

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0136765

(43) 공개일자 2019년12월10일

(51) 국제특허분류(Int. Cl.)

G16B 30/00 (2019.01)

(52) CPC특허분류

G16B 30/00 (2019.02)

(21) 출원번호 10-2018-0062896

(22) 출원일자 2018년05월31일

심사청구일자 2018년05월31일

(71) 출원인

주식회사 테라젠이텍스

경기도 안산시 단원구 산단로68번길 58 (초지동)

(72) 발명자

홍성의

경기도 수원시 팔달구 권광로 373, 103동 1706호 (우만동, 월드메르디앙 아파트)

김해숙

서울특별시 송파구 올림픽로 435, 213동 302호(신천동, 파크리오)

(뒷면에 계속)

(74) 대리인

리엔목특허법인

전체 청구항 수 : 총 9 항

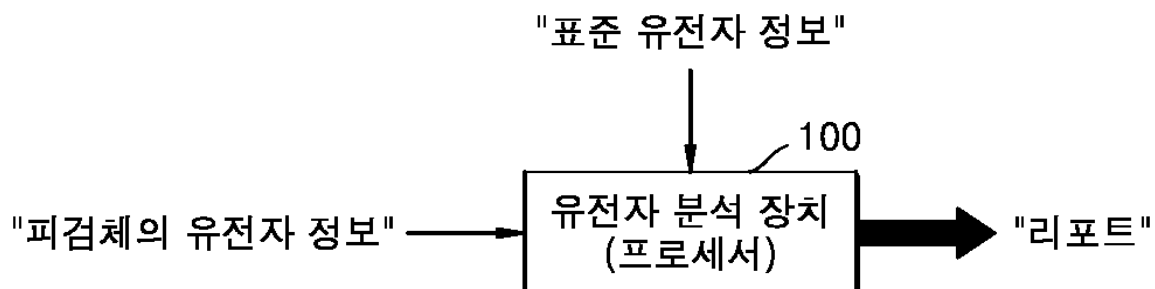
(54) 발명의 명칭 유전자 분석 장치 및 이를 이용한 유전자 분석 방법

(57) 요약

본 명세서는 서버 장치가 사용자로부터 수신된 분석 요청에 의해 상기 분석 요청에 의해 분석되어야 하는 제1 그룹에 포함된 하나 이상의 피검체 중에서, 제1 피검체의 유전자 정보에 포함되는 적어도 하나의 단서열에 대한 품질 점수(quality score)를 산출하고, 기준 품질 점수 미만인 상기 단서열을 제거하는 하나 이상의 전처리 모듈

(뒷면에 계속)

대표도 - 도1



중에서, 상기 제1 그룹의 분석 대상과 대응되는 제1 전처리 모듈을 선택하는 단계; 상기 제1 전처리 모듈에 의해 전처리된 상기 단서열 및 인간 게놈 표준 서열을 비교하여 차이를 산출한 후 선택 기준을 정해 신뢰할 수 있는 변이 정보를 추출하는 하나 이상의 추출 모듈 중에서, 상기 제1 그룹의 분석 대상과 대응되는 제1 추출 모듈을 선택하는 단계; 및 상기 변이 정보 및 상기 인간 게놈 표준 서열에 대한 유사도를 비교하고 상기 변이 정보에 대한 위치정보 및 유전자 정보를 포함하는 하나 이상의 주석달기(annotation)를 수행하고 해당 유전자에 의해 수행하는 기능을 해석하는 하나 이상의 해석 모듈 중에서, 상기 제1 그룹의 분석 대상과 대응되는 제1 해석 모듈을 선택하는 단계; 상기 서버 장치가 상기 제1 전처리 모듈, 상기 제1 추출 모듈, 상기 제1 해석 모듈을 일체의 제1 과정으로 연결한 제1 가상의 영역을 저장 매체에 할당하는 단계; 상기 서버 장치가 상기 일체의 제1 과정을 상기 제1 그룹에 포함되는 하나 이상의 피검체 각각에 대해서 반복적으로 수행하는 단계를 더 포함하는, 유전자 분석 방법을 개시한다.

(72) 발명자

김세훈

경기도 수원시 팔달구 중부대로 96- 19, 101동 70
2호(인계동, 회성 연인 주상복합)

송상훈

경기도 화성시 동탄순환대로 881- 10, 705동 2501
호 (영천동, 동탄역 푸르지오)

황태순

경기도 용인시 수지구 성북2로 251, 207동 402호(
성북동, 버들치마을 성북자이2차)

고진업

경기도 성남시 수정구 대왕판교로 801, 103동 203
호(시흥동, KOLON LINDEN GROVE)

명세서

청구범위

청구항 1

서버 장치가 사용자로부터 수신된 분석 요청에 의해 상기 분석 요청에 의해 분석되어야 하는 제1 그룹에 포함된 하나 이상의 피검체 중에서, 제1 피검체의 유전자 정보에 포함되는 적어도 하나의 단서열에 대한 품질 점수(quality score)를 산출하고, 기준 품질 점수 미만인 상기 단서열을 제거하는 하나 이상의 전처리 모듈 중에서, 상기 제1 그룹의 분석 대상과 대응되는 제1 전처리 모듈을 선택하는 단계; 상기 제1 전처리 모듈에 의해 전처리된 상기 단서열 및 인간 게놈 표준 서열을 비교하여 차이를 산출한 후 선택 기준을 정해 신뢰할 수 있는 변이 정보를 추출하는 하나 이상의 추출 모듈 중에서, 상기 제1 그룹의 분석 대상과 대응되는 제1 추출 모듈을 선택하는 단계; 및 상기 변이 정보 및 상기 인간 게놈 표준 서열에 대한 유사도를 비교하고 상기 변이 정보에 대한 위치정보 및 유전자 정보를 포함하는 하나 이상의 주석달기(annotation)를 수행하고 해당 유전자에 의해 수행하는 기능을 해석하는 하나 이상의 해석 모듈 중에서, 상기 제1 그룹의 분석 대상과 대응되는 제1 해석 모듈을 선택하는 단계;

상기 서버 장치가 상기 제1 전처리 모듈, 상기 제1 추출 모듈, 상기 제1 해석 모듈을 일체의 제1 과정으로 연결한 제1 가상의 영역을 저장 매체에 할당하는 단계;

상기 서버 장치가 상기 일체의 제1 과정을 상기 제1 그룹에 포함되는 하나 이상의 피검체 각각에 대해서 반복적으로 수행하는 단계를 더 포함하는,

유전자 분석 방법.

청구항 2

제 1항에 있어서,

상기 서버 장치가 상기 제1 그룹과 다른 제2 그룹에 대한 분석 요청이 수신되면, 상기 제2 그룹의 분석 대상과 대응되는 제2 전처리 모듈을 선택하고, 상기 제2 그룹의 분석 대상과 대응되는 제2 추출 모듈, 상기 제2 그룹의 분석 대상과 대응되는 제2 해석 모듈을 일체의 제2 과정으로 연결한 제2 가상의 영역을 저장 매체에 할당하고, 상기 제1 그룹에 포함되는 하나 이상의 피검체 각각에 대해서 반복적으로 수행하는 과정과 병렬적으로, 상기 제2 그룹에 포함되는 하나 이상의 피검체 각각에 대해서 반복적으로 상기 제2 과정을 수행하는 단계;를 더 포함하는, 유전 분석 방법.

청구항 3

제 1항에 있어서,

상기 해석 모듈을 선택하는 단계 이후에, 처리 결과를 고객이 알아볼 수 있는 형식의 리포트로 생성하는 하나 이상의 생성 모듈 중에서, 상기 제1 그룹의 분석 대상과 대응되는 제1 생성 모듈을 선택하는 단계;를 더 포함하고,

상기 제1 과정은 상기 제1 생성 모듈을 더 포함하도록 변경되는 점을 특징으로 하는 유전자 분석 방법.

청구항 4

제 1항에 있어서,

상기 제1 전처리 모듈은

상기 단서열에 대한 품질 점수가 상기 기준 품질 점수 이상 나올 때까지 상기 단서열의 제거 및 상기 품질 점수 산출을 반복하는 것을 특징으로 하는 유전자 분석 방법.

청구항 5

컴퓨터를 이용하여 제 1항 내지 제 4항의 방법 중 어느 한 항의 방법을 실행시키기 위하여 상기 컴퓨터로 판독 가능한 기록매체에 저장된 컴퓨터 프로그램.

청구항 6

제1 그룹에 속하는 피검체의 유전자 정보에 포함되는 적어도 하나의 단서열에 대한 품질 점수(quality score)를 산출하고, 기준 품질 점수 미만인 상기 단서열을 제거하는 하나 이상의 전처리 모듈을 포함하고, 상기 피검체의 분석 대상, 분석 그룹, 클라이언트의 성향 중 적어도 하나를 고려하여 상기 하나 이상의 전처리 모듈 중에서, 상기 피검체를 처리할 제1 전처리 모듈을 선택하는 전처리부;

상기 전처리 모듈에 의해 처리된 거친 상기 단서열 및 인간 게놈 표준 서열을 비교하여 차이를 산출한 후 선택 기준을 정해 신뢰할 수 있는 변이 정보를 추출하는 하나 이상의 추출 모듈을 포함하고, 상기 피검체의 분석 대상, 분석 그룹, 클라이언트의 성향 중 적어도 하나를 고려하여 상기 하나 이상의 추출 모듈 중에서, 상기 피검체의 변이 정보를 추출할 제1 추출 모듈을 선택하는 추출부;

상기 변이 정보 및 상기 인간 게놈 표준 서열에 대한 유사도를 비교하고 상기 변이 정보에 대한 위치정보 및 유전자 정보를 포함하는 하나 이상의 주석달기(annotation)를 수행하고 어떤 기능을 수행하는지 해석하는 하나 이상의 해석 모듈을 포함하고, 상기 분석 대상, 상기 분석 그룹, 상기 클라이언트의 성향 중 적어도 하나를 고려하여 상기 하나 이상의 해석 모듈 중에서, 임의의 하나의 제1 해석 모듈을 선택하는 해석부;

상기 제1 전처리 모듈, 상기 제1 추출 모듈, 상기 제1 해석 모듈을 일체의 제1 과정으로 연결한 제1 가상의 영역을 저장 매체에 할당하는 일체 처리부;를 포함하고,

상기 일체 처리부는

상기 일체의 제1 과정을 상기 피검체 및 상기 제1 그룹에 분류되는 하나 이상의 피검체 각각에 대해서 반복적으로 수행하는 점을 특징으로 하는, 유전자 분석 장치.

청구항 7

제 6항에 있어서, 상기 제1 그룹과 다른 제2 그룹에 대한 분석 요청이 수신되면, 상기 제2 그룹의 분석 대상과 대응되는 제2 전처리 모듈, 제2 추출 모듈, 및 제2 해석 모듈을 각각 선택하도록 제어하고, 상기 제2 전처리 모듈, 상기 제2 추출 모듈, 상기 제2 해석 모듈을 일체의 제2 과정으로 연결한 제2 가상의 영역을 저장 매체에 할당하고,

상기 제1 그룹에 포함되는 하나 이상의 피검체 각각에 대해서 반복적으로 수행하는 과정과 병렬적으로 상기 제2 그룹에 포함되는 하나 이상의 피검체 각각에 대해서 반복적으로 상기 제2 과정을 수행하도록 제어하는 병렬 처리부를 더 포함하는 점을 특징으로 하는, 유전자 분석 장치.

청구항 8

제7항에 있어서,

처리 결과를 고객이 알아볼 수 있는 형식의 리포트로 생성하는 하나 이상의 생성 모듈 중에서, 상기 제1 그룹과 대응되는 제1 생성 모듈을 선택하는 생성부;를 더 포함하고,

상기 제1 과정은 상기 일체 처리부에 의해서 상기 제1 생성 모듈을 더 포함하도록 변경되는 점을 특징으로 하는, 유전자 분석 장치.

청구항 9

제 7항에 있어서,

상기 제1 전처리 모듈은

상기 단서열에 대한 품질 점수가 상기 기준 품질 점수 이상 나올 때까지 상기 단서열의 제거 및 상기 품질 점수 산출을 반복하여 수행되도록 제어되는 점을 특징으로 하는, 유전자 분석 장치.

발명의 설명

기술 분야

[0001] 본 발명은 유전자 분석 장치 및 이를 이용한 유전자 분석방법에 관한 것이다.

배경 기술

[0002] 유전체(genome)란 한 생물이 가지는 모든 유전 정보를 말한다. 어느 한 개인의 유전체를 서열화(sequencing)하는 기술로 DNA 칩 및 차세대 시퀀싱(Next Generation Sequencing) 기술, 차차세대 시퀀싱(Next Next Generation Sequencing) 기술 등 여러 기술들이 개발되고 있다. 차세대 시퀀싱은 대규모 병렬 시퀀싱 또는 2세대 시퀀싱과 상호 교환적으로 사용될 수 있다.

[0003] 차세대 시퀀싱 기계에서 생성된 유전체 데이터는 적게는 수십 많게는 수백 기가 바이트에 이른다. 이러한 유전체 데이터의 분석에는 서로 다른 기능을 수행하는 알고리즘을 구현한 툴들의 조합을 사용한다. 유전체 데이터 분석에서 각 툴들은 각 단계별로 사용되며, 각 툴들은 서로 다른 그룹에 의해 개별적으로 개발되었다. 따라서 이들 개별 툴들을 통합하고 입출력을 관리하며 유전체 데이터 분석 전 단계를 자동화 한 소프트웨어인 유전체 분석 파이프라인이 개발되고 있다.

[0004] 전술한 배경기술은 발명자가 본 발명의 도출을 위해 보유하고 있었거나, 본 발명의 도출 과정에서 습득한 기술 정보로서, 반드시 본 발명의 출원 전에 일반 공중에게 공개된 공지기술이라 할 수는 없다.

[0005] [국내 문헌] 국내 공개특허공보 제2017-0064541호

발명의 내용

해결하려는 과제

[0006] 본 발명의 해결하고자 하는 기술적 과제는 기존에 찾기 어려웠던 유전자의 변이를 효율적으로 찾아내는 것이다.

[0007] 또한 유전자 분석에 필요한 부하 및 시간을 단축시키는 것이다.

과제의 해결 수단

[0009] 본 발명의 일 실시 예에 따른 유전자 분석 방법은, 피검체의 유전자 정보에 포함되는 적어도 하나의 단서열에 대한 품질 점수(quality score)를 산출하고, 기준 품질 점수 미만인 상기 단서열을 제거하는 전처리 단계; 상기 전처리 단계를 거친 상기 단서열 및 인간 게놈 표준 서열을 비교하여 차이를 산출한 후 선택 기준을 정해 신뢰할 수 있는 변이 정보를 추출하는 추출단계; 및 상기 변이 정보 및 상기 인간 게놈 표준 서열에 대한 유사도를 비교하고 상기 변이 정보에 대한 위치정보 및 유전자 정보를 포함하는 하나 이상의 주석달기(annotation)를 수행하고 어떤 기능을 수행하는지 해석하는 해석단계;를 포함할 수 있다.

[0010] 상기 방법은, 상기 추출단계 및 상기 해석단계의 처리 결과를 고객이 알아볼 수 있는 형식의 리포트로 생성하는 생성단계;를 더 포함할 수 있다.

[0011] 상기 전처리 단계는, 상기 단서열에 대한 품질 점수가 상기 기준 품질 점수 이상 나올 때까지 상기 단서열의 제거 및 상기 품질 점수 산출을 반복하는 단계;를 포함할 수 있다.

[0012] 상기 추출단계는, 상기 전처리 단계를 거친 상기 단서열 및 상기 인간 게놈 표준 서열을 비교하여 상기 단서열을 상기 인간 게놈 표준 서열대로 정렬하는 맵핑 단계; 상기 맵핑단계를 통하여 정렬한 상기 단서열로부터 중합효소 연쇄 반응(PCR, polymerase chain reaction)에 의한 증폭 시에 발생하는 중복 단서열을 제거하는 제거 단계; 삽입(insert) 또는 결실(indel)로 인하여 맵핑이 이루어 지지 않은 상기 제거 단계를 거친 상기 단서열에 대하여 지역(local) 재배열을 수행하는 재배열 단계; 단서열의 성질과 관련하여 지정된 소정의 변수(covariate)를 바탕으로 재보정 테이블을 생성하고, 염기 당 추측되는 오류에 대한 점수로써의 염기(base)의 품질 점수를 재보정하는 재보정 단계; 및 상기 재보정 단계를 거친 상기 단서열을 상기 인간 게놈 표준 서열과 비교하여 어느 위치에 어느 정도 맵핑되었는지 여부로 변이 정보를 발굴하는 발굴 단계;를 포함할 수 있다.

[0013] 본 발명의 일 실시 예에 따른 유전자 분석 장치는, 피검체의 유전자 정보에 포함되는 적어도 하나의 단서열에

대한 품질 점수(quality score)를 산출하고, 기준 품질 점수 미만인 상기 단서열을 제거하는 전처리부; 상기 전처리 단계를 거친 상기 단서열 및 인간 게놈 표준 서열을 비교하여 차이를 산출한 후 선택 기준을 정해 신뢰할 수 있는 변이 정보를 추출하는 추출부; 및 상기 변이 정보 및 상기 인간 게놈 표준 서열에 대한 유사도를 비교하고 상기 변이 정보에 대한 위치정보 및 유전자 정보를 포함하는 하나 이상의 주석달기(annotation)를 수행하고 어떤 기능을 수행하는지 해석하는 해석부;를 포함할 수 있다.

[0014] 상기 장치는, 상기 추출부 및 상기 해석부의 처리 결과를 고객이 알아볼 수 있는 형식의 리포트로 생성하는 생성부;를 더 포함할 수 있다.

[0015] 상기 전처리부는, 상기 단서열에 대한 품질 점수가 상기 기준 품질 점수 이상 나올 때까지 상기 단서열의 제거 및 상기 품질 점수 산출을 반복할 수 있다.

[0016] 상기 추출부는, 상기 전처리 단계를 거친 상기 단서열 및 상기 인간 게놈 표준 서열을 비교하여 상기 단서열을 상기 인간 게놈 표준 서열대로 정렬하고, 정렬한 상기 단서열로부터 중합효소 연쇄 반응(PCR, polymerase chain reaction)에 의한 증폭 시에 발생하는 중복 단서열을 제거하고, 삽입 또는 결실(indel)로 인하여 맵핑이 이루어 지지 않은 상기 제거를 거친 상기 단서열에 대하여 지역(local) 재배열을 수행하고, 단서열의 성질과 관련하여 지정된 소정의 변수(covariate)를 바탕으로 재보정 테이블을 생성하고, 염기 당 추측되는 오류에 대한 점수로써의 염기(base)의 품질 점수를 재보정하며, 상기 재보정을 거친 상기 단서열을 상기 인간 게놈 표준 서열과 비교하여 어느 위치에 어느 정도 맵핑되었는지 여부로 변이 정보를 발굴할 수 있다.

[0017] 이 외에도, 본 발명을 구현하기 위한 다른 방법, 다른 시스템 및 상기 방법을 실행하기 위한 컴퓨터 프로그램이 더 제공될 수 있다.

[0018] 전술한 것 외의 다른 측면, 특징, 이점이 이하의 도면, 특허청구범위 및 발명의 상세한 설명으로부터 명확해질 것이다.

발명의 효과

[0019] 실시 예들에 따르면, 기존에 찾기 어려웠던 유전자의 변이를 효율적으로 찾아낼 수 있다.

[0020] 또한 유전자 분석에 필요한 부하 및 시간을 단축시킬 수 있다.

[0021] 또한 저렴한 비용으로 대용량의 유전자 분석 정보를 획득할 수 있다.

[0022] 본 발명의 효과는 이상에서 언급된 것들에 한정되지 않으며, 언급되지 아니한 다른 효과들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0023] 본 개시의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 간단한 설명이 제공된다.

도 1은 본 발명의 일 실시 예에 따른 유전자 분석 장치를 개략적으로 설명하기 위하여 도시한 도면이다.

도 2는 도 1의 유전자 분석 장치의 상세 구성을 개략적으로 설명하기 위하여 도시한 도면이다.

도 3 및 도 5는 본 발명의 일 실시 예에 따른 유전자 분석 장치의 동작 방법을 나타내는 흐름도이다.

도 5는 본 발명의 다른 실시 예에 따른 유전자 분석 장치의 동작 방법을 나타내는 흐름도이다.

도 6 내지 도 7은 본 발명의 또 다른 실시 예에 따른 유전자 분석 장치의 동작 방법을 나타내는 흐름도이다.

도 8은 본 발명의 일 실시 예에 따른 컴퓨팅 장치의 하드웨어 구성을 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0024] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 설명되는 실시 예를 참조하면 명확해질 것이다. 그러나 본 발명은 아래에서 제시되는 실시 예들로 한정되는 것이 아니라, 서로 다른 다양한 형태로 구현될 수 있고, 본 발명의 사상 및 기술 범위에 포함되는 모든 변환, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 아래에 제시되는 실시 예들은 본 발명의 개시가 완전하도록 하며, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이다. 본 발명을 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되

는 경우 그 상세한 설명을 생략한다.

- [0025] 본 출원에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다. 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다.
- [0026] 명세서 전체에서 “단서열(read, read)” 이라 함은, 염기 서열 분석(시퀀싱, sequencing) 라이브러리에 포함된 DNA(deoxyribonucleic acid) 또는 cDNA(complementary DNA) 단편에서 생성한 분석량에 대한 염기쌍 정보를 의미하는 것으로, 염기 서열 분석으로 나온 출력 데이터, 시퀀스(염기 서열)의 조각을 포함할 수 있다. 이러한 단서열들은 약 10bp 내지 약 2000bp, 약 15bp 내지 약 1500bp, 약 20bp 내지 약 1000bp, 약 20bp 내지 약 500bp 또는 약 20bp 내지 약 200bp 일 수 있다.
- [0027] 명세서 전체에서 “이중 말단 단서열(paired end read)” 이라 함은, DNA 염기 서열이 삽입(insert)의 이중 가닥 모두의 5' 말단에서 획득되는데, 해당 단서열은 일반적으로 2x"y"로 나타나며, "y"는 염기쌍에서 한 번에 읽을 수 있는 염기 서열의 분석량(길이)(예를 들어, 2x100bp, 2x150bp)를 포함할 수 있다.
- [0028] 명세서 전체에서 “맵핑(mapping)” 이라 함은, 피검체(개인, 환자)의 염기 서열 데이터(sequence read)를 표준 염기 서열(reference genome)과 비교하는 작업을 포함할 수 있다.
- [0029] 명세서 전체에서 “베이트(baits)” 라 함은, 표적 농축(target enrichment)을 수행하는데 있어 관심 부위의 식별 및 결합을 담당하는 올리고 뉴클레오티드 서열(즉, 프로브)의 일반적인 명칭을 포함할 수 있다.
- [0030] 명세서 전체에서 “fastq 파일” 이라 함은, 분석을 하는 기초 데이터이며, 서열 ID, 염기 서열 정보, 염기 서열 품질(quality) 등을 포함할 수 있다.
- [0031] 명세서 전체에서 “bam 파일(binary alignment/map format file)” 이라 함은, fastq 파일의 각 서열들을 표준 염기 서열에 맵핑된 결과 파일로써, SAM 파일(sequence alignment/map format file)의 바이너리(이진법) 버전으로, 데이터 분석의 두 번째 단계에서 얻어지는 결과일 수 있다.
- [0032] 명세서 전체에서 “vcf 파일(variant call format file)” 이라 함은, 변이 정보를 담은 파일을 포함할 수 있다.
- [0033] 명세서 전체에서 “annotation 파일” 이라 함은, vcf 파일에 포함되어 있는 각 변이들이 어떤 유전자(gene)에 해당되는지 정리되어 있는 파일을 포함할 수 있다.
- [0034] 명세서 전체에서 “리포트” 라 함은, 샘플 정보 및 각 변이들의 정보들이 담겨있는 최종 파일을 포함할 수 있다.
- [0036] 이하, 본 발명에 따른 실시 예들을 첨부된 도면을 참조하여 상세히 설명하기로 하며, 첨부 도면을 참조하여 설명함에 있어, 동일하거나 대응하는 구성 요소는 동일한 도면번호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다.
- [0037] 도 1은 본 발명의 일 실시 예에 따른 유전자 분석 장치를 개략적으로 설명하기 위하여 도시한 도면이고, 도 2는 도 1의 유전자 분석 장치의 상세 구성을 개략적으로 설명하기 위하여 도시한 도면이다. 도 1 및 도 2를 참조하면, 유전자 분석 장치(100)는 전처리부(110), 추출부(120), 해석부(130), 생성부(140), 일체 처리부(150), 및 병렬 처리부(160)를 포함할 수 있다.
- [0038] 유전자 분석 장치(100)는 피검체로부터 획득한 유전자 정보 및 정상인 집단으로부터 획득한 표준 유전자 정보를 이용하여 유전자의 서열을 분석하고, 유전자의 서열 분석을 통해 피검체의 피검 유전자에 단일 뉴클레오티드 변이(single nucleotide variant, SNV) 또는 복제수 변이(copy number variant, CNV) 등 유전체의 변이 유전자가 존재하는지 식별할 수 있다. 여기서, 단일 뉴클레오티드 변이(single nucleotide variant, SNV)”는 상대적으로 큰 영역이 결손되거나 증폭되어 반복적으로 나타나는 유전자 내의 복제수 변이(copy number variant, CNV)와는 다르게 단일 뉴클레오티드의 치환 등을 의미할 수 있다.

- [0039] 유전자 분석 장치(100)가 수신하는 유전자 정보는, 차세대 시퀀싱(next generation sequencing, NGS)에 의해 획득한 fastq 파일의 유전자 정보에 해당될 수 있다. fastq 파일은 보통 뉴클레오티드 서열과 같은 생물학적 서열과, 그에 대응되는 퀄리티 스코어를 저장하는 텍스트 기반 포맷(text-based format)일 수 있다. 다만, 본 실시 예에 따른 유전자 분석 장치(100)는, fastq 파일에 제한되지 않고, 다른 파일의 유전자 정보도 분석이 가능하다. fastq 파일을 입력으로 하여 유전자 분석 장치(100)는 bam 파일, SAM 파일, vcf 파일, annotation 파일 및 리포트를 생성할 수 있다.
- [0040] 정상인 집단으로부터 획득한 표준 유전자 정보는, NCBI(National Center for Biotechnology Information), Gene Expression Omnibus (GEO), FDA(Food and Drug Administration), My Cancer Genome, 또는 KFDA(식품의약품 안전처) 등과 같은 당해 기술분야에서 이미 공지된 데이터베이스(DB)로부터 획득되거나, 또는 피검체의 피검 유전자들을 분석하기 위하여 모집된 사람들의 생물학적 샘플로부터 획득한 것일 수 있다. 즉, 정상인 집단으로부터 획득한 표준 유전자 정보는 공개 게놈 데이터 또는 공개 합맵(HapMap) 데이터로부터 획득한 것일 수 있다. 한편, 정상인 집단으로부터 획득한 표준 유전자 정보에 포함된 레퍼런스 유전자들 또는 피검체로부터 획득한 유전자 정보에 포함된 피검 유전자들은, 생검 조직, 포르말린-고정 조직 또는 파라핀-내장(Formalin-fixed, paraffin-embedded) 조직으로부터 획득한 것일 수 있다.
- [0041] 유전자 분석 장치(100)가 피검체의 변이 정보를 생성하기 위해 크게 피검체의 단서열의 필터링 과정, 단서열의 변이체 발굴 과정 및 변이체 해석 과정의 세 가지 과정이 필요하며, 이에 최종 변이정보 확인을 위한 리포트 생성 과정을 더 포함할 수 있다.
- [0042] 전처리부(110)는 상술한 피검체의 단서열 필터링 과정으로써, 피검체로부터 획득한 유전자 정보 즉, 피검체로부터 획득한 하나 이상의 단서열로부터 정확한 변이 정보를 얻기 위해, 단서열 중 어댑터(adapter) 서열이 포함된 부분과 품질(quality)이 떨어지는 염기(base)를 제거할 수 있다. 이를 위해 cutadapt 프로그램을 사용할 수 있다. 이 프로그램은 각 단서열이 포함된 어댑터 서열을 확인하여 제거한 후, 서열의 평균 품질 점수(quality score)를 계산하여 이 값이 사용자가 선택한 품질 기준값 미만(예를 들어, 품질 점수<20)이면, 단서열을 끝에서부터 시작하여 하나씩 트리밍(자르기, trimming)하고 나머지 염기로 다시 품질 점수를 계산하여 평균값이 기준 품질 점수 이상이 나올 때까지 이 작업을 반복할 수 있다. 트리밍은 사용자의 선택에 따라 3' 말단이나 5' 말단으로부터 시작할 수 있다. 최종적으로 얻어진 단서열의 길이가 사용자가 선택한 기준값(예를 들어, 50bp) 보다 작을 경우, 해당 단서열을 제거하고 그렇지 않을 경우 해당 단서열의 필터링을 종료할 수 있다. 전처리부(110)는 상술한 피검체의 단서열 필터링 과정으로써, 피검체로부터 획득한 유전자 정보 즉, 피검체로부터 획득한 하나 이상의 단서열로부터 정확한 변이 정보를 얻기 위해, 단서열 중 어댑터(adapter) 서열이 포함된 부분과 품질(quality)이 떨어지는 염기(base)를 제거하는 하나 이상의 전처리 모듈을 포함할 수 있다. 전처리부(110)는 분석 요청에 응답하여, 분석 대상, 분석 그룹, 클라이언트의 성향 등을 고려하여 하나 이상의 전처리 모듈 중에서, 분석 요청과 대응되는 전처리 모듈을 선택할 수 있다.
- [0043] 추출부(120)는 상술한 단서열의 변이체 발굴 과정으로써, 전처리(필터링) 과정을 거친 단서열 및 인간 게놈 표준 서열을 비교하여 차이를 산출한 후 선택 기준을 정해 신뢰할 수 있는 변이 정보를 추출하기 위해 맵핑, PCR(polymerase chain reaction) duplication 제거, 지역 재배열(local realignment), 염기 재보정(base recalibration), 변이체 발굴 및 변이체 필터링 과정을 수행하는 하나 이상의 추출 모듈을 포함할 수 있다. 추출부(120)는 분석 요청에 응답하여, 분석 대상, 분석 그룹, 클라이언트의 성향 등을 고려하여 하나 이상의 추출 모듈 중에서, 분석 요청과 대응되는 추출 모듈을 선택할 수 있다.
- [0044] 추출부(120)는 피검체의 단서열을 인간 게놈 표준 서열(human reference sequence, UCSC build hg19)에 효과적으로 맵핑하기 위해 BWA(burrows-wheeler aligner)를 이용할 수 있다. 피검체의 단서열 및 인간 게놈 표준 서열의 맵핑 이후의 데이터를 바로 Samtools 프로그램을 이용하여 표준 서열 순서대로 정렬할 수 있다.
- [0045] 추출부(120)는 맵핑을 통하여 정렬한 단서열로부터 중합효소 연쇄 반응(PCR, polymerase chain reaction)에 의한 증폭 시에 발생하는 중복 단서열을 제거할 수 있는데, 정확한 변이체 발굴을 위해 Samtools 0.1.18 버전의 프로그램을 이용하여 PCR에 의한 증폭 시에 발생하는 중복 단서열을 제거할 수 있다.
- [0046] 추출부(120)는 삽입(insert) 또는 결실(indel)로 인하여 맵핑이 이루어지지 않은 상술한 제거 과정을 거친 단서열에 대하여 지역 재배열(local realignment)을 수행할 수 있다. 추출부(120)는 이전 단계에서 생성된 bam 파일로부터 GATK Lite 2.3.9 버전의 프로그램을 이용하여 지역 재배열을 수행할 수 있는데, 단서열의 부분적인 재배열을 통해 전체 단서열 중에 미스매치(mismatch)되는 염기의 수를 최소화 하기 위함이며, 다음과 같은 두

개의 세부 과정을 포함할 수 있다.

- [0047] 먼저 재배열이 필요해 보이는 의심스러운 짧은 인터벌(interval)을 결정(realigner target creator)한다. 다음에 이들 인터벌에 걸쳐 부분적인 재배열(indel realigner)을 수행하며, 결실(indel)로 인하여 미스배열(misalignment)이 발생한 단서열들의 지역 재배열을 수행 할 수 있다
- [0048] 추출부(120)는 GATK Lite 2.3.9 프로그램에서 제공하는 Base Recalibrator 옵션을 이용하여, 단서열의 성질과 관련하여 지정된 소정의 변수(covariate)를 바탕으로 재보정 테이블을 생성하고, 염기 당 추출되는 오류에 대한 점수로써의 염기(base)의 품질 점수(quality score)를 재보정(recalibration) 할 수 있다. 사용자가 지정할 수 있는 변수들로는 단서열 그룹, 기존에 보고된 품질 점수, 머신 사이클(machine cycle) 및 뉴클레오타이드 상황 정보(nucleotide context) 등을 포함할 수 있으며
- [0049] 추출부(120)는 재보정 단계를 거친 단서열을 인간 게놈 표준 서열과 비교하여 어느 위치에 어느 정도 맵핑되었는지 여부로 변이 정보를 발굴할 수 있다. 여기서 추출부(120)는 GATK Lite 2.3.9 프로그램에서 제공하는 UnifiedGenotyper 옵션을 이용하여 변이체를 발굴할 수 있는데, 이 옵션은 Bayesian genotype likelihood model을 이용하여 가장 높은 확률의 변이체 유전형 및 위치를 예측할 수 있다
- [0050] 해석부(130)는 발굴한 변이 정보 및 인간 게놈 표준 서열에 대한 유사도를 비교하고, 발굴한 변이 정보에 대한 위치정보 및 유전자 정보를 포함하는 하나 이상의 주석달기(annotation)를 수행하고 어떤 기능을 수행하는지 해석할 수 있다. 즉, 해석부(130)는 인간 게놈 표준 서열 NCBI build GRCh37 (hg19)의 정보를 근간으로 각 변이의 게놈에서의 위치 정보 및 유전자 정보를 확보하는 주석달기를 수행할 수 있다. 변이체 주석달기를 수행하기 위해서는 보편적으로 이용되는 툴 중 하나인 snpEff을 이용할 수 있다. 사용하는 주석달기 데이터베이스는 clinvar, dbsnp, dbnsfp, cosmic DB 이며, 변이 표기는 HGVS 표준 명명법을 따를 수 있다. 해석부(130)는 상술한 단서열의 변이체 발굴 과정으로써, 전처리(필터링) 과정을 거친 단서열 및 인간 게놈 표준 서열을 비교하여 차이를 산출한 후 선택 기준을 정해 신뢰할 수 있는 변이 정보를 추출하기 위해 맵핑, PCR(polymerase chain reaction) duplication 제거, 지역 재배열(local realignment), 염기 재보정(base recalibration), 변이체 발굴 및 변이체 필터링 과정을 수행하는 하나 이상의 추출 모듈을 포함할 수 있다. 해석부(130)는 분석 요청에 응답하여, 분석 대상, 분석 그룹, 클라이언트의 성향 등을 고려하여 하나 이상의 해석 모듈 중에서, 분석 요청과 대응되는 해석 모듈을 선택할 수 있다.
- [0052] 생성부(140)는 추출 모듈 및 해석 모듈의 처리 결과를 고객이 알아볼 수 있는 형식의 리포트로 생성할 수 있는 하나 이상의 생성 모듈을 포함할 수 있다. 생성부(140)는 분석 요청에 응답하여, 분석 대상, 분석 그룹, 요청한 클라이언트의 성향 등을 고려하여 하나 이상의 생성 모듈 중에서, 분석 요청과 대응되는 생성 모듈을 선택할 수 있다.
- [0053] 일체 처리부(150)는 제1 전처리 모듈, 제1 추출 모듈, 제1 해석 모듈을 일체의 제1 과정으로 연결한 제1 가상의 영역을 저장 매체에 할당한다.
- [0054] 병렬 처리부(160)는 수신된 하나 이상의 분석 요청 각각에 포함된 각 그룹에 포함된 피검체들을 분석하는 각 일체의 과제를 병렬적으로 처리한다.
- [0055] 도 3 및 도 5는 본 발명의 일 실시 예에 따른 유전자 분석 장치의 동작 방법을 나타내는 흐름도이다. 본 발명의 일 실시 예에 따른 유전자 분석 방법은 도 1 및 도 2에 도시된 바와 같이 주변 구성요소들의 도움을 받아 유전자 분석 장치(100)에서 수행될 수 있다. 이하의 설명에서 도 1 및 도 2에 대한 설명과 중복되는 부분은 그 설명을 생략하기로 한다.
- [0056] 도 3을 참조하면, S110 단계에서, 유전자 분석 장치(100)는 사용자로부터 분석 요청을 수신한다. 분석 요청은 분석 대상, 분석되어야 하는 하나 이상의 피검체들, 분석에 추가적으로 요청하고자 하는 요청 사항, 분석 요청을 발신하는 클라이언트 정보 등을 포함할 수 있다.
- [0057] S120에서는, 유전자 분석 장치(100)는 수신된 분석 요청에 의해 분석되어야 하는 제1 그룹에 포함된 하나 이상의 피검체 중에서, 제1 피검체의 유전자 정보에 포함되는 적어도 하나의 단서열에 대한 품질 점수를 산출하고, 기준 품질 점수 미만인 단서열을 제거하는 하나 이상의 전처리 모듈 중에서, 제1 그룹의 분석 대상과 대응되는 제1 전처리 모듈을 선택한다. 제1 전처리 모듈의 동작은 전처리부(110)의 동작과 동일하므로, 중복 설명을 생략한다.

- [0058] S130에서는 유전자 분석 장치(100)는 제1 전처리 모듈에 의해 전처리된 단서열 및 인간 게놈 표준 서열을 비교하여 차이를 산출한 후 선택 기준을 정해 신뢰할 수 있는 변이 정보를 추출하는 하나 이상의 추출 모듈 중에서, 제1 그룹의 분석 대상과 대응되는 제1 추출 모듈을 선택한다.
- [0059] S140에서는 유전자 분석 장치(100)는 상기 변이 정보 및 상기 인간 게놈 표준 서열에 대한 유사도를 비교하고 상기 변이 정보에 대한 위치정보 및 유전자 정보를 포함하는 하나 이상의 주석달기를 수행하고 해당 유전자에 의해 수행하는 기능을 해석하는 하나 이상의 해석 모듈 중에서, 제1 그룹의 분석 대상과 대응되는 제1 해석 모듈을 선택한다.
- [0060] S150에서는 유전자 분석 장치(100)는 제1 전처리 모듈, 제1 추출 모듈, 제1 해석 모듈을 일체의 제1 과정으로 연결한 제1 가상의 영역을 저장 매체에 할당한다.
- [0061] S160에서는 유전자 분석 장치(100)는 일체의 제1 과정을 제1 그룹에 포함되는 하나 이상의 피검체 각각에 대해서 반복적으로 수행한다.
- [0064] 도 4를 참조하면, S310 단계에서, 유전자 분석 장치(100)는 피검체의 유전자 정보에 포함되는 적어도 하나의 단서열에 대한 품질 점수(quality score)를 산출하고, 기준 품질 점수 미만인 상기 단서열을 제거하는 전처리 단계를 수행한다. 유전자 분석 장치(100)는 각 단서열이 포함된 어댑터 서열을 확인하여 제거 한 후, 서열의 평균 품질 점수(quality score)를 계산하여 이 값이 사용자가 선택한 품질 기준값 미만(예를 들어, 품질 점수<20)이면, 단서열을 끝에서부터 시작하여 하나씩 트리밍(자르기, trimming)하고 나머지 염기로 다시 품질 점수를 계산하여 평균값이 기준 품질 점수 이상이 나올 때까지 이 작업을 반복할 수 있다. 최종적으로 얻어진 단서열의 길이가 사용자가 선택한 기준값(예를 들어, 50bp) 보다 작을 경우, 해당 단서열을 제거하고 그렇지 않을 경우 해당 단서열의 필터링을 종료할 수 있다.
- [0065] S320 단계에서, 유전자 분석 장치(100)는 전처리 단계를 거친 단서열 및 인간 게놈 표준 서열을 비교하여 차이를 산출한 후 선택 기준을 정해 신뢰할 수 있는 변이 정보를 추출하는 변이 정보 추출 단계를 수행한다.
- [0066] 도 5에는 변이 정보 추출 단계(S320)의 구체적인 과정을 도시하고 있다. 도 5를 참조하면, S321 단계에서, 유전자 분석 장치(100)는 전처리 단계를 거친 단서열 및 인간 게놈 표준 서열을 비교하여 단서열을 인간 게놈 표준 서열대로 정렬하는 맵핑 단계를 수행한다.
- [0067] S322 단계에서, 유전자 분석 장치(100)는 맵핑단계를 통하여 정렬한 단서열로부터 중합효소 연쇄 반응(PCR, polymerase chain reaction)에 의한 증폭 시에 발생하는 중복 단서열을 제거하는 제거 단계를 수행한다.
- [0068] S323 단계에서, 삽입(insert) 또는 결실(indel)로 인하여 맵핑이 이루어 지지 않은 제거 단계를 거친 단서열에 대하여 지역(local) 재배열을 수행하는 재배열 단계를 수행한다.
- [0069] S324 단계에서, 유전자 분석 장치(100)는 단서열의 성질과 관련하여 지정된 소정의 변수(covariate)를 바탕으로 재보정 테이블을 생성하고, 염기 당 추측되는 오류에 대한 점수로써의 염기(base)의 품질 점수를 재보정하는 재보정 단계를 수행한다.
- [0070] S325 단계에서, 유전자 분석 장치(100)는 재보정 단계를 거친 단서열을 인간 게놈 표준 서열과 비교하여 어느 위치에 어느 정도 맵핑되었는지 여부로 변이 정보를 발굴하는 발굴 단계를 수행한다.
- [0071] 도 4로 돌아와서, S330 단계에서, 유전자 분석 장치(100)는 추출한 변이 정보 및 인간 게놈 표준 서열에 대한 유사도를 비교하고 상기 변이 정보에 대한 위치정보 및 유전자 정보를 포함하는 하나 이상의 주석달기(annotation)를 수행하고 어떤 기능을 수행하는지 해석하는 해석단계를 수행한다.
- [0072] S340 단계에서, 유전자 분석 장치(100)는 추출단계(S320) 및 해석단계(S330)의 처리 결과를 고객이 알아볼 수 있는 형식의 리포트로 생성하는 생성단계를 수행한다.
- [0074] 도 6은 본 발명의 다른 실시 예에 따른 유전자 분석 장치의 동작 방법을 나타내는 흐름도로써 전사체 시퀀싱(RNA sequencing) 방법을 설명하고 있다. 본 발명의 다른 실시 예에 따른 유전자 분석 방법은 도 1 및 도 2에 도시된 바와 같이 주변 구성요소들의 도움을 받아 유전자 분석 장치(100)에서 수행될 수 있다. 이하의 설명에

서 도 1 내지 도 4에 대한 설명과 중복되는 부분은 그 설명을 생략하기로 한다.

- [0075] 전사체 시퀀싱은 전사체의 발현(expression)을 측정하는 매우 예민하고 정확한 방법이다. 전사체 시퀀싱은 샘플 내에 존재하는 모든 전사체(RNA, ribo nucleic acid)에 대한 종합적인 지식을 제공해줄 뿐만 아니라 전사체 각각의 발현 정도를 알 수 있다. 게다가 이중가닥 전사체(stranded RNA) 시퀀싱은 연구자들이 total RNA, mRNA 및 miRNA 염기 서열과 상보적인 염기 서열(sense vs anti-sense)의 방향성 정보를 발견하도록 할 수 있다. 전사체의 발현을 측정하는 매우 예민하고 정확한 수단은 치료법에 대응하여 다른 환경적인 조건과 다른 연구 설계의 넓은 범위 하에 질병 상태에서 나타나는 이전에 발견되지 않은 변화들에 대한 질병 가시성을 제공하고 있다. 또한, 본 실시 예에서는 전사체 시퀀싱에 대한 특별한 응용 기술로, 기존 방식으로는 수행하기 어려웠던 전사체 아형, 유전자 융합, 단일 뉴클레오타이드 변이, 대립형질 특정 유전자 발현과 다른 특징들의 발견을 가능하게 하는 single assay로 기존에 알려진 것과 새로운 특징들 모두에 대한 전사체 데이터 분석을 제한 없이 수행할 수 있다. 기술의 속도는 빨라지고 가격은 낮아지기 때문에 더 많은 특정 연구들에 보편적이고 정확하게 활용될 수 있다.
- [0076] 도 6을 참조하면, S510 단계에서, 유전자 분석 장치(100)는 피검체의 단서열에 대한 품질 점수(quality score), GC 콘텐츠, N 콘텐츠, 길이 분포(length distribution), 중복 레벨(duplication level) 및 over-representative sequence를 분석하고, 품질이 좋지 않은 단서열을 제거함으로써 분석의 정확도를 높이는 단서열 필터링 단계를 수행한다. 단서열 필터링을 수행하기 위해 FastQC 툴을 사용할 수 있다. 단서열 필터링 단계(S510)는 상술한 전처리 단계(S310)에 포함될 수 있다.
- [0077] S520 단계에서, 유전자 분석 장치(100)는 필터링을 완료한 피검체의 단서열을 인간의 cDNA(complementary DNA) 표준에 맵핑시켜, 필터링을 완료한 피검체의 전체 단서열에서 맵핑 비율 및 무작위성(randomness)을 분석하여 종에 맞게 제대로 단서열이 생성되었는지 또는 정확도가 높은 단서열이 생성되었는지 확인하는 제1 맵핑 단계를 수행한다. 제1 맵핑 단계를 수행하기 위해 BWA(burrows-wheeler aligner) 툴을 사용할 수 있다. 제1 맵핑 단계(S520)는 상술한 변이정보 추출 단계(S320)에 포함될 수 있다.
- [0078] S530 단계에서, 유전자 분석 장치(100)는 제1 맵핑을 완료한 피검체의 단서열을 인간 게놈 표준 서열에 맵핑하고, 피검체의 단서열을 표준 서열 순서대로 정렬하는 제2 맵핑 단계를 수행한다. 제2 맵핑 단계를 수행하기 위해 STAR 툴 및 Tophat 툴을 사용할 수 있다.
- [0079] S540 단계에서, 유전자 분석 장치(100)는 제2 맵핑이 종료된 단서열을 잘 알려진 유전자 전사 파일(gene transcript file)과 비교하여 일치하는 전사체가 있는지 또는 새로운 전사체인지 비교하여 새로운 전사체에 대한 위치정보 및 유전자 정보를 포함하는 하나 이상의 주석달기(annotation)를 수행하고 어떤 기능을 수행하는지를 나타내는 유전자 기능 분석(gene ontology) 정보를 해석하는 주석달기 단계를 수행한다. 주석달기 단계를 수행하기 위해 In-house Scripts 툴 및/또는 cufflinks 툴을 사용할 수 있다. S550 단계에서, 유전자 분석 장치(100)는 인간 게놈 표준 서열상의 단서열을 이용하여 전사(transcript)가 어느 정도 발현되었는지 계산하고, 샘플간 발현된 전사체가 어느 정도 차이가 나는지(DEG, differentially expressed genes)를 분석하고, DEG가 유전자 기능 분석(gene ontology) 정보 중 어느 영역에 많이 속해 있는지 분석하는 발현 정도 업데이트 및 농축(enrichment) 단계를 수행한다. 발현 정도 업데이트 및 농축 단계를 수행하기 위해 cuffdiff 툴을 사용할 수 있다.
- [0080] S560 단계에서, 유전자 분석 장치(100)는 분석한 모든 데이터를 HTML 형태로 고객 또는 연구자가 보기 편하게 리포트 형태로 생성하는 리포트 생성 단계를 수행한다.
- [0081] 본 실시 예에 따른 전사체 시퀀싱은 로우 데이터부터 시작해서 DEG, GO까지 손쉽게 한 번에 분석이 가능하게 만들어진 파이프라인이다. 또한 여기서 사용되는 모든 툴 및 순서는 tuxedo protocol을 따르는 것으로 일반적으로 전세계적으로 이용하기 때문에 신뢰성이 높다고 할 수 있다. 파이프라인을 실행하는데 필요한 2개의 파일이 존재하며, 각각의 파일은 샘플 정보 및 사용되는 툴의 옵션을 넣는 파일이다. 우선 샘플 정보를 입력 받는 파일에는 샘플의 이름 및 DEGset, Time-series의 정보를 넣을 수 있는데 한 번 규칙을 알면 누구나 손쉽게 넣을 수 있게 만들어져 있다. 두 번째 파일에서는 툴의 옵션 및 파이프라인이 run하는 순서를 알 수 있게 만들어져 있고, 특정 단계에서 re-run이 필요할 시 이를 지정할 수 있도록 만들어져 있다. 특히 분석 중간에 에러가 발생하더라도 log 폴더에 에러 내용이 저장되어 쉽게 문제점을 파악할 수 있게 만들어져 있고, re-run 하더라도 프로그램 자체에서 분석 결과 파일 생성 여부를 확인해 중간단계부터 실행이 되도록 만들어져 있다. 이는 장치의 사용량 및 시간을 줄여주는데 효과적이라고 할 수 있다. 또한 대량의 샘플이 입력되더라도 클러스터 방식으로 분석을 수행하기 때문에 컴퓨터 자원을 효율적으로 사용하여 빠르게 분석이 가능하다. 그리고 최종 리포트

는 HTML 언어로 제공되므로 대량의 데이터 정보를 사용자가 쉽고 편하게 볼 수 있게 된다.

- [0083] 도 7은 본 발명의 또 다른 실시 예에 따른 유전자 분석 장치의 동작 방법을 나타내는 흐름도로써 전장 엑솜 시퀀싱(whole exome sequencing) 방법을 설명하고 있다. 본 발명의 다른 실시 예에 따른 유전자 분석 방법은 도 1 및 도 2에 도시된 바와 같이 주변 구성요소들의 도움을 받아 유전자 분석 장치(100)에서 수행될 수 있다. 이하의 설명에서 도 1 내지 도 5에 대한 설명과 중복되는 부분은 그 설명을 생략하기로 한다.
- [0084] 단백질을 코딩하고 있는 부분인 엑솜(Exome)은 전체 인간 유전체의 2% 정도 밖에 되지 않지만, 현재까지 알려진 질병 관련 유전자들의 85% 가량이 엑솜에 위치한다고 알려져 있다. 즉 전체 DNA가 아닌 알려진 엑솜 영역만 시퀀싱 함으로써 효율적으로 SNV(single nucleotide variants) 및 IV(indel variants)를 찾을 수 있다.
- [0085] 도 7을 참조하면, S610 단계에서, 유전자 분석 장치(100)는 피검체의 단서열에 대한 품질 점수(quality score), GC 콘텐츠, N 콘텐츠, 길이 분포(length distribution), 중복 레벨(duplication level) 및 over-representative sequence를 분석하고, 품질이 좋지 않은 단서열을 제거 또는 트리밍(trimming)함으로써 분석의 정확도를 높이는 단서열 필터링 또는 트리밍 단계를 수행한다. 단서열 필터링 또는 트리밍 단계를 수행하기 위해 FastQC 툴 및/또는 cutadapt 툴을 사용할 수 있다. 단서열 필터링 또는 트리밍 단계(S610)는 상술한 전처리 단계(S310)에 포함될 수 있다.
- [0086] S620 단계에서, 유전자 분석 장치(100)는 BWA(burrows wheeler aligner) 툴을 이용하여 필터링 또는 트리밍을 완료한 피검체의 단서열을 인간 게놈 표준 서열에 맵핑하고 하나의 파일로 합치는 단서열 배열 단계를 수행한다. 또한 동일한 염기 서열을 가진 단서열이 동일한 위치에 붙었을 때 이를 제거함으로써 정확도를 높일 수 있다. 추가적으로 맵핑 상태를 따로 파일로 기록할 수 있다. 단서열 배열 단계를 수행하기 위해 BWA 툴 및/또는 Picard 툴 및/또는 Samtools를 사용할 수 있다. 단서열 배열 단계(S620)는 상술한 변이정보 추출 단계(S320)에 포함될 수 있다.
- [0087] S630 단계에서, 유전자 분석 장치(100)는 GATK(genome analysis tool kit)를 이용하여 배열을 완료한 피검체의 단서열로부터 변이 정보를 추출하는 변이 정보 추출 단계를 수행한다. 여기서 GATK에 포함되어 있는 옵션을 RealignerTargetCreator → IndelRealigner → BaseRecalibrator → PrintReads → UnifiedGenotyper 순서로 run 하여 최종적으로 변이 정보를 추출한다. 여기서 UnifiedGenotyper 단계에서만 변이 정보를 추출하고, 앞의 4 단계는 변이 정보 추출의 정확도를 높이기 위해 실행할 수 있다. 변이 정보 추출 단계(S630)는 상술한 변이 정보 추출 단계(S320)에 포함될 수 있다.
- [0088] S640 단계에서, 유전자 분석 장치(100)는 추출이 완료된 변이 정보(SNVs, Indels)에 잘 알려진 데이터베이스(dbNSFP, Cosmic, ClinVar, 1000Genomes, ESP6500, Exac)의 정보를 붙여 주는 주석달기 단계를 수행한다. 주석달기 단계를 수행하기 위해 SnpEff 툴을 사용할 수 있다. 주석달기 단계(S640)는 상술한 해석단계(S330)에 포함될 수 있다.
- [0089] S650 단계에서, 유전자 분석 장치(100)는 fastq 파일, bam 파일, 변이 정보추출에 대한 전체적인 통계 파일(statistics) 예를 들어, read depth, read coverage, unique reads information, Variants Count를 생성하는 통계 파일 생성 단계를 수행한다. 통계 파일 생성 단계를 수행하기 위해 InHouse Script 툴 및/또는 Qualimap 툴 및/또는 BedTools 를 이용할 수 있다. 통계 파일 생성 단계(S650)는 상술한 해석단계(S330)에 포함될 수 있다.
- [0090] S660 단계에서, 유전자 분석 장치(100)는 분석한 모든 데이터 예를 들어, 변이 정보, 통계 정보를 PDF 파일 형식으로 고객 또는 연구자가 보기 편하게 리포트 형태로 생성하는 리포트 생성 단계를 수행한다.
- [0091] 본 실시 예에서 상술한 각각의 단계에서 사용되는 툴은 모두 논문에서 잘 알려진 것만 사용하기 때문에 연구용으로 적합할 수 있다. 분석되는 결과 파일은 특정 폴더에 순서대로 저장될 수 있다. 즉, 누구나 보기 쉽게 만들어져 있기 때문에 고객이 추가 분석을 할 때 용이하다고 할 수 있다. 그리고 에러가 발생할 시에 log 내용도 따로 저장되어 쉽게 문제를 해결할 수 있을 뿐만 아니라, 그 시점부터 다시 파이프라인이 실행 가능하기 때문에 분석 시간을 단축할 수 있다. 또한, 확장성이 좋다고 할 수 있는데, 그 이유는 샘플을 분석하기 위해 초기 실행 시 2개의 입력 파일을 받는다. 한 개는 샘플의 정보 및 툴의 위치, 결과 파일 저장소를 입력 받는 것과, 나머지 한 개는 툴이 run 되는 파일이다. 즉, 여기에 추가로 분석하기 위한 기본 정보(툴의 위치, 파이프라인에서의 순서 및 옵션)를 입력하고, python code로 간단히 실행 파일을 만들어 놓으면 파이프라인이 확장되어 다음 분석에서는 자동으로 확장된 부분이 추가되어 분석이 이루어 진다. 마지막으로 대량 분석이 가능한데, 이는 샘플

플 한 개씩 순서대로 분석하는 것이 아니라 컴퓨터의 성능이 최대로 이용 가능하도록 동시에 분석이 이루어진다. 혹 많은 샘플이 분석되더라도 차례대로 이루어지기 때문에 컴퓨터가 다운되는 일은 거의 없다고 할 수 있다. 이처럼 약간의 리눅스 이용이 가능하다면 누구나 쉽게 이용할 수 있게 된다.

- [0093] 도 8은 본 발명의 일 실시 예에 따른 컴퓨팅 장치의 하드웨어 구성을 도시한 도면이다. 도 8을 참고하면, 컴퓨팅 장치(1)는 유전자 분석 장치(프로세서)(100), 데이터 인터페이스(11) 및 메모리(12)를 포함할 수 있다. 한편, 도 7에 도시된 컴퓨팅 장치(1)는 본 실시예의 특징이 흐려지는 것을 방지하기 위하여 본 실시 예에 관련된 구성요소들만이 도시되어 있을 뿐이므로, 도 7에 도시된 구성요소들 외에 다른 범용적인 구성요소들이 더 포함될 수 있다.
- [0094] 데이터 인터페이스(11)는 앞서 도 1에서 설명된, 피검체로부터 획득한 유전자 정보 및 정상인 집단으로부터 획득한 표준 유전자 정보를 수신할 수 있다. 즉, 데이터 인터페이스(11)는 컴퓨팅 장치(1)가 외부의 다른 디바이스들과 통신하기 위한 유/무선 네트워크 인터페이스의 하드웨어로 구현될 수 있다. 데이터 인터페이스(11)는 수신된 피검체로부터 획득한 유전자 정보 및 정상인 집단으로부터 획득한 표준 유전자 정보를 유전자 분석 장치(프로세서)(100)로 전송할 수 있다.
- [0095] 데이터 인터페이스(11)는 데이터베이스(DB)(미도시)로부터 정상인 집단으로부터 획득한 표준 유전자 정보를 수신할 수 있다. 그리고, 데이터 인터페이스(11)는 피검체의 피검 유전자를 시퀀싱하기 위한 외부의 차세대 시퀀싱 장치, 마이크로어레이 등으로부터 피검체로부터 획득한 유전자 정보를 수신할 수 있다.
- [0096] 메모리(12)는 컴퓨팅 장치(1) 내에서 처리될 데이터들 및 처리가 완료된 결과들을 저장하기 위한 하드웨어로서, RAM(random access memory), ROM(read only memory) 등의 메모리 칩들 또는 HDD(hard disk drive), SSD(solid state drive) 등의 스토리지를 포함한다. 즉, 메모리(12)는 데이터 인터페이스(11)에 의해 수신된 피검체로부터 획득한 유전자 정보 및 정상인 집단으로부터 획득한 표준 유전자 정보를 저장할 수 있고, 유전자 분석 장치(프로세서)(100)에 의해 처리된 레퍼런스 데이터 세트, 피검 유전자들에 대한 딥 시퀀싱 데이터, 식별된 변이 정보에 대한 데이터도 저장할 수 있다.
- [0097] 유전자 분석 장치(프로세서)(100)는 하나 이상의 프로세싱 유닛들로 구현된 모듈로서, 다수의 논리 게이트들의 어레이를 갖는 마이크로프로세서와 이 마이크로프로세서에서 실행될 수 있는 프로그램이 저장된 메모리 모듈의 조합으로 구현될 수도 있다. 유전자 분석 장치(프로세서)(100)는 응용 프로그램의 모듈 형태로 구현될 수도 있다. 유전자 분석 장치(프로세서)(100)는 앞서 도 1 내지 도 6에서 설명된 유전자 분석을 처리하는 하드웨어 장치이다.
- [0098] 유전자 분석 장치(프로세서)(100)에 의해 식별된 변이 정보는 데이터 인터페이스(11)를 통해 외부의 다른 디바이스, 예를 들어 디스플레이 디바이스, 다른 컴퓨팅 장치 등으로 전송되거나, 또는 외부 네트워크, 예를 들어 인터넷, 공개 데이터베이스(DB) 서버 상으로 전송될 수 있다.
- [0100] 이상 설명된 본 발명에 따른 실시 예는 컴퓨터 상에서 다양한 구성요소를 통하여 실행될 수 있는 컴퓨터 프로그램의 형태로 구현될 수 있으며, 이와 같은 컴퓨터 프로그램은 컴퓨터로 판독 가능한 매체에 기록될 수 있다. 이때, 매체는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM 및 DVD와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical medium), 및 ROM, RAM, 플래시 메모리 등과 같은, 프로그램 명령어를 저장하고 실행하도록 특별히 구성된 하드웨어 장치를 포함할 수 있다.
- [0101] 한편, 상기 컴퓨터 프로그램은 본 발명을 위하여 특별히 설계되고 구성된 것이거나 컴퓨터 소프트웨어 분야의 당업자에게 공지되어 사용 가능한 것일 수 있다. 컴퓨터 프로그램의 예에는, 컴파일러에 의하여 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용하여 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함될 수 있다.
- [0102] 본 발명의 명세서(특히 특허청구범위에서)에서 "상기"의 용어 및 이와 유사한 지시 용어의 사용은 단수 및 복수 모두에 해당하는 것일 수 있다. 또한, 본 발명에서 범위(range)를 기재한 경우 상기 범위에 속하는 개별적인 값을 적용한 발명을 포함하는 것으로서(이에 반하는 기재가 없다면), 발명의 상세한 설명에 상기 범위를 구성하는 각 개별적인 값을 기재한 것과 같다.
- [0103] 본 발명에 따른 방법을 구성하는 단계들에 대하여 명백하게 순서를 기재하거나 반하는 기재가 없다면, 상기 단

계들은 적당한 순서로 행해질 수 있다. 반드시 상기 단계들의 기재 순서에 따라 본 발명이 한정되는 것은 아니다. 본 발명에서 모든 예들 또는 예시적인 용어(예들 들어, 등등)의 사용은 단순히 본 발명을 상세히 설명하기 위한 것으로서 특허청구범위에 의해 한정되지 않는 이상 상기 예들 또는 예시적인 용어로 인해 본 발명의 범위가 한정되는 것은 아니다. 또한, 당업자는 다양한 수정, 조합 및 변경이 부가된 특허청구범위 또는 그 균등물의 범주 내에서 설계 조건 및 팩터에 따라 구성될 수 있음을 알 수 있다.

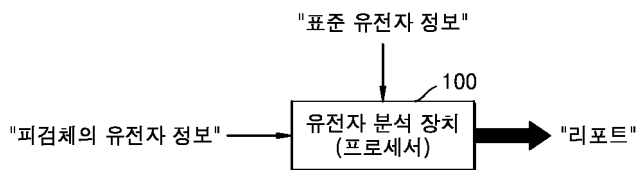
[0104] 따라서, 본 발명의 사상은 상기 설명된 실시 예에 국한되어 정해져서는 아니 되며, 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등한 또는 이로부터 등가적으로 변경된 모든 범위는 본 발명의 사상의 범주에 속한다고 할 것이다.

부호의 설명

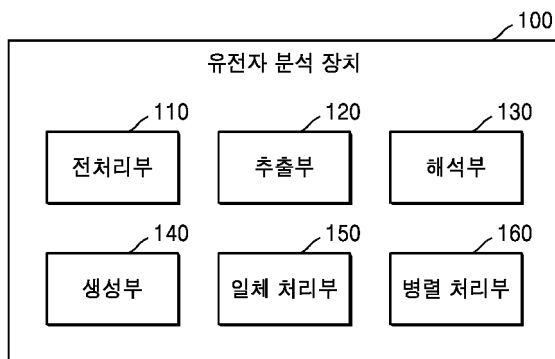
[0105] 100: 유전자 분석 장치
110: 전처리부
120: 추출부
130: 해석부
140: 생성부
150: 일체 처리부
160: 병렬 처리부

도면

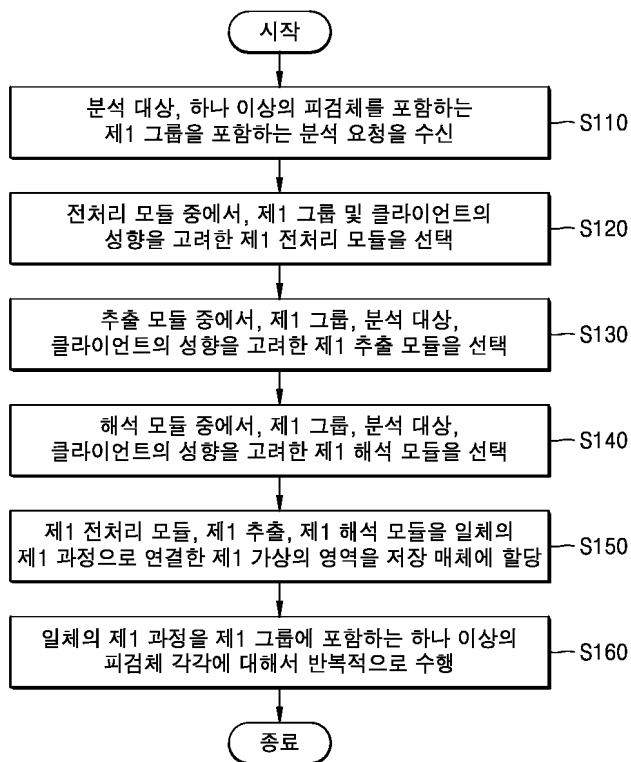
도면1



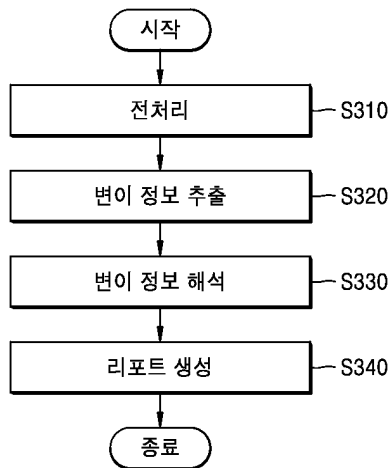
도면2



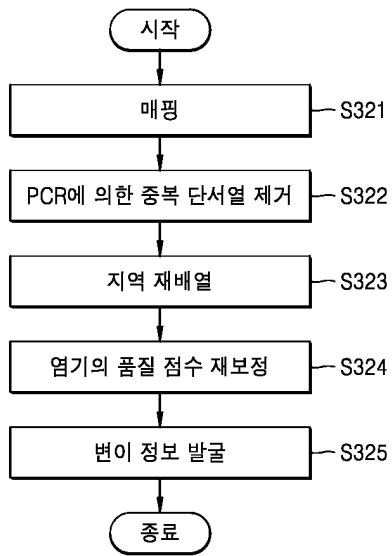
도면3



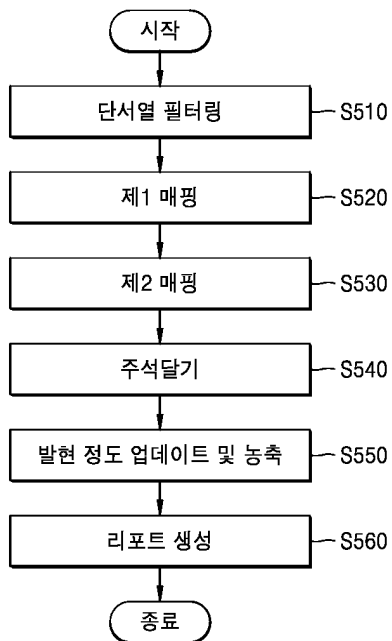
도면4



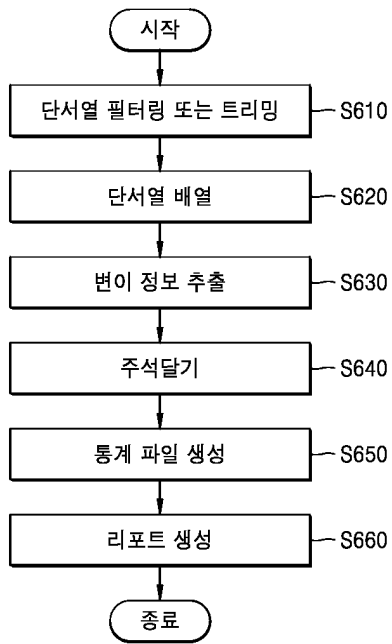
도면5



도면6



도면7



도면8

