

명세서

발명의 명칭: 차세대 염기서열 분석기법을 이용한 유전자 재배열 검출 방법

기술분야

- [1] 본 발명은 차세대 염기서열 분석기법(Next generation sequencing, NGS)를 기반으로 유전자 재배열을 검출하는 방법에 관한 것으로, 더욱 상세하게는 NGS에 의해 생성되는 리드 데이터를 정렬하고, 추출한 다음, 추출한 리드 데이터의 서열 유사성을 분석하여 암 샘플 내에 존재하는 유전자 재배열을 검출하고, 나아가 유전자 재배열의 방향, 마이크로호몰로지 서열, 외부 삽입 서열 및 위치를 검출하는 방법에 관한 것이다.

[2]

배경기술

- [3] 유전자에 발생하는 돌연변이의 종류는 다양하다. DNA의 염기서열 중 일부가 바뀌거나 삭제 또는 삽입되는 경우, 유전자의 숫자가 늘어나거나 줄어드는 경우가 있으며, 정상세포에는 존재하지 않는 유전자들끼리의 새로운 만남이 생길 수 있다. 이러한 유전자끼리의 만남은 특정부위의 DNA가닥이 외부자극에 의해 끊어졌다가 애기치 않은 위치에서 다시 이어져 발생한다. 이렇게 서로 다른 염색체 안에 들어 있는 두개의 유전자가 위치를 이동해 융합 유전자를 만들어내는 것은 이질 염색체 간 융합 유전자(Inter-chromosomal fusion gene) 이라하며, 같은 염색체지만 따로 떨어져 있는 두 개의 유전자가 위치이동을 해 융합 유전자를 만드는 것을 동일 염색체 내 융합 유전자(Intra-chromosomal fusion gene)이라 한다(Rabbitts, T.H. et al. Nature, Vol. 372, pp. 143-149, 1994).
- [4] 대부분의 융합 유전자는 세포의 생존에 치명적인 영향을 미치게 되어 세포단계에서 사멸하게 되지만 우연한 조합으로 만들어진 유전자가 여러 가지 조건이 들어맞는 필연적인 환경이 부가되는 경우, 비정상적인 기능을 가지게 되어 살아남게 되는 경우가 생긴다. 원암유전자(proto-oncogene)의 앞부분(upstream)에 강력한 프로모터를 가지는 유전자가 결합하게 되면 발현양이 급증하거나 상시 발현되는 융합 유전자가 된다. 이러한 융합 유전자는 근래의 연구로 인해 혈액암, 폐암, 대장암, 정신분열증 등의 병을 일으키는 종양유전자(oncogene)으로 작용한다는 것이 알려져 있다(David, R. et al. Cell Physiol Biochem. Vol. 37(1), pp.77-93, 2015). 비소세포암(NSCLC) 중 특히 폐선암(LADC)에서 4%의 빈도로 발생하며, 주로 젊고 비흡연환자에게서 나타나는 EML4-ALK (Echinoderm microtubule-associated protein like 4 - Anaplastic lymphoma kinase) 재배열변이는 2번 염색체의 역전(inversion)에 의한 융합 유전자이다. EML4의 길이에 따라 다양한 EML4-ALK 융합 유전자가 존재하며 TRK-fuses gene(TFG), kinesin light chain(KLC-1), kinesin family member 5b(KIF5B)

와 ALK 유전자의 융합도 보고되어 있다. 상위 유전자에 의해 ALK가 활성화되면 RAS, PI3K, JAK-STAT3등의 신호전달체계를 통해 세포 증식을 증가시키고 세포자멸사(apoptosis)를 억제하여 암의 원인으로 작용하게 된다. (Takashi, K. et al. *Ann Oncol.* Vol. 25(1), pp. 138-42. 2014).

- [5] 체세포상의 유전체 재배열을 NGS 데이터를 통해 알아내는 것은 융합 유전자에 의한 암 발생 원인을 찾는 데 매우 중요하다. 유전체 재배열이나 구조적 변이(Structural variants, SVs)를 찾아내기 위해서는 일반적으로 페어드 엔드(Paired-End) 시퀀싱을 사용한다. 이러한 시퀀싱 데이터에서 유전체 재배열을 알아낼 수 있는 몇 가지 방법이 있다. 그 중 리드 깊이(Read depth, RD) 방식으로는 주로 유전자 복제수 변이 정보를 얻어 낼 수 있으며, 커버 깊이(depth of coverage)와 윈도우 크기(window size)에 의해 해상력이 결정되며, 싱글 엔드 리드(Single end read)와 페어드 엔드 리드(Paired end read) 모두에서 사용가능하다. 하지만 상기 방법은 실제적으로 융합 유전자가 만들어지는 융합 유전자의 위치(fusion break point)는 자세히 알 수 없다는 단점이 있다(Feuk, L. et al. *Nat Rev Genet.* Vol. 7(2), pp. 85-97. 2006).
- [6] 비정상 페어드 엔드(Paired-End) 정렬 방식은 메이트 리드(mate-read)가 다른 유전자, 다른 염색체에 맵핑 되어 있는 불일치 리드 쌍(discordant read pair)을 통해 융합 유전자의 위치(fusion break point)를 찾아내는 방식이다. 융합 유전자의 위치(Break point)가 리드-페어(read-pair) 사이 서열분석이 되지 않는 위치(unsequenced region)에 있는 경우에 사용되며, 이 방식은 리드의 조각 크기(fragment size)와 커버리지(coverage)에 의해 해상력이 결정된다(Chiang, D.Y. et al. *Nat Methods.* Vol.6(1), pp. 99-103 2009).
- [7] 스플릿 리드(Split read, SR) 방식은 리드 내부에 융합 유전자의 위치(fusion break point)가 존재할 때 리드를 참조 유전체(reference genome)에 맵핑시 리드의 일부분을 정렬 프로그램에서 소프트-클리핑(Soft-clipping) 처리하는 것을 이용해 위치(break point)를 찾아내는 방법이다. 싱글 엔드 리드(Single end read)와 페어드 엔드 리드(Paired end read) 모두에서 사용가능하며 페어드-엔드 리드를 사용할 때 보다 정확한 결과를 얻을 수 있다. 하지만 소프트-클리핑 서열(soft-clipped sequence)부분이 시퀀싱 에러에 의해서도 발생 할 수 있고, 마이크로호몰로지(micro-homology)에 의해서도 발생 할 수 있다는 단점이 있다(Chen, K. et al. *Nat Methods.* Vol.6(9) pp.677-81. 2009).
- [8]
- [9] 이에, 본 발명자들은 상기 방식들이 가지는 문제점을 해결하기 위하여 예의 노력한 결과, 리드를 불일치 리드 쌍과 일치 서열 쌍으로 분류한 다음, 각각의 리드를 쿼리로 하여 비교 분석하는 페어-블라스트(pair-blast) 분석을 수행할 경우, 다른 방법으로는 검출할 수 없었던 유전자 재배열을 검출할 수 있을 뿐만 아니라, 유전자 재배열의 위치, 마이크로호몰로지 서열, 외부 삽입 서열 및 유전자 재배열의 방향을 검출할 수 있다는 것을 확인하고, 본 발명을 완성하게

되었다.

[10]

[11] **발명의 요약**

[12] 본 발명의 목적은 차세대 염기서열 분석기법(Next Generation Sequencing, NGS)을 이용한 유전자 재배열의 검출 방법을 제공하는데 있다.

[13] 본 발명의 다른 목적은 차세대 염기서열 분석기법(Next Generation Sequencing, NGS)을 이용한 유전자 재배열의 검출 방법을 수행할 수 있도록 컴퓨팅 시스템을 제어하기 위한 복수의 명령이 암호화된 컴퓨터 리드 가능한 매체를 포함하는 컴퓨터 시스템을 제공하는데 있다.

[14] 상기 목적을 달성하기 위하여, 본 발명은 다음의 단계를 포함하는 샘플 내 유전자 재배열 검출 방법을 제공한다: (a) 피험체의 샘플로부터 복수의 핵산 분자를 포함하는 라이브러리를 획득하는 단계; (b) 상기 라이브러리를 복수의 베이트(bait) 세트와 접촉시킴으로써, 사전선택된 서열에 대해 상기 라이브러리를 풍부화시켜서 선택된 핵산 분자를 제공하고, 이로 인해 라이브러리 캐치(library catch)를 제공하는 단계; (c) 차세대 시퀀싱 방법에 의해, 상기 라이브러리 캐치의 핵산 분자로부터 리드를 획득하는 단계; (d) 정렬 방법에 의해 상기 리드를 정렬하는 단계; 및 (e) 상기 정렬된 리드를 분석하여 유전자 재배열을 검출하는 단계, 여기서 상기 분석 방법은 상기 정렬된 리드를 추출하여 서열 유사성을 분석하는 단계를 포함함.

[15] 본 발명은 또한, 차세대 염기 서열 분석(Next Generation Sequencing, NGS)을 이용한 유전자 재배열 검출 방법을 수행할 수 있도록 컴퓨팅 시스템을 제어하기 위한 복수의 명령이 암호화된 컴퓨터 리드 가능한 매체를 포함하는 컴퓨터 시스템으로서,

[16] 상기 방법은 (a) 피험체의 샘플로부터 복수의 핵산 분자를 포함하는 라이브러리를 획득하는 단계; (b) 상기 라이브러리를 복수의 베이트(bait) 세트와 접촉시킴으로써, 사전선택된 서열에 대해 상기 라이브러리를 풍부화시켜서 선택된 핵산 분자를 제공하고, 이로 인해 라이브러리 캐치(library catch)를 제공하는 단계; (c) 차세대 시퀀싱 방법에 의해, 상기 라이브러리 캐치의 핵산 분자로부터 리드를 획득하는 단계; (d) 정렬 방법에 의해 상기 리드를 정렬하는 단계; 및 (e) 상기 정렬된 리드를 분석하여 유전자 재배열을 검출하는 단계, 여기서 상기 상기 정렬된 리드를 추출하여 서열 유사성을 분석하는 단계를 포함하는 것인 컴퓨터 시스템을 제공한다.

[17]

도면의 간단한 설명

[18] 도 1은 본 발명의 유전자 재배열 검출 기작에 대한 개략도이다.

[19] 도 2는 본 발명의 일 실시예에서 추출한 리드의 종류에 대한 모식도이다.

[20] 도 3은 본 발명의 일 실시예에서 일치 쌍 및 불일치 쌍을 분석하는 방법에 대한

모식도이다.

- [21] 도 4는 본 발명의 일 실시예에서 유전자 재배열을 검출한 방법에 대한 모식도이다.
- [22] 도 5은 본 발명의 융합 유전자 검출 방법의 전체 과정을 도식화 한 것이다.
- [23] 도 6은 본 발명의 일 실시예에 따른 유전자 재배열의 각 과정을 순서도로 도식화 한 것이다.
- [24]
- [25] **발명의 상세한 설명 및 바람직한 구현예**
- [26] 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 명명법은 본 기술 분야에서 잘 알려져 있고 통상적으로 사용되는 것이다.
- [27]
- [28] 본 발명에서의 용어 “차세대 염기서열 분석기법” 또는 “NGS” 또는 “차세대 염기서열 분석”은 개개의 핵산 분자(예를 들어 단일 분자 시퀀싱에서) 또는 고속 대량 방식으로(예를 들어, 10, 100, 1000 이상의 분자가 동시에 시퀀싱됨) 개개의 핵산 분자에 대해 클론으로 확장된 프록시(proxy) 중 하나의 뉴클레오타이드 서열을 결정하는 임의의 시퀀싱 방법을 지칭한다. 일 실시형태에서, 라이브러리 내 핵산 종의 상대적 존재비는 시퀀싱 실험에 의해 만들어진 데이터에서 그것의 동족 서열의 발생의 상대적인 수를 계측함으로써 추정될 수 있다. 차세대 시퀀싱 방법은 당업계에 공지되어 있으며, 예를 들어 Metzker, M. (2010) Nature Biotechnology Reviews11:31-46]에 기재된다. 차세대 시퀀싱은 샘플 내 핵산의 5% 미만으로 존재하는 변이체를 검출할 수 있다.
- [29] 본 발명에서 차세대 염기서열 분석 과정은 하기의 3단계로 구분될 수 있다.
- [30] (1) 타겟의 포획
- [31] 질병의 원인 유전자를 찾기 위하여 차세대 염기서열 분석법을 이용해 전장유전체(Whole-genome)를 시퀀싱하거나, 엑솜 영역만을 목표로 하여 시퀀싱할 수 있으며(Targeted sequencing), 특정 유전자를 타겟으로 수행할 수도 있다. 엑솜 영역또는 특정 타겟 유전자만을 시퀀싱하는 경우에는 비용이나 효율성 측면에서 유리하다. 또한 유전자의 변화가 암과 같은 직접적인 질병으로 나타나는 경우가 많기 때문에 엑솜 영역 또는 타겟 유전자에서의 염기서열의 변화를 검출하는 것이 원인 유전자를 찾는 데 효과적이라고 할 수 있다. 엑솜 또는 타겟 유전자만을 시퀀싱하기 위해서는 엑솜 또는 타겟 유전자만 포획할 수 있는 라이브러리가 필요하다.
- [32] 엑솜 영역 포획을 위해 주로 사용되는 라이브러리는 SureSelect Human All Exon Kits([http:// www.genomics.agilent.com](http://www.genomics.agilent.com))이나, 이에 한정하는 것은 아니다. SureSelect Human All Exon Kits는 CCDS(Consensus CDS, NCBI, EBI, UCSC, Wellcome Trust Sanger Institute가 참여하여 정의한 인간 유전체의 유전자 세트)

엑손을 기초로 디자인 되었으며, 인간유전체의 1.22%에 해당하는 영역을 포함하고 있다.

- [33] 타겟 유전자만을 포획하기 위해서는 특정 타겟 유전자에 특이적인 프로브 또는 베이트를 이용할 수 있다.
- [34] (2) 대용량 병렬 DNA 시퀀싱
- [35] 차세대 염기서열 분석기법(Next Generation Sequencing: NGS)은 기존의 모세관 서열확인법(capillary sequencing)에 비해서 빠르면서 한 번에 더 많은 양의 서열확인을 수행할 수 있고, 기존의 모세관 서열확인법에 사용하는 벡터를 이용한 시료의 증폭 과정이 생략되기 때문에 이 과정에서 발생하는 실험적인 오류를 피할 수 있다는 장점이 있다.
- [36] 3곳의 회사에서 제작한 NGS 시스템이 주로 사용되고 있다. 2004년에 출시된 로슈(Roche)사의 454 GS FLX는 처음 소개된 NGS 장비로, 이 장치는 피로시퀀싱(pyrosequencing) 방법과 유화제-중합효소반응(emulsion polymerase chain reaction)을 사용하여 서열확인을 수행하고, 실험의 최종단계에서 나오는 빛의 세기에 따라서 특정 염기를 확인할 수 있다. 7시간 가동시켰을 때 100Mb 정도의 서열을 확인할 수 있는데, 기존의 ABI 3730 기기가 같은 시간에 440kb의 서열을 확인할 수 있는 것에 비해서 월등히 높은 성능을 나타낸다.
- [37] 일루미나(Illumina)사의 Illumina Genome Analyzer는 합성에 의한 서열확인(sequencing by synthesis)이라는 개념을 도입한 것으로, 유리판 위에 한 가닥만으로 이루어진 DNA 조각을 부착한 후에, 이 조각들을 중합반응을 거쳐서 군집(cluster)을 이루게 한다. 이 과정을 거칠 때 검사하려는 DNA 조각에 붙은 염기의 종류를 확인하면서 서열 분석을 수행하는데, 약 4 일 정도의 작업으로 32-40 개의 염기길이를 가지는 단편이 4-5천만 개가 생산이 된다.
- [38] 라이프 테크놀로지(Life Technologies)사의 SOLiD (Sequencing by Oligo Ligation) 기기는 1 μ m 크기의 자성 구슬에 검사하려는 DNA 조각을 부착시킨 후에 유화제-중합효소연쇄반응을 이용하여 서열확인을 수행한다. 서열 확인을 할 때는 8-mer의 단편들을 반복해서 붙이는 방식을 사용하는데, 이 8-mer의 4, 5번째에 실제 서열확인에 사용될 염기가 위치하고 있다. 그 뒤에 붙은 나머지 부위에는 형광물질이 연결되어 있어서 어느 염기가 검사하려는 DNA 조각에 상보적으로 결합하는 지를 표시해 준다. 한 번의 결합 주기마다 8-mer를 모두 5번 붙이고, 같은 작업을 5번 시행하면 총 25염기로 이루어진 DNA 조각의 서열을 확인할 수 있다. SOLiD 기기의 특징은 두 개의 염기를 이용한(two-base encoding) 서열확인으로, 이 방법은 하나의 염기의 서열을 결정할 때 같은 부위를 두 번의 서열확인을 통해서 확인하는 것이다. 자성구슬에 부착된 부착제(adaptor)쪽으로 한 번의 결합 주기마다 한 염기씩 서열을 이동시키면서 서열확인을 수행한다. 이 과정을 통해서 서열확인 실험에서 발생하는 오류를 제거할 수 있는 장점이 있다.
- [39] (3) 염기서열 데이터의 분석
- [40] 질병의 원인 유전자를 찾기 위해서는 기존의 유전자 염기서열로부터 어떤

변화가 일어났는지 조사해야 하기 때문에 개인(환자)의 염기서열 데이터(sequence reads)를 레퍼런스 염기서열(reference Genome)과 비교하는 작업을 하게 된다. 이 작업을 맵핑(Mapping)이라고 한다. 맵핑을 통해 개인과 레퍼런스 염기서열의 차이를 알아낸 후 이를 적당한 선택 기준을 정해 신뢰할 수 있는 염기서열 변이 정보만 추출(Variant Calling)하게 된다. 이 변이 정보는 단일염기서열변이(SNV: Single Nucleotide Variation), 짧은 삽입/결실(Short Indel), 복제수 변이(copy number variation, CNV) 및 융합 유전자 등을 포함하는 구조변이(structural variation, SV)이다. 그런 다음 염기서열 변이 정보를 기존 데이터베이스와 비교하여 이미 밝혀진 변이인지 새롭게 발견된 변이인지 판단한다. 그리고 그 변이가 아미노산의 변화를 가져올 것인지 아닌지, 또한 단백질 구조에 있어서 어떤 영향을 줄 것인지 예측하게 된다. 이 과정을 주석달기(Annotation)라고 한다. 추출한 단일염기서열변이와 짧은 삽입/결실에 관한 정보는 정보의 품질을 더 높이기 위하여 데이터베이스에 등재하거나 전장유전체연관분석(Genome Wide Association Study; GWAS)과 통합 연구를 통해 질병의 원인 변이를 찾는 연구를 수행할 수도 있다.

[41] 다만, 기존의 방법으로는 콜링하는 변이 정보에 있어서, SNV, Indel, CNV 등의 변이 정보는 높은 정확도로 추출할 수 있었으나, 구조 변이에 대해서는 그 정확도가 낮은 단점이 있었으며, 본원 발명에서는 구조 변이 중, 특히 융합 유전자에 대한 변이 정보를 높은 정확도로 추출하는 방법을 개발한 것이다.

[42]

[43] 본 발명에서의 용어 "암" 또는 "종양"은 본 명세서에서 상호 호환적으로 사용된다. 이들 용어는 암-원인 세포의 전형적인 특징, 예컨대 제어되지 않는 증식, 불멸, 전이 가능성, 빠른 성장 및 증식 속도 및 어떤 특징적 형태학적 특징을 소유하는 세포의 존재를 지칭한다. 암 세포는 종종 종양 형태이지만, 이러한 세포는 동물 내에서 단독으로 존재할 수 있거나 또는 비-종양 암 세포, 예컨대 백혈병 세포일 수 있다. 이들 용어는 고형 종양, 연조직 종양 또는 전이성 병변을 포함한다. 본 명세서에서 사용되는 용어 "암"은 전암성뿐만 아니라 악성 암을 포함한다.

[44] 본 발명에서의 용어 "샘플", "조직 샘플", "암 샘플", "환자 샘플", "환자 세포 또는 조직 샘플" 또는 "표본"은 각각 피험체 또는 환자의 조직 또는 순환 세포로부터 얻은 유사한 세포의 수집을 지칭한다. 조직 샘플의 공급원은 신선한, 냉동 및/또는 보존된 기관, 조직 샘플, 생검 또는 흡입으로부터의 고형 조직; 혈액 또는 임의의 혈액 구성요소; 체액, 예컨대 뇌척수액, 양수, 복막액 또는 세포간질액; 또는 피험체의 임신 또는 발생에서 어느 시점으로부터의 세포일 수 있다. 조직 샘플은 자연에서 조직과 자연적으로 상호혼합되지 않는 화합물, 예컨대 보존제, 항응고제, 완충제, 정착제, 영양제, 항생제 등을 함유할 수 있다. 일 실시형태에서, 샘플은 냉동 샘플로서 또는 포름알데하이드- 또는 파라포름알데하이드-고정 파라핀-포매(paraformaldehyde-fixed

paraffin-embedded: FFPE) 조직 제조물로서 제조된다. 예를 들어, 샘플은 매트릭스, 예를 들어 FFPE 블록 또는 냉동 샘플에서 포매될 수 있다.

- [45] 일 실시형태에서, 샘플은 암 샘플이며, 예를 들어, 하나 이상의 전암성 또는 악성 세포를 포함한다. 특정 실시형태에서, 샘플, 예를 들어 종양 샘플은 고형 종양, 연조직 종양 또는 전이성 병변으로부터 획득된다. 다른 실시형태에서, 샘플, 예를 들어 종양 샘플은 수술절제면으로부터 조직 또는 세포를 포함한다. 다른 실시형태에서, 샘플, 예를 들어 종양 샘플은 하나 이상의 혈중 종양 세포(CTC)(예를 들어, 혈액 샘플로부터 획득한 CTC)를 포함한다.

[46]

- [47] 본 발명에서의 용어 "획득하다" 또는 "획득하는"이 본 명세서에서 사용되며, 물리적 독립체 또는 값을 "직접적으로 획득하거나" 또는 "간접적으로 획득함으로써" 물리적 독립체 또는 값, 예를 들어 수치적 값의 소유를 얻는 것을 지칭한다. "간접적으로 획득하는"은 물리적 독립체 또는 값을 얻기 위한 처리를 수행하는 것(예를 들어, 합성 또는 분석 방법을 수행하는 것)을 의미한다. "간접적으로 획득하는 것"은 다른 관계자 또는 공급원(예를 들어 물리적 독립체 또는 값을 직접적으로 획득한 제3자 연구소)으로부터 물리적 독립체 또는 값을 수용하는 것을 지칭한다.

[48]

물리적 독립체를 간접적으로 획득하는 것은 물리적 물질, 예를 들어 출발 물질에서 물리적 변화를 포함하는 처리를 수행하는 것을 포함한다. 대표적인 변화는 2 이상의 출발 물질로부터 물리적 독립체를 만드는 것, 물질을 전단(shearing) 또는 단편화하는 것, 물질을 분리시키거나 정제하는 것, 2 이상의 별개의 독립체를 혼합물로 합하는 것, 공유 또는 비공유 결합을 파괴하거나 또는 형성하는 것을 포함하는 화학 반응을 수행하는 것을 포함한다. 값을 간접적으로 획득하는 것은 샘플 또는 다른 물질에서 물리적 변화를 포함하는 처리를 수행하는 것, 예를 들어 물질, 예를 들어 샘플, 분석물 또는 시약에서 물리적 변화를 포함하는 분석 과정을 수행하는 것(때때로, 본 명세서에서 "물리적 분석"으로서 지칭됨), 분석 방법, 예를 들어 다음 중 하나 이상을 포함하는 방법을 수행하는 것: 물질, 예를 들어 분석물 또는 이것의 단편 또는 다른 유도체를 다른 물질로부터 분리시키거나 또는 정제하는 것; 분석물 또는 이것의 단편 또는 다른 유도체를 다른 물질, 예를 들어 완충제, 용매 또는 반응물과 합하는 것; 또는, 예를 들어 분석물의 제1 원자와 제2 원자 사이의 공유 또는 비공유 결합을 파괴하거나 또는 형성함으로써 분석물 또는 이것의 단편 또는 다른 유도체의 구조를 변화시키는 것; 또는, 예를 들어 시약의 제1과 제2 원자 사이의 공유 또는 비공유 결합을 파괴하거나 형성함으로써 시약 또는 이것의 단편 또는 다른 유도체의 구조를 변화시키는 것을 포함한다.

[49]

본 발명에서의 용어 "서열을 획득하는 것" 또는 "리드를 획득하는 것"은 본 명세서에서 사용되며, 서열 또는 리드를 "직접적으로 획득하거나" 또는 "간접적으로 획득함으로써" 뉴클레오타이드 서열 또는 아미노산 서열의 소유를

얻는 것을 지칭한다. 서열 또는 리드를 "직접적으로 획득하는 것"은 시퀀싱 방법(예를 들어, 차세대 시퀀싱(NGS) 방법)을 수행하는 것과 같이 서열을 얻기 위한 과정을 수행하는 것(예를 들어, 합성 또는 분석 방법을 수행하는 것)을 의미한다. 서열 또는 리드를 "간접적으로 획득하는"은 다른 관계자 또는 공급원(예를 들어 서열을 직접적으로 획득한 제3자 연구소)으로부터 서열을 수용하거나 또는 서열의 정보 또는 지식을 수용하는 것을 지칭한다. 획득한 서열 또는 리드는 완전한 서열일 필요는 없으며, 예를 들어 적어도 하나의 뉴클레오타이드의 시퀀싱 또는 피험체에서 존재하는 것과 같은 본 명세서에 개시된 변경 중 하나 이상을 확인하는 정보 또는 지식을 얻는 것은 서열을 획득하는 것을 구성한다.

[50] 서열 또는 리드를 직접적으로 획득하는 것은 물리적 물질, 예를 들어 출발 물질, 예컨대 조직 또는 세포 샘플, 예를 들어 생검 또는 분리된 핵산(예를 들어 DNA 또는 RNA) 샘플에서 물리적 변화를 포함하는 과정을 수행하는 것을 포함한다. 대표적인 변화는 2 이상의 출발 물질, 물질을 전단 또는 단편화하는 것, 예컨대 게놈 DNA 단편으로부터 물리적 독립체를 제조하는 것(예를 들어, 조직으로부터 핵산 샘플을 분리시키는 것); 2 이상의 별개의 독립체를 혼합물로 합하는 것, 공유 또는 비-공유 결합을 파괴하거나 또는 형성하는 것을 포함하는 화학 반응을 수행하는 것을 포함한다. 값을 직접적으로 획득하는 것은 상기 기재한 바와 같은 샘플 또는 다른 물질에서 물리적 변화를 포함하는 과정을 수행하는 것을 포함한다.

[51]

[52] 본 발명에서의 용어 "핵산" 또는 "폴리뉴클레오타이드"는 단일 가닥 또는 이중 가닥 형태의 데옥시리보핵산(DNA) 또는 리보핵산(RNA) 및 이들의 중합체를 의미한다. 달리 특별히 제한되지 않는 한, 상기 용어는 기준 핵산과 유사한 결합특성을 갖고 천연 뉴클레오타이드와 유사한 방식으로 대사되는 천연 뉴클레오타이드의 공지된 유사체를 함유하는 핵산을 포함한다. 달리 기재되지 않은 한, 특정 핵산 서열은 또한 명확히 기재된 서열뿐만 아니라 암묵적으로 이의 보존적으로 변형된 변이체(예를 들면, 축퇴성 코돈 치환), 대립유전자, 오소로그, SNP 및 상보적 서열을 포함한다. 구체적으로, 하나 이상의 선택된(또는 모든) 코돈의 3번 위치가 혼합 염기 및/또는 데옥시이노신잔기로 치환되는 서열을 생성함으로써 축퇴성 코돈 치환이 달성될 수 있다(Batzer et al., Nucleic Acid Res. 19:5081 (1991); Ohtsuka et al., J. Biol. Chem. 260:2605-2608 (1985); and Rossolini et al, Mol. Cell. Probes 8:91-98 (1994)). 상기 용어 핵산은 유전자, cDNA, mRNA, 작은 비코딩 RNA, 마이크로 RNA(miRNA), 피위상호작용(Piwi-interacting) RNA 및 유전자 또는 유전자좌에 의해 코딩된 짧은 헤어핀 RNA(shRNA)와 상호 교환적으로 사용된다.

[53] 본 발명에서 베이트는 타겟 유전자만을 포획하는 용도뿐만 아니라, 샘플 유전체의 어떤 영역(예를 들어, 5'-UTR, 인트론, microsatellite 지역, centromere

지역, telomere 지역 등)에서든 목적하는 유전체의 특정 영역을 포획하는 용도로 사용될 수 있다.

- [54] 본 발명에서의 베이트는 다양한 종류의 베이트를 혼합하여 사용하는 것을 특징으로 할 수 있으며, 이에 한정되는 것은 아니나, 바람직하게는 다음 중 2개 이상을 포함하는 것을 특징으로 할 수 있다:
- [55] a) 약 5% 이하로 나타나는 변이를 검출하기 위한 제1베이트 세트로서, 상기 베이트 세트는 약 500X 이상의 read depth가 필요한 변이를 검출할 수 있는 것을 특징으로 할 수 있다;
- [56] b) 약 10% 이상의 빈도로 나타나는 변이를 검출하기 위한 제2베이트 세트로서, 상기 베이트 세트는 약 200X 이상의 read depth가 필요한 변이를 검출할 수 있는 것을 특징으로 할 수 있다;
- [57] c) 약물 대사관련 SNP, 환자 특이적(genomic fingerprint) SNP 및/또는 loss of heterozygosity(LOH) 검출하기 위한 제3베이트 세트로서 상기 베이트 세트는 10-100X read depth가 필요한 변이를 검출할 수 있는 것을 특징으로 할 수 있다;
- [58] d) 구조적 이상(structural variation)을 검출하기 위한 제4베이트 세트로서 상기 베이트 세트는 5-50X read depth가 필요한 변이를 검출할 수 있는 것을 특징으로 할 수 있다; 및
- [59] e) 복제수 변화(copy number variation)를 검출하기 위한 제5베이트 세트로서 상기 베이트 세트는 0.1-300X read depth가 필요한 변이를 검출할 수 있는 것을 특징으로 할 수 있다.
- [60] 본 발명에서 베이트 선택의 효율에 대한 값은 하기 중 하나 이상에 의해 변형될 수 있다: 상이한 베이트 세트의 차별적인 표현, 베이트 서브세트의 차별적인 중복, 차별적인 베이트 변수, 상이한 베이트 세트의 혼합, 및/또는 상이한 유형의 베이트 세트의 사용. 예를 들어, 선택 효율의 변화(예를 들어, 각각의 베이트 세트/표적 카테고리의 상대적 서열 coverage)는 하기 중 하나 이상을 변경함으로써 조절될 수 있다:
- [61] (i) 상이한 베이트 세트의 차별적인 표현 - 주어진 표적(예를 들어, 표적 구성원)을 포획하기 위한 베이트 세트 설계는 더 많은/소수의 복제물에 포함되어 상대적인 표적 coverage depth를 향상시키고/감소시킬 수 있다;
- [62] (ii) 베이트 서브세트의 차별적인 중복 - 주어진 표적(예를 들어, 표적 구성원)을 포획하기 위한 베이트 세트 설계는 이웃하는 베이트 사이의 더 길거나 또는 더 짧은 복제물을 포함하여 상대적인 표적 coverage depth를 향상시키고/감소시킬 수 있다;
- [63] (iii) 차별적인 베이트 변수 - 주어진 표적(예를 들어, 표적 구성원)을 포획하기 위한 베이트 세트 설계는 서열 변형/더 짧은 길이를 포함하여 포획 효율을 감소시키고, 상대적 표적 coverage depth를 낮출 수 있다;
- [64] (iv) 상이한 베이트 세트를 혼합 - 상이한 표적 세트를 포획하기 위해 설계된 베이트 세트는 상이한 물비로 혼합되어 상대적 표적 coverage depth를

향상시키고/감소시킬 수 있다;

- [65] (v) 상이한 유형의 올리고뉴클레오타이드 베이트 세트 사용 - 특정 실시형태에서, 베이트 세트는 하기를 포함할 수 있다:
- [66] (a) 하나 이상의 화학적으로(예를 들어, 비-효소적으로) 합성된(예를 들어, 개별적으로 합성된) 베이트,
- [67] (b) 어레이에서 합성된 하나 이상의 베이트,
- [68] (c) 하나 이상의 효소적으로 제조된, 예를 들어 시험관내 전사된 베이트;
- [69] (d) (a), (b) 및/또는 (c)의 임의의 조합,
- [70] (e) 하나 이상의 DNA 올리고뉴클레오타이드(예를 들어, 자연적으로 또는 비-자연적으로 발생하는 DNA 올리고뉴클레오타이드),
- [71] (f) 하나 이상의 RNA 올리고뉴클레오타이드(예를 들어, 자연적으로 또는 비-자연적으로 발생하는 RNA 올리고뉴클레오타이드),
- [72] (g) (e) 및 (f)의 조합, 또는
- [73] (h) 상기 중 어떤 것의 조합.
- [74] 상이한 올리고뉴클레오타이드 조합은 상이한 비로, 예를 들어 1:1, 1:2, 1:3, 1:4, 1:5, 1:10, 1:20, 1:50; 1:100, 1:1000 등으로부터 선택된 비로 혼합될 수 있다. 일 실시형태에서, 화학적으로 합성된 베이트 대 배열-생성된 베이트의 비는 1:5, 1:10 또는 1:20으로부터 선택된다. DNA 또는 RNA 올리고뉴클레오타이드는 자연적으로- 또는 비-자연적으로 발생될 수 있다.
- [75] 특정 실시형태에서, 베이트는, 예를 들어 용점을 증가시키는 하나 이상의 비-자연적으로 발생하는 뉴클레오타이드를 포함한다. 대표적인 비-자연적으로 발생하는 올리고뉴클레오타이드는 변형된 DNA 또는 RNA 뉴클레오타이드를 포함한다. 대표적인 변형된 뉴클레오타이드(예를 들어, 변형된 RNA 또는 DNA 뉴클레오타이드)는, 이하에 제한되는 것은 아니지만, 잠금 핵산(locked nucleic acid: LNA)을 포함하되, LNA 뉴클레오타이드의 리보스 모이어티(moiety)는 2' 산소와 4' 탄소를 연결하는 추가의 브릿지; 펩타이드 핵산(peptide nucleic acid: PNA), 예를 들어, 펩타이드 결합에 의해 연결된 반복되는 N-(2-아미노에틸)-글라이신 단위로 구성된 PNA; 저 GC 영역을 포획하기 위해 변형된 DNA 또는 RNA 올리고뉴클레오타이드; 바이사이클릭 핵산(bicyclic nucleic acid: BNA); 가교된 올리고뉴클레오타이드; 변형된 5-메틸 테옥시사이티딘; 및 2,6-다이아미노퓨린에 의해 변형된다. 다른 변형된 DNA 및 RNA 뉴클레오타이드는 당업계에서 공지된다.
- [76] 특정 실시형태에서, 표적 서열(예를 들어, 표적 구성원)의 실질적으로 균일하거나 또는 동등한 coverage가 얻어진다. 예를 들어, 각각의 베이트 세트/표적 카테고리 내에서, coverage의 균일성은, 예를 들어 하기 중 하나 이상에 의해 베이트 변수를 변형함으로써 최적화될 수 있다:
- [77] (i) 베이트 표현 또는 중복의 증가/감소는 동일 카테고리 내의 다른 표적에 대해 밑에/위에-cover되어 있는 표적(예를 들어, 표적 구성원)의 coverage를

- 향상/감소시키기 위해 사용될 수 있다;
- [78] (ii) 표적 서열(예를 들어, 고 GC 함량 서열)을 포획하기가 어려운 낮은 coverage에 대해, 예를 들어 인접한 서열(예를 들어, 더 적은 GC-풍부 인접 서열)을 덮는 베이트 세트에 표적화된 영역을 확장한다;
- [79] (iii) 베이트 서열의 변형은 베이트의 2차 구조를 감소시키고, 그것의 선택 효율을 향상시키도록 만들어질 수 있다;
- [80] (iv) 베이트 길이의 변형은 동일 카테고리 내에서 상이한 베이트의 용융 혼성화 역학과 동일하게 되도록 사용될 수 있다. 베이트 길이는 직접적으로(다양한 길이를 갖는 베이트를 생성함으로써) 또는 간접적으로(일정한 길이의 베이트를 생성하고, 베이트 말단을 임의의 서열로 대체함으로써) 변형될 수 있다;
- [81] 동일 표적 영역(즉, 전방 및 후방 가닥)에 대해 상이한 배향의 베이트를 변형시키는 것은 상이한 결합 효율을 가질 수 있다. 각각의 표적에 대해 최적의 coverage를 제공하는 배향 중 하나를 지니는 베이트 세트가 선택될 수 있다;
- [82] (vi) 결합 완전체의 양을 변형시키는 것, 예를 들어 각 베이트에 존재하는 포획 태그(예를 들어, 바이오틴)은 그것의 결합 효율에 영향을 미칠 수 있다. 특이적 표적을 표적화하는 베이트의 태그 수준을 증가/감소시키는 것은 상대적 표적 coverage를 향상/감소시키도록 사용될 수 있다;
- [83] (vii) 상이한 베이트에 대해 사용된 뉴클레오타이드 유형의 변형은 표적의 결합 친화도에 영향을 미치도록 변경될 수 있고, 상대적 표적 coverage를 향상시키고/감소시킬 수 있다; 또는
- [84] (viii) 예를 들어 더 안정한 염기 쌍을 갖는 변형된 올리고뉴클레오타이드 베이트를 사용하는 것은 고 GC 함량에 대해 낮은 또는 정상의 GC 함량의 영역 사이의 용융 혼성화 역학이 동등하게 되도록 사용될 수 있다.
- [85] 예를 들어, 상이한 유형의 올리고뉴클레오타이드 베이트 세트가 사용될 수 있다.
- [86] 일 실시형태에서, 선택의 효율에 대한 값은 상이한 유형의 베이트 올리고뉴클레오타이드를 사용함으로써 변형되어 사전선택된 표적 영역을 포함한다. 예를 들어, 제1 베이트 세트(예를 들어, 10,000 내지 50,000 RNA 또는 DNA 베이트를 포함하는 어레이-기반 베이트 세트)가 거대 표적 영역(예를 들어, 1 내지 2MB 전체 표적 영역)을 덮도록 사용될 수 있다. 제1 베이트 세트는 제2 베이트 세트가 스파이킹되어(spiked)(예를 들어, 5,000 미만의 베이트를 포함하는 개별적으로 합성된 RNA 또는 DNA 베이트 세트) 사전선택된 표적 영역(예를 들어, 250kb 미만의 표적 영역에 걸친, 관심의 선택된 서브게놈 간격) 및/또는 더 높은 2차 구조, 예를 들어 더 고 GC 함량의 영역을 덮는다. 관심의 선택된 서브게놈 간격은 본 명세서에 기재된 유전자 또는 유전자 생성물 또는 이것의 단편 중 하나 이상에 대응될 수 있다. 제2 베이트 세트는 원하는 베이트 중복에 따라서 약 1 내지 5,000, 2 내지 5,000, 3 내지 5,000, 10 내지 5,000, 100 내지 5,000, 500 내지 5,000, 100 내지 5,000, 1000 내지 5,000, 2,000 내지 5,000을 포함할 수

있다. 다른 실시형태에서, 제2 베이트 세트는 제1 베이트 세트에 스파이킹된 선택된 올리고 베이트(예를 들어, 400, 200, 100, 50, 40, 30, 20, 10, 5, 4, 3, 2 또는 1 미만의 베이트)를 포함할 수 있다. 제2 베이트 세트는 개개의 올리고 베이트의 임의의 비로 혼합될 수 있다. 예를 들어, 제2 베이트 세트는 1:1의 등물비로서 존재하는 개개의 베이트를 포함할 수 있다. 대안적으로, 제2 베이트 세트는, 예를 들어 특정 표적의 포획을 최적화하기 위해(예를 들어, 특정 표적은 다른 표적과 비교하여 제2 베이트의 5 내지 10X를 가질 수 있음) 상이한 비(예를 들어, 1:5, 1:10, 1:20)로 존재하는 개개의 베이트를 포함할 수 있다.

- [87] 다른 실시형태에서, 선택 효율은 베이트의 등물 혼합을 사용할 때 관찰된 차별적인 서열 포획 효율에 대해 베이트의 상대적 존재비(abundance) 또는 결합 독립체의 밀도(예를 들어, 헵텐 또는 친화도 태그 밀도)를 조절한 다음, 내부-평균화된(levelled) 그룹 2에 비해 전반적인 베이트 혼합에 대해 내부로-평균화된 그룹 1 만큼 차별적인 과량을 도입함으로써 그룹(예를 들어, 제1, 제2 또는 제3 다수의 베이트) 내에서 개개 베이트의 효율을 평균화함으로써 조절된다.
- [88] 실시형태에서, 해당 방법은 종양 구성원, 예를 들어 종양 세포로부터 표적 서열을 포함하는 핵산 모듈을 선택하는 베이트 세트를 포함하는 복수의 베이트 세트의 사용을 포함한다. 종양 구성원은 종양 또는 암 세포 내에 존재하는 본 명세서에 기재된 종양 세포, 예를 들어 돌연변이된, 야생형, PGx, 기준 또는 인트론 뉴클레오타이드 서열 내에 존재하는 임의의 뉴클레오타이드 서열일 수 있다. 일 실시형태에서, 종양 구성원은 낮은 빈도로 나타나는 변경(예를 들어 하나 이상의 돌연변이)을 포함하며, 예를 들어 종양 샘플로부터의 세포의 약 5% 미만이 그것의 게놈 내 변경을 가지는 것을 의미한다. 다른 실시형태에서, 종양 구성원은 종양 샘플로부터 약 10%의 빈도로 나타나는 변경(예를 들어, 하나 이상의 돌연변이)을 포함한다.
- [89] 다른 실시형태에서, 종양 구성원은 PGx 유전자 또는 유전자 생성물, 인트론 서열, 예를 들어, 본 명세서에 기재된 인트론 서열, 종양 세포 내에 존재하는 기준 서열로부터의 표적 서열을 포함한다.
- [90] 다른 양태에서, 본 발명은 본 명세서에 기재된 베이트 세트, 본 명세서에 기재된 개개의 베이트 세트의 조합, 예를 들어 본 명세서에 기재된 조합을 특징으로 한다. 베이트 세트(들)는 설명서, 표준, 완충제 또는 효소 또는 다른 시약을 선택적으로 포함할 수 있는 키트의 부분일 수 있다.
- [91]
- [92] 본 발명에서의 용어 “페어드 엔드 리드(paired-end read)”는 ‘페어드 엔드’란 동일한 DNA 분자의 양 말단을 의미한다. 한 쪽 말단을 시퀀싱하고, 이를 뒤집어 다른 말단을 시퀀싱했을 경우, 염기서열이 규명된 이들 두 말단을 ‘페어드 엔드 리드’라 한다. 예를 들어 Illumina 시퀀싱은 약 500bps의 리드를 생성하고, 이 리드의 양쪽 끝 75bps의 염기 서열을 읽어낸다. 이때 두 리드(제1리드와

제2리드)를 읽는 방향은 3'와 5'로 각각 반대가 되며, 서로의 페어드 엔드 리드가 된다.

- [93] 본 발명에서의 용어 “제1리드”, “제2리드”는 페어드 엔드 리드 시퀀싱(Paired-End Read Sequencing)을 통해 얻어진 5' 방향의 제1리드와 3'방향의 제2리드를 의미한다.
- [94] 본 발명에서의 용어 “소프트-클립(soft-clip)”, “소프트-클립 조각(soft-clip segment)” 또는 “소프트 클립 리드(soft clipped read)”는 NGS에서 획득한 리드에서 일부만 참조 유전체로 맵핑되고, 나머지는 맵핑이 되지 않은 상태의 리드를 의미한다.
- [95] 본 발명에서의 용어 “불일치 리드 쌍(disconcordant read pair)”은 페어드 엔드 리드 시퀀싱으로 획득한 리드 쌍(제1리드, 제2리드)이 같은 참조 유전자 상에 맵핑되지 않고, 서로 다른 위치 또는 서로 다른 염색체 상에 맵핑되는 리드 쌍을 의미한다.
- [96] 본 발명에서의 용어 “일치 리드 쌍(concordant read pair)”은 페어드 엔드 리드 시퀀싱으로 획득한 리드 쌍(제1리드, 제2리드)이 같은 유전자에 맵핑되었지만, 리드의 소프트 클립 조각 부분이 다른 유전자에 맵핑되는 정보를 가지고 있는 것을 의미한다.
- [97] 본 발명에서 용어 “지원 쌍 개수(supporting pair count)”는 융합 유전자의 제1유전자 및 제2유전자 양쪽 모두에 매칭되는 리드 쌍이 1개 이상일 경우 개수가 1 증가 하는 것을 의미하며, 이 때의 리드 쌍은 불일치 리드 쌍이나, 일치 리드 쌍과 상관없이 2 개 이상이면 가능하다.
- [98]
- [99] 본 발명에서는 암 샘플로부터 핵산을 추출하여, NGS로 리드를 획득한 다음, 불일치 리드쌍과 일치 리드쌍을 모두 이용하여 유전자 재배열을 검출할 수 있는지 확인하고자 하였다(도 1).
- [100] 즉, 본 발명의 일 실시예에서는 폐암 조직 샘플에서 획득한 FFPE 샘플에서 핵산을 추출하여, NGS를 통해 리드를 획득하고 정렬한 다음, 융합 유전자 후보 리드를 추출하여 불일치 리드 쌍과 일치 리드 쌍을 분리한 다음(도 2), 페어-블라스트 검색을 통해 상기 리드 쌍으로부터 융합 유전자 후보군을 도출하여 지원 쌍 개수(supporting pair count)를 결정하였으며(도 3), 상기 획득한 리드 중, 추출하지 않은 리드를 융합 유전자 후보군으로부터 제작한 융합 유전자 모형(fusion gene template)와 매칭시켜 지원 리드 개수(supporting read count)를 결정한 다음, 지원 쌍 개수 및 지원 리드 개수를 고려하여 최종적으로 융합 유전자를 검출하였으며(도 4), 그 결과를 기존의 공지된 프로그램 중 하나인 FACTERA(Fusion gene And Chromosomal Trnslocation Enumeration and Recovery Algorithm, Aaron M. et al., 2014) 프로그램과 비교한 결과, 공지 프로그램이 검출하지 못하는 융합 유전자를 본 발명의 방법으로 검출할 수 있다는 것을 확인하였다(도 5, 표 1).

[101]

[102] 따라서, 본 발명은 일 관점에서, (a) 피험체의 샘플로부터 복수의 핵산 분자를 포함하는 라이브러리를 획득하는 단계; (b) 상기 라이브러리를 복수의 베이트(bait) 세트와 접촉시킴으로써, 사전선택된 서열에 대해 상기 라이브러리를 풍부화시켜서 선택된 핵산 분자를 제공하고, 이로 인해 라이브러리 캐치(library catch)를 제공하는 단계; (c) 차세대 시퀀싱 방법에 의해, 상기 라이브러리 캐치의 핵산 분자로부터 리드를 획득하는 단계; (d) 정렬 방법에 의해 상기 리드를 정렬하는 단계; 및 (e) 상기 정렬된 리드를 분석하여 유전자 재배열을 검출하는 단계, 여기서 상기 분석 방법은 상기 정렬된 리드를 추출하여 서열 유사성을 분석하는 단계를 포함하는 샘플 내 유전자 재배열 검출 방법에 관한 것이다.

[103] 본 발명에서 있어서, 상기 암은 비호지킨 림프종 (non-Hodgkin lymphoma), 호지킨 림프종 (non-Hodgkin lymphoma), 급성 골수성 백혈병 (acute-myeloid leukemia), 급성 림프구성 백혈병 (acute-lymphoid leukemia), 다발성 골수종 (multiple myeloma), 경부암 (head and neck cancer), 폐암, 교모세포종 (glioblastoma), 대장/직장암, 췌장암, 유방암, 난소암, 흑색종 (melanoma), 전립선암, 신장암 및 중피종 (mesothelioma)으로 구성된 군에서 선택되는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[104] 본 발명에 있어서, 상기 샘플은 하나 이상의 전암성(premalignant) 또는 악성 세포; 고형 종양, 연조직 종양 또는 전이 병변으로부터 선택된 세포; 수술절제면으로부터의 조직 또는 세포; 조직학적으로 정상인 조직; 하나 이상의 혈중 종양 세포(CTC); 정상 인접 조직(normal adjacent tumor: NAT) 및 종양을 가지거나 또는 가질 위험에 있는 동일 피험체로부터의 혈액 샘플로 구성된 군에서 선택되는 것을 특징으로 할 수 있으며, 바람직하게는 FFPE 샘플인 것을 특징으로 할 수 있으나 이에 한정되는 것은 아니다.

[105] 본 발명에 있어서, 상기 유전자 재배열(gene rearrangement)은 정상 유전체에 대비하여, 염기서열 상의 위치가 변화된 변이이면 그 종류에 상관없이 모두 의미하며, 유전자 융합(fusion), 전좌(translocation), 역위(inversion), 결실(deletion)로 구성된 군에서 선택될 수 있으나 이에 한정되는 것은 아니다.

[106] 본 발명에 있어서, 상기 방법은 DNA NGS 분석 또는 RNA NGS 분석에 적용되는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니며, 유전자 재배열을 NGS로 분석할 수 있는 모든 방법에 적용가능하다는 것은 당업자에게 자명하다. (내용 추가)

[107] 본 발명에 있어서, 상기 (d) 단계에서 리드의 정렬은 차세대 시퀀싱 방법(NGS)으로 획득한 리드를 유전체 좌표로 정렬할 수 있는 프로그램을 이용한 방식이면 모두 이용가능하나, 바람직하게는 BWA(Burrows-Wheeler Aligner)를 이용하는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[108] 본 발명에 있어서, 상기 BWA 사용 시, 일치 리드 쌍에 대한 정보를 추가하기

위하여 2차 정렬 태그(secondary alignment tag)를 추가하는 mark shorter split hits as secondary(-M) 옵션을 사용하는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

- [109] 본 발명에 있어서, 상기 (d) 단계에서 리드의 정렬을 수행하는 참조 유전체는 정상 세포의 전체 유전체, 예를 들어 hg19를 사용하는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.
- [110] 본 발명에 있어서, 상기 (d) 단계에서 정렬한 리드 파일의 형식은 BAM/SAM 파일인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.
- [111] 본 발명에 있어서, 상기 (d) 단계의 후보 리드 추출은 관심영역(region of interest) 정보로 상기 (a) 단계에서 획득한 리드를 필터하여 추출하는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.
- [112] 본 발명에 있어서, 상기 관심 영역은 이미 공지된 융합 유전자의 위치에 대한 정보 및 타겟 유전자 영역 정보를 포함하는 것을 특징으로 할 수 있고, 상기 관심 영역 정보는 염색체 정보와 염색체 상의 시작 및 끝 위치 정보 형태인 것을 특징으로 할 수 있다.
- [113] 본 발명에 있어서, 상기 관심영역 정보는 하기 표 1에서 개시된 내용을 포함하는 것을 특징으로 할 수 있으나 이에 한정되는 것은 아니다.
- [114]

[115] [Ⅹ1]

관심영역 정보

Chromosome	Start Position	Stop Position	Gene Name
chr1	156843424	156843751	NTRK1_10
chr1	156843751	156844174	NTRK1_i_01
chr1	156844174	156844192	NTRK1_11
chr1	156844192	156844362	NTRK1_i_02
chr1	156844362	156844418	NTRK1_12
chr1	156844418	156844697	NTRK1_i_03
chr1	156844697	156844800	NTRK1_13
chr1	156844800	156845311	NTRK1_i_04
chr1	156845311	156845458	NTRK1_14
chr1	156845458	156845871	NTRK1_i_05
chr1	156845871	156846002	NTRK1_15
chr2	29446207	29446394	ALK_10
chr2	29446394	29448326	ALK_i_01
chr2	29448326	29448431	ALK_11
chr2	29448431	29449787	ALK_i_02
chr2	29449787	29449940	ALK_12
chr2	29449940	29450439	ALK_i_03
chr2	29450439	29450538	ALK_13
chr2	29450538	29451749	ALK_i_04
chr2	29451749	29451932	ALK_14
chr4	1808272	1808410	FGFR3_16
chr4	1808410	1808555	FGFR3_i_01
chr4	1808555	1808661	FGFR3_17
chr4	1808661	1808842	FGFR3_i_03
chr4	1808842	1808989	FGFR3_18
chr6	117641030	117641193	ROS1_08
chr6	117641193	117642421	ROS1_i_01
chr6	117642421	117642557	ROS1_09
chr6	117642557	117645494	ROS1_i_02
chr6	117645494	117645578	ROS1_10
chr6	117645578	117647386	ROS1_i_03
chr6	117647386	117647577	ROS1_11

[116]

chr6	117647577	117650491	ROS1_i_04
chr6	117650491	117650609	ROS1_12
chr6	117650609	117658334	ROS1_i_05
chr6	117658334	117658503	ROS1_13
chr7	55259411	55259567	EGFR_23
chr7	55259567	55260458	EGFR_i_23
chr7	55260458	55260534	EGFR_24
chr7	55260534	55266409	EGFR_i_24
chr7	55266409	55266556	EGFR_25
chr7	55266556	55268008	EGFR_i_25
chr7	55268008	55268106	EGFR_26
chr8	38283639	38283763	FGFR1_13
chr8	38283763	38285438	FGFR1_i_01
chr8	38285438	38285611	FGFR1_14

[117]

chr 10	43609003	43609123	RET_10
chr 10	43609123	43609927	RET_i_01
chr 10	43609927	43610184	RET_11
chr 10	43610184	43612031	RET_i_02
chr 10	43612031	43612179	RET_12
chr 10	43612179	43613820	RET_i_03
chr 10	43613820	43613928	RET_13
chr 10	123239094	123239184	FGFR2_01
chr 10	123239184	123239370	FGFR2_i_01
chr 10	123239370	123239535	FGFR2_02
chr 10	123239535	123241685	FGFR2_i_02
chr 10	123241685	123241691	FGFR2_03
chr 10	123241691	123243211	FGFR2_i_03
chr 10	123243211	123243317	FGFR2_04
chr 1	114935399	115053781	TRIM33
chr 1	154127780	154164611	TPM3
chr 1	156052369	156109880	LMNA
chr 1	156611740	156629324	BCAN
chr 1	204797782	204991950	NFASC
chr 1	205626979	205649630	SLC45A3
chr 10	32297938	32345371	KIF5B

[118]

chr 10	51565108	51590734	NCOA4
chr 10	60272774	60591194	BICC1
chr 10	61548505	61666414	CCDC6
chr 10	75757836	75879918	VCL
chr 10	115438921	115490668	CASP7
chr 10	118642888	118886097	KIAA1598
chr 11	3022152	3078681	CARS
chr 11	62623484	62656355	SLC3A2
chr 12	1100404	1605099	ERC1
chr 12	27677045	27848497	PPFIBP1
chr 12	59265937	59314319	LRIG3
chr 12	122755981	122907179	CLIP1
chr 12	122956146	122985543	ZCCHC8
chr 14	56046925	56151302	KTN1
chr 14	93260576	93306308	GOLGA5
chr 14	104095525	104167888	KLC1

[119]

chr 15	40987327	41024356	RAD51
chr 15	52599480	52821247	MYO5A
chr 17	7571720	7590868	TP53
chr 17	16945790	17095962	MPRIP
chr 17	57697050	57774317	CLTC
chr 17	66507921	66547457	PRKAR1A
chr 18	59854506	59974355	KIAA1468
chr 19	16178317	16213815	TPM4
chr2	24252206	24270296	C2orf44
chr2	37064841	37193673	STRN
chr2	42396490	42559688	EML4
chr2	54683454	54898583	SPTBN1
chr2	74588281	74619214	DCTN1
chr2	100162326	100759037	AFF3
chr2	109335902	109402267	RANBP2
chr2	216176679	216214496	ATIC
chr2	216225177	216300890	FN1
chr20	43953928	43977064	SDC4
chr22	19166986	19279247	CLTCL1

[120]

chr3	100428128	100467811	TFG
chr4	1723217	1746905	TACC3
chr4	25656853	25680735	SLC34A2
chr4	83739814	83812419	SEC31A
chr5	149781200	149792543	CD74
chr5	159502889	159546452	PWWP2A
chr5	170814652	170837888	NPM1
chr5	179233388	179265078	SQSTM1
chr6	28870779	28891768	TRIM27
chr6	29910247	29913661	HLA-A
chr6	117881432	117923705	GOPC
chr6	159186773	159240456	EZR
chr7	44915892	44924960	PURB
chr7	75162619	75368290	HIP1
chr7	97920962	98030427	BAIAP2L1
chr7	101459184	101927250	CUX1
chr7	138145079	138270333	TRIM24
chr8	17780364	17887457	PCM1
chr8	22462145	22477984	KIAA1967
chr8	37553301	37556396	ZNF703
chr8	37593743	37615319	ERLIN2
chr8	38034106	38070819	BAG4
chr8	42752033	42885682	HOOK3
chr9	125703288	125867147	RABGAP1
<u>chrX</u>	13733549	13787480	OPD1
<u>chrX</u>	64808257	64961793	MSN

[121]

[122] 본 발명에 있어서, 상기 (e) 단계는 상기 리드를 추출한 다음, 불일치 리드 쌍(discordant read pair)와 일치 리드 쌍(concordant read pair)으로 분리하는 단계를 포함하는 것을 특징으로 할 수 있다.

[123] 본 발명에 있어서, 상기 불일치 리드 쌍과 일치 리드쌍의 분리는 참조 유전자(RefGene) 정보와 리드를 매칭하여 분리하는 것을 특징으로 할 수 있다.

[124] 본 발명에 있어서, 상기 참조 유전체는 페어드 엔드 리드 시퀀싱으로 얻은 리드

쌍(제1리드 및 제2리드)가 불일치 리드 쌍인지 일치 리드 쌍인지를 결정할 수 있는 유전체 정보이면 모두 이용가능하나, 바람직하게는 USCS genome database 유래 RefGene

정보(<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.2bit>)를 이용할 수 있다.

- [125] 본 발명에 있어서, 상기 불일치 리드 쌍은 소프트 클립을 가지고 있거나, 가지고 있지 않는 것을 특징으로 할 수 있으며(도 2, type 3, 4), 일치 리드 쌍은 소프트 클립을 가지고 있으면서 SA 정보가 없거나, SA 정보를 가지고 있는 것(도 2, type 1, 2)을 특징으로 할 수 있다.
- [126] 본 발명에 있어서, 상기 (e) 단계의 불일치 리드 쌍의 분리 후, 제1리드의 soft-clip segment 부분을 쿼리로 하여, 쌍을 이루는 제2리드의 매칭 영역을 찾거나, 제2리드의 soft-clip segment를 쿼리로 하여 쌍을 이루는 제1리드의 매칭 영역을 찾는 단계를 추가로 포함하는 것을 특징으로 할 수 있다.
- [127] 본 발명에 있어서, 상기 (e) 단계의 일치 리드 쌍의 분리 후, 제1리드의 2차 정렬 태그 정보를 참조하여 다른 유전자에 2차적으로 맵핑 되는지를 확인하여 얻은 2차 맵핑 지역을 가상의 제2리드로 가정하여, 제1리드의 soft-clip segment 부분을 쿼리로 하여, 쌍을 이루는 제2리드의 매칭 영역을 찾거나, 제2리드의 soft-clip segment를 쿼리로 하여 쌍을 이루는 제1리드의 매칭 영역을 찾는 단계(도 3)를 추가로 포함하는 것을 특징으로 할 수 있다.
- [128] 본 발명에 있어서, 상기 (e) 단계의 일치 리드 쌍 및 불일치 리드쌍을 분리한 다음, 각 리드의 매칭영역을 찾는 방법을 “페어-블라스트 검색(pair-blast search)”라고 통합하여 지칭할 수 있다.
- [129] 본 발명에 있어서, 상기 불일치 리드 쌍의 매칭 영역을 찾는 단계에서, 매칭 영역이 발견되었을 경우, 상기 매칭영역을 포함하는 리드를 유전자 재배열 후보군으로 도출하여 지원 쌍 개수(supporting pair count)를 결정하는 것을 특징으로 할 수 있다.
- [130] 본 발명에 있어서, 상기 일치 리드 쌍의 매칭 영역을 찾는 단계에서, 매칭 영역이 발견되었을 경우, 상기 매칭영역을 포함하는 리드를 유전자 재배열 후보군으로 도출하여 지원 쌍 개수(supporting pair count)를 결정하는 것을 특징으로 할 수 있다.
- [131] 본 발명에 있어서, 상기 지원 쌍 개수는 불일치 리드 쌍 및 일치 리드쌍에서 결정한 결과를 통합하는 단계를 거쳐 결정되는 것을 특징으로 할 수 있다.
- [132] 본 발명에 있어서, 상기 지원쌍 개수를 통합하는 단계는 불일치 리드 쌍에서도 지원 쌍 개수로 결정되고, 일치 리드 쌍에서도 지원 쌍 개수로 결정될 경우 증가하는 것을 특징으로 할 수 있고, 유전자 재배열의 종류가 같더라도, 유전자 재배열 위치가 상이할 경우, 서로 다른 것으로 결정하는 것을 특징으로 할 수 있다.
- [133]

- [134] 본 발명에 있어서, 상기 (e) 단계는 유전자 재배열 후보 리드로 추출되지 않는 리드를 추가 분석을 위해 정리하는 단계를 추가로 포함하는 것을 특징으로 할 수 있다.
- [135] 본 발명에 있어서, 상기 (e) 단계는 유전자 재배열 후보군으로 도출된 리드 정보를 바탕으로 유전자 재배열 모형(gene rearrangement template)을 제조하는 단계(도 4, fusion gene template)를 추가로 포함하는 것을 특징으로 할 수 있다.
- [136] 본 발명에 있어서, 상기 유전자 재배열 모형은 유전자 재배열 후보군으로 도출된 리드에서, 유전자 재배열 위치(예를 들어, 유전자 재배열이 융합 유전자 일 경우, 융합 유전자의 breakpoint)로부터 5' 방향으로 300bp 내지 500bp, 3' 방향으로 300bp 내지 500bp를 포함하는 참조 유전체 상의 염기 서열인 것을 특징으로 할 수 있으나 이에 한정되는 것은 아니다.
- [137] 본 발명에 있어서, 상기 (e) 단계의 서열 유사성을 분석하는 단계는 유전자 재배열 모형과 유전자 재배열 후보군으로 분석하기 위해 추출되지 않은 리드를 비교하여 지원 리드 개수(supporting read count)를 결정하는 단계를 추가로 포함하는 것을 특징으로 할 수 있다.
- [138] 본 발명에 있어서, 상기 지원 리드 개수는 상기 정리된 리드를 blastdb로 하고, 상기 유전자 재배열 모형을 쿼리로 하여 블라스트를 수행한 다음, 유전자 재배열 모형에서 유전자 재배열의 위치(breakpoint)를 지나가면서 맵핑되는 리드의 개수를 지원 리드 개수로 결정하는 것을 특징으로 할 수 있다.
- [139] 본 발명에 있어서, 상기 유전자 재배열 후보군으로 분석하기 위해 추출되지 않은 리드는 유전자 재배열 후보군 위치의 5' 방향, 3' 방향 500bp 이내에 존재하는 리드인 것을 특징으로 할 수 있으나 이에 한정되는 것은 아니다.
- [140] 본 발명에 있어서, 상기 유전자 재배열 후보군으로 분석하기 위해 추출되지 않은 리드는 소프트 클립 조각(soft clip segment)를 포함하고 있는 것을 특징으로 할 수 있다.
- [141] 본 발명에 있어서, 상기 유전자 재배열을 검출하는 단계는 지원 리드 개수(supporting read count)가 5개 이상일 때, 유전자 재배열로 결정하는 것을 특징으로 할 수 있다.
- [142] 본 발명에 있어서, 상기 유전자 재배열을 검출하는 단계는 지원 쌍 개수(supporting pair count)가 2개 이상인 기준값을 추가로 포함하는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.
- [143] 본 발명에 있어서, 상기 유전자 재배열을 검출하는 지원 리드 개수 및 지원 쌍 개수는 하기 수식으로 결정되는 것을 특징으로 할 수 있다.
- [144] 수식 1: 지원 쌍 개수(Supporting Pair Score) = Discordant Supporting Pair Count + Concordant Supporting Pair Count
- [145] 수식 2: 지원 리드 개수(Supporting Read Score) = Read1 Supporting Read Count + Read2 Supporting Read Count
- [146] 수식 3: 기준값(Cutoff): Supporting Pair Score $\geq$ 2 AND Supporting Read Score

>= 5

[147]

[148] 본 발명의 다른 실시예에서는 폐암 환자의 FFPE 샘플로부터 NGS를 수행한 리드를 Hg19 참조 유전체에 BWA의 ?M 옵션을 이용하여 정렬한 다음, 관심영역 정보를 바탕으로 리드를 추출하고, UCSC genome database의 유전체 정보에 매칭하여 일치 리드 쌍과 불일치 리드 쌍을 분리하였다. 그 뒤, 불일치 리드 쌍의 경우, 제1리드와 제2리드의 소프트 클립 조각을 서로 매칭하여 매칭되는 부분을 찾고, 일치 리드 쌍의 경우, 가상의 제2리드를 제작하여 제1리드와 매칭하여 매칭되는 부분을 찾아 이를 지원 쌍 개수(supporting pair count)로 결정하였다. 그 다음, 상기 단계에서 결정된 유전자 재배열 후보군으로 유전자 재배열 모형을 제작하여 상기 단계에서 추출되지 않은 리드를 유전자 재배열 모형에 매칭시켜 지원 리드 개수(supporting read count)를 결정하고, 상기 지원 리드 개수가 제1리드 및 제2리드에서 각각 1개 이상일 경우, 이를 유전자 재배열로 결정하는 컴퓨터 시스템을 설계하고 이를 테스트 하여, 기존의 공개된 프로그램에서 찾지 못하는 유전자 재배열을 찾을 수 있다는 것을 확인 하였다.

[149] 따라서, 본 발명은 다른 관점에서, 차세대 염기 서열 분석(Next Generation Sequencing, NGS)을 이용한 유전자 재배열 검출 방법을 수행할 수 있도록 컴퓨팅 시스템을 제어하기 위한 복수의 명령이 암호화된 컴퓨터 판독 가능한 매체를 포함하는 컴퓨터 시스템으로서, 상기 방법은 (a) 피험체의 샘플로부터 복수의 핵산 분자를 포함하는 라이브러리를 획득하는 단계; (b) 상기 라이브러리를 복수의 배이트(bait) 세트와 접촉시킴으로써, 사전선택된 서열에 대해 상기 라이브러리를 풍부화시켜서 선택된 핵산 분자를 제공하고, 이로 인해 라이브러리 캐치(library catch)를 제공하는 단계; (c) 차세대 시퀀싱 방법에 의해, 상기 라이브러리 캐치의 핵산 분자로부터 리드를 획득하는 단계; (d)정렬 방법에 의해 상기 리드를 정렬하는 단계; 및 (e)상기 정렬된 리드를 분석하여 유전자 재배열을 검출하는 단계, 여기서 상기 분석 방법은 상기 정렬된 리드를 추출하여 서열 유사성을 분석하는 단계를 포함하는 것을 특징으로 하는 컴퓨터 시스템에 관한 것이다.

[150]

[151] 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.

[152]

[153] 실시예 1: 폐암 샘플로부터 차세대 염기 서열 분석기법을 수행하는 단계

[154] FISH(fluorescence in situ hybridization)을 통해 이미 어떤 유합 유전자가 존재하는지 알고 있는 폐암 조직 샘플에서 획득한 FFPE 샘플에서 DNA를 추출하여 라이브러리를 제작하고 Illumina 사의 MiSeq을 이용하여 NGS 리드를

수득하였다.

[155] 베이트는 표 1의 관심영역 정보를 포함하는 염색체 상의 위치를 모두 포함할 수 있도록 설계하였다.

[156]

[157] 실시예 2: 시퀀싱된 서열들을 참조 유전체에 서열 정리 및 리드 분류

[158] 실시예 1에서 수득한 리드를 Hg19 참조 유전체에 BWA로 정렬하였으며, 이때 추후 일치 리드 쌍의 분석을 위해 정렬 프로그램(BWA)에서 2차 정렬 태그를 추가하는 옵션(-M)을 추가하여 분석을 진행하였다.

[159] 불일치 리드 쌍과 일치 리드 쌍은 UCSC RefGene 정보(HCSC hg 19)를 바탕으로 리드를 입력하여 분리한 다음, 필터링한 리드를 오름차순으로 정렬하여 순차적으로 제1리드 및 제2리드를 추출하기 쉽도록 정리하였다.

[160]

[161] 실시예 3: 페어-블라스트 검색 및 지원 쌍 개수 결정

[162] 실시예 2에서 분류한 불일치 리드 쌍의 경우, 제1리드(read1)의 soft-clip segment 부분을 추출하여 query를 만들고 mate를 이루는 제2리드(read2)의 matching segment 부분을 subject로 하여 blastn search local alignment 과정을 수행하였다. 이 과정을 통해 align된 read의 strand를 판별하여 read1(gene1)과 read2(gene2)의 방향성을 결정하였으며, read2의 soft-clip segment를 query, read1의 matching segment 부분을 subject로 하여 반대의 방법으로도 blastn search를 같은 방식으로 수행하였다.

[163] 그 결과, nucleotide level의 match, mismatch정보를 얻을 수 있기 때문에 fusion gene 양쪽 모두에 존재하는 micro-homology sequence도 판별할 수 있으며, 외래에서 삽입된 inserted-sequence도 검출할 수 있는 것을 확인할 수 있었다.

[164] 일치 리드 쌍의 경우에는 read1과 read2가 같은 gene에 mapping되어 있기 때문에 read1에서 추가적인 정보를 가지는 SA(secondary alignment) tag 정보를 참조하여 read2가 아니라 다른 gene에 2차적으로 mapping되어 있는지를 다시 확인하여, 2차 mapping 지역을 가상의 read2로 가정하고 불일치 리드 쌍에서 했던 방식과 같은 방법으로 서열 분석을 진행하였다.

[165] 그 결과, fusion gene orientation, micro-homology, inserted-sequence 정보를 가지고 nucleotide basepair resolution으로 fusionbreak point를 하기의 기준으로 결정할 수 있음을 확인할 수 있었다(표 2): 불일치 리드 쌍 및 일치 리드 쌍에서 각각 도출된 융합 유전자 후보군을 통합하고 지원 쌍 개수를 결정할 때, 각각의 리드 쌍에서 융합 유전자 후보군으로 동시에 결정될 경우, 지원 쌍 개수를 증가시키되, 유전자 종류가 같더라도, brakpoint가 micro-homology sequence 개수 이상 차이나는 것은 하나의 fusion gene으로 합치지 않고 다른 fusion gene으로 기록한다.

[166]

[167] [표2]

지원 쌍 개수 결정 단계까지의 결과(일부)

fusion gene	fusion orient	gene1	gene1 transcript	gene1 strand	gene1 chr	gene1 breakpoint	gene1 orient	gene2	gene2 transcript	gene2 strand	gene2 chr	gene2 breakpoint	gene2 orient	support pair cnt	new fusion gene	gene1 query size	gene2 query size	micro homology	discordant pair cnt
SLC34A2>>ROS1	>>	SLC34A2	NM_006424.2	+	chr4	25679213	>	ROS1	NM_002944.2	-	chr6	117656511	>	10	SLC34A2>>ROS1	300	301	0	0
EGFR<>TPK1	<>	EGFR	NM_005228.3	+	chr7	55266908	<	TPK1	NM_022445.3	-	chr7	144478980	>	4	EGFR<>TPK1	300	297	4	0
FGFR1<>PGBD5	><	FGFR1	NM_015850.3	-	chr8	38285389	>	PGBD5	NM_00125831.1	-	chr1	230468742	<	4	FGFR1><PGBD5	300	288	13	0
ULK4>>RET	>>	ULK4	NM_017886.2	-	chr3	41561725	>	RET	NM_020975.4	+	chr10	43611745	>	3	ULK4>>RET	300	301	0	0
ACOT7>>AFF3	>>	ACOT7	NM_007274.3	-	chr1	6408272	>	AFF3	NM_002285.2	-	chr2	100632991	>	2	ACOT7>>AFF3	301	301	1	0
AFF3<>PARP4	<>	AFF3	NM_002285.2	-	chr2	100477189	<	PARP4	NM_006437.3	-	chr13	25062673	>	2	AFF3<>PARP4	300	268	33	0
AFF3<>WISP2	<>	AFF3	NM_002285.2	-	chr2	100632570	<	WISP2	NM_003881.2	+	chr20	43351474	>	2	AFF3<>WISP2	300	301	0	1

[168]

[169] 실시예 4: 지원 리드 개수 결정

[170] 실시예 2에서 추출한 리드를 제외한 나머지 read 들을 fusion gene template 에 BLAST search 를 통해 mapping 시켜서 supporting read count 를 결정하였다. 먼저, 실시예 3의 분석에 이용한 리드를 제외한 나머지 read들을 추출하여 blastdb를 만든 뒤, 실시예 3에서 얻은 융합 유전자 후보군을 바탕으로 5' 방향, 3' 방향 각각 300bp 씩을 추출하여 fusion gene template를 제조한 다음, 이를 query로 하여 blastn 탐색을 수행하였다.

[171] 위 과정에서, 모든 나머지read를 추출하는 것이 아니라 fusion gene breakpoint 5' 방향, 3' 방향 500 basepair이내에 존재하는 read만 추출하며 soft-clip segment를 가지고 있는 read만 추출하여 분석에 이용하였으며, Fusion gene template의 fusion breakpoint를 지나가면서 mapping 되는 read의 갯수만 필터링하여 fusion gene supporting read count로 기록하였다.

[172] 그 결과, 제1리드와 제2리드에서 supporting read count가 각각 결정되는 것을 확인할 수 있었다.

[173]

[174] [표3]

지원 리드 개수 결정 단계까지의 결과(일부)

fusion_gene	fusion_orient	gene1	gene1_tx	gene1_str	gene1_ch	gene1_bp	gene1_orient	inserted_seq	gene2	gene2_tx	gene2_str	gene2_ch	gene2_bp	gene2_orient	Supporting Pair Count	new_fusion_gene	Support Read Count (BLAST)
HIP1>>ALK	>>	HIP1	NM_005338.5	-	chr7	75172169	>		ALK	NM_004304.4	-	chr2	29446393	>	130	HIP1>>ALK	224
NTRK3>>FN1	>>	NTRK3	NM_002530.3	-	chr15	88472629	>	G	FN1	NM_002026.2	-	chr2	216245663	>	5	NTRK3>>FN1	10
MET>>MPRIP	>>	MET	NM_000245.2	+	chr7	116340231	>		MPRIP	NM_015134.3	+	chr17	17035425	>	5	MET>>MPRIP	5
FBXL2>>ROS1	>>	FBXL2	NM_012157.3	+	chr3	33358244	>	A	ROS1	NM_002944.2	-	chr6	117647586	>	4	FBXL2>>ROS1	8
EGFR>>SLC34A2	>>	EGFR	NM_005228.3	+	chr7	55266557	>	CTGGATGATAGACCCAGATAGTCCGCC	SLC34A2	NM_006424.2	+	chr4	25679663	>	3	EGFR>>SLC34A2	500
KIAA1967>>JAK1	>>	KIAA1967	NM_021174.5	+	chr8	22476766	>	ATAAATG	JAK1	NM_002227.2	-	chr1	65349161	>	3	KIAA1967>>JAK1	476
KDM6A>>TP53	>>	KDM6A	NM_021140.2	+	chrX	44790819	>	TGCTCAGATAGCGAT	TP53	NM_001126113.2	-	chr17	7578290	>	3	KDM6A>>TP53	222
EGFR>>YWHAG	>>	EGFR	NM_005228.3	+	chr7	55259563	>	G	YWHAG	NM_013479.3	-	chr7	75958412	>	3	EGFR>>YWHAG	6
FAM20A><ROS1	<<	FAM20A	NM_017565.3	-	chr17	66539615	>	G	ROS1	NM_002944.2	-	chr6	117687239	<	3	FAM20A><ROS1	6
SLC34A2>>EGFR	>>	SLC34A2	NM_006424.2	+	chr4	25658933	>	AT	EGFR	NM_005228.3	+	chr7	55272981	>	3	SLC34A2>>EGFR	6
TSC2><SLC34A2	<<	TSC2	NM_000548.3	+	chr16	2110699	<	G	SLC34A2	NM_006424.2	+	chr4	25678927	>	3	TSC2><SLC34A2	6
ROS1>>AFF3	>>	ROS1	NM_002944.2	-	chr6	117730745	>	AAAATGGAACCTTCCACTGCTCCCTTTGCTTC	AFF3	NM_002285.2	-	chr2	100703237	>	2	ROS1>>AFF3	500

[175]

[176] 실시예 5: 최종 융합 유전자 결정

[177] 상기 실시예 3에서 도출한 지원 쌍 개수가 2 이상인 리드를 바탕으로 융합 유전자 모형을 제조하였고, 이를 바탕으로 지원 리드 개수를 결정한다. 지원 리드 개수가 제1리드 및/또는 제2리드에서 5 이상인 융합 유전자 후보군을 최종적으로 융합 유전자로 결정하였다.

[178] 그 결과, 표 4에 개시된 바와 같이 기존의 공지된 프로그램에서는 검출할 수 없었던 융합 유전자를 검출할 수 있다는 것을 확인하였다.

[179]

[180] [표4]

본원 발명과 기존 프로그램의 융합 유전자 검출 결과 비교

Fusion gene (by FISH)	Fusion gene (by program)	Sample	<u>FindingFusion</u>	FACTERA
ROS1	SLC34A:ROS1	FFPE6	0	0
	HSF2:ROS1 / ROS1:VGLL2	FFPE9	0	0
RET	KIF5B:RET	FFPE22	0	X
	CCDC6:RET	FFPE24	0	X
	CCDC6:RET	FFPE43	0	0
ALK	EML4:ALK	FFPE17	0	0
	EML4:ALK	FFPE28	0	0
	EML4:ALK	FFPE29	0	0
	EML4:ALK	FFPE37	0	0
	EML4:ALK	FFPE45	0	0
	EML4:ALK	FFPE50	0	0
	EML4:ALK	FFPE52	0	0
	EML4:ALK	FFPE53	0	0
	EML4:ALK	FFPE54	0	0
	HIP:ALK	FFPE56	0	0

[181]

[182] 이상으로 본 발명 내용의 특정한 부분을 상세히 기술하였는바, 당업계의 통상의 지식을 가진 자에게 있어서, 이러한 구체적 기술은 단지 바람직한 실시양태일 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.

[183]

산업상 이용가능성

[184] 본 발명에 따른 NGS를 이용한 유전자 재배열 검출 방법은 NGS에서 수득한 리드를 통해 유전자 재배열의 검출뿐만 아니라, 유전자 재배열의 방향,

마이크로호물로지 서열, 외부 삽입 서열 및 유전자 재배열의 위치까지 베이스 페어 단위로 정확히 판별해 낼 뿐만 아니라, 기존의 방식으로는 탐색하지 못했던 일치 리드 쌍(concordant read pair)에서도 탐색을 수행하여 정확도가 높으며, 특정 암이나 종양 관련 유전자 부위만 중점적으로 탐색이 가능하여 검출에 소모되는 시간이 감소되는 장점이 있다. 따라서 본 발명의 방법은 암 샘플에서 유전자 재배열을 효과적으로 검출하는데 유용하다.

[185]

청구범위

- [청구항 1] 다음의 단계를 포함하는 샘플 내 유전자 재배열 분석 방법:
 (a) 피험체의 샘플로부터 복수의 핵산 분자를 포함하는 라이브러리를 획득하는 단계;
 (b) 상기 라이브러리를 복수의 베이트(bait) 세트와 접촉시킴으로써, 사전선택된 서열에 대해 상기 라이브러리를 풍부화시켜서 선택된 핵산 분자를 제공하고, 이로 인해 라이브러리 캐치(library catch)를 제공하는 단계;
 (c) 차세대 시퀀싱 방법에 의해, 상기 라이브러리 캐치의 핵산 분자로부터 리드를 획득하는 단계;
 (d) 정렬 방법에 의해 상기 리드를 정렬하는 단계; 및
 (e) 상기 정렬된 리드를 분석하여 유전자 재배열을 검출하는 단계,
 여기서
 상기 분석 방법은 상기 정렬된 리드를 추출하여 서열 유사성을 분석하는 단계를 포함함.
- [청구항 2] 제1항에 있어서, 상기 리드를 추출하는 단계는 관심영역(region of interest) 정보로 리드를 추출하는 것을 특징으로 하는 방법.
- [청구항 3] 제1항에 있어서, 상기 (e) 단계는 상기 리드를 추출한 다음, 불일치 리드 쌍(discordant read pair)과 일치 리드 쌍(concordant read pair)으로 분리하는 단계를 포함하는 것을 특징으로 하는 방법.
- [청구항 4] 제3항에 있어서, 상기 불일치 리드 쌍과 일치 리드 쌍의 분리는 참조 유전체 정보와 리드를 매칭하여 분리하는 것을 특징으로 하는 방법.
- [청구항 5] 제3항에 있어서, 상기 불일치 리드 쌍의 분리 후, 제1리드의 soft-clip segment 부분을 쿼리로 하여, 쌍을 이루는 제2리드의 매칭 영역을 찾거나, 제2리드의 soft-clip segment를 쿼리로 하여 쌍을 이루는 제1리드의 매칭 영역을 찾는 단계를 추가로 포함하는 것을 특징으로 하는 방법.
- [청구항 6] 제3항에 있어서, 상기 일치 리드 쌍의 분리 후, 제1리드의 2차 정렬 태그 정보를 참조하여 다른 유전자에 2차적으로 맵핑 되는지를 확인하여 얻은 2차 맵핑 지역을 가상의 제2리드로 가정하여, 제1리드의 soft-clip segment 부분을 쿼리로 하여, 쌍을 이루는 제2리드의 매칭 영역을 찾거나, 제2리드의 soft-clip segment를 쿼리로 하여 쌍을 이루는 제1리드의 매칭 영역을 찾는 단계를 추가로 포함하는 것을 특징으로 하는 방법.
- [청구항 7] 제5항에 있어서, 상기 매칭 영역을 찾은 리드를 유전자 재배열 후보군으로 도출하여, 지원 쌍 개수(supporting pair count)를 결정하는 단계.
- [청구항 8] 제6항에 있어서, 상기 매칭 영역을 찾은 리드를 유전자 재배열 후보군으로 도출하여, 지원 쌍 개수(supporting pair count)를 결정하는

- 단계.
- [청구항 9] 제7항 또는 제8항에 있어서, 상기 지원 쌍 개수를 통합하는 단계를 추가로 포함하는 것을 특징으로 하는 방법.
- [청구항 10] 제9항에 있어서, 상기 통합하는 단계는 불일치 리드 쌍 및 일치 리드 쌍에서 지원 쌍 개수가 동시에 결정될 경우, 증가하는 것을 특징으로 하는 방법.
- [청구항 11] 제5항 내지 제10항 중 어느 한 항에 있어서, 상기 지원 쌍 개수는 유전자 재배열의 종류가 같더라도 유전자 재배열 위치가 상이할 경우, 서로 다른 것으로 결정하는 것을 특징으로 하는 방법.
- [청구항 12] 제3항에 있어서, 상기 (e) 단계는 유전자 재배열 후보 리드로 추출되지 않는 리드를 추가 분석을 위해 정리하는 단계를 추가로 포함하는 것을 특징으로 하는 방법.
- [청구항 13] 제12항에 있어서, 상기 (e) 단계는 제5항 또는 제6항에 도출한 유전자 재배열 후보군으로 유전자 재배열 모형(gene rearrangement template)를 제조하는 단계를 추가로 포함하는 것을 특징으로 하는 방법.
- [청구항 14] 제1항에 있어서, 상기 (e) 단계는 제12항에서 정리한 리드의 서열 유사성을 분석하여, 지원 리드 개수(supporting read count)를 결정하는 단계를 추가로 포함하는 것을 특징으로 하는 방법.
- [청구항 15] 제14항에 있어서, 상기 지원 리드 개수는 제12항의 정리된 리드를 blastdb로 하고, 제13항에서 제조한 유전자 재배열 모형을 쿼리로 하여 블라스트를 수행한 다음, 상기 유전자 재배열 모형에서 유전자 재배열의 위치(breakpoint)를 지나가면서 맵핑되는 리드의 개수를 지원 리드 개수로 결정하는 것을 특징으로 하는 방법.
- [청구항 16] 제12항에 있어서, 상기 (e) 단계에서 추출되지 않는 리드는 유전자 재배열 후보군 위치의 5' 방향, 3' 방향 500bp 이내에 존재하는 리드인 것을 특징으로 하는 방법.
- [청구항 17] 제16항에 있어서, 상기 리드는 soft-clip segment를 가지고 있는 것을 특징으로 하는 방법.
- [청구항 18] 제1항에 있어서, 상기 유전자 재배열을 검출하는 단계는 지원 리드 개수(supporting read count)가 5개 이상일 때, 유전자 재배열로 결정하는 것을 특징으로 하는 방법.
- [청구항 19] 차세대 염기 서열 분석(Next Generation Sequencing, NGS)을 이용한 유전자 재배열 검출 방법을 수행할 수 있도록 컴퓨팅 시스템을 제어하기 위한 복수의 명령이 암호화된 컴퓨터 판독 가능한 매체를 포함하는 컴퓨터 시스템으로서,
상기 방법은 (a) 피험체의 샘플로부터 복수의 핵산 분자를 포함하는 라이브러리를 획득하는 단계;
(b) 상기 라이브러리를 복수의 베이트(bait) 세트와 접촉시킴으로써,

사전선택된 서열에 대해 상기 라이브러리를 풍부화시켜서 선택된 핵산 분자를 제공하고, 이로 인해 라이브러리 캐치(library catch)를 제공하는 단계;

(c) 차세대 시퀀싱 방법에 의해, 상기 라이브러리 캐치의 핵산 분자로부터 리드를 획득하는 단계;

(d) 정렬 방법에 의해 상기 리드를 정렬하는 단계; 및

(e) 상기 정렬된 리드를 분석하여 유전자 재배열을 검출하는 단계, 여기서

상기 분석 방법은 상기 정렬된 리드를 추출하여 서열 유사성을 분석하는 단계를 포함함.

[청구항 20]

제1항 내지 제18항 중 어느 한 항에 있어서, 상기 샘플은 하나 이상의 전암성(premalignant) 또는 악성 세포; 고형 종양, 연조직 종양 또는 전이 병변으로부터 선택된 세포; 수술절제면으로부터의 조직 또는 세포; 조직학적으로 정상인 조직; 하나 이상의 혈중 종양 세포(CTC); 정상 인접 조직(normal adjacent tumor: NAT) 및 종양을 가지거나 또는 가질 위험에 있는 동일 피험체로부터의 혈액 샘플로 구성된 군에서 선택되는 것을 특징으로 하는 방법.