(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년10월11일
(11) 등록번호 10-2715713
(24) 등록일자 2024년10월04일

(51) 국제특허분류(Int. Cl.)
G16B 40/20 (2019.01) *G06N 20/00* (2019.01)
G16B 20/20 (2019.01) *G16B 30/10* (2019.01)
(52) CPC특허분류
G16B 40/20 (2019.02)
G06N 20/00 (2021.08)
(21) 출원번호 10-2024-0008609
(22) 출원일자 2024년01월19일
심사청구일자 2024년01월19일
(56) 선행기술조사문헌
KR1020220037376 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
주식회사 이노크라스코리아
대전광역시 유성구 문지로 193, 카이스트 문지캠퍼스 진리관 티331호(문지동)
(72) 발명자
임준오
대전광역시 유성구 문지로 193, 카이스트 문지캠퍼스 진리관 티331호
박성열
대전광역시 유성구 문지로 193, 카이스트 문지캠퍼스 진리관 티331호
(74) 대리인
김한솔, 김세환, 김준식, 안제성

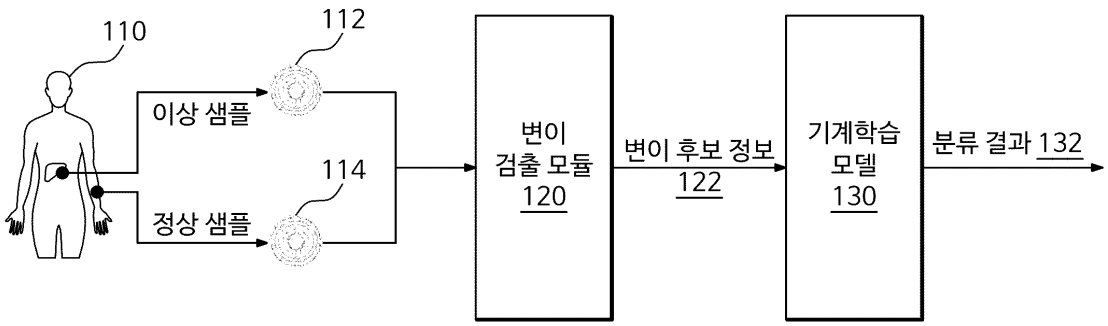
전체 청구항 수 : 총 21 항 심사관 : 성경아

(54) 발명의 명칭 **세포 샘플 내 진양성 변이를 검출하기 위한 기계학습 모델을 학습시키는 방법 및 장치**

(57) 요 약

본 개시는 기계학습 모델을 학습시키는 방법에 관한 것이다. 기계학습 모델을 학습시키는 방법은, 참조 샘플의 참조 변이 후보 정보를 획득하는 단계, 참조 변이 후보와 연관된 어노테이션 정보를 생성하는 단계, 획득된 참조 변이 후보 정보 및 생성된 어노테이션 정보에 기초하여 학습 데이터를 생성하는 단계 및 생성된 학습 데이터를 이용하여 기계학습 모델을 학습시키는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G16B 20/20 (2019.02)
G16B 30/10 (2019.02)

명 세 서

청구범위

청구항 1

적어도 하나의 프로세서에 의해 실행되는, 기계학습 모델을 학습시키는 방법에 있어서,

참조 샘플의 참조 변이 후보 정보를 획득하는 단계;

상기 참조 변이 후보와 연관된 어노테이션 정보를 생성하는 단계;

상기 참조 변이 후보 정보 및 상기 어노테이션 정보에 기초하여, 상기 참조 변이 후보의 특징을 추출하는 단계;

상기 참조 변이 후보 정보 및 상기 추출된 참조 변이 후보의 특징을 포함하는 학습 데이터를 생성하는 단계; 및

상기 생성된 학습 데이터를 이용하여, 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 출력하도록 기계학습 모델을 학습시키는 단계

를 포함하고,

상기 어노테이션 정보를 생성하는 단계는,

FFPE(Formalin-Fixed, Paraffin-Embedded) 처리된 복수의 샘플과 연관된 복수의 시퀀싱 데이터로부터 생성된 FFPE 처리 조직 유전체 데이터(POF: Panel of FFPEs)를 이용하여, 상기 참조 변이 후보 및 상기 FFPE 처리 조직 유전체 데이터와 연관된 제1 어노테이션 정보를 생성하는 단계

를 포함하고,

상기 FFPE 처리된 복수의 샘플은 상기 참조 샘플이 채취된 개체와 상이한 복수의 개체로부터 채취된 샘플을 포함하는, 기계학습 모델을 학습시키는 방법.

청구항 2

제1항에 있어서,

상기 참조 샘플은,

동일한 개체로부터 채취된 참조 정상 샘플 및 참조 이상 샘플을 포함하고,

상기 획득된 참조 변이 후보 정보는,

변이 검출 모듈을 이용하여, 상기 참조 정상 샘플과 연관된 제1 참조 시퀀싱 데이터 및 상기 참조 이상 샘플과 제2 참조 시퀀싱 데이터를 기초로 결정된, 기계학습 모델을 학습시키는 방법.

청구항 3

제2항에 있어서,

상기 변이 검출 모듈은 복수의 검출 모듈을 포함하고,

상기 획득된 참조 변이 후보 정보는,

상기 복수의 검출 모듈을 이용하여 획득된 참조 변이 서브 후보 정보를 통합(union)함으로써 결정되고,

상기 획득된 참조 변이 서브 후보 정보는,

상기 복수의 검출 모듈의 각각에 상기 제1 참조 시퀀싱 데이터 및 상기 제2 참조 시퀀싱 데이터를 적용함으로써 결정되는, 기계학습 모델을 학습시키는 방법.

청구항 4

제1항에 있어서,

상기 어노테이션 정보를 생성하는 단계는,

상기 참조 샘플의 시퀀싱 결과 생성된 리드(read) 중, 레퍼런스 지놈에 매핑된 위치의 적어도 일부가 상기 참조 변이 후보의 위치와 중첩되는 복수의 리드를 결정하는 단계; 및

상기 결정된 복수의 리드와 연관된 제2 어노테이션 정보를 생성하는 단계

를 더 포함하는, 기계학습 모델을 학습시키는 방법.

청구항 5

제4항에 있어서,

상기 복수의 리드는 상기 레퍼런스 지놈과 상이한 복수의 변이 리드(variant read)를 포함하고,

상기 제2 어노테이션 정보는,

상기 복수의 변이 리드의 인서트 사이즈(insert size)의 최솟값, 상기 복수의 변이 리드의 인서트 사이즈의 최댓값, 또는 상기 복수의 변이 리드 중 특정 조건을 만족하는 페어드 리드(paired read)의 개수 중 적어도 하나를 포함하고,

상기 특정 조건은,

상기 페어드 리드의 제1 리드 및 제2 리드가 각각 정방향과 역방향으로 정렬되고, 상기 페어드 리드의 인서트 사이즈가 하한임계치와 상한임계치 사이인 조건을 포함하는, 기계학습 모델을 학습시키는 방법.

청구항 6

제1항에 있어서,

상기 어노테이션 정보를 생성하는 단계는,

복수의 정상 샘플과 연관된 복수의 시퀀싱 데이터로부터 생성된 정상 조직 유전체 데이터(PON: Panel of Normals)를 이용하여, 상기 참조 변이 후보 및 상기 정상 조직 유전체 데이터와 연관된 제3 어노테이션 정보를 생성하는 단계

를 더 포함하는, 기계학습 모델을 학습시키는 방법.

청구항 7

삭제

청구항 8

제1항에 있어서,

상기 FFPE 처리 조직 유전체 데이터는,

상기 FFPE 처리된 복수의 샘플 중, 샘플 내 염기 서열 상의 특정 위치(position)에 대한 VAF(Variant Allele Frequency)가 미리 정해진 임계치 미만인 샘플의 수를 포함하는, 기계학습 모델을 학습시키는 방법.

청구항 9

제1항에 있어서,

상기 FFPE 처리 조직 유전체 데이터는,

상기 FFPE 처리된 복수의 샘플 중, 샘플 내 염기 서열 상의 특정 위치에서 미리 정해진 개수의 변이 리드를 갖는 샘플의 수를 포함하는, 기계학습 모델을 학습시키는 방법.

청구항 10

제1항에 있어서,

상기 어노테이션 정보를 생성하는 단계는,

상기 참조 변이 후보의 변이 유형과 연관된 정보 및 상기 참조 변이 후보의 시퀀스 컨텍스트(sequence context) 정보를 포함하는 제4 어노테이션 정보를 생성하는 단계

를 더 포함하는, 기계학습 모델을 학습시키는 방법.

청구항 11

제1항에 있어서,

상기 학습 데이터를 생성하는 단계는,

상기 참조 변이 후보에 대한 분류 정보를 레이블링하는 단계

를 포함하는, 기계학습 모델을 학습시키는 방법.

청구항 12

제11항에 있어서,

상기 참조 샘플은 FFPE 처리된 샘플이고,

상기 레이블링하는 단계는,

상기 참조 변이 후보 정보의 적어도 일부와, FF(Fresh-Frozen) 처리된 샘플 내 변이 후보 중 어느 하나의 변이 후보와 연관된 정보의 적어도 일부가 서로 대응되는 것으로 판단되는 것에 응답하여, 상기 참조 변이 후보가 진양성(True Positive) 변이인 것으로 레이블링하는 단계

를 포함하고,

상기 FF 처리된 샘플은 상기 FFPE 처리된 샘플에 대응되는 샘플인, 기계학습 모델을 학습시키는 방법.

청구항 13

제12항에 있어서,

상기 레이블링하는 단계는,

상기 참조 변이 후보 정보와, FF 처리된 샘플 내 변이 후보 중 어느 하나의 변이 후보와 연관된 정보가 서로 대응되지 않는 것으로 판단되는 것에 응답하여, 상기 참조 변이 후보가 위양성(False positive) 변이인 것으로 레이블링하는 단계

를 더 포함하는, 기계학습 모델을 학습시키는 방법.

청구항 14

제11항에 있어서,

상기 학습 데이터를 생성하는 단계는,

상기 레이블링된 분류 정보를 상기 학습 데이터에 더 포함시키는 단계

를 더 포함하는, 기계학습 모델을 학습시키는 방법.

청구항 15

제14항에 있어서,

상기 기계학습 모델은 복수의 분류기(classifier)를 포함하고,

상기 기계학습 모델을 학습시키는 단계는,

상기 참조 변이 후보 정보 및 상기 참조 변이 후보의 특징을 상기 복수의 분류기의 각각에 입력하는 단계;

상기 복수의 분류기 중 적어도 하나의 분류기로부터의 출력 결과를 이용하여 상기 참조 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 결정하는 단계; 및

상기 참조 변이 후보의 분류 결과와 상기 참조 변이 후보에 레이블링된 분류 정보에 기초하여 상기 기계학습 모델의 파라미터를 조정하는 단계

를 포함하는, 기계학습 모델을 학습시키는 방법.

청구항 16

제1항에 있어서,

타겟 샘플 내 타겟 변이 후보 정보 및 상기 타겟 변이 후보의 특징이 기계학습 모델에 입력됨으로써, 상기 분류 결과가 출력되는, 기계학습 모델을 학습시키는 방법.

청구항 17

제16항에 있어서,

상기 타겟 샘플은,

동일한 개체로부터 채취된 타겟 정상 샘플 및 타겟 이상 샘플을 포함하고,

상기 타겟 변이 후보 정보는,

변이 검출 모듈을 이용하여, 상기 타겟 정상 샘플과 연관된 제1 타겟 시퀀싱 데이터 및 상기 타겟 이상 샘플과 연관된 제2 타겟 시퀀싱 데이터를 기초로 결정되는, 기계학습 모델을 학습시키는 방법.

청구항 18

제16항에 있어서,

상기 타겟 샘플은 FFPE 처리된 샘플인, 기계학습 모델을 학습시키는 방법.

청구항 19

적어도 하나의 프로세서에 의해 실행되는, 세포 샘플 내 진양성 변이 검출을 통한 유전체 프로파일링 방법에 있어서,

타겟 샘플의 타겟 변이 후보 정보를 획득하는 단계;

기계학습 모델을 이용하여, 상기 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 결정하는 단계; 및

상기 결정된 분류 결과를 기초로 상기 타겟 샘플에 대한 유전체 프로파일링(Genomic Profiling)을 수행하는 단계

를 포함하고,

상기 기계학습 모델은,

참조 샘플의 참조 변이 후보 정보 및 상기 참조 변이 후보와 연관된 어노테이션 정보에 기초하여 추출된 상기 참조 변이 후보의 특징, 및 상기 참조 변이 후보 정보를 이용하여, 상기 타겟 변이 후보가 진양성 변이인지 여부를 결정하도록 학습되고,

상기 어노테이션 정보는,

상기 참조 변이 후보 및 FFPE 처리 조직 유전체 데이터(POF: Panel of FFPEs)와 연관된 어노테이션 정보를 포함하고,

상기 FFPE 처리 조직 유전체 데이터는,

FFPE(Formalin-Fixed, Paraffin-Embedded) 처리된 복수의 샘플과 연관된 복수의 시퀀싱 데이터로부터 생성되고,

상기 FFPE 처리된 복수의 샘플은 상기 참조 샘플이 채취된 개체와 상이한 복수의 개체로부터 채취된 샘플을 포함하는, 유전체 프로파일링 방법.

청구항 20

제19항에 있어서,

상기 유전체 프로파일링의 수행 결과에 기초하여, 상기 타겟 샘플이 채취된 개체의 질병 진단 정보, 치료 전략 정보, 예후 예측 정보 또는 약물 반응성 예측 정보 중 적어도 하나를 제공하는 단계

를 더 포함하는, 유전체 프로파일링 방법.

청구항 21

제1항 내지 제6항 및 제8항 내지 제20항 중 어느 한 항에 따른 방법을 컴퓨터에서 실행하기 위해 컴퓨터 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램.

청구항 22

장치에 있어서,

메모리; 및

상기 메모리와 연결되고, 상기 메모리에 포함된 컴퓨터 판독 가능한 적어도 하나의 프로그램을 실행하도록 구성된 적어도 하나의 프로세서

를 포함하고,

상기 적어도 하나의 프로그램은,

참조 샘플의 참조 변이 후보 정보를 획득하고,

상기 참조 변이 후보와 연관된 어노테이션 정보를 생성하고,

상기 참조 변이 후보 정보 및 상기 어노테이션 정보에 기초하여, 상기 참조 변이 후보의 특징을 추출하고,

상기 획득된 참조 변이 후보 정보 및 상기 추출된 참조 변이 후보의 특징을 포함하는 학습 데이터를 생성하고, 상기 생성된 학습 데이터를 이용하여, 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 출력하도록 기계학습 모델을 학습시키기 위한 명령어를 포함하고,

상기 어노테이션 정보를 생성하는 것은,

FFPE(Formalin-Fixed, Paraffin-Embedded) 처리된 복수의 샘플과 연관된 복수의 시퀀싱 데이터로부터 생성된 FFPE 처리 조직 유전체 데이터(POF: Panel of FFPEs)를 이용하여, 상기 참조 변이 후보 및 상기 FFPE 처리 조직 유전체 데이터와 연관된 어노테이션 정보를 생성하는 것을 포함하고,

상기 FFPE 처리된 복수의 샘플은 상기 참조 샘플이 채취된 개체와 상이한 복수의 개체로부터 채취된 샘플을 포함하는, 장치.

발명의 설명

기 술 분 야

[0001] 본 개시는 기계학습 모델을 학습시키는 방법 및 장치에 관한 것으로, 구체적으로, 참조 변이 후보 정보 및 어노테이션 정보를 이용하여, 기계학습 모델을 학습시키는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 유전정보 분석 기술은 생명체가 가진 유전정보를 파악함으로써 생명체가 어떠한 특성 또는 기질을 가지는지 판단하는 데 사용되는 등 의료 분야에서 폭넓게 사용되고 있다. 최근 종양 등 각종 질병의 원인을 이해하거나 질병을 치료하기 위한 의료 행위는 전통적인 처방 중심의 접근으로부터 정밀 의학(Precision medicine), 다시 말해 개체의 유전정보 및 건강기록 등을 고려한 맞춤형 치료 형태로 진화하고 있다. 정밀 의학 분야에서는 방대한 양의 개체 유전정보를 획득하고 이와 연관된 임상 분석을 수행하는 것이 주요하며, 이러한 주요 요소는 정밀 의학 기술의 발전 속도를 가속화시키는 핵심 요소에 해당한다.

[0003] 특히, 개체로부터 채취한 조직에 대한 전장유전체 분석을 시행하는 경우, 개체로부터 채취된 즉시 조직을 냉동처리하는 이른바 '신선 동결(FF: Fresh Frozen)' 처리 방식이 널리 사용되고 있다. FF 처리된 조직은 채취 직후에 동결됨으로써 조직 내 세포들의 DNA 손상이 적어 전장유전체 분석을 위한 최적의 조직 처리 방식이라고 알려져 있다. 다만, 채취된 조직을 FF 처리하거나, FF 처리된 조직을 보관하기 위해서는 질소 탱크 등 통상적으로 진료 현장에 구비되지 않거나 구비되기 어려운 시설이나 장비가 요구되는 문제가 있다.

[0004] 반면, 개체의 유전정보 분석을 위해 종양조직을 절제하거나 조직검사를 진행하는 경우, 의료기관에서는 개체로부터 채취한 조직(종양조직 등)을 FFPE(Formalin-Fixed, Paraffin-Embedded) 처리하여 장기 보관하고, 후속 검사 또는 학술적인 연구 목적으로 FFPE 처리된 조직을 활용하는 것이 일반적이다. FFPE 방식을 통해 채취된 조직을 처리하는 경우, 채취된 조직의 처리 및 보관에 큰 비용과 노력이 소요되지 않을 뿐만 아니라, 조직 내의 유전적 정보가 대부분 유지된 채 조직이 장기간 보관될 수 있어 채취된 조직을 추후 활용(가령, 조직의 재검사 또는 재분석 등)하기 용이하다는 장점이 있다.

[0005] 그러나, 개체로부터 채취한 조직을 FFPE 처리하여 장기 보관하는 과정에서, DNA의 서로 다른 부분들끼리 화학적으로 엉켜 붙는 교차 결합(cross-linking), DNA가 작은 크기로 절단되는 단편화(fragmentation), 기타 비생물학적 원인으로 인한 DNA 염기의 변이 등 조직 내의 DNA에 다양한 유형의 손상이 발생할 수 있다.

[0006] 위와 같은 DNA의 손상으로 인해, FFPE 처리된 조직을 대상으로 전장유전체 분석 및 변이(mutation) 분석을 수행하는 경우, 변이 검출 데이터에 노이즈가 발생하여 부정확하고 왜곡된 분석 결과가 도출될 수 있다. 이러한 노이즈는 FF 처리된 조직의 전장유전체 분석 데이터에서는 발견되지 않는 것이 일반적이다. 따라서, FFPE 처리된 조직으로부터 왜곡되지 않은 분석 결과를 도출하기 위해 변이 검출 데이터 상의 노이즈를 효과적으로 처리 또는 제거할 필요가 있다.

발명의 내용

해결하려는 과제

[0007] 본 개시는 상기와 같은 문제점을 해결하기 위한 기계학습 모델을 학습시키는 방법, 기록매체에 저장된 컴퓨터 프로그램 및 장치(시스템)를 제공한다.

과제의 해결 수단

[0008] 본 개시는 방법, 시스템(장치) 또는 판독 가능 저장 매체에 저장된 컴퓨터 프로그램을 포함한 다양한 방식으로 구현될 수 있다.

[0009] 적어도 하나의 프로세서에 의해 실행되는, 기계학습 모델을 학습시키는 방법에 있어서, 참조 샘플의 참조 변이 후보 정보를 획득하는 단계, 참조 변이 후보와 연관된 어노테이션 정보를 생성하는 단계, 획득된 참조 변이 후보 정보 및 생성된 어노테이션 정보에 기초하여 학습 데이터를 생성하는 단계 및 생성된 학습 데이터를 이용하여 기계학습 모델을 학습시키는 단계를 포함한다.

[0010] 본 개시의 일 실시예에 있어서, 참조 샘플은, 동일한 개체로부터 채취된 참조 정상 샘플 및 참조 이상 샘플을 포함하고, 획득된 참조 변이 후보 정보는, 변이 검출 모듈을 이용하여, 참조 정상 샘플과 연관된 제1 참조 시퀀싱 데이터 및 참조 이상 샘플과 제2 참조 시퀀싱 데이터를 기초로 결정된다.

[0011] 본 개시의 일 실시예에 있어서, 변이 검출 모듈은 복수의 검출 모듈을 포함하고, 획득된 참조 변이 후보 정보는, 복수의 검출 모듈을 이용하여 획득된 참조 변이 서브 후보 정보를 통합(union)함으로써 결정되고, 획득된 참조 변이 서브 후보 정보는, 복수의 검출 모듈의 각각에 제1 참조 시퀀싱 데이터 및 제2 참조 시퀀싱 데이터를 적용함으로써 결정된다.

[0012] 본 개시의 일 실시예에 있어서, 어노테이션 정보를 생성하는 단계는, 매핑된 위치 중 적어도 일부가 참조 변이 후보의 위치와 중첩되는 복수의 리드(read)를 결정하는 단계 및 결정된 복수의 리드와 연관된 제1 어노테이션 정보를 생성하는 단계를 포함한다.

[0013] 본 개시의 일 실시예에 있어서, 복수의 리드는 레퍼런스 지놈과 상이한 복수의 변이 리드(variant read)를 포함하고, 제1 어노테이션 정보는, 복수의 변이 리드의 인서트 사이즈(insert size)의 최솟값, 복수의 변이 리드의 인서트 사이즈의 최댓값, 또는 복수의 변이 리드 중 특정 조건을 만족하는 페어드 리드(paired read)의 개수 중 적어도 하나를 포함하고, 특정 조건은, 페어드 리드의 제1 리드 및 제2 리드가 각각 정방향과 역방향으로 정렬되고, 페어드 리드의 인서트 사이즈가 하한임계치와 상한임계치 사이인 조건을 포함한다.

[0014] 본 개시의 일 실시예에 있어서, 어노테이션 정보를 생성하는 단계는, 복수의 정상 샘플과 연관된 복수의 시퀀싱 데이터로부터 생성된 정상 조직 유전체 데이터(PON: Panel of Normals)를 수신하는 단계 및 정상 조직 유전체 데이터와 연관된 제2 어노테이션 정보를 생성하는 단계를 포함한다.

[0015] 본 개시의 일 실시예에 있어서, 어노테이션 정보를 생성하는 단계는, FFPE(Formalin-Fixed, Paraffin-Embedded) 처리된 복수의 샘플과 연관된 복수의 시퀀싱 데이터로부터 생성된 FFPE 처리 조직 유전체 데이터(POF: Panel of FFPEs)를 수신하는 단계 및 FFPE 처리 조직 유전체 데이터와 연관된 제3 어노테이션 정보를 생성하는 단계를 포함한다.

[0016] 본 개시의 일 실시예에 있어서, 제3 어노테이션 정보는, FFPE 처리된 복수의 샘플 중, 샘플 내 염기 서열 상의 특정 위치(position)에 대한 VAF(Variant Allele Frequency)가 미리 정해진 임계치 미만인 샘플의 수를 포함한다.

[0017] 본 개시의 일 실시예에 있어서, 제3 어노테이션 정보는, FFPE 처리된 복수의 샘플 중, 샘플 내 염기 서열 상의 특정 위치에서 미리 정해진 개수의 변이 리드를 갖는 샘플의 수를 포함한다.

[0018] 본 개시의 일 실시예에 있어서, 어노테이션 정보를 생성하는 단계는, 참조 변이 후보의 변이 유형과 연관된 정보 및 참조 변이 후보의 시퀀스 컨텍스트(sequence context) 정보를 포함하는 제4 어노테이션 정보를 생성하는 단계를 포함한다.

[0019] 본 개시의 일 실시예에 있어서, 학습 데이터를 생성하는 단계는, 참조 변이 후보에 대한 분류 정보를 레이블링하는 단계를 포함한다.

[0020] 본 개시의 일 실시예에 있어서, 참조 샘플은 FFPE 처리된 샘플이고, 레이블링하는 단계는, 참조 변이 후보 정보의 적어도 일부와, FF(Fresh-Frozen) 처리된 샘플 내 변이 후보 중 어느 하나의 변이 후보와 연관된 정보의 적어도 일부가 서로 대응되는 것으로 판단되는 것에 응답하여, 참조 변이 후보가 진양성(True Positive) 변이인 것으로 레이블링하는 단계를 포함하고, FF 처리된 샘플은 FFPE 처리된 샘플에 대응되는 샘플이다.

[0021] 본 개시의 일 실시예에 있어서, 레이블링하는 단계는, 참조 변이 후보 정보와, FF 처리된 샘플 내 변이 후보 중 어느 하나의 변이 후보와 연관된 정보가 서로 대응되지 않는 것으로 판단되는 것에 응답하여, 참조 변이 후보가 위양성(False positive) 변이인 것으로 레이블링하는 단계를 더 포함한다.

[0022] 본 개시의 일 실시예에 있어서, 학습 데이터를 생성하는 단계는, 참조 변이 후보 정보 및 어노테이션 정보에 기초하여, 참조 변이 후보의 특징을 추출하는 단계 및 참조 변이 후보 정보, 추출된 참조 변이 후보의 특징 및 레이블링된 분류 정보를 포함하는 데이터 세트를 학습 데이터에 포함시키는 단계를 더 포함한다.

[0023] 본 개시의 일 실시예에 있어서, 기계학습 모델은 복수의 분류기(classifier)를 포함하고, 기계학습 모델을 학습시키는 단계는, 참조 변이 후보 정보 및 참조 변이 후보의 특징을 복수의 분류기의 각각에 입력하는 단계, 복수의 분류기 중 적어도 하나의 분류기로부터의 출력 결과를 이용하여 참조 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 결정하는 단계 및 분류 결과와 참조 변이 후보에 레이블링된 분류 정보에 기초하여 기계학습 모델의 파라미터를 조정하는 단계를 포함한다.

[0024] 본 개시의 일 실시예에 있어서, 기계학습 모델은, 타겟 샘플 내 타겟 변이 후보 정보 및 타겟 변이 후보의 특징을 수신하여, 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 출력한다.

[0025] 본 개시의 일 실시예에 있어서, 타겟 샘플은, 동일한 개체로부터 채취된 타겟 정상 샘플 및 타겟 이상 샘플을 포함하고, 타겟 변이 후보 정보는, 변이 검출 모듈을 이용하여, 타겟 정상 샘플과 연관된 제1 타겟 시퀀싱 데이터 및 타겟 이상 샘플과 연관된 제2 타겟 시퀀싱 데이터를 기초로 결정된다.

[0026] 본 개시의 일 실시예에 있어서, 타겟 샘플은 FFPE 처리된 샘플이다.

[0027] 본 개시의 일 실시예에 따른 적어도 하나의 프로세서에 의해 실행되는, 세포 샘플 내 진양성 변이 검출을 통한 유전체 프로파일링 방법에 있어서, 타겟 샘플의 타겟 변이 후보 정보를 획득하는 단계, 기계학습 모델을 이용하여, 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 결정하는 단계 및 결정된 분류 결과를 기초로 타겟 샘플에 대한 유전체 프로파일링(Genomic Profiling)을 수행하는 단계를 포함하고, 기계학습 모델은, 참조 샘플의 참조 변이 후보 정보 및 참조 변이 후보와 연관된 어노테이션 정보를 이용하여, 참조 변이 후보가 진양성 변이인지 여부를 결정하도록 학습된다.

[0028] 본 개시의 일 실시예에 있어서, 유전체 프로파일링의 수행 결과에 기초하여, 타겟 샘플이 채취된 개체의 질병 진단 정보, 치료 전략 정보, 예후 예측 정보 또는 약물 반응성 예측 정보 중 적어도 하나를 제공하는 단계를 더 포함한다.

[0029] 본 개시의 일 실시예에 따른 기계학습 모델을 학습시키는 방법 및/또는 세포 샘플 내 진양성 변이 검출을 통한 유전체 프로파일링 방법을 컴퓨터에서 실행하기 위해 컴퓨터 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램이 제공된다.

[0030] 본 개시의 일 실시예에 따른 장치에 있어서, 메모리 및 메모리와 연결되고, 메모리에 포함된 컴퓨터 판독 가능한 적어도 하나의 프로그램을 실행하도록 구성된 적어도 하나의 프로세서를 포함하고, 적어도 하나의 프로그램은, 참조 샘플의 참조 변이 후보 정보를 획득하고, 참조 변이 후보 정보와 연관된 어노테이션 정보를 생성하고, 획득된 참조 변이 후보 정보 및 생성된 어노테이션 정보에 기초하여 학습 데이터를 생성하고, 생성된 학습 데이터를 이용하여 기계학습 모델을 학습시키기 위한 명령어를 포함한다.

발명의 효과

[0031] 본 개시의 다양한 실시예에 따르면, 이상 샘플 내 특정 변이 후보가 위양성(false positive) 변이인 것으로 판단되는 경우, 변이 후보 리스트로부터 해당 특정 변이 후보가 삭제/필터링됨으로써, 정확도 높은 변이 리스트가 결정될 수 있다.

[0032] 본 개시의 다양한 실시예에 따르면, FFPE 처리된 조직을 대상으로 전장유전체 분석을 수행하는 경우에 발생할 수 있는 노이즈 내지 오차를 보정함으로써, FF 처리된 조직의 전장유전체 분석 데이터에서와 같이 왜곡되지 않은 분석 결과가 도출될 수 있다.

[0033] 본 개시의 다양한 실시예에 따르면, 의료기관 및 인체유래물은행(Biobank) 등이 확보, 축적하고 있는 방대한 양의 FFPE 조직들을 대상으로 전장유전체 분석을 높은 정확도로 시행할 수 있을 뿐만 아니라, 통상적인 진료 현장에 구비된 시설만으로 의료기관의 조직 검체 처리 절차에 대한 변경 없이도 환자 등의 조직 검체에 대한 전장유전체 분석이 진행될 수 있다.

[0034] 본 개시의 다양한 실시예에 따르면, 유전체 프로파일링은 질병의 조기 진단, 개인 맞춤형 의학, 유전 질환 연구, 약물 개발 등 다양한 분야에서 응용될 수 있으며, 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 기초로 유전체 프로파일링을 수행함으로써, 의학 연구와 임상적 응용 등에 있어 정밀하고 광범위한 유전 정보가 제공될 수 있다.

[0035] 본 개시의 효과는 이상에서 언급한 효과로 제한되지 않으며, 언급되지 않은 다른 효과들은 청구범위의 기재로부터 본 개시가 속하는 기술분야에서 통상의 지식을 가진 자("통상의 기술자"라 함)에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0036] 본 개시의 실시예들은, 이하 설명하는 첨부 도면들을 참조하여 설명될 것이며, 여기서 유사한 참조 번호는 유사한 요소들을 나타내지만, 이에 한정되지는 않는다.

도 1은 본 개시의 일 실시예에 따른 기계학습 모델을 이용하여 변이 후보의 분류 결과가 결정되는 예시를 나타내는 도면이다.

도 2는 본 개시의 일 실시예에 따른 기계학습 모델의 학습 프로세스 및 추론 프로세스를 수행하기 위한 컴퓨팅 장치의 내부 구성을 나타내는 블록도이다.

도 3은 본 개시의 일 실시예에 따른 변이 검출 모듈의 예시를 나타내는 도면이다.

도 4는 본 개시의 일 실시예에 따른 어노테이션 모듈의 예시를 나타내는 도면이다.

도 5는 본 개시의 일 실시예에 따른 특징 추출 모듈의 예시를 나타내는 도면이다.

도 6은 본 개시의 일 실시예에 따른 기계학습 모델을 이용하여 변이 후보의 분류 결과가 출력되는 예시를 나타내는 도면이다.

도 7은 본 개시의 일 실시예에 따른 기계학습 모델의 세부 구성을 나타내는 도면이다.

도 8은 본 개시의 일 실시예에 따른 기계학습 모델의 학습 프로세스의 예시를 나타내는 도면이다.

도 9는 본 개시의 일 실시예에 따른 기계학습 모델의 추론 프로세스의 예시를 나타내는 도면이다.

도 10은 본 개시의 일 실시예에 따른 학습 데이터의 예시를 나타내는 도면이다.

도 11은 본 개시의 일 실시예에 따른 학습된 기계학습 모델의 성능 평가 결과를 나타내는 도면이다.

도 12는 본 개시의 일 실시예에 따른 인공신경망 모델을 나타내는 예시도이다.

도 13은 본 개시의 일 실시예에 따른 기계학습 모델을 학습시키는 방법을 나타내는 흐름도이다.

도 14는 본 개시의 일 실시예에 따른 세포 샘플 내 진양성 변이 검출을 통한 유전체 프로파일링 방법을 나타내는 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0037] 이하, 본 개시의 실시를 위한 구체적인 내용을 첨부된 도면을 참조하여 상세히 설명한다. 다만, 이하의 설명에서는 본 개시의 요지를 불필요하게 흐릴 우려가 있는 경우, 널리 알려진 기능이나 구성에 관한 구체적 설명은 생략하기로 한다.

[0038] 첨부된 도면에서, 동일하거나 대응하는 구성요소에는 동일한 참조부호가 부여되어 있다. 또한, 이하의 실시예들의 설명에 있어서, 동일하거나 대응되는 구성요소를 중복하여 기술하는 것이 생략될 수 있다. 그러나, 구성요소에 관한 기술이 생략되어도, 그러한 구성요소가 어떤 실시예에 포함되지 않는 것으로 의도되지는 않는다.

[0039] 개시된 실시예의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 개시는 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로

다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 개시가 완전하도록 하고, 본 개시가 통상의 기술자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것일 뿐이다.

[0040] 본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 개시된 실시예에 대해 구체적으로 설명하기로 한다. 본 명세서에서 사용되는 용어는 본 개시에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 관련 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서, 본 개시에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 개시의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0041] 본 명세서에서의 단수의 표현은 문맥상 명백하게 단수인 것으로 특정하지 않는 한, 복수의 표현을 포함한다. 또한, 복수의 표현은 문맥상 명백하게 복수인 것으로 특정하지 않는 한, 단수의 표현을 포함한다. 명세서 전체에서 어떤 부분이 어떤 구성요소를 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다.

[0042] 또한, 명세서에서 사용되는 '모듈' 또는 '부'라는 용어는 소프트웨어 또는 하드웨어 구성요소를 의미하며, '모듈' 또는 '부'는 어떤 역할들을 수행한다. 그렇지만, '모듈' 또는 '부'는 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. '모듈' 또는 '부'는 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서, '모듈' 또는 '부'는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 또는 변수들 중 적어도 하나를 포함할 수 있다. 구성요소들과 '모듈' 또는 '부'들은 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 '모듈' 또는 '부'들로 결합되거나 추가적인 구성요소들과 '모듈' 또는 '부'들로 더 분리될 수 있다.

[0043] 본 개시의 일 실시예에 따르면, '모듈' 또는 '부'는 프로세서 및 메모리로 구현될 수 있다. '프로세서'는 범용 프로세서, 중앙 처리 장치(CPU), 마이크로프로세서, 디지털 신호 프로세서(DSP), 제어기, 마이크로제어기, 상태 머신 등을 포함하도록 넓게 해석되어야 한다. 몇몇 환경에서, '프로세서'는 주문형 반도체(ASIC), 프로그램가능 로직 디바이스(PLD), 필드 프로그램가능 게이트 어레이(FPGA) 등을 지칭할 수도 있다. '프로세서'는, 예를 들어, DSP와 마이크로프로세서의 조합, 복수의 마이크로프로세서들의 조합, DSP 코어와 결합한 하나 이상의 마이크로프로세서들의 조합, 또는 임의의 다른 그러한 구성들의 조합과 같은 처리 디바이스들의 조합을 지칭할 수도 있다. 또한, '메모리'는 전자 정보를 저장 가능한 임의의 전자 컴포넌트를 포함하도록 넓게 해석되어야 한다. '메모리'는 임의 액세스 메모리(RAM), 판독-전용 메모리(ROM), 비-휘발성 임의 액세스 메모리(NVRAM), 프로그램가능 판독-전용 메모리(PROM), 소거-프로그램가능 판독 전용 메모리(EPROM), 전기적으로 소거가능 PROM(EEPROM), 플래쉬 메모리, 자기 또는 광학 데이터 저장장치, 레지스터들 등과 같은 프로세서-판독가능 매체의 다양한 유형들을 지칭할 수도 있다. 프로세서가 메모리로부터 정보를 판독하고/하거나 메모리에 정보를 기록할 수 있다면 메모리는 프로세서와 전자 통신 상태에 있다고 불린다. 프로세서에 집적된 메모리는 프로세서와 전자 통신 상태에 있다.

[0044] 본 개시에서, '시스템'은 서버 장치와 클라우드 장치 중 적어도 하나의 장치를 포함할 수 있으나, 이에 한정되는 것은 아니다. 예를 들어, 시스템은 하나 이상의 서버 장치로 구성될 수 있다. 다른 예로서, 시스템은 하나 이상의 클라우드 장치로 구성될 수 있다. 또 다른 예로서, 시스템은 서버 장치와 클라우드 장치가 함께 구성되어 동작될 수 있다.

[0045] 또한, 이하의 실시예들에서 사용되는 제1, 제2, A, B, (a), (b) 등의 용어는 어떤 구성요소를 다른 구성요소와 구별하기 위해 사용되는 것일 뿐, 그 용어에 의해 해당 구성요소의 본질이나 차례 또는 순서 등이 한정되지는 않는다.

[0046] 또한, 이하의 실시예들에서, 어떤 구성요소가 다른 구성요소에 '연결', '결합' 또는 '접속'된다고 기재된 경우, 그 구성요소는 그 다른 구성요소에 직접적으로 연결되거나 또는 접속될 수 있지만, 각 구성요소 사이에 또 다른 구성요소가 '연결', '결합' 또는 '접속'될 수도 있다고 이해되어야 한다.

[0047] 또한, 이하의 실시예들에서 사용되는 '포함한다(comprises)' 및/또는 '포함하는(comprising)'은 언급된 구성요소, 단계, 동작 및/또는 소자는 하나 이상의 다른 구성요소, 단계, 동작 및/또는 소자의 존재 또는 추가를 배제하지 않는다.

[0048] 본 개시에서, '복수의 A 각각' 은 복수의 A에 포함된 모든 구성 요소의 각각을 지칭하거나, 복수의 A에 포함된 일부 구성 요소의 각각을 지칭할 수 있다.

[0049] 본 개시의 다양한 실시예들을 설명하기에 앞서, 사용되는 용어에 대하여 설명하기로 한다.

[0050] 본 개시에서 '전장 유전체 시퀀싱(Whole Genome Sequencing: WGS)' 또는 '전체 게놈 시퀀싱'은 게놈의 전체 DNA 서열을 결정하는 데 사용되는 기술을 지칭할 수 있다. 구체적으로, 전장 유전체 시퀀싱은 사람 또는 유기체의 전체 유전 물질 세트 내에서 뉴클레오티드 염기(아데닌, 시토신, 구아닌 및 티민)의 순서를 해독하고 식별하는 것과 연관될 수 있다. 이 때, 전체 유전 물질 세트는 모든 유전자, 비암호화 영역 및 게놈 내에 존재하는 임의의 추가 유전 요소를 포함할 수 있다. 일 실시예에서, 전장 유전체 시퀀싱은 여러 단계를 거쳐 수행될 수 있다. 예를 들어, 전장 유전체 시퀀싱은 특정 세포 등에서 DNA를 추출한 뒤, 추출된 DNA를 더 작은 조각으로 분할하고, 이에 대해 "리드(reads)"로 지칭되는 수백만 또는 수십억 개의 짧은 DNA 시퀀스를 생성함으로써 수행될 수 있다. 생성된 리드는 전체 게놈 시퀀스를 재구성하기 위해 정렬 및 조립될 수 있다.

[0051] 본 개시에서, '시퀀싱 데이터'는 시퀀싱 프로세스를 통해 분석된, 특정 개체의 DNA(Deoxyribo Nucleic Acid) 서열 또는 RNA(RiboNucleic Acid) 서열과 연관된 데이터를 지칭할 수 있다.

[0052] 본 개시에서, 'X 샘플 시퀀싱 데이터(sequencing data)'는 'X 샘플'에 대한 시퀀싱 프로세스를 통해 생성된 시퀀싱 데이터를 지칭할 수 있다.

[0053] 본 개시에서, '이상 세포'는 정상 세포와는 다른 크기, 모양, 구조, 기능 등을 가지는 비정상 세포를 지칭할 수 있고, '이상 샘플'은 이상 세포를 포함하는 샘플을 지칭할 수 있다. 이상 세포는 유전적 돌연변이, 감염, 독소 노출 등 다양한 요인으로 발생할 수 있으며, 암세포, 종양 세포, 괴사 세포, 노화 세포, 이수성 세포(Aneuploid Cell), 과형성 세포(Hyperplastic Cell), 비대 세포(Hypertrophic Cell) 등 다양한 종류의 비정상 세포를 포함할 수 있다.

[0054] 본 개시에서, '변이'는 단일 염기 변이(Single-Nucleotide Variant, 이하 “SNV”라고 지칭될 수 있다)와 염기 삽입 변이 및 결실 변이(short insertion-and-deletion, 이하 “INDEL”이라고 지칭될 수 있다)을 아우르는 점 돌연변이(Point Mutation), 구조적 변이(structural variation) 및/또는 유전자 복제 수 변이(CNV: Copy-Number Variant) 등 다양한 유형의 변이를 지칭할 수 있다.

[0055] 본 개시에서, '변이 후보'는 변이(Mutation)에 해당할 확률이 미리 정해진 임계 확률 이상인 DNA 또는 RNA 시퀀스를 지칭할 수 있다. 예를 들어, 변이 후보는, 점 돌연변이에 해당할 확률이 미리 정해진 임계 확률 이상인 시퀀스일 수 있다.

[0056] 본 개시에서, '어노테이션(annotation) 정보'는 변이 후보의 특징 추출을 위한 정보를 지칭할 수 있으며, 본 개시에서 기계학습 모델의 학습을 위해 데이터에 부여되는 교사 신호(정답)인 '레이블' 내지 '분류 정보'과는 구별될 수 있다.

[0057] 본 개시에서, '시퀀스 컨텍스트(sequence context)'는 특정 DNA 또는 RNA 염기서열을 둘러싸는 하나 이상의 이웃 뉴클레오타이드를 포함하는 염기서열을 지칭할 수 있다.

[0058] 본 개시에서, '통합(union)'은 합집합 연산을 지칭할 수 있다.

[0059] 본 개시에서, 유전체 프로파일링(Genomic profiling)은, 개체의 유전체 내지 DNA 시퀀스를 분석함으로써 유전적 변이, 구조적 변화, 유전자 발현 패턴 및/또는 기타 유전 정보를 조사하는 프로세스를 지칭할 수 있다.

[0060] 추론 과정에서 이용되는 용어 '타겟 X'와 대응하여, 학습 과정에서 이용되는 용어는 '참조 X'로 정의될 수 있다. 예를 들어, '타겟 샘플'은 학습된 기계학습 모델을 이용하여 샘플 내 타겟 변이 후보가 진양성 변이인지 여부를 추론하기 위한 대상이 되는 샘플일 수 있고, '참조 샘플'은 기계학습 모델의 학습에 이용되는 학습 데이터를 생성하기 위해 시퀀싱되는 샘플일 수 있다. 한편, 기계학습 모델의 학습 프로세스 또는 추론 프로세스와 연관되고 '타겟' 또는 '참조' 표현 없이 사용된 용어 'X'는, 반대되는 기재가 없는 한 '타겟 X' 및/또는 '참조 X'를 지칭할 수 있으며, 용어가 사용된 맥락에 따라 해석되어야 할 것이다.

[0061] 본 개시에서, 'X와 연관된 정보' 및 'X 정보'는 서로 동일한 의미로 혼용될 수 있다.

[0062] 이하, 본 개시의 다양한 실시예들에 대하여 첨부된 도면에 따라 상세하게 설명한다.

[0063] 도 1은 본 개시의 일 실시예에 따른 기계학습 모델(130)을 이용하여 변이 후보의 분류 결과(132)가 결정되는 예

시를 나타내는 도면이다.

[0064] 개체(110)는 종양 조직(예를 들어, 암 조직) 등 변이된 조직 내지 세포를 가지는 개체일 수 있다. 개체(110)는 기계학습 모델(130)의 학습 프로세스(learning)에서 학습 데이터 생성의 기초가 되는 참조 샘플을 제공하는 개체이거나, 학습된 기계학습 모델(130)을 이용한 추론 프로세스(inference)에서 세포 샘플 내 진양성 변이(true positive mutation)를 결정하기 위한 타겟(target) 개체일 수 있다. 개체(110)는 사람에 한정되지 않고, 임의의 생물체일 수 있다.

[0065] 개체(110)로부터, 종양 조직 생검(tumor tissue biopsy) 샘플 등 이상 샘플이 채취될 수 있다. 이상 샘플은 변이 검출의 대상이 되는 이상 세포를 포함하는 샘플일 수 있다.

[0066] 또한, 이상 샘플이 채취된 개체와 동일한 개체(110)로부터 정상 샘플이 채취될 수 있다. 정상 샘플에는 이상 세포가 포함되어 있지 않다고 가정될 수 있다. 정상 샘플은 정상 혈액 샘플 또는 정상 세포 샘플 등을 포함할 수 있다. 일 예시로, 정상 혈액 샘플은 개체(110)로부터 채취되어 원심분리된 샘플의 하단에 있는 적혈구 층과 상단에 있는 혈장 층 사이에 형성되는 얇은 버피 코트(buffy coat)일 수 있다.

[0067] 개체(110)로부터 채취된 이상 샘플 및 정상 샘플은, FFPE(Formalin-Fixed, Paraffin-Embedded) 처리되거나 FF(Fresh-Frozen) 처리될 수 있다. 이후, FFPE 또는 FF 처리된 이상 샘플로부터 이상 샘플 시퀀싱 데이터(112)가 생성되고, FFPE 또는 FF 처리된 정상 샘플로부터 정상 샘플 시퀀싱 데이터(114)가 생성될 수 있다. 여기서, 시퀀싱 데이터(112, 114)는 전체 게놈 시퀀싱(Whole Genome Sequencing: WGS) 및/또는 타겟 패널 시퀀싱(Target Panel Sequencing: TPS)을 통해 획득될 수 있다.

[0068] 시퀀싱 데이터(112, 114)는 기계학습 모델(130)의 학습 프로세스에서 학습 데이터 생성의 기초가 되는 참조 시퀀싱 데이터이거나, 학습된 기계학습 모델(130)을 이용한 추론 프로세스에서 세포 샘플 내 진양성 변이(true positive mutation)를 검출/결정하기 위해 분석되는 타겟 시퀀싱 데이터일 수 있다.

[0069] 예를 들어, 시퀀싱 데이터(112, 114)는 기계학습 모델(130)의 학습 프로세스를 설명하는 도 8에서, FFPE 샘플 시퀀싱 데이터(812) 및 FF 샘플 시퀀싱 데이터(814)와 대응될 수 있다. 다른 예에서, 시퀀싱 데이터(112, 114)는 학습된 기계학습 모델(130)의 추론 프로세스를 설명하는 도 9에서, FFPE 샘플 시퀀싱 데이터(910)와 대응될 수 있다. 이에 대해서는 각각 도 8 및 도 9에서 자세히 후술한다.

[0070] 변이 검출 모듈(120)은 이상 샘플 시퀀싱 데이터(112) 및 정상 샘플 시퀀싱 데이터(114)를 이용하여 변이 후보 정보(122)를 결정할 수 있다. 예를 들어, 정상 샘플 시퀀싱 데이터(114)는 이상 샘플 시퀀싱 데이터(112)와 대조되는 시퀀싱 데이터로 사용됨으로써, 이상 샘플 내 DNA/RNA에 존재하는 유전적 변이 또는 돌연변이가 비교되고 식별될 수 있다. 이와 달리, 이상 샘플을 이용한 딥 시퀀싱(Deep Sequencing), 알려진 변이 데이터베이스와의 대조 등을 수행함으로써, 정상 샘플 시퀀싱 데이터(114)를 이용하지 않고 이상 샘플 시퀀싱 데이터(112)만을 이용하여 이상 샘플 내 DNA/RNA에 존재하는 유전적 변이 또는 돌연변이가 식별될 수도 있다.

[0071] 변이 검출 모듈(120)에서 결정되는 변이 후보 정보(122)는 변이 후보의 위치 정보, 변이 후보의 위치에서의 기준 대립유전자(reference allele) 정보, 변이 후보에 대응되는 변형된 대립유전자(altered allele) 정보 등을 포함할 수 있다. 변이 후보의 위치 정보는, 변이 후보가 위치한 염색체 정보(예: 염색체 번호), 염색체 내 위치 정보(예: 1-based position) 등을 포함할 수 있다. 변이 검출 모듈(120)에 대해서는 도 3을 이용하여 자세히 후술한다.

[0072] 기계학습 모델(130)은 변이 후보 정보(122)에 기초하여, 변이 후보가 진양성(true positive) 변이인지 여부를 나타내는 분류 결과(132)를 출력할 수 있다. 이상 샘플 내 존재하는 변이로 추정되는 복수의 변이 후보를 포함하는 변이 후보 리스트 중, 특정 변이 후보가 위양성(false positive) 변이인 것으로 판단되는 경우, 변이 후보 리스트로부터 해당 특정 변이 후보가 삭제/필터링될 수 있다. 이를 통해, 이상 샘플 내 실제 변이로 추정되는 변이를 포함하는 변이 리스트가 결정될 수 있다.

[0073] 기계학습 모델(130)의 세부 구성 및 동작에 대해서는 도 6, 도 7 및 도 12 등을 이용하여 자세히 후술한다.

[0074] 도 2는 본 개시의 일 실시예에 따른 기계학습 모델의 학습 프로세스 및 추론 프로세스를 수행하기 위한 컴퓨팅 장치(200)의 내부 구성을 나타내는 블록도이다. 컴퓨팅 장치(200)는 메모리(210), 프로세서(220), 통신 모듈(230) 및 입출력 인터페이스(240)를 포함할 수 있다. 일 실시예에서, 컴퓨팅 장치(200)는 복수의 분산 컴퓨팅 장치로 구성될 수 있으며, 도 2에 도시된 메모리(210), 프로세서(220), 통신 모듈(230) 및 입출력 인터페이스(240)의 각각은 복수의 분산 컴퓨팅 장치에 포함된 복수의 메모리, 복수의 프로세서 등을 포괄하여 지칭하는 것

일 수 있다.

[0075] 도 2에 도시된 바와 같이, 컴퓨팅 장치(200)는 통신 모듈(230)을 이용하여 네트워크를 통해 정보 및/또는 데이터를 통신할 수 있도록 구성될 수 있다. 일 실시예에서, 컴퓨팅 장치(200)는 통신 모듈(230)을 이용하여 네트워크를 통해 외부 데이터베이스 등과 정보 및/또는 데이터를 통신할 수 있도록 구성될 수 있다. 예를 들어, 컴퓨팅 장치(200)는 복수의 정상 샘플과 연관된 복수의 시퀀싱 데이터로부터 생성된 정상 조직 유전체 데이터(PON: Panel of Normals)를 포함하는 데이터베이스(가령, 도 6의 데이터베이스(630)와 대응), 특징 추출 정보를 포함하는 데이터베이스(가령, 도 7의 데이터베이스(750)와 대응)와 연결됨으로써, 정보 및/또는 데이터를 서로 송수신할 수 있다.

[0076] 메모리(210)는 비-일시적인 임의의 컴퓨터 판독 가능한 기록매체를 포함할 수 있다. 일 실시예에 따르면, 메모리(210)는 RAM(random access memory), ROM(read only memory), 디스크 드라이브, SSD(solid state drive), 플래시 메모리(flash memory) 등과 같은 비소멸성 대용량 저장 장치(permanent mass storage device)를 포함할 수 있다. 다른 예로서, ROM, SSD, 플래시 메모리, 디스크 드라이브 등과 같은 비소멸성 대용량 저장 장치는 메모리와는 구분되는 별도의 영구 저장 장치로서 컴퓨팅 장치(200)에 포함될 수 있다. 또한, 메모리(210)에는 운영체제와 적어도 하나의 프로그램 코드가 저장될 수 있다.

[0077] 이러한 소프트웨어 구성요소들은 메모리(210)와는 별도의 컴퓨터에서 판독 가능한 기록매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록매체는 이러한 컴퓨팅 장치(200)에 직접 연결가능한 기록 매체를 포함할 수 있는데, 예를 들어, 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독 가능한 기록매체를 포함할 수 있다. 다른 예로서, 소프트웨어 구성요소들은 컴퓨터에서 판독 가능한 기록매체가 아닌 통신 모듈(230)을 통해 메모리(210)에 로딩될 수도 있다. 예를 들어, 적어도 하나의 프로그램은 개발자들 또는 어플리케이션의 설치 파일을 배포하는 파일 배포 시스템이 통신 모듈(230)을 통해 제공하는 파일들에 의해 설치되는 컴퓨터 프로그램에 기반하여 메모리(210)에 로딩될 수 있다.

[0078] 프로세서(220)는 기본적인 산술, 로직 및 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령, 데이터 등을 처리하도록 구성될 수 있다. 명령은 메모리(210) 또는 통신 모듈(230)에 의해 사용자 단말(미도시) 또는 다른 외부 시스템으로 제공될 수 있다. 또한, 프로세서(220)는 복수의 사용자 단말 및/또는 복수의 외부 시스템으로부터 수신된 정보 및/또는 데이터를 관리, 처리 및/또는 저장하도록 구성될 수 있다.

[0079] 통신 모듈(230)은 네트워크를 통해 사용자 단말(미도시)과 컴퓨팅 장치(200)가 서로 통신하기 위한 구성 또는 기능을 제공할 수 있으며, 컴퓨팅 장치(200)가 외부 시스템과 통신하기 위한 구성 또는 기능을 제공할 수 있다.

[0080] 또한, 컴퓨팅 장치(200)의 입출력 인터페이스(240)는 컴퓨팅 장치(200)와 연결되거나 컴퓨팅 장치(200)가 포함할 수 있는 입력 또는 출력을 위한 장치(미도시)와의 인터페이스를 위한 수단일 수 있다. 도 2에서는 입출력 인터페이스(240)가 프로세서(220)와 별도로 구성된 요소로서 도시되었으나, 이에 한정되지 않으며, 입출력 인터페이스(240)가 프로세서(220)에 포함되도록 구성될 수 있다.

[0081] 컴퓨팅 장치(200)는 도 2의 구성요소들보다 더 많은 구성요소들을 포함할 수 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다.

[0082] 프로세서(220)는 도 3 내지 도 9 등에 도시되는 변이 검출 모듈, 필터 모듈, 어노테이션 모듈, 레이블링 모듈, 트레이닝 모듈 등 다양한 종류의 모듈을 포함할 수 있다.

[0083] 본 명세서에서는 먼저 도 3 내지 도 7을 이용하여, 기계학습 모델의 학습 프로세스 및 추론 프로세스에서 이용될 수 있는 변이 검출 모듈, 어노테이션 모듈, 특징 추출 모듈 및 기계학습 모델 각각의 동작을 구체적으로 설명한다. 그런 다음, 도 8을 이용하여 기계학습 모델의 학습 프로세스를 설명하고, 도 9를 이용하여 기계학습 모델의 추론 프로세스를 설명한다.

[0084] 도 3은 본 개시의 일 실시예에 따른 변이 검출 모듈(310)의 예시를 나타내는 도면이다. 변이 검출 모듈(310)은 특정 샘플과 연관된 시퀀싱 데이터(300)로부터 특정 샘플 내 변이 후보 정보를 결정할 수 있다. '변이 후보'는 특정 샘플 내 포함된, 변이로 추정되는 시퀀스일 수 있다.

[0085] 변이 검출 모듈(310)은 기계학습 모델의 학습 프로세스 및 추론 프로세스에서 이용될 수 있다. 예를 들어, 변이 검출 모듈(310)은, 기계학습 모델의 학습 프로세스에서, 참조 샘플과 연관된 참조 시퀀싱 데이터로부터 참조 샘플 내 참조 변이 후보 정보를 결정할 수 있다. 이와 유사하게, 변이 검출 모듈(310)은 기계학습 모델의 추론 프로세스에서, 타겟 샘플과 연관된 타겟 시퀀싱 데이터로부터 타겟 샘플 내 타겟 변이 후보 정보를 결정할 수

있다.

[0086] 시퀀싱 데이터(300)는 특정 개체로부터 채취된 이상 샘플과 연관된 시퀀싱 데이터(예를 들어, 도 1의 112) 및 동일한 개체로부터 채취된 정상 샘플과 연관된 시퀀싱 데이터(예를 들어, 도 1의 114)를 포함할 수 있다. 예를 들어, 변이 검출 모듈(310)은, 정상 샘플에는 이상 세포가 포함되어 있지 않다고 가정함으로써, 이상 샘플과 연관된 시퀀싱 데이터와 정상 샘플과 연관된 시퀀싱 데이터를 대조하여 변이 후보 정보(340)를 결정할 수 있다.

[0087] 이상 샘플 및 정상 샘플은, FFPE(Formalin-Fixed, Paraffin-Embedded) 처리된 샘플이거나 FF(Fresh-Frozen) 처리된 샘플일 수 있다. 샘플을 FFPE 처리하는 경우, DNA/RNA에 다양한 유형의 손상이 발생할 수 있다. 이로 인해, FF 처리된 샘플의 시퀀싱 데이터를 이용하여 결정된 변이 후보의 개수보다, FFPE 처리된 샘플의 시퀀싱 데이터를 이용하여 결정된 변이 후보의 개수가 더 많을 수 있다. 즉, FFPE 처리된 샘플의 시퀀싱 데이터를 이용하여 결정된 변이 후보는, FF 처리된 샘플의 시퀀싱 데이터를 이용하여 결정된 변이 후보 및 FFPE 처리 및 보관 과정에서 발생하는 아티팩트(artifact)를 포함할 수 있다. FFPE 처리 및 보관 과정에서 발생하는 아티팩트는 이상 샘플 내 변이 목록 결정 시 노이즈에 해당하므로, 본 개시에 따른 기계학습 모델 등에 의해 이러한 노이즈가 추후 제거될 수 있다.

[0088] 변이 검출 모듈(310)은 복수의 검출 모듈(310_1 내지 310_n)(n은 임의의 자연수)을 포함할 수 있다. 복수의 검출 모듈(310_1 내지 310_n)은 시퀀싱 데이터(300)를 각각 입력받아, 변이 서브 후보 정보(320_1 내지 320_n)를 결정할 수 있다. 즉, '변이 서브 후보'는 개별 검출 모듈에서 각각 결정된, 변이로 추정되는 시퀀스를 지칭할 수 있다. 복수의 검출 모듈(310_1 내지 310_n)의 각각은 시퀀싱 데이터를 대조하는 프로세스가 서로 상이할 수 있으므로, 제1 내지 제n 변이 서브 후보 정보(320_1 내지 320_n)는 서로 상이할 수 있다. 제1 내지 제n 변이 서브 후보 정보(320_1 내지 320_n)는 각각의 변이 서브 후보가 결정된 검출 모듈과 연관된 정보를 포함할 수 있다.

[0089] 변이 후보 정보(340)는 변이 서브 후보 정보(320_1 내지 320_n)에 기초하여 결정될 수 있다. 예를 들어, 변이 검출 모듈(310)은 변이 서브 후보 정보(320_1 내지 320_n)를 통합(union)(330)함으로써, 변이 후보 정보(340)를 결정할 수 있다. 가령, 제1 검출 모듈(310_1)로부터 결정된 변이 서브 후보는 a 변이, b 변이를 포함하고, 제2 검출 모듈(310_2)로부터 결정된 변이 서브 후보는 a 변이, c 변이를 포함하고, 제n 검출 모듈(310_n)로부터 결정된 변이 서브 후보는 b 변이, d 변이를 포함하는 경우, 변이 후보 정보(340)는 변이 서브 후보 정보(320_1 내지 320_n)의 합집합으로써 a 변이, b 변이, c 변이 및 d 변이와 연관된 정보를 포함할 수 있다. 이를 통해, 실제 변이일 가능성이 있는 모든 변이 후보 정보를 취합할 수 있다.

[0090] 이와 달리, 변이 검출 모듈(310)은 복수의 검출 모듈(310_1 내지 310_n) 중 임의의 개수(예: 2개) 이상의 검출 모듈에서 공통적으로 결정된 변이 서브 후보 정보를 변이 후보 정보(340)로 결정할 수 있다. 가령, 위의 변이 서브 후보 예시에 따르면, 변이 후보 정보(340)는 a 변이 및 b 변이와 연관된 정보를 포함할 수 있다. 이를 통해, 실제 변이일 가능성이 높은 변이 후보 정보를 취합할 수 있다.

[0091] 변이 후보 정보(340)는 변이 후보의 위치 정보, 참조 변이 후보의 위치에서의 기준 대립유전자(reference allele) 정보, 참조 변이 후보에 대응되는 변형된 대립유전자(altered allele) 정보, 변이 후보의 신뢰도 정보, 변이 후보의 품질 정보(예를 들어, Phred quality score), 변이 후보의 유전형(genotype) 정보, 변이 후보와 연관된 리드 카운트(read count) 정보 및/또는 복수의 검출 모듈(310_1 내지 310_n)중 해당 변이 후보가 결정된 검출 모듈의 정보 등을 포함할 수 있다. 변이 후보의 위치 정보는, 변이 후보가 위치한 염색체 정보(예: 염색체 번호) 및/또는 염색체 내에서의 위치 정보(예: 1-based position)를 포함할 수 있다.

[0092] 추가적으로, 변이 후보 정보(340)는 필터(filter) 정보를 포함할 수 있다. 예를 들어, 복수의 검출 모듈(310_1 내지 310_n)의 각각은 변이 서브 후보가 미리 정해진 복수의 품질 지표를 충족하는지 여부를 판별하는 필터링을 수행할 수 있으며, 필터 정보는 복수의 검출 모듈(310_1 내지 310_n)의 각각에서 변이 서브 후보가 충족하거나 충족하지 못한 품질 지표와 연관된 정보를 포함할 수 있다. 가령, 필터 정보는 변이 서브 후보가 약한 증거(weak evidence) 기반의 변이 후보인지 여부, 슬리피지(slippage)가 발생한 변이 후보인지 여부, 다른 변이 후보와 인접하여 발생하였는지(clustered events) 여부, 일배체형(haplotype)인지 여부, 생식 세포(germline) 계열의 변이 후보인지 여부 등의 필터를 통과하였는지 여부와 연관된 정보를 포함할 수 있다. 변이 서브 후보가 모든 필터를 통과하는 것에 응답하여, 변이 서브 후보의 필터 정보는 'PASS'로 표시될 수 있다.

[0093] 도 3에서는 변이 검출 모듈(310)이 복수의 검출 모듈(310_1 내지 310_n)을 포함하는 것으로 도시되었으나 이에 한정되지 않으며, 변이 검출 모듈(310)은 하나의 검출 모듈로 구성될 수도 있다. 이 경우, 하나의 검출 모듈로

부터 생성되는 변이 서브 후보 정보는 변이 후보 정보(340)와 동일할 수 있다.

[0094] 도 4는 본 개시의 일 실시예에 따른 어노테이션 모듈(420)의 예시를 나타내는 도면이다. 어노테이션 모듈(420)은 변이 후보 정보(410) 및/또는 데이터베이스(430)로부터 수신한 정보에 기초하여 어노테이션 정보(442, 444, 446, 448, 450)를 생성할 수 있다. 데이터베이스(430)는 서로 다른 유형의 정보를 포함하는 복수의 데이터베이스(예: Ensembl, RefSeq 등의 유전자 데이터베이스)를 포함할 수 있다.

[0095] 예를 들어, 어노테이션 모듈(420)은, 기계학습 모델의 학습 프로세스에서, 참조 변이 후보 정보 및/또는 데이터베이스(430)로부터 수신한 정보에 기초하여 어노테이션 정보(442, 444, 446, 448, 450)를 생성할 수 있다. 이와 유사하게, 어노테이션 모듈(420)은 기계학습 모델의 추론 프로세스에서, 타겟 변이 후보 정보 및/또는 데이터베이스(430)로부터 수신한 정보에 기초하여 어노테이션 정보(442, 444, 446, 448, 450)를 생성할 수 있다. 생성된 어노테이션 정보(442, 444, 446, 448, 450)의 적어도 일부는 참조 변이 후보 또는 타겟 변이 후보의 특징으로서 기계학습 모델에 입력되어, 기계학습 모델의 학습 또는 추론의 기초가 되는 데이터로 이용될 수 있다.

[0096] 변이 후보 정보(410)는 도 3의 변이 후보 정보(340)와 대응될 수 있다. 변이 후보 정보(410)는, 변이 검출 모듈을 이용하여, FFPE 처리된 이상 샘플 및 정상 샘플로부터 결정된 것이거나 FF 처리된 이상 샘플 및 정상 샘플로부터 결정된 것일 수 있다.

[0097] 어노테이션 모듈(420)은 제1 어노테이션 모듈(422) 및 제2 어노테이션 모듈(424)을 포함할 수 있다. 제1 어노테이션 모듈(422) 및 제2 어노테이션 모듈(424)은 설명의 편의를 위해 기능별로 구분하여 설명되나, 이는 발명의 이해를 돕기 위한 것으로서, 반드시 물리적으로 구분되는 것을 의미하지 않고, 이에 한정되지 않는다. 예를 들어, 제1 어노테이션 모듈(422) 및 제2 어노테이션 모듈(424)은 하나의 모듈로 구성될 수 있다. 다른 예에서, 제2 어노테이션 모듈(424)은 제2 내지 제5 어노테이션 정보(444, 446, 448, 450)를 각각 출력하는 복수의 모듈을 포함할 수 있다.

[0098] 제1 어노테이션 모듈(422)은 변이 후보 정보(410)에 기초하여 데이터베이스(430)로부터 변이 후보와 관련하여 알려진 정보를 추출하고, 추출된 정보를 포함하는 제1 어노테이션 정보(442)를 생성할 수 있다. 예를 들어, 제1 어노테이션 정보(442)는 변이 후보가 단백질(또는, 아미노산 서열)에 미치는 영향, 유전 패턴, 대립 유전자 변이, 특정 인구 집단에서의 발생 빈도, 위험도 등과 연관된 정보를 포함할 수 있다.

[0099] 추가적으로 또는 대안적으로, 제1 어노테이션 정보(442)는 변이 후보 생물학적 결과(biological consequence) 정보를 포함할 수 있다. 예를 들어, 제1 어노테이션 모듈(422)은 변이 후보 정보(410)를 기초로 데이터베이스(430)로부터 변이 후보의 시퀀스 컨텍스트(sequence context) 정보를 추출한 뒤, 이를 이용하여 변이 후보에 대한 생물학적 결과를 포함하는 제1 어노테이션 정보(442)를 생성할 수 있다.

[0100] 제1 어노테이션 모듈(422)에서 참조되는 데이터베이스(430)는 Ensembl 또는 RefSeq 등의 유전자 데이터베이스를 포함할 수 있다.

[0101] 제2 어노테이션 모듈(424)은 변이 후보와 연관된 시퀀스 컨텍스트 및/또는 변이 후보의 상태와 연관된 정보 등을 포함하는 어노테이션 정보를 생성할 수 있다.

[0102] 일 실시예에서, 제2 어노테이션 모듈(424)에서 생성되는 제2 어노테이션 정보(444)는, 샘플(예를 들어, 변이 후보를 포함하는 참조 샘플 또는 타겟 샘플)의 시퀀싱 결과, 매핑(mapping) 및 정렬(aligning) 위치가 변이 후보의 위치와 중첩되는 복수의 리드(read)와 연관된 정보를 포함할 수 있다. 예를 들어, 제2 어노테이션 정보(444)는, 특정 변이 후보가 발견된 시퀀싱 리드(sequencing read) 상에서 이 변이 후보의 위치가 시퀀싱 리드의 시작위치로부터 얼마나 떨어져 있는가를 나타내는 position from 5'-end, position from 3'-end 등의 정보를 포함할 수 있다.

[0103] 다른 예에서, 제2 어노테이션 정보(444)는, 변이 후보를 포함하는 샘플의 시퀀싱 데이터에 포함된 복수의 리드 중 레퍼런스 지놈(reference genome)과 상이한(예: 변이를 포함하는) 리드인 변이 리드(variant read)와 연관된 정보 및 변이 리드의 위치와 중첩되는 위치를 가지는 비변이 리드(non-variant read, 이하, “레퍼런스 리드(reference read)”라고도 지칭한다)와 연관된 정보를 포함할 수 있다.

[0104] 레퍼런스/변이 리드와 연관된 정보는, 레퍼런스/변이 리드의 개수, 레퍼런스/변이 리드와 연관된, 매핑 품질(mapping quality)의 최솟값, 중간값, 최댓값, 염기 품질(base quality)의 통곗값(예:중간값, 평균값), 클리핑 베이스(clipping base)의 비율, 매핑된 리드의 매칭되지 않은 염기 수의 통곗값(예:최솟값, 중간값, 최댓값), 인서트 크기(insert size)의 통곗값(예: 제1 사분위수 내지 제3 사분위수), 적절히 짝을 이룬 리드(properly

paired read)의 개수, 리드의 서로 다른 일부가 서로 다른 레퍼런스 지놈에 적절하게 정렬되는 리드를 뜻하는 키메릭 리드(chimeric read)의 개수 등의 정보를 포함할 수 있다. "인서트 크기"란 짝을 이룬 리드 사이의 레퍼런스 지놈 상의 거리를 의미할 수 있다. "인서트 크기"는 리드1의 길이, 리드2의 길이, 그리고 리드1과 리드2 사이의 시퀀싱되지 않은 부분(unsequenced portion)의 길이의 합으로 이해될 수 있다. "적절히 짝을 이룬 리드"란 짝을 이룬 리드1과 리드2가 각각 정방향(forward 방향)과 역방향(reverse 방향)으로 잘 정렬되어 있고 인서트 사이즈가 기대 값에서 크게 벗어나지 않게(가령, 인서트 사이즈가 하한임계치와 상한임계치 사이 이도록) 짝을 이룬 리드를 의미할 수 있다. 인서트 사이즈의 기대 값은 DNA 시퀀싱 라이브러리 제작과정에서 DNA를 일정한 사이즈의 절편(fragment)으로 잘라 진행(size selection)할 때, 잘려진 절편의 평균 길이에 따라 기대되는 리드1과 리드2 사이의 거리를 의미할 수 있다.

[0105] 레퍼런스/변이 리드와 연관된 정보는 다양한 항목명으로 분류되어 저장될 수 있다. 예를 들어, 앞서 설명된 레퍼런스/변이 리드와 연관된 정보는 ref_readN, ref_minMQ, ref_medMQ, ref_maxMQ, ref_medBQ, ref_meanBQ, ref_clip_pct, ref_mismatch_min, ref_mismatch_med, ref_mismatch_max, ref_f1_n, ref_f2_n, ref_r1_n, ref_r2_n, ref_isize_lq, ref_isize_uq, ref_isize_min, ref_isize_max, ref_ppair_n, ref_chim_n, var_readN, var_minMQ, var_medMQ, var_maxMQ, var_medBQ, var_meanBQ, var_clip_pct, var_mismatch_min, var_mismatch_med, var_mismatch_max, var_f1_n, var_f2_n, var_r1_n, var_r2_n, var_isize_lq, var_isize_uq, var_isize_min, var_isize_max, var_ppair_n, var_chim_n, pf5p_med, pf3p_med 등의 항목명으로 저장될 수 있고, 각 항목명은 표 1 및 표 2와 같이 정의될 수 있다. 표 1은 레퍼런스 리드와 연관된 정보의 예시를 나타내고, 표 2는 변이 리드와 연관된 정보의 예시를 나타낸다.

표 1

[0107]

항목명	설명
ref_readN	레퍼런스 리드의 수
ref_minMQ	레퍼런스 리드의 매핑 품질(mapping quality)의 최솟값
ref_medMQ	레퍼런스 리드의 매핑 품질(mapping quality)의 중간값
ref_maxMQ	레퍼런스 리드의 매핑 품질(mapping quality)의 최댓값
ref_minBQ	레퍼런스 리드의 염기 품질(base quality)의 최솟값
ref_medBQ	레퍼런스 리드의 염기 품질(base quality)의 중간값
ref_meanBQ	레퍼런스 리드의 염기 품질(base quality)의 평균값
ref_maxBQ	레퍼런스 리드의 염기 품질(base quality)의 최댓값
ref_clip_pct	레퍼런스 리드의 클리핑 베이스의 비율(percentage)
ref_mismatch_min	레퍼런스 리드에서 레퍼런스 지놈과 매칭되지 않은 염기 수의 최솟값
ref_mismatch_med	레퍼런스 리드에서 레퍼런스 지놈과 매칭되지 않은 염기 수의 중간값
ref_mismatch_max	레퍼런스 리드에서 레퍼런스 지놈과 매칭되지 않은 염기 수의 최댓값
ref_f1_n	리드 쌍 중 첫 번째 리드이면서 동시에 정방향(forward 방향)으로 레퍼런스 지놈에 정렬된, 레퍼런스 리드의 수
ref_f2_n	리드 쌍 중 두 번째 리드이면서 동시에 정방향으로 레퍼런스 지놈에 정렬된, 레퍼런스 리드의 수
ref_r1_n	리드 쌍 중 첫 번째 리드이면서 동시에 역방향(backward 방향)으로 레퍼런스 지놈에 정렬된, 레퍼런스 리드의 수
ref_r2_n	리드 쌍 중 두 번째 리드이면서 동시에 역방향(backward 방향)으로 레퍼런스 지놈에 정렬된, 레퍼런스 리드의 수
ref_isize_lq	레퍼런스 리드의 인서트 사이즈의 제1사분위수(25%)
ref_isize_uq	레퍼런스 리드의 인서트 사이즈의 제3사분위수(75%)
ref_isize_min	레퍼런스 리드의 인서트 사이즈의 최솟값
ref_isize_max	레퍼런스 리드의 인서트 사이즈의 최댓값
ref_ppair_n	적절히 짝을 이룬 레퍼런스 리드 수
ref_chim_n	키메릭 리드(chimeric read)에 해당하는 레퍼런스 리드 수

표 2

[0109]

항목명	설명
var_readN	변이 리드의 수
var_minMQ	변이 리드의 매핑 품질(mapping quality)의 최솟값
var_medMQ	변이 리드의 매핑 품질(mapping quality)의 중간값
var_maxMQ	변이 리드의 매핑 품질(mapping quality)의 최댓값

var_minBQ	변이 리드의 염기 품질(base quality)의 최솟값
var_medBQ	변이 리드의 염기 품질(base quality)의 중간값
var_meanBQ	변이 리드의 염기 품질(base quality)의 평균값
var_maxBQ	변이 리드의 염기 품질(base quality)의 최댓값
var_clip_pct	변이 리드의 클리핑 베이스의 비율(percentage)
var_mismatch_min	변이 리드에서 레퍼런스 지놈과 매칭되지 않은 염기 수의 최솟값
var_mismatch_med	변이 리드에서 레퍼런스 지놈과 매칭되지 않은 염기 수의 중간값
var_mismatch_max	변이 리드에서 레퍼런스 지놈과 매칭되지 않은 염기 수의 최댓값
var_f1_n	리드 쌍 중 첫 번째 리드이면서 동시에 정방향(forward 방향)으로 레퍼런스 지놈에 정렬된, 변이 리드의 수
var_f2_n	리드 쌍 중 두 번째 리드이면서 동시에 정방향으로 레퍼런스 지놈에 정렬된, 변이 리드의 수
var_r1_n	리드 쌍 중 첫 번째 리드이면서 동시에 역방향(backward 방향)으로 레퍼런스 지놈에 정렬된, 변이 리드의 수
var_r2_n	리드 쌍 중 두 번째 리드이면서 동시에 역방향(backward 방향)으로 레퍼런스 지놈에 정렬된, 변이 리드의 수
var_isize_lq	변이 리드의 인서트 사이즈의 제1사분위수(25%)
var_isize_uq	변이 리드의 인서트 사이즈의 제3사분위수(75%)
var_isize_min	변이 리드의 인서트 사이즈의 최솟값
var_isize_max	변이 리드의 인서트 사이즈의 최댓값
var_ppair_n	적절히 짝을 이룬 변이 리드 수
var_chim_n	키메릭 리드(chimeric read)에 해당하는 변이 리드 수
pf5p_med	변이 리드의 변이 위치가 5'-end로부터 떨어진 거리의 중간값
pf3p_med	변이 리드의 변이 위치가 3'-end로부터 떨어진 거리의 중간값

[0110] 표 1 및 표 2에서 설명된 바와 같이, 각 항목명을 구성하는 세부 명칭들 중 "ref_"는 레퍼런스 리드와 연관된 정보임을 나타내고, "var_"는 변이 리드와 연관된 정보임을 나타낼 수 있다. "readN"은 리드의 수를 나타내고, "MQ"와 "BQ"는 각각 매핑 품질(Mapping Quality)와 염기 품질(Base Quality)을 나타내고, "min", "med", "mean", "max"는 각각 최솟값, 중간값, 평균값, 최댓값을 나타내고, "clip_pct"는 클리핑 베이스의 비율(percentage)을 나타내고, "mismatch"는 레퍼런스 지놈과 매칭되지 않은 염기 수를 나타내고, "fk(단, k는 자연수)"은 특정 리드가 리드 쌍 중 k 번째 리드이면서 동시에 정방향(forward 방향)으로 레퍼런스 지놈에 정렬됨을 나타내고, 마찬가지로 "r1(단, 1은 자연수)"는 특정 리드가 리드 쌍 중 1 번째 리드이면서 역방향(reverse 방향)으로 레퍼런스 지놈에 정렬됨을 나타내고, "isize"는 인서트 사이즈(insert size)를 나타내고, "lq"와 "uq"는 각각 제1사분위수(25%)와 제3 사분위수(75%)를 나타내고, "ppair"는 적절히 짝을 이룬 리드(properly paired read), "chim"은 키메릭 리드(chimeric read)를 나타내고, "pf5p", "pf3p"는 각각 변이 리드의 변이 위치가 5'-end로부터 떨어진 거리(position from 5'-end), 변이 리드의 변이 위치가 3'-end로부터 떨어진 거리(position from 3'-end)를 나타낼 수 있다. 그리고, 상술한 세부 명칭들을 조합함으로써 레퍼런스/변이 리드와 연관된 정보의 종류가 정의될 수 있다. 가령, "ref_mismatch_mean"은 레퍼런스 리드에서 레퍼런스 지놈과 매칭되지 않은 염기 수의 평균값을 나타낼 수 있다. 추가적으로, 상술한 정보 종류에 한정되지 않고, 상술한 세부 명칭들을 조합하여 추가적으로 정의되는 레퍼런스/변이 리드와 연관된 정보가 제2 어노테이션 정보(444)에 포함될 수 있다.

[0111] 일 실시예에서, 제2 어노테이션 모듈(424)에서 생성되는 제3 어노테이션 정보(446)는, 복수의 정상 샘플과 연관된 복수의 시퀀싱 데이터로부터 생성된 정상 조직 유전체 데이터(PON: Panel Of Normals)를 포함할 수 있다. 정상 조직 유전체 데이터는 데이터베이스(430)로부터 획득되는 전장 유전체 시퀀싱 데이터일 수 있으며, 복수의 정상 샘플에 걸쳐 공통적으로 나타나는 특성과 연관된 정보, 즉 정상 샘플 군집의 특성을 반영한 정보를 포함할 수 있다.

[0112] 예를 들어, 정상 조직 유전체 데이터는, 정상 조직 유전체 데이터를 구축하는 데 기초가 된 복수의 정상 샘플에서, 임의의 위치(position)에 대한 리드 깊이(read depth)를 모두 합한 값('PON_dpsum'로 지칭), 복수의 정상 샘플 중 해당 위치에 대한 리드 깊이가 0이 아닌 샘플들의 수('PON_dpN'로 지칭), 복수의 정상 샘플 중 해당 위치에 대한 리드 깊이가 10 이상인 샘플들의 수('PON_dp10N'로 지칭), 복수의 정상 샘플의 해당 위치에 대한 변이 리드(variant read) 개수를 모두 합한 값('PON_varsum'로 지칭), 복수의 정상 샘플 중 해당 위치에서 변이 리드를 1개라도 가지고 있는 샘플들의 수('PON_varN'로 지칭), 복수의 정상 샘플 중 해당 위치에 대한 VAF(variant allele frequency)가 0.2 미만인 샘플들의 수('PON_var0.21N'로 지칭), 복수의 정상 샘플 중 해당

위치에 대한 VAF가 0.2 이상인 샘플의 수('PON_var0.2hN'로 지칭) 및/또는 복수의 정상 샘플 중 해당 위치에 대한 변이 리드가 2개인 샘플들의 수('PON_var2N'로 지칭) 등을 포함할 수 있다. 'PON_dp10N', 'PON_var0.21N', 'PON_var0.2hN', 'PON_var2N'의 산출을 위한 비곳값은 각각 10, 0.2, 0.2, 2로 설명되었으나, 이는 임의로 설정될 수 있다. 가령, 정상 조직 유전체 데이터는 복수의 정상 샘플 중 임의의 위치에 대한 VAF가 0.25 미만인 샘플들의 수('PON_var0.251N'로 지칭)를 포함할 수 있다.

[0113] 일 실시예에서, 제2 어노테이션 모듈(424)에서 생성되는 제4 어노테이션 정보(448)는, FFPE(Formalin-Fixed, Paraffin-Embedded) 처리된 복수의 샘플과 연관된 복수의 시퀀싱 데이터로부터 생성된 FFPE 처리 조직 유전체 데이터(POF: Panel of FFPEs)를 포함할 수 있다. FFPE 처리된 복수의 샘플은 정상 샘플일 수 있다.

[0114] 대안적으로, FFPE 처리된 복수의 샘플은 이상 샘플(예를 들어, 종양 세포 샘플)일 수 있다. FFPE 처리된 복수의 샘플이 이상 샘플인 경우, 이상 샘플 내 변이와 연관된 정보가 유전체 데이터에 포함될 수 있다. 따라서, FFPE 처리된 이상 샘플과 대응되는 FF(Fresh-Frozen) 처리된 이상 샘플을 이용하여 변이와 연관된 정보를 제거함으로써, 정상 샘플을 사용하여 FFPE 처리 조직 유전체 데이터를 구현하는 것과 실질적으로 동일한 데이터가 구현될 수 있다. 이 때, FFPE 처리된 이상 샘플은 FF 처리된 이상 샘플로부터 직접적으로 전환(direct conversion)된 것일 수 있다.

[0115] FFPE 처리 조직 유전체 데이터는 데이터베이스(430)로부터 획득되는 전장 유전체 시퀀싱 데이터일 수 있으며, 복수의 FFPE 샘플에 걸쳐 공통적으로 나타나는 특성과 연관된 정보, 즉 FFPE 처리된 샘플 군집의 특성을 반영한 정보를 포함할 수 있다.

[0116] 예를 들어, FFPE 처리 조직 유전체 데이터는, FFPE 처리 조직 유전체 데이터의 기초가 된 복수의 FFPE 처리된 샘플에서, 임의의 위치(position)에 대한 리드 깊이(read depth)를 모두 합한 값('POF_dpsum'로 지칭), 복수의 FFPE 샘플 중 해당 위치에 대한 리드 깊이가 0이 아닌 샘플들의 수('POF_dpN'로 지칭), 복수의 FFPE 샘플 중 해당 위치에 대한 리드 깊이가 10 이상인 샘플들의 수('POF_dp10N'로 지칭), 복수의 FFPE 샘플의 해당 위치에 대한 변이 리드(variant read) 개수를 모두 합한 값('POF_varsum'로 지칭), 복수의 FFPE 샘플 중 해당 위치에 대한 변이 리드를 1개라도 가지고 있는 샘플들의 수('POF_varN'로 지칭), 복수의 FFPE 샘플 중 해당 위치에 대한 VAF가 0.2 미만인 샘플들의 수('POF_var0.21N'로 지칭), 복수의 FFPE 샘플 중 해당 위치에 대한 VAF가 0.2 이상인 샘플들의 수('POF_var0.2hN'로 지칭) 및/또는 복수의 FFPE 샘플 중 해당 위치에 대한 변이 리드가 2개인 샘플들의 수('POF_var2N'로 지칭) 등을 포함할 수 있다. 'POF_dp10N', 'POF_var0.21N', 'POF_var0.2hN', 'POF_var2N'의 산출을 위한 비곳값은 각각 10, 0.2, 0.2, 2로 설명되었으나, 이는 임의로 설정될 수 있다. 가령, FFPE 처리 조직 유전체 데이터는 복수의 FFPE 처리된 샘플 중 임의의 위치에 대한 VAF가 0.25 미만인 샘플들의 수('POF_var0.251N'로 지칭)를 포함할 수 있다.

[0117] 일 실시예에서, 제2 어노테이션 모듈(424)에서 생성되는 제5 어노테이션 정보(450)는, 변이 후보의 변이 유형과 연관된 정보 및 변이 후보의 시퀀스 컨텍스트(sequence context) 정보를 포함할 수 있다.

[0118] 예를 들어, 변이 후보가 SNV인 경우, 변이 유형과 연관된 정보는 SNV의 패턴 정보를 포함할 수 있다. 가령, 패턴 정보는 'A->C', 'C->A', 'G->T', 'T->G', 'G->U'와 같이 변이 전 염기와 변이 후 염기의 정보를 포함할 수 있다.

[0119] 이와 달리, 변이 후보가 INDEL인 경우, 변이 유형과 연관된 정보는, 변이 후보가 결실 변이 또는 삽입 변이 중 어느 변이에 해당하는지 여부, 결실 변이인 경우 결실 서열 길이, 삽입 변이인 경우 삽입 서열 길이 등의 정보를 포함할 수 있다.

[0120] 변이 후보가 SNV인 경우, 시퀀스 컨텍스트 정보는 변이 후보를 포함하는 일정 길이(예: 3bp, 5bp)의 주변 염기 서열(flanking base) 정보를 포함할 수 있다. 가령, 특정 위치의 'A(아데닌)'가 변이 후보인 경우, 시퀀스 컨텍스트 정보는 'CpApG'(단, 'p'는 인산다이에스터 결합을 나타냄) 등으로 표현될 수 있다(즉, 변이 후보와 인접한 염기는 C(사이토신)와 G(구아닌)).

[0121] 이와 달리, 변이 후보가 INDEL인 경우, 시퀀스 컨텍스트 정보는 변이 후보 위치의 앞 또는 뒤의 서열이 반복되는 서열인지 또는 미세상동성(microhomology) 패턴을 갖는지에 대한 정보를 포함할 수 있다. 가령, 반복 서열이 'A'이고 반복 길이가 4인 경우 시퀀스 컨텍스트 정보는 'AAAA'로 표현될 수 있고, 반복 서열이 'ACG'이고 반복 길이가 3인 경우 시퀀스 컨텍스트 정보는 'ACGACGACG'로 표현될 수 있다. 2bp('AG') 미세상동성 패턴을 갖는 예시에서, 시퀀스 컨텍스트 정보는 변이 후보의 앞과 뒤가 같은 'AG{deleted sequence}AG'로 표현될 수 있다.

[0122] 어노테이션 모듈(420)에서 생성되는 어노테이션 정보는 도 4에서 도시되고 설명된 것에 한정되지 않으며, 일부 정보가 생략되거나 추가될 수 있다.

[0123] 도 5는 본 개시의 일 실시예에 따른 특징 추출 모듈(530)의 예시를 나타내는 도면이다. 특징 추출 모듈(530)은 어노테이션 정보(520)(추가적으로, 변이 후보 정보(510))에 기초하여 변이 후보의 특징(540)을 추출할 수 있다. 변이 후보 정보(510)는 도 3의 변이 후보 정보(340)와 대응될 수 있고, FFPE 처리된 이상 샘플 및 정상 샘플로부터 결정된 것일 수 있다. 추가적으로, 변이 후보 정보(510)는 필터 모듈(예를 들어, 도 8의 842, 846)에 의해 필터링된 정보일 수 있다. 어노테이션 정보(520)는 도 4의 어노테이션 모듈(420)에서 생성된 것일 수 있다.

[0124] 특징 추출 모듈(530)은, 기계학습 모델의 학습 프로세스에서, 참조 변이 후보 정보 및 이와 연관된 어노테이션 정보에 기초하여 참조 변이 후보의 특징을 추출할 수 있다. 이와 유사하게, 특징 추출 모듈(530)은 기계학습 모델의 추론 프로세스에서, 타겟 변이 후보 정보 및 이와 연관된 어노테이션 정보에 기초하여 타겟 변이 후보의 특징을 추출할 수 있다.

[0125] 변이 후보의 특징(540)은 범주형 변수(categorical variable)로 표현되는 특징과, 수치형 변수(numeric variable)로 표현되는 특징으로 구분될 수 있다. 일 실시예에서, 범주형 변수는 원-핫 인코딩(one-hot encoding) 등을 통해 수치형 변수로 매핑되어 표현될 수 있다.

[0126] 특징 추출 모듈(530)은 변이 후보 정보(510)에 대응하는 어노테이션 정보(520)의 전부 또는 일부에 기초하여 변이 후보의 특징(540)을 생성할 수 있다. 특징 추출 모듈(530)은 어노테이션 정보(520)의 전부 또는 일부를 그대로 또는 가공/변형하여 이용하는 특징 추출 프로세스를 통해 변이 후보의 특징(540)을 생성할 수 있다. 추가적으로, 변이 후보가 반복적으로 동일한 위치에 동일한 형태로 발생하는 것이 알려진 생물학적 변이(가령, 핫스팟 변이(hotspot mutation))인 경우, 변이 후보는 실제 변이에 해당할 확률이 높으므로, 변이 후보의 특징(540)은 변이 후보가 진양성 변이에 해당함을 나타내는 정보를 포함될 수 있다.

[0127] 특징 추출 모듈(530)은, 어노테이션 정보(520)를 이용하여 변이 후보의 특징(540)을 생성하는 특징 추출 프로세스와 연관된, 특징 추출 정보(550)를 데이터베이스(560)에 저장할 수 있다. 가령, 특징 추출 정보(550)는 추출될 어노테이션 정보의 종류, 어노테이션 정보에 수행될 연산 등과 연관된 정보를 포함할 수 있다. 특징 추출 모듈(530)은, 새로운 변이 후보의 특징 추출 시 데이터베이스(560)를 참조하여 특징 추출 정보(550)를 이용함으로써, 새로 입력되는 변이 후보 및 어노테이션 정보에 대해서도 동일한 특징 추출 프로세스를 수행할 수 있다. 추가적으로, 특징 추출 모듈(530)은 생성된 변이 후보의 특징(540)을 데이터베이스(560)에 저장할 수 있다.

[0128] 특징 추출 모듈(530)은 데이터베이스(560)에 저장된 정보를 기초로 변이 후보의 특징(540)과 관련한 추가적인 프로세스(예: refinement)를 수행할 수 있다. 일 예시로서, 특징 추출 모듈(530)은 변이 후보의 특징(540)에 대해, Z 스코어 표준화(Z-Score Standardization) 등의 데이터 표준화 프로세스를 수행할 수 있다. 가령, 특징 추출 모듈(530)은 변이 후보의 특징(540)의 평균(μ)과 표준편차(σ)를 이용하여 변이 후보의 특징(540)과 연관된 값 x를 $z=(x-\mu)/\sigma$ 으로 치환할 수 있다. 이 때, 변이 후보의 특징(540)에 대한 평균 및 표준편차 값은 데이터베이스(560)에 저장될 수 있다. 추가적으로, 특징 추출 모듈(530)은 데이터 표준화 프로세스 전, 변이 후보의 특징(540)의 정규분포화를 위해 로그 변환(Log Transformation)을 수행할 수 있다.

[0129] 다른 예시로서, 특징 추출 모듈(530)은 변이 후보의 특징(540)에 대해, 도메인 적응(DA: Domain Adaptation) 등의 기계학습 기법을 수행할 수 있다. 가령, 기계학습 모델의 학습 프로세스에서 사용되는 데이터가 포함된 데이터 도메인(Source Domain)과, 새로운 데이터가 포함된 데이터 도메인(Target Domain) 간 임의의 특징 값의 분포에 임계치 이상의 차이가 발생할 수 있다. 이 때, 특징 추출 모듈(530)은 이러한 차이를 줄이는 방향으로 Source Domain 및/또는 Target Domain의 특징 값의 분포를 조정할 수 있다. 이를 통해, 특징 값 분포의 차이에 의한 기계학습 모델의 성능 저하 문제가 완화될 수 있다.

[0130] 상술한 Z 스코어 표준화(Z-Score Standardization) 등의 데이터 표준화 프로세스 또는 도메인 적응(DA: Domain Adaptation) 등의 기계학습 기법은 변이 후보의 특징(540) 중 일부 특징(예를 들어, 원-핫 인코딩 등을 통해 수치형 변수로 매핑된 특징)에 대해서는 적용되지 않을 수 있다.

[0131] 도 6은 본 개시의 일 실시예에 따른 기계학습 모델(630)을 이용하여 변이 후보의 분류 결과(640)가 출력되는 예시를 나타내는 도면이다. 기계학습 모델(630)은 샘플 내 변이 후보 정보(610) 및 변이 후보의 특징(620)을 수신하여, 변이 후보가 진양성(True Positive) 변이인지(또는, 변이 후보가 FFPE 처리에 의한 아티팩트(artifact)인지) 여부를 나타내는 분류 결과(640)를 출력할 수 있다.

[0132] 예를 들어, 기계학습 모델(630)은 학습 프로세스에서, 참조 변이 후보 정보 및 참조 변이 후보의 특징을 수신하여, 참조 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 출력할 수 있다. 참조 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과와, 참조 변이 후보에 레이블링된 분류 정보에 기초하여 기계학습 모델(630)의 파라미터 내지 하이퍼 파라미터가 조정될 수 있다. 이에 대해서는 도 8을 이용하여 자세히 후술한다.

[0133] 이와 유사하게, 상술한 학습 프로세스에 의해 학습된 기계학습 모델(630)은, 추론 프로세스에서 타겟 변이 후보 정보 및 타겟 변이 후보의 특징을 수신하여, 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 출력할 수 있다. 이진 분류 출력 결과는 분류기 내부적으로 원시 출력(raw output) 내지 스코어(score)으로부터 시작하는 일련의 결정 과정을 통해 생성되는 것으로써, 출력 결과가 원시 출력 내지 스코어인 경우가 이를 포함한다고 볼 수 있으므로, 원시 출력 내지 스코어에 대한 후술 내용은 이진 분류인 경우에도 해당할 수 있다.

[0134] 도 7은 본 개시의 일 실시예에 따른 기계학습 모델(700)의 세부 구성을 나타내는 도면이다. 기계학습 모델(700)은 도 6의 기계학습 모델(600)에 대응될 수 있다. 기계학습 모델(700)은 복수의 분류기(classifier)(710_1 내지 710_n) 및 이와 연결된 메타 분류기(730)를 포함할 수 있다.

[0135] 복수의 분류기(710_1 내지 710_n)의 각각에는 변이 후보 정보(도 6의 610) 및 변이 후보의 특징(도 6의 620)이 입력될 수 있다. 복수의 분류기(710_1 내지 710_n)는, 변이 후보 정보 및 변이 후보의 특징에 기초하여, 변이 후보가 진양성 변이인지(또는, 변이 후보가 FFPE 처리에 의한 아티팩트(artifact)인지) 여부를 나타내는 복수의 출력 결과(720_1 내지 720_n)를 출력할 수 있다. 복수의 출력 결과(720_1 내지 720_n)는 TRUE/FALSE 등 이진 분류(binary class) 방식으로 분류되거나, 0~1 사이의 값을 갖는 원시 출력일 수 있다.

[0136] 이후, 메타 분류기(730)는 복수의 분류기(710_1 내지 710_n) 중 적어도 하나의 분류기로부터의 출력 결과를 이용하여, 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과(740)(도 6의 640과 대응)를 결정할 수 있다.

[0137] 메타 분류기(730)는 복수의 출력 결과(720_1 내지 720_n) 중 적어도 일부를 특징(feature)으로 하는 학습 데이터에 기초하여 학습된 분류기일 수 있다. 메타 분류기(730)는, 학습 데이터를 이용하여 산출되는 분류 결과(740) 및 변이 후보에 대해 레이블링된 분류 정보에 기초하여 학습될 수 있다.

[0138] 분류 결과(740)는 변이 후보가 진양성 변이인지 여부를 나타낼 수 있다. 분류 결과(740)는, 메타 분류기(730)로부터 결정되는 원시 출력 내지 스코어(변이 후보가 진양성 변이일 확률과 연관)가, 분류 결과(740)가 결정되기 위한 결정 임계값(decision threshold)보다 큰 것으로 판단되는 것에 응답하여, 변이 후보가 진양성 변이인 것으로 결정될 수 있다. 다른 실시예에서, 변이 후보가 진양성 변이일 확률이 임계 확률보다 높거나, 변이 후보가 아티팩트일 확률이 임계 확률보다 낮은 것으로 판단되는 것에 응답하여 변이 후보가 진양성 변이인 것으로 결정될 수 있다. 변이 후보가 진양성 변이일 확률 또는 변이 후보가 아티팩트일 확률은 메타 분류기(730)로부터 결정되는 원시 출력 내지 스코어로부터 결정될 수 있다. 분류 결과(740)는 변이 후보가 진양성 변이라는 의미의 'TRUE'로 표시되거나, 변이 후보가 아티팩트가 아니라는 의미의 'FALSE'로 표시될 수 있다.

[0139] 분류 결과(740)가 결정되기 위한 결정 임계값은 메타 분류기(730)에서 결정되는 raw score를 레이블링된 분류 정보와 비교함으로써 결정될 수 있다. 가령, 결정 임계값은 0 내지 1 사이의 값 중 F1-score 또는 ROC(Receiver Operating Characteristic) 곡선의 AUC(Area Under the Curve)가 최대가 되는 값으로 결정될 수 있다. 이 때, F1-score 또는 ROC 곡선의 AUC가 최대가 되는 값이 여러 개인 경우, 그 중간값, 평균값 등으로 결정 임계값이 결정될 수 있다.

[0140] 복수의 분류기(710_1 내지 710_n) 및 메타 분류기(730)의 각각은 회귀분석 모델(예: Elastic-Net Logistic Regression), 서포트 벡터 머신(SVM: Support Vector Machine), 랜덤 포레스트(Random Forest), 그라디언트 부스팅(Gradient Boosting, 가령 XGBoost, LightGBM 등) 또는 다층 퍼셉트론(MLP: Multilayer Perceptron) 등의 신경망으로 구현될 수 있다. 이에 한정되지 않고, 복수의 분류기(710_1 내지 710_n) 및 메타 분류기(730)의 각각은 심층 신경망(DNN: Deep Neural Network)를 포함하는 일반적인 기계학습 기반 분류기일 수 있다. 또한, 메타 분류기(730)는 기계학습 뿐만 아니라 규칙 기반 알고리즘(Rule-based Algorithm)을 포함할 수 있다. 추가적으로 또는 대안적으로, 메타 분류기(730)는 복수의 출력 결과(720_1 내지 720_n)를 애버리징(averaging)하거나 보팅(voting)함으로써 분류 결과(740)를 출력할 수 있다.

[0141] 기계학습 모델(700)의 학습 프로세스가 수행되기 위한 학습 데이터는 참조 변이 후보 정보, 변이 후보의 특징 및 참조 변이 후보에 레이블링된 분류 정보를 포함할 수 있다. 학습 데이터는 훈련 세트(training set) 및 검증 세트(validation set)로 분할될 수 있다. 추가적으로, 학습 데이터는 검증 세트와 별개의 테스트 세트(test set)가 생성되도록 더 분할될 수 있다.

[0142] 예를 들어, 학습 데이터는, 데이터를 k개로 분할한 뒤 k-1개를 훈련 세트로, 1개를 검증 세트로 사용하고, 이러한 과정을 k번 반복하여 k개의 성능 지표를 얻어내는 k-겹 교차검증법(k-fold cross validation)에 의해 분할되고 사용될 수 있다. 위 과정은 다시 n번 반복될 수 있으며(n-repeated k-fold cross validation), 매 데이터 분할 전 데이터가 랜덤하게 섞일 수 있다. 전체 학습 데이터는 '에포크(epoch)'로, 학습 데이터가 분할된 k개 데이터 세트의 각각은 '배치(batch)'로 지칭될 수 있다.

[0143] 분할된 복수의 배치의 각각은 복수의 분류기(710_1 내지 710_n) 중 서로 다른 종류의 분류기로 학습될 수 있다. 복수의 분류기(710_1 내지 710_n)의 종류 및 개수는 가변적일 수 있고, 분할된 복수의 배치의 개수에 기초하여 복수의 분류기(710_1 내지 710_n)의 개수가 결정될 수 있다. 예를 들어, 이용 가능한 분류기의 종류가 N개이고, 학습 데이터에 대해 k-겹 교차검증법(k-fold cross validation)이 n번 반복될 때(n-repeated k-fold cross validation), N, k 및 n 값에 기초하여 복수의 분류기(710_1 내지 710_n)의 개수(N*n*k)가 결정될 수 있다. 가령, 분류기의 종류가 5개이고, 10겹 교차검증법이 10번 반복되는 경우, 복수의 분류기(710_1 내지 710_n)의 개수는 5*10*10=500(개)일 수 있다.

[0144] 이와 달리, 기계학습 모델(700)에는 하나의 분류기 및 메타 분류기(730)가 포함될 수 있으며, 이 경우 메타 분류기(730)는 지시 함수(indicator function)일 수 있다.

[0145] k-겹 교차검증법을 이용하여 기계학습 모델(700)의 학습 프로세스가 진행되는 동안, 복수의 분류기(710_1 내지 710_n)의 하이퍼 파라미터(hyperparameter)가 최적화되고, 분류 결과(740)가 실제 분류와 일치하는지 여부에 기초하여 모델 선택(model selection)이 수행될 수 있다. 예를 들어, 분류 결과(740) 산출 시, 복수의 분류기(710_1 내지 710_n) 각각의 출력 결과를 사용할 것인지, 사용하는 경우 어떠한 가중치로 사용할 것인지 결정될 수 있다.

[0146] 일 실시예에서, 학습 데이터가 복수의 샘플에 대한 복수의 데이터 세트를 포함하는 경우, 복수의 샘플의 각각에서 실제 변이 수 및 그 분포가 서로 다른 점을 고려하여, 학습 데이터는 가중 샘플링(Weighted Sampling) 방식으로 분할되거나, 계층적 샘플링(Stratified Sampling) 방식으로 분할될 수 있다. 추가적으로, 각 샘플에 포함된 실제 변이의 수가 서로 달라 각 샘플에 대해 클래스 불균형(class imbalance) 문제가 발생하는 것을 방지하기 위해 일부 데이터가 오버샘플링(oversampling) 또는 언더샘플링(undersampling)되거나, 손실(loss) 계산 시 가중치 밸런싱(weight balancing)이 수행될 수 있다. 또는, 오버샘플링을 할 때 변이 후보의 특징을 변형하거나 새로운 데이터를 생성하는 합성 샘플링(synthetic sampling)을 수행할 수 있다.

[0147] 이와 달리, 학습 데이터가 복수의 배치로 분할되기 이전, 복수의 샘플에 대한 복수의 데이터 세트가 하나의 데이터 세트로 결합(concatenate)된 뒤 분할될 수 있다. 이 때, 데이터 세트 내 복수의 변이 후보(예를 들어, 도 10의 행(row))를 기준으로, 학습 데이터가 복수의 배치로 분할될 수 있다.

[0148] 기계학습 모델의 추론 프로세스에서는, 복수의 분류기(710_1 내지 710_n)의 각각은 타겟 샘플 내 타겟 변이 후보 정보 및 타겟 변이 후보의 특징을 수신하여 타겟 변이 후보가 진양성 변이인지 여부(또는, FFPE 처리에 의한 아티팩트인지 여부)를 나타내는 결과를 출력할 수 있다.

[0149] 복수의 분류기(710_1 내지 710_n)의 각각에서, 가령, 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 결과로서, 원시 출력 내지 스코어가 산출될 수 있다. 산출된 원시 출력 내지 스코어는 변이 후보가 진양성 변이일(또는, 아티팩트일) 확률 값으로 매핑 또는 변환될 수 있다.

[0150] 일 실시예에서, 복수의 분류기(710_1 내지 710_n)의 각각은 학습 프로세스의 교차 검증(cross validation)에 의해 결정된 컷오프 값(cut-off value)을 이용하여, 산출된 원시 출력, 스코어 또는 확률 값을 이진 분류(binary classification)하고, 이진 분류의 결과를 출력 결과(720_1 내지 720_n)로서 출력할 수 있다. 메타 분류기(730)는 출력 결과(720_1 내지 720_n)로서 출력된 이진 분류 결과를 보팅(voting)함으로써 분류 결과(740)를 출력할 수 있다.

[0151] 다른 실시예에서는, 복수의 분류기(710_1 내지 710_n)의 각각에서 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 결과를 나타내는 원시 출력, 스코어 또는 확률 값이 메타 분류기(730)를 이용하여 애버리징(averaging)되고, 학습 데이터의 일부를 이용하여 최적의 컷오프 값이 산출되고, 애버리징된 값 및 산출된 컷오프 값에 기초하여 분류 결과(740)가 출력될 수 있다.

[0152] 또 다른 실시예에서는, 복수의 분류기(710_1 내지 710_n)의 각각에서 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 결과를 나타내는 원시 출력, 스코어 또는 확률 값을 입력 특징으로 수신하여 학습된 메타 러닝 모델

(meta-learner)을 이용하여 분류 결과(740)가 산출될 수 있다.

[0153] 도 8은 본 개시의 일 실시예에 따른 기계학습 모델(870)의 학습 프로세스의 예시를 나타내는 도면이다. FFPE 샘플 시퀀싱 데이터(812)(예를 들어, FFPE 처리된 정상 샘플과 연관된 시퀀싱 데이터 및 FFPE 처리된 이상 샘플과 연관된 시퀀싱 데이터)는 참조 샘플 시퀀싱 데이터로서, FFPE 샘플 시퀀싱 데이터(812)를 이용하여 기계학습 모델(870)의 학습 데이터가 생성될 수 있다.

[0154] 변이 검출 모듈(820)(도 3의 310과 대응)은 FFPE 샘플 시퀀싱 데이터(812)로부터 FFPE 샘플 내 변이 후보 정보(822)를 결정할 수 있다. 마찬가지로, 변이 검출 모듈(820)은 FF 샘플 시퀀싱 데이터(814)(예를 들어, FF 처리된 정상 샘플과 연관된 시퀀싱 데이터 및 FF 처리된 이상 샘플과 연관된 시퀀싱 데이터)로부터 FFPE 샘플 내 변이 후보 정보(824)를 결정할 수 있다.

[0155] FFPE 샘플과 FF 샘플은 서로 대응되는 샘플일 수 있다. 예를 들어, FFPE 샘플과 FF 샘플은 동일 개체로부터 채취된 것일 수 있다. 추가적으로, FFPE 샘플은 FF 샘플로부터 직접적으로 전환된 샘플이거나 FF 샘플과 일부 시간차를 두고 채취된 샘플일 수 있다.

[0156] 어노테이션 모듈(830)(도 4의 420과 대응)은 FFPE 샘플 변이 후보 정보(822)를 입력받아, FFPE 샘플 어노테이션 정보(832)를 출력할 수 있다. FFPE 샘플 어노테이션 정보(832)는 제1 필터 모듈(842)에 전달됨으로써, FFPE 샘플 내 변이 후보 정보(822)를 필터링하는 데 이용될 수 있다.

[0157] 이와 유사하게, 어노테이션 모듈(830)은 FF 샘플 변이 후보 정보(824)를 입력받아, FF 샘플 어노테이션 정보(834)를 출력할 수 있다. FF 샘플 어노테이션 정보(834)는 제2 필터 모듈(846)에 전달됨으로써, FF 샘플 내 변이 후보 정보(824)를 필터링하는 데 이용될 수 있다.

[0158] 제1 필터 모듈(842)은 FFPE 샘플 내 변이 후보 정보(822)를 입력받아, FFPE 샘플 내 필터링된 변이 후보 정보(844)를 출력할 수 있다. 이와 유사하게, 제2 필터 모듈(846)은 FF 샘플 내 변이 후보 정보(824)를 입력받아, FF 샘플 내 필터링된 변이 후보 정보(848)를 출력할 수 있다. 즉, 제1 필터 모듈(842) 및 제2 필터 모듈(846)은 샘플 내 변이 후보의 일부를 필터링하여 노이즈(예를 들어, FFPE 처리 시 발생하는 아티팩트(artifact) 등)를 제거함으로써, 기계학습 모델(870)의 학습 정확도를 향상시킬 수 있다.

[0159] 일 실시예에서, 제1 필터 모듈(842) 및 제2 필터 모듈(846)은 변이 후보 정보(822, 824) 중 변이 검출 모듈(820)에 의해 생성된 필터 정보에 기초하여 필터링을 수행할 수 있다. 필터 정보는 변이 후보가 충족하거나 충족하지 못한 품질 지표와 연관된 정보를 포함할 수 있다. 예를 들어, 필터 정보는 변이 후보가 약한 증거(weak evidence) 기반의 변이 후보인지 여부, 슬리피지(slippage)가 발생한 변이 후보인지 여부, 다른 변이 후보와 인접하여 발생하였는지(clustered events) 여부, 일배체형(haplotype)인지 여부, 생식 세포(germline) 계열의 변이 후보인지 여부 등의 필터 통과 여부와 연관된 정보를 포함할 수 있고, 변이 후보가 모든 필터를 통과하는 것에 응답하여 필터 정보는 'PASS'로 표시될 수 있다.

[0160] 예를 들어, 제1 필터 모듈(842) 및 제2 필터 모듈(846)은 상술한 필터 정보를 참조하여, 특정 품질 지표를 충족하지 못한 변이 후보를 필터링할 수 있다. 가령, 제1 필터 모듈(842) 및 제2 필터 모듈(846)은 필터 정보가 'PASS'로 표시되지 않는 변이 후보를 필터링할 수 있다.

[0161] 일 실시예에서, 제1 필터 모듈(842) 및 제2 필터 모듈(846)은 FFPE 샘플 어노테이션 정보(832) 및 FF 샘플 어노테이션 정보(834) 내 정상 조직 유전체 데이터(PON: Panel of Normals) 및/또는 FFPE 처리 조직 유전체 데이터(POF: Panel of FFPEs)(예를 들어, 도 4의 제2 어노테이션 모듈(424)에서 생성)에 기초하여 아티팩트인 변이 후보의 적어도 일부를 필터링할 수 있다.

[0162] 일 실시예에서, 제1 필터 모듈(842) 및 제2 필터 모듈(846)(대안적으로, 변이검출 모듈(820))은, 변이 검출 모듈(820)에 포함된 복수의 검출 모듈(도 3의 310_1 내지 310_n에 대응)에서 결정된 변이 서브 후보 정보가 통합(union)됨으로써 생성된 변이 후보 정보(822, 824) 중, 변이 검출 모듈(820)에 포함된 복수의 검출 모듈(도 3의 310_1 내지 310_n에 대응) 중 미리 정해진 개수(예: 2개) 이상의 검출 모듈에서 공통적으로 결정된 변이 서브 후보 정보를 필터링된 변이 후보 정보(844, 848)로 결정할 수 있다. 추가적으로, 제1 필터 모듈(842) 및 제2 필터 모듈(846)에 의해 수행되는 필터링은 복수의 검출 모듈의 종류 및/또는 개수에 기초하여 수행될 수 있다.

[0163] 추가적으로, 제2 필터 모듈(846)은 FF 샘플 어노테이션 정보(834)와 연관된 필터링 조건에 기초하여 필터링을 수행할 수 있다. 이 때, FF 샘플 어노테이션 정보(834)와 연관된 필터링 조건은 시퀀싱 플랫폼(sequencing platform), 라이브러리 제작(library preparation) 방식, 시퀀싱 깊이(sequencing depth), 조직 샘플의 상태,

샘플 순도(purity) 등 다양한 환경 맥락적 변수에 기초하여 보정 및 최적화되어 설정될 수 있다. FF 샘플 어노테이션 정보(834)와 연관된 필터링은 규칙 기반 알고리즘 또는 기계학습 기반 모델 등을 이용하여 수행될 수 있다.

[0164] 특징 추출 모듈(850)(도 5의 530과 대응)은 FFPE 샘플 어노테이션 정보(832)(추가적으로, FFPE 샘플 내 필터링된 변이 후보 정보(844))에 기초하여 변이 후보의 특징(852)을 추출할 수 있다. 이와 달리, 제1 필터 모듈(842)이 생략되는 경우, 특징 추출 모듈(850)은 FFPE 샘플 내 변이 후보 정보(822) 및 FFPE 샘플 어노테이션 정보(832)에 기초하여 변이 후보의 특징(852)을 추출할 수 있다. 추출된 변이 후보의 특징(852)은, FFPE 샘플 내 필터링된 변이 후보 정보(844) 와 함께, 기계학습 모델(870)의 학습 데이터의 일부로서 이용될 수 있다.

[0165] 레이블링 모듈(860)은 기계학습 모델(870)의 학습 데이터의 일부인, 참조 변이 후보에 대한 분류 정보(862)를 레이블링할 수 있다. 분류 정보(862)는 참조 변이 후보가 진양성(True Positive) 변이인지 또는 위양성(False Positive) 변이인지 여부를 나타낼 수 있다.

[0166] 일 실시예에서, FFPE 샘플 내 특정 변이 후보 정보의 적어도 일부와, FF 처리된 샘플 내 변이 후보 중 어느 하나의 변이 후보와 연관된 정보의 적어도 일부가 서로 대응되는 것으로 판단되는 것에 응답하여, 레이블링 모듈(860)은 해당 특정 변이 후보가 진양성 변이인 것으로 레이블링할 수 있다. 즉, 해당 특정 변이 후보는, 샘플 처리 방식(FFPE, FF)의 차이에 따라 발생한 것이 아닌 것으로서 실제 변이인 것으로 결정될 수 있다.

[0167] 예를 들어, FFPE 샘플 내 특정 변이 후보 정보 중, 변이 후보의 위치 정보(예: 변이 후보가 위치한 염색체 정보 및 염색체 내 위치 정보), 변이 후보의 위치에서의 기준 대립유전자(reference allele) 정보 및 변이 후보에 대응되는 변형된 대립유전자(altered allele) 정보는, FF 처리된 샘플 내 임의의 변이 후보와 연관된 정보와 서로 대응 또는 일치할 수 있다. 이 경우, 레이블링 모듈(860)은 해당 특정 변이 후보가 진양성 변이인 것으로 레이블링할 수 있다.

[0168] 특정 변이 후보가 진양성 변이인 경우, 해당 특정 변이 후보는 'TRUE'(즉, 진양성 변이임을 나타냄)로 레이블링되거나, 'FALSE'(즉, FFPE 샘플에서 발견되는 아티팩트가 아님을 나타냄)로 레이블링될 수 있다.

[0169] 반면, FFPE 샘플 내 특정 변이 후보 정보와, FF 처리된 샘플 내 변이 후보 중 어느 하나의 변이 후보와 연관된 정보가 서로 대응되지 않는 것으로 판단되는 것에 응답하여, 레이블링 모듈(860)은 해당 특정 변이 후보가 위양성 변이인 것으로 레이블링할 수 있다.

[0170] 예를 들어, FFPE 샘플 내 변이 후보의 위치 정보, 변이 후보의 위치에서의 기준 대립유전자(reference allele) 정보 및 변이 후보에 대응되는 변형된 대립유전자(altered allele) 정보 중 어느 하나라도 FF 처리된 샘플 내 어떠한 임의의 변이 후보와 연관된 정보와도 서로 대응되지 않을 수 있다. 이 경우, 레이블링 모듈(860)은 해당 특정 변이 후보가 위양성 변이인 것으로 레이블링할 수 있다. 이와 달리, 해당 특정 변이 후보는 레이블링되지 않은 상태(unlabeled)로 유지될 수 있다.

[0171] 일 실시예에서, 레이블링 모듈(860)에서 생성된 분류 정보(862)는 생성 후 수정/변경되지 않을 수 있다. 예를 들어, FFPE 샘플 내 모든 변이 후보가 임계치 이상의 신뢰도로 레이블링되거나, 전체 학습 프로세스가 1회 진행되는 경우, 생성된 분류 정보(862)는 생성 뒤 수정/변경되지 않을 수 있다.

[0172] 다른 실시예에서, 레이블링 모듈(860)에서 생성된 분류 정보(862)는 생성 후 수정/변경될 수 있다. 예를 들어, 분류 정보(862) 중 적어도 일부가 잘못 레이블링된 것이거나, 잘못 레이블링되었을 확률이 임계치 이상인 것으로 판단되는 것에 응답하여, 분류 정보(862)는 생성 후 수정/변경될 수 있다.

[0173] 일 예시에서, 기계학습 모델(870)의 학습 프로세스를 여러 번 반복하는 경우 분류 정보(862)는 생성 후 수정/변경될 수 있다. 다른 예시에서, 학습 데이터 중 임계치 이상의 신뢰도로 레이블링된 데이터를 포함하는 서브 세트를 이용하여, 기계학습 모델(870)에 상응하거나 유사한 교사 모델(teacher model)을 구축 및 학습하고, 학습된 교사 모델을 이용하여, 학습 데이터 중 임계치 이하의 신뢰도로 레이블링된 변이 후보 또는 레이블링되지 않은 변이 후보를 레이블링하는 프로세스(예: noisy label training) 등이 수행되는 경우, 분류 정보(862)는 초기 생성 후에도 수정/변경될 수 있다.

[0174] 기계학습 모델(870)은 FFPE 샘플 내 필터링된 변이 후보 정보(844) 및 변이 후보의 특징(852)을 입력받아, 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과(872)를 결정하고 출력할 수 있다. 트레이닝 모듈(880)은 변이 후보의 분류 결과(872) 및 변이 후보에 레이블링된 분류 정보(862)에 기초하여 기계학습 모델(870)의 파라미터 조정(882)을 수행함으로써, 기계학습 모델(870)을 학습시킬 수 있다.

[0175] 도 8에 도시된 일부 구성은 생략될 수 있다. 예를 들어, 제1 필터 모듈(842)이 생략되는 경우 상술한 FFPE 샘플 내 필터링된 변이 후보 정보(844)를 대체하여 FFPE 샘플 내 변이 후보 정보(822)가 사용될 수 있고, 제2 필터 모듈(846)이 생략되는 경우 상술한 FF 샘플 내 필터링된 변이 후보 정보(848)를 대체하여 FF 샘플 내 변이 후보 정보(824)가 사용될 수 있다.

[0176] 도 9는 본 개시의 일 실시예에 따른 기계학습 모델(960)의 추론 프로세스의 예시를 나타내는 도면이다. 도 9의 변이 검출 모듈(920), 어노테이션 모듈(930), 필터 모듈(940) 및 특징 추출 모듈(950)은 각각 도 8의 변이 검출 모듈(820), 어노테이션 모듈(830), 제1 필터 모듈(842) 및 특징 추출 모듈(850)과 대응되며, 이와 관련하여 도 8을 이용하여 설명한 내용은 그 기재가 생략될 수 있다.

[0177] FFPE 샘플 시퀀싱 데이터(910)는 타겟 샘플 시퀀싱 데이터로서, 학습된 기계학습 모델(960)의 추론 프로세스에서 새로 주어지는 시퀀싱 데이터에 해당할 수 있다. 학습된 기계학습 모델(960)을 이용하여, FFPE 샘플 시퀀싱 데이터(910)로부터 FFPE 샘플의 타겟 변이 후보가 진양성 변이인지 여부(또는, FFPE 처리 등에 의한 아티팩트(artifact)인지 여부)를 나타내는 분류 결과(962)가 출력될 수 있다. 즉, 학습된 기계학습 모델(960)은 샘플 내 존재하는 변이 후보가 실제 변이인지, 또는 FFPE 처리 과정 등의 외부 요인으로 발생한 아티팩트인지 추론할 수 있다.

[0178] 구체적으로, 변이 검출 모듈(920)은 FFPE 샘플 시퀀싱 데이터(910)를 이용하여 FFPE 샘플 내 변이 후보 정보(922)를 출력하고, FFPE 샘플 내 변이 후보 정보(922)는 어노테이션 모듈(930) 및 필터 모듈(940)로 전달될 수 있다.

[0179] 어노테이션 모듈(930)은 FFPE 샘플 내 변이 후보 정보(922)에 기초하여 FFPE 샘플 어노테이션 정보(932)를 생성할 수 있다. 필터 모듈(940)은 FFPE 샘플 내 변이 후보 정보(922) 및 FFPE 샘플 어노테이션 정보(932)를 입력받아 복수의 변이 후보 중 일부를 필터링함으로써, FFPE 샘플 내 필터링된 변이 후보 정보(942)를 생성할 수 있다.

[0180] 특징 추출 모듈(950)은 FFPE 샘플 어노테이션 정보(932)(추가적으로, FFPE 샘플 내 필터링된 변이 후보 정보(942))를 이용하여 변이 후보의 특징(952)을 생성할 수 있다.

[0181] 기계학습 모델(960)은 FFPE 샘플 내 필터링된 변이 후보 정보(942) 및 변이 후보의 특징(952)을 입력받아, 변이 후보의 분류 결과(962)를 출력할 수 있다. 특징 추출 모듈(950)은 기계학습 모델(960)의 학습 프로세스에서 사용된 특징 추출 정보를 이용하여, 동일한 특징 추출 프로세스에 의해 변이 후보의 특징(952)을 추출할 수 있다.

[0182] 특정 변이 후보는, 반복적으로 동일한 위치에 동일한 형태로 발생하는 것이 알려진 생물학적 변이(가령, 핫스팟 변이(hotspot mutation))일 수 있다. 이 경우, 해당 특정 변이 후보는 실제 변이에 해당할 확률이 높으므로, 해당 특정 변이 후보가 진양성 변이에 해당함을 나타내는 레이블이 부여된 채 기계학습 모델(960)에 입력될 수 있다.

[0183] 이후, 분류 결과(962)에 기초하여, 복수의 변이 후보 중 일부가 변이 후보 리스트로부터 필터링/제거됨으로써, FFPE 샘플 내 실제 변이로 추론되는 변이 목록이 결정될 수 있다.

[0184] 도 3 내지 도 9에서 도시된 시스템 내 각 모듈은 예시일 뿐이며, 일부 실시예에서는 도시된 모듈 외 다른 모듈 등의 구성을 추가로 포함할 수 있으며, 일부 구성이 생략될 수도 있다. 예를 들어, 위 내부 구성 중 일부가 생략되는 경우, 생략된 일부 내부 구성의 기능을 다른 모듈 또는 다른 컴퓨팅 장치의 프로세서가 수행하도록 구성될 수 있다. 또한, 도 8 및 도 9에서 각 모듈을 기능별로 구분하여 설명하였으나, 이는 발명의 이해를 돕기 위한 것으로서, 반드시 물리적으로 구분되는 것을 의미하지 않고, 이에 한정되지 않는다.

[0185] 도 10은 본 개시의 일 실시예에 따른 학습 데이터(1000)의 예시를 나타내는 도면이다. 도 10에 도시된 바와 같이, 학습 데이터(1000)는 테이블 형태의 데이터 매트릭스(matrix)로 구성될 수 있다. 도 10에서는 데이터 매트릭스의 행(row)이 참조 변이 후보 각각에 부여된 고유 번호(1 내지 13633)를 나타내고, 열(column)은 참조 변이 후보 정보, 참조 변이 후보의 특징 및 분류 정보를 나타내나, 이러한 형식에 한정되는 것은 아니다. 예를 들어, 학습 데이터는 상술한 데이터 매트릭스의 전치(transpose) 매트릭스로 구현되거나, 다차원 벡터, 배열, 데이터프레임 등의 다양한 형식으로 구현될 수 있다.

[0186] 학습 데이터(1000)는 참조 변이 후보 정보, 변이 후보의 특징 및 참조 변이 후보에 레이블링된 분류 정보를 포함할 수 있다.

[0187] 예를 들어, CHROM, POS, REF, ALT 항목은 참조 변이 후보 정보 중, 순서대로 각각 변이 후보가 위치한 염색체

정보, 염색체 내 위치 정보, 변이 후보의 위치에서의 기준 대립유전자(reference allele) 정보 및 변이 후보에 대응되는 변형된 대립유전자(altered allele) 정보를 나타낼 수 있다.

[0188] 'FILTER'로 시작되는 복수의 항목은, 참조 변이 후보 정보 중 변이 검출 모듈(예: 도 3의 310)에서 생성된 필터 정보를 나타낼 수 있다. 예를 들어, 도 10의 학습 데이터(1000)에 표시된 모든 변이 후보는, 변이 검출 모듈의 모든 필터를 통과한 것으로 볼 수 있다(FILTER_PASS: 1).

[0189] 'label' 항목은 참조 변이 후보에 레이블링된 분류 정보를 나타낼 수 있다.

[0190] 상술한 항목 외의 항목은 어노테이션 정보로부터 추출된 변이 후보의 특징을 나타낼 수 있다. 변이 후보의 특징을 나타내는 항목 각각이 지칭하는 정보의 종류는 도 4에서 상술한 바 있다.

[0191] 학습 데이터(1000)는 복수의 샘플에 대한 복수의 데이터 세트가 하나의 데이터 세트로 결합(concatenate)된 것일 수 있다. 이 경우 학습 데이터(1000)의 데이터 매트릭스의 행(row)을 기준으로 학습 데이터가 복수의 배치로 분할된 뒤, 기계학습 모델의 복수의 분류기 각각에 입력될 수 있다.

[0192] 도 11은 본 개시의 일 실시예에 따른 학습된 기계학습 모델의 성능(performance) 평가 결과를 나타내는 도면이다. 그래프(1112, 1114, 1122, 1124, 1132, 1134)에 표시된 복수의 점(dot)의 각각은 기계학습 모델을 이용한 변이 후보의 필터링 전후, FF 샘플을 이용하여 측정된 변량(x축)의 값 및 FF 샘플과 대응되는 FFPE 샘플을 이용하여 측정된 변량(y축)의 값을 하나의 점으로 나타낸 것이다. 즉, 복수의 점(dot)의 각각이 각 그래프에 표시된 y=x 선에 가깝게 위치할수록, FF 샘플을 이용하여 측정된 변량 값과 이에 대응하는 FFPE 샘플을 이용하여 측정된 변량 값이 유사한 것으로 판단될 수 있다. 복수의 점의 각각은, 서로 다른 복수의 개체의 각각으로부터 채취된 FFPE 샘플 및 FF 샘플을 이용하여 측정된 변량을 나타낼 수 있다.

[0193] 구체적으로, 제1 그래프(1112) 및 제2 그래프(1114) 내 복수의 점(dot)의 각각은, 기계학습 모델을 이용한 변이 후보의 필터링 전과 필터링 후, FF 샘플을 이용하여 측정된 SNV의 개수 및 FF 샘플과 대응되는 FFPE 샘플을 이용하여 측정된 SNV의 개수를 나타낸 것이다. 이와 유사하게, 제3 그래프(1122) 및 제4 그래프(1124) 내 복수의 점(dot)의 각각은, 기계학습 모델을 이용한 변이 후보의 필터링 전과 필터링 후, FF 샘플을 이용하여 측정된 INDEL의 개수와, FF 샘플과 대응되는 FFPE 샘플을 이용하여 측정된 INDEL의 개수를 나타낸 것이다.

[0194] 제1 그래프(1112)를 참조하면, 기계학습 모델을 이용하여 변이 후보를 필터링하기 전, FF 샘플을 이용하여 측정된 SNV의 개수보다 이와 대응되는 FFPE 샘플을 이용하여 측정된 SNV의 개수가 많은 경향성을 확인할 수 있다. 이와 유사하게, 제3 그래프(1122)를 참조하면, 기계학습 모델을 이용하여 변이 후보를 필터링하기 전, FF 샘플을 이용하여 측정된 INDEL의 개수보다 이와 대응되는 FFPE 샘플을 이용하여 측정된 INDEL의 개수가 많은 경향성을 확인할 수 있다. 이는 FFPE 처리 과정으로 인해 샘플 내에서 발생하는 교차 결합(cross-linking), 단편화(fragmentation), 기타 비생물학적 원인으로 인한 염기의 변이 등, 다양한 유형의 손상으로부터 기인할 수 있다.

[0195] 반면, 제2 그래프(1114)를 참조하면, 기계학습 모델을 이용하여 변이 후보를 필터링한 뒤, FF 샘플을 이용하여 측정된 SNV의 개수와, 이와 대응되는 FFPE 샘플을 이용하여 측정된 SNV의 개수 간의 차이가 감소한 것을 확인할 수 있다(즉, y=x line 방향으로 데이터가 정렬되는 경향). 마찬가지로, 제4 그래프(1124)를 참조하면, 기계학습 모델을 이용하여 변이 후보를 필터링한 뒤, FF 샘플을 이용하여 측정된 INDEL의 개수와, 이와 대응되는 FFPE 샘플을 이용하여 측정된 INDEL의 개수 간의 차이가 감소한 것을 확인할 수 있다. 즉, 제2 그래프(1114) 및 제4 그래프(1124)를 통해, 기계학습 모델을 이용하여 세포 샘플 내 위양성(False Positive) 변이를 필터링함으로써 FFPE 처리 과정으로 인해 발생하는 노이즈(또는, 아티팩트)가 제거된 것을 확인할 수 있다.

[0196] 제5 그래프(1132) 및 제6 그래프(1134) 내 복수의 점(dot)의 각각은, 기계학습 모델을 이용한 변이 후보의 필터링 전과 필터링 후, FF 샘플을 이용하여 측정된 HRD(Homologous Recombination Deficiency) score 및 FF 샘플과 대응되는 FFPE 샘플을 이용하여 측정된 HRD score를 나타낸 것이다.

[0197] 제6 그래프(1134)를 참조하면, 기계학습 모델을 이용하여 변이 후보를 필터링한 뒤, FF 샘플을 이용하여 측정된 HRD 점수와, 이와 대응되는 FFPE 샘플을 이용하여 측정된 HRD 점수 간의 차이가 감소한 것을 확인할 수 있다.

[0198] 아래 표 3은 본 개시의 기계학습 모델을 이용한 변이 후보 필터링의 수행에 따른, 단일 염기 변이(SNV), 염기 삽입 변이 및 결실 변이(INDEL)와 연관된 성능 평가 지표를 나타낸다.

표 3

[0199]

	민감도	특이도	PPV	F1
단일 염기 변이(SNV)	0.97	0.87	0.91	0.94
염기 삽입 및 결실 변이 (INDEL)	0.91	0.91	0.92	0.91

[0200] 표 3에서, 민감도(Sensitivity)는 TP / (TP + FN), 특이도(specificity)는 TN / (TN + FP), PPV(Positive Predictive Value, 양성 예측도)는 TP / (TP + FP), F1-score는 2 * 민감도 * 양성 예측도 / (민감도 + 양성 예측도)로 정의된다. 이 때, TP(True Positive), TN(True Negative), FP(False Positive) 및 FN(False Negative)의 각각은, FFPE 샘플을 이용하여 결정된 변이 후보에 대해 본 개시에 따른 기계학습 모델을 이용한 분류 결과와 실제 분류 정보(Ground Truth)를 비교한 결과, 실제 변이로 옳게 예측된 개수, 실제 변이가 아닌 것으로 옳게 예측된 개수, 실제 변이인 것으로 틀리게 예측된 개수 및 실제 변이가 아닌 것으로 틀리게 예측된 개수를 지칭할 수 있다.

[0201] 아래 표 4는 본 개시에 따른 기계학습 모델을 이용한 변이 후보 필터링 수행 전 산출된 일치도(Concordance) 지표를 나타내고, 표 5는 본 개시에 따른 기계학습 모델을 이용한 변이 후보 필터링 수행 후 산출된 일치도 지표를 나타낸다.

표 4

[0202]

	일치도	95% 신뢰구간
HRD	0.60	[0.47, 0.70]
TMB	SNV: 0.01 INDEL: 0.00	SNV: [0.00, 0.02] INDEL: [0.00, 0.00]

표 5

[0204]

	일치도	95% 신뢰구간
HRD	0.99	[0.98, 0.99]
TMB	SNV: 0.96 INDEL: 0.87	SNV: [0.93, 0.98] INDEL: [0.78, 0.92]

[0205] 표 4 및 표 5의 일치도는 두 변수가 얼마나 유사한 값과 값의 경향을 보이는지를 나타내는 지표로, Lin's concordance correlation coefficient 값으로 정의될 수 있다. Lin's concordance correlation coefficient 값으로 일치도가 정의되는 경우, 그래프(1112, 1114, 1122, 1124, 1132, 1134) 각각의 복수의 점이 y=x line에 가까울수록 일치도는 1에 가까운 값을 갖고, y=x line에서 멀어질수록 0에 가까운 값을 갖는다. 또한, 설령 복수의 점의 분포 경향이 선형이더라도 y=x line으로부터 복수의 점의 분포가 멀어질수록 일치도 값이 0에 가까워질 수 있다.

[0206] 표 4 및 표 5를 참조하면, 기계학습 모델을 이용한 변이 후보 필터링 수행 전후 HRD의 일치도가 0.60에서 0.99로 비약적으로 향상된 것을 확인할 수 있으며, TMB(Tumor Mutation Burden, 종양 변이 부하) 또한 SNV에 대해 0.01에서 0.96으로, INDEL에 대해 0.00에서 0.87로 일치도가 크게 향상된 것을 확인할 수 있다.

[0207] 도 12는 본 개시의 일 실시예에 따른 인공신경망 모델(1200)을 나타내는 예시도이다. 인공신경망 모델(1200)은, 기계학습 모델의 일 예로서, 기계학습(Machine Learning) 기술과 생물학적 신경망의 구조에 기초하여 구현된 통계학적 학습 알고리즘 또는 그 알고리즘을 실행하는 구조이다.

[0208] 일 실시예에 따르면, 상술한 기계학습 모델(또는, 기계학습 모델 내 분류기)은 인공신경망 모델(1200)의 형태로 생성될 수 있다. 예를 들어, 인공신경망 모델(1200)은 변이 후보 정보 및 변이 후보의 특징을 수신하고, 변이 후보가 진양성(True Positive) 변이인지(또는, 변이 후보가 FFPE 처리에 의한 artifact인지) 여부를 나타내는 분류 결과를 출력할 수 있다.

[0209] 일 실시예에 따르면, 인공신경망 모델(1200)은, 생물학적 신경망에서와 같이 시냅스의 결합으로 네트워크를 형

성한 인공 뉴런인 노드(Node)들이 시냅스의 가중치를 반복적으로 조정하여, 특정 입력에 대응한 올바른 출력과 추론된 출력 사이의 오차가 감소되도록 학습함으로써, 문제 해결 능력을 가지는 기계학습 모델을 나타낼 수 있다. 예를 들어, 인공신경망 모델(1200)은 기계학습, 딥러닝 등의 인공지능 학습법에 사용되는 임의의 확률 모델, 뉴럴 네트워크 모델 등을 포함할 수 있다.

[0210] 일 실시예에서, 인공신경망 모델(1200)은 다층의 노드들과 이들 사이의 연결로 구성된 다층 퍼셉트론(MLP: multilayer perceptron)으로 구현될 수 있다. 본 실시예에 따른 인공신경망 모델(1200)은 MLP를 포함하는 다양한 인공신경망 모델 구조들 중의 하나를 이용하여 구현될 수 있으나, 이에 한정되는 것은 아니다. 도 4에 도시된 바와 같이, 인공신경망 모델(1200)은, 외부로부터 입력 신호 또는 데이터(1210)를 수신하는 입력층(1220), 입력 데이터에 대응한 출력 신호 또는 데이터(1250)를 출력하는 출력층(1240), 입력층(1220)과 출력층(1240) 사이에 위치하며 입력층(1220)으로부터 신호를 받아 특성을 추출하여 출력층(1240)으로 전달하는 n개(여기서, n은 양의 정수)의 은닉층(1230_1 내지 1230_n)으로 구성된다. 여기서, 출력층(1240)은 은닉층(1230_1 내지 1230_n)으로부터 신호를 받아 외부로 출력한다.

[0211] 인공신경망 모델(1200)의 학습 방법에는, 교사 신호(정답)의 입력에 의해서 문제의 해결에 최적화되도록 학습하는 지도 학습(Supervised Learning) 방법과, 교사 신호를 필요로 하지 않는 비지도 학습(Unsupervised Learning) 방법이 있다. 일 실시예에 따르면, 인공신경망 모델(1200)은 참조 변이 후보 정보, 변이 후보의 특징 및 참조 변이 후보에 레이블링된 분류 정보를 포함하는 학습 데이터 세트에 기초하여 학습될 수 있으며, 학습 데이터 세트 중 분류 정보를 교사 신호로 하는 지도 학습 방식에 의해 학습될 수 있다.

[0212] 일 실시예에 따르면, 인공신경망 모델(1200)의 입력변수는, 변이 후보 정보 및 변이 후보의 특징을 포함할 수 있다. 이와 같이 상술된 입력변수가 입력층(1220)을 통해 입력되는 경우, 인공신경망 모델(1200)의 출력층(1240)에서 출력되는 출력변수는 변이 후보가 진양성(True Positive) 변이인지(또는, 변이 후보가 FFPE 처리에 의한 아티팩트(artifact)인지) 여부를 나타내는 분류 결과가 될 수 있다.

[0213] 이와 같이, 인공신경망 모델(1200)의 입력층(1220)과 출력층(1240)에 복수의 입력변수와 대응되는 복수의 출력변수가 각각 매칭되고, 입력층(1220), 은닉층(1230_1 내지 1230_n) 및 출력층(1240)에 포함된 노드들 사이의 시냅스 값이 조정됨으로써, 특정 입력에 대응한 올바른 출력이 추출될 수 있도록 학습될 수 있다. 이러한 학습 과정을 통해, 인공신경망 모델(1200)의 입력변수에 숨겨져 있는 특성을 파악할 수 있고, 입력변수에 기초하여 계산된 출력변수와 목표 출력(예: 레이블링된 분류 정보) 간의 오차가 줄어들도록 인공신경망 모델(1200)의 노드들 사이의 시냅스 값(또는 가중치)를 조정할 수 있다. 또한, 인공신경망 모델(1200)은 변이 후보 정보 및 변이 후보의 특징을 입력으로 받는 알고리즘을 학습하며, 분류 정보와의 손실(loss)을 최소화하는 방식으로 학습될 수 있다.

[0214] 이렇게 학습된 인공신경망 모델(1200)을 이용하여, 변이 후보가 진양성 변이인지(또는, 변이 후보가 FFPE 처리에 의한 아티팩트인지) 여부를 나타내는 분류 결과가 추출될 수 있다.

[0215] 도 13은 본 개시의 일 실시예에 따른 기계학습 모델을 학습시키는 방법(1300)을 나타내는 흐름도이다. 방법(1300)은 적어도 하나의 프로세서에 의해 수행될 수 있다. 방법(1300)은 프로세서가 참조 샘플 내 참조 변이 후보 정보를 결정함으로써 개시될 수 있다(S1310). 참조 변이 후보는, SNV와 INDEL을 아우르는 점 돌연변이(Point Mutation)에 해당할 확률이 미리 정해진 임계 확률 이상인 시퀀스일 수 있다. 참조 변이 후보 정보는, 참조 변이 후보의 위치 정보, 참조 변이 후보의 위치에서의 기준 대립유전자(reference allele) 정보 또는 참조 변이 후보에 대응되는 변형된 대립유전자(altered allele) 정보 중 적어도 하나를 포함할 수 있다.

[0216] 일 실시예에서, 참조 샘플은, 동일한 개체로부터 채취된 참조 정상 샘플 및 참조 이상 샘플을 포함하고, 참조 시퀀싱 데이터는, 참조 정상 샘플과 연관된 제1 참조 시퀀싱 데이터 및 참조 이상 샘플과 연관된 제2 참조 시퀀싱 데이터를 포함하고, 프로세서는 변이 검출 모듈을 이용하여, 제1 참조 시퀀싱 데이터 및 제2 참조 시퀀싱 데이터를 기초로 참조 변이 후보 정보를 결정할 수 있다.

[0217] 일 실시예에서, 변이 검출 모듈은 복수의 검출 모듈을 포함하고, 프로세서는 복수의 검출 모듈의 각각에 제1 참조 시퀀싱 데이터 및 제2 참조 시퀀싱 데이터를 적용함으로써, 참조 변이 서브 후보 정보를 결정하고, 참조 변이 서브 후보 정보를 통합(union)함으로써, 참조 변이 후보 정보를 결정할 수 있다. 이와 달리, 프로세서는 복수의 검출 모듈의 각각에 제1 참조 시퀀싱 데이터 및 제2 참조 시퀀싱 데이터를 적용함으로써, 참조 변이 서브 후보 정보를 결정하고, 복수의 검출 모듈 중 둘 이상의 검출 모듈에서 공통적으로 결정된 참조 변이 서브 후보 정보를 참조 변이 후보 정보로 결정할 수 있다.

[0218] 프로세서는 기계학습 모델의 학습을 위한 어노테이션 정보를 생성할 수 있다(S1330).

[0219] 일 실시예에서, 프로세서는 매핑된 위치 중 적어도 일부가 참조 변이 후보의 위치와 중첩되는 복수의 리드(read)를 결정하고, 결정된 복수의 리드와 연관된 제1 어노테이션 정보를 생성할 수 있다. 예를 들어, 복수의 리드는 레퍼런스 지놈과 상이한 복수의 변이 리드(variant read)를 포함하고, 제1 어노테이션 정보는, 복수의 변이 리드의 인서트 사이즈(insert size)의 최솟값, 복수의 변이 리드의 인서트 사이즈의 최댓값, 또는 복수의 변이 리드 중 특정 조건을 만족하는 페어드 리드(paired read)의 개수 중 적어도 하나를 포함하고, 특정 조건은, 페어드 리드의 제1 리드 및 제2 리드가 각각 정방향과 역방향으로 정렬되고, 페어드 리드의 인서트 사이즈가 하한임계치와 상한임계치 사이인 조건을 포함할 수 있다.

[0220] 일 실시예에서, 프로세서는 복수의 정상 샘플과 연관된 복수의 시퀀싱 데이터로부터 생성된 정상 조직 유전체 데이터(PON: Panel of Normals)를 수신하고, 정상 조직 유전체 데이터와 연관된 제2 어노테이션 정보를 생성할 수 있다.

[0221] 일 실시예에서, 프로세서는 FFPE(Formalin-Fixed, Paraffin-Embedded) 처리된 복수의 샘플과 연관된 복수의 시퀀싱 데이터로부터 생성된 FFPE 처리 조직 유전체 데이터(POF: Panel of FFPEs)를 수신하고, FFPE 처리 조직 유전체 데이터와 연관된 제3 어노테이션 정보를 생성할 수 있다. 제3 어노테이션 정보는, FFPE 처리된 복수의 샘플 중, 샘플 내 염기 서열 상의 특정 위치(position)에 대한 VAF(Variant Allele Frequency)가 미리 정해진 임계치 미만인 샘플의 수를 포함할 수 있다. 제3 어노테이션 정보는, FFPE 처리된 복수의 샘플 중, 샘플 내 염기 서열 상의 특정 위치에서 미리 정해진 개수의 변이 리드를 갖는 샘플의 수를 포함할 수 있다.

[0222] 일 실시예에서, 프로세서는 참조 변이 후보의 변이 유형과 연관된 정보 및 참조 변이 후보의 시퀀스 컨텍스트(sequence context) 정보를 포함하는 제4 어노테이션 정보를 생성할 수 있다.

[0223] 프로세서는 결정된 참조 변이 후보 정보 및 생성된 어노테이션 정보에 기초하여 학습 데이터를 생성할 수 있다(S1340).

[0224] 일 실시예에서, 프로세서는 참조 변이 후보에 대한 분류 정보를 레이블링할 수 있다. 예를 들어, 참조 샘플은 FFPE 처리된 샘플이고, 프로세서는 참조 변이 후보 정보의 적어도 일부와, FF(Fresh-Frozen) 처리된 샘플 내 변이 후보 중 어느 하나의 변이 후보와 연관된 정보의 적어도 일부가 서로 대응되는 것으로 판단되는 것에 응답하여, 참조 변이 후보가 진양성(True Positive) 변이인 것으로 레이블링할 수 있다. 이 때, FF 처리된 샘플은 FFPE 처리된 샘플에 대응되는 샘플일 수 있다.

[0225] 일 실시예에서, 프로세서는 참조 변이 후보 정보와, FF 처리된 샘플 내 변이 후보 중 어느 하나의 변이 후보와 연관된 정보가 서로 대응되지 않는 것으로 판단되는 것에 응답하여, 참조 변이 후보가 위양성(False positive) 변이인 것으로 레이블링할 수 있다.

[0226] 일 실시예에서, 프로세서는 참조 변이 후보 정보 및 어노테이션 정보에 기초하여, 참조 변이 후보의 특징을 추출하고, 참조 변이 후보 정보, 추출된 참조 변이 후보의 특징 및 레이블링된 분류 정보를 포함하는 데이터 세트를 학습 데이터에 포함시킴으로써, 학습 데이터를 생성할 수 있다.

[0227] 프로세서는 생성된 학습 데이터를 이용하여 기계학습 모델을 학습시킬 수 있다(S1350). 일 실시예에서, 기계학습 모델은 복수의 분류기(classifier)를 포함하고, 프로세서는 참조 변이 후보 정보 및 참조 변이 후보의 특징을 복수의 분류기의 각각에 입력하고, 복수의 분류기 중 적어도 하나의 분류기로부터의 출력 결과를 이용하여 참조 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 결정하고, 분류 결과와 참조 변이 후보에 레이블링된 분류 정보에 기초하여 기계학습 모델의 파라미터 내지 하이퍼 파라미터를 조정할 수 있다.

[0228] 일 실시예에서, 기계학습 모델은, 타겟 샘플 내 타겟 변이 후보 정보 및 타겟 변이 후보의 특징을 수신하여, 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 출력할 수 있다. 타겟 샘플은, 동일한 개체로부터 채취된 타겟 정상 샘플 및 타겟 이상 샘플을 포함하고, 타겟 변이 후보 정보는, 변이 검출 모듈을 이용하여, 타겟 정상 샘플과 연관된 제1 타겟 시퀀싱 데이터 및 타겟 이상 샘플과 연관된 제2 타겟 시퀀싱 데이터를 기초로 결정될 수 있다. 타겟 샘플은 FFPE 처리된 샘플일 수 있다.

[0229] 도 14는 본 개시의 일 실시예에 따른 세포 샘플 내 진양성 변이 검출을 통한 유전체 프로파일링 방법(1400)을 나타내는 흐름도이다. 방법(1400)은 적어도 하나의 프로세서에 의해 수행될 수 있다. 방법(1400)은 프로세서가 타겟 샘플의 타겟 변이 후보 정보를 획득함으로써 개시될 수 있다(S1410). 타겟 샘플은 FFPE 처리된 샘플일 수 있다.

[0230] 일 실시예에서, 타겟 샘플은 타겟 이상 샘플을 포함하고, 타겟 변이 후보 정보는 타겟 이상 샘플과 연관된 타겟 시퀀싱 데이터를 기초로 결정될 수 있다. 가령, 타겟 변이 후보 정보는 딥 시퀀싱(Deep Sequencing), 알려진 변이 데이터베이스와의 대조 등을 수행하여, 타겟 이상 샘플과 연관된 타겟 시퀀싱 데이터를 기초로 결정될 수 있다.

[0231] 다른 실시예에서, 타겟 샘플은 동일한 개체로부터 채취된 타겟 정상 샘플 및 타겟 이상 샘플을 포함하고, 타겟 변이 후보 정보는, 변이 검출 모듈을 이용하여, 타겟 정상 샘플과 연관된 제1 타겟 시퀀싱 데이터 및 타겟 이상 샘플과 연관된 제2 타겟 시퀀싱 데이터를 기초로 결정될 수 있다.

[0232] 이후, 프로세서는 기계학습 모델을 이용하여, 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 결정할 수 있다(S1420). 기계학습 모델은, 도 13의 기계학습 모델을 학습시키는 방법(1300)에 의해 학습된 모델일 수 있다. 예를 들어, 기계학습 모델은, 참조 샘플의 참조 변이 후보 정보 및 참조 변이 후보와 연관된 어노테이션 정보를 이용하여, 참조 변이 후보가 진양성 변이인지 여부를 결정하도록 학습된 모델일 수 있다. 이를 통해, 전장유전체 분석 수행 시 발생할 수 있는 노이즈 내지 오차가 보정되어 왜곡되지 않은 분석 결과가 도출될 수 있다.

[0233] 이후, 프로세서는 결정된 분류 결과를 기초로 타겟 샘플에 대한 유전체 프로파일링(Genomic Profiling)을 수행할 수 있다(S1430). 가령, 프로세서는 결정된 분류 결과를 기초로 일부 변이 후보(예: 위양성 변이로 결정된 변이 후보)가 삭제/필터링된 변이 리스트를 이용하여, 타겟 샘플에 대한 유전체 프로파일링(Genomic Profiling)을 수행할 수 있다.

[0234] 결정된 분류 결과를 기초로 수행되는 유전체 프로파일링의 일 예시로서, 프로세서는 타겟 샘플의 유전체에서 SNV, INDEL, 그리고 염색체 재배열과 같은 유전적 변이를 식별하고 분석할 수 있다. 다른 예시로서, 프로세서는 개체의 유전자가 개체 내에서 어떻게, 언제, 얼마나 발현되는지 조사 및 식별할 수 있다. 다른 예시로서, 프로세서는 DNA 메틸화, 히스톤 변형과 같은 에피게네틱(epigenetic) 변화를 분석하여 개체의 유전자 발현에 영향을 미치는 요소를 분석할 수 있다. 다른 예시로서, 프로세서는 큰 규모의 염색체 재배열이나 복제 수 변화 등을 조사하여, 개체의 유전체 구조 변화를 파악할 수 있다. 다른 예시로서, 프로세서는 개체의 유전적 특성을 이해하고, 질병의 위험성, 약물 반응성 등을 예측할 수 있다.

[0235] 상술한 유전체 프로파일링은 질병의 조기 진단, 개인 맞춤형 의학, 유전 질환 연구, 약물 개발 등 다양한 분야에서 응용될 수 있으며, 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 기초로 유전체 프로파일링을 수행함으로써, 의학 연구와 임상적 응용 등에 있어 정밀하고 광범위한 유전 정보가 제공될 수 있다.

[0236] 이후, 프로세서는 유전체 프로파일링의 수행 결과에 기초하여, 타겟 샘플이 채취된 개체의 질병 진단 정보, 치료 전략 정보, 예후 예측 정보 또는 약물 반응성 예측 정보 중 적어도 하나를 제공할 수 있다(S1440). 가령, 프로세서는 S1430 단계에서 수행된 유전체 프로파일링 결과, 타겟 샘플에 BRCA1과 BRCA2 유전자 변이가 존재하는 것을 확인하고, 타겟 샘플이 채취된 개체가 유방암과 난소암의 발병 위험도가 높은 것으로 진단할 수 있다(질병 진단 정보를 제공하는 예시). 프로세서는 개체의 유전체 프로파일링 수행 결과를 기초로 치료 시기, 치료 방법, 사용 약품 등을 결정함으로써 질병을 보다 효과적으로 치료할 수 있다(치료 전략 정보를 제공하는 예시). 프로세서는 질병의 가능성 있는 경과를 예측함으로써 예후 예측 정보를 제공할 수 있다. 프로세서는 개체의 유전자 구성에 따라 서로 다른 약품의 종류, 사용량 등에 개체가 어떻게 반응할 것인지를 예측하고 약품의 처방을 조정/결정함으로써, 약품의 효능을 높이고 부작용을 줄일 수 있다(약물 반응성 예측 정보를 제공하는 예시).

[0237] 도 13 및 14에서 도시한 흐름도 및 상술한 설명은 하나의 예시일 뿐이며, 일부 실시예에서는 다르게 구현될 수 있다. 예를 들어, 하나 이상의 단계가 생략되거나, 각 단계의 순서가 바뀌거나, 하나 이상의 단계가 중첩되어 수행되거나, 하나 이상의 단계가 여러 번 반복 수행될 수 있다.

[0238] 상술한 방법은 컴퓨터에서 실행하기 위해 컴퓨터 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램으로 제공될 수 있다. 매체는 컴퓨터로 실행 가능한 프로그램을 계속 저장하거나, 실행 또는 다운로드를 위해 임시 저장하는 것일 수도 있다. 또한, 매체는 단일 또는 수개 하드웨어가 결합된 형태의 다양한 기록수단 또는 저장수단일 수 있는데, 어떤 컴퓨터 시스템에 직접 접속되는 매체에 한정되지 않고, 네트워크 상에 분산 존재하는 것일 수도 있다. 매체의 예시로는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM 및 DVD 와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical medium), 및 ROM, RAM, 플래시 메모리 등을 포함하여 프로그램 명령어가 저장되도록 구성된 것이 있을 수 있다. 또한, 다른 매체

의 예시로, 애플리케이션을 유통하는 앱 스토어나 기타 다양한 소프트웨어를 공급 내지 유통하는 사이트, 서버 등에서 관리하는 기록매체 내지 저장매체도 들 수 있다.

[0239] 본 개시의 방법, 동작 또는 기법들은 다양한 수단에 의해 구현될 수도 있다. 예를 들어, 이러한 기법들은 하드웨어, 펌웨어, 소프트웨어, 또는 이들의 조합으로 구현될 수도 있다. 본원의 개시와 연계하여 설명된 다양한 예시적인 논리적 블록들, 모듈들, 회로들, 및 알고리즘 단계들은 전자 하드웨어, 컴퓨터 소프트웨어, 또는 양자의 조합들로 구현될 수도 있음을 통상의 기술자들은 이해할 것이다. 하드웨어 및 소프트웨어의 이러한 상호 대체를 명확하게 설명하기 위해, 다양한 예시적인 구성요소들, 블록들, 모듈들, 회로들, 및 단계들이 그들의 기능적 관점에서 일반적으로 위에서 설명되었다. 그러한 기능이 하드웨어로서 구현되는지 또는 소프트웨어로서 구현되는지의 여부는, 특정 애플리케이션 및 전체 시스템에 부과되는 설계 요구사항들에 따라 달라진다. 통상의 기술자들은 각각의 특정 애플리케이션을 위해 다양한 방식들로 설명된 기능을 구현할 수도 있으나, 그러한 구현들은 본 개시의 범위로부터 벗어나게 하는 것으로 해석되어서는 안된다.

[0240] 하드웨어 구현에서, 기법들을 수행하는 데 이용되는 프로세싱 유닛들은, 하나 이상의 ASIC들, DSP들, 디지털 신호 프로세싱 디바이스들(digital signal processing devices; DSPD들), 프로그램가능 논리 디바이스들(programmable logic devices; PLD들), 필드 프로그램가능 게이트 어레이들(field programmable gate arrays; FPGA들), 프로세서들, 제어기들, 마이크로제어기들, 마이크로프로세서들, 전자 디바이스들, 본 개시에 설명된 기능들을 수행하도록 설계된 다른 전자 유닛들, 컴퓨터, 또는 이들의 조합 내에서 구현될 수도 있다.

[0241] 따라서, 본 개시와 연계하여 설명된 다양한 예시적인 논리 블록들, 모듈들, 및 회로들은 범용 프로세서, DSP, ASIC, FPGA나 다른 프로그램 가능 논리 디바이스, 이산 게이트나 트랜지스터 로직, 이산 하드웨어 컴포넌트들, 또는 본원에 설명된 기능들을 수행하도록 설계된 것들의 임의의 조합으로 구현되거나 수행될 수도 있다. 범용 프로세서는 마이크로프로세서일 수도 있지만, 대안으로, 프로세서는 임의의 종래의 프로세서, 제어기, 마이크로제어기, 또는 상태 머신일 수도 있다. 프로세서는 또한, 컴퓨팅 디바이스들의 조합, 예를 들면, DSP와 마이크로프로세서, 복수의 마이크로프로세서들, DSP 코어와 연계한 하나 이상의 마이크로프로세서들, 또는 임의의 다른 구성의 조합으로서 구현될 수도 있다.

[0242] 펌웨어 및/또는 소프트웨어 구현에 있어서, 기법들은 랜덤 액세스 메모리(random access memory; RAM), 판독 전용 메모리(read-only memory; ROM), 비휘발성 RAM(non-volatile random access memory; NVRAM), PROM(programmable read-only memory), EPROM(erasable programmable read-only memory), EEPROM(electrically erasable PROM), 플래시 메모리, 컴팩트 디스크(compact disc; CD), 자기 또는 광학 데이터 스토리지 디바이스 등과 같은 컴퓨터 판독가능 매체 상에 저장된 명령들로서 구현될 수도 있다. 명령들은 하나 이상의 프로세서들에 의해 실행 가능할 수도 있고, 프로세서(들)로 하여금 본 개시에 설명된 기능의 특정 양태들을 수행하게 할 수도 있다.

[0243] 이상 설명된 실시예들이 하나 이상의 독립형 컴퓨터 시스템에서 현재 개시된 주제의 양태들을 활용하는 것으로 기술되었으나, 본 개시는 이에 한정되지 않고, 네트워크나 분산 컴퓨팅 환경과 같은 임의의 컴퓨팅 환경과 연계하여 구현될 수도 있다. 또 나아가, 본 개시에서 주제의 양상들은 복수의 프로세싱 칩들이나 장치들에서 구현될 수도 있고, 스토리지는 복수의 장치들에 걸쳐 유사하게 영향을 받게 될 수도 있다. 이러한 장치들은 PC들, 네트워크 서버들, 및 휴대용 장치들을 포함할 수도 있다.

[0244] 본 명세서에서는 본 개시가 일부 실시예들과 관련하여 설명되었지만, 본 개시의 발명이 속하는 기술분야의 통상의 기술자가 이해할 수 있는 본 개시의 범위를 벗어나지 않는 범위에서 다양한 변형 및 변경이 이루어질 수 있다. 또한, 그러한 변형 및 변경은 본 명세서에 첨부된 특허청구의 범위 내에 속하는 것으로 생각되어야 한다.

부호의 설명

[0245] 110: 개체 112: 이상 샘플 시퀀싱 데이터

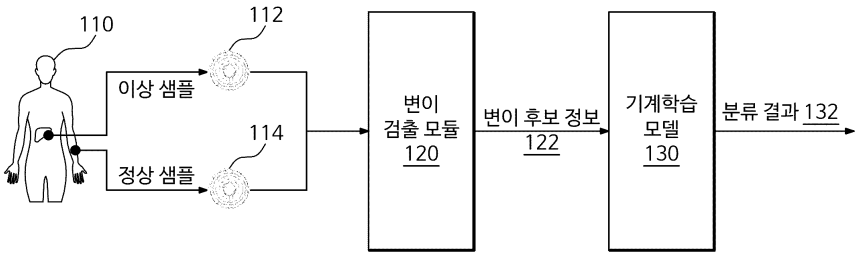
114: 정상 샘플 시퀀싱 데이터

120: 변이 검출 모듈 122: 변이 후보 정보

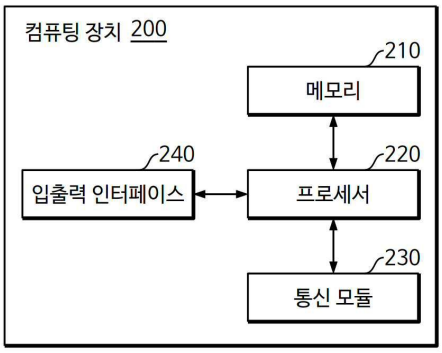
130: 기계학습 모델 132: 분류 결과

도면

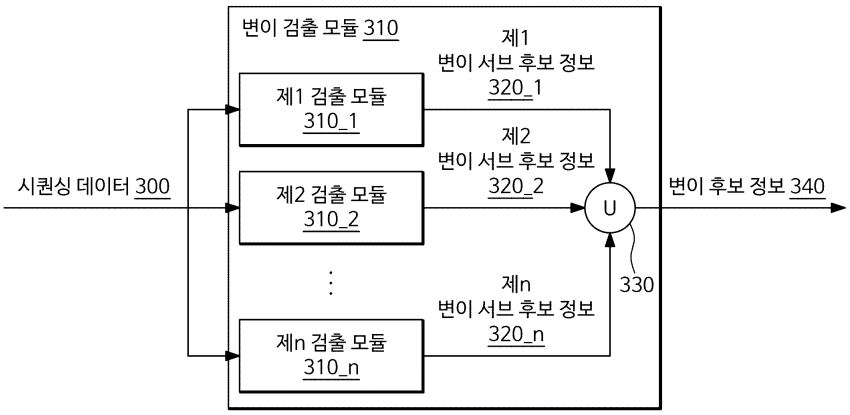
도면1



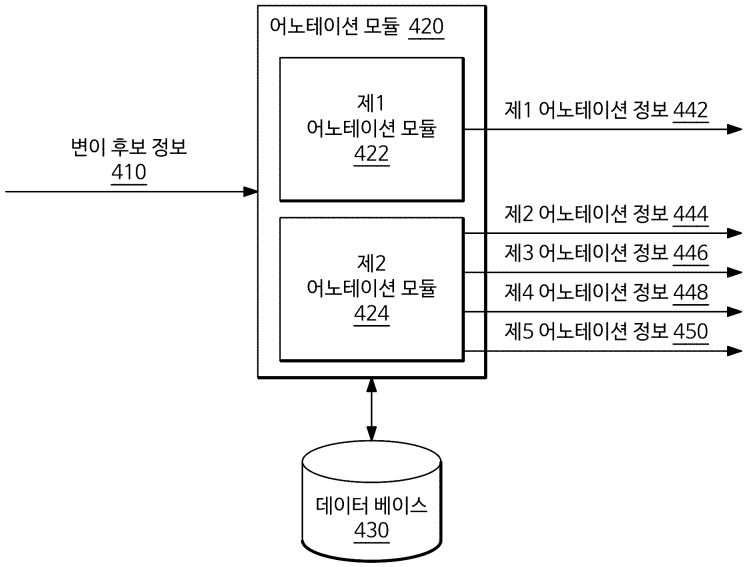
도면2



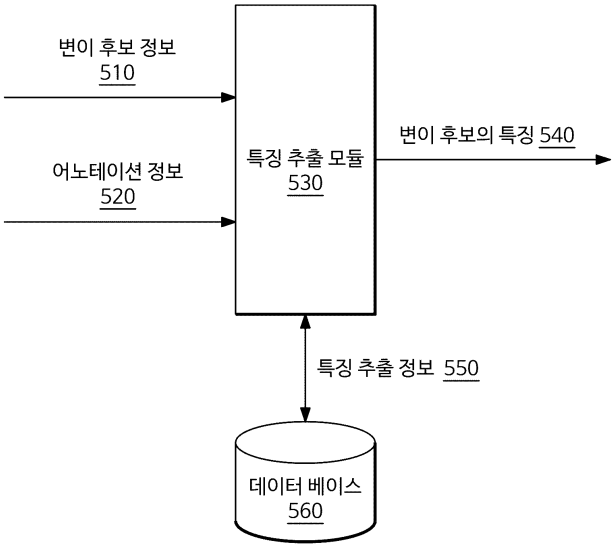
도면3



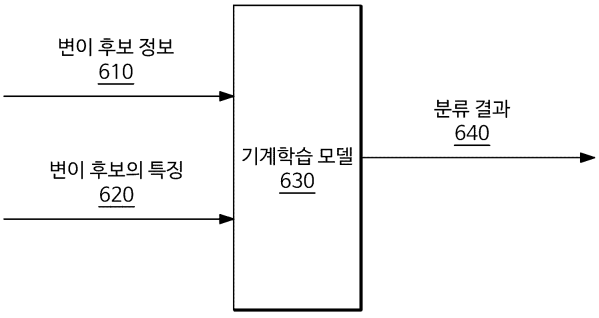
도면4


어노테이션 모듈 420
제1
어노테이션 모듈
422
제2
어노테이션 모듈
424
변이 후보 정보
410
제1 어노테이션 정보 442
제2 어노테이션 정보 444
제3 어노테이션 정보 446
제4 어노테이션 정보 448
제5 어노테이션 정보 450
데이터 베이스
430

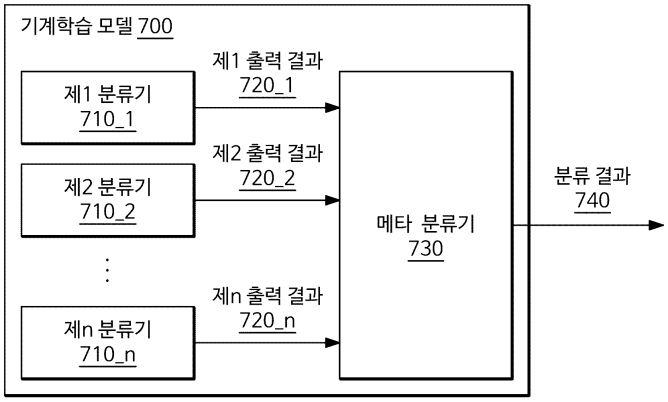
도면5


변이 후보 정보
510
어노테이션 정보
520
특징 추출 모듈
530
변이 후보의 특징 540
특징 추출 정보 550
데이터 베이스
560

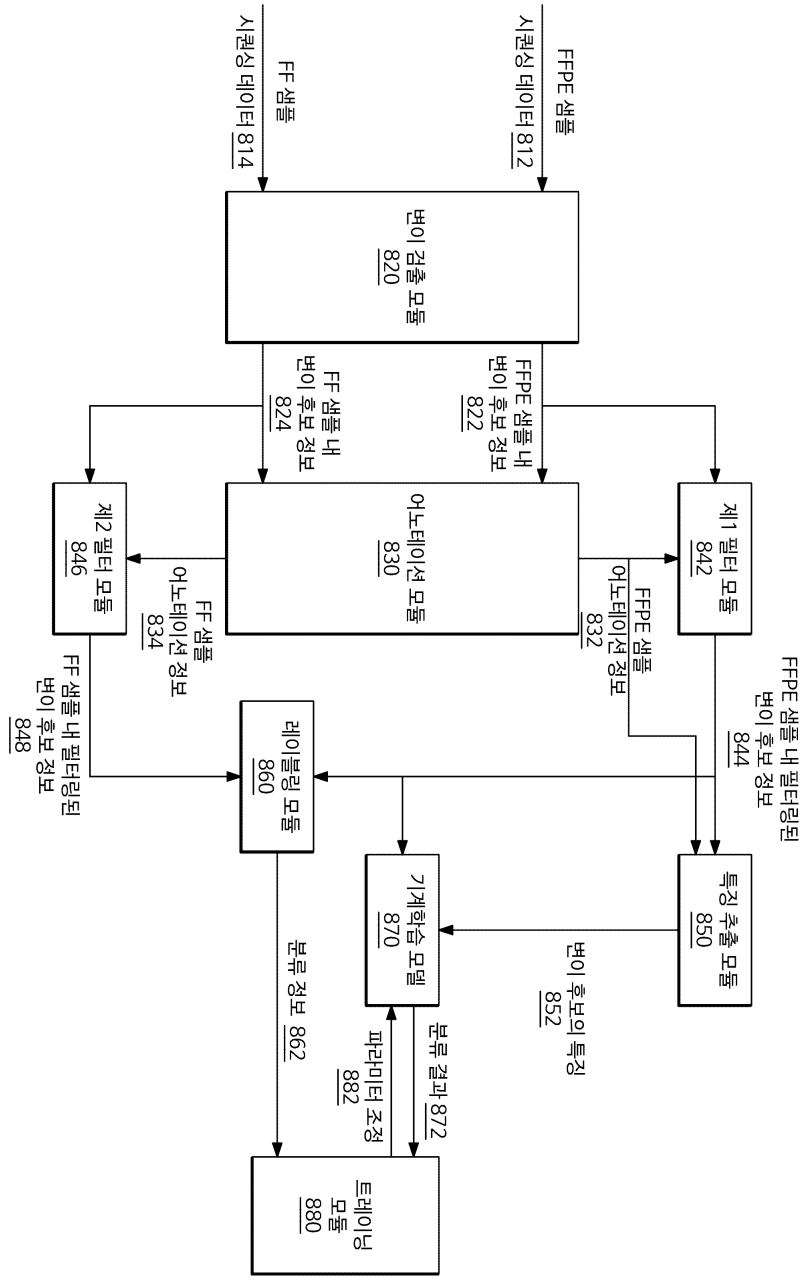
도면6


변이 후보 정보
610
변이 후보의 특징
620
기계학습 모델
630
분류 결과
640

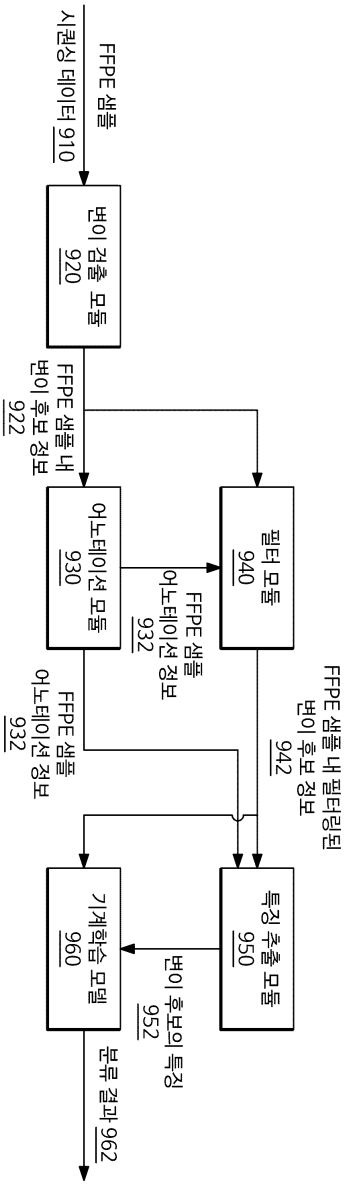
도면7


기계학습 모델 700
제1 분류기 710_1
제2 분류기 710_2
제n 분류기 710_n
제1 출력 결과 720_1
제2 출력 결과 720_2
제n 출력 결과 720_n
메타 분류기 730
분류 결과 740

도면8


시퀀싱 데이터 814
FF 샘플
시퀀싱 데이터 812
FFPE 샘플
변이 검출 모듈 820
FF 샘플 내 변이 후보 정보 824
FFPE 샘플 내 변이 후보 정보 822
제2 필터 모듈 846
어노테이션 모듈 830
제1 필터 모듈 842
FF 샘플 어노테이션 정보 834
FFPE 샘플 어노테이션 정보 832
FF 샘플 내 필터링된 변이 후보 정보 848
레이블링 모듈 860
FFPE 샘플 내 필터링된 변이 후보 정보 844
기계학습 모델 870
특징 추출 모듈 850
분류 정보 862
변이 후보의 특징 852
파라미터 조정 882
분류 결과 872
트레이닝 모듈 880

도면9


시퀀싱 데이터 910
FFPE 샘플
변이 검출 모듈 920
FFPE 샘플 내 변이 후보 정보 922
어노테이션 모듈 930
FFPE 샘플 어노테이션 정보 932
필터 모듈 940
FFPE 샘플 어노테이션 정보 932
FFPE 샘플 내 필터링된 변이 후보 정보 942
기계학습 모델 960
변이 후보의 특징 952
특징 추출 모듈 950
분류 결과 962

도면10

	CHROM	POS	REF	ALT	vaf	label	FILTER_clustered_events	FILTER_fragment	FILTER_germline	FILTER_haplotype	FILTER_multiallelic
1:	chr1	51898	C	A	0.13333333	FALSE	0	0	0	0	0
2:	chr1	773081	G	T	0.15555556	FALSE	0	0	0	0	0
3:	chr1	822235	A	G	0.35000000	FALSE	0	0	0	0	0
4:	chr1	891096	T	A	0.06666667	FALSE	0	0	0	0	0
5:	chr1	896386	C	A	0.20000000	FALSE	0	0	0	0	0

13629:	chrY	56821709	G	A	0.10909091	FALSE	0	0	0	0	0
13630:	chrY	56852302	T	C	0.11864407	FALSE	0	0	0	0	0
13631:	chrM	3481	G	A	0.21287728	FALSE	0	0	0	0	0
13632:	chrM	8115	G	A	0.40480000	FALSE	0	0	0	0	0
13633:	chrM	16292	C	T	0.28990000	FALSE	0	0	0	0	0

	FILTER_normal_artifact	FILTER_PASS	FILTER_position	FILTER_strand_bias	FILTER_weak_evidence	PON_varsum	PON_varN	PON_vafpct	PON_var0.2hN
1:	0	1	0	0	0	1379	143	5.30	43
2:	0	1	0	0	0	1315	371	10.28	67
3:	0	1	0	0	0	37	37	0.22	0
4:	0	1	0	0	0	15	15	0.08	0
5:	0	1	0	0	0	397	148	2.05	0

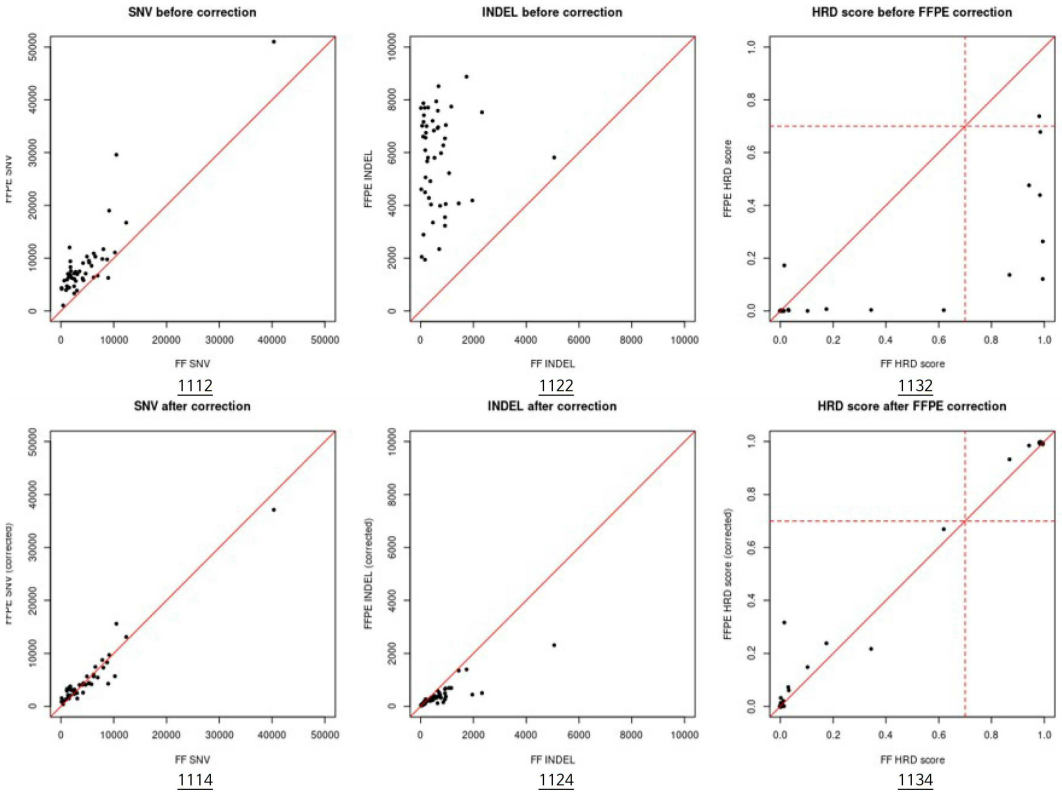
13629:	0	1	0	0	0	6	6	0.01	0
13630:	0	1	0	0	0	5	5	0.01	0
13631:	0	1	0	0	0	116	99	0.01	0
13632:	0	1	0	0	0	129	109	0.01	0
13633:	0	1	0	0	0	12157	184	0.89	3

	PON_var0.2lN	PON_var2N	POF_varsum	POF_varN	POF_vafpct	POF_var0.2hN	POF_var0.2lN	POF_var2N	ref_readN	var_readN	ref_minMQ	ref_medMQ	ref_maxMQ
1:	99	38	1158	65	8.92	24	41	21	78	12	13	60	60
2:	250	130	349	111	5.88	5	87	50	38	7	0	22	40
3:	37	37	33	12	0.45	1	11	11	39	21	0	60	60
4:	15	15	8	6	0.16	0	6	5	42	3	60	60	60
5:	148	85	74	40	1.87	2	37	31	20	5	0	53	60

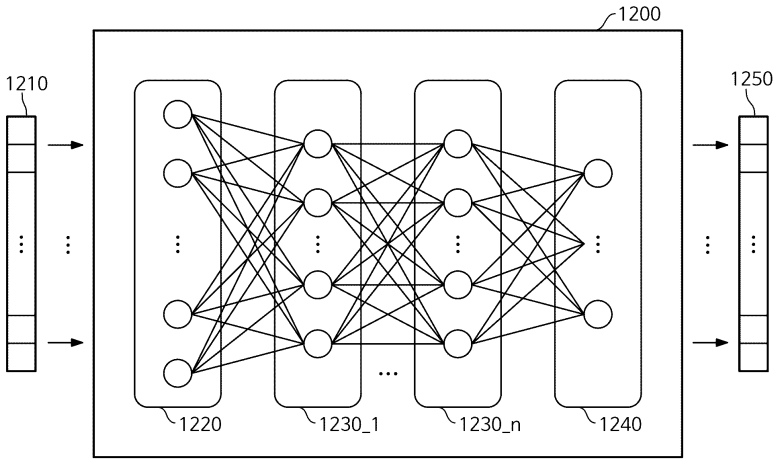
13629:	6	6	12	3	0.12	1	2	1	49	6	20	57	60
13630:	5	5	17	9	0.09	0	9	8	52	7	0	40	60
13631:	99	96	4181	106	0.32	1	105	63	7876	2130	18	60	60
13632:	109	106	4434	95	0.46	1	94	69	5952	4048	0	60	60
13633:	181	158	12874	127	1.01	2	125	61	7101	2899	0	60	60

1000

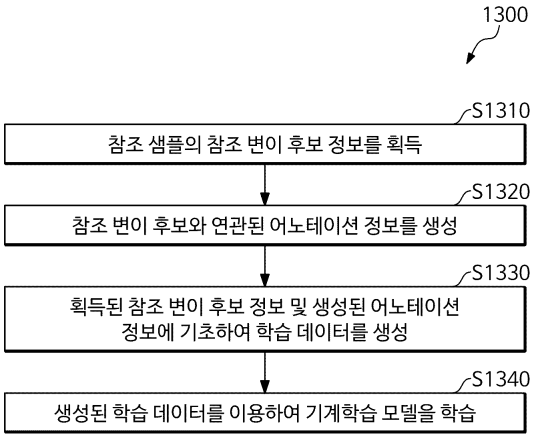
도면11


SNV before correction
FFPE SNV
FF SNV
1112
INDEL before correction
FFPE INDEL
FF INDEL
1122
HRD score before FFPE correction
FFPE HRD score
FF HRD score
1132
SNV after correction
FFPE SNV (corrected)
FF SNV
1114
INDEL after correction
FFPE INDEL (corrected)
FF INDEL
1124
HRD score after FFPE correction
FFPE HRD score (corrected)
FF HRD score
1134

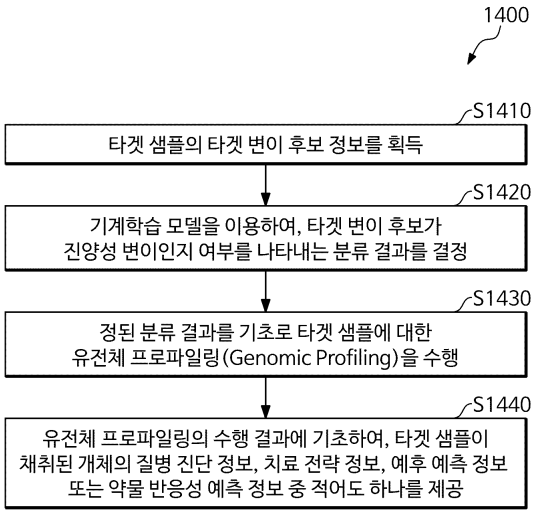
도면12


1200
1210
1250
1220
1230_1
1230_n
1240

도면13


1300
S1310
참조 샘플의 참조 변이 후보 정보를 획득
S1320
참조 변이 후보와 연관된 어노테이션 정보를 생성
S1330
획득된 참조 변이 후보 정보 및 생성된 어노테이션 정보에 기초하여 학습 데이터를 생성
S1340
생성된 학습 데이터를 이용하여 기계학습 모델을 학습

도면14


1400
S1410
타겟 샘플의 타겟 변이 후보 정보를 획득
S1420
기계학습 모델을 이용하여, 타겟 변이 후보가 진양성 변이인지 여부를 나타내는 분류 결과를 결정
S1430
정된 분류 결과를 기초로 타겟 샘플에 대한 유전체 프로파일링(Genomic Profiling)을 수행
S1440
유전체 프로파일링의 수행 결과에 기초하여, 타겟 샘플이 채취된 개체의 질병 진단 정보, 치료 전략 정보, 예후 예측 정보 또는 약물 반응성 예측 정보 중 적어도 하나를 제공