



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2016-0106796
(43) 공개일자 2016년09월13일

(51) 국제특허분류(Int. Cl.)

G06F 19/20 (2011.01)

(52) CPC특허분류

G06F 19/20 (2013.01)

(21) 출원번호 10-2015-0028944

(22) 출원일자 2015년03월02일

심사청구일자 2015년03월02일

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

이재원

서울특별시 강남구 압구정로36길 26, 14동 101호 (신사동, 현대맨션)

정수진

서울특별시 은평구 수색로12가길 9 (수색동)

(뒷면에 계속)

(74) 대리인

김동용, 김홍석

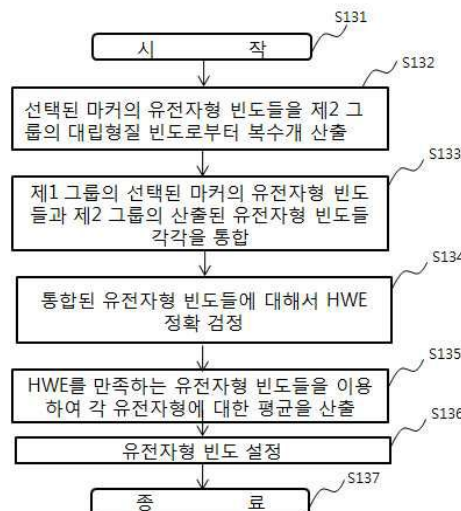
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 대립형질 빈도 추정 방법 및 장치

(57) 요약

본 발명은 (a) 제2 그룹의 대립형질 데이터에 기초하여 HWE(Hardy-Weinberg Equilibrium)를 만족하고 마커에서 나타날 수 있는 유전자형들에 대한 유전자형 빈도들을 복수개 산출하는 단계, (b) 제1 그룹의 유전자형 빈도와 제2 그룹으로부터 산출된 복수개의 유전자형 빈도들 각각을 통합하는 단계 및 (c) 통합된 유전자형 빈도들에 대해서 HWE 정확 검정을 실시하여 각 유전자형 빈도들이 HWE를 만족하는 지를 확인하는 단계를 포함하는 대립형질 빈도 추정 방법 및 장치에 관한 것이다.

대표도 - 도3



(72) 발명자

이효정

서울특별시 노원구 마들로 31 106동 2404호 (월계
동,그랑빌아파트)

이혜승

서울특별시 중구 다산로 32 5동 501호 (신당동,남
산타운아파트)

명세서

청구범위

청구항 1

- (a) 제2 그룹의 대립형질 데이터에 기초하여 유전자의 마커에서 나타날 수 있는 유전자형들에 대한 유전자형 빈도들을 복수개 산출하는 단계;
- (b) 제1 그룹의 유전자형 빈도와 상기 제2 그룹으로부터 산출된 복수개의 유전자형 빈도들 각각을 통합하는 단계; 및
- (c) 통합된 유전자형 빈도들에 대해서 HWE(Hardy-Weinberg Equilibrium) 정확 검정을 실시하여 각 유전자형 빈도들이 HWE를 만족하는 지를 확인하는 단계;를 포함하는,
- 대립형질 빈도 추정 방법.

청구항 2

- 제1항에 있어서,
- (d) 통합된 유전자형 빈도들 중 상기 단계 (c)에 의해 HWE를 만족하는 유전자형 빈도들을 이용하여 각 유전자형에 대한 평균을 산출하는 단계; 및
- (e) 산출된 평균을 각 유전자형에 대한 유전자형 빈도로 설정하는 단계;를 더 포함하는,
- 대립형질 빈도 추정 방법.

청구항 3

- 제2항에 있어서,
- (g) 평균을 통해 마커의 설정된 유전자형 빈도들로부터 제1 그룹 및 제2 그룹을 포함하는 분석대상 집단의 대립형질 빈도를 추정하는 단계;를 더 포함하는,
- 대립형질 빈도 추정 방법.

청구항 4

- 제1항에 있어서,
- 상기 단계 (a)는 제2 그룹의 표본수에 기초해서 상기 마커에서 대립형질 빈도로부터 기대되는 제2 그룹의 유전자형 빈도를 산출하는 다항분포를 통해 유전자형 빈도들 각각을 복수개 산출하는,
- 대립형질 빈도 추정 방법.

청구항 5

- 제1항에 있어서,
- 상기 단계 (a) 내지 (c)는 모든 마커 각각에 대해서 수행되는,
- 대립형질 빈도 추정 방법.

청구항 6

- 제3항에 있어서,
- 마커의 추정된 상기 대립형질 빈도는 개인식별검사, 친자확인검사 또는 형제검사에서 이용되는,
- 대립형질 빈도 추정 방법.

청구항 7

제1 그룹의 유전자형 빈도를 각 마커별로 저장하고 제2 그룹의 대립형질 데이터를 각 마커별로 저장하는 저장부; 및

상기 저장부에 저장된 데이터를 이용하여 유전자형 빈도를 산출하는 제어부;를 포함하며,

상기 제어부는, 유전자형 빈도의 산출을 위해, 제2 그룹의 대립형질 데이터에 기초하여 유전자의 마커에서 나타날 수 있는 유전자형들에 대한 유전자형 빈도들을 복수개 산출하고, 제1 그룹의 유전자형 빈도와 상기 제2 그룹으로부터 산출된 복수개의 유전자형 빈도들 각각을 통합하고, 통합된 유전자형 빈도들에 대해서 HWE 정확 검정을 실시하여 각 유전자형 빈도들이 HWE를 만족하는 지를 확인하는,

대립형질 빈도 추정 장치.

청구항 8

제7항에 있어서,

상기 제어부는, 통합된 유전자형 빈도들 중 HWE를 만족하는 유전자형 빈도들을 이용하여 각 유전자형에 대한 평균을 산출하고, 산출된 평균을 각 유전자형에 대한 유전자형 빈도로 설정하는,

대립형질 빈도 추정 장치.

청구항 9

제8항에 있어서,

상기 제어부는, 평균을 통해 마커의 설정된 유전자형 빈도들로부터 제1 그룹 및 제2 그룹을 포함하는 분석대상 집단의 대립형질 빈도를 추정하고 추정된 대립형질 빈도를 상기 저장부에 저장하는,

대립형질 빈도 추정 장치.

청구항 10

제7항에 있어서,

상기 제어부는, 제2 그룹의 표본수에 기초해서 상기 마커에서 대립형질 빈도로부터 기대되는 제2 그룹의 유전자형 빈도를 산출하는 다항분포를 통해서 상기 유전자형 빈도들 각각을 복수개 산출하는,

대립형질 빈도 추정 장치.

발명의 설명

기술 분야

[0001] 본 발명은 대립형질 빈도 추정 방법 및 장치에 관한 것으로서, 구체적으로는 여러 연구소에서 가지고 있거나 공개된 개별적인 유전자 자료로부터 통합적인 대립형질 빈도를 생성하여 각종 유전자 검사에서 신뢰성 있는 검사 결과를 제공할 수 있도록 하는, 대립형질 빈도 추정 방법 및 장치에 관한 것이다.

배경 기술

[0002] 유전자(DNA)를 이용한 개인 식별은 지난 50년 동안 형사 소송이나 친자확인소송 등의 법적 추론에서 널리 이용되고 있다. 특히 1985년 이후부터는 유전자를 직접 검사하는 유전자 검사가 보편화 되었다. 유전자 검사는 유전자 지문(DNA fingerprint)이라고 불릴 만큼 개인 식별의 유용성을 널리 인식시켰다.

[0003] 유전자 검사를 통한 개인 식별은 실험적인 측면 외에 통계학적 문제들과 밀접히 연관되어 있고 통계적인 분석과정을 통해서 적용가능하다. 유전자 증거를 이용하여 개인 식별을 하기 위해서는 먼저 여러 유전자들 중에서 분석 대상인 유전자를 선택하여 검사 체계를 구성하여야 한다. 또한 분석대상 집단의 표본으로부터 유전자 검사 체계에서 조사된 유전자형을 이용하여 결정된 대립형질 빈도를 바탕으로 유전자 증거를 평가하고 판단할 수 있어야 한다.

[0004] 유전자 검사는 예를 들어 개인식별검사, 친자확인검사, 형제검사 등을 수행한다. 이러한 유전자 검사는 기본적인

으로 특정 분석대상 집단의 대립형질 빈도를 이용하여 그 검사가 이루어진다. 따라서 이용된 표본의 대립형질 빈도의 정확도나 대표성에 따라 유전자 검사의 정확도 또는 신뢰도는 달라진다.

- [0005] 신뢰성 있는 개인 식별이 가능하기 위해서는 분석대상 집단의 유전적 특성이 선행하여 조사되어야 하고 분석 목적에 따른 신뢰성 있는 검사 체계의 구성이 필수적이다. 이를 기준으로 한 유전자 증거의 평가가 이루어져야만 신뢰성있는 평가가 가능하다.
- [0006] 유전자형과 같은 과학적 증거에 대한 평가는 통계적 가설검정의 개념을 일반적으로 활용한다. 이는 비교하고자 하는 사람(예를 들어 용의자나 친부라고 지목된 사람)의 유전자 표본과 유전자 증거(범죄현장의 흔적 또는 부성인자)의 표본이 일치한다고 할 때, 두 표본이 실제로 같은 사람의 것이기 때문에 일치했을 확률이 다른 사람의 것인데 우연히 일치했을 확률에 비해 얼마나 큰가를 기준으로 이루어진다.
- [0007] 또한, 유전자 증거의 평가에서 가장 유의해야 할 점은 유전자 증거를 남긴 사람과 비교된 사람의 유전적 관련성의 문제이다. 기존의 유전자 평가는 무한 크기(infinite size)와 임의교배(random mating)를 가정한 이상집단(ideal population)에서 유전적 관련성을 무시하고 두 표본은 무관한 것임을 가정한 방법에 의존해 왔다. 그러나 유전자 증거의 평가를 위해서 두 표본 간의 친척관계를 설정하는 경우와 집단의 유전적 관련성을 가정하는 경우를 중심으로 한 연구가 활발히 진행되고 있고 두 표본간의 관계를 어떻게 가정하느냐에 따라 다르게 평가될 수 있다. 유전적 관련성은 "무관함으로 가정하는 경우"(unrelated), "집단의 유전적 관련성을 가정하는 경우"(subpopulation) 및 "가까운 친척관계를 가정하는 경우"(relatives)로 나누어서 평가될 수 있다.
- [0008] 이처럼 유전자 증거를 이용한 개인 식별은 유전자 증거 자료와 비교 대상자가 같은 유전자형을 가지고 있다고 할 때, "이 유전자형은 기준 집단에서 볼 때 드물게 나타나는 유전자형이므로 두 표본이 서로 다른 사람의 것일 확률은 매우 작을 것이다"의 논리로 진행된다.
- [0009] 유전자 증거를 개인 식별 등에 이용하고자 할 때에는 집단의 유전적 특성 파악을 위한 각 집단별로 나타나는 유전자형과 대립형질의 빈도에 대한 연구가 선행되어져 있어야 한다. 그러나 유전자형과 대립형질의 빈도 획득은 용이치 않다. 단지 특정 한 연구소 내에서 연구된 유전자 자료만을 이용하여 대립형질 빈도가 획득되거나 알려져 있는 특정 틀에서 제공되는 빈도가 이용된다. 이러한 틀의 유전자 자료는 통상 특정 집단(예를 들어 미국인 등)의 유전자에 최적화되어 있어 다른 집단(예를 들어 한국인)에 적용하는 것이 용이치 않다. 또한 개별 연구소의 유전자 자료는 표본수가 적어서 특정 집단을 대표할 수가 없다.
- [0010] 여러 연구소나 틀에서 공개되는 유전자 자료를 통합하여 특정 분석대상 집단의 대표적인 유전자형 빈도, 나아가 대립형질 빈도를 생성할 수 있다면 산재된 유전자 자료로부터 대표성을 가지고 신뢰성이 높은 빈도로부터 유전자 검사가 가능할 것이다. 이는 유전자 검사의 신뢰성을 높일 수 있고 정확도에 영향을 미치게된다. 따라서 특정 분석대상 집단의 많은 표본을 이용하여 결정되는 대립형질 빈도의 이용이 매우 바람직하다.
- [0011] 그러나 여러 유전자 자료로부터 이 여러 유전자 자료에 대한 대표성을 인정할 수 있는 통합 빈도를 생성하는 방안에 대해서는 알려져 있지 않다. 더욱이 여러 연구소에서 공개하는 유전자 자료는 대립형질 빈도에 관련된 자료만이다.
- [0012] 여러 유전자 자료를 통합하기 위해서는 먼저 유전자 자료로부터 유전자형 빈도를 계산하여야 하고 이로부터 통합된 유전자형 빈도를 계산하여야 한다. 그러나 일반적으로 공개되는 대립형질 빈도 자체만으로는 통합적인 또는 대표성을 가지는 유전자형 빈도를 계산할 수가 없다. 나아가 다른 유형의 유전자 자료를 통합하는 것이 용이치 않다.
- [0013] 이와 같이, 여러 유전자 자료를 통합하기 위해서는 선행하여 대표성을 가지는 유전자형 빈도로의 통합이 필요하다. 이후 유전자형 빈도로부터 용이하게 대립형질 빈도를 생성할 수 있고 신뢰성을 높인 생성된 대립형질 빈도를 여러 유전자 검사에 활용 가능하다.
- [0014] 따라서, 여러 공개된 유전자 자료를 통합하여 다수의 표본으로부터 대표성을 가질 수 있는 대립형질 빈도 추정 방법 및 장치가 필요하다.

발명의 내용

해결하려는 과제

- [0015] 본 발명은 상술한 문제점을 해결하기 위해서 안출한 것으로서, 유전자 검사시 특정 집단에서의 검사 신뢰도를 높일 수 있도록 여러 유전자 자료로부터 특정 집단의 통합된 유전자형 빈도를 생성하고 이후 유전자 검사에서

이용되는 통합된 대립형질 빈도를 생성할 수 있는 대립형질 빈도 추정 방법 및 장치를 제공하는 데 그 목적이 있다.

과제의 해결 수단

[0016] 상기와 같은 목적을 달성하기 위한 대립형질 빈도 추정 방법은 (a) 제2 그룹의 대립형질 데이터에 기초하여 마커에서 나타날 수 있는 유전자형들에 대한 유전자형 빈도들을 복수개 산출하는 단계, (b) 제1 그룹의 유전자형 빈도와 제2 그룹으로부터 산출된 복수개의 유전자형 빈도들 각각을 통합하는 단계 및 (c) 통합된 유전자형 빈도들에 대해서 HWE(Hardy-Weinberg Equilibrium) 정확 검정을 실시하여 각 유전자형 빈도들이 HWE를 만족하는 지를 확인하는 단계를 포함한다.

[0017] 또한 상기와 같은 목적을 달성하기 위한 대립형질 빈도 추정 장치는 제1 그룹의 유전자형 빈도를 각 마커별로 저장하고 제2 그룹의 대립형질 데이터를 각 마커별로 저장하는 저장부 및 저장부에 저장된 데이터를 이용하여 유전자형 빈도를 산출하는 제어부를 포함하며, 제어부는, 유전자형 빈도의 산출을 위해, 제2 그룹의 대립형질 데이터에 기초하여 마커에서 나타날 수 있는 유전자형들에 대한 유전자형 빈도들을 복수개 산출하고, 제1 그룹의 유전자형 빈도와 제2 그룹으로부터 산출된 복수개의 유전자형 빈도들 각각을 통합하고, 통합된 유전자형 빈도들에 대해서 HWE 정확 검정을 실시하여 각 유전자형 빈도들이 HWE를 만족하는 지를 확인한다.

발명의 효과

[0018] 상기와 같은 본 발명에 따른 대립형질 빈도 추정 방법 및 장치는 유전자 검사시 특정 집단에서의 검사 신뢰도를 높일 수 있도록 여러 유전자 자료로부터 특정 집단의 통합된 유전자형 빈도를 생성하고 이후 유전자 검사에서 이용되는 통합된 대립형질 빈도를 생성할 수 있는 효과가 있다.

도면의 간단한 설명

[0019] 도 1은 대립형질 빈도 추정 장치의 예시적인 블록도를 도시한 도면이다.

도 2는 대립형질 빈도 추정을 위한 예시적인 흐름을 도시한 도면이다.

도 3은 도 2의 단계 S130의 구체적 흐름을 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0020] 상술한 목적, 특징 및 장점은 첨부된 도면을 참조하여 상세하게 후술 되어 있는 상세한 설명을 통하여 더욱 명확해 질 것이며, 그에 따라 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 본 발명의 기술적 사상을 용이하게 실시할 수 있을 것이다. 또한, 본 발명을 설명함에 있어서 본 발명과 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에 그 상세한 설명을 생략하기로 한다. 이하, 첨부된 도면을 참조하여 본 발명에 따른 바람직한 실시 예를 상세히 설명하기로 한다.

[0021] 도 1은 대립형질 빈도 추정 장치(100)의 예시적인 블록도를 도시한 도면이다.

[0022] 도 1에 따르면 대립형질 빈도 추정 장치(100)는 입력부(110), 출력부(130), 저장부(150), 연결부(170) 및 제어부(190)를 포함한다. 도 1의 대립형질 빈도 추정 장치(100)는 개인이 이용하는 개인용 컴퓨터, 노트북 등이거나 허가된 개인이 인터넷망을 통해서 접속가능한 서버 등일 수 있다. 대립형질 빈도 추정 장치(100)가 개인 전용인지 공용인지의 유형에 따라 대립형질 빈도 추정 장치(100)를 구성하는 블록, 각 구성 블록들의 하드웨어 구조나 기능은 달라질 수 있다.

[0023] 대표적인 대립형질 빈도 추정 장치(100)의 유형을 고려하여 각 블록들을 살펴보면, 입력부(110)는 본 발명에 따른 대립형질 빈도 추정 장치(100)에서 필요한 각종 데이터와 제어 데이터를 수신한다. 입력부(110)는 특정 분석대상 집단에서의 특정 표본(그룹)에 대하여 실험하여 획득한 유전자 자료를 수신하고 동일한 분석대상 집단에서의 다른 특정 표본(그룹)에 대하여 실험하여 획득한 유전자 자료를 수신 가능하다. 입력부(110)는 예를 들어 키보드, 마우스, 마이크, 휴대형 메모리 인터페이스, 유선랜이나 무선랜 등과 인터페이스를 하기 위한 네트워크 인터페이스 등 중 하나 이상을 포함한다. 입력부(110)를 통해서 대립형질 빈도 추정 장치(100)는 각종 데이터를 수집하고 수집된 데이터로 통합된 유전자형 빈도나 대립형질 빈도를 산출한다.

[0024] 특정 분석대상 집단의 특정 표본의 유전자 자료들의 수신과 통합된 유전자형 빈도나 대립형질 빈도의 생성을 수

행하기 위한 제어 데이터의 수신은 다른 시점에 이루어질 수 있다. 따라서, 입력부(110)를 통해 여러 시점에 유전자 자료들을 수집하고 이후 입력부(110)를 통한 사용자 제어 입력으로 특정 표본들을 통합 가능하다.

[0025] 출력부(130)는 대립형질 빈도 추정 장치(100)에 의해서 생성된 데이터를 출력한다. 출력부(130)는 디스플레이, 스피커, 휴대형 메모리 인터페이스, 네트워크 인터페이스 등 중 하나 이상을 포함한다. 출력부(130)는 대립형질 빈도 추정 장치(100)의 유형에 따라 입력부(110)의 구성과 동일할 수도 있다. 예를 들어 출력부(130)는 입력부(110)와 동일한 네트워크 인터페이스로 구성되거나 휴대형 메모리 인터페이스로 구성될 수도 있다.

[0026] 출력부(130)는 예를 들어 여러 그룹(표본) 간에 통합된 유전자형 빈도 및/또는 대립형질 빈도를 출력한다. 출력된 유전자형 빈도 및/또는 대립형질 빈도는 다른 프로그램 툴에서 이용 가능하다. 예를 들어 출력된 유전자형 빈도 및/또는 대립형질 빈도는 개인식별검사, 친자확인검사, 형제검사와 같은 유전자 검사를 위한 툴에서 이용 가능하다. 이 툴은 본 발명에 따라 제공된 유전자형 빈도 및/또는 대립형질 빈도에 기초해서 특정 유전자 검사를 수행하고 검사 결과를 출력한다.

[0027] 저장부(150)는 각종 데이터를 저장한다. 저장부(150)는 휘발성 메모리, 비휘발성 메모리 및/또는 하드 디스크 등과 같은 대용량 저장 매체를 포함한다. 저장부(150)는 적어도 분석대상 집단에서의 특정 그룹(표본)을 대상으로 조사된 유전자 자료 데이터를 포함하고 저장하고 있다.

[0028] 예를 들어 한 그룹(이하 '제1 그룹'이라고 지칭함)의 유전자 자료는 대립형질 빈도 추정 장치(100)를 운용하는 주체(예를 들어 연구소)에서 실험 및 연구 등을 통해서 생성된 자료일 수 있다. 제1 그룹의 유전자 자료는 제1 그룹을 표본으로 하여 수집된 자료로서 대립형질(Allele) 빈도 데이터와 유전자형(Genotype) 빈도 데이터를 포함한다. 여기서의 빈도 데이터는 제1 그룹의 표본내에서 해당 대립형질이나 유전자형을 가진 개체의 숫자나 확률을 나타내거나 이 숫자나 확률을 도출할 수 있는 데이터를 나타낸다.

[0029] 다른 한 그룹(이하 '제2 그룹'이라고 지칭함)의 유전자 자료는 동일한 주체가 다른 시기에 또는 다른 주체에 의해서 생성된 자료일 수 있다. 제2 그룹의 유전자 자료는 제2 그룹을 표본으로 하여 수집된 자료로서 일반적으로 대립형질 빈도 데이터를 포함한다. 제2 그룹을 대상으로 하는 유전자 자료는 일반적으로 다른 주체에 의해서 공개된 자료로서 통상 대립형질 빈도 데이터를 포함하나 유전자형 빈도 데이터는 포함되지 않는다.

[0030] 따라서, 두 그룹을 통합하여 두 표본을 대표하는 대립형질 빈도를 산출하기 위해서는 먼저 후자의 제2 그룹의 대립형질 빈도로부터 유전자형 빈도의 산출 과정과 이후 제1 그룹의 유전자형 빈도와와의 통합 과정이 필요하다. 이러한 산출 및 통합 과정은 도 2 및 도 3을 통해서 상세히 살펴보도록 한다.

[0031] 여기서, 제1 그룹의 유전자 자료는 특정 마커 그룹(이하 '제1 마커 그룹'이라고도 함)에 대한 유전자형 빈도 데이터를 포함하고 제2 그룹의 유전자 자료는 다른 마커 그룹(이하 '제2 마커 그룹'이라고도 함)에 대한 대립형질 빈도 데이터를 포함할 수 있다. 따라서, 통합된 유전자형 빈도의 산출 과정에서 다른 마커 그룹들을 고려한 통합이 필요하다.

[0032] 이와 같이, 저장부(150)는 적어도 제1 마커 그룹의 각 마커(Marker)에 대한 제1 그룹의 유전자형 빈도를 저장하고 있고 제2 마커 그룹의 각 마커에 대한 제2 그룹의 대립형질 데이터(빈도 데이터)를 저장하고 있다.

[0033] 연결부(170)는 도 1의 대립형질 빈도 추정 장치(100)의 각 블록들 사이의 데이터의 전달한다. 연결부(170)는 대립형질 빈도 추정 장치(100)의 유형에 따라서 달라질 수 있다. 예를 들어 연결부(170)는 병렬 버스, 시리얼 버스, GPIO(General Purpose Input Output) 등으로 구성될 수 있다. 또는 연결부(170)는 유선 랜 칩이나 무선 랜 칩을 통한 근거리 유선/무선 네트워크로 구성될 수도 있다.

[0034] 제어부(190)는 대립형질 빈도 추정 장치(100)의 각 블록들을 제어한다. 제어부(190)는 프로그램의 명령어 코드를 수행할 수 있는 하나 이상의 실행 유닛을 포함한다. 실행 유닛은 MPU(Main Processing Unit), CPU(Central Processing Unit) 등으로 지칭될 수 있다.

[0035] 제어부(190)는 저장부(150)에 저장된 유전자형 빈도 및 대립형질 빈도를 생성하는 생성 프로그램을 제어부(190) 내의 메모리에 로딩하고 이 생성 프로그램의 명령어 코드를 수행하여 유전자형 빈도 및/또는 대립형질 빈도를 생성한다. 통합되는 빈도 데이터의 생성 과정에서 제어부(190)는 저장부(150)에 저장된 제1 및 제2 그룹의 유전자 자료 데이터를 이용한다.

[0036] 유전자형 빈도의 산출을 위해 제어부(190)에서 수행되는 구체적인 제어 흐름은 도 2 및 도 3을 통해서 상세히 살펴보도록 한다.

- [0037] 도 2는 대립형질 빈도 추정을 위한 예시적인 흐름을 도시한 도면이다.
- [0038] 도 2의 흐름은 대립형질 빈도 추정 장치(100)에서 수행되고 제어부(190)에 의한 제어로 수행된다. 바람직하게는 도 2의 흐름은 저장부(150)에 저장된 프로그램을 제어부(190)가 로딩하여 프로그램의 명령어 코드를 제어부(190)가 수행함으로써 진행된다. 도 2의 대립형질 빈도 추정을 위한 흐름은 유전자형 빈도의 생성 후의 대립형질 빈도 추정 과정을 포함한다.
- [0039] 먼저 제어부(190)는 사용자의 입력부(110)를 통한 입력으로 저장부(150)에 저장된 생성 프로그램을 로딩하고 생성 프로그램을 수행(S100)한다.
- [0040] 제어부(190)는 입력부(110)를 통한 입력으로 또는 자동으로 통합할 그룹들을 선택(S110)한다. 예를 들어 제어부(190)는 저장부(150)에 저장되어 있는 모든 그룹들에 대한 정보를 출력부(130)를 통해 출력하고 입력부(110)를 통한 특정 두 그룹 또는 두개 이상의 그룹의 선택 입력을 입력부(110)를 통해 수신한다. 이에 따라 제어부(190)는 제1 그룹과 제2 그룹을 선택 가능하다. 제1 그룹 및 제2 그룹의 선택에 따라 제어부(190)는 각 그룹의 유전자 자료 데이터를 저장부(150)로부터 식별하고 로딩할 수 있다. 제1 그룹 및 제2 그룹은 하나의 동일한 분석 대상 집단에서의 표본들의 집합을 나타낸다. 제1 그룹 및 제2 그룹은 특정 연구소나 툴 등에서 연구되거나 이용되는 유전자 자료의 생성에 이용된 대상 표본들을 나타낸다.
- [0041] 제어부(190)는 세계 이상의 그룹의 선택도 가능하다. 이 때에는 먼저 두 개의 그룹에 대해서 도 2의 흐름을 제어부(190)가 수행한다. 이후 도 2의 흐름의 수행에 따라 통합된 유전자 자료와 나머지 하나의 그룹에 대하여 제어부(190)가 도 2의 흐름을 수행하여 세계의 그룹에 대하여 통합된 유전자형 빈도 및/또는 대립형질 빈도를 산출할 수 있다.
- [0042] 제1 그룹의 유전자 자료 데이터는 적어도 제1 마커 그룹내의 각 마커에서의 유전자형 빈도 데이터를 포함한다. 제2 그룹의 유전자 자료 데이터는 적어도 제2 마커 그룹내의 각 마커에서의 대립형질 빈도 데이터를 포함한다.
- [0043] 제1 마커 그룹과 제2 마커 그룹은 동일할 필요가 없고 다른 경우가 다수 존재한다. 이는 각 연구소나 툴 내에서 이용되는 마커들이 서로 다르기 때문이다. 본 발명은 유전자형 빈도나 대립형질 빈도의 통합에서 서로 다른 마커들을 고려할 수 있도록 설계된다.
- [0044] 마커는 유전자 검사를 위해서 이용되는 유전자좌를 나타낸다. 마커는 예를 들어 STR(Short Tandem Repeat, 또는 Microsatellite라고도 지칭됨)이다. STR은 유전자 변이형으로서 개인식별에 활용된다. STR과 같은 마커는 각 개인간에 차이를 가지고 마커의 서로 다른 형태를 대립형질로 정의된다. 모든 개인은 부모로부터 물려받은 2개의 대립형질(Allele)을 가지게 되고, 2개의 대립형질의 조합을 유전자형(Genotype)이라고 지칭한다. 여기서 유전자형들의 빈도를 알고 있다면 대립형질들의 빈도를 용이하게 산출할 수 있으나 역으로는 용이치 않다. 또한 유전자형의 빈도의 통합을 통해서 통합된 대립형질 빈도를 산출할 수 있다. 유전자형은 두개의 대립형질을 가지므로 유전자형의 개수나 확률이 산출되면 유전자형의 두개의 대립형질의 개수나 확률 산출이 가능하다.
- [0045] 일반적으로 활용되는 마커 그룹은 연구소나 툴마다 서로 다르다. 유전자 통합을 위해 본 발명에서 활용되는 마커 그룹은 제1 마커 그룹과 제2 마커 그룹의 모든 마커를 포함하고 널리 알려져 있는 22개의 모든 마커를 포함한다. 이에 따라 통합된 유전자 자료는 유전자 검사에서 활용되는 마커 그룹에 상관없이 통일적으로 이후 여러 연구소나 툴에서 사용가능하다. 각 그룹에 속한 마커들은 유전자 검사의 검사 체계를 구성한다.
- [0046] 22개의 마커는 D3S1358, D1S1656, D2S441, D10S1248, D13S317, Penta E, D16S539, D18S51, D2S1338, CSF1PO, Penta D, TH01, vWA, D21S11, D7S820, D5S818, TPOX, D8S1179, D12S391, D19S433, FGA, D22S1045이다. 본 발명에서 이용하는 22개의 마커는 미국 연방수사국에서 사용하는 CODIS 마커(D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, vWA, TPOX, D18S51, D5S818, FGA)를 포함하고 다른 알려진 툴에서 사용하는 마커를 포함한다. 이와 같이 본 발명에서 이용하는 마커 그룹의 마커들은 국내 및 국외에서 사용되는 대표적인 마커들을 모두 포함한다.
- [0047] 제1 그룹 및 제2 그룹의 선택에 따라, 제어부(190)는 본 발명에서 활용되는 마커 그룹 중에서 하나의 마커를 선택(S120)한다. 선택된 마커는 STR을 나타내고 유전자 검사 등의 검사 체계에서 이용되는 유전자좌이다.
- [0048] 이후 제어부(190)는 제1 그룹 및 제2 그룹의 유전자 자료 데이터를 이용하여 통합된 유전자형 빈도를 산출(S130)한다. 산출된 유전자형 빈도는 제1 그룹 및 제2 그룹의 통합된 표본들을 대표할 수 있다. 제1 그룹 및 제2 그룹의 유전자 연구 자료를 통합함으로써 특정 분석대상 집단을 대표할 수 있는 유전자형 빈도를 구축할 수

있다. 단계 S130에 대해서는 도 3을 통해서 상세히 살펴보도록 한다.

[0049] 이후 제어부(190)는 본 발명의 마커 그룹의 모든 마커에 대해서 유전자형 빈도의 산출이 이루어졌는 지를 판단(S140)한다. 만일 모든 마커 각각에 대한 유전자형 빈도의 산출이 이루어지지 않은 경우, 제어부(190)는 단계 S120으로 전이하여 다음(다른) 마커를 선택하고 단계 S130을 수행한다. 만일 모든 마커에 대한 유전자형 빈도가 산출되었으면 제어부(190)는 단계 S150으로 전이한다.

[0050] 단계 S150에서 제어부(190)는 제1 그룹 및 제2 그룹으로부터 산출된 유전자형 빈도(평균을 통한 유전자형 빈도)들로부터 제1 그룹 및 제2 그룹을 포함하는 분석대상 집단의 대립형질 빈도를 추정한다.

[0051] 구체적으로, 제어부(190)는 모든 마커 각각에서의 유전자형 빈도를 알고 있다. 유전자형은 일반적으로 두 개의 대립형질로 구성되어 있다. 유전자형 빈도를 산출하였으므로 산출된 유전자형 빈도로부터 하나의 대립형질 빈도를 산출하고 이를 분석대상 집단의 대립형질 빈도로 추정한다. 이러한 대립형질 빈도의 추정을 통해서 제어부(190)는 모든 마커 각각의 모든 대립형질 빈도를 추정가능하다.

[0052] 이후 제어부(190)는 마커 그룹내의 모든 마커의 추정된 대립형질 빈도를 저장부(150)에 저장(S160)한다.

[0053] 저장된 대립형질 빈도 데이터는 이후 출력부(130)를 통해서 배포될 수 있고 특정 다른 응용 툴이나 프로그램 또는 단계 S160 이후에 수행되는 프로그램에 의한 유전자 검사에서 이용된다. 이와 같이 본 발명에 따라 통합된 대립형질 빈도는 이후 다른 유전자 검사에서 활용가능하고 특정 분석대상 집단의 표본을 넓힐 수 있어 대립형질 빈도의 대표성을 점진적으로 확대 및 제공 가능하다.

[0054] 본 발명에 따라 추정된 대립형질 빈도는 각종 유전자 검사에서 이용 가능한데, 이용가능한 유전자 검사로서 대표적으로 개인식별검사, 친자확인검사, 형제검사 등이 있다. 이러한 검사는 프로그램 툴의 형태로 제공되고 특정 마커 그룹을 이용하여 이루어진다. 본 발명의 대립형질 빈도는 알려진 모든 마커를 수용하고 알려진 마커의 대립형질 빈도를 추정하여 여러 상이한 마커 그룹의 프로그램 툴에서 이용 가능하다.

[0055] 또한 본 발명에 따라 제공된 대립형질 빈도는 여러 그룹의 유전자 자료를 통합하여 생성된 자료이기에 이후 각종 검사에서 보다더 정확한 검사 결과를 제공할 수 있도록 한다.

[0056] 도 3은 도 2의 단계 S130의 구체적 흐름을 도시한 도면이다.

[0057] 도 3은 선택된 마커 각각에 대해서 수행되며 모든 마커 각각에 대해서 수행된다. 도 3은 물론 대립형질 빈도 추정 장치(100)(또는 이 장치의 제어부(190))에 의해서 수행된다.

[0058] 도 2의 단계 S130의 진입(S131)에 따라 제어부(190)는 유전자의 선택된 마커의 유전자형 빈도들을 복수개 산출(S132)한다. 선택된 마커는 여러 유전자형(적어도 대립형질이 둘 이상이 있고 이는 적어도 유전자형을 3개를 가짐)을 가진다. 따라서 하나의 마커는 여러 유전자형 각각에 대해서 유전자형 빈도들을 가지고 이 유전자형 빈도들을 제어부(190)가 복수개 산출한다. 유전자형 빈도들 각각을 다른 유전자형 빈도와 구별하기 위해 이하 '산출 슬롯'이라는 용어를 이용한다. 산출 슬롯은 제2 그룹의 대립형질 데이터(구체적으로는 대립형질 빈도 데이터)에 기초하여 산출된다. 각 산출 슬롯의 유전자형 빈도들은 HWE(Hardy-Weinberg Equilibrium, '하디-베인베르크 평형'이라고도 함)를 만족하는 조건하에서 아래의 수학식 1의 다항분포를 통해서 산출된다.

수학식 1

$$P(X_1=x_1, X_2=x_2, \dots, X_k=x_k) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

[0059]

[0060] 여기서, $(0 \leq p_i \leq 1; p_1 + p_2 + \dots + p_k = 1; x_1 + x_2 + \dots + x_k = n)$ 이다. n은 제2 그룹의 총

표본수를 나타내고 p_i (i는 1에서 k)는 HWE 가정하에 제2 그룹의 대립형질 빈도에 의해서 수학식 1의 다항분

포에 따라 계산된 선택된 마커의 각 유전자형(i)의 기대 확률(빈도)이고 x_i 는 선택된 마커의 유전자형의 다항분포에 따른 개수를 나타낸다. 이와 같이, 다항분포의 적용에 의해 각 산출 슬롯은 선택된 마커의 각 유전자형에 대한 유전자형 빈도가 추정될 수 있다.

- [0061] 물론 다항분포에 따라 각각의 산출 슬롯에서 동일한 유전자형 빈도는 서로 다른 빈도를 가질 수 있다. 각 유전자형 빈도는 적어도 HWE 조건을 만족하면서 예측된 값이다.
- [0062] 유전학 관련 업계에서 널리 알려진 바와 같이, 대립형질 빈도가 세대를 거듭해도 변화가 없는 상태를 하디-베인베르그 평형(Hardy-Weinberg equilibrium;HWE)이라고 한다. 인구 집단이 HWE 상태라고 한다면 유전자형에 해당하는 두 종류의 대립형질은 통계학적으로 각각 독립으로 볼 수 있다. 이 때의 유전자형이 나타날 확률은 대립형질이 나타날 확률의 곱으로 표현 가능하다.
- [0063] 예를 들어, 선택된 마커의 대립형질이 12, 13, 15(즉, $k=3$)가 있는 것으로 가정하자. 12, 13, 15 등은 특정 대립형질을 나타내는 식별자이다. 이 마커에서 나타날 수 있는 유전자형은 '12-12', '13-13', '15-15', '12-13', '12-15' 및 '13-15'의 총 6종류가 존재한다. 12, 13, 15의 각 대립형질 빈도를 각각 a,b,c라 할 때, 각 유전자형은 하디-베인베르그 평형을 가정하면 각 기대빈도는 aa, bb, cc, 2ab, 2ac, 2bc 이다. 하디-베인베르그 평형 조건을 만족하는 유전자형 기대 빈도를 가정하여 수학적 식 1에 따른 다항분포를 이용하여 선택된 마커의 각 유전자형 빈도를 제어부(190)가 각각 산출한다.
- [0064] 이와 같이, 단계 S132에서 제어부(190)는 제2 그룹의 표본수에 기초하고 HWE 조건에 부합하며 제2 그룹의 대립형질 빈도로부터 기대되는 유전자형 빈도를 다항분포를 통해 산출한다. 단계 S132를 통해 산출된 복수개의 산출 슬롯은 표본추출에 의한 오류를 최소화하기 위해서 이하 이용된다.
- [0065] 이후 단계 S133에서 제어부(190)는 제1 그룹의 선택된 마커의 유전자형 빈도들과 제2 그룹의 산출된 유전자형 빈도들 각각을 통합한다. 제1 그룹의 유전자형 빈도들과 제2 그룹의 유전자형 빈도들의 통합으로 복수개의 통합된 빈도 데이터(이하 '통합 슬롯'이라고 지칭함)가 생성된다. 예를 들어 산출 슬롯의 개수(예를 들어 산출 슬롯의 개수가 1000개)와 동일한 개수의 통합 슬롯(1000개의 통합 슬롯)이 생성된다. 구체적으로, 제어부(190)는 제1 그룹의 유전자형 빈도들과 하나의 산출 슬롯의 대응하는 유전자형 빈도들 각각을 합산하거나 표본수에 따라 평균하여 두 그룹의 유전자형 빈도를 통합한다.
- [0066] 이러한 통합된 유전자형 빈도들은 HWE를 만족하지 않을 수 있다. HWE를 만족하지 못하는 데이터는 이후 대표적인 유전자형 빈도의 산출에서 필터링될 필요가 있다. 이를 위해, 제어부(190)는 모든 통합 슬롯의 통합된 유전자형 빈도들에 대해서 HWE 정확 검정을 실시(S134)하고 각 통합 슬롯의 유전자형 빈도들이 HWE를 만족하는 지를 확인한다. 만일 HWE를 만족하지 못하는 통합 슬롯이 존재하면, 이 통합 슬롯은 이후에서의 유전자형 빈도 산출의 계산에서 제외된다.
- [0067] HWE 정확 검정(exact test)은 유전학 관련하여 널리 알려져 있다. 정확 검정은 주어진 대립형질 빈도를 만족하는 모든 가능한 유전자형 조합의 발생 확률을 계산하고 이에 대해 실제로 관찰된 유전자형 분포에 대한 확률값보다 작은 확률값을 더하여 누적값을 계산한다. 이 계산된 값이 지정한 수준(예: 5%)보다 작을 때 마커 내 대립형질들의 독립성 가정을 만족하지 못하는 것으로 결정된다.
- [0068] HWE 정확 검정을 통과한 통합 슬롯들을 이용하여 제1 그룹과 제2 그룹에 대한 유전자형 빈도를 산출할 수 있다. 단계 S135에서 제어부(190)는 HWE를 만족하는 통합 슬롯들을 이용하여 마커의 각 유전자형에 대한 평균을 산출(S135)한다.
- [0069] 즉, 제어부(190)는 선택된 마커에서 통과된 통합 슬롯들에서의 대응하는 유전자형 빈도들을 합산하고 이를 HWE를 통과한 통합 슬롯의 개수로 나눔으로써 평균을 산출하거나 평균 확률을 계산하여 평균을 산출한다.
- [0070] 평균 계산에 이용된 통합 슬롯들은 이미 HWE 정확 검정을 통과한 슬롯들이다. 제2 그룹의 대립형질 빈도로부터 예상되는 유전자형 빈도는 다양하게 존재한다. 복수개의 산출 슬롯을 활용하고 이후 통합 슬롯을 생성하며 통과된 통합 슬롯에 대한 평균으로 두 그룹을 대표하는 유전자형 빈도를 산출가능하다.
- [0071] 제어부(190)는 선택된 마커의 특정 유전자형에 대한 유전자형 빈도를 산출된 평균으로 설정(S136)한다. 물론 제어부(190)는 선택된 마커의 모든 유전자형에 대해서 각 유전자형에 대한 평균으로 설정한다. 제어부(190)는 단계 S136의 수행을 완료한 후에 도 3의 수행 흐름을 종료(S137)하고 이후 도 2의 단계 S140을 수행한다.

[0072] 도 3에 따른 유전자형 빈도의 산출에 의해서, 더 많은 표본수(제1 그룹의 표본수 + 제2 그룹의 표본수)를 대표하는 유전자형 빈도를 산출 가능하다. 물론 산출된 유전자형 빈도로부터 대립형질 빈도가 계산되고 이와 같이 계산되는 대립형질 빈도는 더 많은 표본수에 대한 것이므로 이후 유전자 검사에서 신뢰성 있는 검사 결과를 제공할 수 있다. 또한 다항분포를 통해서 HWE를 만족하는 통합된 유전자형 빈도를 산출하여 제1 그룹 및 제2 그룹을 대표하고 유전학적으로 신뢰성을 제공할 수 있는 대립형질 빈도의 제공이 가능하다.

[0073] 이상 살펴본 바와 같이 도 2 및 도 3의 제어 흐름을 통해서, 여러 산재된 유전자 자료를 통합하고 통합에 따라 유전자 검사에서 보다더 높은 신뢰성을 제공할 수 있다. 도 2 및 도 3의 흐름은 프로그램으로 구성되어 배포될 수 있다. 또한 프로그램에 의해서 생성된 대립형질 빈도 또는 유전자형 빈도의 데이터는 데이터베이스나 자료구조를 통해서 배포될 수 있다. 배포는 온라인으로 이루어지거나 오프라인으로 이루어진다.

[0074] 온라인을 통한 배포는 예를 들어 다운로드 서버 등을 통해서 이루어지고 다운로드 서버는 하나 이상의 프로세서, 대용량 저장 매체 및 네트워크 인터페이스를 구비한다. 대용량 저장 매체는 본 발명에 따른 유전자형 빈도/대립형질 빈도의 생성이 가능한 프로그램이나 데이터 구조를 저장하고 프로세서는 네트워크 인터페이스를 통한 다운로드 요청에 따라 이 프로그램이나 대립형질/유전자형 빈도를 나타내는 데이터 구조를 특정 장치로 전송 가능하다.

[0075] 오프라인을 통한 배포는 예를 들어 컴퓨터 프로그램이나 데이터 구조를 저장하는 이동형 메모리, CD, DVD 등을 통해서 이루어질 수 있다.

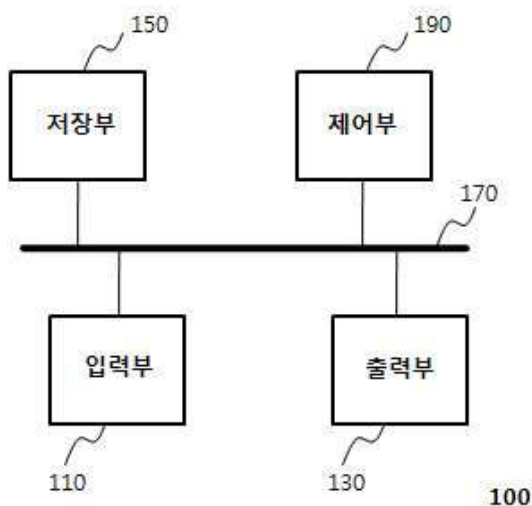
[0076] 이상에서 설명한 본 발명은, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 있어 본 발명의 기술적 사상을 벗어나지 않는 범위 내에서 여러 가지 치환, 변형 및 변경이 가능하므로 전술한 실시 예 및 첨부된 도면에 의해 한정되는 것이 아니다.

부호의 설명

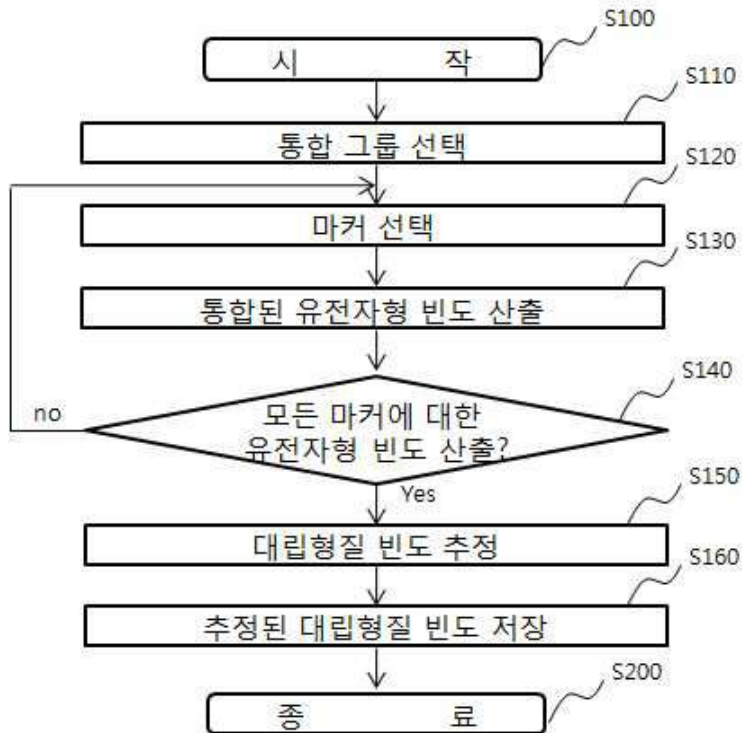
[0077] 100 : 대립형질 빈도 추정 장치
110 : 입력부 130 : 출력부
150 : 저장부 170 : 연결부
190 : 제어부

도면

도면1



도면2



도면3

