

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2021-0055072
(43) 공개일자 2021년05월14일

(51) 국제특허분류(Int. Cl.)
G16B 20/20 (2019.01) *G06N 3/12* (2006.01)
G16B 40/20 (2019.01) *G16H 10/60* (2018.01)
G16H 50/80 (2018.01)
 (52) CPC특허분류
G16B 20/20 (2019.02)
G06N 3/126 (2013.01)
 (21) 출원번호 10-2021-7010041
 (22) 출원일자(국제) 2019년09월06일
 심사청구일자 없음
 (85) 번역문제출일자 2021년04월05일
 (86) 국제출원번호 PCT/US2019/049942
 (87) 국제공개번호 WO 2020/051445
 국제공개일자 2020년03월12일
 (30) 우선권주장
 62/728,536 2018년09월07일 미국(US)

(71) 출원인
 리제너론 파아마슈티컬스, 인크.
 미국 뉴욕 10591-6707 테리타운 올드 소우 밀 리버 로드 777
 (72) 발명자
 스테이플스 제프리
 미국 뉴욕주 10591 테리타운 올드 소우 밀 리버 로드 777
 곤자가-요레귀이 클라우디아
 미국 뉴욕주 10591 테리타운 올드 소우 밀 리버 로드 777
 (뒷면에 계속)
 (74) 대리인
 장훈

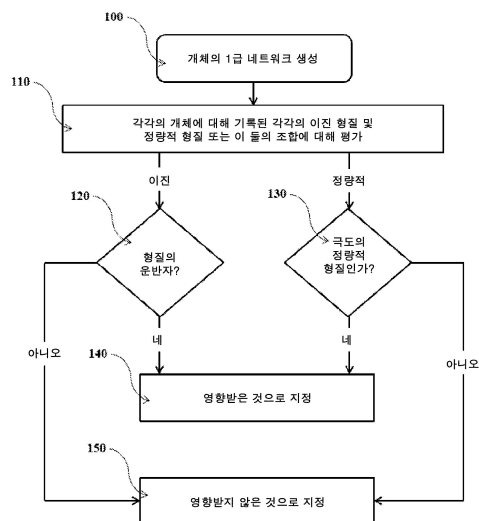
전체 청구항 수 : 총 63 항

(54) 발명의 명칭 **혈통 농축 및 혈통 내 가족-기반 분석을 위한 방법 및 시스템**

(57) 요약

농축된 혈통을 생성하기 위한 방법, 비-일시적 컴퓨터 구현 방법 및 시스템이 제공된다. 또한 질병-유발 변이체를 결정하기 위한 방법, 비-일시적 컴퓨터 구현 방법 및 시스템이 제공된다.

대표도 - 도1



(52) CPC특허분류

G16B 40/20 (2019.02)

G16H 10/60 (2018.01)

G16H 50/80 (2018.01)

(72) 발명자

리드 제프리

미국 뉴욕주 10591 테리타운 올드 소우 밀 리버 로
드 777

해비거 루카스

미국 뉴욕주 10591 테리타운 올드 소우 밀 리버 로
드 777

명세서

청구범위

청구항 1

질병-유발 변이체를 식별하기 위한 시스템으로서, 상기 시스템은,
 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하고, 상기 프로세서는,
 코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하고;
 1급 네트워크에서 개체가 영향받았는지 또는 영향받지 않았는지 식별하고, 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 있는 상기 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질 또는 극도의 정량적 형질이 없는 상기 개체는 영향받지 않은 것으로 식별되고; 및
 영향받거나 영향받지 않는 것으로서의 지정을 포함하는 개체를 함유하는 적어도 하나의 농축된 혈통을 생성하도록 구성되는, 질병-유발 변이체를 식별하기 위한 시스템.

청구항 2

제 1항에 있어서, 상기 프로세서는 적어도 하나의 농축된 혈통 내에서 및 혈통에 걸쳐서 공동 분리되는 변이체 형질 쌍을 식별하기 위한 분리 분석을 수행하도록 추가로 구성되는, 시스템

청구항 3

제 1항에 있어서, 상기 프로세서는 질병-유발 변이체를 결정하기 위해 변이체 형질 쌍을 분석하기 위한 분리 분석을 수행하도록 추가로 구성되는, 시스템

청구항 4

제 1항에 있어서, 상기 시퀀싱 데이터는 엑솜 시퀀싱 데이터 또는 전체 게놈 시퀀싱 데이터 또는 유전자형 배열 데이터로 구성된 군으로부터 선택될 수 있는, 시스템.

청구항 5

제 1항에 있어서, 상기 프로세서는 코호트에서 적어도 하나의 이진 형질의 유병률이 5%를 초과하는 경우 영향받은 것으로 식별된 개체를 영향받지 않은 것으로 식별하도록 추가로 구성되는, 시스템.

청구항 6

제 1항에 있어서, 상기 프로세서는 개체의 적어도 하나의 극도의 정량적 형질이 코호트의 평균 극도의 정량적 형질으로부터 2 표준 편차 미만인 경우 영향받은 것으로 식별된 개체를 영향받지 않은 것으로 식별하도록 추가로 구성되는, 시스템.

청구항 7

제 1항 내지 제 6항 중 어느 한 항에 있어서, 상기 시스템에 의해 생성된 상기 농축된 혈통이 단일의 가능한 구조를 갖는, 시스템.

청구항 8

제 1항 내지 제 7항 중 어느 한 항에 있어서, 상기 시스템에 의해 생성된 상기 농축된 혈통이 공통 조상의 3명 이상의 영향받은 개체들을 포함하는, 시스템.

청구항 9

제 1항 내지 제 8항 중 어느 한 항에 있어서, 상기 시스템에 의해 생성된 상기 농축된 혈통이 영향받지 않은 것으로 결정된 부모를 갖는 하나 이상의 영향받는 개체들을 포함하는, 시스템.

청구항 10

제 1항 내지 제 9항 중 어느 한 항에 있어서, 상기 프로세서는 분리 분석을 수행하기 위해 적어도 하나의 혈통을 우선순위화하도록 추가로 구성되는, 시스템.

청구항 11

제 10항에 있어서, 적어도 하나의 혈통의 상기 우선순위화는 적어도 하나의 관련된 영향받지 않은 개체를 포함하는 적어도 하나의 혈통을 선택하는 것을 포함하는, 시스템.

청구항 12

제 10항에 있어서, 적어도 하나의 혈통의 상기 우선순위화는 적어도 두명의 영향받은 형제자매를 포함하는 적어도 하나의 혈통을 선택하는 것을 포함하는, 시스템.

청구항 13

제 10항에 있어서, 상기 분리 분석은 분리의 우성 유전적 모델을 형성하는, 시스템.

청구항 14

제 10항에 있어서, 상기 분리 분석은 분리의 열성 유전적 모델을 형성하는, 방법.

청구항 15

제 3항에 있어서, 질병-유발 변이체를 결정하기 위한 변이체 형질 쌍의 분석이 가족-기반 연관 분석을 사용하여 수행되는, 방법.

청구항 16

질병-유발 변이체를 식별하는 방법으로서,

코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하는 단계;

1급 네트워크에서 개체를 영향받거나 영향받지 않은 것으로 식별하고, 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합에 의해 영향받는 상기 개체는 영향받는 것으로 식별되고 적어도 하나의 이진 형질 또는 극도의 정량적 형질에 의해 영향받지 않은 상기 개체는 영향받지 않은 것으로 식별되는 단계;

영향받거나 영향받지 않은 것으로의 지정을 포함하는 개체를 함유하는 하나 이상의 농축된 혈통을 생성하는 단계;

적어도 하나의 농축된 혈통 내에서 및 혈통에 걸쳐서 공동 분리되는 변이체 형질 쌍을 식별하기 위해 분리 분석을 수행하는 단계; 및

질병-유발 변이체를 식별하기 위해 변이체 형질 쌍을 분석하는 단계를 포함하는, 질병-유발 변이체를 식별하는 방법.

청구항 17

제 16항에 있어서, 상기 시퀀싱 데이터는 엑솜 시퀀싱 데이터, 전체 게놈 시퀀싱 데이터 및 유전자형 배열 데이터로부터 선택될 수 있는, 방법.

청구항 18

제 16항 또는 제 17항 중 어느 한 항에 있어서, 상기 코호트에서 적어도 하나의 이진 형질의 유병률이 5%를 초과하는 경우 영향받은 것으로 식별된 개체는 영향받지 않은 것으로 비-식별되는, 방법.

청구항 19

제 16항 또는 제 17항 중 어느 한 항에 있어서, 개체의 적어도 하나의 극도의 정량적 형질이 코호트의 평균 극도의 정량적 형질으로부터 2 표준 편차 미만인 경우 영향받은 것으로 식별된 개체는 영향받지 않은 것으로 비-

식별되는, 방법.

청구항 20

제 16항 내지 제 19항 중 어느 한 항에 있어서, 상기 농축된 혈통은 단일의 가능한 구조를 갖는, 방법.

청구항 21

제 16항 내지 제 20항 중 어느 한 항에 있어서, 상기 농축된 혈통은 공통 조상의 3명 이상의 영향받은 개체들을 포함하는, 방법.

청구항 22

제 16항 내지 제 21항 중 어느 한 항에 있어서, 상기 농축된 혈통은 영향받지 않은 것으로 결정된 부모를 갖는 하나 이상의 영향받는 개체들을 포함하는, 방법.

청구항 23

제 16항 내지 제 22항 중 어느 한 항에 있어서, 분리 분석을 수행하기 위해 적어도 하나의 혈통을 우선순위화하는 것을 추가로 포함하는, 방법.

청구항 24

제 23항에 있어서, 적어도 하나의 혈통의 상기 우선순위화는 적어도 하나의 관련된 영향받지 않은 개체를 포함하는 적어도 하나의 혈통을 선택하는 것을 포함하는, 방법.

청구항 25

제 23항에 있어서, 적어도 하나의 혈통의 상기 우선순위화는 적어도 두명의 영향받은 형제자매를 포함하는 적어도 하나의 혈통을 선택하는 것을 포함하는, 방법.

청구항 26

제 23항 내지 제 25항 중 어느 한 항에 있어서, 상기 분리 분석은 분리의 우성 유전적 모델을 형성하는, 방법.

청구항 27

제 23항 내지 제 25항 중 어느 한 항에 있어서, 상기 분리 분석은 분리의 열성 유전적 모델을 형성하는, 방법.

청구항 28

제 16항 내지 제 27항 중 어느 한 항에 있어서, 질병-유발 변이체를 식별하기 위한 변이체 형질 쌍의 분석이 가족-기반 연관 분석을 사용하여 수행되는, 방법.

청구항 29

프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하는 강력한 명령의 비-일시적 컴퓨터 판독가능 매체로서,

코호트의 엑스 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하고;

1급 네트워크에서 개체를 영향받거나 영향받지 않은 것으로 식별하고, 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 있는 상기 개체는 영향받는 것으로 식별되고 적어도 하나의 이진 형질 또는 극도의 정량적 형질이 없는 상기 개체는 영향받지 않은 것으로 식별되고;

영향받거나 영향받지 않은 것으로의 지정을 포함하는 개체를 함유하는 하나 이상의 농축된 혈통을 생성하고;

적어도 하나의 농축된 혈통 내에서 및 혈통에 걸쳐서 공동 분리되는 변이체 형질 쌍을 식별하기 위해 분리 분석을 수행하고; 및

질병-유발 변이체를 결정하기 위해 변이체 형질 쌍을 분석하는 것을 포함하는, 비-일시적 컴퓨터 판독가능 매체.

청구항 30

제 29항에 있어서, 상기 시퀀싱 데이터는 엑솜 시퀀싱 데이터, 전체 게놈 시퀀싱 데이터 및 유전자형 배열 데이터로부터 선택될 수 있는, 비-일시적 컴퓨터 판독가능 매체.

청구항 31

제 29항 또는 제 30항 중 어느 한 항에 있어서, 코호트에서 적어도 하나의 이진 형질의 유병률이 5%를 초과하는 경우 영향받은 것으로 식별된 개체는 영향받지 않은 것으로 비-식별되는, 비-일시적 컴퓨터 판독가능 매체.

청구항 32

제 29항 또는 제 30항 중 어느 한 항에 있어서, 개체의 적어도 하나의 극도의 정량적 형질이 코호트의 평균 극도의 정량적 형질으로부터 2 표준 편차 미만인 경우 영향받은 것으로 식별된 개체는 영향받지 않은 것으로 비-식별되는, 비-일시적 컴퓨터 판독가능 매체.

청구항 33

제 29항 내지 제 32항 중 어느 한 항에 있어서, 상기 농축된 혈통은 단일의 가능한 구조를 갖는, 비-일시적 컴퓨터 판독가능 매체.

청구항 34

제 29항 내지 제 33항 중 어느 한 항에 있어서, 상기 농축된 혈통은 공통 조상의 3명 이상의 영향받은 개체들을 포함하는, 비-일시적 컴퓨터 판독가능 매체.

청구항 35

제 29항 내지 제 34항 중 어느 한 항에 있어서, 상기 농축된 혈통은 영향받지 않은 것으로 결정된 부모를 갖는 하나 이상의 영향받는 개체들을 포함하는, 비-일시적 컴퓨터 판독가능 매체.

청구항 36

제 29항 내지 제 35 중 어느 한 항에 있어서, 상기 방법은 분리 분석을 수행하기 위해 적어도 하나의 혈통을 우선순위화하는 것을 추가로 포함하는, 비-일시적 컴퓨터 판독가능 매체.

청구항 37

제 36항에 있어서, 상기 적어도 하나의 혈통의 상기 우선순위화는 적어도 하나의 관련된 영향받지 않은 개체를 포함하는 적어도 하나의 혈통을 선택하는 것을 포함하는, 비-일시적 컴퓨터 판독가능 매체.

청구항 38

제 36항에 있어서, 상기 적어도 하나의 혈통의 상기 우선순위화는 적어도 두명의 영향받은 형제자매를 포함하는 적어도 하나의 혈통을 선택하는 것을 포함하는, 비-일시적 컴퓨터 판독가능 매체.

청구항 39

제 36항 내지 제 38항 중 어느 한 항에 있어서, 상기 분리 분석이 분리의 우성 유전적 모델을 형성하는, 비-일시적 컴퓨터 판독가능 매체.

청구항 40

제 36항 내지 제 38항 중 어느 한 항에 있어서, 상기 분리 분석이 분리의 열성 유전적 모델을 형성하는, 비-일시적 컴퓨터 판독가능 매체.

청구항 41

제 29항 내지 제 40항 중 어느 한 항에 있어서, 상기 질병-유발 변이체를 식별하기 위한 변이체 형질 쌍의 분석이 가족-기반 연관 분석을 사용하여 수행되는, 비-일시적 컴퓨터 판독가능 매체.

청구항 42

농축된 혈통을 생성하기 위한 시스템으로서, 상기 시스템은,

데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하고, 상기 프로세서는,

코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하고;

1급 네트워크에서 개체가 영향받았는지 또는 영향받지 않았는지 식별하고, 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 있는 상기 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질 또는 극도의 정량적 형질이 없는 상기 개체는 영향받지 않은 것으로 식별되고; 및

영향받거나 영향받지 않는 것으로의 지정을 포함하는 개체를 함유하는 적어도 하나의 농축된 혈통을 생성하도록 구성되는, 농축된 혈통을 생성하기 위한 시스템.

청구항 43

제 42항에 있어서, 상기 시퀀싱 데이터는 엑솜 시퀀싱 데이터, 또는 전체 게놈 시퀀싱 데이터 또는 유전자형 배열 데이터로 구성된 군으로부터 선택될 수 있는, 시스템.

청구항 44

제 42항에 있어서, 상기 프로세서는 코호트에서 적어도 하나의 이진 형질의 유병률이 5%를 초과하는 경우 영향받은 것으로 식별된 개체는 영향받지 않은 것으로 식별하도록 추가로 구성되는, 시스템.

청구항 45

제 42항에 있어서, 상기 프로세서는 개체의 적어도 하나의 극도의 정량적 형질이 코호트의 평균 극도의 정량적 형질으로부터 2 표준 편차 미만인 경우 영향받은 것으로 식별된 개체는 영향받지 않은 것으로 식별하도록 추가로 구성되는, 시스템

청구항 46

제 42항 내지 제 45항 중 어느 한 항에 있어서, 상기 시스템에 의해 생성된 상기 농축된 혈통은 단일의 가능한 구조를 갖는, 시스템.

청구항 47

제 42항 내지 제 46항 중 어느 한 항에 있어서, 상기 시스템에 의해 생성된 상기 농축된 혈통은 공통 조상의 3명 이상의 영향받은 개체들을 포함하는, 시스템.

청구항 48

제 42항 내지 제 47항 중 어느 한 항에 있어서, 상기 시스템에 의해 생성된 상기 농축된 혈통은 영향받지 않은 것으로 결정된 부모를 갖는 하나 이상의 영향받은 개체들을 포함하는, 시스템.

청구항 49

농축된 혈통을 생성하는 방법으로서, 상기 방법은,

코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하는 단계;

1급 네트워크에서 개체가 영향받는지 또는 영향받지 않았는지 식별하고, 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합에 의해 영향받은 상기 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질 또는 극도의 정량적 형질에 의해 영향받지 않은 상기 개체는 영향을 받지 않은 것으로 식별되는 단계; 및

영향받거나 영향받지 않는 것으로서의 지정을 포함하는 개체를 함유하는 적어도 하나의 농축된 혈통을 생성하는 단계를 포함하는, 농축된 혈통을 생성하는 방법.

청구항 50

제 49항에 있어서, 상기 시퀀싱 데이터는 엑솜 시퀀싱 데이터, 전체 게놈 시퀀싱 데이터 및 유전자형 배열 데이

터로부터 선택될 수 있는, 방법.

청구항 51

제 49항 및 제 50항 중 어느 한 항에 있어서, 코호트에서 적어도 하나의 이진 형질의 유병률이 5%를 초과하는 경우 영향받은 것으로 식별된 개체는 영향받지 않은 것으로 비-식별되는, 방법.

청구항 52

제 49항 및 제 50항 중 어느 한 항에 있어서, 개체의 적어도 하나의 극도의 정량적 형질이 코호트의 평균 극도의 정량적 형질으로부터 2 표준 편차 미만인 경우 영향받은 것으로 식별된 개체는 영향받지 않은 것으로 비-식별되는, 방법.

청구항 53

제 49항 내지 제 52항 중 어느 한 항에 있어서, 상기 농축된 혈통은 단일의 가능한 구조를 갖는, 방법.

청구항 54

제 49항 내지 제 53항 중 어느 한 항에 있어서, 상기 농축된 혈통은 공통 조상의 3명 이상의 영향받은 개체들을 포함하는, 방법.

청구항 55

제 49항 내지 제 54항 중 어느 한 항에 있어서, 상기 농축된 혈통은 영향받지 않은 것으로 결정된 부모를 갖는 하나 이상의 영향받은 개체들을 포함하는, 방법.

청구항 56

프로세서로 하여금 농축된 혈통을 생성하는 방법을 수행하게 하는 강력한 명령의 비-일시적 컴퓨터 판독가능 매체로서,

코호트의 엑스 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하고;

1급 네트워크에서 개체가 영향받는지 또는 영향받지 않았는지 식별하고, 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 있는 상기 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질 또는 극도의 정량적 형질이 없는 상기 개체는 영향받지 않은 것으로 식별되고; 및

영향받거나 영향받지 않는 것으로서의 지정을 포함하는 개체를 함유하는 적어도 하나의 농축된 혈통을 생성하는 것을 포함하는, 비-일시적 컴퓨터 판독가능 매체.

청구항 57

제 56항에 있어서, 상기 시퀀싱 데이터는 엑스 시퀀싱 데이터, 전체 게놈 시퀀싱 데이터 및 유전자형 배열 데이터로부터 선택될 수 있는, 비-일시적 컴퓨터 판독가능 매체.

청구항 58

제 56항 또는 제 57항에 있어서, 코호트에서 적어도 하나의 이진 형질의 유병률이 5%를 초과하는 경우 영향받은 것으로 식별된 개체는 영향받지 않은 것으로 비-식별되는, 비-일시적 컴퓨터 판독가능 매체.

청구항 59

제 56항 내지 제 58항 중 어느 한 항에 있어서, 개체의 적어도 하나의 극도의 정량적 형질이 코호트의 평균 극도의 정량적 형질로부터 2 표준 편차 미만인 경우 영향받은 것으로 식별된 개체는 영향받지 않은 것으로 비-식별되는, 비-일시적 컴퓨터 판독가능 매체.

청구항 60

제 56항 내지 제 59항 중 어느 한 항에 있어서, 상기 농축된 혈통은 단일의 가능한 구조를 갖는, 비-일시적 컴퓨터 판독가능 매체.

청구항 61

제 56항 내지 제 60항 중 어느 한 항에 있어서, 상기 농축된 혈통은 공통 조상의 3명 이상의 영향받은 개체들을 포함하는, 비-일시적 컴퓨터 판독가능 매체.

청구항 62

제 56항 내지 제 61항 중 어느 한 항에 있어서, 상기 농축된 혈통은 영향받지 않은 것으로 결정된 부모를 갖는 하나 이상의 영향받은 개체들을 포함하는, 비-일시적 컴퓨터 판독가능 매체.

청구항 63

제 56항 내지 제 62항 중 어느 한 항에 있어서, 상기 방법은 분리 분석을 수행하기 위해 적어도 하나의 혈통을 우선순위화하는 것을 추가로 포함하는, 비-일시적 컴퓨터 판독가능 매체.

발명의 설명

기술 분야

- [0001] 관련 출원에 대한 상호 참조
- [0002] 본 출원은 2018년 9월 7일에 출원된 미국 가특허 출원 번호 62/728,536의 이익을 주장하며; 이 출원의 내용은 그 전체가 여기에 참조로 포함된다.
- [0003] 기술 분야
- [0004] 본 명세서는 일반적으로 대규모 집단 코호트에서 혈통 농축을 위한 방법 및 시스템에 관한 것이다. 보다 구체적으로, 본 명세서는 시퀀싱 데이터를 사용하여 혈통을 농축하기 위해 1급(first-degree) 가족 네트워크에서 영향 받은 개체들을 식별하고 회귀한 유전적 변이를 질병 및 질병 감수성에 연결하기 위해 혈통 내에서 및 혈통에 걸쳐서 공동 분리되는 변이체-형질 쌍을 추가로 식별하는 시스템 및 방법에 관한 것이다.

배경 기술

- [0005] 임상 연구자들은 질병을 일으키는 병원성 변이체를 식별하기 위해 지속적으로 노력하고 있다. 세포 유전체 검정 및 연결 패널의 유전자형분석은 각각 복제수 변이를 식별하고 큰 멘델의(특히 우성의) 질병 가족 내에서 일배체 형의 공동 분리를 식별하는 데 유용한 접근 방식으로 남아 있다. 그러나 복잡한 질병에서 병원성 변이체를 발견하기 위한 최적의 접근법은 아직 명확하지 않다.
- [0006] 계보학을 통한 변이체의 전이는 현대 유전학의 기초가 된다. 대부분의 유전적 장애는 질병을 일으키는 역할을 하는 몇몇 유전자에서 많은 유전자의 범위내에서 이질적이다. 많은 회귀 장애의 유전적 결함은 아직 파악되지 않은 채 남아있다. 고전적인 위치 클로닝 기술을 사용하면 원인 유전자가 있어야 하는 영역을 식별하기 위해 상당한 수의 영향받은 가족이 필요하며, 회귀 장애의 경우 이러한 가족이 항상 이용가능한 것은 아니다. 또한 관심 영역을 식별하는 것만으로는 충분하지 않으며; 이 영역에 있는 유전자는 모두 시퀀싱해야 하는데, 이는 매우 힘들 수 있다. 차세대 시퀀싱의 도래로 후보 유전자 영역을 선택할 필요없이 환자의 전체 게놈 또는 엑솜을 연구할 수 있다. 이제 대규모 연구 코호트에서 회귀 유전적 변이체를 발견하고 유전자형분석을 할 수 있지만, 이러한 변이체의 대부분은 소수의 개체- 집단-기반 유전 연구에서는 변이체의 > 50% 이 단일 개체에서 볼 수 있으며- 예게만 존재하여 연관 증거를 확립하기 어렵다.
- [0007] 대규모 유전 코호트의 게놈-전체 스캔에서 회귀 변이체가 이러한 이중 장애에 미치는 영향을 조사하는 것은 더욱 특히 어렵다. 서열 변이체에 대한 질병 원인의 명확한 할당은 종종 불가능하며, 특히 회귀하고 심각한 질병의 많은 경우에 근본적인 매우 낮은 빈도의 변이체의 경우에는 더욱 그렇다. 그러나, 특정 유전적 장애를 공유하는 관련 개체 세트가 식별되면, 이 이질성이 크게 감소하여 혈통 내에서 영향받은 개체에서 분리되는 특정 표현형을 유도하는 단일 유전자 및 변이체에 집중할 수 있게 된다.
- [0008] 전체 인간 게놈에 걸친 질병 유전자좌에 대한 편견없는 검색을 가능하게 하는 게놈-전체 연관 연구(genome-wide association studies, GWAS)의 잠재력은 유전학에서 전례없는 연구 기회를 제공한다. 동시에 많은 피험자에 걸쳐 수십만 개의 단일 뉴클레오타이드 다형성(SNP)을 조사하면 이러한 연구의 설계 및 분석에서 많은 통계적 문제가 발생한다. 이러한 규모의 유전자형분석은 데이터 품질 문제를 처리하기 위한 새로운 방법론이 필요하며; 마

참가지로 연관 테스트는 수십만 개의 마커에 대해 계산되며, 여러 비교를 위해 결과를 조정해야 한다. 이러한 문제의 규모는 이러한 조밀한 SNP 세트를 유전자형분석하는 새로운 기술 능력이 새로운 유전 질환 유전자좌를 식별하는 것으로 해석될지 또는 기술적 진보가 활용도가 낮게 남아있을지 여부에 대한 의문을 제기한다. 계층-전체 연관 연구에 접근하는 방법에는 집단-기반 및 가족-기반 디자인의 적어도 두 가지가 있다.

[0009] 집단-기반 연구는 수천 명의 피험자의 샘플 크기를 갖는다(Szklo M. Epidemiologic Reviews(1998) 20(1): 81-90). 그러나, 이러한 연구는 비용이 많이 들고 시간이 많이 걸리며 샘플 크기가 크기 때문에 표현형 및 유전형 이질성이 발생할 수 있다(Sorlie 및 Wei. Journal of American College of Cardiology(2011) 58(19): 2010-3; Laird 및 Lange. Statistical Science(2009) 24(4): 388-397).

[0010] 가족-기반 분석은 관심있는 표현형과 공동 분리되는 잠재적인 중간-내지-대형 효과의 회귀 변이체를 조사할 때 특히 유익할 수 있으며, 이러한 변이체는 집단-기반 분석으로는 쉽게 검출되지 않을 수 있다. 가족-기반 연관 연구의 주요 이점은 잠재적인 일률 손실에도 불구하고 인구 계층화로 인한 혼동 편향을 제어하는 것이다(Witte 등 American Journal of epidemiology(1999) 149(8): 693-705; Thomas 등 Cancer(2003) 97(8): 1894-1903).

[0011] 수십만 명의 비-식별화된 개체를 확인하고 시퀀싱하기 위한 대규모 시퀀싱 계획이 많이 존재한다(예를 들어, DiscovEHR, UK Biobank, 미국 정부의 미국 전체 (정밀 의학 계획의 일부), TOPMed, ExAC/gnomAD, 및 Dewey 등. Science(2016); 254, aaf6814; Sudlow 등 PLoS Med.(2015) 12, e1001779; Collins 등(2016) New England Journal of Medicine(2015) 372, 793-795; Lek 등 Nature(2016) 536, 285-291). 혈통은 단백질 시퀀싱 정보의 대규모 데이터세트에서 구성될 수 있으며, 이는 연구자가 형질 및 장애에 대한 유전가능성 및 유전적 모델을 결정하는 데 사용될 수 있다. 정확한 혈통 구조를 알면 질병 계승의 유전적 방식을 정확하게 식별하고 진정한 혈통 구조를 필요로 하거나 그로부터 이익을 얻는 강력한 유전-분석 도구를 활용할 수 있다. 그러나, 많은 강력한 가족-기반 분석을 배제하고 비-식별화된 건강 기록으로부터 정확한 혈통 기록을 직접 얻는 데 어려움이 있다.

[0012] 프리머스(PRIMUS) 및 클래퍼(CLAPPER)와 같은 도구를 사용하여 유전적 데이터로부터 직접 혈통 구조를 재구성하기 위해 밀접한 쌍 관계를 사용할 수 있다 (Staples 등 American Journal of Human Genetics(2014) 95, 553-564 및 Ko 및 Nielson. PLoS Genet.(2017) 13, e1006963). 추정된 관계 및 혈통은 매우 유용하지만, 추정된 관계 및 혈통 구조의 부정확성에 민감한 분석에서 상당한 통계적 불확실성을 가진 추정된 관계 및 혈통의 사용에 대한 우려가 존재한다.

[0013] 정밀 의학 코호트는 혈통 정보를 쉽게 갖지 못할 수 있지만, 유전적 데이터에서 직접 유익한 혈통을 얻어 전통적인 멘델 분석을 위한 대규모 코호트를 생성할 수 있다. 유전적 원인이 가족 단위 내에서 공유될 가능성이 더 높기 때문에, 관심있는 표현형을 가진 영향받는 개체들의 농축된 혈통을 식별하는 것은 이러한 표현형을 유도하는 인과적(회귀) 변이를 식별하는 데 사용할 수 있다. 혈통 농축 분석에 사용되는 영향받는 개체의 세트를 정의하는 것이 중요할 수 있다. 따라서, 혈통 농축을 가능하게 하는 방법이나 시스템이 필요하다. 이러한 농축된 혈통을 활용하면 관심있는 표현형을 가진 관련 참가자의 서브세트를 정의한 다음 이러한 서브세트를 검사하여 형질과 질병의 유전적 조작자(genetic driver)를 식별하는데 도움이 될 수 있다. 대규모로 전통적인 멘델 분석을 가능하게 하는 잠재적으로 유익한 혈통-표현형 쌍을 식별하기 위해 혈통 농축을 위한 개선된 생물정보학 도구에 대한 필요성이 남아 있다.

[0014] 농축된 혈통을 생성하기 위한 방법 및 시스템의 발견은 약물 발견 과학자가 정상적인 생리학 또는 질병의 원인에서 특정 단백질 및 그 변이체가 수행하는 중요한 역할을 이해하고 생화학적 및 생물학적으로 그 기능을 설명하도록 안내할 수 있다(Lele RJ Assoc. Physicians India(2003) 51: 373-380).

[0015] 본 명세서에 기술된 방법 및 시스템은 그러한 질병-유발 변이체(들)를 식별하도록 유도할 수 있는 농축된 혈통을 제공하여 약물 발견 노력 및 임상 조사 노력을 촉진할 것이다.

발명의 내용

[0016] 한 예시적인 측면에서, 본 명세서는 코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하고, 코호트에서 영향받거나 영향받지 않은 것으로 개체를 식별하고, 영향받는 개체 및 영향받지 않은 개체를 함유하는 농축된 혈통을 생성하는 방법을 제공한다.

[0017] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 방법은 혈통에 있는 개체를 영향받거나 영향받지 않은 것으로 식별하는 것을 포함할 수 있으며, 여기서 적어도 하나의 이진 형질(binary trait)을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질이 없는 개체는 영향받지 않은 것으로 식별된 다음 영향받은 개체와

영향받지 않은 개체의 패턴이 멘델의 계승 방식(예: 상염색체 우성, 상염색체 열성, x-연관 우성, x-연관 열성 또는 y-연관)과 일치하는지 평가한다. 일부 특정 예시적인 구현양태에서, 이진 형질은 질병, 징후 및 증상, 비정상적인 조사결과, 불만, 사회적 상황 및 부상이나 질병의 외부 원인에 대한 코드를 포함하는 세계 보건기구(WHO)의 의료 분류 목록인 질병 및 관련 건강 문제의 국제 통계 분류(ICD)를 사용하여 정의될 수 있다. ICD의 9판 또는 10판은 이진 형질을 정의하는 데 사용할 수 있다. 한 예시적인 구현양태에서, 특정 이진 형질에 대해 전자 건강 기록 데이터를 사용할 수 없는 개체 또는 특정 이진 형질에 대해 상충하거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 이진 형질의 부재 또는 존재와 관계없이 미지의 영향받는 것으로 결정될 수 있다.

[0018] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 방법은 혈통에 있는 개체를 영향받거나 영향받지 않은 것으로 식별하는 것을 포함할 수 있으며, 여기서 적어도 하나의 극도의 정량적 형질을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 극도의 정량적 형질이 없는 개체는 영향받지 않은 것으로 식별된 다음 영향받은 개체 및 영향받지 않은 개체의 패턴이 멘델의 계승 방식(예: 상염색체 우성, 상염색체 열성, x-연관 우성, x-연관 열성 또는 y-연관)과 일치하는지 평가한다. 장애의 조기 발병을 정의하기 위한 최대 연령 컷오프, 또는 형질의 정상 집단 측정 편차의 정의된 통계적 컷오프를 초과한(예: 집단 평균보다 2 표준 편차를 벗어난) 정량적 형질의 최소 또는 최대 또는 중앙 측정을 갖는 극도의 정량적 형질의 영향받은 개체가 있는지 여부를 정의하는 데 여러 매개 변수를 사용할 수 있다. 한 예시적인 구현양태에서, 특정 정량적 형질에 대해 전자 건강 기록 데이터를 사용할 수 없거나 특정 정량적 형질에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 정량적 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.

[0019] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 방법은 혈통에 있는 개체를 영향받거나 영향받지 않은 것으로 식별하는 것을 포함할 수 있으며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나가 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않은 것으로 식별된다. 이진 형질은 상기에서 기재된 바와 같이 정의된 ICD 코드일 수 있다. 상기에서 기재된 바와 같이 극도의 정량적 형질을 정의하기 위해 여러 매개 변수를 사용할 수 있다. 한 예시적인 구현양태에서, 특정 이진 형질, 정량적 형질 또는 이들의 조합에 대해 전자 건강 기록 데이터를 사용할 수 없거나 특정 이진 형질, 정량적 형질 또는 이들의 조합에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 정량적 형질의 부재 또는 존재와 관계없이 미지의 영향받는 것으로 결정될 수 있다.

[0020] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 방법은 혈통에 있는 개체를 영향받거나 영향받지 않은 것으로 식별하는 것을 포함할 수 있으며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나가 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향을 받지 않는 것으로 식별되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합은 둘 또는 그 이상의 유사하거나 상보적인 형질을 포함할 수 있다.

[0021] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 방법은 혈통에 있는 개체를 영향받거나 영향받지 않은 것으로 식별하는 것을 포함할 수 있으며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나가 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향을 받지 않는 것으로 식별되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합은 둘 또는 그 이상의 극도의 또는 흥미있는 형질의 교차를 취하는 것을 포함할 수 있다.

[0022] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 방법은 혈통에 있는 개체를 영향받은 것으로 식별하는 것을 포함할 수 있으며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 영향을 받는 것으로 결정된 개체를 외부 분석으로부터의 연관 결과의 영향을 받는 운반자로 정의한다.

[0023] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 방법은 코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하는 것을 포함한다. 시퀀싱 데이터에는 전체 게놈 시퀀싱 데이터, 엑솜 시퀀싱 데이터 또는 유전자형 데이터가 포함될 수 있다.

[0024] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 방법은 엑솜 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하는 것을 포함한다. 엑솜 시퀀싱 데이터를 기반으로 한 개체의 1급 네트워크는 다음과 같은 집단의 관련성을 활용하여 생성할 수 있다; 다수의 인간 피험자로부터 얻은 핵산 서열 샘플의 데이터세트에서 저-품질 서열 변이체를 제거하고, 하나 이상의 샘플 각각에 대해 조상 상위클래스 지정을 설정하고, 데이터세트에서 저-품질 샘플을 제거하고, 조상 상위클래스 내의 피험자에 대한 첫 번째 아이덴티티-바이-디센트(identity-by-

descent) 추정값을 생성하고, 피험자의 조상 상위클래스와 독립적인 피험자의 두 번째 아이덴티티-바이-디센트 추정값을 생성하고 및 두 번째 아이덴티티-바이-디센트 추정값 중 하나 이상을 기반으로 피험자를 주요한 1급 가족 네트워크로 클러스터링한다.

- [0025] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 방법은 코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하는 것을 포함하며, 여기서 코호트는 다수의 피험자를 포함하는 임의의 데이터셋을 포함할 수 있다.
- [0026] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 방법은 p-값을 기반으로 혈통을 농축하는 것을 추가로 포함한다. 농축에는 혈통 내에서 설립자의 모든 후손으로서 혈통의 "설립자 고정 가지" 또는 "가지"를 정의하고 가지가 이진 형질에 대해 농축되었는지 평가하기 위해 이항 테스트를 사용하는 것이 포함될 수 있다. 이진 형질은 상기에서 설명한대로 ICD를 사용하여 정의할 수 있다. 농축은 또한 혈통 내에서 설립자의 모든 후손으로서 혈통의 "설립자 고정 가지" 또는 "가지"를 정의하고, 가지는 극도의 정량적 형질에 대해 농축되었는지 평가하기 위해 t-테스트를 사용하는 것을 포함할 수 있다. 상기에서 설명한대로 극도의 정량적 형질을 정의하기 위해 여러 매개 변수를 사용할 수 있다. 또한, 농축에는 다중-테스트 p-값 컷오프를 적용하는 것도 포함될 수 있다.
- [0027] 한 예시적인 측면에서, 본 명세서는 코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하고, 코호트 내의 개체를 영향받거나 또는 영향받지 않은 것으로 식별하고, 영향받은 개체와 영향받지 않은 개체를 포함하는 하나 이상의 농축 혈통을 생성하고, 적어도 하나의 농축 혈통 내에서 및 혈통에 걸쳐서 공동 분리하는 변이체 형질 쌍을 식별하기 위해 분리 분석을 수행하고, 질병-유발 변이체를 식별하기 위해 변이체 형질 쌍을 분석함으로써 농축된 혈통을 생성하여 질병-유발 변이체를 식별하는 방법을 제공한다.
- [0028] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 혈통에 있는 개체를 영향받거나 영향받지 않은 것으로 식별하는 단계를 포함할 수 있으며, 여기서 적어도 하나의 이진 형질을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질이 없는 개체는 영향받지 않은 것으로 식별되고, 영향받은 개체 및 영향받지 않은 개체의 패턴이 멘델의 계승 방식(예: 상염색체 우성, 상염색체 열성, x-연관 우성, x-연관 열성 또는 y-연관)과 일치하는지 평가한다. 일부 특정 예시적인 구현양태에서, 이진 형질은 질병, 징후 및 증상, 비정상적인 조사결과, 불만, 사회적 상황 및 부상이나 질병의 외부 원인에 대한 코드를 포함하는 세계 보건기구(WHO)의 의료 분류 목록인 질병 및 관련 건강 문제의 국제 통계 분류(ICD)를 사용하여 정의될 수 있다. ICD의 9판 또는 10판은 이진 형질을 정의하는 데 사용할 수 있다. 한 예시적인 구현양태에서, 특정 이진 형질에 대해 전자 건강 기록 데이터를 사용할 수 없는 개체 또는 특정 이진 형질에 대해 상충하거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 이진 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.
- [0029] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 혈통에 있는 개체를 영향받거나 영향받지 않은 것으로 식별하는 단계를 포함할 수 있으며, 여기서 적어도 하나의 극도의 정량적 형질을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 극도의 정량적 형질이 없는 개체는 영향받지 않은 것으로 식별된 다음 영향받은 개체와 영향받지 않은 개체의 패턴이 멘델의 계승 방식(예: 상염색체 우성, 상염색체 열성, x-연관 우성, x-연관 열성 또는 y-연관)과 일치하는지 여부를 평가한다. 장애의 조기 발병을 정의하기 위한 형질의 정상 집단 측정 편차의 정의된 통계적 컷오프를 초과한(예: 집단 평균보다 2 표준 편차를 벗어난) 정량적 형질의 최소 또는 최대 또는 중앙 측정을 갖는 극도의 정량적 형질에 의해 영향받은 개체가 있는지 여부를 정의하는 데 여러 매개 변수를 사용할 수 있다. 한 예시적인 구현양태에서, 특정 정량적 형질에 대해 전자 건강 기록 데이터를 사용할 수 없거나 특정 정량적 형질에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 정량적 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.
- [0030] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 혈통에 있는 개체를 영향받거나 영향받지 않은 것으로 식별하는 것을 포함할 수 있으며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않은 것으로 식별된다. 이진 형질은 상기에서 설명한대로 정의된 ICD 코드일 수 있다. 상기에서 설명한대로 극단적인 정량적 형질을 정의하기 위해 여러 매개 변수를 사용할 수 있다. 한 예시적인 구현양태에서, 특정 이진 형질, 정량적 형질 또는 이들의 조합에 대해 전자 건강 기록 데이터를 사용할 수 없거나 특정 이진 형질, 정량적 형질 또는 이들의 조합에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 정량적 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.
- [0031] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 혈통에 있는 개체를 영향받거나 영향받지 않은 것으로 식별하는 것을 포함할 수 있으며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의

조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않은 것으로 식별되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합은 두 개 또는 그 이상의 유사하거나 상보적인 형질을 포함할 수 있다.

[0032] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 혈통에 있는 개체를 영향받거나 영향받지 않은 것으로 식별하는 것을 포함할 수 있으며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않은 것으로 식별되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합은 두 개 또는 그 이상의 극도의 또는 흥미로운 형질의 교차를 취하는 것을 포함할 수 있다.

[0033] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 혈통에 있는 개체를 영향받거나 영향받지 않은 것으로 식별하는 것을 포함할 수 있으며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 영향받은 것으로 결정된 개체를 외부 분석으로부터의 연관 결과의 영향을 받는 운반자로서 정의한다.

[0034] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 코호트의 시퀀싱 데이터에 기반하여 개체의 1급 네트워크를 생성하는 것을 포함한다. 시퀀싱 데이터에는 전체 게놈 시퀀싱 데이터, 엑솜 시퀀싱 데이터 또는 유전자형 데이터가 포함될 수 있다.

[0035] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 엑솜 시퀀싱 데이터에 기반하여 개체의 1급 네트워크를 생성하는 것을 포함한다. 엑솜 시퀀싱 데이터를 기반으로 한 개체의 1급 네트워크는 다음과 같은 집단의 관련성을 활용하여 생성할 수 있다: 다수의 인간 피험자로부터 얻은 핵산 서열 샘플의 데이터세트에서 저-품질 서열 변이체를 제거하고, 하나 이상의 샘플 각각에 대해 조상 상위클래스 지정을 설정하고, 데이터세트에서 저-품질 샘플을 제거하고, 조상 상위클래스 내의 피험자에 대한 첫 번째 아이덴티티-바이-디센트 추정값을 생성하고, 피험자의 조상 상위클래스와 독립적인 피험자의 두 번째 아이덴티티-바이-디센트 추정값을 생성하고 및 두 번째 아이덴티티-바이-디센트 추정값 중 하나 이상을 기반으로 피험자를 주요한 1급 가족 네트워크로 클러스터링한다.

[0036] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 코호트의 시퀀싱 데이터에 기반하여 개체의 1급 네트워크를 생성하는 것을 포함하며, 여기서 코호트는 다수의 피험자를 포함하는 임의의 데이터세트를 포함할 수 있다.

[0037] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 방법은 p-값을 기반으로 혈통을 농축하는 것을 추가로 포함한다. 농축에는 혈통 내에서 설립자의 모든 후손으로 혈통의 "설립자 고정 가지" 또는 "가지"를 정의하고 가지가 이진 형질에 대해 농축되었는지 평가하기 위해 이항 테스트를 사용하는 것이 포함될 수 있다. 이진 형질은 상기에서 설명한대로 ICD를 사용하여 정의할 수 있다. 농축은 또한 혈통 내에서 설립자의 모든 후손으로 혈통의 "설립자 고정 가지" 또는 "가지"를 정의하고, 가지는 극도의 정량적 형질에 대해 농축되었는지 평가하기 위해 t-테스트를 사용하는 것을 포함할 수 있다. 상기에서 설명한대로 극도의 정량적 형질을 정의하기 위해 여러 매개 변수를 사용할 수 있다. 또한, 농축에는 다중-테스트 p-값 컷오프를 적용하는 것도 포함될 수 있다.

[0038] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 혈통 내에서 영향받은 개체들과 공동 분리되는 변이체 형질 쌍을 식별하고, 표현형 분리에 기반하여 적어도 하나의 농축된 혈통을 찾는 것을 포함하는 분리 분석을 수행하는 것을 포함할 수 있다. 분리에는 우성 및 추가 분리 모델과 열성 분리 모델이 포함될 수 있다. 하나의 예시적인 구현양태에서, 우성 및 추가 분리 모델에 기반하여 적어도 하나의 농축된 혈통을 찾는 것은 하나의 가능한 구조 및 공통 조상의 적어도 3명의 영향받은 개체들을 갖는 혈통을 선택하는 것을 포함한다. 이는 거짓 양성을 감소시키기 위해 하나 이상의 관련된 영향받지 않는 개체들을 갖는 하나 이상의 농축된 혈통을 선택하는 것을 추가로 포함할 수 있다. 또 다른 예시적인 구현양태에서, 열성 분리 모델에 기반하여 적어도 하나의 농축된 혈통을 찾는 것은 하나의 가능한 구조 및 영향받지 않는 부모의 하나 이상의 영향받은 개체들을 갖는 혈통을 선택하는 것을 포함한다. 이는 거짓 양성을 감소시키기 위해 적어도 두명의 영향을 받은 형제자매를 갖는 적어도 하나의 농축된 혈통을 선택하는 것을 추가로 포함할 수 있다.

[0039] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 분리의 특정 유전적 모델을 형성하기 위해 분리 분석을 수행하는 것을 포함한다. 분리의 특정 유전적 모델은 분리의 우성 유전적 모델 또는 분리의 열성 유전적 모델을 포함할 수 있다. 또한, 분리의 특정 유전적 모델은 계승의 Y-연관, 다-인자 또는 미토콘드리아-연관 모드와 같은 다른 계승 모드에 기반한 분리의 유전적 모델을 포함할 수도 있다. 한 예시적인

구현양태에서, 질병-유발 변이체를 식별하는 방법은 분리의 우성 유전적 모델을 형성하기 위해 분리 분석을 수행하는 것을 포함하며, 여기서 질병-유발 변이체는 적어도 하나의 이진 형질, 극도의 정량적 형질, 또는 이들의 조합에 대해 영향받은 개체들로 분리한다. 한 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 분리의 열성 유전적 모델을 형성하기 위해 분리 분석을 수행하는 것을 포함하며, 여기서 질병-유발 변이체는 주어진 유전자에서 이중 대립 유전자 변이체 운반자인 영향받은 개체들과 분리되고, 부모에 대해 유전적 데이터를 사용할 수 있으면, 식별된 질병-유발 변이체에 대해 이형 접합이어야 한다.

- [0040] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 적어도 하나의 농축 혈통 내에서 및 혈통에 걸쳐서 공동 분리하는 변이체 형질 쌍을 식별하기 위해 분리 분석을 수행하는 것을 포함할 수 있다. 하나의 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 다수의 농축된 혈통 내에서 및 혈통에 걸쳐서 공동-분리하는 변이체 형질 쌍을 식별하기 위한 분리 분석을 포함한다.
- [0041] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 가족 구조에 포함되지 않은 관심 표현형에 대해 다른 영향받은 개체에서 변이체 또는 유전자 분리를 식별하기 위해 분리 분석을 수행하는 것을 포함할 수 있다.
- [0042] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 집단-규모 분석으로부터의 연관 결과를 갖는 교차 참조 변이체 및 형질을 포함하는 분리 분석을 수행하는 것을 포함할 수 있다.
- [0043] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 이전에 공지된 원인 변이체 및 유전자를 식별하기 위해 분리 분석을 수행하는 것을 포함 할 수 있다.
- [0044] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 지지 혈통/영향받은 개체의 수 및 후보 원인 변이체 및 유전자의 수에 의해 농축된 혈통을 우선순위화하는 것을 추가로 포함할 수 있다.
- [0045] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 변이체 형질 쌍을 분석하는 것을 포함할 수 있으며, 가족-기반 연관 분석을 보충하기에 충분한 가족 데이터로 영향받은 개체들의 세트를 식별하는 것을 추가로 포함할 수 있다.
- [0046] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 변이체 형질 쌍을 분석하는 것을 포함할 수 있으며, 혈통 및 표현형 정보에 기반하여 적절한 경우 전달 불균형 테스트(Transmission Disequilibrium Test, TDT) 또는 기타 분석을 수행하는 것을 포함할 수 있다.
- [0047] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 여러 생리적 장애에 대한 질병-유발 변이체를 식별하는 방법을 포함할 수 있다.
- [0048] 한 예시적인 측면에서, 본 명세서는 프로세서로 하여금 농축된 혈통을 생성하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체를 제공하며, 코호트의 엑스 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하고, 1급 네트워크에서 개체를 영향받거나 영향받지 않은 것으로 식별하고, 영향받거나 영향받지 않은 것으로의 지정을 포함하는 개체를 함유하는 적어도 하나의 농축된 혈통을 생성하는 것을 포함한다.
- [0049] 일부 예시적인 구현양태에서, 프로세서로 하여금 농축된 혈통을 생성하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하는 것을 포함하며, 여기서 적어도 하나의 이진 형질을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질이 없는 개체는 영향받지 않은 것으로 식별된 다음 영향받은 개체와 영향받지 않은 개체의 패턴이 멘델의 계승 방식(예: 상염색체 우성, 상염색체 열성, x-연관 우성, x-연관 열성 또는 y-연관)과 일치하는지 여부를 평가한다. 일부 특정 예시적인 구현양태에서, 이진 형질은 질병, 징후 및 증상, 비정상적인 조사결과, 불만, 사회적 상황 및 부상이나 질병의 외부 원인에 대한 코드를 포함하는 세계 보건기구(WHO)의 의료 분류 목록인 질병 및 관련 건강 문제의 국제 통계 분류(ICD)를 사용하여 정의될 수 있다. ICD의 9판 또는 10판은 이진 형질을 정의하는 데 사용할 수 있다. 한 예시적인 구현양태에서, 특정 이진 형질에 대해 전자 건강 기록 데이터를 사용할 수 없는 개체 또는 특정 이진 형질에 대해 상충하거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 이진 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.
- [0050] 일부 예시적인 구현양태에서, 프로세서로 하여금 농축된 혈통을 생성하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하는 것을 포함하며, 여기서 적어도 하나의 극도의 정량적 형질을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 극도의 정량적 형질이 없는 개체는 영향받지 않은 것으로 식별된 다음 영향받은 개체와 영향받지 않은 개체

의 패턴이 멘델의 계승 방식(예: 상염색체 우성, 상염색체 열성, x-연관 우성, x-연관 열성 또는 y-연관)과 일치하는지 여부를 평가한다. 장애의 조기 발병을 정의하기 위한 최대 연령 컷오프, 또는 형질의 정상 집단 측정 편차의 정의된 통계적 컷오프를 초과한(예: 집단 평균보다 2 표준 편차를 벗어난) 정량적 형질의 최소 또는 최대 또는 중앙 측정을 갖는 극도의 정량적 형질의 영향받은 개체가 있는지 여부를 정의하는 데 여러 매개 변수를 사용할 수 있다. 한 예시적인 구현양태에서, 특정 정량적 형질에 대해 전자 건강 기록 데이터를 사용할 수 없거나 특정 정량적 형질에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 정량적 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.

[0051] 일부 예시적인 구현양태에서, 프로세서로 하여금 농축된 혈통을 생성하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하는 것을 포함하며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않는 것으로 식별된다. 이진 형질은 상기에서 설명한대로 정의된 ICD 코드일 수 있다. 상기에서 설명한대로 극도의 정량적 형질을 정의하기 위해 여러 매개 변수를 사용할 수 있다. 한 예시적인 구현양태에서, 특정 이진 형질, 정량적 형질 또는 이들의 조합에 대해 전자 건강 기록 데이터를 사용할 수 없거나 특정 이진 형질, 정량적 형질 또는 이들의 조합에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 정량적 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.

[0052] 일부 예시적인 구현양태에서, 프로세서로 하여금 농축 혈통을 생성하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하는 것을 포함하며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않는 것으로 식별되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합은 두개 또는 그 이상의 유사하거나 상보적인 형질을 포함할 수 있다.

[0053] 일부 예시적인 구현양태에서, 프로세서로 하여금 농축 혈통을 생성하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하는 것을 포함하며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않는 것으로 식별되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합은 두개 또는 그 이상의 극도의 또는 흥미로운 형질의 교차를 취하는 것을 포함할 수 있다.

[0054] 일부 예시적인 구현양태에서, 프로세서로 하여금 농축된 혈통을 생성하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 코호트에서 개체가 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 갖는 경우 영향받은 개체로 식별하고 영향받은 것으로 결정된 개체를 외부 분석으로부터 연관 결과의 영향받은 운반자로서 정의하는 것을 추가로 포함한다.

[0055] 일부 예시적인 구현양태에서, 프로세서로 하여금 농축된 혈통을 생성하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하는 것을 포함한다. 시퀀싱 데이터는 전체 게놈 시퀀싱 데이터, 엑솜 시퀀싱 데이터, 또는 유전자형 데이터를 포함할 수 있다.

[0056] 일부 예시적인 구현양태에서, 프로세서로 하여금 농축된 혈통을 생성하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 엑솜 시퀀싱 데이터를 기반으로 한다. 엑솜 시퀀싱 데이터를 기반으로 한 개체의 1급 네트워크는 다음과 같은 집단의 관련성을 활용하여 생성할 수 있다: 다수의 인간 피험자로부터 얻은 핵산 서열 샘플의 데이터세트에서 저-품질 서열 변이체를 제거하고, 하나 이상의 샘플 각각에 대해 조상 상위클래스 지정을 설정하고, 데이터세트에서 저-품질 샘플을 제거하고, 조상 상위클래스 내의 피험자에 대한 첫 번째 아이덴티티-바이-디센트 추정값을 생성하고, 피험자의 조상 상위클래스와 독립적인 피험자의 두 번째 아이덴티티-바이-디센트 추정값을 생성하고 및 두 번째 아이덴티티-바이-디센트 추정치 중 하나 이상을 기반으로 피험자를 주요한 1급 가족 네트워크로 클러스터링한다.

[0057] 일부 예시적인 구현양태에서, 프로세서로 하여금 농축된 혈통을 생성하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 코호트의 시퀀싱 데이터를 기반으로 한 개체의 1급 네트워크를 생성하는 것을 포함하며, 여기서 코호트는 다수의 피험자를 포함하는 임의의 데이터세트를 포함할 수 있다.

- [0058] 일부 예시적인 구현양태에서, 프로세서로 하여금 농축 혈통을 생성하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 p-값을 기반으로 혈통을 농축하는 것을 추가로 포함한다. 농축에는 혈통 내에서 설립자의 모든 후손으로 혈통의 "설립자 고정 가지" 또는 "가지"를 정의하고 가지가 이진 형질에 대해 농축되었는지 평가하기 위해 이항 테스트를 사용하는 것이 포함될 수 있다. 이진 형질은 상기에서 설명한 대로 ICD를 사용하여 정의할 수 있다. 농축은 또한 혈통 내에서 설립자의 모든 후손으로 혈통의 "설립자 고정 가지" 또는 "가지"를 정의하고, 가지는 극도의 정량적 형질에 대해 농축되었는지 평가하기 위해 t-테스트를 사용하는 것을 포함할 수 있다. 상기에서 설명한대로 극도의 정량적 형질을 정의하기 위해 여러 매개 변수를 사용할 수 있다. 또한, 농축에는 다중-테스트 p-값 컷오프를 적용하는 것도 포함될 수 있다.
- [0059] 한 예시적인 측면에서, 본 명세서는 프로세서로 하여금 질병-유발 변이체를 식별하게 하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체를 제공하며, 코호트의 엑스 시퀀싱 데이터에 기반하여 개체의 1급 네트워크를 생성하고, 1급 네트워크의 개체를 영향받거나 영향받지 않는 것으로 식별하고, 영향받거나 영향받지 않는 것으로의 지정을 포함하는 개체를 함유하는 적어도 하나의 농축된 혈통을 생성하고, 적어도 하나의 농축된 혈통 내에서 및 혈통에 걸쳐서 공동 분리하는 변이체 형질 쌍을 식별하기 위해 분리 분석을 수행하고, 질병-유발 변이체를 결정하기 위해 변이체 형질 쌍을 분석하는 것을 포함한다.
- [0060] 한 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하게 하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하는 것을 포함하며, 여기서 적어도 하나의 이진 형질을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질이 없는 개체는 영향받지 않은 것으로 식별된 다음 영향받은 개체와 영향받지 않은 개체의 패턴이 멘델의 계승 방식(예: 상염색체 우성, 상염색체 열성, x-연관 우성, x-연관 열성 또는 y-연관)과 일치하는지 여부를 평가한다. 일부 특정 예시적인 구현양태에서, 이진 형질은 질병, 징후 및 증상, 비정상적인 조사결과, 불만, 사회적 상황 및 부상이나 질병의 외부 원인에 대한 코드를 포함하는 세계 보건기구(WHO)의 의료 분류 목록인 질병 및 관련 건강 문제의 국제 통계 분류(ICD)를 사용하여 정의될 수 있다. ICD의 9판 또는 10판은 이진 형질을 정의하는 데 사용할 수 있다. 한 예시적인 구현양태에서, 특정 이진 형질에 대해 전자 건강 기록 데이터를 사용할 수 없는 개체 또는 특정 이진 형질에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 이진 형질의 부재 또는 존재와 관계없이 미지의 영향받는 것으로 결정될 수 있다.
- [0061] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하는 것을 포함하며, 여기서 적어도 하나의 극도의 정량적 형질을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 극도의 정량적 형질이 없는 개체는 영향받지 않은 것으로 식별된 다음 영향받은 개체와 영향받지 않은 개체의 패턴이 멘델의 계승 방식(예: 상염색체 우성, 상염색체 열성, x-연관 우성, x-연관 열성 또는 y-연관)과 일치하는지 여부를 평가한다. 장애의 조기 발병을 정의하기 위한 최대 연령 컷오프, 또는 형질의 정상 집단 측정 편차의 정의된 통계적 컷오프를 초과한(예: 집단 평균보다 2 표준 편차를 벗어난) 정량적 형질의 최소 또는 최대 또는 중앙 측정을 갖는 극도의 정량적 형질의 영향받는 개체가 있는지 여부를 정의하는 데 여러 매개 변수를 사용할 수 있다. 한 예시적인 구현양태에서, 특정 정량적 형질에 대해 전자 건강 기록 데이터를 사용할 수 없거나 특정 정량적 형질에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 정량적 형질의 부재 또는 존재와 관계없이 미지의 영향받는 것으로 결정될 수 있다.
- [0062] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하는 것을 포함하며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않는 것으로 식별된다. 이진 형질은 상기에서 설명한대로 정의된 ICD 코드일 수 있다. 상기에서 설명한대로 극도의 정량적 형질을 정의하기 위해 여러 매개 변수를 사용할 수 있다. 한 예시적인 구현양태에서, 특정 이진 형질, 정량적 형질 또는 이들의 조합에 대해 전자 건강 기록 데이터를 사용할 수 없거나 특정 이진 형질, 정량적 형질 또는 이들의 조합에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 정량적 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.
- [0063] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하는 것을 포함하며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지

않는 것으로 식별되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합은 두개 또는 그 이상의 유사하거나 상보적인 형질을 포함할 수 있다.

[0064] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하는 것을 포함하며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않는 것으로 식별되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합은 두개 또는 그 이상의 극도의 또는 흥미로운 형질의 교차를 취하는 것을 포함할 수 있다.

[0065] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 코호트에서 개체가 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 갖는다면 영향받는 것으로 개체를 식별하고 영향받은 것으로 결정되는 개체를 외부 분석으로부터 연관 결과의 영향받는 운반자로서 정의하는 것을 추가로 포함한다.

[0066] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하는 것을 포함한다. 시퀀싱 데이터는 전체 게놈 시퀀싱 데이터, 엑솜 시퀀싱 데이터, 또는 유전자형 데이터를 포함할 수 있다.

[0067] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 엑솜 시퀀싱 데이터를 기반으로 한다. 엑솜 시퀀싱 데이터를 기반으로 한 개체의 1급 네트워크는 다음과 같은 집단의 관련성을 활용하여 생성할 수 있다: 다수의 인간 피험자로부터 얻은 핵산 서열 샘플의 데이터세트에서 저-품질 서열 변이체를 제거하고, 하나 이상의 샘플 각각에 대해 조상 상위클래스 지정을 설정하고, 데이터세트에서 저-품질 샘플을 제거하고, 조상 상위클래스 내의 피험자에 대한 첫 번째 아이덴티티-바이-디센트 추정값을 생성하고, 피험자의 조상 상위클래스와 독립적인 피험자의 두 번째 아이덴티티-바이-디센트 추정값을 생성하고 및 두 번째 아이덴티티-바이-디센트 추정값 중 하나 이상을 기반으로 피험자를 주요한 1급 가족 네트워크로 클러스터링한다.

[0068] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 코호트의 시퀀싱 데이터를 기반으로 한 개체의 1급 네트워크를 생성하는 것을 포함하며, 여기서 코호트는 다수의 피험자를 포함하는 임의의 데이터세트를 포함할 수 있다.

[0069] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 p-값을 기반으로 혈통을 농축하는 것을 추가로 포함한다. 농축에는 혈통 내에서 설립자의 모든 후손으로 혈통의 "설립자 고정 가지" 또는 "가지"를 정의하고 가지가 이진 형질에 대해 농축되었는지 평가하기 위해 이항 테스트를 사용하는 것이 포함될 수 있다. 이진 형질은 상기에서 설명한대로 ICD를 사용하여 정의할 수 있다. 농축은 또한 혈통 내에서 설립자의 모든 후손으로 혈통의 "설립자 고정 가지" 또는 "가지"를 정의하고, 가지는 극도의 정량적 형질에 대해 농축되었는지 평가하기 위해 t-테스트를 사용하는 것을 포함할 수 있다. 상기에서 설명한대로 극도의 정량적 형질을 정의하기 위해 여러 매개 변수를 사용할 수 있다. 또한 농축에는 다중-테스트 p-값 컷오프를 적용하는 것도 포함될 수 있다.

[0070] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 혈통 내에서 영향받은 개체들과 공동 분리되는 변이체 형질 쌍을 식별하고, 표현형 분리에 기반하여 적어도 하나의 농축된 혈통을 찾는 것을 포함하는 분리 분석을 수행하는 것을 포함할 수 있다. 분리에는 우성 및 추가 분리 모델과 열성 분리 모델이 포함될 수 있다. 하나의 예시적인 구현양태에서, 우성 및 추가 분리 모델에 기반하여 적어도 하나의 농축된 혈통을 찾는 것은 하나의 가능한 구조 및 공통 조상의 적어도 3명의 영향받은 개체를 갖는 혈통을 선택하는 것을 포함한다. 이는 거짓 양성을 감소시키기 위해 하나 이상의 관련된 영향받지 않는 개체를 갖는 하나 이상의 농축 혈통을 선택하는 것을 추가로 포함할 수 있다. 또 다른 예시적인 구현양태에서, 열성 분리 모델에 기반하여 적어도 하나의 농축된 혈통을 찾는 것은 하나의 가능한 구조 및 영향을 받지 않는 부모의 하나 이상의 영향받는 개체를 갖는 하나 이상의 혈통을 선택하는 것을 포함한다. 이는 거짓 양성을 감소시키기 위해 적어도 두명의 영향을 받은 형제자매를 갖는 적어도 하나의 농축된 혈통을 선택하는 것을 추가로 포함할 수 있다.

[0071] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령

어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 분리의 특정 유전적 모델을 형성하기 위해 분리 분석을 수행하는 것을 포함한다. 분리의 특정 유전적 모델은 분리의 우성 유전적 모델 또는 분리의 열성 유전적 모델을 포함할 수 있다. 또한, 분리의 특정 유전적 모델은 계승의 Y-연관, 다-인자 또는 미토콘드리아-연관 모드와 같은 다른 계승 모드에 기반한 분리의 유전적 모델을 포함할 수도 있다. 한 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 분리의 우성 유전적 모델을 형성하기 위해 분리 분석을 수행하는 것을 포함하며, 여기서 질병-유발 변이체는 적어도 하나의 이진 형질, 극도의 정량적 형질, 또는 이들의 조합에 대해 영향받은 개체로 분리한다. 한 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 분리의 열성 유전적 모델을 형성하기 위해 분리 분석을 수행하는 것을 포함하며, 여기서 질병-유발 변이체는 주어진 유전자에서 이중 대립 유전자 변이체 운반자인 영향받은 개체들과 분리되고 부모에 대해 유전적 데이터를 사용할 수 있으면, 식별된 질병-유발 변이체에 대해 이형 집합이어야 한다.

- [0072] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 적어도 하나의 농축 혈통 내에서 및 혈통에 걸쳐서 공동 분리하는 변이체 형질 쌍을 식별하기 위해 분리 분석을 수행하는 것을 포함할 수 있다. 하나의 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 다수의 농축된 혈통 내에서 및 혈통에 걸쳐서 공동-분리하는 변이체 형질 쌍을 식별하기 위한 분리 분석을 포함한다.
- [0073] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 가족 구조에 포함되지 않은 관심 표현형에 대해 다른 영향받은 개체에서 변이체 또는 유전자 분리를 식별하기 위해 분리 분석을 수행하는 것을 포함할 수 있다.
- [0074] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 집단-규모 분석으로부터의 연관 결과를 갖는 교차 참조 변이체 및 형질을 포함하는 분리 분석을 수행하는 것을 포함할 수 있다.
- [0075] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 이전에 공지된 원인 변이체 및 유전자를 식별하기 위해 분리 분석을 수행하는 것을 포함할 수 있다.
- [0076] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 지지 혈통/영향받은 개체의 수 및 후보 원인 변이체 및 유전자의 수에 의해 농축 혈통을 우선순위화하는 것을 추가로 포함할 수 있다.
- [0077] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 변이체 형질 쌍을 분석하는 것을 포함할 수 있으며, 가족-기반 연관 분석을 보증하기에 충분한 가족 데이터로 영향받은 개체의 세트를 식별하는 것을 추가로 포함할 수 있다.
- [0078] 일부 예시적인 구현양태에서, 프로세서로 하여금 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체는 전달 불균형 테스트(TDT) 또는 혈통 및 표현형 정보에 기반하여 적절한 경우 기타 분석을 수행하는 것을 포함하는 변이체 형질 쌍을 분석하는 것을 포함할 수 있다.
- [0079] 일부 예시적인 구현양태에서, 프로세서로 하여금 여러 생리적 장애에 대한 질병-유발 변이체를 식별하는 방법을 수행하게 하기 위한 명령어를 저장하는 비-일시적 컴퓨터 판독가능 매체를 포함할 수 있다.
- [0080] 한 예시적인 구현양태에서, 본 명세서는 농축된 혈통을 생성하기 위한 시스템을 제공하며, 상기 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하고, 1급 네트워크에서 개체를 영향받거나 영향받지 않는 것으로 식별하고, 영향받거나 영향받지 않는 것으로의 지정을 포함하는 개체를 함유하는 적어도 하나의 농축된 혈통을 생성하도록 구성된다.
- [0081] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하도록 구성되며, 여기서 적어도 하나의 이진 형질을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질이 없는 개체는 영향받지 않는 것으로 식별된 다음 영향받은 개체와 영향받지 않는 개체의 패턴이 멘델의 계승 방식(예: 상염색체 우성, 상염색체 열성, x-연관 우성, x-연관 열성 또는 y-연관)과 일치하는지 여부를 평가한다. 일부 특정 예시적인 구현양태에서, 이진 형질은 질병, 징후 및 증상, 비정상적인 조사결과, 불만, 사회적 상황 및 부상

이나 질병의 외부 원인에 대한 코드를 포함하는 세계 보건기구(WHO)의 의료 분류 목록인 질병 및 관련 건강 문제의 국제 통계 분류(ICD)를 사용하여 정의될 수 있다. ICD의 9판 또는 10판은 이진 형질을 정의하는 데 사용할 수 있다. 한 예시적인 구현양태에서, 특정 이진 형질에 대해 전자 건강 기록 데이터를 사용할 수 없는 개체 또는 특정 이진 형질에 대해 상충하거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 이진 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.

[0082] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하도록 구성되며, 여기서 적어도 하나의 극도의 정량적 형질을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 극도의 정량적 형질이 없는 개체는 영향받지 않은 것으로 식별된 다음 영향받은 개체와 영향받지 않은 개체의 패턴이 멘델의 계승 방식(예: 상염색체 우성, 상염색체 열성, x-연관 우성, x-연관 열성 또는 y-연관)과 일치하는지 여부를 평가한다. 장애의 조기 발병을 정의하기 위한 최대 연령 컷오프, 또는 형질의 정상 집단 측정 편차의 정의된 통계적 컷오프를 초과한(예: 집단 평균보다 2 표준 편차를 벗어난)정량적 형질의 최소 또는 최대 또는 중앙 측정을 갖는 극도의 정량적 형질의 영향받은 개체가 있는지 여부를 정의하는 데 여러 매개 변수를 사용할 수 있다. 한 예시적인 구현양태에서, 특정 정량적 형질에 대해 전자 건강 기록 데이터를 사용할 수 없거나 특정 정량적 형질에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 정량적 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.

[0083] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하도록 구성되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않은 것으로 식별된다. 이진 형질은 상기에서 설명한대로 정의된 ICD 코드일 수 있다. 상기에서 설명한대로 극도의 정량적 형질을 정의하기 위해 여러 매개 변수를 사용할 수 있다. 한 예시적인 구현양태에서, 특정 이진 형질, 정량적 형질 또는 이들의 조합에 대해 전자 건강 기록 데이터를 사용할 수 없거나 특정 이진 형질, 정량적 형질 또는 이들의 조합에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 정량적 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.

[0084] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하도록 구성되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않은 것으로 식별되며, 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합은 두개 또는 그 이상의 유사하거나 상보적인 형질을 포함할 수 있다.

[0085] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하도록 구성되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않은 것으로 식별되며, 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합은 두개 또는 그 이상의 극도의 또는 흥미있는 형질의 교차를 취하는 것을 포함할 수 있다.

[0086] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 코호트에서 개체가 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 갖는다면 영향받은 것으로 개체를 식별하도록 구성되며, 영향받았다고 결정되는 개체를 외부 분석으로부터 연관 결과의 영향받는 운반자로서 정의한다.

[0087] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하도록 구성된다. 시퀀싱 데이터는 전체 게놈 시퀀싱 데이터, 엑솜 시퀀싱 데이터, 또는 유전자형 데이터를 포함할 수 있다.

[0088] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 엑솜 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하도록 구성된다. 엑솜 시퀀싱 데이터를 기반으로 한 개체의 1급 네트워크는 다음과 같은 집단의 관련성을 활용하여 생성할 수 있다: 다수의 인간 피험자로부터 얻은 핵산 서열 샘플의 데이터세트에서 저-품질 서열 변이를 제거하고, 하나 이

상의 샘플 각각에 대해 조상 상위클래스 지정을 설정하고, 데이터세트에서 저-품질 샘플을 제거하고, 조상 상위 클래스 내의 피험자에 대한 첫 번째 아이덴티티-바이-디센트 추정값을 생성하고, 피험자의 조상 상위클래스와 독립적인 피험자의 두 번째 아이덴티티-바이-디센트 추정값을 생성하고 및 두 번째 아이덴티티-바이-디센트 추정값 중 하나 이상을 기반으로 피험자를 주요한 1급 가족 네트워크로 클러스터링한다.

[0089] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하도록 구성되며, 여기서 코호트는 다수의 피험자를 포함하는 임의의 데이터세트를 포함할 수 있다.

[0090] 일부 예시적인 구현양태에서, 농축된 혈통을 생성하는 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 p-값을 기반으로 혈통을 농축하는 것을 추가로 포함하도록 구성된다. 농축은 혈통 내에서 설립자의 모든 후손으로 혈통의 "설립자 고정 가지" 또는 "가지"를 정의하고 가지가 이진 형질에 대해 농축되었는지 평가하기 위해 이항 테스트를 사용하는 것을 포함할 수 있다. 이진 형질은 상기에서 설명한대로 ICD를 사용하여 정의할 수 있다. 농축은 또한 혈통 내에서 설립자의 모든 후손으로 혈통의 "설립자 고정 가지" 또는 "가지"를 정의하고, 가지가 극도의 정량적 형질에 대해 농축되었는지 평가하기 위해 t-테스트를 사용하는 것을 포함할 수 있다. 상기에서 설명한대로 극도의 정량적 형질을 정의하기 위해 여러 매개 변수를 사용할 수 있다. 또한, 농축에는 다중-테스트 p-값 컷오프를 적용하는 것도 포함될 수 있다.

[0091] 한 예시적인 구현양태에서, 본 명세서는 질병-유발 변이체를 식별하기 위한 시스템을 제공하며, 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하고, 1급 네트워크에서 개체를 영향받거나 영향받지 않는 것으로 식별하고, 영향받거나 영향받지 않는 것으로의 지정을 포함하는 개체를 함유하는 적어도 하나의 농축된 혈통을 생성하도록 구성된다.

[0092] 한 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하도록 구성되며, 여기서 적어도 하나의 이진 형질을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질이 없는 개체는 영향받지 않은 것으로 식별된 다음 영향받은 개체와 영향받지 않은 개체의 패턴이 멘델의 계승 방식(예: 상염색체 우성, 상염색체 열성, x-연관 우성, x-연관 열성 또는 y-연관)과 일치하는지 여부를 평가한다. 일부 특정 예시적인 구현양태에서, 이진 형질은 질병, 징후 및 증상, 비정상적인 조사결과, 불만, 사회적 상황 및 부상이나 질병의 외부 원인에 대한 코드를 포함하는 세계 보건기구(WHO)의 의료 분류 목록인 질병 및 관련 건강 문제의 국제 통계 분류(ICD)를 사용하여 정의될 수 있다. ICD의 9판 또는 10판은 이진 형질을 정의하는 데 사용할 수 있다. 한 예시적인 구현양태에서, 특정 이진 형질에 대해 전자 건강 기록 데이터를 사용할 수 없는 개체 또는 특정 이진 형질에 대해 상충하거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 이진 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.

[0093] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하도록 구성되며, 여기서 적어도 하나의 극도의 정량적 형질을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 극도의 정량적 형질이 없는 개체는 영향받지 않은 것으로 식별된 다음 영향받은 개체와 영향받지 않은 개체의 패턴이 멘델의 계승 방식(예: 상염색체 우성, 상염색체 열성, x-연관 우성, x-연관 열성 또는 y-연관)과 일치하는지 여부를 평가한다. 장애의 조기 발병을 정의하기 위한 최대 연령 컷오프, 또는 형질의 정상 집단 측정 편차의 정의된 통계적 컷오프를 초과한(예: 모집단 평균보다 2 표준 편차를 벗어나는) 정량적 형질의 최소 또는 최대 또는 중앙 측정을 갖는 극도의 정량적 형질의 영향받은 개체가 있는지 여부를 정의하는 데 여러 매개 변수를 사용할 수 있다. 한 예시적인 구현양태에서, 특정 정량적 형질에 대해 전자 건강 기록 데이터를 사용할 수 없거나 특정 정량적 형질에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 정량적 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.

[0094] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하도록 구성되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않는 것으로 식별된다. 이진 형질은 상기에서 설명한대로 정의된 ICD 코드일 수 있다. 상기에서 설명한대로 극도의 정량적 형질을 정의하기 위해 여러 매개 변수를 사용할 수 있다. 한 예시적인 구현양태에서, 특정 이진 형질, 정

량적 형질 또는 이들의 조합에 대해 전자 건강 기록 데이터를 사용할 수 없거나 특정 이진 형질, 정량적 형질 또는 이들의 조합에 대해 상충되거나 신뢰할 수 없는 데이터가 있는 개체는 의료 기록에서 특정 정량적 형질의 부재 또는 존재와 관계없이 미지의 영향받은 것으로 결정될 수 있다.

[0095] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하도록 구성되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않는 것으로 식별되고, 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합은 두개 또는 그 이상의 유사하거나 상보적인 형질을 포함할 수 있다.

[0096] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 혈통의 개체가 영향받았는지 또는 영향받지 않았는지 식별하도록 구성되며, 여기서 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 가진 개체는 영향받은 것으로 식별되고 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합이 없는 개체는 영향받지 않는 것으로 식별되고, 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합은 두개 또는 그 이상의 극도의 또는 흥미있는 형질의 교차를 취하는 것을 포함할 수 있다.

[0097] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 개체가 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 갖는다면 코호트에서 영향받은 것으로 개체를 식별하도록 구성되고 영향받았다고 결정되는 개체를 외부 분석으로부터 연관 결과의 영향받는 운반자로서 정의하도록 구성된다.

[0098] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 코호트의 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하도록 구성된다. 시퀀싱 데이터는 전체 게놈 시퀀싱 데이터, 엑솜 시퀀싱 데이터, 또는 유전자형 데이터를 포함할 수 있다.

[0099] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 엑솜 시퀀싱 데이터를 기반으로 개체의 1급 네트워크를 생성하도록 구성된다. 엑솜 시퀀싱 데이터를 기반으로 한 개체의 1급 네트워크는 다음과 같은 집단의 관련성을 활용하여 생성될 수 있다: 다수의 인간 피험자로부터 얻은 핵산 서열 샘플의 데이터세트에서 저-품질 서열 변이를 제거하고, 하나 이상의 샘플 각각에 대해 조상 상위클래스 지정을 설정하고, 데이터세트에서 저-품질 샘플을 제거하고, 조상 상위클래스 내의 피험자에 대한 첫 번째 아이덴티티-바이-디센트 추정값을 생성하고, 피험자의 조상 상위클래스와 독립적인 피험자의 두 번째 아이덴티티-바이-디센트 추정값을 생성하고 및 두 번째 아이덴티티-바이-디센트 추정값 중 하나 이상을 기반으로 피험자를 주요한 1급 가족 네트워크로 클러스터링한다.

[0100] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 코호트의 시퀀싱 데이터를 기반으로 한 개체의 1급 네트워크를 생성하도록 구성되며, 여기서 코호트는 다수의 피험자를 포함하는 임의의 데이터세트를 포함할 수 있다.

[0101] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 p-값을 기반으로 혈통을 농축하는 것을 추가로 포함하도록 구성된다. 농축은 혈통 내에서 설립자의 모든 후손으로 혈통의 "설립자 고정 가지" 또는 "가지"를 정의하고 가지가 이진 형질에 대해 농축되었는지 평가하기 위해 이항 테스트를 사용하는 것을 포함할 수 있다. 이진 형질은 상기에서 설명한대로 ICD를 사용하여 정의할 수 있다. 농축은 또한 혈통 내에서 설립자의 모든 후손으로 혈통의 "설립자 고정 가지" 또는 "가지"를 정의하고, 가지가 극도의 정량적 형질에 대해 농축되었는지 평가하기 위해 t-테스트를 사용하는 것을 포함할 수 있다. 상기에서 설명한대로 극도의 정량적 형질을 정의하기 위해 여러 매개 변수를 사용할 수 있다. 또한, 농축에는 다중-테스트 p-값 컷오프를 적용하는 것도 포함될 수 있다.

[0102] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 혈통 내에서 영향받은 개체들과 공동 분리되는 변이체 형질 쌍을 식별하고, 표현형 분리에 기반하여 적어도 하나의 농축된 혈통을 찾는 것을 포함하는 분리 분석을 수행하도록 구성된다. 분리에는 우성 및 추가 분리 모델과 열성 분리 모델이 포함될 수 있다. 하나의 예시적인 구현양태에서, 우성 및 추가 분리 모델에 기반하여 적어도 하나의 농축된 혈통을 찾는 것은 하나의 가능한 구조 및 공통 조상

의 적어도 3명의 영향받은 개체를 갖는 혈통을 선택하는 것을 포함한다. 이는 거짓 양성을 감소시키기 위해 하나 이상의 관련된 영향받지 않은 개체를 갖는 하나 이상의 농축 혈통을 선택하는 것을 추가로 포함할 수 있다. 또 다른 예시적인 구현양태에서, 열성 분리 모델에 기반하여 적어도 하나의 농축된 혈통을 찾는 것은 하나의 가능한 구조 및 영향받지 않은 부모의 하나 이상의 영향받은 개체를 갖는 하나 이상의 혈통을 선택하는 것을 포함한다. 이는 거짓 양성을 감소시키기 위해 적어도 2명의 영향받은 형제자매를 갖는 적어도 하나의 농축된 혈통을 선택하는 것을 추가로 포함할 수 있다.

[0103] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 분리 분석을 수행하여 분리의 특정 유전적 모델을 형성하도록 구성된다. 분리의 특정 유전적 모델은 분리의 우성 유전적 모델 또는 분리의 열성 유전적 모델을 포함할 수 있다. 또한, 분리의 특정 유전적 모델은 계승의 Y-연관, 다-인자 또는 미토콘드리아-연관 모드와 같은 다른 계승 모드에 기반한 분리의 유전적 모델을 포함할 수도 있다. 한 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 분리 분석을 수행하여 분리의 우성 유전적 모델을 형성하는 것을 포함하며, 여기서 질병-유발 변이체는 적어도 하나의 이진 형질, 극도의 정량적 형질, 또는 이들의 조합에 대해 영향받은 개체로 분리한다. 한 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 분리 분석을 수행하여 분리의 열성 유전적 모델을 형성하는 것을 포함하며, 여기서 질병-유발 변이체는 주어진 유전자에서 이중 대립 유전자 변이체 운반자인 영향받은 개체들과 분리되고 부모에 대해 유전적 데이터를 사용할 수 있으면, 식별된 질병-유발 변이체에 대해 이형 접합이어야 한다.

[0104] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 적어도 하나의 농축 혈통 내에서 및 혈통에 걸쳐서 공동 분리하는 변이체 형질 쌍을 식별하기 위해 분리 분석을 수행하도록 구성된다. 하나의 예시적인 구현양태에서, 질병-유발 변이체를 식별하는 방법은 다수의 농축된 혈통 내에서 및 혈통에 걸쳐서 공동-분리하는 변이체 형질 쌍을 식별하기 위한 분리 분석을 포함한다.

[0105] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 가족 구조에 포함되지 않은 관심 표현형에 대해 다른 영향받은 개체에서 변이체 또는 유전자 분리를 식별하기 위해 분리 분석을 수행하도록 구성된다.

[0106] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 집단-규모 분석으로부터의 연관 결과를 갖는 교차 참조 변이체 및 형질을 포함하는 분리 분석을 수행하도록 구성된다.

[0107] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 이전에 공지된 원인 변이체 및 유전자를 식별하기 위해 분리 분석을 수행하도록 구성된다.

[0108] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 지지 혈통/영향받은 개체의 수 및 후보 원인 변이체 및 유전자의 수에 의해 농축된 혈통을 우선순위화하도록 구성된다.

[0109] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 변이체 형질 쌍을 분석하도록 구성되고, 가족-기반 연관 분석을 보충하기 위해 충분한 가족 데이터를 가진 영향받은 개체의 세트를 식별하는 것을 추가로 포함할 수 있다.

[0110] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 여기서 프로세서는 전달 불균형 테스트(TDT) 또는 혈통 및 표현형 정보에 기반하여 적절한 경우 기타 분석을 수행하는 것을 포함하는 변이체 형질 쌍을 분석하도록 구성된다.

[0111] 일부 예시적인 구현양태에서, 질병-유발 변이체를 식별하기 위한 시스템은 데이터 프로세서 및 데이터 프로세서와 연결된 메모리를 포함하며, 프로세서는 다수의 생리학적 장애에 대해 질병 유발 변이체를 식별하도록 구성된다.

[0112] 본원에 기술된 방법 및 시스템은 (i) 질병을 유발하는 분자 메커니즘에 대한 더 나은 이해를 제공하고, (ii) 질병의 더 나은 분류 및 더 나은 관리로 이어질 수 있으며, (iii) 관련 유전자 변이와 관련된 차등 대사의 식별을 제공할 수 있으며(신약 개발의 표적으로 암세포에서 변경된 대사과 관련된 중요한 효소 또는 단백질 또는 수용체를 사용하여), (iv) 암과 같은 질병에 대한 정제된 클래스 예측을 제공하여 향후 임상 경과 및 생존을 예측하

는 데 도움이 될 수 있으며, (v) 질병을 유발하는 유전적 결함을 식별하여 유전자 치료 설계할 수 있다(바람직 하지만 결핍된 유전자의 증가 또는 유해한 유전자 차단(안티센스 올리고리보뉴클레오타이드 또는 전사 인자 미끼, 또는 특정 앵타머를 통해)).

도면의 간단한 설명

[0113]

도 1은 혈통 농축을 수행하기 위한 본 발명의 예시적인 구현양태의 흐름도이다.

도 2는 혈통 농축을 수행하기 위한 본 발명의 예시적인 구현양태의 흐름도이다.

도 3은 예시적인 운영 환경이다.

도 4는 기재된 방법을 수행하도록 구성된 복수의 시스템 구성요소를 도시한다.

도 5는 예시적인 구현양태에 따라 확인된 DiscovEHR 코호트로부터의 처음 92K 서열화된 개체에 대한 IBD0 대 IDB1를 도시한다.

도 6은 원발성 혈전성향증 표현형(Phe10_D685, ICD10CM D68.5)에 대한 DiscovEHR 코호트로부터의 몇 가지 농축 된 혈통을 도시하며, 여기서 혈통 농축은 예시적인 구현양태에 따라 수행된다.

도 7a 및 7b는 유전성 출혈성 모세혈관확장증 표현형(Phe10_I780, ICD10CM I78.0)에 대한 두 가지 농축 혈통을 도시하며, 여기서 혈통 농축은 예시적인 구현양태에 따라 수행된다.

도 8은 유전성 출혈성 모세혈관확장증 표현형(Phe10_I780, ICD10CM I78.0)에 대한 변이체의 분리를 나타내는 농축된 혈통을 포함하는 DiscovEHR 코호트의 혈통을 도시하며, 여기서 혈통 농축 및 분리 분석은 예시적인 구현양태에 따라 수행된다.

도 9는 폐기종 표현형에 대한 DiscovEHR 코호트로부터의 여러 농축 혈통을 도시하며, 여기서 혈통 농축은 예시적인 구현양태에 따라 수행된다.

도 10은 신장 이식 표현형(Phe9_V420, ICD9CM V42.0)에 대한 DiscovEHR 코호트로부터의 농축 혈통을 도시하며, 여기서 혈통 농축은 예시적인 구현양태에 따라 수행된다.

도 11은 말기 신장 질환 표현형(Phe9_5856, ICD9CM 585.6)에 대한 DiscovEHR 코호트로부터의 여러 농축 혈통을 도시하며, 여기서 혈통 농축은 예시적인 구현양태에 따라 수행된다.

도 12는 유전성 운동 및 감각 신경병증 표현형(샤르코-마리-투스 질환, Charcot-Marie-Tooth Disease)(Phe10_G600, ICD10CM G60.0) 표현형에 대한 DiscovEHR 코호트로부터의 농축 혈통을 도시한다.

도 13은 다양한 조직에서 인코딩된 트로포미오신 2(*tropomyosin 2*, TMP2) 유전자의 백만 당 전사(transcripts per million, TPM)의 유전자 발현 데이터를 보여주는 차트이다.

도 14는 양극성 장애에 대한 DiscovEHR 코호트로부터의 농축 혈통을 도시하며, 여기서 혈통 농축 및 분리 분석은 예시적인 구현양태에 따라 수행된다.

도 15는 다양한 조직에서 인코딩된 염색체 20 오픈 리딩 프레임 203 (*C20orf203*)의 백만 당 전사(TPM)의 유전자 발현 데이터를 보여주는 차트이다.

도 16은 양극성 장애 표현형에 대한 DiscovEHR 코호트로부터의 농축 혈통을 도시하며, 여기서 혈통 농축은 예시적인 구현양태에 따라 수행된다.

도 17은 양극성 장애 표현형에 대한 DiscovEHR 코호트로부터의 농축 혈통을 도시하며, 여기서 혈통 농축은 예시적인 구현양태에 따라 수행된다.

도 18은 양극성 장애 표현형에 대한 DiscovEHR 코호트로부터의 농축 혈통을 도시하며, 여기서 혈통 농축은 예시적인 구현양태에 따라 수행된다.

도 19는 다양한 조직에서 마이크로세팔린 1(*MCPH 1*)의 백만 당 전사(TPM)의 유전자 발현 데이터를 보여주는 차트이다.

도 20은 가족성 지중해빈혈 표현형에 대한 DiscovEHR 코호트로부터의 농축 된 혈통을 도시하며, 여기서 혈통 농축은 예시적인 구현양태에 따라 수행된다.

도 21은 알칼리 포스파타제 외래환자 중심 경향 값에 대한 DiscovEHR 코호트로부터의 농축 혈통을 도시하며, 여기서 혈통 농축은 예시적인 구현양태에 따라 수행된다.

발명을 실시하기 위한 구체적인 내용

- [0114] 용어 "하나"는 "적어도 하나"를 의미하는 것으로 이해되어야 하며; 용어 "약" 및 "대략"은 당업자에 의해 이해되는 바와 같이 표준 변형을 허용하는 것으로 이해되어야 하며; 범위가 제공되는 경우 끝점이 포함된다.
- [0115] 가족-기반 연관 연구는 병원 또는 질병 등록 기관에서 받은 사례와 함께 사례-대조군 디자인을 사용한다. 대조군은 관련이 없거나(예: 집단 또는 병원/레지스트리 기반), 사례의 가족 구성원(예: 부모 또는 형제 자매)일 수 있다. 사례 대 대조군에서 주어진 대립 유전자의 발생을 비교하여 유전자와 질병 사이에 "연관"이 존재하는지 확인한다. 대규모 단일-뉴클레오타이드 다형성(SNP) 유전형분석이 가능해짐에 따라 연관 연구가 점점 보편화되고 있으며 초점을 맞춘 후보 유전자 연구에서 게놈 전반의 연관 연구로 빠르게 확장되고 있다.
- [0116] 차세대 시퀀싱 전략의 출현은 이러한 질병의 유전적 결함을 밝혀낼 가능성을 높여 주었다. 전체 게놈(약 30억 염기쌍)은 현재 며칠 동안 시퀀싱할 수 있으며 비용이 빠르게 감소하여 일상적인 연구 도구로 접근하게 할 수 있다. 엑솜 시퀀싱으로 언급되는 게놈의 단백질 코딩 부분을 시퀀싱하는 것은 질병을 유발하는 유전자를 찾는 데 훨씬 더 효율적이며, 엑솜은 게놈의 작은 부분(약 38Mb)만을 나타내고 엑손은 대부분의 멘델 유전자의 알려진 돌연변이를 보유하기 때문이다(Albert 등 Nature Methods(2007) 4:903-905; Gnirke 등 Nature Biotechnology(2009) 27: 182-189; Hodges 등 Nature Genetics(2007) 9: 1522-1527); Majewski 등 Journal of Medical Genetics(2011) 48: 580-589). 따라서, 엑솜 시퀀싱은 후보 유전자 또는 경로에 대한 사전 지식 없이도 유전적 원인이 의심되는 질환에서 돌연변이를 검색하는 데 매우 적합하다.
- [0117] 많은 대규모 인간 시퀀싱 연구는 변형 발견 및 정밀 의학을 촉매하기 위해 전자 건강 기록(electronic health records, EHR) 및 게놈 서열 데이터를 결합하는 것을 목표로 표현형-풍부한 EHR을 수반하는 통합 건강 관리 집단으로부터 샘플을 수집한다. 이러한 프로젝트의 데이터는 형질과 질병의 특정 유전적 조작자(genetic driver)를 식별하는 데 사용될 수 있다.
- [0118] 사례 및 대조군이 집단 계층화를 유발하는 다양한 대립 유전자 빈도를 갖는 상이한 출처 집단에서 나온 경우 가짜 연관성이 검출될 수 있다(Cardon 및 Palmer. Lancet(2003) 361(9357): 598-604). 이러한 혼란으로 인해 얼마나 많은 편견이 발생할 수 있는지에 대한 논쟁이 있다(Wacholder 등 Cancer Epidemiology, Biomarkers & Prevention(2002) 11(6): 513-520; Thomas 및 Witte. Cancer Epidemiology, Biomarkers & Prevention(2002) 11(6): 502-512; Gorroochurn 등 Human Heredity(2004) 58(1): 40-48). 가족-기반 연구 설계를 사용하여 인구 계층화를 피할 수 있다. 부모와 그 자손 또는 형제 자매를 연구할 때, 각 가족 내의 사례와 대조군은 동일한 출처 집단에서 발생한다. 일반적인 가족-기반 사례-대조군 설계는 부모 트리오(예: 전송 불균형 테스트(TDT) 접근 방식) 및 형제자매 대조군이다. 또한 다른 친척(예: 사촌)을 연구하거나 동시에 많은 수의 다른 가족 구성원을 연구할 수 있다.
- [0119] 대규모 코호트 내에서 가족을 식별하는 것은 가족-기반 유전 연구에 적합하도록 주어진 형질에 대해 충분한 정보를 제공하는 영향받은 개체로 구성된 혈통을 식별하는 것을 포함한다. 혈통은 가족 내에서 주어진 관심 표현형과 공동 분리되는 잠재적 중등 내지 대형 효과의 희귀 변이체를 조사할 때 특히 유용하다. 이러한 혈통은 관심있는 표현형을 가진 관련 참가자의 서브셋을 정의하는 데 도움이 될 수 있으며 이러한 서브셋을 검사하여 형질 및 질병의 유전적 조작자를 식별할 수 있다.
- [0120] 본 명세서는 적어도 부분적으로, 다수의 피험자의 게놈 샘플 데이터세트 내에서 개체의 1급 네트워크에 관한 정보가 다른 것들 중에서 희귀한 유전적 변이와 질병 사이의 연관성을 조사할 수 있다는 인식에 기초한다.
- [0121] 본원에 기술된 방법은 게놈 샘플의 다양한 유형의 데이터세트에 적용될 수 있다. 데이터세트 유형의 비-제한적인 예에는 단일-의료-네트워크 집단; 다중-의료-네트워크 집단; 인종적, 문화적 또는 사회적으로 동질적이거나 이질적인 집단; 혼합-연령 집단 또는 연령 측면에서 동질적인 집단; 지리적으로 집중되거나 분산된 집단; 또는 이들의 조합을 포함한다. 데이터세트에는 다양한 유형의 유전적 변이가 있을 수 있다. 평가될 수 있는 유전적 변이 유형의 비-제한적인 예에는 점 돌연변이, 삽입, 결실, 역위, 복제 및 다량체화가 포함된다. 유전적 변이가 얻어질 수 있는 수단의 비-제한적인 예는 다음 단계를 포함한다:
- [0122] - 샘플 준비 및 시퀀싱(Dewey 등(2016), Science 354, aaf6814-1 내지 aaf6814-10);
- [0123] - 시퀀싱이 완료되면 각 시퀀싱 실행의 원시 데이터를 로컬 버퍼 저장소에 수집하고 자동화 분석을 위해

DNAexus 플랫폼(Reid 등(2014); BMC Bioinformatics 15, 30)에 업로드할 수 있다.

- [0124] - 샘플-레벨 읽기 파일은 카사바(CASAVA)(Illumina Inc., San Diego, CA)를 사용하여 생성할 수 있으며, BWA-
mem(mem)을 사용하여 GRCh38에 정렬할 수 있다(Li 및 Durbin (2009); Bioinformatics 25, 1754-176;
Li(2013);arXiv q-bio.GN).
- [0125] - 결과적인 BAM 파일은 GATK(McKenna 등 (2010); Genome Res. 20, 1297-1303) 및 피카드(Picard)를 사용하여
처리되어 중복을 정렬하고, 중복을 표시하고, 추정 인텔 주변에서 읽기의 로컬 재정렬을 수행할 수 있다.
- [0126] -시퀀싱된 변이체는 전사체 및 유전자에 대한 기능적 영향을 결정하기 위해 엔셈블85(Ensembl85) 유전자 정의를
사용하여 snpEFF(Cingolani 등 (2012); Fly (Austin) 6, 80-92)로 주석을 달 수 있다(annotated).
- [0127] 본원에 기재된 방법은 생리적 장애를 일으키는 질병-유발 변이체를 식별하기 위해 적용될 수 있다. 비-제한적인
예는 심리적 장애, 혈액-관련 장애, 통증-관련 장애, 호르몬-관련 장애, 폐 질환, 치과 장애, 불임-관련 장애,
정신 장애, 운동 장애, 심혈관 장애, 순환계 장애, 자가면역 질환, 염증성 질환, 신장 장애, 간 장애, 유전성
출혈성 모세혈관확장증, 운동 감각 신경병증, 가족성 대동맥류, 갑상선암, 색소성 녹내장, 가족성 고콜레스테롤
혈증, 또는 이들의 조합을 포함한다.
- [0128] 방법은 전술한 단계 중 임의의 것에 제한되지 않으며, 서열 변이체의 획득은 임의의 적절한 수단에 의해 수행될
수 있음이 이해된다.
- [0129] 본 명세서는 또한 적어도 부분적으로, 복수의 피험자의 게놈 샘플 데이터세트 내의 1급 친척에 관한 정보로부터
생성된 혈통이 가족에서 분리되는 희귀 변이체를 식별하기 위한 정보를 제공할 수 있다는 인식에 기초한다.
- [0130] 1급 친척을 식별하는 데 사용할 수 있는 몇 가지 통계적 방법이 개발되었다. 이러한 비-제한적인 예 중 하나는
개체가 데이터세트 내에서 다른 유형의 가족 관계를 식별할 수 있는지 여부로 아이덴티티-바이-디센트(IBM) 추
정값의 계산을 통하는 것이며, 프리머스(PRIMUS)(Staples 등(2014), Am. J. Hum. Genet. 95, 553-564)을 사용
하여 쌍 관계를 다른 가족 부류로 분류하고 혈통을 재구성할 수 있다. 데이터세트 간의 추정된 1급 관계만 포함
되어야 한다. 예를 들어, 엑스 시퀀싱 데이터를 포함하는 데이터세트에서 1급 친척을 식별하기 위해 2018년 9월
7일에 출원된 공동 계류중인 미국 특허 공개 번호 20190205502에 기술된 방법, "유전자 데이터 분석에서 관계를
활용하기 위한 시스템 및 방법" 을 이용할 수 있으며, 이는 그 전체가 여기에 참조로 포함된다.
- [0131] 다수의 피험자의 게놈 샘플 데이터세트에서 혈통을 생성하기 위해, COP (Constructing Outbred Pedigrees) 및
CIP(Constructing Inbred Pedigrees), IPED (Inheritance Path-based Pedigree Reconstruction) 및 IPED2,
PREPARE(친척 분할) 및 최대 비관련 세트의 혈통 재구성 및 식별(Pedigree Reconstruction and Identification
of the Maximally Unrelated Set, PRIMUS)과 같은 몇 가지 접근법을 사용할 수 있다(Riester 등
Bioinformatics(2009) 25: 2134-2139; Hadfield 등 Molecular Ecology(2006) 15: 3715-3730; Marshall 등
Molecular Ecology(1998) 7: 639-655; Cussens 등 Genetic Epidemiology(2013) 37: 69-83; He 등 Journal of
Computational Biology(2013) 20: 780-792 ; Kirkpatrick 등 Journal of Computational Biology(2011) 18:
1481-1493; Staples 등 Genetic Epidemiology (2013) 37: 136-141; Shem-Tov 및 Halperin. PLoS
Computational Biology(2014) 10: e1003610). PLINK, KING 및 KINSHIP과 같은 다른 방법도 사용할 수 있다.
- [0132] 본 명세서는 상술한 데이터세트의 임의의 것, 1급 친척을 식별하고/하거나 혈통을 생성하는 방법 중 어느 것에
제한되지 않으며, 다수의 피험자의 게놈 샘플 데이터세트의 획득 및 처리가 당업계에 알려진 임의의 적절한 수
단에 의해 수행될 수 있음이 이해된다.
- [0133] 본 명세서는 또한 데이터세트에서 영향받은 개체와 영향받지 않은 개체를 결정함으로써 혈통을 생성하고 농축
혈통을 형성하기 위해 혈통을 정제함으로써 혈통을 생성하는 정보가 희귀한 유전적 변이와 질병 사이에서 연관
을 찾기 위한 다운-스트림 분석에 중요하다는 인식에 적어도 부분적으로 기초한다.
- [0134] 데이터세트에서 영향받는 개체들은 적어도 하나의 이진 형질 또는 극도의 정량적 형질 또는 이들의 조합의 존재
에 기반하여 데이터세트에서 개체를 식별함으로써 정의될 수 있다.
- [0135] 일부 예시적인 구현양태에서, 이진 형질은 질병 및 관련 건강 문제의 국제 통계 분류 목록(ICD)으로부터의 3개
의 문자 코드를 사용하여 정의된다. 일부 특정 예시적인 구현양태에서, ICD의 9번째 또는 10번째 개정판으로
부터의 3개의 문자 코드를 사용하여 이진 형질을 정의한다. 이진 형질은 ICD의 9번째 또는 10번째 개정판의 4개
의 문자 코드를 사용하여 추가로 정의할 수 있다. 개체의 표현형에 설명된 이진 형질이 있는 경우, 개체는 "영
향받은" 것으로 결정될 수 있다. 일부 예시적인 구현양태에서, 코호트에서 유병률이 5% 초과인 이진 형질을 갖

는 개체는 이전에 "영향받는"것으로 결정된 경우에도 "영향받지 않은"것으로 결정될 수 있다. 또한, 개체가 의 료 기록에 형질의 부재 또는 존재를 표시하고 있고 상충되는 기록이 있는 경우 해당 개체는 미지의 영향받은 것 으로 결정된다.

[0136] 일부 예시적인 구현양태에서, 극도의 정량적 형질은 집단에서의 형질의 분포에 기반하여 형질의 극히 높거나 낮 은 값을 갖는 개체를 취함으로써 정의되며, 예를 들어, 각 형질 값에 대한 z-점수를 계산하고 형질의 z-점수가 극히 높거나 낮은 특성 값에 대해 각각 2 초과이거나 -2 미만이면 개체를 "영향받은" 것으로 표시한다. 또한, 개체가 의료 기록에 형질의 부재 또는 존재를 표시하고 있고 상충되는 기록이 있는 경우 해당 개체는 미지의 영 향받은 것으로 결정된다.

[0137] 영향받은 개체를 포함하는 혈통은 농축 혈통을 생성하도록 추가로 정제될 수 있다. 혈통은 표현형 분리 또는 p- 값을 기반으로 농축될 수 있다.

[0138] 도 1은 1급 네트워크로부터의 개체가 영향받는 것 및 영향받지 않은 것으로 결정되는 예시적인 구현양태의 흐름 도이다. 개체의 1급 네트워크는 임의의 적절한 수단에 의해 단계(100)에서 복수의 인간 피험자로부터 생성된다. 네트워크의 모든 개체는 기록된 각 이진 형질 또는 기록된 각 정량적 형질 또는 이들의 조합에 대해 단계(110) 에서 평가될 수 있다. 네트워크의 모든 개체는 단계(120)에서 기록된 각 이진 형질에 대해 평가될 수 있으며 단 계(140)에서 이진 형질로 영향받은 경우 "영향받음" 으로 분류된다. 반대로, 개체가 고려중인 특정 이진 형질에 영향받지 않은 경우, 개체는 단계(150)에서 "영향받지 않음"으로 분류된다. 네트워크의 모든 개체는 각각기록된 정량적 형질에 대해 단계(130)에서 평가될 수 있으며 정량적 형질을 가지며 단계(140)에서 정량적 형질에 영향 을 받으면 "영향받음"으로 분류된다. 반대로 개체가 고려중인 특정 정량적 형질에 영향받지 않으면 단계(150)에 서 "영향받지 않음"으로 분류된다.

[0139] 도 2는 1급 네트워크로부터의 개체가 영향받는 것과 영향받지 않은 것으로 결정되는 다른 예시적인 구현양태의 흐름도이다. 임의의 적절한 수단에 의해 단계 (100)에서 복수의 인간 피험자로부터 개체의 1급 네트워크를 생성 한 후, 단계(110)에서 네트워크의 모든 개체는 각각 기록된 이진 형질 또는 기록된 각각의 정량적 형질 또는 이 들의 조합에 대해 평가될 수 있다. 또한, 각각의 기록된 이진 형질 또는 기록된 정량적 형질 또는 이들의 조합 을 가진 모든 개체는 단계(155)에서 이진 형질 또는 정량적 형질의 존재를 기반으로 평가된다. 단계(155)에 이 어, 단계(160)은 개체를 분류할 수 있다: 영향받는 것으로 개체를 식별하는 데 사용된 이진 형질이 코호트에서 5% 초과 의 유병률을 가진다면, 그 후 영향받은 개체는 단계(190)에서 "영향받지 않음"으로 분류될 수 있고; 영 향받는 것으로 개체를 식별하는 데 사용된 이진 형질이 5 % 미만의 유병률을 가진다면 영향받은 개체는 단계 (180)에서 "영향받음"으로 분류될 수 있다. 마찬가지로 단계(170)은 개체를 재분류할 수 있다: 영향받는 것으로 개체를 식별하는 데 사용된 정량적 형질이 코호트의 평균 정량적 형질보다 2 표준 편차를 초과하면 개체는 단계 (180)에서 "영향받음"으로 분류되거나, 그렇지 않으면 단계(190)에서 개체는 "영향받지 않음"으로 분류된다.

[0140] 혈통 내 또는 혈통에 걸친 표현형 분리는 우성 및 부가 분리 모델 또는 열성 분리 모델을 생성할 수 있다. 우성 및 부가 분리 모델로의 표현형 분리를 갖는 혈통에 대한 일부 예시적인 구현양태에서, 하나의 가능한 구조 및 공통 조상의 3명 이상의 영향받은 개체를 갖는 혈통을 사용하여 농축 혈통을 생성할 수 있다. 또한, 거짓 양성 을 줄이기 위해 하나 이상의 관련된 영향받지 않은 개체를 갖는 혈통을 선택하여 분리 분석을 위해 농축 혈통을 우선순위화할 수 있다.

[0141] 열성 분리 모델로의 표현형 분리를 갖는 혈통에 대한 일부 예시적인 구현양태에서, 하나의 가능한 구조 및 영향 받지 않은 부모가 있는 하나 이상의 영향받은 개체를 갖는 혈통은 농축 혈통을 생성하는데 사용된다. 또한, 둘 이상의 영향받는 형제자매를 갖는 혈통을 선택하여 분리 분석을 위해 농축 혈통을 우선순위화할 수 있다.

[0142] 일부 예시적인 구현양태에서, 둘 이상의 표현형적으로 유사하거나 상보적인 이진 또는 극도의 정량적 형질로부 터의 영향받은 개체들을 병합하여 모든 형질을 포함하는 장애에 대한 영향받은 개체들을 형성할 수 있다. 예를 들어, 양극성 장애가 농축된 혈통을 찾을 때, 양극성 장애의 유전적 원인이 일부 개체에서 단극성으로만 나타날 수 있기 때문에 단극성 장애도 또한 고려할 수 있다.

[0143] 일부 예시적인 구현양태에서, 둘 이상의 극도의 또는 흥미로운 이진 또는 극도의 정량적 형질을 갖는 영향받 은 개체들은 이들 둘 이상의 형질 모두를 포함하는 장애에 대해 영향받은 개체들을 형성하도록 선택될 수 있다. 둘 이상의 극도의 또는 흥미로운 형질을 가진 영향받은 개체들의 교차점을 취하면 개체의 더 동질적인 서브셋 을 식별할 수 있다. 예를 들어, 천식과 COPD를 모두 가진 개체를 가진 농축 혈통을 얻으려면 천식과 COPD를 모 두 가진 환자의 교차점을 영향받는 것으로 간주한다.

- [0144] 본 명세서는 전술한 장애 또는 분리 모델 중 임의의 것에 제한되지 않으며 혈통 농축이 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합을 기반으로 임의의 장애 또는 분리 모델에 대해 수행될 수 있음이 이해된다.
- [0145] 대안적으로, 농축 혈통은 p-값을 기반으로 결정될 수 있다. 일부 예시적인 구현양태에서, 혈통의 설립자 고정 가치를 식별할 때, 혈통이 이진 형질에 대해 농축되었다는 것을 평가하기 위해 이항 테스트가 수행된다. 다른 예시적인 구현양태에서, 혈통의 설립자 고정 가치를 식별할 때, 혈통이 극도의 정량적 형질에 대해 농축되었는지 평가하기 위해 t-테스트가 수행된다. 또한, 다중-테스트 수정된 p-값 컷오프가 거짓 양성을 제거하도록 설정된다.
- [0146] 본 명세서는 적어도 부분적으로, 주어진 표현형을 갖는 영향받은 개체에 대해 농축된 혈통이 수반되는(예를 들어, 회귀한) 변이체가 관심 표현형과 분리되고 유도될 수 있다는 인식에 기초한다. 이러한 유전적 원인이 가족 단위 내에서 공유될 가능성이 더 높기 때문에, 관심있는 표현형을 가진 영향받은 개체들이 농축된 혈통을 식별하면 이러한 표현형을 유도하는 우발성(예: 드문) 돌연변이를 식별하는 데 도움이 될 수 있다.
- [0147] 농축된 혈통이 확인되면, 분리 분석 및 가족-기반 연관 분석을 수행하여 근원적인 유전적 원인을 결정할 수 있다. 일부 혈통의 경우, 영향받은 개체들과 분리되는 공지된 질병-유발 돌연변이가 존재할 것이다. 나머지 혈통은 여러 혈통에 걸쳐 영향받은 개체에서 또는 혈통에 포함되지 않은 데이터세트에서 영향받은 개체들로 분리되는 변이체 및 유전자에 의해 우선위화될 수 있다. 그럼에도 불구하고 이러한 분리 분석의 결과에는 후보 변이체 목록이 포함될 수 있다.
- [0148] 다양한 정도의 일반성 모델을 테스트함으로써 분리 분석을 수행할 수 있다. 다양한 제한(예: 우성 또는 열성 계승)이 있는 모델은 가장 일반적인 모델과 비교될 수 있으며 모델의 모든 매개 변수는 데이터에 가장 적합한 모델을 확인하기 위해 추정된다. 큰 혈통의 가족 및 영향받은 많은 개체들은 유전자가 중요하다는 것을 확인하고 특정 유전자를 식별하는 데 특히 유익하다.
- [0149] 주어진 표현형의 유전적 원인을 식별하는 데 도움을 주기 위해 혈통 구조를 사용하는 방법은 일반적으로 연관 매핑, 연결 분석 또는 둘 모두에 대한 혁신적인 변이를 포함한다. 이러한 방법에는 MORGAN, pVAAS, FBAT (www.hsph.harvard.edu/fbat/fbat.htm), QTD (csg.sph.umich.edu/abecasis/qtd/), ROADTRIPS, rareIBD 및 RV-GDT가 포함된다. 사용에 적절한 방법은 표현형, 계승 모드, 조상 배경, 혈통 구조/크기, 혈통 수 및 관련없는 데이터세트의 크기에 의존한다. 관계 및 혈통을 사용하여 유전자-표현형 연관을 직접 조사하는 것 외에도, 추가의 또는 개선된 데이터를 생성하는 여러 다른 방법이 사용될 수도 있다: 혈통-인식 전가, 혈통-인식 페이징, 멘델 오류 검사, 복합 이형 접합 녹아웃 감지(compound heterozygous knockout detection) 및 드 노보(*de novo*) 돌연변이 콜링(calling), 변이체 콜링 검증.
- [0150] 본 발명에 의해 설명되거나 예시된 임의의 방법은 컴퓨터-구현 방법 및/또는 시스템으로서 실행될 수 있다. 당업자에게 공지된 임의의 적합한 컴퓨터 시스템이 이러한 목적을 위해 사용될 수 있다.
- [0151] 도 3은 본 발명의 방법 및 시스템이 작동할 수 있는 예시적인 환경(200)의 다양한 측면을 도시한다. 본 방법은 디지털 및 아날로그 장비를 모두 사용하는 다양한 유형의 네트워크 및 시스템에서 사용될 수 있다. 여기에 기능적 설명이 제공되며 각 기능은 소프트웨어, 하드웨어 또는 소프트웨어 및 하드웨어의 조합에 의해 수행될 수 있다.
- [0152] 환경(200)은 로컬 데이터/처리 센터(210)를 포함할 수 있다. 로컬 데이터/처리 센터(210)는 하나 이상의 컴퓨팅 장치 사이의 통신을 용이하게 하기 위해 근거리 통신망과 같은 하나 이상의 네트워크를 포함할 수 있다. 하나 이상의 컴퓨팅 장치를 사용하여 생물학적 데이터를 저장, 처리, 분석, 출력 및/또는 시각화할 수 있다. 환경(200)은 임의로 의료 데이터 제공자(220)를 포함할 수 있다. 의료 데이터 제공자(220)는 하나 이상의 생물학적 데이터 소스를 포함할 수 있다. 예를 들어, 의료 데이터 제공자(220)는 하나 이상의 환자에 대한 의료 정보에 접근할 수 있는 하나 이상의 건강 시스템을 포함할 수 있다. 의료 정보는, 예를 들어, 병력, 의료 전문가 관찰 및 발언, 실험실 보고서, 진단, 의사의 지시, 처방, 활력 징후, 혈액 균형, 호흡 기능, 혈액 매개 변수, 심전도, X-레이, CT 스캔, MRI 데이터, 실험실 테스트 결과, 진단, 예후, 평가, 입원 및 퇴원 기록, 환자 등록 정보를 포함할 수 있다. 의료 데이터 제공자(220)는 하나 이상의 컴퓨팅 장치 사이의 통신을 용이하게 하기 위해 근거리 통신망과 같은 하나 이상의 네트워크를 포함할 수 있다. 하나 이상의 컴퓨팅 장치는 의료 정보를 저장, 처리, 분석, 출력 및/또는 시각화하는 데 사용될 수 있다. 의료 데이터 제공자(220)는 의료 정보를 비-식별화하고 비식별화된 의료 정보를 로컬 데이터/처리 센터(210)에 제공할 수 있다. 비-식별화된 의료 정보는 의료

정보를 비-식별된 상태로 유지하면서 한 환자의 의료 정보를 다른 환자와 구별하기 위해 각 환자에 대한 고유 식별자를 포함할 수 있다. 비-식별화된 의료 정보는 환자의 신원이 자신의 특정 의료 정보와 연결되는 것을 방지한다. 로컬 데이터/처리 센터(210)는 비-식별화된 의료 정보를 분석하여 각 환자에게 하나 이상의 표현형을 할당할 수 있다(예를 들어, 질병의 국제 분류 "ICD" 및/또는 현재 절차 용어 "CPT" 코드를 할당함으로써).

[0153] 환경(200)은 NGS 시퀀싱 설비(230)를 포함할 수 있다. NGS 시퀀싱 설비(230)는 하나 이상의 시퀀서(예를 들어, Illumina HiSeq 2500, Pacific Biosciences PacBio RS II 등)를 포함할 수 있다. 하나 이상의 시퀀서는 엑솜 시퀀싱, 전체 엑솜 시퀀싱, RNA-seq, 전체-게놈 시퀀싱, 표적 시퀀싱 등을 위해 구성될 수 있다. 예시적인 측면에서, 의료 데이터 제공자(220)는 비-식별된 의료 정보와 연관된 환자로부터의 생물학적 샘플을 제공할 수 있다. 고유 식별자는 생물학적 샘플 및 생물학적 샘플에 상응하는 비-식별화된 의료 정보 간의 연관성을 유지하는 데 사용할 수 있다. NGS 시퀀싱 설비(230)는 생물학적 샘플을 기반으로 각 환자의 엑솜을 시퀀싱할 수 있다. 시퀀싱 전에 생물학적 샘플을 저장하기 위해, NGS 시퀀싱 설비(230)는 바이오 뱅크(예를 들어, Liconic Instruments로부터)를 포함할 수 있다. 생물학적 샘플은 튜브(환자와 연관된 각 튜브)에 수용될 수 있으며, 각 튜브는 로컬 데이터/처리 센터(210)에 샘플을 자동으로 기록하기 위해 스캔할 수 있는 바코드(또는 다른 식별자)를 포함할 수 있다. NGS 시퀀싱 설비(230)는 균일한 데이터 및 효과적인 논-스톱 작동을 보장하기 위해 하나 이상의 시퀀싱 상에서 사용하기 위한 하나 이상의 로봇을 포함할 수 있다. 따라서 NGS 시퀀싱 설비(230)는 연간 수만 개의 엑솜을 시퀀싱할 수 있다. 한 측면에서, NGS 시퀀싱 설비(230)는 매일 적어도 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10,000, 11,000 또는 12,000 전체 엑솜을 시퀀싱할 수 있는 기능적 능력을 갖는다.

[0154] NGS 시퀀싱 설비(230)에 의해 생성된 생물학적 데이터(예를 들어, 원시 시퀀싱 데이터)는 로컬 데이터/처리 센터(210)로 전송될 수 있으며, 그 다음 생물학적 데이터를 원격 데이터/처리 센터(240)로 전송할 수 있다. 원격 데이터/처리 센터(240)는 하나 이상의 컴퓨팅 장치를 포함하는 클라우드-기반 데이터 저장 및 처리 센터를 포함할 수 있다. 로컬 데이터/처리 센터(210) 및 NGS 시퀀싱 설비(230)는 하나 이상의 고용량 광섬유 라인을 통해 원격 데이터/처리 센터(240)와 직접 데이터를 주고받을 수 있지만, 다른 데이터 통신 시스템(예를 들어, 인터넷)이 고려된다. 예시적인 측면에서, 원격 데이터/처리 센터(240)는 제 3자 시스템, 예를 들어 아마존 웹 서비스(DNAexus)를 포함할 수 있다. 원격 데이터/처리 센터(240)는 분석 단계의 자동화를 용이하게 할 수 있고 안전한 방식으로 하나 이상의 협력자(250)와 데이터를 공유할 수 있게 한다. 로컬 데이터/처리 센터(210)로부터 생물학적 데이터를 수신하면, 원격 데이터/처리 센터(240)는 생물정보학 도구를 사용하여 1차 및 2차 데이터 분석을 위한 자동화된 일련의 파이프 라인 단계를 수행할 수 있으며, 결과적으로 각 샘플에 대한 주석이 달린(annotated) 변이체 파일이 생성된다. 이러한 데이터 분석(예를 들어, 유전자형)의 결과는 로컬 데이터/처리 센터(210)로 다시 전달될 수 있으며, 예를 들어 실험실 정보 경영 시스템(Laboratory Information Management System, LIMS)에 통합되어 각 생물학적 샘플의 상태를 유지하도록 구성될 수 있다.

[0155] 로컬 데이터/처리 센터(210)는 유전자형과 표현형 간의 연관성을 식별하기 위해 비-식별된 의료 정보(식별된 표현형 포함)와 조합하여 NGS 시퀀싱 설비(230) 및 원격 데이터/처리 센터(240)를 통해 얻은 생물학적 데이터(예를 들어, 유전자형)를 이용할 수 있다. 예를 들어, 로컬 데이터/처리 센터(210)는 표현형-우선 접근법을 적용할 수 있으며, 여기서 표현형은 특정 질병 영역, 예를 들어 심혈관 질환에 대한 극도의 혈액 지질에서 치료 잠재력을 가질 수 있는 표현형이 정의된다. 또 다른 예는 전형적인 범위의 동반질환으로부터 보호되는 것으로 보이는 개체를 식별하기 위한 비만 환자의 연구이다. 또 다른 접근법은 유전자형과 가설, 예를 들어 유전자 X가 질병 Y를 유발하거나 예방하는 데 관여한다는 것으로 시작하는 것이다.

[0156] 예시적인 측면에서, 하나 이상의 협력자(250)는 인터넷(260)과 같은 네트워크를 통해 생물학적 데이터 및/또는 비-식별화된 의료 정보의 일부 또는 전부에 접근할 수 있다.

[0157] 도 4에 도시된 바와 같이 예시적인 측면에서, 로컬 데이터/처리 센터(210) 및/또는 원격 데이터/처리 센터(240) 중 하나 이상은 유전 데이터 구성요소(300), 표현형 데이터 구성요소(310), 유전적 변이체-표현형 연관 데이터 구성요소(320) 및/또는 데이터 분석 구성요소(330)를 포함하는 하나 이상의 컴퓨팅 장치를 포함할 수 있다. 유전자 데이터 구성요소(300), 표현형 데이터 구성요소(310), 및/또는 유전적 변이체-표현형 연관 데이터 구성요소(320)는 시퀀스 데이터의 품질 평가, 참조 게놈에 대한 판독 정렬, 변이체 식별, 변이체의 주석달기, 표현형 식별, 변이체-표현형 연관 식별, 데이터 시각화, 이들의 조합 등 하나 이상에 대해 구성될 수 있다.

[0158] 예시적인 측면에서, 하나 이상의 구성요소는 전체 하드웨어 구현양태, 전체 소프트웨어 구현양태, 또는 소프트웨어 및 하드웨어 측면을 결합한 구현양태의 형태를 취할 수 있다. 더욱이, 본 방법 및 시스템은 저장 매체에

구현된 컴퓨터-판독가능 프로그램 명령(예를 들어, 비-일시적 컴퓨터 소프트웨어)을 갖는 컴퓨터-판독가능 저장 매체상의 컴퓨터 프로그램 제품의 형태를 취할 수 있다. 보다 구체적으로, 본 방법 및 시스템은 웹-구현 컴퓨터 소프트웨어의 형태를 취할 수 있다. 하드 디스크, CD-ROM, 광학 저장 장치 또는 자기 저장 장치를 포함하는 임의의 적절한 컴퓨터-판독가능 저장 매체가 사용될 수 있다.

[0159] 예시적인 측면에서, 유전 데이터 구성요소(300)는 하나 이상의 유전적 변이체에 기능적으로 주석을 달도록 구성될 수 있다. 유전 데이터 구성요소(300)는 또한 하나 이상의 유전 변이체를 저장, 분석, 수신 등을 위해 구성될 수 있다. 하나 이상의 유전적 변이체는 하나 이상의 환자(피험자)로부터 얻은 서열 데이터(예를 들어, 원시 서열 데이터)로부터 주석을 달 수 있다. 예를 들어, 하나 이상의 유전적 변이체는 적어도 100,000, 200,000, 300,000, 400,000 또는 500,000 피험자 각각으로부터 주석을 달 수 있다. 하나 이상의 유전적 변이체에 기능적으로 주석을 달아서 유전적 변이체 데이터가 생성된다. 예를 들어, 유전적 변이체 데이터는 하나 이상의 변이체 콜 포맷(Variant Call Format, VCF) 파일을 포함할 수 있다. VCF 파일은 SNP, 인델 및/또는 구조적 변형 콜을 나타내는 텍스트 파일 형식이다. 변이체는 전사체/유전자에 대한 기능적 영향을 평가하고 잠재적 기능 손실(potential loss-of-function, pLoF) 후보를 식별한다. 변이체는 엔SEMBL75(Ensembl75) 유전자 정의를 사용하여 snpEff로 주석달기하고 기능 주석달기는 각 변이체(및 유전자)에 대해 추가로 처리된다.

[0160] 숫자 및/또는 문자로 본원에 제공된 바와 같은 방법 단계의 연속적인 라벨링은 방법 또는 그의 임의의 구현양태를 특정 지시된 순서로 제한하는 것을 의미하지 않는다.

[0161] 특허, 특허 출원, 공개된 특허 출원, 수탁 번호, 기술 기사 및 학술 논문을 포함한 다양한 간행물이 명세서 전체에서 인용된다. 이들 인용된 각각의 참고 문헌은 전체적으로 그리고 모든 목적을 위해 본원에 참고로 포함된다.

[0162] 본 명세서는 본 명세서를 보다 상세하게 설명하기 위해 제공되는 하기 실시예를 참조하여 보다 완전히 이해될 것이다. 이들은 예시를 위한 것이며 본 명세서의 범위를 제한하는 것으로 해석되어서는 안된다.

[0163] 실시예

[0164] 실시예 1

[0165] 개체 및 샘플

[0166] 마이코드 커뮤니티 헬스 이니셔티브(MyCode Community Health Initiative)의 일부가 되는 데 동의한 93,368명의 비-식별화된 가이싱거 헬스 시스템(Geisinger Health System, GHS) 참가자를 시퀀싱했다. 이 계획의 일환으로 개체들은 레제네론(Regeneron) GHS DiscovEHR 협업의 일환으로 게놈 분석 및 가이싱거 기관 검토 위원회(Geisinger Institutional Review Board)에서 승인한 프로토콜에 따라 GHS EHR의 데이터에 연결하는 것을 포함하여 광범위한 미래 연구를 위해 혈액 및 DNA 샘플을 제공하는데 동의했다. 수행된 모든 분석은 참가자의 동의 및 IRB 승인에 따라 수행되었다. 각 참가자는 해당하는 비-식별화된 EHR에 연결된 엑숨을 가지고 있다. DiscovEHR 연구는 연구 참가자로서 가족을 구체적으로 대상으로 하지 않았지만, 만성 건강 문제(그리고 서로 관련이 있을 수 있는) 때문에 의료 시스템과 자주 상호 작용하는 성인뿐만 아니라 GHS의 관상동맥 카테터 삽입 실험실(Coronary Catheterization Laboratory) 및 바리야트릭 서비스(Bariatric Service)로부터의 참가자에 대해 암시적으로 농축되었다.

[0167] 실시예 2

[0168] 샘플 준비, 시퀀싱, 변이체 콜링 및 샘플 QC

[0169] 최초 61K 샘플("VCrome 세트")에 대한 샘플 준비 및 시퀀싱은 이전에 설명되었다(Dewey 등 Science(2016) 354:aaf6814). 31K 샘플의 나머지 세트는 님블젠 프로브 캡처(NimbleGen probed capture) 대신에, 추가 프로브가 첨가된 약간 수정된 버전의 IDT xGen 프로브를 사용하여 표준 xGen 프로브로 제대로 커버되지 않지만 님블젠 VCrome 캡처 시약으로 잘 커버되는 게놈 영역을 캡처하여 동일한 프로세스에서 준비했다. 캡처된 단편은 스트렙타비딘-접합된 비드(streptavidin-conjugated beads)에 결합되고, 비-특이적 DNA 단편은 제조업체(IDT)의 권장 프로토콜에 따라 일련의 엄격한 세척으로 제거되었다. 두 번째 샘플 세트는 "xGen 세트"로 지칭되었다. 게놈 분석 툴킷(Genome Analysis Toolkit, GATK; Web Resources)을 사용하여 변이체 콜이 생성되었다. GATK는 추정 인델 주변의 각 샘플에 대해 정렬된 중복-표시 읽기의 로컬 재정렬에 사용되었다. INDEL 재정렬된 중복-표시 판독은 GATK의 하플로타입콜러(HaplotypeCaller)를 사용하여 처리되어 샘플이 게놈 변이체 콜 포맷(gVCF)에서 게놈 참조로부터 변이된 샘플에서 모든 엑소닉 위치를 식별했다. 유전자형 분석은 각 샘플 및 참조와 비교하여 단

일-뉴클레오타이드 변이체(SNV) 및 인텔을 모두 식별하는 단일-샘플 변이체 콜 포맷(VCF) 파일을 출력하는 무작위로 선택된 50개의 샘플로 구성된 트레이닝 세트에 대해 GATK의 게노타입GYCF (GenotypeGYCF)로 수행되었다. 단일-샘플 VCF 파일은 두 세트에서 단일-샘플 VCF 파일로부터 모든 가변 사이트를 포함하는 허위-샘플을 생성하는데 사용되었다. VCRome 세트에 대해 독립적인 pVCF 파일은 허위-샘플과 함께 200개의 단일-샘플 gVCF 파일을 공동 콜링하여 두 캡처 세트에 걸쳐서 모든 가변 사이트에서 각 샘플에 대한 콜 또는 비-콜을 강제 실행함으로써 생성되었다. 모든 200개-샘플 pVCF 파일을 결합하여 VCRome pVCF 파일을 생성한 다음 이 과정을 반복하여 xGen pVCF 파일을 생성했다. VCRome 및 xGen pVCF 파일을 결합하여 통합 pVCF를 생성했다. GRCh38에 대한 서열 판독을 엔SEMBL 85 유전자 정의를 사용하여 정렬하고 변이체에 주석을 달았다. 유전자 정의는 54,214 개의 전사체로 제한되었으며, 이는 주석이 달린 시작 및 종결 코돈이 있는 단백질-코딩 유전자 19,467에 해당한다. 이전에 설명한 샘플 QC 과정 후, 92,455개의 엑솜이 분석을 위해 남아 있었다.

[0170] 실시예 3

[0171] 주요 구성요소 및 조상 추정

[0172] 플링크v1.910(PLINKv1.910)을 사용하여 통합 데이터세트를 맵(HapMap)318과 병합하고, 참조 SNP 더스터 ID를 기반으로 두 데이터세트에 있는 SNP를 유지했다. 분석은 다음 플링크 필터를 적용하여 미소-대립 유전자 빈도 > 10 %, 유전자형 결손 < 5% 및 하디-바인버그 평형(Hardy-Weinberg Equilibrium) p 값 > 0.00001을 가진 고품질 공통 SNP로 제한되었다: "-maf 0.1-geno 0.05 -snps -오직-h we 0.00001". 맵3 샘플에 대해 주요 구성요소(principal components, PC)를 계산한 다음 플링크를 사용하여 데이터세트의 각 단순 요소를 해당 PC에 투영했다. 맵3 샘플에 대해 PC를 사용하여 다섯개의 조상 상위클래스 각각에 대해 커널 밀도 추정기(kernel density estimator, KDE)를 훈련했다: 아프리카인(AFR) 혼합 미국인(AMR), 동아시아인(EAS), 유럽인(EUR) 및 남아시아인(SAS). KDE는 각 샘플이 각 상위클래스에 속할 가능성을 추정하기 위해 계산되었다. 각 샘플에 대해 가능성을 기반으로 한 조상 상위클래스가 지정되었다. 샘플에 확률이 > 0.3 인 두 개의 조상 그룹이 있는 경우, 샘플에는 EUR보다 AFR, EUR보다 AMR, EAS보다 AMR, EUR보다 SAS, AFR보다 AMR이 할당되었다; 그렇지 않으면 "미지의(UNKNOWN)"이다. 0개 또는 두개 이상의 조상 그룹이 충분히 높은 가능성을 가진 경우, 샘플은 조상에 대해 "미지의"로 지정되었다. 미지의 조상을 가진 샘플은 조상 기반 이이덴티티-바이-디센트(IBD) 계산에서 제외되었다.

[0173] 실시예 4

[0174] IBD 추정

[0175] 다음 플래그를 사용하여 완전한 데이터세트에서 플링크를 실행하여 고품질의 공통 변이체를 필터링했다: "--maf 0.1 --geno 0.05 --snps-단독--hwe 0.00001. 그런 다음 엑솜 데이터에서 정확한 IBD 추정값을 얻기 위해 두-갈래 접근 방식을 취했다. 첫째, 개체 간의 IBD 추정값은 조상 분석에서 결정된 것과 동일한 조상 상위클래스(예: AMR, AFR, EAS, EUR 및 SAS) 내에서 계산되었다.

[0176] 둘째, 조상이 다른 개체 간의 1급 관계를 파악하기 위해 --min 0.3 플링크 옵션을 사용하여 모든 개체 간에 IBD 추정값을 계산했다. 그런 다음 개체는 네트워크 노드가 개체이고 에지가 1급 관계인 1급 가족 네트워크로 그룹화되었다. 각 1급 가족 네트워크는 프리프리머스(prePRIMUS) 파이프 라인(Staples 등(2014); Am. J. Hum. Genet. 95, 553-564)을 통해 실행되었으며, IBD 추정을 개선하기 위해 샘플의 조상을 적절한 조상 미소 대립 유전자 빈도와 일치시켰다. 이 과정은 각 가족 네트워크 내 개체 간의 1급 관계를 정확하게 추정했다(최소 PI_HAT 0.15).

[0177] 실시예 5

[0178] 92K 인간 엑솜의 코호트에서 관계 추정 및 관련성 설명

[0179] 92,455명의 개체, 43의 일란성 쌍둥이, 16,476의 부모-자녀 관계, 10,479의 완전-형제자매 관계 및 39,000의 2급 관계가 DiscovEHR 데이터세트로부터 식별되었다(도 5). 무방향 그래프를 생성하기 위해 개체는 노드로서 관계는 에지로서 취급했다. 1급 관계만 사용하여 12,594개의 연결된 구성요소를 식별했으며, 이를 1급 가족 네트워크라고 지칭한다. DiscovEHR 코호트의 개체 중 39%는 데이터세트에 적어도 1급의 친척을 가졌다.

[0180] 표 1 (DiscovEHR 데이터세트의 조상 분석)

조상 클래스	샘플수	사람 %
EUR	88634	95.9%
AFR	1984	2.1%
AMR	959	1.0%
SAS	196	0.2%
EAS	194	0.2%
미지	488	0.5%

[0181]

[0182] 표2 (DiscovEHR 데이터세트에서 1급 관계에 관련된 개체의 조상 배경의 완전한 분석)

관계	조상	총수
MZ 쌍둥이	EUR-EUR	42
MZ 쌍둥이	SAS-SAS	1
부모-자식	EUR-EUR	16028
부모-자식	AFR-AFR	115
부모-자식	AFR-EUR	86
부모-자식	AMR-EUR	83
부모-자식	AMR-AMR	43
부모-자식	EUR-미지	43
부모-자식	미지-미지	20
부모-자식	AFR-미지	13
부모-자식	AMR-미지	13
부모-자식	EAS-미지	13
부모-자식	SAS-SAS	11
부모-자식	AFR-AMR	5
부모-자식	EUR-SAS	2
부모-자식	EAS-SAS	1
완전-형제자매	EUR-EUR	10364
완전-형제자매	AFR-AFR	155
완전-형제자매	AMR-EUR	24
완전-형제자매	AMR-AMR	16
완전-형제자매	미지-미지	10
완전-형제자매	AMR-미지	4
완전-형제자매	SAS-SAS	2
완전-형제자매	EAS-EAS	1
완전-형제자매	EAS-미지	1
완전-형제자매	EUR-SAS	1
완전-형제자매	EUR-미지	1

[0183]

[0184] 실시예 6

[0185] 혈통 재구성

[0186] DiscovEHR 코호트 내에서 식별된 모든 1급 가족 네트워크는 프리머스v1.9.0으로 재구성되었다. 결합된 IBD 추정

값은 유전적으로 파생된 성별 및 EHR 보고된 연령 내에서 프리머스에 제공되었다. $PI_{HAT} > 0.375$ 의 관련성 컷 오프는 재구성을 1급 가족 네트워크로 제한하기 위해 지정되었다.

[0187] 이러한 혈통들 사이에서 멘델 방식으로 분리되는 300개 초과 전자 건강 기록(EHR) 유래 표현형이 데이터세트에서 발견되어, 대규모로 전통적인 멘델 분석을 가능하게 하는 2,000개 이상의 잠재적으로 유익한 혈통-표현형 쌍을 제공한다.

[0188] **실시예 7**

[0189] **혈통 농축**

[0190] 1급 가족 네트워크의 개체는 적어도 하나의 이진 형질, 극도의 정량적 형질 또는 이들의 조합에 대해 "영향받은" 또는 "영향받지 않은" 것으로 결정되었다. 이러한 영향받은 개체들의 집합은 가족-기반 분리 분석을 수행할 수 있을 만큼 충분히 영향받은 개체가 농축된 혈통을 식별하기 위해 혈통과 교차되었다.

[0191] 2,978의 형질-혈통 농축 쌍이 데이터세트로부터 인식되었다(2,596 우성 및 382 열성). 이러한 형질-혈통 농축 쌍 중에, 981개의 혈통에서 1,015개의 다른 형질을 가진 3,975명의 영향받은 개체가 있었다. 50% 이상의 형질이 두개 이상의 혈통에서 농축되고 357의 형질이 3 이상의 혈통에서 농축되었다.

[0192] 추가로, 2,978의 형질-혈통 농축 쌍 중에서, 1,911개는 673개의 혈통을 갖는 809개의 상이한 형질을 갖는 이진 형질-혈통 농축 쌍이었다. 이진 형질-혈통 농축 쌍에서, 가장 농축된 혈통은 충치에 대한 것이었다($N = 46$). 또한, 2,978개의 형질-혈통 농축 쌍 중에서, 1,067개는 581개의 혈통을 가진 206개의 상이한 형질을 가진 정량적 형질-혈통 농축 쌍이었다. 정량적 형질-혈통 농축 쌍에서, 가장 농축된 혈통은 고 트리글리세라이드_Med_Lab 값 ($N = 19$)에 대한 것이었다.

[0193] **7.1 원발성 혈전성향증**

[0194] 원발성 혈전성향증은 혈전 형성(과응고상태)으로 이어지는 지혈 메커니즘의 유전적 장애이다. 이것은 일반적으로 정맥계에 영향을 미친다(예: 심부정맥 혈전증, 폐색전증).

[0195] 집단의 개체는 원발성 혈전성향증에 대한 이진 형질(Phe10_D685, ICD10 4D)에 기반하여 영향받은 것으로 결정되었다.

[0196] 실시예 6에 언급된 방법을 사용하여 재구성된 혈통(표 3 및 4)으로 부터, 1 급 혈통을 필터링하여 하나의 가능한 구조가 없고 공통 조상의 영향받은 3개 미만의 원발성 혈전성향증을 갖는 모든 혈통을 제거하여 원발성 혈전성향증이 농축된 혈통을 생성하였다. 코호트에서 원발성 혈전성향증(Phe10_D685, ICD10CM D68.5)의 유병률은 1.3%였다.

[0197] 따라서 원발성 혈전성향증이 농축된 여러 혈통이 식별되었다(도 6 참조).

[0198] **표3**

일련번호	영향받은 개체 수	공통 조상의 최대 영향받은 개체	공통 조상의 최대 유익한 영향받은 개체	영향받은 형제자매쌍의 수	혈통 세대
1	3	3	3	1	2
2	3	3	4	2	3
3	3	3	3	0	3
4	2	2	2	1	3
5	4	4	4	0	3
6	3	3	3	3	4
7	3	3	3	0	4
8	3	3	3	1	3

[0199]

표4

일련번호	영향받은 세대	영향받지 않은 개체수	미지의 영향받은 개체수	혈통에서 샘플의 수	연령 일관된 가능한 혈통의 수	계승 방식
1	2	3	0	6	1	우성
2	2	11	0	14	1	우성
3	2	6	0	9	1	우성
4	1	5	1	8	1	열성
5	3	0	0	4	1	우성
6	1	29	2	34	1	우성
7	3	23	1	27	0	우성
8	2	6	0	9	1	우성

7. 2 유전성 출혈성 모세혈관확장증

유전성 출혈성 모세혈관확장증(HTT)은 신체 전체의 혈관에 영향을 미치고(혈관 이형성을 유발) 출혈 경향을 유발하는 희귀 상염색체 우성 장애이다(이 상태는 오슬러-베버-렌두 질병(Osler-Weber-Rendu disease, OWRD)으로도 알려져 있으며; 두 용어는 상호교환적으로 사용된다). HHT는 심각한 이환율과 사망률의 잠재적 원인인 점막 피부 모세혈관확장 효소 및 동정맥 기형(AVM)에 의해 나타난다. 병변은 비인두, 중추 신경계(CNS), 폐, 간 및 비장 뿐만 아니라 요로, 위장(GI)관, 결막, 몸통, 팔 및 손가락에 영향을 미칠 수 있다.

집단의 개체는 HTT에 대한 이진 형질(Phe10_I780, ICD10CM I78.0)을 기반으로 영향받은 것으로 결정되었다.

HTT에 대해 실시예 6에 언급된 방법을 사용하여 두개의 혈통을 재구성하였다(표 5 및 6 참조). 두 혈통 모두 공통 조상의 3개의 HHT 영향받은 개체들 및 하나의 가능한 구조를 지녔다. 또한, 코호트에서 HTT의 유병률은 0.0% 였다.

표5

일련번호	영향받은 개체 수	공통 조상의 최대 영향받은 개체	공통 조상의 최대 유익한 영향받은 개체	영향받은 형제자매 쌍의 수	혈통 세대
1	3	3	3	0	3
2	3	3	3	0	3

표6

일련번호	영향받은 세대	영향받지 않은 개체 수	미지의 영향받은 개체 수	혈통에서 샘플 수	연령 일관된 가능한 혈통수	계승 방식
1	3	1	1	5	1	우성
2	2	13	1	17	1	우성

HTT에 대한 이진 형질이 농축된 두개의 혈통을 사용하여 희귀 변이체 분리 분석을 수행했다(도 7a 및 7b 참조).

도 7a에 표시된 HTT가 농축된 혈통에 대해서, 분리 및 연관 분석은 *SMAD4* 유전자의 변이체가 혈통에서 HTT 표현형과 공동 분리되었음을 나타냈다(표 7 참조). *SMAD4*(SMAD 패밀리 구성원 4)는 신호 변환 단백질의 SMAD 패밀리의 구성원이다. Smad 단백질은 형질전환 성장 인자(TGF)-베타 신호에 반응하여 막관통 세린-트레오닌 수용체 키나제에 의해 인산화되고 활성화된다. *SMAD4*는 다른 활성화된 Smad 단백질과 동종 복합체 및 이종 복합체를 형성하고, 이는 핵에 축적되어 표적 유전자의 전사를 조절하며 BMP 신호전달 경로의 중요한 구성요소이다. *SMAD4*의

돌연변이 또는 결실은 유전성 출혈성 모세혈관확장증 증후군(HHT) 및 마이레(Myhre) 증후군의 유전성 장애; 및 청소년 폴립증 증후군(염색체 18q21상의 *SMAD4* 유전자의 이중접합 돌연변이)을 포함하는 가족성 암 감수성 장애와 관련이 있다. *SMAD4*는 종양 억제자로서 작용하며 상피세포 증식을 억제한다. 혈관신생을 줄이고 혈관 과다투과성을 증가시켜 종양에 대한 억제 효과를 또한 가질 수 있다. *SMAD4*의 체세포 돌연변이는 췌장암에서 확인되었다.

표7

유전자	유전자_명칭	NT_변화	AA_변화	TGP_F REQ	EXAC_FRE Q	RGC_F REQ	보존	예측	유전자형
<i>SMAD4</i>	SMAD 패밀리 구성원 4	c.1242_1245 delAGAC	p.D415Efs*20	해당없음	해당없음	0.000011	해당없음	해로운	HET=4/HOM=0

도 7b 및 도 8에 도시된 HHT가 농축된 혈통에 대해 분리 및 연관 분석 결과 액티빈 A 수용체 타입 II-유사 1(ACVRL1) 유전자에 대한 변이가 혈통에서 HHT 표현형과 공동 분리되었음을 나타낸다(표 8 참조). ACVRL1 유전자는 리간드의 TGF-베타 수퍼패밀리에 대한 타입 I 세포-표면 수용체를 인코딩하고 수용체 세린/트레오닌 키나제의 서브패밀리를 형성하는 다른 밀접하게 관련된 ALK 또는 액티빈 수용체-유사 키나제 단백질과 유사한 도메인 구조를 공유한다. ACVRL1에서의 돌연변이는 렌두-오슬러-웨버 증후군 2 및 폐동맥 고혈압으로도 알려진 출혈성 모세혈관확장증 타입 2와 연관된다. 결막 모세혈관확장증, 코 점막 모세혈관확장증이 있는 환자는 종종 질병의 첫 징후로 코 출혈이 종종 생기며, 구강 모세혈관확장증, 다양한 기관의 동정맥 기형, 피부 모세혈관확장증, 빈혈 및 일부 폐동맥 고혈압이 발생한다. HHT2의 내장 발견에는 폐 동정맥 기형(plumonary arteriovenous malformation, PAVM), 대뇌 AVM, 척추 AVM, 간 AVM, AVM으로 인한 위장 출혈 및 간경변이 포함된다. HHT2의 신경학적 증상에는 발작, 허혈성 뇌졸중, 편두통, -대뇌 동정맥 기형 및 뇌내 출혈이 포함된다.

표8

유전자	유전자_명칭	NT_변화	AA_변화	TGP_F REQ	EXAC_F REQ	RGC_F REQ	보존	예측	유전자형
<i>ACVRL1</i>	액티빈 A 수용체 유사 타입 1	c.C853T	p.L285F	해당없음	해당없음	0.000011	보존된	해로운	HET=4/HOM=0

7.3 폐활량측정에 의한 골드 단계(GOLD Stage) 2-4 환자의 폐기종

폐기종은 호흡 곤란을 유발하는 폐질환이며 만성 폐쇄성 폐질환(COPD)을 포함하는 질환 중 하나이다. 폐기종 환자의 경우 폐의 기낭(폐포)이 손상된다. 시간이 지남에 따라 기낭의 내벽이 약해지고 파열되어 많은 작은 공기 공간 대신 더 큰 공기 공간이 만들어진다. 이것은 폐의 표면적을 감소시키고 이어서 이는 혈류에 도달하는 산소의 양을 감소시킨다. 숨을 내쉴 때 손상된 폐포가 제대로 작동하지 않고 오래된 공기가 갇혀 신선하고 산소가 풍부한 공기가 들어갈 여지가 없게 된다.

"폐활량 측정에 의한 골드 단계 2-4 환자의 폐기종"에 대한 이진 형질은 폐 기능 테스트에 대한 정량적 형질에서 파생되었다. 전자 의료 기록에 보고된 다중 발생률을 기반으로 한 높은 신뢰도의 금연 COPD 환자 세트가 사용되었다. 폐 기능 테스트의 정량적 형질 중 하나는 "최근 폐활량 측정에서 50% 강제 활력 용량에서 강제 흡기 흐름에 대한 50% 강제 활력 용량에서 기관지 확장 전 강제 호기 흐름"을 사용하여 정의되었다. 집단의 형질에 대한 평균은 0 이었고 표준 편차는 0.27 이었다. 농축은 정량적 형질의 하한을 사용하여 수행되었다. 폐 기능 검사에 대한 또 다른 정량적 형질은 "최근 폐활량 측정에서 1초 내에 예상되는 기관지 확장 후 강제 호기량의 백분율"을 사용하여 정의되었다. 집단의 형질에 대한 평균은 81.89 였고 표준 편차는 20.84 였다. 농축은 정량

적 형질의 하한을 사용하여 수행되었다.

[0220] 1급 가족 네트워크로부터 폐활량측정에 의해 골드 단계 2-4 환자의 폐기종에 대한 이진 형질이 농축된 혈통이 분리되었다(도 9 참조). 코호트에서 이 특정 표현형의 유병률은 1.8% 였다. 혈통은 단지 하나의 가능한 구조를 가지고 있고 공통 조상의 세 명의 영향받은 개체들로 구성되었다.

[0221] 7.4 신장 이식

[0222] 신장 이식을 위한 이진 형질이 농축된 혈통(Phe9_V420, ICD9DM V42.0)을 1 급 가족 네트워크로부터 분리했다. 이 특정 표현형의 유병률은 0.8 %였다.

[0223] 1급 혈통은 단지 하나의 가능한 구조를 가지며 공통 조상을 갖는 4명의 영향받은 개체를 지녔다. 필요한 기준을 포함하는 혈통이 식별되었다(도 10 및 표 9 참조).

[0224] 표9

형질 카테고리	건강 상태 및 의료 서비스와의 접촉에 영향을 미치는 인자
영향받은 개체수	5
공통 조상의 최대 영향받은 개체	5
공통 조상의 최대 유익한 영향받은 개체	5
영향받은 형제자매 쌍의 수	2
혈통 세대	2
영향받은 세대	2
영향받지 않은 개체 수	3
미지의 영향받은 개체 수	0
혈통에서 샘플 수	8
연령 일관된 가능한 혈통의 수	1
계승 방식	우성

[0225]

[0226] 7.5 말기 신장 질환

[0227] 집단의 개체는 말기 신장 질환에 대한 이진 형질(Phe10_5856, ICD9CM 585.6)을 기반으로 영향받은 것으로 결정되었다. 말기 신장 질환이 농축된 여러 혈통이 식별되었다(도 11).

[0228] 7.6 유전성 운동 및 감각 신경병증(샤르코-마리-투스 질환)

[0229] 샤르코-마리-투스 질환(Charcot-Marie-Tooth disease, CMT)은 미국에서 약 2,500명 중 약 1명에게 영향을 미치는 가장 흔한 유전적 신경 장애 중 하나이다. 유전성 운동 및 감각 신경병증(HMSN) 또는 비골 근육 위축이라고도 하며 말초 신경에 영향을 미치는 장애 그룹을 포함한다.

[0230] 집단의 개체는 유전성 운동 및 감각 신경병증에 대한 이진 형질(Phe10_G600, ICD10CM G60.0)에 기반하여 영향받은 것으로 결정되었다. 코호트에서 이 특정 표현형의 유병률은 0.1% 였다.

[0231] 실시예 6에서 재구성된 혈통으로부터 유전성 운동 및 감각 신경병증에 대한 1급 혈통은 하나의 가능한 구조 및 공통 조상을 갖는 3명의 영향받은 개체를 갖는다(도 12 및 표 10 참조).

표10

형질 카테고리	신경계 및 감각 기관의 질병
영향받은 개체수	3
공통 조상의 최대 영향받은 개체수	3
공통 조상의 최대 유익한 영향받은 개체수	3
영향받은 형제자매 쌍의 수	1
혈통 세대	3
영향받은 세대	2
영향받지 않은 개체 수	1
미지의 영향받은 개체 수	1
혈통에서 샘플의 수	5
연령 일관된 가능한 혈통의 수	1
계승 방식	우성

유전성 운동 및 감각 신경병증이 농축된 혈통에 대해, 분리 및 연관 분석은 혈통에서 트로포미오신 2(베타)(*TPM2*) 유전자에 대한 변이체가 유전성 운동 및 감각 신경병증 표현형과 공동 분리됨을 나타냈다(표 11). *TPM2*는 액틴 필라멘트 결합 단백질 패밀리의 구성원인 베타-트로포미오신을 인코딩하며 주로 느린, 타입 1 근육 섬유로 표현된다. *TPM2*의 돌연변이는 다른 육종 트로포미오신 단백질의 발현을 변화시킬 수 있으며 캡 질환, 네 말린 근육병증 및 원위 관절구축증 증후군을 유발할 수 있다.

표11

유전자	유전자_명칭	NT_변화	AA_변화	TGP_F REQ	EXAC_F REQ	RGC_FREQ	보존	예측	유전자형
<i>TPM2</i>	트로포미오신 2 (베타)	c.G479A	p.R160H	해당 없음	0.00002	0.00001	보존된	해로운	HET=6 /HOM=0

다양한 조직에서 인코딩된 *TPM2*의 백만 당 전사(TPM)의 유전자 발현 데이터는 동맥, S 자-결장, 식도-위장 접합부, 식도-근육 및 골격근에서 높은 발현을 나타낸다(도 13 참조).

혈통의 영향받은 개체에 대한 환자 기록(표 12 참조)은 이 가족이 유전성 운동 및 감각 신경병증의 증거를 보여 주지 않고 오히려 *TPM2*의 돌연변이로 인해 네 말린 근육병증 타입 4를 가지고 있음을 나타냈다(Donner 등 Neuromuscular Disorders (2009) 19: 348-3351).

표12

일련번호	표현형 비교
1	발목 터널 증후군; 유전성 말초 신경병증; 선천성 및 유전성 근이영양증; 후천적 발 변형; 선천성 및 유전성 근이영양증; 관절, 여러 부위 구축; 근력저하 (일반화된); 팔다리통증; 비골 근육 위축증; 숨가쁨
2	비골 근육 위축증; 후천적 발의 요내반 변형; 유전성 및 진행성 근이영양증; 상세불명 부위의 단발성신경염; 관절, 발목, 발에서의 통증; 팔다리통증; 달리 분류된 질환에서의 다발성 신경병증; 상세불명의 유전성 및 특발성 말초 신경병증
3	비골 근육 위축증; 보행의 비정상; 유전성 말초 신경병증; 기타 추체외로질환 및 비정상 운동 장애; 후천적 발 변형

7.7 양극성 장애

[0242] 양극성 장애 또는 "조울병"은 정서적 고조(조증 또는 경조증) 및 저조(우울증)를 포함한 극심한 기분 전환을 유발한다. 인구의 약 2.6%(미국 성인 570만 명)가 해당 연도에 이 질환을 앓고 있다.

[0243] 집단의 개체는 양극성 장애 및 단극성 장애에 기반하여 영향받은 것으로 결정되었다. 양극성 장애의 ICD 10 코드는 F31이고; ICD 9 코드는 296.4 내지 296.7이다. 환자의 서브세트(35 내지 40%)가 리튬 처방을 받는다. 단극성/주요 우울 장애의 ICD 10 코드는 F32, F33, F39이고; ICD-9 코드는 296.2/.3/.9(가족 네트워크 내 2급)이다. 자폐증(ICD-10 코드 F84) 및 정신 지체(ICD-10 코드 F70.9, F71.9, F72.9, F73.9, F79.9)가 있는 개체는 영향받는 세트에서 제외되었다. 코호트에서 양극성 장애(F319-3.2%) 및 단극성 장애(각각 F31, F32 및 F33- 0.0%, 4.1% 및 2.1%)에 대한 이진 형질의 유병률은 5% 미만이었다.

[0244] 양극성 장애에 대한 이진 형질이 농축된 혈통은 1급 가족 네트워크에서 분리되었다.

[0245] 1급 혈통은 단지 하나의 가능한 구조 및 공통 조상을 갖는 적어도 3명의 영향받은 개체가 있는지 확인하기 위해 평가되었다(도 14 참조). 농축 혈통에 대해 수행된 분리 분석은 표현형과 공동 분리되는 가능한 변이체 목록을 생성했다(표 13). 표현형과 공동 분리되는 변이체 C20orf203은 해롭고 보존되지 않는다.

[0246] 표13

유전자	유전자 명칭	NT_ 변화	AA_ 변화	TGP_ FREQ	EXA C_ FRE Q	RGC_ FREQ	보존	예 측	유전자 형
<i>C20orf203</i>	염색체 20 오픈 리딩 프레임 203	c.113 delC	p.T38 fs	해당 없음	해당 없음	1.11E-05	해당 없음	해로운	HET=4/ HOM=0
<i>TUBGCP6</i>	튜불린 감마 복합체 연관 단백질 6	c.C21 19T	p.Q7 07X	해당 없음	해당 없음	1.11E-05	보존된	해로운	HET=4/ HOM=0
<i>SEC14L3</i>	SEC14 유사 지질 결합 3	c.A34 6T	p.K11 6X	해당 없음	8.38 E-06	1.11E-05	보존된	해로운	HET=4/ HOM=0
<i>TMPRSS4</i>	막관통 프로테아제, 세린 4	c.C35 8T	p.Q1 20X	해당 없음	8.27 E-06	2.21E-05	보존된	내성의	HET=8/ HOM=0
<i>CARD10</i>	카스파제 모집 도메인 패밀리 구성원 10	c.C55 0A	p.R18 4S	해당 없음	해당 없음	3.88E-05	보존된	해로운	HET=14/ HOM=0
<i>LMAN1L</i>	렉틴, 만노스 결합 1 유사	c.C56 8T	p.R19 0W	해당 없음	4.79 E-05	0.000 144	보존된	내성의	HET=52/ HOM=0
<i>C10orf71</i>	염색체 10 오픈 리딩	c.C37 15A	p.P21 39T	해당 없음	해당 없음	1.38E-05	보존된	해로운	HET=5/ HOM=0

[0247]

	프레임 71							운	
<i>TMEM</i> 181	막관통 단백질 181	c.G16 57A	p.A5 53T	해당 없음	해당 없음	1.38E -05	보존 된	내 성 의	HET=5/ HOM=0
<i>MAN2</i> B1	매노시다 제 알파 클래스 2B 구성원 1	c.G20 94C	p.Q6 98H	해당 없음	4.94 E- 05	7.75E -05	뉴트 럴	해 로 운	HET=28 /HOM= 0
<i>IFT172</i>	편모내 수송 172	c.G26 81A	p.R89 4H	해당 없음	해당 없음	3.05E -05	보존 된	해 로 운	HET=11 /HOM= 0
<i>PIWIL</i> 1	피위 유사 RNA- 매개된 유전자 발현억제 1	c.G15 2A	p.R51 Q	해당 없음	8.26 E- 06	3.05E -05	보존 된	해 로 운	HET=11 /HOM= 0
<i>ACVR</i> L1	활성 A 수용체 유사 타입 1	c.G94 8C	p.E31 6D	해당 없음	4.16 E- 05	0.000 125	뉴트 럴	해 로 운	HET=45 /HOM= 0
<i>TTLL1</i> 0	튜블린 티로신 리가제 유사 10	c.G21 1A	p.E71 K	0.000 02	3.37 E- 05	1.66E -05	뉴트 럴	내 성 의	HET=6/ HOM=0
<i>CDBF</i> A2T2	CBFA2/RU NX1 전좌 파트너 2	c.C13 5A	p.N4 5K	해당 없음	8.24 E- 06	0.000 105	보존 된	해 로 운	HET=38 /HOM= 0
<i>ZBTB2</i>	아연 핑거 및 BTB 도메인 함유 2	c.A14 93G	p.H4 98R	해당 없음	1.65 E- 05	3.05E -05	보존 된	내 성 의	HET=11 /HOM= 0

[0248]

[0249]

FLJ33706 (대체 유전자 기호 C20orf203)은 니코틴 중독의 원인이 되는 가능한 변이체로 식별되었다. 염색체 20 오픈 리딩 프레임 203(C20orf203)의 백만 당 전사(TPM)의 유전자 발현 데이터는 다양한 조직에서 인코딩되지만, 주로 소뇌 반구 및 뇌의 소뇌에서 발현된다(도 15). 연계 연구에서는 FLJ33706의 3'UTR에 위치한 SNP인 rs17123507이 니코틴 중독에 대한 감수성과 중요하게 연관되어 있음을 식별되었다(Li 등 PLoS Computational Biology(2010) 6:e1000734).

[0250]

또한, 두개의 더 농축된 혈통이 식별되었다(도 16 및 17; 표 14 및 15 참조). 두 혈통 모두 단지 하나의 가능한 구조를 가지며 공통 조상의 3명 이상의 영향받은 개체를 가졌다.

[0251] 표14

일련번호	영향받은 개체의 수	공통 조상의 최대 영향받은 개체	공통 조상의 최대 유익한 영향받은 개체	영향받은 형제자매 쌍의 수	혈통 세대
1	5	3	3	0	2
2	4	3	2	0	2

[0252]

[0253] 표15

일련번호	영향받은 세대	영향받지 않은 개체 수	미지의 영향받은 개체 수	혈통에서 샘플의 수	연령 일관된 가능한 혈통의 수	계승 방식
1	2	2	0	7	1	우성
2	2	3	1	8	1	우성

[0254]

[0255] 추가로, 양극성 장애에 대한 이진 형질이 농축된 또 다른 혈통은 단지 하나의 가능한 구조를 가졌으며 공통 조상의 3명 이상의 영향받은 개체를 지녔다(도 18 참조).

[0256] 농축 혈통에 대해 수행된 변이체 분석은 표현형과 공동 분리되는 가능한 변이체 목록을 생성했다(표 16).

[0257]

표16

유전자	유전자 명칭	NT 변화	AA 변화	TGP FRE Q	EXA C FRE Q	RGC FRE Q	보존	예측	유전자형
<i>MCPH 1</i>	마이크로세팔린 1	c.2453-1G>C)		해당 없음	1.675E-05	3.32E-05	보존된	해로운	HET=12/ HOM=0
<i>NOL8</i>	핵 단백질	c.38delG	p.G13fs	해당 없음	166E-05	3.60E-05	해당 없음	해로운	HET=13/ HOM=0
<i>INSM 2</i>	INSM 전사 억제제 2	c.1031_1035del	p.344_345del	해당 없음	해당 없음	5.54E-05	해당 없음	해로운	HET=20/ HOM=0
<i>CDRT 15L2</i>	CMT1A 중복된 전사 15-유사 2	c.345_346del	p.115_116del	해당 없음	0.0002	0.000235	해당 없음	해로운	HET=85/ HOM=0
<i>SHPK</i>	세도헵톨로키나제	c.C355T	p.R119X	0.000599	0.0005	0.000684	뉴트럴	내성의	HET=245/ HOM=0
<i>INSM 2</i>	INSM 전사 억제제 2	c.A686G	p.D229G	해당 없음	5.56E-05	0.00013	보존된	해로운	HET=47/ HOM=0
<i>DAA M1</i>	형태형성의호트리진 연관된 활성화제 1	c.T2474G	p.F825C	해당 없음	6.59E-05	0.000191	보존된	해로운	HET=69/ HOM=0
<i>APBA 2</i>	아밀로이드 베타 전구체 단백질 결합 패밀리 A 구성원 2	c.C1417T	p.R473C	0.0002	8.46E-06	9.97E-05	보존된	해로운	HET=36/ HOM=0
<i>FAM1 07B</i>	시퀀스 유사성이있는 패밀리 107 멤버B	c.A155G	p.E52G	해당 없음	해당 없음	5.26E-05	보존된	해로운	HET=19/ HOM=0
<i>AKR1 C3</i>	알도-케토 환원 효소 패밀리 1	c.G512T	p.R171M	해당 없음	8.24E-06	6.64E-05	뉴트럴	해로운	HET=24/ HOM=0

[0258]

유전자	유전자 명칭	NT 변화	AA 변화	TGP FRE Q	EXA C FRE Q	RGC FRE Q	보존	예측	유전자형
	구성원 C3								
<i>PLEKH G3</i>	플렉크스트린 상동성 및 G3를 함유하는 RhoGEF 도메인	c.G80T	p.G27V	해당 없음	1.87 E-05	3.05E -05	보존 된	해로운	HET=11/ HOM=0
<i>PDSS 1</i>	프레닐 (데카프레닐) 디포스페이트 합성효소, 서브유닛1	c.G1138 A	p.A380T	해당 없음	4.95 E-05	8.30E -05	보존 된	해로운	HET=30/ HOM=0
<i>CCM2</i>	CCM2 스캐폴딩 단백질	c.G391A	p.D131 N	0.00 02	0.00 03	0.00 0221	보존 된	해로운	HET=80/ HOM=0
<i>LRRC3 7B</i>	37B를 포함하는 류신이 풍부한 반복	c.A2188 C	p.K730 Q	해당 없음	4.98 E-05	9.41E -05	뉴트럴	내성의	HET=34/ HOM=0
<i>EML4</i>	극피 동물 미세 소관 연관된 단백질 유사 4	c.A736T	p.N246Y	해당 없음	1.65 E-05	3.88E -05	보존 된	해로운	HET=14/ HOM=0
<i>ARHG EF10</i>	Rho 구아닌 뉴클레오티드 교환 인자 10	c.C3097T	p.H1033 Y	해당 없음	해당 없음	1.66E -05	보존 된	해로운	HET=6/H OM=0
<i>STAT5 A</i>	신호 변환기 및 전사 5A 의활성 화제	c.G779A	p.R260 Q	해당 없음	3.34 E-05	0.00 0188	보존 된	해로운	HET=66/ HOM=1
<i>RP1L1</i>	색소성 망막염-1-	c.G1493 T	p.S4981	해당 없음	8.76 E-06	4.43E -05	뉴트럴	내성의	HET=16/ HOM=0

[0259]

유전자	유전자 명칭	NT 변화	AA 변화	TGP FRE Q	EXA C FRE Q	RGC FRE Q	보존	예측	유전자형
	유사 1								
<i>ATAD 2</i>	ATPase 패밀리, g2를 함유하는 AAA 도메인	c.C2219 A	p.S740Y	해당 없음	해당 없음	0.00 0102	보존 된	내성의	HET=37/ HOM=0
<i>RBM4 3</i>	RNA 결합 모티프 단백질 43	c.C35G	p.A12G	해당 없음	3.33 E-05	0.00 0147	뉴트럴	내성의	HET=53/ HOM=0

[0260]

[0261]

표 17에 나열된 변이체 중에서, 마이크로세팔린 1(*MCPH1*)은 원발성 소두증에 대해 보고된 병원성 변이체이다. *MCPH1*의 백만 당 전사(TPM)의 유전자 발현 데이터는 다양한 조직에서 인코딩되고 여러 조직에서 높은 발현을 나

타냈다(도 19 참조).

[0262] 원발성 소두증 타입 1은 연령-연관된 평균 미만에서 3 표준 편차 초과를 벗어나는 머리 둘레를 특징으로 한다. 뇌 무게가 현저하게 감소하고 대뇌 피질이 불균형적으로 작다. 영향받은 개체는 심각한 지적 장애가 있다. 일부 *MCHP1* 환자는 세포 유전학적 제제에서 검출된 많은 수의 프로페이스-유사(prophase-like) 세포 및 불량한 메타페이스 G-밴딩(metaphase G-banding)으로 표시되는 바와 같이 성장 지연, 단신 및 잘못 조절된 염색체 응축이 또한 존재한다.

[0263] 표17

유전자	유전자_명칭	NT_변화	AA_변화	TGP_F REQ	EXAC_FR EQ	RGC_F REQ	보존	예측	유전자형
MC PH1	마이크로 세팔린 1	c.2453- 1G>C	스플라 이싱	해당 없음	0.000016 67	0.0000 3321	보존된	해로운	HET=1 2/HO M=0

[0264]

[0265] 7.8 지중해빈혈

[0266] 지중해빈혈은 정상보다 체내 헤모글로빈이 적고 적혈구 수가 적다는 특징이 있는 유전성 혈액 질환이다. 지중해빈혈에서 낮은 헤모글로빈 및 적은 적혈구는 빈혈을 유발하여 환자를 피로하게 한다.

[0267] 지중해빈혈의 ICD 10 코드는 D56이다.

[0268] 지중해빈혈에 대한 이진 형질이 농축된 혈통이 1급 가족 네트워크로부터 분리되었다.

[0269] 1급 혈통은 단지 하나의 가능한 구조 및 공통 조상의 3명 이상의 영향받은 개체를 갖는지 확인하기 위해 평가되었다(도 20 참조). 두개의 농축된 혈통이 식별되었다(도 20 참조). 두 혈통은 모두 단지 하나의 가능한 구조를 가지며 3명 이상의 영향받은 개체를 지녔다.

[0270] 농축된 혈통에 대해 수행된 변이체 분석은 표현형과 공동 분리되는 *HBB* 유전자의 가능한 변이체 목록을 생성했다. *HBB* 유전자는 베타-글로빈이라는 단백질을 만들기 위한 명령을 제공한다. 베타-글로빈은 적혈구 내부에 위치하는 헤모글로빈이라는 더 큰 단백질의 구성요소(서브유닛)이다. 성인에서 헤모글로빈은 일반적으로 4개의 단백질 서브유닛으로 구성된다: 베타-글로빈의 두 서브유닛 및 HBA라는 다른 유전자에서 생성되는 알파-글로빈이라는 다른 단백질의 두 서브유닛. 이들 단백질 서브유닛 각각은 헴이라고 하는 철분-함유 분자에 부착(결합)된다; 각 헴은 하나의 산소 분자에 결합할 수 있는 철 분자를 중심에 포함한다. 적혈구 내의 헤모글로빈은 폐의 산소 분자에 결합한다. 이 세포는 혈류를 통해 이동하여 몸 전체의 조직에 산소를 전달한다. *HBB* 유전자와 연관된 질병에는 베타-지중해빈혈과 겸상 적혈구 빈혈이 포함된다.

[0271] 표현형과 공동 분리된 *HBB* 유전자에서 식별된 두개의 돌연변이는 Gln40에서 종결코돈 추가 돌연변이 및 Gly84에서 구조이동 돌연변이였다(연관 분석 p-값은 $< 3.1 \times 10^{-19}$ 이다). 이러한 식별된 돌연변이를 연구할 수 있으며 이 지식을 사용하여 가족성 지중해빈혈을 치료할 수 있는 가능한 치료 방법을 추가로 개발할 수 있다.

[0272] 7.10 감소된 알칼리 포스파타제 외래 환자 중심 경향 값

[0273] 알칼리 포스파타제에 대한 일상적인 실험실 테스트는 증상이 있는 환자의 진단 목적 및 무증상 환자의 선별 목적 모두를 위해 병원에서 매우 자주 수행된다. 알칼리 포스파타제 효소는 몸 전체의 조직에 존재하지만, 간 및 뼈 질환 환자에서 가장 자주 증가된다.

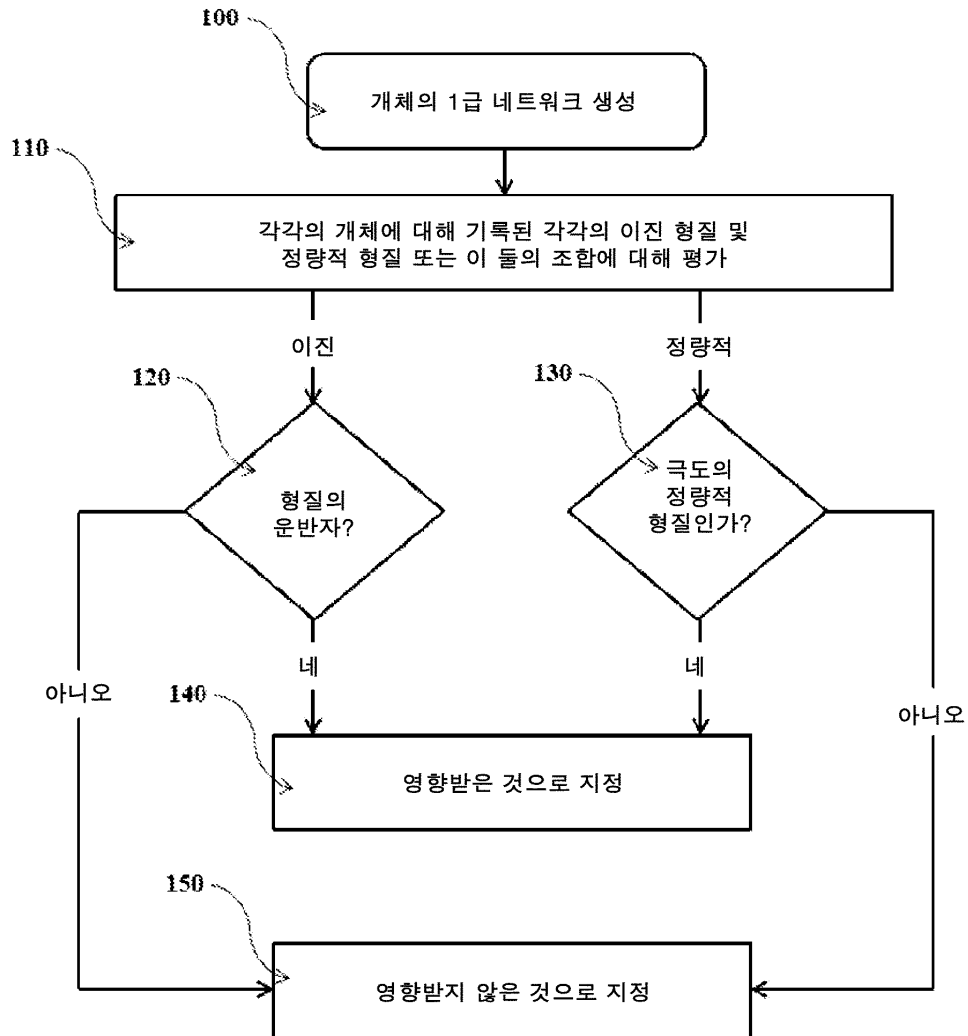
[0274] 감소된 알칼리 포스파타제 수준에 대해 농축된 혈통이 생성되었고, 단지 하나의 가능한 구조를 가지며 공통 조상의 적어도 3명의 영향받은 개체를 가지는지 확인하기 위해 평가되었다(도 21 참조).

[0275] 농축된 혈통에 대해 수행된 변이체 분석은 *ALPL* 유전자의 미스센스 돌연변이가 표현형과 공동 분리되었음을 나타냈다. *ALPL* 유전자는 조직-비특이적 알칼리 포스파타제(TNSALP)라는 효소를 만들기 위한 명령을 제공한다. 이 효소는 뼈와 치아의 성장과 발달에 중요한 역할을 한다. 또한 다른 많은 조직, 특히 간과 신장에서도 활성이 있다. 이 효소는 포스파타제 역할을 하는데, 이는 다른 분자에서 산소 및 인 원자(인산기)의 클러스터를 제거한다는 것을 의미한다. TNSALP는 칼슘 및 인과 같은 미네랄이 발달하는 뼈와 치아에 침착되는 미네랄화 과정에 필수적이다. 미네랄화는 강하고 단단한 뼈와 씹기와 갈기를 견딜 수 있는 치아의 형성에 매우 중요하다. *ALPL* 유전자에서 식별된 이형접합 미스센스 돌연변이는 Leu275(Leu275Pro)(도 21 참조) (연관 분석 p-값은 $< 7.2 \times 10^{-27}$ 이

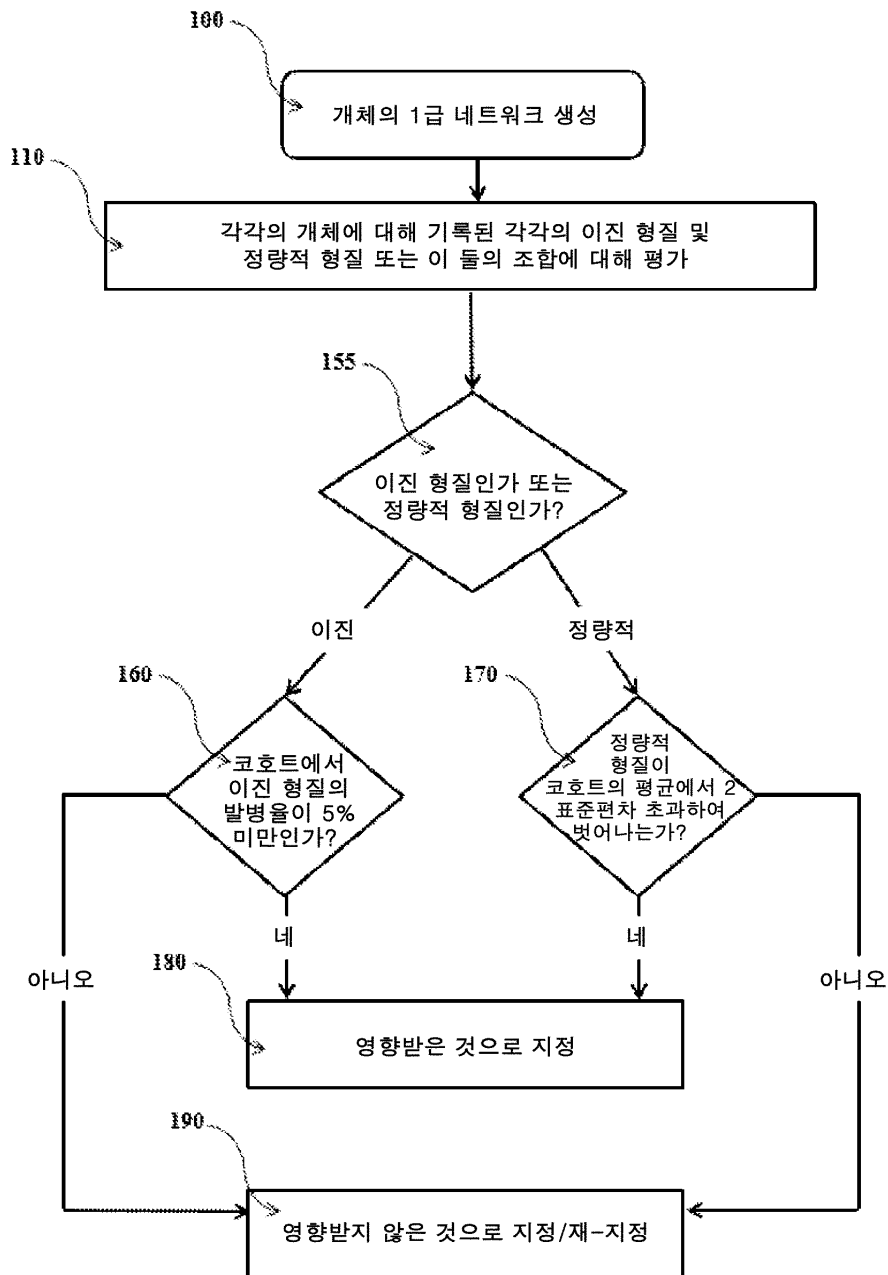
다)에 존재했다. 연관 결과 및 멘델 분리는 변이체 간의 연관성 및 정량적 형질의 감소에 대한 다소 독립적인 증거를 제공한다.

도면

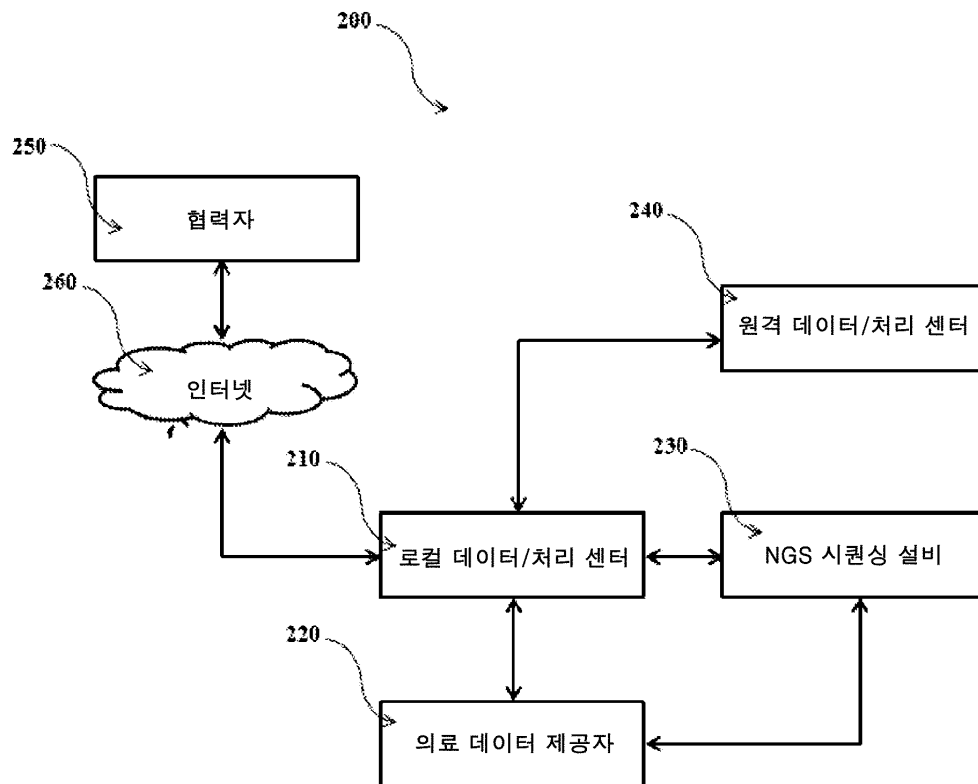
도면1



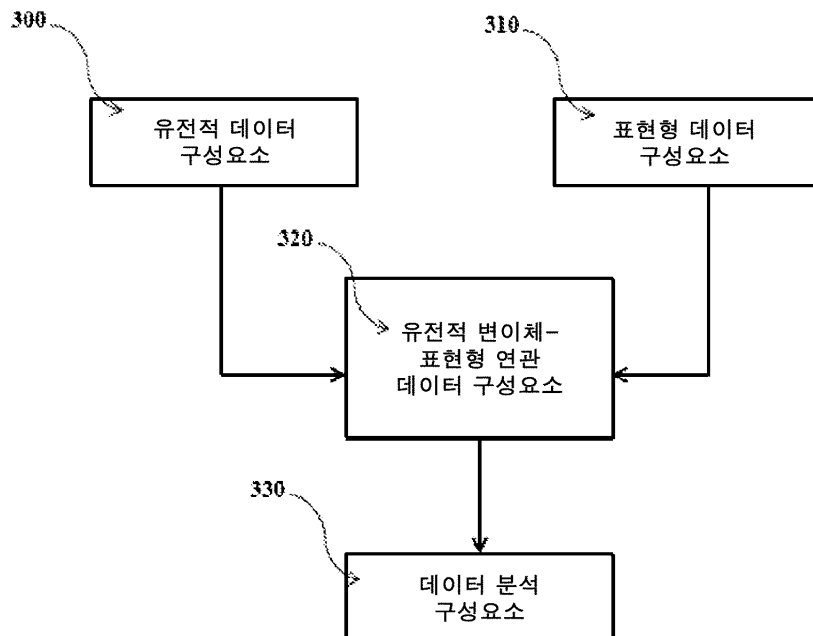
도면2



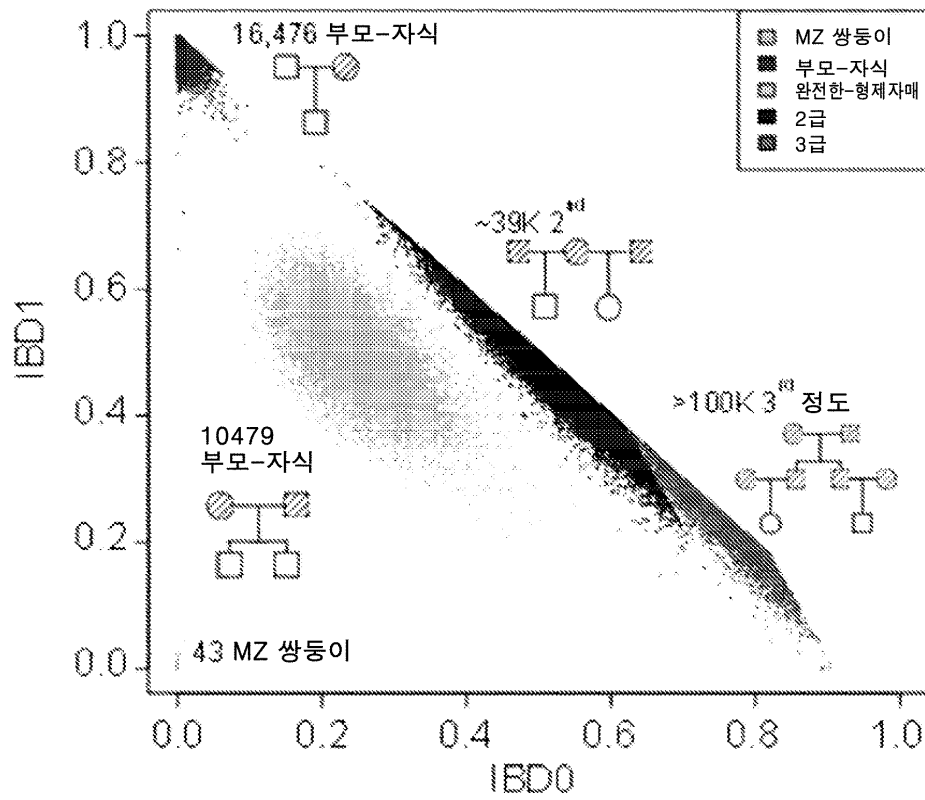
도면3



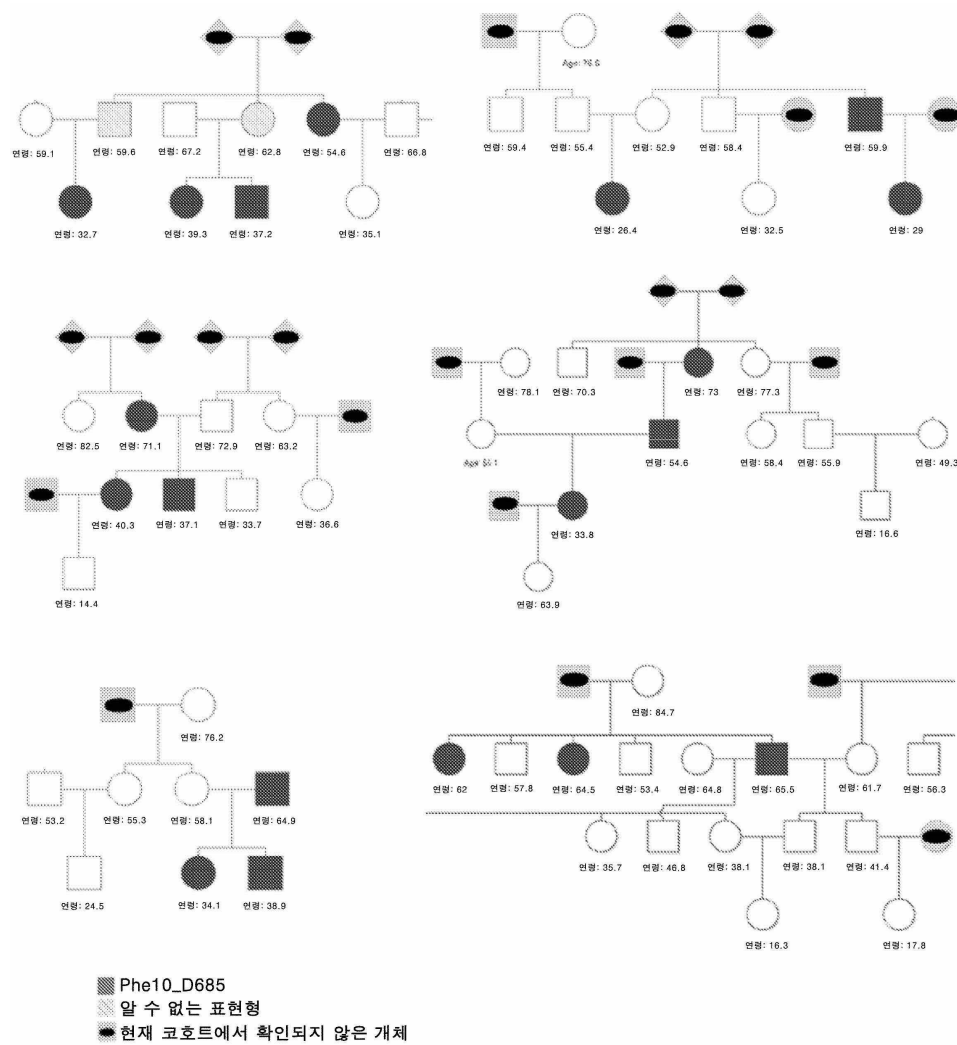
도면4



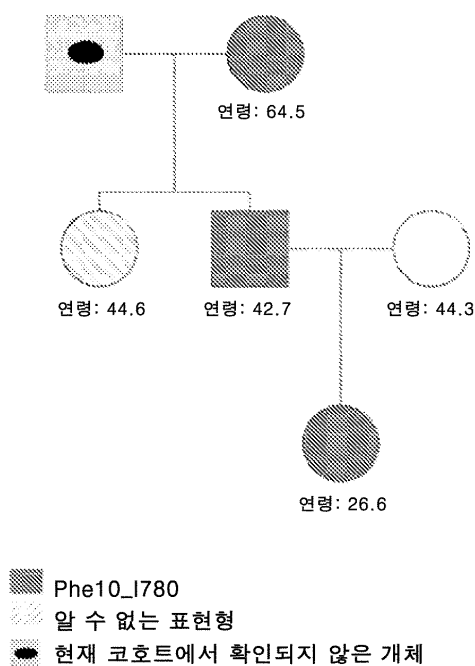
도면5



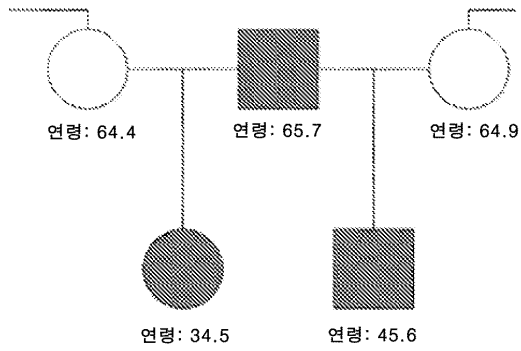
도면6



도면7a

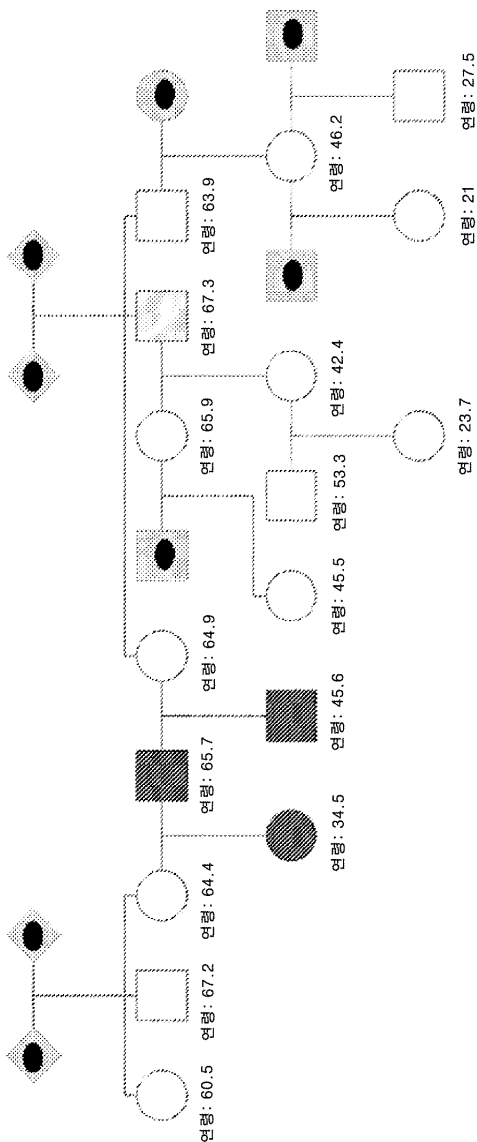


도면7b



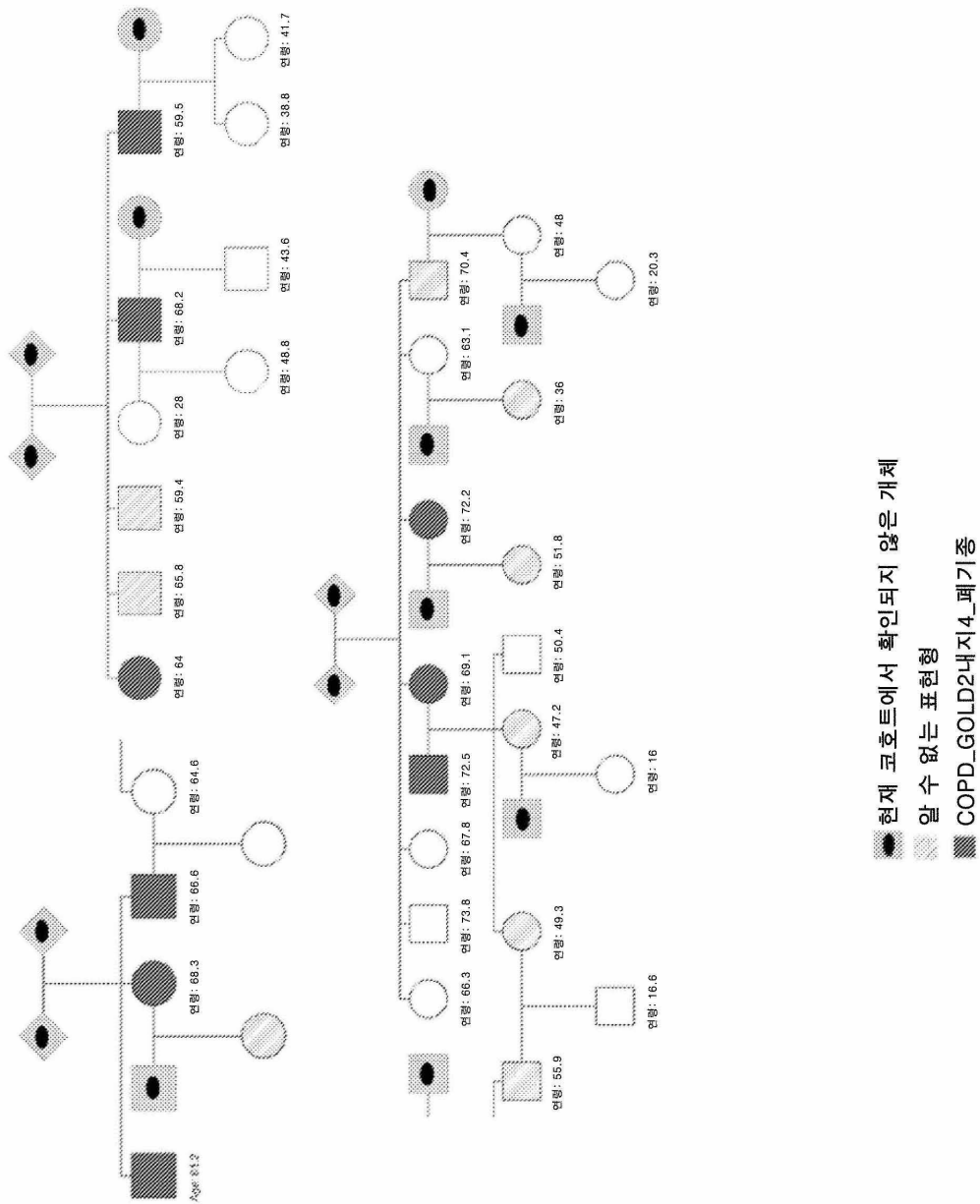
 Phe10_I780
 알 수 없는 표현형
 현재 코호트에서 확인되지 않은 개체

도면8

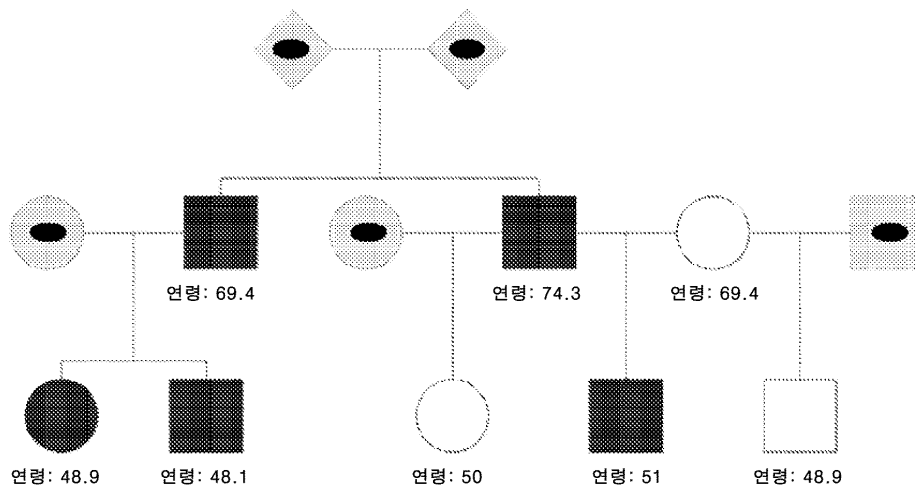


 Phe10_I780
 알 수 없는 표현형
 현재 코호트에서 확인되지 않은 개체

도면9



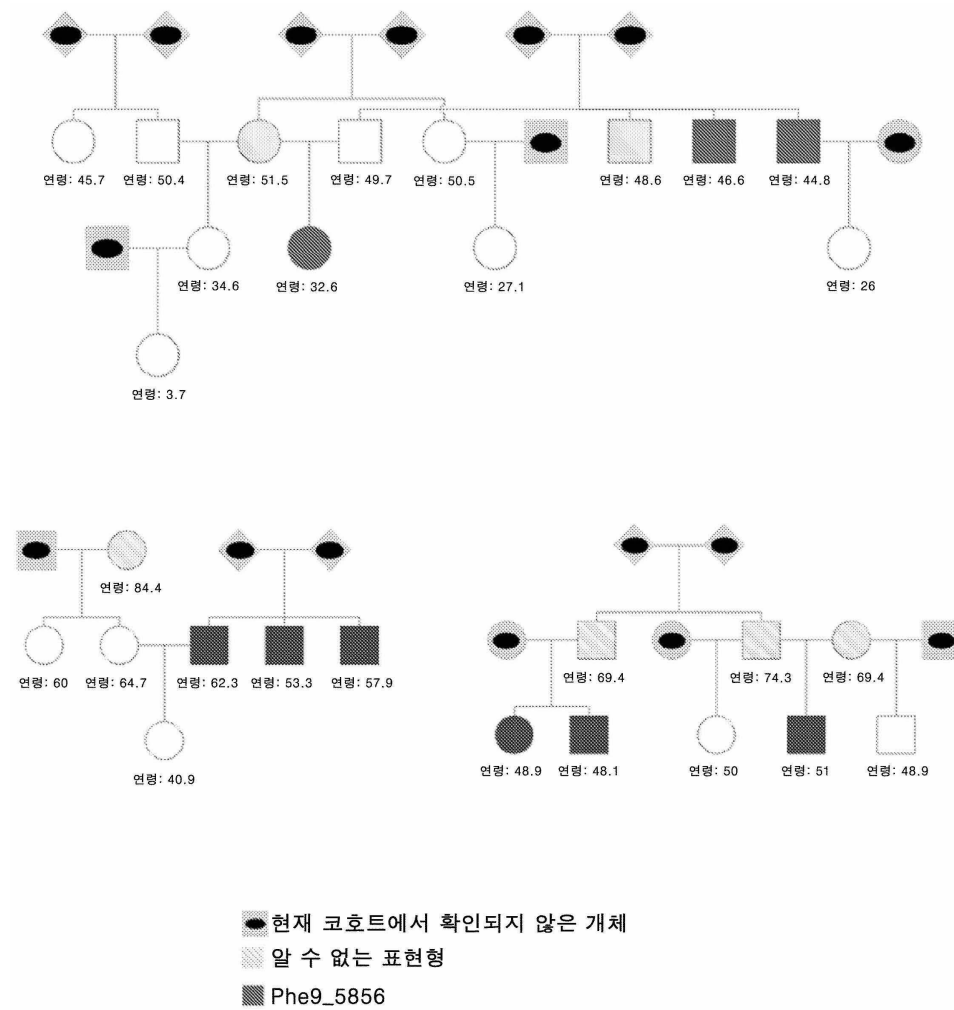
도면10



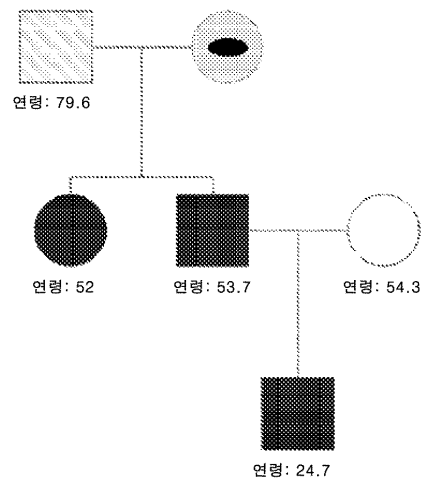
■ Phe9_V420

● 현재 코호트에서 확인되지 않은 개체

도면11

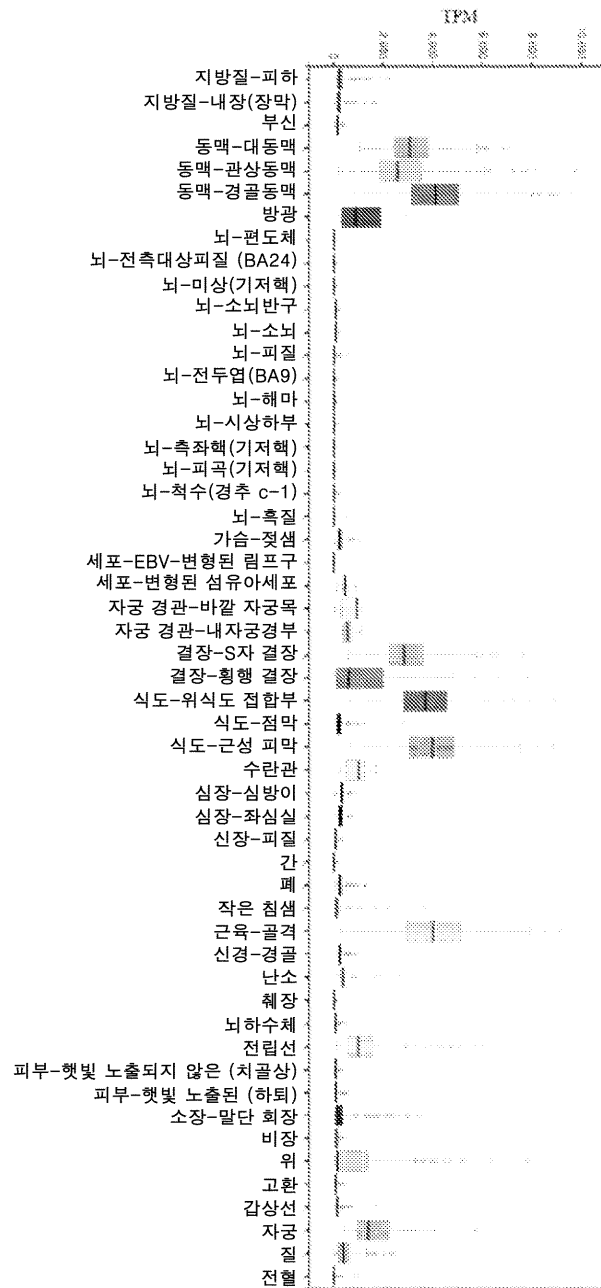


도면12

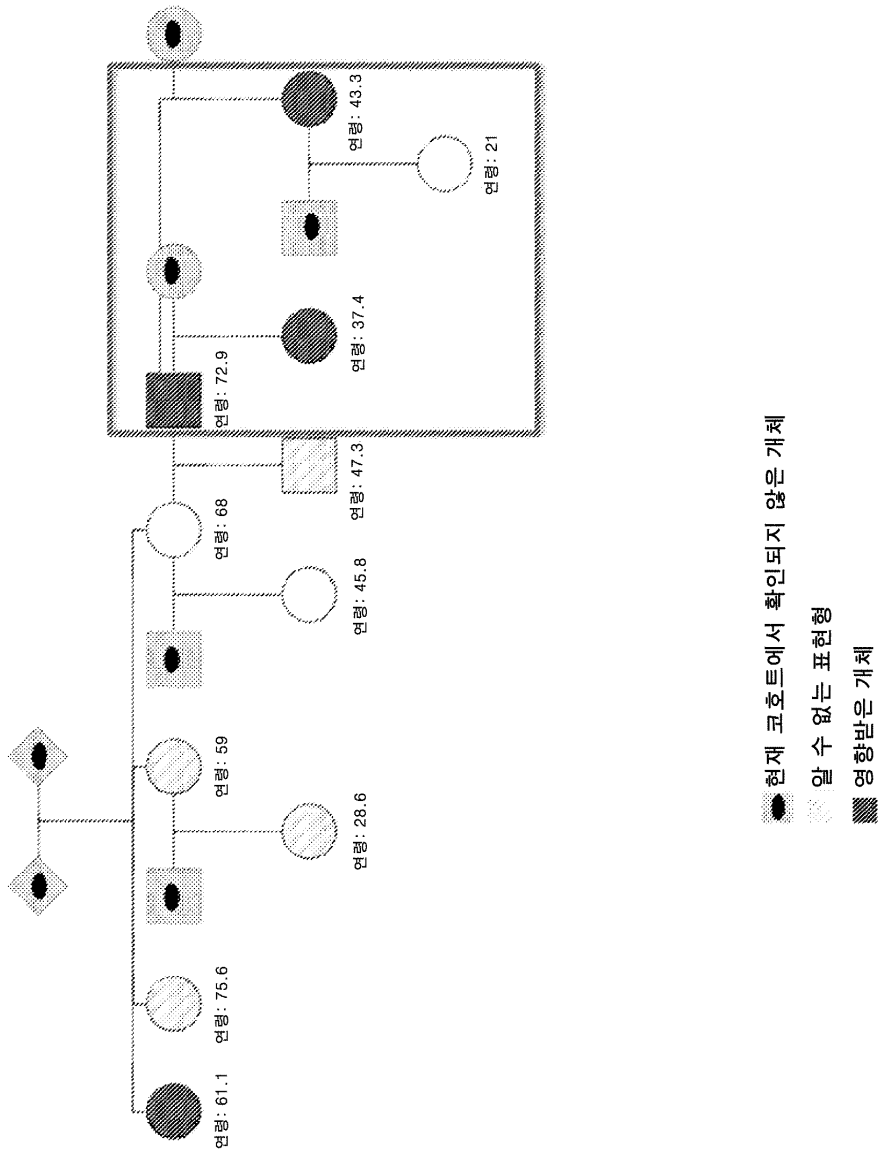


- 현재 코호트에서 확인되지 않은 개체
- 알 수 없는 표현형
- Phe10_G600

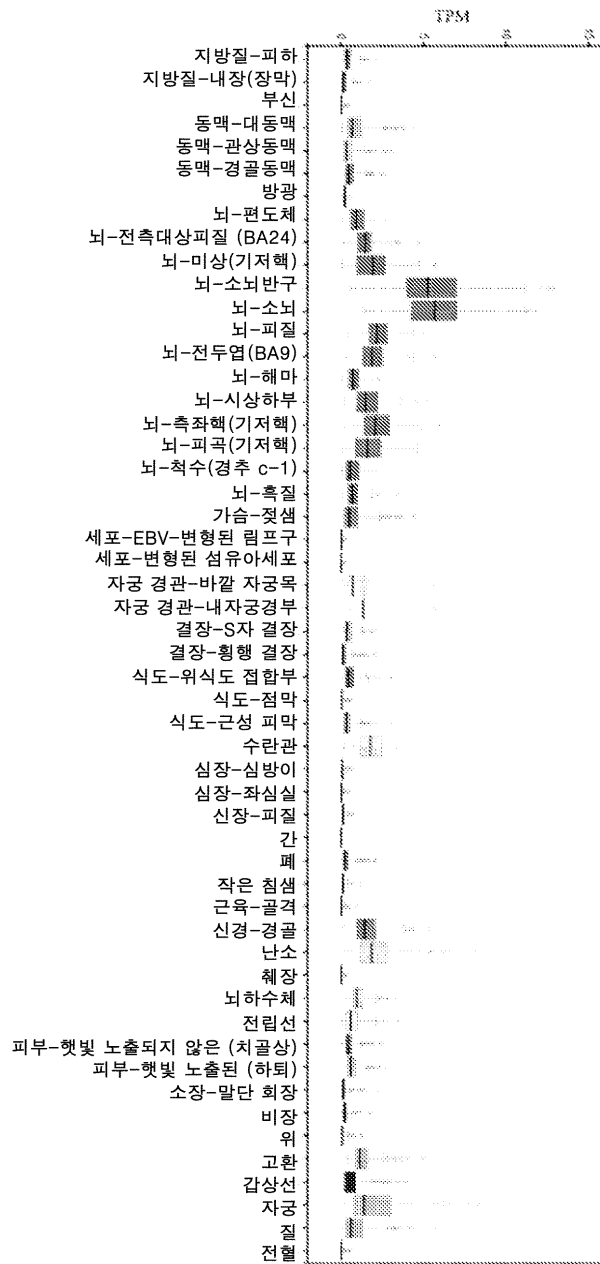
도면13



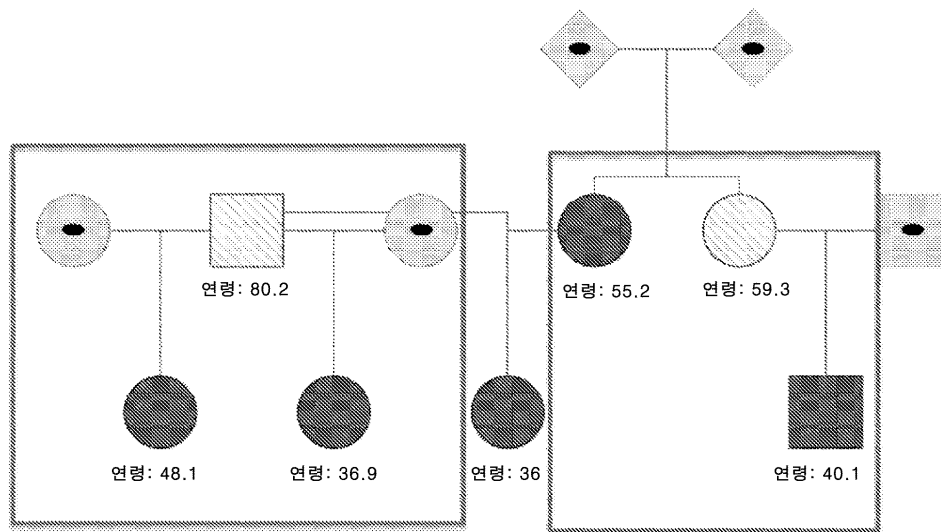
도면14



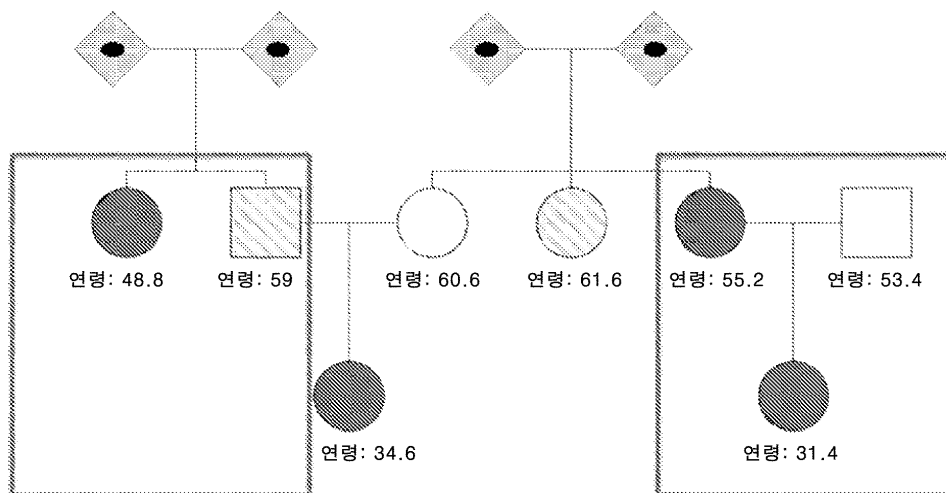
도면15



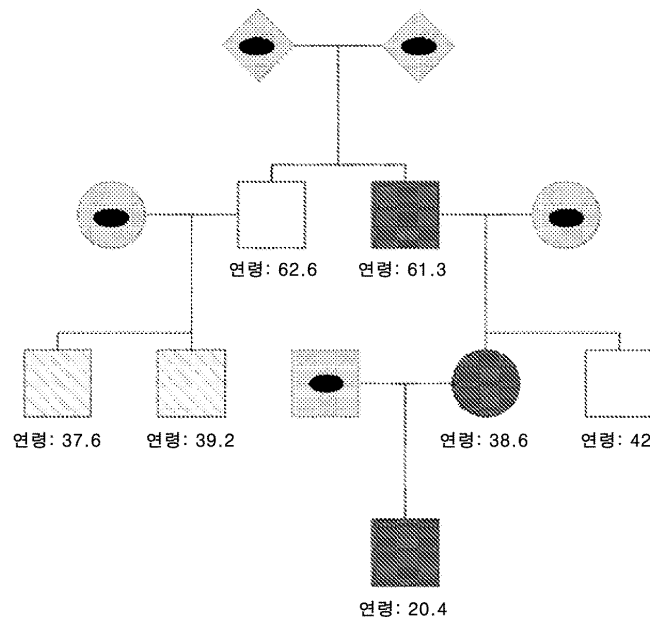
도면16



도면17

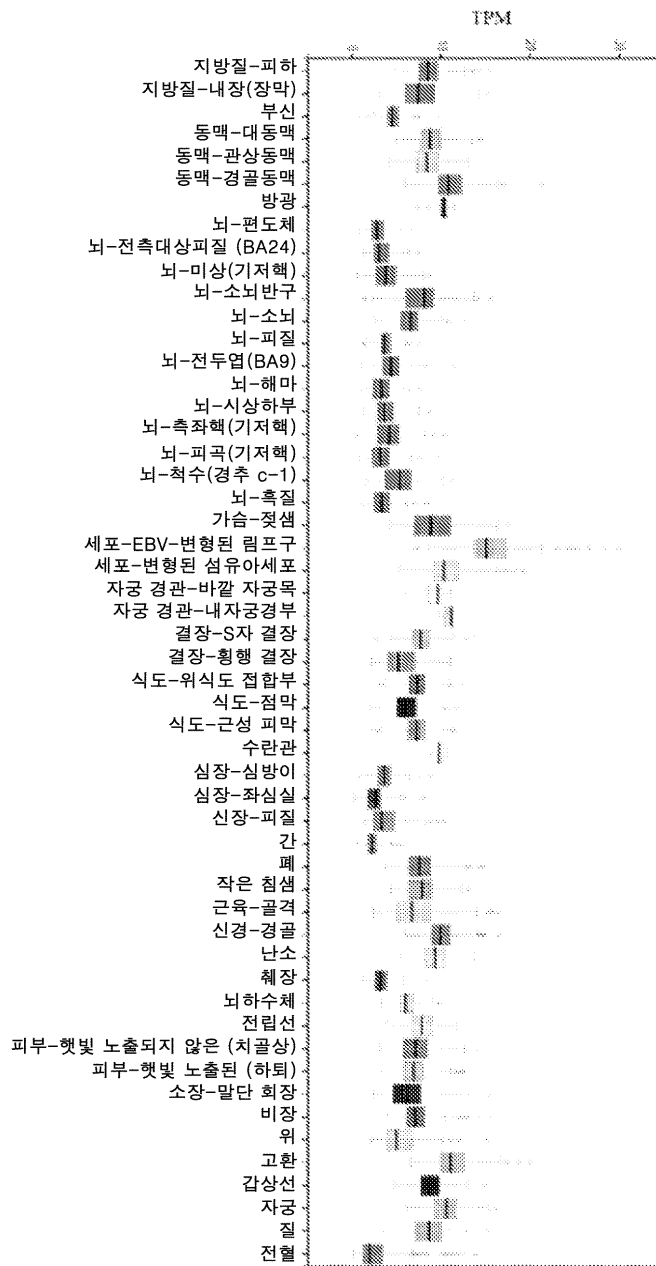


도면18

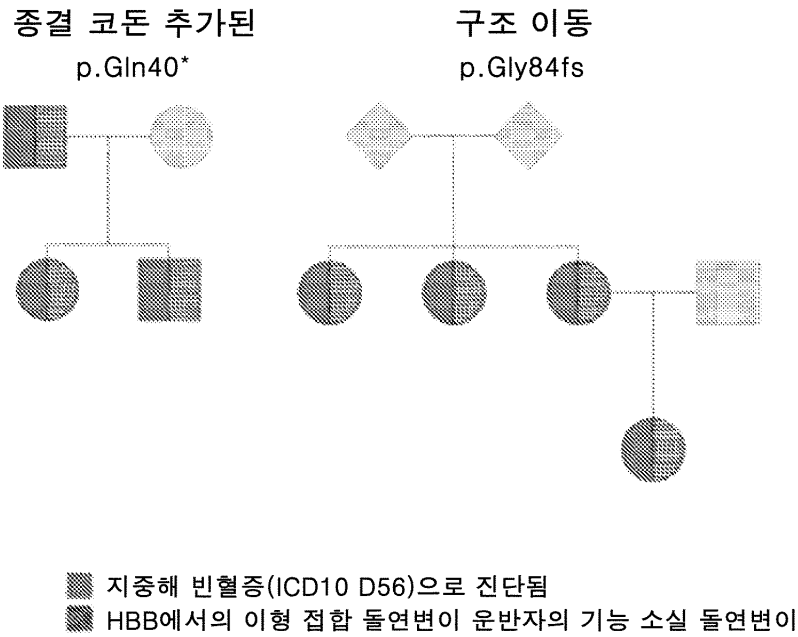


-  현재 코호트에서 확인되지 않은 개체
-  알 수 없는 표현형
-  영향받은 개체

도면19



도면20



도면21

