

명세서

발명의 명칭: 세포유리 핵산단편 말단 서열 모티프 빈도 및 크기를 이용한 암 진단 및 암 종 예측방법

기술분야

- [1] 본 발명은 세포유리 핵산단편 말단 서열 모티프 빈도 및 크기를 이용한 암 진단 및 암 종 예측방법에 관한 것으로, 보다 구체적으로는 생체시료에서 핵산을 추출하여, 서열정보를 획득하여 정렬한 리드를 기반으로 핵산단편의 말단 서열 모티프 빈도와 핵산단편의 크기를 도출한 다음, 이를 벡터화된 데이터로 생성한 후, 학습된 인공지능 모델에 입력하여 계산된 값을 분석하는 방법을 이용한 암 진단 및 암 종 예측방법에 관한 것이다.

[2]

배경기술

- [3] 임상에서의 암 진단은 통상적으로 병력 조사, 물리적 검사 및 임상적 평가 후 조직 생검(tissue biopsy)을 수행 하여 확인하고 있다. 임상 실험에 의한 암 진단은 암 세포의 수가 10억 개 이상이고 암의 직경이 1cm 이상일 경우에만 가능하다. 이 경우, 암 세포는 이미 전이능력을 가지고 있으며, 적어도 이들 중 반은 이미 전이된 상태이다. 또한, 조직생검은 침습적이어서 환자에게 상당한 불편함을 주고, 암 환자를 치료하다 보면 조직생검을 수행할 수 없는 경우도 자주 있다는 문제점이 있다. 이외에, 암 스크리닝에 있어서 암으로부터 직접 또는 간접적으로 생산되는 물질을 모니터링하기 위한 종양 마커가 사용되고 있지만, 암이 존재하는 경우에도 종양 마커 스크리닝 결과 반 이상이 정상으로 나타나고, 암이 없는 경우에도 자주 양성으로 나타나기 때문에, 그 정확성에 한계가 있다.
- [4] 이와 같은 통상적인 암 진단 방법의 문제점을 보완할 만한 비교적 간편하고 비침습적이며 높은 민감도 및 특이도를 가진 암 진단 방법의 요구에 따라, 최근 암의 진단, 추적 검사로 환자의 체액을 활용하는 액상생검(liquid biopsy)이 많이 이용되고 있다. 액상생검은 비침습적(non-invasive)인 방법으로, 기존의 침습적인 진단 및 검사방법의 대안으로 주목 받고 있는 진단기술이다.
- [5] 최근에는 액상생검에서 획득한 세포 유리 DNA (cell free DNA)을 이용하여 암 진단 및 암 종 감별을 수행하는 방법이 개발되고 있으며(US 10975431, Zhou, Xionghui et al., bioRxiv, 2020.07.16.201350), 특히, 세포 유리 핵산 말단 서열의 모티프 빈도 정보를 분석하여 암 진단, 산전진단 또는 장기이식 모니터링에 이용하는 방법이 알려져 있다(WO 2020-125709, Peiyong Jiang et al., cancer discovery, Vol. 10, 2020, pp. 664-673).
- [6] 한편, 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.

- [7] 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).
- [8] 이러한 인공신경망을 이용하여 바이오 분야에 활용하는 다양한 특허(KR KR 10-2018-0124550, KR 10-2019-7038076, KR 10-2019-0003676, KR 10-2019-0001741)가 존재하고 있으나, 혈액 내 무세포 DNA(cell-free DNA, cfDNA)의 서열분석 정보를 기반으로 인공신경망 분석을 통해 암 종을 예측하는 방법에 대해서는 연구가 부족한 실정이다.
- [9] 이에, 본 발명자들은 상기 문제점들을 해결하고, 높은 민감도와 정확도의 인공지능 기반 암 진단 및 암 종 예측방법을 개발하기 위해 예의 노력한 결과, 무세포 핵산단편의 말단 서열 모티프와 핵산단편의 길이 정보를 기반으로 벡터화된 데이터를 생성하고, 이를 학습된 인공지능 모델로 분석할 경우, 높은 민감도와 정확도로 암 진단 및 암 종을 예측할 수 있다는 것을 확인하고, 본 발명을 완성하였다.
- [10]
- [11] **발명의 요약**
- [12] 본 발명의 목적은 세포유리 핵산단편 말단 서열 모티프 빈도 및 크기를 이용한 암 진단 및 암 종 예측방법을 제공하는 것이다.
- [13] 본 발명의 다른 목적은 세포유리 핵산단편 말단 서열 모티프 빈도 및 크기를 이용한 암 진단 및 암 종 예측장치를 제공하는 것이다.

- [14] 본 발명의 또 다른 목적은 상기 방법으로 암 진단 및 암 종을 예측하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공하는 것이다.
- [15] 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 이용하여 핵산단편(fragments)의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하는 단계; (d) 상기 도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 이용하여 벡터화된 데이터를 생성하는 단계; (e) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및 (f) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 포함하는 암 진단 및 암 종 예측을 위한 정보의 제공방법을 제공한다.
- [16] 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)를 이용하여 핵산단편(fragments)의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하는 단계; (d) 상기 도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 이용하여 벡터화된 데이터를 생성하는 단계; (e) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및 (f) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 포함하는 암 진단 및 암 종 예측 방법을 제공한다.
- [17] 본 발명은 또한, 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부; 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 정렬된 서열 기반의 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하는 핵산단편 분석부; 도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 이용한 벡터화된 데이터를 생성하는 데이터 생성부; 생성된 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석하고, 기준값과 비교하여 암 유무를 판정하는 암 진단부; 및 출력된 결과값을 분석하여 암 종을 예측하는 암 종 예측부를 포함하는 암 진단 및 암 종 예측 장치를 제공한다.
- [18] 본 발명은 또한, 컴퓨터 판독 가능한 저장 매체로서, 암 진단 및 암 종을 예측하는 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 이용하여 핵산단편(fragments)의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하는 단계; (d) 상기 도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를

이용하여 벡터화된 데이터를 생성하는 단계; (e) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및 (f) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 통하여, 암 유무 및 암 종을 예측하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공한다.

[19]

도면의 간단한 설명

[20] 도 1은 본 발명의 세포유리 핵산단편 말단 서열 모티프 빈도 및 크기를 이용한 암 진단 및 암 종 예측방법을 수행하기 위한 전체 흐름도이다.

[21] 도 2는 본 발명의 일 실시예에서 건강인과 암 환자, 또는 각 암 종 사이에서 발현 빈도에 차이가 있는 motif를 선별한 과정의 예시이다.

[22] 도 3은 본 발명의 일 실시예에서 선별한 핵산단편들의 크기 분포를 확인한 그래프이다.

[23] 도 4의 왼쪽 패널은 본 발명의 일 실시예에서 제작한 FEMS table을 하나의 핵산단편으로 작성한 예시이고, 오른쪽 패널은 전체 핵산단편으로 작성한 예시이다.

[24] 도 5의 왼쪽 패널은 본 발명의 일 실시예에서 Edge summary를 추가로 수행하여 작성한 FEMS table의 예시이며, 오른쪽 패널은 이를 시각화한 결과이다.

[25] 도 6은 본 발명의 일 실시예에서 사용한 건강인, 간암 환자 및 식도암 환자의 데이터를 바탕으로 작성한 FEMS table의 시각화 예시이다.

[26] 도 7의 (A)는 본 발명의 일 실시예에서 구축한 CNN 모델의 성능을 Accuracy와 micro AUC로 확인한 결과이며, (B)는 혼동행렬(confusion matrix)이다.

[27] 도 8은 본 발명의 일 실시예에서 구축한 CNN 모델에서 예측한 건강인, 간암 환자 및 식도암 환자의 확률값이 실제 환자와 얼마나 일치하는지를 CNN 모델이 출력한 DPI 값의 분포를 통해 확인한 결과이다.

[28] 도 9는 본 발명의 일 실시예에서 구축한 CNN 모델의 구성을 나타낸 개략도이다.

[29]

발명의 상세한 설명 및 바람직한 구현예

[30] 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.

[31] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 해당 구성요소들은 상기 용어들에 의해 한정되지는 않으며, 단지 하나의

구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 이하 설명하는 기술의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

- [33] 본 명세서에서 사용되는 용어에서 단수의 표현은 문맥상 명백하게 다르게 해석되지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함한다" 등의 용어는 설시된 특징, 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 의미하는 것이지, 하나 또는 그 이상의 다른 특징들이나 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다.
- [34] 도면에 대한 상세한 설명을 하기에 앞서, 본 명세서에서의 구성부들에 대한 구분은 각 구성부가 담당하는 주기능 별로 구분한 것에 불과함을 명확히 하고자 한다. 즉, 이하에서 설명할 2개 이상의 구성부가 하나의 구성부로 합쳐지거나 또는 하나의 구성부가 보다 세분화된 기능별로 2개 이상으로 분화되어 구비될 수도 있다. 그리고 이하에서 설명할 구성부 각각은 자신이 담당하는 주기능 이외에도 다른 구성부가 담당하는 기능 중 일부 또는 전부의 기능을 추가적으로 수행할 수도 있으며, 구성부 각각이 담당하는 주기능 중 일부 기능이 다른 구성부에 의해 전담되어 수행될 수도 있음은 물론이다.
- [35] 또, 방법 또는 동작 방법을 수행함에 있어서, 상기 방법을 이루는 각 과정들은 문맥상 명백하게 특정 순서를 기재하지 않은 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 과정들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.
- [36] 본 발명에서는, 샘플에서 획득한 서열 분석 데이터를 참조 유전체에 정렬한 다음, 정렬된 서열정보를 기반으로 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하고, 상기 도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 이용하여 벡터화된 데이터를 생성한 다음, 학습된 인공지능 모델에서 DPI값을 계산하여 분석할 경우, 높은 민감도와 정확도로 암 진단 및 암 종류를 예측할 수 있다는 것을 확인하고자 하였다.
- [37] 즉, 본 발명의 일 실시예에서는, 혈액에서 추출한 DNA를 시퀀싱 한 뒤, 참조 염색체에 정렬한 다음, 이를 이용하여 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하고, 핵산단편의 말단 서열 모티프 빈도를 X축으로 하고, 핵산단편의 크기를 Y축으로 하는 벡터화된 데이터를 생성한 다음, 이를 딥러닝 모델에 학습시켜 DPI 값을 계산하였으며, 이를 기준값과 비교하여 암 진단을 수행한 다음, 각 암 종류로 계산된 DPI 값 중, 가장 높은 DPI값을 나타낸 암 종을 샘플의 암 종으로 결정하는 방법을 개발하였다(도 1)
- [38] 따라서, 본 발명은 일관점에서,
- [39] (a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;

- [40] (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;
- [41] (c) 상기 정렬된 서열정보(reads)를 이용하여 핵산단편(fragments)의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하는 단계;
- [42] (d) 상기 도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 이용하여 벡터화된 데이터를 생성하는 단계;
- [43] (e) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및
- [44] (f) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 포함하는 암 진단 및 암 종 예측을 위한 정보의 제공방법에 관한 것이다.
- [45]
- [46] 본 발명에 있어서, 상기 핵산 단편은 생체시료에서 추출한 핵산의 조각이면 제한없이 이용할 수 있으며, 바람직하게는 세포 유리 핵산 또는 세포 내 핵산의 조각일 수 있으나, 이에 한정되는 것은 아니다.
- [47] 본 발명에 있어서, 상기 핵산 단편은 통상의 기술자에게 알려진 모든 방법으로 얻을 수 있으며, 바람직하게는 직접 서열분석하거나, 차세대 염기서열 분석을 통해 서열분석하거나 또는 비특이적 전장 유전체 증폭(non-specific whole genome amplification)을 통해 서열분석하여 얻거나, 프로브 기반 서열분석을 통해 얻을 수 있으나, 이에 한정되는 것은 아니다.
- [48]
- [49] 본 발명에서 상기 암은 고형암 또는 혈액암일 수 있으며, 바람직하게는 비호지킨 림프종 (non-Hodgkin lymphoma), 호지킨 림프종 (non-Hodgkin lymphoma), 급성 골수성 백혈병 (acute-myeloid leukemia), 급성 림프구성 백혈병 (acute-lymphoid leukemia), 다발성 골수종 (multiple myeloma), 경부암 (head and neck cancer), 폐암, 교모세포종 (glioblastoma), 대장/직장암, 췌장암, 유방암, 난소암, 흑색종 (melanoma), 전립선암, 간암, 갑상선암, 위암, 담낭암, 담도암, 방광암, 소장암, 자궁경부암, 원발부위불명암, 신장암, 식도암 및 중피종 (mesothelioma)으로 구성된 군에서 선택될 수 있으며, 더욱 바람직하게는 간암 또는 식도암 일 수 있으나, 이에 한정되는 것은 아니다.
- [50]
- [51] 본 발명에 있어서,
- [52] 상기 (a) 단계는
- [53] (a-i) 생체시료에서 핵산을 수득하는 단계;
- [54] (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;
- [55] (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;

- [56] (a-iv) 제작된 라이브러리를 차세대 유전자서열검사기(next-generation sequencer)에 반응시키는 단계; 및
- [57] (a-v) 차세대 유전자서열검사기에서 핵산의 서열정보(reads)를 획득하는 단계;
- [58] 를 포함하는 것을 특징으로 할 수 있다.
- [59]
- [60] 본 발명에 있어서, 상기 (a) 단계의 서열정보를 획득하는 단계는 분리된 무세포 DNA를 1백만 내지 1억 리드 깊이로 전장 유전체 시퀀싱을 통해 획득하는 것을 특징으로 할 수 있다.
- [61] 본 발명에 있어서, 상기 생체시료는 개체로부터 얻어지거나 개체로부터 유래된 임의의 물질, 생물학적 체액, 조직 또는 세포를 의미하는 것으로, 예를 들면, 전혈(whole blood), 백혈구(leukocytes), 말초혈액 단핵 세포(peripheral blood mononuclear cells), 백혈구 연층(buffy coat), (혈장(plasma) 및 혈청(serum)을 포함하는) 혈액, 객담(sputum), 눈물(tears), 점액(mucus), 세비액(nasal washes), 비강 흡인물(nasal aspirate), 호흡(breath), 소변(urine), 정액(semen), 침(saliva), 복강 세척액(peritoneal washings), 골반 내 유체액(pelvic fluids), 낭종액(cystic fluid), 뇌척수막 액(meningeal fluid), 양수(amniotic fluid), 선액(glandular fluid), 췌장액(pancreatic fluid), 림프액(lymph fluid), 흉수(pleural fluid), 유두 흡인물(nipple aspirate), 기관지 흡인물(bronchial aspirate), 활액(synovial fluid), 관절 흡인물(joint aspirate), 기관 분비물(organ secretions), 세포(cell), 세포 추출물(cell extract), 정액, 모발, 타액, 소변, 구강세포, 태반세포, 뇌척수액(cerebrospinal fluid) 및 이의 혼합물을 포함할 수 있으나, 이에 제한되는 것은 아니다.
- [62]
- [63] 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.
- [64] 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.

- [65] 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.
- [66] NGS 기술은, 예를 들어 주형 제조, 시퀀싱 및 이미징 및 데이터 분석 단계 중 하나 이상을 포함할 수 있다.
- [67] 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.
- [68] 클론이 증폭된 주형의 제조방법은, 예를 들어 에멀전 PCR(emulsion PCR: emPCR) 및 고체상 증폭을 포함한다.
- [69] 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.

- [70] 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.
- [71] 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.
- [72] 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.
- [73] 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.

- [74] 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.
- [75] 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.
- [76] SBL은 시퀀싱을 위해 DNA 리가제 및 1-염기-암호화된 프로브 또는 2-염기-암호화된 프로브 중 하나를 사용한다.
- [77] 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.
- [78] 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.

- [79] 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)
- [80] 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.
- [81] 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.
- [82] 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.
- [83] 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.
- [84] 폴로니 시퀀싱은 폴로니 증폭 및 다중 단일-염기-연장(FISSEQ)을 통해 시퀀싱에 따르는 것을 기반으로 한다. 폴로니 증폭은 폴리아크릴아마이드 펠름 상에서 인시츄로 DNA를 증폭시키는 방법이다. 대표적인 폴로니 시퀀싱 방법은, 예를 들어 미국특허 출원 공개 제2007/0087362호에 기재된다.
- [85] 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.

- [86] 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.
- [87] 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.
- [88] 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; 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.
- [89]
- [90] NGS 리드가 만들어진 후, 그것들은 공지된 기준 서열에 대해 정렬되거나 데노보 조립된다. 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.
- [91] 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.
- [92] 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.

[93]

[94] 본 발명에 있어서, 상기 (b) 단계의 서열 정렬은 컴퓨터 알고리즘으로서 계층에서 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.

[95] 전형적으로, 정렬 과정에 두 단계가 있을 수 있다: 후보자 검사 및 서열 정렬. 후보자 검사는 가능한 정렬 위치의 더 짧은 열거에 대해 전체 계층으로부터 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-분취(Needleman-Wunsch) 정렬) 또는 국소 정렬(예를 들어, 스미스-워터만 정렬)을 사용하여 수행될 수 있다.

[96] 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.

- [97] 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.
- [98] 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.

[99]

[100] 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.

[101]

[102] 본 발명에 있어서, 상기 (c) 단계의 핵산단편 말단 서열 모티프는 핵산단편 양 말단의 2 내지 30개의 염기서열의 패턴인 것을 특징으로 할 수 있다.

[103] 즉, 하기와 같이 페어드-엔드 시퀀싱에 의해 서열분석된 핵산단편이 있을 시,

[104] Forward strand: 5'-TACAGACTTTGGAAT-3' (서열번호 1)

[105] Reverse strand: 3'-ATGACTGAAACCTTA-5' (서열번호 2)

[106] Forward strand 5' 말단에서부터 순서대로 읽은 TACA와, Reverse strand 5' 말단에서부터 순서대로 읽은 ATTC가 이 핵산단편의 말단 서열 모티프 값이 된다.

[107]

[108] 본 발명에 있어서, 상기 (c) 단계의 핵산단편 말단 서열 모티프의 빈도는 전체 핵산 단편에서 검출된 각각의 모티프 수인 것을 특징으로 할 수 있다.

[109] 즉, 핵산단편 말단 서열 모티프를 양 말단의 4개의 염기를 바탕으로 분석할 경우(4-mer motif), 1, 2, 3, 4 번째 위치에 각각 A, T, G, C 네 종류의 염기 조합이 가능하기 때문에, 총 256 가지 (4*4*4*4) 조합의 motif 값이 분석 대상이 된다.

[110] 시퀀싱으로 생산된 전체 핵산 단편에서 각 motif들이 관측되는 수를 계수한 것이 모티프 빈도이고, 이 값을 생산된 전체 핵산 단편 숫자로 나누어 계산한 값이 각 motif의 상대 빈도(relative frequency)이다.

[111]

[112] [표1]

	AAAA	AAAC	AAAG	AAAT	AACA	AACC	...	TTTT	Row Sum
Forward Strand	62,639	105,142	127,299	75,485	399,505	42,583	...	269,530	63,319,687
Reverse Strand	62,432	105,719	126,493	75,788	400,900	42,467	...	269,802	63,110,437
Merged	125,071	210,861	253,792	151,273	800,405	85,050	...	539,332	126,430,124
End Motif Relative Freq	0.00099	0.00167	0.00201	0.00120	0.00633	0.00067	...	0.00427	-

[113] 상기 표 1에 기재된 바와 같이 전체 핵산 단편의 숫자가 126,430,124개 이고, AAAA가 핵산단편 말단 서열 모티프로 분석되는 핵산단편의 개수가 125,071개이므로, AAAA 핵산단편 말단 서열 모티프의 빈도는 125,071이 되고, 이를 전체 핵산단편 숫자로 나누어 계산한 핵산단편 말단 서열 모티프의 상대 빈도는 0.00099가 되는 것이다.

[114]

[115] 본 발명에 있어서, 상기 (c) 단계의 핵산단편의 크기는 핵산단편의 5' 말단에서 3' 말단까지의 염기 개수인 것을 특징으로 할 수 있다.

[116] 예를 들어, 상기 서열번호 1 및 2로 분석되는 핵산단편의 크기는 15이다.

[117] 본 발명에서, 상기 핵산단편의 크기는 1 내지 10000일 수 있고, 바람직하게는 10 내지 1000일 수 있으며, 더욱 바람직하게는 50 내지 500일 수 있고, 가장 바람직하게는 90 내지 250일 수 있으나, 이에 한정되는 것은 아니다.

[118]

[119] 본 발명에 있어서, 상기 (d) 단계의 벡터화된 데이터는 핵산단편 말단 서열 모티프 종류를 X축으로 하고, 핵산단편의 크기를 Y축으로 하는 것을 특징으로 할 수 있다.

[120] 즉, 아래와 같은 핵산 단편이 하나 있다고 가정했을 때,

[121] Forward strand: 5'-**TACAGACTAGT** ... TTGGAAT-3' (서열번호 3)

[122] Reverse strand: 3'-ATGACTGATCA ... AACCTTA-5' (서열번호 4)

[123] Fragment Size: **176**

[124] 이 핵산단편은 도 4의 왼쪽 패널과 같은 2차원 벡터로 표현될 수 있으며, 이러한 과정을 전체 핵산 단편으로 확장하여 누적하면, 도 4의 오른쪽 패널과 같은 2차원 벡터를 생성하는 것이다.

[125]

[126] 본 발명에 있어서, 상기 벡터화된 데이터는 핵산단편 말단 모티프별 빈도의 총합 및 핵산단편 크기별 빈도의 총합을 추가로 포함하는 것을 특징으로 할 수 있다.

[127] 즉, Fragment Size에 관계 없는 Fragment End Motif별 frequency 정보를 추가하기 위해 열 합 (column sum) 값을 도 4의 2차원 벡터 가장 아래에 4번 추가하고, Fragment End Motif에 관계 없는 Fragment Size 정보를 추가하기 위해 행 합 (row sum) 값을 도 4의 2차원 벡터 가장 우측에 4번 추가하는 Edge Summary를 추가적으로 수행하여, 도 5의 왼쪽 패널과 같은 2차원 벡터를 생성하는 것이다.

[128] 본 발명에서는 상기 2차원 벡터를 Fragment End Motif frequency and Size

(FEMS) table이라 정의하였다. FEMS table을 시각화하면 도 5의 오른쪽 패널 및 도 6과 같이 나타날 수 있다.

[129]

[130] 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.

[131] 본 발명의 벡터화된 데이터는 이미지에 한정되지 않고, 예를 들어 n개의 흑백 이미지 여러 장으로 쌓아서 n차원의 2D 벡터(Multi-dimensional Vector)를 이용하여 인공지능 모델의 입력 데이터로 사용할 수 있다.

[132]

[133] 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.

[134] 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.

[135]

[136] 본 발명에 있어서, 상기 (e) 단계의 인공지능 모델은 암 종류별 이미지를 구별할 수 있도록 학습할 수 있는 모델이면 제한없이 사용가능하며, 바람직하게는 딥러닝 모델인 것을 특징으로 할 수 있다.

[137]

[138] 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) 및 순환 신경망(Recurrent Neural Network, RNN)으로 구성된 군에서 선택되는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[139]

[140] 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(attention recurrent neural network)으로 구성된 군에서 선택되는 것을 특징으로 할 수 있다.

[141]

[142] 본 발명에 있어서, 상기 인공지능 모델이 CNN일 경우, binary classification을 수행하는 손실함수는 하기 수식 1로 표시되는 것을 특징으로 할 수 있고, Multi-class classification을 수행하는 손실함수는 하기 수식 2로 표시되는 것을

특징으로 할 수 있다.

[143] 수식 1: 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

[144]

수식 2: 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

[145] 본 발명에서, 상기 binary classification은 인공지능 모델이 암 유무를 판별하도록 학습하는 것을 의미하며, multi-class classification은 인공지능 모델이 두 가지 이상의 암 종을 판별하도록 학습하는 것을 의미한다.

[146]

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

[148] i) 생산된 벡터 데이터를 training(학습), validation(검증), test(성능평가) 데이터로 분류하는 단계;

[149] 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.

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

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

[152]

- [153] 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.
- [154] 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.
- [155]
- [156] 본 발명에 있어서, 상기 e) 단계에서 인공지능 모델이 입력된 벡터화된 데이터로부터 분석한 결과값은 특정 score 또는 실수이면 제한없이 이용가능하며, 바람직하게는 DPI(Deep Probability Index) 값인 것을 특징으로 할 수 있으나 이에 한정되는 것은 아니다.
- [157]
- [158] 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.
- [159] Binary classification일 경우에는 sigmoid function을 이용하여 암 일 경우 DPI 값이 1이 되게끔 학습을 하게 된다. 예를 들어, 유방암 샘플과 정상 샘플이 입력되면, 유방암 샘플의 DPI 값이 1에 가깝도록 학습하는 것이다.
- [160] Multi-class classification 일 경우에는 softmax function을 이용하여, class 개수만큼의 DPI 값을 뽑게 된다. Class 개수만큼의 DPI값의 합은 1이되고, 실제 해당되는 암 종의 DPI값이 1이 되게끔 학습을 하게 된다. 예를 들어, 3개의 class 유방암, 간암, 정상이 있고, 유방암 sample이 들어오면, 유방암 class를 1에 가깝게 학습하게 되는 것이다.
- [161] 본 발명에서 상기 (e) 단계의 출력 결과값은 암 종별로 도출되는 것을 특징으로 할 수 있다.
- [162]
- [163] 본 발명에서, 상기 인공지능 모델은 학습할 때, 암이 있으면 output 결과가 1에 가깝게 학습하고, 암이 없으면 output 결과가 0에 가깝게 학습을 시켜서, 0.5를 기준으로 0.5 이상이면 암이 있다고 판단하고, 0.5 이하이면 암이 없다고 판단하여 performance 측정을 수행하였다(Training, validation, test accuracy).
- [164] 여기서, 0.5의 기준값은 언제든지 바뀔 수 있는 값이라는 것은 통상의 기술자에게 자명한 것이다. 예를 들어서 False positive(위양성)를 줄이고자 하면, 0.5보다 높은 기준값을 설정하여 암이 있다고 판단되는 기준을 엄격하게 가져 갈 수 있고, False Negative(위음성)를 줄이고자 하면 기준값을 더 낮게 측정하여 암이 있다고 판단되는 기준을 조금 더 약하게 가져 갈 수 있다.

- [165] 가장 바람직하게는 학습된 인공지능 모델을 이용하여 unseen data(학습에 training하지 않은 답을 알고 있는 data)를 적용시켜서, DPI값의 probability를 확인하여 기준값을 정할 수 있다.
- [166]
- [167] 본 발명에 있어서, 상기 (f) 단계의 출력 결과값 비교를 통해 암 종을 예측하는 단계는 출력 결과값 중, 가장 높은 값을 나타내는 암 종을 샘플의 암으로 판정하는 단계를 포함하는 방법으로 수행하는 것을 특징으로 할 수 있다.
- [168]
- [169] 본 발명은 다른 관점에서, 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부;
- [170] 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 및
- [171] 정렬된 서열 기반의 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하는 핵산단편 분석부;
- [172] 도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 이용한 벡터화된 데이터를 생성하는 데이터 생성부;
- [173] 생성된 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석하고, 기준값과 비교하여 암 유무를 판정하는 암 진단부; 및
- [174] 출력된 결과값을 분석하여 암 종을 예측하는 암 종 예측부를 포함하는 암 진단 및 암 종 예측 장치에 관한 것이다.
- [175] 본 발명에서, 상기 해독부는 독립된 장치에서 추출된 핵산을 주입하는 핵산 주입부; 및 주입된 핵산의 서열정보를 분석하는 서열정보 분석부를 포함할 수 있으며, 바람직하게는 NGS 분석 장치일 수 있으나, 이에 한정되는 것은 아니다.
- [176] 본 발명에서, 상기 해독부는 독립된 장치에서 생성된 서열정보 데이터를 수신하여 해독하는 것을 특징으로 할 수 있다.
- [177]
- [178] 본 발명은 또 다른 관점에서, 컴퓨터 판독 가능한 저장 매체로서, 암 진단 및 암 종을 예측하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,
- [179] (a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;
- [180] (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;
- [181] (c) 상기 정렬된 서열정보(reads)를 이용하여 핵산단편(fragments)의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하는 단계;
- [182] (d) 상기 도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 이용하여 벡터화된 데이터를 생성하는 단계;
- [183] (e) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및
- [184] (f) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 통하여, 암 유무 및 암 종을 예측하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는

컴퓨터 판독 가능한 저장 매체에 관한 것이다.

[185]

[186] 다른 양태에서 본원에 따른 방법은 컴퓨터를 이용하여 구현될 수 있다. 일 구현예에서, 컴퓨터는 칩 세트에 연결된 하나 이상의 프로세서를 포함한다. 또한 칩 세트에는 메모리, 저장 장치, 키보드, 그래픽 어댑터(Graphics Adapter), 포인팅 장치(Pointing Device) 및 네트워크 어댑터(Network Adapter) 등이 연결되어 있다. 일 구현예에서, 상기 칩 세트의 성능은 메모리 컨트롤러 허브(Memory Controller Hub) 및 I/O 컨트롤러 허브에 의하여 가능하다. 다른 구현예에서, 상기 메모리는 칩 세트 대신에 프로세서에 직접 연결되어 사용될 수 있다. 저장 장치는 하드 드라이브, CD-ROM(Compact Disk Read-Only Memory), DVD 또는 기타 메모리 장치를 포함하는 데이터를 유지할 수 있는 임의의 장치이다. 메모리는 프로세서에 의하여 사용된 데이터 및 명령에 관여한다. 상기 포인팅 디바이스는 마우스, 트랙볼 (Track Ball) 또는 다른 유형의 포인팅 디바이스일 수 있고, 키보드와 조합하여 입력 데이터를 컴퓨터 시스템으로 전송하는데 사용된다. 상기 그래픽 어댑터는 디스플레이 상에서 이미지 및 다른 정보를 나타낸다. 상기 네트워크 어댑터는 근거리 또는 장거리 통신망으로 컴퓨터 시스템과 연결된다. 본원에 사용되는 컴퓨터는 하지만 위와 같은 구성으로 제한되는 것은 아니고, 일부 구성이 없거나, 추가의 구성을 포함 할 수 있으며, 또한 저장장치영역네트워크(Storage Area Network, SAN)의 일부일 수 있으며, 본원의 컴퓨터는 본원에 따른 방법의 수행을 위한 프로그램에 모듈의 실행에 적합하도록 구성될 수 있다.

[187]

[188] 본원에서 모듈이라 함은, 본원에 따른 기술적 사상을 수행하기 위한 하드웨어 및 상기 하드웨어를 구동하기 위한 소프트웨어의 기능적, 구조적 결합을 의미할 수 있다. 예컨대, 상기 모듈은 소정의 코드와 상기 소정의 코드가 수행되기 위한 하드웨어 리소스(Resource)의 논리적인 단위를 의미할 수 있으며, 반드시 물리적으로 연결된 코드를 의미하거나, 한 종류의 하드웨어를 의미하는 것은 아님은 본원 기술분야의 당업자에게 자명한 것이다.

[189]

[190] 본 발명은 또 다른 관점에서,

[191] (a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;

[192] (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

[193] (c) 상기 정렬된 서열정보(reads)를 이용하여 핵산단편(fragments)의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하는 단계;

[194] (d) 상기 도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 이용하여 벡터화된 데이터를 생성하는 단계;

[195] (e) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한

- 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및
- [196] (f) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 포함하는 암 진단 및 암 종 예측방법에 관한 것이다.
- [197]
- [198] 실시예
- [199] 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.
- [200]
- [201] 실시예 1. 혈액에서 DNA를 추출하여, 차세대 염기서열 분석 수행
- [202] 건강인 349명, 간암 환자 51명 및 식도암 환자 108명의 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, MGIEasy cell-free DNA library prep set kit 를 사용하여 library preparation 과정을 수행 한 다음, DNBseq G400 장비 (MGI) 를 100 base Paired end 모드로 sequencing 하였다. 그 결과, 샘플 당 약 170 million 개의 reads가 생산되는 것을 확인 하였다.
- [203]
- [204] 실시예 2. 핵산단편 말단 모티프 및 핵산단편 크기 선별
- [205] 2-1. 핵산단편 말단 모티프 선별
- [206] 핵산단편 말단 모티프를 4개의 염기(A, T, G, C)로 설정하고, 전체 256가지(4*4*4*4) 가지 종류의 motif 중에는 Normal / HCC / EC 그룹에서 상대빈도 차이가 없는 motif 들이 있다. 이러한 차이가 없는 motif를 포함하여 FEMS table을 생성할 경우, Classification을 위한 의미 있는 정보는 주지 못하면서 모델의 연산 양만 늘리는 Noise가 된다. 따라서, 이러한 의미 없는 Motif는 제외시키기 위해, 세 그룹에서 상대빈도의 차이가 유의미하게 존재했던 특정 motif들만 선별하였다.
- [207] 아울러, Size, Motif 선별 과정에서 모델 과적합 (Overfitting) 이슈가 생기는 것을 방지하기 위해, Size, Motif 선별 과정에는 Training set만을 사용하다.
- [208] 즉, 실시예 1에서 생성한 NGS 데이터를 이용하여, 핵산단편 말단 모티프를 4개의 염기(A, T, G, C)로 설정하고, 전체 256가지(4*4*4*4) 가지 종류의 motif 들 중 건강인(Normal), 간암(HCC), 식도암(EC) 환자 그룹 사이에 통계적으로 유의미한 수준의 (Kruskal-wallis Test, FDR-adjust $p < 0.05$) 상대 빈도(relative frequency) 차이를 나타내는 일부 motif를 선별하였다(도 2).
- [209] 또한, 위 과정에서 선별한 motif 중, overfitting을 방지하기 위하여 건강인

그룹에서의 평균 frequency가 random baseline (1/256, 0.004) 보다 높은 motif를 추가 선별하였다.

[210] 그 결과, 총 84개의 모티프를 선별하였으며, 상세한 모티프 정보는 아래와 같다:

[211] CTGG, ACTT, CCTA, TGGA, TGGG, CAGG, TATA, CCTT, CAGC, TAGA, AGAA, AGAG, CATA, CAGT, CAGA, ACCT, CTGT, ACAT, GCTT, GCTA, TCAG, CTTA, GGCC, ATTT, CCCA, TATC, CCTG, TCTA, GCCT, ACTG, TGAG, GGTA, CATT, TATT, CCAT, CCTC, CCAA, CTTT, TAAG, GCTG, CCCT, TGAA, ACCA, GTTT, TGTA, CTCA, GCCA, TATG, GCAT, AAAG, AAAA, GGCT, TGAC, AGCA, TCTT, CTGA, CATC, ACAA, GACA, AACA, CCCC, CACT, GGAG, GGCA, TCAA, CAAG, TAAA, AAAT, TGCC, GGTT, GGGA, CCAC, TGTG, CATG, TGCA, GAAT, TGTC, TGCT, CAAT, GGAA, AGTG, TACT, CACA, TCCC

[212] 2-2. 핵산단편 크기 선별

[213] 핵산단편 크기 선별의 경우에는, 퀄리티 확인이 끝난 핵산 단편들은 대부분 도 3에 기재된 바와 같이, 90~250 범위의 size를 가지고 있으므로, 이 크기 범위를 벗어나는 '영역을 포함하여 FEMS table을 생성할 경우, 대부분의 영역이 0 값으로 채워지고, 의미 없는 Noise만 증가하게 되므로, 상기 크기를 선별하였다.

[214]

[215] 실시예 3. Fragment End Motif frequency and Size (FEMS) table 생성

[216] 실시예 2에서 선별한 핵산 단편의 Fragment End Motif frequency 값과 Size 정보를 동시에 표현할 수 있도록 X 축에는 motif 종류를, Y 축에는 Fragment Size를 배열하여 2차원의 벡터를 생성하였다. 보다 구체적으로는, 도 4의 왼쪽 패널에 기재된 바와 같이, 하나의 핵산 단편에 대하여 양 말단의 핵산 모티프 종류와 크기를 빈도수로 표현하고, 이를 전체 핵산 단편으로 확장시키고 누적하여, 도 4에 기재된 바와 같은 2차원 벡터를 생성하였다.

[217] 또한, Fragment Size에 관계 없는 Fragment End Motif별 frequency 정보를 추가하기 위해 열 합 (column sum) 값을 위의 2차원 벡터 가장 아래에 4번 추가하고, Fragment End Motif에 관계 없는 Fragment Size 정보를 추가하기 위해 행 합 (row sum) 값을 위의 2차원 벡터 가장 우측에 4번 추가하는 Edge Summary 단계를 수행하여 최종적으로 도 5에 기재된 바와 같은 2차원 벡터를 생성하였다. 이 2차원 벡터를 Fragment End Motif frequency and Size (FEMS) table이라 정의하였으며, 이를 시각화한 예시는 도 5에 기재된 바와 같다.

[218]

[219] 실시예 3. CNN 모델 구축 및 학습 과정

[220] FEMS table 2차원 벡터를 인풋으로 하여 건강인, 간암 환자, 식도암 환자를 구분하는 CNN 인공지능 모델을 학습하였다.

[221] 전체 샘플을 Training, Validation, Test 데이터 세트로 나누어 Training 데이터

세트는 모델 학습에, Validation 데이터 세트는 hyper-parameter tuning에, Test 데이터 세트는 최종 모델 성능 평가에 사용하였다. 각 세트 별 샘플 수는 아래와 같다.

[222]

[223] [표2]

데이터 세트	건강인	간암 환자	식도암 환자	합계
Train	193	26	60	279
Valid	71	12	22	105
Test	85	13	26	124
합계	349	51	108	508

- [224] CNN 모델의 기본적인 구성은 도 9와 같다. 활성화함수는 ReLU (RectifiedLinearunit)을 사용하였고, convolution layer 는 1개를 사용하였고, 5개의 10*10 patch 를 사용하였다. Pooling 방식은 max 를 이용했고 2x2 patch 를 이용하였다. Fully connected layer는 1개를 사용하였고 512개의 hidden node가 포함되어 있다. 마지막으로 softmax 함수값을 이용해 최종 DPI 값을 계산하였다.
- [225] 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 과정에는 Bayesian optimization 및 grid search 기법을 사용하였고, Training loss 대비 validation loss가 증가하기 시작하면 모델이 과적합(Overfitting) 되었다 판단되어 model 학습을 중단하였다.
- [226] Hyper-parameter tuning을 통해서 얻어진 여러 모델의 성능을 Validation 데이터 세트를 이용하여 비교한 다음, Validation 데이터 세트 성능이 가장 좋은 모델을 최적의 모델이라 판단하고, Test 데이터 세트로 최종 성능 평가를 수행하였다.
- [227] 상기 과정을 거쳐서 만들어진 모델에 임의의 샘플의 FEMS table 2차원 벡터를 넣어 주면, CNN 모델의 마지막 layer인 softmax 함수를 통해 해당 샘플의 건강인일 확률, 간암 환자일 확률, 식도암 환자일 확률이 각각 계산되고, 이 확률 값을 Deep Probability Index (DPI)라 정의하였다.
- [228] 임의의 샘플은 세 종류의 DPI 값 중 가장 높은 값을 갖는 그룹으로 판단하게 된다. 예를 들어, 임의의 샘플에서 계산된 건강인, 간암 환자, 식도암 환자의 DPI 값이 각각 0.6, 0.3, 0.1 이었을 경우, 이 샘플은 건강인으로 판단하게 된다.
- [229]
- [230] 실시예 4. 구축한 딥러닝 모델의 성능 확인
- [231] 4-1 성능 확인
- [232] 실시예 3에서 구축한 딥러닝 모델에서 출력한 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.

[233]

[234] [표3]

	Accuracy	micro AUC
Train	0.913	0.991
Validation	0.927	0.990
Test	0.895	0.955

[235] 그 결과, 표 3 및 도 7에 기재된 바와 같이, Accuracy 는 Train, Valid, Test 그룹에서 각각 91.3%, 92.7%, 89.5%인 것을 확인 하였고, Multi-class ROC 분석 결과인 micro AUC 값은 Train, Valid, Test 그룹에서 각각 0.991, 0.990, 0.955로 나타나는 것을 확인하였다. 도 7의 (A)는 Train, Validation, Test 그룹에서 CNN 모델의 성능을 Accuracy와 microAUC로 확인한 것이며, 도 7의 (B)는 Train, Validation, Test 그룹에서 CNN 모델의 성능을 혼동행렬 (confusion matrix)로 확인한 것이다.

[236] 4-2. DPI 분포 확인

[237] 실시예 3에서 구축한 딥러닝 모델의 출력값인 DPI 값이 실제 환자와 얼마나 일치하는 지를 확인하였다. 도 8의 X 축은 실제 샘플의 그룹 (True label) 정보를 나타내고, Y 축은 왼쪽에서부터 순서대로 CNN 모델에서 계산된 건강인(Normal), 간암 환자(HCC), 식도암 환자(EC)일 DPI 값을 나타낸다.

[238] 그 결과, 도 8에 기재된 바와 같이 DPI 분포는 Train, Validation, Test 데이터 세트 모두에서 건강인 샘플들은 건강인일 확률이 가장 높게 분포하는 것을 확인하였고, 간암 환자 샘플들은 간암 환자일 확률이 가장 높게 나타나는 것을 확인하였으며, 식도암 환자 샘플들은 식도암 환자일 확률이 가장 높게 분포하는 것을 확인하였다.

[239]

[240] 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.

[241]

산업상 이용가능성

[242] 본 발명에 따른 세포유리 핵산단편 말단 서열 모티프 빈도 및 크기를 이용한 암 진단 및 암 종 예측방법은 벡터화된 데이터를 생성하여 AI 알고리즘을 이용하여 분석하기 때문에 리드 커버리지가 낮더라도 높은 민감도와 정확도를 나타내어 유용하다.

[243]

서열목록 Free Text

[244] 전자파일 첨부하였음.

청구범위

- [청구항 1] (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 이용하여 핵산단편(fragments)의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하는 단계;
 (d) 상기 도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 이용하여 벡터화된 데이터를 생성하는 단계;
 (e) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및
 (f) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 포함하는 암 진단 및 암 종 예측을 위한 정보의 제공방법
- [청구항 2] (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 이용하여 핵산단편(fragments)의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하는 단계;
 (d) 상기 도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 이용하여 벡터화된 데이터를 생성하는 단계;
 (e) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및
 (f) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 포함하는 암 진단 및 암 종 예측방법.
- [청구항 3] 제1항 또는 제2항에 있어서, 상기 (a) 단계는 다음의 단계를 포함하는 방법으로 수행되는 것을 특징으로 하는 방법:
 (a-i) 혈액, 정액, 질 세포, 모발, 타액, 소변, 구강세포, 태반세포 또는 태아세포를 포함하는 양수, 조직세포 또는 이의 혼합물에서 핵산을 수득하는 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 수득하는 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 4] 제1항에 있어서, 상기 (c) 단계의 말단 서열 모티프는 핵산단편 양 말단의 2 내지 30개의 염기서열의 패턴인 것을 특징으로 하는 방법.
- [청구항 5] 제1항 또는 제2항에 있어서, 상기 (c) 단계의 말단 서열 모티프 빈도는 전체 핵산 단편에서 검출된 각각의 모티프 수인 것을 특징으로 하는 방법.
- [청구항 6] 제1항 또는 제2항에 있어서, 상기 (c) 단계의 핵산단편의 크기는 핵산단편의 5' 말단에서 3' 말단까지의 염기 개수인 것을 특징으로 하는 방법.
- [청구항 7] 제1항 또는 제2항에 있어서, 상기 (d) 단계의 벡터화된 데이터는 핵산단편 말단 서열 모티프 종류를 X축으로 하고, 핵산단편의 크기를 Y축으로 하는 것을 특징으로 하는 방법.
- [청구항 8] 제7항에 있어서, 상기 벡터화된 데이터는 핵산단편 말단 모티프별 빈도의 총합 및 핵산단편 크기별 빈도의 총합을 추가로 포함하는 것을 특징으로 하는 방법.
- [청구항 9] 제1항 또는 제2항에 있어서, 상기 (e) 단계의 인공지능 모델은 건강인 벡터화된 데이터와 암이 있는 벡터화된 데이터를 구별할 수 있도록 학습하는 것을 특징으로 하는 방법.
- [청구항 10] 제9항에 있어서, 상기 인공지능 모델은 합성곱 신경망(convolutional neural network, CNN), 심층 신경망(Deep Neural Network, DNN) 및 순환 신경망(Recurrent Neural Network, RNN)으로 구성된 군에서 선택되는 것을 특징으로 하는 방법.
- [청구항 11] 제10항에 있어서, 상기 인공지능 모델이 CNN이고, binary classification 을 학습할 경우, 손실함수는 하기 수식 1로 표시되며, 상기 인공지능 모델이 CNN이고, Multi-class classification을 학습할 경우, 손실함수는 하기 수식 2으로 표시되는 것을 특징으로 하는 방법:

수식 1: 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

수식 2: 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

- [청구항 12] 제1항 또는 제2항에 있어서, 상기 (e) 단계의 인공지능 모델이 입력된 벡터화된 데이터를 분석하여 출력하는 결과값은 DPI(Deep Probability Index)값인 것을 특징으로 하는 방법.
- [청구항 13] 제1항 또는 제2항에 있어서, 상기 (e) 단계의 기준값은 0.5이며, 0.5 이상일 경우, 암 인 것으로 판정하는 것을 특징으로 하는 방법.
- [청구항 14] 제1항 또는 제2항에 있어서,
상기 (f) 단계의 출력 결과값 비교를 통해 암 종을 예측하는 단계는 출력 결과값 중, 가장 높은 값을 나타내는 암 종을 샘플의 암으로 판정하는 단계를 포함하는 방법으로 수행하는 것을 특징으로 하는 방법.
- [청구항 15] 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부;
해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부;
정렬된 서열 기반의 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하는 핵산단편 분석부;
도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를 이용한 벡터화된 데이터를 생성하는 데이터 생성부;
생성된 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석하고, 기준값과 비교하여 암 유무를 판정하는 암 진단부; 및
출력된 결과값을 분석하여 암 종을 예측하는 암 종 예측부를 포함하는 암 진단 및 암 종 예측 장치.
- [청구항 16] 컴퓨터 판독 가능한 저장 매체로서, 암 진단 및 암 종을 예측하는 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 이용하여 핵산단편(fragments)의 말단 서열 모티프 빈도 및 핵산단편의 크기를 도출하는 단계;
(d) 상기 도출된 핵산단편의 말단 서열 모티프 빈도 및 핵산단편의 크기를

이용하여 벡터화된 데이터를 생성하는 단계;

(e) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및

(f) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 통하여, 암 유무 및 암 종을 예측하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체.