(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년08월10일
(11) 등록번호 10-2566176
(24) 등록일자 2023년08월08일

(51) 국제특허분류(Int. Cl.)
G16B 30/00 (2019.01) *C12Q 1/68* (2018.01)
G16B 20/00 (2019.01)
(52) CPC특허분류
G16B 30/00 (2019.02)
C12Q 1/6858 (2018.05)
(21) 출원번호 10-2016-7036900
(22) 출원일자(국제) 2015년05월29일
심사청구일자 2020년05월27일
(85) 번역문제출일자 2016년12월29일
(65) 공개번호 10-2017-0016393
(43) 공개일자 2017년02월13일
(86) 국제출원번호 PCT/US2015/033403
(87) 국제공개번호 WO 2015/184404
국제공개일자 2015년12월03일
(30) 우선권주장
62/005,877 2014년05월30일 미국(US)
(56) 선행기술조사문헌
US20130029852 A1*
(뒷면에 계속)
전체 청구항 수 : 총 27 항

(73) 특허권자
베리나타 헬스, 인코포레이티드
미국, 캘리포니아 92122, 샌 디에고, 일루미나 웨이 5200
(72) 발명자
추도바, 다리야 아이.
미국, 캘리포니아 95123, 산 조세, 타오미노 애버뉴 5931
아베듀바, 다이아나
미국, 캘리포니아 94563, 오린다, 오차드 로드 227
(74) 대리인
강명구, 이경민

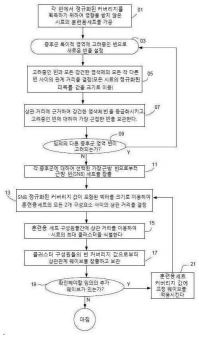
심사관 : 김채연

(54) 발명의 명칭 **태아 아-염색체 홀배수체 및 복제수 변이 탐지**

(57) 요 약

아염색체성 영역들의 복제 수 변이(CNV)에 관련된 증후군이 포함된 다양한 의료상 문제와 연관된 것으로 알려진 또는 의심되는 CNV를 결정하는 방법들이 공개된다. 일부 구체예들에 있어서, 모체 및 태아 무 세포 DNA가 포함된 모체 시료들을 이용하여 태아의 CNV를 결정하는 방법들이 제시된다. 본 명세서에서 공개된 일부 구체예들에서 시료 에서 GC-함량 편향을 제거함으로써, 서열 데이터 분석의 민감성 및/또는 특이성을 개선시키는 방법들이 제시된다. 일부 구체예들에 있어서, 시료내 GC-함량 편향의 제거는 영향을 받지 않은 훈련용 시료들에 걸쳐 공통적인 조직적 변이에 대하여 교정된 서열 데이터에 근거한다. 일부 구체예들에 있어서, 시료 데이터에서 증후군 관련된 편향은 신호 대 잡음비를 증가시키기 위하여 또한 제거된다. 관심 대상 서열들의 CNV 평가를 위한 시스템이 또한 설명된다.

대 표 도



(52) CPC특허분류
C12Q 1/6869 (2018.05)
G16B 20/00 (2019.02)
C12Q 2535/122 (2019.08)
C12Q 2537/16 (2013.01)
C12Q 2537/165 (2013.01)

(56) 선행기술조사문헌
WO2013109981 A1*
D. Y. Chiang 외, "High-resolution mapping of copy-number alterations with massively parallel sequencing", Nature Methods, 6(1):99-103. (2008.11.30.)*
WO2014015319 A1
WO2013015793 A1
KR1020140023847 A
*는 심사관에 의하여 인용된 문헌

명 세 서

청구범위

청구항 1

하나 또는 그 이상의 게놈의 핵산을 포함하는 테스트 시료에서의 관심 대상 서열의 복제 수를 평가하기 위하여 하나 또는 그 이상의 프로세서 및 시스템 메모리를 포함하는 컴퓨터 시스템에서 실행되는 방법에 있어서,

상기 관심 대상 서열은, 복제 수 변이가 유전적 증후군과 관련되는 아염색체 게놈 영역에 있으며, 상기 방법은

(a) 상기 테스트 시료에서 핵산의 서열화에 의해 획득된 서열 리드를 수취하는 공정;

(b) 상기 테스트 시료의 서열 리드를, 상기 관심 대상 서열을 포함하는 참조 게놈에 정렬시키고, 그렇게 함으로써 테스트 서열 태그를 제공하는 공정, 여기서 상기 참조 게놈은 다수의 빈으로 분할되며;

(c) 상기 참조 게놈에서 빈에 대한 테스트 서열 태그의 커버리지를 결정하는 공정;

(d) 테스트 시료와 동일한 방식으로 서열화되고 정렬된 훈련용 시료의 훈련용 세트의 부분 집단으로부터 획득된 관심 대상 서열에서 다수의 빈에서의 다수의 예상 커버리지를 이용하여 관심 대상 서열에서 다수의 빈에서의 다수의 커버리지를 조정하는 공정,

여기서, 상기 훈련용 시료는 관심 대상 서열의 복제수 변이에 의해 영향을 받지 않으며, 서열 밖의 변이는 부분 집단에서의 훈련용 시료간 유사하며, 각 서열 밖의 변이는 관심 대상 서열밖의 다수의 빈에 걸친 커버리지 변이이며, 상기 관심 대상 서열 밖의 다수의 빈의 커버리지는 관심 대상 서열 내의 빈의 커버리지와 상관관계가 있으며, 관심 대상 서열 밖의 빈 커버리지를 이용하여 다수의 예상 커버리지를 결정하는 공정은, (i) 훈련용 시료들의 훈련용 세트로부터 훈련용 부분 집단을 확인하는 공정과, 여기서 상기 훈련용 부분 집단의 시료는 상기 관심 대상 서열 밖의 빈들에서의 그들의 커버리지에서 서로 관련되고, (ii) 훈련용 세트의 훈련용 부분집단의 관심 대상 서열 밖의 빈들의 커버리지로부터 예상 커버리지를 획득하는 공정을 포함하며; 그리고

(e) 상기 (d)공정으로부터 조정된 커버리지에 근거하여 테스트 시료에서의 관심 대상 서열의 복제 수를 평가하는 공정을 포함하는, 방법.

청구항 2

프로그램 코드가 기록되는 컴퓨터-리드가능한 저장 매체에 있어서,

상기 프로그램 코드가

(a) 테스트 시료에서 핵산의 서열화에 의해 획득되는 서열 리드를 수취하는 공정;

(b) 상기 테스트 시료의 서열 리드를 유전적 증후군에 관련된 상기 관심 대상 서열을 포함하는 참조 게놈에 정렬시키고, 그렇게 함으로써 테스트 서열 태그를 제공하는 공정, 여기서 상기 참조 게놈은 다수의 빈으로 분할되며;

(c) 상기 참조 게놈에서 빈에 대한 테스트 서열 태그의 커버리지를 결정하는 공정;

(d) 테스트 시료와 동일한 방식으로 서열화되고 정렬된 훈련용 시료의 훈련용 세트의 부분 집단으로부터 획득된 관심 대상 서열에서 다수의 빈에서의 다수의 예상 커버리지를 이용하여 관심 대상 서열에서 다수의 빈에서의 다수의 커버리지를 조정하는 공정, 여기서 상기 훈련용 시료는 관심 대상 서열의 복제수 변이에 의해 영향을 받지 않으며, 서열 밖의 변이는 부분 집단에서의 훈련용 시료간 유사하며, 각 서열 밖의 변이는 관심 대상 서열밖의 다수의 빈에 걸친 커버리지 변이이며, 상기 관심 대상 서열 밖의 다수의 빈의 커버리지는 관심 대상 서열 내의 빈의 커버리지와 상관관계가 있으며, 상기 다수의 예상 커버리지는 (i) 훈련용 시료들의 훈련용 세트로부터 훈련용 부분 집단을 확인하는 공정과, 여기서 상기 훈련용 부분 집단의 시료는 상기 관심 대상 서열 밖의 빈들에서의 그들의 커버리지에서 서로 관련되고, (ii) 훈련용 세트의 훈련용 부분집단의 관심 대상 서열 밖의 빈들의 커버리지로부터 예상 커버리지를 획득하는 공정에 의하여 결정되며; 그리고

(e) 상기 (d)공정으로부터 조정된 커버리지에 근거하여 테스트 시료에서의 관심 대상 서열의 복제 수를 컴퓨터

시스템에 의해 평가하는 공정에 대한 코드를 포함하는, 컴퓨터-리드가능한 저장 매체.

청구항 3

하나 또는 그 이상의 게놈의 핵산을 포함하는 테스트 시료들에서 빈 커버리지를 조정하는데 사용하기 위한 예상 커버리지를 확인하는 방법에 있어서,

상기 방법이

(a) 테스트 시료와 동일한 방법으로 서열화되고 정렬된 훈련용 시료의 훈련용 세트로부터 데이터를 획득하는 공정; 및

(b) 관심 대상 서열에서 다수의 빈에서의 다수의 예상 커버리지를 결정하는 공정을 포함하며,

여기서, 상기 관심 대상 서열에서 다수의 빈에서의 다수의 예상 커버리지는 훈련용 세트의 부분 집단으로부터 얻어지며,

훈련용 시료는 관심 대상 서열의 복제수 변이에 의해 영향을 받지 않으며,

서열 밖의 변이는 부분 집단에서의 훈련용 시료간 유사하며,

각 서열 밖의 변이는 관심 대상 서열밖의 다수의 빈에 걸친 커버리지 변이이며,

상기 관심 대상 서열 밖의 다수의 빈의 커버리지는 관심 대상 서열 내의 빈의 커버리지와 상관관계가 있으며;

상기 관심 대상 서열은, 복제 수 변이가 유전적 증후군과 관련되는 아염색체 게놈 영역에 있으며, 여기서 관심 대상 서열 밖의 빈 커버리지를 이용하여 다수의 예상 커버리지를 결정하는 공정은, (i) 훈련용 시료들의 훈련용 세트로부터 훈련용 부분 집단을 확인하는 공정과, 여기서 상기 훈련용 부분 집단의 시료는 상기 관심 대상 서열 밖의 빈들에서의 그들의 커버리지에서 서로 관련되고, (ii) 훈련용 세트의 훈련용 부분집단의 관심 대상 서열 밖의 빈들의 커버리지로부터 예상 커버리지를 획득하는 공정을 포함하는, 방법.

청구항 4

청구항 3에 있어서,

각각의 훈련용 시료에 대하여

(i) 훈련용 시료에서 핵산을 서열화함으로써 획득된 서열 리드를 수취하는 공정;

(ii) 상기 훈련용 시료의 서열 리드를, 상기 관심 대상 서열을 포함하는 참조 게놈에 정렬시키고, 그렇게 함으로써 훈련용 서열 태그를 제공하는 공정, 여기서 상기 참조 게놈은 다수의 빈으로 분할되며; 및

(iii) 상기 관심 대상 서열을 포함하는 참조 게놈에서 빈에 대한 훈련용 서열 태그의 커버리지를 결정하는 공정을 더 포함하는 것을 특징으로 하는 방법.

청구항 5

청구항 4에 있어서,

상기 (iii)공정에서 빈에 대한 커버리지를 결정하는 공정은 전체 빈에 걸쳐 서열 태그의 총 수에 관하여 빈 당 태그의 계수를 정규화하는 공정을 포함하며, 상기 (b)공정에서의 예상 커버리지는 정규화된 커버리지인 것을 특징으로 하는 방법.

청구항 6

청구항 3에 있어서,

상기 관심 대상 서열 밖의 다수의 빈들의 커버리지와 상기 관심 대상 서열 내의 다수의 빈들의 커버리지 사이의 상관성은 상관 거리에 의해 측정되는 것을 특징으로 하는 방법.

청구항 7

청구항 6에 있어서,

상기 상관 거리는 훈련용 세트의 시료로부터 창출된 빈 커버리지의 벡터 간의 거리로써 산출되는 것을 특징으로 하는 방법.

청구항 8

삭제

청구항 9

청구항 3에 있어서,

상기 훈련용 부분 집단을 확인하는 공정은, 상기 훈련용 세트에서 시료들의 클러스터를 확인하는 공정을 포함하는 것을 특징으로 하는 방법.

청구항 10

하나 또는 그 이상의 게놈의 핵산이 포함된 테스트 시료를 이용하여 유전적 증후군과 관련된 관심 대상 서열의 복제 수 평가를 위한 시스템에 있어서,

상기 시스템은,

(a) 상기 테스트 시료에서 핵산을 서열화함으로써 획득되는 서열 리드를 수취하는 작동;

(b) 상기 테스트 시료의 서열 리드를, 상기 관심 대상 서열을 포함하는 참조 게놈에 정렬시키고, 그렇게 함으로써 테스트 서열 태그를 제공하는 작동, 여기서, 상기 참조 게놈은 다수의 빈으로 분할되고;

(c) 상기 참조 게놈에서 빈에 대한 테스트 서열 태그의 커버리지를 결정하는 작동;

(d) 테스트 시료와 동일한 방식으로 서열화되고 정렬된 훈련용 시료의 훈련용 세트의 부분 집단으로부터 획득된 관심 대상 서열에서 다수의 빈의 다수의 예상 커버리지를 이용하여 관심 대상 서열에서 다수의 빈에서의 다수의 커버리지를 조정하는 작동, 여기서 상기 훈련용 시료는 관심 대상 서열의 복제수 변이에 의해 영향을 받지 않으며, 서열 밖의 변이는 부분 집단에서의 훈련용 시료간 유사하며, 각 서열 밖의 변이는 관심 대상 서열밖의 다수의 빈에 걸친 커버리지 변이이며, 상기 관심 대상 서열 밖의 다수의 빈의 커버리지는 관심 대상 서열 내의 빈의 커버리지와 상관관계가 있고, 관심 대상 서열 밖의 빈 커버리지를 이용하여 다수의 예상 커버리지를 결정하는 작동은, (i) 훈련용 세트의 부분 집단으로서 상기 관심 대상 서열 밖의 빈들에서의 그들의 커버리지에서 서로 관련하는 훈련용 세트 중의 훈련용 시료를 확인하는 작동과, (ii) 상기 부분 집단의 빈에서의 커버리지로부터 예상 커버리지를 획득하는 작동을 포함하며; 그리고

(e) 컴퓨터 시스템에 의해, 상기 (d)작동으로부터의 조정된 커버리지에 근거하여 테스트 시료에서의 관심 대상 서열의 복제 수를 평가하는 작동을, 실행 또는 야기하도록 설계되거나 구성되는 하나 이상의 프로세서를 포함하는, 시스템.

청구항 11

청구항 10에 있어서,

상기 하나 이상의 프로세서는 테스트 시료에서 하나 또는 그 이상의 염색체의 복제 수를 평가하는 작동을 실행 또는 야기하도록 더 설계되거나 구성되는 것을 특징으로 하는 시스템.

청구항 12

청구항 10에 있어서,

상기 하나 이상의 프로세서는 전체 빈에 걸쳐 서열 태그의 총 수에 관하여 빈 당 태그의 계수를 정규화함으로써 상기 (c)작동에서의 빈에 대한 커버리지를 결정하는 작동을 실행 또는 야기하도록 더 설계되거나 구성되며, 여기서 상기 (d)작동에서의 조정된 커버리지는 정규화된 커버리지인 것을 특징으로 하는 시스템.

청구항 13

청구항 10에 있어서,

상기 (d)작동에서 이용된 관심 대상 서열 밖의 빈은 염색체 13, 18, 및 21 이외의 인간의 상염색체에서의 빈인

것을 특징으로 하는 시스템.

청구항 14

청구항 10에 있어서,

상기 하나 이상의 프로세서가, 관심대상 서열 내의 고려중인 빈에서의 커버리지와 상기 관심 대상 서열 밖의 빈에서의 커버리지 간의 상관 거리를 결정함으로써 관심대상 서열 밖의 빈을 확인하는 작동을 실행하거나 야기하도록 더 설계되거나 또는 구성되는 것을 특징으로 하는 시스템.

청구항 15

청구항 14에 있어서,

상기 하나 이상의 프로세서가 상기 훈련용 세트의 시료로부터 창출된 빈 커버리지의 벡터간의 산출된 거리에 의해 상기 상관 거리를 산출하는 작동을 실행하거나 야기하도록 더 설계되거나 또는 구성되는 것을 특징으로 하는 시스템.

청구항 16

삭제

청구항 17

청구항 10에 있어서,

시료들의 군을 확인하는 작동은, 상기 시료들의 클러스터를 확인하는 작동을 포함하는 것을 특징으로 하는 시스템.

청구항 18

청구항 17에 있어서,

상기 예상 커버리지를 획득하는 작동은 훈련용 시료들의 확인된 군의 커버리지의 중심적 경향을 결정하는 작동을 포함하는 것을 특징으로 하는 시스템.

청구항 19

청구항 10 내지 15, 17 및 18중 어느 한 항에 있어서,

상기 (d)작동에서의 빈의 테스트 서열 태그의 커버리지를 조정하는 작동은,

데이터 점들에 함수를 피팅시키는 작동, 여기서 각각의 데이터 점은 빈에서 테스트 시료에 대하여 대응하는 커버리지에 예상 커버리지를 관련시키고; 그리고

상기 빈에서의 커버리지를 상기 함수에 적용시킴으로써 관심 대상 서열의 빈에서의 커버리지를 조정하는 작동을 포함하는 것을 특징으로 하는 시스템.

청구항 20

청구항 19에 있어서,

상기 함수는 선형 함수인 것을 특징으로 하는 시스템.

청구항 21

청구항 10 내지 15, 17 및 18중 어느 한 항에 있어서,

상기 하나 이상의 프로세서가, 상기 관심 대상 서열의 빈에 대한 측정 커버리지 값으로부터 상기 예상 값을 차감시킴으로써 상기 (d)작동에서의 관심 대상 서열의 빈에서 테스트 서열 태그의 커버리지를 조정하는 작업을 실행 또는 야기하도록 더 설계되거나 구성되는 것을 특징으로 하는 시스템.

청구항 22

청구항 10 내지 15, 17 및 18중 어느 한 항에 있어서,

상기 하나 이상의 프로세서가, 상기 관심 대상 서열로써 증후군 특이적 영역의 시작점과 끝점을 결정하기 위한 분절화를 수행하는 작동을 실행하거나 야기하도록 더 설계되거나 구성되는 것을 특징으로 하는 시스템.

청구항 23

청구항 10 내지 15, 17 및 18중 어느 한 항에 있어서,

상기 테스트 시료는 2종의 상이한 게놈 유래의 핵산의 혼합물을 포함하는 것을 특징으로 하는 시스템.

청구항 24

청구항 10 내지 15, 17 및 18중 어느 한 항에 있어서,

상기 핵산은 cfDNA 분자를 포함하는 것을 특징으로 하는 시스템.

청구항 25

청구항 10 내지 15, 17 및 18중 어느 한 항에 있어서,

상기 테스트 시료는 태아 및 모체의 무-세포 핵산을 포함하는 것을 특징으로 하는 시스템.

청구항 26

청구항 10 내지 15, 17 및 18중 어느 한 항에 있어서,

상기 테스트 시료는 동일한 개체로부터의 암성 세포와 영향이 없는 세포 유래의 핵산을 포함하는 것을 특징으로 하는 시스템.

청구항 27

청구항 10 내지 15, 17 및 18중 어느 한 항에 있어서,

상기 서열화 리드는 초기 복합 서열화에 의해 획득되며,

상기 하나 이상의 프로세서가,

제1 임계치보다 더 높은 증후군 분류 및 복제 수 변이를 소명(calling)하기 위하여 테스트 시료가 관심 대상 서열의 제1커버리지을 갖는 것을 결정하는 작동;

초기 복합 서열화보다 더 심화된 서열화 깊이에서 테스트 시료를 재서열화시켜, 재서열화된 데이터를 획득하는 작동; 그리고

상기 재서열화된 데이터를 이용하여 증후군 분류 또는 복제 수 변이를 결정하는 작동을, 실행 또는 야기하도록 더 설계되거나 구성되는 것을 특징으로 하는 시스템.

청구항 28

청구항 27에 있어서,

상기 하나 이상의 프로세서가,

재서열화된 데이터로부터 증후군 분류 또는 복제 수 변이를 소명하는 관심 대상 서열의 제2 커버리지를 획득하는 작동; 그리고

제2 커버리지를, 제1 임계치보다 더 높은 제2 임계치에 비교하는 작동에 의해서 재서열화된 데이터를 이용하여 증후군 분류 또는 복제 수 변이를 결정하는 작동을 실행하거나 야기하도록 더 설계되거나 또는 구성되는 것을 특징으로 하는 시스템.

청구항 29

청구항 10 내지 15, 17 및 18중 어느 한 항에 있어서,

상기 유전적 증후군은 1p36 결손 증후군, Wolf-Hirschhorn 증후군, Cri-du-Chat 증후군, Angelman 증후군,

Williams 증후군 및 DiGeorge 증후군으로 구성되는 군으로부터 선정되는 것을 특징으로 하는 시스템.

청구항 30

삭제

청구항 31

삭제

청구항 32

삭제

청구항 33

삭제

청구항 34

삭제

청구항 35

삭제

청구항 36

삭제

청구항 37

삭제

청구항 38

삭제

청구항 39

삭제

청구항 40

삭제

청구항 41

삭제

청구항 42

삭제

청구항 43

삭제

청구항 44

삭제

청구항 45

삭제

청구항 46

삭제

청구항 47

삭제

청구항 48

삭제

청구항 49

삭제

청구항 50

삭제

청구항 51

삭제

청구항 52

삭제

청구항 53

삭제

청구항 54

삭제

청구항 55

삭제

청구항 56

삭제

청구항 57

삭제

청구항 58

삭제

청구항 59

삭제

청구항 60

삭제

청구항 61

삭제

청구항 62

삭제

청구항 63

삭제

청구항 64

삭제

청구항 65

삭제

청구항 66

삭제

청구항 67

삭제

청구항 68

삭제

청구항 69

삭제

청구항 70

삭제

청구항 71

삭제

청구항 72

삭제

청구항 73

삭제

청구항 74

삭제

청구항 75

삭제

청구항 76

삭제

청구항 77

삭제

청구항 78

삭제

청구항 79

삭제

청구항 80

삭제

청구항 81

삭제

청구항 82

삭제

청구항 83

삭제

청구항 84

삭제

청구항 85

삭제

청구항 86

삭제

발명의 설명

기 술 분 야

[0001] **관련 출원에 대한 교차 참조**

[0002] 본 출원은 35 USC § 119(e) 하에 2014년 5월 30일자로 제출된 "태아 아염색체 홀배수체 탐지"를 제목으로 하는 미국 가출원 번호 62/005,877을 우선권으로 주장하며, 이의 전문은 모든 목적을 위하여 본 출원에 참고로 포함된다.

배 경 기 술

[0003] **배경**

[0004] 인간의 의학 연구에 있어서 중요한 시도중의 하나가 유해한 건강 결과를 낳는 유전적 비정상의 발견이다. 많은 경우에 있어서, 특이적 유전자들 및/또는 임계적 진단학적 표지들이 비정상적 복제 수에 존재하는 게놈 부분들에서 확인되었다. 예를 들면, 출생전 진단에서 전체 염색체의 잉여(extra) 또는 유실(missing) 복사체는 유전적 병소들의 빈번한 발생이다. 암에서 전체 염색체 또는 염색체 세그먼트들(segment)의 복사체들의 결손(deletion) 또는 증식(multiplication), 그리고 상기 게놈의 특이적 영역들(regions)의 더 높은 수준의 증폭(amplifications)이 흔히 발생하는 것이다.

[0005] 복제 수 변이 (CNV)에 대한 대부분의 정보는 구조적 비정상의 인지가 허용되는 세포유전적 해상력(resolution)에 의해 제공되어 왔었다. 유전적 스크리닝(screening) 및 생물학적 양측정(dosimetry)을 위한 통상적인 과정은 핵형(karyotypes) 분석용 세포를 얻기 위하여 침습적(invasive) 과정들, 가령, 양막천자(amniocentesis), 탯줄천자(cordocentesis), 또는 융모막 융모 샘플링 (CVS)을 이용하여 왔었다. 세포 배양물, 형광 핵산 혼성화

(FISH), 정량적 형광 PCR (QF-PCR) 및 배열-비교 게놈 혼성화 (어레이-CGH)를 요구하지 않는 좀더 신속한 테스트 방법들의 필요에 대한 인식은 복제 수 변이 분석을 위한 분자-세포유전적 방법들로 개발되어 왔었다.

[0006] 인간의 의학 연구에 있어서 중요한 시도 중의 하나가 유해한 건강 결과를 낳는 유전적 비정상의 발견이다. 많은 경우에 있어서, 특이적 유전자들 및/또는 임계적 진단학적 표지들이 비정상적 복제 수에 존재하는 게놈 부분들에서 확인되었다. 예를 들면, 출생전 진단에서 전체 염색체의 잉여(extra) 또는 유실(missing) 복사체는 유전적 병소들의 빈번한 발생이다. 암에서 전체 염색체 또는 염색체 세그먼트들(segment)의 복사체들의 결손(deletion) 또는 증식(multiplication), 그리고 상기 게놈의 특이적 영역들(regions)의 더 높은 수준의 증폭(amplifications)이 흔히 발생하는 것이다.

[0008] 상대적으로 짧은 시간에 전체 게놈의 서열화(sequencing)를 허용하는 기술의 도래, 그리고 순환하는 무-세포(cell-free) DNA (cfDNA)의 발견은 하나의 염색체에서 기인된 유전적 물질을 침습적 샘플링 방법들과 연관된 위험없이, 또다른 염색체로부터 기인된 유전적 물질과 비교하는 기회를 제공하였으며, 이것은 관심 대상의 유전적 서열들의 다양한 종류의 복제 수 변이를 진단하는 도구를 제공한다.

[0009] 일부 적용에서 복제 수 변이 (CNV)는 고조된 기술적 도전과 관련된다. 예로써, 형제간 다태 (또는 다접합체) 임신에 대한 비-침습성 출생전 진단 (NIPD)은 단일 임신보다 더 어려운데, 그 이유는 태아 cfDNA의 전체 분획(fraction)이 단일 임신과 다중 임신이 유사하고, 태아의 수에 의해 태아 분획이 낮아지고, 이는 다시 분석에서 신호 대 잡음비를 감소시키기 때문이다. 추가적으로, Y 염색체 기반의 진단 이를 테면, 성별 식별은 Y 염색체와 관련된 한정에 의해 영향을 받는다. 구체적으로, Y 염색체의 커버리지(Y 염색체)는 상염색체(autosomes)의 것보다 더 낮고, Y 염색체 상에 반복된 서열들은 이들의 정확한 위치에 리드의 매핑(mapping)을 복잡하게 만든다. 더욱이, 현행 일부 서열화 프로토콜은 초단(ultra-short) 리드, 이를 테면 25mer 리드 및 태그를 이용하며, 이는 25mer 태그가 대부분의 편재된 반복가능한 요소들의 전형적인 크기보다 더 짧기 때문에 여전히 또다른 정렬상 난제를 제기한다. 본 명세서에서 공개된 일부 구체예들은 CNV 평가용 서열 데이터의 분석에 있어서 민감성(sensitivity) 및/또는 특이성(specificity)을 개선시키는 방법들을 제시한다.

[0010] 공개된 공정들의 일부 구체예들은 전체 염색체 또는 염색체의 세그먼트들의 복제 수 변이의 탐지에 적합하다. 그러나, 더 짧은 유전적 서열들이 관련된 유전적 질환들의 경우, 기존 방법들의 신호대 잡음비가 너무 낮아서 복제 수 변이의 믿을만한 탐지가 어렵다. 예로써, 많은 아염색체성(subchromosomal) 즉 염색체부분유전적 증후군은 어느 정도 메가베이스(megabases) 크기의 서열과 관련되어, CNV를 판단하기 위한 분석에 이용가능한 신호를 제한시킨다.

[0011] 출생전 비침습성 진단학에 있어서 증후군 관련된 짧은 서열로부터 파생되는 불충분한 민감성, 제한된 수준의 cfDNA, 그리고 게놈 정보의 내재된 성질로부터 파생된 기술에 의한 서열화 편향(bias)이 포함된 기존 방법들에서의 한계는, 다양한 임상적 환경에서 복제 수 변화를 확실하게 진단하기 위하여 특이성, 민감성, 그리고 적용가능성 중 어느 하나 또는 모두를 제공하는 비침습성 방법들의 지속적인 필요성의 근저가 된다. 본 명세서에서 공개된 구체예들은 전술한 필요의 일부를 충족시키고, 특히 출생전 비침습성 진단학 실행에 적용가능한 믿을만한 방법들의 제시함에 있어서, 장점을 제시한다.

[0012] **요약**

[0013] 다양한 구체예들에 있어서, 임의의 태아 이수성(異數性) 즉 홀배수체(aneuploidy)의 복제 수 변이 (CNV), 그리고 다양한 의료상 문제와 연관된 것으로 공지된 또는 의심되는 CNVs를 판단하는 방법들이 제시된다. 상기 방법들은 관심대상의 CNVs과 무관한 조직적 변동(variance), 이를 테면 게놈 서열들에서 GC 격변(fluctuation)에 대한 잡음 및 오류를 감소시키는 기전을 포함한다. 본 방법에 따라 결정될 수 있는 CNV는 염색체 1-22의, X와 Y의 임의의 하나 또는 그 이상의 삼체성(trisomies) 및 단체성(monosomies), 기타 염색체 다체성(polysomies), 그리고 상기 염색체의 임의의 하나 또는 그 이상의 세그먼트들의 결손 및/또는 중복을 포함한다.

[0014] 한 구체예는 테스트 시료에서 관심 대상 서열, 가령, 특이적 증후군과 관련된 상대적으로 짧은 세그먼트들의 복제 수 변이 (CNV)를 식별해내는 방법을 제시한다. 상기 방법은 완전한 염색체 또는 염색체의 세그먼트들 대신 관심 대상 서열들의 복제 수 변이를 평가한다.

[0015] 본 공개의 한 측면은 하나 또는 그 이상의 게놈 핵산이 포함된 테스트 시료 에서 관심 대상 서열의 복제 수를 평가하는 방법들을 제시한다. 상기 관심 대상 서열은 아염색체 게놈 영역에 있으며, 이때 복제 수 변이는 유전적 증후군과 연관된다. 상기 방법들은 하나 또는 그 이상의 프로세서와 메모리가 포함된 컴퓨터 시스템에서 실행될 수 있다. 한 가지 실행에 있어서, 상기 방법은 다음 공정을 수반한다: (a) 테스트 시료에서 세포 없는 DNA

를 서열화함으로써 서열 리드를 얻고; (b) 상기 관심 대상 서열이 함유된 참조 게놈에 테스트 시료의 서열 리드를 나란하게 정렬하고, 그렇게 함으로써 테스트 서열 태그가 제공되며, 이때 상기 참조 게놈은 다수의 빈(bin)으로 분할되며; (c) 상기 관심 대상 서열이 함유된 참조 게놈에서 빈들에 대한 테스트 서열 태그의 커버리지를 결정하고; (d) 실질적으로 테스트 시료와 동일한 방법으로 서열화되고, 영향을 받지 않은 정렬된 훈련용(training) 시료의 훈련용 세트의 부분 집단(subset)에서 획득된 예상 커버리지를 이용하여 빈에서 테스트 서열 태그의 커버리지를 조정하고, 그리고 이때 상기 예상된 커버리지(coverage)는 상기 관심 대상 서열내의 빈의 커버리지와 관련된 것으로 밝혀진 관심 대상 서열 밖에 있는 빈의 커버리지를 이용하여 획득하고; 그리고 (e) 상기(d)의 조정된 커버리지에 근거하여 테스트 시료에서 관심 대상 서열의 복제 수를 평가하는 공정을 포함한다.

[0016] 일부 실행에 있어서, 상기 방법에는 하나 또는 그 이상의 게놈들에서 하나 또는 그 이상이 염색체 홀배수체를 보유하는지를 결정하기 위하여 테스트 시료에서 하나 또는 그 이상의 염색체의 복제 수를 평가하는 공정을 더 수반한다. 일부 실행에 있어서, 하나 또는 그 이상의 염색체의 복제 수의 평가는 (d)공정 이후에 실행된다. 일부 실행에 있어서, 상기 방법에는 (d)공정에 앞서, 상기 훈련용 세트로부터 획득된 포괄적 웨이브 프로파일(global wave profile)을 적용함으로써, 테스트 서열 태그의 커버리지를 조정하는 공정이 더 수반되고, 이때 상기 포괄적 웨이브 프로파일에는 훈련용 세트에 걸쳐 평균된 참조 게놈에서 빈의 커버리지를 포함한다. 일부 실행에 있어서, 상기 방법에는 (d)공정에 앞서, 테스트 시료의 빈들간의 GC 함량 수준과 커버리지와의 상관관계에 근거하여 테스트 서열 태그의 커버리지를 조정하는 공정을 더 수반한다.

[0017] 일부 실행에 있어서, 하나 또는 그 이상의 염색체의 복제 수의 평가에는 테스트 시료의 하나 또는 그 이상의 염색체 각각의 서열 분량(dose) 산출이 수반되며; 이때 상기 서열 분량은 하나 또는 그 이상의 염색체에서 테스트 서열 태그의 테스트 서열 커버리지를 정규화(normalizing) 서열에서 테스트 서열 태그의 커버리지로 나누어 산출된다. 일부 실행에 있어서, 상기 방법에는 상기 서열 분량을 훈련용 세트의 서열 분량의 표준 편차로 나누어 정규화된 서열 값을 획득하는 공정을 더 수반한다.

일부 실행에 있어서, (c)에서 빈에 대한 커버리지 결정은 전체 빈에 걸쳐 서열 태그의 총 수에 있어서, 빈 당 태그의 계수(counts)를 정규화시키는 공정을 포함하며, 여기에서 (d)에서 조정된 커버리지는 정규화된 커버리지이다. 일부 실행에 있어서, (d)에 이용된 관심 대상 서열 밖에 있는 빈은 염색체 13, 18, 및 21이외의 인간의 상(常) 염색체에서의 빈들이다. 일부 실행에 있어서, 상기 관심 대상 서열내의 고려중인 빈에서 커버리지와 관심 대상 서열 밖에 있는 빈에서 커버리지 사이의 상관 거리(correlation distances)를 결정함으로써 관심 대상 서열 밖에 있는 빈이 식별된다. 일부 실행에 있어서, 상기 상관 거리는 훈련용 세트의 시료로부터 창출된 빈 커버리지의 벡터 간에 거리로 산출된다.

[0018] 삭제

[0019] 일부 실행에 있어서, 상기 예상된 커버리지는 (i) 훈련용 세트의 부분 집단으로서 상기 관심 대상 서열 밖에서의 빈들의 커버리지에서 서로간 관련된 훈련용 세트 중의 훈련용 시료를 확인하고, 그리고 (ii) 상기 부분집단의 빈에서의 커버리지로부터 예상된 커버리지를 획득함으로써, 예상된 커버리지가 획득되었다. 일부 실행에 있어서, 시료들 집단(group)의 식별은 전술한 시료들의 클러스터(cluster)를 식별해내는 공정을 포함한다. 일부 실행에 있어서, 상기 예상된 커버리지의 획득은 훈련용 시료들의 식별된 집단의 커버리지의 중심적 경향(central tendency) 결정과 관련된다.

[0020] 상기 설명된 방법들의 일부 실행에서 여러번 반복(iterations)을 위하여 (d)를 되풀이 하는 공정을 더 수반한다. 각 반복에는 앞선 반복에서의 조정된 커버리지가 현재 반복에서 조정되는 커버리지로서 이용된다. 더욱이, 각 반복에는 전술한 영향을 받지 않은 시료들의 상이한 부분 집단으로부터 획득된 예상된 커버리지가 이용된다.

[0021] 일부 실행에 있어서, (d)작동에 있어서 빈의 테스트 서열 태그의 커버리지 조정에는 다음공정이 수반된다: 데이터 점들에 함수, 가령, 선형 함수를 피팅(fitting)시키고, 각 데이터 점은, 빈에서 테스트 시료에 대하여 대응하는 커버리지에 예상된 커버리지를 관련시키며; 그리고 전술한 빈에서의 커버리지를 상기 함수에 적용시킴으로써 관심 대상 서열의 빈에서의 커버리지가 조정된다. 다른 시행들에 있어서, (d)작동에 있어서 관심 대상 서열의 빈에서 테스트 서열 태그의 커버리지 조정에는 상기 관심 대상 서열의 빈에 대하여 측정된 커버리지 값으로부터 상기 예상된 값을 차감시키는 것이 수반된다.

[0022] 일부 실행에 있어서, 상기 방법들에는 관심 대상 서열로써 증후군 특이적 영역의 시작점과 끝점을 결정하기 위

한 분절화(segmentation)의 실행이 더 수반된다.

[0023] 일부 실행에 있어서, 상기 테스트 시료에는 2개의 상이한 게놈들로부터의 핵산 혼합물이 포함된다. 일부 실행에 있어서, 전술한 핵산은 cfDNA 분자들이다. 일부 실행에 있어서, 상기 테스트 시료에는 태아와 모체의 무-세포 핵산이 함유된다. 일부 실행에 있어서, 상기 테스트 시료에는 동일한 개체로부터 암에 걸린 세포와 영향을 받지 않은 세포의 핵산이 포함된다. 일부 실행에 있어서, 상기 방법들에는 영향을 받지 않은 다수의 개체 및/또는 상기 테스트 시료로부터 무 세포 DNA의 추출이 더 수반된다. 일부 실행에 있어서, 상기 방법들에는 서열화기(sequencer)를 이용하여 테스트 시료로부터 핵산이 서열화되고, 그렇게 함으로써 테스트 시료의 서열 리드가 생성되는 것이 더 수반된다. 일부 실행에 있어서, 상기 서열 리드에는 개체의 전체 게놈의 임의의 부위로부터 약 20 내지 50-bp의 서열들이 포함된다. 일부 실행에 있어서, 상기 서열 리드에는 바-코드화된 25-mers가 포함된다. 일부 실행에 있어서, 상기 테스트 서열 태그와 훈련용 서열 태그의 커버리지는 배제되지 않은(non-excluded) 부위 계수 (NES 계수)로 제공된다. NES 계수는 배제되지 않은 부위들에 매핑된 비-겹침(non-redundant) 서열 태그의 수이다. 일부 실행에 있어서, 또는 이 계수는 배제되지 않은 부위들에 대하여 매핑된 독특하게 나열된 비-겹침 서열 태그의 수이다.

[0024] 전술한 방법들중 임의의 것에 대한 일부 실행에 있어서, 빈 크기는 약 1000 bp 내지 1,000,000 bp, 또는 약 100,000 bp이다. 일부 실행에 있어서, 상기 방법에는 테스트 시료의 서열 리드의 수를 이용한 계산에 의해 빈 크기를 결정하는 것이 더 수반된다. 일부 실행에 있어서, 상기 유전적 증후군은 1p36 결손 증후군, Wolf-Hirschhorn 증후군, Cri-du-Chat 증후군, Angelman 증후군, Williams 증후군, 그리고 DiGeorge 증후군으로 구성된 군에서 선택된다.

[0025] 전술한 방법들중 임의의 것에 대한 일부 실행에 있어서, 상기 서열화 리드는 초기 복합 서열화에 의해 획득되는데, 이는 다음의 것들이 더 포함된다: 제1 임계치(threshold)보다 더 큰 증후군 분류 또는 복제 수 변이를 소명(calling)하기 위한 제1값을 갖는 테스트 시료를 식별 내지 확인하고; 초기 복합 서열화보다 더 심층화된 서열화에서 확인된 테스트 시료를 재서열화시켜 재서열화된 데이터를 획득하고; 그리고 상기 재서열화된 데이터를 이용하여 증후군 분류 또는 복제 수 변이를 결정하는 공정이 더 포함된다. 일부 실행에 있어서, 재서열화된 데이터를 이용하여 증후군 분류 또는 복제 수 변이를 결정하는 공정은 다음을 수반한다: 재서열화된 데이터로부터 증후군 분류 또는 복제 수 변이를 소명하기 위하여 제2 값을 획득하고; 그리고 제2 값을 제2 임계치와 비교하는 공정을 더 포함하며, 상기 제2 임계치는 제1 임계치보다 더 높다. 일부 실행에 있어서, 식별된 테스트 시료의 제1 값은 사전 설정된(preset) 값보다 더 낮고, 여기에서 사전 설정된 값은 제1 임계치보다 더 높고, 그리고 여기에서 제1 임계치보다 더 낮은 시료는 영향을 받지 않은 시료로 판단되며, 사전설정된 값보다 더 높은 시료는 영향을 받은 것으로 판단되며, 그리고 제1 임계치에서 사전설정된 값 사이 범위의 시료는 재서열화용으로 확인된다. 일부 실행에 있어서, 식별된 테스트 시료의 제1 값은 공지의 영향을 받은 시료들과 비교하였을 때 상대적으로 낮다. 일부 실행에 있어서, 확인된 테스트 시료의 제1 값은 공지의 영향을 받은 시료들의 약 90%보다 더 낮다.

[0026] 본 명세서의 또다른 측면은 하나 또는 그 이상의 게놈들의 핵산이 포함된 테스트 시료들에서 빈의 커버리지 조절에 이용되는 예상 커버리지를 확인하는 방법들을 제시한다. 일부 실행에 있어서, 상기 방법은 다음을 수반한다: (a) 상기 테스트 시료들과 실질적으로 동일한 방식으로 서열화되고, 정렬된 영향을 받지 않은 훈련용 시료들의 훈련용 세트로부터 데이터를 획득하고; 그리고 (b) 관심 대상 서열 밖에 있는 빈의 커버리지를 이용하여 예상 커버리지를 결정하고, 여기에서 관심 대상 서열 밖에 있는 빈의 커버리지는 관심 대상 서열내에서의 빈의 커버리지와 연관되며, 그리고 여기에서 상기 관심 대상 서열은 이의 복제 수 변이가 유전적 증후군과 연관된 아염색체 게놈 영역이다. 일부 실행에 있어서, (d)에 이용된 관심 대상 서열 밖에 있는 빈은 염색체 13, 18, 및 21이외의 인간의 상염색체에 있는 빈들이다. 일부 실행에 있어서, 각 훈련용 시료의 경우 상기 방법은 다음을 더 수반한다: (i) 훈련용 시료에서 무 세포 DNA의 서열화에 의해 획득된 서열 리드를 얻고; (ii) 훈련용 시료의 서열 리드는 상기 관심 대상 서열이 포함된 참조 게놈에 대하여 정렬되고, 그렇게 함으로써 훈련용 서열 태그가 제공되며, 여기에서 상기 참조 게놈은 다수의 빈으로 분리되며; 그리고 (iii) 상기 관심 대상 서열이 포함된 참조 게놈에서 빈에 대한 훈련용 서열 태그의 커버리지가 결정된다. 일부 실행에 있어서, (iii)에서 빈에 대한 커버리지 결정은 전체 빈에 걸쳐 서열 태그의 총 수에 있어서, 빈 당 태그의 계수를 정규화시키는 것을 포함하며, 그리고 여기에서 (b)에서 예상 커버리지는 정규화된 커버리지이다.

[0027] 일부 실행에 있어서, 상기 방법은 상기 관심 대상 서열내의 고려중인 빈에서 커버리지와 관심 대상 서열 밖의 빈에서 커버리지 사이의 상관 거리를 결정함으로써, 상기 관심 대상 서열내 빈에서 커버리지와 관련된 커버리지를 보유하는 상기 관심 대상 서열 밖에서의 빈을 더 확인하는 공정을 더 수반한다. 일부 실행에 있어서, 상기

상관 거리는 훈련용 세트의 시료로부터 창출된 빈 커버리지의 벡터 간의 거리로 산출된다.

[0028] 일부 실행에 있어서, 상기 방법의 (b)에 있어서 관심 대상 서열 밖에 있는 빈의 커버리지를 이용한 예상 커버리지의 결정은 다음을 수반한다: (i) 영향을 받지 않은 훈련용 시료들의 훈련용 세트로부터 훈련용 부분 집단을 확인해내고, 여기에서 훈련용 부분 집단의 시료는 상기 관심 대상 서열 밖에 있는 빈에서 이들의 커버리지에서 서로 연관되며, 그리고 (ii) 훈련용 부분 집단의 빈들의 커버리지로부터 예상 커버리지를 획득한다. 일부 실행에 있어서, 훈련용 부분 집단의 확인은 훈련용 세트에서 시료들의 클러스터의 확인공정을 포함한다. 일부 실행에 있어서, 상기 예상된 커버리지의 획득은 식별된 훈련용 부분 집단의 커버리지의 중심적 경향 (가령, 평균, 중앙, 또는 최빈치(最頻値, mode)를 결정하는 것을 수반한다.

[0029] 일부 실행에 있어서, 상기 방법은 다음에 의해 훈련용 세트에서 훈련용 시료 서열의 커버리지를 조절하는 것을 더 수반한다: 함수 (가령, 선형 함수 또는 이차 함수)를 데이터 점들에 피팅시키고, 각각은 훈련용 부분 집단에 걸쳐 예상된 커버리지을 전술한 특정 빈에서 훈련용 시료 서열에 대하여 대응하는 관찰된 커버리지에 관련시키며; 그리고 전술한 빈에서 관찰된 커버리지를 상기 함수에 적용시킴으로써 훈련용 시료 서열의 빈에서 상기 커버리지를 조절하는 공정을 포함한다. 일부 실행에 있어서, 상기 함수는 선형 함수다.

[0030] 본 명세서의 또다른 측면에는 컴퓨터 시스템의 하나 또는 그 이상의 프로세서에 의해 실행될 때, 컴퓨터 시스템에 의해 유전적 증후군에 관련된 관심 대상 서열의 복제 수 평가를 하기 위한 방법이 실행되도록 하게 만드는 비-일시적인 기계 리드가능 매체 저장 프로그램 코드가 포함된 컴퓨터 프로그램 제품이 제공된다. 일부 실행에 있어서, 상기 프로그램 코드는 (a) 상기 테스트 시료에서 무 세포 DNA의 서열화에 의해 획득된 서열 리드를 수취(受取)하고; (b) 상기 테스트 시료의 서열 리드를 상기 관심 대상 서열이 포함된 참조 게놈에 정렬시키고, 그렇게 함으로써 테스트 서열 태그가 제공되며, 여기에서 상기 참조 게놈이 다수의 빈으로 분할되고; (c) 상기 참조 게놈에서 상기 관심 대상 서열이 포함된, 빈에 테스트 서열 태그의 커버리지를 측정하고; (d) 상기 테스트 시료에서와 실질적으로 동일한 방식으로 서열화되고, 그리고 정렬된 영향을 받지 않은 훈련용 시료들의 훈련용 세트로부터 획득된 예상된 커버리지를 이용하여 빈에서 테스트 서열 태그의 커버리지를 조절하고, 그리고 여기에서 상기 예상된 커버리지는 관심 대상 서열내에서의 빈의 커버리지와 관련된 것으로 밝혀진 관심 대상 서열 밖의 빈의 커버리지를 이용하여 획득되며; 그리고 (e) 컴퓨터 시스템에서 (d)의 조정된 커버리지에 기초하여 테스트 시료에서 관심 대상 서열의 복제 수를 평가하는 코드를 포함한다.

[0031] 다양한 구체예들에 있어서, 컴퓨터 프로그램 제품은 컴퓨터 시스템들이 상기 설명된 방법들중 임의의 것을 실행하게 할 수 있도록 하기 위한 지침을 제공할 수 있다.

[0032] 본 명세서의 또다른 측면은 하나 또는 그 이상의 게놈의 핵산이 포함된 테스트 시료를 이용하여 유전적 증후군과 관련된 관심 대상 서열의 복제 수 평가를 위한 시스템을 제공한다. 상기 시스템은 시료로부터 핵산 서열 정보를 제공하는 테스트 시료로부터 핵산을 수용하는 서열화기, 그리고 유전적 증후군과 관련된 관심 대상 서열의 복제 수를 평가하기 위한 작동(operation)을 실행 또는 야기하도록 설계 또는 구성된 로직(logic)을 포함한다. 일부 실행에 있어서, 상기 작동은 (a) 상기 테스트 시료에서 무 세포 DNA의 서열화에 의해 획득된 서열 리드를 수취하고; (b) 상기 테스트 시료의 서열 리드를 상기 관심 대상 서열이 포함된 참조 게놈에 정렬시키고, 그렇게 함으로써 테스트 서열 태그가 제공되며, 여기에서 상기 참조 게놈이 다수의 빈으로 분할되고; (c) 상기 관심 대상 서열이 포함된 상기 참조 게놈에서, 빈에 대하여 테스트 서열 태그의 커버리지를 측정하고; (d) 상기 테스트 시료에서와 실질적으로 동일한 방식으로 서열화되고, 그리고 정렬된 영향을 받지 않은 훈련용 시료들의 훈련용 세트로부터 획득된 예상된 커버리지를 이용하여 빈에서 테스트 서열 태그의 커버리지를 조절하고, 그리고 여기에서 상기 예상된 커버리지는 관심 대상 서열에서의 빈의 커버리지와 관련된 것으로 밝혀진 관심 대상 서열 밖의 빈의 커버리지를 이용하여 획득되며; 그리고 (e) 컴퓨터 시스템에 의해 (d)로부터의 조정된 커버리지에 기초하여 테스트 시료에서 관심 대상 서열의 복제 수를 평가하는 공정을 포함한다.

[0033] 일부 실행에 있어서, 상기 시스템의 로직은 프로세서; 그리고 전술한 작동들의 실행을 위한 지침이 보관된 하나 또는 그 이상의 컴퓨터-리드가능 저장 매체를 포함한다. 일부 실행에 있어서, 상기 시스템은 다음을 더 포함한다: 모체 테스트 시료에서 태아 핵산과 모체 핵산으로부터 최소한 약 10,000개의 서열 리드를 수취하기 위한 인터페이스(interface), 여기에서 상기 서열 리드는 전자 포멧(format)으로 제공되며; 그리고 최소한 일시적으로 다수의 전술한 서열 리드를 저장하기 위한 메모리. 일부 실행에 있어서, 상기 시스템은 모체 테스트 시료로부터 무 세포 DNA를 추출하기 위한 장치를 더 포함한다. 일부 실행에 있어서, 무 세포 DNA를 추출하기 위한 장치는 상기 서열화기와 함께 동일한 설비에 위치되며, 그리고 여기에서 모체 테스트 시료를 취하는 장치는 멀리있는 원격 설비에 위치한다. 일부 실행에 있어서, 상기 로직은 하나 또는 그 이상의 게놈이 염색체 홀배수체를 가지

는 지를 판단하기 위하여 상기 테스트 시료에서 하나 또는 그 이상의 염색체의 복제 수 평가를 실행 또는 야기하도록 더 설계 또는 구성된다. 일부 실행에 있어서, 상기 로직은 (d) 이후 하나 또는 그 이상의 염색체의 복제 수를 평가하도록 실행 또는 야기하도록 더 설계 또는 구성된다.

[0034] 다양한 구체예들에 있어서, 상기 시스템은 컴퓨터 시스템이 유전적 증후군과 관련된 관심 대상 서열의 복제 수를 평가하기 위하여 상기 설명된 방법중 임의의 것을 실행하도록 하기 위한 지침 내지 명령(instruction)을 제공할 수 있다.

[0035] 본 명세서의 실시예들은 인간에 관련되기는 하지만, 언어는 인간 게놈들에 주로 관계되며, 본 명세서에서 설명된 개념은 임의의 식물 또는 동물의 게놈에 적용가능하다. 본 명세서의 이들 목적 및 다른 목적과 특징은 다음의 설명 및 첨부된 청구범위로 부터 더욱 명백하게 할 것이며, 또는 이후에서 제시되는 명세서의 실행에 의해 알게 될 것이다.

[0036] **참고로 포함**

[0037] 본 명세서에 언급된 이들 참고문헌에 개시된 모든 서열을 포함하는 모든 특허, 특허 출원 및 기타 간행물은 각각 개별 간행물, 특허, 특허 출원이 본원에 참고로 포함되는 것을 구체적으로 개별적으로 나타낸 것과 동일한 정도로 참고에 의해 본 명세서에 명시적으로 포함된다. 관련 부분에서 인용된 모든 문서는 본 명세서에서 그들 인용의 맥락에 의해 지시된 목적을 위해 본원에 참고로 그 전체 포함된다. 그러나 어떠한 문서의 인용도 본 개시에 대한 선행 기술임을 인정하는 것으로 해석되어서는 안된다.

도면의 간단한 설명

[0038] **도면의 간단한 설명**

도 1은 핵산 혼합물이 포함된 테스트 시료에서 복제 수 변이의 존재 또는 부재를 결정하는 방법(100)의 순서도다.

도 2는 복제 수의 평가를 위하여 이용된 관심 대상 서열의 커버리지를 결정하기 위한 공정의 순서도를 나타낸다.

도 3a는 테스트 시료의 서열 데이터에서 노이즈를 감소시키기 위한 공정의 일 실시예의 공정도를 나타낸다.

도 3b-3k는 도 3a에서 설명된 공정의 다양한 단계에서 획득된 데이터 분석을 제시한다.

도 4a는 서열 데이터에서 노이즈를 감소시키는 서열 마스크(mask)를 만들기 위한 공정의 순서도를 보여준다.

도 4b는 MapQ 스코어(score)는 정규화된 커버리지 양의 CV와 함께 강력한 변화없는 단조한(monotonous) 상관관계를 갖는다는 것을 보여준다.

도 5는 상기 게놈 관련된 특이적 증후군의 영역(region)에 도입되는 조직적 편향을 제거 또는 감소하는 공정을 보여준다.

도 6은 중합과정의 2개 패스를 보여주며, 패스(pass) 1은 일반적 CNV 탐지용이고, 패스 2는 상대적으로 짧은 아염색체성 서열들에 관련된 증후군에 관한 CNV를 탐지하기 위한 것이다.

도 7은 테스트 시료를 처리하고, 그리고 궁극적으로 진단을 하기 위한 분산된(dispersed) 시스템의 블록 구성도(block diagram)이다.

도 8은 테스트 시료들을 처리함에 있어서, 상이한 작동들이 어떤 방식에 의해 집단화되고 시스템의 상이한 요소들에 의해 취급되어질 수 있는 지를 도식적으로 설명한다.

도 9a 및 9b는 실시예 1a (도 9a)에서 설명된 단축 프로토콜, 그리고 실시예 1b (도 9b)에서 설명된 프로토콜에 따라 준비된 cfDNA 서열화 라이브러리의 전기영동도(electropherogram)를 보여준다.

도 10은 118명의 쌍태 임산부로부터 모체 혈장 시료에서 정규화된 염색체 값 (NCV) 분포를 나타낸다. (A) 염색체 21 및 18의 NCV 분포; 3개 시료는 T21 영향을 받은 (T21에 대한 모자이크(mosaic)인 태아 포함) 것으로 분류되었으며, 한 개 시료는 T18 영향을 받은 것으로 분류되었다. (B) 염색체 Y의 NCV 분포. 상기 코호트는 여성/여성으로 임상적으로 분류되는 시료들 또는 최소한 하나의 남성 태아 (남성/여성 및 남성/남성)가 포함된 시료들로 분리되며, 그리고 Y 염색체의 존재는 염색체 Y에 대한 NCV를 이용하여 결정되었다.

도 11은 NIPT 연구에서 분석된 쌍태(twin) 시료를 나타낸다. 상업적으로 이용가능한 NIPT 테스트의 수행을 평가하기 위하여 다양한 연구에서 쌍태 시료가 이용되었다.

도 12은 2D 100kb 정규화된 커버리지 히트맵(heatmap)에 의해 제시되는 12plex 훈련용 데이터에서 신호 이질성(signal heterogeneity)을 설명한다 (x-축 = chr 22 게놈 순서 그리고 y-축은 비-교사(unsupervised) 계층별 클러스터링에 의해 구동된 순서를 가진 영향을 받지 않은 CLIA 시료들을 나타낸다).

도 13은 DiGeorge 증후군 커버리지 빈도를 나타낸다. 직선(Lines)은 공적(公的) DB/lit 검토에서 증후군 커버리지 빈도를 나타내고; 보다 저밀도의 사선영역은 5.5Mb 증후군 조사 경계를 나타내며, 그리고 2.7Mb 콘센수스(consensus) 영역은 보다 고밀도의 사선 영역을 나타낸다.

도 14는 AS/PW 증후군 커버리지 빈도를 나타낸다. 직선은 공적 DB/lit 검토에서 증후군 커버리지 빈도를 나타내고; 보다 저밀도의 사선 영역은 10Mb 증후군 조사 경계를 나타내며, 그리고 5.8Mb 콘센수스 영역은 보다 고밀도의 사선 영역을 나타낸다.

도 15는 CdC 증후군 커버리지 빈도를 나타낸다. 직선은 공적 DB/lit 검토에서 증후군 커버리지 빈도를 나타내고; 보다 저밀도의 사선 영역은 26.5Mb 증후군 조사 경계를 나타내며, 그리고 9.8Mb 콘센수스 영역은 보다 저밀도의 사선 영역을 나타낸다.

도 16은 CdC 증후군 커버리지 빈도를 나타낸다. 직선은 공적 DB/lit 검토에서 증후군 커버리지 빈도를 나타내고; 보다 저밀도의 사선 영역은 8.6Mb 증후군 조사 경계를 나타내며, 그리고 1.58Mb 콘센수스 영역은 보다 고밀도의 사선 영역을 나타낸다.

도 17은 Wolf-Hirschhorn 증후군 커버리지 빈도를 나타낸다. 직선은 공적 DB/lit 검토에서 증후군 커버리지 빈도를 나타내고; 보다 저밀도의 사선 영역은 14.5Mb 증후군 조사 경계를 나타내며, 그리고 3.6Mb 콘센수스 영역은 보다 고밀도의 사선 영역을 나타낸다.

도 18은 1p36 증후군 커버리지 빈도를 나타낸다. 직선은 대중 DB/lit 검토에서 증후군 커버리지 빈도를 나타내고; 보다 저밀도의 사선 영역은 13.5Mb 증후군 조사 경계를 나타내며, 그리고 5Mb 콘센수스 영역은 보다 고밀도의 사선 영역을 나타낸다.

도 19는 증후군에 의한 SNB 인구학을 나타낸다. Cri du Chat의 경우 chr 19 상에 SNBs 분획이 상승되었음을 보여주거나 또는 chr 22에 속하는 p36 SNBs의 예상밖 높은 비율(>30%)을 보여준다.

도 20은 증후군에 의한 SNB 중첩(overlap)을 나타낸다. 1p36와 DiGeorge 그리고 Cri du Chat와 Angelman 간에 SNBs에서 상당한 중첩이 있다.

도 21은 CriDuChat 증후군의 경우 SSS-BER에서 SNB 웨이브 #에 대한 함수로써 CV 드롭의 개관을 나타낸다.

도 22는 CriDuChat 증후군의 경우 SNB 크기에 대하여 콘센수스 증후군 비율 CV를 나타낸다.

도 23은 오리지널(original) (v4) 대(vs) 프로토타입화된 (v5) 검정(verifi) 파이프라인(pipeline)에서 콘센수스 증후군 비율 CV 감소를 나타낸다.

도 24는 염색체 21에 대한 상이한 서열화 깊이(depths)에 대한 예상된 가음성 (FN) 대(vs) 가양성 (FP) 비율을 나타낸다.

도 25는 염색체 18에 대한 상이한 서열화 심도(depths)에 대한 예상된 FN 대 FP 비율을 나타낸다.

도 26은 Cri-du-Chat 증후군을 가지는 것으로 공지된 높은 태아 분획 임상 시료에 대하여 본 명세서에서 설명된 분절화 및 판단 분석을 이용하여 증후군 소명 성능을 보여준다.

도 27은 Cri-du-Chat 증후군을 가지는 것으로 공지된 낮은 태아 분획 임상 시료에 대하여 본 명세서에서 설명된 분절화 및 판단 분석을 이용하여 증후군 소명 성능을 보여준다.

발명을 실시하기 위한 구체적인 내용

[0039] **상세한 설명**

[0040] 공개된 구체예들은 태아 및 모체 무-세포 핵산이 포함된 테스트 시료에서 Y 염색체의 복제 수 평가를 위한 방법들, 장치, 그리고 시스템에 관계한다. 일부 구체예들에 있어서, 관심 대상 서열들은 유전적 또는 질환 상태와

연관된 것으로 알려진 또는 연관된 것으로 의심되는 전체 염색체에 대하여 가령, 킬로베이스 (kb) 내지 메가베이스 (Mb) 범위로부터 전체 염색체에 걸친 게놈 세그먼트 서열들을 포함한다. 일부 구체예들에 있어서, Y 염색체의 복제 수는 태아 성별을 결정하는데 이용된다. 일부 구체예들에 있어서, 본 방법에 따라 결정되는 CNV는 성염색체 Y의 단체성 및 삼체성(가령 47,XXY 및 47,XYY), 성염색체의 다른 다체성 이를 테면 사염색체성(tetrasomy) 및 오염색체성(pentasomies) (가령 XXXXY 및 XYYYY), 그리고 임의의 하나 또는 그 이상의 성염색체의 세그먼트들의 결손 및/또는 중복을 포함한다. 관심 대상 서열들의 다른 예로는 공지의 홀배수체와 연관된 염색체, 가령, 삼염색체성 XXX, 삼염색체성 21과 연관된 염색체 및 암 등의 질환에서 증배하고 있는 염색체의 세그먼트들, 가령 급성 골수성 백혈병에서 부분적 삼염색체성 8이 포함된다.

[0041] 다른 명시가 없는 한, 본 명세서의 방법 및 시스템의 실시는 분자 생물학, 미생물학, 단백질 정제, 단백질 가공, 단백질 및 DNA 서열화 그리고 당기술 범위 에서 있는 재조합 DNA 분야에서 통상적으로 이용되는 통상적 기술 및 장치를 포함한다. 이러한 기술 및 장치는 당업자들에게 공지된 것이며, 다수의 책 및 참고 작업에서 설명된다 (가령, Sambrook et al., "Molecular Cloning: A Laboratory Manual," Third Edition (Cold Spring Harbor), [2001]); and Ausubel et al., "Current Protocols in Molecular Biology" [1987] 참고).

수치적 범위는 그 범위를 한정하는 수를 포함한다. 본 명세서를 통해 주어진 모든 최대 수치 한정은, 모든 보다 낮은 수치 한정을, 마치 그런 보다 낮은 수치 한정이 본 명세서에 명시적으로 기재된 것처럼 포함하는 것으로 의도되며, 본 명세서를 통하여 주어진 모든 최소 수치 한정은 모든 더 높은 수치 한정을, 마치 그런 더 높은 수치 한정이 본 명세서에 명시적으로 기재된 것처럼 포함할 것이다. 본 명세서를 통하여 제공된 모든 수치적 범위는 보다 더 넓은 수치적 범위내에 속하는 모든 보다 더 좁은 수치적 범위를, 마치 그러한 보다 더 좁은 수치적 범위가 본 명세서에서 모두 명시적으로 표현된 것처럼 포함할 것이다.

본 명세서에서 제공된 머릿글은 본 공개 내용을 제한시키려는 의도는 아니다.

[0042] 삭제

[0043] 삭제

[0044] 명시적으로 다른 언급이 없는 한, 본 명세서에서 이용된 모든 기술적 그리고 과학적 용어는 당업계 숙련자들에 의해 공통적으로 이해되는 것과 동일한 의미를 가진다. 본 명세서에 포함된 용어들이 포함된 다양한 과학적 사전은 공지되어 있고, 당업자들에게 이용가능하다. 비록 본 발명의 실시 또는 테스트에 있어서 본 명세서에서 설명된 것들과 유사한 또는 대등한 임의의 방법 및 재료들이 또한 이용될 수 있지만, 일부 방법들과 재료들이 지금 설명된다.

[0045] 하기에서 바로 정의되는 용어들은 대체로 명세서를 참고하여 더욱 자세히 설명된다. 본 발명은 설명된 특정 방법, 프로토콜, 설명된 시약들은 매우 다변할 수 있기 때문에, 이들에 본 발명이 한정되지 않는다는 것을 인지해야 한다.

[0046] **정의**

[0047] 본 명세서에서 이용된 바와 같이, 단수("a", "an" 및 "the")는 다른 명시적인 언급이 없는 한 복수 개념을 포함한다.

[0048] 다른 언급이 없는 한, 핵산은 5'에서 3' 방향으로 좌에서 우로 기재되며 아미노산 서열은 아미노부터 카르복시 방향으로 좌에서 우로 각각 기재된다.

[0049] CNV를 위한 핵산 시료 분석 내용에서 이용될 때 "평가하는(assessing)"이란, 세가지 소명(call)중 하나: "정상" 또는 "영향없음", "영향있음", 그리고 "무-소명(no-call)"에 의해 염색체 또는 세그먼트 홀배수체의 상태를 특징화하는 것을 말한다. 정상 및 영향있음(발병)의 소명(호출, calling)을 위한 임계치는 전형적으로 설정된다. 홀배수체 또는 다른 복제 수 변이와 관련된 매개변수는 시료에서 측정되며, 이 측정된 값은 임계치와 비교된다. 중복 유형 홀배수체, 염색체 또는 세그먼트 분량 (또는 다른 측정된 값 서열 함량)이 영향있음(본 명세서에서는 '영향을 받은' 또는 '발병'이라고도 표현됨)의 시료들의 정의(규정)된 임계치 설정보다 우위에 있다면, 영향있음이라고 하는 소명이 행해진다. 이러한 홀배수체의 경우, 염색체 또는 세그먼트 분량이 정상 시료의 임계치 설정보다 아래에 있다면, 정상이라는 소명이 행해진다. 결손 유형 홀배수체와 대조적으로,

염색체 또는 세그먼트 분량이 영향을 받은 또는 영향없는 시료의 정의된 임계치 아래에 있는 경우 영향있음이라고 하는 발병 소명이 행해지며, 염색체 또는 세그먼트 분량이 정상 시료의 임계치 설정 보다 위에 있다면 정상이라는 소명이 행해진다. 예를 들면, 삼염색체성 존재하에서 "정상" 소명은 신뢰성에 대한 사용자-정의된 임계치를 하회하는 매개변수 가령 테스트 염색체 분량에 의해 결정되며, 그리고 "영향있음" 소명은 신뢰성에 대한 사용자-정의된 임계치를 상회하는 매개변수 가령 테스트 염색체 분량에 의해 결정된다. "무-소명" 결과는 "정상" 또는 "영향있음" 이라는 소명은 행하기 위한 임계치 사이에 있는 매개변수 가령 테스트 염색체 분량에 의해 결정된다. 용어 "무-소명"은 "미분류(unclassified)"와 호환가능하게 이용된다.

[0050] 본 명세서에서 용어 "복제 수 변이"는 참조 시료에 존재하는 핵산 서열의 복제 수와 비교하여, 테스트 시료에 존재하는 핵산 서열의 복제 수에서의 변이를 지칭한다. 특정 구체예들에 있어서, 상기 핵산 서열은 1 kb 또는 그 보다 더 크다. 일부 경우들에 있어서, 상기 핵산 서열은 염색체 전체 또는 그의 상당부분이다. "복제 수 변종(copy number variant)"는 테스트 시료에서 관심 대상의 핵산 서열과 이 관심 대상 핵산서열의 예상된 수준을 비교하여 복제 수 차이가 발견된 핵산의 서열을 말한다. 예를 들면, 테스트 시료에서 있는 관심 대상 핵산서열의 수준은 검증된 적격시료(qualified sample)에 존재하는 것과 비교된다. 복제 수 변종/변이에는 미세결손(microdeletions)이 포함된 결손, 미세삽입(microinsertions)이 포함된 삽입, 중복, 다중복(multiplications), 그리고 전위가 포함된다. CNVs는 염색체 홀배수체 및 부분적 홀배수체를 포괄한다.

[0051] 본 명세서에서 용어 "홀배수체(aneuploidy)"는 전체 염색체, 또는 염색체의 일부의 상실 또는 획득에 의해 야기되는 유전적 물질의 불균형을 말한다.

[0052] 본 명세서에서 용어 "염색체 홀배수체" 및 "완전한 염색체 홀배수체"란 전체 염색체의 상실 또는 증대(gain)에 의해 야기되는 유전적 물질의 불균형을 말하며, 그리고 생식계열 홀배수체 및 모자이크 홀배수체가 포함된다.

[0053] 본 명세서에서 용어 "부분적 홀배수체" 및 "부분적 염색체 홀배수체"란 염색체의 일부의 상실 또는 증대에 의해 야기되는 유전적 물질의 불균형을 말하며, 가령, 부분적 일염색체성(monosomy) 및 부분적 삼염색체성을 말하며, 그리고 전위, 결손 및 삽입로 인한 불균형이 포괄된다.

[0054] 용어 "다수"는 하나 이상의 요소를 지칭한다. 예를 들면, 이 용어는 명세서에서 공개된 방법들을 이용하여 테스트 시료 및 검증된 적격시료에서 복제 수 변이에 있어서 유의적인 차이를 식별해내는데 충분한 핵산 분자들 또는 서열 태그의 수를 지칭하는데 이용된다. 일부 구체예들에 있어서, 각 테스트 시료에 대하여 약 20 내지 40bp의 최소한 약 3 x 10^6 서열 태그가 획득된다. 일부 구체예들에 있어서, 각 테스트 시료는 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 제공하며, 각 서열 태그는 약 20 내지 40bp를 포함한다.

[0055] 용어 "폴리뉴클레오티드", "핵산" 및 "핵산 분자들"은 호환이용되며, 하나의 뉴클레오티드의 펜토즈의 3' 위치가 포스포디에스테르기에 의해 그 다음 펜토즈의 5' 위치에 연결되는 공유결합으로 연결된 뉴클레오티드 서열(가령, RNA의 경우 리보뉴클레오티드, 그리고 DNA의 경우 데옥시리보뉴클레오티드)을 말한다. 상기 뉴클레오티드는 cfDNA 분자 등의 DNA 및 RNA 분자들에 제한되지 않고, 임의의 형태의 핵산 서열들을 포함한다. 용어 "폴리뉴클레오티드"는 제한없이, 단일- 및 이중-가닥으로된 폴리뉴클레오티드를 포함한다.

[0056] 용어 "일부분(portion)"이란 생물학적 시료에서 있는 태아 및 모체 핵산 분자들의 서열 정보의 양이 1 인간 게놈의 서열 정보보다 총량에서 적은 양을 언급할 때 이용된다.

[0057] 본 명세서에서 용어 "테스트 시료"란 생물학적 유체, 세포, 조직, 장기, 또는 유기체로부터 전형적으로 유도된, 복제 수 변이를 위하여 스크리닝되는 핵산 또는 최소한 하나의 핵산 서열이 포함된 핵산 혼합물이 포함된 시료를 지칭한다. 특정 구체예들에 있어서, 상기 시료는 복제 수가 변이를 겪고 있는 것으로 의심되는 최소한 하나의 핵산 서열을 포함한다. 이러한 시료들에는 가래/구강 유체, 양수, 혈액, 혈액 분획, 또는 세침(fine needle) 생검 시료들 (가령, 외과적 생검, 세침 생검, 등등), 소변, 복막 유체, 흉막 유체, 그리고 이와 유사한 것들이 포함되나, 이에 국한되지 않는다. 시료는 대개 인간 개체 (가령, 환자)로부터 채집되지만, 개, 고양이, 말, 염소, 양, 소, 돼지, 등등이 포함되나, 이에 국한되지 않는 임의의 포유류로부터 취한 시료가 복제 수 변이(CNVs) 분석에서 이용될 수 있다. 상기 시료는 생물학적 원천으로부터 획득된 것으로 바로 이용되거나 또는 상기 시료의 성질을 변형시키기 위하여 전처리후 이용될 수 있다. 예를 들면, 이러한 전처리는 혈액으로부터 혈장을 준비하고, 점성 유체를 희석 하는 등등을 포함할 수 있다. 전처리 방법은 여과, 침전, 희석, 증류, 혼합, 원심분리, 동결, 동결건조, 농축, 증폭, 핵산 분절화, 간섭 성분들의 비활성화, 시약의 추가, 용해 등이

또한 관련될 수 있지만, 이에 국한되지 않는다. 이러한 전처리 방법이 테스트 시료에 이용된다면, 이러한 전처리 방법은 전형적으로 때로는 관심 대상의 핵산(들)이, 처리안된 테스트 시료(가령, 즉, 이러한 전처리 방법(들)을 받지 않은 시료)에 핵산에 비례하는 농도로, 테스트 시료에 남아있는 그러한 것이다. 이러한 "처리된" 또는 "가공된(processed)" 시료들은 본 명세서에서 설명된 방법들에 있어서 생물학적 생물학적 "테스트" 시료들로 여전히 간주된다.

[0058] 본 명세서에서 용어 "검정된 시료" 또는 "영향없는 시료"는 테스트 시료 에서 핵산과 비교되는 공지의 복제 수로 존재하는 핵산 혼합물이 포함된 시료를 지칭하며, 그리고 이는 상기 관심 대상 서열에 대하여 정상인, 가령, CNV 또는 홀배수체도 없는 시료이다. 일부 구체예들에 있어서, 검정된 시료들은 서열 마스크 또는 서열 프로파일을 유도하기 위하여 훈련용 세트의 영향없는 훈련용 시료들로 이용된다. 특정 구체예들에 있어서, 검정된 시료들은 고려중인 염색체에 대하여 하나 또는 그 이상의 정규화 염색체 또는 세그먼트들을 식별하는데 이용된다. 예를 들면, 검정된 시료들은 염색체 21에 대하여 정규화 염색체를 식별하는데 이용될 수 있다. 이러한 경우에서, 상기 검정된 시료, 달리 말해서 검증된 시료는 삼염색체성 21 시료가 아닌 시료이다. 또 다른 실시예는 염색체 X에 대한 시료를 검정하는 만큼 오직 여성만을 이용한다. 또한 검정된 시료들은, 영향있는 시료들을 소명하기 위한 임계치 결정, 참조 서열 상에 마스크 영역들을 규정하기 위한 임계치 확인, 게놈의 상이한 영역들에 대한 예상된 커버리지 양을 결정하는 등의 다른 목적들을 위하여 이용될 수 있다.

[0059] 본 명세서에서 용어 "훈련용 세트(training set)"는 영향을 받은 및/또는 영향을 받지 않은 시료들을 포함할 수 있고, 그리고 테스트 시료들의 분석을 위한 모델을 개발하는데 이용되는 훈련용 시료 세트를 지칭한다. 일부 구체예들에 있어서, 상기 훈련용 세트는 영향을 받지 않은 시료들을 포함한다. 이들 구체예들에 있어서, CNV 결정을 위한 임계치는 관심 대상의 복제 수 변이에 대하여 발병되지 않는 시료의 훈련용 세트를 이용하여 확립된다. 훈련용 세트에서 영향을 받지 않은 시료들은 정규화 서열들, 가령, 정규화 염색체를 식별해내기 위한 검증된 검정된 시료로 이용될 수 있고, 그리고 영향을 받지 않은 시료들의 염색체 분량은 상기 서열들, 가령, 관심 대상의 염색체 각각에 대한 임계치를 설정하는데 이용된다. 일부 구체예들에 있어서, 상기 훈련용 세트는 영향을 받은 시료들을 포함한다. 훈련용 세트에서 영향을 받은 시료들은 영향을 받은 테스트 시료들이 영향을 받지 않은 시료들로부터 용이하게 분화될 수 있다는 것을 증명하는데 이용될 수 있다.

[0060] "훈련용 세트(Training set)"는 또한 본 명세서에서 관심 대상 집단의 통계학적 시료의 개별 세트 지칭에서도 이용되며, 각 개별 데이터는 집단에 대하여 일반화시킬 수 있는 관심 대상의 하나 또는 그 이상의 정량적 값을 결정하는데 이용된다. 통계학적 시료는 관심 집단에서 개체의 부분집단이다. 상기 개체는 사람, 동물, 조직, 세포, 다른 생물학적 시료들 (가령, 통계학적 시료는 다중 생물학적 시료들을 포함할 수 있고), 그리고 통계학적 분석을 위한 데이터 점들을 제공하는 다른 개별 엔터티일 수 있다.

[0061] 일부 구체예들에 있어서, 훈련용 세트는 확증 세트와 함께 이용된다. 본 명세서에서 용어 "검증 세트(validation set)"는 통계학적 시료에서 개체 세트 관련해서 이용되는데, 개체의 데이터 세트는 훈련용 세트를 이용하여 결정된 관심 대상의 정량적 값을 검증 또는 평가하는데 이용된다. 일부 구체예들에 있어서, 예로써, 훈련용 세트는 참조 서열용 마스크를 산출하기 위한 데이터를 제공하며; 검증 세트는 마스크를 검증 또는 평가하기 위한 데이터를 제공한다.

[0062] 게놈 데이터가 빈으로 조직화되는 일부 구체예들에 있어서, 영향을 받지 않은 시료들의 훈련용 세트를 이용하여 상기 게놈의 빈에서 데이터 변이를 나타내는 프로파일 또는 웨이브를 획득하며, 이의 프로파일 또는 웨이브는 영향을 받지 않은 시료들에 대해 공통적이나, 관심 대상 서열의 CNV에는 무관한 것으로 보인다. 일부 구체예들에 있어서, 훈련용 세트의 부분집단을 이용하여 관심대상의 CNV에 무관한 편향 또는 변이를 교정하기 위한 프로파일 또는 웨이브가 생성된다. 일부 구체예들에 있어서, 상기 부분집단은 프로파일 또는 웨이브에서 조직적 그리고 지속적 변이를 갖는 훈련용 세트에서 최대 시료 군(group, 群)을 포함한다.

[0063] 본 명세서에서 용어 "프로파일(profile)" 및 "웨이브(wave)"는 빈들에 걸쳐 커버리지의 변이를 지칭하기 위하여 혼용된다. 일부 구체예들에 있어서, 프로파일 또는 웨이브는 CNV를 갖지 않는 공지의 영향을 받지 않은 시료들로부터 획득된다. 이와 같이, 프로파일 및 웨이브는 CNV와 무관한 변이를 나타낸다. 다양한 구체예들에 있어서, 상기 프로파일 또는 웨이브는 CNV 소명이 있기전, 커버리지 측정으로부터 이동된다.

[0064] 본 명세서에서 "복제 수의 평가"는 상기 서열의 복제 수와 관련된 유전적 서열 상태의 통계학적 평가에 관련하여 이용된다. 예를 들면, 일부 구체예들에서, 상기 평가는 유전적 서열의 존재 또는 부재의 결정을 수반한다. 일부 구체예들에 있어서, 상기 평가는 상기 유전적 서열의 부분적 또는 완전한 홀배수체의 결정을 포함한다. 다른 구체예들에서 상기 평가는 유전적 서열의 복제 수에 근거하여 2개 또는 그 이상의 시료들간을

식별하는 것을 포함한다. 일부 구체예들에 있어서, 상기 평가는 통계학적 분석, 가령, 상기 유전적 서열의 복제 수에 근거하여 정규화 및 비교를 포함한다.

[0065] 용어 "검정된 핵산"은 "검정된 서열"과 호환되며, 이는 테스트 서열 또는 테스트 핵산의 양이 비교되는 서열이다. 검정된 서열은 바람직하게는 공지의 구성에서 생물학적 시료중에 존재하는 것으로, 즉 검정된 서열의 양은 공지이다. 일반적으로, 검정된 서열은 "검정된 시료"에 존재하는 서열이다. "관심 대상의 검정된 서열"은 검정된 시료중에서 그 양이 공지된 검정된 서열이며, 의료적 상태를 가진 개체에서의 서열 표현의 차이와 연관되는 서열이다.

[0066] 본 명세서에서 용어 "관심 대상 서열"은 질환이 있는 개체에 대해 건강한 개체에서의 서열 표현의 차이와 연관된 핵산 서열을 지칭한다. 관심 대상 서열은 질환 또는 유전적 상태에서 오인표시된(misrepresented) 가령, 과다- 또는 과소-표시된, 염색체 상의 서열일 수 있다. 관심 대상 서열은 염색체의 일부분, 가령, 염색체 세그먼트, 또는 염색체 전체일 수 있다. 예를 들면, 관심 대상 서열은 홀배수체 상태에서 과다-구성된 염색체이거나, 또는 암에서 과소-구성된 종양-억제자(suppressor)를 코드화하는 유전자일 수 있다. 관심 대상 서열들은 세포의 전체 집단, 또는 부분 집단에서 과다- 또는 과소-구성된 서열들을 포함한다. "관심 대상의 검정된 서열"은 검정된 시료중의 관심 대상 서열이다. "관심 대상의 검사 서열"은 테스트 시료에서의 관심 대상 서열이다.

[0067] 본 명세서에서 용어 "정규화 서열(normalizing sequence)"은 정규화 서열과 연관된 관심 대상 서열에 매핑된 서열 태그의 수를 정규화하는데 이용되는 서열을 지칭한다. 일부 구체예들에 있어서, 정규화 서열은 강건한(robust) 염색체를 포함한다. "강건한 염색체"는 홀배수체일 것 같지 않는 것이다. 인간 염색체가 관련된 일부 경우들에 있어서, 강건한 염색체는 X 염색체, Y 염색체, 염색체 13, 염색체 18, 그리고 염색체 21 이외의 다른 임의의 염색체다. 일부 구체예들에 있어서, 정규화 서열은 시료들 중에 그것이 매핑된 서열 태그 수의 변동성과 정규화 매개변수로써 이용되는 관심 대상 서열의 변동성에 가까운 서열화 운용(sequencing runs) 수에서 변동성(variability)을 나타낸다. 상기 정규화 서열은 하나 또는 그 이상의 영향을 받지 않은 시료들로부터 영향을 받은 시료를 구별해낼 수 있다. 일부 실행에 있어서, 상기 정규화 서열은 다른 잠재적 정규화 서열들 이를 테면 다른 염색체와 비교하였을 때, 하나 또는 그 이상의 영향을 받지 않은 시료들로부터 영향을 받은 시료를 가장 또는 효과적으로 구별해낸다. 일부 구체예들에 있어서, 상기 정규화 서열의 변동성은 시료들과 서열화 운용에 걸쳐 관심 대상 서열에 대하여 염색체 분량에서 변동성으로 산출된다. 일부 구체예들에 있어서, 정규화 서열들은 영향을 받지 않은 시료들의 세트에서 식별된다.

[0068] "정규화 염색체", "정규화 분모(denominator) 염색체", 또는 "정규화 염색체 서열"은 "정규화 서열"의 예들이다. "정규화 염색체 서열"은 단일 염색체 또는 염색체 집단으로 구성될 수 있다. 일부 구체예들에 있어서, 정규화 서열은 2개 또는 그 이상의 강건한 염색체를 포함한다. 특정 구체예들에 있어서, 상기 강건한 염색체는 염색체, X, Y, 13, 18, 그리고 21 이외의 모든 상염색체들이다. "정규화 세그먼트"는 "정규화 서열"의 또다른 예들이다. "정규화 세그먼트 서열"는 염색체의 단일 세그먼트로 구성되거나, 또는 동일한 또는 상이한 염색체의 2개 또는 그 이상의 세그먼트들로 구성될 수 있다. 특정 구체예들에 있어서, 정규화 서열은 변동성, 이를 테면 공정-관련된, 염색체간 (intra-run), 그리고 서열화 간 (inter-run) 변동성에 대하여 정규화되도록 의도된다.

[0069] 본 명세서에서 용어 "구별가능성(differentiability)"이란 하나 또는 그 이상의 영향을 받은, 가령, 홀배수체, 시료들로부터 하나 또는 그 이상의 영향을 받지 않은, 가령, 정상적, 시료들을 구별해낼 수 있도록 정규화 염색체의 특징을 지칭한다. 최대 "구별가능성"을 나타내는 정규화 염색체는 검정된 시료들의 세트에서 관심 염색체에 대한 염색체 분량 분포와 하나 또는 그 이상의 영향을 받은 시료들에서 대응하는 염색체 에서 동일한 관심 염색체에 대한 염색체 분량 사이에 통계학적으로 최대 차이를 제공하는 염색체 또는 염색체 집단이다.

[0070] 본 명세서에서 용어 "변동성(variability)"이란 하나 또는 그 이상의 영향을 받은, 가령, 홀배수체, 시료들로부터 하나 또는 그 이상의 영향을 받지 않은, 가령, 정상적, 시료들을 구별해낼 수 있도록 정규화 염색체의 또다른 특징을 지칭한다. 검정된 시료들의 세트에서 측정된 정규화 염색체의 변동성은 정규화 매개변수로 기능하는 관심 염색체에 매핑된 서열 태그 수에서의 변동성에 근접한 이에 매핑된 서열 태그 수의 변동성을 지칭한다.

[0071] 본 명세서에서 용어 "서열 태그 밀도(sequence tag density)"란 참조 게놈 서열에 매핑된 서열 리드 수를 지칭하며, 가령, 염색체 21에 대한 서열 태그 밀도는 상기 참조 게놈의 염색체 21에 매핑된 서열화 방법에 의해

생성된 서열 리드 수가 된다.

[0072] 본 명세서에서 용어 "서열 태그 밀도 비율"이란 참조 게놈 염색체의 길이에 대한 상기 참조 게놈의 염색체, 가령, 염색체 21에 매핑된 서열 태그의 수의 비율을 말한다.

[0073] 본 명세서에서 용어 "서열 분량(sequence dose)"이란 관심 대상 서열의 경우에 확인된 서열 태그 수와 정규화 서열의 경우에 확인된 서열 태그 수와 관련된 매개변수를 말한다. 일부 경우들에 있어서, 상기 서열 분량은 정규화 서열에 대한 서열 태그 커버리지에 대하여 관심 대상 서열의 서열 태그 커버리지의 비율이다. 일부 경우들에 있어서, 상기 서열 분량이란 정규화 서열의 서열 태그 밀도에 대하여 관심 대상 서열의 서열 태그 밀도와 관련된 매개변수를 지칭한다. "테스트 서열 분량"은 테스트 시료에서 결정된 정규화 서열, 가령, 염색체 9의 밀도에 대한 관심 대상 서열, 가령, 염색체 21의 서열 태그 밀도와 관련된 매개 변수다. 유사하게, "검정된 서열 분량"은 검정된 시료에서 결정된 정규화 서열에 대한 관심 대상 서열의 서열 태그 밀도와 관련된 매개변수다.

[0074] 용어 "커버리지(coverage)"란 규정된 서열에 매핑된 서열 태그의 존재도(abundance)를 지칭한다. 커버리지는 서열 태그 밀도 (또는 서열 태그의 계수), 서열 태그 밀도 비율, 정규화된 커버리지 양, 조정된 커버리지 값, 등등에 의해 정량적으로 표시될 수 있다.

[0075] 용어 "커버리지 분량(coverage quantity)"이란 미가공(raw) 커버리지의 변형(modification)으로써, 빈과 같은 게놈 영역에서의 서열 태그의 상대적 분량(때로는 계수로도 칭함)을 흔히 나타낸다. 커버리지 분량은 상기 게놈 영역에 대한 미가공 커버리지 또는 계수의 정규화, 조정 및/또는 교정에 의해 획득될 수 있다. 예를 들면, 한 영역에 대한 정규화된 커버리지 분량은 그 영역에 매핑된 서열 태그 계수를 전체 게놈에 매핑된 전체 수 서열 태그로 나눔으로써 획득될 수 있다. 정규화된 커버리지 분량은 상이한 시료들간에 빈의 커버리지 비교를 허용하고, 시료들은 상이한 서열화 깊이를 가질 수 있다. 후자는 전체 게놈의 부분 집단에 매핑된 태그 계수로 나눔으로써 전형적으로 획득되기 때문에, 서열 분량과는 상이할 수 있다. 상기 부분 집단은 정규화 세그먼트 또는 염색체다. 정규화안된 것이든 상관없이, 커버리지 분량은 상기 강건한 염색체에서 게놈, G-C 분획 변이, 아웃라이어(outliers), 등등에서 영역간 포괄적 프로파일 변이에 대해 수정될 수 있다.

[0076] 본 명세서에서 용어 "차세대 서열화 (NGS)"란 클론적으로 증폭된 분자들과 단일 핵산 분자들의 대량 병행(parallel) 서열화를 허용하는 서열화 방법을 지칭한다. NGS의 비-제한적 예에는 가역적 염료 종료물질들을 이용하는 합성에 의한 서열화(sequencing-by-synthesis), 그리고 결찰에 의한 서열화(sequencing-by-ligation)가 포함된다.

[0077] 본 명세서에서 용어 "매개변수(parameter)"는 물리적 성질을 특징화시키는 수치적 값을 지칭한다. 빈번하게, 매개변수는 정량적 데이터 세트를 수치적으로 특징화하거나 및/또는 정량적 데이터 세트간의 수치적 관계를 특징화한다. 예를 들면, 염색체에 매핑된 서열 태그의 수와 태그가 매핑된 염색체의 길이 사이의 비율 (또는 비율 함수)이 매개변수다.

[0078] 본 명세서에서 용어 "임계치 값(threshold value)" 및 "검정된 임계치 값"이란 시료, 이를 테면 의료적 상태를 가진 것으로 의심되는 유기체의 핵산이 포함된 테스트 시료를 특징화하기 위한 컷오프(cutoff)로 이용되는 임의의 수를 말한다. 이러한 매개변수 값을 야기하는 시료는 이 유기체가 의료적 상태를 가지는 지를 판단하기 위하여 상기 임계치는 매개변수 값과 비교될 수 있다. 특정 구체예들에 있어서, 검증된 즉 검정된 임계치 값은 검증된 데이터 세트를 이용하여 산출되며, 그리고 유기체 에서 복제 수 변이, 가령, 홀배수체의 진단 한계로 제시된다. 임계치가 본 명세서에서 공개된 방법들의 의해 획득된 결과를 초과하는 경우, 개체는 복제 수 변이, 가령, 삼염색체성 21를 가진 것으로 진단받을 수 있다. 본 명세서에서 설명된 방법들을 위한 적절한 임계치 값은 훈련용 시료 세트에 대하여 산출된 정규화된 값 (가령 염색체 분량, NCVs 또는 NSVs)을 분석함으로써 확인될 수 있다. 임계치 값은 검증된 (가령, 영향을 받지 않은) 시료들과 영향을 받은 시료들을 모두 포함하는 훈련용 세트 에서 검증된 (가령, 영향을 받지 않은) 시료들을 이용하여 확인될 수 있다. 염색체 홀배수체를 가진 것으로 알려진 훈련용 세트 (가령, 영향을 받은 시료들) 에서 시료를 이용하여 상기 선택된 임계치가 테스트 세트에서영향을 받지 않은 시료들로부터 영향을 받은 시료를 구별해내는데 (본 명세서의 실시예들 참고) 유용하다는 것을 확인할 수 있다. 임계치의 선택은 사용자가 분류하고자 원하는 신뢰 수준에 따라 달라진다. 일부 구체예들에 있어서, 적절한 임계치 값를 확인하기 위하여 이용된 훈련용 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 검증된 시료들을 포함한다. 임계치 값의 진단학적 유용성을 개

선시키기 위하여 더 큰 검증된 시료 세트를 이용하는 것이 유익할 수 있다.

[0079] 용어 "빈(bin)"은 서열의 세그먼트 또는 게놈의 세그먼트를 지칭한다. 일부 구체예들에 있어서, 빈은 상기 게놈 또는 염색체에서 서로 연속되거나, 분리되어 떨어져 있다. 각 빈은 참조 게놈에서 뉴클레오티드의 서열을 특정시킬 수 있다. 상기 빈의 크기는 특정 응용에 요구되는 분석 및 서열 태그 밀도에 따라 1 kb, 100 kb, 1Mb, 등등이 될 수 있다. 참조 서열에서 이들 위치에 추가하여, 빈은 다른 특징들, 이를 테면 시료 커버리지 및 서열 구조 특징들, 이를 테면 G-C 분획을 가질 수 있다.

[0080] 본 명세서에서 용어 "마스킹 임계치(masking threshold)"는 서열 빈에서 서열 태그 수에 기반된 값이 비교되는 분량을 지칭하며, 여기에서 마스킹 임계치를 초과하는 값을 가진 빈은 감춰진다. 일부 구체예들에 있어서, 마스킹 임계치는 백분위 점수, 절대적 계수, 매핑 품질 스코어, 또는 다른 적절한 값일 수 있다. 일부 구체예들에 있어서, 마스킹 임계치는 다중 영향을 받지 않은 시료들에서 변이계수의 백분위로 정의될 수 있다. 다른 구체예들에서, 마스킹 임계치는 매핑 품질 스코어, 가령, 참조 게놈에 서열 리드를 나열시키는 신뢰도와 관련된 MapQ 스코어로 정의될 수 있다. 마스킹 임계치 값은 복제 수 변이 (CNV) 임계치 값과 상이할 수 있으며, 후자는 CNV와 관련된 의료적 상태를 갖는 것으로 의심되는 유기체 즉 생물 유래의 핵산이 포함된 시료를 특징화하는 컷오프이다. 일부 구체예에서, CNV 임계치 값은 본 명세서의 도처에서 설명된 정규화된 염색체 값 (NCV) 또는 정규화된 세그먼트 값 (NSV)에 상대적으로 정의된다.

[0081] 본 명세서에서 용어 "정규화된 값(normalized value)"이란 관심 대상의 서열 (가령 염색체 또는 염색체 세그먼트)에 대하여 확인된 서열 태그 수를, 정규화 서열 (가령 정규화 염색체 또는 정규화 염색체 세그먼트)에 대하여 확인된 서열 태그 수에 관련된 수치를 말한다. 예를 들면, "정규화된 값"은 본 명세서의 도처에서 설명된 염색체 분량일 수 있거나, 또는 NCV일 수 있고, 또는 본 명세서의 도처에서 설명된 NSV일 수 있다.

[0082] 용어 "리드(read)"는 핵산 시료의 일부분으로부터 서열 리드를 지칭한다. 반드시 그런 것은 아니지만, 전형적으로 리드는 시료에서 연속 염기쌍들의 짧은 서열을 나타낸다. 리드는 시료 부분의 염기쌍 서열에 의해 기호(ATCG에서)로 나타낼 수 있다. 리드는 메모리 장치에 보관될 수 있고, 참조 서열과 일치되는지 또는 다른 기준에 부합되는 지를 판단하기 위하여 적절하게 가공될 수 있다. 리드는 서열화 장치로부터 직접적으로 획득될 수 있거나 또는 이 시료와 관련된 저장된 서열 정보로부터 간접적으로 획득될 수 있다. 일부 경우들에 있어서, 리드는 더 큰 서열 또는 영역을 식별해내기 위하여 이용되는, 가령, 염색체 또는 게놈 영역 또는 유전자에 나열되는, 구체적으로 할당될 수 있는 충분한 길이(가령, 최소한 약 25 bp)의 DNA 서열이다.

[0083] 용어 "게놈 리드"는 개체의 전체 게놈에서 임의의 세그먼트들의 리드와 관련하여 이용된다.

[0084] 본 명세서에서 용어 "서열 태그(sequence tag)"는 용어 "매핑된 서열 태그(mapped sequence tag)"와 호환되며, 정렬에 의해 더 큰 서열, 가령, 참조 게놈에 구체적으로 할당된, 가령, 매핑된 서열 리드를 지칭한다. 매핑된 서열 태그는 참조 게놈에 독특하게 매핑되며, 가령, 태그들은 상기 참조 게놈의 단일 위치에 할당된다. 다른 언급이 없는 한, 참조 서열 상에서 동일한 서열에 매핑된 태그는 한번 계수된다. 태그는 데이터 구조 또는 다른 데이터 집합(assemblages)으로 제공될 수 있다. 특정 구체예들에 있어서, 태그는 리드 서열과 이 리드에 대한 연관 정보, 이를 테면 상기 게놈에서 서열의 위치, 가령, 염색체 상에서 위치를 포함한다. 특정 구체예들에 있어서, 상기 위치는 양성 가닥 방향에 대하여 명시된다. 태그는 참조 게놈에 나열함에 있어서 불합치(mismatch)의 한계량을 제공하기 위하여 정의된다. 일부 구체예들에 있어서, 참조 게놈 상에 하나 이상의 위치에 매핑될 수 있는 태그, 가령, 독특하게 매핑되지 않는 태그는 이 분석에 포함되지 않을 수 있다.

[0085] 용어 "비-겹침 서열 태그(non-redundant sequence tag)"란 동일한 부위로 매핑되지 않는 서열 태그를 지칭하는데, 일부 구체예들에서 정규화된 염색체 값 (NCVs)을 결정하기 위한 목적으로 계수된다. 때로, 다중 서열 리드는 참조 게놈 상에 동일한 위치에 나열되어, 겹침(redundant) 또는 중복된 서열 태그가 생성된다. 일부 구체예들에 있어서, 동일한 위치에 매핑된 중복 서열 태그는 생략되거나 또는 NCVs를 측정하기 위한 목적으로 하나의 "비-겹침 서열 태그"로 계수된다. 일부 구체예들에 있어서, 배제되지 않은 부위들에 나열된 비-겹침 서열 태그는 NCVs를 결정하기 위하여 "배제되지 않은-부위 계수" (NES 계수)가 생성되도록 계수된다.

[0086] 용어 "부위(site)"는 참조 게놈 상에 독특한 위치 (가령 염색체 ID, 염색체 위치 및 방향)를 지칭한다. 일부 구체예들에 있어서, 부위는 서열 상의 잔기(residue), 서열 태그, 또는 세그먼트의 위치가 될 수 있다.

[0087] "배제된 부위들"은 서열 태그를 계수하기 위한 목적으로 배제된 참조 게놈 영역에서 볼 수 있는 부위들이다. 일부 구체예들에 있어서, 배제된 부위들은 반복적 서열들이 포함된 염색체 영역들, 가령, 중심체(centromeres) 및 말단소체(telomeres), 그리고 하나 이상의 염색체에 공통적인 염색체 영역들, 가령, Y-염색체 상에 있는 영

역들이 X 염색체에도 존재하는 영역에서 발견된다.

[0088] "비-배제된 부위들" (NESs)은 서열 태그를 계수하기 위한 목적으로 배제되지 않은 참조 게놈 영역에 있는 부위들이다.

[0089] "비-배제된-부위 계수" (NES 계수)는 참조 게놈 상에 NESs에 매핑된 서열 태그의 수다. 일부 구체예들에 있어서, NES 계수는 NESs에 매핑된 비-겹침(non-redundant) 서열 태그의 숫자들이다. 일부 구체예들에 있어서, 커버리지 및 관련된 매개변수들, 이를 테면, 정규화된 커버리지 분량, 포괄적 프로파일 제거된 커버리지 분량, 그리고 염색체 분량은 NES 계수에 근거한다. 한 실시예에서, 염색체 분량은 정규화 염색체에 대한 NES 계수의 수에 대한 관심 염색체의 NES 계수의 수의 비율로 산출된다.

[0090] 정규화된 염색체 값 (NCV)은 훈련/검증된 시료들 세트의 커버리지에 대한 테스트 시료의 커버리지와 관계한다. 일부 구체예들에 있어서, NCV는 염색체 분량에 근거한다. 일부 구체예들에 있어서, NCV는 테스트 시료 에서 관심 염색체의 염색체 분량과 검증된 시료들 세트에 대응하는 염색체 분량의 평균 간의 차이에 관한 것으로, 다음과 같이 산출될 수 있다:

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

[0091]

[0092] 이때 $\hat{\mu}_j$ 및 $\hat{\sigma}_j$ 는 검증된 시료들의 세트에서 j-번째 염색체 분량에 대하여 각각 차례로 추정된 평균 및 표준편차이며, 그리고 x_{ij} 는 테스트 시료 i에 대하여 관찰된 j-번째 염색체 비율(분량)이다.

[0093] 일부 구체예들에 있어서, NCV는 테스트 시료에서의 관심 대상 염색체의 염색체 분량이 동일한 플로우 쎌(flow cell) 상에서 서열화된 복합화된 시료들에서 대응하는 염색체 분량의 중앙값(median)에 관련됨으로써 다음과 같이 "즉시(on the fly)" 산출될 수 있다:

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

[0094]

[0095] 이때 M_j 는동일한 플로우 쎌 상에서 서열화된 복합화된 시료들의 세트에서 j-번째 염색체 분량에 대하여 추정된 중앙값이며; $\hat{\sigma}_j$ 는 하나 또는 그 이상의 플로우 쎌 상에서 서열화된 복합화된 시료들의 하나 또는 그 이상의 세트 에서 j-번째 염색체 분량에 대한 표준 편차이며, 그리고 x_i 는 테스트 시료 i에 대한 관찰된 j-번째 염색체 분량이다. 이 구체예에 있어서, 테스트 시료 i는 M_j 가 측정되는 동일한 플로우 쎌 상에서 서열화된 복합된 시료들중에 하나이다.

[0096] 예를 들면, 한 개의 플로우 쎌 상에 64개의 복합화된 시료들중 하나로 서열화되는, 테스트 시료 A 에서 관심 염색체 21의 경우, 테스트 시료 A 에서 염색체 21의 NCV는 시료 A에서 염색체 21의 분량으로부터 64개의 복합화된 시료들에서 측정된 염색체 21의 분량의 중앙값을 빼고, 플로우 쎌 1 또는 추가 플로우 쎌, 가령, 20 상에서 64개의 복합된 시료들에 대해 측정된 염색체 21에 대한 분량의 표준 편차로 나누어 산출된다.

[0097] 본 명세서에서 이용된 바와 같이, 용어 "정렬된(aligned)", "정렬(alignment)", 또는 "정렬되는(aligning)"이란 리드 또는 태그를 참조 서열과 비교하고, 그렇게 함으로써 상기 리드 서열이 참조 서열에 포함되는 지를 판단하는 공정을 지칭한다. 상기 참조 서열에 상기 리드가 포함되어 있는 경우, 이 리드는 상기 참조 서열에 매핑될 수 있거나, 또는 특정 구체예들에 있어서, 상기 참조 서열에서 특정 위치에 매핑될 수 있다. 일부 경우들에 있어서, 정렬은 리드가 특정 참조 서열의 구성원인지 아닌지 (가령, 상기 리드가 상기 참조 서열에서 존재하는지 또는 존재하지 않는지)를 간단히 말해준다. 예를 들면, 인간 염색체 13의 경우 상기 참조 서열에 대한 리드의 정렬은 염색체 13의 경우 상기 리드가 참조 서열에 존재하는 지를 알려줄 것이다. 이 정보를 제공

하는 도구는 세트 구성원 검사기로 불릴 수 있다. 일부 경우들에 있어서, 정렬은 상기 리드 또는 태그가 매핑되는 참조 서열에서의 위치를 추가로 나타낸다. 예를 들면, 상기 참조 서열이 전체 인간 게놈 서열인 경우, 정렬은 리드가 염색체 13 상에 존재함을 나타낼 수 있고, 그리고 상기 리드가 염색체 13의 특정 가닥(strand) 및/또는 부위 상에 존재함을 더 나타낼 수도 있다.

[0098] 정렬된 리드 또는 태그는 참조 게놈의 공지의 서열에 대하여 이들 핵산 분자들의 순서로 정합(match)으로 식별되는 하나 또는 그 이상의 서열들이다. 정렬은 전형적으로 컴퓨터 알고리즘에 의해 실행되지만, 본 명세서에서 공개된 방법들을 실행하기 위한 합당한 시간 동안 리드를 정렬하는 것이 불가능할 수도 있기 때문에 수작업으로 실행될 수 있다. 정렬되는 서열들로부터 알고리즘의 한 예는 Illumina Genomics Analysis pipeline의 일부분으로 분포된 뉴클레오티드 데이터 (ELAND) 컴퓨터 프로그램의 효과적인 국소 정렬(Efficient Local Alignment)이다. 대안으로, Bloom 필터 또는 유사한 세트 구성원 검사기는 참조 게놈들에 리드를 정렬하는데 이용될 수 있다. 2011년 10월 27일자로 제출된 US 특허 출원 번호 61/552,374 참고, 이의 전문이 본 명세서에 참고로 포함된다. 정렬에서 서열 리드의 정합은 100% 서열 정합 또는 100% 미만 (불완전한 정합)일 수 있다.

[0099] 용어 "정렬 프로파일(alignment profile)"은 참조 관심 대상 서열에서 염기쌍 빈으로 식별될 수 있는 위치에 정렬된 서열 태그의 분포와 관련하여 이용된다.

[0100] 본 명세서에서 용어 "매핑(mapping)"은 정렬에 의해 더 큰 서열, 가령, 참조 게놈에 서열 리드를 특이적으로 할당하는 것을 지칭한다.

[0101] 본 명세서에서 이용된 바와 같이, 용어 "참조 게놈(reference genome)" 또는 "참조 서열(reference sequence)"은 임의의 유기체 또는 바이러스의 부분적 또는 온전한 임의의 특정 공지의 게놈 서열을 지칭하며, 이는 개체로부터 확인된 서열들을 언급할 때 이용될 수 있다. 예를 들면, 인간 개체들 뿐만 아니라 많은 다른 유기체들에 대한 참조 게놈은 ncbi.nlm.nih.gov.의 National Center for Biotechnology Information에서 볼 수 있다. "게놈(genome)"은 핵산 서열들이 발현된 유기체 또는 바이러스의 완전한 유전적 정보를 지칭한다.

[0102] 다양한 구체예들에 있어서, 상기 참조 서열은 이에 정렬된 리드보다 상당히 더 크다. 예를 들면, 이는 최소한 약 100 배 더 크고, 또는 최소한 약 1000 배 더 크고, 또는 최소한 약 10,000 배 더 크고, 또는 최소한 약 105 배 더 크고, 또는 최소한 약 106 배 더 크고, 또는 최소한 약 107 배 더 클 수 있다.

[0103] 한 실시예에서, 상기 참조 서열은 전장의 인간 게놈의 서열이다. 이러한 서열들은 게놈 참조 서열들로 지칭될 수 있다. 또다른 실시예에 있어서, 상기 참조 서열은 특이적 인간 염색체, 이를 테면 염색체 13로 한정된다. 일부 구체예들에 있어서, 참조 Y 염색체는 인간 게놈 형태(version) hg19의 Y 염색체 서열이다. 이러한 서열들은 염색체 참조 서열들로 지칭될 수 있다. 참조 서열들의 다른 예로는 다른 종들의 게놈들, 뿐만 아니라 임의의 종의 염색체, 아염색체 영역들 (이를 테면 가닥), 등등을 포함한다.

[0104] 다양한 구체예들에 있어서, 상기 참조 서열은 다중 개체들로부터 유래된 콘센수스 서열 또는 다른 조합이다. 그러나, 특정 적용에 있어서, 상기 참조 서열은 특정 개체로부터 취할 수도 있다.

[0105] 본 명세서에서 용어 "임상적으로-관련된 서열"이란 유전적 또는 질환 상태와 연관된 또는 연루된 것으로 공지된 또는 의심되는 핵산 서열을 지칭한다. 임상적으로-관련된 서열의 부재 또는 존재의 결정은 의료적 상태를 진단하거나 또는 진단을 확인하거나, 또는 질환의 발달에 대한 예후를 제공하는데 유용할 수 있다.

[0106] 본 명세서에서 핵산 또는 핵산 혼합물 문맥에서 이용될 경우 용도 "유래된(derived)"이란 상기 핵산(들)이, 그들이 기인하는 원천으로부터 획득되었음을 의미한다. 예를 들면, 한 구체예에서 2개의 상이한 게놈들로부터 유래된 핵산 혼합물은 상기 핵산, 가령, cfDNA는 자연 발생적 공정들, 이를 테면 괴사 또는 자가사멸(apoptosis)을 통하여 세포로부터 자연적으로 방출된다는 의미다. 또다른 구체예에서, 2개의 상이한 게놈들로부터 유래된 핵산 혼합물은 상기 핵산이 개체에서 2개의 상이한 유형의 세포로부터 추출되었음을 의미한다.

[0107] 본 명세서에서 특이적 정량적 값을 획득하는 내용에서 이용된 용어 "기반된(based on)"이란 산출량(output)으로 특이적 정량적 값을 산출하기 위하여 입력량(input)으로 또다른 수량을 이용한다는 것을 말한다.

[0108] 본 명세서에서 용어 "환자 시료(patient sample)"란 환자, 가령, 의학적 주목 내지 배려, 간호 또는 치료를 받는 자로부터 획득된 생물학적 시료를 말한다. 상기 환자 시료는 본 명세서에서 설명된 임의의 시료들이 될 수 있다. 특정 구체예들에 있어서, 상기 환자 시료는 비-침습성 과정들, 가령, 말초 혈액 시료 또는 대변 시료에 의해 획득된다. 본 명세서에서 설명된 방법들은 인간에 한정될 필요가 없다. 따라서, 상기 환자 시료가 비-인간 포유류 (가령, 고양이, 돼지, 말, 소, 그리고 이와 유사한 것들)로부터 얻은 시료가 되는 경우들과 같이

다양한 수의학적 응용들이 고려된다.

[0109] 본 명세서에서 용어 "혼합된 시료(mixed sample)"란 상이한 게놈들로부터 유래된 핵산 혼합물이 포함된 시료를 지칭한다.

[0110] 본 명세서에서 용어 "모체 시료(maternal sample)"란 임신한 개체, 가령, 여성으로부터 획득된 생물학적 시료를 지칭한다.

[0111] 본 명세서에서 용어 "생물학적 유체"란 생물학적 원천으로부터 취한 유체를 말하며, 예를 들면, 혈액, 혈청, 혈장, 가래, 세척(lavage)액, 뇌척수액, 소변, 정액, 땀, 눈물, 타액, 그리고 이와 유사한 것들을 포함한다. 본 명세서에서 이용된 바와 같이, 용어 "혈액", "혈장" 및 "혈청"은 분획 또는 이의 가공된 부분들을 명시적으로 포괄한다. 유사하게, 시료가 생검, 스왑(swab), 도말표본(smear), 등등인 경우, 상기 "시료"는 생검, 스왑, 도말표본, 등등으로부터 유도된 가공된 분획 또는 부분을 명시적으로 포괄한다.

[0112] 본 명세서에서 용어 "모체 핵산" 및 "태아 핵산"은 임신한 여성 개체의 핵산과 이 임신한 여성이 출산하게 되는 태아의 핵산을 각각 지칭한다.

[0113] 본 명세서에서 이용된 바와 같이, 용어 "~에 대응하는"이란 상이한 대상의 게놈에 존재하고, 그리고 모든 게놈들에서 반드시 동일한 서열을 갖지는 않지만, 관심 대상 서열, 가령, 유전자 또는 염색체의 유전적 정보라기보다는 신원(identity)을 제공하는 핵산 서열, 가령, 유전자 또는 염색체를 흔히 지칭한다.

[0114] 본 명세서에서 바람직한 시료과 연계하여 이용된 바와 같이, 바람직한 시료와 관련되어 이용된 용어 "실질적으로 무 세포(cell free)"란 세포 성분들은 시료와 정상적으로 연관된 세포 성분들이 제거된 바람직한 시료 조제물이 포괄된다. 예를 들면, 혈장 시료는 시료에 정상적으로 연관된 혈액 세포, 가령, 적혈구를 제거함으로써, 실질적으로 무세포를 만든다. 일부 구체예들에 있어서, 실질적으로 무 세포 시료들은 세포를 제거하기 위하여 가공되어, CNV를 테스트하기 위한 바람직한 유전적 물질이 될 수 있다.

[0115] 본 명세서에서 이용된 바와 같이, 용어 "태아 분획(fetal fraction)"은 태아 핵산 및 모체 핵산이 포함된 시료에 존재하는 태아 핵산 분획을 지칭한다. 태아 분획은 다른 혈액 에서 cfDNA를 특징화하는데 흔히 이용된다.

[0116] 본 명세서에서 이용된 바와 같이 용어 "염색체(chromosome)"는 DNA 및 단백질 성분들 (구체적으로 히스톤)이 포함된 크로마틴 가닥으로부터 유래된 살아있는 세포의 유전체질-함유한 유전자를 지칭한다. 본 명세서에서는 관례적으로 국제적으로 인지된 개별 인간 게놈 염색체 넘버링 시스템이 이용된다.

[0117] 본 명세서에서 이용된 바와 같이, 용어 "폴리뉴클레오티드 길이"는 참조 게놈의 서열 또는 영역에서 핵산 분자들 (뉴클레오티드)의 절대적 수를 지칭한다. 용어 "염색체 길이"란 World Wide Web상에서 |genome|.|ucsc|.|edu/cgi-bin/hgTracks?hgsid=167155613&chromInfoPage=에서 볼 수 있는 인간 염색체의 NCBI36/hg18 어셈블리에서 제공되는 염기쌍에 있는 염색체의 공지 길이를 말한다.

[0118] 본 명세서에서 용어 "대상(subject)"이란 인간 대상 뿐만 아니라 비-인간 대상, 이를 테면 포유류, 무척추동물, 척추동물, 곰팡이, 효모, 박테리아, 그리고 바이러스를 지칭한다. 본 명세서의 실시예들은 인간에 관련되기는 하지만, 언어는 인간 게놈들에 주로 관계되며, 본 명세서에서 설명된 개념은 임의의 식물 또는 동물의 게놈에 적용가능하고, 그리고 수의학적 의학, 동물 과학, 연구 실험실 및 이와 같은 분야에 유용하다.

[0119] 본 명세서에서 용어 "상태(condition)"란 모든 질환 및 장애가 포함된 광범위한 의미로 "의료적 상태"를 말하지만, [부상] 및 정상적 건강 상황, 이를 테면 개인의 건강, 의학 지원으로부터 영향을 줄 수 있는 임신, 또는 의학 처치를 위한 암시를 가질 수 있는 것들이 포함될 수 있다.

[0120] 본 명세서에서 염색체 홀배수체에서 언급되는 용어 "완전한(complete)"이란 전체 염색체의 획득 또는 상실을 지칭한다.

[0121] 본 명세서에서 염색체 홀배수체에서 언급되는 용어 "부분적(partial)"이란 일부분, 가령, 염색체의 세그먼트의 증대 또는 상실을 지칭한다.

[0122] 본 명세서에서 용어 "모자이크(mosaic)"는 단일 수정란으로부터 발달된 하나의 개체에서 상이한 핵형을 가진 두 집단의 세포가 존재함을 표시할 때 언급된다. 모자이크현상(Mosaicism)은 발생 동안 성체 세포의 오직 부분 집단으로만 전파되는 돌연변이로부터 야기될 수 있다.

[0123] 본 명세서에서 용어 "비-모자이크(non-mosaic)"란 한 가지 핵형의 세포로 구성된 유기체, 가령, 인간 태아를 지칭한다.

[0124] 본 명세서에서 염색체 분량를 결정할 때 언급되는 용어 "염색체를 이용하는"이란 염색체에서 획득되는 서열 정보, 가령, 염색체로부터 획득된 서열 태그 수를 이용한다는 것을 말한다.

[0125] 본 명세서에서 이용된 용어 "민감성(sensitivity)"은 참양성(true positives)의 수를 참양성과 가음성의 합으로 나눈 것과 같다.

[0126] 본 명세서에서 이용된 바와 같이 용어 "특이성(specificity)"은 참음성(true negatives)의 수를 참 음성과 가양성의 합으로 나눈 것과 같다.

[0127] 본 명세서에서 용어 "보강하다(enrich)"란 모체 시료의 일부분에 포함된 다형태(polymorphic) 표적 핵산을 증폭시키고, 그리고 상기 증폭된 산물은 해당 부분이 빠진 모체 시료의 나머지와 복합시키는 공정을 지칭한다. 예를 들면, 모체 시료의 나머지는 원래 모체 시료일 수 있다.

[0128] 본 명세서에서 용어 "원래 모체 시료(original maternal sample)"란 다형태 표적 핵산을 증폭시키기 위하여 일부가 분리된 원천이 되는 임신한 개체, 가령 여성으로부터 획득된 비-보강된 생물학적 시료를 지칭한다. 상기 "원래 시료"란 임신한 개체, 그리고 이의 가공된 분획, 가령, 모체 혈장 시료로부터 추출된 정제된 cfDNA 시료로부터 회득된 임의의 시료일 수 있다.

[0129] 본 명세서에서 이용된 바와 같이, 용어 "프라이머(primer)"는 연장 산물의 합성 유도 조건 (가령, 상기 조건은 뉴클레오티드, 유도물질, 이를 테면 DNA 중합효소, 그리고 적절한 온도 및 pH)하에 두었을 때, 합성의 개시점으로 작용할 수 있는 단리된 올리고뉴클레오티드를 지칭한다. 상기 프라이머는 증폭에서 최대 효과를 위하여 바람직하게는 단일 가닥이지만, 대안으로 이중 가닥일 수 있다. 이중 가닥인 경우, 상기 프라이머는 우선 연장 산물을 만들기 위하여 이용되기 전 이의 가닥들을 분리시키기 위하여 우선 처리된다. 바람직하게는, 상기 프라이머는 올리고데옥시리보뉴클레오티드이다. 상기 프라이머는 유도 물질 존재 하에 연장 산물의 합성을 프라임시킬 수 있을 정도로 충분히 길어야 한다. 상기 프라이머의 정확한 길이는 온도, 프라이머의 원천, 방법의 용도, 그리고 프라이머 기획에 이용되는 매개변수들이 포함된 많인 인자들에 따라 달라진다.

[0130] 구절 "투여되는 원인(cause to be administered)"이란 의학 전문인 (가령, 의사), 또는 개체에게 문제가 되는 물질(들)/화합물(들)의 투여를 제어 및/또는 허용하는 개체의 의학적 관리를 제어 또는 지시하는 자에 의해 취해지는 활동을 지칭한다. 투여되는 원인은 진단 및/또는 적절한 치료요법적 또는 예방학적 섭생의 결정, 및/또는 개체를 위한 특정 물질(들)/화합물들의 처방이 관련될 수 있다. 이러한 처방에는 예를 들면, 처방전의 초안 작성, 의학 기록에 주석달기, 그리고 이와 유사한 것들이 포함될 수 있다. 유사하게, 가령, 진단학적 과정에 대하여 "실시되는 원인(cause to be performed)"은 전문인 (가령, 의사), 또는 개체에서 하나 또는 그 이상의 진단 프로토콜을 제어 및/또는 허용하는 개체의 의학적 관리를 제어 또는 지시하는 자에 의해 취해지는 활동을 지칭한다.

[0131] **개요**

[0132] 다양한 출생전 비침습성 진단 (NIPD) 방법들은 모체 유체, 이를 테면 말초 혈액에서 이용가능한 태아 기원의 cfDNA를 이용한다. 많은 NIPD 방법들은 임산부의 태아의 cfDNA가 질환들 또는 표현형과 관련된 유전적 서열들에서 복제 수 변이를 품고 있는지를 판단하기 위하여 모체 말초 혈액으로부터 cfDNA를 추출하고, 서열화하고, 그리고 정렬한다. 상기 추출된 그리고 서열화된 cfDNA는 서열 리드를 제공하며, 그 다음 참조 게놈에 매핑된다. 상기 참조 게놈들 상의 독특한 위치 또는 부위에 매핑된 서열 리드는 서열 태그로 지칭된다. 관심 대상 서열에 매핑된 서열 태그 수를 이용하여 상기 관심 대상 서열의 복제 수 또는 복제 수 변이를 결정할 수 있다.

[0133] 관심 대상 서열에 매핑된 서열 태그 수는 커버리지로 지칭된다. 유전적 서열의 빈 또는 영역에 대한 커버리지는 또다른 영역에 대한 하나의 영역의 상대적인 존재도(abundance) 또는 또다른 시료에 대한 하나의 시료의 상대적 존재도를 계산하기 위한 데이터를 제공한다. 관심 대상 서열의 커버리지가 비정상적으로 낮거나 또는 높을 때, 서열의 복제 수 변이를 암시할 수 있다.

[0134] 2개 또는 그 이상의 상이한 게놈들로부터 유래된 핵산 혼합물이 포함되고, 또한 하나 또는 그 이상의 관심 대상 서열의 양이 공지되어 있거나 또는 상이한 것으로 의심되는 테스트 시료에 있어서 상이한 관심 대상 서열의 복제수 및 복제 수 변이(CNV)를 측정하기 위하여 본 명세서에서 방법들, 장치, 그리고 시스템이 설명된다. 본 명세서에서 공개된 방법들과 장치들에 의해 결정된 복제 수 변이는 전체 염색체의 증대 또는 상실, 현미경적으로

볼 수 있는 매우 큰 염색체 세그먼트들이 관련된 변이, 그리고 크기가 단일 뉴클레오티드부터 킬로베이스 (kb) 까지, 메가베이스 (Mb) 범위의 DNA 세그먼트들의 초현미경적 복제 수 변이의 존재도를 포함한다.

[0135] 일부 구체예들에 있어서, 모체 및 태아 무 세포 DNA가 포함된 모체 시료들을 이용하여 태아의 복제 수 변이 (CNV)를 결정하는 방법들이 제시된다. 본 명세서에서 공개된 일부 구체예들에서는 시료내 GC-함량 편향을 제거함으로써, 서열 데이터 분석의 민감성 및/또는 특이성을 개선시키는 방법들이 제시된다. 일부 구체예들에 있어서, 시료내 GC-함량 편향의 제거는 영향을 받지 않은 훈련용 시료들에 걸쳐 잘 알려진 조직적 변이에 대하여 교정된 서열 데이터에 근거한다.

[0136] 공개된 일부 구체예들은 잡음이 낮고, 신호가 높은 서열 커버리지 수량을 결정하는 방법들이 제시되는데, 통상적인 방법들에 의해 획득된 서열 커버리지 분량과 비교하여 개선된 민감성, 선택성, 및/또는 효과로 다양한 유전적 상태와 관련된 복제수를 결정하기 위하여 데이터를 제시한다. 설명된 공정은 고려중인 게놈 (가령, 태아의 게놈)으로부터 상대적으로 낮은 DNA 분획을 보유하는 시료에서 신호를 개선시키는데 특히 효과적인 것으로 밝혀졌다. 이러한 시료의 예로는 형제간 쌍태, 세쌍둥이, 등등을 가진 임신 개체의 모체 혈액 시료이며, 이때 상기 공정은 태아중 하나의 게놈에서 복제 수 변이를 평가한다. 또다른 실시예는 임상 증후군과 관련된 몇 메가베이스 크기의 상대적으로 짧은 아염색체성 영역들의 CNV이다.

[0137] 상기 방법들은 임의의 태아 홀배수체의 CNV, 그리고 다양한 의료적 상태와 연관된 것으로 공지된 또는 의심받는 CNVs를 결정하는데 적용가능하다. 인간 개체와 관련된 일부 구체예들에 있어서, 본 방법에 따라 결정될 수 있는 CNV는 염색체 1-22, X와 Y의 임의의 하나 또는 그 이상의 삼체성(trisomies) 및 단체성(monosomies), 기타 염색체 다체성(polysomies), 그리고 상기 염색체의 임의의 하나 또는 그 이상의 세그먼트들의 결손 및/또는 중복을 포함하며, 이는 테스트 시료의 상기 핵산을 단지 한번 서열화함으로써 탐지될 수 있다. 임의의 홀배수체는 테스트 시료의 상기 핵산의 오직 한번의 서열화에 의해 획득된 서열화 정보로부터 결정될 수 있다.

[0138] 인간 게놈에서 CNV는 인간 다양성(diversity) 및 질환에 대한 질병소인에 상당히 영향을 끼친다(Redon et al., Nature 23:444-454 [2006], Shaikh et al. Genome Res 19:1682-1690 [2009]). CNVs는 상이한 기전들을 통하여 유전적 질환에 기여하는 것으로 알려져 있으며, 대부분의 경우 유전자 정량(dosage)의 불균형 또는 유전자 붕괴(disruption)를 야기한다. 유전적 장애들에 대하여 CNVs의 직접적인 상관관계에 추가하여, CNVs는 유해할 수 있는 표현형 변화를 중개하는 것으로 알려져 있다. 최근, 복합 장애들, 이를 테면 자폐증, ADHD, 그리고 정신분열병에서 정상적인 대조와 비교하였을 때 희귀한 또는 데노보(de novo) CNVs의 부하 증가에 대한 몇몇 연구 보고가 있었고, 희귀한 또는 독특한 CNVs의 잠재적 병원성을 강조한다 (Sebat et al., 316:445 - 449 [2007]; Walsh et al., Science 320:539 - 543 [2008]). CNV는 결손, 중복, 삽입, 그리고 불균형 전위 사건들에 의해 주로 게놈 재배열로부터 발생된다.

[0139] 본 명세서에서 설명된 방법들 및 장치들은 차세대 서열화 기술 (NGS)을 이용할 수 있는데, 이는 대량 병행 서열화다. 특정 구체예들에 있어서, 클론적으로 증폭된 DNA 주형 또는 단일 DNA 분자들은 플로우 쎌 에서서 대량 병행 방식으로 서열화된다 (가령 Volkerding et al. Clin Chem 55:641-658 [2009]; Metzker M Nature Rev 11:31-46 [2010]에서 설명됨). 대량-처리 서열 정보에 추가하여, NGS는 정량적 정보를 제공하는데, 이때 각 서열 리드는 개별 클론성 DNA 주형 또는 단일 DNA 분자를 나타내는 계수가능한 "서열 태그"다. NGS의 서열화 기술은 파이로서열화(pyrosequencing), 가역적 염료 종료물질들과 함께 합성에 의한 서열화, 올리고뉴클레오티드 프로브 결찰에 의한 서열화 및 이온 반도체 서열화를 포함한다. 최대 수백만개의 DNA 서열 리드를 만들기 위하여, 개별 시료의 DNA는 개별적으로 서열화될 수 있고 (가령, 단일화(singleplex) 서열화) 또는 다중 시료들로부터 DNA가 푸울(pooled)되고, 단일 서열화 운영에서 지수화된(indexed) 게놈 분자들로써 서열화될 수 있다 (가령, 복합 서열화). 본 방법에 따른 서열 정보를 획득하기 위하여 이용될 수 있는 서열화 기술의 예는 하기에서 설명된다.

[0140] DNA 시료들을 이용한 다양한 CNV 분석은 서열화기의 서열 리드를 참조 서열에 대하여 정렬시키거나 또는 매핑하는 것이 수반된다. 참조 서열은 전체 게놈의 서열, 염색체의 서열, 준 염색체 영역의 서열, 등등이 될 수 있다. 참조 서열의 특징들로 인하여, Y 염색체의 CNV의 진단은 상염색체와 비교하여 고조된 기술적 난제가 수반되는데, 그 이유는 Y 염색체의 커버리지는 상염색체의 것보다 더 낮고, 그리고 Y 염색체 상에 반복 서열은 이들의 정확한 위치에 리드을 매핑하는 것을 복합하게 만들기 때문이다. 현재 NGS 기술에 의해 접근가능한 약 10 Mb의 독특한 Y 서열이 있지만, 성별 탐지는 태아 진단학적 부분에서 여전히 도전 과제로 남아있으며, 이때 모체 시료에서 태아 cfDNA의 양은 모체 DNA의 것보다 최소한 한 자릿수가 더 낮고, 비특이적 매핑 문제가 강조된다.

[0141] 추가적으로, 일부 현재 서열화 프로토콜은 초단(ultra-short) 리드, 이를 테면 25mer 리드 및 태그를 이용한다. 서열화 프로토콜의 공정에서 이용되는 초단 서열화는 서열 정렬에서 기술적 난제가 되는 짧은 리드 길이를 만들고, 인간 게놈의 거의 절단이 반복(repeats)으로 덮혀있고, 이들 중 많은 것들은 대략 수십년간 공지되어 왔었다. 컴퓨터를 이용한 관점에서, 반복은 정렬에서 불확실성을 낳고, 이는 다시 전체 염색체 계수화 수준에서 편향 및 오류를 낳을 수 있다.

[0142] 더욱이, 더 짧은 관심 대상 서열, 가령, 메가베이스 크기의 서열의 경우, 신호대 잡음비가 흔히 너무 낮아, CNV의 믿을 만한 탐지를 제시하지 못한다. 본 명세서는 CNV 탐지에서 이러한 난제들을 극복하는 방법들이 제시된다.

[0143] **CNV 평가**

[0144] *CNV 측정을 위한 방법*

[0145] 본 명세서에서 공개된 방법들에 의혜 제시된 서열 커버리지 값을 이용하여, 통상적인 방법들에 의해 획득된 서열 커버리지 값을 이용한 것과 비교하여, 개선된 민감성, 선택성, 및/또는 효과를 가지면서 서열들, 염색체, 또는 염색체 세그먼트들의 복제수 및 CNV에 관련된 다양한 유전적 질환을 측정할 수 있다. 예를 들면, 일부 구체예들에서 태아 및 모체 핵산 분자들이 포함된 모체 테스트 시료에서 임의의 2개 또는 그 이상의 상이한 완전한 태아 염색체 홀배수체의 존재 또는 부재를 판단하는데 마스크된 참조 서열이 이용된다. 참조 서열들 (참조 게놈들이 포함)에 리드를 정렬시키는 예시적인 방법들이 제공된다. 상기 정렬은 마스크안된 또는 마스크된 참조 서열 상에서 실행되고, 그렇게 함으로써 상기 참조 서열에 매핑된 서열 태그가 생성된다. 일부 구체예들에 있어서, 상기 참조 서열의 마스크안된 세그먼트에 속하는 서열 태그만 계수하여 복제 수 변이를 결정한다.

[0146] 일부 구체예들에 있어서, 모체 테스트 시료에서 임의의 완전한 태아 염색체 홀배수체의 존재 또는 부재를 결정하는 방법은 다음 공정을 수반한다: (a) 모체 테스트 시료에서 태아 및 모체 핵산에 대한 서열 정보를 획득하고; (b) 염색체 1-22, X와 Y에서 선택된 각 관심 대상 염색체에 대하여 서열 태그 수 또는 그로부터 도출된 서열 커버리지 분량을 확인하고, 그리고 하나 또는 그 이상의 정규화 염색체 서열들에 대한 서열 태그 수를 확인하기 위하여 상기에서 설명된 서열 정보 및 방법을 이용하고; (c) 각 관심 대상 염색체에 대한 단일 염색체 분량을 산출하기 위하여 각 관심 염색체에서 확인된 서열 태그 수와 각 정규화 염색체에서 확인된 서열 태그 수를 이용하고; 그리고 (d) 각 염색체 분량을 임계치 값과 비교하고, 그렇게 함으로써 모체 테스트 시료에서 임의의 완전한 태아 염색체 홀배수체의 존재 또는 부재가 결정된다.

[0147] 일부 구체예들에 있어서, 상기 설명된 단계 (a)는 상기 테스트 시료의 태아 및 모체 핵산 분자들에 대한 서열 정보를 획득하기 위하여 테스트 시료의 핵산 분자들의 최소한 일부분의 서열화를 수반할 수 있다. 일부 구체예들에 있어서, 단계 (c)는 각 관심 염색체에서 확인된 서열 태그의 수와 상기 정규화 염색체 서열(들)에서 확인된 서열 태그 수의 비율로써 각 관심 염색체에 대한 단일 염색체 분량을 산출하는 것을 포함한다. 일부 다른 구체예들에 있어서, 염색체 분량은 서열 태그의 수로부터 유래된 가공된 서열 커버리지 분량에 근거한다. 일부 구체예들에 있어서, 오직 독특한 비-겹침 서열 태그만을 이용하여 가공된 서열 커버리지 분량을 산출한다. 일부 구체예들에 있어서, 상기 가공된 서열 커버리지 분량은 서열 태그 밀도 비율이며, 이는 서열 길이에 의해 표준화된 서열 태그의 수이다. 일부 구체예들에 있어서, 상기 가공된 서열 커버리지 분량은 정규화된 서열 태그이며, 이는 게놈의 전부 또는 실질적인 부분으로 나눈 관심 대상 서열의 서열 태그의 수이다. 일부 구체예들에 있어서, 상기 가공된 서열 커버리지 분량은 상기 관심 대상 서열의 포괄적 프로파일에 따라 조정된다. 일부 구체예들에 있어서, 상기 가공된 서열 커버리지 분량은 테스트되는 시료의 GC 함량과 서열 커버리지의 상관관계에 따라 조정된다. 일부 구체예들에 있어서, 상기 가공된 서열 커버리지 분량은 본 명세서의 다른 부분에서 추가 설명되는 공정들의 조합으로부터 산출된다.

[0148] 일부 구체예들에 있어서, 염색체 분량은 각 관심 대상 염색체의 가공된 서열 커버리지 분량과 정규화 염색체 서열(들)의 가공된 서열 커버리지 분량의 비율로써 산출된다.

[0149] 상기 구체예들중 임의의 하나에 있어서, 완전한 염색체 홀배수체는 완전한 염색체 삼체성, 완전한 염색체 단체성 및 완전한 염색체 다체성에서 선택된다. 완전한 염색체 홀배수체는 염색체 1-22, X, 그리고 Y중 임의의 하나의 완전한 홀배수체에서 선택된다. 예를 들면, 전술한 상이한 완전한 태아 염색체 홀배수체는 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.

[0150] 상기 구체예들중 임의의 하나에 있어서, 단계 (a)-(d)는 상이한 모체 개체의 테스트 시료들에서 반복되며, 그리

고 상기 방법은 상기의 각 테스트 시료들에서 임의의 2개 또는 그 이상의 상이한 완전한 태아 염색체 홀배수체의 존재 또는 부재를 결정하는 것을 포함한다.

[0151] 상기 구체예들중 임의의 하나에 있어서, 상기 방법은 정규화된 염색체 값 (NCV)을 산출하는 것을 더 포함할 수 있으며, 여기에서 상기 NCV는 다음과 같이 염색체 분량을 검정된 시료들 세트에서 대응하는 염색체 분량의 평균에 연관시킨다:

[0152] $$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

[0153] 이때 $\hat{\mu}_j$ 및 $\hat{\sigma}_j$ 는 검정된 시료들의 세트에서 *j*-번째 염색체 분량에 대하여 각각 차례로 추정된 평균 및 표준편차이며, 그리고 x_{ij} 테스트 시료 *i*에 대하여 관찰된 *j*-번째 염색체 분량이다.

[0154] 일부 구체예들에 있어서, NCV는 테스트 시료에서 관심 대상 염색체의 염색체 분량을 동일한 플로우 쎌(flow cell) 상에 서열화된 복합화된 시료들에서 대응하는 염색체 분량의 중앙값(median)에 연관시킴으로써 다음과 같이 "즉시(on the fly)" 산출될 수 있다:

[0155] $$NCV_{ij} = \frac{x_{ij} - \mathrm{M}_j}{\hat{\sigma}_j}$$

[0156] 이때 $\mathrm{M_j}$ 는 동일한 플로우 쎌 상에서 서열화된 복합화된 시료들의 세트에서 *j*-번째 염색체 분량에 대하여 추정된 중앙값이며; $\hat{\sigma}_\mathrm{j}$ 는 하나 또는 그 이상의 플로우 쎌 상에서 서열화된 복합화된 시료들의 하나 또는 그 이상의 세트 에서 *j*-번째 염색체 분량에 대한 표준 편차이며, 그리고 $\mathrm{x_i}$ 테스트 시료 *i*에 대한 관찰된 *j*-번째 염색체 분량이다. 이 구체예에 있어서, 테스트 시료 i는 $\mathrm{M_j}$ 가 측정되는 동일한 플로우 쎌 상에서 서열화된 복합된 시료들중에 하나이다.

[0157] 일부 구체예들에 있어서, 태아 및 모체 핵산 분자들이 포함된 모체 테스트 시료에서 상이한 부분적 태아 염색체 홀배수체의 존재 또는 부재를 결정하는 방법이 제시된다. 상기 방법은 상기에서 개요가 설명된 바와 같이 완전한 홀배수체를 탐지하기 위한 상기 방법과 유사한 과정들을 수반한다. 그러나, 완전한 염색체를 분석하는 대신, 염색체의 세그먼트가 분석된다. US 특허 출원 공개 번호 2013/0029852를 참조하고, 이는 참고로 본 명세서에 포함된다.

[0158] 도 1은 일부 구체예들에 따른 복제 수 변이의 존재를 결정하는 방법을 보여준다. 작동(130) 및 (135)에서, 검증된 서열 태그 커버리지 및 테스트 서열 태그 커버리지가 결정된다. 본 명세서는 통상적인 방법들과 비교하여 개선된 민감성 및 선택성을 제공하는 커버리지 분량을 결정하는 공정들을 제공한다. 작동(130) 및 (135)는 별표 표시되어 있고, 굵은 선의 네모 박스로 강조되어 있는데, 이는 이들 작동이 선행 기술을 상회하는 향상에 기여함을 나타내기 위한 것이다. 일부 구체예들에 있어서, 상기 서열 태그 커버리지 분량은 분석의 민감성 및 선택성을 개선시키기 위하여 정규화되고, 조정되고, 손질되고, 그리고 그렇지 않으면 가공된다. 이들 공정들은 본 명세서의 도처에서 추가 설명된다.

[0159] 전체적인 관점에서, 상기 방법은 테스트 시료들의 CNV 결정에 있어서 검정된 훈련용 시료들의 정규화 시료를 이용한다. 일부 구체예들에 있어서, 상기 검정된 훈련용 시료들은 영향을 받지 않은, 그리고 정상적 복제 수를 갖는다. 정규화 서열들은 운용내(intra-run) 및 운용간(inter-run) 가변성에 대한 측정들을 정규화시키는 기전을 제공한다. 임의의 하나의 관심 대상 서열, 가령, 염색체 또는 이의 세그먼트에 대하여 정상 복제 수를 갖는 세포를 포함하는 것으로 알려진 개체에서 획득된 검정된 시료 세트의 서열 정보를 이용하여 정규화 서열들이 식별된다. 도 1에서 설명된 방법의 구체예의 단계(110), (120), (130), (145) 및 (146)에서 개요가 설명된다. 일부 구체예들에 있어서, 상기 정규화 서열들을 이용하여 테스트 서열들에 대한 서열 분량이 산출된다. 단계(150) 참고. 일부 구체예들에 있어서, 정규화 서열들을 이용하여 임계치를 산출하고, 이는 상기 테스트 서열들

의 서열 분량과 비교된다. 단계(150) 참고. 상기 정규화 서열 및 테스트 서열로부터 획득된 서열 정보는 테스트 시료들 에서 염색체 홀배수체의 통계학적으로 의미있는 식별을 결정하는데 이용된다(단계 160).

[0160] 일부 구체예들에 따른 복제 수 변이 존재를 결정하기 위한 방법의 세부사항으로 돌아가면, 도 1은 생물학적 시료에서 관심 대상 서열, 가령, 염색체 또는 이의 세그먼트의 CNV를 결정하는 흐름도 구체예(100)를 제공한다. 일부 구체예들에 있어서, 생물학적 시료는 개체로부터 획득되며, 그리고 상이한 게놈들에 의해 기증된 핵산 혼합물을 포함한다. 상기 상이한 게놈들은 2명의 개체에 의해 시료로 기증될 수 있는데, 가령, 상이한 게놈들은 태아와 태아를 출산하는 엄마에 의해 기증된다. 또한, 상기 상이한 게놈들은 3명 또는 그 이상의 개체에 의해 시료로 기증될 수 있는데, 가령, 상이한 게놈들은 2명 또는 그 이상의 태아 및 이들 태아를 출산하는 엄마에 의해 기증된다. 대안으로, 상기 게놈들은 동일한 개체로부터 홀배수체 암에 걸린 세포와 정상적 정배수체(euploid) 세포, 가령, 암 환자의 혈장 시료에 의해 시료로 기증된다.

[0161] 환자의 테스트 시료를 분석하는 것과 별도로, 각 가능한 관심 염색체에 대하여 하나 또는 그 이상의 정규화 염색체 또는 하나 또는 그 이상의 정규화 염색체 세그먼트들이 선택된다. 상기 정규화 염색체 또는 세그먼트들은 환자 시료들의 정상 테스트로부터 비동기적으로 확인되며, 이는 임상 환경에서 일어날 수 있다. 환언하면, 상기 정규화 염색체 또는 세그먼트들은 환자 시료들의 테스트에 앞서 식별된다. 정규화 염색체 또는 세그먼트들과 관심 대상의 염색체 또는 세그먼트들 사이의 연관은 테스트하는 동안 사용을 위하여 보관된다. 하기에서 설명되는 바와 같이, 이러한 연관은 많은 시료들의 테스트 기간에 걸쳐 전형적으로 유지된다. 다음의 논의는 관심대상의 개별 염색체 또는 세그먼트들에 대하여 정규화 염색체 또는 염색체 세그먼트들을 선택하는 구체예들에 관계한다.

[0162] 검정된 정규화 서열들을 확인하고, 그리고 테스트 시료들에서 CNV의 통계학적으로 의미있는 확인을 결정하는데 사용되는 변동 값을 제공하기 위하여 검증된 시료들 세트가 획득된다. 단계(110)에서, 임의의 하나의 관심 대상 서열에 있어서 정상적 복제 수를 갖는 세포들이 포함된 것으로 알려진 다수의 대상으로부터 다수의 생물학적 검정된 시료들이 획득된다. 한 구체예에서, 상기 검정된 시료들은 염색체의 정상적 복제 수를 가지기 위하여 세포유전적 평균을 이용하여 확인된 태아를 임신한 엄마로부터 획득된다. 상기 생물학적 검정된 시료들은 생물학적 유체, 가령, 혈장, 또는 하기에서 설명된 바와 같은 임의의 적절한 시료일 수 있다. 일부 구체예들에 있어서, 검정된 시료는 핵산 분자들, 가령, cfDNA 분자들의 혼합물을 포함한다. 일부 구체예들에 있어서, 상기 검정된 시료는 태아 및 모체 cfDNA 분자들의 혼합물이 포함된 모체 혈장 시료다. 정규화 염색체 및/또는 이의 세그먼트들에 대한 서열 정보는 임의의 공지의 서열화 방법을 이용하여 상기 핵산, 가령, 태아 및 모체 핵산의 최소한 일부분을 서열화함으로써 획득된다. 바람직하게는 본 명세서의 도처에서 설명된 차세대 서열화 (NGS) 방법들중 하나는 단일 또는 클론적으로 증폭된 분자들로써 태아 및 모체 핵산을 서열화하는데 이용된다. 다양한 구체예들에 있어서, 상기 검정된 시료들은 서열화 전 또는 서열화 동안 하기에서 공개된 바와 같이 가공된다. 이들은 본 명세서에서 공개된 바와 같은 장치, 시스템 및 키트를 이용하여 가공될 수 있다.

[0163] 단계(120)에서, 상기 검정된 시료들 에서 포함된 모든 상기 검정된 핵산의 최소한 일부가 서열화되어 수백만개의 서열 리드, 가령, 36bp 리드가 생성되고, 이들은 참조 게놈, 가령, hg18에 대하여 정렬된다. 일부 구체예들에 있어서, 상기 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, 약90bp, 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. 기술적 진보는 500bp 이상의 단일-말단 리드가 쌍을 이룬 말단 리드가 생성될 때, 약 1000bp 이상의 리드를 가능하게 할 것으로 예상된다. 한 구체예에서, 상기 매핑된 서열 리드는 36bp을 포함한다. 또다른 구체예에서, 상기 매핑된 서열 리드는 25bp을 포함한다.

[0164] 서열 리드는 참조 게놈에 정렬되고, 그리고 상기 참조 게놈에 독특하게 매핑된 리드는 서열 태그로 공지된다. 마스크된 참조 서열의 마스크된 세그먼트들 상에 속하는 서열 태그는 CNV 분석에 계수되지 않는다.

[0165] 한 구체예에서, 20 내지 40bp 리드가 포함된 최소한 약 3 x 10^6 검증된 서열 태그, 최소한 약 5 x 10^6 검증된 서열 태그, 최소한 약 8 x 10^6 검증된 서열 태그, 최소한 약 10 x 10^6 검증된 서열 태그, 최소한 약 15 x 10^6 검증된 서열 태그, 최소한 약 20 x 10^6 검증된 서열 태그, 최소한 약 30 x 10^6 검증된 서열 태그, 최소한 약 40 x 10^6 검증된 서열 태그, 또는 최소한 약 50 x 10^6 검증된 서열 태그가 참조 게놈에 독특하게 매핑된 리드로부터 획득된다.

[0166] 단계(130)에서, 상기 검증된 시료들에서 핵산의 서열화로부터 획득된 모든 태그를 계수하여 검정된 서열 태그 커버리지를 획득한다. 유사하게, 작동(135)에서 테스트 시료로부터 획득된 모든 태그를 계수하여 테스트 서열 태그 커버리지를 획득한다. 본 명세서는 통상적인 방법들과 비교하여 개선된 민감성 및 선택성을 제공하는 커버리지 분량을 결정하는 공정들을 제공한다. 작동(130) 및 (135)는 별표 표시되어 있고, 굵은 선의 네모 박스로 강조되어 있는데, 이는 이들 작동이 선행 기술을 상회하는 개선에 기여함을 나타낸다. 일부 구체예들에 있어서, 상기 서열 태그 커버리지 분량은 분석의 민감성 및 선택성을 개선시키기 위하여 정규화되고, 조정되고, 손질되고, 그리고 그렇지 않으면 가공된다. 이들 공정들은 본 명세서의 도처에서 추가 설명된다.

[0167] 모든 검정된 서열 태그가 각각의 검정된 시료들에 매핑되고 계수되기 때문에, 후속적으로 정규화 서열들이 확인되는 추가 서열들에 대한 서열 태그 커버리지와 같이 상기 검정된 시료들에서 관심 대상 서열, 가령, 임상적으로-관련된 서열에 대한 서열 태그 커버리지가 결정된다.

[0168] 일부 구체예들에 있어서, 상기 관심 대상 서열은 완전한 염색체 홀배수체, 가령, 염색체 21과 연관된 염색체이며, 그리고 상기 검정된 정규화 서열은 염색체 홀배수체와 연관되지 않고, 서열 태그 커버리지에서 이의 변이는 관심 대상 가령, (가령, 염색체) 염색체 21의 서열의 것에 근사한 완전한 염색체다. 선택된 정규화 염색체(들)은 상기 관심 대상 서열의 서열 태그 커버리지에서 변이에 가장 근사한 하나 또는 집단일 수 있다. 염색체 1-22, X, 그리고 Y중 임의의 하나 또는 그 이상은 관심 대상 서열일 수 있고, 그리고 하나 또는 그 이상의 염색체는 상기 검정된 시료들에서 임의의 하나의 염색체 1-22, X와 Y 각각에 대한 정규화된 서열로써 확인될 수 있다. 상기 정규화 염색체는 개별 염색체일 수 있거나 또는 본 명세서의 도처에서 설명된 염색체 집단일 수 있다.

[0169] 또다른 구체예에서, 상기 관심 대상 서열은 부분적 홀배수체, 가령, 염색체 결손 또는 삽입, 또는 불균형 염색체 전위와 연관된 염색체의 세그먼트이며, 그리고 상기 정규화 서열은 부분적 홀배수체와 연관되지 않고, 서열 태그 커버리지에서 이의 변이가 부분적 홀배수체와 연관된 염색체 세그먼트의 것에 근사한 염색체 세그먼트 (또는 세그먼트들의 집단)이다. 선택된 정규화 염색체 세그먼트(들)은 상기 관심 대상 서열의 서열 태그 커버리지에서 변이에 가장 근사한 하나 또는 그 이상일 수 있다. 염색체 1-22, X, 그리고 Y의 임의의 하나 또는 그 이상의 세그먼트들은 관심 대상 서열일 수 있다.

[0170] 다른 구체예들에서, 상기 관심 대상 서열은 부분적 홀배수체와 연관된 염색체의 세그먼트이며, 상기 정규화 서열은 전체 염색체 또는 염색체이다. 또 다른 구체예들에 있어서, 상기 관심 대상 서열은 홀배수체와 연관된 전체 염색체이며, 상기 정규화 서열은 홀배수체와 연관되지 않은 염색체 세그먼트 또는 세그먼트들이다.

[0171] 검정된 시료들에서, 단일 서열 또는 서열들의 집단이 임의의 하나 또는 그 이상의 관심 대상 서열들에 대하여 정규화 서열(들)로서 확인되는지의 여부에 관계없이, 상기 검정된 정규화 서열은, 상기 검정된 시료들에서 결정될 때의 관심 대상 서열의 것에 가장 또는 효과적으로 근사한 서열 태그 커버리지에서의 변이를 가지도록 선택될 수 있다. 예를 들면, 검정된 정규화 서열은 관심 대상 서열을 정규화하는데 이용될 경우 상기 검정된 시료들에 걸쳐 최소한의 변동성을 만드는 서열이며, 가령, 상기 정규화 서열의 변동성은 검증된 시료들에서 결정된 관심 대상 서열의 것에 가장 근접한다. 환언하면, 상기 검정된 정규화 서열은 상기 검정된 시료들에 걸쳐 서열 분량 (상기 관심 대상 서열의)에서 최소 변이를 만들도록 선택된 서열이다. 따라서, 상기 공정은 정규화 염색체로 이용될 때, 상기 관심 대상 서열에 대하여 운용간 염색체 분량에 있어서 최소 변동성을 낳는 것으로 예상되는 서열을 선택한다.

[0172] 서열화 라이브러리를 만들고, 그리고 이 시료들을 서열화하기 위하여 필요한 과정들이 시간 경과에 따라 기본적으로 변경되지 않는다면, 임의의 하나 또는 그 이상의 관심 대상 서열들에 대하여 검정된 시료들에서 확인된 정규화 서열은 수일간, 수주간, 수개월간, 그리고 어쩌면 수년간에 걸쳐 테스트 시료들에서 홀배수체의 존재 또는 부재를 결정하기 위하여 선택되는 정규화 서열로 남아있다. 상기에서 설명된 바와 같이, 시료들, 가령, 상이한 시료들 중에서 이에 매핑되는 서열 태그의 수에서 변동성과, 서열화 운용, 가령, 동일한 날짜 및/또는 상이한 날짜들에서 일어나는 서열화 운용을 위하여(물론 다른 이유들 중에서 아마도)에 대하여 정규화 매개변수로써 이용되는 관심 대상 서열의 변동성에 가장 근접하는 홀배수체의 존재를 결정하기 위한 정규화 서열들이 선택된다. 이들 과정에서 실질적인 변경은 모든 서열들에 매핑된 태그 수에 영향을 줄 것이며, 이는 다시 동일한 날짜 또는 상이한 날짜에 동일한 및/또는 상이한 서열화 운용에 있어서 시료들에 걸쳐 관심 대상 서열(들)의 것에 가장 근접한 변동성을 갖는 서열 또는 서열 집단이 결정될 것이고, 정규화 서열들의 세트는 재결정이 요구될 것이다. 과정들에서 실질적인 변경은 상기 서열화 라이브러리를 준비하는데 이용되는 실험실 프로토콜의 변화를 포함하며, 이 변경에는 단일화된 서열화 대신 복합 서열화를 위하여 시료들을 준비하는데 관련된 변화, 그리고 서열화

에 이용되는 화학물질에서의 변화가 포함된, 서열화 플렛포옴에서의 변화가 포함된다.

[0173] 일부 구체예들에 있어서, 특정 관심 대상 서열을 정규화하기 위하여 선택된 정규화 서열은 하나 또는 그 이상의 영향을 받은 시료들로부터 하나 또는 그 이상의 검정된 시료들을 가장 잘 구별해내는 서열이며, 이는 상기 정규화 서열이 최대 구별가능성(differentiablity)을 갖는다는 것을 의미하는 것으로, 가령 정규화 서열의 구별가능성은 영향을 받은 테스트 시료에서 관심 대상 서열에 대하여 최적의 구별을 제공하여 영향을 받은 테스트 시료를 다른 영향을 받지 않은 시료들로부터 용이하게 구별해내는 것이다. 다른 구체예들에서, 상기 정규화 서열은 최소의 변동성과 최대의 구별가능성의 조합을 갖는 서열이다.

[0174] 구별가능성의 수준은 하기에서 설명되고, 실시예들에서 보여진 바와 같이, 검정된 시료들 집단에서 서열 분량, 가령, 염색체 분량 또는 세그먼트 분량과 하나 또는 그 이상의 테스트 시료들에서 염색체 분량(들) 간의 통계학적 차이로 결정될 수 있다. 예를 들면, 구별가능성은 t-테스트 값으로 수치적으로 나타낼 수 있는데, 이는 검정된 시료들 집단에서 염색체 분량과 하나 또는 그 이상의 테스트 시료들에서 염색체 분량(들)들 간의 통계학적 차이를 나타낸다. 유사하게, 구별가능성은 염색체 분량 대신 세그먼트 분량에 근거할 수 있다. 대안으로, 구별가능성은 정규화된 염색체 값 (NCV)으로 수치적으로 나타낼 수 있는데, 이는 NCV에 대한 분포가 정상적이기만 한다면 염색체 분량의 z-점수이다. 유사하게, 염색체 세그먼트들이 관심 대상 서열들인 경우에서, 세그먼트 분량의 구별가능성은 정규화된 세그먼트 값 (NSV)으로 수치적으로 나타낼 수 있고, 이는 NSV에 대한 분포가 정상적이기만 한다면 염색체 분량의 z-점수이다. z-점수를 결정하는데 있어서, 검증된 시료들 세트에서 염색체 또는 세그먼트 분량의 평균 및 표준 편차가 이용될 수 있다. 대안으로, 검정된 시료들과 영향을 받은 시료들이 포함된 훈련용 세트 에서 염색체 또는 세그먼트 분량의 평균 및 표준 편차가 이용될 수 있다. 다른 구체예들에서, 상기 정규화 서열은 최소의 변동성과 최대의 구별가능성 또는 작은 변동성과 큰 구별가능성의 최적의 조합을 갖는 서열이다.

[0175] 상기 방법은 유전적으로 유사한 특징들을 갖고, 시료들과 서열화 운용에서 유사한 변이를 하는 경향이 있는 서열들을 식별해내고, 이는 테스트 시료들에서 서열 분량을 결정하는데 유용하다.

[0176] *서열 분량의 결정*

[0177] 일부 구체예들에 있어서, 하나 또는 그 이상의 관심대상의 염색체 또는 세그먼트들의 염색체 또는 세그먼트 분량은 도 1에 나타낸 단계(146)에서 설명된 바와 같이 모든 검정된 시료들에서 결정되고, 그리고 정규화 염색체 또는 세그먼트 서열은 단계(145)에서 식별된다. 서열 분량이 산출되기 전, 일부 정규화 서열들이 제시된다. 그 다음, 하기에서 추가로 설명되는 바와 같이 다양한 기준에 따라 하나 또는 그 이상의 정규화 서열들이 식별된다, 단계(145) 참고. 일부 구체예들에 있어서, 가령, 식별된 정규화 서열은 모든 검증된 시료들에 걸쳐 관심 대상 서열에 대한 서열 분량에서 최소 변동성을 낳는다.

[0178] 단계(146)에서 산출된 검증된 태그 밀도에 근거하여, 검증된 서열 분량, 가령, 관심 대상 서열에 대한 염색체 분량 또는 세그먼트 분량은 상기 관심 대상 서열에 대한 서열 태그 커버리지와 추가 서열에 대한 검증된 서열 태그 커버리지의 비율로 결정되며, 이로부터 단계(145)에서 후속적으로 이의 정규화 서열들이 확인된다. 확인된 정규화 서열들이 후속적으로 이용되어 테스트 시료들에서서 서열 분량이 결정된다.

[0179] 한 구체예에서, 검증된 시료들에서 서열 분량은 관심 염색체에 대한 서열 태그 수와 검증된 시료 에서 정규화 염색체 서열에 대한 서열 태그의 수의 비율로 산출되는 염색체 분량이다. 상기 정규화 염색체 서열은 단일 염색체, 염색체의 집단, 한 염색체의 세그먼트, 또는 상이한 염색체의 세그먼트들의 집단일 수 있다. 따라서, 관심 염색체에 대한 염색체 분량은 검증된 시료에서 관심 염색체에 대한 태그의 수와 다음의 것들에 대한 태그 수에 대한 비율로 결정된다: (i) 단일 염색체로 구성된 정규화 염색체 서열, (ii) 2개 또는 그 이상의 염색체로 구성된 정규화 염색체 서열, (iii) 단일 염색체의 세그먼트로 구성된 정규화 세그먼트 서열, (iv) 하나의 염색체로부터 2개 또는 그 이상의 세그먼트들로 구성된 정규화 세그먼트 서열, 또는 (v) 2개 또는 그 이상의 염색체의 2개 또는 그 이상의 세그먼트들로 구성된 정규화 세그먼트 서열. (i)-(v)에 따라 관심 염색체 21의 염색체 분량을 결정하는 예는 다음과 같다: 관심 염색체, 가령, 염색체 21의 염색체 분량은 상기 염색체 21의 서열 태그 커버리지와 다음 서열 태그 커버리지중 하나의 비율로 결정된다: (i) 남아있는 모든 각 염색체, 가령, 염색체 1-20, 염색체 22, 염색체 X, 그리고 염색체 Y 각각; (ii) 2개 또는 그 이상의 남아있는 염색체의 모든 가능한 조합; (iii) 또다른 염색체, 가령, 염색체 9의 세그먼트; (iv) 하나의 다른 염색체의 2개의 세그먼트들, 가령, 염색체 9의 2개 세그먼트들 9; (v) 2개의 상이한 염색체의 2개의 세그먼트들, 가령, 염색체 9의 세그먼트와 염색체 14의 세그먼트.

[0180] 또다른 구체예에서, 상기 검증된 시료들에서 서열 분량은 염색체 분량에 반대되는 세그먼트 분량이며, 이 세그먼트 분량은 전체 염색체가 아닌 관심대상의 세그먼트에 대한 서열 태그의 수 그리고 검증된 시료에서 정규화 세그먼트 서열에 대한 서열 태그의 수의 비율로 산출된다. 상기 정규화 세그먼트 서열은 상기에서 논의된 상기 정규화 염색체 또는 세그먼트 서열들중 임의의 것일 수 있다.

[0181] *정규화 서열들의 식별*

[0182] 단계(145)에서, 정규화 서열은 관심 대상 서열에 대하여 확인된다. 일부 구체예들에 있어서, 가령, 상기 정규화 서열은 산출된 서열 분량, 가령, 검증된 모든 훈련용 시료들에 걸쳐 관심 대상 서열에 대한 서열 분량에서 최소한의 변동성을 초래하는 서열 분량에 기초한 서열이다. 상기 방법은 유전적으로 유사한 특징들을 갖고, 시료들과 서열화 운용에서 유사한 변이를 하는 경향이 있는 서열들을 식별해내고, 이는 테스트 시료들에서 서열 분량을 결정하는데 유용하다.

[0183] 하나 또는 그 이상의 관심 대상 서열들에 대한 정규화 서열들은 검증된 시료들 세트에서 확인될 수 있으며, 그리고 상기 검증된 시료들에서 확인된 서열들을 후속적으로 이용하여 상기의 각 테스트 시료들에서 홀배수체의 존재 또는 부재를 결정하기 위하여 상기의 각 테스트 시료들에서 하나 또는 그 이상의 관심 대상 서열들의 서열 분량이 산출된다(단계 150). 상이한 서열화 플랫포옴이 이용될 때 및/또는 서열화되는 핵산의 정제 및/또는 상기 서열화 라이브러리의 준비에 차이가 존재할 때 관심대상의 염색체 또는 세그먼트들에 대하여 확인된 정규화 서열은 상이할 수 있다. 본 명세서에서 설명된 방법들에 따라 정규화 서열들의 사용은 시료 조제물 및/또는 이용되는 서열화 플랫포옴과 무관하게, 염색체 또는 이의 세그먼트의 복제 수에서 특이적 그리고 민감한 변이 측정을 제공한다.

[0184] 일부 구체예들에 있어서, 하나 이상의 정규화 서열이 확인되고, 가령, 하나의 관심 대상 서열에 대하여 상이한 정규화 서열들이 결정될 수 있고, 그리고 하나의 관심 대상 서열에 대하여 다중 서열 분량이 결정될 수 있다. 예를 들면, 염색체 14의 서열 태그 커버리지가 이용될 때, 관심 염색체 21에 대한 염색체 분량에서 변이, 가령, 변이계수 (CV= 표준 편차/평균)는 최소이다. 그러나, 테스트 시료에서 관심 대상 서열에 대한 서열 분량 결정에 사용을 위한 2, 3, 4, 5, 6, 7, 8 또는 그 이상의 정규화 서열들이 확인될 수 있다. 예를 들면, 염색체 7, 염색체 9, 염색체 11 또는 염색체 12를 정규화 염색체 서열로 이용하여 임의의 하나의 테스트 시료에서 염색체 21에 대한 제 2 분량이 결정될 수 있는데, 이들 염색체은 모두 염색체 14의 것에 가까운 CV를 가지기 때문이다.

[0185] 일부 구체예들에 있어서, 관심 염색체에 대한 정규화 염색체 서열로 단일 염색체가 선택될 때, 상기 정규화 염색체 서열은 테스트된 모든 시료들, 가령, 검증된 시료들에 걸쳐 최소의 변동성을 갖는 관심 염색체에 대한 염색체 분량을 야기하는 염색체일 것이다. 일부 경우들에 있어서, 최고의 정규화 염색체는 최소 변이를 보유하지 않을 수 있지만, 검증된 시료들로부터 테스트 시료 또는 시료들을 가장 잘 구별해내는 검증된 분량의 분포를 가질 수 있고, 가령, 최고의 정규화 염색체는 최저 변이를 보유하지 않을 수 있지만, 최대의 구별가능성을 보유할 수는 있다.

[0186] *테스트 시료들에서 홀배수체의 결정*

[0187] 검증된 시료들에서 정규화 서열(들)의 식별에 기초하여, 하나 또는 그 이상의 관심 대상 서열들이 상이한 게놈들로부터 유도된 핵산 혼합물이 포함된 테스트 시료에서 관심 대상 서열에 대하여 서열 분량이 측정된다.

[0188] 단계(115)에서 관심 대상 서열의 임상적으로-관련된 CNV를 가지고 있는 것으로 의심되는 또는 가지고 있는 것으로 알려진 개체로부터 테스트 시료가 획득된다. 상기 테스트 시료는 생물학적 유체, 가령, 혈장, 또는 하기에서 설명된 바와 같은 임의의 적절한 시료일 수 있다. 설명된 바와 같이, 이 시료는 비-침습성 과정, 이를 테면 단순 채혈을 이용하여 획득될 수 있다. 일부 구체예들에 있어서, 테스트 시료는 핵산 분자들, 가령, cfDNA 분자들의 혼합물을 포함한다. 일부 구체예들에 있어서, 상기 테스트 태아 및 모체 cfDNA 분자들의 혼합물이 포함된 모체 혈장 시료이다.

[0189] 단계(125)에서 테스트 시료에서 테스트 핵산의 최소한의 일부가 서열화되어, 상기 검증된 시료들에서 설명된 바와 같이, 수백만개의 서열 리드, 가령, 36bp 리드가 생성된다. 단계(120)에서와 같이, 테스트 시료에서 핵산의 서열화로부터 생성된 리드는 참조 게놈에 독특하게 매핑되거나 또는 정렬되어 태그가 생성된다. 단계(120)에서 설명된 바와 같이, 20 내지 40bp 리드가 포함된 최소한 약 3×10^6 검증된 서열 태그, 최소한 약 5×10^6 검증된 서열 태그, 최소한 약 8×10^6 검증된 서열 태그, 최소한 약 10×10^6 검증된 서열 태그, 최소한 약 15×10^6 검증된 서열 태그, 최소한 약 20×10^6 검증된 서열 태그, 최소한 약 30×10^6 검증된 서열 태그, 최소한 약 40

x 10^6 검증된 서열 태그, 또는 최소한 약 50 x 10^6 검증된 서열 태그가 참조 게놈에 독특하게 매핑된 리드로부터 획득된다. 특정 구체예들에 있어서, 서열화 장치에 의해 생성된 리드는 전자 포멧으로 제시된다. 하기에서 논의되는 바와 같이 컴퓨터 장치를 이용하여 정렬이 이루어진다. 개별 리드는 상기 참조 게놈과 비교되는데, 상기 리드가 상기 참조 게놈과 독특하게 대응하는 부위들을 식별해내기 위하여 이 작업은 간혹 방대하다(수백만개의 염기쌍). 일부 구체예들에 있어서, 이 정렬 과정으로 리드와 상기 참조 게놈 사이의 제한적인 불합치가 허용된다. 일부 경우들에 있어서, 리드에서 1, 2, 또는 3개의 염기쌍이 참조 게놈에서 대응하는 염기쌍에 대하여 불합치가 허용되지만, 여전히 매핑이 된다.

[0190] 단계(135)에서 상기 테스트 시료들에서 핵산의 서열화로부터 획득된 모든 또는 대부분의 태그가 계수되어 하기에서 설명되는 컴퓨터 장치를 이용한 테스트 서열 태그 커버리지가 결정된다. 일부 구체예들에 있어서, 각 리드는 상기 참조 게놈의 특정 영역 (대부분의 경우에 있어서 염색체 또는 세그먼트)에 정렬되고, 그리고 상기 리드에 부위 정보를 부착시킴으로써 이 리드가 태그로 전환된다. 이 공정이 펼쳐질 때, 상기 컴퓨터 장치는 상기 참조 게놈의 각 영역 (대부분의 경우에 있어서 염색체 또는 세그먼트)에 매핑되는 태그/리드 수를 계수하는 작업을 지속할 수 있다. 관심대상의 각 염색체 또는 세그먼트와 각 대응하는 정규화 염색체 또는 세그먼트에 대한 계수가 저장된다.

[0191] 특정 구체예들에 있어서, 상기 참조 게놈은 진짜 생물학적 게놈의 일부이지만, 상기 참조 게놈에서 포함되지 않은 하나 또는 그 이상의 배제된 영역들을 보유한다. 이들 배제된 영역들에 대하여 잠재적으로 정렬되는 리드는 계수되지 않는다. 배제된 영역들의 예로는 긴 반복된 서열들의 영역, X와 Y 염색체 사이에 유사한 영역들, 등등이 포함된다. 상기에서 설명된 마스킹 기술에 의해 획득된 마스크된 참조 서열을 이용하여, CNV 분석에는 상기 참조 서열 상에 마스크안된 세그먼트들에 있는 태그만 계수된다.

[0192] 일부 구체예들에 있어서, 상기 방법은 다중 리드가 참조 게놈 또는 서열 상의 동일 부위에 정렬될 때 태그를 한번 이상 계수해야할 지를 결정한다. 2개 태그가 동일한 서열을 갖고, 따라서, 참조 서열 상에 동일 부위에 정렬되는 경우들이 있다. 계수 태그를 계수하는데 이용되는 방법은 특정 상황하에 동일한 서열화된 동일한 시료로부터 유도된 동일한 태그 계수를 배제할 수 있다. 주어진 시료 에서 태그의 균형에 맞지 않는 수가 동일하다면, 이는 이 과정에서 강력한 편향 또는 다른 결함이 있다는 것을 암시한다. 따라서, 특정 구체예들에 따라 상기 계수 방법은 주어진 시료에서 이미 계수된 시료의 태그와 동일한 주어진 태그를 계수하지 않는다.

[0193] 단일 시료에서 동일한 태그를 무시할 때를 선택하기 위하여 다양한 기준이 설정될 수 있다. 특정 구체예들에 있어서, 계수된 태그의 확정된 백분율은 특유적이어야 한다. 이 임계치 이상의 태그가 특유하지 않다면, 이들은 무시된다. 예를 들면, 확정된 백분율이 최소한 50% 독특함을 요구하는 경우, 이 시료에 대하여 특유한 태그의 백분율이 50%를 초과할 때까지 동일한 태그는 계수되지 않는다. 다른 구체예들에서, 특유한 태그의 임계치수는 최소한 약 60%이다. 다른 구체예들에서, 특유한 태그의 임계치 백분율은 최소한 약 75%, 또는 최소한 약 90%, 또는 최소한 약 95%, 또는 최소한 약 98%, 또는 최소한 약 99%이다. 염색체 21에 대한 임계시는 90%에서 설정될 수 있다. 30M 태그가 염색체 21에 대하여 정렬될 경우, 그 다음 이중 최소한 27M은 특유해야만 한다. 3M 계수된 태그가 특유하지 않고, 30 백만 및 제1 태그가 특유하지 않다면, 이는 계수되지 않는다. 동일한 태그를 더 카운트하지 않을 때를 결정하는데 이용되는 특정 임계치 또는 다른 기준의 선택은 적절한 통계학적 분석을 이용하여 선택될 수 있다. 이 임계치 또는 다른 기준에 영향을 주는 하나의 인자는 태그가 정렬되는 게놈의 크기에 대한 서열화되는 시료의 상태적인 양이다. 다른 인자들은 리드의 크기 및 유사한 고려들을 포함한다.

[0194] 한 구체예에서, 관심 대상 서열에 매핑된 테스트 서열 태그의 수는 테스트 서열 태그 밀도 비율을 제공하기 위하여 관심 대상 서열의 공지의 길이로 정규화된다. 상기 검증된 시료들에 대하여 설명된 바와 같이, 관심 대상 서열의 공지의 길이로의 정규화는 요구되지 않으며, 그리고 인간 해독을 위하여 이를 단순화하기 위하여 수에서 숫자 수를 줄이기 위한 단계로써 포함될 수 있다. 상기 테스트 시료 에서 모든 매핑된 테스트 서열 태그가 계수되기 때문에, 상기 검증된 시료들에서 확인된 최소한 하나의 정규화 서열에 상응하는 추가 서열에 대한 서열 태그 커버리지와 같이, 관심 대상 서열, 가령, 테스트 시료 에서 임상적으로-관련된 서열에 대한 서열 태그 커버리지가 결정된다.

[0195] 단계(150)에서, 상기 검증된 시료들에서 최소한 하나의 정규화 서열에 근거하여, 상기 테스트 시료에서 관심 대상 서열에 대한 테스트 서열 분량이 결정된다. 다양한 구체예들에 있어서, 상기 테스트 서열 분량은 본 명세서에서 설명된 바와 같이 상기 관심 대상 서열과 대응하는 정규화 서열에 대한 서열 태그 커버리지를 이용하여 컴퓨터에 의해 결정된다. 이 작업을 하는 컴퓨터 장치는 상기 관심 대상 서열과 이의 연관된 정규화 서열 사이

의 연관성에 전자적으로 접근할 수 있고, 이는 데이터베이스, 표, 그래프 에서 저장되거나 또는 프로그램 지시에 코드로 포함될 수 있다.

[0196] 본 명세서의 도처에서 설명된, 최소한 하나의 정규화 서열은 단일 서열 또는 서열들의 집단일 수 있다. 테스트 시료에서 관심 대상 서열에 대한 서열 분량은 테스트 시료에서 관심 대상 서열에 대해 결정된 서열 태그 커버리지와 상기 테스트 시료에서 결정된 최소한 하나의 정규화 서열의 서열 태그 커버리지의 비율이며, 여기에서 상기 테스트 시료에서 정규화 서열은 특정 관심 대상 서열에 대하여 검정된 시료들에서 확인된 정규화 서열에 대응한다. 예를 들면, 상기 검증된 시료들에서 염색체 21에 대하여 확인된 정규화 서열은 염색체, 가령, 염색체 14로 결정된다면, 염색체 21 (관심 대상 서열)에 대한 테스트 서열 분량은 테스트 서열에서 결정된 염색체 21의 서열 태그 커버리지와 테스트 시료에서 결정된 염색체 14의 서열 태그 커버리지의 비율로 결정된다. 유사하게, 염색체 홀배수체와 연관된 염색체 13, 18, X, Y, 그리고 다른 염색체에 대한 염색체 분량이 결정된다. 관심 염색체에 대한 정규화 서열은 하나의 염색체 또는 염색체 집단이거나, 또는 염색체 세그먼트들중 하나 또는 집단일 수 있다. 이미 설명한 바와 같이, 관심 대상 서열은 염색체의 일부, 가령, 염색체 세그먼트일 수 있다. 따라서, 염색체 세그먼트에 대한 분량은 테스트 시료에서 세그먼트에 대하여 결정된 서열 태그 커버리지와 테스트 시료에서 상기 정규화 염색체 세그먼트에 대한 서열 태그 커버리지의 비율로 결정될 수 있으며, 여기에서 상기 테스트 시료 에서 정규화 세그먼트는 관심 대상의 특정 세그먼트에 대하여 검정된 시료들에서 확인된 상기 정규화 세그먼트 (단일 세그먼트 또는 세그먼트들의 집단)에 대응한다. 염색체 세그먼트들의 크기는 킬로베이스 (kb) 내지 메가베이스 (Mb) (가령, 약 1kb 내지 10 kb, 또는 약 10 kb 내지 100 kb, 또는 약 100kb 내지 1 Mb)이다.

[0197] 단계(155)에서, 임계치 값은 다수의 검증된 시료들에서 결정된 검증된 서열 분량 및 관심 대상 서열에 대하여 홀배수체인 것으로 공지된 시료들에 대한 서열 분량으로부터 확립된 표준 편차 값으로부터 유도된다. 이 작동은 전형적으로 환자 테스트 시료들의 분석과 비동기적으로 실행된다. 예를 들면, 검증된 시료들로부터 정규화 서열의 선택과 동시에 실행될 수 있다. 정확한 분류는 상이한 부류, 가령, 홀배수체의 유형에 대한 확률분포 사이의 차이에 따라 달라진다. 일부 실시예들에 있어서, 각 유형의 홀배수체, 가령, 삼염색체성 21에 대한 실증적인 분포로부터 임계치들이 선택된다. cfDNA의 서열화에 의해 염색체 홀배수체를 결정하는 방법을 설명하는 실시예들에서 설명된 바와 같이 삼염색체성 13, 삼염색체성 18, 삼염색체성 21, 그리고 일염색체성 X 홀배수체를 분류하기 위하여 확립된 가능 임계치는 태아 핵산 및 모체 핵산 혼합물이 포함된 모체 시료로부터 추출되었다. 염색체의 홀배수체에 대하여 영향을 받은 시료들을 구별하기 위하여 결정된 임계치는 상이한 홀배수체에 대한 임계치와 동일하거나 또는 상이할 수 있다. 실시예들에서 나타낸 것과 같이, 각 관심 염색체에 대한 임계치는 시료들에 걸쳐 그리고 서열화 운용들에 걸쳐 관심 염색체의 분량에 있어서 변동성으로부터 결정된다. 임의의 관심 염색체에 대한 염색체 분량의 변수가 적을 수록, 모든 영향을 받지 않은 시료들에 걸쳐 관심 염색체에 대한 분량에서 스프레드는 더 좁아지며, 이를 이용하여 상이한 홀배수체를 결정하는 임계치가 설정된다.

[0198] 단계(160)에서 환자 테스트 시료의 분류와 연관된 공정 플로우로 돌아가서, 상기 관심 대상 서열에 대한 테스트 서열을 상기 검증된 서열 분량에서 확립된 최소한 하나의 임계치 값에 비교함으로써, 테스트 시료에서 관심 대상 서열의 복제 수 변이가 결정된다. 이 작동은 서열 태그 커버리지를 측정하고 및/또는 세그먼트 분량을 산출하는데 이용된 동일한 컴퓨터 장치에 의해 실행될 수 있다.

[0199] 단계(160)에서, 테스트 관심 대상 서열에 대한 산출된 분량은 이 시료를 "정상", "영향을 받은(발병)", 또는 "무소명(no call)"로 분류하기 위하여 사용자-정의된 "신뢰 임계치"에 따라 선택된 임계치 값으로써 이 세트에 비교된다. "무소명(no call)" 시료들은 신뢰성을 가지고 확정적 진단이 만들어질 수 없는 시료들이다. 영향을 받은 시료의 각 유형 (가령, 삼염색체성 21, 부분적 삼염색체성 21, 일염색체성 X)는 이의 고유한 임계치, 정상적 (영향을 받지 않은) 시료로 소명되는 하나 그리고 영향을 받은 시료들로 소명되는 또 다른 하나 (일부 경우들에 있어서 2개의 임계치는 일치한다)를 가진다. 본 명세서의 도처에서 설명된, 일부 환경 하에 상기 테스트 시료에서 핵산의 태아 분획이 충분히 높다면, 무-소명은 소명 (영향을 받은 또는 정상적)으로 전환될 수 있다. 상기 테스트 서열의 분류는 이 공정 플로우의 다른 작동에 이용된 컴퓨터 장치에 의해 보고될 수 있다. 일부 경우들에 있어서, 이 분류는 전자 포멧으로 보고되며, 관심 대상에게 보여지거나, 이메일로 보내지거나, 문자로 보내어질 수 있다.

[0200] 일부 구체예들에 있어서, CNV의 결정은 염색체 또는 세그먼트 분량이 상기에서 설명된 검증된 시료들 세트 에서 대응하는 염색체 또는 세그먼트 분량에 관련된 NCV 또는 NSV의 산출을 포함한다. 그 다음 CNV는 예정된 복제 수 평가 임계치 값에 상기 NCV/NSV를 비교함으로써 결정될 수 있다.

[0201] 가양성과 가음성 비율을 최적화시키기 위하여 복제 수 평가 임계치가 선택될 수 있다. 복제 수 평가 임계치가 높을 수록, 가양성 발생은 더 적을 것이다. 유사하게, 임계치가 낮을 수록, 가음성 발생 가능성이 더 적을 것 같다. 따라서, 오직 참양성으로 분류되는 것 이상의 제1 이상적 임계치와 오직 참음성으로 분류되는 이하의 제2 이상적 임계치 사이에 균형(trade-off)이 존재한다.

[0202] 임계치는 영향을 받지 않은 시료들 세트에서 결정된 바와 같이, 특정 관심 염색체에 대한 염색체 분량에서 변동성에 따라 대개 달라진다. 상기 변동성은 시료 에서 존재하는 태아 cDNA의 분획이 포함된 다수의 인자들에 의존적이다. 상기 변동성 (CV)은 영향을 받지 않은 시료들의 집단에 걸쳐 염색체 분량에 대한 평균 또는 중앙값 그리고 표준 편차에 의해 결정된다. 따라서, 홀배수체를 분류하기 위한 임계치 (들)은 다음에 따라 NCVs를 이용한다:

[0203] $$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

[0204] (이때 $\hat{\mu}_j$ 및 $\hat{\sigma}_j$ 검증된 시료들의 세트에서 j-번째 염색체 분량에 대하여 각각 차례로 추정된 평균 및 표준편차이며, 그리고 x_{ij} 테스트 시료 i에 대하여 관찰된 j-번째 염색체 분량이다.)

[0205] 연관된 태아 분획:

[0206] $$FF_{ij} = 2 \times \left| \frac{NCV_{ij} \times \hat{\sigma}_j}{\hat{\mu}_j} \right| = 2 \times NCV \times CV$$

[0207] 따라서, 관심 염색체의 모든 NCV의 경우, 주어진 NCV 값과 연관된 예상된 태아 분획은 영향을 받지 않은 시료들의 집단에 걸쳐 관심 염색체에 대한 염색체 비율의 평균과 표준 편차에 근거하여 CV로부터 산출될 수 있다.

[0208] 후속적으로, 태아 분획과 NCV 값 사이의 상관관계에 근거하여, 정상적 분포 분위수(quantiles)에 근거하여 양성(영향을 받은)으로 결정되는 시료 이상에서 판단 경계가 선택될 수 있다. 상기에서 설명된 바와 같이, 참양성의 탐지와 가음성 결과의 비율 사이에 최적의 균형에 대하여 임계치가 설정된다. 따라서, 설정된 임계치는 가양성과 가음성을 최적화시키기 위하여 선택된다.

[0209] 특정 구체예들은 태아 핵산 분자들과 모체 핵산 분자들이 포함된 생물학적 시료 에서 태아 염색체 홀배수체의 출생전 진단을 제공하는 방법을 제시한다. 생물학적 테스트 시료, 가령, 모체 혈장 시료로부터 유도된 태아 핵산 분자들과 모체 핵산 분자들의 혼합물의 최소한 일부분으로부터 서열 정보를 획득하고, 서열화 데이터로부터 하나 또는 그 이상의 관심 염색체에 대한 정규화 염색체 분량이 산출되며, 및/또는 관심대상의 하나 또는 그 이상의 세그먼트들에 대한 세그먼트 분량을 정규화시키고, 그리고 테스트 시료 에서 각각 관심 염색체에 대한 염색체 분량 및/또는 관심 세그먼트에 대한 세그먼트 분량과 다수의 검증된 (정상적) 시료들에서 확립된 임계치 값 간에 통계학적으로 유의한 차이를 결정하고, 그리고 통계학적 차이에 근거하여 출생전 진단을 제공하여 진단이 이루어진다. 이 방법의 단계(160)에서 설명된 바와 같이, 정상 또는 발병 진단이 이루어진다. 정상 또는 발병의 진단이 확신이 없을 경우 "무소명"이 제시된다.

[0210] 일부 구체예들에 있어서, 2개의 임계치가 선택될 수 있다. 시료들이 "영향을 받은(Affected)"으로 분류되는 것 이상의 가양상 비율을 최소화시키도록 제1 임계치가 선택되며, 그리고 시료들이 "영향을 받지 않은" 것보다 낮은 가음성 비율을 최소화시키도록 제2 임계치가 선택된다. 제2 임계치 이상 그러나 제1 임계치 아래의 NCV를 갖는 시료는 "홀배수체로 의심되거나" 또는 "무소명" 시료들로 분류될 수 있고, 홀배수체의 존재 또는 부재는 독립적 평균에 의해 확인될 수 있다. 제1 임계치와 제2 임계치 사이 영역은 "무소명" 영역으로 지칭될 수 있다.

[0211] 일부 구체예들에 있어서, 표 1에 의심되는 임계치와 무소명 임계치들이 나타나있다. 볼 수 있는 바와 같이, NCV의 임계치들은 상이한 염색체 간에 변화된다. 일부 구체예들에 있어서, 임계치들은 상기에서 설명된 바와 같이, 이 시료에 대한 FF에 따라 변화된다. 본 명세서에서 적용되는 임계치 기술은 일부 구체예들에서 개선된 민감성 및 선택성에 기여한다.

표 1

표 1: 무소명 범위를 제시하는 의심되는 그리고 영향을 받은 NCV 임계치

	의심되는	영향을 받은
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

[0212]

[0213] **서열 커버리지 결정**

[0214] *서열 커버리지 결정을 위한 일반적 공정*

[0215] 공개된 일부 구체예들은 잡음이 낮고, 신호가 높은 서열 커버리지 분량을 결정하는 방법들이 제시되는데, 통상적인 방법들에 의해 획득된 서열 커버리지 분량과 비교하여 개선된 민감성, 선택성, 및/또는 효과로 다양한 유전적 상태와 관련된 복제수를 결정하기 위하여 데이터를 제시한다. 특정 구체예들에 있어서, 서열 커버리지 분량을 획득하기 위하여 테스트 시료로부터 서열들이 가공된다.

[0216] 이 공정은 다른 원천으로부터 이용가능한 특정 정보를 이용한다. 일부 실행에 있어서, 모든 정보는 영향을 받지 않은 것으로 공지된 훈련용 세트 시료 (가령, 홀배수체가 아님)로부터 획득된다. 다른 구체예들에서, 다른 테스트 시료들로부터 일부 정보 또는 모든 정도가 획득되는데, 동일한 공정에서 다중 시료들이 분석되는 것과 같이 "그때그때(on-the-fly)" 제공될 수 있다.

[0217] 특정 구체예들에 있어서, 데이터 잡음을 감소시키기 위하여 서열 마스크가 이용된다. 일부 구체예들에 있어서, 상기 관심 대상 서열과 이의 정규화 서열들이 마스크된다. 일부 구체예들에 있어서, 관심대상의 상이한 염색체 또는 세그먼트들이 고려될 때, 상이한 마스크들이 이용될 수 있다. 예를 들면, 염색체 13이 관심 염색체일 때, 하나의 마스크 (또는 마스크 집단)가 이용될 수 있고, 염색체 21이 관심 염색체일 때, 상이한 마스크 (또는 마스크 집단)가 이용될 수 있다. 특정 구체예들에 있어서, 빈의 해상력(resolution)에서 마스크가 정의된다. 따라서, 한 실시예에서, 마스크 해상력(resolution)은 100 kb이다. 일부 구체예들에 있어서, 분명한 마스크는 염색체 Y에 적용될 수 있다. 염색체 Y의 마스크된 배제 영역들은 다른 관심 염색체의 경우보다 더 뛰어난 해상도(1kb)로 제공될 수 있으며, 이는 2013년 6월 17일자로 제출된 US 가특허 출원 번호 61/836,057에서 설명된다[attorney docket no. ARTEP008P]. 마스크들은 배제된 게놈 영역들을 식별하는 파일 형태로 제공된다.

[0218] 특정 구체예들에 있어서, 상기 공정은 관심 대상 서열의 프로파일에서 빈간(bin-to-bin)의 변이를 제공하기 위하여 정규화된 커버리지의 예상 값을 이용하는데, 빈간의 변이는 상기 테스트 시료에 대한 CNV 결정에 있어서 충분한 정보를 제공하지 못한다. 이 공정은 전체 게놈에 걸쳐 각 빈에 대하여 또는 상기 참조 게놈 에서 최소한 강건한 염색체의 빈들에 대하여 정규화된 커버리지의 예상 값에 따라 커버리지 수량을 조절한다(하기 작동(317)에서의 사용을 위하여). 예상 값은 영향을 받지 않은 시료들의 훈련용 세트로부터 결정될 수 있다. 예로써, 예상 값은 상기 훈련용 세트 시료들의 중앙값일 수 있다. 이 시료들의 예상된 커버리지 값은 빈에 대하여 정렬된 특유한 비-다중 태그의 수를 참조 게놈의 강건한 염색체 에서 모든 빈에 대하여 정렬된 특유한 비-다중 태그 전체 수로 나눠서 결정될 수 있다.

[0219] 도 2는 관심 대상 서열의 커버리지를 결정하기 위한 공정(200)의 순서도를 나타내는데, 이는 블록(214)에서 테스트 시료에서 관심 대상 서열의 복제 수를 평가하는데 이용된다. 이 공정은 영향을 받지 않은 훈련용 시료들

에 걸쳐 공통적인 조직적 변이를 제거하는데, 이 변이는 CNV 평가를 위한 분석에서 잡음을 증가시킨다. 이 공정은 테스트 시료에 특이적인 GC도 또한 제거하는데, 그렇게 함으로써 데이터 분석에서 잡음에 대한 신호 비율이 증가된다.

[0220] 이 공정은 블록(202)에서 나타낸 바와 같이 테스트 시료의 서열 리드를 제공함으로써 시작된다. 일부 구체예들에 있어서 엄마와 태아의 cfDNA가 포함된 임산부 혈액으로부터 획득된 DNA 세그먼트들를 서열화함으로써 서열 리드가 획득된다. 이 공정은 상기 관심 대상 서열이 함유된 참조 게놈에 서열 리드를 정렬하도록 진행되어, 테스트 서열 태그가 제공된다. 블록(204). 상기 참조 서열 상에서 각 빈에 테스트 서열 태그 계수는 이 빈의 커버리지를 한정시킨다. 블록(206). 일부 구체예들에 있어서, 하나 이상의 부위에 대하여 정렬된 리드는 배제된다. 일부 구체예들에 있어서, 동일한 부위에 정렬된 다중 리드는 배제되거나 또는 단일 리드 계수로 감소된다. 일부 구체예들에 있어서, 배제된 부위들에 대하여 정렬된 리드 또한 배제된다. 따라서, 일부 구체예들에서, 오직 특유하게 정렬된, 배제되지 않은 부위들에 대하여 정렬된 비-다중 태그만 계수되어 각 빈에 대한 커버리지를 결정하는 배제되지 않은 부위 계수 (NES 계수)가 제공된다. 일부 구체예들에 있어서, 각 빈의 커버리지는 동일한 시료에서 정규화 서열의 커버리지로 나누어, 정규화된 커버리지 분량이 제공된다.

[0221] 그 다음 공정(200)은 관심 대상 서열의 포괄적 프로파일을 제공한다. 상기 포괄적 프로파일은 영향을 받지 않은 훈련용 시료들의 훈련용 세트로부터 획득된 각 빈에서 예상된 커버리지을 포함한다. 블록(208). 공정(200)은 포괄적-프로파일-제거된 커버리지를 획득하기 위하여 예상된 커버리지에 따라 테스트 서열 태그의 정규화된 커버리지 분량을 조절함으로써, 훈련용 시료에 공통적인 변이를 제거한다. 블록(210). 일부 구체예들에 있어서, 블록(208)에서 제시되는 훈련용 세트로부터 획득된 예상된 커버리지는 상기 훈련용 시료들에 걸친 중앙값이다. 일부 구체예들에 있어서, 작동(2010)은 상기 정규화된 커버리지로부터 예상된 커버리지를 차감함으로써 정규화된 커버리지 분량을 조정한다. 다른 구체예들에서, 작동(2010)은 상기 정규화된 커버리지 분량을 상기 예상된 커버리지로 나눈다.

[0222] 더욱이, 공정(200)은 포괄적 프로파일을 제거하기 위하여 조정된 커버리지 분량을 추가 조절함으로써 테스트 시료에 특이적인 GC 편향을 제거한다. 블록(212)에 나타낸 것과 같이, 상기 공정은 테스트 시료에서 GC 함량 수준과 포괄적-프로파일-제거된 커버리지 사이의 상관관계에 근거하여 포괄적-프로파일-제거된 커버리지를 조절하고, 그렇게 함으로써 시료-GC-교정된 커버리지가 획득된다. 영향을 받지 않은 훈련용 시료들에서 공통적인 조직적 변이에 대한 조정과 개체-내 GC 편향의 조정 후, 이 공정은 개선된 민감성 및 특이성으로 이 시료의 CNV를 평가하기 위한 커버리지 분량을 제공한다.

[0223] *서열 커버리지 결정을 위한 예시적인 공정에 대한 상세*

[0224] 도 3a는 테스트 시료로부터 서열 데이터에서 잡음을 줄이기 위하여 공정(301)의 예를 제시한다. 도 3b-3j는 공정의 다양한 단계에서 데이터 분석을 제시한다. 도 3a에 나타낸 바와 같이, 도시된 공정은 하나 또는 그 이상의 시료들로부터 cfDNA 추출로 시작된다. 블록(303) 참고. 적절한 추출 공정들 및 장치는 본 명세서의 도처에서 설명된다. 일부 구체예들에 있어서, 2013년 3월 15일자 제출된 US 특허 출원 번호 61/801,126 (이의 전문이 본 명세서에 참고로 포함된다)에서 설명된 공정은 cfDNA를 추출한다. 일부 실행에 있어서, 상기 장치는 복합화된 라이브러리 및 서열 데이터를 함께 제공하기 위하여 다중 시료들로부터 cfDNA를 처리한다. 도 3a에서 블록(305) 및 (307) 참고. 일부 구체예들에 있어서, 상기 장치는 나란한 8개 또는 그 이상의 테스트 시료들로 부터 cfDNA를 처리한다. 본 명세서의 도처에서 설명된 바와 같이, 서열화 시스템은 cfDNA 단편들의 코드화된 (가령, 바코드화된) 라이브러리를 만들기 위하여 추출된 cfDNA를 처리할 수 있다. 서열화기는 cfDNA의 라이브러리를 서열화하여 매우 많은 수의 서열 리드를 만든다. 복합화된 시료들에서 시료당 코딩은 리드의 탈-복합화를 허용한다. 8개 또는 그 이상의 시료들 각각은 수십만 또는 수백만 리드를 보유할 수 있다. 상기 공정은 도 3a에서 추가 작동 전에 리드를 필터링할 수 있다. 일부 구체예들에 있어서, 리드 필터링은 오류 및 저품질 리드를 걸러내기 위하여 서열화기에서 실행되는 소프트웨어 프로그램에 의해 실행되는 품질-필터링 공정이다. 예를 들면, Illumina의 Sequencing Control Software (SCS) 및 Consensus Assessment of Sequence and 변이 소프트웨어 프로그램은 하류 분석을 위하여 생물학적으로 관련된 정보를 제공하기 위하여, 상기 서열화 반응에 의해 생성된 미가공 이미지 데이터를 강도(intensity) 점수, 염기 소명(base calls), 품질 점수화된 정렬, 그리고 추가 포멧으로 전환시킴으로써, 오류 및 저품질 리드를 걸러낸다.

[0225] 상기 서열화기 또는 다른 장치에 의해 시료에 대한 리드가 생성된 후, 상기 시스템의 요소는 컴퓨터에 의해 이 리드를 참조 게놈에 정렬시킨다. 블록(309) 참고. 정렬은 본 명세서의 도처에서 설명된다. 상기 정렬은 태그를 만들고, 이 태그는 상기 참조 게놈 상에 특유 위치를 명시하는 주해가 붙은 위치 정보를 갖는 리드 서열을

포함한다. 특정 실행에 있어서, 상기 시스템은 중복 리드없이 제1 패스 정렬을 실행하고 - 동일한 서열들을 갖는 2개 또는 그 이상의 리드 - 그리고 후속적으로 비-중복된 서열 태그를 만들기 위하여 중복된 리드를 제거하거나 또는 중복 리드는 동일 리드로 계수한다. 다른 시행들에 있어서, 상기 시스템은 중복된 리드를 제거하지 않는다. 일부 구체예들에 있어서, 이 공정은 특유하게 정렬된 태그를 만들기 위하여 게놈 상에 다중 위치에 정렬된 리드를 고려사항으로부터 제거한다. 일부 구체예들에 있어서, 배제되지 않은 부위들 (NESs)에 매핑된 특유하게 정렬된, 비-다중 서열 태그는 배제되지 않은 부위 계수 (NES 계수)를 얻기 위하여 계수되고, 이는 커버리지를 예상하기 위한 데이터를 제공한다.

[0226] 도처에서 설명된 바와 같이, 배제된 부위들은 서열 태그를 계수하기 위한 목적으로 배제되었는 참조 게놈의 영역들에서 발견된 부위들이다. 일부 구체예들에 있어서, 배제된 부위들은 반복적 서열들이 포함된 염색체 영역들, 가령, 중심체(centromeres) 및 말단소체(telomeres), 그리고 하나 이상의 염색체에 공통적인 염색체 영역들, 가령, Y-염색체 상에 있는 영역들이 X 염색체에도 존재하는 영역에서 발견된다. 비-배제된 부위들(NESs)은 서열 태그를 계수하기 위한 목적으로 배제되지 않은 참조 게놈 영역에 있는 부위들이다.

[0227] 그 다음, 상기 시스템은 정렬된 태그를 상기 참조 게놈 상의 빈으로 분할한다. 블록(311) 참고. 이 빈들은 상기 참조 게놈의 길이를 따라 공간배열되어 있다. 일부 구체예들에 있어서, 전체 참조 게놈은 연속 빈 안으로 분할되는데, 이는 정의된 동일한 크기 (가령, 100 kb)를 가질 수 있다. 대안으로, 이 빈들은 역학적으로, 아마도 시료 기반에 근거하여 결정된 길이를 가질 수 있다. 서열화 깊이는 최적의 빈 크기 선택에 영향을 준다. 역학적 크기를 가진 빈은 라이브러리 크기에 의해 결정된 이들의 크기를 가질 수 있다. 예를 들면, 상기 빈 크기는 평균적으로 1000개 태그를 수용하는데 필요한 서열 길이가 되는 것으로 결정될 수 있다.

[0228] 각 빈은 고려중인 시료로부터 다수의 태그를 보유한다. 정렬된 서열의 "커버리지"를 반영하는 태그 수는 필터링을 위한 시작점으로서 기능을 하며, 그렇지 않으면 이 시료에서 복제 수 변이를 확실하게 결정하기 위하여 시료 데이터를 클리닝하는 기능을 한다. 도 3a는 블록(313)에서 (321)에 클리닝 작동을 보여준다.

[0229] 도 3a에서 도시된 구체예에서, 이 공정은 상기 참조 게놈의 빈에 마스크를 적용한다. 블록(313) 참고. 상기 시스템은 다음의 공정 작동 일부 또는 전부에서 고려중인 마스크된 빈에서 커버리지는 배제할 수 있다. 많은 경우에 있어서, 도 3a에서 나머지 작동중 어느 것에서도 마스크된 빈의 커버리지 값은 고려되지 않는다.

[0230] 다양한 실행에 있어서, 시료 간의 높은 변동성을 나타내기 위하여 발견된 게놈 영역들에 대한 빈을 제거하기 위하여 하나 또는 그 이상의 마스크가 적용된다. 관심 염색체 (가령, chr13, 18, 그리고 21) 및 다른 염색체 모두에 대하여 이러한 마스크들이 제공된다. 도처에서 설명된 바와 같이, 관심 염색체는 복제 수 변이 또는 다른 이상을 잠재적으로 품고있는 것으로 고려되는 염색체다.

[0231] 일부 실행에 있어서, 다음의 방식을 이용하여 훈련용 검증된 시료들 세트로부터 마스크들이 확인된다. 우선, 각 훈련용 세트 시료는 도 3a에서 작동(315)-(319)에 따라 가공되고, 필터된다. 상기 정규화된 그리고 교정된 커버리지 수량은 각 빈에 표시되고, 통계, 이를 테면 표준 편차, 중앙값 절대 편차, 및/또는 변이계수가 각 빈에 대하여 산출된다. 각 관심 염색체에 대한 다양한 필터 조합이 평가될 수 있다. 상기 필터 조합은 관심 염색체의 빈에 대하여 하나의 필터를 제공하고, 모든 다른 염색체의 빈에 대하여 상이한 필터를 제공한다.

[0232] 일부 실행에 있어서, 마스크를 획득한 후 (가령, 상기에서 설명된 바와 같이 관심 염색체에 대한 컷-오프를 선택함으로써), 정규화 염색체(또는 염색체 집단) 선택이 재고려된다. 상기 서열 마스크를 적용한 후, 정규화 염색체 또는 염색체를 선택하는 공정은 본 명세서의 도처에서 설명된 바와 같이 실행될 수 있다. 예를 들면, 염색체의 모든 가능한 조합이 정규화 염색체로 평가되고, 영향을 받은 시료들과 영향을 받지 않은 시료들을 구별하는 이들의 능력에 따라 등급화된다. 이 공정은 상이한 최적의 정규화 염색체 또는 염색체 집단을 찾을 수 있다(또는 찾지 않을 수 있다). 다른 구체예들에서, 정규화 염색체는 모든 검증된 시료들에 걸쳐 관심 대상 서열에 대한 서열 분량에서 최소 변동성을 야기하는 것들이다. 상이한 정규화 염색체 또는 염색체 집단이 확인되면, 이 공정은 빈의 전술한 식별의 필터를 선택적으로 실행한다. 아마도 새로운 정규화 염색체(들)이 상이한 컷오프(cut-offs)를 야기할 수 있다.

[0233] 특정 구체예들에 있어서, 상이한 마스크가 염색체 Y에 적용된다. 적절한 염색체 Y 마스크의 예는 2013년 6월 17일자로 제출된 US 가특허 출원 번호 61/836,057에서 설명된다[attorney docket no. ARTEP008P], 이는 모든 목적을 위하여 본 명세서에 참고로 포함된다.

[0234] 상기 시스템이 계산적으로 빈들을 마스크한 후, 마스크에 의해 배제되지 않은 빈들에서 커버리지 값을 계산적으로 정규화시킨다. 블록(315) 참고. 특정 구체예들에 있어서, 상기 시스템은 참조 게놈 또는 이의 일부분에서

대부분의 또는 모든 커버리지 (가령, 상기 참조 게놈의 강건한 염색체에서 커버리지)에 대응하여 각 빈에서 테스트 시료 커버리지 값을 정규화시킨다 (가령, 빈 당 NES 계수). 일부 경우들에 있어서, 상기 시스템은 고려되는 빈의 계수를 참조 게놈 안의 모든 강건한 염색체에 정렬되는 모든 배제되지 않은 부위들의 총 수로 나눔으로써 테스트 시료 커버리지 값 (하나의 빈에 대한)을 정규화시킨다. 설명된 바와 같이, 강건한 염색체는 홀배수체일 것 같지 않은 염색체이다. 특정 구체예들에 있어서, 상기 강건한 염색체는 염색체 13, 18, 그리고 21이외의 모든 상염색체이다.

[0235] 빈의 변환된 계수 값 또는 커버리지는 추가 공정을 위한 "정규화된 커버리지 분량"이라고 지칭된다. 상기 정규화는 각 시료에 특유한 정보를 이용하여 실행된다. 전형적으로, 훈련용 세트의 정보는 이용되지 않는다. 정규화는 상이한 라이브러리 크기 (그리고 결과적으로 상이한 수의 리드 및 태그)를 보유한 시료들의 커버리지 분량이 동일한 기초에서 처리되도록 한다. 일부 후속 공정 작동은 고려중인 테스트 시료에 이용되는 라이브러리보다 더 큰 또는 더 작은 라이브러리로부터 서열화될 수 있는 훈련용 시료들로부터 유도된 커버리지 분량을 이용한다. 전체 참조 게놈 (또는 최소한 강건한 염색체)에 대하여 정렬된 리드 수에 기초한 정규화없이, 훈련용 세트로부터 유도된 매개변수를 이용한 처리는 일부 실행에 있어서 신뢰성이 없거나 또는 일반화시키지 못할 수 있다.

[0236] 도 3b는 많은 시료들에 대하여 염색체 21, 13, 그리고 18에 걸친 커버리지를 설명한다. 일부 시료들은 서로 상이하게 가공되었다. 그 결과, 임의의 주어진 게놈 위치에서 광범위한 시료간(sample-to-sample) 변이를 볼 수 있다. 정규화는 이러한 시료간 변이의 일부를 제거한다. 도 3c의 좌측 패널은 전체 게놈에 걸쳐 정규화된 커버리지 분량을 나타낸다.

[0237] 도 3a의 구체예에 있어서, 상기 시스템은 작동(315)에서 생성된 상기 정규화된 커버리지 분량으로부터 "포괄적 프로파일"을 제거 또는 감소시킨다. 블록 (317) 참고. 이 작동은 상기 게놈의 구조, 라이브러리 생성 공정, 그리고 상기 서열화 공정으로부터 발생되는 정규화된 커버리지 분량에서 조직적 편향을 제거한다. 추가로, 이 작동은 임의의 주어진 시료에서 예상된 프로파일로부터 임의의 조직적 선형 편차에 대하여 교정하도록 기획된다.

[0238] 일부 실행에 있어서, 상기 포괄적 프로파일 제거는 각 빈의 상기 정규화된 커버리지 분량을 각 빈의 대응하는 예상된 값으로 나누는 것이 관련된다. 다른 구체예들에서, 상기 포괄적 프로파일 제거는 각 빈의 정규화된 커버리지 분량으로부터 각 빈의 예상된 값을 차감시키는 것이 관련된다. 상기 예상된 값은 영향을 받지 않은 시료들 (또는 X염색체의 경우 영향을 받지 않은 여성 시료들)의 훈련용 세트로부터 획득된다. 발병안된 시료들은 관심 염색체에 대하여 홀배수체를 보유하지 않은 것으로 공지된 개체의 시료들이다. 일부 실행에 있어서, 상기 포괄적 프로파일 제거는 각 빈의 정규화된 커버리지 분량으로부터 각 빈의 예상된 값(훈련용 세트로부터 획득)을 차감시키는 것이 관련된다. 일부 구체예들에 있어서, 상기 공정은 훈련용 세트를 이용하여 결정하였을 때 각 빈에 대한 정규화된 커버리지 분량의 중앙값을 이용한다. 환언하면, 상기 중앙 값은 상기 예상된 값이다.

[0239] 일부 구체예들에 있어서, 상기 포괄적 프로파일 제거는 이 포괄적 프로파일에서 시료 커버리지 의존성에 대한 선형 교정을 이용하여 실행된다. 표시된 바와 같이, 상기 포괄적 프로파일은 상기 훈련용 세트에서 측정된 각 빈의 예상값 (예를 들면 각 빈에 대한 중앙 값)이다. 이들 구체예들은 각 빈에 대하여 획득된 포괄적 중앙 프로파일에 대항하여 테스트 시료의 정규화된 커버리지 분량을 피팅함으로써 획득된 강건한 선형 모델을 이용할 수 있다. 일부 구체예들에 있어서, 상기 선형 모델은 포괄적 중앙 (또는 다른 예상 값) 프로파일에 대항하여 이 시료의 관찰된 정규화된 커버리지 분량을 회귀시킴으로써 획득된다.

[0240] 상기 선형 모델은 시료 커버리지 분량이 포괄적 프로파일 예상 값과 선형 관계를 보유한다는 가정에 근거한다. 도 3d 참고. 이러한 경우에서, 포괄적 프로파일의 예상된 커버리지 분량에 이 시료 정규화된 커버리지 분량을 회귀시키면 기울기(slope) 및 절편(intercept)을 갖는 선이 생성된다. 특정 구체예들에 있어서, 이러한 선의 기울기 및 절편은 각 빈의 포괄적 프로파일 값으로부터 "예상된" 커버리지 분량을 산출하는데 이용된다. 일부 실행에 있어서, 포괄적 프로파일 교정은 각 빈의 정규화된 커버리지 분량을 빈에 대한 예상된 커버리지 분량으로 나누는 것이 관련된다. 빈에 대한 예상된 값은 기울기 산물에 절편을 추가하고, 빈에 대한 포괄적 프로파일 예상 값 (가령, 중앙값)에 의해 결정될 수 있다. 환언하면, 이 작동은 각 빈의 정규화된 커버리지 분량을 다음의 식에 따라 산출된 대응하는 예측으로 나눔으로써 실행될 수 있다:

[0241] *예상된 커버리지 분량 = 시료 정규화된 빈 커버리지/(기울기*포괄적 프로파일 + 절편) - 1*

[0242] 기울기 및 절편은 도 3d에 나타낸 바와 같은 선으로부터 획득된다. 도 3c에서는 포괄적 프로파일 제거의 예시

가 설명된다. 좌측 패널은 많은 시료에 걸쳐 정규화된 커버리지 분량에서 높은 빈간 변이를 보여준다. 우측 패널은 상기에서 설명된 바와 같이, 포괄적 프로파일 제거 후 동일한 정규화된 커버리지 분량을 보여준다.

[0243] 상기 시스템이 블록(317)에서 포괄적 프로파일 변이를 제거 또한 감소시킨 후, 시스템은 시료-안 GC (구아닌-시토신) 함량 변이에 대하여 교정한다. 블록 (319) 참고. 모든 빈은 GC로부터 이의 고유 분획성 기여(contribution)를 갖는다. 이 분획은 빈에서 G와 C 뉴클레오티드의 수를 빈에서 전체 뉴클레오티드 수(가령, 100,000)로 나눔으로써 결정된다. 일부 빈은 다른 빈보다 더 큰 GC 분획을 보유할 것이다. 도 3e 및 3f에서 나타낸 것과 같이, 상이한 시료들은 상이한 GC 편향을 나타낸다. 이들 차이 및 이들의 교정은 하기에서 더 설명될 것이다. 도 3e-g는 GC 분획(하나의 빈에 있어서)에 대한 함수로써 포괄적 프로파일 교정된, 정규화된 커버리지 분량 (하나의 빈에 있어서)을 보여준다. 놀랍게도, 상이한 시료들은 상이한 GC 의존성을 나타낸다. 일부 시료들은 점감적(monotonically decreasing) 의존성 (도 3e에서와 같이)을 보이며, 한편 다른 시료들은 쉼표 모양의 의존성 (도 3f 및 3g에서와 같이)을 나타낸다. 이들 프로파일은 각 시료에 대하여 특유할 수 있기 때문에, 이 단계에서 설명된 교정은 각 시료에 대하여 별도로 그리고 특유적으로 실행된다.

[0244] 일부 구체예들에 있어서, 상기 시스템은 도 3e-g에서 설명된 바와 같이 GC 분획에 기초하여 빈을 계산적으로 배열한다. 그 다음 유사한 GC 함량을 가진 다른 빈의 정보를 이용하여 빈의 포괄적 프로파일 교정된, 정규화된 커버리지 분량을 교정한다. 이 교정은 마스크안된 각 빈에 적용된다.

[0245] 일부 공정들에 있어서, 다음과 같은 방식으로 GC 함량에 대하여 각 빈이 교정된다. 상기 시스템은 고려중인 빈의 것들과 유사한 GC 분획을 보유한 빈을 계산적으로 선택하고, 그 다음 선택된 빈에서의 정보로부터 교정 매개변수를 결정한다. 일부 구체예들에 있어서, 유사한 GC 분획을 보유한 이들 빈들은 임의적으로 정의된 유사한 컷-오프 값을 이용하여 선택된다. 한 실시예에서, 모든 빈의 2%가 선택된다. 이들 빈은 고려중인 빈에 가장 유사한 GC 함량을 갖는 빈의 2% 이다. 예를 들면, 약간 더 많은 GC 함량을 가진 빈의 1%와 약간 더 적은 GC 함량을 가진 빈의 1%가 선택된다.

[0246] 선택된 빈을 이용하여, 상기 시스템은 계산적으로 교정 매개변수를 결정한다. 한 실시예에서, 상기 교정 매개변수는 선택된 빈에서 상기 정규화된 커버리지 분량의 대표 값이다 (포괄적 프로파일 제거 후). 이러한 대표 값의 예로는 선택된 빈에서 정규화된 커버리지 분량의 중앙값 또는 평균이 포함된다. 상기 시스템은 고려중인 빈에 대한 산출된 교정 매개변수를 고려중인 빈에 대한 정규화된 커버리지 분량 (포괄적 프로파일 제거 후)에 적용시킨다. 일부 실행에 있어서, 대표 값 (가령, 중앙 값)은 고려중인 빈의 상기 정규화된 커버리지 분량으로부터 차감된다. 일부 구체예들에 있어서, 정규화된 커버리지 분량의 중앙 값 (또는 다른 대표 값)은 강건한 상 염색체 (염색체 13, 18, 그리고 21이외의 모든 상염색체)에 대한 커버리지 분량만을 이용하여 선택된다.

[0247] 가령, 100kb 빈을 이용하는 한 실시예에서, 각 빈은 특유한 값의 GC 분획을 보유할 것이며, 그리고 이 빈은 이들의 GC 분획 함량에 기초하여 집단으로 분할된다. 예를 들면, 이 빈은 50개 집단으로 분할되며, 이때 집단 경계는 GC% 분포의 (0, 2, 4, 6, …, 그리고 100) 분위수에 대응된다. 정규화된 커버리지 분량의 중앙값은 동일한 GC 집단 (이 시료 에서서)에 매핑된 강건한 상염색체로부터 빈의 각 집단에 대하여 산출되며, 그리고 그 다음 이 중앙 값은 상기 정규화된 커버리지 분량 (동일한 GC 집단 에서서 전체 게놈에 걸쳐 모든 빈에 대한) 정규화된 커버리지 분량으로부터 차감된다. 이것은 임의의 주어진 시료 에서 강건한 염색체로부터 추정된 GC 교정을 동일한 시료에서 잠재적으로 영향을 받은 염색체에 적용된다. 예를 들면, 0.338660 내지 0.344720 사이의 GC 함량을 보유한 강건한 염색체 상의 모든 빈은 함께 집단화되며, 중앙값은 이 집단으로부터 산출되고, 그리고 이 GC 범위에서 빈의 정규화된 커버리지로부터 차감되는데, 이 빈은 게놈상의 임의의 위치에서 발견될 수 있다 (염색체 13, 18, 21, 그리고 X 제외). 특정 구체예들에 있어서, 염색체 Y는 이 GC 교정 공정에서 배제된다.

[0248] 도 3g는 방금 설명된 바와 같이, 정규화된 커버리지 분량의 중앙값을 교정 매개변수로 이용하여 GC 교정의 응용을 보여준다. 좌측 패널은 GC 분획 프로파일에 대하여 교정안된 커버리지 분량을 보여준다. 나타낸 바와 같이, 상기 프로파일은 비-선형 모양을 갖는다. 우측 패널은 교정된 커버리지 분량을 나타낸다. 도 3h는 GC 분획 교정 (좌측 패널) 전, 그리고 GC 분획 교정 후(우측 패널) 많은 시료들에 대한 정규화된 커버리지를 보여준다. 도 3i는 GC 분획 교정 전 (적색) 그리고 GC 분획 교정 후 (녹색) 많은 테스트 시료들에 대한 정규화된 커버리지의 변이계수(CV)를 보여주는데, 이때 GC 교정으로 정규화된 커버리지에서 실질적으로 더 적은 변이를 유도한다.

[0249] 상기 공정은 상대적으로 단순하게 실행되는 GC 교정이다. GC 편향을 교정하기 위한 대한적 방식은 스플라인(spline) 또는 다른 비-선형 피팅 기술을 이용하는데, 이는 연속성 GC 공간에 적용되며, 그리고 GC 함량에 의한 커버리지 분량의 폐기(binning)와 관련되지 않는다. 적절한 기술의 예로는 연속성 뢰스(loess) 교정 및 스무스

(smooth) 스플라인 교정을 포함한다. 피팅 함수은 고려중인 시료의 경우 GC 함량에 대하여 빈간(bin-by-bin) 정규화된 커버리지 분량으로부터 유도될 수 있다. 각 빈에 대한 교정은 고려 중인 빈의 GC 함량을 피팅 함수에 적용시킴으로써 산출된다. 예로써, 상기 정규화된 커버리지 수량은 고려중인 빈의 GC 함량에서 스플라인의 예상된 커버리지 값을 차감시킴으로써 조정될 수 있다. 대안으로, 상기 조정은 스플라인 피트에 따라 예상된 커버리지 값의 분할에 의해 이루어질 수 있다.

[0250] 작동(319)에서 GC-의존성의 교정 후, 상기 시스템은 고려중인 시료에서 아웃라이어 빈을 계산적으로 제거한다 - 블록(321) 참고. 이 작동은 단일 시료 필터링 또는 트리밍(trimming)이라고도 불릴 수 있다. 도 3j는 GC 교정 후에라도, 상기 커버리지는 작은 영역들에서 여전히 시료-특이적 변이를 갖는다는 것을 보여준다. 예를 들면, 상기 예상된 값으로부터 예상치 못한 높은 편차가 발생되는 염색체 12 상의 위치 1.1 e8에서의 커버리지를 참고하라. 물질 게놈에서 작은 복제 수 변이로 발생될 가능성도 있다. 대안으로, 이는 복제 수 변이와 무관한 서열화에서 기술적인 이유들로 인한 것일 수 있다. 전형적으로, 이 작동은 오직 강건한 염색체에만 적용된다.

[0251] 한 가지 예로써, 상기 시스템은 필터링을 위하여 고려중인 빈을 품고있는 염색체에서 모든 빈에 걸쳐 GC 교정된 정규화된 커버리지 수량의 중앙값으로부터 3이상의 중앙값 절대 편차의 GC 교정된 정규화된 커버리지 수량을 보유한 임의의 빈들을 계산적으로 필터한다. 한 실시예에서, 컷-오프 값은 표준 편차와 일치되도록 조정된 3 중앙값 절대적 편차로 정의되며, 따라서 실질적으로 컷오프는 중앙값으로부터 1.4826*중앙 절대적 편차다. 특정 구체예들에 있어서, 이 작동은 강건한 염색체와 홀배수체로 의심되는 염색체 모두가 포함된 시료 에서 모든 염색체에 적용된다.

[0252] 특정 실행에 있어서, 품질 관리로 특징될 수 있는 추가 작동이 실행된다. 블록(323) 참고. 일부 구체예들에 있어서, 품질 관리 계측은 임의의 잠재적 분모 염색체, 가령 "정규화 염색체" 또는 "강건한 염색체"가 홀배수체인지 또는 그렇지않으면 상기 테스트 시료가 관심 대상 서열 에서 복제 수 변이를 갖고 있는지를 결정하는데 부적절한지를 탐지하는 것과 관련된다. 이 공정에서 강건한 염색체가 부적절하다고 판단되면, 이 공정은 상기 테스트 시료를 폐기하고, 소명하지 않을 것이다. 대안으로, 이런 QC 계측의 실패는 소명용으로 정규화 염색체의 대체 세트의 사용을 촉발시킬 수 있다. 한 실시예에서, 품질 관리 방법은 강건한 상 염색체의 예상 값에 대하여 강건한 염색체의 실질적인 정규화된 커버리지 값을 비교한다. 상기 예상 값은 다변량 정상적 모델을 영향을 받지 않은 훈련용 시료들의 정규화된 프로파일에 피팅시키고, 데이터 또는 Bayesian 기준의 확률(likelihood)에 따라 최고 모델 구조를 선택하고(가령, 이 모델은 Akaike 정보 기준 또는 있을수 있는 Bayesian 정보 기준을 이용하여 선택되고), 그리고 QC에 사용을 위하여 최고 모델을 고정시킴으로써, 획득될 수 있다. 강건한 염색체의 정상적 모델은 예를 들면, 정상적 시료들 에서 염색체 커버리지에 대한 평균 및 표준편차를 가진 확률 함수를 확인하는 클러스터링 기술을 이용하여 획득될 수 있다. 물론, 다른 모델 형태들이 이용될 수 있다. 이 공정은 고정된 모델 매개변수들이 제공되면, 임의의 유입되는 테스트 시료에서 관찰된 정규화된 커버리지의 확률을 평가한다. 확률을 획득하고, 그렇게 함으로써 정상적 시료 세트에 대한 아웃라이어를 확인하기 위한 모델로 각 유입되는 테스트 시료를 득점화시킴로써, 이를 실행할 수 있다. 상기 훈련용 시료들의 확률로부터 테스트 시료의 확률에서 편차는 정규화 염색체 또는 시료 취급 / 인공물 가공 분석에서 비정상을 암시하는데, 이러한 비정상은 부정확한 시료 분류를 초래할 수 있다. 이러한 QC 계측을 이용하여 이들 시료 인공물과 연관된 분류상 오류를 감소시킬 수 있다. 도 3k, 우측 패널은 x-축 상에 염색체 수를, 그리고 y-축에는 상기에서 설명된 바와 같이 획득된 QC 모델과 비교에 근거한 정규화된 염색체 커버리지를 보여준다. 이 그래프는 염색체 2에 대한 과도한 커버리지를 갖는 한 가지 지료와 염색체 20에 대한 과도한 커버리지을 갖는 또다른 시료를 보여준다. 이들 시료들은 여기에서 설명된 QC 계측을 이용하여 제거될 수 있고, 또는 대체 세트의 정규화 염색체를 이용하기 위하여 전환될 수 있다. 도 3k의 좌측 패널은 염색체에 대한 확률대 NCV를 보여준다.

[0253] 도 3a에 나타낸 서열은 게놈에서 모든 염색체의 모든 빈에 이용될 수 있다. 특정 구체예들에 있어서, 상이한 공정이 염색체 Y에 적용된다. 염색체 또는 세그먼트 분량, NCV, 및/또는 NSV를 산출하기 위하여, 분량, NCV, 및/또는 NSV에 대한 표현으로 이용되는 염색체 또는 세그먼트들에서 빈의 교정된 정규화된 커버리지 분량(도 3a에서 결정된 바와 같은)이 이용된다. 블록(325) 참고. 특정 구체예들에 있어서, 평균 정규화된 커버리지 분량은 관심 염색체, 정규화 염색체, 관심대상의 세그먼트, 및/또는 정규화 세그먼트에서 모든 빈으로부터 산출되고, 본 명세서의 도처에서 설명된 서열 분량, NCV, 및/또는 NSV를 산출하기 위하여 이용된다.

[0254] 특정 구체예들에 있어서, 염색체 Y는 상이하게 처리된다. Y 염색체에 특유한 빈 세트를 마스킹함으로써 필터될 수 있다. 일부 구체예들에 있어서, Y 염색체 필터는 US 가특허 출원 번호 61/836,057에 따라 결정되며, 이는 이미 참고로 포함되어 있다. 일부 구체예들에 있어서, 상기 필터는 다른 염색체의 필터에서 보다 더 작은 빈을 마스크한다. 예를 들면, Y 염색체 마스크는 1 kb 수준에서 필터될 수 있지만, 다른 염색체 마스크는 100 kb

수준에서 필터될 수 있다. 그럼에도불구하고, Y 염색체는 다른 염색체와 동일한 빈 크기 (가령, 100 kb)에서 정규화될 수 있다.

[0255] 특정 구체예들에 있어서, 필터된 Y 염색체는 도 3a의 작동(315)에서 상기에서 설명된 바와 같이 정규화된다. 그러나, 그렇지 않으면, Y 염색체는 더 교정되지 않는다. 따라서, Y 염색체 빈은 포괄적 프로파일 제거를 겪지 않는다. 유사하게, Y 염색체 빈은 그 이후 실행되는 GC 교정 또는 다른 필터링 단계를 겪지 않는다. 그 이유는 이 시료가 가공될 때, 이 공정에서 이 시료가 남성 또는 여성인지를 모르기 때문이다. 여성 시료는 Y 참조 염색체에 대하여 정렬되는 리드를 보유하지 않아야만 한다.

[0256] **서열 마스크의 창출**

[0257] 본 명세서에서 공개된 일부 구체예들은 서열 마스크를 이용하여 관심 대상 서열 상에 비-판별 서열 리드를 걸러내기 위한(또는 마스킹) 전략을 이용하는데, 이로써 통상적인 방법들에 의해 산출된 값에 비하여 CNV 평가를 위하여 이용되는 커버리지 값에서 상대적으로 더 높은 신호, 더 낮은 잡음을 갖게 한다. 이러한 마스크들은 다양한 기술에 의해 식별될 수 있다. 한 구체예에서, 하기에서 더 상세하게 설명되는 바와 같이, 도 4a-4b에서 설명된 기술을 이용하여 마스크가 식별된다.

[0258] 일부 실행에 있어서, 관심 대상 서열의 정상적 복제 수를 보유한 것으로 알려진 대표 시료들의 훈련용 세트를 이용하여 마스크가 식별된다. 상기 훈련용 세트 시료들을 우선 정규화시키고, 그 다음 서열 범위 (가령, 프로파일)에 걸쳐 조직적 변이에 대하여 교정하고, 그리고 그 다음 하기에서 설명되는 것과 같이 GC 변동성에 대하여 이를 교정하는 기술을 이용하여 마스크가 식별될 수 있다. 테스트 시료들이 아닌 훈련용 세트 시료 상에서 정규화 및 교정이 실행된다. 일단 마스크가 식별되면, 그 다음 많은 테스트 시료들에 적용된다.

[0259] 도 4a는 이러한 서열 마스크를 창출하기 위한 공정(400)의 순서도를 보여주는데, 이는 하나 또는 그 이상의 테스트 시료들에 적용시켜, 관심 대상 서열의 빈들을 복제 수 평가의 고려에서 제거된다. 이 공정은 다수의 영향을 받지 않은 훈련용 시료들로부터 서열리드를 포함하는 훈련용 세트가 제공함으로써 시작된다. 블록(402). 그 다음 이 공정은 상기 훈련용 세트의 서열 리드를 상기 관심 대상 서열을 포함하는 참조 게놈에 정렬시키고, 그렇게 함으로써 상기 훈련용 시료들에 대한 훈련용 서열 태그를 제공한다. 블록(404). 일부 구체예들에 있어서, 배제되지 않은 부위들에 대하여 매핑된 특유하게 정렬된 비-다중 태그 만이 추가 분석에 이용된다. 이 공정은 참조 게놈을 다수의 빈으로 분할하고, 그리고 각각 영향을 받지 않은 훈련용 시료에 대하여, 각 훈련용 시료에 대한 각 빈에서 훈련용 서열 태그의 커버리지를 결정하는 공정을 수반한다. 블록(406). 이 공정은 또한 모든 훈련용 시료들에 걸쳐 각 빈에 대한 훈련용 서열 태그의 예상된 커버리지를 결정한다. 블록(408). 일부 구체예들에 있어서, 각 빈의 예상된 커버리지는 상기 훈련용 시료들에 걸쳐 중앙값 또는 평균이다. 상기 예상된 커버리지는 포괄적 프로파일을 구성한다. 그 다음 이 공정은 포괄적 프로파일에서 변이를 제거함으로써 각 훈련용 시료에 대한 각 빈에서 훈련용 서열 태그의 커버리지를 조절하고, 그에 의해서 각 훈련용 시료에 대하여 빈에서 훈련용 서열 태그의 포괄적-프로파일-제거된 커버리지를 획득한다. 블록(410). 일부 실행은 각 훈련용 시료에 존재하는 포괄적-프로파일-제거된 커버리지와 GC함량 수준 사이의 상관관계에 근거하여 각 훈련용 시료에 대한 포괄적-프로파일-제거된 커버리지를 조절하여, 그에 의해서 각각의 훈련용 시료에 대하여 훈련용 서열 태그의 시료-GC-교정된 커버리지를 획득하는 공정을 수반한다. 블록(412). 그 다음 이 공정은 참조 게놈에 걸쳐 마스크안된 빈과 마스크된 빈이 포함된 서열 마스크를 창출한다. 각 마스크된 빈은 마스킹 임계치를 초과하는 특징적 분포를 갖는다. 블록(414). 훈련용 시료들에 걸쳐 빈에서 훈련용 서열 태그의 조정된 커버리지에 대하여 이러한 분포 특징이 제시된다. 일부 실행에 있어서, 마스킹 임계치는 훈련용 시료들에 걸쳐 빈에서 정규화된 커버리지에서 관찰된 변이와 관련될 수 있다. 시료에 걸쳐 정규화된 커버리지의 높은 변이 계수 또는 중앙값 절대 편차를 가진 빈들은 각 계측의 실질적인 분포에 기초하여 식별될 수 있다. 일부 대체 실행에 있어서, 마스킹 임계치는 훈련용 시료들에 걸쳐 빈에서 정규화된 커버리지에서 관찰된 변이와 관련될 수 있다. 시료에 걸쳐 정규화된 커버리지의 높은 변이 계수 또는 중앙값 절대 편차를 가진 빈들은 각 계측의 실질적인 분포에 기초하여 마스크될 수 있다.

[0260] 일부 실행에 있어서, 마스크된 빈을 식별하기 위한 별도 컷-오프, 가령, 마스킹 임계치는 관심 염색체와 모든 다른 염색체에 대하여 정의된다. 더욱이, 각 관심 염색체에 별도로 별도의 마스킹 임계치가 정의될 수 있고, 그리고 모든 비-영향을 받은 염색체 세트에 대하여 단일 마스킹 임계치가 정의될 수 있다. 예를 들면, 염색체 13의 경우 특정 마스킹 임계치에 기초된 마스크가 정의되고, 다른 염색체의 경우 마스크를 정의하기 위하여 또 다른 마스킹 임계치가 이용된다. 비-영향을 받은 염색체는 염색체당 정의된 이들의 마스킹 임계치를 보유할 수 있다.

[0261] 각 관심 염색체에 대하여 다양한 마스킹 임계치 조합이 평가될 수 있다. 상기 마스킹 임계치 조합은 관심 염색체의 빈에 대하여 하나의 마스크를 제공하고, 모든 다른 염색체의 빈에 대하여 상이한 마스크를 제공한다.

[0262] 한 접근 방식에 있어서, 변이계수 (CV)의 값에 대한 범위 또는 시료 분포 컷-오프 측정 범위는 빈 CV 값의 실질적인 분포의 백분위수 (가령, 95, 96, 97, 98, 99)로 정의되며, 이들 컷-오프 값은 관심 염색체를 제외한 모든 상염색체에 적용된다. 더욱이, CV에 대한 컷-오프 값 백분위수 범위는 실험적 CV 분포에 대하여 정의되며, 이들 컷-오프 값은 관심 염색체 (가령, chr 21)에 적용된다. 일부 구체예들에 있어서, 관심 염색체는 X 염색체, 그리고 염색체 13, 18, 그리고 21이다. 물론, 다른 접근방식이 고려될 수 있는데; 예를 들면, 각 염색체에 대하여 별도의 최적화가 실행될 수 있다. 이와 함께, 병행하여 최적화되는 범위 (가령, 고려되는 관심 염색체의 경우 한 범위와 모든 다른 염색체에 대한 또다른 범위)는 CV 컷-오프 조합의 그리드를 정의한다. 도 4b 참고. 훈련용 세트 상에서 시스템의 수행은 2개의 컷-오프 (정규화 염색체 (또는 관심 염색체이외의 상염색체)의 것 하나와 관심 염색체의 것 하나)에 걸쳐 평가되며, 그리고 최종 배열을 위하여 최고의 실행 조합이 선택된다. 이 조합은 각 관심 염색체에 있어서 상이할 수 있다. 특정 구체예들에 있어서, 훈련용 세트 대신 확증 세트에서 수행이 평가되며, 즉 실행을 평가하기 위하여 교차-확증이 이용된다.

[0263] 일부 구체예들에 있어서, 컷-오프 범위를 결정하기 위하여 최적화되는 수행은 염색체 분량의 변이계수 (정규화 염색체의 잠정적 선별에 기초하여)이다. 이 공정은 현재 선택된 정규화 염색체 (또는 염색체)를 이용하여 관심 염색체의 염색체 분량 (가령, 비율)의 CV를 최소화시키는 컷-오프 조합을 선택한다. 한 가지 접근 방법에 있어서, 이 공정은 다음과 같이 그리드에서 컷-오프의 각 조합의 수행을 테스트한다: (1) 모든 염색체에 대한 마스크를 정의하기 위하여 컷-오프 조합을 적용시키고, 이들 마스크를 적용시켜 훈련용 세트의 태그를 필터하고; (2) 도 3a의 공정을 필터된 태그에 적용시킴으로써, 영향을 받지 않은 시료들의 훈련용 세트에 걸쳐 정규화된 커버리지를 산출하고; (3) 가령, 고려중인 염색체에 대한 빈의 정규화된 커버리지를 합산함으로써, 염색체당 대표적인 정규화된 커버리지를 결정하고; (4) 현재 정규화 염색체를 이용하여 염색체 분량을 산출하고, 그리고 (5) 염색체 분량의 CV를 결정한다. 이 공정은 선택된 필터들의 수행을 상기 훈련용 세트의 원래 부분으로부터 떨어져 나온 테스트 시료에 적용시킴으로써, 선택된 필터들의 수행을 평가할 수 있다. 즉, 이 공정은 원래 훈련용 세트를 훈련용과 테스트 부분집단으로 분할시킨다. 상기 훈련용 부분집단은 상기에서 설명된 바와 같이, 마스크 컷-오프를 정의하는데 이용된다.

[0264] 대체 구체예들에서, 커버리지의 CV 기초한 마스크를 정의하는 대신, 빈에서 훈련용 시료들에 걸쳐 정렬 결과로부터 품질 점수의 매핑 분포에 의해 마스크가 정의될 수 있다. 매핑 품질 점수는 리드가 참조 게놈에 매핑되는 특유성을 반영한다. 환언하면, 매핑 품질 점수는 리드가 잘못 정렬되는 가능성을 정량화한다. 낮은 매핑 품질 점수는 낮은 특유성 (정렬불량 가능성이 높음)과 관련된다. 상기 특유성은 리드 서열(상기 서열화기에 의해 생성된)에서 하나 또는 그 이상의 오류를 설명한다. 매핑 품질 점수의 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 제시되며, 이는 전문이 본 출원에 참고로 포함된다. 일부 실행에 있어서, 본 명세서의 매핑 품질 점수는 MapQ 스코어라고 지칭된다. 도 4b는 MapQ 스코어가 가공된 커버리지의 CV와 강력한 변함없는 상관관계를 가짐을 보여준다. 예로써, 0.4 이상의 CV를 가진 빈은 약 4 미만의 MapQ 스코어를 갖는 도 4b에서 플롯의 좌측 상에 거의 완전하게 모여있다. 따라서, 작은 MapQ를 갖는 마스킹 빈은 높은 CV를 갖는 마스킹 빈에 의해 정의된 것과 상당히 유사한 마스크를 산출한다.

[0265] **<u>증후군 특이적 조직적 편향의 제거</u>**

[0266] 상기에서 설명하는 기술은 일반적으로 CNV를 탐지하기 위한 커버리지를 결정하는데 적용될 수 있다. 더 짧은 관심 대상 서열들과 관련된 일부 구체예들에 있어서, CNV 탐지를 위한 신호는 더 낮고, 이로써 증후군 특이적 편향을 제거하기 위한 추가 공정이 요구된다. 하기에서 설명되는 공정은 게놈의 하나 또는 그 이상의 증후군-특이적 영역들에서 빈에 대한 교정을 결정하는데 이용될 수 있다. 도 3a의 작동(317)에 적용된 포괄적 프로파일 교정과 같이, 시료 공정화, 서열화, 및/또는 상기 게놈 구조로부터 조직적 편향을 제거 또는 감소시키기 위하여 교정이 필요하지만, 이 경우에서는 게놈의 증후군 영역(들) 에서 도입된 편향에 집중한다. 이 교정은 테스트 시료들에 대하여 결정된 빈 커버리지를 교정하는데 이용된 하나 또는 그 이상의 "웨이브" 형태의 값을 취한다. 영향을 받지 않은 시료들의 클러스터로부터 획득된 커버리지 값에서 웨이브가 결정될 수 있다.

[0267] 도 5의 블록(01)에 나타낸 바와 같이, 상기 분석 방법은 영향을 받지 않은 시료들의 훈련용 세트 안에 각 빈에 대한 커버리지를 획득한다. 상기 훈련용 세트 데이터는 비-영향을 받은 시료들의 서열화로부터 획득된 각 빈에 대한 커버리지 값(가령, 주어진 100kb 빈 안에서 발견된 배제되지 않은 부위들의 수)을 포함한다. 상기 서열화

는 이 분야에서 테스트 시료들과 함께 이용될 수 있는 동일한 장치/프로토콜을 이용한다. 특정 구체예들에 있어서, 최소한 백개, 또는 최소한 300개 훈련용 세트 시료들이 이용된다. 특정 구체예들에 있어서, 커버리지 값은 복합 서열화 (가령, 12-plex 서열화)로부터 획득된다. 일부 구체예들에 있어서, 상기 참조 게놈은 각각 약 100kb 크기의 동일한 빈으로 분할된다. 본 명세서의 도처에서 설명된 다른 크기도 가능하다.

[0268] 분석의 초기 부분은 증후군 특이적 영역 또는 증후군 특이적 영역들의 집단에서 모든 필터안된 빈들이 분석될 때까지 한번에 하나의 빈에서 실행되며, 그리고 이렇게 실행됨으로써 이 분석은 동일한 일반적 방식에서 증후군 특이적 영역(들) 안에 빈과 동일한 일반적 방식에서와 같이 증후군 특이적 영역(들) 밖의 빈 집단을 식별한다. 블록(03)에서 설명된 바와 같이, 이 분석 방법은 증후군-특이적 영역에서 새로운 빈을 고려중인 빈으로 설정함으로써 빈 분석 일정을 제어한다. 특정 구체예들에 있어서, 증후군 특이적 영역들의 경계는 이후 본 명세서에서 설명된 바와 같이 콘센수스 영역으로부터 선택된다. 일부 구체예들에 있어서, 증후군 특이적 영역들의 경계는 이후 본 명세서에서 설명된 바와 같이 콘센수스 영역을 벗어난 조사 영역에서 선택된다.

[0269] 그 다음, 블록(05)에서 설명된 바와 같이, 이 분석 공정은 고려중인 빈과 모든 강건한 염색체에서 다른 이용가능한 빈 사이의 상관 거리를 결정한다. 한 가지 접근방식에 있어서, 이 공정은 모든 시료들로부터 정규화된 커버리지 값을 이용함으로써 상관 거리를 확인한다.

[0270] 블록(05)의 조작 (블록(07)의 조작과 함께)은 고려중인 빈과 동일한 방식으로 일반적으로 변화되는 빈을 식별한다. 환언하면, 상기 방법은 빈 커버리지를 분석하여 증후군 영역들 안에서 관찰된 조직적 변동성을 공유하는 빈을 식별해낸다. 강건한 염색체에 속하는 모든 상염색체 빈(가령, 염색체 13, 18, 그리고 21를 배제한 모든 인간 염색체의 빈) 사이의 쌍별(pair wise) 거리를 나타내는 거리 매트릭스를 구축할 수 있다. 물론, 상기 상염색체 빈은 상기에서 논의된 바와 같이 NESs를 만들기 위하여 필터링 후에 남아있는 것들로 제한될 수 있다.

[0271] 빈간의 상관관계는 다양한 기술들에 의해 확인될 수 있다. 본 명세서에서 논의된 접근방식에서, 상기 상관관계는 2개의 벡터, 고려중인 빈에 대하여 모든 훈련용 시료 세트에 걸친 커버리지 값에 의해 정의된 벡터와 증후군 영역 밖에 비교용 빈에 대하여 모든 훈련용 시료 세트에 걸쳐 커버리지 값을 포함하는 다른 벡터 사이의 거리로 산출된다. 이들 2개의 벡터 사이의 서리는 많은 기술, 이를 테면 벡터 사이의 각을 측정함으로써, 가령, 이들 벡터 사이 각의 코사인에 의해 산출될 수 있다.

[0272] 여전히 고려중인 빈에 집중하여, 이 분석 공정은 상관 거리에 기반을 두고 강건한 염색체(증후군 특이적 영역 밖)의 빈을 등급화하고, 그리고 고려중인 빈에 대하여 가장 이웃하는 빈 (SNB - 증후군 근린 빈) 세트의 구성원을 식별해낸다. 블록(07) 참고. 이 조작은 고려중인 빈과 최대 상관관계를 갖는 증후군 영역 밖의 빈을 수집한다. 일부 구체예들에 있어서, 관심대상의 증후군 특이적 영역에서 각 100kb 빈에 대하여 가장 크게 관련된 빈의 약 5%가 선택된다. 다른 구체예들에서, 상이한 양의 빈이 선택될 수 있다. 이들 빈은 SNB 세트에 함입용으로 떼어둔다.

[0273] 이것은 증후군 특이적 영역에서 고려중인 빈의 가공을 끝내며, 이로써 이 분석은 임의의 다른 증후군-특이적 빈이 여전히 고려되는 것으로 유지되는 지를 판단한다. 블록(09) 참고. 그럴 경우, 공정 제어는 블록(03)으로 복귀되며, 이때 증후군 특이적 영역에서 다음 빈이 고려중인 빈으로 설정된다. 이와 같은 방식으로, 증후군-특이적 영역들에 속하는 필터안된 각 빈에 대하여 이 분석이 반복된다.

[0274] 고려중인 각 증후군 특이적 빈에 대하여 식별되고, 보관된 이웃 빈들을 푸울링하고, 증후군 영역(들) 밖의 특유한 빈 수집으로 보관된다. 고려중인 각 증후군 빈에 대하여 실행된 공정 되풀이를 통하여 이웃 빈들이 수집된다. 모든 증후군 빈들이 분석된 후에 생성된 수집물이 SNB 세트이다. 이 조작은 블록(11)에서 설명된다. 이 점으로부터 앞으로, 이 분석은 SNB로부터 오직 커버리지 값만 고려되는 제2 단계를 착수한다. 일부 구체예들에 있어서, SNB 세트에는 약 6000개의 빈이 포함된다. 일부 구체예들에 있어서, 전체 게놈의 대표 부분집단이 포함된 SNB 세포를 보유하는 것이 유익하다. 예를 들면, 전체 게놈의 약 10%를 계수하는 SNB 세포를 보유하는 것이 유익하다.

[0275] 제2 단계를 시작으로, 조작(13)은 범위(dimension)로써 SNB 커버리지 값이 포함된 벡터를 이용하여 훈련용 세트의 모든 2개 구성원들 간에 상관거리를 결정한다. 이 공정은 블록(15)의 공정과 함께, SNB 세트에서 모든 빈에 걸쳐 동일한 방식으로 일반적으로 변화되는 훈련용 세트 구성원(영향을 받지 않은 시료들)을 식별해낸다. 이러한 훈련용 세트 구성원들은 그 다음 고려중 빈에 대하여 다중-웨이브 교정 공정에서 단일 웨이브인 교정을 식별해내는데 이용된다.

[0276] 한 가지 실행에 있어서, 한 가지 크기의 SNB 세트 빈과 다른 크기의 훈련용 시료들의 매트릭스를 형성함으로써,

훈련용 세트 구성원들이 식별된다. 이 매트릭스에서 각각의 위치는 훈련용 시료 j에서 증후군 근방 빈 i에 대한 정규화된 커버리지에 의해 확보된다. 이 결과는 효과적인 벡터 세트이며, 상이한 시료에 대하여 각각의 벡터이며, 그리고 각 벡터는 크기로써 SNB 빈을 갖는다. 특정 구체예들에 있어서, 상기에서 설명된 바와 같이, 최소한 포괄적 프로파일 교정 및 GC 교정이 포함된 완전한 상류 공정으로부터 커버리지 값이 획득된다.

[0277] 그 다음 작동(15)에서 이 분석 공정은 시료들의 최대 클러스트를 식별해내기 위하여 훈련용 세트 구성원들 사이의 상관 "거리"를 이용한다. 이 접근방식에서, 상기 훈련용 세트의 각 구성원은 SNB 빈에 대한 이의 커버리지 값으로 나타낸다. 다양한 클러스터링 과정들이 이용될 수 있다. 한 가지 실행에 있어서, 이 과정들은 코사인-각 거리에 근거하여 입력량으로 부동성(dissimilarity) 매트릭스 D를 취하는 HOPACH-PAM 알고리즘 운영한다 (상관 거리를 결정하는데 이용된 접근방식과 유사하지만, 단 벡터들은 상관관계 산출에서 평균 0을 가지도록 재설정되기 보다는 이들의 고유 평균 주변에 집중된다). M. van der Laan and K. Pollard, *A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap*, Journal of Statistical Planning and Inference, 117:275-303, 2003 참고, 이는 본 명세서에 그 전문이 참고로 포함된다.

[0278] 부동성 매트릭스는 작동(13)에서 생성된 매트릭스로부터 획득될 수 있으며, (훈련용 시료들의 수)/(훈련용 시료들의 수)로 나타내며, 이때 각 지표 쌍은 이들 지표들을 가진 2개의 훈련용 시료들에 대한 시료 벡터의 코사인-각 거리 (작동 13에서 획득된)를 나타낸다. 코사인-각 거리의 논의와 관련하여 가령, brenocon.com/blog/2012/03/cosine-similarity-pearson-correlation-and-ols-coefficients/를 참고하며, 이의 전문이 본 명세서에 참고로 포함된다. 영향을 받지 않은 시료들을 선택하기 위하여 이 공정이 이용하는 접근방식에 무관하게, 선택된 시료들은 정규화 커버리지에서 공통적인 조직적 변동성을 갖는다.

[0279] 그 다음 조작(15)에서 이 분석 공정은 시료들의 최대 클러스트를 식별해내기 위하여 훈련용 세트 구성원들 사이의 상관 "거리"를 이용한다. 이 접근방식에서, 상기 훈련용 세트의 각 구성원은 SNB 빈에 대한 이의 커버리지 값으로 나타낸다. 다양한 클러스터링 과정들이 이용될 수 있다. 한 가지 실행에 있어서, 이 과정들은 코사인-각 거리에 근거하여 입력량으로 부동성(dissimilarity) 매트릭스 D를 취하는 HOPACH-PAM 알고리즘 운영한다 (상관 거리를 결정하는데 이용된 접근방식과 유사하지만, 단 벡터들은 상관관계 산출에서 평균 0을 가지도록 재설정되기 보다는 이들의 고유 평균 주변에 집중된다). M. van der Laan and K. Pollard, *A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap*, Journal of Statistical Planning and Inference, 117:275-303, 2003 참고, 이는 본 명세서에 그 전문이 참고로 포함된다.

[0280] 이 시점에서 증후군 특이적 영역(들)의 배제되지 않은 각 빈과 증후군 특이적 영역(들) 밖 매우 관련된 빈에 대하여 교정 값이 창출되었다. 이들 값은 교정의 "제1 웨이브"로 나타낸다. 도시된 순서도에서 분석 공정은 이 교정 웨이브를 훈련용 세트 커버리지 값에 적용시킨다. 블록(19) 참고. 이 공정은 임의의 많은 이용가능한 기술을 이용하여 교정 웨이브를 적용시킬 수 있다. 한 가지 접근방식에 있어서, 시료들의 커버리지 값은 빈마다(bin-by-bin) 교정 값으로 나뉜다. 또다른 접근방식에 있어서, 교정 값은 빈마다(bin-by-bin) 이 시료 값으로부터 차감된다. 여전히 또다른 접근방식에 있어서, 빈 예상된 값은 빈마다(bin-by-bin) 이 시료 값으로부터 차감된다.

[0281] 한 실시예에서, 이 공정은 하나의 크기로 웨이브 교정 값이 포함된 점을 피팅시키고, 그리고 각 빈에 대하여 다른 크기의 대응하는 시료 값을 대응시킴으로써 빈 예상 값을 결정한다. 따라서 이점들의 수집이 생성되고, 각 시료에 피팅된다. 한 가지 접근방식에 있어서, 이 공정은 회귀 기술(regression technique)을 이용하여 이러한 점들을 연결하는 선을 만들고, 그 다음 이 선은 예를 들면 다음을 이용하여 시료 커버리지 값으로부터 예상된 커버리지 값을 획득하기 위한 함수로 이용한다

[0282] 예상된 빈 커버리지 값 = 선 기울기×(시료 빈 커버리지 값) + 선 절편

[0283] 생성된 예상된 빈 커버리지 값은 이 시료 빈 커버리지 값으로부터 차감될 수 있거나 또는 제수(divisor)로 이용될 수 있다.

[0284] 그 다음 이 공정은 임의의 더 많은 증후군 특이적 교정 웨이브가 적용되어야 하는 지를 판단한다. 블록(21) 참고. 일부 실행에 있어서, 이 공정은 오직 단일 교정 웨이브만을 생성한다. 좀더 전형적으로, 다중 교정 웨이브가 생성되는데, 각 교정 웨이브는 이 시료 분석을 더 개선시키는데; 가령, 조직적 변이를 감소시키고, 변이계수를 개선시킨다. 일부 구체예들에 있어서, 충분한 많은 수의 웨이브가 유도되는데, 가령 10-20개의 웨이브가 유도된다. 이 웨이브의 수행은 웨이브 수에 대한 함수로써 증후군 CVs를 산출함으로써 독립적 테스트 세트에서 평가된다. 웨이브의 바람직한 수는 테스트 데이터에서 CV를 통계학적으로 유의적으로 감소시키는 모든 웨이브

를 발견함으로써 결정된다.

[0285] 만약 이 공정에서 또다른 웨이브가 생성되어야 한다고 판단된다면, 작동(13)-(19)를 반복하고, 이때 교정된 훈련용 세트가 이용되어, 그 다음 교정 웨이브가 생성된다. 환언하면, 가장 최근에 생성된 교정 웨이브를 적용시킴으로써 변형된 훈련용 세트를 이용한다.

[0286] **CNV 탐지를 위한 2개 패스 공정**

[0287] 본 명세서에서 공개된 공정들은 전체 염색체 및 아염색체성 영역들의 CNV를 결정하는데 적절하다. 상기에서 설명된 바와 같이, 상대적으로 짧은 아염색체성 영역들이 유전적 증후군에 관련될 때, 증후군 특이적 프로파일 CNV와 무관한 증후군 특이적 프로파일은 제거되어 탐지 민감성이 개선될 수 있다.

[0288] 일부 구체예들에 있어서, 유사한 공정들이 전체 염색체 및 증후군 영역 분석에 적용될 수 있다. 도 6은 작업 흐름에서 2개의 중첩되는 패스, 일반적 CNV 탐지를 위한 패스 1, 그리고 증후군 특이적 CNV 탐지를 위한 패스 2의 순서도를 보여준다. 2개의 패스는 비교할만한 작동을 포함하여 조정된 커버리지 정보를 획득할 수 있는데, 이 정보에 근거하여 CNV가 결정된다. 증후군 분석에 특이적인 작동들중 하나는 커버리지 데이터를 조정할 때 증후군-특이적 웨이브의 제거다, 블록(620) 참고, 이는 본 명세서의 도처에서 더 상세하게 설명된다.

[0289] 이 공정의 초기 단일 패스 부분은 서열화 데이터를 수용함으로써 시작되며, 블록(602) 참고, 그리고 상기에서 설명된 바와 같이 전산(computing) 계수를 통하여 지속된다, 블록(612) 참고. 이 시점 후, 도시된 공정은 상기에서 설명된 바와 같이 2개의 패스로 갈린다. 이 공정의 초기 부분으로 되돌아가서, 작업흐름은 서열화 데이터를 서열 리드로 전환시킨다. 상기 서열화 데이터가 복합 서열화로부터 유도될 때, 상기 서열 리드는 또한 탈-복합화되어, 데이트의 원천이 확인된다. 블록(604) 참고. 그 다음 상기 서열 리드가 참조 서열에 정렬되고, 이때 정렬된 서열 리드는 서열 태그로 제시된다. 블록(606) 참고. 그 다음 서열 태그는 필터되어, 배제되지 않은 부위들 (NESs)이 확득되며, 이들은 분명하게 매핑된 비-중복된 서열 태그들이다. 서열 태그는 특이적 서열 길이, 이를 테면 1 kb, 100 kb, 또는 1 Mb의 빈으로 정리된다. 블록(610) 참고. 증후군 특이적 영역들의 분석 관련 일부 구체예들에 있어서, 빈은 100 kb이다. 일부 구체예들에 있어서, 도 3a, 블록(313)에서 설명된 것과 같은 방식으로 다수의 영향을 받지 않은 시료들로부터 획득된 서열 마스크를 이용하여 높은 변동성을 나타내는 빈들을 마스크할 수 있다. 그 다음 NESs에서 태그가 계수되어, CNV 분석을 위한 정규화된 그리고 조정된 커버리지가 제시된다. 블록(612) 참고.

[0290] 도시된 구체예에 있어서, 작동(604), (606), (610), 그리고 (612)는 일회 실행되며, 나머지 작동 대부분은 2회 실행되는데, 전체 염색체 CNV 분석 패스 (pass 1) 동안 1회 그리고 증후군-특이적 CNV 분석 패스 (pass 2) 동안 다시 1회 실행된다. 다른 구체예들에서, 2개의 패스에서 실행된 바와 같이 보여진 하나 또는 그 이상의 작동은 1회 실행되고, 그 결과는 두 공정에서 공유된다. 이러한 공유된 작동의 예로는 작동 (614), (616), 그리고 (618)이 포함된다.

[0291] 도시된 구체예들에 있어서, NESs의 획득된 계수는 가령, 빈의 NES 계수를 상기 게놈 또는 정규화 염색체의 세트의 전체 NESs로 나눔으로써, 정규화된다. 블록 (614) 참고. 그 다음, 일부 구체예들에서, 영향을 받지 않은 시료들이 포함된 훈련용 세트에 공통적인 변동은 제거되며, 이 변동은 관심 대상의 CNV와 무관하다. 도시된 구체예들에 있어서, 공통 변동은 상기에서 설명된 포괄적 웨이브 프로파일에 유사한 방식으로, 영향을 받지 않은 시료들로부터 획득된 포괄적 웨이브 프로파일로 나타낸다. 도 6에서 설명된 바와 같이 일부 구체예들에 있어서, 포괄적 웨이브 프로파일을 획득하기 위하여 이용된 영향을 받지 않은 시료들은 동일한 플로우 쎌 또는 공정 배취(batch)로부터 유입된 시료들을 포함한다. 블록(616) 참고. 플로우 쎌 특이적 포괄적 웨이브의 산출은 하기에서 더 설명된다. 도시된 구체예들에 있어서, 포괄적 웨이브 프로파일이 제거된 후, 시료-특이적 기반에 근거하여 GC 수준에 대하여 커버리지가 교정된다. 블록(616) 참고. GC 교정을 위한 일부 알고리즘은 도 3a, 블록(319)과 연관된 내용에서 더욱 상세하게 설명된다.

[0292] 도시된 구체예들에 있어서, 증후군 분석을 위한 패스 2에서 GC 교정된 커버리지는 증후군과 연관된 영역내 CNV와 무관한 증후군 특이적 변동에 대하여 조정된다. 이 증후군 특이적 변동은 본 명세서 도처에서 더 상세하게 설명되는 바와 같이 증후군 특이적 웨이브 프로파일로 획득될 수 있다. 블록(620) 참고. 이후, 전체 염색체 분석용 패스 1과 증후군 분석용 패스 2 모두에서, 데이터는 개별 시료에 특이적인 잡음에 대하여 추가 필터될 수 있으며, 가령, 다른 빈과 상당히 상이한 커버리지를 갖는 아웃라이어 빈의 데이터는 분석으로부터 제거될 수 있는데, 이러한 차이는 관심대상의 복제 수 변이에 기인되지 않을 수 있다. 블록(622) 참고. 이러한 시료-내 필터링 작동은 도 3a의 블록(321)에 대응한다.

[0293] 일부 구체예들에 있어서, 단일 시료 필터링 후, 전체 염색체 분석용 패스 1의 공정은 전체 염색체 수준에서 커버리지를 응집시킴으로써 진행되며, 염색체에 대한 전체 커버리지 값이 제시된다. 도시된 구체예들에 있어서, 전체 커버리지는 관심대상의 CNV에 무관한 변동 또는 잡음의 다중 원천에 대하여 정규화되고, 조정된다. 블록(624) 참고. 그 다음 염색체의 커버리지를 이용하여 상기에서 설명된 바와 같이 염색체 분량 및 정규화된 염색체 값 (NCV)이 산출된다. 그 다음 상기 NCV는 기준 점수와 비교되어, 전체 염색체가 관련된 CNV가 소명되어야 하는 지가 판단된다. 블록(626) 및 (632) 참고.

[0294] 도시된 구체예들에 있어서, 증후군 특이적 분석용 패스 2의 공정은 상기 테스트 시료에 대한 증후군 조사 영역에서 세그먼트를 확인하기 위한 분절화 일정을 실행하는 것에 관련된다. 블록(628) 참고. 상기 증후군 조사 영역은 관심 증후군이 관련된 신호들을 조사하는 최대 영역이다. 조사 영역에서 세그먼트는 CNV 소명을 위한 최대 신호를 보유하는 것으로 결정된다. 그 다음 이 증후군 조사 영역에서 세그먼트의 점수가 획득되며, 판단 기준과 비교될 수 있다. 블록(630) 참고. 판단 기준 이상의 점수는 증후군과 연루된 서열의 CNV가 존재함을 암시한다. 블록 (632) 참고.

[0295] 일부 구체예들에 있어서, 하기에서 더 설명되는 바와 같이 2-단계 서열화 접근방식이 일부 테스트 시료들에게 적용된다. 간략하게 설명하자면, 시료의 초기 점수화는 민감성을 증가시키기 위하여 기획된 상대적으로 낮은 제1 임계치와 비교되고, 그리고 만약 이 시료가 제1 임계치 이상으로 점수를 얻으면, 이 시료는 제1 서열화보다 더 심도 있는 제2 라운드의 서열화를 겪게 된다. 그 다음 시료는 재-가공되고, 상기에서 설명된 것과 유사한 작업흐름에서 분석된다. 그 다음 생성된 점수는 민감성을 개선시키기 위하여 기획된 상대적으로 높은, 제2 임계치와 비교된다. 일부 구체예들에 있어서, 제2 라운드의 서열화를 거치는 시료들은 제1 임계치 이상의 점수를 얻는 것들 중에서 상대적으로 낮은 점수를 얻고, 그렇게 함으로써 재서열화가 필요한 시료의 수가 감소된다.

[0296] **시료들 및 시료 가공**

[0297] *시료들*

[0298] CNV, 가령, 염색체 홀배수체, 부분적 홀배수체, 그리고 이와 유사한 것들을 결정하는데 이용되는 시료들은 임의의 세포, 조직, 또는 장기에서 취한 시료들을 포함할 수 있으며, 이때 하나 또는 그 이상의 관심 대상 서열들의 복제 수 변이가 결정된다. 바람직하게는, 이 시료들은 세포 에서 존재하는 핵산 및/또는 "무-세포" (가령, cfDNA) 핵산을 포함한다.

[0299] 일부 구체예들에 있어서 무-세포 핵산, 가령, 무-세포 DNA (cfDNA)을 획득하는 것이 유익하다. 무-세포 DNA가 포함된 무-세포 핵산은 혈장, 혈청, 그리고 소변이 포함되나, 이에 국한되지 않은 생물학적 시료로부터 당분야에 공지된 다양한 방법들에 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] 참고). 시료에서 세포로부터 무-세포 DNA를 분리시키기 위하여, 분획화, 원심분리 (가령, 밀도 구배 원심분리), DNA-특이적 침전, 또는 대량-처리 세포 소팅 및/또는 다른 분리 방법들이 포함되나, 이에 국한되지 않는 다양한 방법들이 이용될 수 있다. cfDNA의 수작업 및 자동화 분리를 위한 키트가 상업적으로 이용가능하다(Roche Diagnostics, Indianapolis, IN, Qiagen, Valencia, CA, Macherey-Nagel, Duren, DE). cfDNA가 포함된 생물학적 시료들은 염색체 홀배수체 및/또는 다양한 다형성(polymorphisms)을 탐지할 수 있는 서열화 분석에 의해 염색체 비정상의 존부, 가령, 삼염색체성 21의 존부를 판단하는 분석에 이용된다.

[0300] 다양한 구체예들에 있어서 이 시료에 존재하는 cfDNA는 사용전(가령, 서열화 라이브러리를 준비하기 전) 특이적으로 또는 비-특이적으로 농축될 수 있다. 시료 DNA의 비-특이적 농축(enrichment)은 시료의 게놈 DNA 단편의 전체 게놈 증폭을 지칭하는데, cfDNA 서열화 라이브러리를 준비하기 전, 시료 DNA의 수준을 증가시키는데 이용될 수 있다. 비-특이적 농축은 하나 이상의 게놈이 포함된 시료에서 2개의 게놈중 하나의 선택적 농축이 될 수 있다. 예를 들면, 비-특이적 농축은 모체 시료에서 태아 게놈에 대하여 선택적일 수 있으며, 시료에서 모체 DNA에 대하여 태아 DNA의 상대적 비율을 증가시키는 공지의 방법들에 의해 획득될 수 있다. 대안으로, 비-특이적 농축은 이 시료중에 존재하는 양쪽 게놈들의 비-선택적 증폭일 수 있다. 예를 들면, 비-특이적 증폭은 태아 게놈과 모체 게놈으로부터 DNA 혼합물이 포함된 시료에서 태아 및 모체 DNA의 증폭일 수 있다. 전체 게놈 증폭 방법들은 당업계에 공지되어 있다. 축퇴(Degenerate) 올리고뉴클레오티드-프라임된 PCR (DOP), 프라이머 연장 PCR 기술 (PEP) 및 다중 전위 증폭 (MDA)이 전체 게놈 증폭 방법들의 예시들이다. 일부 구체예들에 있어서, 상이한 게놈들로부터 cfDNA 혼합물이 포함된 시료는 이 혼합물중에 존재하는 게놈의 cfDNA에 대하여 농축되지 않

는다. 다른 구체예들에서, 상이한 게놈들로부터 cfDNA의 혼합물이 포함된 시료는 이 시료중에 존재하는 게놈들중 임의의 하나에 대하여 비-특이적으로 농축된다.

[0301] 본 명세서에서 설명된 방법들이 적용되는 핵산(들)이 포함된 시료는 전형적으로 상기에서 설명한 바와 같이 가령, 생물학적 시료 ("테스트 시료")를 포함한다. 일부 구체예들에 있어서, 하나 또는 그 이상의 CNVs에 대하여 스크리닝되는 핵산(들)은 잘 공지된 다수의 방법들중 임의의 것에 의해 정제 또는 단리된다.

[0302] 따라서, 특정 구체예들에 있어서 이 시료는 정제된 또는 단리된 폴리뉴클레오티드를 포함하거나 또는 이들로 구성되며, 또는 시료는 시료들, 이를 테면 조직 시료, 생물학적 유체 시료, 세포 시료, 그리고 이와 유사한 것들을 포함할 수 있다. 적절한 생물학적 유체 시료들은 혈액, 혈장, 혈청, 땀, 눈물, 가래, 소변, 가래, 귀 플로우, 림프, 침, 뇌척수액, 세척액(ravages), 골수 현탁액, 질 흐름, 경부-경유 세척액, 뇌 유체, 복수, 젖, 호흡기, 내장 및 비뇨생식기 관에서 분비물, 양수, 젖, 그리고 백혈구영양증(leukophoresis) 시료들이 포함되나, 이에 국한되지 않는다. 일부 구체예들에 있어서, 이 시료는 비-침습성 과정들, 가령, 혈액, 혈장, 혈청, 땀, 눈물, 가래, 소변, 가래, 귀 플로우, 침 또는 대변으로부터 용이하게 획득가능한 시료다. 특정 구체예들에 있어서, 이 시료는 말초 혈액 시료, 또는 말초 혈액 시료의 혈장 및/또는 혈청 분획이다. 다른 구체예들에서, 상기 생물학적 시료는 스왑 또는 도말표본, 생검 견본, 또는 세포 배양물이다. 또다른 구체예에서, 이 시료는 2개 또는 그 이상의 생물학적 시료들의 혼합물, 가령, 생물학적 시료는 생물학적 유체 시료, 조직 시료, 그리고 세포 배양물 시료중 2개 또는 그 이상을 포함할 수 있다. 본 명세서에서 이용된 바와 같이, 용어 "혈액", "혈장" 및 "혈청"은 분획 또는 이의 가공된 부분들을 명시적으로 포괄한다. 유사하게, 시료가 생검, 스왑(swab), 도말표본(smear), 등등인 경우, 상기 "시료"는 생검, 스왑, 도말표본, 등등으로부터 유도된 가공된 분획 또는 부분을 명시적으로 포괄한다.

[0303] 특정 구체예들에 있어서, 상이한 개체로부터의 시료들, 동일한 또는 상이한 개체의 상이한 발달 단계로부터의 시료들, 상이한 질환을 앓는 개체 (가령, 암 환자 또는 유전적 장애를 가진 것으로 의심되는 개체)의 시료들, 정상적 개체, 개체의 질환의 상이한 단계에서 획득된 시료들, 질환에 대하여 상이한 치료를 받은 개체로부터 획득된 시료들, 상이한 환경 인자에 놓인 개체 시료들, 병인에 대한 질병 소인을 가진 개체 시료들, 감염성 질환물질 (가령, HIV)에 노출된 개체의 시료들, 그리고 이와 유사한 것들이 포함되나, 이에 국한되지 않은 시료들이 원천으로부터 획득될 수 있다.

[0304] 한 가지 예시적인, 그러나 비-제한적 구체예에서, 이 시료는 임신한 암컷, 예를 들면 임신한 여성에서 획득된 모체 시료다. 이 경우에 있어서, 태아의 출생전 잠재적 염색체 비정상 진단을 하기 위하여 본 명세서에서 설명된 방법들을 이용하여 이 시료들이 분석될 수 있다. 모체 시료는 조직 시료, 생물학적 유체 시료, 또는 세포 시료일 수 있다. 생물학적 유체는 비-제한적 예로써, 혈액, 혈장, 혈청, 땀, 눈물, 가래, 소변, 가래, 귀 플로우, 림프, 침, 뇌척수액, 세척액(ravages), 골수 현탁액, 질 흐름, 경부경유 세척액, 뇌 유체, 복수, 젖, 호흡기, 내장 및 비뇨생식기 관에서 분비물, 양수, 젖, 그리고 백혈구영양증 시료들이 포함되나, 이에 국한되지 않는다.

[0305] 또다른 예시적인, 그러나 비-제한적 구체예에 있어서, 모체 시료는 2개 또는 그 이상의 생물학적 시료들의 혼합물, 가령, 생물학적 시료는 생물학적 유체 시료, 조직 시료, 그리고 세포 배양물 시료중 2개 또는 그 이상을 포함할 수 있다. 일부 구체예들에 있어서, 이 시료는 비-침습성 과정들, 가령, 혈액, 혈장, 혈청, 땀, 눈물, 젖, 가래, 소변, 가래, 귀 플로우, 침 및 대변으로부터 용이하게 획득가능한 시료다. 일부 구체예들에 있어서, 상기 생물학적 시료는 말초 혈액 시료, 및/또는 혈장 및 이의 혈청 분획이다. 다른 구체예들에서, 상기 생물학적 시료는 스왑 또는 도말표본, 생검 견본, 또는 세포 배양물의 시료다. 상기에서 공개된 바와 같이, 용어 "혈액", "혈장" 및 "혈청"은 분획 또는 이의 가공된 부분들을 명시적으로 포괄한다. 유사하게, 시료가 생검, 스왑(swab), 도말표본(smear), 등등인 경우, 상기 "시료"는 생검, 스왑, 도말표본, 등등으로부터 유도된 가공된 분획 또는 부분을 명시적으로 포괄한다.

[0306] 특정 구체예들에 있어서, 시료들은 시험관 배양된 조직, 세포, 또는 다른 폴리뉴클레오티드-포함 원천으로부터 또한 획득될 수 있다. 상이한 매체 및 상태 (가령, pH, 압력, 또는 온도)에서 유지된 배양물 (가령, 조직 또는 세포), 상이한 기간 동안 유지된 배양물 (가령, 조직 또는 세포), 상이한 인자들 또는 시약 (가령, 약물 후보물질, 또는 조절물질)로 처리된 배양물(가령, 조직 또는 세포), 또는 상이한 유형의 조직 및/또는 세포의 배양물이 포함되나, 이에 국한되지 않은 배양된 시료들이 원천으로부터 취해질 수 있다.

[0307] 생물학적 원천으로부터 핵산을 단리하는 방법들은 잘 공지되어 있고, 원천의 성질에 따라 달라질 수 있다. 당업자는 본 명세서에서 설명된 방법에서 요구되는 바와 같이 원천으로부터 용이하게 핵산(들)을 단리시킬 수 있

다. 일부 경우들에 있어서, 핵산 시료 에서 상기 핵산 분자들을 단편화시키는 것이 유익할 수 있다. 단편화는 무작위적이거나, 또는 예를 들면, 제한 엔도뉴클레아제 절단을 이용하여 획득되는 것과 같이 특이적일 수 있다. 무작위 단편화를 위한 방법들은 당업계에 공지되어 있으며, 그리고 예를 들면, 제한된 DNAse 절단, 알칼리 처리 및 물리적 전단을 포함한다. 한 구체예에서, 시료 핵산은 cfDNA와 같이 획득되며, 이는 단편화 대상이 아니다.

[0308] 다른 예시적인 구체예들, 이 시료 핵산(들)은 게놈 DNA로 획득되며, 대략적으로 300 또는 그 이상의, 대략적으로 400 또는 그 이상의, 또는 대략적으로 500 또는 그 이상의 염기쌍의 단편으로 단편화되며, 그리고 이 단편들에게 NGS 방법들이 바로 적용될 수 있다.

[0309] *서열화 라이브러리 준비*

[0310] 한 구체예에서, 본 명세서에서 설명된 방법들은 차세대 서열화 기술 (NGS)을 이용할 수 있는데, 이 기술은 게놈 분자들로써 개별적으로 서열화되거나 (가령, 단일 서열화) 또는 단일 서열화 운용에서 지수화된 게놈 분자들이 포함된 푸울된 시료들로 서열화(가령, 복합 서열화)되도록 한다. 이들 방법들은 최대 700백만개 DNA 서열 리드를 생산할 수 있다. 다양한 구체예들에 있어서 게놈 핵산의 서열, 및/또는 지수화된 게놈 핵산의 서열은 본 명세서에서 설명된 예를 들면, 차세대 서열화 기술 (NGS)을 이용하여 결정될 수 있다. 다양한 구체예들에 있어서, NGS를 이용하여 획득된 서열 데이터의 방대한 양의 분석은 본 명세서에서 설명된 바와 같이 하나 또는 그 이상의 프로세서를 이용하여 실행될 수 있다.

[0311] 다양한 구체예들에 있어서 이러한 서열화 기술의 사용은 서열화 라이브러리의 준비를 수반하지 않는다.

[0312] 그러나, 특정 구체예들에 있어서 본 명세서에서 고려되는 서열화 방법들은 서열화 라이브러리 준비와 관련되어 있다. 한 가지 예시적인 접근방식에 있어서, 서열화 라이브러리 준비는 서열화될 준비가 어뎁터-변형된 DNA 단편 (가령, 폴리뉴클레오티드)의 무작위 수집물의 생성과 관련된다. 폴리뉴클레오티드의 서열화 라이브러리는 DNA 또는 cDNA의 등가체, 유사체, 예를 들면, RNA 주형으로부터 역 전사효소의 작용에 의해 생성된 복제 DNA 또는 상보적 DNA 또는 cDNA가 포함된, DNA 또는 RNA로부터 준비된다. 폴리뉴클레오티드는 이중-가닥으로된 형태 (가령, dsDNA 이를 테면 게놈 DNA 단편, cDNA, PCR 증폭 제품, 그리고 이와 유사한 것들)로 유래될 수 있거나 또는, 특정 구체예들에 있어서, 폴리뉴클레오티드는 단일-가닥으로된 형태 (가령, ssDNA, RNA, 등등)로 유래될 수 있고, dsDNA 형태로 전환되었다. 설명하자면, 특정 구체예들에 있어서, 단일 가닥으로된 mRNA 분자들는 서열화 라이브러리 준비에 이용하는데 적절한 이중-가닥으로된 cDNAs로 복제될 수 있다. 일차 폴리뉴클레오티드 분자들의 정확한 서열은 일반적으로 라이브러리 준비 방법에 물질이 아니며, 그리고 공지의 것이거나 또는 미지의 것일 수 있다. 한 구체예에서, 상기 폴리뉴클레오티드 분자들은 DNA 분자들이다. 더욱 구체적으로, 특정 구체예들에 있어서, 상기 폴리뉴클레오티드 분자들은 유기체의 전체 유전적 보체 또는 실질적으로 유기체의 전체 유전적 보체를 나타내며, 그리고 인트론 서열과 엑손 서열 (코딩 서열), 뿐만 아니라 비-코딩 조절 서열들, 이를 테면 프로모터 및 인헨서 서열들을 전형적으로 포함하는 게놈 DNA 분자들 (가령, 세포의 DNA, 무 세포 DNA (cfDNA), 등등)이다. 특정 구체예들에 있어서, 일차 폴리뉴클레오티드 분자들은 인간 게놈 DNA 분자들, 가령, 임신한 개체의 말초 혈액 에서 존재하는 cfDNA 분자들을 포함한다.

[0313] 일부 NGS 서열화 플랫포옴을 위한 서열화 라이브러리의 준비는 특정 범위의 단편 크기들이 포함된 폴리뉴클레오티드를 사용하여 실행된다. 이러한 라이브러리의 준비는 바람직한 크기 범위의 폴리뉴클레오티드를 획득하기 위하여, 큰 폴리뉴클레오티드 (가령, 세포의 게놈 DNA)의 단편화가 일반적으로 수반된다.

[0314] 당분야의 숙련자들에게 공지된 수많은 방법들중 임의의 것에 의해 단편화가 실행될 수 있다. 예를 들면, 단편화는 분무(nebulization), 초음파분쇄(sonication) 및 유체전단(hydroshear)이 포함되나, 이에 국한되지 않은 기계적 수단에 의해 이루어질 수 있다. 그러나 기계적 단편화는 C-O, P-O 및 C-C 결합에서 DNA 기본골격을 전형적으로 절단하고, 이로써 절단된 C-O, P-O 및/ C-C 결합을 갖는 블런트(blunt) 및 3'와 5'-오버행 말단의 이질적 혼합물이 생성되는데 (가령, Alnemri and Liwack, J Biol. Chem 265:17323-17333 [1990]; Richards and Boyer, J Mol Biol 11:327-240 [1965]) 이들은 서열화용 DNA 준비에 필요한 후속적인 효소 반응, 가령, 서열화 어뎁터의 결찰을 위한 필요조건인 5'-인산염이 부족하기 때문에, 다시 쌍을 이루게해주어야 할 필요가 있을 것이다.

[0315] 대조적으로, cfDNA를 이용하면, 전형적으로 약 300개 미만의 염기쌍 단편으로 존재하고, 결과적으로 cfDNA 시료들을 이용하여 서열화 라이브러리를 만들에 단편화는 필요하지 않다.

[0316] 전형적으로, 폴리뉴클레오티드가 강제적으로 단편화되건 (가령, 시험관에서 단편화), 또는 자연적으로 단편으로

존재하건, 이들은 5'-인산염 및 3'-히드록실을 갖는 블런트-단부 DNA로 전환된다. 표준 프로토콜, 가령, 예를 들면, 본 명세서의 도처에서 설명된 Illumina 플랫포옴을 이용한 서열화 프로토콜은 사용자에게 시료 DNA를 단부-복구시키고, dA-꼬리붙이기(tailing)전 단부-복구된 생성물을 정제시키고, 그리고 라이브러리 준비의 어뎁터-결찰 단계에 앞서 dA-꼬리붙이기(tailing)된 생성물을 정제시키라고 지시한다.

[0317] 본 명세서에서 설명된 서열 라이브러리 준비를 위한 방법들의 다양한 구체예들은 NGS에 의해 서열화될 수 있는 변형된 DNA 산물을 획득하기 위한 표준 프로토콜에 의해 전형적으로 지시되는 하나 또는 그 이상의 단계를 실행할 필요성을 제거시킨다. 2012년 7월 20일자로 제출된 특허 출원 3/555,037에서 볼 수 있는 서열화 라이브러리 준비를 위한 방법들의 예로써 단축 방법 (ABB 방법), 1-단계 방법, 그리고 2-단계 방법이 있으며, 상기 특허 출원은 본 명세서에 그 전문이 참고로 포함된다.

[0318] *시료 진실성을 추적 및 입증하기 위한 표지 핵산*

[0319] 다양한 구체예들에 있어서 이 시료들의 진실성을 입증하고, 시료 추적은 시료 게놈 핵산, 가령, cfDNA의 서열화 혼합물, 그리고 가령, 공정에 앞서, 이 시료 안으로 도입되는 동반 표지 핵산 혼합물에 의해 실행될 수 있다.

[0320] 표지 핵산은 상기 테스트 시료 (가령, 생물학적 원천 시료)와 복합되고, 그리고 하나 또는 그 이상의 단계, 예를 들면, 상기 생물학적 원천 시료를 분획하고, 가령, 전체 혈액 시료로부터 기본적으로 무-세포 혈장 분획을 획득하고, 분획된 시료, 가령, 혈장, 또는 분획안된 생물학적 원천 시료, 가령, 조직 시료로부터 핵산을 정제하고, 그리고 서열화하는 하나 또는 그 이상의 단계들이 포함된 공정을 거치게된다. 일부 구체예들에 있어서, 서열화는 서열화 라이브러리를 준비하는 것을 포함한다. 상기 서열 또는 원천 시료와 복합되는 표지 분자들의 서료 조합은 원천 시료에 특유적인 것으로 선택된다. 일부 구체예들에 있어서, 시료 에서 특유한 표지 분자들은 모두 동일한 서열을 갖는다. 다른 구체예들에서, 시료에서 특유한 표지 분자들은 다수의 서열들, 가령, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 또는 그 이상의 상이한 서열들의 조합이다.

[0321] 한 구체예에서, 동일한 서열들을 가진 다수의 표지 핵산 분자들을 이용하여 시료의 진실성이 증명될 수 있다. 대안으로, 시료의 정체성(identity)은 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 서열들을 갖는 다수의 표지 핵산 분자들을 이용하여 증명될 수 있다. 다수의 생물학적 시료들, 가령, 2개 또는 그 이상의 생물학적 시료들의 진실성 입증은 표시되는 다수 테스트 시료 각각에 특유한 서열들을 갖는 표지 핵산으로 2개 또는 그 이상의 시료 각각이 표시되는 것을 요구한다. 예를 들면, 제1 시료는 서열 A를 갖는 표지 핵산으로 표시될 수 있고, 그리고 제2 시료는 서열 B를 갖는 표지 핵산으로 표시될 수 있다. 대안으로, 제1 시료는 서열 A를 모두 갖는 표지 핵산 분자들로 표시될 수 있고, 그리고 제2 시료는 서열 B 와 C의 혼합물로 표시될 수 있으며, 여기에서 서열 A, B 및 C는 상이한 서열들을 갖는 표지 분자들이다.

[0322] 라이브러리 준비 (만일 라이브러리가 준비된다면) 및 서열화에 앞서 일어나는 시료 준비의 임의의 단계에서 시료에 표지 핵산(들)이 추가될 수 있다. 한 구체예에서, 표지 분자들은 가공안된 원천 시료와 복합될 수 있다. 예를 들면, 상기 표지 핵산은 수집 튜브로 제공되고, 이는 혈액 시료를 수집하는데 이용된다. 대안으로, 상기 표지 핵산은 채혈 후 혈액 시료에 추가될 수 있다. 한 구체예에서, 상기 표지 핵산은 관에 추가되고, 이 관은 생물학적 유체 시료를 수집하는데 이용되며, 가령, 상기 표지 핵산(들)은 혈액 수집 튜브에 추가되며, 이 튜브는 혈액 시료를 수집하는데 이용된다. 또다른 구체예에서, 상기 표지 핵산(들)은 상기 생물학적 유체 시료의 분획에 추가된다. 예를 들면, 상기 표지 핵산은 혈장 및/또는 혈액 시료의 혈청 분획, 가령, 모체 혈장 시료에 추가된다. 여전히 또다른 구체예에서, 상기 표지 분자들은 정제된 시료, 가령, 생물학적 시료로부터 정제된 핵산 시료에 추가된다. 예를 들면, 상기 표지 핵산은 정제된 모체 및 태아 cfDNA의 시료에 추가된다. 유사하게, 상기 표지 핵산은 견본을 가공하기 전, 생검 견본에 추가될 수 있다. 일부 구체예들에 있어서, 상기 표지 핵산은 생물학적 시료의 세포로 표지 분자들을 운반하는 운반체와 복합될 수 있다. 세포-전달 운반체는 pH-민감성 및 양이온 리포좀을 포함한다.

[0323] 다양한 구체예들에 있어서, 상기 표지 분자들은 상기 생물학적 원천 시료의 게놈에 없는 서열인, 항게놈(antigenomic) 서열들을 보유한다. 예시적인 구체예에 있어서, 인간 생물학적 원천 시료의 진실성을 입증하는데 이용되는 표지 분자들은 인간 게놈에 없는 서열들을 보유한다. 대체 구체예에 있어서, 상기 표지 분자들은 원천 시료에 없고, 임의의 하나 또는 그 이상의 다른 공지의 게놈들에 없는 서열들을 보유한다. 예를 들면, 인간 생물학적 원천 시료의 진실성을 입증하는데 이용되는 표지 분자들은 인간 게놈에도 없고, 마우스 게놈에도

없는 서열을 보유한다. 이러한 대안은 2개 또는 그 이상의 게놈들이 포함된 테스트 시료의 진실성을 입증한다. 예를 들면, 병원체, 가령, 박테리아에 의해 영향을 받은 대상에서 획득된 인간 무-세포 DNA 시료의 진실성은 인간 게놈 및 영향을 미치는 박테리아의 게놈 모두에 존재하지 않는 서열을 가진 표지 분자들을 이용하여 입증될 수 있다. 다수의 병원균, 가령, 박테리아, 바이러스, 효모, 곰팡이, 프로토조아 등등 다수의 병원균의 게놈 서열은 World Wide Web at ncbi.nlm.nih.gov/genomes에서 공개적으로 이용가능하다. 또다른 구체예에서, 표지 분자들은 임의의 공지의 게놈에 없는 서열들을 보유한 핵산이다. 표지 분자들의 서열들은 알고리즘에 의해 무작위로 생성될 수 있다.

[0324] 다양한 구체예들에 있어서 상기 표지 분자들은 자연-발생적 데옥시리보핵산 (DNA), 리보핵산 또는 자연-발생적 DNA 또는 RNA와는 분자의 기본골격에 변화로 구별되는 펩티드 핵산 (PMA), 몰포리노 핵산, 잠금 핵산, 글리콜 핵산, 그리고 트레오스 핵산이 포함된 인공 핵산 유사체 (핵산 모방체) 또는 포스포디에스테르 기본골격을 보유하지 않는 DNA 모방체일 수 있다. 데옥시리보핵산은 자연-발생적 게놈들의 것일 수 있거나 또는 효소의 사용을 통하여 또는 고형상 화학 합성에 의해 실험실에서 만들어질 수 있다. 자연에서 볼 수 없는 DNA 모방체는 화학적 방법들을 이용하여 또한 만들 수 있다. 포스포디에스테르 연계는 대체되었지만, 데옥시리보스는 유지된 이용가능한 DNA 유도체들은 티오포름아세탈에 의해 형성된 기본 골격 또는 가르복사미드 연계를 보유한 DNA 모방체들을 포함하나, 이에 국한되지 않으며, 이들은 양호한 구조적 DNA 모방체로 밝혀졌다. 기타 DNA 모방체는 몰포리노 유도체들과 N-(2-아미노에틸) 글리신-기반의 슈도펩티드 기본골격이 포함된 펩티드 핵산 (PNA)을 포함한다 (Ann Rev Biophys Biomol Struct 24:167-183 [1995]). PNA는 상당히 구조적으로 우수한 DNA 모방체 (또는 리보핵산 [RNA]의)이며, 그리고 PNA 올리고머는 Watson-Crick 상보적 DNA 및 RNA (또는 PNA) 올리고머와 함께 매우 안정적인 듀플렉스 구조를 형성할 수 있고, 그리고 이들은 나선 침입에 의해 듀플렉스 DNA 안의 표적에 또한 결찰할 수 있다 (Mol Biotechnol 26:233-248 [2004]. 표지 분자로 이용될 수 있는 또다른 구조적으로 양호한 DNA 모방/유사체는 포스포로티오에이트 DNA인데, 이때 다리연결되지 않은 산소중 하나가 황으로 대체되어 있다. 이와 같은 변형은 5' 에서 3' 그리고 3' 에서 5' DNA POL 1 엑소뉴클레아제, 뉴클레아제 S1 및 P1, RNases, 혈청 뉴클레아제 그리고 뱀 독 포스포디에스테라제가 포함된 엔도- 및 엑소뉴클레아제2의 작용을 감소시킨다.

[0325] 상기 표지 분자들의 길이는 이 시료 핵산의 길이와 차이가 있거나 또는 차이가 없을 수 있는데, 가령, 상기 표지 분자들의 길이는 이 시료 게놈 분자들의 길이와 유사할 수도 있거나, 또는 이 시료 게놈 분자들의 길이보다 더 크거나 또는 더 작을 수 있다. 상기 표지 분자들의 길이는 상기 표지 분자를 구성하는 뉴클레오티드 또는 뉴클레오티드 유사체 염기 수에 의해 결정된다. 이 시료 게놈 분자들의 길이와 상이한 길이를 가진 표지 분자들은 당분야에 공지된 분리 방법을 이용하여 원천 핵산으로부터 구별될 수 있다. 예를 들면, 상기 표지 분자와 시료 핵산 분자의 길이에서 차이는 전기영동적 분리, 가령, 모세관 전기영동에 의해 측정될 수 있다. 크기 차등은 상기 표지와 시료 핵산을 정량화하고, 질을 평가하는데 유익할 수 있다. 바람직하게는, 상기 표지 핵산은 게놈 핵산보다 더 짧고, 그리고 이 시료의 게놈에 표지 핵산이 매핑되지 않도록 하는 충분한 길이를 가진다. 예를 들면, 30개 염기로써 인간 서열은 인간 게놈에 이를 특유하게 매핑될 필요가 있다. 따라서 특정 구체예들에 있어서, 인간 시료들의 서열화 생물학적 분석에 이용되는 표지 분자들은 길이가 최소한 30 bp이어야 한다.

[0326] 상기 표지 분자의 길이 선택은 원천 시료의 진실성을 입증하는데 이용되는 서열화 기술에 의해 주로 결정된다. 서열화되는 시료 게놈 핵산의 길이 또한 고려될 수 있다. 예를 들면, 일부 서열화 기술은 폴리뉴클레오티드의 클론 증폭을 이용하는데, 이 기술은 클론적으로 증폭되는 게놈 폴리뉴틀레오티드의 길이를 최소로 요구할 수 있다. 예를 들면, Illumina GAII 서열 분석기를 이용한 서열화는 클론적으로 증폭되고, 서열화될 수 있는 최소 200bp 내지 600bp 미만의 핵산을 제공하기 위하여, 110bp의 최소 길이를 갖는 폴리뉴클레오티드의 PCR (클러스터 증폭으로도 공지됨)을 결찰된 어뎁터에 연결시키는 시험관 클론 증폭을 포함한다. 일부 구체예들에 있어서, 상기 어뎁터-결찰된 표지 분자의 길이는 약 200bp 내지 약 600bp, 약 250bp 내지 550bp, 약 300bp 내지 500bp, 또는 약 350 내지 450이다. 다른 구체예들에서, 어뎁터-결찰된 표지 분자의 길이는 약 200bp이다. 예를 들면, 모체 시료 에서 존재하는 태아 cfDNA를 서열화할 때, 상기 표지 분자의 길이는 태아 cfDNA 분자들의 길이와 유사하도록 선택될 수 있다. 따라서, 한 구체예에서, 태아 염색체 홀배수체의 존재 또는 부재를 결정하기 위하여 모체 시료에서 cfDNA의 대량 병행 서열화를 포함하는 분석에 이용되는 표지 분자의 길이는 약 150 bp, 약 160bp, 170 bp, 약 180bp, 약 190bp 또는 약 200bp일 수 있으며; 바람직하게는, 상기 표지 분자는 약 170 pp이다. 기타 서열화 접근방식, 가령, SOLiD 서열화, Polony 서열화 및 454 서열화는 서열화를 위한 DNA 분자를 클론적으로 증폭시키기 위하여 에멀젼(emulsion) PCR을 이용하며, 그리고 각 기술은 증폭되는 분자들의 최소 길이와 최대 길이를 지시한다. 클론적으로 증폭된 핵산으로서 서열화되는 표지 분자들의 길이는 최대 약 600bp

일 수 있다. 일부 구체예들에 있어서, 서열화되는 상기 표지 분자의 길이는 600bp이상일 수 있다.

[0327] 분자들의 클론 증폭을 이용하지 않고, 그리고 대부분의 상황에서 주형 길이의 광범위한 범위에 걸쳐 핵산을 서열화시킬 수 있는 단일 분자 서열화 기술은 서열화되는 분자들이 임의의 특이적 길이이어야 함을 요구하지 않는다. 그러나, 단위량당 서열 생산량은 3' 말단 히드록실 기 수에 따라 달라지며, 그리고 따라서 서열화를 위하여 상대적으로 짧은 주형을 가지면 긴 주형을 가진 것보다 더 효과적이다. 1000 nt보다 긴 핵산으로 시작한다면, 100 내지 200 nt의 평균 길이로 핵산을 전단시키는 것을 일반적으로 권하며, 이로써 동일한 양의 핵산으로부터 더 많은 서열 정보가 생성될 수 있다. 따라서, 상기 표지 분자의 길이는 10개 염기에서 수천개 염기 범위가 될 수 있다. 단일 분자 서열화에 이용되는 표지 분자 길이는 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 이상의 길이일 수 있다.

[0328] 표지 분자를 위하여 선택된 길이는 서열화되는 게놈 핵산의 길이에 의해 또한 결정된다. 예를 들면, cfDNA는 세포의 게놈 DNA의 게놈 단편과 같이 인간 혈류 안을 순환한다. 임신한 여성의 혈장에서 발현되는 태아 cfDNA 분자들은 모체 cfDNA 분자들보다 일반적으로 더 짧다(Chan et al., Clin Chem 50:8892 [2004]). 순환하는 태아 DNA의 크기 분획화에 의해 순환하는 태아 DNA 단편의 평균 길이가 <300 bp임이 확인되며, 한편 모체 DNA는 약 0.5 내지 1 Kb인 것으로 추정된다 (Li et al., Clin Chem, 50: 1002-1011 [2004]). 이러한 발견은 NGS를 이용하여 태아 cfDNA가 >340bp인 경우는 드물다고 판단한 Fan et al.,의 것과 일관된다(Fan et al., Clin Chem 56:1279-1286 [2010]). 표준 실리카-기반의 방법에의해 소변으로부터 단리된 DNA는 2개의 분획, 탈락(shed) 세포로부터 기인된 고분자량 DNA와 트란스레날(transrenal) DNA (Tr-DNA)의데옥시리보핵산, DNA 저분자량 (150-250 염기쌍) 분획의 총 2개의 분획으로 구성된다 (Botezatu et al., Clin Chem. 46: 1078-1084, 2000; 그리고 Su et al., J Mol. Diagn. 6: 101-107, 2004). 체액으로부터 무-세포 핵산을 단리시키는 새로 개발된 기술을 트랜스레날 핵산을 분리하는데 적용시키면 소변에서 150 염기쌍보다 훨씬 더 짧은 DNA와 RNA 단편이 존재한다는 것이 밝혀졌다(U.S. 특허 출원 공개 No. 20080139801). cfDNA가 서열화되는 게놈 핵산인 구체예들에 있어서, 선택되는 표지 분자들은 최대 cfDNA의 길이가 될 수 있다. 예를 들면, 서열화되는 단일 핵산 분자들 또는 클론적으로 증폭된 핵산와 같이 모체 cfDNA 시료들 에서 이용되는 표지 분자의 길이는 약 100 bp 내지 600bp이다. 다른 구체예들에서, 이 시료 게놈 핵산은 더 큰 분자들의 단편이다. 예를 들면, 서열화되는 시료 게놈 핵산은 단편화된 세포의 DNA이다. 단편화된 세포 DNA가 서열화되는 구체예들에 있어서, 상기 표지 분자의 길이들은 최대 DNA 단편의 길이가 될 수 있다. 일부 구체예들에 있어서, 상기 표지 분자의 길이들은 최소한 적절한 참조 게놈에 대하여 독특하게 서열 리드을 매핑하는데 요구되는 최소 길이가 된다. 다른 구체예들에서, 상기 표지 분자의 길이는 이 시료 참조 게놈에 매핑되는 표지 분자를 배제시키는데 요구되는 최소 길이다.

[0329] 또한, 핵산 서열화에 의해 분석되지 않고, 그리고 서열화이외의 통상적인 생물학적 기술, 가령, 실시간 PCR에 의해 입증될 수 있는 시료들을 입증하는데 표지 분자들이 이용될 수 있다.

[0330] *시료 대조 (가령, 서열화 및/또는 분석에서 공정 양성 대조).*

[0331] 다양한 구체예들에 있어서 상기에서 설명된 바와 같이, 가령 이 시료 안으로 도입되는 표지 서열들은 서열화 그리고 후속적인 공정 및 분석의 정확성 및 효과를 입증하기 위한 양성 대조로 기능을 할 수 있다.

[0332] 따라서, 시료에서 DNA 서열화를 위한 공정-내 양성 대조 양성 대조조성물 및 방법들이 제시된다. 특정 구체예들에 있어서, 게놈 혼합물이 포함된 시료 에서 cfDNA 서열화를 위한 양성 대조가 제공된다. IPC는 상이한 시료들의 세트, 가령, 상이한 서열화 운용에서 상이한 시점에서 서열화되는 시료들로부터 획득된 서열 정보에서 기선 이동(baseline shifts)을 결부시키는데 이용될 수 있다. 따라서, 예를 들면, IPC는 모체 테스트 시료에서 획득된 서열 정보를 상이한 시점에서 서열화되는 검증된 시료들 세트에서 획득된 서열 정보에 결부시킬 수 있다.

[0333] 유사하게, 세그먼트 분석의 경우에 있어서, IPC는 개체로부터 특정 세그먼트(들)에 관한 서열 정보를 상이한 시점에서 서열화된 검증된 시료들 세트 (유사한 서열들의)로부터 획득된 서열에 결부시킬 수 있다. 특정 구체예들에 있어서 IPC는 개체로부터 특정 암-관련된 좌(loci)에 대하여 획득된 서열 정보를 검증된 시료들 세트 (가령, 공지의 증폭/결손, 그리고 이와 유사한 것들로부터)로부터 획득된 서열 정보에 결부시킬 수 있다.

[0334] 또한, IPCs는 상기 서열화 공정을 통하여 시료(들)을 추적하기 위한 표지로 이용될 수 있다. IPCs는 적절한 해독을 제공하고, 그리고 데이터의 의존성 및 정확성을 확보하기 위하여, 관심 염색체, 가령, 삼염색체성 21, 삼염색체성 13, 삼염색체성 18의 하나 또는 그 이상의 홀배수체에 대한 정량적인 양성 서열 분량 값, 가령, NCV를

또한 제공할 수 있다. 특정 구체예들에 있어서 태아가 남성인지를 판단하기 위하여 모체 시료에서 염색체 X와 Y의 분량을 제공하기 위한 남성 및 여성 게놈들의 핵산을 포함하는 IPCs가 탐지될 수 있다.

[0335] 공정-내 대조의 유형 및 수는 필요한 테스트의 유형 또는 성질에 의존적이다. 예를 들면, 염색체 홀배수체가 존재하는 지를 판단하기 위하여 게놈 혼합물이 포함된 시료로부터 DNA의 서열화를 요구하는 테스트의 경우에 있어서, 공정-내 대조는 테스트되는 동일한 염색체 홀배수체가 포함된 공지의 시료로부터 획득된 DNA를 포함할 수 있다. 일부 구체예들에 있어서, IPC는 관심 염색체의 홀배수체를 포함하는 것으로 알려진 시료로부터 DNA를 포함한다. 예를 들면, 모체 시료 에서 태아 삼염색체성, 가령, 삼염색체성 21의 존재 또는 부재를 결정하기 위한 테스트에서 IPC는 삼염색체성 21를 가진 개인으로부터 획득된 DNA를 포함한다. 일부 구체예들에 있어서, IPC는 상이한 홀배수체를 가진 2명 또는 그 이상의 개인으로부터 획득된 DNA 혼합물을 포함한다. 예를 들면, 삼염색체성 13, 삼염색체성 18, 삼염색체성 21, 그리고 일염색체성 X의 존재 또는 부재를 판단하기 위한 테스트의 경우, IPC는 테스트되는 삼체성중 하나를 가진 태아를 가진 임산부에서 획득된 DNA 시료 조합을 포함한다. 완전한 염색체 홀배수체에 추가하여, IPCs는 부분적 홀배수체의 존재 또는 부재를 판단하기 위한 테스트의 양성 대조를 제공하도록 만들어질 수 있다.

[0336] 단일 홀배수체를 탐지하기 위한 대조로 작용하는 IPC는 홀배수체 게놈의 기여자인 2명의 개체로부터 획득된 세포 게놈 DNA 혼합물을 이용하여 만들 수 있다. 예를 들면, 태아 삼염색체성, 가령, 삼염색체성 21를 결정하는 테스트의 대조로 만들어진 IPC는 삼염색체성 염색체를 휴대하는 남성 또는 여성 개체의 게놈 DNA를 삼염색체성 염색체를 휴대하지 않은 것으로 알려진 여성 개체의 게놈 DNA에 복합시킴으로써 창출될 수 있다. 게놈 DNA는 이들 두 개체의 세포로부터 추출될 수 있고, 그리고 모체 시료들 에서 순환하는 cfDNA 단편을 모의실험하기 위하여 약 100 - 400 bp, 약 150-350 bp, 또는 약 200-300 bp의 단편을 제공하기 위하여 전단될 수 있다. 홀배수체, 가령, 삼염색체성 21을 휴대하는 개체로부터 단편화된 DNA의 비율은 단편화된 DNA이 포함된 혼합물을 포함하는 IPC가 홀배수체를 휴대하는 개체로부터 DNA의 약 5%, 약 10%, 약 15%, 약 20%, 약 25%, 약 30%가 단편화된 DNA를 포함하도록 하기 위하여, 모체 시료 에서 발견되는 순환하는 태아 cfDNA의 비율을 흉내내도록 선택된다. 상기 IPC는 상이한 홀배수체를 각각 휴대하는 상이한 개체들의 DNA를 포함할 수 있다. 예를 들면, 상기 IPC는 영향을 받지 않은 여성 DNA를 약 80%로 포함할 수 있고, 그리고 나머지 20%는 삼염색체성 염색체 21, 삼염색체성 염색체 13, 그리고 삼염색체성 염색체 18를 각각 휴대하는 3명의 상이한 개체의 DNA일 수 있다. 서열화를 위하여 단편화된 DNA 혼합물이 준비된다. 단편화된 DNA의 혼합물의 가공은 서열화 라이브러리를 준비하는 것을 포함할 수 있으며, 이 라이브러리는 단일 또는 복합 방식으로 임의의 대량 병행 방법들을 이용하여 서열화될 수 있다. 게놈 IPC의 원액은 보관될 수 있고, 그리고 다중 진단학적 테스트에 이용될 수 있다.

[0337] 대안으로 상기 IPC는 공지의 염색체 홀배수체를 가진 태아를 출산한 것으로 알려진 여성으로부터 획득된 cfDNA를 이용하여 창출될 수 있다. 예를 들면, cfDNA는 삼염색체성 21를 가진 태아를 임신한 여성으로부터 획득될 수 있다. cfDNA는 모체 시료로부터 추출되고, 그리고 박테리아 벡터로 클로닝되고, 박테리아에서 성장되어, 상기 IPC의 현행 원천이 제공된다. 제한효소를 이용하여 박테리아 벡터로부터 상기 DNA가 추출될 수 있다. 대안으로, 상기 클론된 cfDNA는 가령, PCR에 의해 증폭될 수 있다. 상기 IPC DNA는 염색체 홀배수체의 존재 또는 부재에 대하여 분석되는 테스트 시료에서 cfDNA와 같은 동일한 운용에서 서열화를 위하여 가공될 수 있다.

[0338] 삼체성에 있어서 IPCs의 창출은 상기에서 설명되지만, 다른 부분적 홀배수체, 예를 들면, 다양한 세그먼트 증폭 및/또는 결손이 반영된 IPCs가 창출될 수 있음을 인지할 수 있을 것이다. 따라서, 예를 들면, 이때 다양한 암들은 특정 증폭 (가령, 유방 암은 20Q13와 연관됨)과 연관되었음이 공지되어 있고, IPCs는 이들 공지의 증폭이 포함되도록 창출될 수 있다.

[0339] **서열화 방법들**

[0340] 상기에서 표시된 바와 같이, 준비된 시료들 (가령, 서열화 라이브러리)은 복제 수 변이(들)를 식별해내는 과정의 일부분으로써 서열화된다. 다수의 서열화 기술중 임의의 것이 이용될 수 있다.

[0341] 일부 서열화 기술은 상업적으로 이용가능한데, 이를 테면 Affymetrix Inc의 혼성화 플랫포옴에 의한 서열화 (Sunnyvale, CA) 및 454 Life Sciences (Bradford, CT), Illumina/Solexa (Hayward, CA) 및 Helicos Biosciences (Cambridge, MA)의 상기 합성 플랫포옴에 의한 서열화, 그리고 Applied Biosystems (Foster City, CA)의 결찰 플랫포옴에 의한 서열화가 있고, 하기에서 설명된다. Helicos Biosciences의 합성에 의한 서열화를 이용하여 실행되는 단일 분자 서열화에 추가하여, 다른 단일 분자 서열화 기술은 Pacific Biosciences의 SMRT™ 기술, ION TORRENTTM 기술, 그리고 예를 들면, Oxford Nanopore Technologies에 의해 개발된 나노포어 서열화를 포함하나 이에 국한되지 않는다.

[0342] 자동화된 Sanger 방법은 '제1 세대' 기술로 간주되지만, 자동화된 Sanger 서열화가 포함된 Sanger 서열화 또한 본 명세서에서 설명된 방법들에서 이용될 수 있다. 추가적으로 적절한 서열화 방법들은 핵산 영상화 기술, 가령, 원자력 현미경 (AFM) 또는 투과 전자 현미경 (TEM)이 포함되나, 이에 국한되지 않는다. 예시적인 서열화 기술은 하기에서 더 상세하게 설명된다.

[0343] 한 가지 예시적인, 그러나 비-제한적인 구체예에서, 본 명세서에서 설명된 방법들은 테스트 시료, 가령, 모체 시료에서 cfDNA, Helicos True Single Molecule Sequencing (tSMS) 기술 (가령 Harris T.D. et al., Science 320:106-109 [2008]에서 설명된 바와 같이)의 단일 분자 서열화 기술을 이용하여, 암, 그리고 이와 유사한 것들에 대하여 스크리닝되는 개체에서 cfDNA 또는 세포 DNA에서 핵산에 대한 서열 정보를 획득하는 것을 포함한다. tSMS 기술에 있어서, DNA 시료는 대략적으로 100 내지 200개 뉴클레오티드의 가닥으로 절단되며, 그리고 polyA 서열은 각 DNA 가닥의 3' 말단에 추가된다. 각 가닥은 형광적으로 라벨된 아데노신 뉴클레오티드를 추가함으로써 라벨된다. 그 다음 DNA 가닥들은 플로우 쎌에 혼성화되는데, 이 쎌은 플로우 쎌 표면에 고정된 수백만개의 올리고-T 포획 부위을 포함하고 있다. 특정 구체예들에 있어서, 상기 주형은 약 100 백만개 주형/cm2의 밀도일 수 있다. 그 다음 플로우 쎌을 기구, 가령, HeliScope™ 서열화기에 얹고, 그리고 레이져로 플로우 쎌의 표면을 비추면, 각 주형의 위치가 드러난다. CCD 카메라는 플로우 쎌 표면 상에 주형의 위치 지도를 만들 수 있다. 그 다음 상기 주형 형광 라벨은 절단되고, 씻겨나간다. 상기 서열화 반응은 DNA 중합효소와 형광적으로 라벨된 뉴클레오티드를 도입시킴으로써 시작된다. 올리고-T 핵산은 프라이머로 작용한다. 중합효소는 라벨된 뉴클레오티드를 주형 지향된 방식으로 상기 프라이머에 혼입시킨다. 중합효소와 혼입안된 뉴클레오티드는 제거된다. 형광적으로 라벨된 뉴클레오티드의 혼입을 지시하는 주형은 플로우 쎌 표면의 영상화에 의해 포착된다. 영상화 후, 절단 단계에서 형광 라벨이 제거되고, 그리고 바람직한 리드 길이가 획득될 때까지 형광적으로 라벨된 다른 뉴클레오티드를 가지고 상기 공정은 반복된다. 각 뉴클레오티드 추가 단계에서 서열 정보가 수집된다. 단일 분자 서열화 기술에 의한 전체 게놈 서열화는 상기 서열화 라이브러리 준비에서 PCR-기반의 증폭을 배제하거나 또는 전형적으로 회피하고, 그리고 상기 방법들은 시료의 복사체 측정보다는 시료의 직접적인 측정을 허용한다.

[0344] 또다른 예시적인, 그러나 비-제한적 구체예에 있어서, 본 명세서에서 설명된 방법들은 상기 테스트 시료, 가령, 모체 테스트 시료 에서 cfDNA, 암, 그리고 이와 유사한 것들에 대하여 스크리닝되는 개체에서 cfDNA 또는 세포 DNA의 핵산에 대한 서열 정보를 454 서열화 (Roche) (가령 Margulies, M. et al. Nature 437:376-380 [2005]에서 설명된)를 이용하여 획득하는 것을 포함한다. 454 서열화는 전형적으로 2 단계를 수반한다. 제1 단계에서, DNA는 대략적으로 300-800개 염기쌍의 단편으로 전단되고, 그리고 이 단편들은 블런트-말단이다. 그 다음 올리고뉴클레오티드 어뎁터는 단편의 말단에 결합된다. 어뎁터는 단편의 증폭 및 서열화를 위한 프라이머로 작용한다. 상기 단편은 가령, 5'-비오틴 태그가 포함된 어뎁터 B를 이용하여 DNA 포획 비드, 가령, 스트렙타아비딘-피복된 비드에 부착될 수 있다. 비드에 부착된 단편은 오일-물 에멀젼의 방울 에서서 PCR 증폭된다. 이 결과는 각 비드 상에 클론적으로 증폭된 DNA 단편의 다중 복사체다. 제2 단계에서, 상기 비드는 웰 에서 포획된다 (가령, 피코리터-크기의 웰). 파이로서열화가 각 DNA 단편 상에서 병행 실행된다. 하나 또는 그 이상의 뉴클레오티드의 추가로 서열화 기구에서 CCD 카메라에 의해 기록되는 광 신호가 발생된다. 상기 신호 강도는 혼입된 뉴클레오티드의 수에 비례한다. 파이로서열화는 뉴클레오티드 추가 시 방출되는 피로포스페이트(PPi)를 이용한다. PPi는 아데노신 5' 포스포술페이트 존재하에 ATP 술푸릴라제에 의해 ATP로 전환된다. 루시퍼라제는 ATP를 이용하여 루시페린을 옥시루시페린으로 전환시키고, 그리고 이 반응은 측정되고, 분석되는 빛을 생성한다.

[0345] 또다른 예시적인, 그러나 비-제한적 구체예에 있어서, 본 명세서에서 설명된 방법들은 상기 테스트 시료, 가령, 모체 테스트 시료 에서 cfDNA, 암, 그리고 이와 유사한 것들에 대하여 스크리닝되는 개체에서 cfDNA 또는 세포 DNA의 핵산에 대한 서열 정보를 SOLiD™ 기술 (Applied Biosystems)를 이용하여 획득하는 것을 포함한다. SOLiD™ 결찰에 의한 서열화에서, 게놈 DNA는 단편으로 전단되며, 그리고 어뎁터들은 상기 단편의 5' 및 3' 단부에 부착되어 단편 라이브러리가 생성된다. 대안으로, 어뎁터들을 상기 단편의 5' 및 3' 단부에 결찰시키고, 상기 단편을 원형화시키고, 원형화된 단편을 절단하여 내부 어뎁터, 그리고 어뎁터들을 생성된 단편의 5' 및 3' 단부에 부착시켜 짝으로-쌍(mate-paired)을 이룬 라이브러리가 생성됨으로써, 내부 어뎁터들이 도입될 수 있다. 그 다음, 비드, 프라이머, 주형, 그리고 PCR 성분들이 포함된 마이크로반응기 에서 클론 비드 집단이 준비된다. PCR 후, 상기 주형이 변성되고, 비드는 농축되어, 연장된 주형을 가진 비드를 분리시킨다. 선택된 비드 상에 있는 주형은 유리 슬라이드 상에 결합이 허용되도록 3' 변형을 겪게 된다. 특이적 형광단에 의해 식별되는 중심 결정된 염기(또는 염기쌍)을 가진 부분적으로 무작위 올리고뉴클레오티드의 순차적 혼성화 및 결

찰에 의해 서열이 결정될 수 있다. 색깔이 기록된 후, 결찰된 올리고뉴클레오티드는 절단되고, 제거되며, 이 공정은 반복된다.

[0346] 또다른 예시적인, 그러나 비-제한적 구체예에 있어서, 본 명세서에서 설명된 방법들은 상기 테스트 시료, 가령, 모체 테스트 시료 에서 cfDNA, 암, 그리고 이와 유사한 것들에 대하여 스크리닝되는 개체에서 cfDNA 또는 세포 DNA의 핵산에 대한 서열 정보를 Pacific Biosciences의 단일 분자, 실시간 (SMRT™) 서열화 기술을 이용하여 획득하는 것을 포함한다. SMRT 서열화에 있어서, 염료-라벨된 뉴클레오티드의 연속적인 혼입은 DNA 합성 동안 영상화된다. 단일 DNA 중합효소 분자들은 개별 제로-방식 웨이브길이 검출기 (ZMW 검출기)의 바닥에 부착되어, 포스포연계된 뉴클레오티드가 커지는 프라이머 가닥 안으로 혼입되는 동안 서열 정보가 획득된다. ZMW 검출기는 ZMW 밖으로 신속하게 확산되는 (가령, 마이크로세컨드에서) 형광 뉴클레오티드 배경에 대하여 DNA 중합효소에 의한 단일 뉴클레오티드의 혼입이 관찰되는 유폐(confinement) 구조를 포함한다. 이 검출기는 뉴클레오티드를 커져가는 가닥으로 혼입시키기 위하여 전형적으로 7백만개를 취한다. 이 시기 동안, 상기 형광 라벨이 여기되어(excited) 형광 신호를 생성시키고, 그리고 형광 태그는 절단된다. 상기 염료의 대응하는 형광 측정은 어느 염기가 혼입되었는 지를 나타낸다. 이 공정이 반복되어 서열이 제공된다.

[0347] 또다른 예시적인, 그러나 비-제한적 구체예에 있어서, 본 명세서에서 설명된 방법들은 상기 테스트 시료, 가령, 모체 테스트 시료 에서 cfDNA, 암, 그리고 이와 유사한 것들에 대하여 스크리닝되는 개체에서 cfDNA 또는 세포 DNA의 핵산에 대한 서열 정보를 나노포어 서열화 (가령 Soni GV and Meller A. Clin Chem 53: 1996-2001 [2007]에서 설명된 바와 같은)를 이용하여 획득하는 것을 포함한다. 나노포어 서열화 DNA 분석 기술은 예를 들면, Oxford Nanopore Technologies (Oxford, United Kingdom), Sequenom, NABsys, 그리고 이와 유사한 회사들이 포함된 다수의 회사들에 의해 개발된다. 나노포어 서열화는 DNA의 단일 분자가 이 분자가 통과하는 나노포어를 통하여 직접적으로 서열화되는, 단일-분자 서열화 기술이다. 나노포어는 전형적으로 직경이 1 나노미터인 작은 구멍이다. 전도성(conducting) 유체 에서 나노포어를 침수시키고, 나노포어에 걸쳐 전위(전압)를 제공하면 나노포어를 통한 이온들의 전도로 인하여 약간의 전류가 생성된다. 흐르는 전류의 양은 나노포어의 크기 및 모양에 민감하다. DNA 분자가 나노포어를 통과할 때, DNA 분자상의 각 뉴클레오티드는 나노포어를 상이한 정도로 방해하여, 나노포어를 통한 전류 크기가 상이한 정도로 변화된다. 따라서, 상기 DNA 분자가 나노포어를 통과할 때 전류에서 이러한 변화는 이 DNA 서열의 리드를 제공한다.

[0348] 또다른 예시적인, 그러나 비-제한적 구체예에 있어서, 본 명세서에서 설명된 방법들은 상기 테스트 시료, 가령, 모체 테스트 시료 에서 cfDNA, 암, 그리고 이와 유사한 것들에 대하여 스크리닝되는 개체에서 cfDNA 또는 세포 DNA의 핵산에 대한 서열 정보를 (가령, U.S. 특허 출원 공개 No. 2009/0026082에서 설명된 바와 같은)를 이용하여 획득하는 것을 포함한다. 이 기술의 한 실시예에서, DNA 분자들을 반응 챔버 에서 배치시킬 수 있고, 그리고 주형 분자들은 중합효소에 결합된 서열화 프라이머에 혼성화될 수 있다. 하나 또는 그 이상의 삼인산염이 상기 서열화 프라이머의 3' 말단에서 새로운 핵산 가닥에 혼입되는 것은 chemFET에 의한 전류 변화로 포작될 수 있다. 어레이는 다중 chemFET 센서를 보유할 수 있다. 또다른 실시예에 있어서, 단일 핵산은 비드에 부착될 수 있고, 그리고 상기 핵산은 비드 상에서 증폭될 수 있고, 그리고 개별 비드는 chemFET 어레이 상의 개별 반응 챔버로 전달될 수 있고, 각 챔버는 chemFET 센서를 보유하며, 그리고 상기 핵산이 서열화될 수 있다.

[0349] 또다른 구체예에서, 상기 본 방법은 투과 전자 현미경 (TEM)을 이용하는 Halcyon Molecular 기술을 이용하여 테스트 시료, 가령, 모체 테스트 시료 에서 cfDNA에 대한 서열 정보를 획득하는 것을 포함한다. Individual Molecule Placement Rapid Nano Transfer (IMPRNT)이라고 불리는 상기 방법은 중원자(heavy atom) 표지들로 선택적으로 라벨된 고분자량 (150kb 또는 그 이상) DNA의 단일 원자 해리 투과 전자현미경 영상을 이용하고, 그리고 일관된 염기간 거리를 가진 초-조밀 (가닥간에 3nm) 병행 어레이의 초박 필름 상에 이들을 배열시키는 것을 포함한다. 상기 전자 현미경은 중원자 표지들의 위치를 결정하고, DNA로부터 염기 서열 정보를 추출하기 위하여 필름상에서 이들 분자들을 영상화하는데 이용된다. 상기 방법은 PCT 특허 공개 WO 2009/046445에서 더 설명된다. 상기 방법은 10분 이내에 완전한 인간 게놈들을 서열화시킨다.

[0350] 또다른 구체예에서, DNA 서열화 기술은 Ion Torrent 서열화인데, 이것은 반도체 기술을 단순 서열화 화학과 쌍을 이뤄 화학적으로 인코드된 정보 (A, C, G, T)를 반도체 칩 상에 디지털 정보 (0, 1)로 바로 해독시킨다. 실제, 뉴클레오티드가 중합효소에 의해 DNA 가닥으로 혼입될 때, 부산물로써 수소 이온이 방출된다. Ion Torrent는 대량 병행 방식으로 이러한 생화학적 공정을 실행하기 위하여 마이크로-기계화된 웰의 고-밀도 어레이를 이용한다. 각 웰은 상이한 DNA 분자를 잡고 있다. 이들 웰 아래 이온-민감 층이 있고, 이 아래에 이온 센서가 있다. 뉴클레오티드, 예를 들면 C가 DNA 주형에 부가되고, 그 다음 DNA 가닥 안으로 혼힙될 때, 수소 이온이 방출될 것이다. 이 이온으로부터 변화는 용액의 pH를 변화시킬 것이며, 이는 Ion Torrent의 이온 센서

에 의해 감지될 수 있다. 상기 서열화기-기본적으로 세계 최소 고형-상대 pH 측정기-는 이 염기를 소명하고, 화학적 정보가 바로 디지털 정보로 간다. 그 다음 상기 이온 개별적 게놈 기계 (PGM™) 서열화기는 이 칩에 차례로 뉴클레오티드를 보낸다. 만약 이 칩에 보내진 그 다음 뉴클레오티드가 정합이 아니라면, 전압 변화가 기록되지 않을 것이고, 소명되는 염기는 없을 것이다. DNA 가닥 상에 2개의 동일한 염기가 있다면, 전압은 배가 될 것이고, 그리고 칩은 2개의 동일한 염기가 소명된다고 기록할 것이다. 직접적인 탐지는 뉴클레오티드 혼입의 기록을 몇초에서 허용한다.

[0351] 또다른 구체예에서, 상기 본 방법은 혼성화에 의한 서열화를 이용하여 테스트 시료, 가령, 모체 테스트 시료 에서 cfDNA에 대한 서열 정보를 획득하는 것을 포함한다. 혼성화에 의한 서열화는 다수의 폴리뉴클레오티드 서열들에 다수의 폴리뉴클레오티드 프로브를 접촉시키는 것을 포함하고, 여기에서 다수의 폴리뉴클레오티드 프로브 각각은 임의선택적으로 기질에 묶여있을 수 있다. 상기 기질은 공지의 뉴클레오티드 서열들의 어레이를 포함하는 편평한 표면일 수 있다. 이 어레이에 혼성화되는 패턴을 이용하여 시료에 존재하는 상기 폴리뉴클레오티드 서열들이 결정될 수 있다. 다른 구체예들에서, 각 프로브는 비드, 가령, 자석 비드 또는 이와 유사한 것들에 연결된다. 상기 비드에 대한 혼상화가 결정되고, 이 시료 에서 다수의 폴리뉴클레오티드 서열들을 식별하는데 이용된다.

[0352] 또다른 구체예에서, 본 방법은 상기 테스트 시료 에서 핵산, 가령, 모체 테스트 시료 에서 cfDNA에 대한 서열 정보를 Illumina의 합성에 의한 서열화 및 가역적 종료물질-기반의 서열화 화학 (가령 Bentley et al., Nature 6:53-59 [2009]에서 설명된 바와 같이)을 이용하여 수백만개의 DNA 단편의 대량 병행 서열화에 의해 획득하는 것을 포함한다. 주형 DNA는 게놈 DNA, 가령, cfDNA일 수 있다. 일부 구체예들에 있어서, 단리된 세포의 게놈 DNA는 주형으로 이용하고, 그리고 이는 몇백개의 염기쌍 길이로 단편화된다. 다른 구체예들에서, cfDNA는 주형으로 이용되고, 그리고 cfDNA가 짧은 단편으로 존재할 때 단편화는 요구되지 않는다. 예를 들면 태아 cfDNA는 길이가 대략적으로 170 염기쌍 (bp)인 단편으로 혈류에서 순환할 때(Fan et al., Clin Chem 56:1279-1286 [2010]), 그리고 서열화 전, DNA의 단편화는 요구되지 않는다. Illumina의 서열화 기술은 올리고뉴클레오티드 엥커가 결합되는 평면의 광학적으로 투명한 표면에 단편화된 게놈 DNA의 부착을 필요로 한다. 주형 DNA는 말단-복구되어 5'-포스포릴화된 블런트 말단이 생성되고, Klenow 단편의 중합효소 활성을 이용하여 단일 A 염기를 이 블런트 포스포릴화된 DNA 단편의 3' 말단에 추가한다. 이러한 추가는 올리고뉴클레오티드 어뎁터들에 결찰용 DNA 단편을 준비하고, 이들 단편은 결찰 효과를 증가시키기 위하여 이들의 3' 말단에 단일 T 염기의 오버행을 보유한다 상기 어뎁터 올리고뉴클레오티드는 플로우-쎌 엥커에 상보적이다. 제한-희석 조건하에서 어뎁터-변형된, 단일-가닥으로된 주형 DNA는 플로우 쎌에 추가되고, 엥커에 혼성화됨으로써 고정된다. 부착된 DNA 단편은 연장되고, 다리 증폭되어 수억개의 클러스트를 가진 초 고 밀도 서열화 플로우 쎌이 창출되며, 이 쎌은 각각 동일한 주형의 ~1,000개 복사체를 포함하고 있다. 한 구체예에서, 무작위로 단편화된 게놈 DNA, 가령, cfDNA는 클러스터 증폭되기에 앞서 PCR에 의해 증폭된다. 대안으로, 증폭-없는 게놈 라이브러리 준비가 이용되는데, 그리고 상기 무작위로 단편화된 게놈 DNA, 가령, cfDNA는 클러스터 증폭 만을 이용하여 농축된다 (Kozarewa et al., Nature Methods 6:291-295 [2009]). 상기 주형은 제거가능한 형광 염료와 함께 가역적 종료물질들을 이용하는 합성 기술에 의한 강건한 4색 DNA 서열화를 이용하여 서열화된다. 고감도 형광 탐지는 레이저 여기(excitation) 및 전체 내부 반사 광학을 이용하여 이루어진다. 짧은 서열 리드는 약 20-40 bp, 가령, 36 bp의 짧은 서열 리드는 반복-마스크된 참조 게놈에 대응하여 정렬되고, 상기 참조 게놈에 대한 짧은 서열 리드의 특유 매핑은 특별히 개발된 데이터 분석 파이프라인 소프트웨어를 이용하여 식별된다. 비-반복-마스크된 참조 게놈들이 또한 이용될 수 있다. 반복-마스크된 또는 비-반복-마스크된 참조 게놈들이 이용될 때, 상기 참조 게놈에 특유하게 매핑된 리드만이 계수된다. 제1 리드가 완료된 후, 상기 단편의 반대 말단으로부터 제2 리드가이 가능하도록 주형은 즉석에서 생성될 수 있다. 따라서, DNA 단편의 단일-말단 또는 쌍을 이룬 말단 서열화가 이용될 수 있다. 시료에서 존재하는 DNA 단편의 부분적 서열화가 실행되고, 예정된 길이, 가령, 36 bp의 리드이 포함된 서열 태그는 공지의 참조 게놈에 매핑되고, 그리고 계수된다. 한 구체예에서, 상기 참조 게놈 서열은 world wide web, genome.ucsc.edu/cgi-bin/hgGateway?org=Human&db=hg18&hgsid=166260105)에서 이용가능한 NCBI36/hg18 서열이다. 대안으로, 상기 참조 게놈 서열은 world wide web, genome.ucsc.edu/cgi-bin/hgGateway에서 이용가능한 GRCh37/hg19이다. 대중적 서열 정보의 다른 출처는 GenBank, dbEST, dbSTS, EMBL (the European 분자 생물학 실험실), 그리고 DDBJ (the DNA Databank of Japan)을 포함한다. 서열들의 정렬에 대한 다수의 컴퓨터 알고리즘이 이용가능한데, BLAST (Altschul et al., 1990), BLITZ (MPsrch) (Sturrock & Collins, 1993), FASTA (Person & Lipman, 1988), BOWTIE (Langmead et al., Genome 생물학 10:R25.1-R25.10 [2009]), 또는 ELAND (Illumina, Inc., San Diego, CA, USA)가 포함되나, 이에 국한되지 않는다. 한 구체예에서, 혈장 cfDNA 분자들의 클론적으로 확장된 복사체의 한 말단이 서열화되고,

Illumina Genome Analyzer에서 생물정보 정렬 분석에 의해 가공되는데, 이것은 Efficient Efficient Large-Scale Alignment of Nucleotide 데이터베이스 (ELAND) 소프트웨어를 이용한다.

[0353] 본 명세서에서 설명된 방법들의 일부 구체예들에 있어서, 상기 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 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 500bp의 서열 리드를 포함한다. 기술적 진보는 500bp 이상의 단일-말단 리드가 쌍을 이룬 말단 리드가 생성될 때, 약 1000bp 이상의 리드를 가능하게 할 것으로 예상된다. 한 구체예에서, 상기 매핑된 서열 태그는 36bp의 서열 리드를 포함한다. 상기 서열 태그의 메칭은 상기 태그의 서열을 참조의 상기 서열과 비교하여, 상기 서열화된 핵산 (가령 cfDNA) 분자의 염색체 기원을 판단함으로써 이루어지고, 그리고 특이적 유전적 서열 정보는 필요하지 않다. 소규모의 불합치 (서열 태그당 0-2개의 불합치)는 참조 게놈과 혼합된 시료 에서 게놈들 사이에 존재할 수 있는 소수 다형성의 계수를 허용할 수 있다.

[0354] 다수의 서열 태그는 전형적으로 시료마다 획득된다. 일부 구체예들에 있어서, 20 내지 40bp 리드, 가령, 36bp이 포함된 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 태그는 시료 당 참조 게놈에 대한 상기 리드의 매핑에 의해 획득된다. 한 구체예에서, 모든 상기 서열 리드는 상기 참조 게놈의 모든 영역에 매핑된다. 한 구체예에서, 모든 영역들, 가령, 상기 참조 게놈의 모든 염색체에 매핑된 태그가 계수되고, 그리고 혼합된 DNA 시료 에서 관심 대상 서열, 가령, 염색체 또는 이의 일부분의 CNV, 가령 과다 또는 과소-현시가 결정된다. 상기 방법은 2개의 게놈 간에 분화를 요구하지 않는다.

[0355] CNV, 가령, 홀배수체가 시료에서 존재하는지 또는 존재하지 않는지에 대한 정확한 판단을 위하여 요구되는 정확성은 서열화 운영에서 시료들중에서 참조 게놈에 매핑된 서열 태그의 수의 변이(염색체간의 변동성), 그리고 상이한 서열화 운영들에서 참조 게놈에 매핑된 서열 태그 수의 변이 (서열화간의 변동성)에서 예측된다. 예를 들면, 상기 변이는 GC-많은 또는 GC-적은 참조 서열들에 매핑된 태그에 대하여 특별히 표명될 수 있다. 상기 핵산의 추출 및 정제, 상기 서열화 라이브러리의 준비에 상이한 프로토콜 이용, 상기 서열화 라이브러리의 준비에, 그리고 상이한 서열화 플랫포옴의 사용으로 인하여 기타 변이가 발생될 수 있다. 본 방법은 염색체간 (운용내), 그리고 서열화간에 (운용간) 그리고 플랫포옴-의존적 변동성으로부터 기안된 증가된 변동성을 본질적으로 계수하기 위하여 정규화 (정규화 염색체 서열들 또는 정규화 세그먼트 서열들) 서열의 지식에 근거한 서열 분량 (염색체 분량, 또는 세그먼트 분량)을 이용한다. 염색체 분량은 단일 염색체, 또는 염색체 1-22, X, 그리고 Y에서 선택된 2개 또는 그 이상의 염색체를 포함할 수 있는 정규화 염색체 서열의 지식에 근거한다. 대안으로, 정규화 염색체 서열들은 하나의 염색체 또는 2개 또는 그 이상의 염색체의 단일 염색체 세그먼트 또는 2개 또는 그 이상의 세그먼트들로 구성될 수 있다. 세그먼트 분량은 정규화 세그먼트 서열의 지식에 근거하는데, 정규화 세그먼트는 임의의 하나의 염색체의 단일 세그먼트, 또는 염색체 1-22, X, 그리고 Y중 임의의 2개 또는 그 이상의 염색체들의 2개 또는 그 이상의 세그먼트들을 포함할 수 있다.

[0356] **CNV 및 출생전 진단**

[0357] 모체 혈액중을 순환하는 무-세포 태아 DNA 및 RNA는 임신 관리 및 출산 판단에 모두 도움이 되도록 되기 위하여 유전적 질환의 수가 증가되는 조기 비-침습성 출생전 진단 (NIPD)에 이용될 수 있다. 혈류중에서 순환하는 무-세포 DNA는 50여년 이상 동안 공지되어 왔었다. 좀더 최근에, 임신 동안 모체 혈류중에서 순환하는 태아 DNA 소량이 존재한다는 것이 발견되었다 (Lo et al., Lancet 350:485-487 [1997]). 죽어가는 태반 세포로부터 유래된 것으로 생각되는데, 무-세포 태아 DNA (cfDNA)는 길이가 200bp미만의 전형적으로 짧은 단편으로 구성되었고, Chan et al., Clin Chem 50:88-92 [2004]), 임신 4주차와 같은 조기에 식별될 수 있고 (Illanes et al., Early Human Dev 83:563-566 [2007]), 그리고 분만의 수시간이내에 모체 순환계로부터 제거되는 것으로 공지되어 있다 (Lo et al., Am J Hum Genet 64:218-224 [1999]). cfDNA에 추가하여, 무-세포 태아 RNA (cfRNA)의 단편은 태아 또는 태반에서 전사되는 유전자들로부터 기인되는데, 모체 혈류에서 또한 포착될 수 있다. 모체 혈액 시료로부터 이들 태아 유전자 요소들의 추출 및 후속적인 분석은 NIPD에 대한 새로운 기회를 부여한다.

[0358] 본 방법은 NIPD에 사용하기 위한 다형성-독립적 방법으로, 그리고 태아 홀배수체의 판단을 위하여 태아 cfDNA가 모체 cfDNA와 구별되어야 하는 것을 요구하지 않는다. 일부 구체예들에 있어서, 상기 홀배수체는 완전한 염색체 삼염색체성 또는 일염색체성, 또는 부분적 삼염색체성 또는 일염색체성이다. 염색체의 일부의 상실 또는 획

득에 의해 부분적 홀배수체가 야기되고, 그리고 불균형 전위, 불균형 역전, 결손 및 삽입으로 인한 염색체 불균형이 포괄된다. 단연코, 생활에서 양립되는 가장 흔한 공지의 홀배수체는 삼염색체성 21, 가령, 다운 증후군(DS)으로써, 염색체 21의 전부 또는 일부의 존재로 인한 것이다. 드물게는, DS는 염색체 21의 전부 또는 일부의 잉여분 복사체가 또다른 염색체 (보통 염색체 14)에 부착되어 단일 이상(aberrant) 염색체가 형성되는, 유전적 또는 산발적 결함에 의해 야기될 수 있다. DS는 지적 장애, 심각한 학습 곤란 및 장기간 건강 문제, 이를테면 심장 질환에 의한 초과 사망률과 연관된다. 공지의 임상적 유의성을 갖는 기타 홀배수체에는 Edward 증후군 (삼염색체성 18) 및 Patau 증후군 (삼염색체성 13)이 포함되는데 출생 첫 몇개월 이내에 흔히 치명적이다. 성염색체의 수와 연관된 비정상 또한 공지되어 있으며, 일염색체성 X, 가령, 여성 출산시 Turner 증후군 (XO), 그리고 삼중 X 증후군 (XXX) 그리고 남성 출산시 증후군 (XXY) 및 XYY 증후군이 포함되며, 이는 모두 불임이 포함된 다양한 표현형과 지적 기능 감소와 연관된다. 일염색체성 X [45, X]는 자연 유산의 약 7%에 해당하는 초기 유산의 가장 흔한 원인이다. 1-2/10,000의 45,X (Turner 증후군이라고도 불림) 출생 빈도에 근거하여, 45,X 임신의 1% 미만이 만삭까지 생존할 것으로 추정된다. Turners 증후군 환자의 약 30%는 45,X 세포 계통과 46,XX 세포 계통 또는 재배열된 X 염색체를 포함하는 계통 모두를 가진 모자이크다 (Hook and Warburton 1983). 출생한 영아의 표현형은 높은 배아 치사률을 고려하면 상대적으로 약하고, Turner 증후군이 있는 모든 출생 여아는 2개의 성 염색체가 포함된 세포 계통을 휴대하는 것으로 가설되었다. 일염색체성 X는 여성에서 45,X 또는 45,X/46XX로, 그리고 남성에서 45,X/46XY로 발생될 수 있다. 인간 상염색체 단체성은 일반적으로 생명과 양립되지 않는 것으로 제시되었지만; 그러나, 출생 어린이에서 하나의 염색체 21의 완전한 일염색체성을 설명한 세포유전적 보고서가 꽤 있다 (Vosranova Iet al., Molecular Cytogen. 1:13 [2008]; Joosten et al., Prenatal Diagn. 17:271-5 [1997]. 본 명세서에서 설명된 방법을 이용하여 출생전 이들 비정상 및 다른 염색체 비정상을 진단할 수 있다.

[0359] 일부 구체예들에 따르면, 상기 본 명세서에서 공개된 방법들은 염색체 1-22, X와 Y중 임의의 하나의 염색체 삼체성의 존재 또는 부재를 결정할 수 있다. 본 방법에 따라 탐지되는 염색체 삼체성의 예로는 삼염색체성 21 (T21; Down 증후군), 삼염색체성 18 (T18; Edward 증후군), 삼염색체성 16 (T16), 삼염색체성 20 (T20), 삼염색체성 22 (T22; Cat Eye 증후군), 삼염색체성 15 (T15; Prader Willi 증후군), 삼염색체성 13 (T13; Patau 증후군), 삼염색체성 8 (T8; Warkany 증후군), 삼염색체성 9, 그리고 the XXY (Kleinefelter 증후군), XYY, 또는 XXX 삼체성이 포함되나, 이에 국한되지 않는다. 비-모자이크 상태에 존재하는 다른 상염색체의 완전한 삼체성은 치명적이나, 모자이크 상태에 존재할 때 생명과 양립될 수 있다. 모자이크 또는 비-모자이크 상태로 존재하던 간에, 다양한 완전한 삼체성과 부분적 삼체성은 본 명세서에서 제공되는 교시에 따라 태아 cfDNA에서 판단됨을 인지할 것이다.

[0360] 본 방법에 의해 결정되는 부분적 삼체성의 비-제한적 예로는 부분적 삼염색체성 1q32-44, 삼염색체성 9 p, 삼염색체성 4 모자이크현상, 삼염색체성 17p, 부분적 삼염색체성 4q26-qter, 부분적 2p 삼염색체성, 부분적 삼염색체성 1q, 및/또는 부분적 삼염색체성 6p/일염색체성 6q이 포함되나, 이에 국한되지 않는다.

[0361] 상기 본 명세서에서 공개된 방법들을 또한 이용하여 염색체 일염색체성 X, 염색체 일염색체성 21, 그리고 부분적 단체성 이를 테면, 일염색체성 13, 일염색체성 15, 일염색체성 16, 일염색체성 21, 그리고 일염색체성 22를 판단할 수 있는데, 이들은 유산과 관련된 것으로 알려져 있다. 완전한 홀배수체에서 전형적으로 수반되는 염색체의 부분적 일염색체성은 또한 본 명세서에서 설명된 방법에 의해 판단될 수 있다. 본 발명에 따라 결정될 수 있는 결손 증후군의 비-제한적 예로는 염색체의 부분적 결손에 의해 야기되는 증후군을 포함한다. 본 명세서에서 설명된 방법들에 따라 결정되는 부분적 결손의 예로는 염색체 1, 4, 5, 7, 11, 18, 15, 13, 17, 22 및 10의 부분적 결손을 포함하나, 이에 국한되지 않으며, 이들은 다음에서 설명된다.

[0362] 1q21.1 결손 증후군 또는 1q21.1 (재발) 극소결손(microdeletion)은 염색체 1의 희귀한 이상이다. 상기 결손 증후군 다음으로, 1q21.1 중복 증후군이 또한 존재한다. 특정 점 상에 결손 증후군이 있는 DNA 유실의 일부분이 있는 한편, 중복 증후군을 가진 동일 점 상에 이 DNA의 유사한 부분의 2 또는 3 복사체가 있다. 문헌에서는 이들 결손과 중복 모두를 1q21.1 복사체-수 변이 (CNV)로 언급한다. 1q21.1 결손은 TAR 증후군 (요골결여가 수반된 혈소판감소증)과 연관될 수 있다.

[0363] Wolf-Hirschhorn 증후군 (WHS) (OMIN #194190)은 염색체 4p16.3의 반접합성 결손과 연관된 연속 유전자 결손 증후군이다. Wolf-Hirschhorn 증후군은 출생전 및 출생후 성장 결핍, 다양한 정도의 발달 장애, 특징적인 두개안면 특징('그리스 전사 투구' 모양의 코, 솟은 전두, 두드러진 이마, 두눈먼거리증(hypertelorism), 고궁(high-arched) 눈썹, 돌출 눈, 내안각 주름(epicanthal folds), 짧은 안중, 입꼬리가 처진(downturned

corners) 특유의 입, 그리고 소악증), 그리고 발작 장애를 특징으로 하는 선천적 기형 증후군이다.

[0364] 5p- 또는 5p 마이너스로 공지된, 그리고 Cris du Chat 증후군 (OMIN#123450)으로 명명된 염색체 5의 부분적 결손은 염색체 5의 짧은 팔(p 아암) (5p15.3-p15.2)의 결손으로 야기된다. 이런 상태를 가진 유아는 고양이와 같은 소리를 내는 고음의 울음소리를 갖는다. 상기 장애는 지적 장애 및 지연 발달, 작은 머리 크기(소두증), 저체중 출산, 그리고 유아기 약한 근긴장(긴장저하증), 독특한 안면 특징 및 있음직한 심장 결함을 특징으로 한다.

[0365] Williams-Beuren 증후군은 염색체 7q11.23 결손 증후군 (OMIN 194050)으로 또한 공지된 연속 유전자 결손 증후군으로써, 대략적으로 28개의 유전자가 포함된 염색체 7q11.23 상에 1.5 내지 1.8 Mb의 반접합성 결손에 의한 다중시스템 장애를 초래한다.

[0366] 11q 결손 장애로 알려진 Jacobsen 증후군은 밴드 11q24.1이 포함된 염색체 11의 말단 영역의 결손으로 발생되는 희귀 선천적 장애다. 이는 지적 장애, 독특한 안면 외양, 그리고 심장 결함 및 출혈 장애가 포함된 다양한 신체적 문제를 야기할 수 있다.

[0367] 일염색체성 18p로 알려진 염색체 18의 부분적 일염색체성은 염색체 18의 짧은 팔(p)의 전부 또는 일부가 결손된 (단염색체성) 희귀한 염색체 장애다. 상기 장애는 작은 키, 다양한 정도의 정신 지체, 말이 느림, 구개골 및 안면(두개안면) 영역의 기형, 및/또는 추가적인 물리적 비정상에 의해 일반적으로 특징화된다. 연관된 두개안면 결함은 사례마다 그 범위 및 정도가 다양할 수 있다.

[0368] 염색체 15의 구조 또는 복제 수의 변화에 의해 야기되는 질환은 Angelman 증후군 및 Prader-Willi 증후군을 포함하는데, 염색체 15의 동일 부분, 15q11-q13 영역에서 유전자 활성 상실이 수반된다. 몇몇 전위 및 미세결손은 편부모에서 미정상일 수 있고, 이는 후손에게 주요 유전적 질환의 원인이 될 수 있는 것으로 인지될 것이다. 예를 들면, 15q11-q13 극소결손을 휴대하는 건강한 엄마는 심각한 신경변성 장애인, Angelman 증후군을 가진 아이를 낳을 수 있다. 따라서, 본 명세서에서 설명된 방법들, 장치 및 시스템들은 태아에서 이러한 부분적 결손 및 기타 결손을 확인하는데 이용될 수 있다.

[0369] 부분적 일염색체성 13q는 염색체 13의 긴 팔(q)의 조각이 유실 (일염색체)될 때 나타나는 희귀 염색체 장애다. 부분적 일염색체성 13q를 가지고 태어난 영아는 출생 저체중, 머리와 안면(두개안면 영역)의 기형, 골격 비정상 (구체적으로 손과 발), 그리고 다른 물리적 비정상을 나타낼 수 있다. 정신 지체가 이 상태의 특징이다. 이 장애를 가진 아이들중에서 영아기 사망률이 높다. 부분적 일염색체성 13q의 거의 모든 경우는 명백한 이유없이(산발적) 무작위로 발생된다.

[0370] Smith-Magenis 증후군 (SMS - OMIM #182290)은 염색체 17의 하나의 복사체 상에 유전적 물질의 결손 또는 상실에 의한 것이다. 잘-알려진 이 증후군은 발달 지연, 정신 지체, 선천적 기형,이상, 이를 테면 심장 및 신장 결함, 그리고 신경행동적 비정상, 이를 테면 심각한 수면 곤란, 및 자해 거동과 연관된다. Smith-Magenis 증후군 (SMS)은 대부분의 경우 (90%) 염색체 17p11.2에서 3.7-Mb의 간질 결손에 의한 것이다.

[0371] DiGeorge 증후군으로 또한 알려진 22q11.2 결손 증후군은 염색체 22의 작은 조각의 결손으로 야기되는 증후군이다. 이 결손 (22 q11.2)은 염색체 쌍중 하나의 긴 팔 상에 염색체 거의 중간에서 일어난다. 이 증후군의 특징은 심지어 같은 가족 구성원들간에도 매우 광범위하고, 신체의 많은 부분에 영향을 준다. 특징적인 신호 및 증후군은 출생 결함, 이를 테면 선천적 심장 질환, 닫힌 상태에서 신경근육 문제와 가장 흔히 관련된 입천장에서 결함(비-인강적 부전), 학습 불능, 안면 특징의 약한 차이, 그리고 재발성 감염을 포함할 수 있다. 염색체 영역 22q11.2에서 극소결손은 20 내지 30배 증가된 정신분열병 위험과 연관된다.

[0372] 염색체 10의 짧은 팔 상에 결손은 DiGeorge 증후군 유사 표현형과 연관된다. 염색체 10p의 부분적 일염색체성은 희귀하지만, DiGeorge 증후군의 특징을 보이는 일부 환자들에서 관찰되었다.

[0373] 한 구체예에서, 본 명세서에서 설명된 상기 방법들, 장치, 그리고 시스템을 이용하여 염색체 1, 4, 5, 7, 11, 18, 15, 13, 17, 22 및 10의 부분적 일염색체성, 가령, 부분적 일염색체성 1q21.11, 부분적 일염색체성 4p16.3, 부분적 일염색체성 5p15.3-p15.2, 부분적 일염색체성 7q11.23, 부분적 일염색체성 11q24.1, 부분적 일염색체성 18p, 염색체 15 (15q11-q13)의 부분적 일염색체성, 부분적 일염색체성 13q, 부분적 일염색체성 17p11.2, 염색체 22(22q11.2)의 부분적 일염색체성이 포함되나, 이에 국한되지 않은 부분적 단체성이 탐지될 수 있으며, 그리고 부분적 일염색체성 10p 또한 상기 방법을 이용하여 결정될 수 있다.

[0374] 본 명세서에서 설명된 방법들에 따라 결정될 수 있는 기타 부분적 단체성은 불균형 전위 t(8;11)(p23.2;p15.5);

11q23 극소결손; 17p11.2 결손; 22q13.3 결손; Xp22.3 극소결손; 10p14 결손; 20p 극소결손, [del(22)(q11.2q11.23)], 7q11.23 and 7q36 결손; 1p36 결손; 2p 극소결손; 신경섬유종증 유형 1 (17q11.2 극소결손), Yq 결손 ; 4p16.3 극소결손; 1p36.2 극소결손; 11q14 결손; 19q13.2 극소결손; Rubinstein-Taybi (16p13.3 극소결손); 7p21 극소결손; Miller-Dieker 증후군 (17p13.3); 그리고 2q37 극소결손을 포함한다. 부분적 결손은 염색체의 일부의 작은 결손일 수 있고, 또는 이들 결손은 염색체의 미세결손일 수 있으며, 이때 단일 유전자의 결손이 발생될 수 있다.

[0375] 염색체 팔중 일부의 중복에 의해 야기되는 몇 가지 중복 증후군이 확인되었다(OMIN [Online Mendelian Inheritance in Man viewed online at ncbi.nlm.nih.gov/omim] 참고). 한 구체예에서, 본 방법을 이용하여 염색체 1-22, X와 Y중 임의의 하나의 세그먼트들의 중복 및/또는 다중복의 존재 또는 부재가 결정될 수 있다. 본 발명의 방법에 따라 결정될 수 있는 중복 증후군의 비-제한적 예로는 염색체 8, 15, 12, 그리고 17의 일부의 중복을 포함하며, 이는 하기에서 설명된다.

[0376] 8p23.1 중복 증후군은 인간 염색체 8의 영역의 중복에 의해 야기되는 희귀 유전 장애다. 이 중복 증후군은 64,000명 출생중 1명의 추정 빈도를 가지고, 상호적 8p23.1 결손 증후군이다. 8p23.1 중복은 말 지연, 발달 지연, 약한 동일이형, 그리고 두드러진 전두 및 곡형 눈썹, 그리고 선천적 심장 질환 (CHD)중 하나 또는 그 이상이 포함된 가변적 표현형과 연관된다.

[0377] 염색체 15q 중복 증후군 (Dup15q)은 염색체 15q11-13.1의 중복으로 발생된 임상적으로 식별가능한 증후군이다. Dup15q를 가진 아기들은 근긴장저하 (부족한 근긴장도), 성장 지체를 보통 가지고 있으며; 구순열 및/또는 구개를 가지고 태어나거나, 또는 심장, 신장 또는 다른 장기의 기형을 가지고 태어날 수 있으며; 어느 정도의 인지 지체/불능 (정신 지체), 말 및 언어 지체, 그리고 감각 과정 장애들을 나타낸다.

[0378] Pallister Killian 증후군은 잉여분 #12 염색체 물질의 결과다. 일부는 잉여분 #12 물질을 가지고, 그리고 일부는 정상적인(잉여분 #12 물질이 없는 46개 염색체) 세포의 혼합물(모자이크현상)이 있다 . 이 증후군을 가진 아기들은 심각한 정신 지체, 부족한 근육긴강, "거친(coarse)" 안면 특징, 그리고 두드러진 전두가 포함된 많은 문제점들을 가지고 있다. 이들은 매우 얇은 윗 입술과 더 두터운 아래 입술 그리고 짧은 코를 가지는 경향이 있다. 기타 건강 문제는 발작, 나쁜 영양, 뻣뻣한 관절, 성인 백내장, 청력장애 및 심장 결함을 포함한다. Pallister Killian이 있는 사람은 수명이 짧다.

[0379] dup(17)(p11.2p11.2) 또는 dup 17p으로 지칭된 유전적 상태를 가진 개체는 염색체 17의 짧은 팔에 잉여분 유전적 정보 (중복으로 공지된)를 휴대한다. 염색체 17p11.2 중복은 Potocki-Lupski 증후군 (PTLS)의 기저가 되며, 이는 새로 인지된 유전적 상태로써 의학 문헌에 보고된 경우가 겨우 수십건 정도이다. 이러한 중복을 가진 환자들은 낮은 근육긴장, 나쁜 영양공급, 그리고 유아기에 잘 자라지 못하고, 그리고 운동 및 언어 발달 시점이 지체된다. PTLS을 가진 많은 개체는 말로 표현 및 언어 과정에 어려움을 가진다. 또한, 환자은 자폐증 또는 자폐증-범위 장애들을 가진 환자에서 볼 수 있는 것과 유사한 거동 특징을 가질 수 있다. PTLS를 가진 개체는 심장 결함 및 수면 무호흡을 가질 수 있다. 상기 유전자 PMP22가 포함된 염색체 17p12에서 큰 영역의 중복이 Charcot-Marie Tooth 질환을 야기하는 것으로 알려져 있다.

[0380] CNV는 사산과 연관되어 있다. 그러나, 통상적인 세포유전학의 고유한 제약으로 인하여, 사산에 CNV의 기여는 실제 보다 과소표시된 것으로 보인다 (Harris et al., Prenatal Diagn 31:932-944 [2011]). 실시예들 및 본 명세서의 도처에 나타낸 것과 같이, 본 방법은 부분적 홀배수체, 가령, 염색체 세그먼트들의 결손 및 다중복의 존재를 결정할 수 있고, 그리고 사산과 연관된 CNV의 존재 또는 부재를 확인하고 판단하는데 이용될 수 있다.

[0381] **CNV를 판단하는 장치 및 시스템**

[0382] 상기 서열화 데이터의 분석, 그리고 이로부터 유도된 진단은 다양한 컴퓨터 이행된 알고리즘 및 프로그램을 이용하여 전형적으로 실행된다. 따라서, 특정 구체예들은 하나 또는 그 이상의 컴퓨터 시스템 또는 다른 공정 시스템에 저장된 또는 이를 통하여 전달된 공정 관련 데이터를 이용한다. 본 명세서에서 공개된 구체예들은 이런 작동을 실행하기 위한 장치에 또한 관계한다. 이 장치는 필요한 목적을 위하여 특별히 구축될 수 있거나, 또는 이 장치는 컴퓨터 프로그램에 의해 선택적으로 활성화되거나 또는 재구성된 일반적-목적의 컴퓨터 (또는 컴퓨터 집단) 및/또는 이 컴퓨터에 저장된 데이터 구조일 수 있다. 일부 구체예들에 있어서, 프로세서 집단은 인용된 모든 또는 일부 분석학적 작동을 통합적으로 (가령, 네트워크를 통하여 또는 클라우드 컴퓨팅을 통하여) 및/또는 병행 실행한다. 본 명세서에서 설명된 방법들을 실행하는 프로세서 또는 프로세서 집단은 마이크로콘트롤러 및 마이크로프로세서, 이를 테면 프로그램작동이 가능한 장치들 (가령, CPLDs 및 FPGAs) 그리고 프로그

램작동이 불가능한 장치들, 이를 테면 게이트 어레이 ASICs 또는 일반적 목적의 마이크로프로세서가 포함된 다양한 유형일 수 있다.

[0383] 또한, 특정 구체예들은 다양한 컴퓨터-시행된 작동을 실행하기 위한 프로그램 명령 및/또는 데이터 (데이터 구조 포함)를 포함하는 구체적인 및/또는 비-일시적인 컴퓨터 리드가능 매체 또는 컴퓨터 프로그램 제품에 관계한다. 컴퓨터-리드가능 매체의 예로는 반도체 메모리 장치들, 자석 매체, 이를 테면 디스크 드라이브, 자석 테이프, 광학 매체 이를 테면 CDs, 자석-광학 매체, 그리고 프로그램 명령을 저장 및 실행하도록 특별히 기획된 하드웨어 장치들, 이를 테면 리드-전용 메모리 장치들 (ROM) 및 임의 접근 메모리 (RAM)를 포함하나, 이에 국한되지 않는다. 컴퓨터 리드가능 매체는 최종 사용자에 의해 직접적으로 조절될 수 있거나 또는 상기 매체는 최종 사용자에 의해 간접적으로 조절될 수 있다. 직접적으로 조절되는 매체의 예로는 다른 엔터티와 공유되지 않는 사용자 설비 및/또는 매체에 위치한 매체를 포함한다. 간접적으로 조절되는 매체의 예로는 외부 네트워크 및/또는 공유된 자료를 제공하는 서비스, 이를 테면 "클라우드"를 통하여 사용자에게 간접적으로 접근가능한 매체를 포함한다. 프로그램 명령의 예로는 기계 코드, 이를 테면 컴파일러(compiler)에 의해 생산되는 기계 코드, 그리고 해석자를 이용하여 컴퓨터에 의해 이행될 수 있는 더 높은 수준 코드가 포함된 파일 모두를 포함한다.

[0384] 다양한 구체예들에 있어서, 공개된 방법들 및 장치에서 이용되는 데이터 또는 정보는 전자 포멧으로 제공된다. 이러한 데이터 또는 정보는 핵산 시료로부터 유도된 리드 및 태그, 참조 서열의 특정 영역들과 함께 정렬되는 (가령, 염색체 또는 염색체 세그먼트에 정렬되는) 이러한 태그의 계수 또는 밀도 , 참조 서열들 (단독으로 또는 일차적으로 다형성을 제공하는 참조 서열들을 포함), 염색체 및 세그먼트 분량, 소명(calls) 이를 테면 홀배수체 소명, 정규화된 염색체 및 세그먼트 값, 염색체 또는 세그먼트들과 대응하는 정규화 염색체 또는 세그먼트의 쌍, 상담 권고, 진단, 그리고 이와 유사한 것들을 포함할 수 있다. 본 명세서에서 이용된 바와 같이, 전자 포멧으로 제공되는 데이터 또는 다른 정보는 기계 상에 저장 및 기계 간의 전달에 이용가능하다. 통상적으로, 전자 포멧의 데이터는 디지털로 제공되며, 다양한 데이터 구조, 목록, 데이터베이스, 등등에서 비트 및/또는 바이트로 저장될 수 있다. 데이터는 전자적으로, 광학적으로, 기타 등등으로 구체화될 수 있다.

[0385] 한 구체예는 테스트 시료에서 홀배수체, 가령, 태아 홀배수체 또는 암의 존재 또는 부재를 나타내는 산출량을 생성하는 컴퓨터 프로그램 생성물을 제공한다. 상기 컴퓨터 생성물은 염색체 이상을 판단하기 위하여 임의의 하나 또는 그 이상의 상기에서 설명된 방법들을 실행하는 명령을 포함할 수 있다. 설명된 바와 같이, 상기 컴퓨터 생성물은 프로세서를 통하여 염색체 분량을 결정하고, 그리고 일부 경우에 있어서 태아 홀배수체가 존재 또는 부재하는 지를 판단하기 위하여 비-일시적인 및/또는 유형의 컴퓨터 리드가능 매체 상에 컴퓨터 실행가능한 또는 편집가능한 로직 (가령, 명령)을 보유하는 비-일시적인 및/또는 유형의 컴퓨터 리드가능 매체를 포함할 수 있다. 한 실시예에서, 상기 컴퓨터 생성물은 프로세서에 의해 태아 홀배수체를 진단할 수 있도록 컴퓨터 실행가능한 또는 편집가능한 로직 (가령, 명령)을 갖는 컴퓨터 리드가능 매체를 포함하는데: 모체 생물학적 시료로부터 핵산 분자들의 최소한 일부분으로부터 서열화 데이터를 수용하는 수용 과정, 여기에서 전술한 서열화 데이터는 산출된 염색체 및/또는 세그먼트 분량을 포함하고; 전술한 수용된 데이터로부터 태아 홀배수체를 분석하기 위한 컴퓨터 지원된 로직; 그리고 전술한 태아 홀배수체의 존재, 부재 또는 종류를 표시하는 산출량을 생성시키기 위한 산출량 과정을 포함한다.

[0386] 고려중인 시료로부터 서열 정보는 염색체 참조 서열들에 매핑되어 임의의 하나 또는 그 이상의 관심 염색체 각각에 대한 서열 태그 수를 확인하고, 그리고 전술한 임의의 하나 또는 그 이상의 관심 염색체 각각에 대한 정규화 세그먼트 서열에 대한 서열 태그 수를 확인할 수 있다. 다양한 구체예들에 있어서, 상기 참조 서열들은 데이터베이스, 이를 테면 관계(relational) 데이터베이스 또는 객체(object) 데이터베이스에 저장된다.

[0387] 대부분의 경우 도움을 받지 않은 인간이 본 명세서에서 공개된 방법들의 컴퓨터 작동을 실행하는 것이 현실적이지 않거나 또는 가능하지 않을 수 있음을 인지해야 한다. 예를 들면, 시료의 단일 30 bp 리드를 인간 염색체중 임의의 하나에 매핑시키는 것도 컴퓨터 장치의 지원 없이는 수년 간의 시도를 필요로 할지도 모른다. 물론 믿을만한 홀배수체 소명은 매핑 수천 (가령, 최소한 약 10,000) 또는 심지어 백만번의 리드를 하나 또는 그 이상의 염색체에 매핑시키는 것을 일반적으로 요구하기 때문에, 이 문제는 더 복잡하다.

[0388] 상기 본 명세서에서 공개된 방법들은 테스트 시료에서 유전적 관심 대상 서열의 복제수 평가를 위한 시스템을 이용하여 실행될 수 있다. 상기 시스템은 다음을 포함한다: (a) 시료로부터 핵산 서열 정보를 제공하기 위하여 테스트 시료로부터 핵산을 수용하는 서열화기; (b) 프로세서; 그리고 (c) 마스크에 의해 필터된 Y 염색체의 참조 서열을 이용하여 테스트 시료에서 Y 염색체의 복제 수를 평가하기 위하여 전술한 프로세서에서 실행을 위하

여 명령이 저장된 하나 또는 그 이상의 컴퓨터-리드가능 저장 매체. 상기 마스크는 Y 염색체의 참조 서열 상에 특정 크기의 빈들을 포함한다. 상기 빈들은 이에 정렬된 훈련용 서열 태그의 임계치 수 이상을 갖는다. 상기 훈련용 서열 태그는 Y 염색체의 참조 서열에 개별적으로 정렬된 다수의 여성 개체의 서열 리드를 포함한다.

[0389] 일부 구체예들에 있어서, 상기 방법들은 임의의 CNV, 가령, 염색체 또는 부분적 홀배수체를 식별해내기 위한 방법을 실행하기 위하여 컴퓨터-리드가능 명령이 저장된 컴퓨터-리드가능 매체에 의해 지시된다. 따라서 한 구체예는 태아 및 모체 무-세포 핵산이 포함된 테스트 시료에서 관심 대상 서열의 복제 수 평가를 위한 방법을 컴퓨터 시스템이 실행할 수 있도록 컴퓨터 시스템의 하나 또는 그 이상의 프로세서에 의해 이행될 때, 컴퓨터-실행가능한 명령이 저장된 하나 또는 그 이상의 컴퓨터-리드가능 비-일시적인 저장 매체를 포함하는 컴퓨터 프로그램 산물을 제공한다. 상기 방법은 다음 공정을 포함한다: (a) 테스트 시료의 서열 리드를 제공하고; (b) 상기 관심 대상 서열이 포함된 참조 게놈에 상기 테스트 시료의 서열 리드를 정렬시키고, 그렇게 함으로써 테스트 서열 태그가 제공되며; (c) 각 빈에 위치한 테스트 서열 태그의 커버리지를 결정하고, 여기에서 상기 참조 게놈은 다수의 빈으로 분할되고; (d) 상기 관심 대상 서열에 대한 포괄적 프로파일이 제공되며, 여기에서 포괄적 프로파일은 각 빈 에서 예상된 커버리지를 포함하고, 그리고 여기에서 상기 예상된 커버리지는 상기 테스트 시료와 실질적으로 동일한 방식으로 서열화되고, 정렬된 영향을 받지 않은 훈련용 시료들의 훈련용 세트로부터 획득되며, 상기 예상된 커버리지는 빈간의 변이를 나타내고; (e) 각 빈에서 상기 예상된 커버리지에 따라 테스트 서열 태그의 커버리지를 조절하고, 그렇게 함으로써 상기 테스트 서열 태그의 각 빈에서 포괄적-프로파일-제거된 커버리지가 획득되며; (f) GC 함량 수준과 상기 테스트 서열 태그의 빈에 대하여 포괄적-프로파일-제거된 커버리지 사이의 상관관계에 기초하여 포괄적-프로파일-제거된 커버리지를 조절하고, 그렇게 함으로써 상기 관심 대상 서열 상에서 상기 테스트 서열 태그의 시료-GC-교정된 커버리지가 획득되며; 그리고 (g) 이 시료-GC-교정된 커버리지에 기초하여, 테스트 서열에서 관심 대상 서열의 복제수가 평가된다.

[0390] 일부 구체예들에 있어서, 상기 명령은 상기 방법에 관련된 자동적으로 기록된 정보, 이를 테면 모체 테스트 시료를 제공하는 인간 개체에 대한 환자의 의무 기록에서 염색체 분량 및 태아 염색체 홀배수체의 존재 또는 부재를 추가도 더 포함할 수 있다. 상기 환자 의무 기록은 예를 들면, 실험실, 전문의실, 병원, 건강 유지 기구, 보험 회사, 또는 개인 의무 기록 웹사이트에 의해 유지될 수 있다. 더욱이, 프로세서-시행된 분석에 근거하여, 상기 방법은 모체 테스트 시료를 얻은 인간 개체의 처방, 치료의 개시 및/또는 변경을 더 수반할 수 있다. 이는 상기 개체로부터 취한 추가 시료에서 하나 또는 그 이상의 추가적인 테스트 또는 분석을 실행하는 것을 수반할 수 있다.

[0391] 공개된 방법들은 임의의 CNV, 가령, 염색체 또는 부분적 홀배수체를 식별하기 위한 방법을 실행하도록 채택된 또는 구성된 컴퓨터 공정 시스템을 이용하여 또한 실행될 수 있다. 한 구체예는 본 명세서에서 설명된 방법을 실행하기 위하여 채택된 또는 구성된 컴퓨터 공정 시스템을 제공한다. 한 구체예에서, 상기 장치는 본 명세서의 도처에서 설명된 서열 정보의 유형을 획득하기 위하여 시료 에서 핵산 분자들의 최소한 일부분을 서열화시키기 위하여 채택된 또는 구성된 서열화 장치를 포함한다. 상기 장치는 이 시료를 가공하기 위한 성분들을 또한 포함할 수 있다. 이러한 성분들은 본 명세서의 도처에서 설명된다.

[0392] 서열 또는 다른 데이터는 컴퓨터에 입력되거나 또는 컴퓨터 리드가능 매체에 직접적으로 또는 간접적으로 저장될 수 있다. 한 구체예에서, 컴퓨터 시스템은 시료의 핵산 서열을 리드 및/또는 분석하는 서열화 장치에 바로 연결된다. 이러한 도구로부터 서열들 또는 다른 정보는 상기 컴퓨터 시스템에서 인터페이스를 통하여 제공된다. 대안으로, 시스템에 의해 가공된 서열들은 서열 저장 출처, 이를 테면 데이터베이스 또는 다른 저장소로부터 제공된다. 공정 장치에 일단 이용가능하다면, 메모리 장치 또는 집단 저장 장치는 상기 핵산의 서열을 완충시키거나, 또는 최소한 일시적으로 저장한다. 또한, 상기 메모리 장치는 다양한 염색체 또는 게놈들, 등등의 태그 계수를 저장할 수 있다. 상기 메모리는 또한 제시되는 상기 서열 또는 매핑된 데이터를 분석하기 위하여 다양한 루틴 및/또는 프로그램을 저장할 수 있다. 이러한 프로그램/루틴은 통계학적 분석, 등등을 실행하기 위한 프로그램을 포함할 수 있다.

[0393] 한 실시예에서, 사용자는 시료를 서열화 장치로 제공한다. 데이터는 수집되고 및/또는 컴퓨터에 연결된 서열화 장치에 의해 분석된다. 컴퓨터 상의 소프트웨어는 데이터 수집 및/또는 분석을 허용한다. 데이터는 저장되거나, 전시되거나 (모니터 또는 다른 유사한 장치를 통하여), 및/또는 또다른 위치로 보내질 수 있다. 상기 컴퓨터는 인터넷에 연결되어, 원격 사용자 (가령, 의사, 과학적 또는 분석자)가 이용하는 포켓용 장치로 데이터를 전송하는데 이용된다. 전송전 상기 데이터는 저장되거나 및/또는 분석될 수 있음을 인지할 것이다. 일부 구체예들에 있어서, 미가공 데이터가 수집되고, 사용자 또는 장치로 전송되어, 상기 데이터를 분석하거나 및/또는 저장할 수 있다. 전송은 인터넷을 통하여 일어날 수 있지만, 위성 또는 다른 연결을 통하여 일어날 수도 있다.

대안으로, 데이터는 컴퓨터-리드가능 매체 상에 저장되고, 상기 매체는 최종 사용자 (가령,우편을 통하여)에게 전달될 수 있다. 원격 사용자는 건물, 도시, 주, 나라 또는 대륙이 포함되나, 이에 국한되지 않는 동일한 또는 상이한 지리학적 위치에 있을 수 있다.

[0394] 일부 구체예들에 있어서, 상기 방법들은 다수의 폴리뉴클레오티드 서열들 (가령, 리드, 태그 및/또는 참조 염색체 서열들)에 관한 데이터를 수집하고, 상기 데이터를 컴퓨터 또는 다른 컴퓨터 시스템으로 전송하는 것을 또한 포함한다. 예를 들면, 상기 컴퓨터는 실험실 장비, 가령, 시료 수집 장치, 뉴클레오티드 증폭 장치, 뉴클레오티드 서열화 장치, 또는 혼성화 장치에 연결될 수 있다. 상기 컴퓨터는 그 다음 실험실 장치에 의해 모여진 적용가능한 데이터를 수집할 수 있다. 상기 데이터는 임의의 단계에서, 가령, 실시간으로 수집되고, 전송전, 또는 전송하는 동안, 전송과 병행하여, 또는 전송 후 컴퓨터에 저장될 수 있다. 상기 데이터는 상기 컴퓨터로부터 추출될 수 있는 컴퓨터-리드가능 매체 상에 저장될 수 있다. 수집된 또는 저장된 데이터는 상기 컴퓨터로부터 원격 위치까지, 가령, 지역 네트워크 또는 광역 네트워크, 이를 테면 인터넷을 통하여 전송될 수 있다. 원격 위치에서 하기에서 설명되는 바와 같이, 전송된 데이터에 다양한 작동이 실시될 수 있다.

[0395] 본 명세서에서 공개된 시스템, 장치 및 방법에서 저장, 전송, 분석 및/또는 작동될 수 있는 전자적으로 포멧화된 데이터의 유형들중에서 다음의 것들이 있다.

[0396] 테스트 시료에서 핵산의 서열화에 의해 획득되는 리드

[0397] 리드를 참조 게놈 또는 다른 참조 서열 또는 서열들에 정렬시켜 획득되는 태그

[0398] 참조 게놈 또는 서열

[0399] 서열 태그 밀도 - 참조 게놈 또는 다른 참조 서열들의 2개 또는 그 이상의 영역들 (전형적으로 염색체 또는 염색체 세그먼트들) 각각에 대한 태그의 계수 또는 숫자

[0400] 관심 대상의 특정 염색체 또는 염색체 세그먼트들에 대한 정규화 염색체 또는 염색체 세그먼트들의 정체성

[0401] 관심대상의 염색체 또는 세그먼트들과 대응하는 정규화 염색체 또는 세그먼트들로부터 획득된 염색체 또는 염색체 세그먼트들 (또는 다른 영역들)에 대한 분량

[0402] 염색체 분량을 영향을 받은, 비-영향을 받은 것으로 소명 또는 소명 없음을 위한 임계치

[0403] 염색체 분량의 실질 소명

[0404] 진단 (상기 소명과 연관된 임상적 상태)

[0405] 상기 소명 및/또는 진단으로부터 유래된 추가 테스트 권고

[0406] 상기 소명 및/또는 진단으로부터 유래된 치료 및/또는 감시 계획

[0407] 이러한 다양한 유형의 데이터는 별개의 장치를 이용하여 하나 또는 그 이상의 위치에서 획득, 저장, 전송, 분석 및/또는 작동될 수 있다. 상기 공정 선택은 광범위한 범위로 뻗어있다. 상기 범위의 한 단부에서, 이러한 모든 또는 대부분의 정보는 상기 테스트 시료가 가공되는 위치, 가령, 의사 진료실 또는 다른 임상 환경에서 저장, 이용된다. 다른 극단의 경우, 이 시료는 한 지역에서 획득되고, 상이한 위치에서 가공되고, 임의선택적으로 서열화되고, 리드는 하나 또는 그 이상의 상이한 위치에서 정렬되고, 소명이 이루어지며, 그리고 여전히 또다른 위치(이때 이 위치는 시료가 획득된 위치일 수 있음)에서 진단, 권고 및 기획이 준비될 수 있다.

[0408] 다양한 구체예들에 있어서, 상기 리드는 상기 서열화 장치와 함께 생성되고, 그 다음 원격 위치로 전송되어, 홀배수체 소명이 만들어지도록 가공된다. 이 원격 위치에서, 예로써, 상기 리드는 참조 서열에 정렬되어 태그가 만들어지고, 이 태그는 계수되고, 관심 대상의 염색체 또는 세그먼트들에 할당된다. 또한 상기 원격 위치에서, 상기 계수는 연관된 정규화 염색체 또는 세그먼트들을 이용하여 분량으로 전환된다. 여전히 추가적으로, 상기 원격 위치에서, 상기 분량을 이용하여 홀배수체 소명이 이루어진다.

[0409] 별개의 장소에서 이용될 수 있는 공정 작동들중에 다음의 것들이 있다:

[0410] 시료 수집

[0411] 서열화에 앞서 시료 가공

[0412] 서열화

[0413] 서열 데이터 분석 및 홀배수체 소명 유도

[0414] 진단

[0415] 진단 및/또는 소명을 환자 또는 건강 관리 제공자에게 보고

[0416] 추가 치료, 테스트, 및/또는 감시를 위한 계획 개발

[0417] 상기 계획 실행

[0418] 상담

[0419] 이들 작동중 임의의 하나 또는 그 이상은 본 명세서의 도처에서 설명된 바와 같이 자동화될 수 있다. 전형적으로, 상기 서열화 및 서열 데이터의 분석 그리고 홀배수체 소명 유도는 자동적으로 실행될 수 있다. 다른 작동들은 수작업으로 또는 자동으로 실행될 수 있다.

[0420] 시료 수집이 실행될 수 있는 위치의 예로는 건강 요원 사무실, 진료소, 환자의 집 (이때 시료 수집 도구 또는 키트가 제공됨), 그리고 이동식 건강 관리 차량을 포함한다. 서열화에 앞서 시료를 가공하는 위치의 예로는 건강 요원 사무실, 진료소, 환자의 집 (이때 시료 가공 도구 또는 키트가 제공됨), 그리고 이동식 건강 관리 차량 그리고 홀배수체 분석 제공자의 설비를 포함한다. 서열화가 실행되는 위치의 예로는 건강 요원 사무실, 진료소, 환자의 집 (이때 시료 서열화 도구 또는 키트가 제공됨), 그리고 이동식 건강 관리 차량 그리고 홀배수체 분석 제공자의 설비를 포함한다. 상기 서열화가 일어나는 위치는 전자 포멧의 서열 데이터 (전형적으로 리드)를 전달하기 위한 전용(dedicated) 네트워크 연결로 제공될 수 있다. 이러한 연결은 컴퓨터 시스템에 연결될 수 있거나 또는 무선일 수 있고, 공정 부위로 전성되기 전, 데이터가 가공되거나 및/또는 응집되는 부위로 데이터를 보내도록 구성될 수 있다. 데이터 수집자(aggregators)는 건강 기구, 이를 테면 건강 유지 기구(HMOs)에 의해 유지될 수 있다.

[0421] 분석 및/또는 유도 작동은 임의의 전술한 위치에서 또는 대안으로 핵산 서열 데이터를 분석하는 서비스 전용부위로부터 더 먼 부위에서 실행될 수 있다. 이러한 위치는 예를 들면, 클러스터, 이를 테면 일반 목적 서버 팜, 홀배수체 분석 서비스 비지니드 설비, 그리고 이와 유사한 것들을 포함한다. 일부 구체예들에 있어서, 이 분석을 실행하는데 이용되는 컴퓨터 장치는 대여되거나 또는 임차된다. 상기 계산 재원은 프로세서의 인터넷 접근 가능한 수집, 이를 테면 일상적인 말로 클라우드로 알려진 공정 재원일 수 있다. 일부 경우들에 있어서, 계산은 서로 제휴된 또는 제휴안된 평행 또는 대량적으로 평행 프로세스 집단에 의해 실행된다. 이 공정은 분산 공정, 이를 테면 클러스터 계산(computing), 그리드 계산(grid computing), 그리고 이와 유사한 것들을 이용하여 실행될 수 있다. 이러한 구체예들에 있어서, 집합적 계산 재원의 클러스터 또는 그리드는 본 명세서에서 설명된 분석 및/또는 유도를 실행하기 위하여 함께 작용하는 다중 프로세스 또는 컴퓨터를 포함하는 슈퍼 가상 컴퓨터를 형성한다. 이들 기술 뿐만 아니라 더욱 통상적인 슈퍼컴퓨터는 본 명세서에서 설명된 바와 같이 서열 데이터를 가공하는데 이용될 수 있다. 각각은 프로세서 또는 컴퓨터에 의존하는 병행 계산 형태다. 그리드의 경우에, 이들 프로세서 (대개 전체 컴퓨터)를 계산은 통상적인 네트워크 프로토콜, 이를 테면 이더넷(Ethernet)에 의해 네트워크 (개인, 대중, 또는 인터넷)에 연결된다. 대조적으로, 슈퍼컴퓨터는 로컬 고속 컴퓨터 버스에 연계된 많은 프로세서를 가진다.

[0422] 특정 구체예들에 있어서, 분석 작동과 동일한 동일한 위치에서 진단 (가령, 태아는 다운 증후군을 가지거나 또는 상기 환자는 특정 유형의 암을 가짐)이 이루어진다. 다른 구체예들에서, 진단은 상이한 위치에서 실행된다. 일부 실시예들에 있어서, 상기 진단의 보고는 반드시 그럴 필요는 없지만, 시료를 취하는 위치에서 실행된다. 진단이 이루어지는 또는 보고되는 위치 및/또는 계획 개발이 실행되는 위치의 예로는 건강 요원 사무실, 진료소, 그리고 포켓용 장치들 이를 테면 이동전화, 테블릿, 스마트폰, 네트워크에 무선으로 연결된 또는 컴퓨터에 연결된 이동전화, 테블릿, 스마트폰, 등등에 의해 접근가능한 인터넷 주소를 포함한다. 상담이 실행되는 위치의 예로는 건강 요원 사무실, 진료소, 컴퓨터, 포켓용 장치들 등에 의해 접근가능한 인터넷 주소등을 포함한다.

[0423] 일부 구체예들에 있어서, 이 시료 수집, 시료 공정, 그리고 서열화 작동은 제1 위치에서 실행되고, 분석 및 유도 작동은 제2 위치에서 실행된다. 그러나, 일부 경우들에 있어서, 이 시료 수집은 한 (가령, 건강 요원 사무실 또는 진료소)에서 수집되고, 이 시료 공정 및 서열화는 상이한 위치에서 실행되는데, 이 위치는 임의선택적으로 분석 및 유도가 이루어지는 위치와 동일하다.

[0424] 다양한 구체예들에 있어서, 상기 열거된 작동의 순서는 시료 수집, 시료 공정 및/또는 서열화를 시작하는 사용

자 또는 엔터티에 의해 촉발될 수 있다. 이들 작동의 하나 또는 그 이상이 실행되기 시작한 후, 다른 작동이 자연적으로 이어질 수 있다. 예를 들면, 상기 서열화 작동은 리드가 자동적으로 수집되고, 대개 자동적으로 그리고 가능하면 사용자 중재, 상기 서열 분석 그리고 홀배수체 작동의 유도없이 실행되는 공정 장치로 보내지도록 할 수 있다. 일부 실행에 있어서, 그 다음 이 공정 작동의 결과는 가능하면 진단으로 다시 포멧되어, 건강 전문가 및/또는 환자에게 상기 정보를 보고하는 시스템 성분 또는 엔터티로 자동적으로 전달된다. 설명된 바와 같이 이러한 정보는 자동적으로 또한 가공되어, 가능하면 상담 정보와 함께, 치료, 테스트, 및/또는 감시 계획이 만들어질 수 있다. 따라서, 초기 단계 작동의 개시는 건강 전문가, 환자 또는 다른 관련 부분에 진단, 계획, 상담 및/또는 물리적 상태에서 작용하는데 유용한 다른 정보를 제공하는 최종 순서를 촉발시킬 수 있다. 전체 시스템의 일부들이 물리적으로 분리되어 있고, 가령, 이 시료와 서열 장치와 멀리 있어도 이는 가능하다.

[0425] 도 7은 테스트 시료로부터 소명 또는 진단을 하기 위한 분산된 시스템의 한 가지 실행을 보여준다. 시료 수집 위치 01은 환자, 이를 테면 임신한 여성 또는 가상 암 환자로부터 테스트 시료를 획득하는데 이용된다. 그 다음 공정 및 테스트 시료가 가공되는 서열화 위치 03로 에 제공되는 시료들은 상기에서 설명된 바와 같이 가공되고, 서열화된다. 위치 03은 이 시료의 가공을 위한 장치 뿐만 아니라 상기 가공된 시료의 서열화를 위한 장치를 포함한다. 본 명세서의 도처에서 설명된 바와 같이 서열화 결과는 전형적으로 전자 포멧 형태로 제공되고, 네트워크, 이를 테면 도 7에서 참고 숫자 05로 표시된 인터넷으로 제공되는 리드의 수집이다.

[0426] 상기 서열 데이터는 멀리있는 위치 07로 제공되며, 이곳에서 분석 및 소명 생성이 실행된다. 이 위치는 하나 또는 그 이상의 강력한 컴퓨터 장치들, 이를 테면 컴퓨터 또는 프로세서를 포함할 수 있다. 위치 07에서 계산 재원이 이들의 분석을 완료하고, 제공받은 서열 정보로부터 소명을 생성한 후, 이 소명은 네트워크 05로 다시 전달된다. 일부 실행에 있어서, 위치 07에서 소명이 생성될 뿐만 아니라, 연관된 진단이 또한 이루어진다. 그 다음 소명 및/또는 진단은 네트워크를 통하여 도 7에서 설명된 바와 같이, 다시 시료 수집 위치 01로 되돌아간다. 설명된 바와 같이, 이것은 소명 또는 진단을 만드는 것과 연관된 다양한 작동들이 다양한 위치로 분할되는지를 보여주는 많은 변형중의 단순한 하나이다. 한 가지 공통적인 변이는 단일 위치에서 시료 수집 및 공정 그리고 서열화를 제공하는 것이 관련된다. 또다른 변이는 분석 및 소명 생성과 동일한 위치에서 공정 및 서열화를 제공하는 것과 관련된다.

[0427] 도 8은 별개의 위치들에서 다양한 작동을 실행하는 옵션에 대해 상술한다. 도 8에서 도시된 가장 세분화된 감각에서, 다음의 각 작동은 별도의 위치에서 실행된다: 시료 수집, 시료 공정, 서열화, 리드 정렬, 소명, 진단, 그리고 보고 및/또는 계획 개발.

[0428] 이들 작동중 일부를 합치는 한 구체예에서, 시료 가공 및 서열화가 한 위치에서 실행되고, 그리고 리드 정렬, 소명, 그리고 진단이 별도의 위치에서 실행된다. 참조 문자 A로 표시된 도 8 부분 참고. 또다른 시행에서, 도 8에서 문자 B로 표시된 바와 같이, 시료 수집, 시료 공정, 그리고 서열화는 모두 동일한 위치에서 실행된다. 이 시행에서, 리드 정렬 및 소명은 제2 위치에서 실행된다. 끝으로, 진단 및 보고 및/또는 계획 개발은 제 3의 위치에서 실행된다. 도 8에서 문자 C로 표시된 시행에 있어서, 시료 수집은 제1 위치에서 실행되며, 시료 공정, 서열화, 리드 정렬, 소명, 그리고 진단은 모두 제2 위치에서 함께 실행되며, 그리고 보고 및/또는 계획 개발은 제 3 위치에서 실행된다. 끝으로, 도 8에서 D로 라벨된 시행에 있어서, 시료 수집은 제1 위치에서 실행되며, 시료 공정, 서열화, 리드 정렬, 그리고 소명은 모두 제2 위치에서 실행되며, 그리고 진단 및 보고 및/또는 계획 관리는 제 3 위치에서 실행된다.

[0429] 한 구체예는 태아 및 모체 핵산이 포함된 모체 테스트 시료 에서서 임의의 하나 또는 그 이상의 상이한 완전한 태아 염색체 홀배수체의 존재 또는 부재를 결정하는데 이용되는 시스템을 제공하는데, 상기 시스템은 핵산 시료를 수용하고, 이 시료로부터 태아 및 모체 핵산 서열 정보를 제공하는 서열화기; 프로세서; 그리고 전술한 프로세서에서 실행을 위한 명령이 포함된 기계 리드가능 저장 매체를 포함하며, 상기 명령은 다음을 포함한다:

[0430] (a) 이 시료 에서 태아 및 모체 핵산에 대한 서열 정보를 획득하기 위한 코드;

[0431] (b) 염색체 1-22, X, 그리고 Y로부터 선택된 임의의 하나 또는 그 이상의 관심 염색체 각각에 대하여 태아 및 모체 핵산으로부터 서열 태그 수를 계산적으로 식별하기 위하여 전술한 서열 정보를 이용하고, 그리고 전술한 임의의 하나 또는 그 이상의 관심 염색체 각각에 대한 최소한 하나의 정규화 염색체 서열 또는 정규화 염색체 세그먼트 서열의 서열 태그 수를 확인하기 위한 코드;

[0432] (c) 그리고 임의의 하나 또는 그 이상의 관심 염색체 각각에 대한 단일 염색체 분량을 산출하기 위하여, 전술한 임의의 하나 또는 그 이상의 관심 염색체 각각에 대하여 확인된 서열 태그 수, 그리고 정규화 염색체 서열 또는

정규화 염색체 세그먼트 서열 각각에 대하여 확인된 서열 태그의 수를 이용하는 코드; 그리고

[0433] (d) 임의의 하나 또는 그 이상의 관심 염색체 각각에 대한 각 단일 염색체 분량을 하나 또는 그 이상의 관심 염색체 각각에 대한 대응하는 임계치 값에 비교하고, 그리고 그렇게 함으로써 이 시료내의 임의의 하나 또는 그 이상의 완전한 상이한 태아 염색체 홀배수체의 존재 또는 부재를 결정하기 위한 코드

[0434] 일부 구체예들에 있어서, 임의의 하나 또는 그 이상의 관심 염색체 각각에 대한 단일 염색체 분량을 산출하기 위한 코드는 관심 염색체의 선택된 하나에 대한 염색체 분량을 선택된 관심 염색체에 대하여 확인된 서열 태그의 수와 선택된 관심 염색체에 대한 대응하는 최소한 하나의 정규화 염색체 서열 또는 정규화 염색체 세그먼트 서열에 대하여 확인된 서열 태그의 수의 비율로 산출하는 코드를 포함한다.

[0435] 일부 구체예들에 있어서, 상기 시스템은 임의의 하나 또는 그 이상의 관심 염색체의 임의의 하나 또는 그 이상의 세그먼트들의 임의의 나머지 염색체 세그먼트들 각각에 대한 염색체 분량의 산출을 반복하기 위한 코드를 더 포함한다.

[0436] 일부 구체예들에 있어서, 염색체 1-22, X, 그리고 Y로부터 선택된 하나 또는 그 이상의 관심 염색체는 염색체 1-22, X, 그리고 Y에서 선택된 최소 20개 염색체를 포함하고, 그리고 여기에서 상기 명령은 최소한 20개의 상이한 완전한 태아 염색체 홀배수체의 존재 또는 부재 결정을 위한 명령을 포함한다.

[0437] 일부 구체예들에 있어서, 최소한 하나의 정규화 염색체 서열은 염색체 1-22, X, 그리고 Y로부터 선택된 염색체 집단이다. 다른 구체예들에서, 최소한 하나의 정규화 염색체 서열은 염색체 1-22, X, 그리고 Y로부터 선택된 단일 염색체다.

[0438] 또 다른 구체예는 태아 및 모체 핵산이 포함된 모체 테스트 시료에서 임의의 하나 또는 그 이상의 상이한 부분적 태아 염색체 홀배수체의 존재 또는 부재를 결정하는데 이용되는 시스템을 제공하는데, 상기 시스템은 핵산 시료를 수용하고, 이 시료로부터 태아 및 모체 핵산 서열 정보를 제공하는 서열화기; 프로세서; 그리고 전술한 프로세서에서 실행을 위한 명령이 포함된 기계 리드가능 저장 매체를 포함하며, 상기 명령은 다음을 포함한다:

[0439] (a) 전술한 시료 에서 태아 및 모체 핵산에 대한 서열 정보를 획득하기 위한 코드;

[0440] (b) 염색체 1-22, X, 그리고 Y로부터 선택된 임의의 하나 또는 그 이상의 관심 염색체 각각에 대하여 태아 및 모체 핵산으로부터 서열 태그 수를 계산적으로 식별하기 위하여 전술한 서열 정보를 이용하고, 그리고 관심대상의 임의의 하나 또는 그 이상의 염색체의 전술한 임의의 하나 또는 그 이상의 세그먼트 각각에 대한 최소한 하나의 정규화 세그먼트 서열에 대한 서열 태그 수를 확인하기 위한 코드;

[0441] (c) 그리고 임의의 하나 또는 그 이상의 관심 염색체 각각에 대한 전술한 하나 또는 그 이상의 세그먼트에 대하여 확인된 서열 태그의 수, 그리고 임의의 하나 또는 그 이상의 관심 염색체의 전술한 임의의 하나 또는 그 이상의 세그먼트 각각에 대한 단일 염색체 세그먼트 분량을 산출하기 위하여 전술한 정규화 세그먼트 서열에 대하여 확인된 서열 태그의 수를 이용하는 코드; 그리고

[0442] (d) 임의의 하나 또는 그 이상의 관심 염색체의 전술한 임의의 하나 또는 그 이상의 세그먼트들 각각에 대한 전술한 단일 염색체 세그먼트 분량 각각을 임의의 하나 또는 그 이상의 관심 염색체의 전술한 임의의 하나 또는 그 이상의 염색체 세그먼트들 각각에 대한 대응하는 임계치 값에 비교하고, 그리고 그렇게 함으로써 전술한 시료에서 하나 또는 그 이상의 상이한 부분적 태아 염색체 홀배수체의 존재 또는 부재를 결정하는 코드

[0443] 일부 구체예들에 있어서, 단일 염색체 세그먼트 분량을 산출하기 위한 코드는 염색체 세그먼트들중 선택된 하나에 대한 염색체 세그먼트 분량을 선택된 염색체 세그먼트에 대하여 확인된 서열 태그의 수와 선택된 염색체 세그먼트에 대하여 대응하는 정규화 세그먼트 서열에 대하여 확인된 서열 태그의 수의 비율로 산출하는 코드를 포함한다.

[0444] 일부 구체예들에 있어서, 상기 시스템은 임의의 하나 또는 그 이상의 관심 염색체의 임의의 하나 또는 그 이상의 세그먼트들의 임의의 나머지 염색체 세그먼트들 각각에 대한 염색체 세그먼트 분량의 산출을 반복하기 위한 코드를 더 포함한다.

[0445] 일부 구체예들에 있어서, 상기 시스템은 (i) 상이한 모체 개체로부터 테스트 시료에 대하여 (a)-(d)를 반복하는 코드, 그리고 (ii) 전술한 각 시료에서 임의의 하나 또는 그 이상의 상이한 부분적 태아 염색체 홀배수체의 존재 또는 부재를 결정하는 코드를 더 포함한다.

[0446] 본 명세서에서 제공되는 시스템중 임의의 다른 구체예들에서, 상기 코드는 모체 테스트 시료를 제공하는 인간

개체를 위하여 환자 의무 기록에서 (d)에서 결정된 바와 같이 태아 염색체 홀배수체의 존재 또는 부재를 자동으로 기록하는 코드를 더 포함하고, 여기에서 기록은 프로세서를 이용하여 실행된다.

[0447] 본 명세서에서 제공되는 시스템중 임의의 다른 구체예들에서, 상기 서열화기는 차세대 서열화 (NGS)를 실행하도록 설정된다. 일부 구체예들에 있어서, 상기 서열화기는 가역적 염료종료물질들과 함께, 합성에 의한 서열화를 이용하여 대량 병행 서열화를 시행하도록 설정된다. 다른 구체예들에서, 상기 서열화기는 결찰에 의한 서열화를 시행하도록 설정된다. 여전히 다른 구체예들에 있어서, 상기 서열화기는 단일 분자 서열화를 실행하도록 설정된다.

[0448] **실험**

[0449] **실시예 1**

[0450] **일차 및 보강된 서열화 라이브러리의 준비 및 서열화**

[0451] *a. 서열화 라이브러리의 준비 - 단축 프로토콜 (ABB)*

[0452] 모든 서열화 라이브러리, 가령, 일차 라이브러리 및 보강된 라이브러리는 모체 혈장으로부터 추출된 대략적으로 2 ng의 정제된 cfDNA로부터 준비되었다. 라이브러리 준비는 다음과 같이, Illumina®의 NEBNext™ DNA Sample Prep DNA Reagent Set 1 (Part No. E6000L; New England Biolabs, Ipswich, MA) 시약들을 이용하여 실행되었다. 무-세포 혈장 DNA는 자연상태에서 단편화되기 때문에, 혈장 DNA 시료의 분무 또는 초음파분쇄에 의한 추가 단편화는 실행되지 않았다. 40 μl 에서 포함된 대략적으로 2 ng의 정제된 cfDNA 단편의 오버행은 NEBNext® End Repair Module에 따라 1.5 ml 마이크로원심분리 튜브에서 상기 cfDNA와 함께 5μl 10X 포스포릴화 완충액, NEBNext™ DNA Sample Prep DNA Reagent Set 1에서 제공되는 2 μl 데옥시뉴클레오티드 용액 믹스 (10 mM의 각 dNTP), 1μl의 1:5 희석의 DNA 중합효소 I, 1 μl T4 DNA 중합효소 그리고 1 μl T4 폴리뉴클레오티드 키나제와 함께 15분 동안 20℃에서 항온처리함으로써 포스포릴화된 블런트 말단으로 전환되었다. 그 다음 이들 효소들은 상기 반응 혼합물을75℃에서 5 분 동안 항온처리함으로써 열 비활성화되었다. 상기 혼합물은 4℃로 냉각되었고, 그리고 블런트-말단 DNA의 dA 꼬리화는 Klenow 단편 (3' 에서 5' 엑소 마이너스) (NEBNext™ DNA Sample Prep DNA Reagent Set 1)가 포함된 10μl의 dA-테일링 마스터 믹스를 이용하고, 그리고 15 분 동안 37℃에서 항온처리함으로써 성취되었다. 후속적으로, 상기 Klenow 단편은 상기 반응 혼합물을 75℃에서 5 분 동안 항온처리함으로써 열불활성화되었다. Klenow 단편의 불활성화 후, 1 μl의 1:5 희석물의 Illumina 게놈 어뎁터 올리고 믹스 (Part No. 1000521; Illumina Inc., Hayward, CA)를 이용하여 Ilumina 어뎁터들 (Non-Index Y-Adaptors)을 NEBNext™ DNA Sample Prep DNA Reagent Set 1에서 제공되는 4 μl의 T4 DNA 리게이즈를 이용하여 dA-꼬리의 DNA에 결찰시키고, 상기 반응 혼합물을 15 분 동안 25℃에서 항온처리하였다. 상기 혼합물은 4℃로 냉각시키고, 그리고 어뎁터-결찰된 cfDNA는 Agencourt AMPure XP PCR 정제 시스템 (Part No. A63881; Beckman Coulter Genomics, Danvers, MA)에서 제공되는 자석 비드를 이용하여 결찰안된 어뎁터들, 어뎁터 이량체들, 그리고 다른 시약들로부터 정제되었다. Phusion ® High-Fidelity Master Mix (25μl; Finnzymes, Woburn, MA) 및 상기 어뎁터에 상보적인 Illumina's PCR 프라이머 (0.5 μM 각) (Part No. 1000537 및 1000537)를 이용하여 어뎁터-결찰된 cfDNA (25 μl)를 선택적으로 농축시키기 위하여 18회 PCR이 실행되었다. 상기 어뎁터-결찰된 DNA는 Illumina 게놈 PCR 프라이머 (Part Nos. 100537 및 1000538) 그리고 NEBNext™ DNA Sample Prep DNA Reagent Set 1에서 제공되는 Phusion HF PCR Master Mix를 이용하여 제조업자의 지시에 따라 PCR (98℃에서 30 초; 98℃에서 10 초, 65℃에서 30 초, 그리고 72℃에서 30초를 18회; 72℃에서 5 분 동안 최종 연장, 그리고 4℃에서 유지)을 거쳤다. 상기 증폭된 산물은 www.beckmangenomics.com/products/AMPureXPProtocol_000387v001.pdf에서 이용가능한 제조업자의 명령에 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 농도 및 크기 분포는 Agilent DNA 1000 Kit for the 2100 Bioanalyzer (Agilent technologies Inc., Santa Clara, CA)을 이용하여 분석되었다.

[0453] *b. 서열화 라이브러리의 준비 -전장 프로토콜*

[0454] 본 명세서에서 전장 프로토콜은 기본적으로 Illumina에서 제공하는 표분 프로토콜이며, 그리고 증폭된 라이브러리의 정제에서 Illumina 프로토콜과 단지 상이하다. 본 명세서에서 설명된 프로토콜은 동일한 정제 단계를 위하여 자석 비드를 사용하지만, 상기 Illumina 프로토콜은 증폭된 라이브러리를 겔 전기영동을 이용하여 정제하도록 지시한다. 모체 혈장으로부터 추출된 대략적으로 2 ng의 정제된 cfDNA는 제조업자의 지시에 기본적으로

따라 Illumina®의 NEBNext™ DNA Sample Prep DNA Reagent Set 1 (Part No. E6000L; New England Biolabs, Ipswich, MA)를 이용하여 일차 서열화 라이브러리를 준비하는데 이용된다. Agencourt 자석 비드와 정제 컬럼 대신 시약들을 이용하여 실행되는 어뎁터-결찰된 제품의 최종 정제를 제외한 모든 단계는 Illumina® GAII를 이용하여 서열화된 게놈 DNA라이브러리를 위한 시료 준비용 NEBNext™ 시약들을 수반하는 프로토콜에 따라 실행된다. NEBNext™ 프로토콜은 기본적으로 Illumina에서 제공되는 것을 따르며, 이는 grcf.jhml.edu/hts/프로토콜/11257047_ChIP_Sample_Prep.pdf에서 이용가능하다.

[0455] NEBNext® End Repair Module에 따라, 열 순환기 에서 200 μl 마이크로원심분리 튜브에서 40μl cfDNA를 5μl 10X 포스포릴화 완충액, 2 μl 데옥시뉴클레오티드 용액 믹스 (10 mM 각 dNTP), 1 μl의 1:5 희석액의 DNA 중합효소 I, 1 μl T4 DNA 중합효소 그리고 NEBNext™ DNA Sample Prep DNA Reagent Set 1에서 제공되는 1 μl T4 폴리뉴클레오티드 Kinase를 30 분 20℃에서 항온처리함으로써, 40 μl 에서 포함된 대략적으로 2 ng 정제된 cfDNA 단편은 포스포릴화된 블런트 단부로 전환되었다. 이 시료는 4℃로 냉각시키고, 그리고 QIAQuick PCR 정제 키트 (QIAGEN Inc., Valencia, CA)에서 제공되는 QIAQuick 컬럼을 이용하여 다음과 같이 정제되었다. 50 μl 반응물은 1.5 ml 마이크로원심분리 튜브로 옮겨지고, 그리고 250 μl의 Qiagen 완충액 PB가 추가되었다. 생성된 300 μl는 QIAquick 컬럼으로 옮겨졌고, 마이크로원심분리기에서 13,000 RPM에서 1 분동안 원심분리되었다. 상기 컬럼은 750 μl Qiagen 완충액 PE로 세척되었고, 그리고 재-원심분리되었다. 잔류 에탄올은 5 분 동안 13,000 RPM에서 추가 원심분리에 의해 제거되었다. 상기 DNA는 39 μl Qiagen 완충액 EB으로 원심분리에 의해 용리되었다. 34 μl의 블런트-말단의 DNA에 dA 꼬리붙이기는 Klenow 단편 (3' 에서 5' 엑소 마이너스) (NEBNext™ DNA Sample Prep DNA Reagent Set 1)가 포함된 dA-테일링 마스터 믹스 16 μl을 이용하고, 그리고 제조업자의 NEBNext® dA-Tailing Module에 따라 30 분 동안 37℃에서 항온처리함으로써 이루어졌다. 이 시료는 4℃로 냉각되었고, 그리고 MinElute PCR 정제 키트 (QIAGEN Inc., Valencia, CA)에서 제공되는 컬럼을 이용하여 다음과 같이 정제되었다. 50 μl 반응물은 1.5 ml 마이크로원심분리 튜브로 옮겨지고, 그리고 250 μl의 Qiagen 완충액 PB가 추가되었다. 300 μl는 MinElute 컬럼으로 옮겨졌고, 마이크로원심분리기에서 13,000 RPM에서 1 분동안 원심분리되었다. 상기 컬럼은 750 μl Qiagen 완충액 PE로 세척되었고, 그리고 재-원심분리되었다. 잔류 에탄올은 5 분 동안 13,000 RPM에서 추가 원심분리에 의해 제거되었다. 상기 DNA는 15 μl Qiagen 완충액 EB에서 원심분리에 의해 용리되었다. NEBNext® Quick Ligation Module에 따라 10 μl의 DNA 용출물은 1 μl의 1:5 희석액의 Illumina 게놈 어뎁터 올리고 믹스 (Part No. 1000521), 15 μl의 2X Quick 결찰 반응 완충액, 그리고 4 μl Quick T4 DNA 리게이즈와 함께 15 분 동안 25℃에서 항온처리되었다. 이 시료는 4℃로 냉각되었고, 그리고 MinElute 컬럼을 이용하여 다음과 같이 정제되었다. 150 μl의 Qiagen 완충액 PE를 상기 30 μl 반응물에 추가하였고, 그리고 상기 전체 용적은 MinElute 컬럼으로 옮겨졌으며, 마이크로원심분리기에서 13,000 RPM에서 1 분동안 원심분리되었다. 상기 컬럼은 750 μl Qiagen 완충액 PE로 세척되었고, 그리고 재-원심분리되었다. 잔류 에탄올은 5 분 동안 13,000 RPM에서 추가 원심분리에 의해 제거되었다. 상기 DNA는 28 μl Qiagen 완충액 EB에서 원심분리에 의해 용리되었다. 23 μl의 어뎁터-결찰된 DNA 용출물은 Illumina 게놈 PCR 프라이머 (Part Nos. 100537 및 1000538) 그리고 Phusion HF PCR Master Mix provided in the NEBNext™ DNA Sample Prep DNA Reagent Set 1에서 제공되는 Phusion HF PCR Master Mix를 이용하여 제조업자의 지시에 따라 18회의 PCR (98℃에서 30 초; 98℃에서 10 초, 65℃에서 30 초, 그리고 72℃에서 30초를 18회; 72℃에서 5 분 동안 최종 연장, 그리고 4℃에서 유지)을 거쳤다. 상기 증폭된 산물은 www.beckmangenomics.com/products/AMPureXPProtocol_

000387v001.pdf에서 이용가능한 제조업자의 명령에 따라 Agencourt AMPure XP PCR 정제 시스템 (Agencourt Bioscience Corporation, Beverly, MA)을 이용하여 정제되었다. Agencourt AMPure XP PCR 정제 시스템은 혼입 안된 dNTPs, 프라이머, 프라이머 이량체들, 염 및 다른 오염물질을 제거하고, 그리고 100 bp 이상의 앰플리콘을 회수한다. 상기 정제된 증폭된 산물은 40 μl의 Qiagen EB 완충액에서 Agencourt 비드로부터 용리되었고, 그리고 라이브러리의 크기 분포는 2100 Bioanalyzer (Agilent technologies Inc., Santa Clara, CA)용 Agilent DNA 1000 Kit를 이용하여 분석되었다.

[0456] *c. 단축 (a) 프로토콜과 전장 (b) 프로토콜에 따라 준비된 서열화 라이브러리의 분석*

[0457] Bioanalyzer에 의해 생성된 전지영동도는 도 9a 및 9b에 나타낸다. 도 9a는 (a)에서 설명된 전장 프로토콜을 이용하여 혈장 시료 M24228로부터 정제된 cfFNA로부터 준비된 라이브러리 DNA의 전기영동도이며, 그리고 도 9b는 (b)에서 설명된 전장 프로토콜을 이용하여 혈장 시료 M24228로부터 정제된 cfDNA로부터 준비된 라이브러리 DNA의 전기영동도를 보여준다. 두 도면 모두에서, 피크 1과 4는 차례로 15 bp 하위 표지, 그리고 1,500 상위 표지를 나타내며; 피크 이상의 수는 라이브러리 단편의 이동 시간을 나타내고; 그리고 수평선은 통합을 위한 설

정 임계치를 나타낸다. 도 9a에서 전기영동도는 187 bp 단편의 작은 피크를 그리고 263 bp의 주요 피크를 나타내며, 도 9b에서 전기영동도는 265 bp에서 오직 하나의 피크만을 보여준다. 피크 면적의 통합으로 도 9a에서 187bp의 DNA의 산출된 농도 0.40 ng/μl, 도 9a에서 263bp 피크 DNA의 농도 7.34 ng/μl, 그리고 도 9b에서 265 bp 피크 DNA의 농도 14.72 ng/μl를 결과하였다. cfDNA에 결찰된 Illumina 어뎁터들은 265 bp로부터 차감될 때 92 bp으로 공지되며, 이는 cfDNA의 피크 크기가 173 bp 임을 나타낸다. 187 bp에서 작은 피크는 결찰된 말단에서 말단으로 결찰된 2개의 프라이머의 단편을 나타낼 가능성이 있다. 상기 선형의 2개-프라이머 단편은 단축 프로토콜이 이용될 때 최종 라이브러리 산물로부터 제거된다. 상기 단축 프로토콜은 187 bp 미만의 다른 더 작은 단편들을 또한 제거한다. 이 실시예에서, 정제된 어뎁터-결찰된 cfDNA의 농도는 전장 프로토콜을 이용하여 만든 어뎁터-결찰된 cfDNA의 배가 된다. 어뎁터-결찰된 cfDNA 단편의 농도는 전장 프로토콜을 이용하여 획득된 것보다 항상 더 크다는 것을 주지해야 한다(데이터 제시하지 않음).

[0458] 따라서, 단축 프로토콜을 이용하여 서열화 라이브러리를 준비하는 장점은 획득된 라이브러리는 일관되게 262-267 bp 범위의 오직 하나의 주요 피크만으로 구성되며, 한편 전장 프로토콜을 이용하여 준비된 라이브러리의 품질은 cfDNA를 나타내는 것보다는 피크의 수와 이동성이 반영되어 가변적이다. 비-cfDNA 제품은 플로우 쎌 상에 공간을 차지할 수 있고, 클러스터 증폭의 품질을 감소시키고, 후속적으로 서열화 반응물의 영상화를 감소시키고, 이는 홀배수체 상태의 전반적인 할당의 근간이 된다. 단축 프로토콜은 상기 라이브러리의 서열화에 영향을 주지 않는 것으로 나타났다.

[0459] 단축 프로토콜을 이용하여 서열화 라이브러리를 준비하는 또다른 장점은 블런트-말단화의 3개 효소적 단계, d-A 꼬리붙이기, 그리고 어뎁터-결찰이 신속한 홀배수체 진단 서비스의 확증 및 시행을 뒷받침하는 것이 종료될 때까지 1시간이 채 안 걸린다는 점이다.

[0460] 또다른 장점은 블런트-말단화의 3개 효소적 단계, d-A 꼬리붙이기, 그리고 어뎁터-결찰이 동일한 반응 튜브에서 실행되며, 따라서 다수의 시료 이동을 피할 수 있는데, 시료 이동은 잠재적으로 물질의 상실, 그리고 더 중요한 것은 시료의 섞임 및 오염의 가능성으로 이어진다.

[0461] **실시예 2**

[0462] **쌍태 임신에서 정확한 홀배수체 탐지**

[0463] *개요*

[0464] 전체-게놈 대량 병행 서열화를 이용하여 총 무 세포 DNA (cfDNA)의 비-침습성 출생전 테스트 (NIPT)는 태아 염색체 홀배수체를 탐지하는 매우 정확하고 강건한 방법으로 드러났다. 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; 57:1042-9 참고. 본 테스트는 단일 모체 혈액 시료로부터 삼염색체성 21, 18, 13 및 성염색체 홀배수체를 탐지한다. 본 테스트는 현재 10+ 주에서 외둥이 임산부, 그리고 태아 홀배수체 고위험 임산부를 위한 것이다. 최근, the American College of Obstetricians and Gynecologists (ACOG), the International Society for Prenatal Diagnosis (ISPD), the American College of Medical Genetics and Genomics (ACMG) 및 the National Society of Genetic Counselors (NSGC)는 태아 홀배수체의 고위험군 여성을 위하여 NIPT의 사용을 고려할 것을 권고하였다.

[0465] 미국에서, 쌍태는 30명의 생존 출생중 대략적으로 1명꼴로 계수되며, 쌍태 출산 비율은 증가추세에 있다 (National Center for Health Statistics Data Brief, No. 80, January 2012). 여성이 나이들 수록 월경 주기당 하나 이상의 난자를 배출할 가능성이 더 크고, 그렇기 때문에 30세 이상의 여성의 쌍태 임신 가능성이 약 1/3 증가되는 것으로 계수된다. 시험관 수정 동안 흔히 하나 이상의 배아가 이식되는, 지원된 생식 기술은 쌍태 임신에서 잔존적 증가의 대부분을 설명한다.

[0466] 예비 증거는 모체 순환계에 있는 태아 DNA의 양은 단태 임신과 비교하였을 때 쌍태 임신에서 대략적으로 35% 증가된다고 제시하지만, 연구는 각 태아로부터 유도된 cfDNA의 양을 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. 순환하는 태아 DNA in 쌍태 임신에서 순환

하는 태아 DNA의 양이 전반적으로 증가되지만, 각 태아의 cfDNA의 양은 감소된다고 연구자들은 설명하였다. 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. 따라서, 쌍태 임신에서 홀배수체의 정확한 분류를 확실히 하기 위한 민감한 방법들이 요구된다.

[0467] 홀배수체 시료들을 정확하게 분류하는 NIPT 능력을 최대화하는 인자들은 분석에 이용된 서열화 리드의 수가 증가되어, 따라서 통계학적 잡음이 최소화되고, 그리고 운용간 변동성이 감소되도록 염색체 신호를 정규화시키는 능력이다. 최근, 출원인은 시료당 이용가능한 리드 수를 증가시키는 개선된, 자동화된 시료 준비 작업흐름과 홀배수체 염색체의 특이적 신호를 증가시키는 개선된 분석학적 방법을 개발하였다. 이러한 보강은 홀배수체 영향을 받은 시료들을 분류하는 전반적인 정확성을 개선시킨다.

[0468] 본 실시예는 현재까지 이용된 최대 쌍태 확증 코호트에 개선된 분류 알고리즘을 적용하는 것을 설명한다. 개선된 SAFeR (Selective Algorithm for Fetal Results) 알고리즘은 쌍태 시료들에서 태아당 무세포 DNA의 양이 감소된 것으로 공지된 홀배수체의 정확한 탐지를 허용한다는 것을 우리는 설명한다.

[0469] *방법들*

[0470] 고위험 모체 집단과 평균 위험 모체 집단 모두 관련된 2개의 독립 임상 연구의 일부분으로 시료들이 수집되었다. MatErnal BLood IS Source to Accurately Diagnose Fetal Aneuploidy study (MELISSA; NCT01122524)는 고위험 임산부에서 전체 염색체 홀배수체를 탐지하도록 기획되었다. 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 trial (CARE; NCT01663350)는 평균 위험 모체 집단에서 삼염색체성 21과 삼염색체성 18에 대하여 통상적인 출생전 혈청 스크리닝 방법들과 비교하여 본 테스트의 우수한 특이성을 설명하기 위하여 기획되었다(공개를 위하여 제출됨). 데이터 설정의 세부 사항은 표 2에 제시된다. 임상 결과는 출생전 침습적 과정의 핵형 또는 신생아 신체 검사에 의해 결정되었다.

표 2

표 2: **쌍태 시료들의 핵형 및 즉각적 분류.** 염색체 21, 18 및 13 의 홀배수체 및 Y 염색체의 존재에 대한 출생전 즉석 테스트를 이용하여 118 명의 쌍태 임산부들의 모체 시료들이 분석되었다. 즉석 데이터는 핵형 분석 또는 신생아 신체 검사에 의해 결정된 임상 결과와 비교되었다.

연구된 수	태아 1	태아 2	즉석 홀배수체 분류	즉석 염색체 Y 분류
24	46,XX	46,XX	영향없음	탐지 안됨
48	46,XX	46,XY	영향없음	Y 탐지됨
42	46,XY	46,XY	영향없음	Y 탐지됨
2	47,XY,+21	46,XY	T21 영향을 받음	Y 탐지됨
1	Mos 47,XY,+21[7]/46,XY[11]	46,XX	T21 영향을 받음	Y 탐지됨
1	47,XY,+ 18	47,XY,+18	T18 영향을 받음	Y 탐지됨

[0471]

[0472] 무-세포 DNA는 냉동된 혈장 시료들로부터 추출되었고, 이미 설명한 바와 같이, HiSeq2000 서열화기에서 서열화되었다. 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 병행 서열화 (MPS) 서열 태그는 인간 게놈 참조 빌드 hg19에 매핑되었고, 정규화된 염색체 값 (NCVs)은 신호대 잡음비를 최대화시키고, 탐지의 전반적인 민감성을 개선시켰던, 개선된 분석 작업 흐름을 이용하여 염색체 21, 18, 13, X와 Y에 대하여 산출되었다. 알고리즘 성분들은 개선된 게놈 필터링, 분자 생물학 단계들을 통하여 도입되는 조직적 편향 제거 및 개선된 정규화 및 분류 방법들을 포함하였다. 서열화를 실행하는 실험실 직원들은 임상 결과에 대하여 모르게하였다.

[0473] *결과*

[0474] 임상적으로 정의된 결과를 가진 118명의 쌍태 임산부로부터 모체 혈장 시료들을 본 연구에서 조사하였다 (표 2). 본 연구의 모든 시료들에 대하여 염색체 21, 18 및 13에 대한 홀배수체 분류가 생성되었고, 하나 또는 그 이상의 홀배수체 태아를 가진 임산부의 4개 시료가 정확하게 식별되었다 (도 10). 이들 시료중 2개는 두융모막 쌍태 쌍으로부터 온 것이며, 각각 하나는 T21 영향을 받은 남성 태아이며, 하나는 비-영향을 받지 않은 남성 태아 (47,XY+21/46,XY)이고; 하나는 홀융모막 쌍태 시료로써 47,XY+18 핵형을 가지고; 그리고 하나의 시료는 두융모막 쌍태이며, 이때 하나의 쌍태는 모자이크 핵형 47,XY+T21[7]/46,XY[11]을 가졌다. 본 연구에서 임상적으로-정의된 영향을 받지 않은 시료들 (N=114)중 홀배수체에 있어서 영향을 받은 것으로 분류된 것은 없었다.

[0475] 태아의 성별은 cfDNA에서 Y 염색체의 존재에 의해 결정될 것이다. 본 명세서에서 공개된 테스트는 최소한 하나의 남성 태아를 가진 모든 시료에서 Y 염색체의 존재를 양성적으로 식별해낼 수 있었다 (도 10). 더욱이, 상기 테스트는 또한 2개의 여성 태아를 가진 시료 에서 Y 염색체의 부재를 정확하게 식별하였다.

[0476] *결론*

[0477] 현 연구는 쌍태 시료들의 가장 민감한 상염색체 홀배수체 테스트를 가능하게 하는 개선된 분석 방법을 설명한다. 강화된 분석 방법은 게놈 필터링, 조직적 잡음 감소에서 개선점을 가지고, 개선된 분류 방법들을 가진다. 개선된 분석 작업흐름의 유용성은 118개의 쌍태 시료들 세트에서 설명되었으며; 상염색체 홀배수체를 탐지하고, 쌍태에서 Y 염색체의 존재를 탐지하기 위하여 임의의 MPS 확증에 최대 시료수가 이용되었다 (도 11). 도 11은 NIPT 연구에서 분석된 쌍태 시료들을 나타낸다. 상업적으로 이용가능한 NIPT 테스트의 수행을 평가하기 위하여 다양한 연구에서 쌍태 시료가 이용되었다. 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. 상기 개선된 분석 방법은 임의의 가양성 결과의 생성 없이, 삼염색체성 21에 대하여 모자이크인 영향을 받은 태아가 포함된, 코호트에서 삼염색체성 21 및 삼염색체성 18 시료들의 존재를 정확하게 탐지함으로써 정확하게 실행되는 것을 보여주었다. 추가적으로, 개선된 분석 방법은 최소한 하나의 남성 태아를 가진 모든 쌍태 임신에서 Y 염색체의 존재를 정확하게 탐지하였고, 그리고 2개의 여성 태아를 갖는 임의의 쌍태 임신에서 Y 염색체는 탐지되지 않았다.

[0478] 감각적 방법의 한 가지 특징은 조직적 잡음을 최소화시키고, 전반적인 신호대 잡음비를 증가시키는 능력이다. 현 연구는 임의의 다른 상업적으로-이용가능한 NIPT 분석보다 시료당 더 많은 서열화 리드를 생성하고 (대략적으로 28M 서열화 리드/시료) 그리고 복합 DNA 시료들의 생화학적 작동에 수반되는 조직적 잡음을 더 잘 취급하는 분석 방법을 개선시킴으로써 성취되었다. 개선된 분석 작업흐름은 궁극적으로 정규화된 염색체 계수 분포의 폭을 감소시킴으로써, 영향을 받지 않은 집단과 영향을 받은 집단을 더 잘 분리하게 하고, 그리고 소량의 태아 DNA로, 홀배수체 영향을 받은 태아를 정확하게 식별해내는 능력이 개선되었다.

[0479] 쌍태 임신에서 홀배수체를 탐지하는 매우 정확하고, 민감한 방법을 가지는 능력은 쌍태 임신에서 무 세포 태아 DNA의 총량은 비록 증가하지만, 각 태아에 기인하는 양은 감소하기 때문에 중요하다. 따라서, A) 단태 임신에 등가이거나 가음성 결과 확률(likelihood)이 증가되면 이러한 발견 및 테스트 시료들을 무시할 수 있고, B) 불충분한 DNA로 인하여 시료 수의 증가를 거부할 수 있고 또는 C) 더 민감한 방법을 구축할 수 있다 (표 3).

표 3

표 3: 상업적으로 이용가능한 NIPT 테스트를 이용한 쌍태 임신 처리 전략

	옵션	결과
A	존재하는 cfDNA 가 단태 임신과 동일한지에 대한 쌍태 임신 테스트.	가음성 확률 증가
B	테스트 쌍태 임신을 테스트하기 위하여 현재 방법 이용.	불충분한 DNA 로 시료 거부
C	개별 cfDNA 농도에 더욱 민감한 개선된 방법 이용	쌍태 및 더 적은 가음성을 가진 낮은 수준의 단태에 대한 더 정확한 테스트

[0480]

[0481] SAFeR™ 알고리즘에 대한 분석 개선은 쌍태 임신에서 홀배수체 분류의 정확하게 하는 것 이상으로 확장된다. 영향을 받지 않은 집단과 영향을 받은 집단의 개선된 분리는 의심되는 홀배수체로 분류되는 시료들의 전반적인 빈도를 또한 감소시킨다. 추가적으로, 개선된 분석 작업흐름은 홀배수체 탐지 및 성별분류에서 유사한 개선으로 단태 임신에 적용될 수 있다.

[0482] 결론적으로, 현재 연구는 홀배수체 영향을 받지 않은 시료와 영향을 받은 시료를 더 잘 분리하고, 소량의 태아 DNA가 포함된 시료들로부터 상 염색체 홀배수체 분류를 좀더 정확하게 할 수 있는 개선된 분석 방법을 설명한다. 이러한 개선을 통합함으로써, 출생전 테스트의 능력은 쌍태 임신 테스트로 확장되었다.

[0483] **실시예 3: 증후군 특이적 조직적 편향 제거 일정 (SSS-BER)**

[0484] *개요*

[0485] 다양한 출생전 비침습성 진단 (NIPD) 방법들은 모체 유체, 이를 테면 말초 혈액에서 이용가능한 태아 기원의 cfDNA를 이용한다. 많은 NIPD 방법들은 임산부의 태아의 cfDNA가 질환들 또는 표현형과 관련된 유전적 서열들에서 복제 수 변이를 품고 있는지를 판단하기 위하여 모체 말초 혈액으로부터 cfDNA를 추출하고, 서열화하고, 그리고 정렬한다. 상기 추출된 그리고 서열화된 cfDNA는 서열 리드을 제공하며, 그 다음 참조 게놈에 매핑된다. 상기 참조 게놈들 상의 독특한 위치 또는 부위에 매핑된 서열 리드는 서열 태그로 지칭된다. 관심 대상 서열에 매핑된 서열 태그 수를 이용하여 상기 관심 대상 서열의 복제 수 또는 복제 수 변이를 결정할 수 있다.

[0486] 관심 대상 서열에 매핑된 서열 태그 수는 커버리지로 지칭된다. 유전적 서열의 빈 또는 영역에 대한 커버리지은 또다른 영역에 대한 하나의 영역의 상대적인 존재도(abundance) 또는 다른 시료에 대한 하나의 시료의 상대적 존재도를 계산하기 위한 데이터를 제공한다. 관심 대상 서열의 커버리지이 비정상적으로 낮거나 또는 높을 때, 서열의 복제 수 변이를 암시할 수 있다. 다양한 유전적 질환들은 복제 수 변이와 연관된다. 예로써, 6개의 유전적 질환들 뿐만 아니라 이들의 관련된 아염색체성 서열들이 표 4에 열거된다.

표 4

표 4: 1 및 12 plex 테스트 세트에서 증후군 비율 CV.

[0487]

증후군	Chr	시작	종료	12plex	1plex
1p36	1	0	5.00E+06	1.13%	1.07%
WolfHirschhorn	4	0	3573000	1.14%	0.76%
CriduChat	5	0	9800000	1.19%	0.79%
Williams	7	72701000	74285000	1.77%	1.47%
Angelman	15	22699000	28520000	1.57%	1.32%
DiGeorge	22	18891000	21561000	1.16%	0.96%

[0488] 복제 수 변이를 결정하는 신호는 다양한 인자들에 의해 영향을 받는다. 예로써, 주어진 시료에서 심화(deeper) 서열화는 관심 대상 서열 상에서 각 빈 또는 영역에 대한 더 많은 리드 및 태그를 제공하고, 이 시료 측정에서 변이를 낮추고, 이는 다시 복제 수 변이 결정을 위한 잡음을 감소시키거나 및/또는 신호를 증가시킨다. 더욱이, 상기 관심 대상 서열의 길이 또한 유사한 방식으로 신호에 영향을 주는데, 관심 대상 서열이 더 길수록, 이 시료로부터 더 많은 서열 태그가 상기 관심 대상 서열에 매핑되기 때문이다. 더욱이, 태아 분획, 가령, 엄마와 태아 모두의 cfDNA에 대한 태아의 cfDNA 비율은 태아에서 CNV를 결정하는 신호에 영향을 준다. 태아 분획이 더 높을 수록, 태아 DNA에서 동일한 변화로 인하여 혼합된 cfDNA에서 관찰되는 변화의 정도는 더

클 수 있다.

[0489] 본 명세서에서 설명된 그리고 기존에 이용가능한 일부 방법들은 전체 염색체 또는 염색체의 세그먼트들의 복제 수 변이를 탐지하는데 적절하다. 그러나, 더 작은 유전적 서열들 스트레취와 관련된 유전적 질환들의 경우, 기존 방법의 신호대 잡음비는 일반적으로 너무 낮아서 복제 수 변이의 믿을만한 탐지를 허용하지 않는다. 예로써, 표 4에 나타낸 6가지 유전적 증후군의 경우, 증후군 비율의 변이계수 (CV)는 12-plex 서열화를 이용하면 1.13% 내지 1.77% 범위가 된다. 증후군 비율은 문제의 증후군과 연관된 서열 영역에서 서열 태그의 커버리지에 근거한 계측이며, 이는 하기에서 추가 설명된다. 단일 서열화의 이용은 증후군 비율에서 변이계수를 감소시키고, 신호대 잡음비를 증가시킨다. 표 4의 가장 우측 컬럼에 나타낸 바와 같이, CV는 0.076% 내지 1.47% 범위가 된다. 그러나, 일부 연구에 따르면, 증후군과 연관된 서열들의 복제 수 변이를 신뢰적으로 결정하기 위해서 0.7% 또는 더 낮은 CV를 요구한다. 따라서, 신호를 증가시키고 및/또는 잡음을 감소시키는 방법들을 개발해야할 필요가 있다. 한 가지 접근방식은 데이터에서 상기 관심 대상 서열의 무관한 복제 수 변이인 잡음을 감소시킨다. 잡음은 서열화 편향으로 간주될 수 있다.

[0490] 서열 편향에 기여하는 몇 가지 인자들이 조사되었다. GC 함량, 서열화 리드의 매핑능력(mappability), 그리고 로컬 구조에 의해 생성될 수 있는 지역적 편향. 예를 들면, 헤테로크로마틴(heterochromatin) 단편은 유크로마틴(euchromatin)과 비교하여 상이하게 시료 준비/서열화 분석을 겪을 수 있다. 뉴클레오티드 조성물로 인한 더 작은 규모의 구조적 차이 또한 DNA를 손상에 대하여 차등적으로 민감하게 하도록 할 수 있다. 이러한 크로마틴 또는 DNA 구조적 효과는 커버리지 불균질성(inhomogeneity)에 기여하고, 자체가 정규화된 커버리지의 조직적 편차에서 1로부터 멀어지게 될 수 있다.

[0491] 대조 시료들과 테스트 시료들 모두에서 변이가 존재한다. 이들 일부 변이는 증후군-관련된 서열들과 연관된다. 기타 변이는 전체 게놈에서 일반적이며, 증후군-관련된 서열에 특이적이지 않다. 관심 대상 서열의 복제 수 변이에 기여하지 않는 이들 변이의 제거로 잡음 감소 및 신호대 잡음비의 증가를 도울 것이다. 아래 첫 단락은 증후군 관련된 영역의 안과 밖에 있는 빈 사이의 일관적인 변이가 포함된 영향을 받지 않은 시료들의 변이에 기초하여 테스트 시료의 게놈을 통한 변이 제거를 설명한다. 이 변이는 또한 증후군 특이적 조직적 편향으로 지칭된다. 상기 게놈에 일반적이거나 또는 테스트 시료에 특이적이지만, 그러나, 증후군 관련된 영역과 무관한 변이 제거가 또한 설명된다. 일반적 변이의 제거는 증후군 특이적 조직적 편향 제거에 앞서 우선적으로 적용될 수 있다. 그러나, 이 내용은 후자를 강조하는데, 따라서 증후군-특이적 조직적 편향 제거 일정 (SSS-BER)으로 하기에서 우선적으로 설명된다.

[0492] *증후군-특이적 조직적 편향 제거 일정 (SSS-BER)*

[0493] *동기*

[0494] 아염색체성 데이터 분석에서 데이터, 가령 데이터 동일한 시약 로트로 가공된 주어진 플로우 쎌/플레이트 또는 시료들의 집합, 등등 로부터 벗어난(coming off) 데이터의 상이한 "배취" 간에 정규화된 커버리지 패턴의 차이로 나타나는 조직적 편향이 드러났다. 이들 시료-, 시약-, 분석-의존적 실험 변이에 의해 도입되는 누적 오류를 "배취 효과(batch effects)"로 부른다. 이러한 편향들은 시료 공정/서열화에서 명백한 차이, 가령 시약 공급자의 변화 또는 상기 서열화 운용에서 무작위 순환 오류에 의해 때때로 식별될 수 있다. 그러나, 개체내 GC 편향 제거 후에도, 정규화된 커버리지 변이는 남아있다. 남아있는 변이의 원천은 모르거나, 측정되지 않거나, 또는 너무 복잡하여 단순 모델을 통하여 잡아내기 어렵다. 그럼에도 불구하고, 이질성의 이들 원천을 분석으로의 통합 실패는 아염색체성 테스트의 민감성 및 특이성에 있어서 광범위하고, 유해한 영향을 끼칠 수 있다.

[0495] 정규화된 커버리지 이질성은 예상된 정규화된 커버리지 = 1로부터 공조적 편차(coordinated deviation) ("게놈 웨이브"로 지칭됨)를 보이는 뚜렷한 100kb 커버리지 빈과 연관되는 경향이 있음을 아염색체성 프로파일은 보여준다, 도 12 참고. 이들 게놈 웨이브의 존재는 탐지 알고리즘의 수행에 불리하게 영향을 끼치고, 과장된 가양성/음성 소명을 초래할 수 있다. 이 문제를 해결하기 위하여, 대표적인 영향을 받지 않은 시료들의 정규화된 커버리지를 이용하여 가장 공통적인 게놈 웨이브를 나타내는 직교(orthogonal) 프로파일의 형태로 조직적 잔류 커버리지 변동성을 파악한다.

[0496] *SSS-BER 공정의 개요*

[0497] (I) 각 증후군에 대하여 증후군 콘센수스 영역 (또는 콘센수스 구역으로 지칭됨) 및 증후군 조사 영역 (또는 조사 구역으로 지칭됨) 확인. 상기 증후군 콘센수스 영역은 해당 증후군에 대하여 다양한 존재하는 데이터베이스 및 문헌에서 설명된 서열들의 큰 부분에 걸쳐 공통적인 영역이다. 증후군 조사 영역은 전형적으로 상기 데이터

베이스 및 문헌에서 설명된 모든 또는 많은 서열을 포함하며, 이는 증후군 콘센수스 영역보다 더 넓다.

[0498] (II) 각 증후군에 대한 증후군 근방 빈 (SNB 세트) 확인. 상기 SNB는 증후군 콘센수스 영역에 있는 빈과 밀접한 관계가 있는 강건한 염색체 (가령, chr 13, 18, 그리고 21를 제외한 염색체)에 있는 빈들이다.

[0499] (III) 이들의 SNB 세트 데이터를 이용하여 영향을 받지 않은 시료들의 훈련용 세트 T의 무리를 형성하게 하고, 그렇게 함으로써 훈련용 부분 집단 S1이 획득된다. 상기 훈련용 부분 집단은 SNB 세트에서 유사한 변동성을 갖는 영향을 받지 않은 시료들을 포함한다.

[0500] (IV) 증후군 웨이브 프로파일, w1프로파일은 상기 게놈에서 모든 빈에 있는 중앙 정규화된 커버리지 값으로 정의하고, 상기 중앙값은 S1 시료들에 걸쳐 취해진다.

[0501] (V) 상기 훈련용 세트 T에서 모든 시료들에 대하여 상기 게놈 안의 모든 빈에 대한 정규화된 커버리지 ~ w1프로파일의 강건한 선형 피트의 잔류(residual)를 획득한다.

[0502] (VI) 단계 (III)에서 SNB 세트 데이터와 같이, 단계 (V)로부터 획득된 SNB의 잔류를 이용하여 작동 (III)에서 (V)를 되풀이 반복하고, 그렇게 함으로써 추가적인 증후군 웨이브 프로파일 (w2프로파일, w3프로파일, 등등)를 규정한다.

[0503] (VII) 상기 테스트 시료의 게놈에서 빈의 커버리지를 조절하기 위하여 테스트 시료에 하나 또는 그 이상의 증후군 웨이브 프로파일을 반복적으로 적용한다. 조정의 반복은 정규화된 커버리지 ~ w#프로파일의 강건한 선형 피트의 잔류를 획득하는 것을 포함한다. 상기 잔류는 바람직한 웨이브가 제거될 때까지 그 다음 조정 반복에서 입력량 데이터로 이용된다. 그 다음 상기 잔류는 CNV 분석에 이용된다.

[0504] *SSS-BER 공정의 상세*

[0505] **(I) 증후군 경계 확인**

[0506] 증후군 경계(boundaries)는 증후군 콘센수스 영역과 증후군 조사 영역을 규정한다. 상기 증후군 콘센수스 영역은 해당 증후군에 대하여 다양하게 존재하는 데이터베이스 및 문헌에서 설명된 많은 서열들에 걸쳐 공통적인 영역이다. 본 실시예에서 증후군 콘센수스 영역은 해당 증후군에 대하여 연구된 데이터베이스 및 문헌의 최소한 절반에 공동적인 서열 위치를 포함한다. 증후군 조사 영역은 많은 데이터 베이스 또는 문헌 에서 연구들에 대하여 공통적이지 않는 것들이 포함된, 데이터베이스와 문헌에 설명된 모든 또는 많은 서열들을 전형적으로 포함한다. 상기 증후군 조사 영역은 전형적으로 상기 증후군 콘센수스 영역보다 더 넓다. 본 실시예에서 증후군 조사 영역은 최소한 하나의 데이터베이스 또는 참조 문헌에 공개된 서열들을 포함한다. 당업자는 콘센수스 또는 조사 영역을 선택하는 기준은 조정될 수 있음을 인지할 것이다.

[0507] 게놈 데이터는 몇 가지 공개 데이터베이스로부터 추출되었다, 표 5 참고. 모든 증후군에 있어서, 데이터베이스에서 보고된 변이들은 대표적인 증후군 변곡점 세트를 획득하기 위하여 조사되었다. 추가적으로, 게놈 구조를 설명하는 주요 연구를 확인하고, 증후군 변이들의 유병률을 확인하기 위하여 문헌 검토가 실행되었다. 복합된 이들 원천은 콘센수스 증후군 영역을 확립하고, 뿐만 아니라 주어진 증후군에 대하여 관찰된 변곡점의 대부분이 포괄되는 더 큰 증후군 변곡점 조사 구역을 확립하는데 기여하였다. 텍스트에서 게놈 위치는 참조 인간 게놈 (UCSC Genome Browser GRCh37/hg19)에 따라 명명된다.

표 5

데이터베이스 이름	데이터베이스 링크	목적
DGV: 게놈 변이체들의 데이터베이스(**D**atabase of **G**enomic **V**ariants)	projects|.|tcag|.|ca/variation/	증후군 경계안에 양성 CNV 다형성의 빈도 및 길이 검사
DECIPHER: 앙상블 원천을 이용하여 인간에서 염색체 불균형 및 표현형의 데이터베이스(**D**atabas**E** of **C**hromosomal **I**mbalance and **P**henotype in **H**umans using **E**nsembl **R**esources)	decipher|.|sanger|.|ac|.|uk/	증후군 경계에서 자연적 변이 확립
ClinVar: 서열 변이 및 인간 표현형중에서 상관관계의 공공보존 기록(public archive of relationships among sequence variation and human phenotype)	www|.|ncbi.nlm.nih|.|gov/clinvar/	

표 5: 증후군 경계를 확립하는데 이용된 공공 데이터베이스

[0508]

[0509] **DiGeorge 증후군**

[0510] 세포유전적 위치: 22q11.21

[0511] 발생률: 6,000명에서 1명 (93% 새로이)

[0512] 때때로 두음글자 CATCH22로 불리고, DiGeorge 증후군 (DGS; 188400), 입천장심장얼굴증후군 (VCFS; 192430), 뿔줄기(conotruncal) 기형 안면 증후군 (CTAFS), 그리고 일부 가족성 또는 특발성 뿔줄기 심장 결함 (217095)이 포함된, 발달 장애 집단은 22q11.2의 극소손실과 연관되어 있다. 이 증후군의 임상 이질성과 대조적으로, del22q11 유전적 병소는 단지 몇 안되는 예외적인 것이 있지만, 영향을 받은 개인에서 상당히 균질하다. 대략적으로 환자의 90%는 전형적으로 ~3 Mb의 결손된 영역(TDR)를 갖고, 이는 추정된 30개 유전자들을 포괄하고, 반면 환자의 약 8%는 ~1.5 Mb의 더 작은, 내포된(~1.5 Mb) 결손을 갖는다. Lindsay, E. A. *et al. Velo-cardio-facial syndrome: frequency* and extent of 22q11 deletions. *Am. J. Med. Genet.* ***57,***514-522 (1995). Carlson, C. *et al. Molecular definition of 22q11 deletions in* 151 velo-cardio-facial syndrome patients. *Am. J. Hum. Genet.* ***61, 620-629 (1997).*** Shaikh, T. H. *et al. Chromosome 22-specific low copy* repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum. Mol. Genet.* ***9, 489-501 (2000).***

[0513] 표 5에 열거된 공적 데이터베이스에 추가하여, 또다른 상대적으로 큰 연구에서 환자들의 경계는 전형적으로 결손된 영역 크기를 독립적으로 확인하는데 이용되었다. Adeyinka A1, Stockero KJ, Flynn HC, Lorentz CP, Ketterling RP, Jalal SM. *Familial 22q11.2 deletions in DiGeorge/velocardiofacial syndrome are predominantly smaller than the commonly observed 3Mb.* Genet Med. 2004 Nov-Dec;6(6):517-20.

[0514] 도 13은 DiGeorge 증후군 지형(geography)을 요약하고, 확립된 콘센수스/조사 구역 경계를 그린다.

[0515] **Angelman 증후군/ Prader-Willi 증후군 (15q11.2-q13)**

[0516] 세포유전적 위치: 15q11.2-q13

[0517] 발생률: 12,000명에서 1명

[0518] Angelman 증후군 (AS) 및 Prader-Willi 증후군 (PWS)은 염색체 15q11-q13의 세포유전적 결손을 공유한다. Williams et al. Clinical and genetic aspects of Angelman syndrome. Genet Med. 2010 Jul; 12(7):385-95.

[0519] 도 14는 대중 데이터베이스에서 15q11-q13 증후군의 지형을 보여주고, 확립된 콘센수스/조사 구역 경계를 그린다.

[0520] **Cri-du-Chat (5p- 증후군)**

[0521] 세포유전적 위치: 5p- (5p15.2)

[0522] 발생률: 20,000 - 50,000에서 1명

[0523] Cri-du-chat 증후군은 염색체 5의 짧은 팔의 일부분의 결손과 연관된 유전적 선천적 증후군이다. 상기 결손은 단지 밴드 5p15.2만 관련된 지극히 작은 것에서부터 짧은 팔 전체까지 크기가 다양할 수 있다. 분자 세포유전적 분석에서 62명의 환자 (77.50%)는 p13 (D5S763) 내지 p15.2 (D5S18)의 변곡점 간격으로 특징화된 5p 말단 결손을 가지고 있었다. Mainardi et al. *Clinical and molecular characterisation of 80 patients with 5p deletion: genotype-phenotype correlation.* J Med Genet. 2001 Mar;38(3):151-8.

[0524] 도 15는 관찰된 Cri-du-Chat 증후군 빈도 및 제안된 콘센수스/조사 구역 경계를 열거한다.

[0525] **Williams/Williams-Beuren**

[0526] 세포유전적 위치: 7q11.2 결손

[0527] 발생률: 7,500명에 1명- 20,000명에 1명

[0528] 대부분 환자들은 (95%) ELN 좌의 1.55-Mb 결손을 나타내고, D7S489B, D7S2476, D7S613, D7S2472, 그리고 D7S1870 좌에서 반접합성 (또는 비-정보성)이다. 대중 데이터베이스 마이닝(mining) 결과는 도 5 참고. Bayés, et al., Mutational mechanisms of Williams-Beuren syndrome deletions. Am J Hum Genet. 2003 Jul;73(1):131-51. Epub 2003 Jun 9. Jurado et al., Molecular definition of the chromosome 7 deletion in Williams syndrome and parent-of-origin effects on growth. Am J Hum Genet. Oct 1996; 59(4): 781-792.

[0529] 도 16은 CdC 증후군 커버리지 빈도를 나타낸다. 선은 공적 데이터베이스 또는 문헌 검토에서 증후군 커버리지 빈도를 나타내며, 음영 회색 영역은 8.6Mb 증후군 조사 경계를 나타내며, 1.58Mb 콘센수스 영역들은 녹색 음영으로 나타낸다.

[0530] **Wolf-Hirschhorn 증후군**

[0531] 세포유전적 위치: 4p16.3

[0532] 발생률: 50,000명에 1명

[0533] Wolf-Hirschhorn 증후군은 부분적 4p 결함이 원인이 되는 세그먼트성 2상염색체(aneusomy)이다. 상기 결손은 전형적으로 1.9-3.5 Mb이며, 그리고 주로 말단이다. 최근, 상기 증후군의 기전이 검토되었고, 새로운 임계적 영역이 확립되었다. Zollino, et al. Mapping the Wolf-Hirschhorn Syndrome Phenotype Outside the Currently *Accepted WHS Critical Region and Defining a New Critical Region, WHSCR-2* Am J Hum Genet. 2003 March; 72(3): 590-597.

[0534] 도 17은 관찰된 WH 증후군 커버리지 빈도 및 제안된 콘센수스/조사 구역 경계를 열거한다.

[0535] **1p36 결손**

[0536] 세포유전적 위치: 1p36

[0537] 발생률: 5,000명에 1명

[0538] 염색체 1p36 결손 증후군은 가장 흔한 인간 말단 결손 증후군으로, 5,000명의 출생중 1명꼴로 발생된다. 60개의 가족으로부터 일염색체성 1p36을 가진 61명의 개체에서 결손 크기를 평가한 연구는 1p36에 대한 콘센수스 영역을 선택하는데 기여하고 있었다. Heilstedt et al. Physical map of 1p36, placement of breakpoints in *monosomy 1p36, and clinical characterization of the syndrome.* Am J Hum Genet. 2003 May;72(5):1200-12.

[0539] 도 18은 1p36 증후군 지형(geography)을 요약하고, 확립된 콘센수스/조사 구역 경계를 그린다.

[0540] 하기 표 6은 모든 제안된 콘센수스/조사 구역 경계를 요약한다.

표 6

표 6: 증후군 경계 요약

[0541]

			콘센수스 영역			조사 영역		
증후군	발생률	Chr	시작	중단	Size (MB)	시작	중단	Size (MB)
1p36	5,000명에 1명	1	0	5,000,000	5	0	13,488,000	13.5
WolfHirschhorn	50,000명에 1명	4	0	3,573,000	3.6	0	14,501,000	14.5
CriduChat	1 in 20,000	5	0	9,800,000	9.8	0	26,541,000	26.5
Williams	1 in 7,500	7	72,701,000	74,285,000	1.6	68,848,000	77,453,000	8.6
Angelman	12,000명에 1명	15	22,699,000	28,520,000	5.8	20,576,000	30,641,000	10.1
DiGeorge	6,000명에 1명	22	18,891,000	21,561,000	2.7	17,398,000	22,916,000	5.5

[0542] **(II) 증후군 근방 빈 식별**

[0543] 증후군 근방 빈 (SNB)은 증후군 콘센수스 영역에 있는 빈과 밀접한 관계가 있는 강건한 염색체 (가령, chr 13, 18, 그리고 21를 제외한 염색체)에 있는 빈들이다. 증후군 콘센수스 영역들에서 관찰된 조직적 변동성을 공유하는 100kb 빈을 식별해내기 위하여, 훈련용 데이터를 이용하여 콘센수스 영역에서 각 빈과 강건한 염색체 (가령 염색체 13, 18, 그리고 21 제외)에 속하는 모든 100kb 상염색체 빈 사이의 쌍별(pairwise) 거리를 나타내는 거리 매트릭스를 구축한다. 상기 거리는, 대상간 변이가 콘센수스 영역에서 빈과 강건한 염색체에서 빈과의 사이에서, 얼마나 유사한 지를 나타낸다.

[0544] 상기 거리는 Euclidian 거리, 상관, 절대적 상관, 코사인 각, 절대적 코사인 각, 또는 임의의 다른 적절한 계측

으로 산출될 수 있다. 상관관계는 벡터들이 유사한 모양을 갖도록 요구한다. 코사인-각 거리는 상기 상관관계 산출에서 평균 제로를 가지도록 재-설계되기 보다는 이들 고유 평균 가까이 집중된다는 점을 제외하고, 상관관계에 유사하다. 이 실시예에서, 상기 거리는 100kb 커버리지 계수에서 코사인-각 (aka 상관 거리)으로 산출된다.

[0545] 그 다음, 증후군 콘센수스 영역에서 각 100kb 빈의 경우, 가장 근접한 (그리고 가장 관련이 높은) 빈이 선택된다. 그 다음 증후군 근방 빈 (SNB) 세트를 창출하기 위하여 증후군 콘센수스 영역에서 모든 100kb 빈에 대하여 가장 가까운 빈들을 모았다. 일부 구체예들에 있어서, 다중 증후군에 대한 가장 가까운 빈들을 복합시켜 다중 증후군에 대한 마스터 SNB 세트를 창출시킨다. 다른 구체예들에서, 도 19에서 보여준 실시예에서와 같이 상이한 증후군에 대한 별도 SNB 세트들이 창출된다.

[0546] SNB의 인구학 검정에서 복합 증후군 - 염색체 관계가 드러났다. 도 19는 증후군에 의한 SNB 염색체 자격을 요약한 것을 보여준다. 도 19의 수평 축은 염색체 수이며, 그리고 수직 축은 SNB 세트에서 빈의 백분율이다. 일부 증후군은 특정 염색체에서 특징적으로 높은 빈의 비율을 갖는 것으로 보인다. 예로써, Cri-du-Chat 증후군은 염색체 19 상에 25% 이상의 SNBs를 갖고, 염색체 16 상에 15% 이상의 빈을 갖는다. 염색체 1p36 결손 증후군은 염색체 22 상에 SNBs의 30% 이상의 놀라운 높은 비율을 갖는다.

[0547] 추가적으로, 증후군간의 SNB의 관계가 연구되었고, 증후군들중에서 SNB 영역들에서 사소하지 않은 연계(non-trivial ties) 및 상관관계가 다시 드러난다. SNB의 크기는 영향을 받지 않은 훈련용 코호트에서 증후군 수행 관점으로부터 또한 연구되었고, 본 내용의 후반에 논의된다.

[0548] **(III) 발병안된 시료들의 훈련용 세트로부터 훈련용 부분 집단 획득**

[0549] 이들의 SNB 세트 데이터를 이용하여 영향을 받지 않은 시료들의 훈련용 세트 T의 무리를 형성하게 하고, 그렇게 함으로써 훈련용 부분집단 S1이 획득된다. 상기 훈련용 부분집단은 SNB 세트에서 유사한 변동성을 갖는 영향을 받지 않은 시료들을 포함한다.

[0550] 이 실시예에서와 같이 일부 구체예들에 있어서, Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH) 알고리즘을 이용하여 부분집단 S1가 획득된다. 기타 클러스터링 기술, 이를 테면 K-평균s 클러스터링은 훈련용 부분집단을 획득하기 위하여 대체 구체예들에 또한 적용될 수 있다. HOPACH는 입력량 매트릭스를 취하고, 2개의 시료들 간에 쌍별 거리에 근거하여 분할(partitioning)을 실행한다. 상기에서 열거된 바와 같이 2개의 시료 사이의 거리를 산출하는 상이한 방식들이 존재한다.

[0551] 여기 실시예에서와 같이, 한 가지 실행에 있어서, 분할 기술은 메도이드 주변 분할 (PAM)을 통한 기술이며, 그리고 클러스터링 알고리즘은 HOPACH-PAM으로 지칭되며, 이는 클러스터의 계층별 트리이다. M. van der Laan and K. Pollard, A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. M. van der Laan and K. Pollard, A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. HOPACH 방법은 분할 및 응집성(agglomerative) 클러스터링 방법들의 강도를 복합시키고, 연구원이 증가된 상세 수준에서 클러스터를 검토하도록 허용한다.

[0552] HOPACH-PAM는 트리의 각 수준에서 다음 단계의 반복이 연관된 계층별 클러스터링 분석이다:

[0553] 1. 분할(Partition): 각 클러스터에서 요소들에 대하여 PAM 적용;

[0554] 2. 정리(Order): PAM으로부터 획득된 새로운 클러스터들을 정리; 그리고

[0555] 3. 붕괴(Collapse): 일부 클러스터들을 가능한 합병.

[0556] 각 클러스터가 3 요소들을 포함할 때, 또는 평균 실루엣(silhouette)이 최대화될 때, 이 공정은 중단된다. 최종 수준은 상기 요소들의 정돈된 목록이다.

[0557] **1. 분할(Partition)**: A HOPACH-PAM 공정은 상기 훈련용 세트를 2개의(또는 그 이상의) 클러스터로 분할시킴으로써 시작될 수 있다. 자동화된 클러스터링 공정에서, 2개의 시작 클러스터는 임의로 선택될 수 있다. 그 다음 각 클러스터에 대한 메도이드(medoid)가 획득된다. 메도이드는 클러스터에서 모든 대상에 대한 평균 부동성 (또는 거리)가 최소인 클러스터 구성원이다. 메도이드는 평균 (또는 중심(centroids))과 개념적으로 유사하지만, 메도이드는 항상 데이터 세트의 구성원들이다. 그 다음, 이 공정은 평균 실루엣이 최대화될 때까지 구성원들과 메도이드들을 재할당함으로써 클러스터를 업데이트한다. 클러스터 k의 수가 명시된 범위에서서 명시되거나 또는 자동적으로 조사될 수 있다. 후자의 경우에 있어서, 최대 평균 실루엣을 갖는 클러스터의 수에 의해 k가 결정될

수 있다.

[0558] 개념적으로 상기 실루엣은 한 대상이 이의 고유 클러스터에서 다른 대상에 얼마나 잘 정합되는지 그리고 이 대상이 또다른 클러스터로 이동된다면 잘 정합되는지에 대하여 측정한다. 요소 j의 실루엣은 다음과 같이 정의된다:

[0559] $$S_j = \frac{b_j - a_j}{\text{최대}(a_j, b_j)}$$

[0560] 여기에서 a_j는 이 클러스터의 다른 요소들과 요소 j의 평균 부동성이며, 그리고 b_j는 요소 j가 속하지 않는 임의의 하나의 클러스터의 요소들과 요소 j의 최소 평균 부동성이다.

[0561] **2. 정리(Order):** HOPACH-PAM는 상기 분할 단계에서와 같이 동일한 거리 계측을 이용하여 약간의 감각적 방식으로 클러스터들을 각 수준에서 정리한다. 정리하는 한 가지 방식은 이들의 메도이드의 거리에 근거하여 부모들을 트리 수준으로 정리하는 것이다. 차일드 클러스터에 대하여 이웃 클러스터는 상기 차일드 클러스터의 부모 클러스터의 우측에 있는 클러스터로 정의한다. 그 다음 차일드 클러스터들과 이들의 이웃 클러스터 간의 거리에 근거하여 차일드 클러스터를 정리한다. 차일드 클러스터의 최우측 세트의 경우, 이웃 클러스는 이들 부모의 좌측에 있고, 최소 거리에서부터 최대 거리로 정리된다.

[0562] **3. 붕괴(Collapse):** 클러스터 트리 수준에서 2개 또는 그 이상의 클러스터는 유사할 수 있으며, 그리고 이들은 부모를 공유하거나 또는 공유하지 않을 수 있다. 본 실시예에서와 같이, 붕괴시킬 것인가 판단은 붕괴 단계 전후 비교되는 기준에 근거한다. 일부 구체예들에 있어서, 붕괴로 트리 수준의 평균 실루엣이 개선되는 클러스터쌍이 없어질 때까지 붕괴는 지속된다. 대안으로, 붕괴되는 클러스터는 정리된 거리 매트릭스의 시각적 관찰에 의해 식별될 수 있다. 하나의 클러스터의 라벨을 다른 클러스터에 제공하여, 트리 구조가 보존되도록 함으로써, 붕괴가 실행된다. 라벨의 선택은 임의적일 수 있거나 (가령: 가장 우측 클러스터) 또는 예전 메도이드가 새로운 인접 클러스터에 대한 유사성에 근거하여 또는 일부 다른 기준에 근거하여 이루어질 수 있다. 다양한 옵션에 의해 병합된 클러스트에 대하여 새로운 메도이드가 선택되며, 이를 테면 2개의 예전 메도이드의 가장 인접한 이웃의 평균(아마도 중량) 및 클러스터의 평균 실루엣의 최대인 것이 선택된다.

[0563] 본 실시예에서 HOPACH-PAM 적용에 있어서, the HOPACH-PAM 알고리즘은 매트릭스 SNBNormCov = SNB 빈의 # x 훈련용 시료들의 #에서 훈련용 시료들의 SNB 커버리지을 입력량으로 취하고, SNBNormCovij = 훈련용 시료 j에서 증후군 근방 빈 i의 정규화된 커버리지. 상기 알고리즘은 코사인-각 거리 (상관관계 산출에서 평균 제로를 가지도록 재-설계되기 보다는 이들 고유 평균 가까이 집중된다는 점을 제외하고, 상관관계에 유사함)에 근거한 입력량 매트릭스로부터 부동성 매트릭스 D를 계산한다. 상기에서 설명된 바와 같이, 코사인 각은 2개의 시료 사이의 거리를 계산하기 위한 가능한 방식중 하나이다.

[0564] 영향을 받지 않은 시료들은 클러스터의 계층별 트리로 분할되며, 그리고 전체 훈련용 세트 T로부터 최대 클러스터, S1이 선택된다. 클러스터링 분석에 있어서, 정규화된 커버리지에서 공통의 조직적 변동성을 공유하는 훈련용 영향을 받지 않은 시료들의 집단을 식별해내는 것이 흥미롭다. 계층별 트리 내용에서, 이것은 클러스터가 여전히 중요한 트리 수준을 선택하는 것에 상응한다. 이것은 트리를 따라 하향 이동하면서, 평균 실루엣이 트리의 기존 수준의 평균 실루엣을 개선시키는 경우에만 이 트리의 붕괴된 다음 수준을 수용함으로써 실행된다.

[0565] **(IV) 증후군 웨이브 프로파일의 규정**

[0566] 영향을 받지 않은 시료들 T의 전체 훈련용 세트중에서 HOPACH-PAM로부터 최대 클러스터를 훈련용 부분집단 S1로 선택한 후, S1의 시료들은 데이터를 제공하고, 증후군 웨이브 프로파일, **w1프로파일**은 상기 게놈에서 모든 빈의 중앙 정규화된 커버리지 중앙값으로 정의하며, S1 시료들중에서 중앙값이 취해진다. 이 실시예에서와 같이 일부 구체예들에 있어서, 참조 서열, 이를 테면 전체 게놈 또는 정규화 상염색체의 세트의 커버리지에 대하여 상기 정규화된 커버리지 값이 정규화된다. 더욱이, 일부 구체예들에서, 상기 커버리지는 본 명세서의 도처에서 설명된 포괄적 웨이브 프로파일로 지칭되는 포괄적 변동에 대하여 조정되며, 포괄적 웨이브 프로파일은 전체 훈련용 세트 T로부터 획득되었다.

[0567] **(V) 강건한 선형 피트의 잔류 획득**

[0568] 상기 훈련용 세트 T에서 모든 시료에 대하여 커버리지 ~ **w1프로파일**의 강건한 선형 피트를 획득하기 위하여 Huber M-추정량(estimator)을 이용한다. 그 다음 모든 시료에 대한 잔류 정규화된 커버리지는 모든 빈의 커버리

지로부터 선형 피트 값을 차감함으로써 획득되며, 그렇게 함으로써 조정된 커버리지가제공되고, 증후군 웨이브 프로파일은 제거된다.

[0569] **(VI) 추가적인 증후군 웨이브 프로파일 규정**

[0570] 하나의 증후군 웨이브 프로파일 w1프로파일이 웨이브 #1에 대하여 제거된 후, 유사한 방식으로 추가 증후군 웨이브 프로파일이 제거될 수 있다. 이는 작동 V에서 획득된 SNB의 잔류를 작동 III을 위한 SNB 세트로 이용하여 작동 III-V를 반복하고, 그렇게 함으로써 추가적인 증후군 웨이브 프로파일 (**w2프로파일**, **w3프로파일**, 등등)이 규정됨으로써 실행될 수 있다. 웨이브 #1가 제거된 후 잔류 정규화된 커버리지 매트릭스는 제2 라운드의 계층별 클러스터링를 거치고, 최대 클러스터의 중앙 프로파일이 다시 추출되고, 웨이브 #1에 직교(orthogonal) 성분이 웨이브 #2로 저장된다. 이와 같은 반복은 증후군 웨이브 프로파일의 부동(stationary) 세트의 일부분으로 축적될 때까지 지속되며, 이는 상이한 시료들을 테스트하는데 적용될 수 있다.

[0571] **(VII) 증후군 웨이브 프로파일을 적용**

[0572] 웨이브 프로파일이 획득된 후, 이들은 상기 게놈에서 빈의 커버리지를 조절하기 위하여 테스트 시료에 반복적으로 적용될 수 있다. 상기 테스트 시료의 조정은 상기 훈련용 시료들로부터 웨이브를 획득할 때, 훈련용 시료에 대한 조정과 동일한 방식으로 실행될 수 있다. 조정의 반복은 커버리지 ~ **w#프로파일**의 강건한 선형 피트의 잔류를 획득하는 것을 포함한다. 그 다음 잔류 정규화된 커버리지는 그 다음 조정 반복에서 입력량 데이터로 이용된다.

[0573] 일단 SNB 웨이브가 확립되면, 이들의 적용은 한번에 하나의 웨이브를 이용하여 반복적 강건한 선형 피트 공정이 되고, SSS-BER의 다음 반복에 이용되는 잔류 정규화된 커버리지 매트릭스가 형성된다.

[0574] *SSS-BER 공정의 효과*

[0575] 모든 반복 단계에서 SSS-BER의 효과 연구에서 웨이브 수가 N= 4를 초과할때 수행 개선에서 신속한 감소가 드러났다, Cri-du-Chat에 대한 CV의 경우 도 21 참고. 여기에서, 콘센수스 증후군 영역에서 중앙값 커버리지은 증후군 염색체의 중앙값 정규화된 커버리지에 의해 정규화되어, 수행 계측으로써 콘센수스 증후군 비율을 확립한다.

[0576] 추가적으로, SNB의 크기를 연구하여 증후군 근방 빈의 가장 대표적인 수집을 결정한다. 구체적으로, 증후군 콘센수스 영역 에서 모든 100kb 빈에 대하여 상위 가장 근접한 빈의 변수가 선택되었고 (상위 2% 내지 10%) 그리고 SS-BER-이후 중앙 정규화된 커버리지의 CV는 모든 SNB 형태에 대하여 산출되었다, Cri-du-Chat에서 보여진 효과는 도 22 참고.

[0577] *Pre-SSS-BER 단계*

[0578] *플로우 쎌-특이적 포괄적 웨이브 프로파일 제거*

[0579] 일부 구체예들에 있어서, SSS-BER 공정에 앞서, 빈의 커버리지는 플로우-세포 포괄적 웨이브 프로파일과 GC 프로파일에 대하여 조정된다.

[0580] 조직적 변동성을 더 감소시키기 위하여, 일부 구체예들은 전체 훈련용 세트의 시료에 근거하여 포괄적 웨이브 교정을 대체하는 동일한 플로우 쎌의 시료에 근거한 플로우 쎌-특이적 배취 교정의 장점을 취한다. 이러한 교정은 플로우 쎌-특이적 포괄적 웨이브 프로파일 (FC-GWP) 제거로 지칭되며, 이때 플로우 쎌-기반의 포괄적 중앙 프로파일은 증후군 웨이브 프로파일 제거 전, 미가공 커버리지 교정에 이용된다. xjk는 염색체 j = 1,...,h, 그리고 빈 k=1,...,nj인 임상 시료로부터 계수된 관찰된 NES로 추정하고, 이때 nj는 염색체 j의 빈의 수이다.

[0581] 단계 1: 정규화 커버리지를 획득하기 위하여, 모든 빈 k와 염색체 j에 대하여 다음을 산출한다

[0582] $$nx_{jk} = \frac{x_{jk}}{\sum_{\{염색체의\ 강건한\ 세트\}에서\ j} x_{jk}}$$

[0583] 단계 2: FC-GWP 산출: FC-GWP는 계수 품질 관리 임계치를 통과하는 FC 에서 모든 시료에 걸친 nx_{jk}의 중앙값으로 규정.

[0584] 단계 3: nx_{jk} 를 $FC\text{-}GWP_{jk}$ 로 나눔으로써, FC-GWP에 의한 정규화된 커버리지를 조절. 등급(scale) 매개

변수는 $FC - GWP_{jk}$ into nx_{jk} Huber M − 추정량이용의 강건한 선형 피트를 통하여 $\{\alpha,\beta\}=nx_{jk}=rlm(FC - GWP_{jk})$: 의 회귀 계수를 이용하여 결정된다.

$$r_{jk} = \frac{nx_{jk}}{\alpha \cdot GWP_{jk} + \beta}$$

[0585]

[0586] 도 23은 모든 훈련용 시료들에 기초한 포괄적 프로파일을 제거하는 공정 파이프라인에서 중앙 증후군 CVs을 FC-GWP를 이용하는 공정, 그리고 FC-GWP 및 SSS-BER을 포함하는 공정에 비교한다. 표 7은 다양한 깊이의 커버리지에서 FC-GWP/SSS-BER 파이프라인 수행을 요약한다.

표 7

표 7: 1 및 12plex 테스트 세트에서 증후군 비율 CV.

증후군	Chr	시작	중단	길이 (Mb)	12plex	1plex
1p36	1	0	5,000,000	5	0.88%	0.43%
WolfHirschhorn	4	0	3,573,000	3.57	1.08%	0.38%
CriduChat	5	0	9,800,000	9.8	0.58%	0.36%
Williams	7	72,701,000	74,285,000	1.58	1.90%	0.75%
Angelman	15	22,699,000	28,520,000	5.82	0.88%	0.58%
DiGeorge	22	18,891,000	21,561,000	2.67	1.23%	0.49%

[0587]

[0588] *GC 교정*

[0589] 다양한 구체예들에 있어서, 커버리지는 상기에서 설명된 바와 같이, 증후군 특이적 변동에 대하여 조절되기에 앞서, GC 함량 편향에 대하여 교정될 수 있다. 간략하게 설명하자면, 이는 상기 테스트 시료의 빈에서 GC 함량 수준과 커버리지 사이의 관계에 근거하여 커버리지을 조절하여 실행될 수 있다. 일부 구체예들에 있어서, GC 함량 수준과 커버리지 간의 상관관계는 GC 함량과 커버리지 사이의 비-선형 모델에 의해 표시될 수 있다. 일부 구체예들에 있어서, 상기 조정은 비-선형 피트 값을 커버리지로부터 차감함으로써 이루어진다. 다른 구체예들에서, 상기 조정은 커버리지를 비-선형 피트 값으로 나눔으로써 이루어질 수 있다. 일부 더 단순한 시행에서, 조정 값을 획득하기 위하여 비-선형 피트를 이용하는 대신, 커버리지에 의해 정리된 빈의 평균 또는 중앙 값을 조정 값으로 이용할 수 있다.

[0590] *SSS-BER 단계 이후*

[0591] *정규화된 증후군 값 획득*

[0592] 증후군 영역에 대한 커버리지가 획득된 후, 이 영역에 대한 커버리지가 복합될 수 있다. 일부 구체예들에 있어서, 상기 영역은 상이한 테스트 시료들에 대하여 동일하며, 이 경우 복합된 커버리지 값은 증후군 영역을 주최하는 염색체의 커버리지로 나눠질 수 있다. 그 다음 상기 커버리지 값을 이용하여 영향을 받지 않은 시료들의 커버리지 값에 근거한 표전화된 스코어를 제공할 수 있다. 표준화된 값은 정규화된 증후군 값으로 지칭될 수 있다:

[0593] $$NSV_j = \frac{x_j - \hat{\mu}_j}{\hat{\sigma}_j}$$

[0594] 이때 $\hat{\mu}_j$ 및 $\hat{\sigma}_j$ 는 영향을 받지 않은 시료들 세트에서 j-번째 증후군 분량에 대한 차례로 추정된 평균 및 표준 편차이다. 이들 구체예들에 있어서, NSV는 특정 신뢰 구간을 갖는 판단 기준에 대응하는 z-점수에 개념적으로 유사한 판단 임계치와 비교된다.

[0595] 다른 구체예들에서, 각 개별 테스트 시료는 문제의 증후군과 관련된 가장 강력한 신호를 포함하는 시료별(sample-wise) 증후군 세그먼트를 갖는다. 상기 증후군 세그먼트는 하기에서 더 설명되는 바와 같이 획득될 수 있다. 이러한 경우에서, 증후군 소명을 위한 계측이 상기 증후군 세그먼트에 대하여 획득되고, 판단 기준에 비교된다.

[0596] *분절화*

[0597] 일부 구체예들에 있어서, 상기에서 설명된 증후군 영역은 하기에서 더 설명되는 분절화 일정에 의해 획득된 증후군 조사 영역 에서 있는 증후군 세그먼트일 수 있다. 일부 관련된 구체예들에서, 상기 조정-이후 단계는 하기에서 설명되는 대체 방법으로 치환될 수 있다. 일부 구체예들에 있어서, 상기 분석은 증후군 세그먼트에 대한 NSV 대신, 태아 분획(FF) 또는 이로부터 유도된 값을 산출하고, FF 또는 유도된 값은 판단 기준과 비교하여 CNV가 증후군 세그먼트에서 존재하는 지를 결정한다. 판단 공정 및 상세한 실시예의 고차원적 설명이 이후에 제공된다.

[0598] **실시예 4: 민감성 및 선택성을 최대화시키기 위한 2-단계 서열화**

[0599] 주어진 데이터 세트의 경우, 영향을 받은 시료와 영향을 받지 않은 시료를 구별하는 주어진 판단 기준은 항상 가양성 소명과 가음성 소명 사이에 균형을 유지한다. 영향을 받은 시료들이 영향을 받지 않은 시료보다 더 높은 평균을 가진 정상적 분포를 가진다고 가정하면, 기준이 더 낮게 설정되어 있다면, 민감성은 증가되고, 선택성은 감소된다. 역으로, 기준이 더 높게 설명된다면, 민감성은 감소되고, 선택성은 증가된다. 반복하기 위하여, 민감성은 실질적으로 모든 양성 (가령, 영향을 받은 또는 환자) 시료들에 걸쳐 참양성 소명의 비율이며, 한편 선택성은 실질적으로 모든 음성 (가령, 영향을 받지 않은 또는 대조) 시료들에 걸쳐 참음성 소명의 비율이다. 본 내용은 서열들의 CNV의 소명의 민감성 및 선택성 모두를 증가시키는 방법을 제시한다. 이는 측정된 값이 상대적으로 낮은 영향을 받은 시료들, 그리고 측정된 값이 상대적으로 높은 영향을 받지 않은 시료들에서 특히 유용하다.

[0600] 일부 구체예들에 있어서, 복합 서열화에 의해 서열화된 시료들의 경우 제1 양성 임계치와 제1 음성 임계치가 선택된다. 제1 양성 임계치보다 더 높은 시료들은 영향을 받은 것으로 판단되며, 그리고 제1 음성 임계치보다 더 낮은 시료들은 영향을 받지 않은 것으로 판단된다. 제1 양성 임계치와 제1 음성 임계치 사이에 있는 시료들은 무-소명 구역으로 간주된다. 가양성을 감소시키기 위하여, 초기 양성 임계치는 상대적으로 높게, 가령, z=4로 설정된다. 한편, 가음성을 감소시키기 위하여 초기 음성 임계치는 상대적으로 낮게, 가령, z=1로 설정된다. 초기 양성 임계치와 초기 음성 임계치 사이의 범위는 "무소명(no call)" 구역으로 간주된다. 그 다음 무-소명 구역에 있는 시료들은 제2 라운드의 더 높은 서열화 심도의 서열화(가령, 단일 서열화)를 거쳐 제2 NCV 또는 NSV, 또는 판단을 위한 다른 측정된 값 (가령, z 값)을 얻는다.

[0601] 제1 양성 임계치와 제1 음성 임계치 사이의 시료들이 추가적인 심화 서열화를 거치는 것에 추가적으로 또는 이를 대신하여, 이를 테면 하기에서 설명되는 실시예에서와 같이, 재서열화를 위하여 제1 임계치 (더 낮은, 가령, NCV=1.3)를 설정하고, 최저 등급 시료의 일부를 선택할 수 있다. 예로써, 제1 임계치 이상의 시료의 10%는 재서열화를 거친다. 그 다음, 제2 임계치 (더 높은, 가령, NCV=4)를 이용하여 재서열화된 시료가 영향을 받았는지를 결정할 수 있다.

[0602] 심화 서열화는 더 많은 리드를 만들고, 상기 환자 분포의 폭을 줄이며, 이는 상기 환자 분포와 대조 분포 사이의 중첩을 감소시킨다. 따라서, 공개된 방법은 제2 측정된 값이 비교되는 제 2 임계치에 적용될 수 있다. 가음성의 과다한 팽창없이 가양성을 상대적으로 낮게 유지시키기 위하여 제 2 임계치가 선택된다. 심화된 서열화 깊이로 인하여, 참양성 시료들의 폭을 효과적으로 조일 수 있다는 것도 가능하다. 일부 구체예에서, z 스코어 또

는 NCV는 예로써 3 또는 4로 설정될 수 있다. 이 2-단계 서열화 방법은 선택성 및 선택성을 모두 증가시킨다.

[0603] 본 명세서는 NIPT 또는 암에서 염색체 홀배수체 분류 및 혼합물에서 복제 수 변이에 있어서 높은 민감성 및 높은 특이성을 얻을 수 있는 2단계-서열화 및 서열 태그 계수를 위한 것이다. 상기 2-단계 방법은 서열 태그의 계수에 의존하는 임의의 진단학적 방법에 적용가능하다.

[0604] NIPT에서 진단학적 결과의 정확성을 최대화시키는 것에 추가하여, 2-단계 방법은 제1 저가 비용 서열화 단계에서 시료의 90% 또는 그 이상이 잠재적 양성으로 제거되기 때문에 합당한 비용으로 평균 위험 시장에 진입을 허용한다. 그 다음 양성 예측 값 (PPV)은 가양성 비율을 낮춤으로써 단일 단계 서열화 방법보다 개선될 수 있으며, 여기에서 PPV = 참양성/ (참양성 + 가양성)이다.

[0605] 상기 방법은 제1 저가 비용 서열화 단계에서 시료의 80-90%가 잠재적 양성으로 제거되고, 한편 제2 서열화 단계로부터 높은 특이성이 유지되면서 합당한 비용에서 암을 스크리닝 및/또는 식별하기 위한 시작으로의 진입을 잠재적으로 허용한다.

[0606] 임의의 좌 또는 대립유전자-특이적 계수 적용은 최대 특이성 및 민감성을 위하여 최적화될 수 있다.

[0607] 서열 태그의 계수에 의한 염색체 홀배수체가 탐지될 때:

[0608] (1) 가양성 비율은 이용되는 태그 수에 의존하지 않으며-시료를 "탐지된 홀배수체"로 분류하기 위하여 선택된 오직 컷-오프에만 의존한다

[0609] 예를 들면, 만일 "홀배수체 탐지됨"은 다음의 z-점수 컷-오프에서 분류된다면:

[0610] - 3 - FP 비율은 0.13%이고

[0611] - 2 - FP 비율은 2.30%이고

[0612] - 1.65 - 5% 등등이며; 그리고

[0613] (2) 가음성 비율은 3개 인자에 의존된다:

[0614] - 이용되는 태그 수

[0615] - 시료를 "탐지된 배수체"로 분류하기 위하여 선택된 컷-오프(z-점수)

[0616] - 분류되는 특정 염색체에 대한 측정의 상대적 변동 (CV - 변이계수)

[0617] ■ 예를 들면, 염색체 21의 CV는 ~0.4%이며, 염색체 18의 경우는 ~0.2%이다.

[0618] - 집단에서 cfDNA 태아 분획 분포

[0619] ■ 낮은 태아 분획을 가진 시료를 제거하기 위하여 태아 분획 측정은 없는 것으로 추정한다.

[0620] 따라서, 예상된 태아 분획 분포 및 특이적 변이계수 (CV)를 위하여, 도 24에 나타낸 민감성/특이성 곡선으로써 계수가능한 서열 태그 수의 범위에 대하여 분류된 가양성 (FP)에 예상된 가음성 (FN)을 결부시킬 수 있다. 도 24는 염색체 21에 있어서 상이한 서열화 심도에 있어서 가양성 비율에 대한 가음성 비율을 나타낸다. 도 25는 염색체 18에 있어서 상이한 서열화 심도에 있어서 가양성 비율에 대한 가음성 비율을 나타낸다. 이들 곡선은 z(가양성)에 대하여 z(참양성)을 플롯시킨 수용자 작동 특징(ROC)에 상보적이며, 여기에서 ROC 곡선 아래 면적은 특정 d' 민감성을 나타낸다 (여기에서 보여지지는 않음).

[0621] **실시예 2-단계 서열화 공정**

[0622] **단계 1**

[0623] 많은 시료들/라인을 서열, 그리고 NCV 컷-오프를 낮게 설정하여 민감성은 >99.5%이다.

[0624] ■ 예를 들면, NCV = 1.3, FP 비율은 ~10%이다.

[0625] ■ T21의 경우 5M 태그 탐지율은 >99.3%이며, T18의 경우 > 99.8%이다.

[0626] 단계 1은 분석의 민감성을 최대화시키는데, 가령 참양성을 탐지하는 능력을 최대화시킨다.

[0627] 예를 들면, 제1 단계에서 제1 (더 큰) 수의 시료들 가령 48개로부터 DNA는 한개의 플로우 쎌 라인에서 서열화되어, 각 시료에 대한 서열 태그의 첫 수 가령 5M을 생성시키고, 그리고 이 시료에 대한 NCV가 산출된다. 상기

NCV는 참양성의 최대 비율이 탐지되는 것 이상에서 제1 (더 낮은) 사전설정된 임계치 값과 비교되는데, 가령 제1 임계치는 이 분석의 민감성을 최대로 하기 위하여 가령 T21의 경우 >99.5% 그리고 T18의 경우 99.8%가 되도록 사전설정된다. 이 경우에서 NCV =1.3에서 집단의 약 90%는 참음성으로 분류될 것이고, 이렇게 보고된다.

[0628] **단계 2**

[0629] 몇 개의 시료들/라인 (~8)으로 제1 서열 운용으로부터 10% 추정 양성을 서열화하고, 특이성은 >99.9%이 되도록 NCV 컷-오프를 설정한다.

[0630] ■ 예를 들면, NCV = 4, FP 비율은 <<0.1%이다.

[0631] ■ 삼염색체성의 경우 25 M 태그에 대한 탐지 비율은 >99.9%이다.

[0632] 단계 2는 이 분석의 특이성을 최대화시키는데, 가령 참음성을 탐지하는 능력을 최대화시킨다.

[0633] 제2 단계에서, 99.5% 민감성으로 양성으로 결정되었던 시료들만, 가령 제1의 사전설정된 더 낮은 임계치 (1 SD)보다 더 큰 NCV를 보유한 시료들은 시료당 더 큰 서열 태그 수, 가령 24M 태그/시료를 획득하도록 재서열화된다.

[0634] 이 시료의 태아 분획이 상기 예상된 분포 에서 속한다는 가정하에, 시료당 서열 태그의 수를 증가시키고, 그리고 탐지되는 가양성 시료들의 수를 최소화시키는 이상으로 임계치를 사전설정함으로써, 이 분석의 특이성이 최대화될 수 있다. 시료당 태그 수가 클수록 분류되는 특정 염색체의 측정에 대한 상대적 변동(CV-변이계수)는 더 낮아지고, 테스트의 정확성은 더 커진다.

[0635] 플로우 쎌 라인당 더 적은 시료를 서열화함으로써, 시료당 더 큰 서열화 심도를 이룰 수 있다. 예를 들면, 단계 2에서 시료당 24M 태그를 획득하기 위하여 플로우 쎌 라인당 오직 8개 시료만 서열화된다. 각 시료에 대한 NCV가 산출되고, 제2의 (더 높은) 사전설정된 임계치와 비교된다. 이 경우에 있어서 제2 임계치는 NCV=4에서 사전설정되고, 그리고 오직 NCV>4를 보유한 시료들만 양성으로 분류된다. NCV>4를 갖는 시료들에 대하여 탐지된 가양성 비율은 0.1% 미만임을 보여주었다.

[0636] 상기에서 설명된 포괄적 프로파일과 증후군 프로파일 제거 공정이 복합된, 유사한 2-단계 서열화 과정을 이용하여, 아염색체성 영역들의 CNV가 탐지될 수 있다. 표 8에서는 1p36 결손 증후군, Wolf-Hirschhorn 증후군, Cri-du-Chat 증후군, Angelman 증후군, 그리고 DiGeorge 증후군의 경우 상기 기술을 이용한 NCV 탐지의 민감성은 높은 민감성과 양성 예측 값과 함께 획득될 수 있음을 보여준다.

표 8

표 8. 5 개 증후군에 대하여 상기 기술들을 이용한 NCV 탐지의 민감성.

		Estimated performance at NCV = 4 cutoff			
Syndrome	**Length, M**	**1X (N=160), sensitivity**	**12X (N=384), sensitivity**	**12X + 1X re-seq, sensitivity**	**12X + 1X re-seq, PPV**
1p36	5	88.41%	78.21%	**86.18%**	14.70%
WolfHirschhorn	3.57	90.65%	67.35%	**86.96%**	4.17%
CriduChat	9.8	91.36%	89.26%	**90.34%**	5.68%
Angelman/Prader-Willi	5.82	77.96%	78.12%	**76.96%**	6.03%
DiGeorge	2.67	91.09%	58.31%	**85.70%**	17.65%

Assumptions
Distribution of fetal fraction for affected samples is the same as that for unaffecteds
Cutoff at NCV = 4 ensures sufficient false positive rate control
Aberrations span the minimum syndrome region
PPV: Sensitivity / specificity assumptions correspond to re-sequencing workflow performance

[0637]

[0638] **실시예 5: 분절화 및 아염색체성 CNV의 측정**

[0639] 일부 구체예들에 있어서, 증후군 탐지는 이 증후군과 연관된 빈에서 관찰된 정규화된 커버리지 (Y)에 대한 영가설 (M0) 및 대립 가설 (M1)을 평가한다. M0: 결손 (또는 다른 CNV)이 존재하지 않고, 이 커버리지은 증후군과 연관된 영역에서 두배수체 게놈의 예측과 일치된다. M1: 결손은 빈 's'에서 시작하고, 빈 "e"에서 종료되는 'ff'의 태아 분획에서 존재한다.

[0640] 각 빈 (Y_i)에 대한 정규화된 커버리지의 확률은 매개변수로써 ff, s, 그리고 e를 갖는 t-분포로써 모형화될 수 있다. 상기 모델 관계는 하기에서 더 설명된다.

[0641] M0에서 태아 분획은 Y에 영향을 주지 않기 때문에, ff는 모형화 및 산출에서 임의적으로 ff=0로 설정될 수 있다. 일부 구체예들에 있어서, M1에서 ff는 영향을 받지 않은 남성의 훈련용 세트로부터 획득된 분포 상에 값을 추정한다.

[0642] 3가지 매개변수 (ff, s, 그리고 e - 증후군-특이적 콘센수스 및 문헌의 조사 값을 이용하여 결합된 s 및 e와 함께)의 모든 가능한 값에 대하여 다중 확률이 산출될 수 있다. 그 다음 다중 확률(likelihood)이 복합되어, 관찰된 정규화된 커버리지이 주어진 M0에 대한 개연성(probability)과 동일한 커버리지이 주어진 M1에 대한 개연성이 수득된다. 상기 2개의 개연성의 비교로 스코어가 산출된다. 상기 스코어가 임계치이상이면, 이 분석에서 결손이 존재한다고 결정된다.

[0643] 증후군 특이적 영역 분석에 대하여 매개변수들에게 약간의 구속이 적용된다:

[0644] --출발 위치 's'는 이 증후군의 조사 영역 에서 있다.

[0645] --상기 결손이 "말단"으로 규정된다면, 출발 위치는 이 염색체의 시작에서 항상 고정된다.

[0646] --종료 위치 'e'는 이 증후군의 조사 영역 에서 있다.

[0647] --태아 분획 "ff"는 모체 혈장 시료의 태아 성분에 대하여 예상된 분포 에서 있다.

[0648] --결손 크기 ('e' - 's')는 최소한 콘센수스 영역의 크기에 걸쳐있다.

[0649] 2개의 각 대에서서, 관찰된 정규화된 커버리지은 t-분포에 의해 모형화된다. 결손 존재에 대하여 주어진 시료를 점수화하기 위하여, 제1 대립되는 주어진 정규화된 커버리지의 개연성은 제2 대립되는 주어진 동일한 커버리지의 개연성에 비교된다.

[0650] 증후군 경계 및 태아 분획 (ff)에 대한 시작(s)/종료(e) 위치들이 제공되면, 빈에서 정규화된 커버리지 프로파일의 확률은 t-분포에 의해 모형화된다:

$$P(Y_i \mid s, e, ff) = \begin{cases} t(\mu - 0.5 * ff, \sigma, df), & s \le i \le e \\ t(\mu, \sigma, df), & i < s \text{ or } i > e \end{cases}$$

[0651]

[0652] 증후군에 영향을 받은 태아 분획의 분포를 유도하기 위하여 모든 가능한 시작/종료 값에 걸쳐 확률이 통합될 수 있다:

$$P(ff \mid Y) = \frac{P(Y \mid ff) * P(ff)}{\int_{ff} P(Y \mid ff) dff} = \frac{\sum_{s,e} P(Y \mid s, e, ff) * P(ff) * P(s, e)}{\int_{ff} \sum_{s,e} P(Y \mid s, e, ff) * P(ff) * P(s, e) dff}$$

[0653]

[0654] 최종 결손 존재 분류 (M0 및 M1)에 대하여 2개의 대립 모델의 스코어가 비교될 수 있다. M0: No 결손 존재 (ff = 0). M1: 일부 태아 분획 ff > 0에서 결손 존재

[0655] *단계 I: 주어진 증후군에 대한 증후군 위치 시작/종료 위치의 보편성(universe) 규정*

[0656] 모든 가능한 시작/종료 위치는 증후군 경계에 근거하여 규정된다. 태아 분획 분포는 후반 산출로 연결될 수 있다. 최소 증후군 영역이 주어지면, 유효한 시작/종료 위치가 같은 개연성으로 추정된다. 이 분석은 조사 영역 및 태아 분획에서 모든 타당한 증후군 위치 하에 확률 평가를 수반한다. 증후군 영역을 규정하기 위한 허위-코드(Pseudo-code)가 하기에 제시된다. 허위코드에 이용된 매개변수들은 표 9에서 나타낸다.

[0657] if (slideStart){# non-terminal deletion

[0658] sseq = seq(start, (stop - minbins), minshift); # potential starting positions

[0659] eseq = sseq + minbins # potential stop positions

[0660] } else {# terminal deletion

[0661] sseq = start; # potential starting positions - always fixed

[0662] eseq = seq((start + minbins), (stop), minshift) # potential stop pos allowed to vary

[0663] }

[0664] gr = expand.grid(sseq, eseq) # consider all combinations of start / stop

[0665] gr = gr[gr[,2] >= gr[,1] + minbins, drop=FALSE] #enforce min. deletion region

[0666] colnames(gr) = c("start", "end")

표 9

표 9. 분절화 매개변수들

	매개변수	명명법
한 개 빈마다 (훈련용 데이터로부터 추정된)	t-분포의 평균	Params[“m”,]
	t-분포의 SD	Params[“s”,]
	t-분포의 DF	Params[“df”,]
	t-분포의 NCP	Params[“ncp”,] =0
한 증후군마다 (증후군 콘센수스 및 조사 영역에 의해 규정된)	염색체	
	조사영역의 최소 위치	Start
	조사영역의 최대 위치	Stop
	증후군 경계 조사의 경우 슬라이딩 윈도우 vs. 말단 위치	slideStart
	증후군 영역 후보에서 빈의 최소 수 (콘센수스 영역의 크기)	Minbins
	후보 증후군 영역에 대한 Min shift (#빈)	Minshift
태아 분획 분포 (훈련용 데이터로부터 추정됨)	FF 조사 공간의 통 마다 태아 분획 값 및 태아 분획의 예상 빈도에 대한 공간 조사	alphaSeq

[0667]

[0668] *단계 II: 고정된 시작 / 종료 / ff 값에 대한 로그-확률의 산출*

[0669] 고정된 시작 / 종료 / ff 값에 대한 로그-확률을 위한 입력령: y - 증후군 조사 영역 에서 주어진 시료에 대한 관찰된 정규화된 커버리지. Theta - (시작, 종료, 알파 (태아 분획)); Params - 증후군에 대한 t-분포 매개변

수. 허위-코드는 다음과 같다.

[0670] pyGivenSEA <- function (y, theta, params){

[0671] # 결손된 영역

[0672] i1 = c(theta$start : theta$end);

[0673] # 결손안된 영역

[0674] i2 = setdiff(1:length(y), i1)

[0675] ll =

[0676] # 태아 분획 @ 에서 결손 에서 영역 로그-확률 heta$alpha

[0677] sum(dt((y[i1] - params["m",i1] + 0.5*theta$alpha)/params["s",i1], params["df",i1], params["ncp",i1], log = TRUE), na.rm = TRUE) +

[0678] # @ 정상적 두배수체 상태@에서 결손 외부의 영역의 로그-확률

[0679] sum(dt((y[i2] - params["m",i2])/params["s",i2], params["df",i2], params["ncp",i2], log = TRUE), na.rm = TRUE);

[0680] return (ll)

[0681] }

[0682] 입력량: bincov (정규화된 커버리지, 조사영역에 있어서 염색체당 중앙값에 대하여 교정된; 열(column)은 시료, 행(rows)은 빈)

[0683] ll = array(NA, dim = c(length(alphaSeq), nrow(gr), ncol(bincov)));

[0684] for (k in 1:ncol(bincov)){ # for each sample

[0685] ll[,,k] = sapply(1:nrow(gr), function(i){ # for each combination of valid potential start / end positions

[0686] # calculate log-lik for each potential fetal fraction defined in the alphaSeq vector

[0687] loglik = sapply(alphaSeq, function(a) pyGivenSEA(

[0688] bincov[,k],

[0689] list(start = gr$start[i], end = gr$end[i], alpha =a),

[0690] params

[0691]))

[0692] return (loglik)

[0693] })

[0694] }

[0695] *단계 III: 실험적으로 추정된 컷-오프 값에 근거하여 분류 결정*

[0696] 상기 분석은 결손에 대하여 추정된 태아 분획을 산출하고 (결손이 없으면 0), 실험적으로 추정된 컷-오프 값에 근거하여 분류를 결정하는데 이용되는 결손 존재에 우호적인(in favor of) 스코어를 산출함으로써 진행된다. M0 및 M1의 스코어는 다음과 같이 얻을 수 있다. 이 시료 소명을 위한 스코어는 2개 대안적 가설의 스코어 비율이 될 수 있다.

$$스코어(M_0|Y) = \prod_{s,e} P(Y|s,e,ff=0)P(s,e)$$

$$스코어(M_1|Y) = \prod_{s,e} P(Y|s,e,ff)P(ff)$$

$$스코어(Y) = 스코어(M_1|Y)/스코어(M_0|Y)$$

[0697]

[0698] 이때 모든 시작 및 종료 위치는 동일한 개연성으로 추정되며,그리고 M_1에 대한 ff는 영향을 받지 않은 남성 훈련용 시료들로부터 추정된다.

[0699] 대안 모델 및 시료 소명을 위한 점수는 일부 구체예들에서 다음의 허위-코드 에서 나타낸 것과 같은 알고리즘에 의해 시행될 수 있다.

[0700] pAlpha = apply(ll, 3, apply, 1, function(x) {v = max(x); v + log(sum(exp(x - v)))})

[0701] estFf = alphaSeq[unlist(apply(pAlpha, 2, which.max))] # 결손에 대하여 가장 있음직한 ff 추정.

[0702] 비율 = estFf

[0703] logScoreM0 = apply(ll, 3, function(x){# M0 하에 스코어 계산(zero ff (alphaSeq))

[0704] y = x[which.min(abs(alphaSeq)),]

[0705] v = max(y)

[0706] v + log(sum(exp(y-v)))

[0707] })

[0708] logScoreM1 = apply(ll, 3, function(x){# 스코어 M1 산출

[0709] logPDGivenFF = apply(x, 1, function(y) {

[0710] v = max(y)

[0711] v + log(sum(exp(y-v)))

[0712] })

[0713] z = max(logPDGivenFF)

[0714] z + log(sum(exp(logPDGivenFF-z)[-1]*pff))

[0715] })

[0716] logScore = (logScoreM1 - logScoreM0)/nrow(bincov)

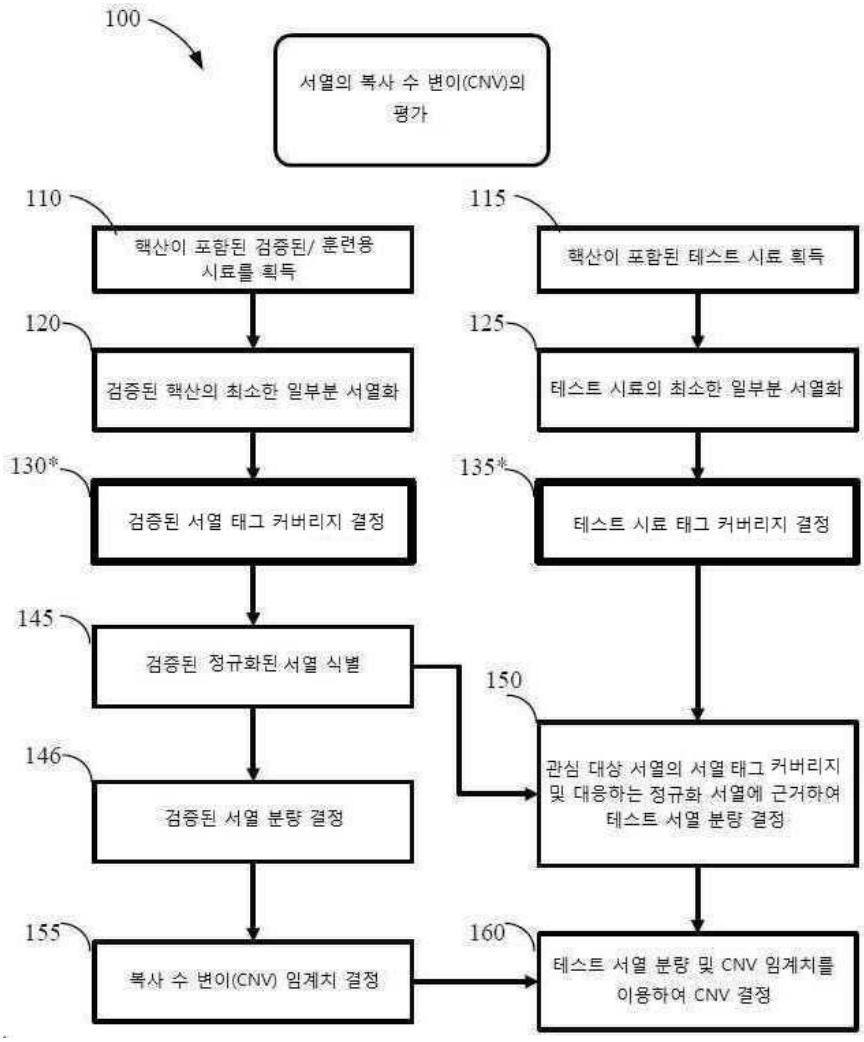
[0717] 끝으로, 증후군 특이적 CNV에 대한 소명이 만들어져야 하는지를 판단하기 위하여 테스트 시료 스코어 (Y) 에 대한 스코어는 기준과 비교된다. 일부 구체예들에 있어서, 상기 기준은 영향을 받지 않은 및/또는 영향을 받은 개체의 훈련용 시료에 기반을 둘 수 있다.

[0718] 도 26은 Cri-du-Chat 증후군을 가지는 것으로 공지된 높은 태아 분획 임상 시료에 대하여 상기에서 설명된 분절화 및 판단 분석을 이용하여 증후군 소명 수행을 보여준다. 도 27은 Cri-du-Chat 증후군을 가지는 것으로 공지된 낮은 태아 분획 임상 시료에 대한 실행을 보여준다. 이들 도면들에서 나타낸 데이터로부터 상기 분석은 높은 ff 및 낮은 ff 시료들 모두에 대하여 증후군에 의해 영역들 영향을 받은 시작점과 끝점의 우수한 예측을 제공한다. 각 도면에서 추정된 증후군 특이적 영역은 2개의 수직선 사이에 있다. 좌측 점선은 콘센수스 영역의 우측 경계를 나타낸다. 우측 점선은 조사 영역의 우측 경계를 나타낸다. 2개의 시료들에 대한 ff 값은 증후군 특이적 영역 에서 있는 수평선 세그먼트에 의해 표시된 바와 같이 또한 정확하게 추정된다.

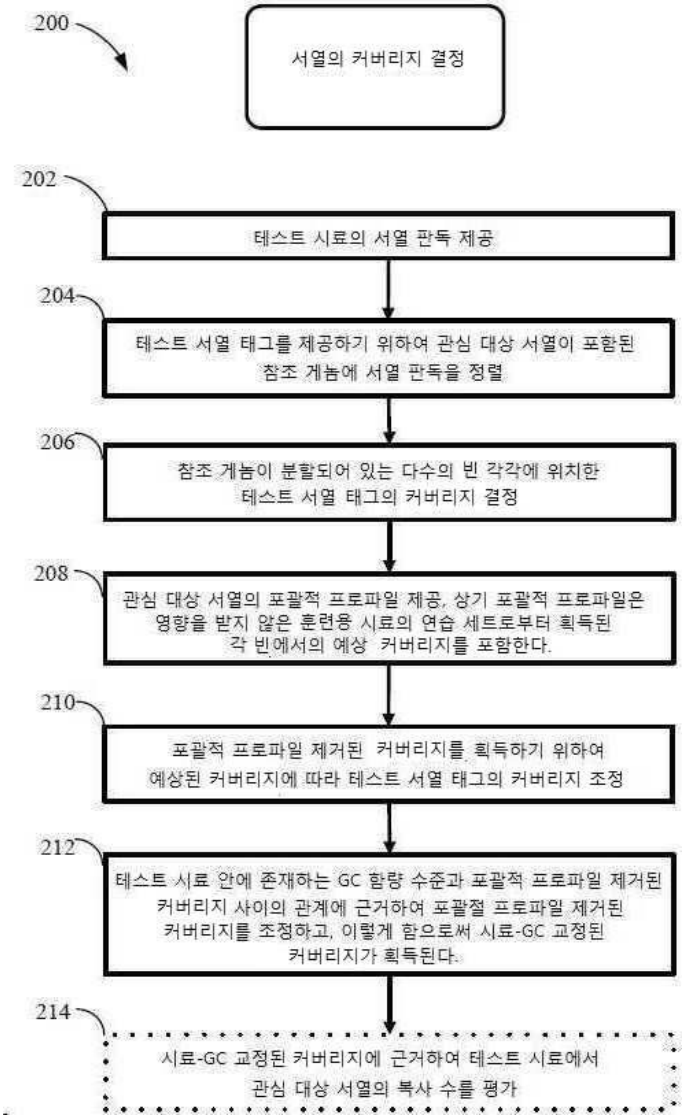
[0719] 본 명세서는 이의 사상 및 기본적인 특징을 벗어나지 않고 다른 특이적 형태로 구체화될 수 있다. 설명된 구체예들은 모든 측면에 있어서 예를 든 것이며, 이에 제한되는 것은 아니다. 따라서, 본 명세서의 범위는 전술한 설명보다는 첨부된 청구범위에 의해 나타낸다. 청구범위의 등가의 의미 및 범위 에서 있는 모든 변화는 이 청구범위의 영역에서 포함된다.

도면

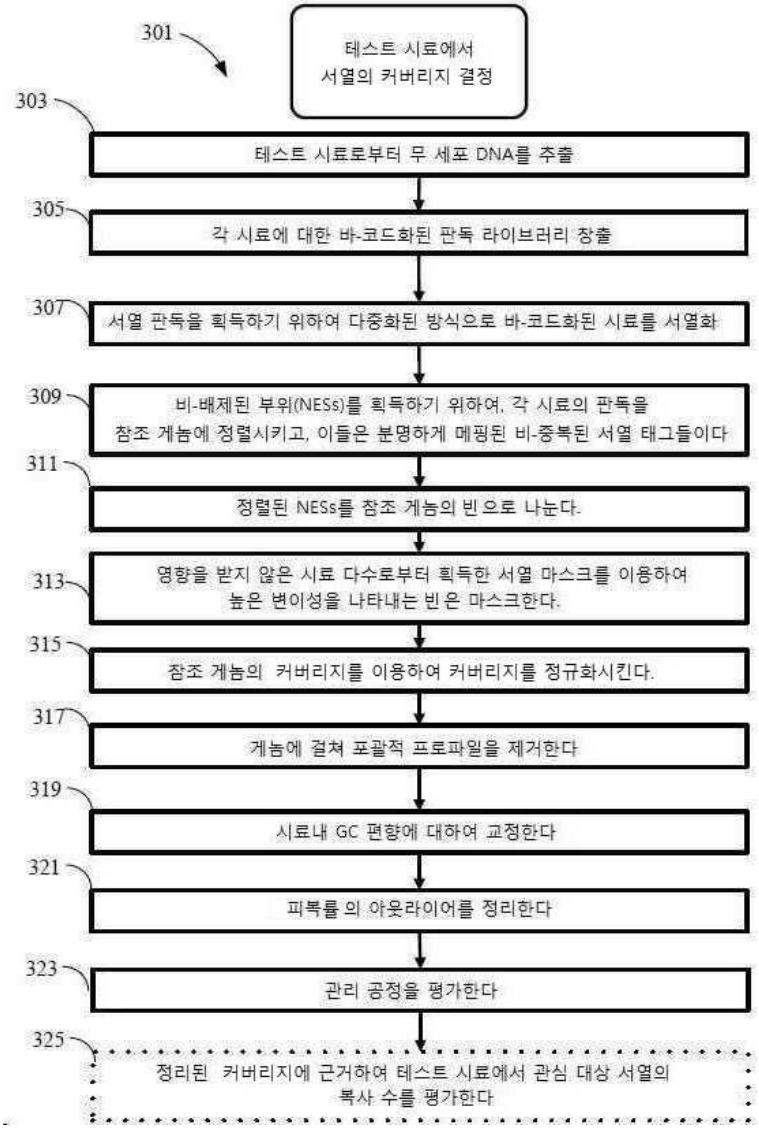
도면1



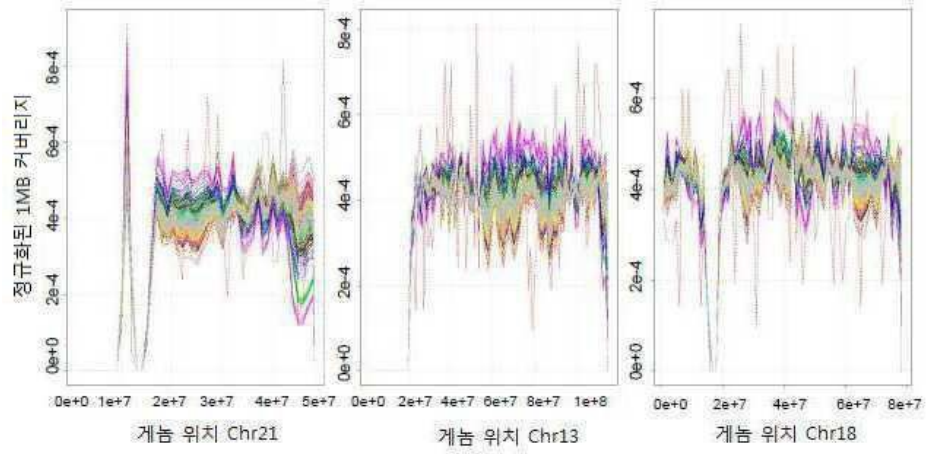
도면2


200
서열의 커버리지 결정
202
테스트 시료의 서열 판독 제공
204
테스트 서열 태그를 제공하기 위하여 관심 대상 서열이 포함된 참조 게놈에 서열 판독을 정렬
206
참조 게놈이 분할되어 있는 다수의 빈 각각에 위치한 테스트 서열 태그의 커버리지 결정
208
관심 대상 서열의 포괄적 프로파일 제공, 상기 포괄적 프로파일은 영향을 받지 않은 훈련용 시료의 연습 세트로부터 획득된 각 빈에서의 예상 커버리지를 포함한다.
210
포괄적 프로파일 제거된 커버리지를 획득하기 위하여 예상된 커버리지에 따라 테스트 서열 태그의 커버리지 조정
212
테스트 시료 안에 존재하는 GC 함량 수준과 포괄적 프로파일 제거된 커버리지 사이의 관계에 근거하여 포괄절 프로파일 제거된 커버리지를 조정하고, 이렇게 함으로써 시료-GC 교정된 커버리지가 획득된다.
214
시료-GC 교정된 커버리지에 근거하여 테스트 시료에서 관심 대상 서열의 복사 수를 평가

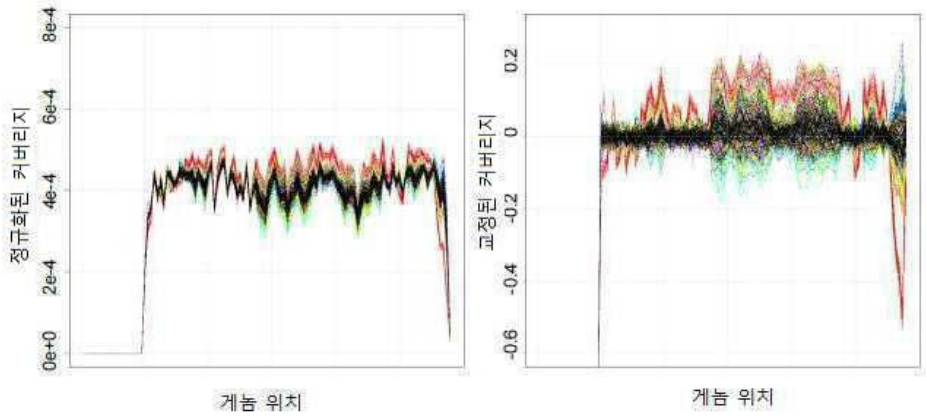
도면3a


301
테스트 시료에서
서열의 커버리지 결정
303
테스트 시료로부터 무 세포 DNA를 추출
305
각 시료에 대한 바-코드화된 판독 라이브러리 창출
307
서열 판독을 획득하기 위하여 다중화된 방식으로 바-코드화된 시료를 서열화
309
비-배제된 부위(NESs)를 획득하기 위하여, 각 시료의 판독을
참조 게놈에 정렬시키고, 이들은 분명하게 메핑된 비-중복된 서열 태그들이다
311
정렬된 NESs를 참조 게놈의 빈으로 나눈다.
313
영향을 받지 않은 시료 다수로부터 획득한 서열 마스크를 이용하여
높은 변이성을 나타내는 빈은 마스크한다.
315
참조 게놈의 커버리지를 이용하여 커버리지를 정규화시킨다.
317
게놈에 걸쳐 포괄적 프로파일을 제거한다
319
시료내 GC 편향에 대하여 교정한다
321
피복률의 아웃라이어를 정리한다
323
관리 공정을 평가한다
325
정리된 커버리지에 근거하여 테스트 시료에서 관심 대상 서열의
복사 수를 평가한다

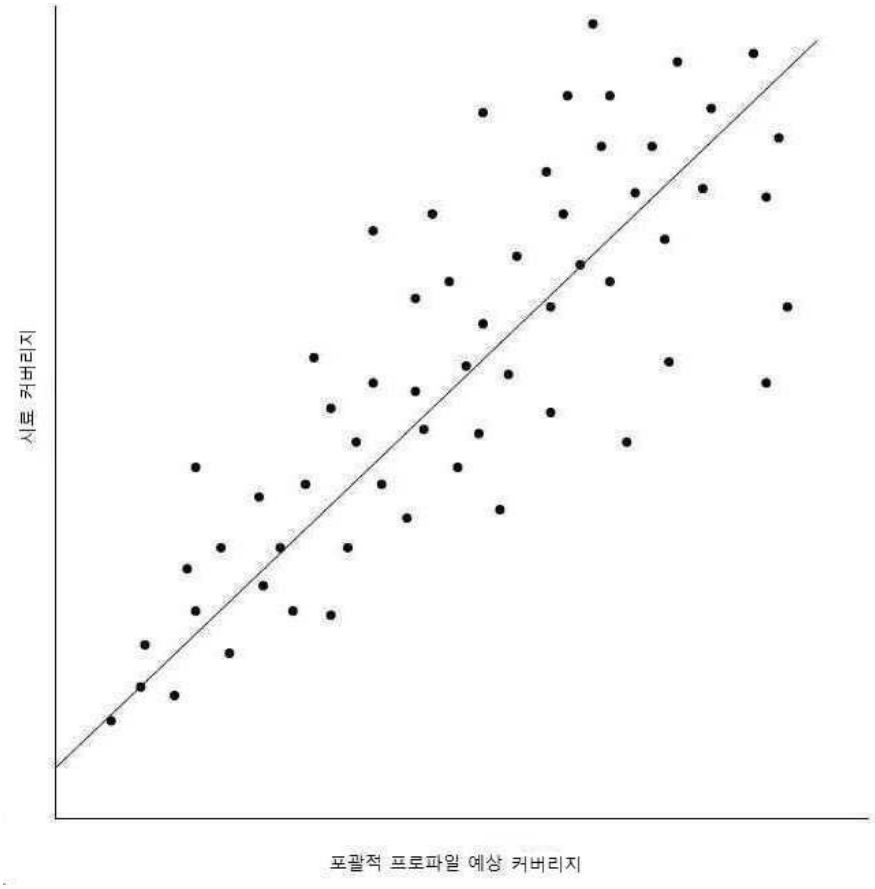
도면3b


정규화된 1MB 커버리지
0e+0
2e-4
4e-4
6e-4
8e-4
0e+0 1e+7 2e+7 3e+7 4e+7 5e+7
게놈 위치 Chr21
0e+0 2e+7 4e+7 6e+7 8e+7 1e+8
게놈 위치 Chr13
0e+0 2e+7 4e+7 6e+7 8e+7
게놈 위치 Chr18

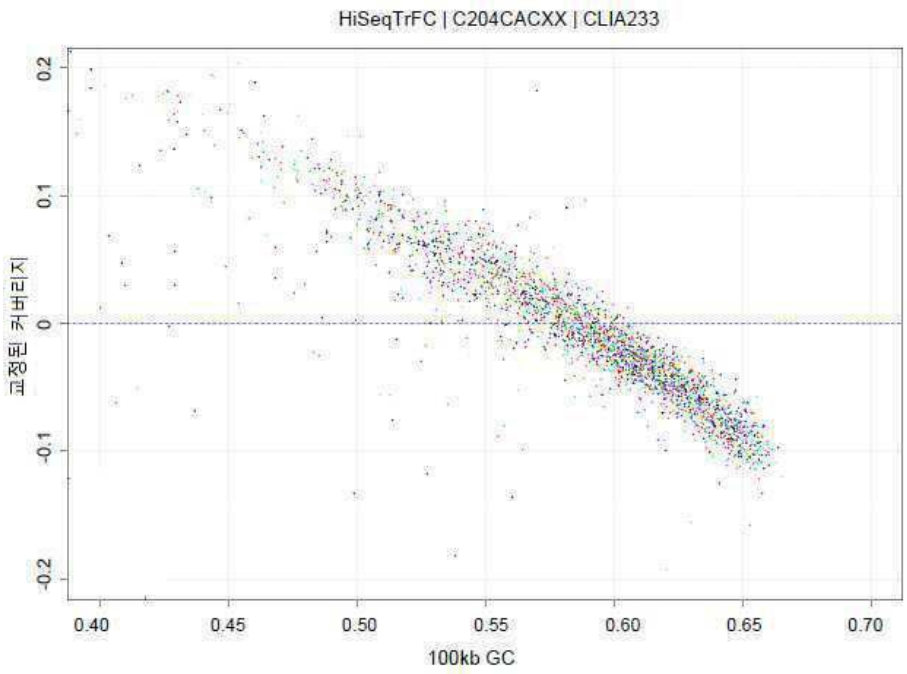
도면3c


정규화된 커버리지
0e+0
2e-4
4e-4
6e-4
8e-4
게놈 위치
교정된 커버리지
-0.6
-0.4
-0.2
0
0.2
게놈 위치

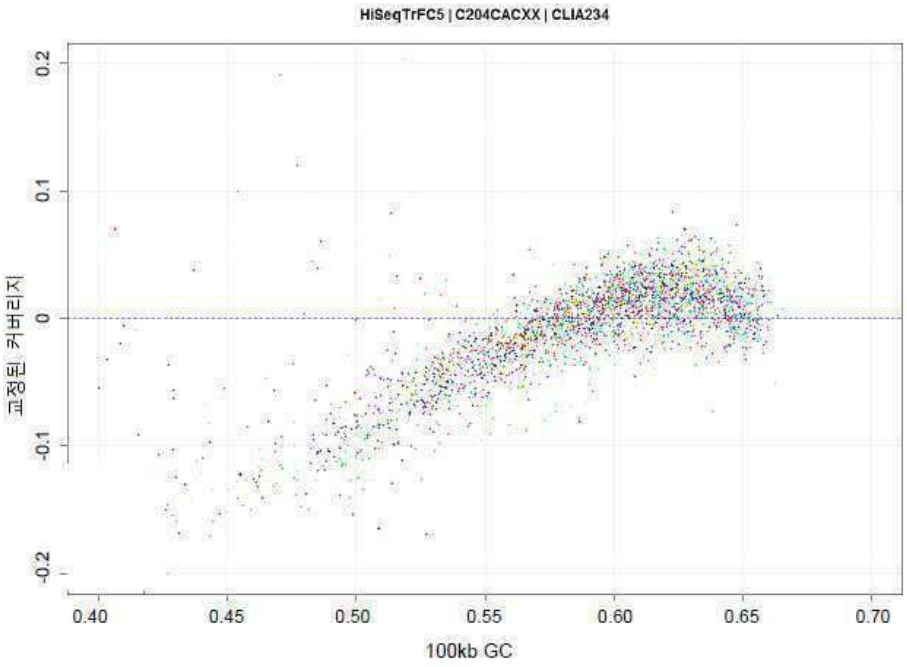
도면3d


시료 커버리지
포괄적 프로파일 예상 커버리지

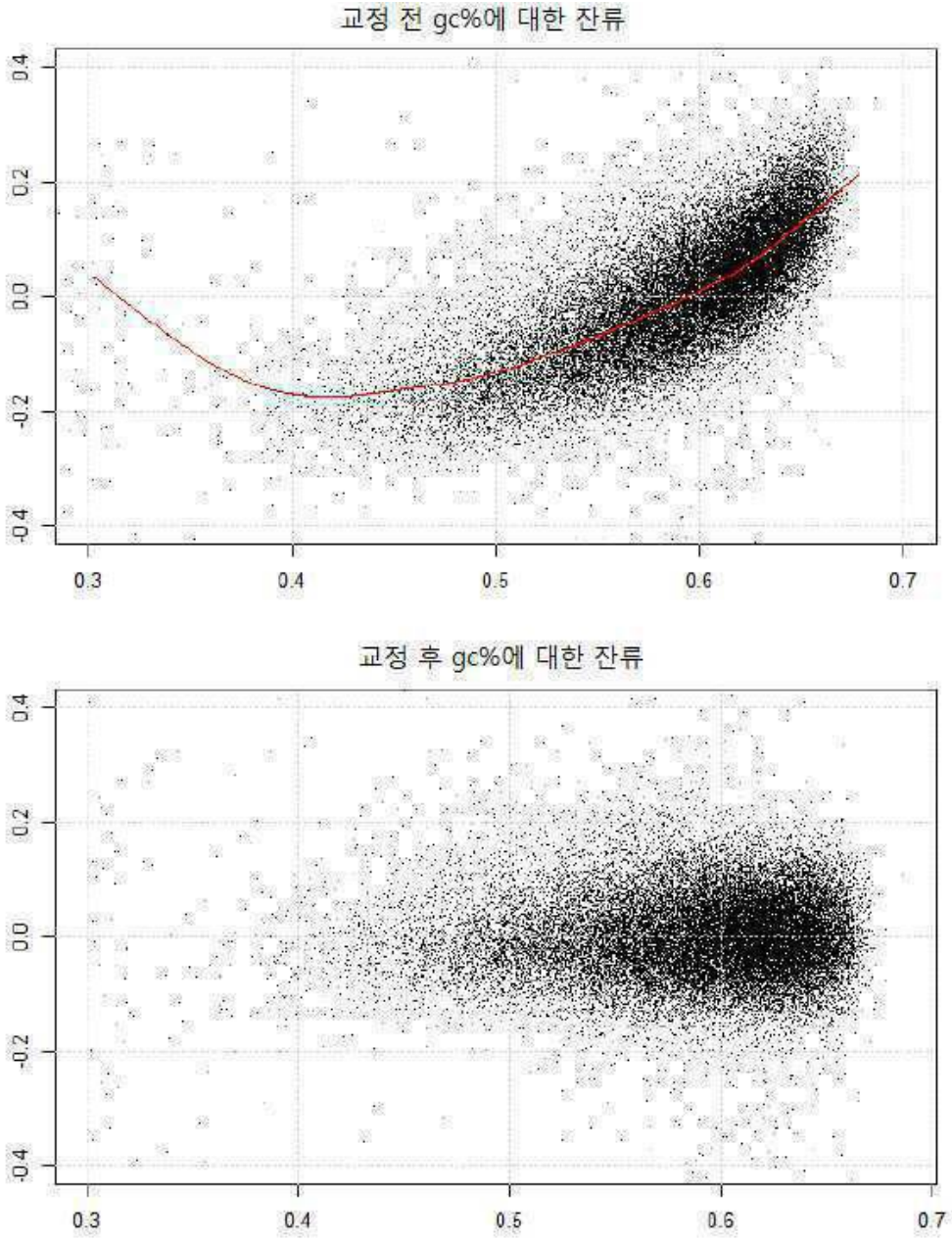
도면3e


HiSeqTrFC | C204CACXX | CLIA233
교정된 커버리지
0.2
0.1
0
-0.1
-0.2
0.40
0.45
0.50
0.55
0.60
0.65
0.70
100kb GC

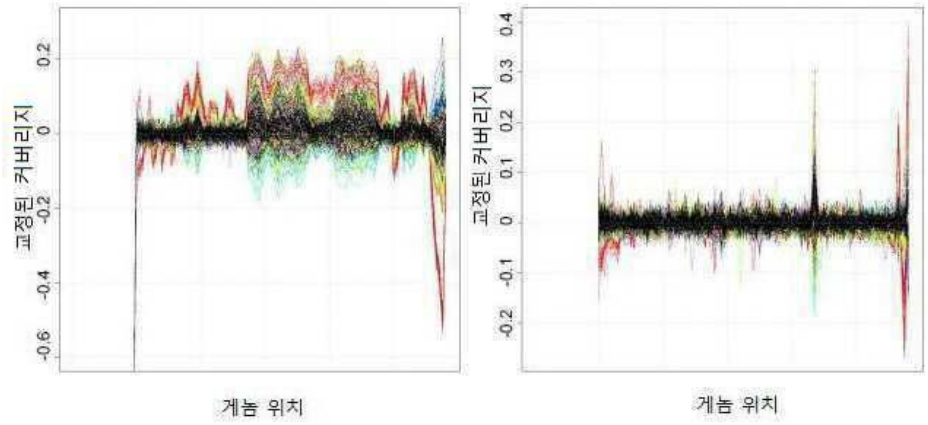
도면3f


HiSeqTrFC5 | C204CACXX | CLIA234
교정된 커버리지
0.2
0.1
0
-0.1
-0.2
0.40
0.45
0.50
0.55
0.60
0.65
0.70
100kb GC

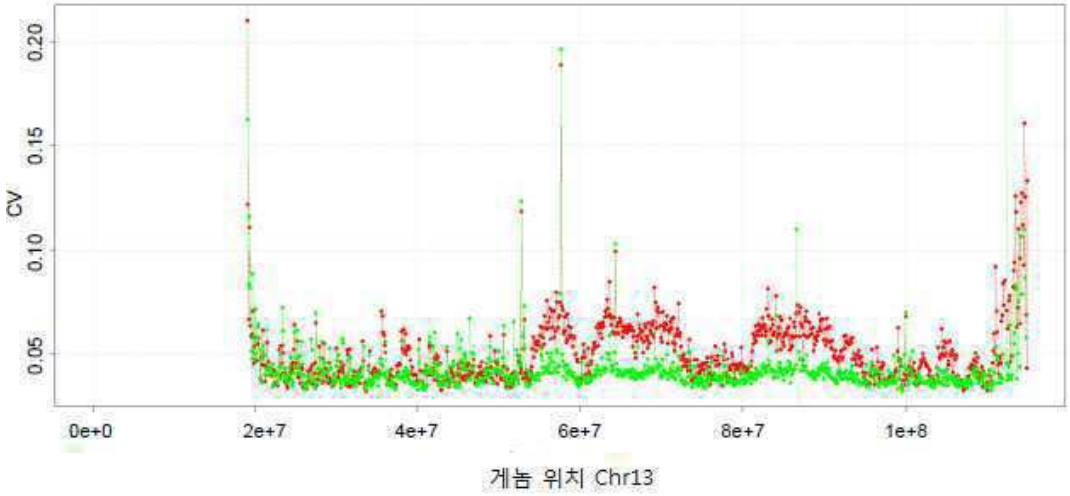
도면3g


교정 전 gc%에 대한 잔류
0.4
0.2
0.0
-0.2
-0.4
0.3
0.4
0.5
0.6
0.7
교정 후 gc%에 대한 잔류
0.4
0.2
0.0
-0.2
-0.4
0.3
0.4
0.5
0.6
0.7

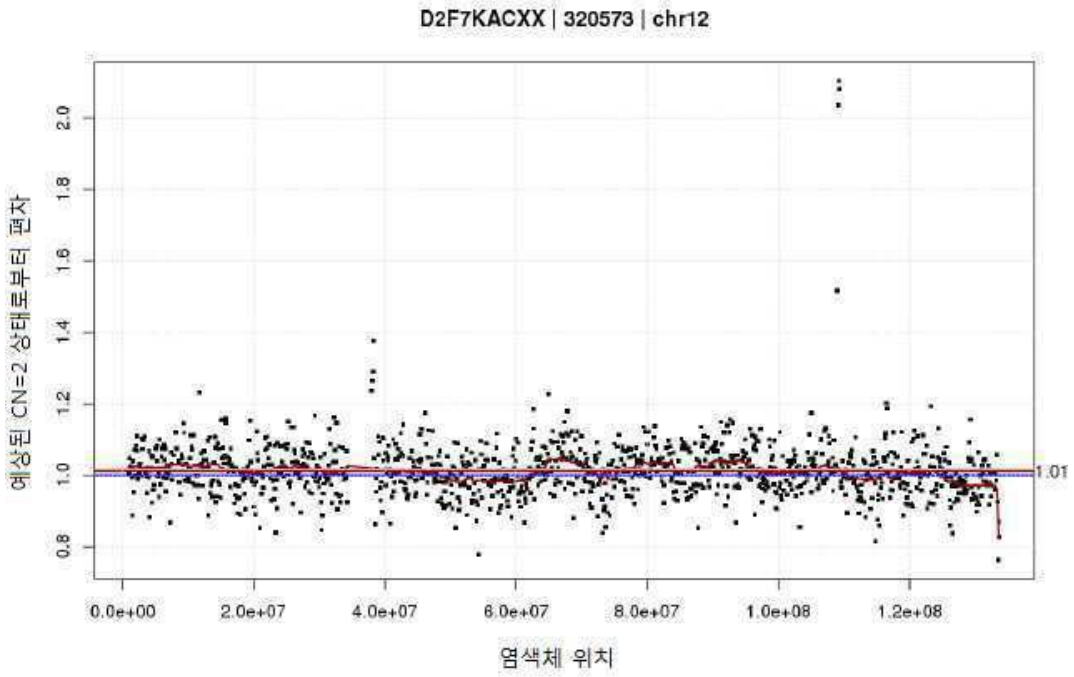
도면3h


교정된 커버리지
0.2
0
-0.2
-0.4
-0.6
게놈 위치
교정된 커버리지
0.4
0.3
0.2
0.1
0
-0.1
-0.2
게놈 위치

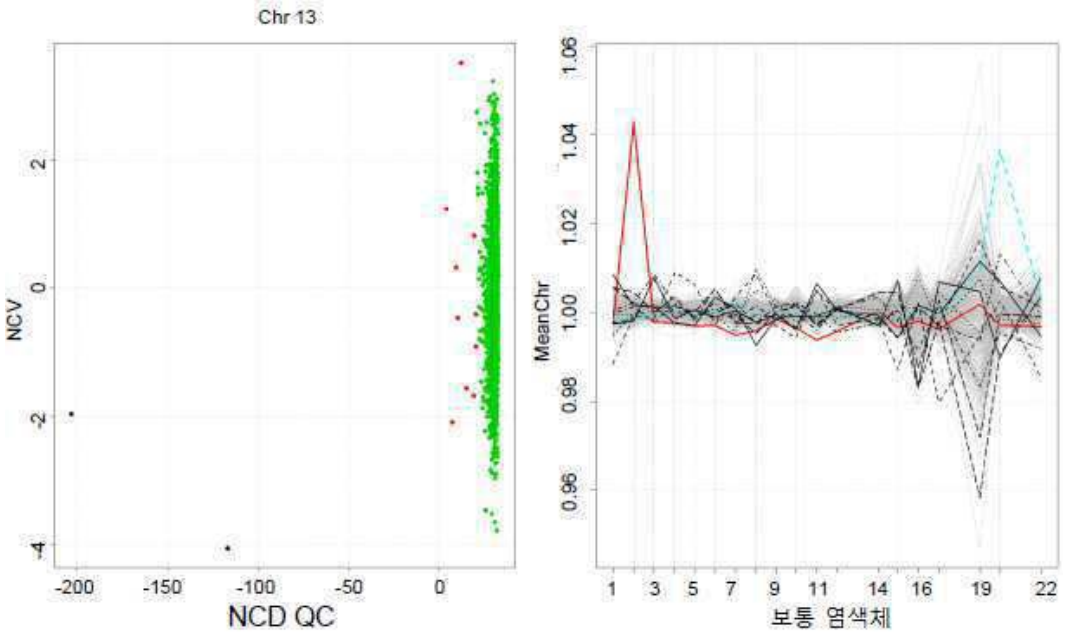
도면3i


CV
0.20
0.15
0.10
0.05
0e+0
2e+7
4e+7
6e+7
8e+7
1e+8
게놈 위치 Chr13

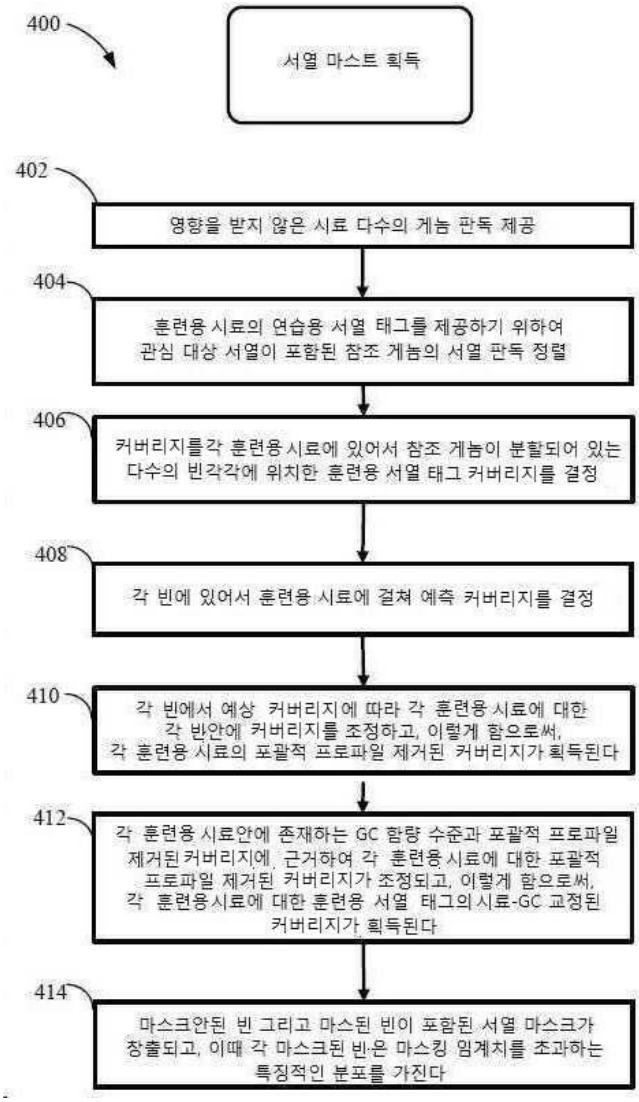
도면3j


D2F7KACXX | 320573 | chr12
예상된 CN=2 상태로부터 편차
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

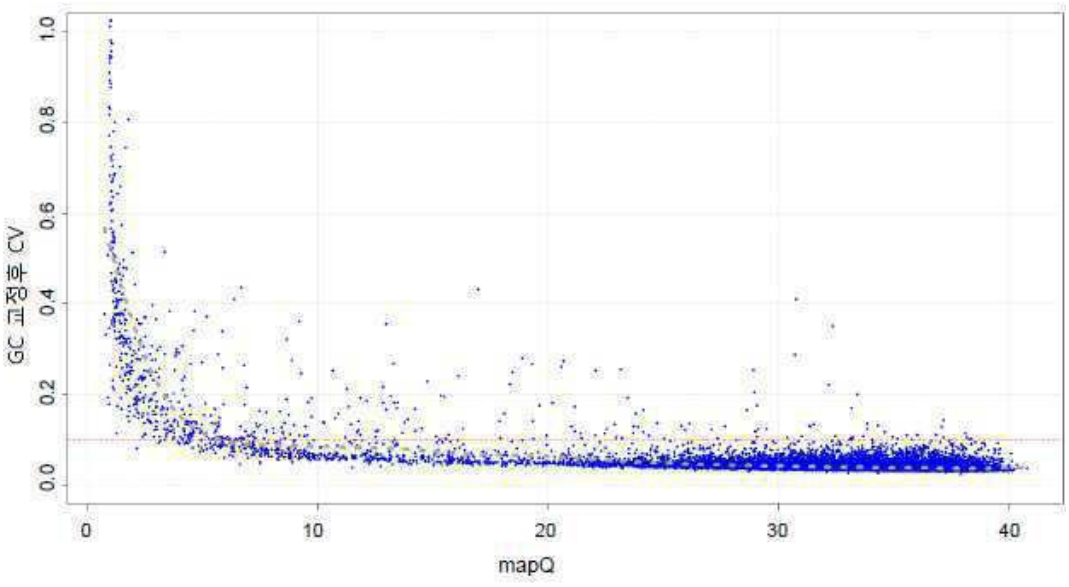
도면3k


Chr 13
NCV
2
0
-2
-4
-200
-150
-100
-50
0
NCD QC
MeanChr
0.96
0.98
1.00
1.02
1.04
1.06
1
3
5
7
9
11
14
16
19
22
보통 염색체

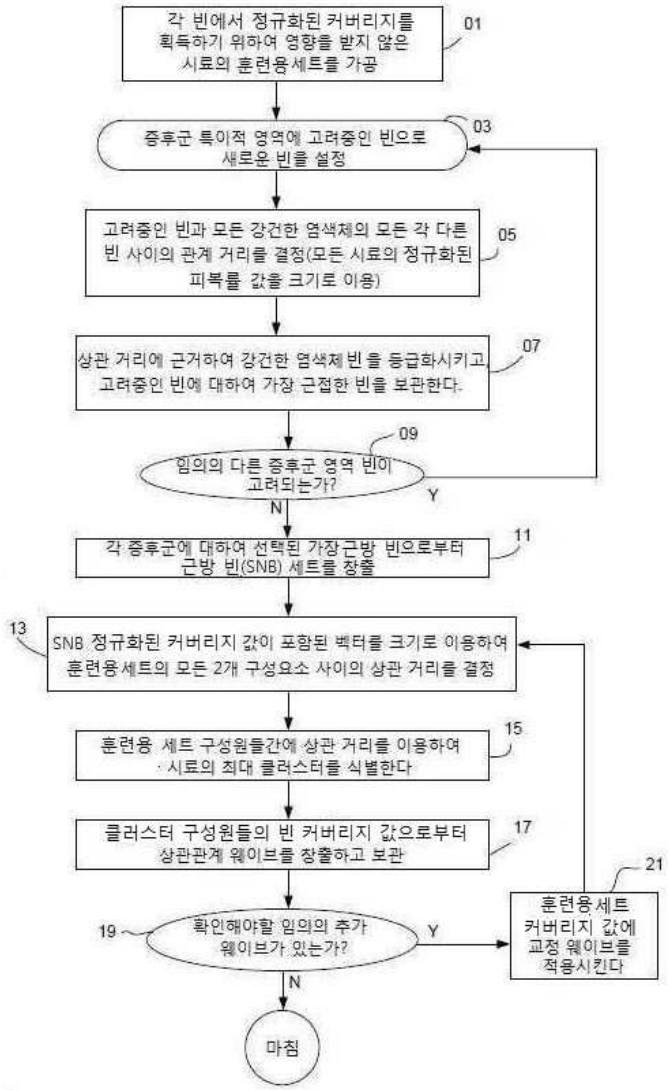
도면4a


400
서열 마스트 획득
402
영향을 받지 않은 시료 다수의 게놈 판독 제공
404
훈련용 시료의 연습용 서열 태그를 제공하기 위하여
관심 대상 서열이 포함된 참조 게놈의 서열 판독 정렬
406
커버리지를각 훈련용 시료에 있어서 참조 게놈이 분할되어 있는
다수의 빈각각에 위치한 훈련용 서열 태그 커버리지를 결정
408
각 빈에 있어서 훈련용 시료에 걸쳐 예측 커버리지를 결정
410
각 빈에서 예상 커버리지에 따라 각 훈련용 시료에 대한
각 빈안에 커버리지를 조정하고, 이렇게 함으로써,
각 훈련용 시료의 포괄적 프로파일 제거된 커버리지가 획득된다
412
각 훈련용 시료안에 존재하는 GC 함량 수준과 포괄적 프로파일
제거된커버리지에 근거하여 각 훈련용 시료에 대한 포괄적
프로파일 제거된 커버리지가 조정되고, 이렇게 함으로써,
각 훈련용시료에 대한 훈련용 서열 태그의시료-GC 교정된
커버리지가 획득된다
414
마스크안된 빈 그리고 마스된 빈이 포함된 서열 마스크가
창출되고, 이때 각 마스크된 빈은 마스킹 임계치를 초과하는
특징적인 분포를 가진다

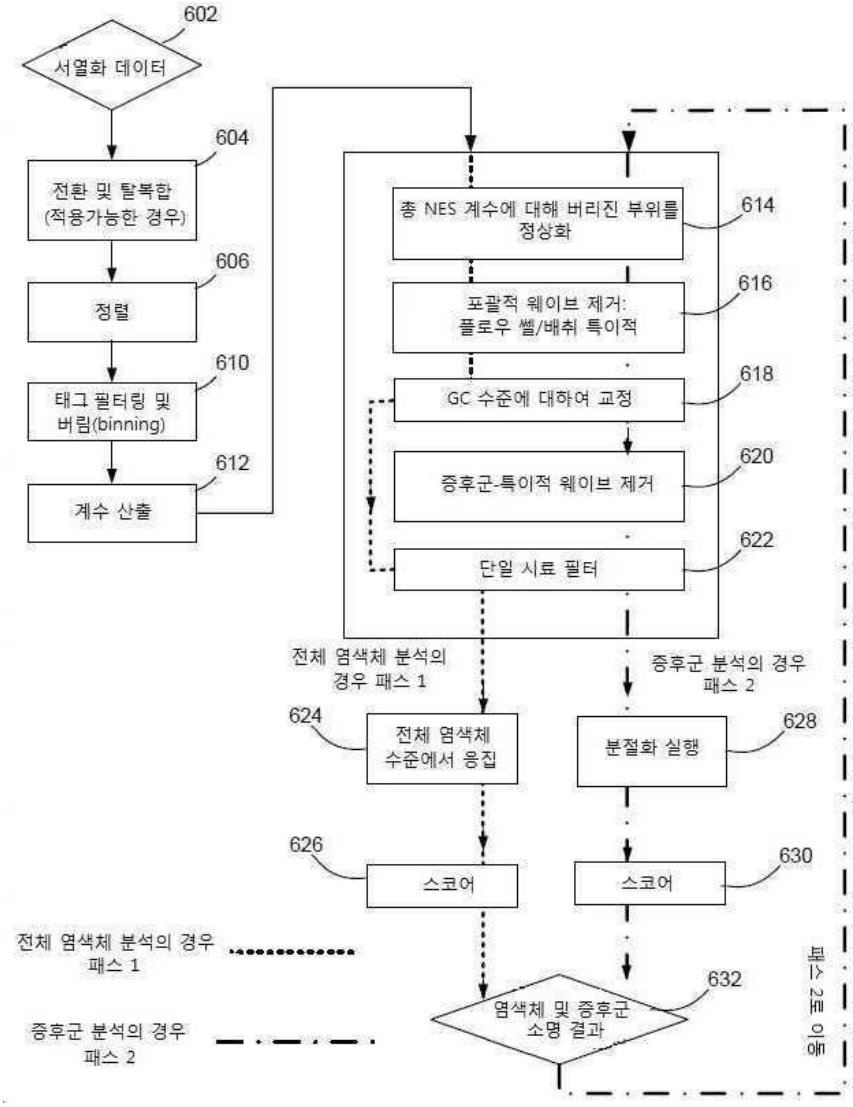
도면4b


GC 교정후 CV
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
mapQ

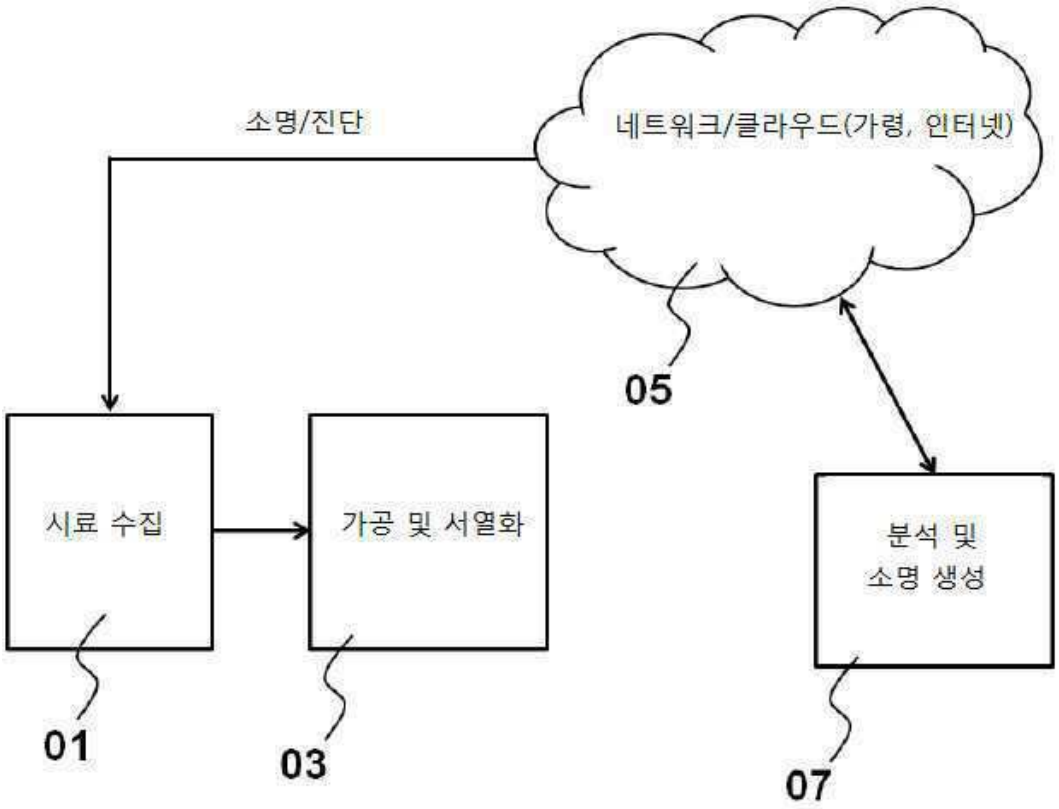
도면5


각 빈에서 정규화된 커버리지를 획득하기 위하여 영향을 받지 않은 시료의 훈련용세트를 가공
01
증후군 특이적 영역에 고려중인 빈으로 새로운 빈을 설정
03
고려중인 빈과 모든 강건한 염색체의 모든 각 다른 빈 사이의 관계 거리를 결정(모든 시료의 정규화된 피복률 값을 크기로 이용)
05
상관 거리에 근거하여 강건한 염색체 빈을 등급화시키고 고려중인 빈에 대하여 가장 근접한 빈을 보관한다.
07
임의의 다른 증후군 영역 빈이 고려되는가?
09
Y
N
각 증후군에 대하여 선택된 가장근방 빈으로부터 근방 빈(SNB) 세트를 창출
11
SNB 정규화된 커버리지 값이 포함된 벡터를 크기로 이용하여 훈련용세트의 모든 2개 구성요소 사이의 상관 거리를 결정
13
훈련용 세트 구성원들간에 상관 거리를 이용하여 시료의 최대 클러스터를 식별한다
15
클러스터 구성원들의 빈 커버리지 값으로부터 상관관계 웨이브를 창출하고 보관
17
확인해야할 임의의 추가 웨이브가 있는가?
19
Y
N
훈련용세트 커버리지 값에 교정 웨이브를 적용시킨다
21
마침

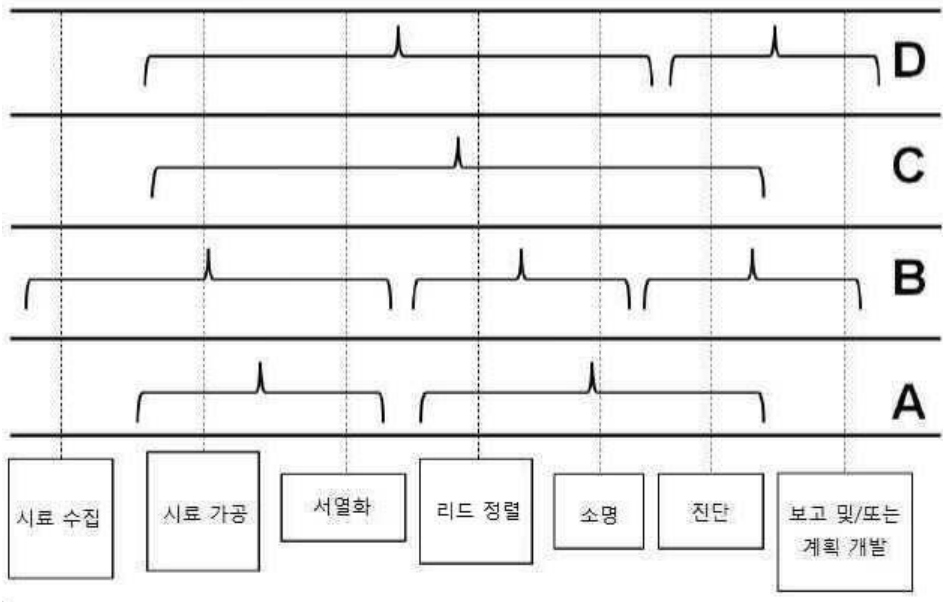
도면6


602
서열화 데이터
604
전환 및 탈복합
(적용가능한 경우)
606
정렬
610
태그 필터링 및
버림(binning)
612
계수 산출
총 NES 계수에 대해 버리진 부위를
정상화
614
포괄적 웨이브 제거:
플로우 셀/배취 특이적
616
GC 수준에 대하여 교정
618
증후군-특이적 웨이브 제거
620
단일 시료 필터
622
전체 염색체 분석의
경우 패스 1
증후군 분석의 경우
패스 2
624
전체 염색체
수준에서 응집
628
분절화 실행
626
스코어
630
스코어
632
염색체 및 증후군
소명 결과
패스 2로 이동
전체 염색체 분석의 경우
패스 1
증후군 분석의 경우
패스 2

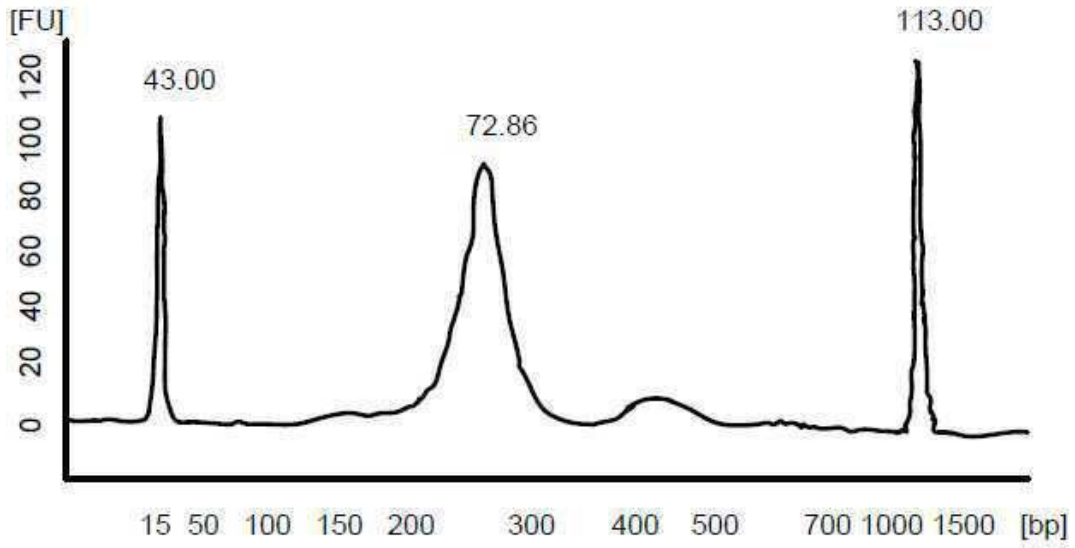
도면7


소명/진단
네트워크/클라우드(가령, 인터넷)
05
시료 수집
가공 및 서열화
분석 및
소명 생성
01
03
07

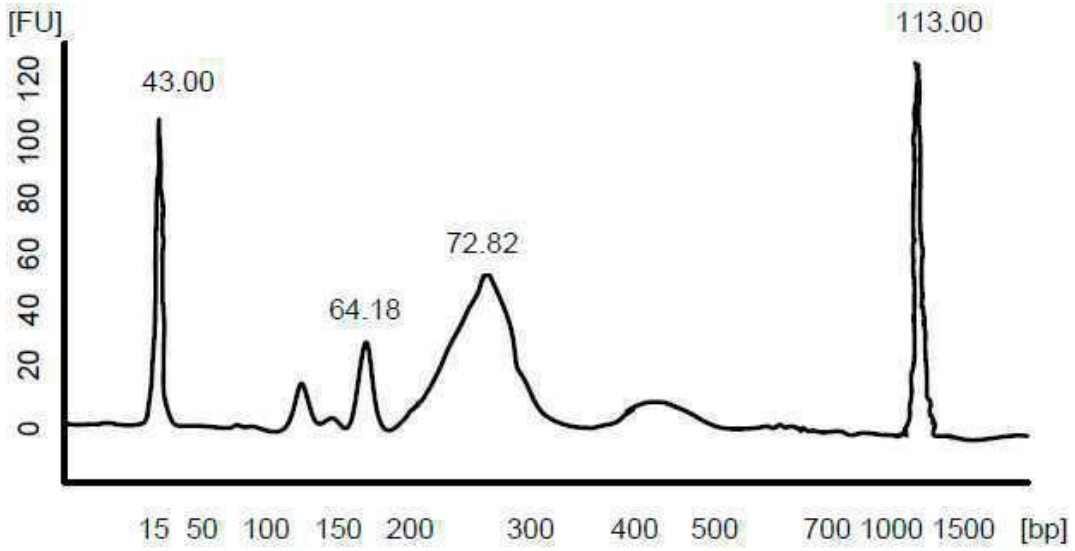
도면8


D
C
B
A
시료 수집
시료 가공
서열화
리드 정렬
소명
진단
보고 및/또는
계획 개발

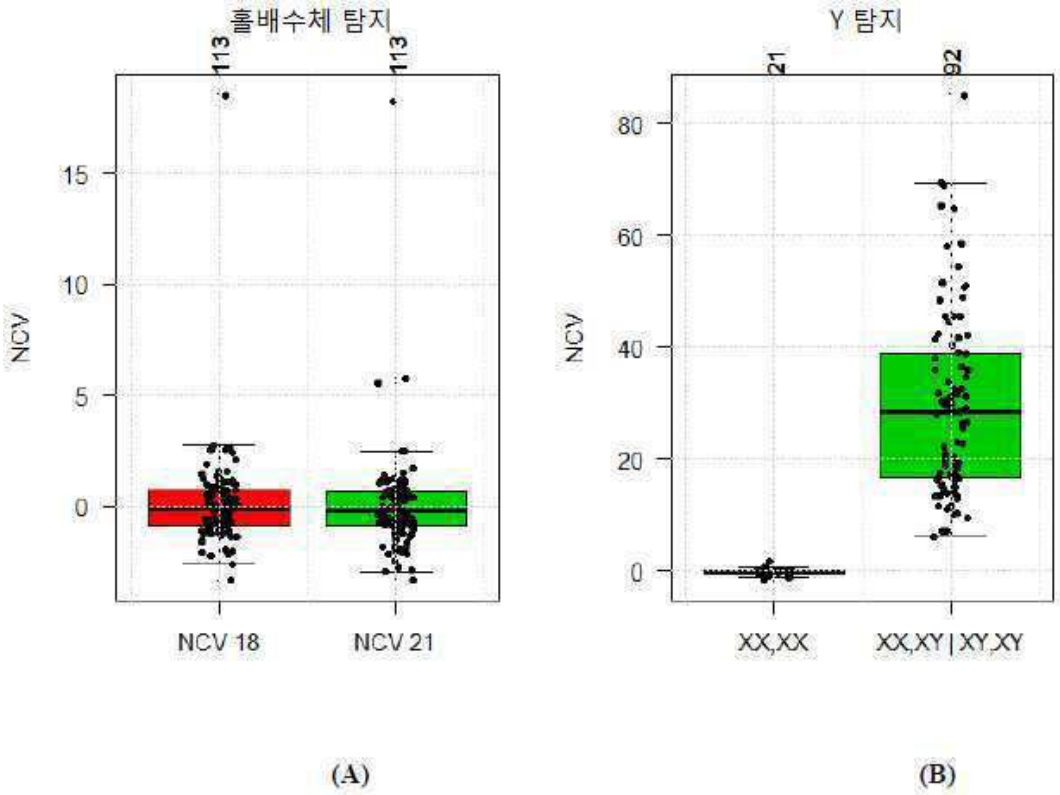
도면9a


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

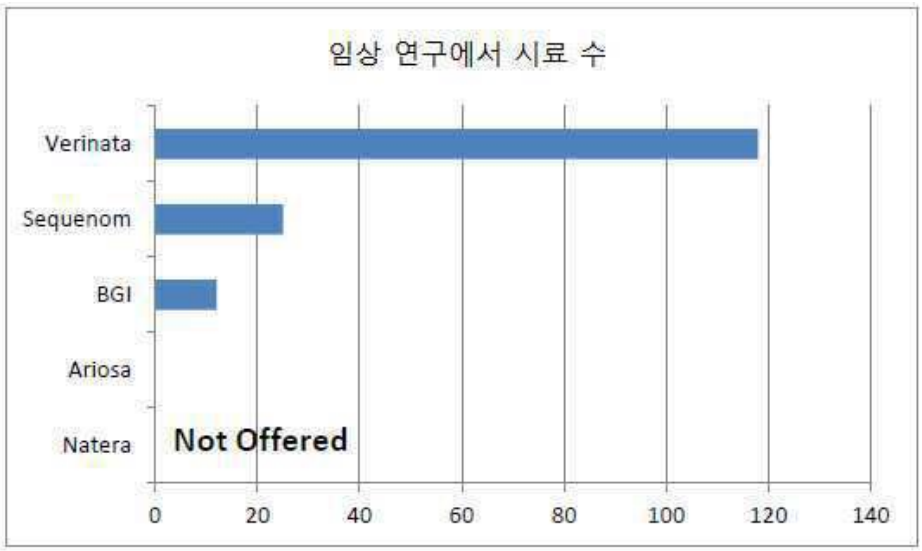
도면9b


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

도면10

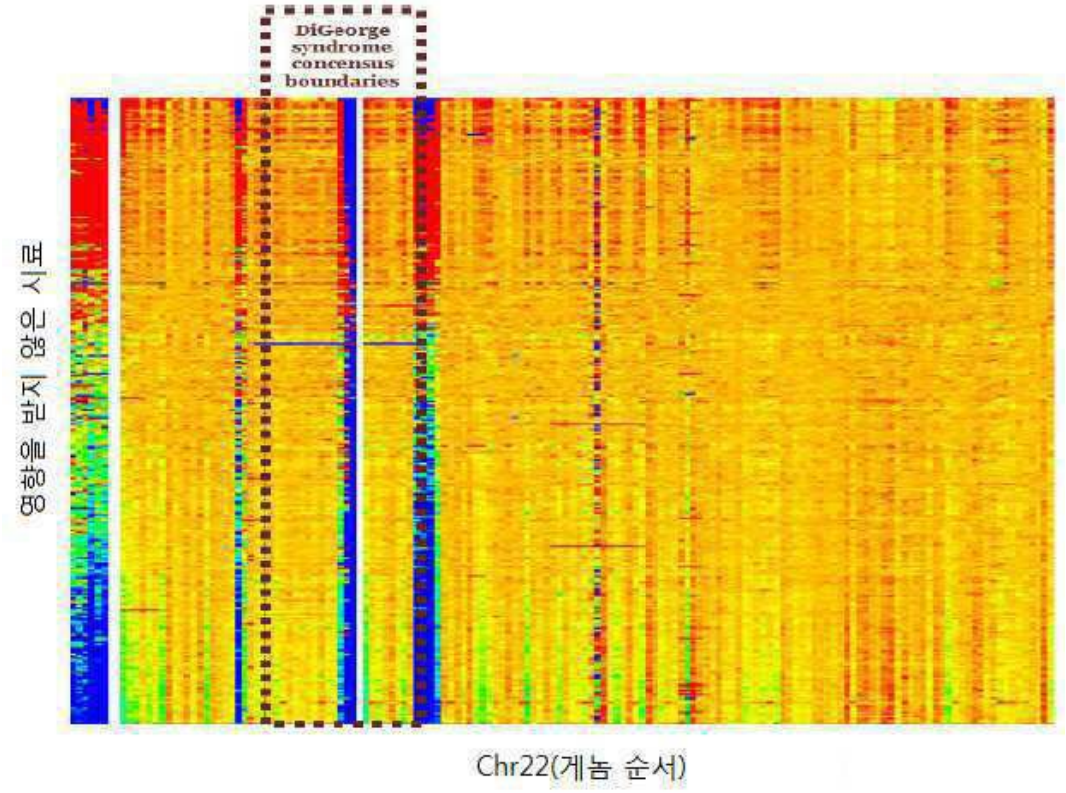


도면11

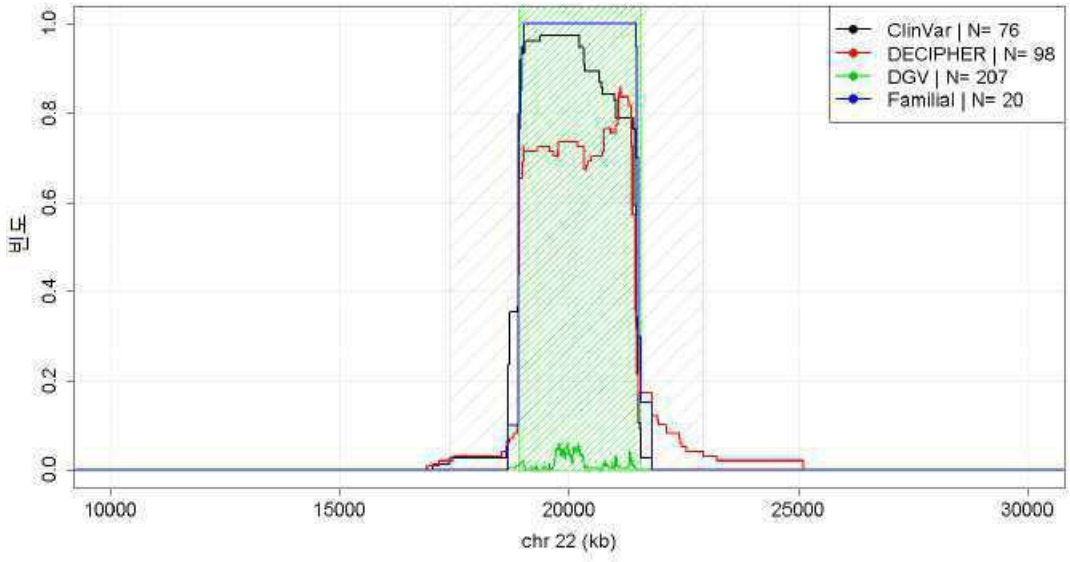


Natera	제공되지 않음
Ariosa	0
BGI	12
Sequenom	25
Verinata	118

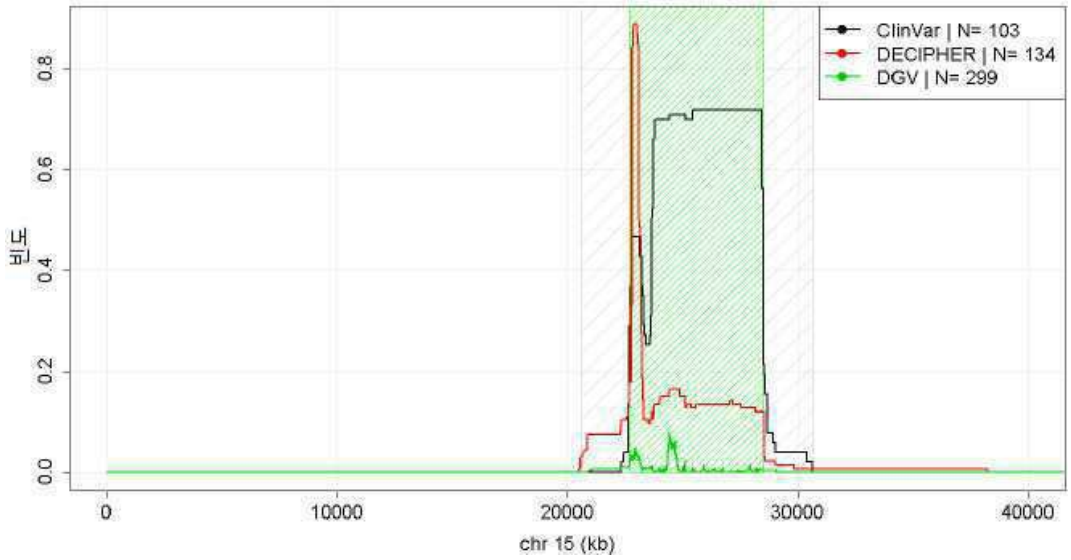
도면12


DiGeorge
syndrome
concensus
boundaries
영향을 받지 않은 시료
Chr22(게놈 순서)

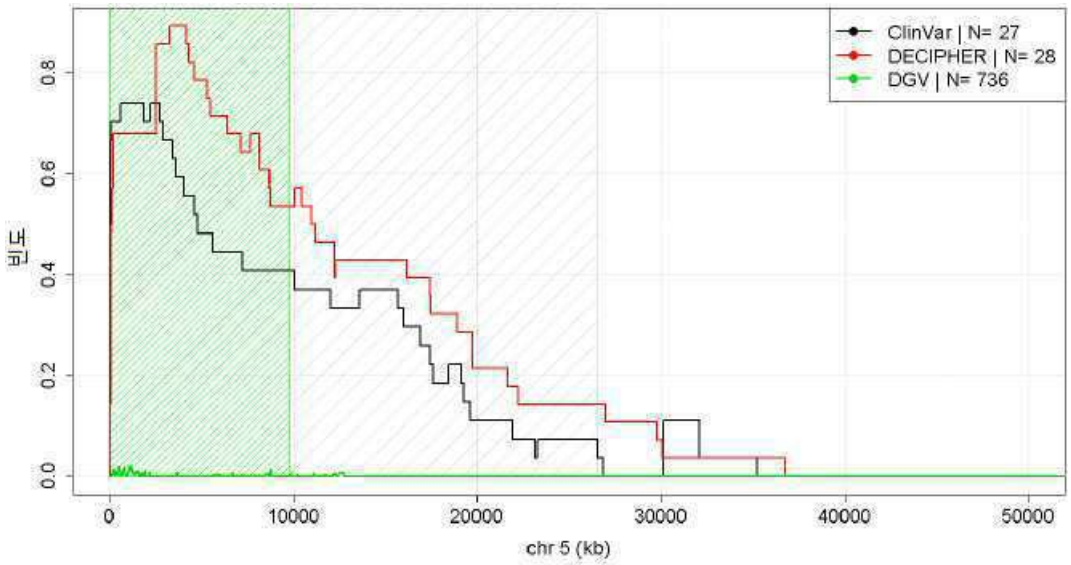
도면13


ClinVar | N= 76
DECIPHER | N= 98
DGV | N= 207
Familial | N= 20
빈도
0.0
0.2
0.4
0.6
0.8
1.0
10000
15000
20000
25000
30000
chr 22 (kb)

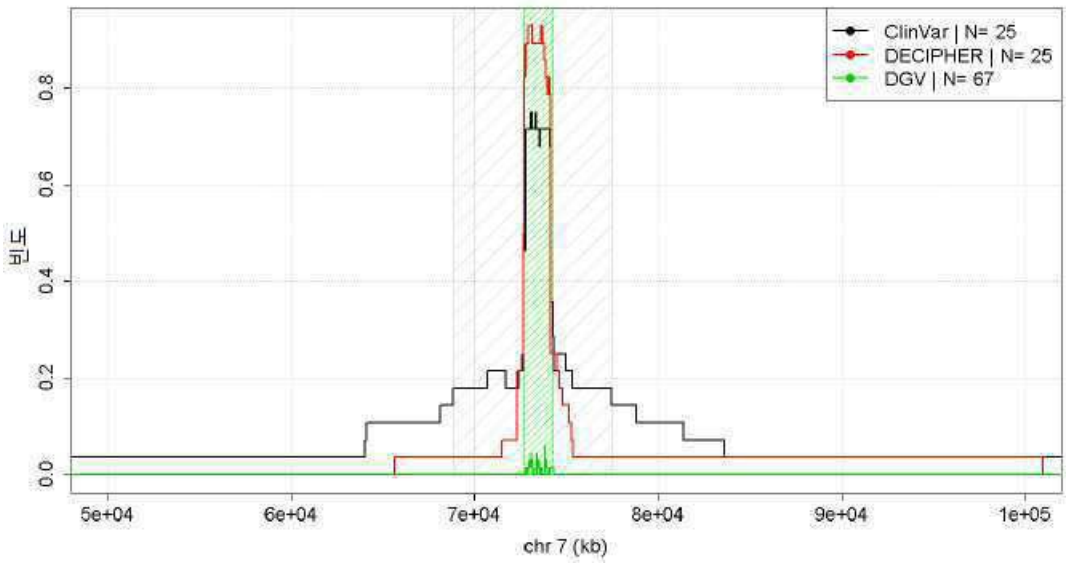
도면14


ClinVar | N= 103
DECIPHER | N= 134
DGV | N= 299
빈도
0.0
0.2
0.4
0.6
0.8
0
10000
20000
30000
40000
chr 15 (kb)

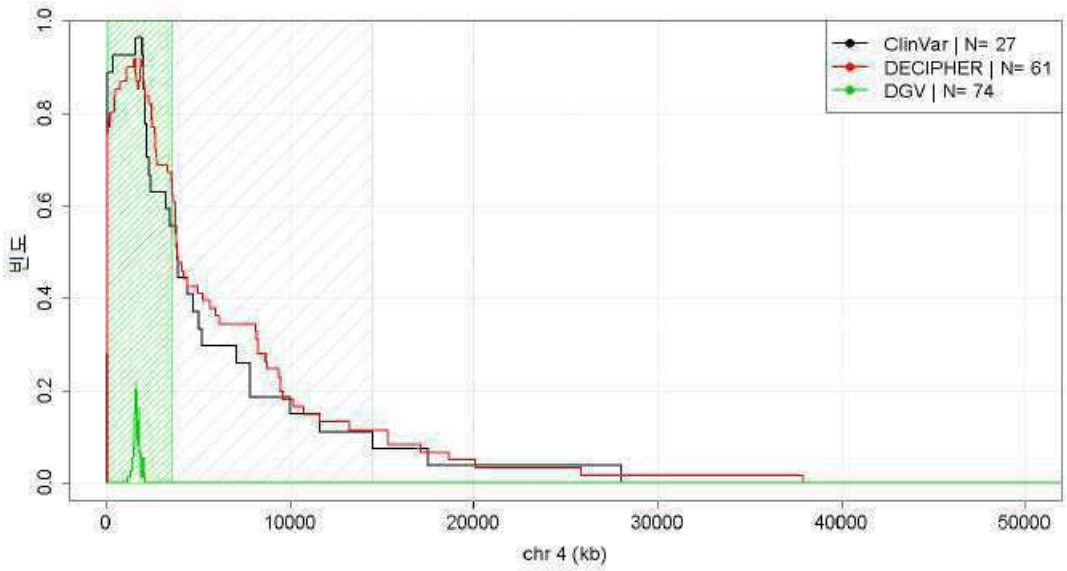
도면15


ClinVar | N= 27
DECIPHER | N= 28
DGV | N= 736
빈도
0.0
0.2
0.4
0.6
0.8
0
10000
20000
30000
40000
50000
chr 5 (kb)

도면16


ClinVar | N= 25
DECIPHER | N= 25
DGV | N= 67
빈도
0.0
0.2
0.4
0.6
0.8
5e+04
6e+04
7e+04
8e+04
9e+04
1e+05
chr 7 (kb)

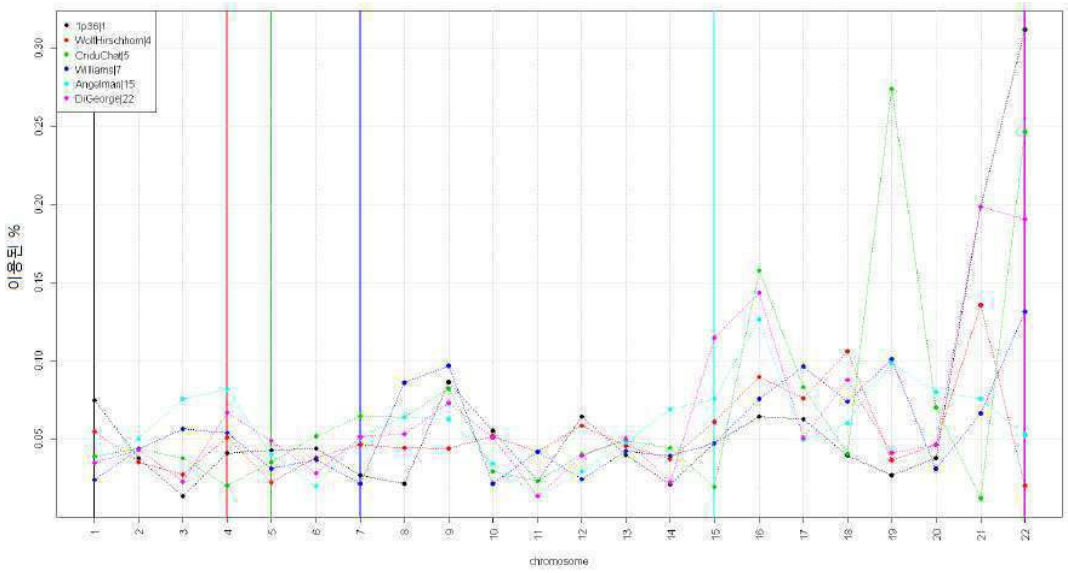
도면17


ClinVar | N= 27
DECIPHER | N= 61
DGV | N= 74
빈도
0.0
0.2
0.4
0.6
0.8
1.0
0
10000
20000
30000
40000
50000
chr 4 (kb)

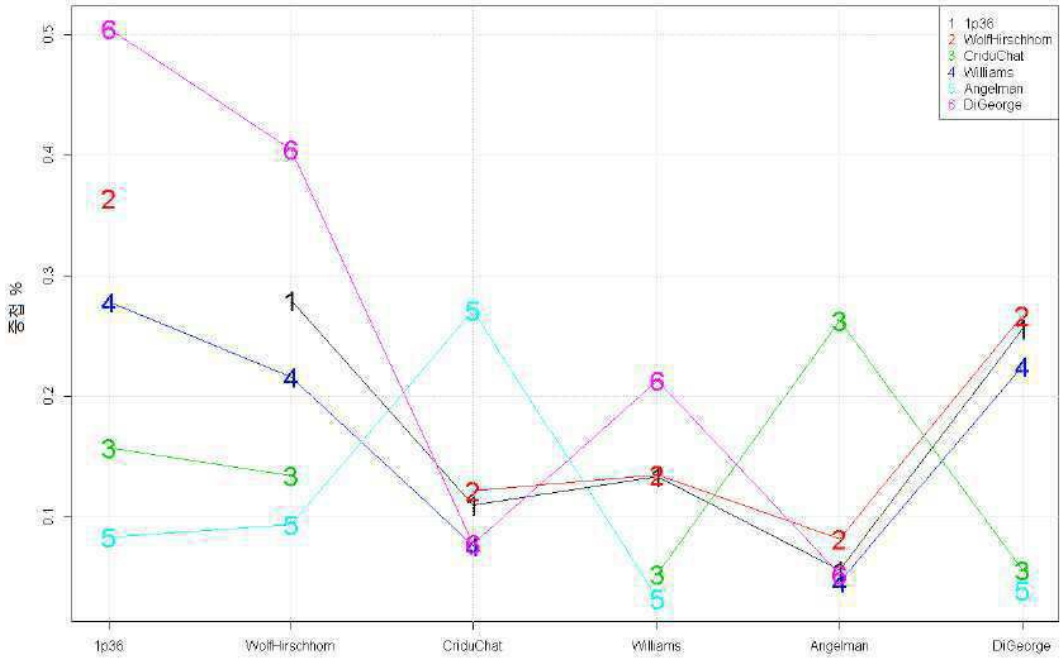
도면18


ClinVar | N= 67
DECIPHER | N= 98
DGV | N= 997
Familial | N= 67
빈도
0.0
0.1
0.2
0.3
0.4
0.5
0
10000
20000
30000
40000
50000
chr 1 (kb)

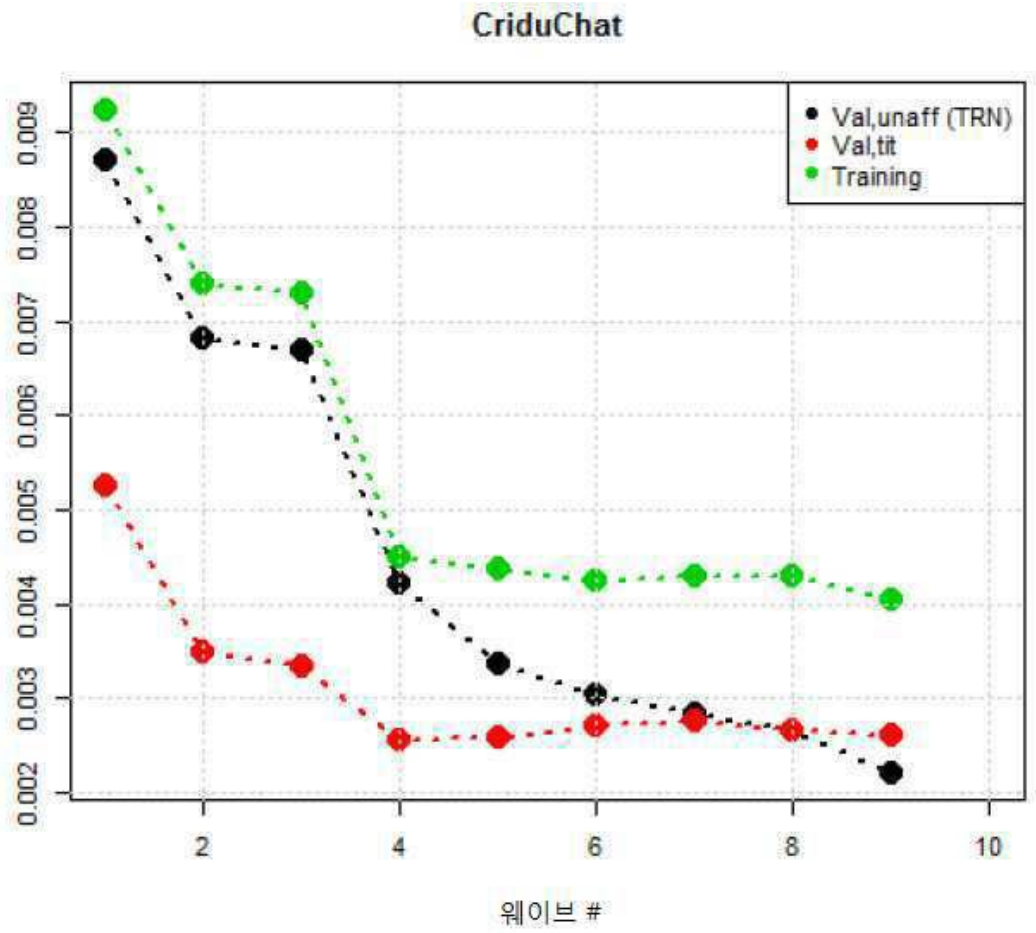
도면19


Tp36[1
WolfHirschhorn[4
CriduChat[5
Williams[7
Angelman[15
DiGeorge[22
이용된 %
chromosome

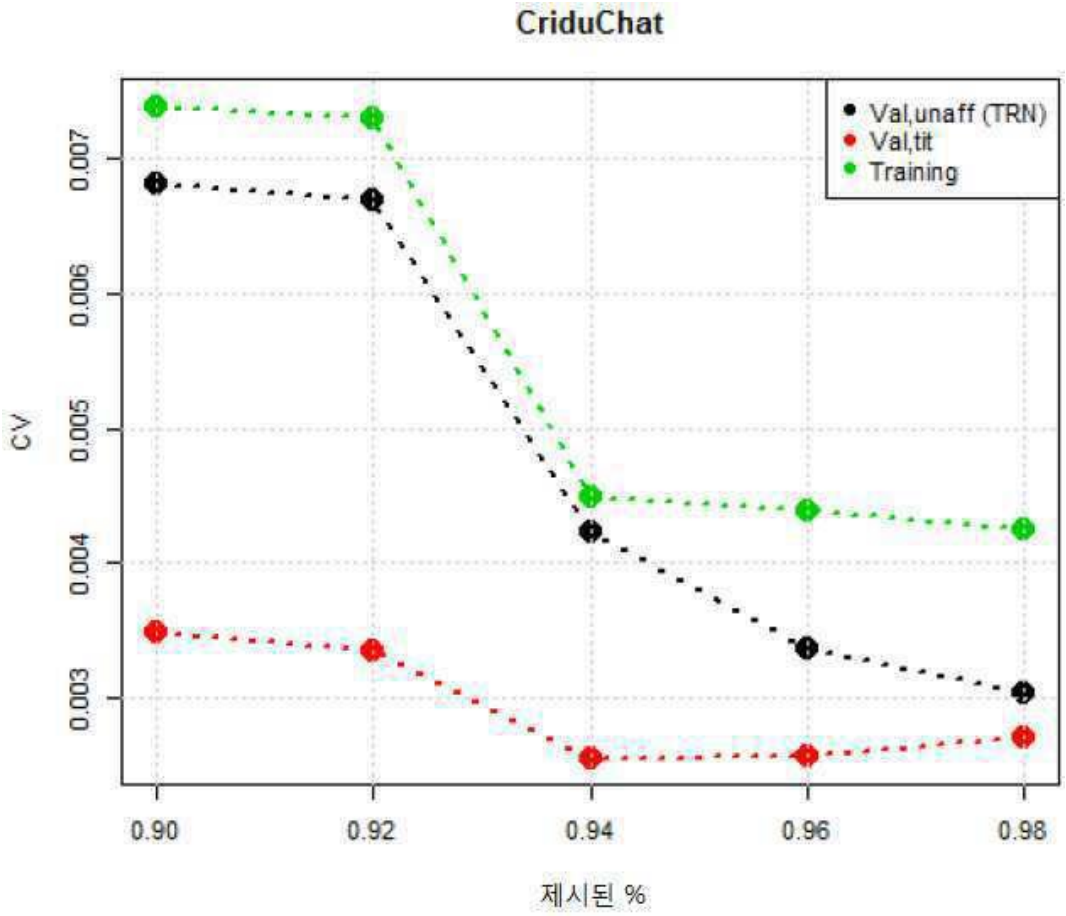
도면20


1 1p36
2 WolfHirschhorn
3 CriduChat
4 Williams
5 Angelman
6 DiGeorge
Kb 절 %
0.1
0.2
0.3
0.4
0.5
1p36
Wolf-Hirschhorn
CriduChat
Williams
Angelman
DiGeorge

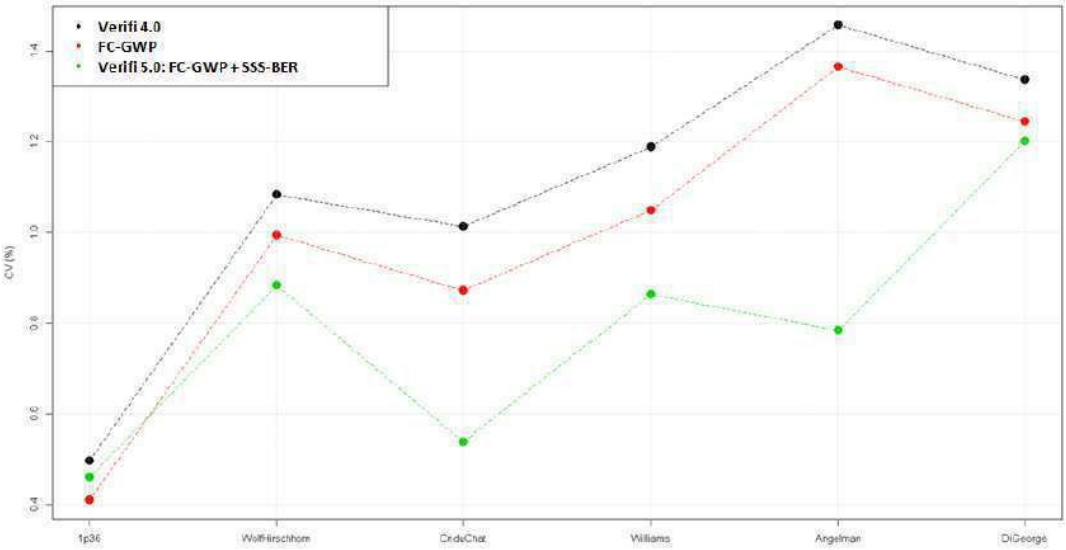
도면21


CriduChat
Val,unaff (TRN)
Val,tit
Training
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
2
4
6
8
10
웨이브 #

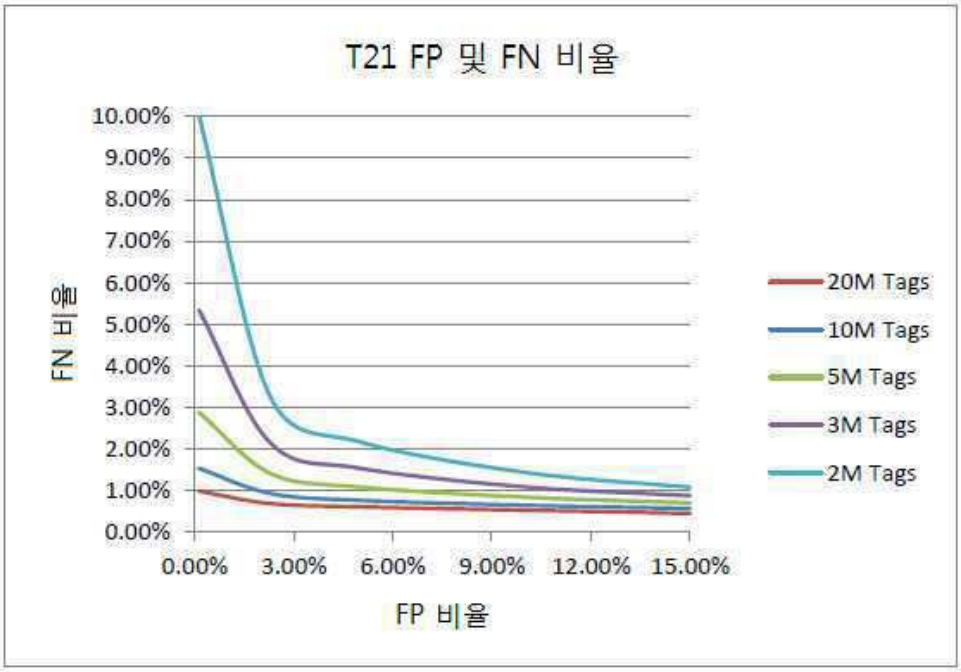
도면22


CriduChat
Val,unaff (TRN)
Val,tit
Training
CV
0.003
0.004
0.005
0.006
0.007
0.90
0.92
0.94
0.96
0.98
제시된 %

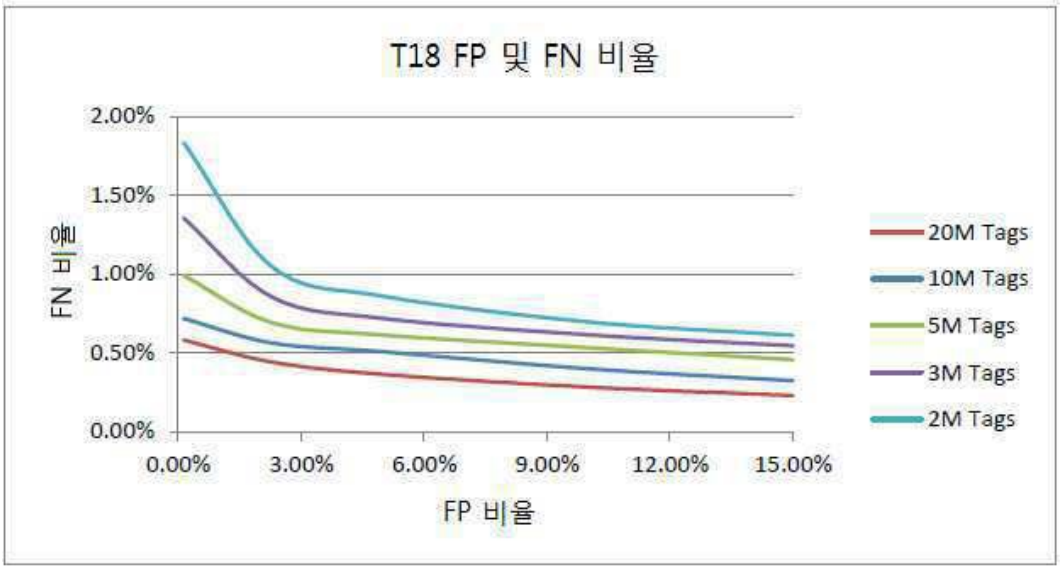
도면23


Verifi 4.0
FC-GWP
Verifi 5.0: FC-GWP + SSS-BER
CV (%)
0.4
0.6
0.8
1.0
1.2
1.4
tp36
Wolff-Hirschhorn
CriduChat
Williams
Angelman
DiGeorge

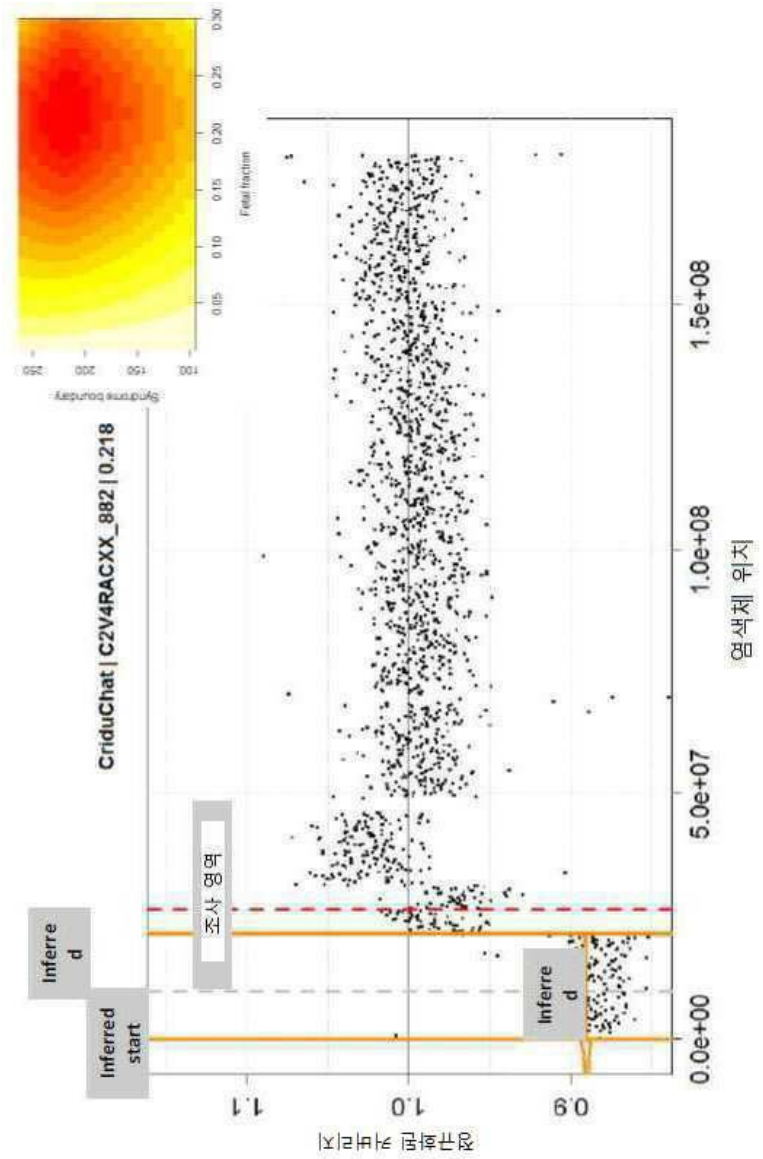
도면24


T21 FP 및 FN 비율
10.00%
9.00%
8.00%
7.00%
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
FN 비율
0.00%
3.00%
6.00%
9.00%
12.00%
15.00%
FP 비율
20M Tags
10M Tags
5M Tags
3M Tags
2M Tags

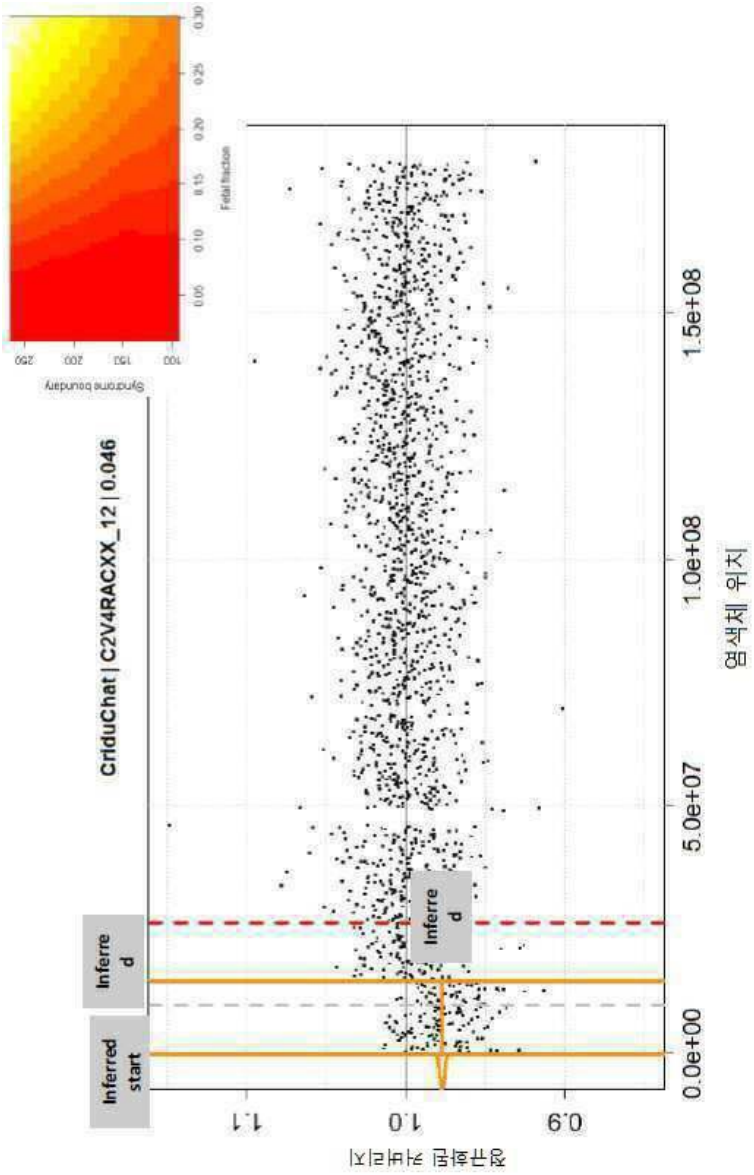
도면25


T18 FP 및 FN 비율
2.00%
1.50%
1.00%
0.50%
0.00%
FN 비율
0.00%
3.00%
6.00%
9.00%
12.00%
15.00%
FP 비율
20M Tags
10M Tags
5M Tags
3M Tags
2M Tags

도면26


CriduChat | C2V4RACXX_882 | 0.218
조사 영역
Inferred start
Inferred
Inferred
염색체 위치
0.0e+00
5.0e+07
1.0e+08
1.5e+08
정규화된 커버리지
0.9
1.0
1.1
Fetal fraction
Gyndrome boundary

도면27



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 19

【변경전】

청구항 10 내지 15, 17 및 18중 어느 한 항에 있어서,

상기 (d)작동에서의 빈의 테스트 서열 태그의 커버리지를 조정하는 작동은,

데이터 점들에 함수를 피팅시키는 작동, 여기서 각각의 데이터 점은 빈에서 테스트 시료에 대하여 대응하는 커버리지에 예상 커버리지를 관련시키고; 그리고

상기 빈에서의 커버리지를 상기 함수에 적용시킴으로써 관심 대상 서열의 빈에서의 커버리지를 조정하는 작동을 포함하는 것을 특징으로 시스템.

【변경후】

청구항 10 내지 15, 17 및 18중 어느 한 항에 있어서,

상기 (d)작동에서의 빈의 테스트 서열 태그의 커버리지를 조정하는 작동은,

데이터 점들에 함수를 피팅시키는 작동, 여기서 각각의 데이터 점은 빈에서 테스트 시료에 대하여 대응하는 커버리지에 예상 커버리지를 관련시키고; 그리고

상기 빈에서의 커버리지를 상기 함수에 적용시킴으로써 관심 대상 서열의 빈에서의 커버리지를 조정하는 작동을 포함하는 것을 특징으로 하는 시스템.