(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년02월16일
(11) 등록번호 10-2638152
(24) 등록일자 2024년02월14일

(51) 국제특허분류(Int. Cl.)
G16B 20/00 (2019.01)
(52) CPC특허분류
G16B 20/00 (2019.02)
(21) 출원번호 10-2019-7017174
(22) 출원일자(국제) 2017년11월14일
심사청구일자 2020년10월20일
(85) 번역문제출일자 2019년06월14일
(65) 공개번호 10-2019-0077097
(43) 공개일자 2019년07월02일
(86) 국제출원번호 PCT/US2017/061554
(87) 국제공개번호 WO 2018/093780
국제공개일자 2018년05월24일
(30) 우선권주장
62/422,841 2016년11월16일 미국(US)
62/447,076 2017년01월17일 미국(US)
(56) 선행기술조사문헌
US20150324519 A1*
US20150337388 A1*
WO2013070634 A1*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
일루미나, 인코포레이티드
미국 캘리포니아 92122 샌디에고 일루미나 웨이 5200
(72) 발명자
천, 샤오
미국 92122 캘리포니아주 샌디에고 일루미나 웨이 5200
자오, 천
미국 92122 캘리포니아주 샌디에고 일루미나 웨이 5200
(뒷면에 계속)
(74) 대리인
양영준, 김연송, 백만기

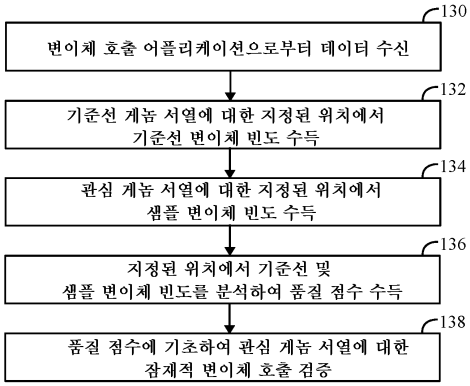
전체 청구항 수 : 총 23 항 심사관 : 정태수

(54) 발명의 명칭 **서열 변이체 호출을 위한 검증 방법 및 시스템**

(57) 요 약

변이체 호출을 검증하기 위한 방법 및 시스템이 제공된다. 관심 게놈 서열을 따라 샘플 판독물에 대한 서열분석 데이터가 수신된다. 관심 게놈 서열을 따라 뉴클레오티드 서열 내의 지정된 위치에서 잠재적 변이체 호출의 표시가 수신된다. 방법 및 시스템은 하나 이상의 기준선 게놈 서열 내의 지정된 위치에서 기준선 변이체 빈도를 수득하고, 관심 게놈 서열에 대한 지정된 위치에서 샘플 변이체 빈도를 결정하고, 지정된 위치에서 기준선 및 샘플 변이체 빈도를 분석하여 품질 점수를 수득하고, 품질 점수를 기초로 하여 관심 게놈 서열에 대한 잠재적 변이체 호출을 검증한다.

대 표 도 - 도1c



(72) 발명자

고든, 제시카

미국 92122 캘리포니아주 샌디에고 일루미나 웨이 5200

장, 샤일

미국 92122 캘리포니아주 샌디에고 일루미나 웨이 5200

장, 팅팅

미국 92122 캘리포니아주 샌디에고 일루미나 웨이 5200

베리, 그웬

미국 92122 캘리포니아주 샌디에고 일루미나 웨이 5200

샤, 예샤

미국 92122 캘리포니아주 샌디에고 일루미나 웨이 5200

창, 한-유

미국 92122 캘리포니아주 샌디에고 일루미나 웨이 5200

명 세 서

청구범위

청구항 1

변이체 호출을 검증하기 위한 컴퓨터 구현 방법으로서,

하기에 대한 프로그램 명령어를 실행하는 1개 이상의 프로세서의 제어 하에,

관심 게놈 서열을 따라 상응하는 뉴클레오티드 서열을 갖는 샘플 판독물(read)을 포함하는 서열분석 데이터를 수신하는 단계;

관심 게놈 서열을 따라 뉴클레오티드 서열 내의 지정된 위치에서 잠재적 변이체 호출의 표시를 수신하는 단계;

하나 이상의 기준선 게놈 서열 내의 지정된 위치에서 기준선 변이체 빈도를 수득하는 단계;

관심 게놈 서열에 대해 지정된 위치에서 샘플 변이체 빈도를 수득하는 단계;

지정된 위치에서 기준선 및 샘플 변이체 빈도를 분석하여 품질 점수를 수득하는 단계이며, 여기서 분석은 기준선 변이체 빈도의 분포와 관련하여 샘플 변이체 빈도를 인덱싱하는 것을 포함하는 것인 단계; 및

품질 점수를 기초로 하여 관심 게놈 서열에 대한 잠재적 변이체 호출을 검증하는 단계

를 포함하는 방법.

청구항 2

제1항에 있어서, 분석 작업이 샘플 변이체 빈도와 기준선 변이체 빈도의 분포 사이의 관계를 수득하는 것을 추가로 포함하며, 품질 점수는 관계에 기초하는 것인 방법.

청구항 3

제2항에 있어서, 관계가 비-파라미터 윌콕슨 순위 합계 검정에 기초하는 것인 방법.

청구항 4

제1항에 있어서, 기준선 변이체 빈도가 기준선 게놈 서열을 따라 상응하는 위치에서의 배경 노이즈의 정도를 표시하는 것인 방법.

청구항 5

제1항에 있어서, 검증하는 단계가 품질 점수를 한계값과 비교하는 단계; 및 품질 점수가 한계값을 초과할 때 잠재적 변이체 호출을 유효한 변이체 호출인 것으로 판정하는 단계를 추가로 포함하는 것인 방법.

청구항 6

제1항에 있어서, 기준선 변이체 빈도가 하나 초과의 유형의 대립유전자와 연관된 다수의 기준선 게놈 서열로부터 유래된 것인 방법.

청구항 7

제1항에 있어서, 기준선 게놈 서열을 따라 뉴클레오티드 서열의 복수의 참조 판독물을 포함하는 서열분석 데이터를 수신하는 단계, 및 지정된 위치에서 참조 판독물에 대한 기준선 변이체 빈도를 결정하는 단계를 추가로 포함하는 방법.

청구항 8

제7항에 있어서, 기준선 변이체 빈도를 결정하는 단계가

a. 참조 판독물을 따라 현재 염기 쌍 윈도우 내의 일련의 위치들에 대한 서열분석 데이터를 수신하는

단계;

b. 현재 염기 쌍 윈도우 내의 위치 세트 내의 1개 이상의 위치에 대한 후보 변이체 빈도를 확인하는 단계;

c. 참조 판독물 내의 지정된 위치에 대한 기준선 변이체 빈도로서 후보 변이체 빈도 중 하나를 선택하는 단계; 및

d. 염기 쌍 윈도우를 기준선 게놈 서열을 따라 이동시키고, a, b 및 c의 작업을 반복하는 단계

를 추가로 포함하는 것인 방법.

청구항 9

변이체 호출을 검증하기 위한 컴퓨터 구현 방법으로서,

하기에 대한 프로그램 명령어를 실행하는 1개 이상의 프로세서의 제어 하에,

관심 게놈 서열에 대한 잠재적 변이체 호출의 표시를 수신하는 단계;

관심 게놈 서열을 따라 샘플 및 미가공 뉴클레오티드 단편에 대한 판독물을 포함하는 서열분석 데이터를 수신하는 단계이며, 상기 판독물은 관심 게놈 서열을 따라 지정된 위치에서 뉴클레오티드 서열에 상응하는 샘플 단편에 대한 샘플 판독물을 포함하는 것인 단계;

기준선 게놈 서열에 대한 지정된 위치에서의 기준선 변이체 빈도와 관련하여 관심 게놈 서열에 대한 지정된 위치에서의 샘플 변이체 빈도를 분석하여 품질 점수를 수득하는 단계이며, 여기서 분석은 기준선 변이체 빈도의 분포와 관련하여 샘플 변이체 빈도를 인덱싱하는 것을 포함하는 것인 단계; 및

하기:

A) 미가공 단편이 잠재적 변이체 호출을 확인하는지 여부;

B) 샘플 판독물이 샘플 단편에 대한 미리 결정된 양의 커버리지를 제공하는지 여부; 및

C) 잠재적 변이체 호출이 미리 규정된 집단에 걸쳐 나타난 이전 변이체 호출과 매칭되는지 여부

중 적어도 하나를 결정하는 단계; 및

품질 점수 및 결정 작업에 기초하여 잠재적 변이체 호출을 검증하는 단계

를 포함하는 방법.

청구항 10

제9항에 있어서, 결정하는 단계가 작업 A)를 포함하고, 작업 A)는

지지 변이체 호출을 표시하는 미가공 단편을 식별하는 단계;

지지 변이체 호출을 표시하는 미가공 단편에 대해 가중 단편 점수를 수득하는 단계; 및

가중 단편 점수를 미가공 단편 한계값과 비교하여 잠재적 변이체 호출을 확인하는지 여부를 결정하는 단계

를 포함하는 것인 방법.

청구항 11

제9항에 있어서, 결정하는 단계가 작업 A)를 포함하고, 미가공 단편이 듀플렉스 스티칭된 단편, 심플렉스 스티칭된 단편, 듀플렉스 비-스티칭된 단편, 또는 심플렉스 비-스티칭된 단편 중 적어도 하나에 상응하는 것인 방법.

청구항 12

제9항에 있어서, 결정하는 단계가 작업 A)를 포함하고, 작업 A)는 듀플렉스 스티칭된 단편을 미가공 단편으로서 식별하는 단계; 및 듀플렉스 스티칭된 단편이 잠재적 변이체 호출을 확인하는지 여부를 결정하기 위한 지지 변이체 호출을 표시하는지 여부를 결정하는 단계를 포함하는 것인 방법.

청구항 13

제9항에 있어서, 결정하는 단계가 작업 B)를 포함하고, 샘플 판독물이 미리 결정된 양의 커버리지 미만으로 떨어질 때 잠재적 변이체 호출이 호출-부재인 것으로 판정되는 것인 방법.

청구항 14

제9항에 있어서, 결정하는 단계가 작업 C)를 포함하고:

i. 잠재적 및 이전 변이체 호출 사이에 매치가 결정되는 경우, 결정하는 작업은 제1 한계값을 이용하여 작업 A) 또는 작업 B) 중 적어도 하나를 수행하고;

ii. 잠재적 및 이전 변이체 호출 사이에 매치가 존재하지 않는 경우, 결정하는 작업은 제2 한계값을 이용하여 작업 A) 또는 작업 B) 중 적어도 하나를 수행하는 것인

방법.

청구항 15

제9항에 있어서, 분석 작업 및 결정 작업 A), B) 및 C)가 계층식으로 수행되는 것인 방법.

청구항 16

변이체 호출을 검증하기 위한 시스템으로서,

프로그램 명령어 및 관심 게놈 서열을 따라 상응하는 뉴클레오티드 서열을 갖는 샘플 판독물을 포함하는 서열분석 데이터를 저장하는 메모리;

프로그램 명령어를 실행할 때,

관심 게놈 서열을 따라 뉴클레오티드 서열 내의 지정된 위치에서 잠재적 변이체 호출의 표시를 수신하고;

복수의 기준선 게놈 서열 내의 지정된 위치에서 기준선 변이체 빈도를 수득하고;

관심 게놈 서열에 대한 지정된 위치에서 샘플 변이체 빈도를 수득하고;

지정된 위치에서 기준선 및 샘플 변이체 빈도를 분석하여 품질 점수를 수득하고, 여기서 분석은 기준선 변이체 빈도의 분포와 관련하여 샘플 변이체 빈도를 인덱싱하는 것을 포함하는 것이며;

품질 점수를 기초로 하여 관심 게놈 서열에 대한 잠재적 변이체 호출을 검증하는

1개 이상의 프로세서

를 포함하는 시스템.

청구항 17

제16항에 있어서, 1개 이상의 프로세서가 샘플 변이체 빈도와 기준선 변이체 빈도의 분포 사이의 관계를 수득하며, 품질 점수는 관계에 기초하는 것인 시스템.

청구항 18

제16항에 있어서, 기준선 변이체 빈도가 기준선 게놈 서열을 따라 상응하는 유전자좌에서의 배경 노이즈의 정도를 표시하는 것인 시스템.

청구항 19

제16항에 있어서, 1개 이상의 프로세서가 품질 점수를 한계값과 비교하고; 비교에 기초하여 잠재적 변이체 호출을 무효인 것으로 판정하는 표시를 출력하는 것인 시스템.

청구항 20

제16항에 있어서, 기준선 게놈 서열을 따라 뉴클레오티드의 복수의 참조 판독물을 포함하는 서열분석 데이터를 저장하는 메모리, 지정된 위치에서 참조 판독물에 대해 기준선 변이체 빈도를 결정하기 위한 1개 이상의 프로세

서를 추가로 포함하는 시스템.

청구항 21

변이체 호출을 검증하기 위한 시스템으로서,

프로그램 명령어 및 관심 게놈 서열을 따라 샘플 및 미가공 뉴클레오티드 단편에 대한 판독물을 포함하는 서열 분석 데이터를 저장하는 메모리이며, 상기 판독물은 관심 게놈 서열을 따라 지정된 위치에서 뉴클레오티드 서열에 상응하는 샘플 단편에 대한 샘플 판독물을 포함하는 것인 메모리;

프로그램 명령어를 실행할 때,

관심 게놈 서열에 대한 잠재적 변이체 호출의 표시를 수신하고;

기준선 게놈 서열에 대한 지정된 위치에서의 기준선 변이체 빈도와 관련하여 관심 게놈 서열에 대한 지정된 위치에서의 샘플 변이체 빈도를 분석하여 품질 점수를 수득하고, 여기서 분석은 기준선 변이체 빈도의 분포와 관련하여 샘플 변이체 빈도를 인덱싱하는 것을 포함하며;

하기:

A) 미가공 단편이 잠재적 변이체 호출을 확인하는지 여부;

B) 샘플 판독물이 샘플 단편에 대한 미리 결정된 양의 커버리지를 제공하는지 여부; 및

C) 잠재적 변이체 호출이 미리 규정된 집단에 걸쳐 나타난 이전 변이체 호출과 매칭되는지 여부

중 적어도 하나를 결정하고;

품질 점수 및 결정 작업에 기초하여 잠재적 변이체 호출을 검증하는

1개 이상의 프로세서

를 포함하는 시스템.

청구항 22

제21항에 있어서, 결정이 작업 A)를 포함하고, 작업 A)는

지지 변이체 호출을 표시하는 미가공 단편을 식별하고;

지지 변이체 호출을 표시하는 미가공 단편에 대해 가중 단편 점수를 수득하고;

가중 단편 점수를 미가공 단편 한계값과 비교하여 잠재적 변이체 호출을 확인하는지 여부를 결정하는 것

을 포함하는 것인 시스템.

청구항 23

제21항에 있어서, 결정이 작업 A)를 포함하고, 미가공 단편이 듀플렉스 스티칭된 단편, 심플렉스 스티칭된 단편, 듀플렉스 비-스티칭된 단편, 또는 심플렉스 비-스티칭된 단편 중 적어도 하나에 상응하는 것인 시스템.

청구항 24

삭제

청구항 25

삭제

발명의 설명

기 술 분 야

[0001] 본 출원은 2017년 1월 17일에 출원된, 발명의 명칭이 "서열 변이체 호출을 위한 검증 방법 및 시스템"인 미국 특허 가출원 번호 62/447,076 및 2016년 11월 16일에 출원된, 발명의 명칭이 "변이체 호출에서의 정확도를 개선

시키기 위한 방법 및 시스템"인 미국 특허 가출원 번호 62/422,841에 관한 것이고, 그에 대해 우선권의 이익을 주장하며, 이들의 완전한 대상은 그 전문이 참조로 포함된다.

배 경 기 술

[0002] 본원의 대상은 일반적으로 서열분석 작업을 분석하기 위한 시스템 및 방법, 및 보다 특히 서열분석 작업과 관련하여 만들어진 변이체 호출을 검증하기 위한 시스템 및 방법에 관한 것이다.

[0003] 오늘날 암 치료 기술은 어디든-적합한-약물(one-drug-fits-all) 접근법으로부터 멀어지고 있다. 포괄적인 진단 접근법은 DNA 변이체를 검출하는 방법에 대한 문제점을 발생시킨다. DNA 검출 방법의 비제한적 예는 생어 서열분석, 면역조직화학(IHC), 형광 계내 혼성화(FISH), 및 정량적 PCR(qPCR)을 포함한다. 일부 검출 방법은 한 번에 하나 또는 몇몇의 유전자 변화에 대해 DNA를 분석한다. 그러나, 암 진행을 유도하는 것으로 공지된 DNA 변경의 수가 계속해서 늘어남에 따라, 1회의 서열분석 작업 동안 더욱 더 많은 유전자 변화에 대해 DNA를 분석하기 위한 새로운 방법이 제안된다.

[0004] 차세대 서열분석(NGS)은 1종의 샘플을 사용하여 다수의 유전자 내 변이체를 평가하는 능력을 제공한다. 이를 위해, NextSeq™ 및 HiSeq™ 플랫폼 상에서의 서열분석을 위한 다수의 암 유전자(예를 들어, 100개 초과의 암 유전자)를 표적화하는 하이브리드 포획 기반 NGS 시험(검정)을 제공하는 시스템이 제안되었다. NGS 시험은 단일 뉴클레오티드 변이체(SNV), 소형 삽입 및 결실(indel), 다수의 뉴클레오티드 변이체(MNV), 유전자 증폭(CNV)의 식별을 위한 DNA 작업흐름을 포함한다. NGS 시험은 또한 스플라이스 변이체 및 유전자 융합체의 식별을 위한 RNA 작업흐름을 포함한다. 변이체는 샘플 핵산 서열이 서열에 따라 1개 이상의 염기 쌍 위치에서 참조 서열과 상이한 것으로 결정되는 경우에 식별된다. 단일 샘플을 서열분석하면서 공통 시점에 상이한 유전자로부터의 DNA 및 RNA 변경 둘 다를 평가할 수 있다.

[0005] 급속 전환 시간 내에 변이체 호출을 전달하는 생물정보학 분석 기술이 개발되고 있다. 상기 기술은 특히 정렬기 및 변이체 호출기를 포함한다. 정렬 프로세스는 긴 indel을 정렬시켜, 짧은 판독물 길이를 갖는 샷건 서열분석으로부터의 문제점을 극복할 수 있다. 추가로, NGS 시험은 다양한 샘플 제조 기술을 이용하며, 그의 한 예는 포르말린-고정, 파라핀-포매(FFPE)이다. FFPE 샘플은 관심 DNA가 낮은 존재비를 갖고/거나 분해되었을 수 있는 경우에 DNA 변이체를 검출함에 있어서 또 다른 어려움을 제시한다.

[0006] 그러나, 현재 제안된 서열분석 기술은 변이체 호출과 관련하여 거짓 양성을 경험한다. 예를 들어, 상기 기술은 변이체가 특정한 위치(염기 쌍)에서 샘플 서열에 존재하는 것을 부정확하게 결정하고/거나 변이체의 유형을 부정확하게 식별할 수 있다(일반적으로 거짓 양성으로 지칭됨). 시스템적 오류는, 예컨대 FFPE 아티팩트, 서열분석 오류 또는 PCR 오류로 인해 거짓 양성을 유발할 수 있다.

[0007] 과도하게 많은 수의 거짓 양성을 판정하지 않으면서 단일 샘플에서 간단하고 비용-효과적인 방식으로 다수의 유형의 DNA 변이체를 동시에 조사할 수 있는 서열분석 방법 및 시스템에 대한 필요가 남아있다.

발명의 내용

[0008] 정의

[0009] 특허, 특허 출원, 기사, 서적, 논문 및 웹페이지를 포함하나 이에 제한되지 않는 본 출원에 인용된 모든 문헌 및 유사 자료는 이러한 문헌 및 유사 자료의 포맷에 관계없이 그 전문이 명백하게 참조로 포함된다. 정의된 용어, 용어 용법, 기재된 기술 등을 포함하나 이에 제한되지 않는, 포함된 문헌 및 유사 자료 중 하나 이상이 본 출원과 상이하거나 모순되는 경우에, 본 출원이 우선한다.

[0010] 본원에 사용된 하기 용어는 나타낸 의미를 갖는다.

[0011] 용어 "염색체"는 DNA 및 단백질 성분(특히 히스톤)을 포함하는 염색질 가닥으로부터 유래된, 살아있는 세포의 유전성-보유 유전자 캐리어를 지칭한다. 통상적인 국제적으로 인식되는 개별 인간 게놈 염색체 넘버링 시스템이 본원에 사용된다.

[0012] 용어 "부위"는 참조 게놈 상의 특유한 위치(예를 들어, 염색체 ID, 염색체 위치 및 배향)를 지칭한다. 일부 실시양태에서, 부위는 서열 상의 잔기, 서열 태그, 또는 절편의 위치일 수 있다. 용어 "유전자좌"는 참조 염색체 상의 핵산 서열 또는 다형성의 특정 위치를 지칭하기 위해 사용될 수 있다.

[0013] 본원에서 용어 "샘플"은, 전형적으로 서열분석되고/거나 위상화될 적어도 1개의 핵산 서열을 함유하는 핵산 또

는 핵산들의 혼합물을 함유하는 생물학적 유체, 세포, 조직, 기관, 또는 유기체로부터 유래된 샘플을 지칭한다. 이러한 샘플은 객담/타액, 양수, 혈액, 혈액 분획, 미세 바늘 생검 샘플(예를 들어, 외과적 생검, 미세 바늘 생검 등), 소변, 복막액, 흉막액, 조직 외식편, 기관 배양물 및 임의의 다른 조직 또는 세포 표본, 또는 그의 또는 그로부터 단리된 분획 또는 유도체를 포함하나 이에 제한되지는 않는다. 샘플은 종종 인간 대상체(예를 들어, 환자)로부터 취해지지만, 개, 고양이, 말, 염소, 양, 소, 돼지 등을 포함하나 이에 제한되지는 않는, 염색체를 갖는 임의의 유기체로부터 샘플을 취할 수 있다. 상기 샘플은 직접적으로 생물학적 공급원으로부터 수득한 대로 사용할 수 있거나 또는 샘플의 특징을 변형시키는 전처리 후에 사용할 수 있다. 예를 들어, 이러한 전처리는 혈액으로부터 혈장을 제조하는 것, 점성 유체를 희석하는 것 등을 포함할 수 있다. 전처리 방법은 또한 여과, 침전, 희석, 증류, 혼합, 원심분리, 동결, 동결건조, 농축, 증폭, 핵산 단편화, 간섭 성분의 불활성화, 시약의 첨가, 용해 등을 수반할 수 있으나 이에 제한되지는 않는다.

[0014] 용어 "서열"은 서로 커플링된 뉴클레오티드의 가닥을 포함하거나 이를 나타낸다. 뉴클레오티드는 DNA 또는 RNA에 기초할 수 있다. 1개의 서열이 다수의 하위-서열을 포함할 수 있는 것으로 이해되어야 한다. 예를 들어, 단일 서열(예를 들어, PCR 앰플리콘의 것)은 350개의 뉴클레오티드를 가질 수 있다. 샘플 판독물은 이들 350개 뉴클레오티드 내의 다수의 하위-서열을 포함할 수 있다. 예를 들어, 샘플 판독물은 예를 들어 20-50개의 뉴클레오티드를 갖는 제1 및 제2 플랭킹 하위-서열을 포함할 수 있다. 제1 및 제2 플랭킹 하위-서열은 상응하는 하위-서열(예를 들어, 40-100개 뉴클레오티드)을 갖는 반복 절편의 어느 한 쪽 상에 위치할 수 있다. 각각의 플랭킹 하위-서열은 프라이머 하위-서열(예를 들어, 10-30개 뉴클레오티드)을 포함할 수 있다(또는 그의 부분을 포함할 수 있다). 판독의 용이성을 위해, 용어 "하위-서열"은 "서열"로 지칭될 것이지만, 2개의 서열이 공통 가닥 상에서 서로 반드시 분리되는 것은 아닌 것으로 이해된다. 본원에 기재된 다양한 서열을 구별하기 위해, 서열에 상이한 표지가 주어질 수 있다(예를 들어, 표적 서열, 프라이머 서열, 플랭킹 서열, 참조 서열 등). 다른 용어, 예컨대 "대립유전자"에는 유사한 대상들 사이를 구별하기 위해 상이한 표지가 주어질 수 있다.

[0015] 용어 "쌍형성된-말단 서열분석"은 표적 단편의 양쪽 말단을 서열분석하는 서열분석 방법을 지칭한다. 쌍형성된-말단 서열분석은 게놈 재배열 및 반복 절편, 뿐만 아니라 유전자 융합체 및 신규 전사체의 검출을 용이하게 할 수 있다. 쌍형성된-말단 서열분석을 위한 방법론은 PCT 공개 WO07010252, PCT 출원 일련 번호 PCTGB2007/003798 및 미국 특허 출원 공개 US 2009/0088327에 기재되어 있으며, 이들 각각은 본원에 참조로 포함된다. 한 예에서, 일련의 작업은 하기와 같이 수행될 수 있다; (a) 핵산의 클러스터를 생성하고; (b) 핵산을 선형화하고; (c) 제1 서열분석 프라이머를 혼성화하고, 상기 제시된 바와 같은 연장, 스캐닝 및 탈차단의 반복 사이클을 수행하고; (d) 상보적 카피를 합성함으로써 유동 셀 표면 상의 표적 핵산을 "역전"시키고; (e) 재합성된 가닥을 선형화하고; (f) 제2 서열분석 프라이머를 혼성화하고, 상기 제시된 바와 같은 연장, 스캐닝 및 탈차단의 반복 사이클을 수행한다. 역전 작업은 브릿지 증폭의 단일 사이클에 대해 상기 제시된 바와 같은 시약을 전달하는 것으로 수행될 수 있다.

[0016] 용어 "참조 게놈" 또는 "참조 서열"은 대상체로부터의 식별된 서열을 참조하는데 사용될 수 있는 임의의 유기체의 임의의 특정한 공지된 게놈 서열(부분적이든 완전한 것이든)을 지칭한다. 예를 들어, 인간 대상체뿐만 아니라 많은 다른 유기체에 사용되는 참조 게놈은 국립 생물 정보 센터(ncbi.nlm.nih.gov)에서 찾아볼 수 있다. "게놈"은 핵산 서열에서 발현되는, 유기체 또는 바이러스의 완전한 유전 정보를 지칭한다. 게놈은 유전자 및 DNA의 비-코딩 서열 둘 다를 포함한다. 참조 서열은 그에 대해 정렬된 판독물보다 더 클 수 있다. 예를 들어, 그것은 적어도 약 100배 더 크거나, 또는 적어도 약 1000배 더 크거나, 또는 적어도 약 10,000배 더 크거나, 또는 적어도 약 10^5배 더 크거나, 또는 적어도 약 10^6배 더 크거나, 또는 적어도 약 10^7배 더 클 수 있다. 한 예에서, 참조 게놈 서열은 전장 인간 게놈의 것이다. 또 다른 예에서, 참조 게놈 서열은 특정 인간 염색체 예컨대 염색체 13으로 제한된다. 일부 실시양태에서, 참조 염색체는 인간 게놈 버전 hg19로부터의 염색체 서열이다. 이러한 서열은 염색체 참조 서열로 지칭될 수 있지만, 용어 참조 게놈은 이러한 서열을 커버하는 것으로 의도된다. 참조 서열의 다른 예는 다른 종의 게놈, 뿐만 아니라 임의의 종의 염색체, 하위-염색체 영역(예컨대 가닥) 등을 포함한다. 다양한 실시양태에서, 참조 게놈은 다수의 개체로부터 유래된 컨센서스 서열 또는 다른 조합이다. 그러나, 특정 적용에서, 참조 서열은 특정한 개체로부터 취해질 수 있다.

[0017] 용어 "판독물(read)"은 뉴클레오티드 샘플 또는 참조물의 단편을 기재하는 서열 데이터의 집합을 지칭한다. 용어 "판독물"은 샘플 판독물 및/또는 참조 판독물을 지칭할 수 있다. 전형적으로, 반드시는 아니지만, 판독물은 샘플 또는 참조물에서의 인접한 염기 쌍의 짧은 서열을 나타낸다. 판독물은 샘플 또는 참조 단편의 염기 쌍 서열에 의해 기호로 (ATCG로) 나타내어질 수 있다. 이는 메모리 디바이스에 저장될 수 있고, 판독물이 참조 서열과 매칭되는지 또는 다른 기준을 충족시키는지 결정하기 위해 적절하게 프로세싱될 수 있다. 판독물은 서열분

석 장치로부터 직접적으로 또는 저장된 샘플 관련 서열 정보로부터 간접적으로 수득될 수 있다. 일부 경우에, 판독물은 더 큰 서열 또는 영역을 식별하는데 사용될 수 있는, 예를 들어 염색체 또는 게놈 영역 또는 유전자에 대해 정렬되고 특이적으로 할당될 수 있는, 충분한 길이(예를 들어, 적어도 약 25 bp)의 DNA 서열이다.

[0018] 차세대 서열분석 방법은 예를 들어 합성 기술에 의한 서열분석(일루미나(Illumina)), 파이로시퀀싱(454), 이온 반도체 기술(이온 토렌트 서열분석), 단일-분자 실시간 서열분석(퍼시픽 바이오사이언시즈(Pacific Biosciences)) 및 라이게이션에 의한 서열분석(SOLiD 서열분석)을 포함한다. 서열분석 방법에 따라, 각각의 판독물의 길이는 약 30 bp에서 10,000 bp 초과로 다양할 수 있다. 예를 들어, SOLiD 서열분석기를 사용하는 일루미나 서열분석 방법은 약 50 bp의 핵산 판독물을 생성한다. 또 다른 예의 경우, 이온 토렌트 서열분석은 최대 400 bp의 핵산 판독물을 생성하고, 454 파이로시퀀싱은 약 700 bp의 핵산 판독물을 생성한다. 또 다른 예의 경우, 단일-분자 실시간 서열분석 방법은 10,000 bp 내지 15,000 bp의 판독물을 생성할 수 있다. 따라서, 특정 실시양태에서, 핵산 서열 판독물은 30-100 bp, 50-200 bp, 또는 50-400 bp의 길이를 갖는다.

[0019] 용어 "샘플 판독물", "샘플 서열" 또는 "샘플 단편"은 샘플로부터의 관심 게놈 서열에 대한 서열 데이터를 지칭한다. 예를 들어, 샘플 판독물은 정방향 및 역방향 프라이머 서열을 갖는 PCR 앰플리콘으로부터의 서열 데이터를 포함한다. 서열 데이터는 임의의 선택된 서열 방법론으로부터 수득할 수 있다. 샘플 판독물은 예를 들어 합성에-의한-서열분석(SBS) 반응, 라이게이션에-의한-서열분석 반응, 또는 반복 요소의 길이 및/또는 정체성을 결정하는데 바람직한 임의의 다른 적합한 서열분석 방법론으로부터의 것일 수 있다. 샘플 판독물은 다수의 샘플 판독물로부터 유래된 컨센서스(예를 들어, 평균 또는 가중) 서열일 수 있다. 특정 실시양태에서, 참조 서열을 제공하는 것은 PCR 앰플리콘의 프라이머 서열에 기초하여 관심 유전자좌를 식별하는 것을 포함한다.

[0020] 용어 "미가공 단편"은 샘플 판독물 또는 샘플 단편 내의 관심 있는 지정된 위치 또는 2차적 위치와 적어도 부분적으로 중첩되는, 관심 게놈 서열의 부분에 대한 서열 데이터를 지칭한다. 미가공 단편의 비제한적 예는 듀플렉스 스티칭된 단편, 심플렉스 스티칭된 단편, 듀플렉스 비-스티칭된 단편 및 심플렉스 비-스티칭된 단편을 포함한다. 용어 "미가공"은 미가공 단편이 샘플 판독물 내 잠재적 변이체에 상응하고 그를 인증하거나 또는 그를 확인하는 지지 변이체를 나타내는지와 관계없이, 샘플 판독물 내 서열 데이터와 일부 관련된 서열 데이터를 포함한다는 것을 나타내는데 사용된다. 용어 "미가공 단편"은 단편이 샘플 판독물 내 변이체 호출을 검증하는 지지 변이체를 반드시 포함한다는 것을 나타내지는 않는다. 예를 들어, 샘플 판독물이 변이체 호출 어플리케이션에 의해 제1 변이체를 나타내는 것으로 결정된 경우에, 변이체 호출 어플리케이션은 1개 이상의 미가공 단편이, 그렇지 않은 경우 샘플 판독물 내 변이체를 고려하여 발생할 것으로 예상될 수 있는 "지지" 변이체의 상응하는 유형이 결여되어 있다는 것을 결정할 수 있다.

[0021] 용어 "정렬된", "정렬", 또는 "정렬하는"은 판독물 또는 태그를 참조 서열과 비교하여 참조 서열이 판독물 서열을 함유하는지를 결정하는 프로세스를 지칭한다. 참조 서열이 판독물을 함유하는 경우, 판독물은 참조 서열에 맵핑될 수 있거나 또는, 특정 실시양태에서, 참조 서열 내 특정한 위치에 맵핑될 수 있다. 일부 경우에, 정렬은 판독물이 단순하게 특정한 참조 서열의 구성원인지 아닌지(즉, 판독물이 참조 서열 내에 존재하는지 부재하는지)를 말해준다. 예를 들어, 인간 염색체 13에 대한 참조 서열에 대한 판독물의 정렬은 판독물이 염색체 13에 대한 참조 서열 내에 존재하는지를 말해줄 것이다. 이러한 정보를 제공하는 도구는 세트 멤버쉽 테스터로 불릴 수 있다. 일부 경우에, 정렬은 추가적으로 판독물 또는 태그가 맵핑되는 참조 서열 내 위치를 나타낸다. 예를 들어, 참조 서열이 전체 인간 게놈 서열인 경우, 정렬은 판독물이 염색체 13 상에 존재한다는 것을 나타낼 수 있고, 판독물이 염색체 13의 특정한 가닥 및/또는 부위 상에 있다는 것을 추가로 나타낼 수 있다.

[0022] 용어 "indel"은 유기체의 DNA에서의 염기의 삽입 및/또는 결실을 지칭한다. 마이크로-indel은 1 내지 50개 뉴클레오티드의 순 변화를 발생시키는 indel을 나타낸다. 게놈의 코딩 영역에서, indel의 길이가 3의 배수가 아닌 경우, 이는 프레임시프트 돌연변이를 생성할 것이다. indel은 점 돌연변이와 대조될 수 있다. indel은 서열로부터 뉴클레오티드를 삽입 및 결실시키는 반면, 점 돌연변이는 DNA 내 전체 수를 변화시키지 않으면서 뉴클레오티드 중 1개를 대체하는 치환의 형태이다. indel은 또한 탠덤 염기 돌연변이(TBM)와 대조될 수 있으며, 이는 인접한 뉴클레오티드에서의 치환(주로 2개의 인접한 뉴클레오티드에서의 치환이지만, 3개의 인접한 뉴클레오티드에서의 치환이 관찰됨)으로 정의될 수 있다.

[0023] 용어 "변이체"는 핵산 참조물과 상이한 핵산 서열을 지칭한다. 전형적인 핵산 서열 변이체는 비제한적으로 단일 뉴클레오티드 다형성(SNP), 짧은 결실 및 삽입 다형성(Indel), 카피 수 변이(CNV), 미소위성체 마커 또는 짧은 탠덤 반복물 및 구조적 변이를 포함한다. 체세포 변이체 호출은 DNA 샘플에서 낮은 빈도로 존재하는 변이체를 식별하기 위한 것이다. 체세포 변이체 호출은 암 치료와 관련하여 관심 대상이다. 암은 DNA에서의 돌연변

이의 축적에 의해 유발된다. 일부 정상 세포, 암 진행의 초기 단계에 있는 일부 세포(보다 적은 돌연변이를 가짐), 및 일부 후기-단계 세포(보다 많은 돌연변이를 가짐)를 포함한 종양으로부터의 DNA 샘플은 일반적으로 불균질하다. 이러한 불균질성 때문에, (예를 들어, FFPE 샘플로부터) 종양을 서열분석하는 경우에, 체세포 돌연변이는 종종 낮은 빈도로 나타날 것이다. 예를 들어, SNV는 주어진 염기를 커버하는 판독물의 단지 10%에서 관찰될 수 있다.

[0024] 용어 "변이체 빈도"는 분율 또는 백분율로서 표현되는, 집단 내 특정한 유전자좌에서의 대립유전자(유전자의 변이체)의 상대적 빈도를 나타낸다. 예를 들어, 분율 또는 백분율은 그 대립유전자를 보유하는 집단 내의 모든 염색체의 분율일 수 있다. 예로서, 샘플 변이체 빈도는 개체로부터의 관심 게놈 서열에 대해 수득된 판독물 및/또는 샘플의 수에 상응하는 "개체수" 대비, 관심 게놈 서열에 따라 특정한 유전자좌/위치에서의 대립유전자/변이체의 상대적 빈도를 나타낸다. 또 다른 예로서, 기준선 변이체 빈도는 1개 이상의 기준선 게놈 서열에 따라 특정한 유전자좌/위치에서의 대립유전자/변이체의 상대적 빈도를 나타내며, 여기서 "개체수"는 정상 개체의 집단으로부터의 1개 이상의 기준선 게놈 서열에 대해 수득된 판독물 및/또는 샘플의 수에 상응한다.

[0025] 용어 "위치", "지정된 위치", 및 "유전자좌"는 뉴클레오티드 서열 내의 1개 이상의 뉴클레오티드의 위치 또는 좌표를 지칭한다. 용어 "위치", "지정된 위치", 및 "유전자좌"는 또한 뉴클레오티드 서열 내의 1개 이상의 염기 쌍의 위치 또는 좌표를 지칭한다.

[0026] 용어 "반수체형"은 함께 유전되는, 염색체 상의 인접한 부위에서의 대립유전자의 조합을 지칭한다. 반수체형은, 만약 발생한다면 주어진 유전자좌 세트 사이에 발생한 재조합 사건의 수에 따라 1개의 유전자좌, 여러 유전자좌, 또는 전체 염색체일 수 있다.

[0027] 본원에서 용어 "한계값"은 샘플, 핵산, 또는 그의 부분(예를 들어, 판독물)을 특징화하기 위한 컷오프로서 사용되는 수치 또는 비-수치 값을 지칭한다. 한계값은 실험적 분석에 기초하여 달라질 수 있다. 한계값을 측정된 값 또는 계산된 값과 비교하여 이러한 값을 생성하는 공급원이 특정한 방식으로 분류되어야 함을 제안하는지를 결정할 수 있다. 한계값은 경험적으로 또는 분석적으로 식별될 수 있다. 한계값의 선택은 사용자가 분류를 함에 있어서 갖고자 하는 신뢰 수준에 따라 좌우된다. 한계값은 특정한 목적을 위해(예를 들어, 민감도 및 선택성의 균형을 이루기 위해) 선택될 수 있다. 본원에 사용된 용어 "한계값"은 분석 과정이 변화될 수 있는 지점 및/또는 작용이 촉발될 수 있는 지점을 나타낸다. 한계값은 미리 결정된 수일 것으로 요구되지 않는다. 대신에, 한계값은 예를 들어 복수의 인자에 기초한 함수일 수 있다. 한계값은 상황에 적응될 수 있다. 더욱이, 한계값은 상한, 하한, 또는 그 사이의 범위를 나타낼 수 있다.

[0028] 일부 실시양태에서, 서열분석 데이터에 기초한 메트릭 또는 점수를 한계값과 비교할 수 있다. 본원에 사용된 용어 "메트릭" 또는 "점수"는 서열분석 데이터로부터 결정된 값 또는 결과를 포함할 수 있거나, 또는 서열분석 데이터로부터 결정된 값 또는 결과에 기초한 함수를 포함할 수 있다. 한계값과 마찬가지로, 메트릭 또는 점수는 상황에 적응될 수 있다. 예를 들어, 메트릭 또는 점수는 정규화된 값일 수 있다. 점수 또는 메트릭의 예로서, 하나 이상의 실시양태는 데이터를 분석하는 경우에 카운트 점수를 사용할 수 있다. 카운트 점수는 샘플 판독물의 수에 기초할 수 있다. 샘플 판독물은 샘플 판독물이 적어도 1개의 공통 특징 또는 품질을 갖도록 1회 이상의 필터링 단계를 거칠 수 있다. 예를 들어, 카운트 점수를 결정하는데 사용되는 각각의 샘플 판독물은 참조 서열과 정렬되어 있을 수 있거나 또는 잠재적 대립유전자로 할당될 수 있다. 공통 특징을 갖는 샘플 판독물의 수를 카운팅하여 판독물 카운트를 결정할 수 있다. 카운트 점수는 판독물 카운트에 기초할 수 있다. 일부 실시양태에서, 카운트 점수는 판독물 카운트와 동일한 값일 수 있다. 다른 실시양태에서, 카운트 점수는 판독물 카운트 및 다른 정보에 기초할 수 있다. 예를 들어, 카운트 점수는 유전자좌의 특정한 대립유전자에 대한 판독물 카운트 및 유전자좌에 대한 판독물의 총수에 기초할 수 있다. 일부 실시양태에서, 카운트 점수는 유전자좌에 대한 판독물 카운트 및 사전 수득된 데이터에 기초할 수 있다. 일부 실시양태에서, 카운트 점수는 미리 결정된 값들 사이의 정규화된 점수일 수 있다. 카운트 점수는 또한 샘플의 다른 유전자좌로부터의 판독물 카운트의 함수 또는 관심 샘플과 공동으로 실행되는 다른 샘플로부터의 판독물 카운트의 함수일 수 있다. 예를 들어, 카운트 점수는 특정한 대립유전자의 판독물 카운트 및 샘플에서의 다른 유전자좌의 판독물 카운트 및/또는 다른 샘플로부터의 판독물 카운트의 함수일 수 있다. 한 예로서, 다른 유전자좌로부터의 판독물 카운트 및/또는 다른 샘플로부터의 판독물 카운트를 사용하여 특정한 대립유전자에 대한 카운트 점수를 정규화할 수 있다.

[0029] 용어 "커버리지" 또는 "단편 커버리지"는 서열의 동일한 단편에 대한 다수의 샘플 판독물의 카운트 또는 다른 측정값을 지칭한다. 판독물 카운트는 상응하는 단편을 커버하는 판독물의 수의 카운트를 나타낼 수 있다. 대안적으로, 커버리지는 판독물 카운트를 역사적 지식, 샘플 지식, 유전자좌 지식 등에 기초한 지정된 인자로 곱

함으로써 결정될 수 있다.

[0030] 요약

[0031] 본원의 실시양태에 따르면, 변이체 호출을 검증하기 위한 컴퓨터 구현 방법이 제공된다. 상기 방법은 하기에 대한 프로그램 명령어를 실행하는 1개 이상의 프로세서의 제어 하에, 관심 게놈 서열을 따라 상응하는 뉴클레오티드 서열을 갖는 샘플 판독물을 포함하는 서열분석 데이터를 수신하는 단계; 관심 게놈 서열을 따라 뉴클레오티드 서열 내의 지정된 위치에서 잠재적 변이체 호출의 표시를 수신하는 단계; 및 하나 이상의 기준선 게놈 서열 내의 지정된 위치에서 기준선 변이체 빈도를 수득하는 단계로 작동한다. 상기 방법은 관심 게놈 서열에 대해 지정된 위치에서 샘플 변이체 빈도를 수득한다. 상기 방법은 지정된 위치에서 기준선 및 샘플 변이체 빈도를 분석하여 품질 점수를 수득하고; 품질 점수를 기초로 하여 관심 게놈 서열에 대한 잠재적 변이체 호출을 검증한다.

[0032] 임의로, 상기 분석 작업은 샘플 변이체 빈도와 기준선 변이체 빈도의 분포 사이의 관계를 수득하는 것을 포함하며, 상기 품질 점수는 관계에 기초한다. 임의로, 상기 분석 작업은 기준선 변이체 빈도의 분포와 관련하여 샘플 변이체 빈도를 인덱싱하는 것을 포함한다. 관계는 비-파라미터 윌콕슨 순위 합계 검정에 기초할 수 있다. 기준선 변이체 빈도는 기준선 게놈 서열을 따라 상응하는 위치에서의 배경 노이즈의 정도를 표시한다.

[0033] 임의로, 검증하는 단계는 품질 점수를 한계값과 비교하는 단계; 및 품질 점수가 한계값을 초과할 때 잠재적 변이체 호출을 유효한 변이체 호출인 것으로 판정하는 단계를 추가로 포함한다. 기준선 변이체 빈도는 하나 초과의 유형의 대립유전자와 연관된 다수의 기준선 게놈 서열로부터 유래될 수 있다. 임의로, 상기 방법은 기준선 게놈 서열을 따라 뉴클레오티드 서열의 복수의 참조 판독물을 포함하는 서열분석 데이터를 수신하는 단계, 및 지정된 위치에서 참조 판독물에 대한 기준선 변이체 빈도를 결정하는 단계를 추가로 포함한다. 기준선 변이체 빈도를 결정하는 단계는 참조 판독물을 따라 현재 염기 쌍 윈도우 내의 일련의 위치들에 대한 서열분석 데이터를 수신하는 단계; 현재 염기 쌍 윈도우 내의 위치 세트 내의 1개 이상의 위치에 대한 후보 변이체 빈도를 확인하는 단계; 참조 판독물 내의 지정된 위치에 대한 기준선 변이체 빈도로서 후보 변이체 빈도 중 하나를 선택하는 단계; 염기 쌍 윈도우를 기준선 게놈 서열을 따라 이동시키고, 작업을 반복하는 단계를 추가로 포함할 수 있다.

[0034] 한 실시양태에 따르면, 변이체 호출을 검증하기 위한 컴퓨터 구현 방법이 제공된다. 상기 방법은 하기에 대한 프로그램 명령어를 실행하는 1개 이상의 프로세서의 제어 하에, 관심 게놈 서열에 대한 잠재적 변이체 호출의 표시를 수신하는 단계; 관심 게놈 서열을 따라 샘플 및 미가공 뉴클레오티드 단편에 대한 판독물을 포함하는 서열분석 데이터를 수신하는 단계이며, 상기 판독물은 관심 게놈 서열을 따라 지정된 위치에서 뉴클레오티드 서열에 상응하는 샘플 단편에 대한 샘플 판독물을 포함하는 것인 단계로 작동한다. 상기 방법은 기준선 게놈 서열에 대한 지정된 위치에서의 기준선 변이체 빈도와 관련하여 관심 게놈 서열에 대한 지정된 위치에서의 샘플 변이체 빈도를 분석하여 품질 점수를 수득하고; 하기: A) 미가공 단편이 잠재적 변이체 호출을 확인하는지 여부; B) 샘플 판독물이 샘플 단편에 대한 미리 결정된 양의 커버리지를 제공하는지 여부; 및 C) 잠재적 변이체 호출이 미리 규정된 집단에 걸쳐 나타난 이전 변이체 호출과 매칭되는지 여부 중 적어도 하나를 결정한다. 상기 방법은 품질 점수 및 결정 작업에 기초하여 잠재적 변이체 호출을 검증한다.

[0035] 임의로, 결정하는 단계는 작업 A)를 포함할 수 있고, 작업 A)는 지지 변이체 호출을 표시하는 미가공 단편을 식별하는 단계; 지지 변이체 호출을 표시하는 미가공 단편에 대해 가중 단편 점수를 수득하는 단계; 및 가중 단편 점수를 미가공 단편 한계값과 비교하여 잠재적 변이체 호출을 확인하는지 여부를 결정하는 단계를 포함한다.

[0036] 임의로, 결정하는 단계는 작업 A)를 포함할 수 있고, 여기서 미가공 단편은 듀플렉스 스티칭된 단편, 심플렉스 스티칭된 단편, 듀플렉스 비-스티칭된 단편, 또는 심플렉스 비-스티칭된 단편 중 적어도 하나에 상응한다. 임의로, 결정하는 단계는 작업 A)를 포함할 수 있고, 작업 A)는 듀플렉스 스티칭된 단편을 미가공 단편으로서 식별하는 단계; 및 듀플렉스 스티칭된 단편이 잠재적 변이체 호출을 확인하는지 여부를 결정하기 위한 지지 변이체 호출을 표시하는지 여부를 결정하는 단계를 포함할 수 있다. 임의로, 결정하는 작업은 작업 B)를 포함할 수 있고, 샘플 판독물이 미리 결정된 양의 커버리지 미만으로 떨어질 때 상기 잠재적 변이체 호출은 호출-부재인 것으로 판정된다. 임의로, 결정하는 작업은 작업 C)를 포함할 수 있다. 잠재적 및 이전 변이체 호출 사이에 매치가 결정되는 경우, 결정하는 작업은 제1 한계값을 이용하여 작업 A) 또는 작업 B) 중 적어도 하나를 수행한다. 잠재적 및 이전 변이체 호출 사이에 매치가 존재하지 않는 경우, 결정하는 작업은 제2 한계값을 이용하여 작업 A) 또는 작업 B) 중 적어도 하나를 수행한다.

[0037] 임의로, 방법은 분석 작업 및 결정 작업 A), B) 및 C)를 계층식으로 수행할 수 있다.

[0038] 본원의 실시양태에 따르면, 변이체 호출을 검증하기 위한 시스템이 제공된다. 상기 시스템은 프로그램 명령어 및 관심 게놈 서열을 따라 상응하는 뉴클레오티드 서열을 갖는 샘플 판독물을 포함하는 서열분석 데이터를 저장하는 메모리 및 프로그램 명령어를 실행하는 1개 이상의 프로세서를 포함한다. 프로세서는 관심 게놈 서열을 따라 뉴클레오티드 서열 내의 지정된 위치에서 잠재적 변이체 호출의 표시를 수신하고; 하나 이상의 기준선 게놈 서열 내의 지정된 위치에서 기준선 변이체 빈도를 수득한다. 프로세서는 관심 게놈 서열에 대한 지정된 위치에서 샘플 변이체 빈도를 결정하고; 지정된 위치에서 기준선 및 샘플 변이체 빈도를 분석하여 품질 점수를 수득한다. 프로세서는 품질 점수를 기초로 하여 관심 게놈 서열에 대한 잠재적 변이체 호출을 검증한다.

[0039] 임의로, 프로세서는 샘플 변이체 빈도와 기준선 변이체 빈도의 분포 사이의 관계를 수득할 수 있으며, 상기 품질 점수는 관계에 기초한다. 임의로, 프로세서는 기준선 변이체 빈도의 분포와 관련하여 샘플 변이체 빈도를 인덱싱한다. 임의로, 기준선 변이체 빈도는 기준선 게놈 서열을 따라 상응하는 유전자좌에서의 배경 노이즈의 정도를 표시한다. 임의로, 프로세서는 품질 점수를 한계값과 비교하고; 비교에 기초하여 잠재적 변이체 호출을 무효인 것으로 판정하는 표시를 출력한다. 임의로, 시스템은 기준선 게놈 서열을 따라 뉴클레오티드의 복수의 참조 판독물을 포함하는 서열분석 데이터를 저장하는 메모리, 지정된 위치에서 참조 판독물에 대해 기준선 변이체 빈도를 결정하기 위한 1개 이상의 프로세서를 추가로 포함한다.

[0040] 본원의 실시양태에 따르면, 변이체 호출을 검증하기 위한 시스템이 제공된다. 시스템은, 프로그램 명령어 및 관심 게놈 서열을 따라 샘플 및 미가공 뉴클레오티드 단편에 대한 판독물을 포함하는 서열분석 데이터를 저장하는 메모리이며, 상기 판독물은 관심 게놈 서열을 따라 지정된 위치에서 뉴클레오티드 서열에 상응하는 샘플 단편에 대한 샘플 판독물을 포함하는 것인 메모리; 및 프로그램 명령어를 실행하는 1개 이상의 프로세서를 포함한다. 프로세서는 관심 게놈 서열에 대한 잠재적 변이체 호출의 표시를 수신하고; 기준선 게놈 서열에 대한 지정된 위치에서의 기준선 변이체 빈도와 관련하여 관심 게놈 서열에 대한 지정된 위치에서의 샘플 변이체 빈도를 분석하여 품질 점수를 수득한다. 프로세서는 하기: A) 미가공 단편이 잠재적 변이체 호출을 확인하는지 여부; B) 샘플 판독물이 샘플 단편에 대한 미리 결정된 양의 커버리지를 제공하는지 여부; 및 C) 잠재적 변이체 호출이 미리 규정된 집단에 걸쳐 나타난 이전 변이체 호출과 매칭되는지 여부 중 적어도 하나를 결정한다. 프로세서는 품질 점수 및 결정 작업에 기초하여 잠재적 변이체 호출을 검증한다.

[0041] 임의로, 프로세서는 지지 변이체 호출을 표시하는 미가공 단편을 식별하고; 지지 변이체 호출을 표시하는 미가공 단편에 대해 가중 단편 점수를 수득하고; 가중 단편 점수를 미가공 단편 한계값과 비교하여 잠재적 변이체 호출을 확인하는지 여부를 결정함으로써 미가공 단편이 잠재적 변이체 호출을 확인하는지 여부를 결정할 수 있다. 미가공 단편은 듀플렉스 스티칭된 단편, 심플렉스 스티칭된 단편, 듀플렉스 비-스티칭된 단편, 또는 심플렉스 비-스티칭된 단편 중 적어도 하나에 상응할 수 있다.

[0042] 본원의 실시양태에 따르면, 시스템적 오류로부터의 거짓 양성 변이체 호출을 감소시키기 위한 시스템 및 방법이 기재된다. 시스템적 오류는 FFPE 아티팩트, 서열분석 오류, 라이브러리 제조 오류, PCR 오류 등과 같은 다양한 인자로 인해 발생할 수 있다. 변이체 호출은 NGS 기반 검정에 의해 서열분석된 다양한 조직으로부터의 다양한 DNA 품질을 갖는 FFPE 정상 샘플의 패널로부터 컴파일링될 수 있는 유전자좌 특이적 배경 오류 분포에 통계적으로 적용된다. FFPE 정상 샘플의 동일한 서열분석 데이터는 또한 PCR, DNA 품질, 프로브 풀-다운 효율, 또는 서열 GC 함량에 의해 야기되는 판독 커버리지에서 시스템적 바이어스를 정규화하여 시험 샘플에서 진정한 카피수 변경을 밝혀내는데 사용될 수 있다. CNV 호출에서 신호 대 노이즈 비를 추가로 확대하기 위해, 추가적인 인핸서 프로브를 하이브리드 포획에 부가하여 유전자 증폭의 강건한 추정을 제공할 수 있다. 본원의 실시양태에 따르면, 노이즈 문제를 해결하고 시스템적 오류가 거짓 양성 변이체 호출에 기여하는 것을 방지하는 방법 및 시스템이 기재된다. 이와 관련하여, 시스템이 높은 배경 노이즈를 갖는 영역에서 종양 샘플 내 호출 엄격성을 증가시키도록 시스템적 바이어스를 식별하기 위해 정상 샘플 세트가 사용된다. FFPE 샘플의 경우, 정상 FFPE 샘플을 사용하여 기준선을 구축할 수 있다. ctDNA 샘플의 경우, 정상 게놈 DNA 데이터를 사용하여 기준선을 구축할 수 있다. 본원에 기재된 통상적인 접근법의 문제점은, 종양 샘플의 변이체 호출을 시도하려고 하는 경우와 같이 낮은 빈도 분산을 나타내는 샘플에서 특히 문제가 된다.

[0043] 본원의 실시양태에 따르면, 거짓 양성 체세포 변이체 호출을 감소시키는 문제를 해결하는 체세포 변이체에 대한 후속 프로세싱 방법 및 시스템이 제공된다. 상기 방법 및 시스템은 정상 샘플로부터 변이체 빈도 기준선을 구축하여 상이한 배경 노이즈 수준을 갖는 게놈 영역에서 변이체 호출 신뢰를 조정한다. 노이즈로부터 신호를 분리하는 것은 노이지 FFPE 데이터에서의 체세포 변이체의 검출 및 무세포 DNA 데이터에서의 매우 낮은 빈도의 종

양 변이체 발견 둘 다를 위한 도전이다. 시스템적 거짓 양성은 소규모 변이체 호출에서 흔히 발견된다(도 1). 시스템적 거짓 양성은 라이브러리 제조 및 서열분석을 포함하는 실험 단계 동안 도입될 수 있다. 또한, 정렬 오류에 민감한 인간 게놈 내의 영역, 예컨대 반복 영역 또는 서열 복잡성이 낮은 영역이 존재한다.

도면의 간단한 설명

[0044] 본 출원은 방법을 수행하기 위한 다양한 방법 및 시스템을 기재한다. 방법의 적어도 일부가 복수의 작업으로서 도면에 도시되어 있다. 그러나, 실시양태는 도면에 도시된 작업으로 제한되지 않는 것으로 이해되어야 한다. 작업은 생략될 수 있고/거나, 작업은 변형될 수 있고/거나, 다른 작업이 첨가될 수 있다. 또한, 본원에 기재된 작업은 조합될 수 있거나, 작업은 동시에 수행될 수 있거나, 작업은 공동으로 수행될 수 있거나, 작업은 다수의 하위-작업으로 분할될 수 있거나, 작업은 상이한 순서로 수행될 수 있거나, 또는 작업(또는 일련의 작업)은 반복적 방식으로 재-수행될 수 있다. 또한, 상이한 방법들이 본원에서 제시되지만, 상이한 방법들(또는 상이한 방법의 작업들)은 다른 실시양태들에서 조합될 수 있는 것으로 이해되어야 한다.

도 1a는 본원의 실시양태에 따라 샘플을 분석하는 방법을 도시한다.

도 1b는 본원의 실시양태에 따라 서열분석 데이터를 분석하여 잠재적 변이체 호출을 식별하는 방법에 대한 흐름도를 도시한다.

도 1c는 본원의 실시양태에 따라 변이체 호출을 검증하기 위한 컴퓨터 구현 방법을 도시한다.

도 1d는 본원의 실시양태에 따라 기준선 변이체 빈도를 수득하는 프로세스를 도시한다.

도 2a는 본원의 실시양태에 따라 이용되는 기준선 집단 내의 상이한 개체에 대한 예시적인 패널을 도시한다.

도 2b는 도 2a의 패널 중 하나의 작은 부분에 대한 확대를 나타내는 염기 쌍 패널을 도시한다.

도 3은 본원의 실시양태에 따라 정상 기준선 게놈 서열을 이용하는 배경 보정의 원리와 관련된 그래프를 도시한다.

도 4는 본원의 실시양태에 따라 구현된 변이체 호출을 필터링하기 위한 계층-기반 결정 트리를 도시한다.

도 5a는 본원의 대안적인 실시양태에 따라 구현된 변이체 호출을 필터링하기 위한 계층-기반 결정 트리를 도시한다.

도 5b는 본원의 실시양태에 따라 가중 점수를 이용하는 것과 관련하여 수행되는 프로세스를 도시한다.

도 6은 본원의 실시양태에 따라 형성된 컴퓨팅 시스템 환경의 블록 다이어그램을 도시한다.

도 7은 본원의 실시양태에 따라 형성된 대안적 시스템의 블록 다이어그램을 도시한다.

발명을 실시하기 위한 구체적인 내용

[0045] 다양한 실시양태의 상세한 설명은 첨부된 도면과 함께 읽을 때 더 잘 이해될 것이다. 도면이 다양한 실시양태의 기능 블록의 다이어그램을 도시한다는 점에서, 기능 블록은 반드시 하드웨어 회로 사이의 분할을 표시할 필요는 없다. 따라서, 예를 들어, 기능 블록(예를 들어, 모듈, 프로세서, 또는 메모리) 중 하나 이상은 단일 피스의 하드웨어(예를 들어, 일반적 목적 신호 프로세서 또는 랜덤 액세스 메모리의 블록, 하드 디스크 등) 또는 다수의 피스의 하드웨어에서 구현될 수 있다. 유사하게, 프로그램은 독립형 프로그램일 수 있고, 운영 시스템 내 서브루틴으로서 포함될 수 있고, 설치된 소프트웨어 패키지 내의 기능일 수 있고, 기타이다. 다양한 측면은 도면에 제시된 배열 및 수단에 제한되지 않는 것으로 이해되어야 한다.

[0046] 본원에 제시된 실시양태는 서열 변이를 식별하기 위해 핵산 서열을 분석하는데 적용될 수 있다. 실시양태는 유전자 위치/유전자좌의 잠재적 변이체/대립유전자를 분석하고 유전자좌의 유전자형을 결정하거나, 다시 말해서 유전자좌에 대한 유전자형 호출을 제공하는데 사용될 수 있다. 예를 들어, 핵산 서열은 미국 특허 출원 공개 2016/0085910 및 미국 특허 출원 공개 2013/0296175에 기재된 방법 및 시스템에 따라 분석될 수 있으며, 이들의 완전한 대상은 그 전문이 본원에 명백하게 참조로 포함된다.

[0047] 본원에 기재된 방법은 본원에 기재된 바와 같은 메모리에 저장된 프로그램 명령어를 실행하는 다양한 시스템 및 디바이스의 1개 이상의 프로세서에 의해 구현될 수 있다. 달리 명시되지 않는 한, 본원의 방법에서의 작업 순서는 다양할 수 있다. 또한, 본원에 기재된 방법에서 작업 중 하나 이상은 완전히 생략될 수 있고, 추가의 작

업으로 보충될 수 있다.

[0048] 서열분석 프로세스

[0049] 도 1a는 한 실시양태에 따른 방법(100)을 도시한다. 방법(100)은, (102)에서 핵산, 예컨대 DNA를 포함하거나 또는 포함하는 것으로 의심되는 샘플을 수신하는 것을 포함한다. 샘플은 공지된 또는 미지의 공급원, 예컨대 동물(특히, 인간), 식물, 박테리아 또는 진균으로부터의 것일 수 있다. 샘플은 공급원으로부터 직접 채취될 수 있다. 예를 들어, 혈액 또는 타액은 개체로부터 직접 채취될 수 있다. 대안적으로, 샘플은 공급원으로부터 직접 수득되지 않을 수 있다. (104)에서, 1개 이상의 프로세서는 시스템이 서열분석을 위한 샘플을 제조하도록 지시한다. 제조(104)는 이물 물질을 제거하고/거나, 특정 물질(예를 들어, DNA)을 단리하는 것을 포함할 수 있다. 생물학적 샘플은 특정한 검정에 대한 특색을 포함하도록 제조될 수 있다. 예를 들어, 생물학적 샘플은 합성에-의한-서열분석(SBS)을 위해 제조될 수 있다. 특정 실시양태에서, 제조는 게놈의 특정 영역의 증폭을 포함할 수 있다. 예를 들어, (104)에서 제조는 STR 및/또는 SNP를 포함하는 것으로 공지된 미리 결정된 유전자좌를 증폭시키는 것을 포함할 수 있다. 유전자좌는 미리 결정된 프라이머 서열을 사용하여 증폭될 수 있다.

[0050] (106)에서, 1개 이상의 프로세서는 시스템이 샘플을 서열분석하도록 지시한다. 서열분석은 다양한 공지된 서열분석 프로토콜을 통해 수행될 수 있다. 특정한 실시양태에서, 서열분석은 SBS를 포함한다. SBS에서, 복수의 형광-표지된 뉴클레오티드는 광학 기판의 표면(예를 들어, 유동 셀 내의 채널을 적어도 부분적으로 규정하는 표면) 상에 존재하는 증폭된 DNA의 복수의 클러스터(아마도 수백만개의 클러스터)를 서열분석하기 위해 사용된다. 유동 셀은 서열분석을 위한 핵산 샘플을 함유할 수 있고, 여기서 유동 셀은 적절한 유동 셀 홀더 내에 위치한다.

[0051] 핵산은 미지의 표적 서열에 인접한 공지된 프라이머 서열을 포함하도록 제조될 수 있다. 제1 SBS 서열분석 사이클을 개시하기 위해, 하나 이상의 상이하게 표지된 뉴클레오티드, 및 DNA 폴리머라제 등은 유체 유동 하위시스템에 의해 유동 셀 내로/이를 통해 유동될 수 있다(제시되지 않음). 단일 유형의 뉴클레오티드가 한번에 첨가될 수 있거나, 또는 서열분석 절차에 사용되는 뉴클레오티드가 가역적 종결 특성을 보유하도록 특별히 설계될 수 있고, 따라서 서열분석 반응의 각각의 사이클이 여러 유형의 표지된 뉴클레오티드(예를 들어, A, C, T, G)의 존재 하에 동시에 일어나게 할 수 있다. 뉴클레오티드는 검출가능한 표지 모이어티, 예컨대 형광단을 포함할 수 있다. 4개의 뉴클레오티드가 함께 혼합된 경우, 폴리머라제는 혼입되는 정확한 염기를 선택할 수 있고, 각각의 서열은 단일 염기에 의해 연장된다. 비-혼입된 뉴클레오티드는 유동 셀을 통해 세척 용액을 유동시킴으로써 세척 제거될 수 있다. 하나 이상의 레이저는 핵산을 여기시키고 형광을 유도할 수 있다. 핵산으로부터 방출된 형광은 혼입된 염기의 형광단에 기초하고, 상이한 형광단은 상이한 파장의 방출 광을 방출할 수 있다. 디블록킹 시약을 유동 셀에 첨가하여 연장되고 검출된 DNA 가닥으로부터 가역적 종결인자 기를 제거할 수 있다. 이어서, 디블록킹 시약은 유동 셀을 통해 세척 용액을 유동시킴으로써 세척 제거될 수 있다. 이어서, 유동 셀은 상기 제시된 바와 같은 표지된 뉴클레오티드의 도입으로 시작하여 추가의 서열분석 사이클을 위해 준비된다. 유체 및 검출 작업을 수회 반복하여 서열분석 실행을 완료할 수 있다. 예시적인 서열분석 방법은 예를 들어 문헌 [Bentley et al., Nature 456:53-59 (2008)], 국제 공개 번호 WO 04/018497; 미국 특허 번호 7,057,026; 국제 공개 번호 WO 91/06678; 국제 공개 번호 WO 07/123744; 미국 특허 번호 7,329,492; 미국 특허 번호 7,211,414; 미국 특허 번호 7,315,019; 미국 특허 번호 7,405,281, 및 미국 공개 번호 2008/0108082에 기재되어 있고, 이들 각각은 본원에 참조로 포함된다.

[0052] 일부 실시양태에서, 핵산은 표면에 부착되고 서열분석 전 또는 서열분석 동안 증폭될 수 있다. 예를 들어, 가교 증폭을 사용하여 증폭을 수행하여 표면 상에 핵산 클러스터를 형성할 수 있다. 유용한 가교 증폭 방법은 예를 들어, 미국 특허 번호 5,641,658; 미국 특허 공개 번호 2002/0055100; 미국 특허 번호 7,115,400; 미국 특허 공개 번호 2004/0096853; 미국 특허 공개 번호 2004/0002090; 미국 특허 공개 번호 2007/0128624; 및 미국 특허 공개 번호 2008/0009420에 기재되어 있고, 이들 각각은 그 전문이 본원에 참조로 포함된다. 표면 상의 핵산을 증폭시키기 위한 또 다른 유용한 방법은 예를 들어 이들 각각이 본원에 참조로 포함된 문헌 [Lizardi et al., Nat. Genet. 19:225-232 (1998)] 및 미국 특허 공개 번호 2007/0099208 A1에 기재된 바와 같은 롤링 써클 증폭(RCA)이다.

[0053] 하나의 예시적인 SBS 프로토콜은, 예를 들어 각각 본원에 참조로 포함되는 국제 특허 공개 번호 WO 04/018497, 미국 특허 공개 번호 2007/0166705A1, 및 미국 특허 번호 7,057,026에 기재된 바와 같은, 제거가능한 3' 블록을 갖는 변형된 뉴클레오티드를 이용한다. 예를 들어, 반복 사이클의 SBS 시약이, 예를 들어 가교 증폭 프로토콜의 결과로서 표적 핵산이 부착된 유동 셀로 전달될 수 있다. 핵산 클러스터는 선형화 용액을 사용하여 단일 가

닥 형태로 전환될 수 있다. 선형화 용액은 예를 들어 각 클러스터의 한 가닥을 절단할 수 있는 제한 엔도뉴클레아제를 함유할 수 있다. 다른 절단 방법이 제한 효소 또는 닉킹 효소에 대한 대안으로서 사용될 수 있고, 특히 화학적 절단(예를 들어, 퍼아이오데이트에 의한 디올 연결의 절단), 엔도뉴클레아제에 의한 절단에 의한(예를 들어, 미국 매사추세츠주 입시치 소재의 NEB에 의해 공급되는 바와 같은 'USER', 파트 번호 M5505S), 열 또는 알칼리에의 노출에 의한 무염기 부위의 절단, 다른 방식으로 데옥시리보뉴클레오티드로 구성된 증폭 생성물에 혼입된 리보뉴클레오티드의 절단, 광화학적 절단 또는 펩티드 링커의 절단을 포함한다. 선형화 작업 후에, 서열분석 프라이머가 서열분석하고자 하는 표적 핵산에 혼성화하는 조건 하에 서열분석 프라이머를 유동 셀에 전달할 수 있다.

[0054] 이어서, 유동 셀을, 단일 뉴클레오티드 첨가에 의해 각각의 표적 핵산에 혼성화된 프라이머를 연장시키는 조건 하에, 제거가능한 3' 블록 및 형광 표지를 갖는 변형된 뉴클레오티드를 갖는 SBS 연장 시약과 접촉시킬 수 있다. 일단 변형된 뉴클레오티드가 서열분석될 주형 영역에 상보적인 성장하는 폴리뉴클레오티드 쇄 내로 혼입되면 추가의 서열 연장을 지시하는데 이용가능한 유리 3'-OH 기가 존재하지 않고, 따라서 폴리머라제는 추가의 뉴클레오티드를 부가할 수 없기 때문에, 단일 뉴클레오티드만이 각각의 프라이머에 부가된다. SBS 연장 시약을 제거하고, 샘플을 방사선에 의한 여기 하에 보호하는 성분을 함유하는 스캔 시약으로 대체할 수 있다. 스캔 시약에 대한 예시적인 성분은 미국 공개 US 2008/0280773 A1 및 미국 일련 번호 13/018,255에 기재되어 있고, 이들 각각은 본원에 참조로 포함된다. 이어서, 스캔 시약의 존재 하에 연장된 핵산을 형광으로 검출할 수 있다. 일단 형광이 검출되면, 사용된 블록 기에 적절한 디블록킹 시약을 사용하여 3' 블록을 제거할 수 있다. 각각의 블록 기에 유용한 예시적인 디블록킹 시약은 각각 본원에 참조로 포함되는 WO0004018497, US 2007/0166705A1 및 미국 특허 번호 7,057,026에 기재되어 있다. 디블록킹 시약은 세척 제거되어, 이제 추가의 뉴클레오티드의 부가에 적격인 3' OH 기를 갖는 연장된 프라이머에 혼성화된 표적 핵산을 남길 수 있다. 따라서, 하나 이상의 작업 사이의 임의적 세척과 함께, 연장 시약, 스캔 시약 및 디블록킹 시약을 첨가하는 사이클은 목적하는 서열이 수득될 때까지 반복될 수 있다. 상기 사이클은 각각의 변형된 뉴클레오티드가 특정한 염기에 상응하는 것으로 공지된 상이한 표지가 부착되어 있는 경우에, 사이클당 단일 연장 시약 전달 작업을 사용하여 수행될 수 있다. 상이한 표지는 각각의 혼입 작업 동안 부가된 뉴클레오티드들 사이의 구별을 용이하게 한다. 대안적으로, 각각의 사이클은 연장 시약 전달의 개별 작업, 이어서 스캔 시약 전달 및 검출의 개별 작업을 포함할 수 있고, 이 경우에 뉴클레오티드 중 2개 이상은 동일한 표지를 가질 수 있고, 공지된 전달 순서에 기초하여 구별될 수 있다.

[0055] 서열분석 작업이 (106)에서 특정한 SBS 프로토콜과 관련하여 상기에 도시되었지만, 다양한 다른 분자 분석을 서열분석하기 위한 다른 프로토콜이 원하는 경우에 수행될 수 있다는 것을 이해할 것이다.

[0056] (108)에서, 시스템의 1개 이상의 프로세서는 (110)에서의 후속 분석을 위한 서열분석 데이터를 수신한다. 서열분석 데이터는 .BAMS 파일과 같은 다양한 방식으로 포맷될 수 있다. 서열분석 데이터는 예를 들어 다수의 샘플 판독물을 포함할 수 있다. 서열분석 데이터는 뉴클레오티드의 상응하는 샘플 서열을 갖는 복수의 샘플 판독물을 포함할 수 있다. 오직 하나의 샘플 판독물만이 논의되지만, 서열분석 데이터는 예를 들어 수백, 수천, 수십만 또는 수백만 개의 샘플 판독물을 포함할 수 있는 것으로 이해되어야 한다. 상이한 샘플 판독물은 상이한 수의 뉴클레오티드를 가질 수 있다. 예를 들어, 샘플 판독물은 10개 뉴클레오티드 내지 약 500개 뉴클레오티드 또는 그 초과의 범위일 수 있다. 샘플 판독물은 공급원(들)의 전체 게놈에 걸칠 수 있다. 한 예로서, 샘플 판독물은 미리 결정된 유전자좌, 예컨대 의심되는 STR 또는 의심되는 SNP를 갖는 유전자좌에 관한 것이다.

[0057] 각각의 샘플 판독물은 샘플 서열, 샘플 단편 또는 표적 서열로서 지칭될 수 있는 뉴클레오티드의 서열을 포함할 수 있다. 샘플 서열은 예를 들어 프라이머 서열, 플랭킹 서열, 및 표적 서열을 포함할 수 있다. 샘플 서열 내의 뉴클레오티드의 수는 30, 40, 50, 60, 70, 80, 90, 100개 또는 그 초과를 포함할 수 있다. 일부 실시양태에서, 1개 이상의 샘플 판독물(또는 샘플 서열)은 적어도 150개 뉴클레오티드, 200개 뉴클레오티드, 300개 뉴클레오티드, 400개 뉴클레오티드, 500개 뉴클레오티드, 또는 그 초과를 포함한다. 일부 실시양태에서, 샘플 판독물은 1000개 초과의 뉴클레오티드, 2000개 초과의 뉴클레오티드, 또는 그 초과를 포함할 수 있다. 샘플 판독물(또는 샘플 서열)은 하나의 말단 또는 둘 다의 말단에서 프라이머 서열을 포함할 수 있다.

[0058] (110)에서, 1개 이상의 프로세서는 서열분석 데이터를 분석하여, 잠재적 변이체 호출(들) 및 샘플 변이체 호출(들)의 샘플 변이체 빈도를 수득한다. (110)에서의 작업은 또한 변이체 호출 어플리케이션 또는 변이체 호출기로 지칭될 수 있다. 대안적 변이체 호출기가 본원의 실시양태에 따라 사용될 수 있고, 여기서 상이한 변이체 호출기는 수행될 서열분석 작업의 유형에 기초하여, 관심 샘플의 특색 등에 기초하여 사용될 수 있다. 변이체 호출 어플리케이션의 하나의 비제한적인 예는 일루미나 인크.(Illumina Inc.) (캘리포니아주 샌디에고)의 피스

세스(Pisces)™ 어플리케이션이다. 추가적으로 또는 대안적으로, (110)에서의 작업은 문헌 ["Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs" by Sanders et al., Bioformatics, 2012, July 15, Volume 28, No. 14, pages 1811-1817]에 기재된 변이체 호출 어플리케이션을 이용할 수 있고, 그의 완전한 대상은 그 전문이 본원에 참조로 포함된다.

[0059] 적어도 일부 실시양태에 따르면, (110)에서 변이체 호출 어플리케이션은 낮은 빈도 변이체에 대한 호출, 배선 호출 등을 제공한다. 비제한적인 예로서, (110)에서 변이체 호출 어플리케이션은 종양-단독 샘플 및/또는 종양-정상 쌍형성된 샘플에 대해 실행될 수 있다. (110)에서 변이체 호출 어플리케이션은 단일 뉴클레오티드 변이(SNV), 다수의 뉴클레오티드 변이(MNV), indel 등을 검색할 수 있다. 변이체 호출 어플리케이션은 변이체를 식별하고, 서열분석 또는 샘플 제조 오류로 인한 미스매치에 대한 필터링을 수행한다. 각각의 변이체에 대해, 변이체 호출기는 참조 서열, 변이체의 위치, 및 잠재적 변이체 서열(들) (예를 들어, A에서 C로의 SNV, 또는 AG에서 A로의 결실)을 식별한다. 변이체 호출 어플리케이션은 샘플 서열(또는 샘플 단편), 참조 서열/단편, 및 변이체가 존재하는 표시로서 변이체 호출을 식별한다. 도 1b와 관련하여 이하에서 설명되는 바와 같이, (110)에서 변이체 호출 어플리케이션은 미가공 단편을 식별할 수 있고, 미가공 단편의 지정, 잠재적 변이체 호출을 검증하는 미가공 단편의 수의 카운트, 지지 변이체가 발생한 미가공 단편 내의 위치 및 다른 관련 정보를 출력할 수 있다. 미가공 단편의 비제한적인 예는 듀플렉스 스티칭된 단편, 심플렉스 스티칭된 단편, 듀플렉스 비-스티칭된 단편, 및 심플렉스 비-스티칭된 단편을 포함한다.

[0060] 변이체 호출 어플리케이션은, 예컨대 .VCF 또는 .GVCF 파일과 같은 다양한 포맷으로 호출을 출력할 수 있다. 단지 예로서, 변이체 호출 어플리케이션은 MiSeq리포터(MiSeqReporter) 파이프라인에 포함될 수 있다(예를 들어, MiSeq® 서열분석기 기기에서 구현되는 경우). 임의로, 어플리케이션은 다양한 작업흐름으로 구현될 수 있다. 분석은, (110)에서, 원하는 정보를 수득하기 위해 단일 프로토콜 또는 지정된 방식으로 샘플 판독물을 분석하는 프로토콜의 조합을 포함할 수 있다.

[0061] (111)에서, 1개 이상의 프로세서는 잠재적 변이체 호출과 관련하여 검증 작업을 수행한다. 검증 작업은 도 1c-5b와 관련하여 하기에서 보다 상세하게 기재된다. 검증 작업은 이하에서 설명되는 바와 같이, 품질 점수 및/또는 계층 계단식 시험에 기초할 수 있다. 검증 작업이 잠재적 변이체 호출을 인증하거나 또는 확인하는 경우, 검증 작업은 변이체 호출 정보((110)에서 변이체 호출 어플리케이션으로부터)를 샘플 보고 생성기(112)로 전송한다. 대안적으로, 검증 작업이 잠재적 변이체 호출을 무효화하거나 또는 실격시키는 경우, 검증 작업은 상응하는 표시(예를 들어, 음의 지표, 호출 지표 없음, 무효 호출 지표)를 샘플 보고 생성기(112)로 전송한다. 또한, (111)에서의 검증 작업은 변이체 호출이 정확한지 또는 무효 호출 지정이 정확한지의 신뢰도와 관련된 신뢰 점수를 전송할 수 있다.

[0062] (112)에서, 1개 이상의 프로세서는 샘플 보고를 생성하고 저장한다. 샘플 보고는, 예를 들어, 샘플에 대한 복수의 유전자좌에 관한 정보를 포함할 수 있다. 예를 들어, 미리 결정된 세트의 유전자좌의 각각의 유전자좌에 대해, 샘플 보고는 유전자형 호출을 제공하거나; 유전자형 호출이 만들어질 수 없음을 표시하거나; 유전자형 호출의 확실성에 신뢰 점수를 제공하거나; 또는 하나 이상의 유전자좌와 관련된 검정에서 잠재적 문제점을 표시하는 것 중 적어도 하나일 수 있다. 샘플 보고는 또한 샘플을 제공한 개체의 성별을 표시할 수 있고/거나 샘플이 다수의 공급원을 포함한다는 것을 표시할 수 있다. 본원에 사용된 "샘플 보고"는 유전자좌 또는 유전자좌의 미리 결정된 세트의 디지털 데이터(예를 들어, 데이터 파일) 및/또는 유전자좌 또는 유전자좌 세트의 인쇄된 보고를 포함할 수 있다. 따라서, (112)에서의 생성 또는 제공은 데이터 파일을 생성하고/거나 샘플 보고를 인쇄하거나, 또는 샘플 보고를 디스플레이하는 것을 포함할 수 있다.

[0063] 샘플 보고는 변이체 호출이 결정되었지만 검증되지 않았음을 나타낼 수 있다. 변이체 호출이 무효인 것으로 결정될 때, 샘플 보고는 변이체 호출을 검증하지 않는 결정에 대한 기준에 관한 추가의 정보를 표시할 수 있다. 예를 들어, 보고에서 추가의 정보는 미가공 단편들의 설명 및 미가공 단편들이 변이체 호출을 지지하거나 그와 모순되는 정도(예를 들어, 카운트)를 포함할 수 있다. 추가적으로 또는 대안적으로, 보고에서 추가의 정보는 본원에 기재된 실시양태에 따라 수득된 품질 점수를 포함할 수 있다.

[0064] 변이체 호출 어플리케이션

[0065] 도 1b는 본원의 실시양태에 따른 잠재적 변이체 호출을 식별하기 위해 서열분석 데이터를 분석하기 위한 방법(114)에 대한 흐름도를 도시한다. 예로서, 도 1b의 작업은 도 1a에서 (110) 동안 수행될 수 있다. 도 1b의 작업은 미리 수행된 서열분석 작업에 대해 저장된 데이터에 따라 작동할 수 있는 것으로 이해된다. 추가적으로 또는 대안적으로, 도 1b의 작업은 서열분석 작업이 수행되는 동안 실시간으로 수행될 수 있다. (118)에서, 각

각의 샘플 판독물은 상응하는 유전자좌에 할당된다. 샘플 판독물은 샘플 판독물의 뉴클레오티드의 서열, 또는 다시 말해서 샘플 판독물 내의 뉴클레오티드의 순서 (예를 들어, A, C, G, T)를 기초로 하여 상응하는 유전자좌에 할당될 수 있다. 이러한 분석을 기초로, 샘플 판독물은 특정한 유전자좌의 가능한 변이체/대립유전자를 포함하는 것으로 지정될 수 있다. 샘플 판독물은 유전자좌의 가능한 변이체/대립유전자를 포함하는 것으로 지정된 다른 샘플 판독물과 함께 수집(또는 통합 또는 비닝)될 수 있다. 할당 작업은, (118)에서, 샘플 판독물이 특정한 유전자 위치/유전자좌와 연관될 가능성이 있는 것으로 식별된다는 점에서 또한 호출 작업으로 지칭될 수 있다. 샘플 판독물은 다른 샘플 판독물로부터 샘플 판독물을 구별시켜주는 뉴클레오티드의 1개 이상의 식별 서열(예를 들어, 프라이머 서열)을 위치시키기 위해 분석될 수 있다. 보다 구체적으로, 식별 서열(들)은 샘플 판독물을 다른 샘플 판독물로부터 특정한 유전자좌와 연관된 것으로서 식별할 수 있다.

[0066] 할당 작업은, (118)에서, 식별 서열의 일련의 n개의 뉴클레오티드를 분석하여, 식별 서열의 일련의 n개의 뉴클레오티드가 선택 서열 중 하나 이상과 효과적으로 매칭되는지를 결정하는 것을 포함할 수 있다. 특정 실시양태에서, 할당 작업은, (118)에서, 샘플 서열의 제1 n개의 뉴클레오티드를 분석하여 샘플 서열의 제1 n개의 뉴클레오티드가 선택 서열 중 하나 이상과 효과적으로 매칭되는지를 결정하는 것을 포함할 수 있다. 수 n은 다양한 값을 가질 수 있고, 이는 프로토콜로 프로그래밍되거나 또는 사용자에 의해 입력될 수 있다. 예를 들어, 수 n은 데이터베이스 내의 최단 선택 서열의 뉴클레오티드 수로서 정의될 수 있다. 수 n은 미리 결정된 수일 수 있다. 미리 결정된 수는 예를 들어, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 또는 30개 뉴클레오티드일 수 있다. 그러나, 다른 실시양태에서는 보다 적거나 또는 보다 많은 뉴클레오티드가 사용될 수 있다. 수 n은 또한 시스템의 사용자와 같은 개체에 의해 선택될 수 있다. 수 n은 하나 이상의 조건에 기초할 수 있다. 예를 들어, 수 n은 데이터베이스 내의 최단 프라이머 서열의 뉴클레오티드의 수 또는 지정된 수로서 정의될 수 있고, 어느 쪽이든 더 작은 수이다. 일부 실시양태에서, n에 대해 최소값, 예컨대 15가 사용될 수 있고, 15개 미만의 뉴클레오티드의 임의의 프라이머 서열이 예외로 지정될 수 있다.

[0067] 일부 경우에, 식별 서열의 일련의 n개의 뉴클레오티드는 선택 서열의 뉴클레오티드에 정확하게 매칭되지 않을 수 있다. 그럼에도 불구하고, 식별 서열이 선택 서열과 거의 동일한 경우에 식별 서열은 선택 서열에 효과적으로 매칭될 수 있다. 예를 들어, 식별 서열의 일련의 n개의 뉴클레오티드(예를 들어, 제1 n개의 뉴클레오티드)가 선택 서열과 지정된 수 이하의 미스매치(예를 들어, 3개) 및/또는 지정된 수의 시프트(예를 들어, 2개)로 매칭되는 경우, 샘플 판독물은 유전자좌에 대해 호출될 수 있다. 규칙은 각각의 미스매치 또는 시프트가 샘플 판독물과 프라이머 서열 사이의 차이로서 카운팅될 수 있도록 확립될 수 있다. 차이의 수가 지정된 수 미만인 경우, 샘플 판독물은 상응하는 유전자좌에 대해 호출될 수 있다(즉, 상응하는 유전자좌에 할당될 수 있다). 일부 실시양태에서, 샘플 판독물의 식별 서열과 유전자좌와 연관된 선택 서열 사이의 차이의 수를 기초로 하여 매칭 점수가 결정될 수 있다. 매칭 점수가 지정된 매칭 한계값을 통과하는 경우, 선택 서열에 상응하는 유전자좌는 샘플 판독물에 대한 잠재적 유전자좌로서 지정될 수 있다. 일부 실시양태에서, 샘플 판독물이 유전자좌에 대해 호출되는지 여부를 결정하기 위해 후속 분석을 수행할 수 있다.

[0068] 샘플 판독물이 데이터베이스 내의 선택 서열 중 하나와 효과적으로 매칭하는 경우(즉, 상기 기재된 바와 같이 정확하게 매칭되거나 거의 매칭되는 경우), 샘플 판독물은 선택 서열과 상관되는 유전자좌에 할당되거나 지정된다. 이는 유전자좌 호출 또는 잠정-유전자좌 호출로 지칭될 수 있고, 여기서 샘플 판독물은 선택 서열과 상관되는 유전자좌에 대해 호출된다. 그러나, 상기 논의된 바와 같이, 샘플 판독물은 하나 초과의 유전자좌에 대해 호출될 수 있다. 이러한 실시양태에서, 잠재적 유전자좌 중 오직 하나에 샘플 판독물을 호출하거나 할당하기 위해 추가의 분석을 수행할 수 있다. 일부 실시양태에서, 참조 서열의 데이터베이스와 비교되는 샘플 판독물은 쌍형성된-말단 서열분석으로부터의 제1 판독물이다. 쌍형성된-말단 서열분석을 수행할 때, 샘플 판독물과 상관된 제2 판독물(미가공 단편을 나타냄)이 수득된다. 할당 후에, (118)에서, 할당된 판독물을 사용하여 수행되는 후속 분석은 할당된 판독물에 호출된 유전자좌의 유형에 기초할 수 있다.

[0069] (120)에서, 샘플 판독물을 분석하여 잠재적 변이체 호출을 식별한다. 특히, 분석 결과는 잠재적 변이체 호출, 샘플 변이체 빈도, 참조 서열 및 변이체가 발생하는 관심 게놈 서열 내의 위치를 식별한다. 예를 들어, 유전자좌가 SNP를 포함하는 것으로 공지된 경우에, 할당된 판독물의 SNP를 식별하기 위해, 유전자좌에 대해 호출된 할당된 판독물은 (120)에서 분석을 거칠 수 있다. 유전자좌가 다형성 반복적 DNA 요소를 포함하는 것으로 공지된 경우에, 샘플 판독물 내의 다형성 반복적 DNA 요소를 식별하거나 특징화하기 위해, 할당된 판독물이 (120)에서 분석될 수 있다. 일부 실시양태에서, 할당된 판독물이 STR유전자좌 및 SNP유전자좌와 효과적으로 매칭하는 경우, 경고 또는 플래그가 샘플 판독물에 할당될 수 있다. 샘플 판독물은 STR유전자좌 및 SNP 유전자좌 둘 다로

서 지정될 수 있다. 분석하는 것은, (1202)에서, 할당된 판독물들을 정렬 프로토콜에 따라 정렬하여 할당된 판독물들의 서열 및/또는 길이를 결정하는 것을 포함할 수 있다. 정렬 프로토콜은 2013년 3월 15일자로 출원된 국제 출원 번호 PCT/US2013/030867 (공개 번호 WO 2014/142831)에 기재된 방법을 포함할 수 있고, 이는 그 전문이 본원에 참조로 포함된다.

[0070] (122)에서, 1개 이상의 프로세서는 미가공 단편을 분석하여 미가공 단편 내의 상응하는 위치에서 지지 변이체가 존재하는지 여부를 결정한다. 다양한 유형의 미가공 단편이 식별될 수 있다. 예를 들어, 변이체 호출기는 원래의 변이체 호출을 검증하는 변이체를 나타내는 미가공 단편의 유형을 식별할 수 있다. 예를 들어, 미가공 단편의 유형은 듀플렉스 스티칭된 단편, 심플렉스 스티칭된 단편, 듀플렉스 비-스티칭된 단편, 또는 심플렉스 비-스티칭된 단편을 나타낼 수 있다. 임의로, 상기 실시예 대신 또는 그에 더하여 다른 미가공 단편이 식별될 수 있다. 미가공 단편의 각각의 유형을 식별하는 것과 관련하여, 변이체 호출기는 또한 지지 변이체가 발생한 미가공 단편 내의 위치, 뿐만 아니라 지지 변이체를 나타내는 미가공 단편의 수의 카운트를 식별한다. 예를 들어, 변이체 호출기는 미가공 단편의 10개의 판독물이 특정한 위치 X에서 지지 변이체를 갖는 듀플렉스 스티칭된 단편을 나타내는 것으로 식별된다는 표시를 출력할 수 있다. 변이체 호출기는 또한 미가공 단편의 5개의 판독물이 특정한 위치 Y에서 지지 변이체를 갖는 심플렉스 비-스티칭된 단편을 나타내는 것으로 식별된다는 표시할 수 있다. 변이체 호출기는 또한, 참조 서열에 상응하여 따라서 관심 게놈 서열에서 잠재적 변이체 호출을 검증하는 증거를 달리 제공할 지지 변이체를 포함하지 않는, 다수의 미가공 단편을 출력할 수 있다.

[0071] (122)에서, 지지 변이체, 뿐만 아니라 지지 변이체가 발생한 위치를 포함한 미가공 단편의 카운트는 유지된다. 추가적으로 또는 대안적으로, 관심 위치에서 지지 변이체를 포함하지 않는 미가공 단편의 카운트는 (샘플 판독물 또는 샘플 단편 내의 잠재적 변이체 호출의 위치 대비) 유지될 수 있다. 추가적으로 또는 대안적으로, 참조 서열에 상응하고 잠재적 변이체 호출을 인증하거나 확인하지 않는 미가공 단편의 카운트가 유지될 수 있다. 잠재적 변이체 호출을 지지하는 미가공 단편의 카운트 및 유형, 미가공 단편에서의 지지 분산의 위치, 잠재적 변이체 호출을 지지하지 않는 미가공 단편의 카운트 등을 포함한 (122)에서 결정된 정보는 변이체 호출 검증 어플리케이션에 출력된다.

[0072] 잠재적 변이체 호출이 식별되면, 도 1b의 프로세스는 잠재적 변이체 호출, 변이체 서열, 변이체 위치 및 그와 연관된 참조 서열의 표시를 출력한다. 오류는 거짓 변이체를 식별하기 위해 호출 프로세스를 유발할 수 있으므로 변이체 호출은 "잠재적인" 변이체를 나타내는 것으로 지정된다. 본원의 실시양태에 따르면, 거짓 변이체 또는 거짓 양성을 감소시키고 제거하기 위해 잠재적 변이체 호출을 분석한다. 추가적으로 또는 대안적으로, 도 1b의 프로세스는 샘플 판독물과 연관된 하나 이상의 미가공 단편을 분석하고, 미가공 단편과 연관된 상응하는 변이체 호출을 출력한다.

[0073] 변이체 호출 검증 어플리케이션

[0074] 도 1c는 본원의 실시양태에 따른 변이체 호출을 검증하기 위한 컴퓨터 구현 방법을 도시한다. (130)에서, 1개 이상의 프로세서는 지정된 위치에서의 잠재적 변이체 호출의 표시, 관심 게놈 서열 내의 잠재적 변이체 호출의 위치를 포함한, 변이체 호출 어플리케이션으로부터의 데이터를 수신한다(도 1b). (130)에서, 1개 이상의 프로세서는 또한 잠재적 변이체 호출을 지지하는 미가공 단편의 카운트 및 유형, 미가공 단편 내의 지지 변이체의 위치, 잠재적 변이체 호출 및 다른 관련 정보를 지지하지 않는 미가공 단편의 카운트를 수신한다. (130)에서의 작업은 이전에 수행된 서열에 대한 저장된 데이터에 접근하는 하나 이상의 작업에 상응할 수 있는 것으로 이해된다. 추가적으로 또는 대안적으로, (130)에서의 작업은 서열분석 작업이 수행되는 동안 실시간으로 수행될 수 있다.

[0075] 임의로, (130)에서, 1개 이상의 프로세서는 참조 서열의 표시를 수신할 수 있고, 이에 대해 잠재적 변이체 호출이 이루어진다. 참조 서열은 하나 이상의 기준선 게놈 서열(들) 내의 뉴클레오티드의 서열에 상응한다.

[0076] (132)에서, 1개 이상의 프로세서는 하나 이상의 기준선 게놈 서열 내의 지정된 위치에서 하나 이상의 기준선 변이체 빈도를 수득한다(예를 들어, 도 1d의 프로세스에 따름). 한 예로서, 기준선 변이체 빈도는 하나 초과의 유형의 대립유전자와 연관된 다수의 기준선 게놈 서열로부터 유래된다. 임의로, 기준선 변이체 빈도는 하나의 유형의 대립유전자와 연관될 수 있지만, 이는 다수의 개체로부터의 샘플로부터 유래된 것이다. 기준선 변이체 빈도는 도 1c의 다른 작업과 동시에 결정될 수 있다. 대안적으로, 기준선 변이체 빈도는 관심 샘플에 대한 서열분석 작업 또는 도 1c의 검증 프로세스의 구현 전에 미리 결정될 수 있다. 기준선 변이체 빈도는 다양한 참조 기준선 게놈 서열을 따라 다양한 위치와 관련하여(예를 들어 라이브러리 또는 데이터베이스에서) 저장될 수 있다. 예를 들어, 라이브러리 또는 데이터베이스는 다수의 기준선 게놈 서열에 대해 유지될 수 있다. 라이브

러리 또는 데이터베이스(예를 들어, 도 7에서 (784))는 또한 기준선 게놈 서열을 따라 각각의 위치와 연관된 기준선 변이체 빈도를 보유할 수 있다. 예컨대 상이한 유형의 서열분석 방법, 제조 방법, 서열분석 장비 등과 관련하여, 기준선 게놈 서열이 상이한 유형의 기준선 변이체 빈도를 나타낼 수 있는 것으로 인식된다. 따라서, 라이브러리 또는 데이터베이스는 기준선 게놈 서열을 따라 임의의 주어진 위치에 대해 기준선 변이체 빈도의 세트를 유지할 수 있다. (132)에서의 작업은 라이브러리 또는 데이터베이스를 관리하는 네트워크 서버 또는 원격 컴퓨터(예를 들어, 780)에 요청을 송신하는 것을 포함할 수 있다. 요청은 관심 기준선 게놈 서열, 그의 지정된 위치, 뿐만 아니라 추가의 정보(예를 들어, 서열분석 작업의 유형, 사용되는 장비의 유형, 또는 라이브러리 제조 프로토콜)를 지정할 수 있다.

[0077] 기준선 변이체 빈도는 기준선 게놈 서열을 따라 1개 이상의 위치에서의 변이를 설명하는 관심있는 상이한 특징을 식별하기 위해 상이한 방식으로 결정될 수 있다. 예로서, 기준선 변이체 빈도는 기준선 게놈 서열을 따라 상응하는 위치에서의 배경 노이즈의 정도(또는 범위)를 나타낸다. 배경 노이즈의 정도/범위는, 예를 들어 서열분석 작업의 유형, 장비의 유형, 라이브러리 제조 프로세스 등에 따라 달라질 수 있다.

[0078] (134)에서, 1개 이상의 프로세서는 관심 게놈 서열에 대한 지정된 위치에서 샘플 변이체 빈도를 수득한다. 예를 들어, 프로세서는 변이체 호출 어플리케이션으로부터 샘플 변이체 빈도를 수득하고, 여기서 샘플 변이체 빈도는 샘플 집단에 걸쳐 관심 게놈 서열 내의 특정한 위치/유전자좌에서의 대립유전자(유전자의 변이체)의 상대적 빈도를 나타낸다. 예를 들어, 샘플 변이체 빈도는 분석된 개체로부터의 관심있는 모든 샘플(예를 들어 염색체)의 분획 또는 백분율로 표현될 수 있다.

[0079] (136)에서, 1개 이상의 프로세서는 지정된 위치에서 기준선 및 샘플 변이체 빈도를 분석하여 품질 점수를 수득한다. 예로서, 분석은 샘플 변이체 빈도와 기준선 변이체 빈도의 분포 사이의 관계를 수득하는 것을 포함할 수 있으며, 품질 점수는 관계에 기초하여 결정된다. 예를 들어, 다수의 기준선 변이체 빈도가 기준선 게놈 서열의 집단에 대해 저장될 수 있다. 다수의 기준선 변이체 빈도는 분포로 조직화된다. 프로세서는 기준선 변이체 빈도의 분포에 대해 샘플 변이체 빈도를 인덱싱할 수 있다. 한 예로서, 관계는 윌콕슨 순위 합계 검정과 같은 비-파라미터 검정에 기초할 수 있다. 윌콕슨 순위 합계 검정은 샘플 변이체 빈도와 기준선 변이체 빈도의 분포 사이의 관계를 나타내는 p-값을 생성한다. p-값은 샘플 변이체 호출이 노이즈에 기인한 것인지 또는 배경 노이즈에 기인하지 않은 것인지의 신뢰도의 수치 표시를 나타낸다.

[0080] 프로세서는 수치 p-값을 품질 점수(예를 들어 Q-점수로 지칭됨)로 전환시킨다. 예를 들어, 품질 점수는 단순히 p-값과 동일할 수 있다. 대안적으로, 품질 점수는 미리 결정된 수학적 오퍼레이터를 p-값에 적용(예를 들어 p-값을 정규화, p-값을 정수로 전환)함으로써 형성될 수 있다. 임의로, p-값은 다른 정보/인자와 조합되어 품질 점수를 형성할 수 있다. 예를 들어, p-값은 수득된 수많은 샘플 및/또는 샘플 판독물에 기초하여 변형될 수 있다. 임의로, 기준선 변이체 빈도와 샘플 변이체 빈도 사이의 관계를 결정하기 위해 대안적 검정이 적용될 수 있고, 여기서 이러한 대안적 검정은 샘플 변이체 호출에서의 신뢰도의 표시를 제공한다. 예를 들어, 신뢰도는 샘플 변이체 호출이 배경 노이즈에 기인한 것인지 또는 배경 노이즈에 기인하지 않은 것인지를 나타낼 수 있다. 대안적으로, 신뢰도는 샘플 변이체 호출이 예컨대 FFPE 아티팩트, 서열분석 오류 또는 PCR 오류로 인한 거짓 변이체를 유발하는 다양한 유형의 시스템적 오류를 포함한다는 것을 나타낼 수 있다.

[0081] (138)에서, 1개 이상의 프로세서는 품질 점수를 기초로 하여 관심 게놈 서열에 대한 잠재적 변이체 호출을 검증한다. 예를 들어, 검증 작업은 추가로 품질 점수를 한계값과 비교하고, 품질 점수가 한계값을 초과할 때 잠재적 변이체 호출이 유효한 변이체 호출인 것으로 판정하는 것을 포함할 수 있다. 한계값은 잠재적 거짓 변이체를 필터링하기 위한 Q-점수 컷-오프를 나타낸다. 유효한 변이체 호출은 호출이 정확한 "높은" 신뢰도가 존재하는 변이체 호출을 나타낼 수 있다. 높은 신뢰도 vs 낮은 신뢰도를 나타내는 것의 척도는 구현의 특정한 사실 및 상황에 기초하여 달라질 수 있는 상대적인 용어를 나타내는 것으로 인식된다. 한 예로서, 한계값은 실험적 분석에 기초하여 변경될 수 있다.

[0082] 도 1d는 본원의 실시양태에 따라 기준선 변이체 빈도를 수득하는 프로세스를 도시한다. 도 1d의 프로세스는 다양한 시스템 및 디바이스의 1개 이상의 프로세서에 의해 수행될 수 있다. 도 1d 의 작업 순서는 하나의 예를 나타내며, 작업은 대안적 순서로 수행될 수 있는 것으로 이해된다. 또한, 도 1d의 작업 중 하나 이상은 완전히 생략될 수 있다. 도 1d의 작업은 도 2와 관련하여 기재될 것이다.

[0083] NGS 기반 검정에 의해 서열분석된 다양한 조직으로부터의 다양한 DNA 품질을 갖는 FFPE 정상 샘플에 대한 복수의 패널로부터의 유전자좌 특이적 배경 오류 분포를 컴파일링하기 위해 도 1d의 방법을 주기적으로 반복할 수 있다. 도 1d의 방법은 FFPE 정상 샘플의 복수의 패널로부터 동일한 서열분석 데이터를 이용하여, PCR, DNA 품

질, 프로브 풀-다운 효율, 또는 서열 GC 함량에 의해 유발되는 판독물 커버리지에서의 시스템적 바이어스를 정규화하여, 시험 샘플 내의 변이체의 정확한 수를 밝혀낼 수 있다.

[0084] 도 2a는 기준선 집단 내의 상이한 개체에 대한 예시적인 패널(202-206)을 도시한다. 본 실시예에서, 기준선 집단은 정상/건강한 개체를 나타내지만, 기준선 집단은 관심 특정한 돌연변이, 예컨대 특정한 유형의 암 등을 나타내는 개체를 나타낼 수 있는 것으로 인식된다. 패널(202-206)의 수직축은 분석된 1000개의 염기 당 발견된 거짓 양성 또는 거짓 변이체 호출의 개수를 나타낸다. 각각의 패널(202-206)은 다수의 참조 판독물, 예를 들어 1000, 5000, 10,000개 등의 누적에 상응하는 것으로 이해된다. 수평축은 기준선 게놈 서열 내의 염색체의 전부 또는 일부를 나타낸다. 도 2의 예에서, 기준선 게놈 서열을 따라 19개의 염색체가 도시되었지만, 기준선 게놈 서열은 더 많거나 더 적은 염색체에 대해 유지될 수 있는 것으로 이해된다. 일련의 수직 막대가 수평축을 따라 지정되고, 각각의 막대는 상응하는 염색체를 따라 상응하는 위치 또는 부위에서의 거짓 변이체의 수에 상응하는 양만큼 상향 연장된다. 수직 막대는 연관된 염색체 내의 상응하는 위치 또는 좌표의 대략적인 위치에서 수평축을 따라 위치한다. 예를 들어, 수직 막대(221)는 제10 염색체가 1000개 염기 당 대략 25개의 거짓 변이체를 경험하였음을 나타내기 위해 제10 염색체와 관련하여 제시되고, 여기서 염색체 내의 위치 또는 좌표는 염색체의 중심보다 약간 아래이다. 또 다른 예로서, 수직 막대(223)는 제1 염색체가 제1 염색체의 중심보다 약간 위의 위치 또는 좌표에서 1000개 염기 당 대략 7개의 거짓 변이체를 경험하였음을 나타내기 위해 제1 염색체와 관련하여 제시된다.

[0085] 도 2b는 도 2a의 패널(202-206) 중 하나의 작은 부분에 대한 확대도를 나타내는 염기 쌍 패널을 도시한다. 예로서, 도 2b는 패널(202)의 염색체(16) 내의 염기 쌍 1-9에 대한 참조 판독물에 상응할 수 있다(도 2a). 염기 쌍 패널(208)은 분석된 1000개의 염기 당 발견되는 거짓 변이체 호출의 수를 나타내는 수직축을 포함한다. 도 2a와 관련하여 상기 설명된 바와 같이, 염기 쌍 패널(208)은 다수의 참조 판독물, 예를 들어 1000, 5000, 10,000개 등의 누적에 상응한다. 수평축은 기준선 게놈 서열 내의 단일 염색체와 연관된 짧은 염기 쌍의 서열을 나타낸다. 도 2b의 예에서, 기준선 게놈 서열을 따라 9개 염기 쌍이 도시되지만, 기준선 게놈 서열이 본 염색체 및 다른 염색체에 대해 더 많은 염기 쌍을 유지하는 것으로 이해된다. 일련의 수직 막대가 수평축을 따라 지정되고, 각각의 막대는 상응하는 염색체를 따라 상응하는 염기 쌍 위치 또는 부위에서의 거짓 변이체의 수에 상응하는 양만큼 상향 연장된다. 수직 막대는 연관된 염색체 내의 상응하는 염기 쌍의 위치에서 수평축을 따라 위치한다. 예를 들어, 수직 막대는 염기 쌍 #1이 1000개 염기 당 대략 10개의 거짓 변이체를 경험한다는 것을 표시하기 위해 염기 쌍 #2에서 제시된다. 또 다른 예로서, 수직 막대는 염기 쌍 #5가 1000개 염기당 대략 25개의 거짓 변이체를 경험한다는 것을 표시하기 위해 염기 쌍 #5와 관련하여 제시된다.

[0086] 도 1d로 돌아가면, (150)에서, 1개 이상의 프로세서는 하나 이상의 기준선 게놈 서열에 대한 복수의 참조 판독물에 대한 서열 데이터를 수신한다. 도 2a와 관련하여, 각각의 패널(202-206)이 1000개의 참조 판독물의 누적을 나타내는 것으로 가정되면, 시스템은 약 3000개의 참조 판독물을 수득할 것이다. (151)에서, 1개 이상의 프로세서는 복수의 참조 판독물로부터 참조 판독물 세트를 선택한다. 도 2a와 관련하여, 참조 판독물 세트는 WGDNA04 (패널 202)로 지정된 개체에 대해 염색체 16 내의 염기 쌍 1-9에 상응하여 선택될 수 있다. 도 2b는 염색체 16 내의 염기 쌍 1-9에 대한 예를 도시한다. 152에서, 시스템의 1개 이상의 프로세서는 참조 판독물의 현재 세트에서 현재 위치를 선택한다. 예를 들어, 현재 위치는 패널(208)을 따라 위치 #2에 중심을 맞출 수 있다. (154)에서, 시스템의 1개 이상의 프로세서는 참조 판독물 세트 내의 현재 위치를 염기 쌍 윈도우로 오버레이한다.

[0087] 도 2b에서, 브라켓(210)은 염기 쌍 윈도우의 예를 도시하기 위해 제시된다. 염기 쌍 윈도우의 길이는 예를 들어 1개의 염기 쌍으로부터 임의의 원하는 수의 다수의 염기 쌍까지 다양할 수 있다. 도 2b의 예에서, 염기 쌍 윈도우는 3개의 염기 쌍에 상응하도록 지정되었지만, 다른 길이가 이용될 수 있다. 따라서, 도 1d의 작업을 통한 본 반복에서, 염기 쌍 윈도우(210)는 위치 #1-#3을 포괄한다. 1개 초과의 염기 쌍을 포괄하는 염기 쌍 윈도우를 이용하는 것은 노이즈 위치가 함께 발생할 수 있는 상황을 고려한다. 예를 들어, 라이브러리 제조 또는 다른 인자가 뉴클레오티드 서열 내의 한 위치에서 노이즈를 발생시키는 경우, 동일한 인자는 하나의 위치에 인접한 1개 이상의 위치에서 노이즈를 발생시킬 수 있다. 염기 쌍 윈도우의 길이는 부분적으로, 노이즈가 다수의 염기 쌍 위치에 걸쳐있을 것으로 예상되는 정도에 기초하여 정의될 수 있다.

[0088] (156)에서, 1개 이상의 프로세서는 염기 쌍 윈도우 내의 참조 판독물의 각각의 위치에 대한 후보 변이체 빈도를 식별한다. 상기 예를 계속하여, 후보 변이체 빈도는 염기 쌍 윈도우(210) 내의 위치 #1-#3 각각에 대해 식별될 것이다(예를 들어, #1에서 약 2개의 거짓 변이체, 위치 #2에서 약 10개의 거짓 변이체, 및 위치 #3에서 약 0개

의 거짓 변이체).

[0089] (158)에서, 1개 이상의 프로세서는 후보 변이체 빈도 중 하나를 선택하여 기준선 게놈 서열 내의 현재 위치에 대한 기준선 변이체 빈도(또한 생성된 변이체 빈도라고도 지칭됨)를 나타낸다. 예를 들어, 현재 위치는 염기 쌍 윈도우(210)의 중심에 상응할 수 있다. 대안적으로, 현재 위치는 염기 쌍 윈도우(210)의 리딩 에지 또는 트레일링 에지에 상응할 수 있다. 기준선/생성된 변이체 빈도에 대한 선택은 다양한 기준에 기초할 수 있다. 예를 들어, 프로세서는 염기 쌍 윈도우 내에서 가장 높은 후보 변이체 빈도를 선택할 수 있다. 대안적으로, 프로세서는 평균화, 가중치 합 등과 같은 후보 변이체 빈도의 수학적 조합을 형성하여 기준선/생성된 변이체 빈도를 형성할 수 있다. 기준선/생성된 변이체 빈도는 현재의 염기 쌍 위치와 관련하여 저장된다. 예를 들어, 기준선/생성된 변이체 빈도는 기준선 게놈 서열 데이터베이스(도 7에서 (784))에 저장될 수 있다. 도 2b를 참조하면, 염기 쌍 윈도우(210)는 10의 기준선 변이체 빈도(윈도우 내에서 가장 높은 변이체 빈도에 상응함)로 할당될 수 있다.

[0090] (162)에서, 시스템의 1개 이상의 프로세서는 현재 위치와 관련하여 기준선 변이체 빈도를 저장한다. 기준선 변이체 빈도는 상이한 방식으로 현재 위치에 대해 기록될 수 있다. 염기 쌍 #1 및 염기 쌍 #3(둘 다 윈도우(210) 내)에 대한 변이체 빈도는 변하지 않고 남아있을 수 있다. 도 2b와 관련하여, 10의 기준선 변이체 빈도 값이 #2에 위치하는 염기 쌍과 관련하여 기록될 수 있다. 대안적으로, 기준선 변이체 빈도는 염기 쌍 윈도우 내의 리딩 위치 #1 또는 트레일링 위치 #3에 할당될 수 있는 반면, 인접한 위치는 널 값에 할당되거나 변화되지 않은 채로 유지된다. 임의로, 기준선 변이체 빈도는 염기 쌍 윈도우 내의 각각의 위치에 할당될 수 있다.

[0091] 임의로, 기준선 변이체 빈도는 미리 규정된 수준으로 설정될 수 있고, 여기서 상이한 미리 규정된 수준은 상이한 후보 변이체 빈도와 관련하여 이용된다. 예를 들어, 후보 변이체 빈도가 비교적 높을 때(예를 들어, 약 30% 초과), 기준선 변이체 빈도는 0 또는 또 다른 미리 규정된 더 낮은 수준으로 설정될 수 있다. 예로서, 배세포 유사 변이체를 배경 노이즈로서 카운팅하는 것을 피하기 위해 후보 변이체 빈도가 한계값을 초과할 때 기준선 변이체 빈도는 0 또는 미리 규정된 수준으로 설정될 수 있다.

[0092] (164)에서, 1개 이상의 프로세서는 염기 쌍 윈도우를 참조 판독물의 현재 세트를 따라 또 다른 위치로 이동시킬지 여부를 결정한다. 염기 쌍 윈도우를 이동시키는 결정은 참조 판독물의 추가의 부분이 분석되는 것으로 남아있는지 여부에 기초할 수 있다. (164)에서의 결정이 염기 쌍 윈도우를 이동시키는 것인 경우, 흐름은 (166)으로 이동한다. 그렇지 않으면, 흐름은 계속해서 (168)이다.

[0093] (166)에서, 시스템의 1개 이상의 프로세서는 염기 쌍 윈도우를 참조 판독물 세트를 따라 다음 위치로 이동시키고, 흐름은 (154)로 되돌아간다. 그 후, (154) 및 (156)에서의 작업을 반복한다. 예로서, (166)에서의 이동은 염기 쌍 윈도우를 단일 염기 쌍만큼, 또는 다수의 염기 쌍만큼 이동시킬 수 있다. 예를 들어, 도 2b를 참조하면, 이동은 단일 염기 쌍에 상응할 수 있다. 따라서, 염기 쌍 윈도우는 브라켓 (210)에서 지정된 위치로부터 브라켓 (212)에서 지정된 위치 (염기 쌍 위치 #2-#4를 포괄함)로 이동된다. (164-166)에서의 작업을 통한 후속 반복 동안, 염기 쌍 윈도우는 (212)에 지정된 위치로부터 (214)에 지정된 위치 (염기 쌍 위치 #3-#5를 포괄함)로 이동될 수 있다.

[0094] 본 실시예에서, 염기 쌍 윈도우는 연속적인 중첩 위치로 이동된다. 염기 쌍 윈도우가 3개 초과의 염기 쌍을 포함하는 경우, (166)에서의 이동은 유사하게 1개 염기 쌍보다 클 수 있다. 예를 들어, 염기 쌍 윈도우가 5개 염기 쌍 길이이고, 현재 위치가 윈도우(위치 #1-#5)에 상응하면, 166에서, 염기 쌍 윈도우는 염기 쌍 윈도우의 연속적인 위치가 중첩되지만 연속적인 염기 쌍 윈도우들 사이에서 다수의 염기 쌍만큼 앞으로 나아가도록, 2개 위치 앞으로 이동되어 위치 #3-#7에 중첩될 수 있다.

[0095] 대안적으로, 염기 쌍 윈도우는 참조 판독물을 따라 연속적인 비-중첩 위치로 이동될 수 있다. 예를 들어, 염기 쌍 윈도우가 3개의 염기 쌍 길이이고, 현재 위치가 참조 판독물 세트에서 위치 #1, #2 및 #3에 상응하면, 166에서, 염기 쌍 윈도우는 염기 쌍 윈도우의 연속적인 위치가 비-중첩되도록, 3개 위치 앞으로 이동되어 위치 #4, #5 및 #6에 중첩될 수 있다.

[0096] 다음으로, (154-162)에서의 작업을 다음 위치에 대해 반복한다. 흐름이 (164)에서 (168)로 진행하면, 시스템의 1개 이상의 프로세서는 복수의 참조 판독물로부터 참조 판독물의 추가 세트를 분석할지 여부를 결정한다. 추가의 참조 판독물을 분석하고자 하는 경우, 흐름은 (170)으로 되돌아간다. 달리, 도 1d의 프로세스는 종료되고, 흐름은 도 1c로 되돌아간다. (170)에서, 1개 이상의 프로세서는 분석될 참조 판독물의 다음 세트를 선택한다. 그 후, 흐름은 (152)로 되돌아간다. 이어서, (152-166)에서의 작업을 반복한다.

[0097] 도 3은 정상 기준선 게놈 서열을 이용하는 배경 보정의 원리와 관련된 그래프를 도시한다. 수평축은 뉴클레오티드 서열을 따라 일부 지정된 위치에서 샘플 내의 변이체의 빈도에 상응하는 변이체 대립유전자 빈도(VAF)를 나타낸다. 각각의 원(302)은 샘플과 연관된 변이체를 나타내고, 변이체(302)의 클러스터(304)는 샘플 클러스터에 대한 게놈 서열에 따른 위치에 상응한다. 클러스터(306)는 기준선 집단에 걸친 비교적 다량의 노이즈(또한 노이지 유전자좌로도 지칭됨)를 나타낸다. 원(306)은 또한 상이한 샘플과 연관된 변이체를 나타내고, 클러스터(308)는 샘플 클러스터에 대한 게놈 서열에 따른 위치에 상응한다. 클러스터 (308)는 기준선 집단에 걸쳐 비교적 소량의 노이즈(또한 클린(clean) 유전자좌로도 지칭됨)를 나타낸다.

[0098] 도 1d의 방법은 관심 게놈 서열을 따라 관심 위치에서 변이체 빈도를 수집함으로써 정상 샘플 집단에 대한 배경 노이즈에 대한 프로파일을 생성한다. 배경 집단은 대립유전자 빈도가 클러스터(308)에서와 같이 0 근처의 영역에 비교적 밀접하게/가까이 분포되어 있는 "클린 유전자좌"를 나타낸다. 따라서, 심지어 관심 샘플 게놈 서열에 대한 대립유전자 빈도가 낮은 경우에도(예를 들어, 종양 샘플 번호 1) 종양 변이체 호출(시험되는 개체가 가짐)은 노이즈로부터 쉽게 구별될 수 있다. 대안적으로, 배경 집단은 대립유전자 빈도가 클러스터 (304)에서와 같이 관심 게놈 서열을 따라 다양한 관심 지점에서 보다 광범위하게 분포/확산되는 "노이지 유전자좌"를 나타낼 수 있다. 따라서, 배경 노이즈로부터 낮은 변이체 빈도(예를 들어, 종양 샘플 1과 같음)를 나타내는 종양 변이체 호출(관심 샘플에 대한 것)을 구별하는 것이 더욱 어려울 수 있다. 그러나, 높은 빈도의 변이체 호출(예를 들어, 종양 샘플 2와 같음)은 관심 위치가 클린 또는 노이지 배경 특징을 나타내는지 여부와 관계없이 여전히 신뢰성 있게 호출될 수 있다. 본원의 실시양태는 정상 샘플의 배경 노이즈 수준을 이용하여, 종양 샘플과 관련된 호출 엄격성을 조정한다.

[0099] 다음으로, 본원의 실시양태의 측면은 변이체 호출 민감도 및 특이성을 개선하기 위해 상기 프로세스로부터의 정보 및 미가공 단편으로부터의 정보를 이용하는 것과 관련하여 기재된다.

[0100] 도 4는 변이체 호출을 필터링하기 위한 본원의 실시양태에 따라 구현된 계층-기반 결정 트리를 도시한다. (402)에서, 1개 이상의 프로세서는 착신 호출에 의해 변이체 (대립유전자)가 관찰되었다는 것이 표시되는지 여부를 결정한다. 변이체가 관찰되지 않는 경우, 이것은 호출이 어떠한 변이체도 없는 참조 서열에 상응하는 것으로 확인되었다는 것을 나타낼 수 있다. 착신 호출이 참조 서열에 상응하는 것으로 표시되는 경우, 흐름은 (404)로 분지된다. 그렇지 않으면, 착신 호출이 변이체/대립유전자를 포함하는 것으로 표시되는 경우, 흐름은 (406)으로 분지된다.

[0101] (404)에서, 1개 이상의 프로세서는 현재 변이체/대립유전자에 대해 미리 결정된 양의 커버리지가 수득되었는지 여부를 결정한다. 커버리지는 현재 위치/유전자좌를 포함하는 관심 게놈 서열의 단편을 커버하는 수집된 다수의 판독물의 표시에 기초하여 결정될 수 있다. 커버리지는 다양한 방식으로 정의될 수 있다. 예를 들어, 커버리지는 현재 위치/유전자좌에 대한 판독물을 수득하기 위해 분석된 다수의 분자에 기초하여 정의될 수 있다. 추가적으로 또는 대안적으로, 커버리지는 현재 위치/유전자좌에 대해 판독물이 수득된 다수의 샘플에 기초하여 정의될 수 있다. (404)에서, 커버리지 한계값이 정의된다 (예를 들어, 10X, 100X 등). 현재 유전자좌를 포함하는 판독물의 수가 커버리지 한계값을 초과하는 경우, 흐름이 (408)로 진행되고, 여기서 "음성 통과"가 판정된다. "음성 통과"는 프로세스가 원래 형태의 착신 호출을 검증한다는 것 (예를 들어, 음성 호출을 확인한다는 것)을 표시한다. 예를 들어, 착신 호출이 특정한 참조 서열에 상응하는 샘플 서열을 식별하는 경우, (408)에서, 프로세스는 변형 없이 호출을 통과시킨다.

[0102] 그렇지 않으면, 프로세서에 의해 커버리지가 커버리지 한계값과 동일하지 않거나 그를 초과하는 것으로 결정되는 경우, 흐름은 (410)으로 진행되고, 여기서 "호출 부재" 조건이 지정된다. 현재 유전자좌의 커버리지가 너무 낮아서 "변이체 호출 부재"가 정확하다는 충분한 신뢰도를 나타낼 수 없는 경우에 호출 부재 조건이 지정된다. 따라서, 착신 호출은 샘플 서열이 임의의 참조 서열에 대한 어떠한 상관관계의 표시도 제공하지 않는 것과 같은 "호출 부재"로 변경된다. 특정 커버리지 컷오프 미만의 참조 호출을 필터링하여 낮은 깊이로의 변이체-운반 판독물로 인해 변이체 부재를 판정하는 것을 피한다.

[0103] (406)에서, 1개 이상의 프로세서는 현재 변이체가 이전 변이체에 상응하는지 여부를 결정한다. 예를 들어, 현재 변이체는 집단에 걸쳐 나타난 공통 변이체에 관한 정보를 저장하는 변이체의 수집물 (도 7에서의 이전 데이터베이스(782) 내의 것)과 비교될 수 있다. 예를 들어, 집단에서 발생하는 것으로 공지된 변이체 호출에서의 신뢰도를 증가시키기 위해 공지된 데이터베이스 (예를 들어, 1000G 상 3 및 코스믹 데이터베이스)로부터의 정보가 사용될 수 있다. 현재 변이체가 이전 변이체 데이터베이스 내의 이전 변이체와 유사한 경우, 흐름은 분지(412)를 따라 이동하고, 여기서 일련의 계층 기반 계단식 시험이 수행된다. (406)에서, 데이터베이스로부터 이

전 변이체가 식별되지 않는 경우, 흐름은 분지(414)를 따라 이동하고, 여기서 상이한 일련의 계층 기반 계단식 시험이 수행된다. 계단식 시험 중 1개 이상은 각각의 분지(412 및 414)를 따라 수행될 수 있다. 추가로, 다양한 시험이 도 4에 도시된 특정한 배열 이외의 대안적인 순서로 및 대안적인 조합으로 수행될 수 있다. 도 4에서의 각각의 시험은 변이체가 발생하거나 발생하지 않았다는 상응하는 정도/수준의 2차적 지지 증거를 제공하고, 시험은 2차적 지지 증거로서 상이한 정도의 신뢰도를 나타낸다.

[0104] 분지(414) 후에, (416)에서, 1개 이상의 프로세서는 현재 변이체에 대한 변이체 품질 점수가 품질 점수 한계값을 초과하는지 여부를 결정한다. 변이체 품질 점수는 도 1c 및 1d와 관련하여 기재된 바와 같이, 본원의 실시양태에 따라 결정된다. 본원에서 설명되는 바와 같이, 변이체 품질 점수는 샘플 변이체 빈도와 다수의 기준선 변이체 빈도 사이의 관계를 나타낸다. 변이체 품질 점수가 품질 점수 한계값 미만인 경우, 흐름은 (418)로 이동하고, 여기서 현재 변이체는 무효화되며, 예컨대 거짓 변이체로 지정된다 (예를 들어, 호출을 낮은 품질 점수로 인해 "음성"을 나타내는 것으로 판정함). 변이체 품질 점수가 품질 점수 한계값 이상인 경우, 흐름은 (420)으로 이동한다.

[0105] (420)에서, 1개 이상의 프로세서는 변이체 호출을 지지하는 듀플렉스 단편 (예를 들어, 듀플렉스 스티칭 또는 비-스티칭)의 카운트를 수득하고 검토한다. 카운트는 샘플 판독물에서 잠재적 변이체에 상응하는 지지 변이체를 수득하기 위해 변이체 호출 어플리케이션에 의해 식별된 다수의 듀플렉스 미가공 단편을 표시한다. 서열분석 동안, DNA의 두 가닥을 분석하고, 각각의 가닥으로부터의 단편에 대해 판독물을 수득한다. 예로서, 적어도 제1 및 제2 판독물은 제1 가닥의 중첩 단편에 대해 수득될 수 있고, 반면 제3 및 제4 판독물은 제2 가닥의 단편에 대해 수득된다. 1-4개의 판독물의 세트를 검토하여 몇개의 판독물이 현재 위치/유전자좌에서 변이체를 표시하는지를 결정한다.

[0106] 변이체가 샘플 판독물에 존재하는 경우, 각각의 제1 및 제2 가닥 내의 현재 유전자좌에 대한 모든 판독물이 관련 또는 "지지" 변이체를 나타낼 것으로 예상될 것이다. 예를 들어, 샘플 판독물이 제1 가닥에 상응하고 잠재적 변이체에 대해 호출되는 경우 (예를 들어, 유전자좌가 "A"를 나타내는 경우), 제2 가닥 (미가공 단편에 상응함) 내의 상응하는 유전자좌는 "T"인 것으로 예상될 것이다. 제2 가닥 (미가공 단편) 내의 상응하는 유전자좌가 "T"인 변이체를 포함하는 경우, 미가공 단편의 현재 판독물은 지지 변이체를 갖는 것으로 카운팅될 것이다. 이중 가닥 DNA 내의 각각의 가닥 (샘플 단편 판독물 및 미가공 단편 판독물)에서 매칭 지지 변이체가 발생하는 경우, 이것은 샘플 판독물 내 지정된 위치에서 변이체가 발생하였다는 양호한 지표로 간주될 수 있다.

[0107] 서열분석 분석 (변이체 호출 어플리케이션에 의한 것)이 지지 변이체를 나타내는 미가공 단편 판독물의 높은 카운트를 생성하는 경우 ((420)의 블록에서 "DS>=한계값"으로 표시됨), 흐름은 (422)로 이동한다. (422)에서, 착신 호출이 검증되고, 즉 변이체로서 통과되거나 또는 양성 변이체 호출로 지정된다. 서열분석 분석이 지지 변이체를 나타내는 미가공 단편 판독물의 낮은 카운트를 생성하는 경우, 이러한 상황은 잠재적 변이체 호출에서 오류가 발생하였다는 것 및 변이체 호출 검증 적용은 착신 호출이 변이체 호출을 나타내는 신뢰도의 낮은 수준을 표시한다는 것의 표시로서 해석된다. 따라서, 흐름은 (424)로 이동하고, 여기서 착신 변이체 호출은 무효화되며, 즉 "음성 통과"로 지정된다.

[0108] (416 및 420)에서의 변이체 시험은 그의 결과에서 높은 수준의 신뢰도 (다른 유형의 시험에서의 신뢰도에 비해)를 나타내는 시험을 나타낸다. 추가적인 또는 대안적인 시험이 적용될 수 있을 뿐만 아니라 (416 및 420)에서의 시험의 순서를 역전시킬 수 있는 것으로 인식된다.

[0109] (406)으로 돌아가, 이어서 계층 기반 결정 트리가 분지(412)와 관련하여 기재된다. 흐름이 분지(412)를 따라 이동하는 경우, (426)에서, 1개 이상의 프로세서는 착신 변이체 호출에 할당된 변이체 품질 점수를 검토한다. 변이체 품질 점수가 미리 결정된 품질 점수 한계값 미만인 경우, 흐름은 (428)로 이동하고, 여기서 착신 변이체 호출은 거짓 변이체로 판정된다. 예를 들어, 착신 변이체 호출은 통과 거부되거나 또는 "음성" (낮은 품질로 인함)으로 지정된다. 변이체 품질 점수가 미리 결정된 품질 점수 한계값 이상인 경우, 흐름은 (430)으로 이동한다. (426)에서 이용되는 품질 점수 한계값은 (416)에서 이용되는 품질 점수 한계값과 동일하거나 상이할 수 있다.

[0110] (430)에서, 1개 이상의 프로세서는 1개 이상의 유형의 미가공 단편에 대한 카운트와 연관된 가중 점수(WS)를 결정한다. 예를 들어, 변이체 호출 어플리케이션은 다수의 상이한 미가공 단편 유형, 예컨대 듀플렉스 스티칭된 단편, 심플렉스 스티칭된 단편, 듀플렉스 비-스티칭된 단편 및 심플렉스 비-스티칭된 단편에 대한 카운트를 출력할 수 있다. 카운트는 지지 변이체를 포함한, 상응하는 유형으로부터의 미가공 단편의 판독물의 수를 표시한다. (430)에서, 프로세서는 카운트에 가중치를 적용하고, 결과를 합산하여 상이한 유형의 미가공 단편들을 조

합하는 가중 점수를 수득한다. 가중 합계를 가중 점수(WS) 한계값과 비교한다. 가중 점수의 합계가 WS 한계값을 초과하는 경우, 흐름은 (432)로 진행된다. 그렇지 않으면, 흐름은 (434)로 이동하고, 여기서 착신 변이체 호출은 확인되지 않고, 무효로 판정되며, "음성 통과"로 나타내어진다.

[0111] (432)에서, 1개 이상의 프로세서는 착신 변이체 호출을 검증한 미가공 단편의 카운트 합산을 수행한다. 예를 들어, 프로세서는 지지 변이체를 포함한, 상응하는 유형으로부터의 미가공 단편의 판독물의 수를 표시하는 카운트를 합산할 수 있다. 합산 카운트를 합산 카운트(SC) 한계값과 비교한다. 합산 카운트가 SC 한계값 미만에 속하는 경우, 흐름은 (436)으로 이동한다. 그렇지 않으면, 흐름은 (438)로 이동하고, 여기서 착신 변이체 호출이 검증되며, 즉 "양성 통과"로 판정된다.

[0112] (436)에서, 1개 이상의 프로세서는 지지 변이체를 포함하는 듀플렉스 스티칭된 단편(들)의 카운트를 검토하여 ((420)에서의 프로세스와 유사하게) 듀플렉스 스티칭된 단편(들)이 착신 변이체 호출을 인증/확인하는지 여부를 결정한다. 듀플렉스 스티칭된 단편(들)과 연관된 변이체 결정을 듀플렉스 스티칭(DS) 한계값과 비교한다. DS 지지 변이체의 카운트가 한계값 미만에 속하는 경우, 흐름이 (440)으로 이동하고, 여기서 착신 변이체 호출이 무효화되며, 즉 "음성 경로"가 판정된다. 듀플렉스 스티칭된 미가공 단편의 카운트가 DS 한계값 이상인 경우, (436)에서, 흐름은 (442)로 이동하고, 여기서 착신 변이체 호출이 검증되며, 즉 "양성 통과"를 나타내는 것으로 판정된다.

[0113] 상기 계층 기반 결정 트리는 변이체 관련 인자가 분석될 수 있는 방식의 한 예를 나타내는 것으로 인식된다. 예를 들어, 특정 예에서, 변이체 관련 인자는 결정 트리 내에서 상이한 순서로 고려되도록 재배열될 수 있다. 도 4와 관련하여, 미가공 단편에 관한 정보 (예를 들어, 듀플렉스 스티칭된 미가공 단편, 미가공 단편에 대한 가중 점수, 변이체 미가공 단편 카운트 등)에 기초하여 다양한 결정이 이루어진다. 추가로, 변이체 품질 점수는 또 다른 필터 기준으로서 사용된다. 계층 기반 결정 트리는 상기 정보를 사용하여, 예컨대 ctDNA 데이터를 사용하여 변이체 호출 민감도 및 특이성을 개선시킨다. 계층 기반 결정 트리를 사용하여 변이체 호출기로 처음 불리는 변이체를 필터링한다. 도 4의 실시양태에서, 가중 단편 지지는 변이체 호출에서 신뢰도를 유도하는 것을 돕는데 사용되며, 단 착신 변이체 호출은 이전 변이체와 매칭되고, 품질 점수는 품질 점수 한계값을 초과한다. 점수는 듀플렉스/심플렉스 및 스티칭/비-스티칭된 단편을 포함한, 각각의 변이체에 대한 상이한 미가공 단편 유형에 대해 상이한 가중치를 제공한다. 가중치는 공지된 결과를 갖는 데이터로부터 훈련될 수 있다. 통과하는 변이체 호출은 정상 기준선 (이전 섹션 참조)과 비교하여 계산된 충분히 높은 변이체 Q-점수 뿐만 아니라 충분한 가중 점수를 가져야 한다. 그러나, 가중 단편 지지는 결정 트리에서의 상이한 지점에서 수행될 수 있거나 또는 전적으로 생략될 수 있는 것으로 인식된다.

[0114] 도 4의 실시양태에서, 계층 기반 결정 트리는 착신 변이체 호출을 검증하기 전에 적용되는 적어도 3개 층의 지지 단편 시험을 포함한다. 예를 들어, (422)에서 착신 변이체 호출을 검증하기 전에, 미가공 단편 시험은 3개 층, 즉 (406, 416 및 420)에서 양성 결과를 제공한다. (438)에서 착신 변이체 호출을 검증하기 전에, 미가공 단편 시험은 4개 층, 즉 (406, 426, 430 및 432)에서 양성 결과를 제공한다. (442)에서 착신 변이체 호출을 검증하기 전에, 미가공 단편 시험은 5개 층, 즉 (406, 426, 430, 432 및 436)에서 양성 결과를 제공한다. 미가공 단편에 적용되는 시험의 층의 수는 임의의 특정한 시험에 의해 제공되는 신뢰도에 기초하여 달라질 수 있다.

[0115] 도 5a는 변이체 호출을 필터링하기 위해 본원의 대안적인 실시양태에 따라 구현된 계층-기반 결정 트리를 도시한다. 도 5a는 일반화된 프로세스이며, 이 중 하나의 구현은 도 4에 제시되어 있다. (550)에서, 1개 이상의 프로세서는 관심 게놈 서열에 대한 잠재적 변이체 호출의 표시를 수신한다. (552)에서, 1개 이상의 프로세서는 관심 게놈 서열을 따라 샘플 및 미가공 뉴클레오티드 단편에 대한 판독물을 포함한 서열분석 데이터를 수신한다. 판독물은 관심 게놈 서열을 따라 지정된 위치에서의 뉴클레오티드의 서열에 상응하는 샘플 단편에 대한 샘플 판독물을 포함한다. 4

[0116] (554)에서, 1개 이상의 프로세서는 품질 점수를 수득하기 위해, 기준선 게놈 서열에 대한 지정된 위치에서의 기준선 변이체 빈도에 비해, 관심 게놈 서열에 대한 지정된 위치에서의 샘플 변이체 빈도를 분석한다. 다음으로, 1개 이상의 프로세서는 미가공 단편에 관한 다수의 시험 중 1개 이상을 수행한다. 도 5a의 예에서, (556, 558 및 560)에서의 3개의 결정이 도시되어 있다. 임의로, (556, 558 및 560)에서의 결정이 반복될 수 있다 (분지 (564)에서 나타낸 바와 같음). 추가적으로 또는 대안적으로, (556, 558 및 560)에서의 결정은 상이한 정보와 관련하여 1회 초과로 반복될 수 있고, 임의의 순서로 수행될 수 있으며, 잠재적 변이체 호출을 검증하는 것과 관련하여 상이한 가중치가 제공될 수 있다. 임의로, (556, 558 및 560)에서의 결정 중 1개 이상이 완전히 생략될 수 있다.

[0117] (556)에서, 1개 이상의 프로세서는 미가공 단편이 잠재적 변이체 호출을 확인하는지 여부를 결정한다. 예를 들어, 미가공 단편은 듀플렉스 스티칭된 단편, 심플렉스 스티칭된 단편, 듀플렉스 비-스티칭된 단편 또는 심플렉스 비-스티칭된 단편 중 적어도 1개에 상응할 수 있다. 한 예로서, 결정은 지지 변이체를 포함한 미가공 단편으로서 다수의 듀플렉스 스티칭된 단편의 카운트를 식별하는 것을 포함할 수 있다. 프로세서는 지지 변이체를 포함한 듀플렉스 스티칭된 단편의 카운트가 DS 한계값을 초과하는지 여부를 결정함으로써, 잠재적 변이체 호출을 확인할 것이다. 임의로, (556)에서, 프로세서는 도 4에서의 (420, 430, 432 및 436)에서의 작업과 관련하여 상기 논의된 바와 같은 듀플렉스, 가중 점수 및 합산 카운트 결정을 수행할 수 있다.

[0118] (558)에서, 1개 이상의 프로세서는 샘플 판독물이 샘플 단편에 대한 미리 결정된 양의 커버리지를 제공하는지 여부를 결정한다. 샘플 판독물이 미리 결정된 양의 양의 커버리지 미만에 속하는 경우에 잠재적 변이체 호출은 호출-부재로 판정될 수 있다.

[0119] (560)에서, 1개 이상의 프로세서는 잠재적 변이체 호출이 미리 규정된 집단에 걸쳐 나타난 이전 변이체 호출과 매칭되는지 여부를 결정한다. 이전 변이체와의 상관관계는 다른 시험에서 이용되는 단지 1개 이상의 한계값에 대한 정보로서 이용될 수 있다. 예를 들어, 잠재적 변이체 호출과 이전 변이체 호출 사이에서 매치가 결정되는 경우, 작업(556) 또는 작업(558) 중 적어도 1개를 수행하는 결정 작업은 제1 한계값을 이용한다. 추가적으로, 잠재적 변이체 호출과 이전 변이체 호출 사이에서 매치가 존재하지 않는 경우, 작업(556) 또는 작업(558) 중 적어도 1개를 수행하는 결정 작업은 제2 한계값을 이용한다.

[0120] (562)에서, 1개 이상의 프로세서는 품질 점수 및 결정 작업(들)에 기초하여 잠재적 변이체 호출을 검증한다. 예를 들어, 품질 점수가 품질 점수 한계값을 초과하고 (556 - 560)에서의 결정 중 적어도 1개가 잠재적 변이체 호출을 지지하는 경우에 잠재적 변이체 호출이 검증될 수 있다. 추가적으로 또는 대안적으로, (556 - 560)에서의 결정은, 예컨대 각각의 결정의 가중치가 동일하거나 상이할 수 있는 가중 합계를 통해 조합될 수 있다. (556 - 560)에서의 결정의 가중 합계는 한계값과 비교될 수 있고, 가중 합계가 한계값을 초과하는 경우에 잠재적 변이체 호출만이 검증될 수 있다. 임의로, (556 - 560)에서의 결정에 제공되는 중요성 및/또는 가중치는 품질 점수의 수준에 기초하여 달라질 수 있다. 예를 들어, 매우 높은 품질 점수가 결정되는 경우, 잠재적 변이체 호출을 검증하기 위해, (556 - 560)에서의 결정 중 단지 1개만이 잠재적 변이체 호출을 지지하는 것으로 결정될 수 있다. 대안적으로, 중간 수준의 품질 점수가 결정되는 경우, 잠재적 변이체 호출을 검증하기 전에, (556 - 560)에서의 결정 중 1개 이상이 잠재적 변이체 호출을 지지하는데 필요할 수 있다. 또 다른 예로서, (556 - 560)에서의 결정에 적용된 한계값은 품질 점수의 수준에 기초하여 달라질 수 있다.

[0121] 도 5b는 본원의 실시양태에 따라 가중 점수를 이용하는 것 (예를 들어, 도 4에서의 (430) 및/또는 도 5a에서의 (556)의 작업)과 관련하여 수행된 프로세스를 도시한다. (570)에서, 1개 이상의 프로세서는 잠재적 변이체 호출과 관련된 지지 변이체를 포함하는 미가공 단편의 카운트를 식별한다. (572)에서, 1개 이상의 프로세서는 지지 변이체 호출을 표시하는 미가공 단편에 대한 가중 단편 점수를 수득한다. 예를 들어, 듀플렉스 스티칭된 미가공 단편과 연관된 카운트에는 제1 가중 인자를 곱할 수 있고, 심플렉스 스티칭된 미가공 단편과 연관된 카운트에는 제2 가중 인자를 곱할 수 있고, 듀플렉스 비-스티칭된 단편과 연관된 카운트에는 제3 가중 인자를 곱할 수 있고, 심플렉스 비스티칭된 단편과 연관된 카운트에는 제4 가중 인자를 곱할 수 있다. 이어서, 제1 내지 제4 가중 인자를 합산하거나 또는 일부 다른 방식으로 조합하여 가중 단편 점수를 수득할 수 있다. 추가적으로 또는 대안적으로, 가중 단편 점수는 지지 변이체를 포함하지 않은 미가공 단편의 수에 기초하여 변형될 수 있다. 예를 들어, 많은 수의 지지 단편이 참조 서열에 상응하는 경우, 미가공 단편 (참조 서열에 매칭됨)의 카운트는 가중 단편 점수를 감소시키기 위한 인자로서 이용될 수 있다. (574)에서, 1개 이상의 프로세서는 가중 단편 점수를 미가공 단편 한계값과 비교하여 잠재적 변이체 호출을 확인할 수 있는지 여부를 결정한다.

[0122] 컴퓨터-구현 양태

[0123] 관련 기술분야의 통상의 기술자에 의해 이해되는 바와 같이, 본원에 기재된 방법 및 정보는 공지된 컴퓨터 판독가능 매체 상의 컴퓨터 실행가능 명령어로서 전부 또는 부분적으로 구현될 수 있다. 예를 들어, 본원에 기재된 방법은 하드웨어에서 구현될 수 있다. 대안적으로, 방법은, 예를 들어 1개 이상의 메모리 또는 다른 컴퓨터 판독가능 매체에 저장되어 있는 소프트웨어에서 구현될 수 있고 1개 이상의 프로세서 상에서 구현될 수 있다. 공지된 바와 같이, 프로세서는 1개 이상의 제어기, 계산 유닛 및/또는 컴퓨터 시스템의 다른 유닛과 연관될 수 있거나 또는 원하는 경우 펌웨어에 심어질 수 있다. 소프트웨어에서 구현되는 경우, 루틴은 또한 공지된 바와 같이, 임의의 컴퓨터 판독가능 메모리, 예컨대 RAM, ROM, 플래시 메모리, 자기 디스크, 레이저 디스크 또는 다른 저장 매체에 저장될 수 있다. 마찬가지로, 이러한 소프트웨어는, 예를 들어 전화선, 인터넷, 무선 접속 등과

같은 통신 채널을 통한 것 또는 컴퓨터 판독가능 디스크, 플래쉬 드라이브 등과 같은 수송가능 매체를 통한 것을 포함한 임의의 공지된 전달 방법을 통해 컴퓨팅 디바이스로 전달될 수 있다.

[0124] 보다 일반적으로, 및 관련 기술분야의 통상의 기술자에 의해 이해되는 바와 같이, 상기 기재된 다양한 단계는 다양한 블록, 작업, 도구, 모듈 및 기술로서 구현될 수 있으며, 이는 차례로, 하드웨어, 펌웨어, 소프트웨어, 또는 하드웨어, 펌웨어 및/또는 소프트웨어의 임의의 조합에서 구현될 수 있다. 하드웨어에서 구현되는 경우, 블록, 작업, 기술 등 중 일부 또는 전부는, 예를 들어 주문형 집적 회로(IC), 응용 특이적 집적 회로(ASIC), 필드 프로그램가능 논리 배열(FPGA), 프로그램가능 논리 배열(PLA) 등에서 구현될 수 있다.

[0125] 소프트웨어에서 구현되는 경우, 소프트웨어는 임의의 공지된 컴퓨터 판독가능 매체, 예컨대 자기 디스크, 광학 디스크 또는 다른 저장 매체, 컴퓨터 또는 프로세서의 RAM 또는 ROM 또는 플래시 메모리, 하드 디스크 드라이브, 광학 디스크 드라이브, 테이프 드라이브 등에 저장될 수 있다. 마찬가지로, 소프트웨어는, 예를 들어 컴퓨터 판독가능 디스크 또는 다른 수송가능 컴퓨터 저장 메카니즘을 포함한 임의의 공지된 전달 방법을 통해 사용자 또는 컴퓨팅 시스템으로 전달될 수 있다.

[0126] 도 6은 본원의 실시양태에 따라 형성된 컴퓨팅 시스템 환경(600)의 블록 다이어그램을 도시한다. 컴퓨팅 시스템 환경(600)은 적합한 컴퓨팅 환경의 단지 하나의 예이고, 본원의 용도 또는 기능의 범주에 대한 임의의 제한을 시사하는 것으로 의도되지 않는다. 컴퓨팅 환경(600)은 예시적인 작동 환경(600)에서 예시된 구성요소들 중 어느 하나 또는 조합에 관한 임의의 의존성 또는 요건을 갖는 것으로 해석되지 않아야 한다. 방법 및 시스템은 수많은 다른 일반 목적 또는 특수 목적 컴퓨팅 시스템 환경 또는 구성에 의해 작동한다. 청구범위의 방법 또는 시스템과 함께 사용하기에 적합할 수 있는 컴퓨팅 시스템, 환경 및/또는 구성의 예는 개인용 컴퓨터, 서버 컴퓨터, 핸드-헬드 또는 랩탑 디바이스, 멀티프로세서 시스템, 마이크로프로세서-기반 시스템, 셋탑 박스, 프로그램가능 소비자 전자장치, 네트워크 PC, 미니컴퓨터, 메인프레임 컴퓨터, 상기 시스템 또는 디바이스 중 임의의 것을 포함하는 분산 컴퓨팅 환경 등을 포함하나 이에 제한되지는 않는다.

[0127] 방법 및 시스템의 작업은 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터-실행가능 명령어의 일반 컨텍스트로 기재될 수 있다. 일반적으로, 프로그램 모듈은 특정한 과제를 수행하거나 또는 특정한 추상 데이터 유형을 구현하는 루틴, 프로그램, 객체, 구성요소, 데이터 구조 등을 포함한다. 방법 및 장치는 또한 과제가 통신 네트워크를 통해 연결되는 원격 프로세싱 디바이스에 의해 수행되는 분산 컴퓨팅 환경에서 실시될 수도 있다. 집적 및 분산 컴퓨팅 환경 둘 다에서, 프로그램 모듈은 메모리 저장 디바이스를 포함한 로컬 및 원격 컴퓨터 저장 매체 둘 다에 위치될 수 있다.

[0128] 도 6을 참조하면, 컴퓨터(610)의 구성요소는 프로세싱 유닛(620), 시스템 메모리(630), 및 시스템 메모리를 포함한 다양한 시스템 구성요소를 프로세싱 유닛(620)에 커플링시키는 시스템 버스(621)를 포함할 수 있으나 이에 제한되지는 않는다. 시스템 버스(621)는 메모리 버스 또는 메모리 제어기, 주변 버스, 및 다양한 버스 아키텍처들 중 임의의 것을 사용하는 로컬 버스를 포함한 여러 유형의 버스 구조물 중 임의의 것일 수 있다. 예로서 및 비제한적으로, 이러한 아키텍처는 산업 표준 아키텍처(USA) 버스, 마이크로 채널 아키텍처(MCA) 버스, 향상된 ISA(EISA) 버스, 비디오 전자공학 표준 위원회(VESA) 로컬 버스, 및 메자닌(Mezzanine) 버스로도 공지된 주변 구성요소 상호접속(PCI) 버스를 포함한다.

[0129] 컴퓨터(610)는 전형적으로 다양한 컴퓨터 판독가능 매체를 포함한다. 컴퓨터 판독가능 매체는 컴퓨터(610)에 의해 액세스될 수 있고 휘발성 및 비휘발성 매체, 착탈식 및 비-착탈식 매체 둘 다를 포함하는 임의의 이용가능한 매체일 수 있다. 예로서 및 비제한적으로, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체 및 통신 매체를 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 다른 데이터와 같은 정보의 저장을 위해 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 착탈식 및 비-착탈식 매체 둘 다를 포함한다. 컴퓨터 저장 매체는 RAM, ROM, EEPROM, 플래시 메모리 또는 다른 메모리 기술, CD-ROM, 디지털 다기능 디스크(DVD) 또는 다른 광학 디스크 저장장치, 자기 카세트, 자기 테이프, 자기 디스크 저장장치 또는 다른 자기 저장 디바이스, 또는 원하는 정보를 저장하는데 사용될 수 있으며 컴퓨터(610)에 의해 액세스될 수 있는 임의의 다른 매체를 포함하나 이에 제한되지는 않는다. 통신 매체는 전형적으로 반송파 또는 다른 전송 메카니즘과 같은 변조된 데이터 신호로 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 다른 데이터를 구현하고, 임의의 정보 전달 매체를 포함한다. 용어 "변조된 데이터 신호"는 신호의 특징 중 1개 이상을 그 신호 내의 정보를 코딩하는 방식으로 설정 또는 변경시킨 신호를 의미한다. 예로서 및 비제한적으로, 통신 매체는 유선 네트워크 또는 직접-유선 접속과 같은 유선 매체, 및 음향, RF, 적외선 및 다른 무선 매체와 같은 무선 매체를 포함한다. 상기 중 임의의 것의 조합이 또한 컴퓨터 판독가능 매체의 범주 내에 포함될 것이다.

[0130] 시스템 메모리(630)는, 판독 전용 메모리(ROM)(631) 및 랜덤 액세스 메모리(RAM)(632)와 같은 휘발성 및/또는 비휘발성 메모리의 형태의 컴퓨터 저장 매체를 포함한다. 시동 동안과 같이 컴퓨터(610) 내의 요소들 사이의 정보 전달을 돕는 기본 루틴을 함유하는 기본 입력/출력 시스템(633)(BIOS)은 전형적으로 ROM(631)에 저장된다. RAM(632)은 전형적으로 프로세싱 유닛(620)에 의해 즉시 액세스가능하고/거나 현재 작동되고 있는 데이터 및/또는 프로그램 모듈을 함유한다. 예로서 및 비제한적으로, 도 6은 작동 시스템(634), 애플리케이션 프로그램(635), 다른 프로그램 모듈(636) 및 프로그램 데이터(637)를 도시한다.

[0131] 컴퓨터(610)는 또한 다른 착탈식/비-착탈식, 휘발성/비휘발성 컴퓨터 저장 매체를 포함할 수 있다. 단지 예로서, 도 6은 비-착탈식, 비휘발성 자기 매체로부터 판독하거나 이러한 자기 매체에 기록하는 하드 디스크 드라이브(640), 착탈식, 비휘발성 자기 디스크(652)로부터 판독하거나 이러한 자기 디스크에 기록하는 자기 디스크 드라이브(651), 및 CD ROM이나 다른 광학 매체와 같은 착탈식, 비휘발성 광학 디스크(656)로부터 판독하거나 이러한 광학 디스크에 기록하는 광학 디스크 드라이브(655)를 도시한다. 예시적인 작동 환경에서 사용될 수 있는 다른 착탈식/비-착탈식, 휘발성/비휘발성 컴퓨터 저장 매체는 자기 테이프 카세트, 플래시 메모리 카드, 디지털 다기능 디스크, 디지털 비디오 테이프, 고체 상태 RAM, 고체 상태 ROM 등을 포함하나 이에 제한되지는 않는다. 하드 디스크 드라이브(641)는 전형적으로 인터페이스(640)와 같은 비-착탈식 메모리 인터페이스를 통해 시스템 버스(621)에 접속되고, 자기 디스크 드라이브(651) 및 광학 디스크 드라이브(655)는 전형적으로 인터페이스(650)와 같은 착탈식 메모리 인터페이스에 의해 시스템 버스(621)에 접속된다.

[0132] 상기에서 논의되고 도 6에 도시된 드라이브 및 연관된 컴퓨터 저장 매체는 컴퓨터(610)를 위한 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 및 다른 데이터의 저장을 제공한다. 도 6에서, 예를 들어, 하드 디스크 드라이브(641)는 작동 시스템(644), 애플리케이션 프로그램(645), 다른 프로그램 모듈(646) 및 프로그램 데이터(647)를 저장하는 것으로 도시되어 있다. 이들 구성요소가 작동 시스템(634), 애플리케이션 프로그램(635), 다른 프로그램 모듈(636) 및 프로그램 데이터(637)와 동일하거나 상이할 수 있다는 것을 주목한다. 작동 시스템(644), 애플리케이션 프로그램(645), 다른 프로그램 모듈(646) 및 프로그램 데이터(647)는 최소한 이들이 상이한 카피라는 것을 예시하기 위해 여기서 상이한 번호가 주어진다. 사용자는 통상적으로 마우스, 트랙볼 또는 터치 패드로 지칭되는 포인팅 디바이스(661) 및 키보드(662)와 같은 입력 디바이스를 통해 명령 및 정보를 컴퓨터 내로 입력할 수 있다. 다른 입력 디바이스 (제시되지 않음)는 마이크로폰, 조이스틱, 게임 패드, 위성 안테나, 스캐너 등을 포함할 수 있다. 이들 및 다른 입력 디바이스는 종종 시스템 버스에 커플링되는 사용자 입력 인터페이스(660)를 통해 프로세싱 유닛(620)에 접속되지만, 병렬 포트, 게임 포트 또는 범용 직렬 버스(USB)와 같은 다른 인터페이스 및 버스 구조물에 의해 접속될 수 있다. 모니터(691) 또는 다른 유형의 디스플레이 디바이스가 또한 비디오 인터페이스(690)와 같은 인터페이스를 통해 시스템 버스(621)에 접속된다. 모니터 이외에, 컴퓨터는 또한 스피커(697) 및 프린터(696)와 같은 다른 주변 출력 디바이스를 포함할 수 있으며, 이들은 출력 주변 인터페이스(690)를 통해 접속될 수 있다.

[0133] 컴퓨터(610)는 원격 컴퓨터(680)와 같은 1개 이상의 원격 컴퓨터에 대한 논리적 접속을 사용하여 네트워킹 환경에서 작동할 수 있다. 원격 컴퓨터(680)는 개인용 컴퓨터, 서버, 라우터, 네트워크 PC, 피어 디바이스 또는 다른 통상의 네트워크 노드일 수 있고, 도 6에는 단지 메모리 저장 디바이스(681)만이 도시되었지만, 전형적으로 컴퓨터(610)에 대하여 상기 기재된 많은 또는 모든 요소를 포함한다. 도 6에 도시된 논리적 접속은 로컬 영역 네트워크(LAN)(671) 및 광역 네트워크(WAN)(673)를 포함하지만, 또한 다른 네트워크를 포함할 수도 있다. 이러한 네트워킹 환경은 사무실, 기업의 컴퓨터 네트워크, 인트라넷 및 인터넷에서 통상적이다.

[0134] LAN 네트워킹 환경에서 사용되는 경우, 컴퓨터(610)는 네트워크 인터페이스 또는 어댑터(670)를 통해 LAN(671)에 연결된다. WAN 네트워킹 환경에서 사용되는 경우, 컴퓨터(610)는 전형적으로 인터넷과 같은 WAN(673)을 통해 통신을 확립하기 위한 모뎀(672) 또는 다른 수단을 포함한다. 내장형 또는 외장형일 수 있는 모뎀(672)은 사용자 입력 인터페이스(660) 또는 다른 적절한 메카니즘을 통해 시스템 버스(621)에 접속될 수 있다. 네트워킹 환경에서, 컴퓨터(610)에 대하여 도시된 프로그램 모듈 또는 그의 부분은 원격 메모리 저장 디바이스에 저장될 수 있다. 예로서 및 비제한적으로, 도 6은 원격 애플리케이션 프로그램(685)이 메모리 디바이스(681)에 존재하는 것으로 도시한다. 제시된 네트워크 접속은 예시적이고, 컴퓨터 사이에 통신 연결을 확립하는 다른 수단이 사용될 수 있는 것으로 인지될 것이다.

[0135] 위험 평가 시스템 및 방법 및 다른 요소가 바람직하게는 소프트웨어에서 구현되는 것으로 기재되었지만, 이들은 하드웨어, 펌웨어 등에서 구현될 수 있고, 임의의 다른 프로세서에 의해 구현될 수 있다. 따라서, 본원에 기재된 요소는 표준 다목적 CPU 또는 특이적으로 설계된 하드웨어 또는 펌웨어, 예컨대 도 6의 컴퓨터(610)를 포함하나 이에 제한되지는 않는, 원하는 바와 같은 애플리케이션-특이적 집적 회로(ASIC) 또는 다른 하드-유선 디바

이스에서 구현될 수 있다. 소프트웨어에서 구현되는 경우, 소프트웨어 루틴은 임의의 컴퓨터 판독가능 메모리, 예컨대 자기 디스크, 레이저 디스크 또는 다른 저장 매체, 컴퓨터 또는 프로세서의 RAM 또는 ROM, 임의의 데이터베이스 등에 저장될 수 있다. 마찬가지로, 이러한 소프트웨어는, 예를 들어 컴퓨터 판독가능 디스크 또는 다른 수송가능 컴퓨터 저장 메카니즘 상 또는 전화선, 인터넷, 무선 통신 등과 같은 통신 채널을 통한 것 (이는 수송가능 저장 매체를 통해 이러한 소프트웨어를 제공하는 것과 동일하거나 또는 상호교환가능한 것으로 간주됨)을 포함한 임의의 공지된 또는 원하는 전달 방법을 통해 사용자 또는 진단 시스템으로 전달될 수 있다. 따라서, 본 발명의 취지 및 범주로부터 벗어나지 않고 본원에 기재 및 도시된 기술 및 구조에 대한 많은 변형 및 변동이 이루어질 수 있다. 따라서, 본원에 기재된 방법 및 장치는 단지 예시적인 것이고 본 발명의 범주를 제한하지 않는 것으로 이해되어야 한다.

[0136] 도 7은 본원의 실시양태에 따라 형성된 대안적 시스템의 블록 다이어그램을 도시한다. 점선 윤곽에 제시된 구성요소는 기술적으로 컴퓨터(710)의 일부가 아니라 도 7의 예시적인 실시양태를 도시하기 위해 사용된다. 컴퓨터(710)의 구성요소는 프로세서(720), 시스템 메모리(730), 노스브리지 칩(Northbridge chip)으로도 공지된 메모리/그래픽 인터페이스(721), 및 사우스브리지 칩(Southbridge chip)으로도 공지된 I/O 인터페이스(722)를 포함할 수 있으나 이에 제한되지는 않는다. 시스템 메모리(730) 및 그래픽 프로세서(790)는 메모리/그래픽 인터페이스(721)에 커플링될 수 있다. 모니터(791) 또는 다른 그래픽 출력 디바이스는 그래픽 프로세서(790)에 커플링될 수 있다.

[0137] 프로세서(720), 메모리/그래픽 인터페이스(721) 및 I/O 인터페이스(722) 사이의 고속 시스템 버스 (723), 메모리/그래픽 인터페이스(721)와 시스템 메모리(730) 사이의 전면 버스(724), 및 메모리/그래픽 인터페이스(721)와 그래픽 프로세서(790) 사이의 진보된 그래픽 프로세싱(AGP) 버스(725)를 비롯하여, 일련의 시스템 버스는 다양한 시스템 구성요소를 커플링시킬 수 있다. 시스템 버스(723)는 여러 유형의 버스 구조물 중 임의의 것을 포함할 수 있으며, 예로서 및 비제한적으로, 이러한 아키텍처는 산업 표준 아키텍처(USA) 버스, 마이크로 채널 아키텍처(MCA) 버스 및 향상된 ISA(EISA) 버스를 포함한다. 시스템 아키텍처가 진화됨에 따라, 다른 버스 아키텍처 및 칩 세트가 사용될 수 있지만, 종종 일반적으로 이 패턴을 따른다. 예를 들어, 인텔 및 AMD와 같은 회사는 각각 인텔 허브 아키텍처(IHA) 및 하이퍼트랜스포스(Hypertransport)™ 아키텍처를 지원한다.

[0138] 컴퓨터(710)는 전형적으로 다양한 컴퓨터-판독가능 매체를 포함한다. 컴퓨터-판독가능 매체는 컴퓨터(710)에 의해 액세스될 수 있고 휘발성 및 비휘발성 매체, 착탈식 및 비-착탈식 매체 둘 다를 포함하는 임의의 이용가능한 매체일 수 있다. 예로서 및 비제한적으로, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체를 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 다른 데이터와 같은 정보의 저장을 위해 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 착탈식 및 비-착탈식 매체 둘 다를 포함한다. 컴퓨터 저장 매체는 RAM, ROM, EEPROM, 플래시 메모리 또는 다른 메모리 기술, CD-ROM, 디지털 다기능 디스크(DVD) 또는 다른 광학 디스크 저장장치, 자기 카세트, 자기 테이프, 자기 디스크 저장장치 또는 다른 자기 저장 디바이스, 또는 원하는 정보를 저장하는데 사용될 수 있으며 컴퓨터(710)에 의해 액세스될 수 있는 임의의 다른 물리적 매체를 포함하나 이에 제한되지는 않는다.

[0139] 시스템 메모리(730)는 판독 전용 메모리(ROM)(731) 및 랜덤 액세스 메모리(RAM)(732)와 같은 휘발성 및/또는 비휘발성 메모리 형태의 컴퓨터 저장 매체를 포함한다. 시스템 ROM(731)은 정보를 식별하고 만들어내는 것과 같은 영구적인 시스템 데이터(743)를 함유할 수 있다. 일부 실시양태에서, 기본 입력/출력 시스템(BIOS)은 또한 시스템 ROM(731)에 저장될 수 있다. RAM(732)은 전형적으로 프로세서(720)에 의해 즉시 액세스가능하고/거나 현재 작동 중인 데이터 및/또는 프로그램 모듈을 함유한다. 예로서 및 비제한적으로 도 5는 작동 시스템(734), 애플리케이션 프로그램(735), 다른 프로그램 모듈(736) 및 프로그램 데이터(737)를 도시한다.

[0140] I/O 인터페이스(722)는 컴퓨터(710)에 다양한 내장형 및 외장형 디바이스를 커플링시키는 다수의 다른 버스(726, 727 및 728)와 시스템 버스(723)를 커플링시킬 수 있다. 직렬 주변 인터페이스(SPI) 버스(726)는 시동 동안과 같이 컴퓨터(710) 내의 요소들 사이의 정보 전달을 돕는 기본 루틴을 함유하는 기본 입력/출력 시스템(BIOS) 메모리(733)에 접속될 수 있다.

[0141] 슈퍼 입력/출력 칩(760)은 예로서 플로피 디스크(752), 키보드/마우스(762) 및 프린터(796)와 같은 다수의 '레거시' 주변장치에 접속하는데 사용될 수 있다. 슈퍼 I/O 칩(760)은, 일부 실시양태에서, 낮은 핀 카운트(LPC) 버스와 같은 버스(727)로 I/O 인터페이스(722)에 접속될 수 있다. 슈퍼 I/O 칩(760)의 다양한 실시양태는 상업적 시장에서 널리 입수가능하다. 한 실시양태에서, 버스(728)는 주변 구성요소 상호접속(PCI) 버스 또는 그의 변형체일 수 있으며, 보다 높은 속도의 주변장치를 I/O 인터페이스(722)에 접속하는데 사용될 수 있다. PCI 버

스는 또한 메자닌 버스로도 공지되어 있을 수 있다. PCI 버스의 변형은 주변 구성요소 상호접속-익스프레스(PCI-E) 및 주변 구성요소 상호접속--확장(PCI-X) 버스를 포함하며, 전자는 직렬 인터페이스를 갖고 후자는 역방향 호환가능 병렬 인터페이스이다. 다른 실시양태에서, 버스(728)는 직렬 ATA 버스(SATA) 또는 병렬 ATA(PATA)의 형태의 진보된 기술 부착(ATA) 버스일 수 있다.

[0142] 컴퓨터(710)는 또한 다른 착탈식/비-착탈식, 휘발성/비휘발성 컴퓨터 저장 매체를 포함할 수 있다. 단지 예로서, 도 7은 비-착탈식, 비휘발성 자기 매체로부터 판독하거나 그에 기록하는 하드 디스크 드라이브(740)를 도시한다. 하드 디스크 드라이브(740)는 통상적인 하드 디스크 드라이브일 수 있다. 범용 직렬 버스(USB) 메모리(753), 파이어와이어(IEEE 7394) 또는 CD/DVD 드라이브(756)와 같은 착탈식 매체는 PCI 버스(728)에 직접적으로 또는 인터페이스(750)를 통해 접속될 수 있다. 저장 매체(754)는 인터페이스(750)를 통해 커플링될 수 있다. 예시적인 작동 환경에서 사용될 수 있는 다른 착탈식/비-착탈식, 휘발성/비휘발성 컴퓨터 저장 매체는 자기 테이프 카세트, 플래시 메모리 카드, 디지털 다기능 디스크, 디지털 비디오 테이프, 고체 상태 RAM, 고체 상태 ROM 등을 포함하나 이에 제한되지는 않는다.

[0143] 상기에서 논의되고 도 7에 도시된 드라이브 및 그의 연관된 컴퓨터 저장 매체는 컴퓨터(710)를 위한 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 및 다른 데이터의 저장을 제공한다. 도 7에서, 예를 들어, 하드 디스크 드라이브(740)는 작동 시스템(744), 애플리케이션 프로그램(745), 다른 프로그램 모듈(746) 및 프로그램 데이터(747)를 저장하는 것으로 도시되어 있다. 이들 구성요소가 작동 시스템(734), 애플리케이션 프로그램(735), 다른 프로그램 모듈(736) 및 프로그램 데이터(737)와 동일하거나 상이할 수 있다는 것을 주목한다. 작동 시스템(744), 애플리케이션 프로그램(745), 다른 프로그램 모듈(746) 및 프로그램 데이터(747)는 최소한 이들이 상이한 카피라는 것을 예시하기 위해 여기서 상이한 번호가 주어진다. 사용자는 마우스/키보드(762) 또는 다른 입력 디바이스 조합과 같은 입력 디바이스를 통해 명령 및 정보를 컴퓨터 내로 입력할 수 있다. 다른 입력 디바이스 (제시되지 않음)는 마이크로폰, 조이스틱, 게임 패드, 위성 안테나, 스캐너 등을 포함할 수 있다. 이들 및 다른 입력 디바이스는 종종 SPI(726), LPC(727) 또는 PCI(728)와 같은 I/O 인터페이스 버스 중 하나를 통해 프로세서(720)에 접속되지만, 다른 버스도 사용될 수 있다. 일부 실시양태에서, 다른 디바이스가 슈퍼 I/O 칩(760)을 통해 병렬 포트, 적외선 인터페이스, 게임 포트 등 (도시되지 않음)에 커플링될 수 있다. 컴퓨터(710)는 네트워크 인터페이스 제어기(NIC)(770)를 통해 원격 컴퓨터(780)와 같은 1개 이상의 원격 컴퓨터에 대한 논리적 접속을 사용하여 네트워킹 환경에서 작동할 수 있다. 원격 컴퓨터(780)는 개인용 컴퓨터, 서버, 라우터, 네트워크 PC, 피어 디바이스 또는 다른 공통 네트워크 노드일 수 있고, 전형적으로 컴퓨터(710)에 대하여 상기 기재된 많은 또는 모든 요소를 포함한다. 도 7에 도시된 NIC(770)와 원격 컴퓨터(780) 사이의 논리적 접속은 로컬 영역 네트워크(LAN), 광역 네트워크(WAN) 또는 둘 다를 포함할 수 있지만, 또한 다른 네트워크도 포함할 수 있다. 이러한 네트워킹 환경은 사무실, 기업의 컴퓨터 네트워크, 인트라넷 및 인터넷에서 통상적이다. 원격 컴퓨터(780)는 또한 컴퓨터(710)와의 상호작용 세션을 지지하는 웹 서버를 나타낼 수 있거나, 위치-기반 애플리케이션의 특정 경우에는 위치 서버 또는 애플리케이션 서버일 수 있다. 원격 컴퓨터(780)는 기준선 게놈 서열 데이터베이스(784), 이전 데이터베이스(782) 등을 관리하는 서버를 나타낼 수 있다. 예를 들어, 본원에서 설명된 바와 같이, 기준선 게놈 서열 데이터베이스(784)는 새로운 기준선 정보로 주기적으로 업데이트된다. 데이터베이스(784)는 샘플에 대한 잠재적 변이체 호출을 검증하는 것과 관련하여 (예를 들어, 지정된 위치에서 기준선 변이체 빈도를 수득하기 위해) 액세스된다. 또 다른 예로서, 이전 데이터베이스(782)는 또한 샘플에 대한 잠재적 변이체 호출을 검증하는 것과 관련하여 (예를 들어, 잠재적 변이체 호출이 이전 변이체에 상응하는지 여부를 결정하기 위해) 액세스될 수 있다.

[0144] 일부 실시양태에서, 네트워크 인터페이스는 광대역 접속이 이용가능하지 않거나 사용되지 않는 경우에 모뎀을 사용할 수 있다 (도시되지 않음). 제시된 네트워크 접속은 예시적이고, 컴퓨터 사이에 통신 연결을 확립하는 다른 수단이 사용될 수 있는 것으로 인지될 것이다. 예시적인 프로세서(프로세싱 유닛)는 컴퓨팅 디바이스에 사용되는 모든 다양한 마이크로프로세서 및 다른 프로세싱 유닛을 포함한다. 예시적인 컴퓨터-판독가능 매체는 상기에 기재되어 있다. 시스템의 2개 이상의 구성요소가 프로세서 또는 컴퓨터-판독가능 매체를 포함하는 경우, 일반적으로 단일 프로세서 및/또는 컴퓨터 판독가능 매체가 시스템의 단일 구성요소에 전용이거나; 또는 2개 이상의 기능이 단일 프로세서를 공유하고/거나 단일 컴퓨터 판독가능 매체를 공유하여 시스템이 적게는 1개의 프로세서 및/또는 1개의 컴퓨터 판독가능 매체를 함유하도록 하는 시스템이 생성될 수 있다. 일부 변형에서, 예를 들어 상이한 위치에서 시스템의 구성요소를 갖는 것이 편리한 경우, 다수의 프로세서 또는 매체를 사용하는 것이 유리하다. 예를 들어, 시스템의 일부 구성요소는 실험실 또는 데이터 분석에 전용인 시험 실험실에 위치할 수 있고, 반면 입력 정보를 공급하거나 출력 통신을 수득하기 위한 구성요소 (임의적임)를 포함한 다른 구성요소는 시험 서비스가 수행되는 인간 대상체 (환자)의 의학적 치료 또는 권고 시설 (예를 들어, 의

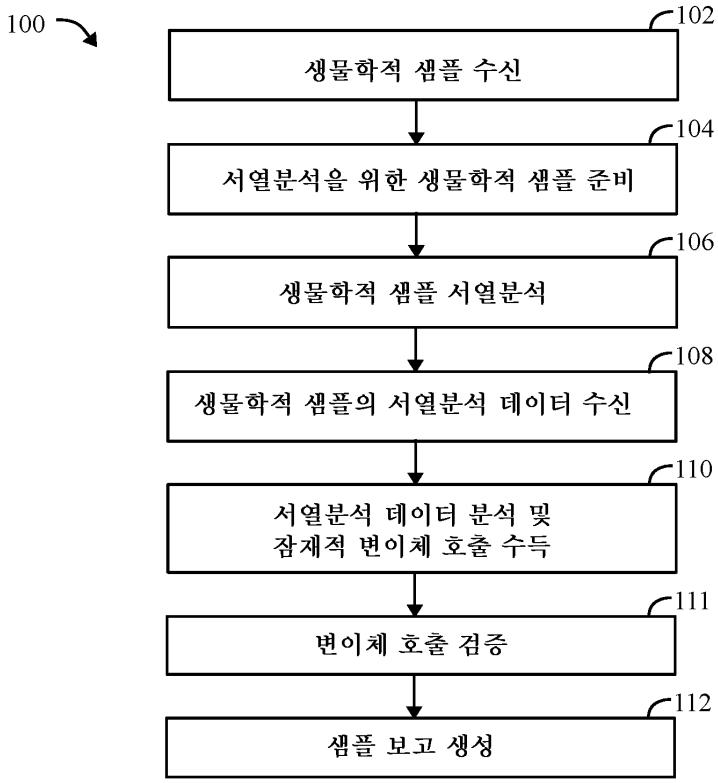
사의 진료실, 건강 클리닉, HMO, 약사, 유전학자, 병원) 및/또는 가정 또는 사업체에 위치할 수 있다.

[0145] 원격 컴퓨터(780)는 이전 데이터베이스(782)에 통신 커플링되고 인간 집단에서 변이체/대립유전자의 존재 또는 부재와 상관관계가 있는 집단 정보를 함유하는 서버 또는 다른 네트워크 자원을 나타낼 수 있다. 예를 들어, 1개 이상의 변이체/대립유전자는 선택 결함을 유발하거나 그를 나타내는 돌연변이체 대립유전자를 포함한다. 간단한 변형에서, 이전 데이터베이스(782)는, 예를 들어 방광암을 갖는 인간 집단, 및 예를 들어 방광암이 없는 인간 집단에서 선택 대립유전자가 관찰된 빈도에 관한 데이터를 함유한다. 추가적으로 또는 대안적으로, 이전 데이터베이스는 2개 이상의 대립유전자와 관련한 유사한 데이터를 포함하여, 이에 의해 인간 대상체가 2개 이상의 대립유전자 중 임의의 것을 갖는 경우에 유용한 참조를 제공할 수 있다. 추가적으로 또는 대안적으로, 이전 데이터베이스는 병태가 있거나 없는 것으로 진단된 개체에 관한 추가의 정량적인 개인, 의학적 또는 유전적 정보를 데이터베이스에 포함할 수 있다. 이러한 정보는 인간에서의 연령, 성별, 민족, 인종, 병력, 체중, 당뇨병 상태, 혈압, 방광암의 가족력, 흡연 이력 및 알콜 사용과 같은 파라미터에 관한 정보, 및 이전 경우에 관한 적어도 1개의 파라미터의 영향을 포함하나 이에 제한되지는 않는다. 이전 경우는 또한 대립유전자에 대한 다른 유전적 위험 인자에 관한 정보를 포함할 수 있다.

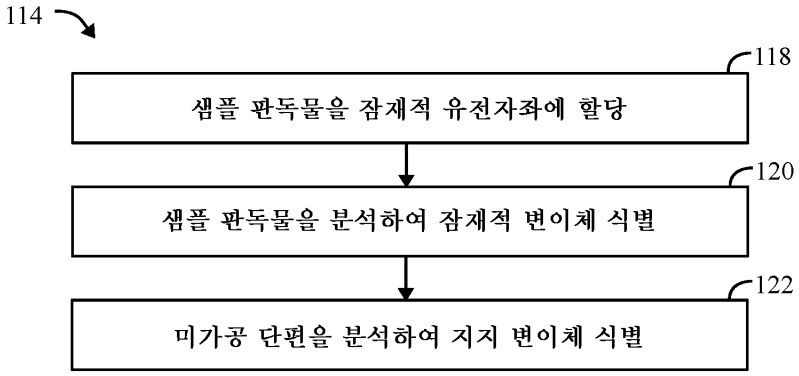
[0146] 기준선 변이체 빈도 및 다른 정보를 식별하기 위해, 본원의 실시양태에 따라 이용되는 기준선 게놈 서열을 저장하기 위해 기준선 게놈 서열 데이터베이스 또는 라이브러리(784)가 제공될 수 있다.

도면

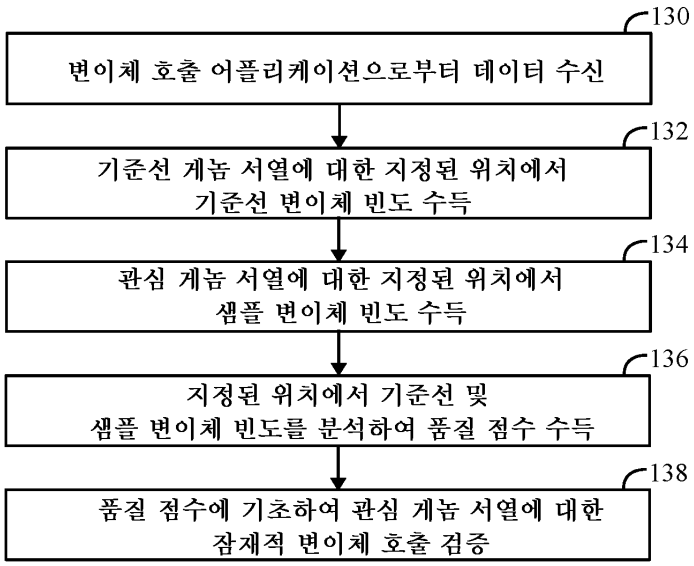
도면1a



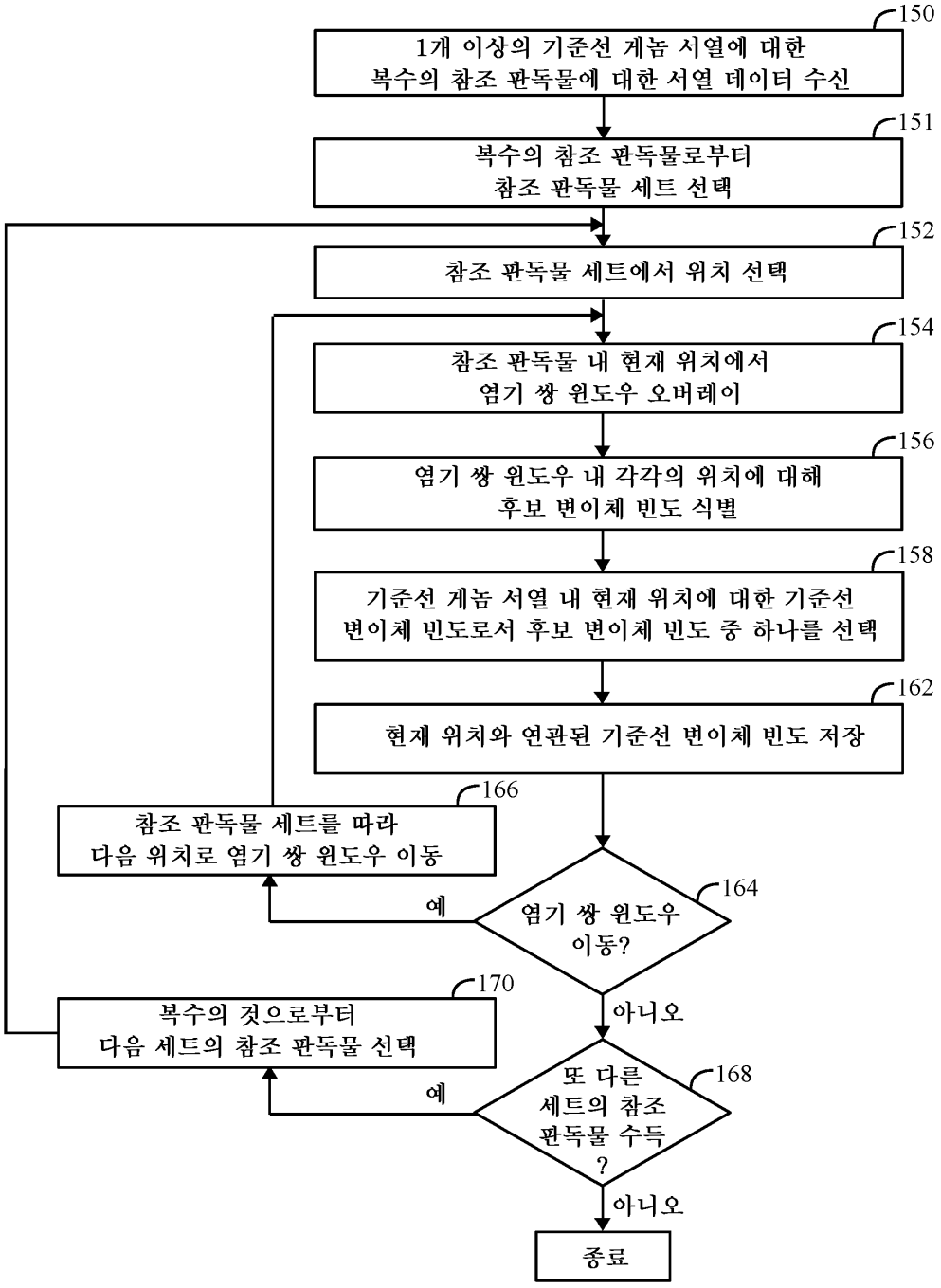
도면1b


114
118
샘플 판독물을 잠재적 유전자좌에 할당
120
샘플 판독물을 분석하여 잠재적 변이체 식별
122
미가공 단편을 분석하여 지지 변이체 식별

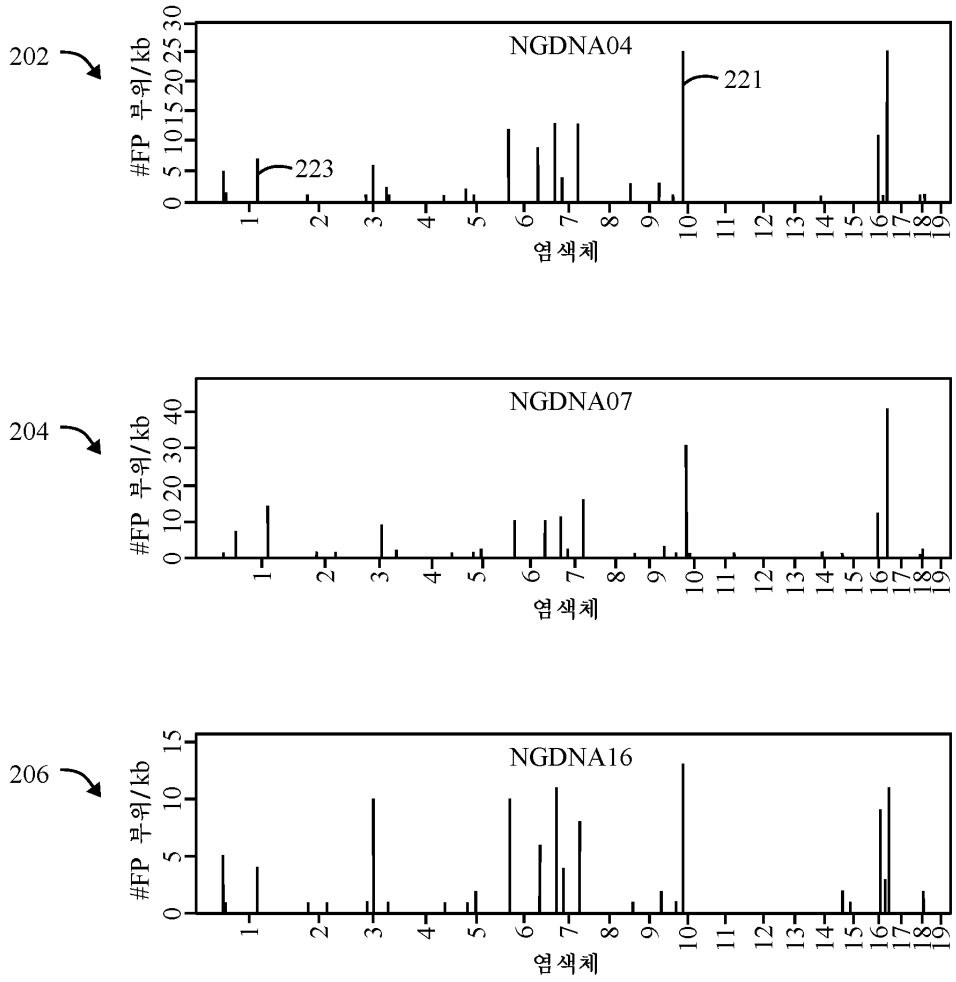
도면1c


130
변이체 호출 어플리케이션으로부터 데이터 수신
132
기준선 게놈 서열에 대한 지정된 위치에서
기준선 변이체 빈도 수득
134
관심 게놈 서열에 대한 지정된 위치에서
샘플 변이체 빈도 수득
136
지정된 위치에서 기준선 및
샘플 변이체 빈도를 분석하여 품질 점수 수득
138
품질 점수에 기초하여 관심 게놈 서열에 대한
잠재적 변이체 호출 검증

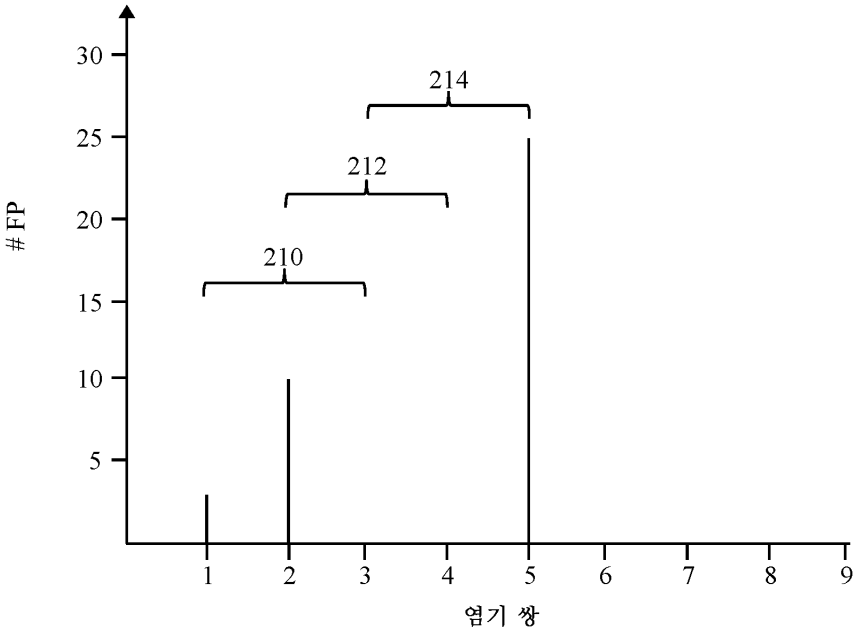
도면1d


150
1개 이상의 기준선 게놈 서열에 대한
복수의 참조 판독물에 대한 서열 데이터 수신
151
복수의 참조 판독물로부터
참조 판독물 세트 선택
152
참조 판독물 세트에서 위치 선택
154
참조 판독물 내 현재 위치에서
염기 쌍 윈도우 오버레이
156
염기 쌍 윈도우 내 각각의 위치에 대해
후보 변이체 빈도 식별
158
기준선 게놈 서열 내 현재 위치에 대한 기준선
변이체 빈도로서 후보 변이체 빈도 중 하나를 선택
162
현재 위치와 연관된 기준선 변이체 빈도 저장
164
염기 쌍 윈도우
이동?
예
166
참조 판독물 세트를 따라
다음 위치로 염기 쌍 윈도우 이동
아니오
168
또 다른
세트의 참조
판독물 수득
?
예
170
복수의 것으로부터
다음 세트의 참조 판독물 선택
아니오
종료

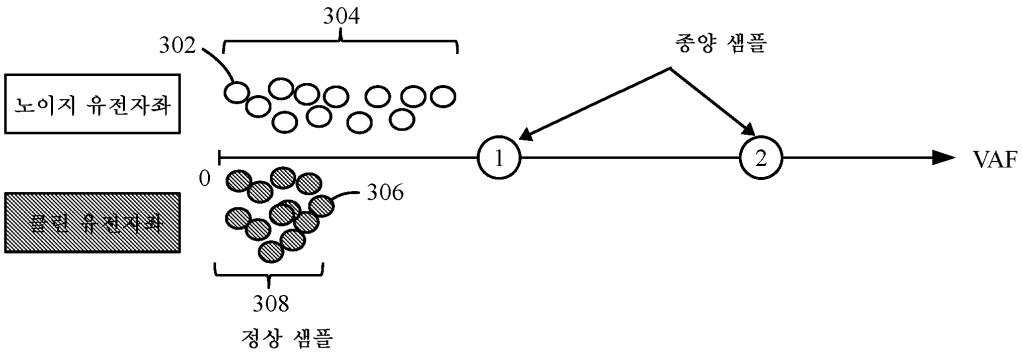
도면2a


202
NGDNA04
221
223
#FP 부위/kb
0 5 10 15 20 25 30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
염색체
204
NGDNA07
#FP 부위/kb
0 10 20 30 40
염색체
206
NGDNA16
#FP 부위/kb
0 5 10 15
염색체

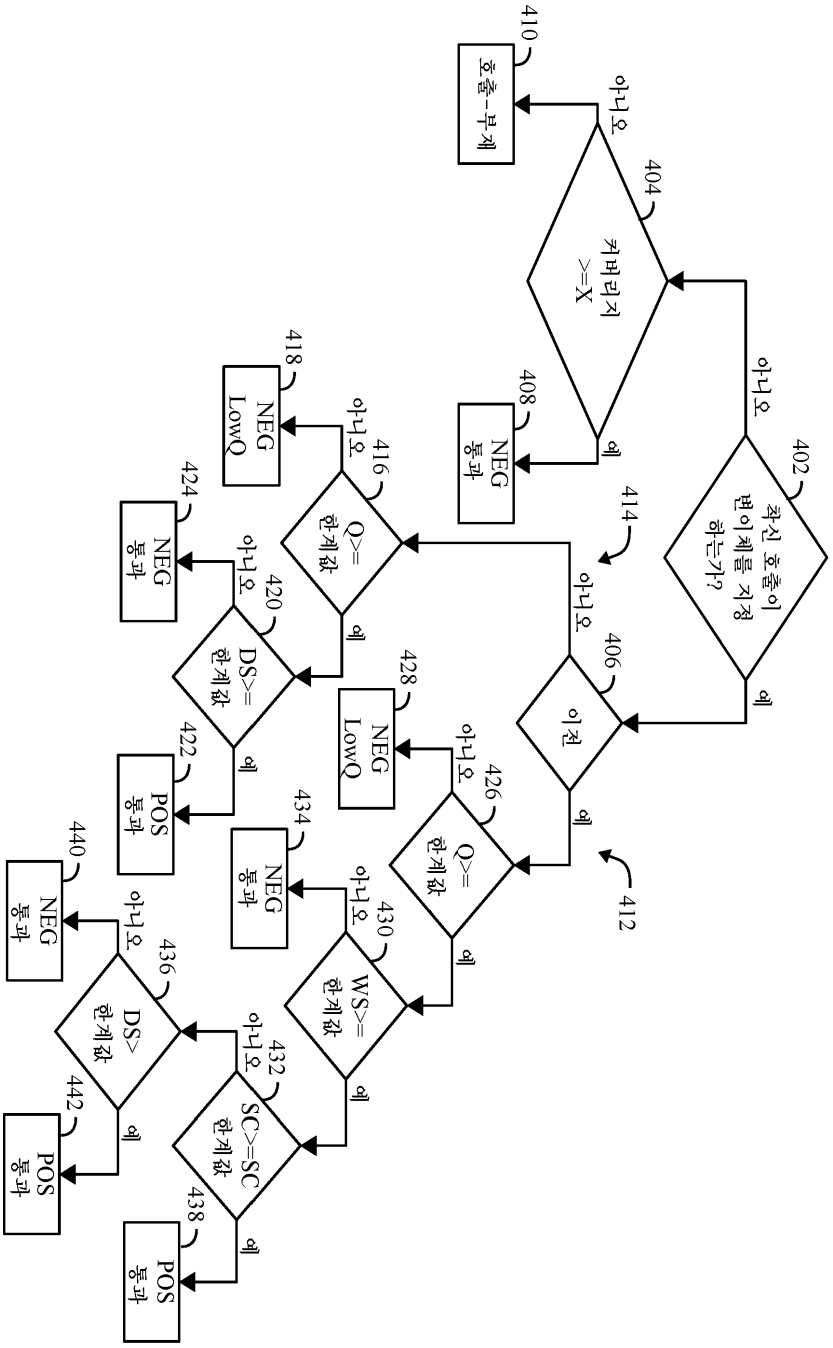
도면2b


214
212
210
FP
30 25 20 15 10 5
1 2 3 4 5 6 7 8 9
염기 쌍

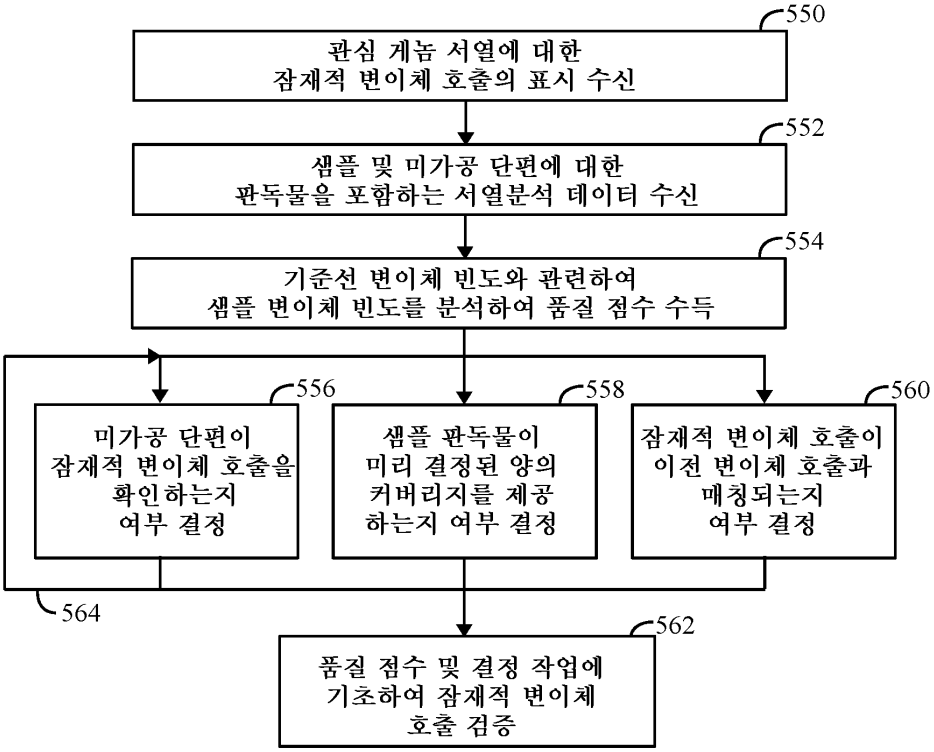
도면3



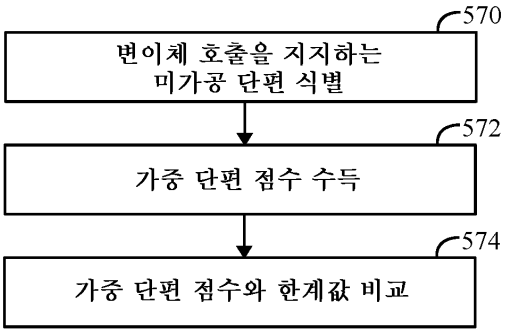
도면4


402
착신 호출이 변이체를 지정하는가?
아니오
예
404
커버리지 >=X
아니오
예
406
이전
408
NEG 통과
410
호출-부재
412
414
아니오
예
416
Q>= 한계값
아니오
예
418
NEG LowQ
420
DS>= 한계값
아니오
예
422
POS 통과
424
NEG 통과
426
Q>= 한계값
아니오
예
428
NEG LowQ
430
WS>= 한계값
아니오
예
432
SC>=SC 한계값
아니오
예
434
NEG 통과
436
DS> 한계값
아니오
예
438
POS 통과
440
NEG 통과
442
POS 통과

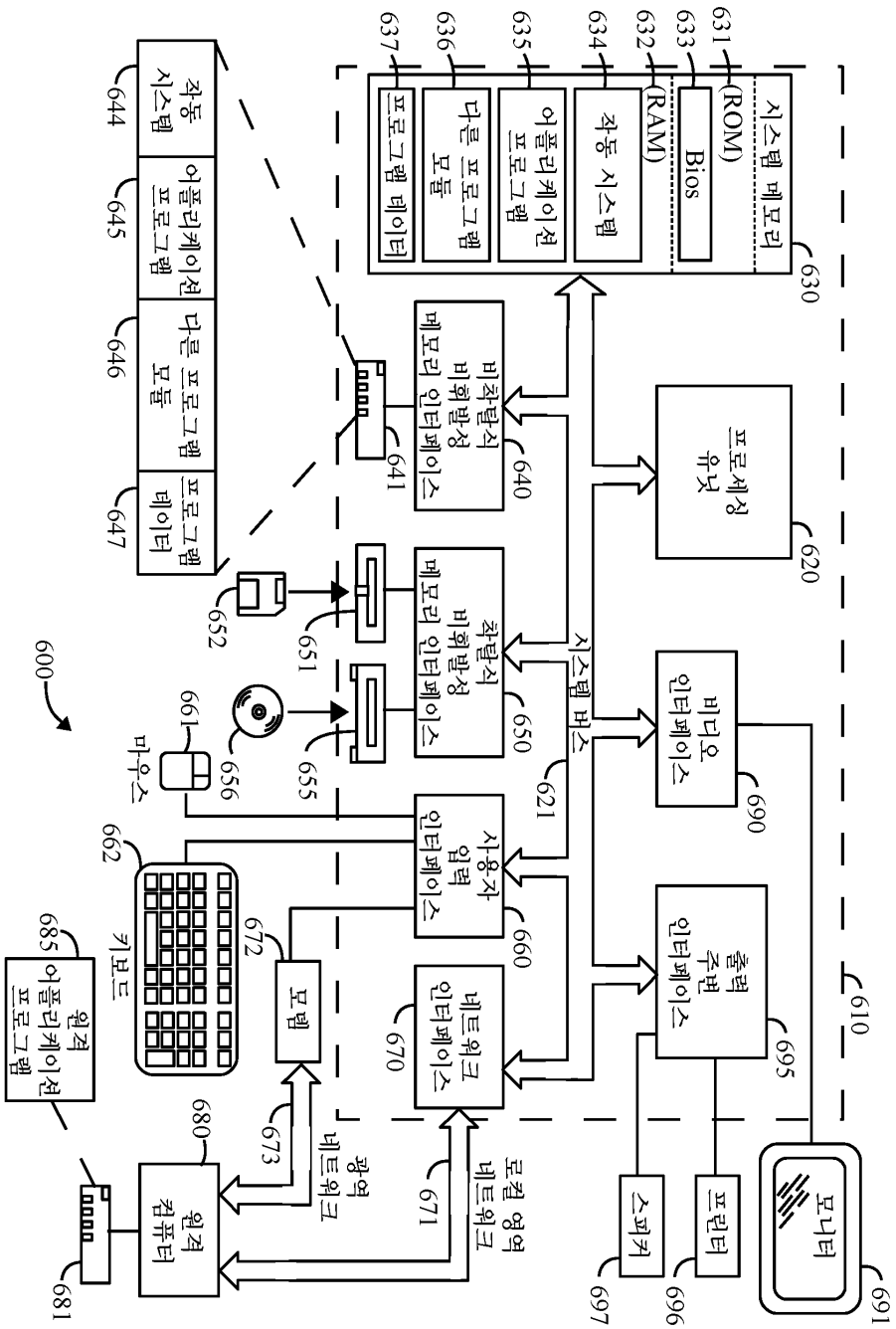
도면5a


550
관심 게놈 서열에 대한
잠재적 변이체 호출의 표시 수신
552
샘플 및 미가공 단편에 대한
판독물을 포함하는 서열분석 데이터 수신
554
기준선 변이체 빈도와 관련하여
샘플 변이체 빈도를 분석하여 품질 점수 수득
556
미가공 단편이
잠재적 변이체 호출을
확인하는지
여부 결정
558
샘플 판독물이
미리 결정된 양의
커버리지를 제공
하는지 여부 결정
560
잠재적 변이체 호출이
이전 변이체 호출과
매칭되는지
여부 결정
564
562
품질 점수 및 결정 작업에
기초하여 잠재적 변이체
호출 검증

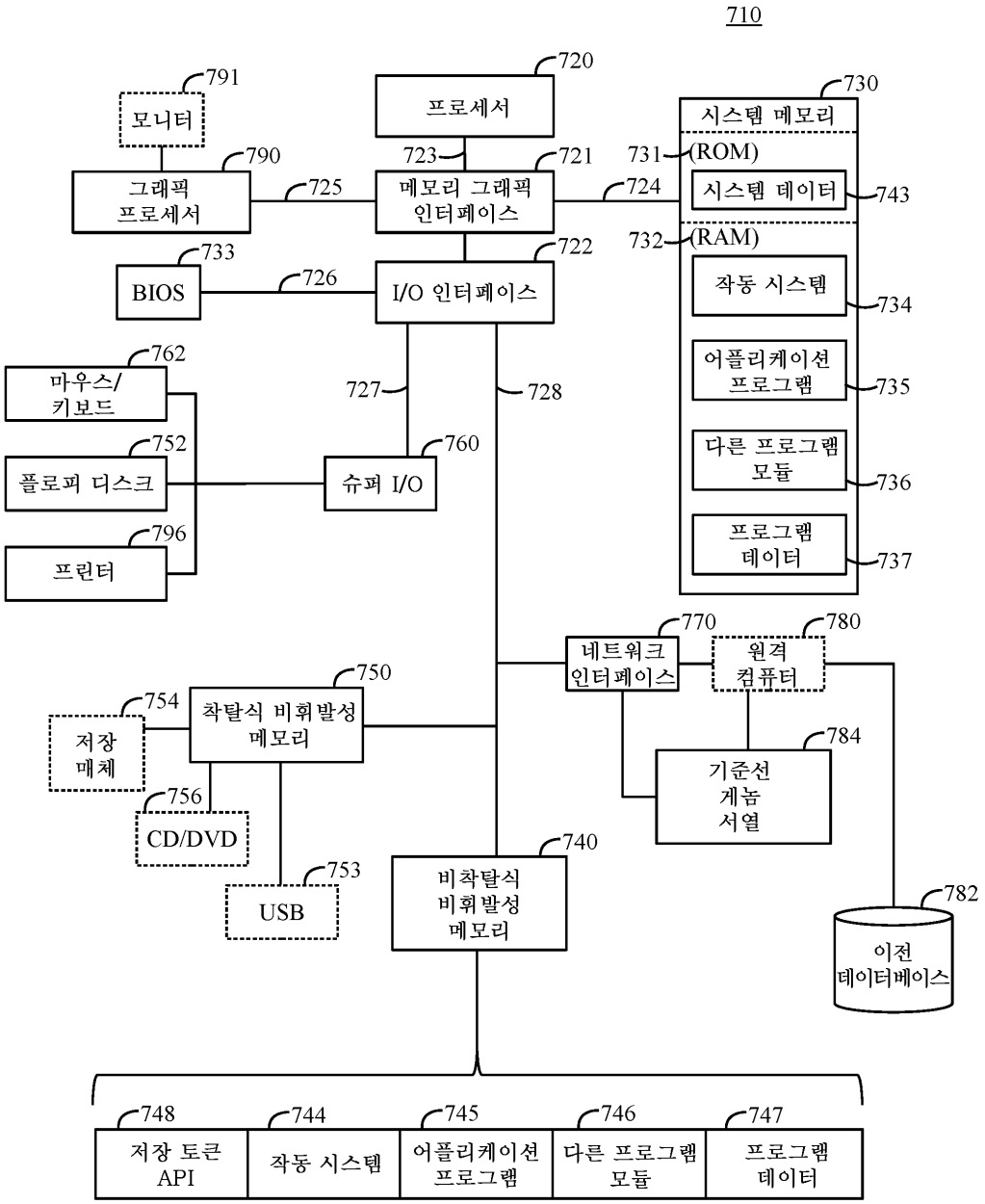
도면5b


570
변이체 호출을 지지하는
미가공 단편 식별
572
가중 단편 점수 수득
574
가중 단편 점수와 한계값 비교

도면6


600
610
620
621
630
631
632
633
634
635
636
637
시스템 메모리
(ROM)
Bios
(RAM)
작동 시스템
어플리케이션 프로그램
다른 프로그램 모듈
프로그램 데이터
프로세싱 유닛
비디오 인터페이스
출력 주변 인터페이스
시스템 버스
비착탈식 비휘발성 메모리 인터페이스
착탈식 비휘발성 메모리 인터페이스
사용자 입력 인터페이스
네트워크 인터페이스
모니터
프린터
스피커
모뎀
키보드
마우스
원격 컴퓨터
원격 어플리케이션 프로그램
광역 네트워크
로컬 영역 네트워크
641
640
650
651
652
655
656
660
661
662
670
671
672
673
680
681
685
690
691
695
696
697
644
645
646
647

도면7


710
720
프로세서
791
모니터
790
그래픽 프로세서
725
723
721
메모리 그래픽 인터페이스
724
731
(ROM)
730
시스템 메모리
시스템 데이터
743
732
(RAM)
733
BIOS
726
722
I/O 인터페이스
작동 시스템
734
762
마우스/ 키보드
727
728
어플리케이션 프로그램
735
752
플로피 디스크
760
슈퍼 I/O
다른 프로그램 모듈
736
796
프린터
프로그램 데이터
737
770
780
네트워크 인터페이스
원격 컴퓨터
750
754
착탈식 비휘발성 메모리
저장 매체
784
기준선 계놈 서열
756
CD/DVD
753
USB
740
비착탈식 비휘발성 메모리
782
이전 데이터베이스
748
744
745
746
747
저장 토큰 API
작동 시스템
어플리케이션 프로그램
다른 프로그램 모듈
프로그램 데이터