



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2018-0083706
(43) 공개일자 2018년07월23일

- | | |
|---|--|
| <p>(51) 국제특허분류(Int. Cl.)
G06F 19/26 (2011.01) C12Q 1/68 (2018.01)</p> <p>(52) CPC특허분류
G06F 19/26 (2013.01)
C12Q 1/6869 (2018.05)</p> <p>(21) 출원번호 10-2017-0006310</p> <p>(22) 출원일자 2017년01월13일
심사청구일자 2017년01월13일</p> | <p>(71) 출원인
주식회사 클리노믹스
울산광역시 울주군 언양읍 유니스트길 50, 경영관 806-6호</p> <p>재단법인 계놈연구재단
충청북도 청주시 흥덕구 오송읍 오송생명1로 194-41, 604호(기업연구관2)</p> <p>울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50</p> <p>(72) 발명자
조윤성
울산광역시 울주군 범서읍 모두박길 9-3, 302호</p> <p>김현호
경기도 수원시 영통구 에듀타운로 101, 103동 1407호(이의동)</p> <p>(74) 대리인
김순용</p> |
|---|--|

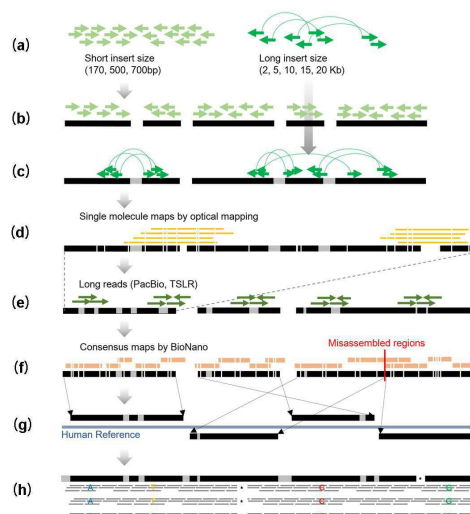
전체 청구항 수 : 총 17 항

(54) 발명의 명칭 공통서열을 포함한 참조표준 게놈지도 구축 장치 및 방법

(57) 요약

공통서열을 포함한 참조표준 게놈지도 구축 장치 및 방법이 개시된다. 본 발명은, NGS(next generation sequencing) 단서열 해독, PacBio 긴서열 해독, Illumina TSLR 긴서열 해독 등의 시퀀싱 및 맵핑 기술을 활용하여 참조표준 게놈지도(reference genome map)를 구축하고, 공통 돌연변이(SNP, indel)들의 정보를 구축된 참조표준 게놈지도에 통합한다. 본 발명에 따르면, NGS(next generation sequencing) 단서열 해독, PacBio 긴서열 해독, Illumina TSLR 긴서열 해독 등의 3개 이상의 시퀀싱 및 맵핑 기술을 활용하여 참조표준 게놈지도(reference genome map)를 구축함으로써 짧은 시간에 적은 비용으로 참조표준 게놈지도를 조립할 수 있고, 공통 돌연변이(SNP, indel)들의 정보를 구축된 참조표준 게놈지도에 통합함으로써, 게놈 상의 각 포지션에 집단 공통(consensus)의 서열을 보유하도록 적용할 수 있다.

대표도 - 도2



(52) CPC특허분류

C12Q 2535/122 (2013.01)

(72) 발명자

김학민

울산광역시 중구 종가2길 5-13, 202호 (유곡동)

전재훈

충청북도 청주시 흥덕구 오송읍 만수1길 4-13, 104호

조성웅

대전광역시 서구 월평북로 11, 209동 503호(월평동, 주공아파트)

박중화

울산광역시 울주군 언양읍 유니스트길 50, 401동 601호(유니스트 교수아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호 10050164

부처명 산업통상자원부

연구관리전문기관 한국산업기술평가관리원

연구사업명 포스트게놈다부처유전체사업

연구과제명 한국인 표준 게놈지도 작성: 한국인 유전체 대동여지도 사업

기 여 율 30/100

주관기관 재단법인 게놈연구재단

연구기간 2014.11.01 ~ 2015.10.31

이 발명을 지원한 국가연구개발사업

과제고유번호 10063239

부처명 산업통상자원부

연구관리전문기관 한국산업기술평가관리원

연구사업명 국가참조표준데이터개발보급사업

연구과제명 국가참조표준데이터개발보급

기 여 율 40/100

주관기관 재단법인 게놈연구재단

연구기간 2016.01.01 ~ 2016.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711045598

부처명 미래창조과학부

연구관리전문기관 울산과학기술원

연구사업명 울산과학기술원 연구운영비 지원사업

연구과제명 대량 집단게놈 연구용 표게놈지도 작성 및 활용 사업

기 여 율 30/100

주관기관 울산과학기술원

연구기간 2016.01.01 ~ 2016.12.31

명세서

청구범위

청구항 1

NGS(next generation sequencing)를 이용하여 미리 설정된 값보다 작은 단편 크기(insert size)를 가지는 단서열과 상기 미리 설정된 값보다 큰 단편 크기(insert size)를 가지는 긴작서열을 생산하는 서열 생산부;

상기 단서열과 상기 긴작서열에서 미리 설정된 리드(read)를 필터링하는 서열 필터링부;

드 부루인(De Bruijn) 그래프를 이용하여 상기 단서열을 기반으로 컨티그(contig)를 조립하는 컨티그 조립부;

상기 컨티그와 상기 긴작서열을 기반으로 스캐폴드(scaffold)를 조립하는 스캐폴드 조립부;

단일 분자 지도(single molecule map)와 상기 스캐폴드(scaffold)를 기반으로 제한 효소(restriction enzyme) 비교를 통해 슈퍼-스캐폴드(super-scaffold)를 조립하고, PacBio 긴서열 해독 방법과 Illumina TSLR 합성 긴서열 해독 방법을 이용하여 상기 슈퍼-스캐폴드(super-scaffold) 상의 갭(gap)을 메꾸며, 상기 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증하는 슈퍼-스캐폴드 조립부; 및

상기 슈퍼-스캐폴드(super-scaffold)의 위치와 방향(strand) 정보를 기반으로 염색체(chromosome) 서열을 조립하는 염색체 서열 조립부;

를 포함하는 공통서열을 포함한 참조표준 게놈지도 구축 장치.

청구항 2

제1항에서,

상기 컨티그 조립부는,

상기 단서열을 기반으로 K-mer 분석을 수행하여 K-mer 빈도 테이블(frequency table)을 획득하고, 상기 K-mer 빈도 테이블(frequency table)을 이용하여 상기 단서열의 에러(error)를 보정하며, 상기 K-mer 빈도 테이블(frequency table)과 상기 드 부루인(De Bruijn) 그래프를 이용하여 에러 보정된 상기 단서열을 기반으로 상기 컨티그(contig)를 조립하는 공통서열을 포함한 참조표준 게놈지도 구축 장치.

청구항 3

제1항에서,

상기 서열 필터링부는,

상기 단서열과 상기 긴작서열에서 중복된 리드(read)를 필터링하고, 상기 단서열과 상기 긴작서열에서 어댑터(adapter) 서열이 포함된 리드(read)를 필터링하며, 상기 단서열과 상기 긴작서열에서 퀄리티 점수(quality score) 값이 미리 설정된 값보다 작은 서열을 필터링하고, 상기 긴작서열에서 접합 어댑터(junction adapter)가 포함된 리드(read)를 필터링하며, 상기 단서열과 상기 긴작서열이 미리 설정된 길이를 가지도록 자르는(trimming) 공통서열을 포함한 참조표준 게놈지도 구축 장치.

청구항 4

제1항에서,

상기 스캐폴드 조립부는,

상기 컨티그에 상기 긴작서열을 정렬(alignment)하여 실제 단편 크기(insert size)를 측정하고, 상기 실제 단편 크기(insert size)를 이용하여 상기 컨티그와 상기 긴작서열을 기반으로 상기 스캐폴드(scaffold)를 조립하며, 상기 스캐폴드(scaffold) 상의 갭(gap)을 메꾸는 공통서열을 포함한 참조표준 게놈지도 구축 장치.

청구항 5

제1항에서,

상기 슈퍼-스캐폴드 조립부는,

상기 제한 효소(restriction enzyme)를 선택하고, 옵티컬 맵핑(Optical mapping)에 의해 생성된 상기 단일 분자 지도(single molecule map)와 상기 스캐폴드(scaffold)를 기반으로 선택된 상기 제한 효소(restriction enzyme) 패턴의 비교를 통해 상기 슈퍼-스캐폴드(super-scaffold)를 조립하는 공통서열을 포함한 참조표준 게놈 지도 구축 장치.

청구항 6

제1항에서,

상기 슈퍼-스캐폴드 조립부는,

나노채널-기반 게놈 맵핑 데이터(nanochannel-based genome mapping data)를 이용하여 상기 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증하는 공통서열을 포함한 참조표준 게놈 지도 구축 장치.

청구항 7

제1항에서,

상기 염색체 서열 조립부는,

인간게놈지도(GRCh38)에 정렬(aligned)된 상기 슈퍼-스캐폴드(super-scaffold)의 위치와 방향(strand) 정보를 기반으로 상기 염색체(chromosome) 서열을 조립하는 공통서열을 포함한 참조표준 게놈 지도 구축 장치.

청구항 8

제1항에서,

전장 게놈 서열 데이터(whole genome re-sequencing data)를 이용하여 상기 염색체(chromosome) 서열을 치환하는 서열 치환부를 더 포함하는 공통서열을 포함한 참조표준 게놈 지도 구축 장치.

청구항 9

공통서열을 포함한 참조표준 게놈 지도 구축 장치의 게놈 지도 구축 방법으로서,

NGS(next generation sequencing)를 이용하여 미리 설정된 값보다 작은 단편 크기(insert size)를 가지는 단서열과 상기 미리 설정된 값보다 큰 단편 크기(insert size)를 가지는 긴단서열을 생산하는 단계;

상기 단서열과 상기 긴단서열에서 미리 설정된 리드(read)를 필터링하는 단계;

드 부루인(De Bruijn) 그래프를 이용하여 상기 단서열을 기반으로 컨티그(contig)를 조립하는 단계;

상기 컨티그와 상기 긴단서열을 기반으로 스캐폴드(scaffold)를 조립하는 단계;

단일 분자 지도(single molecule map)와 상기 스캐폴드(scaffold)를 기반으로 제한 효소(restriction enzyme) 비교를 통해 슈퍼-스캐폴드(super-scaffold)를 조립하는 단계;

PacBio 긴서열 해독 방법과 Illumina TSLR 합성 긴서열 해독 방법을 이용하여 상기 슈퍼-스캐폴드(super-scaffold) 상의 갭(gap)을 메꾸는 단계;

상기 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증하는 단계; 및

상기 슈퍼-스캐폴드(super-scaffold)의 위치와 방향(strand) 정보를 기반으로 염색체(chromosome) 서열을 조립하는 단계;

를 포함하는 공통서열을 포함한 참조표준 게놈 지도 구축 방법.

청구항 10

제9항에서,

상기 컨티그(contig) 조립 단계는,

상기 단서열을 기반으로 K-mer 분석을 수행하여 K-mer 빈도 테이블(frequency table)을 획득하는 단계;

상기 K-mer 빈도 테이블(frequency table)을 이용하여 상기 단서열의 에러(error)를 보정하는 단계; 및

상기 K-mer 빈도 테이블(frequency table)과 상기 드 부루인(De Bruijn) 그래프를 이용하여 에러 보정된 상기 단서열을 기반으로 상기 컨티그(contig)를 조립하는 단계;

를 포함하는 공통서열을 포함한 참조표준 게놈지도 구축 방법.

청구항 11

제9항에서,

상기 리드(read) 필터링 단계는,

상기 단서열과 상기 긴작서열에서 중복된 리드(read)를 필터링하는 단계;

상기 단서열과 상기 긴작서열에서 어댑터(adapter) 서열이 포함된 리드(read)를 필터링하는 단계;

상기 단서열과 상기 긴작서열에서 퀄리티 점수(quality score) 값이 미리 설정된 값보다 작은 서열을 필터링하는 단계;

상기 긴작서열에서 접합 어댑터(junction adapter)가 포함된 리드(read)를 필터링하는 단계; 및

상기 단서열과 상기 긴작서열이 미리 설정된 길이를 가지도록 자르는(trimming) 단계;

를 포함하는 공통서열을 포함한 참조표준 게놈지도 구축 방법.

청구항 12

제9항에서,

상기 스캐폴드(scaffold) 조립 단계는,

상기 컨티그에 상기 긴작서열을 정렬(alignment)하여 실제 단편 크기(insert size)를 측정하는 단계;

상기 실제 단편 크기(insert size)를 이용하여 상기 컨티그와 상기 긴작서열을 기반으로 상기 스캐폴드(scaffold)를 조립하는 단계; 및

상기 스캐폴드(scaffold) 상의 갭(gap)을 메꾸는 단계;

를 포함하는 공통서열을 포함한 참조표준 게놈지도 구축 방법.

청구항 13

제9항에서,

상기 슈퍼-스캐폴드(super-scaffold) 조립 단계는,

상기 제한 효소(restriction enzyme)를 선택하는 단계; 및

옵티컬 맵핑(Optical mapping)에 의해 생성된 상기 단일 분자 지도(single molecule map)와 상기 스캐폴드(scaffold)를 기반으로 선택된 상기 제한 효소(restriction enzyme) 패턴의 비교를 통해 상기 슈퍼-스캐폴드(super-scaffold)를 조립하는 단계;

를 포함하는 공통서열을 포함한 참조표준 게놈지도 구축 방법.

청구항 14

제9항에서,

상기 슈퍼-스캐폴드(super-scaffold) 검증 단계는,

나노채널-기반 게놈 맵핑 데이터(nanochannel-based genome mapping data)를 이용하여 상기 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증하는 것으로 이루어진 공통서열을 포함한 참조표준 게놈지도 구축 방법.

청구항 15

제9항에서,

상기 염색체(chromosome) 서열 조립 단계는,

인간게놈지도(GRCh38)에 정렬(aligned)된 상기 슈퍼-스캐폴드(super-scaffold)의 위치와 방향(strand) 정보를 기반으로 상기 염색체(chromosome) 서열을 조립하는 것으로 이루어진 공통서열을 포함한 참조표준 게놈지도 구축 방법.

청구항 16

제9항에서,

전장 게놈 서열 데이터(whole genome re-sequencing data)를 이용하여 상기 염색체(chromosome) 서열을 치환하는 단계를 더 포함하는 공통서열을 포함한 참조표준 게놈지도 구축 방법.

청구항 17

제9항 내지 제16항 중 어느 한 항에 기재된 공통서열을 포함한 참조표준 게놈지도 구축 방법을 컴퓨터에서 실행시키기 위하여 컴퓨터로 읽을 수 있는 기록 매체에 저장된 컴퓨터 프로그램.

발명의 설명

기술 분야

[0001] 본 발명은 공통서열을 포함한 참조표준 게놈지도 구축 장치 및 방법에 관한 것으로, 더욱 상세하게는 NGS(next generation sequencing) 단서열 해독, PacBio 긴서열 해독, Illumina TSLR 긴서열 해독 등의 시퀀싱 및 맵핑 기술을 활용하여 참조표준 게놈지도(reference genome map)를 구축하고, 공통 돌연변이(SNP, indel)들의 정보를 구축된 참조표준 게놈지도에 통합하는 하는 장치 및 방법에 관한 것이다.

배경 기술

[0002] 게놈(genome)이란 한 개체의 유전 정보의 총합을 의미한다. 참조표준 게놈 지도(reference genome map)는 한 종 혹은 한 개체의 게놈 상의 모든 염기 서열 정보 및 이의 위치 정보를 의미한다. 서로 다른 종(species)은 진화적으로 서로 다른 게놈 서열을 보유하고 있으며, 같은 종의 다른 개체 간에도 유사도는 높지만 서로 다른 게놈 서열을 보유하고 있다. 예컨대, 인간과 침팬지는 진화적으로 조상을 공통으로 공유하고, 약 6백만년 전에 분기했다고 알려져 있는데, 게놈 서열 뿐만 아니라 게놈의 구조도 서로 상이(인간의 염색체 2번은 침팬지와 같은 유인원에서는 두 개의 염색체(2A, 2B)로 나뉘어져 존재함)하다. 또한, 인간 내에서도 구조 변이(structural variation) 및 카피수 변이(copy number variation)와 같이 서로 다른 게놈 구조가 존재한다. 2009년 12월경에는 중국의 BGI-Shenzhen 연구소에서 NGS(next generation sequencing)를 이용하여 아시아인(중국인) 한 명과 흑인 한 명의 참조표준 게놈지도 초안(스캐폴드 서열)을 발표하였다. 그 결과 백인의 게놈 대비 아시아인 게놈의 다른 부분을 제시하였고, 이 결과는 각 인종별 게놈지도의 필요성을 의미한다.

[0003] 참조표준 게놈지도는 다른 게놈 서열 분석을 함에 있어, 기준/표준 역할을 하는데, 전장 게놈 서열(whole genome re-sequencing) 방식에 의해 생성된 짧은 DNA 단편 서열의 위치를 참조표준 게놈지도에 맵핑(mapping)함으로써, 돌연변이를 발굴할 수 있게 된다. SNV(Single nucleotide variation)란 참조표준 게놈지도 대비 한 개의 염기가 다른 돌연변이이고, indel(insertion or deletion)은 짧은(~50bp) 염기 서열의 길이 차이가 발생하는 돌연변이를 의미한다. 이러한 SNV 및 indel은 개개인/개체별 차이를 만드는 원인이며, 집단의 특징, 질병 발생 등과 관련이 있어, 질병 예측 및 질환 마커 발굴 등에 매우 중요하다.

[0004] 현재까지 공개된 모든 de novo 게놈 조립 방식으로 생성된 참조표준 게놈지도는 1명 또는 1개체로부터 추출한 시료를 주로 사용하여 조립되었는데, 특정 개인/개체는 개인/개체 특이적인 게놈 구조 및 서열을 보유하고 있다. 즉, 1명의 DNA 서열을 활용하여 게놈지도를 구축하는 경우, 집단의 대표성을 띄는 참조표준 게놈지도를 만들 수 없을 뿐만 아니라, 개인 특이적인 서열이 게놈지도 상에 존재하게 되어, 이후 전장 게놈 서열 데이터(whole genome re-sequencing data)를 분석하는데 있어, 많은 수의 불필요한 돌연변이들이 찾아지게 된다.

[0005] 인간의 참조표준 게놈지도의 경우, 지금까지 약 10여개의 게놈지도가 발표되어 있다. 가장 품질이 좋고, 인간 게놈 분석의 표준 역할을 하는 인간게놈지도(human reference; 현재 버전 GRCh38)의 경우, 약 50명의 참여자들의 시료를 대상으로 BAC(bacterial artificial chromosome) 클론을 제작하여, 다양한 인종 및 개인을 포함한 게놈지도를 구축하였으나, 실제로는 RPCI-11(또는 RP11)이라는 개인의 BAC 클론이 주로(74.3%) 사용되어 해독 및 조립되었다(아래의 [표 1] 참조). 또한, 인간 게놈지도는 여러명의 대표 서열로 조립한 것이 아니고, 각 게놈 포지션에는 1명의 BAC 클론만이 사용되었다.

표 1

[0006]

Library	Fraction	Ethnicity
CTB	0.016	Caucasian
CTC	0.021	Caucasian
CTD	0.043	East Asian
RP1	0.028	Caucasian
RP3	0.016	Caucasian
RP4	0.022	Caucasian
RP5	0.027	Caucasian
RP11	0.743	Caucasian/African
total	0.916	

[0007] 다른 인간 게놈지도의 경우에도 모두 1명의 DNA 시료를 활용하여 조립되었다. 아래의 [표 2]는 현재까지 발표된 주요 인간 참조표준 게놈지도이다. 더불어, 지금까지 공개된 여러 게놈지도들의 경우, 주로 1개 또는 2개 이내의 주요 게놈 실험 방식을 활용하였다는 특징이 있다.

표 2

[0008]

게놈지도	게놈 해독 방식	참고논문
GRCh38	인간게놈지도. BAC라이브를 구축하여, Sanger 해독방식으로 게놈해독 및 조립됨	Lander, E. S. <i>et al.</i> Initial sequencing and analysis of the human genome. Nature 409, 860921 (2001).
AK1	PacBio 긴서열 및 BioNano 맵으로 조립	Seo, J. S. <i>et al.</i> De novo assembly and phasing of a Korean human genome. Nature 538, 243247 (2016).
HX1	PacBio 긴서열 및 BioNano map으로 조립	Shi, L. <i>et al.</i> Long-read sequencing and de novo assembly of a Chinese genome. Nat. Commun. 7, 12065 (2016).
ASM101398v1	PacBio 긴서열 및 BioNano map으로 조립	Pendleton, M. <i>et al.</i> Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat. Methods 12, 780786 (2015).
CHM1_PacBio_r2	PacBio 긴서열로 조립	Chaisson, M. J. <i>et al.</i> Resolving the complexity of the human genome using single-molecule sequencing. Nature 517, 608611 (2015).
HsapALLPATHS1	NGS 방식으로 해독 조립	Gnerre, S. <i>et al.</i> High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl Acad. Sci. USA 108, 15131518(2011).
HuRef	생어해독방식으로 조립	Levy, S. <i>et al.</i> The diploid genome sequence of an individual human. PLoS Biol. 5, e254 (2007).
Mongolian	NGS 방식으로 해독 조립	Bai, H. <i>et al.</i> The genome of a Mongolian individual reveals the genetic imprints of Mongolians on modern human populations. Genome Biol. Evol. 6, 31223136 (2014).
YH / YH_2.0	NGS 방식으로 해독 조립	Li, R. <i>et al.</i> De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20, 265272 (2010).
African	NGS 방식으로 해독 조립	

[0009] Dewey et al. 논문(Dewey, F. E. et al. Phased whole-genome genetic risk in a family quartet using a

major allele reference sequence. PLoS Genet. 7, e1002280 (2011).)에서는 기 공개된 인간게놈지도를 대상으로, 1000 Genome Project에서 생산한 수십 ~ 수백명의 전장 게놈 서열 데이터(whole genome re-sequencing data)로부터 규명한 SNV를 인간 게놈지도에 통합시켜, 인종별(백인, 흑인, 동아시아인)로 공통(consensus) 서열을 만들어서 발표를 하였다. 그러나, Dewey et al. 논문의 경우, 사용한 전장 게놈 서열 데이터(whole genome re-sequencing data)의 시퀀싱 배수(깊이)가 낮은 데이터를 활용하였고, 또한, 공통(consensus) 서열을 SNV에 대해서만 적용시켰다. 또한, Dewey et al. 논문의 내용은 de novo assembly에 대한 것이 아니고, 기존 발표된 게놈지도에 공통 서열을 적용시킨 방법에 대한 것이다. 이에 반면, 본 발명은 de novo assembly된 게놈지도에 high depth의 전장 게놈 서열 데이터(whole genome re-sequencing data)를 사용하여, SNV와 더불어 indel에 대해서도 집단의 공통 서열을 적용시켰다는 점에서 차이가 있다.

[0010] 본 발명은 현재까지 발표된 다양한 게놈 실험 방식(NGS 단서열 해독, PacBio 긴서열 해독, Illumina TSLR 긴서열 해독, OpGen whole genome maps, BioNano maps)을 동시에 적용하여, 고품질의 1명의 참조표준 게놈지도를 구축하는 방법에 대한 것이다. 그리고, 본 발명은 집단의 공통 서열을 보유하는 참조표준 게놈지도를 구축하기 위해, 많은 수의 high-depth 전장 게놈 서열 데이터(whole genome re-sequencing data)로부터 추출된 돌연변이(SNV, indel)를 통합시켜 집단공통 참조표준 게놈지도를 구축하는 방법에 대한 것이다.

선행기술문헌

특허문헌

[0011] (특허문헌 0001) 한국공개특허 제2006-0052710호 (더 리전트 오브 더 유니버시티 오브 캘리포니아) 2006. 5. 19. 특허문헌 1은 기능성 DNA 요소와 세포 단백질의 게놈 지도작성으로서, 특허문헌 1에는 전체 게놈(가령, 하나 이상의 크로모솜 또는 크로모솜 영역과 같은 전체 게놈 또는 일부 구역)에서 DNA에 대한 단백질의 결합을 검사하는 방법에 대한 내용이 개시되어 있다.

발명의 내용

해결하려는 과제

[0012] 본 발명이 이루고자 하는 기술적 과제는, NGS(next generation sequencing) 단서열 해독, PacBio 긴서열 해독, Illumina TSLR 긴서열 해독 등의 시퀀싱 및 맵핑 기술을 활용하여 참조표준 게놈지도(reference genome map)를 구축하고, 공통 돌연변이(SNP, indel)들의 정보를 구축된 참조표준 게놈지도에 통합하는 공통서열을 포함한 참조표준 게놈지도 구축 장치 및 방법을 제공하는 데 있다.

과제의 해결 수단

[0013] 상기의 기술적 과제를 달성하기 위한 본 발명에 따른 공통서열을 포함한 참조표준 게놈지도 구축 장치는, NGS(next generation sequencing)를 이용하여 미리 설정된 값보다 작은 단편 크기(insert size)를 가지는 단서열과 상기 미리 설정된 값보다 큰 단편 크기(insert size)를 가지는 긴서열을 생산하는 서열 생산부; 상기 단서열과 상기 긴서열에서 미리 설정된 리드(read)를 필터링하는 서열 필터링부; 드 부루인(De Bruijn) 그래프를 이용하여 상기 단서열을 기반으로 컨티그(contig)를 조립하는 컨티그 조립부; 상기 컨티그와 상기 긴서열을 기반으로 스캐폴드(scaffold)를 조립하는 스캐폴드 조립부; 단일 분자 지도(single molecule map)와 상기 스캐폴드(scaffold)를 기반으로 제한 효소(restriction enzyme) 비교를 통해 슈퍼-스캐폴드(super-scaffold)를 조립하고, PacBio 긴서열 해독 방법과 Illumina TSLR 합성 긴서열 해독 방법을 이용하여 상기 슈퍼-스캐폴드(super-scaffold) 상의 갭(gap)을 메꾸며, 상기 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증하는 슈퍼-스캐폴드 조립부; 및 상기 슈퍼-스캐폴드(super-scaffold)의 위치와 방향(strand) 정보를 기반으로 염색체(chromosome) 서열을 조립하는 염색체 서열 조립부;를 포함한다.

[0014] 상기 컨티그 조립부는, 상기 단서열을 기반으로 K-mer 분석을 수행하여 K-mer 빈도 테이블(frequency table)을 획득하고, 상기 K-mer 빈도 테이블(frequency table)을 이용하여 상기 단서열의 에러(error)를 보정하며, 상기 K-mer 빈도 테이블(frequency table)과 상기 드 부루인(De Bruijn) 그래프를 이용하여 에러 보정된 상기 단서열을 기반으로 상기 컨티그(contig)를 조립할 수 있다.

[0015] 상기 서열 필터링부는, 상기 단서열과 상기 긴서열에서 중복된 리드(read)를 필터링하고, 상기 단서열과 상기

긴짜서열에서 어댑터(adapter) 서열이 포함된 리드(read)를 필터링하며, 상기 단서열과 상기 긴짜서열에서 퀄리티 점수(quality score) 값이 미리 설정된 값보다 작은 서열을 필터링하고, 상기 긴짜서열에서 접합 어댑터(junction adapter)가 포함된 리드(read)를 필터링하며, 상기 단서열과 상기 긴짜서열이 미리 설정된 길이를 가지도록 자를(trimming) 수 있다.

- [0016] 상기 스캐폴드 조립부는, 상기 컨티그에 상기 긴짜서열을 정렬(alignment)하여 실제 단편 크기(insert size)를 측정하고, 상기 실제 단편 크기(insert size)를 이용하여 상기 컨티그와 상기 긴짜서열을 기반으로 상기 스캐폴드(scaffold)를 조립하며, 상기 스캐폴드(scaffold) 상의 갭(gap)을 메꿀 수 있다.
- [0017] 상기 슈퍼-스캐폴드 조립부는, 상기 제한 효소(restriction enzyme)를 선택하고, 옵티컬 맵핑(Optical mapping)에 의해 생성된 상기 단일 분자 지도(single molecule map)와 상기 스캐폴드(scaffold)를 기반으로 선택된 상기 제한 효소(restriction enzyme) 패턴의 비교를 통해 상기 슈퍼-스캐폴드(super-scaffold)를 조립할 수 있다.
- [0018] 상기 슈퍼-스캐폴드 조립부는, 나노채널-기반 게놈 맵핑 데이터(nanochannel-based genome mapping data)를 이용하여 상기 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증할 수 있다.
- [0019] 상기 염색체 서열 조립부는, 인간게놈지도(GRCh38)에 정렬(aligned)된 상기 슈퍼-스캐폴드(super-scaffold)의 위치와 방향(strand) 정보를 기반으로 상기 염색체(chromosome) 서열을 조립할 수 있다.
- [0020] 전장 게놈 서열 데이터(whole genome re-sequencing data)를 이용하여 상기 염색체(chromosome) 서열을 치환하는 서열 치환부를 더 포함할 수 있다.
- [0021] 상기의 기술적 과제를 달성하기 위한 본 발명에 따른 공통서열을 포함하는 참조표준 게놈지도 구축 방법은, 참조표준 게놈지도 구축 장치의 게놈지도 구축 방법으로서, NGS(next generation sequencing)를 이용하여 미리 설정된 값보다 작은 단편 크기(insert size)를 가지는 단서열과 상기 미리 설정된 값보다 큰 단편 크기(insert size)를 가지는 긴짜서열을 생산하는 단계; 상기 단서열과 상기 긴짜서열에서 미리 설정된 리드(read)를 필터링하는 단계; 드 부루인(De Bruijn) 그래프를 이용하여 상기 단서열을 기반으로 컨티그(contig)를 조립하는 단계; 상기 컨티그와 상기 긴짜서열을 기반으로 스캐폴드(scaffold)를 조립하는 단계; 단일 분자 지도(single molecule map)와 상기 스캐폴드(scaffold)를 기반으로 제한 효소(restriction enzyme) 비교를 통해 슈퍼-스캐폴드(super-scaffold)를 조립하는 단계; PacBio 긴서열 해독 방법과 Illumina TSLR 합성 긴서열 해독 방법을 이용하여 상기 슈퍼-스캐폴드(super-scaffold) 상의 갭(gap)을 메꾸는 단계; 상기 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증하는 단계; 및 상기 슈퍼-스캐폴드(super-scaffold)의 위치와 방향(strand) 정보를 기반으로 염색체(chromosome) 서열을 조립하는 단계;를 포함한다.
- [0022] 상기 컨티그(contig) 조립 단계는, 상기 단서열을 기반으로 K-mer 분석을 수행하여 K-mer 빈도 테이블(frequency table)을 획득하는 단계; 상기 K-mer 빈도 테이블(frequency table)을 이용하여 상기 단서열의 에러(error)를 보정하는 단계; 및 상기 K-mer 빈도 테이블(frequency table)과 상기 드 부루인(De Bruijn) 그래프를 이용하여 에러 보정된 상기 단서열을 기반으로 상기 컨티그(contig)를 조립하는 단계;를 포함할 수 있다.
- [0023] 상기 리드(read) 필터링 단계는, 상기 단서열과 상기 긴짜서열에서 중복된 리드(read)를 필터링하는 단계; 상기 단서열과 상기 긴짜서열에서 어댑터(adapter) 서열이 포함된 리드(read)를 필터링하는 단계; 상기 단서열과 상기 긴짜서열에서 퀄리티 점수(quality score) 값이 미리 설정된 값보다 작은 서열을 필터링하는 단계; 상기 긴짜서열에서 접합 어댑터(junction adapter)가 포함된 리드(read)를 필터링하는 단계; 및 상기 단서열과 상기 긴짜서열이 미리 설정된 길이를 가지도록 자르는(trimming) 단계;를 포함할 수 있다.
- [0024] 상기 스캐폴드(scaffold) 조립 단계는, 상기 컨티그에 상기 긴짜서열을 정렬(alignment)하여 실제 단편 크기(insert size)를 측정하는 단계; 상기 실제 단편 크기(insert size)를 이용하여 상기 컨티그와 상기 긴짜서열을 기반으로 상기 스캐폴드(scaffold)를 조립하는 단계; 및 상기 스캐폴드(scaffold) 상의 갭(gap)을 메꾸는 단계;를 포함할 수 있다.
- [0025] 상기 슈퍼-스캐폴드(super-scaffold) 조립 단계는, 상기 제한 효소(restriction enzyme)를 선택하는 단계; 및 옵티컬 맵핑(Optical mapping)에 의해 생성된 상기 단일 분자 지도(single molecule map)와 상기 스캐폴드(scaffold)를 기반으로 선택된 상기 제한 효소(restriction enzyme) 패턴의 비교를 통해 상기 슈퍼-스캐폴드(super-scaffold)를 조립하는 단계;를 포함할 수 있다.
- [0026] 상기 슈퍼-스캐폴드(super-scaffold) 검증 단계는, 나노채널-기반 게놈 맵핑 데이터(nanochannel-based genome

mapping data)를 이용하여 상기 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증하는 것으로 이루어질 수 있다.

[0027] 상기 염색체(chromosome) 서열 조립 단계는, 인간게놈지도(GRCh38)에 정렬(aligned)된 상기 슈퍼-스캐폴드(super-scaffold)의 위치와 방향(strand) 정보를 기반으로 상기 염색체(chromosome) 서열을 조립하는 것으로 이루어질 수 있다.

[0028] 전장 게놈 서열 데이터(whole genome re-sequencing data)를 이용하여 상기 염색체(chromosome) 서열을 치환하는 단계를 더 포함할 수 있다.

[0029] 상기의 기술적 과제를 달성하기 위한 본 발명에 따른 컴퓨터 프로그램은 컴퓨터로 읽을 수 있는 기록 매체에 저장되어 상기한 방법 중 어느 하나를 컴퓨터에서 실행시킨다.

발명의 효과

[0030] 본 발명에 따른 공통서열을 포함한 참조표준 게놈지도 구축 장치 및 방법에 의하면, NGS(next generation sequencing) 단서열 해독, PacBio 긴서열 해독, Illumina TSLR 긴서열 해독 등의 3개 이상의 시퀀싱 및 맵핑 기술을 활용하여 참조표준 게놈지도(reference genome map)를 구축함으로써, 짧은 시간에 적은 비용으로 참조표준 게놈지도를 조립할 수 있다. 아울러, 공통 돌연변이(SNP, indel)들의 정보를 구축된 참조표준 게놈지도에 통합함으로써, 게놈 상의 각 포지션에 집단 공통(consensus)의 서열을 보유하도록 적용할 수 있다.

도면의 간단한 설명

[0031] 도 1은 본 발명의 바람직한 실시예에 따른 공통서열을 포함한 참조표준 게놈지도 구축 장치를 설명하기 위한 블록도이다.

도 2는 본 발명의 바람직한 실시예에 따른 참조표준 게놈지도 구축 과정의 일례를 설명하기 위한 도면이다.

도 3은 본 발명의 바람직한 실시예에 따른 슈퍼-스캐폴드의 검증 결과의 일례를 설명하기 위한 도면이다.

도 4는 본 발명의 바람직한 실시예에 따른 염색체 서열의 조립 결과의 일례를 설명하기 위한 도면이다.

도 5는 본 발명의 바람직한 실시예에 따른 참조표준 게놈 지도의 효과를 설명하기 위한 그래프이다.

도 6은 본 발명의 바람직한 실시예에 따른 공통서열을 포함한 참조표준 게놈지도 구축 방법을 설명하기 위한 흐름도이다.

도 7은 도 6에 도시한 서열 필터링 단계를 보다 자세히 나타낸 흐름도이다.

도 8은 도 6에 도시한 컨티그 조립 단계를 보다 자세히 나타낸 흐름도이다.

도 9는 도 6에 도시한 스캐폴드 조립 단계를 보다 자세히 나타낸 흐름도이다.

도 10은 도 6에 도시한 슈퍼-스캐폴드 조립 단계를 보다 자세히 나타낸 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0032] 이하에서 첨부한 도면을 참조하여 본 발명에 따른 공통서열을 포함한 참조표준 게놈지도 구축 장치 및 방법의 바람직한 실시예에 대해 상세하게 설명한다.

[0033] 먼저, 도 1 내지 도 4를 참조하여 본 발명의 바람직한 실시예에 따른 공통서열을 포함한 참조표준 게놈지도 구축 장치에 대하여 설명한다.

[0034] 도 1은 본 발명의 바람직한 실시예에 따른 공통서열을 포함한 참조표준 게놈지도 구축 장치를 설명하기 위한 블록도이고, 도 2는 본 발명의 바람직한 실시예에 따른 참조표준 게놈지도 구축 과정의 일례를 설명하기 위한 도면이며, 도 3은 본 발명의 바람직한 실시예에 따른 슈퍼-스캐폴드의 검증 결과의 일례를 설명하기 위한 도면이고, 도 4는 본 발명의 바람직한 실시예에 따른 염색체 서열의 조립 결과의 일례를 설명하기 위한 도면이다.

[0035] 도 1을 참조하면, 본 발명의 바람직한 실시예에 따른 공통서열을 포함한 참조표준 게놈지도 구축 장치(100)(이하 '게놈지도 구축 장치'라 한다)는 NGS(next generation sequencing) 단서열 해독, PacBio 긴서열 해독, Illumina TSLR 긴서열 해독 등의 3개 이상의 시퀀싱 및 맵핑 기술을 활용하여 참조표준 게놈지도(reference genome map)를 구축한다. 그리고, 게놈지도 구축 장치(100)는 공통 돌연변이(SNP, indel)들의 정보를 구축된

참조표준 게놈지도에 통합할 수 있다.

[0036] 이를 위해, 게놈지도 구축 장치(100)는 서열 생산부(110), 서열 필터링부(120), 컨티그 조립부(130), 스캐폴드 조립부(140), 슈퍼-스캐폴드 조립부(150), 염색체 서열 조립부(160) 및 서열 치환부(170)를 포함할 수 있다.

[0037] 서열 생산부(110)는 NGS(next generation sequencing)를 이용하여 미리 설정된 값보다 작은 단편 크기(insert size)를 가지는 단서열(short insert[short paired-end])과 미리 설정된 값보다 큰 단편 크기(insert size)를 가지는 긴쌍서열(long insert[long-mate pair])을 생산한다. 여기서, 미리 설정된 값은 1Kb(단편 크기)일 수 있다. 즉, 서열 생산부(110)는 1Kb 보다 작은 단편 크기(insert size)를 가지는 단서열과 1Kb 보다 큰 단편 크기(insert size)를 가지는 긴쌍서열을 생산할 수 있다.

[0038] 다시 설명하면, 도 2의 (a)에 도시된 바와 같이, 서열 생산부(110)는 NGS 방식을 이용하여 게놈지도의 초안(scaffold)을 구축하기 위해 단서열(short insert [short paired-end], insert size 1 Kb 미만)과 긴쌍서열(long insert [long-mate pair], insert size 1 Kb 이상)을 모두 생산할 수 있다. 이때, DNA 라이브러리 제작은, 단서열의 경우 Illumina TruSeq® DNA Sample Preparation Guide, July 2012, 긴쌍서열의 경우 Illumina Nextera® Mate Pair Sample Preparation Guide, January 2013에 기재된 방식을 활용할 수 있다. 본 발명에서는 단서열의 경우 단편 크기(insert size) 기준 170bp, 500bp, 700bp의 라이브러리, 긴쌍서열의 경우 단편 크기(insert size) 기준 2Kb, 5Kb, 10Kb, 15Kb, 20Kb 라이브러리가 구축 및 해독되었다. 이에 따른 NGS 서열 해독 결과는 아래의 [표 3]과 같다.

표 3

[0039]

Type	Insert size	Read length (bp)	Number of read pairs	Total data (Gb)	Sequence depth (X)	
Short-insert size libraries	170bp	101	254,562,947	51.42	16.59	48.69
			246,624,330	49.82	16.07	
			246,007,078	49.70	16.03	
	500bp	101	246,418,836	49.78	16.06	46.71
			230,109,465	46.48	14.99	
			240,361,539	48.55	15.66	
	700bp	101	207,193,678	41.85	13.50	39.17
			188,159,956	38.01	12.26	
			205,873,335	41.59	13.41	
Long-mate pair libraries	2Kb	101	196,290,337	39.65	12.79	38.22
			232,858,099	47.04	15.17	
			157,507,662	31.82	10.26	
	5Kb	101	152,201,289	30.74	9.92	32.81
			177,874,430	35.93	11.59	
			173,383,733	35.02	11.30	
	10Kb	101	205,215,277	41.45	13.37	40.05
			209,859,354	42.39	13.67	
			199,617,521	40.32	13.01	
	15Kb	101	156,336,183	31.58	10.19	30.65
			166,036,249	33.54	10.82	
			147,927,209	29.88	9.64	
	20Kb	101	181,506,276	36.66	11.83	34.72
			177,434,679	35.84	11.56	
			173,929,946	35.13	11.33	
Total			4,773,289,408	964.19	311.02	311.02

[0041] 서열 필터링부(120)는 단서열과 긴쌍서열에서 미리 설정된 리드(read)를 필터링한다. 여기서, 미리 설정된 리드(read)는 중복된 리드(read), 어댑터(adapter) 서열이 포함된 리드(read), 퀄리티 점수(quality score) 값이 미리 설정된 값보다 작은 서열, 접합 어댑터(junction adapter)가 포함된 리드(read) 등을 말한다.

[0042] 다시 설명하면, 서열 필터링부(120)는 Illumina 해독 기계에서 생산된 단서열과 긴쌍서열로부터 정확한 표준 게

놈지도를 얻기 위해, 해독된 단서열 및 긴작서열 중 PCR 과정에서 생기는 중복된 리드(read)와 어댑터(adapter)가 포함된 리드(read), 퀄리티(quality)가 낮은 리드(read) 등을 제거할 수 있다. 이와 같은 과정을 위해 본 발명에서는 각각 PrinSeq, SOAPfilter, cutadapt 프로그램들이 사용되었다.

[0044] 즉, 서열 필터링부(120)는 단서열과 긴작서열에서 중복된 리드(read)를 필터링할 수 있다. 다시 설명하면, 단서열과 긴작서열의 라이브러리제작 과정 중, 서열 해독을 용이하게 하기 위해 PCR을 이용하여 DNA 조각들을 증폭하는 과정이 있다. 이 과정에서 라이브러리에 동일한 서열이 중복으로 서열 해독이 되어 정확한 참조표준 게놈지도를 만드는데 불필요한 요소로 작용되게 된다. 이러한 중복된 리드(read)를 제거하기 위해 서열 필터링부(120)는 필터링을 수행한다. 이때, 본 발명에서 사용된 프로그램은 PrinSeq(lite-0.20.4)이다.

[0045] 그리고, 서열 필터링부(120)는 단서열과 긴작서열에서 어댑터(adapter) 서열이 포함된 리드(read)를 필터링할 수 있다. 다시 설명하면, 단서열과 긴작서열의 서열 해독 과정 중 리드(read)에 어댑터(adapter) 서열이 (Forward adapter : GATCGGAAGAGCACACGTCTGAACTCCAGTCAC, Reverse adapter : GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT) 포함된 리드(read)가 존재하며, 정확한 참조표준 게놈지도 작성을 위해 서열 필터링부(120)는 이와 같은 리드(read)들을 제거한다. 이때, 본 발명에서 사용된 프로그램은 SOAPfilter이다.

[0046] 또한, 서열 필터링부(120)는 단서열과 긴작서열에서 퀄리티 점수(quality score) 값이 미리 설정된 값보다 작은 서열을 필터링할 수 있다. 다시 설명하면, 해독된 단서열과 긴작서열은 각 서열마다 해당하는 퀄리티 점수(quality score) 값을 가지게 된다. 정확도 99% 이상의 서열을 얻기 위해 퀄리티 점수(Quality Score) 20을 기준으로 퀄리티 점수(Quality Score) 값이 리드(read)당 평균 20 미만일 때, 서열 필터링부(120)는 해당 서열을 필터링한다. 또한, 각 리드(read)에 포함된 N base(ambiguous base) 개수의 비율이 해당 리드(read) 길이의 5% 이상일 때, 서열 필터링부(120)는 해당 서열을 필터링한다. 이때, 본 발명에서 사용된 프로그램은 PrinSeq(lite-0.20.4)이다.

[0047] 그리고, 서열 필터링부(120)는 긴작서열에서 접합 어댑터(junction adapter)가 포함된 리드(read)를 필터링할 수 있다. 다시 설명하면, 해독된 긴작서열의 경우 실험 과정상 접합 어댑터(junction adapter)가 랜덤하게 서열에 포함될 수 있다. 서열 필터링부(120)는 접합 어댑터(junction adapter)가 포함된 리드(read)를 잘라(trimming) 참조표준 게놈지도 오조립에 영향을 줄 수 있는 요소를 제거한다. 이때, 본 발명에서 사용된 프로그램은 cutadapt(v1.1)이다.

[0048] 또한, 서열 필터링부(120)는 단서열과 상기 긴작서열이 미리 설정된 길이를 가지도록 자를(trimming) 수 있다. 다시 설명하면, 서열 필터링부(120)는 서열 해독기의 특성상, 해독된 서열의 5' 말단 및 3' 말단 부분의 서열 해독 퀄리티(quality)가 떨어지는 부분과 양 말단에 나타나는 Poly-A tail과 같은 서열 조립에 영향을 줄 수 있는 부분을 일괄적으로 잘라낸다. 예컨대, 단서열의 경우 90bp가 되도록 5' 말단 부분의 3base 및 3' 말단 부분의 8base를 잘라(trimming)내고, 긴작서열의 경우 3' 말단 부분을 기준으로 리드(read) 길이가 49bp가 되도록 잘라(trimming)낸다. 이때, 본 발명에서 사용된 프로그램은 자체적으로 개발한 script를 사용하였다.

[0049] 이에 따른 서열 필터링 결과는 아래의 [표 4]와 같다.

표 4

[0050]

Type	Insert size	Read length (bp)	Number of read pairs	Total data (Gb)	Sequence Depth (X)	
Short-insert size libraries	170bp	90	238,901,578	43.00	13.87	40
			225,934,916	40.67	13.12	
			224,145,725	40.35	13.01	
	500bp	90	220,100,704	39.62	12.78	37.57
			207,716,033	37.39	12.06	
			219,165,329	39.45	12.73	
	700bp	90	189,043,000	34.03	10.98	32.24
			173,545,699	31.24	10.08	
			192,535,557	34.66	11.18	

Long-mate pair libraries	2Kb	49	102,368,796	10.03	3.24	9.64
			118,485,351	11.61	3.75	
			83,704,400	8.20	2.65	
	5Kb	49	74,199,538	7.27	2.35	8.08
			93,060,115	9.12	2.94	
			88,156,446	8.64	2.79	
	10Kb	49	52,521,514	5.15	1.66	5.03
			54,759,429	5.37	1.73	
			51,874,811	5.08	1.64	
	15Kb	49	60,904,413	5.97	1.93	5.3
			55,631,632	5.45	1.76	
			51,042,581	5.00	1.61	
	20Kb	49	20,374,949	2.00	0.64	2.08
			26,561,512	2.60	0.84	
			19,032,195	1.87	0.60	
Total			2,843,766,223	433.77	139.94	139.94

[0052] 컨티그 조립부(130)는 드 부루인(De Bruijn) 그래프를 이용하여 단서열을 기반으로 컨티그(contig)를 조립한다.

[0053] 즉, 컨티그 조립부(130)는 단서열을 기반으로 K-mer 분석을 수행하여 K-mer 빈도 테이블(frequency table)을 획득한다. 다시 설명하면, 정확한 참조표준 게놈지도 작성을 위해서 해독된 단서열을 이용하여 게놈의 크기를 측정한다. 게놈 크기 측정을 위해서는 K-mer 분석을 수행하여야 하는데, 정해진 K 값에 따라 단서열의 리드(read)들을 K 값의 길이만큼 쪼개서, 쪼개진 조각의 개수를 이용하여 K-mer 빈도 테이블(frequency table)을 획득하고 게놈 크기를 측정한다. K-mer=17일 경우, A, T, G, C가 우연히 존재할 확률인 4^{17} , 약 16Gb의 게놈 크기까지 분석이 되는데, 시퀀싱 반복 배수가 충분한 경우 K-mer 크기가 클수록 분석의 정확도가 올라간다. 본 발명에 따른 참조표준 게놈지도 작성과정에서는 더 정확한 결과를 얻기 위해, K-mer=23을 이용하여 분석하였다. 이때, 본 발명에서 사용된 프로그램은 SOAPec이다.

[0054] 그리고, 컨티그 조립부(130)는 K-mer 빈도 테이블(frequency table)을 이용하여 상기 단서열의 에러(error)를 보정한다. 다시 설명하면, 정확한 컨티그(contig) 조립을 위해 K-mer 분석 과정에서 생성된 K-mer 빈도 테이블(frequency table)을 이용하여, 배수(depth)가 낮은 부분의 K-mer 조각들에 해당되는 단서열은 이형 접합적(heterozygous)이거나, 시퀀싱 과정의 에러(error)로 생각할 수 있기에 보정(correction)한다. 이때, 본 발명에서는 단서열로 해독한 170bp, 500bp, 700bp 라이브러리에 각각 적용하였으며, 사용된 프로그램은 SOAPec이다. 이에 따른 에러 보정(error correction) 결과는 아래의 [표 5]와 같다.

표 5

Insert Size	Library	Error corrected bases ratio
170bp	KR01_PE_170_L1_1	0.0569%
	KR01_PE_170_L1_2	0.0640%
	KR01_PE_170_L2_1	0.0725%
	KR01_PE_170_L2_2	0.1675%
	KR01_PE_170_L3_1	0.0716%
	KR01_PE_170_L3_2	0.1715%
500bp	KR01_PE_500_L1_1	0.0729%
	KR01_PE_500_L1_2	0.2081%
	KR01_PE_500_L2_1	0.0684%
	KR01_PE_500_L2_2	0.1718%
	KR01_PE_500_L3_1	0.0840%
	KR01_PE_500_L3_2	0.1615%

700bp	KR01_PE_700_L1_1	0.1074%
	KR01_PE_700_L1_2	0.2794%
	KR01_PE_700_L2_1	0.1182%
	KR01_PE_700_L2_2	0.2401%
	KR01_PE_700_L3_1	0.0757%
	KR01_PE_700_L3_2	0.2625%

[0056] 또한, 컨티그 조립부(130)는 K-mer 빈도 테이블(frequency table)과 드 부루인(De Bruijn) 그래프를 이용하여 에러 보정(error correction)된 단서열을 기반으로 컨티그(contig)를 조립한다. 다시 설명하면, 도 2의 (b)에 도시된 바와 같이, 컨티그 조립부(130)는 서열 필터링 및 에러 보정(error correction)된 단서열을 이용하여 컨티그(contig)를 조립한다. 이때, 본 발명에서는 여러 조립 알고리즘 중 드 부루인(de Bruijn) 그래프 방식을 이용한 프로그램인 SOAPdenovo2(r240)를 사용하여 컨티그(contig)를 조립한다. 즉, 총 두 가지의 과정(pregraph 과정 : K-mer 값에 따라 K-mer의 빈도 테이블(frequency table)을 생성함 / contig 과정 : 버비빈도 테이블(frequency table) 및 드 부루인(de Bruijn) 그래프 알고리즘을 이용하여 컨티그(contig)를 생성함)을 거쳐서 컨티그(contig)가 조립된다. 최적의 K-mer 값을 찾기 위해, 29, 39, 49, 55, 59, 63, 69, 75, 79의 K-mer 값에 대해 컨티그(contig) 조립을 수행(다양한 K-mer 값에 따라 조립 결과를 확인 후, K=55를 활용하여 이후 분석을 수행함)한다. 이에 따른 컨티그(contig) 조립 결과는 아래의 [표 6]과 같다.

표 6

[0057]

K-mer size	All sequences			Longer than 100 bp		
	Total size	Longest	N50	Total size	Longest	N50
29	5,187,304,717	16,946	90	2,275,359,750	16,946	1,099
39	4,459,796,947	35,726	300	2,529,816,579	35,726	1,939
49	4,066,593,737	51,838	980	2,740,134,913	51,838	2,375
55	3,860,731,497	44,789	1,447	2,915,054,629	44,789	2,559
59	3,744,446,380	48,982	1,773	2,990,197,206	48,982	2,735
63	3,641,677,654	54,683	2,113	3,029,961,853	54,683	2,964
69	3,524,281,519	54,689	2,589	3,072,247,309	54,689	3,295
75	3,429,622,648	62,488	2,918	3,097,380,667	62,488	3,466
79	3,343,414,611	80,399	2,789	3,086,359,621	80,399	3,187

[0059] 스캐폴드 조립부(140)는 컨티그(contig)와 긴작서열을 기반으로 스캐폴드(scaffold)를 조립한다.

[0060] 즉, 스캐폴드 조립부(140)는 컨티그(contig)에 긴작서열을 정렬(alignment)하여 실제 단편 크기(insert size)를 측정한다. 다시 설명하면, 해독된 단서열, 긴작서열의 실제 단편 크기(insert size)를 구하기 위해 앞서 생성된 컨티그(contig)에 단서열과 긴작서열을 정렬(alignment)하여 각 라이브러리별 실제 단편 크기(insert size)의 평균과 표준 편차를 측정한다. 이때, 본 발명에서 사용된 정렬(alignment) 프로그램은 Burrows-Wheeler Aligner(BWA, v0.7.7)와 Samtools(v.0.1.19)이다. 생성된 BAM 파일의 9번째 컬럼에 해당하는 값(1번 리드(read)와 2번 리드(read) 간의 정렬(alignment) 거리)을 이용하여 각 라이브러리별 단편 크기(insert size)를 측정한다. 이때, 본 발명에서 사용된 프로그램은 자체적으로 개발한 script를 사용하였다.

[0061] 그리고, 스캐폴드 조립부(140)는 실제 단편 크기(insert size)를 이용하여 컨티그(contig)와 긴작서열을 기반으로 스캐폴드(scaffold)를 조립한다. 다시 설명하면, 도 2의 (c)에 도시된 바와 같이, 앞서 생성된 컨티그(contig) 서열과 단서열 및 긴작서열, 그리고 앞서 계산된 단편 크기(insert size)의 값을 이용하여 스캐폴드(scaffold)를 조립한다. 이때, 입력해야 하는 단편 크기(insert size) 값의 경우, 단편 크기(insert size) 평균의 20%에 해당하는 값을 표준 편차로 사용하였다. 스캐폴드 조립부(140)는 총 두 가지 과정(map 과정 : 컨티그(contig) 서열에 단서열 및 긴작서열을 정렬(alignment)함 / scaff 과정 : 컨티그(contig)에 정렬(alignment)된 결과와 단편 크기(insert size)의 정보를 이용하여 스캐폴드(scaffold)를 조립)을 거쳐서 스캐폴드(scaffold)를 조립한다. 이때, 본 발명에서 사용된 프로그램은 SOAPdenovo2(r240)이다. 이에 따른 스캐폴드(scaffold) 조립 결과는 아래의 [표 7]과 같다.

표 7

[0062]

	Scaffold	
	Size (Mb)	No.
N90	3.09	178
N80	6.45	116
N70	10.45	81
N60	16.16	59
N50	19.85	42
Longest	81.91	-
Gaps	1.65 %	-
Total (≥ 200bp)	2.92 Gb	68,170
Total (≥10 Kb)	2.88 Gb	1,243

[0063]

또한, 스캐폴드 조립부(140)는 스캐폴드(scaffold) 상의 갭(gap)을 메꾼다. 다시 설명하면, 생성된 스캐폴드(scaffold) 서열은 단서열 및 긴작서열의 단편 크기(insert size) 만큼의 거리를 채우지 못한 부분(N base)이 다수 존재하게 되는데, 이 부분을 갭(gap)이라고 한다. 이러한 갭(gap) 부분의 서열을 채우기 위해 gap closing 작업을 2번 반복해서 수행한다. 이때, 본 발명에서 사용된 프로그램은 gapcloser이다.

[0065]

슈퍼-스캐폴드 조립부(150)는 단일 분자 지도(single molecule map)와 스캐폴드(scaffold)를 기반으로 슈퍼-스캐폴드(super-scaffold)를 조립한다. 다시 설명하면, 스캐폴드(scaffold)를 슈퍼-스캐폴드(super-scaffold)로 확장시키기 위해, OpGen사의 전장 게놈 옵티컬 맵핑 데이터(whole genome optical mapping data)를 획득한다.

[0066]

즉, 슈퍼-스캐폴드 조립부(150)는 단일 분자 지도(single molecule map)와 상기 스캐폴드(scaffold)를 기반으로 제한 효소(restriction enzyme) 비교를 통해 슈퍼-스캐폴드(super-scaffold)를 조립한다.

[0067]

이때, 슈퍼-스캐폴드 조립부(150)는 제한 효소(restriction enzyme)를 선택할 수 있다. 다시 설명하면, 조립된 스캐폴드(scaffold) 서열을 대상으로 적합한 제한 효소(restriction enzyme)를 선택한다. 이때, 본 발명에서는 여러 제한효소를 대상으로 average fragment size(AFS), usable sequence information의 양, large fragments의 크기와 양을 측정하여 최종적으로 적합한 제한 효소(restriction enzyme)를 Genome-Builder™ 분석을 통하여 선정한다.

[0068]

제한 효소(restriction enzyme)의 구체적인 선정 방법은 다음과 같다.

[0069]

- usable sequence information % (5-20 Kb) : 90 % 이상

[0070]

- usable sequence information % (6-15 Kb) : 70 % 이상

[0071]

- usable sequence information % (6-12 Kb) : 60 % 이상

[0072]

- average fragment size (AFS) : 5 Kb 이상

[0073]

위의 조건을 만족하는 제한 효소(restriction enzyme) 중, number of fragments > 100 Kb의 개수와 large fragment의 크기가 큰 것을 선정한다.

[0074]

이에 따른 제한 효소(restriction enzyme)의 선택 결과(SpeI 제한 효소 선택)는 아래의 [표 8]과 같다.

표 8

[0075]

Enzyme	Usable% 5Kb-20Kb	Usable% 6Kb-15Kb	Usable% 6Kb-12Kb	Ave. Frags size (kb)	# of Frags > 100kb	Max Frag size (Kb)
AflIII	25.12	10.31	10.07	4.58	4	117.49
BamHI	94.94	82.36	72.76	8.08	19	159.82
KpnI	98.76	91.89	69.64	10.35	50	154.09
NcoI	17.1	3.37	3.35	3.85	0	84.46
NheI	98.08	89.26	65.1	10.67	62	149.61
SpeI	94.8	73.17	67.9	7.44	63	196.12
BglII	7.01	2.12	2.07	3.79	1	104.69

EcoRI	7.86	2.87	2.85	3.65	0	71.37
MluI	0.76	0.23	0.09	130.62	9422	1529.97
NdeI	12.35	6.4	6.21	3.25	3	105.73
PvuII	2.2	0.4	0.4	2.7	3	149.7
XbaI	9.27	3.33	3.26	3.64	3	147.38
XhoI	26.46	11.1	4.88	23.64	2612	372.38

[0076]

그런 다음, 슈퍼-스캐폴드 조립부(150)는 옵티컬 맵핑(Optical mapping)에 의해 생성된 단일 분자 지도(single molecule map)와 스캐폴드(scaffold)를 기반으로 선택된 제한 효소(restriction enzyme) 패턴의 비교를 통해 슈퍼-스캐폴드(super-scaffold)를 조립할 수 있다. 다시 설명하면, 도 2의 (d)에 도시된 바와 같이, 스캐폴드(scaffold) 서열과 옵티컬 맵핑(optical mapping) 방식으로부터 생성된 단일 분자 지도(single molecule map)를 제한 효소 패턴(restriction enzyme pattern) 비교를 통하여 함께 조립한다. 이때, 본 발명에서는 Genome Builder™를 활용하여, 스캐폴드(scaffold) 상의 선택된 제한 효소(restriction enzyme)의 위치를 규명한다. 정확한 슈퍼-스캐폴드(super-scaffold) 조립을 위하여 스캐폴드(scaffold)는 200 Kb 초과 크기만 대상으로 하며, 단일 분자 지도(single molecule map)는 250 Kb 이상의 크기만 활용한다. 이후, 스캐폴드(scaffold)의 제한 효소(restriction enzyme)의 위치와 단일 분자 지도(single molecule map)의 제한 효소(restriction enzyme)의 위치 정보를 활용하여, extension이 가능한 스캐폴드(scaffold)와 스캐폴드(scaffold)를 서로 연결시켜, 슈퍼-스캐폴드(super-scaffold)를 조립한다. 이에 따른 단일 분자 지도(single molecule map) 생성 결과는 아래의 [표 9]와 같고, 슈퍼-스캐폴드(super-scaffold) 조립 결과는 아래의 [표 10]과 같다.

표 9

[0077]

Summary of SMRM data	Maps used in analysis
Total Size (Gb)	745.51
Number of Molecules	2,071,951
Average Size of Molecules (Kb)	359.81
Minimum molecule size (Kb)	250
Average Size of Fragments (Kb)	13.24

표 10

[0078]

	Whole-genome optical mapping을 활용한 super-scaffold 조립결과	
	Size (Mb)	No.
N90	3.86	140
N80	9.45	92
N70	14.47	67
N60	19.56	49
N50	25.93	36
Longest	101.22	-
Gaps	1.75 %	-
Total (≥ 200bp)	2.92 Gb	68,103
Total (≥10 Kb)	2.88 Gb	1,176

[0080]

그리고, 슈퍼-스캐폴드 조립부(150)는 PacBio 긴서열 해독 방법과 Illumina TSLR 합성 긴서열 해독 방법을 이용하여 상기 슈퍼-스캐폴드(super-scaffold) 상의 갭(gap)을 메꾼다. 다시 설명하면, 도 2의 (e)에 도시된 바와 같이, 슈퍼-스캐폴드(super-scaffold) 상의 갭(gap)(염기 서열 'N'으로 표기)을 메꾸기 위하여, 두 종류의 긴서열 해독을 수행한다. 이때, 본 발명에서는 PacBio 긴서열 해독 방식과 Illumina TSLR 합성긴서열 해독 방식을 활용한다. 여러 긴서열을 동시에 입력하며, consensus sequence(여러 서열 정보를 합치는 방식)를 만드는 방식으로 긴서열의 오류가 최소화 되도록 PBJelly2 프로그램(version 14.9.9)을 default option으로 수행한다. 이에 따른 PacBio 긴서열 해독 결과는 아래의 [표 11](P4C2 chemistry) 및 [표 12](P5C3 chemistry)와 같고,

Illumina TSLR 합성 긴서열 해독 결과는 아래의 [표 13]과 같으며, 슈퍼-스캐폴드(super-scaffold) 상의 갭(gap)을 메꾼 결과는 아래의 [표 14]와 같다.

표 11

[0081]

Size	Number of bases (bp)	Number of reads	Mean length (bp)
~2kb	2,200,375,125	2,023,326	1,088
~3kb	2,598,138,881	1,054,927	2,463
~4kb	2,253,729,183	650,819	3,463
~5kb	1,993,913,569	445,503	4,476
~6kb	1,868,335,867	341,037	5,478
~7kb	1,692,679,373	261,244	6,479
~8kb	1,490,151,540	199,293	7,477
~9kb	1,264,147,938	149,166	8,475
~10kb	1,025,254,470	108,261	9,470
10kb~	2,404,653,532	202,921	11,850
Total	18,791,379,478	5,436,497	3,457

표 12

[0082]

Region	Number of bases (bp)	Number of reads	Mean length (bp)
~2kb	376,691,922	352,650	1,068
~3kb	448,189,058	179,744	2,493
~4kb	581,090,138	166,158	3,497
~5kb	707,030,086	157,272	4,496
~6kb	815,006,427	148,315	5,495
~7kb	905,881,157	139,481	6,495
~8kb	978,965,060	130,607	7,496
~9kb	1,063,290,046	125,158	8,496
~10kb	1,084,089,752	114,232	9,490
10kb~	5,347,185,274	406,019	13,170
Total	12,307,418,920	1,919,636	6,411

표 13

[0083]

Region	Number of bases (bp)	Number of reads	Mean length (bp)
~2kb	1,745,885,089	1,627,362	1,073
~3kb	1,227,839,348	498,112	2,465
~4kb	1,200,052,670	345,449	3,474
~5kb	1,170,624,980	261,313	4,480
~6kb	1,141,935,546	208,259	5,483
~7kb	1,132,652,780	174,578	6,488
~8kb	1,358,992,691	181,044	7,506
~9kb	2,532,232,743	294,819	8,589
~10kb	2,879,791,577	304,656	9,453
10kb~	1,910,098,184	181,128	10,546
Total	16,300,105,608	4,076,720	3,998

표 14

[0084]

	긴서열을 활용한 갭 클로징(gap closing) (PacBio and TSLR)	
	Size (Mb)	No.
N90	3.53	143
N80	9.26	93

N70	14.53	67
N60	19.36	50
N50	26.08	36
Longest	101.48	-
Gaps	1.06 %	-
Total (≥ 200bp)	2.94 Gb	68,451
Total (≥10 Kb)	2.90 Gb	1,369

[0086]

또한, 슈퍼-스캐폴드 조립부(150)는 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증한다. 즉, 슈퍼-스캐폴드 조립부(150)는 나노채널-기반 게놈 맵핑 데이터(nanochannel-based genome mapping data)를 이용하여 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증할 수 있다. 다시 설명하면, 도 2의 (f)에 도시된 바와 같이, BioNano Genomics의 나노채널-기반 게놈 맵핑 데이터(nanochannel-based genome mapping data)를 활용하여, 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증한다. 이때, 슈퍼-스캐폴드(super-scaffold)는 10 Kb 초과 크기만을 대상으로 한다. 게놈 맵핑 데이터(genome mapping data)를 BioNano Genomics Irys genome mapping system을 활용하여 consensus genome maps으로 조립을 하고, irysView 소프트웨어(version 2.1.0.30787)의 리눅스 명령어를 사용하여 BioNano consensus genome maps과 조립된 슈퍼-스캐폴드(super-scaffold)들의 게놈 구조를 비교한다. 오조립 영역을 검증하기 위해서, BioNano consensus genome maps은 인간게놈지도(GRCh38)와 비교한다. 모든 정렬(alignments) 결과를 수작업으로(manually) 확인하여, 오조립 영역을 규명하고 분리시킨다. 또한, BioNano consensus map도 오조립이 발생할 수 있으므로, align confidence score 20 이상인 consensus map을 대상으로 오조립 검증을 수행하며, 슈퍼-스캐폴드(super-scaffold), BioNano consensus map, 및 GRCh38의 구조를 서로 비교하여, 오조립 영역을 규명한다. 이에 따른 BioNano 게놈 지도(genome map) 생산 결과는 아래의 [표 15]와 같고, BioNano 게놈 지도(genome map)를 이용한 슈퍼-스캐폴드(super-scaffold)의 검증 결과는 도 3과 같다. 도 3에 도시된 바와 같이, 가장 길게 조립된 슈퍼-스캐폴드(super-scaffold)(도 3의 초록색 블록)에 대한 BioNano 게놈 지도(genome map)(도 3의 파란색 블록)를 통해 슈퍼-스캐폴드(super-scaffold)를 검증하였다.

표 15

[0087]

	BioNano single molecules	BioNano consensus maps
Total data	210 Gb	-
Single molecule N50	273 Kb	-
Molecules above 150Kb	145 Gb	-
Coverage depth	45 ×	-
Assembly size	-	2.78 Gb
Consensus map N50	-	1.12 Mb

[0089]

염색체 서열 조립부(160)는 슈퍼-스캐폴드(super-scaffold)의 위치와 방향(strand) 정보를 기반으로 염색체(chromosome) 서열을 조립한다. 즉, 염색체 서열 조립부(160)는 인간게놈지도(GRCh38)에 정렬(aligned)된 슈퍼-스캐폴드(super-scaffold)의 위치와 방향(strand) 정보를 기반으로 염색체(chromosome) 서열을 조립할 수 있다.

[0090]

다시 설명하면, 도 2의 (g)에 도시된 바와 같이, 조립 및 검증을 한 슈퍼-스캐폴드(super-scaffold) 서열을 염색체(chromosome) 서열로 확장시키기 위하여, 검증한 서열을 인간게놈지도(GRCh38)를 대상으로 SyMap 프로그램(v4.2)을 디폴트 비교 파라미터(default comparison parameters)로 전장 게놈 정렬(whole genome alignment)을 수행한다. 이때, 게놈 상의 반복 서열에 의한 bias를 제거하기 위하여, 10 Kb 이상의 슈퍼-스캐폴드(super-scaffold)를 대상으로 전장 게놈 정렬(whole genome alignment)을 수행한다. 맵핑되지 않은(unmapped) 슈퍼-스캐폴드들(super-scaffolds)은 GRCh38에 mapped anchor number를 4 이상의 조건으로 하여, 재정렬(re-align)시킨다. 크기가 작은 슈퍼-스캐폴드들(super-scaffolds)(200 bp ~ 10 Kb)은 BLASR 프로그램을 default option으로 수행하여, GRCh38에 정렬(aligned)시킨다. 이때, 맵핑 퀄리티(mapping quality) = 254인 경우만 사용한다. 인간게놈지도(GRCh38)에 정렬(aligned)된 슈퍼-스캐폴드(super-scaffold)의 위치와 방향(strand) 정보

를 활용하여, 염색체(chromosome) 서열로 확장한다. 슈퍼-스캐폴드(super-scaffold) 간의 갭(gap) 정보는 인간게놈지도 상에서의 빈 영역의 길이를 사용하며, 만약, 슈퍼-스캐폴드(super-scaffold)의 위치가 겹치는 경우에는 10 Kb의 임의의 갭(gap)을 할당한다. 염색체(chromosome)의 양 끝에 10 Kb 갭(gap) 서열을 추가하여 telomeric region의 서열을 할당한다. GRCh38에 정렬(aligned)되지 않아, 염색체(chromosome) 상의 위치를 알 수 없는 슈퍼-스캐폴드들(super-scaffolds)은 chrUn 그룹에 위치시킨다. 이때, 본 발명에서 염색체(chromosome) 서열 조립은 자체적으로 개발한 script를 사용하였다. 이에 따른 인간게놈지도(GRCh38)와 조립/검증된 슈퍼-스캐폴드(super-scaffold)의 구조 비교 결과는 도 4와 같고, 염색체(chromosome) 조립 결과는 아래의 [표 16]과 같다.

표 16

[0091]

	Chromosomes	
	*Unplaced scaffolds were excluded.	
	Size (Mb)	No.
N90	81.54	19
N80	103.05	16
N70	136.43	13
N60	137.59	11
N50	155.88	8
Longest	251.92	-
Gaps	9.44 %	-
Total (≥ 200bp)	3.12 Gb	24
Total (≥10 Kb)	3.12 Gb	24

[0093]

서열 치환부(170)는 전장 게놈 서열 데이터(whole genome re-sequencing data)를 이용하여 염색체(chromosome) 서열을 치환한다.

[0094]

다시 설명하면, 도 2의 (h)에 도시된 바와 같이, 집단 공통서열 치환에 활용한 전장 게놈 서열 데이터(whole genome re-sequencing data)를 이용하여, 조립된 염색체(chromosome) 서열을 치환한다. 본 발명에서는 아래의 [표 17]과 같은 40명의 전장 게놈 서열 데이터(whole genome re-sequencing data)를 활용하였다.

표 17

[0095]

Sample ID	Total number of raw reads	Mapped read depth (except 'N')	Read mapping rate (%)	Homozygous SNVs	Homozygous INDELs	Heterozygous SNVs	Heterozygous INDELs	All variants
KPGP-00002	98,317,515,960	27.64	99.29	962,066	146,462	2,958,707	292,082	4,359,317
KPGP-00006	93,448,081,980	24.73	99.28	1,431,527	204,234	2,915,971	276,219	4,827,951
KPGP-00032	112,190,946,660	30.36	99.29	1,444,163	215,475	2,955,815	296,145	4,911,598
KPGP-00033	108,196,466,760	29.95	99.30	1,406,058	211,651	2,961,708	297,035	4,876,452
KPGP-00039	101,141,448,400	30.19	99.16	1,391,102	212,028	2,991,047	315,678	4,909,855
KPGP-00056	111,361,334,200	32.24	99.34	1,419,373	230,317	3,100,438	340,429	5,090,557
KPGP-00086	102,626,322,600	29.88	99.34	1,423,097	228,216	3,074,640	335,156	5,061,109
KPGP-00125	118,670,365,980	33.12	99.31	1,438,747	211,687	2,932,733	291,074	4,874,241
KPGP-00127	118,883,354,760	32.81	99.33	1,416,527	206,959	2,948,523	288,104	4,860,113
KPGP-00128	117,849,278,700	32.76	99.29	1,407,530	208,532	2,941,634	292,805	4,850,501

KPGP-00129	107,124,150, 780	29.96	99.28	1,440,746	203,979	2,908,731	271,108	4,824,564
KPGP-00131	120,142,829, 340	33.36	99.29	1,432,319	211,261	2,970,372	289,604	4,903,556
KPGP-00132	122,237,363, 160	33.93	99.30	1,411,276	210,946	2,946,694	297,988	4,866,904
KPGP-00134	119,540,641, 320	32.54	99.28	1,416,157	207,904	2,931,855	288,305	4,844,221
KPGP-00136	114,984,689, 940	30.71	99.30	1,429,777	204,804	2,940,492	274,170	4,849,243
KPGP-00137	118,027,255, 140	32.97	99.28	1,403,331	207,581	2,940,643	289,256	4,840,811
KPGP-00138	123,868,546, 380	33.39	99.32	1,398,902	207,327	2,938,964	289,045	4,834,238
KPGP-00139	105,730,760, 700	29.32	99.28	1,397,287	207,216	2,918,240	291,707	4,814,450
KPGP-00141	111,508,577, 820	31.41	99.24	1,405,400	207,892	2,926,108	288,957	4,828,357
KPGP-00142	125,024,326, 200	32.62	99.29	1,443,241	211,075	2,943,175	292,818	4,890,309
KPGP-00144	127,001,127, 600	33.96	99.30	1,422,369	211,512	2,973,541	296,396	4,903,818
KPGP-00145	111,861,808, 380	31.18	99.29	1,438,003	210,730	2,953,375	293,052	4,895,160
KPGP-00205 -B01-G	123,835,438, 866	37.24	98.41	1,422,423	221,835	3,072,207	332,313	5,048,778
KPGP-00220	106,317,727, 560	28.21	99.28	1,411,132	201,485	2,931,702	284,397	4,828,716
KPGP-00227	115,164,844, 920	34.39	99.30	1,419,518	217,159	3,039,274	308,248	4,984,199
KPGP-00228	112,898,405, 520	33.34	99.30	1,455,818	221,343	3,052,488	303,008	5,032,657
KPGP-00230	110,458,697, 940	32.86	99.31	1,414,415	214,448	3,031,789	301,182	4,961,834
KPGP-00232	109,620,112, 860	32.01	99.29	1,442,223	214,897	3,020,544	292,548	4,970,212
KPGP-00233	107,091,428, 940	32.08	99.27	1,421,451	216,917	3,014,334	302,473	4,955,175
KPGP-00235	114,400,539, 900	34.74	99.31	1,414,391	218,911	3,047,216	309,518	4,990,036
KPGP-00245 -B01-G-PE5 00	102,078,086, 860	31.40	99.11	1,465,527	223,235	3,031,190	322,301	5,042,253
KPGP-00254	122,277,928, 000	34.56	99.24	1,427,301	221,720	3,080,569	313,709	5,043,299
KPGP-00255	102,221,657, 600	29.67	99.34	1,414,140	227,857	3,083,228	336,527	5,061,752
KPGP-00256	127,033,362, 000	36.61	99.35	1,422,753	235,874	3,174,628	355,538	5,188,793
KPGP-00265 -B01-G-P50 0	90,922,729,4 00	27.53	99.29	1,414,977	216,811	2,964,359	306,126	4,902,273
KPGP-00266 -B01-G-P50 0	91,666,078,8 00	27.38	99.32	1,374,215	212,665	2,962,424	307,516	4,856,820
KPGP-00269 -B01-G-PE5 00	100,240,975, 874	30.81	99.32	1,449,250	219,822	3,052,622	324,886	5,046,580
KPGP-00317 -B01-G-PE5 00	103,075,371, 660	26.76	87.15	1,400,454	208,300	3,002,602	306,055	4,917,411
KPGP-00318 -B01-G-PE5 00	101,805,865, 370	28.22	95.42	1,440,304	218,383	2,971,844	319,451	4,949,982

KPGP-00319 -B01-G-PE5 00	100,957,938, 100	27.77	97.17	1,403,626	213,564	3,063,114	315,785	4,996,089
--------------------------------	---------------------	-------	-------	-----------	---------	-----------	---------	-----------

[0096]

전장 게놈 서열(whole genome re-sequencing)을 BWA-MEM 프로그램(version 0.7.8)을 default option으로 조립된 염색체(chromosome) 서열에 맵핑(mapping)한다. 이에 따른 맵핑 결과는 아래의 [표 18]과 같다.

표 18

[0097]

Sample ID	Total amount of raw reads	Mapped read depth (except 'N')	Read mapping rate (%)	Homozygous SNP	Homozygous INDEL	Heterozygous SNP	Heterozygous INDEL	All variants
KPGP-00002	98,317,515,960	27.64	99.29	962,066	146,462	2,958,707	292,082	4,359,317
KPGP-00006	93,448,081,980	24.73	99.28	1,431,527	204,234	2,915,971	276,219	4,827,951
KPGP-00032	112,190,946,660	30.36	99.29	1,444,163	215,475	2,955,815	296,145	4,911,598
KPGP-00033	108,196,466,760	29.95	99.30	1,406,058	211,651	2,961,708	297,035	4,876,452
KPGP-00039	101,141,448,400	30.19	99.16	1,391,102	212,028	2,991,047	315,678	4,909,855
KPGP-00056	111,361,334,200	32.24	99.34	1,419,373	230,317	3,100,438	340,429	5,090,557
KPGP-00086	102,626,322,600	29.88	99.34	1,423,097	228,216	3,074,640	335,156	5,061,109
KPGP-00125	118,670,365,980	33.12	99.31	1,438,747	211,687	2,932,733	291,074	4,874,241
KPGP-00127	118,883,354,760	32.81	99.33	1,416,527	206,959	2,948,523	288,104	4,860,113
KPGP-00128	117,849,278,700	32.76	99.29	1,407,530	208,532	2,941,634	292,805	4,850,501
KPGP-00129	107,124,150,780	29.96	99.28	1,440,746	203,979	2,908,731	271,108	4,824,564
KPGP-00131	120,142,829,340	33.36	99.29	1,432,319	211,261	2,970,372	289,604	4,903,556
KPGP-00132	122,237,363,160	33.93	99.30	1,411,276	210,946	2,946,694	297,988	4,866,904
KPGP-00134	119,540,641,320	32.54	99.28	1,416,157	207,904	2,931,855	288,305	4,844,221
KPGP-00136	114,984,689,940	30.71	99.30	1,429,777	204,804	2,940,492	274,170	4,849,243
KPGP-00137	118,027,255,140	32.97	99.28	1,403,331	207,581	2,940,643	289,256	4,840,811
KPGP-00138	123,868,546,380	33.39	99.32	1,398,902	207,327	2,938,964	289,045	4,834,238
KPGP-00139	105,730,760,700	29.32	99.28	1,397,287	207,216	2,918,240	291,707	4,814,450
KPGP-00141	111,508,577,820	31.41	99.24	1,405,400	207,892	2,926,108	288,957	4,828,357
KPGP-00142	125,024,326,200	32.62	99.29	1,443,241	211,075	2,943,175	292,818	4,890,309
KPGP-00144	127,001,127,600	33.96	99.30	1,422,369	211,512	2,973,541	296,396	4,903,818
KPGP-00145	111,861,808,380	31.18	99.29	1,438,003	210,730	2,953,375	293,052	4,895,160
KPGP-00205-B01-G	123,835,438,866	37.24	98.41	1,422,423	221,835	3,072,207	332,313	5,048,778

KPGP-00220	106,317,727,560	28.21	99.28	1,411,132	201,485	2,931,702	284,397	4,828,716
KPGP-00227	115,164,844,920	34.39	99.30	1,419,518	217,159	3,039,274	308,248	4,984,199
KPGP-00228	112,898,405,520	33.34	99.30	1,455,818	221,343	3,052,488	303,008	5,032,657
KPGP-00230	110,458,697,940	32.86	99.31	1,414,415	214,448	3,031,789	301,182	4,961,834
KPGP-00232	109,620,112,860	32.01	99.29	1,442,223	214,897	3,020,544	292,548	4,970,212
KPGP-00233	107,091,428,940	32.08	99.27	1,421,451	216,917	3,014,334	302,473	4,955,175
KPGP-00235	114,400,539,900	34.74	99.31	1,414,391	218,911	3,047,216	309,518	4,990,036
KPGP-00245-B01-G-PE500	102,078,086,860	31.40	99.11	1,465,527	223,235	3,031,190	322,301	5,042,253
KPGP-00254	122,277,928,000	34.56	99.24	1,427,301	221,720	3,080,569	313,709	5,043,299
KPGP-00255	102,221,657,600	29.67	99.34	1,414,140	227,857	3,083,228	336,527	5,061,752
KPGP-00256	127,033,362,000	36.61	99.35	1,422,753	235,874	3,174,628	355,538	5,188,793
KPGP-00265-B01-G-P500	90,922,729,400	27.53	99.29	1,414,977	216,811	2,964,359	306,126	4,902,273
KPGP-00266-B01-G-P500	91,666,078,800	27.38	99.32	1,374,215	212,665	2,962,424	307,516	4,856,820
KPGP-00269-B01-G-PE500	100,240,975,874	30.81	99.32	1,449,250	219,822	3,052,622	324,886	5,046,580
KPGP-00317-B01-G-PE500	103,075,371,660	26.76	87.15	1,400,454	208,300	3,002,602	306,055	4,917,411
KPGP-00318-B01-G-PE500	101,805,865,370	28.22	95.42	1,440,304	218,383	2,971,844	319,451	4,949,982
KPGP-00319-B01-G-PE500	100,957,938,100	27.77	97.17	1,403,626	213,564	3,063,114	315,785	4,996,089

[0098] 맵핑 결과 중 동일한 포지션에 일치하게 맵핑된 결과를 제거한다. 맵핑 퀄리티를 향상시키기 위하여 IndelRealigner를 수행하고, base quality scores는 GATK의 TableRecalibration 알고리즘을 사용하여 재교정(recalibration)한다. 한국인 전장 게놈 서열(whole genome re-sequencing)의 단일 염기 서열 변이(single nucleotide variation; SNV)와 삽입 및 결실(small insertion or deletion; indel) 변이를 Genome Analysis Toolkit(GATK, version 2.3.9) 프로그램을 이용하여 규명한다. 이에 따른 공통 서열 치환에 사용된 변이의 결과는 아래의 [표 19]와 같다.

표 19

SNVs	indels	Total
1,951,986	219,728	2,171,714

[0100] SNV의 경우, 각 염기 위치별 allele 비율을 측정하고, 전체 전장 게놈 서열(n) * 2 haploids = 2n 중, 가장 높은 빈도로 발견되는 염기 서열이 KOREF 염기 서열과 다를 경우 치환한다. indel의 경우, 전체 전장 게놈 서열

(n) * 2 haploids = 2n 중, 50% 이상에서 공통적으로 발견되는 경우에 치환한다. 성염색체(X, Y 염색체)의 경우, X 염색체는 여성 1인당 2n, 남성 1인당 1n으로부터 공통 변이를 측정하고, Y 염색체는 남성 1인당 1n으로부터 공통 변이를 측정한다. 이때, 본 발명에서 사용된 프로그램은 자체적으로 개발한 script를 사용하였다.

< 상염색체 (1 ~ 22 염색체) >

- 상염색체 대상 haploids 수 : 전장게놈서열 샘플수 (n) * 2 haploids = 2n

- 대표 SNV 선정 : 2n 개의 haploid 중 가장 높은 빈도로 발생하는 염기 서열(최고 높은 빈도의 염기가 2개 이상일 경우, KOREF 샘플과 동일한 염기 서열 선택)

- 공통 indel 선정 : 2n 개의 haploid 중 50% 이상 공통적으로 발견되는 indel 변이

< 성염색체 (X, Y 염색체) >

- X 염색체 대상 haploids 수 : (남성 전장게놈서열 샘플수 (n) * 1 haploid) + (여성 전장게놈서열 샘플수 (m) * 2 haploids) = n + 2m

- Y 염색체 대상 haploids 수 : (남성 전장게놈서열 샘플수 (n) * 1 haploid) = n

- 대표 SNV 선정 : n + 2m 개의 haploid 중 가장 높은 빈도로 발생하는 염기 서열(최고 높은 빈도의 염기가 2개 이상일 경우, KOREF 샘플과 동일한 염기 서열 선택)

- 공통 indel 선정 : n 개의 haploid 중 50% 이상 공통적으로 발견되는 indel 변이

위와 같은 단계별 참조표준 게놈지도의 조립 결과를 정리하면 아래의 [표 20]과 같다.

표 20

	Scaffold		Whole-genome optical mapping		Super-scaffold (Long reads)		Chromosomes (Assessment using BioNano maps)	
	Size (Mb)	No.	Size (Mb)	No.	Size (Mb)	No.	Size (Mb)	No.
N90	3.09	178	3.86	140	3.53	143	81.54	19
N80	6.45	116	9.45	92	9.26	93	103.05	16
N70	10.45	81	14.47	67	14.53	67	136.43	13
N60	16.16	59	19.56	49	19.36	50	137.59	11
N50	19.85	42	25.93	36	26.08	36	155.88	8
Longest	81.91	-	101.22	-	101.48	-	251.92	-
Gaps	1.65 %	-	1.75 %	-	1.06 %	-	9.44 %	-
Total (≥ 200bp)	2.92 Gb	68,170	2.92 Gb	68,103	2.94 Gb	68,451	3.12 Gb	24
Total (≥ 10 Kb)	2.88 Gb	1,243	2.88 Gb	1,176	2.90 Gb	1,369	3.12 Gb	24

그러면, 도 5를 참조하여 본 발명의 바람직한 실시예에 따른 참조표준 게놈지도의 효과에 대하여 설명한다.

도 5는 본 발명의 바람직한 실시예에 따른 참조표준 게놈 지도의 효과를 설명하기 위한 그래프이다.

도 5의 (a)와 (b)에 도시된 그래프의 X 축은 흑인 5개(Mandeka, Yoruba, San, Mbuti, Dinka), 백인 5개(Sardinian, French, CEU), 비한국 아시아인 5개(Mongolian, Chinese, Japanese), 한국인 5개(Korean)의 전장 게놈 서열 데이터(whole genome re-sequencing data)를 나타낸다. 그리고, 도 5의 (a)에 도시된 그래프의 Y 축은 homozygous SNV의 개수를 나타내고, (b)에 도시된 그래프의 Y 축은 homozygous indel의 개수를 나타낸다.

도 5에 도시된 그래프를 참조하면, 인간게놈지도(GRCh38)와 KOREF_S(본 발명에서, 1명으로 조립한 염색체),

KOREF_C(본 발명에서, 1명으로 조립한 염색체 서열을 40명의 공통 변이로 치환)를 reference 서열로 활용했을 경우 20명의 전장 게놈 서열 데이터(whole genome re-sequencing data)에서 돌연변이 수를 확인할 수 있다. 1명으로 조립된 참조표준 게놈지도 대비(GRCh38 및 KOREF_S), 집단 공통(consensus) 서열을 포함한 참조표준 게놈지도(KOREF_C)가 더 적은 수의 변이를 발견하게 됨을 확인할 수 있다. 이와 같이, 집단 공통 서열을 포함한 참조표준 게놈지도는 개인 특이적으로 보유하고 있는 서열을 제거함으로써, 더 참조표준 게놈지도에 적합한 공통 서열을 보유하고 있음을 확인할 수 있다.

[0120] 그리고, 본 발명에 따른 참조표준 게놈지도('KOREF'라 한다)의 품질을 비교한 결과는 아래의 [표 21]과 같다.

표 21

[0121]

게놈지도	조립서열 길이(bp)	스캐폴드/컨티그 N50 (Mb)	인간게놈지도 (GRCh38) 복원율 (%)	단편중복(segmental duplication) 영역		반복(repeat) 서열		NCBI 유전자 복원	
				길이(bp)	%	길이(bp)	%	수	%
인간게놈지도 GRCh38 (염색체)	3,209,286,105	67.79	-	212,777,868	-	1,564,209,365	-	20,135	-
KOREF (염색체)	3,211,075,818	26.46	88.47	149,353,191	70.19	1,452,404,484	92.85	17,758	88.19
AK1	2,904,207,228	44.85	87.90	144,868,735	68.08	1,454,888,506	93.01	17,759	88.20
CHM1_PacBio_r2	2,996,426,293	26.90	88.02	205,559,250	96.61	1,541,211,387	98.53	17,657	87.69
ASM101398v1	3,176,574,379	26.83	88.26	168,652,649	79.26	1,545,168,387	98.78	6,610	32.83
HsapALLPATHS1	2,786,258,565	12.08	82.89	90,343,965	42.46	1,250,655,296	79.95	16,995	84.41
HuRef (염색체)	2,844,000,504	17.66	85.85	134,317,812	63.13	1,411,487,301	90.24	16,968	84.27
Mongolian	2,881,945,563	7.63	86.54	121,384,034	57.05	1,399,420,366	89.47	17,189	85.37
YH_2.0	2,911,235,363	20.52	86.31	127,254,909	59.81	1,397,013,571	89.31	17,125	85.05
African	2,676,008,911	0.062	69.47	55,830,170	26.24	968,988,149	61.95	9,167	45.53

[0122] 여기서, 조립 서열 길이가 인간게놈지도의 조립 서열 길이와 유사할수록, 스캐폴드/컨티그 N50이 길수록, 인간게놈지도 복원율/단편중복영역 복원율, 반복서열 복원율, NCBI 유전자 복원율이 높을수록 조립 품질이 좋음을 의미한다.

[0125] 그러면, 도 6을 참조하여 본 발명의 바람직한 실시예에 따른 공통서열을 포함한 참조표준 게놈지도 구축 방법에 대하여 설명한다.

[0126] 도 6은 본 발명의 바람직한 실시예에 따른 공통서열을 포함한 참조표준 게놈지도 구축 방법을 설명하기 위한 흐름도이다.

[0127] 도 6을 참조하면, 게놈지도 구축 장치(100)는 단서열과 긴쌍서열을 생산한다(S110). 즉, 게놈지도 구축 장치(100)는 NGS(next generation sequencing)를 이용하여 미리 설정된 값보다 작은 단편 크기(insert size)를 가지는 단서열(short insert[short paired-end])과 미리 설정된 값보다 큰 단편 크기(insert size)를 가지는 긴쌍서열(long insert[long-mate pair])을 생산한다. 여기서, 미리 설정된 값은 1Kb(단편 크기)일 수 있다.

[0128] 그리고, 게놈지도 구축 장치(100)는 단서열과 긴쌍서열을 필터링한다(S120). 즉, 게놈지도 구축 장치(100)는

서열 필터링부(120)는 단서열과 긴쌍서열에서 미리 설정된 리드(read)를 필터링한다. 여기서, 미리 설정된 리드(read)는 중복된 리드(read), 어댑터(adapter) 서열이 포함된 리드(read), 퀄리티 점수(quality score) 값이 미리 설정된 값보다 작은 서열, 접합 어댑터(junction adapter)가 포함된 리드(read) 등을 말한다.

- [0129] 그런 다음, 게놈지도 구축 장치(100)는 단서열을 기반으로 컨티그(contig)를 조립한다(S130). 즉, 게놈지도 구축 장치(100)는 드 부루인(De Bruijn) 그래프를 이용하여 단서열을 기반으로 컨티그(contig)를 조립한다.
- [0130] 그리고, 게놈지도 구축 장치(100)는 컨티그(contig)와 긴쌍서열을 기반으로 스캐폴드(scaffold)를 조립한다(S140).
- [0131] 그런 다음, 게놈지도 구축 장치(100)는 단일 분자 지도(single molecule map)와 스캐폴드(scaffold)를 기반으로 슈퍼-스캐폴드(super-scaffold)를 조립한다(S150). 즉, 게놈지도 구축 장치(100)는 단일 분자 지도(single molecule map)와 상기 스캐폴드(scaffold)를 기반으로 제한 효소(restriction enzyme) 비교를 통해 슈퍼-스캐폴드(super-scaffold)를 조립한다.
- [0132] 그리고, 게놈지도 구축 장치(100)는 슈퍼-스캐폴드(super-scaffold) 상의 갭(gap)을 메꾼다(S160). 즉, 게놈지도 구축 장치(100)는 PacBio 긴서열 해독 방법과 Illumina TSLR 합성 긴서열 해독 방법을 이용하여 상기 슈퍼-스캐폴드(super-scaffold) 상의 갭(gap)을 메꾼다.
- [0133] 또한, 게놈지도 구축 장치(100)는 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증한다(S170). 즉, 게놈지도 구축 장치(100)는 나노채널-기반 게놈 맵핑 데이터(nanochannel-based genome mapping data)를 이용하여 슈퍼-스캐폴드(super-scaffold)의 오조립 영역을 검증할 수 있다.
- [0134] 그런 다음, 게놈지도 구축 장치(100)는 염색체(chromosome) 서열을 조립한다(S180). 즉, 게놈지도 구축 장치(100)는 인간게놈지도(GRCh38)에 정렬(aligned)된 슈퍼-스캐폴드(super-scaffold)의 위치와 방향(strand) 정보를 기반으로 염색체(chromosome) 서열을 조립할 수 있다.
- [0135] 이후, 게놈지도 구축 장치(100)는 염색체(chromosome) 서열을 치환할 수 있다(S190). 즉, 게놈지도 구축 장치(100)는 전장 게놈 서열 데이터(whole genome re-sequencing data)를 이용하여 염색체(chromosome) 서열을 치환할 수 있다.
- [0137] 그러면, 도 7을 참조하여 본 발명의 바람직한 실시예에 따른 서열 필터링 단계에 대하여 보다 자세하게 설명한다.
- [0138] 도 7은 도 6에 도시한 서열 필터링 단계를 보다 자세히 나타낸 흐름도이다.
- [0139] 도 7을 참조하면, 게놈지도 구축 장치(100)는 단서열과 긴쌍서열에서 중복된 리드(read)를 필터링할 수 있다(S121).
- [0140] 그리고, 게놈지도 구축 장치(100)는 단서열과 긴쌍서열에서 어댑터(adapter) 서열이 포함된 리드(read)를 필터링할 수 있다(S123).
- [0141] 또한, 게놈지도 구축 장치(100)는 퀄리티 점수(quality score) 값을 기반으로 단서열과 긴쌍서열을 필터링할 수 있다(S125). 즉, 게놈지도 구축 장치(100)는 단서열과 긴쌍서열에서 퀄리티 점수(quality score) 값이 미리 설정된 값보다 작은 서열을 필터링할 수 있다.
- [0142] 그리고, 게놈지도 구축 장치(100)는 긴쌍서열에서 접합 어댑터(junction adapter)가 포함된 리드(read)를 필터링할 수 있다(S127).
- [0143] 또한, 게놈지도 구축 장치(100)는 단서열과 상기 긴쌍서열이 미리 설정된 길이를 가지도록 자를(trimming) 수 있다(S129).
- [0145] 그러면, 도 8을 참조하여 본 발명의 바람직한 실시예에 따른 컨티그 조립 단계에 대하여 보다 자세하게 설명한다.
- [0146] 도 8은 도 6에 도시한 컨티그 조립 단계를 보다 자세히 나타낸 흐름도이다.
- [0147] 도 8을 참조하면, 게놈지도 구축 장치(100)는 단서열을 기반으로 K-mer 분석을 수행하여 K-mer 빈도 테이블

(frequency table)을 획득할 수 있다(S131).

- [0148] 그리고, 게놈지도 구축 장치(100)는 K-mer 빈도 테이블(frequency table)을 이용하여 상기 단서열의 에러(error)를 보정할 수 있다(S133).
- [0149] 또한, 게놈지도 구축 장치(100)는 K-mer 빈도 테이블(frequency table)과 드 부루인(De Bruijn) 그래프를 이용하여 에러 보정(error correction)된 단서열을 기반으로 컨티그(contig)를 조립할 수 있다(S135).
- [0151] 그러면, 도 9를 참조하여 본 발명의 바람직한 실시예에 따른 스캐폴드 조립 단계에 대하여 보다 자세하게 설명한다.
- [0152] 도 9는 도 6에 도시한 스캐폴드 조립 단계를 보다 자세히 나타낸 흐름도이다.
- [0153] 도 9를 참조하면, 게놈지도 구축 장치(100)는 컨티그(contig)에 긴삽서열을 정렬(alignment)하여 실제 단편 크기(insert size)를 측정할 수 있다(S141).
- [0154] 그리고, 게놈지도 구축 장치(100)는 실제 단편 크기(insert size)를 이용하여 컨티그(contig)와 긴삽서열을 기반으로 스캐폴드(scaffold)를 조립할 수 있다(S143).
- [0155] 또한, 게놈지도 구축 장치(100)는 스캐폴드(scaffold) 상의 갭(gap)을 메꿀 수 있다(S145).
- [0157] 그러면, 도 10을 참조하여 본 발명의 바람직한 실시예에 따른 슈퍼-스캐폴드 조립 단계에 대하여 보다 자세하게 설명한다.
- [0158] 도 10은 도 6에 도시한 슈퍼-스캐폴드 조립 단계를 보다 자세히 나타낸 흐름도이다.
- [0159] 도 10을 참조하면, 게놈지도 구축 장치(100)는 제한 효소(restriction enzyme)를 선택할 수 있다(S151).
- [0160] 그리고, 게놈지도 구축 장치(100)는 옵티컬 맵핑(Optical mapping)에 의해 생성된 단일 분자 지도(single molecule map)와 스캐폴드(scaffold)를 기반으로 선택된 제한 효소(restriction enzyme) 패턴의 비교를 통해 슈퍼-스캐폴드(super-scaffold)를 조립할 수 있다(S153).

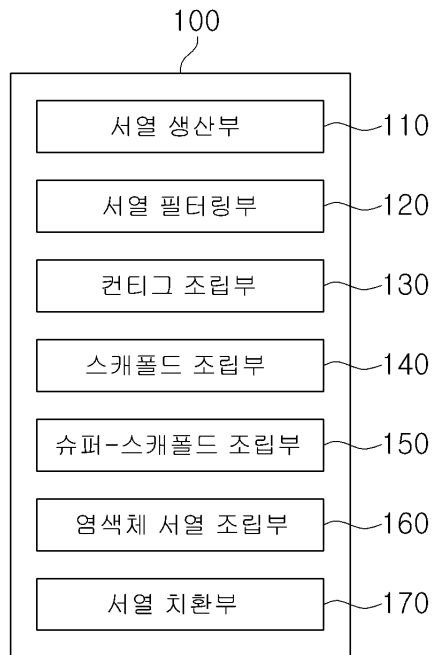
- [0163] 본 발명은 또한 컴퓨터로 읽을 수 있는 기록 매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터로 읽을 수 있는 기록 매체는 컴퓨터에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터가 읽을 수 있는 기록 매체의 예로는 롬(ROM), 램(RAM), 씨디-롬(CD-ROM), 자기 테이프, 플로피 디스크, 광 데이터 저장장치 등이 있으며, 또한 캐리어 웨이브(인터넷을 통한 전송)의 형태로 구현되는 것도 포함한다. 또한, 컴퓨터가 읽을 수 있는 기록 매체는 유무선 통신망으로 연결된 컴퓨터 장치에 분산되어 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.
- [0164] 이상에서 본 발명의 바람직한 실시예에 대하여 상세하게 설명하였지만 본 발명은 상술한 특정의 바람직한 실시예에 한정되지 아니하며, 다음의 청구범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 기술분야에서 통상의 지식을 가진 자라면 누구든지 다양한 변형 실시가 가능한 것은 물론이고, 그와 같은 변경은 청구범위 기재의 범위 내에 있게 된다.

부호의 설명

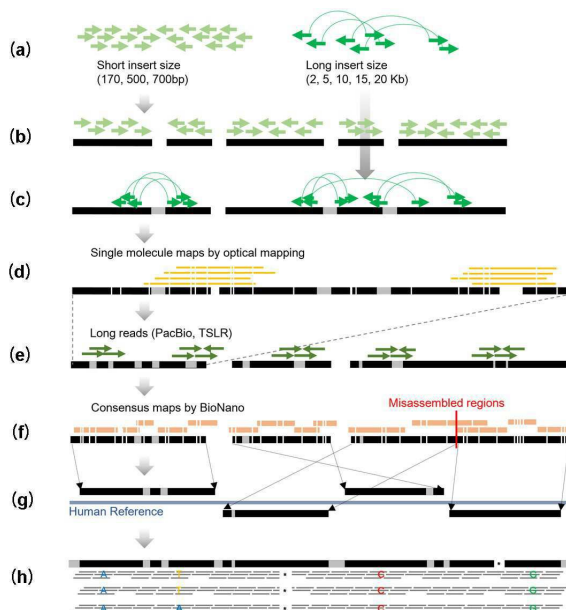
- [0165] 100 : 참조표준 게놈지도 구축 장치, 110 : 서열 생산부,
120 : 서열 필터링부, 130 : 컨티그 조립부,
140 : 스캐폴드 조립부, 150 : 슈퍼-스캐폴드 조립부,
160 : 염색체 서열 조립부, 170 : 서열 치환부

도면

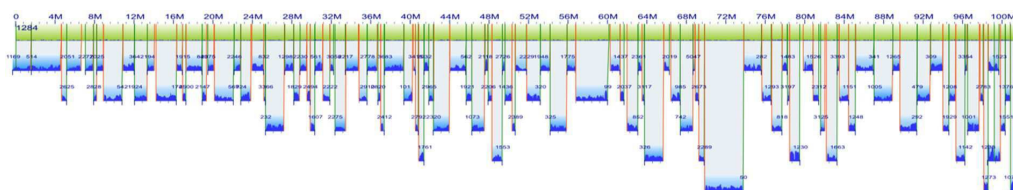
도면1



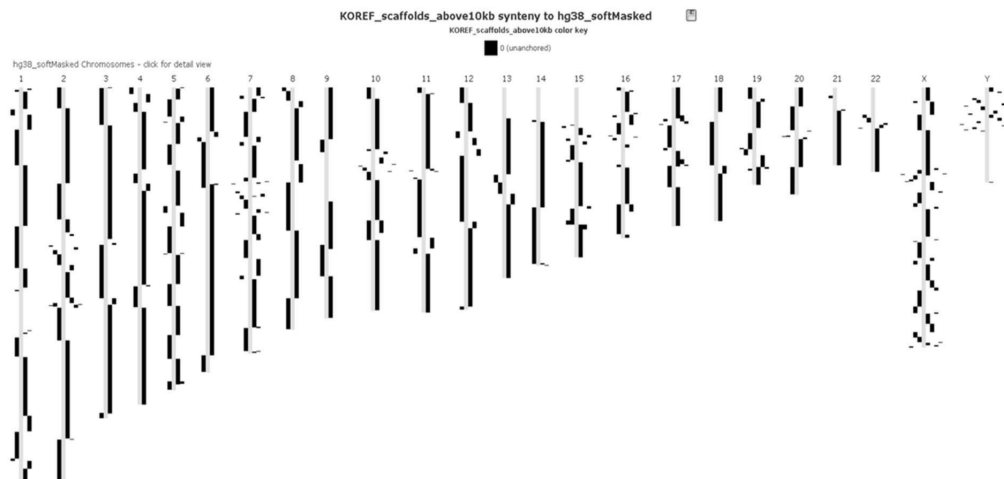
도면2



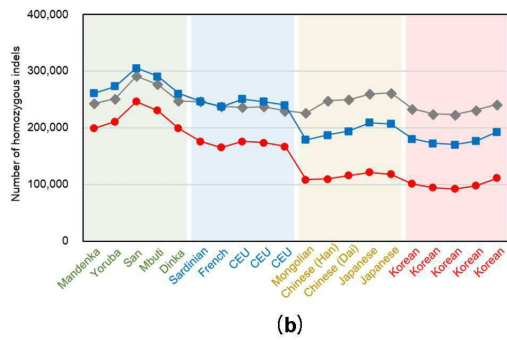
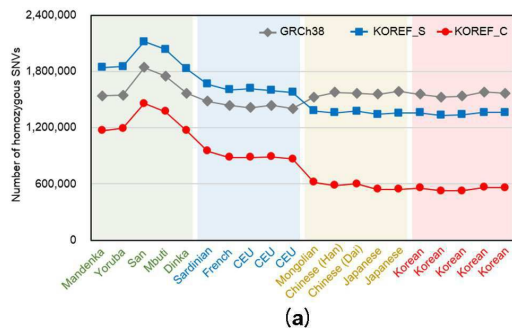
도면3



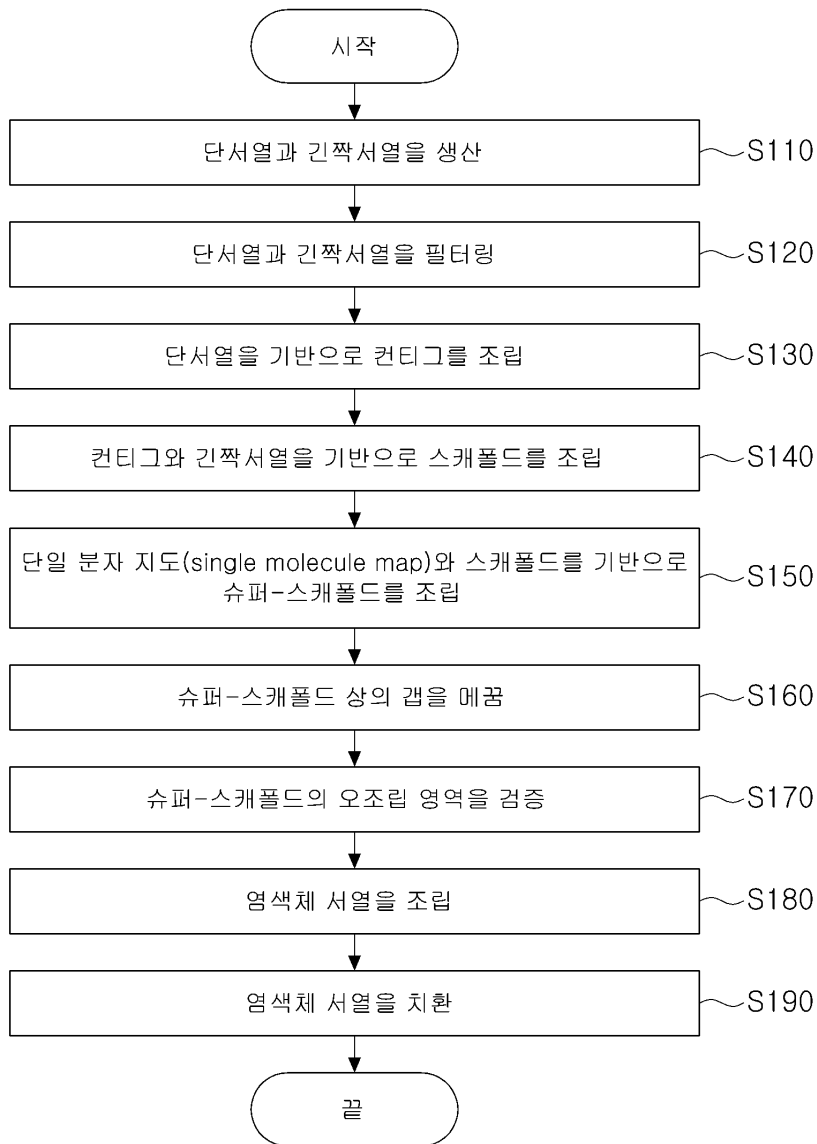
도면4



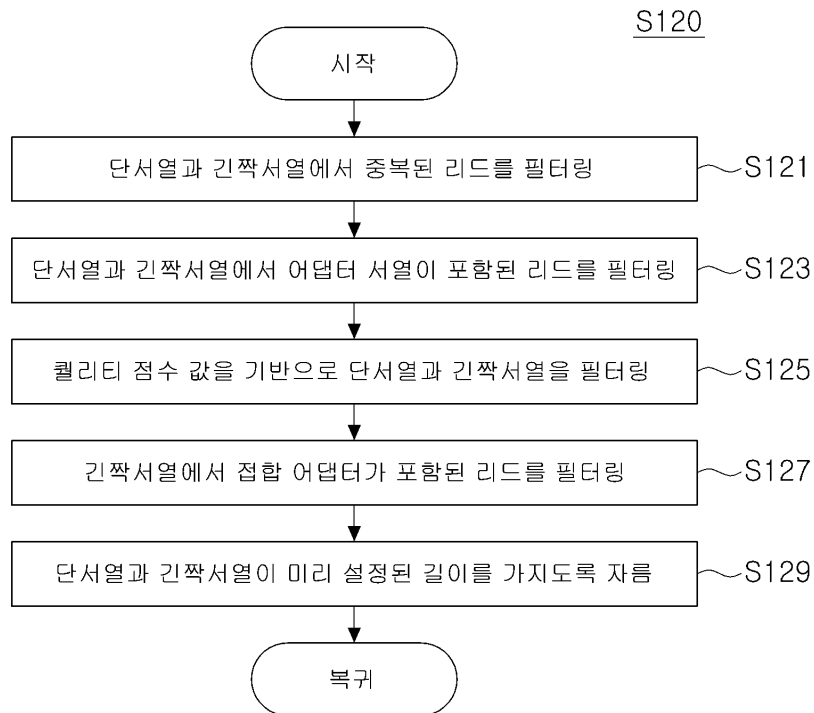
도면5



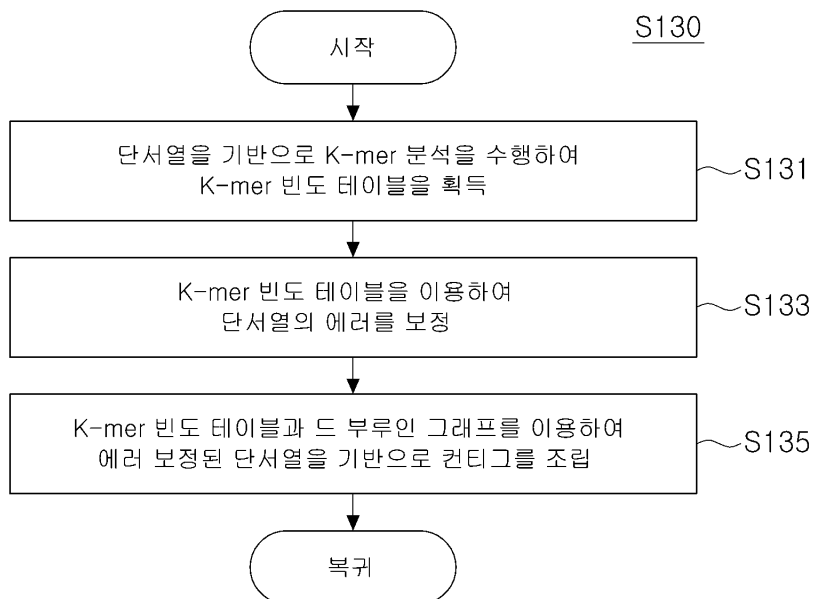
도면6



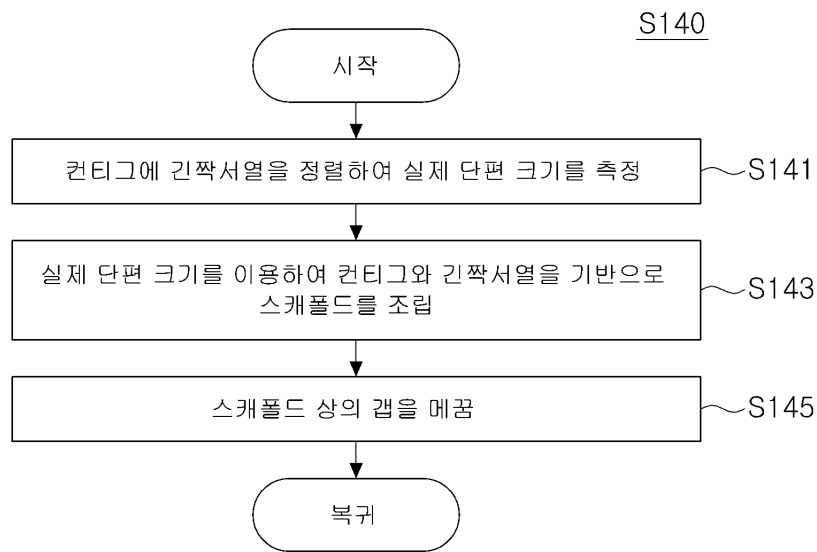
도면7



도면8



도면9



도면10

