(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2018-0098438
(43) 공개일자 2018년09월04일

(51) 국제특허분류(Int. Cl.)
G06F 19/18 (2011.01) *G06F 19/28* (2011.01)
(52) CPC특허분류
G06F 19/18 (2013.01)
G06F 19/28 (2013.01)
(21) 출원번호 10-2017-0024789
(22) 출원일자 2017년02월24일
심사청구일자 2017년02월24일

(71) 출원인
에스디지노믹스 주식회사
서울특별시 강남구 개포로 619 ,11층(개포동,서울강남우체국)
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
사회복지법인 삼성생명공익재단
서울특별시 용산구 이태원로55길 48 (한남동)
(72) 발명자
박인호
서울특별시 강남구 개포로 619 서울강남우체국 11층 에스디지노믹스
강이옥
서울특별시 관악구 남부순환로 1430 107동 504호
(뒷면에 계속)
(74) 대리인
손민

전체 청구항 수 : 총 5 항

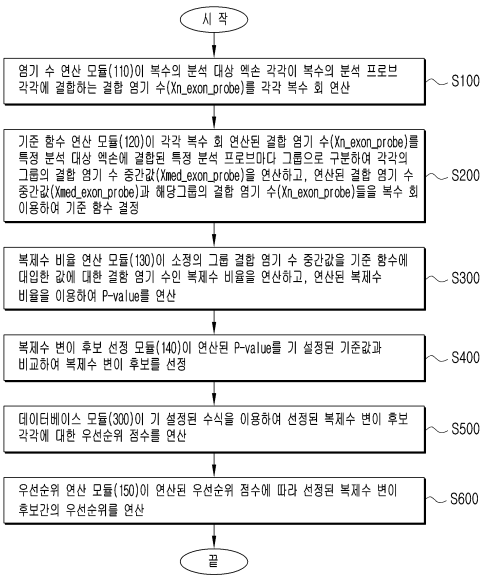
(54) 발명의 명칭 **복제수 변이 후보 우선순위 연산 방법**

(57) 요 약

복제수 변이 후보 우선순위 연산 방법을 제공한다. 본 발명의 실시예에 따른 복제수 변이 후보 우선순위 연산 방법은 (a) 염기 수 연산 모듈(110)이 복수의 분석 대상 엑손(exon) 각각이 복수의 분석 프로브(probe) 각각에 결합하는 결합 염기 수(X_n_exon_probe)를 각각 복수 회 연산하는 단계 (b) 기준 함수 연산 모듈(120)이 각각 복수

(뒷면에 계속)

대 표 도 - 도2



회 연산된 결합 염기 수(X_n_exon_probe)를 특정 분석 대상 엑손에 결합된 특정 분석 프로브마다 그룹으로 구분하여 각각의 그룹의 결합 염기 수 중간값(X_{med}_exon_probe)을 연산하고, 상기 연산된 결합 염기 수 중간값(X_{med}_exon_probe)과 해당 그룹의 결합 염기 수(X_n_exon_probe)들을 복수 회 이용함으로써 기준 함수를 결정하는 단계 (c) 복제수 비율 연산 모듈(130)이 소정의 그룹의 결합 염기 수 중간값을 상기 기준함수에 대입한 값에 대한 상기 결합 염기 수인 복제수 비율을 연산하고, 상기 연산된 복제수 비율을 이용하여 P-value를 연산하는 단계 (d) 복제수 변이 후보 선정 모듈(140)이 상기 연산된 P-value을 기 설정된 기준값과 비교하여 상기 복수의 분석 대상 엑손 중 복제수 변이 후보를 선정하는 단계 (e) 데이터베이스 모듈(300)이 기 설정된 수식을 이용하여 상기 선정된 복제수 변이 후보 각각에 대한 우선순위 점수를 연산하는 단계 및 (f) 우선순위 연산 모듈(150)이 상기 연산된 우선순위 점수에 따라 상기 선정된 복제수 변이 후보 간의 우선순위를 연산하는 단계를 포함할 수 있다.

(72) 발명자

이경아

서울특별시 강동구 고덕로 130 110동 2401호 (암사동,프라이어팰리스아파트)

김종원

서울특별시 용산구 장문로 27 5동 102호 (이태원동,청화아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호	HI12C0022
부처명	보건복지부
연구관리전문기관	한국보건산업진흥원
연구사업명	질환극복기술개발
연구과제명	통합선별검사 해석시스템 및 임상진료/치료 지침의 개발
기 여 율	1/1
주관기관	연세대학교 산학협력단
연구기간	2016.04.01 ~ 2017.03.31

명 세 서

청구범위

청구항 1

(a) 염기 수 연산 모듈(110)이 복수의 분석 대상 엑손(exon) 각각이 복수의 분석 프로브(probe) 각각에 결합하는 결합 염기 수(X_n_exon_probe)를 각각 복수 회 연산하는 단계;

(b) 기준 함수 연산 모듈(120)이 각각 복수 회 연산된 결합 염기 수(X_n_exon_probe)를 특정 분석 대상 엑손에 결합된 특정 분석 프로브마다 그룹으로 구분하여 각각의 그룹의 결합 염기 수 중간값(X_{med}_exon_probe)을 연산하고, 상기 연산된 결합 염기 수 중간값(X_{med}_exon_probe)과 해당 그룹의 결합 염기 수(X_n_exon_probe)들을 복수 회 이용함으로써 기준 함수를 결정하는 단계;

(c) 복제수 비율 연산 모듈(130)이 소정의 그룹의 결합 염기 수 중간값을 상기 기준함수에 대입한 값에 대한 상기 결합 염기 수인 복제수 비율을 연산하고, 상기 연산된 복제수 비율을 이용하여 P-value를 연산하는 단계;

(d) 복제수 변이 후보 선정 모듈(140)이 상기 연산된 P-value을 기 설정된 기준값과 비교하여 상기 복수의 분석 대상 엑손 중 복제수 변이 후보를 선정하는 단계;

(e) 데이터베이스 모듈(300)이 기 설정된 수식을 이용하여 상기 선정된 복제수 변이 후보 각각에 대한 우선순위 점수를 연산하는 단계; 및

(f) 우선순위 연산 모듈(150)이 상기 연산된 우선순위 점수에 따라 상기 선정된 복제수 변이 후보 간의 우선순위를 연산하는 단계;를 포함하는,

복제수 변이 후보 우선순위 연산 방법.

청구항 2

제 1 항에 있어서,

상기 (a) 단계에서,

상기 복수의 분석 대상 엑손이 여성(Woman)의 엑손을 포함하는 경우,

상기 염기 수 연산 모듈(110)은 상기 복수 회 연산된 결합 염기 수(X_n_exon_probe) 중에서, X 염색체가 상기 복수의 분석 프로브에 결합하는 결합 염기 수를 보정하는 단계를 더 포함하는,

복제수 변이 후보 우선순위 연산 방법.

청구항 3

제 1 항에 있어서,

상기 (d) 단계에서,

상기 기 설정된 기준값은 0.05이고,

상기 복제수 변이 후보 선정 모듈(140)은 상기 연산된 P-value가 0.05 미만인 분석 대상 엑손을 복제수 변이 후보로 선정하는,

복제수 변이 후보 우선순위 연산 방법.

청구항 4

제 1 항에 있어서,

상기 (e) 단계에서,

상기 기 설정된 수식은,

Probe Cnt In Region + Average Of Read Depth Ratios + STD of Read Depth Ratios + Average Of CIs + Average Of R2val + Sign Prove Ratio 또는 상기 수식에 포함된 변수 중 적어도 하나의 변수로 이루어진 수식이며,

상기 Probe Cnt In Region는 상기 P-value 값이 0.05 미만인 분석 대상 엑손의 개수이고, 상기 Average Of Read Depth Ratios는 상기 복제수 비율의 평균 값이며, 상기 STD of Read Depth Ratios는 상기 복제수 비율의 표준편차 값이고, 상기 Average Of CIs는 상기 복제수 비율의 신뢰 구간(confidence interval)의 평균 값이며, 상기 Average Of R2val는 상기 함수의 R-squared value의 평균 값이고, 상기 Sign Probe Ratio는 선정된 분석 대상 엑손에 대한 P-value 값이 0.05 미만인 분석 프로브의 개수의 값이며,

상기 Probe Cnt In Region은 2 이상이면 "1", 2 미만이면 "0"이고, 상기 Average Of Read Depth Ratios는 0.65 미만이면 "1", 0.65 이상 1.35 이하이면 "0", 1.35 초과이면 "1"이며, 상기 STD of Read Depth Ratios는 0.25 미만이면 "1", 0.25 이상이면 "0"이고, 상기 Average Of CIs는 0.3 미만이면 "1", 0.3 이상이면 "0"이고, 상기 Average Of R2val은 0.8 이상이면 "1", 0.8 미만이면 "0"이며, 상기 Sign Probe Ratio는 0.9 이상이면 "1", 0.9 미만이면 "0"인,

복제수 변이 후보 우선순위 연산 방법.

청구항 5

제 1 항에 있어서,

(g) 출력부(200)가 상기 우선순위 연산 모듈(150)에 의해 연산된 우선순위를 출력하는 단계;를 더 포함하는,

복제수 변이 후보 우선순위 연산 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은, 복제수 변이(Copy Number Variation) 후보들 간의 우선순위를 연산하는 방법에 관한 것이다.

배 경 기 술

[0002] 유전체(genome)란 한 생물이 가지는 모든 유전 정보를 말한다. 어느 한 개인의 유전체의 시퀀싱(sequencing)을 위해, DNA 칩 및 차세대 서열화(Next Generation Sequencing) 기술, 차차세대 서열화(Next Next Generation Sequencing) 기술 등 여러 기술들이 개발되고 있다. 핵산 서열, 단백질 등과 같은 유전 정보들의 분석은 당뇨병, 암과 같은 질병을 발현시키는 유전자를 찾거나, 유전적 다양성과 개체의 발현 특성 간의 상관관계 등을 파악하기 위하여 폭넓게 활용된다. 특히, 개인으로부터 수집된 유전 데이터는 서로 다른 증상이나 질병의 진행과 관련된 개인의 유전적인 특징을 규명하는데 있어서 중요하다. 따라서, 개인의 핵산 서열, 단백질 등과 같은 유전 데이터는 현재와 미래의 질병 관련 정보를 파악하여 질병을 예방하거나 질병의 초기 단계에서 최적의 치료 방법을 선택할 수 있도록 하는 핵심적인 데이터이다. 생물의 유전 정보들로서 SNP(Single Nucleotide Polymorphsim), CNV(Copy Number Variation) 등을 검출하는 유전체 검출 장비를 활용하여 개인의 유전 데이터를 정확히 분석하고, 개인의 질병을 진단하는 기술들이 연구 중에 있다.

[0004] 여기서 CNV(Copy Number Variation)는 복제수 변이라 불리우며, 대표적인 게놈과 비교해서 특정 염색체의 상대적으로 큰 영역이 결손되거나 증폭되어 반복적으로 나타나는 유전체 DNA의 변이를 말한다. 인간 유전체 변이 중에 약 12%를 차지하며, 그 크기는 1 kilobase에서 여러 megabase까지에 이른다. 복제수 변이는 유전에 의한 것일 수도 있고 새로 생길수도 있다. 최근 제안된 복제수 변이의 원인으로는 fork stalling, 주형 변화(template switching), 복제착오(replication misstep)가 있다.

[0005] 복제수 변이는 결손, 중복, 역위, 전좌와 같은 유전체 재배열에 의해 발생할 수 있다. 저단위반복서열(low copy repeats, LCRs)은 유전체 재배열이 일어나기 쉬운데, 크기, 방향, 유사성, 복제본 사이의 거리와 같은 요인이 유전체 재배열에 영향을 준다.

[0006] 복제수 변이는 형광가시적 분자결합화(fluorescent in situ hybridization), 단순 비교유전체 부합법(comparative genomic hybridization), 비교유전체 부합법(arry comparative genomic hybridization), SNP array(단일염기변이 배열)를 이용한 가상의 유전체분석(virtual karyotype)과 같은 세포유전학 기술을 이용해 발견될 수 있다. 인간게놈프로젝트의 완성으로 인해 유전자 복제수 변이가 인간에게 나타나는 광범위하고 일반적인 현상이라는 것이 밝혀졌다.

[0007] 친족인 아닌 사람들간의 유전체의 약 0.4%가 복제수 차이로 인한 것이라고 추정된다. 일란성쌍둥이로부터 복제수 변이가 새로 생길 수 있다는 사실이 발견되었으며, 다른 유전적 변이들과 같이 복제수 변이도 질병과 관련이 있다는 것이 밝혀졌다. 유전자 복제수 변이는 암세포를 증가시킬 수 있는데, 예를 들어 EGFR 복제수는 폐암세포에서 더 많이 나타난다. CCL3L1 복제수가 많으면 HIV 감염에 저항성이 높고, FCGR3B 복제수가 많으면 전신성 홍반성 루푸스와 그와 비슷한 자가면역성 질환에 걸리기 쉽다.

[0008] 복제수 변이는 또한 자폐증, 정신분열증, 특발성 학습장애와 연관이 있다. 복제수 변이는 단일 유전자에 한정되거나 인접한 여러 유전자에서 일어날 수 있다. 빠르게 성장하는 Escherichia coli 세포같이 어떤 경우에는 DNA 복제 기점 근처에 위치한 유전자가 DNA 복제말단에 있는 유전자의 유전자 복제수보다 4배 클수도 있다. 유전자 복제수의 증가는 그것이 암호화하는 단백질의 발현을 증가시킨다.

[0010] 이러한 이유에서, DNA에서 복제수 변이를 검출하는 과정은 중요하다. 복제수 변이를 검출하는 과정에서 다양한 복제수 변이 후보들이 선정되는데, 모든 복제수 변이 후보를 검증하여 복제수 변이를 찾는 것은 비용면에서나 시간면에서나 비효율적이다. 다양한 복제수 변이 후보 중 신뢰도가 높은 복제수 후보들을 제공해준다면 사용자는 최소한의 비용과 시간으로 복제수 변이 후보를 검증할 수 있으므로, 이에 대한 수요는 크게 늘고 있는 실정이다.

선행기술문헌

특허문헌

[0011] (특허문헌 0001) 한국공개특허문헌 제10-2017-0000744호 (2017.01.03)

(특허문헌 0002) 한국공개특허문헌 제10-2016-0073405호 (2016.06.24)

발명의 내용

해결하려는 과제

[0012] 본 발명은 분석 대상 엑손에서 발생할 수 있는 복제수 변이를 검출하기 위해 사용되는 복제수 변이 후보들 간의 우선순위를 연산하는 방법을 제공하고자 한다.

과제의 해결 수단

[0013] 상기와 같은 과제를 해결하기 위한 본 발명의 일 실시예는, (a) 염기 수 연산 모듈(110)이 복수의 분석 대상 엑손(exon) 각각이 복수의 분석 프로브(probe) 각각에 결합하는 결합 염기 수(X_n_exon_probe)를 각각 복수 회 연산하는 단계, (b) 기준 함수 연산 모듈(120)이 각각 복수 회 연산된 결합 염기 수(X_n_exon_probe)를 특정 분석 대상 엑손에 결합된 특정 분석 프로브마다 그룹으로 구분하여 각각의 그룹의 결합 염기 수 중간값(X_{med}_exon_probe)을 연산하고, 상기 연산된 결합 염기 수 중간값(X_{med}_exon_probe)과 해당 그룹의 결합 염기 수(X_n_exon_probe)들을 복수 회 이용함으로써 기준 함수를 결정하는 단계, (c) 복제수 비율 연산 모듈(130)이 소정의 그룹의 결합 염기 수 중간값을 상기 기준함수에 대입한 값에 대한 상기 결합 염기 수인 복제수 비율을 연산하고, 상기 연산된 복제수 비율을 이용하여 P-value를 연산하는 단계, (d) 복제수 변이 후보 선정 모듈(140)이 상기 연산된 P-value을 기 설정된 기준값과 비교하여 상기 복수의 분석 대상 엑손 중 복제수 변이 후보를 선정하는 단계, (e) 데이터베이스 모듈(300)이 기 설정된 수식을 이용하여 상기 선정된 복제수 변이 후보 각각에 대한 우선순위 점수를 연산하는 단계 및 (f) 우선순위 연산 모듈(160)이 상기 연산된 우선순위 점수에 따라 상기 선정된 복제수 변이 후보 간의 우선순위를 연산하는 단계를 포함하는, 복제수 변이 후보 우선순위 연산 방법

을 제공한다.

[0014] 일 실시예에 있어서, 상기 (a) 단계에서, 상기 복수의 분석 대상 엑손이 여성(Woman)의 엑손을 포함하는 경우, 상기 염기 수 연산 모듈(110)은 상기 복수 회 연산된 결합 염기 수(X_n_exon_probe) 중에서, X 염색체가 상기 복수의 분석 프로브에 결합하는 결합 염기 수를 보정하는 단계를 더 포함하는 것이 바람직하다.

[0015] 일 실시예에 있어서, 상기 (d) 단계에서, 상기 기 설정된 기준값은 0.05이고, 상기 복제수 변이 후보 선정 모듈(140)은 상기 연산된 P-value가 0.05 미만인 분석 대상 엑손을 복제수 변이 후보로 선정하는 것이 바람직하다.

[0016] 일 실시예에 있어서, 상기 (e) 단계에서, 상기 기 설정된 수식은 Probe Cnt In Region + Average Of Read Depth Ratios + STD of Read Depth Ratios + Average Of CIs + Average Of R2val + Sign Prove Ratio 또는 상기 수식에 포함된 변수 중 적어도 하나의 변수로 이루어진 수식이며, 상기 Probe Cnt In Region는 상기 P-value 값이 0.05 미만인 분석 대상 엑손의 개수이고, 상기 Average Of Read Depth Ratios는 상기 복제수 비율의 평균 값이며, 상기 STD of Read Depth Ratios는 상기 복제수 비율의 표준편차 값이고, 상기 Average Of CIs는 상기 복제수 비율의 신뢰 구간(confidence interval)의 평균 값이며, 상기 Average Of R2val는 상기 함수의 R-squared value의 평균 값이고, 상기 Sign Probe Ratio는 선정된 분석 대상 엑손에 대한 P-value 값이 0.05 미만인 분석 프로브의 개수의 값이며, 상기 Probe Cnt In Region은 2 이상이면 "1", 2 미만이면 "0"이고, 상기 Average Of Read Depth Ratios는 0.65 미만이면 "1", 0.65 이상 1.35 이하이면 "0", 1.35 초과이면 "1"이며, 상기 STD of Read Depth Ratios는 0.25 미만이면 "1", 0.25 이상이면 "0"이고, 상기 Average Of CIs는 0.3 미만이면 "1", 0.3 이상이면 "0"이고, 상기 Average Of R2val은 0.8 이상이면 "1", 0.8 미만이면 "0"이며, 상기 Sign Probe Ratio는 0.9 이상이면 "1", 0.9 미만이면 "0"인 것이 바람직하다.

[0017] 일 실시예에 있어서, (g) 출력부(200)가 상기 우선순위 연산 모듈(150)에 의해 연산된 우선순위를 출력하는 단계를 더 포함하는 것이 바람직하다.

발명의 효과

[0018] 많은 유전적 질병은 하나 혹은 그 이상의 엑손의 복제수에 문제가 생긴 것이 원인이 되는데, 본 발명에 이용되는 분석 프로브는 하나의 분석 대상 엑손과 결합하게 되어 하나의 엑손에서 발생하는 복제수 변이를 검출할 수 있는 장점을 갖는다.

[0019] 다수의 통계 과정을 거쳐 복제수 변이 후보 간의 우선순위를 연산하는 것이므로 그 정확도와 신뢰도가 높다.

[0020] 복제수 변이 후보 간의 우선순위가 연산되면, 사용자는 모든 복제수 변이 후보를 검증하지 않고 높은 점수를 갖는 복제수 변이 후보만을 검증하면 되므로, 시간면에서나 비용면에서 효과적으로 복제수 변이를 검출할 수 있게 된다.

도면의 간단한 설명

[0021] 도 1은 본 발명의 실시예에 따른 복제수 변이 후보 우선순위 연산 방법에 사용되는 구성요소를 설명하기 위한 도면이다.

도 2는 도 1의 복제수 변이 후보 우선순위 연산 방법의 순서도를 나타낸 도면이다.

도 3은 복수의 분석 대상 엑손 각각이 복수의 분석 프로브 각각에 결합하는 결합 염기 수를 연산하는 과정을 개략적으로 설명하기 위한 도면이다.

도 4는 각 분석 프로브별 복제수 변이가 없는 상태(Normal)일 때의 결합 염기 수를 예측할 수 있는 기준 함수를 나타낸 도면이다.

도 5는 도 4의 기준 함수를 이용하여 복제수 비율을 연산하는 것을 설명하기 위한 도면이다.

도 6은 하나의 분석 대상 엑손에서 하나의 분석 프로브에 대한 결과를 정리한 예시를 나타낸 도면이다.

도 7 내지 도 10은 선정된 복제수 변이 후보 간의 우선순위를 연산하기 위해 이용되는 수식의 각 항목의 개념을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0022] 첨부된 도면을 참조하여, 본 발명의 실시예에 따른 복제수 변이 후보 우선순위 연산 방법에 대해 구체적으로 설

명한다.

[0023] 먼저, 본 발명은 엑손(exon)에 대한 복수의 복제수 변이 후보 간의 우선순위를 연산하게 된다. 엑손은 진핵생물에서 실제로 단백질을 만들어내는 염기 서열 부분을 말한다. 엑손에 대한 복제수 변이 후보 간의 우선순위를 연산함으로써, 복제수 변이로 인해 실제로 발생할 수 있는 장애 현상 등을 정확하게 예측할 수 있는 장점이 있다.

[0025] 복제수 변이 후보 우선순위 연산을 위해 다음과 같은 과정이 선행된다.

[0026] 사람의 혈액이나 다른 조직으로부터 추출한 DNA를 가지고 DNA의 염기서열을 알아내는 시퀀서 기기를 이용하여 염기서열 데이터(fastq format)를 생성한다. 염기서열 데이터는 인간이 읽을 수 있는 텍스트 형태의 파일 형태로서, 읽어낸 염기서열(A, T, G, C)과 각 염기에 상응하는 qulity score로 구성되어 있다. 염기서열 데이터를 읽어서 각 염기서열을 reference genome에 나열(alignment) 후, sorting, de-duplication, re-alignment, recalibration 등의 과정을 거쳐 유전변이 분석을 할 수 있는 형태인 bam 파일을 생성해낸다. 본 발명의 실시예는 bam 파일을 input으로 입력받아 복제수 변이 후보 우선순위를 연산하게 된다.

[0028] 도 1을 참조하면, 연산부(100)와 출력부(200) 그리고 데이터베이스 모듈(300)이 본 발명의 실시예에 따른 복제수 변이 후보 우선순위 연산 방법에 사용된다.

[0029] 연산부(100)는 염기 수 연산 모듈(110), 기준 함수 연산 모듈(120), 복제수 비율 연산 모듈(130), 복제수 변이 후보 선정 모듈(140) 및 우선순위 연산 모듈(150)을 포함한다.

[0030] 염기 수 연산 모듈(110)은 복수의 분석 대상 엑손 각각이 복수의 분석 프로브(probe) 각각에 결합하는 결합 염기 수(X_n_exon_probe)를 각각 복수 회 연산한다.(S100) 분석 프로브는 분석 대상 엑손의 복제수 변이를 검출하기 위한 부분으로서, 다수의 염기 서열로 이루어진다. 많은 유전적 질병은 하나 혹은 그 이상의 엑손의 복제수에 문제가 생긴 것이 원인이 되는데, 본 발명에 이용되는 각 분석 프로브는 하나의 분석 대상 엑손과 결합하게 되어 하나의 엑손에서 발생하는 복제수 변이를 검출할 수 있는 장점을 갖는다.

[0031] 도 3에는 6개의 분석 프로브(BAIT1, BAIT2, BAIT3, BAIT4, BAIT5, BAIT6)가 도시된다. 염기 수 연산 모듈(110)은 복수의 분석 프로브(BAIT1, BAIT2, BAIT3, BAIT4, BAIT5, BAIT6)의 염기 서열과 결합하는 분석 대상 엑손의 염기 수를 복수 회 연산하게 된다.

[0032] 여성의 경우 XX 염색체를 가지며, 남성의 XY보다 X 염색체가 1개 더 많다. 따라서 X 염색체에 대한 보정이 필요한데, 염기 수 연산 모듈(110)은 분석 대상 엑손이 여성의 엑손을 포함하는 경우, X 염색체가 복수의 분석 프로브에 결합하는 결합 염기 수(X_n_exon_probe)를 2로 나누어 준다. 이로써, 남성과 여성의 경우 모두 동일한 결과를 얻을 수 있다.

[0034] 기준 함수 연산 모듈(120)은 염기 수 연산 모듈(110)이 연산한 결합 염기 수(X_n_exon_probe)를 이용하여 기 설정된 방법에 의해 기준 함수를 결정한다.(S200)

[0035] 구체적으로, 각각 복수 회 연산된 결합 염기 수(X_n_exon_probe)를 특정 분석 대상 엑손에 결합된 특정 분석 프로브마다 그룹으로 구분하여 각각의 그룹의 결합 염기 수 중간값(X_{med}_exon_probe)을 연산하고, 연산된 결합 염기 수 중간값(X_{med}_exon_probe)과 해당 그룹의 결합 염기 수(X_n_exon_probe)들을 복수 회 이용하여 기준 함수를 결정하게 된다.(도 4)

[0036] 다시 말해, 결합 염기 수 중간값(X_{med}_exon_probe)으로부터 각 분석 프로브별 복제수 변이가 없는 상태(Normal)일 때의 결합 염기 수를 예측할 수 있는 기준 함수를 결정하게 된다.(도 4)

[0037] 이 때, 기준 함수를 연산하기 위해 선형 회귀(linear regression) 방정식을 이용하고, bootstrapping 방법을 통해 해당 함수가 얼마나 신뢰성(R-squared value)이 있으며, 그 함수로부터 예측되는 값들의 신뢰구간(confidence interval)을 평가할 수 있다. 수집된 데이터로부터 선형 회귀(linear regression) 방정식을 세우고, 그 과정에서 bootstrapping을 이용하는 기술은 통계 분석에서 널리 사용되고 있으므로, 자세한 설명한 설명

은 생략하기로 한다.

[0038] 기준 함수의 x 값은 결합 염기 수 중간값(X_{med}_exon_probe)이며, y 값은 결합 염기 수(X_n_exon_probe)이다. 본 발명에서는 n개의 분석 대상 엑손을 n개의 중복을 허용하여 랜덤으로 선별한 뒤 선형 회귀 방정식을 생성하는 bootstrapping 방법을 1000번 반복하게 된다. 다시 말해, 하나의 분석 프로브에 대한 선형 회귀 방정식이 1000개씩 만들어지는 것이다. 도 4-(a)에서 표시된 부분은 bootstrapping을 통해 랜덤 추출된 데이터를 의미한다. 여기서 빨간색으로 표시된 이상점(outlier)을 제거하고, 파란색으로 표시된 데이터들은 모두 복제수 변이가 없는 상태(normal)라고 가정한다. 다시 말해, 기준 함수 연산 모듈(120)은 각 분석 프로브별 결합 염기 수(X_n_exon_probe)를 예측할 수 있는 선형 회귀 함수인 기준 함수를 결정하는 것이다. (도 4-(b))

[0040] 복제수 비율 연산 모듈(130)은 기준 함수를 이용하여, 소정의 그룹의 결합 염기 수 중간값을 기준 함수에 대입한 값에 대한 결합 염기 수인 복제수 비율을 연산한다. (S300) 어떠한 분석 대상 엑손이 있는 경우, 분석 대상 엑손의 결합 염기 수 중간값(X_{med}_exon_probe)을 x값으로 입력하면, y 값인 복제수 변이 없는 상태(normal)에서의 결합 염기 수(X_n_exon_probe)를 예측할 수 있다. 다시 말해, 염기 수 연산 모듈(110)에서 연산된 결합 염기 수(X_n_exon_probe)를 기준 함수의 y 값으로 나누어주면 복제수 비율을 구할 수 있다.

[0041] 복제수 비율이 1이면, normal 상태일 때의 결합 염기 수를 그대로 갖는 것이므로, 그 분석 대상 엑손은 해당 분석 프로브 위치에서는 normal이다. 복제수 비율이 0.5이면, normal 상태일 때보다 결합 염기 수가 반 밖에 되지 않으므로 복제수가 줄어들었음을 알 수 있다. 반대로, 복제수 비율이 1.5이면 normal 상태보다 결합 염기 수가 1.5배 증가한 것이므로 복제수가 증가하였음을 알 수 있다.

[0042] 다음, 하나의 분석 대상 엑손 당 복수의 분석 프로브 모두에 대한 복제수 비율 1000개를 이용하여 복제수 비율의 중간값을 구하고, P-value값을 계산한다. 다수의 데이터로부터 P-value를 구하는 방법은 이미 공지된 기술이므로 자세한 설명은 생략한다. 구체적으로, 복제수 비율이 역치값(threshold)을 넘을 확률로 P-value를 연산한다. 도 6은 하나의 분석 대상 엑손에서 하나의 분석 프로브에 대한 결과를 정리한 예시이다. 해당 과정을 통해 도 6의 값을 연산할 수 있다.

[0044] 복제수 변이 후보 선정 모듈(140)은 복제수 비율 연산 모듈(130)이 연산한 P-value를 이용하여 복수의 분석 대상 엑손 중 복제수 변이 후보를 선정한다. (S400) 구체적으로, 연산된 P-value가 0.05 미만인 경우, 그 분석 대상 엑손을 복제수 변이 후보로 선정하게 된다. 복제수 변이 후보 선정 모듈(140)에 의해 적어도 하나의 복제수 변이 후보가 선정될 수 있다.

[0046] 데이터베이스 모듈(300)은 기 설정된 수식을 이용하여 선정된 복제수 변이 후보 각각에 대한 우선순위 점수를 연산한다. (S500)

[0047] 여기서 기 설정된 수식은 Probe Cnt In Region + Average Of Read Depth Ratios + STD of Read Depth Ratios + Average Of CIs + Average Of R2val + Sign Prove Ratio 또는 상기 수식에 포함된 변수 중 적어도 하나의 변수로 이루어진 수식이다.

[0048] 다시 말해, 기 설정된 수식은 Probe Cnt In Region + Average Of Read Depth 일 수 있고, Probe Cnt In Region + Average Of Read Depth Ratios + STD of Read Depth Ratios 일 수도 있으며, Probe Cnt In Region + Average Of Read Depth Ratios + STD of Read Depth Ratios + Average Of CIs + Average Of R2val + Sign Prove Ratio 일 수도 있다.

[0049] 수식에 포함된 변수에 대해 구체적으로 설명한다. Probe Cnt In Region는 복수의 분석 대상 엑손에 대한 P-value 값이 0.05 미만인 분석 프로브의 개수이다. 도 7의 좌측 도면을 참고하면, 해당 복제수 변이 후보를 지지하는 신호(빨간 네모로 도시됨)의 수는 40개이고, 우측 도면은 11개이다. 이 값이 클수록 복제수 변이 후보의 신뢰성이 상승하게 된다.

[0050] Average Of Read Depth Ratios는 복제수 비율의 평균 값이다. 해당 복제수 변이 후보를 지지하는 복제 수 비율이 normal 값(=1)에서 얼마나 멀리 떨어져 있는지를 의미한다. 도 7에서 빨간 네모로 도시된 부분은 복제 수 비

율의 중간값을 의미하는데, 이 값들의 평균을 구하면 Average Of Read Depth Ratios를 구할 수 있다. 이 값은 복제수 감소의 경우 1보다 작을수록 복제수 변이 후보의 신뢰성이 상승하고, 복제수 증가의 경우 1보다 클수록 복제수 변이 후보의 신뢰성이 상승한다.

[0051] STD of Read Depth Ratios는 복제수 비율의 표준편차 값이다. 해당 복제수 변이 후보를 지지하는 복제 수 비율이 얼마나 일정한가를 의미한다. 복제수 비율의 중간값이 1에서 멀리 떨어져 있더라도, 그 값들의 표준편차가 크다면 안정적이라고 볼 수 없다. 복제수 비율의 중간값들의 표준편차를 구하면 STD of Read Depth Ratios가 된다. 이 값은 0에 가까울수록 복제수 변이 후보의 신뢰성이 상승한다.

[0052] Average Of CIs는 복제수 비율의 신뢰 구간(confidence interval)의 평균 값이다. 도 7에서 각 빨간 네모의 수염들이 신호의 신뢰구간(confidence interval)을 의미한다. 우측의 경우 빨간 네모의 수염들이 좌측보다 더 긴 것을 볼 수 있다. 이것은 우측 후보의 신호들이 좌측에 비해 부정확한 복제 수 비율 예측을 했다는 것을 의미한다. 따라서, 각 복제 수 비율의 신뢰 구간의 길이의 평균을 구하면 Average Of Cis가 된다. 이 값은 0에 가까울수록 복제수 변이 후보의 신뢰도가 상승한다.

[0053] Average Of R2val는 함수의 R-squared value의 평균 값이다. 도 7에서 각 빨간 네모에 해당하는 값들을 구하기 위해 1000개의 선형 회귀 방정식이 사용되었다. 각 데이터의 R-squared value는 해당 선형 회귀 방정식이 얼마나 예측을 잘하는지를 의미한다. 예를 들어, 도 8의 두 선형 회귀 방정식은 비슷한 regression(빨간 실선으로 도시)을 갖지만, R-squared value가 다르다. R-squared value가 낮은 우측 선형 회귀 방정식에서 예측된 복제 수 비율은 noise일 확률이 높으므로 신뢰성이 낮아지게 된다. 따라서, 1000개의 선형 회귀 방정식의 R-squared value의 평균을 구하고, 하나의 복제수 변이 후보를 지지하는 신호의 평균 R-squared value의 평균을 구하면 Average Of R2val이 된다. 이 값은 1에 가까울수록 복제수 변이 후보의 신뢰도가 상승한다.

[0054] Sign Probe Ratio는 선정된 복제수 변이 후보 개수에 대한 P-value 값이다. 도 9에 도시된 것처럼 넓은 영역의 복제수 변이 후보를 찾는 과정을 통해 복제수 변이 후보 안에 P-value 값이 0.05 이상인 복제 수 비율 신호가 포함될 수 있다. 복제수 변이 후보 안에 P-value 값이 0.05 미만으로 유의한 신호가 많을수록 해당 후보의 신뢰성이 높다고 볼 수 있다. 도 9에서는 총 15개의 신호 중 유의한 신호가 10개이므로 Sign Probe Ratio는 10/15=0.67이 된다. 이 값은 1에 가까울수록 복제수 변이 후보의 신뢰도가 상승한다.

[0055] 더 구체적으로, 각 항목에 대한 점수는 Probe Cnt In Region은 2 이상이면 "1", 2 미만이면 "0"이고, Average Of Read Depth Ratios는 0.65 미만이면 "1", 0.65 이상 1.35 이하이면 "0", 1.35 초과이면 "1"이며, STD of Read Depth Ratios는 0.25 미만이면 "1", 0.25 이상이면 "0"이고, Average Of CIs는 0.3 미만이면 "1", 0.3 이상이면 "0"이고, Average Of R2val은 0.8 이상이면 "1", 0.8 미만이면 "0"이며, Sign Probe Ratio는 0.9 이상이면 "1", 0.9 미만이면 "0"이다.

[0056] 선정된 복제수 변이 후보 각각에 대해 최소 0점부터 최대 6점의 우선순위 점수가 연산될 수 있다.

[0058] 우선순위 연산 모듈(150)은 연산된 우선순위 점수에 따라 선정된 복제수 변이 후보 간의 우선순위를 연산한다.(S600)

[0059] 기 설정된 수식에서 모든 항목이 1점이면 최대 6점의 점수를 갖게 되는데, 우선순위 연산 모듈(150)은 수식을 통해 연산된 점수가 높은 복제수 변이 후보를 최우선순위로 연산하게 된다. 적어도 하나의 복제수 변이 후보 간의 우선순위를 연산하는 것이다. 우선순위 연산 모듈(150)에 의해 적어도 하나의 복제수 변이 후보 간의 우선순위가 연산되면, 사용자는 모든 복제수 변이 후보를 검증하지 않고 높은 점수를 갖는 복제수 변이 후보만을 검증하면 되므로, 시간면에서나 비용면에서 효과적으로 복제수 변이를 검출할 수 있게 된다.

[0061] 출력부(200)는 연산부(100)와 유선 또는 무선으로 연결되어, 연산부(100)에 의해 연산된 결과를 출력한다. 출력부(200)에 출력되는 결과는 복제수 변이 후보 간의 우선순위 결과일 수도 있고, 연산 과정에서 이용되는 기준 함수나 복제수 비율의 결과값일 수도 있다.

[0063] 이상, 본 명세서에는 본 발명을 당업자가 용이하게 이해하고 재현할 수 있도록 도면에 도시한 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 당업자라면 본 발명의 실시예로부터 다양한 변형 및 균등한 타 실

시예가 가능하다는 점을 이해할 것이다. 따라서 본 발명의 보호범위는 청구범위에 의해서 정해져야 할 것이다.

부호의 설명

[0065] 100: 연산부

110: 염기 수 연산 모듈

120: 기준 함수 연산 모듈

130: 복제수 비율 연산 모듈

140: 복제수 변이 후보 선정 모듈

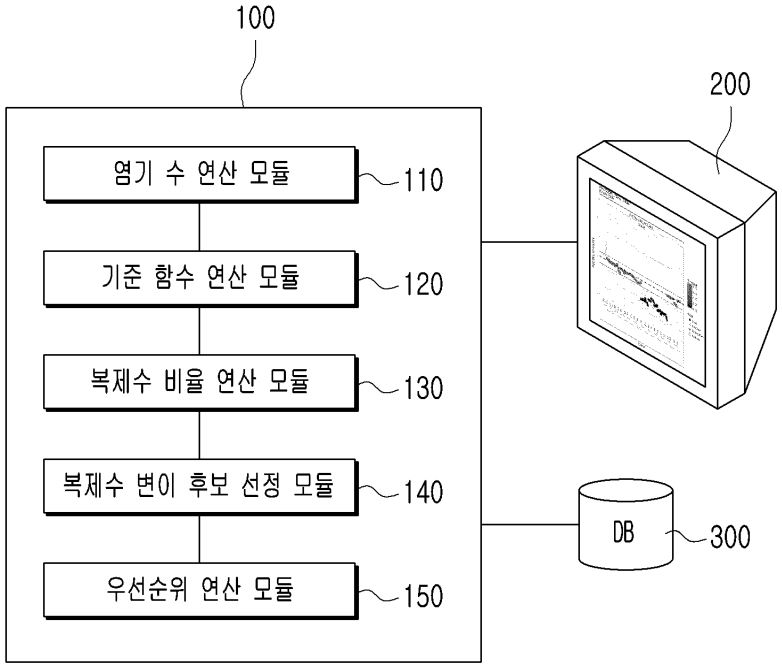
150: 우선순위 연산 모듈

200: 출력부

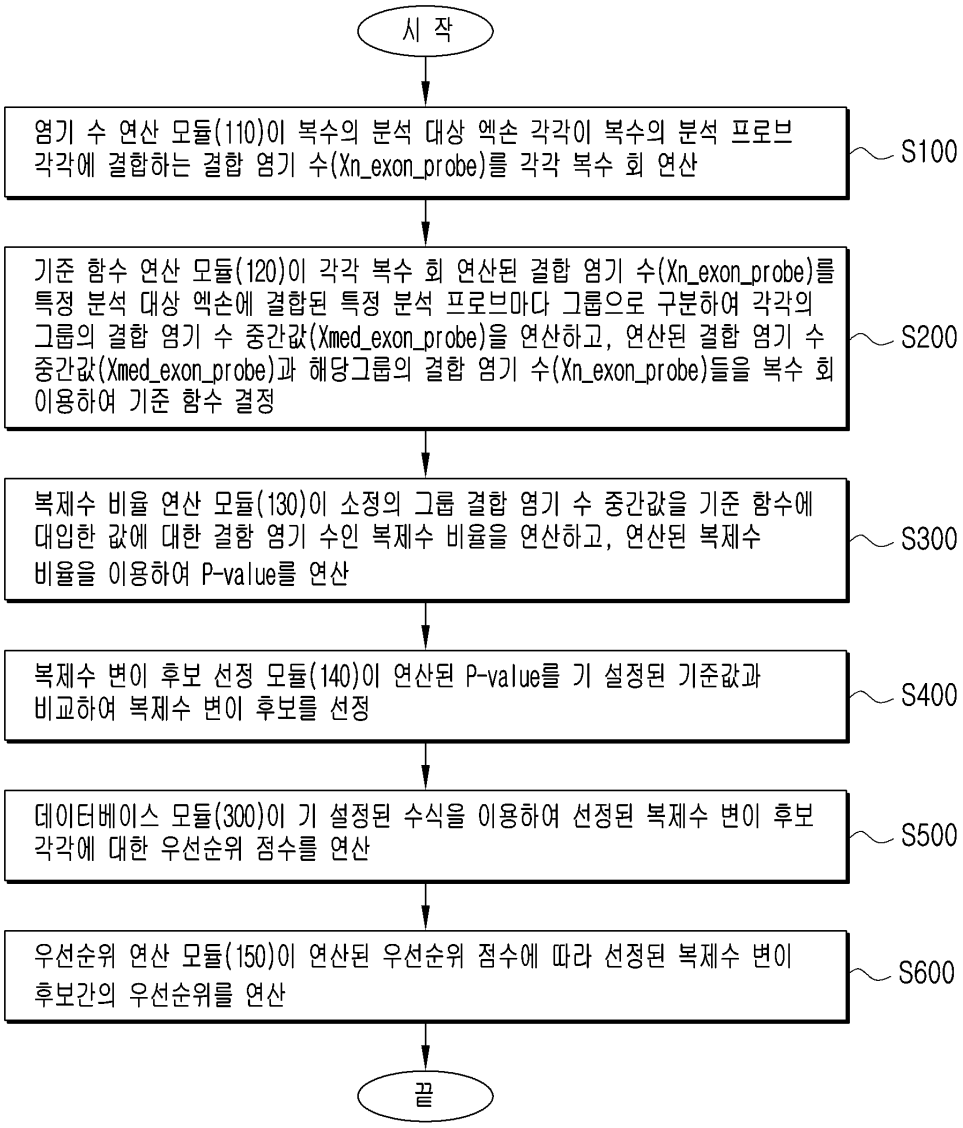
300: 데이터베이스 모듈

도면

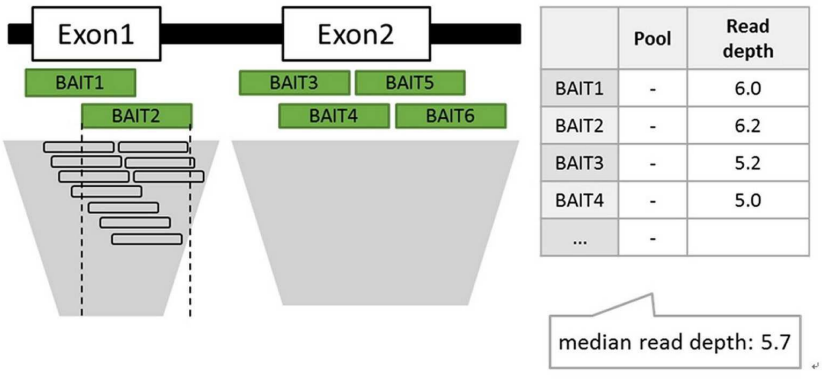
도면1



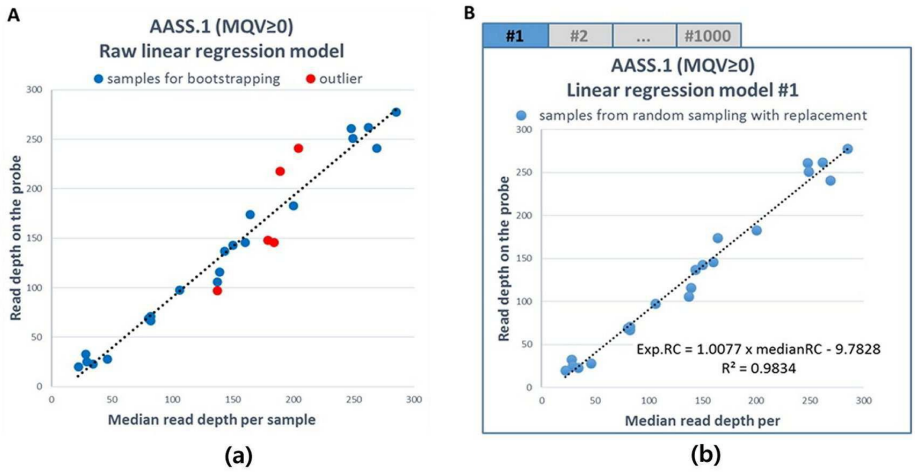
도면2


시 작
염기 수 연산 모듈(110)이 복수의 분석 대상 엑손 각각이 복수의 분석 프로브 각각에 결합하는 결합 염기 수(Xn_exon_probe)를 각각 복수 회 연산
S100
기준 함수 연산 모듈(120)이 각각 복수 회 연산된 결합 염기 수(Xn_exon_probe)를 특정 분석 대상 엑손에 결합된 특정 분석 프로브마다 그룹으로 구분하여 각각의 그룹의 결합 염기 수 중간값(Xmed_exon_probe)을 연산하고, 연산된 결합 염기 수 중간값(Xmed_exon_probe)과 해당그룹의 결합 염기 수(Xn_exon_probe)들을 복수 회 이용하여 기준 함수 결정
S200
복제수 비율 연산 모듈(130)이 소정의 그룹 결합 염기 수 중간값을 기준 함수에 대입한 값에 대한 결함 염기 수인 복제수 비율을 연산하고, 연산된 복제수 비율을 이용하여 P-value를 연산
S300
복제수 변이 후보 선정 모듈(140)이 연산된 P-value를 기 설정된 기준값과 비교하여 복제수 변이 후보를 선정
S400
데이터베이스 모듈(300)이 기 설정된 수식을 이용하여 선정된 복제수 변이 후보 각각에 대한 우선순위 점수를 연산
S500
우선순위 연산 모듈(150)이 연산된 우선순위 점수에 따라 선정된 복제수 변이 후보간의 우선순위를 연산
S600
끝

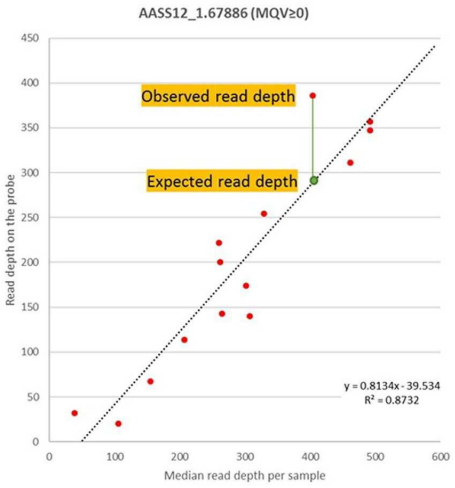
도면3


Exon1
Exon2
BAIT1
BAIT2
BAIT3
BAIT5
BAIT4
BAIT6
Pool
Read depth
BAIT1 - 6.0
BAIT2 - 6.2
BAIT3 - 5.2
BAIT4 - 5.0
... -
median read depth: 5.7

도면4


A
AASS.1 (MQV≥0)
Raw linear regression model
samples for bootstrapping
outlier
Read depth on the probe
Median read depth per sample
(a)
B
#1
#2
...
#1000
AASS.1 (MQV≥0)
Linear regression model #1
samples from random sampling with replacement
Read depth on the probe
Exp.RC = 1.0077 x medianRC - 9.7828
R² = 0.9834
Median read depth per
(b)

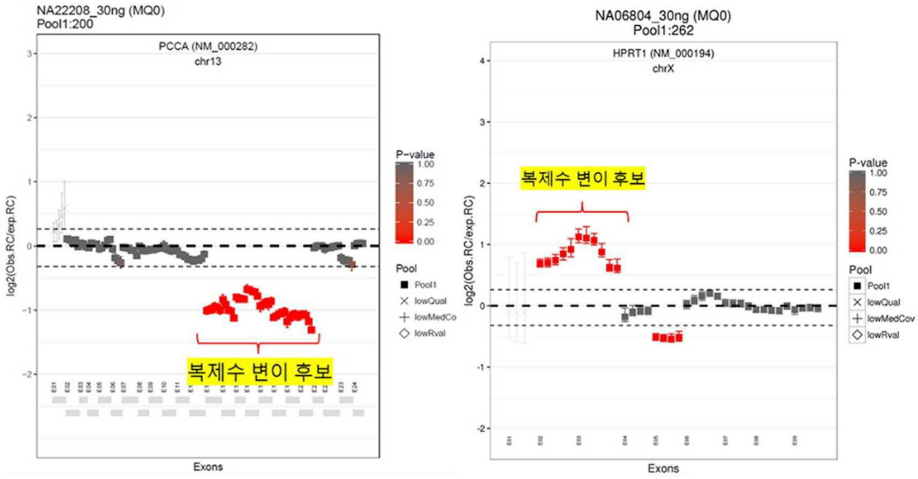
도면5


AASS12_1.67886 (MQV≥0)
Observed read depth
Expected read depth
Read depth on the probe
y = 0.8134x - 39.534
R² = 0.8732
Median read depth per sample

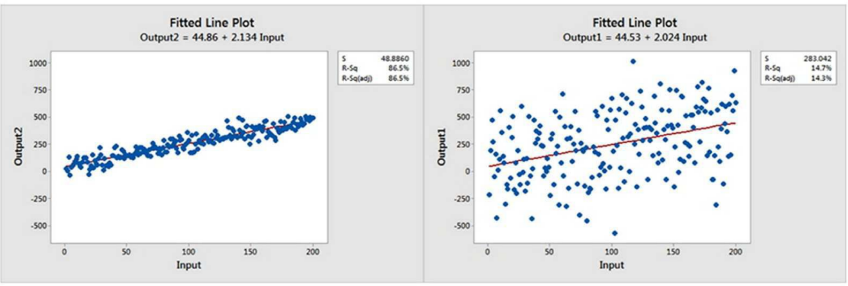
도면6

항목	설명	예시
Median read depth	해당 sample 의 median read depth	278
Observed read depth on the probe	해당 sample 의 해당 probe 의 read depth	282
Expected read depth on the probe (top2.5)	Bootstrapping 을 통해 반복한 1000 개의 model 로부터 구한 expected read depths 의 top 2.5 값	265.6
Expected read depth on the probe (median)	Bootstrapping 을 통해 반복한 1000 개의 model 로부터 구한 expected read depths 의 median 값	275.4
Expected read depth on the probe (top97.5)	Bootstrapping 을 통해 반복한 1000 개의 model 로부터 구한 expected read depths 의 top 97.5 값	283.9
Read depth ratio on the probe (top2.5)	1000 개의 expected read depth 로부터 구한 read depth ratios 의 top 2.5 값	0.993
Read depth ratio on the probe (median)	1000 개의 expected read depth 로부터 구한 read depth ratios 의 median 값	1.024
Read depth ratio on the probe (top97.5)	1000 개의 expected read depth 로부터 구한 read depth ratios 의 top 97.5 값	1.062
P-value	복제수 변이 증가/감소 에 대한 p-value	1
CNVType	복제수 변이 증가/감소/정상	정상
R-squared value	Bootstrapping 을 통해 반복한 1000 개의 model 의 R-squared value 의 평균값	0.85

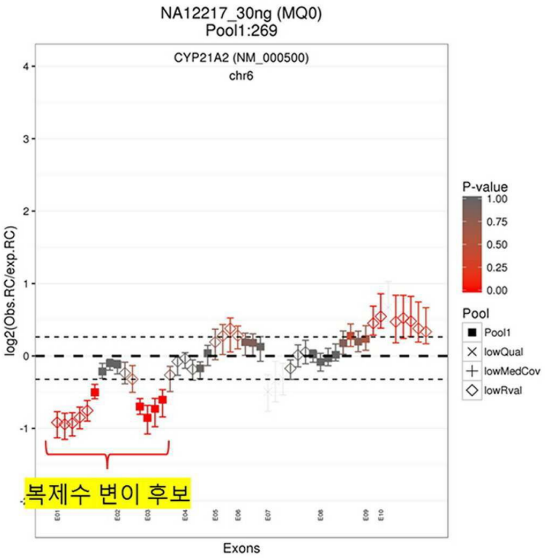
도면7



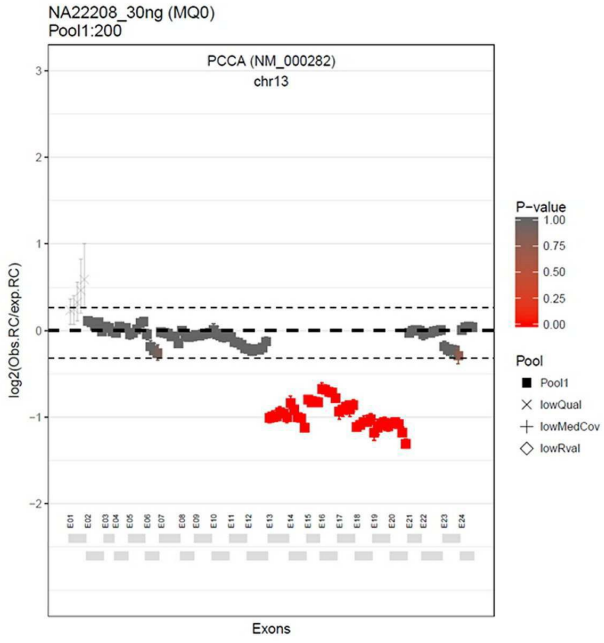
도면8



도면9


NA12217_30ng (MQ0)
Pool1:269
CYP21A2 (NM_000500)
chr6
log2(Obs.RC/exp.RC)
P-value
1.00
0.75
0.50
0.25
0.00
Pool
Pool1
lowQual
lowMedCov
lowRval
복제수 변이 후보
Exons

도면10


NA22208_30ng (MQ0)
Pool1:200
PCCA (NM_000282)
chr13
log2(Obs.RC/exp.RC)
P-value
1.00
0.75
0.50
0.25
0.00
Pool
Pool1
lowQual
lowMedCov
lowRval
Exons