(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년12월18일
(11) 등록번호 10-2192864
(24) 등록일자 2020년12월14일

(51) 국제특허분류(Int. Cl.)
G16B 20/20 (2019.01) *G16B 20/40* (2019.01)
G16B 40/00 (2019.01)
(52) CPC특허분류
G16B 20/20 (2019.02)
G16B 20/40 (2019.02)
(21) 출원번호 10-2019-0036490
(22) 출원일자 2019년03월29일
심사청구일자 2019년03월29일
(65) 공개번호 10-2020-0114546
(43) 공개일자 2020년10월07일
(56) 선행기술조사문헌
Bioinformatics (2016) 32(17):2699-2701*
(뒷면에 계속)
전체 청구항 수 : 총 23 항

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
김상우
서울특별시 마포구 신촌로 170 이대역푸르지오시티 917호
전혜인
경기도 고양시 일산동구 탄중로 398 중산마을8단지아파트 805동 401호
(74) 대리인
특허법인인벤싱크

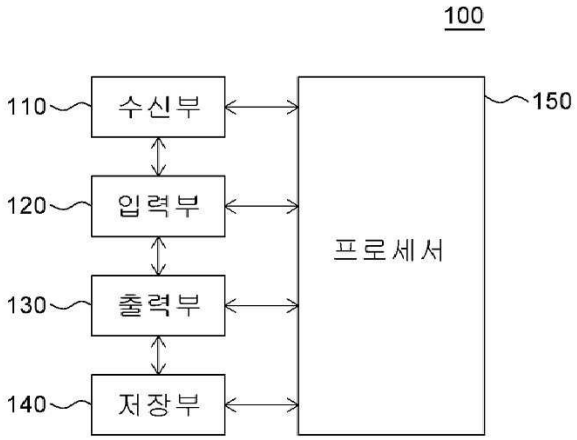
심사관 : 김승범

(54) 발명의 명칭 **NGS 샘플 검증 방법 및 이를 이용한 디바이스**

(57) 요 약

본 명세서에서는, 프로세서를 포함하는 NGS 디바이스에 구현되는, NGS 샘플 검증 방법으로서, 미리 결정된 표적 SNP (single-nucleotide polymorphism) 사이트에 대한 서열 정보를 포함하는, 대상 샘플에 대한 복수의 NGS 데이터를 수신하는 단계, 표적 SNP 사이트에 대한 서열 정보를 기초로, 복수의 NGS 데이터 중 동일 개체에서 유래한 샘플 여부를 결정하는 단계, 및 동일 개체 유래 샘플 여부를 제공하는 단계를 포함하는 NGS 샘플 검증 방법 및 이를 이용한 디바이스를 제공한다.

대 표 도 - 도1a



(52) CPC특허분류
G16B 25/10 (2019.02)
G16B 40/00 (2019.02)

(56) 선행기술조사문헌
Nucleic Acids Research (2017) 45(11):e103*
Bioinformatics (2019.06.14.) 35(22):4806-4808
KR1020170098648 A
전혜인, "Development of sample mismatch detection algorithm in genome sequencing cohorts", 석사학위논문, 연세대학교 (2019.08.)
KR1020170133079 A
*는 심사관에 의하여 인용된 문헌

이 발명을 지원한 국가연구개발사업

과제고유번호	2016M3A9B6903439
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	바이오.의료기술개발사업
연구과제명	멀티오믹스 분석 기반 반려동물-인간 공통 비침습적 암 진단 기술 개발
기 여 율	1/1
과제수행기관명	연세대학교
연구기간	2016.05.01 ~ 2021.10.31

명 세 서

청구범위

청구항 1

프로세서를 포함하는 NGS (Next Generation Sequencing) 디바이스에 구현되는, NGS 샘플 검증 방법으로서,

미리 결정된 표적 SNP (single-nucleotide polymorphism) 사이트에 대한 서열 정보를 포함하는, 대상 샘플에 대한 복수의 NGS 데이터를 수신하는 단계;

상기 표적 SNP 사이트에 대한 서열 정보를 기초로, 상기 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계, 및

상기 동일 개체 유래 샘플 여부를 제공하는 단계를 포함하고,

상기 복수의 NGS 데이터는 미리 결정된 파일명을 포함하고,

상기 파일명 및 상기 표적 SNP 사이트에 대한 서열 정보를 기초로, 상기 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계를 더 포함하는, NGS 샘플 검증 방법.

청구항 2

제1항에 있어서,

상기 복수의 NGS 데이터는, WGS (whole genome sequencing), WES (whole exome sequencing) 파일, RNA 시퀀싱 (RNA sequencing) 파일, 및 표적 시퀀싱 (targeted sequencing) 파일 중 적어도 하나의 NGS 분석 방법에 따른 파일을 포함하고,

상기 복수의 NGS 데이터를 수신하는 단계 이후에,

GMAF (global minor allele frequency) 의 수준, 개별 인구 내의 MAF, 및 SNP 사이트의 맵핑 능력 (mappability) 중 적어도 하나, 및 상기 NGS 분석 데이터의 종류를 기초로, 표적 SNP 사이트를 결정하는 단계를 더 포함하는, NGS 샘플 검증 방법.

청구항 3

제2항에 있어서,

상기 복수의 NGS 데이터가, 상기 WGS 분석 파일, 또는 상기 WES 분석 파일, 또는 상기 RNA 시퀀싱 분석 파일일 경우,

상기 표적 SNP 사이트를 결정하는 단계는,

상기 맵핑 능력이 미리 결정된 수준 이상인 SNP 사이트 중, 상기 GMAF가 0.45 내지 0.55인 SNP 사이트, 또는 상기 개별 인구 내의 MAF가 0.1 내지 0.9인 SNP 사이트를 상기 표적 SNP 사이트로 결정하는 단계를 포함하는, NGS 샘플 검증 방법.

청구항 4

제2항에 있어서,

상기 복수의 NGS 데이터가 상기 표적 시퀀싱 분석 파일일 경우,

상기 표적 SNP 사이트를 결정하는 단계는,

상기 맵핑 능력이 미리 결정된 수준 이상인 SNP 사이트 중, 상기 개별 인구 내의 GMAF가 0.1 내지 0.9인 SNP 사이트를 상기 표적 SNP 사이트로 결정하는 단계를 포함하는, NGS 샘플 검증 방법.

청구항 5

제1항에 있어서,

상기 동일 개체 유래 샘플 여부를 결정하는 단계는,

상기 복수의 NGS 데이터 각각의 상기 표적 SNP 사이트에 대한 서열 정보를 기초로, 유전자형 일치 점수(genotype concordance score)를 산출하는 단계, 및

상기 유전자형 일치 점수를 기초로 상기 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계 포함하고,

상기 제공하는 단계는,

상기 복수의 NGS 데이터 중 동일 개체 유래 샘플로 결정된 NGS 데이터를 제공하는 단계를 포함하는, NGS 샘플 검증 방법.

청구항 6

제5항에 있어서,

상기 유전자형 일치 점수를 산출하는 단계는,

상기 복수의 NGS 데이터 중 선택된 두 개의 NGS 데이터의 상기 표적 SNP 사이트에 대한 서열의 일치율을 산출하는 단계를 포함하고,

상기 유전자형 일치 점수를 기초로 상기 동일 개체 유래 샘플 여부를 결정하는 단계는,

상기 두 개의 NGS 데이터의 표적 SNP 사이트에 대한 서열의 일치율이 0.7 이상인 경우, 상기 두 개의 NGS 데이터를 하나의 개체로부터 유래한 것으로 결정하는 단계를 더 포함하는, NGS 샘플 검증 방법.

청구항 7

제5항에 있어서,

상기 유전자형 일치 점수를 산출하는 단계는,

상기 복수의 NGS 데이터 중 선택된 두 개의 NGS 데이터의 상기 표적 SNP 사이트에 대한 서열의 일치율을 산출하는 단계를 포함하고,

상기 두 개의 NGS 데이터의 표적 SNP 사이트에 대한 서열의 일치율이 0.7 미만인 경우, 상기 두 개의 NGS 데이터를 샘플과 매칭되지 않는 NGS 데이터로 결정하는 단계를 더 포함하는, NGS 샘플 검증 방법.

청구항 8

제7항에 있어서,

상기 제공하는 단계는,

상기 샘플과 매칭되지 않는 NGS 데이터를 더 제공하는 단계를 포함하는, NGS 샘플 검증 방법.

청구항 9

삭제

청구항 10

제1항에 있어서,

상기 파일명을 기초로, 상기 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계는,

상기 파일명을 기초로, 상기 복수의 NGS 데이터에 대한 유사도 점수를 산출하는 단계, 및

상기 유사도 점수를 기초로 상기 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계를 포함하는, NGS 샘플 검증 방법.

청구항 11

프로세서를 포함하는 NGS 디바이스에 구현되는 NGS 샘플 검증 방법으로서,

복수의 부분으로 이루어진 파일명을 포함하는, 대상 샘플에 대한 복수의 NGS 데이터를 수신하는 단계;

상기 복수의 NGS 데이터 중 선택된 두 개의 NGS 데이터 각각의 상기 복수의 부분에 대한 유사도 점수를 산출하는 단계;

상기 유사도 점수를 기초로 상기 두 개의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계, 및

상기 동일 개체 유래 샘플 여부를 제공하는 단계를 포함하는, NGS 샘플 검증 방법.

청구항 12

제11항에 있어서,

상기 파일명은 서로 상이한 구분 문자를 포함하고,

상기 유사도 점수를 산출하는 단계는,

상기 구분 문자를 기초로, 상기 두 개의 NGS 데이터 각각의 상기 파일명을 복수의 부분으로 분할하는 단계;

상기 복수의 부분에 대한 값의 출연 빈도를 산출하는 단계, 및

상기 출연 빈도를 기초로 상기 두 개의 NGS 데이터의 유사도 점수를 산출하는 단계를 포함하는, NGS 샘플 검증 방법.

청구항 13

제11항에 있어서,

상기 파일명은 서로 상이한 구분 문자를 포함하고,

상기 유사도 점수를 산출하는 단계는,

상기 구분 문자를 기초로, 상기 두 개의 NGS 데이터 각각의 상기 파일명을 복수의 부분으로 분할하는 단계;

상기 복수의 부분에 대한 값의 분산 정도를 산출하는 단계, 및

상기 분산 정도를 기초로 상기 두 개의 NGS 데이터의 유사도 점수를 산출하는 단계를 포함하는, NGS 샘플 검증 방법.

청구항 14

제13항에 있어서,

상기 분산 정도를 산출하는 단계는,

상기 복수의 부분에 대한 값의 출연 빈도를 산출하는 단계,

상기 출연 빈도를 기초로 분산 정도를 산출하는 단계를 포함하는, NGS 샘플 검증 방법.

청구항 15

제13항에 있어서,

상기 유사도 점수를 산출하는 단계는,

상기 두 개의 NGS 데이터에 대하여, 상기 복수의 부분 중 선택된 하나의 부분에 대한 값의 동일 여부 및 상기 분산 정도를 기초로, 상기 유사도 점수를 산출하는 단계를 포함하는, NGS 샘플 검증 방법.

청구항 16

제11항에 있어서,

상기 동일 개체 유래 샘플 여부를 결정하는 단계는,

상기 두 개의 NGS 데이터에 대하여 산출된 유사도 점수가, 상기 복수의 NGS 데이터 중 선택된 NGS 데이터 쌍의

유사도 점수 중 가장 가장 높을 경우,

상기 두 개의 NGS 데이터를 동일 개체 유래 샘플로 결정하는 단계를 더 포함하는, NGS 샘플 검증 방법.

청구항 17

제11항에 있어서,

상기 두 개의 NGS 데이터는, 미리 결정된 표적 SNP 사이트에 대한 서열 정보를 포함하고,

상기 SNP 사이트에 대한 서열 정보를 기초로 상기 두 개의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계를 더 포함하는, NGS 샘플 검증 방법.

청구항 18

제17항에 있어서,

상기 SNP 사이트에 대한 서열 정보를 기초로, 상기 두 개의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계는,

상기 두 개의 NGS 데이터 각각의 상기 표적 SNP 사이트에 대한 서열 정보를 기초로, 유전자형 일치 점수를 산출하는 단계, 및

상기 유전자형 일치 점수를 기초로 상기 두 개의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계를 포함하는, NGS 샘플 검증 방법.

청구항 19

프로세서를 포함하는 NGS 디바이스에 구현되는 NGS 샘플 검증 방법으로서,

미리 결정된 표적 SNP 사이트에 대한 서열 정보 및 미리 결정된 파일명을 포함하는, 대상 샘플에 대한 복수의 NGS 데이터를 수신하는 단계;

상기 표적 SNP 사이트에 대한 서열 정보를 기초로, 상기 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 제1 결정하는 단계;

상기 파일명을 기초로, 상기 복수의 NGS 데이터에 대한 유사도 점수를 산출하는 단계, 및

상기 유사도 점수를 기초로 상기 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 제2 결정하는 단계를 포함하는, NGS 샘플 검증 방법.

청구항 20

NGS 샘플 검증용 디바이스로서,

미리 결정된 표적 SNP 사이트에 대한 서열 정보를 포함하는, 대상 샘플에 대한 복수의 NGS 데이터를 수신하도록 구성된 수신부, 및

상기 수신부와 통신하도록 구성된 프로세서를 포함하고,

상기 프로세서는, 상기 표적 SNP 사이트에 대한 서열 정보를 기초로, 상기 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하고, 상기 동일 개체 유래 샘플 여부를 제공하도록 구성되고,

상기 복수의 NGS 데이터는 미리 결정된 파일명을 더 포함하고,

상기 프로세서는, 상기 표적 SNP 사이트에 대한 서열 정보 및 상기 파일명을 기초로, 상기 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하도록 더 구성된, NGS 샘플 검증용 디바이스.

청구항 21

제20항에 있어서,

상기 복수의 NGS 데이터는 WGS 파일, WES 파일, RNA 시퀀싱 파일, 및 표적 시퀀싱 파일 중 적어도 하나의 NGS 분석 방법에 따른 파일을 포함하고,

상기 프로세서는,

GMAF의 수준, 개별 인구 내의 MAF, 및 SNP 사이트의 맵핑 능력 중 적어도 하나, 및 상기 NGS 분석 데이터의 종류 기초로, 상기 표적 SNP 사이트를 결정하도록 더 구성된, NGS 샘플 검증용 디바이스.

청구항 22

NGS 샘플 검증용 디바이스로서,

복수의 부분으로 이루어진 파일명을 포함하는, 대상 샘플에 대한 복수의 NGS 데이터를 수신하도록 구성된 수신부, 및

상기 수신부와 통신하도록 구성된 프로세서를 포함하고,

상기 프로세서는,

상기 복수의 NGS 데이터 중 선택된 두 개의 NGS 데이터 각각의 상기 복수의 부분에 대한 유사도 점수를 산출하고, 상기 유사도 점수를 기초로 상기 두 개의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하고, 상기 동일 개체 유래 샘플 여부를 제공하도록 구성된, NGS 샘플 검증용 디바이스.

청구항 23

제22항에 있어서,

상기 파일명은 서로 상이한 구분 문자를 포함하고,

상기 프로세서는,

상기 구분 문자를 기초로, 상기 두 개의 NGS 데이터 각각의 상기 파일명을 복수의 부분으로 분할하고, 상기 복수의 부분에 대한 값의 출연 빈도를 산출하고, 상기 출연 빈도를 기초로 상기 두 개의 NGS 데이터에 대한 유사도 점수를 산출하도록 더 구성된, NGS 샘플 검증용 디바이스.

청구항 24

제22항에 있어서,

상기 파일명은 서로 상이한 구분 문자를 포함하고,

상기 프로세서는,

상기 구분 문자를 기초로, 상기 두 개의 NGS 데이터 각각의 상기 파일명을 복수의 부분으로 분할하고, 상기 복수의 부분에 대한 값의 분산 정도를 산출하고, 상기 분산 정도를 기초로 상기 두 개의 NGS 데이터에 대한 유사도 점수를 산출하도록 더 구성된, NGS 샘플 검증용 디바이스.

발명의 설명

기 술 분 야

[0001] 본 발명은 NGS 샘플 검증 방법 및 이를 이용한 디바이스에 관한 것으로, 보다 구체적으로 대상 샘플에 대하여 NGS 분석된 데이터를 기초로 NGS 샘플을 검증하는 방법 및 이를 이용한 디바이스에 관한 것이다.

배 경 기 술

[0002] NGS (Next-generation sequencing,) 는 임상적 연구를 포함하는 다양한 목적을 달성하기 위해 유전체 및 전체 분석에 적용되고 있다. 이때, 대부분의 NGS에 기초한 연구들은 제한된 시간 내에 다수의 샘플에 대한 NGS 동시 분석을 진행하고 있다.

[0003] 한편, NGS 분석에 있어서, 알고리즘의 개선으로 인해 유전체 변이에 대한 검출의 정확도는 증가하였지만, 사용자의 핸들링에 따른 오류 즉, 휴먼 오류 (human errors) 는 여전히 빈번하게 발생하고 있는 실정이다.

[0004] 이때, 임상 연구 환경에서, NGS 분석 데이터의 출처를 보장하는 것은 필수일 수 있다. 따라서, NGS 분석 결과와 대상 샘플간의 불일치는 임상 연구에 있어서 큰 문제를 야기할 수 있다.

[0005] 보다 구체적으로, 연구 환경에서 샘플간 혼합에 따른, NGS 분석 데이터 및 대상 샘플의 불일치는, 인과 관계에 있는 변이 검출의 신뢰도 떨어뜨릴 수 있다. 나아가, 임상 환경에서 NGS 분석 데이터 및 대상 샘플의 불일치는, 임상 결과를 보고하는 시간을 지연시키거나, 부정확한 결과를 환자에게 보고할 수도 있다.

[0006] 예를 들어, 하나의 개체 (또는, 샘플) 로부터 획득된 암 조직 및 정상 조직에 대한 NGS 분석 결과는 암 연구에 있어서 중요한 임상적 의의를 가질 수 있다. 이때, 타 개체의 NGS 분석 결과와 미스 매칭될 경우 거짓 체성 돌연변이에 대한 콜 (call) 이 증가하는 등의 부정확한 임상적 결과가 초래될 수 있다.

[0007] 이에, 정확한 NGS 분석 결과를 제공하는 것에 있어서, NGS 분석 과정에서 발생하는 오류의 개선뿐만 아니라, NGS 분석 결과의 보증을 위한 추가적인 검증 절차가 필수적일 수 있다.

[0008] 따라서, NGS 분석의 사후 검증을 위한 방법에 대한 개발이 요구되고 있는 실정이다.

[0009] 발명의 배경이 되는 기술은 본 발명에 대한 이해를 보다 용이하게 하기 위해 작성되었다. 발명의 배경이 되는 기술에 기재된 사항들이 선행기술로 존재한다고 인정하는 것으로 이해되어서는 안 된다.

발명의 내용

해결하려는 과제

[0010] 한편, 대상 샘플 및 NGS 분석 결과의 매칭 오류를 해결하기 위한 방법으로, 염색체 내 특정 위치의 단일 염기서열에서 인구 집단에 대하여 1 % 이상의 빈도로 변이가 발생하는 SNP (single nucleotide polymorphism) 사이트에 기초하여 NGS 데이터에 대한 샘플의 매칭의 오류 및 교차 오염을 확인하는 방법이 제안되었다.

[0011] 그러나, 앞서 제안된 샘플 검증 방법들은, 대량의 샘플일 경우 개별 러닝이 요구되거나 미스 매칭된 NGS 분석 데이터 및 샘플 쌍을 발견하기 어려울 수 있다. 또한, 상기 검증 방법들은, NGS 분석 방법을 고려하지 않고 다수의 SNP 사이트를 이용함에 따라 긴 러닝 타임이 소요될 수 있고, 프리-프로세싱 (pre-processing) 단계가 요구될 수 있으며, 정확도가 낮은 동일 개체 유래 샘플 여부를 제공할 수 있다.

[0012] 이에, 본 발명의 발명자들은, NGS 분석 데이터의 빠르고 정확한 검증이 가능한, 새로운 NGS 샘플 검증 시스템을 개발하고자 하였다.

[0013] 특히, 본 발명의 발명자들은, 다량의 샘플에 대한 NGS 분석 결과에 대하여, 사용자의 개입 없이 샘플 및 NGS 데이터를 정확하고 빠르게 매칭해주는 NGS 샘플 검증 시스템을 개발하고자 하였다.

[0014] 나아가, 본 발명의 발명자들은, 분석 패널의 종류에 따라 SNP 사이트를 결정하여 러닝 타임을 줄이고, 소형 패널에 대하여 높은 검증의 정확도를 유지하도록 구성된 NGS 샘플 검증 시스템을 제공하고자 하였다.

[0015] 본 발명의 발명자들은 특히, 맵핑 능력 (Mappability) 을 고려하여 SNP 사이트를 결정하고자 하였다. 보다 구체적으로, 본 발명의 발명자들은 미리 결정된 수준 이상의 맵핑 능력을 갖는 SNP 사이트를 선별하고, 이들 사이트에 대하여 GMAF (global minor allele frequency) 와 함께 개별 인구 (population) 의 대립 유전자 빈도 (allele-frequency) 를 고려하여, 세부 군집 내에서 자주 변이가 발생된다고 보고된 SNP 사이트를 선별하고자 하였다.

[0016] 나아가, 본 발명의 발명자들은, 분석된 NGS 데이터의 파일명의 유사도 정도를 기초로 NGS 분석 결과를 동일 개체 유래의 데이터 별로 매칭하고자 하였다.

[0017] 그 결과, 본 발명의 발명자들은, 다양한 염기서열 분석 패널, 예를 들어 WES (whole exome sequencing), RNA-seq (RNA sequencing) 와 같은 대량의 분석 패널, 또는 표적 염기서열 분석 (targeted sequencing) 과 같은 소형 분석 패널에 적용한 NGS 샘플 검증 시스템을 개발하기에 이르렀다.

[0018] 나아가, 본 발명의 발명자들은, NGS 데이터의 파일명에 기초하여 유사도 점수를 산출하고, 산출된 유사도 점수를 기초로 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하고, 결과를 제공하도록 구성된 NGS 샘플 검증 시스템을 개발할 수 있었다.

[0019] 보다 구체적으로, 본 발명의 발명자들은, SNP 사이트의 염기서열을 고려하여 NGS 분석 결과를 동일 개체 유래의 데이터 별로 매칭하고, 매칭 또는 미스 매칭된 NGS 분석 결과 또는, 어느 결과와도 매칭되지 않은 NGS 분석 결과를 분류하도록 구성된, 유전자형에 기초한 NGS 샘플 검증 시스템을 개발할 수 있었다.

[0020] 이때, 본 발명의 발명자들은, 새로이 개발된 NGS 샘플 검증 시스템에 대하여, 다량의 샘플에 대한 NGS 분석 데

이터를 적용하거나 다양한 분석 패널에 따른 NGS 분석 데이터를 적용했을 때 종래의 샘플 검증 방법들보다 빠르고 정확한 검증이 가능한 것을 확인할 수 있었다.

[0021] 결과적으로 본 발명의 발명자들은, 새로운 NGS 샘플 검증 시스템의 개발을 통해 NGS 분석 결과를 샘플별로 매칭하고 그 결과를 제공할 수 있었다.

[0022] 본 발명의 발명자들은, NGS 샘플 검증 시스템을 통해 미스 매칭되거나 매칭되지 않은 NGS 분석 결과를 사용자에게 제공함에 따라, NGS 분석 결과 제공에 있어서 빈번하게 발생할 수 있는 문제점들을 해결할 수 있음을 기대할 수 있었다.

[0023] 특히, 본 발명의 발명자들은, 샘플간 혼합에 따라 발생하는 검출의 신뢰도 떨어뜨리거나 부정확한 결과를 제공하는 문제점들을 해결할 수 있음을 기대할 수 있었다.

[0024] 이에, 본 발명이 해결하고자 하는 과제는, SNP 사이트에 대한 서열 정보를 기초하여, NGS 데이터간 동일 개체 유래 샘플 여부를 판단하여 제공하는, NGS 샘플 검증 방법 및 이를 이용한 디바이스를 제공하는 것이다.

[0025] 본 발명이 해결하고자 하는 다른 과제는, 파일명에 기초하여, 동일 개체 유래 샘플 별로 NGS 데이터를 매칭하여 제공하는, NGS 샘플 검증 방법 및 이를 이용한 디바이스를 제공하는 것이다.

[0026] 본 발명이 해결하고자 하는 또 다른 과제는, SNP 사이트 및 파일명에 기초하여, 동일 개체 유래 샘플 별로 NGS 데이터를 매칭하여 제공하는, NGS 샘플 검증 방법 및 이를 이용한 디바이스를 제공하는 것이다.

[0027] 본 발명의 과제들은 이상에서 언급한 과제들로 제한되지 않으며, 언급되지 않은 또 다른 과제들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0028] 전술한 바와 같은 과제를 해결하기 위해, 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법이 제공된다. 이때, 본 발명의 NGS 샘플 검증 방법은, 프로세서를 포함하는 NGS 디바이스에 구현된다. 보다 구체적으로, 본 발명의 NGS 샘플 검증 방법은, 미리 결정된 표적 SNP 사이트에 대한 서열 정보를 포함하는 대상 샘플에 대한 복수의 NGS 데이터를 수신하는 단계, 표적 SNP 사이트에 대한 서열 정보를 기초로 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계, 및 결과를 제공하는 단계를 포함한다.

[0029] 본 발명의 특징에 따르면, 복수의 NGS 데이터는 WES (whole exome sequencing) 파일, RNA 시퀀싱 (RNA sequencing) 파일, 및 표적 시퀀싱 (targeted sequencing) 파일 중 적어도 하나의 NGS 분석 방법에 따른 파일을 포함할 수 있다. 나아가, 본 발명의 NGS 샘플 검증 방법은, 복수의 NGS 데이터를 수신하는 단계 이후에, GMAF (global minor allele frequency), 개별 인구 내의 MAF 및 맵핑 능력 중 적어도 하나를 기초로, 표적 SNP 사이트를 결정하는 단계를 더 포함할 수 있다.

[0030] 본 발명의 다른 특징에 따르면, 복수의 NGS 데이터가, WGS 분석 파일, WES 분석 파일 또는 RNA 시퀀싱 분석 파일일 경우, 표적 SNP 사이트를 결정하는 단계는, 맵핑 능력이 미리 결정된 수준 이상인 SNP 사이트 중 GMAF가 0.45 이상 0.55 이하이고 개별 인구 내 MAF가 0.35 이상 0.65 이하인 SNP 사이트를 표적 SNP 사이트로 결정하는 단계를 포함할 수 있다.

[0031] 본 발명의 또 다른 특징에 따르면, 복수의 NGS 데이터가 표적 시퀀싱 분석 파일일 경우, 표적 SNP 사이트를 결정하는 단계는, 맵핑 능력이 미리 결정된 수준 이상인 SNP 사이트 중 GMAF가 0.1 이상 0.9 이하인 SNP 사이트를 표적 SNP 사이트로 결정하는 단계를 포함할 수 있다.

[0032] 본 발명의 또 다른 특징에 따르면, 동일 개체 유래 샘플 여부를 결정하는 단계는, 복수의 NGS 데이터 각각의 표적 SNP 사이트에 대한 서열 정보를 기초로, 유전자형 일치 점수 (genotype concordance score) 를 산출하는 단계, 및 유전자형 일치 점수를 기초로 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계를 포함할 수 있다. 또한, 제공하는 단계는, 동일 개체 유래 샘플 여부를 제공하는 단계를 포함할 수 있다.

[0033] 본 발명의 또 다른 특징에 따르면, 유전자형 일치 점수를 기초로 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계는, 복수의 NGS 데이터 중 상기 유전자형 일치 점수가 0.7 이상인 NGS 데이터를 동일 개체 유래 샘플로 결정하는 단계를 더 포함할 수 있다.

[0034] 본 발명의 또 다른 특징에 따르면, 복수의 NGS 데이터 중 유전자형 일치 점수가 0.7 미만인 NGS 데이터를 샘플과 매칭되지 않는 NGS 데이터로 결정하는 단계를 더 포함할 수 있다.

[0035] 본 발명의 또 다른 특징에 따르면, 제공하는 단계는, 샘플과 매칭되지 않은 NGS 데이터를 더 제공하는 단계를 더 포함할 수 있다.

[0036] 본 발명의 또 다른 특징에 따르면, 복수의 NGS 데이터는 미리 결정된 파일명을 포함하고, 파일명을 기초로, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계를 더 포함할 수 있다.

[0037] 본 발명의 또 다른 특징에 따르면, 파일명을 기초로, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계는, 파일명을 기초로, 복수의의 NGS 데이터에 대한 유사도 점수를 산출하는 단계, 및 유사도 점수를 기초로 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계를 포함할 수 있다.

[0038] 전술한 바와 같은 과제를 해결하기 위해, 본 발명의 다른 실시예에 따른 NGS 샘플 검증 방법이 제공된다. 본 발명의 다른 실시예에 따른 NGS 샘플 검증 방법은, 프로세서를 포함하는 NGS 디바이스에 구현되고, 미리 결정된 파일명을 포함하는, 대상 샘플에 대한 복수의 NGS 데이터를 수신하는 단계, 파일명을 기초로 복수의 NGS 데이터에 대한 유사도 점수를 산출하는 단계, 및 유사도 점수를 기초로 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계, 및 동일 개체 유래 샘플 여부를 제공하는 단계를 포함한다.

[0039] 본 발명의 특징에 따르면, 파일명은 서로 상이한 구분 문자를 포함할 수 있다. 또한, 유사도 점수를 산출하는 단계는, 구분 문자를 기초로, 복수의 NGS 데이터의 파일명을 복수의 부분으로 분할하는 단계, 복수의 부분에 대한 값의 출연 빈도를 산출하는 단계, 및 출연 빈도를 기초로 복수의 NGS 데이터 중 선택된 두 개의 NGS 데이터의 유사도 점수를 산출하는 단계를 포함할 수 있다.

[0040] 본 발명의 다른 특징에 따르면, 파일명은 서로 상이한 구분 문자를 포함할 수 있다. 또한, 유사도 점수를 산출하는 단계는, 분문자를 기초로, 복수의 NGS 데이터의 파일명을 복수의 부분으로 분할하는 단계, 복수의 부분에 대한 값의 분산 정도를 산출하는 단계, 및 분산 정도를 기초로 복수의 NGS 데이터 중 선택된 두 개의 NGS 데이터의 유사도 점수를 산출하는 단계를 포함할 수 있다.

[0041] 본 발명의 또 다른 특징에 따르면, 분산 정도를 산출하는 단계는, 복수의 부분에 대한 값의 출연 빈도를 산출하는 단계, 출연 빈도를 기초로 분산 정도를 산출하는 단계를 포함할 수 있다.

[0042] 본 발명의 또 다른 특징에 따르면, 분산 정도를 기초로 유사도 점수를 산출하는 단계는, 복수의 NGS 데이터 중 선택된 두 개의 NGS 데이터에 대하여, 복수의 부분 중 선택된 하나의 부분에 대한 값의 동일 여부 및 분산 정도를 기초로, 유사도 점수를 산출하는 단계를 포함할 수 있다.

[0043] 본 발명의 또 다른 특징에 따르면, 동일 개체 유래 샘플 여부를 결정하는 단계는, 복수의 NGS 데이터 중 선택된 두 개의 NGS 데이터에 대하여 산출된 유사도 점수가 가장 높을 경우, 두 개의 NGS 데이터를 동일 개체 유래 샘플 여부를 결정하는 단계를 더 포함할 수 있다.

[0044] 본 발명의 또 다른 특징에 따르면, 복수의 NGS 데이터는, 미리 결정된 표적 SNP 사이트에 대한 서열 정보를 포함하고, 상기 방법은 SNP 사이트에 대한 서열 정보를 기초로 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계를 더 포함할 수 있다.

[0045] 본 발명의 또 다른 특징에 따르면, SNP 사이트에 대한 서열 정보를 기초로, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계는, 복수의 NGS 데이터 각각의 표적 SNP 사이트에 대한 서열 정보를 기초로, 유전자형 일치 점수를 산출하는 단계, 및 유전자형 일치 점수를 기초로 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계를 포함할 수 있다.

[0046] 전술한 바와 같은 과제를 해결하기 위해, 본 발명의 또 다른 실시예에 따른 NGS 샘플 검증 방법이 제공된다. 본 발명의 또 다른 실시예에 따른 NGS 샘플 검증 방법은, 프로세서를 포함하는 NGS 디바이스에 구현되고, 미리 결정된 표적 SNP 사이트에 대한 서열 정보 및 미리 결정된 파일명을 포함하는, 대상 샘플에 대한 복수의 NGS 데이터를 수신하는 단계, 표적 SNP 사이트에 대한 서열 정보를 기초로, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 제1 결정하는 단계, 파일명을 기초로 복수의 NGS 데이터에 대한 유사도 점수를 산출하는 단계, 및 유사도 점수를 기초로 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 제2 결정하는 단계를 포함할 수 있다.

[0047] 전술한 바와 같은 과제를 해결하기 위해, 본 발명의 일 실시예에 따른 NGS 샘플 검증용 디바이스가 제공된다. 상기 NGS 샘플 검증용 디바이스는, 미리 결정된 표적 SNP 사이트에 대한 서열 정보를 포함하는, 대상 샘플에 대한 복수의 NGS 데이터를 수신하도록 구성된 수신부, 및 수신부와 통신하도록 구성된 프로세서를 포함한다. 이때, 프로세서는, 표적 SNP 사이트에 대한 서열 정보를 기초로, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부

를 결정하고, 동일 개체 유래 샘플 여부를 제공하도록 구성된다.

[0048] 본 발명의 특징에 따르면, 복수의 NGS 데이터는 WGS 파일, WES 파일, RNA 시퀀싱 파일, 및 표적 시퀀싱 파일 중 적어도 하나의 NGS 분석 방법에 따른 파일을 포함하고, 프로세서는 GMAF (global minor allele frequency) 의 수준, 개별 인구 내의 MAF, 및 SNP 사이트의 맵핑 능력 (mappability) 중 적어도 하나, 및 NGS 분석 데이터의 종류를 기초로, 표적 SNP 사이트를 결정하도록 더 구성될 수 있다.

[0049] 전술한 바와 같은 과제를 해결하기 위해, 본 발명의 다른 실시예에 따른 NGS 샘플 검증용 디바이스가 제공된다. 상기 NGS 샘플 검증용 디바이스는, 미리 결정된 파일명을 포함하는, 대상 샘플에 대한 복수의 NGS 데이터를 수신하도록 구성된 수신부, 및 수신부와 통신하도록 구성된 프로세서를 포함한다. 이때, 프로세서는, 파일명을 기초로 복수의 NGS 데이터에 대한 유사도 점수를 산출하고, 유사도 점수를 기초로 복수의 NGS 데이터를 매칭하고, 동일 개체 유래 샘플 여부를 제공하도록 구성된다.

[0050] 본 발명의 특징에 따르면, 파일명은 서로 상이한 구분 문자를 포함하고, 프로세서는 구분 문자를 기초로, 복수의 NGS 데이터의 파일명을 복수의 부분으로 분할하고, 복수의 부분에 대한 값의 출연 빈도를 기초로 복수의 NGS 데이터의 유사도 점수를 산출하도록 더 구성될 수 있다.

[0051] 본 발명의 다른 특징에 따르면, 파일명은 서로 상이한 구분 문자를 포함하고, 프로세서는 구분 문자를 기초로, 복수의 NGS 데이터의 파일명을 복수의 부분으로 분할하고, 복수의 부분에 대한 값의 분산 정도를 산출하고, 분산 정도를 기초로 복수의 NGS 데이터의 유사도 점수를 산출하도록 더 구성될 수 있다.

[0052] 이하, 실시예를 통하여 본 발명을 보다 상세히 설명한다. 다만, 이들 실시예는 본 발명을 예시적으로 설명하기 위한 것에 불과하므로 본 발명의 범위가 이들 실시예에 의해 한정되는 것으로 해석되어서는 아니 된다.

발명의 효과

[0053] 본 발명은, 다량의 샘플에 대한 NGS 분석 결과에 대하여, 사용자의 개입 없이 샘플 및 NGS 데이터를 정확하고 빠르게 매칭해주는 NGS 샘플 검증 시스템을 제공할 수 있는 효과가 있다.

[0054] 보다 구체적으로, 본 발명은 사용자의 핸들링에 따라 발생하는 휴먼 오류, 예를 들어 샘플 및 분석 데이터의 미스매칭과 같은 오류를 개선하여, 샘플 및 NGS 데이터를 매칭하여 제공해주는 NGS 샘플 검증 시스템을 제공할 수 있다.

[0055] 특히, 본 발명은, NGS 데이터의 파일명에 기초하여 유사도 점수를 산출하고, 산출된 유사도 점수를 기초로 동일 개체 유래 샘플 여부에 따라 NGS 데이터를 매칭하고, 동일 개체 유래 샘플 여부를 제공하도록 구성된 NGS 샘플 검증 시스템을 제공할 수 있다.

[0056] 이에, 본 발명은, 샘플의 유전자형 및/또는 NGS 데이터의 파일명을 기초로 NGS 샘플 검증 시스템을 통해 미스매칭되거나 매칭되지 않은 NGS 분석 결과를 사용자에게 제공함에 따라, NGS 분석 결과 제공에 있어서 빈번하게 발생할 수 있는 문제점들을 해결할 수 있다.

[0057] 이에, 본 발명은, 샘플간 혼합에 따라 발생하는 검출의 신뢰도 떨어뜨리거나 부정확한 결과를 제공하는 문제점들을 해결할 수 있다.

[0058] 또한, 본 발명은, 맵핑 능력, GMAF, 개별 인구 내 MAF, 분석 패널의 종류에 따라 SNP 사이트를 결정하여 러닝 타임을 줄이고, 소형 패널에 대하여 높은 검증의 정확도를 유지하도록 구성된 NGS 샘플 검증 시스템을 제공할 수 있는 효과가 있다.

[0059] 특히, 본 발명은 다양한 염기서열 분석 패널, 예를 들어 WGS, WES, RNA-seq 와 같은 대량의 분석 패널, 또는 표적 염기서열 분석과 같은 소형 분석 패널에 적용한 NGS 샘플 검증 시스템을 제공할 수 있다.

[0060] 본 발명은, 에 따른 효과는 이상에서 예시된 내용에 의해 제한되지 않으며, 더욱 다양한 효과들이 본 명세서 내에 포함되어 있다.

도면의 간단한 설명

[0061] 도 1a는 본 발명의 일 실시예에 따른 NGS 샘플 검증용 디바이스 구성을 예시적으로 도시한 것이다.

도 1b는 본 발명의 일 실시예에 따른 NGS 샘플 검증용 디바이스의 출력부를 예시적으로 도시한 것이다.

도 2a 및 2b는 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법의 절차를 예시적으로 도시한 것이다.

도 3a 내지 3g는 본 발명의 다른 실시예에 따른 NGS 샘플 검증 방법의 절차를 예시적으로 도시한 것이다.

도 4a 및 4b는 본 발명의 또 다른 실시예에 따른 NGS 샘플 검증 방법의 절차를 예시적으로 도시한 것이다.

도 5는 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법에 의해, 유전자형에 따라 매칭된 NGS 데이터 쌍 또는 미스매칭된 NGS 데이터 쌍의 유전자형 일치 점수를 도시한 것이다.

도 6a 내지 6f는 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법에 의해, 파일명에 따라 매칭된 NGS 데이터 쌍 또는 미스매칭된 NGS 데이터 쌍의 평가 결과를 도시한 것이다.

도 7은 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법에 의해, 유전자형 및 파일명에 따라 매칭된 NGS 데이터 쌍 또는 미스매칭된 NGS 데이터 쌍의 평가 결과를 도시한 것이다.

도 8a 내지 8d는 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법 및 종래의 샘플 검증 방법의 평가 결과를 비교하여 도시한 것이다.

발명을 실시하기 위한 구체적인 내용

[0062] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 것이며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하며, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0063] 본 발명의 실시예를 설명하기 위한 도면에 개시된 형상, 크기, 비율, 각도, 개수 등은 예시적인 것이므로 본 발명이 도시된 사항에 한정되는 것은 아니다. 또한, 본 발명을 설명함에 있어서, 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명은 생략한다. 본 명세서 상에서 언급된 '포함한다', '갖는다', '이루어진다' 등이 사용되는 경우, '~만'이 사용되지 않는 이상 다른 부분이 추가될 수 있다. 구성요소를 단수로 표현한 경우에 특별히 명시적인 기재 사항이 없는 한 복수를 포함하는 경우를 포함한다.

[0064] 구성요소를 해석함에 있어서, 별도의 명시적 기재가 없더라도 오차 범위를 포함하는 것으로 해석한다.

[0065] 본 발명의 여러 실시예들의 각각 특징들이 부분적으로 또는 전체적으로 서로 결합 또는 조합 가능하며, 당업자가 충분히 이해할 수 있듯이 기술적으로 다양한 연동 및 구동이 가능하며, 각 실시예들이 서로에 대하여 독립적으로 실시 가능할 수도 있고 연관 관계로 함께 실시 가능할 수도 있다.

[0066] 본 명세서의 해석의 명확함을 위해, 이하에서는 본 명세서에서 사용되는 용어들을 정의하기로 한다.

[0067] 본 명세서에서 이용되는 용어, "NGS (Next Generation Sequencing)"는 차세대 염기서열 분석으로서, 유전체의 염기서열의 고속 분석 방법 중 하나이다.

[0068] 이때, NGS는 임상적 연구를 포함하는 다양한 목적을 달성하기 위해 유전체 및 전체 분석에 적용될 수 있다. 한편, 대부분의 NGS 분석은 다수의 대상 샘플에 대한 동시 분석이 수행될 수 있다. 이때, NGS 분석 결과와 대상 샘플간의 불일치는 임상 연구에 있어서 큰 문제를 야기할 수 있음에 따라, NGS 데이터의 출처에 대한 검증은 보다 중요할 수 있다.

[0069] 본 명세서에서 이용되는 용어, "대상 샘플"은 NGS 분석의 표적이 되는 샘플을 의미한다. 이때, 대상 샘플은 개체로부터 분리된 생물학적 시료일 수 있다.

[0070] 예를 들어, 대상 샘플은, 하나의 개체로부터 분리된 암 조직 또는 정상 조직일 수 있다. 이때, 하나의 개체로부터 분리된 암 조직과 정상 조직에 대한 NGS 데이터는, 한 쌍의 파일로서 제공될 수 있다.

[0071] 이때, 대상 샘플은 조직에 제한되는 것이 아니며 NGS 분석이 가능한 다양한 시료가 될 수 있다. 예를 들어, 대상 샘플은, 조직, 전혈, 혈장, 혈청, 침, 안구액, 뇌척수액, 땀, 뇨, 젖, 복수액, 활액, 복막액, 세포 용해물, 모, 손발톱의 피부계 하위기관일 수 있다.

[0072] 한편, 본원 명세서 내에 개시된 대상 샘플은, 개체와 동일한 의미로도 해석될 수 있다.

[0073] 본 명세서에서 이용되는 용어, "NGS 데이터"는, 대상 샘플에 대하여 분석된 염기서열 데이터를 포함하는 파일을 의미할 수 있다.

[0074] 이때, NGS 데이터는 유전체 전체에 대하여 염기서열 분석된 WGS (whole genome sequencing) 파일, WES (whole exome sequencing) 파일, RNA 시퀀싱 (RNA sequencing) 파일, 및 특정 영역에 대하여 염기서열 분석된 표적 시퀀싱 (targeted sequencing) 파일을 포함할 수 있다.

[0075] 본 발명의 일 실시예에 따르면, NGS 데이터는 BAM 파일로 제공될 수 있다. 그러나, 이에 제한되지 않고, NGS 데이터는 특정 대상 영역에 대한 위치 정보를 포함하는 BED 파일로 제공될 수도 있다.

[0076] 한편, NGS 데이터는, 동일한 대상 샘플 또는 동일한 개체에 따라 한 쌍의 파일로 존재할 수 있다.

[0077] 이때, '한 쌍의 파일'은, 동일한 개체 또는 샘플 유래한 두 개 이상의 파일을 의미할 수 있다.

[0078] 본 명세서에서 이용되는 용어, "복수의 NGS 데이터"은 복수의 대상 샘플 또는 복수의 개체에 대하여 염기서열 분석된 데이터를 포함하는 파일을 의미할 수 있다.

[0079] 본 발명의 일 실시예에 따르면, NGS 데이터는, SNP 사이트에 대하여 분석된 염기서열을 포함할 수 있다.

[0080] 본 명세서에서 이용되는 용어, "SNP (single-nucleotide polymorphism)"는 염기서열 중 한 개의 염기 서열에 변이가 일어나 유전적 다양성이 관찰되는 현상을 의미할 수 있고, SNP 사이트는 염색체가 갖고 있는 염기서열 중 개인의 편차가 나타나는 하나의 개별 염기 서열 위치를 의미한다. 즉, SNP에 의해 개체들은, 유전적 다양성을 가질 수 있다.

[0081] 본 발명의 일 실시예에 따르면, SNP 사이트는 유전자형에 기초하여 복수의 NGS 데이터 중 동일한 샘플 또는 동일한 개체에 대한 한 쌍의 NGS 데이터를 탐색하기 위해 이용될 수 있다.

[0082] 본 명세서에서 이용되는 용어, "표적 SNP 사이트"는 유전체 내에 존재하는 복수의 SNP 중 선택된 SNP 사이트를 의미할 수 있다.

[0083] 이때, 표적 SNP 사이트는 수신된 분석 패널의 다양성에 따른 NGS 데이터의 종류 (예를 들어, WGS 파일, WES 파일, RNA 시퀀싱 파일, 또는 표적 시퀀싱 파일) 에 따라 가변적으로 선택될 수 있다. NGS 샘플을 검증에 있어서 이러한 표적 SNP 사이트의 선택은, 모든 SNP 사이트에 대한 분석을 진행하여 NGS 샘플을 검증하는 방법보다 빠른 분석 결과를 제공할 수 있다.

[0084] 본 발명의 일 실시예에 따르면 표적 SNP 사이트는, NGS 데이터의 종류에 따라 미리 결정된 맵핑 능력, GMAF (global minor allele frequency) 의 수준, 개별 인구의 MAF 중 적어도 하나에 기초하여 선택될 수 있다.

[0085] 본 명세서에서 이용되는 용어, "맵핑 능력"은, 유전자 맵핑 능력을 의미하며, 맵핑 퀄리티 (mapping quality) 수준에 의해 결정될 수 있다.

[0086] 본 발명의 일 실시예에 따르면, 표적 SNP 사이트는 맵핑 퀄리티 실효값 (root mean square, RMS) 이 50 이상인 SNP 사이트일 수 있다.

[0087] 본 명세서에서 이용되는 용어, "GMAF"는 글로벌 대립 유전자 빈도로, 특정 유전자좌 (gene locus) 에 대립 유전자가 발생하는 빈도를 의미할 수 있다.

[0088] 본 발명의 일 실시예에 따르면, NGS 데이터가 WES 분석 파일 또는 RNA 시퀀싱 분석 파일일 경우, 표적 SNP 사이트는 GMAF가 0.45 내지 0.55인 SNP 사이트일 수 있다.

[0089] 바람직하게, NGS 데이터가 WES 분석 파일 또는 RNA 시퀀싱 분석 파일일 경우, 표적 SNP 사이트는 GMAF가 0.45 내지 0.55이고, 그 개별 인구의 MAF가 0.1 내지 0.9인 SNP 사이트일 수 있다.

[0090] 예를 들어, NGS 데이터가 WGS 분석 파일 또는 WES 분석 파일 또는 RNA 시퀀싱 분석 파일일 경우, 표적 SNP 사이트는 상기와 같은 GMAF 및/또는 MAF의 조건에 따라 853 개의 SNP 사이트가 선택될 수 있으나, 이에 제한되는 것은 아니다.

[0091] 본 발명의 다른 실시예에 따르면, NGS 데이터가 표적 시퀀싱 분석 파일일 경우, 표적 SNP 사이트는 GMAF가 0.1 내지 0.9인 SNP 사이트일 수 있다.

[0092] 예를 들어, NGS 데이터가 WES 분석 파일 또는 RNA 시퀀싱 분석 파일일 경우, 표적 SNP 사이트는 상기와 같은

GMAF의 조건에 따라 최소 200 개의 SNP 사이트가 선택될 수 있으나, 이에 제한되는 것은 아니다.

[0093] 본 발명의 또 다른 실시예에 따르면, 표적 SNP 사이트는, 맵핑 능력 (mappability) 을 더욱 고려하여 결정될 수 있다.

[0094] 보다 구체적으로, 표적 SNP 사이트는, 코딩 영역에서의 변이, 콜링 필터 (Calling filter) 를 통과한 변이, gnomAD 데이터 베이스를 위해 랜덤 포레스트 모델을 이용하여 생성된 필터에 의해 통과된 변이, dbSNP로 보고된 변이, QD (QualByDepth) 가 2.0 이상이고, RMSMQ (root mean square mapping quality) 가 50 이상인 변이 중 적어도 하나를 만족할 수 있다.

[0095] 또한, 표적 SNP 사이트는, 낮은 복합 영역 (low complex region) 가 아닌 변이, 세그먼트 복제 영역 (segment duplicated region) 이 아닌 변이, 단순 반복 영역 (simple repeat region) 이 아닌 변이 중 적어도 하나를 만족할 수 있다.

[0096] 본 명세서에서 이용되는 용어, "유전자형 일치 점수 (genotype concordance score)"는 복수의 NGS 데이터 중 선택된 두 개의 NSG 파일에 대한 SNP 사이트의 유사율을 의미할 수 있다.

[0097] 예를 들어, NGS 데이터의 표적 SNP 사이트에 대한 유전자형 일치 점수가 0.7 이상일 경우, 바람직하게 0.8 이상인 NGS 데이터는 동일한 개체 유래의 샘플인 한 쌍의 파일로 매칭될 수 있다.

[0098] 본 발명의 일 실시예에 따르면, NGS 데이터는, 미리 결정된 파일명을 포함할 수 있다.

[0099] 본 명세서에서 이용되는 용어, "파일명"은 특정한 파일을 구성하는 파일을 구별하기 위해서 사용하는 문자와 숫자로 구성된 기호화된 이름을 의미한다.

[0100] 즉, 파일명은, 복수의 대상 샘플, 또는 복수의 개체별로 상이하게 설정될 수 있다.

[0101] 본 발명의 일 실시예에 따르면, 파일명은 복수의 NGS 데이터 중 동일한 샘플 또는 동일한 개체에 대한 한 쌍의 NGS 데이터를 탐색하기 위해 이용될 수 있다.

[0102] 본 발명의 일 실시예에 따르면, 파일명은 구분 문자를 포함할 수 있다.

[0103] 본 명세서에서 이용되는 용어, "구분 문자"는 임의의 기호로 이루어지는 열을 구성 요소로 구분 짓기 위한 문자를 의미한다. 이때 구성 요소는 영문자나 정수와 같은 값을 가질 수 있고, 구분 문자는 동일 또는 다른 뜻을 갖는 구성 요소의 배열로부터 각각의 구성 요소를 분리하는 데 쓰일 수 있다.

[0104] 예를 들어, 구분 문자는, bar (-, _), dot (.), slash (/), colon (:), semicolon (;), apostrophe (') 중 적어도 하나일 수 있으나, 이에 제한되지 않고 각 구성 요소를 구분하는 한 보다 다양한 문자일 수 있다.

[0105] 본 명세서에서 이용되는 용어, "복수의 부분"은 구분 문자에 의해 구성 요소 별로 구분된 영역을 의미할 수 있다. 이에, 본 명세서에서 이용되는 용어, "복수의 부분에 대한 값"은, 복수의 부분 중 선택된 특정 부분에 해당하는 구성 요소를 의미할 수 있다.

[0106] 본 명세서에서 이용되는 용어, "유사도 점수"는 복수의 NGS 데이터 중 선택된 임의의 두 개의 NSG 파일에 대한 파일명의 유사한 정도를 의미할 수 있다.

[0107] 본 발명의 다른 실시예에 따르면, 유사도 점수는, 미리 결정된 수준 이상의 유전자형 점수를 갖는 복수의 NGS 데이터 중 선택된 임의의 두 개의 NGS 분석 데이터에 대하여 산출될 수 있다.

[0108] 즉, 유사도 점수는, 복수의 NGS 데이터를 이루는 모든 쌍의 NGS 분석 데이터에 대하여 산출될 수 있다.

[0109] 본 발명의 다른 실시예에 따르면, 유사도 점수는, 복수의 부분에 대한 값의 출연 빈도 및/또는 복수의 부분 중 선택된 하나의 부분에 대한 상기 값의 동일 여부 및 상기 분산 정도에 기초하여 산출될 수 있다.

[0110] 본 명세서에서 이용되는 용어, "출연 빈도"는 복수의 부분 중 선택된 특정 부분에 대하여 나타나는 구성 요소의 빈도를 의미할 수 있다.

[0111] 본 명세서에서 이용되는 용어, "분산 정도"는 복수의 부분 중 선택된 특정 부분에 대하여 나타나는 구성 요소의 분산 정도를 의미한다.

[0112] 본 발명의 일 실시예에 따르면, 복수의 NGS 데이터 중 동일한 샘플 또는 동일한 개체에 대한 한 쌍의 NGS 데이터는, 다른 샘플 또는 다른 개체로부터 유래된 NGS 데이터와 매칭되었을 때 보다 높은 유사도 점수를 가질 수

있다.

[0113] 따라서, 두 개의 NGS 데이터에 대하여 산출된 상기 유사도 점수가 가장 높을 경우, 두 개의 NGS 데이터가 한 쌍의 파일로 매칭될 수 있다.

[0114] 그러나, 한 쌍의 파일은 두 개의 NGS 데이터만을 의미하는 것이 아니다.

[0115] 예를 들어, 복수의 개체에 대하여 분리된 세 개의 샘플에 대한 NGS 분석이 수행될 경우, 세 개의 샘플 각각에 대한 세 개의 NGS 데이터가 동일한 개체로부터 유래된 한 쌍의 파일로서 매칭될 수 있다.

[0116] 본 발명의 일 실시예에 따르면, NGS 데이터에 대한 파일명의 유사도 점수는, 구분 문자에 의해 분할된 파일명의 벡터화를 통해 산출될 수 있다. 그러나, 유사도 점수의 산출 방법은 이에 제한되는 것이 아니다.

[0117] 본 명세서에서 이용되는 용어, "NGS 디바이스"는 NGS 샘플 검증 디바이스 또는 NGS 분석 디바이스 자체로 해석될 수 있다.

[0118] 즉, 본 발명의 NGS 샘플 검증 디바이스는, 독립된 디바이스로 존재할 수 있고, 다양한 NGS 분석 디바이스의 구성으로서 수반될 수 있다.

[0119] 이하에서는 도 1a 및 도 1b를 참조하여, 본 발명의 다양한 실시예에 따른 NGS 샘플 검증용 디바이스에 관하여 구체적으로 설명한다.

[0120] 도 1a는 본 발명의 일 실시예에 따른 NGS 샘플 검증용 디바이스 구성을 예시적으로 도시한 것이다. 도 1b는 본 발명의 일 실시예에 따른 NGS 샘플 검증용 디바이스의 출력부를 예시적으로 도시한 것이다.

[0121] 도 1a을 참조하면, 본 발명의 다양한 실시예에 따른 NGS 샘플 검증용 디바이스 (100) 는 수신부 (110), 입력부 (120), 출력부 (130), 저장부 (140) 및 프로세서 (150) 를 포함한다.

[0122] 구체적으로 수신부 (110) 는 미리 결정된 표적 SNP (single-nucleotide polymorphism) 사이트에 대한 서열 정보 및/또는 미리 결정된 파일명을 포함하는, 대상 샘플에 대한 복수의 NGS 데이터를 수신할 수 있다. 본 발명의 특징에 따르면, NGS 데이터는, WGS 파일, WES 파일, RNA 시퀀싱 파일, 및 표적 시퀀싱 파일 중 적어도 하나일 수 있다. 본 발명의 다른 특징에 따르면, NGS 데이터는 BAM 파일로서 제공될 수 있다. 그러나, 수신부 (110) 는 이에 제한되지 않고, 선택 영역에 대한 BED 파일을 더 수신할 수 있다.

[0123] 입력부 (120) 는 키보드, 마우스, 터치 스크린 패널 등 제한되지 않는다. 입력부 (120) 는 NGS 샘플 검증용 디바이스 (100) 를 설정하고, NGS 샘플 검증용 디바이스 (100) 의 동작을 지시할 수 있다. 한편, 사용자는 입력부 (120) 를 통해, NGS 샘플 검증을 위한 추가 정보들을 더욱 입력할 수 있다.

[0124] 다음으로, 출력부 (130) 는 후술할 프로세서 (150) 에 따른 검증 결과들을 표시하도록 구성될 수 있다. 나아가, 출력부 (130) 는 입력부 (120) 에 입력된 추가 정보들을 표시하도록 구성될 수 있다.

[0125] 예를 들어, 도 1b의 (a)를 참조하면, 출력부 (130) 는 매칭되지 않은, 낯 개의 NGS 데이터들 및 매칭된 NGS 데이터를 포함하는 검증 결과를 제공하도록 구성될 수 있다. 보다 구체적으로, 도 1b의 (b)를 참조하면, 출력부 (130) 는 SNP 사이트에 대한 일치 정도의 유전자형 및 파일명의 일치 정도에 기초하여 매칭된 한 쌍의 NGS 데이터들의 유전자형 일치 점수와 함께 제공하도록 구성될 수 있다. 도 1b의 (c)를 더욱 참조하면, 출력부 (130) 는 SNP 사이트에 대한 일치 정도의 유전자형 및 파일명의 일치 정도에 기초하여 매칭되지 않은 NGS 데이터들의 유전자형 일치 점수와 함께 제공하도록 구성될 수 있다. 예를 들어, 출력부 (130) 는 SNP 사이트에 대한 유전자형 높은 일치 점수에 따라 한 쌍의 파일로 매칭되었으나, 파일명의 일치 정도에 따라 최종적으로 매칭되지 않은 NGS 데이터를 제공할 수 있다. 또한, 출력부 (130) 는 파일명의 높은 일치 정도에 따라 한 쌍의 파일로 매칭되었으나, SNP 사이트에 대한 유전자형 일치 여부에 따라 최종적으로 매칭되지 않은 NGS 데이터를 제공할 수 있다.

[0126] 본 발명의 특징에 따르면, 출력부 (130) 는 html 형식으로 검증 결과를 제공하도록 더욱 구성될 수 있다. 그러나, 검증 결과는 출력부 (130) 에 의해 더욱 다양한 형식으로 출력될 수 있다.

[0127] 사용자는 출력부 (130) 를 통해 NGS 분석에 대한 사후 검증 결과를 확인할 수 있다.

[0128] 이때, 출력부 (130) 는 전술한 것에 제한되지 않고, 입력부 (120) 에 입력된 다양한 추가 정보들, 프로세서 (150) 에 의해 산출되거나 결정된 다양한 정보들을 디스플레이적으로 표시하도록 구성될 수 있다.

[0129] 저장부 (140) 는 수신부 (110) 를 통해 수신한 SNP 사이트 및/또는 파일명을 포함하는 복수의 NGS 데이터들을

저장하고, 입력부 (120) 를 통해 설정된 NGS 샘플 검증용 디바이스 (100) 의 지시를 저장하도록 구성될 수 있다. 나아가, 저장부 (140) 는 입력부 (120) 에 입력된, 다양한 추가 정보들을 저장하도록 더 구성될 수 있다. 또한, 저장부 (140) 는 후술될 프로세서 (150) 에 의해 산출된 유전자형 일치 점수, 유사도 점수, 프로세서 (150) 에 의해 결정된 매칭 결과를 저장하도록 구성될 수 있다.

[0130] 그러나, 전술한 것에 제한되지 않고 저장부 (140) 는, 프로세서 (150) 에 의해 결정된 다양한 정보들을 저장할 수 있다.

[0131] 프로세서 (150) 는 NGS 샘플 검증용 디바이스 (100) 의 NGS 샘플의 검증을 위한 구성 요소일 수 있다.

[0132] 이때, 프로세서 (150) 는 복수의 NGS 데이터에 대한 SNP 사이트의 서열 정보 일치 정도 및/ 복수의 NGS 데이터 중 선택된 임의의 두 개의 NGS 데이터에 또는 파일명의 유사 정도를 기초로 파일을 동일한 개체 유래의 샘플 여부를 결정하도록 구성된 알고리즘에 기초할 수도 있다. 그러나, 이에 제한되는 것은 아니다.

[0133] 프로세서 (150) 는, 표적 SNP 사이트에 대한 서열 정보를 기초로, 복수의 NGS 데이터의 동일한 개체 유래의 샘플 여부를 결정하도록 구성될 수 있다. 또한, 본 프로세서 (150) 는 미리 결정된, 맵핑 능력, GMAF의 수준, 개별 인구 내 MAF를 기초로 표적 SNP 사이트를 결정하도록 더 구성될 수 있다.

[0134] 본 발명의 특징에 따르면, 프로세서 (150) 는, 복수의 NGS 데이터의 표적 SNP 사이트에 대한 서열 정보를 기초로, 유전자형 일치 점수를 산출하고, 유전자형 일치 점수를 기초로 복수의 NGS 데이터를 동일 개체 유래의 샘플 여부에 따라 매칭하도록 구성될 수 있다.

[0135] 본 발명의 다른 특징에 따르면, 프로세서 (150) 는 복수의 NGS 데이터 중 상기 유전자형 일치 점수가 0.7 이상인 NGS 데이터, 바람직하게 유전자형 일치 점수가 0.8 이상인 NGS 데이터를 동일 개체 유래의 샘플 파일로 매칭하도록 구성될 수 있다.

[0136] 한편, NGS 데이터에 대한 상기 유전자형 일치 점수가 0.7 미만으로 산출될 경우 프로세서 (150) 는 상기 NGS 데이터를 샘플과 매칭되지 않는 NGS 데이터로 결정하거나, 복수의 NGS 데이터에 대한 매칭을 반복적으로 수행하도록 구성될 수 있다.

[0137] 본 발명의 또 다른 특징에 따르면, 프로세서 (150) 는 파일명을 기초로 복수의 NGS 데이터 중 임의로 선택된 두 개의 NGS 데이터에 대한 유사도 점수를 산출하고, 유사도 점수를 기초로 두 개의 NGS 데이터를 매칭하도록 구성될 수 있다.

[0138] 본 발명의 또 다른 특징에 따르면, 프로세서 (150) 는 파일명에 포함된 구분 문자를 기초로 NGS 데이터의 파일명 각각을 복수의 부분으로 분할하고, 복수의 부분에 대한 값의 출연 빈도를 산출하고, 출연 빈도를 기초로 두 개의 NGS 데이터의 유사도 점수를 산출하도록 구성될 수 있다.

[0139] 본 발명의 또 다른 특징에 따르면, 프로세서 (150) 는 파일명에 포함된 구분 문자를 기초로 NGS 데이터의 파일명 각각을 복수의 부분으로 분할하고, 복수의 부분에 대한 값의 분산 정도를 산출하고, 분산 정도를 기초로 유사도 점수를 산출하도록 구성될 수 있다.

[0140] 이때, 분산 정도는, 수의 부분에 대한 값의 출연 빈도를 기초로 산출될 수 있다.

[0141] 본 발명의 또 다른 특징에 따르면, 프로세서 (150) 는 두 개의 NGS 데이터에 대하여 산출된 유사도 점수가 가장 높을 경우, 두 개의 NGS 데이터를 한 쌍의 파일로 매칭하도록 구성될 수 있다.

[0142] 본 발명의 또 다른 특징에 따르면, 프로세서 (150) 는 표적 SNP 사이트에 대한 서열 정보를 기초로, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 제1 결정하고, 파일명을 기초로, NGS 데이터에 대한 유사도 점수를 산출하고, 유사도 점수를 기초로 NGS 데이터의 동일 개체 유래 샘플 여부를 제2 결정하도록 더욱 구성될 수 있다.

[0143] 이에, 본 발명의 NGS 샘플 검증용 디바이스 (100) 는, NGS 분석된 파일의 유전자형 및/또는 파일명에 기초하여, 샘플 또는 개체별로 매칭된 NGS 데이터를 제공할 수 있다. 나아가, 본 발명의 NGS 샘플 검증용 디바이스 (100) 는, 어느 NGS 데이터와도 매칭되지 않거나 잘못 매칭된 NGS 데이터를 더욱 제공할 수 있다. 이에, 사용자는, 복수의 대상 샘플에 대한 NGS 분석 결과와 함께, 샘플 또는 개체의 종류에 따른 매칭 여부의 검증 결과를 제공받을 수 있다.

[0144] 이하에서는 도 2a 및 2b를 참조하여, 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법을 구체적으로 설명한다.

도 2a 및 2b는 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법의 절차를 예시적으로 도시한 것이다.

[0145] 도 2a를 참조하면, 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법은 먼저, 미리 결정된 표적 SNP 사이트에 대한 서열 정보를 포함하는 복수의 NGS 데이터를 수신하고 (S210), 수신된 NGS 데이터의 종류에 따라 SNP 사이트를 결정하고 (S220), SNP 사이트에 대한 서열 정보를 기초로 복수의 NGS 데이터에 대한 동일 개체 유래의 샘플 여부를 결정하고 (S230), 최종적으로 결과를 제공한다 (S240).

[0146] 보다 구체적으로, 도 2b를 참조하면, 복수의 NGS 데이터를 수신하는 단계 (S210) 에서, WES (whole exome sequencing) 파일 또는 RNA 시퀀싱 (RNA sequencing) 파일 (212) 이 수신될 수 있다.

[0147] 본 발명의 특징에 따르면, 복수의 NGS 데이터를 수신하는 단계 (S210) 에서, 표적 시퀀싱 (targeted sequencing) 파일 (214) 이 수신될 수 있다.

[0148] 본 발명의 다른 특징에 따르면, 복수의 NGS 데이터를 수신하는 단계 (S210) 에서, BED 파일 (216) 이 수신될 수 있다.

[0149] 이때, WES 파일, RNA 시퀀싱 파일 및 표적 시퀀싱 파일은 BAM 파일의 포맷을 가질 수 있으나, NGS 데이터의 형식은 이에 제한되는 것이 아니다.

[0150] 다음으로, SNP 사이트를 결정하는 단계 (S220) 에서, 수신된 NGS 데이터의 종류 및 미리 결정된 맵핑 능력, GMAF, 개별 인구 MAF에 따라 표적 SNP 사이트가 결정된다.

[0151] 본 발명의 특징에 따르면, SNP 사이트를 결정하는 단계 (S220) 에서 WES 분석 파일 또는 RNA 시퀀싱 분석 파일일 경우, 미리 결정된 수준 이상의 맵핑 능력을 갖는 SNP 사이트 중, GMAF가 0.45 내지 0.55이고, 개별 인구 MAF 0.1 내지 0.9인 SNP 사이트가 표적 SNP 사이트로 결정될 수 있다.

[0152] 본 발명의 다른 특징에 따르면, SNP 사이트를 결정하는 단계 (S220) 에서 NGS 데이터가 표적 시퀀싱 분석 파일일 경우, GMAF가 0.1 내지 0.9인 SNP 사이트가 표적 SNP 사이트로 결정될 수 있다.

[0153] 보다 구체적으로, 도 2b를 다시 참조하면, SNP 사이트를 결정하는 단계 (S220) 에서는, WES 파일 또는 RNA 시퀀싱 파일 (212) 내에서 GMAF가 0.45 내지 0.55인 853 개의 대형 분석 패널에 따른 표적 SNP 사이트 (222) 가 결정될 수 있다. 또한, SNP 사이트를 결정하는 단계 (S220) 에서는, 표적 시퀀싱 파일 (214) 및 BED 파일 (216) 내에서 GMAF가 0.1 내지 0.9인, 최소 200 개 이상의 소형 분석 패널에 따른 표적 SNP 사이트 (224) 가 결정될 수 있다.

[0154] 본 발명의 또 다른 특징에 따르면, SNP 사이트를 결정하는 단계 (S220) 에서, 코딩 영역에서의 변이, 콜링 필터 (Calling filter) 를 통과한 변이, 렌덤 포레스트 모델 (Random forest model) 에 의해 통과된 변이, dbSNP로 보고된 변이, QB (QualByDepth) 가 2.0 이상이고, RMSMQ (root mean square mapping quality) 가 50 이상인 변이 중 적어도 하나를 만족하는 SNP가 표적 SNP 사이트로 결정될 수 있다.

[0155] 본 발명의 또 다른 특징에 따르면, SNP 사이트를 결정하는 단계 (S220) 에서, 낮은 복합 영역 (low complex region) 가 아닌 변이, 세그먼트 복제 영역 (segment duplicated region) 이 아닌 변이, 단순 반복 영역 (simple repeat region) 이 아닌 변이 중 적어도 하나를 SNP가 표적 SNP 사이트로 결정될 수 있다.

[0156] 다음으로, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S230) 에서, 복수의 NGS 데이터에 대한 표적 SNP 사이트의 염기 서열의 일치 정도를 기초로 두 개의 NGS 데이터가 매칭될 수 있다.

[0157] 본 발명의 특징에 따르면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S230) 에서, 표적 SNP 사이트에 대한 서열 정보에 기초하여 유전자형 일치 점수 (genotype concordance score) 가 산출되고, 유전자형 일치 점수에 기초하여 NGS 데이터가 매칭될 수 있다.

[0158] 본 발명의 다른 특징에 따르면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S230) 에서, NGS 데이터에 대한 상기 유전자형 일치 점수가 0.7 인 NGS 데이터 바람직하게 0.8 이상인 NGS 데이터는 동일한 개체 유래의 한 쌍의 파일로 매칭될 수 있다.

[0159] 보다 구체적으로, 도 2b를 다시 참조하면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S230) 에서, 표적 SNP 사이트의 염기 서열의 일치 정도, 즉 유전자형에 따라 동일한 개체 유래된 것으로 매칭된 한 쌍의 파일 (232) 또는 매칭되지 않은 파일 (234) 이 결정될 수 있다.

[0160] 예를 들어, 복수의 NGS 데이터 중 유전자형 일치 점수가 0.7 미만인 NGS 데이터는 샘플과 매칭되지 않는 NGS 데

이터로 결정될 수 있다.

[0161] 한편, 본 발명의 또 다른 특징에 따르면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S230)에서, 매칭되지 않은 파일 (234) 이 결정될 경우, 즉 모든 파일이 매칭되지 않을 경우, 복수의 NGS 데이터에 대한 동일 개체 유래 샘플 여부를 결정하는 단계 (S230) 가 다시 수행될 수도 있다.

[0162] 마지막으로, 동일 개체 유래 샘플 여부를 제공하는 단계 (S240) 에서, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계 (S230) 의 결과가 제공될 수 있다.

[0163] 보다 구체적으로, 도 2b를 다시 참조하면, 동일 개체 유래 샘플 여부를 제공하는 단계 (S240) 에서, 유전자형에 따라 매칭된 한 쌍의 파일 (232) 또는 매칭되지 않은 파일 (234) 이 제공될 수 있다.

[0164] 이상의 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법의 결과로, 매칭되거나 매칭되지 않은 NGS 분석 결과가 사용자에게 제공될 수 있다. 이에, 본 발명은, NGS 분석 결과 제공에 있어서 빈번하게 발생할 수 있는 샘플간 혼합에 따른 검출의 신뢰도 떨어뜨리거나 부정확한 결과를 제공하는 문제점들을 해결할 수 있어 NGS 샘플 검증 시스템에 적용될 수 있다. 나아가, 본 발명은, NGS 분석 디바이스에 소프트웨어의 구성으로서 포함되어 제공될 수도 있다.

[0165] 이하에서는 도 3a 내지 3g를 참조하여, 본 발명의 다른 실시예에 따른 NGS 샘플 검증 방법을 구체적으로 설명한다. 도 3a 내지 3g는 본 발명의 다른 실시예에 따른 NGS 샘플 검증 방법의 절차를 예시적으로 도시한 것이다.

[0166] 도 3a를 참조하면, 본 발명의 다른 실시예에 따른 NGS 샘플 검증 방법은 먼저, 미리 결정된 파일명을 포함하는 복수의 NGS 데이터를 수신하고 (S310), 파일명을 기초로 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하고 (S320), 최종적으로 동일 개체 유래 샘플 여부를 제공한다 (S330).

[0167] 보다 구체적으로, 도 3b를 참조하면, 복수의 NGS 데이터를 수신하는 단계 (S310) 에서, WES (whole exome sequencing) 파일 또는 RNA 시퀀싱 (RNA sequencing) 파일 (312) 이 수신될 수 있다.

[0168] 본 발명의 특징에 따르면, 복수의 NGS 데이터를 수신하는 단계 (S310) 에서, 표적 시퀀싱 (targeted sequencing) 파일 (314) 이 수신될 수 있다.

[0169] 본 발명의 다른 특징에 따르면, 복수의 NGS 데이터를 수신하는 단계 (S310) 에서, BED 파일 (316) 이 수신될 수 있다.

[0170] 이때, WES 파일, RNA 시퀀싱 파일 및 표적 시퀀싱 파일은 BAM 파일의 포맷을 가질 수 있으나, NGS 데이터의 형식은 이에 제한되는 것이 아니다.

[0171] 다음으로, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 에서, 파일명을 기초로 복수의 NGS 데이터에 대한 유사도 점수를 산출되고 유사도 점수를 기초로 복수의 NGS 데이터가 매칭될 수 있다.

[0172] 본 발명의 특징에 따르면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 에서, 파일명에 포함된 구분 문자에 기초하여 NGS 데이터의 파일명 각각이 복수의 부분으로 분할되고, 복수의 부분에 대한 값의 출연 빈도가 산출되며, 출연 빈도에 기초하여 복수의 NGS 데이터 중 선택된 두 개의 NGS 데이터의 유사도 점수가 산출될 수 있다.

[0173] 본 발명의 다른 특징에 따르면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 에서, 파일명에 포함된 구분 문자에 기초하여 NGS 데이터의 파일명 각각이 복수의 부분으로 분할되고, 복수의 부분에 대한 값의 분산 정도가 산출되며, 분산 정도에 기초하여 복수의 NGS 데이터 중 선택된 두 개의 NGS 데이터의 유사도 점수가 산출될 수 있다.

[0174] 이때, 분산 정도는, 복수의 부분에 대한 값의 출연 빈도에 기초하여 산출될 수 있다.

[0175] 본 발명의 또 다른 특징에 따르면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 에서, 복수의 NGS 데이터 중 선택된 두 개의 NGS 데이터에 대하여, 복수의 부분 중 선택된 하나의 부분에 대한 값의 동일 여부 및 분산 정도에 기초하여 유사도 점수가 산출될 수 있다.

[0176] 본 발명의 또 다른 특징에 따르면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 에서, 두 개의 NGS 데이터에 대한 파일명의 유사도 점수는, 구분 문자에 의해 분할된 파일명의 벡터화를 통해 산출될 수 있다.

[0177] 예를 들어, 도 3c를 참조하면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 에서, 복

수의 NGS 데이터, 예를 들어 28 개의 NGS 데이터는, bar (-, _), dot (.), slash (/), colon (:), semicolon (;), apostrophe (')와 같은 구분 문자에 의해 8 개의 부분으로 분할될 수 있다. 그 다음, 복수의 NGS 데이터 각각은 벡터화될 수 있다. 다음으로, 도 3d를 참조하면, 복수의 NGS 데이터의 8 개의 복수의 부분 각각에 대한 값의 출연 빈도에 기초하여 복수의 부분 각각에 대한 카운트 벡터 (count vector) 가 생성될 수 있다. 보다 구체적으로, 파일명에서 분할된 8 개의 부분 중 첫 번째 부분을 참조하면, 총 5 개의 값의 출연 빈도가 각각 8 회, 4 회, 4 회, 6 회, 6 회 총 28 회) 로 나타남에 따라, 첫 번째 부분에 대한 카운트 벡터, '[8, 4, 4, 6, 6]'가 생성될 수 있다. 만약, 선택된 부분에 대하여 한 개의 값이 존재한다면, 상기 선택된 부분에 대한 카운트 벡터는 '[28]'일 수 있다. 다음으로, 도 3e를 참조하면, 파일명의 8 개의 복수의 부분 각각에 대하여 생성된 카운트 벡터에 기초하여, 28 개의 NGS 데이터의 파일명 전체에 대한 엔트로피 벡터 (entropy vector) 가 생성된다. 이때, 엔트로피 벡터 (S) 는 하기 [수학식 1] 에 의해 산출될 수 있다.

[0178] [수학식 1]

[0179] $S= -sum(p_k * log(p_k))$

[0180] 보다 구체적으로, 28 개의 NGS 데이터의 8 개의 복수의 부분 각각에 대하여 [1.5741, 1.4900, 06931, 0,0,0,0,0] 의 엔트로피 벡터 (S) 가 산출될 수 있다. 이때, 엔트로피 벡터 (S) 를 참조하면, 8 개의 복수의 부분 중, 첫 번째 내지 세 번째 부분의 값이 0 보다 큰 결과는, 해당 부분에서 28 개의 파일의 각각의 값에 대한 차이가 나머지 부분 (네 번째 내지 여덟 번째 부분) 보다 큰 것을 의미할 수 있다. 다음으로, 도 3f를 참조하면, 28 개의 NGS 데이터 중 선택된 두 개의 NGS 데이터에 대하여 산출된 엔트로피 벡터 값 및 복수의 부분에 대한 값의 동일 여부에 기초하여 유사도 점수가 최종적으로 산출될 수 있다. 보다 구체적으로, 두 개의 NGS 데이터의 파일명에 대한 유사도 점수의 산출에 있어서, 첫 번째, 두 번째, 네 번째 내지 여덟 번째 부분의 값이 서로 동일함에 따라, 상기 부분들에 해당하는 엔트로피 벡터의 값들 (1.5741, 1.4900, 0, 0, 0, 0, 0) 이 합산된다. 나아가, 세 번째 부분의 값이 각각 상이함에 따라 (FP 및 BL), 세 번째 부분에 해당하는 엔트로피 벡터의 값 (0.6931) 이 감산된다. 최종적으로, 두 개의 NGS 데이터에 대한 유사도 점수는 2.371점으로 산출될 수 있다.

[0181] 본 발명의 또 다른 특징에 따르면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 에서, 복수의 NGS 데이터 중 선택된 두 개의 NGS 데이터에 대하여 산출된 유사도 점수가 가장 높을 경우, 두 개의 NGS 데이터가 한 쌍의 파일로 매칭될 수 있다.

[0182] 예를 들어, 도 3f를 다시 참조하면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 에서, '24-S006-FP.RGadded.marked.realigned.fixed.bam', '24-S006-BL.RGadded.marked.realigned.fixed.bam'의 두 개의 파일에 대하여 산출된 유사도 점수가 2.371로 가장 높음에 따라, 상기 두 개의 파일이 한 쌍의 파일로 결정될 수 있다.

[0183] 즉, 도 3b를 다시 참조하면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 의 결과로, 파일명의 유사도 정도에 따라 매칭된 한 쌍의 파일 (322), 미스매칭된 파일 (324), 또는 매칭되지 않은 파일 (326) 이 결정될 수 있다.

[0184] 예를 들어, 도 3g의 (a)를 참조하면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 에서, NGS 데이터 (342 (a)) 은 NGS 데이터 (342 (b)) 과 매칭 되었을 때 가장 높은 유사도 점수 (3.0641) 가 나타남에 따라, 두 개의 NGS 데이터 (342 (a), 342 (b)) 은 동일한 샘플 (또는, 개체) 로부터 유래된 한 쌍의 NGS 데이터 (“Match “) 로 결정될 수 있다. 도 3g의 (b)를 참조하면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 에서, NGS 데이터 (342 (a)) 은 NGS 데이터 (342 (b)) 과 매칭 되었을 때 가장 높은 유사도 점수 (3.0641) 가 나타남에 따라, 두 개의 NGS 데이터 (342 (a), 344 (a)) 는 서로 상이한 샘플 (또는, 개체) 로부터 유래된 한 쌍의 NGS 데이터로서 미스매칭된 파일 (“Swapped “) 로 결정될 수 있다. 도 3g의 (c)를 참조하면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 에서, NGS 데이터 (342 (a)) 은 NGS 데이터 (342 (b)) 과 매칭 되었을 때 가장 높은 유사도 점수 (3.0641) 가 나타남에 따라, 단일로 존재하는 NGS 데이터 (342 (a)) 은 어느 파일과 도 매칭되지 않은 것 (“Orphan “) 으로 결정될 수 있다.

[0185] 한편, 본 발명의 또 다른 특징에 따르면, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부 결정하는 단계 (S320) 에서, 매칭되지 않은 파일 (324) 이 결정될 경우, 즉 모든 파일이 매칭되지 않을 경우, 복수의 NGS 데이터 중 두 개의 NGS 데이터를 결정하고, SNP 사이트의 서열 정보를 기초로 새로운 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계 (S320) 가 다시 수행될 수도 있다.

[0186] 예를 들어, 두 개의 NGS 데이터에 대한 유사도 점수가 미리 결정된 수준 미만일 경우, 복수의 NGS 데이터 중 선택된 임의의 새로운 두 개의 NGS 데이터가 결정되고 파일명을 기초로 새로운 두 개의 파일이 재 매칭될 수도 있다.

[0187] 마지막으로, 동일 개체 유래 샘플 여부를 제공하는 단계 (S330) 에서, 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 결정하는 단계 (S320) 의 결과가 제공될 수 있다.

[0188] 보다 구체적으로, 도 3b를 다시 참조하면, 동일 개체 유래 샘플 여부를 제공하는 단계 (S330) 에서, 검증 결과 (332) 가 제공될 수 있다. 이때, 검증 결과 (332) 는 파일명의 유사도에 따라 매칭된 한 쌍의 파일 (322), 미스매칭된 파일 (324), 또는 매칭되지 않은 파일 (326) 을 포함할 수 있다.

[0189] 본 발명의 특징에 따르면, 동일 개체 유래 샘플 여부를 제공하는 단계 (S330) 에서, 검증 결과 (322) 는 html 파일의 포맷으로 제공될 수 있으나 이에 제한되는 것이 아니다.

[0190] 이상의 본 발명의 다른 실시예에 따른 NGS 샘플 검증 방법의 결과로, 매칭되거나 매칭되지 않은 NGS 분석 결과, 미스 매칭된 결과가 사용자에게 제공될 수 있다. 이에, 본 발명은, NGS 분석 결과 제공에 있어서 빈번하게 발생할 수 있는 샘플간 혼합에 따른 검출의 신뢰도 떨어뜨리거나 부정확한 결과를 제공하는 문제점들을 해결할 수 있어 NGS 샘플 검증 시스템에 적용될 수 있다. 나아가, 본 발명은, NGS 분석 디바이스에 소프트웨어의 구성으로서 포함되어 제공될 수도 있다.

[0191] 이하에서는 도 4a 및 4b를 참조하여, 도 4a 및 4b는 본 발명의 또 다른 실시예에 따른 NGS 샘플 검증 방법의 절차를 예시적으로 도시한 것이다.

[0192] 도 4a를 참조하면, 본 발명의 또 다른 실시예에 따른 NGS 샘플 검증 방법은 먼저, 미리 결정된 표적 SNP 사이트에 대한 서열 정보 및 파일명을 포함하는 복수의 NGS 데이터를 수신하고 (S410), 수신된 NGS 데이터의 종류에 따라 SNP 사이트를 결정하고 (S420), SNP 사이트에 대한 서열 정보를 기초로 복수의 NGS 데이터의 동일 개체 유래 샘플 여부를 제1 결정하고 (S430), 파일명을 기초로 제2 결정하여 (S440) 최종적으로 동일 개체 유래 샘플 여부를 제공한다 (S450).

[0193] 보다 구체적으로, 도 4b를 참조하면, 복수의 NGS 데이터를 수신하는 단계 (S410) 에서, WES 파일 또는 RNA 시퀀싱 파일 (412) 이 수신될 수 있다.

[0194] 본 발명의 특징에 따르면, 복수의 NGS 데이터를 수신하는 단계 (S410) 에서, 표적 시퀀싱 파일 (414) 이 수신될 수 있다.

[0195] 본 발명의 다른 특징에 따르면, 복수의 NGS 데이터를 수신하는 단계 (S410) 에서, BED 파일 (416) 이 수신될 수 있다.

[0196] 이때, WGS 파일, WES 파일, RNA 시퀀싱 파일 및 표적 시퀀싱 파일은 BAM 파일의 포맷을 가질 수 있으나, NGS 데이터의 형식은 이에 제한되는 것이 아니다.

[0197] 다음으로, SNP 사이트를 결정하는 단계 (S420) 에서, 수신된 NGS 데이터의 종류 및 SNP 사이트의 맵핑 능력, 분석 패널에 따라 미리 결정된 GMAF, 개별 인구 내 MAF에 따라 표적 SNP 사이트가 결정된다.

[0198] 본 발명의 특징에 따르면, SNP 사이트를 결정하는 단계 (S420) 에서 WES 분석 파일 또는 RNA 시퀀싱 분석 파일일 경우, 맵핑 능력이 미리 결정된 수준 이상인 SNP 중 GMAF가 0.45 내지 0.55이거나, 개별 인구 MAF가 0.1 내지 0.9인 SNP 사이트가 표적 SNP 사이트로 결정될 수 있다.

[0199] 본 발명의 다른 특징에 따르면, SNP 사이트를 결정하는 단계 (S420) 에서 NGS 데이터가 표적 시퀀싱 분석 파일일 경우, GMAF가 0.1 내지 0.9인 SNP 사이트가 표적 SNP 사이트로 결정될 수 있다.

[0200] 보다 구체적으로, 도 4b를 다시 참조하면, SNP 사이트를 결정하는 단계 (S420) 에서는, WES 파일 또는 RNA 시퀀싱 파일 (412) 내에서 GMAF가 0.45 내지 0.55이고, 개별 인구 MAF가 0.1 내지 0.9인 853 개의 대형 분석 패널에 따른 표적 SNP 사이트 (422) 가 결정될 수 있다. 또한, SNP 사이트를 결정하는 단계 (S420) 에서는, 표적 시퀀싱 파일 (414) 및 BED 파일 (416) 내에서 GMAF가 0.1 내지 0.9인, 최소 200 개 이상의 소형 분석 패널에 따른 표적 SNP 사이트 (424) 가 결정될 수 있다.

[0201] 본 발명의 또 다른 특징에 따르면, SNP 사이트를 결정하는 단계 (S420) 에서, 코딩 영역에서의 변이, 콜링 필터 (Calling filter) 를 통과한 변이, 렌덤 포레스트 모델 (Random forest model) 에 의해 통과된 변이, dbSNP로

보고된 변이, QB (QualByDepth) 가 2.0 이상이고, RMSMQ (root mean square mapping quality) 가 50 이상인 변이 중 적어도 하나를 만족하는 SNP가 표적 SNP 사이트로 결정될 수 있다.

[0202] 본 발명의 또 다른 특징에 따르면, SNP 사이트를 결정하는 단계 (S420) 에서, 낮은 복합 영역 (low complex region) 가 아닌 변이, 세그먼트 복제 영역 (segment duplicated region) 이 아닌 변이, 단순 반복 영역 (simple repeat region) 이 아닌 변이 중 적어도 하나를 SNP가 표적 SNP 사이트로 결정될 수 있다.

[0203] 다음으로, 동일 개체 유래 샘플 여부를 제1 결정하는 단계 (S430) 에서, 복수의 NGS 데이터에 대한 표적 SNP 사이트의 염기 서열의 일치 정도를 기초로 NGS 데이터가 매칭될 수 있다.

[0204] 본 발명의 특징에 따르면, 동일 개체 유래 샘플 여부를 제1 결정하는 단계 (S430) 에서, 표적 SNP 사이트에 대한 서열 정보에 기초하여 유전자형 일치 점수가 산출되고, 유전자형 일치 점수에 기초하여 NGS 데이터가 매칭될 수 있다.

[0205] 본 발명의 다른 특징에 따르면, 동일 개체 유래 샘플 여부를 제1 결정하는 단계 (S430) 에서, NGS 데이터에 대한 상기 유전자형 일치 점수가 0.7 이상일 경우, 바람직하게 0.8 이상인 NGS 데이터는 경우 한 쌍의 파일로 매칭될 수 있다.

[0206] 보다 구체적으로, 도 4b를 다시 참조하면, 동일 개체 유래 샘플 여부를 제1 결정하는 단계 (S430) 에서, 표적 SNP 사이트의 염기 서열의 일치 정도, 즉 유전자형에 따라 매칭된 한 쌍의 파일 (432, 434) 또는 매칭되지 않은 파일 (436) 이 결정될 수 있다.

[0207] 예를 들어, NGS 데이터에 대한 상기 유전자형 일치 점수가 0.7 미만인 NGS 데이터는, 샘플과 매칭되지 않는 NGS 데이터로 결정될 수도 있다.

[0208] 본 발명의 또 다른 특징에 따르면, 동일 개체 유래 샘플 여부를 제1 결정하는 단계 (S430) 에서, 매칭되지 않은 파일 (436) 이 결정될 경우, 즉 모든 파일이 매칭되지 않을 경우, 동일 개체 유래 샘플 여부를 제1 결정하는 단계 (S430) 가 다시 수행될 수도 있다.

[0209] 다음으로, 동일 개체 유래 샘플 여부를 제2 결정하는 단계 (S440) 에서 제1 결정된 복수의 NGS 중 선택된 임의의 두 개의 NGS 데이터의 파일명을 기초로 두 개의 NGS 데이터에 대한 유사도 점수를 산출되고 유사도 점수를 기초로 두 개의 NGS 데이터가 매칭될 수 있다.

[0210] 본 발명의 특징에 따르면, 동일 개체 유래 샘플 여부를 제2 결정하는 단계 (S440) 에서, 파일명에 포함된 구분 문자에 기초하여 NGS 데이터의 파일명 각각이 복수의 부분으로 분할되고, 복수의 부분에 대한 값의 출연 빈도가 산출되며, 출연 빈도에 기초하여 복수의 NGS 데이터 중 선택된 임의의 두 개의 NGS 데이터의 유사도 점수가 산출될 수 있다.

[0211] 본 발명의 다른 특징에 따르면, 동일 개체 유래 샘플 여부를 제2 결정하는 단계 (S440) 에서, 파일명에 포함된 구분 문자에 기초하여 NGS 데이터의 파일명 각각이 복수의 부분으로 분할되고, 복수의 부분에 대한 값의 분산 정도가 산출되며, 분산 정도에 기초하여 복수의 NGS 데이터 중 선택된 임의의 두 개의 NGS 데이터의 유사도 점수가 산출될 수 있다.

[0212] 이때, 분산 정도는, 복수의 부분에 대한 값의 출연 빈도에 기초하여 산출될 수 있다.

[0213] 본 발명의 또 다른 특징에 따르면, 동일 개체 유래 샘플 여부를 제2 결정하는 단계 (S440) 에서, 복수의 NGS 데이터 중 선택된 임의의 두 개의 NGS 데이터에 대하여, 복수의 부분 중 선택된 하나의 부분에 대한 값의 동일 여부 및 분산 정도에 기초하여 유사도 점수가 산출될 수 있다.

[0214] 본 발명의 또 다른 특징에 따르면, 동일 개체 유래 샘플 여부를 제2 결정하는 단계 (S440) 에서, 복수의 NGS 데이터 중 선택된 임의의 두 개의 NGS 데이터에 대한 파일명의 유사도 점수는, 구분 문자에 의해 분할된 파일명의 벡터화를 통해 산출될 수 있다.

[0215] 본 발명의 또 다른 특징에 따르면, 동일 개체 유래 샘플 여부를 제2 결정하는 단계 (S440) 에서, 복수의 NGS 데이터 중 선택된 임의의 두 개의 NGS 데이터에 대하여 산출된 유사도 점수가 가장 높을 경우, 두 개의 NGS 데이터가 한 쌍의 파일로 매칭될 수 있다.

[0216] 보다 구체적으로, 도 4b를 다시 참조하면, 동일 개체 유래 샘플 여부를 제2 결정하는 단계 (S440) 에서, 동일 개체 유래 샘플 여부를 제1 결정하는 단계 (S430) 의 결과로 유전자형에 따라 매칭된 한 쌍의 파일 (432, 434)

은 파일명의 유사도에 따라, 최종 매칭된 한 쌍의 파일 (442) 로 결정되거나, 최종 미스매칭된 한 쌍의 파일 (444) 로 결정될 수 있다. 나아가, 동일 개체 유래 샘플 여부를 제2 결정하는 단계 (S440) 에서, 동일 개체 유래 샘플 여부를 제1 결정하는 단계 (S430) 의 결과로 매칭되지 않은 파일 (436) 은, 최종 매칭되지 않은 파일 (446) 로 결정될 수 있다.

[0217] 본 발명의 또 다른 특징에 따르면, 동일 개체 유래 샘플 여부를 제2 결정하는 단계 (S440) 에서, 최종 매칭되지 않은 파일 (446) 이 결정될 경우, 즉 모든 파일이 매칭되지 않을 경우, 동일 개체 유래 샘플 여부를 제2 결정하는 단계 (S430) 가 다시 수행될 수도 있다.

[0218] 예를 들어, 두 개의 NGS 데이터에 대한 유사도 점수가 미리 결정된 수준 미만일 경우, 복수의 NGS 데이터 중 선택된 임의의 새로운 두 개의 NGS 데이터가 결정되고 파일명을 기초로 새로운 두 개의 파일이 재 매칭될 수도 있다.

[0219] 마지막으로, 동일 개체 유래 샘플 여부를 제공하는 단계 (S450) 에서, 동일 개체 유래 샘플 여부를 제2 결정하는 단계 (S440) 의 결과가 제공될 수 있다.

[0220] 보다 구체적으로, 도 4b를 다시 참조하면, 동일 개체 유래 샘플 여부를 제공하는 단계 (S450) 에서, 최종 매칭에 따른 검증 결과 (452) 가 제공될 수 있다. 이때, 검증 결과 (452) 는 유전자형의 일치 정도 및 파일명의 유사도에 따라 최종 매칭된 한 쌍의 파일 (442), 최종 미스매칭된 한 쌍의 파일 (444), 및 최종 매칭되지 않은 파일 (446) 을 포함할 수 있다.

[0221] 본 발명의 특징에 따르면, 동일 개체 유래 샘플 여부를 제공하는 단계 (S450) 에서, 검증 결과 (452) 는 html 파일의 포맷으로 제공될 수 있으나 이에 제한되는 것이 아니다.

[0222] 이상의 본 발명의 또 다른 실시예에 따른 NGS 샘플 검증 방법의 결과로, 매칭되거나 매칭되지 않은 NGS 분석 결과, 미스 매칭된 결과가 사용자에게 제공될 수 있다. 특히 본 발명은 사용자의 핸들링에 따른 휴먼 오류, 예를 들어 샘플 및 분석 데이터의 미스매칭과 같은 오류를 개선하여, 샘플 및 NGS 데이터를 매칭하여 제공해주는 NGS 샘플 검증 시스템을 제공할 수 있다.

[0223] 이에, 본 발명은, NGS 분석 결과 제공에 있어서 샘플간 혼합에 따른 검출의 신뢰도 떨어뜨리거나 부정확한 결과를 제공하는 문제점들을 해결할 수 있어 다양한 NGS 샘플 검증 시스템에 적용될 수 있다. 나아가, 본 발명은, NGS 분석 디바이스에 소프트웨어의 구성으로서 포함되어 제공될 수도 있다.

[0224] **실시예 1: 유전자형 일치 점수 (concordance score) 에 따른 NGS 샘플 검증**

[0225] 이하에서는 도 5를 참조하여, 본 발명의 다양한 실시예에 적용되는 SNP 사이트의 염기 서열에 기초한 NGS 샘플 검증에 대한 평가 결과를 구체적으로 설명한다.

[0226] 도 5는 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법에 의해, 유전자형에 따라 매칭된 NGS 데이터 쌍 또는 미스매칭된 NGS 데이터 쌍의 유전자형 일치 점수를 도시한 것이다.

[0227] 이때, 유전자형 일치 점수는, NGS 데이터에 대한, 분석 패널의 종류에 따라 미리 결정된 표적 SNP 사이트의 염기 서열에 기초하여 산출될 수 있다.

[0228] 도 5의 (a)를 참조하면, WES의 대형 분석 패널에 따른 NGS 데이터에 대한, 표적 SNP 사이트의 개수 및 매칭여부에 따른 유전자형 일치 점수의 분포가 도시된다. 보다 구체적으로, 매칭된 두 개의 NGS 데이터 대부분은 1에 가까운 유전자형 일치 점수를 갖는 것으로 나타난다. 이와 대조적으로, 매칭되지 않은 NGS 데이터 대부분은 0.4의 낮은 유전자형 일치 점수를 갖는 것으로 나타난다.

[0229] 도 5의 (b) 및 (c)를 참조하면, RNA-seq 및 WES/RNA-seq 각각의 분석 패널에 따른 NGS 데이터에 대한, 표적 SNP 사이트의 개수 및 매칭여부에 따른 유전자형 일치 점수의 분포가 도시된다. 보다 구체적으로, 매칭된 두 개의 NGS 데이터 대부분은 1에 가까운 유전자형 일치 점수를 갖는 것으로 나타난다. 이와 대조적으로, 매칭되지 않은 NGS 데이터 대부분은 0.25 내지 0.5의 낮은 유전자형 일치 점수를 갖는 것으로 나타난다.

[0230] 도 5의 (d) 및 (e)를 참조하면, KCSG (Korean Cancer Study Group), KLCC (Korean Lung Cancer Consortium) 코호트의 표적 시퀀싱 분석 패널에 따른 NGS 데이터에 대한, 표적 SNP 사이트의 개수 및 매칭여부에 따른 유전자형 일치 점수의 분포가 도시된다. 보다 구체적으로, 매칭된 두 개의 NGS 데이터 대부분은 1에 가까운 유전자형 일치 점수를 갖는 것으로 나타난다. 이와 대조적으로, 매칭되지 않은 NGS 데이터 대부분은 0.4 내지 0.6의 낮은 유전자형 일치 점수를 갖는 것으로 나타난다.

[0231] 이상의 실시예 1에 따르면, 유전자형 일치 점수는, 매칭된 두 개의 NGS 데이터 또는 매칭되지 않은 NGS 데이터에 대하여 서로 상이한 점수 영역에서 분포되어 있는 것으로 나타난다. 특히, KCSG 및 KLCC와 같은 소형 분석 패널에 따른 NGS 데이터는 소수의 SNP 사이트를 고려했음에도 매칭여부에 따라 유전자형 일치 점수가 상이하게 분포한 것으로 나타난다.

[0232] 이러한 결과는, 본 발명의 다양한 실시예에서 제공되는 SNP 사이트의 유전자형 일치 점수에 따라 NGS 데이터를 매칭할 경우 매칭 결과에 대한 신뢰도가 높다는 것을 의미할 수 있다. 나아가, 상기와 같은 결과는, 본 발명의 다양한 실시예에 따른 NGS 샘플 검증 방법이, 소형 분석 패널을 포함한 다양한 염기서열 분석 패널에 적용될 수 있음을 의미할 수 있다.

[0233] **실시예 2: 파일명에 따른 NGS 샘플 검증**

[0234] 이하에서는 도 6a 내지 6f를 참조하여, 본 발명의 다양한 실시예에 적용되는 파일명에 기초한 NGS 샘플 검증에 대한 평가 결과를 구체적으로 설명한다.

[0235] 도 6a 내지 6f는 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법에 의해, 파일명에 따라 매칭된 NGS 데이터 쌍 또는 미스매칭된 NGS 데이터 쌍의 평가 결과를 도시한 것이다.

[0236] 본 평가에 이용된 NGS 데이터는, SNP 사이트의 염기 서열 즉, 유전자형에 따른 유전자형 일치 점수가 미리 산출된 파일일 수 있다. 그러나, 이에 제한되는 것은 아니다.

[0237] 도 6a를 참조하면, WES, RNA-seq와 KCSG 및 KLCC의 표적 시퀀 시퀀싱의 모든 분석 패널에서, NGS 데이터의 파일명에 기초한 매칭의 정확도가 100 %로 나타난다.

[0238] 이러한 결과는, 파일명에 기초하여 2 개의 NGS 데이터를 매칭한 결과가 높은 신뢰도를 가진다는 것을 의미할 수 있다.

[0239] 도 6b를 참조하면, WES의 분석 패널에 따른 NGS 데이터의 파일명에 기초하여 산출된 유사도 점수에 기초하여 매칭된 파일들이 도시된다.

[0240] 보다 구체적으로, 'C347.TCGA-05-4244-01A-01D-1105-08.5_gdc_realn.bam', 'C347.TCGA-05-4244-10A-01D-1105-08.5_gdc_realn.bam', 'C347.TCGA-05-4249-01A-01D-1105-08.5_gdc_realn.bam', 'C347.TCGA-05-4249-10A-01D-1105-08.5_gdc_realn.bam' 및 'C347.TCGA-05-4250-01A-01D-1105-08.5_gdc_realn.bam'의 5 개의 파일 각각에 대하여, 'C347.TCGA-05-4244-10A-01D-1105-08.5_gdc_realn.bam', 'C347.TCGA-05-4244-01A-01D-1105-08.5_gdc_realn.bam', 'C347.TCGA-05-4249-10A-01D-1105-08.5_gdc_realn.bam', 'C347.TCGA-05-4249-01A-01D-1105-08.5_gdc_realn.bam' 및 'C347.TCGA-05-4250-10A-01D-1105-08.5_gdc_realn.bam'의 각각의 파일이 9.229892145 점의 가장 높은 유사도 점수를 갖는 파일 (Best scored file) 로 결정되었다.

[0241] 즉, 파일명의 유사도가 높은 두 개의 NGS 데이터가 한 쌍의 파일로서 매칭되어 제공될 수 있다.

[0242] 도 6c를 참조하면, RNA_seq의 분석 패널에 따른 NGS 데이터의 파일명에 기초하여 산출된 유사도 점수에 기초하여 매칭된 파일들이 도시된다.

[0243] 보다 구체적으로, 'TCGA-BC-A10Q-01A.RNA.bam', 'TCGA-BC-A10Q-11A.RNA.bam', 'TCGA-BC-A10R-01A.RNA.bam', 'TCGA-BC-A10R-11A.RNA.bam' 및 'TCGA-BC-A10T-01A.RNA.bam'의 5 개의 파일 각각에 대하여, 'TCGA-BC-A10Q-11A.RNA.bam', 'TCGA-BC-A10Q-01A.RNA.bam', 'TCGA-BC-A10R-11A.RNA.bam', 'TCGA-BC-A10R-01A.RNA.bam' 및 'TCGA-BC-A10T-11A.RNA.bam'의 각각의 파일이 5.313510942 점의 가장 높은 유사도 점수를 갖는 파일로 결정되었다.

[0244] 즉, 파일명의 유사도가 높은 두 개의 NGS 데이터가 한 쌍의 파일로서 매칭되어 제공될 수 있다.

[0245] 도 6d를 참조하면, KCSG의 표적 시퀀싱 분석 패널에 따른 NGS 데이터의 파일명에 기초하여 산출된 유사도 점수에 기초하여 매칭된 파일들이 도시된다.

[0246] 보다 구체적으로, '02-S001-BL.RGadded.marked.realigned.fixed.bam', '02-S001-FP.RGadded.marked.realigned.fixed.bam', '02-S002-BL.RGadded.marked.realigned.fixed.bam', '02-S002-FP.RGadded.marked.realigned.fixed.bam' 및 '02-S003-BL.RGadded.marked.realigned.fixed.bam'의 5 개의 파일 각각에 대하여, '02-S001-FP.RGadded.marked.realigned.fixed.bam', '02-S001-BL.RGadded.marked.realigned.fixed.bam', '02-S002-FP.RGadded.marked.realigned.fixed.bam', '02-S002-

BL.RGadded.marked.realigned.fixed.bam' 및 '02-S003-FP.RGadded.marked.realigned.fixed.bam'의 각각의 파일이 4.385255342 점의 가장 높은 유사도 점수를 갖는 파일로 결정되었다.

[0247] 즉, 파일명의 유사도가 높은 두 개의 NGS 데이터가 한 쌍의 파일로서 매칭되어 제공될 수 있다.

[0248] 도 6e를 참조하면, KLCC의 표적 시퀀싱 분석 패널에 따른 NGS 데이터의 파일명에 기초하여 산출된 유사도 점수에 기초하여 매칭된 파일들이 도시된다.

[0249] 보다 구체적으로, 'TCGA-BC-A10Q-01A.RNA.bam', 'TCGA-BC-A10Q-11A.RNA.bam', 'TCGA-BC-A10R-01A.RNA.bam', 'TCGA-BC-A10R-11A.RNA.bam' 및 'TCGA-BC-A10T-01A.RNA.bam'의 5 개의 파일 각각에 대하여, 'TCGA-BC-A10Q-11A.RNA.bam', 'TCGA-BC-A10Q-01A.RNA.bam', 'TCGA-BC-A10R-11A.RNA.bam', 'TCGA-BC-A10R-01A.RNA.bam' 및 'TCGA-BC-A10T-11A.RNA.bam'의 각각의 파일이 5.313510942 점의 가장 높은 유사도 점수를 갖는 파일로 결정되었다.

[0250] 즉, 파일명의 유사도가 높은 두 개의 NGS 데이터가 한 쌍의 파일로서 매칭되어 제공될 수 있다.

[0251] 도 6f를 참조하면, 전체 NGS 데이터에 대하여 10 %의 미스매칭된 파일 을 포함하는 NGS 데이터 (swapped), 전체 NGS 데이터에 대하여 10 %의 미스매칭된 파일 과 10 %의 매칭되지 않은 파일을 포함하는 NGS 데이터 (swapped + orphan) 및 전체 NGS 데이터에 대하여 10 %의 매칭되지 않은 파일을 포함하는 NGS 데이터 orphan) 에 대한, 파일명에 따른 매칭의 정확도가 도시된다.

[0252] 보다 구체적으로, 본 발명의 다양한 실시예에 따른 NGS 샘플 검증 방법에 따라, WES와 KCSG 및 KLCC의 표적 시퀀싱의 모든 분석 패널에 따른 NGS 데이터가, 100 %의 정확도로 매칭의 여부 또는 미스매칭 (swapped, swapped + orphan, 또는 orphan) 에 따라 분류된 것으로 나타난다.

[0253] 이러한 결과는, 파일명에 기초하여 2 개의 NGS 데이터를 매칭한 결과가 높은 신뢰도를 가진다는 것을 의미할 수 있다.

[0254] 이상의 실시예 2의 결과에 따르면, 본 발명의 다양한 실시예에서 파일명에 따라 NGS 데이터를 매칭할 경우, 매칭 결과에 대한 신뢰도가 높은 것으로 나타난다. 이러한 결과는, 본 발명의 다양한 실시예에 따른 NGS 샘플 검증 방법이 다양한 염기서열 분석 패널에 적용될 수 있음을 의미할 수 있다.

[0255] **실시예 3: 유전자형 일치 점수 및 파일명에 따른 NGS 샘플 검증**

[0256] 이하에서는 도 7을 참조하여, 본 발명의 다양한 실시예에 적용되는 파일명에 기초한 NGS 샘플 검증에 대한 평가 결과를 구체적으로 설명한다.

[0257] 도 7은 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법에 의해, 유전자형 및 파일명에 따라 매칭된 NGS 데이터 쌍 또는 미스매칭된 NGS 데이터 쌍의 평가 결과를 도시한 것이다.

[0258] 도 7의 (a), (b), (c) 및 (d)를 참조하면, 본 발명의 다양한 실시예에 따른 NGS 샘플 검증 방법에 따른 검증 결과가 도시된다.

[0259] 보다 구체적으로, 도 7의 (a)를 참조하면, 유전자형 또는 파일명 중 어느 하나의 조건에 따라 매칭된 결과가 도시된다. 이때, 'S1254_N.bam'의 파일 및 'S1345_T.bam'파일의 두 개의 NGS 데이터와 'S1345_N.bam'의 파일 및 'S1254_T.bam'의 두 개의 NGS 데이터 유전자형 일치 점수 (일치율) 이 각각 0.97, 0.95임에도 파일명의 유사도가 낮은 파일로 결정된 것으로 나타난다. 즉, 이러한 두 개의 NGS 데이터는, 미스매칭된 (swapped) 파일일 수 있다.

[0260] 도 7의 (b)를 참조하면, 낮은 유전자형의 일치 점수를 갖는, 어느 파일과도 매칭되지 않은 (orphan) 파일들이 도시된다.

[0261] 도 7의 (c)를 참조하면, 높은 유전자형 일치 점수를 갖고 및 파일명의 유사도가 높음에 따라 최종적으로 매칭된 파일이 도시된다.

[0262] 도 7의 (d)를 참조하면, 매칭된 파일, 미스매칭된 파일, 및 매칭되지 않은 파일을 모두 포함하는, NGS 데이터에 대한 검증 결과가 통합적으로 도시된다.

[0263] 이상의 실시예 3에 따르면, 본 발명은, 유전자형 및 파일명에 따라 매칭된 다양한 NGS 데이터의 검증 결과를 제공할 수 있다. 이에, 사용자는, 복수의 대상 샘플에 대한 NGS 분석 결과와 함께, 샘플 또는 개체의 종류에 따

른 매칭 여부의 검증 결과를 제공 받을 수 있다.

[0264] 이상의 실시예 1 내지 3의 결과에 따르면, 본 발명의 다양한 실시예에 따른 NGS 검증 방법 및 이를 이용한 디바이스는, 사용자의 핸들링에 따라 발생하는 휴먼 오류, 예를 들어 샘플 및 분석 데이터의 미스매칭과 같은 오류를 개선하여, 샘플 및 NGS 데이터를 매칭하여 제공할 수 있다.

[0265] 보다 구체적으로, NGS 검증 방법 및 이를 이용한 디바이스는, 유전자형 및/또는 샘플의 NGS 데이터의 파일명을 기초로 미스 매칭되거나 매칭되지 않은 NGS 분석 결과를 사용자에게 제공함에 따라, NGS 분석 결과 제공에 있어서 빈번하게 발생할 수 있는 문제점들을 해결할 수 있다.

[0266] 특히, NGS 검증 방법 및 이를 이용한 디바이스는, 샘플간 혼합에 따라 발생하는 검출의 신뢰도 떨어뜨리거나 부정확한 결과를 제공하는 문제점들을 해결할 수 있다.

[0267] 나아가, 본 발명의 NGS 검증 방법 및 이를 이용한 디바이스는, 분석 패널의 종류에 따라 SNP 사이트를 결정하고, 소형 패널에 대하여 높은 검증의 정확도를 제공할 수 있다.

[0268] 특히, NGS 검증 방법 및 이를 이용한 디바이스는 다양한 염기서열 분석 패널, 예를 들어 WES, RNA-seq 와 같은 대량의 분석 패널, 또는 표적 염기서열 분석과 같은 소형 분석 패널에 적용 가능할 수 있다.

[0269] **비교예: 종래의 SNP 사이트에 기초한 NGS 샘플 검증 방법 및 본 발명의 NGS 샘플 검증용 디바이스의 평가**

[0270] 이하에서는, 도 8a 내지 8d를 참조하여, 종래의 SNP 사이트에 기초한 NGS 샘플 검증 방법 및 본 발명의 NGS 샘플 검증용 디바이스의 평가 결과를 설명한다.

[0271] 도 8a 내지 8d는 본 발명의 일 실시예에 따른 NGS 샘플 검증 방법 및 종래의 샘플 검증 방법의 평가 결과를 비교하여 도시한 것이다.

[0272] 이때, 종래의 SNP 사이트에 기초한 NGS 샘플 검증 방법으로서, NGSCheckMate-BAM, NGSCheckMate-FASTQ, BAM-matcher 및 Conpair이 선택되었다.

[0273] 도 8a를 참조하면, 평가에 이용된 데이터 세트들이 도시된다. 보다 구체적으로, 본 평가에서, WES 분석 패널에 대한 202 개 (101 쌍) 의 NGS 데이터, RNA-seq 분석 패널에 대한 130 개 (65 쌍) 의 NGS 데이터, WES/RNA-seq 분석 패널에 대한 168 개 (84 쌍) 의 NGS 데이터, 및 표적 시퀀싱 분석 패널의 KCSG의 192 개 (96 쌍) 의 NGS 데이터와 KLCC의 402 개 (201 쌍) 의 NGS 데이터가 이용되었다.

[0274] 도 8b를 참조하면, 본 발명의 NGS 샘플 검증용 디바이스는, WES, RNA-seq, WES/RNA-seq 및 표적 시퀀싱 분석 (KCSG 및 KLCC) 의 모든 분석 패널에 따른 NGS 데이터에 대하여, 100 %의 정확도, 민감도 및 특이도로 파일 쌍 (매칭 또는 매칭되지 않은 파일 쌍)을 결정하는 것으로 나타난다.

[0275] 그러나, NGSCheckMate-BAM 및 NGSCheckMate-FASTQ의 경우, 표적 시퀀싱 분석의 NGS 데이터에 대하여 정확도, 민감도 및 특이도가 본 발명의 디바이스에 비하여 떨어지는 것으로 나타난다. 나아가, BAM-matcher의 경우, WES/RNA-seq 및 표적 시퀀싱 분석DM NGS 데이터에 대하여 정확도, 민감도 및 특이도가 본 발명의 디바이스에 비하여 떨어지는 것으로 나타난다. 특히, Conpair는 RNA-seq, WES/RNA-seq의 분석 패널에을 제외한 모든 NGS 데이터에서 정확도, 민감도 및 특이도가 낮은 것으로 나타난다.

[0276] 도 8c를 더욱 참조하면, 본 발명의 NGS 샘플 검증용 디바이스는, WES, RNA-seq, WES/RNA-seq 및 표적 시퀀싱 분석 (KCSG 및 KLCC) 의 모든 분석 패널에 따른 NGS 데이터에 대하여, 100 %의 정확도를 유지하는 것으로 나타난다. 그러나, 종래의 SNP 사이트에 기초한 NGS 샘플 검증 방법, 특히 Conpair의 경우, 표적 시퀀싱 분석에 따른 NGS 데이터에 대하여 분류의 정확도가 떨어지는 것으로 나타난다.

[0277] 나아가, 본 발명의 NGS 샘플 검증용 디바이스는, 409 개의 유전자로 구성된 Ion-CCP (Ion AmpliSeq Comprehensive Cancer Panel), 315 개의 유전자로 구성된 FONE (Foundation One), 127 개의 유전자로 구성된 xGen-PCP (xGen Pan-Cancer Panel) 및 46 개의 유전자로 구성된 CCCP (Comprehensive Common Cancer Panel) 의 다양한 암 패널을 적용했을 때, 패널의 유전자의 수가 적어짐에도 100 %의 정확도를 유지하는 것으로 나타난다.

[0278] 그러나, NGSCheckMate-BAM, NGSCheckMate-FASTQ, BAM-matcher 및 Conpair의 종래의 NGS 검증 방법은, 패널의 유전자수가 적어질수록, 정확도가 낮아지는 것으로 나타난다.

[0279] 이상의 결과에 따르면, 본 발명의 NGS 샘플 검증용 디바이스는, 소형 분석 패널에서도 높은 정확도로 NGS 데이터를 매칭하고, 신뢰도 높은 동일 개체 유래 샘플 여부를 제공하는 것으로 나타난다. 특히, 본 발명의 NGS 샘

플 검증용 디바이스는 파일명을 더욱 고려함에 따라, SNP 사이트만을 고려한 종래의 NGS 검증 방법들에 비하여 높은 매칭 성능을 가질 수 있다.

[0280] 도 8d를 더욱 참조하면, 본 발명의 NGS 샘플 검증용 디바이스는, WES 또는 표적 시퀀싱의 분석 패널에 대하여, 종래의 NGS 검증 방법에 비하여 현저하게 낮은 러닝 타임을 갖는 것으로 나타난다.

[0281] 보다 구체적으로, 본 발명의 NGS 샘플 검증용 디바이스는, WES의 NGS 데이터에 대하여 단 5.3 분 (또는 9.9 분)의 러닝 타임이 소요되는 것으로 나타난다. 이와 대조적으로, NGSCheckMate-BAM, NGSCheckMate-FASTQ, BAM-matcher 및 Conpair의 러닝 타임은 본 발명의 NGS 샘플 검증용 디바이스에 비하여 현저하게 높은 것으로 나타난다. 특히, NGSCheckMate-FASTQ (p1) 의 러닝 타임은, 본 발명의 NGS 샘플 검증용 디바이스 보다 약 145 배 긴 러닝 타임이 소요되는 것으로 나타난다.

[0282] 또한, 본 발명의 NGS 샘플 검증용 디바이스는, 표적 시퀀싱 분석된 NGS 데이터에 대하여 단 7.6 분 (또는 16.7 분) 의 러닝 타임이 소요되는 것으로 나타난다. 이와 대조적으로, NGSCheckMate-BAM, NGSCheckMate-FASTQ, BAM-matcher의 러닝 타임은 본 발명의 NGS 샘플 검증용 디바이스에 비하여 높은 것으로 나타난다. 특히, NGSCheckMate-FASTQ (p1) 의 러닝 타임은, 본 발명의 NGS 샘플 검증용 디바이스 보다 약 33 배 긴 러닝 타임이 소요되는 것으로 나타난다.

[0283] 이상의 결과에 따르면, 본 발명의 NGS 샘플 검증용 디바이스는, 분석 패널의 종류에 따라 SNP 사이트를 결정하여 러닝 타임을 줄이고, 이에 짧은 러닝 타임이 소요되는 것으로 나타난다. 특히, 본 발명의 NGS 샘플 검증용 디바이스는, GMAF의 수준, 개별 인구 내의 MAF, 및 SNP 사이트의 맵핑 능력에 기초하여 SNP 사이트를 결정하도록 구성됨에 따라, 분석 패널에 관계 없이 고정된 SNP 사이트만을 고려한 종래의 NGS 검증 방법들에 비하여 빠른 러닝 타임을 제공할 수 있는 효과가 있다.

[0284] 본 발명의 여러 실시예들의 각각 특징들이 부분적으로 또는 전체적으로 서로 결합 또는 조합 가능하며, 당업자가 충분히 이해할 수 있듯이 기술적으로 다양한 연동 및 구동이 가능하며, 각 실시예들이 서로에 대하여 독립적으로 실 시 가능할 수도 있고 연관 관계로 함께 실시 가능할 수도 있다.

[0285] 이상 첨부된 도면을 참조하여 본 발명의 실시예들을 더욱 상세하게 설명하였으나, 본 발명은 반드시 이러한 실시예로 국한되는 것은 아니고, 본 발명 의 기술사상을 벗어나지 않는 범위 내에서 다양하게 변형 실시될 수 있다. 따라서, 본 발명에 개시된 실시예들은 본 발명의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 발명의 기술 사상의 범 위가 한정되는 것은 아니다. 그러므로, 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 본 발명의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 권리 범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0286] 100: NGS 샘플 검증용 디바이스

110: 수신부

120: 입력부

130: 출력부

140: 저장부

150: 프로세서

212, 312, 412: WES 파일 또는 RNA 시퀀싱 파일

214, 314, 414: 표적 시퀀싱 파일

216, 316, 416: BED 파일

222, 422: 대형 분석 패널에 따른 표적 SNP 사이트

224, 424: 소형 분석 패널에 따른 표적 SNP 사이트

232, 322, 432, 434: 매칭된 한 쌍의 파일

234, 326, 436: 매칭되지 않은 파일

324: 미스매칭된 파일

332, 452: 검증 결과

342 (a), 342 (b), 344 (a): NGS 데이터

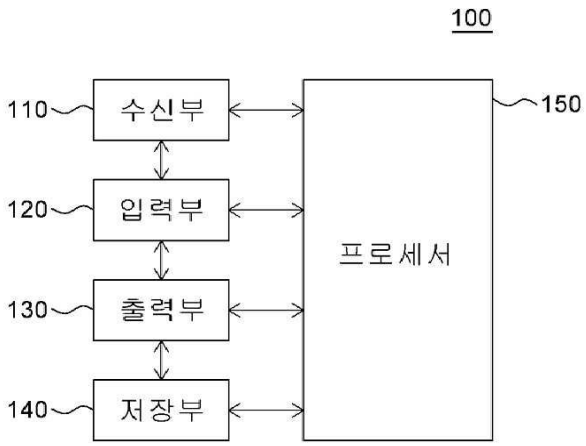
442: 최종 매칭된 한 쌍의 파일

444: 최종 미스매칭된 한 쌍의 파일

446: 최종 매칭되지 않은 파일

도면

도면1a



도면1b

(a)

```
03-S006-BL.RGadded.marked.realigned.fixed.gvcf  03-S006-FP.RGadded.marked.realigned.fixed.gvcf  0.4384  unmatched
03-S006-BL.RGadded.marked.realigned.fixed.gvcf  03-S007-BL.RGadded.marked.realigned.fixed.gvcf  0.4242  unmatched
03-S006-BL.RGadded.marked.realigned.fixed.gvcf  03-S007-FP.RGadded.marked.realigned.fixed.gvcf  0.4264  unmatched
03-S006-BL.RGadded.marked.realigned.fixed.gvcf  03-S008-BL.RGadded.marked.realigned.fixed.gvcf  0.4359  unmatched
03-S006-BL.RGadded.marked.realigned.fixed.gvcf  03-S008-FP.RGadded.marked.realigned.fixed.gvcf  0.4472  unmatched
03-S006-BL.RGadded.marked.realigned.fixed.gvcf  07-S001-BL.RGadded.marked.realigned.fixed.gvcf  0.4611  unmatched
03-S006-BL.RGadded.marked.realigned.fixed.gvcf  07-S001-FP.RGadded.marked.realigned.fixed.gvcf  0.472   unmatched
03-S006-FP.RGadded.marked.realigned.fixed.gvcf  03-S007-BL.RGadded.marked.realigned.fixed.gvcf  0.455   unmatched
03-S006-FP.RGadded.marked.realigned.fixed.gvcf  03-S007-FP.RGadded.marked.realigned.fixed.gvcf  0.4615  unmatched
03-S006-FP.RGadded.marked.realigned.fixed.gvcf  03-S008-BL.RGadded.marked.realigned.fixed.gvcf  1.0     matched
03-S006-FP.RGadded.marked.realigned.fixed.gvcf  03-S008-FP.RGadded.marked.realigned.fixed.gvcf  0.9905  matched
```

(b)

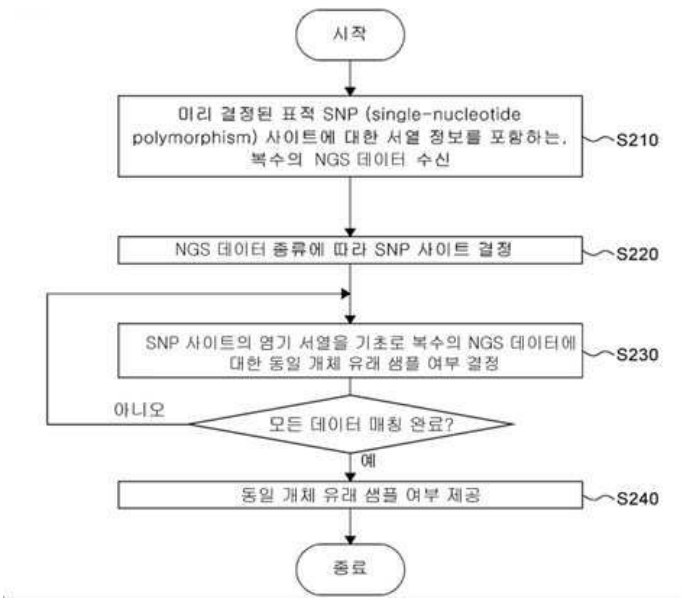
```
#Matched pair by genotype and name.
02-S001-FP.RGadded.marked.realigned.fixed.gvcf  02-S001-BL.RGadded.marked.realigned.fixed.gvcf  1.0
02-S002-FP.RGadded.marked.realigned.fixed.gvcf  02-S002-BL.RGadded.marked.realigned.fixed.gvcf  1.0
02-S003-FP.RGadded.marked.realigned.fixed.gvcf  02-S003-BL.RGadded.marked.realigned.fixed.gvcf  1.0
02-S004-FP.RGadded.marked.realigned.fixed.gvcf  02-S004-BL.RGadded.marked.realigned.fixed.gvcf  1.0
03-S001-FP.RGadded.marked.realigned.fixed.gvcf  03-S001-BL.RGadded.marked.realigned.fixed.gvcf  1.0
03-S004-FP.RGadded.marked.realigned.fixed.gvcf  03-S004-BL.RGadded.marked.realigned.fixed.gvcf  1.0
03-S005-FP.RGadded.marked.realigned.fixed.gvcf  03-S005-BL.RGadded.marked.realigned.fixed.gvcf  0.9873
```

(c)

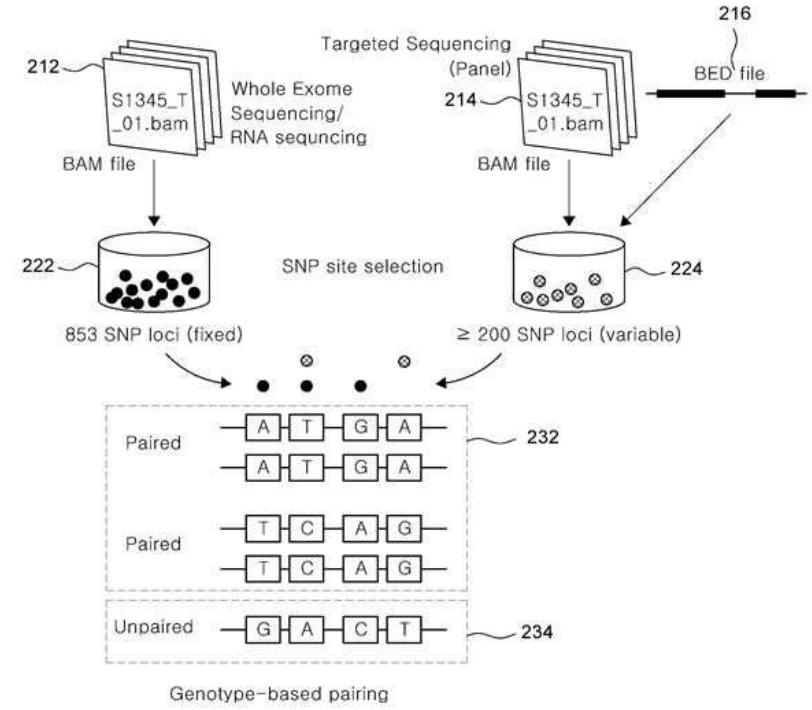
```
#List of pairs are not matched by name but by genotype.
03-S006-FP.RGadded.marked.realigned.fixed.gvcf  03-S008-BL.RGadded.marked.realigned.fixed.gvcf  1.0
03-S006-FP.RGadded.marked.realigned.fixed.gvcf  03-S008-FP.RGadded.marked.realigned.fixed.gvcf  0.9905

#List of samples are matched with nothing by genotype.
03-S006-BL.RGadded.marked.realigned.fixed.gvcf
-> pair by name with 03-S006-FP.RGadded.marked.realigned.fixed.gvcf ( score : 0.4384)
```

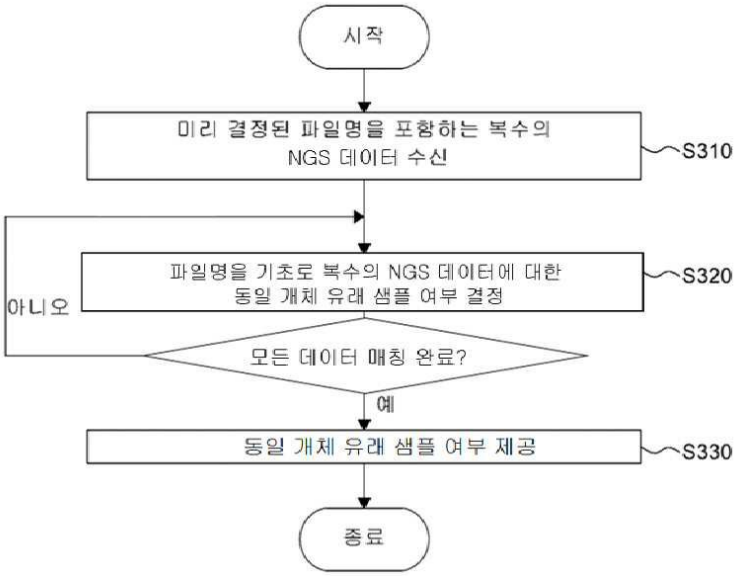
도면2a



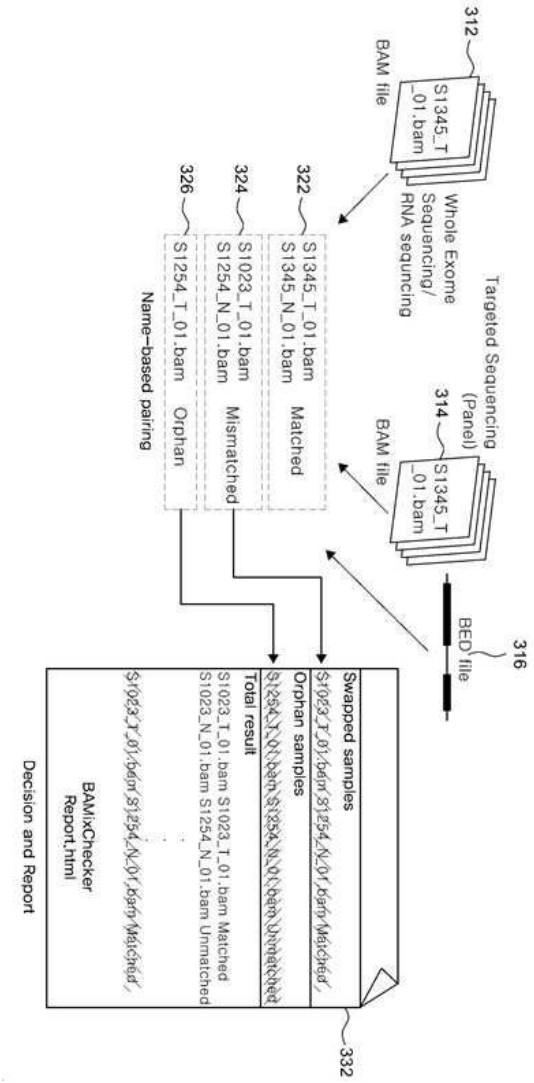
도면2b


212
S1345_T
_01.bam
Whole Exome
Sequencing/
RNA sequncing
BAM file
Targeted Sequencing
(Panel)
214
S1345_T
_01.bam
BAM file
216
BED file
222
SNP site selection
224
853 SNP loci (fixed)
≥ 200 SNP loci (variable)
Paired
A T G A
A T G A
232
Paired
T C A G
T C A G
Unpaired
G A C T
234
Genotype-based pairing

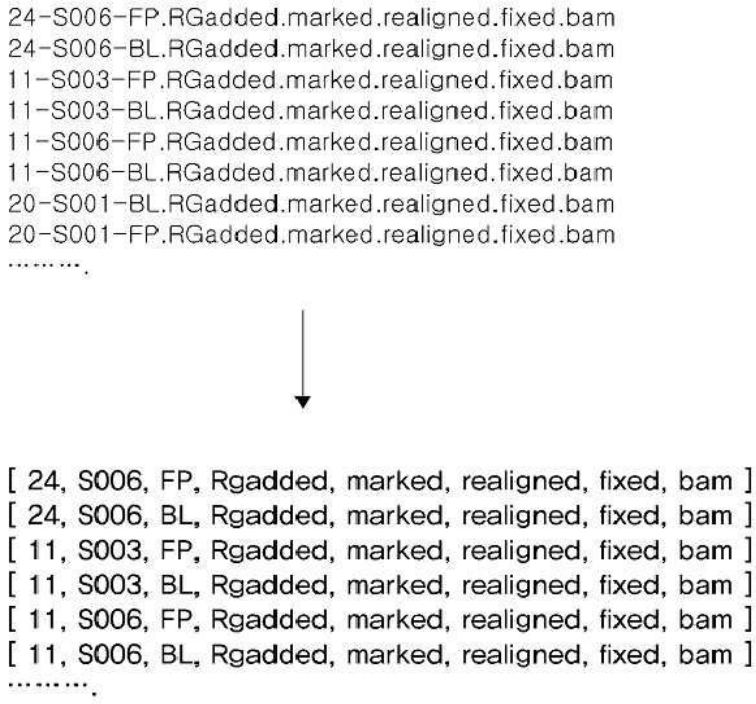
도면3a


시작
미리 결정된 파일명을 포함하는 복수의
NGS 데이터 수신
S310
파일명을 기초로 복수의 NGS 데이터에 대한
동일 개체 유래 샘플 여부 결정
S320
아니오
모든 데이터 매칭 완료?
예
동일 개체 유래 샘플 여부 제공
S330
종료

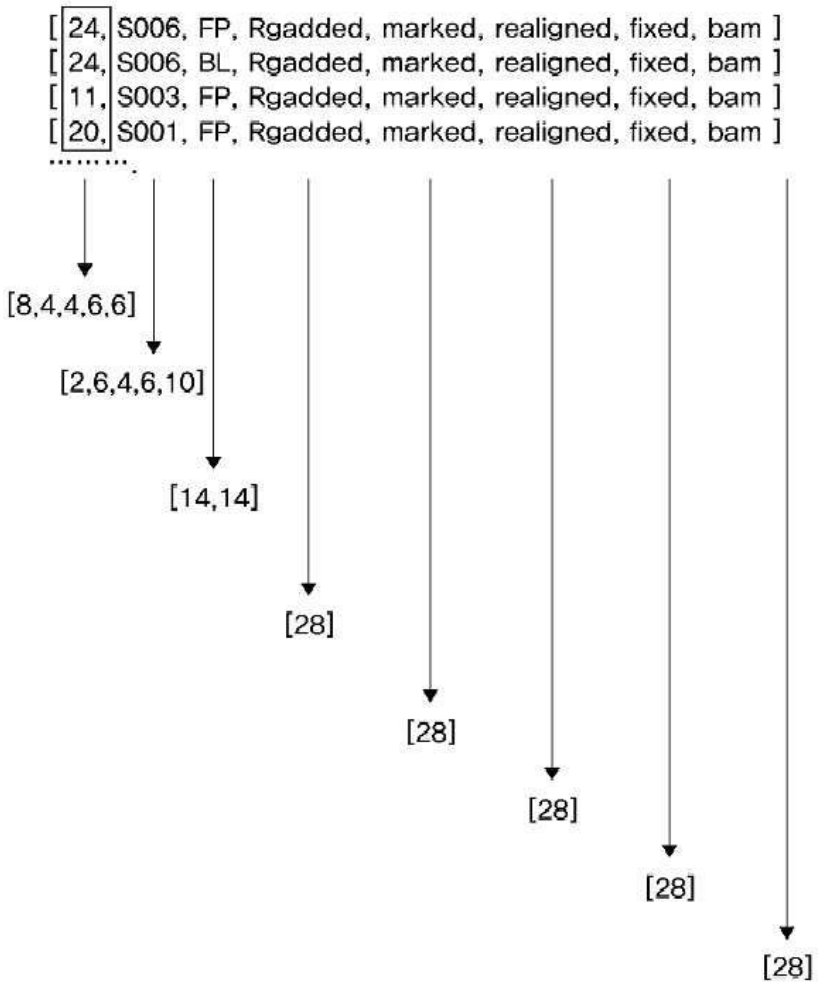
도면3b


312
S1345_T_01.bam
BAM file
Whole Exome Sequencing/ RNA sequncing
Targeted Sequencing (Panel)
314
S1345_T_01.bam
BAM file
316
BED file
322
S1345_T_01.bam
S1345_N_01.bam
Matched
324
S1023_T_01.bam
S1254_N_01.bam
Mismatched
326
S1254_T_01.bam
Orphan
Name-based pairing
Swapped samples
Orphan samples
Total result
S1023_T_01.bam S1023_T_01.bam Matched
S1023_N_01.bam S1254_N_01.bam Unmatched
BAMixChecker Report.html
Decision and Report
332

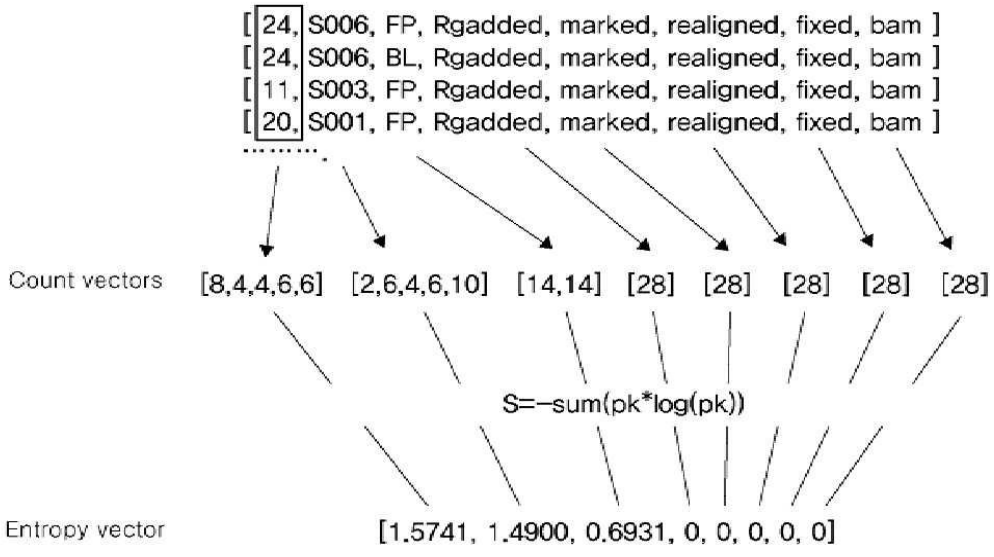
도면3c


24-S006-FP.RGadded.marked.realigned.fixed.bam
24-S006-BL.RGadded.marked.realigned.fixed.bam
11-S003-FP.RGadded.marked.realigned.fixed.bam
11-S003-BL.RGadded.marked.realigned.fixed.bam
11-S006-FP.RGadded.marked.realigned.fixed.bam
11-S006-BL.RGadded.marked.realigned.fixed.bam
20-S001-BL.RGadded.marked.realigned.fixed.bam
20-S001-FP.RGadded.marked.realigned.fixed.bam
……
[24, S006, FP, Rgadded, marked, realigned, fixed, bam]
[24, S006, BL, Rgadded, marked, realigned, fixed, bam]
[11, S003, FP, Rgadded, marked, realigned, fixed, bam]
[11, S003, BL, Rgadded, marked, realigned, fixed, bam]
[11, S006, FP, Rgadded, marked, realigned, fixed, bam]
[11, S006, BL, Rgadded, marked, realigned, fixed, bam]
……

도면3d


[24, S006, FP, Rgadded, marked, realigned, fixed, bam]
[24, S006, BL, Rgadded, marked, realigned, fixed, bam]
[11, S003, FP, Rgadded, marked, realigned, fixed, bam]
[20, S001, FP, Rgadded, marked, realigned, fixed, bam]
[8,4,4,6,6]
[2,6,4,6,10]
[14,14]
[28]
[28]
[28]
[28]
[28]

도면3e


[24, S006, FP, Rgadded, marked, realigned, fixed, bam]
[24, S006, BL, Rgadded, marked, realigned, fixed, bam]
[11, S003, FP, Rgadded, marked, realigned, fixed, bam]
[20, S001, FP, Rgadded, marked, realigned, fixed, bam]
Count vectors
[8,4,4,6,6] [2,6,4,6,10] [14,14] [28] [28] [28] [28] [28]
S=−sum(pk*log(pk))
Entropy vector
[1.5741, 1.4900, 0.6931, 0, 0, 0, 0, 0]

도면3f

[24, S006, FP, Rgadded, marked, realigned, fixed, bam]
[24, S006, BL, Rgadded, marked, realigned, fixed, bam]
.....

Entropy vector

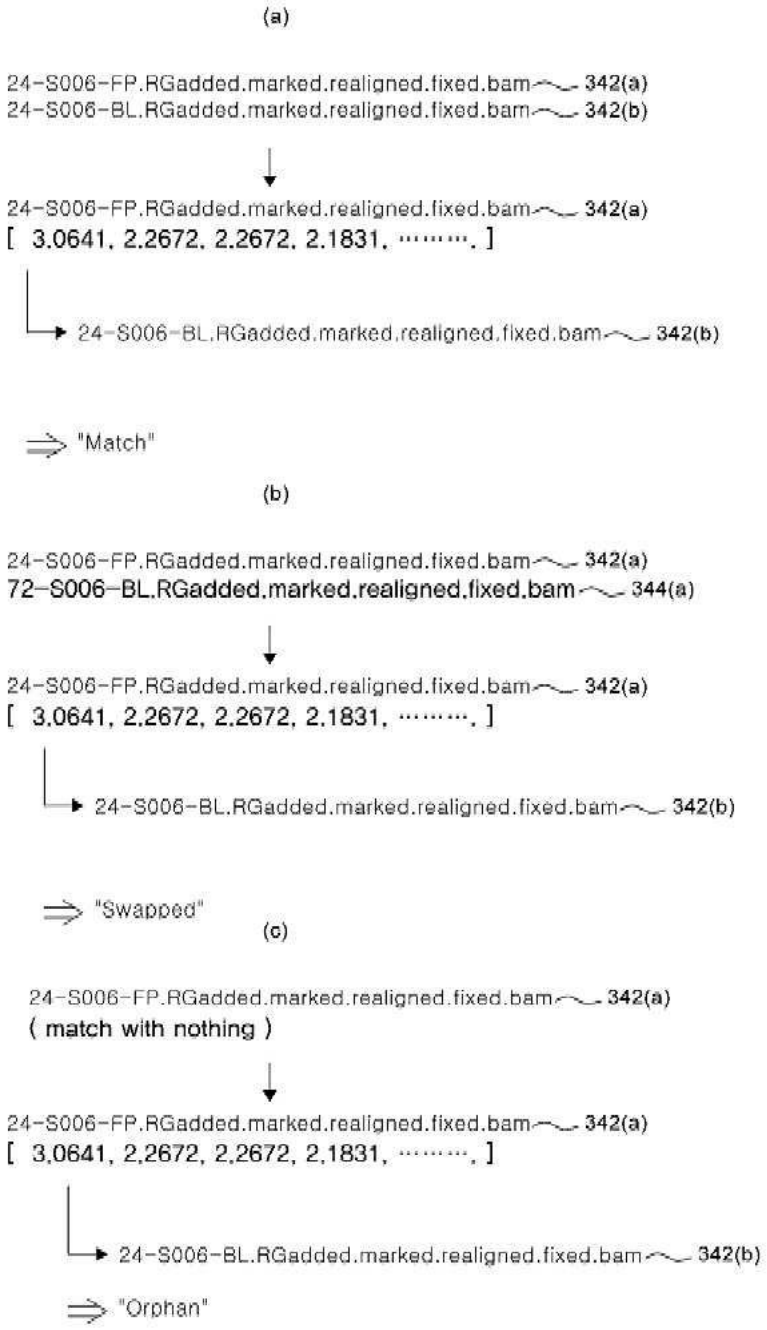
[1.5741, 1.4900, 0.6931, 0, 0, 0, 0, 0]

Score = 1.5741 + 1.4900 − 0.6931 + 0 + 0 + 0 + 0 + 0
= 2.371

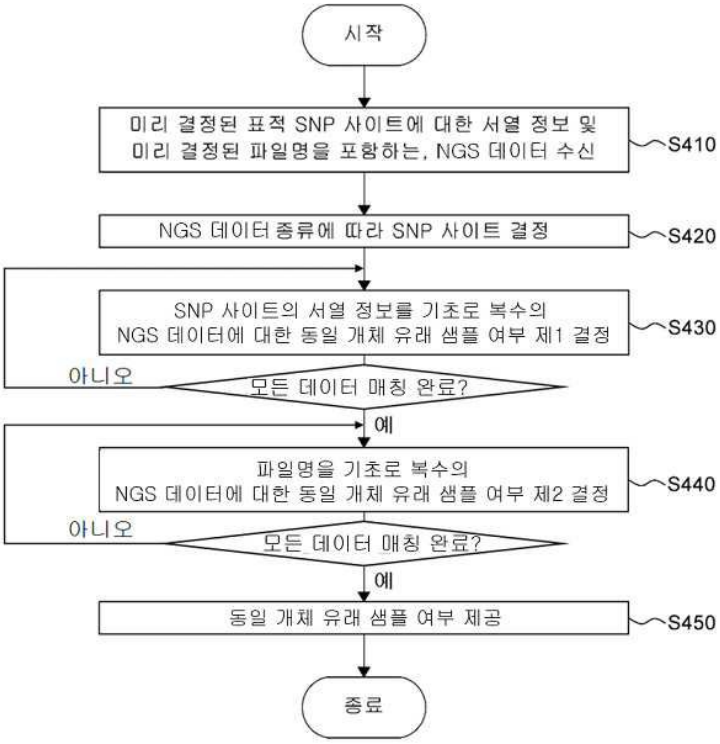
.....

24-S006-FP.RGadded.marked.realigned.fixed.bam [2.371, 0.7772, 0.7772, −0.609, ...]
24-S006-BL.RGadded.marked.realigned.fixed.bam [2.371, −0.609, 0.7772, −0.609, ...]

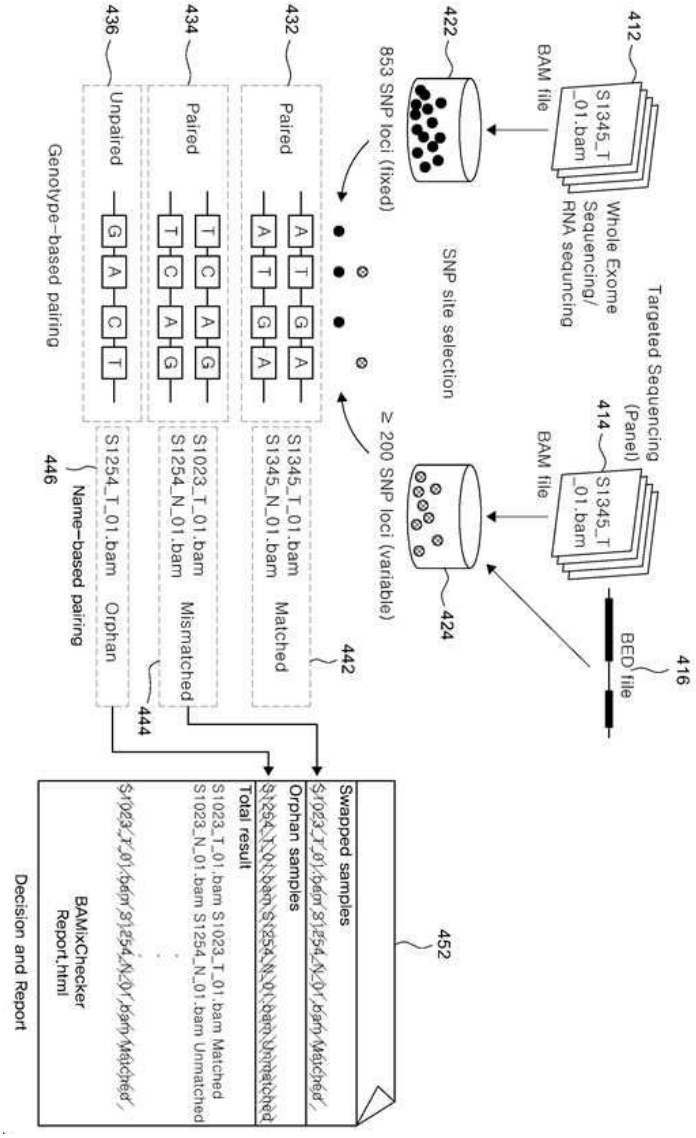
도면3g


(a)
24-S006-FP.RGadded.marked.realigned.fixed.bam 342(a)
24-S006-BL.RGadded.marked.realigned.fixed.bam 342(b)
24-S006-FP.RGadded.marked.realigned.fixed.bam 342(a)
[3.0641, 2.2672, 2.2672, 2.1831, …….]
24-S006-BL.RGadded.marked.realigned.fixed.bam 342(b)
"Match"
(b)
24-S006-FP.RGadded.marked.realigned.fixed.bam 342(a)
72-S006-BL.RGadded.marked.realigned.fixed.bam 344(a)
24-S006-FP.RGadded.marked.realigned.fixed.bam 342(a)
[3.0641, 2.2672, 2.2672, 2.1831, …….]
24-S006-BL.RGadded.marked.realigned.fixed.bam 342(b)
"Swapped"
(c)
24-S006-FP.RGadded.marked.realigned.fixed.bam 342(a)
(match with nothing)
24-S006-FP.RGadded.marked.realigned.fixed.bam 342(a)
[3.0641, 2.2672, 2.2672, 2.1831, …….]
24-S006-BL.RGadded.marked.realigned.fixed.bam 342(b)
"Orphan"

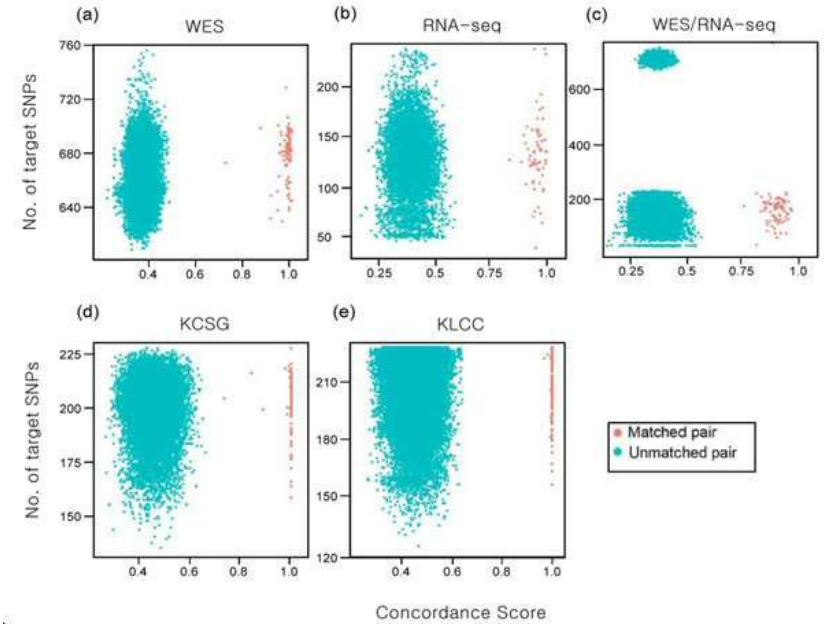
도면4a


시작
미리 결정된 표적 SNP 사이트에 대한 서열 정보 및 미리 결정된 파일명을 포함하는, NGS 데이터 수신
S410
NGS 데이터 종류에 따라 SNP 사이트 결정
S420
SNP 사이트의 서열 정보를 기초로 복수의 NGS 데이터에 대한 동일 개체 유래 샘플 여부 제1 결정
S430
모든 데이터 매칭 완료?
아니오
예
파일명을 기초로 복수의 NGS 데이터에 대한 동일 개체 유래 샘플 여부 제2 결정
S440
모든 데이터 매칭 완료?
아니오
예
동일 개체 유래 샘플 여부 제공
S450
종료

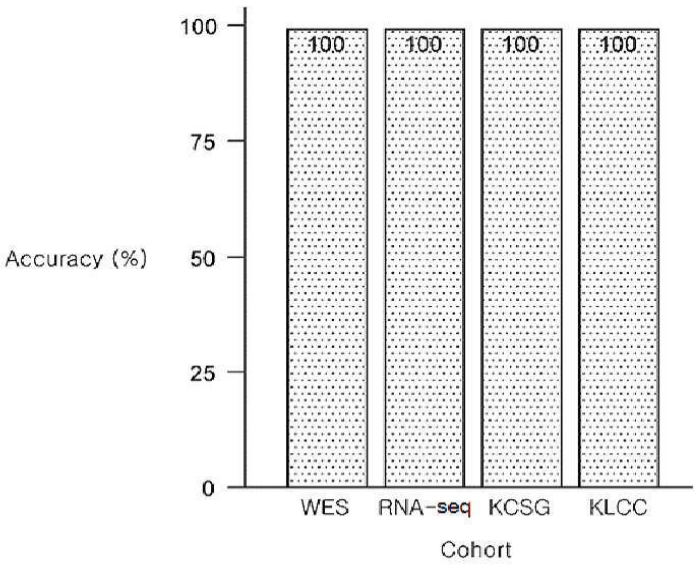
도면4b


412
S1345_T_01.bam
Whole Exome Sequencing/ RNA sequncing
BAM file
422
853 SNP loci (fixed)
SNP site selection
≥ 200 SNP loci (variable)
424
Targeted Sequencing (Panel)
414
S1345_T_01.bam
BAM file
416
BED file
Genotype-based pairing
432
Paired
434
Paired
436
Unpaired
442
S1345_T_01.bam
S1345_N_01.bam
Matched
S1023_T_01.bam
S1254_N_01.bam
Mismatched
444
S1254_T_01.bam
Orphan
446
Name-based pairing
452
Swapped samples
Orphan samples
Total result
S1023_T_01.bam S1023_T_01.bam Matched
S1023_N_01.bam S1254_N_01.bam Unmatched
BAMixChecker Report.html
Decision and Report

도면5



도면6a



도면6b

File name (WES)	Best scored file
C347.TCGA-05-4244-01A-01D-1105-08.5_gdc_realn.bam	C347.TCGA-05-4244-10A-01D-1105-08.5_gdc_realn.bam 9.229892145
C347.TCGA-05-4244-10A-01D-1105-08.5_gdc_realn.bam	C347.TCGA-05-4244-01A-01D-1105-08.5_gdc_realn.bam 9.229892145
C347.TCGA-05-4249-01A-01D-1105-08.5_gdc_realn.bam	C347.TCGA-05-4249-10A-01D-1105-08.5_gdc_realn.bam 9.229892145
C347.TCGA-05-4249-10A-01D-1105-08.5_gdc_realn.bam	C347.TCGA-05-4249-01A-01D-1105-08.5_gdc_realn.bam 9.229892145
C347.TCGA-05-4250-01A-01D-1105-08.5_gdc_realn.bam	C347.TCGA-05-4250-10A-01D-1105-08.5_gdc_realn.bam 9.229892145

도면6c

File name (RNA)	Best scored file
TCGA-BC-A10Q-01A.RNA.bam	TCGA-BC-A10Q-11A.RNA.bam 5.313510942
TCGA-BC-A10Q-11A.RNA.bam	TCGA-BC-A10Q-01A.RNA.bam 5.313510942
TCGA-BC-A10R-01A.RNA.bam	TCGA-BC-A10R-11A.RNA.bam 5.313510942
TCGA-BC-A10R-11A.RNA.bam	TCGA-BC-A10R-01A.RNA.bam 5.313510942
TCGA-BC-A10T-01A.RNA.bam	TCGA-BC-A10T-11A.RNA.bam 5.313510942

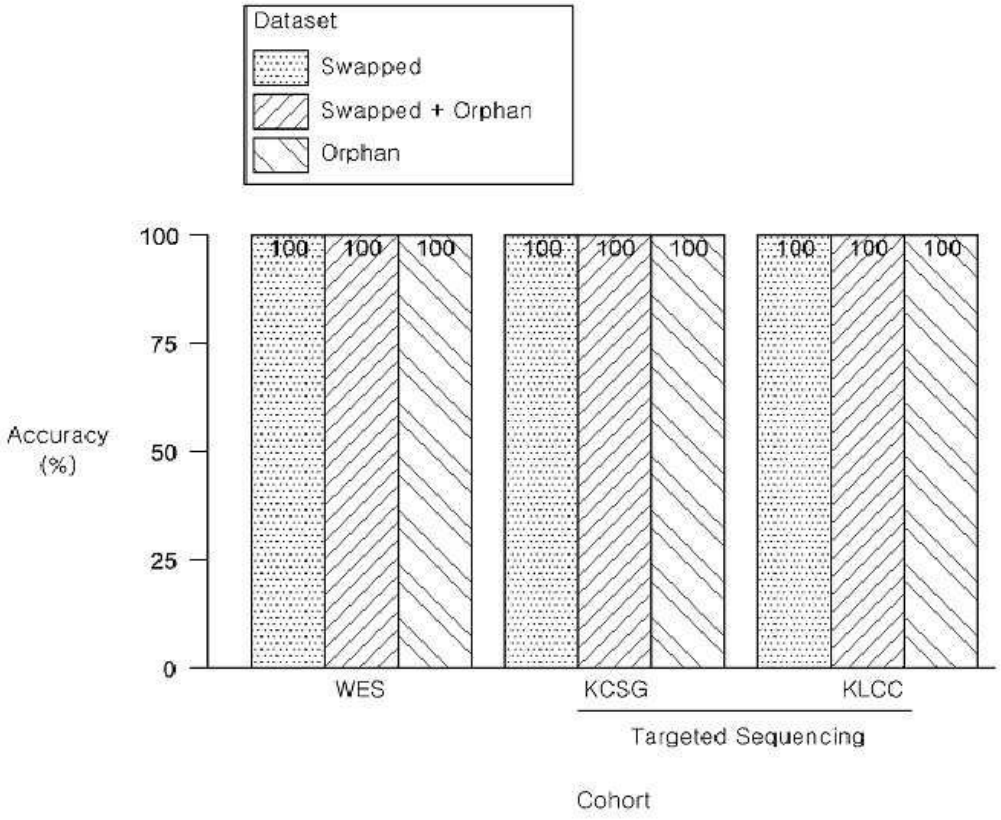
도면6d

File name (Targeted seq_KCSG)	Best scored file
02-S001-BL.RGadded.marked.realigned.fixed.bam	02-S001-FP.RGadded.marked.realigned.fixed.bam 4.385255342
02-S001-FP.RGadded.marked.realigned.fixed.bam	02-S001-BL.RGadded.marked.realigned.fixed.bam 4.385255342
02-S002-BL.RGadded.marked.realigned.fixed.bam	02-S002-FP.RGadded.marked.realigned.fixed.bam 4.385255342
02-S002-FP.RGadded.marked.realigned.fixed.bam	02-S002-BL.RGadded.marked.realigned.fixed.bam 4.385255342
02-S003-BL.RGadded.marked.realigned.fixed.bam	02-S003-FP.RGadded.marked.realigned.fixed.bam 4.385255342

도면6e

File name (Targeted seq_KLCC)	Best scored file
TCGA-BC-A10Q-01A.RNA.bam	TCGA-BC-A10Q-11A.RNA.bam 5.313510942
TCGA-BC-A10Q-11A.RNA.bam	TCGA-BC-A10Q-01A.RNA.bam 5.313510942
TCGA-BC-A10R-01A.RNA.bam	TCGA-BC-A10R-11A.RNA.bam 5.313510942
TCGA-BC-A10R-11A.RNA.bam	TCGA-BC-A10R-01A.RNA.bam 5.313510942
TCGA-BC-A10T-01A.RNA.bam	TCGA-BC-A10T-11A.RNA.bam 5.313510942

도면6f


Dataset
Swapped
Swapped + Orphan
Orphan
Accuracy (%)
100
75
50
25
0
100
100
100
100
100
100
100
100
100
WES
KCSG
KLCC
Targeted Sequencing
Cohort

도면7

(a)

File1	File2	Concordance rate	Conclusion
S1254_N.bam	S1254_T.bam	0.38	Unmatched
S1254_N.bam	S1345_T.bam	0.97	Matched
S1345_N.bam	S1254_T.bam	0.95	Matched
S1345_N.bam	S1345_T.bam	0.36	Unmatched

(b)

Orphan sample	Best match by file name	Concordance rate	Conclusion
S1983_N.bam	S1983_T.bam	0.37	Unmatched
S1983_T.bam	S1983_N.bam	0.37	Unmatched

(c)

File1	File2	Concordance rate	Conclusion
S1023_N.bam	S1023_T.bam	0.92	Matched

(d)

File1	File2	Concordance rate	Conclusion
S1023_N.bam	S1023_T.bam	0.92	Matched
S1023_N.bam	S1254_N.bam	0.37	Unmatched
S1023_N.bam	S1254_T.bam	0.34	Unmatched
S1023_N.bam	S1345_N.bam	0.34	Unmatched
S1023_N.bam	S1345_T.bam	0.36	Unmatched
S1023_N.bam	S1983_N.bam	0.36	Unmatched
S1023_N.bam	S1983_T.bam	0.35	Unmatched
S1023_T.bam	S1254_N.bam	0.35	Unmatched
S1023_T.bam	S1254_T.bam	0.32	Unmatched
S1023_T.bam	S1345_N.bam	0.33	Unmatched
S1023_T.bam	S1345_T.bam	0.35	Unmatched
S1023_T.bam	S1983_N.bam	0.34	Unmatched
S1023_T.bam	S1983_T.bam	0.35	Unmatched
S1254_N.bam	S1254_T.bam	0.38	Unmatched
S1254_N.bam	S1345_N.bam	0.37	Unmatched
S1254_N.bam	S1345_T.bam	0.97	Matched
S1254_N.bam	S1983_N.bam	0.36	Unmatched
S1254_N.bam	S1983_T.bam	0.33	Unmatched
S1254_T.bam	S1345_N.bam	0.95	Matched
S1254_T.bam	S1345_T.bam	0.37	Unmatched
S1254_T.bam	S1983_N.bam	0.39	Unmatched
S1254_T.bam	S1983_T.bam	0.35	Unmatched
S1345_N.bam	S1345_T.bam	0.36	Unmatched
S1345_N.bam	S1983_N.bam	0.38	Unmatched
S1345_N.bam	S1983_T.bam	0.34	Unmatched
S1345_T.bam	S1983_N.bam	0.36	Unmatched
S1345_T.bam	S1983_T.bam	0.32	Unmatched
S1983_N.bam	S1983_T.bam	0.37	Unmatched

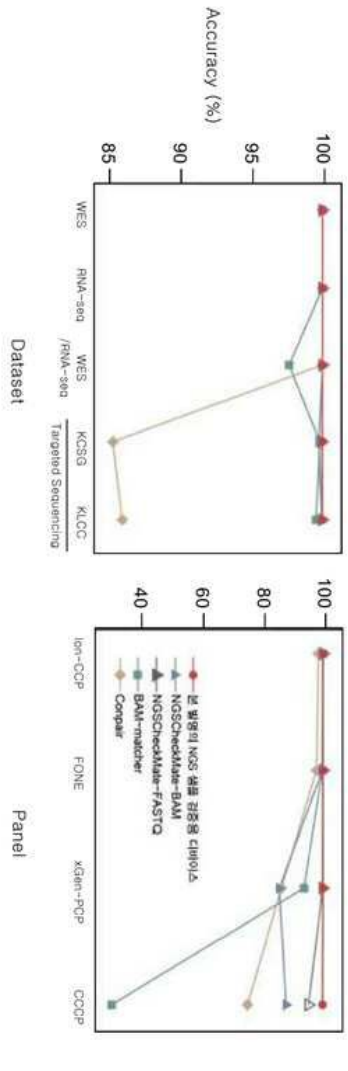
도면8a

Dataset	Data type	#Sample	#Individual	Sequencing Depth
WES	Whole exome sequencing	202	101	40~160X
RNA-seq	RNA sequencing	130	65	60~260X
WES/ RNA-seq	Whole exome sequencing / RNA sequencing	168	84	30~160X /60~270X
KCSG	Targeted sequencing	192	96	120~2310X
KLCC	Targeted sequencing	402	201	70~1810X

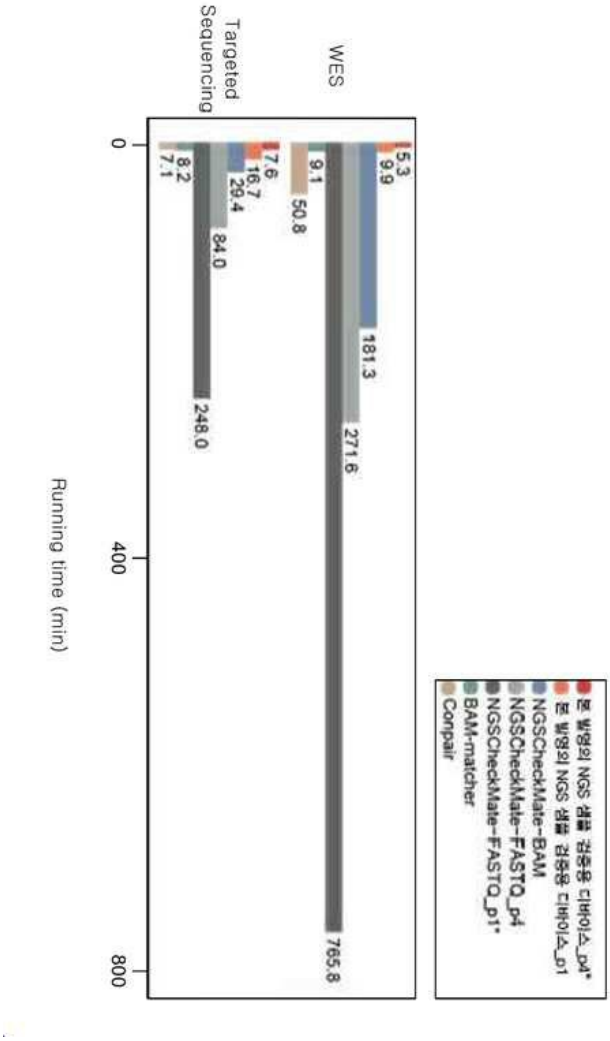
도면8b

		본 발명의 NGS 샘플 검증용 디바이스	NGSCheckMate -BAM	NGSCheckMate -FASTQ	BAM-matcher	Conpair
WES	Accuracy (%)	100	100	100	100	99.99
	Sensitivity (%)	100	100	100	100	97.33
	Specificity (%)	100	100	100	100	100
RNA-seq	Accuracy (%)	100	100	100	100	100
	Sensitivity (%)	100	100	100	100	100
	Specificity (%)	100	100	100	100	100
WES /RNA-seq	Accuracy (%)	100	100	100	99.99	100
	Sensitivity (%)	100	100	100	97.62	100
	Specificity (%)	100	100	100	100	100
KCSG	Accuracy (%)	100	99.73	99.96	99.81	85.07
	Sensitivity (%)	100	48.96	91.67	100	100
	Specificity (%)	100	100	100	99.81	84.99
KLCC	Accuracy (%)	100	99.94	99.96	99.54	85.73
	Sensitivity (%)	100	97.50	96.02	100	100
	Specificity (%)	100	99.97	99.97	99.54	85.09

도면8c


Accuracy (%)
100
95
90
85
WES
RNA-seq
WES
/RNA-seq
KCSG
KLCC
Targeted Sequencing
Dataset
100
80
60
40
Ion-CCP
FONE
xGen-PCP
CCCP
본 발명의 NGS 분석 검증용 데이터셋
NGSCheckMate-BAM
NGSCheckMate-FASTQ
BAM-matcher
Conpair
Panel

도면8d


WES
Targeted Sequencing
5.3
9.9
181.3
271.6
765.8
9.1
50.8
7.6
16.7
29.4
84.0
248.0
8.2
7.1
0
400
800
Running time (min)
NGSCheckMate-BAM
NGSCheckMate-FASTQ_p4
NGSCheckMate-FASTQ_p1*
BAM-matcher
Conpair