(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0103819
(43) 공개일자 2022년07월22일

(51) 국제특허분류(Int. Cl.)
G16B 20/20 (2019.01) *A24F 42/00* (2020.01)
C12Q 1/6876 (2018.01) *G16B 25/10* (2019.01)
G16B 40/00 (2019.01) *G16H 70/60* (2018.01)
(52) CPC특허분류
G16B 20/20 (2019.02)
A24F 42/00 (2022.01)
(21) 출원번호 10-2022-7023834(분할)
(22) 출원일자(국제) 2017년05월30일
심사청구일자 없음
(62) 원출원 특허 10-2019-7009475
원출원일자(국제) 2017년05월30일
심사청구일자 2020년05월19일
(85) 번역문제출일자 2022년07월11일
(86) 국제출원번호 PCT/EP2017/063073
(87) 국제공개번호 WO 2018/050299
국제공개일자 2018년03월22일
(30) 우선권주장
62/394,551 2016년09월14일 미국(US)

(71) 출원인
필립모리스 프로덕츠 에스.에이.
스위스, 씨에이취-2000, 네우차텔, 쿠아이 얀레나 우드 3
(72) 발명자
푸생, 카린
스위스, 씨에이치-2000 네우차텔, 쿠아이 얀레나 우드 3
벨카스트로, 빈센조
스위스, 씨에이치-1400 이베르동-레-뱅, 뤼 뒤 푸르 4
(뒷면에 계속)
(74) 대리인
강철중

전체 청구항 수 : 총 65 항

(54) 발명의 명칭 **개인의 생물학적 상태를 예측하기 위한 시스템, 방법 및 유전자 시그니처**

(57) 요 약

흡연자 상태와 같은, 피험자의 생물학적 상태를 예측하기 위한 피험자의 샘플 평가용 시스템 및 방법. 컴퓨터 실행 방법은, 샘플과 연관된 데이터 세트를 적어도 하나의 하드웨어 프로세서를 포함하는 컴퓨터 시스템에 의해 수신하는 단계를 포함한다. 데이터 세트는, 전체 유전체보다 적은 유전자 세트(AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B 및 TLR5를 포함함)에 대한 정량적 발현 데이터를 포함한다. 적어도 하나의 하드웨어 프로세서는 수신된 데이터 세트 내의 유전자 세트에 대한 정량적 발현 데이터에 기초하여 점수를 생성하는데, 점수는 40 개 미만의 유전자에 기초하고, 피험자의 예측된 흡연 상태를 나타낸다.

(52) CPC특허분류
C12Q 1/6876 (2018.05)
G16B 25/10 (2019.02)
G16B 40/00 (2019.02)
G16H 70/60 (2021.08)
C12Q 2600/158 (2013.01)
(72) 발명자
마틴, 플로리안
스위스, 2000 네우차텔, 쿠아이 얀레나우드 3
부에, 스테파니
스위스, 씨에이치-2068 오뜨히브, 베흐게 끌롯 3
피취, 마누엘, 클로드
스위스, 씨에이치-2034 뻬슈, 체민 가브리엘 5

명 세 서

청구범위

청구항 1

피험자로부터 수득된 샘플을 평가하기 위한 컴퓨터 실행 방법으로서:

상기 샘플과 연관된 데이터 세트를 적어도 하나의 하드웨어 프로세서를 포함하는 컴퓨터 시스템에 의해 수신하되, 상기 데이터 세트는 전체 유전체보다 적은 유전자 세트(AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, 및 TLR5를 포함함)에 대한 정량적 발현 데이터를 포함하는 단계; 및

상기 수신된 데이터 세트 내의 상기 유전자 세트에 대한 상기 정량적 발현 데이터에 기초하여 상기 적어도 하나의 하드웨어 프로세서에 의해 점수를 생성하되, 상기 점수는 40 개 미만의 유전자에 기초하고 상기 피험자의 예측된 흡연 상태를 나타내는 단계를 포함하는, 컴퓨터 실행 방법.

청구항 2

제1항에 있어서, 상기 유전자 세트는 AK8, FSTL1, RGL1 및 VSIG4를 더 포함하는, 컴퓨터 실행 방법.

청구항 3

제1항 내지 제2항 중 어느 한 항에 있어서, 상기 유전자 세트는 C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG 및 PTGFRN을 더 포함하는, 컴퓨터 실행 방법.

청구항 4

제1항 내지 제3항 중 어느 한 항에 있어서, 상기 점수는 상기 데이터 세트에 적용된 분류 체계의 결과이고, 상기 분류 체계는 상기 데이터 세트 내의 상기 정량적 발현 데이터에 기초하여 결정되는, 컴퓨터 실행 방법.

청구항 5

제1항 내지 제4항 중 어느 한 항에 있어서, AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, 및 TLR5 각각에 대한 배수 변화값을 연산하는 단계를 더 포함하는, 컴퓨터 실행 방법.

청구항 6

제5항에 있어서, 각각의 배수 변화값이 적어도 하나의 기준을 충족하는지 결정하는 단계를 더 포함하되, 상기 적어도 하나의 기준은 각각의 연산된 배수 변화값이 적어도 2개의 독립적인 모집단 데이터 세트에 대한 소정의 임계치를 초과할 것을 요구하는 기준인, 컴퓨터 실행 방법.

청구항 7

제1항에 있어서, 상기 유전자 세트는 AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, 및 TLR5로 구성되는, 컴퓨터 실행 방법.

청구항 8

컴퓨터 프로그램 제품으로서, 적어도 하나의 프로세서를 포함하는 컴퓨터화된 시스템에서 실행될 때, 상기 프로세서가 제1항 내지 제7항 중 어느 한 항의 방법의 하나 이상의 단계를 수행하게 하는 컴퓨터 판독 가능 명령어를 포함하는, 컴퓨터 프로그램 제품.

청구항 9

개인의 흡연자 상태 예측용 키트로서:

40 개 미만의 유전자를 갖는 유전자 시그니처(AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, 및 TLR5를 테스트 샘플 내에 포함함) 내에서 상기 유전자의 발현 수준을 검출하는

시약 세트; 및

흡연자 상태 예측용 상기 키트를 상기 개인에서 사용하기 위한 설명서를 포함하는, 키트.

청구항 10

제9항에 있어서, 상기 키트는 흡연 제품의 대안이 개인에 미치는 효과를 평가하기 위해 사용되는, 키트.

청구항 11

제10항에 있어서, 상기 흡연 제품의 대안은 가열식 담배 제품인, 키트.

청구항 12

제9항 내지 제11항 중 어느 한 항에 있어서, 상기 대안이 상기 개인에 미치는 효과는 상기 개인을 비흡연자로서 분류하기 위한 것인, 키트.

청구항 13

제9항 내지 제12항 중 어느 한 항에 있어서, 상기 유전자 시그니처는 AK8, FSTL1, RGL1, 및 VSIG4를 더 포함하는, 키트.

청구항 14

제9항 내지 제13항 중 어느 한 항에 있어서, 상기 유전자 시그니처는 C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG 및 PTGFRN을 더 포함하는, 키트.

청구항 15

피험자로부터 수득된 샘플을 평가하기 위한 컴퓨터 실행 방법으로서:

상기 샘플과 연관된 데이터 세트를 적어도 하나의 하드웨어 프로세서를 포함하는 컴퓨터 시스템에 의해 수신하되, 상기 데이터 세트는 전체 유전체보다 적은 유전자 세트(LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2, 및 GPR63을 포함함)에 대한 정량적 발현 데이터를 포함하는 단계; 및

상기 수신된 데이터 세트 내의 상기 유전자 세트에 대한 상기 정량적 발현 데이터에 기초하여 상기 적어도 하나의 하드웨어 프로세서에 의해 점수를 생성하되, 상기 점수는 40 개 미만의 유전자에 기초하고 상기 피험자의 예측된 흡연 상태를 나타내는 단계를 포함하는, 컴퓨터 실행 방법.

청구항 16

제15항에 있어서, 상기 점수는 상기 데이터 세트에 적용된 분류 체계의 결과이고, 상기 분류 체계는 상기 데이터 세트 내의 상기 정량적 발현 데이터에 기초하여 결정되는, 컴퓨터 실행 방법.

청구항 17

제15항 내지 제16항 중 어느 한 항에 있어서, LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2, 및 GPR63 각각에 대한 배수 변화값을 연산하는 단계를 더 포함하는, 컴퓨터 실행 방법.

청구항 18

제17항에 있어서, 각각의 배수 변화값이 적어도 하나의 기준을 충족시키는지 결정하는 단계를 더 포함하되, 상기 적어도 하나의 기준은 각각의 배수 변화값이 적어도 2 개의 독립적인 모집단 데이터 세트에 대한 소정의 임계치를 초과할 것을 요구하는 기준인, 컴퓨터 실행 방법.

청구항 19

제15항에 있어서, 상기 유전자 세트는 LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2 및 GPR63으로 구성되는, 컴퓨터 실행 방법.

청구항 20

컴퓨터 프로그램 제품으로서, 적어도 하나의 프로세서를 포함하는 컴퓨터화된 시스템에서 실행될 때, 상기 프로세서가 제15항 내지 제19항 중 어느 한 항의 방법의 하나 이상의 단계를 수행하게 하는 컴퓨터 판독 가능 명령어를 포함하는, 컴퓨터 프로그램 제품.

청구항 21

개인의 흡연자 상태 예측용 키트로서:

40 개 미만의 유전자를 갖는 유전자 시그니처(LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2, 및 GPR63을 테스트 샘플 내에 포함함) 내에서 상기 유전자의 발현 수준을 검출하는 시약 세트, 및

흡연자 상태 예측용 상기 키트를 상기 개인에서 사용하기 위한 설명서를 포함하는, 키트.

청구항 22

제21항에 있어서, 상기 키트는 흡연 제품의 대안이 개인에 미치는 효과를 평가하기 위해 사용되는, 키트.

청구항 23

제22항에 있어서, 상기 흡연 제품의 대안은 가열식 담배 제품인, 키트.

청구항 24

제21항 내지 제23항 중 어느 한 항에 있어서, 상기 대안이 상기 개인에 미치는 효과는 상기 개인을 비흡연자로서 분류하기 위한 것인, 키트.

청구항 25

생물학적 상태 예측용 유전자 시그니처를 수득하기 위한 컴퓨터 실행 방법으로서, 상기 방법은:

컴퓨터 시스템에 의해 트레이닝 데이터 세트 및 테스트 데이터 세트를 복수의 사용자 장치에 네트워크를 통해 제공하는 단계로서, 상기 컴퓨터 시스템은 통신 포트 및 적어도 하나의 컴퓨터 프로세서를 포함하고, 상기 적어도 하나의 컴퓨터 프로세서는 트레이닝 데이터 세트 및 테스트 데이터 세트를 포함하는 적어도 하나의 전자 데이터베이스를 저장하는 적어도 하나의 비일시적 컴퓨터 판독 가능 매체와 통신하되,

상기 트레이닝 데이터 세트는 트레이닝 샘플의 세트를 포함하고, 상기 테스트 데이터 세트는 테스트 샘플의 세트를 포함하며, 각 트레이닝 샘플 및 각 테스트 샘플은 유전자의 발현 데이터를 포함하고, 생물학적 상태의 세트로부터 선택된 알려진 생물학적 상태를 갖는 환자에 상응하는, 단계;

상기 트레이닝 데이터 세트에 기초하여 분류기(classifier)를 수득함으로써 각각 생성된 후보 유전자 시그니처를 상기 네트워크로부터 수신하는 단계로서, 각각의 후보 유전자 시그니처는 상기 트레이닝 데이터 세트 내의 상이한 생물학적 상태들 사이에서 판별되도록 결정되는 유전자 세트를 포함하는, 단계;

상기 테스트 샘플의 알려진 생물학적 상태를 예측함에 있어서 상기 각각의 후보 유전자 시그니처의 성과에 기초하여 각 후보 유전자 시그니처 각각에 점수를 할당하는 단계;

상기 할당된 점수에 기초하여 상기 후보 유전자 시그니처의 서브세트를 식별하는 단계;

후보 유전자 시그니처의 적어도 임계 수에 포함된 유전자를 상기 서브세트 내에서 식별하는 단계; 및

상기 식별된 유전자를 상기 유전자 시그니처로서 저장하는 단계를 포함하는, 방법.

청구항 26

제25항에 있어서, 각각의 후보 유전자 시그니처에서 허용된 유전자의 최대 임계 수를 상기 복수의 사용자 장치에 제공하는 단계를 더 포함하는, 방법.

청구항 27

제25항 또는 제26항에 있어서, 네트워크를 통해 상기 테스트 데이터 세트의 일부를 상기 네트워크를 통해 상기 복수의 사용자 장치에 제공하는 단계를 더 포함하되, 상기 테스트 데이터 세트의 상기 일부는 알려진 생물학적 상태를 갖는 환자에 대한 상기 유전자의 발현 데이터를 포함하고, 상기 환자의 상기 알려진 생물학적 상태는 포함하지 않는, 방법.

청구항 28

제27항에 있어서, 각각의 후보 유전자 시그니처에 대해, 상기 테스트 데이터 세트 내의 각각의 샘플에 대한 신뢰 수준을 수신하는 단계를 더 포함하는, 방법.

청구항 29

제28항에 있어서, 상기 신뢰 수준은, 상기 테스트 데이터 세트 내의 샘플이 상기 생물학적 상태 중 하나에 속하는 예측 우도를 나타내는 값인, 방법.

청구항 30

제28항 또는 제29항에 있어서, 상기 점수는 상기 신뢰 수준에 적어도 부분적으로 기초하는, 방법.

청구항 31

제30항에 있어서, 상기 점수는 상기 신뢰 수준 및 상기 테스트 데이터 세트 내의 환자의 알려진 생물학적 상태로부터 연산된 정밀도 재현율 아래 면적(AUPR) 기준에 적어도 부분적으로 기초하는, 방법.

청구항 32

제25항 내지 제31항 중 어느 한 항에 있어서, 상기 점수는 상응 후보 유전자 시그니처가 상기 테스트 데이터 세트 내의 환자의 알려진 생물학적 상태와 일치하는 예측을 제공하는지 여부에 적어도 부분적으로 기초하는, 방법.

청구항 33

제32항에 있어서, 상기 상응 후보 유전자 시그니처가 상기 테스트 데이터 세트 내의 환자의 알려진 생물학적 상태와 일치하는 예측을 제공하는지 여부는 매튜 상관 계수(MCC)를 사용하여 결정되는, 방법.

청구항 34

제25항 내지 제33항 중 어느 한 항에 있어서, 상기 후보 유전자 시그니처는 적어도 2개의 상이한 기준에 따라 순위가 매겨져, 각각의 후보 유전자 시그니처에 대한 제1 순위 및 제2 순위를 획득하는, 방법.

청구항 35

제34항에 있어서, 각각의 후보 유전자 시그니처에 대한 상기 제1 순위 및 상기 제2 순위로 평균을 내어 각각의 후보 유전자 시그니처에 대한 상기 점수를 획득하는, 방법.

청구항 36

제25항 내지 제35항 중 어느 한 항에 있어서, 상기 생물학적 상태 세트는 흡연자 상태를 포함하는, 방법.

청구항 37

제36항에 있어서, 상기 흡연자 상태는 현재 흡연자 및 비흡연자를 포함하는, 방법.

청구항 38

제25항 내지 제37항 중 어느 한 항에 있어서, 상기 유전자 시그니처는 전체 유전체 보다 적으며 AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, 및 TLR5를 포함하는, 방법.

청구항 39

제38항에 있어서, 상기 유전자 시그니처는 AK8, FSTL1, RGL1, 및 VSIG4를 더 포함하는, 방법.

청구항 40

제39항에 있어서, 상기 유전자 시그니처는 C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG, 및 PTGFRN을 더 포함하는, 방법.

청구항 41

제40항에 있어서, 상기 유전자 시그니처는 ASGR2, B3GALT2, CYP4F22, FUCA1, GPR63, GUCY1B3, MB21D2, NLK, NR4A1, P2RY1, PF4, PTGFR, SH2D1B, ST6GALNAC1, TMEM163, TPPP3, 및 ZNF618을 더 포함하는, 방법.

청구항 42

제25항 내지 제37항 중 어느 한 항에 있어서, 상기 유전자 시그니처는 전체 유전체 보다 적으며 LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2, 및 GPR63을 포함하는, 방법.

청구항 43

제42항에 있어서, 상기 유전자 시그니처는 DSC2, TLR5, RGL1, FSTL1, VSIG4, AK8, GUCY1A3, GSE1, MIR4697HG, PTGFRN, LOC200772, FANK1, C15orf54, MARC2, TPPP3, ZNF618, PTGFR, P2RY1, TMEM163, ST6GALNAC1, SH2D1B, CYP4F22, PF4, FUCA1, MB21D2, NLK, B3GALT2, ASGR2, NR4A1 및 GUCY1B3을 더 포함하는, 방법.

청구항 44

제25항 내지 제37항 중 어느 한 항에 있어서, 상기 유전자 시그니처는 전체 유전체보다 적으며 AHHR, P2RY6, KLRG1, LRRN3, COX6B2, CTTNBP2, DSC2, F2R, GUCY1B3, MT2, NGFRAP1, REEP6, SASH1, 및 TBX21을 포함하는, 방법.

청구항 45

컴퓨터 프로그램 제품으로서, 적어도 하나의 프로세서를 포함하는 컴퓨터화된 시스템에서 실행될 때, 상기 프로세서가 제25항 내지 제44항 중 어느 한 항의 방법의 하나 이상의 단계를 수행하게 하는 컴퓨터 판독 가능 명령어를 포함하는, 컴퓨터 프로그램 제품.

청구항 46

피험자로부터 수득된 샘플을 평가하기 위한 컴퓨터 실행 방법으로서:

상기 샘플과 연관된 데이터 세트를 적어도 하나의 하드웨어 프로세서를 포함하는 컴퓨터 시스템에 의해 수신하는 단계로서, 상기 데이터 세트는 전체 유전체보다 적은 유전자 세트(AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, TLR5, AK8, FSTL1, RGL1, VSIG4, C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG, PTGFRN, ASGR2, B3GALT2, CYP4F22, FUCA1, GPR63, GUCY1B3, MB21D2, NLK, NR4A1, P2RY1, PF4, PTGFR, SH2D1B, ST6GALNAC1, TMEM163, TPPP3, 및 ZNF618을 포함함)에 대한 정량적 발현 데이터를 포함하는 것인 단계; 및

상기 수신된 데이터 세트에 기초하여 상기 적어도 하나의 하드웨어 프로세서에 의해 점수를 생성하는 단계로서, 상기 점수는 상기 피험자의 예측된 흡연 상태를 나타내는 것인 단계를 포함하는, 방법.

청구항 47

제46항에 있어서, 상기 점수는 상기 데이터 세트에 적용된 분류 체계의 결과이고, 상기 분류 체계는 상기 데이터 세트 내의 상기 정량적 발현 데이터에 기초하여 결정되는, 컴퓨터 실행 방법.

청구항 48

제46항 내지 제47항 중 어느 한 항에 있어서, AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, TLR5, AK8, FSTL1, RGL1, VSIG4, C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG, PTGFRN, ASGR2, B3GALT2, CYP4F22, FUCA1, GPR63, GUCY1B3, MB21D2, NLK, NR4A1, P2RY1, PF4, PTGFR, SH2D1B, ST6GALNAC1, TMEM163, TPPP3, 및 ZNF618 각각에 대한 배수 변화값을 연산

하는 단계를 더 포함하는, 컴퓨터 실행 방법.

청구항 49

제48항에 있어서, 각각의 배수 변화값이 적어도 하나의 기준을 충족하는지 결정하는 단계를 더 포함하되, 상기 적어도 하나의 기준은 각각의 연산된 배수 변화값이 적어도 2개의 독립 모집단 데이터 세트에 대한 소정의 임계치를 초과하는 것을 요구하는 기준인, 컴퓨터 실행 방법.

청구항 50

제46항 내지 제49항 중 어느 한 항에 있어서, 상기 유전자 세트는 AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, TLR5, AK8, FSTL1, RGL1, VSIG4, C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG, PTGFRN, ASGR2, B3GALT2, CYP4F22, FUCA1, GPR63, GUCY1B3, MB21D2, NLK, NR4A1, P2RY1, PF4, PTGFR, SH2D1B, ST6GALNAC1, TMEM163, TPPP3, and ZNF618로 구성되는, 컴퓨터 실행 방법.

청구항 51

컴퓨터 프로그램 제품으로서, 적어도 하나의 프로세서를 포함하는 컴퓨터화된 시스템에서 실행될 때, 상기 프로세서가 제46항 내지 제50항 중 어느 한 항의 방법의 하나 이상의 단계를 수행하게 하는 컴퓨터 판독 가능 명령어를 포함하는, 컴퓨터 프로그램 제품.

청구항 52

개인의 흡연자 상태 예측용 키트로서:

테스트 샘플 내 유전자 시그니처(AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, TLR5, AK8, FSTL1, RGL1, VSIG4, C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG, PTGFRN, ASGR2, B3GALT2, CYP4F22, FUCA1, GPR63, GUCY1B3, MB21D2, NLK, NR4A1, P2RY1, PF4, PTGFR, SH2D1B, ST6GALNAC1, TMEM163, TPPP3, 및 ZNF618을 포함함) 내의 상기 유전자의 발현 수준을 검출하는 시약 세트; 및

흡연자 상태 예측용 상기 키트를 상기 개인에서 사용하기 위한 설명서를 포함하는, 키트.

청구항 53

제52항에 있어서, 상기 키트는 흡연 제품의 대안이 개인에 미치는 효과를 평가하기 위해 사용되는, 키트.

청구항 54

제53항에 있어서, 흡연 제품의 대안은 가열식 담배 제품인, 키트.

청구항 55

제52항 내지 제54항 중 어느 한 항에 있어서, 상기 대안이 상기 개인에 미치는 효과는 상기 개인을 비흡연자로서 분류하기 위한 것인, 키트.

청구항 56

피험자로부터 수득된 샘플을 평가하기 위한 컴퓨터 실행 방법으로서:

상기 샘플과 연관된 데이터 세트를 적어도 하나의 하드웨어 프로세서를 포함하는 컴퓨터 시스템에 의해 수신하는 단계로서, 상기 데이터 세트는 전체 유전체보다 적은 유전자 세트(AHHR, P2RY6, KLRG1, LRRN3, COX6B2, CTTNBP2, DSC2, F2R, GUCY1B3, MT2, NGFRAP1, REEP6, SASH1, 및 TBX21을 포함함)에 대한 정량적 발현 데이터를 포함하는 것인 단계; 및

상기 수신된 데이터 세트 내의 상기 유전자 세트에 대한 상기 정량적 발현 데이터에 기초하여 상기 적어도 하나의 하드웨어 프로세서에 의해 점수를 생성하되, 상기 점수는 40 개 미만의 유전자에 기초하고, 상기 피험자의 예측된 흡연 상태를 나타내는 단계를 포함하는, 컴퓨터 실행 방법.

청구항 57

제56항에 있어서, 상기 점수는 상기 데이터 세트에 적용된 분류 체계의 결과이고, 상기 분류 체계는 상기 데이터 세트 내의 상기 정량적 발현 데이터에 기초하여 결정되는, 컴퓨터 실행 방법.

청구항 58

제56항 내지 제57항 중 어느 한 항에 있어서, AHHR, P2RY6, KLRG1, LRRN3, COX6B2, CTTNBP2, DSC2, F2R, GUCY1B3, MT2, NGFRAP1, REEP6, SASH1, 및 TBX21 각각에 대한 배수 변화값을 연산하는 단계를 더 포함하는, 컴퓨터 실행 방법.

청구항 59

제58항에 있어서, 각각의 배수 변화값이 적어도 하나의 기준을 충족시키는지 결정하는 단계를 더 포함하되, 상기 적어도 하나의 기준은 각각의 연산된 배수 변화값이 적어도 2개의 독립 모집단 데이터 세트에 대한 소정의 임계치를 초과할 것을 요구하는 기준인, 컴퓨터 실행 방법.

청구항 60

제56항에 있어서, 상기 유전자 세트는 AHHR, P2RY6, KLRG1, LRRN3, COX6B2, CTTNBP2, DSC2, F2R, GUCY1B3, MT2, NGFRAP1, REEP6, SASH1, 및 TBX21로 구성되는, 컴퓨터 실행 방법.

청구항 61

컴퓨터 프로그램 제품으로서, 적어도 하나의 프로세서를 포함하는 컴퓨터화된 시스템에서 실행될 때, 상기 프로세서가 제56항 내지 제60항 중 어느 한 항의 방법의 하나 이상의 단계를 수행하게 하는 컴퓨터 판독 가능 명령어를 포함하는, 컴퓨터 프로그램 제품.

청구항 62

개인의 흡연자 상태 예측용 키트로서:

테스트 샘플 내 유전자 시그니처(AHHR, P2RY6, KLRG1, LRRN3, COX6B2, CTTNBP2, DSC2, F2R, GUCY1B3, MT2, NGFRAP1, REEP6, SASH1, 및 TBX21을 포함하고 40 개 미만의 유전자를 포함함)에서 유전자의 발현 수준을 검출하는 시약 세트; 및

흡연자 상태 예측용 상기 키트를 상기 개인에서 사용하기 위한 설명서를 포함하는, 키트.

청구항 63

제62항에 있어서, 상기 키트는 흡연 제품의 대안이 개인에 미치는 효과를 평가하기 위해 사용되는, 키트.

청구항 64

제63항에 있어서, 상기 흡연 제품의 대안은 가열식 담배 제품인, 키트.

청구항 65

제63항 내지 제64항 중 어느 한 항에 있어서, 상기 대안이 상기 개인에 미치는 효과는 상기 개인을 비흡연자로서 분류하기 위한 것인, 키트.

발명의 설명

기 술 분 야

[0001] 관련 출원에 대한 참조

[0002] 본 출원은 35 U.S.C § 119 하에, 2016년 9월 14일자로 출원된 미국 가출원 제62/394,551호에 대한 우선권을 주장하며, 그 전체는 본원에 참조로서 통합된다. 본 출원은 2014년 12월 11일자로 출원된 PCT 출원 제

PCT/EP2014/077473호 및 2014년 8월 12일자로 출원된 PCT 출원 제PCT/EP2014/067276호에 관한 것이며, 이들 각각은 그 전체가 본원에 참조로서 통합된다.

배 경 기 술

[0003] 인간은 유해한 분자 변화를 유발할 수 있는 외부 독성 물질(예, 담배 연기, 살충제)에 끊임없이 노출된다. 21세기 독성학의 맥락에서의 위험 평가는 독성 메커니즘의 설명 및 고 처리량 데이터로부터의 노출 반응 마커의 식별에 의존한다. 전체 유전체 마이크로 어레이(whole genome microarray)와 같은 신기술이 독성 테스트에 통합되어 효율성을 높이고, 노출 반응 평가에 보다 데이터 중심의 접근법을 제공하였다. 전사 유전자 조절에 대한 유전체 규모의 추정은 마이크로 어레이 및 RNA 시퀀싱과 같은 고 처리량 기술의 출현으로 가능해졌는데, 이는 이러한 기술이 테스트된 많은 실험 조건하에서 전사체의 스냅샷을 제공하기 때문이다.

[0004] 바이오메디컬 연구 커뮤니티는 질병 진단을 위한 확고한 시그니처를 찾는데 일반적으로 관심이 있다. 질병의 분자 분류가 형태학적 분류보다 더 정확할 수 있다는 일부 증거가 있다. 그러나, 노출의 주된 부위(예: 연기 또는 공기 오염물질에 노출되는 경우의 기도)로부터 샘플을 획득하는 것은 일반적으로 침습적이므로, 노출 평가 및 모니터링이 편리하지 않다. 최소 침습적인 대안으로서, 말초 혈액 샘플링을 일반 개체군에서 사용하여 전신 바이오마커를 수립할 수 있다. 혈액은 많은 상이한 세포 아개체군(sub-population)을 함유하고 있기 때문에 분석하기에 복잡하다. 그러나, 혈액은 독성 물질에 보다 직접적으로 노출되는 모든 기관 내에서 순환하고 쉽게 접근할 수 있기 때문에 마커 식별을 조사하는 데 관련성이 높은 조직이다. 게다가, 조직학적 이상이 보이지 않더라도 연기 노출에 대한 분자 반응이 검출될 수 있다.

발명의 내용

[0005] 크라우드 소싱(crowd-sourcing) 방법을 사용하여, 개인의 흡연자 상태를 예측하는데 사용될 수 있는 확고한 혈액 기반 유전자 시그니처를 확인하는 연산 시스템 및 방법이 제공된다. 본원에 기술된 유전자 시그니처는 흡연 비경험자로부터 현재 흡연하는 피험자를 구별하는 능력에 의해 개인의 흡연자 상태를 정확하게 예측할 수 있다.

[0006] 크라우드 소싱(crowd-sourcing) 방법을 사용하여, 개인의 흡연자 상태를 예측하는데 사용될 수 있는 확고한 혈액 기반 유전자 시그니처를 확인하는 연산 시스템 및 방법이 제공된다. 본원에 기술된 유전자 시그니처는 흡연 비경험자로부터 현재 흡연하는 피험자를 구별하는 능력에 의해 개인의 흡연자 상태를 정확하게 예측할 수 있다.

[0007] 특정 양태에서, 본 개시의 시스템 및 방법은 피험자로부터 수득한 샘플을 평가하기 위한 컴퓨터 실행 방법을 제공한다. 컴퓨터 실행 방법은, 샘플과 연관된 데이터 세트를 적어도 하나의 하드웨어 프로세서를 포함하는 컴퓨터 시스템에 의해 수신하는 단계를 포함한다. 데이터 세트는, 전체 유전체보다 적은 유전자 세트(AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B 및 TLR5를 포함함)에 대한 정량적 발현 데이터를 포함한다. 적어도 하나의 하드웨어 프로세서는 수신된 데이터 세트 내의 유전자 세트에 대한 정량적 발현 데이터에 기초하여 점수를 생성하는데, 점수는 40 개 미만의 유전자에 기초하고, 피험자의 예측된 흡연 상태를 나타낸다.

[0008] 특정 구현예에서, 유전자 세트는 AK8, FSTL1, RGL1, 및 VSIG4를 더 포함한다. 특정 구현예에서, 유전자 세트는 C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG, 및 PTGFRN을 더 포함한다.

[0009] 특정 구현예에서, 점수는 데이터 세트에 적용된 분류 체계의 결과이고, 분류 체계는 데이터 세트 내의 정량적 발현 데이터에 기초하여 결정된다. 특정 구현예에서, 컴퓨터 실행 방법은 AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, 및 TLR5 각각에 대한 배수 변화값을 연산하는 단계를 더 포함한다. 컴퓨터 실행 방법은 각각의 연산된 배수 변화값이 적어도 2개의 독립적인 모집단 데이터 세트에 대한 소정의 임계치를 초과하는 것을 요구하는 적어도 하나의 기준을 각각의 배수 변화값이 충족하는지 결정하는 단계를 더 포함할 수 있다.

[0010] 특정 구현예에서, 유전자 세트는 AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, 및 TLR5로 구성된다.

[0011] 특정 양태에서, 본 개시의 시스템 및 방법은 개인의 흡연자 상태 예측용 키트를 제공한다. 키트는 40 개 미만의 유전자를 갖는 유전자 시그니처(AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, 및 TLR5를 테스트 샘플 내에 포함함) 내에서 유전자의 발현 수준을 검출하는 시약 세트, 및 흡연자 상태 예측용 상기 키트를 개인에서 사용하기 위한 설명서를 포함한다.

[0012] 특정 구현예에서, 키트는 흡연 제품의 대안이 개인에 미치는 효과를 평가하기 위해 사용된다. 흡연 제품의 대안은 가열식 담배 제품을 포함할 수 있다. 대안이 개인에 미치는 효과는 개인을 비흡연자로서 분류하는 것일 수 있다. 특정 구현예에서, 유전자 시그니처는 AK8, FSTL1, RGL1, 및 VSIG4를 더 포함한다. 특정 구현예에서, 유전자 시그니처는 C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG, 및 PTGFRN을 더 포함한다.

[0013] 특정 양태에서, 본 개시의 시스템 및 방법은 피험자로부터 수득한 샘플을 평가하기 위한 컴퓨터 실행 방법을 제공한다. 컴퓨터 실행 방법은, 샘플과 연관된 데이터 세트를 적어도 하나의 하드웨어 프로세서를 포함하는 컴퓨터 시스템에 의해 수신하는 단계를 포함하고, 데이터 세트는 전체 유전체보다 적은 유전자 세트(LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2, 및 GPR63)에 대한 정량적 발현 데이터를 포함한다. 적어도 하나의 하드웨어 프로세서는 수신된 데이터 세트 내의 유전자 세트에 대한 정량적 발현 데이터에 기초하여 점수를 생성하는데, 점수는 40 개 미만의 유전자에 기초하고, 피험자의 예측된 흡연 상태를 나타낸다.

[0014] 특정 구현예에서, 점수는 데이터 세트에 적용된 분류 체계의 결과이고, 분류 체계는 데이터 세트 내의 정량적 발현 데이터에 기초하여 결정된다.

[0015] 특정 구현예에서, 적어도 하나의 하드웨어 프로세서는 LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2, 및 GPR63 각각에 대한 배수 변화값을 연산한다. 컴퓨터 실행 방법은 각각의 연산된 배수 변화값이 적어도 2개의 독립적인 모집단 데이터 세트에 대한 소정의 임계치를 초과하는 것을 요구하는 적어도 하나의 기준을 각각의 배수 변화값이 충족하는지 결정하는 단계를 더 포함할 수 있다.

[0016] 특정 구현예에서, 유전자 세트는 LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2, 및 GPR63로 구성된다.

[0017] 특정 양태에서, 본 개시의 시스템 및 방법은 개인의 흡연자 상태 예측용 키트를 제공한다. 키트는 테스트 샘플 내의 유전자 시그니처(40개 미만의 유전자를 갖고, LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2, 및 GPR63을 포함함)에서 유전자의 발현 수준을 검출하는 시약 세트, 및 흡연자 상태 예측용 상기 키트를 개인에서 사용하기 위한 설명서를 포함한다.

[0018] 특정 구현예에서, 키트는 흡연 제품의 대안이 개인에 미치는 효과를 평가하기 위해 사용된다. 흡연 제품의 대안은 가열식 담배 제품을 포함할 수 있다. 대안이 개인에 미치는 효과는 개인을 비흡연자로서 분류하는 것일 수 있다.

[0019] 특정 양태에서, 본 개시의 시스템 및 방법은 생물학적 상태 예측용 유전자 시그니처를 수득하기 위한 컴퓨터 실행 방법을 제공한다. 컴퓨터 실행 방법은 네트워크를 통해 테스트 데이터 세트를 복수의 사용자 장치에 제공하는 단계를 포함하되, 컴퓨터 시스템은 통신 포트 및 적어도 하나의 컴퓨터 프로세서를 포함하고, 상기 적어도 하나의 컴퓨터 프로세서는 트레이닝 데이터 세트 및 테스트 데이터 세트를 포함하는 적어도 하나의 전자 데이터베이스를 저장하는 적어도 하나의 비일시적 컴퓨터 판독 가능 매체와 통신한다. 트레이닝 데이터 세트는 한 세트의 트레이닝 샘플을 포함하고, 테스트 데이터 세트는 한 세트의 테스트 샘플을 포함한다. 각각의 트레이닝 샘플 및 각각의 테스트 샘플은 유전자 발현 데이터를 포함하고, 한 세트의 생물학적 상태로부터 선택된 알려진 생물학적 상태를 갖는 환자에 상응한다. 컴퓨터 실행 방법은 트레이닝 데이터 세트에 기초하여 분류기(classifier)를 수득함으로써 각각 생성된 후보 유전자 시그니처를 네트워크로부터 수신하는 단계를 더 포함하되, 각각의 후보 유전자 시그니처는 트레이닝 데이터 세트 내의 상이한 생물학적 상태들 사이에서 판별되도록 결정되는 한 세트의 유전자를 포함한다. 점수는 테스트 샘플의 알려진 생물학적 상태를 예측할 때 각각의 후보 유전자 시그니처의 성과에 기초하여 각각의 후보 유전자 시그니처에 할당된다. 후보 유전자 시그니처의 서브세트(또는 후보 유전자 시그니처의 전체 세트를 포함할 수 있는 후보 유전자 시그니처의 일부)는 할당된 점수에 기초하여 식별되고, 후보 유전자 시그니처의 적어도 임계 수에 포함된 유전자가 서브세트 내에서 식별된다. 식별된 유전자는 유전자 시그니처로서 저장된다.

[0020] 특정 구현예에서, 컴퓨터 실행 방법은, 각각의 후보 유전자 시그니처에서 허용된 유전자의 최대 임계 수를 대표하는 수를 복수의 사용자 장치에 제공하는 단계를 더 포함한다.

[0021] 특정 구현예에서, 컴퓨터 실행 방법은, 네트워크를 통해 테스트 데이터 세트의 일부를 복수의 사용자 장치에 제공하는 단계를 더 포함하되, 테스트 데이터 세트의 일부는 알려진 생물학적 상태를 갖는 환자에 대한 유전자의 발현 데이터를 포함하고, 환자의 알려진 생물학적 상태는 포함하지 않는다. 컴퓨터 실행 방법은 각각의 후보 유

전자 시그니처에 대해, 테스트 데이터 세트 내의 각각의 샘플에 대한 신뢰 수준을 수신하는 단계를 더 포함할 수 있다. 신뢰 수준은, 테스트 데이터 세트 내의 샘플이 생물학적 상태 중 하나에 속하는 예측 우도를 나타내는 값일 수 있다. 점수는 신뢰 수준에 적어도 부분적으로 기초할 수 있다. 특히, 점수는 신뢰 수준 및 테스트 데이터 세트 내의 환자의 알려진 생물학적 상태로부터 연산된 정밀도 재현율 아래 면적(AUPR) 기준에 적어도 부분적으로 기초할 수 있다.

[0022] 특정 구현예에서, 점수는 상응하는 후보 유전자 시그니처가 테스트 데이터 세트 내의 환자의 알려진 생물학적 상태와 일치하는 예측을 제공하는지 여부에 적어도 부분적으로 기초한다. 상응하는 후보 유전자 시그니처가 테스트 데이터 세트 내의 환자의 알려진 생물학적 상태와 일치하는 예측을 제공하는지 여부는 매튜 상관 계수(MCC)를 사용하여 결정될 수 있다.

[0023] 특정 구현예에서, 후보 유전자 시그니처는 적어도 2개의 상이한 기준에 따라 순위가 매겨져, 각각의 후보 유전자 시그니처에 대한 제1 순위 및 제2 순위를 획득한다. 각각의 후보 유전자 시그니처에 대한 제1 순위 및 제2 순위로 평균을 내어 각각의 후보 유전자 시그니처에 대한 점수를 획득할 수 있다.

[0024] 특정 구현예에서, 생물학적 상태의 세트는 흡연자 상태를 포함한다. 흡연자 상태에는 현재 흡연자와 비흡연자가 포함될 수 있다.

[0025] 특정 구현예에서, 유전자 시그니처는 전체 유전체 보다 적으며 AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, 및 TLR5를 포함한다. 또한, 유전자 시그니처는 AK8, FSTL1, RGL1, 및 VSIG4를 더 포함할 수 있다. 또한, 유전자 시그니처는 C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG, 및 PTGFRN을 더 포함할 수 있다. 또한, 유전자 시그니처는 ASGR2, B3GALT2, CYP4F22, FUCA1, GPR63, GUCY1B3, MB21D2, NLK, NR4A1, P2RY1, PF4, PTGFR, SH2D1B, ST6GALNAC1, TMEM163, TPPP3, 및 ZNF618을 더 포함할 수 있다. 일부 구현예에서, 유전자 시그니처는 임계 수의 유전자, 예컨대 10, 15, 20, 25, 30, 35, 40 개, 또는 전체 유전체 내의 유전자 수보다 적은 임의의 다른 적절한 수의 유전자로 제한될 수 있다.

[0026] 특정 구현예에서, 유전자 시그니처는 전체 유전체보다 적으며 LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2 및 GPR63을 포함한다. 또한, 유전자 시그니처는 DSC2, TLR5, RGL1, FSTL1, VSIG4, AK8, GUCY1A3, GSE1, MIR4697HG, PTGFRN, LOC200772, FANK1, C15orf54, MARC2, TPPP3, ZNF618, PTGFR, P2RY1, TMEM163, ST6GALNAC1, SH2D1B, CYP4F22, PF4, FUCA1, MB21D2, NLK, B3GALT2, ASGR2, NR4A1, 및 GUCY1B3를 더 포함할 수 있다. 일부 구현예에서, 유전자 시그니처는 임계 수의 유전자, 예컨대 10, 15, 20, 25, 30, 35, 40 개, 또는 전체 유전체 내의 유전자 수보다 적은 임의의 다른 적절한 수의 유전자로 제한될 수 있다.

[0027] 특정 구현예에서, 유전자 시그니처는 전체 유전체보다 적으며 AHHR, P2RY6, KLRG1, LRRN3, COX6B2, CTTNBP2, DSC2, F2R, GUCY1B3, MT2, NGFRAP1, REEP6, SASH1, 및 TBX21을 포함한다. 일부 구현예에서, 유전자 시그니처는 임계 수의 유전자, 예컨대 10, 15, 20, 25, 30, 35, 40 개, 또는 전체 유전체 내의 유전자 수보다 적은 임의의 다른 적절한 수의 유전자로 제한될 수 있다.

[0028] 특정 양태에서, 본 개시의 시스템 및 방법은 피험자로부터 수득한 샘플을 평가하기 위한 컴퓨터 실행 방법을 제공한다. 컴퓨터 실행 방법은, 샘플과 연관된 데이터 세트를 적어도 하나의 하드웨어 프로세서를 포함하는 컴퓨터 시스템에 의해 수신하는 단계를 포함한다. 테이터 세트는, 전체 유전체보다 적으며 AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, TLR5, AK8, FSTL1, RGL1, VSIG4, C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG, PTGFRN, ASGR2, B3GALT2, CYP4F22, FUCA1, GPR63, GUCY1B3, MB21D2, NLK, NR4A1, P2RY1, PF4, PTGFR, SH2D1B, ST6GALNAC1, TMEM163, TPPP3, 및 ZNF618을 포함하는 한 세트의 유전자에 대한 정량적 발현 데이터를 포함한다. 적어도 하나의 하드웨어 프로세서는 수신된 데이터 세트에 기초하여 점수를 생성하는데, 점수는 피험자의 예측된 흡연 상태를 나타낸다.

[0029] 특정 구현예에서, 점수는 데이터 세트에 적용된 분류 체계의 결과이고, 분류 체계는 데이터 세트 내의 정량적 발현 데이터에 기초하여 결정된다.

[0030] 특정 구현예에서, 컴퓨터 실행 방법은 AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, TLR5, AK8, FSTL1, RGL1, VSIG4, C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG, PTGFRN, ASGR2, B3GALT2, CYP4F22, FUCA1, GPR63, GUCY1B3, MB21D2, NLK, NR4A1, P2RY1, PF4, PTGFR, SH2D1B, ST6GALNAC1, TMEM163, TPPP3, 및 ZNF618 각각에 대한 배수 변화값을 연산하는 단계를 더

포함한다. 컴퓨터 실행 방법은 각각의 연산된 배수 변화값이 적어도 2개의 독립적인 모집단 데이터 세트에 대한 소정의 임계치를 초과하는 것을 요구하는 적어도 하나의 기준을 각각의 배수 변화값이 충족하는지 결정하는 단계를 더 포함할 수 있다.

[0031] 특정 구현예에서, 유전자 세트는 AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, TLR5, AK8, FSTL1, RGL1, VSIG4, C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG, PTGFRN, ASGR2, B3GALT2, CYP4F22, FUCA1, GPR63, GUCY1B3, MB21D2, NLK, NR4A1, P2RY1, PF4, PTGFR, SH2D1B, ST6GALNAC1, TMEM163, TPPP3, 및 ZNF618로 구성된다.

[0032] 특정 양태에서, 본 개시의 시스템 및 방법은 개인의 흡연자 상태 예측용 키트를 제공한다. 키트는 유전자 시그니처(AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, TLR5, AK8, FSTL1, RGL1, VSIG4, C15orf54, CTTNBP2, RANK1, GSE1, GUCY1A3, LOC200772, MARC2, MIR4697HG, PTGFRN, ASGR2, B3GALT2, CYP4F22, FUCA1, GPR63, GUCY1B3, MB21D2, NLK, NR4A1, P2RY1, PF4, PTGFR, SH2D1B, ST6GALNAC1, TMEM163, TPPP3, 및 ZNF618을 테스트 샘플 내에 포함함) 내에서 유전자의 발현 수준을 검출하는 시약 세트, 및 흡연자 상태 예측용 키트를 개인에서 사용하기 위한 설명서를 포함한다.

[0033] 특정 구현예에서, 키트는 흡연 제품의 대안이 개인에 미치는 효과를 평가하기 위해 사용된다. 흡연 제품의 대안은 가열식 담배 제품을 포함할 수 있다. 대안이 개인에 미치는 효과는 개인을 비흡연자로서 분류하는 것일 수 있다.

[0034] 특정 양태에서, 본 개시의 시스템 및 방법은 피험자로부터 수득한 샘플을 평가하기 위한 컴퓨터 실행 방법을 제공한다. 컴퓨터 실행 방법은, 샘플과 연관된 데이터 세트를 적어도 하나의 하드웨어 프로세서를 포함하는 컴퓨터 시스템에 의해 수신하는 단계를 포함하고, 데이터 세트는 AHHR, P2RY6, KLRG1, LRRN3, COX6B2, CTTNBP2, DSC2, F2R, GUCY1B3, MT2, NGFRAP1, REEP6, SASH1, 및 TBX21를 포함하는, 전체 유전체보다 적은 유전자 세트에 대한 정량적 발현 데이터를 포함한다. 적어도 하나의 하드웨어 프로세서는 수신된 데이터 세트 내의 유전자 세트에 대한 정량적 발현 데이터에 기초하여 점수를 생성하는데, 점수는 40 개 미만의 유전자에 기초하고, 피험자의 예측된 흡연 상태를 나타낸다.

[0035] 특정 구현예에서, 점수는 데이터 세트에 적용된 분류 체계의 결과이고, 분류 체계는 데이터 세트 내의 정량적 발현 데이터에 기초하여 결정된다.

[0036] 특정 구현예에서, 컴퓨터 실행 방법은 AHHR, P2RY6, KLRG1, LRRN3, COX6B2, CTTNBP2, DSC2, F2R, GUCY1B3, MT2, NGFRAP1, REEP6, SASH1, 및 TBX21 각각에 대한 배수 변화값을 연산하는 단계를 더 포함한다. 컴퓨터 실행 방법은 각각의 연산된 배수 변화값이 적어도 2개의 독립적인 모집단 데이터 세트에 대한 소정의 임계치를 초과하는 것을 요구하는 적어도 하나의 기준을 각각의 배수 변화값이 충족하는지 결정하는 단계를 더 포함할 수 있다.

[0037] 특정 구현예에서, 유전자 세트는 AHHR, P2RY6, KLRG1, LRRN3, COX6B2, CTTNBP2, DSC2, F2R, GUCY1B3, MT2, NGFRAP1, REEP6, SASH1, 및 TBX21로 구성된다.

[0038] 특정 양태에서, 본 개시의 시스템 및 방법은 개인의 흡연자 상태 예측용 키트를 제공한다. 키트는 유전자 시그니처(AHHR, P2RY6, KLRG1, LRRN3, COX6B2, CTTNBP2, DSC2, F2R, GUCY1B3, MT2, NGFRAP1, REEP6, SASH1, 및 TBX21을 테스트 샘플 내에 포함하고, 40 개 미만의 유전자를 포함함) 내에서 유전자의 발현 수준을 검출하는 시약 세트, 흡연자 상태 예측용 키트를 개인에서 사용하기 위한 설명서를 포함한다.

[0039] 특정 구현예에서, 키트는 흡연 제품의 대안이 개인에 미치는 효과를 평가하기 위해 사용된다. 흡연 제품의 대안은 가열식 담배 제품을 포함할 수 있다. 대안이 개인에 미치는 효과는 개인을 비흡연자로서 분류하는 것일 수 있다.

도면의 간단한 설명

[0040] 본 개시의 추가 특징, 본질 및 다양한 장점은 첨부된 도면과 함께 다음의 상세한 설명을 고려하면 명백해질 것이고,

명세서 전체에 걸쳐 도면의 동일한 참조 부호는 동일한 부분을 나타내며,

도면 중:

도 1은 크라우드 소싱을 사용하여 유전자 시그니처의 식별을 수행하기 위해 컴퓨터화된 시스템의 블록다이어그

램이고;

도 2는 본원에 설명된 컴퓨터화된 시스템 중 임의의 구성 요소를 구현하는데 사용될 수 있는 예시적인 컴퓨팅 장치의 블록다이어그램이고;

도 3은 개인의 생물학적 상태를 예측하기 위해 크라우드 소싱을 사용하여 유전자 시그니처를 식별하는 프로세스의 순서도이고;

도 4a 및 4b는 인간 데이터(도 4a) 및 종-독립 데이터(도 4b)에 대해 상이한 팀에 걸친 동시 발생을 나타내는 표이고;

도 5는 피험자의 예측된 흡연 상태를 나타내는 점수의 평가 방법에 대한 흐름도이고;

도 6은 상이한 연구에 대한 샘플 그룹/분류, 크기 및 특성을 요약한 표이고;

도 7a는 인간 및 마우스의 전혈 유전자 발현 데이터로부터 화학 노출 반응 마커를 식별하고, 새로운 혈액 샘플을 노출 그룹 또는 비노출 그룹의 부분으로서 예측 분류하기 위한 연산 모델에서 이러한 마커를 시그니처로서 차용하는 것을 보여주는 다이어그램이고;

도 7b는 (i) 흡연자와 비흡연자를 구별(과업 1)하고, 이어서 (ii) 현재 비흡연자를 이전 흡연자 및 흡연 비경험자로 분류(과업 2)하기 위해 확고하고 희소한 인간(하위 도전 1, SC1) 및 종-독립적(하위 도전 2, SC2) 혈액 기반 유전자 시그니처 분류 모델을 개발하는 것을 도시하는 다이어그램이고;

도 8은 트레이닝 데이터 세트, 테스트 데이터 세트, 및 혈액 유전자의 발현 데이터 중 검증 데이터 세트를 방출하는 것을 나타내는 다이어그램이고;

도 9a는 흡연자와 비흡연자 간의 명확한 분리를 보여주는 상자도이고;

도 9b는 흡연 그룹에 대해 0 일차와 5 일차 세션 간에 유의한 차이가 없지만, 세스(Cess)와 스위치(Switch) 그룹의 경우 0 일차에 각각의 베이스라인과 비교해 유의한 감소를 보여주는 2 개의 상자도를 포함하고;

도 10은 클래스 예측을 위한 유전자 시그니처 분류 모델의 클래스 예측 성능을 보여주는 2 개의 표를 포함하고;

도 11a 및 11b는 테스트 및 검증 데이터 세트에 대한 참가자별 혈액 샘플 클래스 예측을 보여주는 상자도이고;

도 12는 검증 데이터 세트에 대해 구금 상태에서의 0 일차와 5 일차 간의 크라우드 로그 오즈비(crowd log odds ratio)를 보여주는 상자도를 포함하고;

도 13은 그룹/클래스 당 크라우드 로그 오즈 분포 스플릿 및 pMRTP 또는 후보 MRTP에 대한 노출 시간, 또는 pMRTP 또는 후보 MRTP로 전환한 후의 노출 시간을 보여주는 상자도이며;

도 14 및 15는 ML 기반의 클래스 예측을 사용하여 길이가 2 내지 18인 시그니처의 모든 가능한 조합의 성능을 평가하기 위한 MCC 및 AUPR 점수의 플롯이다.

발명을 실시하기 위한 구체적인 내용

[0041] 개인의 생물학적 상태를 예측하는 데 사용될 수 있는 확고한 유전자 시그니처 식별용 연산 시스템 및 방법이 본원에 기술된다. 특히, 생물학적 상태는 개인의 흡연 노출 반응 상태에 상응할 수 있다. 본원에 기술된 유전자 시그니처는 비흡연자 또는 금연자로부터 현재 흡연을 하는 피험자를 구별할 수 있다. 본원에 기술된 실시예는 주로 흡연자 상태 또는 흡연 노출 반응 상태에 관한 것이지만, 당업자는 본 개시의 시스템 및 방법이 개인의 생물학적 상태를 예측하기 위한 유전자 시그니처를 식별하는 크라우드 소싱 접근법을 사용하는 데 적용될 수 있다는 것을 이해할 것이다(여기서, 생물학적 상태는 흡연 노출 반응 상태, 흡연자 상태, 질병 상태, 생리학적 상태, 화학 물질 노출 상태, 또는 개인의 생물학적 데이터와 관련된 임의의 다른 적절한 상태 또는 임의의 상태를 지칭할 수 있음).

[0042] 본원에서 사용된 바와 같이, 개인의 생물학적 상태는 질병에서 발생하거나 하나 이상의 독성 물질, 약물, 환경 변화(예를 들어 온도, 미세 중력, 압력 및 방사선), 또는 이들의 임의의 적절한 조합에 대한 노출에 반응하여 발생할 수 있는 다양한 분자 변화를 대표하는 것일 수 있다. 기준(criteria)은 예측 분류 모델에 대해 정의되며 예측 분류 모델의 개발 및 트레이닝을 위한 연산 분석에 사용된다. 클래스를 구별하는 특징들이 추출되어 클래스 예측을 위한 분류 모델 내에 삽입된다. 본원에서 사용된 바와 같이, 분류기(classifier)는 클래스 예측을 위해 사용되는 판별 특징 및 규칙을 포함한다.

[0043] 본원에 기술된 크라우드 소싱 접근법은 확고한 유전자 시그니처를 식별하는데 사용되어 하나 이상의 화학 물질에 대한 개인의 노출 상태를 예측할 수 있다. 하기 실시예 1과 관련하여 기술된 연구는 연기에 대한 개인의 노출을 예측하기 위한 유전자 시그니처를 식별하기 위한 하나의 이러한 크라우드 소싱 접근법의 예시적인 도시를 포함한다. 아래에 기술된 실시예 1의 연구는 대중(예, 다수의 도전 참가자)으로부터 수득한 인간 혈액 기반의 흡연 노출 반응 유전자 시그니처에 대한 유전자 목록과, 대중으로부터 수득한 종 독립적 혈액 기반의 흡연 노출 반응 유전자 시그니처를 위한 유전자 목록을 제공한다. 본원에 기술된 유전자 시그니처는 개인이 흡연에 노출되었는지 여부를 예측하기 위해 새로운 인간(인간 시그니처) 또는 인간 및 설치류(종 독립적 시그니처) 혈액 유전자의 발현 샘플 데이터에 적용될 수 있는 하나 이상의 분류 모델에 적용될 수 있다. 본원에 기술된 시스템 및 방법은 개인이 하나 이상의 화학 물질에 노출되었는지 여부를 예측하기 위해 유전자 시그니처 및 하나 이상의 분류 모델을 식별하도록 확장될 수 있다. 하기 실시예 1과 관련하여 기술된 연구는 혈액 기반 유전자 시그니처를 식별하는 것에 관한 것이지만, 당업자는 본 개시의 시스템 및 방법이 크라우드 소싱 접근법을 사용하여 혈액에만 의존하지 않는 유전자 시그니처를 식별하는데 적용할 수 있다는 것을 이해할 것이다. 대신에, 본 개시는 예를 들어 단백질 및 메틸화 변화와 같은, 조직 및 다른 특징에 기초하여 유전자 시그니처를 식별하는데 적용될 수 있다.

[0044] 본원의 시스템 및 방법은 독성 물질에 대한 노출을 예측할 수 있는 마커를 식별하는데 사용될 수 있다. 실제로, 새로운 샘플에 적용된 견고한 마커 기반 분류 모델은 (i) 피험자가 화학 물질에 노출되었는지 여부를 예측가능하게 할 수 있고, (ii) 시간에 따른 노출 반응의 강도를 제품을 테스트하거나 회수하는 동안에 모니터링하도록 할 수 있다.

[0045] 본원에서 사용된 바와 같이, "확고한" 유전자 시그니처는 연구, 실험실, 샘플 공급원 및 기타 인구 통계학적 요인에 걸쳐 강력한 성과를 유지하는 것이다. 중요하게는, 큰 개인 편차를 포함하는 모집단 데이터 집합에서도 확고한 시그니처를 검출할 수 있어야 한다. 데이터 세트 전반의 강인성은 시그니처 성능에 대한 지나치게 낙관적인 보고를 피하기 위해 적절히 검증되어야 한다.

[0046] 시스템 생물학은 생물학적 시스템이 외부 자극(예, 약물, 영양 및 온도) 및 유전적 변형(예, 돌연변이, 후생적 변형)에 반응하거나 적응하는 메커니즘에 대한 자세한 이해를 생성하는 것을 목표로 한다. 새로운 기계론적 통찰력은 오믹스(omics) 또는 고 함량 스크리닝(high content screening)과 같은 첨단 기술을 사용하여 생성된 다량의 분자 및 기능적 데이터의 분석 및 통합을 통해 얻어진다. 독성학 분야에 적용될 경우, 시스템 독성학으로 지칭되는 전반적인 접근법은 생체 이물질(예, 살충제, 화학 물질)에 의해 유발된 생물학적 시스템 혼란을 정량화하고, 독성의 작용 모드를 설명하고, 관련 위험을 평가할 수 있게 한다. 시스템 독성학은 단기 관측치로 장기 결과로 추정하고, 실험적인 시스템으로부터 식별된 잠재 위험을 인간에 대해 해석하는 능력을 가지고 있는데, 이는 이를 응용하는 것이 위험 평가 및 의사 결정을 위한 새로운 표준이 될 수 있음을 시사한다. 예측 독성학적 결과 및 위험 추정치에 대한 외삽 및 해석을 비롯하여 시스템 독성학 데이터는 고급 연산 방법론의 개발을 필요로 한다. 새로운 연산 접근법의 개선된 성능과 신뢰성을 입증하기 위해, 연구자들은 최첨단 방법에 대해 자신의 기술을 벤치마킹 할 수 있지만, 편향된 평가를 초래하는 소위 "자체 평가의 덫"에 종종 빠진다. 또한, 시스템 생물학/독성학에서 생성되고 분석되는 데이터가 쇄도하면 심사원은 공개된 결과와 결론에 대한 지루한 검토를 하게 된다. 검토자가 원칙적으로 공개 저장소에 저장된 원시 데이터에 접근할 수 있지만 전체 분석을 스스로 재현하는 것은 종종 어렵다. 그러므로, 외부의 제삼자가 참여하는, 방법 및 데이터에 대한 독립적이고 객관적인 평가 또는 검증에 대한 분명한 요구가 있다. 본 개시의 시스템 및 방법은 이러한 요구를 다루고, 연구원으로부터 제출물을 받는 크라우드 소싱 방식을 제공하고, 최선의 수행 기술을 식별하고, 이들의 결과를 집계하여 생물학적 상태를 예측하기 위한 확고한 유전자 시그니처를 생성한다.

[0047] 도 1은 본원에 개시된 시스템 및 방법을 구현하는데 사용될 수 있는 컴퓨터 네트워크 및 데이터베이스 구조의 예를 나타낸다. 도 1은, 예시적인 구현예에 따라, 크라우드 소싱을 사용하여 유전자 시그니처의 식별을 수행하기 위한 컴퓨터 시스템(100)의 구성도이다. 시스템(100)은 서버(104) 및 컴퓨터 네트워크(102)를 통해 서버(104)에 접속된 2개의 사용자 장치(108a 및 108b)(사용자 장치(108)로 통칭함)를 포함한다. 서버(104)는 프로세서(105)를 포함하고, 각 사용자 장치(108)는 프로세서(110a 또는 110b) 및 사용자 인터페이스(112a 또는 112b)를 포함한다. 본원에서 사용된 바와 같이, "프로세서" 또는 "연산 장치"라는 용어는 본원에 기술된 하나 이상의 컴퓨터 기술을 수행하기 위해 하드웨어, 펌웨어 및 소프트웨어로 구성된 하나 이상의 컴퓨터, 마이크로 프로세서, 논리 장치, 서버 또는 기타 장치를 지칭한다. 프로세서 및 처리 장치는 현재 처리되는 입력, 출력 및 데이터를 저장하기 위한 하나 이상의 메모리 장치를 포함할 수도 있다. 본원에 기술된 프로세서 및 서버들 중 임의의 것을 구현하는데 사용될 수 있는 예시적인 연산 장치(200)는 도 2를 참조하여 아래에서 상세히 기술된다. 본

원에서 사용된 바와 같이, "사용자 인터페이스"는 하나 이상의 입력 장치(예, 키패드, 터치 스크린, 트랙볼, 음성 인식 시스템, 등) 및/또는 하나 이상의 출력 장치(예, 시각 디스플레이, 스피커, 촉각 디스플레이, 인쇄 장치, 등)의 임의의 적절한 조합을 제한없이 포함한다. 본원에서 사용된 바와 같이, "사용자 인터페이스"는 본원에 기술된 하나 이상의 컴퓨터화된 동작 또는 기술을 수행하기 위해 하드웨어, 펌웨어 및 소프트웨어로 구성된 하나 이상의 장치의 임의의 적절한 조합을, 제한없이 포함한다. 사용자 장치의 예로는 개인용 컴퓨터, 랩톱 및 모바일 장치(예컨대 스마트폰, 태블릿 컴퓨터, 등)를 제한없이 포함한다. 도면이 복잡해지는 것을 피하기 위해, 도 1에는 하나의 서버, 하나의 데이터베이스, 및 2개의 사용자 장치만이 도시되지만, 당업자는 시스템(100)이 다수의 서버 및 임의의 수의 데이터베이스 또는 사용자 장치를 지원할 수 있음을 이해할 것이다.

[0048] 컴퓨터화된 시스템(100)은 개인의 생물학적 상태를 예측하기 위한 유전자 시그니처를 식별하는데 있어서 대중의 지혜를 이용하는데 사용될 수 있다. 전술한 바와 같이, 시스템 생물학을 연구하는 과학자는 종종 자체 평가의 덫에 빠져 편향된 평가를 초래한다. 본원에 기술된 크라우드 소싱 방식은, 해결 과제를 설계하고, (유전자의 발현 및 알려진 생물학적 상태 데이터베이스(106)에 대한 데이터를 사용자 장치(108)에 이용 가능하게 함으로써) 이를 과학계에 공개하고, (예를 들어, 사용자 장치(108a 및 108b)로부터) 독립된 과학자 또는 그룹으로부터의 제출물을 수신하고, 최선의 수행 결과 또는 예측을 집계함으로써 이러한 편향을 피하는데 도움을 준다. 광범위한 참여를 보장하기 위해, 과제는 공통 관심사의 과학적 문제(예: 개인의 생물학적 상태 또는 흡연자 상태를 예측하기 위한 혈액 기반 유전자 시그니처의 식별)와 관련된 질문을 다루는 것을 목표로 할 수 있다.

[0049] 과제는 개인의 그룹으로부터 수득한 혈액 샘플 데이터와 관련된 특정 데이터를 과학계가 이용할 수 있게 한다. 특히, 유전자 발현 및 알려진 생물학적 상태 데이터베이스(106)(데이터베이스(106)로 통칭함)는 한 세트의 개인의 알려진 생물학적 상태 및 유전자 발현 데이터(환자 세트로부터의 혈액 샘플로부터 수득됨)를 대표하는 데이터를 포함하는 데이터베이스이다. (혈액 샘플이 데이터베이스(106)에 저장된) 한 세트의 개인에서의 각 개인은 트레이닝 샘플 또는 테스트 샘플로서 무작위로 배정될 수 있다. 일부 구현예에서, 트레이닝 샘플 또는 테스트 샘플로서의 개인을 배정하는 것은 완전한 무작위 배정이 아닐 수 있다. 이 경우, 할당하는 동안에 하나 이상의 기준이 사용될 수 있다 (예컨대, 서로 다른 생물학적 상태를 가진 비슷한 수의 개인이 트레이닝 및 테스트 데이터 세트 각각에 있도록 하는 것을 포함함). 일반적으로, 생물학적 상태의 분포가 트레이닝 데이터 세트 및 테스트 데이터 세트에서 다소 유사함을 보장하는 한편, 임의의 적합한 방법이 개인을 트레이닝 또는 테스트 샘플로서 할당하는 데 사용될 수 있다.

[0050] 각 트레이닝 샘플 및 테스트 샘플은 개인의 혈액 샘플뿐만 아니라 개인의 알려진 생물학적 상태(예, 개인의 알려진 흡연자 상태)로부터 측정된 유전자 발현 수준을 포함한다. 트레이닝 샘플은 트레이닝 데이터 세트를 구성하고, 테스트 샘플은 테스트 데이터 세트를 구성한다. 전체 트레이닝 데이터 세트가 데이터베이스(106)로부터 사용자 장치(108)에 제공되는 반면, 테스트 데이터 세트의 일부만이 사용자 장치(108)에 제공된다. 특히, 테스트 샘플로부터의 측정된 유전자 발현 수준이 사용자 장치(108)에 제공되지만, 테스트 샘플에 상응하는 알려진 생물학적 상태는 사용자 장치(108)로부터 숨겨진 채로 유지된다.

[0051] 사용자 장치(108)의 과학자는 트레이닝 데이터 세트 내의 개인의 생물학적 상태 및 측정된 유전자 발현 수준 간의 임의의 의존성, 연관성 또는 상관 관계를 식별하기 위해 트레이닝 샘플을 분석할 수 있다. 식별된 상관 관계는 후보 유전자 시그니처 및 분류기의 형태를 가질 수 있다. 후보 유전자 시그니처는 상이한 생물학적 상태(예, 현재 흡연자 대 현재 비흡연자)와 관련되는 샘플에 대해 차별적으로 발현되는 유전자의 목록을 포함한다. 과학자는 필터, 래퍼 및 내재된 방법과 같은 임의의 특징 선택 기술을 사용하여 후보 유전자 시그니처를 적절한 컴퓨터 기술을 사용해 식별할 수 있다. 추출된 특징은 판별 분석, 지원 벡터 머신, 선형 회귀, 로지스틱 회귀, 의사 결정 트리, 나이브 베이즈, k-최근접 이웃, K-평균, 랜덤 포레스트 또는 임의의 적합한 기술과 같은 기계 학습(machine learning) 접근법을 사용하여 트레이닝된 분류 모델에서 결합된다. 분류기는, 개인의 예측된 생물학적 상태를 지칭할 수 있는 클래스에 샘플을 배정하기 위해, 후보 유전자 시그니처에서 유전자의 발현 수준을 사용하는 결정 규칙 또는 매핑을 포함한다. 이러한 방식으로, 각 사용자 장치(108)에서의 각 과학자는 트레이닝 데이터 세트에 기초하여 후보 유전자 시그니처 및 분류기를 식별한다.

[0052] 사용자 장치(108)의 과학자는 그들의 후보 유전자 시그니처 및 분류기를 사용하여 테스트 데이터 세트 내에서 테스트 샘플의 생물학적 상태를 예측한다. 후보 유전자 시그니처 및 각 테스트 샘플에 대해 수득된 결과는 네트워크(102)를 통해 사용자 장치(108)로부터 서버(104)에 제공된다. 과학자로부터의 제출물은 익명일 수 있다. 일 실시예에서, 각각의 테스트 샘플에 대한 결과는 상응하는 테스트 샘플이 예측된 생물학적 상태에 속할 우도 또는 확률에 상응하는 신뢰 수준을 포함한다. 신뢰 수준은 도 3의 단계(308)와 관련하여 상세히 설명된다. 또 다른 실시예에서, 결과는 신뢰 수준을 포함하지 않고 오히려 각 테스트 샘플에 대한 예측된 생물학적 상태만을 포

함한다.

[0053] 서버(104)는 각각의 테스트 샘플에 대해 수득된 결과를 각각의 테스트 샘플에 대한 알려진 생물학적 상태와 비교함으로써 최고 수행 후보 유전자 시그니처를 식별할 수 있다. 일반적으로, 최고 수행 후보 유전자 시그니처는 알려진 생물학적 상태와 밀접하게 일치하는 결과를 가진다. 그런 뒤에, 서버(104)는 개인의 생물학적 상태를 예측하는데 사용될 수 있는 확고한 유전자 시그니처를 얻기 위해 최고 수행 후보 유전자 시그니처에 걸쳐 집계한다. 이 프로세스는 도 3의 단계(314, 316, 및 318)와 관련하여 보다 자세히 기술된다.

[0054] 도 1의 시스템(100) 구성 요소는 다수의 방식 중 하나의 방식으로 배치, 분산 및 결합될 수 있다. 예를 들어, 네트워크(102)를 통해 접속된 다수의 처리 장치 및 저장 장치에 대해 시스템(100)의 구성 요소를 분산하는 컴퓨터 시스템이 사용될 수 있다. 이러한 구현예는 공통 네트워크 자원에 대한 액세스를 공유하는 무선 및 유선 통신 시스템을 포함하는 다중 통신 시스템을 통한 분산 컴퓨팅에 적합할 수 있다. 일부 구현예에서, 시스템(100)은 하나 이상의 컴포넌트가 인터넷 또는 다른 통신 시스템을 통해 접속된 상이한 처리 서비스 및 저장 서비스에 의해 제공되는 클라우드 컴퓨팅 환경에서 구현된다. 서버(104)는 예를 들어 클라우드 컴퓨팅 환경에서 인스턴스화된 하나 이상의 가상 서버일 수 있다. 일부 구현예에서, 서버(104)는 데이터베이스(106)와 결합되어 하나의 구성 요소가 된다.

[0055] 도 3은 개인의 생물학적 상태를 예측하기 위해 크라우드 소싱을 사용하여 유전자 시그니처를 식별하는 방법(300) 에 대한 흐름도이다. 상기 방법(300)은 서버(104)에 의해 실행될 수 있으며, 유전자의 발현 데이터 및 알려진 생물학적 상태를 포함하는 트레이닝 데이터 세트를 사용자 장치 세트에 제공하는 단계(단계(302)), 유전자의 발현 데이터를 포함하는 테스트 데이터 세트를 사용자 장치 세트에 제공하는 단계(단계(304)), 트레이닝 데이터 세트 내의 상이한 생물학적 상태들 사이에서 판별될 것으로 결정되는 유전자 세트를 포함하는 후보 유전자 시그니처를 수신하는 단계(단계(306)), 및 각 후보 유전자 시그니처에 대해, 트레이닝 데이터 세트 내의 각 샘플에 대한 신뢰 수준을 수신하는 단계(단계(308))를 포함한다. 상기 방법(300)은, 신뢰 수준과 테스트 데이터 세트 내의 알려진 생물학적 상태 간의 비교에 기초하여 제1 성과 기준에 따라 후보 유전자 시그니처를 순위 매김하는 단계(단계(310)), 각각의 후보 유전자 시그니처에 대해, 신뢰 수준을 사용하여 테스트 데이터 세트의 각 샘플을 예측된 생물학적 상태로 배정하는 단계(단계 312), 예측된 생물학적 상태가 테스트 데이터 세트 내의 알려진 생물학적 상태와 일치하는지 여부에 기초하여 후보 유전자 시그니처를 제2 성과 기준에 따라 순위 매김하는 단계(단계(314)), 단계(310 및 314)에서 할당된 순위에 기초하여 제3 성과 기준에 따라 후보 유전자 시그니처의 순위 매김하는 단계(단계 316), 최상위 후보 유전자 시그니처에서 후보 유전자 시그니처의 적어도 임계수에 포함되는 유전자를 식별하는 단계(단계(318))를 더 포함한다.

[0056] 단계(302)에서, 유전자의 발현 데이터 및 트레이닝 샘플의 세트에 대한 알려진 생물학적 상태를 포함하는 트레이닝 데이터 세트가 사용자 장치(108) 세트에 제공된다. 도 1과 관련하여 기술된 바와 같이, 단계(302)에서 제공되는 트레이닝 데이터 세트는 개인의 혈액 샘플뿐만 아니라 개인의 알려진 생물학적 상태로부터 측정된 유전자의 발현 수준을 포함하는 트레이닝 샘플을 포함한다. 사용자 장치(108)의 과학자는 트레이닝 데이터 세트를 수신하고 트레이닝 데이터 세트를 사용하여 측정된 유전자의 발현 수준과 알려진 생물학적 상태 사이에서 맵핑을 제공하는 분류기를 트레이닝 한다. 단계(304)에서, 유전자의 발현 데이터를 포함하는 테스트 데이터 세트가 사용자 장치 세트(108)에 제공된다. 도 1과 관련하여 기술된 바와 같이, 단계(304)에서 제공되는 테스트 데이터 세트는, 개인의 혈액 샘플로부터 측정된 유전자의 발현 수준만을 포함하되 개인의 알려진 생물학적 상태는 포함하지 않는 테스트 샘플을 포함한다. 다시 말해, 테스트 샘플의 알려진 생물학적 상태는 사용자 장치(108)의 과학자로부터 숨겨진다.

[0057] 단계(306)에서, 트레이닝 데이터 세트 내의 상이한 생물학적 상태들 사이에서 판별되도록 결정되는 유전자 세트를 포함하는 후보 유전자 시그니처가 수신된다. 사용자 장치(108)에서 각 과학자 또는 과학자 팀은 후보 유전자 시그니처를 서버(104)에 제공할 수 있는데, 과학자는 후보 유전자 시그니처에서의 유전자 발현 수준의 조합이 하나 이상의 기준(예컨대 생물학적 상태 또는 트레이닝 반응 데이터 세트 내의 샘플에 대한 노출 반응 상태)에 대해 판별되는 것으로 결정했다. 트레이닝 데이터 세트가 제공되는 사용자 장치는 과학자가 후보 유전자 시그니처를 제공하는 사용자 장치와 동일하거나 상이할 수 있다.

[0058] 단계(308)에서, 각각의 후보 유전자 시그니처에 대해, 테스트 데이터 세트 내의 각 테스트 샘플에 대한 신뢰 수준이 수신된다. 신뢰 수준은 0 내지 1의 값일 수 있으며, 이는 상응하는 테스트 샘플이 특정 생물학적 상태에 속할 우도를 나타낸다. 일 실시예에서, 2개의 생물학적 상태(예, 제1 생물학적 상태 및 제2 생물학적 상태)가 있는 경우, 신뢰 수준은, 특정 테스트 샘플이 제1 생물학적 상태에 속할 우도를 의미하는 p 값에 대응할 수 있

다. 이 경우, 1-p 값은 특정 테스트 샘플이 제2 생물학적 상태에 속할 우도를 나타낼 수 있다. 일반적으로, 3개 이상의 생물학적 상태가 존재할 때, 다수의 신뢰 수준이 각각의 테스트 샘플 및 각 후보 유전자 시그니처에 제공될 수 있다.

[0059] 단계(310)에서, 서버(104)는 신뢰 수준((단계(308)에서 수신됨)과 테스트 데이터 세트 내의 알려진 생물학적 상태 간의 비교에 기초하여 제1 성과 기준에 따라 후보 유전자 시그니처(단계(306)에서 수신됨)를 순위 매김한다. 단계(310)에서 수행된 순위 매김은 각각의 후보 유전자 시그니처에 제1 순위 값이 배정되게 한다.

[0060] 후보 유전자 시그니처의 성과를 평가하는 하나의 방법은 예측된 생물학적 상태의 행(row)과 실제 생물학적 상태의 열(column)을 포함하는 표에 예측 결과를 표시하는 것이다. 아래 도시된 표 1은 예측 결과를 표시하는 하나의 방법의 예이다. 표의 제1 행은 실제로 제1 생물학적 상태(예, 진짜 현재 흡연자)를 가진 개인의 수와 샘플이 제1 생물학적 상태(예, 예측된 현재 흡연자)와 관련이 있다고 예측되는, 실제로 제2 생물학적 상태(예, 현재 비흡연자)를 가진 개인의 수를 나타낸다. 표의 제2 행은 실제로 제1 생물학적 상태(예, 진짜 현재 흡연자)를 가진 개인의 수와 샘플이 제2 생물학적 상태(예, 예측된 현재 비흡연자)와 관련이 있다고 예측되는, 실제로 제2 생물학적 상태(예, 현재 비흡연자)를 가진 개인의 수를 나타낸다.

표 1

[0061]

	실제 생물학적 상태 1	**실제 생물학적 상태 2**
예측 생물학적 상태 1	진양성	위양성
예측 생물학적 상태 2	위음성	진음성

[0062] 완벽한 예측 변수(predictor)는 모든 개인이 실제로 제1 생물학적 상태를 갖는 것으로 정확하게 예측되는 제1 생물학적 상태를 가지며(진양성은 100%일 것이고 위음성은 0%일 것임), 실제로 제2 생물학적 상태를 갖는 모든 개인은 제2 생물학적 상태를 갖는 것으로 정확히 예측될 것이다(진음성은 100%일 것이고 위양성은 0%일 것임). 본원에 기술된 바와 같이, 개인은 흡연 상태(예, 현재 흡연자, 현재 비흡연자, 이전 흡연자, 흡연 비경험자, 등)와 같은 다수의 생물학적 상태로 분류될 수 있지만, 일반적으로 당업자는 본원에 기술된 시스템 및 방법이 임의의 분류 체계에 적용 가능하다는 것을 이해할 것이다. 예측 변수(예, 분류기 및 후보 유전자 시그니처)의 강도를 평가하기 위해, 예측 결과 표의 값에 기초한 다양한 기준이 사용될 수 있다. 제1 실시예에서, 일 기준은 본원에서 제1 생물학적 상태를 실제로 갖는 개인들의 세트 중에서 제1 생물학적 상태(예, 현재 흡연자)로 정확하게 분류된 개인들의 비율인 "민감도" 또는 "재현율"로 언급된다. 다시 말해, 민감도(또는 재현율) 기준은 진양성의 수를 진양성과 위음성의 합으로 나눈 값, 또는 TP / (TP+FN)과 같다. 민감도 값 1은, 제1 생물학적 상태에 속하는 모든 샘플이 실제로 제1 생물학적 상태에 속하는 것으로 정확히 예측되었음을 나타내지만, 얼마나 많은 기타 샘플이 제1 생물학적 상태(FP)에 속하는 것으로 잘못 예측되었는지에 관한 정보는 제공하지 않는다.

[0063] 제2 실시예에서, 일 기준은 제2 생물학적 상태를 실제로 갖는 개인들의 세트중에서 제2 생물학적 상태(예, 현재 비흡연자)로 정확하게 분류된 개인들의 비율인 "특이도"로서 본원에서 지칭된다. 다시 말해, 특이도는 진음성의 수를 진음성과 위양성의 합으로 나눈 값, 또는 TN / (TN+FP)과 같다. 특이도 값 1은, 제2 생물학적 상태에 속하는 모든 샘플이 실제로 제2 생물학적 상태에 속하는 것으로 정확히 예측된 것을 나타내지만, 제2 생물학적 상태(FN)를 갖는 것으로 잘못 예측된 제1 생물학적 상태를 갖는 샘플의 수에 관한 정보는 제공하지 않는다.

[0064] 제3 실시예에서, 일 기준은 제1 생물학적 상태를 가질 것으로 예측되는 개인들의 세트중에서 제1 생물학적 상태(예, 현재 흡연자)로 정확하게 분류된 개인들의 비율인 "정밀도"로서 본원에서 지칭된다. 다시 말해, 정밀도 기준은 진양성의 수를 진양성과 위음성의 합으로 나눈 값, 또는 TP / (TP+FP)와 같다. 정밀도 값 1은, 특정 클래스에 속한다고 예측된 모든 샘플이 실제로 그 클래스에 속하는 것을 나타내지만, 제2 생물학적 상태(FN)를 갖는 것으로 잘못 예측된 제1 생물학적 상태를 갖는 샘플의 수에 관한 정보는 제공하지 않는다.

[0065] 강력한 예측 변수로 간주되기 위해서는 민감도와 특이도 모두, 민감도와 정밀도 모두, 또는 민감도, 특이도 및 정밀도 모두에서 높은 값이 바람직할 수 있다. 후보 유전자 시그니처의 성과를 평가하기 위해 본원에서 민감도,

특이도 및 정밀도 기준을 사용할 수 있지만, 일반적으로, 음성 테스트(TN / (TN+FN))의 예측 값과 같은 본 개시의 범위를 벗어나지 않는, 임의의 기타 기준이 사용될 수도 있다.

[0066] 일 실시예에서, 제1 성과 기준은 곡선 하 면적(AUC) 기준과 관련된다. 특히, 곡선은 수신기 동작 특성(ROC) 곡선 또는 정밀도 재현율(PR) 곡선에 해당할 수 있다. ROC 곡선의 축은 민감도(또는 진양성률: TP / (TP + FN))과 위음성률(FP / (FP+TN))에 해당한다. PR 곡선의 축은 민감도(TP / (TP+FN))와 정밀도(TP / (TP FP))에 해당한다. 일 실시예에서, PR 곡선 하 면적(AUPR)은 특정 후보 유전자 시그니처에 대한 제1 순위를 획득하도록 제1 성과 기준으로서 사용된다. 또 다른 실시예에서, ROC 곡선 하 면적은 제1 성과 기준으로서 사용된다. PR 곡선 및/또는 ROC 곡선은 연속적일 수 있지만, 본 발명은(임계치가 변화됨에 따라) 불연속 값을 사용할 수 있고, 하나 이상의 보간(interpolation) 기술이 곡선 아래의 영역을 연산하는데 사용될 수 있다.

[0067] 단계(312)에서, 각각의 후보 유전자 시그니처에 대해, 서버(104)는 신뢰 수준을 사용하여 테스트 데이터 세트의 각 샘플을 예측된 생물학적 상태로 할당한다. 특히, 과학자들의 각 제출물에 대해, 각 테스트 샘플은 제출물의 신뢰 수준을 기반으로 예측된 생물학적 상태에 할당된다. 일 실시예에서, 2개의 생물학적 상태(제1 생물학적 상태 및 제2 생물학적 상태)가 있는 경우, 신뢰 수준은 테스트 샘플이 제1 생물학적 상태에 속할 확률을 나타내는 p 값을 가질 수 있다. 또한, 1-p 값은 테스트 샘플이 제2 생물학적 상태에 속할 확률에 대응할 수 있다. 일반적으로, 과학자는 여러 생물학적 상태가 있을 때 여러 신뢰 수준을 제출할 수 있으며 특정 후보 유전자 시그니처에 대한 예측된 생물학적 상태는 가장 높은 신뢰 수준을 갖는 생물학적 상태와 일치할 수 있다.

[0068] 단계(314)에서, 서버는 예측된 생물학적 상태(단계(312)에서 수득됨)가 테스트 데이터 세트의 알려진 생물학적 상태와 일치하는지 여부에 기초하여 제2 성과 기준에 따라 후보 유전자 시그니처를 순위 매김한다. 단계(314)에서 수행된 순위 매김은 각각의 후보 유전자 시그니처에 제2 순위 값을 할당하게 한다.

[0069] 또 다른 실시예에서, 제2 성과 기준은 매튜(Mathews) 상관 계수(MCC) 기준에 해당할 수 있다. MCC 측정 항목은 모든 진/위 양성비와 음성비를 결합하여, 단일 값의 공정한 측정 기준을 제공한다. MCC는 종합 성과 점수로 사용될 수 있는 성과 기준이다. MCC는 -1 내지 +1의 값이며 본질적으로, 알려진 이진 분류와 예측된 이진 분류 간의 상관 계수이다. MCC는 다음 방정식을 사용하여 연산할 수 있다:

[0070] $$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

[0071] TP: 진양성; FP: 위음성; TN: 진음성; FN: 위음성 그러나, 일반적으로, 성과 기준의 세트에 기초하여 합성 성과 기준을 생성하기 위한 임의의 적절한 기술은 후보 유전자 시그니처 및 그것의 대응하는 예측의 성능을 평가하는데 사용될 수 있다. MCC 값이 +1이면 모델이 완벽한 예측을 획득한 것을 나타내며, MCC 값이 0이면 모델 예측이 무작위보다 낫지 않게 수행함을 나타내고, MCC 값이 -1이면 모델 예측이 완벽하게 부정확함을 나타낸다. MCC는 분류기 함수가 단지 클래스 예측만이 이용 가능하도록 코딩될 때, 쉽게 연산할 수 있다는 이점이 있다. 일반적으로, TP, FP, TN 및 FN을 설명하는 임의의 기준이 본 개시에 따라 제2 성과 기준으로서 사용될 수 있다.

[0072] 단계(316)에서, 서버(104)는 단계(310 및 314)에서 할당된 순위에 기초하여 제3 성과 기준에 따라 후보 유전자 시그니처를 순위 매김한다. 특히, 단계(310)에서의 제1 순위는 원(raw) 신뢰 수준과 테스트 샘플의 알려진 생물학적 상태 간의 비교에 기초하여 획득되며, 단계(314)에서 제2 순위는 예측된 생물학적 상태(신뢰 수준으로부터 평가됨)와 테스트 샘플의 알려진 생물학적 상태 간의 비교에 기초하여 획득된다. 제1 및 제2 순위는 제3 성과 기준을 얻기 위해 평균화(또는 어떤 식으로든 결합)될 수 있다.

[0073] 단계(318)에서, 서버(104)는 N개의 최상위 후보 유전자 시그니처에서 후보 유전자 시그니처의 적어도 하나의 임계 수(예, M)에 포함되는 유전자 세트를 식별한다. 실시예에서, 제3 성과 기준에 따라 N개의 가장 높은 순위의 후보 유전자 시그니처가 결정된다. 이들 N 후보 유전자 시그니처 중 적어도 M개에 나타나는 임의의 유전자는 단계(318)에서 식별된 유전자에 포함되며, 여기서 M은 N 미만이다. 일부 구현에서, (N,M) = (3,2), (4,3), (4,2), (5,4), (5,3), (5,2), (6,5), (6,4), (6,3), (6,2) 또는 N 및 M에 대한 값의 임의의 다른 적절한 조합을 포함하며, 여기서 N은 2 내지 후보 유전자 시그니처 총수 범위의 정수이고, M은 2 내지 N 범위의 정수이다.

[0074] 실시예 1 - 서론

[0075] 개개인의 흡연자 상태를 정확하게 예측하기 위한 확고한 유전자 시그니처를 얻기 위해 크라우드 소싱 방법이 사용되는 예시적인 연구가 본원에 기술된다. 본 연구의 일 목적은 인간과 종에 의존하지 않는 혈액 노출 반응 마커와 흡연 및 중단 상태를 예측하는 모델의 식별을 위한 연산 방법을 벤치마킹하여 혈액 내 화학 물질 노출 반

응의 마커를 식별하는 것이다.

[0076] 실시예 1 - 연구 모집단 및 설계

[0077] 전혈 샘플은 임상 및 생체 내 연구 중에 PAXgeneTM 튜브에 수집하거나, 바이오뱅크(Biobank) 보관소에서 구입한다. 다양한 연구에 대한 샘플 그룹/클래스, 크기 및 특성이 도 6의 표에 요약된다. 간략하게는, 인간 혈액 샘플은 (i) 영국 런던의 Queen Ann Street Medical Center (QASMC)에서 시행되고 ClinicalTrials.gov에 식별자 NCT01780298로 등록된 임상 증례 대조 연구; (ii) Biobank 보관소(BioServe Biotechnologies Ltd., 미국, 메릴랜드주, 벨츠빌)(데이터 세트 BLD-SMK-01)로부터 수득할 수 있다. 이 두 가지 출처의 샘플에는, 잘 정의된 포함 기준(도 6)에서 선택된 흡연자(S), 이전 흡연자(FS) 및 흡연 비경험자(NS); (iii) 무작위 대조군, 대조군, 3 군 병행군 및 단일 센터 연구에 해당하는 임상적 ZRHR-감소 노출(REX) C-03-EU 및 04-JP 연구가 포함된다. REX 연구는 흡연에서 선택된 연기 성분에 대한 노출 감소를 입증하는 것을 목표로 하며, 건강한 피험자는 기존의 담배(흡연자)를 5일 동안 구금 상태에서 계속 사용하는 것과 비교하여 위험감소담배제품("MRTP") 또는 흡연 금욕/중단("Cess")으로 전환한다. 일반적으로, MRTP는 가열식 담배 제품일 수 있다. 본원에서 사용된 바와 같이, 가열식 담배 제품은 사용 동안 담배를 태우거나 연소시키지 않고 담배를 포함하는 담배 또는 혼합물을 가열하여 에어로졸을 발생시키는 제품을 포함한다. 마우스 혈액 샘플은 암컷 C57BL/6 및 ApoE-/*?*-마우스에서 각각 7개월 및 8개월 동안 실시한 2가지 독립적인 담배 연기("CS") 흡입 연구로부터 수득하였다. 연구에는 5개의 그룹으로 무작위로 추출된 마우스가 포함되며, 5개의 그룹은: 가짜(Sham)(공기에 노출), 3R4F(기준 담배(reference cigarette) 3R4F로부터의 CS에 노출), 프로토타입/후보 MRTP(3R4F와 일치하는 니코틴 수준의 프로토타입/후보 MRTP로부터의 주류 에어로졸에 노출), 흡연 중단(Cess), 및 3R4F에 2 개월 노출 후 프로토타입/후보 MRTP로 전환(Switch)이다. 혈액 샘플은 상이한 시점에서 수집된다.

[0078] 실시예 1 - 혈액 전사체학(Transcriptomics) 데이터 세트

[0079] 전사체학 데이터 세트는 PAXgeneTM 튜브에서 수집된 전혈 샘플로부터 생성된다.

[0080] 인간 및 마우스 혈액 샘플로부터의 데이터 생성

[0081] 총 RNA는 PAXgene 혈액 키트를 사용하여 분리된다. RNA 샘플의 농도와 순도는, UV 분광 광도계(NanoDrop® 1000 또는 Nanodrop 8000; Thermo Fisher Scientific, 미국, 매사추세츠주, 월섬)를 사용하여, 230, 260 및 280 nm에서 흡광도를 측정하여 결정된다. RNA 무결성은 Agilent 2100 Bioanalyzer(애질런트 테크놀로지스 사, 미국, 캘리포니아주, 산타클라라)를 사용하여 추가 검사한다. RNA 무결성 수가 6을 초과하는 RNA만 추가 분석을 위해 처리된다.

[0082] 제조사의 지침(퀴아젠 사)에 따라 PAXgeneTM 튜브의 샘플로부터 총 RNA를 분리한다. 추출된 RNA의 품질, Ovation® 전혈 시약 및 Ovation RNA 증폭 시스템 V2(누젠 사, 네덜란드, AC Leek)를 사용하여 표적 제조 후 cDNA 품질, 및 파쇄물(예, 최종 파쇄 및 비오티닐화된 제품의 크기 분포는 전기영동도를 사용하여 모니터링된다)은 Agilent 2100 Bioanalyzer(미국, 캘리포니아 주, 산타클라라)를 사용하여 점검된다. cDNA의 양은 SpectraMax® 384Plus 마이크로 플레이트 리더(몰레큘러 디바이스 사, 미국, 캘리포니아 주, 서니베일)로 측정한다. cDNA 품질은 Fragment analyzer(어드밴스트 애널리티컬, 미국, 아이오와 주, 엔케니)를 사용하여 단편화되지 않은 cDNA의 크기를 평가하여 결정된다. 단편화 및 라벨링 후 cDNA 단편을 제조사의 지침에 따라 GeneChip® 인간 유전체 U133 플러스 2.0 어레이(Human Genome U133 Plus 2.0 Array)(아피매트릭스 사)에서 하이브리드화 한다. 원(raw) 전사체학 데이터는 마이크로 어레이 이미지 분석에서 획득한다. QASMC 연구에서 혈액 전사체학 데이터는 AROS 어플라이드 바이오테크놀로지 AS 사(덴마크, 오르후스)에서 생산된다.

[0083] *데이터 처리*

[0084] 각 데이터 세트의 원(raw) 데이터(CEL 파일)는 동결 로부스트 마이크로어레이 분석(frozen Robust Microarray Analysis), fRMA v1.1을 사용하여 R 환경(v3.1.2)에서 처리되고 표준화된다. Frma 및 GNUSE 함수는 인간 동결 변수 벡터(hgu133plus2frmavecs v1.3.0)를 사용한다. 인간(hgu133plus2hsentrezgcdf v16.0.0)에 대한 맞춤형 브레인어레이 cdf 파일은, 아피매트릭스 사의 프로브-대-앙트레(probe-to-entrez) 유전자 ID 매핑에 사용되어 일 유전자 관계에 대해 일 프로브가 설정된다.

[0085] 데이터는, 본원에 기술된 기준에 따라 다음 컷오프 중 하나를 통과하지 못한 모든 CEL 파일이 제거되는, 품질 점검 단계를 거친다. 첫 번째, 주어진 프로브 세트j에 대해, 표준화된 비눈금 표준 오차(NUSE)는 주어진 배열 i에 대한 발현의 추정치의 정밀도를 기타 어레이와 비교하여 제공한다. 문제가 있는 어레이는 중간값 SE보다 표준 오차(SE)가 높게 된다. NUSE 중간값 1을 초과하거나 어레이가 큰 사분위수 범위(IQR)를 갖는 경우, 어레이의

품질이 나쁠것으로 추정된다. NUSE 값이 1.05보다 높은 어레이는 제거된다. 두 번째, RLE(Relative Log Expression)는 모든 j 어레이에 대해 해당 프로브에 대한 강도의 중간값 수준에 상대적인 특정 프로브의 강도 수준을 각 어레이에 대해 비교한다. RLE의 어레이-특정 분포는 특정 어레이에 주로 낮거나 높은 발현된 특징이 있는지 결정하는데 사용된다. 0에 가깝지 않은 중앙값 RLE는 상향 조절된 유전자의 수가 하향 조절된 유전자의 수와 거의 같지 않음을 나타내며, 큰 RLE IQR은 대부분의 유전자가 차별적으로 발현된다는 것을 나타낸다. 중간값이 RLE> 0.1(절대 값)인 어레이는 이상치(outlier)로 간주되어 제거된다. 세 번째, 모든 어레이 데이터 세트의 평균 절대 편차(MARLE)가, 0.01의 제곱근으로 나뉘어진 값(또는 중간값(MARLE)/(1.4826*mad(MARLEs)> 1/0.01의 제곱근))을 초과하는 중앙 절대 RLE(MARLE)를 갖는 어레이는 품질이 나쁜 칩으로 간주되어 제거된다.

[0086] 마우스와 인간에 대한 맞춤형 브레인어레이 CDF 파일은, 아피매트릭스 사의 프로브-대-앙트레(probe-to-Entrez) 유전자 ID 매핑에 사용되어, 일 유전자 관계에 대해 일 프로브가 설정된다(HGU133Plus2_Hs_ENTREZG v16.0, Mouse4302_Mm_ENTREZG v16.0 각각). 품질 검사는 최소 품질 기준을 통과하지 못하는 CEL 파일을 배제한다. 데이터 세트 처리를 용이하게 하기 위해, 인간 및 마우스 유전자의 발현 데이터 세트는, 둘 모두 인간 유전자 시그니처를 구비한다. 마우스 유전자는 NCBI/HCOP 매핑 파일을 사용하여 인간 유전자와 일치된다. 마우스 유전자가 여러 인간 유전자에 매핑되는 경우, 대문자로된 마우스 유전자와 일치하는 인간 유전자만 보유된다.

[0087] 실시예 1 - 도전 개요

[0088] 이러한 도전에 대하여, 흡연자(S) 및 현재 비흡연자(NCS) 피험자의 혈액으로부터의 유전자의 발현 프로파일이 예컨대 도 1과 관련하여 기술된 네트워크(102)를 통해 과학계에 제공된다. 유전자의 발현 프로파일 세트는 트레이닝 세트와 테스트 세트로 균등하게 나뉜다. 트레이닝 데이터 세트(피험자: 흡연자, 이전 흡연자, 흡연 비경험자 클래스의 생물학적 상태에 대한 정보가 가득함)는 테스트 데이터 세트(피험자의 생물학적 상태에 대한 정보 없음)가 발표되기 전에 발표된다. 135명의 등록된 과학자가 61개 팀으로 그룹화된다. 61개 팀 중 23개 팀이 도전 규칙에 따라 제출물을 제공하고, 23개 팀 중 12개 팀이 적격한 제출물을 제공한다. 도 7a는 도전의 목적이, 인간 및 마우스의 전혈 유전자의 발현 데이터로부터 화학적 노출 반응 마커를 식별하고, 노출되거나 비노출된 그룹의 부분으로서 새로운 혈액 샘플의 예측 분류를 위한 연산 모델에서 이러한 마커를 시그니처로서 활용하는 것임을 나타낸다.

[0089] 데이터는 인간과 설치류에서의 CS 노출 및 중단과 관련된 독립적인 임상 및 생체 내 연구로부터 수집된 혈액 샘플로부터 수득된다. 실험 그룹은 또한 일정 기간 동안 CS에 노출된 후 프로토타입/후보 MRTP에 노출되거나 프로토타입/후보 MRTP로 전환된 개인을 포함한다. 참가자는 혈액 샘플에서 생성된 대상의 유전자의 발현 프로파일에 기초하여 흡연 노출을 예측하는 모델을 개발하도록 요청받는다. 구체적으로, 참가자는 2가지 과업을 해결하도록 요청받으며, 2가지 과업은: (1) 흡연자 대 현재 비흡연자를 식별, 및 (2) 현재 비흡연자로서 예측되는 각 피험자에 대해 피험자가 이전 흡연자(FS)이거나 흡연 비경험자(NS)인지 여부를 식별하는 것이다. 득점에 적격하기 위해, 팀은 2가지 작업에 대한 예측(예, 각 테스트 샘플의 신뢰 수준)과 후보 유전자 시그니처(최대 40개의 유전자 포함)를 제출해야 한다. 도전이 끝나면 익명의 예측은 외부 전문가 위원회로 수립된 경로(pipeline)라인에 따라 채점된다. 이 도전에서 최선의 수행자는 흡연자와 현재 비흡연자를 구별하기 위한 완벽에 가까운 예측을 달성했다.

[0090] *도전 목표 및 규칙*

[0091] 참가자는 (i)흡연자와 현재 비흡연자를 구별(과업 1)하고, 이어서 (ii) 현재 비흡연자를 이전 흡연자 및 흡연 비경험자로 분류(과업 2, 도 7b)하기 위해 확고하고 희소한 인간(하위 도전 1, SC1) 및 종 독립적인(하위 도전 2, SC2) 혈액 기반 유전자 시그니처 분류 모델을 개발하도록 요청받는다. 첫 번째 제약으로, 예측 모델은 모델을 재트레이닝/정제할 필요 없이 단일의 새로운 개인 혈액 샘플이 속한 클래스를 예측할 수 있는 능력을 갖도록 귀납적(형질 전환과는 반대로서)일 것을 요청받거나 트레이닝 데이터 세트와 테스트 데이터 세트를 결합한 준감독(semi-supervised) 접근법을 사용하여 샘플 클래스를 예측하도록 요청받는다. 두 번째 제약으로, 시그니처는 40개 이하의 유전자가 포함될 수 있다.

[0092] *트레이닝, 테스트, 및 검증 데이터 세트로서 공개된 데이터*

[0093] 도 8은 혈액 유전자의 발현 데이터의 트레이닝 데이터 세트, 테스트 데이터 세트, 및 검증 데이터 세트를 공개하는 방법을 도시한다. 혈액 샘플 처리 및 유전자의 발현 데이터 생성 후, 독립적인 연구의 데이터는 트레이닝, 테스트 및 검증 데이터 세트로 나뉜다. 트레이닝 데이터 세트로부터의 데이터 및 클래스 라벨은 혈액 기반 유전자 시그니처 분류 모델의 개발 및 교육을 위해 제공된다. 트레이닝된 모델은 혈액 샘플의 클래스 예측을 위한

무작위 테스트 및 검증 유전자의 발현 데이터 세트에 맹목적으로 적용된다.

[0094] 구체적으로, QASMC 임상(도 7b, 데이터 세트 H1) 및 마우스 C57BL/6 흡입(도 7b, 데이터 세트 M1a) 연구로부터 표준화된 유전자의 발현 데이터 및 클래스 라벨이 트레이닝 데이터 세트로서 제공된다. 인간 BLD-SMK-01 및 마우스 ApoE-/*?*- 데이터(도 7b, 데이터 세트 H2 및 M2a 각각)는 테스트 데이터 세트로서 사용된다. REX C-03-EU(도 7b, 데이터 세트 H3) / -04-JP(도 7b, 데이터 세트 H4) 임상 연구 및 마우스 C57BL/6 (도 7b, 데이터 세트 M1b) 및 ApoE-/-(도 7b, 데이터 세트 M2b) 흡입 연구는 검증 데이터 세트로서 공개된다. 테스트 및 검증 세트로부터의 샘플 데이터는 완전히 무작위로 추출되어 클래스 라벨 예측을 위해 순차적으로 공개된 2개의 클래스 균형 서브세트로 분할된다(도 8). 테스트 데이터 세트의 샘플을 사용하여 참가자의 예측을 점수화하고 각 하위 도전에서 팀 수행을 평가한다. 참가자가 흡연자 또는 현재 비흡연자에게 더 가깝다고 샘플을 예측했는지 여부를 평가하는 데 검증 세트가 사용된다. 인간 데이터만, 및 인간과 마우스 데이터는 각각 SC1 및 SC2에 대해 공개된다(도 7b).

[0095] *예측 유전자 시그니처 분류 모델*

[0096] 선택 편향을 피하거나 일반적으로 전체 어레이 기반 유전자 시그니처의 성능에 영향을 미치는 차원의 폐해를 줄이기 위해, 2개의 공개 독립 데이터 세트가 필터링 및 유전자 선택을 안내하는 데 사용된다. 독립적인 연구에서 가장 높은 배수 변화 유전자는, 2개의 연구의 N번째 가장 높은 배수 변화(절대 값)의 교차점에 있는 유전자를 기반으로 선형 판별 모델을(각 N=1에 대해) 평가함으로써 공동으로 사용된다. 최상의 N은 5-배 교차 검증(100회 반복)에 의해 선택되고 11-유전자 시그니처를 이끌어낸다.

[0097] 도전을 위해, 참가자는 다양한 기능 선택 및 기계 학습 방법을 사용하여 차별화된 특징(유전자)을 식별하고 샘플을 분류한다. 랜덤 포레스트(random forest)는, 부분 최소 제곱 판별 분석, 선형 판별 분석(LDA) 및 로지스틱 회귀는 2가지 하위 도전에서 상위 3개의 최선의 성과 팀이 사용한 분류 방법이다. 테스트 및 검증 데이터 세트의 각 샘플에 대해 참가자는 샘플이 클래스 1(예, 흡연자)에 속한 신뢰 값 P (0 내지 1)와, 샘플이 클래스 2에 속하는 신뢰 값(예, 현재 비흡연자)에 해당하는 신뢰 값 1-P를 제공하도록 요청받는다. P 및 1-P는 같지 않도록 요청받는다.

[0098] *성과 평가를 위한 채점*

[0099] 검증 데이터 세트가 아닌 테스트 데이터 세트 내에 있는 샘플은 각 하위 도전에서 팀 실적을 평가하는 데 사용된다. 익명화된 참가자의 클래스 예측은 매튜 상관 계수와 정밀도 재현율 곡선 기준 아래 영역을 사용하여 채점된다. 전반적인 팀 실적은 측정 기준 및 과업(과업 1: 흡연자 대 현재 비흡연자; 과업 2: 이전 흡연자 대 흡연 비경험자)을 통해 연산된 평균 순위에 기초한다. 채점 결과와 최종 순위는, 현장 전문가의 외부 및 독립적인 채점 검토 패널에 의해 검토되고 승인된다. 본 출원의 검증 데이터 세트에서 팀 성과를 평가하기 위해 REX 연구에서 흡연자와 이전 흡연자(Cess) 샘플을 사용하여 동일한 채점 방식이 적용된다.

[0100] *도전 이후 분석*

[0101] 혈액 샘플이 흡연자 또는 3R4F 그룹에 속하는지 여부에 상응하는 신뢰 값은 로그 오즈(odds) (log(P/(1-P)))로 변환된다. 개별적인 상위 3개의 팀(검증 데이터 세트를 사용하여 다시 점수를 매김) 또는 모든 자격을 갖춘 팀의 중간값으로 집계된 로그 오즈는 상자도의 클래스별로 시각화된다. 핵심 비교를 위해 짝(paired)(길이 방향 REX연구에 대해 0일 대 5일) 및 웰치 t-검정(Welch t-test)가 수행하였다(즉, 모든 그룹은 흡연자/3R4F 그룹과 비교되었다). 모든 통계 및 그래픽 시각화는 R 소프트웨어 v3.1.2를 사용하여 수행된다.

[0102] 실시예 1 - 결과

[0103] 본 실시예의 사례 연구는 MRTP 평가와 관련된 시스템 독성학에서의 방법 및 데이터의 독립적 검증 결과를 보고한다. 연구의 일 목적은 흡연 노출 또는 중단 상태를 예측하는 능력을 가진 혈액 기반의 인간 및 종 독립적인 유전자 발현 시그니처 분류 모델의 개발을 위한 계산 방법을 평가하는 것이다(도 7). 참가자는 흡연자/3R4F 및 현재 비흡연자(이전 흡연자/Cess 및 흡연 비경험자/가짜) 데이터 및 프로토타입/후보 MRTP에 노출된 마우스 또는 종래의 CS에 노출된 후, 후보 MRTP로 전환한 인간 및 쥐로부터의 데이터를 포함하는 독립적인 유전자 발현 데이터 세트에 그들의 트레이닝된 모델을 맹목적으로 적용했다. 참가자는 각 샘플에 대해, 샘플이 흡연에 노출되거나 현재 비흡연 노출 그룹에 속하는지 여부에 대한 신뢰 값을 제출한다.

[0104] 인간 흡연 노출 유전자 시그니처 분류 모델을 사용한 흡연자(S) 그룹과 5 일간 중단 및 후보 MRTP 그룹으로 전환한 샘플의 연관성 감소.

[0105] 인간 흡연 노출 반응 유전자 시그니처 분류 모델은 흡연자, 이전 흡연자 및 흡연 비경험자를 포함하는 QASMC 데이터 세트에서 트레이닝된다. 식별된 시그니처는 11 개의 유전자 세트를 포함한다: LRRN3, SASH1, TNFRSF17, DDX43, RGL1, DST, PALLD, CDKN1C, IFI44L, IGJ, 및 LPAR1. 흡연자와 현재 비흡연자를 구별하기 위한 시그니처의 능력을 테스트하기 위해, 모델은 흡연자 그룹에 속한 샘플이 각 샘플에 대해 연산되는 확률로 테스트 데이터 세트(BLD-SMK-01) 및 LDA 점수에 적용된다. 샘플이 흡연자 그룹(P)과 NCS 그룹(1-P)에 속하는 확률은 로그 오즈(log odds) (P/(1-P))로 연산되고 변환되어 흡연자 또는 비 흡연자 그룹과 샘플의 연관을 정량화한다. 그룹/클래스 당 로그 오즈 분포는 상자도(도 9A, 웰치 t-검정 p 값 3*<0.001 대 S 그룹)으로 시각화된다. 흡연자 클래스에 대한 로그 오즈 분포의 중간값은 약 +3.0인 반면, 이전 흡연자 및 흡연 비경험자 클래스의 중간값은 각각 -3.8 및 -5.8이다. 흡연자와 현재 비흡연자의 중간값의 편차가 클수록, 유전자 시그니처 분류 모델의 차별성이 커진다. 상자도는 일측의 흡연자와 타측의 현재 비흡연자로서 정의된 이전 흡연자와 흡연 비경험자 사이의 명확한 분리를 나타낸다(도 9a).

[0106] 동일한 모델 및 절차가 전환(Switch) 또는 세스(Cess) 피험자의 데이터가 흡연자 또는 비현재 흡연자에 더 가깝게 분류되었는지 여부를 결정하기 위해 검증 데이터 세트(REX C-03-EU 및 REX C-04-JP)에 직접 적용된다(도 9a). 특히, 전환 피험자는 후보 MRTP로 전환한 대상이며, 세스 피험자는 5 일 동안의 구금 상태에서 금연을 한 대상이다. 단지 5 일 중단 또는 전환 후에, 이들 그룹과 관련된 로그 오즈는 흡연자 그룹과 비교하여 유의하게 감소하지만, 세스 및 스위치 그룹간에 차이는 발견되지 않았다(도 9a). 0 일 내지 5 일 간 유의한 차이(로그 오즈비)는 흡연 그룹에서 발견되지 않은 반면, 0 일에서 각각의 기준선과 비교하여 세스 및 전환 그룹에서 유의한 감소가 관찰되었다 (도 9b, 짝비교 t검정(Paired t-test) p값 3*<0.001).

[0107] 크라우드 소싱된 데이터 검증은 5 일간의 중단 및 후보 MRTP 그룹으로 전환한 혈액 샘플이 흡연자 그룹에 속한다는 감소된 신뢰도 예측을 확인했다

[0108] 흡연자의 흡연 노출 반응 유전자 시그니처 분류 모델을 트레이닝한 후 참가자들은 무작위 테스트 및 검증 데이터 세트에 모델을 적용하고 흡연자 그룹에 속한 각 피험자의 신뢰 값(확률)을 연산했다. 도전이 종료된 후, 흡연자, 이전 흡연자 및 흡연 비경험자가 아닌 테스트 데이터 세트에 대해 채점이 수행되었다. 참가자의 예측 제출물은 검증 코호트에 대해서만 재채점되고, 팀 225, 264 및 257은 SC1에 대한 상위 3 개 팀으로 식별된다(도 10에 도시된 표). 클래스 예측용 유전자 시그니처 분류 모델의 클래스 예측 성능은 흡연자 및 세스(성과 평가에서 이전 흡연자로서 고려됨) 진 클래스 레이블을 골드 기준(gold standard)으로서 평가되며 AUPR 곡선 값은 상위 3 개 최선의 성과 우수한 팀에서 0.90 이상인 것으로 나타났다.(도 10에 도시된 표)

[0109] 도 11 테스트 및 검증 데이터 세트에 대한 참가자에 의한 인간 및 마우스 혈액 샘플 클래스 예측을 나타낸다. 특히, 참가자는 흡연 노출(S는 인간 3R4F는 마우스) 및 비현재 흡연(NCS) 노출(이전 흡연자 및 FS/Cess 및 흡연 비경험자 NS/Sham) 인간 피험자 및 마우스를 구별하기 위해 인종(도 11a) 및 종 독립적인(도 11b) 혈액 기반의 흡연 노출 유전자 시그니처 모델을 트레이닝했다. 각 샘플에 대해 참가자는 샘플이 S/3R4F 그룹에 속하는 신뢰 값 P와, 샘플이 NCS 그룹에 속하는 신뢰 값 1-P를 제공하도록 요청받는다. 신뢰 값은 로그 오즈(log (P/(1-P)))로 변환되고 모든 12개의 적격 팀에서 각 샘플의 중간값을 연산하여 집계되며 상자도로서 클래스 당 분포로 표시된다(도 11a). 모든 결과는 테스트 데이터 세트에 대해 흡연자와 현재 비흡연자(이전 흡연자 및 흡연 비경험자) 간의 명확한 구별을 나타낸다. 검증 데이터 세트에 대해, 모델을 사용하여 얻은 흡연자 그룹과 5 일간의 Cess 및 스위치 그룹으로부터의 샘플의 감소 된 연관성의 관찰은 개인 또는 집단 참가자의 유사한 결과를 산출한 예측에 의해 분명히 확인되었다 (도 11a). 웰치 t 검정 p 값은 * 0.05, 2 * <0.01, 3 * <0.001 대 S / 3R4F 그룹이다. 이전/비 클래스에 대한 신뢰도 감소는 시그니처 유전자 발현의 변형이 일어나고, 후보 MRTP 로의 전환 또는 중지 5 일 후에 혈액 세포에서 이미 검출 가능하다는 것을 반영한다.

[0110] 크라우드 소싱된 기술 벤치마킹은 인간 및 설치류 종에 관계없이 혈액 샘플 클래스 예측에 대한 최고 성능의 흡연 노출 모델을 식별했다

[0111] SC2의 경우, 참가자들은 인간과 설치류 데이터 모두에 직접적으로 적용될 수 있는 종 예측에 대한 종 독립적인 흡연 노출 반응 유전자 시그니처 모델을 개발하도록 요청받았다. 검증 데이터 세트를 사용하여 참가자들의 예측 제출의 재채점은 SC2에 대한 상위 3 개의 팀(도 10의 표)으로서 팀(219, 250 및 264)을 식별한다. SC1의 경우, 가장 우수한 수행 팀에 의해 또는 모든 팀 값의 집합 후에 얻어진 신뢰 값은 클래스 당 로그 오즈 분포로 시각화된다(도 11b). CS/3R4F에 노출된 코호트와 노출되지 않은(흡연 비경험자/가짜 및 이전의 흡연자/중단) 코호트 사이의 명확한 분리는 인간과 마우스 둘 모두의 상자도에서 관찰할 수 있으며 모델이 종과 관계없이 혈액 샘플을 분류할 수 있음을 나타낸다(도 10,도 11b에 도시된 표). 두 개의 독립적 인 마우스 생체 내 연구의 검증 샘

플에 모델을 맹목적으로 적용 할 경우, 프로토 타입 MRTP (pMRTP) 또는 후보 MRTP에 노출 된 그룹에 해당하는 샘플은 가짜와 비슷한 수준의 로그 오즈 값을 가지며 마우스 및 인간 데이터 세트 (도 11B).

[0112] 도 12는 검증 데이터 세트에 대한 0 일 내지 5 일의 감금 상태에서의 크라우드 로그 오즈비를 나타낸다. 로그 오즈 비율은 세스 및 전환 그룹의 경우 0 일 내지 5 일에 상당한 차이가 있지만 예상대로 흡연자 그룹에서는 상당한 차이가 없었다(짝 비교 t 검정 p 값 3*<0.001).

[0113] 도 13은 그룹/클래스 당 크라우드 로그 오즈 분포 스플릿 및 pMRTP 또는 후보 MRTP에 대한 노출 시간, 또는 pMRTP 또는 후보 MRTP로 전환한 후의 시간을 나타낸다. 특히, 2 개월간의 CS 노출에서 pMRTP로 전환한 후, 시간대에 따라 클래스가 나뉘어질 때 로그 오즈 값의 점진적인 감소가 관찰되며(예: pMRTP에 1, 3 및 4 개월 노출된 것에 해당하는 전환 3, 전환 5 및 전환 7), 이는 시간이 지남에 따라 혈액 세포에서 일어나는 점진적인 유전자 발현 변화의 지표이다.

[0114] 흡연 노출 상태를 예측하는 혈액의 인간 및 종 독립적인 반응 마커는 공통점을 나타내며 팀간에 매우 일관된 핵심 유전자 서브세트를 포함한다

[0115] 흡연 노출 핵심 유전자 서브세트는 적어도 3 개의 팀 및 PMI 시그니처를 통해 적어도 2 개의 동시 발생 유전자를 추출함으로써 식별된다(도 4). 사이클린 의존성 키나아제 억제제 1C(CDKN1C), 류신이 풍부한 반복 뉴런(neuronal) 3((LRRN3) 및 1을 함유하는 SAM 및 SH3도메인(SASH1)은 인간의 시그니처(도 4a)에서 가장 자주 나타나는 유전자이며, 아릴-탄화수소 수용체 리프레저(AHRR), 피리미딘 작용성 수용체 P2Y6(P2RY6)를 코딩하는 유전자는 종 독립적인 시그니처(도 4b)에서 가장 높은 동시 발생을 갖는다. 두 핵심 유전자 서브세트 사이의 비교는 LRRN3, SASH1, AHRR 및 P2RY6 (도 4)를 코딩하는 4 개의 공통 유전자 세트를 나타낸다.

[0116] 실시예 1 - 유전자 시그니처 길이, 유전자 발현의 공동 직선성 수준 및 분류 방법의 상위 6 개 팀의 인간에 근거한 흡연 노출 공감 시그니처 영향의 모든 유전자 조합에 대한 성능 분석

[0117] *방법*

[0118] 공감 시그니처로부터 모든 가능한 유전자의 조합을 고려한다. 이 유전자 분석에 필요한 컴퓨터 집약적 연산의 한계로 인해 18 개 유전자에 기반한 인간의 흡연 노출 공감(consensus) 시그니처는 상위 6 개 팀(12 개 자격을 갖춘 팀 대신)으로 제한된다. DSC2, FSTL1, GPR63, GSE1, GUCY1A3, RGL1, CTTNBP2, F2R, SEMA6B, CDKN1C, CLEC10A, GPR15, LINC00599, P2RY6, PID1, SASH1, AHRR, 및 LRRN3를 포함하는 혈액에서 18 유전자 기반의 공감 시그니처는 상위 6 개 팀의 시그니처를 통해 적어도 2 개의 동시 발생 유전자를 선택함으로써 확인된다. 분류 특성에 미치는 유전자 시그니처 크기 및 공동 직선성 수준의 영향을 조사하였다. 분석은 SC1의 테스트 데이터 세트와 별도로 5 회 교차 검증된 교육(10 회 반복)을 사용하여 수행된다. 도전에서 가장 널리 적용되는 기계 학습(ML) 방법은 랜덤 포레스트(RF), 선형 커널(svmLinear)이 있는 지원 벡터 머신, 부분 최소 판별 분석(PLS), 나이브 베이즈(NB), k-최근접 이웃, 선형 판별 분석(LDA) 및 로지스틱 회귀 분석(LR)을 포함한다. 길이 2 내지 18의 18 개 유전자(즉, 262, 125 유전자 세트)의 가능한 모든 조합이 생성된다. 각 유전자 세트에 7 가지 ML 방법을 적용하면 총 1,834,875 개의 테스트된 분류 전략이 도출된다. 유전자 세트 내의 유전자의 공통 직선성 수준은 해당 유전자 세트로 제한된 발현 매트릭스(matrix)의 제1 주성분의 분산의 백분율로 반영된다. 1,834,875 유전자 세트-ML 예측("Top"이라고 불림)의 성능은 MCC 및 AUPR 점수를 연산하여 평가된다. 이들 "Top"유전자 세트의 성과는 차별적으로 발현된 유전자(DEG, 거짓 발견율, 또는 FDR<=0.5) 또는 또는 HG-U133_Plus_2 칩에 표시된 모든 유전자 중에서 무작위로 선택된 유전자 세트(2-18 유전자)의 성과와 비교된다. 샘플링 과정은 각 유전자 세트 크기에 대해 1,000 번 반복되어 총 17,000 개의 무작위 "DEG"또는 "모든 유전자" 유전자 세트가 생성된다.

[0119] *결과: 상위 6 개 팀의 18 개 유전자 기반 공감 시그니처 유전자 세트 조합은 유익하며 흡연 노출 상태 클래스 예측을 위한 "DEG"및 "모든 유전자"유래 유전자 세트를 능가한다*

[0120] 유전자 시그니처 크기와 공통 직선성 수준이 흡연 노출 상태 클래스 예측의 성능에 미치는 영향은 상위 6 개 팀의 예측에서 18 가지 유전자 기반의 공감 시그니처를 사용하여 조사한다. MCC 및 AUPR 점수는 ML 기반 클래스 예측(도 14 및 15)을 사용하여 길이 2 내지 18의 모든 가능한 서명 조합의 성능을 평가하기 위해 계산된다. 도 14 및 15는 MCC 점수(도 14) 및 AUPR 점수(도 15)에 대한 결과를 나타낸다. 두 그림에서, 패널 A는 교차 검증 및 테스트 데이터 세트에 대한 점수 대 유전자 시그니처 크기를 나타낸다. 특징은 (i) "탑"유전자(즉, 시그니처의 일부로서 참가자에 의해 빈번하게 선택된 유전자;(ii) "DEGs", 차별적으로 발현된 유전자의 목록; (iii) "모든 유전자", 모든 측정된 유전자, 목록으로부터 선택된다. 두 그림 모두에서, 패널 B는 점수 대 시그니처의 유

전자 간 유사성 계수를 나타낸다. 7 가지 기계 학습 분류기가 테스트된다: 랜덤 포레스트(RF), 선형 커널(svmLinear), 부분 최소 판별 분석 (PLS), 나이브 베이즈(NB), k-최근접 이웃(kNN), 선형 판별 분석(LDA) 및 로지스틱 회귀 분석(LR). 두 그림에서, 패널 C는 CV 및 테스트 세트 데이터의 점수 분포와 "Top"(상위), "DEG"(중간) 및 "모든 유전자"(하단) 선택에 대한 차이 분포를 나타낸다.

[0121] 도 14 및 15의 데이터에 의해 표시된 바와 같이, 예측 성과는 유전자 세트 크기에 따라 증가하고 트레이닝 2 가지 트레이닝 모두(교차 검증, CV) (CV의 경우, 크기=2에 대한 MCC = 0.57, 및 크기=18 에 대한 MCC=0.91) 및 테스트 세트(테스트의 경우, 크기=2의 경우 MCC=0.42 및 크기=18의 경우 MCC=0.77)에서 최대 18 개의 유전자를 포함하여 더 긴 세트로 점진적으로 안정화된다(도 14a). 예측 성과는 50% 내지 60% 범위의 "Top" 유전자 세트의 유전자의 공동 직선성 수준(유전자 세트 발현 행렬로부터 연산된 제1 주성분에 의해 대표되는 분산 백분율에 의해 반영됨)이 최대가 될 때까지 도달했고, 그런뒤에 증가된 공동 직선성과 함께 감소하였다(도 14b). "Top" 유전자 세트가 다른 팀의 시그니처 유전자로 구성되어 있고 이미 상당히 다양했기 때문에 어느 정도 일치하는 유전자를 결합하면 예측을 강화할 수 있다. 성과는 DEG로부터의 유전자 세트 내의 유전자의 공통 직선성이 증가함에 따라 감소하였다(도 14b). 일반적으로 "Top", "DEG"및 "All Genes"의 유전자 세트가 각각 최상, 중간 및 최악의 성과를 나타낸다.(도 14). 또한, CV로부터 파생된 성과는 테스트 세트에 대해 연산된 성능보다 우수했다(도 14). 다양한 ML 방법으로 얻어진 성과 기준은 유사한 패턴(도 14b)을 나타내었고, 따라서, 결과의 시각화를 용이하게 하기 위해 집계되었다.(도 14a 및 도 14c). 전반적으로, 결과는 18 유전자 기반의 공감 시그니처에서 얻은 혈액 유전자가 정보를 제공하고 결합되었을 때 흡연 노출 상태에 대한 예측력이 높음을 나타낸다.

[0122] 실시예 1 - 논의

[0123] 이 실시예 연구에서 수득한 결과는 후보 MRTP에 노출된 피험자 또는 기존 CS 노출 후, 후보 MRTP로 전환한 피험자가 흡연 노출 그룹 또는 현재 비흡연 노출 그룹에 속한다고 예측된 신뢰를 제공한다.

[0124] 결과는 명확하게 흡연자와 비흡연자를 분리한다. 참가자들은 인간과 마우스 종에 관계없이 흡연 노출 상태 예측에 매우 우수한 성과를 보이는 종 독립적 혈액 기반 유전자 시그니처 모델을 성공적으로 개발했다. 인간의 테스트 데이터 세트에서, 이전 흡연자 그룹은 흡연 비경험자 그룹과 매우 흡사하지만 흡연자 그룹과 흡연 비경험자 그룹 사이의 중간에 머물러 있었으며, 이는 이전 흡연자의 유전자 시그니처에서 유전자의 발현이 완전히 흡연 비경험자의 발현 수준으로 완전히 되돌아 갈 수 없다는 것을 나타낸다. 변화의 회귀는 피험자마다 다른 흡연 내역 및 종료 시간에 따라 달라질 수 있으며 이 그룹에 대한 예측의 더 높은 변동성을 설명한다. 이전 흡연자의 혈액 세포의 경우, DNA 메틸화 수준(예, F2RL3 유전자)은 팩(pack) 햇수(year)와 절연 후 시간에 따라 달라질 수 있다.

[0125] 마우스 데이터 세트에서, 세스(Cess) 그룹의 발현 수준은 가짜(Sham) 그룹의 수준에 도달하여 더 유전적으로 그리고 실험적으로 균질한 마우스 품종(strain)의 혈액 세포에서 특이적 유전자 발현 변화의 회귀(reversion)를 제안한다. 흥미롭게도, 이 회귀는 시간이 지남에 따라 점차적으로 발생하는데, 이는 그룹이 중단 시간을 기준으로 분할될 때 관찰된다. 이는 유전자 시그니처 분류 접근법이 이진 분류에 유용할 뿐 아니라 변화의 크기와 속도(kinetics)를 따르기 위해 보다 정량적인 방법(예, LDA 점수 또는 관련 신뢰도와 같은 모델 매개 변수의 크기)에서도 사용될 수 있음을 제시한다. 사실, 이것은 흡연자 그룹과 비교하여 흡연 비경험자 그룹의 값에 대하여 감소하는 것을 나타내는 검증 인간 REX 데이터 세트로부터의 전환(Switch) 및 세스(Cess) 그룹의 경우이다. 이 관찰은 흡연 노출 시그니처 유전자에 의해 반영된 분자적 변화가 단지 MRTP 후보로 전환하거나 기존의 담배를 끊은지 5일만에 혈액 세포에서 발생함을 나타낸다. 이러한 결과는 임상적 "하루 감량 담배" 감금 상태 연구에서 1 주일 후에 측정된 노출 반응성 바이오 마커의 감소와 일치한다. 마우스 검증 데이터 세트의 경우, 3R4F 그룹과 프로토타입/후보 MRTP 또는 스위치 그룹(가짜와 유사한 레벨) 간의 로그 오즈의 차이는, 전환 후에 후보 MRTP 또는 pMRTP에 더 오래(수개월) 노출될 때 설명될 수 있고, MRTP의 생물학적 효과가 기존 CS와 비교하여 혈액 세포에 미친 영향을 반영하기 때문에 더 중요하다.

[0126] 혈액 기반의 흡연 노출 반응 분류 모델을 개발하고 트레이닝하는 데 사용되는 계산 방법이 다르더라도, 상위 실적 팀이 획득한 샘플 분류 성과는 높다. 흡연 노출에 의해 유발된 유전자 발현 변화가 인간 또는 인간 및 마우스(종 독립적인 시그니처)의 흡연 노출 상태를 예측할 수 있는 특이적이고 강력한 혈액 시그니처를 구성하는 유전자를 선택하는 데 충분한 정보와 일관성을 갖는다는 것을 나타내는 핵심 유전자 시그니처가 팀간에 일관되게 식별된다.

[0127] 흡연자와 비흡연자로부터의 세포 특이적 백혈구에 대해 보고된 DNA 메틸화 분석과 유사한 혈액 세포 유형 특이적(type-specific) 전사체 분석은 흡연 반응 반응 특성에 대한 각 혈액 세포 유형의 기여도를 보다 잘 이해하는

데 도움이 될 수 있다. 일부 유전자는 특정 혈액 세포 아집단과 관련될 수 있다. 전반적으로 핵심 시그니처의 일부인 이러한 흡연 노출 관련 유전자는 기존 담배와 비교하여 후보 MRTP와 같은 신제품의 영향을 모니터링하고 가능하면 정량화할 수 있는 강력한 혈액 마커 세트를 구성한다.

[0128] 실시예 1과 관련하여 설명한 연구는 대중의 힘을 활용하여 시스템 방법을 평가하고 시스템 독성학에서 데이터를 검증하는 방법을 나타낸다. 고전적 동등 심의 프로세스(peer review process)를 보완하는 것 외에도, 제품 위험 평가 데이터에 대한 독립적이고 편견없는 평가를 통해 과학적 결론을 확인하고 신뢰를 제공하는데 사용될 수 있고 의사 결정을 위한 규제 기관을 지원할 수 있다. 본원에 기재된 실시예는 개개인의 흡연자 상태 예측용 확고한 유전자 시그니처를 확인하기 위해 크라우드 소싱 접근법을 주로 사용하는 것에 관한 것이지만, 당업자라면 본 개시의 시스템 및 방법을 질병 상태, 생리학적 상태, 노출 상태, 또는 개인의 생물학적 상태와 관련된 개인의 다른 적절한 상태 또는 상태를 포함하는 개인의 생물학적 상태 예측용 유전자 시그니처를 포함할 수 있다.

[0129] 하기 표 2는 실시예 1에 따라 수행된 연구 결과를 포함한다. 특히, 표 2에 제시된 결과는 인간의 흡연 시그니처에서 추출되었으며 제1 열에 유전자 세트가 나열된다. 제2 열에는 시그니처에 해당 유전자가 포함된 팀 또는 참가자의 수(12 개 중)가 나열된다. 제3 열에는 시그니처에 해당 유전자가 포함된 상위 3개 팀 수(테스트 데이터 세트에 따라 평가됨)가 나열된다. 제4 열에는 시그니처에 해당 유전자가 포함된 상위 3 개 팀 수(검증 데이터 세트에 따라 평가됨)가 나열된다. 제5 열에는 제3 및 제4 열의 값의 평균이 나열된다.

표 2

[0130]

테스트 세트 채점	합계 (12 개 팀 중)	상위 3 개의 테스트 세트 합계	상위 3 개의 검증 세트 합계	테스트+검증의 평균
LRRN3	9	3	3	3
AHRR	9	3	3	3
CDKN1C	9	3	3	3
PID1	8	3	3	3
SASH1	7	3	3	3
GPR15	7	3	3	3
P2RY6	6	3	3	3
LINC00599	6	2	3	2.5
CLEC10A	6	3	2	2.5
SEMA6B	5	2	3	2.5
F2R	5	2	2	2
DSC2	5	1	0	0.5
TLR5	5	0	1	0.5
RGL1	4	1	2	1.5
FSTL1	4	1	0	0.5
VSIG4	4	0	0	0
AK8	4	0	0	0
CTTNBP2	3	2	2	2
GUCY1A3	3	1	1	1
GSE1	3	1	0	0.5
MIR4697HG	3	0	0	0
PTGFRN	3	0	0	0
LOC200772	3	0	0	0
FANK1	3	0	0	0
C15orf54	3	0	0	0
MARC2	3	0	0	0
GPR63	2	2	1	1.5
TPPP3	2	1	1	1
ZNF618	2	1	1	1
PTGFR	2	1	0	0.5
GUCY1B3	2	0	1	0.5
P2RY1	2	0	0	0
TMEM163	2	0	0	0
ST6GALNAC1	2	0	0	0
SH2D1B	2	0	0	0
CYP4F22	2	0	0	0

PF4	2	0	0	0
FUCA1	2	0	0	0
MB21D2	2	0	0	0
NLK	2	0	0	0
B3GALT2	2	0	0	0
ASGR2	2	0	0	0
NR4A1	2	0	0	0
RTN1	1	1	1	1
MAFB	1	1	1	1
ARHGEF10L	1	1	1	1
CLDN23	1	1	1	1
TGFBI	1	1	1	1
LOC284837	1	1	1	1
SYCE1L	1	1	1	1
SEZ6L	1	1	1	1
KLF4	1	1	1	1
NOD1	1	1	1	1
FAM225A	1	1	1	1
CRACR2B	1	1	0	0.5

[0131] 일부 구현예에서, 흡연 노출 반응 상태를 결정하기 위해 사용되는 유전자 시그니처는 표 2에 나열된 유전자를 포함하며, 이는 상위 3 개 수행 유전자 시그니처 중 2 개 이상에 나타나는 유전자에 해당한다. 테스트 데이터 세트(예, 표 2의 제3 열에 도시됨)에 따라 평가한 경우 LRRN3, AHRR, CDKN1C, PID1, SASH1, GPR15, P2RY6, LINC00599, CLEC10A, SEMA6B, F2R, CTTNBP2 및 GPR63이 포함된다. 테스트 데이터 세트(예, 표 2의 제4 열에 도시됨)에 따라 평가한 경우 LRRN3, AHRR, CDKN1C, PID1, SASH1, GPR15, P2RY6, LINC00599, CLEC10A, SEMA6B, F2R, RGL1 및 CTTNBP2가 포함된다. 테스트 및 검증 데이터 세트 간의 평균에 따라 평가한 경우(예, 표 2의 제5 열에 표시)에는 LRRN3, AHRR, CDKN1C, PID1, SASH1, GPR15, P2RY6, LINC00599, CLEC10A, SEMA6B, F2R, 및 CTTNBP2가 포함된다. 일부 구현예에서, 흡연 노출 반응 상태를 결정하기 위해 사용된 유전자 시그니처는 표 2에 나열된 유전자를 포함하며, 이는 12 개 후보 유전자 시그니처 중 적어도 M 개에서 나타나는 유전자에 해당하며, 여기서 M은 1, 2, 3, 4, 5, 6, 7, 8, 또는 9이다. 예를 들어, M이 9인 경우 유전자 시그니처는 제2 열에 9 이상의 값을 갖는 유전자, 즉: LRRN3, AHRR, 및 CDKN1C이 포함된다. 다른 실시예로서, M이 8인 경우, 유전자 시그니처는 제2 열에 8 이상의 값을 갖는 유전자, 즉: LRRN3, AHRR, CDKN1C, 및 PID1이 포함된다. 다른 실시예로서, M이 7인 경우, 유전자 시그니처는 제2 열에 7 이상의 값을 갖는 유전자, 즉: LRRN3, AHRR, CDKN1C, PID1, SASH1, 및 GPR15이 포함된다. 다른 실시예로서, M이 6인 경우, 유전자 시그니처는 제2 열에 6 이상의 값을 갖는 유전자, 즉: LRRN3, AHRR, CDKN1C, PID1, SASH1, GPR15, P2RY6, LINC00599, 및 CLEC10A이 포함된다. 다른 실시예로서, M이 5인 경우, 유전자 시그니처는 제2 열에 5 이상의 값을 갖는 유전자, 즉: LRRN3, AHRR, CDKN1C, PID1, SASH1, GPR15, P2RY6, LINC00599, CLEC10A, SEMA6B, F2R, DSC2, 및 TLR5이 포함된다. 다른 실시예로서, M이 4인 경우, 유전자 시그니처는 제2 열에 4 이상의 값을 갖는 유전자, 즉: LRRN3, AHRR, CDKN1C, PID1, SASH1, GPR15, P2RY6, LINC00599, CLEC10A, SEMA6B, F2R, DSC2, TLR5, RGL1, FSTL1, VSIG4, 및 AK8이 포함된다. 다른 실시예로서, M이 3인 경우, 유전자 시그니처는 제2 열에 3 이상의 값을 갖는 유전자, 즉: LRRN3, AHRR, CDKN1C, PID1, SASH1, GPR15, P2RY6, LINC00599, CLEC10A, SEMA6B, F2R, DSC2, TLR5, RGL1, FSTL1, VSIG4, AK8, CTTNBP2, GUCY1A3, GSE1, MIR4697HG, PTGFRN, LOC200772, FANK1, C15orf54, 및 MARC2이 포함된다. 다른 실시예로서, M이 2인 경우, 유전자 시그니처는 제2 열에 2 이상의 값을 갖는 유전자, 즉: LRRN3, AHRR, CDKN1C, PID1, SASH1, GPR15, P2RY6, LINC00599, CLEC10A, SEMA6B, F2R, DSC2, TLR5, RGL1, FSTL1, VSIG4, AK8, CTTNBP2, GUCY1A3, GSE1, MIR4697HG, PTGFRN, LOC200772, FANK1, C15orf54, MARC2, GPR63, TPPP3, ZNF618, PTGFR, GUCY1B3, P2RY1, TMEM163, ST6GALNAC1, SH2D1B, CYP4F22, PF4, FUCA1, MB21D2, NLK, B3GALT2, ASGR2, 및NR4A1이 포함된다. 또 다른 실시예로서, M이 1인 경우, 유전자 시그니처는 상기 표 2에 나열된 모든 유전자를 포함한다.

[0132] 하기 표 3은 실시예 1에 따라 수행된 연구 결과를 포함한다. 특히, 표 2에 제시된 결과는 종 독립적인 흡연 시그니처에서 추출한 것이며 제1 열에 유전자 세트가 나열된다. 제2 열에는 시그니처에 해당 유전자가 포함된 팀 또는 참가자의 수(12 개 중)가 나열된다. 제3 열에는 시그니처에 해당 유전자가 포함된 상위 3개 팀 수(테스트 데이터 세트에 따라 평가됨)가 나열된다. 제4 열에는 시그니처에 해당 유전자가 포함된 상위 3 개 팀 수(검증

데이터 세트에 따라 평가됨)가 나열된다. 제5 열에는 제3 및 제4 열의 값의 평균이 나열된다.

표 3

[0133]

테스트 세트 채점	합계 (12 개 팀 중)	상위 3개의 테스트 세트 합계	상위 3 개의 검증 세트 합계	테스트+검증의 평균
AHRR	5	3	3	3
P2RY6	4	3	3	3
COX6B2	2	2	2	2
DSC2	2	2	2	2
KLRG1	3	2	2	2
LRRN3	3	2	2	2
SASH1	2	2	2	2
TBX21	2	2	2	2
ADORA3	1	1	1	1
AF529169	1	1	1	1
AKAP5	1	1	1	1
ASGR2	1	1	1	1
B3GALT2	1	1	1	1
BCL3	1	1	1	1
BIRC2	1	1	1	1
CCR4	1	1	1	1
CDKN1C	1	1	1	1
CLEC10A	1	1	1	1
CLEC5A	1	1	1	1
CNNM1	1	1	1	1
COL6A3	1	1	1	1
COX6C	1	1	1	1
CRACR2B	1	1	1	1
CTNNAL1	1	1	1	1
CTTNBP2	2	1	1	1
DCAF8	1	1	1	1
EIF5A2	1	1	1	1
ELOVL7	1	1	1	1
ENDOU	1	1	1	1
ERI1	1	1	1	1
ESAM	1	1	1	1
EVA1B	1	1	1	1
F2R	2	1	1	1
FANK1	1	1	1	1
FKRP	1	1	1	1
FSTL1	1	1	1	1
GGT7	1	1	1	1
GLCCI1	1	1	1	1
GNAZ	1	1	1	1
GNPDA2	1	1	1	1
GP1BA	1	1	1	1
GPR63	1	1	1	1
GSE1	1	1	1	1
GUCY1B3	2	1	1	1
HES1	1	1	1	1
HPGD	1	1	1	1
HSPB6	1	1	1	1
IRF7	1	1	1	1
JARID2	1	1	1	1
KCNQ1OT1	1	1	1	1
KISS1R	1	1	1	1
LIMS1	1	1	1	1
LRRK1	1	1	1	1

LTBP1	1	1	1	1
MBTD1	1	1	1	1
MCEMP1	1	1	1	1
MKNK1	1	1	1	1
MPP2	1	1	1	1
MRAS	1	1	1	1
MT2	2	1	1	1
NDUFA3	1	1	1	1
NGFRAP1	2	1	1	1
NR4A1	1	1	1	1
PF4	1	1	1	1
PGRMC1	1	1	1	1
PHACTR3	1	1	1	1
PID1	1	1	1	1
PTGFR	1	1	1	1
R3HDM4	1	1	1	1
RBM43	1	1	1	1
REEP6	2	1	1	1
REXO2	1	1	1	1
RUNDC3A	1	1	1	1
SAMD11	1	1	1	1
SDR16C5	1	1	1	1
SIAH1A	1	1	1	1
SLPI	1	1	1	1
SPINK2	1	1	1	1
STAR	1	1	1	1
SYTL4	1	1	1	1
TCEAL8	1	1	1	1
TLR2	1	1	1	1
TMEM163	1	1	1	1
TRIB3	1	1	1	1
UBE2B	1	1	1	1
VCAN	1	1	1	1
VSIG4	1	1	1	1
WDFY1	1	1	1	1
ZFP704	1	1	1	1

[0134] 일부 구현예에서, 흡연 노출 반응 상태를 결정하기 위해 사용되는 유전자 시그니처는 표 3에 나열된 유전자를 포함하며, 이는 상위 3 개 수행 유전자 시그니처 중 2 가지 이상에 나타나는 유전자에 해당한다. 표 3에 도시된 바와 같이, 이것이 테스트 데이터 세트 (예: 표 3의 제3 열에 표시), 검증 데이터 세트 (예: 표 3의 제4 열에 표시)에 따라 평가되는지 여부에 관계없이 테스트 데이터와 검증 데이터 사이의 평균값 (예: 표 3의 제5 열에 표시)에는 AHRR, P2RY6, COX6B2, DSC2, KLRG1, LRRN3, SASH1 및 TBX21이 포함된다. 일부 구현예에서, 흡연 노출 반응 상태를 결정하기 위해 사용되는 유전자 시그니처는 표 3에 열거된 유전자를 포함하며, 12 개의 제출된 유전자 시그니처 중 M 개 이상(M은 1, 2, 3, 4 또는 5임)에 나타나는 유전자에 해당한다. 예를 들어, M이 5일 때, 유전자 시그니처는 제2 열에서 5 이상의 값을 갖는 유전자를 포함한다. 즉: AHRR. 다른 실시예로서, M이 4일 때, 유전자 시그니처는 제2 열에서 4 이상의 값을 갖는 유전자를 포함한다. 즉: AHRR 및 P2RY6. 다른 실시예로서, M이 3일 때, 유전자 시그니처는 제2 열에서 3 이상의 값을 갖는 유전자를 포함한다. 즉: AHRR, P2RY6, KLRG1, 및 LRRN3. 다른 실시예로서, M이 2 일 때, 유전자 시그니처는 제2 열에서 2 이상의 값을 갖는 유전자를 포함한다. 즉: AHRR, P2RY6, KLRG1, LRRN3, COX6B2, DSC2, SASH1, TBX21, CTTNBP2, F2R, GUCY1B3, MT2, NGFRAP1, 및 REEP6. 또 다른 실시예로서, M이 1인 경우, 유전자 시그니처는 표 3에 나열된 모든 유전자를 포함한다.

[0135] 일부 구현예에서, 본원에 기재된 유전자 시그니처는 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40 또는 전체 유전자에 있는 유전자의 수 미만의 임의의 적합한 수를 갖도록 제한된다. 여기에 기술된 유전자 시그니처는 전체 유전자에 비해 상대적으로 적은 수의 유전자로 제한된다. 더 긴 유전자 시그니처가 트레이닝 데이터 세트에

과하게 적합하다면, 더 긴 유전자 시그니처는 짧은 유전자 시그니처보다 악화될 수 있다. 이 경우 더 긴 유전자 시그니처는 학습 데이터 세트의 임의의 오류 또는 노이즈를 나타낼 수 있다. 테스트 데이터 세트의 클래스를 예측하는 데 사용되는 경우, 더 짧은 유전자 시그니처가 초과된 긴 유전자 시그니처를 능가할 수 있다. 표 2 및 3과 관련하여 기술된 유전자 시그니처을 포함하여, 본원에 기술된 임의의 유전자 시그니처는 특정 최대 유전자 수를 갖는 것으로 제한될 수 있다.

[0136] 도 5는 본 개시의 예시적인 실시예에 따라, 환자로부터 수득한 샘플을 평가하기 위한 프로세스(500)의 흐름도이다. 프로세스(500)는 LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2 및 GPR63에 대한 정량적 발현 데이터를 포함하는 샘플과 관련된 데이터 세트를 수신하는 단계(단계 502), 수신된 데이터 세트에 기초하여 점수를 생성하며, 점수는 피험자의 예측된 흡연 상태를 나타낸다(단계 504). 일부 구현예에서, 단계(502)에서 수신된 데이터 세트는 다음의 임의의 수에 대한 정량적 발현 데이터를 더 포함한다: DSC2, TLR5, RGL1, FSTL1, VSIG4, AK8, GUCY1A3, GSE1, MIR4697HG, PTGFRN, LOC200772, FANK1, C15orf54, MARC2, TPPP3, ZNF618, PTGFR, P2RY1, TMEM163, ST6GALNAC1, SH2D1B, CYP4F22, PF4, FUCA1, MB21D2, NLK, B3GALT2, ASGR2, NR4A1, 및 GUCY1B3. 일부 구현예에서, 단계(502)에서 수신된 데이터 세트는 표 2 및 표 3과 관련하여 기술된 임의의 유전자 시그니처 또는 본원에 기술된 임의의 다른 유전자 시그니처에 대한 정량적 발현 데이터를 더 포함한다.

[0137] 단계(504)에서 생성된 점수는 데이터 세트에 적용된 분류 체계의 결과이며, 분류 체계는 데이터 세트의 정량적 발현 데이터에 기초하여 결정된다. 특히, 본 명세서에 기술된 예에서, 기계 학습 기술을 사용하여 트레이닝 된 분류자는 502에서 수신된 데이터 세트에 적용되어 개인에 대한 예측된 분류를 결정할 수 있다.

[0138] 본원에 기재된 유전자 시그니처는 대상으로부터 수득된 샘플을 평가하기 위한 컴퓨터 실행 방법에 사용될 수 있다. 특히, 샘플과 관련된 데이터 세트가 수득될 수 있고, 데이터 세트는 핵심 유전자 시그니처에 대한 정량적 발현 데이터(LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2 및 GPR63)를 포함할 수 있다. 일반적으로, 표 2 및 3과 관련하여 기술된 유전자 시그니처 중 어느 것이 핵심 유전자 시그니처로 사용될 수 있다. 핵심 유전자 시그니처는 전체 유전자에서 유전자의 수보다 적은 수의 유전자를 포함하며 전체적으로 함께 고려할 때 흡연 상태와 같은 생물학적 상태를 예측하는 데 유익한 유전자 세트를 포함한다. 적어도 하나의 하드웨어 프로세서는 수신된 데이터 세트에 기초하여 점수를 발생시키고, 점수는 피험자의 예측된 흡연 상태를 나타낸다. 특히, 점수는 본원에 기술된 크라우드 소싱 접근법을 사용하여 구축된 분류기에 기초할 수 있다. 데이터 세트는 확장된 유전자 시그니처에 포함될 수 있는 추가의 마커(DSC2, TLR5, RGL1, FSTL1, VSIG4, AK8, GUCY1A3, GSE1, MIR4697HG, PTGFRN, LOC200772, FANK1, C15orf54, MARC2, TPPP3, ZNF618, PTGFR, P2RY1, TMEM163, ST6GALNAC1, SH2D1B, CYP4F22, PF4, FUCA1, MB21D2, NLK, B3GALT2, ASGR2, NR4A1, 및 GUCY1B3)의 임의의 적합한 조합에 대한 정량적 발현 데이터를 더 포함할 수 있다. 데이터 세트는 위의 표 2 및 3과 관련하여 기술된 임의의 유전자 시그니처에 대한 정량적 발현 데이터를 더 포함할 수 있다.

[0139] 일부 구현예에서, 데이터 세트는 마커 세트 LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2 및 GPR63의 임의의 수를 포함한다. 상기 부분 집합은 이들 확인 된 유전자들 모두를 포함하지 않을 수 있다. 핵심 세트 내에 있는 마커의 적어도 3 개(또는 4, 5, 6, 7, 8, 9, 10, 11 또는 12와 같은 임의의 다른 적절한 수)를 포함하는 것과 같은 하나 이상의 기준이 시그니처에 포함되도록 마커에 적용될 수 있다. 핵심 세트: LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2 및 GPR63, 및 표 2 또는 표3과 관련하여 기술된 유전자 시그니처의 마커 중 임의의 하나의 적어도 2종(예컨대 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 또는 12와 같은 임의의 적절한 수) 전술한 바와 같이, 일부 구현예에서, 시그니처는 전체 게놈에서 유전자의 수보다 적은 수의 유전자로 제한되고, 최대 유전자 수가 예컨대 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 또는 전체 유전자에서 유전자의 수보다 적은 임의의 수로 제한될 수 있다. 일반적으로, 이들 마커의 조합을 사용하는 임의의 시그니처는 본 개시의 범위를 벗어나지 않고, 흡연 상태와 같은 대상의 생물학적 상태를 예측하는데 사용될 수 있다.

[0140] 일부 구현예에서, 본원에 기술된 특성의 유전자는 개체의 흡연자 상태 예측용 키트를 조립하는데 사용된다. 특히, 키트에는 테스트 샘플의 유전자 시그니처에서 유전자의 발현 수준을 검출하는 시약 세트와 개인의 흡연자 상태 예측용 키트 사용 지침이 포함된다. 이 키트는 HTP와 같은 개인의 흡연 제품에 대한 중단 또는 대안의 효과를 평가하는 데 사용될 수 있다.

[0141] 도 2는, 도 1 및 2와 관련하여 기술된 프로세스들과 같이 본원에 기술된 프로세스들 중 임의의 프로세스를 수행하거나 핵심 유전자 시그니처, 연장된 유전자 시그니처, 또는 본원에 기술된 임의의 기타 유전자 시그니처를 저

장하기 위해 사용될 수 있다. 특히, 컴퓨터 판독 가능 매체에 저장된 유전자 시그니처는 LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2 및 GPR63에 대한 발현 데이터를 포함한다. 또 다른 실시예에서, 컴퓨터 판독 가능 매체는 (a)~(d) 중 어느 하나의 항체로 이루어진 군으로부터 선택된 적어도 4, 5, 6, 7, 8, 9, 10, 11 또는 12 마커에 대한 발현 데이터를 포함하는 유전자 시그니처를 포함한다. LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2, 및 GPR63. 또 다른 실시예에서, 컴퓨터 판독 가능 매체는 본원에 기술된 임의의 유전자 시그니처 또는 마커 세트에 관련된 데이터를 포함한다.

[0142] 특정 구현예에서, 컴포넌트 및 데이터베이스는 여러 컴퓨팅 장치(200)에 걸쳐 구현될 수 있다. 컴퓨팅 장치(200)는 적어도 하나의 통신 인터페이스 유닛, 입/출력 제어기(210), 시스템 메모리 및 하나 이상의 데이터 저장 장치를 포함한다. 시스템 메모리는 적어도 하나의 랜덤 액세스 메모리(RAM (202)) 및 적어도 하나의 판독 전용 메모리(ROM (204))를 포함한다. 이들 요소 모두는 중앙 처리 장치(CPU(206))와 통신하여 컴퓨팅 장치(200)의 작동을 용이하게 한다. 컴퓨팅 장치(200)는 많은 다른 방식으로 구성될 수 있다. 예를 들어, 컴퓨팅 장치(200)는 종래의 독립형 컴퓨터일 수 있거나 대안적으로, 컴퓨팅 장치(200)의 기능은 다수의 컴퓨터 시스템 및 아키텍처에 걸쳐 분산될 수 있다. 컴퓨팅 장치(200)는 모델링, 채점 및 집합 동작 중 일부 또는 전부를 수행하도록 구성될 수 있다. 도 2에서, 컴퓨팅 장치(200)는 네트워크 또는 로컬 네트워크를 통해 기타 서버 또는 시스템에 링크된다.

[0143] 컴퓨팅 장치(200)는 분산 아키텍처로 구성될 수 있으며, 데이터베이스 및 프로세서는 별개의 유닛 또는 위치에 하우징된다. 이러한 일부 유닛은 1차 처리 기능을 수행하고 최소한 일반 제어기 또는 프로세서 및 시스템 메모리를 포함한다. 그러한 양태에서, 이들 유닛 각각은 통신 인터페이스 유닛(208)을 통해 다른 서버, 클라이언트 또는 사용자 컴퓨터 및 다른 관련 장치와의 주요 통신 링크로서 기능하는 통신 허브 또는 포트(도시되지 않음)에 부착된다. 통신 허브 또는 포트는 처리 기능 자체가 최소일 수 있으며 주로 통신 라우터로 사용된다. 다양한 통신 프로토콜은 시스템의 일부일 수 있되, 이더넷, SAP, SAS ™, ATP, BLUETOOTH ™, GSM 및 TCP/IP에 한정되지 않는다.

[0144] CPU(206)는 하나 이상의 종래의 마이크로 프로세서와 같은 프로세서 및 CPU(206)로부터 작업 부하를 오프로딩하기 위한 수학 협업-프로세서와 같은 하나 이상의 보조 협업-프로세서를 포함한다. CPU(206)는 통신 인터페이스 유닛(208) 및 입/출력 제어기(210)와 통신하며, 이 인터페이스를 통해 CPU(206)는 다른 서버, 사용자 단말 또는 장치와 같은 다른 장치와 통신한다. 통신 인터페이스 유닛(208) 및 입/출력 제어기(210)는 예를 들어 다른 프로세서, 서버 또는 클라이언트 단말과 동시에 통신하기위한 다수의 통신 채널을 포함할 수 있다. 서로 통신하는 장치는 서로 지속적으로 서로에게 전송할 필요는 없다. 반대로, 그러한 장치는 필요에 따라 서로에게만 전송할 필요가 있으며, 실제로 대부분의 시간 동안 데이터를 교환하지 못하도록 하고, 장치들간의 통신 링크를 설정하기 위해 여러 단계를 수행할 필요가 있을 수 있다.

[0145] CPU(206)는 또한 데이터 저장 장치와 통신한다. 데이터 저장 장치는 자기, 광학 또는 반도체 메모리의 적절한 조합을 포함할 수 있으며, 예를 들어 RAM (202), ROM (204), 플래시 드라이브, 컴팩트 디스크 또는 하드 디스크 또는 드라이브와 같은 광학 디스크를 포함할 수 있다. CPU(206) 및 데이터 저장 장치는 각각 예를 들어 단일 컴퓨터 또는 다른 컴퓨팅 장치 내에 완전히 위치할 수 있으며; USB 포트, 직렬 포트 케이블, 동축 케이블, 이더넷 유형 케이블, 전화선, 무선 주파수 송수신기 또는 다른 유사한 무선 또는 유선 매체 또는 이들의 조합과 같은 통신 매체에 의해 서로 접속될 수 있다. 예를 들어, CPU(206)는 통신 인터페이스 유닛(208)을 통해 데이터 저장 장치에 접속될 수 있다. CPU(206)는 하나 이상의 특정 처리 기능을 수행하도록 구성될 수 있다.

[0146] (예, 컴퓨터 프로그램 코드 또는 컴퓨터 프로그램 제품)데이터 저장 장치는 예를 들어, (i) 컴퓨팅 장치(200)용 운영 체제(212); (ii) 본원에 기술된 시스템 및 방법에 따라 CPU(206)를 지시하도록 적응된 하나 이상의 애플리케이션(214) (예를 들어, 컴퓨터 프로그램 코드 또는 컴퓨터 프로그램 제품)을 포함하며, 특히 CPU(206); 또는 (iii) 프로그램에 의해 요구되는 정보를 저장하는데 이용될 수 있는 정보를 저장하도록 구성된 데이터베이스(들)(216)를 포함할 수 있다. 일부 양태에서, 데이터베이스(들)는 실험 데이터를 저장하는 데이터베이스 및 공개된 문헌 모델을 포함한다.

[0147] 운영 체제(212) 및 애플리케이션들(214)은 예를 들어 압축된, 비 컴파일된 및 암호화된 포맷으로 저장될 수 있으며, 컴퓨터 프로그램 코드를 포함할 수 있다. 프로그램의 명령어는 ROM(204) 또는 RAM(202)과 같은 데이터 저장 장치 이외의 컴퓨터 판독 가능 매체로부터 프로세서의 주 메모리로 판독될 수 있다. 프로그램 내의 명령들의 시퀀스의 실행은 CPU(206)로 하여금 본 명세서에서 기술된 프로세스 단계들을 수행하게 하지만, 하드 - 와이어

드 회로는 본 개시의 프로세스의 구현을 위한 소프트웨어 명령 대신에 또는 소프트웨어 명령과 함께 사용될 수 있다. 따라서, 기술된 시스템 및 방법은 하드웨어 및 소프트웨어의 특정 조합으로 제한되지 않는다.

[0148] 적합한 컴퓨터 프로그램 코드는 여기에 기술된 바와 같은 하나 이상의 기능을 수행하기 위해 제공될 수 있다. (예, 비디오 디스플레이, 키보드, 컴퓨터 마우스 등)프로그램은 또한 프로세서가 컴퓨터 주변 장치(예를 들어, 비디오 디스플레이, 키보드, 컴퓨터 마우스 등)와 인터페이스 할 수 있게 하는 운영 시스템(212), 데이터베이스 관리 시스템 및 "장치 드라이버"와 같은 프로그램 요소를 포함할 수 있다. 입/출력 제어기(210)를 통해 수신된다.

[0149] 본 명세서에서 사용되는 "컴퓨터 판독 가능 매체"라는 용어는 실행을 위해 컴퓨팅 장치(200)(또는 본 명세서에 기술된 장치의 임의의 다른 프로세서)의 프로세서에 명령을 제공하거나 제공하는데 참여하는 임의의 비 일시적인 매체를 지칭한다. 그러한 매체는 비 휘발성 매체 및 휘발성 매체를 포함하지만 이에 한정되지 않는 많은 형태를 취할 수 있다. 비 휘발성 매체는 예를 들어, 광학, 자기 또는 광 자기 디스크, 또는 플래시 메모리와 같은 집적 회로 메모리를 포함한다. 휘발성 매체는 일반적으로 주 메모리를 구성하는 동적 랜덤 액세스 메모리(DRAM)를 포함한다. 컴퓨터 판독 가능 매체의 일반적인 형태는 예를 들어 플로피 디스크, 플렉시블 디스크, 하드 디스크, 자기 테이프, 임의의 다른 자기 매체, CD-ROM, DVD, 임의의 다른 광학 매체, 펀치 카드, 페이퍼 테이프, RAM, PROM, EPROM 또는 EEPROM(전기적으로 지워질 수 있는 프로그램가능한 판독 전용 메모리), FLASH-EEPROM, 임의의 다른 메모리 칩 또는 카트리지, 또는 그 밖의 임의의 컴퓨터가 판독 가능할 수 있는 비일시적인 매체를 포함할 수 있다.

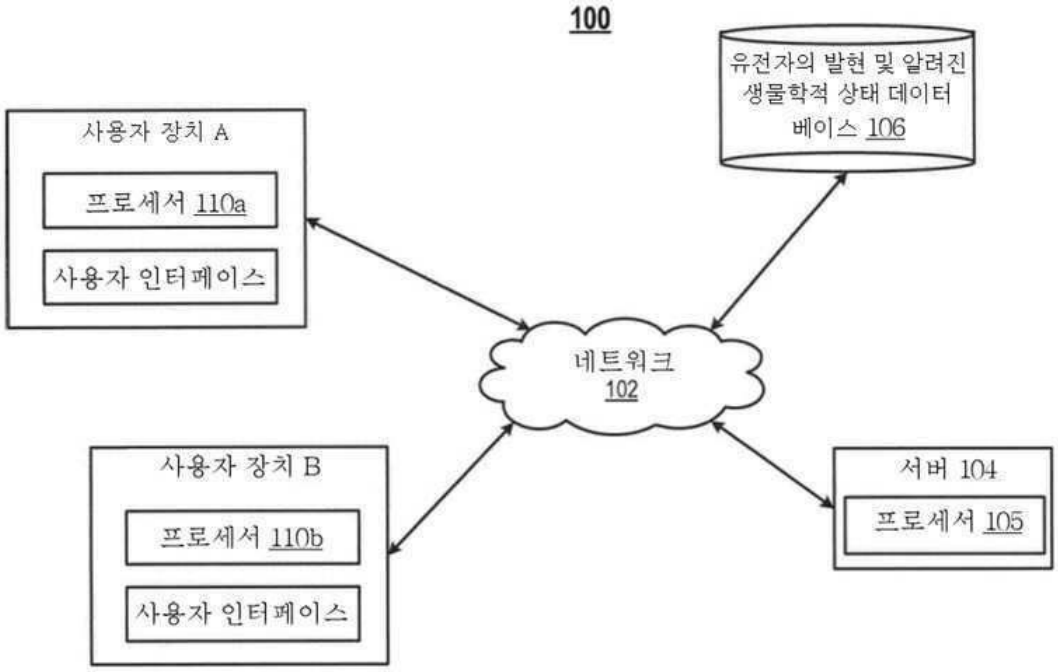
[0150] 컴퓨터 판독 가능 매체의 다양한 형태는 실행을 위해 하나 이상의 명령의 하나 이상의 시퀀스를 CPU(206)(또는 본원에 기술된 장치의 임의의 다른 프로세서)로 운반하는데 포함될 수 있다. 예를 들어, 명령어들은 초기에 원격 컴퓨터(미도시)의 자기 디스크 상에 포함될 수 있다. 원격 컴퓨터는 명령어를 동적 메모리에 로드하고 모뎀을 사용하여 이더넷 연결, 케이블 회선 또는 전화선을 통해 지시를 전송할 수 있다. 컴퓨팅 장치(200)(예, 서버)에 로컬인 통신 장치는 각각의 통신 회선상에서 데이터를 수신하고 프로세서에 대한 시스템 버스 상에 데이터를 배치할 수 있다. 시스템 버스는 데이터를 주 메모리로 전달하며, 프로세서는 이를 통해 명령어를 검색하고 실행한다. 주 메모리에 의해 수신된 명령은 선택적으로 프로세서에 의한 실행 전후에 메모리에 저장될 수 있다. 또한, 지시들은 통신 포트를 통해 다양한 형태의 정보를 운반하는 무선 통신 또는 데이터 스트림의 예시적인 형태인 전기, 전자기 또는 광학 신호로서 수신될 수 있다.

[0151] 본원에서 언급된 각각의 참조는 그 전체가 본원에 참조로서 통합된다.

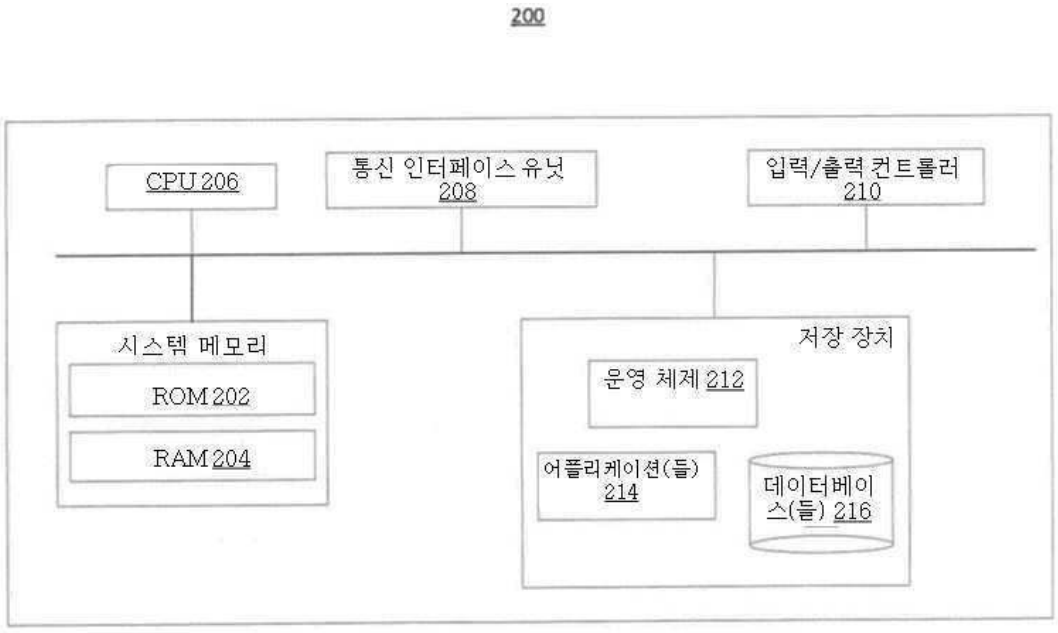
[0152] 본 개시의 구현예가 특정 실시예를 참조하여 구체적으로 도시되고 기술되었지만, 당업자는 첨부된 청구범위에 의해 정의된 바와 같이 본 개시의 범위를 벗어나지 않고 형태 및 세부 사항에서 다양한 변경이 이루어질 수 있음을 이해해야한다. 따라서, 개시된 범위는 첨부된 청구범위에 의해 표시되고, 청구범위의 등가물의 의미 및 범위 내에 있는 모든 변경은 그러므로 받아들여지도록 의도된다.

도면

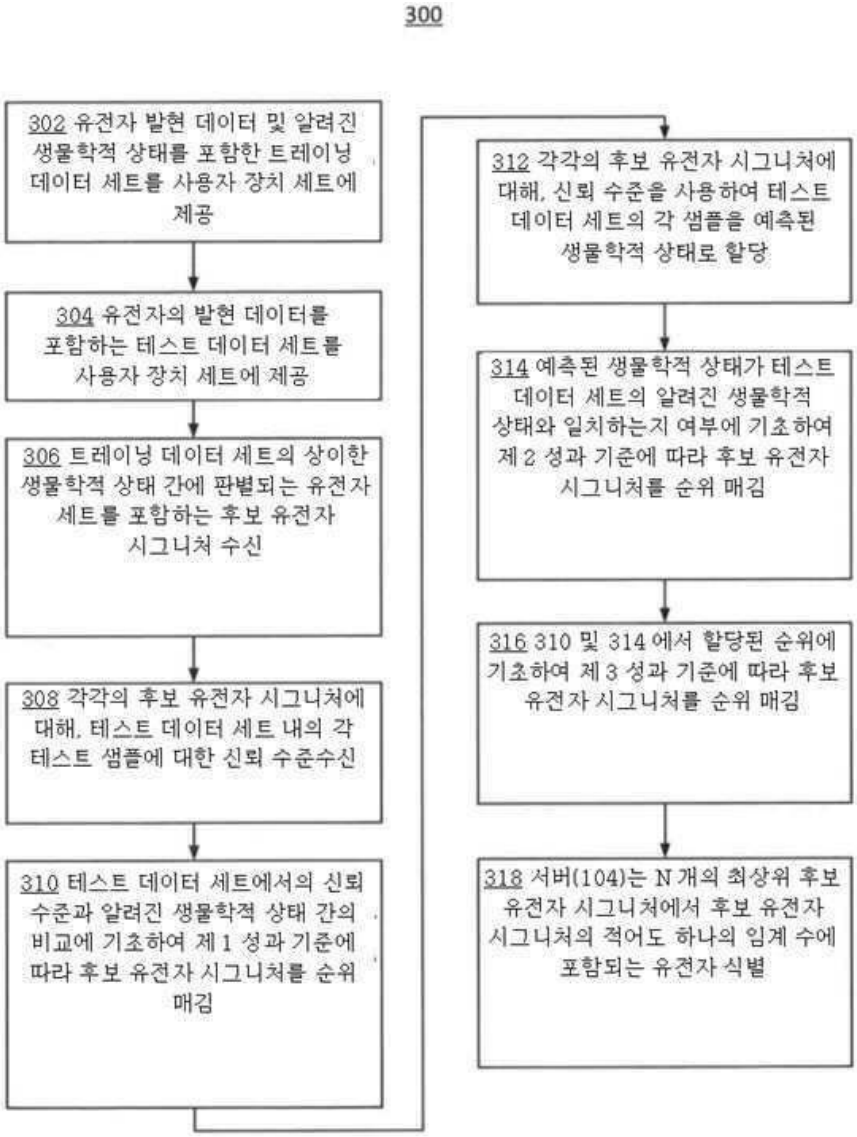
도면1


100
유전자의 발현 및 알려진 생물학적 상태 데이터베이스 106
사용자 장치 A
프로세서 110a
사용자 인터페이스
네트워크 102
사용자 장치 B
프로세서 110b
사용자 인터페이스
서버 104
프로세서 105

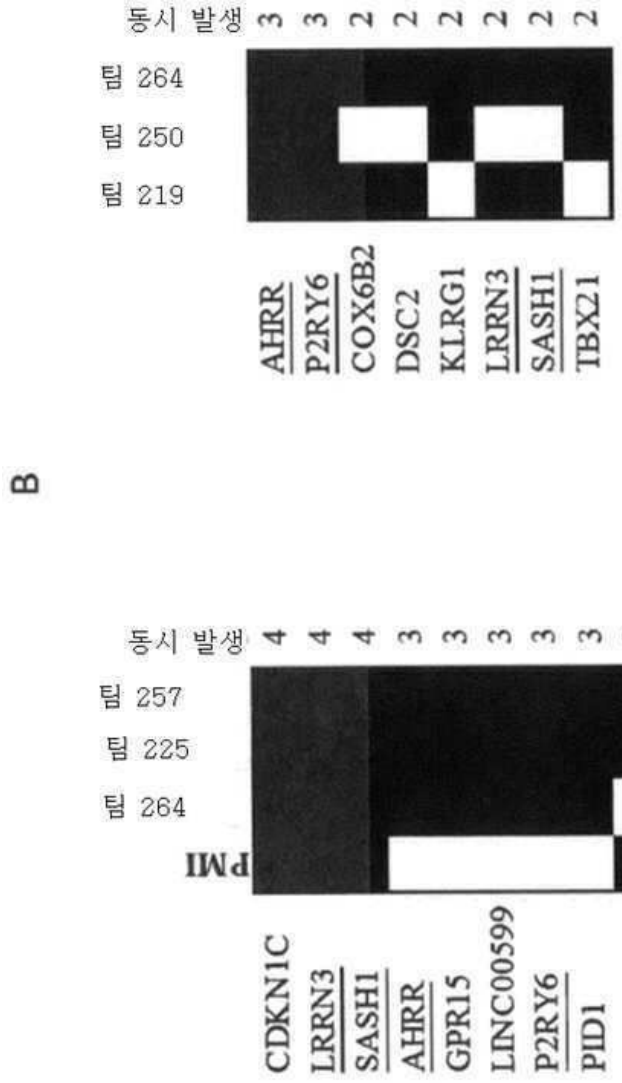
도면2


200
CPU 206
통신 인터페이스 유닛 208
입력/출력 컨트롤러 210
시스템 메모리
ROM 202
RAM 204
저장 장치
운영 체제 212
어플리케이션(들) 214
데이터베이스(들) 216

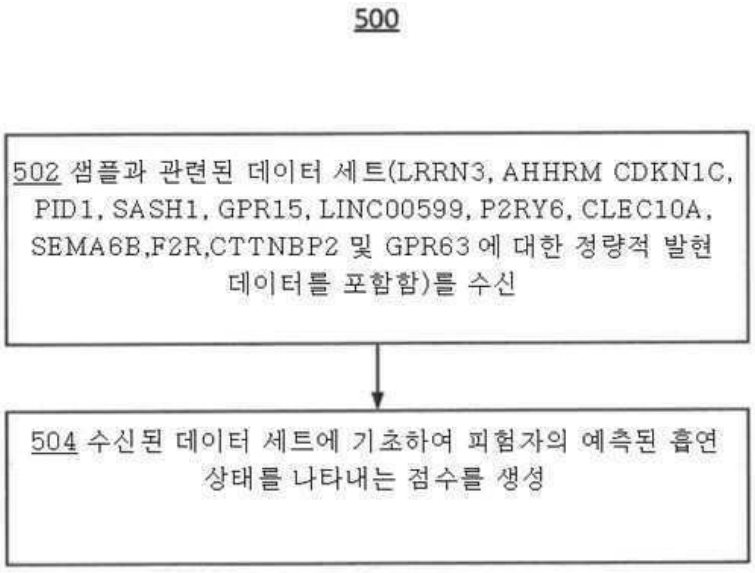
도면3


300
302 유전자 발현 데이터 및 알려진 생물학적 상태를 포함한 트레이닝 데이터 세트를 사용자 장치 세트에 제공
304 유전자의 발현 데이터를 포함하는 테스트 데이터 세트를 사용자 장치 세트에 제공
306 트레이닝 데이터 세트의 상이한 생물학적 상태 간에 판별되는 유전자 세트를 포함하는 후보 유전자 시그니쳐 수신
308 각각의 후보 유전자 시그니쳐에 대해, 테스트 데이터 세트 내의 각 테스트 샘플에 대한 신뢰 수준수신
310 테스트 데이터 세트에서의 신뢰 수준과 알려진 생물학적 상태 간의 비교에 기초하여 제 1 성과 기준에 따라 후보 유전자 시그니쳐를 순위 매김
312 각각의 후보 유전자 시그니쳐에 대해, 신뢰 수준을 사용하여 테스트 데이터 세트의 각 샘플을 예측된 생물학적 상태로 할당
314 예측된 생물학적 상태가 테스트 데이터 세트의 알려진 생물학적 상태와 일치하는지 여부에 기초하여 제 2 성과 기준에 따라 후보 유전자 시그니쳐를 순위 매김
316 310 및 314 에서 할당된 순위에 기초하여 제 3 성과 기준에 따라 후보 유전자 시그니쳐를 순위 매김
318 서버(104)는 N 개의 최상위 후보 유전자 시그니쳐에서 후보 유전자 시그니쳐의 적어도 하나의 임계 수에 포함되는 유전자 식별

도면4


B
동시 발생 3 3 2 2 2 2 2 2
팀 264
팀 250
팀 219
AHRR
P2RY6
COX6B2
DSC2
KLRG1
LRRN3
SASH1
TBX21
A
동시 발생 4 4 4 3 3 3 3 3 3 3 2 2 2
팀 257
팀 225
팀 264
PMI
CDKN1C
LRRN3
SASH1
AHRR
GPR15
LINC00599
P2RY6
PID1
RGL1
SEMA6B
CLEC10A
CTTNBP2
F2R

도면5



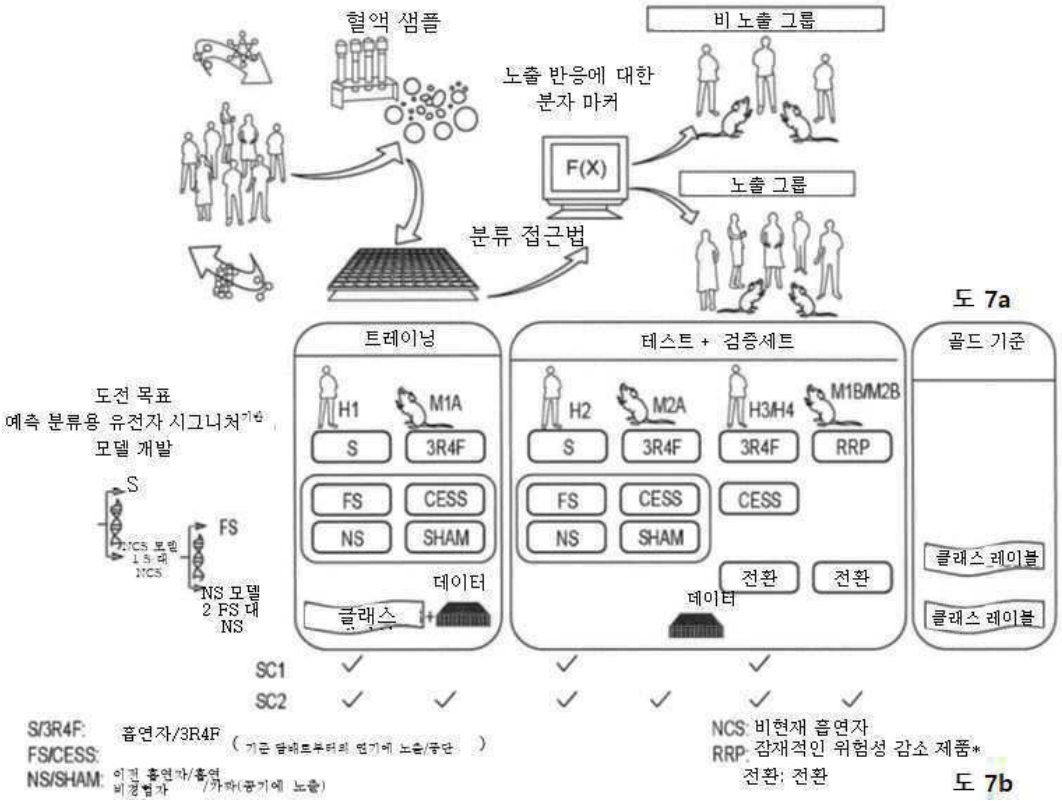
도면6

			흡연자	현재 비흡연자		기타 그룹	
데이터 세트 이름	데이터 세트 코드	일반적인 설명	S / 3R4F	FS / CESS	NS / SHAM	P/CMRTP	전환
GASMC NCT01760298	H1	임상 사례 통제 연구 40~70 세의 60 명의 피험자 그룹 당 남성(58%) 및 여성(42%)	N=109 COPD 골드 스테이지 1 및 2 ≥ 10 팩/년 흡연 이력	N=57 일치하는 흡연 이력 1 년 이상 금연	N=58 연령 성별 및 민족에 일치	-	-
BLD-SMK-01	H2	:바이오 뱅크 혈액 샘플 23~65 세 제외 질병의 병력 및 약물	N=27 3 년 이상 ≥10 개비/일	N=26 2 년 이상 금연	N=28 연령 및 성별에 의해 일치	-	-
REX C-03-EU NCT01959932	H3	구금 상태에서 무작위 임상 통제 연구.	N=180	N=37 5 일 동안 금연	-	-	N=70 이 22 로 5 일 동안 전환
REX C-04-JP NCT01970982	H4		N=176	N=31	-	-	N=63
마우스 C57BL/6J	M1AM1B	SR4P 접기 또는 MRTP 에어로졸에 마일 노출 노출에 의해 폴리핀 ↓ 개체 (C57BL/6) 또는 3 개체 (APOE -/-) 1-HR 분획은 신선한 공기와 함께 혹의. SR4P 에 2 개월 노출 후 중단 또는 검출 없음	N=40	N=27	N=45	P/MRTP N=45	N=28
마우스 APOE-/-15	M2AM2B		N=12	N=8	N=13	CMRTP (THS2.2) N=9	N=8

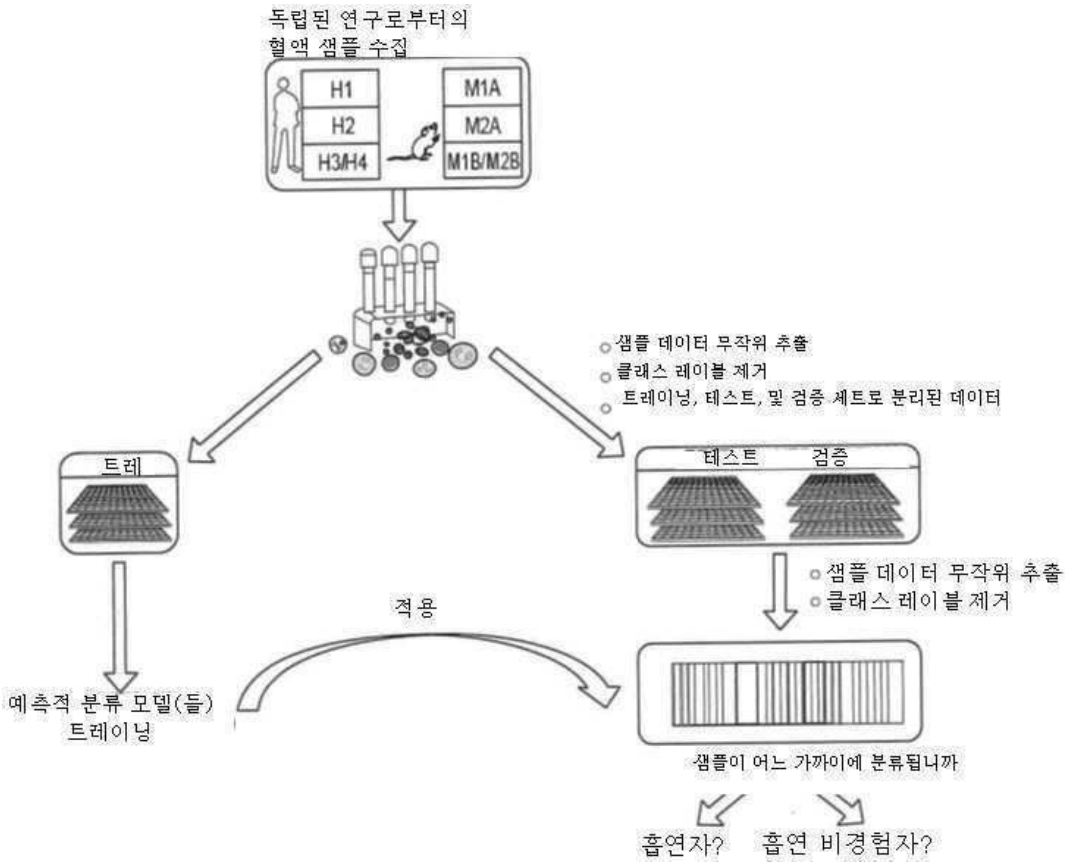
약어: REX, 임상적 ZRHR-감소된 노출; IS, 흡입 연구; S, 흡연자; Fs, 이전 흡연자; NS, 흡연 미경험자, P/CMRTP, 잠재적/후보 완화된 위험성 담배 제품; 세스(CESS), 중단(CESSATION); 전환(Switch), 전환(Switching). NCT 로 시작하는 번호는 임상 시험에 등록된 임상 연구의 고유 식별자에 해당한다. GOV.

색상 코드: 트레이닝 데이터 세트 / 테스트 데이터 세트 / 검증 데이터 세트

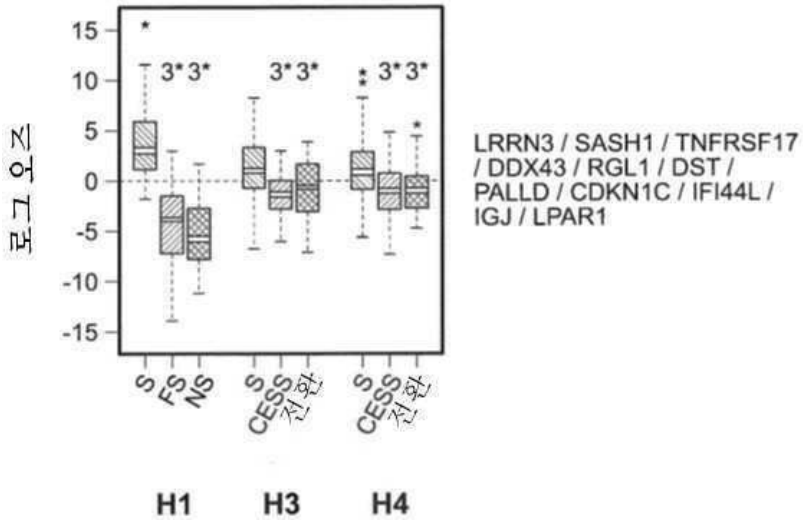
도면7


혈액 샘플
비 노출 그룹
노출 반응에 대한
분자 마커
F(X)
노출 그룹
분류 접근법
도 7a
트레이닝
테스트 + 검증세트
골드 기준
도전 목표
예측 분류용 유전자 시그니처
모델 개발
H1
M1A
H2
M2A
H3/H4
M1B/M2B
S
3R4F
RRP
FS
CESS
NS
SHAM
전환
데이터
클래스
클래스 레이블
SC1
SC2
S/3R4F:
FS/CESS:
NS/SHAM:
NCS: 비현재 흡연자
RRP: 잠재적인 위험성 감소 제품*
전환: 전환
도 7b

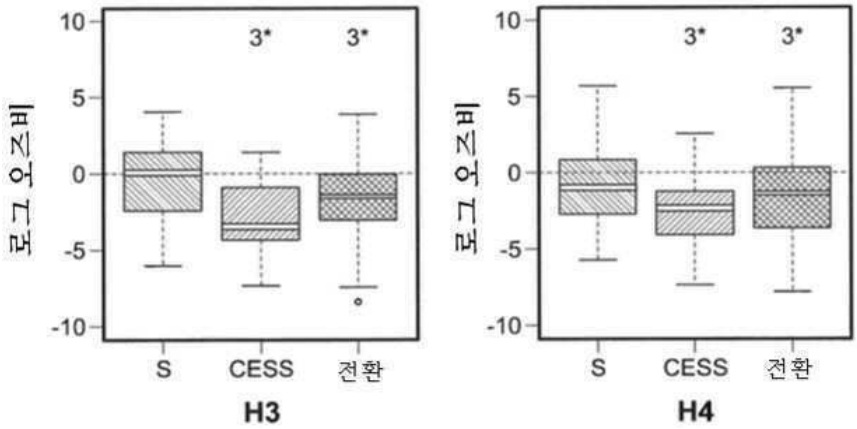
도면8


독립된 연구로부터의
혈액 샘플 수집
H1
H2
H3/H4
M1A
M2A
M1B/M2B
샘플 데이터 무작위 추출
클래스 레이블 제거
트레이닝, 테스트, 및 검증 세트로 분리된 데이터
트레
테스트
검증
샘플 데이터 무작위 추출
클래스 레이블 제거
적용
예측적 분류 모델(들)
트레이닝
샘플이 어느 가까이에 분류됩니까
흡연자?
흡연 비경험자?

도면9a


로그 오즈
15
10
5
0
-5
-10
-15
3* 3*
3* 3*
3* 3*
S
FS
NS
S
CESS
전환
S
CESS
전환
H1
H3
H4
LRRN3 / SASH1 / TNFRSF17
/ DDX43 / RGL1 / DST /
PALLD / CDKN1C / IFI44L /
IGJ / LPAR1

도면9b


로그 오즈비
10
5
0
-5
-10
3*
3*
S
CESS
전환
H3
로그 오즈비
10
5
0
-5
-10
3*
3*
S
CESS
전환
H4

도면10

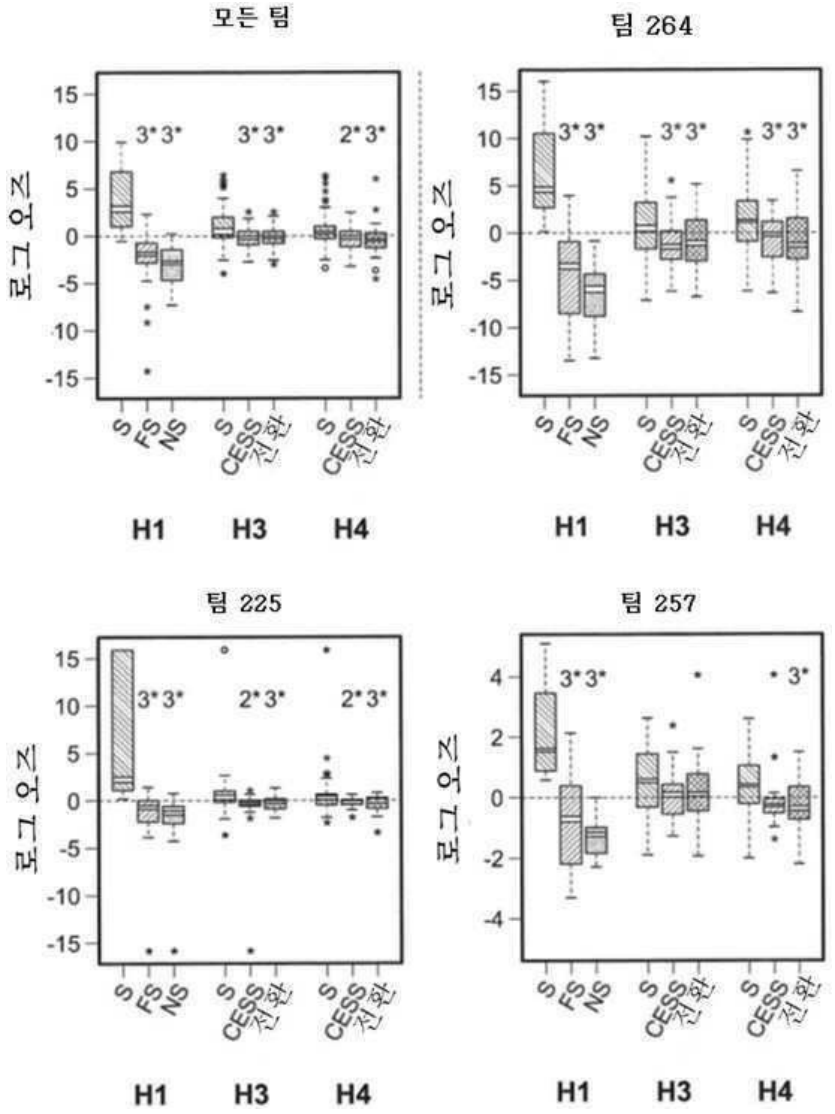
SC1

팀	AUPR	MCC	평균 순위
225	**0.94**	**0.24**	**1**
264	**0.92**	**0.18**	**2**
257	**0.91**	**0.16**	**3**
259	0.90	0.15	4
269	0.90	0.10	6.5
222	0.89	0.13	7
250	0.89	0.12	7
247	0.90	0.09	8
283	0.90	0.08	8
290	0.87	0.11	8.5
221	0.85	0.06	11
215	0.82	-0.07	12
PMI	0.93	0.24	

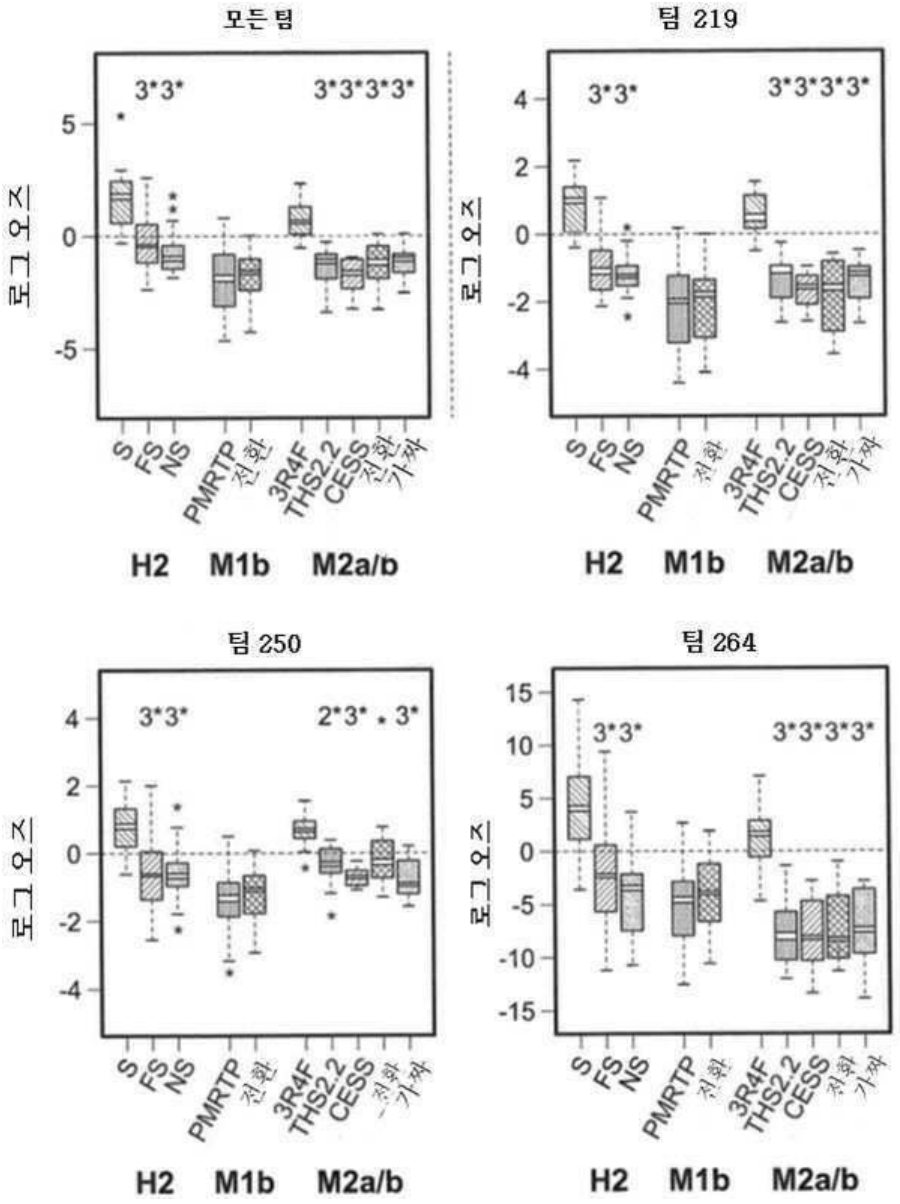
SC2

팀	AUPR	MCC	평균 순위
219	**0.99**	**0.87**	**1**
250	**0.96**	**0.81**	**2**
264	**0.96**	**0.75**	**3**
225	0.73	0.38	4
247	0.62	0.20	5.5
221	0.46	0.39	5.5

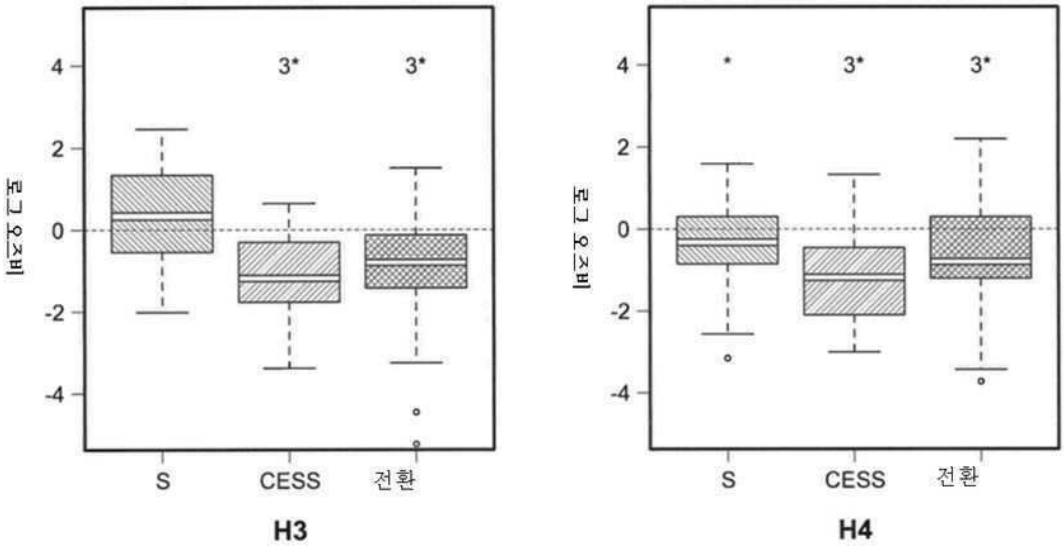
도면11a



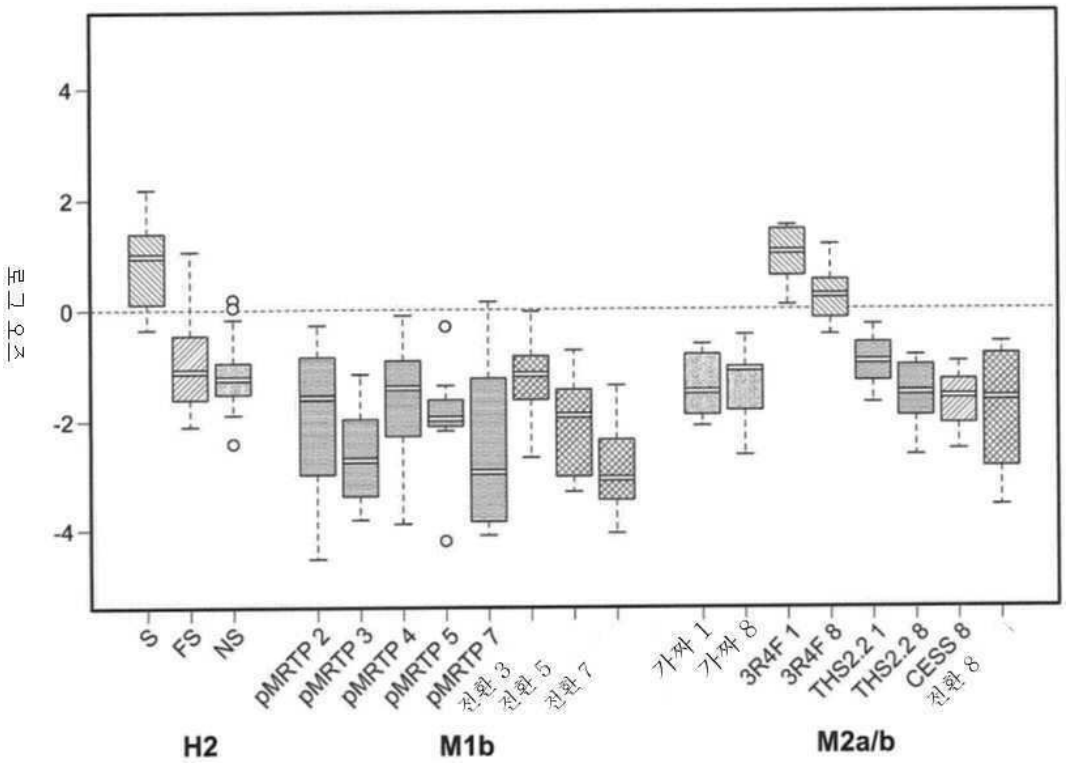
도면11b


모든 팀
팀 219
팀 250
팀 264
로그 오즈
3*3*
3*3*3*3*
2*3* * 3*
S
FS
NS
PMRTP
전환
3R4F
THS2.2
CESS
전환
가짜
H2
M1b
M2a/b

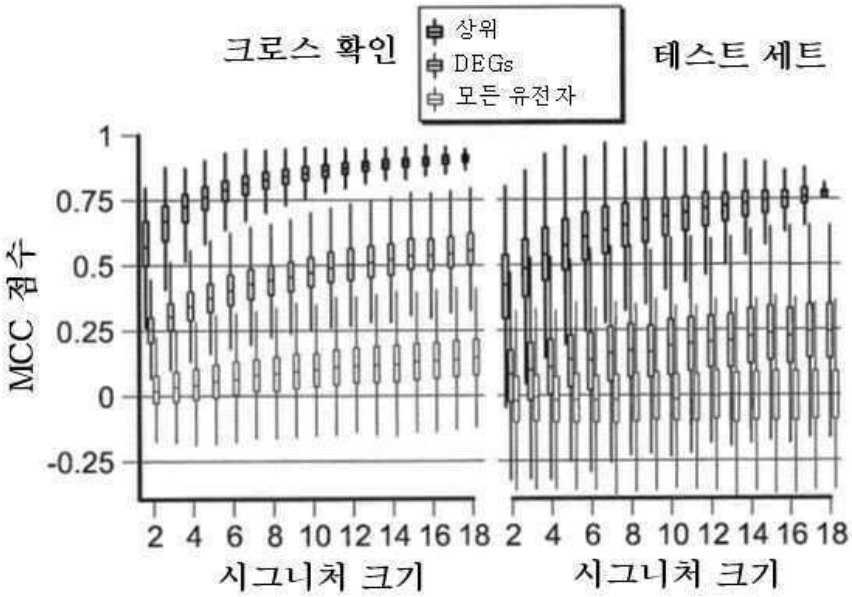
도면12


로그 오즈비
4
2
0
-2
-4
3*
3*
S
CESS
전환
H3
로그 오즈비
4
2
0
-2
-4
*
3*
3*
S
CESS
전환
H4

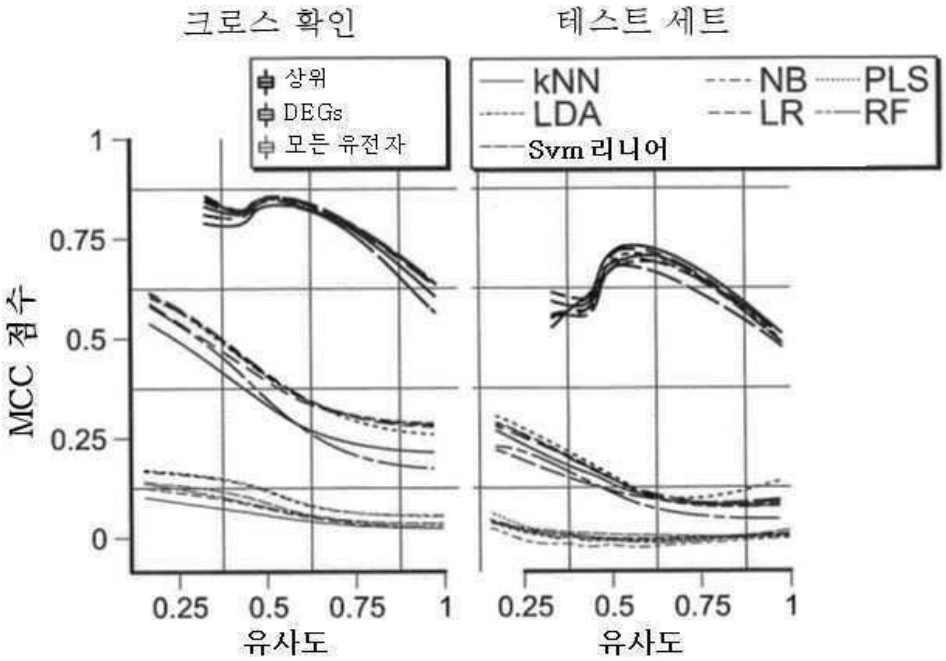
도면13


로그 오즈
4
2
0
-2
-4
8
S
FS
NS
pMRTP 2
pMRTP 3
pMRTP 4
pMRTP 5
pMRTP 7
전환 3
전환 5
전환 7
가짜 1
가짜 8
3R4F 1
3R4F 8
THS2.2 1
THS2.2 8
CESS 8
전환 8
H2
M1b
M2a/b

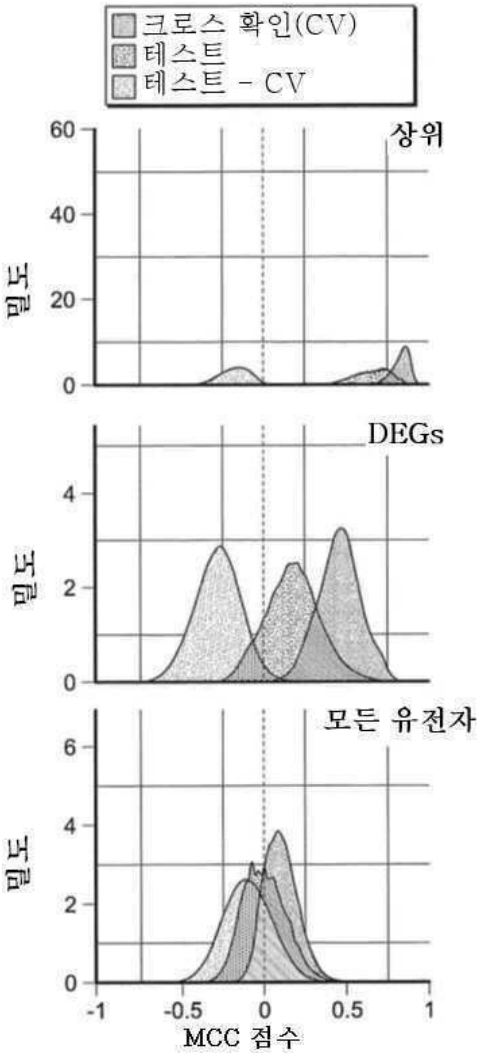
도면14a


크로스 확인
상위
DEGs
모든 유전자
테스트 세트
MCC 점수
1
0.75
0.5
0.25
0
-0.25
2 4 6 8 10 12 14 16 18
시그니처 크기
2 4 6 8 10 12 14 16 18
시그니처 크기

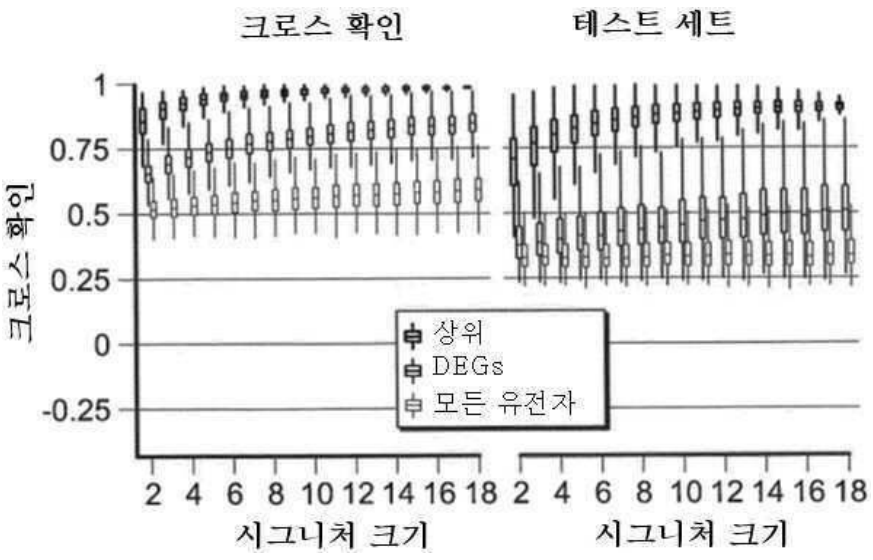
도면14b


크로스 확인
테스트 세트
상위
DEGs
모든 유전자
kNN
LDA
Svm 리니어
NB
LR
PLS
RF
MCC 점수
1
0.75
0.5
0.25
0
0.25 0.5 0.75 1
유사도
0.25 0.5 0.75 1
유사도

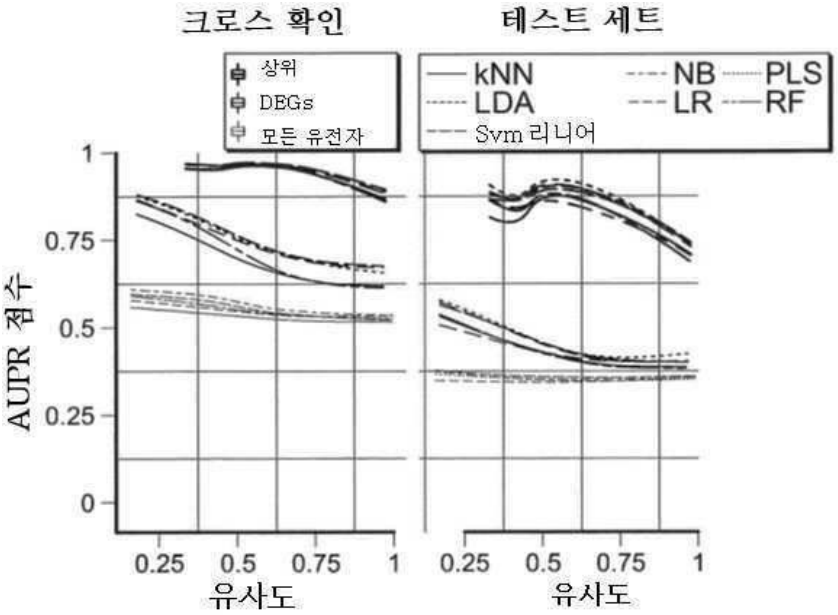
도면14c


크로스 확인(CV)
테스트
테스트 - CV
상위
DEGs
모든 유전자
밀도
MCC 점수

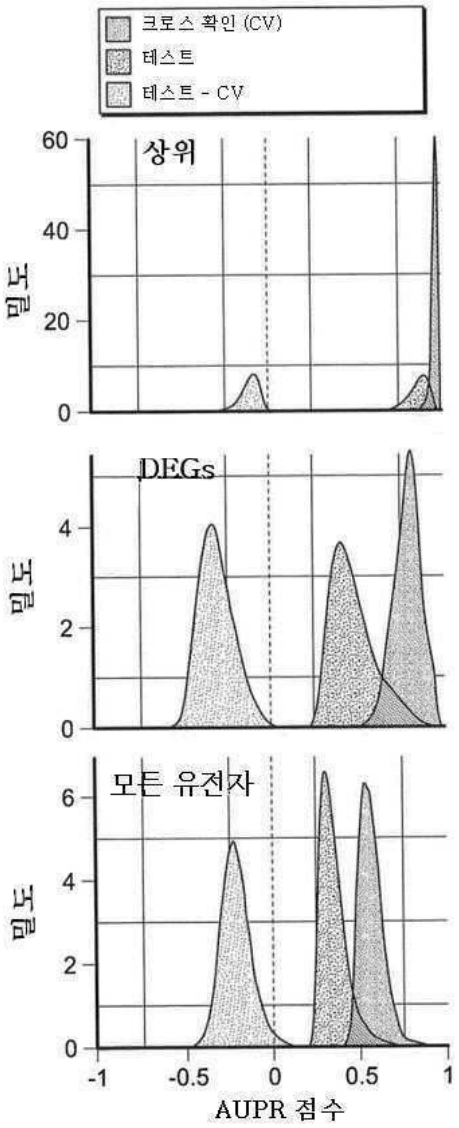
도면15a


크로스 확인
테스트 세트
크로스 확인
상위
DEGs
모든 유전자
시그니처 크기
시그니처 크기

도면15b


크로스 확인
테스트 세트
상위
DEGs
모든 유전자
kNN
LDA
Svm 리니어
NB
LR
PLS
RF
AUPR 점수
1
0.75
0.5
0.25
0
0.25
0.5
0.75
1
유사도
0.25
0.5
0.75
1
유사도

도면15c


크로스 확인 (CV)
테스트
테스트 - CV
상위
DEGs
모든 유전자
밀도
60
40
20
0
4
2
6
-1
-0.5
0
0.5
1
AUPR 점수