 일부 또는 전부가 다른 시험 표본으로부터 획득되고, 이것은 복수 표본이 동일한 과정에서 분석될 때 "온 더 플라이"로서 제공될 수 있다.

[0228] 일정한 구체예에서, 서열 마스크가 데이터 잡음을 감소시키는데 이용된다. 일부 구체예에서, 관심되는 서열 및 이의 정규화 서열 둘 모두 마스크가 적용된다. 일부 구체예에서, 관심되는 상이한 염색체 또는 분절이 고려될 때 상이한 마스크가 이용될 수 있다. 예로서, 염색체 13이 관심되는 염색체일 때, 하나의 마스크 (또는 일군의 마스크)가 이용될 수 있고, 그리고 염색체 21이 관심되는 염색체일 때, 상이한 마스크 (또는 일군의 마스크)가 이용될 수 있다. 일정한 구체예에서, 마스크는 빈의 분해능에서 규정된다. 이런 이유로, 한 가지 실례에서, 마스크 분해능은 100 kb이다. 일부 구체예에서, 상이한 마스크가 염색체 Y에 적용될 수 있다. 염색체 Y에 대한 마스크가 적용된 배제 영역은 2013년 6월 17일자 제출된 US 특허가출원 번호 61/836,057 [attorney docket no. ARTEP008P]에서 설명된 바와 같이, 관심되는 다른 염색체에서보다 더욱 미세한 분해능 (1kb)에서 제공될 수 있다. 마스크는 배제된 유전체 영역을 확인하는 파일의 형태에서 제공된다.

[0229] 일정한 구체예에서, 상기 과정은 정규화된 리드 커버리지의 예상 값을 활용하여 관심되는 서열의 프로필에서 빈마다 변동을 제거하는데, 상기 변동은 시험 표본에 대한 CNV를 결정하기에는 정보가 충분하지 않다. 상기 과정은 전체 유전체 전체에 대하여 각 빈에 대한, 또는 최소한, 참조 유전체에서 로버스트 염색체의 빈에 대한 정규화된 리드 커버리지의 예상 값에 따라 정규화된 리드 커버리지 양을 조정한다 (아래 작업 317에서 이용을 위해). 예상 값은 영향을 받지 않은 표본의 훈련 세트로부터 결정될 수 있다. 실례로서, 예상 값은 훈련 세트 표

본 전체에 대하여 중앙값일 수 있다. 표본의 예상된 리드 커버리지 값은 참조 유전체의 로버스트 염색체에서 모든 빈에 정렬된 독특한 비다중 태그의 총수에 의해 나눗셈된, 빈에 정렬된 독특한 비다중 태그의 개수로서 결정될 수 있다.

[0230] 도면 2는 관심되는 서열의 리드 커버리지를 결정하기 위한 과정 200의 흐름도를 묘사하는데, 이것은 블록 214에서 시험 표본 내에 관심되는 서열의 사본수를 평가하는데 이용된다. 이러한 과정은 영향을 받지 않은 훈련 표본 전체에 대하여 공통적인 체계적 변동을 제거하고, 상기 변동은 CNV 평가에 대한 분석에서 잡음을 증가시킨다. 이것은 또한, 시험 표본에 특정한 GC 바이어스를 제거하고, 따라서 데이터 분석에서 신호 대 잡음 비율을 증가시킨다.

[0231] 상기 과정은 블록 202에서 지시된 바와 같이 시험 표본의 서열 리드를 제공함으로써 시작된다. 일부 구체예에서, 서열 리드는 모체와 태아의 cfDNA를 포함하는 임신 여성의 혈액으로부터 획득된 DNA 분절을 염기서열결정함으로써 획득된다. 상기 과정은 진행하여 서열 리드를 관심되는 서열을 포함하는 참조 유전체에 맞춰 정렬하고, 시험 서열 태그를 제공한다. 블록 204. 참조 서열 상에서 각 빈에서 시험 서열 태그 수치는 빈의 리드 커버리지를 규정한다. 블록 206. 일부 구체예에서, 하나 이상의 부위에 정렬되는 리드는 배제된다. 일부 구체예에서, 동일한 부위에 정렬되는 복수 리드는 배제되거나 또는 단일 리드 수치로 감소된다. 일부 구체예에서, 배제된 부위에 정렬되는 리드 역시 배제된다. 이런 이유로, 일부 구체예에서, 비-배제된 부위에 정렬되는 독특하게 정렬된, 비다중 태그만 계수되고, 각 빈의 리드 커버리지를 결정하기 위한 비-배제된 부위 수치 (NES 수치)를 제공한다. 일부 구체예에서, 각 빈의 리드 커버리지는 동일한 표본 내에 정규화 서열의 리드 커버리지에 의해 나눗셈되고, 정규화된 리드 커버리지 양을 제공한다.

[0232] 과정 200은 이후, 관심되는 서열의 전역 프로필을 제공한다. 전역 프로필은 영향을 받지 않은 훈련 표본의 훈련 세트로부터 획득된 각 빈에서 예상된 리드 커버리지를 포함한다. 블록 208. 과정 200은 예상된 리드 커버리지에 따라 시험 서열 태그의 정규화된 리드 커버리지 양을 조정함으로써 훈련 표본에서 공통적인 변동을 제거하여 전역-프로필-교정된 리드 커버리지를 획득한다. 블록 210. 일부 구체예에서, 블록 208에서 제공된 훈련 세트로부터 획득된 예상된 리드 커버리지는 훈련 표본 전체에 대하여 중앙값이다. 일부 구체예에서, 작업 2010은 예상된 리드 커버리지를 정규화된 리드 커버리지로부터 감산함으로써 정규화된 리드 커버리지 양을 조정한다. 다른 구체예에서, 작업 2010은 정규화된 리드 커버리지 양을 각 빈의 예상된 리드 커버리지로 나눗셈하여 전역-프로필 교정된 리드 커버리지를 산출한다.

[0233] 게다가, 과정 200은 전역 프로필을 제거하기 위해 조정되었던 리드 커버리지 양을 더욱 조정함으로써 시험 표본에 특정한 GC 바이어스를 제거한다. 블록 212에서 보여지는 바와 같이, 상기 과정은 시험 표본 내에 현존하는 GC 함량 수준 및 전역-프로필-교정된 리드 커버리지 사이의 관계에 근거하여 전역-프로필-교정된 리드 커버리지를 조정하고, 따라서 표본-GC-교정된 리드 커버리지를 획득한다. 영향을 받지 않은 훈련 표본에서 공통적인 체계적 변동 및 개체내 GC 바이어스에 대해 조정한 후에, 상기 과정은 향상된 감수성과 특이성을 갖는 표본의 CNV를 평가하기 위한 리드 커버리지 양을 제공한다.

[0234] *서열 리드 커버리지를 결정하기 위한 예시적인 과정의 상세*

[0235] 도면 3a는 시험 표본으로부터 서열 데이터에서 잡음을 감소시키기 위한 과정 301의 실례를 제공한다. 도면 3b-3j는 상기 과정의 다양한 시기에서 데이터 분석을 제공한다. 도면 3a에서 보여지는 바와 같이, 묘사된 과정은 하나 또는 그 이상의 표본으로부터 cfDNA의 추출로 시작된다. 블록 303을 참조한다. 적합한 추출 과정과 기구는 본원의 다른 곳에서 설명된다. 일부 구체예에서, 2013년 3월 15일자 제출된 US 특허 출원 번호 61/801,126 (전체적으로 본원에 참조로서 편입됨)에서 설명된 과정은 cfDNA를 추출한다. 일부 실행에서, 상기 기구는 복수 표본으로부터 cfDNA를 함께 처리하여 다중화된 라이브러리 및 서열 데이터를 제공한다. 도면 3a에서 블록 305와 307을 참조한다. 일부 구체예에서, 상기 기구는 8개 또는 그 이상의 시험 표본으로부터 cfDNA를 병렬적으로 처리한다. 본원의 다른 곳에서 설명된 바와 같이, 염기서열결정 시스템은 추출된 cfDNA를 처리하여 코딩된 (가령, 바코드화된) cfDNA 단편의 라이브러리를 생산할 수 있다. cfDNA의 서열분석기 서열 라이브러리는 매우 다수의 서열 리드를 생산한다. 표본당 코딩은 다중화된 표본에서 리드의 역다중화를 허용한다. 8개 또는 그 이상의 표본 각각은 수십만 개 또는 수백만 개의 리드를 가질 수 있다. 상기 과정은 도면 3a에서 추가 작업에 앞서 리드를 필터링할 수 있다. 일부 구체예에서, 리드 필터링은 잘못된 및 낮은 품질 리드를 걸러 내기 위해 서열분석기에서 실행되는 소프트웨어 프로그램에 의해 가능한 품질-필터링 과정이다. 가령, Illumina의 염기서열결정 대조 소프트웨어 (SCS) 및 서열과 변동의 공통 사정 소프트웨어 프로그램은 염기서열결정 반응에 의해 산출된 미가공 이미지 데이터를 강도 점수, 염기 호출, 품질평가점수 정렬, 그리고 하류 분석을 위한 생물학적으로 유관한 정

보를 제공하는 추가 형식으로 전환함으로써 잘못된 및 낮은 품질 리드를 걸러 낸다.

[0236] 서열분석기 또는 다른 기구가 표본에 대한 리드를 산출한 후에, 시스템의 요소는 리드를 참조 유전체에 연산적으로 정렬한다. 블록 309를 참조한다. 정렬은 본원의 다른 곳에서 설명된다. 정렬은 태그를 산출하고, 이들은 참조 유전체 상에서 독특한 위치를 명시하는 위치 정보가 주해된 리드 서열을 내포한다. 일정한 실행에서, 시스템은 중복 리드 - 동일한 서열을 갖는 2개 또는 그 이상의 리드 -를 고려하지 않고 첫 번째 통과 정렬을 수행하고, 그리고 차후에, 중복 리드를 제거하거나 또는 중복 리드를 단일 리드로서 계수하여 비-중복 서열 태그를 생산한다. 다른 실행에서, 시스템은 중복 리드를 제거하지 않는다. 일부 구체예에서, 상기 과정은 고려 사항으로부터, 유전체 상에서 복수 위치에 정렬되는 리드를 제거하여 독특하게 정렬된 태그를 생산한다. 일부 구체예에서, 비-배제된 부위 (NESs)에 매핑된 독특하게 정렬된, 비다중 서열 태그는 비-배제된 부위 수치 (NES 수치)를 산출하기 위해 파악되는데, 이들은 리드 커버리지를 추정하는 데이터를 제공한다.

[0237] 다른 곳에서 설명된 바와 같이, 배제된 부위는 서열 태그를 계수하는 목적으로 배제되었던 참조 유전체의 영역에서 발견되는 부위이다. 일부 구체예에서, 배제된 부위는 반복 서열을 내포하는 염색체의 영역, 예를 들면, 동원체와 말단소립, 그리고 하나 이상의 염색체에 공통적인 염색체의 영역, 예를 들면, Y 염색체 상에 존재하고, 또한 X 염색체 상에 존재하는 영역에서 발견된다. 비-배제된 부위 (NESs)는 서열 태그를 계수하는 목적으로 참조 유전체에서 배제되지 않는 부위이다.

[0238] 그 다음, 시스템은 정렬된 태그를 참조 유전체 상에서 빈 내로 분할한다. 블록 311을 참조한다. 빈은 참조 유전체의 길이에 따라서 이격된다. 일부 구체예에서, 전체 참조 유전체는 인접한 빈으로 분할되는데, 이들은 규정된 동등한 크기 (가령, 100 kb)를 가질 수 있다. 대안으로, 빈은 아마도 표본당 기초에서 역동적으로 결정된 길이를 가질 수 있다. 염기서열결정 깊이는 최적 빈 크기 선별에 충격을 준다. 역동적으로 크기산정된 빈은 그들의 크기가 라이브러리 크기에 의해 결정될 수 있다. 가령, 빈 크기는 평균적으로, 1000개 태그를 수용하는데 필요한 서열 길이가 되도록 결정될 수 있다.

[0239] 각 빈은 고려 중인 표본으로부터 다수의 태그를 갖는다. 정렬된 서열의 "리드 커버리지"를 반영하는 태그의 이러한 숫자는 표본 데이터를 필터링하고, 만약 그렇지 않으면 세정하여 표본 내에 사본수 변동을 확실하게 결정하기 위한 출발점으로서 역할을 한다. 도면 3a는 블록 313 내지 321에서 세정 작업을 보여준다.

[0240] 도면 3a에서 묘사된 구체예에서, 상기 과정은 마스크를 참조 유전체의 빈에 적용한다. 블록 313을 참조한다. 시스템은 다음 과정 작업의 일부 또는 전부에서 마스크가 적용된 빈에서 리드 커버리지를 고려 사항으로부터 배제할 수 있다. 많은 경우에, 마스크가 적용된 빈으로부터 리드 커버리지 값은 도면 3a에서 나머지 작업 중에서 어느 것에도 고려되지 않는다.

[0241] 다양한 실행에서, 하나 또는 그 이상의 마스크가 표본간 높은 가변성을 표시하는 것으로 밝혀진 유전체의 영역에 대한 빈을 제거하기 위해 적용된다. 이런 마스크는 관심되는 염색체 (가령, chr13, 18, 그리고 21) 및 다른 염색체 둘 모두에 대해 제공된다. 다른 곳에서 설명된 바와 같이, 관심되는 염색체는 사본수 변동 또는 다른 이상을 잠재적으로 품는 것으로 고려 중인 염색체이다.

[0242] 일부 실행에서, 마스크는 다음의 접근법을 이용하여 유자격 표본의 훈련 세트로부터 확인된다. 초기에, 각 훈련 세트 표본은 도면 3a에서 작업 315 내지 319에 따라 처리되고 필터링된다. 정규화되고 교정된 리드 커버리지 양은 이후, 각 빈에 대해 알려지고, 그리고 통계, 예를 들면, 표준 편차, 중위 절대 편차, 및/또는 변동 계수가 각 빈에 대해 계산된다. 다양한 필터 조합이 관심되는 각 염색체에 대해 평가될 수 있다. 필터 조합은 관심되는 염색체의 빈에 대한 한 가지 필터 및 모든 다른 염색체의 빈에 대한 상이한 필터를 제공한다.

[0243] 일부 실행에서, 정규화 염색체 (또는 일군의 염색체)의 선택은 마스크를 획득한 후에 재고된다 (가령, 앞서 설명된 바와 같이 관심되는 염색체에 대한 컷오프를 선택함으로써). 서열 마스크를 적용한 후에, 정규화 염색체 또는 염색체들을 선택하는 과정은 본원의 다른 곳에서 설명된 바와 같이 수행될 수 있다. 가령, 염색체의 모든 가능한 조합이 정규화 염색체로서 평가되고, 그리고 영향을 받은 표본 및 영향을 받지 않은 표본을 구별하는 그들의 능력에 따라 순위매겨진다. 이러한 과정은 상이한 최적 정규화 염색체 또는 일군의 염색체를 찾을 수 있다 (또는 찾지 못할 수도 있다). 다른 구체예에서, 정규화 염색체는 모든 유자격 표본 전체에 대하여 관심되는 서열에 대한 서열 도스에서 가장 작은 가변성을 유발하는 것들이다. 상이한 정규화 염색체 또는 일군의 염색체가 확인되면, 상기 과정은 빈의 상기 설명된 확인을 필터에 임의선택적으로 이행한다. 아마도 새로운 정규화 염색체(들)는 상이한 컷오프를 유발한다.

[0244] 일정한 구체예에서, 염색체 Y에 대해 상이한 마스크가 적용된다. 적합한 염색체 Y 마스크의 실례는 2013년 6월

17일자 제출된 US 특허가출원 번호 61/836,057 [attorney docket no. ARTEP008P]에서 설명되고, 이들은 모든 점에서 본원에 참조로서 편입된다.

[0245] 시스템은 빈에 연산적으로 마스크를 적용한 후에, 마스크에 의해 배제되지 않는 빈에서 리드 커버리지 값을 연산적으로 정규화한다. 블록 315를 참조한다. 일정한 구체예에서, 시스템은 참조 유전체 또는 이의 부분에서 리드 커버리지 (가령, 참조 유전체의 로버스트 염색체에서 리드 커버리지) 중에서 대부분 또는 전부에 대하여, 각 빈에서 시험 표본 리드 커버리지 값 (가령, 빈마다 NES 수치)을 정규화한다. 일부 경우에, 시스템은 고려 중인 빈에 대한 수치를 참조 유전체 내에 모든 로버스트 염색체에 정렬하는 모든 비-배제된 부위의 총수로 나눗셈함으로써 시험 표본 리드 커버리지 값 (빈마다)을 정규화한다. 일부 구체예에서, 시스템은 선형 회귀를 수행함으로써 시험 표본 리드 커버리지 값 (빈마다)을 정규화한다. 가령, 시스템은 먼저 y_a = 인터셉트 + 기울기 * gwp_a로서 로버스트 염색체에서 빈의 부분집합에 대한 리드 커버리지를 계산하고, 여기서 y_a는 빈 a에 대한 리드 커버리지이고, 그리고 gwp_a는 동일한 빈에 대한 전역 프로필이다. 시스템은 이후, z_b = y_b / (인터셉트 + 기울기 * gwp_b) - 1로서 정규화된 리드 커버리지 z_b를 계산한다.

[0246] 앞서 설명된 바와 같이, 로버스트 염색체는 이수체일 개연성이 낮은 염색체이다. 일정한 구체예에서, 로버스트 염색체는 염색체 13, 18과 21 이외에 모든 상염색체이다. 일부 구체예에서, 로버스트 염색체는 정상적인 이배체 유전체로부터 일탈하는 것으로 결정된 염색체 이외에 모든 상염색체이다.

[0247] bin의 변환된 수치 값 또는 리드 커버리지는 추가 처리를 위한 "정규화된 리드 커버리지 양"으로서 지칭된다. 정규화는 각 표본에 독특한 정보를 이용하여 수행된다. 전형적으로, 훈련 세트로부터 정보 없음이 이용된다. 정규화는 상이한 라이브러리 크기 (및 결과적으로 상이한 숫자의 리드와 태그)를 갖는 표본으로부터 리드 커버리지 양이 동등한 기초에서 처리되도록 허용한다. 차후 과정 작업 중에서 일부는 고려 중인 시험 표본에 대해 이용된 라이브러리보다 크거나 또는 작은 라이브러리로부터 염기서열결정될 수 있는 훈련 표본으로부터 유래된 리드 커버리지 양을 이용한다. 전체 참조 유전체 (또는 최소한 로버스트 염색체)에 정렬된 리드의 개수에 근거된 정규화 없이, 훈련 세트로부터 유래된 파라미터를 이용한 처리는 일부 실행에서 신뢰할 만한 하거나 또는 일반화가능하지 않을지도 모른다.

[0248] 도면 3b는 많은 표본에 대해 염색체 21, 13과 18 전체에 대하여 리드 커버리지를 도해한다. 표본 중에서 일부는 서로 다르게 처리되었다. 결과로서, 임의의 소정의 유전체 위치에서 넓은 표본간 변동을 목격할 수 있다. 정규화는 표본간 변동 중에서 일부를 제거한다. 도면 3c의 왼쪽 패널은 전체 유전체 전체에 대하여 정규화된 리드 커버리지 양을 묘사한다.

[0249] 도면 3a의 구체예에서, 시스템은 작업 315에서 산출된 정규화된 리드 커버리지 양으로부터 "전역 프로필"을 제거하거나 또는 감소시킨다. 블록 317을 참조한다. 이러한 작업은 유전체의 구조, 라이브러리 산출 과정, 그리고 염기서열결정 과정으로부터 발생하는 정규화된 리드 커버리지 양에서 체계적인 바이어스를 제거한다. 이에 더하여, 이러한 작업은 임의의 소정의 표본에서 예상된 프로필로부터 임의의 체계적인 선형 편차를 교정하도록 설계된다.

[0250] 일부 실행에서, 전역 프로필 제거는 각 빈의 정규화된 리드 커버리지 양을 각 빈의 상응하는 기댓값으로 나눗셈하는 것을 수반한다. 다른 구체예에서, 전역 프로필 제거는 각 빈의 기댓값을 각 빈의 정규화된 리드 커버리지 양으로부터 감산하는 것을 수반한다. 기댓값은 영향을 받지 않은 표본 (또는 X 염색체에 대한 영향을 받지 않은 암컷 표본)의 훈련 세트로부터 획득될 수 있다. 영향을 받지 않은 표본은 관심되는 염색체에 대한 이수성을 갖지 않는 것으로 알려진 개체로부터 표본이다. 일부 실행에서, 전역 프로필 제거는 각 빈의 기댓값 (훈련 세트로부터 획득됨)을 각 빈의 정규화된 리드 커버리지 양으로부터 감산하는 것을 수반한다. 일부 구체예에서, 상기 과정은 훈련 세트를 이용하여 결정된 바와 같이 각 빈에 대한 정규화된 리드 커버리지 양의 중앙값을 이용한다. 다시 말하면, 중앙값은 기대값이다.

[0251] 일부 구체예에서, 전역 프로필 제거는 전역 프로필에 대한 표본 리드 커버리지의 의존에 대한 선형 교정을 이용하여 실행된다. 지시된 바와 같이, 전역 프로필은 훈련 세트로부터 결정된 바와 같은 각 빈에 대한 기대값 (가령, 각 빈에 대한 중앙값)이다. 이들 구체예는 시험 표본의 정규화된 리드 커버리지 양을 각 빈에 대해 획득된 전역 중앙 프로필에 대하여 적합시킴으로써 획득된 로버스트 선형 모형을 이용할 수 있다. 일부 구체예에서, 선형 모형은 전역 중앙값 (또는 다른 예상 값) 프로필에 대하여 표본의 관찰된 정규화된 리드 커버리지 양을 회귀시킴으로써 획득된다.

[0252] 선형 모형은 표본 리드 커버리지 양이 전역 프로필 값과 선형 관계를 갖는다는 가정에 근거되고, 상기 선형 관계는 로버스트 염색체/영역 및 관심되는 서열 둘 모두에 대해 유효해야 한다. 도면 3d를 참조한다. 이런 경우에, 전역 프로필의 예상된 리드 커버리지 양에서 표본 번호 리드 커버리지 양의 회귀는 기울기 및 인터셉트를 갖는 라인을 산출할 것이다. 일정한 구체예에서, 이런 라인의 기울기 및 인터셉트는 빈에 대한 전역 프로필 값으로부터 "예측된" 리드 커버리지 양을 계산하는데 이용된다. 일부 실행에서, 전역 프로필 교정은 각 빈의 정규화된 리드 커버리지 양을 빈에 대한 예측된 리드 커버리지 양으로 모형화하는 것을 수반한다. 일부 실행에서, 시험 서열 태그의 리드 커버리지는 (i) 하나 또는 그 이상의 로버스트 염색체 또는 영역 내에 복수의 빈에서 예상된 리드 커버리지와 대비하여 시험 서열 태그의 리드 커버리지 사이에 수학적 관계를 획득하고, 그리고 (ii) 수학적 관계를 관심되는 서열에서 빈에 적용함으로써 조정된다. 일부 실행에서, 시험 표본 내에 리드 커버리지는 영향을 받지 않은 훈련 표본으로부터 예상된 리드 커버리지 값 및 유전체의 로버스트 염색체 또는 다른 로버스트 영역에서 시험 표본에 대한 리드 커버리지 값 사이에 선형 관계를 이용하여 변동에 대해 교정된다. 조정은 전역-프로필-교정된 리드 커버리지를 유발한다. 일부 경우에, 조정은 아래와 같이, 로버스트 염색체 또는 영역 내에 빈의 부분집합에 대한 시험 표본에 대한 리드 커버리지를 획득하는 것을 수반한다:

[0253] $y_a = 인터셉트 + 기울기 * gwp_a$

[0254] 여기서 ya는 하나 또는 그 이상의 로버스트 염색체 또는 영역 내에 시험 표본을 위한 빈 a의 리드 커버리지이고, 그리고 gwpa는 영향을 받지 않은 훈련 표본을 위한 빈 a의 전역 프로필이다. 이후, 상기 과정은 아래와 같이, 서열 또는 관심 영역에 대한 전역-프로필-교정된 리드 커버리지 zb를 연산한다:

[0255] $z_b = y_b / (인터셉트 + 기울기 * gwp_b) - 1$

[0256] 여기서 yb는 관심되는 서열 (이것은 로버스트 염색체 또는 영역 외부에 체류할 수 있다)에서 시험 표본을 위한 빈 b의 관찰된 리드 커버리지이고, 그리고 gwpb는 영향을 받지 않은 훈련 표본을 위한 빈 b에 대한 전역 프로필이다. 분모 (인터셉트 + 기울기 * gwpb)는 유전체의 로버스트 영역으로부터 추정된 관계에 근거하여, 영향을 받지 않은 시험 표본에서 관찰될 것으로 예측되는 빈 b에 대한 리드 커버리지이다. 사본수 변동을 품는 관심되는 서열의 경우에, 빈 b에 대한 관찰된 리드 커버리지 및 따라서, 전역-프로필-교정된 리드 커버리지 값은 영향을 받지 않은 표본의 리드 커버리지로부터 유의미하게 일탈할 것이다. 가령, 교정된 리드 커버리지 zb는 영향을 받은 염색체에서 빈에 대한 삼염색체성 표본의 경우에 태아 분율에 비례할 것이다. 이러한 과정은 로버스트 염색체에서 인터셉트와 기울기를 연산함으로써 표본 내에서 정규화하고, 그리고 이후, 관심되는 유전체 영역이 동일한 표본 내에서 로버스트 염색체에 대해 유지하는 관계 (기울기 및 인터셉트에 의해 설명된 바와 같이)로부터 어떻게 일탈하는 지를 평가한다.

[0257] 기울기 및 인터셉트는 도면 3d에서 보여지는 바와 같은 라인으로부터 획득된다. 전역 프로필 제거의 실례는 도면 3c에서 묘사된다. 왼쪽 패널은 많은 표본 전체에 대하여 정규화된 리드 커버리지 양에서 높은 빈마다 변동을 보여준다. 오른쪽 패널은 앞서 설명된 바와 같이, 전역 프로필 제거 후 동일한 정규화된 리드 커버리지 양을 보여준다.

[0258] 시스템은 블록 317에서 전역 프로필 변동을 제거하거나 또는 감소시킨 후에, 표본내 GC (구아닌-시토신) 함량 변동을 교정한다. 블록 319를 참조한다. 모든 빈은 GC로부터 자체 분율 기여를 갖는다. 분율은 빈에서 G와 C 뉴클레오티드의 숫자를 빈에서 뉴클레오티드의 총수 (가령, 100,000)로 나눗셈함으로써 결정된다. 일부 빈은 다른 것들보다 더욱 큰 GC 분율을 가질 것이다. 도면 3e와 3f에서 보여지는 바와 같이, 상이한 표본은 상이한 GC 바이어스를 표시한다. 이들 차이 및 이들의 교정은 아래에 더욱 설명될 것이다. 도면 3e-g는 전역 프로필 교정된, 정규화된 리드 커버리지 양 (빈마다)을 GC 분율 (빈마다)의 함수로서 보여준다. 놀랍게도, 상이한 표본은 상이한 GC 의존을 표시한다. 일부 표본은 단조적으로 감소시키는 의존 (도면 3e에서처럼)을 보여주고, 반면 다른 것들은 콤마 모양 의존 (도면 3f와 3g에서처럼)을 표시한다. 이들 프로필이 각 표본에 대해 독특할 수 있기 때문에, 이러한 단계에서 설명된 교정은 각 표본에 대해 별도로 및 독특하게 수행된다.

[0259] 일부 구체예에서, 시스템은 도면 3e-g에서 예시된 바와 같이 GC 분율의 기초에서 빈을 연산적으로 배열한다. 이것은 이후, 유사한 GC 함량을 갖는 다른 빈으로부터 정보를 이용하여 빈의 전역 프로필 교정된, 정규화된 리드 커버리지 양을 교정한다. 이러한 교정은 마스크가 적용되지 않은 빈 각각에 적용된다.

[0260] 일부 과정에서, 각 빈은 다음의 방식으로 GC 함량에 대해 교정된다. 시스템은 고려 중인 빈의 것들과 유사한 GC 분율을 갖는 빈을 연산적으로 선별하고, 그리고 이후, 선별된 빈에서 정보로부터 교정 파라미터를 결정한다. 일

부 구체예에서, 유사한 GC 분율을 갖는 빈은 유사성의 임의적으로 규정된 컷오프 값을 이용하여 선별된다. 한 가지 실례에서, 모든 빈 중에서 2%가 선별된다. 이들 빈은 고려 중인 빈과 가장 유사한 GC 함량을 갖는 2% 빈이다. 가령, 약간 더 많은 GC 함량을 갖는 1%의 빈 및 약간 더 적은 GC 함량을 갖는 1%의 빈이 선별된다.

[0261] 선별된 빈을 이용하여, 시스템은 교정 파라미터를 연산적으로 결정한다. 한 가지 실례에서, 교정 파라미터는 선별된 빈에서 정규화된 리드 커버리지 양 (전역 프로필 제거 후)의 대표값 이다. 이런 대표값의 실시예는 선별된 빈에서 정규화된 리드 커버리지 양의 중앙값 또는 평균을 포함한다. 시스템은 고려 중인 빈에 대한 계산된 교정 파라미터를 고려 중인 빈에 대한 정규화된 리드 커버리지 양 (전역 프로필 제거 후)에 적용한다. 일부 실행에서, 대표값 (가령, 중앙값)은 고려 중인 빈의 정규화된 리드 커버리지 양으로부터 감산된다. 일부 구체예에서, 정규화된 리드 커버리지 양의 중앙값 (또는 다른 대푯값)은 로버스트 상염색체 (염색체 13, 18과 21 이외에 모든 상염색체)에 대한 리드 커버리지 양만을 이용하여 선별된다.

[0262] 예로서, 100kb 빈을 이용한 한 가지 실례에서, 각 빈은 GC 분율의 독특한 값을 가질 것이고, 그리고 이들 빈은 그들의 GC 분율 함량에 근거하여 군으로 분할된다. 가령, 이들 빈은 50가지 군으로 분할되고, 여기서 군 경계는 %GC 분포의 (0, 2, 4, 6, …, 그리고 100) 사분위수에 상응한다. 중앙 정규화된 리드 커버리지 양은 동일한 GC 군 (표본 내에)에 매핑하는 로버스트 상염색체로부터 빈의 각 군에 대해 계산되고, 그리고 이후, 중앙값이 정규화된 리드 커버리지 양 (동일한 GC 군에서 전체 유전체 전체에 대하여 모든 빈에 대한)으로부터 감산된다. 이것은 임의의 소정의 표본 내에 로버스트 염색체로부터 추정된 GC 교정을 동일한 표본 내에 잠재적으로 영향을 받는 염색체에 적용한다. 가령, 0.338660 및 0.344720 사이에 GC 함량을 갖는 로버스트 염색체 상에서 모든 빈은 함께 그룹화되고, 중앙값이 이러한 군에 대해 계산되고 이러한 GC 범위 내에 빈의 정규화된 리드 커버리지로부터 감산되는데, 이들 빈은 유전체 상에서 어디든지 발견될 수 있다 (염색체 13, 18, 21, 그리고 X를 제외). 일정한 구체예에서, 염색체 Y는 이러한 GC 교정 과정으로부터 배제된다.

[0263] 도면 3g는 바로 앞서 설명된 바와 같이, 중앙값 정규화된 리드 커버리지 양을 교정 파라미터로서 이용한 GC 교정의 적용을 보여준다. 왼쪽 패널은 GC 분율 프로필과 대비하여 교정되지 않은 리드 커버리지 양을 보여준다. 보여지는 바와 같이, 프로필은 비선형 모양을 갖는다. 오른쪽 패널은 교정된 리드 커버리지 양을 보여준다. 도면 3h는 GC 분율 교정 전 (왼쪽 패널) 및 GC 분율 교정 후 (오른쪽 패널), 많은 표본에 대한 정규화된 리드 커버리지를 보여준다. 도면 3i는 GC 분율 교정 전 (적색) 및 GC 분율 교정 후 (녹색), 많은 시험 표본에 대한 정규화된 리드 커버리지의 변동 계수 (CV)를 보여주는데, 여기서 GC 교정은 정규화된 리드 커버리지에서 훨씬 작은 변동을 야기한다.

[0264] 상기 과정은 GC 교정의 상대적으로 단순한 실행이다. GC 바이어스를 교정하는 대안적 접근법은 스플라인 또는 다른 비선형 적합 기술을 이용하는데, 이것은 연속적 GC 공간에서 적용될 수 있고 GC 함량에 의해 리드 커버리지 양을 빈닝하는 것을 수반하지 않는다. 적합한 기술의 실례는 연속적 뢰스 교정 및 부드러운 스플라인 교정을 포함한다. 적합 함수는 고려 중인 표본에 대한 GC 함량과 대비하여 빈별로 정규화된 리드 커버리지 양으로부터 파생될 수 있다. 각 빈에 대한 교정은 고려 중인 빈에 대한 GC 함량을 적합 함수에 적용함으로써 계산된다. 가령, 정규화된 리드 커버리지 양은 고려 중인 빈의 GC 함량에서 스플라인의 예상된 리드 커버리지 값을 감산함으로써 조정될 수 있다. 대안으로, 조정은 스플라인 적합에 따라 예상된 리드 커버리지 값의 나눗셈에 의해 달성될 수 있다.

[0265] 작업 319에서 GC-의존을 교정한 후에, 시스템은 고려 중인 표본에서 이상점 빈을 연산적으로 제거한다 - 블록 321를 참조한다. 이러한 작업은 단일 표본 필터링 또는 손질로서 지칭될 수 있다. 도면 3j는 GC 교정 이후에도, 리드 커버리지가 여전히, 작은 영역 내에서 표본-특이적 변동을 갖는다는 것을 보여준다. 가령, 기댓값으로부터 예상치 않게 높은 편차가 발생하는 염색체 12 상에서 위치 1.1 e8에서 리드 커버리지를 참조한다. 이러한 편차는 물질 유전체에서 작은 사본수 변동으로부터 발생하는 것이 가능하다. 대안으로, 이것은 사본수 변동에 관련 없는 염기서열결정에서 기술적인 이유에 기인할 수 있다. 전형적으로, 이러한 작업은 단지 로버스트 염색체에만 적용된다.

[0266] 한 가지 실례로서, 시스템은 필터링에 대해 고려 중인 빈을 품는 염색체에서 모든 빈 전체에 대하여 GC 교정된 정규화된 리드 커버리지 양의 중앙값으로부터 3 중위 절대 편차보다 많은 GC 교정된 정규화된 리드 커버리지 양을 갖는 임의의 빈을 연산적으로 필터링한다. 한 가지 실례에서, 컷오프 값은 표준 편차와 일치하도록 조정된 3 중위 절대 편차로서 규정되고, 따라서 실질적으로 컷오프는 중앙값으로부터 1.4826* 중위 절대 편차이다. 일정한 구체예에서, 이러한 작업은 로버스트 염색체 및 이수성인 것으로 의심되는 염색체 둘 모두를 비롯하여, 표본 내에 모든 염색체에 적용된다.

[0267] 일정한 실행에서, 품질 관리로서 특징화될 수 있는 추가 작업이 수행된다. 블록 323을 참조한다. 일부 구체예에서, 품질 관리 계량은 임의의 잠재적 분모 염색체, 다시 말하면, "정규화 염색체" 또는 "로버스트 염색체"가 이수체인지 또는 만약 그렇지 않으면, 시험 표본이 관심되는 서열에서 사본수 변동을 갖는 지를 결정하는데 부적절한 지를 검출하는 것을 수반한다. 상기 과정이 로버스트 염색체가 부적절하다고 결정할 때, 상기 과정은 시험 표본을 무시하고 호출 없음을 설정할 수 있다. 대안으로, 이러한 QC 계량의 실패는 호출을 위한 정규화 염색체의 대체 세트의 이용을 촉발할 수 있다. 한 가지 실례에서, 품질 관리 방법은 로버스트 염색체에 대한 실제 정규화된 리드 커버리지 값을 로버스트 상염색체에 대한 예상 값에 대하여 비교한다. 예상 값은 다변량 정상적인 모형을 영향을 받지 않은 훈련 표본의 정규화된 프로필에 적합시키고, 데이터의 가능성 또는 베이지안 규준 (가령, 모형은 아카이케 정보 규준 또는 아마도 베이지안 정보 규준을 이용하여 선별된다)에 따라 최적 방식 구조를 선별하고, 그리고 QC에서 이용을 위한 최적 방식을 고정함으로써 획득될 수 있다. 로버스트 염색체의 정상적인 모형은 예로서, 정상적인 표본에서 염색체 리드 커버리지에 대한 평균 및 표준 편차를 갖는 확률 함수를 확인하는 군집화 기술을 이용함으로써 획득될 수 있다. 당연히, 다른 모형 형태가 이용될 수 있다. 상기 과정은 고정된 모형 파라미터를 고려하여, 임의의 인입 시험 표본에서 관찰된 정규화된 리드 커버리지의 가능성을 평가한다. 각 인입 시험 표본을 상기 모형으로 채점하여 가능성을 획득하고, 따라서 정상적인 표본 세트에 비하여 이상점을 확인함으로써 이것이 가능할 수 있다. 훈련 표본의 것으로부터 시험 표본의 가능성에서 편차는 부정확한 표본 분류를 유발할 수 있는 정규화 염색체 또는 표본 취급에서 이상 / 검정 처리 인공물을 암시할 수 있다. 이러한 QC 계량은 이들 표본 인공물 중에서 어느 한쪽과 연관된 분류에서 오차를 감소시키는데 이용될 수 있다. 도면 3k, 오른쪽 패널은 x 축 상에 염색체 번호를 보여주고, 그리고 y 축은 앞서 설명된 바와 같이 획득된 QC 모형과의 비교에 근거하여 정규화된 염색체 리드 커버리지를 보여준다. 이들 그래프는 염색체 2에 대한 과도한 리드 커버리지를 갖는 한 가지 표본 및 염색체 20에 대한 과도한 리드 커버리지를 갖는 다른 표본을 보여준다. 이들 표본은 여기에서 설명된 QC 계량을 이용하여 제거되거나 또는 정규화 염색체의 대체 세트를 이용하도록 전환될 것이다. 도면 3k의 왼쪽 패널은 염색체에 대한 가능성과 대비하여 NCV를 보여준다.

[0268] 도면 3a에서 묘사된 서열이 유전체 내에서 모든 염색체의 모든 빈에 대해 이용될 수 있다. 일정한 구체예에서, 상이한 과정이 염색체 Y에 적용된다. 염색체 또는 분절 도스, NCV, 및/또는 NSV를 계산하기 위해, 도스, NCV, 및/또는 NSV에 대한 발현에서 이용된 염색체 또는 분절 내에 빈으로부터 교정된 정규화된 리드 커버리지 양 (도면 3a에서 결정된 바와 같이)이 이용된다. 블록 325를 참조한다. 일정한 구체예에서, 평균 정규화된 리드 커버리지 양이 관심되는 염색체, 정규화 염색체, 관심되는 분절, 및/또는 정규화 분절 내에 모든 빈으로부터 계산되고, 본원의 다른 곳에서 설명된 바와 같이 서열 도스, NCV, 및/또는 NSV를 계산하는데 이용된다.

[0269] 일정한 구체예에서, 염색체 Y는 상이하게 처리된다. 이것은 Y 염색체에 독특한 한 세트의 빈에 마스크를 적용함으로써 필터링될 수 있다. 일부 구체예에서, Y 염색체 필터는 앞서 참조로서 편입된 US 특허가출원 번호 61/836,057에서 과정에 따라 결정된다. 일부 구체예에서, 필터는 다른 염색체의 필터에서 것들보다 작은 빈에 마스크를 적용한다. 가령, Y 염색체 마스크는 1 kb 수준에서 필터링할 수 있고, 반면 다른 염색체 마스크는 100 kb 수준에서 필터링할 수 있다. 그럼에도 불구하고, Y 염색체는 다른 염색체와 동일한 빈 크기 (가령, 100 kb)에서 정규화될 수 있다.

[0270] 일정한 구체예에서, 필터링된 Y 염색체는 도면 3a의 작업 315에서 앞서 설명된 바와 같이 정규화된다. 하지만, 만약 그렇지 않으면, Y 염색체는 더욱 교정되지 않는다. 따라서, Y 염색체 빈은 전역 프로필 제거에 종속되지 않는다. 유사하게, Y 염색체 빈은 그 후에 수행된 GC 교정 또는 다른 필터링 단계에 종속되지 않는다. 이것은 표본이 처리될 때, 상기 과정이 표본이 수컷 또는 암컷 표본인지를 알지 못하기 때문이다. 암컷 표본은 Y 참조 염색체에 정렬하는 어떤 리드도 갖지 않을 것이다.

[0271] **서열 마스크 창출**

[0272] 본원에서 개시된 일부 구체예는 서열 마스크를 이용하여 관심되는 서열 상에서 비-판별 서열 리드를 걸러 내기 위한 (또는 마스크를 적용하기 위한) 전략을 이용하는데, 이것은 CNV 평가에 이용된 리드 커버리지 값에서, 전통적인 방법에 의해 계산된 값에 비하여 더욱 높은 신호 및 더욱 낮은 잡음을 야기한다. 이런 마스크는 다양한 기술에 의해 확인될 수 있다. 한 구체예에서, 마스크는 아래에 더욱 상세하게 설명된 바와 같이 도면 4a-4b에서 예시된 기술을 이용하여 확인된다.

[0273] 일부 실행에서, 마스크는 관심되는 서열의 정상적인 사본수를 갖는 것으로 알려진 대표적인 표본의 훈련 세트를 이용하여 확인된다. 마스크는 먼저 훈련 세트 표본을 정규화하고, 이후 다양한 서열 (가령, 프로필) 전체에 대하여 체계적 변동을 교정하고, 그리고 이후 이들을 아래에 설명된 바와 같이 GC 가변성에 대해 교정하는 기술을

이용하여 확인될 수 있다. 정규화 및 교정은 시험 표본이 아닌 훈련 세트로부터 표본에서 수행된다. 마스크는 1회 확인되고, 그리고 이후, 많은 시험 표본에 적용된다.

[0274] 도면 4a는 이런 서열 마스크를 창출하기 위한 과정 400의 흐름도를 보여주는데, 이것은 사본수의 평가에서 고려사항으로부터 관심되는 서열 상에서 빈을 제거하기 위해 하나 또는 그 이상의 시험 표본에 적용될 수 있다. 상기 과정은 복수의 영향을 받지 않은 훈련 표본으로부터 서열 리드를 포함하는 훈련 세트를 제공함으로써 시작된다. 블록 402. 상기 과정은 이후, 훈련 세트의 서열 리드를 관심되는 서열을 포함하는 참조 유전체에 맞춰 정렬하고, 따라서 훈련 표본에 대한 훈련 서열 태그를 제공한다. 블록 404. 일부 구체예에서, 비-배제된 부위에 매핑된 단지 독특하게 정렬된 비다중 태그만 추가 분석에 이용된다. 상기 과정은 참조 유전체를 복수의 빈으로 분할하고, 그리고 각 훈련 표본을 위한 각 빈에서 훈련 서열 태그의 리드 커버리지를 각 영향을 받지 않은 훈련 표본에 대해 결정하는 것을 수반한다. 블록 406. 상기 과정은 또한, 모든 훈련 표본 전체에 대하여 훈련 서열 태그의 예상된 리드 커버리지를 각 빈에 대해 결정한다. 블록 408. 일부 구체예에서, 각 빈의 예상된 리드 커버리지는 훈련 표본 전체에 대하여 중앙값 또는 평균이다. 예상된 리드 커버리지는 전역 프로필을 구성한다. 상기 과정은 이후, 전역 프로필에서 변동을 제거함으로써 각 훈련 표본을 위한 각 빈에서 훈련 서열 태그의 리드 커버리지를 조정하고, 따라서 각 훈련 표본을 위한 빈에서 훈련 서열 태그의 전역-프로필-교정된 리드 커버리지를 획득한다. 상기 과정은 이후, 참조 유전체 전체에 대하여 마스크가 적용되지 않은 및 마스크가 적용된 빈을 포함하는 서열 마스크를 창출한다. 마스크가 적용된 빈 각각은 마스크 적용 역치를 초과하는 분포 특징을 갖는다. 분포 특징은 훈련 표본 전체에 대하여 빈 내에서 훈련 서열 태그의 조정된 리드 커버리지에 대해 제공된다. 일부 실행에서, 마스크 적용 역치는 훈련 표본 전체에 대하여 빈 내에서 정규화된 리드 커버리지에서 관찰된 변동에 관계할 수 있다. 표본 전체에 대하여 정규화된 리드 커버리지의 높은 변동 계수 또는 중위 절대 편차를 갖는 빈은 개별 메트릭스의 경험적 분포에 근거하여 확인될 수 있다. 일부 대안적 실행에서, 마스크 적용 역치는 훈련 표본 전체에 대하여 빈 내에서 정규화된 리드 커버리지에서 관찰된 변동에 관계할 수 있다. 표본 전체에 대하여 정규화된 리드 커버리지의 높은 변동 계수 또는 중위 절대 편차를 갖는 빈은 개별 메트릭스의 경험적 분포에 근거하여 마스크가 적용될 수 있다.

[0275] 일부 실행에서, 마스크가 적용된 빈을 확인하기 위한 별개의 컷오프, 다시 말하면, 마스크 적용 역치가 관심되는 염색체 및 모든 다른 염색체에 대해 규정된다. 게다가, 별개의 마스크 적용 역치가 관심되는 각 염색체에 대해 별개로 규정될 수 있고, 그리고 단일 마스크 적용 역치가 모든 비-영향을 받은 염색체의 세트에 대해 규정될 수 있다. 실례로서, 일정한 마스크 적용 역치에 근거된 마스크가 염색체 13에 대해 규정되고, 그리고 다른 마스크 적용 역치가 다른 염색체에 대한 마스크를 규정하는데 이용된다. 비-영향을 받은 염색체 역시 그들의 마스크 적용 역치가 염색체마다 규정될 수 있다.

[0276] 다양한 마스크 적용 역치 조합이 관심되는 각 염색체에 대해 평가될 수 있다. 마스크 적용 역치 조합은 관심되는 염색체의 빈에 대한 한 가지 마스크 및 모든 다른 염색체의 빈에 대한 상이한 마스크를 제공한다.

[0277] 한 가지 접근법에서, 변동 계수 (CV) 또는 표본 분포 컷오프의 척도에 대한 값의 범위가 빈 CV 값의 경험적 분포의 백분위수 (가령, 95, 96, 97, 98, 99)로서 규정되고, 그리고 이들 컷오프 값은 관심되는 염색체를 제외한 모든 상염색체에 적용된다. 게다가, CV에 대한 백분위수 컷오프 값의 범위가 경험적 CV 분포에 대해 규정되고, 그리고 이들 컷오프 값은 관심되는 염색체 (가령, chr 21)에 적용된다. 일부 구체예에서, 관심되는 염색체는 X 염색체 및 염색체 13, 18과 21이다. 당연히, 다른 접근법이 고려될 수 있다; 가령, 각 염색체에 대해 별개의 최적화가 수행될 수 있다. 종합하면, 병렬적으로 최적화되는 범위 (가령, 고려 중인 관심되는 염색체에 대한 한 가지 범위 및 모든 다른 염색체에 대한 다른 범위)가 CV 컷오프 조합의 격자를 규정한다. 도면 4b를 참조한다. 훈련 세트에서 시스템의 성과는 이들 2가지 컷오프 (정규화 염색체 (또는 관심되는 염색체 이외에 상염색체)에 대한 컷오프 및 관심되는 염색체에 대한 컷오프) 전체에 대하여 평가되고, 그리고 최적 수행 조합이 최종 배치를 위해 선택된다. 이러한 조합은 관심되는 염색체 각각에 대해 상이할 수 있다. 일정한 구체예에서, 성과는 훈련 세트 대신에 검증 세트에서 평가된다, 다시 말하면, 교차 검증이 성과를 평가하는데 이용된다.

[0278] 일부 구체예에서, 컷오프 범위를 결정하는데 최적화된 성과는 염색체 도스의 변동 계수 (정규화 염색체의 잠정적인 선별에 근거됨)이다. 상기 과정은 현재 선별된 정규화 염색체 (또는 염색체)를 이용하여 관심되는 염색체의 염색체 도스 (가령, 비율)의 CV를 최소화하는 컷오프의 조합을 선별한다. 한 가지 접근법에서, 상기 과정은 아래와 같이 격자 내에 컷오프의 각 조합의 성과를 시험한다: (1) 컷오프의 조합을 적용하여 모든 염색체에 대한 마스크를 규정하고, 그리고 이들 마스크를 적용하여 훈련 세트의 태그를 필터링한다; (2) 도면 3a의 과정을 필터링된 태그에 적용함으로써, 영향을 받지 않은 표본의 훈련 세트 전체에 대하여 정규화된 리드 커버리지를 계산한다; (3) 예로서, 고려 중인 염색체에 대한 빈의 정규화된 리드 커버리지를 가산함으로써 염색체마다 대표

적인 정규화된 리드 커버리지를 결정한다; (4) 현재 정규화 염색체를 이용하여 염색체 도스를 계산한다, 그리고 (5) 염색체 도스의 CV를 결정한다. 상기 과정은 선별된 필터를 훈련 세트의 본래 부분으로부터 분리된 한 세트의 시험 표본에 적용함으로써, 이들 선별된 필터의 성과를 사정할 수 있다. 다시 말하면, 상기 과정은 본래 훈련 세트를 훈련과 시험 부분집합으로 분할한다. 훈련 부분집합은 앞서 설명된 바와 같이, 마스크 컷오프를 규정하는데 이용된다.

[0279] 대안적 구체예에서, 리드 커버리지의 CV에 근거하여 마스크를 규정하는 대신에, 마스크는 빈 내에서 훈련 표본 전체에 대하여 정렬 결과로부터 매핑 품질평가점수의 분포에 의해 규정될 수 있다. 매핑 품질평가점수는 리드가 참조 유전체에 매핑되는 독특성을 반영한다. 다시 말하면, 매핑 품질평가점수는 리드가 오정렬되는 확률을 정량한다. 낮은 매핑 품질평가점수는 낮은 독특성 (높은 확률의 오정렬)과 연관된다. 독특성은 리드 서열 (서열분석기에 의해 산출될 때) 내에 하나 또는 그 이상의 오차를 설명한다. 매핑 품질평가점수의 상세한 설명은 Li H, Ruan J, Durbin R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* 18:1851-8에서 제공되고, 이것은 전체적으로 본원에 참조로서 편입된다. 일부 실행에서, 본원에서 매핑 품질평가점수는 MapQ 점수로서 지칭된다. 도면 4b는 MapQ 점수가 처리된 리드 커버리지의 CV와 강한 단조로운 상관을 갖는다는 것을 보여준다. 가령, 0.4보다 높은 CV를 갖는 빈은 도면 4b에서 플롯의 왼쪽에 거의 완전하게 무리를 이루고, 약 4보다 낮은 MapQ 점수를 갖는다. 이런 이유로, 작은 MapQ를 갖는 마스크가 적용된 빈은 높은 CV를 갖는 마스크가 적용된 빈에 의해 규정된 것과 매우 유사한 마스크를 산출할 수 있다.

[0280] **표본 및 표본 처리**

[0281] *표본*

[0282] CNV, 예를 들면, 염색체 이수성, 부분적인 이수성 등을 결정하는데 이용되는 표본은 하나 또는 그 이상의 관심되는 서열에 대한 사본수 변동이 결정되는 임의의 세포, 조직, 또는 장기로부터 채취된 표본을 포함할 수 있다. 바람직하게, 표본은 세포 내에 존재하는 핵산 및/또는 "무세포"인 핵산 (가령, cfDNA)을 내포한다.

[0283] 일부 구체예에서, 무세포 핵산, 예를 들면, 무세포 DNA (cfDNA)를 획득하는 것이 유리하다. 무세포 DNA를 비롯한 무세포 핵산은 혈장, 혈청, 그리고 소변을 포함하지만 이들에 한정되지 않는 생물학적 표본으로부터 다양한 당분야에서 공지된 방법에 의해 획득될 수 있다 (가령, Fan et al., Proc Natl Acad Sci 105:16266-16271 [2008]; Koide et al., Prenatal Diagnosis 25:604-607 [2005]; Chen et al., Nature Med. 2: 1033-1035 [1996]; Lo et al., Lancet 350: 485-487 [1997]; Botezatu et al., Clin Chem. 46: 1078-1084, 2000; 그리고 Su et al., J Mol. Diagn. 6: 101-107 [2004]를 참조한다). 표본 내에 세포로부터 무세포 DNA를 분리하기 위해, 분별, 원심분리 (가령, 밀도 기울기 원심분리), DNA-특이적 침전, 또는 고처리량 세포 분류 및/또는 다른 분리 방법을 포함하지만 이들에 한정되지 않는 다양한 방법이 이용될 수 있다. cfDNA의 수동과 자동화된 분리를 위한 상업적으로 가용한 키트가 가용하다 (Roche Diagnostics, Indianapolis, IN, Qiagen, Valencia, CA, Macherey-Nagel, Duren, DE). cfDNA를 포함하는 생물학적 표본은 염색체 이수성 및/또는 다양한 다형성을 검출할 수 있는 염기서열결정 검정에 의해, 염색체 비정상, 예를 들면, 삼염색체성 21의 존재 또는 부재를 결정하기 위한 검정에서 이용되었다.

[0284] 다양한 구체예에서, 표본 내에 존재하는 cfDNA는 이용에 앞서 (가령, 염기서열결정 라이브러리를 제조하기에 앞서), 특이적으로 또는 비특이적으로 농축될 수 있다. 표본 DNA의 비특이적 농축은 cfDNA 염기서열결정 라이브러리를 제조하기에 앞서, 표본 DNA의 수준을 증가시키는데 이용될 수 있는, 표본의 유전체 DNA 단편의 전체 유전체 증폭을 지칭한다. 비특이적 농축은 하나 이상의 유전체를 포함하는 표본 내에 존재하는 두 유전체 중에서 한 가지의 선택적 농축일 수 있다. 가령, 비특이적 농축은 표본 내에 태아 대 모계 DNA의 상대적 비율을 증가시키는 공지된 방법에 의해 획득될 수 있는, 모계 표본 내에 태아 유전체의 선택적 농축일 수 있다. 대안으로, 비특이적 농축은 표본 내에 존재하는 양쪽 유전체의 비선별적인 증폭일 수 있다. 가령, 비특이적 증폭은 태아와 모계 유전체로부터 DNA의 혼합물을 포함하는 표본에서 태아와 모계 DNA의 비특이적 증폭일 수 있다. 전체 유전체 증폭을 위한 방법은 당분야에서 공지된다. 축중성 올리고뉴클레오티드-기폭된 PCR (DOP), 프라이머 연장 PCR 기술 (PEP) 및 다중 변위 증폭 (MDA)이 전체 유전체 증폭 방법의 실례이다. 일부 구체예에서, 상이한 유전체로부터 cfDNA의 혼합물을 포함하는 표본은 혼합물 내에 존재하는 유전체의 cfDNA에 대해 농축되지 않는다. 다른 구체예에서, 상이한 유전체로부터 cfDNA의 혼합물을 포함하는 표본은 표본 내에 존재하는 유전체 중에서 임의의 한 가지에 대해 비특이적으로 농축된다.

[0285] 본원에서 설명된 방법이 적용되는 핵산(들)을 포함하는 표본은 전형적으로, 예로서 앞서 설명된 바와 같이, 생물학적 표본 ("시험 표본")을 포함한다. 일부 구체예에서, 하나 또는 그 이상의 CNV에 대해 선별검사되는 핵산

(들)은 다수의 널리 공지된 방법 중에서 한 가지에 의해 정제되거나 또는 단리된다.

[0286] 따라서, 일정한 구체예에서, 표본은 정제된 또는 단리된 폴리뉴클레오티드를 포함하거나 또는 이것으로 구성되고, 또는 이것은 조직 표본, 생물학적 유체 표본, 세포 표본 등과 같은 표본을 포함할 수 있다. 적합한 생물학적 유체 표본은 혈액, 혈장, 혈청, 땀, 눈물, 객담, 소변, 객담, 귀 흐름, 림프, 타액, 뇌척수액, 손상, 골수 현탁액, 질 흐름, 경경부 세척액, 뇌 유체, 복수, 우유, 호흡기, 장관과 비뇨생식기로의 분비액, 양수, 모유, 그리고 류코포레시스 표본을 포함하지만 이들에 한정되지 않는다. 일부 구체예에서, 표본은 비침습성 시술에 의해 쉽게 획득가능한 표본, 예를 들면, 혈액, 혈장, 혈청, 땀, 눈물, 객담, 소변, 객담, 귀 흐름, 타액 또는 대변이다. 일정한 구체예에서, 표본은 말초혈 표본, 또는 말초혈 표본의 혈장 및/또는 혈청 분획물이다. 다른 구체예에서, 생물학적 표본은 면봉 또는 도말, 생검 검체, 또는 세포 배양액이다. 다른 구체예에서, 표본은 2개 또는 그 이상 생물학적 표본의 혼합물이다, 예를 들면, 생물학적 표본은 생물학적 유체 표본, 조직 표본, 그리고 세포 배양 표본 중에서 2개 또는 그 이상을 포함할 수 있다. 본원에서 이용된 바와 같이, 용어 "혈액," "혈장" 및 "혈청"은 분획물 또는 이들의 처리된 부분을 명시적으로 포괄한다. 유사하게, 표본이 생검, 면봉, 도말 등으로부터 채취되는 경우에, "표본"은 생검, 면봉, 도말 등으로부터 유래된 처리된 분획물 또는 부분을 명시적으로 포괄한다.

[0287] 일정한 구체예에서, 표본은 상이한 개체로부터 표본, 동일하거나 상이한 개체의 상이한 발달 단계로부터 표본, 상이한 병든 개체 (가령, 암을 앓거나 또는 유전 질환을 앓는 것으로 의심되는 개체)로부터 표본, 정상적인 개체, 개체에서 질환의 상이한 시기에서 획득된 표본, 질환에 대한 상이한 치료에 종속된 개체로부터 획득된 표본, 상이한 환경적 인자에 종속된 개체로부터 표본, 병리에 대한 소인을 갖는 개체로부터 표본, 감염성 질환 병원체 (가령, HIV)에 노출된 표본 개체, 기타 등등을 포함하지만 이들에 한정되지 않는 공급원으로부터 획득될 수 있다.

[0288] 예시적인, 하지만 무제한적 구체예에서, 표본은 임신 암컷, 예를 들면, 임신 여성으로부터 획득되는 모계 표본이다. 이러한 사례에서, 표본은 태아에서 잠재적 염색체 비정상의 출생전 진단을 제공하는 본원에서 설명된 방법을 이용하여 분석될 수 있다. 모계 표본은 조직 표본, 생물학적 유체 표본, 또는 세포 표본일 수 있다. 생물학적 유체는 무제한적 실례로서, 혈액, 혈장, 혈청, 땀, 눈물, 객담, 소변, 객담, 귀 흐름, 림프, 타액, 뇌척수액, 손상, 골수 현탁액, 질 흐름, 경경부 세척액, 뇌 유체, 복수, 모유, 호흡기, 장관과 비뇨생식기로의 분비액, 그리고 류코포레시스 표본을 포함한다.

[0289] 다른 예시적인, 하지만 무제한적 구체예에서, 모계 표본은 2개 또는 그 이상 생물학적 표본의 혼합물이다, 예를 들면, 생물학적 표본은 생물학적 유체 표본, 조직 표본, 그리고 세포 배양 표본 중에서 2개 또는 그 이상을 포함할 수 있다. 일부 구체예에서, 표본은 비침습성 시술에 의해 쉽게 획득가능한 표본, 예를 들면, 혈액, 혈장, 혈청, 땀, 눈물, 객담, 소변, 모유, 객담, 귀 흐름, 타액 및 대변이다. 일부 구체예에서, 생물학적 표본은 말초혈 표본, 및/또는 이의 혈장과 혈청 분획물이다. 다른 구체예에서, 생물학적 표본은 면봉 또는 도말, 생검 검체, 또는 세포 배양액의 표본이다. 상기 개시된 바와 같이, 용어 "혈액," "혈장" 및 "혈청"은 분획물 또는 이들의 처리된 부분을 명시적으로 포괄한다. 유사하게, 표본이 생검, 면봉, 도말 등으로부터 채취되는 경우에, "표본"은 생검, 면봉, 도말 등으로부터 유래된 처리된 분획물 또는 부분을 명시적으로 포괄한다.

[0290] 일정한 구체예에서, 표본은 또한, 시험관내 배양된 조직, 세포, 또는 다른 폴리뉴클레오티드-내포 공급원으로부터 획득될 수 있다. 배양된 표본은 상이한 배지와 조건 (가령, pH, 압력, 또는 온도)에서 유지된 배양액 (가령, 조직 또는 세포), 상이한 길이의 기간 동안 유지된 배양액 (가령, 조직 또는 세포), 상이한 인자 또는 시약 (가령, 약물 후보, 또는 조절인자)으로 처리된 배양액 (가령, 조직 또는 세포), 또는 상이한 유형의 조직 및/또는 세포의 배양액을 포함하지만 이들에 한정되지 않는 공급원으로부터 채취될 수 있다.

[0291] 생물학적 공급원으로부터 핵산을 단리하는 방법은 널리 공지되어 있고, 그리고 공급원의 성격에 따라 달라질 것이다. 당업자는 본원에서 설명된 방법에 대한 필요에 따라, 공급원으로부터 핵산(들)을 쉽게 단리할 수 있다. 일부 경우에, 핵산 표본 내에 핵산 분자를 단편화하는 것이 유리할 수 있다. 단편화는 무작위이거나, 또는 예로서, 제한 엔도뉴클레아제 소화를 이용하여 달성될 때 특정할 수 있다. 무작위 단편화를 위한 방법은 당분야에서 널리 공지되고, 그리고 예로서, 제한된 DNA분해효소 소화, 알칼리 처리 및 물리적 전단을 포함한다. 한 구체예에서, 표본 핵산은 cfDNA로서 획득되고, 이것은 단편화에 종속되지 않는다.

[0292] 다른 예시적인 구체예에서, 표본 핵산(들)은 유전체 DNA로서 획득되고, 이것은 대략 300 또는 그 이상, 대략 400 또는 그 이상, 또는 대략 500 또는 그 이상 염기쌍의 단편으로의 단편화에 종속되고, 그리고 여기에 NGS 방법이 쉽게 적용될 수 있다.

[0293] *염기서열결정 라이브러리 제조*

[0294] 한 구체예에서, 본원에서 설명된 방법은 복수 표본이 단일 염기서열결정 실행에서, 유전체 분자로서 개별적으로 (즉, 단일플렉스 염기서열결정) 또는 색인된 유전체 분자를 포함하는 모아진 표본으로서 (가령, 멀티플렉스 염기서열결정) 염기서열결정되도록 허용하는 차세대 염기서열결정 기술 (NGS)을 활용할 수 있다. 이들 방법은 DNA 서열의 최대 수억 개 리드를 산출할 수 있다. 다양한 구체예에서, 유전체 핵산, 및/또는 색인된 유전체 핵산의 서열은 예로서, 본원에서 설명된 차세대 염기서열결정 기술 (NGS)을 이용하여 결정될 수 있다. 다양한 구체예에서, NGS를 이용하여 획득된 많은 양의 서열 데이터의 분석이 본원에서 설명된 바와 같은 하나 또는 그 이상의 프로세서를 이용하여 수행될 수 있다.

[0295] 다양한 구체예에서, 이런 염기서열결정 기술의 이용은 염기서열결정 라이브러리의 제조를 수반하지 않는다.

[0296] 하지만, 일정한 구체예에서, 본원에서 예기된 염기서열결정 방법은 염기서열결정 라이브러리의 제조를 수반한다. 한 예시적인 접근법에서, 염기서열결정 라이브러리 제조는 염기서열결정될 준비가 된 어댑터-변형된 DNA 단편 (가령, 폴리뉴클레오티드)의 무작위 수집물의 생산을 수반한다. 폴리뉴클레오티드의 염기서열결정 라이브러리는 DNA 또는 RNA뿐만 아니라 DNA 또는 cDNA 중에서 어느 한쪽의 등가물, 유사체, 예를 들면, 역전사효소의 작용에 의해 RNA 주형으로부터 생산된 상보성 또는 사본 DNA인 DNA 또는 cDNA로부터 제조될 수 있다. 폴리뉴클레오티드는 이중 가닥 형태 (가령, dsDNA, 예를 들면, 유전체 DNA 단편, cDNA, PCR 증폭 산물 등)에서 기원하거나, 또는 일정한 구체예에서, 폴리뉴클레오티드는 단일 가닥 형태 (가령, ssDNA, RNA 등)에서 기원하고 dsDNA 형태로 전환될 수 있다. 실례로서, 일정한 구체예에서, 단일 가닥 mRNA 분자는 염기서열결정 라이브러리를 제조하는데 이용하기 적합한 이중 가닥 cDNA로 복제될 수 있다. 일차성 폴리뉴클레오티드 분자의 정밀한 서열은 일반적으로, 라이브러리 제조의 방법에 재료가 아니고, 그리고 공지이거나 또는 미지일 수 있다. 한 구체예에서, 폴리뉴클레오티드 분자는 DNA 분자이다. 더욱 구체적으로, 일정한 구체예에서, 폴리뉴클레오티드 분자는 생물체의 전체 유전자 보체 또는 생물체의 실제적으로 전체 유전자 보체를 나타내고, 그리고 인트론 서열과 엑손 서열 (코딩 서열) 둘 모두 뿐만 아니라 비코딩 조절 서열, 예를 들면, 프로모터와 인핸서 서열을 전형적으로 포함하는 유전체 DNA 분자 (가령, 세포 DNA, 무세포 DNA (cfDNA) 등)이다. 일정한 구체예에서, 일차성 폴리뉴클레오티드 분자는 인간 유전체 DNA 분자, 예를 들면, 임신 개체의 말초혈 내에 존재하는 cfDNA 분자를 포함한다.

[0297] 일부 NGS 염기서열결정 플랫폼을 위한 염기서열결정 라이브러리의 제조는 특정한 범위의 단편 크기를 포함하는 폴리뉴클레오티드의 이용에 의해 가능해진다. 이런 라이브러리의 제조는 전형적으로, 원하는 크기 범위에서 폴리뉴클레오티드를 획득하기 위한 큰 폴리뉴클레오티드 (가령, 세포 유전체 DNA)의 단편화를 수반한다.

[0298] 단편화는 당업자에게 공지된 다수의 방법 중에서 한 가지에 의해 달성될 수 있다. 가령, 단편화는 분무, 초음파처리 및 수리전단을 포함하지만 이들에 한정되지 않는 기계적 수단에 의해 달성될 수 있다. 하지만 기계적 단편화는 전형적으로, C-O, P-O 및 C-C 결합에서 DNA 중추를 개열하여, 파괴된 C-O, P-O 및/또는 C-C 결합을 갖는 평활 단부 및 3'- 및 5'-돌출 단부의 이질성 혼합물을 유발하는데 (가령, Alnemri and Liwack, J Biol. Chem 265:17323-17333 [1990]; Richards and Boyer, J Mol Biol 11:327-240 [1965]를 참조한다), 이들 결합은 염기서열결정을 위한 DNA를 준비하는데 필요한 차후 효소적 반응, 예를 들면, 염기서열결정 어댑터의 결찰을 위해 필수적인 5'-인산염을 결여할 수 있기 때문에, 수복될 필요가 있을 수도 있다.

[0299] 대조적으로, cfDNA는 전형적으로, 약 300개보다 적은 염기쌍의 단편으로서 존재하고, 그리고 결과적으로, cfDNA 표본을 이용하여 염기서열결정 라이브러리를 산출하는 경우에 단편화가 전형적으로 필요하지 않다.

[0300] 전형적으로, 폴리뉴클레오티드가 강제적으로 단편화되거나 (가령, 시험관내 단편화된), 또는 자연적으로 단편으로서 존재하는 지에 상관없이, 이들은 5'-인산염 및 3'-히드록실을 갖는 평활 말단 DNA로 전환된다. 표준 프로토콜, 예를 들면, 예로서, 본원의 다른 곳에서 설명된 바와 같은 Illumina 플랫폼을 이용하여 염기서열결정하기 위한 프로토콜은 사용자에게 표본 DNA를 단부 수복하고, 단부 수복된 산물을 dA-테일링에 앞서 정제하고, 그리고 dA-테일링 산물을 라이브러리 제조물의 어댑터-결찰 단계에 앞서 정제할 것을 지시한다.

[0301] 본원에서 설명된 서열 라이브러리 제조의 방법의 다양한 구체예는 NGS에 의해 염기서열결정될 수 있는 변형된 DNA 산물을 획득하기 위한 표준 프로토콜에 의해 전형적으로 위임된 단계 중에서 하나 또는 그 이상을 수행해야 하는 필요를 배제한다. 단축된 방법 (ABB 방법), 1-단계 방법, 그리고 2-단계 방법은 염기서열결정 라이브러리의 제조 방법의 실례인데, 이것은 전체적으로 참조로서 편입되는, 2012년 7월 20일자 제출된 특허 출원 13/555,037에서 발견될 수 있다.

[0302] *표본 완전성을 추적하고 실증하기 위한 마커 핵산*

[0303] 다양한 구체예에서, 표본의 완전성 및 표본 추적의 실증은 표본 유전체 핵산, 예를 들면, cfDNA, 그리고 예로서, 처리에 앞서 표본 내로 도입되었던 동행하는 마커 핵산의 혼합물을 염기서열결정함으로써 달성될 수 있다.

[0304] 마커 핵산은 시험 표본 (가령, 생물학적 공급원 표본)과 결합되고, 그리고 예로서, 생물학적 공급원 표본을 분획하는 단계, 예를 들면, 전혈 표본으로부터 본질적으로 무세포 혈장 분획물을 획득하는 단계, 분획된, 예를 들면, 혈장, 또는 미분획된 생물학적 공급원 표본, 예를 들면, 조직 표본으로부터 핵산을 정제하는 단계, 그리고 염기서열결정하는 단계 중에서 하나 또는 그 이상을 포함하는 과정에 종속될 수 있다. 일부 구체예에서, 염기서열결정은 염기서열결정 라이브러리를 제조하는 것을 포함한다. 공급원 표본과 결합되는 마커 분자의 서열 또는 서열의 조합은 공급원 표본에 독특하도록 선택된다. 일부 구체예에서, 표본 내에 독특한 마커 분자 모두 동일한 서열을 갖는다. 다른 구체예에서, 표본 내에 독특한 마커 분자는 복수의 서열, 예를 들면, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20개, 또는 그 이상의 상이한 서열의 조합이다.

[0305] 한 구체예에서, 표본의 완전성은 동일한 서열을 갖는 복수의 마커 핵산 분자를 이용하여 실증될 수 있다. 대안으로, 표본의 동일성은 최소한 2, 최소한 3, 최소한 4, 최소한 5, 최소한 6, 최소한 7, 최소한 8, 최소한 9, 최소한 10, 최소한 11, 최소한 12, 최소한 13, 최소한 14, 최소한 15, 최소한 16, 최소한 17, 최소한 18, 최소한 19, 최소한 20, 최소한 25, 최소한 30, 최소한 35, 최소한 40, 최소한 50개, 또는 그 이상의 상이한 서열을 갖는 복수의 마커 핵산 분자를 이용하여 실증될 수 있다. 복수의 생물학적 표본, 다시 말하면, 2개 또는 그 이상 생물학적 표본의 완전성의 실증은 2개 또는 그 이상의 표본 각각이 표시되는 복수의 시험 표본 각각에 독특한 서열을 갖는 마커 핵산으로 표시될 것을 요구한다. 가령, 첫 번째 표본은 서열 A를 갖는 마커 핵산으로 표시될 수 있고, 그리고 두 번째 표본은 서열 B를 갖는 마커 핵산으로 표시될 수 있다. 대안으로, 첫 번째 표본은 서열 A를 갖는 마커 핵산 분자 모두로 표시될 수 있고, 그리고 두 번째 표본은 서열 B와 C의 혼합물로 표시될 수 있고, 여기서 서열 A, B와 C는 상이한 서열을 갖는 마커 분자이다.

[0306] 마커 핵산(들)은 라이브러리 제조 (라이브러리가 제조되면) 및 염기서열결정에 앞서 발생하는 표본 제조의 임의의 시기에서 표본에 첨가될 수 있다. 한 구체예에서, 마커 분자는 처리되지 않은 공급원 표본과 결합될 수 있다. 가령, 마커 핵산은 혈액 표본을 수집하는데 이용되는 수집물 튜브에서 제공될 수 있다. 대안으로, 마커 핵산은 채혈 이후에 혈액 표본에 첨가될 수 있다. 한 구체예에서, 마커 핵산은 생물학적 유체 표본을 수집하는데 이용되는 용기에 첨가된다, 예를 들면, 마커 핵산(들)은 혈액 표본을 수집하는데 이용되는 혈액 수집 튜브에 첨가된다. 다른 구체예에서, 마커 핵산(들)은 생물학적 유체 표본의 분획물에 첨가된다. 가령, 마커 핵산은 혈액 표본, 예를 들면, 모계 혈장 표본의 혈장 및/또는 혈청 분획물에 첨가된다. 또 다른 구체예에서, 마커 분자는 정제된 표본, 예를 들면, 생물학적 표본으로부터 정제되었던 핵산의 표본에 첨가된다. 가령, 마커 핵산은 정제된 모계와 태아 cfDNA의 표본에 첨가된다. 유사하게, 마커 핵산은 검체를 처리하기 전에, 생검 검체에 첨가될 수 있다. 일부 구체예에서, 마커 핵산은 마커 분자를 생물학적 표본의 세포 내로 전달하는 운반체와 결합될 수 있다. 세포-전달 운반체는 pH-민감성 및 양이온성 리포솜을 포함한다.

[0307] 다양한 구체예에서, 마커 분자는 생물학적 공급원 표본의 유전체로부터 부재하는 서열인 항유전체성 서열을 갖는다. 한 예시적인 구체예에서, 인간 생물학적 공급원 표본의 완전성을 실증하는데 이용되는 마커 분자는 인간 유전체로부터 부재하는 서열을 갖는다. 대안적 구체예에서, 마커 분자는 공급원 표본으로부터 및 임의의 한 가지 또는 그 이상의 다른 공지된 유전체로부터 부재하는 서열을 갖는다. 가령, 인간 생물학적 공급원 표본의 완전성을 실증하는데 이용되는 마커 분자는 인간 유전체로부터 및 생쥐 유전체로부터 부재하는 서열을 갖는다. 대안은 2개 또는 그 이상의 유전체를 포함하는 시험 표본의 완전성을 실증하는 것을 허용한다. 가령, 병원체, 예를 들면, 세균에 의해 영향을 받는 개체로부터 획득된 인간 무세포 DNA 표본의 완전성은 인간 유전체 및 영향을 주는 세균의 유전체 둘 모두로부터 부재하는 서열을 갖는 마커 분자를 이용하여 실증될 수 있다. 다양한 병원체, 예를 들면, 세균, 바이러스, 효모, 균류, 원생동물 등의 유전체의 서열은 월드와이드웹 상에 ncbi.nlm.nih.gov/genomes에서 공개적으로 가용하다. 다른 구체예에서, 마커 분자는 임의의 공지된 유전체로부터 부재하는 서열을 갖는 핵산이다. 마커 분자의 서열은 알고리즘적으로 무작위 산출될 수 있다.

[0308] 다양한 구체예에서, 마커 분자는 자연발생 데옥시리보핵산 (DNA), 리보핵산, 또는 펩티드 핵산 (PNA), 모르폴리노 핵산, 잠금된 핵산, 글리콜 핵산, 그리고 트레오스 핵산을 비롯한 인공 핵산 유사체 (핵산 모방체)일 수 있는데, 이들은 포스포디에스테르 중추를 갖지 않는 분자 또는 DNA 모방체의 중추로의 변화에 의해 자연발생 DNA 또는 RNA와 구별된다. 데옥시리보핵산은 자연발생 유전체로부터 유래될 수 있거나 또는 효소의 이용을 통해 또

는 고체상 화학적 합성에 의해 실험실에서 산출될 수 있다. 화학적 방법은 또한, 자연에서 발견되지 않는 DNA 모방체를 산출하는데 이용될 수 있다. 포스포디에스테르 연쇄가 대체되었지만 데옥시리보오스가 유지되는 가용한 DNA의 유도체는 우수한 구조적 DNA 모방체인 것으로 밝혀진, 티오포름아세탈 또는 카르복스아미드 연쇄에 의해 형성된 중추를 갖는 DNA 모방체를 포함하지만 이들에 한정되지 않는다. 다른 DNA 모방체는 모르폴리노 유도체 및 펩티드 핵산 (PNA)을 포함하는데, 이들은 N-(2-아미노에틸)글리신-기초된 슈도펩티드 중추를 내포한다 (Ann Rev Biophys Biomol Struct 24:167-183 [1995]). PNA는 DNA (또는 리보핵산 [RNA])의 극히 우수한 구조적 모방체이고, 그리고 PNA 소중합체는 왓슨 크릭 상보성 DNA와 RNA (또는 PNA) 소중합체와 매우 안정된 이중나선 구조를 형성할 수 있고, 그리고 이들은 또한, 나선 침입에 의해 이중나선 DNA에서 표적에 결합할 수 있다 (Mol Biotechnol 26:233-248 [2004]. 마커 분자로서 이용될 수 있는 DNA 유사체의 다른 우수한 구조적 모방체/유사체는 비-가교화 산소 중에서 하나가 황에 의해 대체되는 포스포로티오에이트 DNA이다. 이러한 변형은 5'에서 3' 및 3'에서 5' DNA POL 1 엑소뉴클레아제, 뉴클레아제 S1와 P1, RNA분해효소, 혈청 뉴클레아제 및 뱀독 포스포디에스테라아제를 비롯한 엔도-와 엑소뉴클레아제 2의 작용을 감소시킨다.

[0309] 마커 분자의 길이는 표본 핵산의 길이와 상이하거나 또는 상이하지 않을 수 있다, 다시 말하면, 마커 분자의 길이는 표본 유전체 분자의 길이와 유사할 수 있거나, 또는 표본 유전체 분자의 길이보다 크거나 또는 작을 수 있다. 마커 분자의 길이는 마커 분자를 구성하는 뉴클레오티드 또는 뉴클레오티드 유사체 염기의 숫자에 의해 계측된다. 표본 유전체 분자의 것들과 상이한 길이를 갖는 마커 분자는 당분야에서 공지된 분리 방법을 이용하여 공급원 핵산과 구별될 수 있다. 가령, 마커와 표본 핵산 분자의 길이에서 차이는 전기영동 분리, 예를 들면, 모세관 전기영동에 의해 결정될 수 있다. 크기 차등화는 마커와 표본 핵산의 품질을 정량하고 사정하는데 유리할 수 있다. 바람직하게는, 마커 핵산은 유전체 핵산보다 짧고, 그리고 이들이 표본의 유전체에 매핑되는 것을 배제할 만큼 충분한 길이를 갖는다. 가령, 30개 염기 인간 서열이 이를 인간 유전체에 독특하게 매핑하는데 필요하다. 따라서 일정한 구체예에서, 인간 표본의 생물학적 검정을 염기서열결정하는데 이용되는 마커 분자는 길이에서 최소한 30 bp이어야 한다.

[0310] 마커 분자의 길이의 선택은 일차적으로, 공급원 표본의 완전성을 실증하는데 이용되는 염기서열결정 기술에 의해 결정된다. 염기서열결정되는 표본 유전체 핵산의 길이 역시 고려될 수 있다. 가령, 일부 염기서열결정 기술은 폴리뉴클레오티드의 클론 증폭을 이용하는데, 이것은 클론에 의해 증폭되는 유전체 폴리뉴클레오티드가 최소 길이일 것을 요구할 수 있다. 가령, Illumina GAII 서열 분석기를 이용한 염기서열결정은 110bp의 최소 길이를 갖고, 어댑터가 결찰되면 클론에 의해 증폭되고 염기서열결정될 수 있는 최소한 200 bp 및 600 bp 이하의 핵산을 제공하는 폴리뉴클레오티드의 가교 PCR에 의한 시험관내 클론 증폭 (클러스터 증폭으로서 또한 알려져 있음)을 포함한다. 일부 구체예에서, 어댑터-결찰된 마커 분자의 길이는 약 200bp 내지 약 600bp, 약 250bp 내지 550bp, 약 300bp 내지 500bp, 또는 약 350bp 내지 450bp이다. 다른 구체예에서, 어댑터-결찰된 마커 분자의 길이는 약 200bp이다. 가령, 모계 표본 내에 존재하는 태아 cfDNA를 염기서열결정할 때, 마커 분자의 길이는 태아 cfDNA 분자의 것과 유사하도록 선택될 수 있다. 따라서, 한 구체예에서, 태아 염색체 이수성의 존재 또는 부재를 결정하기 위해 모계 표본 내에 cfDNA의 대량으로 병렬 염기서열결정을 포함하는 검정에서 이용되는 마커 분자의 길이는 약 150 bp, 약 160bp, 170 bp, 약 180bp, 약 190bp 또는 약 200bp일 수 있다; 바람직하게는, 마커 분자는 약 170 bp이다. 다른 염기서열결정 접근법, 예를 들면, SOLiD 염기서열결정, 폴로니 염기서열결정 및 454 염기서열결정은 유제 PCR을 이용하여, 염기서열결정하기 위한 DNA 분자를 클론에 의해 증폭하고, 그리고 각 기술은 증폭되는 분자의 최소와 최고 길이를 강제한다. 클론에 의해 증폭된 핵산으로서 염기서열결정되는 마커 분자의 길이는 최대 약 600bp일 수 있다. 일부 구체예에서, 염기서열결정되는 마커 분자의 길이는 600bp보다 클 수 있다.

[0311] 분자의 클론 증폭을 이용하지 않고, 그리고 매우 광범위한 범위의 주형 길이에 걸쳐 핵산을 염기서열결정할 수 있는 단일 분자 염기서열결정 기술은 대부분의 환경에서, 염기서열결정되는 분자가 임의의 특정한 길이일 것을 요구하지 않는다. 하지만, 단위 질량당 서열의 수율은 3' 단부 히드록실 기의 숫자에 의존하고, 그리고 따라서, 염기서열결정하기 위한 상대적으로 짧은 주형을 갖는 것은 긴 주형을 갖는 것보다 효율적이다. 1000 nt보다 긴 핵산으로 시작하면, 더욱 많은 서열 정보가 동일한 질량의 핵산으로부터 산출될 수 있도록, 핵산을 100 내지 200 nt의 평균 길이로 전단하는 것이 일반적으로 권장된다. 따라서, 마커 분자의 길이는 수십 개의 염기 내지 수천 개의 염기 범위에서 변할 수 있다. 단일 분자 염기서열결정에 이용되는 마커 분자의 길이는 길이에서 최대 약 25bp, 최대 약 50bp, 최대 약 75bp, 최대 약 100bp, 최대 약 200bp, 최대 약 300bp, 최대 약 400bp, 최대 약 500bp, 최대 약 600bp, 최대 약 700bp, 최대 약 800 bp, 최대 약 900bp, 최대 약 1000bp, 또는 그 이상일 수 있다.

[0312] 마커 분자에 대해 선택되는 길이는 또한, 염기서열결정되는 유전체 핵산의 길이에 의해 결정된다. 가령, cfDNA는 인간 혈류에서 세포 유전체 DNA의 유전체 단편으로서 순환한다. 임신 여성의 혈장에서 발견되는 태아 cfDNA 분자는 일반적으로, 모계 cfDNA 분자보다 짧다 (Chan et al., Clin Chem 50:8892 [2004]). 순환하는 태아 DNA의 크기 분별은 순환하는 태아 DNA 단편의 평균 길이가 <300 bp라는 것을 확증하였고, 반면 모계 DNA는 약 0.5와 1 Kb 사이인 것으로 추정되었다 (Li et al., Clin Chem, 50: 1002-1011 [2004]). 이들 조사 결과는 Fan 등의 것들과 일치하는데, 이들은 NGS를 이용하여, 태아 cfDNA가 드물게 >340bp라는 것을 확인하였다 (Fan et al., Clin Chem 56:1279-1286 [2010]). 표준 실리카-기초된 방법으로 소변으로부터 단리된 DNA는 2가지 분획물, shed 세포로부터 기원하는 고분자량 DNA 및 경신장 DNA (Tr-DNA)의 저분자량 (150-250 염기쌍) 분획물로 구성된다 (Botezatu et al., Clin Chem. 46: 1078-1084, 2000; 그리고 Su et al., J Mol. Diagn. 6: 101-107, 2004). 체액으로부터 무세포 핵산의 단리를 위해 새로 개발된 기술을 경신장 핵산의 단리에 적용하는 것은 소변 내에 150개 염기쌍보다 훨씬 짧은 DNA와 RNA 단편의 존재를 드러냈다 (U.S. 특허 출원 공개 번호 20080139801). cfDNA가 염기서열결정되는 유전체 핵산인 구체예에서, 선택되는 마커 분자는 대략 cfDNA의 길이까지일 수 있다. 가령, 단일 핵산 분자로서 또는 클론에 의해 증폭된 핵산으로서 염기서열결정되는 모계 cfDNA 표본에서 이용된 마커 분자의 길이는 약 100 bp와 600bp 사이일 수 있다. 다른 구체예에서, 표본 유전체 핵산은 더욱 큰 분자의 단편이다. 가령, 염기서열결정되는 표본 유전체 핵산은 단편화된 세포 DNA이다. 단편화된 세포 DNA가 염기서열결정되는 구체예에서, 마커 분자의 길이는 DNA 단편의 길이까지일 수 있다. 일부 구체예에서, 마커 분자의 길이는 최소한, 서열 리드를 적절한 참조 유전체에 독특하게 매핑하는데 필요한 최소 길이이다. 다른 구체예에서, 마커 분자의 길이는 마커 분자가 표본 참조 유전체에 매핑되는 것을 배제하는데 필요한 최소 길이이다.

[0313] 이에 더하여, 마커 분자는 핵산 염기서열결정에 의해 검정되지 않고, 그리고 염기서열결정 이외에 통상적인 생물-기술, 예를 들면, 실시간 PCR에 의해 실증될 수 있는 표본을 실증하는데 이용될 수 있다.

[0314] *표본 대조 (가령, 염기서열결정 및/또는 분석을 위한 과정중 양성 대조).*

[0315] 다양한 구체예에서, 예로서 앞서 설명된 바와 같이 표본 내로 도입된 마커 서열은 염기서열결정 및 차후 처리와 분석의 정확도와 효력을 실증하기 위한 양성 대조로서 기능할 수 있다.

[0316] 따라서, 표본 내에 DNA를 염기서열결정하기 위한 과정중 양성 대조 (IPC)를 제공하기 위한 조성물과 방법이 제공된다. 일정한 구체예에서, 유전체의 혼합물을 포함하는 표본 내에 cfDNA를 염기서열결정하기 위한 양성 대조가 제공된다. IPC는 표본, 예를 들면, 상이한 염기서열결정 실행에서 상이한 시점에서 염기서열결정되는 표본의 상이한 세트로부터 획득된 서열 정보에서 기준선 이동을 관련시키는데 이용될 수 있다. 따라서, 예로서, IPC는 모계 시험 표본에 대해 획득된 서열 정보를 상이한 시점에서 염기서열화되었던 한 세트의 유자격 표본으로부터 획득된 서열 정보에 관련시킬 수 있다.

[0317] 유사하게, 분절 분석의 경우에, IPC는 특정 분절(들)에 대해 개체로부터 획득된 서열 정보를 상이한 시점에서 염기서열화되었던 한 세트의 유자격 표본 (유사한 서열의)으로부터 획득된 서열과 관련시킬 수 있다. 일정한 구체예에서, IPC는 특정 암-관련된 좌위에 대해 개체로부터 획득된 서열 정보를 한 세트의 유자격 표본으로부터 (가령, 공지된 증폭/결실, 기타 등등으로부터) 획득된 서열 정보와 관련시킬 수 있다.

[0318] 이에 더하여, IPC는 염기서열결정 과정을 통해 표본(들)을 추적하는 마커로서 이용될 수 있다. IPC는 또한, 관심되는 염색체의 하나 또는 그 이상의 이수성, 예를 들면, 삼염색체성 21, 삼염색체성 13, 삼염색체성 18에 대한 정성적 양성 서열 도스 값, 예를 들면, NCV를 제공하여, 적절한 해석을 제공하고 데이터의 의존성과 정확도를 담보할 수 있다. 일정한 구체예에서, IPC는 태아가 수컷인지를 결정하기 위한 모계 표본 내에 염색체 X와 Y에 대한 도스를 제공하기 위해, 수컷과 암컷 유전체로부터 핵산을 포함하도록 창출될 수 있다.

[0319] 과정중 대조의 유형과 숫자는 필요한 시험의 유형 또는 성격에 의존한다. 가령, 염색체 이수성이 존재하는 지를 결정하기 위해 유전체의 혼합물을 포함하는 표본으로부터 DNA의 염기서열결정을 필요로 하는 시험의 경우에, 과정중 대조는 시험되는 동일한 염색체 이수성을 포함하는 공지된 표본으로부터 획득된 DNA를 포함할 수 있다. 일부 구체예에서, IPC는 관심되는 염색체의 이수성을 포함하는 것으로 알려진 표본으로부터 DNA를 포함한다. 가령, 모계 표본 내에 태아 삼염색체성, 예를 들면, 삼염색체성 21의 존재 또는 부재를 결정하는 시험에 대한 IPC는 삼염색체성 21을 갖는 개체로부터 획득된 DNA를 포함한다. 일부 구체예에서, IPC는 상이한 이수성을 갖는 2명 또는 그 이상의 개체로부터 획득된 DNA의 혼합물을 포함한다. 가령, 삼염색체성 13, 삼염색체성 18, 삼염색체성 21, 그리고 일염색체성 X의 존재 또는 부재를 결정하기 위한 시험의 경우에, IPC는 시험되는 삼염색체성 중에서 한 가지를 갖는 태아를 각각 잉태하는 임신 여성으로부터 획득된 DNA 표본의 조합을 포함한다. 완전한 염색체 이수성에 더하여, IPC는 부분적인 이수성의 존재 또는 부재를 결정하는 시험에 대한 양성 대조를 제공하

도록 창출될 수 있다.

[0320] 단일 이수성을 검출하기 위한 대조로서 역할을 하는 IPC는 2명의 개체 (이들 중에서 한 명은 이수체 유전체의 기여자이다)로부터 획득된 세포 유전체 DNA의 혼합물을 이용하여 창출될 수 있다. 가령, 태아 삼염색체성, 예를 들면, 삼염색체성 21을 결정하는 시험에 대한 대조로서 창출되는 IPC는 삼염색체성 염색체를 보유하는 수컷 또는 암컷 개체로부터 유전체 DNA를 삼염색체성 염색체를 보유하지 않는 것으로 알려진 암컷 개체로부터 유전체 DNA와 결합함으로써 창출될 수 있다. 유전체 DNA는 양쪽 개체의 세포로부터 추출되고, 그리고 모계 표본 내에 순환하는 cfDNA 단편을 모의하기 위한 약 100 내지 400 bp, 약 150 내지 350 bp, 또는 약 200 내지 300 bp의 단편이 제공되도록 전단될 수 있다. 이수성, 예를 들면, 삼염색체성 21을 보유하는 개체로부터 단편화된 DNA의 비율은 모계 표본 내에 발견되는 순환하는 태아 cfDNA의 비율을 모의하여, 이수성을 보유하는 개체로부터 DNA의 약 5%, 약 10%, 약 15%, 약 20%, 약 25%, 약 30%를 포함하는 단편화된 DNA의 혼합물을 포함하는 IPC가 제공되도록 선택된다. IPC는 상이한 이수성을 각각 보유하는 상이한 개체로부터 DNA를 포함할 수 있다. 가령, IPC는 약 80%의 영향을 받지 않은 암컷 DNA를 포함할 수 있고, 그리고 나머지 20%는 삼염색체성 염색체 21, 삼염색체성 염색체 13, 그리고 삼염색체성 염색체 18을 각각 보유하는 3명의 상이한 개체로부터 DNA일 수 있다. 단편화된 DNA의 혼합물은 염기서열결정을 위해 준비된다. 단편화된 DNA의 혼합물의 처리는 염기서열결정 라이브러리를 제조하는 것을 포함할 수 있고, 이것은 단일플렉스 또는 멀티플렉스 방식으로 임의의 대량으로 병렬 방법을 이용하여 염기서열결정될 수 있다. 유전체 IPC의 원액은 저장되고 복수 진단적 시험에서 이용될 수 있다.

[0321] 대안으로, IPC는 공지된 염색체 이수성을 갖는 태아를 잉태하는 것으로 알려진 모체로부터 획득된 cfDNA를 이용하여 창출될 수 있다. 가령, cfDNA는 삼염색체성 21을 갖는 태아를 잉태하는 임신 여성으로부터 획득될 수 있다. cfDNA는 모계 표본으로부터 도출되고, 그리고 세균 벡터 내로 클로닝되고 세균에서 성장되어 IPC의 진행 중인 공급원을 제공한다. DNA는 제한 효소를 이용하여 세균 벡터로부터 추출될 수 있다. 대안으로, 클로닝된 cfDNA는 예로서, PCR에 의해 증폭될 수 있다. IPC DNA는 염색체 이수성의 존재 또는 부재에 대해 분석되는 시험 표본으로부터 cfDNA와 동일한 실행에서 염기서열결정을 위해 처리될 수 있다.

[0322] IPC의 창출이 삼염색체성에 대하여 전술되긴 하지만, 예로서 다양한 분절 증폭 및/또는 결실을 포함하는 다른 부분적인 이수성을 반영하는 IPC가 창출될 수 있는 것으로 인지될 것이다. 따라서, 예로서 다양한 암이 특정 증폭과 연관되는 것으로 알려져 있는 경우에 (가령, 20Q13과 연관된 유방암), 이들 공지된 증폭을 통합하는 IPC가 창출될 수 있다.

[0323] **염기서열결정 방법**

[0324] 앞서 지시된 바와 같이, 제조된 표본 (가령, 염기서열결정 라이브러리)은 사본수 변동(들)을 확인하기 위한 절차의 일부로서 염기서열결정된다. 다수의 염기서열결정 기술 중에서 한 가지가 활용될 수 있다.

[0325] 일부 염기서열결정 기술, 예를 들면, Affymetrix Inc. (Sunnyvale, CA)로부터 혼성화에 의한 염기서열결정 플랫폼 및 454 Life Sciences (Bradford, CT), Illumina/Solexa (Hayward, CA) 및 Helicos Biosciences (Cambridge, MA)로부터 합성에 의한 염기서열결정 플랫폼, 그리고 아래에 설명된 바와 같이, Applied Biosystems (Foster City, CA)로부터 결찰에 의한 염기서열결정 플랫폼이 상업적으로 가용하다. Helicos Biosciences의 합성에 의한 염기서열결정을 이용하여 수행된 단일 분자 염기서열결정에 더하여, 다른 단일 분자 염기서열결정 기술은 Pacific Biosciences의 SMRT™ 기술, ION TORRENT™ 기술, 그리고 예로서, Oxford Nanopore Technologies에 의해 개발된 나노구멍 염기서열결정을 포함하지만 이들에 한정되지 않는다.

[0326] 자동화된 생어 방법이 '1세대' 기술로서 고려되긴 하지만, 자동화된 생어 염기서열결정을 포함하는 생어 염기서열결정 역시 본원에서 설명된 방법에서 이용될 수 있다. 추가 적합한 염기서열결정 방법은 핵산 영상 기술, 예를 들면, 원자력 현미경검사 (AFM) 또는 전파 전자 현미경검사 (TEM)를 포함하지만 이들에 한정되지 않는다. 예시적인 염기서열결정 기술은 아래에 더욱 상세하게 설명된다.

[0327] 한 예시적인, 하지만 무제한적 구체예에서, 본원에서 설명된 방법은 Helicos의 단일 분자 염기서열결정 기술, 진정한 단일 분자 염기서열결정 (tSMS) 기술을 이용하여 시험 표본 내에 핵산, 예를 들면, 모계 표본 내에 cfDNA, 암에 대해 선별검사되는 개체에서 cfDNA 또는 세포 DNA, 기타 등등에 대한 서열 정보를 획득하는 것을 포함한다 (가령, Harris T.D. et al., Science 320:106-109 [2008]에서 설명된 바와 같이). tSMS 기술에서, DNA 표본은 대략 100 내지 200개 뉴클레오티드의 가닥으로 개열되고, 그리고 polyA 서열이 각 DNA 가닥의 3' 단부에 부가된다. 각 가닥은 형광으로 표지화된 아데노신 뉴클레오티드의 부가에 의해 표지화된다. DNA 가닥은 이후, 흐름 셀에 혼성화되는데, 이것은 흐름 셀 표면에 고정되는 수백만 개의 올리고-T 포획 부위를 내포한다. 일

정한 구체예에서, 주형은 약 100 백만 주형/cm2의 밀도에 있을 수 있다. 흐름 셀은 이후, 기기, 예를 들면, HeliScope™ 서열분석기 내로 적하되고, 그리고 레이저가 흐름 셀의 표면을 조명하고, 각 주형의 위치를 드러낸다. CCD 카메라가 흐름 셀 표면 상에서 주형의 위치를 매핑할 수 있다. 주형 형광 라벨은 이후, 개열되고 씻겨 내려간다. 염기서열결정 반응은 DNA 중합효소 및 형광으로 표지화된 뉴클레오티드를 도입함으로써 시작된다. 올리고-T 핵산은 프라이머로서 역할을 한다. 중합효소는 표지화된 뉴클레오티드를 주형 주도된 방식으로 프라이머에 통합한다. 중합효소 및 통합되지 않은 뉴클레오티드는 제거된다. 형광으로 표지화된 뉴클레오티드의 통합을 주도한 주형은 흐름 셀 표면을 영상함으로써 구별된다. 영상한 후에, 개열 단계는 형광 라벨을 제거하고, 그리고 이러한 과정은 원하는 리드 길이가 달성될 때까지 다른 형광으로 표지화된 뉴클레오티드로 반복된다. 서열 정보는 각 뉴클레오티드 부가 단계에서 수집된다. 단일 분자 염기서열결정 기술에 의한 전체 유전체 염기서열결정은 염기서열결정 라이브러리의 제조에서 PCR-기초된 증폭을 배제하거나 또는 전형적으로 배제하고, 그리고 이들 방법은 표본의 사본의 계측보다는 표본의 직접적인 계측을 허용한다.

[0328] 다른 예시적인, 하지만 무제한적 구체예에서, 본원에서 설명된 방법은 454 염기서열결정 (Roche)을 이용하여, 시험 표본 내에 핵산, 예를 들면, 모계 시험 표본 내에 cfDNA, 암에 대해 선별검사되는 개체에서 cfDNA 또는 세포 DNA, 기타 등등에 대한 서열 정보를 획득하는 것을 포함한다 (가령, Margulies, M. et al. Nature 437:376-380 [2005]에서 설명된 바와 같이). 454 염기서열결정은 전형적으로, 2 단계를 수반한다. 첫 번째 단계에서, DNA는 대략 300-800개 염기쌍의 단편으로 전단되고, 그리고 이들 단편은 평활 말단이다. 올리고뉴클레오티드 어댑터가 이후, 이들 단편의 단부에 결찰된다. 이들 어댑터는 단편의 증폭과 염기서열결정을 위한 프라이머로서 역할을 한다. 단편은 예로서, 5'-비오틴 태그를 내포하는 어댑터 B를 이용하여 DNA 포획 비드, 예를 들면, 스트렙타비딘-코팅된 비드에 부착될 수 있다. 비드에 부착된 단편은 오일-물 유제의 비말 내에서 PCR 증폭된다. 결과는 각 비드 상에서 클론에 의해 증폭된 DNA 단편의 복수 사본이다. 두 번째 단계에서, 비드는 웰 (가령, 피코리터-크기산정된 웰)에서 포획된다. 파이로시퀀싱이 각 DNA 단편에서 병렬적으로 수행된다. 하나 또는 그 이상의 뉴클레오티드의 부가는 염기서열결정 기기에서 CCD 카메라에 의해 기록되는 광 신호를 산출한다. 신호 강도는 통합된 뉴클레오티드의 숫자에 비례한다. 파이로시퀀싱은 뉴클레오티드 부가 시에 방출되는 피로인산염 (PPi)을 이용한다. PPi는 아데노신 5' 포스포황산염의 존재에서 ATP 술푸릴라아제에 의해 ATP로 전환된다. 루시페라아제는 ATP를 이용하여 루시페린을 옥시루시페린으로 전환시키고, 그리고 이러한 반응은 계측되고 분석되는 광을 산출한다.

[0329] 다른 예시적인, 하지만 무제한적 구체예에서, 본원에서 설명된 방법은 SOLiD™ 기술 (Applied Biosystems)을 이용하여, 시험 표본 내에 핵산, 예를 들면, 모계 시험 표본 내에 cfDNA, 암에 대해 선별검사되는 개체에서 cfDNA 또는 세포 DNA, 기타 등등에 대한 서열 정보를 획득하는 것을 포함한다. SOLiD™ 결찰에 의한 염기서열결정에서, 유전체 DNA는 단편으로 전단되고, 그리고 어댑터가 단편 라이브러리를 산출하기 위해 단편의 5'와 3' 단부에 부착된다. 대안으로, 내부 어댑터가 어댑터를 단편의 5'와 3' 단부에 결찰하고, 이들 단편을 원형으로 만들고, 원형으로 만들어진 단편을 소화하여 내부 어댑터를 산출하고, 그리고 어댑터를 결과의 단편의 5' 와 3' 단부에 부착하여 메이트-짝짓기된 라이브러리를 산출함으로써 도입될 수 있다. 그 다음, 클론 비드 개체군이 비드, 프라이머, 주형, 그리고 PCR 성분을 내포하는 마이크로반응기에서 제조된다. PCR 이후에, 주형이 변성되고 비드가 농축되어 연장된 주형을 갖는 비드가 분리된다. 선별된 비드 상에서 주형은 유리 슬라이드에 결합을 허용하는 3' 변형에 종속된다. 서열은 순차적 혼성화, 그리고 특정한 형광단에 의해 확인되는 중심 결정된 염기 (또는 염기의 쌍)로 부분적으로 무작위 올리고뉴클레오티드의 결찰에 의해 결정될 수 있다. 칼라가 기록된 후에, 결찰된 올리고뉴클레오티드는 개열되고 제거되고, 그리고 이러한 과정은 이후 반복된다.

[0330] 다른 예시적인, 하지만 무제한적 구체예에서, 본원에서 설명된 방법은 Pacific Biosciences의 단일 분자, 실시간 (SMRT™) 염기서열결정 기술을 이용하여, 시험 표본 내에 핵산, 예를 들면, 모계 시험 표본 내에 cfDNA, 암에 대해 선별검사되는 개체에서 cfDNA 또는 세포 DNA, 기타 등등에 대한 서열 정보를 획득하는 것을 포함한다. SMRT 염기서열결정에서, 염료-표지화된 뉴클레오티드의 연속적 통합이 DNA 합성 동안 영상화된다. 단일 DNA 중합효소 분자는 서열 정보를 획득하는 개별 제로-방식 파장 검출기 (ZMW 검출기)의 아래쪽 표면에 부착되고, 반면 인산기에 의해 연결된 뉴클레오티드는 성장 프라이머 가닥 내로 통합된다. ZMW 검출기는 ZMW의 외부에서 급속히 확산하는 (가령, 마이크로초 내에) 형광 뉴클레오티드의 배경에 대하여 DNA 중합효소에 의한 단일 뉴클레오티드의 통합을 관찰할 수 있게 하는 밀폐 구조를 포함한다. 뉴클레오티드를 성장 가닥 내로 통합하는 것은 전형적으로, 수 밀리초가 소요된다. 이러한 시간 동안, 형광 라벨이 여기되고 형광 신호를 발생시키고, 그리고 형광 태그가 쪼개진다. 염료의 상응하는 형광의 계측은 염기가 통합되었다는 것을 지시한다. 이러한 과정은 서열을 제공하기 위해 반복된다.

[0331] 다른 예시적인, 하지만 무제한적 구체예에서, 본원에서 설명된 방법은 나노구멍 염기서열결정을 이용하여, 시험 표본 내에 핵산, 예를 들면, 모계 시험 표본 내에 cfDNA, 암에 대해 선별검사되는 개체에서 cfDNA 또는 세포 DNA, 기타 등등에 대한 서열 정보를 획득하는 것을 포함한다 (가령, Soni GV and Meller A. Clin Chem 53: 1996-2001 [2007]에서 설명된 바와 같이). 나노구멍 염기서열결정 DNA 분석 기술은 예로서, Oxford Nanopore Technologies (Oxford, United Kingdom), Sequenom, NABsys 등을 비롯한 다수의 기업에 의해 개발된다. 나노구멍 염기서열결정은 단일-분자 염기서열결정 기술인데, 여기서 DNA의 단일 분자가 나노구멍을 통과하는 동안 직접적으로 염기서열결정된다. 나노구멍은 전형적으로, 직경에서 1 나노미터의 차수의 작은 구멍이다. 전도성 유체에 나노구멍의 담금 및 이것 전체에 대하여 전위 (전압)의 적용은 나노구멍을 통한 이온의 전도로 인한 약간 전류를 유발한다. 흘러가는 전류의 양은 나노구멍의 크기와 모양에 민감하다. DNA 분자가 나노구멍을 통과할 때, DNA 분자 상에 각 뉴클레오티드는 나노구멍을 상이한 정도로 방해하고, 나노구멍을 통한 전류의 크기를 상이한 정도에서 변화시킨다. 따라서, DNA 분자가 나노구멍을 통과할 때 전류에서 이러한 변화는 DNA 서열의 리드를 제공한다.

[0332] 다른 예시적인, 하지만 무제한적 구체예에서, 본원에서 설명된 방법은 화학적-민감성 필드 효과 트랜지스터 (chemFET) 어레이를 이용하여, 시험 표본 내에 핵산, 예를 들면, 모계 시험 표본 내에 cfDNA, 암에 대해 선별검사되는 개체에서 cfDNA 또는 세포 DNA, 기타 등등에 대한 서열 정보를 획득하는 것을 포함한다 (가령, U.S. 특허 출원 공개 번호 2009/0026082에서 설명된 바와 같이). 이러한 기술의 한 가지 실례에서, DNA 분자가 반응 챔버 내로 배치될 수 있고, 그리고 주형 분자가 중합효소에 결합된 염기서열결정 프라이머에 혼성화될 수 있다. 염기서열결정 프라이머의 3' 단부에서 새로운 핵산 가닥 내로 하나 또는 그 이상의 삼인산염의 통합은 chemFET에 의한 전류에서 변화로서 식별될 수 있다. 어레이는 복수 chemFET 센서를 가질 수 있다. 다른 실례에서, 단일 핵산이 비드에 부착될 수 있고, 그리고 이들 핵산은 비드 상에서 증폭될 수 있고, 그리고 개별 비드는 chemFET 어레이 상에서 개별 반응 챔버로 이전될 수 있고, 각 챔버는 chemFET 센서를 갖고, 그리고 이들 핵산은 염기서열결정될 수 있다.

[0333] 다른 구체예에서, 본 발명 방법은 투과 전자 현미경검사 (TEM)를 이용하는 Halcyon Molecular의 기술을 이용하여, 시험 표본 내에 핵산, 예를 들면, 모계 시험 표본 내에 cfDNA에 대한 서열 정보를 획득하는 것을 포함한다. 개별 분자 배치 급속 나노 전달 (IMPRNT)로 명명된 상기 방법은 중원자 마커로 선별적으로 표지화된 높은-분자량 (150kb 또는 그 이상) DNA의 단일 원자 분해능 투과 전자 현미경 영상을 활용하고, 그리고 일관된 염기간 이격을 갖는 초밀집 (3nm 가닥간) 병렬 어레이에서 초박막 상에 이들 분자를 배열하는 것을 포함한다. 전자 현미경은 분자를 필름 상에 영상하여, 중원자 마커의 위치를 결정하고 DNA로부터 염기 서열 정보를 도출하는데 이용된다. 상기 방법은 PCT 특허 공개 WO 2009/046445에서 더욱 설명된다. 상기 방법은 완전한 인간 유전체를 10 분 이내에 염기서열결정하는 것을 허용한다.

[0334] 다른 구체예에서, DNA 염기서열결정 기술은 이온 토렌트 단일 분자 염기서열결정인데, 이것은 반도체 기술을 단순 염기서열결정 화학과 짝지어, 화학적으로 인코딩된 정보 (A, C, G, T)를 반도체 칩 상에 디지털 정보 (0, 1)로 직접적으로 번역한다. 자연에서, 뉴클레오티드가 중합효소에 의해 DNA의 가닥 내로 통합될 때, 수소 이온이 부산물로서 방출된다. 이온 토렌트는 마이크로-절삭된 웰의 고밀도 어레이를 이용하여, 이러한 생화학적 과정을 대량으로 병렬 방식으로 수행한다. 각 웰은 상이한 DNA 분자를 유지한다. 웰 아래에 이온-민감성 층이 있고, 그리고 그 아래에 이온 센서가 있다. 뉴클레오티드, 예를 들면, C가 DNA 주형에 부가되고, 그리고 이후, DNA의 가닥 내로 통합될 때, 수소 이온이 방출될 것이다. 상기 이온으로부터 전하는 용액의 pH를 변화시킬 것이고, 이것은 이온 토렌트의 이온 센서에 의해 검출될 수 있다. 이러한 서열분석기 - 본질적으로 세계에서 가장 작은 고체-상태 pH 측정기 -는 염기를 호출하고, 화학적 정보에서 디지털 정보로 직접적으로 진행할 것이다. 이온 개인 유전체 기계 (PGM™) 서열분석기는 이후, 뉴클레오티드를 연달아 칩에 순차적으로 쏟아 붇는다. 칩에 쏟아 부어지는 다음 뉴클레오티드가 정합이 아니면, 어떤 전압 변화도 기록되지 않을 것이고 어떤 염기도 호출되지 않을 것이다. DNA 가닥 상에 2개의 동일한 염기가 있으면, 전압은 2배가 될 것이고, 그리고 칩은 호출된 2개의 동일한 염기를 기록할 것이다. 직접적인 검출은 수초 내에 뉴클레오티드 통합의 기록을 허용한다.

[0335] 다른 구체예에서, 본 발명 방법은 혼성화에 의한 염기서열결정을 이용하여, 시험 표본 내에 핵산, 예를 들면, 모계 시험 표본 내에 cfDNA에 대한 서열 정보를 획득하는 것을 포함한다. 혼성화에 의한 염기서열결정은 복수의 폴리뉴클레오티드 서열을 복수의 폴리뉴클레오티드 프로브와 접촉시키는 것을 포함하고, 여기서 복수의 폴리뉴클레오티드 프로브 각각은 기질에 임의선택적으로 묶일 수 있다. 기질은 공지된 뉴클레오티드 서열의 어레이를 포함하는 편평한 표면일지도 모른다. 어레이에 혼성화의 패턴은 표본 내에 존재하는 폴리뉴클레오티드 서열을 결정하는데 이용될 수 있다. 다른 구체예에서, 각 프로브는 비드, 예를 들면, 자성 비드 또는 기타 유사한 것에

묶인다. 비드에 혼성화가 결정되고, 그리고 표본 내에서 복수의 폴리뉴클레오티드 서열을 확인하는데 이용될 수 있다.

[0336] 다른 구체예에서, 본 발명 방법은 Illumina의 합성에 의한 염기서열결정 및 가역성 종결인자-기초된 염기서열결정 화학을 이용한 수백만 개 DNA 단편의 대량으로 병렬 염기서열결정에 의해, 시험 표본 내에 핵산, 예를 들면, 모계 시험 표본 내에 cfDNA에 대한 서열 정보를 획득하는 것을 포함한다 (가령, Bentley et al., Nature 6:53-59 [2009]에서 설명된 바와 같이). 주형 DNA는 유전체 DNA, 예를 들면, cfDNA일 수 있다. 일부 구체예에서, 단리된 세포로부터 유전체 DNA가 주형으로서 이용되고, 그리고 수백 개 염기쌍의 길이로 단편화된다. 다른 구체예에서, cfDNA가 주형으로서 이용되고, 그리고 단편화가 필요하지 않은데, 그 이유는 cfDNA가 짧은 단편으로서 존재하기 때문이다. 가령, 태아 cfDNA는 혈류 내에 대략 170 염기쌍 (bp) 길이의 단편으로 순환하고 (Fan et al., Clin Chem 56:1279-1286 [2010]), 그리고 염기서열결정에 앞서, DNA의 단편화가 필요하지 않다. Illumina의 염기서열결정 기술은 올리고뉴클레오티드 앵커가 결합되는 평면의 광학적으로 투명한 표면에 단편화된 유전체 DNA의 부착에 의존한다. 주형 DNA는 단부 수복되어 5'-인산화된 평활 말단이 산출되고, 그리고 클레노브 단편의 중합효소 활성이 단일 A 염기를 평활 인산화된 DNA 단편의 3' 단부에 부가하는데 이용된다. 이러한 부가는 결찰 효율을 증가시키기 위해 3' 단부에서 단일 T 염기의 오버행을 갖는 올리고뉴클레오티드 어댑터에 결찰을 위한 DNA 단편을 제조한다. 어댑터 올리고뉴클레오티드는 흐름 셀 앵커에 상보적이다. 제한-희석 조건 하에, 어댑터-변형된, 단일 가닥 주형 DNA가 흐름 셀에 첨가되고 앵커에 혼성화에 의해 고정된다. 부착된 DNA 단편은 연장되고 가교 증폭되어, 동일한 주형의 ~1,000개 사본을 각각 내포하는 수억 개의 클러스터를 갖는 초고밀도 염기서열결정 흐름 셀이 창출된다. 한 구체예에서, 무작위로 단편화된 유전체 DNA, 예를 들면, cfDNA는 클러스터 증폭에 종속되기 전에, PCR을 이용하여 증폭된다. 대안으로, 증폭-없는 유전체 라이브러리 제조가 이용되고, 그리고 무작위로 단편화된 유전체 DNA, 예를 들면, cfDNA는 클러스터 증폭 단독을 이용하여 농축된다 (Kozarewa et al., Nature Methods 6:291-295 [2009]). 주형은 제거가능한 형광 염료와 함께 가역성 종결인자를 이용하는 견실한 4-칼라 DNA 합성에 의한 염기서열결정 기술을 이용하여 염기서열결정된다. 높은-감수성 형광 검출은 레이저 여기 및 전체 내부 반사 광학을 이용하여 달성된다. 약 20-40 bp, 예를 들면, 36 bp의 짧은 서열 리드가 반복-마스크 적용된 참조 유전체에 대항하여 정렬되고, 그리고 참조 유전체에 짧은 서열 리드의 독특한 매핑은 특별히 개발된 데이터 분석 파이프라인 소프트웨어를 이용하여 확인된다. 비-반복-마스크 적용된 참조 유전체 또한 이용될 수 있다. 반복-마스크 적용된 또는 비-반복-마스크 적용된 참조 유전체가 이용되는 지에 상관없이, 참조 유전체에 독특하게 매핑하는 리드만 계수된다. 첫 번째 리드의 완결 후, 주형은 단편의 반대쪽으로부터 두 번째 리드를 할 수 있게 하기 위해 원지 재건될 수 있다. 따라서, DNA 단편의 싱글 엔드 또는 페어드 엔드 염기서열결정이 이용될 수 있다. 표본 내에 존재하는 DNA 단편의 부분적인 염기서열결정이 수행되고, 그리고 미리 결정된 길이, 예를 들면, 36 bp의 리드를 포함하는 서열 태그가 공지된 참조 유전체에 매핑되고 계수된다. 한 구체예에서, 참조 유전체 서열은 NCBI36/hg18 서열인데, 이것은 월드와이드웹 상에서 genome.ucsc.edu/cgi-bin/hgGateway?org=Human&db=hg18&hgsid=166260105)에서 가용하다. 대안으로, 참조 유전체 서열은 GRCh37/hg19인데, 이것은 월드와이드웹 상에서 genome.ucsc.edu/cgi-bin/hgGateway에서 가용하다. 공개 서열 정보의 다른 출처는 GenBank, dbEST, dbSTS, EMBL (the European Molecular Biology Laboratory), 그리고 the DDBJ (the DNA Databank of Japan)를 포함한다. 제한 없이, BLAST (Altschul et al., 1990), BLITZ (MPsrch) (Sturrock & Collins, 1993), FASTA (Person & Lipman, 1988), BOWTIE (Langmead et al., Genome Biology 10:R25.1-R25.10 [2009]), 또는 ELAND (Illumina, Inc., San Diego, CA, USA)를 비롯한 다수의 컴퓨터 알고리즘이 서열을 정렬하는데 가용하다. 한 구체예에서, 혈장 cfDNA 분자의 클론에 의해 확대된 사본의 한쪽 단부가 염기서열결정되고, 그리고 뉴클레오티드 데이터베이스의 효율적인 대규모 정렬 (ELAND) 소프트웨어를 이용하는 Illumina 유전체 분석기에 대한 생물정보학 정렬 분석에 의해 처리된다.

[0337] 본원에서 설명된 방법의 일부 구체예에서, 매핑된 서열 태그는 약 20bp, 약 25bp, 약 30bp, 약 35bp, 약 40bp, 약 45bp, 약 50bp, 약 55bp, 약 60bp, 약 65bp, 약 70bp, 약 75bp, 약 80bp, 약 85bp, 약, 약 95bp, 약 100bp, 약 110bp, 약 120bp, 약 130, 약 140bp, 약 150bp, 약 200bp, 약 250bp, 약 300bp, 약 350bp, 약 400bp, 약 450bp, 또는 약 500bp의 서열 리드를 포함한다. 기술적인 진전은 페어드 엔드 리드가 산출될 때 약 1000bp보다 큰 리드를 실시가능하게 하는 500bp보다 큰 싱글 엔드 리드를 가능하게 할 것으로 것으로 예상된다. 한 구체예에서, 매핑된 서열 태그는 36bp인 서열 리드를 포함한다. 서열 태그의 매핑은 태그의 서열을 참조의 서열과 비교하여 염기서열결정된 핵산 (가령, cfDNA) 분자의 염색체 기원을 결정함으로써 달성되고, 그리고 특이적 유전자 서열 정보가 필요하지 않다. 작은 정도의 부정합 (서열 태그마다 0-2개 부정합)은 참조 유전체 및 혼합된 표본 내에 유전체 사이에 존재할 수 있는 가벼운 다형성을 설명하기 위해 허용될 수 있다.

[0338] 표본당 복수의 서열 태그가 전형적으로 획득된다. 일부 구체예에서, 표본당20 내지 40bp 리드, 예를 들면, 36bp

를 포함하는 최소한 약 3 x 10^6 서열 태그, 최소한 약 5 x 10^6 서열 태그, 최소한 약 8 x 10^6 서열 태그, 최소한 약 10 x 10^6 서열 태그, 최소한 약 15 x 10^6 서열 태그, 최소한 약 20 x 10^6 서열 태그, 최소한 약 30 x 10^6 서열 태그, 최소한 약 40 x 10^6 서열 태그, 또는 최소한 약 50 x 10^6 서열 태그가 이들 리드를 참조 유전체에 매핑하는 것으로부터 획득된다. 한 구체예에서, 모든 서열 리드가 참조 유전체의 모든 영역에 매핑된다. 한 구체예에서, 참조 유전체의 모든 영역, 예를 들면, 모든 염색체에 매핑된 태그가 계수되고, 그리고 CNV, 다시 말하면, 혼합된 DNA 표본 내에 관심되는 서열, 예를 들면, 염색체 또는 이의 부분의 과다표현 또는 과소표현이 결정된다. 상기 방법은 두 유전체 사이에 구별을 필요로 하지 않는다.

[0339] 표본 내에 CNV, 예를 들면, 이수성이 존재하거나 또는 부재하는 지를 정확하게 결정하는데 필요한 정확도는 염기서열결정 실행 이내에 표본 사이에서 참조 유전체에 매핑하는 서열 태그의 개수의 변동 (염색체간 가변성), 그리고 상이한 염기서열결정 실행에서 참조 유전체에 매핑하는 서열 태그의 개수의 변동 (염기서열결정간 가변성)에 입각된다. 가령, 변동은 GC-풍부한 또는 GC-불량한 참조 서열에 매핑하는 태그에 대해 특히 확연할 수 있다. 다른 변동은 핵산의 추출과 정제, 염기서열결정 라이브러리의 제조, 그리고 상이한 염기서열결정 플랫폼의 이용을 위한 상이한 프로토콜을 이용하는 것으로부터 발생할 수 있다. 본 발명 방법은 염색체간 (실행내), 그리고 염기서열결정간 (실행간) 및 플랫폼-의존성 가변성에 기인하는 발생된 가변성을 내재적으로 설명하기 위해, 정규화 서열 (정규화 염색체 서열 또는 정규화 분절 서열)의 지식에 근거된 서열 도스 (염색체 도스, 또는 분절 도스)를 이용한다. 염색체 도스는 정규화 염색체 서열의 지식에 근거되는데, 상기 서열은 단일 염색체, 또는 염색체 1-22, X와 Y에서 선택되는 2개 또는 그 이상의 염색체로 구성될 수 있다. 대안으로, 정규화 염색체 서열은 단일 염색체 분절, 또는 한 염색체 또는 2개 또는 그 이상 염색체의 2개 또는 그 이상 분절로 구성될 수 있다. 분절 도스는 정규화 분절 서열의 지식에 근거되는데, 상기 서열은 임의의 한 가지 염색체의 단일 분절, 또는 염색체 1-22, X와 Y 중에서 임의의 2개 또는 그 이상의 2개 또는 그 이상 분절로 구성될 수 있다.

[0340] **CNV 및 출생전 진단**

[0341] 모체 혈액 내에 순환하는 무세포 태아 DNA와 RNA가 임신 관리 및 생식 의사 결정의 보조 둘 모두를 위해, 증가하는 숫자의 유전적 장애의 초기 비침습성 출생전 진단 (NIPD)에 이용될 수 있다. 혈류 내에 순환하는 무세포 DNA의 존재는 50 여 년 동안 알려져 있었다. 더욱 최근에, 소량의 순환하는 태아 DNA의 존재가 임신 동안 모체 혈액에서 발견되었다 (Lo et al., Lancet 350:485-487 [1997]). 죽어가는 태반 세포로부터 기원하는 것으로 생각되는 무세포 태아 DNA (cfDNA)는 전형적으로 길이에서 200 bp보다 적은 짧은 단편으로 구성되는 것으로 나타났는데 (Chan et al., Clin Chem 50:88-92 [2004]), 이들은 빠르게는 임신 4주차에 구별될 수 있고 (Illanes et al., Early Human Dev 83:563-566 [2007]), 그리고 전달의 수시간 이내에 모계 순환으로부터 소실되는 것으로 알려져 있다 (Lo et al., Am J Hum Genet 64:218-224 [1999]). cfDNA에 더하여, 무세포 태아 RNA (cfRNA)의 단편 역시 모체 혈액에서 구별될 수 있고, 태아 또는 태반에서 전사되는 유전자로부터 유래한다. 모체 혈액 표본으로부터 이들 태아 유전자 요소의 추출과 차후 분석은 NIPD에 대한 신규한 기회를 제공한다.

[0342] 본 발명 방법은 NIPD에서 이용을 위한 다형성-독립된 방법이고, 그리고 태아 cfDNA가 태아 이수성의 결정을 할 수 있게 하기 위해 모계 cfDNA로부터 구별될 것을 요구하지 않는다. 일부 구체예에서, 이수성은 완전한 염색체 삼염색체성 또는 일염색체성, 또는 부분적인 삼염색체성 또는 일염색체성이다. 부분적인 이수성은 염색체의 부분의 상실 또는 획득에 의해 유발되고, 그리고 불균형 전위, 불균형 전도, 결실 및 삽입으로부터 발생하는 염색체 불균형을 포괄한다. 지금까지, 수명과 양립성인 가장 흔한 공지된 이수성은 삼염색체성 21, 다시 말하면, 다운 증후군 (DS)인데, 이것은 염색체 21의 부분 또는 전부의 존재에 의해 유발된다. 드물게, DS는 유전된 또는 산발성 결함에 의해 유발될 수 있는데, 여기서 염색체 21의 전부 또는 일부의 추가의 사본이 다른 염색체 (통상적으로 염색체 14)에 부착되어 단일 일탈적 염색체를 형성한다. DS는 지적 장애, 심각한 학습 어려움 및 장기간 건강 문제, 예를 들면, 심장병에 의해 유발된 초과 사망률과 연관된다. 공지된 임상적 유의성을 갖는 다른 이수성은 에드워드 증후군 (삼염색체성 18) 및 파타우 증후군 (삼염색체성 13)을 포함하는데, 이들은 빈번하게, 수명의 첫 수 개월 이내에 치명적이다. 성염색체의 숫자와 연관된 비정상 역시 알려져 있고, 그리고 일염색체성 X, 예를 들면, 터너 증후군 (XO), 그리고 암컷 신생아에서 삼중 X 증후군 (XXX) 및 수컷 신생아에서 클라인펠터 증후군 (XXY)과 XYY 증후군을 포함하는데, 이들 모두 불임 및 지적 기량에서 감소를 비롯한 다양한 표현형과 연관된다. 일염색체성 X [45, X]는 자연 유산의 약 7%를 차지하는 초기 유산의 공통 원인이다. 1-2/10,000의 45,X (또한 터너 증후군으로 불림)의 생존 출산 빈도에 근거하여, 1% 보다 적은 45,X 수태가 임신 기간까지 생존할 것으로 추정된다. 약 30%의 터너 증후군 환자는 45,X 세포주 및 46,XX 세포주 또는 재배열된 X 염색체를 내포하는 것 중에서 어느 한 가지로 모자이크이다 (Hook and Warburton 1983). 생존 출산 영아에서 표현형은 높은 태

아 치사율을 고려하면 상대적으로 경미하고, 그리고 아마도, 터너 증후군을 앓는 모든 생존 출산 암컷이 2개의 성염색체를 내포하는 세포주를 보유하는 것으로 가정되었다. 일염색체성 X는 암컷에서 45,X 또는 45,X/46XX로서 일어날 수 있고, 그리고 수컷에서 45,X/46XY로서 일어날 수 있다. 인간에서 상염색체 일염색체성은 일반적으로, 수명과 양립하지 않는 것으로 제안된다; 하지만, 생존 출산 아동에서 한 염색체 21의 완전한 일염색체성을 설명하는 매우 많은 세포유전학적 보고가 있다 (Vosranova Iet al., Molecular Cytogen. 1:13 [2008]; Joosten et al., Prenatal Diagn. 17:271-5 [1997]. 본원에서 설명된 방법은 출생전에 이런 저런 염색체 비정상을 진단하는데 이용될 수 있다.

[0343] 일부 구체예에 따라, 본원에서 개시된 방법은 염색체 1-22, X와 Y 중에서 임의의 한 가지의 염색체 삼염색체성의 존재 또는 부재를 결정할 수 있다. 본 발명 방법에 따라 검출될 수 있는 염색체 삼염색체성의 실례는 제한없이, 삼염색체성 21 (T21; 다운 증후군), 삼염색체성 18 (T18; 에드워드 증후군), 삼염색체성 16 (T16), 삼염색체성 20 (T20), 삼염색체성 22 (T22; 고양이눈 증후군), 삼염색체성 15 (T15; 프레더 윌리 증후군), 삼염색체성 13 (T13; 파타우 증후군), 삼염색체성 8 (T8; 와르카니 증후군), 삼염색체성 9, 그리고 XXY (클라인펠터 증후군), XYY, 또는 XXX 삼염색체성을 포함한다. 비-모자이크 상태에서 존재하는 다른 상염색체의 완전한 삼염색체성은 치명적이지만, 모자이크 상태에서 존재할 때 수명과 양립할 수 있다. 다양한 완전한 삼염색체성 (모자이크 또는 비-모자이크 상태에서 존재하는 지에 상관없이), 그리고 부분적인 삼염색체성은 본원에서 제공된 교시에 따라, 태아 cfDNA에서 결정될 수 있는 것으로 인지될 것이다.

[0344] 본 발명 방법에 의해 결정될 수 있는 부분적인 삼염색체성의 무제한적 실례는 부분적인 삼염색체성 1q32-44, 삼염색체성 9 p, 삼염색체성 4 모자이크현상, 삼염색체성 17p, 부분적인 삼염색체성 4q26-qter, 부분적인 2p 삼염색체성, 부분적인 삼염색체성 1q, 및/또는 부분적인 삼염색체성 6p/일염색체성 6q를 포함하지만 이들에 한정되지 않는다.

[0345] 본원에서 개시된 방법은 또한, 염색체 일염색체성 X, 염색체 일염색체성 21, 그리고 부분적인 일염색체성, 예를 들면, 일염색체성 13, 일염색체성 15, 일염색체성 16, 일염색체성 21, 그리고 일염색체성 22를 결정하는데 이용될 수 있는데, 이들은 임신 유산에 관련되는 것으로 알려져 있다. 완전한 이수성에 전형적으로 관련된 염색체의 부분적인 일염색체성 역시 본원에서 설명된 방법에 의해 결정될 수 있다. 본 발명 방법에 따라 결정될 수 있는 결실 증후군의 무제한적 실례는 염색체의 부분적인 결실에 의해 유발된 증후군을 포함한다. 본원에서 설명된 방법에 따라 결정될 수 있는 부분적인 결실의 실례는 제한 없이, 다음에서 설명되는 염색체 1, 4, 5, 7, 11, 18, 15, 13, 17, 22와 10의 부분적인 결실을 포함한다.

[0346] 1q21.1 결실 증후군 또는 1q21.1 (재발성) 미세결실은 염색체 1의 희귀한 이상이다. 결실 증후군의 바로 옆에, 1q21.1 중복 증후군이 또한 있다. 결실 증후군에는 특정 스팟 상에서 DNA 결여의 부분이 있는 반면, 중복 증후군에는 동일한 스팟 상에서 DNA의 유사한 부분의 2개 또는 3개 사본이 있다. 기존 문헌은 결실 및 중복 둘 모두를 1q21.1 사본수 변동 (CNV)으로서 지칭한다. 1q21.1 결실은 타르 증후군 (요골 무형성을 동반한 혈소판감소증)과 연관될 수 있다.

[0347] 울프 허쉬호른 증후군 (WHS) (OMIN #194190)은 염색체 4p16.3의 반접합성 결실과 연관된 인접한 유전자 결실 증후군이다. 울프 허쉬호른 증후군은 출산전과 출산후 성장 결핍, 가변적 정도의 발달 장애, 특징적인 두개안면 특질 (코의 '그리스 전사 헬멧' 모습, 높은 이마, 두드러진 미간, 격리증, 높은 아지형 눈썹, 돌출되는 눈, 내안각 주름, 짧은 인중, 하강된 코너를 갖는 뚜렷한 입, 그리고 소하악증), 그리고 발작 장애에 의해 특징화되는 선천성 기형 증후군이다.

[0348] 5p- 또는 5p 마이너스로서 또한 알려져 있고, 그리고 고양이 울음 증후군 (OMIN#123450)으로 명명된 염색체 5의 부분적인 결실은 염색체 5의 짧은 팔 (p 팔)의 결실 (5p15.3-p15.2)에 의해 유발된다. 이러한 상태를 갖는 영아는 종종, 고양이의 울음처럼 들리는 고음의 울음을 갖는다. 상기 장애는 지적 장애 및 지연된 발달, 작은 머리 크기 (소두증), 낮은 출생 체중, 그리고 영아기에서 약한 근육 긴장도 (저혈압증), 특유한 안면 특질 및 아마도 심장 결함에 의해 특징화된다.

[0349] 염색체 7q11.23 결실 증후군 (OMIN 194050)으로서 또한 알려져 있는 윌리암스 보이렌 증후군은 대략 28개 유전자를 내포하는 염색체 7q11.23 상에서 1.5 내지 1.8 Mb의 반접합성 결실에 의해 유발된 다중시스템 장애를 유발하는 인접한 유전자 결실 증후군이다.

[0350] 11q 결실 장애로서 또한 알려져 있는 제이콥슨 증후군은 띠 11q24.1을 포함하는 염색체 11의 말단 영역의 결실로부터 발생하는 희귀한 선천성 장애이다. 이것은 지적 장애, 특유한 안면 모습, 그리고 심장 결함 및 출혈 장

애를 비롯한 다양한 신체적 문제를 유발할 수 있다.

[0351] 일염색체성 18p로서 알려져 있는 염색체 18의 부분적인 일염색체성은 염색체 18의 짧은 팔 (p)의 전부 또는 일부가 결실되는 희귀한 염색체 장애 (일염색체성)이다. 상기 장애는 전형적으로, 짧은 신장, 가변적 정도의 정신 지체, 언어 능력 지연, 두개골 및 안면 (두개안면) 영역의 기형, 및/또는 추가 신체 비정상에 의해 특징화된다. 연관된 두개안면 결함은 사례별로 범위 및 심각도에서 크게 변할 수 있다.

[0352] 염색체 15의 사본의 구조 또는 숫자에서 변화에 의해 유발된 질환은 안젤만 증후군 및 프래더-윌리 증후군을 포함하는데, 이들은 염색체 15의 동일한 부분, 15q11-q13 영역에서 유전자 활성의 상실을 수반한다. 여러 전위와 미세결실은 보유 부모에서 무증후성일 수 있지만, 자손에서 주요 유전성 질환을 유발할 수 있는 것으로 인지될 것이다. 가령, 15q11-q13 미세결실을 보유하는 건강한 모체는 심각한 신경변성 장애인 안젤만 증후군을 앓는 아동을 출산할 수 있다. 따라서, 본원에서 설명된 방법, 기구와 시스템은 태아에서 이런 부분적인 결실 및 다른 결실을 확인하는데 이용될 수 있다.

[0353] 부분적인 일염색체성 13q는 염색체 13의 긴 팔 (q)의 조각이 결여될 때 발생하는 희귀한 염색체 장애 (일염색체성)이다. 부분적인 일염색체성 13q를 갖고 태어난 영아는 낮은 출생 체중, 머리와 얼굴 (두개안면 영역)의 기형, 골격 비정상 (특히, 손발의), 그리고 다른 신체 비정상을 표시할 수 있다. 정신 지체는 이러한 질환의 특징이다. 영아기 동안 사망률은 이러한 장애를 갖고 태어난 개체 사이에서 높다. 부분적인 일염색체성 13q의 거의 모든 사례는 명확한 이유 없이 무작위로 발생한다 (산발성).

[0354] 스미스 마제니스 증후군 (SMS - OMIM #182290)은 염색체 17의 한 사본에서 유전 물질의 결실, 또는 상실에 의해 유발된다. 이러한 널리 공지된 증후군은 발달 지연, 정신 지체, 선천성 기형, 예를 들면, 심장과 신장 결함, 그리고 신경행동 비정상, 예를 들면, 심각한 수면 장애 및 자해 행동과 연관된다. 스미스 마제니스 증후군 (SMS)은 많은 경우에 (90%), 염색체 17p11.2에서 3.7-Mb 염색체내 결손에 의해 유발된다.

[0355] 디조지 증후군으로서 또한 알려져 있는 22q11.2 결실 증후군은 염색체 22의 작은 조각의 결실에 의해 유발된 증후군이다. 결실 (22 q11.2)은 한 쌍의 염색체 중에서 하나의 긴 팔 상에서 염색체의 중앙에 근접하게 발생한다. 이러한 증후군의 특질은 심지어 동일한 패밀리의 구성원 사이에서도 폭넓게 변하고, 그리고 많은 신체 부위에 영향을 준다. 특징적인 징후와 증상은 출생 결함, 예를 들면, 선천성 심장 질환, 폐쇄에서 신경근 문제와 가장 흔히 관련되는 구개에서 결함 (연인두폐쇄부전), 학습 장애, 안면 특질에서 경미한 차이, 그리고 재발성 감염을 포함할 수 있다. 염색체 영역 22q11.2에서 미세결실은 정신분열병의 20 내지 30-배 증가된 위험과 연관된다.

[0356] 염색체 10의 짧은 팔에서 결실은 디조지 증후군 유사 표현형과 연관된다. 염색체 10p의 부분적인 일염색체성은 희귀하지만, 디조지 증후군의 특질을 보여주는 환자의 부분에서 관찰되었다.

[0357] 한 구체예에서, 본원에서 설명된 방법, 기구, 그리고 시스템은 염색체 1, 4, 5, 7, 11, 18, 15, 13, 17, 22와 10의 부분적인 일염색체성, 예를 들면, 부분적인 일염색체성 1q21.11, 부분적인 일염색체성 4p16.3, 부분적인 일염색체성 5p15.3-p15.2, 부분적인 일염색체성 7q11.23, 부분적인 일염색체성 11q24.1, 부분적인 일염색체성 18p, 염색체 15의 부분적인 일염색체성 (15q11-q13), 부분적인 일염색체성 13q, 부분적인 일염색체성 17p11.2, 염색체 22의 부분적인 일염색체성 (22q11.2)을 포함하지만 이들에 한정되지 않는 부분적인 일염색체성을 결정하는데 이용되고, 그리고 부분적인 일염색체성 10p 또한 상기 방법을 이용하여 결정될 수 있다.

[0358] 본원에서 설명된 방법에 따라 결정될 수 있는 다른 부분적인 일염색체성은 불균형 전위 t(8;11)(p23.2;p15.5); 11q23 미세결실; 17p11.2 결실; 22q13.3 결실; Xp22.3 미세결실; 10p14 결실; 20p 미세결실, [del(22)(q11.2q11.23)], 7q11.23 및 7q36 결실; 1p36 결실; 2p 미세결실; 신경섬유종증 유형 1 (17q11.2 미세결실), Yq 결실; 4p16.3 미세결실; 1p36.2 미세결실; 11q14 결실; 19q13.2 미세결실; 루빈스타인 테이비 (16p13.3 미세결실); 7p21 미세결실; 밀러 디커 증후군 (17p13.3); 그리고 2q37 미세결실을 포함한다. 부분적인 결실은 염색체의 부분의 작은 결실일 수 있거나, 또는 이들은 단일 유전자의 결실이 일어날 수 있는 염색체의 미세결실일 수 있다.

[0359] 염색체 팔의 부분의 중복에 의해 유발된 여러 중복 증후군이 확인되었다 (OMIN [ncbi.nlm.nih.gov/omim에서 온라인 개관된 Online Mendelian Inheritance in Man] 참조). 한 구체예에서, 본 발명 방법은 염색체 1-22, X와 Y 중에서 임의의 한 가지의 분절의 중복 및/또는 증가의 존재 또는 부재를 결정하는데 이용될 수 있다. 본 발명 방법에 따라 결정될 수 있는 중복 증후군의 무제한적 실례는 다음에서 설명되는, 염색체 8, 15, 12와 17의 부분의 중복을 포함한다.

[0360] 8p23.1 중복 증후군은 인간 염색체 8로부터 영역의 중복에 의해 유발된 희귀한 유전 질환이다. 이러한 중복 증

후군은 64,000 신생아 중에서 1명의 추정된 유병률을 갖고 8p23.1 결실 증후군과 상호적이다. 8p23.1 중복은 언어 능력 지연, 발달 지연, 두드러진 이마와 아치형 눈썹을 동반한 경미한 형태이상, 그리고 선천성 심장 질환(CHD) 중에서 하나 또는 그 이상을 포함하는 가변적 표현형과 연관된다.

[0361] 염색체 15q 중복 증후군 (Dup15q)은 염색체 15q11-13.1의 중복으로부터 발생하는 임상적으로 인지가능한 증후군이다. Dup15q를 앓는 아기는 통상적으로, 저혈압증 (불량한 근육 긴장도), 성장 지연을 갖는다; 이들은 구순열 및/또는 구개 파열 또는 심장, 신장 또는 다른 장기의 기형을 갖고 태어날 수 있다; 이들은 어느 정도의 인지 지연/장애 (정신 지체), 언어 능력과 언어 지연, 그리고 감각 처리 장애를 보여준다.

[0362] 팔리스터 킬리안 증후군은 추가 #12 염색체 물질의 결과이다. 통상적으로 세포의 혼합물 (모자이크현상)이 존재하는데, 일부는 추가의 #12 물질을 갖고, 그리고 일부는 정상적이다 (추가의 #12 물질이 없는 46 염색체). 이러한 증후군을 앓는 아기는 심각한 정신 지체, 불량한 근육 긴장도, "거친" 안면 특질, 그리고 두드러진 이마를 비롯한 많은 문제점을 갖는다. 이들은 매우 얇은 윗입술과 더욱 두꺼운 아랫입술, 그리고 짧은 코를 갖는 경향이 있다. 다른 건강 문제는 발작, 불량한 영양공급, 뻣뻣한 관절, 성인기에 백내장, 청력 상실, 그리고 심장 결함을 포함한다. 팔리스터 킬리안을 앓는 개체는 단축된 수명을 갖는다.

[0363] dup(17)(p11.2p11.2) 또는 dup 17p로서 지정된 유전적 장애를 앓는 개체는 염색체 17의 짧은 팔에서 추가의 유전 정보 (중복으로서 알려져 있음)를 보유한다. 염색체 17p11.2의 중복은 포토스키 럽스키 증후군 (PTLS)의 원인가 되는데, 이것은 의학적 문헌에서 단지 수십 건의 사례만 보고된 새로 인정된 유전적 장애이다. 이러한 중복을 갖는 환자는 종종, 낮은 근육 긴장도, 불량한 영양공급, 그리고 영아기 동안 성장 장애를 갖고, 그리고 또한, 운동과 언어 이정표의 지연된 발달을 나타낸다. PTLS를 앓는 많은 개체는 관절 및 언어 처리에서 어려움을 갖는다. 이에 더하여, 환자는 자폐증 또는 자폐증-스펙트럼 장애를 앓는 개체에서 목격되는 것들과 유사한 행동 특징을 가질 수 있다. PTLS를 앓는 개체는 심장 결함 및 수면 무호흡을 앓을 수 있다. 유전자 PMP22를 포함하는 염색체 17p12에서 큰 영역의 중복은 샤르코 마리 치아 질환을 유발하는 것으로 알려져 있다.

[0364] CNV는 사산과 연관되었다. 하지만, 전통적인 세포유전학의 내재하는 한계로 인해, 사산에 대한 CNV의 기여는 불충분하게 표현되는 것으로 생각된다 (Harris et al., Prenatal Diagn 31:932-944 [2011]). 실시예에서 보여지고 본원의 다른 곳에서 설명되는 바와 같이, 본 발명 방법은 부분적인 이수성의 존재, 예를 들면, 염색체 분절의 결실과 증가를 결정할 수 있고, 그리고 사산과 연관되는 CNV의 존재 또는 부재를 확인하고 결정하는데 이용될 수 있다.

[0365] **CNV를 결정하기 위한 기구와 시스템**

[0366] 염기서열결정 데이터의 분석 및 그것으로부터 유래된 진단은 전형적으로, 다양한 컴퓨터 실행된 알고리즘과 프로그램을 이용하여 수행된다. 이런 이유로, 일정한 구체예는 하나 또는 그 이상의 컴퓨터 시스템 또는 다른 처리 시스템에서 저장된 또는 이들 시스템을 통하여 이전된 데이터를 수반하는 과정을 이용한다. 본원에서 개시된 구체예는 또한, 이들 작업을 수행하기 위한 기구에 관계한다. 이러한 기구는 필요한 목적을 위해 특별히 구축되거나, 또는 이것은 컴퓨터 내에 저장된 컴퓨터 프로그램 및/또는 데이터 구조에 의해 선별적으로 활성화되거나 또는 변경된 일반 목적 컴퓨터 (또는 일군의 컴퓨터)이다. 일부 구체예에서, 일군의 프로세서는 언급된 분석적 작업의 일부 또는 전부를 협력적으로 (가령, 네트워크 또는 클라우드 컴퓨팅을 통해) 및/또는 병렬적으로 수행한다. 본원에서 설명된 방법을 수행하기 위한 프로세서 또는 일군의 프로세서는 마이크로컨트롤러 및 마이크로프로세서, 예를 들면, 프로그램가능 장치 (가령, CPLDs 및 FPGAs) 및 비-프로그램가능 장치, 예를 들면, 게이트 어레이 ASICs 또는 일반 목적 마이크로프로세서를 비롯한 다양한 유형일 수 있다.

[0367] 이에 더하여, 일정한 구체예는 다양한 컴퓨터-실행된 작업을 수행하기 위한 프로그램 명령 및/또는 데이터 (데이터 구조 포함)를 포함하는 실재하는 및/또는 비-일시적 컴퓨터 판독가능 매체 또는 컴퓨터 프로그램 산물에 관계한다. 컴퓨터-판독가능 매체의 실례는 반도체 기억 장치, 자성 매체, 예를 들면, 디스크 드라이브, 자성 테이프, 광학 매체, 예를 들면, CDs, 광자기 매체, 그리고 프로그램 명령을 저장하고 수행하도록 특수하게 설정되는 하드웨어 장치, 예를 들면, 리드 전용 메모리 장치 (ROM) 및 무작위 접근 메모리 (RAM)를 포함하지만 이들에 한정되지 않는다. 컴퓨터 판독가능 매체는 최종 사용자에 의해 직접적으로 제어되거나 또는 이들 매체는 최종 사용자에 의해 간접적으로 제어될 수 있다. 직접적으로 제어된 매체의 실례는 사용자 시설에서 위치된 매체 및/또는 다른 실체와 공유되지 않는 매체를 포함한다. 간접적으로 제어된 배지의 실례는 외부 네트워크를 통해 및/또는 공유된 자원을 제공하는 서비스, 예를 들면, "클라우드"를 통해 사용자에게 간접적으로 접근가능한 매체를 포함한다. 프로그램 명령의 실례는 예로서, 컴파일러에 의해 생산된 양쪽 기계 코드, 그리고 번역기를 이용하여 컴퓨터에 의해 실행될 수 있는 더욱 높은 수준 코드를 내포하는 파일을 포함한다.

[0368] 다양한 구체예에서, 개시된 방법과 기구에서 이용된 데이터 또는 정보는 전자 형식으로 제공된다. 이런 데이터 또는 정보는 핵산 표본으로부터 유래된 리드와 태그, 참조 서열의 특정 영역에 맞추어 정렬하는 (가령, 염색체 또는 염색체 분절에 맞춰 정렬하는) 이런 태그의 수치 또는 밀도, 참조 서열 (오로지 또는 일차적으로 다형성을 제공하는 참조 서열 포함), 염색체와 분절 도스, 호출, 예를 들면, 이수성 호출, 정규화된 염색체와 분절 값, 염색체 또는 분절 및 상응하는 정규화 염색체 또는 분절의 쌍, 카운슬링 권고, 진단, 기타 등등을 포함할 수 있다. 본원에서 이용된 바와 같이, 전자 형식으로 제공된 데이터 또는 다른 정보는 기계에서 저장 및 기계 사이에 전달에 가용하다. 전통적으로, 전자 형식에서 데이터는 디지털 방식으로 제공되고 다양한 데이터 구조, 목록, 데이터베이스 등에서 비트 및/또는 바이트로서 저장될 수 있다. 데이터는 전자적으로, 광학적으로 등으로 구현될 수 있다.

[0369] 한 구체예는 시험 표본 내에 이수성, 예를 들면, 태아 이수성 또는 암의 존재 또는 부재를 지시하는 출력을 산출하기 위한 컴퓨터 프로그램 산물을 제공한다. 컴퓨터 산물은 염색체 이상을 결정하기 위한 상기 설명된 방법 중에서 임의의 한 가지 또는 그 이상을 수행하기 위한 명령을 내포할 수 있다. 설명된 바와 같이, 컴퓨터 산물은 프로세서가 염색체 도스 및 일부 경우에, 태아 이수성이 존재하거나 또는 부재하는 지의 여부를 결정하는 것을 가능하게 하는 컴퓨터 실행가능한 또는 편집가능한 논리 (가령, 명령)가 기록된 비-일시적 및/또는 실재적인 컴퓨터 판독가능 매체를 포함할 수 있다. 한 가지 실례에서, 컴퓨터 산물은 다음을 포함하는, 프로세서가 태아 이수성을 진단하는 것을 가능하게 하는 컴퓨터 실행가능한 또는 편집가능한 논리 (가령, 명령)가 기록된 컴퓨터 판독가능 매체를 포함한다: 모계 생물학적 표본으로부터 핵산 분자 중에서 최소한 일부로부터 염기서열결정 데이터를 제공받기 위한 제공받는 절차, 여기서 상기 염기서열결정 데이터는 계산된 염색체 및/또는 분절 도스를 포함하고; 상기 제공받은 데이터로부터 태아 이수성을 분석하기 위한 컴퓨터 보조된 논리; 그리고 상기 태아 이수성의 존재, 부재 또는 종류를 지시하는 출력을 산출하기 위한 출력 절차.

[0370] 고려 중인 표본으로부터 서열 정보는 관심되는 임의의 한 가지 또는 그 이상 염색체 각각에 대한 서열 태그의 개수를 확인하고, 그리고 관심되는 상기 임의의 한 가지 또는 그 이상 염색체 각각에 대한 정규화 분절 서열에 대한 서열 태그의 개수를 확인하기 위해, 염색체 참조 서열에 매핑될 수 있다. 다양한 구체예에서, 참조 서열은 데이터베이스, 예를 들면, 예로서 관계 또는 목적 데이터베이스에서 저장된다.

[0371] 도움을 받지 못한 인간이 본원에서 개시된 방법의 연산 작업을 수행하는 것은 많은 경우에 실질적이지 않거나, 또는 심지어 가능하지 않은 것으로 이해되어야 한다. 가령, 인간 염색체 중에서 임의의 한 가지에 대한 표본으로부터 단일 30 bp 리드를 매핑하는 것은 연산 기구의 도움이 없으면, 수년의 노력을 필요로 할지도 모른다. 당연히, 문제는 악화되는데, 그 이유는 신뢰할 만한 이수성 호출이 일반적으로, 하나 또는 그 이상의 염색체에 수천 개 (가령, 최소한 약 10,000개) 또는 심지어 수백만 개의 리드를 매핑하는 것을 필요로 하기 때문이다.

[0372] 본원에서 개시된 방법은 시험 표본 내에 관심되는 유전자 서열의 사본수의 평가를 위한 시스템을 이용하여 수행될 수 있다. 상기 시스템은 다음을 포함한다: (a) 표본으로부터 핵산 서열 정보를 제공하는 시험 표본으로부터 핵산을 제공받기 위한 서열분석기; (b) 프로세서; 그리고 (c) 상기 프로세서가 임의의 CNV, 예를 들면, 염색체 또는 부분적인 이수성을 확인하기 위한 방법을 실행하도록 이행을 위한 명령이 그 안에 저장된 하나 또는 그 이상의 컴퓨터-판독가능 저장 매체.

[0373] 일부 구체예에서, 이들 방법은 임의의 CNV, 예를 들면, 염색체 또는 부분적인 이수성을 확인하기 위한 방법을 수행하기 위한 컴퓨터-판독가능 명령이 그 안에 저장된 컴퓨터-판독가능 매체에 의해 지시된다. 따라서 한 구체예는 컴퓨터-실행가능 명령이 그 안에 저장된 하나 또는 그 이상의 컴퓨터-판독가능 비-일시적 저장 매체를 포함하는 컴퓨터 프로그램 산물을 제공하고, 상기 명령은 컴퓨터 시스템의 하나 또는 그 이상의 프로세서에 의해 실행될 때, 컴퓨터 시스템이 태아와 모계 무세포 핵산을 포함하는 시험 표본에서 관심되는 서열의 사본수의 평가를 위한 방법을 실행하도록 유발한다. 상기 방법은 다음을 포함한다: (a) 시험 표본의 서열 리드를 제공하고; (b) 시험 표본의 서열 리드를 관심되는 서열을 포함하는 참조 유전체에 맞춰 정렬하고, 따라서 시험 서열 태그를 제공하고; (c) 각 빈 내에 위치된 시험 서열 태그의 리드 커버리지를 결정하고, 여기서 참조 유전체는 복수의 빈으로 분할되고; (d) 관심되는 서열에 대한 전역 프로필을 제공하고, 여기서 전역 프로필은 각 빈에서 예상된 리드 커버리지를 포함하고, 그리고 여기서 예상된 리드 커버리지는 시험 표본과 실제적으로 동일한 방식으로 염기서열결정되고 정렬된 영향을 받지 않은 훈련 표본의 훈련 세트로부터 획득되고, 예상된 리드 커버리지는 빈마다 변동을 표시하고; (e) 각 빈에서 예상된 리드 커버리지에 따라 시험 서열 태그의 리드 커버리지를 조정하고, 따라서 시험 서열 태그의 각 빈에서 전역-프로필-교정된 리드 커버리지를 획득하고; (f) GC 함량 수준 및 시험 서열 태그의 빈에 대한 전역-프로필-교정된 리드 커버리지 사이의 관계에 근거하여 전역-프로필-교정된 리드 커버리지를 조정하고, 따라서 관심되는 서열 상에서 시험 서열 태그의 표본-GC-교정된 리드 커버리지를 획득

하고; 그리고 (g) 표본-GC-교정된 리드 커버리지에 근거하여 시험 표본 내에 관심되는 서열의 사본수를 평가한다. 일부 구체예에서, 단계 (c)에서 결정된 리드 커버리지는 정규화된다. 정규화는 리드 커버리지를 로버스트 염색체에 매핑하는 리드의 총수로 나눗셈하거나 또는 이것으로부터 리드 커버리지를 모형화하는 것을 수반할 수 있다 (때때로, 라이브러리 깊이 정규화로서 또한 지칭됨).

[0374] 일부 구체예에서, 명령은 상기 방법에 관련된 정보, 예를 들면, 염색체 도스 및 모계 시험 표본을 제공하는 인간 개체에 대한 환자 병력에서 태아 염색체 이수성의 존재 또는 부재를 자동적으로 기록하는 것을 더욱 포함할 수 있다. 환자 병력은 예로서, 실험실, 의사의 사무실, 병원, 건강 관리 기관, 보험 회사, 또는 개인 병력 웹사이트 에 의해 유지될 수 있다. 게다가, 프로세서-실행된 분석의 결과에 근거하여, 상기 방법은 모계 시험 표본이 채취되었던 인간 개체의 치료를 처방하고, 개시하고, 및/또는 변경하는 것을 더욱 수분할 수 있다. 이것은 개체로부터 채취된 추가 표본 상에서 하나 또는 그 이상의 추가 시험 또는 분석을 수행하는 것을 수반할 수 있다.

[0375] 개시된 방법은 또한, 임의의 CNV, 예를 들면, 염색체 또는 부분적인 이수성을 확인하기 위한 방법을 수행하도록 적응되거나 또는 설정된 컴퓨터 처리 시스템을 이용하여 수행될 수 있다. 한 구체예는 본원에서 설명된 바와 같은 방법을 수행하도록 적응되거나 또는 설정된 컴퓨터 처리 시스템을 제공한다. 한 구체예에서, 기구는 본원의 다른 곳에서 설명된 서열 정보의 유형을 획득하기 위해 표본 내에 핵산 분자 중에서 최소한 일부를 염기서열결정하도록 적응된 또는 설정된 염기서열결정 장치를 포함한다. 기구는 또한, 표본의 처리를 위한 성분을 포함할 수 있다. 이런 성분은 본원의 다른 곳에서 설명된다.

[0376] 서열 또는 다른 데이터는 컴퓨터 내로 입력되거나, 또는 직접적으로 또는 간접적으로 컴퓨터 판독가능 매체에 저장될 수 있다. 한 구체예에서, 컴퓨터 시스템은 표본으로부터 핵산의 서열을 판독하고 및/또는 분석하는 염기서열결정 장치에 직접적으로 연계된다. 이런 도구로부터 서열 또는 다른 정보는 컴퓨터 시스템 내에 인터페이스를 통해 제공된다. 대안으로, 시스템에 의해 처리된 이들 서열은 서열 저장 출처, 예를 들면, 데이터베이스 또는 다른 저장소로부터 제공된다. 일단 처리 기구에 가용하면, 기억 장치 또는 질량 저장 장치는 핵산의 서열을 최소한 일시적으로 완충하거나 또는 저장한다. 이에 더하여, 기억 장치는 다양한 염색체 또는 유전체 등에 대한 태그 수치를 저장할 수 있다. 메모리는 또한, 서열 또는 매핑된 데이터를 분석하고 제공하기 위한 다양한 루틴 및/또는 프로그램을 저장한다. 이런 프로그램/루틴은 통계학적 분석을 수행하기 위한 프로그램 등을 포함할 수 있다.

[0377] 한 가지 실례에서, 사용자는 표본을 염기서열결정 기구 내로 제공한다. 데이터는 컴퓨터에 연결되는 염기서열결정 기구에 의해 수집되고 및/또는 분석된다. 컴퓨터 상에서 소프트웨어는 데이터 수집 및/또는 분석을 허용한다. 데이터는 저장되고, 표시되고 (모니터 또는 다른 유사한 장치를 통해), 및/또는 다른 위치로 보내질 수 있다. 컴퓨터는 데이터를 원격 사용자 (가령, 의사, 과학자 또는 분석가)에 의해 활용되는 손바닥크기 장치로 전송하는데 이용되는 인터넷에 연결될 수 있다. 데이터는 전송에 앞서 저장되고 및/또는 분석될 수 있는 것으로 이해된다. 일부 구체예에서, 미가공 데이터는 수집되고, 그리고 데이터를 분석하고 및/또는 저장할 원격 사용자 또는 기구에 보내진다. 전송은 인터넷을 통해 일어날 수 있지만, 위성 또는 다른 연결을 통해 발생할 수도 있다. 대안으로, 데이터는 컴퓨터-판독가능 매체 상에 저장될 수 있고, 그리고 상기 매체가 최종 사용자에게 발송될 수 있다 (가령, 메일을 통해). 원격 사용자는 건물, 시, 주, 국가 또는 대륙을 포함하지만 이들에 한정되지 않는 동일한 또는 상이한 지리학적 위치 내에 있을 수 있다.

[0378] 일부 구체예에서, 이들 방법은 또한, 복수의 폴리뉴클레오티드 서열에 관한 데이터 (가령, 리드, 태그 및/또는 참조 염색체 서열)을 수집하고 상기 데이터를 컴퓨터 또는 다른 연산 시스템으로 전송하는 하는 것을 포함한다. 가령, 컴퓨터는 실험실 설비, 예를 들면, 표본 수집 기구, 뉴클레오티드 증폭 기구, 뉴클레오티드 염기서열결정 기구, 또는 혼성화 기구에 연결될 수 있다. 컴퓨터는 이후, 실험실 장치에 의해 모아진 적용가능한 데이터를 수집할 수 있다. 데이터는 임의의 단계에서, 예를 들면, 실시간으로 수집되는 동안, 전송에 앞서, 전송 동안 또는 전송과 함께, 또는 전송 이후에 컴퓨터에 저장될 수 있다. 데이터는 컴퓨터로부터 추출될 수 있는 컴퓨터-판독가능 매체 상에 저장될 수 있다. 수집되거나 또는 저장된 데이터는 예로서, 국부 네트워크 또는 광역 통신망, 예를 들면, 인터넷을 통해 컴퓨터로부터 원격 위치로 전송될 수 있다. 원격 위치에서, 전송된 데이터에서 다양한 작업이 아래에 설명된 바와 같이 수행될 수 있다.

[0379] 본원에서 개시된 시스템, 기구, 그리고 방법에서 저장되고, 전송되고, 분석되고, 및/또는 조작될 수 있는 전자적으로 형식화된 데이터의 유형에는 다음이 포함된다:

[0380] 시험 표본 내에 핵산을 염기서열결정함으로써 획득된 리드

[0381] 리드를 참조 유전체 또는 다른 참조 서열 또는 서열에 맞춰 정렬함으로써 획득된 태그

[0382] 참조 유전체 또는 서열

[0383] 서열 태그 밀도 - 참조 유전체 또는 다른 참조 서열의 2개 또는 그 이상 영역 (전형적으로 염색체 또는 염색체 분절) 각각에 대한 태그의 수치 또는 숫자

[0384] 관심되는 특정 염색체 또는 염색체 분절에 대한 정규화 염색체 또는 염색체 분절의 동일성

[0385] 관심되는 염색체 또는 분절 및 상응하는 정규화 염색체 또는 분절로부터 획득된 염색체 또는 염색체 분절 (또는 다른 영역)에 대한 도스

[0386] 염색체 도스를 영향을 받음, 영향을 받지 않음, 또는 호출 없음으로 호출하기 위한 역치

[0387] 염색체 도스의 실제 호출

[0388] 진단 (호출과 연관된 임상적 상태)

[0389] 호출 및/또는 진단으로부터 도출된 추가 시험에 대한 권고

[0390] 호출 및/또는 진단으로부터 도출된 치료 및/또는 모니터링 계획

[0391] 이들 다양한 유형의 데이터는 상이한 기구를 이용하여 하나 또는 그 이상의 위치에서 획득되고, 저장되고, 전송되고, 분석되고, 및/또는 조작될 수 있다. 처리 옵션은 넓은 스펙트럼에 걸친다. 스펙트럼의 한쪽 단부에서, 모든 또는 많은 이러한 정보는 시험 표본이 처리되는 위치, 예를 들면, 의사의 사무소 또는 다른 임상적 세팅에서 저장되고 이용된다. 다른 극단에서, 표본이 한 위치에서 획득되고, 상기 표본이 상이한 위치에서 처리되고 임의 선택적으로 염기서열결정되고, 리드가 정렬되고 호출이 하나 또는 그 이상의 상이한 위치에서 만들어지고, 그리고 진단, 권고, 및/또는 계획이 또 다른 위치에서 준비된다 (이것은 표본이 획득되었던 위치일 수도 있다).

[0392] 다양한 구체예에서, 리드가 염기서열결정 기구로 산출되고, 그리고 이후, 이들이 처리되는 원격 부위로 전송되어 이수성 호출이 산출된다. 이러한 원격 위치에서, 실례로서, 리드는 참조 서열에 맞춰 정렬되어 태그가 생산되고, 이들은 계수되고 관심되는 염색체 또는 분절에 배정된다. 또한 원격 위치에서, 이들 수치는 연관된 정규화 염색체 또는 분절을 이용하여 도스로 전환된다. 더 나아가, 원격 위치에서, 이들 도스는 이수성 호출을 산출하는데 이용된다.

[0393] 상이한 위치에서 이용될 수 있는 처리 작업에는 다음이 포함된다:

[0394] 표본 수집

[0395] 염기서열결정에 예비적인 표본 처리

[0396] 염기서열결정

[0397] 서열 데이터를 분석하고 이수성 호출을 도출

[0398] 진단

[0399] 진단 및/또는 호출을 환자 또는 건강 관리 제공자에게 보고

[0400] 추가 치료, 시험, 및/또는 모니터링을 위한 계획을 개발

[0401] 계획을 이행

[0402] 카운슬링

[0403] 이들 작업 중에서 임의의 한 가지 또는 그 이상이 본원의 다른 곳에서 설명된 바와 같이 자동화될 수 있다. 전형적으로, 염기서열결정 및 서열 데이터의 분석 및 이수성 호출의 도출은 연산적으로 수행될 것이다. 다른 작업은 수동으로 또는 자동적으로 수행될 수 있다.

[0404] 표본 수집이 수행될 수 있는 위치의 실례는 건강 요원의 사무소, 진료소, 환자의 집 (여기서 표본 수집 도구 또는 키트가 제공된다), 그리고 이동 건강 관리 차량을 포함한다. 염기서열결정에 앞서 표본 처리가 수행될 수 있는 위치의 실례는 건강 요원의 사무실, 진료소, 환자의 집 (여기서 표본 처리 기구 또는 키트가 제공된다), 이동 건강 관리 차량, 그리고 이수성 분석 제공자의 시설을 포함한다. 염기서열결정이 수행될 수 있는 위치의 실례는 건강 요원의 사무실, 진료소, 건강 요원의 사무실, 진료소, 환자의 집 (여기서 표본 염기서열결정 기구 및

/또는 키트가 제공된다), 이동 건강 관리 차량, 그리고 이수성 분석 제공자의 시설을 포함한다. 염기서열결정이 발생하는 위치는 데이터 (전형적으로 리드)를 전자 형식으로 전송하기 위한 서열 전용 네트워크 연결이 제공될 수 있다. 이런 연결은 유선 또는 무선일 수 있고, 그리고 데이터가 처리될 수 있고 및/또는 처리 위치로 전송에 앞서 병합되는 위치를 갖고 이러한 위치에 데이터를 보내도록 설정될 수 있다. 데이터 병합은 의료 기관, 예를 들면, 건강 관리 기관 (HMOs)에 의해 유지될 수 있다.

[0405] 분석 및/또는 도출 작업은 전술한 위치 중에서 한 곳 또는 대안으로, 연산 및/또는 핵산 서열 데이터를 분석하는 서비스만을 목적으로 하는 더욱 먼 위치에서 수행될 수 있다. 이런 위치는 예로서, 클러스터, 예를 들면, 일반 목적 서버 팜, 이수성 분석 서비스 사업의 시설 등을 포함한다. 일부 구체예에서, 분석을 수행하는데 이용된 연산 기구는 임차되거나 또는 임대된다. 연산 자원은 프로세서의 인터넷 접근가능한 수집, 예를 들면, 구어적으로 클라우드로서 알려져 있는 처리 자원의 부분일 수 있다. 일부 경우에, 연산은 서로 연계되거나 또는 연계되지 않은 프로세서의 병렬 또는 대량으로 병렬 군에 의해 수행된다. 처리는 분산 처리, 예를 들면, 클러스터 컴퓨팅, 그리드 컴퓨팅, 기타 등등을 이용하여 달성될 수 있다. 이런 구체예에서, 연산 자원의 클러스터 또는 격자는 본원에서 설명된 분석 및/또는 도출을 수행하기 위해 함께 작동하는 복수 프로세서 또는 컴퓨터로 구성된 슈퍼 가상 컴퓨터를 집합적으로 형성한다. 이들 기술뿐만 아니라 더욱 전통적인 슈퍼컴퓨터가 본원에서 설명된 바와 같이 서열 데이터를 처리하는데 이용될 수 있다. 각각은 프로세서 또는 컴퓨터에 의존하는 병렬 컴퓨팅의 형태이다. 그리드 컴퓨팅의 경우에, 이들 프로세서 (종종, 전체 컴퓨터)는 전통적인 네트워크 프로토콜, 예를 들면, 이더넷에 의한 네트워크 (민간, 공공, 또는 인터넷)에 의해 연결된다. 대조적으로, 슈퍼컴퓨터는 많은 프로세서가 국부 고속 컴퓨터 버스에 의해 연결된다.

[0406] 일정한 구체예에서, 진단 (가령, 태아가 다운 증후군을 앓거나 또는 환자가 특정 유형의 암을 앓는다)은 분석 작업과 동일한 위치에서 산출된다. 다른 구체예에서, 이것은 상이한 위치에서 수행된다. 일부 실례에서, 진단을 보고하는 것이 표본이 채취되었던 위치에서 수행되긴 하지만, 이것이 기정 사실인 것은 아니다. 진단이 산출되거나 또는 보고될 수 있고 및/또는 계획을 개발하는 것이 수행되는 위치의 실례는 건강 요원의 사무소, 진료소, 컴퓨터에 의해 접근가능한 인터넷 부위, 그리고 네트워크에 유선 또는 무선 연결을 갖는 손바닥크기 장치, 예를 들면, 휴대폰, 태블릿, 스마트폰 등을 포함한다. 카운슬링이 수행되는 위치의 실례는 건강 요원의 사무소, 진료소, 컴퓨터에 의해 접근가능한 인터넷 부위, 손바닥크기 장치 등을 포함한다.

[0407] 일부 구체예에서, 표본 수집, 표본 처리, 그리고 염기서열결정 작업은 첫 번째 위치에서 수행되고, 그리고 분석과 도출 작업은 두 번째 위치에서 수행된다. 하지만, 일부 경우에, 표본 수집은 한 위치 (가령, 건강 요원의 사무소 또는 클리닉)에서 수집되고, 그리고 표본 처리와 염기서열결정은 임의선택적으로 분석 및 도출이 발생하는 위치와 동일한 상이한 위치에서 수행된다.

[0408] 다양한 구체예에서, 상기 열거된 작업의 서열은 사용자, 또는 표본 수집, 표본 처리 및/또는 염기서열결정을 개시하는 실체에 의해 촉발될 수 있다. 하나 또는 그 이상의 이들 작업이 이행을 시작한 후에, 다른 작업이 자연적으로 뒤따를 것이다. 가령, 염기서열결정 작업은 리드가 자동적으로 수집되고 처리 기구로 보내지도록 유발할 수 있고, 이것은 이후, 종종 자동적으로 및 아마도 추가 이용 개입 없이, 이수성 작업의 서열 분석과 도출을 수행한다. 일부 실행에서, 이러한 과정 작업의 결과는 이후, 아마도 진단 결과로서 재포맷으로, 시스템 성분 또는 실체에 자동적으로 전달되고, 이의 처리는 정보를 건강 전문가 및/또는 환자에게 보고한다. 설명된 바와 같이, 이런 정보는 또한, 아마도 카운슬링 정보와 함께 치료, 시험, 및/또는 모니터링 계획을 산출하기 위해 자동적으로 처리될 수 있다. 따라서, 초기 단계 작업을 개시하는 것은 끝과 끝 서열을 촉발할 수 있고, 여기에서 건강 전문가, 환자 또는 다른 당사자는 신체 상태에 작용하는데 유용한 진단, 계획, 카운슬링 및/또는 다른 정보가 제공된다. 이것은 전체 시스템의 부분이 물리적으로 분리되고 아마도 예로서, 표본 및 서열 기구의 위치로부터 멀리 떨어진 경우에도 달성된다.

[0409] 도면 5는 시험 표본으로부터 호출 또는 진단을 산출하기 위한 분산된 시스템의 한 가지 실행을 보여준다. 표본 수집 위치 01이 환자, 예를 들면, 임신 암컷 또는 추정 암 환자로부터 시험 표본을 획득하는데 이용된다. 표본은 이후, 처리와 염기서열결정 위치 03에 제공되고, 여기서 시험 표본은 앞서 설명된 바와 같이 처리되고 염기서열결정될 수 있다. 위치 03은 표본을 처리하기 위한 기구뿐만 아니라 처리된 표본을 염기서열결정하기 위한 기구를 포함한다. 염기서열결정의 결과는 본원의 다른 곳에서 설명된 바와 같이, 전형적으로 전자 형식으로 제공되고, 그리고 도면 5에서 참조 번호 05에 의해 표시되는 네트워크, 예를 들면, 인터넷에 제공되는 리드의 수집물이다.

[0410] 서열 데이터는 원격 위치 07에 제공되고, 여기서 분석과 호출 산출이 수행된다. 이러한 위치는 하나 또는 그 이

상의 유력한 연산 장치, 예를 들면, 컴퓨터 또는 프로세서를 포함할 수 있다. 위치 07에서 연산 자원이 그들의 분석을 완결하고, 그리고 제공받은 서열 정보로부터 호출을 산출한 후에, 호출은 네트워크 05에 역으로 중계된다. 일부 실행에서, 위치 07에서 호출이 산출될 뿐만 아니라 연관된 진단 역시 산출된다. 호출 및 또는 진단은 이후, 도면 5에서 예시된 바와 같이 네트워크 전체에 대하여 및 표본 수집 위치 01에 역으로 전송된다. 설명된 바와 같이, 이것은 호출 또는 진단을 산출하는 것과 연관된 다양한 작업이 다양한 위치 사이에서 어떻게 분할될 수 있는 지에 관한 많은 변동 중에서 단지 한 가지일 뿐이다. 한 가지 통상적인 변동은 단일 위치에서 표본 수집과 처리 및 염기서열결정을 제공하는 것을 수반한다. 다른 변동은 분석 및 호출 산출과 동일한 위치에서 처리 및 염기서열결정을 제공하는 것을 수반한다.

[0411] 도면 6은 상이한 위치에서 다양한 작업을 수행하기 위한 옵션을 상술한다. 도면 6에서 묘사된 가장 일반적인 의미에서, 다음의 작업 각각은 별개의 위치에서 수행된다: 표본 수집, 표본 처리, 서열 리드 정렬, 호출, 진단, 그리고 보고 및/또는 계획 개발.

[0412] 이들 작업 중에서 일부를 병합하는 한 구체예에서, 표본 처리 및 염기서열결정은 한 위치에서 수행되고, 그리고 리드 정렬, 호출 및 진단은 별개의 위치에서 수행된다. 참조 문자 A에 의해 확인된 도면 6의 부분을 참조한다. 도면 6에서 문자 B에 의해 확인되는 다른 실행에서, 표본 수집, 표본 처리, 그리고 염기서열결정 모두 동일한 위치에서 수행된다. 이러한 실행에서, 리드 정렬 및 호출은 두 번째 위치에서 수행된다. 최종적으로, 진단 및 보고 및/또는 계획 개발은 세 번째 위치에서 수행된다. 도면 6에서 문자 C에 의해 묘사된 실행에서, 표본 수집은 첫 번째 위치에서 수행되고, 표본 처리, 서열 리드 정렬, 호출 및 진단 모두 두 번째 위치에서 함께 수행되고, 그리고 보고 및/또는 계획 발달은 세 번째 위치에서 수행된다. 최종적으로, 도면 6에서 문자 D로 표지화된 실행에서, 표본 수집은 첫 번째 위치에서 수행되고, 표본 처리, 서열 리드 정렬 및 호출 모두 두 번째 위치에서 수행되고, 그리고 진단 및 보고 및/또는 계획 관리는 세 번째 위치에서 수행된다.

[0413] 한 구체예는 태아와 모계 핵산을 포함하는 시험 표본 내에 임의의 한 가지 또는 그 이상의 상이한 완전한 태아 염색체 이수성 모계의 존재 또는 부재를 결정하는데 이용하기 위한 시스템을 제공하고, 상기 시스템은 핵산 표본을 제공받고 표본으로부터 태아와 모계 핵산 서열 정보를 제공하기 위한 서열분석기; 프로세서; 그리고 상기 프로세서에서 이행을 위한 명령을 포함하는 기계 판독가능한 저장 매체를 포함하고, 상기 명령은 다음을 포함한다:

[0414] (a) 표본에서 상기 태아와 모계 핵산에 대한 서열 정보를 획득하기 위한 코드;

[0415] (b) 염색체 1-22, X와 Y에서 선택되는 관심되는 임의의 한 가지 또는 그 이상 염색체 각각에 대한 태아와 모계 핵산으로부터 서열 태그의 개수를 연산적으로 확인하고, 그리고 관심되는 상기 임의의 한 가지 또는 그 이상 염색체 각각에 대한 최소한 하나의 정규화 염색체 서열 또는 정규화 염색체 분절 서열에 대한 서열 태그의 개수를 확인하기 위해, 상기 서열 정보를 이용하기 위한 코드;

[0416] (c) 관심되는 임의의 한 가지 또는 그 이상 염색체 각각에 대한 단일 염색체 도스를 계산하기 위해, 관심되는 상기 임의의 한 가지 또는 그 이상 염색체 각각에 대해 확인된 서열 태그의 상기 숫자 및 각 정규화 염색체 서열 또는 정규화 염색체 분절 서열에 대해 확인된 서열 태그의 상기 숫자를 이용하기 위한 코드; 그리고

[0417] (d) 관심되는 임의의 한 가지 또는 그 이상 염색체 각각에 대한 각각의 단일 염색체 도스를 관심되는 하나 또는 그 이상의 염색체 각각에 대한 상응하는 역치값과 비교하고, 그리고 따라서, 표본 내에 임의의 한 가지 또는 그 이상의 완전한 상이한 태아 염색체 이수성의 존재 또는 부재를 결정하기 위한 코드.

[0418] 일부 구체예에서, 관심되는 임의의 한 가지 또는 그 이상 염색체 각각에 대한 단일 염색체 도스를 계산하기 위한 코드는 관심되는 염색체 중에서 선별된 한 가지에 대한 염색체 도스를 관심되는 선별된 염색체에 대해 확인된 서열 태그의 개수 및 관심되는 선별된 염색체에 대한 상응하는 최소한 하나의 정규화 염색체 서열 또는 정규화 염색체 분절 서열에 대해 확인된 서열 태그의 개수의 비율로서 계산하기 위한 코드를 포함한다.

[0419] 일부 구체예에서, 시스템은 관심되는 임의의 한 가지 또는 그 이상 염색체의 임의의 한 가지 또는 그 이상 분절의 임의의 남아있는 염색체 분절 각각에 대한 염색체 도스의 계산을 반복하기 위한 코드를 더욱 포함한다.

[0420] 일부 구체예에서, 염색체 1-22, X와 Y에서 선택되는 관심되는 하나 또는 그 이상의 염색체는 염색체 1-22, X와 Y에서 선택되는 최소한 20개 염색체를 포함하고, 그리고 여기서 명령은 최소한 20개의 상이한 완전한 태아 염색체 이수성의 존재 또는 부재를 결정하기 위한 명령을 포함한다.

[0421] 일부 구체예에서, 최소한 하나의 정규화 염색체 서열은 염색체 1-22, X와 Y에서 선택되는 일군의 염색체이다.

다른 구체예에서, 최소한 하나의 정규화 염색체 서열은 염색체 1-22, X와 Y에서 선택되는 단일 염색체이다.

[0422] 다른 구체예는 태아와 모계 핵산을 포함하는 시험 표본 내에 임의의 한 가지 또는 그 이상의 상이한 부분적인 태아 염색체 이수성 모계의 존재 또는 부재를 결정하는데 이용하기 위한 시스템을 제공하고, 상기 시스템은 핵산 표본을 제공받고 표본으로부터 태아와 모계 핵산 서열 정보를 제공하기 위한 서열분석기; 프로세서; 그리고 상기 프로세서에서 이행을 위한 명령을 포함하는 기계 판독가능한 저장 매체를 포함하고, 상기 명령은 다음을 포함한다:

[0423] (a) 상기 표본에서 상기 태아와 모계 핵산에 대한 서열 정보를 획득하기 위한 코드;

[0424] (b) 염색체 1-22, X와 Y에서 선택되는 관심되는 임의의 한 가지 또는 그 이상 염색체의 임의의 한 가지 또는 그 이상 분절 각각에 대한 태아와 모계 핵산으로부터 서열 태그의 개수를 연산적으로 확인하고, 그리고 관심되는 임의의 한 가지 또는 그 이상 염색체의 상기 임의의 한 가지 또는 그 이상 분절 각각에 대한 최소한 하나의 정규화 분절 서열에 대한 서열 태그의 개수를 확인하기 위해, 상기 서열 정보를 이용하기 위한 코드;

[0425] (c) 관심되는 임의의 한 가지 또는 그 이상 염색체의 상기 임의의 한 가지 또는 그 이상 분절 각각에 대한 단일 염색체 분절 도스를 계산하기 위해, 관심되는 임의의 한 가지 또는 그 이상 염색체의 상기 임의의 한 가지 또는 그 이상 분절 각각에 대해 확인된 서열 태그의 상기 숫자 및 상기 정규화 분절 서열에 대해 확인된 서열 태그의 상기 숫자를 이용하는 코드; 그리고

[0426] (d) 관심되는 임의의 한 가지 또는 그 이상 염색체의 상기 임의의 한 가지 또는 그 이상 분절 각각에 대한 각각의 상기 단일 염색체 분절 도스를 관심되는 임의의 한 가지 또는 그 이상 염색체의 상기 임의의 한 가지 또는 그 이상 염색체 분절 각각에 대한 상응하는 역치값과 비교하고, 그리고 따라서 상기 표본 내에 하나 또는 그 이상의 상이한 부분적인 태아 염색체 이수성의 존재 또는 부재를 결정하기 위한 코드.

[0427] 일부 구체예에서, 단일 염색체 분절 도스를 계산하기 위한 코드는 염색체 분절 중에서 선별된 한 가지에 대한 염색체 분절 도스를 선별된 염색체 분절에 대해 확인된 서열 태그의 개수 및 선별된 염색체 분절에 대한 상응하는 정규화 분절 서열에 대해 확인된 서열 태그의 개수의 비율로서 계산하기 위한 코드를 포함한다.

[0428] 일부 구체예에서, 시스템은 관심되는 임의의 한 가지 또는 그 이상 염색체의 임의의 한 가지 또는 그 이상 분절의 임의의 남아있는 염색체 분절 각각에 대한 염색체 분절 도스의 계산을 반복하기 위한 코드를 더욱 포함한다.

[0429] 일부 구체예에서, 시스템은 (i) 상이한 모계 개체로부터 시험 표본에 대한 (a)-(d)를 반복하기 위한 코드, 그리고 (ii) 상기 표본 각각에서 임의의 한 가지 또는 그 이상의 상이한 부분적인 태아 염색체 이수성의 존재 또는 부재를 결정하기 위한 코드를 더욱 포함한다.

[0430] 본원에서 제공된 시스템 중에서 한 가지의 다른 구체예에서, 코드는 모계 시험 표본을 제공하는 인간 개체에 대한 환자 병력에서 (d)에서 결정된 바와 같은 태아 염색체 이수성의 존재 또는 부재를 자동적으로 기록하기 위한 코드를 더욱 포함하고, 여기서 기록은 프로세서를 이용하여 수행된다.

[0431] 본원에서 제공된 시스템 중에서 한 가지의 일부 구체예에서, 서열분석기는 차세대 염기서열결정 (NGS)을 수행하도록 구성된다. 일부 구체예에서, 서열분석기는 가역성 염료 종결인자로 합성에 의한 염기서열결정을 이용한 대량으로 병렬 염기서열결정을 수행하도록 구성된다. 다른 구체예에서, 서열분석기는 결찰에 의한 염기서열결정을 수행하도록 구성된다. 또 다른 구체예에서, 서열분석기는 단일 분자 염기서열결정을 수행하도록 구성된다.

[0432] **실험적**

[0433] **실시예 1**

[0434] **일차성 및 농축된 염기서열결정 라이브러리의 제조와 염기서열결정**

[0435] *a. 염기서열결정 라이브러리의 제조 - 단축된 프로토콜 (ABB)*

[0436] 모든 염기서열결정 라이브러리, 다시 말하면, 일차성 및 농축된 라이브러리는 모계 혈장으로부터 추출되었던 대략 2 ng의 정제된 cfDNA로부터 제조되었다. 라이브러리 제조는 아래와 같이 Illumina®에 대해, NEBNext™ DNA 표본 Prep DNA 시약 세트 1 (부품 번호 E6000L; New England Biolabs, Ipswich, MA)의 시약을 이용하여 수행되었다. 무세포 혈장 DNA가 자연에서 단편화되기 때문에, 혈장 DNA 표본 상에서 분무 또는 초음파처리에 의한 어떤 추가 단편화도 행위되지 않았다. 40 μl에서 내포된 대략 2 ng 정제된 cfDNA 단편의 오버행은 1.5 ml 마이크로퓨즈 튜브에서, cfDNA를 NEBNext™ DNA 표본 Prep DNA 시약 세트 1에서 제공된 5μl 10X 인산화 완충액, 2

μl 데옥시뉴클레오티드 용액 믹스 (10 mM 각 dNTP), 1μl의 DNA 중합효소 I의 1:5 희석액, 1 μl T4 DNA 중합효소 및 1 μl T4 폴리뉴클레오티드 키나아제와 함께 20℃에서 15 분 동안 항온처리함으로써, NEBNext® 단부 수복 모듈에 따라 인산화된 평활 말단으로 전환되었다. 이들 효소는 이후, 반응 혼합물을 75℃에서 5 분 동안 항온처리함으로써 열 비활성화되었다. 혼합물은 4℃로 냉각되었고, 그리고 평활 말단 DNA의 dA 테일링이 클레노브 단편 (3'에서 5' 엑소 마이너스)을 내포하는 10μl의 dA-테일링 마스터 믹스 (NEBNext™ DNA 표본 Prep DNA 시약 세트 1)를 이용하고, 그리고 37℃에서 15 분 동안 배양함으로써 달성되었다. 차후에, 클레노브 단편은 반응 혼합물을 75℃에서 5 분 동안 항온처리함으로써 열 비활성화되었다. 클레노브 단편의 비활성화 이후에, 1 μl의 Illumina 유전체 어댑터 올리고 믹스 (부품 번호 1000521; Illumina Inc., Hayward, CA)의 1:5 희석액이 반응 혼합물을 25℃에서 15 분 동안 항온처리함으로써, NEBNext™ DNA 표본 Prep DNA 시약 세트 1에서 제공된 4 μl의 T4 DNA 연결효소를 이용하여, Illumina 어댑터 (비-지수 Y-어댑터)를 dA-측 DNA에 결찰하는데 이용되었다. 혼합물은 4℃로 냉각되었고, 그리고 어댑터-결찰된 cfDNA는 Agencourt AMPure XP PCR 정제 시스템 (부품 번호 A63881; Beckman Coulter Genomics, Danvers, MA)에서 제공된 자성 비드를 이용하여, 결찰되지 않은 어댑터, 어댑터 이합체, 그리고 다른 시약으로부터 정제되었다. 18회 주기의 PCR이 Phusion® 높은-충실성 마스터 믹스 (25 μl; Finnzymes, Woburn, MA) 및 이들 어댑터에 상보적인 Illumina의 PCR 프라이머 (각각 0.5 μM) (부품 번호 1000537과 1000537)를 이용하여 어댑터-결찰된 cfDNA (25 μl)를 선별적으로 농축하기 위해 수행되었다. 어댑터-결찰된 DNA는 제조업체의 사용설명서에 따라, Illumina 유전체 PCR 프라이머 (부품 번호 100537과 1000538) 및 NEBNext™ DNA 표본 Prep DNA 시약 세트 1에서 제공된 Phusion HF PCR 마스터 믹스를 이용하여 PCR (30 초 동안 98℃; 10 초 동안 98℃, 30 초 동안 65℃ 및 30 초 동안 72℃의 18회 주기; 5 분 동안 72℃에서 최종 연장, 그리고 4℃에서 유지)에 종속되었다. 증폭된 산물은 www.beckmangenomics.com/products/AMPureXPProtocol_000387v001.pdf에서 가용한 제조업체의 사용설명서에 따라, Agencourt AMPure XP PCR 정제 시스템 (Agencourt Bioscience Corporation, Beverly, MA)을 이용하여 정제되었다. 정제된 증폭된 산물은 40 μl의 Qiagen EB 완충액에서 용리되었고, 그리고 증폭된 라이브러리의 농도와 크기 분포가 2100 바이오분석기 (Agilent technologies Inc., Santa Clara, CA)에 대한 Agilent DNA 1000 키트를 이용하여 분석되었다.

[0437] *b. 염기서열결정 라이브러리의 제조 -전장 프로토콜*

[0438] 여기에서 설명된 전장 프로토콜은 본질적으로, Illumina에 의해 제공된 표준 프로토콜이고, 그리고 증폭된 라이브러리의 정제에서만 Illumina 프로토콜과 상이하다. Illumina 프로토콜은 증폭된 라이브러리가 겔 전기영동을 이용하여 정제될 것을 지시하는 반면, 본원에서 설명된 프로토콜은 동일한 정제 단계에 대해 자성 비드를 이용한다. 모계 혈장으로부터 추출된 대략 2 ng의 정제된 cfDNA가 본질적으로 제조업체의 사용설명서에 따라, Illumina®에 대한 NEBNext™ DNA 표본 Prep DNA 시약 세트 1 (부품 번호 E6000L; New England Biolabs, Ipswich, MA)을 이용하여, 일차성 염기서열결정 라이브러리를 제조하는데 이용되었다. 정제 칼럼 대신에 Agencourt 자성 비드 및 시약을 이용하여 수행된 어댑터-결찰된 산물의 최종 정제를 제외한 모든 단계는 Illumina® GAII을 이용하여 염기서열결정되는 유전체 DNA 라이브러리에 대한 표본 제조를 위한 NEBNext™ 시약을 동행하는 프로토콜에 따라 수행되었다. NEBNext™ 프로토콜은 본질적으로, grcf.jhml.edu/hts/protocols/11257047_ChIP_Sample_Prep.pdf에서 가용한 Illumina에 의해 제공된 것을 따른다.

[0439] 40 μl에서 내포된 대략 2 ng 정제된 cfDNA 단편의 오버행은 40μl cfDNA를 200 μl 마이크로퓨즈 튜브에서 유전자 증폭기에서, NEBNext™ DNA 표본 Prep DNA 시약 세트 1에서 제공된 5μl 10X 인산화 완충액, 2 μl 데옥시뉴클레오티드 용액 믹스 (10 mM 각 dNTP), 1 μl의 DNA 중합효소 I의 1:5 희석액, 1 μl T4 DNA 중합효소 및 1 μl T4 폴리뉴클레오티드 키나아제와 함께 20℃에서 30 분 동안 항온처리함으로써, NEBNext® 단부 수복 모듈에 따라 인산화된 평활 말단으로 전환되었다. 표본은 4℃로 냉각되고, 그리고 아래와 같이 QIAQuick PCR 정제 키트 (Qiagen Inc., Valencia, CA)에서 제공된 QIAQuick 칼럼을 이용하여 정제되었다. 50 μl 반응물은 1.5 ml 마이크로퓨즈 튜브로 이전되고, 그리고 250 μl의 Qiagen 완충액 PB가 첨가되었다. 결과의 300 μl는 QIAquick 칼럼으로 이전되었고, 이것은 마이크로퓨즈에서 13,000 RPM에서 1 분 동안 원심분리되었다. 칼럼은 750 μl Qiagen 완충액 PE으로 세척되고, 그리고 재원심분리되었다. 잔여 에탄올은 13,000 RPM에서 5 분 동안 추가 원심분리에 의해 제거되었다. DNA는 원심분리에 의해 39 μl Qiagen 완충액 EB에서 용리되었다. 34 μl의 평활 말단 DNA의 dA 테일링이 제조업체의 NEBNext® dA-테일링 모듈에 따라, 클레노브 단편 (3'에서 5'엑소 마이너스)을 내포하는 16 μl의 dA-테일링 마스터 믹스 (NEBNext™ DNA 표본 Prep DNA 시약 세트 1)를 이용하고, 그리고 37℃에서 30 분 동안 항온처리하여 달성되었다. 표본은 4℃로 냉각되고, 그리고 아래와 같이 MinElute PCR 정제 키트 (Qiagen Inc., Valencia, CA)에서 제공된 칼럼을 이용하여 정제되었다. 50 μl 반응물은 1.5 ml 마이크로퓨즈

튜브로 이전되고, 그리고 250 μl의 Qiagen 완충액 PB가 첨가되었다. 300 μl가 MinElute 칼럼으로 이전되었고, 이것은 마이크로퓨즈에서 1 분 동안 13,000RPM에서 원심분리되었다. 칼럼은 750 μl Qiagen 완충액 PE으로 세척되고, 그리고 재원심분리되었다. 잔여 에탄올은 13,000 RPM에서 5 분 동안 추가 원심분리에 의해 제거되었다. DNA는 원심분리에 의해 15 μl Qiagen 완충액 EB에서 용리되었다. 10 마이크로리터의 DNA 용출물이 NEBNext® Quick Ligation 모듈에 따라, 1 μl의 Illumina 유전체 어댑터 올리고 믹스 (부품 번호 1000521)의 1:5 희석액, 15 μl의 2X Quick Ligation 반응 완충액, 그리고 4 μl Quick T4 DNA 연결효소와 함께 25℃에서 15 분 동안 항온처리되었다. 표본은 4℃로 냉각되고, 그리고 아래와 같이 MinElute 칼럼을 이용하여 정제되었다. 150 마이크로리터의 Qiagen 완충액 PE가 30 μl 반응물에 첨가되었고, 그리고 전체 용적은 MinElute 칼럼으로 이전되었고, 이것은 마이크로퓨즈에서 13,000RPM에서 1 분 동안 원심분리되었다. 칼럼은 750 μl Qiagen 완충액 PE으로 세척되고, 그리고 재원심분리되었다. 잔여 에탄올은 13,000 RPM에서 5 분 동안 추가 원심분리에 의해 제거되었다. DNA는 원심분리에 의해 28 μl Qiagen 완충액 EB에서 용리되었다. 23 마이크로리터의 어댑터-결찰된 DNA 용출물은 제조업체의 사용설명서에 따라, Illumina 유전체 PCR 프라이머 (부품 번호 100537과 1000538) 및 NEBNext™ DNA 표본 Prep DNA 시약 세트 1에서 제공된 Phusion HF PCR 마스터 믹스를 이용한 18회 주기의 PCR (30 초 동안 98℃; 10 초 동안 98℃, 30 초 동안 65℃ 및 30 초 동안 72℃의 18회 주기; 5 분 동안 72℃에서 최종 연장, 그리고 4℃에서 유지)에 종속되었다. 증폭된 산물은 www.beckmangenomics.com/products/AMPureXPProtocol_000387v001.pdf에서 가용한 제조업체의 사용설명서에 따라, Agencourt AMPure XP PCR 정제 시스템 (Agencourt Bioscience Corporation, Beverly, MA)을 이용하여 정제되었다. Agencourt AMPure XP PCR 정제 시스템은 통합되지 않은 dNTPs, 프라이머, 프라이머 이합체, 염 및 다른 오염물을 제거하고, 그리고 100 bp보다 큰 앰플리콘을 회수한다. 정제된 증폭된 산물은 40 μl의 Qiagen EB 완충액에서 Agencourt 비드로부터 용리되었고, 그리고 라이브러리의 크기 분포가 2100 바이오분석기 (Agilent technologies Inc., Santa Clara, CA)에 대한 Agilent DNA 1000 키트를 이용하여 분석되었다.

[0440] *c. 단축된 (a) 및 전장 (b) 프로토콜에 따라 제조된 염기서열결정 라이브러리의 분석*

[0441] 바이오분석기에 의해 산출된 전기영동도는 도면 7a와 7b에서 도시된다. 도면 7a는 (a)에서 설명된 전장 프로토콜을 이용하여 혈장 표본 M24228로부터 정제된 cfDNA로부터 제조된 라이브러리 DNA의 전기영동도를 보여주고, 그리고 도면 7b는 (b)에서 설명된 전장 프로토콜을 이용하여 혈장 표본 M24228로부터 정제된 cfDNA로부터 제조된 라이브러리 DNA의 전기영동도를 보여준다. 양쪽 도면에서, 피크 1과 4는 각각, 15 bp 아래쪽 마커 및 1,500 위쪽 마커를 나타낸다; 피크 위에 숫자는 라이브러리 단편에 대한 이주 시간을 지시한다; 그리고 수평선은 통합에 대한 설정 역치를 지시한다. 도면 7a에서 전기영동도는 187 bp의 단편의 마이너 피크 및 263 bp의 단편의 주요 피크를 보여주고, 반면 도면 7b에서 전기영동도는 265 bp에서 단지 하나만 피크만 보여준다. 피크 구역의 통합은 도면 7a에서 187 bp 피크의 DNA에 대한 0.40 ng/μl의 계산된 농도, 도면 7a에서 263bp 피크의 DNA에 대한 7.34 ng/μl의 농도, 그리고 도면 7b에서 265 bp 피크의 DNA에 대한 14.72 ng/μl의 농도를 유발하였다. cfDNA에 결찰되었던 Illumina 어댑터는 92 bp것으로 알려져 있는데, 이것은 265 bp로부터 감산될 때, cfDNA의 피크 크기가 173 bp라는 것을 지시한다. 187 bp에서 마이너 피크가 끝과 끝 결찰되었던 두 프라이머의 단편을 나타내는 것이 가능하다. 선형 2-프라이머 단편은 단축된 프로토콜이 이용될 때, 최종 라이브러리 산물로부터 제거된다. 단축된 프로토콜은 또한, 187 bp보다 적은 다른 작은 단편을 제거한다. 본 실시예에서, 정제된 어댑터-결찰된 cfDNA의 농도는 전장 프로토콜을 이용하여 생산된 어댑터-결찰된 cfDNA의 농도의 2배이다. 어댑터-결찰된 cfDNA 단편의 농도는 전장 프로토콜을 이용하여 획득된 것보다 항상 큰 것으로 알려졌다 (데이터 제시되지 않음).

[0442] 따라서, 단축된 프로토콜을 이용하여 염기서열결정 라이브러리를 제조하는 이점은 획득된 라이브러리가 262-267 bp 범위에서 단지 하나의 주요 피크만 일관되게 포함하는 반면, 전장 프로토콜을 이용하여 제조된 라이브러리의 품질이 cfDNA를 나타내는 것 이외에 피크의 숫자와 이동성에 의해 반영되는 바와 같이 변한다는 점이다. 비-cfDNA 산물은 흐름 셀 상에서 공간을 점유하고, 그리고 이수성 상태의 전반적인 배정의 기초가 되는 염기서열결정 반응의 클러스터 증폭 및 차후 영상의 품질을 떨어뜨릴 것이다. 단축된 프로토콜은 라이브러리의 염기서열결정에 영향을 주지 않는 것으로 나타났다.

[0443] 단축된 프로토콜을 이용하여 염기서열결정 라이브러리를 제조하는 다른 이점은 평활 말단화, d-A 테일링, 그리고 어댑터-결찰의 3가지 효소적 단계가 신속한 이수체 진단 서비스의 검증과 실행을 위한 뒷받침을 완결하는데 1 시간 보다 적게 소요된다는 것이다.

[0444] 다른 이점은 평활 말단화, d-A 테일링, 그리고 어댑터 결찰의 3 가지 효소적 단계가 동일한 반응 튜브에서 수행되고, 따라서 물질의 상실, 그리고 더욱 중요하게는, 가능한 표본 뒤섞임 및 표본 오염을 잠재적으로 야기할 복

수 표본 전달을 방지한다는 것이다.

[0445] **실시예 2**

[0446] **쌍둥이 임신에서 정확한 이수성 검출**

[0447] *도입*

[0448] 전체-유전체 대량으로 병렬 염기서열결정을 이용한 전체 무세포 DNA (cfDNA)의 비침습성 출생전 시험 (NIPT)은 태아 염색체 이수성을 검출하는 매우 정확하고 견실한 방법인 것으로 나타났다. Bianchi DW, Platt LD, Goldberg JD, et al. Genome-wide fetal aneuploidy detection by maternal plasma DNA sequencing. Obstet Gynecol 2012;119:890-901; Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A 2008;105:16266-71; Sehnert AJ, Rhees B, Comstock D, et al. Optimal detection of fetal chromosomal abnormalities by massively parallel DNA sequencing of cell-free fetal DNA from maternal blood. Clin Chem 2011;57:1042-9를 참조한다. 본 시험은 단일 모체 혈액 표본으로부터 삼염색체성 21, 18, 13 및 성염색체 이수성을 검출한다. 본 시험은 현재, 10+ 주에서, 그리고 태아 이수성에 대한 높은 위험에서 단태 임신한 임신 여성에 대해 필요하다. 최근에, American College of Obstetricians and Gynecologists (ACOG), International Society for Prenatal Diagnosis (ISPD), American College of Medical Genetics and Genomics (ACMG) 및 National Society of Genetic Counselors (NSGC)는 태아 이수성의 높은 위험을 갖는 여성의 경우에 NIPT의 이용을 고려하도록 권장하였다.

[0449] 미국에서, 쌍둥이는 30건의 생존 출산 중에서 대략 1건을 차지하고, 그리고 쌍둥이 출산율이 증가하고 있다 (National Center for Health Statistics Data Brief, No. 80, January 2012). 여성이 노화함에 따라서, 이들은 월경 주기마다 하나 이상의 난자를 방출할 개연성이 더욱 높고, 그리고 따라서, 30세 이상의 여성은 쌍둥이 임신에서 증가의 약 1/3을 차지한다. 종종 하나 이상의 배아가 시험관내 수정 동안 이전되는 보조 생식 기술은 쌍둥이 임신에서 나머지 증가의 대부분을 차지한다.

[0450] 예비적 증거는 모계 순환 내에 존재하는 태아 DNA의 양이 단태 임신과 비교할 때 쌍둥이 임신에서 대략 35% 증가한다는 것을 암시하지만, 이 연구는 각 태아로부터 유래된 cfDNA의 양을 자세히 살피지 않았다. Canick JA, Kloza EM, Lambert-Messerlian GM, et al. DNA sequencing of maternal plasma to identify Down syndrome and other trisomies in multiple gestations. Prenat Diagn 2012;32:730-4. 연구자들은 쌍둥이 임신에서 순환하는 태아 DNA의 양에서 전반적인 증가가 있긴 하지만, 각 태아에 대한 cfDNA의 양이 감소한다는 것을 증명하였다. Srinivasan A, Bianchi D, Liao W, Sehnert A, Rava R. 52: Maternal plasma DNA sequencing: effects of multiple gestation on aneuploidy detection and the relative cell-free fetal DNA (cffDNA) per fetus. American journal of obstetrics and gynecology 2013;208:S31. Srinivasan A, Bianchi DW, Huang H, Sehnert AJ, Rava RP. Noninvasive detection of fetal subchromosome abnormalities via deep sequencing of maternal plasma. American journal of human genetics 2013;92:167-76. 이런 이유로, 쌍둥이 임신에서 이수성의 정확한 분류를 담보하는 민감한 방법이 필요하다.

[0451] 이수성 표본을 정확하게 분류하는 NIPT의 능력을 최대화하는 인자는 통계학적 잡음이 최소화되도록 분석에서 이용된 서열 리드의 개수에서 증가, 그리고 실행간 가변성이 감소되도록 염색체 신호를 정규화하는 능력이다. 최근에, 출원인은 표본당 이용가능한 리드의 개수를 증가시키는 향상된, 자동화된 표본 제조 작업 흐름 및 이수체 염색체의 특정한 신호를 증가시키는 향상된 분석법을 개발하였다. 이들 증강은 이수체 영향을 받은 표본을 분류하는 전반적인 정확도를 향상시킨다.

[0452] 본 실시예는 현재까지 이용된 가장 큰 쌍둥이 검증 코호트에 향상된 분류 알고리즘의 적용을 설명한다. 우리는 향상된 SAFeR (태아 결과에 대한 선택적 알고리즘) 알고리즘이 태아마다 감소된 양의 무세포 DNA를 갖는 것으로 알려져 있는 쌍둥이 표본에서 정확한 이수성 검출을 허용한다는 것을 증명한다.

[0453] *방법*

[0454] 표본은 높은 위험 및 평균 위험 모계 개체군 둘 모두를 수반하는 2가지 독립된 임상적 연구의 일부로서 수집되었다. MatErnal BLood IS Source to Accurately Diagnose Fetal Aneuploidy 연구 (MELISSA; NCT01122524)는 높은 위험 임신에서 전체 염색체 이수성을 검출하도록 설계되었다. Bianchi DW, Platt LD, Goldberg JD, et al. Genome-wide fetal aneuploidy detection by maternal plasma DNA sequencing. Obstet Gynecol 2012;119:890-901. Comparison of Aneuploidy Risk Evaluations 시험 (CARE; NCT01663350)은 평균 위험 모계

개체군 (공개를 위해 제출됨)에서 삼염색체성 21 및 삼염색체성 18에 대한 전통적인 출생전 혈청 선별검사 방법과 비교하여, 본 시험의 우수한 특이성을 증명하도록 설계되었다. 데이터세트의 상세는 표 3에서 도시된다. 임상적 결과는 출생전 침습성 시술로부터 핵형에 의해 또는 신생아 신체 검사에 의해 결정되었다.

표 3

[0455] **쌍둥이 표본의 핵형 분류 및 본 발명 분류.** 118건의 쌍둥이 임신으로부터 모계 표본은 염색체 21, 18과 13의 이수성에 대해 및 Y 염색체의 존재에 대해 본 발명 출생전 시험을 이용하여 분석되었다. 본 데이터는 핵형 분석 또는 신생 신체 검사에 의해 결정된 임상적 결과와 비교되었다.

연구된 숫자	태아 1	태아 2	본 발명 이수성 분류	본 발명 염색체 Y 분류
24	46,XX	46,XX	영향받지 않음	검출되지 않음
48	46,XX	46,XY	영향 받지 않음	Y 검출됨
42	46,XY	46,XY	영향 받지 않음	Y 검출됨
2	47,XY,+21	46,XY	T21 영향 받음	Y 검출됨
1	Mos 47,XY,+21[7]/46,XY[11]	46,XX	T21 영향 받음	Y 검출됨
1	47,XY,+ 18	47,XY,+18	T18 영향 받음	Y 검출됨

[0456] 무세포 DNA는 동결된 혈장 표본으로부터 추출되고 앞서 설명된 바와 같이 HiSeq2000 서열분석기에서 염기서열결정되었다. Sehnert AJ, Rhees B, Comstock D, et al. Optimal detection of fetal chromosomal abnormalities by massively parallel DNA sequencing of cell-free fetal DNA from maternal blood. Clin Chem 2011;57:1042-9. 대량 병렬 염기서열결정 (MPS) 서열 태그가 인간 유전체 참조 빌드 hg19에 매핑되었고, 그리고 정규화된 염색체 값 (NCVs)이 신호 대 잡음 비율을 최대화하고 검출의 전반적인 감수성을 향상시킨 향상된 분석적 작업 흐름을 이용하여 염색체 21, 18, 13, X와 Y에 대해 계산되었다. 알고리즘 성분은 향상된 유전체 필터링, 분자생물학 단계를 통해 도입된 체계적인 바이어스의 제거 및 향상된 정규화와 분류 방법을 포함하였다. 염기서열결정을 수행한 실험실 인원은 임상적 결과에 맹검이었다.

[0457] *결과*

[0458] 임상적 확진 결과를 갖는 118개 쌍둥이 임신으로부터 모계 혈장 표본이 본 연구에서 조사되었다 (표 3). 염색체 21, 18과 13에 대한 이수성 분류가 연구 중인 모든 표본에 대해 산출되었고, 그리고 하나 또는 그 이상의 이수성 태아의 임신으로부터 4개 표본이 정확하게 확인되었다 (도면 8). 이들 표본 중에서 2개는 각각 1개의 T21 영향을 받은 수컷 태아 및 1개의 영향을 받지 않은 수컷 태아를 갖는 두융모막 쌍둥이 쌍 (47,XY+21/46,XY)으로부터 유래되었다; 1개는 47,XY+18 핵형을 갖는 단일융모막 쌍둥이 표본이었다; 그리고 1개의 표본은 한 쌍둥이가 모자이크 핵형 47,XY+T21[7]/46,XY[11]을 갖는 두융모막 쌍둥이이었다. 본 연구에서 임상적으로-규정된 영향을 받지 않은 표본 (N=114) 중에서 어느 것도 이수성에 대해 영향을 받는 것으로서 분류되지 않았다.

[0459] 태아의 성별은 cfDNA에서 Y 염색체의 존재에 의해 결정될 수 있다. 본원에서 개시된 시험은 최소한 하나의 수컷 태아를 갖는 모든 표본에서 Y 염색체의 존재를 양성으로 확인할 수 있었다 (도면 8). 게다가, 상기 시험은 또한, 2개의 암컷 태아를 갖는 표본에서 Y 염색체의 부재를 정확하게 확인하였다.

[0460] *결론*

[0461] 본 연구는 쌍둥이 표본의 가장 민감한 상염색체 이수성 시험을 할 수 있게 하는 향상된 분석법을 증명한다. 증강된 분석법은 유전체 필터링에서 향상, 체계적인 잡음 감소 및 향상된 분류 방법을 이용한다. 향상된 분석적 작업 흐름의 유용성은 한 세트의 118개 쌍둥이 표본; 쌍둥이에서 상염색체 이수성 및 Y 염색체의 존재를 검출하기 위한 MPS의 임의의 검증에서 이용된 표본의 가장 큰 숫자에서 증명되었다 (도면 9). 도면 9는 NIPT 연구에서 분석된 쌍둥이 표본을 보여준다. 상업적으로 가용한 NIPT 시험의 성과를 사정하기 위한 다양한 연구에서 이용된 쌍둥이 표본의 숫자. Canick JA, Kloza EM, Lambert-Messerlian GM, et al. DNA sequencing of maternal plasma to identify Down syndrome and other trisomies in multiple gestations. Prenat Diagn 2012;32:730-4. Lau TK, Jiang F, Chan MK, Zhang H, Lo PSS, Wang W. Non-invasive prenatal screening of fetal Down syndrome by maternal plasma DNA sequencing in twin pregnancies. Journal of Maternal-Fetal and Neonatal Medicine 2013;26:434-7. 향상된 분석법은 임의의 가양성 결과를 산출하지 않으면서, 삼염색체성 21에 대해 모

자이크인 영향을 받은 태아를 포함하는 코호트에서 모든 삼염색체성 21과 삼염색체성 18 표본의 존재를 정확하게 검출함으로써 정밀하게 수행하는 것으로 나타났다. 추가적으로, 향상된 분석법은 최소한 하나의 수컷 태아를 갖는 모든 쌍둥이 임신에서 Y 염색체의 존재를 정확하게 검출하였고, 그리고 2개의 암컷 태아를 갖는 어떤 쌍둥이 임신에서도 Y 염색체를 검출하지 못하였다.

[0462] 민감한 방법의 한 가지 특징은 체계적인 잡음을 최소화하고 전반적인 신호 대 잡음 비율을 증가시키는 능력이다. 본 연구는 임의의 다른 상업적으로-가용한 NIPT 검정보다 많은 표본당 서열 리드 (대략 28M 서열 리드/표본)를 생산함으로써, 그리고 복합적 DNA 표본의 생화학적 조작을 동행하는 체계적인 잡음을 더욱 우수하게 처리하도록 분석법을 향상시킴으로써, 이것을 달성하였다. 향상된 분석적 작업 흐름은 궁극적으로, 정규화된 염색체 수치 분포의 폭을 감소시켜, 영향을 받지 않은 개체군 및 영향을 받은 개체군의 더욱 우수한 분리 및 낮은 양의 태아 DNA를 갖는 이수성 영향을 받은 태아를 정확하게 확인하는 향상된 능력을 허용한다.

[0463] 쌍둥이 임신에서 이수성을 검출하는 매우 정확하고 민감한 방법을 갖는 능력이 중요한데, 그 이유는 비록 무세포 태아 DNA의 총량이 쌍둥이 임신에서 증가하지만, 각 태아에 기인한 양이 감소하기 때문이다. 이런 이유로, A) 마치 시험 표본이 단태 임신에 동등한 것처럼 이러한 발견 및 시험 표본을 무시하고, 가음성 결과의 가능성을 증가시키거나, B) 불충분한 DNA로 인해 표본의 증가된 숫자를 거부하거나 또는 C) 더욱 민감한 방법을 구축할 수 있다 (표 4).

표 4

[0464] **상업적으로 가용한 NIPT 시험을 이용하여 쌍둥이 임신을 처리하기 위한 전략**

	옵션		결과
A	마치 존재하는 cfDNA가 단태 임신과 동일한 것처럼 쌍둥이 임신을 시험한다.		가음성의 증가된 가능성.
B	쌍둥이 임신을 시험하기 위해 현재 방법론을 이용한다.		불충분한 DNA로 인해 표본을 거부한다
C	개별 cfDNA 농도에 더욱 민감한 향상된 방법을 이용한다		더욱 적은 가음성에서 쌍둥이와 낮은 수준 단태에 대한 더욱 정확한 시험.

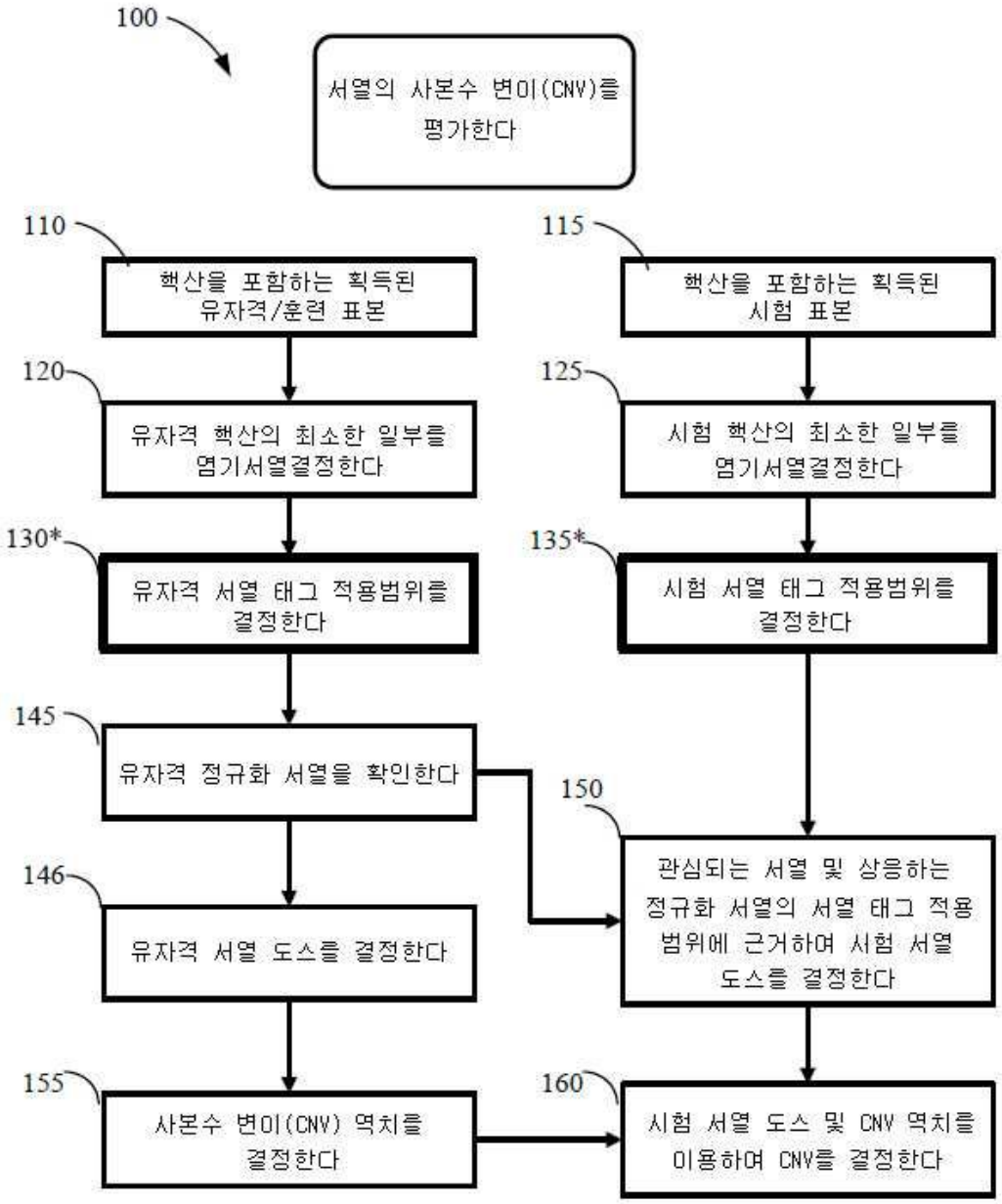
[0465] SAFeR™ 알고리즘에 분석적 향상은 쌍둥이 임신에서 정확한 이수성 분류를 실시가능하게 하는 것 너머까지 미친다. 영향을 받지 않은 개체군 및 영향을 받은 개체군의 향상된 분리는 또한, 이수성 의심됨으로서 분류되는 표본의 전체 빈도를 감소시킨다. 추가적으로, 향상된 분석적 작업 흐름이 이수성 검출 및 성별 분류에서 유사한 향상으로 단태 임신에 적용될 수 있다.

[0466] 결론적으로, 본 연구는 이수성 영향을 받지 않은 표본 및 영향을 받은 표본의 더욱 우수한 분리, 그리고 낮은 양의 태아 DNA를 내포하는 표본으로부터 더욱 정확한 상염색체 이수성 분류를 야기하는 향상된 분석법을 설명한다. 이들 향상을 통합함으로써, 출생전 시험의 능력이 쌍둥이 임신을 시험하는 데까지 확대되었다.

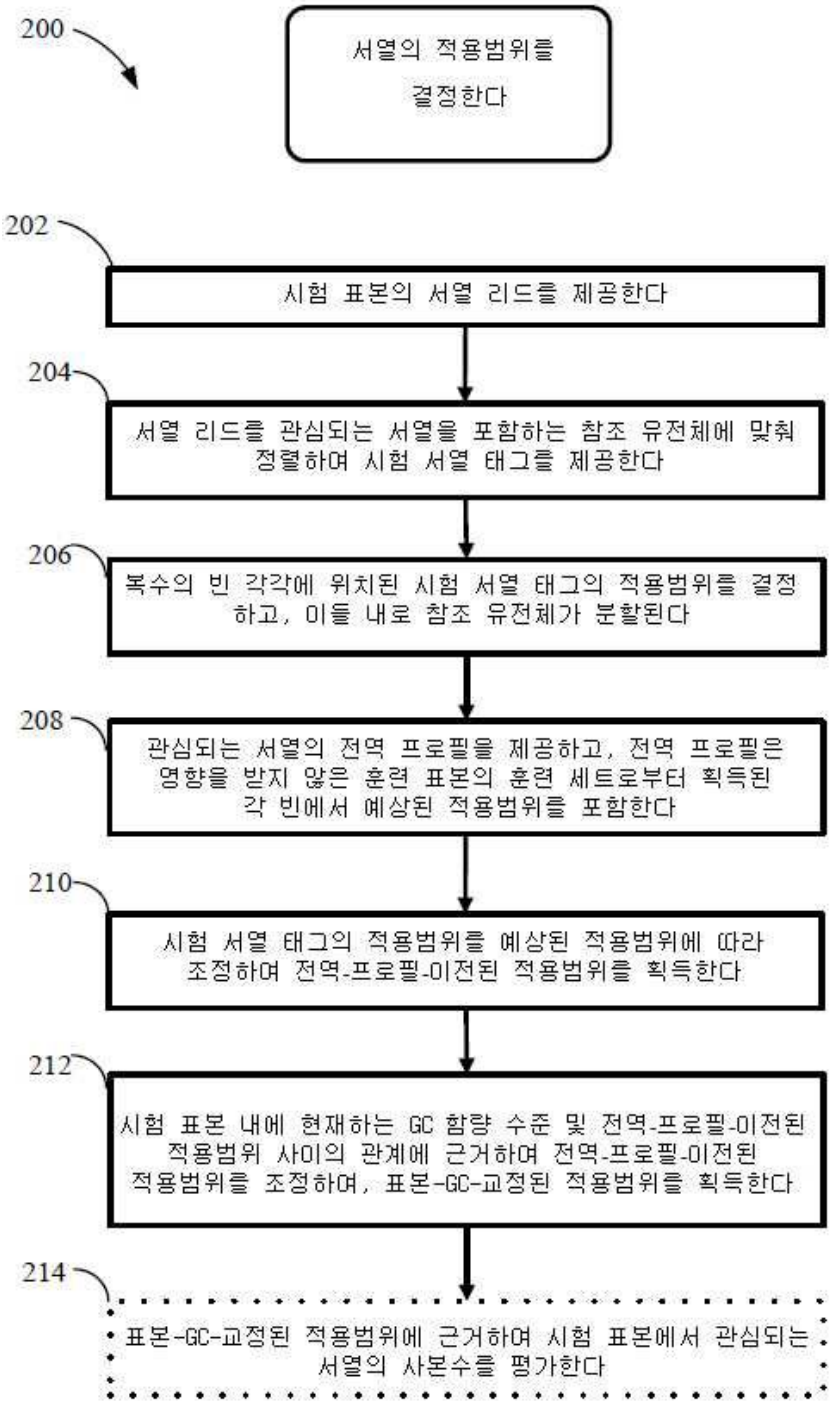
[0467] 본 발명은 이의 사상 또는 필수적인 특징으로부터 벗어나지 않으면서 다른 특정한 형태에서 구현될 수 있다. 설명된 구체예는 모든 양상에서 단지 예시에 불과하고 제한하지 않는 것으로 고려된다. 발명의 범위는 이런 이유로, 전술한 설명보다는 첨부된 청구항에 의해 지시된다. 청구항의 등가의 의미와 범위 안에 있는 모든 변화는 범위 내에 포섭된다.

도면

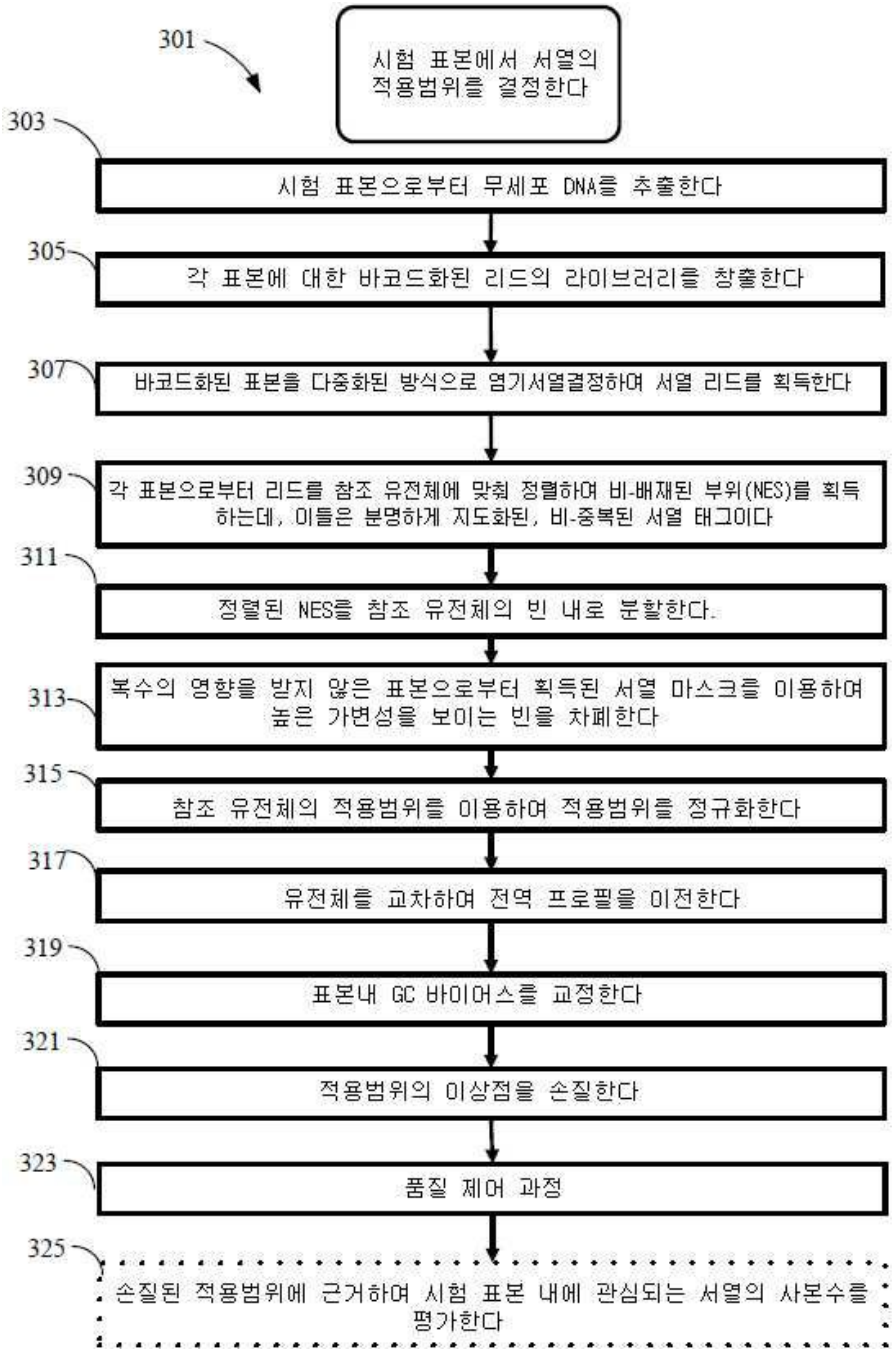
도면1


100
서열의 사본수 변이(CNV)를 평가한다
110
핵산을 포함하는 획득된 유자격/훈련 표본
115
핵산을 포함하는 획득된 시험 표본
120
유자격 핵산의 최소한 일부를 염기서열결정한다
125
시험 핵산의 최소한 일부를 염기서열결정한다
130*
유자격 서열 태그 적용범위를 결정한다
135*
시험 서열 태그 적용범위를 결정한다
145
유자격 정규화 서열을 확인한다
150
관심되는 서열 및 상응하는 정규화 서열의 서열 태그 적용 범위에 근거하여 시험 서열 도스를 결정한다
146
유자격 서열 도스를 결정한다
155
사본수 변이(CNV) 역치를 결정한다
160
시험 서열 도스 및 CNV 역치를 이용하여 CNV를 결정한다

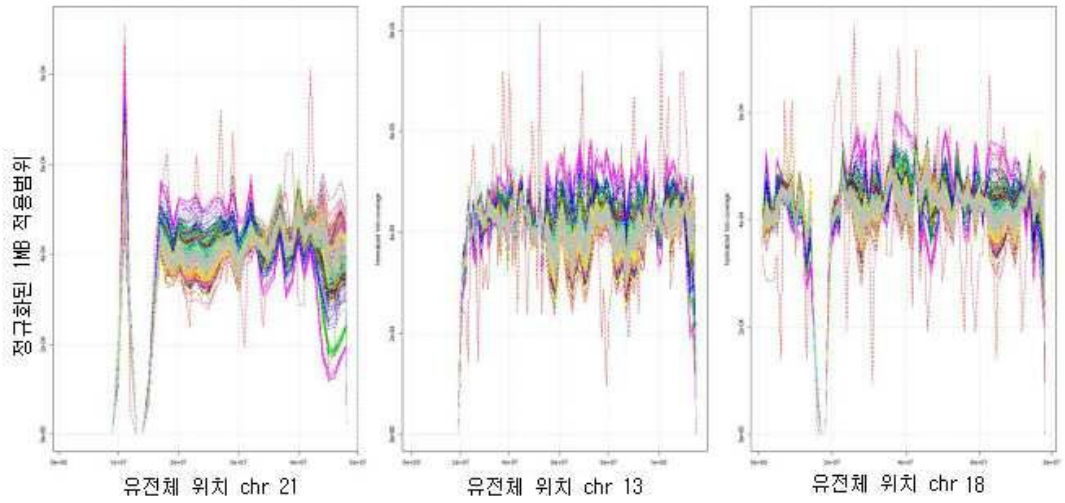
도면2


200
서열의 적용범위를 결정한다
202
시험 표본의 서열 리드를 제공한다
204
서열 리드를 관심되는 서열을 포함하는 참조 유전체에 맞춰 정렬하여 시험 서열 태그를 제공한다
206
복수의 빈 각각에 위치된 시험 서열 태그의 적용범위를 결정하고, 이들 내로 참조 유전체가 분할된다
208
관심되는 서열의 전역 프로필을 제공하고, 전역 프로필은 영향을 받지 않은 훈련 표본의 훈련 세트로부터 획득된 각 빈에서 예상된 적용범위를 포함한다
210
시험 서열 태그의 적용범위를 예상된 적용범위에 따라 조정하여 전역-프로필-이전된 적용범위를 획득한다
212
시험 표본 내에 현재하는 GC 함량 수준 및 전역-프로필-이전된 적용범위 사이의 관계에 근거하여 전역-프로필-이전된 적용범위를 조정하여, 표본-GC-교정된 적용범위를 획득한다
214
표본-GC-교정된 적용범위에 근거하여 시험 표본에서 관심되는 서열의 사본수를 평가한다

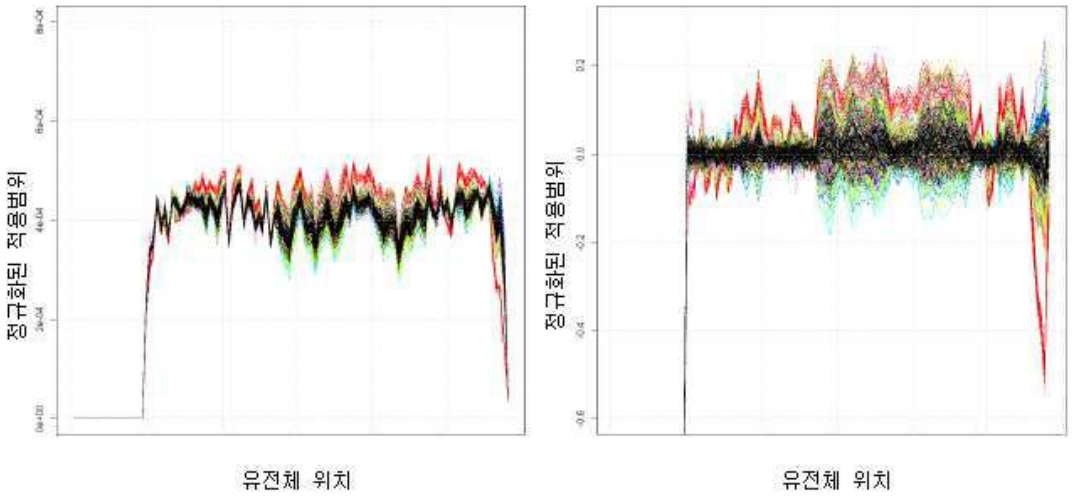
도면3a


301
시험 표본에서 서열의 적용범위를 결정한다
303
시험 표본으로부터 무세포 DNA를 추출한다
305
각 표본에 대한 바코드화된 리드의 라이브러리를 창출한다
307
바코드화된 표본을 다중화된 방식으로 염기서열결정하여 서열 리드를 획득한다
309
각 표본으로부터 리드를 참조 유전체에 맞춰 정렬하여 비-배재된 부위(NES)를 획득하는데, 이들은 분명하게 지도화된, 비-중복된 서열 태그이다
311
정렬된 NES를 참조 유전체의 빈 내로 분할한다.
313
복수의 영향을 받지 않은 표본으로부터 획득된 서열 마스크를 이용하여 높은 가변성을 보이는 빈을 차폐한다
315
참조 유전체의 적용범위를 이용하여 적용범위를 정규화한다
317
유전체를 교차하여 전역 프로필을 이전한다
319
표본내 GC 바이어스를 교정한다
321
적용범위의 이상점을 손질한다
323
품질 제어 과정
325
손질된 적용범위에 근거하여 시험 표본 내에 관심되는 서열의 사본수를 평가한다

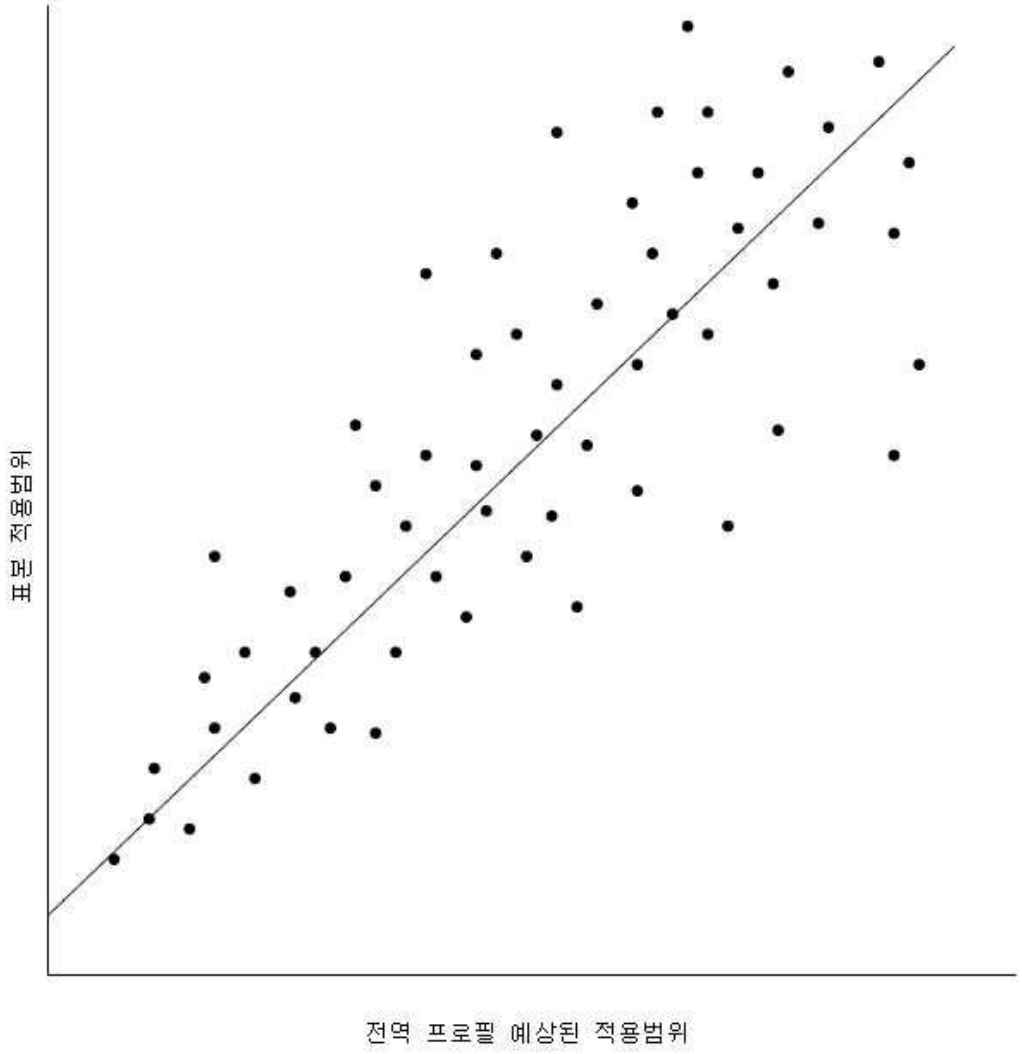
도면3b


정규화된 1MB 적용범위
유전체 위치 chr 21
유전체 위치 chr 13
유전체 위치 chr 18

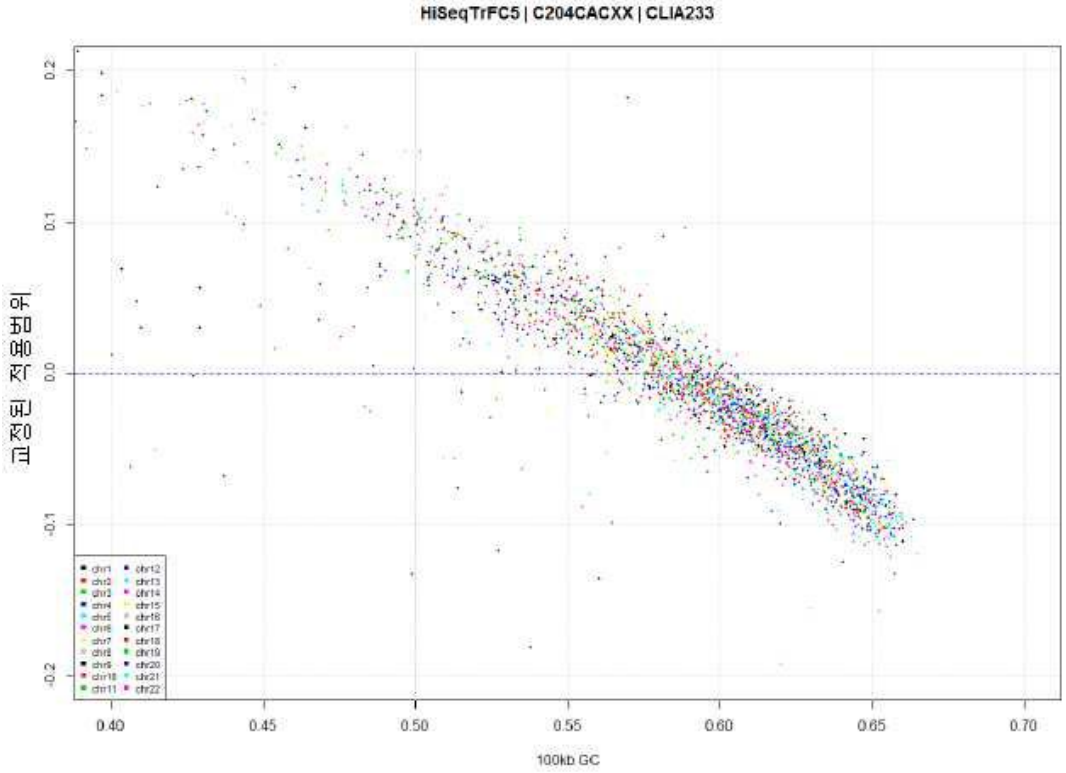
도면3c


정규화된 적용범위
유전체 위치
정규화된 적용범위
유전체 위치

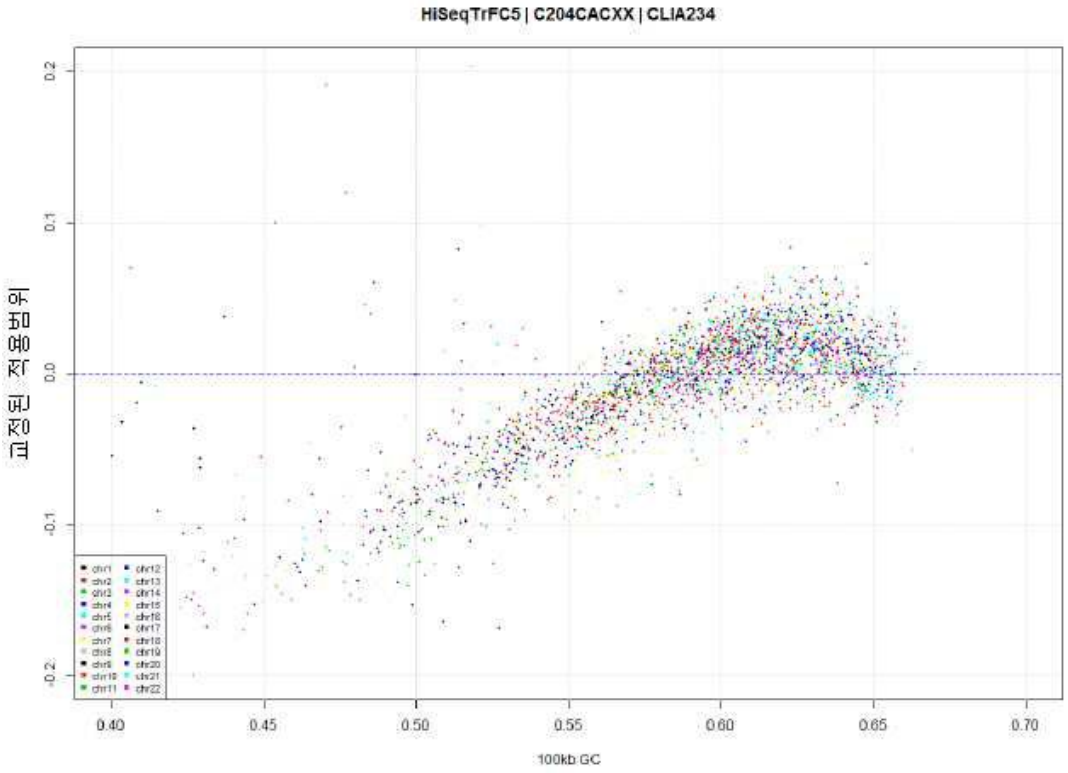
도면3d



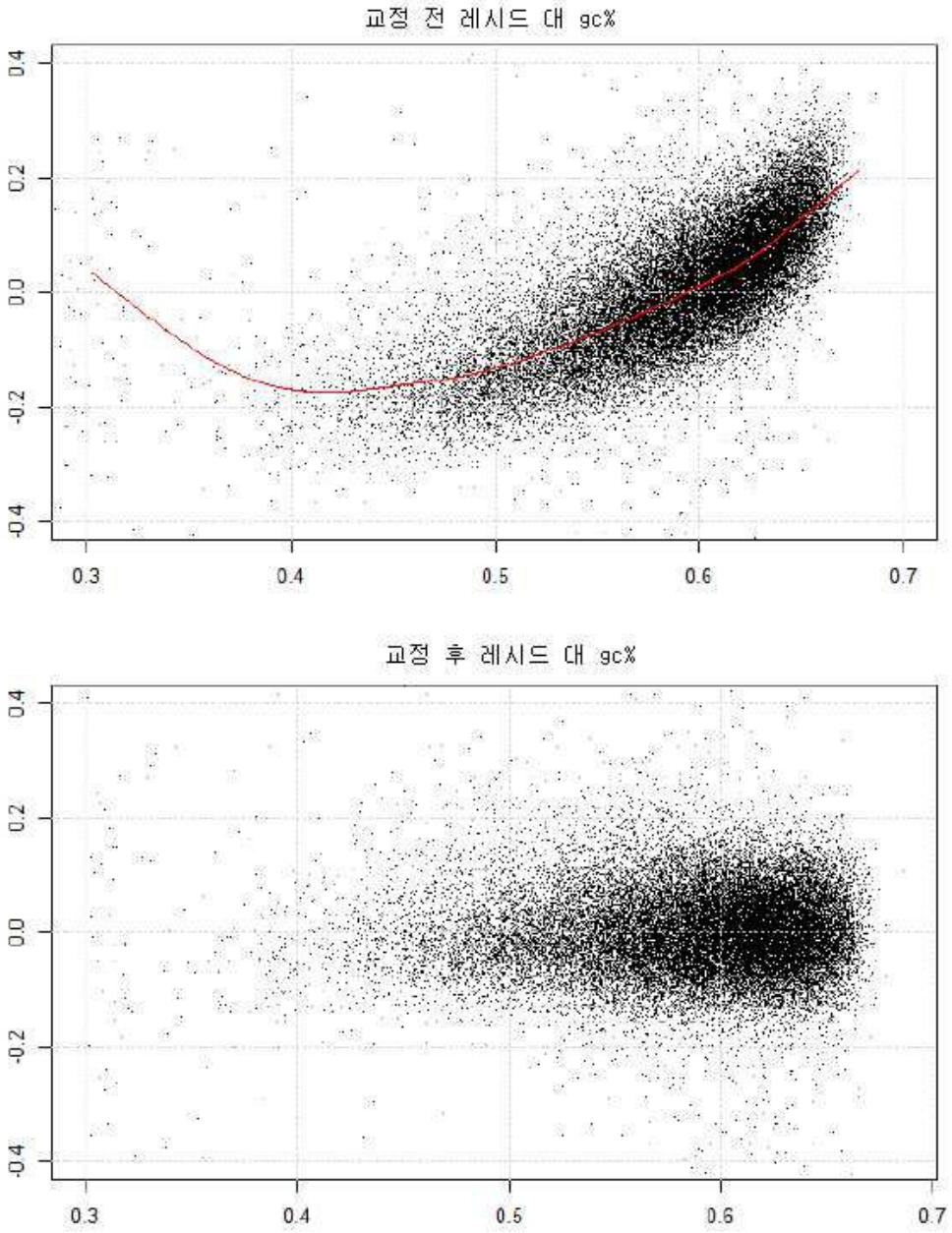
도면3e


HiSeqTrFC5 | C204CACXX | CLIA233
교정된 적용범위
100kb GC
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.2
0.1
0.0
-0.1
-0.2

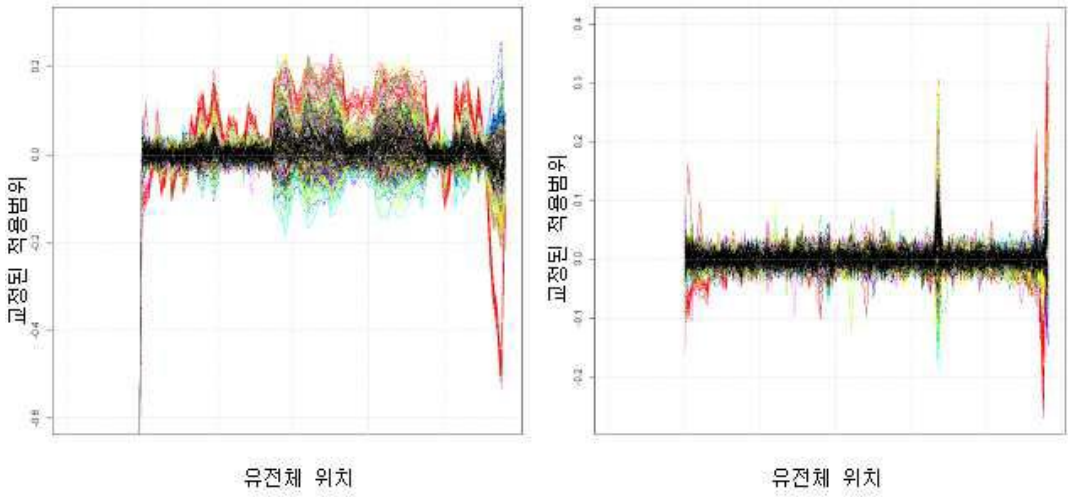
도면3f


HiSeqTrFC5 | C204CACXX | CLIA234
교정된 적용범위
100kb GC
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.2
0.1
0.0
-0.1
-0.2

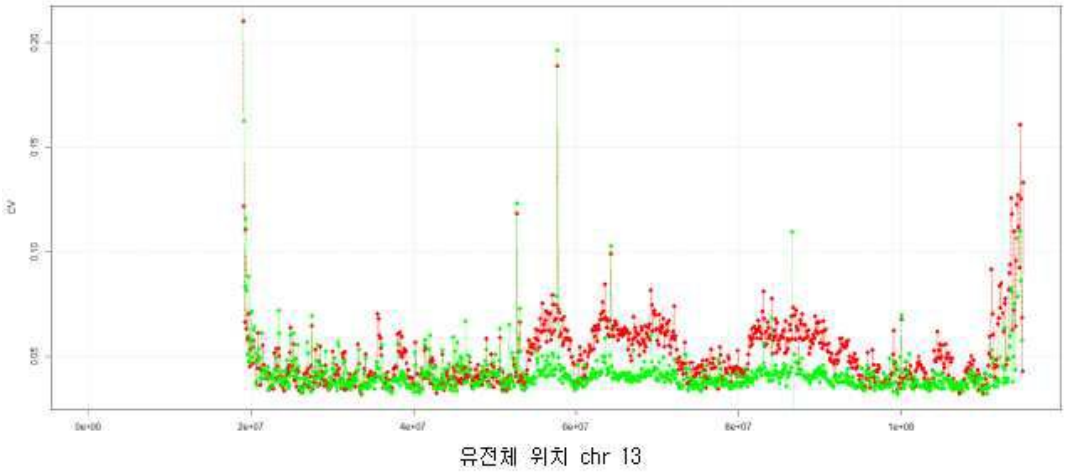
도면3g


교정 전 레시드 대 gc%
0.4
0.2
0.0
-0.2
-0.4
0.3
0.4
0.5
0.6
0.7
교정 후 레시드 대 gc%
0.4
0.2
0.0
-0.2
-0.4
0.3
0.4
0.5
0.6
0.7

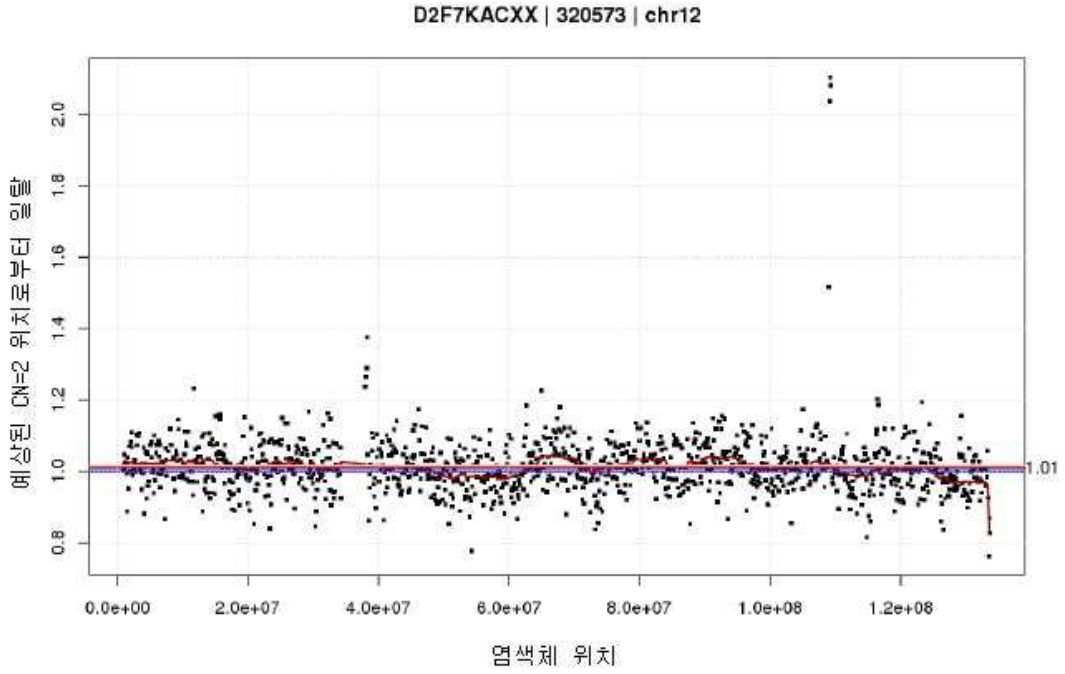
도면3h



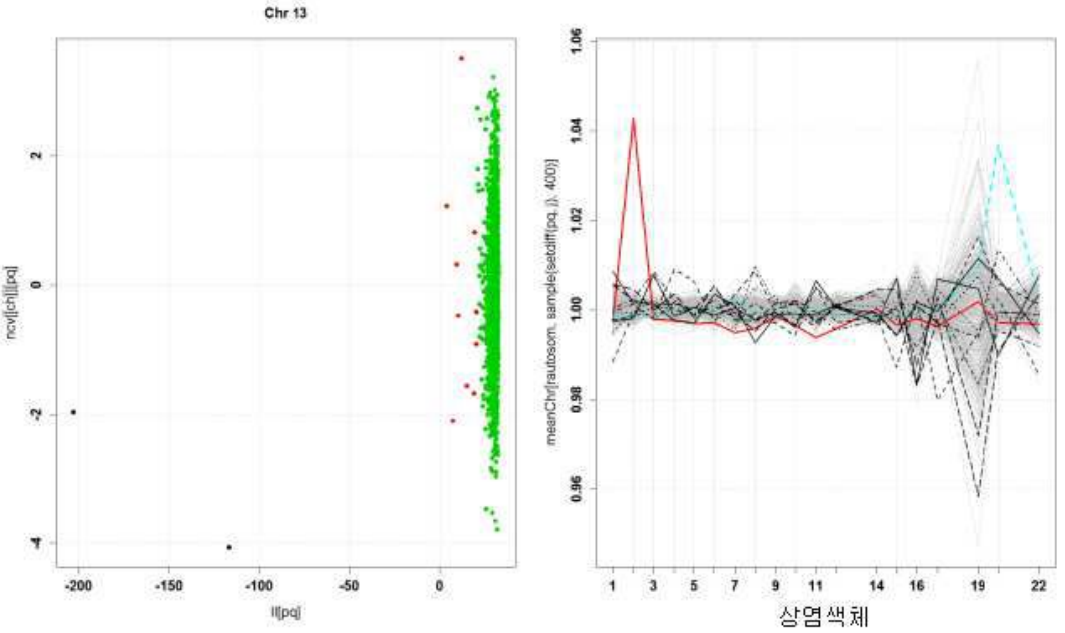
도면3i



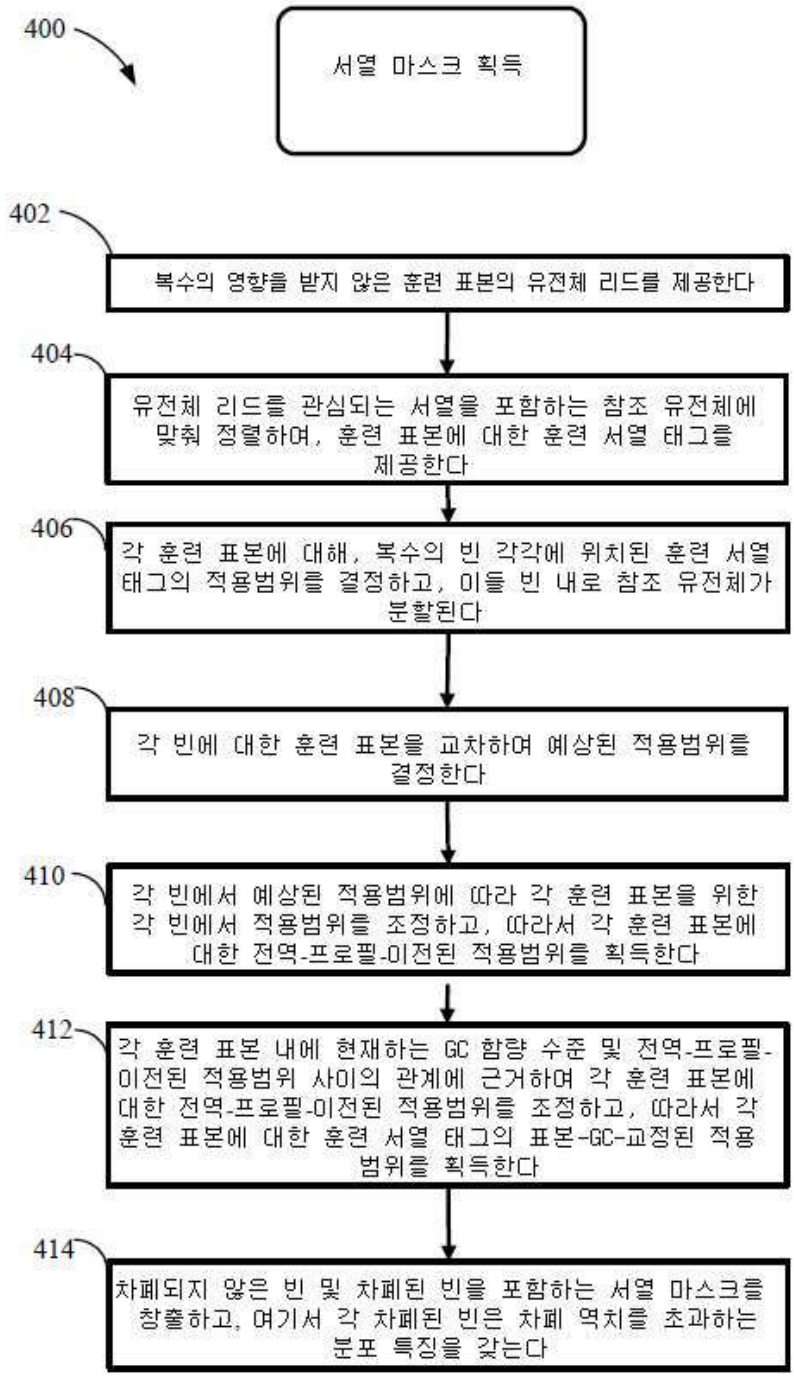
도면3j


D2F7KACXX | 320573 | chr12
예상된 CN=2 위치로부터 일탈
0.8
1.0
1.2
1.4
1.6
1.8
2.0
1.01
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
염색체 위치

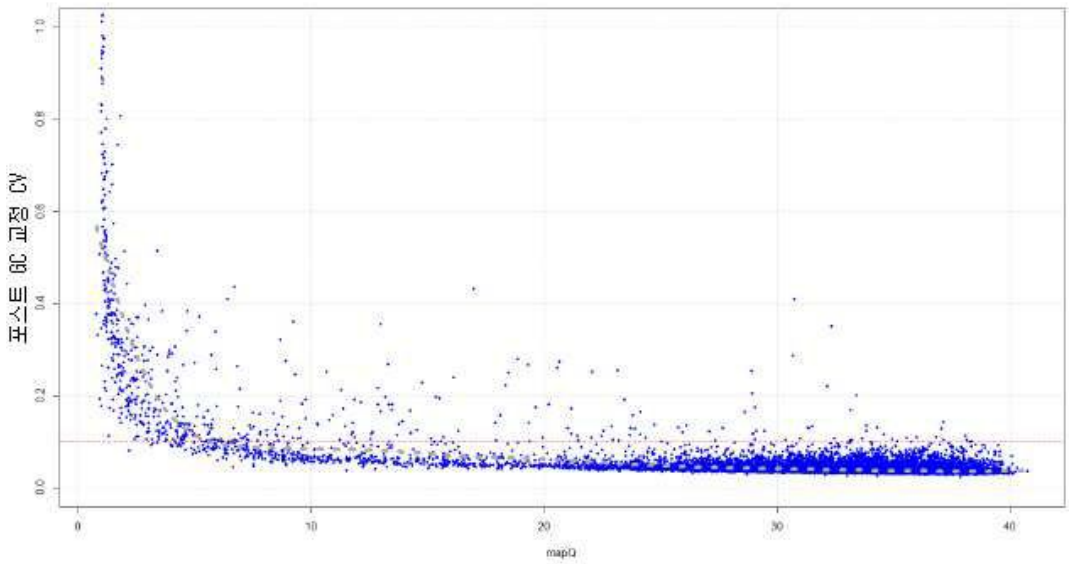
도면3k


Chr 13
nov[ch1][pq]
2
0
-2
-4
-200
-150
-100
-50
0
ll[pq]
meanChr[autosom. sample[setdiff[pq, j], 400]]
1.06
1.04
1.02
1.00
0.98
0.96
1
3
5
7
9
11
14
16
19
22
상염색체

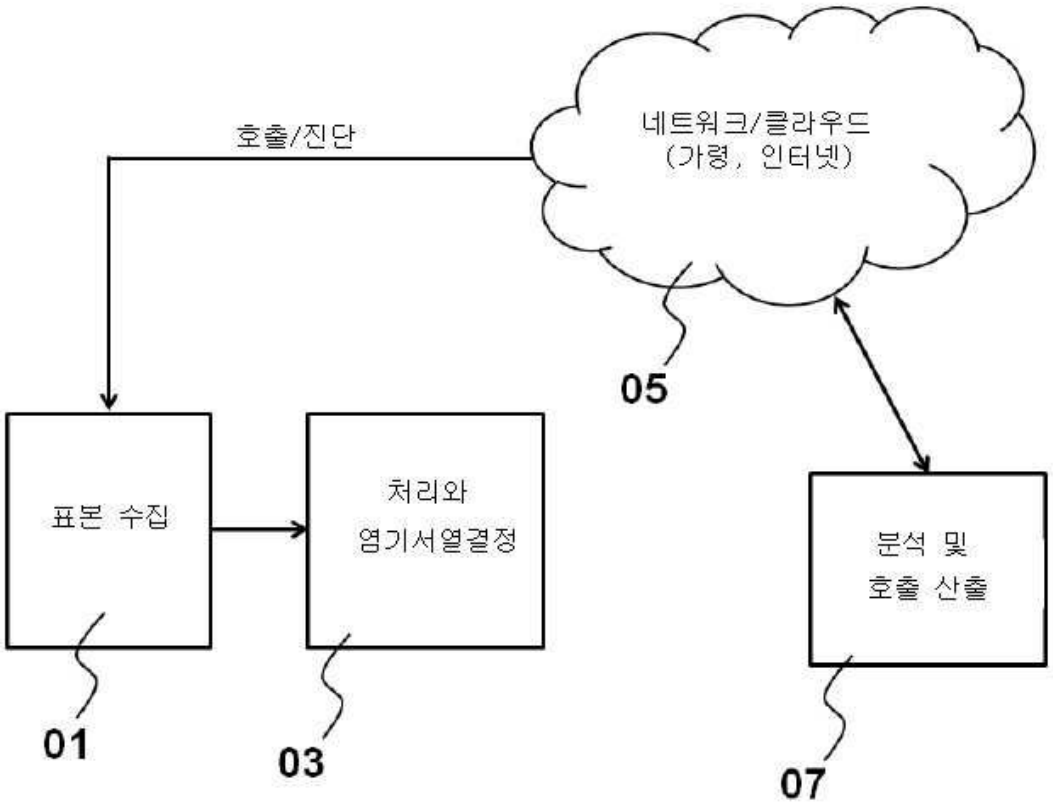
도면4a


400
서열 마스크 획득
402
복수의 영향을 받지 않은 훈련 표본의 유전체 리드를 제공한다
404
유전체 리드를 관심되는 서열을 포함하는 참조 유전체에 맞춰 정렬하여, 훈련 표본에 대한 훈련 서열 태그를 제공한다
406
각 훈련 표본에 대해, 복수의 빈 각각에 위치된 훈련 서열 태그의 적용범위를 결정하고, 이들 빈 내로 참조 유전체가 분할된다
408
각 빈에 대한 훈련 표본을 교차하여 예상된 적용범위를 결정한다
410
각 빈에서 예상된 적용범위에 따라 각 훈련 표본을 위한 각 빈에서 적용범위를 조정하고, 따라서 각 훈련 표본에 대한 전역-프로필-이전된 적용범위를 획득한다
412
각 훈련 표본 내에 현재하는 GC 함량 수준 및 전역-프로필-이전된 적용범위 사이의 관계에 근거하여 각 훈련 표본에 대한 전역-프로필-이전된 적용범위를 조정하고, 따라서 각 훈련 표본에 대한 훈련 서열 태그의 표본-GC-교정된 적용범위를 획득한다
414
차폐되지 않은 빈 및 차폐된 빈을 포함하는 서열 마스크를 창출하고, 여기서 각 차폐된 빈은 차폐 역치를 초과하는 분포 특징을 갖는다

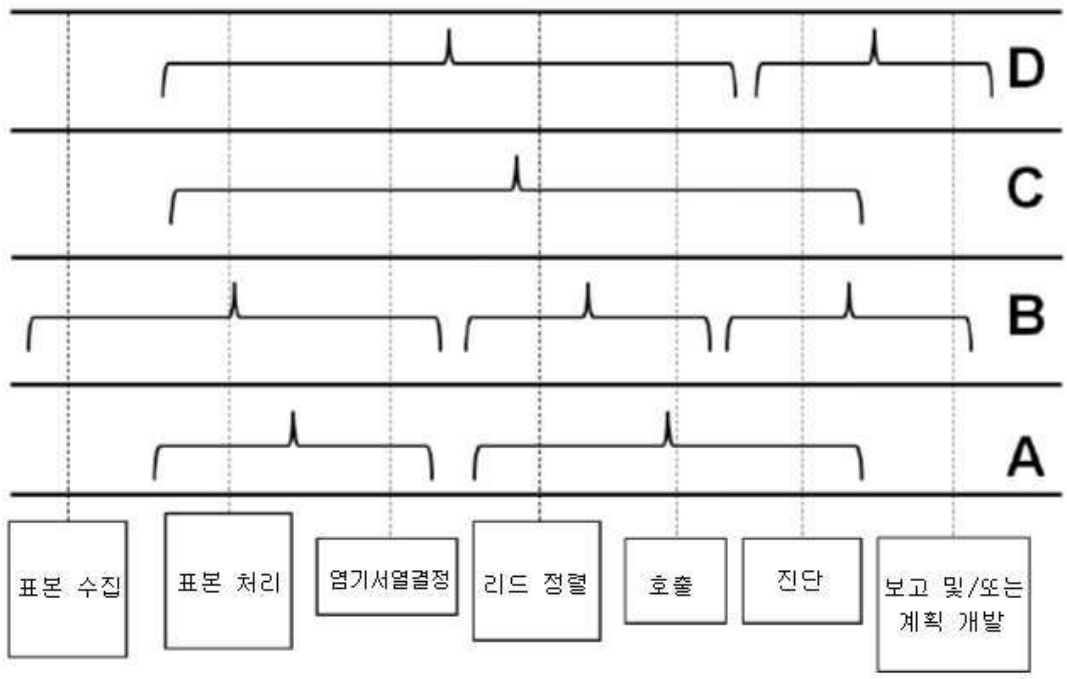
도면4b


포스트 GC 교정 QV
mapQ
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0

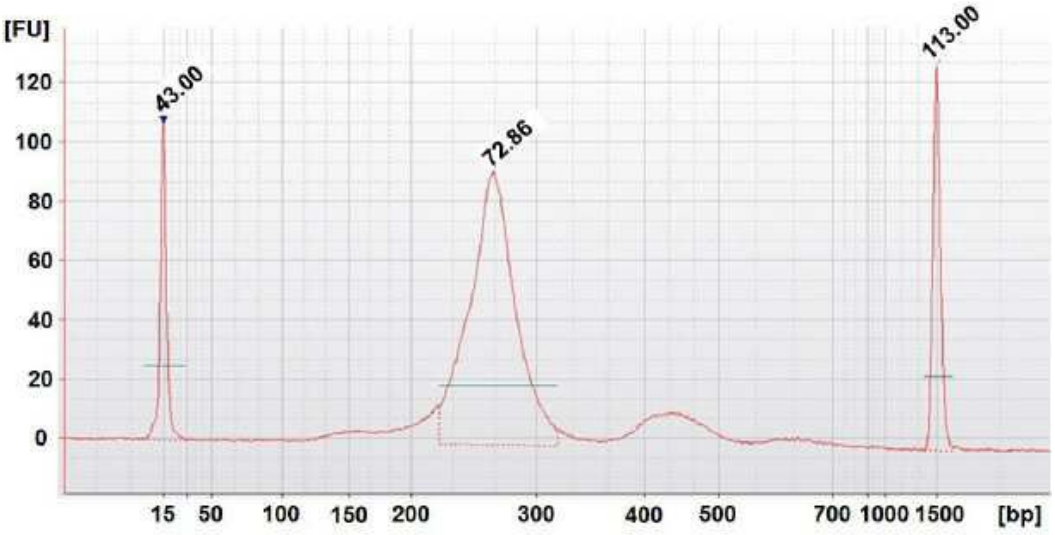
도면5


네트워크/클라우드
(가령, 인터넷)
호출/진단
05
표본 수집
처리와
염기서열결정
분석 및
호출 산출
01
03
07

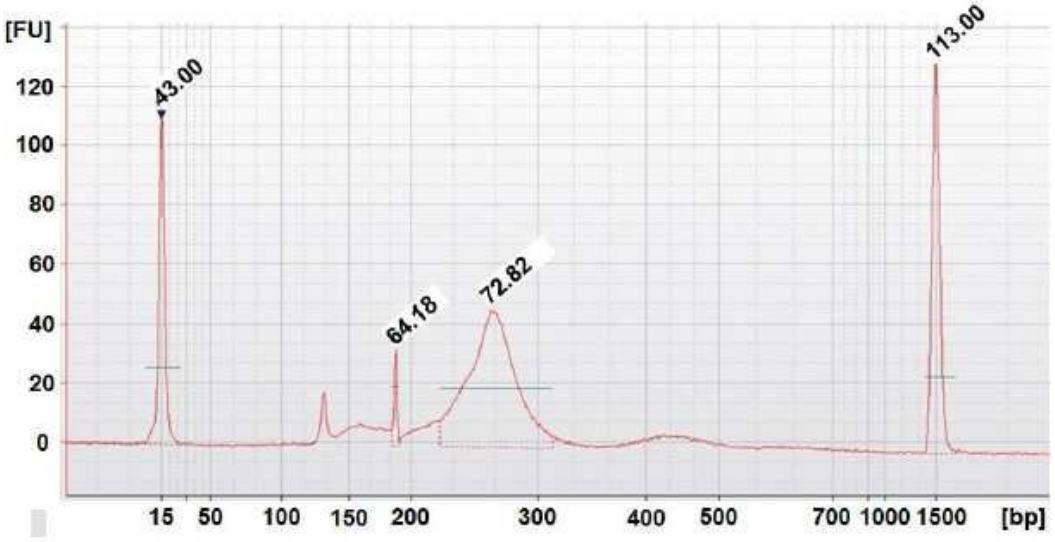
도면6


D
C
B
A
표본 수집
표본 처리
염기서열결정
리드 정렬
호출
진단
보고 및/또는 계획 개발

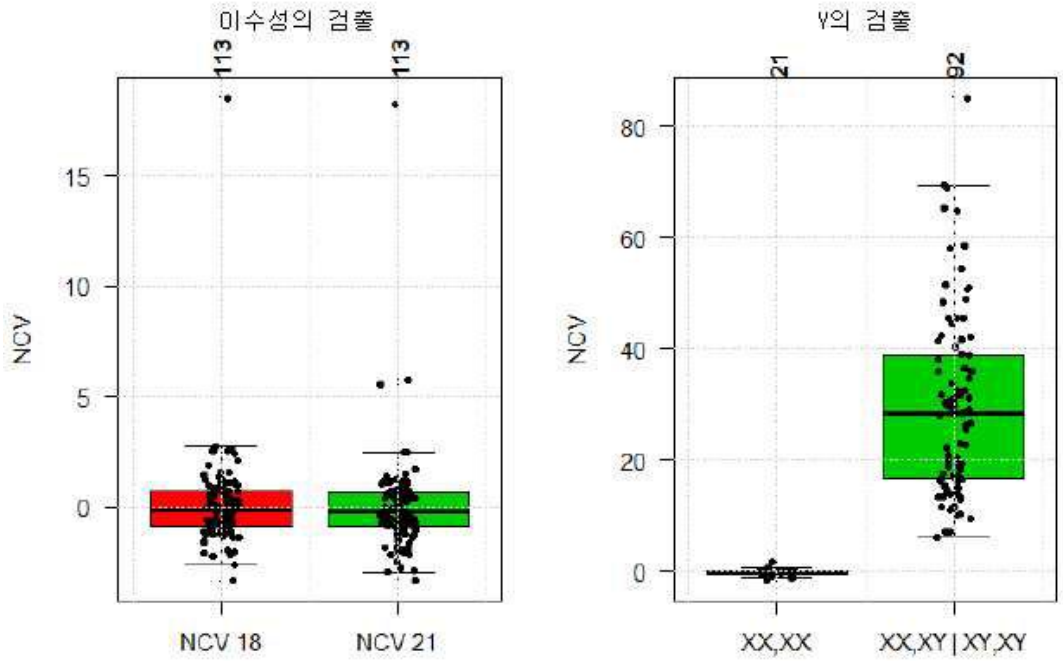
도면7a


[FU]
120
100
80
60
40
20
0
43.00
72.86
113.00
15
50
100
150
200
300
400
500
700
1000
1500
[bp]

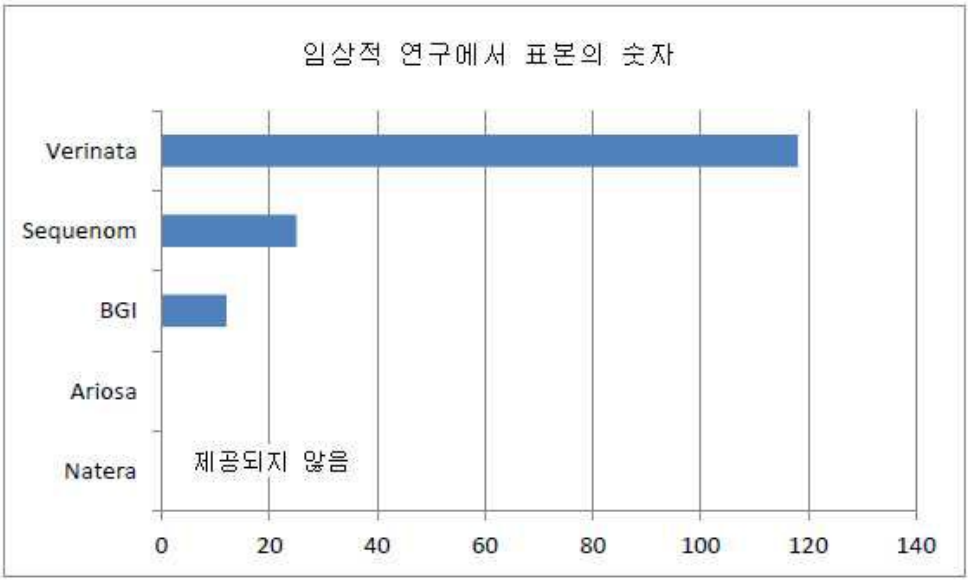
도면7b


[FU]
120
100
80
60
40
20
0
43.00
64.18
72.82
113.00
15
50
100
150
200
300
400
500
700
1000
1500
[bp]

도면8


이수성의 검출
113
113
NCV
15
10
5
0
NCV 18
NCV 21
Y의 검출
21
92
NCV
80
60
40
20
0
XX,XX
XX,XY | XY,XY

도면9



	제공되지 않음
Natera	
Ariosa	0
BGI	12
Sequenom	25
Verinata	118