(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0071122
(43) 공개일자 2022년05월31일

(51) 국제특허분류(Int. Cl.)
G16B 35/10 (2019.01) *C12Q 1/6869* (2018.01)
G16B 30/10 (2019.01) *G16H 50/50* (2018.01)
(52) CPC특허분류
G16B 35/10 (2019.02)
C12Q 1/6869 (2018.05)
(21) 출원번호 10-2021-0161004
(22) 출원일자 2021년11월22일
심사청구일자 2021년11월22일
(30) 우선권주장
1020200158049 2020년11월23일 대한민국(KR)

(71) 출원인
주식회사 지씨지놈
경기도 용인시 기흥구 이현로30번길 107 (보정동)
재단법인 아산사회복지재단
서울특별시 송파구 올림픽로43길 88 (풍납동)
울산대학교 산학협력단
울산광역시 남구 대학로 93(무거동)
(72) 발명자
조은해
경기도 용인시 기흥구 이현로 30번길 107 (보정동)
이준남
경기도 용인시 기흥구 이현로 30번길 107 (보정동)
박숙련
서울특별시 송파구 양재대로 1218, 204동 2102호 (방이동, 올림픽선수기자촌아파트)
(74) 대리인
이처영, 장제환

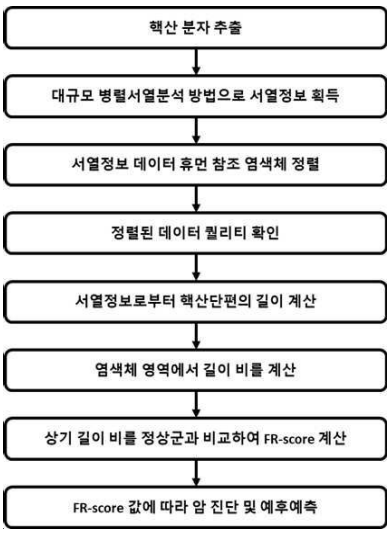
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **핵산 길이 비를 이용한 암 진단 및 예후예측 방법**

(57) 요 약

본 발명은 핵산 길이 비를 이용한 암 진단 및 예후예측 방법에 관한 것으로, 보다 구체적으로는 생체시료에서 핵산을 추출하여, 서열정보를 획득한 다음, 정렬된 핵산 단편의 길이 비를 이용한 암 진단 및 예후예측 방법에 관한 것이다. 본 발명에 따른 암 진단 및 예후예측 방법은 기존의 리드 개수(read count) 기반으로 염색체 양을 결정하는 단계를 이용하는 방식과는 달리, 정렬된 리드(reads)를 기반으로 핵산단편의 길이 비를 이용하는 검출 방법으로, 기존 방법이 리드 개수가 감소하면 정확도가 떨어지나, 본 발명의 방법은 리드 개수가 감소하더라도, 검출의 정확도를 높일 수 있을 뿐만 아니라, 모든 염색체 구간이 아닌 일정 구간의 핵산단편 길이 비를 사용하여도 검출 정확도가 높아 유용할 뿐만 아니라, 기존의 리드 개수(read count) 로는 검출할 수 없었던 염색체 이상 샘플에도 적용 가능하다.

대 표 도 - 도1



(52) CPC특허분류
G16B 30/10 (2019.02)
G16H 50/50 (2018.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1415168006
과제번호 20006145
부처명 산업통상자원부
과제관리(전문)기관명 한국산업기술평가관리원
연구사업명 현장수요의료기기고도화기술개발-현장수요반영의료기기고도화기술개발
연구과제명 간세포암 치료반응 예측을 위한 AI기술 및 유전체분석데이터 융합형 진단보조기기 개발
기 여 율 1/1
과제수행기관명 녹십자지놈
연구기간 2019.06.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

(a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;

(b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

(c) 상기 정렬된 서열정보(reads)에 대하여 핵산단편의 길이(Fragment length)를 계산하는 단계;

(d) 상기 (c) 단계에서 계산한 핵산단편의 길이를 기반으로 염색체 전체 영역 또는 특정 영역 별로 핵산단편 길이 비(Fragment ratio)를 계산하는 단계; 및

(e) 상기 길이 비를 정상 샘플군과 비교하여 FR-score를 계산하여, FR-score가 기준 값 또는 범위 미만 혹은 초과일 경우, 암이 있는 것으로 판정하거나, 예후를 예측하는 단계를 포함하는 암 진단 또는 예후예측을 위한 정보의 제공 방법.

청구항 2

제1항에 있어서, 상기 (a) 단계는 다음의 단계를 포함하는 방법으로 수행되는 것을 특징으로 암 진단 또는 예후예측을 위한 정보의 제공 방법:

(a-i) 생체시료에서 핵산을 수득하는 단계;

(a-ii) 채취된 핵산에서 솔팅-아웃 방법(salting-out method), 컬럼 크로마토그래피 방법(column chromatography method) 또는 비드 방법(beads method)을 사용하여 단백질, 지방, 및 기타 잔여물을 제거하고 정제된 핵산을 수득하는 단계;

(a-iii) 정제된 핵산 또는 효소적 절단, 분쇄, 수압 절단 방법(hydroshear method)으로 무작위 단편화(random fragmentation)된 핵산에 대하여, 싱글 엔드 시퀀싱(single-end sequencing) 또는 페어 엔드 시퀀싱(pair-end sequencing) 라이브러리(library)를 제작하는 단계;

(a-iv) 제작된 라이브러리를 차세대 유전자서열검사기(next-generation sequencer)에 반응시키는 단계; 및

(a-v) 차세대 유전자서열검사기에서 핵산의 서열정보(reads)를 획득하는 단계.

청구항 3

제1항에 있어서, 상기 리드는 페어드 엔드(paired-end) 시퀀싱으로 수득하는 것을 특징으로 하는 암 진단 또는 예후예측을 위한 정보의 제공 방법.

청구항 4

제1항에 있어서, 상기 (c) 단계의 핵산단편의 길이는 핵산단편의 양 말단에 정렬되는 리드의 정렬 위치를 통해 산출하는 것을 특징으로 하는 암 진단 또는 예후예측을 위한 정보의 제공 방법.

청구항 5

제1항에 있어서, 상기 (d) 단계는 다음의 단계를 포함하는 방법으로 수행되는 것을 특징으로 하는 암 진단 또는 예후예측을 위한 정보의 제공 방법:

(d-i) 염색체 전체 영역 또는 특정 영역별로 핵산단편을 긴 핵산단편(long fragment) 및 짧은 핵산단편(short fragment)으로 분류하는 단계;

(d-ii) 하기 수식 1을 바탕으로 핵산단편 길이 비(Fragment ratio)를 계산하는 단계;

수식 1: Fragment ratio(FR) = Number of short fragment group / Number of long fragment group

청구항 6

제5항에 있어서, 상기 (d-i) 단계는 기준점을 중심으로 기준점 이하 길이의 핵산단편은 짧은 핵산단편으로, 기준점 초과 길이의 핵산단편은 긴 핵산단편으로 분류하는 것을 특징으로 하는 암 진단 또는 예후예측을 위한 정보의 제공 방법.

청구항 7

제6항에 있어서, 상기 기준점은 50~200bp 일 수 있고, 바람직하게는 150 내지 170bp인 것을 특징으로 하는 암 진단 또는 예후예측을 위한 정보의 제공 방법.

청구항 8

제1항에 있어서, 상기 (e) 단계는 다음의 단계를 포함하는 방법으로 수행되는 것을 특징으로 하는 암 진단 또는 예후예측을 위한 정보의 제공 방법:

(e-i) 정상 샘플군에서 샘플군과 동일한 염색체 전체 영역 또는 특정 영역별로 핵산단편 길이 비를 계산하여 하기 수식 2로 상대빈도(Relative Frequency) 값을 계산하는 단계;

수식2:

상대빈도i = FR(핵산단편) 비i / ∑ {FR(핵산단편) 비}

(e-ii) 각 영역에서의 상대 빈도 값의 평균과 표준편차를 계산하는 단계;

(e-iii) 제1항의 d) 단계에서 도출한 FR 값의 상대빈도를 수식 2로 계산하여, 하기 수식 3으로 FR Z-score(FRZ)를 계산하는 단계;

수식 3: FR Z-score $_{i\ bin}$ = (분석 샘플 상대빈도$_{i\ bin}$ - 정상인 샘플의 상대빈도 평균$_{i\ bin}$) / 정상인 샘플의 상대빈도 표준편차$_{i\ bin}$

(e-iv) 각 유전영역에 해당하는 GC 값으로 LOESS regression을 수행하고, 잔차를 계산하는 단계;

(e-v) 각 유전영역별로 GC값으로 보정된 FRZ 값을 LOESS 알고리즘을 통해 정규화 하는 단계; 및

(e-iv) 하기 수식 4로 FR-score를 계산하는 단계;

수식 4: $$FR-score = \ln\{\sum_{i=all\ genomic\ position(bin)} abs(LOESS\ smoothed\ GC\ normalized\ FRZ\ i)\ \}$$

청구항 9

제1항에 있어서, 암 진단 또는 예후예측을 위한 상기 FR-score의 기준값은 5 내지 50인 것을 특징으로 하는 암 진단 또는 예후예측을 위한 정보의 제공 방법.

청구항 10

제1항에 있어서, 상기 (e) 단계의 FR-score가 기준 값 또는 범위 미만일 경우에는 예후가 나쁠 것으로 예측하고, FR-score가 기준 값 또는 범위 초과일 경우에는 예후가 좋을 것으로 예측하는 것을 특징으로 하는 암 진단 또는 예후예측을 위한 정보의 제공 방법.

청구항 11

생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부;

해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 및

선별된 서열정보(reads)를 기반으로 핵산단편의 길이를 계산하고, 이를 기반으로 핵산단편 길이 비를 측정한 다음, 정상 샘플군과 비교하여 FR-score를 계산하고, 계산한 FR-score를 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 FR-score가 기준 값 또는 구간 미만 또는 초과 일 경우, 암이 있는 것으로 판정하거나, 예후를 예측하는 암 진단 또는 예후예측부를 포함하는 암 진단 또는 예후예측 장치.

청구항 12

컴퓨터 판독 가능한 저장 매체로서, 암 진단 또는 예후예측을 위한 정보를 제공하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,

(a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;

(b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

(c) 상기 정렬된 서열정보(reads)에 대하여 핵산단편의 길이(Fragment length)를 계산하는 단계;

(d) 상기 (c) 단계에서 계산한 핵산단편의 길이를 기반으로 염색체 전체 영역 또는 특정 영역 별로 핵산단편 길이 비(Fragment ratio)를 계산하는 단계; 및

(e) 상기 길이 비를 정상 샘플군과 비교하여 FR-score를 계산하여, FR-score가 기준 값 또는 범위 미만 혹은 초과일 경우, 암이 있는 것으로 판정하거나, 예후를 예측하기 위한 정보를 제공하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 핵산 길이 비를 이용한 암 진단 및 예후예측 방법에 관한 것으로, 보다 구체적으로는 생체시료에서 핵산을 추출하여, 서열정보를 획득한 다음, 정렬된 핵산 단편의 길이 비를 이용한 암 진단 및 예후예측 방법에 관한 것이다.

배 경 기 술

[0003] 염색체 이상(chromosomal abnormality)은 유전적 결함과 종양 질환과 관련 있다. 염색체 이상은 염색체의 결실 또는 중복, 염색체 중 일부의 결실 또는 중복, 또는 염색체 내의 손상(break), 전위(translocation), 또는 역위(inversion)를 의미하는 것일 수 있다. 염색체 이상은 유전적 균형의 장애 중 하나로, 태아 사망 또는 육체 및 정신 상태의 심각한 결함 및 종양 질환을 유발한다. 예컨대, 다운증후군(Down's syndrome)은 21번 염색체가 3개 존재하여(trisomy 21) 유발되는 염색체 수 이상의 흔한 형태이다. 에드워드증후군(Edwards syndrome) (trisomy 18), 파타우 증후군(Patau syndrome) (trisomy 13), 터너증후군(Turner syndrome) (XO), 및 클라인펠터 증후군(Klinefelter syndrome) (XXY) 또한 염색체 수 이상에 해당한다. 또한 종양 환자에서도 염색체 이상이 발견 된다. 예컨대 간암 환자(Liver Adenomas and adenocarcinomas) 에서 4q, 11q, 22q 영역의 중복과 13q 영역의 결실이 확인되었고, 췌장암 환자에서는 2p, 2q, 6p, 11q 영역의 중복과 6q, 8p, 9p, 21 번 염색체 영역의 결실이 확인 되었다. 이러한 영역들은 종양과 관련된 Oncogene, Tumor suppressor gene 영역과 관련이 되어 있다.

[0004] 염색체 이상은 핵형 검사(Karyotype), FISH(Fluorescent In Situ Hybridization)를 사용하여 검출 가능하다. 이러한 검출법은 시간, 노력 및 정확도 측면에서 불리하다. 또한, DNA 마이크로어레이를 염색체 이상 검출에 사용할 수 있다. 특히, 게놈 DNA 마이크로어레이 시스템의 경우, 프로브의 제작이 용이하고 염색체의 확장된 영역뿐 아니라 염색체의 인트론 영역에서의 염색체 이상을 검출할 수 있지만, 염색체 내의 위치화 및 기능이 확인된 DNA 단편을 많은 수로 제작하기에 곤란하다.

[0005] 최근, 차세대 시퀀싱 기술이 염색체 수 이상 분석에 사용되고 있다(Park, H., Kim et al., Nat Genet 2010, 42, 400-405.; Kidd, J. M. et al., Nature 2008, 453, 56-64). 그러나 이 기술은 염색체 수 이상 분석을 위한 높은 coverage reading을 요구하며, CNV 측정은 독립적인 입증(validation)을 또한 필요로 한다. 따라서 비용이 매우 높고, 결과를 이해하기 어려우므로, 그 당시 일반적인 유전자 검색분석으로서 적절하지 못하였다.

[0006] 실시간 qPCR이 현재 정량적인 유전자 분석용 첨단 기술로서 사용되는데, 이는 넓은 동역학범위(Weaver, S. et al, Methods 2010, 50, 271-276) 및 역치 주기(threshold cycle)와 초기 타겟 양 사이에 선형적인 상관관계가 재현성 있게 관찰되기 때문이다(Deepak, S. et al., Curr Genomics 2007,8, 234-251). 그러나 qPCR 분석의 민감도는 복제수 차이를 구별할 만큼 충분히 높지 않다.

[0007] 한편, 태아 염색체 이상에 대한 기존 산전 검사 항목에는 초음파 검사, 혈중 표지자 검사, 양수검사, 융모막검사, 경피제대혈검사 등이 존재한다(Mujezinovic F, et al. Obstet Gynecol. 2007, 110(3):687-94.). 이 중 초음파 검사와 혈중 표지자 검사는 선별검사, 양수 염색체 검사는 확진 검사로 분류한다. 비침습적 방법인 초음파 검사와 혈중 표지자 검사는 태아에 대한 직접적인 시료 채취를 하지 않아 안전한 방법이지만 검사의 민감도가 80% 이하로 떨어진다(ACOG Committee on Practice Bulletins. 2007). 침습적 방법인 양수검사, 융모막검사, 경피제대혈 검사는 태아 염색체 이상을 확진할 수 있으나, 침습적 의료행위로 인한 태아의 소실 확률이 존재한다는 단점이 있다.

[0008] 1997년 Lo 등이 모체 혈장 및 혈청에서 태아 유래 유전물질을 Y 염색체 염기서열분석에 성공하여 모체 내 태아 유전물질을 산전 검사에 이용하게 되었다(Lo YM, et al. Lancet. 1997, 350(9076):485-7). 모체 혈액 내의 태아 유전물질은 태반 재형성 과정 중 세포사멸과정을 겪은 영양막 세포의 일부분이 물질교환 기전을 통해 모체 혈액으로 들어간 것으로 실제로는 태반으로부터 유래하고 이를 cff DNA(cell-free fetal DNA)라 정의한다.

[0009] cff DNA는 빠르면 배아 이식 18일째부터, 37일째에는 대부분의 모체 혈액 내에서 발견된다. cff DNA는 300bp 이하의 짧은 가닥이며 모체혈액 내 소량으로 존재하는 특징을 가지고 있기 때문에 이를 태아염색체 이상 검출에 적용하기 위하여 차세대염기서열분석기법(NGS)을 이용한 대규모 병렬 염기분석 기술이 사용되고 있다. 대규모 병렬 염기분석 기술을 이용한 비침습적 태아 염색체 이상 검출 성능은 염색체에 따라 90-99% 이상의 검출 민감도를 나타내고 있으나, 위양성 및 위음성 결과가 1-10%에 해당하고 있어 이에 대한 교정 기술이 필요한 시점이다(Gil MM, et al. Ultrasound Obstet Gynecol. 2015, 45(3):249-66).

[0010] 또한, 세포유리 핵산의 길이 데이터와 염색체 암(arm) 복제수 변이 데이터 및 미토콘드리아 복제수 변이 데이터를 함께 기계학습하여 암 진단에 활용하는 기술(Cristiano S. et al., Nature. 2019, Vol. 570(7761), pp. 385-389), 세포 유리 핵산의 조각 패턴(fragmentation pattern)을 학습하여 암 환자를 분류하는 기술(Mouliere F et al., Sci Transl Med. 2018, Vol.10(466). pii: eaat4921) 및 세포 유리 핵산 단편의 패턴, 위치를 이용하여 세포유리 핵산의 기원 또는 유전자 이상을 검출하는 기술(KR 10-2017-0044660, KR 10-2019-0026837, KR 10-2019-0132558) 등이 공지되었으나, 세포유리 핵산의 길이 비(fragment ratio) 정보만을 기반으로 높은 정확도와 민감도로 염색체 이상을 검출하는 기술은 아직 알려져 있지 않은 실정이다.

[0011] 이에, 본 발명자들은 상기 문제점들을 해결하고, 높은 민감도와 정확도의암 진단 및 예후 예측 방법을 개발하기 위해 예의 노력한 결과, 염색체 영역에 정렬되는 리드를 기반으로 핵산단편의 길이 비를 계산하여 정상인 그룹과 비교할 경우, 높은 민감도와 정확도로 암 진단 및 예후 예측을 수행할 수 있다는 것을 확인하고, 본 발명을 완성하였다.

발명의 내용

해결하려는 과제

[0013] 본 발명의 목적은 핵산 길이 비를 이용한 암 진단 및 예후예측 방법을 제공하는 것이다.

[0014] 본 발명의 다른 목적은 핵산 길이 비를 이용한 암 진단 및 예후예측 장치를 제공하는 것이다.

[0015] 본 발명의 또 다른 목적은 상기 방법으로 암 진단 및 예후예측 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공하는 것이다.

과제의 해결 수단

[0017] 상기 목적을 달성하기 위하여, 본 발명은 a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계; b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계; c) 상기 정렬된 서열정보(reads)에 대하여 핵산단편의 길이(Fragment length)를 계산하는 단계; d) 상기 c) 단계에서 계산한 핵산단편의 길이를 기반으로 염색체 전체 영역 또는 특정 영역 별로 핵산단편 길이 비(Fragment ratio)를 계산하는 단계; 및 e) 상기 길이 비를 정상 샘플군과 비교하여 FR-score를 계산하여, FR-score가 기준 값 또는 범위 미만 혹은 초과일 경우, 암이 있는 것으로 판정하거나, 예후를 예측하는 단계를 포함하는 암 진단 또는 예후예측을 위한 정보의 제공 방법을 제공한다.

[0019] 본 발명은 또한, 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부; 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 및 선별된 서열정보(reads)를 기반으로 핵산단편의 길이를 계산하고, 이를 기반으로 핵산단편 길이 비를 측정한 다음, 정상 샘플군과 비교하여 FR-score를 계산하고, 계산한 FR-score를 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 FR-score가 기준 값 또는 구간 미만 또는 초과 일 경우, 암이 있는 것으로 판정하거나 예후를 예측하는 암 진단 또는 예후예측부를 포함하는 암 진단 또는 예후예측 장치를 제공한다.

[0020] 본 발명은 또한, 컴퓨터 판독 가능한 저장 매체로서, 암 진단 및 예후예측을 위한 정보를 제공하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되, a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계; b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계; c) 상기 정렬된 서열정보(reads)에 대하여 핵산단편의 길이(Fragment length)를 계산하는 단계; d) 상기 c) 단계에서 계산한 핵산단편의 길이를 기반으로 염색체 전체 영역 또는 특정 영역 별로 핵산단편 길이 비(Fragment ratio)를 계산하는 단계; 및 e) 상기 길이 비를 정상 샘플군과 비교하여 FR-score를 계산하여, FR-score가 기준 값 또는 범위 미만 혹은 초과일 경우, 암이 있는 것으로 판정하거나, 예후를 예측하기 위한 정보를 제공하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공한다.

발명의 효과

[0022] 본 발명에 따른 암 진단 및 예후예측 방법은, 기존의 리드 개수(read count) 기반으로 염색체 양을 결정하는 단계를 이용하는 방식과는 달리, 정렬된 리드(reads)를 기반으로 핵산단편의 길이 비를 이용하는 검출 방법으로, 기존 방법이 리드 개수가 감소하면 정확도가 떨어지나, 본 발명의 방법은 리드 개수가 감소하더라도, 검출의 정확도를 높일 수 있을 뿐만 아니라, 모든 염색체 구간이 아닌 일정 구간의 핵산단편 길이 비를 사용하여도 검출 정확도가 높아 유용할 뿐만 아니라, 기존의 리드 개수(read count) 로는 검출할 수 없었던 염색체 이상 샘플에도 적용 가능하다.

도면의 간단한 설명

[0024] 도 1은 본 발명의 염색체 이상을 판정하기 위한 전체 흐름도이다.

도 2는 본 발명에서 계산하는 핵산단편 길이를 계산하는 방법을 도식화 한 것이다.

도 3은 본 발명에서 계산하는 FR-score를 도출하는 과정을 도식화한 것으로, 정상인 샘플에서 FR Ratio, 상대빈도, 상대빈도의 표준값, 평균을 계산한 다음, 샘플군에서도 동일한 값을 계산하고 GC 값으로 보정 한 뒤, LOESS smoothing을 수행한 다음 수식으로 계산하는 과정을 나타낸 것이다.

도 4는 본 발명의 일 실시예에 따라 도출한 FR-score의 예시를 나타낸 것이다.

도 5는 본 발명의 일 실시예에 따른 정상인과 HCC 환자의 세포유리 핵산 길이 분포를 관찰한 결과이다.

도 6은 본 발명의 일 실시예에 따른 insert size별 누적 길이 값(A)과 평균값의 차이를 delta로 정의하고 그 분포를 관찰한 결과(B)이다.

도 7은 본 발명의 일 실시예에 따른 insert size 별 delta의 최대 값을 도출한 결과이다.

도 8은 본 발명에서 개발한 방법으로 정상인과 HCC 환자군을 구별하는 민감도를 측정한 결과이다.

도 9는 본 발명에서 개발한 방법으로 정상인과 HCC 환자군을 구별하는 ROC 분석 결과이다.

도 10는 본 발명의 일 실시예에 따른 reads 수에 따른 FR-score의 분포이다.

도 11는 본 발명의 일 실시예에 따른 FR-score 분포에 따른 식도암 환자의 생존 데이터의 분석 결과로서, (A)와 (B)는 FR-score가 기준값보다 높은 환자의 TTP(Time to Progression) 및 OS(Overall Survival)를 의미하고, (C) 및 (D)는 FR-socre가 기준값보다 낮은 환자의 TPP 및 OS를 의미한다.

도 12는 본 발명의 일 실시예에 따른 FR-score 에 따라 간암 환자를 두 그룹으로 나눈 후, 생존 데이터를 분석한 결과로서, (A) 환자의 TTP(Time to Progression)를 의미하고, (B)는 OS(Overall Survival)를 의미한다.

도 13은 본 발명의 일 실시예에 따른 FR-score 에 따라 간암 환자를 네 그룹으로 나눈 후, 생존 데이터를 분석한 결과로서, (A) 환자의 TTP(Time to Progression)를 의미하고, (B)는 OS(Overall Survival)를 의미한다.

도 14는 본 발명의 일 실시예에 따른 FR-score 에 따라 간암 환자를 여섯 그룹으로 나눈 후, 생존 데이터를 분석한 결과로서, (A) 환자의 TTP(Time to Progression)를 의미하고, (B)는 OS(Overall Survival)를 의미한다.

발명을 실시하기 위한 구체적인 내용

[0025] 다른 식으로 정의되지 않는 한, 본 명세서에서 사용된 모든 기술적 및 과학적 용어들은 본 발명이 속하는 기술 분야에서 숙련된 전문가에 의해서 통상적으로 이해되는 것과 동일한 의미를 갖는다. 일반적으로 본 명세서에서 사용된 명명법 및 이하에 기술하는 실험 방법은 본 기술 분야에서 잘 알려져 있고 통상적으로 사용되는 것이다.

[0026] 본 발명에서는, 샘플에서 획득한 서열 분석 데이터를 참조 유전체에 정렬한 다음, 정렬된 리드를 기반으로 핵산 단편의 길이 비를 계산하여 정상인 집단과 실험 대상자의 분석하고자 하는 염색체에서의 길이 비를 비교하여 염색체 이상을 검출할 경우, 높은 민감도와 정확도로 염색체 이상을 검출할 수 있다는 것을 확인하였다.

[0028] 즉, 본 발명의 일 실시예에서는, 혈액에서 추출한 DNA를 시퀀싱 한 뒤, 참조 염색체에 정렬한 다음, 정렬된 리드를 기반으로 핵산단편의 길이를 계산하고, 짧은 핵산단편과 긴 핵산단편의 길이 비를 도출한 다음, 정상인 참조 집단과 비교하여 FR-score를 도출하였으며, FR-score가 기준값 미만 또는 초과일 경우, 실험 대상자의 염색체 이상이 있다고 결정하는 방법을 개발하였다(도 1)

[0029] 따라서, 본 발명은 일관점에서,

[0030] (a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;

[0031] (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

[0032] (c) 상기 정렬된 서열정보(reads)에 대하여 핵산단편의 길이(Fragment length)를 계산하는 단계;

[0033] (d) 상기 (c) 단계에서 계산한 핵산단편의 길이를 기반으로 염색체 전체 영역 또는 특정 영역 별로 핵산단편 길이 비(Fragment ratio)를 계산하는 단계; 및

[0034] (e) 상기 길이 비를 정상 샘플군과 비교하여 FR-score를 계산하여, FR-score가 기준 값 또는 범위 미만 혹은 초과일 경우, 암이 있는 것으로 판정하거나, 예후를 예측하는 단계를 포함하는 암 진단 또는 예후예측을 위한 정보의 제공 방법에 관한 것이다.

[0036] 본 발명에서 용어 “암” 또는 “악성종양”은 체내 세포의 세포 주기가 조절되지 않아 세포분열을 계속하여 발생하는 질병을 의미하며, 그 원인은 정상적인 세포의 유전자나 암 억제 유전자의 돌연변이가 누적되어 염색체

이상이 생겨 발생하는 것으로 알려져 있다.

[0037] 상기 염색체 이상은 염색체에서 발생하는 다양한 변이를 의미하는데, 크게 수 이상과 구조 이상, 미세결실, 염색체 불안정성 등으로 구분될 수 있다.

[0038] 예를 들어, 간암에서는 염색체 1q21, 1q21-23, 1q21-q22, 1q21.1-q23.2, 1q24.1-24.2, 8q-24.21-24.22, 8q21.13, 8q22.3, 8q24.3 및 7q21.3 등에서 염색체 중복(gain)이 발생하는 것으로 알려져 있으며, 염색체 4q34.3-35, 4q13.1-q35.2, 8p, 8p22-p23 및 6q26-q217에서는 이형접합체 상실(Loss of Heterozygosity, LOH)가 발생하는 것으로 알려져 있다(Zhao-Shan Niu et al., World J Gastroenterol, Vol. 722(41), pp. 9069-9095, 2016).

[0039] 또한, 교모세포종(glioblastoma)에서는 염색체 7의 중복(gain)과 염색체 10의 결실(loss)가 관찰되고, 두경부 편평세포암종(Head and neck squamous cell carcinoma, HNSCC)에서는 염색체 3q, 5p, 8p 및 11q 중복(gain) 또는 염색체 3(3q26-29)의 중복이 관찰되며, 구강 편평세포암종(Oral squamous cell carcinoma, OSCC)에서는 염색체 11q22.1-q22.2의 증폭이 관찰되고, 폐암에서는 염색체 1q의 중복 또는 염색체 7p의 중복이 관찰되며, 유방암에서는 염색체 1q21.3의 증폭, 염색체 16q의 결실 또는 염색체 17의 복제수 이상이 관찰되고, 불응성 B세포 전구체 급성림프모구성 백혈병(B cell precursor acute lymphoblastic leukemia, B-ALL)에서는 염색체 21의 중복이 관찰되는 것으로 알려져 있다(Fan Kou et al., Molecular Therapy: Oncolytics, Vol. 17, pp. 562-570, 2020).

[0040] 본 발명에서 상기 암은 고형암 또는 혈액암일 수 있으며, 바람직하게는 간암, 교모세포종, 난소암, 대장암, 두경부암, 방광암, 신장세포암, 위암, 유방암, 전이암, 전립선암, 췌장암, 갑상선암, 담낭암, 담도암, 폐암, 구강암, 흑색종, 자궁경부암, 골육종, 뇌종양, 소장암, 식도암, 직장암, 안암, 요도암, 후두암, 비호지킨 림프종, 다발성골수종, 급성 골수성 백혈병, 림프종, 급성 림프모구 백혈병 및 만성 골수성 백혈병으로 구성된 군으로부터 선택될 수 있으며, 더욱 바람직하게는 간암일 수 있으나, 이에 한정되는 것은 아니다.

[0042] 본 발명에 있어서,

[0043] 상기 (a) 단계는

[0044] (a-i) 생체시료에서 핵산을 수득하는 단계;

[0045] (a-ii) 채취된 핵산에서 솔팅-아웃 방법(salting-out method), 컬럼 크로마토그래피 방법(column chromatography method) 또는 비드 방법(beads method)을 사용하여 단백질, 지방, 및 기타 잔여물을 제거하고 정제된 핵산을 수득하는 단계;

[0046] (a-iii) 정제된 핵산 또는 효소적 절단, 분쇄, 수압 절단 방법(hydroshear method)으로 무작위 단편화(random fragmentation)된 핵산에 대하여, 싱글 엔드 시퀀싱(single-end sequencing) 또는 페어 엔드 시퀀싱(pair-end sequencing) 라이브러리(library)를 제작하는 단계;

[0047] (a-iv) 제작된 라이브러리를 차세대 유전자서열검사기(next-generation sequencer)에 반응시키는 단계; 및

[0048] (a-v) 차세대 유전자서열검사기에서 핵산의 서열정보(reads)를 획득하는 단계를 포함하는 방법으로 수행되는 것을 특징으로 할 수 있다.

[0050] 본 발명에 있어서, 상기 생체시료는 개체로부터 얻어지거나 개체로부터 유래된 임의의 물질, 생물학적 체액, 조직 또는 세포를 의미하는 것으로, 예를 들면, 전혈(whole blood), 백혈구(leukocytes), 말초혈액 단핵 세포(peripheral blood mononuclear cells), 백혈구 연층(buffy coat), (혈장(plasma) 및 혈청(serum)을 포함하는) 혈액, 객담(sputum), 눈물(tears), 점액(mucus), 세비액(nasal washes), 비강 흡인물(nasal aspirate), 호흡(breath), 소변(urine), 정액(semen), 침(saliva), 복강 세척액(peritoneal washings), 골반 내 유체액(pelvic fluids), 낭종액(cystic fluid), 뇌척수막 액(meningeal fluid), 양수(amniotic fluid), 선액(glandular fluid), 췌장액(pancreatic fluid), 림프액(lymph fluid), 흉수(pleural fluid), 유두 흡인물(nipple aspirate), 기관지 흡인물(bronchial aspirate), 활액(synovial fluid), 관절 흡인물(joint aspirate), 기관 분비물(organ secretions), 세포(cell), 세포 추출물(cell extract), 정액, 모발, 타액, 소변, 구강세포, 태반세포, 뇌척수액(cerebrospinal fluid) 및 이의 혼합물을 포함할 수 있으나, 이에 제한되는 것은 아니다.

[0052] 본 발명에 있어서, 상기 차세대 유전자서열검사기(next-generation sequencer)는 당업계에 공지된 임의의 시퀀싱 방법으로 사용될 수 있다. 선택 방법에 의해 분리된 핵산의 시퀀싱은 전형적으로는 차세대 시퀀싱(NGS)을 사용하여 수행된다. 차세대 시퀀싱은 개개의 핵산 분자 또는 고도로 유사한 방식으로 개개의 핵산 분자에 대해 클론으로 확장된 프록시 중 하나의 뉴클레오타이드 서열을 결정하는 임의의 시퀀싱 방법을 포함한다(예를 들어, 10^5개 이상의 분자가 동시에 시퀀싱된다). 일 실시형태에서, 라이브러리 내 핵산 종의 상대적 존재비는 시퀀싱 실험에 의해 만들어진 데이터에서 그것의 동족 서열의 상대적 발생 수를 계측함으로써 추정될 수 있다. 차세대 시퀀싱 방법은 당업계에 공지되어 있고, 예를 들어 본 명세서에 참조로서 포함된 문헌(Metzker, M. (2010) Nature Biotechnology Reviews 11:31-46)에 기재된다.

[0053] 일 실시형태에서, 차세대 시퀀싱은 개개의 핵산 분자의 뉴클레오타이드 서열을 결정하기 위해 한다(예를 들어, 헬리코스 바이오사이언스(Helicos BioSciences)의 헬리스코프 유전자 시퀀싱 시스템(HeliScope Gene Sequencing system) 및 퍼시픽바이오사이언스의 팩바이오 알에스 시스템(PacBio RS system)). 다른 실시형태에서, 시퀀싱, 예를 들어, 더 적지만 더 긴 리드를 만들어내는 다른 시퀀싱 방법보다 시퀀싱 단위 당 서열의 더 많은 염기를 만들어내는 대량병렬의 짧은-리드 시퀀싱(예를 들어, 캘리포니아주 샌디에고에 소재한 일루미나 인코포레이티드(Illumina Inc.) 솔렉사 시퀀서(Solexa sequencer)) 방법은 개개의 핵산 분자에 대해 클론으로 확장된 프록시의 뉴클레오타이드 서열을 결정한다(예를 들어, 캘리포니아주 샌디에고에 소재한 일루미나 인코포레이티드(Illumina Inc.) 솔렉사 시퀀서(Solexa sequencer); 454 라이프 사이언스(Life Sciences)(코네티컷주 브랜포드에 소재) 및 아이온 토렌트(Ion Torrent)). 차세대 시퀀싱을 위한 다른 방법 또는 기계는, 이하에 제한되는 것은 아니지만, 454 라이프 사이언스(Life Sciences)(코네티컷주 브랜포드에 소재), 어플라이드 바이오시스템스(캘리포니아주 포스터 시티에 소재; SOLiD 시퀀서), 헬리코스 바이오사이언스 코포레이션(매사추세츠주 캠브릿지에 소재) 및 에멀젼 및 마이크로 유동 시퀀싱 기법 나노 점적(예를 들어, 지누바이오(GnuBio) 점적)에 의해 제공된다.

[0054] 차세대 시퀀싱을 위한 플랫폼은, 이하에 제한되는 것은 아니지만, 로슈(Roche)/454의 게놈 시퀀서(Genome Sequencer: GS) FLX 시스템, 일루미나(Illumina)/솔렉사(Solexa) 게놈 분석기(Genome Analyzer: GA), 라이프(Life)/APG의 서포트 올리고(Support Oligonucleotide Ligation Detection: SOLiD) 시스템, 폴로네이터(Polonator)의 G.007 시스템, 헬리코스 바이오사이언스의 헬리스코프 유전자 시퀀싱 시스템(Helicos BioSciences' HeliScope Gene Sequencing system) 및 퍼시픽 바이오사이언스(Pacific Biosciences)의 팩바이오 알에스(PacBio RS) 시스템을 포함한다.

[0055] NGS 테크놀로지스는, 예를 들어 주형 제조, 시퀀싱 및 이미징 및 데이터 분석 단계 중 하나 이상을 포함할 수 있다.

[0056] 주형 제조. 주형 제조를 위한 방법은 핵산(예를 들어, 게놈 DNA 또는 cDNA)을 작은 크기로 무작위로 파괴하는 단계 및 시퀀싱 주형(예를 들어, 단편 주형 또는 메이트-쌍 주형)을 만드는 단계와 같은 단계들을 포함할 수 있다. 공간적으로 분리된 주형은 고체 표면 또는 지지체에 부착되거나 또는 고정될 수 있는데, 이는 대량의 시퀀싱 반응이 동시에 수행되도록 한다. NGS 반응을 위해 사용될 수 있는 주형의 유형은, 예를 들어 단일 DNA 분자로부터 유래된 클론이 증폭된 주형 및 단일 DNA 분자 주형을 포함한다.

[0057] 클론이 증폭된 주형의 제조방법은, 예를 들어 에멀젼 PCR(emulsion PCR: emPCR) 및 고체상 증폭을 포함한다.

[0058] EmPCR은 NGS를 위한 주형을 제조하기 위해 사용될 수 있다. 전형적으로, 핵산 단편의 라이브러리가 만들어지며, 보편적 프라이밍 부위를 함유하는 어댑터는 단편의 말단에 결찰된다. 그 다음에 단편은 단일 가닥으로 변성되고, 비드에 의해 포획된다. 각 비드는 단일 핵산 분자를 포획한다. 증폭 및 emPCR 비드의 풍부화 후, 다량의 주형이 부착될 수 있고, 표준 현미경 슬라이드(예를 들어, 폴로네이터(Polonator)) 상에서 폴리아크릴아마이드 겔에 고정되며, 아미노-코팅된 유리 표면(예를 들어, Life/APG; 폴로네이터(Polonator))에 화학적으로 가교되거나, 또는 개개의 피코타이터플레이트(PicoTiterPlate: PTP) 웰(예를 들어, 로슈(Roche)/454) 상에 증착되는데, 이때 NGS 반응이 수행될 수 있다.

[0059] 고체상 증폭이 또한 사용되어 NGS를 위한 주형을 생성할 수 있다. 전형적으로, 전방 및 후방 프라이머는 고체지지체에 공유적으로 부착된다. 증폭된 단편의 표면 밀도는 지지체 상에서 프라이머 대 주형의 비로써 정의된다. 고체상 증폭은 수백만개의 공간적으로 분리된 주형 클러스터(예를 들어, 일루미나/솔렉사(Illumina/Solexa))를 생성할 수 있다. 주형 클러스터의 말단은 NGS 반응을 위한 보편적 프라이머에 혼성화될 수 있다.

[0060] 클론으로 증폭된 주형의 제조를 위한 다른 방법은, 예를 들어 다중 치환 증폭(Multiple Displacement Amplification: MDA)(Lasken R. S. Curr Opin Microbiol. 2007; 10(5):510-6)을 포함한다. MDA는 비-PCR 기반 DNA 증폭 기법이다. 반응은 주형에 대해 무작위 헥사머 프라이머를 어닐링하는 단계 및 일정한 온도에서 고충실도 효소, 전형적으로 Φ에 의해 DNA를 합성하는 단계를 수반한다. MDA는 더 낮은 오류 빈도로 거대한 크기의 생성물을 만들 수 있다.

[0061] PCR과 같은 주형 증폭 방법은 표적에 NGS 플랫폼을 결합시킬 수 있거나 또는 게놈의 특이적 영역을 풍부화할 수 있다(예를 들어, 엑손). 대표적인 주형 풍부화 방법은, 예를 들어 마이크로점적 PCR 기법(Tewhey R. et al., Nature Biotech. 2009, 27:1025-1031), 맞춤-설계된 올리고뉴클레오타이드 마이크로어레이(예를 들어, 로슈(Roche)/님블젠(NimbleGen) 올리고뉴클레오타이드 마이크로어레이) 및 용액-기반 혼성화 방법(예를 들어, 분자 역위 프로브(molecular inversion probe: MIP))(Porreca G. J. et al., Nature Methods, 2007, 4:931-936; Krishnakumar S. et al., Proc. Natl. Acad. Sci. USA, 2008, 105:9296-9310; Turner E. H. et al., Nature Methods, 2009, 6:315-316) 및 바이오틴화된 RNA 포획 서열(Gnirke A. et al., Nat. Biotechnol. 2009;27(2):182-9)을 포함한다.

[0062] 단일-분자 주형은 NGS 반응을 위해 사용될 수 있는 주형의 다른 유형이다. 공간적으로 분리된 단일 분자 주형은 다양한 방법에 의해 고체 지지체 상에 고정될 수 있다. 한 접근에서, 개개의 프라이머 분자는 고체 지지체에 공유적으로 부착된다. 어댑터는 주형에 첨가되고, 주형은 그 다음에 고정된 프라이머에 혼성화된다. 다른 접근에서, 단일-분자 주형은 고정된 프라이머로부터 단일-가닥의 단일-분자 주형을 프라이밍하고 연장시킴으로써 고체 지지체에 공유적으로 부착된다. 그 다음에 보편적 프라이머는 주형에 혼성화된다. 또 다른 접근에서, 단일 폴리머라제 분자는 프라이밍된 주형이 결합된 고체 지지체에 부착된다.

[0063] 시퀀싱 및 이미징. NGS를 위한 대표적인 시퀀싱 및 이미징 방법은, 이하에 제한되는 것은 아니지만, 사이클릭 가역적 종결(cyclic reversible termination: CRT), 결찰에 의한 시퀀싱(sequencing by ligation: SBL), 단일-분자 첨가(파이로시퀀싱(pyrosequencing)) 및 실시간 시퀀싱을 포함한다.

[0064] CRT는 뉴클레오타이드 포함, 형광 이미징 및 절단 단계를 최소로 포함하는 사이클릭 방법에서 가역 종결자를 사용한다. 전형적으로, DNA 폴리머라제는 프라이머에 주형 염기의 상보적 뉴클레오타이드에 대해 상보적인 단일의 형광으로 변형된 뉴클레오타이드를 포함시킨다. DNA 합성은 단일 뉴클레오타이드의 첨가 후 종결되고, 미포함된 뉴클레오타이드는 세척된다. 포함된 표지 뉴클레오타이드의 동일성을 결정하기 위해 이미징이 수행된다. 그 다음에, 절단 단계에서, 종결/억제기 및 형광 염료는 제거된다. CRT 방법을 사용하는 대표적인 NGS 플랫폼은, 이하에 제한되는 것은 아니지만, 전체 내부 반사 형광(total internal reflection fluorescence: TIRF)에 의해 검출된 4-색 CRT 방법과 결합된 클론으로 증폭된 주형 방법을 사용하는 일루미나(Illumina)/솔렉사(Solexa) 게놈 분석기(GA); 및 TIRF에 의해 검출된 1-색 CRT 방법과 결합된 단일-분자 주형 방법을 사용하는 헬리코스 바이오사이언스(Helicos BioSciences)/헬리스코프(HeliScope)를 포함한다.

[0065] SBL은 시퀀싱을 위해 DNA 리가제 및 1-염기-암호화된 프로브 또는 2-염기-암호화된 프로브 중 하나를 사용한다.

[0066] 전형적으로, 형광 표지된 프로브는 프라이밍된 주형에 인접한 상보적 서열에 혼성화된다. DNA 리가제는 프라이머에 염료-표지된 프로브를 결찰시키기 위해 사용된다. 비-결찰 프로브가 세척된 후 결찰된 프로브의 동일성을 결정하기 위하여 형광 이미징이 수행된다. 형광 염료는 후속의 결찰 주기를 위해 5'-PO4 기를 재생하는 절단가능한 프로브를 사용하여 제거될 수 있다. 대안적으로, 새로운 프라이머는 오래된 프라이머가 제거된 후 주형에 혼성화될 수 있다. 대표적인 SBL 플랫폼은, 이하에 제한되는 것은 아니지만, 라이프(Life)/APG/SOLiD(지지체 올리고뉴클레오타이드 결찰 검출)를 포함하는데, 이는 2-염기-암호화된 프로브를 사용한다.

[0067] 파이로시퀀싱 방법은 다른 화학발광 효소로 DNA 폴리머라제의 활성을 검출하는 단계를 기반으로 한다. 전형적으로, 해당 방법은 한 번에 하나의 염기쌍을 따라 상보적 가닥을 합성하고, 각 단계에서 실제로 첨가된 염기를 검출함으로써 DNA의 단일 가닥을 시퀀싱시킨다. 주형 DNA는 고정적이며, A, C, G 및 T 뉴클레오타이드의 용액은 순차적으로 첨가되고, 반응으로부터 제거된다. 빛은 단지 뉴클레오타이드 용액이 주형의 짝지어지지 않은 염기를 보충할 때에만 생성된다. 화학발광 신호를 생성하는 용액의 서열은 주형의 서열을 결정하게 한다. 대표적인 파이로시퀀싱 플랫폼은, 이하에 제한되는 것은 아니지만, PTP 웰에 증착된 백만 내지 2백만개의 비드에 의한 emPCR에 의해 제조된 DNA 주형을 사용하는 로슈(Roche)/454를 포함한다.

[0068] 실시간 시퀀싱은 DNA 합성 동안 염료-표지된 뉴클레오타이드의 연속적 포함을 이미징하는 단계를 수반한다. 대표적인 실시간 시퀀싱 플랫폼은, 이하에 제한되는 것은 아니지만, 포스페이트 연결된 뉴클레오타이드가 성장되

는 프라이머 가닥에 포함될 때 서열 정보를 얻기 위한 개개의 0-모드 웨이브가이드(zero-mode waveguide, ZMW) 검출기의 표면에 부착된 DNA 폴리머라제 분자를 사용하는 퍼시픽 바이오사이언스 플랫폼(Pacific Biosciences); 형광 공명 에너지 전달(fluorescence resonance energy transfer, FRET)에 의한 뉴클레오타이드 포함 후 향상된 신호를 만들기 위해 부착된 형광 염료와 함께 유전자 조작된 DNA 폴리머라제를 사용하는 라이프(Life)/비시겐(VisiGen) 플랫폼; 및 시퀀싱 반응에서 염료-퀀처 뉴클레오타이드를 사용하는 LI-COR 바이오사이언스(Biosciences) 플랫폼을 포함한다.

[0069] NGS의 다른 시퀀싱 방법은, 이하에 제한되는 것은 아니지만, 나노포어 시퀀싱, 혼성화에 의한 시퀀싱, 나노-트랜지스터 어레이 기반 시퀀싱, 폴로니(polony) 시퀀싱, 주사형전자 터널링 현미경(scanning tunneling microscopy, STM) 기반 시퀀싱 및 나노와이어-분자 센서 기반 시퀀싱을 포함한다.

[0070] 나노포어 시퀀싱은 단일-핵산 폴리머에서 분석될 수 있는 고도로 밀폐된 공간을 제공하는 나노-규모 포어를 통해서 용액 중의 핵산 분자의 전기영동을 수반한다. 나노포어 시퀀싱의 대표적인 방법은, 예를 들어 문헌(Branton D. et al., Nat Biotechnol. 2008; 26(10):1146-53]에 기재된다.

[0071] 혼성화에 의한 시퀀싱은 DNA 마이크로어레이를 사용하는 비-효소적 방법이다. 전형적으로, DNA의 단일 풀은 형광으로 표지되며, 공지된 서열을 함유하는 어레이에 혼성화된다. 어레이 상의 주어진 스팟으로부터 혼성화 신호는 DNA 서열을 확인할 수 있다. DNA 이중-가닥에서 DNA 중 한 가닥의 그것의 상보적 가닥에 결합은 혼성체 영역이 짧거나 또는 구체된 미스매치 검출 단백질이 존재할 때, 단일-염기 미스매치에 대해서 조차도 민감하다. 혼성화에 의한 시퀀싱의 대표적인 방법은, 예를 들어 문헌(Hanna G.J. et al., J. Clin. Microbiol. 2000; 38(7): 2715-21; 및 Edwards J.R. et al., Mut. Res. 2005; 573(1-2): 3-12)에 기재된다.

[0072] 폴로니 시퀀싱은 폴로니 증폭 및 다중 단일-염기-연장(FISSEQ)을 통해 시퀀싱에 따르는 것을 기반으로 한다. 폴로니 증폭은 폴리아크릴아마이드 필름 상에서 인시츄로 DNA를 증폭시키는 방법이다. 대표적인 폴로니 시퀀싱 방법은, 예를 들어 미국특허 출원 공개 제2007/0087362호에 기재된다.

[0073] 탄소나노튜브 전계 효과 트랜지스터(Carbon NanoTube Field Effect Transistor: CNTFET)와 같은 나노-트랜지스터 어레이 기반 장치가 또한 NGS를 위해 사용될 수 있다. 예를 들어, DNA 분자는 신장되고, 마이크로-제작된 전극에 의해 나노튜브에 걸쳐 구동된다. DNA 분자는 탄소 나노튜브 표면과 순차적으로 접촉하게 되고, DNA 분자와 나노튜브 사이의 전하 전달에 기인하여 각 염기로부터의 전류 흐름의 차이가 만들어진다. DNA는 이들 차이를 기록함으로써 시퀀싱된다. 대표적인 나노-트랜지스터 어레이 기반 시퀀싱 방법은, 예를 들어 미국특허 공개 제2006/0246497호에 기재된다.

[0074] 주사형전자 터널링 현미경(STM)은 또한 NGS를 위해 사용될 수 있다. STM은 표본의 래스터 주사(raster scan)를 수행하는 피에조-전자-제어 프로브를 사용하여 그것 표면의 이미지를 형성한다. STM은, 예를 들어 작동기-구동 가요성 갭과 주사형전자 터널링 현미경을 통합시킴으로써 일관된 전자 터널링 이미징 및 분광학을 만드는 단일 DNA 분자의 물리적 특성을 이미징하기 위해 사용될 수 있다. STM을 사용하는 대표적인 시퀀싱 방법은, 예를 들어 미국특허출원 공개 제2007/0194225호에 기재된다.

[0075] 나노와이어-분자 센서로 구성된 분자-분석 장치가 또한 NGS를 위해 사용될 수 있다. 이러한 장치는 DNA와 같은 나노와이어 및 핵산 분자에 배치된 질소성 물질의 상호작용을 검출할 수 있다. 분자 가이드는 상호작용 및 후속하는 검출을 허용하기 위해 분자 센서 근처의 분자를 가이딩하기 위해 배치된다. 나노와이어-분자 센서를 사용하는 대표적인 시퀀싱 방법은 예를 들어 미국특허 출원 공개 제2006/0275779호에 기재된다.

[0076] 이중 말단의 시퀀싱 방법이 NGS를 위해 사용될 수 있다. 이중 말단 시퀀싱은 DNA의 센스와 안티센스 가닥 둘 다를 시퀀싱하기 위해 차단 및 미차단 프라이머를 사용한다. 전형적으로, 이들 방법은 핵산의 제1 가닥에 미차단 프라이머를 어닐링시키는 단계; 핵산의 제2 가닥에 제2의 차단 프라이머를 어닐링 시키는 단계; 폴리머라제로 제1 가닥을 따라 핵산을 연장시키는 단계; 제1 시퀀싱 프라이머를 종결시키는 단계; 제2 프라이머를 차단해제(deblocking)하는 단계; 및 제2 가닥을 따라 핵산을 연장시키는 단계를 포함한다. 대표적인 이중 가닥 시퀀싱 방법은, 예를 들어 미국특허 제7,244,567호에 기재된다.

[0077] 데이터 분석 단계.

[0078] NGS 리드가 만들어진 후, 그것들은 공지된 기준 서열에 대해 정렬되거나 데노보 조립된다.

[0079] 예를 들어, 샘플(예를 들어, 종양 샘플)에서 단일-뉴클레오타이드 다형성 및 구조적 변이체와 같은 유전적 변형을 확인하는 것은 기준 서열(예를 들어, 야생형 서열)에 대해 NGS 리드를 정렬함으로써 수행될 수 있다. NGS에

대한 서열 정렬방법은, 예를 들어 문헌(Trapnell C. and Salzberg S.L. Nature Biotech., 2009, 27:455-457)에 기재된다.

[0080] 드노보 조립체의 예는, 예를 들어 문헌(Warren R. et al., Bioinformatics, 2007, 23:500-501; Butler J. et al., Genome Res., 2008, 18:810-820; 및 Zerbino D.R. 및 Birney E., Genome Res., 2008, 18:821-829)에 기재된다.

[0081] 서열 정렬 또는 어셈블리는 하나 이상의 NGS 플랫폼으로부터의 리드 데이터를 사용하여, 예를 들어 로슈(Roche)/454 및 일루미나(Illumina)/솔렉사(Solexa) 리드 데이터를 혼합하여 수행될 수 있다.

[0082] 본 발명에 있어서, 상기 정렬단계는 이에 제한되지는 않으나, BWA 알고리즘 및 hg19 서열을 이용하여 수행되는 것일 수 있다.

[0084] 본 발명에 있어서, 상기 서열 정렬은 컴퓨터 알고리즘으로서 게놈에서 리드 서열(예를 들어, 차세대 시퀀싱으로부터의, 예를 들어 짧은-리드 서열)이 대부분 리드 서열과 기준 서열 사이의 유사성을 평가함으로써 유래될 가능성이 있는 경우로부터 동일성에 대해 사용되는 컴퓨터적 방법 또는 접근을 포함한다. 서열 정렬 문제에 다양한 알고리즘이 적용될 수 있다. 일부 알고리즘은 상대적으로 느리지만, 상대적으로 높은 특이성을 허용한다. 이들은, 예를 들어 역동적 프로그래밍-기반 알고리즘을 포함한다. 역동적 프로그래밍은 그것들이 더 간단한 단계로 나누어짐으로써 복잡한 문제를 해결하는 방법이다. 다른 접근은 상대적으로 더 효율적이지만, 전형적으로 철저하지 않다. 이는, 예를 들어 대량 데이터베이스 검색을 위해 설계된 휴리스틱(heuristic) 알고리즘 및 확률적(probabilistic) 방법을 포함한다.

[0085] 전형적으로, 정렬 과정에 두 단계가 있을 수 있다: 후보자 검사 및 서열 정렬. 후보자 검사는 가능한 정렬 위치의 더 짧은 열거에 대해 전체 게놈으로부터 서열 정렬을 위한 검색 공간을 감소시킨다. 용어가 시사하는 바와 같이 서열 정렬은 후보자 검사 단계에 제공된 서열을 갖는 서열을 정렬시키는 단계를 포함한다. 이는 광역 정렬(예를 들어, 니들만-분쉬(Needleman-Wunsch) 정렬) 또는 국소 정렬(예를 들어, 스미스-워터만 정렬)을 사용하여 수행될 수 있다.

[0086] 대부분의 속성 정렬 알고리즘은 색인 방법에 기반한 3가지 유형 중 하나를 특징으로 할 수 있다: 해쉬 테이블(예를 들어, BLAST, ELAND, SOAP), 접미사트리(예를 들어, Bowtie, BWA) 및 병합 정렬(예를 들어, 슬라이더(Slider))에 기반한 알고리즘. 짧은 리드 서열은 정렬을 위해 전형적으로 사용된다. 짧은-리드 서열에 대한 서열 정렬 알고리즘/프로그램의 예는, 이하에 제한되는 것은 아니지만, BFAST (Homer N. et al., PLoS One. 2009;4(11):e7767), BLASTN(월드 와이드 웹상의 blast.ncbi.nlm.nih.gov에서), BLAT(Kent W.J. Genome Res. 2002;12(4):656-64), 보타이(Bowtie)(Langmead B. et al., Genome Biol. 2009;10(3):R25), BWA(Li H. and Durbin R. Bioinformatics, 2009, 25:1754-60), BWA-SW(Li H. and Durbin R. Bioinformatics, 2010;26(5):589-95), 클라우드버스트(CloudBurst)(Schatz M.C. Bioinformatics. 2009;25(11):1363-9), 코로나 라이트(Corona Lite)(Applied Biosystems, Carlsbad, California, USA), CASHX(Fahlgren N. et al., RNA, 2009; 15, 992-1002), CUDA-EC (Shi H. et al., J Comput Biol. 2010;17(4):603-15), ELAND(월드 와이드 웹상의 bioit.dbi.udel.edu/howto/eland에서), GNUMAP(Clement N.L. et al., Bioinformatics. 2010;26(1):38-45), GMAP(Wu T.D. and Watanabe C.K. Bioinformatics. 2005;21(9):1859-75), GSNAP(Wu T.D. and Nacu S., Bioinformatics. 2010;26(7):873-81), 제니오스 어셈블러(Geneious Assembler)(뉴질랜드 오클랜드에 소재한 Biomatters Ltd.), LAST, MAQ(Li H. et al., Genome Res. 2008;18(11):1851-8), Mega-BLAST(월드 와이드 웹상의 ncbi.nlm.nih.gov/blast/megablast.shtml에서), MOM(Eaves H.L. and Gao Y. Bioinformatics. 2009;25(7):969-70), MOSAIK(월드 와이드 웹 상의 bioinformatics.bc.edu/marthlab/Mosaik에서), 노보얼라인(Novoalign)(월드 와이드 웹 상의 novocraft.com/main/index.php에서), 팔맵퍼(PALMapper)(월드 와이드 웹 상의 fml.tuebingen.mpg.de/raetsch/suppl/palmapper에서), PASS(Campagna D. et al., Bioinformatics. 2009;25(7):967-8), PatMaN(Prufer K. et al., Bioinformatics. 2008; 24(13):1530-1), PerM(Chen Y. et al., Bioinformatics, 2009, 25 (19): 2514-2521), ProbeMatch(Kim Y.J. et al., Bioinformatics. 2009;25(11):1424-5), QPalma(de Bona F. et al., Bioinformatics, 2008, 24(16): i174), RazerS(Weese D. et al., Genome Research, 2009, 19:1646-1654), RMAP (Smith A.D. et al., Bioinformatics. 2009;25(21):2841-2), SeqMap(Jiang H. et al. Bioinformatics. 2008;24:2395-2396.), Shrec(Salmela L., Bioinformatics. 2010;26(10):1284-90), SHRiMP(Rumble S.M. et al., PLoS Comput. Biol., 2009, 5(5):e1000386), SLIDER(Malhis N. et al., Bioinformatics, 2009, 25 (1): 6-13), 슬림 서치(SLIM Search)(Muller T. et al.,

Bioinformatics. 2001;17 Suppl 1:S182-9), SOAP(Li R. et al., Bioinformatics. 2008;24(5):713-4), SOAP2(Li R. et al., Bioinformatics. 2009;25(15):1966-7), SOCS(Ondov B.D. et al., Bioinformatics, 2008; 24(23):2776-7), SSAHA(Ning Z. et al., Genome Res. 2001;11(10):1725-9), SSAHA2(Ning Z. et al., Genome Res. 2001;11(10):1725-9), 스탬피(Stampy)(Lunter G. and Goodson M. Genome Res. 2010, epub ahead of print), 타이판(Taipan)(월드 와이드 웹 상의 taipan.sourceforge.net에서), UGENE(월드 와이드 웹 상의 ugene.unipro.ru에서), XpressAlign(월드 와이드 웹 상의 bcgsc.ca/platform/bioinfo/software/XpressAlign에서), 및 ZOOM(캐나다 온타리오주 워터루에 소재한 바이오인포매틱스 솔루션 인코포레이티드(Bioinformatics Solutions Inc.))을 포함한다.

[0087] 서열 정렬 알고리즘은, 예를 들어 시퀀싱 기법, 리드 길이, 리드 수, 입수가능한 컴퓨팅 자료 및 민감성/스코어링 필요조건을 포함하는 다수의 인자에 기반하여 선택될 수 있다. 상이한 서열 정렬 알고리즘은 상이한 속도 수준, 정렬 민감성 및 정렬 특이성을 달성할 수 있다. 정렬 특이성은 예측된 정렬과 비교하여 정확하게 정렬된 전형적으로 서브미션에서 발견되는 바와 같이 정렬된 표적 서열 잔기의 백분율을 지칭한다. 정렬 민감성은 또한 서브미션에서 정확하게 정렬된 보통 예측된 정렬에서 발견되는 바와 같이 정렬된 표적 서열 잔기의 백분율을 지칭한다.

[0088] 정렬 알고리즘, 예컨대 ELAND 또는 SOAP는 속도가 고려되는 제1 인자일 때 기준 게놈에 대해 짧은 리드(예를 들어, 일루미나(Illumina)/솔렉사(Solexa) 시퀀서제)을 정렬하는 목적으로 사용될 수 있다. BLAST 또는 Mega-BLAST와 같은 정렬 알고리즘은 특이성이 가장 중요한 인자일 때, 이들 방법이 상대적으로 더 느리지만, 짧은 판독(예를 들어, 로슈(Roche) FLX제)을 사용하여 유사성 조사의 목적을 위해 사용될 수 있다. MAQ 또는 노보얼라인(Novoalign)와 같은 정렬 알고리즘은 품질 스코어를 고려하며, 따라서 정확성이 본질을 가질 때 단일- 또는 짝지어진-말단 데이터에 대해 사용될 수 있다(예를 들어, 고속-대량 SNP 검색에서). 보타이(Bowtie) 또는 BWA와 같은 정렬 알고리즘은 버로우즈-휠러 변환(Burrows-Wheeler Transform: BWT)을 사용하며, 따라서 상대적으로 작은 메모리 풋프린트(memory footprint)를 필요로 한다. BFAST, PerM, SHRiMP, SOCS 또는 ZOOM과 같은 정렬 알고리즘은 색공간 리드를 맵핑하며, 따라서 ABI의 SOLiD 플랫폼과 함께 사용될 수 있다. 일부 적용에서, 2 이상의 정렬 알고리즘으로부터의 결과가 조합될 수 있다.

[0090] 본 발명에 있어서, 상기 (b) 단계의 서열정보(reads)의 길이는, 5 내지 5000 bp이고, 사용하는 서열정보의 수는 5천 내지 500만개가 될 수 있으나, 이에 한정되는 것은 아니다.

[0092] 본 발명에 있어서, 상기 리드는 페어드 엔드(paired-end) 시퀀싱으로 수득하는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0093] 본 발명에 있어서, 상기 (c) 단계의 핵산단편의 길이는 핵산단편의 양 말단에 정렬되는 리드의 정렬 위치를 통해 산출하는 것을 특징으로 할 수 있다.

[0094] 즉, 도 2에 기재된 바와 같이, 양 말단 끝의 유전적 위치 정보를 이용해 세포유리핵산의 길이를 추론 할 수 있다. 만약 5` 리드의 위치가 chr1:12001-12050 이고, 반대쪽 말단에서 부터 생산된 리드의 위치가 chr1:12112:12161 이라면, 이 세포유리 핵산의 길이는 12161-12001+1로 계산하여, 161bp이다.

[0096] Paired-End(PE) 모드로 생산된 세포유리핵산 데이터는 양 말단 끝으로부터 특정 base 만큼 존재하게 되는데, 예를 들어 50 base PE 모드로 생산된 데이터에는 세포유리 핵산의 양 말단 끝으로부터 50bp씩 총 100bp 의 정보를 포함한다. 양 말단 끝의 유전적 위치 정보를 이용해 세포유리핵산의 길이를 계산할 수 있다. 만약 5` 리드의 위치가 chr1:12001-12050 이고, 반대쪽 말단에서 부터 생산된 리드의 위치가 chr1:12112:12161 이라면, 이 세포유리 핵산의 길이는 161bp 로 계산된다(12161-12001+1).

[0097] 본 발명에서, 상기 (c) 단계를 수행하기에 앞서 정렬된 리드의 정렬 일치도 점수(mapping quality score)를 만족하는 리드를 따로 분류하는 단계를 추가로 포함하는 것을 특징으로 할 수 있다.

[0098] 본 발명에서 상기 정렬 일치도 점수(mapping quality score)는 원하는 기준에 따라 달라질 수 있으나, 바람직하게는 15-70점, 더욱 바람직하게는 50~70점 일 수 있고, 가장 바람직하게는 60점일 수 있다.

[0100] 본 발명에 있어서, 상기 (d) 단계는 다음의 단계를 포함하는 방법으로 수행되는 것을 특징으로 할 수 있다:

[0101] (d-i) 염색체 전체 영역 또는 특정 영역별로 핵산단편을 긴 핵산단편(long fragment) 및 짧은 핵산단편(short fragment)으로 분류하는 단계;

[0102] (d-ii) 하기 수식 1을 바탕으로 핵산단편 길이 비(Fragment ratio)를 계산하는 단계;

[0103] 수식 1: Fragment ratio(FR) = Number of short fragment group / Number of long fragment group

[0105] 본 발명에 있어서, 상기 (d-i) 단계는 기준점을 중심으로 기준점 이하 길이의 핵산단편은 짧은 핵산단편으로, 기준점 초과 길이의 핵산단편은 긴 핵산단편으로 분류하는 것을 특징으로 할 수 있다.

[0106] 본 발명에 있어서, 상기 기준점은 핵산단편을 나눌 수 있는 특정 길이이면 제한없이 사용할 수 있고, 50 내지 200bp 일 수 있으며, 바람직하게는 150 내지 170bp일 수 있고, 더욱 바람직하게는 160 내지 170bp일 수 있으며, 가장 바람직하게는 168bp인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0107] 예를 들어, 세포 유리 핵산의 경우, 일반적으로 생성되는 핵산 단편의 길이는 최소 118bp에서 최대 220bp일 수 있으며, 이들의 중간값이 168bp를 기준점으로 하여 168bp 이하인 핵산단편은 짧은 핵산단편으로, 168bp 초과인 핵산단편은 긴 핵산단편으로 분류할 수 있다.

[0108] 본 발명에서, 상기 핵산단편 길이 비는 특정 유전체 영역에 위치한 핵산 단편들 길이의 비를 나타내는 값이다. 예를 들어, 짧은 핵산단편의 기준을 100-150bp, 긴 핵산단편의 기준을 151-200bp 로 설정을 하고, 길이가 90, 104, 122, 133, 149, 161, 199, 204 인 핵산단편들이 있다고 할 경우, 짧은 핵산단편 그룹에 속하는 핵산단편은 104, 122, 133, 149 이며, 긴 핵산단편 그룹에 속하는 핵산단편은 161, 199 이다. 따라서, 짧은 핵산단편 그룹의 핵산단편 개수는 4이고, 긴 핵산단편 그룹에 속하는 핵산단편 개수는 2 이므로, 본 발명의 수식 1에 따라 계산한 핵산단편 길이 비는 4/2로 계산하여 2가 된다.

[0110] 본 발명에 있어서, 상기 (e) 단계는 다음의 단계를 포함하는 방법으로 수행되는 것을 특징으로 할 수 있다:

[0111] (e-i) 정상 샘플군에서 샘플군과 동일한 염색체 전체 영역 또는 특정 영역별로 핵산단편 길이 비를 계산하여 하기 수식 2로 상대빈도(Relative Frequency) 값을 계산하는 단계;

[0112] 수식2:

[0113] 상대빈도i = FR(핵산단편) 비i / ∑ FR(핵산단편) 비

[0114] (e-ii) 각 영역에서의 상대 빈도 값의 평균과 표준편차를 계산하는 단계;

[0115] (e-iii) 제1항의 d) 단계에서 도출한 FR 값의 상대빈도를 수식 2로 계산하여, 하기 수식 3으로 FR Z-score(FRZ)를 계산하는 단계;

[0116] 수식 3: FR Z-score $_{i\ bin}$ = (분석 샘플 상대빈도$_{i\ bin}$ - 정상인 샘플의 상대빈도 평균$_{i\ bin}$) / 정상인 샘플의 상대빈도 표준편차$_{i\ bin}$

[0117] (e-iv) 각 유전영역에 해당하는 GC값으로 LOESS regression을 수행하고, 잔차를 계산하는 단계;

[0118] (e-v) 각 유전영역 별로 GC값으로 보정된 FRZ값을 LOESS 알고리즘을 통해 정규화하는 단계; 및

[0119] (e-vi) 하기 수식 4로 FR-score를 계산하는 단계;

[0120] 수식 4: $$FR-score = \ln\{\sum_{i=all\ genomic\ position(bin)} abs(LOESS\ smoothed\ FRZ\ i)\ \}$$

[0122] 본 발명에서 용어 “잔차”는 종속변수와 독립변수와의 관계를 밝히는 통계모형에서 모형에 의하여 추정된 종속변수의 값과 실제 관찰된 종속변수 값과의 차이를 의미하는 것으로, 본 발명에서는 각 유전영역에 해당하는 GC

값으로 LOESS regression을 수행한 다음, 실제 관찰된 FRZ 값과 통계 모형으로 추정된 LOESS regression 값의 잔차를 계산하는 것을 의미한다.

[0123] 본 발명에서, 상기 생체시료가 암 환자에서 유래할 경우에는 예후예측을 위한 용도로 사용될 수 있고, 일반 환자에서 유래할 경우에는 암 진단을 위한 용도로 사용될 수 있으나, 이에 한정되는 것은 아니다.

[0124] 본 발명에 있어서, 상기 FR-score의 기준값은 5 내지 50인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0125] 본 발명에 있어서, 상기 (e) 단계의 FR-score를 이용하여 예후를 예측하는단계는 상기 FR-score가 기준 값 또는 범위 미만일 경우에는 예후가 나쁠 것으로 예측하고, FR-score가 기준 값 또는 범위 초과일 경우에는 예후가 좋을 것으로 예측하는 것을 특징으로 할 수 있다.

[0127] 본 발명에서 상기 염색체 전체 영역 또는 특정 유전 영역은 인간 핵산 서열의 집합이면 제한없이 이용가능하나, 바람직하게는 염색체 단위 또는 일부 염색체의 특정 영역일 수 있으며, 예를 들어, 수적 이상 여부 검출을 위한 특정 영역에는 정배수체로 생각되는 상염색체가 될 수 있고, 구조적 이상 여부 검출을 위한 특정 영역에는 고유성이 떨어지는 영역(centromere, telomere)을 제외한 모든 유전적 영역이 될 수 있으나, 이에 한정되는 것은 아니다.

[0129] 본 발명은 다른 관점에서, 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부;

[0130] 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 및

[0131] 선별된 서열정보(reads)를 기반으로 핵산단편의 길이를 계산하고, 이를 기반으로 핵산단편 길이 비를 측정한 다음, 정상 샘플군과 비교하여 FR-score를 계산하고, 계산한 FR-score를 기반으로 염색체 전체 영역 또는 특정 유전 영역 별로 FR-score가 기준 값 또는 구간 미만 또는 초과 일 경우, 암이 있는 것으로 판정하거나 예후를 예측하는 암 진단 또는 예후예측부를 포함하는 암 진단 또는 예후예측 장치에 관한 것이다.

[0133] 본 발명에서, 상기 해독부는 독립된 장치에서 추출된 핵산을 주입하는 핵산 주입부; 및 주입된 핵산의 서열정보를 분석하는 서열정보 분석부를 포함할 수 있으며, 바람직하게는 NGS 분석 장치일 수 있으나, 이에 한정되는 것은 아니다.

[0134] 본 발명에서, 상기 해독부는 독립된 장치에서 생성된 서열정보 데이터를 수신하여 해독하는 것을 특징으로 할 수 있다.

[0136] 본 발명은 또 다른 관점에서, 컴퓨터 판독 가능한 저장 매체로서, 암 진단 또는 예후예측을 위한 정보를 제공하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,

[0137] (a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;

[0138] (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

[0139] (c) 상기 정렬된 서열정보(reads)에 대하여 핵산단편의 길이(Fragment length)를 계산하는 단계;

[0140] (d) 상기 c) 단계에서 계산한 핵산단편의 길이를 기반으로 염색체 전체 영역 또는 특정 영역 별로 핵산단편 길이 비(Fragment ratio)를 계산하는 단계; 및

[0141] (e) 상기 길이 비를 정상 샘플군과 비교하여 FR-score를 계산하여, FR-score가 기준 값 또는 범위 미만 혹은 초과일 경우, 암이 있는 것으로 판정하거나, 예후를 예측하기 위한 정보를 제공하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체에 관한 것이다.

[0143] 다른 양태에서 본원에 따른 방법은 컴퓨터를 이용하여 구현될 수 있다. 일 구현예에서, 컴퓨터는 칩 세트에 연

결된 하나 이상의 프로세서를 포함한다. 또한 칩 세트에는 메모리, 저장 장치, 키보드, 그래픽 어댑터(Graphics Adapter), 포인팅 장치(Pointing Device) 및 네트워크 어댑터(Network Adapter) 등이 연결되어 있다. 일 구현예에서, 상기 칩 세트의 성능은 메모리 컨트롤러 허브(Memory Controller Hub) 및 I/O 컨트롤러 허브에 의하여 가능하다. 다른 구현예에서, 상기 메모리는 칩 세트 대신에 프로세서에 직접 연결되어 사용될 수 있다. 저장 장치는 하드 드라이브, CD-ROM(Compact Disk Read-Only Memory), DVD 또는 기타 메모리 장치를 포함하는 데이터를 유지할 수 있는 임의의 장치이다. 메모리는 프로세서에 의하여 사용된 데이터 및 명령에 관여한다. 상기 포인팅 디바이스는 마우스, 트랙볼 (Track Ball) 또는 다른 유형의 포인팅 디바이스일 수 있고, 키보드와 조합하여 입력 데이터를 컴퓨터 시스템으로 전송하는데 사용된다. 상기 그래픽 어댑터는 디스플레이 상에서 이미지 및 다른 정보를 나타낸다. 상기 네트워크 어댑터는 근거리 또는 장거리 통신망으로 컴퓨터 시스템과 연결된다. 본원에 사용되는 컴퓨터는 하지만 위와 같은 구성으로 제한되는 것은 아니고, 일부 구성이 없거나, 추가의 구성을 포함 할 수 있으며, 또한 저장장치영역네트워크(Storage Area Network, SAN)의 일부일 수 있으며, 본원의 컴퓨터는 본원에 따른 방법의 수행을 위한 프로그램에 모듈의 실행에 적합하도록 구성될 수 있다.

[0145] 본원에서 모듈이라 함은, 본원에 따른 기술적 사상을 수행하기 위한 하드웨어 및 상기 하드웨어를 구동하기 위한 소프트웨어의 기능적, 구조적 결합을 의미할 수 있다. 예컨대, 상기 모듈은 소정의 코드와 상기 소정의 코드가 수행되기 위한 하드웨어 리소스(Resource)의 논리적인 단위를 의미할 수 있으며, 반드시 물리적으로 연결된 코드를 의미하거나, 한 종류의 하드웨어를 의미하는 것은 아님은 본원 기술분야의 당업자에게 자명한 것이다.

[0147] **실시예**

[0148] 이하, 실시예를 통하여 본 발명을 더욱 상세히 설명하고자 한다. 이들 실시예는 오로지 본 발명을 예시하기 위한 것으로서, 본 발명의 범위가 이들 실시예에 의해 제한되는 것으로 해석되지는 않는 것은 당업계에서 통상의 지식을 가진 자에게 있어서 자명할 것이다.

[0150] **실시예 1. 혈액에서 DNA를 추출하여, 차세대 염기서열 분석 수행**

[0151] 간암환자(hepatocellular carcinoma, HCC) 70명과 정상인 109명의 혈액을 10mL씩 채취하여 EDTA Tube에 보관하였으며, 채취 후 2시간 이내에 1200g, 4℃15분의 조건으로 혈장 부분만 1차 원심분리한 다음, 1차 원심분리된 혈장을 16000g, 4℃10분의 조건으로 2차 원심분리하여 침전물을 제외한 혈장 상층액을 분리하였다. 분리된 혈장에 대해 Chemagic DNA kit (Tiangen)을 사용하여 cell-free DNA를 추출하고, MGIEasy cell-free DNA library prep set kit 를 사용하여 library preparation 과정을 수행 한 다음, DNBseq G400 장비 (MGI) 를 100 base Paired end 모드로 sequencing 하였다.

[0152] 그 결과, 샘플 당 약 196.8 million 개의 reads가 생산되는 것을 확인 하였다.

[0154] **실시예 2. 서열정보 데이터의 품질관리**

[0155] 염기서열 정보를 전처리하고, FR-score를 계산하기 전에 다음 일련의 과정을 진행하였다. 차세대염기서열분석기(NGS) 장비에서 생성된 fastq 파일을 BWA-mem 알고리즘을 사용하여 참조 염색체 Hg19 서열을 기준으로 라이브러리 서열을 정렬하였다. 라이브러리 서열의 정렬 시 오류가 발생할 확률이 있어 오류를 교정하는 두 가지 과정을 수행하였다. 우선, 중복된 라이브러리 서열에 대하여 제거 작업을 실시한 다음, BWA-mem 알고리즘에 의해 정렬된 라이브러리 서열 중 Mapping Quality Score가 60에 도달하지 못하는 서열을 제거하였다.

[0157] **실시예 3. FR-score 계산**

[0158] 3-1. 핵산단편 비(Fragment ratio, FR) 계산

[0159] 핵산단편비를 계산하기 위해서, 염색체 영역을 한정하고(bin, gene, 염색체 arm 단위), 한정된 영역에서 세포유리핵산을 그 길이에 따라 Long Fragment group, Short Fragment group 으로 나누었다. Long Fragment group 의 값은 169 < 세포유리핵산길이 < 220, Short Fragment group 은 118 < 세포유리핵산 길이 < 168 로 정의하였다.

[0160] 이후 핵산단편 비(Fragment Ratio, FR)는 수식 1로 계산하였다.

[0161] 수식 1: Fragment ratio(FR) = Number of short fragment group / Number of long fragment group

[0163] 3-2. FR-score 계산

[0164] 정상인 그룹에서, 3-1과 같은 핵산단편비(FR)를 각 유전영역(bin)의 계산하고, FR의 상대빈도(Relative Frequency) 값을 수식 2로 계산하였다.

[0165] 수식 2: 상대빈도i = FR(핵산단편) 비i / ∑ {FR(핵산단편) 비}

[0166] 각 염색체 영역에서의 상대빈도 값의 평균과 표준편차를 계산한 다음, 불안정성 여부를 확인하고자 하는 샘플 역시 3-1과 같이 각 유전영역(bin)의 FR의 상대빈도값을 구하고, 상기 I 과정에서 계산한 정상인 그룹에서 계산된 평균과 표준편차를 이용해 하기 수식 3으로 FR Z-score (FRZ) 를 계산하였다.

[0167] 수식 3: FR Z-score $_{i\ bin}$ = (분석 샘플 상대빈도$_{i\ bin}$ - 정상인 샘플의 상대빈도 평균$_{i\ bin}$) / 정상인 샘플의 상대빈도 표준편차$_{i\ bin}$

[0168] 그 뒤, 각 유전영역별(bin)로 계산된 FRZ 값과 GC값 사이의 LOESS regression line을 이용해 normalization 하였다. 그 뒤 GC 값으로 보정된 FRZ를 LOESS 알고리즘을 이용해 smoothing한 다음, 모든 유전체 위치의 LOESS 알고리즘으로 smoothing 된 값의 절대값을 모두 더하고 자연로그를 취해 하기 수식 4로 FR-score를 계산하였다(도 3, 도 4).

[0169] 수식 4: $$FR-score = \ln\{\sum_{i=all\ genomic\ position(bin)} abs(LOESS\ smoothed\ GCnormalized\ FRZ\ i)\ \}$$

[0170] 그 결과, 하기 표 1과 2와 같이 정상인 샘플군과 HCC 환자 군에서 FR-score의 분포가 차이가 나는 것을 확인할 수 있었다.

표 1

[0171]

sample	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
normal	6.78	8.38	8.83	8.73	9.18	9.90

표 2

[0172]

sample	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
HCC	7.19	9.00	10.67	10.99	13.19	15.31

[0174] 두 그룹간의 FR-score 분석 결과, 통계적으로 유의한 수준의 값이 분포하는 것을 확인하였고(P-value = $4.1 * 10^{-11}$) (도 8), ROC 분석 결과, 0.793의 AUC값을 확인하였다(도 9).

[0175] 또한, ROC 분석을 통해 얻은 민감도와 특이도의 균형을 갖춘 임계치 값도 9.9로 계산되는 것을 확인하였다(도 9).

[0177] **실시예 4. 세포유리 핵산(Fragment)을 분류하는 기준값 설정**

[0178] DELFI 논문(Cristiano S et al., Nature, Vol.570(7761), pp. 385-389, 2019) 에서 short fragment range 는 100-150bp, long fragment range 는 151-220bp 로 정의 되어 있으며, 본 발명의 실시예에서는 정상인과 HCC (Hepatocellular carcinoma) 환자의 세포유리핵산 길이 정보를 이용해 short, long range 값의 범위를 새롭게 정의 하고자 하였다.

[0179] 정상인 20명, HCC 환자 76명의 fragment 길이의 빈도를 관찰한 결과, 도 5에 기재된 바와 같이, Major peak의 경우 166bp 정상인과 HCC 환자에서 비슷하나, 150bp 주변에서 HCC 환자에서 좀 더 많은 세포유리 핵산이 존재하는 것을 확인하였다.

[0180] 각 세포유리핵산 길이(insert size)의 정상그룹과 HCC 환자 그룹의 평균값을 계산하고 그 누적 분포를 관찰한 결과, 도 6의 A와 같은 분포가 나타나는 것을 확인하였으며, 상기 과정에서 계산한 각 insert size 별 평균값의 차이를 delta로 정의하고 그 분포를 관찰한 결과, 도 6의 B 와 같은 분포가 나타나는 것을 확인하였다.

[0181] DELFI 에서 정의된 값의 범위(100,150,220) 중, 150 bp 에서 정상인과 HCC 환자의 delta 값이 가장 큰 것을 확인 할 수 있었으나, Long과 short 을 구분하는 값의 범위로는 적당하지 않은 값으로 판단하였으며, 누적 Delta 값 분석 결과, delta 값이 상승하는 값은 118bp 이며, 가장 차이를 많이 보일 것으로 예측되는 short 과 long fragment 를 나누는 값은 168bp 임을 확인하고, Short Fragment group 범위는 118~168, Long Fragment group 은 169~220 으로 설정하였다(도 7).

[0183] **실시예 5. 핵산 단편 개수에 따른 FR-score 값의 변화**

[0184] 핵산단편 개수에 따른 FR-score의 변화를 확인하기 위해, 랜덤 핵산 단편 추출 방식을 통해 down sampling을 과정을 진행 했다. Down sampling 핵산단편수는 2천만개, 3천만개, 4천만개, 5천만개, 6천만개, 7천만개를 사용했다(도 10). 간암 5명 샘플에 대해서 down sampling 한 결과, 핵산단편 수가 줄어듦에도 FR-score 값의 커다란 차이가 없었고, 모두 간암으로 판별이 가능한 수치로 분포하는 것을 확인하였다(표 3, 도 10).

표 3

[0185]

핵산단편수	AMC0089	AMC0121	AMC0126	AMC0161	AMC0286
20M	15.600	15.491	15.873	16.167	16.386
30M	15.631	15.469	15.825	16.144	16.403
40M	15.620	15.523	15.819	16.127	16.399
50M	15.606	15.478	15.838	16.133	16.371
60M	15.609	15.580	15.847	16.136	16.397
70M	15.623	15.527	15.843	16.129	16.384
all	15.612	15.524	15.833	16.128	16.381

[0187] **실시예 6. FR-score 를 이용한 식도암 환자 예후 예측**

[0188] 실시예 1,2,3 의 방법으로 식도암 환자 61 명의 FR-score를 계산하였다. 식도암 환자를 대상으로 Chemoradiotherapy (CRT) 를 시행한 뒤 수술 여부와, FR-score 분포에 따른 예후 결과를 분석했다.

[0189] FR-score 기준값을 10.31 로 설정하고, 환자의 FR-score 가 기준값 보다 높은 그룹 (도 11 A, B) 과, 낮은 그룹 (도 11 C, D) 으로 나눴다. 그리고 그룹별 Kaplan-Meier curve를 Time To Progression (TTP) (도 11 A, C), Overall Survaval (OS) (도 11 B, D) 를 수술 여부와 함께 분석 했다. FR-score, 수술 여부에 따른 TTP와 OS 분석 결과, FR-score가 높은 그룹에서 CRT 후 수술 여부에 따라 예후가 유의한 수준의 차이를 확인 했다. 즉, CRT 후 수술을 진행한 그룹이 수술을 하지 않은 그룹에 비해 더 좋은 예후를 보이는 것을 확인하였다(median TTP, 12.7 vs 3.45 months; P=0.011; OS, not reched vs. 12.9 month; P=0.02). 반면 FR-score가 낮은 그룹에서는 CRT 후 수술 여부에 따른 예후에 차이가 없는 것으로 확인되었다.

[0190] 이를 통해 식도암 환자에서 CRT와 수술 후 예후를 예측하는 biomarker 로써 FR-score를 활용할 수 있다는 것을 확인하였다.

[0192] **실시예 7. FR-score 를 이용한 간암 환자 예후 예측**

[0193] 위 실시예 방법으로 임상정보가 확인된 간암 환자 75 명의 FR-score를 계산하였다. 분석 샘플을 FR-score 기준으로 2,4,6 개의 그룹으로 나누고, Kaplan-Meier 추정 분석 분석(TTP;Time To Progression, OS;Overall

Survival)을 진행 했다.

[0194] 2개 그룹의 FR-score 기준값은 11.64, 4개 그룹의 기준값은 10.36, 11.64, 13.69, 그리고 6개 그룹의 기준값은 10.01, 10.75, 11.71, 13.15, 14.15 를 사용했다.

[0195] 2개 그룹 분석 결과, OS, TTP 결과 모두 유의미한 결과를 확인하였다. (OS p-value : 0.0001, TTP p-value : 0.03665) (도 12)

[0196] 4개 그룹 분석 결과, OS 결과 유의미한 차이를 보였지만, TTP 분석 결과는 유의미한 결과는 확인되지 않았다. (OS p-value : 0.0001, TTP p-value : 0.01964) (도 13)

[0197] 6개 그룹 분석 결과, OS 결과 유의미한 차이를 보였지만, TTP 분석 결과는 유의미한 결과는 확인되지 않았다. (OS p-value : 0.02891, TTP p-value : 0.68211) (도14)

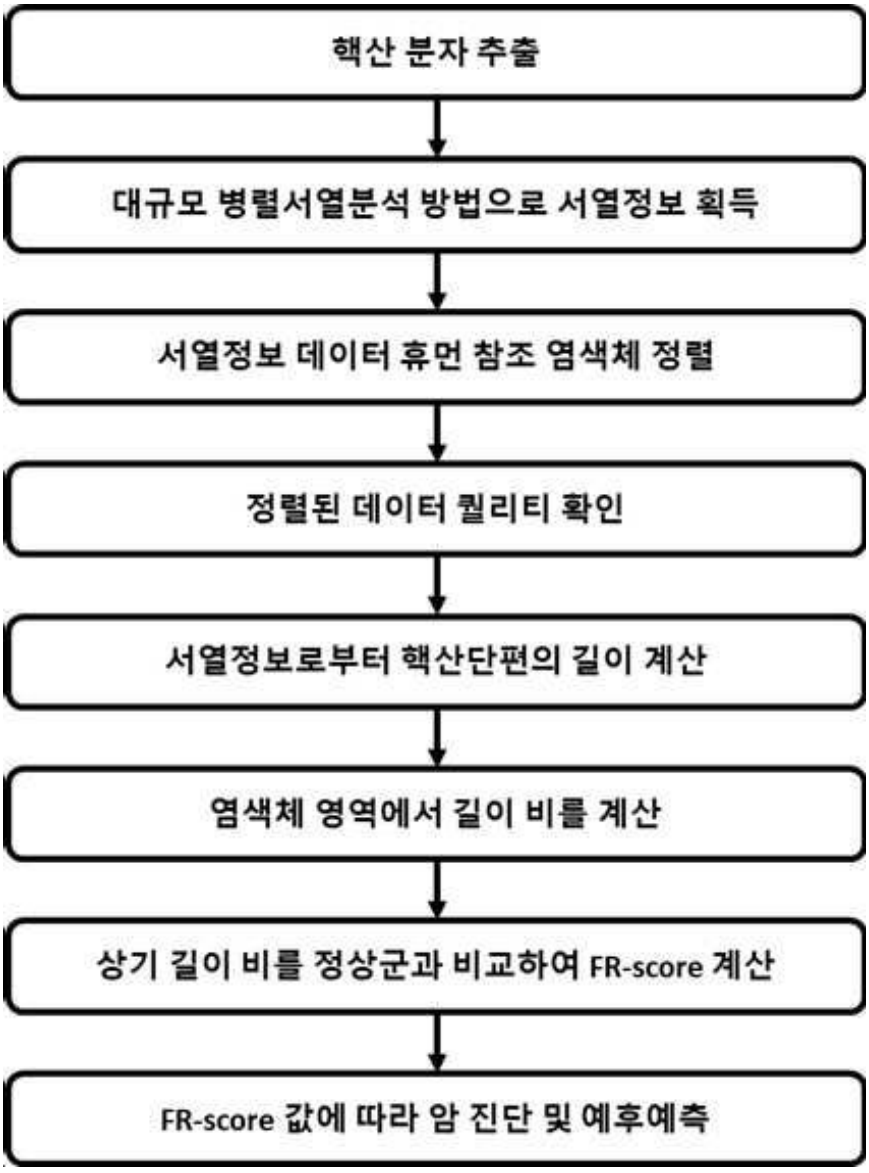
[0198] TTP 분석 결과, 2개의 그룹으로 나눴을때만 유의미한 결과를 확인했다. 반면 OS 분석 결과 2,4,6개 그룹으로 나눴을 때 모두 유의미한 결과를 확인 했다. 간암에서 FR-score 가 상대적으로 높을수록 환자의 Overall Survival이 좋지 않았다.

[0199] 이를 통해 간암 환자에서 예후를 예측하는 biomarker로써 FR-score를 활용할 수 있다는 것을 확인 했다.

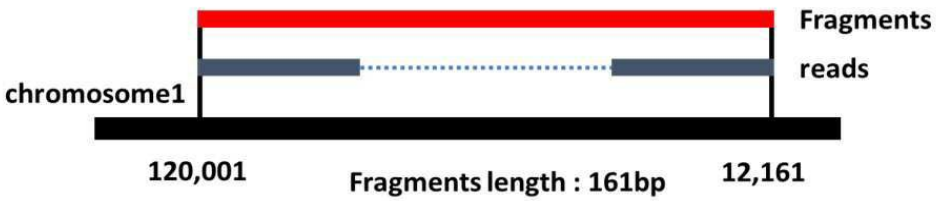
[0201] 이상으로 본 발명 내용의 특정한 부분을 상세히 기술하였는 바, 당업계의 통상의 지식을 가진 자에게 있어서 이러한 구체적 기술은 단지 바람직한 실시 양태일 뿐이며, 이에 의해 본 발명의 범위가 제한되는 것이 아닌 점은 명백할 것이다. 따라서, 본 발명의 실질적인 범위는 첨부된 청구항들과 그것들의 등가물에 의하여 정의된다고 할 것이다.

도면

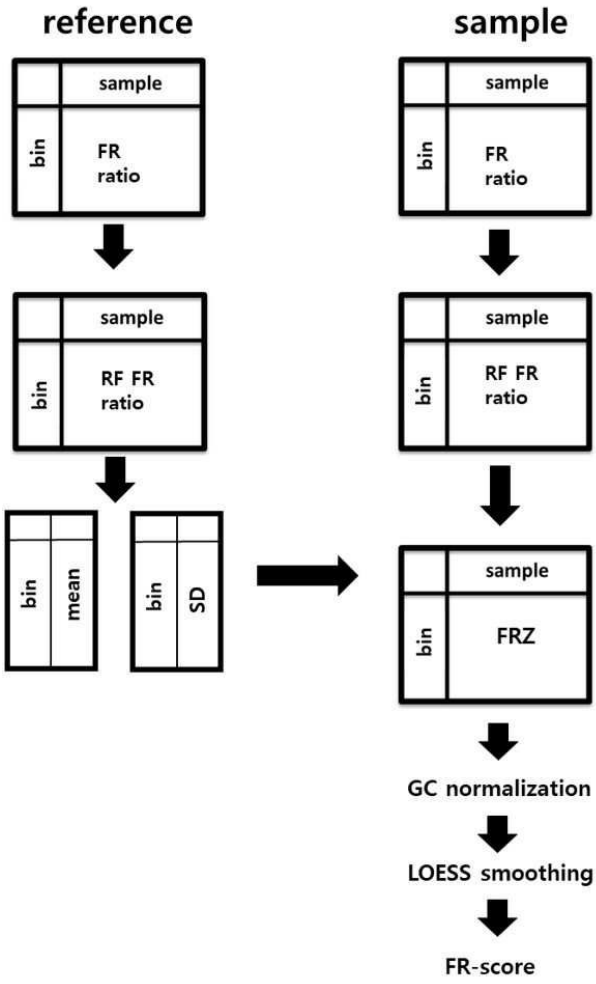
도면1


핵산 분자 추출
대규모 병렬서열분석 방법으로 서열정보 획득
서열정보 데이터 휴먼 참조 염색체 정렬
정렬된 데이터 퀄리티 확인
서열정보로부터 핵산단편의 길이 계산
염색체 영역에서 길이 비를 계산
상기 길이 비를 정상군과 비교하여 FR-score 계산
FR-score 값에 따라 암 진단 및 예후예측

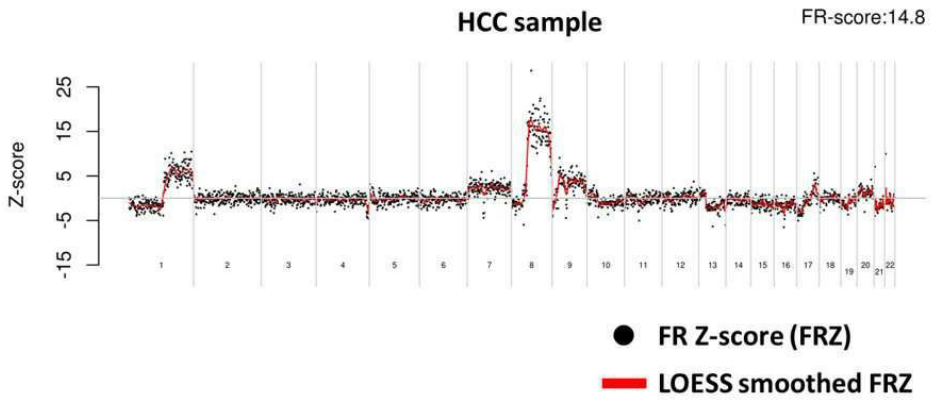
도면2


Fragments
reads
chromosome1
120,001
Fragments length : 161bp
12,161

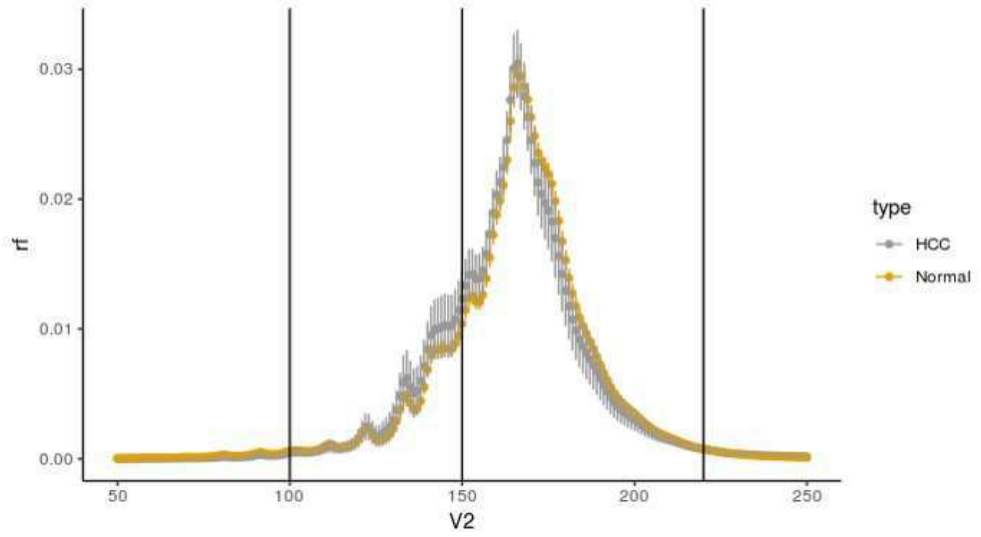
도면3


reference
sample
sample
bin
FR ratio
RF FR ratio
mean
SD
FRZ
GC normalization
LOESS smoothing
FR-score

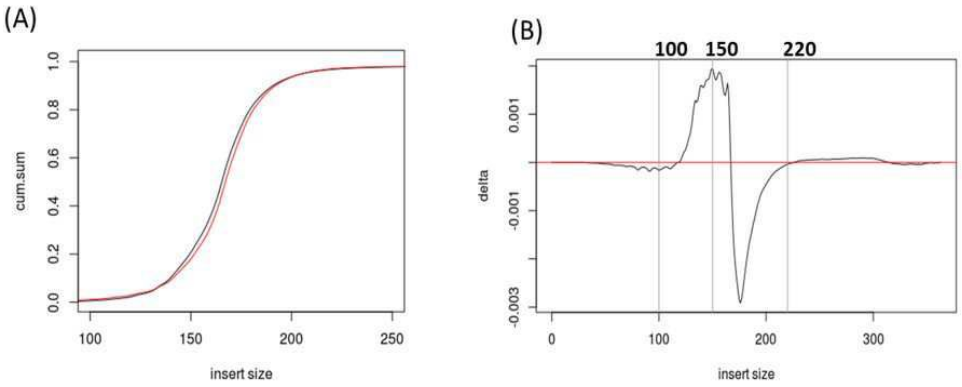
도면4


HCC sample
FR-score:14.8
Z-score
25
15
5
-5
-15
FR Z-score (FRZ)
LOESS smoothed FRZ

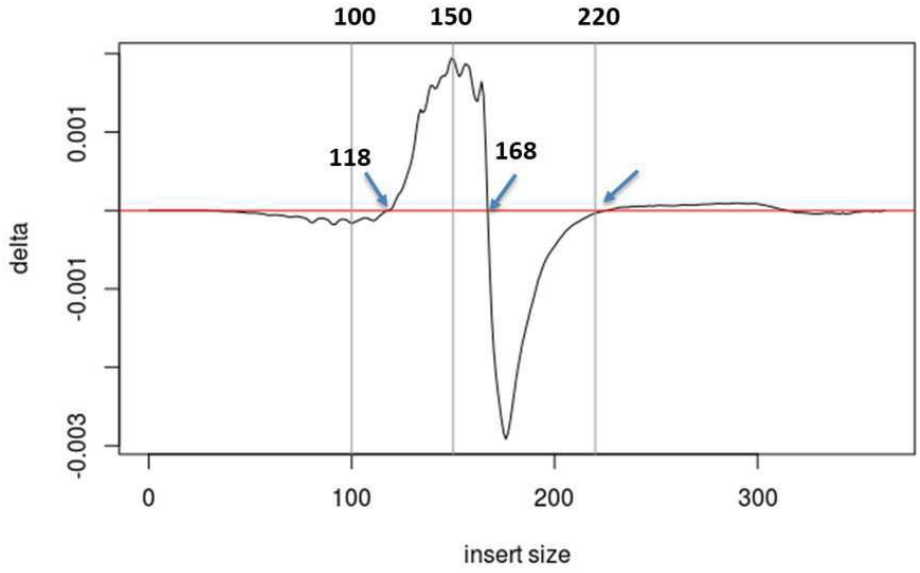
도면5


0.03
0.02
0.01
0.00
rf
50
100
150
200
250
V2
type
HCC
Normal

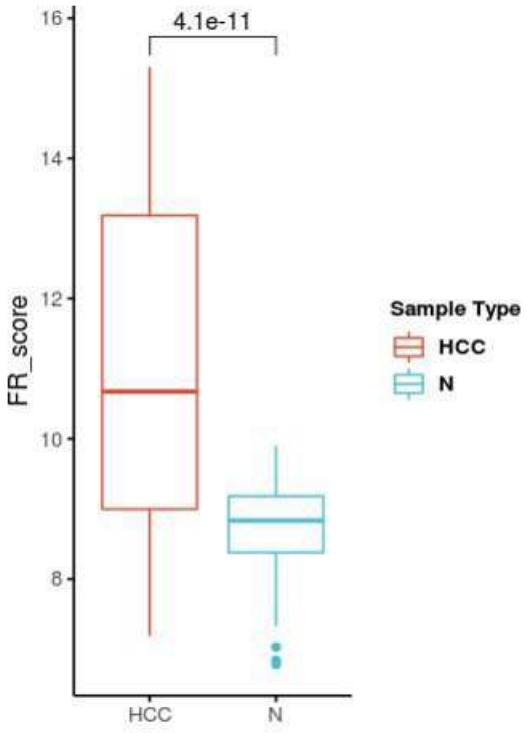
도면6


(A)
cum.sum
0.0
0.2
0.4
0.6
0.8
1.0
100
150
200
250
insert size
(B)
100
150
220
delta
0.001
-0.001
-0.003
0
100
200
300
insert size

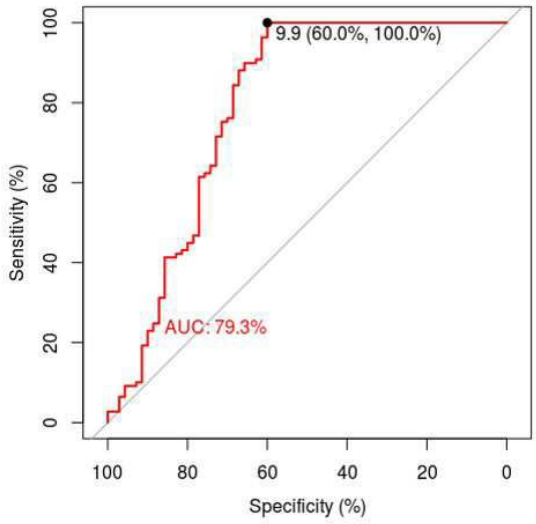
도면7


100
150
220
118
168
0.001
-0.001
-0.003
delta
0
100
200
300
insert size

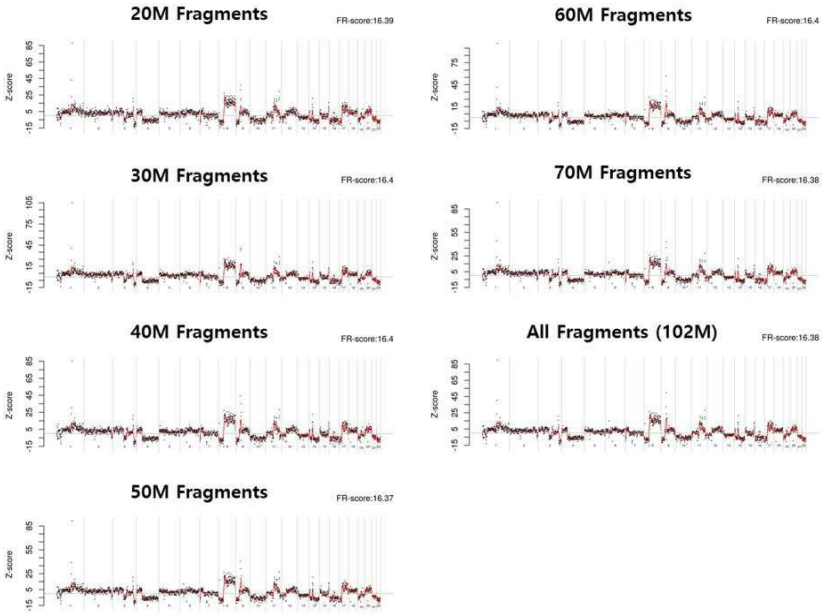
도면8


4.1e-11
16
14
12
10
8
FR_score
HCC
N
Sample Type
HCC
N

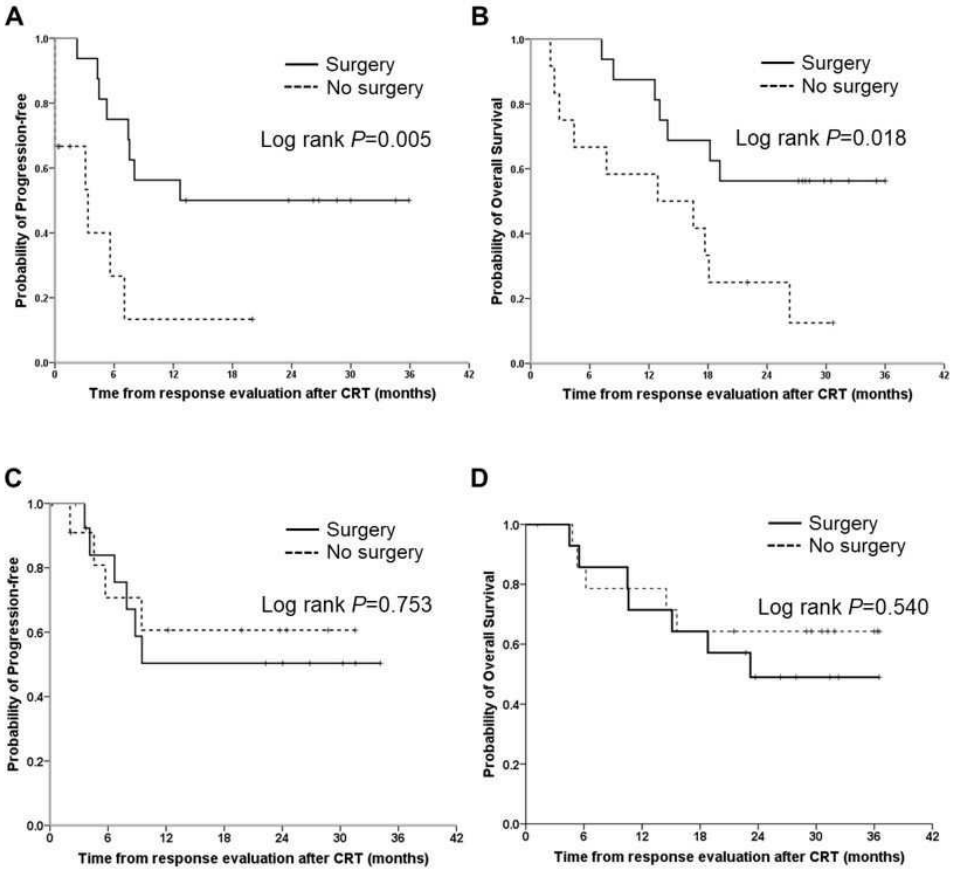
도면9


9.9 (60.0%, 100.0%)
AUC: 79.3%
Sensitivity (%)
Specificity (%)
100
80
60
40
20
0

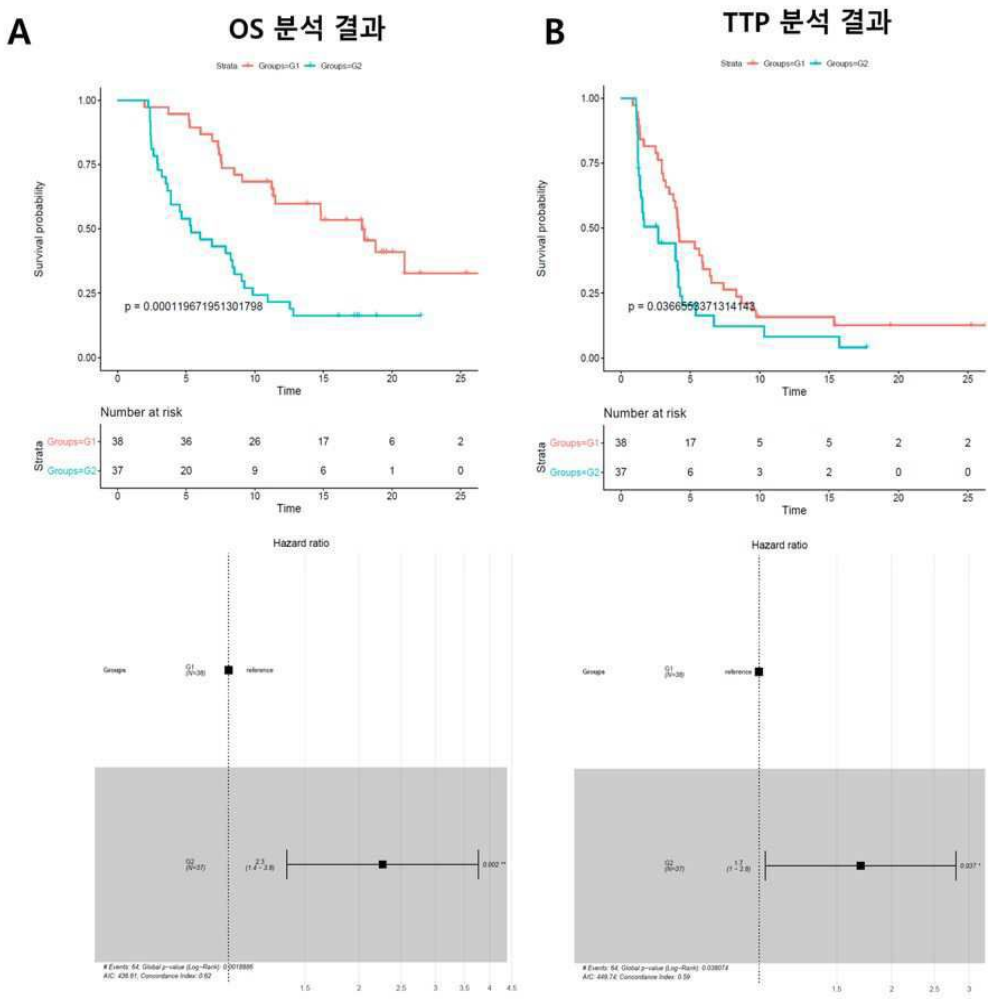
도면10


20M Fragments
60M Fragments
30M Fragments
70M Fragments
40M Fragments
All Fragments (102M)
50M Fragments
Z-score

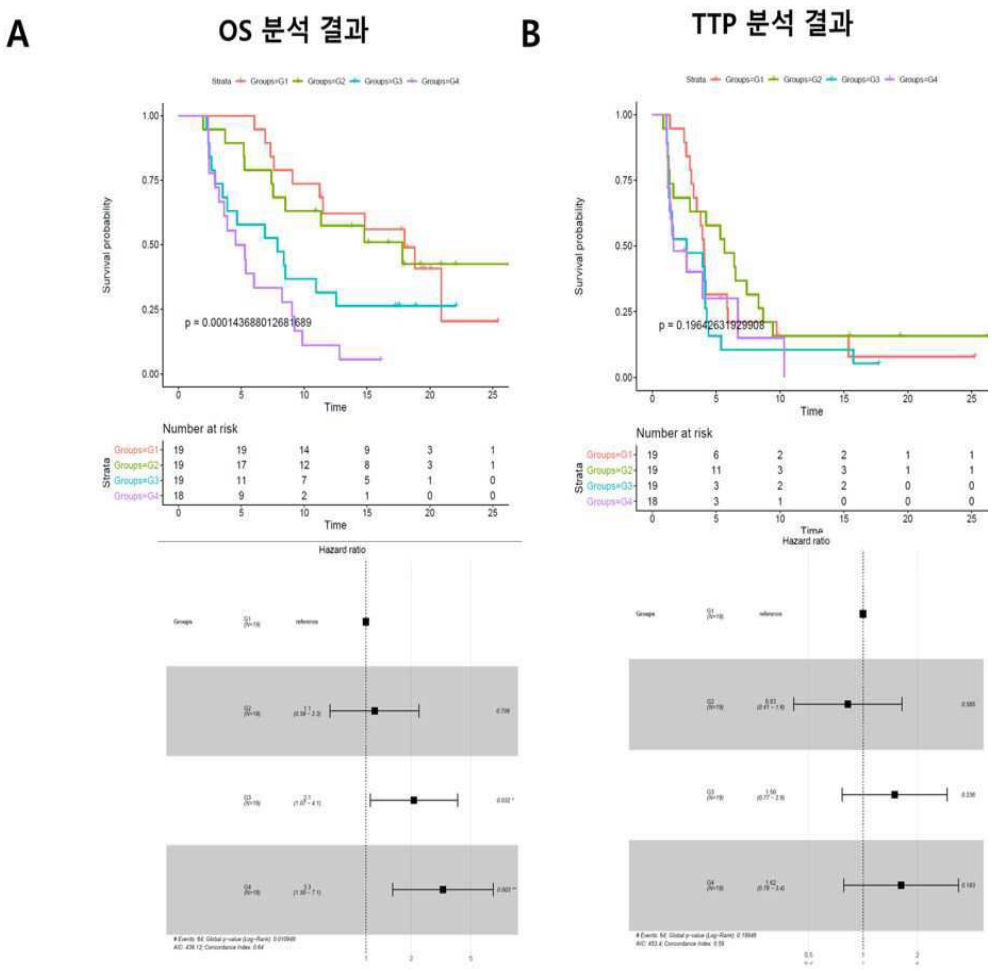
도면11


A
Probability of Progression-free
Surgery
No surgery
Log rank P=0.005
Tme from response evaluation after CRT (months)
B
Probability of Overall Survival
Surgery
No surgery
Log rank P=0.018
Time from response evaluation after CRT (months)
C
Probability of Progression-free
Surgery
No surgery
Log rank P=0.753
Time from response evaluation after CRT (months)
D
Probability of Overall Survival
Surgery
No surgery
Log rank P=0.540
Time from response evaluation after CRT (months)

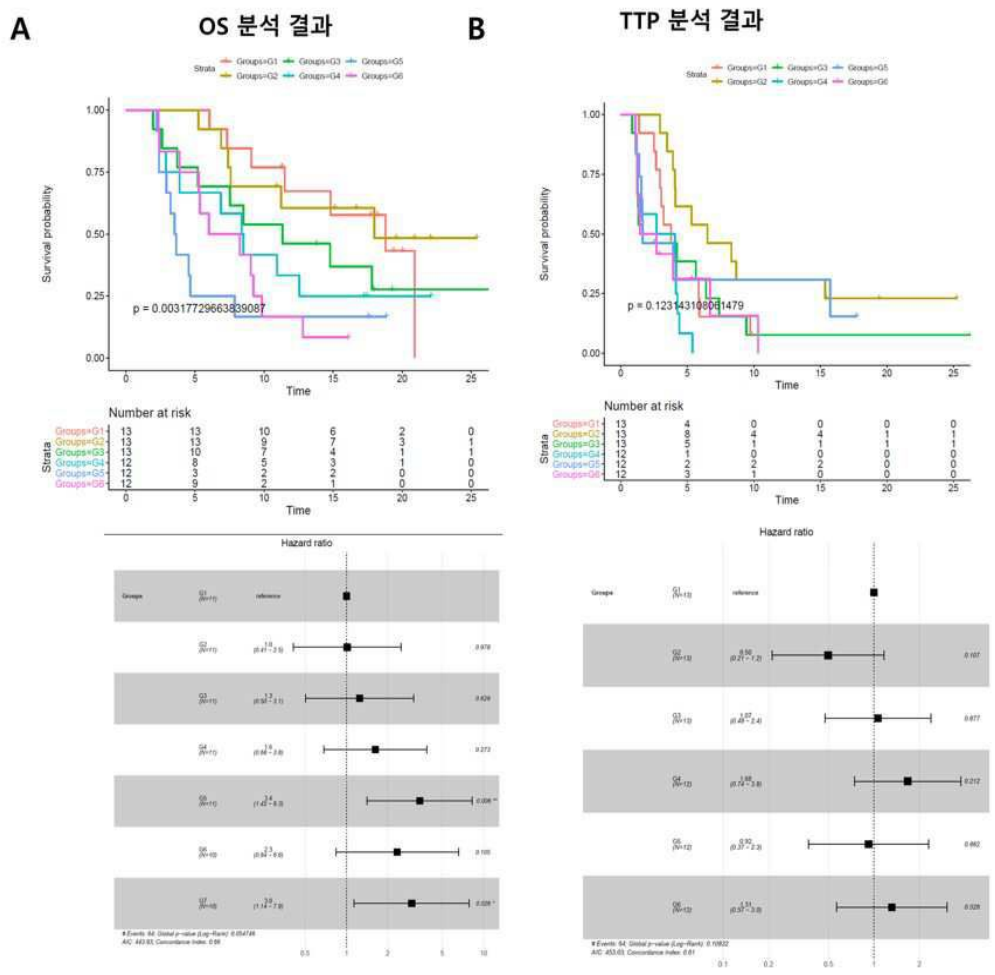
도면12


A
OS 분석 결과
Strata Groups=G1 Groups=G2
Survival probability
p = 0.000119671951301798
Time
Number at risk
Groups=G1 38 36 26 17 6 2
Groups=G2 37 20 9 6 1 0
Hazard ratio
B
TTP 분석 결과
Strata Groups=G1 Groups=G2
Survival probability
p = 0.0366559371314143
Time
Number at risk
Groups=G1 38 17 5 5 2 2
Groups=G2 37 6 3 2 0 0
Hazard ratio

도면13


A
OS 분석 결과
B
TTP 분석 결과
Survival probability
Time
p = 0.00014368801268689
p = 0.19642631929908
Number at risk
Strata
Hazard ratio

도면14


A
OS 분석 결과
B
TTP 분석 결과
Survival probability
Time
Number at risk
Strata
Hazard ratio
p = 0.0031772966338390 87
p = 0.123143108061479