(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0124391
(43) 공개일자 2024년08월16일

(51) 국제특허분류(Int. Cl.)
G16B 25/10 (2019.01) *G16B 20/20* (2019.01)
G16B 30/10 (2019.01) *G16B 40/20* (2019.01)
G16H 10/40 (2018.01) *G16H 20/10* (2018.01)
G16H 50/20 (2018.01) *G16H 50/30* (2018.01)
G16H 50/70 (2018.01)
(52) CPC특허분류
G16B 25/10 (2019.02)
G16B 20/20 (2019.02)
(21) 출원번호 10-2024-7024939
(22) 출원일자(국제) 2022년12월28일
심사청구일자 없음
(85) 번역문제출일자 2024년07월23일
(86) 국제출원번호 PCT/US2022/054196
(87) 국제공개번호 WO 2023/129622
국제공개일자 2023년07월06일
(30) 우선권주장
63/294,813 2021년12월29일 미국(US)
(뒷면에 계속)

(71) 출원인
일루미나, 인코포레이티드
미국 캘리포니아 92122 샌디에고 일루미나 웨이 5200
(72) 발명자
피지에프 페트코 플라메노프
미국 캘리포니아주 92122 샌디에고 일루미나 웨이 5200 일루미나, 인코포레이티드 내
맥레이 제레미 프란시스
미국 캘리포니아주 92122 샌디에고 일루미나 웨이 5200 일루미나, 인코포레이티드 내
파 카이-하우
미국 캘리포니아주 92122 샌디에고 일루미나 웨이 5200 일루미나, 인코포레이티드 내
(74) 대리인
특허법인아주김장리

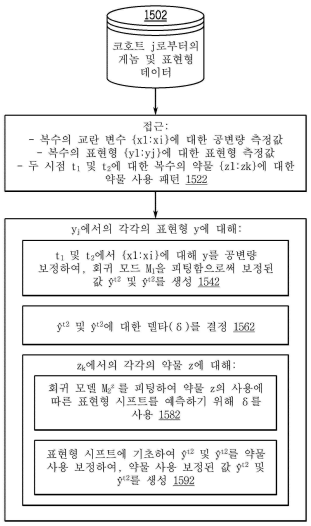
전체 청구항 수 : 총 34 항

(54) 발명의 명칭 **상이한 약물 사용 패턴에 대한 표현형 측정값으로부터의 시간 데이터에 대한 공변량 보정**

(57) 요 약

복수의 교란 변수를 갖는 개체 코호트의 복수의 표현형에 대한 복수의 약물의 사용에 따른 표현형 시프트를 예측하는 컴퓨터 구현 방법. 개체 코호트는 두 개의 개별 시점에 대한 연관된 표현형 측정값, 공변량 측정값 및 약물 사용 패턴을 갖는다. 제1 시점 및 제2 시점에 대한 표현형 측정값은 생물 통계학의 사용을 통해 공변량 보정되고 약물 사용 보정된다.

대 표 도 - 도15



(52) CPC특허분류
G16B 30/10 (2019.02)
G16B 40/20 (2019.02)
G16H 10/40 (2021.08)
G16H 20/10 (2021.08)
G16H 50/20 (2018.01)
G16H 50/30 (2018.01)
G16H 50/70 (2018.01)

(30) 우선권주장
63/294,816 2021년12월29일 미국(US)
63/294,820 2021년12월29일 미국(US)
63/294,827 2021년12월29일 미국(US)
63/294,828 2021년12월29일 미국(US)
63/294,830 2021년12월29일 미국(US)
63/351,283 2022년06월10일 미국(US)
63/351,299 2022년06월10일 미국(US)
63/351,317 2022년06월10일 미국(US)
17/968,766 2022년10월18일 미국(US)

명 세 서

청구범위

청구항 1

복수의 교란 변수를 갖는 개체 코호트의 복수의 표현형에 대한 복수의 약물의 사용에 따른 표현형 시프트를 예측하는 컴퓨터 구현 방법으로서,

상기 개체 코호트에 대해 그리고 제1 시점 및 제2 시점에 대해:

상기 복수의 표현형에 대한 표현형 측정값에 접근하는 단계;

상기 복수의 교란 변수에 대한 공변량 측정값에 접근하는 단계; 및

상기 복수의 약물에 대한 약물 사용 패턴에 접근하는 단계를 포함하고,

표현형별 기준으로:

상기 공변량 측정값에 기초하여 상기 제1 시점 및 제2 시점에 대한 상기 표현형 측정값을 공변량 보정하여, 상기 제1 시점 및 제2 시점에 대한 공변량 보정 표현형 측정값을 생성하는 단계;

상기 제1 시점과 제2 시점에 대한 상기 공변량 보정 표현형 측정값 사이의 차이에 기초하여 델타를 결정하는 단계;

각각의 상기 약물 사용 패턴에 대해, 상기 공변량 보정 표현형 측정값에서의 상기 복수의 약물의 사용에 따른 표현형 시프트를 예측하기 위해 상기 델타를 사용하는 제2 회귀 모델을 피팅하는 단계; 및

상기 표현형 시프트에 기초하여 상기 제1 시점 및 제2 시점에 대한 상기 표현형 측정값을 약물 사용 보정하여, 상기 제1 시점에 대한 약물 사용 보정 표현형 측정값을 생성하는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 2

제1항에 있어서, 상기 복수의 교란 변수는 연령, 성별, 유전적 주요 성분, 식이 및 흡연 상태를 포함하는, 컴퓨터 구현 방법.

청구항 3

제1항에 있어서, 상기 공변량 보정은 제1 회귀 모델을 피팅하여 상기 공변량 측정값을 회귀시킴으로써 구현되는, 컴퓨터 구현 방법.

청구항 4

제1항에 있어서, 상기 표현형 시프트 예측은 각각의 상기 약물 사용 패턴에 대한 표현형 시프트를 모델링하는 제2 회귀 모델을 피팅함으로써 구현되는, 컴퓨터 구현 방법.

청구항 5

제4항에 있어서, 상기 제2 회귀 모델은 각각의 상기 약물 사용 패턴에 대한 상기 표현형 시프트를 연속적으로 그리고 누적적으로 포함함으로써 상기 델타를 반복적으로 예측하는 전진 선택 단계적 회귀 모델인, 컴퓨터 구현 방법.

청구항 6

제4항에 있어서, 상기 약물 사용 패턴은,

상기 제1 시점 및 제2 시점에 약물을 사용하지 않는 것,

상기 제1 시점과 제2 시점 사이에 상기 약물을 사용하기 시작하는 것,

상기 제1 시점과 제2 시점 사이에 상기 약물을 사용하는 것을 중단하는 것, 및

상기 제1 시점 및 제2 시점에 상기 약물을 사용하는 것을 포함하는, 컴퓨터 구현 방법.

청구항 7

제6항에 있어서, 상기 제2 회귀 모델은 각각의 상기 약물 사용 패턴에 대한 이진 지시자 독립 변수를 갖는, 컴퓨터 구현 방법.

청구항 8

제1항에 있어서, 상기 공변량 보정, 상기 델타 결정, 상기 표현형 시프트 예측 및 상기 약물 사용 보정은 상기 복수의 약물에서의 약물에 대해 약물별 기준으로 실행되는, 컴퓨터 구현 방법.

청구항 9

제8항에 있어서, 상기 제2 회귀 모델은 각각의 상기 약물에 대해 반복적으로 피팅되는, 컴퓨터 구현 방법.

청구항 10

제8항에 있어서, 상기 약물을 약물 카테고리 세트로 그룹화하는 단계를 더 포함하는 컴퓨터 구현 방법.

청구항 11

제10항에 있어서, 상기 공변량 보정, 상기 델타 결정, 상기 표현형 시프트 예측 및 상기 약물 사용 보정은 상기 약물 카테고리 세트에서의 약물 카테고리에 대해 약물 카테고리별 기준으로 실행되는, 컴퓨터 구현 방법.

청구항 12

제11항에 있어서, 상기 제2 회귀 모델은 각각의 상기 약물 카테고리에 대해 반복적으로 피팅되는, 컴퓨터 구현 방법.

청구항 13

제1항에 있어서, 상기 제2 회귀 모델은 상기 개체 코호트 내의 개체에 대한 개체별 기준으로 상기 제1 시점과 제2 시점 사이에 경과된 시간에 따라 표현형 시프트를 더 모델링하는, 컴퓨터 구현 방법.

청구항 14

제1항에 있어서, 상기 제2 회귀 모델은 상기 제1 시점과 제2 시점 사이의 평균으로의 회귀에 따라 표현형 시프트를 더 모델링하는, 컴퓨터 구현 방법.

청구항 15

제8항에 있어서, 상기 제2 회귀 모델은 상기 복수의 약물에서 관련 약물 세트에 대해 공동으로 피팅되는, 컴퓨터 구현 방법.

청구항 16

제1항에 있어서, 상기 약물 사용 보정은 제3 회귀 모델을 피팅함으로써 구현되는, 컴퓨터 구현 방법.

청구항 17

제16항에 있어서, 상기 제1 시점과 제2 시점 사이에 상기 약물을 사용하기 시작하는 제1 약물 사용 패턴에 대한 제1 이진 지시자 독립 변수, 상기 제1 시점 및 제2 시점에 상기 약물을 사용하지 않는 제2 약물 사용 패턴에 대한 제2 이진 지시자 독립 변수 및 개체가 상기 제1 시점에 특정 약물을 사용하고 있었는지 여부를 인코딩하는 약물 특정 이진 지시자 독립 변수에 기초하여 상기 제1 시점에 대한 상기 표현형 측정값을 약물 사용 보정하는 단계를 더 포함하는 컴퓨터 구현 방법.

청구항 18

제1항에 있어서, 상기 약물 사용 보정은 제4 회귀 모델을 피팅함으로써 구현되는, 컴퓨터 구현 방법.

청구항 19

제18항에 있어서, 상기 제1 시점과 제2 시점 사이에 상기 약물을 사용하는 것을 중단하는 제3 약물 사용 패턴에 대한 제3 이진 지시자 독립 변수, 상기 제1 시점 및 제2 시점에 상기 약물을 사용하는 제4 약물 사용 패턴에 대한 제4 이진 지시자 독립 변수 및 개체가 상기 제2 시점에 특정 약물을 사용하고 있었는지 여부를 인코딩하는 약물 특정 이진 지시자 독립 변수에 기초하여 상기 제2 시점에 대한 상기 표현형 측정값을 약물 사용 보정하는 단계를 더 포함하는 컴퓨터 구현 방법.

청구항 20

제1항에 있어서, 순위 기반 역정규 변환을 상기 제1 시점 및 제2 시점에 대한 상기 약물 사용 보정 표현형 측정값에 적용하여 상기 제1 시점 및 제2 시점에 대한 정규화된 약물 사용 보정 표현형 측정값을 생성하는 단계를 더 포함하는 컴퓨터 구현 방법.

청구항 21

제20항에 있어서, 상기 제1 시점 및 제2 시점에 대한 상기 정규화된 약물 사용 보정 표현형 측정값을 공변량 보정하여, 상기 제1 시점 및 제2 시점에 대한 공변량 보정되고 정규화된 약물 사용 보정 표현형 측정값을 생성하는 단계를 더 포함하는 컴퓨터 구현 방법.

청구항 22

제21항에 있어서, 상기 공변량 보정되고 정규화된 약물 사용 보정 표현형 측정값을 사용하여 희귀 변이체 다유전자 위험 점수를 생성하는 단계를 더 포함하는 컴퓨터 구현 방법.

청구항 23

제1항에 있어서, 상기 복수의 표현형은 복수의 정량적 표현형에 대응하는, 컴퓨터 구현 방법.

청구항 24

제23항에 있어서, 상기 복수의 정량적 표현형에서의 정량적 표현형은 정량적 바이오마커 측정값인, 컴퓨터 구현 방법.

청구항 25

제23항에 있어서, 상기 공변량 보정, 상기 델타 결정, 상기 표현형 시프트 및 상기 약물 사용 보정에 사용하기 위해 상기 복수의 정량적 표현형을 중복되지 않은 세트로 프루닝하는 단계를 더 포함하는 컴퓨터 구현 방법.

청구항 26

제25항에 있어서, 상기 중복되지 않은 세트 내의 정량적 표현형의 각각의 쌍은 상한 임계값보다 낮은 절대 쌍별 피어슨 상관 관계를 갖는, 컴퓨터 구현 방법.

청구항 27

제26항에 있어서, 상기 상한 임계값은 0.95인, 컴퓨터 구현 방법.

청구항 28

제26항에 있어서, 상기 복수의 정량적 표현형 내의 중복된 정량적 표현형의 각각의 그룹 중에서, 상기 중복되지 않은 세트에 포함시키기 위해 가장 많은 샘플을 갖는 표현형을 선택하는 단계를 더 포함하는 컴퓨터 구현 방법.

청구항 29

제1항에 있어서, 상기 복수의 표현형은 복수의 범주형 표현형에 대응하는, 컴퓨터 구현 방법.

청구항 30

제29항에 있어서, 상기 복수의 범주형 표현형에서의 범주형 표현형은 임상 진단인, 컴퓨터 구현 방법.

청구항 31

제4항에 있어서, 상기 제2 회귀 모델을 사용하여 약물-표현형 연관성을 검출하는 단계를 더 포함하는 컴퓨터 구현 방법.

청구항 32

제31항에 있어서, 상기 약물-표현형 연관성은 잠재적인 원치 않는 부작용 및 원하는 표적 효과를 포함하는, 컴퓨터 구현 방법.

청구항 33

시스템으로서, 메모리에 커플링된 하나 이상의 프로세서를 포함하고, 상기 메모리에는 복수의 교란 변수를 갖는 개체 코호트의 복수의 표현형에 대한 복수의 약물의 사용에 따른 표현형 시프트를 예측하기 위한 컴퓨터 명령어가 로딩되고, 상기 명령어는, 상기 프로세서에서 실행될 때,

상기 개체 코호트에 대해 그리고 제1 시점 및 제2 시점에 대해:

상기 복수의 표현형에 대한 표현형 측정값에 접근하는 것;

상기 복수의 교란 변수에 대한 공변량 측정값에 접근하는 것; 및

상기 복수의 약물에 대한 약물 사용 패턴에 접근하는 것을 포함하고,

표현형별 기준으로:

상기 공변량 측정값에 기초하여 상기 제1 시점 및 제2 시점에 대한 상기 표현형 측정값을 공변량 보정하여, 상기 제1 시점 및 제2 시점에 대한 공변량 보정 표현형 측정값을 생성하는 것;

상기 제1 시점과 제2 시점에 대한 상기 공변량 보정 표현형 측정값 사이의 차이에 기초하여 델타를 결정하는 것;

각각의 상기 약물 사용 패턴에 대해, 상기 공변량 보정 표현형 측정값에서의 상기 복수의 약물의 사용에 따른 표현형 시프트를 예측하기 위해 상기 델타를 사용하는 것; 및 상기 표현형 시프트에 기초하여 상기 제1 시점에 대한 상기 표현형 측정값을 약물 사용 보정하여, 상기 제1 시점 및 제2 시점에 대한 약물 사용 보정 표현형 측정값을 생성하는 것을 포함하는 동작들을 구현하는, 시스템.

청구항 34

복수의 교란 변수를 갖는 개체 코호트의 복수의 표현형에 대한 복수의 약물의 사용에 따른 표현형 시프트를 예측하기 위한 컴퓨터 프로그램 명령어가 저장된 비일시적 컴퓨터 판독 가능 저장 매체로서, 상기 명령어는, 프로세서에서 실행될 때,

상기 개체 코호트에 대해 그리고 제1 및 제2 시점에 대해:

상기 복수의 표현형에 대한 표현형 측정값에 접근하는 단계;

상기 복수의 교란 변수에 대한 공변량 측정값에 접근하는 단계; 및

상기 복수의 약물에 대한 약물 사용 패턴에 접근하는 단계를 포함하고,

표현형별 기준으로:

상기 공변량 측정값에 기초하여 상기 제1 시점 및 제2 시점에 대한 상기 표현형 측정값을 공변량 보정하여, 상기 제1 시점 및 제2 시점에 대한 공변량 보정 표현형 측정값을 생성하는 단계;

상기 제1 시점과 제2 시점에 대한 상기 공변량 보정 표현형 측정값 사이의 차이에 기초하여 델타를 결정하는 단계;

각각의 상기 약물 사용 패턴에 대해, 상기 공변량 보정 표현형 측정값에서의 상기 복수의 약물의 사용에 따른 표현형 시프트를 예측하기 위해 상기 델타를 사용하는 단계; 및 상기 표현형 시프트 예측에 기초하여 상기 제1 시점 및 제2 시점에 대한 상기 표현형 측정값을 약물 사용 보정하여, 상기 제1 시점 및 제2 시점에 대한 약물 사용 보정 표현형 측정값을 생성하는 단계를 포함하는 방법을 구현하는, 비일시적 컴퓨터 판독 가능 저장 매체.

발명의 설명

기 술 분 야

[0001] 우선권 출원

[0002] 본 출원은 다음의 이익 및 그에 대한 우선권을 주장한다.

[0003] 2021년 12월 29일자로 출원되고 발명의 명칭이 "PERIODIC MASK PATTERN FOR REVELATION LANGUAGE MODELS"인 미국 임시 특허 출원 제63/294,813호(대리인 문서 번호 ILLM 1063-1/IP-2296-PRV);

[0004] 2021년 12월 29일자로 출원되고 발명의 명칭이 "CLASSIFYING MILLIONS OF VARIANTS OF UNCERTAIN SIGNIFICANCE USING PRIMATE SEQUENCING AND DEEP LEARNING"인 미국 임시 특허 출원 제63/294,816호(대리인 문서 번호 ILLM 1064-1/IP-2297-PRV);

[0005] 2021년 12월 29일자로 출원되고 발명의 명칭이 "IDENTIFYING GENES WITH DIFFERENTIAL SELECTIVE CONSTRAINT BETWEEN HUMANS AND NON-HUMAN PRIMATES"인 미국 임시 특허 출원 제63/294,820호(대리인 문서 번호 ILLM 1065-1/IP-2298-PRV);

[0006] 2021년 12월 29일자로 출원되고 발명의 명칭이 "DEEP LEARNING NETWORK FOR EVOLUTIONARY CONSERVATION"인 미국 임시 특허 출원 제63/294,827호(대리인 문서 번호 ILLM 1066-1/IP-2299-PRV);

[0007] 2021년 12월 29일자로 출원되고 발명의 명칭이 "INTER-MODEL PREDICTION SCORE RECALIBRATION"인 미국 임시 특허 출원 제63/294,828호(대리인 문서 번호 ILLM 1067-1/IP-2301-PRV);

[0008] 2021년 12월 29일자로 출원되고 발명의 명칭이 "SPECIES-DIFFERENTIABLE EVOLUTIONARY PROFILES"인 미국 임시 특허 출원 제63/294,830호(대리인 문서 번호 ILLM 1068-1/IP-2302-PRV);

[0009] 2022년 6월 10일자로 출원되고 발명의 명칭이 "OPTIMIZED BURDEN TEST BASED ON NESTED T-TESTS THAT MAXIMIZE SEPARATION BETWEEN CARRIERS AND NON-CARRIERS" 미국 임시 특허 출원 제63/351,283호(대리인 문서 번호 ILLM 1070-1/IP-2368-PRV);

[0010] 2022년 6월 10일자로 출원되고 발명의 명칭이 "RARE VARIANT POLYGENIC RISK SCORES"인 미국 임시 특허 출원 제63/351,299호(대리인 문서 번호 ILLM 1071-1/IP-2378-PRV); 및

[0011] 2022년 6월 10일자로 출원되고 발명의 명칭이 "COVARIATE CORRECTION INCLUDING DRUG USE FROM TEMPORAL DATA"인 미국 임시 특허 출원 제63/351,317호(대리인 문서 번호 ILLM 1073-1/IP-2387-PRV).

[0012] 상기 우선권 출원들은 본원에 완전히 기재된 것처럼 인용되어 포함된다.

[0013] **기술분야**

[0014] 개시된 기술은 인공 지능 유형 컴퓨터 및 디지털 처리 시스템 및 대응하는 데이터 처리 방법 및 지능 에뮬레이션 제품(즉, 지식 기반 시스템, 추론 시스템 및 지식 획득 시스템)에 관한 것이며; 불확실성이 있는 추론을 위한 시스템(예를 들어, 퍼지 로직 시스템), 적응 시스템, 기계 학습 시스템 및 인공 신경망을 포함한다. 특히, 개시된 기술은 심층 컨볼루션 신경망을 사용하여 정렬된 데이터를 분석하는 것에 관한 것이다.

[0015] **관련 출원**

[0016] 본 출원은 2022년 10월 18일자로 출원되고 발명의 명칭이 "OPTIMIZED BURDEN TEST BASED ON NESTED T-TESTS THAT MAXIMIZE SEPARATION BETWEEN CARRIERS AND NON-CARRIERS"인 미국 정규 특허 출원 제17/968,285호(대리인 문서 번호 ILLM 1070-2/IP-2368-US)와 관련된다.

[0017] 관련 출원은 모든 목적을 위해 본원에 인용되어 포함된다.

[0018] 본 출원은 2022년 10월 18일자로 출원되고 발명의 명칭이 "RARE VARIANT POLYGENIC RISK SCORES"인 미국 정규 특허 출원 제17/968,723호(대리인 문서 번호 ILLM 1071-2/IP-2378-US)와 관련된다.

[0019] 관련 출원은 모든 목적을 위해 본원에 인용되어 포함된다.

[0020] **참조에 의한 통합**

[0021] 다음은 본원에 완전히 기재된 것처럼 모든 목적을 위해 인용되어 포함된다:

[0022] Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. Nat. Genet. 50, 1161-1170 (2018);

[0023] Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. Cell 176, 535-548 (2019);

[0024] 2017년 10월 16일자로 출원되고 발명의 명칭이 "TRAINING A DEEP PATHOGENICITY CLASSIFIER USING LARGE-SCALE BENIGN TRAINING DATA"인 미국 특허 출원 제62/573,144호(대리인 문서 번호 ILLM 1000-1/IP-1611-PRV);

[0025] 2017년 10월 16일자로 출원되고 발명의 명칭이 "PATHOGENICITY CLASSIFIER BASED ON DEEP CONVOLUTIONAL NEURAL NETWORKS (CNNs)"인 미국 특허 출원 제62/573,149호(대리인 문서 번호 ILLM 1000-2/IP-1612-PRV);

[0026] 2017년 10월 16일자로 출원되고 발명의 명칭이 "DEEP SEMI-SUPERVISED LEARNING THAT GENERATES LARGE-SCALE PATHOGENIC TRAINING DATA"인 미국 특허 출원 제62/573,153호(대리인 문서 번호 ILLM 1000-3/IP-1613-PRV);

[0027] 2017년 11월 7일자로 출원되고 발명의 명칭이 "PATHOGENICITY CLASSIFICATION OF GENOMIC DATA USING DEEP CONVOLUTIONAL NEURAL NETWORKS (CNNs)"인 미국 특허 출원 제62/582,898호(대리인 문서 번호 ILLM 1000-4/IP-1618-PRV);

[0028] 2018년 10월 15일자로 출원되고 발명의 명칭이 "DEEP LEARNING-BASED TECHNIQUES FOR TRAINING DEEP CONVOLUTIONAL NEURAL NETWORKS"인 미국 특허 출원 제16/160,903호(대리인 문서 번호 ILLM 1000-5/IP-1611-US);

[0029] 2018년 10월 15일자로 출원되고 발명의 명칭이 "DEEP CONVOLUTIONAL NEURAL NETWORKS FOR VARIANT CLASSIFICATION"인 미국 특허 출원 제16/160,986호(대리인 문서 번호 ILLM 1000-6/IP-1612-US);

[0030] 2018년 10월 15일자로 출원되고 발명의 명칭이 "SEMI- SUPERVISED LEARNING FOR TRAINING AN ENSEMBLE OF DEEP CONVOLUTIONAL NEURAL NETWORKS"인 미국 특허 출원 제16/160,968호(대리인 문서 번호 ILLM 1000-7/IP-1613-US);

[0031] 2018년 10월 15일자로 출원되고 발명의 명칭이 "DEEP LEARNING-BASED SPLICE SITE CLASSIFICATION"인 미국 특허 출원 제16/160,978호(대리인 문서 번호 ILLM 1001-4/IP-1680-US);

[0032] 2019년 5월 8일자로 출원되고 발명의 명칭이 "DEEP LEARNING-BASED TECHNIQUES FOR PRE-TRAINING DEEP CONVOLUTIONAL NEURAL NETWORKS"인 미국 특허 출원 제16/407,149호(대리인 문서 번호 ILLM 1010-1 /IP-1734-US);

[0033] 2021년 4월 15일자로 출원되고 발명의 명칭이 "DEEP CONVOLUTIONAL NEURAL NETWORKS TO PREDICT VARIANT PATHOGENICITY USING THREE-DIMENSIONAL (3D) PROTEIN STRUCTURES"인 미국 특허 출원 제17/232,056호(대리인 문서 번호 ILLM 1037-2/IP-2051-US);

[0034] 2021년 4월 15일자로 출원되고 발명의 명칭이 "MULTI-CHANNEL PROTEIN VOXELIZATION TO PREDICT VARIANT PATHOGENICITY USING DEEP CONVOLUTIONAL NEURAL NETWORKS"인 미국 특허 출원 제63/175,495호(대리인 문서 번호 ILLM 1047-1/IP-2142-PRV);

[0035] 2021년 4월 16일자로 출원되고 발명의 명칭이 "EFFICIENT VOXELIZATION FOR DEEP LEARNING"인 미국 특허 출원 제63/175,767호(대리인 문서 번호 ILLM 1048-1/IP-2143-PRV);

[0036] 2021년 9월 7일자로 출원되고 발명의 명칭이 "ARTIFICIAL INTELLIGENCE-BASED ANALYSIS OF PROTEIN THREE-DIMENSIONAL (3D) STRUCTURES"인 미국 특허 출원 제17/468,411호(대리인 문서 번호 ILLM 1037-3/IP-2051A-US);

[0037] 2021년 10월 6일자로 출원되고 발명의 명칭이 "PROTEIN STRUCTURE-BASED PROTEIN LANGUAGE MODELS"인 미국 임시 특허 출원 제63/253,122호(대리인 문서 번호 ILLM 1050-1/IP-2164-PRV);

[0038] 2021년 11월 19일자로 출원되고 발명의 명칭이 "PREDICTING VARIANT PATHOGENICITY FROM EVOLUTIONARY CONSERVATION USING THREE-DIMENSIONAL (3D) PROTEIN STRUCTURE VOXELS"인 미국 임시 특허 출원 제63/281,579호(대리인 문서 번호 ILLM 1060-1/IP-2270-PRV);

[0039] 2021년 11월 19일자로 출원되고 발명의 명칭이 "COMBINED AND TRANSFER LEARNING OF A VARIANT PATHOGENICITY

PREDICTOR USING GAPED AND NON-GAPED PROTEIN SAMPLES"인 미국 임시 특허 출원 제63/281,592호(대리인 문서 번호 ILLM 1061-1/IP-2271-PRV);

[0040] 2021년 12월 29일자로 출원되고 발명의 명칭이 "PERIODIC MASK PATTERN FOR REVELATION LANGUAGE MODELS"인 미국 임시 특허 출원 제63/294,813호(대리인 문서 번호 ILLM 1063-1/IP-2296-PRV);

[0041] 2021년 12월 29일자로 출원되고 발명의 명칭이 "CLASSIFYING MILLIONS OF VARIANTS OF UNCERTAIN SIGNIFICANCE USING PRIMATE SEQUENCING AND DEEP LEARNING"인 미국 임시 특허 출원 제63/294,816호(대리인 문서 번호 ILLM 1064-1/IP-2297-PRV);

[0042] 2021년 12월 29일자로 출원되고 발명의 명칭이 "IDENTIFYING GENES WITH DIFFERENTIAL SELECTIVE CONSTRAINT BETWEEN HUMANS AND NON-HUMAN PRIMATES"인 미국 임시 특허 출원 제63/294,820호(대리인 문서 번호 ILLM 1065-1/IP-2298-PRV);

[0043] 2021년 12월 29일자로 출원되고 발명의 명칭이 "DEEP LEARNING NETWORK FOR EVOLUTIONARY CONSERVATION"인 미국 임시 특허 출원 제63/294,827호(대리인 문서 번호 ILLM 1066-1/IP-2299-PRV);

[0044] 2021년 12월 29일자로 출원되고 발명의 명칭이 "INTER-MODEL PREDICTION SCORE RECALIBRATION"인 미국 임시 특허 출원 제63/294,828호(대리인 문서 번호 ILLM 1067-1/IP-2301-PRV); 및

[0045] 2021년 12월 29일자로 출원되고 발명의 명칭이 "SPECIES-DIFFERENTIABLE EVOLUTIONARY PROFILES"인 미국 임시 특허 출원 제63/294,830호(대리인 문서 번호 ILLM 1068-1/IP-2302-PRV).

배 경 기 술

[0046] 이 섹션에서 논의되는 주제는 단지 이 섹션 내에서의 이의 언급의 결과로서 종래 기술이라고 가정되어서는 안 된다. 유사하게, 이 섹션에서 언급되거나 배경기술로서 제공되는 주제와 연관된 문제는 종래 기술에서 이전에 인식되었다고 가정되어서는 안 된다. 이 섹션에서의 주제는 단지 상이한 접근법을 표현할 뿐이며, 그 접근법 자체는 청구되는 기술의 구현예에 또한 대응할 수 있다.

[0047] 광범위한 의미에서, 기능 유전체학으로도 지칭되는 유전체학은 게놈 서열 분석, 전사체 프로파일링 및 단백질체학과 같은 게놈 스케일 분석을 사용하여 유기체의 모든 게놈 요소의 기능을 특성화하는 것을 목표로 한다. 유전체학은 데이터 중심(data-driven) 과학으로서 발생하였으며 그것은 선입견이 있는 모델 및 가설을 검정하기보다는 게놈 스케일 데이터의 탐구로부터 신규 속성을 발견함으로써 작동한다. 유전체학의 응용은 유전자형과 표현형 사이의 연관성을 찾는 것, 환자 계층화에 대한 바이오마커를 발견하는 것, 유전자의 기능을 예측하는 것 및 전사 생물 통계학적 분석을 사용하여 인핸서(transcriptional enhancer) 및 단일 염기 다형성(SNP: single nucleotide polymorphism)과 같은 생물화학적 활성 게놈 영역 및 잔류물을 차트화하는 것을 포함한다.

[0048] 유전체학 데이터는 쌍별 상관(pairwise correlation)의 시각적 연구만으로 조사하기에는 너무 크고 너무 복잡하다. 대신, 예상되지 않은 관계의 발견을 지원하기 위해, 신규한 가설 및 모델을 도출하기 위해, 그리고 예측을 행하기 위해 분석 툴이 요구된다. 가정 및 도메인 전문 지식이 하드 코딩되는 일부 알고리즘과는 달리, 기계 학습 알고리즘은 데이터에서 패턴을 자동으로 검출하도록 설계된다. 따라서, 기계 학습 알고리즘은 데이터 중심 과학, 특히 유전체학에 적합하다. 그러나, 기계 학습 알고리즘의 성능은 데이터가 표현되는 방법, 즉 각각의 변수(특징부로도 불림)가 계산되는 방법에 강하게 의존할 수 있다. 예를 들어, 형광 현미경 이미지로부터 종양을 악성 또는 양성으로 분류하기 위해, 전처리 알고리즘이 세포를 검출할 수 있고, 세포 유형을 식별할 수 있고, 각각의 세포 유형에 대한 세포 카운트의 목록을 생성할 수 있다.

[0049] 기계 학습 모델은 추정된 세포 카운트를 취할 수 있으며, 이러한 카운트는 종양을 분류하기 위한 입력 특징부로서, 수작업으로 작성된 특징부의 예이다. 중심 문제는 분류 성능이 이러한 특징부의 품질 및 관련성에 크게 의존한다는 것이다. 예를 들어, 세포 형태학, 세포들 사이의 거리 또는 기관 내의 국지성과 같은 관련 시각적 특징부는 세포 카운트에서 캡처되지 않고, 데이터의 이러한 불완전한 표현은 분류 정확도를 감소시킬 수 있다.

[0050] 기계 학습의 하위 구분인 심층 학습은 기계 학습 모델 자체에 특징부의 계산을 임베딩하여 엔드-투-엔드(end-to-end) 모델을 산출함으로써 이러한 문제를 다룬다. 이러한 결과는 심층 신경망, 즉 연속적인 기본 동작을 포함하는 기계 학습 모델의 개발을 통해 실현되었으며, 이들은 선행 동작의 결과를 입력으로서 취함으로써 점점 더 복잡한 특징부를 계산한다. 심층 신경망은 위의 예에서 세포의 세포 형태학 및 공간 구성과 같은 높은 복잡도의 관련 특징부를 발견함으로써 예측 정확도를 개선할 수 있다. 심층 신경망의 구성 및 훈련은 데이터의

폭증, 알고리즘 진보 및 계산 용량의 실질적인 증가에 의해, 특히 그래픽 처리 장치(GPU: graphical processing unit)의 사용을 통해 가능하게 되었다.

[0051] 지도 학습(supervised learning)의 목표는, 특징부를 입력으로서 취하고 이른바 표적 변수에 대한 예측을 반환하는 모델을 획득하는 것이다. 지도 학습 문제의 한 예는 정규(canonical) 스플라이스 부위 서열의 존재 여부 및 스플라이싱 분기점의 위치 또는 인트론 길이와 같은 RNA 상의 특징부를 고려하여 인트론이 스플라이스-아웃(splice out)되는지의 여부를 예측하는 것(표적)이다. 기계 학습 모델을 훈련시키는 것은 이의 매개변수를 학습하는 것을 지칭하며, 이는 보통 보이지 않은 데이터에 대한 정확한 예측을 행하는 목적으로 훈련 데이터에 대한 손실 함수를 최소화하는 것을 수반한다.

[0052] 계산 생물학에서의 많은 지도 학습 문제의 경우, 입력 데이터는 예측을 행하는 데 잠재적으로 유용한 수치 또는 카테고리 데이터를 각각 포함하는 다수의 열 또는 특징부를 갖는 표로서 표현될 수 있다. 일부 입력 데이터는 (온도 또는 시간과 같이) 표 내의 특징부로서 자연적으로 표현되는 반면, 다른 입력 데이터는 (*k*-량체 카운트로의 데옥시리보핵산(DNA) 서열과 같이) 표로 나타낸 표현에 피팅하기 위해 특징부 추출로 불리는 프로세스를 사용하여 먼저 변환될 필요가 있다. 인트론 스플라이싱 예측 문제의 경우, 표준 스플라이스 부위 서열의 존재 여부, 스플라이싱 분기점의 위치 및 인트론 길이는 표로 나타낸 포맷으로 수집된 전처리된 특징부일 수 있다. 표로 나타낸 데이터는, 로지스틱 회귀(logistic regression)와 같은 단순한 선형 모델로부터 신경망과 같은 더 유연한 비선형 모델에 이르기는 광범위한 지도 기계 학습 모델 및 많은 다른 것에 대한 표준이다.

[0053] 로지스틱 회귀는 이진 분류자, 즉 이진 표적 변수를 예측하는 지도 학습 모델이다. 구체적으로, 로지스틱 회귀는 한 유형의 활성화 함수인 시그모이드 함수를 사용하여 [0, 1] 구간에 맵핑된 입력 특징부의 가중 합을 계산함으로써 포지티브 클래스의 확률을 예측한다. 로지스틱 회귀의 매개변수 또는 상이한 활성화 함수를 사용하는 다른 선형 분류자는 가중 합에서의 가중치이다. 선형 분류자는 클래스, 예를 들어 스플라이스-아웃되거나 스플라이스-아웃되지 않은 인트론의 클래스가 입력 특징부의 가중 합으로 잘 구별될 수 없을 때 실패한다. 예측 성능을 개선하기 위해, 예를 들어, 거듭제곱 또는 쌍별 곱을 취함으로써 새로운 방식으로 기존의 특징부를 변환 또는 조합함으로써 새로운 입력 특징부가 수동으로 추가될 수 있다.

[0054] 신경망은 은닉 계층들을 사용하여 이러한 비선형 특징부 변환을 자동으로 학습한다. 각각의 은닉 계층은 시그모이드 함수 또는 더 대중적인 정류 선형 유닛(ReLU: rectified-linear unit)과 같은 비선형 활성화 함수에 의해 변환된 출력을 갖는 다수의 선형 모델로서 생각될 수 있다. 이러한 계층들은 함께 입력 특징부를 관련된 복잡한 패턴으로 구성하며, 이는 2개의 클래스를 구별하는 태스크를 용이하게 한다.

[0055] 심층 신경망은 많은 은닉 계층들을 사용하고, 계층은 각각의 뉴런이 선행 계층의 모든 뉴런으로부터 입력을 수신할 때 완전 연결된 것으로 간주된다. 신경망은 일반적으로, 매우 큰 데이터 세트에 대해 모델을 훈련시키는 데 적합한 알고리즘인 확률론적 경사 하강법(stochastic gradient descent)을 사용하여 훈련된다. 최신 심층 학습 프레임워크를 사용한 신경망의 구현예는 상이한 아키텍처 및 데이터 세트로 신속한 프로토타이핑을 가능하게 한다. 완전 접속 신경망은 다수의 유전체학 응용예에 사용될 수 있으며, 이는 서열 보존 또는 스플라이스 인자의 결합 모티프의 존재와 같은 서열 특징부로부터의 주어진 서열에 대해 스플라이스-인(splice in)된 엑손의 백분율을 예측하는 것; 잠재적인 질환 유발 유전자 변이체를 우선 순위화하는 것; 및 염색질 마크, 유전자 발현 및 진화적 보존과 같은 특징부를 사용하여 주어진 게놈 영역 내의 *cis*-조절 요소를 예측하는 것을 포함한다.

[0056] 효과적인 예측을 위해 공간적 및 종방향 데이터에서의 로컬 종속성이 고려되어야 한다. 예를 들어, DNA 서열 또는 이미지의 픽셀을 셔플링하는 것은 정보성 패턴을 심하게 파괴한다. 이러한 로컬 종속성은 표로 나타낸 데이터와는 분리된 공간적 또는 종방향 데이터를 설정하며, 이를 위한 특징부의 순서화는 임의적이다. 특정 전사 인자에 의해 게놈 영역을 결합(bound) 대 비결합(unbound)으로 분류하는 문제를 고려한다 - 이때 결합 영역은 서열 분석(ChIP-seq) 데이터가 뒤에 오는 염색질 면역 침전(chromatin immunoprecipitation)에서의 고신뢰 결합 이벤트로서 정의된다. 서열 모티프를 인식함으로써 전사 인자가 DNA에 결합된다. 서열 내의 *k*-량체 인스턴스(instance)의 수 또는 위치 가중치 행렬(PWM: position weight matrix) 매칭과 같은 서열 파생 특징부에 기초한 완전 연결 계층(fully-connected layer)이 이러한 태스크에 사용될 수 있다. 따라서, *k*-량체 또는 PWM 인스턴스 빈도는 서열 내에서 모티프를 시프트하는 것에 강건하기 때문에, 이러한 모델은 상이한 위치에 위치한 동일한 모티프를 갖는 서열에 대해 양호하게 일반화될 수 있다. 그러나, 이는 전사 인자 결합이 잘 정의된 간격을 갖는 다수의 모티프의 조합에 의존하는 패턴을 인식하지 못할 것이다. 또한, 가능한 *k*-량체의 수는 *k*-량체 길이에 따라 기하급수적으로 증가하며, 이는 저장 및 오버피팅 문제를 모두 제기한다.

[0057] 컨볼루션 계층(convolutional layer)은 완전 연결 계층의 특수 형태이며, 동일한 완전 연결 계층은, 예를 들어,

6 bp 윈도우에서, 모든 서열 위치에 국부적으로 적용된다. 이러한 접근법은 또한, 예를 들어, 전사 인자 GATA1 및 TAL1에 대해, 다수의 PWM을 사용하여 서열을 스캐닝하는 것으로 보일 수 있다. 위치들에 걸쳐 동일한 모델 매개변수를 사용함으로써, 매개변수의 총 수는 급격히 감소되고, 네트워크는 훈련 동안 보이지 않는 위치에서 모티프를 검출할 수 있다. 각각의 컨볼루션 계층은 모든 위치에서 스칼라 값을 생성함으로써 여러 필터로 서열을 스캐닝하며, 이는 필터와 서열 사이의 매칭을 정량화한다. 완전 연결 신경망에서와 같이, 비선형 활성화 함수(일반적으로, ReLU)가 각각의 계층에 적용된다. 다음으로, 풀링(pooling) 동작이 적용되며, 이는 위치 축에 걸친 인접 빈(bin)에서의 활성화를 집약하여, 일반적으로, 각각의 채널에 대해 최대 또는 평균 활성화를 취한다. 풀링은 유효 서열 길이를 감소시키고, 신호를 덜 세밀하게 만든다(coarsen). 후속 컨볼루션 계층은 이전 계층의 출력을 조정하며, GATA1 모티프 및 TAL1 모티프가 일부 거리 범위에 존재하였는지의 여부를 검출할 수 있다. 마지막으로, 컨볼루션 계층의 출력은 최종 예측 태스크를 수행하기 위해 완전 접속 신경망에 대한 입력으로서 사용될 수 있다. 따라서, 상이한 유형의 신경망 계층(예를 들어, 완전 연결 계층 및 컨볼루션 계층)이 단일 신경망 내에서 조합될 수 있다.

[0058] 컨볼루션 신경망(CNN: convolutional neural network)은 DNA 서열 하나에만 기초하여 다양한 분자 표현형을 예측할 수 있다. 응용예는 전사 인자 결합 부위를 분류하는 것과, 염색질 특징부, DNA 접촉 맵, DNA 메틸화, 유전자 발현, 번역 효율, RBP 결합 및 마이크로RNA(miRNA) 표적과 같은 분자 표현형을 예측하는 것을 포함한다. 서열로부터 분자 표현형을 예측하는 것에 더하여, 컨볼루션 신경망은 수작업으로 작성된 생물정보학 파이프라인에 의해 전통적으로 다루어진 더 많은 기술적 태스크에 적용될 수 있다. 예를 들어, 컨볼루션 신경망은 가이드 RNA의 특이성을 예측할 수 있고, ChIP-seq를 잡음 제거할 수 있고, Hi-C 데이터 해상도를 향상시킬 수 있고, 근원이 DNA 서열로부터인 실험을 예측할 수 있고, 유전자 변이체를 호출할 수 있다. 컨볼루션 신경망은 또한 게놈에서 장거리 종속성을 모델링하기 위해 채용되었다. 상호 작용하는 조절 요소가 전개된 선형 DNA 서열 상에서 원거리에 위치될 수 있지만, 이러한 요소는 종종 실제 3D 염색질 형태에서 근위에 있다. 따라서, 선형 DNA 서열로부터 분자 표현형을 모델링하는 것은, 염색질의 대충의 근사화에도 불구하고, 장거리 종속성을 허용하고 모델이 프로모터-인핸서 루핑과 같은 3D 구성의 양태를 암시적으로 학습할 수 있게 함으로써 개선될 수 있다. 이것은 최대 32 kb의 수용 필드를 갖는 확장된 컨볼루션을 사용하여 달성된다. 확장된 컨볼루션은 또한 스플라이스 부위가 10 kb의 수용 필드를 사용하여 서열로부터 예측될 수 있도록 하며, 이로 인해, 전형적인 인간 인트론만큼 긴 거리에 걸쳐 유전자 서열의 통합이 가능하게 한다(문헌[Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. Cell 176, 535--548 (2019)] 참조).

[0059] 상이한 유형의 신경망들은 이들의 매개변수 공유 스킴에 의해 특성화될 수 있다. 예를 들어, 완전 연결 계층들은 매개변수 공유를 갖지 않는 반면, 컨볼루션 계층들은 이들의 입력의 모든 위치에서 동일한 필터를 적용함으로써 번역 불변성을 부과한다. 순환 신경망(RNN: recurrent neural network)은 상이한 매개변수 공유 스킴을 구현하는, DNA 서열 또는 시계열과 같은 순차적 데이터를 처리하기 위한 컨볼루션 신경망에 대한 대안이다. 순환 신경망은 각각의 서열 요소에 동일한 동작을 적용한다. 동작은 이전 서열 요소의 메모리 및 새로운 입력을 입력으로서 취한다. 이는 메모리를 업데이트하고, 선택적으로는 후속 계층으로 전달되거나 모델 예측으로서 직접 사용되는 출력을 방출한다. 각각의 서열 요소에 동일한 모델을 적용함으로써, 순환 신경망은 처리된 서열에서 위치 인덱스에 대해 불변이다. 예를 들어, 순환 신경망은 서열 내의 위치에 관계없이 DNA 서열에서 개방 판독 프레임을 검출할 수 있다. 이러한 태스크는 인프레임(in-frame) 정지 코돈이 이어지는 시작 코돈과 같은 특정 입력 시리즈의 인식을 요구한다.

[0060] 컨볼루션 신경망에 비해 순환 신경망의 주요 이점은, 이론적으로, 메모리를 통해 무한히 긴 서열을 거쳐 정보를 전달할 수 있다는 것이다. 또한, 순환 신경망은 mRNA 서열과 같은 광범위하게 변화하는 길이의 서열을 자연적으로 처리할 수 있다. 그러나, 다양한 트릭(예를 들어, 확장된 컨볼루션)과 조합된 컨볼루션 신경망은 오디오 합성 및 기계 번역과 같은 서열 모델링 태스크에 대해 순환 신경망과 유사하거나 심지어 그보다 더 양호한 성능에 도달할 수 있다. 순환 신경망은 단일 세포 DNA 메틸화 상태, RBP 결합, 전사 인자 결합 및 DNA 접근성을 예측하기 위한 컨볼루션 신경망의 출력을 집계할 수 있다. 더욱이, 순환 신경망이 순차적인 동작을 적용하기 때문에, 이는 쉽게 병렬화될 수 없고, 따라서, 컨볼루션 신경망보다 계산하는 데 훨씬 더 느리다.

[0061] 각각의 인간은 고유한 유전자 코드를 갖지만, 인간 유전자 코드의 대부분은 모든 인간에 대해 공통적이다. 일부 경우에 있어서, 인간 유전자 코드는 비교적 작은 그룹의 인간 모집단의 개체들 사이에서 공통적일 수 있는 유전자 변이체로 불리는 이상치(outlier)를 포함할 수 있다. 예를 들어, 특정 인간 단백질은 특정 서열의 아미노산을 포함할 수 있는 반면, 그 단백질의 변이체는 그 외의 동일한 특정 서열 내의 하나의 아미노산만큼 상이할 수 있다.

[0062] 유전자 변이체는 병원성이어서, 질환으로 이어질 수 있다. 이러한 유전자 변이체의 대부분이 자연적인 선택에 의해 게놈으로부터 대폭 감소되었지만, 어느 유전자 변이체가 병원성일 가능성이 있는지를 식별하는 능력은 연구자가 이러한 유전자 변이체에 초점을 맞추어 대응하는 질환 및 이의 진단, 처치 또는 치유에 대한 이해를 얻는 데 도움이 될 수 있다. 수백만 개의 인간 유전자 변이체의 임상적 해석은 여전히 불명확하다. 가장 빈번한 병원성 변이체 중 일부는 단백질의 아미노산을 변화시키는 단일 뉴클레오티드 미스센스(missense) 돌연변이이다. 그러나, 모든 미스센스 돌연변이가 병원성인 것은 아니다.

[0063] 생물학적 서열로부터 직접적으로 분자 표현형을 예측할 수 있는 모델은 유전자 변이체와 표현형 변이체 사이의 연관성을 프로브하기 위해 컴퓨터 내(in silico) 섭동 툴로서 사용될 수 있고, 정량적 형질 유전자좌(quantitative trait loci) 식별 및 변이체 우선 순위화를 위한 새로운 방법으로서 부상하였다. 복잡한 표현형의 전장 유전체 연관성 분석(genome-wide association study)에 의해 식별된 변이체의 대부분이 비-코딩이라면, 이러한 접근법은 매우 중요하며, 이는 표현형에 대한 이의 효과 및 기여도를 추정하는 것을 어렵게 만든다. 더욱이, 연관 불균형(linkage disequilibrium)은 변이체의 블록이 동시-유전되는 결과를 초래하며, 이는 개별 인과 변이체를 정확하게 찾아내는 것에 어려움을 야기한다. 따라서, 이러한 변이체의 영향을 평가하기 위한 심문 툴로서 사용될 수 있는 서열 기반 심층 학습 모델은 복잡한 표현형의 잠재적인 드라이버를 찾는 것에 대한 유망한 접근법을 제공한다. 하나의 예는 전사 인자 결합, 염색질 접근성 또는 유전자 발현 예측의 측면에서 2개의 변이체 사이의 차이로부터 간접적으로 짧은 삽입 또는 결실(인델) 및 비-코딩 단일 뉴클레오티드 변이체의 효과를 예측하는 것을 포함한다. 다른 예는 스플라이싱에 대한 유전자 변이체의 서열 또는 정량적 효과로부터 신규한 스플라이스 부위 생성을 예측하는 것을 포함한다.

[0064] 변이체 효과 예측을 위한 엔드-투-엔드 심층 학습 접근법은 서열 보존 데이터 및 단백질 서열로부터의 미스센스 변이체의 병원성을 예측하기 위해 적용된다(본원에서 "PrimateAI"로 지칭되는 문헌[Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* 50, 1161-1170 (2018)] 참조). PrimateAI는 종간(cross-species) 정보를 사용한 데이터 증강에 의해 공지된 병원성의 변이체에 대해 훈련된 심층 신경망을 사용한다. 특히, PrimateAI는 차이를 비교하고 훈련된 심층 신경망을 사용하여 돌연변이의 병원성을 결정하기 위해 야생형 및 변종 단백질의 서열을 사용한다. 병원성 예측을 위해 단백질 서열을 활용하는 이러한 접근법은, 환상성(circularity) 문제 및 이전 지식에 대한 과적합화를 회피할 수 있기 때문에 유망하다. 그러나, 심층 신경망을 효과적으로 훈련시키기 위한 적절한 수의 데이터와 비교하면, ClinVar에서 이용 가능한 임상 데이터의 수는 비교적 작다. 이러한 데이터 부족을 극복하기 위해, PrimateAI는 공통적인 인간 변이체 및 영장류로부터의 변이체를 양성 데이터로서 사용하지만, 트리뉴클레오티드 콘텍스트에 기초한 시뮬레이션된 변이체가 라벨링되지 않은 데이터로서 사용되었다.

[0065] PrimateAI는 서열 정렬에 대해 직접적으로 훈련될 때 이전 방법을 능가한다. PrimateAI는 약 120,000개의 인간 샘플로 이루어진 훈련 데이터로부터 직접적으로 중요한 단백질 도메인, 보존 아미노산 위치 및 서열 종속성을 학습한다. PrimateAI는 후보 발달 장애 유전자에서 양성 및 병원성 신생 돌연변이를 구별하고 ClinVar에서 이전 지식을 재생하는 데 있어서 다른 변이체 병원성 예측 툴의 성능을 실질적으로 초과한다. 이러한 결과는 PrimateAI가 이전 지식에 대한 임상 보고의 의존을 줄일 수 있는 변이체 분류 툴을 위해 중요한 진전임을 시사한다.

[0066] 구조 요소가 관찰된 기능을 발생시키는 방법에 대한 이해는 단백질 생물학에 대해 중심이 된다. 방대한 단백질 구조 데이터는 구조적-기능적 관계를 지배하는 규칙을 체계적으로 도출하기 위한 계산 방법의 개발을 가능하게 한다. 그러나, 이러한 방법의 성능은 단백질 구조 표현의 선택에 중대하게 의존한다.

[0067] 단백질 부위는 이의 구조적 또는 기능적 역할에 의해 구별되는 단백질 구조 내의 미세 환경이다. 부위는 3차원(3D) 위치 및 구조 또는 기능이 존재하는 이러한 위치 주위의 국부적 이웃에 의해 정의될 수 있다. 아미노산의 구조적 배열이 단백질 부위 내에서 기능적 특성을 생성하는 방법에 대한 이해는 합리적인 단백질 공학에 대해 중심이 된다. 단백질 내의 개별 아미노산의 구조적 및 기능적 역할의 결정은 공학자를 돕고 단백질 기능을 변경하기 위한 정보를 제공한다. 기능적으로 또는 구조적으로 중요한 아미노산을 식별하는 것은 표적화된 단백질 기능적 속성을 변경하기 위한 부위 유도 돌연변이 유발과 같은 집중된 공학 노고를 허용한다. 대안적으로, 이러한 지식은 원하는 기능을 무효화할 공학 설계를 회피하는 데 도움이 될 수 있다.

[0068] 구조가 서열보다 훨씬 더 많이 보존된다는 것이 확립되었기 때문에, 단백질 구조 데이터에서의 증가는 데이터 중심 접근법을 사용하여 구조적-기능적 관계를 지배하는 기본 패턴을 체계적으로 연구할 기회를 제공한다. 임의의 계산 단백질 분석의 기본 양태는 단백질 구조 정보가 표현되는 방법이다. 기계 학습 방법의 성능은 종종 채

용된 기계 학습 알고리즘보다 데이터 표현의 선택에 더 많이 의존한다. 양호한 표현은 가장 중대한 정보를 효율적으로 캡처하는 반면, 불량한 표현은 기본 패턴이 없는 잡음 분포를 생성한다.

[0069] 방대한 단백질 구조 및 심층 학습 알고리즘의 최근의 성공은 단백질 구조의 태스크 특정적 표현을 자동으로 추출하기 위한 툴을 개발할 기회를 제공한다.

[0070] 인간 게놈에 있는 7천만 개 이상의 가능한 미스센스 변이체 중, 대다수는 임상적 유의성이 알려지지 않았으며 약 0.1%만이 임상적 변이체 데이터베이스에 주석이 달려있다. 일반적인 질병에서 희귀 침투 변이체의 역할을 이해할 수 있는 기회가 생긴다. 양성 결과를 초래하는 변이체로부터 유해 변이체를 정확하게 구별하는 것은 정밀 의학 및 표적 약물 개발 모두에 도움이 될 수 있다.

도면의 간단한 설명

[0071] 특허 또는 출원 파일은 컬러로 작성된 적어도 하나의 도면을 포함한다. 컬러 도면(들)을 갖는 이러한 특허 또는 특허 출원 공보의 사본은 요청 및 필요한 수수료의 지불에 따라 특허청(Office)에 의해 제공될 것이다. 컬러 도면은, 또한, Supplemental Content 탭을 통해 PAIR에서 입수 가능할 수 있다.

도면에서, 유사한 도면 부호는 일반적으로 상이한 도면 전체에 걸쳐서 유사한 부분을 지칭한다. 또한, 도면은 반드시 일정한 비율로 표시될 필요는 없으며, 대신 개시된 기술의 원리를 예시하는 데 일반적으로 중점을 두고 있다. 다음 설명에서, 개시된 기술의 다양한 구현예가 다음의 도면들을 참조하여 설명된다.

도 1a는 가중된 희귀 변이체 PRS를 결정하는 방법을 예시하는 개략도이다.

도 1b는 특정 복수의 유전자 및 특정 표현형에 대한 희귀 변이체 다유전자 위험 점수의 예시적인 계산을 예시한다.

도 2는 개시된 기술을 구현하는 데 사용될 수 있는 예시적인 컴퓨터 시스템을 도시한다.

도 3은 유전자 변이체의 한 예를 예시한다.

도 4는 유전자형에서의 변화에 따른 표현형 효과를 예시하는 한 예이다.

도 5는 심혈관 질환 환자 X에 대응하는 복수의 표현형 발현을 예시한다.

도 6은 공통 변이체 대 희귀 변이체에 대해 전장 유전체 연관성 분석을 그래프로 대조한다.

도 7은 각각 유전자 분해 시 특정 개별 공통 변이체 또는 특정 집계 희귀 변이체에 대한 유전자 연관성 검정을 예시한다.

도 8은 희귀 변이체 부담 검정을 최적화하기 위한 방법의 개략도이다.

도 9는 변이체의 병원성을 결정하기 위한 시스템의 프로세스를 예시하는 흐름도이다.

도 10은 개시된 기술의 하나의 구현예에 따른 병원성 분류자의 예시적인 처리 아키텍처를 도시한다.

도 11은 개시된 기술을 구현하는 데 사용될 수 있는 예시적인 컴퓨터 시스템을 도시한다.

도 12는 최종 병원성 점수를 결정하는 하나의 구현예를 예시한다.

도 13은 각각의 유전자 내의 다중 검정을 보정하기 위한 프로세스의 흐름도이다.

도 14는 공변량에 대한 표현형 값을 보정하는 방법에 대한 개략도이다.

도 15는 복수의 표현형에 대한 복수의 약물 사용에 따라 표현형 시프트를 예측하는 것을 나타내는 개략도이다.

도 16은 특정 코호트에 대한 약물 사용 패턴 및 표현형 데이터를 획득하기 위한 실험 설정을 나타내는 개략도이다.

도 17은 심혈관 질환 환자 X에 대응하는 복수의 표현형을 예시한다.

도 18은 다양한 유형의 부담 검정에 대해 식별된 유의한 유전자-표현형 쌍의 총 수를 정량화하는 그래프이다.

도 19는 병원성 분류자 PrimateAI-3D에 의해 식별된 질병 중증도 및 발병 연령에 영향을 미치는 희귀 유해 변이체를 예시하는 그래프 모음이다.

도 20은 다양한 병원성 점수와 표현형 값의 평균 절대 Spearman 상관 관계의 그래프이다.

도 21은 희귀 유해 변이체와 공통 전장 유전체 연관성 분석 변이체의 히트맵 비교이다.

도 22는 희귀 유해 변이체와 공통 전장 유전체 연관성 분석 변이체의 추가 비교이다.

도 23은 영국 Biobank 코호트 내의 모든 개체에 걸친 콜레스테롤 경로와 총 콜레스테롤 분포를 예시한다.

도 24는 희귀 변이체 PRS 성능의 측정값을 포함한다.

도 25는 PRS 이상치의 농축(enrichment)의 그래프이다.

도 26은 정량적 표현형에 대한 PRS 이상치의 그래프이다.

도 27은 2개의 개별 코호트로부터의 정규화된 총 콜레스테롤 분포의 그래프를 포함한다.

도 28은 2개의 개별 코호트 사이에서 희귀 변이체 PRS 이상치와 표현형을 비교하는 그래프를 포함한다.

도 29는 인종별 성과 결과를 예시하는 그래프를 포함한다.

도 30은 정상 콜레스테롤 레벨에 사용되는 공통 PRS 변이체와 희귀 PRS 유전자에 대한 효과 크기 및 빈도를 비교한 표이다.

도 31은 희귀 미스센스 및 LoF 변이체에 대한 농축에 기초하여 선택된 34개 유전자-표현형 쌍 세트에 대한 다양한 병원성 스코어링 방법에 의해 설명되는 표현형 분산의 평균 분율의 그래프이다.

도 32는 정량적 표현형과 임상적 표현형 사이의 모든 쌍별 비교를 위한 GWAS 유전자에서의 희귀 변이체의 농축에 대한 히트맵이다.

도 33은 동일한 유전자에 대한 단일 LoF 변이체의 평균 효과 크기와 가장 유의한 GWAS 변이체의 효과 크기 사이의 비율의 절대값에 대한 분포 그래프를 포함한다.

도 34는 5%의 오류 발견율에서 유의미한 유전자-표현형 쌍으로부터의 임의의 유전자에서 개체당 변이체의 수를 도시한다.

도 35는 훈련 데이터 대 검정 데이터 분할에서의 효과 크기의 비교를 도시한다.

도 36은 공통 변이체와 희귀 변이체 PRS 서브세트를 비교하는 그래프를 포함한다.

발명을 실시하기 위한 구체적인 내용

[0072] 아래의 논의는 어느 당업자라도 개시된 기술을 제조하고 사용할 수 있게 하도록 제시되며, 특정의 응용예 및 이의 요건과 관련하여 제공된다. 개시된 구현예에 대한 다양한 수정은 당업자에게 용이하게 명백할 것이며, 본원에 정의된 일반 원리는 개시된 기술의 사상 및 범위로부터 벗어나지 않으면서 다른 구현예 및 응용예에 적용될 수 있다. 따라서, 개시된 기술은 도시된 구현예로 제한되도록 의도된 것이 아니라, 본원에 개시된 원리 및 특징과 일치하는 가장 넓은 범위에 부합되어야 한다.

[0073] 다양한 구현예에 대한 상세한 설명은 첨부된 도면과 함께 읽을 때 더 잘 이해될 것이다. 도면이 다양한 구현예의 기능 블록의 다이어그램을 예시하는 범위에서, 기능 블록은 반드시 하드웨어 회로부 사이의 분할을 나타내는 것은 아니다. 따라서, 예를 들어, 하나 이상의 기능 블록(예를 들어, 모듈, 프로세서 또는 메모리)은 단일 하드웨어(예를 들어, 범용 신호 프로세서 또는 랜덤 액세스 메모리의 블록, 하드 디스크 등) 또는 다수의 하드웨어에서 구현될 수 있다. 유사하게, 프로그램은 독립형 프로그램일 수 있고, 운영 체제에서 서브루틴으로서 통합될 수 있으며, 설치된 소프트웨어 패키지에서의 기능일 수 있는 등이다. 다양한 구현예가 도면에 도시된 배열 및 수단으로 제한되지 않는다는 것이 이해될 것이다.

[0074] 모듈로서 지정된 도면의 처리 엔진 및 데이터베이스는 하드웨어 또는 소프트웨어로 구현될 수 있고, 도면에 도시된 바와 같이 정확하게 동일한 블록으로 분할될 필요는 없다. 모듈 중 일부는 또한 상이한 프로세서, 컴퓨터 또는 서버 상에서 구현될 수 있거나, 다수의 상이한 프로세서, 컴퓨터 또는 서버 사이에 분산될 수 있다. 또한, 모듈 중 일부가, 달성된 기능에 영향을 주지 않고서 도면에 도시된 것과 조합되거나, 이와 병렬로 또는 상이한 순서로 동작될 수 있다는 것이 이해될 것이다. 도면 내의 모듈은 또한 방법에서의 흐름도 단계로서 생각될 수 있다. 모듈은 또한 이의 코드 전부가 반드시 메모리에 인접하게 배치될 필요가 없고; 코드의 일부 부분은 코드의 다른 부분과는 분리될 수 있으며, 이때 다른 모듈 또는 다른 기능으로부터의 코드가 사이에 배치된다.

[0075] 개시된 기술은 희귀 변이체에 대한 다유전자 위험 스코어링의 품질을 향상시키는 데 사용될 수 있다. 개시된 기술은 심각한 조기 발병 질병에 대한 위험이 가장 크고 임상적 개입으로 가장 큰 혜택을 받는 환자를 표현형 스펙트럼의 극단에서 식별하는 데 사용될 수 있다. 이러한 개시내용을 제시하는 연구자들은 심각한 조기 발병 질병과 연관된 이상치 표현형이 작은 효과 크기를 갖는 많은 공통 변이체의 집합적 작용보다는 희귀 침투 변이체에 의해 더 잘 설명된다는 것을 보여준다. 여러 세대 동안 자연 선택에 의해 유해한 결과가 주로 대폭 감소된 공통 변이체와는 달리, 희귀 변이체는 대부분 필터링되지 않으며, 복잡한 형질(trait)과 질병에 높은 침투 효과를 발휘할 가능성을 보존한다. 개시된 기술은 표현형 반응에 대한 희귀 병원성 변이체의 기여도의 정량화를 위한 새로운 가중 합 모델을 포함한다.

[0076] **서론**

[0077] 개시된 기술은, 다수의 표현형 연관 유전자로부터의 희귀 유해 변이체의 가중 합에 기초하여, 상보적인 희귀 변이체 다유전자 위험 점수 모델을 개발한다. 전장 유전체 연관성 분석(GWAS: genome-wide association study)은 유전적 돌연변이 또는 변이체를 지연 발달 장애(DDD: delayed development disorder)와 같은 유전 질환의 표현형 및 복잡한 형질과 연관시키기 위한 툴이다. GWAS 툴은 매우 희귀한 변이체보다는 공통 또는 준공통 변이체에서 더 잘 작동한다. 예를 들어, 영국 Biobank 코호트는 200,643개의 엑솜 및 연관된 환자 데이터를 포함한다. 유해한 것으로 간주되는 희귀 변이체 중 거의 절반이 데이터 세트 내의 한 개체에서만 나타났다. 대규모 코호트에서 희귀 변이체의 인스턴스가 하나만 있는 경우 전통적인 통계 분석 및 부담 검정은 실패한다.

[0078] 개별 전장 유전체 연관성 분석(GWAS: genome-wide association study) 변이체가 임상적 실행 가능성이 너무 가벼운 경향이 있는 효과를 부여하지만, 수백에서 수백만 개의 공통 변이체로부터의 신호를 조합한 다유전적 위험 점수는 질병에 대해 위험이 높은 표현형 극단에서 환자를 예측하는 데 상당한 효능을 입증했다. 그러나 기존의 공통 변이체 다유전자 위험 점수 모델은 유의성이 알려지지 않은 변이체를 해석하는 데 있어서의 어려움과 효과 크기 추정에서의 부정확성 때문에 희귀 변이체를 대체로 제외해 왔다. 기존 공통 변이체 다유전자 위험 점수 모델과 비교하여, 개시된 기술은 개체가 각각의 유전자에서 희귀 유해 변이체를 보유하고 있는지 여부에 기초하여 유전자 전반에 걸쳐 위험을 집계하도록 구성된 희귀 변이체 다유전자 위험 점수 모델을 포함한다.

[0079] 개체의 유전자형은 개체의 표현형(표현형은 개체의 하나 이상의 관찰 가능한 신체적 특성으로 정의됨)에 기여한다; 따라서 특정 변이체 또는 변이체 그룹을 가진 개체에게서 신체적 형질, 이상 또는 질병의 존재의 예측에 도움이 되도록 사용될 수 있는 유전자형과 표현형 사이의 통계적 상관 관계가 있다. 특정 유전 질환은 단일 유전자에서의 변이체로 인해 발생될 수 있지만(즉, 단일 유전자 유전 질환), 다수의 유전자에서의 변이체로 인해 발생되는 유전 질환(즉, 다유전자 유전 질환)이 더 흔하다. 따라서, 심각한 질병을 일으킬 가능성이 있는 희귀 유해 변이체에 대해 다유전적 위험 점수를 결정하기 위한 강력한 방법이 필요하다.

[0080] 일반 모집단의 건강한 개체를 대상으로 한 최근의 대규모 게놈 및 엑솜 서열 분석 연구는 평균적인 사람은 최근 돌연변이를 통해 발생한 수십 개의 잠재적으로 유해한 희귀 변이체를 보유하고 있는 것을 밝혀냈다. 여러 세대 동안 자연 선택에 의해 유해한 결과가 주로 대폭 감소된 공통 변이체와는 달리, 희귀 변이체는 대부분 필터링되지 않으며, 복잡한 형질(trait)과 질병에 높은 침투 효과를 발휘할 가능성을 보존한다. 200,643개의 영국 Biobank 엑솜의 공개 자료는, 변이체 병원성 예측의 정확성에서의 급속한 향상과 함께, 공통 인간 질병 및 복잡한 형질의 포괄적인 세트에 대한 희귀 침투성 변이체의 영향을 조사할 뿐만 아니라, 일반 모집단에 대한 사람 게놈 서열 분석의 잠재적 유용성을 엿볼 수 있는 기회를 제공한다.

[0081] 최근의 기하급수적인 인구 증가는, GWAS에서 식별된 공통 변이체와 대조적으로, 자연 선택이 유해한 결과를 갖는 것을 대폭 감소시킬 적절한 시간을 제공하지 못한 채 무작위로 발생하는 돌연변이를 통해 과다한 희귀 변이체를 생성해 왔다. 우리는 경미한 임상적 위험과 연관된 공통 변이체를 포함하는 각각의 GWAS 유전자좌에서 훨씬 더 큰 중증도를 갖는 희귀 유해 변이체도 또한 존재할 것이어서, 변이체의 중증도가 모집단에서의 이의 빈도와 반비례하는 대립 유전자 시리즈를 형성한다는 것을 검정하기 시작했다.

[0082] 희귀 유해 변이체의 높은 침투율에도 불구하고, 대다수의 개체는 주어진 표현형에 대해 희귀 유해 변이체를 보유하고 있지 않기 때문에, 이의 희귀성은 표현형 분산을 설명하는 능력을 모집단의 작은 일부로 제한한다. 따라서, 공통 변이체 PRS뿐만 아니라, 개시된 기술에 설명된 희귀 변이체 PRS의 하나의 구현예는 전체 모집단에 걸쳐 설명된 분산의 측면에서 약 1/20만 수행한다. 그러나, 표현형 극단에 있는 개체를 고려할 때, 이러한 경향은 반전되었다: 이상치 표현형(z-점수 3 이상)을 가진 개체는 기준 모집단에 비해 1번째 또는 99번째 백분위수에서 희귀 변이체 PRS 점수를 가질 가능성이 3배 더 높았으며, 공통 변이체 PRS의 경우의 1.8배에 비교된다.

[0083] 공통 변이체 PRS 모델의 임상적 채택에 대한 중요한 장벽은 서로 다른 조상을 가진 모집단 사이에 제한된 일반화 가능성을 가진다는 것이다. 이러한 문제는 연관 불균형으로 인해 공통 변이체 PRS 모델에서의 예측 변수(predictors)로서 질병과 연관되어 있지만 인과 관계가 없는 변이체의 통합으로부터 비롯된다. 인과 변이체의 효과가 코호트 사이에 일반화될 것으로 예상될 수 있지만, 실제 인과 변이체와 모델에 사용된 변이체 사이의 상관 관계가 보존될 것이라는 보장은 없고; 오히려 이러한 상관 관계는 모집단 조상에서의 차이뿐만 아니라 유전자형 분석 및 전가(imputation)에서의 기술적 아티팩트에 의해 영향을 받는다. 이에 비해, 희귀 변이체 PRS 모델은 희귀 유해 변이체를 예측 변수로 직접 사용하고, 이는 모집단에서의 희귀성과 최근 이력을 고려할 때, 공통 변이체에 대한 인과 관계를 상관 관계로부터 구별하기 어렵게 만드는 연관 불균형 문제에 의한 영향을 크게 받지 않는다. 오히려, 희귀 변이체 다유전자 위험을 예측하기 위한 과제는 VUS(variant of uncertain significance)의 효과를 정확하게 해석하는 데 있으며, 이는 주어진 임상 표현형에 대한 희귀 유해 변이체의 효과 크기를 예측하는 것뿐만 아니라 유의한 희귀 변이체 연관성을 갖는 유전자를 식별하는 것 모두에 중요하다. 희귀 변이체 PRS 유전자에 나타나는 희귀 유해 변이체의 거의 절반이 전체 200,643개의 영국 Biobank 엑솜 코호트에서 단지 한 개인에게서만 나타나기 때문에, 이 문제는 전통적인 통계 분석으로는 다루기 힘든 것처럼 보일 수 있지만, 심층 학습, 높은 처리량의 실험 분석 및 밀접하게 관련된 영장류 종으로부터의 변이체 정보를 사용하는 최근의 발전은 전장 유전체 규모에서 VUS 문제를 해결하는 방향으로의 진전을 보여주었다. 희귀 변이체 PRS 유전자의 기본이 되는 예외적인 대립 유전자 이질성은 변이체 효과 예측의 관점에서 볼 때 어려울 수 있지만, 역설적으로, 이는 또한 서로 다른 코호트에 걸쳐 희귀 변이체 PRS 모델의 견고성과 이식성의 핵심이다: 유전자당 수천개의 고유 희귀 유해 변이체에 걸쳐 신호를 통합함으로써, 희귀 변이체 PRS는 데이터에 존재할 수 있는 임의의 변이체 특정 아티팩트의 효과를 완화하는 바람직한 특성을 얻는다.

[0084] 희귀 및 공통 유전자 변이체가 일반적인 질병과 복잡한 형질에 대한 위험에 기여하는 정도는 인간 유전학 분야에서 수십 년 동안 지속된 논쟁이었다. 개시된 기술은 GWAS 유전자좌의 상당 부분에서 희귀 침투 변이체의 존재를 입증하고 전체 모집단 분산은 물론 심각한 조기 발병 질병에 대한 위험이 가장 큰 이상치 개체에 대한 희귀 및 공통 변이체의 상대적 기여도를 정량화한다.

[0085] 정밀 의학의 관점에서 볼 때, 전체적인 영국 모집단으로부터의 성인의 대표적인 면인 영국 Biobank 코호트는 또한 희귀 침투 변이체의 부담과 일반 모집단에서의 이의 효과를 특성화할 수 있는 고유 기회를 제시한다. 개시된 기술의 하나의 구현예에서 연구된 90개의 임상적 및 정량적 표현형 중 하나 이상에 대해 유의한 500개 유전자에 걸쳐, 개체의 86%가 적어도 하나의 희귀 유해 변이체를 보유했으며, 개체당 평균 2.03개의 희귀 유해 변이체가 있었다. 개시된 기술의 일부 구현예는 이러한 희귀 유해 변이체가 매우 높은 침투 효과를 가지며, 동일한 유전자좌에 존재하는 공통 GWAS 변이체보다 평균 10배 더 큰 효과 크기에 기여한다는 것을 보여준다. 평균적으로, 개체의 5.2%가 주어진 표현형에 대해 희귀 유해 변이체를 보유했으며, 개체의 0.4%가 주어진 유전자에 대해 희귀 침투 변이체를 보유했다.

[0086] 개시된 기술은 희귀 변이체 부담 검정 접근법을 강화하여 성능을 최대화하고 양성 미스센스 변수를 필터링한다. 개시된 기술의 일부 구현예에서, 부담 검정 성능은 부담 검정의 유의성을 최대화하는 최대 대립 유전자 카운트에 대한 컷오프를 찾기 위해 관찰된 변이체의 대립 유전자 카운트에 대한 제1 그리드 검색을 수행함으로써 최대화되고, 양성 미스센스 변이체는 부담 검정의 유의성을 최대화하는 병원성 점수에 대한 최적 임계값을 찾기 위하여 병원성 점수(예를 들어, PrimateAI-3D 점수) 임계값에 대한 복수의 컷오프에 대해 제2 그리드 검색을 수행함으로써 필터링된다.

[0087] 개시된 기술은, 다수의 표현형 연관 유전자로부터의 희귀 유해 변이체의 가중 합에 기초하여, 희귀 유해 변이체 다유전자 위험 점수 모델을 개발한다. 개별 전장 유전체 연관성 분석 변이체는 단일 유전자 장애와 비교하여 다수의 유전자에서의 변이체로 인해 유발되는 유전 장애 및 임상 실행 가능성에 비해 너무 가벼운 효과를 부여하는 경향이 있다. 따라서, 복수의 변이체로부터의 신호를 조합하는 다유전자 위험 점수 모델은 질병의 위험이 높은 환자를 예측하는 데 종종 더 효과적이다. 그러나 기존 다유전자 위험 점수 모델은 주로 공통 변이에 대해 최적화되어 있으며, 유의성이 알려지지 않은 변이체를 해석하는 어려움, 검정력이 낮은 연구 및 희귀 변이체의 작은 샘플 크기로 인한 부정확한 효과 크기 추정 때문에 희귀 변이체를 제외한다. 개시된 기술은 개체가 특정 유전자 내에 적어도 하나의 희귀 유해 변이체를 보유하는지 여부에 기초하여 유전자 분해 기준으로 위험을 집계함으로써 희귀 변이체의 포함을 위해 이전 다유전자 위험 점수 모델을 개선한다.

[0088] 개시된 희귀 다유전자 위험 점수는 표현형 반응에 대한 희귀 변이체의 기여도 및 표현형과 연관된 유전자의 연관성 강도가 정량화되는 부담 검정으로서 계산된다. 개체 코호트에 대한 변이체 보인자 상태 및 특정 측정 표현형 값을 나타내는 데이터가 획득된다. 획득된 데이터를 사용하여 연관된 유전자 각각에 대한 부담 점수가 계산

되며, 부담 점수는 연관된 유전자 각각의 보인자 상태와 하나 이상의 희귀 병원성 변이체의 연관된 유전자에서의 존재에 대한 표현형 반응 사이의 코호트에서의 결과적이고 비무작위적인 연관성을 식별한다. 보인자 상태는 특정 유전자에서의 하나 이상의 희귀 병원성 변이체의 존재 또는 부재에 의해 결정되는 부울(Boolean) 변수이다. 특정 희귀 변이체의 병원성은 유전자가 단백질로 발현될 때 기능에 대한 특정 희귀 변이체의 예측된 영향에 의해 결정되며, 예측된 영향은 컨볼루션 신경망 병원성 분류자에 의해 병원성 점수로 결정된다. 특정 희귀 변이체의 희귀성은 모집단 내의 특정 희귀 변이체의 사전 결정된 임계값(즉, 최대 대립 유전자 카운트) 미만의 발생에 의해 결정된다.

[0089] 연관성 강도는 유전자당 분해에서 각각의 유전자에 대한 희귀 병원성 변이체의 보인자 상태에 대한 각각의 표현형 반응 및 각각의 유전자에 대한 희귀 병원성 변이체의 보인자 상태에 대한 유효 강도 점수로 정량화된다. 표현형 반응에 대한 연관된 유전자 전반에 걸쳐 코호트 내의 대상체의 보인자 상태의 기여도에 대한 점수는 유효 강도 점수에 기초하여 결정될 수 있다. 개시된 기술의 일부 구현예에서, 특정 표현형은 정량적 바이오마커 표현형이고, 유전자에 대한 결과적이고 비무작위적인 연관성에 대한 유효 강도 점수는 부담 점수의 선형 회귀 성분에 대한 양측(two-tailed) t-검정을 사용하여 정량적 바이오마커 표현형에 대해 결정된다. 양측 t-검정은 보인자와 비보인자의 평균 표현형 측정값 사이의 차이가 사전 결정된 유의 수준에서 유의한지 여부를 결정하기 위해 p-값을 생성한다.

[0090] 개시된 기술의 다른 구현예에서, 특정 표현형은 범주형 임상 진단 표현형이고, 유전자에 대한 결과적이고 비무작위적인 연관성에 대한 유효 강도 점수는 보인자 상태에 대한 베타 계수로서 범주형 임상 진단 표현형에 대해 결정된다. 베타 계수는 부담 점수의 로지스틱 회귀 성분을 사용하여 결정되며, 로지스틱 회귀 성분은 보인자 상태 및 복수의 공변량에 대응하는 이진 지시자 변수로부터 임상 진단 라벨을 예측하도록 피팅된다.

[0091] 개시된 기술의 일부 구현예에서, 특정 유전자 및 특정 표현형에 대한 각각의 부담 검정은 모두 특정 유전자에 대해 최적화되고 특정 표현형 값에 대해 보정된다. 개시된 기술의 일부 구현예는 특정 유전자에 대한 최적화 또는 특정 표현형 값에 대한 보정의 어느 하나를 포함할 수 있지만 다른 것은 포함하지 않을 수 있다.

[0092] 개시된 기술의 추가 구현예는 보인자와 비보인자 사이의 분리를 최대화하는 네스팅된(nested) t-검정을 사용하여 희귀 변이체 부담 검정을 최적화하고, 부담 검정 매개변수는 특정 유전자에 대해 최적화된다. 부담 검정 성능은 부담 검정의 유의성을 최대화하는 최대 대립 유전자 카운트에 대한 컷오프를 찾기 위해 관찰된 변이체의 대립 유전자 카운트에 대한 제1 그리드 검색을 수행함으로써 최대화되고, 양성 미스센스 변이체는 부담 검정의 유의성을 최대화하는 병원성 점수에 대한 최적 임계값을 찾기 위하여 병원성 점수(예를 들어, PrimateAI-3D 점수) 임계값에 대한 복수의 컷오프에 대해 제2 그리드 검색을 수행함으로써 필터링된다.

[0093] 개시된 기술의 일부 구현예에서, 그리드 검색 절차는 복수의 대립 유전자 카운트 및 복수의 병원성 점수 임계값을 검색하는 것을 포함하며, 그리드 검색은 복수의 대립 유전자 카운트 및 복수의 병원성 점수 임계값으로부터 대립 유전자 카운트 및 병원성 점수 임계값의 복수의 조합의 생성, 대립유전자 카운트 및 병원성 점수 임계값의 복수의 조합에 대응하는 복수의 희귀 병원성 변이체 그룹의 식별, 개체 코호트 내의 특정 희귀 병원성 변이체 그룹의 보인자를 개체 코호트 내의 특정 희귀 병원성 변이체 그룹의 비보인자로부터 분리하는 보인자 상태에 따른 복수의 희귀 병원성 변이체 그룹의 부담 검정 및 대립 유전자 카운트와 병원성 점수 임계값의 복수의 조합에 대응하는 복수의 효과 크기 및 p-값의 결정을 포함한다.

[0094] 가장 유의한 p-값을 갖는 대립 유전자 카운트와 병원성 점수 임계값의 특정 조합이 특정 유전자에 대한 최적 조합으로 사용되도록 선택된다. 각각의 유전자는 각각의 유전자의 보인자와 비보인자 사이의 최적 분리를 제공하는 유사한 최대 대립 유전자 카운트 임계값 및/또는 최소 병원성 점수 임계값을 가질 수도 있고 가지지 않을 수도 있다. 병원성 점수 임계값은 희귀 병원성 변이체 그룹에서 희귀 병원성 변이체에 대해 결정되는 병원성 점수의 병원성 점수 변위치(quantile)에 대응한다. 개시된 기술의 일부 구현예에서, 병원성 점수는 PrimateAI-3D와 같은 컨볼루션 신경망 병원성 분류자에 의해 생성된다. 개시된 기술의 다른 구현예에서, 병원성 점수를 생성하기 위해 광범위한 추가 AI, 기계 학습 및 심층 학습 모델이 사용된다.

[0095] 전술된 그리드 탐색 절차는 생성된 오류 발견율(false discovery rate) 보정 p-값에 대해 다중 검정에 대한 보정을 필요로 한다. 개시된 기술의 일부 구현예에서, 다중 검정 보정은 Benjamini-Hochberg 절차를 사용하여 수행된다. 개시된 기술의 다른 구현예에서, 다중 검정 보정은 적응형 순열 검정 절차를 사용하여 수행된다.

[0096] 개시된 기술의 추가 구현예는 시간적으로 분포된 검출 시점에 걸쳐 약물 사용에 공변량 보정을 적용한다. 부담 검정 입력 데이터는 측정된 표현형 값에 대해 보정된다. 표현형 반응 값은 복수의 공변량에 대해 공변량 보정되

고, 복수의 약물에 대해 약물 사용량 보정된다. 개시된 기술의 일부 구현예에서, 복수의 교란 변수(confounder)를 갖는 개체 코호트의 복수의 표현형에 대한 복수의 약물에 따른 표현형 시프트는 공변량 보정되고 약물 사용 보정된 표현형 측정값을 생성할 것으로 예측된다.

[0097] 개시된 기술의 일부 구현예에서, 표현형 시프트는 개체 코호트에 대한 표현형별 기준 및 복수의 표현형에 대한 표현형 측정값에서의 복수의 교란 변수(예를 들어, 연령, 성별, 유전적 주요 성분, 식이, 흡연 상태)에 따른 개체 코호트의 복수의 표현형에 대한 복수의 약물의 사용, 복수의 교란 변수에 대한 공변량 측정값 및 2개의 별개 시점에서의 코호트 내의 각각의 개체에 대한 복수의 약물에 대한 약물 사용 패턴에 따라 예측된다.

[0098] 표현형 측정값은 제1 회귀 모델을 피팅하여 공변량 측정값을 회귀시킴으로써 공변량 측정값에 기초하여 제1 시점 및 제2 시점에 대해 공변량 보정되고, 이에 따라 제1 시점 및 제2 시점에 대한 공변량 보정 표현형 측정값을 생성한다. 델타는 제1 시점 및 제2 시점에 대한 공변량 보정 표현형 측정값 사이의 차이에 기초하여 결정된다. 각각의 약물 사용 패턴에 대해, 공변량 보정 표현형 측정값에서의 복수의 약물 사용에 따른 표현형 시프트를 예측하기 위해 델타를 사용하는 제2 회귀 모델이 피팅된다.

[0099] 제2 회귀 모델은 각각의 약물 사용 패턴에 대한 표현형 시프트를 연속적으로 그리고 누적적으로 포함함으로써 델타를 반복적으로 예측하는 전진 선택 단계적 회귀 모델(forward selection stepwise regression model)이다. 제2 회귀 모델은 각각의 약물 사용 패턴에 대해 이진 지시자 독립 변수를 가지며, 약물 사용 패턴은 제1 시점 및 제2 시점에 약물을 사용하지 않은 것, 제1 시점과 제2 시점 사이에 약물을 사용하기 시작하는 것, 제1 시점과 제2 시점 사이에 약물을 사용하는 것을 중단하는 것 및 제1 시점 및 제2 시점에 약물을 사용하는 것을 포함한다. 개시된 기술의 일부 구현예에서, 공변량 보정, 델타 결정, 표현형 시프트 예측 및 약물 사용 보정은 복수의 약물에서의 약물에 대해 약물별 기준으로 실행된다. 개시된 기술의 다른 구현예에서, 공변량 보정, 델타 결정, 표현형 시프트 예측 및 약물 사용 보정은 복수의 약물 카테고리에서의 약물 카테고리에 대해 약물 카테고리별 기준(예를 들어, 스타틴, NSAID, 오피오이드 등)으로 실행된다.

[0100] 제2 회귀 모델은 개체 코호트에서의 개체에 대해 개체별 기준으로 제1 시점과 제2 시점 사이의 시간에 경과된 시간에 따른 표현형 시프트 및 제1 시점과 제2 시점 사이의 평균으로의 회귀에 따른 표현형 시프트를 더 모델링한다. 제2 회귀 모델은 복수의 약물에서의 관련 약물 세트에 공동으로 피팅된다.

[0101] 개시된 기술의 일부 구현예에서, 약물 사용 보정은, 제1 시점과 제2 시점 사이에 약물을 사용하기 시작하는 제1 약물 사용 패턴에 대한 제1 이진 지시자 독립 변수, 제1 시점 및 제2 시점에 약물을 사용하지 않는 제2 약물 사용 패턴에 대한 제2 이진 지시자 독립 변수 및 개체가 제1 시점에 특정 약물을 사용하고 있었는지 여부를 인코딩하는 약물 특정 이진 지시자 독립 변수에 기초하여 제1 시점에 대한 표현형 측정값을 약물 사용 보정하는 것을 포함하는 제3 회귀 모델을 피팅함으로써 구현된다.

[0102] 제1 시점과 제2 시점 사이에 약물을 사용하는 것을 중단하는 제3 약물 사용 패턴에 대한 제3 이진 지시자 독립 변수, 제1 시점 및 제2 시점에 약물을 사용하는 제4 약물 사용 패턴에 대한 제4 이진 지시자 독립 변수 및 개체가 제2 시점에 특정 약물을 사용하고 있었는지 여부를 인코딩하는 약물 특정 이진 지시자 독립 변수에 기초하여 제2 시점에 대한 표현형 측정값을 약물 사용 보정하는 것을 포함하는 제4 회귀 모델이 피팅된다.

[0103] 순위 기반 역정규 변환은 제1 시점 및 제2 시점에 대한 약물 사용 보정 표현형 측정값에 적용되어 제1 시점 및 제2 시점에 대한 정규화된 약물 사용 보정 표현형 측정값을 생성할 수 있다. 그 다음, 정규화된 약물 사용 보정 표현형 측정값은 제1 시점 및 제2 시점에 대해 공변량 보정되어, 제1 시점 및 제2 시점에 대한 공변량 보정되고 정규화된 약물 사용 보정 표현형 측정값을 생성한다. 공변량 보정되고 정규화된 약물 사용 보정 표현형 측정값은 희귀 변이체 다유전자 위험 점수를 생성하는 데 사용될 수 있고, 관심 표현형에 대한 측정값이 공변량 및 약물 사용 패턴에 따른 표현형 시프트에 대해 보정된다.

[0104] 개시된 기술의 일부 구현예에서, 구축된 회귀 모델은 원치 않는 부작용 및 원하는 표적 효과와 같은 약물-표현형 연관성을 검출하는 데 사용될 수 있다.

[0105] **<u>희귀 변이체 다유전자 위험 점수(PRS: Polygenic Risk Score)</u>**

[0106] 도 1a는 가중된 희귀 변이체 PRS를 결정하는 방법을 예시하는 개략도이다. 데이터베이스(102)는 코호트 *i*에 속하는 개체 그룹에 대응하는 게놈 및 표현형 데이터를 포함한다. 코호트 *i*(104)는 복수의 표현형 측정값을 위해 유전자 서열 분석 및 검정을 거친 *N*명의 개체를 포함한다. 코호트 *i*(104) 내의 개체에 대한 서열 분석 데이터는 복수의 변이체 보인자 상태에 대응하는 데이터를 포함한다. 특정 유전자는 3 가지 알려진 가능성 있는 희귀 유해 변이체(예를 들어, 변이체 A, 변이체 B 및 변이체 C)를 보유할 수 있으며, 개체는 보인자(즉, 개체는 해당

변이체를 보유하므로 그 변이체의 보인자가 된다) 또는 비보인자(즉, 개체는 해당 변이체를 보유하지 않으므로 그 변이체의 보인자가 아니다)가 될 수 있다. 당업자는 변이체 A, 변이체 B 및 변이체 C가 예시적인 예로서 주어지고 유전자가 임의의 수의 희귀 유해 변이체를 보유할 수 있다는 것을 인식할 것이다. 개시된 기술은 변이체를 모집단 내의 변이체 발생률이 사전 결정된 임계값 미만인 경우 '희귀(rare)'로서 정의하고 유전자가 단백질로 발현될 때 변이체의 병원성이 사전 결정된 임계값을 초과하는 측정된 영향을 가질 것으로 예측되는 경우 '유해(deleterious)'로서 정의한다. 희귀성과 병원성 모두에 대한 각각의 사전 결정된 임계값은 유전자 특이적이다(즉, 각각의 개별 유전자는 다른 유전자와 다를 수 있는 개별적으로 결정된 희귀성 임계값 및 병원성 임계값에 대응할 것이다).

[0107] 코호트 *i*(104) 내의 각각의 개체에 대한 게놈 및 엑솜 서열 분석으로부터 획득된 게놈 데이터 외에도, 표현형 데이터도 각각의 개체에 대해 제공된다(즉, 개체 N은 표현형 D에 대해 x_n의 측정된 값을 갖는다). 특정 유전자형과 특정 표현형 사이의 관계에 대한 모델(122)을 구축함으로써, 특정 표현형에 대한 특정 유전자형의 효과 크기를 측정하는 것이 가능하다. 상자 수염 플롯(Box-and-whisker plot)(124)은 표현형 D에 대한 하나 이상의 특정 희귀 유해 변이체에 대한 보인자 상태의 효과가 *p*-값(즉, 보인자와 비보인자 사이의 평균 표현형 값을 비교할 때의 유의도)과 β 계수(즉, 독립 변수 및 종속 변수의 분산이 1이 되도록 기본 데이터가 표준화되는 보인자의 평균 표현형 값과 비보인자의 평균 표현형 값을 연결하는 회귀선으로부터 유래하는 가중치 추정값) 모두의 형태로 측정되는 대표적인 유전자 연관성 검정을 예시한다.

[0108] 특정 표현형이 정량적 바이오마커 표현형인 기술의 일부 구현예에서, 연관성 강도는 양측 *t*-검정에 의해 결정된 *p*-값으로 정량화된다. 특정 표현형이 정성적 표현형(예를 들어, 범주형 임상 진단 표현형)인 개시된 기술의 다른 구현예에서, 연관성 강도는 로지스틱 회귀의 β 계수로 정량화된다.

[0109] 개시된 기술에서, 보인자 상태는 변이체 분해 기준이 아닌 유전자 분해 기준으로 정의된다. 특정 유전자에 대해, 개체가 특정 유전자 내에 적어도 하나의 희귀 유해 변이체를 보유하고 있는 경우 보인자로 정의된다(즉, 2개의 희귀 유해 변이체에 대한 보인자인 개체가 하나의 희귀 유해 변이체의 보인자와 구별되지 않는 반면, 0개의 희귀 유해 변이체에 대한 보인자인 개체는 비보인자로서 구별된다). 복수의 희귀 변이체의 보인자 상태와 표현형 반응 사이의 코호트 *i*(104) 내의 연관성 강도의 유효 강도 점수에 대해 결정(142)이 이루어질 수 있다.

[0110] 그래프(144)는 복수의 희귀 변이체의 보인자 상태와 표현형 반응 사이의 코호트 i(104) 내의 연관성 강도의 유효 강도 점수의 결정을 위해 *t*-검정을 포함한다. *t*-검정에 대한 귀무 가설(null hypothesis)은 코호트 *i* 내 보인자의 표본 평균과 코호트 *i*(104) 내 비보인자의 표본 평균 사이의 차이의 절대값이 0과 같다고 명시한다. *t*-검정에 대한 대립 가설(alternative hypothesis)은 코호트 *i*(104) 내 보인자의 표본 평균과 코호트 *i*(104) 내 비보인자의 표본 평균 사이의 차이의 절대값이 0이 아니라고 명시한다. 귀무 가설을 수락하거나 거부하는 결정은 특정 유의 수준 α와 α에 대한 결과적인 *p*-값에 따라 결정된다. (차트(124)에 표시된 바와 같이, *p*-값 또는 β 계수로 측정된) 특정 공유 표현형 반응에 대한 복수의 유전자 특정 유효 강도 점수를 획득하기 위해 복수의 개별 유전자에 대해 개별 *t*-검정이 수행될 수 있다.

[0111] 가중 부담 점수(162)는 복수의 개별 유전자 특정 유효 강도 점수로부터 생성될 수 있다. 희귀 변이체 PRS로서 설명되는 이 계산은 식(164)에 도시되며, 식에서 희귀 변이체 PRS는 복수의 유전자에 대한 특정 유전자에 대한 효과 크기와 보인자 상태의 곱의 합산과 동일하다.

[0112] 이제 논의는 특정 개체에 대한 희귀 변이체 PRS, 특정 개체에 대해 서열 분석된 특정 복수의 개별 유전자 및 특정 관심 표현형을 생성하기 위해 복수의 개별 유전자 특정 유효 강도 점수를 사용하는 일 실시예를 제공한다.

[0113] 도 1b는 특정 복수의 유전자 및 특정 표현형에 대한 희귀 변이체 다유전자 위험 점수의 예시적인 계산을 예시한다. 헤더 행(108)은 특정 개체에 대한 유전자, 효과 크기 및 보인자 상태에 대응하는 열 라벨 그룹을 포함한다. 유전자 열에 포함된 유전자는 유전자 A(행 118), 유전자 B (행 128), 유전자 C(행 138), 유전자 D (행 148), 유전자 E(행 158)이고, 유전자 A 내지 E는 특정 표현형 Y와 연루된 유전자로 가정될 수 있다. 각각의 유전자에 대한 효과 크기에서, 효과 크기 A는 유전자 A(행 118)에 대한 것이고, 효과 크기 B는 유전자 B(행 128)에 대한 것이고, 효과 크기 C는 유전자 C(행 138)에 대한 것이고, 효과 크기 D는 유전자 D(행 148)에 대한 것이고, 효과 크기 E는 유전자 E(행 158)에 대한 것이다. 보인자 상태는 특정 유전자 내의 적어도 하나의 식별된 희귀 유해 변이체의 존재를 설명하는 부울 변수이다(예를 들어, 코호트 *i*(104)에서 검출된 변이체 A, B 및 C가 유전자 A(행 118)에 특정하다고 가정하면, 이에 따라 코호트 *i*(104)에 대한 표에 나열된 각각의 개체는 유전자 A에 대한 보인자로 간주된다). 게놈이 도 1b에서 특성화된 개체는 각각 유전자 A(행 118), 유전자 C(행 138), 유전자 D(행 148), 유전자 E(행 158)에서 적어도 하나의 희귀 유해 변이체에 대해 보인자이다. 개체는 유전자 B(행 12

8)에서 어떠한 희귀 유해 변이체에 대해서도 보인자가 아니다.

[0114] 가중 부담 점수(162)를 출력하기 위해, 식(164)이 적용된다. 특정 개체에 대한 식(164)은 식(182)과 동등한 것으로 간주될 수 있으며, 특정 개체는 유전자 A(행 118)에 대한 효과 크기 A, 유전자 C(행 138)에 대한 효과 크기 C, 유전자 D(행 148)에 대한 효과 크기 D, 유전자 E(행 158)에 대한 효과 크기 E의 합과 동일한 특정 표현형 Y에 대한 희귀 변이체 PRS를 갖는다. 특정 개체가 적어도 하나의 희귀 유해 변이체에 대한 보인자인 각각의 해당 유전자에 대해, 보인자 상태는 1과 같다(즉, 효과 크기 * 1 = 효과 크기). 유전자 B(행 128)에 대한 효과 크기 B는 특정 개체가 적어도 하나의 희귀 유해 변이체에 대한 보인자가 아니기 때문에 식(182)에 표시되지 않는다; 따라서, 보인자 상태는 0과 같고(즉, 효과 크기 *0 = 0), 유전자 B(행 128)에 대한 효과 크기 B는 식(182)에 존재하지 않는다.

[0115] 도 2은 개시된 기술을 구현하는 데 사용될 수 있는 예시적인 컴퓨터 시스템(200)을 도시한다. 컴퓨터 시스템(200)은 버스 서브시스템(242)을 통해 다수의 주변 디바이스와 통신하는 적어도 하나의 중앙 처리 장치(CPU: central processing unit)(244)를 포함한다. 이러한 주변 디바이스는, 예를 들어 메모리 디바이스 및 파일 저장 서브시스템(236)을 포함하는 저장 서브시스템(210), 사용자 인터페이스 입력 디바이스(238), 사용자 인터페이스 출력 디바이스(248) 및 네트워크 인터페이스 서브시스템(246)을 포함할 수 있다. 입력 및 출력 디바이스는 컴퓨터 시스템(200)과의 사용자 상호 작용을 가능하게 한다. 네트워크 인터페이스 서브시스템(246)은 다른 컴퓨터 시스템에서의 대응하는 인터페이스 디바이스에 대한 인터페이스를 포함하는 인터페이스를 외부 네트워크에 제공한다.

[0116] 하나의 구현예에서, 희귀 변이체 PRS 모델(240)은 저장 서브시스템(210) 및 사용자 인터페이스 입력 디바이스(238)에 통신 가능하게 연결된다.

[0117] 사용자 인터페이스 입력 디바이스(238)는, 키보드; 마우스, 트랙볼, 터치패드 또는 그래픽 태블릿과 같은 포인팅 디바이스; 스캐너; 디스플레이 내에 통합된 터치 스크린; 음성 인식 시스템 및 마이크와 같은 오디오 입력 디바이스; 및 다른 유형의 입력 디바이스를 포함할 수 있다. 일반적으로, "입력 디바이스"라는 용어의 사용은 정보를 컴퓨터 시스템(200)에 입력하기 위한 모든 가능한 유형의 디바이스 및 방식을 포함하도록 의도된다.

[0118] 사용자 인터페이스 출력 디바이스(248)는 디스플레이 서브시스템, 프린터, 팩스 기계 또는 오디오 출력 디바이스와 같은 비시각적 디스플레이를 포함할 수 있다. 디스플레이 서브시스템은 LED 디스플레이, 음극선관(CRT), 액정 디스플레이(LCD)와 같은 평면 디바이스, 프로젝션 디바이스 또는 가시적인 이미지를 생성하기 위한 일부 다른 메커니즘을 포함할 수 있다. 디스플레이 서브시스템은 또한 오디오 출력 디바이스와 같은 비시각적 디스플레이를 제공할 수 있다. 일반적으로, "출력 디바이스"라는 용어의 사용은 정보를 컴퓨터 시스템(200)으로부터 사용자에게 또는 다른 기계 또는 컴퓨터 시스템에 출력하기 위한 모든 가능한 유형의 디바이스 및 방식을 포함하도록 의도된다.

[0119] 저장 서브시스템(210)은 본원에 설명된 모듈 및 방법 중 일부 또는 전부의 기능을 제공하는 프로그래밍 및 데이터 구성을 저장한다. 이러한 소프트웨어 모듈은 일반적으로 프로세서(249)에 의해 실행된다.

[0120] 프로세서(249)는 그래픽 처리 장치(GPU: graphics processing unit), 필드 프로그래밍 가능 게이트 어레이(FPGA: field-programmable gate array), 주문형 반도체(ASIC: application-specific integrated circuit) 및/또는 CGRA(coarse-grained reconfigurable architecture)일 수 있다. 프로세서(249)는 Google Cloud Platform™, Xilinx™ 및 Cirrascale™과 같은 심층 학습 클라우드 플랫폼에 의해 호스팅될 수 있다. 프로세서(249)의 예는 Google의 Tensor Processing Unit(TPU)™, GX4 Rackmount Series™, GX2 Rackmount Series™와 같은 랙마운트 솔루션, NVIDIA DGX-1™, Microsoft의 Stratix V FPGA™, Graphcore의 Intelligent Processor Unit(IPU)™, Snapdragon processors™을 갖는 Qualcomm의 Zeroth Platform™, NVIDIA의 Volta™, NVIDIA의 DRIVE PX™, NVIDIA의 JETSON TX1/TX2 MODULE™, Intel의 Nirvana™, Movidius VPU™, Fujitsu DPI™, ARM의 DynamicIQ™, IBM TrueNorth™, Testa V100s™를 갖는 Lambda GPU 서버 등을 포함한다.

[0121] 저장 서브시스템(210)에 사용되는 메모리 서브시스템(222)은 프로그램 실행 동안의 명령어 및 데이터의 저장을 위한 메인 랜덤 액세스 메모리(RAM: random-access memory)(232) 및 고정된 명령어가 저장되는 판독 전용 메모리(ROM: read only memory)(234)를 포함하는 다수의 메모리를 포함할 수 있다. 파일 저장 서브시스템(236)은 프로그램 및 데이터 파일을 위한 영구 저장소를 제공할 수 있고, 하드 디스크 드라이브, 연관된 탈착식 매체와 함께하는 플로피 디스크 드라이브, CD-ROM 드라이브, 광학 드라이브 또는 탈착식 매체 카트리지를 포함할 수 있다. 특정 구현예의 기능을 구현하는 모듈은 저장 서브시스템(210) 내의 파일 저장 서브시스템(236)에 의해,

또는 프로세서에 의해 접근 가능한 다른 기계에 저장될 수 있다.

[0122] 버스 서브시스템(242)은 컴퓨터 시스템(200)의 다양한 컴포넌트 및 서브시스템이 의도된 바와 같이 서로 통신하게 하기 위한 메커니즘을 제공한다. 버스 서브시스템(242)이 단일 버스로서 개략적으로 도시되어 있지만, 버스 서브시스템의 대안적인 구현예는 다수의 버스를 사용할 수 있다.

[0123] 컴퓨터 시스템(200) 자체는 개인용 컴퓨터, 휴대용 컴퓨터, 워크스테이션, 컴퓨터 단말기, 네트워크 컴퓨터, 텔레비전, 메인프레임, 서버 팜, 느슨하게 네트워킹된 컴퓨터의 광범위하게 분포된 세트 또는 임의의 다른 데이터 처리 시스템 또는 사용자 디바이스를 포함한 다양한 유형의 것일 수 있다. 컴퓨터 및 네트워크의 지속적으로 변화하는 특성으로 인해, 도 2에 묘사된 컴퓨터 시스템(200)의 설명은 본 발명의 바람직한 구현예를 예시하기 위한 특정 실시예로서만 의도된다. 도 2에 묘사된 컴퓨터 시스템보다 더 많은 또는 더 적은 컴포넌트를 갖는 컴퓨터 시스템(200)의 많은 다른 구성이 가능하다.

[0124] **대안적인 희귀 유해 다유전자 위험 점수 모델**

[0125] 하나의 구현예에서, 개시된 기술은 위에서 논의된 것과는 다른 희귀 유해 다유전자 위험 점수 모델을 사용할 수 있다. 유전자에서 변이체에 할당된 가중치가 그 유전자 내의 모든 희귀 유해 변이체에 걸쳐 관찰된 효과 크기와 동일한, 즉 유전자에서의 모든 변이체에 대해 동일한 희귀 유해 변이체의 가중 합을 사용하는 대신, 개시된 기술은 각각의 유의한 유전자 내의 변이체별 가중치를 추정할 수 있다. 변이체별 가중치는, 변이체 병원성 점수(예를 들어, PrimateAI-3D)의 선형 회귀 분석과 개체의 형질 값에 대한 log10 변환 대립 유전자 빈도를 통해, 유의한 유전자에 대해 희귀 유해 변이체를 보유하는 개체로부터 추정될 수 있다. 이는, 일부 구현예에서, 설명된 분산(variance explained)의 측면에서 성능을 18%만큼 향상시킬 수 있다.

[0126] **유전자형**

[0127] 도 1에 설명된 다유전자 위험 점수를 계산하는 방법은 유전자형 분석 어레이(genotyping array)(즉, 특정 개체의 전체 게놈 서열 분석) 및 엑솜 서열 분석(즉, 특정 개체의 게놈의 단백질 코딩 영역의 서열 분석)에 의해 결정된 개체의 유전자형과 연관된 데이터를 포함한다. 이제 논의는 특정 개체에 대한 게놈 데이터, 이의 연관된 특징부 및 유전자형과 표현형 사이의 관계에 대한 세부 사항으로 넘어간다.

[0128] 도 3은 유전자 변이체의 한 예를 예시한다. 예시적인 기준 유전자 서열 A(302)는 *N*의 길이를 보유하며, 각각의 서열 위치는 티민, 아데닌, 구아닌 또는 시토신의 뉴클레오티드 염기를 포함한다. 기준 유전자 서열 A(302)는 특정 유전자의 세그먼트에 대한 천연(native) 또는 정상 유전자 서열로 간주될 수 있다. 기준 유전자 서열 A(302)는 변이체 서열이 5번째 염기 위치에 각자의 단일 뉴클레오티드 변이체를 보유하지만 그 외에는 기준 서열과 동일한 조성을 보유하는 2개의 예시적인 변이체 서열 1A(322) 및 1B(342)와 비교된다. 예를 들어, 기준 유전자 서열 A(302)에서의 시토신(304)과 비교하여 단일 뉴클레오티드 치환은 변이체 1A(322)에서의 아데닌(324) 및 변이체 1B(342)에서의 티민(344)으로 표시된다.

[0129] 기준 유전자 서열 A(302), 변이체 1A(322) 및 변이체 1B(342)가 단백질의 아미노산 단위로 전사되고 번역되는 경우, 단일 뉴클레오티드 변이체(324, 344)는 각각 천연 단백질에 존재하는 아미노산(들)과 비교하여 적어도 하나의 유사하거나 상이한 아미노산을 생성할 수 있다. 변이체 1A(322) 또는 변이체 1B(342)가 각각 번역된 단백질에서 적어도 하나의 변경된 아미노산을 생성하는 경우, 각각의 변이체는 또한 해당 변이체를 보유하는 개체에 대한 변경된 표현형을 생성할 수도 있다.

[0130] 도 4는 유전자형에서의 변화에 따른 표현형 효과를 예시하는 한 예이다. 예시적인 기준 유전자 서열 A(402)는 N의 길이를 보유하며, 각각의 서열 위치는 티민, 아데닌, 구아닌 또는 시토신의 뉴클레오티드 염기를 포함한다. 기준 유전자 서열 A(402)는 특정 유전자의 세그먼트에 대한 천연(native) 또는 정상 유전자 서열로 간주될 수 있다. 기준 유전자 서열 A(402)는 변이체 서열이 5번째 염기 위치에 각자의 단일 뉴클레오티드 변이체를 보유하지만 그 외에는 기준 서열과 동일한 조성을 보유하는 3개의 예시적인 변이체 서열 1A(422), 1B(442) 및 1C(462)와 비교된다. 예를 들어, 기준 유전자 서열 A(402)에서의 시토신(404)과 비교하여 단일 뉴클레오티드 치환은 변이체 1A(422)에서의 아데닌(404), 변이체 1B(442)에서의 티민(444) 및 변이체 1C(462)에서 구아닌(464)으로 표시된다.

[0131] 기준 유전자 서열 A(402)는 기준 유전자 서열 A(402)를 포함하는 특정 유전자에 대한 천연 단백질 조성인 단백질 A(406)를 생성한다. 천연 단백질 A(406)를 소유한 개체는 건강한 표현형인 표현형 A(408)를 나타낼 것이다. 변이체 1A(422)는 돌연변이 단백질 1A(426)를 생성하며, 돌연변이 단백질 1A(426)는 천연 단백질 A(406)와 다른 단백질 구조 및 기능을 제공하는 미스센스 돌연변이를 포함한다. 미스센스 단백질 1A(406)를 보유한 개체는 질

병 표현형인 표현형 1A(428)를 나타낼 것이다. 변이체 1B(442)는 돌연변이 단백질 1B(446)를 생성하며, 돌연변이 단백질 1B(446)는 천연 단백질 A(406)와 비교하여 단백질 구조 및 기능에 변화가 없는 동의(synonymous) 돌연변이를 포함한다(즉, 변이체 1B(442)의 위치 5에 있는 단일 뉴클레오티드 변이체에 따라 아미노산 변화가 발생하지 않는다). 동의 단백질 1B(446)를 소유한 개체는 표현형 A(408)와 유사한 건강한 표현형인 표현형 1B(448)를 나타낼 것이다. 변이체 1C(462)는 돌연변이 단백질 1C(466)를 생성하며, 돌연변이 단백질 1C(466)는 천연 단백질 A(406)와 비교하여 잘린 비기능적 단백질 구조 및 기능을 생성하는 넌센스(nonsense) 돌연변이를 포함한다. 넌센스 단백질 1C(466)를 보유한 개체는 표현형 1C(468)를 나타낼 것이며, 생존할 수 없는 배아 또는 상당히 단축된 수명을 초래한다.

[0132] 당업자는 변이체 1A(422), 1B(442) 및 1C(462)가 단순화된 예로서 나열되어 있으며 잠재적인 표현형 표현이 제한된 수의 개별 표현보다는 넓은 스펙트럼에 걸쳐 있다는 것을 인식할 것이다. 더욱이, 당업자는 단일 유전자 내의 복수의 변이체 또는 조합된 다유전자 변이체 효과로 인해 많은 표현형 반응이 발생한다는 것을 인식할 것이다. 개시된 기술의 많은 구현예는 개체의 삶의 질 및/또는 기대 수명에 상당한 해를 끼치는 것으로 알려진 심각한 유전 질환과 연관된 특정 표현형 반응에 대한 다유전자 위험 스코어링을 구체적으로 다룬다.

[0133] **표현형**

[0134] 이제 논의는 더욱 상세한 표현형 표시로 넘어간다. 도 4는 광범위한 유기체 수준에서 복수의 표현형 예를 예시하는 반면, 도 5는 이제 정량적 바이오마커 값 또는 범주형 임상 진단으로 설명될 수 있는 표적화되고 측정 가능한 수준에서 복수의 표현형 예를 예시한다.

[0135] 본 출원은 "정량적 바이오마커 값(quantitative biomarker value)"과 "정량적 표현형(quantitative phenotype)"을 상호 교환 가능하게 사용한다는 점에 유의한다. 또한, 본 출원은 "범주형 임상 진단(categorical clinical diagnosis)"과 "정성적 표현형(qualitative phenotype)"을 상호 교환 가능하게 사용한다는 점에 유의한다.

[0136] 도 5는 심혈관 질환 환자 X(500)에 대응하는 복수의 표현형을 예시한다. 심혈관 질환 환자 X(500)는 환자 X의 게놈 내의 복수의 유전적 변이체에 따른 복수의 관찰 가능한 표현형으로 설명될 수 있다. 이러한 표현형은 정성적이거나 정량적일 수 있다. 정성적 표현형은 인구 통계학적 출력(502)(예를 들어, 혈통, 생물학적 성별 등), 범주형 바이오마커(504)(예를 들어, 혈액 내 특정 단백질의 존재에 대한 이진 변수, 결정 임계값에 의해 결정되는 혈액 내 환자의 대사 물질 레벨의 이진 분류 또는 특정 조직학적 형태의 존재에 대한 멀티클래스 변수), 임상 진단(506)(예를 들어, 심혈관 질환 진단에 대한 이진 변수, 특정 암 진단에 대한 이진 변수 등), 또는 질병 중증도 측정(예를 들어, 종양 전이의 진행 단계, 난치성 복강병 1형 대 2형과 같은 질병 하위 유형의 분류 등)을 포함할 수 있다.

[0137] 정량적 표현형은 정확한 혈액 또는 소변 바이오마커 값(522)(예를 들어, 크레아틴, 콜레스테롤, 저밀도 지질 단백질(LDL), 중성 지방, 포도당 등), 체질량 측정값(542)(예를 들어, 키, 체중, 체지방률, 체질량 지수 등) 또는 활력 징후(예를 들어, 안정시 심박수, 수축기 및 확장기 혈압, 호흡수 등)를 포함할 수 있다. 당업자는 심혈관 질환 환자 X(500)에 대해 나열된 정량적 및 정성적 표현형이 비제한적인 예이며 개체의 건강 및 조성을 설명하기 위해 사용될 수 있는 무제한적인 수의 관찰 가능한 값이 있다는 것을 인식할 것이다.

[0138] **유전자-표현형 연관성 검정**

[0139] 지금까지 논의는 특정 개체를 설명하는 데 사용되는 게놈 및 표현형 데이터에 대한 정의 및 배경을 포괄한다. 도 1은 특정 표현형에 대한 유전자 분해 수준에서 특정 희귀 유해 변이체 보인자 상태에 대한 유전적 연관성을 결정하는 방법을 높은 수준에서 설명한다. 이제 논의는 전장 유전체 연관성 분석의 소개와 공통 변이체 및 희귀 변이체 모두에 대한 구현예를 통해 이 방법을 더욱 자세히 재검토한다. 희귀 변이체에 대한 전통적인 전장 유전체 연관성 분석과 대조적으로, 논의는 적어도 하나의 특정 희귀 유해 변이체를 포함하는 복수의 특정 유전자에 대한 복수의 희귀 변이체 부담 검정이 구현되어 복수의 특정 유전자 및 특정 표현형에 대한 희귀 변이체 다유전자 위험 점수를 결정하는 개선된 유용성을 갖는 새로운 방법론을 다룬다. 희귀 유해 변이체가 심각한 조기 발병 유전 질환에 연루될 통계적 가능성으로 인해, 논의는 구체적으로 희귀 유해 변이체에 초점을 맞춘다.

[0140] 도 6은 공통 변이체 대 희귀 변이체에 대해 전장 유전체 연관성 분석을 그래프로 대조한다. 그래프(602) 및 (604)는 각각 콜레스테롤 레벨과 공통 변이체 또는 희귀 변이체에 대한 전장 유전체 연관성을 설명한다. 그래프(602) 및 (604) 모두에 대해, 교대하는 색상은 연속적인 염색체를 나타낸다. 그래프(602)는 개별 공통 변이체를 데이터 포인트로 도시하는 반면, 그래프(604)는 개별 유전자(즉, 모델(122)에 나타낸 바와 같이 단일 보인자 상

태로 응축된 특정 유전자에 대한 모든 희귀 변이체)를 도시한다. 그래프(602)에 도시된 공통 변이체 전장 유전체 연관성 분석은 콜레스테롤 레벨과 강하게 연관된 게놈 내의 많은 위치를 찾아낸다. 서로 가까운 공통 변이체들은 빈번하게 상관되기 때문에, 연관성은 많은 연관된 변이체가 있는 피크로서 발생한다. 그 결과, 특정 연관 유전자좌 내에서 임의의 인과 변이체를 검출하는 것은 상당한 노력을 필요로 할 수 있다.

[0141] 그래프(604)는 인간 게놈에서의 19,500개 단백질 코딩 유전자에 대한 희귀 변이체 전장 유전체 연관성 분석 결과를 도시하고, 소수의 희귀 변이체가 강하게 연관된다. 희귀성으로 인해, 희귀 변이체 전장 유전체 연관성 분석은 그래프(602)에 도시된 공통 변이체 전장 유전체 연관성 분석에 비해 검정력이 좋지 않다. 희귀 변이체는 통상적으로 서로 상관되지 않는다; 따라서, 효과는 인근 유전자로 확장되지 않는다. 따라서, 유의한 유전자는 그래프(602) 내의 클러스터가 아닌 분리된 것으로 나타난다. 단백질 서열을 변경하는 변이체가 구체적으로 검정되며, 인과 변이체가 희귀 변이체 전장 유전체 연관성 분석에 의해 식별된 유의한 유전자에 대해 식별될 수 있다는 것을 의미한다.

[0142] 도 6은 전장 유전체 레벨에서 복수의 유전적 연관성을 예시하는 반면, 이제 논의는 특정 개별 변이체(즉, 특정 개별 공통 변이체 대립 유전자 투여량과 특정 개별 공통 변이체 대립 유전자 투여량 값의 연관된 유전자 내 존재에 대한 표현형 반응의 결과적이고 무작위적인 연관성) 또는 변이체의 특정 집계(즉, 각각의 연관된 유전자의 보인자 상태와 하나 이상의 희귀 병원성 변이체의 연관된 유전자 내 존재에 대한 표현형 반응 사이의 코호트에서의 결과적이고 비무작위적인 연관성)에 대한 유전적 연관성 검정으로 넘어간다.

[0143] 도 7은 각각 유전자 분해 시 특정 개별 공통 변이체 또는 특정 집계 희귀 변이체에 대한 유전자 연관성 검정을 예시한다. 유전자형 분석 어레이로부터의 단일 공통 변이체 유전 연관성 검정(702)의 경우, 개별 공통 변이체에 대한 소수 대립 유전자의 수(여기서 상자 플롯 704, 706 및 708은 각각 0, 1 또는 2개의 소수 대립 유전자에 대한 표현형 값 분포이다) 및 관심 표현형(예를 들어, 콜레스테롤) 사이의 관계를 질문하기 위하여 각각의 개별 변이체가 평가될 수 있다. 특정 변이체 및 표현형 반응에 대한 유효 연관성 강도 점수를 측정하는 *p*-값과 β 계수를 획득하기 위해 각각의 소수 대립 유전자 카운트 값에 대한 평균값 사이에 선형 회귀 모델이 구축될 수 있다.

[0144] 대조적으로, 엑솜 서열 분석을 통해 획득된 희귀 변이체의 경우, 개별 변이체별 검정은 각각의 희귀 변이체에 대한 발생률이 너무 낮기 때문에 유용하지 않다. 희귀 변이체에 대한 유전적 연관성을 검정하기 위해, 차트(122)에서 입증된 바와 같이 특정 유전자 내의 집계된 변이체에 대해 부담 검정(722)이 수행된다. 단일 공통 변이체 유전적 연관성 검정(702)과 비교하여, 집계된 희귀 변이체 부담 검정(722)은 유전자 내의 적어도 하나의 희귀 변이체에 대한 보인자 상태(여기서 상자 플롯(724) 및 (726)은 각각 보인자 및 비보인자에 대한 표현형 값 분포에 대한 것이다) 및 관심 표현형 사이의 연관성 강도를 측정한다. 희귀성을 필터링하는 것에 더하여, 병원성 측정 임계값에 의해 결정되는 유해한 변이체만 포함된다.

[0145] 개시된 기술의 일부 구현예에서, 관심 표현형 반응은 범주형 임상 진단이다. 따라서, 유전자에 대한 결과적이고 비무작위적인 연관성에 대한 유효 강도 점수는 보인자 상태에 대한 베타 계수로서 범주형 임상 진단 표현형에 대해 결정되며, 베타 계수는 부담 점수의 로지스틱 회귀 성분을 사용하여 결정된다.

[0146] *희귀 변이체 부담 검정의 최적화*

[0147] 논의는 이제 개별적으로 결정된 희귀성 임계값(즉, 최대 대립 유전자 카운트)과 특정 유전자에 대한 병원성 점수 임계값의 개념을 재검토한다. 이전 도면들은 특정 표현형에서의 특정 유전자에 대한 특정 복수의 희귀 유해 변이체의 유전적 연관성 검정을 설명했으며, 최대 희귀성 임계값 및 최소 병원성 점수 임계값은 특정 유전자에 특이적이며 다른 특정 유전자에 대한 최대 희귀성 임계값 및 최소 병원성 점수 임계값과 다를 수 있다.

[0148] 도 8에 대해 아래에서 설명되는 구현예에서, 표적 목표가 특정 유전자에 대해 가장 유의한 t-검정 통계량을 생성하는 매개변수 조합을 식별하는 것이기 때문에, "최적화(optimization)"는 희귀성에 대한 최대 임계값과 병원성에 대한 최소 점수를 식별하는 것을 의미한다. 그러나, 다른 구현예에서, 유전적 의학적 맥락으로 인해 최적화에 대한 표적 목표는 다를 수 있다. 당업자는 검정 통계량(예를 들어, p-값)의 해석이 상황에 따라 다르므로 유의성과 연관된 목표 값 또는 목표 값의 범위가 표본 크기, 통계적 검정력, 유전자별 또는 변이체별 효과 크기에서의 변화, 오류 분산, 실제 효과의 기본 비율 및 가능한 부담 검정 결과와 연관된 페이오프(payoff) 계산과 같은 다양한 매개변수에 의해 영향을 받는다는 것을 인식할 것이다.

[0149] 본 출원은 "희귀성 임계값(rarity threshold)", "최대 대립 유전자 카운트(maximum allele count)", "대립 유전자 카운트 임계값(allele count threshold)" 및 "AC"라는 용어를 상호 교환 가능하게 사용한다는 점에 유의한

다. 또한 본 출원은 "병원성 임계값(pathogenicity threshold)", "병원성 점수 임계값(pathogenicity score threshold)", "최소 병원성 점수 임계값(minimum pathogenicity score threshold)" 및 "PST"라는 용어를 상호 교환 가능하게 사용한다는 점에 유의한다.

[0150] 이제 논의는 가장 유의한 p-값을 갖는 대립 유전자 카운트와 병원성 점수 임계값의 특정 조합을 결정하고 특정 조합을 최적 조합으로서 사용하기 위한 개시된 기술로 넘어간다. 대립 유전자 카운트 및 병원성 점수 임계값의 복수의 조합 내의 대립 유전자 카운트 및 병원성 점수 임계값의 각각의 조합에 대한 각각의 p-값은 다중 검정(즉, 대립 유전자 카운트 및 병원성 점수 임계값의 복수의 조합에 걸친 그리드 검색 내에 수행된 특정 표현형에서의 특정 유전자에 대한 특정 복수의 희귀 유해 변이체의 유전적 연관성 검정을 위한 t-검정)을 처리하기 위해 보정된다.

[0151] 도 8은 희귀 변이체 부담 검정을 최적화하기 위한 방법의 개략도이다. 데이터베이스(802)는 코호트 i에 속하는 개체 그룹에 대응하는 게놈 및 표현형 데이터를 포함한다. 코호트 i(804)는 복수의 표현형 측정값에 대한 유전자 서열 분석 및 검정을 거친 N명의 개체로 구성된다. 코호트 i(804) 내의 개체에 대한 서열 분석 데이터는 복수의 변이체 보인자 상태에 대응하는 데이터를 포함한다. 특정 유전자는 3 가지 알려진 가능성 있는 희귀 유해 변이체(예를 들어, 변이체 A, 변이체 B 및 변이체 C)를 보유할 수 있으며, 개체는 보인자(즉, 개체는 해당 변이체를 보유하므로 그 변이체의 보인자가 된다) 또는 비보인자(즉, 개체는 해당 변이체를 보유하지 않으므로 그 변이체의 보인자가 아니다)가 될 수 있다. 당업자는 변이체 A, 변이체 B 및 변이체 C가 예시적인 예로서 주어지고 유전자가 임의의 수의 희귀 유해 변이체를 보유할 수 있다는 것을 인식할 것이다. 개시된 기술은 변이체를 모집단 내의 변이체 발생률이 사전 결정된 임계값 미만인 경우 '희귀(rare)'로서 정의하고 유전자가 단백질로 발현될 때 변이체의 병원성이 사전 결정된 임계값을 초과하는 측정된 영향을 가질 것으로 예측되는 경우 '유해(deleterious)'로서 정의한다. 희귀성과 병원성 모두에 대한 각각의 사전 결정된 임계값은 유전자 특이적이다(즉, 각각의 개별 유전자는 다른 유전자와 다를 수 있는 개별적으로 결정된 희귀성 임계값 및 병원성 임계값에 대응할 것이다).

[0152] 코호트 i(804) 내의 각각의 개체에 대한 게놈 및 엑솜 서열 분석으로부터 획득된 게놈 데이터 외에도, 표현형 데이터도 각각의 개체에 대해 제공된다(즉, 개체 N은 표현형 D에 대해 x_n의 측정된 값을 갖는다).

[0153] 희귀성 및 병원성에 대한 최적의 임계값을 결정하기 위해, 모든 가능한 병원성 점수 임계값(PST) 및 최대 대립 유전자 카운트 임계값(AC)을 포함하는 공간을 검색하기 위해 그리드 검색(822)이 수행된다. 그리드(822) 내에서, 각각의 개별 검정(PST 값 m 및 AC 값 N에 대응)은 t-검정(824)이다. 각각의 t-검정(824)에 대해, 귀무 가설은 변이체 그룹에 대해 보인자인 개체에 대한 표현형 값과 변이체 그룹에 대해 보인자가 아닌 개체에 대한 표현형 값 사이의 차이는 통계적으로 유의하지 않다고 명시한다. 대립 가설은 변이체 그룹에 대해 보인자인 개체에 대한 표현형 값과 변이체 그룹에 대해 보인자가 아닌 개체에 대한 표현형 값 사이의 차이는 통계적으로 유의하다고 명시한다. t-검정은 PST 값 및 AC 값의 각각의 가능한 조합에 대해 전체 그리드에 걸쳐 반복적으로 수행되어 균일한 유의 수준에서 각각의 t-검정에 대한 p-값을 획득한다. PST 값 및 AC 값의 복수의 조합 및 연관된 p-값(842) 내에서, 하위 코호트(PST_m, AC_n)(844)는 특정된 PST 값 m 및 AC 값 N을 충족하는 N명의 개체에 대한 게놈 및 표현형 데이터를 포함한다(즉, 변이체 B가 특정 t-검정에 대해 특정된 병원성 임계값 또는 대립 유전자 카운트를 충족하지 않는 경우, 변이체 B에 대한 보인자 상태는 집계된 이진 보인자 상태 변수 내에서 고려되지 않는다).

[0154] 가장 유의한 p-값에 대응하는 PST 및 AC의 조합은 특정 유전자에 대한 최적의 PST 및 AC 조합(862)으로 출력된다. 개시된 기술의 일부 구현예에서, 결정된 최적의 PST 및 AC 조합(862)에서의 최적화 후에 희귀 변이체 부담 검정이 수행된다. 따라서, 최적화된 부담 검정을 수행하는 방법은, 특정 유전자에 대한 모든 최적화된 부담 검정이 시작되어 특정 표현형에 대해 특정 유전자 내의 희귀 병원성 변이체의 부담 검정 효과의 유의성을 최대화하는 최대 대립 유전자 카운트 및 최소 병원성 점수 임계값의 최적 조합이 특정 유전자에 대한 부담 검정을 위한 매개변수로서 사용되도록 수행된다.

[0155] **단백질 구조 기반 병원성 결정**

[0156] 지금까지의 논의는 개체가 각각의 유전자에서 희귀 유해 변이체를 가지고 있는지 여부에 기초하여 유전자 전반에 걸쳐 위험을 집계하도록 구성된 희귀 변이체 다유전자 위험 점수 모델을 설명하였다. 도 1, 도 6, 도 7 및 도 8에 설명된 희귀 변이체 다유전자 위험 점수 모델은 특정 표현형에 대해 특정 유전자 내의 희귀 병원성 변이체의 부담 검정 효과의 유의성을 최대화하는 최대 대립 유전자 카운트 및 최소 병원성 점수 임계값의 최적 조합

에서 수행된다.

[0157] 이제 논의는 단백질 구조 기반 병원성 결정을 위한 시스템에 대한 설명으로 넘어가며, 개시된 기술의 한 구현예는 변이체의 병원성을 결정하도록 구성된 컨볼루션 신경망 병원성 분류자를 구현한다. 특정 유전자에 대한 특정 변이체에 대응하는 병원성 점수 출력은 특정 t-검정(824)에서 특정 유전자에 대한 병원성 점수 임계값으로서 구현될 수 있다. 복수의 m개 병원성 점수 임계값은 최적 병원성 점수 임계값(862)과 같은 최적의 최소 병원성 점수 임계값을 결정하기 위해 그리드 검색(822) 내에서 검정될 수 있다.

[0158] 도 9는 변이체의 병원성을 결정하기 위한 시스템의 프로세스를 예시하는 흐름도(900)이다. 단계 902에서, 시스템의 서열 접근자(904)가 기준 및 대안 아미노산 서열에 접근한다. 912에서, 시스템의 3D 구조 생성자(914)가 기준 아미노산 서열에 대한 3D 단백질 구조를 생성한다. 일부 구현예에서, 3D 단백질 구조는 인간 단백질의 상동성 모델이다. 하나의 구현예에서, 이른바 SwissModel 상동성 모델링 파이프라인이 예측된 인간 단백질 구조의 공개 리포지토리를 제공한다. 다른 구현예에서, 이른바 HHpred 상동성 모델링이 Modeller로 불리는 툴을 사용하여 주형 구조로부터 표적 단백질의 구조를 예측한다.

[0159] 단백질은 3D 공간에서 원자의 집합 및 이의 좌표로 표현된다. 아미노산은 탄소 원자, 산소(O) 원자, 질소(N) 원자 및 수소(H) 원자와 같은 다양한 원자를 가질 수 있다. 원자는 측쇄 원자 및 백본(backbone) 원자로서 추가로 분류될 수 있다. 백본 탄소 원자는 알파-탄소(Cα) 원자 및 베타-탄소(Cβ) 원자를 포함할 수 있다.

[0160] 단계 922에서, 시스템의 좌표 분류자(924)가 아미노산 기준으로 3D 단백질 구조의 3D 원자 좌표를 분류한다. 하나의 구현예에서, 아미노산별 분류는 3D 원자 좌표를 21개의 아미노산 카테고리(정지 또는 갭 아미노산 카테고리를 포함함)에 귀속시키는 것을 수반한다. 하나의 예에서, 알파-탄소 원자의 아미노산별 분류는 21개의 아미노산 카테고리 각각 하에 알파-탄소 원자를 각각 열거할 수 있다. 다른 예에서, β-탄소 원자의 아미노산별 분류는 21개의 아미노산 카테고리 각각 하에 β-탄소 원자를 각각 열거할 수 있다.

[0161] 또 다른 예에서, 산소 원자의 아미노산별 분류는 21개의 아미노산 카테고리 각각 하에 산소 원자를 각각 열거할 수 있다. 또 다른 예에서, 질소 원자의 아미노산별 분류는 21개의 아미노산 카테고리 각각 하에 질소 원자를 각각 열거할 수 있다. 또 다른 예에서, 수소 원자의 아미노산별 분류는 21개의 아미노산 카테고리 각각 하에 수소 원자를 각각 열거할 수 있다.

[0162] 당업자는, 다양한 구현예에서, 아미노산별 분류가 21개의 아미노산 카테고리의 서브세트 및 상이한 원자 원소의 서브세트를 포함할 수 있다는 것을 이해할 것이다.

[0163] 본 출원은 복셀(voxel)과 복셀화(voxelization)를 비제한적인 예로 사용한다. 당업자는, 다양한 구현예에서, 다양한 차원의 특징부, 벡터, 어레이와 같은 데이터를 배열하고 처리하는 다양한 포맷이 대안으로 또는 조합으로 사용될 수 있다는 것을 이해할 것이다.

[0164] 단계 932에서, 시스템의 복셀 그리드 생성자(934)가 복셀 그리드를 인스턴스화한다. 복셀 그리드는 임의의 해상도, 예를 들어 3x3x3, 5x5x5, 7x7x7 등을 가질 수 있다. 복셀 그리드 내의 복셀은, 예를 들어 각각의 측에서 1 옹스트롬(Å), 각각의 측에서 2 Å, 각각의 측에서 3 Å 등등인, 임의의 크기를 가질 수 있다. 당업자는, 복셀이 정육면체이기 때문에 이러한 예시적인 차원이 입방 차원을 지칭한다는 것을 이해할 것이다. 또한, 당업자는, 이러한 예시적인 차원이 비제한적이고 복셀이 임의의 입방 차원을 가질 수 있다는 것을 이해할 것이다.

[0165] 단계 942에서, 시스템의 복셀 그리드 센터러(944)는 아미노산 수준에서 표적 변이체를 경험하는 기준 아미노산에 복셀 그리드의 중심을 설정한다. 하나의 구현예에서, 복셀 그리드는 표적 변이체를 경험하는 기준 아미노산의 특정 원자의 원자 좌표, 예를 들어, 표적 변이체를 경험하는 기준 아미노산의 알파-탄소 원자의 3D 원자 좌표에 중심이 설정된다.

[0166] **<u>거리 채널</u>**

[0167] 복셀 그리드 내의 복셀은 복수의 채널(또는 특징부)을 가질 수 있다. 하나의 구현예에서, 복셀 그리드 내의 복셀은 복수의 거리 채널(예를 들어, 각각, 21개의 아미노산 카테고리(정지 또는 갭 아미노산 카테고리를 포함함)에 대한 21개의 거리 채널)을 갖는다. 단계 952에서, 시스템의 거리 채널 생성자(954)는 복셀 그리드 내의 복셀에 대한 아미노산별 거리 채널을 생성한다. 거리 채널은 21개의 아미노산 카테고리 각각에 대해 독립적으로 생성된다.

[0168] 예를 들어, 알라닌(A) 아미노산 카테고리를 고려한다. 예를 들어, 복셀 그리드가 크기 3x3x3를 갖고 27개의 복셀을 갖는다는 것을 추가로 고려한다. 이어서, 하나의 구현예에서, 알라닌 거리 채널이 복셀 그리드 내의 27개

의 복셀에 대한 27개의 거리 값을 각각 포함한다. 알라닌 거리 채널에서의 27개의 거리 값은 복셀 그리드 내의 27개의 복셀의 각자의 중심으로부터 알라닌 아미노산 카테고리 내의 각자의 가장 가까운 원자까지 측정된다.

[0169] 하나의 예에서, 알라닌 아미노산 카테고리는 알파-탄소 원자만을 포함하고, 따라서, 가장 가까운 원자는 각각 복셀 그리드 내의 27개의 복셀에 가장 근접한 이러한 알라닌 알파-탄소 원자이다. 다른 예에서, 알라닌 아미노산 카테고리는 β-탄소 원자만을 포함하고, 따라서, 가장 가까운 원자는 각각 복셀 그리드 내의 27개의 복셀에 가장 근접한 이러한 알라닌 β-탄소 원자이다.

[0170] 또 다른 예에서, 알라닌 아미노산 카테고리는 산소 원자만을 포함하고, 따라서, 가장 가까운 원자는 각각 복셀 그리드 내의 27개의 복셀에 가장 근접한 이러한 알라닌 산소 원자이다. 또 다른 예에서, 알라닌 아미노산 카테고리는 질소 원자만을 포함하고, 따라서, 가장 가까운 원자는 각각 복셀 그리드 내의 27개의 복셀에 가장 근접한 이러한 알라닌 질소 원자이다. 또 다른 예에서, 알라닌 아미노산 카테고리는 수소 원자만을 포함하고, 따라서, 가장 가까운 원자는 각각 복셀 그리드 내의 27개의 복셀에 가장 근접한 이러한 알라닌 수소 원자이다.

[0171] 알라닌 거리 채널과 마찬가지로, 거리 채널 생성자(954)는 나머지 아미노산 카테고리 각각에 대한 거리 채널(즉, 복셀별 거리 값의 세트)을 생성한다. 다른 구현예에서, 거리 채널 생성자(954)는 21개의 아미노산 카테고리의 서브세트에 대해서만 거리 채널을 생성한다.

[0172] 다른 구현예에서, 가장 가까운 원자의 선택은 특정 원자 유형으로 한정되지 않는다. 즉, 대상 아미노산 카테고리 내에서, 특정 복셀에 대해 가장 가까운 원자가, 가장 가까운 원자의 원자 원소 및 대상 아미노산 카테고리에 대한 거리 채널에 포함시키기 위해 계산된 특정 복셀에 대한 거리 값과 관계없이 선택된다.

[0173] 또 다른 구현예에서, 거리 채널은 원자 원소 기준으로 생성된다. 아미노산 카테고리에 대한 거리 채널을 갖는 대신에 또는 그에 더하여, 원자가 속하는 아미노산에 관계없이 원자 원소 카테고리에 대해 거리 값이 생성될 수 있다. 예를 들어, 기준 아미노산 서열 내의 아미노산의 원자는 7개의 원자 원소, 즉 탄소, 산소, 질소, 수소, 칼슘, 요오드 및 황에 걸쳐 있다는 것을 고려한다. 이어서, 복셀 그리드 내의 복셀은 7개의 거리 채널을 갖도록 구성되어, 7개의 거리 채널 각각은 대응하는 원자 원소 카테고리 내의 가장 가까운 원자까지만의 거리를 특정하는 27개의 복셀별 거리 값을 갖도록 한다. 다른 구현예에서, 7개의 원자 원소의 서브세트만을 위한 거리 채널이 생성될 수 있다. 또 다른 구현예에서, 원자 원소 카테고리 및 거리 채널 생성은 동일한 원자 원소, 예를 들어, 알파-탄소(C_α) 원자 및 β-탄소(C_β) 원자의 변이체로 추가로 계층화될 수 있다.

[0174] 또 다른 구현예에서, 거리 채널은 원자 유형 기준으로 생성되며, 예를 들어, 측쇄 원자에 대해서만의 거리 채널 및 백본 원자에 대해서만의 거리 채널이 생성될 수 있다.

[0175] 가장 가까운 원자는 복셀 중심으로부터 사전 정의된 최대 스캔 반경(예를 들어, 6 옹스트롬(Å)) 내에서 검색될 수 있다. 또한, 다수의 원자가 복셀 그리드 내의 동일한 복셀에 가장 가까울 수 있다.

[0176] 거리는 복셀 중심의 3D 좌표와 원자의 3D 원자 좌표 사이에서 계산된다. 또한, 거리 채널은 동일한 위치에 중심이 설정된(예를 들어, 표적 변이체를 경험하는 기준 아미노산의 알파-탄소 원자의 3D 원자 좌표에 중심이 설정된) 복셀 그리드로 생성된다.

[0177] 거리는 유클리드 거리일 수 있다. 또한, 거리는 (예를 들어, 논의 대상 원자의 Lennard-Jones 전위 및/또는 Van der Waals 원자 반경을 사용하여) 원자 크기(또는 원자 영향)에 의해 매개변수화될 수 있다. 또한, 거리 값은 최대 스캔 반경에 의해, 또는 대상 아미노산 카테고리 또는 대상 원자 원소 카테고리 또는 대상 원자 유형 카테고리 내의 최대한 가장 가까운 원자의 최대 관찰된 거리 값에 의해 정규화될 수 있다. 일부 구현예에서, 복셀과 원자 사이의 거리는 복셀 및 원자의 극좌표에 기초하여 계산된다. 극좌표는 복셀과 원자 사이의 각도에 의해 매개변수화된다. 하나의 구현예에서, 이러한 각도 정보는 복셀에 대한 각도 채널을 생성하는 데 사용된다(즉, 거리 채널로부터 독립적임). 일부 구현예에서, 가장 가까운 원자와 이웃 원자(예를 들어, 백본 원자) 사이의 각도는 복셀로 인코딩되는 특징부로서 사용될 수 있다.

[0178] **<u>기준 대립 유전자 및 대안 대립 유전자 채널</u>**

[0179] 복셀 그리드 내의 복셀은 또한 기준 대립 유전자 및 대안 대립 유전자 채널을 가질 수 있다. 단계 962에서, 시스템의 원-핫 인코더(964)가 기준 아미노산 서열 내의 기준 아미노산의 기준 원-핫 인코딩 및 대안 아미노산 서열 내의 대안 아미노산의 대안 원-핫 인코딩을 생성한다. 기준 아미노산은 표적 변이체를 경험한다. 대안 아미노산은 표적 변이체이다. 기준 아미노산 및 대안 아미노산은 기준 아미노산 서열 및 대안 아미노산 서열에서 각각 동일한 위치에 위치된다. 기준 아미노산 서열 및 대안 아미노산 서열은 하나의 예외를 갖는 동일한 위치별

아미노산 조성을 갖는다. 예외는, 기준 아미노산 서열에서는 기준 아미노산을 갖고 대안 아미노산 서열에서는 대안 아미노산을 갖는 위치이다.

[0180] 단계 972에서, 시스템의 연결자(974)가 아미노산별 거리 채널과 기준 및 대안 원-핫 인코딩을 연결한다. 다른 구현예에서, 연결자(974)는 원자 원소별 거리 채널과 기준 및 대안 원-핫 인코딩을 연결한다. 또 다른 구현예에서, 연결자(974)는 원자 유형별 거리 채널과 기준 및 대안 원-핫 인코딩을 연결한다.

[0181] 단계 982에서, 시스템의 런타임 로직(984)은 병원성 분류자(병원성 결정 엔진)를 통해, 연결된 아미노산별/원자 원소별/원자 유형별 거리 채널과 기준 및 대안 원-핫 인코딩을 처리하여 표적 변이체의 병원성을 결정하며, 이는 결국, 아미노산 수준에서 표적 변이체를 생성하는 기본 뉴클레오티드 변이체의 병원성 결정으로서 추론된다. 병원성 분류자는 양성 및 병원성 변이체의 라벨링된 데이터세트를 사용하여, 예를 들어 역전파 알고리즘을 사용하여, 훈련된다. 양성 및 병원성 변이체의 라벨링된 데이터세트 및 병원성 분류자의 예시적인 아키텍처 및 훈련에 관한 추가적인 세부 사항은 공동 소유의 미국 특허 출원 제16/160,903호; 제16/160,986호; 제16/160,968호; 및 제16/407,149호에서 찾을 수 있다.

[0182] 도 10은 개시된 기술의 하나의 구현예에 따른, 병원성 분류자(900)의 예시적인 처리 아키텍처(1000)를 도시한다. 처리 아키텍처(1000)는 처리 모듈(1006, 1010, 1014, 1018, 1022, 1010, 1030, 1034, 1038, 1042)의 캐스케이드를 포함하며, 이들 각각은 1D 컨볼루션(1x1x1 CONV), 3D 컨볼루션(3x3x3 CONV), ReLU 비선형성 및 배치 정규화(BN: batch normalization)를 포함할 수 있다. 처리 모듈의 다른 예는 완전 연결(FC: fully-connected) 계층, 드롭아웃 계층, 평탄화 계층, 및 양성 클래스 및 병원성 클래스에 속하는 표적 변이체에 대한 지수적으로 정규화된 점수를 생성하는 최종 소프트맥스 계층을 포함한다. 도 10에서, "64"는 특정 처리 모듈에 의해 적용된 컨볼루션 필터의 수를 표시한다. 도 10에서, 입력 복셀(1002)의 크기는 15x15x15x8이다. 도 10은 또한 처리 아키텍처(1000)에 의해 생성된 중간 입력(1004, 1008, 1012, 1016, 1020, 1024, 1028, 1032, 1036, 1040)의 각자의 체적 차원수를 도시한다.

[0183] 도 11은 개시된 기술을 구현하는 데 사용될 수 있는 예시적인 컴퓨터 시스템(1100)을 도시한다. 컴퓨터 시스템(1100)은 버스 서브시스템(1142)을 통해 다수의 주변 디바이스와 통신하는 적어도 하나의 중앙 처리 장치(CPU)(1144)를 포함한다. 이러한 주변 디바이스는, 예를 들어, 메모리 디바이스 및 파일 저장 서브시스템(1136)을 포함하는 저장 서브시스템(1110), 사용자 인터페이스 입력 디바이스(1138), 사용자 인터페이스 출력 디바이스(1148) 및 네트워크 인터페이스 서브시스템(1146)을 포함할 수 있다. 입력 및 출력 디바이스는 컴퓨터 시스템(1100)과의 사용자 상호 작용을 가능하게 한다. 네트워크 인터페이스 서브시스템(1146)은 다른 컴퓨터 시스템에서의 대응하는 인터페이스 디바이스에 대한 인터페이스를 포함하는 인터페이스를 외부 네트워크에 제공한다.

[0184] 하나의 구현예에서, 병원성 분류자(1000)는 저장 서브시스템(1110) 및 사용자 인터페이스 입력 디바이스(1138)에 통신 가능하게 연결된다.

[0185] 사용자 인터페이스 입력 디바이스(1138)는, 키보드; 마우스, 트랙볼, 터치패드 또는 그래픽 태블릿과 같은 포인팅 디바이스; 스캐너; 디스플레이 내에 통합된 터치 스크린; 음성 인식 시스템 및 마이크와 같은 오디오 입력 디바이스; 및 다른 유형의 입력 디바이스를 포함할 수 있다. 일반적으로, "입력 디바이스"라는 용어의 사용은 정보를 컴퓨터 시스템(1100)에 입력하기 위한 모든 가능한 유형의 디바이스 및 방식을 포함하도록 의도된다.

[0186] 사용자 인터페이스 출력 디바이스(1148)는 디스플레이 서브시스템, 프린터, 팩스 기계 또는 오디오 출력 디바이스와 같은 비시각적 디스플레이를 포함할 수 있다. 디스플레이 서브시스템은 LED 디스플레이, 음극선관(CRT), 액정 디스플레이(LCD)와 같은 평면 디바이스, 프로젝션 디바이스 또는 가시적인 이미지를 생성하기 위한 일부 다른 메커니즘을 포함할 수 있다. 디스플레이 서브시스템은 또한 오디오 출력 디바이스와 같은 비시각적 디스플레이를 제공할 수 있다. 일반적으로, "출력 디바이스"라는 용어의 사용은 정보를 컴퓨터 시스템(1100)으로부터 사용자에게 또는 다른 기계 또는 컴퓨터 시스템에 출력하기 위한 모든 가능한 유형의 디바이스 및 방식을 포함하도록 의도된다.

[0187] 저장 서브시스템(1110)은 본원에 설명된 모듈 및 방법 중 일부 또는 전부의 기능을 제공하는 프로그래밍 및 데이터 구성을 저장한다. 이러한 소프트웨어 모듈은 일반적으로 프로세서(1148)에 의해 실행된다.

[0188] 프로세서(1148)는 그래픽 처리 장치(GPU), 필드 프로그래밍 가능 게이트 어레이(FPGA), 주문형 반도체(ASIC) 및/또는 CGRA(coarse-grained reconfigurable architecture)일 수 있다. 프로세서(1148)는 Google Cloud Platform™, Xilinx™ 및 Cirrascale™과 같은 심층 학습 클라우드 플랫폼에 의해 호스팅될 수 있다. 프로세서(1148)의 예는 Google의 Tensor Processing Unit(TPU)™, GX4 Rackmount Series™, GX11 Rackmount Series™

와 같은 랙마운트 솔루션, NVIDIA DGX-1™, Microsoft의 Stratix V FPGA™, Graphcore의 Intelligent Processor Unit(IPU)™, Snapdragon processors™을 갖는 Qualcomm의 Zeroth Platform™, NVIDIA의 Volta™, NVIDIA의 DRIVE PX™, NVIDIA의 JETSON TX1/TX11 MODULE™, Intel의 Nirvana™, Movidius VPU™, Fujitsu DPI™, ARM의 DynamicIQ™, IBM TrueNorth™, Testa V100s™를 갖는 Lambda GPU 서버 등을 포함한다.

[0189] 저장 서브시스템(1110)에 사용되는 메모리 서브시스템(1122)은 프로그램 실행 동안의 명령어 및 데이터의 저장을 위한 메인 랜덤 액세스 메모리(RAM)(1132) 및 고정된 명령어가 저장되는 판독 전용 메모리(ROM)(1134)를 포함하는 다수의 메모리를 포함할 수 있다. 파일 저장 서브시스템(1136)은 프로그램 및 데이터 파일을 위한 영구 저장소를 제공할 수 있고, 하드 디스크 드라이브, 연관된 탈착식 매체와 함께하는 플로피 디스크 드라이브, CD-ROM 드라이브, 광학 드라이브 또는 탈착식 매체 카트리지를 포함할 수 있다. 특정 구현예의 기능을 구현하는 모듈은 저장 서브시스템(1110) 내의 파일 저장 서브시스템(1136)에 의해, 또는 프로세서에 의해 접근 가능한 다른 기계에 저장될 수 있다.

[0190] 버스 서브시스템(1140)은 컴퓨터 시스템(1100)의 다양한 컴포넌트 및 서브시스템이 의도된 바와 같이 서로 통신하게 하기 위한 메커니즘을 제공한다. 버스 서브시스템(1142)이 개략적으로 단일 버스로서 도시되어 있지만, 버스 서브시스템의 대안적인 구현예는 다수의 버스를 사용할 수 있다.

[0191] 컴퓨터 시스템(1100) 자체는 개인용 컴퓨터, 휴대용 컴퓨터, 워크스테이션, 컴퓨터 단말기, 네트워크 컴퓨터, 텔레비전, 메인프레임, 서버 팜, 느슨하게 네트워킹된 컴퓨터의 광범위하게 분포된 세트 또는 임의의 다른 데이터 처리 시스템 또는 사용자 디바이스를 포함한 다양한 유형의 것일 수 있다. 컴퓨터 및 네트워크의 지속적으로 변화하는 특성으로 인해, 도 11에 묘사된 컴퓨터 시스템(1100)의 설명은 본 발명의 바람직한 구현예를 예시하기 위한 특정 예로서만 의도된다. 도 11에 묘사된 컴퓨터 시스템보다 더 많은 또는 더 적은 컴포넌트를 갖는 컴퓨터 시스템(1100)의 많은 다른 구성이 가능하다.

[0192] 도 12는 최종 병원성 점수를 결정하는 하나의 구현예(1200)를 예시한다. 동작 1202에서, 하나의 구현예에서, 병원성 분류자(1000)는 제1 기준 아미노산과 동일한 제1 대안 아미노산에 대한 제1 병원성 점수를 생성한다. 동작 1212에서, 하나의 구현예에서, 병원성 분류자(1000)는 제1 기준 아미노산과 다른 제2 대안 아미노산에 대한 제2 병원성 점수를 생성한다. 동작 1222에서, 하나의 구현예에서, 제2 대안 아미노산에 대한 최종 병원성 점수는 제2 대안 아미노산에 대한 제2 병원성 점수이다.

[0193] 다른 대안에서, 제2 대안 아미노산에 대한 최종 병원성 점수는 제1 병원성 점수와 제2 병원성 점수의 조합에 기초한다. 1222a에서의 제1 대안에서, 하나의 구현예에서, 제2 대안 아미노산에 대한 최종 병원성 점수는 제1 병원성 점수와 제2 병원성 점수의 합에 대한 제2 병원성 점수의 비이다. 1222b에서의 제2 대안에서, 하나의 구현예에서, 제2 대안 아미노산에 대한 최종 병원성 점수는 제2 병원성 점수에서 제1 병원성 점수를 빼서 결정된다.

[0194] 당업자는 다른 현재 및 미래의 인공 지능, 기계 학습, 및 심층 학습 모델, 데이터세트 및 훈련 기술이 개시된 기술의 사상으로부터 벗어나지 않고 개시된 변이체 병원성 분류자에 통합될 수 있다는 것을 이해할 것이다.

[0195] **다중 검정에 대한 보정**

[0196] 지금까지의 논의는, 연관성 강도가 표현형과 연관된 유전자와 표현형 반응에 대한 희귀 변이체의 기여도 사이에 정량화되고 가장 유의한 *p*-값을 갖는 대립 유전자 카운트와 병원성 점수 임계값의 특정 조합에 최적화되는, 최적화된 부담 검정을 수행하는 방법을 포함하는 개시된 기술의 구현예를 설명했다. 최적화된 부담 검정은 특정 유전자 내에서 적어도 하나의 희귀 유해 변이체의 보인자와 비보인자 사이의 분리를 최대화하는 복수의 네스팅된 *t*-검정에 기초한다. 개시된 기술의 이러한 구현예에서, 최적화 방법 내에서 수행된 다중 *t*-검정은 각각의 유전자 내에서 다중 검정을 보정할 필요성을 보증한다.

[0197] 이제 PST 값 및 AC 값의 최적 조합에 대응하는 가장 유의한 *p*-값을 수정하는 것에 대한 설명으로 넘어간다.

[0198] 도 13은 각각의 유전자 내의 다수의 검정을 보정하기 위한 프로세스의 흐름도(1300)이다. 오류 발견율(FDR: false discovery rate) 보정에 이어, PST 및 AC 모두에 대한 부담 검정 최적화를 처리하기 위해 다중 검정에 대해 *p* 값이 추가로 보정된다. FDR 보정 *p*-값(1302)은 값 범위로 분리된다. 개시된 기술의 하나의 구현예에서, 다중 검정 보정(1300)을 위한 다음의 알고리즘이 이어진다. 개시된 기술의 다른 구현예에서는, 다른 형태의 다중 검정 보정이 사용될 수 있다. 범위 (0, 1e-5] 내의 값에 대해, FDR 보정 *p*-값(1302)은 Benjamini-Hochberg FDR 보정(1310)을 거친다.

[0199] 범위 (0.01, 1] 내의 값에 대해, FDR 보정 p-값(1302)은 순열 검정 단계 1312를 거친다. 단계 1312에서, 하나

의 구현예에서, 생성된 순열의 총 수의 카운트(N)는 0으로 설정된다. 다른 구현예에서, N은 임의의 사전 정의된 시작 값으로 설정될 수 있다. 다른 구현예에서, N은 임의의 사전 정의된 시작 값으로 설정될 수 있다. 단계 1312에 이어서, 단계 1322는 표현형 라벨의 1,000개 순열의 생성을 포함하고, N을 N + 1000으로 설정한다. 다음 단계 1342에서, 데이터의 각각의 순열에 대해 부담 검정이 수행되고, 원래 관찰된 데이터보다 더 유의한 p-값을 산출하는 순열의 비율이 p로 카운트된다. N이 100/p 보다 크면 단계 1362로 진행한다. 단계 1362에서, 중지하고 p를 출력한다. 단계 1362에서의 중지 지점은 N<10,000을 보장하여 계산 효율성을 유지한다. 즉, 순열의 수(N)가 10,000개 미만으로 제한되어 계산이 불가능하거나 계산이 비효율적이거나 계산 비용이 많이 드는 카운트에 도달하는 것을 방지한다.

[0200] 범위 (1e-5, 0.01] 내의 값에 대해, FDR 보정 p-값(1302)은 순열 검정 단계 1314를 거친다. 단계 1314에서, 표현형 라벨의 10,000개의 순열이 생성된다. 단계 1314에 이어서, 단계 1324는 순열된 데이터로부터의 각각의 해당하는 부담 검정에 대한 검정 통계의 절대값에 대한 일반화된 극단값 분포의 피팅을 포함한다. 다음 단계 1344에서, 보정된 p-값이 피팅된 분포의 곡선 아래 영역으로부터 추정된다. 즉, 단계 1362에서, 초기 p-값이 출력되고, 단계 1344에서, 보정된 p-값이 추정된다.

[0201] **공변량에 대한 표현형 값의 보정**

[0202] 지금까지 논의는 표현형과 연관된 유전자의 연관성 강도 및 표현형 반응에 대한 희귀 변이체의 기여도를 정량화하는 방법을 다루었으며, 여기서 희귀 다유전자 위험 점수에 기여하는 복수의 유전자의 유효 강도 점수는 부울 보인자 상태 변수에 의해 결정되고, 희귀 다유전자 위험 점수 모델(즉, 부담 검정)은 가장 유의한 p-값을 갖는 대립 유전자 카운트 및 표현형 점수 임계값의 특정 최적 조합에 대해 최적화된다. 이제 논의는 표현형 측정값의 개념을 재검토하여 표현형 측정값이 공변량 보정되고 약물 사용 보정되는 추가적인 최적화 방법을 소개한다.

[0203] 본 출원은 "약물 사용", "약물 사용 패턴", "약물 카테고리 사용" 및 "약물 카테고리 사용 패턴"이라는 용어를 상호 교환 가능하게 사용한다는 점에 유의한다. 또한 본 출원은 하나 이상의 다른 공변량 및 관심 결과 모두와 연관된 공변량을 설명하기 위해 "교란 변수(confounder)"라는 용어를 사용한다는 점에 유의한다(예를 들어, 담배 사용은 심혈관 질환에 대한 약물 사용 패턴 및 표현형 측정값과 연관된 교란 변수이다). 당업자는 개시된 기술이 종양 질환, 당뇨병 등과 같은 임의의 질환에 적용될 수 있다는 것을 이해할 것이다.

[0204] 도 14는 공변량에 대한 표현형 값을 보정하는 방법(1400)에 대한 개략도이다. 단계 1402에서, 정량적 표현형의 데이터세트는 중복되지 않은 데이터세트로 프루닝된다(pruned). 단계 1404에서, 복수의 관심 약물이 약물 카테고리 세트로 그룹화된다. 단계 1406에서, 중복되지 않은 정량적 표현형 데이터세트 내의 각각의 표현형은 약물 카테고리 세트 내의 각각의 약물 카테고리(예를 들어, 스타틴, 혈압 약 등)에 대한 약물 사용에 대해 보정된다. 단계 1408에서, 각각의 보정된 표현형은 역순위 정규 변환된다. 단계 1410에서, 보정된 표현형은 복수의 교란 변수(예를 들어, 연령, 성별, 유전적 주요 성분, 식이, 흡연 상태 등)에 대해 추가로 보정된다. 교란 변수와 관련하여, 교란 변수는 시험 약물과 결과 사이의 관계를 교란하는 요인이다. 일부 시험은 시험 동안의 보조 약물 또는 심리 치료와 같은 보조 치료를 허용할 수 있다. 결과가 검정되고 있는 약물보다는 부가제에 기인할 수 있기 때문에 이는 교란 변수이다. 한 예로서, 진통제 시험은 종종 부목, 크림, 마사지, 온욕 및 척추 지압사의 촉진(manipulation)과 같은 참가자의 비약물 진통제 사용을 무시한다. 이러한 추가 시험 치료에 대한 반응은 시험 결과를 상당히 교란할 수 있다. 일반적인 교란 변수는 참가자의 속성이고; 예를 들어, 체질량 지수, 흡연 상태, 질병 발병 연령, 사회 경제적 상태, 교육 상태 및 지원 네트워크 범위가 있다. 또한, 생활 사건(life event)도 잠재적인 교란 변수이다. 이는 참가자의 기분과 증상 수준에 큰 변화를 일으킬 수 있으며, 이는 참가자와 평가자 사이의 관계에서의 사건의 경우 아무리 작더라도 특히 그렇다. 실험군과 대조군 사이에 일치되지 않는 모든 알려진 교란 변수가 통계 분석에서 고려될 수 있다.

[0205] 이제 논의는 제1 시점 및 제2 시점에서 개체 코호트의 복수의 표현형에 대한 복수의 약물의 사용에 따라 표현형 시프트를 예측하도록 구성된 개시된 기술의 하나의 구현예에 대한 상세한 설명으로 넘어가고, 코호트 내의 개체는 두 시점에서 측정된 순열된 약물 사용 패턴(예를 들어, 제1 시점 및 제2 시점에 약물을 사용하지 않는 것, 제1 시점과 제2 시점 사이에 약물을 사용하기 시작하는 것, 제1 시점과 제2 시점 사이에 약물을 사용하는 것을 중단하는 것, 제1 시점 및 제2 시점에 약물을 사용하는 것)에 기초하여 그룹화된다. 생물 통계학 및 인구 유전학 연구에서 일반적으로 사용되는 공변량 보정을 위한 다른 전략과 달리, 단일 시점이 아닌 다수의 시점의 사용은 특정 교란 변수가 표현형 시프트에 미치는 영향에 대해 더 낮은 오류 분산 및 더 정확한 근사치를 제공한다.

[0206] 도 15는 복수의 표현형에 대한 복수의 약물 사용에 따라 표현형 시프트를 예측하는 것을 나타내는 개략도이다. 데이터베이스(1502)는 코호트 j에 속하는 개체 그룹에 대응하는 게놈 및 표현형 데이터를 포함한다. 단계 1522

에서, 복수의 교란 변수에 대한 공변량 측정값, 복수의 표현형에 대한 표현형 측정값 및 복수의 약물 카테고리에 대한 약물 사용 패턴이 데이터베이스(1502)로부터 두 시점 t_1 및 t_2에서 접근된다. 단계 1542, 1562, 1582 및 1592는 복수의 표현형 내의 각각의 표현형에 대해 수행된다. 당업자는 개시된 기술이 3, 4, 5 등의 임의의 수의 시점으로 확장될 수 있다는 것을 이해할 것이다.

[0207] 단계 1542에서, 표현형 측정값은 t_1 및 t_2 에서 복수의 교란 변수에 대해 공변량 보정되어, 제1 회귀 모델을 피팅함으로써 교란 변수 보정 값을 생성한다. 단계 1562에서, 제1 회귀 모델이 교란 변수 보정 값에 대한 델타(δ)를 결정하는 데 사용된다. 따라서 공변량 보정은 제1 회귀 모델을 피팅하여 공변량 측정값을 회귀시킴으로써 구현된다. 단계 1582 및 1592는 복수의 약물 카테고리 내의 각각의 약물 카테고리에 대해 수행된다. 단계 1582에서, δ는 제2 회귀 모델을 피팅함으로써 특정 약물 카테고리의 사용에 따른 표현형 시프트를 예측하는 데 사용된다. 단계 1592에서, 교란 변수 보정 값은 특정 약물 카테고리에 대한 약물 사용에 대해 추가로 보정되어 특정 표현형에 대한 교란 변수 보정되고 약물 카테고리 사용 보정된 값을 생성한다. 따라서, 표현형 시프트 예측은 각각의 약물 사용 패턴에 대한 표현형 시프트를 모델링하는 제2 회귀 모델을 피팅함으로써 구현된다. 제2 회귀 모델은 각각의 약물 사용 패턴에 대한 이진 지시자 변수를 가진다.

[0208] 도 16은 특정 코호트에 대한 약물 사용 패턴 및 표현형 데이터를 획득하기 위한 실험 설정을 나타내는 개략도이다. 데이터베이스(1602)는 코호트 j에 속하는 개체 그룹에 대응하는 게놈 및 표현형 데이터를 포함한다. 데이터베이스(1602)는 제1 시점 t_1 에서 특정 약물 카테고리 z를 사용하지 않는 개체(1622)와 특정 약물 카테고리 z를 사용하는 개체(1624)를 각각 포함하는 2개의 하위 코호트로 분할된다. 특정 약물 카테고리 z를 사용하지 않는 개체(1622)를 포함하는 하위 코호트는 특정 약물 카테고리 z를 계속 사용하지 않은 개체(1642)와 제1 시점 t_1 이후 제2 시점 t_2 이전에 특정 약물 카테고리 z를 사용하기 시작한 개체(1644)를 각각 포함하는 더 작은 하위 코호트로 더 분할될 수 있다.

[0209] 특정 약물 카테고리 z를 사용하는 개체(1646)를 포함하는 하위 코호트는 특정 약물 카테고리 z의 사용을 중단한 개체(1646)와 제1 시점 t_1 이후 제2 시점 t_2 이전에 특정 약물 카테고리 z를 계속 사용한 개체(1648)를 각각 포함하는 더 작은 하위 코호트로 더 분할될 수 있다.

[0210] 표현형 값 추세 그래프(1662)는 하위 코호트(1642)(즉, t_1 또는 t_2에서 약물 카테고리 z를 사용하고 있지 않았던 환자)에 대해 t_1 및 t_2 에서 측정된 표현형 값의 한 예를 예시하며, 여기서 특정 표현형 값은 t_1에서 y-축에서 높게 시작하고 t_2에서 약간 증가한다. 표현형 값 추세 그래프(1664)는 하위 코호트(1644)(즉, t_1에 약물 카테고리 z를 사용하지 않고 있었지만 t_2 이전에 약물 카테고리 z를 사용하기 시작한 환자)에 대해 t_1 및 t_2에서 측정된 표현형 값의 일례를 예시하며, 여기서 특정 표현형 값은 t_1에서 y-축에서 높게 시작하고 t_2에서 감소한다. 표현형 값 추세 그래프(1666)는 하위 코호트(1642)(즉, t_1에서 약물 카테고리 z를 사용하고 있었지만, t_2 이전에 약물 카테고리 z 사용을 중단한 환자)에 대해 t_1 및 t_2 에서 측정된 표현형 값의 한 예를 예시하며, 여기서 특정 표현형 값은 t_1에서 y-축에서 낮게 시작하고 t_2에서 증가한다. 표현형 값 추세 그래프(1668)는 하위 코호트(1648)(즉, t_1 및 t_2에서 약물 카테고리 z를 사용하고 있었던 환자)에 대해 t_1 및 t_2에서 측정된 표현형 값의 일례를 예시하며, 여기서 특정 표현형 값은 t_1 에서 y-축에서 낮게 시작하고 t_2 에서 낮게 유지된다.

[0211] 개시된 기술의 일부 구현예에서, 표현형 값에 대한 약물 효과(즉, 약물 카테고리 z에 따른 표현형 시프트)는 모델(1682)에 의해 학습될 수 있으며, 여기서 t_1 및 t_2에서의 표현형 값에 대한 델타(δ_y)는 약물 카테고리 z에 대한 β 계수를 사용한 회귀로 모델링된다. 이 모델은 귀무 가설이 약물 카테고리 z에 대한 β 계수가 0과 같다고 명시하고 대립 가설이 약물 카테고리 z에 대한 β 계수가 0과 같지 않다고 명시하는 경우 유의성에 대해 검정될 수 있다.

[0212] 개시된 기술의 일부 구현예에서, 데이터베이스(1602)로부터의 데이터에 구축된 단계 1542 및 1582로부터의 회귀 모델은 다음 프로토콜을 통해 구성된다:

[0213] 34개 약물 카테고리(예를 들어, 스타틴, NSAID, 오피오이드 등)의 표현형 시프트 효과는 5년만큼 떨어진 t_1 및 t_2에서 측정된 정량적 표현형 값 및 대응하는 공변량을 가진 참가자 코호트로부터 추정되었다. 각각의 정량적 표

현형에 대해, t_1 및 t_2에서의 공변량 보정 값은 약물 사용을 제외한 모든 공변량(예를 들어, 연령, 성별, 유전적 주요 성분, 식이, 흡연 상태 등)을 회귀하여 생성된다. 두 시점 간의 차이는 다음과 같이 계산된다:

$$\delta = \tilde{Y}_{t2} - \tilde{Y}_{t1}$$

[0214]

[0215] 각각의 약물 카테고리 X에 대해, 코호트 내의 모든 개체는 (하위 코호트 1642, 1644, 1646 및 1648에 설명된 바와 같이) t_1 및 t_2에서의 약물 사용에 따라 4개 그룹으로 분할되고, 개체가 그 그룹에 속하는지 여부를 인코딩하는 이진 지시자 변수가 각각의 그룹에 대해 도입된다:

[0216] 1) 시점 t_1 및 t_2 모두에서 약물 X를 사용하고 있지 않았던 개체(지시자 변수: X_{00})

[0217] 2) 시점 t_1 및 t_2 사이에 약물 X를 사용하기 시작한 개체(지시자 변수: X_{01})

[0218] 3) 시점 t_1에서 t_2 사이에 약물 X를 사용하는 것을 중단한 개체(지시자 변수: X_{10})

[0219] 4) 시점 t_1 및 t_2 모두에서 약물 X를 사용하고 있던 개체(지시자 변수: X_{11})

[0220] 어느 약물이 표현형 Y에 유의한 영향을 미치는지 결정하기 위해, 모든 약물 카테고리에 걸쳐 반복되는 아래 형식의 전진 선택 단계적 회귀가 피팅된다:

$$\delta = \beta_{00}X_{00} + \beta_{01}X_{01} + \beta_{10}X_{10} + \beta_{11}X_{11} + \beta_t t + \beta_m\left(\tilde{Y}_{t1} - mean\left(\tilde{Y}_{t1}\right)\right) + \beta_0$$

[0221]

[0222] 위 식에서, 항 $\beta_t t$는 t_1과 t_2 사이에 경과된 시간(코호트 내의 각각의 개체에 대해 $t = t_2 - t_1$)의 효과이고, 항 $\beta_m\left(\tilde{Y}_{t1} - mean\left(\tilde{Y}_{t1}\right)\right)$는 t_1과 t_2 사이의 평균으로의 회귀로 인한 효과를 설명하고, β_0은 절편이다. 약물 X를 시작하는 효과를 모델링하는 계수 β_{01}의 p-값이 10^{-3} 이상으로 증가하면 단계적 절차가 중단된다.

[0223] 관련 약물 D 세트가 결정된 후, 다음 회귀 분석을 피팅함으로써 개별 효과가 공동으로 추정될 수 있다.

$$\delta = \sum_{d \in D}\left(\beta_{00}^{(d)}X_{00}^{(d)} + \beta_{01}^{(d)}X_{01}^{(d)} + \beta_{10}^{(d)}X_{10}^{(d)} + \beta_{11}^{(d)}X_{11}^{(d)}\right) + \beta_t t + \beta_m\left(\tilde{Y}_{t1} - mean\left(\tilde{Y}_{t1}\right)\right) + \beta_0$$

[0224]

[0225] 코호트 내의 모든 개체에 대한 t_1에서의 표현형 Y에 대한 원시 값은 다음과 같이 보정될 수 있다:

$$\tilde{Y}_{t1} = Y_{t1} - \sum_{d \in D}\left(\beta_{01}^{(d)} - \beta_{00}^{(d)}\right)X^{(d)}$$

[0226]

[0227] $X^{(d)}$는 개체가 초기 방문 시(t_1) 약물 X를 사용하고 있었는지 여부를 인코딩하는 이진 지시자 변수이다. 약물 사용에 대해 보정한 후, 값 $\tilde{Y}_{t1}$가 역순위 정규 변환될 수 있으며 모든 다른 공변량이 회귀 분석될 수 있다. 공변량 보정된 정규화된 약물 사용 보정 표현형 측정값은 회귀 변이체 다유전자 위험 점수를 생성하는 데 사용될 수 있다.

[0228] 개시된 기술의 일부 구현예에서, 약물 사용 보정은, 제1 시점과 제2 시점 사이에 약물을 사용하기 시작하는 제1 약물 사용 패턴에 대한 제1 이진 지시자 독립 변수, 제1 시점 및 제2 시점에 약물을 사용하지 않는 제2 약물 사용 패턴에 대한 제2 이진 지시자 독립 변수 및 개체가 제1 시점에 특정 약물을 사용하고 있었는지 여부를 인코딩하는 약물 특정 이진 지시자 독립 변수에 기초하여 제1 시점에 대한 표현형 측정값을 약물 사용 보정하는 것을 포함하는 제3 회귀 모델을 피팅함으로써 구현된다. 제1 및 제2 시점 사이에 약물을 사용하는 것을 중단하는 제3 약물 사용 패턴에 대한 제3 이진 지시자 독립 변수, 제1 및 제2 시점에 약물을 사용하는 제4 약물 사용 패턴에 대한 제4 이진 지시자 독립 변수 및 개체가 제2 시점에 특정 약물을 사용하고 있었는지 여부를 인코딩하는 약물 특정 이진 지시자 독립 변수에 기초하여 제2 시점에 대한 표현형 측정값을 약물 사용 보정하는 것을 포함하는 제4 회귀 모델이 피팅된다.

[0229] 순위 기반 역정규 변환은 제1 시점 및 제2 시점에 대한 약물 사용 보정 표현형 측정값에 적용되어 제1 시점 및 제2 시점에 대한 정규화된 약물 사용 보정 표현형 측정값을 생성할 수 있다. 그 다음, 정규화된 약물 사용 보정 표현형 측정값은 제1 시점 및 제2 시점에 대해 공변량 보정되어, 제1 시점 및 제2 시점에 대한 공변량 보정되고 정규화된 약물 사용 보정 표현형 측정값을 생성한다. 공변량 보정되고 정규화된 약물 사용 보정 표현형 측정값은 희귀 변이체 다유전자 위험 점수를 생성하는 데 사용될 수 있고, 관심 표현형에 대한 측정값이 공변량 및 약물 사용 패턴에 따른 표현형 시프트에 대해 보정된다.

[0230] 도 17은 심혈관 질환 환자 X(1700)에 대응하는 복수의 표현형을 예시한다. 심혈관 질환 환자 X(1700)에 대해 측정된 표현형 값은 복수의 교란 요인에 의해 영향을 받을 것이다. 이러한 교란 요인의 예는 인구 통계학(1702)(예를 들어, 연령, 성별, 인종 등), 심리 사회적 요인(1722)(정신 건강 관련 교란 변수, 사회 경제적 계층 등), 담배 및 술 소비(1742), 약물 사용(1704)(예를 들어, 불법 약물과 같은 공식 약물 효과 분석에 포함되지 않은 약물), 체질량 측정값(1724)(예를 들어, 체질량 지수, 체지방률, 복부 지방 밀도 등) 또는 식이 및 생활 방식 요인(1744)(예를 들어, 식이 제한 사항, 식이 습관, 운동 등)을 포함한다. 당업자는 이것이 비한정적인 교란 변수이고 측정된 표현형 값에 영향을 미칠 수 있는 광범위한 복수의 추가 교란 변수가 존재한다는 것을 인식할 것이다.

[0231] **비자명성 및 독창성의 객관적 표시로서의 성능 측정 결과**

[0232] 지금까지의 논의는 특정 표현형에 대한 특정 유전자 내의 희귀 유해 변이체 보인자 상태에 대해 구축된 다유전자 위험 점수 모델을 포함하는 희귀 변이체 부담 검정을 위한 개시된 기술의 복수의 구현예를 다루었으며, 여기서 모델 매개변수(즉, 병원성 점수 임계값 및 최대 대립 유전자 카운트)는 네스팅된 *t*-검정(즉, 그리드 검색)을 통해 최적화되고 두 표현형 값 모두 복수의 공변량에 대해 공변량 보정되고 복수의 약물 사용 패턴에 대해 약물 사용 보정된다. 이제 논의는 개시된 기술의 다양한 구현예에 대한 성능 결과로 넘어간다.

[0233] 도 18은 다양한 유형의 부담 검정에 대해 식별된 유의한 유전자-표현형 쌍의 총 수를 정량화하는 그래프(1800)이다. 유의한 유전자-표현형 쌍의 총 수가 다양한 유형의 부담 검정에 대해 90가지 표현형에 걸쳐 식별되었으며, PrimateAI-3D에 의해 우선 순위가 부여된 희귀 LoF 및 유해 미스센스 변이체에 대한 조합된 검정이 다른 접근 방식보다 성능이 우수함을 보여준다. 음성 대조군으로서, 동의 변이체에 대해 수행된 부담 검정에 대한 유의한 유전자형-표현형 쌍의 수도 도시된다.

[0234] 도 19는 병원성 분류자 PrimateAI-3D에 의해 식별된 질병 중증도 및 발병 연령에 영향을 미치는 희귀 유해 변이체를 예시하는 그래프 모음이다. 그래프(1902)는 *LDLR* 유전자와, LDL 콜레스테롤 및 지질 단백질 대사 장애와의 이의 연관성에 대한 것이다. *LDLR* 유전자 내의 희귀 미스센스 변이체의 보인자의 경우, LDL 콜레스테롤 레벨(y-축)은 PrimateAI-3D 백분위수(percentile) 점수(x-축)와 양으로 상관된다. 이하, PrimateAI-3D 점수는 유전자 사이의 비교를 용이하게 하기 위해 각각의 유전자 내에서 0 내지 1로 정규화된 PrimateAI-3D 백분위수 점수를 의미한다. 그래프(1904)는 PrimateAI-3D 점수가 *LDLR* 유전자 내의 희귀 미스센스 변이체 보인자에 대한 이상 지질혈증 발병 연령을 예측한다는 것을 예시한다.

[0235] 그래프(1922)는 *LDLR*의 하향 조절 인자인 *PCSK9* 유전자에 대한 것이다. *LDLR*의 하향 조절 인자인 *PCSK9*에서의 미스센스 변이체 보인자의 LDL 콜레스테롤 레벨은 PrimateAI-3D 점수와 음으로 상관된다. 그래프(1924)는 보인자의 LDL 콜레스테롤 레벨이 비보인자와 유사한 비율로 연령에 따라 증가하지만 평균적으로 동일한 연령 그룹의 비보인자보다 낮다는 것을 도시한다. 그래프(1942)는 *GCK* 유전자에 대한 것이며, *GCK* 유전자 내의 희귀 미스센스 변이체의 보인자에 대해 HbA1c 레벨이 PrimateAI-3D와 상관된다는 것을 도시한다. 그래프(1944)는 보인자의 HbA1c 레벨이 비보인자와 비슷한 비율로 연령에 따라 증가하지만 평균적으로 일생 동안 당뇨병 전단계 임계값에 더 빨리 도달한다는 것을 도시한다.

[0236] 도 20은 다양한 병원성 점수와 표현형 값의 평균 절대 Spearman 상관 관계의 그래프(2000)이다. 그래프(2000)은 34개 유전자-표현형 쌍의 벤치마킹 세트에서 표현형 값과 다양한 병원성 점수의 평균 절대 Spearman 상관 관계를 동일한 변이체의 보인자에 의해 설정된 이론적 상한과 비교한 것을 도시한다.

[0237] 도 21은 희귀 유해 변이체와 공통 전장 유전체 연관성 분석 변이체의 히트맵(2100) 비교이다. 각각의 표현형 쌍에 대해, 히트맵은 y-축에서의 표현형과 연관된 GWAS 유전자와 x-축에서의 표현형과 연관된 희귀 변이체 유전자 사이의 중첩의 통계적 유의성을 보여준다.

[0238] 도 22는 희귀 유해 변이체와 공통 전장 유전체 연관성 분석 변이체의 추가 비교이다. 그래프(2200A)는 LoF, 유해 미스센스(PrimateAI-3D > 0.5) 및 크립틱(cryptic) 스플라이스(SpliceAI > 0.2) 변이체에 대한 변이체 대립

유전자 빈도와 변이체 효과 크기의 관계를 도시한다. 동의 및 양성 미스센스(PrimateAI-3D < 0.5) 변이체는 음성 대조군으로 도시된다. 도트 크기는 각각의 도트에서의 변이체 수의 제곱근에 비례한다. 그래프(2200B)는 동일한 유전자에 대한 선두 GWAS 변이체의 효과 크기로 나눈 가장 희귀한 LoF 변이체의 효과 크기의 분포를 포함한다. 높은 pLI(상단 플롯) 및 낮은 pLI(하단 플롯) 유전자 모두에 대한 히스토그램이 도시된다. 빨간색 수직선은 각각의 분포의 평균을 나타낸다. 그래프(2200C)는 다양한 GWAS 유의성을 갖는 유전자에 의해 계층화된 희귀 변이체 농축(명목 p-값 ≤ 0.05)을 갖는 명확하게 매핑된 GWAS 유전자의 백분율을 도시한다. 결과는 높은 pLI 유전자(pLI > 0.5)와 낮은 pLI 유전자(pLI < 0.5)에 대해 별도로 도시된다. 신뢰도가 높은 GWAS 유전자는 p-값이 10 내지 100 미만인 선두 변이체 및 연관된 유전자 내에 코딩 변이체를 갖는 강한 LD(r^2 >= 0.9)를 갖는 것으로 정의되었다. 점선은 우연히 명목 p-값 임계값(p < 0.05)을 충족할 것으로 예상되는 유전자의 백분율을 보여준다.

[0239] 도 23은 영국 Biobank 코호트 내의 모든 개체에 걸친 콜레스테롤 경로와 총 콜레스테롤 분포를 예시한다. 도해(2302)는 콜레스테롤 경로를 포함하며, 희귀 변이체 PRS 모델에서의 유전자가 그림에 겹쳐 놓여 있다. 각각의 유전자에 대해, 숫자와 화살표는 효과 크기와 효과 방향을 나타낸다. 그래프(2304)는 영국 Biobank 코호트 내의 모든 개체에 대한 총 콜레스테롤의 분포를 도시하며, 공통 변이체와 희귀 변이체의 평균 효과 크기가 도시된다.

[0240] 도 24는 희귀 변이체 PRS 성능의 측정값을 포함한다. 그래프(2402)는, PRS 훈련을 위해 UKBB의 90%를 사용하고 검정을 위해 10%를 사용한, 공통 변이체 PRS(AF > 1%), 희귀 변이체 PRS(AF < 0.1%) 및 조합된 PRS와 72개 표현형의 평균 상관 관계를 도시한다. 상자 플롯은 형질 사이의 피어슨(Pearson) 상관 관계(R)의 분포를 도시한다. 그래프(2404)는 유전자 또는 유전자좌를 모델에 포함하는 것을 결정하기 위해 다양한 유의성 컷오프를 사용하여 PRS 성능(표현형과의 평균 상관 관계)을 비교한 것이다. 데이터포인트 크기는 각각의 PRS 모델에 사용되는 유전자 또는 유전자좌의 수에 대응한다.

[0241] 도 25는 PRS 이상치의 농축(enrichment)의 그래프(2500)이다. 그래프(2500)는 표현형 이상치인 개체 내의 PRS 이상치의 농축을 도시한다. x-축은 표현형 이상치를 정의하는 데 사용되는 z-점수 임계값을 도시하고, y-축은 기준선 모집단 대비 표현형 이상치 개체에서의 PRS 이상치에 대한 농축을 도시한다.

[0242] 도 26은 정량적 표현형에 대한 PRS 이상치의 그래프(2600)이다. 그래프(2600)는 56개의 정량적 표현형에 대한 PRS 이상치를 도시한다. 막대는 99번째 및 99.9번째 백분위수의 표현형 이상치 임계값에 있는 개체를 설명하기 위해 공통 변이체 PRS 또는 희귀 변이체 PRS에 대한 통계적 유의성을 나타낸다.

[0243] 도 27은 2개의 개별 코호트로부터의 정규화된 총 콜레스테롤 분포의 그래프(2702, 2704)를 포함한다. 그래프(2702)는 콜레스테롤 PRS에 대해 하위 0.5%(낮음), 0.5% 내지 99.5%(중간) 및 상위 0.5%(높음) 그룹에 있는 영국 Biobank 개체로부터의 정규화된 총 콜레스테롤 분포를 도시하며, 코호트의 50%가 PRS를 훈련하는 데 사용되고, 나머지가 검정을 위해 사용된다. 그래프(2704)는 콜레스테롤 PRS에 대해 하위 0.5%(낮음), 0.5% 내지 99.5%(중간) 및 상위 0.5%(높음) 그룹에 있는 MGB 개체로부터의 정규화된 총 콜레스테롤 분포를 도시하며, 코호트의 50%가 PRS를 훈련하는 데 사용되고, 나머지가 검정을 위해 사용된다.

[0244] 도 28은 2개의 개별 코호트 사이에서 희귀 변이체 PRS 이상치와 표현형을 비교하는 그래프(2802, 2804)를 포함한다. 그래프(2802)는 두 코호트 모두에서 측정된 17개의 정량적 표현형에 대한 희귀 변이체 PRS 이상치 검정을 도시한다. PRS와 표현형 이상치는 모집단의 상위 0.5%와 하위 0.5%에 있는 개체로 정의되었다. 그래프(2804)는 17개의 정량적 표현형에 대한 희귀 변이체 PRS와 표현형 사이의 상관 관계를 도시한다.

[0245] 도 29는 인종별 성과 결과를 예시하는 그래프(2902 및 2904)를 포함한다. 그래프(2902)는 MGB 코호트 내의 낮은 PRS 그룹과 높은 PRS 그룹에 대한 정규화된 콜레스테롤 분포를 도시하며, 결과는 백인과 비백인 개체 모두에 대해 도시된다. 그래프(2904)는 백인 및 비백인 개체 모두에 대해 MGB에서의 11개 표현형 각각에 대해 낮은 PRS(0.5% 미만) 그룹과 높은 PRS(99.5% 초과) 그룹 사이의 평균 z-점수 거리의 비교를 도시한다.

[0246] 도 30은 정상 콜레스테롤 레벨에 사용되는 공통 PRS 변이체와 희귀 PRS 유전자에 대한 효과 크기 및 빈도를 비교한 표(3000)이다.

[0247] 도 31은 희귀 미스센스 및 LoF 변이체에 대한 농축에 기초하여 선택된 34개 유전자-표현형 쌍 세트에 대한 다양한 병원성 스코어링 방법에 의해 설명되는 표현형 분산의 평균 분율의 그래프(3100)이다. 설명된 분율 표현형 분산은 각각의 스코어링 방법으로부터의 각각의 점수와 대응하는 변이체의 보인자의 표현형 값 사이의 제곱 Spearman 상관 관계로 계산된다. 방법은 그래프(3100)에서 동일한 미스센스 변이체의 보인자의 표현형 값을 사

용하여 계산된 이론적 상한과 두 개의 이론적 하한(동일한 동의 변이체의 보인자 및 무작위 점수)과 비교된다.

[0248] 도 32는 정량적 표현형과 임상적 표현형 사이의 모든 쌍별 비교를 위한 GWAS 유전자에서의 희귀 변이체의 농축에 대한 히트맵(3200)이다. 각각의 표현형 쌍에 대해, 히트맵(3200)은 제2 표현형(x-축)에 영향을 미치는 희귀 유해 변이체에 대한 농축에 의해 하나의 표현형(y-축)의 GWAS 유전자의 가장 유의한 서브세트의 순위에 대한 통계적 유의성을 도시한다. 희귀 유해 변이체에 대한 부담 검정은 각각의 표현형으로부터 모든 유의한 GWAS 변이체의 효과를 회귀 분석한 후 계산된다. 히트맵(2100)은 히트맵(3200) 내의 주 대각선에 농축된 표현형의 서브세트를 도시한다.

[0249] 도 33은 동일한 유전자에 대한 단일 LoF 변이체의 평균 효과 크기와 가장 유의한 GWAS 변이체의 효과 크기 사이의 비율의 절대값에 대한 분포 그래프(3300)를 포함한다. 그래프(3300A)는 그래프(2200B)로부터의 높은 pLI 유전자에 대해 동일한 유전자에 대한 단일 LoF 변이체의 평균 효과 크기와 가장 유의한 GWAS 변이체의 효과 크기 사이의 비율의 절대값에 대한 분포를 도시한다. 그래프(3300B)는 그래프(2200B)로부터의 낮은 pLI 유전자에 대해 동일한 유전자에 대한 단일 LoF 변이체의 평균 효과 크기와 가장 유의한 GWAS 변이체의 효과 크기 사이의 비율의 절대값에 대한 분포를 도시한다.

[0250] 도 34는 5%의 오류 발견율에서 유의미한 유전자-표현형 쌍으로부터의 임의의 유전자에서 개체당 변이체의 수를 도시한다. 그래프(3400)는 유의성 임계값을 통과한 1031개의 유전자-표현형 쌍 중 500개의 유전자를 포함한다. 1인당 평균 변이체 수는 2.03개이고 표준편차는 1.51이다.

[0251] 도 35는 훈련 데이터 대 검정 데이터 분할에서의 효과 크기의 비교(3500)를 도시한다. 희귀 변이체 검정은 코호트의 50%를 훈련 세트로 사용하고 나머지 50%를 검정 세트로 사용하여 수행되었다. 그래프(3500)에서, 훈련 세트에서 p-값 < 10^{-7}인 유전자에 대한 훈련 세트 대 검정 세트로부터 효과 크기가 비교된다.

[0252] 도 36은 공통 변이체와 희귀 변이체 PRS 서브세트를 비교하는 그래프(3600)를 포함한다. 그래프(3600A)는 유전자좌가 공통 검정과 희귀 검정 모두에서 유의해야 할 필요가 있었던 경우의 공통 및 희귀 PRS 상관 관계를 도시한다. PRS는 90%에 대해 훈련되었고 나머지에 대해 검정되었다. 그래프(3600B)는 훈련 샘플 수별로 최대값에 대한 공통 및 희귀 PRS 상관 관계를 도시한다. 그래프(3600C)는 PrimateAI-3D 없이 구축된 희귀 변이체 PRS 모델에 대한 PrimateAI-3D를 사용한 희귀 PRS 모델로부터의 설명된 효과 크기 및 분산의 비율을 도시한다.

[0253] **조항**

[0254] 개시된 기술, 특히, 이 섹션에 개시된 조항은 시스템, 방법 또는 제조 물품으로서 실시될 수 있다. 하나의 구현예의 하나 이상의 특징은 기본 구현예와 조합될 수 있다. 상호 배타적이지 않은 구현예들은 조합 가능한 것으로 교시된다. 하나의 구현예의 하나 이상의 특징은 다른 구현예와 조합될 수 있다. 본 개시내용은 이러한 옵션을 사용자에게 주기적으로 리마인드한다. 이러한 옵션을 반복하는 인용의 일부 구현예로부터의 생략은 선행하는 섹션에 교시된 조합을 제한하는 것으로서 간주되어서는 안 된다 - 이러한 인용은 본원에 의해 다음의 구현예 각각에 참조로 통합된다.

[0255] 개시된 기술의 하나 이상의 구현예 및 조항 또는 이들의 요소는, 나타낸 방법 단계를 수행하기 위한 컴퓨터 사용 가능 프로그램 코드를 갖는 비일시적 컴퓨터 판독 가능 저장 매체를 포함하는 컴퓨터 제품의 형태로 구현될 수 있다. 더욱이, 개시된 기술의 하나 이상의 구현예 및 조항 또는 이들의 요소는, 메모리 및 메모리에 커플링되고 예시적인 방법 단계를 수행하기 위해 동작하는 적어도 하나의 프로세서를 포함하는 장치의 형태로 구현될 수 있다. 또한, 추가로, 다른 양태에서, 개시된 기술의 하나 이상의 구현예 및 조항 또는 이들의 요소는, 본원에 설명된 방법 단계 중 하나 이상을 수행하기 위한 수단의 형태로 구현될 수 있고; 수단은 (i) 하드웨어 모듈(들), (ii) 하나 이상의 하드웨어 프로세서 상에서 실행되는 소프트웨어 모듈(들), 또는 (iii) 하드웨어와 소프트웨어 모듈의 조합을 포함할 수 있고; (i) 내지 (iii) 중 임의의 것이 본원에 제시된 특정 기술을 구현하고, 소프트웨어 모듈은 컴퓨터 판독 가능 저장 매체(또는 다수의 이러한 매체)에 저장된다.

[0256] 이 섹션에 설명된 조항은 특징으로 조합될 수 있다. 간결함을 위해, 특징의 조합은 개별적으로 열거되지 않으며 각각의 기본 특징 세트에서 반복되지 않는다. 독자는 이 섹션에 설명된 조항에서 식별된 특징이 본 출원의 다른 섹션에서 구현예로서 식별된 기본 특징 세트와 어떻게 쉽게 조합될 수 있는지 이해할 것이다. 이러한 조항들은 상호 배타적이거나 포괄적이거나 제한적이라는 의미가 아니며; 개시된 기술은 이러한 조항들에 제한되지 않으며 오히려 청구된 기술 및 그 등가물의 범위 내에서 가능한 모든 조합, 수정 및 변형을 포함한다.

[0257] 이 섹션에 설명된 조항의 다른 구현예는 이 섹션에 설명된 조항 중 임의의 조항을 수행하기 위해 프로세서에 의

해 실행 가능한 명령어를 저장하는 비일시적 컴퓨터 판독 가능 저장 매체를 포함할 수 있다. 이 섹션에 설명된 조항의 또 다른 구현예는 메모리 및 이 섹션에 설명된 조항 중 임의의 조항을 수행하기 위해 메모리에 저장된 명령어를 실행하도록 동작 가능한 하나 이상의 프로세서를 포함하는 시스템을 포함할 수 있다.

[0258] 우리는 다음 조항들을 개시한다:

[0259] 1. 복수의 교란 변수를 갖는 개체 코호트의 복수의 표현형에 대한 복수의 약물의 사용에 따른 표현형 시프트를 예측하는 컴퓨터 구현 방법으로서,

[0260] 개체 코호트에 대해 그리고 제1 시점 및 제2 시점에 대해:

[0261] 복수의 표현형에 대한 표현형 측정값에 접근하는 단계;

[0262] 복수의 교란 변수에 대한 공변량 측정값에 접근하는 단계; 및

[0263] 복수의 약물에 대한 약물 사용 패턴에 접근하는 단계를 포함하고,

[0264] 표현형별 기준으로:

[0265] 공변량 측정값에 기초하여 제1 시점 및 제2 시점에 대한 표현형 측정값을 공변량 보정하여, 제1 시점 및 제2 시점에 대한 공변량 보정 표현형 측정값을 생성하는 단계;

[0266] 제1 시점과 제2 시점에 대한 공변량 보정 표현형 측정값 사이의 차이에 기초하여 델타를 결정하는 단계;

[0267] 각각의 약물 사용 패턴에 대해, 공변량 보정 표현형 측정값에서의 복수의 약물의 사용에 따른 표현형 시프트를 예측하기 위해 델타를 사용하는 제2 회귀 모델을 피팅하는 단계; 및

[0268] 표현형 시프트에 기초하여 제1 시점 및 제2 시점에 대한 표현형 측정값을 약물 사용 보정하여, 제1 시점에 대한 약물 사용 보정 표현형 측정값을 생성하는 단계를 포함하는, 컴퓨터 구현 방법.

[0269] 2. 조항 1에 있어서, 복수의 교란 변수는 연령, 성별, 유전적 주요 성분, 식이 및 흡연 상태를 포함하는, 컴퓨터 구현 방법.

[0270] 3. 조항 1에 있어서, 공변량 보정은 제1 회귀 모델을 피팅하여 공변량 측정값을 회귀시킴으로써 구현되는, 컴퓨터 구현 방법.

[0271] 4. 조항 1에 있어서, 표현형 시프트 예측은 각각의 약물 사용 패턴에 대한 표현형 시프트를 모델링하는 제2 회귀 모델을 피팅함으로써 구현되는, 컴퓨터 구현 방법.

[0272] 5. 조항 4에 있어서, 제2 회귀 모델은 각각의 약물 사용 패턴에 대한 표현형 시프트를 연속적으로 그리고 누적적으로 포함함으로써 델타를 반복적으로 예측하는 전진 선택 단계적 회귀 모델인, 컴퓨터 구현 방법.

[0273] 6. 조항 4에 있어서, 약물 사용 패턴은,

[0274] 제1 시점 및 제2 시점에 약물을 사용하지 않는 것,

[0275] 제1 시점과 제2 시점 사이에 약물을 사용하기 시작하는 것,

[0276] 제1 시점과 제2 시점 사이에 약물을 사용하는 것을 중단하는 것, 및

[0277] 제1 시점 및 제2 시점에 약물을 사용하는 것을 포함하는, 컴퓨터 구현 방법.

[0278] 7. 조항 6항에 있어서, 제2 회귀 모델은 각각의 약물 사용 패턴에 대한 이진 지시자 독립 변수를 갖는, 컴퓨터 구현 방법.

[0279] 8. 조항 1에 있어서, 공변량 보정, 델타 결정, 표현형 시프트 예측 및 약물 사용 보정은 복수의 약물에서의 약물에 대해 약물별 기준으로 실행되는, 컴퓨터 구현 방법.

[0280] 9. 조항 8에 있어서, 제2 회귀 모델은 각각의 약물에 대해 반복적으로 피팅되는, 컴퓨터 구현 방법.

[0281] 10. 조항 8에 있어서, 약물을 약물 카테고리 세트로 그룹화하는 단계를 더 포함하는 컴퓨터 구현 방법.

[0282] 11. 조항 10에 있어서, 공변량 보정, 델타 결정, 표현형 시프트 예측 및 약물 사용 보정은 약물 카테고리 세트에서의 약물 카테고리에 대해 약물 카테고리별 기준으로 실행되는, 컴퓨터 구현 방법.

[0283] 12. 조항 11에 있어서, 제2 회귀 모델은 각각의 약물 카테고리에 대해 반복적으로 피팅되는, 컴퓨터 구현 방법.

[0284] 13. 조항 1에 있어서, 제2 회귀 모델은 개체 코호트 내의 개체에 대한 개체별 기준으로 제1 시점과 제2 시점 사이에 경과된 시간에 따라 표현형 시프트를 더 모델링하는, 컴퓨터 구현 방법.

[0285] 14. 조항 1에 있어서, 제2 회귀 모델은 제1 시점과 제2 시점 사이의 평균으로의 회귀에 따라 표현형 시프트를 더 모델링하는, 컴퓨터 구현 방법.

[0286] 15. 조항 8에 있어서, 제2 회귀 모델은 복수의 약물에서 관련 약물 세트에 대해 공동으로 피팅되는, 컴퓨터 구현 방법.

[0287] 16. 조항 1에 있어서, 약물 사용 보정은 제3 회귀 모델을 피팅함으로써 구현되는, 컴퓨터 구현 방법.

[0288] 17. 조항 16에 있어서, 제1 시점과 제2 시점 사이에 약물을 사용하기 시작하는 제1 약물 사용 패턴에 대한 제1 이진 지시자 독립 변수, 제1 시점 및 제2 시점에 약물을 사용하지 않는 제2 약물 사용 패턴에 대한 제2 이진 지시자 독립 변수 및 개체가 제1 시점에 특정 약물을 사용하고 있었는지 여부를 인코딩하는 약물 특정 이진 지시자 독립 변수에 기초하여 제1 시점에 대한 표현형 측정값을 약물 사용 보정하는 단계를 더 포함하는 컴퓨터 구현 방법.

[0289] 18. 조항 1에 있어서, 약물 사용 보정은 제4 회귀 모델을 피팅함으로써 구현되는, 컴퓨터 구현 방법.

[0290] 19. 조항 18에 있어서, 제1 시점과 제2 시점 사이에 약물을 사용하는 것을 중단하는 제3 약물 사용 패턴에 대한 제3 이진 지시자 독립 변수, 제1 시점 및 제2 시점에 약물을 사용하는 제4 약물 사용 패턴에 대한 제4 이진 지시자 독립 변수 및 개체가 제2 시점에 특정 약물을 사용하고 있었는지 여부를 인코딩하는 약물 특정 이진 지시자 독립 변수에 기초하여 제2 시점에 대한 표현형 측정값을 약물 사용 보정하는 단계를 더 포함하는 컴퓨터 구현 방법.

[0291] 20. 조항 1에 있어서, 순위 기반 역정규 변환을 제1 시점 및 제2 시점에 대한 약물 사용 보정 표현형 측정값에 적용하여 제1 시점 및 제2 시점에 대한 정규화된 약물 사용 보정 표현형 측정값을 생성하는 단계를 더 포함하는 컴퓨터 구현 방법.

[0292] 21. 조항 12에 있어서, 제1 시점 및 제2 시점에 대한 정규화된 약물 사용 보정 표현형 측정값을 공변량 보정하여, 제1 시점 및 제2 시점에 대한 공변량 보정되고 정규화된 약물 사용 보정 표현형 측정값을 생성하는 단계를 더 포함하는 컴퓨터 구현 방법.

[0293] 22. 조항 12에 있어서, 공변량 보정되고 정규화된 약물 사용 보정 표현형 측정값을 사용하여 희귀 변이체 다유전자 위험 점수를 생성하는 단계를 더 포함하는 컴퓨터 구현 방법.

[0294] 23. 조항 1에 있어서, 복수의 표현형은 복수의 정량적 표현형에 대응하는, 컴퓨터 구현 방법.

[0295] 24. 조항 23에 있어서, 복수의 정량적 표현형에서의 정량적 표현형은 정량적 바이오마커 측정값인, 컴퓨터 구현 방법.

[0296] 25. 조항 23에 있어서, 공변량 보정, 델타 결정, 표현형 시프트 및 약물 사용 보정에 사용하기 위해 복수의 정량적 표현형을 중복되지 않은 세트로 프루닝하는 단계를 더 포함하는 컴퓨터 구현 방법.

[0297] 26. 조항 25에 있어서, 중복되지 않은 세트 내의 정량적 표현형의 각각의 쌍은 상한 임계값보다 낮은 절대 쌍별 피어슨 상관 관계를 갖는, 컴퓨터 구현 방법.

[0298] 27. 조항 26에 있어서, 상한 임계값은 0.95인, 컴퓨터 구현 방법.

[0299] 28. 조항 26에 있어서, 복수의 정량적 표현형 내의 중복된 정량적 표현형의 각각의 그룹 중에서, 중복되지 않은 세트에 포함시키기 위해 가장 많은 샘플을 갖는 표현형을 선택하는 단계를 더 포함하는 컴퓨터 구현 방법.

[0300] 29. 조항 1에 있어서, 복수의 표현형은 복수의 범주형 표현형에 대응하는, 컴퓨터 구현 방법.

[0301] 30. 조항 29에 있어서, 복수의 범주형 표현형에서의 범주형 표현형은 임상 진단인, 컴퓨터 구현 방법.

[0302] 31. 조항 4에 있어서, 제2 회귀 모델을 사용하여 약물-표현형 연관성을 검출하는 단계를 더 포함하는 컴퓨터 구현 방법.

[0303] 32. 조항 31에 있어서, 약물-표현형 연관성은 잠재적인 원치 않는 부작용 및 원하는 표적 효과를 포함하는, 컴퓨터 구현 방법.

[0304] 33. 시스템으로서, 메모리에 커플링된 하나 이상의 프로세서를 포함하고, 메모리에는 복수의 교란 변수를 갖는 개체 코호트의 복수의 표현형에 대한 복수의 약물의 사용에 따른 표현형 시프트를 예측하기 위한 컴퓨터 명령어가 로딩되고, 명령어는, 프로세서에서 실행될 때,

[0305] 개체 코호트에 대해 그리고 제1 시점 및 제2 시점에 대해:

[0306] 복수의 표현형에 대한 표현형 측정값에 접근하는 것;

[0307] 복수의 교란 변수에 대한 공변량 측정값에 접근하는 것; 및

[0308] 복수의 약물에 대한 약물 사용 패턴에 접근하는 것을 포함하고,

[0309] 표현형별 기준으로:

[0310] 공변량 측정값에 기초하여 제1 시점 및 제2 시점에 대한 표현형 측정값을 공변량 보정하여, 제1 시점 및 제2 시점에 대한 공변량 보정 표현형 측정값을 생성하는 것;

[0311] 제1 시점과 제2 시점에 대한 공변량 보정 표현형 측정값 사이의 차이에 기초하여 델타를 결정하는 것;

[0312] 각각의 약물 사용 패턴에 대해, 공변량 보정 표현형 측정값에서의 복수의 약물의 사용에 따른 표현형 시프트를 예측하기 위해 델타를 사용하는 것; 및 표현형 시프트에 기초하여 제1 시점에 대한 표현형 측정값을 약물 사용 보정하여, 제1 시점 및 제2 시점에 대한 약물 사용 보정 표현형 측정값을 생성하는 것을 포함하는 동작들을 구현하는, 시스템.

[0313] 34. 조항 33에 있어서, 복수의 교란 변수는 연령, 성별, 유전적 주요 성분, 식이 및 흡연 상태를 포함하는, 시스템.

[0314] 35. 조항 33에 있어서, 공변량 보정은 제1 회귀 모델을 피팅하여 공변량 측정값을 회귀시킴으로써 구현되는, 시스템.

[0315] 36. 조항 33항에 있어서, 표현형 시프트 예측은 각각의 약물 사용 패턴에 대한 표현형 시프트를 모델링하는 제2 회귀 모델을 피팅함으로써 구현되는, 시스템.

[0316] 37. 조항 36에 있어서, 제2 회귀 모델은 각각의 약물 사용 패턴에 대한 표현형 시프트를 연속적으로 그리고 누적적으로 포함함으로써 델타를 반복적으로 예측하는 전진 선택 단계적 회귀 모델인, 시스템.

[0317] 38. 조항 36에 있어서, 약물 사용 패턴은,

[0318] 제1 시점 및 제2 시점에 약물을 사용하지 않는 것,

[0319] 제1 시점과 제2 시점 사이에 약물을 사용하기 시작하는 것,

[0320] 제1 시점과 제2 시점 사이에 약물을 사용하는 것을 중단하는 것, 및

[0321] 제1 시점 및 제2 시점에 약물을 사용하는 것을 포함하는, 시스템.

[0322] 39. 조항 38항에 있어서, 제2 회귀 모델은 각각의 약물 사용 패턴에 대한 이진 지시자 독립 변수를 갖는, 시스템.

[0323] 40. 조항 33에 있어서, 공변량 보정, 델타 결정, 표현형 시프트 예측 및 약물 사용 보정은 복수의 약물에서의 약물에 대해 약물별 기준으로 실행되는, 시스템.

[0324] 41. 조항 40에 있어서, 제2 회귀 모델은 각각의 약물에 대해 반복적으로 피팅되는, 시스템.

[0325] 42. 조항 40에 있어서, 약물을 약물 카테고리 세트로 그룹화하는 동작을 더 구현하는 시스템.

[0326] 43. 조항 42에 있어서, 공변량 보정, 델타 결정, 표현형 시프트 예측 및 약물 사용 보정은 약물 카테고리 세트에서의 약물 카테고리에 대해 약물 카테고리별 기준으로 실행되는, 시스템.

[0327] 44. 조항 43에 있어서, 제2 회귀 모델은 각각의 약물 카테고리에 대해 반복적으로 피팅되는, 시스템.

[0328] 45. 조항 33에 있어서, 제2 회귀 모델은 개체 코호트 내의 개체에 대한 개체별 기준으로 제1 시점과 제2 시점 사이에 경과된 시간에 따라 표현형 시프트를 더 모델링하는, 시스템.

[0329] 46. 조항 33에 있어서, 제2 회귀 모델은 제1 시점과 제2 시점 사이의 평균으로의 회귀에 따라 표현형 시프트를

더 모델링하는, 시스템.

[0330] 47. 조항 40에 있어서, 제2 회귀 모델은 복수의 약물에서 관련 약물 세트에 대해 공동으로 피팅되는, 시스템.

[0331] 48. 조항 33에 있어서, 약물 사용 보정은 제3 회귀 모델을 피팅함으로써 구현되는, 시스템.

[0332] 49. 조항 48에 있어서, 제1 시점과 제2 시점 사이에 약물을 사용하기 시작하는 제1 약물 사용 패턴에 대한 제1 이진 지시자 독립 변수, 제1 시점 및 제2 시점에 약물을 사용하지 않는 제2 약물 사용 패턴에 대한 제2 이진 지시자 독립 변수 및 개체가 제1 시점에 특정 약물을 사용하고 있었는지 여부를 인코딩하는 약물 특정 이진 지시자 독립 변수에 기초하여 제1 시점에 대한 표현형 측정값을 약물 사용 보정하는 동작을 더 구현하는 시스템.

[0333] 50. 조항 33에 있어서, 약물 사용 보정은 제4 회귀 모델을 피팅함으로써 구현되는, 시스템.

[0334] 51. 조항 50에 있어서, 제1 시점과 제2 시점 사이에 약물을 사용하는 것을 중단하는 제3 약물 사용 패턴에 대한 제3 이진 지시자 독립 변수, 제1 시점 및 제2 시점에 약물을 사용하는 제4 약물 사용 패턴에 대한 제4 이진 지시자 독립 변수 및 개체가 제2 시점에 특정 약물을 사용하고 있었는지 여부를 인코딩하는 약물 특정 이진 지시자 독립 변수에 기초하여 제2 시점에 대한 표현형 측정값을 약물 사용 보정하는 동작을 더 구현하는 시스템.

[0335] 52. 조항 33에 있어서, 순위 기반 역정규 변환을 제1 시점 및 제2 시점에 대한 약물 사용 보정 표현형 측정값에 적용하여 제1 시점 및 제2 시점에 대한 정규화된 약물 사용 보정 표현형 측정값을 생성하는 동작을 더 구현하는 시스템.

[0336] 53. 조항 44에 있어서, 제1 시점 및 제2 시점에 대한 정규화된 약물 사용 보정 표현형 측정값을 공변량 보정하여, 제1 시점 및 제2 시점에 대한 공변량 보정되고 정규화된 약물 사용 보정 표현형 측정값을 생성하는 동작을 더 구현하는 시스템.

[0337] 54. 조항 44에 있어서, 공변량 보정되고 정규화된 약물 사용 보정 표현형 측정값을 사용하여 희귀 변이체 다유전자 위험 점수를 생성하는 동작을 더 구현하는 시스템.

[0338] 55. 조항 33에 있어서, 복수의 표현형은 복수의 정량적 표현형에 대응하는, 시스템.

[0339] 56. 조항 55에 있어서, 복수의 정량적 표현형에서의 정량적 표현형은 정량적 바이오마커 측정값인, 시스템.

[0340] 57. 조항 55에 있어서, 공변량 보정, 델타 결정, 표현형 시프트 예측 및 약물 사용 보정에 사용하기 위해 복수의 정량적 표현형을 중복되지 않은 세트로 프루닝하는 동작을 더 구현하는 시스템.

[0341] 58. 조항 57에 있어서, 중복되지 않은 세트 내의 정량적 표현형의 각각의 쌍은 상한 임계값보다 낮은 절대 쌍별 피어슨 상관 관계를 갖는, 시스템.

[0342] 59. 조항 58에 있어서, 상한 임계값은 0.95인, 시스템.

[0343] 60. 조항 58에 있어서, 복수의 정량적 표현형 내의 중복된 정량적 표현형의 각각의 그룹 중에서, 중복되지 않은 세트에 포함시키기 위해 가장 많은 샘플을 갖는 표현형을 선택하는 동작을 더 구현하는 시스템.

[0344] 61. 조항 33에 있어서, 복수의 표현형은 복수의 범주형 표현형에 대응하는, 시스템.

[0345] 62. 조항 61에 있어서, 복수의 범주형 표현형에서의 범주형 표현형은 임상 진단인, 시스템.

[0346] 63. 조항 36에 있어서, 제2 회귀 모델을 사용하여 약물-표현형 연관성을 검출하는 동작을 더 구현하는 시스템.

[0347] 64. 조항 63에 있어서, 약물-표현형 연관성은 잠재적인 원치 않는 부작용 및 원하는 표적 효과를 포함하는, 시스템.

[0348] 65. 복수의 교란 변수를 갖는 개체 코호트의 복수의 표현형에 대한 복수의 약물의 사용에 따른 표현형 시프트를 예측하기 위한 컴퓨터 프로그램 명령어가 저장된 비일시적 컴퓨터 판독 가능 저장 매체로서, 명령어는, 프로세서에서 실행될 때,

[0349] 개체 코호트에 대해 그리고 제1 시점 및 제2 시점에 대해:

[0350] 복수의 표현형에 대한 표현형 측정값에 접근하는 단계;

[0351] 복수의 교란 변수에 대한 공변량 측정값에 접근하는 단계; 및

[0352] 복수의 약물에 대한 약물 사용 패턴에 접근하는 단계를 포함하고,

[0353] 표현형별 기준으로:

[0354] 공변량 측정값에 기초하여 제1 시점 및 제2 시점에 대한 표현형 측정값을 공변량 보정하여, 제1 시점 및 제2 시점에 대한 공변량 보정 표현형 측정값을 생성하는 단계;

[0355] 제1 시점과 제2 시점에 대한 공변량 보정 표현형 측정값 사이의 차이에 기초하여 델타를 결정하는 단계;

[0356] 각각의 약물 사용 패턴에 대해, 공변량 보정 표현형 측정값에서의 복수의 약물의 사용에 따른 표현형 시프트를 예측하기 위해 델타를 사용하는 단계; 및

[0357] 표현형 시프트 예측에 기초하여 제1 시점 및 제2 시점에 대한 표현형 측정값을 약물 사용 보정하여, 제1 시점 및 제2 시점에 대한 약물 사용 보정 표현형 측정값을 생성하는 단계를 포함하는 방법을 구현하는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0358] 66. 조항 65에 있어서, 복수의 교란 변수는 연령, 성별, 유전적 주요 성분, 식이 및 흡연 상태를 포함하는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0359] 67. 조항 65에 있어서, 공변량 보정은 제1 회귀 모델을 피팅하여 공변량 측정값을 회귀시킴으로써 구현되는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0360] 68. 조항 65항에 있어서, 표현형 시프트 예측은 각각의 약물 사용 패턴에 대한 표현형 시프트를 모델링하는 제2 회귀 모델을 피팅함으로써 구현되는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0361] 69. 조항 68에 있어서, 제2 회귀 모델은 각각의 약물 사용 패턴에 대한 표현형 시프트를 연속적으로 그리고 누적적으로 포함함으로써 델타를 반복적으로 예측하는 전진 선택 단계적 회귀 모델인, 비일시적 컴퓨터 판독 가능 저장 매체.

[0362] 70. 조항 68에 있어서, 약물 사용 패턴은,

[0363] 제1 시점 및 제2 시점에 약물을 사용하지 않는 것,

[0364] 제1 시점과 제2 시점 사이에 약물을 사용하기 시작하는 것,

[0365] 제1 시점과 제2 시점 사이에 약물을 사용하는 것을 중단하는 것, 및

[0366] 제1 시점 및 제2 시점에 약물을 사용하는 것을 포함하는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0367] 71. 조항 70항에 있어서, 제2 회귀 모델은 각각의 약물 사용 패턴에 대한 이진 지시자 독립 변수를 갖는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0368] 72. 조항 65에 있어서, 공변량 보정, 델타 결정, 표현형 시프트 예측 및 약물 사용 보정은 복수의 약물에서의 약물에 대해 약물별 기준으로 실행되는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0369] 73. 조항 72에 있어서, 제2 회귀 모델은 각각의 약물에 대해 반복적으로 피팅되는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0370] 74. 조항 72에 있어서, 약물을 약물 카테고리 세트로 그룹화하는 단계를 더 포함하는 방법을 구현하는 비일시적 컴퓨터 판독 가능 저장 매체.

[0371] 75. 조항74에 있어서, 공변량 보정, 델타 결정, 표현형 시프트 예측 및 약물 사용 보정은 약물 카테고리 세트에서의 약물 카테고리에 대해 약물 카테고리별 기준으로 실행되는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0372] 76. 조항 75에 있어서, 제2 회귀 모델은 각각의 약물 카테고리에 대해 반복적으로 피팅되는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0373] 77. 조항 65에 있어서, 제2 회귀 모델은 개체 코호트 내의 개체에 대한 개체별 기준으로 제1 시점과 제2 시점 사이에 경과된 시간에 따라 표현형 시프트를 더 모델링하는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0374] 78. 조항 65에 있어서, 제2 회귀 모델은 제1 시점과 제2 시점 사이의 평균으로의 회귀에 따라 표현형 시프트를 더 모델링하는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0375] 79. 조항 72에 있어서, 제2 회귀 모델은 복수의 약물에서 관련 약물 세트에 대해 공동으로 피팅되는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0376] 80. 조항 65에 있어서, 약물 사용 보정은 제3 회귀 모델을 피팅함으로써 구현되는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0377] 81. 조항 80에 있어서, 제1 시점과 제2 시점 사이에 약물을 사용하기 시작하는 제1 약물 사용 패턴에 대한 제1 이진 지시자 독립 변수, 제1 시점 및 제2 시점에 약물을 사용하지 않는 제2 약물 사용 패턴에 대한 제2 이진 지시자 독립 변수 및 개체가 제1 시점에 특정 약물을 사용하고 있었는지 여부를 인코딩하는 약물 특정 이진 지시자 독립 변수에 기초하여 제1 시점에 대한 표현형 측정값을 약물 사용 보정하는 단계를 더 포함하는 방법을 구현하는 비일시적 컴퓨터 판독 가능 저장 매체.

[0378] 82. 조항 65에 있어서, 약물 사용 보정은 제4 회귀 모델을 피팅함으로써 구현되는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0379] 83. 조항 82에 있어서, 제1 시점과 제2 시점 사이에 약물을 사용하는 것을 중단하는 제3 약물 사용 패턴에 대한 제3 이진 지시자 독립 변수, 제1 시점 및 제2 시점에 약물을 사용하는 제4 약물 사용 패턴에 대한 제4 이진 지시자 독립 변수 및 개체가 제2 시점에 특정 약물을 사용하고 있었는지 여부를 인코딩하는 약물 특정 이진 지시자 독립 변수에 기초하여 제2 시점에 대한 표현형 측정값을 약물 사용 보정하는 단계를 더 포함하는 방법을 구현하는 비일시적 컴퓨터 판독 가능 저장 매체.

[0380] 84. 조항 65에 있어서, 순위 기반 역정규 변환을 제1 시점 및 제2 시점에 대한 약물 사용 보정 표현형 측정값에 적용하여 제1 시점 및 제2 시점에 대한 정규화된 약물 사용 보정 표현형 측정값을 생성하는 단계를 더 포함하는 방법을 구현하는 비일시적 컴퓨터 판독 가능 저장 매체.

[0381] 85. 조항 76에 있어서, 제1 시점 및 제2 시점에 대한 정규화된 약물 사용 보정 표현형 측정값을 공변량 보정하여, 제1 시점 및 제2 시점에 대한 공변량 보정되고 정규화된 약물 사용 보정 표현형 측정값을 생성하는 단계를 더 포함하는 방법을 구현하는 비일시적 컴퓨터 판독 가능 저장 매체.

[0382] 86. 조항 76에 있어서, 공변량 보정되고 정규화된 약물 사용 보정 표현형 측정값을 사용하여 희귀 변이체 다유전자 위험 점수를 생성하는 단계를 더 포함하는 방법을 구현하는 비일시적 컴퓨터 판독 가능 저장 매체.

[0383] 87. 조항 65에 있어서, 복수의 표현형은 복수의 정량적 표현형에 대응하는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0384] 88. 조항 87에 있어서, 복수의 정량적 표현형에서의 정량적 표현형은 정량적 바이오마커 측정값인, 비일시적 컴퓨터 판독 가능 저장 매체.

[0385] 89. 조항 87에 있어서, 공변량 보정, 델타 결정, 표현형 시프트 예측 및 약물 사용 보정에 사용하기 위해 복수의 정량적 표현형을 중복되지 않은 세트로 프루닝하는 단계를 더 포함하는 방법을 구현하는 비일시적 컴퓨터 판독 가능 저장 매체.

[0386] 90. 조항 89에 있어서, 중복되지 않은 세트 내의 정량적 표현형의 각각의 쌍은 상한 임계값보다 낮은 절대 쌍별 피어슨 상관 관계를 갖는, 비일시적 컴퓨터 판독 가능 저장 매체.

[0387] 91. 조항 90에 있어서, 상한 임계값은 0.95인, 비일시적 컴퓨터 판독 가능 저장 매체.

[0388] 92. 조항 90에 있어서, 복수의 정량적 표현형 내의 중복된 정량적 표현형의 각각의 그룹 중에서, 중복되지 않은 세트에 포함시키기 위해 가장 많은 샘플을 갖는 표현형을 선택하는 단계를 더 포함하는 방법을 구현하는 비일시적 컴퓨터 판독 가능 저장 매체.

[0389] 93. 조항 65에 있어서, 복수의 표현형은 복수의 범주형 표현형에 대응하는, 비일시적 컴퓨터 판독 가능 저장 매체.

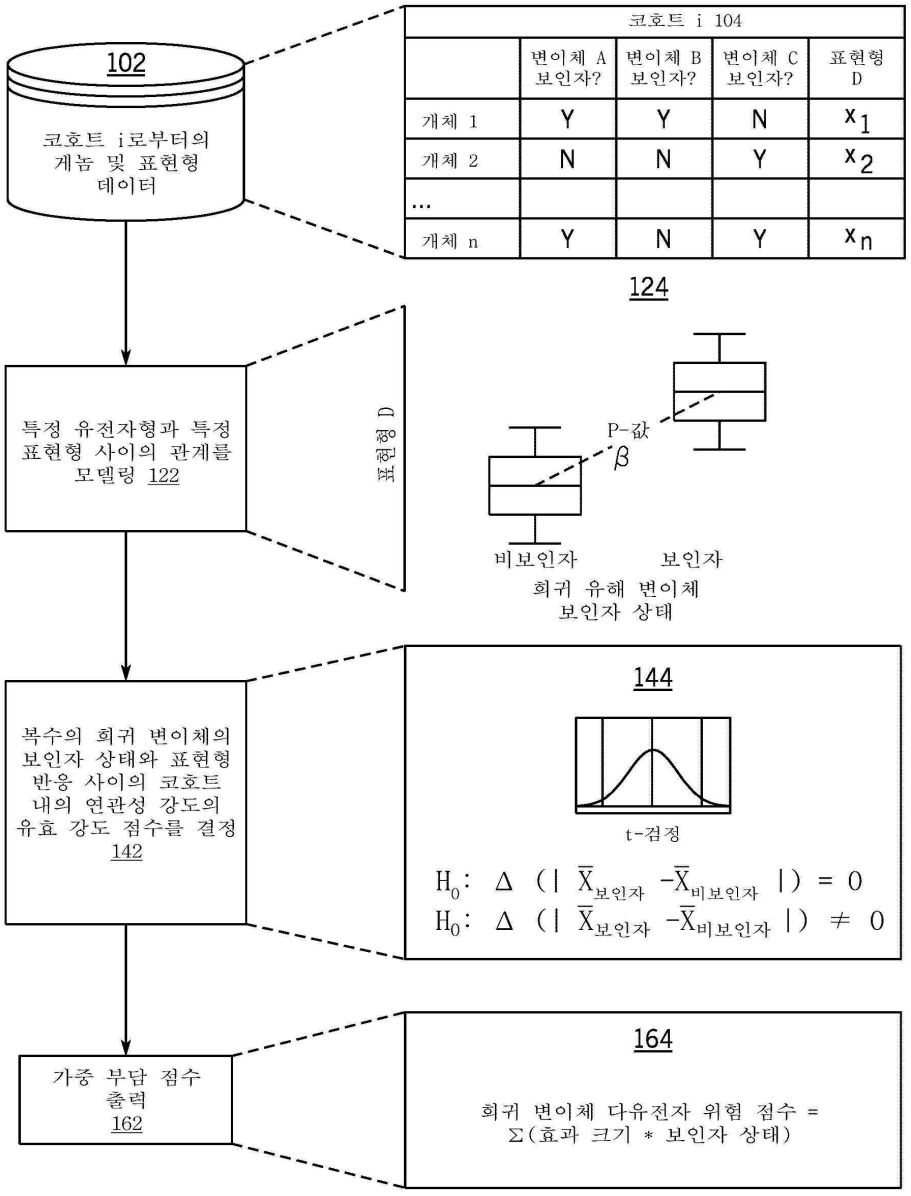
[0390] 94. 조항 93에 있어서, 복수의 범주형 표현형에서의 범주형 표현형은 임상 진단인, 비일시적 컴퓨터 판독 가능 저장 매체.

[0391] 95. 조항 68에 있어서, 제2 회귀 모델을 사용하여 약물-표현형 연관성을 검출하는 단계를 더 포함하는 방법을 구현하는 비일시적 컴퓨터 판독 가능 저장 매체.

[0392] 96. 조항 95에 있어서, 약물-표현형 연관성은 잠재적인 원치 않는 부작용 및 원하는 표적 효과를 포함하는, 비일시적 컴퓨터 판독 가능 저장 매체.

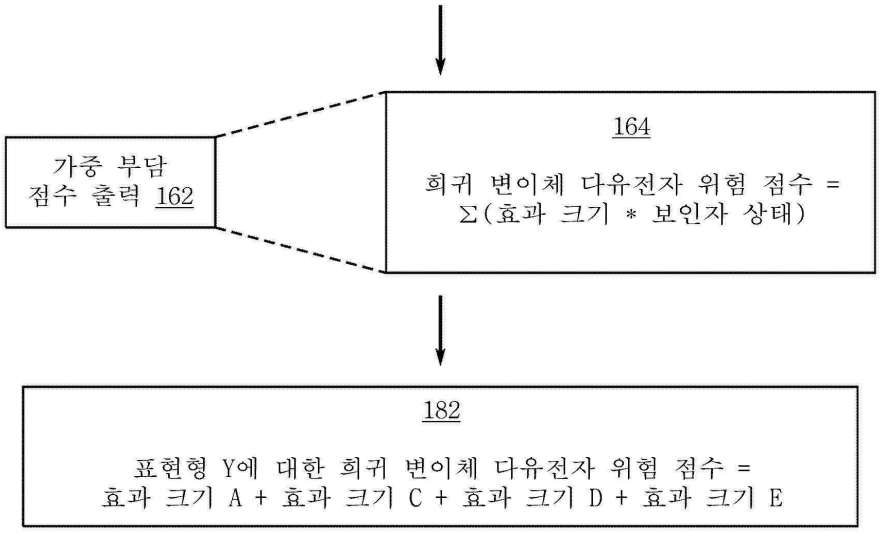
도면

도면1a

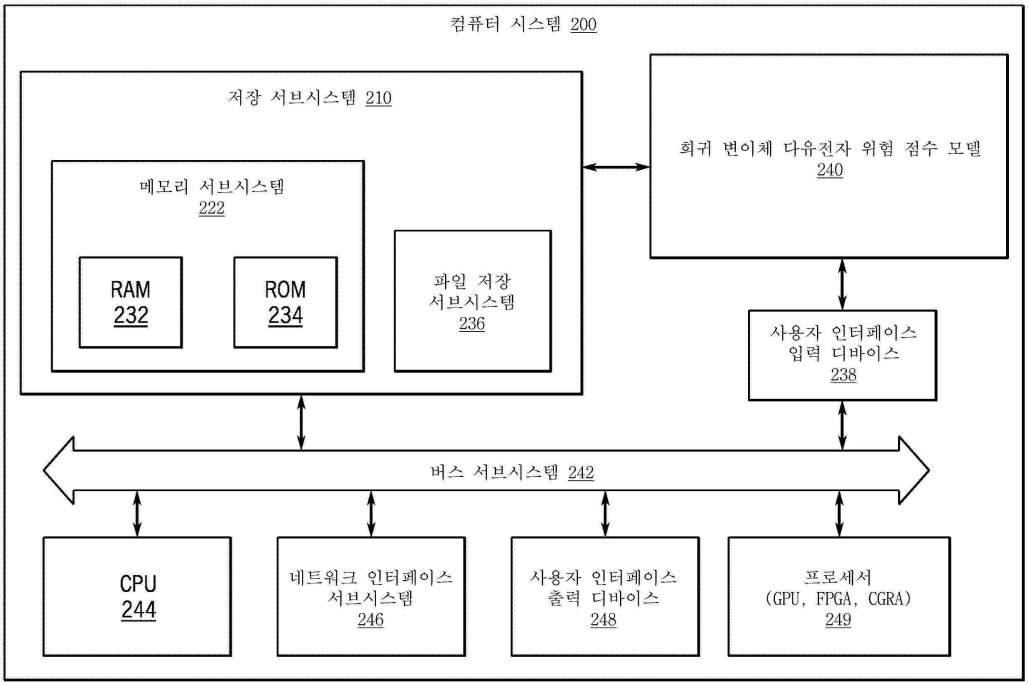

코호트 i 104
변이체 A 보인자?
변이체 B 보인자?
변이체 C 보인자?
표현형 D
개체 1
개체 2
...
개체 n
102
코호트 i로부터의 게놈 및 표현형 데이터
124
특정 유전자형과 특정 표현형 사이의 관계를 모델링 122
표현형 D
P-값
β
비보인자
보인자
희귀 유해 변이체 보인자 상태
복수의 희귀 변이체의 보인자 상태와 표현형 반응 사이의 코호트 내의 연관성 강도의 유효 강도 점수를 결정 142
144
t-검정
가중 부담 점수 출력 162
164
희귀 변이체 다유전자 위험 점수 = Σ(효과 크기 * 보인자 상태)

도면1b

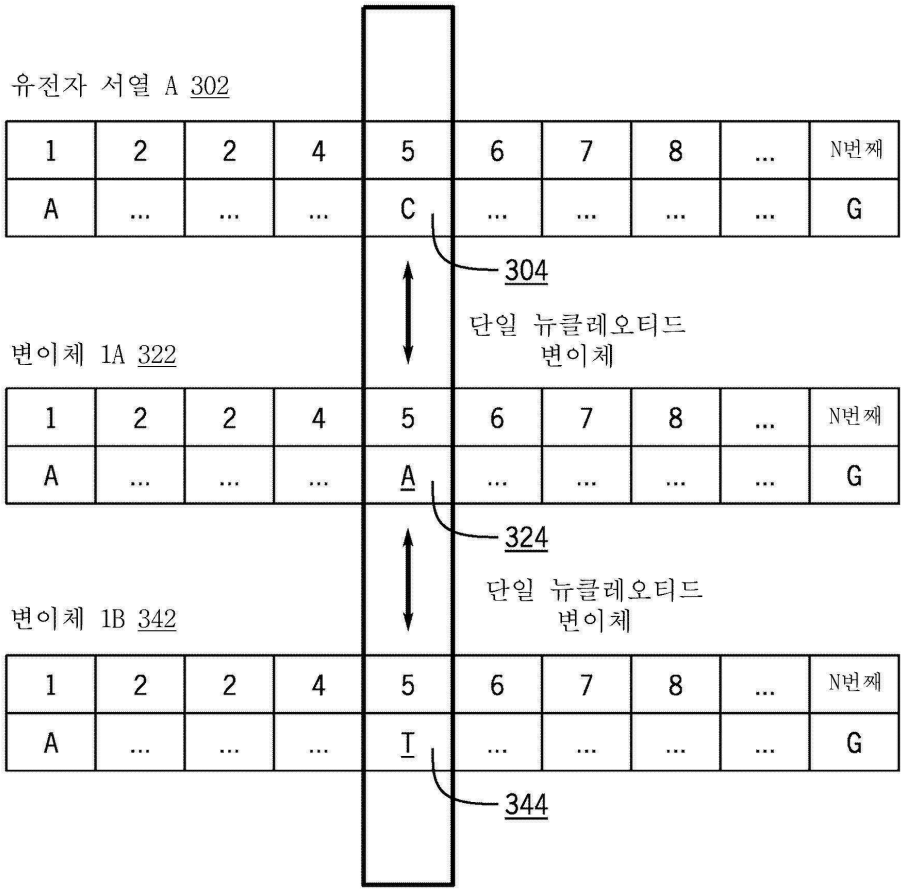
108	유전자	효과 크기	보인자 상태
118	유전자 A	효과 크기 A	보인자
128	유전자 B	효과 크기 B	비보인자
138	유전자 C	효과 크기 C	보인자
148	유전자 D	효과 크기 F	보인자
158	유전자 E	효과 크기 E	보인자



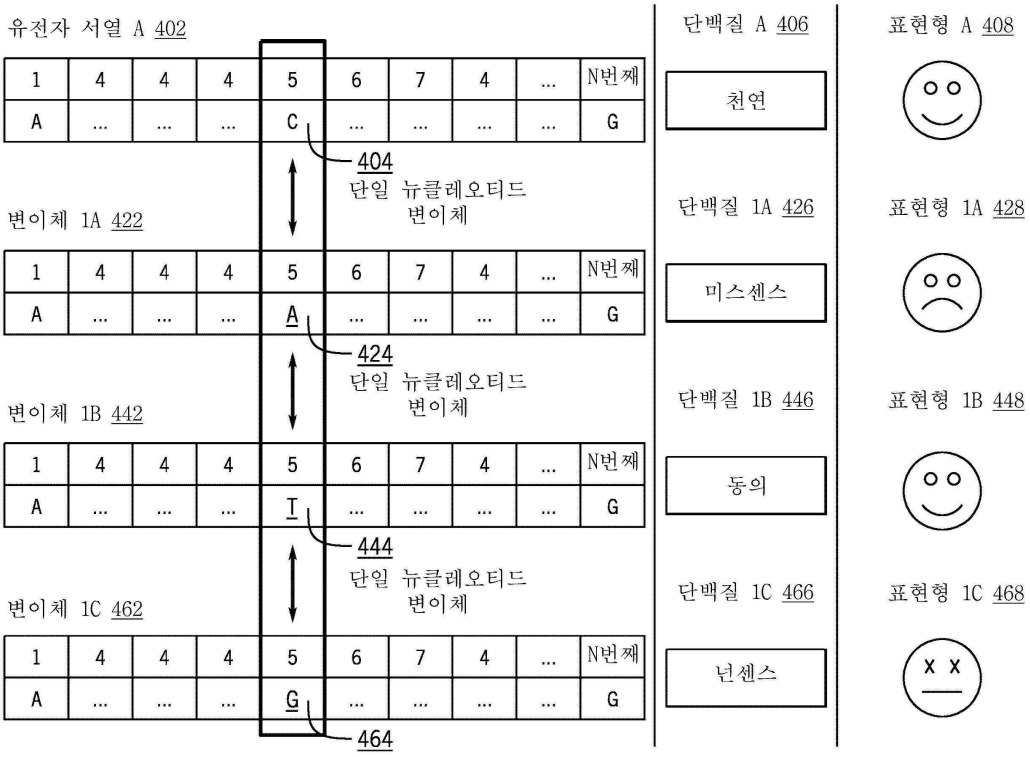
도면2


컴퓨터 시스템 200
저장 서브시스템 210
메모리 서브시스템 222
RAM 232
ROM 234
파일 저장 서브시스템 236
희귀 변이체 다유전자 위험 점수 모델 240
사용자 인터페이스 입력 디바이스 238
버스 서브시스템 242
CPU 244
네트워크 인터페이스 서브시스템 246
사용자 인터페이스 출력 디바이스 248
프로세서 (GPU, FPGA, CGRA) 249

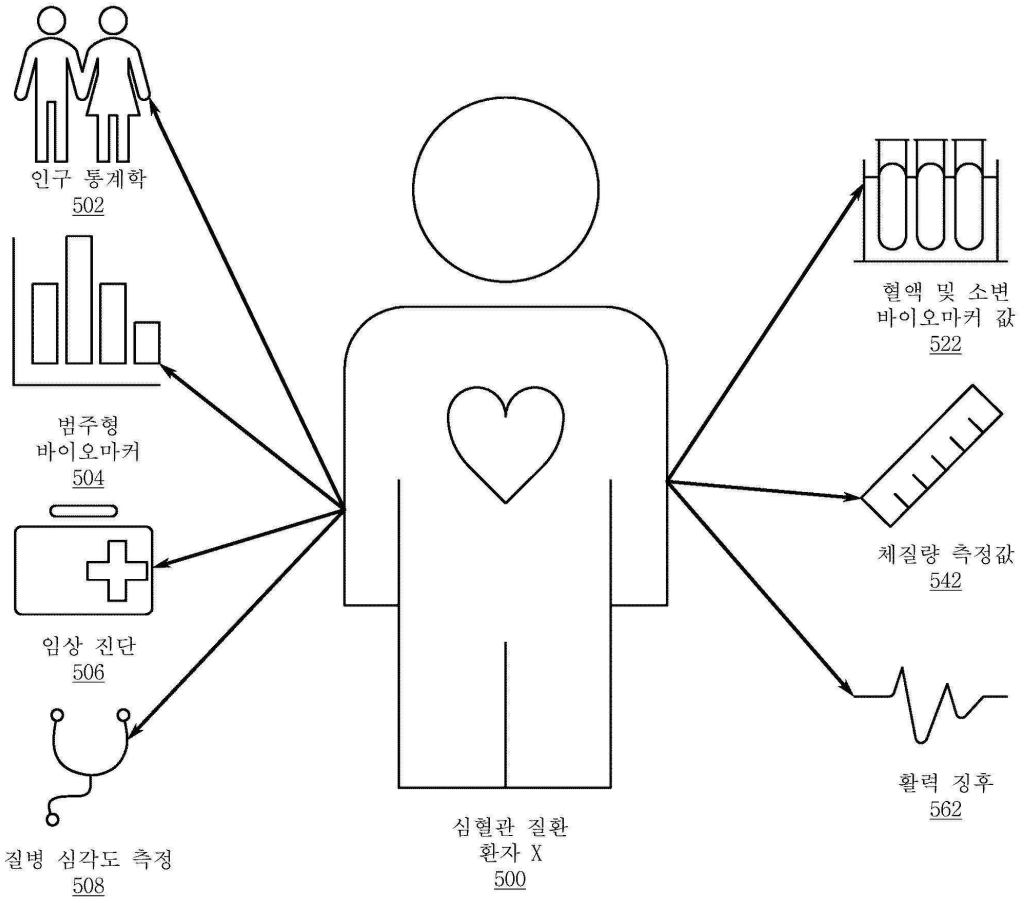
도면3


유전자 서열 A 302
1 2 2 4 5 6 7 8 ... N번째
A ... ... C ... ... G
304
단일 뉴클레오티드 변이체
변이체 1A 322
1 2 2 4 5 6 7 8 ... N번째
A ... ... A ... ... G
324
단일 뉴클레오티드 변이체
변이체 1B 342
1 2 2 4 5 6 7 8 ... N번째
A ... ... T ... ... G
344

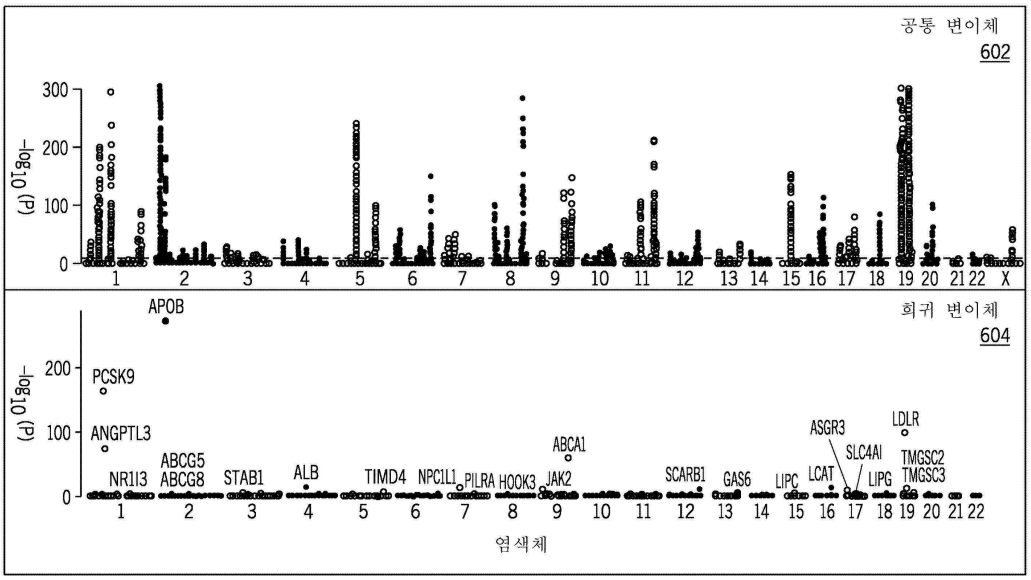
도면4


유전자 서열 A 402
1 4 4 4 5 6 7 4 ... N번째
A ... ... C ... ... G
404
단일 뉴클레오티드 변이체
변이체 1A 422
1 4 4 4 5 6 7 4 ... N번째
A ... ... A ... ... G
424
단일 뉴클레오티드 변이체
변이체 1B 442
1 4 4 4 5 6 7 4 ... N번째
A ... ... T ... ... G
444
단일 뉴클레오티드 변이체
변이체 1C 462
1 4 4 4 5 6 7 4 ... N번째
A ... ... G ... ... G
464
단백질 A 406
천연
표현형 A 408
단백질 1A 426
미스센스
표현형 1A 428
단백질 1B 446
동의
표현형 1B 448
단백질 1C 466
넌센스
표현형 1C 468

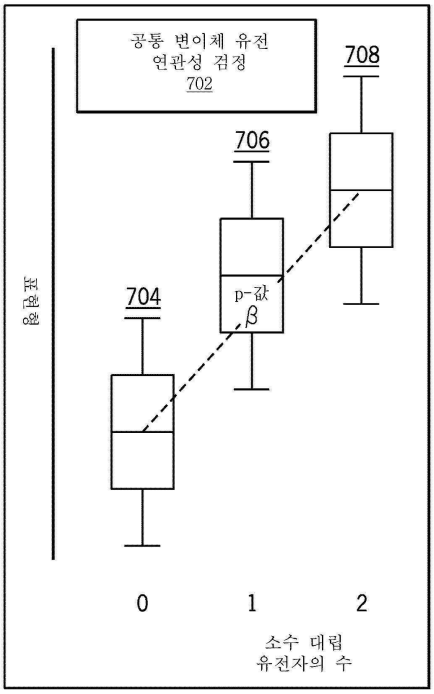
도면5

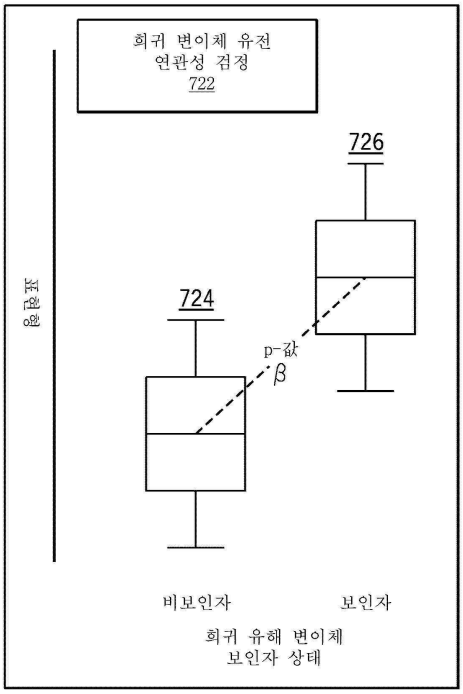

인구 통계학
502
범주형 바이오마커
504
임상 진단
506
질병 심각도 측정
508
심혈관 질환 환자 X
500
혈액 및 소변 바이오마커 값
522
체질량 측정값
542
활력 징후
562

도면6

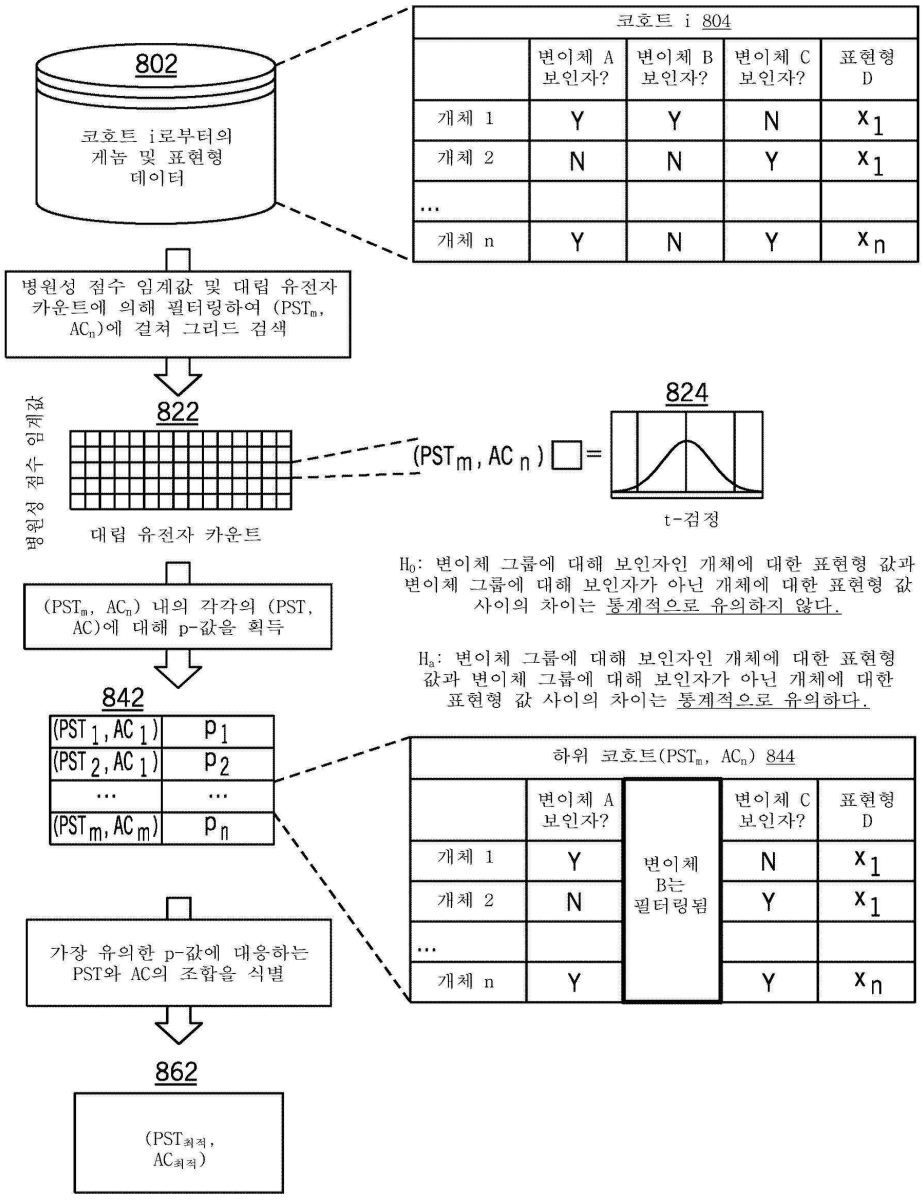

공통 변이체
602
희귀 변이체
604
염색체
APOB
PCSK9
ANGPTL3
NR1I3
ABCG5
ABCG8
STAB1
ALB
TIMD4
NPC1L1
PILRA
HOOK3
ABCA1
JAK2
SCARB1
GAS6
LIPC
LCAT
ASGR3
SLC4A1
LIPG
LDLR
TMGSC2
TMGSC3

도면7

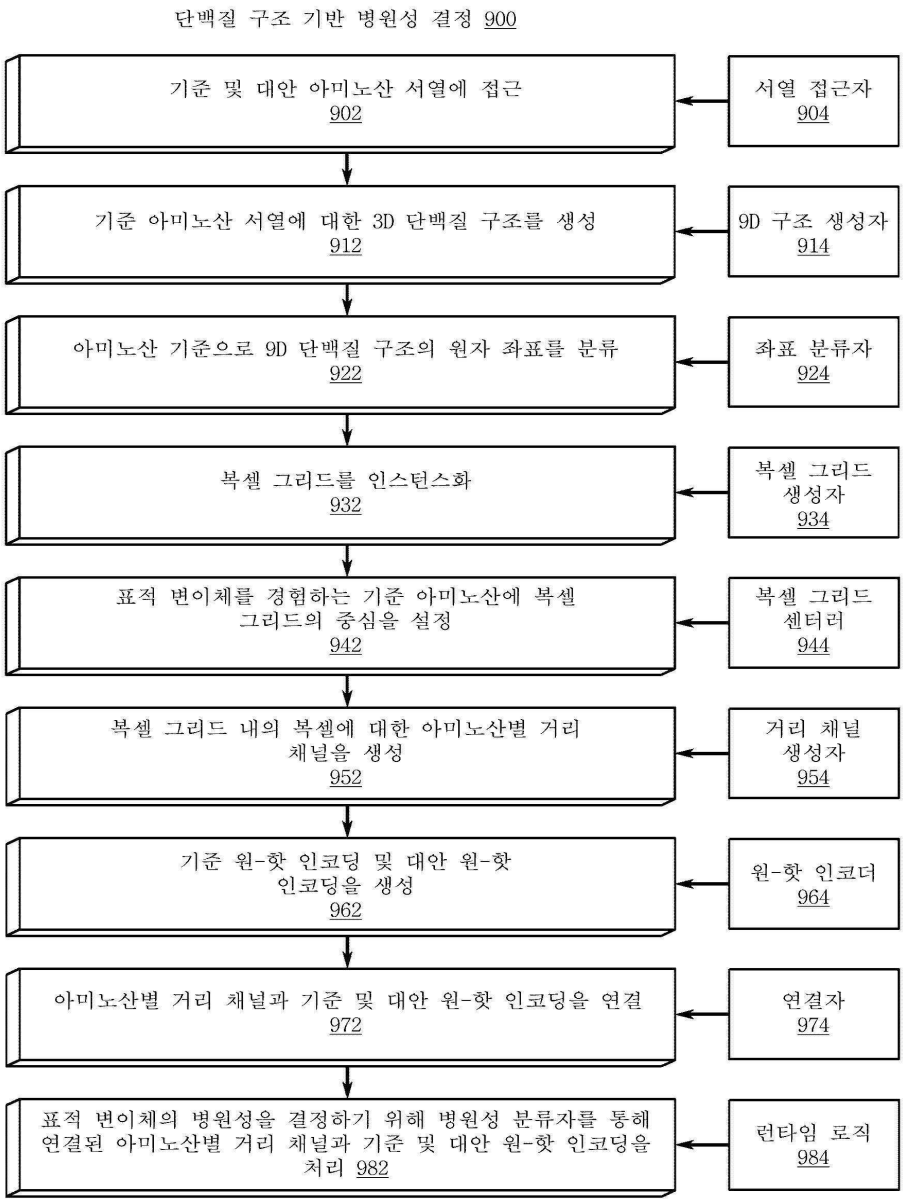

공통 변이체 유전
연관성 검정
702
704
706
708
p-값
β
표현형
0
1
2
소수 대립
유전자의 수


희귀 변이체 유전
연관성 검정
722
724
726
p-값
β
표현형
비보인자
보인자
희귀 유해 변이체
보인자 상태

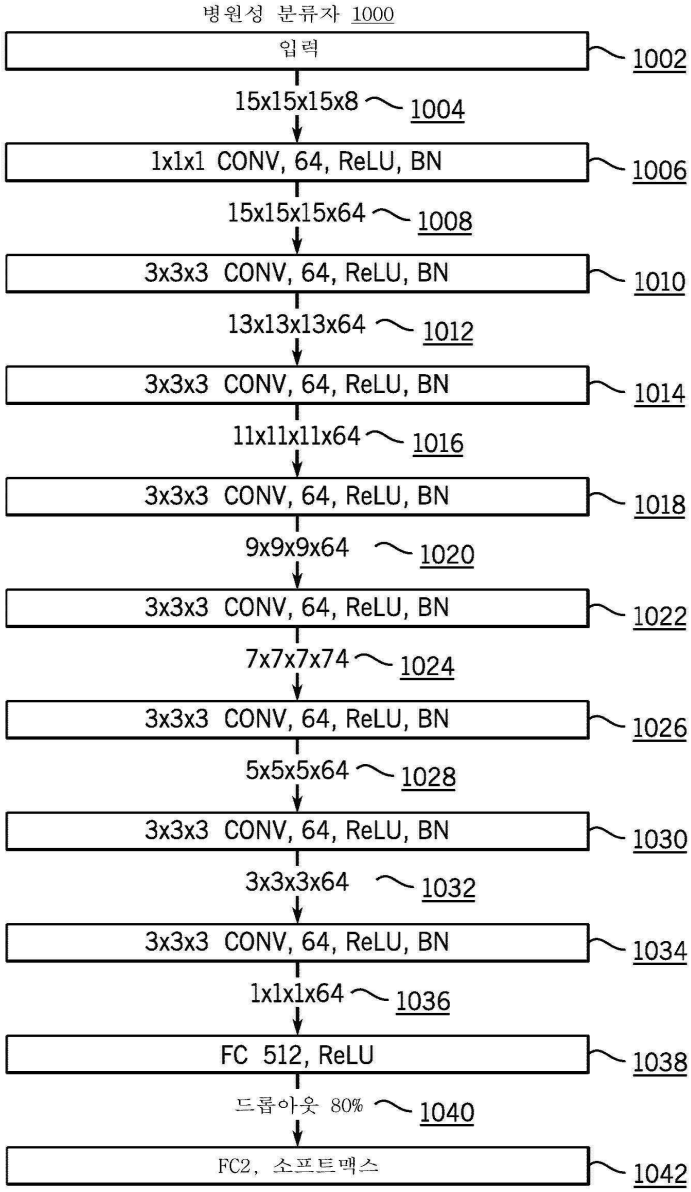
도면8


802
코호트 i로부터의 게놈 및 표현형 데이터
코호트 i 804
변이체 A 보인자?
변이체 B 보인자?
변이체 C 보인자?
표현형 D
개체 1
개체 2
개체 n
병원성 점수 임계값 및 대립 유전자 카운트에 의해 필터링하여 (PSTm, ACn)에 걸쳐 그리드 검색
822
병원성 점수 임계값
대립 유전자 카운트
824
t-검정
H0: 변이체 그룹에 대해 보인자인 개체에 대한 표현형 값과 변이체 그룹에 대해 보인자가 아닌 개체에 대한 표현형 값 사이의 차이는 통계적으로 유의하지 않다.
Ha: 변이체 그룹에 대해 보인자인 개체에 대한 표현형 값과 변이체 그룹에 대해 보인자가 아닌 개체에 대한 표현형 값 사이의 차이는 통계적으로 유의하다.
(PSTm, ACn) 내의 각각의 (PST, AC)에 대해 p-값을 획득
842
가장 유의한 p-값에 대응하는 PST와 AC의 조합을 식별
862
(PST최적, AC최적)
하위 코호트(PSTm, ACn) 844
변이체 B는 필터링됨

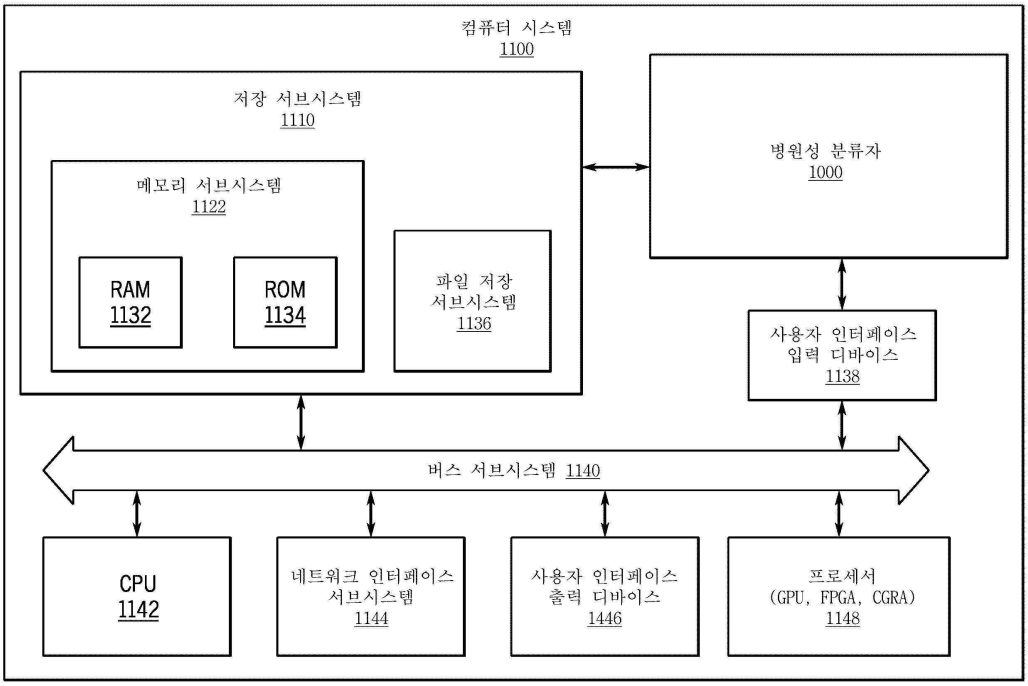
도면9


단백질 구조 기반 병원성 결정 900
기준 및 대안 아미노산 서열에 접근 902
서열 접근자 904
기준 아미노산 서열에 대한 3D 단백질 구조를 생성 912
9D 구조 생성자 914
아미노산 기준으로 9D 단백질 구조의 원자 좌표를 분류 922
좌표 분류자 924
복셀 그리드를 인스턴스화 932
복셀 그리드 생성자 934
표적 변이체를 경험하는 기준 아미노산에 복셀 그리드의 중심을 설정 942
복셀 그리드 센터러 944
복셀 그리드 내의 복셀에 대한 아미노산별 거리 채널을 생성 952
거리 채널 생성자 954
기준 원-핫 인코딩 및 대안 원-핫 인코딩을 생성 962
원-핫 인코더 964
아미노산별 거리 채널과 기준 및 대안 원-핫 인코딩을 연결 972
연결자 974
표적 변이체의 병원성을 결정하기 위해 병원성 분류자를 통해 연결된 아미노산별 거리 채널과 기준 및 대안 원-핫 인코딩을 처리 982
런타임 로직 984

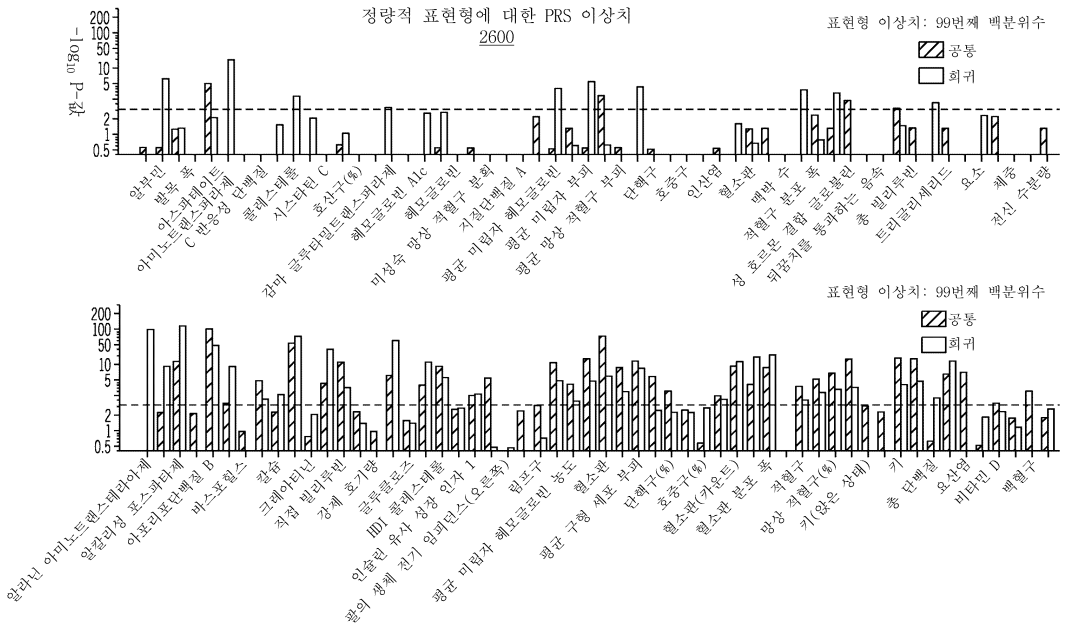
도면10


병원성 분류자 1000
입력
1002
15x15x15x8
1004
1x1x1 CONV, 64, ReLU, BN
1006
15x15x15x64
1008
3x3x3 CONV, 64, ReLU, BN
1010
13x13x13x64
1012
3x3x3 CONV, 64, ReLU, BN
1014
11x11x11x64
1016
3x3x3 CONV, 64, ReLU, BN
1018
9x9x9x64
1020
3x3x3 CONV, 64, ReLU, BN
1022
7x7x7x74
1024
3x3x3 CONV, 64, ReLU, BN
1026
5x5x5x64
1028
3x3x3 CONV, 64, ReLU, BN
1030
3x3x3x64
1032
3x3x3 CONV, 64, ReLU, BN
1034
1x1x1x64
1036
FC 512, ReLU
1038
드롭아웃 80%
1040
FC2, 소프트맥스
1042

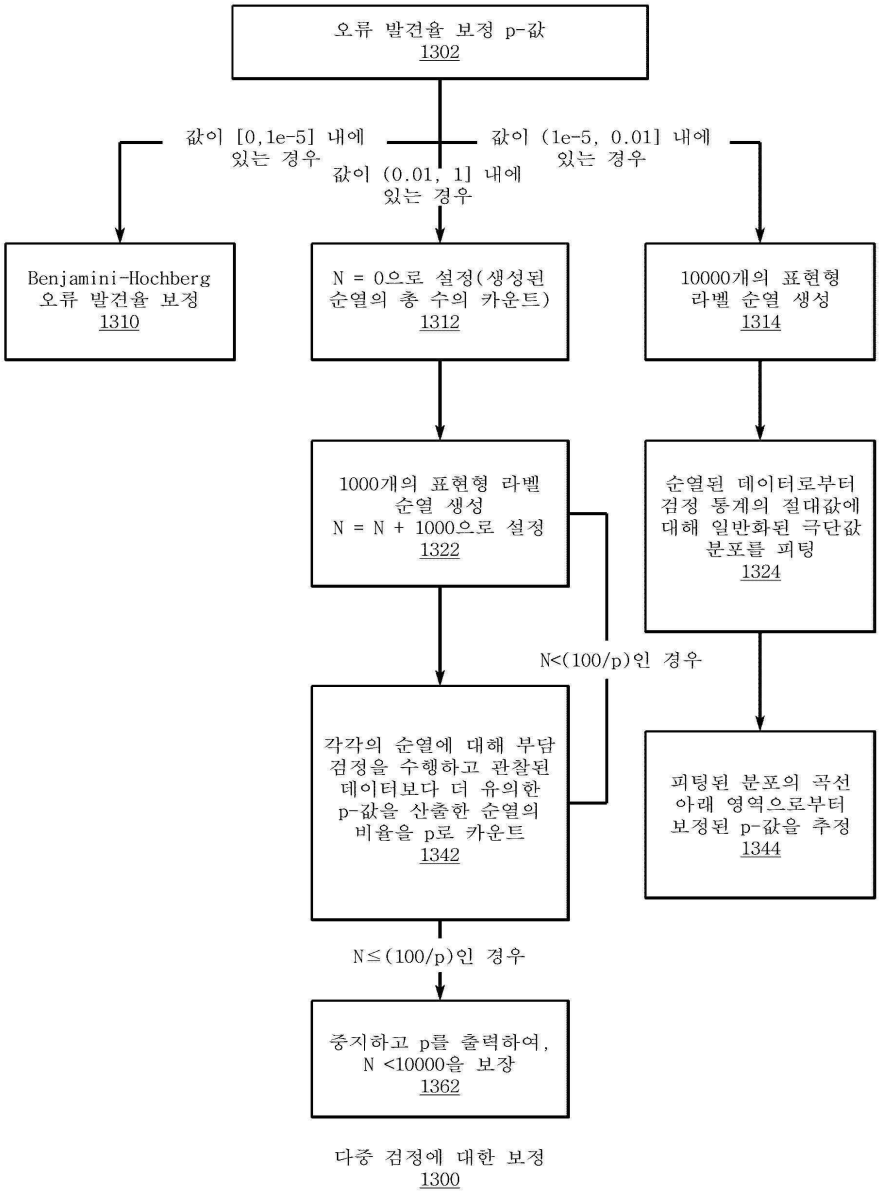
도면11


컴퓨터 시스템
1100
저장 서브시스템
1110
메모리 서브시스템
1122
RAM
1132
ROM
1134
파일 저장
서브시스템
1136
병원성 분류자
1000
사용자 인터페이스
입력 디바이스
1138
버스 서브시스템 1140
CPU
1142
네트워크 인터페이스
서브시스템
1144
사용자 인터페이스
출력 디바이스
1446
프로세서
(GPU, FPGA, CGRA)
1148

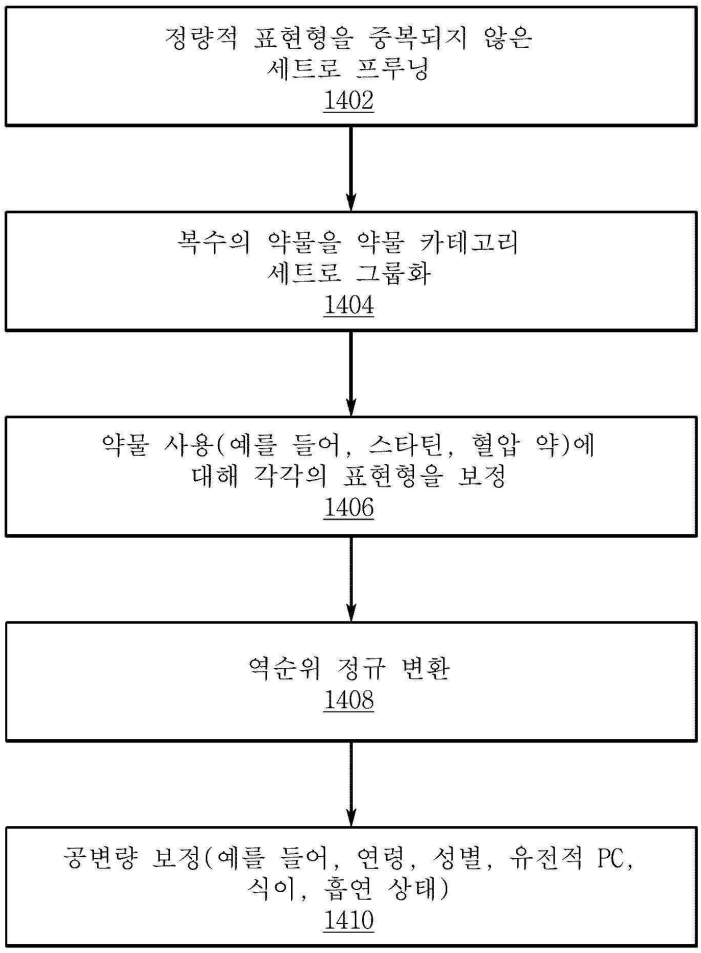
도면12


정량적 표현형에 대한 PRS 이상치
2600
표현형 이상치: 99번째 백분위수
공통
희귀
−log10 P-값
표현형 이상치: 99번째 백분위수
공통
희귀

도면13

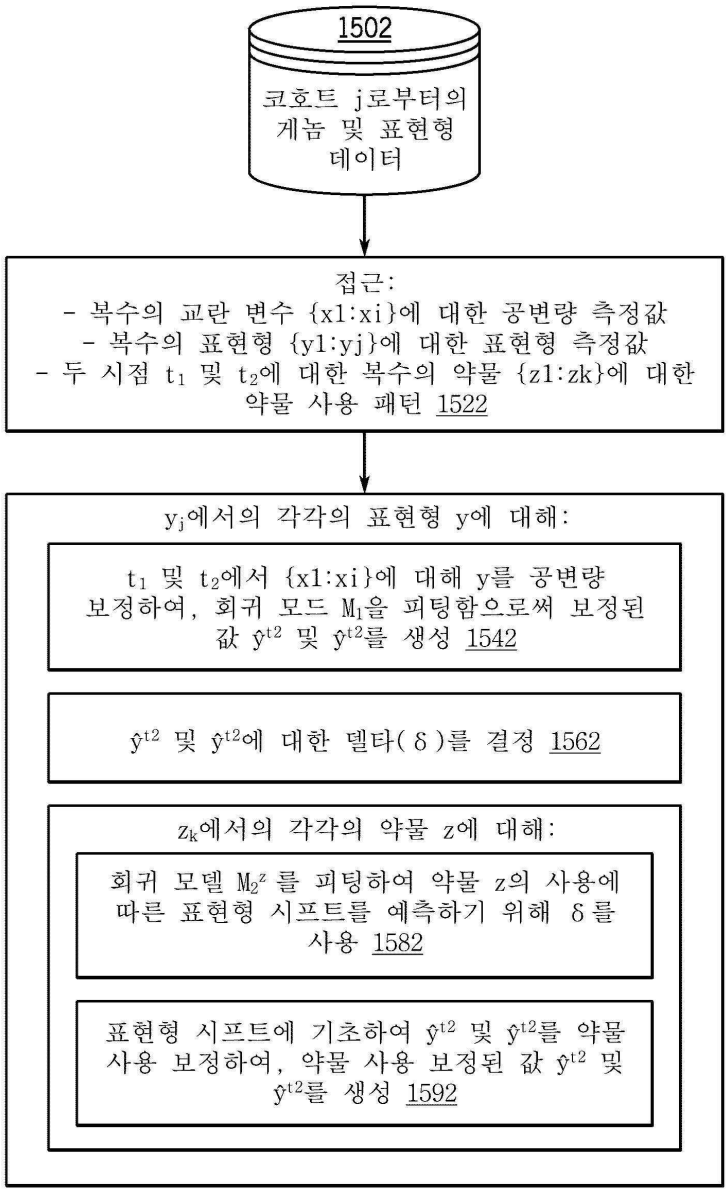

오류 발견율 보정 p-값
1302
값이 [0,1e-5] 내에 있는 경우
값이 (1e-5, 0.01] 내에 있는 경우
값이 (0.01, 1] 내에 있는 경우
Benjamini-Hochberg 오류 발견율 보정
1310
N = 0으로 설정(생성된 순열의 총 수의 카운트)
1312
10000개의 표현형 라벨 순열 생성
1314
1000개의 표현형 라벨 순열 생성
N = N + 1000으로 설정
1322
순열된 데이터로부터 검정 통계의 절대값에 대해 일반화된 극단값 분포를 피팅
1324
N<(100/p)인 경우
각각의 순열에 대해 부담 검정을 수행하고 관찰된 데이터보다 더 유의한 p-값을 산출한 순열의 비율을 p로 카운트
1342
피팅된 분포의 곡선 아래 영역으로부터 보정된 p-값을 추정
1344
N≤(100/p)인 경우
중지하고 p를 출력하여, N <10000을 보장
1362
다중 검정에 대한 보정
1300

도면14

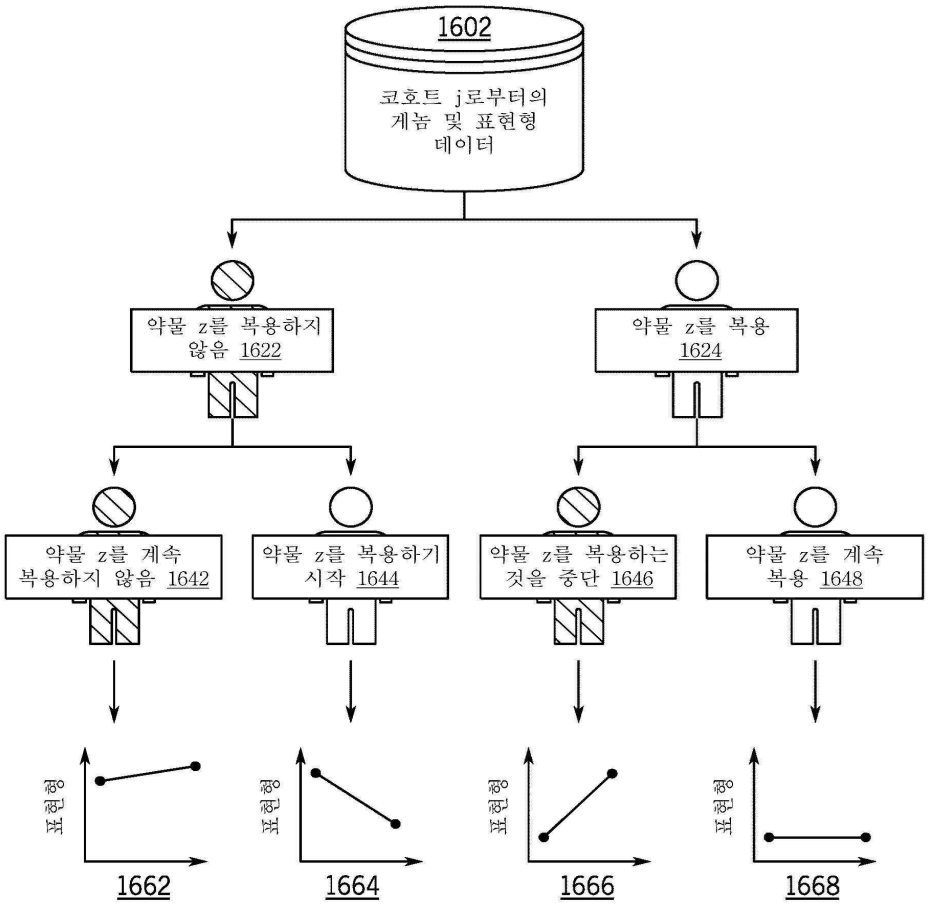


공변량에 대한 보정
1400

도면15


1502
코호트 j로부터의 게놈 및 표현형 데이터
접근:
- 복수의 교란 변수 {x1:xi}에 대한 공변량 측정값
- 복수의 표현형 {y1:yj}에 대한 표현형 측정값
- 두 시점 t_1 및 t_2에 대한 복수의 약물 {z1:zk}에 대한 약물 사용 패턴 1522
y_j에서의 각각의 표현형 y에 대해:
t_1 및 t_2에서 {x1:xi}에 대해 y를 공변량 보정하여, 회귀 모드 M_1을 피팅함으로써 보정된 값 $\hat{y}^{t2}$ 및 $\hat{y}^{t2}$를 생성 1542
$\hat{y}^{t2}$ 및 $\hat{y}^{t2}$에 대한 델타(δ)를 결정 1562
z_k에서의 각각의 약물 z에 대해:
회귀 모델 M_2^z를 피팅하여 약물 z의 사용에 따른 표현형 시프트를 예측하기 위해 δ를 사용 1582
표현형 시프트에 기초하여 $\hat{y}^{t2}$ 및 $\hat{y}^{t2}$를 약물 사용 보정하여, 약물 사용 보정된 값 $\hat{y}^{t2}$ 및 $\hat{y}^{t2}$를 생성 1592

도면16



1. 약물 효과에 대한 학습 모델:

$$\delta_y = y_2 - y_1 = \beta_1 S + \varepsilon$$

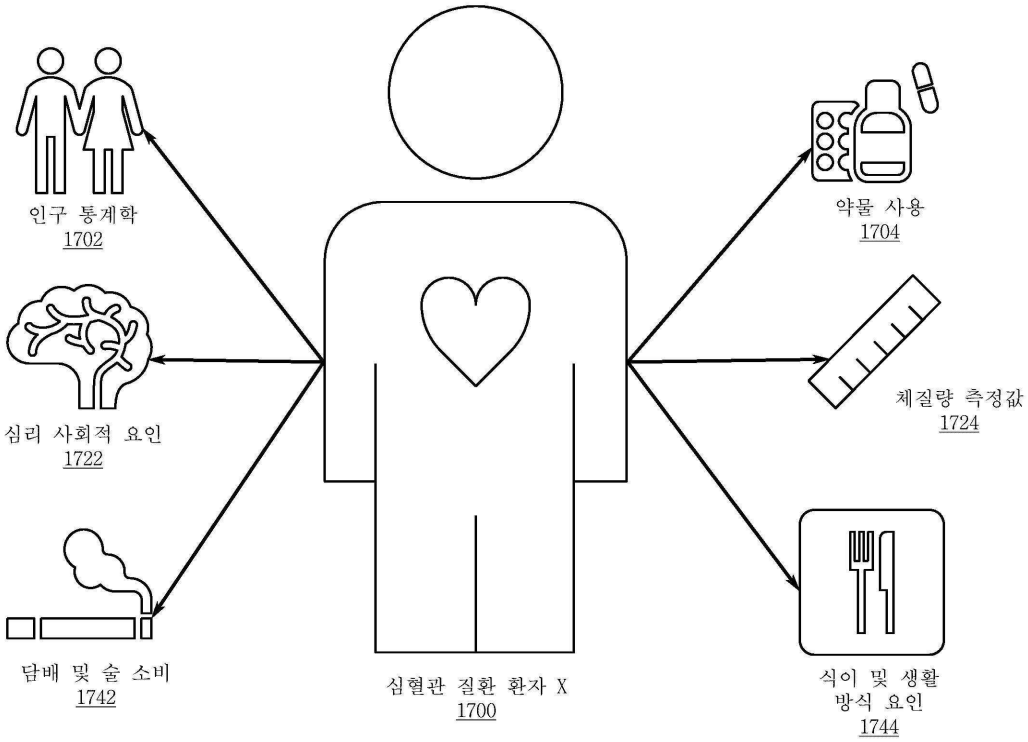
1682

2. 유의성에 대한 검정 효과:

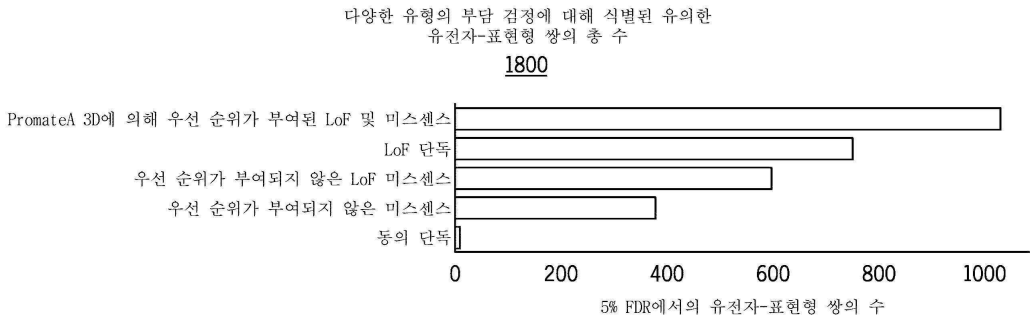
$$\beta_z \neq 0$$

1692

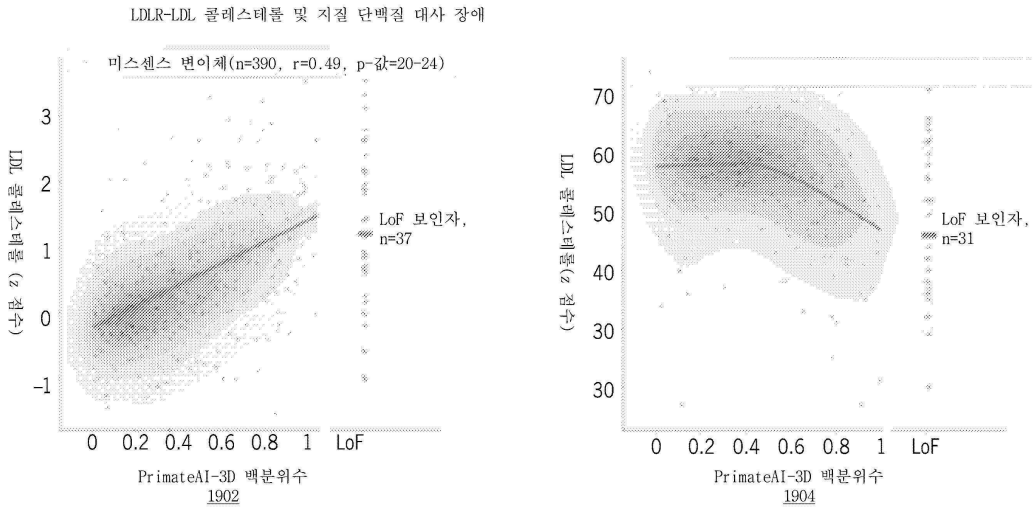
도면17



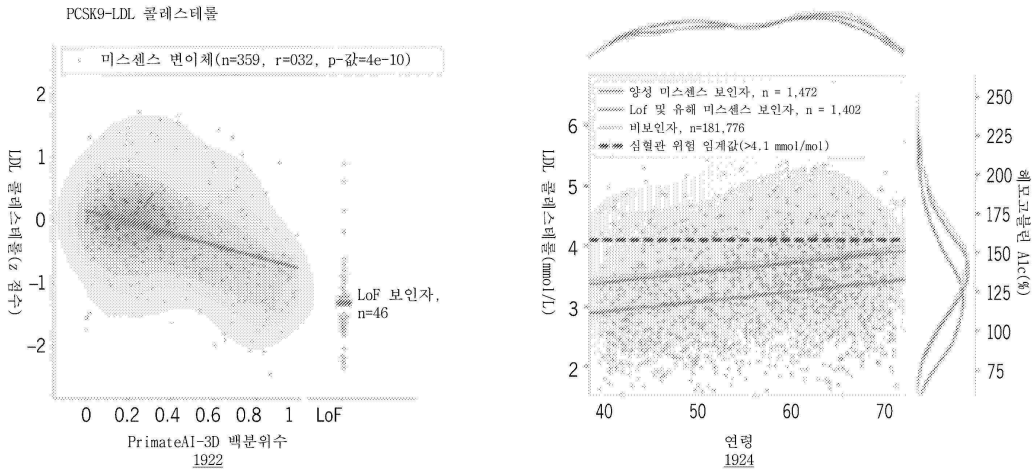
도면18



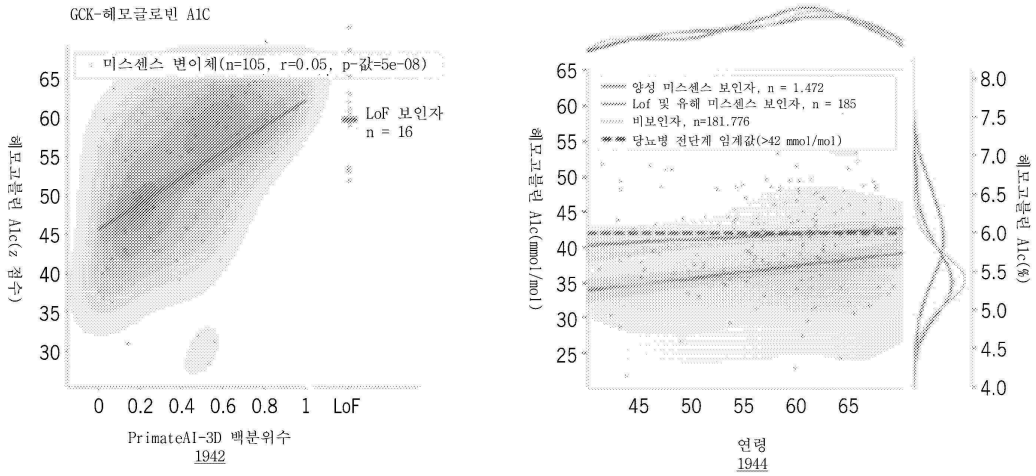
도면19a



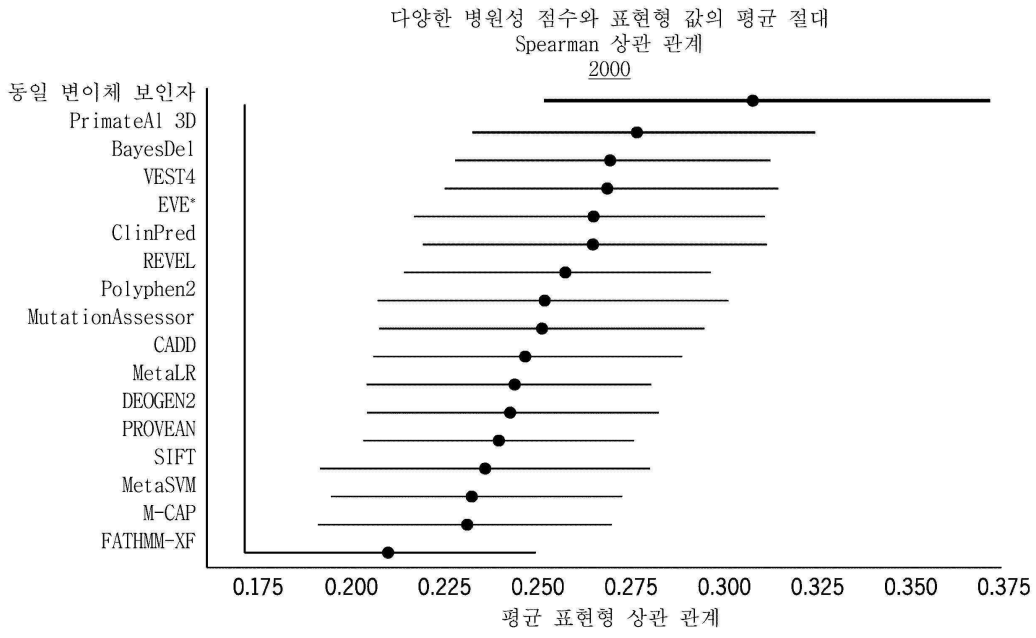
도면19b



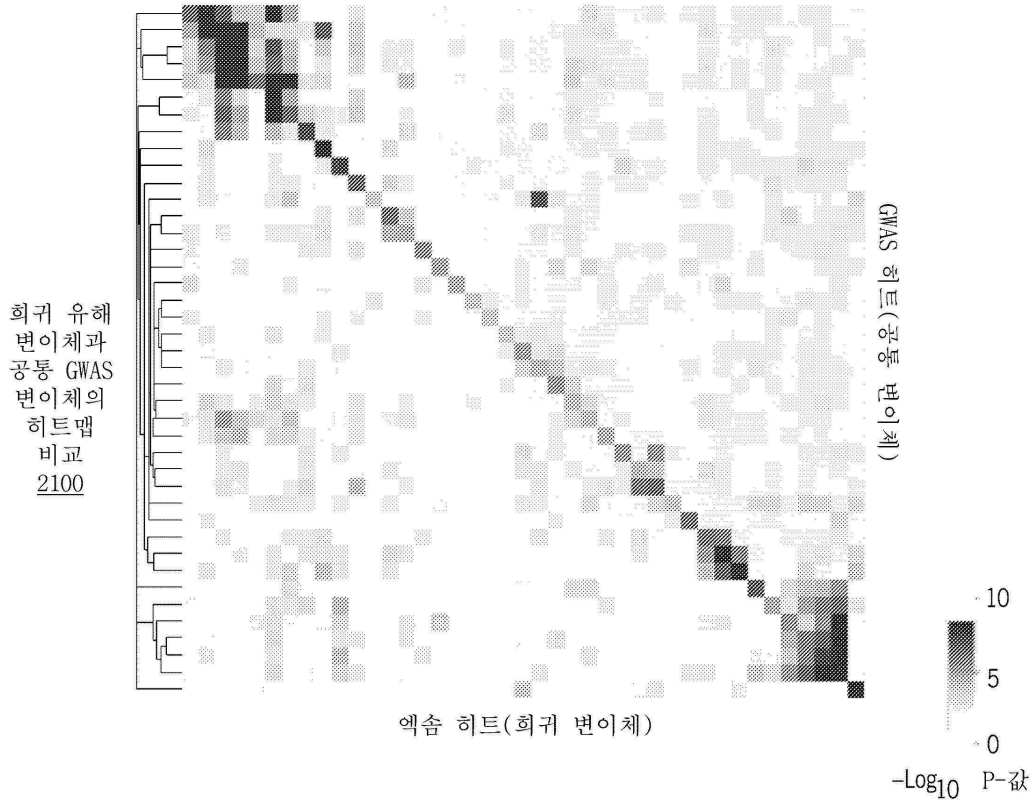
도면19c



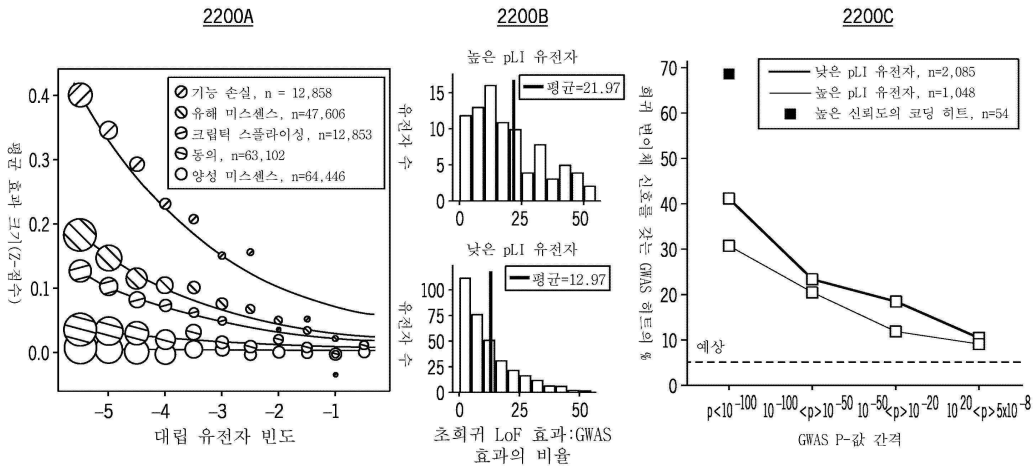
도면20



도면21

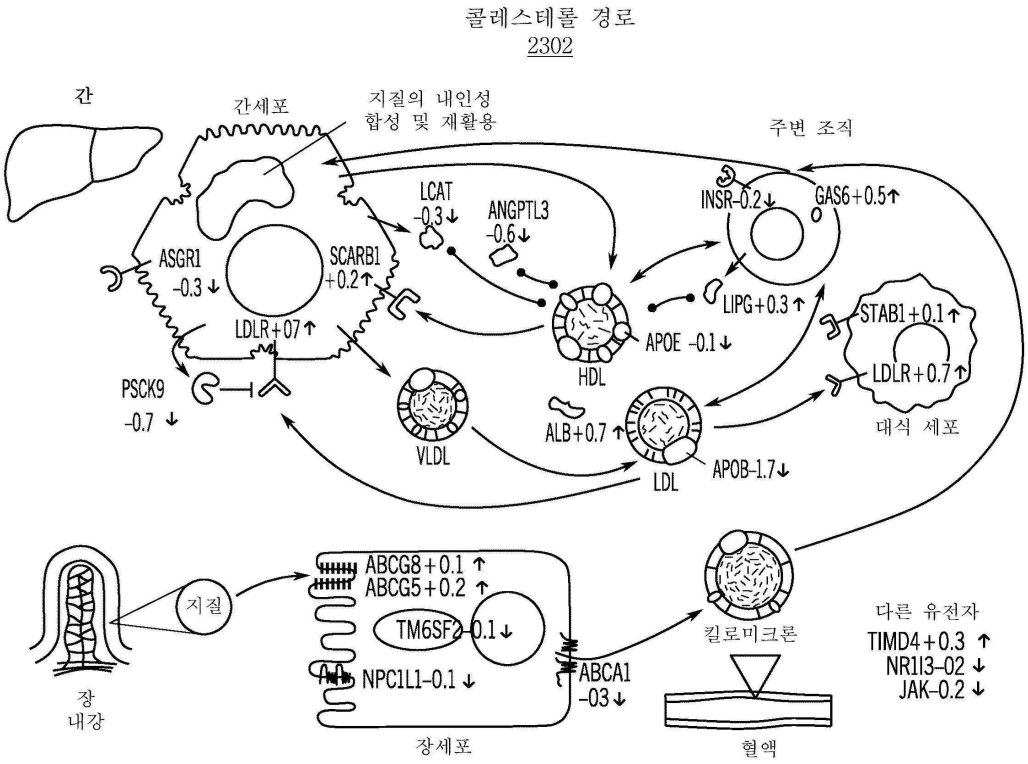


도면22

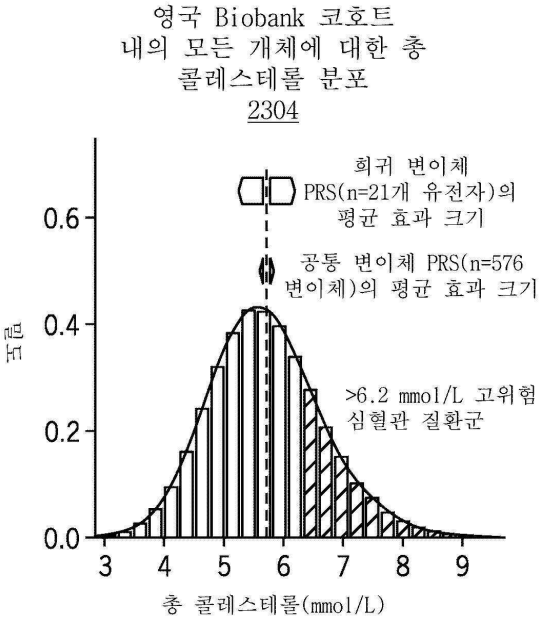


희귀 유해 변이체와 공통 GWAS 변이체의 비교 2200

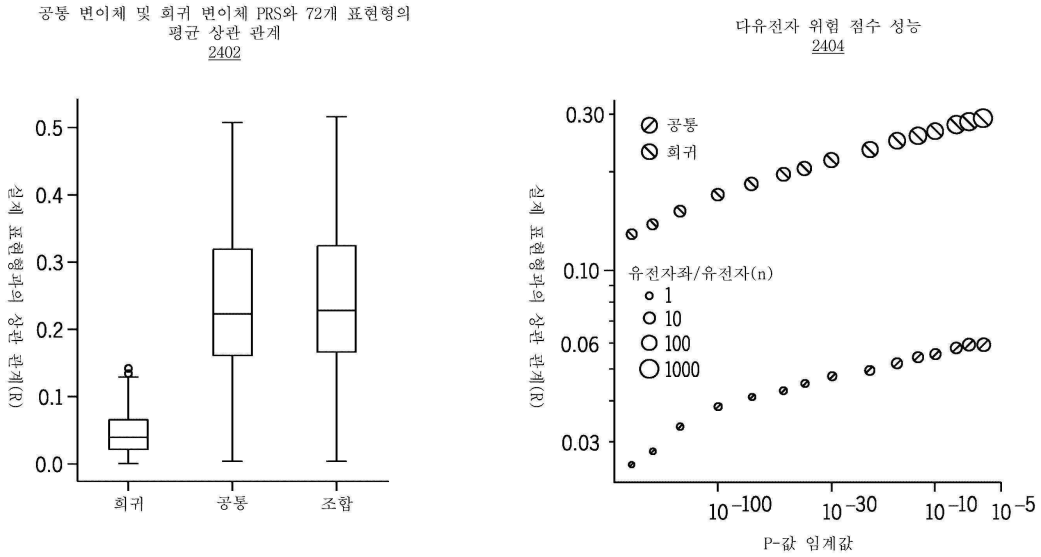
도면23a


콜레스테롤 경로
2302
간
간세포
지질의 내인성
합성 및 재활용
주변 조직
LCAT
-0.3↓
ANGPTL3
-0.6↓
INSR-0.2↓
GAS6+0.5↑
ASGR1
-0.3↓
SCARB1
+0.2↑
LDLR+07↑
LIPG+0.3↑
STAB1+0.1↑
APOE -0.1↓
HDL
PSCK9
-0.7↓
VLDL
ALB+0.7↑
LDL
APOB-1.7↓
LDLR+0.7↑
대식 세포
지질
ABCG8+0.1↑
ABCG5+0.2↑
TM6SF2-0.1↓
NPC1L1-0.1↓
ABCA1
-03↓
킬로미크론
다른 유전자
TIMD4+0.3↑
NR1I3-02↓
JAK-0.2↓
장
내강
장세포
혈액

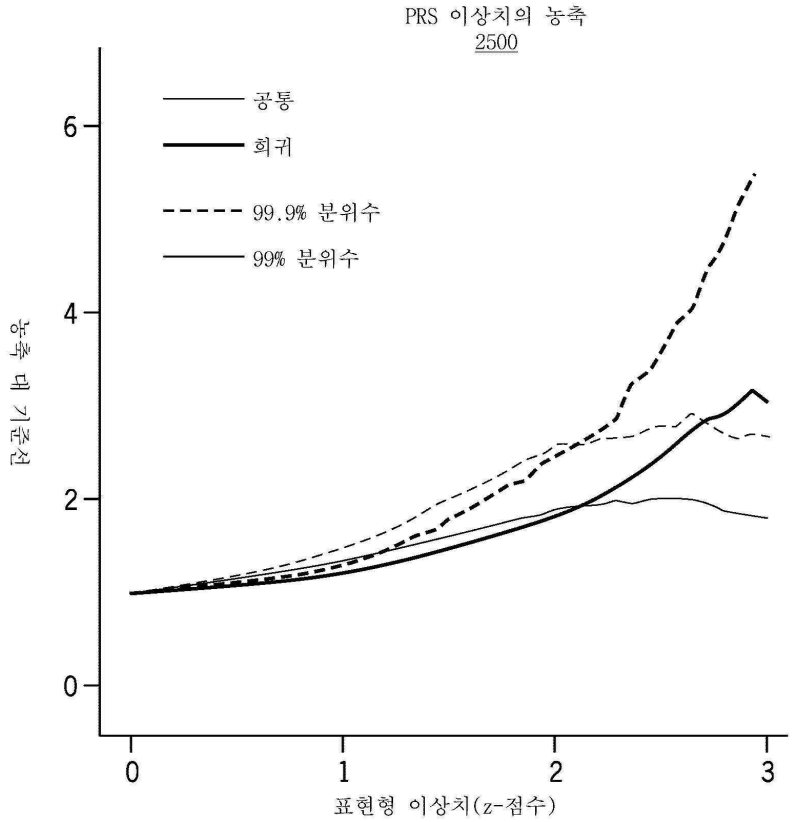
도면23b


영국 Biobank 코호트
내의 모든 개체에 대한 총
콜레스테롤 분포
2304
희귀 변이체
PRS(n=21개 유전자)의
평균 효과 크기
공통 변이체 PRS(n=576
변이체)의 평균 효과 크기
>6.2 mmol/L 고위험
심혈관 질환군
밀도
0.6
0.4
0.2
0.0
3
4
5
6
7
8
9
총 콜레스테롤(mmol/L)

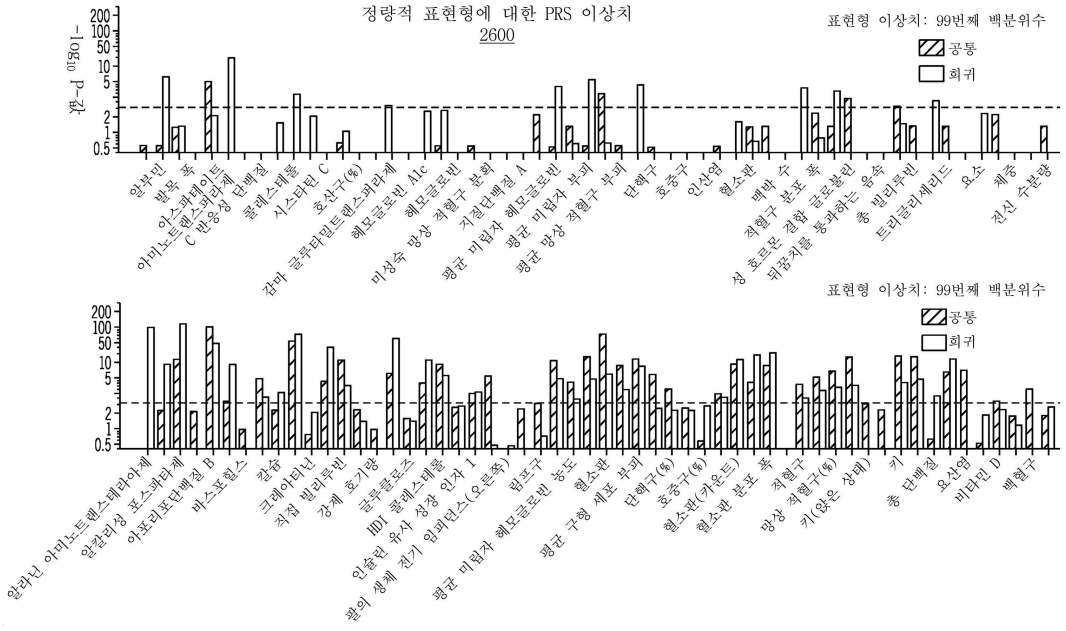
도면24


공통 변이체 및 희귀 변이체 PRS와 72개 표현형의 평균 상관 관계
2402
실제 표현형과의 상관 관계(R)
0.5
0.4
0.3
0.2
0.1
0.0
희귀
공통
조합
다유전자 위험 점수 성능
2404
실제 표현형과의 상관 관계(R)
0.30
0.10
0.06
0.03
공통
희귀
유전자좌/유전자(n)
1
10
100
1000
10-100
10-30
10-10
10-5
P-값 임계값

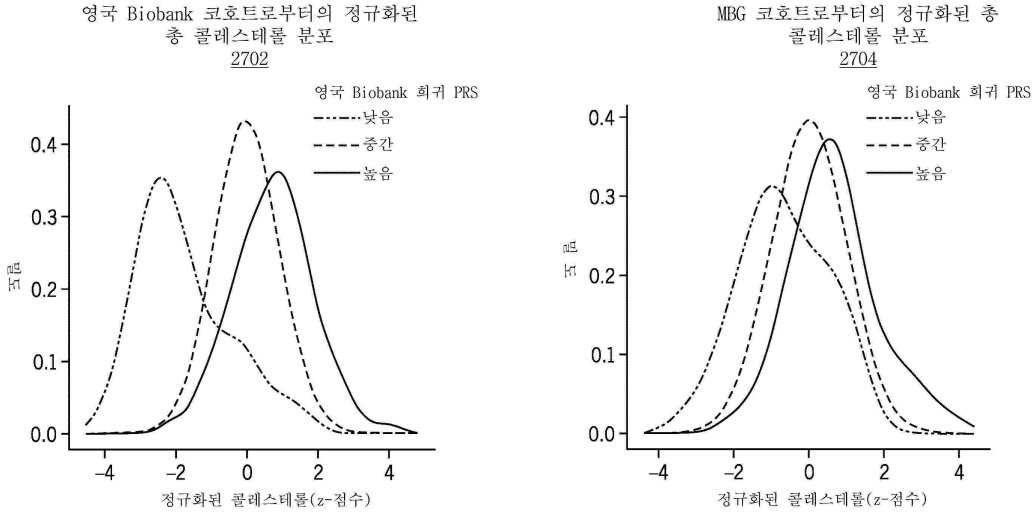
도면25


PRS 이상치의 농축
2500
공통
희귀
99.9% 분위수
99% 분위수
농축 대 기준선
6
4
2
0
0
1
2
3
표현형 이상치(z-점수)

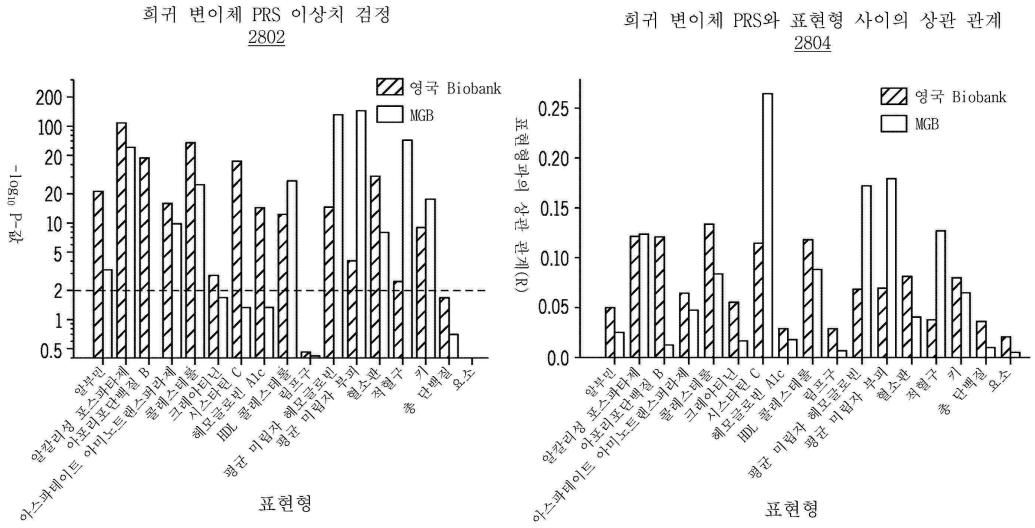
도면26


정량적 표현형에 대한 PRS 이상치
2600
표현형 이상치: 99번째 백분위수
공통
희귀
−log10 p-값
표현형 이상치: 99번째 백분위수
공통
희귀

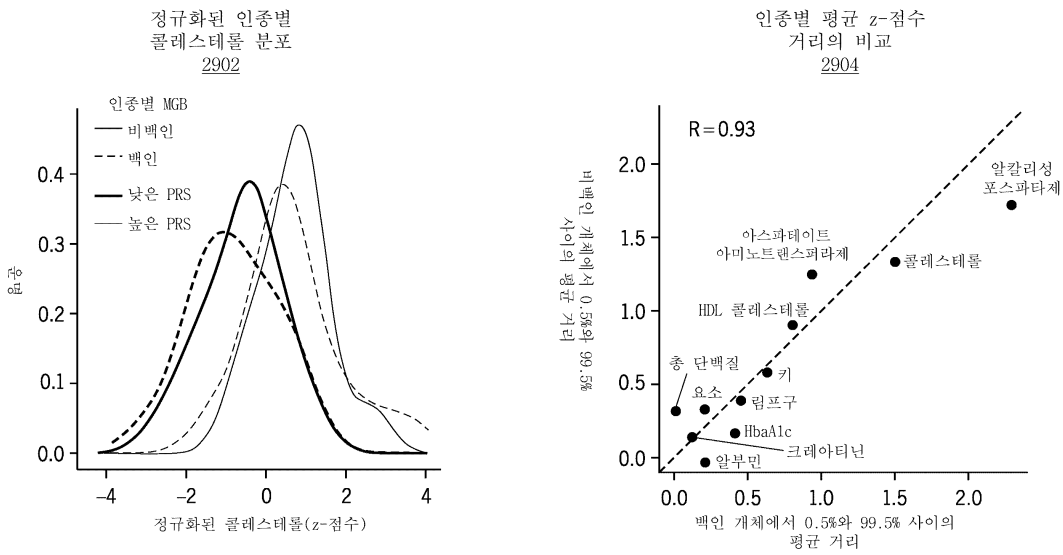
도면27


영국 Biobank 코호트로부터의 정규화된
총 콜레스테롤 분포
2702
영국 Biobank 희귀 PRS
낮음
중간
높음
밀도
정규화된 콜레스테롤(z-점수)
MBG 코호트로부터의 정규화된 총
콜레스테롤 분포
2704
영국 Biobank 희귀 PRS
낮음
중간
높음
밀도
정규화된 콜레스테롤(z-점수)

도면28


회귀 변이체 PRS 이상치 검정
2802
영국 Biobank
MGB
-log10 p-값
표현형
회귀 변이체 PRS와 표현형 사이의 상관 관계
2804
영국 Biobank
MGB
표현형과의 상관 관계(R)
표현형

도면29


정규화된 인종별
콜레스테롤 분포
2902
인종별 MGB
비백인
백인
낮은 PRS
높은 PRS
밀도
정규화된 콜레스테롤(z-점수)
인종별 평균 z-점수
거리의 비교
2904
R=0.93
비백인 개체에서 0.5%와 99.5% 사이의 평균 거리
백인 개체에서 0.5%와 99.5% 사이의
평균 거리
알칼리성
포스파타제
아스파테이트
아미노트랜스퍼라제
콜레스테롤
HDL 콜레스테롤
총 단백질
키
요소
림프구
HbaA1c
크레아티닌
알부민

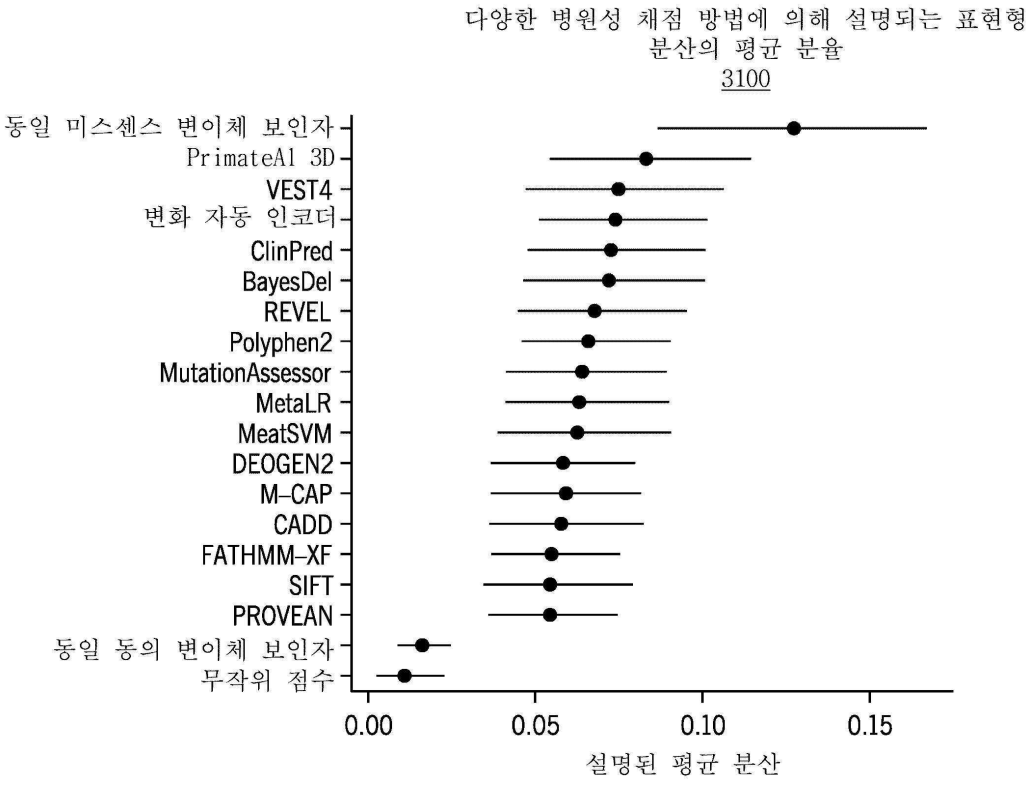
도면30

정상 콜레스테롤 레벨에 사용되는 공통 PRS 변이체와
희귀 PRS 유전자에 대한 효과 크기 및 빈도의 비교

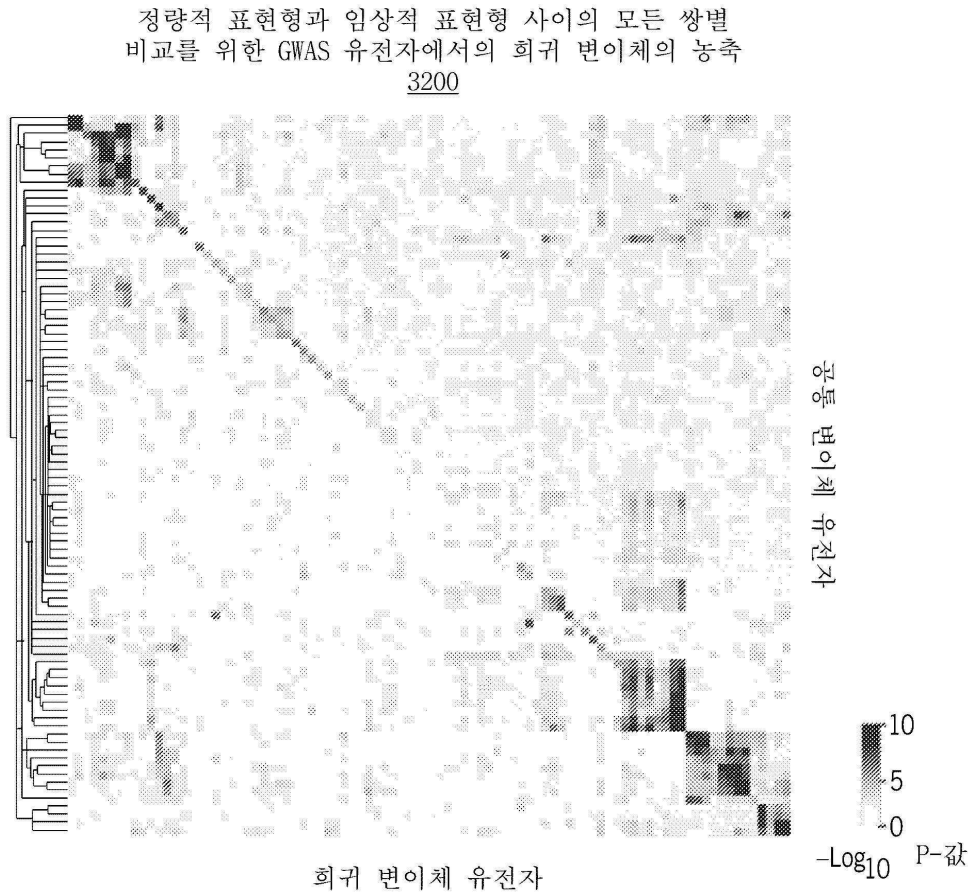
3000

	공통 PRS 변이체						희귀 PRS 유전자			
크롬	위치 (GRCh37)	주요 대립 유전자	소수 대립 유전자	대립 유전자 빈도	효과 크기 (z 점수)	P-값	유전자	집계 대립 유전자 빈도	효과 크기 (z-점수)	P-값
18	47109955	A	G	0.014	0.198	2.6×10^{-84}	*ABCA1*	0.010	−0.344	1.0×10^{-57}
6	32644523	T	C	0.171	0.045	1.8×10^{-54}	*ABCG5*	0.009	0.156	3.5×10^{-13}
6	160575985	C	T	0.373	−0.044	1.8×10^{-127}	*ABCG8*	0.007	0.139	8.1×10^{-8}
13	114544024	T	C	0.297	0.033	1.4×10^{-67}	*ALB*	0.001	0.709	1.2×10^{-16}
6	29925268	C	A	0.068	0.042	4.9×10^{-20}	*ANGPTL3*	0.004	−0.569	1.4×10^{-67}
8	28843833	G	A	0.041	0.032	3.2×10^{-8}	*APOB*	0.002	−1.686	9.7×10^{-268}
11	2963612	C	T	0.391	−0.014	2.1×10^{-14}	*APOE*	0.007	−0.138	7.5×10^{-8}
19	7827830	G	A	0.152	−0.023	1.1×10^{-15}	*ASGR1*	0.002	−0.330	3.9×10^{-11}
14	94847262	T	A	0.049	0.038	8.3×10^{-13}	*GAS6*	0.001	0.479	4.3×10^{-7}
20	62362116	G	C	0.245	0.018	5.1×10^{-13}	*INSR*	0.003	−0.212	3.0×10^{-8}
17	38175462	C	T	0.368	0.016	1.7×10^{-20}	*JAK2*	0.003	−0.213	2.2×10^{-7}
19	46436564	C	T	0.016	−0.115	1.4×10^{-32}	*LCAT*	0.003	−0.339	7.1×10^{-17}
2	43987888	CTT	C	0.067	−0.042	4.7×10^{-21}	*LDLR*	0.004	0.721	1.7×10^{-107}
9	78730766	A	G	0.423	0.014	1.3×10^{-20}	*LIPG*	0.001	0.347	2.4×10^{-8}
5	156369171	T	C	0.347	−0.046	1.0×10^{-154}	*NPC1L1*	0.015	−0.128	8.4×10^{-14}
9	107589744	C	T	0.153	0.055	3.6×10^{-71}	*NR1I3*	0.002	−0.239	2.7×10^{-7}
			...				*PCSK9*	0.005	−0.714	2.8×10^{-134}
			558개 변이체 생략				*SCARB1*	0.008	0.157	6.7×10^{-11}
			...				*STAB1*	0.014	0.095	2.6×10^{-7}
4	104250412	C	T	0.132	−0.018	4.7×10^{-9}	*TIMD4*	0.002	0.309	2.8×10^{-9}
9	110530324	G	A	0.288	0.013	2.4×10^{-10}	*TM6SF2*	0.009	−0.141	1.7×10^{-10}

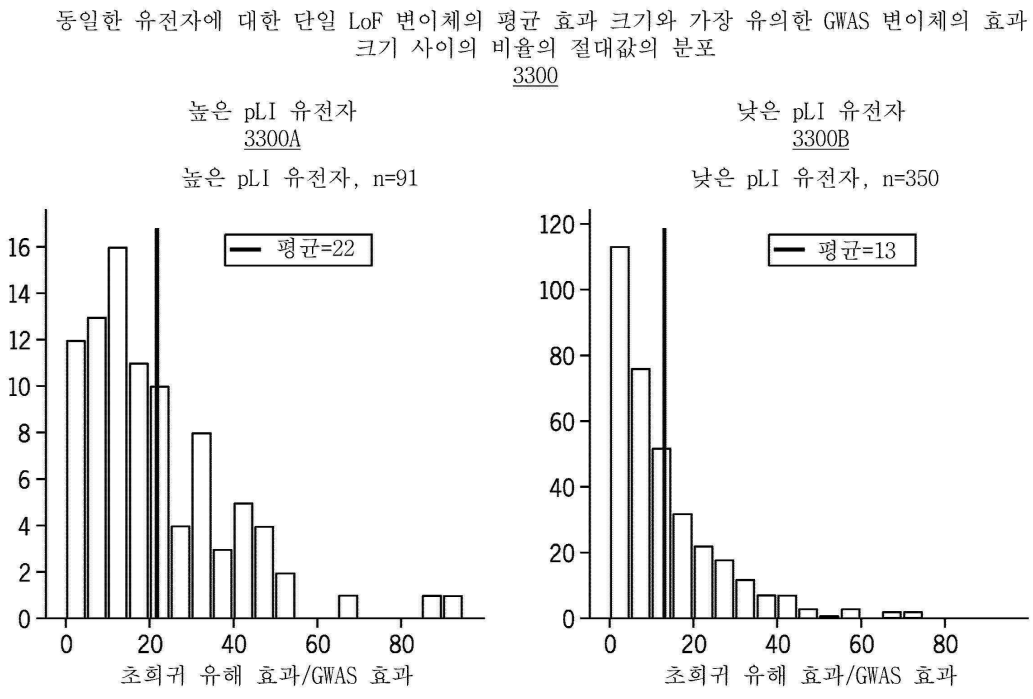
도면31



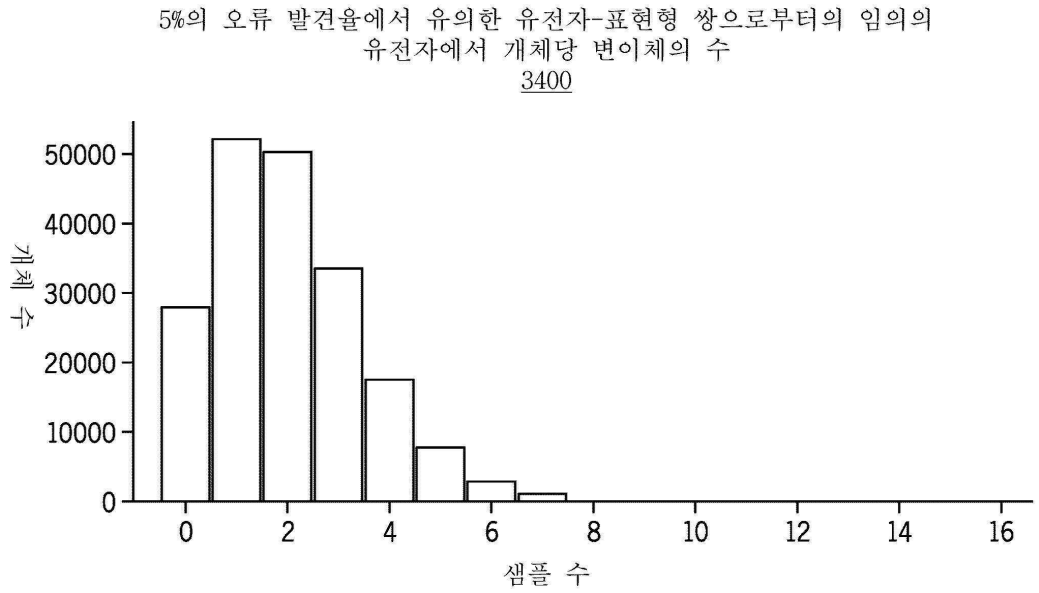
도면32


정량적 표현형과 임상적 표현형 사이의 모든 쌍별
비교를 위한 GWAS 유전자에서의 희귀 변이체의 농축
3200
공통 변이체 유전자
희귀 변이체 유전자
10
5
0
-Log10 P-값

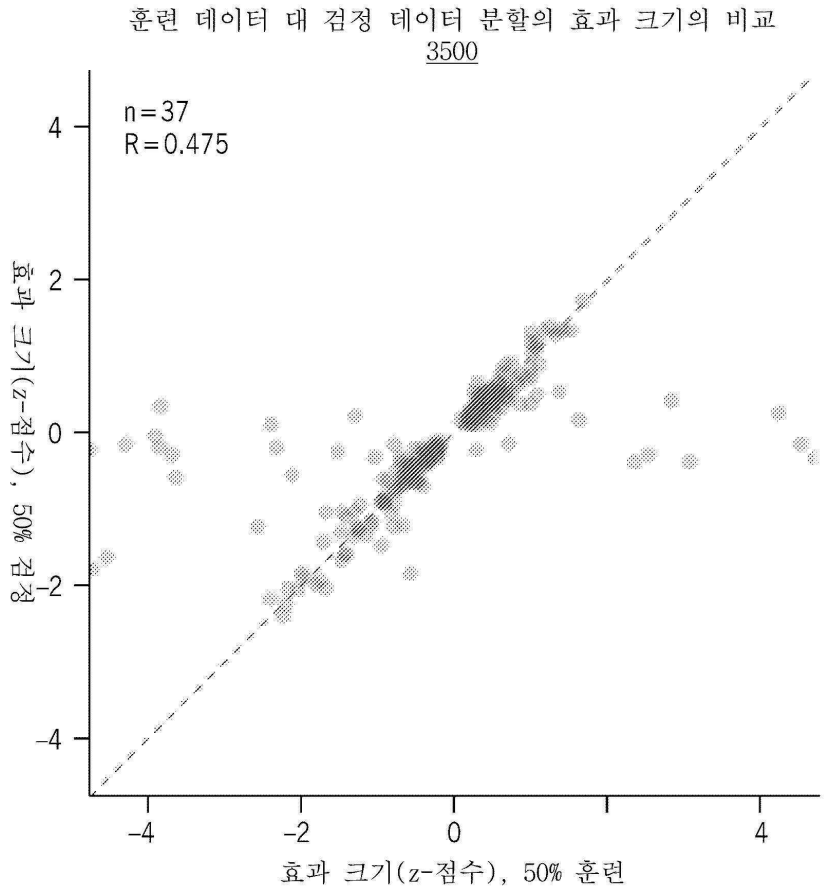
도면33


동일한 유전자에 대한 단일 LoF 변이체의 평균 효과 크기와 가장 유의한 GWAS 변이체의 효과
크기 사이의 비율의 절대값의 분포
3300
높은 pLI 유전자
3300A
낮은 pLI 유전자
3300B
높은 pLI 유전자, n=91
낮은 pLI 유전자, n=350
평균=22
평균=13
초희귀 유해 효과/GWAS 효과
초희귀 유해 효과/GWAS 효과

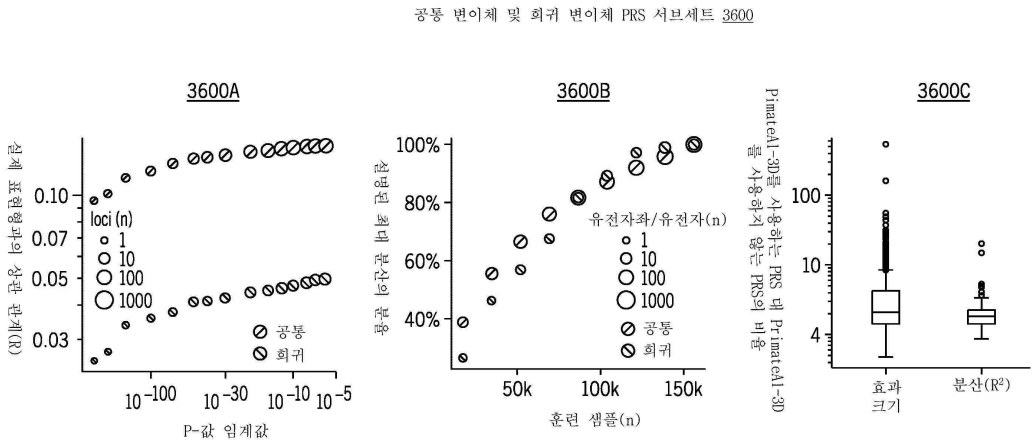
도면34


5%의 오류 발견율에서 유의한 유전자-표현형 쌍으로부터의 임의의 유전자에서 개체당 변이체의 수
3400
개체 수
50000
40000
30000
20000
10000
0
0
2
4
6
8
10
12
14
16
샘플 수

도면35


훈련 데이터 대 검정 데이터 분할의 효과 크기의 비교
3500
n=37
R=0.475
효과 크기(z-점수), 50% 검정
4
2
0
-2
-4
-4
-2
0
4
효과 크기(z-점수), 50% 훈련

도면36


공통 변이체 및 희귀 변이체 PRS 서브세트 3600
3600A
3600B
3600C
표현형과의 상관관계(R²)
0.10
0.07
0.05
0.03
loci (n)
1
10
100
1000
공통
희귀
10⁻¹⁰⁰
10⁻³⁰
10⁻¹⁰
10⁻⁵
P-값 임계값
설명된 분산의 비율
100%
80%
60%
40%
유전자좌/유전자(n)
50k
100k
150k
훈련 샘플(n)
PrimateAI-3D를 사용하는 PRS 대 PrimateAI-3D를 사용하지 않는 PRS의 비율
100
10
4
효과 크기
분산(R²)