

## 명세서

### 발명의 명칭: 핵산 단편간 거리 정보를 이용한 염색체 이상 검출 방법

#### 기술분야

- [1] 본 발명은 핵산 단편간 거리 정보를 이용한 염색체의 이상을 검출하는 방법에 관한 것으로, 보다 구체적으로는 생체시료에서 핵산을 추출하여 서열정보를 획득한 다음 핵산 단편 기준값 사이의 거리를 계산하는 방법을 이용한 염색체의 이상 검출 방법에 관한 것이다.

[2]

#### 배경기술

- [3] 염색체 이상(chromosomal abnormality)은 유전적 결함과 종양 질환과 관련이 있다. 염색체 이상은 염색체의 결실 또는 중복, 염색체 중 일부의 결실 또는 중복, 또는 염색체 내의 손상(break), 전위(translocation), 또는 역위(inversion)를 의미하는 것일 수 있다. 염색체 이상은 유전적 균형의 장애 중 하나로, 태아 사망 또는 육체 및 정신 상태의 심각한 결함 및 종양 질환을 유발한다. 예컨대, 다운증후군(Down's syndrome)은 21번 염색체가 3개 존재하여(trisomy 21) 유발되는 염색체 수 이상의 혼한 형태이다. 에드워드증후군(Edwards syndrome) (trisomy 18), 파타우 증후군(Patau syndrome) (trisomy 13), 터너증후군(Turner syndrome) (XO), 및 클라인펠터 증후군(Klinefelter syndrome) (XXY) 또한 염색체 수 이상에 해당한다. 또한 종양 환자에서도 염색체 이상이 발견 된다. 예컨대 간암 환자(Liver Adenomas and adenocarcinomas) 에서 4q, 11q, 22q 영역의 중복과 13q 영역의 결실이 확인되었고, 췌장암 환자에서는 2p, 2q, 6p, 11q 영역의 중복과 6q, 8p, 9p, 21 번 염색체 영역의 결실이 확인 되었다. 이러한 영역들은 종양과 관련된 Oncogene, Tumor suppressor gene 영역과 관련이 되어 있다.
- [4] 염색체 이상은 핵형 검사(Karyotype), FISH(Fluorescent In Situ Hybridization)를 사용하여 검출 가능하다. 이러한 검출법은 시간, 노력 및 정확도 측면에서 불리하다. 또한, DNA 마이크로어레이를 염색체 이상 검출에 사용할 수 있다. 특히, 게놈 DNA 마이크로어레이 시스템의 경우, 프로브의 제작이 용이하고 염색체의 확장된 영역뿐 아니라 염색체의 인트론 영역에서의 염색체 이상을 검출할 수 있지만, 염색체 내의 위치 및 기능이 확인된 DNA 단편을 많은 수로 제작하기에 곤란하다.
- [5] 최근, 차세대 시퀀싱 기술이 염색체 수 이상 분석에 사용되고 있다(Park, H., Kim et al., Nat Genet 2010, 42, 400-405.; Kidd, J. M. et al., Nature 2008, 453, 56-64). 그러나 이 기술은 염색체 수 이상 분석을 위한 높은 coverage reading을 요구하며, CNV 측정은 독립적인 입증(validation)을 또한 필요로 한다. 따라서 비용이 매우 높고, 결과를 이해하기 어려우므로, 그 당시 일반적인 유전자 검색분석으로서

적절하지 못하였다.

- [6] 실시시간 qPCR이 현재 정량적인 유전자 분석용 첨단 기술로서 사용되는데, 이는 넓은 동역학 범위(Weaver, S. et al, Methods 2010, 50, 271-276) 및 역치 주기(threshold cycle)와 초기 타겟 양 사이에 선형적인 상관관계가 재현성 있게 관찰되기 때문이다(Deepak, S. et al., Curr Genomics 2007,8, 234-251). 그러나 qPCR 분석의 민감도는 복제수 차이를 구별할 만큼 충분히 높지 않다.
- [7] 한편, 태아 염색체 이상에 대한 기존 산전 검사 항목에는 초음파 검사, 혈중 표지자 검사, 양수검사, 융모막검사, 경피제대혈검사 등이 존재한다(Mujezinovic F, et al. Obstet Gynecol. 2007, 110(3):687-94.). 이 중 초음파 검사와 혈중 표지자 검사는 선별검사, 양수 염색체 검사는 확진 검사로 분류한다. 비침습적 방법인 초음파 검사와 혈중 표지자 검사는 태아에 대한 직접적인 시료 채취를 하지 않아 안전한 방법이지만 검사의 민감도가 80% 이하로 떨어진다(ACOG Committee on Practice Bulletins. 2007). 침습적 방법인 양수검사, 융모막검사, 경피제대혈 검사는 태아 염색체 이상을 확진할 수 있으나, 침습적 의료행위로 인한 태아의 소실 확률이 존재한다는 단점이 있다.
- [8] 1997년 Lo 등이 모체 혈장 및 혈청에서 태아 유래 유전물질을 Y 염색체 염기서열분석에 성공하여 모체 내 태아 유전물질을 산전 검사에 이용하게 되었다(Lo YM, et al. Lancet. 1997, 350(9076):485-7). 모체 혈액 내의 태아 유전물질은 태반 재형성 과정 중 세포사멸과정을 겪은 영양막 세포의 일부분이 물질 교환 기전을 통해 모체 혈액으로 들어간 것으로 실제로는 태반으로부터 유래하고 이를 cff DNA(cell-free fetal DNA)라 정의한다.
- [9] cff DNA는 빠르면 배아 이식 18일째부터, 37일째에는 대부분의 모체 혈액 내에서 발견된다. cff DNA는 300bp 이하의 짧은 가닥이며 모체 혈액 내 소량으로 존재하는 특징을 가지고 있기 때문에 이를 태아 염색체 이상 검출에 적용하기 위하여 차세대염기서열분석기법(NGS)을 이용한 대규모 병렬 염기분석 기술이 사용되고 있다. 대규모 병렬 염기분석 기술을 이용한 비침습적 태아 염색체 이상 검출 성능은 염색체에 따라 90-99% 이상의 검출 민감도를 나타내고 있으나, 위양성 및 위음성 결과가 1-10%에 해당하고 있어 이에 대한 교정 기술이 필요한 시점이다(Gil MM, et al. Ultrasound Obstet Gynecol. 2015, 45(3):249-66).
- [10] 이에, 본 발명자들은 상기 문제점들을 해결하고, 높은 민감도와 정확도의 염색체 이상 검출 방법을 개발하기 위해 예의 노력한 결과, 염색체 영역에 정렬되는 핵산 단편(fragments)을 그룹화한 다음, 핵산 단편 기준값 사이의 거리를 계산하여 정상인 그룹과 비교할 경우, 높은 민감도와 정확도로 염색체 이상을 검출할 수 있다는 것을 확인하고, 본 발명을 완성하였다.
- [11]
- [12] **발명의 요약**
- [13] 본 발명의 목적은 핵산 단편간 거리 정보를 이용한 염색체의 이상을 판정하는 방법을 제공하는 것이다.

- [14] 본 발명의 다른 목적은 핵산 단편간 거리 정보를 이용한 염색체의 이상을 판정하는 장치를 제공하는 것이다.
- [15] 본 발명의 또 다른 목적은 상기 방법으로 염색체의 이상을 판정하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공하는 것이다.
- [16]
- [17] 상기 목적을 달성하기 위하여, 본 발명은 생체시료에서 추출한 핵산 단편(fragments) 기준값 사이의 거리를 계산하여 염색체 이상을 검출하는 방법을 제공한다.
- [18] 본 발명은 또한, 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부; 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 및 선별된 핵산 단편(fragments)에 대하여 정렬된 핵산 단편의 기준값 사이의 거리를 측정하여, FD 값(Fragments Distance)을 계산하고, 계산한 FD 값을 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 FDI 값(Fragments Distance Index)를 계산하여, FDI 값이 기준값 범위 미만 또는 초과 일 경우, 염색체 이상이 있는 것으로 판정하는 염색체 이상 판정부를 포함하는 염색체 이상 검출 장치를 제공한다.
- [19] 본 발명은 또한, 컴퓨터 판독 가능한 저장 매체로서, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,
- [20] (A) 생체시료에서 핵산을 추출하여 핵산 단편을 획득하여 서열정보를 취득하는 단계;
- [21] (B) 취득한 서열정보(reads)에 기반하여 핵산 단편을 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;
- [22] (C) 선별된 핵산 단편(fragments)의 기준값 사이의 거리를 측정하여, FD 값(Fragments Distance)을 계산하는 단계; 및
- [23] (D) 상기 (C) 단계에서 계산한 FD 값을 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 FDI 값(Fragments Distance Index)를 계산하여, FDI 값이 기준 값 또는 범위 미만 또는 초과 일 경우, 염색체 이상이 있는 것으로 판정하는 단계를 통하여, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공한다.
- [24]
- [25] 본 발명은 또한, (A) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계; (B) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계; (C) 상기 정렬된 서열정보(reads)에 대하여 정렬된 리드 사이의 거리를 측정하여, RD 값(Read Distance)을 계산하는 단계; 및 (D) 상기 (C) 단계에서 계산한 RD 값을 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 RDI 값(Read Distance Index)를 계산하여, RDI 값이 기준 값 또는 범위 미만 혹은 초과일 경우, 염색체 이상이 있는 것으로 판정하는 단계를

포함하는 염색체 이상 검출 방법을 제공한다.

- [26] 본 발명은 또한, 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부; 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 및 선별된 서열정보(reads)에 대하여 정렬된 리드 사이의 거리를 측정하여, RD 값(Read Distance)을 계산하고, 계산한 RD 값을 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 RDI 값(Read Distance Index)를 계산하여, RDI 값이 기준 값 범위 미만 또는 초과일 경우, 염색체 이상이 있는 것으로 판정하는 염색체 이상 판정부를 포함하는 염색체 이상 검출 장치를 제공한다.
- [27] 본 발명은 또한, 컴퓨터 판독 가능한 저장 매체로서, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되, (A) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계; (B) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계; (C) 선별된 서열정보(reads)에 대하여 정렬된 리드 사이의 거리를 측정하여, RD 값(Read Distance)을 계산하는 단계; 및 (D) 상기 (C) 단계에서 계산한 RD 값을 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 RDI 값(Read Distance Index)를 계산하여, RDI 값이 기준 값 또는 범위 미만 또는 초과 일 경우, 염색체 이상이 있는 것으로 판정하는 단계를 통하여, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공한다.

[28]

### 도면의 간단한 설명

- [29] 도 1은 본 발명의 일 실시예에 따른 FD 값에 기반한 염색체 이상을 판정하기 위한 전체 흐름도이다.
- [30] 도 2는 싱글 엔드 시퀀싱(single-end sequencing) 방법으로 생산된 리드에서 본 발명의 FD 값을 계산하는 방법을 나타낸 개념도이다.
- [31] 도 3은 페어드 엔드 시퀀싱(paired-end sequencing) 방법으로 생산된 리드에서 본 발명의 FD 값을 계산하는 방법을 나타낸 개념도이다.
- [32] 도 4는 본 발명에서 리드 외 위치 정보를 활용하여 FD 값을 보정하는 방법에 대한 개념도이다.
- [33] 도 5는 본 발명의 일 실시예에서 페어드 엔드 시퀀싱 방법으로 생산된 리드 데이터를 기반으로 리드 외 위치 정보를 활용하지 않을 경우와 활용한 경우 계산되는 FD 값의 차이를 측정한 그래프이다.
- [34] 도 6은 본 발명의 일 실시예에 따른 RD 값에 기반한 염색체 이상을 판정하기 위한 전체 흐름도이다.
- [35] 도 7은 본 발명에의 일 실시예에 따른 RD 값 기반의 방법에서 계산하는 Read Distance에 대한 개념을 도식화한 것이다. Reads Distance 계산에 사용되는 Reads의 경우, 정렬된 방향에 관계없이 사용할 수 있고(도 7. A), 정렬된 방향을

고려하여 사용할 수 있다. (양의 방향: 도 7. B, 음의 방향: 도 7. C)

- [36] 도 8은 본 발명의 일 실시예에 따른 RD값 기반의 방법에서 X 염색체의 리드 개수(read count)와 RepRD의 분포를 도식화 한 것으로, 두 값의 관계가 선형이 아닌 비선형 관계임을 확인했다.
- [37] 도 9는 본 발명의 일 실시예에 따른 RD값 기반의 방법에서 염색체별 리드 개수(read count)와 RepRD의 분포를 도식화한 것으로, (A)는 정상 염색체, (B)는 삼염색체성(trisomy) 21번 염색체, (C)는 삼염색체성(trisomy) 18번 염색체 및 (D)는 삼염색체성(trisomy) 13번 염색체를 나타낸 것이다.
- [38] 도 10은 본 발명의 일 실시예에 따른 RD값 기반의 방법에서 계산한 RDI 값과 태아분획(도 10의 A), 임신 주수(도 10의 B) 및 G-score 값(대한민국 특허 제10-1686146호, 도 10의 C)의 관계를 각각 확인한 결과이다.
- [39] 도 11은 본 발명의 일 실시예에 따른 RD값 기반의 방법에서 정상군과 각 염색체 이수성으로 확인된 샘플에 대한 ROC 분석의 결과이다.
- [40] 도 12는 본 발명의 일 실시예에 따른 RD값 기반의 방법에서 리드 수에 따른 정확도를 확인한 결과로서, X 축은 리드의 수이며, Y축은 AUC를 의미한다.
- [41] 도 13은 본 발명의 일 실시예에 따른 RD값 기반의 방법과 리드의 수 및 염색체 이상과의 관련도를 확인한 결과이다.
- [42] 도 14는 본 발명의 일 실시예에 따른 RD값 기반의 방법과 microarray 분석 결과를 비교한 결과이다.
- [43] 도 15는 본 발명의 일 실시예에 따른 RD값 기반의 방법에서 정상인과 21번 염색체 이수성 샘플의 RepRD를 중앙값의 역수로 설정한 RDI 값 분포를 확인한 결과이다.
- [44] 도 16는 본 발명의 일 실시예에 따른 RD값 기반의 방법에서 정상인과 21번 염색체 이수성 샘플의 RepRD를 평균값으로 설정한 RDI 값 분포를 확인한 결과이다.
- [45] 도 17는 본 발명의 일 실시예에 따른 RD값 기반의 방법에서 정상인과 21번 염색체 이수성 샘플의 RepRD를 평균값의 역수로 설정한 RDI 값 분포를 확인한 결과이다.

[46]

[47] **발명의 상세한 설명 및 바람직한 구현예**

[48] 다른 식으로 정의되지 않는 한, 본 명세서에서 사용된 모든 기술적 및 과학적 용어들은 본 발명이 속하는 기술 분야에서 숙련된 전문가에 의해서 통상적으로 이해되는 것과 동일한 의미를 갖는다. 일반적으로 본 명세서에서 사용된 명명법 및 이하에 기술하는 실험 방법은 본 기술 분야에서 잘 알려져 있고 통상적으로 사용되는 것이다.

[49] 본 발명에서는, 샘플에서 획득한 서열 정보(read) 데이터를 참조 유전체에 정렬한 다음, 정렬된 핵산 단편(fragments)를 그룹화한 다음, 핵산 단편 기준값 사이의 거리를 계산하여 정상인 집단과 실험 대상자의 분석하고자 하는

염색체에서의 대표값을 비교하여 염색체 이상을 검출할 경우, 높은 민감도와 정확도로 염색체 이상을 검출할 수 있다는 것을 확인하고자 하였다.

[50] 본 발명에 따른 염색체 이상 검출 방법은 이수성 등의 태아의 염색체 이상뿐 아니라, 중양의 검출, 즉 중양의 진단이나 예후의 예측에도 이용될 수 있다.

[51]

[52] 즉, 본 발명의 일 실시예에서는, 혈액에서 추출한 DNA를 시퀀싱 한 뒤, 참조 염색체에 정렬한 다음, 핵산 단편을 전체 그룹, 정방향 그룹 및 역방향 그룹으로 그룹화한 다음, 각 그룹별로 핵산 단편 기준값 사이의 거리(fragment distance, FD)를 계산하고, 각 유전 영역당 핵산 단편 기준값 사이 거리의 대표값(RepFD)을 도출한 다음, 정규화 요소를 이용하여 RepFD ratio를 계산하고, 정상인 참조 집단에서의 RepFD ratio와 비교하여 그룹별 FDI(Fragment Distance Index) 값을 도출하였으며, 모든 그룹별 FDI 값이 기준값 미만 또는 초과일 경우, 실험 대상자의 염색체 이상이 있다고 결정하는 방법을 개발하였다(도 1).

[53]

[54] 따라서, 본 발명은 일 관점에서, 생체시료에서 추출한 핵산 단편(fragments) 기준값 사이의 거리를 계산하여 염색체 이상 검출 방법에 관한 것이다.

[55] 본 발명에 있어서, 상기 핵산 단편은 생체시료에서 추출한 핵산의 조각이면 제한없이 이용할 수 있으나, 바람직하게는 세포 유리 핵산 또는 세포 내 핵산의 조각일 수 있으나, 이에 한정되는 것은 아니다.

[56] 본 발명에 있어서, 상기 핵산 단편은 직접 서열분석하거나, 차세대 염기서열 분석을 통해 서열분석하거나 또는 비특이적 전장 유전체 증폭(non-specific whole genome amplification)을 통해 서열분석하여 얻은 것임을 특징으로 할 수 있다.

[57] 본 발명에서 상기 핵산 단편을 직접 서열분석하는 방법은 기존의 공지된 기술을 모두 사용할 수 있다.

[58] 본 발명에서 비특이적 전장 유전체 증폭을 통해 서열분석하는 방법은 랜덤 프라이머를 이용하여 핵산을 증폭한 다음, 서열 분석을 수행하는 모든 방법을 의미한다.

[59] 본 발명에 있어서, 차세대 염기서열 분석을 통한 서열분석을 이용하여 핵산 단편 기준값 사이의 거리를 계산하고 이로부터 염색체 이상 여부를 판정하는 방법은,

[60] (A) 생체시료에서 핵산을 추출하여 핵산 단편을 획득하여 서열정보를 획득하는 단계;

[61] (B) 획득한 서열정보(reads)에 기반하여 표준 염색체 서열 데이터베이스(reference genome database)에서 핵산 단편의 위치를 확인 하는 단계;

[62] (C) 상기 서열정보(reads)를 전체 서열, 정방향 서열 및 역방향 서열로 그룹화 하는 단계;

[63] (D) 상기 그룹화된 서열정보를 이용하여, 각 핵산단편의 기준값을 정의하고,

기준값 사이의 거리를 측정하여, 각 그룹별 FD 값(Fragments Distance)을 계산하는 단계; 및

- [64] (E) 상기 (D) 단계에서 계산한 각 그룹별 FD 값을 기반으로 염색체 전체 영역 또는 특정 영역 별로 각각의 FDI 값(Fragments Distance Index)을 계산하여, 각각의 FDI 값이 모두 기준값 범위 미만 혹은 초과일 경우, 염색체 이상이 있는 것으로 판정하는 단계;
- [65] 를 포함하는 방법으로 수행하는 것을 특징으로 할 수 있지만, 이에 한정되는 것은 아니다.
- [66]
- [67] 본 발명에서 용어 “염색체 이상”은 염색체에서 발생하는 다양한 변이를 의미하는데, 크게 수 이상과 구조 이상, 미세결실, 염색체 불안정성 등으로 구분될 수 있다.
- [68] 염색체의 수 이상은 염색체의 개수에서 이상이 발생하는 경우로서 예를 들어 다운 증후군(Down Syndrome, 21번째 염색체가 1개 더 많은 전체 염색체의 수가 47개), 터너 증후군(Turner Syndrome, 단일 X를 가지고 있어 염색체의 수가 45개) 및 클라인펠터 증후군(Klinefelter Syndrome, XXYY, XXXY, XXXXY 등의 염색체 수를 가짐)과 같이 전체 염색체의 개수가 23쌍 46개에서 이상이 발생하는 경우를 모두 포함할 수 있다.
- [69] 염색체의 구조 이상은 결실, 중복, 역위, 전좌, 융합(Fusion), 미세부수체 불안정성(MSI-H) 등 염색체의 개수에는 변화가 없으나 염색체의 구조에 변화가 발생하는 모든 경우를 의미한다. 예를 들어, 5번 염색체의 일부분 결실(고양이울음 증후군), 7번 염색체의 일부분 결실(필리암스 증후군), 12번 염색체의 일부분 중복(월프-허쉬호른 증후군) 등이 있을 수 있다. 종양 환자에서 발견되는 염색체 구조적 이상에는 9번과 22번 염색체 사이의 전좌(만성 골수 백혈병), 4q, 11q, 22q 영역의 중복과 13q 영역의 결실(간암), 2p, 2q, 6p, 11q 영역의 중복과 6q, 8p, 9p, 21번 염색체 영역의 결실(췌장암), TMPRSS2-TRG 유전자 융합(전립선암), 염색체 전반에 걸친 미세부수체 불안정성(대장암) 등이 확인 되었다. 이러한 영역들은 종양과 관련된 종양 유전자(oncogene), 종양 억제 유전자(tumor suppressor gene) 영역과 관련이 되어 있으나 위에 기술된 내용에 한정되는 것은 아니다.
- [70]
- [71] 본 발명에 있어서,
- [72] 상기 (A) 단계는
- [73] (A-i) 혈액, 정액, 질 세포, 모발, 타액, 소변, 구강세포, 태반세포 또는 태아세포를 포함하는 양수, 조직세포 및 이의 혼합물에서 핵산을 수득하는 단계;
- [74] (A-ii) 채취된 핵산에서 솔팅-아웃 방법(salting-out method), 컬럼 크로마토그래피 방법(column chromatography method) 또는 비드 방법(beads method)을 사용하여 단백질, 지방, 및 기타 잔여물을 제거하고 정제된 핵산을

수득하는 단계;

- [75] (A-iii) 정제된 핵산 또는 효소적 절단, 분쇄, 수압 절단 방법(hydroshear method)으로 무작위 단편화(random fragmentation)된 핵산에 대하여, 싱글 엔드 시퀀싱(single-end sequencing) 또는 페어 엔드 시퀀싱(pair-end sequencing) 라이브러리(library)를 제작하는 단계;
- [76] (A-iv) 제작된 라이브러리를 차세대 유전자서열검사기(next-generation sequencer)에 반응시키는 단계; 및
- [77] (A-v) 차세대 유전자서열검사기에서 핵산의 서열정보(reads)를 획득하는 단계를 포함하는 것을 특징으로 할 수 있다.
- [78]
- [79] 본 발명에 있어서, 상기 차세대 유전자서열검사기(next-generation sequencer)는 당업계에 공지된 임의의 시퀀싱 방법으로 사용될 수 있다. 선택 방법에 의해 분리된 핵산의 시퀀싱은 전형적으로는 차세대 시퀀싱(NGS)을 사용하여 수행된다. 차세대 시퀀싱은 개개의 핵산 분자 또는 고도로 유사한 방식으로 개개의 핵산 분자에 대해 클론으로 확장된 프록시 중 하나의 뉴클레오타이드 서열을 결정하는 임의의 시퀀싱 방법을 포함한다(예를 들어, 105개 이상의 분자가 동시에 시퀀싱된다). 일 실시형태에서, 라이브러리 내 핵산 중의 상대적 존재비는 시퀀싱 실험에 의해 만들어진 데이터에서 그것의 동족 서열의 상대적 발생 수를 계측함으로써 추정될 수 있다. 차세대 시퀀싱 방법은 당업계에 공지되어 있고, 예를 들어 본 명세서에 참조로서 포함된 문헌(Metzker, M. (2010) Nature Biotechnology Reviews 11:31-46)에 기재된다.
- [80] 일 실시형태에서, 차세대 시퀀싱은 개개의 핵산 분자의 뉴클레오타이드 서열을 결정하기 위해 한다(예를 들어, 헬리코스 바이오사이언스(Helicos BioSciences)의 헬리스코프 유전자 시퀀싱 시스템(HeliScope Gene Sequencing system) 및 퍼시픽바이오사이언스의 팍바이오 알에스 시스템(PacBio RS system)). 다른 실시형태에서, 시퀀싱, 예를 들어, 더 적지만 더 긴 리드를 만들어내는 다른 시퀀싱 방법보다 시퀀싱 단위 당 서열의 더 많은 염기를 만들어내는 대량병렬의 짧은-리드 시퀀싱(예를 들어, 캘리포니아주 샌디에고에 소재한 일루미나 인코포레이티드(Illumina Inc.) 솔렉사 시퀀서(Solexa sequencer)) 방법은 개개의 핵산 분자에 대해 클론으로 확장된 프록시의 뉴클레오타이드 서열을 결정한다(예를 들어, 캘리포니아주 샌디에고에 소재한 일루미나 인코포레이티드(Illumina Inc.) 솔렉사 시퀀서(Solexa sequencer); 454 라이프 사이언스(Life Sciences)(코네티컷주 브랜포드에 소재) 및 아이온 토렌트(Ion Torrent)). 차세대 시퀀싱을 위한 다른 방법 또는 기계는, 이하에 제한되는 것은 아니지만, 454 라이프 사이언스(Life Sciences)(코네티컷주 브랜포드에 소재), 어플라이드 바이오시스템스(캘리포니아주 포스터 시티에 소재; SOLiD 시퀀서), 헬리코스 바이오사이언스 코포레이션(매사추세츠주 캄브릿지에 소재) 및 에멀전 및 마이크로 유동 시퀀싱 기법 나노 점적(예를 들어, 지누바이오(GnuBio))

점적)에 의해 제공된다.

- [81] 차세대 시퀀싱을 위한 플랫폼은, 이하에 제한되는 것은 아니지만, 로슈(Roche)/454의 게놈 시퀀서(Genome Sequencer: GS) FLX 시스템, 일루미나(Illumina)/솔렉사(Solexa) 게놈 분석기(Genome Analyzer: GA), 라이프(Life)/APG의 서포트 올리고(Support Oligonucleotide Ligation Detection: SOLiD) 시스템, 폴로네이터(Polonator)의 G.007 시스템, 헬리코스 바이오사이언스의 헬리스코프 유전자 시퀀싱 시스템(Helicos BioSciences' HeliScope Gene Sequencing system) 및 퍼시픽 바이오사이언스(Pacific Biosciences)의 팍바이오알에스(PacBio RS) 시스템, MGI 사의 DNBseq를 포함한다.
- [82] NGS 테크놀로지는, 예를 들어 주형 제조, 시퀀싱 및 이미징 및 데이터 분석 단계 중 하나 이상을 포함할 수 있다.
- [83] 주형 제조 단계. 주형 제조를 위한 방법은 핵산(예를 들어, 게놈 DNA 또는 cDNA)을 작은 크기로 무작위로 파괴하는 단계 및 시퀀싱 주형(예를 들어, 단편 주형 또는 메이트-쌍 주형)을 만드는 단계와 같은 단계들을 포함할 수 있다. 공간적으로 분리된 주형은 고체 표면 또는 지지체에 부착되거나 또는 고정될 수 있는데, 이는 대량의 시퀀싱 반응이 동시에 수행되도록 한다. NGS 반응을 위해 사용될 수 있는 주형의 유형은, 예를 들어 단일 DNA 분자로부터 유래된 클론이 증폭된 주형 및 단일 DNA 분자 주형을 포함한다.
- [84] 클론이 증폭된 주형의 제조방법은, 예를 들어 에멀전 PCR(emulsion PCR: emPCR) 및 고체상 증폭을 포함한다.
- [85] EmPCR은 NGS를 위한 주형을 제조하기 위해 사용될 수 있다. 전형적으로, 핵산 단편의 라이브러리가 만들어지며, 보편적 프라이밍 부위를 함유하는 어댑터는 단편의 말단에 결합된다. 그 다음에 단편은 단일 가닥으로 변성되고, 비드에 의해 포획된다. 각 비드는 단일 핵산 분자를 포획한다. 증폭 및 emPCR 비드의 풍부화 후, 다량의 주형이 부착될 수 있고, 표준 현미경 슬라이드(예를 들어, 폴로네이터(Polonator)) 상에서 폴리아크릴아마이드 겔에 고정되며, 아미노-코팅된 유리 표면(예를 들어, Life/APG; 폴로네이터(Polonator))에 화학적으로 가교되거나, 또는 개개의 피코타이터플레이트(PicoTiterPlate: PTP) 웰(예를 들어, 로슈(Roche)/454) 상에 증착되는데, 이때 NGS 반응이 수행될 수 있다.
- [86] 고체상 증폭이 또한 사용되어 NGS를 위한 주형을 생성할 수 있다. 전형적으로, 전방 및 후방 프라이머는 고체지지체에 공유적으로 부착된다. 증폭된 단편의 표면 밀도는 지지체 상에서 프라이머 대 주형의 비로써 정의된다. 고체상 증폭은 수백만개의 공간적으로 분리된 주형 클러스터(예를 들어, 일루미나/솔렉사(Illumina/Solexa))를 생성할 수 있다. 주형 클러스터의 말단은 NGS 반응을 위한 보편적 프라이머에 혼성화될 수 있다.
- [87] 클론으로 증폭된 주형의 제조를 위한 다른 방법은, 예를 들어 다중 치환

- 증폭(Multiple Displacement Amplification: MDA)(Lasken R. S. Curr Opin Microbiol. 2007; 10(5):510-6)을 포함한다. MDA는 비-PCR 기반 DNA 증폭 기법이다. 반응은 주형에 대해 무작위 헥사머 프라이머를 어닐링하는 단계 및 일정한 온도에서 고충실도 효소, 전형적으로  $\Phi$ 29 중합효소에 의해 DNA를 합성하는 단계를 수반한다. MDA는 더 낮은 오류 빈도로 거대한 크기의 생성물을 만들 수 있다.
- [88] PCR과 같은 주형 증폭 방법은 표적에 NGS 플랫폼을 결합시킬 수 있거나 또는 게놈의 특이적 영역을 풍부화할 수 있다(예를 들어, 엑손). 대표적인 주형 풍부화 방법은, 예를 들어 마이크로점적 PCR 기법(Tewhey R. et al., Nature Biotech. 2009, 27:1025-1031), 맞춤-설계된 올리고뉴클레오타이드 마이크로어레이(예를 들어, 로슈(Roche)/nimbleGen(NimbleGen) 올리고뉴클레오타이드 마이크로어레이) 및 용액-기반 혼성화 방법(예를 들어, 분자역위 프로브(molecular inversion probe: MIP))(Porreca G. J. et al., Nature Methods, 2007, 4:931-936; Krishnakumar S. et al., Proc. Natl. Acad. Sci. USA, 2008, 105:9296-9310; Turner E. H. et al., Nature Methods, 2009, 6:315-316) 및 바이오틴화된 RNA 포획 서열(Gnirke A. et al., Nat. Biotechnol. 2009;27(2):182-9)을 포함한다.
- [89] 단일-분자 주형은 NGS 반응을 위해 사용될 수 있는 주형의 다른 유형이다. 공간적으로 분리된 단일 분자 주형은 다양한 방법에 의해 고체 지지체 상에 고정될 수 있다. 한 접근에서, 개개의 프라이머 분자는 고체 지지체에 공유적으로 부착된다. 어댑터는 주형에 첨가되고, 주형은 그 다음에 고정된 프라이머에 혼성화된다. 다른 접근에서, 단일-분자 주형은 고정된 프라이머로부터 단일-가닥의 단일-분자 주형을 프라이밍하고 연장시킴으로써 고체 지지체에 공유적으로 부착된다. 그 다음에 보편적 프라이머는 주형에 혼성화된다. 또 다른 접근에서, 단일 폴리머라제 분자는 프라이밍된 주형이 결합된 고체 지지체에 부착된다.
- [90] 시퀀싱 및 이미징. NGS를 위한 대표적인 시퀀싱 및 이미징 방법은, 이하에 제한되는 것은 아니지만, 사이클릭 가역적 종결(cyclic reversible termination: CRT), 절찰에 의한 시퀀싱(sequencing by ligation: SBL), 단일-분자 첨가(파이로시퀀싱(pyrosequencing)) 및 실시간 시퀀싱을 포함한다.
- [91] CRT는 뉴클레오타이드 포함, 형광 이미징 및 절단 단계를 최소로 포함하는 사이클릭 방법에서 가역 종결자를 사용한다. 전형적으로, DNA 폴리머라제는 프라이머에 주형 염기의 상보적 뉴클레오타이드에 대해 상보적인 단일의 형광으로 변형된 뉴클레오타이드를 포함시킨다. DNA 합성은 단일 뉴클레오타이드의 첨가 후 종결되고, 미포함된 뉴클레오타이드는 세척된다. 포함된 표지 뉴클레오타이드의 동일성을 결정하기 위해 이미징이 수행된다. 그 다음에, 절단 단계에서, 종결/억제기 및 형광 염료는 제거된다. CRT 방법을 사용하는 대표적인 NGS 플랫폼은, 이하에 제한되는 것은 아니지만, 전체 내부 반사 형광(total internal reflection fluorescence: TIRF)에 의해 검출된 4-색 CRT 방법과 결합된 클론으로 증폭된 주형 방법을 사용하는

일루미나(Illumina)/솔렉사(Solexa) 게놈 분석기(GA); 및 TIRF에 의해 검출된 1-색 CRT 방법과 결합된 단일-분자 주형 방법을 사용하는 헬리코스 바이오사이언스(Helicos BioSciences)/헬리스코프(HeliScope)를 포함한다.

- [92] SBL은 시퀀싱을 위해 DNA 리가제 및 1-염기-암호화된 프로브 또는 2-염기-암호화된 프로브 중 하나를 사용한다.
- [93] 전형적으로, 형광 표지된 프로브는 프라이밍된 주형에 인접한 상보적 서열에 혼성화된다. DNA 리가제는 프라이머에 염료-표지된 프로브를 결합시키기 위해 사용된다. 비-결찰 프로브가 세척된 후 결찰된 프로브의 동일성을 결정하기 위하여 형광 이미징이 수행된다. 형광 염료는 후속의 결찰 주기를 위해 5'-PO4 기를 재생하는 절단가능한 프로브를 사용하여 제거될 수 있다. 대안적으로, 새로운 프라이머는 오래된 프라이머가 제거된 후 주형에 혼성화될 수 있다. 대표적인 SBL 플랫폼은, 이하에 제한되는 것은 아니지만, 라이프(Life)/APG/SOLiD(지지체 올리고뉴클레오타이드 결찰 검출)를 포함하는데, 이는 2-염기-암호화된 프로브를 사용한다.
- [94] 파이로시퀀싱 방법은 다른 화학발광 효소로 DNA 폴리머라제의 활성을 검출하는 단계를 기반으로 한다. 전형적으로, 해당 방법은 한 번에 하나의 염기쌍을 따라 상보적 가닥을 합성하고, 각 단계에서 실제로 첨가된 염기를 검출함으로써 DNA의 단일 가닥을 시퀀싱시킨다. 주형 DNA는 고정적이며, A, C, G 및 T 뉴클레오타이드의 용액은 순차적으로 첨가되고, 반응으로부터 제거된다. 빛은 단지 뉴클레오타이드 용액이 주형의 짝지어지지 않은 염기를 보충할 때에만 생성된다. 화학발광 신호를 생성하는 용액의 서열은 주형의 서열을 결정하게 한다. 대표적인 파이로시퀀싱 플랫폼은, 이하에 제한되는 것은 아니지만, PTP 웰에 증착된 백만 내지 2백만개의 비드에 의한 emPCR에 의해 제조된 DNA 주형을 사용하는 로슈(Roche)/454를 포함한다.
- [95] 실시간 시퀀싱은 DNA 합성 동안 염료-표지된 뉴클레오타이드의 연속적 포함을 이미징하는 단계를 수반한다. 대표적인 실시간 시퀀싱 플랫폼은, 이하에 제한되는 것은 아니지만, 포스페이트 연결된 뉴클레오타이드가 성장되는 프라이머 가닥에 포함될 때 서열 정보를 얻기 위한 개개의 0-모드 웨이브가이드(zero-mode waveguide, ZMW) 검출기의 표면에 부착된 DNA 폴리머라제 분자를 사용하는 퍼시픽 바이오사이언스 플랫폼(Pacific Biosciences); 형광 공명 에너지 전달(fluorescence resonance energy transfer, FRET)에 의한 뉴클레오타이드 포함 후 향상된 신호를 만들기 위해 부착된 형광 염료와 함께 유전자 조작된 DNA 폴리머라제를 사용하는 라이프(Life)/비시젠(VisiGen) 플랫폼; 및 시퀀싱 반응에서 염료-쿼터 뉴클레오타이드를 사용하는 LI-COR 바이오사이언스(Biosciences) 플랫폼을 포함한다.
- [96] NGS의 다른 시퀀싱 방법은, 이하에 제한되는 것은 아니지만, 나노포어 시퀀싱, 혼성화에 의한 시퀀싱, 나노-트랜지스터 어레이 기반 시퀀싱, 폴로니(polony)

시퀀싱, 주사형전자 터널링 현미경(scanning tunneling microscopy, STM) 기반 시퀀싱 및 나노와이어-분자 센서 기반 시퀀싱을 포함한다.

- [97] 나노포어 시퀀싱은 단일-핵산 폴리머에서 분석될 수 있는 고도로 밀폐된 공간을 제공하는 나노-규모 포어를 통해서 용액 중의 핵산 분자의 전기영동을 수반한다. 나노포어 시퀀싱의 대표적인 방법은, 예를 들어 문헌[Branton D. et al., Nat Biotechnol. 2008; 26(10):1146-53]에 기재된다.
- [98] 혼성화에 의한 시퀀싱은 DNA 마이크로어레이를 사용하는 비-효소적 방법이다. 전형적으로, DNA의 단일 풀은 형광으로 표지되며, 공지된 서열을 함유하는 어레이에 혼성화된다. 어레이 상의 주어진 스팟으로부터 혼성화 신호는 DNA 서열을 확인할 수 있다. DNA 이중-가닥에서 DNA 중 한 가닥의 그것의 상보적 가닥에 결합은 혼성체 영역이 짧거나 또는 구체된 미스매치 검출 단백질이 존재할 때, 단일-염기 미스매치에 대해서 조차도 민감하다. 혼성화에 의한 시퀀싱의 대표적인 방법은, 예를 들어 문헌(Hanna G.J. et al., J. Clin. Microbiol. 2000; 38(7): 2715-21; 및 Edwards J.R. et al., Mut. Res. 2005; 573(1-2): 3-12)에 기재된다.
- [99] 플로니 시퀀싱은 플로니 증폭 및 다중 단일-염기-연장(FISSEQ)을 통해 시퀀싱에 따르는 것을 기반으로 한다. 플로니 증폭은 폴리아크릴아마이드 필름 상에서 인시츄로 DNA를 증폭시키는 방법이다. 대표적인 플로니 시퀀싱 방법은, 예를 들어 미국특허 출원 공개 제2007/0087362호에 기재된다.
- [100] 탄소나노튜브 전계 효과 트랜지스터(Carbon NanoTube Field Effect Transistor: CNTFET)와 같은 나노-트랜지스터 어레이 기반 장치가 또한 NGS를 위해 사용될 수 있다. 예를 들어, DNA 분자는 신장되고, 마이크로-제작된 전극에 의해 나노튜브에 걸쳐 구동된다. DNA 분자는 탄소 나노튜브 표면과 순차적으로 접촉하게 되고, DNA 분자와 나노튜브 사이의 전하 전달에 기인하여 각 염기로부터의 전류 흐름의 차이가 만들어진다. DNA는 이들 차이를 기록함으로써 시퀀싱된다. 대표적인 나노-트랜지스터 어레이 기반 시퀀싱 방법은, 예를 들어 미국특허 공개 제2006/0246497호에 기재된다.
- [101] 주사형전자 터널링 현미경(STM)은 또한 NGS를 위해 사용될 수 있다. STM은 표본의 래스터 주사(raster scan)를 수행하는 피에조-전자-제어 프로브를 사용하여 그것 표면의 이미지를 형성한다. STM은, 예를 들어 작동기-구동 가요성 갭과 주사형전자 터널링 현미경을 통합시킴으로써 일관된 전자 터널링 이미징 및 분광학을 만드는 단일 DNA 분자의 물리적 특성을 이미징하기 위해 사용될 수 있다. STM을 사용하는 대표적인 시퀀싱 방법은, 예를 들어 미국특허출원 공개 제2007/0194225호에 기재된다.
- [102] 나노와이어-분자 센서로 구성된 분자-분석 장치가 또한 NGS를 위해 사용될 수 있다. 이러한 장치는 DNA와 같은 나노와이어 및 핵산 분자에 배치된 질소성 물질의 상호작용을 검출할 수 있다. 분자 가이드는 상호작용 및 후속하는 검출을 허용하기 위해 분자 센서 근처의 분자를 가이드하기 위해 배치된다.

나노와이어-분자 센서를 사용하는 대표적인 시퀀싱 방법은 예를 들어 미국특허출원 공개 제2006/0275779호에 기재된다.

- [103] 이중 말단의 시퀀싱 방법이 NGS를 위해 사용될 수 있다. 이중 말단 시퀀싱은 DNA의 센스와 안티센스 가닥 둘 다를 시퀀싱하기 위해 차단 및 미차단 프라이머를 사용한다. 전형적으로, 이들 방법은 핵산의 제1 가닥에 미차단 프라이머를 어닐링시키는 단계; 핵산의 제2 가닥에 제2의 차단 프라이머를 어닐링시키는 단계; 폴리머라제로 제1 가닥을 따라 핵산을 연장시키는 단계; 제1 시퀀싱 프라이머를 종결시키는 단계; 제2 프라이머를 차단해제(deblocking)하는 단계; 및 제2 가닥을 따라 핵산을 연장시키는 단계를 포함한다. 대표적인 이중 가닥 시퀀싱 방법은, 예를 들어 미국특허 제7,244,567호에 기재된다.
- [104] 데이터 분석 단계. NGS 리드가 만들어진 후, 그것들은 공지된 기준 서열에 대해 정렬되거나 데노보 조립된다.
- [105] 예를 들어, 샘플(예를 들어, 종양 샘플)에서 단일-뉴클레오타이드 다형성 및 구조적 변이체와 같은 유전적 변형을 확인하는 것은 기준 서열(예를 들어, 야생형 서열)에 대해 NGS 리드를 정렬함으로써 수행될 수 있다. NGS에 대한 서열 정렬방법은, 예를 들어 문헌(Trapnell C. and Salzberg S.L. Nature Biotech., 2009, 27:455-457)에 기재된다.
- [106] 드노보 조립체의 예는, 예를 들어 문헌(Warren R. et al., Bioinformatics, 2007, 23:500-501; Butler J. et al., Genome Res., 2008, 18:810-820; 및 Zerbino D.R. 및 Birney E., Genome Res., 2008, 18:821-829)에 기재된다.
- [107] 서열 정렬 또는 어셈블리는 하나 이상의 NGS 플랫폼으로부터의 리드 데이터를 사용하여, 예를 들어 로슈(Roche)/454 및 일루미나(Illumina)/솔렉사(Solexa) 리드 데이터를 혼합하여 수행될 수 있다. 본 발명에 있어서, 상기 정렬단계는 이에 제한되지는 않으나, BWA 알고리즘 및 hg19 서열을 이용하여 수행되는 것일 수 있다.
- [108]
- [109] 본 발명에 있어서, 상기 (B) 단계의 핵산 단편의 위치를 확인 하는 단계는 바람직하게는 서열 정렬(Sequence alignment)를 통해 수행되는 것을 특징으로 할 수 있으며, 상기 서열 정렬은 컴퓨터 알고리즘으로서 게놈에서 리드 서열(예를 들어, 차세대 시퀀싱으로부터의, 예를 들어 짧은-리드 서열)이 대부분 리드 서열과 기준 서열 사이의 유사성을 평가함으로써 유래될 가능성이 있는 경우로부터 동일성에 대해 사용되는 컴퓨터적 방법 또는 접근을 포함한다. 서열 정렬 문제에 다양한 알고리즘이 적용될 수 있다. 일부 알고리즘은 상대적으로 느리지만, 상대적으로 높은 특이성을 허용한다. 이들은, 예를 들어 역동적 프로그래밍-기반 알고리즘을 포함한다. 역동적 프로그래밍은 그것들이 더 간단한 단계로 나누어짐으로써 복잡한 문제를 해결하는 방법이다. 다른 접근은 상대적으로 더 효율적이지만, 전형적으로 철저하지 않다. 이는, 예를 들어 대량

데이터베이스 검색을 위해 설계된 휴리스틱(heuristic) 알고리즘 및 확률적(probabilistic) 방법을 포함한다.

- [110] 전형적으로, 정렬 과정에 두 단계가 있을 수 있다: 후보자 검사 및 서열 정렬. 후보자 검사는 가능한 정렬 위치의 더 짧은 열거에 대해 전체 게놈으로부터 서열 정렬을 위한 검색 공간을 감소시킨다. 용어가 시사하는 바와 같이 서열 정렬은 후보자 검사 단계에 제공된 서열을 갖는 서열을 정렬시키는 단계를 포함한다. 이는 광역 정렬(예를 들어, 니들만-분쉬(Needleman-Wunsch) 정렬) 또는 국소 정렬(예를 들어, 스미스-워터만 정렬)을 사용하여 수행될 수 있다.
- [111] 대부분의 속성 정렬 알고리즘은 색인 방법에 기반한 3가지 유형 중 하나를 특징으로 할 수 있다: 해쉬 테이블(예를 들어, BLAST, ELAND, SOAP), 접미사트리(예를 들어, Bowtie, BWA) 및 병합 정렬(예를 들어, 슬라이더(Slider))에 기반한 알고리즘. 짧은 리드 서열은 정렬을 위해 전형적으로 사용된다. 짧은-리드 서열에 대한 서열 정렬 알고리즘/프로그램의 예는, 이하에 제한되는 것은 아니지만, BFAST (Homer N. et al., PLoS One. 2009;4(11):e7767), BLASTN(월드 와이드 웹상의 blast.ncbi.nlm.nih.gov에서), BLAT(Kent W.J. Genome Res. 2002;12(4):656-64), 보타이(Bowtie)(Langmead B. et al., Genome Biol. 2009;10(3):R25), BWA(Li H. and Durbin R. Bioinformatics, 2009, 25:1754-60), BWA-SW(Li H. and Durbin R. Bioinformatics, 2010;26(5):589-95), 클라우드버스트(CloudBurst)(Schatz M.C. Bioinformatics. 2009;25(11):1363-9), 코로나 라이트(Corona Lite)(Applied Biosystems, Carlsbad, California, USA), CASHX(Fahlgren N. et al., RNA, 2009; 15, 992-1002), CUDA-EC (Shi H. et al., J Comput Biol. 2010;17(4):603-15), ELAND(월드 와이드 웹상의 bioit.dbi.udel.edu/howto/eland에서), GNUMAP(Clement N.L. et al., Bioinformatics. 2010;26(1):38-45), GMAP(Wu T.D. and Watanabe C.K. Bioinformatics. 2005;21(9):1859-75), GSNAP(Wu T.D. and Nacu S., Bioinformatics. 2010;26(7):873-81), 제니오스 어셈블러(Geneious Assembler)(뉴질랜드 오클랜드에 소재한 Biomatters Ltd.), LAST, MAQ(Li H. et al., Genome Res. 2008;18(11):1851-8), Mega-BLAST(월드 와이드 웹 상의 ncbi.nlm.nih.gov/blast/megablast.shtml에서), MOM(Eaves H.L. and Gao Y. Bioinformatics. 2009;25(7):969-70), MOSAIK(월드 와이드 웹 상의 bioinformatics.bc.edu/marthlab/Mosaik에서), 노보얼라인(Novoalign)(월드 와이드 웹 상의 novocraft.com/main/index.php에서), 팔맵퍼(PALMapper)(월드 와이드 웹 상의 fml.tuebingen.mpg.de/raetsch/suppl/palmapper에서), PASS(Campagna D. et al., Bioinformatics. 2009;25(7):967-8), PatMaN(Prufer K. et al., Bioinformatics. 2008; 24(13):1530-1), PerM(Chen Y. et al., Bioinformatics, 2009, 25 (19): 2514-2521), ProbeMatch(Kim Y.J. et al., Bioinformatics. 2009;25(11):1424-5), QPalma(de Bona F. et al., Bioinformatics, 2008, 24(16): i174), RazerS(Weese D. et al., Genome Research, 2009, 19:1646-1654), RMAP (Smith A.D. et al., Bioinformatics. 2009;25(21):2841-2),

SeqMap(Jiang H. et al. Bioinformatics. 2008;24:2395-2396.), Shrec(Salmela L., Bioinformatics. 2010;26(10):1284-90), SHRiMP(Rumble S.M. et al., PLoS Comput. Biol., 2009, 5(5):e1000386), SLIDER(Malhis N. et al., Bioinformatics, 2009, 25 (1): 6-13), 슬림 서치(SLIM Search)(Muller T. et al., Bioinformatics. 2001;17 Suppl 1:S182-9), SOAP(Li R. et al., Bioinformatics. 2008;24(5):713-4), SOAP2(Li R. et al., Bioinformatics. 2009;25(15):1966-7), SOCS(Ondov B.D. et al., Bioinformatics, 2008; 24(23):2776-7), SSAHA(Ning Z. et al., Genome Res. 2001;11(10):1725-9), SSAHA2(Ning Z. et al., Genome Res. 2001;11(10):1725-9), 스탬피(Stampy)(Lunter G. and Goodson M. Genome Res. 2010, epub ahead of print), 타이판(Taipan)(월드 와이드 웹 상의 taipan.sourceforge.net에서), UGENE(월드 와이드 웹 상의 ugene.unipro.ru에서), XpressAlign(월드 와이드 웹 상의 bcgsc.ca/platform/bioinfo/software/XpressAlign에서), 및 ZOOM(캐나다 온타리오주 워터루에 소재한 바이오인포매틱스 솔루션 인코포레이티드(Bioinformatics Solutions Inc.))을 포함한다.

- [112] 서열 정렬 알고리즘은, 예를 들어 시퀀싱 기법, 리드 길이, 리드 수, 입수가 가능한 컴퓨팅 자료 및 민감성/스코어링 필요조건을 포함하는 다수의 인자에 기반하여 선택될 수 있다. 상이한 서열 정렬 알고리즘은 상이한 속도 수준, 정렬 민감성 및 정렬 특이성을 달성할 수 있다. 정렬 특이성은 예측된 정렬과 비교하여 정확하게 정렬된 전형적으로 서브미션에서 발견되는 바와 같이 정렬된 표적 서열 잔기의 백분율을 지칭한다. 정렬 민감성은 또한 서브미션에서 정확하게 정렬된 보통 예측된 정렬에서 발견되는 바와 같이 정렬된 표적 서열 잔기의 백분율을 지칭한다.
- [113] 정렬 알고리즘, 예컨대 ELAND 또는 SOAP는 속도가 고려되는 제1 인자일 때 기준 게놈에 대해 짧은 리드(예를 들어, 일루미나(Illumina)/솔렉사(Solexa) 시퀀서제)를 정렬하는 목적으로 사용될 수 있다. BLAST 또는 Mega-BLAST와 같은 정렬 알고리즘은 특이성이 가장 중요한 인자일 때, 이들 방법이 상대적으로 더 느리지만, 짧은 판독(예를 들어, 로슈(Roche) FLX제)를 사용하여 유사성 조사의 목적을 위해 사용될 수 있다. MAQ 또는 노보얼라인(Novoalign)와 같은 정렬 알고리즘은 품질 스코어를 고려하며, 따라서 정확성이 본질을 가질 때 단일- 또는 짝지어진-말단 데이터에 대해 사용될 수 있다(예를 들어, 고속-대량 SNP 검색에서). 보타이(Bowtie) 또는 BWA와 같은 정렬 알고리즘은 버로우즈-휠러 변환(Burrows-Wheeler Transform: BWT)을 사용하며, 따라서 상대적으로 작은 메모리 풋프린트(memory footprint)를 필요로 한다. BFAST, PerM, SHRiMP, SOCS 또는 ZOOM과 같은 정렬 알고리즘은 색공간 리드를 맵핑하며, 따라서 ABI의 SOLiD 플랫폼과 함께 사용될 수 있다. 일부 적용에서, 2 이상의 정렬 알고리즘으로부터의 결과가 조합될 수 있다.
- [114]
- [115] 본 발명에 있어서, 상기 B) 단계의 서열정보(reads)의 길이는, 5 내지 5000

bp이고, 사용하는 서열정보의 수는 5천 내지 500만개가 될 수 있으나, 이에 한정되는 것은 아니다.

[116]

[117] 본 발명에 있어서, 상기 (C) 단계의 서열정보를 그룹화하는 단계는, 서열정보(reads)의 어댑터 서열을 바탕으로 수행할 수 있다. 정방향으로 정렬된 핵산 단편과 역방향으로 정렬된 핵산 단편으로 별도로 구분하여서 선별된 서열정보에 대해서 FD 값을 계산하거나, 전체 그룹에 대하여 FD 값을 계산할 수 있다.

[118]

[119] 본 발명에서, 상기 (C) 단계를 수행하기에 앞서 정렬된 핵산 단편의 정렬 일치도 점수(mapping quality score)를 만족하는 핵산 단편을 따로 분류하는 단계를 추가로 포함하는 것을 특징으로 할 수 있다.

[120] 본 발명에서 상기 정렬 일치도 점수(mapping quality score)는 원하는 기준에 따라 달라질 수 있으나, 바람직하게는 15-70점, 더욱 바람직하게는 50~70점 일 수 있고, 가장 바람직하게는 60점일 수 있다.

[121]

[122] 본 발명에 있어서, 상기 (D) 단계의 FD 값은 수득한 n개의 핵산 단편에 대하여, i 번째 핵산 단편과 i+1 내지 n 번째 핵산 단편에서 선택되는 어느 하나 이상의 핵산 단편의 기준값 사이의 거리로 정의되는 것을 특징으로 할 수 있다.

[123] 본 발명에서, 상기 FD 값은 수득한 n개의 핵산 단편에 대하여, 제1핵산 단편과 제2내지 제n개 핵산 단편으로 구성된 군에서 선택되는 어느 하나 이상의 핵산 단편의 기준값과의 거리를 계산하여 이들의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값과, 가중치가 포함된 계산 결과 및 이에 한정되지 않는 통계치를 FD 값으로 사용할 수 있으나 이에 한정되는 것은 아니다.

[124] 본 발명에서 “하나 이상의 값 및/또는 하나 이상의 이들의 역수값”이라는 기재는 앞서 기재된 수치값들 중에서 하나 또는 2 이상이 조합되어 사용될 수 있다는 의미로 해석된다.

[125] 본 발명에서, 상기 “핵산 단편의 기준값”은 핵산 단편의 중앙값으로부터 임의의 값을 더하거나 빼 값인 것인 것을 특징으로 할 수 있다.

[126]

[127] 상기 FD 값은 수득한 n개의 핵산 단편에 대하여, 다음과 같이 정의 할 수 있다.

[128]  $FD = \text{Dist}(R_i \sim R_j) \ (1 \leq i < j \leq n),$

[129] 여기서 Dist 함수는 선별된  $R_i$ 와  $R_j$  두 핵산 단편 사이에 포함되는 모든 핵산 단편의 정렬 위치값 차이의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값과,

가중치가 포함된 계산 결과 및 이에 한정되지 않는 통계치를 계산한다.

- [130] 즉, 본 발명에서 FD 값(Fragment Distance Value)는 정렬된 핵산 단편 사이의 거리를 의미한다. 여기서 거리 계산을 위한 핵산 단편의 선별 경우의 수는 다음과 같이 정의 할 수 있다. 총 N개의 핵산 단편이 존재할 경우  $\sum_{k=1}^{n-1} k$  개의

핵산 단편 간 거리 조합이 가능하다. 즉, i가 1일 경우, i+1은 2가 되어 2 내지 n 번째 핵산 단편에서 선택되는 어느 하나 이상의 핵산 단편과의 거리를 정의 할 수 있다.

- [131] 본 발명에 있어서, 상기 FD 값은 상기 i 번째 핵산 단편 내부의 특정 위치와 i+1 내지 n 번째 중 어느 하나 이상의 핵산 단편 내부의 특정 위치 사이의 거리를 계산하는 것을 특징으로 할 수 있다.
- [132] 예를 들어 어떤 핵산 단편의 길이가 50bp 이며, 염색체 1번의 4,183 위치에 정렬 되었다고 하면, 이 핵산 단편의 거리 계산에 사용할 수 있는 유전적 위치값은 염색체 1번의 4,183 ~ 4,232 이다.
- [133] 상기 핵산 단편과 인접한 50bp 길이의 핵산 단편이 염색체 1번의 4,232번째 위치에 정렬되면, 이 핵산 단편의 거리 계산에 사용할 수 있는 유전적 위치값은 염색체 1번의 4,232 ~ 4,281이고, 두 핵산 단편 사이의 FD 값은 1에서 99가 될 수 있다.
- [134] 또 다른 인접한 50bp 길이의 핵산 단편이 염색체 1번의 4123번째 위치에 정렬되면, 이 핵산 단편의 거리 계산에 사용할 수 있는 유전적 위치값은 염색체 1번의 4,123~4,172이며, 두 핵산 단편 사이의 FD 값은 61에서 159이며, 첫 번째 예시 핵산 단편과의 FD 값은 12에서 110으로, 상기 두 FD 값 범위의 한 값의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값, 및 가중치가 포함된 계산결과 및 이에 한정되지 않는 통계치를 FD 값으로 사용할 수 있으며, 바람직하게는 두 FD 값 범위의 한 값의 역수값인 것을 특징으로 할 수 있으나 이에 한정되는 것은 아니다
- [135]
- [136] 바람직하게는 본 발명에 있어서, 상기 FD 값은 핵산 단편의 중앙값으로부터 임의의 값을 더하거나 빼 값인 것을 특징으로 할 수 있다.
- [137] 본 발명에서, FD의 중앙값은 계산된 FD 값들을 크기의 순서대로 정렬했을 때 가장 중앙에 위치하는 값을 의미한다. 예를 들어 1, 2, 100 과 같이 세 개의 값이 있을 때, 2가 가장 중앙에 있기 때문에 2가 중앙값이 된다. 만약 짝수 개의 FD 값이 있을 경우 가운데 있는 두 값의 평균으로 중앙값을 결정 한다. 예를 들어 1, 10, 90, 200의 FD 값이 있을 경우 중앙값은 10과 90의 평균인 50이 된다.
- [138] 본 발명에 있어서, 상기 임의의 값은 핵산 단편의 위치를 나타낼 수 있으면

제한없이 이용가능하나, 바람직하게는 0 내지 5 kbp 또는 핵산 단편 길이의 0 내지 300%, 0 내지 3 kbp 또는 핵산 단편 길이의 0 내지 200%, 0 내지 1 kbp 또는 핵산 단편 길이의 0 내지 100%, 더욱 바람직하게는 0 내지 500 bp 또는 핵산 단편 길이의 0 내지 50% 일 수 있으나, 이에 한정되는 것은 아니다.

- [139] 본 발명에 있어서, 상기 FD값은 페어드 엔드 시퀀싱(paired-end sequencing)일 경우, 정방향 및 역방향 서열정보(reads)의 위치값을 기반으로 도출하는 것을 특징으로 할 수 있다.
- [140] 예를 들어, 50bp 길이의 페어드 엔드 리드 쌍에서, 정방향 리드는 염색체 1번의 4183번째 위치에 정렬되고, 역방향 리드는 4349번째 위치에 정렬되면, 이 핵산단편의 양 말단은 4183, 4349가 되고, 핵산 단편 거리에 사용할 수 있는 기준값은 4183~4349이다. 이 때 상기 핵산 단편과 인접한 다른 페어드 엔드 리드 쌍에서, 정방향 리드는 염색체 1번의 4349번째 위치에 정렬되고, 역방향 리드는 4515번째 정렬되면, 이 핵산 단편의 위치 값은 4349~4515 이다. 이 두 핵산 단편의 거리는 0~333이 될 수 있고, 가장 바람직하게 각 핵산 단편의 중앙값의 거리인 166이 될 수 있다.
- [141] 본 발명에 있어서, 상기 페어드 엔드 시퀀싱으로 서열정보를 획득할 경우, 서열정보(reads)의 정렬 점수가 기준값 미만인 핵산 단편의 경우, 계산과정에서 제외하는 단계를 추가로 포함하는 것을 특징으로 할 수 있다.
- [142]
- [143] 본 발명에 있어서, 상기 FD 값은 싱글 엔드 시퀀싱(single-end sequencing)일 경우, 정방향 또는 역방향 서열정보(read)의 위치값 한 종류를 기반으로 도출하는 것을 특징으로 할 수 있다.
- [144] 본 발명에 있어서, 상기 싱글 엔드 시퀀싱의 경우, 정방향으로 정렬된 서열정보를 기반으로 위치값을 도출할 경우에는 임의의 값을 더해주며, 역방향으로 정렬된 서열정보를 기반으로 위치값을 도출할 경우에는 임의의 값을 빼주는 것을 특징으로 할 수 있으며, 상기 임의의 값은 FD 값이 핵산 단편의 위치를 명확하게 나타내도록 하는 값이면 제한없이 이용가능하나, 바람직하게는 0 내지 5kbp 또는 핵산 단편 길이의 0 내지 300%, 0 내지 3kbp 또는 핵산 단편 길이의 0 내지 200%, 0 내지 1kbp 또는 핵산 단편 길이의 0 내지 100%, 더욱 바람직하게는 0 내지 500bp 또는 핵산 단편 길이의 0 내지 50% 일 수 있으나, 이에 한정되는 것은 아니다.
- [145]
- [146] 본 발명에서 분석하고자 하는 핵산은 시퀀싱 되어 리드(reads)라는 단위로 표현될 수 있다. 이 리드는 시퀀싱 방법에 따라 싱글 엔드 시퀀싱(single end sequencing read, SE) 및 페어드 엔드 시퀀싱(paired end sequencing read, PE)으로 나눌 수 있다. SE 방식의 리드는 핵산 분자의 5'과 3' 중 한 곳을 랜덤한 방향으로 일정한 길이만큼 시퀀싱 한 것을 의미하고, PE 방식의 리드는 5'과 3' 을 모두 일정 한 길이만큼 시퀀싱 하게 된다. 이러한 차이 때문에, SE 모드로 시퀀싱 할

경우 한 개의 핵산 단편으로부터 1개의 리드가 생기고, PE 모드에서는 1개의 핵산 단편으로부터 2개의 리드가 쌍으로 생성되는 것은 통상의 기술자에게 잘 알려진 사실이다.

- [147] 핵산 단편 사이의 정확한 거리를 계산 하기 위한 가장 이상적인 방식은 핵산 분자를 처음부터 끝까지 시퀀싱하고, 그 리드를 정렬하고, 정렬된 값의 중앙값(센터)을 이용하는 것이다. 그러나 기술적으로 위 방식은 시퀀싱 기술의 한계 및 비용적인 측면 때문에 제약이 있는 실정이다. 따라서 SE, PE와 같은 방식으로 시퀀싱을 하게 되는데, PE 방식의 경우 핵산 분자의 시작과 끝 위치를 알 수 있기 때문에 이 값들의 조합을 통해 핵산 단편의 정확한 위치(중앙값)를 파악 할 수 있으나, SE 방식의 경우 핵산 단편의 한쪽 끝 정보만을 이용할 수 있기 때문에 정확한 위치(중앙값) 계산에 한계가 있다.
- [148] 또한 정방향, 역방향의 양 방향으로 시퀀싱 된(정렬된), 모든 리드의 말단 정보를 이용해 핵산 분자의 거리 계산시, 시퀀싱 방향 이라는 요소 때문에 정확하지 않은 값이 계산 될 수 있다.
- [149] 따라서, 시퀀싱 방식의 기술적 이유로 정방향 리드의 5'말단은, 핵산 분자의 중심 위치 보다 작은 위치 값을 갖고, 역방향 리드의 3'말단은 큰 값을 갖게 된다. 이러한 특징을 이용해, 정방향 리드의 경우 임의의 값(Extended bp)을 더해주고, 역방향 리드는 빼주게 되면 핵산 분자의 중심 위치에 가까운 값을 추정 할 수 있다.
- [150] 즉, 임의의 값(Extended bp)은 사용하는 시료에 따라 달라질 수 있으며, 세포유리 핵산의 경우 그 핵산의 평균 길이가 166bp 정도로 알려져 있기 때문에 약 80bp 정도로 설정을 할 수 있다. 만일 단편화 장비(ex; sonication)를 통해 실험이 진행 된 경우 단편화 과정에서 설정한 타겟 길이의 절반 정도를 extended bp으로 설정 할 수 있다.
- [151]
- [152] 본 발명에 있어서, 상기 (E) 단계의 염색체의 이상을 판정하는 단계는
- [153] (E-i) 염색체 전체 영역 또는 특정 영역별로 FD 값의 대표값(RepFD)을 결정하는 단계;
- [154] (E-ii) 분석하고자 하는 염색체 전체 영역 또는 특정 영역 외의 샘플 내 특정 영역의 RepFD 값의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값을 계산하여, 정규화 요소(Normalized Factor)를 도출하는 단계;
- [155] (E-iii) 하기 식 1을 바탕으로 대표값 비율(RepFD ratio)를 계산하는 단계;
- [156] 식 1:  $\text{RepFD ratio} = \text{RepFD Target genomic region} / \text{Normalized Factor}$
- [157] (E-iv) 정상인 참조집단과 샘플의 RepFD ratio 값을 비교하여, FDI (Fragments Distance Index)를 계산하는 단계;
- [158] 를 포함하여 수행되는 것을 특징으로 할 수 있다.

[159]

[160] 본 발명에 있어서, 상기 (E-i) 단계의 대표값(RepFD)은 FD 값의 합, 차, 곱, 평균, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값인 것을 특징으로 할 수 있으며, 바람직하게는 FD 값들의 중앙값, 평균값 또는 이의 역수값인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[161]

본 발명에서 상기 염색체 전체 영역 또는 특정 유전 영역은 인간 핵산 서열의 집합이면 제한없이 이용가능하나, 바람직하게는 염색체 단위 또는 일부 염색체의 특정 영역일 수 있으며, 예를 들어, 수적 이상 여부 검출을 위한 특정 영역에는 정배수체로 생각되는 상염색체가 될 수 있고, 구조적 이상 여부 검출을 위한 특정 영역에는 고유성이 떨어지는 영역(centromere, telomere)을 제외한 모든 유전적 영역이 될 수 있으나, 이에 한정되는 것은 아니다.

[162]

[163] 상기 (E-ii) 단계의 분석하고자 하는 염색체 전체 영역 또는 특정 유전 영역 외의 샘플 내 특정 영역은

[164]

a) 무작위로 분석하고자 하는 염색체 전체 영역 또는 특정 유전 영역 외의 영역을 무작위로 선별하는 단계;

[165]

b) 상기 a) 단계에서 선별한 유전 영역의 RepFD 값의 대표값을 사전 정규화 요소(Pre Normalized Factor, PNF)로 결정하는 단계;

[166]

c) 하기 식 2를 바탕으로 대표값 비율(RepFD ratio)를 계산하는 단계:

[167]

식 2:  $\text{RepFD ratio} = \text{RepFD Target genomic region} / \text{PNF}$

[168]

d) 정상인 참조 집단의 RepFD ratio 값의 변동 계수(Coefficient of Variance: SD / Mean)를 계산하는 단계; 및

[169]

e) 상기 a) 내지 d) 단계를 반복 시행하여 수득한 변동 계수 중 가장 작은 값을 갖는 유전 영역을 염색체 전체 영역 또는 특정 유전 영역 외의 샘플내 특정 영역으로 결정하는 단계를 포함하는 방법으로 선별하는 것을 특징으로 할 수 있다.

[170]

본 발명에서 상기 e) 단계의 반복 시행은 100회 이상, 바람직하게는 1만에서 100만회 사이, 가장 바람직하게는 10만회인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[171]

[172]

본 발명에 있어서, 상기 (E-iv) 단계는 정상인 참조집단의 RepFD ratio 값을 샘플의 RepFD ratio 값과 비교하는 것을 특징으로 할 수 있다.

[173]

[174]

본 발명에 있어서, 상기 정상인 참조집단의 RepFD ratio 값과 샘플의 RepFD ratio를 비교하는 방식은 두 값이 통계적으로 유의미하게 차이가 나는 것을 확인할 수 있는 방법이면 제한없이 사용가능하나, 바람직하게는 평균 및 표준편차 기반의 Z-score나 중앙값 기반의 Log 비, 그 외 분류 알고리즘을 통해

산출된 우도비(Likelihood) 등이 선택되는 방법일 수 있으며, 가장 바람직하게는 평균 및 표준편차 기반의 Z점수 계산방식일 수 있으나, 이에 한정되는 것은 아니다.

- [175] 본 발명에 있어서, 상기 Fragments Distance Index 는 정상참조 집단과 분석하고자 하는 샘플의 Rep FD ratio 값의 비교를 통해 계산되는데, 비교를 하는 방법에는 Z점수와 같은 표준 점수 방식을 사용 할 수 있고, 임계치는 무한대의 양수 음수등의 정수 또는 범위가 가능하고, 바람직하는 3이 될 수 있으나, 이에 한정되는 것은 아니다.
- [176]
- [177] 본 발명은 다른 관점에서, 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부;
- [178] 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 및
- [179] 선별된 핵산 단편(fragments)에 대하여 정렬된 핵산 단편 사이의 거리를 측정하여, FD 값(Fragments Distance)을 계산하고, 계산한 FD 값을 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 FDI 값(Fragments Distance Index)를 계산하여, FDI 값이 기준값 또는 구간 미만 또는 초과 일 경우, 염색체 이상이 있는 것으로 판정하는 염색체 이상 판정부를 포함하는 염색체 이상 검출 장치에 관한 것이다.
- [180]
- [181] 본 발명은 또 다른 관점에서, 컴퓨터 판독 가능한 저장 매체로서, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,
- [182] (A) 생체시료에서 핵산을 추출하여 핵산 단편을 획득하여 서열정보를 취득하는 단계;
- [183] (B) 취득한 서열정보(reads)에 기반하여 핵산 단편을 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;
- [184] (C) 선별된 핵산 단편(fragments) 사이의 거리를 측정하여, FD 값(Fragments Distance)을 계산하는 단계; 및
- [185] (D) 상기 (C) 단계에서 계산한 FD 값을 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 FDI 값(Fragments Distance Index)를 계산하여, FDI 값이 기준 값 범위 미만 또는 초과일 경우, 염색체 이상이 있는 것으로 판정하는 단계;
- [186] 를 통하여, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체에 관한 것이다.
- [187]
- [188] 구체적으로는 본 발명에 따른 컴퓨터 판독 가능한 저장 매체는, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,
- [189] (A) 생체시료에서 핵산을 추출하여 핵산 단편을 획득하여 서열정보를 취득하는 단계;
- [190] (B) 취득한 서열정보(reads)에 기반하여 핵산 단편을 표준 염색체 서열

- 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;
- [191] (C) 상기 서열정보(reads)에 기반하여 정렬된 핵산 단편(fragments)을 전체 서열, 정방향 서열 및 역방향 서열로 그룹화 하는 단계;
- [192] (D) 상기 그룹화된 각각의 핵산 단편에 대하여 정렬된 핵산 단편 기준값 사이의 거리를 측정하여, 각 그룹별 FD 값(Fragments Distance)을 계산하는 단계; 및
- [193] (E) 상기 (D) 단계에서 계산한 각 그룹별 FD 값을 기반으로 염색체 전체 영역 또는 특정 영역 별로 각각의 FDI 값(Fragments Distance Index)을 계산하여, 각각의 FDI 값이 모두 기준값 범위 미만 혹은 초과일 경우, 염색체 이상이 있는 것으로 판정하는 단계;
- [194] 를 통하여, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 것을 특징으로 할 수 있지만 이에 한정되는 것은 아니다.
- [195]
- [196] 본 발명의 다른 실시예에서는 상기 정렬된 핵산 단편의 중앙값에서 분석대상 핵산 단편 평균 길이의 50%를 더하거나 빼서 리드의 양 말단의 위치값을 계산하여, 리드 사이의 거리를 계산할 수 있다는 것을 확인하였다(도 6).
- [197]
- [198] 따라서, 본 발명은 또 다른 관점에서,
- [199] (A) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;
- [200] (B) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;
- [201] (C) 상기 정렬된 서열정보(reads)에 대하여 정렬된 리드 사이의 거리를 측정하여, RD 값(Read Distance)을 계산하는 단계; 및
- [202] (D) 상기 (C) 단계에서 계산한 RD 값을 기반으로 염색체 전체 영역 또는 특정 영역 별로 RDI 값(Read Distance Index)을 계산하여, RDI 값이 기준 값 범위 미만 혹은 초과일 경우, 염색체 이상이 있는 것으로 판정하는 단계를 포함하는 염색체 이상 검출 방법에 관한 것이다.
- [203]
- [204] 본 발명에 있어서,
- [205] 상기 A) 단계는
- [206] (A-i) 혈액, 정액, 질 세포, 모발, 타액, 소변, 구강세포, 태반세포 또는 태아세포를 포함하는 양수, 조직세포 및 이의 혼합물에서 핵산을 수득하는 단계;
- [207] (A-ii) 채취된 핵산에서 솔팅-아웃 방법(salting-out method), 컬럼 크로마토그래피 방법(column chromatography method) 또는 비드 방법(beads method)을 사용하여 단백질, 지방, 및 기타 잔여물을 제거하고 정제된 핵산을 수득하는 단계;
- [208] (A-iii) 정제된 핵산 또는 효소적 절단, 분쇄, 수압 절단 방법(hydroshear method)으로 무작위 단편화(random fragmentation)된 핵산에 대하여, 싱글 엔드 시퀀싱(single-end sequencing) 또는 페어 엔드 시퀀싱(pair-end sequencing)

- 라이브러리(library)를 제작하는 단계;
- [209] (A-iv) 제작된 라이브러리를 차세대 유전자서열검사기(next-generation sequencer)에 반응시키는 단계; 및
- [210] (A-v) 차세대 유전자서열검사기에서 핵산의 서열정보(reads)를 획득하는 단계포함하는 방법으로 수행되는 것을 특징으로 할 수 있다.
- [211]
- [212] 본 발명에 있어서, 상기 (B) 단계의 서열정보(reads)의 길이는, 5 내지 5000 bp이고, 사용하는 서열정보의 수는 5천 내지 500만개가 될 수 있으나, 이에 한정되는 것은 아니다.
- [213]
- [214] 본 발명에 있어서, 상기 (C) 단계는 정렬된 리드를 정렬된 방향에 따라 그룹화하는 단계를 추가로 사용할 수 있는 것을 특징으로 할 수 있다.
- [215] 본 발명에서는 리드를 그룹화하는 단계를 추가로 사용할 수 있으며, 이때 그룹화 기준은, 정렬된 리드의 어댑터 서열을 바탕으로 수행할 수 있다. 정방향으로 정렬된 리드와 역방향으로 정렬된 리드로 별도로 구분하여서 선별된 서열정보에 대해서 RD 값을 계산할 수 있다.
- [216]
- [217] 본 발명에서, 상기 (C) 단계를 수행하기에 앞서 정렬된 리드의 정렬 일치도 점수(mapping quality score)를 만족하는 리드를 따로 분류하는 단계를 추가로 포함하는 것을 특징으로 할 수 있다.
- [218] 본 발명에서 상기 정렬 일치도 점수(mapping quality score)는 원하는 기준에 따라 달라질 수 있으나, 바람직하게는 15-70점, 더욱 바람직하게는 50~70점 일 수 있고, 가장 바람직하게는 60점일 수 있다.
- [219]
- [220] 본 발명에 있어서, 상기 (C) 단계의 RD 값은 수득한 n개의 리드에 대하여, i 번째 리드와 i+1 내지 n 번째 리드에서 선택되는 어느 하나 이상의 리드의 양 말단 값 중 하나의 값에 핵산 평균 길이의 50%를 더하거나 뺀 값 사이의 거리를 통해 산출하는 것을 특징으로 할 수 있다.
- [221] 본 발명에서, 상기 RD 값은 수득한 n개의 리드에 대하여, 제1리드와 제2 내지 제n개 리드로 구성된 군에서 선택되는 어느 하나 이상의 리드와의 거리를 계산하여 이들의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차, 변동 계수, 이들의 역수값 및 조합, 가중치가 포함된 계산 결과 및 이에 한정되지 않는 통계치를 RD 값으로 사용할 수 있으나 이에 한정되는 것은 아니다.
- [222] 본 발명에서, RD의 중앙값은 계산된 RD 값들을 크기의 순서대로 정렬했을 때 가장 중앙에 위치하는 값을 의미한다. 예를 들어 1, 2, 100 과 같이 홀수 개의 값이 있을 때, 2가 가장 중앙에 있기 때문에 2가 중앙값이 된다. 만약 짝수 개의 RD 값이 있을 경우 가운데 있는 두 값의 평균으로 중앙값을 결정 한다. 예를 들어 1,

10, 90, 200의 RD 값이 있을 경우 중앙값은 10과 90의 평균인 50이 된다.

[223]

[224] 본 발명에 있어서, 상기 RD 값은 상기 i 번째 리드 내부의 5' 또는 3' 말단과 i+1 내지 n 번째 중 어느 하나 이상의 리드의 5' 또는 3' 말단 사이의 거리를 계산하는 것을 특징으로 할 수 있다.

[225] 예를 들어, 50bp 길이의 페어드 엔드 리드 쌍에서, 정방향 리드는 염색체 1번의 4183번째 위치에 정렬되고, 역방향 리드는 4349번째 위치에 정렬되면, 이 핵산단편의 양 말단은 4183, 4349가 되고, 핵산 단편 거리에 사용할 수 있는 기준값은 4183~4349이다. 이 때 상기 핵산 단편과 인접한 다른 페어드 엔드 리드 쌍에서, 정방향 리드는 염색체 1번의 4349번째 위치에 정렬되고, 역방향 리드는 4515번째 정렬되면, 이 핵산 단편의 위치 값은 4349~4515 이다. 이 두 핵산 단편의 거리는 0~333이 될 수 있고, 가장 바람직하게 각 핵산 단편의 중앙값의 거리인 166이 될 수 있다. 상기 예시에서 핵산 단편의 평균 길이가 166 일 경우, 핵산 단편 평균 길이의 50% 값을 중앙값(4266)에서 빼는 경우, 첫번째 핵산단편의 위치값은 4183 이 되고, 두번째 핵산 단편의 위치값은 4349 이며, 이때 리드 사이의 거리는 166이 된다(4349 - 4183). 반면 50% 값을 중앙값에서 더하는 경우, 첫번째 핵산단편의 위치값은 4349, 두번째 핵산단편의 위치값은 4515 이며 이때 리드 사이의 거리는 166이 된다(4515 - 4349).

[226]

[227] 본 발명에 있어서, 상기 (D) 단계의 염색체의 이상을 판정하는 단계는

[228] (D-i) 각 염색체 전체 영역 또는 특정 유전 영역별로 RD 값의 대표값(RepRD)을 결정하는 단계;

[229] (D-ii) 분석하고자 하는 염색체 전체 영역 또는 특정 유전 영역 외의 샘플 내 영역의 RepRD 값의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값을 계산하여, 정규화 요소(Normalized Factor)를 도출하는 단계;

[230] (D-iii) 하기 식 10을 바탕으로 대표값 비율(RepRD ratio)을 계산하는 단계;

[231] 식 10:  $\text{RepRD ratio} = \text{RepRD Target genomic region} / \text{Normalized Factor}$

[232] (D-iv) 정상인 참조집단과 샘플의 RepRD ratio 값을 비교하여, RDI (Read Distance Index)를 계산하는 단계;

[233] 를 포함하여 수행되는 것을 특징으로 할 수 있다.

[234]

[235] 본 발명에 있어서, 상기 (D-i) 단계의 대표값(RepRD)은 RD 값의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값, 및 이에 한정하지 않는 통계치로 구성된 군에서 선택되는 하나 이상인 것을 특징으로 할 수 있으며, 바람직하게는 RD 값들의

중앙값, 평균값 또는 이의 역수값인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[236]

[237] 본 발명에서, RepRD 값의 중앙값은 계산된 RepRD 값들을 크기의 순서대로 정렬했을 때 가장 중앙에 위치하는 값을 의미한다. 예를 들어 1, 2, 100 과 같이 홀수 개의 값이 있을 때, 2가 가장 중앙에 있기 때문에 2가 중앙값이 된다. 만약 짝수 개의 RepRD 값이 있을 경우 가운데 있는 두 값의 평균으로 중앙값을 결정한다. 예를 들어 1, 10, 90, 200의 RepRD 값이 있을 경우 중앙값은 10과 90의 평균인 50이 된다.

[238]

[239] 본 발명에서 상기 염색체 전체 영역 또는 특정 유전 영역(specific genomic region)은 인간 핵산 서열의 집합이면 제한없이 이용가능하나, 바람직하게는 염색체 단위 또는 일부 염색체의 특정 영역일 수 있으며, 예를 들어, 수직 이상 여부 검출을 위한 특정 영역에는 정배수체로 생각되는 상염색체가 될 수 있고, 구조적 이상 여부 검출을 위한 특정 영역에는 고유성이 떨어지는 영역(centromere, telomere)을 제외한 모든 유전적 영역이 될 수 있으나, 이에 한정되는 것은 아니다.

[240]

[241] 상기 (D-ii) 단계의 분석하고자 하는 염색체 전체 영역 또는 특정 유전 영역 외의 샘플 내 특정 영역은

[242] a) 무작위로 분석하고자 하는 염색체 전체 영역 또는 특정 유전 영역 외의 영역을 선별하는 단계;

[243] b) 상기 a)단계에서 선별한 유전 영역의 RepRD 값의 대표값을 사전 정규화 요소(Pre Normalized Factor, PNF)로 결정하는 단계;

[244] c) 하기 식 11을 바탕으로 대표값 비율(RepRD ratio)를 계산하는 단계:

[245] 식 11:  $\text{RepRD ratio} = \text{RepRD Target genomic region} / \text{PNF}$

[246] d) 정상인 참조 집단의 RepRD ratio 값의 변동 계수(Coefficient of Variance: SD / Mean)를 계산하는 단계; 및

[247] e) 상기 a) 내지 d) 단계를 반복 시행하여 수득한 변동 계수 중 가장 작은 값을 갖는 유전 영역을 염색체 전체 영역 또는 특정 유전 영역 외 영역으로 결정하는 단계를 포함하는 방법으로 선별하는 것을 특징으로 할 수 있다.

[248] 본 발명에서 상기 e) 단계의 반복 시행은 100회 이상, 바람직하게는 1만에서 100만회 사이, 가장 바람직하게는 10만회인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[249]

[250] 본 발명에 있어서, 상기 (iv) 단계는 정상인 참조집단의 RepRD ratio 값을 샘플의 RepRD ratio 값과 비교하는 것을 특징으로 할 수 있다.

[251]

- [252] 본 발명에 있어서, 상기 정상인 참조집단의 RepRD ratio 값과 샘플의 RepRD ratio를 비교하는 방식은 두 값이 통계적으로 유의미하게 차이가 나는 것을 확인할 수 있는 방법이면 제한없이 사용가능하나, 바람직하게는 평균 및 표준편차 기반의 Z-score나 중앙값 기반의 Log 비, 그 외 분류 알고리즘을 통해 산출된 우도비(Likelihood) 등이 선택되는 방법일 수 있으며, 가장 바람직하게는 평균 및 표준편차 기반의 Z점수 계산방식일 수 있으나, 이에 한정되는 것은 아니다.
- [253] 본 발명에 있어서, 상기 Reads Distance Index 는 정상참조 집단과 분석하고자 하는 샘플의 Rep RD ratio 값의 비교를 통해 계산되는데, 비교를 하는 방법에는 Z점수와 같은 표준 점수 방식을 사용 할 수 있고, 임계치는 무한대의 양수 음수 등의 정수 또는 범위가 가능하고, 바람직하는 -3 또는 3 이 될 수 있으나, 이에 한정되는 것은 아니다.
- [254]
- [255] 본 발명은 또 다른 관점에서, 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부;
- [256] 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 및
- [257] 선별된 서열정보(reads)에 대하여 정렬된 리드 사이의 거리를 측정하여, RD 값(Read Distance)을 계산하고, 계산한 RD 값을 기반으로 유전 영역 별로 RDI 값(Read Distance Index)를 계산하여, RDI 값이 기준값 또는 구간 미만 또는 초과 일 경우, 염색체 이상이 있는 것으로 판정하는 염색체 이상 판정부를 포함하는 염색체 이상 검출 장치에 관한 것이다.
- [258]
- [259] 본 발명은 또 다른 관점에서, 컴퓨터 판독 가능한 저장 매체로서, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,
- [260] (A) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;
- [261] (B) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;
- [262] (C) 선별된 서열정보(reads)에 대하여 정렬된 리드 사이의 거리를 측정하여, RD 값(Read Distance)을 계산하는 단계; 및
- [263] (D) 상기 (C) 단계에서 계산한 RD 값을 기반으로 유전 영역 별로 RDI 값(Read Distance Index)를 계산하여, RDI 값이 기준 값 범위 미만 또는 초과 일 경우, 염색체에 이상이 있는 것으로 판정하는 단계를 통하여, 염색체의 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체에 관한 것이다.
- [264]
- [265] 실시예
- [266] 이하, 실시예를 통하여 본 발명을 더욱 상세히 설명하고자 한다. 이들 실시예는 오로지 본 발명을 예시하기 위한 것으로서, 본 발명의 범위가 이들 실시예에

의해 제한되는 것으로 해석되지는 않는 것은 당업계에서 통상의 지식을 가진 자에게 있어서 자명할 것이다.

[267]

[268] 실시예 1. 혈액에서 DNA를 추출하여, 차세대 염기서열 분석 수행

[269] 샘플의 혈액을 10mL씩 채취하여 EDTA Tube에 보관하였으며, 채취 후 2시간 이내에 1200g, 4°C 15분의 조건으로 혈장 부분만 1차 원심분리한 다음, 1차 원심분리된 혈장을 16000g, 4°C 10분의 조건으로 2차 원심분리하여 침전물을 제외한 혈장 상층액을 분리하였다. 분리된 혈장에 대해 Tiangenmicro DNA kit (Tiangen)을 사용하여 cell-free DNA를 추출하였다. Paired-end(PE) 데이터의 생산은, MGIEasy Cell-free DNA Library Prep set kit (MGI)를 사용해 Library preparation 과정을 수행 후, DNBseq G400장비 (MGI)를 이용했고 (50cycle \* 2), Single-end(SE) 데이터는 Truseq Nano DNA HT library prep kit (Illumina)를 사용해 Library preparation 과정 후, Nextseq500 (Illumina) 장비를 이용해 생산했다.

[270] PE 데이터는 약 천만개의 핵산단편에 대한 서열 정보를 얻을 수 있었고, SE 데이터는 약 백삼십만개의 핵산 단편에 대한 서열 정보를 얻을 수 있었다.

[271]

[272] 실시예 2. 서열정보 데이터의 품질관리 및 FD값 계산

[273] 염기서열 정보를 전처리하고, FD 값을 계산하기 전에 다음 일련의 과정을 진행하였다. 차세대염기서열분석기(NGS) 장비에서 생성된 fastq 형식의 파일을 BWA-mem 알고리즘을 사용하여 참조 염색체 Hg19 서열을 기준으로 라이브러리 서열을 정렬하였다. 라이브러리 서열의 정렬 시 오류가 발생할 확률이 있어 오류를 교정하는 두 가지 과정을 수행하였다. 우선, 중복된 라이브러리 서열에 대하여 제거 작업을 실시한 다음, BWA-mem 알고리즘에 의해 정렬된 라이브러리 서열 중 Mapping Quality Score가 60에 도달하지 못하는 서열을 제거하였다.

[274] 선별한 리드를 정렬된 방향에 따라 정방향 리드와 역방향 리드로 그룹화 한 다음, 가장 인접한 리드와의 거리를 FD 값으로 하기 식 3을 이용하여 계산하였고, 이의 개념은 도 2 및 도 3에 나타내었다. 하기 식 3의 D 함수는 유전체 위치의 차이값을 계산하는 함수이다. 하기 식 3의 a와 b는 핵산단편의 위치값이며, PE 시퀀싱의 경우 2개의 서열 정보의 정렬된 위치값의 최소값부터 최대값 중 어느 한 값이 될 수 있고, SE 시퀀싱의 경우 서열 정보의 정렬된 위치값이거나 위치값에 특정 값을 extension 한 값일 수 있다.

[275] 식 3:  $Fragment\ Distance\ (FD) = D(a,b) \mid a \in Fi, b \in Fi$

[276]

[277] 실시예 3. Extension에 따른 FD값의 차이 확인

[278] PE로 생산된 데이터는 핵산 단편의 시작과 말단의 위치정보를 알 수 있고, 중간 위치를 기준으로 각 핵산 단편간 거리를 계산 할 수 있다. PE로 생산된

데이터에서 무작위 적으로 For, Rev 리드로 그룹화 하고 For 로 분류된 리드는 For 리드의 5' 위치를 기준으로, Rev 로 분류된 리드는 Rev 리드의 3' 위치를 기준으로 FD를 계산한 다음, For 리드에는 80bp를 더하고, Rev 리드에는 80bp를 빼주는 extension을 수행하였다.

[279] 상기 과정의 FD 값과 extension을 수행한 과정의 FD 값의 차이를 비교한 결과, 도 5에 기재된 바와 같이 extension 한뒤 계산된 FD 값이 PE의 centered FD 값과 유사한 것이 확인을 확인하였으며, extension을 수행하지 않은 FD 값은 +166, -166 의 FD 값의 차이가 나는 것을 확인하였다.

[280]

[281] 실시예 4. FDI 값 계산

[282] 4-1. 염색체 수적 이상 검출을 위한 FDI 값 계산

[283] SE 시퀀싱 데이터를 이용해 FDI 값을 계산했으며, extension 값은 80bp를 설정했다. 이수성 여부를 확인 하고자 하는 염색체를, 각각 선별된 염색체의 집합의 RepFD 값의 중앙값 비를 Normalized Factor로 정의하고, 하기 식 4로 계산 하였다.

[284]

[285] [표1]

| 이수성<br>확인을 위한<br>염색체 | 염색체 집합 |
|----------------------|--------|
|----------------------|--------|

|    |                                                 |
|----|-------------------------------------------------|
| 13 | 1,2,3,4,5,6,7,8,9,10,11,12,14,15,16,17,19,20,22 |
| 18 | 1,2,3,4,5,6,7,8,9,10,11,12,14,15,16,17,19,20,22 |
| 21 | 1,2,3,4,5,6,7,8,9,10,11,12,14,15,16,17,19,20,22 |

[286] 식 4:  $Normalized\ Factor = Median\ of\ RepFD_{selected\ chromosome\ set}$

[287] (selected chromosome set: 상기 표에서 염색체 집합에 해당하는 부분임)

[288] 식 3과 식 4에서 계산한 FD와 Normalized Factor 를 이용해서 RepFD ratio를 식 5로 계산하였다.

[289] 식 5:  $RepFD\ ratio = RepFD_{Target\ chromosome} / Normalized\ Factor$

[290]

[291] 2000명의 정상인 참조 집단에서 RepFD ratio의 평균과 표준편차 값을 계산하고, 분석 하고자 하는 샘플의 Fragments Distance Index (FDI) 값을 식 6으로 계산하였다.

[292] 식 6:  $FDI = (MEAN(RepFD\ Ratio_{reference} - RepFD\ Ratio_{sample}) / SD(RepFD\ Ratio_{reference}))$

[293] 상기 식 6의 과정을 모든 핵산단편을 사용하는 경우(식 7), 정방향으로 정렬된

핵산단편을 사용하는 경우(식 8), 역방향으로 정렬된 핵산단편을 사용하는 경우(식 9) 각각 수행하였다.

[294] 식 7:  $FDI^{all} = \frac{\text{mean}(\text{RepFD}^{all} \text{ Ratio}_{reference}) - \text{RepFD}^{all} \text{ Ratio}_{sample}}{SD(\text{RepFD}^{all} \text{ Ratio}_{reference})}$

식 8:  $FDI^{For} = \frac{\text{mean}(\text{RepFD}^{For} \text{ Ratio}_{reference}) - \text{RepFD}^{For} \text{ Ratio}_{sample}}{SD(\text{RepFD}^{For} \text{ Ratio}_{reference})}$

식 9:  $FDI^{Rev} = \frac{\text{mean}(\text{RepFD}^{Rev} \text{ Ratio}_{reference}) - \text{RepFD}^{Rev} \text{ Ratio}_{sample}}{SD(\text{RepFD}^{Rev} \text{ Ratio}_{reference})}$

[295]

[296] 4-2. 염색체 수적 이상 검출을 위한 FDI 값의 성능 확인

[297] 정상 표준 집단 2000명의 샘플과 Trisomy 6 샘플을 포함한 임상 88 샘플의 분석 결과, 100% sensitivity와 100% specificity를 확인하였다.

[298] 양성 판별을 위한 임계치 값은  $FDI21^{all}$ ,  $FDI21^{For}$ ,  $FDI21^{Rev}$  모두 3을 사용하였다.

[299] 각각 샘플은 양성 판별을 위해 3개의 FDI 값을 계산하며, 모두 3 이상일 경우 최종 양성으로 판별하였다. 분석된 88 샘플 중 3개 샘플(G19NIPT261-3, G19NIPT261-10, G19NIPT261-13)은 1개의 FDI 값에서 양성으로 판별 되었지만, 나머지 2개의 FDI 값은 음성을 판별 되어, 최종적으로 음성으로 판정하였다.

[300]

[301] [표2]

| sample        | Result | FDI21.All | FDI21.For | FDI21.rev |
|---------------|--------|-----------|-----------|-----------|
| G19NIPT264-22 | T21    | 13.5189   | 18.9706   | 22.7449   |
| G19NIPT261-27 | T21    | 11.3797   | 15.7714   | 16.3134   |
| G19NIPT262-11 | T21    | 9.8365    | 13.4254   | 12.6415   |
| G19NIPT263-21 | T21    | 9.6024    | 13.3521   | 13.8184   |
| G19NIPT264-29 | T21    | 9.3665    | 12.3367   | 14.4694   |
| G19NIPT261-3  | T21    | 6.2652    | 7.9757    | 8.6841    |
| G19NIPT263-12 | N      | 2.8936    | 3.5516    | 1.5855    |
| G19NIPT263-10 | N      | 1.9594    | 2.6681    | 3.1888    |
| G19NIPT263-13 | N      | 1.7782    | 2.0879    | 3.8235    |

[302]

[303] 실시예 5. RD 값 기반의 분석을 위한 혈액에서 DNA를 추출하여, 차세대 염기서열 분석 수행

[304] 정상인 400명, Trisomy 21 175명, Trisomy 18 67명 및 Trisomy 13 26명의 혈액을 10mL씩 채취하여 EDTA Tube에 보관하였으며, 채취 후 2시간 이내에 1200g, 4°C, 15분의 조건으로 혈장 부분만 1차 원심분리한 다음, 1차 원심분리된 혈장을 16000g, 4°C, 10분의 조건으로 2차 원심분리하여 침전물을 제외한 혈장 상층액을 분리하였다. 분리된 혈장에 대해 Tiangenmicro DNA kit (Tiangen)을 사용하여

cell-free DNA를 추출하고, Truseq Nano DNA HT library prep kit (Illumina)를 사용해 Library preparation 과정을 수행한 다음, Nextseq500 장비 (Illumina) 를 75 Single-end 모드로 sequencing을 진행하였다.

[305] 그 결과, 샘플 당 약 13 million 개의 reads가 생산되는 것을 확인 하였다.

[306]

[307] 실시예 6. 서열정보 데이터의 품질관리 및 RD값 계산

[308] 염기서열 정보를 전처리하고, RD 값을 계산하기 전에 다음 일련의 과정을 진행하였다. 차세대염기서열분석기(NGS) 장비에서 생성된 Bcl 파일(염기서열정보 포함)을 fastq 형식으로 변환한 다음, fastq 파일을 BWA-mem 알고리즘을 사용하여 참조 염색체 Hg19 서열을 기준으로 라이브러리 서열을 정렬하였다. 라이브러리 서열의 정렬 시 오류가 발생할 확률이 있어 오류를 교정하는 두 가지 과정을 수행하였다. 우선, 중복된 라이브러리 서열에 대하여 제거 작업을 실시한 다음, BWA-mem 알고리즘에 의해 정렬된 라이브러리 서열 중 Mapping Quality Score가 60에 도달하지 못하는 서열을 제거하였다.

[309] 선별한 리드를 정렬된 방향에 따라 정방향 리드와 역방향 리드로 그룹화 한 다음, 가장 인접한 리드와의 거리를 RD 값으로 하기 식 12를 이용하여 계산하였고, 이의 개념은 도 7에 나타내었다. 하기 12의 D 함수는 유전체 위치의 차이값을 계산하는 함수이다.

[310] 식 12:  $Read\ Distance\ (RD) = D(a,b) \mid a \in Ri, b \in Ri$

[311]

[312] 실시예 7. RDI 값 계산

[313] 7-1. 염색체 수적 이상 검출을 위한 RDI 값 계산

[314] 각 염색체 별로 RD 값의 중앙값을 RepRD로 정의하였다. 이수성 여부를 확인 하고자 하는 염색체를 각각 선별된 염색체의 집합의 RepRD 값의 중앙값 비를 Normalized Factor로 정의하고, 하기 식 13으로 계산하였다.

[315]

[316] [표3]

| 이수성 확인을 위한<br>염색체 | 염색체 집합    |
|-------------------|-----------|
| 13                | 4,6       |
| 18                | 5,8       |
| 21                | 2,4,14,20 |

[317] 식 13:  $Normalized\ Factor = Median\ of\ RepRD_{selected\ chromosome\ set}$

[318] (selected chromosome set: 상기 표에서 염색체 집합에 해당하는 부분임)

[319] 식 12와 식 13에서 계산한 RD와 Normalized Factor 를 이용해서 RepRD ratio를 식 14로 계산하였다.

- [320] 식 14:  $RepRD\ ratio = RepRD_{Target\ chromosome} / Normalized\ Factor$
- [321] 400명의 정상인 참조 집단에서 RepRD ratio의 평균과 표준편차 값을 계산하고, 분석하고자 하는 샘플의 Reads Distance Index (RDI) 값을 식 15로 계산하였다.
- [322] 식 15:  $RDI = RepRD\ Ratio_{sample} - MEAN(RepRD\ Ratio_{reference}) / SD(RepRD\ Ratio_{reference})$
- [323]
- [324] 7-2. 염색체 구조적 이상을 위한 RDI 값 계산
- [325] 염색체를 50kbase 로 일정하게 나눈 뒤, 각 영역별로 RD 값의 중앙값을 RepRD로 정의 하였다. 또한 Normalized Factor 는 상염색체 RD 값의 중앙값을 사용하였다. 참조 집단은 정상 여성 437명 데이터를 이용했고, 염색체 각 영역별로 RepRD Ratio의 평균과 표준편차를 계산했다. RDI 값은 식 15을 인용해 계산하였다.
- [326]
- [327] 7-3. RD 대표값(RepRD) 계산 방식에 따른 성능 확인 (중앙값의 역수 이용)
- [328] 각 유전 영역(염색체 별)에 정렬된 서열정보들의 RD 값을 계산한 뒤, 이 값들의 중앙값의 역수를 RD 대표값(RepRD)으로 정의 하였다. 여기서 중앙값이란, 계산된 RD 값들을 크기의 순서대로 정렬 했을 때 가장 중앙에 위치하는 값을 의미한다. 예를 들어 1,2,100 과 같이 세 개의 값이 있을 때, 2가 가장 중앙에 있기 때문에 2가 중앙값이 된다.
- [329] 만약 짝수개의 RD 값이 있을 경우, 가운데 있는 두 값의 평균으로 중앙값을 결정 한다.. 예를 들어 1,10,90,200의 RD 값이 있을 경우 중앙에 위치한 10과 90의 평균인 50이 중앙값이 된다.. 분석 샘플은 Trisomy 21 로 확인된 49 샘플과 정상으로 확인된 3,448 샘플을 이용하였고, RepRD 값은 RD 값들의 중앙값의 역수를 사용하였다. 분석 방법은 정상인 3,448 샘플의 RepRD 값의 평균과 표준편차를 이용한 Z-score 방식으로 RDI 값을 계산 하였다. 분석 결과, 약 0.999 의 정확도로 샘플의 염색체 수 이상 여부를 검출 할 수 있었다(표 4, 도 15).
- [330]
- [331] [표4]

| Chromosome | Accuracy (95%CI)       | Sensitivity | Specificity | AUC    |
|------------|------------------------|-------------|-------------|--------|
| T21        | 0.9994(0.9979, 0.9999) | 0.9795      | 0.9997      | 1.0000 |

- [332]
- [333] 7-4. RD 대표값 (RepRD) 계산 방식에 따른 성능 확인 (평균 이용)
- [334] 각 유전 영역(염색체 별)에 정렬된 서열정보들의 RD 값을 계산한 뒤, 이 값들의 평균값을 RD 대표값(RepRD)으로 정의 하였다. 여기서 평균 값이란, 계산된 RD 값들의 산술평균값으로 만약 10,50,90 의 RD 값이 있다면,  $(10+50+90)/3$  인 50이

RD 대표값이 된다. 정상인 1,999 과 T21 163 샘플을 이용해, 정상인 집단의 RepRD 평균과 표준편차를 이용한 Z-score 방식으로 RDI 값을 계산 하였다. Normalized Factor로 사용한 염색체는 2,7,9,12,14 였다. 분석 결과, 약 0.9995 의 정확도로 샘플의 염색체 수 이상 여부를 검출 할 수 있었고, 임계값을 4.0으로 설정시 민감도는 0.999 특이도는 1.000인 것을 확인할 수 있었다(표 5, 도 16).

[335]

[336] [표5]

| Chromosome | Accuracy              | Sensitivity | Specificity | AUC    |
|------------|-----------------------|-------------|-------------|--------|
| T21        | 0.9995(0.9974,1.0000) | 0.9999      | 1.0000      | 1.0000 |

[337]

[338] 7-5. RD 대표값 (RepRD) 계산 방식에 따른 성능 확인(평균의 역수값 이용)

[339] 각 유전 영역(염색체 별)에 정렬된 서열정보들의 RD 값을 계산한 뒤, 이 값들의 평균값의 역수값을 RD 대표값(RepRD)으로 정의 하였다. 여기서 평균 값이란, 계산된 RD 값들의 산술평균값으로 만약 10,50,90 의 RD 값이 있다면,  $(10+50+90)/3$  인 50 이 평균값이 되고, 이 값의 역수인  $1/50 = 0.02$  를 RD 대표값으로 이용했다. 정상인 1,999 과 T21 163 샘플을 이용해, 정상인 집단의 RepRD 평균과 표준편차를 이용한 Z-score 방식으로 RDI 값을 계산 하였다. Normalized Factor로 사용한 염색체는 2,7,8,9,12,14 였다. 분석 결과, 약 0.9995 의 정확도로 샘플의 염색체 수 이상 여부를 검출 할 수 있었고, 임계값을 4.3으로 설정시 민감도는 0.993 특이도는 1.000인 것을 확인할 수 있었다(표 6, 도 17).

[340]

[341] [표6]

| Chromosome | Accuracy              | Sensitivity | Specificity | AUC   |
|------------|-----------------------|-------------|-------------|-------|
| T21        | 0.9995(0.9974,1.0000) | 0.9939      | 1.000       | 1.000 |

[342]

[343] 실시예 8. 염색체 수적 이상 검출을 위한 RDI 값의 성능 확인

[344] 8-1. Read count 와 Read Distance의 분포

[345] 정렬된 Reads의 거리 개념을 이용한 분석에서, 생산된 reads의 수가 많을수록, 각 Reads사이의 거리는 짧게 유지가 될 것이다. 이를 확인하기 위해, 각 염색체별로 Reads 수와 RepRD값의 분포를 분석하였다.

[346] 그 결과, 전체적으로 Reads 수가 많을수록 RepRD이 감소되는 것을 확인했다. 특히, Reads 수와 RepRD 값의 관계가 선형 관계가 아닌 비 선형 관계임을 확인했고, 이는 Reads 거리 개념이 단순 Reads의 수뿐만 아니라 정렬된 위치를 모두 반영하는 결과이다(도 8).

[347] 정상 샘플과 비교하면 Trisomy 13,18 그리고 21번 샘플에서 각각 이수성으로

확인된 염색체의 RepRD 값이 낮게 분포하는 것을 확인할 수 있었다(도 9).

[348]

[349] 8-2. Reads Distance Index 의 성능 및 태아분획, 임상정보, 기존 G-score 와의 관계

[350] 산모 혈액을 이용한 태아 이수성 검사에서, 태아분획, 임신 주수는 검사의 정확도에 많은 영향을 미친다. 임신주수가 높을수록 태아 분획은 높아지는 경향성이 있고, 태아분획이 높을수록 검사의 정확도는 높아진다. Trisomy 21 샘플의  $RDI_{chr21}$  값과 산모의 임신주수, 태아분획과의 분포를 분석한 결과, 태아분획이 높아지면서  $RDI_{chr21}$  값이 떨어지는 것을 확인하였다. 또 임신주수와  $RDI_{chr21}$  값의 관계에서는, 15주 이상의 샘플에서 값이 떨어지는 경향성을 확인하였으며, 기존 Reads count 에 기반한 값인, G-score(대한민국 특허 제10-1686146호)와의 관계를 살펴본 결과, 유사한 경향성을 확인할 수 있었다(도 10).

[351]

[352] 8-3. 양성 판별 임상 검체 분석 결과

[353] RDI 값을 이용해, 정상군과 각 염색체 이수성으로 확인된 샘플을 분석 성능을 검증하였다. RDI 값을 일정기준 cutoff (-3) 으로 설정 후 정상, 이수성 샘플간의 분석성능을 비교해본 결과, Trisomy 13 은 0.991, Trisomy 18 은 0.989, Trisomy 21 은 0.998의 정확도가 확인 되었다(표 7). 또한 AUC 값은 Trisomy 13,18,21 에서 각각 0.999, 0.984, 1.000인 것을 확인할 수 있었다(도 11).

[354]

[355] [표7]

| Chromosome | Accuracy (95%CI)   | Sensitivity | Specificity | AUC   |
|------------|--------------------|-------------|-------------|-------|
| T13        | 0.991(0.976-0.997) | 0.846       | 1.000       | 0.999 |
| T18        | 0.989(0.975-0.997) | 0.925       | 1.000       | 0.984 |
| T21        | 0.998(0.99-1)      | 1.000       | 0.998       | 1.000 |

[356]

[357] 8-4. RDI 계산방식에 따른 성능 확인

[358] 정상 참조 집단 RDI ratio 의 평균과 표준편차를 이용하는 Z-score 방식과 다른, 중앙값을 이용한 Log ratio 분석 결과를 확인 하였다. Log ratio의 분석방법은 식 16을 이용하였다.

[359] 식 16:  $RDI = \log_{10}(\text{RepRD Ratio}_{\text{sample}} / \text{Median}(\text{RepRD Ratio}_{\text{reference}}))$

[360] 실시예 8-3 에서 이용한 것과 같은 샘플을 이용하였고, RDI 값을 일정기준 cutoff (-0.0045) 으로 설정 후 분석성능을 비교해 봤다. 성능은 양성 종류에 따라 조금씩 차이가 있었으며, 정확도는 Trisomy 21 은 0.976, Trisomy 18 은 0.994, 그리고 Trisomy 13 은 0.991로 확인 되었다(표 8).

[361]

[362] [표8]

| Chromosome | Accuracy (95%CI)   | Sensitivity | Specificity | AUC   |
|------------|--------------------|-------------|-------------|-------|
| T13        | 0.991(0.976-0.997) | 0.846       | 1.000       | 0.999 |
| T18        | 0.994(0.981-0.999) | 0.955       | 1.000       | 0.984 |
| T21        | 0.976(0.959-0.987) | 1.000       | 0.965       | 1.000 |

[363]

[364] 8-5. Down sampling 성능 확인

[365] 차세대 염기서열 분석 기술을 이용해, 비 침습적 방법의 태아 이수성 여부를 확인하는 검사에서 생산되는 데이터의 양(리드 수)은 정확도의 중요한 요소로 알려져 있다. 본 실시예에서, 리드 수에 따른 RDI 방식의 분석성능을 계산해 보았다. 분석 성능의 기준은 ROC 분석의 AUC 값을 이용했고, 리드의 수는 in-silico 방식의 무작위 리드 선별 방법을 이용했다. 무작위적으로 리드를 백만 개에서 천만 개를 선별했다. 21번 이수성 샘플을 이용한 분석 결과 리드 수가 줄어들수록 분석 성능이 낮아지는 것을 확인할 수 있었다(도 12).

[366]

[367] 실시예 9. 염색체 구조적 이상 검출을 위한 RDI 값의 성능 확인

[368] 9-1. Read count 와 Read Distance의 분포

[369] RDI를 이용한 염색체 구조적 이상 여부를 보기 위해, 염색체를 적당한 크기로 나누는 작업이 필요하며, 본 실시예에선 50k base 의 크기로 염색체의 구간을 나눴다. 리드의 거리는 리드의 수가 많을수록 작고, 리드의 수가 많을수록 길게 분포한다. 나뉜 구간에 해당하는 리드의 수와 거리의 관계를 살펴본 결과, 염색체의 구조적 이상인 결실이 확인된 영역의 리드 거리가 구조적 이상이 없는 영역보다 리드 거리가 길게 분포하는 것을 확인할 수 있었다(도 13).

[370]

[371] 9-2. Microarray 의 결과와 비교

[372] 염색체 구조적 이상 여부를 검출하는 Microarray 검사와 RDI 의 분석 결과를 비교 하였다. 분석 샘플은 1번 염색체의 끝단에 3,897,640 bp 길이의 결실이 확인되었던 샘플이고, RDI를 이용한 분석 결과 비슷한 지역의 3,700,000 bp 사이즈에서 구조적 이상(결실)이 검출 되는 것을 확인할 수 있었다(도 14).

[373]

[374] 이상으로 본 발명 내용의 특정한 부분을 상세히 기술하였는 바, 당업계의 통상의 지식을 가진 자에게 있어서 이러한 구체적 기술은 단지 바람직한 실시 양태일 뿐이며, 이에 의해 본 발명의 범위가 제한되는 것이 아닌 점은 명백할 것이다. 따라서, 본 발명의 실질적인 범위는 첨부된 청구항들과 그것들의 등가물에 의하여 정의된다고 할 것이다.

[375]

### 산업상 이용가능성

- [376] 본 발명에 따른 염색체 이상 판정 방법은, 기존의 리드 개수(read count) 기반으로 염색체 양을 결정하는 단계를 이용하는 방식과는 달리, 정렬된 핵산 단편(fragments)을 그룹화한 다음, 핵산 단편 기준값 사이의 거리 개념을 이용한 방법으로, 기존 방법이 리드 개수가 감소하면 정확도가 떨어지나, 본 발명의 방법은 리드 개수가 감소하더라도, 검출의 정확도를 높일 수 있을 뿐만 아니라, 모든 염색체 구간이 아닌 일정 구간의 핵산 단편 사이의 거리를 분석하여도 검출 정확도가 높아 유용하다.

## 청구범위

- [청구항 1] 생체시료에서 추출한 핵산 단편(fragments) 기준값 사이의 거리를 계산하여 염색체 이상을 검출하는 방법.
- [청구항 2] 제1항에 있어서, 상기 핵산 단편은 세포 유리 핵산 또는 세포 내 핵산인 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 3] 제1항에 있어서, 상기 핵산 단편은 직접 서열분석하거나, 차세대 염기서열 분석을 통해 서열분석하거나 또는 비특이적 전장 유전체 증폭(non-specific whole genome amplification)을 통해 서열분석하여 얻은 것임을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 4] 제3항에 있어서,  
 (A) 생체시료에서 핵산을 추출하여 핵산 단편을 획득하여 서열정보를 획득하는 단계;  
 (B) 획득한 서열정보(reads)에 기반하여 표준 염색체 서열 데이터베이스(reference genome database)에서 핵산 단편의 위치를 확인하는 단계;  
 (C) 상기 서열정보(reads)를 전체 서열, 정방향 서열 및 역방향 서열로 그룹화 하는 단계;  
 (D) 상기 그룹화된 서열정보를 이용하여, 각 핵산단편의 기준값을 정의하고, 기준값 사이의 거리를 측정하여, 각 그룹별 FD 값(Fragments Distance)을 계산하는 단계; 및  
 (E) 상기 (D) 단계에서 계산한 각 그룹별 FD 값을 기반으로 염색체 전체 영역 또는 특정 영역 별로 각각의 FDI 값(Fragments Distance Index)을 계산하여, 각각의 FDI 값이 모두 기준값 범위 미만 혹은 초과일 경우, 염색체 이상이 있는 것으로 판정하는 단계를 포함하는 염색체 이상을 검출하는 방법.
- [청구항 5] 제4항에 있어서, 상기 (A) 단계는 다음의 단계를 포함하는 방법으로 수행되는 것을 특징으로 하는 염색체 이상을 검출하는 방법:  
 (A-i) 혈액, 정액, 질 세포, 모발, 타액, 소변, 구강세포, 태반세포 또는 태아세포를 포함하는 양수, 조직세포 및 이의 혼합물에서 핵산을 획득하는 단계;  
 (A-ii) 채취된 핵산에서 솔팅-아웃 방법(salting-out method), 컬럼 크로마토그래피 방법(column chromatography method) 또는 비드 방법(beads method)을 사용하여 단백질, 지방, 및 기타 잔여물을 제거하고 정제된 핵산을 획득하는 단계;  
 (A-iii) 정제된 핵산 또는 효소적 절단, 분쇄, 수압 절단 방법(hydroshear method)으로 무작위 단편화(random fragmentation)된 핵산에 대하여, 싱글 엔드 시퀀싱(single-end sequencing) 또는 페어 엔드 시퀀싱(pair-end

sequencing) 라이브러리(library)를 제작하는 단계;

(A-iv) 제작된 라이브러리를 차세대 유전자서열검사기(next-generation sequencer)에 반응시키는 단계; 및

(A-v) 차세대 유전자서열검사기에서 핵산의 서열정보(reads)를 획득하는 단계.

- [청구항 6] 제4항에 있어서, 상기 (D) 단계의 FD 값은 수득한 n개의 핵산 단편에 대하여, i 번째 핵산 단편과 i+1 내지 n 번째 핵산 단편에서 선택되는 어느 하나 이상의 핵산 단편의 기준값 사이의 거리를 통해 산출하는 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 7] 제6항에 있어서, 상기 핵산 단편의 기준값은 핵산 단편의 중앙값으로부터 임의의 값을 더하거나 빼 값인 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 8] 제7항에 있어서, 상기 핵산 단편의 기준값은 페어드 엔드 시퀀싱(paired-end sequencing)일 경우, 정방향 및 역방향 서열정보(reads)의 위치값을 기반으로 도출하는 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 9] 제8항에 있어서, 서열정보(reads)의 정렬 점수가 기준값 미만인 핵산 단편의 경우, 계산과정에서 제외하는 단계를 추가로 포함하는 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 10] 제6항에 있어서, 상기 핵산 단편의 기준값은 싱글 엔드 시퀀싱(single-end sequencing)일 경우, 정방향 또는 역방향 서열정보(read)의 위치값 한 종류를 기반으로 도출하는 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 11] 제10항에 있어서, 정방향으로 정렬된 서열정보를 기반으로 위치값을 도출할 경우에는 임의의 값을 더해주며, 역방향으로 정렬된 서열정보를 기반으로 위치값을 도출할 경우에는 임의의 값을 빼주는 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 12] 제7항에 있어서, 상기 임의의 값은 분석대상 핵산 평균 길이의 30 내지 70%인 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 13] 제7항에 있어서, 상기 임의의 값은 0 내지 5kbp 또는 핵산단편 길이의 0 내지 300%인 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 14] 제4항에 있어서, 상기 (E) 단계는 다음의 단계를 포함하는 방법으로 수행되는 것을 특징으로 하는 염색체 이상을 검출하는 방법:  
 (E-i) 염색체 전체 영역 또는 특정 영역별로 FD 값의 대표값(RepFD)을 결정하는 단계;  
 (E-ii) 분석하고자 하는 염색체 전체 영역 또는 특정 영역 외의 샘플 내 특정 영역의 RepFD 값의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 절대 편차, 변동 계수, 이들의

- 역수값 및 조합으로 구성된 군에서 선택되는 하나 이상의 값을 계산하여, 정규화 요소(Normalized Factor)를 도출하는 단계;
- (E-iii) 하기 식 1을 바탕으로 대표값 비율(RepFD ratio)를 계산하는 단계;  
 식 1:  $\text{RepFD ratio} = \text{RepFD Target genomic region} / \text{Normalized Factor}$
- (E-iv) 정상인 참조집단과 샘플의 RepFD ratio 값을 비교하여, FDI (Fragments Distance Index)를 계산하는 단계.
- [청구항 15] 제14항에 있어서, 상기 (E-i) 단계의 대표값(RepFD)은 FD 값들의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값인 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 16] 제15항에 있어서, 상기 (E-i) 단계의 대표값(RepFD)은 FD 값들의 중앙값, 평균값 또는 이의 역수값인 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 17] 제14항에 있어서, 상기 (E-ii) 단계의 분석하고자 하는 염색체 전체 영역 또는 특정 유전 영역 외의 샘플 내 특정 영역은 하기의 단계를 포함하는 방법으로 선별하는 것을 특징으로 하는 염색체 이상을 검출하는 방법:  
 a) 무작위로 분석하고자 하는 염색체 전체 영역 또는 특정 유전 영역 외의 영역을 무작위로 선별하는 단계;  
 b) 상기 a) 단계에서 선별한 유전 영역의 대표값(RepFD)을 사전 정규화 요소(Pre Normalized Factor, PNF)로 결정하는 단계;  
 c) 하기 식 2를 바탕으로 대표값 비율(RepFD ratio)를 계산하는 단계:  
 식 2:  $\text{RepFD ratio} = \text{RepFD Target genomic region} / \text{PNF}$   
 d) 정상인 참조 집단의 RepFD ratio 값의 변동 계수(Coefficient of Variance:  $\text{SD} / \text{Mean}$ )를 계산하는 단계; 및  
 e) 상기 a) 내지 d) 단계를 반복 시행하여 수득한 변동 계수 중 가장 작은 값을 갖는 유전 영역을 염색체 전체 영역 또는 특정 유전 영역 외의 샘플내 특정 영역으로 결정하는 단계.
- [청구항 18] 제14항에 있어서, 상기 (E-iv) 단계는 정상인 참조집단의 RepFD ratio 값을 샘플의 RepFD ratio 값과 비교하는 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 19] 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부; 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 및 선별된 핵산 단편(fragments)에 대하여 정렬된 핵산 단편 사이의 거리를 측정하여, FD 값(Fragments Distance)을 계산하고, 계산한 FD 값을 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 FDI 값(Fragments Distance Index)를 계산하여, FDI 값이 기준 값 또는 구간 미만 또는 초과 일 경우, 염색체 이상이 있는 것으로 판정하는 염색체 이상 판정부를

포함하는 염색체 이상 검출 장치.

- [청구항 20] 컴퓨터 판독 가능한 저장 매체로서, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,  
 (A) 생체시료에서 핵산을 추출하여 핵산 단편을 획득하여 서열정보를 획득하는 단계;  
 (B) 획득한 서열정보(reads)에 기반하여 핵산 단편을 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;  
 (C) 선별된 핵산 단편(fragments) 사이의 거리를 측정하여, FD 값(Fragments Distance)을 계산하는 단계; 및  
 (D) 상기 (C) 단계에서 계산한 FD 값을 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 FDI 값(Fragments Distance Index)를 계산하여, FDI 값이 기준값 범위 미만 또는 초과 일 경우, 염색체 이상이 있는 것으로 판정하는 단계를 통하여, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체.
- [청구항 21] 제7항에 있어서, 상기 임의의 값이 분석대상 핵산 평균 길이의 50%일 경우, 계산한 FD값은 RD값(Read Distance)인 것을 특징으로 하는 염색체 이상을 검출하는 방법.
- [청구항 22] (A) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;  
 (B) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;  
 (C) 상기 정렬된 서열정보(reads)에 대하여 정렬된 리드 사이의 거리를 측정하여, RD 값(Read Distance)을 계산하는 단계; 및  
 (D) 상기 (C) 단계에서 계산한 RD 값을 기반으로 염색체 전체 영역 또는 특정 영역 별로 RDI 값(Read Distance Index)를 계산하여, RDI 값이 기준값 범위 미만 혹은 초과일 경우, 염색체 이상이 있는 것으로 판정하는 단계를 포함하는 염색체 이상 검출방법.
- [청구항 23] 제22항에 있어서, 상기 (A) 단계는 다음의 단계를 포함하는 방법으로 수행되는 것을 특징으로 하는 염색체 이상 검출방법:  
 (A-i) 혈액, 정액, 질 세포, 모발, 타액, 소변, 구강세포, 태반세포 또는 태아세포를 포함하는 양수, 조직세포 및 이의 혼합물에서 핵산을 획득하는 단계;  
 (A-ii) 채취된 핵산에서 솔팅-아웃 방법(salting-out method), 컬럼 크로마토그래피 방법(column chromatography method) 또는 비드 방법(beads method)을 사용하여 단백질, 지방, 및 기타 잔여물을 제거하고 정제된 핵산을 획득하는 단계;  
 (A-iii) 정제된 핵산 또는 효소적 절단, 분쇄, 수압 절단 방법(hydroshear method)으로 무작위 단편화(random fragmentation)된 핵산에 대하여, 싱글 엔드 시퀀싱(single-end sequencing) 또는 페어 엔드 시퀀싱(pair-end

sequencing) 라이브러리(library)를 제작하는 단계;

(A-iv) 제작된 라이브러리를 차세대 유전자서열검사기(next-generation sequencer)에 반응시키는 단계; 및

(A-v) 차세대 유전자서열검사기에서 핵산의 서열정보(reads)를 획득하는 단계.

[청구항 24] 제22항에 있어서, 상기 (C) 단계 이전에 정렬된 리드를 정렬된 방향에 따라 그룹화하는 단계를 추가로 사용할 수 있는 것을 특징으로 하는 염색체 이상 검출 방법.

[청구항 25] 제22항에 있어서, 상기 (C) 단계의 RD 값은 수득한 n개의 리드에 대하여, i 번째 리드와 i+1 내지 n 번째 리드에서 선택되는 어느 하나 이상의 리드의 양 말단 값 중 하나의 값에 핵산 평균 길이의 50%를 더하거나 빼 값 사이의 거리를 통해 산출하는 것을 특징으로 하는 염색체 이상을 검출하는 방법.

[청구항 26] 제22항에 있어서, RD 값은 상기 i 번째 리드 내부의 5' 또는 3' 말단과 i+1 내지 n 번째 중 어느 하나 이상의 리드의 5' 또는 3' 말단 사이의 거리를 계산하는 것을 특징으로 하는 염색체 이상 검출 방법.

[청구항 27] 제22항에 있어서, 상기 (D) 단계는 다음의 단계를 포함하는 방법으로 수행되는 것을 특징으로 하는 염색체 이상 검출방법:

(D-i) 염색체 전체 영역 또는 특정 영역별로 RD 값의 대표값(RepRD)을 결정하는 단계;

(D-ii) 분석하고자 하는 염색체 전체 영역 또는 특정 영역 외의 샘플 내 특정 영역의 RepRD 값의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값을 계산하여, 정규화 요소(Normalized Factor)를 도출하는 단계;

(D-iii) 하기 식 10을 바탕으로 대표값 비율(RepRD ratio)를 계산하는 단계;  
식 10:  $\text{RepRD ratio} = \text{RepRD Target genomic region} / \text{Normalized Factor}$

(D-iv) 정상인 참조집단과 샘플의 RepRD ratio 값을 비교하여, RDI (Read Distance Index)를 계산하는 단계;

[청구항 28] 제27항에 있어서, 상기 (D-i) 단계의 대표값(RepRD)은 RD값의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값인 것을 특징으로 하는 염색체 이상 검출방법.

[청구항 29] 제28항에 있어서, 상기 (D-i) 단계의 대표값(RepRD)은 RD값들의 중앙값, 평균값 또는 이의 역수값인 것을 특징으로 하는 염색체 이상 검출방법.

[청구항 30] 제27항에 있어서, 상기 (D-ii) 단계의 분석하고자 하는 염색체 전체 영역 또는 특정 유전 영역 외의 샘플 내 특정 영역은 하기의 단계를 포함하는

방법으로 선별하는 것을 특징으로 하는 염색체 이상 검출방법:

a) 무작위로 분석하고자 하는 염색체 전체 영역 또는 특정 유전 영역 외의 영역을 무작위로 선별하는 단계;

b) 상기 a) 단계에서 선별한 유전 영역의 RepRD 값의 대표값을 사전 정규화 요소(Pre Normalized Factor, PNF)로 결정하는 단계;

c) 하기 식 11을 바탕으로 대표값 비율(RepRD ratio)를 계산하는 단계:

식 11:  $\text{RepRD ratio} = \text{RepRD Target genomic region} / \text{PNF}$

d) 정상인 참조 집단의 RepRD ratio 값의 변동 계수(Coefficient of Variance:  $\text{SD} / \text{Mean}$ )를 계산하는 단계; 및

e) 상기 a) 내지 d) 단계를 반복 시행하여 수득한 변동 계수 중 가장 작은 값을 갖는 유전 영역을 염색체 전체 영역 또는 특정 유전 영역 외의 샘플내 특정 영역으로 결정하는 단계.

[청구항 31] 제27항에 있어서, 상기 (D-iv) 단계는 정상인 참조집단의 RepRD ratio 값을 샘플의 RepRD ratio 값과 비교하는 것을 특징으로 하는 염색체 이상 검출방법.

[청구항 32] 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부; 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 및 선별된 서열정보(reads)에 대하여 정렬된 리드 사이의 거리를 측정하여, RD 값(Read Distance)을 계산하고, 계산한 RD 값을 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 RDI 값(Read Distance Index)를 계산하여, RDI 값이 기준 값 또는 구간 미만 또는 초과 일 경우, 염색체 이상이 있는 것으로 판정하는 염색체 이상 판정부를 포함하는 염색체 이상 검출 장치.

[청구항 33] 컴퓨터 판독 가능한 저장 매체로서, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,

(A) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;

(B) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

(C) 선별된 서열정보(reads)에 대하여 정렬된 리드 사이의 거리를 측정하여, RD 값(Read Distance)을 계산하는 단계; 및

(D) 상기 (C) 단계에서 계산한 RD 값을 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 RDI 값(Read Distance Index)를 계산하여, RDI 값이 기준 값 범위 미만 또는 초과 일 경우, 염색체 이상이 있는 것으로 판정하는 단계를 통하여, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체.