

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0120782

(43) 공개일자 2021년10월07일

(51) 국제특허분류(Int. Cl.)

G16B 20/40 (2019.01) C12Q 1/6869 (2018.01)

G16B 20/20 (2019.01) G16B 40/00 (2019.01)

(52) CPC특허분류

G16B 20/40 (2019.02)

C12Q 1/6869 (2018.05)

(21) 출원번호 10-2020-0080295

(22) 출원일자 2020년06월30일

심사청구일자 2020년06월30일

(30) 우선권주장

1020200037631 2020년03월27일 대한민국(KR)

(71) 출원인

울산과학기술원

울산광역시 울주군 언양읍 유니스트길 50

주식회사 클리노믹스

울산광역시 울주군 언양읍 유니스트길 50, 110동
3층 301-3호

(72) 발명자

박종화

울산광역시 울주군 언양읍 유니스트길 50

이세민

울산광역시 울주군 언양읍 유니스트길 50

(뒷면에 계속)

(74) 대리인

김순용

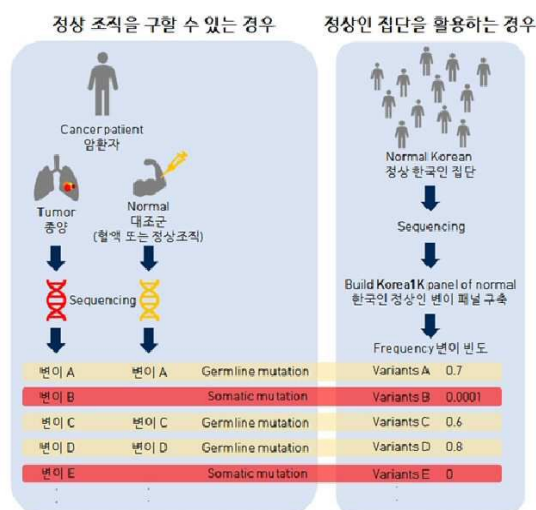
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법

(57) 요약

본 발명은 집단의 변이체 특성이 반영된 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트 구축 방법에 관한 것이다. 본 발명에 따른 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트는 피검 대상과 특정 정보를 공유하는 서브집단의 변이체 정보를 토대로 구축된 것이므로, 피검 대상의 변이체 정보를 보다 빠르고 정확하게 분석하여 질병 등에 대한 변이체 정보를 제공할 수 있다. 이에, 본 발명에 따른 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트는 유전체 분석 분야; 피검 대상의 질병 진단 또는 예후 예측 분야; 및 개인 맞춤형 치료 분야;에서 다양하게 활용될 수 있다.

대표도 - 도1



(52) CPC특허분류

G16B 20/20 (2019.02)

G16B 40/00 (2019.02)

(72) 발명자

최연송

울산광역시 울주군 언양읍 유니스트길 50

전성원

울산광역시 울주군 언양읍 유니스트길 50

박영준

울산광역시 울주군 언양읍 유니스트길 50

김병철

경기도 수원시 영통구 봉영로1744번길 11, 222동
802호(영통동, 벽산.풍림아파트)

조윤성

경기도 용인시 기흥구 흥덕1로79번길 37, 503동
1702호(영덕동, 흥덕마을5단지 호반베르디움)

김병철

울산광역시 울주군 언양읍 유니스트길 50

명세서

청구범위

청구항 1

집단의 유전체 정보로부터 피검 대상에 대응되는 서브집단의 유전체 정보를 추출하는 단계; 및
상기 서브집단의 유전체 정보로부터 서브집단의 변이체 정보를 수집하는 단계;를 포함하는, 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법.

청구항 2

제1항에 있어서,

상기 집단의 유전체 정보는 단일 염기 변이(single nucleotide variants, SNV), 삽입 결실(indels), 복제수 변형(copy number variations, CNV), 전이인자(transposable element, TE), 삽입(insertion) 및 HLA(human leukocyte antigen)로 이루어진 군에서 선택된 1종 이상의 정보가 반영된 것인, 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법.

청구항 3

제1항에 있어서,

상기 추출 단계는 집단의 유전체 정보로부터 피검 대상과 국가, 인종, 성별, 나이, 표현형 및 임상정보로 이루어진 군에서 선택된 1 이상의 특징을 공유하는 서브집단을 추출하는 것인, 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법.

청구항 4

제1항에 있어서,

상기 방법은 서브집단의 변이체 정보 중 질병 관련 변이체 정보에 주석(annotation)을 다는 단계;를 더 포함하는, 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법.

청구항 5

제4항에 있어서,

상기 방법은 주석 표시된 질병 관련 변이체 정보를 기반으로 피검 대상의 유전체 중 질병 관련 변이체를 분류하는 단계를 더 포함하는, 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법.

청구항 6

제5항에 있어서,

상기 변이체 분류는 대립 유전자 발생 빈도에 따라 생식세포 변이체(germline variant) 또는 체세포 변이체(somatic variant)로 분류하는 것인, 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법.

청구항 7

제5항에 있어서, 상기 변이체 분류는 대립 유전자 발생 빈도에 따라 생식세포 변이체(germline variant)를 분류하는 것인, 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법.

청구항 8

제1항 내지 제7항 중 어느 한 항에 따른 방법으로 구축된, 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트.

청구항 9

제8항의 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트 및 피검 대상의 데이터 세트의 변이체를 비교하는 단계;를 포함하는, 피검 대상 맞춤형 변이체 정보 제공 방법.

청구항 10

제9항에 있어서,

상기 비교 단계는 피검 대상에서 특이적으로 발생된 변이체 및 공통으로 발생된 변이체를 확인하는 것인, 피검 대상 맞춤형 변이체 정보 제공 방법.

청구항 11

제9항에 있어서,

상기 비교 단계는 변이체의 대립 유전자 발생 빈도를 비교하는 것인, 피검 대상 맞춤형 변이체 정보 제공 방법.

청구항 12

제11항에 있어서,

상기 대립 유전자 발생 빈도는 컷오프 값이 0.5 내지 3%인, 피검 대상 맞춤형 변이체 정보 제공 방법.

청구항 13

제11항에 있어서,

상기 방법은 대립 유전자 발생 빈도에 따라 생식세포 변이체(germline variant) 또는 체세포 변이체(somatic variant)로 구분하는 단계를 더 포함하는, 피검 대상 맞춤형 변이체 정보 제공 방법.

청구항 14

제10항에 있어서,

상기 방법은 대립 유전자 발생 빈도에 따라 생식세포 변이체(germline variant)를 분류하는 것인, 피검 대상 맞춤형 변이체 정보 제공 방법.

청구항 15

제1항 내지 제7항 중 어느 한 항의 방법으로 구축된 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트에서, 주석 표시된 질병 관련 변이체를 확인하는 단계;를 포함하는 피검 대상 맞춤형 질병 관련 변이체 정보 제공 방법.

청구항 16

제1항 내지 제7항 중 어느 한 항에 따른 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

청구항 17

제9항에 따른 피검 대상 맞춤형 변이체 정보 제공 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

청구항 18

한국인 집단의 유전체 정보로부터 한국인 피검 대상에 대응되는 서브집단의 유전체 정보를 추출하는 단계; 및

상기 서브집단의 유전체 정보로부터 서브집단의 변이체 정보를 수집하는 단계;를 포함하는, 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법.

발명의 설명

기술 분야

[0001] 본 발명은 집단의 변이체 특성이 반영된 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트 구축 방법에 관한 것이다.

배경 기술

[0002] 전 세계 약 8,500만 명으로 추산되는 한국 인종의 유전자의 총합인 유전체에 존재하는 변이의 양 및 구조는 대규모 유전적 혼합(few large-scale admixture events) 사건들이 적어 매우 균질한 것으로 추측된다. 하지만 이러한 추측에 대한 공식적인 대규모 유전체 연구가 거의 이루어지지 않는 상황이다.

[0003] 그러나 한국 최초의 참조표준 유전체의 해독(KOREF_S)과 40명의 친척 관계가 없는 한국인 그룹표준(KOREF_C)에 대한 표준 유전체 연구가 최근에 진행되어, 표준화된 소수의 한국인 유전체는 존재했다. 2008년 발표된 최초의 한국인 유전체인 김성진유전체(SJK) 이후, 여러 종류의 한국의 전장 유전체와 발현유전체(엑솜, exomes) 연구가 보고된바 있으며, 이들 연구를 토대로 한국 유전체의 변이체 데이터베이스인 코베리움(KoVariome) 데이터 베이스가 형성되었다.

[0004] 현재, 전 세계적으로 인간 유전체 다양성을 연구하는 다인종 유전체 프로젝트 등이 진행되고 있다. 영국에서 시작된 1,000개의 유전체 프로젝트(1,000 genome project, 1KGP)는 중국 및 일본을 포함한 2,500개 이상의 유전체가 포함되어 있으나, 한국인의 유전체 데이터는 포함되어 있지 않다.

[0005] 그러나 많은 인간 샘플을 대상으로 하여 수행된 선행 대규모 유전체 시퀀싱 프로젝트에도 불구하고, 분석된 유전체 정보와 질병 간의 상관관계에 대한 임상 데이터가 거의 없거나, 참가자의 유전자-표현형 연관에 관한 제한적인 정보만이 수집되었으며, 이를 통해 생성된 인구의 건강 및 질병 상태(population's health and disease states)가 제한적으로 수집된 상황이다.

[0006] 한편, 희귀질환이나 암과 같은 특정 질환에 관련된 유전자 변이를 찾는 것은 해당 질병을 갖는 사람을 검진하고 치료함에 있어서 첫 단계이자 가장 중요한 단계이다. 유전자 변이 분석을 정밀하고 정확하게 하기 위해서는 질병을 가진 사람 또는 질병 조직의 유전정보뿐만 아니라 건강한 사람 또는 정상적인 조직의 유전정보가 반드시 필요하다. 그러나, 건강한 사람 또는 정상적인 조직의 유전정보를 확보하는 것에는 추가적인 비용이 소요되며, 경우에 따라서는 건강한 사람 또는 정상적인 조직을 구하지 못하는 경우도 발생한다.

발명의 내용

해결하려는 과제

[0007] 이에 본 발명자들은 집단의 유전체 정보 및 표현형 정보에 기초하여 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트를 구축함으로써 본 발명을 완성하게 되었다.

[0008] 따라서 본 발명의 목적은, 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법 및 이의 방법으로 구축된 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트를 제공하는 것이다.

[0009] 본 발명의 다른 목적은 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트 및 피검 대상의 데이터 세트의 변이체를 비교하는 단계;를 포함하는 피검 대상 맞춤형 변이체 정보 제공 방법을 제공하는 것이다.

[0010] 본 발명의 또 다른 목적은 구축된 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트에서, 주석 표시된 질병 관련 변이체를 확인하는 단계;를 포함하는 피검 대상 맞춤형 질병 관련 변이체 정보 제공 방법을 제공하는 것이다.

[0011] 본 발명의 또 다른 목적은 상기 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법; 또는 피검 대상 맞춤형 변이체 정보 제공 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공하는 것이다.

[0012] 본 발명의 또 다른 목적은 한국인 집단의 유전체 정보로부터 한국인 피검 대상에 대응되는 서브집단의 유전체 정보를 추출하는 단계; 및 상기 서브집단의 유전체 정보로부터 서브집단의 변이체 정보를 수집하는 단계;를 포함하는, 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법을 제공하는 것이다.

과제의 해결 수단

- [0013] 상기 목적을 달성하기 위하여, 본 발명은 집단의 유전체 정보로부터 피검 대상에 대응되는 서브집단의 유전체 정보를 추출하는 단계; 및 상기 서브집단의 유전체 정보로부터 서브집단의 변이체 정보를 수집하는 단계;를 포함하는 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법을 제공한다.
- [0014] 또한 본 발명은 상기 방법으로 구축된 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트를 제공한다.
- [0015] 또한 본 발명은 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트 및 피검 대상의 데이터 세트의 변이체를 비교하는 단계;를 포함하는, 피검 대상 맞춤형 변이체 정보 제공 방법을 제공한다.
- [0016] 또한 본 발명은 상기 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공한다.
- [0017] 또한 본 발명은 상기 피검 대상 맞춤형 변이체 레퍼런스 데이터 세트에서, 주식 표시된 발명 관련 변이체를 확인하는 단계;를 포함하는 피검 대상 맞춤형 질병 관련 변이체 정보 제공 방법을 제공한다.
- [0018] 또한 본 발명은 상기 피검 대상 맞춤형 변이체 정보 제공 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공한다.
- [0019] 또한 본 발명은 한국인 집단의 유전체 정보로부터 한국인 피검 대상에 대응되는 서브집단의 유전체 정보를 추출하는 단계; 및 상기 서브집단의 유전체 정보로부터 서브집단의 변이체 정보를 수집하는 단계;를 포함하는, 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법을 제공한다.

발명의 효과

- [0020] 본 발명에 따른 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트는 피검 대상과 특정 정보를 공유하는 서브집단의 변이체 정보를 토대로 구축된 것이므로, 피검 대상의 변이체 정보를 보다 빠르고 정확하게 분석하여 질병 등에 대한 변이체 정보를 제공할 수 있다. 이에, 본 발명에 따른 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트는 유전체 분석 분야; 피검 대상의 질병 진단 또는 예후 예측 분야; 및 개인 맞춤형 치료 분야;에서 다양하게 활용될 수 있다.

도면의 간단한 설명

- [0021] 도 1은 정상조직을 구할 수 없는 경우에 정상인 집단의 변이체 정보를 이용하여 암환자의 생식세포 돌연변이를 대립유전자 발생 빈도수를 이용하여 발굴하는 과정을 나타낸 도이다.
- 도 2는 인종별 정상인 데이터 세트를 사용하여 변이체 분류(variant classification)의 퍼포먼스를 확인한 결과를 나타낸 도이다(a: 분류의 정확성(ACC); b: Matthews correlation coefficient values, c: Germline recovery rate). 도 2에 있어서, X 축의 숫자는 각 패널에서 변이체를 선택하기 위한 대립유전자 발생 빈도 컷오프를 나타낸다.
- 도 3은 대립유전자 발생 빈도 설정을 달리하면서 인종별 정상인 데이터 세트가 갖는 CGC(Cancer Gene Census) 유전자 밀도를 측정한 결과를 나타낸 도이다.

발명을 실시하기 위한 구체적인 내용

- [0022] 이하, 본 발명을 상세히 설명한다.
- [0023] 본 발명의 양태에 따르면, 본 발명은 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법 및 상기 방법으로 구축된 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트를 제공한다.
- [0024] 상기 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법은 집단의 유전체 정보로부터 피검 대상에 대응되는 서브집단의 유전체 정보를 추출하는 단계; 및 상기 서브집단의 유전체 정보로부터 서브집단의 변이체 정보를 수집하는 단계;를 포함한다.
- [0025] 본 발명의 구체예에서, 상기 피검 대상은 1 이상의 개체인 것이 바람직하다.
- [0026] 본 발명에 구체예에서, 상기 집단의 유전체 정보는 단일 염기 변이(single nucleotide variants, SNV), 삽입결실(indels), 복제수 변형(copy number variations, CNV), 전이인자(transposable element, TE), 삽입(insertion) 및 HLA(human leukocyte antigen)로 이루어진 군에서 선택된 1종 이상의 정보가 반영된 것일 수 있다.

- [0027] 본 발명에 있어서, 변이체(variome)는 발생된 모든 유전자의 변이 정보를 의미한다. 상기 변이체는 동일한 유전자 변이라 할지라도, 집단의 특징에 따라 서로 다른거나 정반대의 영향을 나타낼 수 있다. 일례로, 결핵(tuberculosis)의 감수성과 연관성이 있는 것으로 보고된 18q11 유전자 영역의 rs4331426 SNP 위치의 경우, 중국인 집단에서는 대립유전자 A(아데닌)이 결핵에 대한 높은 감수성과 연관되어 있음이 보고되는 반면, 아프리카인 집단에서는 대립유전자 G(구아닌)이 결핵에 대한 높은 감수성(즉, A allele은 아프리카인 집단에서 결핵에 대한 낮은 감수성)과 연관되어 있음이 보고된다. 이에, 해당 집단에 기반한 변이체 데이터 세트를 활용함으로써, 연구, 질병 예측 및 진단 등의 정확성을 높일 수 있다.
- [0028] 본 발명의 구체예에서, 상기 추출 단계는 집단의 유전체 정보로부터 피검 대상과 공통적인 특징을 갖는 집단의 유전체 정보를 추출하는 것일 수 있으며, 바람직하게는 피검 대상과 국가, 인종, 성별, 나이, 표현형 및 임상정보로 이루어진 군에서 선택된 1 이상의 특징을 공유하는 서브집단을 추출한 것이 바람직하나, 이에 제한되지 않는다. 또한 상기 서브집단은 상기 1 이상의 특징 중 다른 특징을 통해 피검대상과 구별될 수 있으며, 이에 제한되지 않는다. 또한 상기 서브집단은 상기 선택된 1 이상의 특징을 피검대상과 공유하고, 다른 1 이상의 특징을 통해 피검대상과 구별되는 것일 수 있다.
- [0029] 상기와 같은 서브집단을 추출하여 레퍼런스 데이터 세트를 구축함으로써, 피검 대상, 즉 환자에 특이적인 변이를 보다 정확하고 빠르게 획득할 수 있다.
- [0030] 본 발명에 있어서, 인종(ethnic group)은 민족의 과학적인 용어로, 인종 내의 다양한 족을 의미한다. 민족은 문화적·사회적 용어이므로 유전적 다양성이 큰 모든 사회구성원을 뜻할 수 있다. 반면에 인종은 상대적으로 단일한 족을 뜻한다. 인종은 같은 언어를 쓸 수도 있고, 다른 언어를 쓸 수 있는바, 어족(language family)과 구분되는 개념이다. 예를 들어, 현재의 유럽인들은 거대한 유럽인족이며, 서로간의 유전적 거리가 가깝다. 중국에는 약 50개의 인종이 있다고 일반적으로 기술되나, 실제로, 유전적 다양성을 고려하면, 서로 거리가 가까워서, 그 수가 50개보다 적다. 따라서 중국은 50개의 민족이 있는데, 서로 문화적으로 달라보일지라도 하나의 인종으로 묶일 수 있다. 반대로, 문화적으로는 같아 보이나 실제 유전상의 거리는 매우 멀 수도 있다. 한국의 경우 다양한 지방이 있으나, 하나의 인종이다.
- [0031] 본 발명에서는 인종 정보를 반영한 서브집단을 이용하므로, 보다 정확한 인종특이적 변이체 정보를 제공할 수 있다.
- [0032] 본 발명의 구체예에서, 상기 방법은 서브집단의 변이체 정보 중 질병 관련 변이체 정보에 주석(annotation)을 다는 단계를 더 포함할 수 있다.
- [0033] 본 발명의 바람직한 구체예에서, 상기 방법은 주석 표시된 질병 관련 변이체 정보를 기반으로 피검 대상의 유전체 중 질병 관련 변이체를 분류하는 단계를 더 포함할 수 있다.
- [0034] 본 발명의 보다 바람직한 구체예에서, 상기 변이체 분류는 대립 유전자 발생 빈도에 따라 생식세포 변이체(germline variant) 또는 체세포 변이체(somatic variant)로 분류하는 것일 수 있다. 분류된 생식세포 변이체 및 체세포 변이체의 비교를 통해 암 등과 같은, 체세포 변이체의 영향을 받는 질환과 관련된 변이체 정보를 제공하는 데에 활용될 수 있다.
- [0035] 본 발명의 보다 바람직한 구체예에서, 상기 변이체 분류는 대립 유전자 발생 빈도에 따라 생식세포 변이체를 분류하는 것일 수 있다. 분류된 생식세포 변이체 간의 비교를 통해 알츠하이머, 당뇨, 심혈관계 질환 등과 같은, 생식세포 변이체의 영향을 받는 질환과 관련된 변이체 정보를 제공하는 데에 활용될 수 있다.
- [0036] 본 발명에 있어서, 생식세포 변이체는 생식세포에서 돌연변이(mutation)가 나타난 것으로, 부모가 물려주는 유전적인 경향을 의미한다.
- [0037] 본 발명에 있어서, 체세포 변이체는 체세포에서 돌연변이가 나타난 것으로, 배아가 분화되는 과정 또는 분화된 이후에 조상의 생식세포에 없던 돌연변이가 무작위적으로 발생한 것을 의미한다. 체세포 변이체는 자손에 전달되지 않는다.
- [0038] 본 발명에 따른 방법으로 구축된 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트는 (i) 단일 염기 변이, 삽입 결실, 복제수 변형, 전이인자 삽입, HLA 등의 유전적 특성이 반영된 집단의 유전체 정보를 반영하고 있고, (ii) 피검 대상과 국가, 인종, 성별 나이 및 표현형 등의 정보를 공유하고 있다. 따라서 상기 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트는 피검 대상의 변이체 정보를 보다 빠르고 정확하게 분석하여 질병 등에 대한 변이체 정보를 제공할 수 있는바, 유전체 분석 분야; 피검 대상의 질병 진단 또는 예후 예측 분야; 및 개인

맞춤형 치료 분야;에서 다양하게 활용될 수 있다.

- [0040] 본 발명의 다른 양태에 따르면, 본 발명은 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트 및 피검 대상의 데이터 세트의 변이체를 비교하는 단계;를 포함하는 피검 대상 맞춤형 변이체 정보 제공 방법을 제공한다.
- [0041] 본 발명의 구체예에서, 상기 비교 단계는 피검 대상에서 특이적으로 발생된 변이체 및 공통으로 발생된 변이체를 확인하는 것이 바람직하다.
- [0042] 본 발명의 구체예에서, 상기 비교 단계는 변이체의 대립 유전자 발생 빈도를 비교하는 것이 바람직하다.
- [0043] 본 발명에 있어서, 대립 유전자 발생 빈도는 한 개체군 내에서 특정 대립유전자가 차지하는 비율을 의미한다.
- [0044] 본 발명의 바람직한 구체예에서, 상기 대립 유전자 발생 빈도는 컷오프 값이 0.5 내지 3%인 것이 바람직하며, 더 바람직하게는 1 내지 2%이며, 더욱 바람직하게는 1.5%이나, 이에 제한되지 않는다.
- [0046] 본 발명의 바람직한 구체예에서, 상기 방법은 대립 유전자 발생 빈도에 따라 생식세포 변이체(germline variant) 또는 체세포 변이체(somatic variant)로 구분하는 단계를 더 포함할 수 있다. 분류된 생식세포 변이체 및 체세포 변이체의 비교를 통해 암 등과 같은, 체세포 변이체의 영향을 받는 질환과 관련된 변이체 정보를 제공하는 데에 활용될 수 있다.
- [0047] 본 발명의 바람직한 구체예에서, 상기 변이체 분류는 대립 유전자 발생 빈도에 따라 생식세포 변이체를 분류하는 것일 수 있다. 분류된 생식세포 변이체 간의 비교를 통해 알츠하이머, 당뇨, 심혈관계 질환 등과 같은, 생식세포 변이체의 영향을 받는 질환과 관련된 변이체 정보를 제공하는 데에 활용될 수 있다.
- [0048] 본 발명의 보다 바람직한 구체예에서, 상기 컷오프 값은 카이스퀘어 검정(chisquared test), 피셔 검정(fisher test), 로지스틱 회귀분석(logistic regression)등을 포함하는 통계 분석 방법을 이용하여 측정하는 것이 바람직하다.
- [0049] 본 발명의 보다 바람직한 구체예에서, 상기 컷오프 값은 그 유의성(p-value)이 0.05인 것이 바람직하다.
- [0050] 본 발명에 따른 변이체 분석을 통한 피검 대상 맞춤형 변이체 정보 제공 방법은 피검 대상이 속하는 집단의 특성을 반영하고 있는바, 분석 정확도가 현저히 높다. 또한 정상 시료 없이도 특정 질병의 진단을 위한 목적으로 유전체 정보를 분석할 경우 정상 시료의 유전체 분석 데이터가 필요하다. 그러나 정상 시료를 취득할 수 없는 경우가 다수 존재한다는 한계가 존재한다. 본 발명의 피검 대상 맞춤형 변이체 정보 제공 방법은 정상 시료의 취득이 어렵더라도, 변이체 정보를 제공할 수 있다는 장점이 있다. 즉, 본 발명의 방법은 구축된 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트와 피검 대상의 변이체 데이터 세트를 비교함으로써 피검 대상에 대한 맞춤형 변이체 정보를 제공할 수 있다.
- [0052] 본 발명의 또 다른 양태에 따르면, 본 발명은 구축된 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트에서, 주석 표시된 질병 관련 변이체를 확인하는 단계;를 포함하는 피검 대상 맞춤형 질병 관련 변이체 정보 제공 방법을 제공한다.
- [0053] 상기 구축된 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트는 (i) 질병 관련 변이체 정보에 주석 표시되어 있으며, (ii) 상기 주석 표시된 질병 관련 변이체 정보를 기반으로 피검 대상의 유전체 중 질병 관련 변이체가 생식세포 변이체 또는 체세포 변이체로 분류되어 있다.
- [0055] 본 발명의 또 다른 양태에 따르면, 본 발명은 상기 피검 대상 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법; 또는 피검 대상 맞춤형 변이체 정보 제공 방법;을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공한다.
- [0056] 본 발명의 또 다른 양태에 따르면, 본 발명은 변이체 분석을 통한 질병의 진단에 대한 정보 제공 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공한다.
- [0057] 본 발명에 있어서, 상기 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 컴퓨터가 가독 가능한 기록매체의 예로는 마그네틱 저장매체(예를 들면, 플로피 디스크, 하드 디스크 등), 광학적 판독 매체(예를 들면, CD, DVD, USB 등)와 같은 저장매체를 포함한다. 또한 캐리어 웨이브(예를 들어 인터넷을 통한 전송)의 형태로 구현되는 것도 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

- [0059] 본 발명의 또 다른 양태에 따르면, 본 발명은 한국인 집단의 유전체 정보로부터 서브집단의 유전체 정보를 추출하는 단계; 및 상기 서브집단의 유전체 정보로부터 서브집단의 변이체 정보를 수집하는 단계;를 포함하는 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트의 구축 방법을 제공한다.
- [0060] 상기 방법은 한국인 집단의 단일 염기 변이, 삽입 결실, 복제수 변형, 전이인자 삽입, HLA 등의 유전적 특성이 반영된 집단의 유전체 정보; 및 피검 대상과 한국인(즉, 인종) 정보;를 공유하는 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트를 구축할 수 있다. 상기 방법으로 구축된 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트는 피검 대상이 한국인 집단에 속할 경우 변이체 정보를 보다 빠르고 정확하게 분석하여 정상 시료를 수득하기 위한 어려움 없이 질병 등에 대한 변이체 정보를 제공할 수 있다.
- [0062] 이하, 실시예를 통하여 본 발명을 더욱 상세히 설명하기로 한다. 이들 실시예는 단지 본 발명을 예시하기 위한 것이므로, 본 발명의 범위가 이들 실시예에 의해 제한되는 것으로 해석되지는 않는다.
- [0064] **실시예 1. 한국인의 샘플 데이터 수집을 통한 Korea1K 데이터 세트의 구축**
- [0065] 1-1. 샘플 모집 및 서열 분석
- [0066] 실험은 두 개의 하위 프로젝트로 구성된 한국 울산 유전체 프로젝트(Korean Ulsan genome project)에 참여한 모든 사람의 동의를 얻어 진행되었다. 79개의 형질(trait)과 1인당 160개의 건강 관련 문항(health-related question)을 포함하는 설문 조사를 기반으로 하는 모든 임상 정보는 울산 대학교 병원에서 검사하였다. 조사된 임상 정보는 서브집단 추출에 사용되었다. 상기 임상정보는 건강검진 및 설문조사 결과가 반영된 것으로, 구체적인 건강검진 및 설문조사 항목은 표 1 및 2에 각각 나타내었다.

표 1

분류	no.	항목	분류	no.	항목	분류	no.	항목
신체계측	1	신장	소변 검사	15	상피세포 (Transitional)	혈액검사	22	RH
신체계측	2	체중	소변 검사	16	상피세포 (Renal)	혈액검사	23	ABO
신체계측	3	표준체중	소변 검사	17	기타	혈액검사	24	총단백
신체계측	4	비만도	시력 검사	1	시력:나안(우)	혈액검사	25	알부민
신체계측	5	신체계측검사	시력 검사	2	시력:나안(좌)	혈액검사	26	A/G Ratio
신체계측	6	제지방율	시력 검사	3	시력:교정(우)	혈액검사	27	Globulin
신체계측	7	BMI	시력 검사	4	시력:교정(좌)	혈액검사	28	총빌리루빈
신체계측	8	허리둘레	시력 검사	5	시력:안압(우)	혈액검사	29	직접빌리루빈
신체계측	9	수축기혈압	시력 검사	6	시력:안압(좌)	혈액검사	30	AST(SGOT)
신체계측	10	이완기혈압	시력 검사	7	안저(우)	혈액검사	31	ALT(SGPT)
신체계측	11	맥박수	시력 검사	8	안저(좌)	혈액검사	32	ALP
신체계측	12	기타	시력 검사	9	안과	혈액검사	33	감마지티피
폐기능검사	1	노력성폐활량 (FVC)	시력 검사	10	기타	혈액검사	34	S항원(HBsAg)
폐기능검사	2	노력성폐활량 백분율 (%FVC)	부인과 검사	1	액상세포 도말 검사	혈액검사	35	S항체 (AntiHBs)
폐기능검사	3	1초간노력성폐활량(FEV1)	부인과 검사	2	기타	혈액검사	36	AntiHCV
폐기능검사	4	1초율(FEV1%)	심장 검사	1	심전도	혈액검사	37	AFP
폐기능검사	5	1초율노력성폐활량백분율 (FEV1%%)	심장 검사	2	기타	혈액검사	38	혈액요소질소
폐기능검사	6	폐기능	영상 검사	1	L-spine (X-ray)	혈액검사	39	크레아티닌(혈중)
폐기능검사	7	기타	영상 검사	2	복부초음파	혈액검사	40	Na
청력	1	1000Hz(좌)	영상 검사	3	기타	혈액검사	41	K
청력	2	2000Hz(좌)	기타 기록	1	수술 기록	혈액검사	42	총칼슘
청력	3	3000Hz(좌)	기타 기록	2	기타	혈액검사	43	무기인
청력	4	4000Hz(좌)	혈액검사	1	총콜레스테롤	혈액검사	44	T3
청력	5	1000Hz(우)	혈액검사	2	중성지방	혈액검사	45	Free T4
청력	6	2000Hz(우)	혈액검사	3	HDL-Cholesterol	혈액검사	46	TSH
청력	7	3000Hz(우)	혈액검사	4	LDL-Cholesterol	혈액검사	47	CRP(정량)
청력	8	4000Hz(우)	혈액검사	5	LP(a)	혈액검사	48	ESR
청력	9	기타	혈액검사	6	혈당(공복)	혈액검사	49	RA(정성)
소변 검사	1	요산	혈액검사	7	Hemoglobin A _{1c}	혈액검사	50	CEA
소변 검사	2	백혈구반응	혈액검사	8	백혈구	혈액검사	51	CA19-9
소변 검사	3	요질산염	혈액검사	9	적혈구	혈액검사	52	CA125
소변 검사	4	산도	혈액검사	10	헤모글로빈	혈액검사	53	PA(PSA)
소변 검사	5	요단백	혈액검사	11	헤마토크리트	혈액검사	54	Amylase
소변 검사	6	요비중	혈액검사	12	RDW	혈액검사	55	CPK
소변 검사	7	요당	혈액검사	13	MCV	혈액검사	56	LDH
소변 검사	8	케톤체	혈액검사	14	MCH	혈액검사	57	호모시스테인
소변 검사	9	유로빌리노겐	혈액검사	15	MCHC	혈액검사	58	헤리코박터균
소변 검사	10	오빌리루빈	혈액검사	16	혈소판수	혈액검사	59	기타
소변 검사	11	오장혈반응	혈액검사	17	Polyseg neut	내시경	1	위내시경
소변 검사	12	백혈구	혈액검사	18	Lymphocyte	내시경	2	대장내시경
소변 검사	13	적혈구	혈액검사	19	Monocyte	내시경	3	기타
소변 검사	14	상피세포 (Squamous)	혈액검사	20	Eosinophil	기타 검사	1	기타
소변 검사	15	기타	혈액검사	21	Basophil			

[0067]

표 2

대분류	소분류	대분류	소분류
참여자 기초 정보	생년월일	치아/턱 관절에 관한 문항	중지
	태어난 시간		양치질
	성별		지열
	성씨 및 본관		치아 보철물
	동거형태		턱관절장애
	직업		치과 치료 여부
	가족관계		스케일링 여부
신체적특성에 관한 문항	인종/민족	식습관에 관한 문항	식사 패턴
	혈액형		음식 종류
	키		흡연
	몸무게		음주
	왼손잡이 또는 오른손잡이		매운맛 잘 먹는 정도
	쌍둥이 여부		우유 소화력
	귀지 형태		대변 패턴
신체활동에 관한 문항	눈 색	수면에 관한 문항	소변 패턴
	땀흘림		기상/취침 시간
	심장 두근거림		낮잠 빈도
	길을 걸을 때 숨이 차는 정도		코골이
정신건강에 관한 문항	운동량	알러지에 관한 문항	불면증 척도
	우울척도		알러지
감각에 관한 문항	스트레스척도	질환에 관한 문항	알 질환
	자살 사고		대사, 영양, 면역 질환
	시각		혈액장애 질환
	청각		신경계 질환
한열/체질에 관한 문항	미각		시각, 청각 질환
	후각		순환계 질환
	촉각		호흡계 질환
	한의학 건강평가 설문		전염성 질환
성적 유형에 관한 문항	대범 vs 성세		소화계 질환
	빠른 행동 vs 느린 행동		비노생식기계 질환
	전극전 vs 소극전		피부질환
	외향적 vs 내향적		근골격계, 결합조직 질환
	다혈질 vs 이성적		선천적 특성 및 이상
	계획적 vs 즉흥적		복용약
	정리정돈 vs 어지름		항생제 여부
	장외적 vs 현실적	가족병력에 관한 항목	당뇨
	인간관계 중시 vs 일 관계 중시		고혈압
	낙천적 vs 비관적		B형 간염
머리카락에 관한 문항	머리카락 색		암
	직모 또는 곱슬		뇌심혈관
	탈모		류마티스(관절염) 질환
피부에 관한 문항	피부 타입	여성건강에 관한 문항	희귀질환
	피부색		기타등등
	햇빛 반응 정도		임신
	주근깨		월경
	여드름		
	탈		
	흉터		

[0068]

[0069]

울산 대학교 병원 바이오 뱅크에서 총 721개의 샘플이 선별되었으며, 이후 추가적으로 샘플이 접수되었다. 또한, 286개의 샘플을 별도로 수집하였다.

[0070]

또한, 본 연구진들은 한반도 전역에서 희망자를 모아 제작된 KoVariome 데이터의 87개의 한국인 샘플로부터 데이터를 수집하였다.

[0071]

시료 수집 및 시퀀싱은 울산과학기술원(UNIST)의 기관 검토위원회(IRB)에 의해 승인되었다.

[0072]

DNeasy 혈액 및 조직 키트(독일 Qiagen)를 사용하여 인간의 혈액 샘플로부터 유전체 DNA를 제조 프로토콜에 따라 분리하였고, 추출된 DNA를 Quant-iT BR 분석 키트(Invitrogen)를 사용하여 정량하였다. Covaris S2 초음파 처리 시스템을 사용하여 고분자량 유전체 DNA를 전단(shear)하여 적절한 크기의 단편을 획득하였다.

- [0073] 페어드 엔드 리드(paired-end reads)를 위한 짧은 350-bp 인서트(insert)를 갖는 라이브러리를 TruSeq Nano DNA 샘플 프렙 키트(TruSeq Nano DNA sample prep kit)를 사용하여 일루미나 기반 서열분석을 위한 제조사 프로토콜에 따라 준비하였다.
- [0074] 상기 라이브러리 생성물을 Bioanalyzer 2100(Agilent, Santa Clara, USA, USA)을 사용하여 정량하였고, Illumina Hi-Seq X10 플랫폼(Illumina)을 사용하여 기초 자료(raw data)를 생성하였다. 재시퀀싱을 통해 PE 2 × 150-bp 사이클 시퀀싱 관독을 사용하여 클러스터를 생성하였다. 기본 매개 변수(default parameters)가 있는 Illumina 파이프라인을 사용하여 추가 이미지 분석 및 베이스 호출(base calling)을 수행하였다.
- [0075] 수집된 한국인 집단의 유전체 정보는 단일 염기 변이(single nucleotide variants, SNV), 삽입 결실(indels), 복제수 변형(copy number variations, CNV), 전이인자(transposable element, TE) 삽입(insertion) 및 HLA(human leukocyte antigen) 등이 반영되어 있다.
- [0077] 1-2. 변이 콜링(Variant calling)
- [0078] 포워드 어댑터('GATCGGAAGAGCACACGTCTGAACTCCAGTCAC'), 리버스 어댑터('GATCGGAAGAGCGTCGTAGGGAAAGAGTGT'), 및 트리밍 후 최소 50bp 리드 길이를 갖는 어댑터 오염(Adapter contamination)은 Cutadapt(ver 1.9.1) 프로그램을 사용하여 트리밍하였다.
- [0079] 그런 다음, '-M' 옵션을 갖는 BWA-MEM(ver. 0.7.16a) 프로그램을 사용하여 트리밍된 리드(read)가 hg38 레퍼런스에 맵핑되었다. 맵핑된 BAM 파일을 Sortsam 모듈을 갖는 Picard(ver.2.14.0) 프로그램을 사용하여 조정(coordination)을 통해 분류하였다.
- [0080] 중복 리드(Duplicate read)는 Markduplicates 모듈을 갖는 Picard(ver.2.14.0) 프로그램을 사용하여 표시하였다. GATK(Genome Analysis Tool Kit)(ver. 3.7) 키트의 BaseRecalibrator 툴(tool)을 사용하여 맵핑 품질을 재보정하였다.
- [0081] SNV(single nucleotide variants)와 Indel 부위는 'stand_call_conf 30' 옵션을 갖는 GATK(Genome Analysis Tool Kit)(ver. 3.7) 키트의 'HaplotypeCaller' 기능으로 공동으로 콜링하였다.
- [0082] VEP(variants Effect Predictor, ver. 92)를 사용하여 콜링된 변이(SNV와 Indel)에 주석을 달았고, 문 등의 스크립트를 사용하여 네거티브 선발(negative selection) 하에서 프랙션(fraction)을 추정하였으며(<https://github.com/moon-s/fraction-under-selection>), 다음과 같은 주석이 달린 변이체를 기능 손실 돌연변이로 할당하였다: "Frame_Shift_Del", "Frame_Shift_Ins", "In_Frame_Del", "In_Frame_Ins", "Nonsense_Mutation", "Nonstop_Mutation" 또는 "Splice_Site".
- [0084] 1-3. 배치 효과 제거(Batch effect removal)
- [0085] 각 샘플은 시퀀싱 라이브러리 준비 프로토콜과, 시퀀싱 회사, 또는 혈액 샘플 또는 라이브러리를 회사에 보내는 날짜에 따라 레이블을 지정하였다.
- [0086] 12개의 테크니컬 배치(technical batches)가 확인되었고, 배치 효과는 하기 기준에 따라 변이체 및 샘플을 사용하여 EIGENSOFT(ver 6.1.4) 프로그램을 사용하여 주성분 분석(PCA, principal component analysis)을 통해 평가되었다:
- [0087] 1) 마이너 대립 유전자 빈도(MAF, minor allele frequency)를 갖는 이중 대립(Biallelic) SNV(single nucleotide variants) $\geq 5\%$,
- [0088] 2) 하디-바인베르그 평형(HWE, Hardy-Weinberg Equilibrium) 테스트의 P-밸류 값 > 0.05 ,
- [0089] 3) 유전자형 결손률(Genotype missing rate) < 0.01
- [0090] 이 후, 필터링된 변이체는 'indep-pairwise 200 4 0.1' 모듈을 갖는 PLINK(ver 1.9b) 프로그램을 사용하여 101,326개의 SNV(single nucleotide variants)만을 남기고 연관 불균형(LD, linkage disequilibrium)에 기초하여 정리하였다. 개체 선택(individual selection)을 위해 PLINK 프로그램에서 추정되는 IBD(ident-by-descent) 정보를 기반으로 식별된 밀접하게 연관된 개체를 필터링하였다. IBD(ident-by-descent) 값이 0.125를 초과하는 모든 쌍들을 추출하고 가족 그룹에 클러스터링하였다.
- [0091] 연관된 쌍들이 남아 있지 않을 때까지 각 패밀리 그룹은 다음의 방법으로 줄여나갔다:

- [0092] 1) 패밀리 그룹에서 가장 많은 수의 쌍을 가진 샘플을 제거, 및
- [0093] 2) 패밀리 그룹에 동일한 수의 쌍을 가진 여러 샘플이 있는 경우 연관 불균형이 제거된 SNV(single nucleotide variants) 중에서 가장 누락된 콜(highest missing calls)을 갖는 샘플을 제거.
- [0094] 패밀리 그룹 안에서 같은 수의 쌍을 갖는 여러 샘플이 존재하였다. 배치 효과를 나타내는 변이체(variants)를 식별하기 위해, 다음과 같이 모든 변이체에 대해 로지스틱 회귀 모델(logistic regression models)을 사용하였다:
- [0095] 1) 변이체가 다른 모든 배치와 비교하여 배치-특이적 변이체인 경우는 제거, 및 2) 각 배치는 다른 배치와 짝을 이루어 모든 가능한 조합을 만들었고, 임의의 조합에서 변이체가 유의한 경우, 변이체를 제거.
- [0096] 그 결과, 총 6,348,049개의 변이체 위치가 테크니컬 배치($P\text{-value} \leq 0.01$)와 유의하게 관련되어 있었으며, 오리지널 변이체 1차 정보에서 제거되었고 변이체 2차 정보로 가공되었다. 변이 품질 분포(variants quality distribution)를 플로팅하기 위해 조인트 VCF 파일에서 QD 값을 사용하였다.
- [0098] 1-4. 변이체의 분류
- [0099] 모든 샘플의 변이체(Variants)는 조인트 콜링 모드(joint calling mode)와 함께 GATK HaplotypeCaller 프로그램을 사용하여 콜링하였다. 식별된 모든 변이체는 Ensembl Variants Effect Predictor(VEP, ver. 92.1) 프로그램을 사용하여 주석을 달았다. 또한, 변이체는 CGC(Cancer Gene Census) 데이터베이스에서 암 관련 유전자에 주석을 달았다. 암이 있는 개인의 샘플에서 변이체가 배타적으로 확인된 경우 해당 변형을 체세포로 처리하였다.
- [0100] 일본인 집단의 유전적 특성이 반영된 대조군인 3.5KJPN 데이터베이스는 hg19 코디네이트(coordinates)를 가진 변이체에 관한 정보만을 제공하기 때문에, 변이체는 UCSC 유전체 브라우저에서 'liftover tool' 을 사용하여 hg38 코디네이트(coordinates)로 변환되었다.
- [0101] 이후 Korea1K 및 3.5KJPN 세트의 대립유전자 발생 빈도(allele frequencies)에 관한 정보를 유전체 위치 및 대립유전자 정보를 기반으로 하여 주석이 달린 변이체에 병합하였다. 주석이 달린 변이체는 각 패널에서 대립유전자 발생 빈도 컷오프(Korea1K, 3.5KJPN 및 1KGP)를 기준으로 분류되었다.
- [0102] 패널에서 겹쳐진 변이체는 생식세포 변이체(germline variants)로 분류되었고, 겹치지 않은 변이체는 체세포 변이체(somatic variants)로 분류되었다. 그 후, 분류된 세트(classification set)와 실제 세트(true set)(즉, 피검자의 데이터 세트)를 비교하였다.
- [0104] 1-5. 데이터 가용성(Data availability)
- [0105] 울산 국립 과학기술원 KOGIC 윤리위원회의 승인 후, 요청에 따라 모든 유전체 변이체 데이터를 이용할 수 있다.
- [0107] **실시예 2. 피검 대상의 데이터 세트 준비**
- [0108] 2-1. 위암을 앓는 피검 대상의 데이터 세트 분비
- [0109] 본 연구 이전에 보고된 19명의 위암을 앓고 있는 한국인 개체의 전체 유전체 서열 데이터는 NCBI(SRP014574 및 SRA057772)에서 입수하여 "-M" 옵션과 함께 BWA-mem(ver 0.7.15) 프로그램을 사용하여 hg38에 맵핑하였고 SAMtools(ver 1.4)를 사용하여 SAM 포맷을 BAM 포맷으로 변환하였다.
- [0110] BAM 파일은 SAMtools(버전 1.4) 프로그램을 사용하여 정렬되었고, 복제된 리드(reads)는 Picard tool의 'Markduplicates' 모듈을 사용하여 표시하였다.
- [0111] GATK(3.7) 프로그램을 사용하여 베이스 품질 점수(base quality score)의 베이스(base) 재정렬(quality score) 및 재보정(recalibration)을 수행하였다.
- [0113] 2-2. 데이터 가용성(Data availability)
- [0114] 울산 국립 과학기술원 KOGIC 윤리위원회의 승인 후, 요청에 따라 모든 유전체 변이체 데이터를 이용할 수 있다.
- [0116] **실시예 3. 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트를 이용한 암 유전체학 연구**
- [0117] 베리움(variomes) 데이터베이스들은 유전체 와이드 임상 연구(genome-wide clinical studies)를 위한 레퍼런스 패널(reference panels)로 사용될 수 있다. 실제 사례 중 하나는 암 연구이다. 많은 암 관련 연구는 전체 콜 세트(entire call set)에서 생식세포 변이체(germline variants)를 필터링하기 위하여 정상 대조군 샘플(normal

control samples)과 대조할 필요가 있다. 그러나, 암 환자와 일치하는 정상 시료 모두에 대하여 시퀀싱 데이터를 분석하고 보유하는 것은 매우 번거로운 일이며, 실제로 불가능한 경우도 있다. 이와 같은 한계를 해결하기 위하여, 상기 실시예 1에서 한국인의 샘플 데이터를 수집하여 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트(후술되는 실험 및 도면에서는 ‘Korea1K’라 기재함)를 구축하였다. 종래 기술(정상 조직과 비교) 및 레퍼런스 세트(본 실시예)를 이용한 암 유전체학 연구의 개념도는 도 1에 나타내었다.

[0119] 국가 또는 민족의 특성을 나타내는 레퍼런스 패널로써, 상기 실시예 1에서 구축된 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트의 능력을 평가(즉, 변이체 분석)하였다. 구체적으로, 상기 평가에 이용한 레퍼런스 패널 및 피검자 데이터 및 피검자 데이터 세트는 표 3과 같다.

표 3

	데이터 세트	인종	비고
레퍼런스 패널	Korea1K	Korean	실시예 1
	3.5KJPN	Japanese	
	EAS	East Asian	
	SAS	South Asian	
	AMR	American	
	EUR	European	
	AFR	African	
	1KGP	Different ethnic group	1000 Genomes Project
시험군	위암 변이체 데이터세트	Korean	실시예 2

[0122] 먼저 암 조직과 매칭되는 레퍼런스 데이터 세트(즉, 정상인 패널)에서 WGS(Whole-genome sequencing) 데이터를 비교하여 체세포 및 생식세포 변이체를 동정하였다. 그 후, 피검자 데이터 세트의 변이체는 레퍼런스 패널의 대립유전자 발생 빈도 컷오프(allele frequency cut-offs)에 기초하여, 체세포 또는 생식세포의 변이체로 분류하였다. 상기 분류는 표적 변이체(target variant)가 레퍼런스 패널에서 컷오프 값보다 낮은 대립유전자 발생 빈도 값을 갖는 경우 체세포 변이체로 분류하였고, 표적 변이체가 더 높은 대립유전자 발생 빈도 값을 갖는 경우 잠정적으로 생식세포 변이체로 분류하였다.

[0123] 그 후 여러 단계의 대립 유전자 빈도 컷오프(multiple step-wise allele frequency cut-offs)를 기반으로 예측된 변형체 카테고리(variant categories)를 실제 변형체 카테고리들과 비교하여 분류 퍼포먼스(classification performance)의 통계적 측정치(statistical measures)를 생성하였다. 통계적 측정치에 기초하여 분류의 정확성(ACC), 매튜 상관 관계 계수(Matthews correlation coefficient value, MCC) 및 생식계열 회복율(Germline recovery rate)을 도출하였고, 그 결과는 도 2a 내지 2c에 각각 나타내었다.

[0124] 도 2a에 나타난 바와 같이, 일본인의 변이체를 기반으로 한 3.5KJPN 데이터 세트가 모집단의 크기로 인해 가장 많은 변이체를 포함하고 있지만, 실시예 1의 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트는 생식세포 및 체세포 변이체의 예측 정확도가 가장 높았다(대립유전자 발생 빈도 컷오프 0.01 Korea1K: 96.42%; 대립유전자 발생 빈도 컷오프 값 0.01 3.5KJPN: 평균 93.83%).

[0125] 도 2b에 나타난 바와 같이, 실시예 1의 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트는 유사한 매튜 상관 계수가 일본의 3.5KJPN 데이터 세트와 유사하여 유사한 분류 퍼포먼스(classification performance)를 나타내는 것을 확인하였다(0.01:0.38의 대립유전자 발생 빈도 컷오프를 갖는 Korea1K; 평균 0.01:0.37의 대립 유전자 빈도 컷오프를 가진 3.5KJPN).

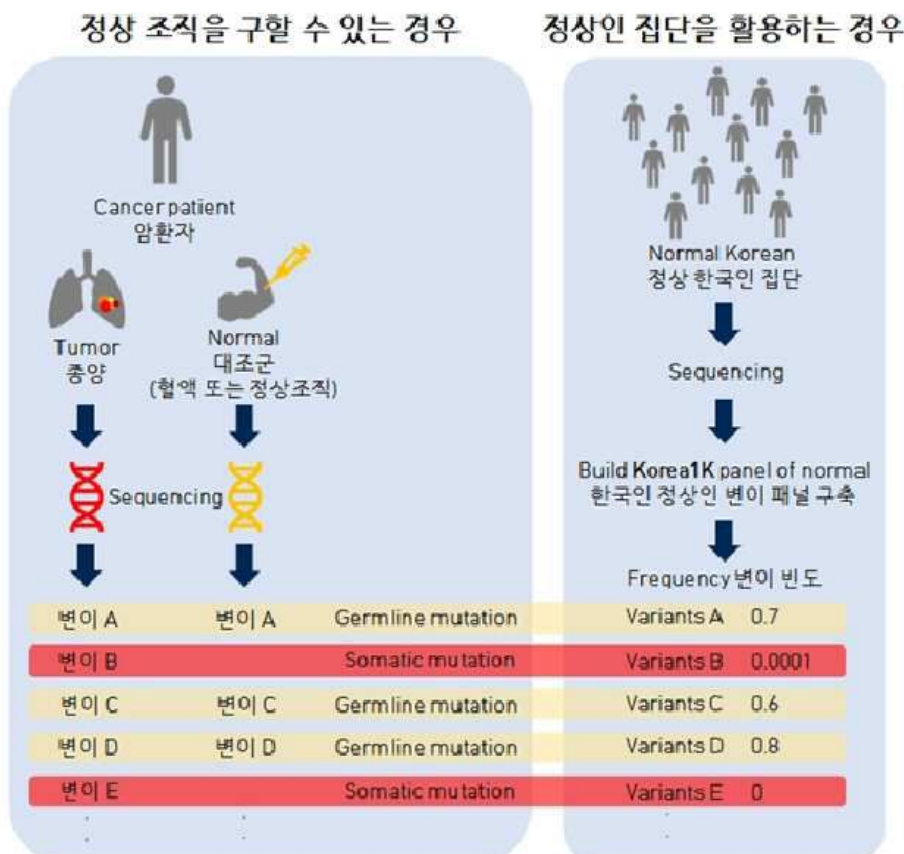
[0126] 도 2c에 나타난 바와 같이, 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트를 이용하여 분석하였을 때 생식세포 계열의 변이체의 예측 정확도가 가장 높았다.

[0127] 상기 결과는 본 실험에서 한국인의 위암 변이체 데이터 세트를 사용했기 때문에 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트와의 유사성이 전반적으로 증가한 것으로 나타났으며, 분석하고자 하는 대상 개체와 정상인 레퍼런스 패널의 모집단이 모두 한국인에 기초하기 때문에 분석에 대한 유사성 경향은 증가할 수 있음을 의미한다. 또한 상기의 결과를 통해 다른 인구집단의 변이체 데이터 세트를 분석이 이용하는 것과 비교하여, 한국인 인구집단의 변이체 정보에 기반한 실시예 1에서 구축된 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트를 사용함으로써 대부분의 체세포 변이체가 필터링될 수 있음을 확인하였다.

- [0129] 상기 실시예 1에서 구축된 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트를 활용하여 암과 관련된 변이를 검출할 때 가장 효과적으로 필터링을 할 수 있는 대립유전자 발생 빈도 컷오프(allele frequency cutoff) 값을 측정하였으며, 그 결과는 도 3에 나타내었다.
- [0130] 도 3에 나타낸 바와 같이, CGC(Cancer Gene Census) 데이터베이스의 암 관련 유전자 변이체의 밀도는 Korea1K 데이터에서 필터링된 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트(실시예 1)에서 가장 높은 것을 확인하였다. 또한, 생식세포 필터링 기준(germline filtering criterion)을 대립유전자 발생 빈도가 1.5%로 설정하였을 때, 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트는 CGC 유전자의 밀도가 가장 높았다. 또한 외국인 정상인 베리움 기반의 데이터 세트로 필터링을 했을 때보다 높은 CGC 비율 값을 갖는 범위 내의 변이들을 이용하면, 한국인의 암 관련 유전자의 변이를 더 잘 검출할 수 있는 것을 확인하였다.
- [0131] 상기 결과는 실시예 1에서 구축된 한국인 맞춤형 변이체 기반 레퍼런스 데이터 세트를 이용하는 것이 다른 인종의 정상인 변이체를 사용하는 것과 비해 더 효과적이고, 상대적으로 더 적은 수의 한국인 정상인 변이체를 사용해도 더 좋은 필터링 결과를 취득할 수 있다는 것을 의미한다.
- [0133] 이상, 본 발명내용의 특정한 부분을 상세히 기술하였는바, 당업계의 통상의 지식을 가진 자에게 있어서, 이러한 구체적인 기술은 단지 바람직한 실시양태일 뿐이며, 이에 의해 본 발명의 범위가 제한되는 것이 아닌 점은 명백할 것이다. 따라서 본 발명의 실질적인 범위는 첨부된 청구항들과 그것들의 등가물에 의해 정의된다고 할 것이다.

도면

도면1



도면3

