

명세서

발명의 명칭: 인공지능 기반 염색체 이상 검출 방법

기술분야

- [1] 본 발명은 인공지능 기반 염색체 이상 검출 방법에 관한 것으로, 보다 구체적으로는 생체시료에서 핵산을 추출하여, 서열정보를 획득하여 정렬한 리드를 기반으로 벡터화된 데이터를 생성한 후, 학습된 인공지능 모델에 입력하여 계산된 값과 기준값을 비교하는 방법을 이용한 인공지능 기반 염색체의 이상 검출방법에 관한 것이다.

[2]

배경기술

- [3] 염색체 이상(chromosomal abnormality)은 유전적 결함과 중앙 질환과 관련 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 중앙 질환을 유발한다. 예컨대, 다운증후군(Down's syndrome)은 21번 염색체가 3개 존재하여(trisomy 21) 유발되는 염색체 수 이상의 흔한 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 중앙 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 위치화 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] 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 확진검사로 분류한다. 침습적 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.
- [7] 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 재형성과정 중 세포사멸과정을 겪은 영양막 세포의 일부분이 물질교환 기전을 통해 모체 혈액으로 들어간 것으로 실제로는 태반으로부터 유래하고 이를 cff DNA(cell-free fetal DNA)라 정의한다.
- [8] cff DNA는 빠르면 배아 이식 18일째부터, 37일째에는 대부분의 모체 혈액 내에서 발견된다. cff DNA는 300bp 이하의 짧은 가닥이며 모체혈액 내 소량으로 존재하는 특징을 가지고 있기 때문에 이를 태아염색체 이상 검출에 적용하기 위하여 차세대염기서열분석기법(NGS)을 이용한 대규모 병렬 염기분석 기술이 사용되고 있다. 대규모 병렬 염기분석 기술을 이용한 침습적 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).
- [9] 한편, 공학분야에서 빈번하게 접하게 되는 입력 패턴을 특정 그룹으로 분류하는 문제를 해결하는 방안으로서, 인간이 지니고 있는 효율적인 패턴 인식 방법을 실제 컴퓨터에 적용시키려는 연구가 활발히 진행되고 있다.
- [10] 여러 가지 컴퓨터 적용 연구들 중에는 효율적인 패턴 인식 작용이 일어나는 인간두뇌 세포구조를 공학적으로 모델링한 인공신경망(Artificial Neural Network)에 대한 연구가 있다. 입력 패턴을 특정 그룹으로 분류하는 문제를 해결하기 위해, 인공신경망은 인간이 가지고 있는 학습이라는 능력을 모방한 알고리즘을 사용한다. 이 알고리즘을 통하여 입력 패턴과 출력 패턴들 사이의 사상(Mapping)을 인공신경망이 생성해낼 수 있는데, 이를 인공신경망이 학습 능력이 있다고 표현한다. 또한, 인공신경망은 학습된 결과를 바탕으로 학습에 사용되지 않았던
- [11] 입력 패턴에 대하여 비교적 올바른 출력을 생성할 수 있는 일반화 능력이 있다. 학습과 일반화라는 두 가지 대표적인 성능 때문에 인공신경망은 기존의 순차적 프로그래밍 방법에 의해서는 좀처럼 해결하기 힘든 문제에 적용되고 있다.

인공신경망은 그 사용범위가 넓어 패턴 분류 문제, 연속 사상, 비선형 시스템 식별, 비선형 제어 및 로봇 제어 분야 등에 활발히 응용되고 있다.

- [12] 인공 신경망이란 연결선으로 연결된 많은 수의 인공 뉴런들을 이용하여 생물학적인 시스템의 계산 능력을 모방하는 소프트웨어나 하드웨어로 구현된 연산모델을 나타낸다. 인공 신경망에서는 생물학적인 뉴런의 기능을 단순화시킨 인공 뉴런을 사용하게 된다. 그리고 연결강도를 갖는 연결선을 통해 상호 연결시켜 인간의 인지작용이나 학습과정을 수행하게 된다. 연결강도는 연결선이 갖는 특정 값으로, 연결가중치라고도 한다. 인공신경망의 학습은 지도 학습과 비지도 학습으로 나눌 수 있다. 지도 학습이란 입력 데이터와 그에 대응하는 출력 데이터를 함께 신경망에 넣고, 입력 데이터에 대응하는 출력 데이터가 출력되도록 연결선들의 연결강도를 갱신시키는 방법이다. 대표적인 학습 알고리즘으로는 델타규칙(Delta Rule)과 오류 역전파 학습(Back propagation Learning)이 있다. 비지도 학습이란 목표 값 없이 입력 데이터만을 사용하여 인공신경망이 스스로 연결강도를 학습시키는 방법이다. 비지도 학습은 입력 패턴들 사이의 상관관계에 의해 연결가중치들을 갱신시켜 나가는 방법이다.
- [13] 기계학습에서 적용되는 많은 데이터는 복잡해지고 차원이 늘어남에 따라 차원의 저주(curse of dimensionality)의 문제가 발생한다. 즉 이는, 필요한 데이터의 차원이 무한으로 갈수록 임의의 두 점간의 거리가 무한대로 발산하며 데이터의 존재량, 즉 밀도가 고차원의 공간에서는 다소 낮아져 데이터의 특성(Feature)을 제대로 반영하지 못하게 되는 것이다(Richard Bellman, Dynamic Programming, 2003, chapter 1). 최근 심층신경망(deep learning)의 발달은 입력층(input layer)과 출력층(output layer) 사이에 숨겨진 층(hidden layer)이 있는 구조로, 입력층으로부터 전달되는 변수 값의 선형 결합(linear combination)을 비선형 함수로 처리하면서 이미지, 영상, 신호데이터 등의 고차원의 데이터에서의 분류기(classifier)의 성능을 크게 향상시켰다고 보고되었다(Hinton, Geoffrey, et al., IEEE Signal Processing Magazine Vol. 29.6, pp. 82-97, 2012).
- [14] 이러한 인공신경망을 이용하여 바이오 분야에 활용하는 다양한 특허(KR 10-2017-0185041, KR 10-2017-0144237, KR 10-2018-124550)가 존재하고 있으나, cfDNA의 서열분석 정보를 기반으로 인공신경망 분석을 통해 염색체 이상을 검출하는 방법에 대해서는 연구가 부족한 실정이다.
- [15] 이에, 본 발명자들은 상기 문제점들을 해결하고, 높은 민감도와 정확도의 인공지능 기반 염색체 이상 검출방법을 개발하기 위해 예의 노력한 결과, 염색체 영역에 정렬되는 리드를 기반으로 벡터화된 데이터를 생성하고, 이를 학습된 인공지능 모델로 분석할 경우, 높은 민감도와 정확도로 염색체 이상을 검출할 수 있다는 것을 확인하고, 본 발명을 완성하였다.
- [16]
- [17] **발명의 요약**
- [18] 본 발명의 목적은 인공지능 기반 염색체 이상 검출 방법을 제공하는 것이다.

- [19] 본 발명의 다른 목적은 인공지능 기반 염색체 이상 판정 장치를 제공하는 것이다.
- [20] 본 발명의 또 다른 목적은 상기 방법으로 염색체의 이상을 판정하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공하는 것이다.
- [21] 상기 목적을 달성하기 위하여, 본 발명은 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)를 기반으로 벡터화된 데이터를 생성하는 단계; d) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석하고, 기준값(cut-off value)과 비교하여 염색체 이상 유무를 판정하는 단계를 포함하는 인공지능 기반 염색체 이상 검출방법을 제공한다.
- [22] 본 발명은 또한, 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부; 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 정렬된 서열정보(read)를 기반으로 벡터화된 데이터를 생성하는 데이터 생성부; 생성된 벡터화된 데이터를 학습된 인공지능 모델에 입력하고 분석하여 염색체 이상 유무를 판정하는 염색체 이상 판정부를 포함하는 인공지능 기반 염색체 이상 검출 장치를 제공한다.
- [23] 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) 상기 정렬된 서열정보(reads)를 기반으로 벡터화된 데이터를 생성하는 단계; d) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석하고, 기준값(cut-off value)과 비교하여 염색체 이상 유무를 판정하는 단계를 통하여, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공한다.

[24]

도면의 간단한 설명

- [25] 도 1은 본 발명의 인공지능 기반 염색체 이상을 판정하기 위한 전체 흐름도이다.
- [26] 도 2는 본 발명의 일 실시예에 따라 생성한 GC plot의 예시로서, X축은 구간별 염색체이며, Y축은 각 구간에 해당되는 핵산단편 사이의 거리값을 의미한다.
- [27] 도 3의 (A)는 본 발명의 일 실시예에 따라 생성한 정상 샘플의 GCW plot의 예시이며, (B)는 21번 염색체가 3개인 샘플의 GCW plot의 예시로서, X축은 염색체, Y축은 염색체에 해당하는 핵산단편 사이의 거리값을 Z점수로 변환한 것을 의미한다.

- [28] 도 4의 (A)는 본 발명의 일 실시예에 따라 생성한 정상 샘플의 GCW plot의 예시이며, (B)는 21번 염색체가 3개인 샘플의 GCW plot의 예시로서, X축은 염색체, Y축은 염색체에 해당하는 리드의 수를 Z점수로 변환한 것을 의미한다.
- [29] 도 5의 (A)는 본 발명의 일 실시예에 따라, 핵산단편 사이의 거리를 기반으로 생성한 GC plot 이미지 데이터를 학습한 딥러닝 모델에 대하여 trisomy 21 판정의 정확도를 확인한 결과이며, (B)는 각 데이터 세트별 확률분포를 나타낸 결과이다.
- [30] 도 6의 (A)는 본 발명의 일 실시예에 따라, 핵산단편 사이의 거리를 기반으로 생성한 GC plot 이미지 데이터를 학습한 딥러닝 모델에 대하여 trisomy 18 판정의 정확도를 확인한 결과이며, (B)는 각 데이터 세트별 확률분포를 나타낸 결과이다.
- [31] 도 7의 (A)는 본 발명의 일 실시예에 따라, 핵산단편 사이의 거리를 기반으로 생성한 GC plot 이미지 데이터를 학습한 딥러닝 모델에 대하여 trisomy 13 판정의 정확도를 확인한 결과이며, (B)는 각 데이터 세트별 확률분포를 나타낸 결과이다.
- [32] 도 8의 윗 패널은 본 발명의 일 실시예에 따라, 핵산단편 사이의 거리를 기반으로 생성한 GCW plot 이미지 데이터를 학습한 딥러닝 모델에 대하여 trisomy 21 판정의 정확도를 확인한 결과이며, 아래 패널은 각 데이터 세트별 확률분포를 나타낸 결과이다.
- [33] 도 9의 왼쪽 패널은 도 8에서 학습한 딥러닝 모델에서 18번 염색체의 이상을 검출한 결과의 정확도를 확인한 결과이며, 오른쪽 패널은 확률 분포를 나타낸 결과이다.
- [34] 도 10의 왼쪽 패널은 도 8에서 학습한 딥러닝 모델에서 13번 염색체의 이상을 검출한 결과의 정확도를 확인한 결과이며, 오른쪽 패널은 확률 분포를 나타낸 결과이다.
- [35] 도 11의 왼쪽 패널은 도 8에서 학습한 딥러닝 모델에서 21번 염색체의 이상을 검출하기 위해 down sampling을 시도한 결과의 정확도를 확인한 결과이며, 오른쪽 패널은 확률 분포를 나타낸 결과이다.
- [36] 도 12의 윗 패널은 본 발명의 일 실시예에 따라, 핵산단편 사이의 거리를 계산한 값의 역수값을 기반으로 생성한 GCW plot 이미지 데이터를 학습한 딥러닝 모델에 대하여 trisomy 21 판정의 정확도를 확인한 결과이며, 아래 패널은 각 데이터 세트별 확률분포를 나타낸 결과이다.
- [37] 도 13의 위쪽 패널은 도 12에서 학습한 딥러닝 모델에서 18번 염색체의 이상을 검출한 결과의 정확도를 확인한 결과이며, 아래쪽 패널은 확률 분포를 나타낸 결과이다.
- [38] 도 14의 위쪽 패널은 도 12에서 학습한 딥러닝 모델에서 13번 염색체의 이상을 검출한 결과의 정확도를 확인한 결과이며, 아래쪽 패널은 확률 분포를 나타낸 결과이다.

- [39]
- [40] **발명의 상세한 설명 및 바람직한 구현예**
- [41] 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.
- [42] 본 발명에서는, 샘플에서 획득한 서열 분석 데이터를 참조 유전체에 정렬한 다음, 정렬된 핵산단편을 기반으로 벡터화된 데이터를 생성한 다음, 학습된 인공지능 모델에서 DPI값을 계산하여 기준값과 비교하여 염색체 이상을 검출할 경우, 높은 민감도와 정확도로 염색체 이상을 검출할 수 있다는 것을 확인하고자 하였다.
- [43]
- [44] 즉, 본 발명의 일 실시예에서는, 혈액에서 추출한 DNA를 시퀀싱 한 뒤, 참조 염색체에 정렬한 다음, 핵산단편 사이의 거리 또는 양을 일정한 염색체 구간 별로 계산하고, 각 유전영역을 X축으로 하고, 핵산단편 사이의 거리 또는 양을 Y축으로 하는 벡터화된 데이터를 생성한 다음, 이를 딥러닝 모델에 학습시켜 DPI 값을 계산하였으며, DPI 값이 기준값 이상일 경우, 염색체 이상이 있다고 결정하는 방법을 개발하였다(도 1)
- [45] 따라서, 본 발명은 일관점에서,
- [46] a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;
- [47] b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;
- [48] c) 상기 정렬된 핵산단편(fragments)을 기반으로 벡터화된 데이터를 생성하는 단계;
- [49] d) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석하고, 기준값(cut-off value)과 비교하여 염색체 이상 유무를 판정하는 단계를 포함하는 인공지능 기반 염색체 이상 검출방법에 관한 것이다.
- [50]
- [51] 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.
- [52] 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.
- [53] 본 발명에서 상기 핵산 단편은 차세대 염기서열 분석을 이용할 경우에는 리드를 의미할 수 있다.
- [54] 본 발명에서 용어 “염색체의 이상”은 염색체에서 발생하는 다양한 변이를

의미하는데, 크게 수 이상과 구조 이상, 미세결실, 염색체 불안정성 등으로 구분될 수 있다.

- [55] 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.
- [56] 염색체의 구조 이상은 결실, 중복 역위, 전좌 등 염색체의 개수에는 변화가 없으나 염색체의 구조에 변화가 발생하는 모든 경우를 의미한다. 예를 들어, 5번 염색체의 일부분이 결실(고양이울음 증후군), 7번 염색체의 일부분이 결실(필리암스 증후군), 12번 염색체의 일부분 중복(월프-허쉬호른 증후군), 9번과 22번 염색체 사이의 전좌(만성 골수 백혈병) 등이 있을 수 있고, 중앙 환자에서 발견되는 일부 염색체 영역의 미세중복과 미세결실도 포함될 수 있다. 그러나 위에 기술된 내용에 한정되는 것은 아니다.
- [57]
- [58] 본 발명에 있어서,
- [59] 상기 a) 단계는
- [60] (a-i) 혈액, 정액, 질 세포, 모발, 타액, 소변, 구강세포, 태반세포 또는 태아세포를 포함하는 양수, 조직세포 및 이의 혼합물에서 핵산을 수득하는 단계;
- [61] (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;
- [62] (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;
- [63] (a-iv) 제작된 라이브러리를 차세대 유전자서열검사기(next-generation sequencer)에 반응시키는 단계; 및
- [64] (a-v) 차세대 유전자서열검사기에서 핵산의 서열정보(reads)를 획득하는 단계;를 포함하는 것을 특징으로 할 수 있다.
- [65]
- [66] 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.

- [67] 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.
- [68] 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) 시스템을 포함한다.
- [69] NGS 테크놀로지는, 예를 들어 주형 제조, 시퀀싱 및 이미징 및 데이터 분석 단계 중 하나 이상을 포함할 수 있다.
- [70] 주형 제조. 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.

- [71] 클론이 증폭된 주형의 제조방법은, 예를 들어 에멀전 PCR(emulsion PCR: emPCR) 및 고체상 증폭을 포함한다.
- [72] 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.
- [73] 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.
- [74] 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 핵사머 프라이머를 어닐링하는 단계 및 일정한 온도에서 고충실도 효소, 전형적으로 $\Phi 29$ 에 의해 DNA를 합성하는 단계를 수반한다. MDA는 더 낮은 오류 빈도로 거대한 크기의 생성물을 만들 수 있다.
- [75] 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), 맞춤형-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.
- [76] 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.

- [77] 시퀀싱 및 이미징. NGS를 위한 대표적인 시퀀싱 및 이미징 방법은, 이하에 제한되는 것은 아니지만, 사이클릭 가역적 종결(cyclic reversible termination: CRT), 결찰에 의한 시퀀싱(sequencing by ligation: SBL), 단일-분자 첨가(파이로시퀀싱(pyrosequencing)) 및 실시간 시퀀싱을 포함한다.
- [78] 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.
- [79] SBL은 시퀀싱을 위해 DNA 리가제 및 1-염기-암호화된 프로브 또는 2-염기-암호화된 프로브 중 하나를 사용한다.
- [80] 전형적으로, 형광 표지된 프로브는 프라이밍된 주형에 인접한 상보적 서열에 혼성화된다. DNA 리가제는 프라이머에 염료-표지된 프로브를 결찰시키기 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.
- [81] 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.

- [82] 실시시간 시퀀싱은 DNA 합성 동안 염료-표지된 뉴클레오타이드의 연속적 포함을 이미징하는 단계를 수반한다. 대표적인 실시시간 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-퀀처 뉴클레오타이드를 사용하는 LI-COR 바이오사이언스(Biosciences) 플랫폼을 포함한다.
- [83] 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.
- [84] 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.
- [85] 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.
- [86] 폴로니 시퀀싱은 폴로니 증폭 및 다중 단일-염기-연장(FISSEQ)을 통해

시퀀싱에 따르는 것을 기반으로 한다. 폴로니 증폭은 폴리아크릴아마이드 필름 상에서 인시츄로 DNA를 증폭시키는 방법이다. 대표적인 폴로니 시퀀싱 방법은, 예를 들어 미국특허 출원 공개 제2007/0087362호에 기재된다.

- [87] 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.
- [88] 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 깎과 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.
- [89] 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 가이딩하기 위해 배치된다. 나노와이어-분자 센서를 사용하는 대표적인 시퀀싱 방법은 예를 들어 미국특허출원 공개 제2006/0275779호에 기재된다.
- [90] 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 어닐링 시키는 단계; 폴리메라제로 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.
- [91]
- [92] 데이터 분석 단계.
- [93] NGS 리드가 만들어진 후, 그것들은 공지된 기준 서열에 대해 정렬되거나 테노보 조립된다.
- [94] 예를 들어, 샘플(예를 들어, 중량 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.

- [95] 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.
- [96] 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.
- [97]
- [98] 본 발명에 있어서, 상기 b) 단계의 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.
- [99] 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.
- [100] 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 웹 상의 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 인코퍼레이티드(Bioinformatics Solutions Inc.))을 포함한다.

- [101] 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.

[102] 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)을 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.

[103]

[104] 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.

[105]

[106] 본 발명에 있어서, 상기 c) 단계의 벡터화된 데이터는 정렬된 핵산단편을 기반으로 생성할 수 있는 벡터화된 데이터이면 제한없이 이용가능하나 바람직하게는 그랜드 캐년 플롯(Grand Canyon plot, GC plot) 또는 지노믹 캐슬 월 플롯(Genomic Castle Wall plot, GCW plot)인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[107]

[108] 본 발명에서 벡터화된 데이터는 이에 한정되지는 않으나 바람직하게는 이미지화된 것을 특징으로 할 수 있다. 이미지는 기본적으로 픽셀로 구성되는데, 픽셀로 구성된 이미지를 벡터화 시키면, 이미지의 종류에 따라서 1차원 2D 벡터(흑백), 3차원 2D 벡터(color(RGB)) 또는 4차원 2D 벡터(color(CMYK))로 표현될 수 있다.

[109] 본 발명의 벡터화된 데이터는 이미지에 한정되지 않고, 예를 들어 n개의 흑백 이미지 여러장으로 쌓아서 n차원의 2D 벡터(Multi-dimensional Vector)를

이용하여 인공지능 모델의 입력 데이터로 사용할 수 있다.

- [110] 본 발명에서 GCW plot은 정상 염색체와 판별 염색체의 구간 별 차이점을 명확히 나타내기 위하여 정상 염색체와 판별 염색체를 교차 배열하여 작성한 그래프를 의미하여, 예를 들어 염색체별 RepFD 값을 이미지화 할 경우, 정상 샘플에서는 모든 염색체에서 RepFD값이 비슷하게 나오기 때문에 GCW plot에서 Up-Down 패턴이 생기지 않으나, 염색체 이수성 샘플에서는 이수성이 있는 염색체에서 정상 염색체들보다 상대적으로 낮은 RepFD 값이 나오기 때문에 GCW plot에서 명확한 Up-Down 패턴이 생성된다. 이러한 패턴의 차이를 극대화하기 위해 (정상염색체, 이수성 염색체) n 를 반복, 교차 배열하여 인공지능 모델에 학습시켜 정상 샘플과 염색체 이수성 샘플을 구분할 수 있게 하는 것이다.
- [111] 본 발명에서 GC plot은 특정 구간을 (일정한 bin 또는 크기가 다른 bin) X축으로 두고 핵산단편 사이의 거리 또는 수와 같은 핵산단편으로 표현 할 수 있는 수치를 Y축으로 생성한 plot이다.
- [112]
- [113] 본 발명에서, 상기 c) 단계를 수행하기에 앞서 정렬된 핵산단편의 정렬 일치도 점수(mapping quality score)를 만족하는 핵산단편을 따로 분류하는 단계를 추가로 포함하는 것을 특징으로 할 수 있다.
- [114] 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.
- [115]
- [116] 본 발명에 있어서, 상기 c) 단계의 GC plot은 정렬된 핵산단편의 염색체 구간 별 분포를 구간 별 핵산단편의 수 또는 핵산단편 사이의 거리를 계산하여 벡터화된 데이터로 생성하는 것을 특징으로 할 수 있다.
- [117] 본 발명에서 핵산단편의 수 또는 핵산단편 사이의 거리 계산값을 벡터화하는 방법은 계산값을 벡터화하는 공지된 기술이면 제한없이 이용할 수 있다.
- [118] 본 발명에 있어서, 상기 정렬된 서열정보의 염색체 구간 별 분포를 핵산단편의 수로 계산하는 것은 하기의 단계를 포함하여 수행하는 것을 특징으로 할 수 있다:
 - [119] i) 염색체를 일정구간(bin)으로 구분하는 단계;
 - [120] ii) 각 구간에 정렬된 핵산단편의 수를 결정하는 단계;
 - [121] iii) 각 구간에 결정된 핵산단편 수를 샘플의 전체 핵산단편 수로 나누어 정규화(normalization)하는 단계; 및
 - [122] iv) 각 구간의 순서를 X 축 값으로 하고, 상기 iii) 단계에서 계산한 정규화 값을 Y축 값으로 하여 GC plot을 생성하는 단계.
- [123]
- [124] 본 발명에 있어서, 상기 정렬된 서열정보의 염색체 구간 별 분포를 핵산단편

사이의 거리로 계산하는 것은 하기의 단계를 포함하여 수행하는 것을 특징으로 할 수 있다:

- [125] i) 염색체를 일정구간(bin)으로 구분하는 단계;
- [126] ii) 각 구간에 정렬된 핵산단편 사이의 거리(Fragments Distance, FD)를 계산하는 단계;
- [127] iii) 각 구간별로 계산된 거리값을 기반으로 각 구간의 거리의 대표값(RepFD)을 결정하는 단계;
- [128] iv) 상기 iii) 단계에서 계산된 대표값을 전체 핵산단편 거리 값의 대표값으로 나누어 정규화(normalization)하는 단계; 및
- [129] iv) 각 구간의 순서를 X 축 값으로 하고, 상기 iv) 단계에서 계산한 정규화 값을 Y축 값으로 하여 GC plot을 생성하는 단계.
- [130] 본 발명에 있어서, 상기 대표값(RepFD)은 FD값의 합, 차, 곱, 평균, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대편차, 변동계수 및 이들의 조합으로 구성된 군에서 선택되는 하나 이상인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.
- [131] 본 발명에 있어서, 상기 일정구간(bin)은 1Kb 내지 3Gb인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.
- [132] 본 발명에서는 핵산단편을 그룹화하는 단계를 추가로 사용할 수 있으며, 이때 그룹화 기준은, 정렬된 핵산단편의 어댑터 서열을 바탕으로 수행할 수 있다. 정방향으로 정렬된 핵산단편과 역방향으로 정렬된 핵산단편으로 별도로 구분하여서 선별된 서열정보에 대해서 핵산단편 사이의 거리를 계산할 수 있다.
- [133] 본 발명에서, 상기 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.
- [134] 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.
- [135] 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.
- [136] 본 발명에서, 상기 “핵산 단편의 기준값”은 핵산 단편의 중앙값으로부터 임의의 값을 더하거나 빼 값인 것인 것을 특징으로 할 수 있다.
- [137]
- [138] 상기 FD 값은 수득한 n개의 핵산 단편에 대하여, 다음과 같이 정의 할 수 있다.

- [139] $FD = \text{Dist}(R_i \sim R_j) (1 < i < j < n)$,
- [140] 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.
- [141] 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, i 가 1일 경우, $i+1$ 은 2가 되어 2 내지 n 번째 핵산 단편에서 선택되는 어느 하나 이상의 핵산 단편과의 거리를 정의 할 수 있다.
- [142] 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.
- [143] 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.
- [144] 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.
- [145] 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
- [146]
- [147] 바람직하게는 본 발명에 있어서, 상기 FD 값은 핵산 단편의 중앙값으로부터 임의의 값을 더하거나 빼 값인 것을 특징으로 할 수 있다.
- [148] 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.

- [149] 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.
- [150] 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.
- [151] 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.
- [152] 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.
- [153]
- [154] 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.
- [155] 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.
- [156]
- [157] 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 일정한 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.

- [158] 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.
- [159] 또한 정방향, 역방향의 양 방향으로 시퀀싱 된(정렬된), 모든 리드의 말단 정보를 이용해 핵산 분자의 거리 계산시, 시퀀싱 방향이라는 요소 때문에 정확하지 않은 값이 계산 될 수 있다.
- [160] 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.
- [161] 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.
- [162] 본 발명에 있어서, 상기 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.
- [163]
- [164] 본 발명에 있어서, 상기 GCW plot은 정렬된 핵산단편 사이의 거리 또는 구간 별 양을 염색체 단위로 계산하고, 정상 염색체와 이수성을 판별하고자 하는 염색체의 거리 값을 교차 배열하여 벡터화된 데이터로 생성하는 것을 특징으로

할 수 있다.

- [165] 본 발명에서 상기 교차배열은, x 축을 따라 ([정상 염색체의 계산값], [이수성을 판별할 염색체의 계산값])_n 값을 배열하는 것을 의미한다. 예를 들어, Trisomy 21 판별을 할 경우, x 축을 따라 1번 염색체의 RepFD 값, 21번 염색체의 RepFD, 2번 염색체의 RepFD, 21번 염색체의 RepFD, 3번 염색체의 RepFD, 21번 염색체의 RepFD ... 이런식으로 배열하는 방식을 의미한다.
- [166] 본 발명에 있어서, 상기 GCW plot은 핵산단편 사이의 거리를 계산값으로 사용할 경우, 하기의 단계를 포함하는 방법으로 수행되는 것을 특징으로 할 수 있다.
- [167] i) 각 염색체 별로 정렬된 핵산단편 사이의 거리(FD)를 계산하는 단계;
- [168] ii) 상기 i) 단계에서 계산된 거리값을 기반으로 각 구간의 거리의 대표값(RepFD)을 결정하는 단계;
- [169] iii) 상기 ii) 단계에서 결정된 염색체별 RepFD 값을 표준화(standardization) 하는 단계;
- [170] iv) 이수성을 판별할 염색체를 제외한 염색체들 중 일부를 비교염색체군으로 선발하는 단계; 및
- [171] iv) 비교염색체군에 속하는 염색체들과 이수성을 판별할 염색체를 X축을 따라 순차적으로 교차 배열하고, 각 염색체의 표준화된 계산값을 Y값으로 하는 GCW plot을 생성하는 단계.
- [172]
- [173] 본 발명에 있어서, 상기 표준화 단계는
- [174] 1) 염색체 이수성이 검출되지 않은 정상 샘플로 구성된 표준 집단을 설정하는 단계;
- [175] 2) 상기 표준 집단에서 관찰되는 염색체별 RepFD 값의 평균(Reference_Mean)과 표준편차(Reference_Standard_Deviation)를 계산하는 단계;
- [176] 3) 상기 2) 단계에서 계산한 평균과 표준편차를 하기 수식 1에 대입하여 Z 표준화하는 단계;
- [177] 수식 1:
- $$Z_{chr} = ((RepFD_{chr} - Reference_Mean_{chr}) / Reference_Standard_Deviation_{chr}) + 5$$
- [178] 여기서, 시퀀싱 양(리드 수)이 많은 샘플에서는 RepFD 값이 낮아지면서 Z_{chr} 값이 음수를 갖는 경우가 생길 수 있는데, Z_{chr} 값이 음수가 되면 하기 단계의 이수성을 판별하고자 하는 염색체의 Z_{chr} 값으로 나누어주는 과정에서 문제가 발생하므로(Up-Down 패턴이 Down-Up 패턴으로 뒤집어서 나올 수 있음), 전체적으로 +5를 해서 모두 양수 값으로 만들어 주기 위해 +5를 더해준다.
- [179] 4) 상기 3) 단계에서 계산한 염색체별 Z_{chr} 값을 이수성을 판별하고자 하는 염색체의 Z_{chr} 값으로 나누는 단계.

- [180] 여기서, 상기 4) 단계는 샘플의 시퀀싱 양에 따라 GCW plot의 y-scale이 변하는 것을 보정하기 위해, 이수성을 판별하고자 하는 염색체의 Z_{chr} 값으로 나누는 것으로, 이 과정을 통해 샘플의 시퀀싱 양에 관계 없이 y-scale이 일정한 GCW plot을 생성할 수 있다.
- [181] 본 발명에 있어서, 상기 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.
- [182]
- [183] 본 발명에서, 상기 GCW plot을 핵산단편의 양으로 나타낸 경우, 핵산단편의 양으로 GCW plot을 그리면, 비교염색체군 막대의 상대적인 높이에서 차이를 보여 이를 인공지능 모델이 학습하도록 할 수 있다(도 4).
- [184]
- [185] 본 발명에서, 상기 비교염색체 군은 판별염색체군을 제외한 어떤 염색체의 조합이든 이용 가능하며, 바람직하게는 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 13, 14, 15, 18, 20 번 염색체일 수 있으나, 이에 한정되는 것은 아니다.
- [186] 본 발명에서, 상기 벡터화된 데이터는 하나의 이미지에 다수의 염색체 별 plot을 포함하는 것을 특징으로 할 수 있다.
- [187]
- [188] 본 발명에 있어서, 상기 (d) 단계의 인공지능 모델은 염색체 상태가 정상인 이미지와 염색체 이상이 있는 이미지를 구별할 수 있도록 학습할 수 있는 모델이면 제한없이 사용가능하며, 바람직하게는 딥러닝 모델인 것을 특징으로 할 수 있다.
- [189]
- [190] 본 발명에 있어서, 상기 인공지능 모델은 인공신경망 기반으로 벡터화된 데이터를 분석할 수 있는 인공신경망 알고리즘이면 제한없이 이용할 수 있으나, 바람직하게는 합성곱 신경망(convolutional neural network, CNN), 심층 신경망(Deep Neural Network, DNN), 순환 신경망(Recurrent Neural Network, RNN) 및 오토 인코더(autoencoder)로 구성된 군에서 선택되는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.
- [191]
- [192] 본 발명에 있어서, 상기 순환 신경망은 LSTM(Long-short term memory) 신경망, GRU(Gated Recurrent Unit) 신경망, 바닐라 순환 신경망(Vanilla recurrent neural network) 및 집중적 순환 신경망(attentive recurrent neural network)으로 구성된 군에서 선택되는 것을 특징으로 할 수 있다.
- [193]
- [194] 본 발명에 있어서, 상기 인공지능 모델이 CNN일 경우, binary classification을

수행하는 손실함수는 하기 수식 2로 표시되는 것을 특징으로 할 수 있고,
Multi-class classification을 수행하는 손실함수는 하기 수식 3으로 표시되는 것을
특징으로 할 수 있다.

[195] 수식 2: Binary classification

$$\text{loss}(\text{model}(x), y) = -\frac{1}{n} \left[\sum_{i=1}^n (y_i \log(\text{model}(x_i)) + (1 - y_i) \log(1 - \text{model}(x_i))) \right]$$

$\text{model}(x_i) = i$ 번째 input에 인공지능 model output

$y =$ 실제 label 값

$n =$ input data 수

[196] 수식 3: Multi-class classification

$$\text{loss}(\text{model}(x), y) = -\frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^c (y_{ij} \log(\text{model}(x_i)_j)) \right)$$

$\text{model}(x_i)_j = i$ 번째 input에 j 번째 인공지능 model output

$y =$ 실제 label 값

$n =$ input data 수

$c =$ class 수

[197] 본 발명에서, 상기 binary classification은 인공지능 모델이 염색체 이상 한
종류를 판별하도록 학습하는 것을 의미하며, multi-class classification은 인공지능
모델이 두 가지 이상의 염색체 이상을 판별하도록 학습하는 것을 의미한다.

[198]

[199] 본 발명에서, 상기 인공지능 모델이 CNN일 경우, 학습은 하기 단계를 포함하여
수행되는 것을 특징으로 할 수 있다:

[200] i) 생산된 GC, GCW plot을 training(학습), validation(검증), test(성능평가)
데이터로 분류하는 단계;

[201] 이 때, Training 데이터는 CNN 모델을 학습할 때 사용되고, Validation 데이터는
hyper-parameter tuning 검증에 사용되며, Test 데이터는 최적의 모델 생산 후, 성능
평가로 사용되는 것을 특징으로 함.

[202] ii) Hyper-parameter tuning 및 학습 과정을 통해서 최적의 CNN 모델을 구축하는
단계;

[203] iii) Hyper-parameter tuning을 통해서 얻어진 여러 모델의 성능을 validation
data를 이용하여 비교하여, validation data 성능이 가장 좋은 모델을 최적의

모델로 결정하는 단계;

[204]

[205] 본 발명에서, 상기 Hyper-parameter tuning 과정은 CNN 모델을 이루는 여러 parameter(convolution layer 수, dense layer 수, convolution filter 수 등) 값을 최적화하는 과정으로 Hyper-parameter tuning 과정으로는 Bayesian optimization 및 grid search 기법을 사용하는 것을 특징으로 할 수 있다.

[206] 본 발명에서, 상기 학습 과정은 정해진 hyper-parameter들을 이용하여 CNN 모델의 내부 parameter(weights)들을 최적화 시켜, Training loss 대비 validation loss가 증가하기 시작하면 모델이 과적합(Overfitting) 되었다 판단하고, 그전에 model 학습을 중단하는 것을 특징으로 할 수 있다.

[207]

[208] 본 발명에 있어서, 상기 d) 단계에서 인공지능 모델이 입력된 벡터화된 데이터로부터 분석한 결과값은 (특정 score 또는 실수)이면 제한없이 이용가능하며, 바람직하게는 DPI(Deep Probability Index) 값인 것을 특징으로 할 수 있으나 이에 한정되는 것은 아니다.

[209]

[210] 본 발명에서, Deep probability Index는 인공지능 model의 마지막 layer에 binary classification일 경우 sigmoid function, multi-class classification일 경우 softmax function을 사용하여 인공지능의 output을 0 ~ 1 scale로 조정하여 확률값으로 표현한 값을 의미한다.

[211] 본 발명에 있어서, 상기 d) 단계의 기준값은 0.5이며, 인공지능 모델에서 분석하여 도출한 값이 0.5 이상일 경우, 염색체 이상이 있는 것으로 판정하는 것을 특징으로 할 수 있다.

[212] 본 발명에서, 상기 인공지능 모델은 학습할 때, 이수성이 있으면 output 결과가 1에 가깝게 학습하고, 이수성이 없으면 output 결과가 0에 가깝게 학습을 시켜서, 0.5를 기준으로 0.5 이상이면 이수성이 있다고 판단하고, 0.5 이하이면 이수성이 없다고 판단하여 performance 측정을 수행하였다(Training, validation, test accuracy).

[213] 여기서, 0.5의 기준값은 언제든지 바뀔수 있는 값이라는 것은 통상의 기술자에게 자명한 것이다. 예를 들어서 False positive(위양성)를 줄이고자 하면, 0.5보다 높은 기준값을 설정하여 이수성이 있다고 판단되는 기준을 엄격하게 가져 갈 수 있고, False Negative(위음성)를 줄이고자 하면 기준값을 더 낮게 측정하여 이수성이 있다고 판단되는 기준을 조금 더 약하게 가져 갈 수 있다.

[214] 가장 바람직하게는 학습된 인공지능 모델을 이용하여 unseen data(학습에 training하지 않은 답을 알고 있는 data)를 적용시켜서, DPI값의 probability를 확인하여 기준값을 정할 수 있다.

[215]

[216] 본 발명은 다른 관점에서, 생체시료에서 핵산을 추출하여 서열정보를 해독하는

- 해독부;
- [217] 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부;
- [218] 정렬된 핵산단편(fragments)를 기반으로 벡터화된 데이터를 생성하는 데이터 생성부;
- [219] 생성된 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석하고, 기준값과 비교하여 염색체 이상 유무를 판정하는 염색체 이상 판정부를 포함하는 인공지능 기반 염색체 이상 검출 장치에 관한 것이다.
- [220]
- [221] 본 발명은 또 다른 관점에서, 컴퓨터 판독 가능한 저장 매체로서, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,
- [222] a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;
- [223] b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;
- [224] c) 상기 정렬된 핵산단편(fragments)을 기반으로 벡터화된 데이터를 생성하는 단계;
- [225] d) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석하고, 기준값(cut-off value)을 비교하여 염색체 이상 유무를 판정하는 단계를 통하여, 염색체 이상을 검출하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체에 관한 것이다.
- [226]
- [227] 실시예
- [228] 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.
- [229]
- [230] 실시예 1. 혈액에서 DNA를 추출하여, 차세대 염기서열 분석 수행
- [231] 정상인 7,931명, Trisomy 21 170명, Trisomy 18 58명 및 Trisomy 13 16명의 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. 그 결과, 샘플 당 약 13 million 개의 reads가 생산되는 것을 확인하였다.
- [232]
- [233] 실시예 2. 핵산단편 거리 기반 GC plot을 이용한 trisomy 21 딥러닝 모델 구축 및

성능 확인

- [234] 정상 샘플 (n=961)과 T21 샘플 (n=170) 을 이용해 DPI 값의 성능을 테스트 하였다. 모든 샘플은 Train, Validation, Test 그룹으로 나눠 진행했고, Train 샘플을 이용하여 model을 구축한 다음 Validation 그룹 및 Test 그룹의 샘플을 이용해서, Train 샘플을 이용해 만든 모델의 성능을 확인하였다.

- [235] [표1]

	Normal	T21	Total	Baseline accuracy
Train	589	93	682	86.4%
Validation	156	34	190	82.1%
Test	216	43	259	83.4%
Total	961	170	1131	

- [236] 그 결과, 표 2 및 도 5에 기재된 바와 같이, Accuracy 는 Train, Valid, Test 그룹에서 각각 100%, 100%, 100% 로 확인 되었고, ROC 분석 결과인 AUC 값은 Train, Valid, Test 그룹에서 각각 1.0, 1.0, 1.0인 것을 확인하였다.

- [237] [표2]

	Accuracy	AUC
Train	100%	1.0
Validation	100%	1.0
Test	100%	1.0

- [238] 도 5의 (A)는 정확도를 측정하는 방법 중 ROC(Receiver Operating Characteristic) curve 를 활용한 분석으로, 커브 아래 면적의 넓이인 AUC(Area Under the Curve) 값이 높을 수록 정확도가 높다고 해석한다. AUC 값은 0-1 사이의 값을 갖으며, 랜덤으로 레이블 값을 예측했을 때 (baseline) 기대되는 AUC 값이 0.5, 완벽히 정확하게 예측했을 때 기대되는 AUC 값이 1이다.

- [239] 도 5의 (B)는 본 발명의 인공지능 모델에서 계산된 염색체 이수성이 있을 확률값(DPI 값을 정상 샘플과 trisomy 21 샘플 그룹에서 boxplot으로 나타낸 것으로, 붉은 선이 DPI cutoff인 0.5를 나타낸다.

- [240]

- [241] 실시예 3. 핵산단편 거리 기반 GC plot을 이용한 trisomy 18 딥러닝 모델 구축 및 성능 확인

- [242] 정상 샘플 (n=961)과 T18 샘플 (n=58) 을 이용해 DPI 값의 성능을 테스트 하였다. 모든 샘플은 Train, Validation, Test 그룹으로 나눠 진행했고, Train 샘플을 이용하여 model을 구축한 다음 Validation 그룹 및 Test 그룹의 샘플을 이용해서, Train 샘플을 이용해 만든 모델의 성능을 확인하였다.

[243] [표3]

	Normal	T18	Total	Baseline accuracy
Train	589	28	617	95.5%
Validation	156	18	174	89.7%
Test	216	12	228	94.7%
Total	961	58	1019	

[244] 그 결과, 표 4 및 도 6에 기재된 바와 같이, Accuracy 는 Train, Valid, Test 그룹에서 각각 100%, 100%, 100% 로 확인 되었고, ROC 분석 결과인 AUC 값은 Train, Valid, Test 그룹에서 각각 1.0, 1.0, 1.0인 것을 확인하였다.

[245] [표4]

	Accuracy	AUC
Train	100%	1.0
Validation	100%	1.0
Test	99.6%	1.0

[246]

[247] 실시예 4. 핵산단편 거리 기반 GC plot을 이용한 trisomy 13 딥러닝 모델 구축 및 성능 확인

[248] 정상 샘플 (n=961)과 T13 샘플 (n=16) 을 이용해 DPI 값의 성능을 테스트 하였다. 모든 샘플은 Train, Validation, Test 그룹으로 나눠 진행했고, Train 샘플을 이용하여 model을 구축한 다음 Validation 그룹 및 Test 그룹의 샘플을 이용해서, Train 샘플을 이용해 만든 모델의 성능을 확인하였다.

[249] [표5]

	Normal	T13	Total	Baseline accuracy
Train	589	7	596	98.8%
Validation	156	5	161	96.9%
Test	216	4	220	98.2%
Total	961	16	977	

[250] 그 결과, 표 6 및 도 7에 기재된 바와 같이, Accuracy 는 Train, Valid, Test 그룹에서 각각 100%, 100%, 100% 로 확인 되었고, ROC 분석 결과인 AUC 값은 Train, Valid, Test 그룹에서 각각 1.0, 1.0, 1.0인 것을 확인하였다.

[251] [표6]

	Accuracy	AUC
Train	100%	1.0
Validation	100%	1.0
Test	99.5%	1.0

[252]

[253] 실시예 5. 핵산단편 거리 기반 GCW plot을 이용한 trisomy 21 딥러닝 모델 구축 및 성능 확인

[254] 정상 샘플 (n=716) 명을 참조집단으로 설정하고 Z점수를 계산한 뒤, 이 값을 이용해 GCW plot을 생산했다. Z점수 계산을 위한 핵산단편 거리값은 핵산단편 사이 거리값들의 중앙값을 이용했다. Z점수 계산을 위한 참조집단에 포함되지 않는 정상 샘플 (n=7,215)과 T21 샘플 (n=168) 을 이용해 DPI 값의 성능을 테스트 하였다. 모든 샘플은 Train, Validation, Test 그룹으로 나뉘 진행했고, Train 샘플을 이용하여 model을 구축한 다음 Validation 그룹 및 Test 그룹의 샘플을 이용해서, Train 샘플을 이용해 만든 모델의 성능을 확인하였다.

[255] [표7]

	Normal	T21	Total	Baseline accuracy
Train	4,362	92	4,454	97.9%
Validation	1,110	34	1,144	97.0%
Test	1,743	42	1,785	97.7%
Total	7,215	168	7,383	97.7%

[256] 그 결과, 표 8 및 도 8에 기재된 바와 같이, Accuracy 는 Train, Valid, Test 그룹에서 각각 99.8%, 100%, 99.8% 로 확인 되었고, ROC 분석 결과인 AUC 값은 Train, Valid, Test 그룹에서 각각 0.9999, 1.0, 0.9998인 것을 확인하였다.

[257] [표8]

	Accuracy	AUC
Train	99.8%	0.9999
Validation	100%	1.0
Test	99.8%	0.9998

[258]

[259] 실시예 6. 핵산단편 거리 기반 GCW plot을 이용한 trisomy 18 딥러닝 모델 구축 및 성능 확인

[260] 정상 샘플 (n=716) 명을 참조집단으로 설정하고 Z점수를 계산한 뒤, 이 값을 이용해 GCW plot을 생산했다. Z점수 계산을 위한 핵산단편 거리값은 핵산단편 사이 거리값들의 중앙값을 이용했다. 실시예 5에서 구축한 딥러닝 모델을

사용하여 18번 염색체의 이수성을 예측하였다. Z점수 계산을 위한 참조집단에 포함되지 않는 정상 샘플 (n=1743)과 T18 샘플 (n=58)을 이용해 DPI 값의 성능을 테스트 하였다.

[261] [표9]

	Normal	T18	Total	Baseline accuracy
Test	1,743	58	1,801	96.7%
Total	1,743	58	1,801	96.7%

[262] 그 결과, 표 10 및 도 9에 기재된 바와 같이, Test 그룹에서 Accuracy 99.8% 이며, ROC 분석 결과인 AUC 값은 0.9994 인 것을 확인하였다.

[263] [표10]

	Accuracy	AUC
Test	99.8%	0.9994

[264]

[265] 실시예 7. 핵산단편 거리 기반 GCW plot을 이용한 trisomy 13 딥러닝 모델 구축 및 성능 확인

[266] 정상 샘플 (n=716) 명을 참조집단으로 설정하고 Z점수를 계산한 뒤, 이 값을 이용해 GCW plot을 생산했다. Z점수 계산을 위한 핵산단편 거리값은 핵산단편 사이 거리값들의 중앙값을 이용했다. 실시예 5에서 구축한 딥러닝 모델을 사용하여 13번 염색체의 이수성을 예측하였다. Z점수 계산을 위한 참조집단에 포함되지 않는 정상 샘플 (n=1,743)과 T13 샘플 (n=16)을 이용해 DPI 값의 성능을 테스트 하였다.

[267] [표11]

	Normal	T13	Total	Baseline accuracy
Test	1,743	16	1,759	99.9%
Total	1,743	16	1,759	99.9%

[268] 그 결과, 표 12 및 도 10에 기재된 바와 같이, Test 그룹에서 Accuracy 99.7% 이며, ROC 분석 결과인 AUC 값은 0.9943 인 것을 확인하였다.

[269] [표12]

	Accuracy	AUC
Test	99.7%	0.9943

[270]

[271] 실시예 8. 핵산단편 거리 기반 GCW plot을 이용한 Down sampling 성능확인

[272] 리드 커버리지를 절반 수준으로 낮추고 (약 10,000,000 리드 이상 -> 5,000,000 리드) GCW plot을 그려 염색체 이수성 검출을 시도한 결과, 낮추기 전과 유사한

수준의 성능을 유지하는 것을 확인하였다. 커버리지를 낮추었을 때의 Trisomy 21 판별 성능은 하기 표 13과 같다.

[273] [표13]

	Normal	T21	Total	Baseline accuracy
Test	158	49	207	76.3%
Total	158	49	207	76.3%

[274] 분석 결과, Test 그룹에서 Accuracy 98.6% 로 확인 되었고, ROC 분석 결과인 AUC 값은 0.998인 것을 확인하였다(표 14, 도 11).

[275] [표14]

	Accuracy	AUC
Test	98.6%	0.9981

[276]

[277] 실시예 9. 핵산단편 거리 역수값 기반 GCW plot을 이용한 trisomy 21 딥러닝 모델 구축 및 성능 확인

[278] 정상 샘플 (n=700) 명을 참조집단으로 설정하고 Z점수를 계산한 뒤, 이 값을 이용해 GCW plot을 생산했다. Z점수 계산을 위한 핵산단편 거리값은 핵산단편 사이 거리값들의 중앙값의 역수를 이용했다. Z점수 계산을 위한 참조집단에 포함되지 않는 정상 샘플 (n=991)과 T21 샘플 (n=163) 을 이용해 DPI 값의 성능을 테스트 하였다. 모든 샘플은 Train, Validation, Test 그룹으로 나눠 진행했고, Train 샘플을 이용하여 model을 구축한 다음 Validation 그룹 및 Test 그룹의 샘플을 이용해서, Train 샘플을 이용해 만든 모델의 성능을 확인하였다.

[279] [표15]

	Normal	T21	Total	Baseline accuracy
Train	485	79	564	84.5%
Validation	208	35	243	85.6%
Test	298	49	347	85.9%
Total	991	163	1154	85.9%

[280] 그 결과, 표 16 및 도 12에 기재된 바와 같이, Accuracy 는 Train, Valid, Test 그룹에서 각각 99.1%, 99.5%, 99.1% 로 확인 되었고, ROC 분석 결과인 AUC 값은 Train, Valid, Test 그룹에서 각각 0.9999, 0.999, 0.999인 것을 확인하였다.

[281] [표16]

	Accuracy	AUC
Train	99.1%	0.999
Validation	99.5%	0.999
Test	99.1%	0.999

[282]

[283] 실시예 10. 핵산단편 거리 역수값 기반 GCW plot을 이용한 trisomy 18 딥러닝 모델 구축 및 성능 확인

[284] 정상 샘플 (n=700) 명을 참조집단으로 설정하고 Z점수를 계산한 뒤, 이 값을 이용해 GCW plot을 생산했다. Z점수 계산을 위한 핵산단편 거리값은 핵산단편 사이 거리값들의 중앙값의 역수를 이용했다. Z점수 계산을 위한 참조집단에 포함되지 않는 정상 샘플 (n=991)과 T18 샘플 (n=57) 을 이용해 DPI 값의 성능을 테스트 하였다. 모든 샘플은 Train, Validation, Test 그룹으로 나눠 진행했고, Train 샘플을 이용하여 model을 구축한 다음 Validation 그룹 및 Test 그룹의 샘플을 이용해서, Train 샘플을 이용해 만든 모델의 성능을 확인하였다.

[285] [표17]

	Normal	T18	Total	Baseline accuracy
Train	485	27	512	94.7%
Validation	208	12	220	94.5%
Test	298	18	316	94.3%
Total	991	57	1048	94.6%

[286] 그 결과, 표 17 및 도 13에 기재된 바와 같이, Accuracy 는 Train, Valid, Test 그룹에서 각각 100%, 100%, 99.6% 로 확인 되었고, ROC 분석 결과인 AUC 값은 Train, Valid, Test 그룹에서 각각 1.0, 1.0, 0.972인 것을 확인하였다.

[287] [표18]

	Accuracy	AUC
Train	100%	1
Validation	100%	1
Test	99.6%	0.972

[288]

[289] 실시예 11. 핵산단편 거리 역수값 기반 GCW plot을 이용한 trisomy 13 딥러닝 모델 구축 및 성능 확인

[290] 정상 샘플 (n=700) 명을 참조집단으로 설정하고 Z점수를 계산한 뒤, 이 값을 이용해 GCW plot을 생산했다. Z점수 계산을 위한 핵산단편 거리값은 핵산단편 사이 거리값들의 중앙값의 역수를 이용했다. Z점수 계산을 위한 참조집단에 포함되지 않는 정상 샘플 (n=991)과 T13 샘플 (n=16) 을 이용해 DPI 값의 성능을

테스트 하였다. 모든 샘플은 Train, Validation, Test 그룹으로 나눠 진행했고, Train 샘플을 이용하여 model을 구축한 다음 Validation 그룹 및 Test 그룹의 샘플을 이용해서, Train 샘플을 이용해 만든 모델의 성능을 확인하였다.

[291] [표19]

	Normal	T13	Total	Baseline accuracy
Train	485	7	492	98.6%
Validation	208	4	212	98.1%
Test	298	5	303	98.3%
Total	991	16	1007	98.4%

[292] 그 결과, 표 20 및 도 13에 기재된 바와 같이, Accuracy 는 Train, Valid, Test 그룹에서 각각 98.7%, 98.5%, 99.0% 로 확인 되었고, ROC 분석 결과인 AUC 값은 Train, Valid, Test 그룹에서 각각 0.996, 0.995, 1.0인 것을 확인하였다.

[293] [표20]

	Accuracy	AUC
Train	98.7%	0.996
Validation	98.5%	0.995
Test	99.0%	1.0

[294]

[295] 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.

[296]

산업상 이용가능성

[297] 본 발명에 따른 인공지능 기반 염색체 이상 검출방법은, 기존의 리드 개수(read count) 기반으로 염색체 양을 결정하는 단계를 이용하는 방식 또는 정렬된 리드(reads) 사이의 거리 개념을 이용하는 검출 방법등에서 리드와 관련된 값을 하나하나의 정형화된 값으로 활용하는 데 비하여 벡터화된 데이터를 생성하여 AI 알고리즘을 이용하여 분석하기 때문에, 리드 커버리지가 낮더라도 유사한 효과를 발휘할 수 있어 유용하다.