(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0029001
(43) 공개일자 2022년03월08일

(51) 국제특허분류(Int. Cl.)
G16B 20/20 (2019.01) *G16B 30/10* (2019.01)
G16B 35/10 (2019.01) *G16B 40/20* (2019.01)
(52) CPC특허분류
G16B 20/20 (2019.02)
G16B 30/10 (2019.02)
(21) 출원번호 10-2020-0110773
(22) 출원일자 2020년09월01일
심사청구일자 2020년09월01일

(71) 출원인
주식회사 아이엠비디엑스
서울특별시 금천구 가산디지털1로 131 에이동 21층(가산동, 비와이씨하이시티)
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
허성훈
서울특별시 서대문구 연희로 82
이동인
서울특별시 서대문구 연희로 89-9
(뒷면에 계속)
(74) 대리인
박원미

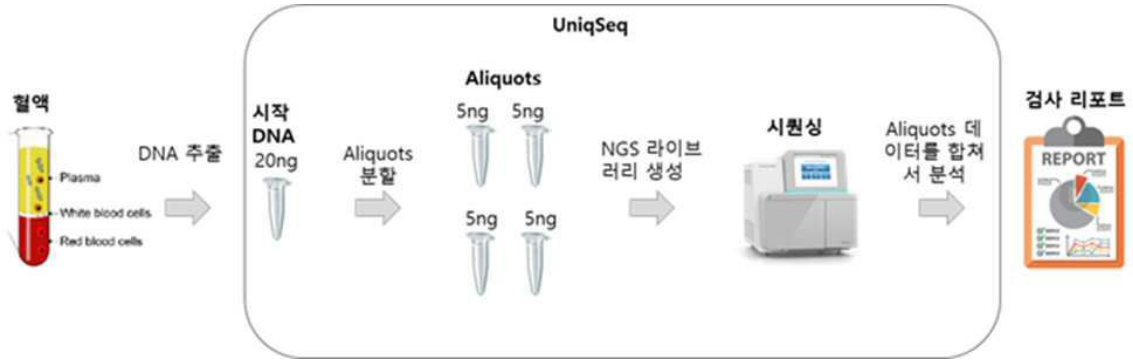
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **cfDNA의 저빈도 변이 검출을 위해 NGS 분석에 사용되는 고유 단편의 비율을 증가시키는 방법**

(57) 요 약

본원은 NGS (Next Generation Sequencing) 분석을 이용한 cfDNA (cell free DNA)의 저빈도 변이 검출에 있어서, 상기 NGS 분석에 사용되는 고유 단편의 비율을 향상시키는 방법을 개시한다. 본원에 따른 방법은 우연히 서열이 동일한 사건 의 비가 일정 이하가 되도록 시료를 2개 이상의 알리쿼트로 분할하여 알리쿼트 당 DNA 농도를 낮춰서 NGS 라이브러리를 만들고, 각 알리쿼트에서 생성된 NGS 데이터는 분석 단계에서 합쳐서 분석함으로써 대부분의 고유한 DNA 단편을 구분할 수 있어 일반적인 NGS 보다 예를 들면 cfDNA에 매우 적은 양으로 존재하는 ctDNA의 저빈도 유전자 변이의 검출이 가능하다.

대 표 도 - 도1



(52) CPC특허분류
G16B 35/10 (2019.02)
G16B 40/20 (2019.02)
(72) 발명자
방두희
서울특별시 종로구 송월길 99
노한성
서울특별시 동작구 사당로28길 55
김황필
경기도 용인시 기흥구 동백죽전대로527번길 67
문성태
서울특별시 강동구 고덕로 333

이 발명을 지원한 국가연구개발사업
과제고유번호 1711104379
과제번호 2016M3A9B6948494
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 원천기술개발사업
연구과제명 CRISPR 기반 유전자 진단기술 개발
기 여 율 3/10
과제수행기관명 연세대학교
연구기간 2020.03.01 ~ 2020.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1465030520
과제번호 HI18C2282030020
부처명 보건복지부
과제관리(전문)기관명 한국보건산업진흥원
연구사업명 포스트게놈신산업육성을위한다부처유전체사업(R&D)(복지부)
연구과제명 말초혈액 유래 면역세포 프로파일링을 통한 암 진단도구 개발
기 여 율 4/10
과제수행기관명 경희대학교
연구기간 2019.01.01 ~ 2019.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711104629
과제번호 2018M3A9H3024850
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 바이오.의료기술개발(R&D)
연구과제명 엔지니어드 박테리아 기반 정밀 면역 치료법 개발
기 여 율 3/10
과제수행기관명 연세대학교
연구기간 2020.03.01 ~ 2020.12.31

명 세 서

청구범위

청구항 1

NGS (Next Generation Sequencing) 분석을 이용한 cfDNA (cell free DNA)의 저빈도 변이 검출에 있어서, 상기 NGS 분석에 사용되는 고유 단편의 비율을 향상시키는 방법으로,

(a) 특정 양의 cfDNA 시료를 제공하는 단계;

(b) 상기 특정 양의 cfDNA에 포함된 유전체 단편 중 51 내지 330bp 길이에 해당하는 총 유전체 단편의 수, 및 고유 단편(unique fragment)의 수를 계산하는 단계로, 상기 고유 단편의 수는 상기 총 유전체 단편의 수에서 상기 51 내지 330bp 길이에 해당하는 각 유전체 단편의 collision count의 합을 제외한 값이고,

상기 collision count 합은 다음 식으로부터 계산되며,

$$\sum_{k=1}^{n} q(k-1;d) = n - d + d\left(\frac{d-1}{d}\right)^{n}.$$

.

상기 식에서 q(k-1;d) : [1,d]의 범위의 n 개의 숫자 중 k 와 같은 숫자가 있을 확률, k : 특정 숫자, d : 숫자의 범위, n : 숫자의 개수임.

(c) 상기 총 유전체 단편 수에서 상기 고유 단편 수가 차지하는 비를 계산하고, 상기 고유 단편의 비를 증가시키도록 상기 특정 양의 cfDNA 시료를 복수 개의 aliquot로 나누는 단계; 및

(d) 상기 복수개의 각 aliquot 별로, 각각 상이한 인덱스를 포함하는 어뎁터를 태깅하여 라이브러리를 제조하고, NGS (Next Generation Sequencing) 분석을 수행한 후 상기 각 aliquot의 NGS 결과를 통합하는 단계.

청구항 2

제 1 항에 있어서,

상기 저빈도 변이는 1% 미만의 빈도인 것인, 방법.

청구항 3

제 1 항 또는 제 2 항에 있어서,

상기 (d) 단계에서 상기 고유 단편의 비가 최소 93% 이상이 되도록 하는 것인, 방법.

청구항 4

제 1 항 또는 제 2 항에 있어서,

상기 특정 양의 cfDNA는 20ng이고, 이 경우 상기 cfDNA는 상기 고유 단편이 비가 93.9%가 되도록 상기 cfDNA를 4개의 aliquot로 분할하는 것인, 방법.

청구항 5

NGS (Next Generation Sequencing) 분석을 이용한 cfDNA (cell free DNA)의 저빈도 변이 검출에 있어서, 상기 NGS 분석에 사용되는 고유 단편의 비율을 향상시키기 위한 라이브러리 제조방법으로, 상기 방법은

(a) 특정 양의 cfDNA 시료를 제공하는 단계;

(b) 상기 특정 양의 cfDNA에 포함된 유전체 단편 중 51bp 내지 330bp 길이에 해당하는 총 유전체 단편의 수, 및 고유 단편 (unique fragment)의 수를 계산하는 단계로, 상기 고유 단편의 수는 상기 총 유전체 단편의 수에서 상기 51~330bp 길이에 해당하는 각 유전체 단편의 collision count의 합을 제외한 값이고,

상기 collision count 합은 다음 식으로부터 계산되며,

$$\sum_{k=1}^{n} q(k-1;d) = n - d + d\left(\frac{d-1}{d}\right)^{n}.$$

상기 식에서 q(k-1;d) : [1,d]의 범위의 n 개의 숫자 중 k 와 같은 숫자가 있을 확률, k : 특정 숫자, d : 숫자의 범위, n : 숫자의 개수임.

(c) 상기 단계 (b)로부터 상기 총 유전체 단편 수에서 상기 고유 단편의 수가 차지하는 비를 계산하고, 상기 고유 단편의 비를 증가시키도록 상기 특정 양의 cfDNA 시료를 복수 개의 aliquot로 나누는 단계; 및

(d) 상기 복수개의 각 aliquot 별로, 각각 상이한 인덱스를 포함하는 어댑터를 태깅하여 NGS용 라이브러리를 제조하는 단계를 포함하는, 방법.

청구항 6

제 5 항에 있어서,

상기 저빈도는 1% 미만의 빈도인 것인, 방법.

청구항 7

제 5 항 또는 제 6 항에 있어서,

상기 (d) 단계에서 상기 고유 단편의 비가 최소 93% 이상이 되도록 하는 것인, 방법.

청구항 8

제 5 항 또는 제 6 항에 있어서,

상기 특정 양의 cfDNA는 20ng이고, 이 경우 상기 cfDNA는 상기 고유 단편이 비가 93.9%가 되도록 상기 cfDNA를 4개의 aliquot로 분할하는 것인, 방법.

발명의 설명

기 술 분 야

[0001] 본원은 NGS (Next Generation Sequencing)를 이용한 cfDNA의 유전자 변이 분석 기술과 관련된 것이다.

배 경 기 술

[0003] 혈액속에 존재하는 세포유리 DNA (cell-free DNA, cfDNA)에는 건강한 사람들의 경우 조혈 세포 (haematopoietic cell)로부터 방출된 DNA가 대부분이다. 하지만 암 환자의 경우 cfDNA에는 암세포 사멸로 파괴된 세포로부터 혈액으로 방출된 순환 종양 DNA (circulating tumor DNA, ctDNA)가 포함되어 있다. 이 ctDNA는 특정 암과 관련된 유전적 변이를 포함하고 있으며, 이러한 유전적 변이의 모니터링을 통해, 병변 발생 전 암의 조기 발견, 특정 암치료법에 대한 반응 분석, 항암제에 대한 저항성 생성 기전 발견, 잔존 암의 존재 등의 확인이 가능하다.

[0004] 이러한 ctDNA의 검출을 위한 방법의 하나는 droplet digital PCR (ddPCR)로 이는 0.001%의 ctDNA까지 검사할 수 있다. 암을 유발하는 DNA 마커는 매우 다양한데, ddPCR의 경우 검사 범위가 제한적인 단점이 있다.

[0005] 다른 방법은 표적화 NGS (Targeted next-generation sequencing) 방법 (Corcoran, R. B., & Chabner, B. A. (2018). New England Journal of Medicine, 379(18), 1754-1765) 이다. 이 기술은 다수의 종양 관련 유전자의 전체 엑손 또는 특정 마커를 한 번에 검사할 수 있는 특징으로 인해 종양에 대한 유전학적 프로파일을 얻을 수 있는 장점이 있다.

[0006] 하지만, 이러한 NGS 방법에는 주로 cfDNA가 사용되는데, 이에 포함된 ctDNA의 양이 매우 제한적이라는 문제점이 있다. ctDNA는 cfDNA의 단지 <0.1 ~ 10% 양으로 포함되어 있다. 나아가, NGS의 서열분석에서 통계적으로 유의한 결과를 얻기 위해서는 에러를 고려하여 최소 10 X read depth가 필요하고, 그 결과 0.5%의 변이 수준을 검출하기 위해서는 총 2000 X depth가 필요하고, 1ng 당 330 genome equivalent인 것을 고려하면 최소 6ng의 DNA가 필요하다. NGS 분석시 실험 단계마다 정보 양이 소실되어, 최종적으로 얻을 수 있는 DNA 정보의 양 (conversion rate)은 30% 수준이다. 그러므로 NGS 분석에서 ctDNA 6ng에 해당하는 정보량을 얻기 위해서는 20ng의 DNA가 필요하나, 임상에서 NGS 검사에 이용할 수 있는 DNA는 매우 제한적이다.

[0007] 이에 더하여 최신 NGS 분석에 적용되는 molecular barcode 방식 적용시 어댑터에 의해 형성된 이량체(dimer)의 증가로 생산한 데이터 중 가용 데이터 비율의 저하, 암세포의 유전체 DNA가 잘리는 과정에서 서로 다른 세포에서 유래했지만 우연히 동일한 부위가 잘려서 NGS를 통해서 PCR로 증폭된 된 것과 구분하지 못하는 경우의 발생으로 인한 데이터 소실, 그리고 수십억개의 판독서열(reads) 중 동일한 서열을 제거하기 위해서 판독서열을 참조유전체에 맵핑(mapping)하는 과정에서 서열 품질이 가장 좋은 하나를 대표 판독서열로 삼고 나머지는 판독서열은 제외하는 중복제거 등으로 인한 가용 데이터의 소실로 인해 실제 더 많은 양의 DNA가 필요하다. 이는 ctDNA에 존재하는 암과 관련된 유전자 변이의 검출을 어렵게 만든다.

[0008] 따라서 cfDNA에 저빈도로 존재하는 유전자 변이 검출을 위해 제한된 양의 cfDNA를 이용한 NGS 검사에 있어서, 분석에 사용될 수 있는 가용 ctDNA 분자 정보의 양을 증가시킬 수 있는 방법의 개발이 필요하다.

발명의 내용

해결하려는 과제

[0010] 본원은 저빈도 유전자 변이 검출을 위해 제한된 양의 cfDNA를 이용한 NGS 검사에 있어서, 분석에 사용될 수 있는 가용 ctDNA 단편 정보의 양을 증가시킬 수 있는 방법을 제공하고자 한다.

과제의 해결 수단

[0012] 한 양태에서 본원은 NGS 분석을 이용한 cfDNA (cell free DNA)의 저빈도 변이 검출에 있어서, 상기 NGS 분석에 사용되는 고유 단편의 비율을 향상시키는 방법을 제공한다.

[0013] 일 구현예에서 상기 방법은 (a) 특정 양의 cfDNA 시료를 제공하는 단계; (b) 상기 특정 양의 cfDNA에 포함된 유전체 단편 중 51 내지 330bp 길이에 해당하는 총 유전체 단편의 수, 및 고유 단편(unique fragment)의 수를 계산하는 단계로, 상기 고유 단편의 수는 상기 총 유전체 단편의 수에서 상기 51 내지 330bp 길이에 해당하는 각 유전체 단편의 collision count의 합을 제외한 값이고, 상기 collision count 합은 본원에 개시된 [식 1]로부터 계산되고, (c) 상기 총 유전체 단편 수에서 상기 고유 단편 수가 차지하는 비를 계산하고, 상기 고유 단편의 비를 증가시키도록 상기 특정 양의 cfDNA 시료를 복수 개의 aliquot로 나누는 단계; 및 (d) 상기 복수개의 각 aliquot 별로, 각각 상이한 인덱스를 포함하는 어뎁터를 태깅하여 라이브러리를 제조하고, NGS 분석을 수행한 후 상기 각 aliquot의 NGS 결과를 통합하는 단계를 포함한다.

[0014] 일 구현예에서 저빈도 변이 검출은 cfDNA에 포함된 ctDNA의 변이 검출을 의미한다.

[0015] 일 구현예에서 cfDNA에 포함된 DNA 유전체 단편 중에서, 약 1% 미만의 저빈도로 존재하는 암세포의 ctDNA의 유전자 변이를 검출하고자 한다.

[0016] 일 구현예에서 상기 (d) 단계에서 상기 고유 단편의 비가 최소 약 93% 이상이 되도록 또는 우연히 서열이 동일

한 사건 (collision으로 표현) 비가 일정 이하가 되도록 시료를 2개 이상의 aliquot로 분할하여 aliquot 당 DNA 농도를 낮춰서 NGS 라이브러리를 만들고, 각 aliquots에서 생성된 NGS 데이터는 분석 단계에서 합쳐서 분석한다.

[0017] 일 구현예에서 특정 양의 cfDNA는 20ng이고, 이 경우 상기 cfDNA는 상기 고유 단편이 비가 약 93.9%가 되도록 상기 cfDNA를 4개의 aliquot로 분할한다.

[0018] 다른 양태에서 본원은 NGS 분석을 이용한 cfDNA의 저빈도 변이 검출에 있어서, 상기 NGS 분석에 사용되는 고유 단편의 비율을 향상시키기 위한 라이브러리 제조방법을 제공한다.

[0019] 일 구현예에서 상기 방법은 (a) 특정 양의 cfDNA 시료를 제공하는 단계;

[0020] (b) 상기 특정 양의 cfDNA에 포함된 유전체 단편 중 51bp 내지 330bp 길이에 해당하는 총 유전체 단편의 수, 및 고유 단편(unique fragment)의 수를 계산하는 단계로, 상기 고유 단편의 수는 상기 총 유전체 단편의 수에서 상기 51~330bp 길이에 해당하는 각 유전체 단편의 collision count의 합을 제외한 값이고, 상기 collision count 합은 본원에 개시된 식 1로부터 계산되며, (c) 상기 단계 (b)로부터 상기 총 유전체 단편 수에서 상기 고유 단편의 수가 차지하는 비를 계산하고, 상기 고유 단편의 비를 증가시키도록 상기 특정 양의 cfDNA 시료를 복수 개의 aliquot로 나누는 단계; 및 (d) 상기 복수개의 각 aliquot 별로, 각각 상이한 인덱스를 포함하는 어뎁터를 태깅하여 NGS용 라이브러리를 제조하는 단계를 포함한다.

[0021] 일 구현예에서 cfDNA에 포함된 DNA 유전체 단편 중에서, 약 1% 미만의 저빈도로 존재하는 암세포의 ctDNA의 유전자 변이를 검출하고자 한다.

[0022] 일 구현예에서 상기 (d) 단계에서 상기 고유 단편의 비가 최소 약 93% 이상이 되도록 또는 우연히 서열이 동일한 사건 (collision으로 표현) 비가 일정 이하가 되도록 시료를 2개 이상의 aliquot로 분할하여 aliquot 당 DNA 농도를 낮춰서 NGS 라이브러리를 만들고, 각 aliquots에서 생성된 NGS 데이터는 분석 단계에서 합쳐서 분석한다.

[0023] 일 구현예에서 특정 양의 cfDNA는 20ng이고, 이 경우 상기 cfDNA는 상기 고유 단편이 비가 약 93.9%가 되도록 상기 cfDNA를 4개의 aliquot로 분할한다.

발명의 효과

[0025] 본원에 따른 방법은 실제 의료현장에서 혈액으로부터 얻을 수 있는 cfDNA 양이 제한적인 상황에서 가용 ctDNA 정보양을 증가시켜 read depth를 향상시켜 차세대 염기서열 분석 (NGS)을 이용한 ctDNA 검사의 성능을 높일 수 있다.

[0026] NGS 방법에 사용되는 일반적인 프로토콜에 따른 라이브러리 제조에 있어서, 암세포의 유전체 DNA가 잘리는 과정에서 서로 다른 세포에서 유래했지만 우연히 동일한 부위가 잘려서 서열이 동일한 단편이 발생하고, 이를 PCR로 증폭된 것과 구분하지 못해, NGS 분석과정에서의 데이터가 소실된다. DNA 단편의 길이가 작을수록, DNA 농도가 높을수록 우연히 동일한 DNA 단편이 발생할 확률이 높다. ctDNA의 경우 정상 DNA보다 평균적인 길이가 더 짧아서 이 사건이 더 높은 확률로 일어난다. 이로 인해 특히 cfDNA에 적은 양으로 존재하는 ctDNA에서 발견되는 유전자변이의 검출을 어렵게 한다. 하지만 본원에 따른 방법은 우연히 서열이 동일한 사건 (collision으로 표현) 비가 일정 이하가 되도록 시료를 2개 이상의 aliquot로 분할하여 aliquot 당 DNA 농도를 낮춰서 NGS 라이브러리를 만들고, 각 aliquots에서 생성된 NGS 데이터는 분석 단계에서 합쳐서 분석함으로써 대부분의 고유한 DNA 단편을 구분할 수 있어 일반적인 NGS 보다 예를 들면 cfDNA에 매우 적은 양으로 존재하는 ctDNA의 저빈도 유전자 변이의 검출이 가능하다.

도면의 간단한 설명

[0028] 도 1은 본원의 일 구현예에 따른 방법을 도식적으로 나타낸 것이다.

도 2a는 실제 cfDNA에 존재하는 유전체 DNA 단편의 길이에 따른 단편의 개수 (fragments count)를 나타낸 그래프로, 2개의 봉우리를 갖는 분포를 나타낸다. 각각 봉우리는 166 bp 와 315 bp에서 최빈값이 확인된다. DNA 단

편 분포는 전체 중의 비율을 계산하여 DNA 단편이 나타날 확률로 간주할 수 있다.

도 2b는 유전체의 특정 loci (좌위)에서 6,600개의 DNA 단편이 있을 때 (6,600 X depth 로 표현됨), 도 2a에서 계산된 DNA 단편 길이의 확률로부터 특정 길이의 DNA 단편의 개수를 계산하고, 해당하는 길이에서 발생가능한 collision을 계산한 그래프이다. 도 2a에서와 마찬가지로 166 bp 길이에서 단편이 가장 많이 분포하고 이 길이에서 DNA의 collision fragments count의 비율이 40.3%로 매우 높고, 누적 collision fraction의 대부분이 100~200 bp 사이 길이에서 일어나고, 이는 일반적인 NGS 분석에서는 사용되지 못하고 버려지는 데이터를 나타낸다.

도 2c는 시작 DNA의 양에 따른 고유 단편의 비율을 나타낸 그래프이다. 20ng에서 21.4%의 단편이 중복서열(duplicates)로 분류되어 이는 일반적인 NGS 분석에서는 사용되지 못한다.

도 3은 본원의 일 구현예에 따른 방법에 따라 20ng의 시작 DNA를 2, 4 및 8개의 aliquot로 나눈 후, aliquot 갯수에 따른 FMD (Fragment Mean Depth) 증가를 나타내는 것으로 4개에서 포화되는 것을 나타낸다.

도 4는 본원에 따른 방법에 의한 FMD 값과 기존 분석 (모든 duplicate 제거)의 FMD 값을 비교한 그래프로, 20ng의 시작 DNA를 2, 4 및 8개의 aliquot로 나누어 분석한 경우, 2개 이상의 모든 aliquot를 이용한 분석에서 FMD 값이 기존 분석 보다 높을 것을 나타낸다.

도 5는 본원의 일 구현예에 따른 방법에 따라 20ng의 시작 DNA를 2, 3 및 8개의 aliquot로 나눈 후, 각 aliquot 별로 library 제조 단계인 pre-PCR 단계에서 증폭되는 DNA 양을 나타내며, aliquot 개수가 증가할수록, 최종적으로 얻을 수 있는 DNA양은 증가하나, 4개 aliquot로 나눴을 때 포화되는 것으로 나타났다.

도 6a 및 도 6b는 오류 수정 전 (a)/후 (b)의 변이의 VAF (Variant allele frequency) (1% 미만) calling 결과를 나타내는 것으로, 오류 수정 전 변이 (a)는 VAF 1%에서 다수의 false positive (FP) 변이가 검출되나, consensus DNA 생성과 aliquot 정보를 이용한 오류 수정 후(b)는 true positive (TP)만 남고 모든 FP가 제거되는 것을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0029] 본원은 cfDNA를 사용한 NGS 분석에 있어서, 특정 시료의 cfDNA 풀에 서로 다른 세포에서 유래했지만 우연히 동일한 부위가 잘려서 이를 PCR로 증폭된 것과 구분하지 못해 NGS 분석과정에서의 데이터가 소실되나, 특정 양의 cfDNA의 collision count를 최소로 하는 방식으로 특정 양의 cfDNA를 복수개의 aliquot로 나누어 라이브러리를 제조할 경우, 상기 데이터 소실을 최소화할 수 있고, 궁극적으로 저빈도 유전자 변이의 검출이 가능하다는 발견에 근거한 것이다.

[0030] 본원에서 “cfDNA(cell-free DNA, cfDNA)”는 혈액 속에 존재하는 다양한 길이의 유전체 단편을 포함하나, 히스톤 단백질에 의해 보호되지 않는 크로마틴 부분이 주로 잘려서 166 bp 길이에서 최빈값을 보인다. cfDNA는 건강한 사람들의 경우 조혈 세포(haematopoietic cell)로부터 방출된 DNA가 대부분이고, 암환자의 경우, 암세포 사멸로 인해 후술하는 바와 같이 암세포 유래의 ctDNA를 포함한다. cfDNA는 혈액으로부터 추출될 수 있으며, 이를 추출하는 시약/키트는 시중에서 구입할 수 있고, 그 방법은 공지되어 있다 (Clara Perez-Barrios *et al.* Traansl Lung Cancer Res 5 (2016).

[0031] 본원에서 “ctDNA (circulating tumor DNA)”는 cfDNA에 포함된, 암세포에서 유래된 유전체 단편이다. 총 cfDNA의 단지 <0.1 ~ 10% 양으로 포함되어 있다. ctDNA는 암세포의 급격한 자기복제로 인해 히스톤 의해 보호되는 부위가 더 적고 결과적으로 건강한 세포유래 cfDNA보다 더 짧아져 주로 90 ~ 150 bp의 길이로 보통의 cfDNA보다 약 20-40 bp (Mouliere, F. *et al*. Sci Transl Med 10, (2018).) 짧다. 이러한 ctDNA는 특정 암과 관련된 유전적 변이를 포함하고 있어 혈액을 이용한 이러한 유전적 변이의 모니터링을 통해, 병변 발생 전 암의 조기 발견, 특정 암치료법에 대한 반응 분석, 항암제에 대한 저항성 생성 기전 발견, 잔존 암의 존재 등의 확인에 유용하게 사용될 수 있다. 본원에 따른 일 구현예에서는 cfDNA에 포함된 DNA 유전체 단편 중에서 약 1% 미만의 저빈도로 존재하는 암세포 유래의 ctDNA의 유전자 변이를 검출하고자 한다.

[0032] 본원에서 “NGS (Next Generation Sequencing)”란, 유전체의 염기서열 분석기술 중 하나로, 유전체 유래의 DNA 단편을 병렬로 처리함으로써 염기서열을 고속으로 분석할 수 있다. 이를 위해 단편에 인덱스, 분자바코드 등을 추가하고 증폭하는 과정을 포함하는 라이브러리 제조 및 산출된 원(raw) 데이터의 정렬(alignment) 및 참조 염기서열에의 맵핑을 통한 오류 처리 및 염기서열 도출 등의 데이터 분석 과정이 필요하다. 차세대 염기서열 분석은 목적에 따라 다양한 분석 플랫폼으로 이용될 수 있다. 예를 들어, 차세대 염기서열 분석의 분석 플랫폼

은 Illumina NextSeq, Illumina NovaSeq, ThermoFisher Ion Proton, Pacific Biosciences Sequel II, BGI MGI 등을 들 수 있고, 각 플랫폼에 사용되는 라이브러리 제조 키트 및 방법은 해당 플랫폼 제조사로부터 입수할 수 있다.

[0033] 이러한 NGS는 그 특징으로 인해 다음과 같은 본질적 문제점이 있다. 먼저 오류 처리 방법으로, 오류는 실험 방법과 NGS 플랫폼에 따라 다른데, 예를 들면 Illumina Inc.의 장비에서는 평균적으로 뉴클레오타이드 당 0.1 ~ 1%의 error rate을 가지고 있다. 일반적으로 cfDNA 검사는 AF (Allele Frequency) 1% 미만의 LOD(Limit of Detection)를 목표로 하기 때문에 전통적인 NGS 실험으로는 진짜 변이와 에러를 구분할 수 없다. 한편 NGS 실험은 어떤 방법을 사용하더라도 반드시 PCR을 이용한 DNA 증폭 단계를 포함한다. 그런데 DNA 증폭은 DNA의 GC 함량, DNA 길이 등 여러가지 요소로 DNA 단편마다 증폭 효율이 다르기 때문에 모든 단편이 균일한 정도로 증폭되는 결과를 얻을 수 없다. 그렇기 때문에 분석단계에서 duplicates (하나의 DNA에서 증폭된 복제물로 PCR duplicate와 collision을 모두 포함)를 제거하여 이 효과를 보정한다. 이때 일반적으로 picard 툴을 사용하는데, 참조 유전체와 동일한 read (NGS로 읽힌 판독서열을 칭함)를 남기고 나머지 duplicates는 제외한다. 만약 duplicates 중에 무작위로 특정 염기(nucleotide)에 오류가 발생하면 비록 하나의 DNA에서 복제되었지만 서열이 다른 reads로 보인다. 전통적인 NGS에서는 이 reads 들은 대체로 무시되고 참조유전체와 가장 가까운 read만 분석에 사용된다. 그런데 ctDNA는 길이가 짧은 특성상 우연히 서로 다른 세포에서 유래한 DNA이지만 완전히 동일한 서열을 갖는 경우가 자주 발생하고, 전통적인 NGS 방법에서는 이 것이 PCR duplicates 인지 서로 다른 세포에서 유래했는지를 구분할 수 없다. 따라서 이로 인해 변이가 발생한 DNA는 무시되는 경우가 발생한다. 이후 Molecular barcoding을 이용한 ctDNA 검사에서는 위의 문제를 극복하기 위해 Barcode sequence 또는 UMI (unique molecular identifier) 기술이 개발되었다. 이 기술을 이용하면 barcode sequence를 사용해 서로 다른 세포에서 유래한 reads를 구분할 수 있고, 이 reads 중에 PCR 오류와 진짜 변이를 구분할 수 있다. 이 과정을 일반적으로 error correction이라 부른다. 이 경우 시퀀싱 오류는 각 base마다 Q 값으로 계산된다. 일반적으로 Illumina 시퀀싱의 전체 단계 중 시작과 끝이 error rates이 더 크다. Molecular barcode sequence는 시퀀싱의 처음에 읽히기 때문에 상대적으로 더 오류에 취약할 수밖에 없다. Barcode sequence가 잘못 읽히는 문제로 인해 고유 분자를 구분하기 위한 본래 목적이 크게 훼손된다. 이런 문제를 극복하기 위해 barcode 서열의 길이를 조절하고, 서열을 정교하게 조합하는 등 많은 시도들이 있었다 (Smith et al., Somervuo et al, *Genome Res.* 27, 491-499 (2017); Somervuo, P. *et al. BMC Bioinformatics* 19, 257 (2018)). 또한 Barcode가 포함된 adapter dimer에 의해 conversion rate가 저하된다. 보통 ctDNA의 길이는 cfDNA 보다 짧고 adapter dimer 보다는 길다. Adapter dimer 제거 시 DNA 길이를 이용하는데, barcode sequence로 인해 DNA 길이가 더 길어졌기 때문에 ctDNA와 구분이 더 어려워지고 adapter dimer 제거시 ctDNA가 더 많이 유실된다. 결과적으로 ctDNA 자체의 conversion rate는 전체 cfDNA의 값보다 더 낮아진다.

[0034] 즉, 세포에서 방출된 유전체 DNA가 잘리는 과정 중 서로 다른 세포에서 유래했지만 우연히 동일한 부위가 잘려서 NGS 결과에서 서열이 동일하게 나타날 수 있다. 한편, DNA는 NGS 실험과정 중 라이브러리 제조과정에서 PCR로 증폭되고 최종적으로 중복서열로 나타나기 때문에, 우연히 서열이 동일한 경우도 일반적인 분석 과정에서 중복서열로 제거된다. 특히 ctDNA는 암세포에 특이적 유전자 변이를 포함하고 있어 이를 검출하는 것이 중요하다. 하지만, ctDNA는 cfDNA에 매우 적은 양으로 포함되어 있으며 중복서열 제거 과정에서 정보가 소실되어 검출이 되지 않는 문제점이 있다.

[0035] 이러한 NGS의 문제점으로 인해, NGS를 이용한 혈액의 cfDNA 분석에 있어서, 분석에 사용될 수 있는 ctDNA의 정보는 더욱 제한된다. 또한 앞서 언급한 바와 같이, 혈액에 포함된 cfDNA의 양, 이에 포함된 ctDNA의 양의 매우 제한적이고 임상에서 채취할 수 있는 혈액의 양도 매우 제한적이어서 DNA 양 자체를 증가시키는 것은 한계가 있다. 예를 들면 정상인의 경우 혈액내에 DNA양은 평균적으로 약 4.4 ng/ml 정도이다 (Raymond, C. K., Hernandez, J., Karr, R., Hill, K. & Li, M. Collection of cell-free DNA for genomic analysis of solid tumors in a clinical laboratory setting. PLoS One 12, (2017).). 이 중 다른 검사 (예를 들어, ddPCR: 25ng, Real-Time PCR: 1~100ng)를 위해 DNA를 남겨두어야 하기 때문에, 실제 임상에서 NGS 검사에 이용할 수 있는 DNA 양은 더욱 제한적일 수밖에 없다.

[0036] 본원에 따른 방법은 cfDNA에 매우 적은 양으로 포함된 ctDNA에 존재하는 암과 연관된 저빈도 변이 예를 들면 1% 미만의 변이를 검출하기 위해, NGS 분석과정에서 에러 처리로 인해 소실되는 서열을 최소화할 수 있다.

[0037] 이에 한 양태에서 본원은 NGS (Next Generation Sequencing) 분석을 이용한 cfDNA (cell free DNA)의 저빈도 변이 검출에 있어서, 상기 NGS 분석에 사용되는 고유 단편의 비율을 향상시키는 방법에 관한 것이다.

[0038] 일 구현예에서 상기 방법은 (a) 특정 양의 cfDNA 시료를 제공하는 단계; (b) 상기 특정 양의 cfDNA에 포함된 유전체 단편 중 51 내지 330bp 길이에 해당하는 총 유전체 단편의 수, 및 고유 단편 (unique fragment)의 수를 계산하는 단계로, 상기 고유 단편의 수는 상기 총 유전체 단편의 수에서 상기 51 내지 330bp 길이에 해당하는 각 유전체 단편의 collision count의 합을 제외한 값이고, 상기 collision count 합은 다음 식으로부터 계산되며,

[0039] $$\sum_{k=1}^{n} q(k-1;d) = n - d + d\left(\frac{d-1}{d}\right)^{n}.$$

[0040] 상기 식에서 q(k-1;d) : [1,d]의 범위의 n 개의 숫자 중 k 와 같은 숫자가 있을 확률, k : 특정 숫자, d : 숫자의 범위, n : 숫자의 개수임.

[0041] (c) 상기 총 유전체 단편 수에서 상기 고유 단편 수가 차지하는 비를 계산하고, 상기 고유 단편의 비를 증가시키도록 상기 특정 양의 cfDNA 시료를 복수 개의 aliquot로 나누는 단계; 및 (d) 상기 복수개의 각 aliquot 별로, 각각 상이한 인덱스를 포함하는 어뎁터를 태깅하여 라이브러리를 제조하고, NGS (Next Generation Sequencing) 분석을 수행한 후 상기 각 aliquot의 NGS 결과를 통합하는 단계를 포함한다.

[0042] 본원에 따른 방법은 도 1을 참조하면, 시료를 2개 이상의 aliquot로 분할하여 NGS 라이브러리를 만드는 것이 특징이다. 각 aliquots에서 생성된 NGS 데이터는 분석 단계에서 합쳐서 분석함으로써 일반적인 NGS 보다 더 많은 양의 ctDNA로부터 서열정보를 얻을 수 있기 때문에, ctDNA의 VAF (Variant Allele Frequency)가 낮은 유전자 변이의 정확한 검출을 가능하게 한다.

[0043] 본원에서 가용 ctDNA 데이터란, 정상 세포와 종양 세포 유래의 DNA가 섞여 있어서 서로 구분할 수 없는 상태에서, 변이가 발생한 종양 세포 유래 DNA의 변이 검출에 사용가능한 NGS 데이터를 의미한다.

[0044] 본원에 따른 방법에서 특정 양의 DNA는 통상적으로 환자로부터 채취된 혈액으로부터 얻을 수 있는, 혈액 샘플에서 추출된 cfDNA로서, NGS 분석에 일반적으로 할당되는 cfDNA의 양을 의미한다. 예를 들면 정상인의 경우 혈액 내에 DNA양은 평균적으로 약 4.4 ng/ml 정도이다 (Raymond, C. K., Hernandez, J., Karr, R., Hill, K. & Li, M. Collection of cell-free DNA for genomic analysis of solid tumors in a clinical laboratory setting. PLoS One 12, (2017).). NGS 분석을 위해 환자 한 명으로부터 일반적으로 약 5ml의 혈액을 채취한다면, 약 20ng의 cfDNA를 수득할 수 있으나, 구체적인 양은 암종이나 암의 진행 상태에 달라질 수 있다.

[0045] 다음 단계로 본원에 따른 방법은 특정 양의 cfDNA에 포함된 유전체 단편의 개수 및 collision count를 이용하여 상기 단편 중 고유 단편의 개수를 수득하는 단계를 포함한다.

[0046] 본원에서는 특정 양의 cfDNA에 포함된 유전체 단편 중 특히 51 내지 330bp 길이에 해당하는 총 유전체 단편의 수 및 고유 단편(unique fragment)의 수를 계산한다. cfDNA 유전체 단편은 5bp ~ 991bp의 길이로 2개의 봉우리를 갖는 분포를 나타낸다. 각각 봉우리는 166 bp와 315 bp에서 최빈값을 갖고, 166 bp의 첫번째 봉우리는 315 bp의 두번째 봉우리보다 16배 정도 큰 비율을 갖는다. cfDNA 단편 중 50bp 이하로 과도하게 많이 잘려서 정보를 잃거나, 330 bp 이상으로 아주 긴 단편은 대부분 ctDNA가 아니므로 검사에 유용하지 않다. 따라서 51 ~ 330bp의 단편에 대부분의 ctDNA 정보가 포함되어 있으며, 본원에 따른 일 구현예에서는 다양한 길이의 유전체 단편을 포함하는 cfDNA에서 51 ~ 330bp 길이의 단편이 분석에 사용된다.

[0047] cfDNA에 포함된 유전체 단편의 개수는 단편의 분자수에 상응하는 개념으로, 인간의 경우 통상 1ng DNA는 330개 Genome Equivalents (한 개 세포에 포함된 모든 유전자가 존재하는 DNA의 양으로 이 수는 특정 생물의 유전체의 크기에 따라 다르며 유전체 염기쌍을 ug의 DNA로 변환하여 계산된다)를 포함한다. 본원에 따른 일 구현예에서 상기 cfDNA의 유전체 단편의 길이는 약 51bp 내지 330bp로, 예를 들면 20ng의 cfDNA에서 51 - 330bp 길이를 갖는 단편의 개수는 표 1 및 2를 참조하면 6,173개이다.

[0048] 본원에 따른 방법에서 collision count를 이용해서 다양한 길이의 유전체 단편을 포함하는 특정 양의 cfDNA에서 우연히 서열이 동일한 단편의 개수가 계산되고, 전체 단편의 개수에서 우연히 서열이 동일한 단편의 개수를 빼면 고유 단편의 수가 된다. 다양한 길이의 유전체 단편을 포함하는 특정 DNA 시료에서 DNA 서열이 우연히 동일한 경우를 collision이라고 하고, collision counting 방법에 따라 우연히 서열이 동일한 DNA 단편의 개수를 계산할 수 있다.

[0049] Collison count는 Birthday paradox 에 근거한 것으로 특정한 크기의 집단에서 우연히 동일한 개체의 확률로, [1,d] 범위로부터 무작위로 선택된 k번째 정수가 적어도 하나의 앞선 선택이 반복될 확률은 q(k-1;d)와 같다는

것으로 다음과 같은 식 1으로 표시될 수 있다 (Might, Matt. "Collision hash collisions with the birthday paradox". Matt Might's blog. Retrieved 17 July 2015).

[0050] [식 1]

$$\sum_{k=1}^{n} q(k-1;d) = n - d + d\left(\frac{d-1}{d}\right)^{n}.$$

[0051]

[0052] 상기 식에서, q(k-1;d) : [1,d]의 범위의 n 개의 숫자 중 k 와 같은 숫자가 있을 확률, k : 특정 숫자, d : 숫자의 범위, n : 숫자의 개수 이다.

[0053] 예를 들어 cfDNA 20ng 에는 6,600개 Genome Equivalents가 존재하는 것은 알려진 사실이고, 이 경우, 어떤 특정한 loci (좌위)를 기준으로 보면, 6,600개의 다른 genome에서 유래된 DNA 단편이 존재하고 (6,600 X depth로 표현), 이 단편은 도 2a에서 계산된 확률과 동일한 분포로 다양한 길이로 존재하고 그 개수는 2b의 분포와 표 1에 표시된 값과 같다. 하지만, 이 중 실제 저빈도 서열 변이를 포함하는 ctDNA의 정보를 반영하는 길이를 51 ~ 330bp 범위로 가정했을 때, 해당하는 유전체 DNA 단편의 총 합은 표 2에 계산된 6,173개이다. 그리고 각 길이의 단편마다 상이한 collision count 값이 계산되는데, 예를 들어 166bp 단편의 경우 표 1에 따르면 그 길이 분포 확률은 0.02874이고 이는 6,600개 중 189.687개에 해당하고, 이때의 collision counts는 다음과 같이 계산되어 76.4513개이다.

$$189.687 - 166 + 166\left(\frac{166-1}{166}\right)^{189.687} = 76.4513$$

[0054]

[0055] 이러한 방식으로, 식 1에 따라 51 ~ 330bp 길이의 각 단편의 collision counts를 계산하여 합하면 1,319개이고 이 것은 6,173개 중 21.4%에 해당한다 (표 2 참조).

[0056] 앞서 언급한대로 cfDNA 특성상 서로 다른 세포에서 유래한 DNA이지만 우연히 완전히 동일한 서열을 갖는 경우가 자주 발생하고, 전통적인 NGS 방법에서는 PCR duplicates 인지 서로 다른 세포에서 유래했는지를 구분할 수 없다. 따라서 이로 인해 변이가 발생한 DNA는 무시되는 경우가 발생하여 가용 ctDNA 정보가 소실된다.

[0057] 따라서 본원에서는 우연히 동일한 DNA가 발생할 확률을 계산하고, 이러한 확률을 최소화하는 방법으로 시료를 분할하여 NGS 검사가 진행된다.

[0058] 예를 들면 이론적으로 계산했을 때, 앞서 언급한 바와 같이 약 20ng DNA에는 6,600개 Genome Equivalents가 존재하고, 이중 분석적으로 의미가 있는 51 ~ 330 bp 길이를 갖는 단편은 6,173개이다. 이 중 Collision count에 의하면 1,319개 copy (21.4%)가 DNA 서열이 우연히 서로 동일해서 PCR duplicates와 구분이 불가능하고 이 데이터는 NGS 분석에서 사용되지 못하고 버려진다. 따라서 서로 다른 세포에서 유래한 DNA 정보라고 하더라도, 분석 과정에서 PCR duplicates를 제거하는 단계에서 버려진다. 이로 인해 실제 DNA 중 서로 다른 고유한 DNA 단편은 4,854개 (6,173-1,319) 이고, 고유 단편의 비는 0.786 (4,854/6,173)이다. 즉 78.6%의 단편만이 분석에 사용된다는 것이다.

[0059] 고유 DNA 단편은 서로 다른 세포에서 유래한 것이고, 우연히 서열이 동일한 것으로 간주되어 제거되지 않는 것이 저빈도 변이 검출에 중요하기 때문에, 고유 단편의 비를 가능한 증가시키는 것이 유리하다. 앞서 언급한 바와 같이 특정 DNA 시료에서 DNA 서열이 우연히 동일한 경우를 collision이라고 하고, collision counting 방법에 따라 특정 길이의 DNA 중에서 우연히 서열이 동일한 DNA 단편의 개수를 계산할 수 있다.

[0060] 본원에서는 cfDNA 단편의 길이별로 발생할 수 있는 확률을 계산하기위해 실제 혈액으로부터 cfDNA를 검사하여 단편 길이의 분포를 얻었다 (도 2a 참조). 그리고 이 확률을 바탕으로 특정 loci에 6,600 X depth가 있을 때 각 길이별 DNA 단편의 개수를 계산했다 (도 2b, 표 1 참조). 이 중 너무 짧거나, 긴 DNA 서열을 제외하고 51bp 내지 330bp까지 길이의 단편에 대하여 길이별로 collision counting을 계산한 결과는 아래 표 1과 같다. 표 1에 의하면 20ng을 사용하는 경우, 예를 들면 166bp 길이에서 collision count의 비율이 40.3% 정도로 높게 나타나고, 이는 기존 NGS 분석에서는 사용되지 못하고 버려지는 데이터이다.

[0061] [표 1-1]

Fragment 길이	Fragment 존재 확률	해당 길이의 fragment 중 position 이 다른 경우의 수	20ng (6600 개 fragment) 중 해당 길이를 갖는 단편의 수	20ng Collision Count	Collision count 의 비율
51	0.000016	51	0.11	0	0.0%
52	0.000021	52	0.14	0	0.0%
53	0.000019	53	0.12	0	0.0%
54	0.000022	54	0.15	0	0.0%
55	0.000023	55	0.15	0	0.0%
56	0.000031	56	0.21	0	0.0%
57	0.000032	57	0.21	0	0.0%
58	0.000038	58	0.25	0	0.0%
59	0.000043	59	0.29	0	0.0%
60	0.000046	60	0.30	0	0.0%
61	0.000053	61	0.35	0	0.0%
62	0.000048	62	0.32	0	0.0%
63	0.000046	63	0.31	0	0.0%
64	0.000043	64	0.28	0	0.0%
65	0.000050	65	0.33	0	0.0%
66	0.000053	66	0.35	0	0.0%
67	0.000059	67	0.39	0	0.0%
68	0.000070	68	0.46	0	0.0%
69	0.000081	69	0.54	0	0.0%
70	0.000084	70	0.56	0	0.0%
71	0.000087	71	0.57	0	0.0%
72	0.000079	72	0.52	0	0.0%
73	0.000085	73	0.56	0	0.0%
74	0.000088	74	0.58	0	0.0%
75	0.000085	75	0.56	0	0.0%
76	0.000091	76	0.60	0	0.0%
77	0.000109	77	0.72	0	0.0%
78	0.000124	78	0.82	0	0.0%
79	0.000160	79	1.05	0	0.0%
80	0.000200	80	1.32	0	0.0%

[0062]

[0063] [표 1-2]

Fragment 길이	Fragment 존재 확률	해당 길이의 fragment 중 position 이 다른 경우의 수	20ng (6600 개 fragment) 중 해당 길이를 갖는 단편의 수	20ng Collision Count	Collision count 의 비율
81	0.000195	81	1.29	0	0.0%
82	0.000153	82	1.01	0	0.0%
83	0.000141	83	0.93	0	0.0%
84	0.000120	84	0.79	0	0.0%
85	0.000129	85	0.85	0	0.0%
86	0.000141	86	0.93	0	0.0%
87	0.000157	87	1.04	0.000219	0.0%
88	0.000165	88	1.09	0.0005495	0.1%
89	0.000211	89	1.40	0.0031037	0.2%
90	0.000274	90	1.81	0.0081559	0.5%
91	0.000348	91	2.29	0.0162852	0.7%
92	0.000324	92	2.14	0.0132215	0.6%
93	0.000279	93	1.84	0.0083328	0.5%
94	0.000225	94	1.49	0.0038582	0.3%
95	0.000227	95	1.50	0.0039528	0.3%
96	0.000228	96	1.50	0.0039428	0.3%
97	0.000242	97	1.60	0.0049544	0.3%
98	0.000295	98	1.95	0.0094001	0.5%
99	0.000359	99	2.37	0.0163691	0.7%
100	0.000435	100	2.87	0.026709	0.9%
101	0.000459	101	3.03	0.0303714	1.0%
102	0.000481	102	3.17	0.0336411	1.1%
103	0.000467	103	3.08	0.0310534	1.0%
104	0.000410	104	2.71	0.0221575	0.8%
105	0.000404	105	2.66	0.0210665	0.8%
106	0.000440	106	2.90	0.0259807	0.9%
107	0.000468	107	3.09	0.0301259	1.0%
108	0.000518	108	3.42	0.0381896	1.1%
109	0.000593	109	3.92	0.0520643	1.3%
110	0.000677	110	4.47	0.0698032	1.6%
111	0.000775	111	5.12	0.094019	1.8%
112	0.000760	112	5.02	0.0892082	1.8%
113	0.000662	113	4.37	0.0646737	1.5%

[0064]

[0065] [표 1-3]

Fragment 길이	Fragment 존재 확률	해당 길이의 fragment 중 position 이 다른 경우의 수	20ng (6600 개 fragment) 중 해당 길이를 갖는 단편의 수	20ng Collision Count	Collision count 의 비율
114	0.000625	114	4.12	0.0561159	1.4%
115	0.000605	115	3.99	0.0515675	1.3%
116	0.000680	116	4.49	0.0670939	1.5%
117	0.000743	117	4.90	0.0811063	1.7%
118	0.000789	118	5.21	0.0921151	1.8%
119	0.000922	119	6.08	0.128439	2.1%
120	0.001127	120	7.44	0.1964112	2.6%
121	0.001449	121	9.56	0.3315454	3.5%
122	0.001755	122	11.58	0.4893183	4.2%
123	0.001637	123	10.80	0.4202291	3.9%
124	0.001353	124	8.93	0.2803854	3.1%
125	0.001132	125	7.47	0.1905513	2.6%
126	0.001135	126	7.49	0.1901254	2.5%
127	0.001230	127	8.12	0.2237167	2.8%
128	0.001392	128	9.18	0.288204	3.1%
129	0.001641	129	10.83	0.4035681	3.7%
130	0.001932	130	12.75	0.5605187	4.4%
131	0.002531	131	16.70	0.9644321	5.8%
132	0.003236	132	21.36	1.5690623	7.3%
133	0.003943	133	26.02	2.3070925	8.9%
134	0.004084	134	26.95	2.4549264	9.1%
135	0.003750	135	24.75	2.0601001	8.3%
136	0.003444	136	22.73	1.7272533	7.6%
137	0.003890	137	25.67	2.1841482	8.5%
138	0.004538	138	29.95	2.9390325	9.8%
139	0.005535	139	36.53	4.3038944	11.8%
140	0.006545	140	43.19	5.9139093	13.7%
141	0.007160	141	47.25	6.9824241	14.8%
142	0.007428	142	49.03	7.4449552	15.2%
143	0.007467	143	49.28	7.4732752	15.2%
144	0.007673	144	50.64	7.8223846	15.4%
145	0.007750	145	51.15	7.9242909	15.5%
146	0.008047	146	53.11	8.4609704	15.9%

[0066]

[0067] [표 1-4]

Fragment 길이	Fragment 존재 확률	해당 길이의 fragment 중 position 이 다른 경우의 수	20ng (6600 개 fragment) 중 해당 길이를 갖는 단편의 수	20ng Collision Count	Collision count 의 비율
147	0.008270	147	54.58	8.8578699	16.2%
148	0.008755	148	57.78	9.8117274	17.0%
149	0.009384	149	61.94	11.122706	18.0%
150	0.010203	150	67.34	12.941604	19.2%
151	0.010751	151	70.96	14.193123	20.0%
152	0.011284	152	74.47	15.446494	20.7%
153	0.011379	153	75.10	15.602162	20.8%
154	0.011305	154	74.61	15.327489	20.5%
155	0.011804	155	77.91	16.520818	21.2%
156	0.012833	156	84.70	19.18153	22.6%
157	0.014089	157	92.99	22.654828	24.4%
158	0.015611	158	103.03	27.172231	26.4%
159	0.016979	159	112.06	31.465095	28.1%
160	0.017922	160	118.29	34.501785	29.2%
161	0.018822	161	124.22	37.47267	30.2%
162	0.020235	162	133.55	42.40592	31.8%
163	0.022411	163	147.91	50.510008	34.1%
164	0.025778	164	170.13	64.067668	37.7%
165	0.028364	165	187.20	75.078845	40.1%
166	0.028740	166	189.69	76.45129	40.3%
167	0.027921	167	184.28	72.49149	39.3%
168	0.026564	168	175.32	66.305669	37.8%
169	0.024722	169	163.16	58.334911	35.8%
170	0.023241	170	153.39	52.165388	34.0%
171	0.021241	171	140.19	44.33719	31.6%
172	0.020003	172	132.02	39.673125	30.1%
173	0.019359	173	127.77	37.251886	29.2%
174	0.018912	174	124.82	35.562301	28.5%
175	0.018207	175	120.17	33.063259	27.5%
176	0.017414	176	114.93	30.365704	26.4%
177	0.016323	177	107.73	26.867792	24.9%
178	0.014823	178	97.83	22.409685	22.9%
179	0.013142	179	86.74	17.844961	20.6%

[0068]

[0069] [표 1-5]

Fragment 길이	Fragment 존재 확률	해당 길이의 fragment 중 position 이 다른 경우의 수	20ng (6600 개 fragment) 중 해당 길이를 갖는 단편의 수	20ng Collision Count	Collision count 의 비율
180	0.011824	180	78.04	14.575045	18.7%
181	0.010648	181	70.28	11.904888	16.9%
182	0.009721	182	64.16	9.9654176	15.5%
183	0.008745	183	57.72	8.1005821	14.0%
184	0.008255	184	54.48	7.2148401	13.2%
185	0.007945	185	52.44	6.6693924	12.7%
186	0.007292	186	48.13	5.621881	11.7%
187	0.006927	187	45.72	5.0640192	11.1%
188	0.006413	188	42.33	4.3361794	10.2%
189	0.005969	189	39.39	3.7494572	9.5%
190	0.005404	190	35.67	3.0697123	8.6%
191	0.004867	191	32.12	2.4846634	7.7%
192	0.004451	192	29.37	2.0707941	7.0%
193	0.004009	193	26.46	1.6735759	6.3%
194	0.003750	194	24.75	1.4571476	5.9%
195	0.003461	195	22.84	1.2346114	5.4%
196	0.003300	196	21.78	1.1166498	5.1%
197	0.003167	197	20.90	1.0228044	4.9%
198	0.002935	198	19.37	0.8728059	4.5%
199	0.002821	199	18.62	0.8019103	4.3%
200	0.002671	200	17.63	0.714144	4.1%
201	0.002455	201	16.20	0.5985785	3.7%
202	0.002208	202	14.57	0.4795369	3.3%
203	0.002112	203	13.94	0.435555	3.1%
204	0.001918	204	12.66	0.3553651	2.8%
205	0.001829	205	12.07	0.3207386	2.7%
206	0.001720	206	11.35	0.2810645	2.5%
207	0.001591	207	10.50	0.2377054	2.3%
208	0.001484	208	9.80	0.2045332	2.1%
209	0.001406	209	9.28	0.1817837	2.0%
210	0.001384	210	9.14	0.1750144	1.9%
211	0.001263	211	8.33	0.143394	1.7%
212	0.001152	212	7.60	0.1172997	1.5%

[0070]

[0071] [표 1-6]

Fragment 길이	Fragment 존재 확률	해당 길이의 fragment 중 position 이 다른 경우의 수	20ng (6600 개 fragment) 중 해당 길이를 갖는 단편의 수	20ng Collision Count	Collision count 의 비율
213	0.001026	213	6.77	0.0910415	1.3%
214	0.000966	214	6.38	0.0795505	1.2%
215	0.000943	215	6.22	0.0750397	1.2%
216	0.000872	216	5.76	0.0629987	1.1%
217	0.000801	217	5.29	0.0520222	1.0%
218	0.000783	218	5.17	0.0491292	1.0%
219	0.000674	219	4.45	0.0349461	0.8%
220	0.000701	220	4.63	0.0380348	0.8%
221	0.000653	221	4.31	0.0321531	0.7%
222	0.000637	222	4.21	0.0302749	0.7%
223	0.000596	223	3.93	0.025814	0.7%
224	0.000538	224	3.55	0.020143	0.6%
225	0.000531	225	3.50	0.0194665	0.6%
226	0.000507	226	3.35	0.0173487	0.5%
227	0.000442	227	2.91	0.0122772	0.4%
228	0.000466	228	3.08	0.0140133	0.5%
229	0.000386	229	2.55	0.0086128	0.3%
230	0.000405	230	2.67	0.0096883	0.4%
231	0.000404	231	2.67	0.0096095	0.4%
232	0.000398	232	2.62	0.0091792	0.3%
233	0.000374	233	2.47	0.0077798	0.3%
234	0.000345	234	2.28	0.0062302	0.3%
235	0.000322	235	2.12	0.0050699	0.2%
236	0.000324	236	2.14	0.0051765	0.2%
237	0.000319	237	2.11	0.0049276	0.2%
238	0.000286	238	1.89	0.0035212	0.2%
239	0.000309	239	2.04	0.0044208	0.2%
240	0.000300	240	1.98	0.004055	0.2%
241	0.000295	241	1.95	0.0038254	0.2%
242	0.000283	242	1.87	0.0033458	0.2%
243	0.000296	243	1.95	0.0038375	0.2%
244	0.000260	244	1.72	0.0025351	0.1%
245	0.000282	245	1.86	0.0032673	0.2%

[0072]

[0073] [표 1-7]

Fragment 길이	Fragment 존재 확률	해당 길이의 fragment 중 position 이 다른 경우의 수	20ng (6600 개 fragment) 중 해당 길이를 갖는 단편의 수	20ng Collision Count	Collision count 의 비율
246	0.000286	246	1.89	0.0033993	0.2%
247	0.000289	247	1.91	0.0035159	0.2%
248	0.000258	248	1.70	0.0024146	0.1%
249	0.000263	249	1.74	0.0025738	0.1%
250	0.000276	250	1.82	0.002987	0.2%
251	0.000290	251	1.92	0.0034968	0.2%
252	0.000289	252	1.91	0.0034378	0.2%
253	0.000282	253	1.86	0.0031519	0.2%
254	0.000250	254	1.65	0.0021205	0.1%
255	0.000282	255	1.86	0.0031481	0.2%
256	0.000244	256	1.61	0.0019266	0.1%
257	0.000269	257	1.78	0.0026842	0.2%
258	0.000267	258	1.76	0.0026065	0.1%
259	0.000264	259	1.74	0.002487	0.1%
260	0.000285	260	1.88	0.0031854	0.2%
261	0.000273	261	1.80	0.0027773	0.2%
262	0.000278	262	1.83	0.0029166	0.2%
263	0.000321	263	2.12	0.0044933	0.2%
264	0.000291	264	1.92	0.0033527	0.2%
265	0.000293	265	1.94	0.0034148	0.2%
266	0.000297	266	1.96	0.0035262	0.2%
267	0.000294	267	1.94	0.0034103	0.2%
268	0.000297	268	1.96	0.003509	0.2%
269	0.000323	269	2.13	0.0044915	0.2%
270	0.000321	270	2.12	0.0043824	0.2%
271	0.000345	271	2.28	0.005359	0.2%
272	0.000371	272	2.45	0.0065194	0.3%
273	0.000384	273	2.53	0.0071123	0.3%
274	0.000362	274	2.39	0.0060642	0.3%
275	0.000394	275	2.60	0.0075554	0.3%
276	0.000390	276	2.58	0.0073582	0.3%
277	0.000401	277	2.65	0.0078563	0.3%
278	0.000421	278	2.78	0.0088756	0.3%

[0074]

[0075] [표 1-8]

Fragment 길이	Fragment 존재 확률	해당 길이의 fragment 중 position 이 다른 경우의 수	20ng (6600 개 fragment) 중 해당 길이를 갖는 단편의 수	20ng Collision Count	Collision count 의 비율
279	0.000418	279	2.76	0.0086796	0.3%
280	0.000477	280	3.15	0.0120771	0.4%
281	0.000511	281	3.37	0.0141941	0.4%
282	0.000555	282	3.66	0.0172579	0.5%
283	0.000577	283	3.81	0.0188333	0.5%
284	0.000584	284	3.86	0.0193495	0.5%
285	0.000592	285	3.91	0.0199063	0.5%
286	0.000595	286	3.93	0.0200309	0.5%
287	0.000666	287	4.40	0.0259505	0.6%
288	0.000697	288	4.60	0.0286504	0.6%
289	0.000752	289	4.96	0.033932	0.7%
290	0.000804	290	5.31	0.0392815	0.7%
291	0.000828	291	5.47	0.0417634	0.8%
292	0.000860	292	5.68	0.045296	0.8%
293	0.000863	293	5.70	0.0454681	0.8%
294	0.000964	294	6.36	0.0576814	0.9%
295	0.001023	295	6.75	0.0654103	1.0%
296	0.001064	296	7.02	0.0709781	1.0%
297	0.001133	297	7.48	0.0810601	1.1%
298	0.001179	298	7.78	0.0879563	1.1%
299	0.001250	299	8.25	0.0993978	1.2%
300	0.001277	300	8.43	0.1036694	1.2%
301	0.001325	301	8.74	0.1116373	1.3%
302	0.001356	302	8.95	0.1169193	1.3%
303	0.001372	303	9.05	0.1193929	1.3%
304	0.001451	304	9.57	0.1339202	1.4%
305	0.001494	305	9.86	0.1420285	1.4%
306	0.001595	306	10.53	0.1623699	1.5%
307	0.001606	307	10.60	0.1642491	1.5%
308	0.001627	308	10.74	0.1680851	1.6%
309	0.001667	309	11.00	0.1762472	1.6%
310	0.001662	310	10.97	0.174698	1.6%
311	0.001721	311	11.36	0.1872885	1.6%

[0076]

[0077] [표 1-9]

Fragment 길이	Fragment 존재 확률	해당 길이의 fragment 중 position 이 다른 경우의 수	20ng (6600 개 fragment) 중 해당 길이를 갖는 단편의 수	20ng Collision Count	Collision count 의 비율
312	0.001736	312	11.46	0.1901055	1.7%
313	0.001739	313	11.48	0.1901182	1.7%
314	0.001742	314	11.50	0.1901899	1.7%
315	0.001794	315	11.84	0.2015678	1.7%
316	0.001686	316	11.13	0.1765444	1.6%
317	0.001750	317	11.55	0.1901907	1.6%
318	0.001702	318	11.24	0.1791083	1.6%
319	0.001724	319	11.38	0.1832094	1.6%
320	0.001715	320	11.32	0.1807096	1.6%
321	0.001660	321	10.95	0.1682735	1.5%
322	0.001678	322	11.08	0.1716508	1.5%
323	0.001588	323	10.48	0.1523997	1.5%
324	0.001617	324	10.67	0.1579097	1.5%
325	0.001577	325	10.41	0.1492874	1.4%
326	0.001552	326	10.24	0.1440361	1.4%
327	0.001487	327	9.82	0.131255	1.3%
328	0.001484	328	9.80	0.1302963	1.3%
329	0.001432	329	9.45	0.1204847	1.3%
330	0.001442	330	9.52	0.1219421	1.3%

[0078]

[0080] 한편 본원에서는 보다 적은 양의 cfDNA를 사용하는 경우, 166bp에서 collision count가 6.06% 수준으로 낮아지며, 즉 NGS 실험에 사용되는 input cfDNA의 양이 증가할수록 collision count의 비율이 증가하는 것을 발견하였다 (표 2 참조). 시작 DNA 양이 적을수록 더 많은 비율의 데이터를 분석에 사용할 수 있다. 따라서 특정 양의 DNA를 일정 숫자로 나눠서 더 작은 시작 DNA 양으로 검사할 경우 더 많은 DNA 정보를 분석에 활용할 수 있음을 발견하였다.

[0081] [표 2]

DNA input amount	Fragments (genome equivalents)	Fragments in 51~330 bp range	Collision counts	Collision fragment ratio	Unique fragments	Unique fragments ratio
1 ng	330	309	3	0.011	305	0.989
5 ng	1,650	1,543	95	0.061	1,448	0.939
10 ng	3,300	3,087	365	0.118	2,721	0.882
20 ng	6,600	6,173	1,319	0.214	4,855	0.786
40 ng	13,200	12,346	4,338	0.351	8,008	0.649
100 ng	33,000	30,866	17,327	0.561	13,539	0.439
150 ng	49,500	46,299	29,825	0.644	16,474	0.356
200 ng	66,000	61,732	42,929	0.695	18,803	0.305

[0082]

[0083] 이에 본원에 따른 방법은 고유 단편의 비가 증가하도록 cfDNA 시료를 복수개의 aliquot로 분할하여 각 aliquot에 대하여 NGS 분석을 수행한다. 예를 들면 표 1을 참조하면, 시작 DNA가 20ng인 경우 1,319개 (21.4%) 단편이 우연히 동일한 서열로 판단되어 분석에 사용되지 못하지만, 시작 cfDNA를 4개로 나누어 각각 약 5ng를 시작 DNA

양으로 하는 경우에는 380 (=95*4)개 단편 (6.1%)이 우연히 동일할 수 있다. 그 결과 4개로 분할하여 NGS 라이브러리를 준비하면 93.9%를 분석에 사용할 수 있고, 분할하지 않은 경우 78.6%에 대비해 15.2% 더 많은 DNA 단편의 서열을 분석에 사용할 수 있다.

[0084] 본원에 따른 방법에서는 하나의 시작 cfDNA 시료에 우연히 서열이 동일한 DNA가 최소화되도록 시작 cfDNA 시료를 적절하게 나누어 분석함으로써, molecular barcode 없이도 각 aliquot를 구분하는 것만으로 고유 단편에서 유래된 저빈도 변이를 발견할 수 있다.

[0085] 일 구현예에서는 고유 단편의 비가 최소 93% 이상이 되도록 특정 양의 시작 DNA를 적절히 분할하여 라이브러리를 제조한다.

[0086] 임상에서 한 환자로부터 혈액을 채취하여 얻을 수 있는 cfDNA의 양이 20ng인 것을 고려하며, 일 구현예에서 특정 양의 cfDNA는 20ng이고, 고유 단편이 비가 93.9%가 되도록 상기 cfDNA를 4개의 aliquot로 분할한다.

[0087] 본원에 따른 방법의 다음단계에서는 분할된 복수개의 각 aliquot를 구분하기 위해, 각 aliquot 별로 각각 상이한 인덱스를 포함하는 어뎁터를 태깅하여 라이브러리를 제조하고, NGS분석을 수행한 후 상기 각 aliquot의 NGS 결과를 통합하는 단계를 포함한다. 본원에서 상기 각 aliquots를 구분하는 인덱스를 tube barcode라고 칭한다. 상이한 인덱스를 포함하는 어뎁터의 선택, 태깅 방법을 포함하는 라이브러리 제조는 채용되는 구체적 플랫폼에 따라 상이할 수 있으며, 본원 실시예 등의 기재를 참조하여 당업자라면 적절한 것을 선택할 수 있다.

[0088] 일 구현예에서는 Illumina와 같은 NGS 플랫폼에서 멀티플렉스 시퀀싱(multiplex sequencing) 방법으로 시퀀싱된다.

[0089] 본원에 따른 방법의 다음 단계에서 각 aliquot의 NGS 데이터는 독립적으로 참조 유전체에 맵핑되고 error correction을 진행한다. 이 단계의 결과물로 각 aliquot 마다 하나의 bam 파일이 생성된다. 이 bam 파일들을 한 개의 bam 파일로 통합한다. 이후 일반적인 변이 분석 프로그램 (Mutect2, Varscan, Vardict, Strelka2, 등)을 이용해 상기 통합된 bam 파일로부터 유전자 변이를 검출한다.

[0090] 또한 다른 양태에서 본원은 NGS 분석을 이용한 cfDNA의 저빈도 변이 검출에 있어서, 상기 NGS 분석에 사용되는 고유 단편의 비율을 향상시키기 위한 라이브러리 제조방법에 관한 것이다.

[0091] 일 구현예에서 상기 방법은 (a) 특정 양의 cfDNA 시료를 제공하는 단계; (b) 상기 특정 양의 cfDNA에 포함된 유전체 단편 중 51bp 내지 330bp 길이에 해당하는 총 유전체 단편의 수, 및 고유 단편 (unique fragment)의 수를 계산하는 단계로, 상기 고유 단편의 수는 상기 총 유전체 단편의 수에서 상기 51~330bp 길이에 해당하는 각 유전체 단편의 collision count의 합을 제외한 값이고, 상기 collision count 합은 식 1로부터 계산되고, (c) 상기 단계 (b)로부터 상기 총 유전체 단편 수에서 상기 고유 단편의 수가 차지하는 비를 계산하고, 상기 고유 단편의 비를 증가시키도록 상기 특정 양의 cfDNA 시료를 복수 개의 aliquot로 나누는 단계; 및 (d) 상기 복수개의 각 aliquot 별로, 각각 상이한 인덱스를 포함하는 어뎁터를 태깅하여 NGS용 라이브러리를 제조하는 단계를 포함한다.

[0092] 상기 방법에 포함된 각 단계는 앞서 언급한 바를 참조할 수 있다.

[0094] 이하, 본 발명의 이해를 돕기 위해서 실시예를 제시한다. 그러나 하기의 실시예는 본 발명을 보다 쉽게 이해하기 위하여 제공되는 것일 뿐 본 발명이 하기의 실시예에 한정되는 것은 아니다.

[0096] **실시예**

[0097] **실시예 1. 상이한 인덱스를 사용한 시료 분할 NGS 분석**

[0098] 본원에서 발견된 것을 다음과 같이 실험으로 증명하였다.

[0099] 실험에 사용한 cfDNA는 SeraseqTM ctDNA mutation mix v2 및 SeraseqTM cfDNA mutation mix v2 WT (SeraCare, Milford, MA)에서 구입하여 사용하였다.

[0100] 실험 디자인은 다음 표와 같다.

[0101] [표 3]

Sample ID	Input DNA (ng)	Reference Material	The number of Aliquot	Dual Index List
LAH001	50	Seraseq® ctDNA Mutation Mix v2 AF1%	4	501-701, 501-702, 501-703, 501-704
LAH002	20	Seraseq® ctDNA Mutation Mix v2 AF1%	4	502-705, 502-706, 502-707, 502-708
LAH003	10	Seraseq® ctDNA Mutation Mix v2 AF1%	4	503-709, 503-710, 503-711, 503-712
LAH004	2	Seraseq® ctDNA Mutation Mix v2 AF1%	4	504-701, 504-702, 504-703, 504-704
LAH005	20	Seraseq® ctDNA Mutation Mix v2 WT	4	505-705, 505-706, 505-707, 505-708
LAH006	20	Seraseq® ctDNA Mutation Mix v2 AF1%	1	501-701
LAH007	20	Seraseq® ctDNA Mutation Mix v2 AF1%	2	502-701, 502-702
LAH008	20	Seraseq® ctDNA Mutation Mix v2 AF1%	4	503-701, 503-702, 503-703, 503-704
LAH009	20	Seraseq® ctDNA Mutation Mix v2 AF1%	8	504-701, 504-702, 504-703, 504-704, 504-705, 504-706, 504-707, 504-708

[0102]

[0103] cfDNA는 TapeStation (Agilent Inc.)을 이용하여 QC (Quality control)를 진행하였고, TapeStation 기준 cfDNA 20 ng을 사용하였다. cfDNA의 양쪽 말단에 A (아데노신)을 결합하고 어뎁터 (Illumina, Inc)를 라이게이션 (ligation)으로 결합하였다. 그리고 극미량의 샘플 swap도 방지할 수 있는 듀얼 인덱스(dual index) 방법 적용을 위해, 5` 말단과 3`말단에 각각 i7과 i5 인덱스를 포함하는 PCR 프라이머 (Illumina, Inc)를 상기 어뎁터에 상보적으로 결합시켰다. 상기 PCR 프라이머는 다음과 같은 공통적인 서열과 각 aliquot를 구별할 수 있는 인덱스 서열([i7]과 [i5]로 표시)이 포함되어 있다.

[0104] 5'-CAAGCAGAAGACGGCATACGAGAT**[i7]**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC-s-T-3'

[0105] 5´-AATGATACGGCGACCACCGAGATCTACAC**[i5]**ACACTCTTTCCCTACACGACGCTCTTCCGATC-s-T-3´

[0106] [표 4]

Index ID	Index Sequence	Index ID	Index Sequence
Dual index 501 primer	TATAGCCT	Dual index 701 primer	ATTACTCG
Dual index 502 primer	ATAGAGGC	Dual index 702 primer	TCCGGAGA
Dual index 503 primer	CCTATCCT	Dual index 703 primer	CGCTCATT
Dual index 504 primer	GGCTCTGA	Dual index 704 primer	GAGATTCC
Dual index 505 primer	AGGCGAAG	Dual index 705 primer	ATTCAGAA
Dual index 506 primer	TAATCTTA	Dual index 706 primer	GAATTCGT
Dual index 507 primer	CAGGACGT	Dual index 707 primer	CTGAAGCT
Dual index 508 primer	GTACTGAC	Dual index 708 primer	TAATGCGC
		Dual index 709 primer	CGGCTATG
		Dual index 710 primer	TCCGCGAA
		Dual index 711 primer	TCTCGCGC
		Dual index 712 primer	AGCGATAG

[0107]

[0108] 일차적으로 만들어진 library는 106개 인간 유전자 (표 5 참조)의 probe (Celemics, Inc. custom panel)를 사용하여 관심이 있는 유전자 만을 캡처하여 최종 library를 기존 방법대로 제조하였다 (Kang, JK et al. Plos one 2020.May).

[0109] [표 5-1]

Gene symbol	Gene full name
ARAF	A-Raf proto-oncogene, serine/threonine kinase
ABL1	ABL proto-oncogene 1, non-receptor tyrosine kinase
AKT1	AKT serine/threonine kinase 1
AKT2	AKT serine/threonine kinase 2
APC	APC, WNT signaling pathway regulator
ARID1A	AT-rich interaction domain 1A
ATM	ATM serine/threonine kinase
BRAF	B-Raf proto-oncogene, serine/threonine kinase
BCR	BCR, RhoGEF and GTPase activating protein
BRCA1	BRCA1, DNA repair associated
BRCA2	BRCA2, DNA repair associated
BTK	Bruton tyrosine kinase
CEBPA	CCAAT/enhancer binding protein alpha
CD274	CD274 molecule
CBL	Cbl proto-oncogene
FBXW7	F-box and WD repeat domain containing 7
GNA11	G protein subunit alpha 11
GNAQ	G protein subunit alpha q
GATA3	GATA binding protein 3
GNAS	GNAS complex locus
HRAS	HRas proto-oncogene, GTPase
JAK2	Janus kinase 2
JAK3	Janus kinase 3
KIT	KIT proto-oncogene receptor tyrosine kinase
KRAS	KRAS proto-oncogene, GTPase
MDM2	MDM2 proto-oncogene
MET	MET proto-oncogene, receptor tyrosine kinase
MPL	MPL proto-oncogene, thrombopoietin receptor
PMS2	PMS1 homolog 2, mismatch repair system component
RB1	RB transcriptional corepressor 1
ROS1	ROS proto-oncogene 1, receptor tyrosine kinase
RAF1	Raf-1 proto-oncogene, serine/threonine kinase
RHEB	Ras homolog enriched in brain
RIT1	Ras like without CAAX 1
SETD2	SET domain containing 2
SMAD4	SMAD family member 4

[0110]

[0111] [표 5-2]

Gene symbol	Gene full name
U2AF1	U2 small nuclear RNA auxiliary factor 1
UGT1A1	UDP glucuronosyltransferase family 1 member A1
ALK	anaplastic lymphoma receptor tyrosine kinase
AR	androgen receptor
CDH1	cadherin 1
CTNNB1	catenin beta 1
CSF1R	colony stimulating factor 1 receptor
CCND1	cyclin D1
CCND2	cyclin D2
CCNE1	cyclin E1
CDK4	cyclin dependent kinase 4
CDK6	cyclin dependent kinase 6
CDKN2A	cyclin dependent kinase inhibitor 2A
DPYD	dihydropyrimidine dehydrogenase
DDR2	discoidin domain receptor tyrosine kinase 2
EGFR	epidermal growth factor receptor
ERBB2	erb-b2 receptor tyrosine kinase 2
ERBB3	erb-b2 receptor tyrosine kinase 3
ESR1	estrogen receptor 1
FGFR1	fibroblast growth factor receptor 1
FGFR2	fibroblast growth factor receptor 2
FGFR3	fibroblast growth factor receptor 3
FLT3	fms related tyrosine kinase 3
IGF1R	insulin like growth factor 1 receptor
IDH1	isocitrate dehydrogenase (NADP(+)) 1, cytosolic
IDH2	isocitrate dehydrogenase (NADP(+)) 2, mitochondrial
KEAP1	kelch like ECH associated protein 1
KDR	kinase insert domain receptor
KDM6A	lysine demethylase 6A
MTOR	mechanistic target of rapamycin
MAPK1	mitogen-activated protein kinase 1
MAPK3	mitogen-activated protein kinase 3
MAP2K1	mitogen-activated protein kinase kinase 1
MAP2K2	mitogen-activated protein kinase kinase 2
MLH1	mutL homolog 1
MSH2	mutS homolog 2

[0112]

[0113] [표 5-3]

Gene symbol	Gene full name
MSH6	mutS homolog 6
NRAS	neuroblastoma RAS viral oncogene homolog
NF1	neurofibromin 1
NF2	neurofibromin 2
NTRK1	neurotrophic receptor tyrosine kinase 1
NTRK2	neurotrophic receptor tyrosine kinase 2
NTRK3	neurotrophic receptor tyrosine kinase 3
NOTCH1	notch 1
NFE2L2	nuclear factor, erythroid 2 like 2
NPM1	nucleophosmin
PTEN	phosphatase and tensin homolog
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
PIK3R1	phosphoinositide-3-kinase regulatory subunit 1
PDGFRA	platelet derived growth factor receptor alpha
PDGFRB	platelet derived growth factor receptor beta
PDCD1LG2	programmed cell death 1 ligand 2
PPP2R1A	protein phosphatase 2 scaffold subunit Aalpha
PTPN11	protein tyrosine phosphatase, non-receptor type 11
RHOA	ras homolog family member A
RET	ret proto-oncogene
RNF43	ring finger protein 43
RUNX1	runt related transcription factor 1
STK11	serine/threonine kinase 11
SMO	smoothened, frizzled class receptor
STAG2	stromal antigen 2
TERT	telomerase reverse transcriptase
TOP2A	topoisomerase (DNA) II alpha
TCF7L2	transcription factor 7 like 2
TSC1	tuberous sclerosis 1
TSC2	tuberous sclerosis 2
TP53	tumor protein p53
MYC	v-myc avian myelocytomatosis viral oncogene homolog
MYCN	v-myc avian myelocytomatosis viral oncogene neuroblastoma derived homolog
VHL	von Hippel-Lindau tumor suppressor

[0114]

[0115] 이어 상기 제조한 라이브러리는 Nextseq550 Dx (Illumina, SanDiego, CA, USA) 장비를 사용하여 2x150 bp paired end 로 sequencing 하였고, bcl2fastq (v2.19.0.316, Illumina Inc.) 프로그램을 이용해 demultiplexing 하여 각 aliquot에 해당하는 fastq 파일을 생성했다. Aliquot마다 forward 방향, reverse 방향의 pair 로 2개의 fastq 파일이 생성된다. 각각 파일은 fastp (version 0.20.0, Shifu Chen et al.)를 사용해 insert (DNA fragment)와 함께 read 말단에 읽힌 adapter 서열을 제거하여 새로운 fastq 파일을 만들었다. 이어서 adapter 서열이 제거된 fastq 파일 (trimmed fastq)을 FastQC (v0.11.8, Babraham Institute)를 사용해 per base sequence quality, overrepresented sequences, adapter content 항목이 'good' 인 경우 QC를 통과한 것으로 판단하여 다음 단계로 진행했다. Trimmed fastq 파일은 GRCh38 버전의 참조유전체에 대하여 bwa (version 0.7.17-r1188)의 BWA-MEM 알고리즘을 사용해 맵핑(mapping)하여 bam 파일을 만들었다. Bam에서 sequencing error를 수정하고 PCR duplicate를 제거하기 위해 gencore (version 0.14.0, Shifu Chen et al.)를 사용했고, 이 결과로 새로운 bam 파일 (collapsed bam)을 만들었다. 이 collapsed bam 파일에 기록된 reads는 error가 수정되고 PCR duplicate가 제거되었기 때문에 혈액에 존재하는 DNA fragment의 정보를 반영한다. 이 과정까지 진행하여 각 샘플별 aliqout로 나눈 개수만큼 bam 파일이 생성되었다. 다음 단계로 진행하기 위해 각 aliquot bam 파일은 sambamba (version 0.7.0, Artem Tarasov et al.)의 merge 기능을 사용해 하나의 bam 파일로 합쳐졌다. 그리고 캡쳐한 106개 유전자에 대해 평균적으로 몇 개의 DNA fragment 정보가 있는지 알기 위해 sambamba (version 0.7.0, Artem Tarasov et al.)의 depth 기능을 이용해 per base depth를 계산하고, 전체

106개 유전자의 base에 대한 평균값을 구해 fragment mean depth(FMD) 값을 구했다.

[0116] 결과는 도 3에 기재되어 있다. 동일한 20ng DNA를 aliquot로 나누지 않거나(1 aliquot), 2개, 4개, 8개로 나눠서 실험했을 때 검사에 사용할 수 있는 고유 단편의 개수가 상이하였다. 고유 단편의 수는 Aliquot 4개로 나눴을 때 포화되고, 이 이상 8개로 나누는 경우에는 큰 효과를 보기는 어려운 것으로 나타났다. 한편 aliquot를 많이 나눌수록 실험적 복잡성이 증가하므로 휴먼 에러를 유발할 가능성도 높아진다. 그러므로, 시작 DNA 양이 20ng인 경우에는 각 aliquot 별로 5ng에 해당하도록 4개의 aliquot로 나눴을 때 ctDNA 정보의 양을 최적으로 얻을 수 있다.

[0117] 또한 Aliquots 로 분할하여 얻은 DNA 정보를 이용해 본원에 따른 방법으로 분석을 하면 그렇지 않은 경우 (모든 duplicate 제거)보다 FMD (Fragment mean depth) 값이 높아지는 것을 확인하였다 (도 4 및 표 7 참조). 즉 동일한 시작 DNA를 사용하여 더 많은 DNA 정보를 검사에 사용할 수 있다. FMD는 NGS 검사 영역에서 대하여 고유 단편을 맵핑 했을 때 계산되는 평균적인 시퀀싱 정도(sequencing depth)로 그 값을 표 6에 나타냈다.

[0118] [표 6]

[0119]

Aliquot	Fragment mean depth (FMD)
1	1,572
2	1,487
4	1,967
4	1,908
4	1,835
8	1,963

[0120] [표 7]

[0121]

#Aliquot	FMD	Aliquot 1 개 대비 FMD 증가
1	1,572	-
2	1,487	-5.4%
4	1,967	25.1%
8	1,963	24.9%

[0122] 기존의 library를 제작하는 방법은 앞에서 기술한 방법과 유사하다. 단 차이점은 본원에 따른 방법에서는 하나의 샘플을 4개의 aliquot로 나눈 뒤 각각 다른 index를 사용하나, 기존의 방법은 하나의 샘플을 하나의 tube로 실험하기 때문에 하나의 인덱스만을 사용하여 라이브러리를 제작하였다.

[0123] 또한 표 8을 참조하면, 동일한 input DNA 양 (20ng) 일 때, aliquot의 개수를 늘리는 만큼 개별 aliquot당 DNA 양은 줄어들고, 각 aliquot의 DNA는 pre-PCR 단계에서 증폭되는데, 1ng 당 pre-PCR 양은 aliquot에 5ng의 DNA가 있을 때 포화된다. 결과적으로 전체 pre-PCR DNA 양은 aliquot 개수를 늘릴수록 증가하고 4개로 늘렸을 때 포화되는 것으로 나타났다 (도 5 참조). 이 것은 FMD 값이 aliquots 4개인 경우 포화되는 것과 비슷한 현상으로 aliquots를 4개로 나눌 때의 특장점으로 볼 수 있다. 부가적으로 한 번의 NGS 실험에서는 증폭된 DNA 중 1000~2000 ng을 사용하기 때문에, aliquot를 나누지 않은 경우엔 1회 분석 분량의 DNA밖에 얻을 수 없는 반면, aliquot로 나누어서 증폭한 경우, 4000~6000 ng를 얻을 수 있기 때문에, 향후 검증과정에서 재검사에 활용하거나, 다른 실험에 사용할 수 있는 추가의 장점이 생긴다.

[0124] [표 8]

ID	Input DNA (ng)	#Aliquot	DNA 양(ng)/ aliquot	pre-PCR DNA 양(ng)/ aliquot	pre-PCR DNA 양/ng	Total pre-PCR DNA 양 (ng)
LAH006	20	1	20	1254	62.7	1254
LAH007	20	2	10	1355	135.5	2710
LAH002	20	4	5	1070	214	4280
LAH005	20	4	5	1572.5	314.5	6290
LAH008	20	4	5	1400	280	5600
LAH009	20	8	2.5	622.5	249	4980

[0125]

[0127] **실시예 2. 본원에 따른 방법을 이용한 분석에서 오류 수정 전/후 변이의 VAF (Variant allele frequency) 향상**

[0128] cfDNA 분석에 있어서 molecular barcode를 사용한 경우, barcode를 사용해 동일한 세포에서 유래한 DNA의 PCR duplicates로 확인되면 서로 염기서열을 비교하고, 다른 부분이 있을 경우 다수결의 원칙으로 오류가 생긴 염기를 수정하고 전체 PCR duplicates를 대표하는 하나의 consensus DNA 서열을 만든다. 이 방법으로 error rate를 1/10000 (10e-4)까지 낮출 수 있다.

[0129] 본원에 따른 방법은 적절하게 aliquot로 나눈 경우 서로 다른 세포에서 온 DNA가 우연히 생기지 않는 것으로 가정하고, 모든 duplicates를 PCR duplicates로 가정하여 molecular barcode를 사용할 때와 유사하게 다수결 원칙으로 consensus DNA를 만든다. 하지만 실제로 다른 세포 유래의 우연히 동일한 DNA가 있을 수 있기 때문에 오류 수정 과정에서 서로 다른 염기가 많은 경우 분석에서 제외시킨다.

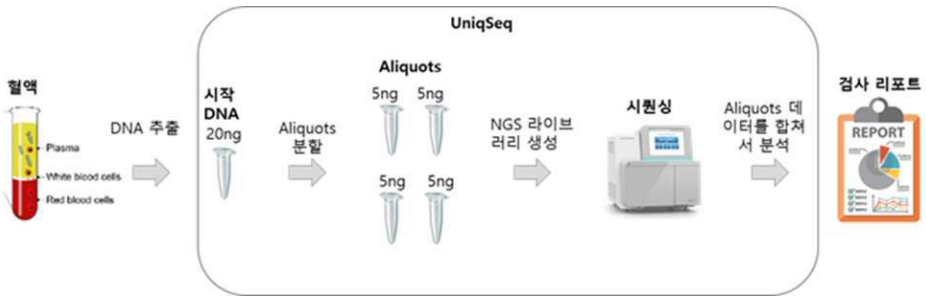
[0130] 한편, DNA를 aliquot로 나누는 것 자체는 반복실험의 효과가 있다. 이 점을 이용해 aliquots 사이에 통계적으로 유의미한 불일치가 발생하는 경우 분석 결과에서 제외한다. 결과적으로 error rate를 10배 더 낮춰서 1/100000 (10e-5)까지 낮출 수 있었다. 도 6a 및 도 6b에 나타난 바와 같이 오류 수정 전 변이는 VAF 1%에서 다수의 false positive (FP) 변이가 검출된다 (도 6a). 반면 오류 수정 후 true positive (TP)만 남고 모든 FP가 사라졌다 (도 6b).

[0132] 이상에서 본원의 예시적인 실시예에 대하여 상세하게 설명하였지만 본원의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본원의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본원의 권리범위에 속하는 것이다.

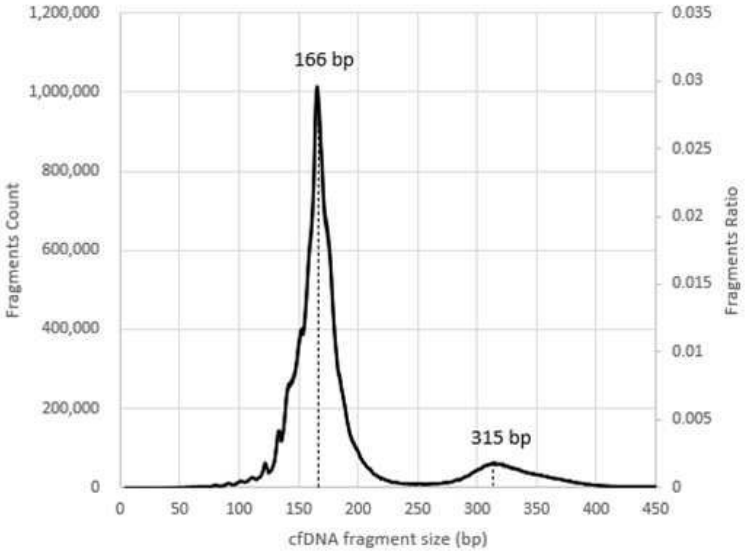
[0134] 본 발명에서 사용되는 모든 기술용어는, 달리 정의되지 않는 이상, 본 발명의 관련 분야에서 통상의 당업자가 일반적으로 이해하는 바와 같은 의미로 사용된다. 본 명세서에 참고문헌으로 기재되는 모든 간행물의 내용은 본 발명에 도입된다.

도면

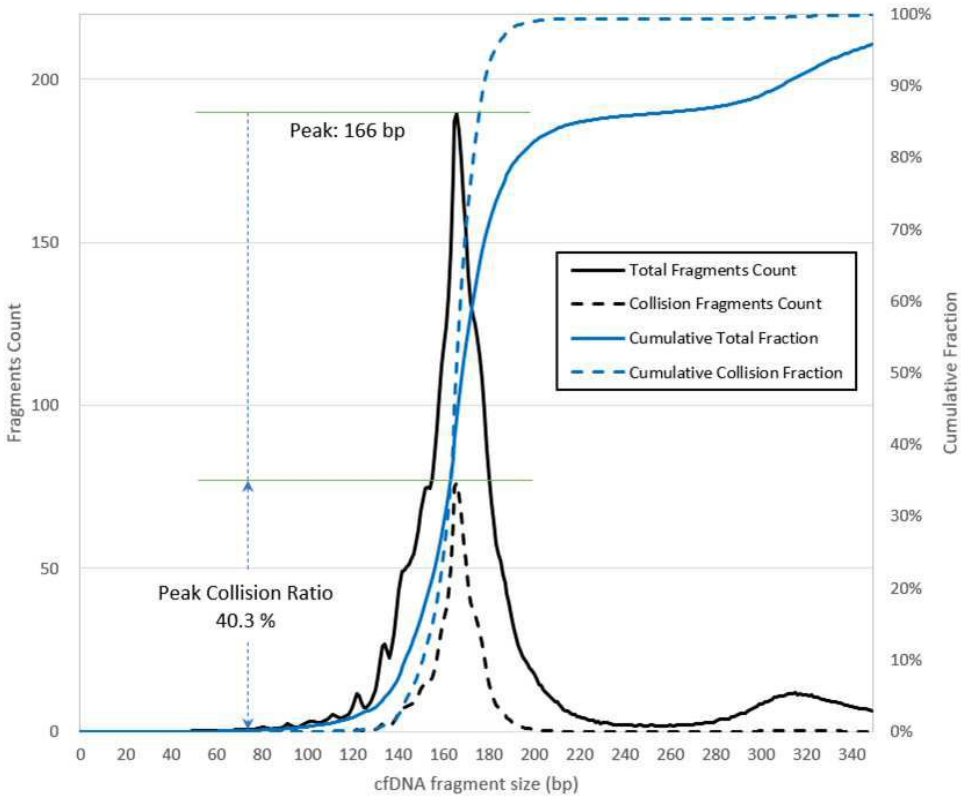
도면1



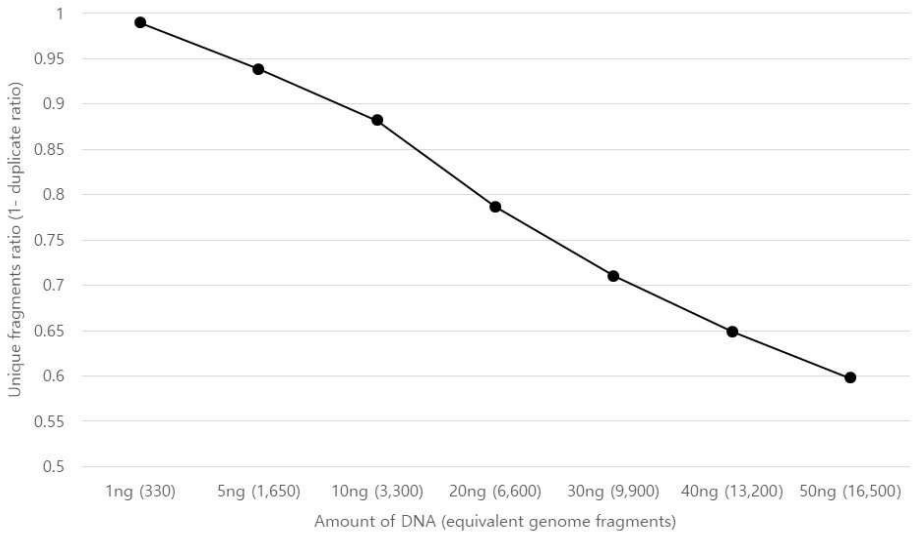
도면2a


166 bp
315 bp
Fragments Count
Fragments Ratio
cfDNA fragment size (bp)
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0 50 100 150 200 250 300 350 400 450

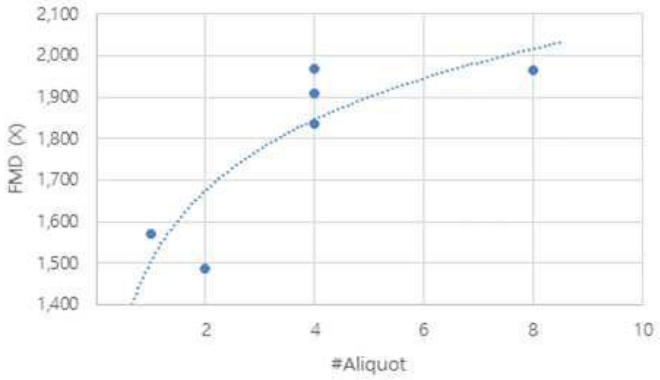
도면2b


Peak: 166 bp
Peak Collision Ratio
40.3 %
Total Fragments Count
Collision Fragments Count
Cumulative Total Fraction
Cumulative Collision Fraction
Fragments Count
Cumulative Fraction
cfDNA fragment size (bp)
200
150
100
50
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340

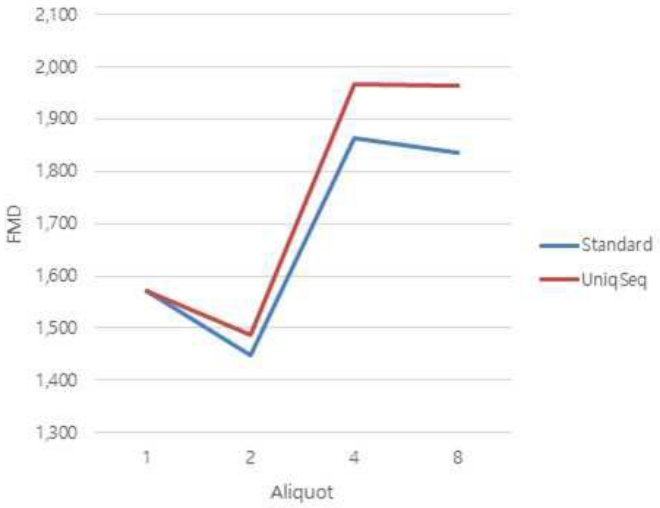
도면2c


1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Unique fragments ratio (1- duplicate ratio)
1ng (330)
5ng (1,650)
10ng (3,300)
20ng (6,600)
30ng (9,900)
40ng (13,200)
50ng (16,500)
Amount of DNA (equivalent genome fragments)

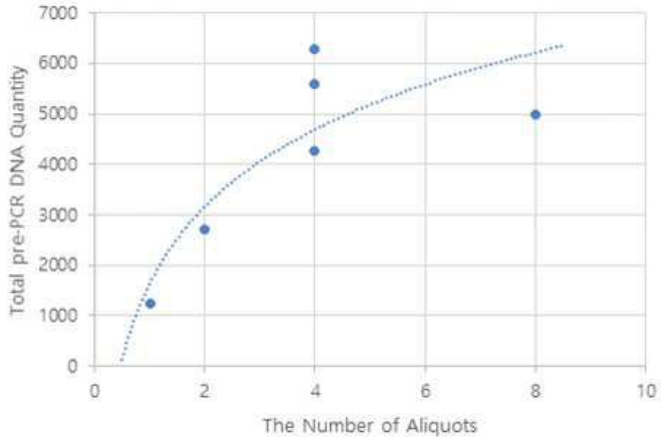
도면3


2,100
2,000
1,900
1,800
1,700
1,600
1,500
1,400
FMD (X)
2
4
6
8
10
#Aliquot

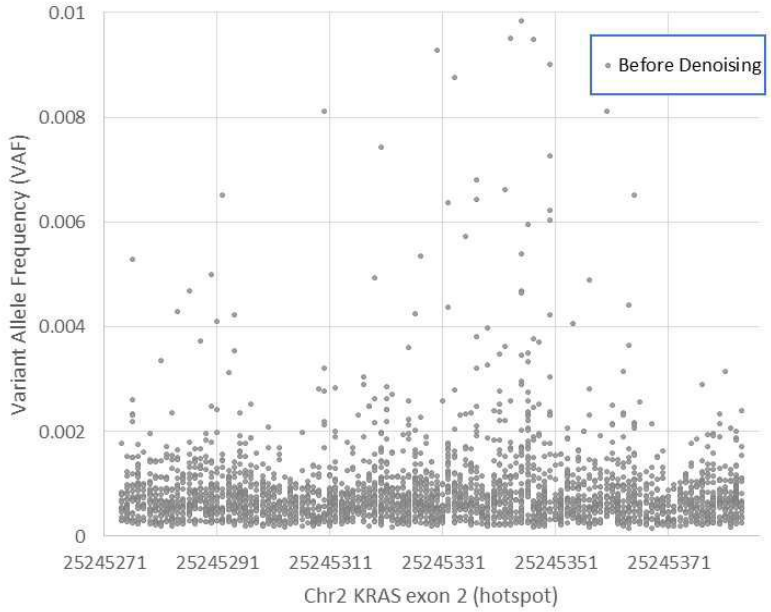
도면4


2,100
2,000
1,900
1,800
1,700
1,600
1,500
1,400
1,300
FMD
1
2
4
8
Aliquot
Standard
UniqSeq

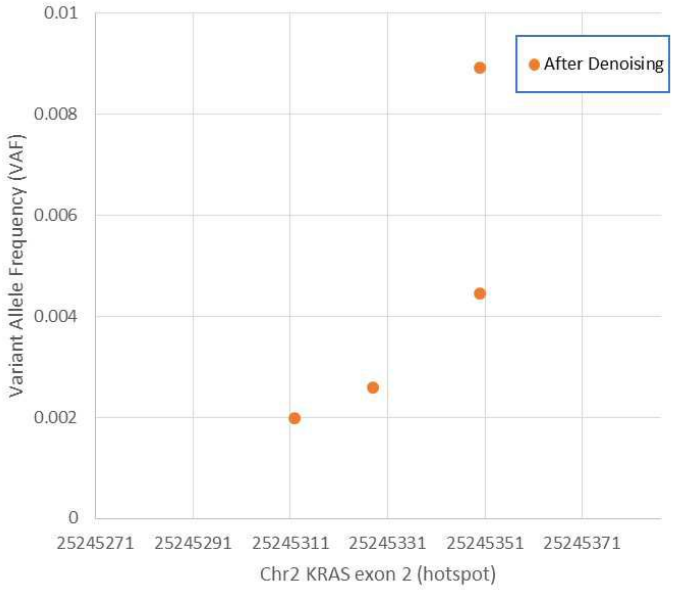
도면5


Total pre-PCR DNA Quantity
7000
6000
5000
4000
3000
2000
1000
0
0
2
4
6
8
10
The Number of Aliquots

도면6a


Variant Allele Frequency (VAF)
0.01
0.008
0.006
0.004
0.002
0
Before Denoising
25245271
25245291
25245311
25245331
25245351
25245371
Chr2 KRAS exon 2 (hotspot)

도면6b


0.01
0.008
0.006
0.004
0.002
0
Variant Allele Frequency (VAF)
After Denoising
25245271
25245291
25245311
25245331
25245351
25245371
Chr2 KRAS exon 2 (hotspot)