



(19) 대한민국특허청(KR)  
(12) 공개특허공보(A)

(11) 공개번호 10-2017-0053147  
(43) 공개일자 2017년05월15일

- (51) 국제특허분류(Int. Cl.)  
G06F 19/22 (2017.01) G06F 19/28 (2011.01)  
G06F 9/38 (2006.01)  
(52) CPC특허분류  
G06F 19/22 (2013.01)  
G06F 19/28 (2013.01)  
(21) 출원번호 10-2016-0147444  
(22) 출원일자 2016년11월07일  
심사청구일자 없음  
(30) 우선권주장  
1020150155211 2015년11월05일 대한민국(KR)

- (71) 출원인  
고려대학교 산학협력단  
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)  
(72) 발명자  
최인걸  
서울특별시 동작구 남부순환로263길 34, 501호(사당동, 태영아트빌)  
민병남  
서울특별시 노원구 중계로 230, 501동 112호(중계동, 중계5단지주공아파트)  
(74) 대리인  
특허법인다나

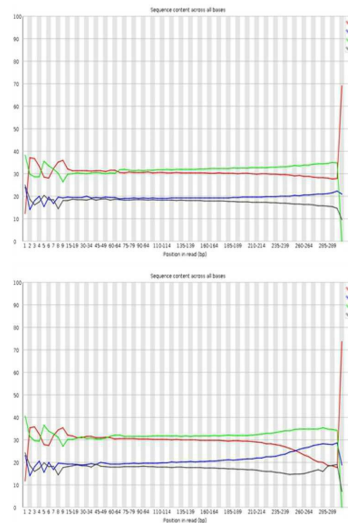
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **균류 유전체 서열 조립 및 평가를 위한 시스템 및 방법**

(57) 요약

본 발명은 차세대 염기서열분석방법으로 얻어진 짧은 DNA서열(sequence reads)로부터 유전체 서열 조립(genome sequence assembly)을 수행하는 시스템에 관한 것이다.

대표도 - 도27



(52) CPC특허분류

**G06F 9/3867** (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 PJ01044003

부처명 미래창조과학부

연구관리전문기관 고려대학교 산학협력단

연구사업명 포스트 게놈 다부처 유전체 사업

연구과제명 버섯류 유전체 해독에 특화된 유전체 조립 스캐폴딩 기술과 해독 파이프라인 개발

기 여 율 1/1

주관기관 농촌진흥청

연구기간 2014.04.01 ~ 2017.12.31

---

## 명세서

### 청구범위

#### 청구항 1

수집된 염기서열 조각을 페이드-엔드(paired-end) 시퀀싱하여 하나의 조각 당 두 개의 리드(read) 서열을 수집 및 저장하는 수집부;

하나 이상의 유전체 조립 소프트웨어를 사용하여 하나 이상의 후보 유전체를 조립하는 유전체 조립부; 및

후보 유전체를 평가하여 최적 유전체를 선별하는 유전체 조립 평가부를 포함하며,

상기 유전체 조립부는 ALLPATHS 파이프라인부 및 Non-ALLPATHS 파이프라인부를 포함하는 군류 유전체 서열 조립 및 평가 시스템.

#### 청구항 2

제 1 항에 있어서,

수집부에서 수집된 리드 서열을 전처리하는 리드 전처리부를 추가로 포함하는 군류 유전체 서열 조립 및 평가 시스템.

#### 청구항 3

제 1 항에 있어서,

유전체 조립 평가부에서 선별된 최적 유전체를 후처리하는 유전체 조립 후처리부를 추가로 포함하는 군류 유전체 서열 조립 및 평가 시스템.

#### 청구항 4

(A) 수집된 염기서열 조각을 페이드-엔드(paired-end) 시퀀싱하여 하나의 조각 당 두 개의 리드(read) 서열을 수집하는 단계;

(B) 하나 이상의 유전체 조립 소프트웨어를 사용하여 하나 이상의 후보 유전체를 조립하는 단계; 및

(C) 후보 유전체를 평가하여 최적 유전체를 선별하는 단계를 포함하며,

상기 하나의 조각에서 수집된 두 개의 리드 서열의 합이 조각의 서열보다 길 경우 ALLPATHS 파이프라인의 소프트웨어를 사용하고,

상기 하나의 조각에서 수집된 두 개의 리드 서열의 합이 조각의 서열보다 짧을 경우 Non-ALLPATHS 파이프라인의 소프트웨어를 사용하는 군류 유전체 서열 조립 및 평가 방법.

#### 청구항 5

제 4 항에 있어서,

수집된 리드 서열을 전처리하는 단계를 추가로 포함하는 군류 유전체 서열 조립 및 평가 방법.

#### 청구항 6

제 5 항에 있어서,

전처리 단계는 리드 품질요약, 어댑터 서열제거, 리드 품질보정 및 k-mer 빈도 계산으로 이루어진 그룹 중에서 하나 이상을 수행하는 군류 유전체 서열 조립 및 평가 방법.

#### 청구항 7

제 4 항에 있어서,

Non-ALLPATHS 파이프라인의 소프트웨어는 ABySS, SOAPdenovo, Velvet 또는 Celera 소프트웨어인 군류 유전체 서열 조립 및 평가 방법.

#### 청구항 8

제 4 항에 있어서,

후보 유전체의 평가는 우도 평가, N50값 평가, 스캐폴드 개수 평가 또는 최대 스캐폴드 길이 평가로 이루어진 그룹 중에서 하나 이상의 평가를 수행하고,

각각의 평가 별로 순위를 부여하여 순위의 합이 가장 적은 유전체를 최적 유전체로 선별하는 군류 유전체 서열 조립 및 평가 방법.

#### 청구항 9

제 4 항에 있어서,

최적 유전체를 후처리하는 단계를 추가로 포함하는 군류 유전체 서열 조립 및 평가 방법.

#### 청구항 10

제 9 항에 있어서,

최적 유전체의 후처리는 외래 DNA에 의한 오염 확인, 28S rDNA 서열 확인, 미토콘드리아 서열 조립 및 웹서버 업로드로 이루어진 그룹 중에서 하나 이상을 수행하는 군류 유전체 서열 조립 및 평가 방법.

#### 청구항 11

제 10 항에 있어서,

외래 DNA에 의한 오염 확인은 GC 함량과 시퀀스 커버리지(sequence coverage)의 분산형 그래프를 이용하여 수행하는 군류 유전체 서열 조립 및 평가 방법.

#### 청구항 12

제 10 항에 있어서,

28S rDNA 서열 확인은 RDP 데이터베이스에 서열 정렬을 하여 분석대상 군류와 유전체 서열이 일치하는지 확인하는 군류 유전체 서열 조립 및 평가 방법.

### 발명의 설명

## 기술 분야

- [0001] 본 발명은 차세대 염기서열분석방법으로 얻어진 짧은 DNA서열(sequence reads)로부터 유전체 서열 조립(genome sequence assembly)을 수행하는 시스템에 관한 것이다.
- [0002] 구체적으로는 고품위균류 유전체에 특화된 시스템으로 일루미나 방식의 DNA 시퀀서에서 생성된 염기서열 리드에 대한 전처리, 유전체 조립, 평가, 후처리를 일련의 과정으로 자동수행하는 유전체 조립 파이프라인 소프트웨어 시스템에 관한 것이다.

## 배경 기술

- [0004] 현대 생명과학에서 DNA 염기서열 분석은 생명과학연구에 유전자와 유전체의 서열정보를 획득하는 필수적인 기술이다. 생명과학 연구자들이 염기서열 데이터를 손쉽게 이용할 수 있게 된 것은 염기서열분석의 가격이 혁신적으로 감소했기 때문이다. 차세대 염기서열분석방법과 같은 서열분석 기술의 비약적인 발전에 의해 서열분석 비용은 현저하게 감소하였다. 2001년도 인간 유전체(약 3 Gbp)크기의 서열분석을 이전 기술인 생어서열분석방법을 사용하면 1억 달러 이상의 비용이 들었던 반면, 2015년에는 약 1,300 달러로 가능해졌다. 15년전에 비해 약 1/77,000의 비용으로 동일한 양의 데이터를 생산할 수 있게 된 것이다.
- [0005] 또한 염기서열을 응용하는 분야도 다양하게 발달하였는데, 환경시료로부터 미생물군집을 분석하는 메타유전체(metagenomics)분석, 유전체 사이의 구조적 변이체 분석, 유전자 발현, 전사체의 변이체 분석 등이 대표적인 예이다.
- [0006] 하지만 DNA 시퀀서에 의해 생산되는 리드(read)의 길이는 유전체의 염기서열 길이에 비해 상당히 짧기 때문에, 한 번의 시퀀싱으로 전체 유전체 서열을 얻는 것은 현재 기술로는 불가능하다. 현재 가장 긴 리드를 생산할 수 있는 팍바이오(PacBio) 플랫폼에서 생산되는 리드의 길이는 최대 20 Kbp인데, 이는 단순한 균류 중 하나인 효모의 유전체 크기(12.07 Mbp)와 비교하면 약 1/600에 해당하는 길이이다. 가장 널리 사용되는 DNA 시퀀싱 플랫폼 중 하나인 일루미나(Illumina) 플랫폼의 경우 한방향으로 최대 300 bp 길이의 리드만을 생산할 수 있다.
- [0007] 유전체 조립이란 DNA 시퀀서로부터 생산된 짧은 리드를 조립하여 원래의 긴 유전체 염기서열로 복원하는 일련의 계산 과정이다.
- [0008] 유전체를 조립하는 알고리즘은 서열이 상당부분 유사한 두 리드는 유전체 내 동일한 위치에서 유래한다는 가정을 기초로 설계되었으며 크게 세 가지로 분류된다.
- [0009] Greedy 알고리즘은 최대로 겹치는 두 리드를 연결함으로써 즉각적인 연결을 취하는 방식이다. 이 알고리즘은 전체 유전체를 한 번에 다룬다기 보다는 부분적인 리드들의 연결을 우선시 하기 때문에, mate-pair read의 연결 등 전체 유전체 정보를 이용할 때에는 어려움이 있다. phrap, TIGR Assembler, VCAKE가 Greedy 알고리즘을 이용하는 대표적인 오픈소스 소프트웨어이다.
- [0010] Overlap-layout-consensus 알고리즘은 먼저 모든 리드 사이에서 충분히 포개어지는 리드의 쌍을 찾아내고, 이를 토대로 그래프를 그린다. 그래프의 노드는 각 리드가 되며 엣지는 서열이 유사한 두 리드의 쌍을 연결한다. 이로써 Greedy 알고리즘과 달리 전체 유전체를 한 번에 다룰 수 있게 된다. Celera Assembler가 이 패러다임을 이용하는 대표적인 오픈소스 소프트웨어이다.
- [0011] De Bruijn graph 알고리즘은 overlap-layout-consensus처럼 그래프를 그리지만 모든 리드들을 k-mer로 분산시킨다는 점에서 그 차이가 있다. De Bruijn graph 알고리즘에서의 엣지는 두 k-mer 사이에 (k-1)의 서열이 정확히 일치하는 두 노드들을 연결시키기 때문에, 시퀀싱 에러를 수정하는 것이 최종 어셈블리의 품질을 결정하는 주요 요인이 된다. Velvet, SOAPdenovo, ALLPATHS와 같은 오픈 소프트웨어에서 이 방식을 이용하고 있다.
- [0012] 이와 같이 유전체를 조립하는 데에는 다양한 알고리즘을 적용할 수 있는데, 어떤 알고리즘이 최적의 어셈블리를 조립하는지에 대해서는 Assemblathon와 Gage 대회를 통해 간접적으로 가늠해볼 수 있다. Assemblathon 대회는 새, 어류, 뱀 유전체에 대하여 어떤 유전체 조립 알고리즘이 원래의 유전체 서열에 제일 가까운 어셈블리를 조립하는지를 평가하는 대회이며, 총 21개 팀이 참여하였다. 두 대회 모두 ALLPATHS가 수상하여 유전체 조립에서의 우수한 알고리즘임을 검증하였다.
- [0013] 하지만 반복 서열의 분포 등 유전체의 구조적 특성이나 시퀀스 라이브러리의 종류에 따라 최적 어셈블리를 조립

하는 알고리즘이 달라질 수 있다. 이와 같은 문제는 사용 가능한 소프트웨어를 이용해 유전체를 각각 조립한 뒤, 각 어셈블리에 대하여 평가를 하고 최적 어셈블리를 선별함으로써 해결할 수 있다. 하지만 이런 단계들은 컴퓨팅 환경에 익숙하지 않은 연구자들에게 상당한 학습 시간을 요구한다.

- [0014] 이에 따라 주어진 유전체의 리드 서열에 대하여 여러 가지의 알고리즘을 동시에 적용하고 이를 평가하여 최적의 어셈블리를 생산하는 소프트웨어적 파이프라인이 요구된다.
- [0015] 균류(Fungi)는 세균(Bacteria)에 비해 크고 복잡한 유전체 구조를 지니고 있으며, 2차 대사산물 생산을 위한 산업 균주로서의 활용 및 환경복원에의 다양한 응용 가능성때문에 이들 균주에 대한 유전체 분석이 활발히 진행되고 있다.
- [0016] 미국 국립생물정보센터(NCBI)에는 약 700개의 균류 유전체가 공개되어 있으며, 조인트 게놈 인스티튜트에서 제공하는 마이코코즘(<http://genome.jgi.doe.gov/programs/fungi/index.jsf>)에는 약 500여개의 균류 유전체가 공개되어 있다. 균류 유전체의 분석 및 데이터 공개의 속도는 해마다 증가하는 추세이다.
- [0017] 균류 유전체의 크기는 2.1 Mbp(Microsporidia에 속하는 기생균류)부터 215 Mbp (Uromyces viciae-fabae, 담자균류)까지 다양하며 평균 35 Mbp의 크기를 가지고 있다. 이는 대장균(4.6 Mbp)과 같은 세균류의 유전체 크기보다는 크며 인간(3 Gbp)과 같은 영장류의 유전체의 크기보다는 훨씬 작은 크기이다.
- [0019] 공개된 어셈블리 파이프라인은 하기와 같다.
- [0020] 1) Orione: a web-based framework for NGS analysis in microbiology
- [0021] 미생물 수준 유전체에 대한 생명정보분석을 위한 종합 도구로서 서열 재분석(resequencing), 유전체 조립, 스캐폴딩, RNA-seq, 유전자 주석달기, 메타유전체분석, 메타전사체분석을 위한 도구들을 사용자가 웹을 통해 직접 파이프라인을 구축하고 실행할 수 있도록 한다. Orione에서 제공하는 De novo assembly 소프트웨어는 Velvet, ABySS, SPAdes, SSAKE, Edena가 있다. 여러 바이오인포매틱스 도구들의 사용을 위한 플랫폼을 제공하는 것을 주 기능으로 하며 자체적인 기능은 제공하지 않고 있다.
- [0023] 2) A5, A5-miseq: An Integrated Pipeline for de Novo Assembly of Microbial Genomes
- [0024] 미생물 수준 유전체에 대하여 Illumina 시퀀스를 받아 유전체 조립을 수행한다. 총 다섯 단계 (1) 리드 전처리, (2) 컨티그 조립, (3) 1차 스캐폴딩, (4) 어셈블리 수정, (5) 최종 스캐폴딩으로 구성되어 있으며 어셈블리를 위한 사용소프트웨어로 IDBA-UD 한 종류만을 사용한다(도 1).
- [0026] 3) CG-Pipeline: A computational genomics pipeline for prokaryotic sequencing projects
- [0027] 원핵생물 유전체에 대하여 유전체 조립, 유전자 기능 유추, 주석달기 등의 작업을 수행하는 자동으로 수행하는 파이프라인이다. 또 다른 차세대염기서열분석 플랫폼인 454 리드에 특화되어 있다. 사용하는 유전체 조립 소프트웨어는 Newbler와 AMOScp이며 Minimus를 이용하여 두 소프트웨어에서 조립한 어셈블리를 합친다(도 2).
- [0029] 4) DSP (denovo\_solid\_pipeline): a semi automated pipeline for short genome assembly using SOLiD sequencing data
- [0030] SOLiD 플랫폼에 특화되어 있는 유전체 조립 파이프라인으로서 동적계획법(dynamic programming)을 이용한 리드 수정을 최적화하는 알고리즘을 주요 기능으로 지니고 있다.
- [0032] 5) iMetAMOS: Automated ensemble assembly and validation of microbial genome
- [0033] iMetAMOS는 세균 유전체에 대하여 여러 오픈소스 소프트웨어를 실행하여 최적 어셈블리를 선별하는 파이프라인이다. 주요 기능은 파라미터를 자동으로 선택하는 것과 각 어셈블리의 품질을 여러 수치를 이용해 점수화하는 것이다. ABySS, CABOG, IDBA-UD, MaMaSuRCA, MetaVelvet, MIRA, Ray/RayMeta, SGA, SOAPdenovo2, SPAdes,

SparseAssembler, Velvet, Velvet-SC를 이용한다(도 3).

[0035] 6) PBcR pipeline: Hybrid error correction and de novo assembly of single-molecule sequencing reads

[0036] 팍바이오(PacBio) 플랫폼에서 생산된 리드의 시퀀싱 에러 수정 알고리즘 및 de novo 어셈블리 기능을 제공한다. 파지, 원핵생물, 진핵생물의 유전체 조립에 대해 테스트되었다(도 4).

[0038] 7) RAMPART: a workflow management system for de novo genome assembly.

[0039] 사용자 수준에서 제공되는 오픈소스 어셈블리 소프트웨어를 이용해 직접 workflow를 설정할 수 있는 시스템이다. 사용자가 직접 각 소프트웨어에 필요한 파라미터를 조절할 수 있고, 다른 시스템과는 달리 큰 진핵생물의 유전체를 다룰 수 있도록 컴퓨터 클러스터나 메모리 공유 시스템을 이용할 수 있도록 한다(도 5).

[0041] 8) VirAmp: a galaxy-based viral genome assembly pipeline.

[0042] 바이러스 유전체 조립에 특화된 파이프라인으로 유전체 조립, 분석, 해석을 순차적으로 수행한다(도 6).

## 발명의 내용

### 해결하려는 과제

[0044] 본 발명은 DNA 시퀀서로부터 생산된 균류 유전체의 리드로부터 여러 오픈소스 조립 소프트웨어를 동시에 이용하여 후보군 어셈블리를 생성하고, 각 어셈블리를 평가 및 비교하여 최종 어셈블리는 선별하는 과정을 자동으로 수행하는 소프트웨어적 파이프라인을 구축하는 것을 목적으로 한다.

### 과제의 해결 수단

[0046] 본 발명에서는 수집된 염기서열 조각을 페이드-엔드(paired-end) 시퀀싱하여 하나의 조각 당 두 개의 리드(read) 서열을 수집 및 저장하는 수집부;

[0047] 하나 이상의 유전체 조립 소프트웨어를 사용하여 하나 이상의 후보 유전체를 조립하는 유전체 조립부; 및

[0048] 후보 유전체를 평가하여 최적 유전체를 선별하는 유전체 조립 평가부를 포함하며,

[0049] 상기 유전체 조립부는 ALLPATHS 파이프라인부 및 Non-ALLPATHS 파이프라인부를 포함하는 균류 유전체 서열 조립 및 평가 시스템을 제공한다.

[0050] 또한, 본 발명에서는 (a) 수집된 염기서열 조각을 페이드-엔드(paired-end) 시퀀싱하여 하나의 조각 당 두 개의 리드(read) 서열을 수집하는 단계;

[0051] (b) 하나 이상의 유전체 조립 소프트웨어를 사용하여 하나 이상의 후보 유전체를 조립하는 단계; 및

[0052] (c) 후보 유전체를 평가하여 최적 유전체를 선별하는 단계를 포함하며,

[0053] 상기 하나의 조각에서 수집된 두 개의 리드 서열의 합이 조각의 서열보다 길 경우 ALLPATHS 파이프라인의 소프트웨어를 사용하고,

[0054] 상기 하나의 수집된 두 개의 리드 서열의 합이 조각의 서열보다 짧을 경우 Non-ALLPATHS 파이프라인의 소프트웨어를 사용하는 균류 유전체 서열 조립 및 평가 방법을 제공한다.

### 발명의 효과

[0056] 본 발명의 균류 유전체 서열 조립 및 평가 방법에 따르면 다음과 같은 효과를 가진다.

- [0057] 첫째, De novo 유전체 조립에 대하여 어떤 알고리즘이 적합한지에 대한 사전 결정 없이, 여러 알고리즘을 한번의 파이프라인 실행으로 적용하고 체계적인 결과물을 얻을 수 있다.
- [0058] 둘째, 서열정보를 컴퓨팅을 이용해 다루는 방법이 익숙하지 않은 연구자들이 일정 수준이상 품질의 어셈블리를 조립할 수 있게 할 수 있다.
- [0059] 셋째, 다양한 어셈블리 평가 기준을 계산하고 시각화하여 보여줄 뿐만 아니라, 최적 어셈블리를 선별하는 점수화 알고리즘을 제공하여 전문적인 배경지식 없이도 손쉽게 최적 어셈블리를 구할 수 있다.
- [0060] 넷째, 파이프라인의 사용으로 각 소프트웨어를 독립적으로 시행할 때보다 인력을 최소화할 수 있다.
- [0061] 다섯째, 최적 어셈블리의 후처리 과정을 통해 외래 서열의 확인, 미토콘드리아 유전체 조립, 여러 서열 데이터베이스에 정렬 등 사용 가능한 어셈블리 수준에서의 다양한 정보를 한 번의 파이프라인 실행으로 얻을 수 있다.

### 도면의 간단한 설명

- [0063] 도 1은 A5 파이프라인 모식도이다.
- 도 2는 CG-Pipeline 모식도이다.
- 도 3은 iMetAMOS 모식도이다.
- 도 4는 PBCR pipeline 모식도이다.
- 도 5는 RAMPART 모식도이다.
- 도 6은 VirAmp 모식도이다.
- 도 7은 본 발명에 따른 군류 유전체 서열 조립 및 평가 방법의 각 단계를 나타내는 모식도이다.
- 도 8은 본 발명에 따른 각 단계를 실행하기 위해 필요한 오픈소스 소프트웨어의 목록을 나타낸다.
- 도 9는 본 발명에 따른 리드 전처리 단계를 나타내는 모식도이다.
- 도 10은 본 발명에 따른 유전체 조립 단계를 나타내는 모식도이다.
- 도 11은 본 발명에 따른 유전체 조립 평가 단계를 나타내는 모식도이다.
- 도 12는 본 발명에 따른 유전체조립 후처리 단계를 나타내는 모식도이다.
- 도 13은 페어드-엔드(paired-end) 시퀀싱을 나타내는 모식도이다.
- 도 14는 페어드-엔드 라이브러리를 나타내는 모식도이다.
- 도 15는 FastQC에 의해 생성된 퀄리티 분포 그래프이다.
- 도 16은 타일당 퀄리티 히트맵이다.
- 도 17은 리드당 퀄리티 평균 그래프이다.
- 도 18은 염기서열 위치당 A, T, G, C의 분포 그래프이다.
- 도 19는 리드당 GC 함량 그래프 및 리드의 GC 함량에 대한 빈도 그래프이다.
- 도 20은 염기서열 위치당 'N' 함량 그래프이다.
- 도 21은 리드 길이 분포 그래프이다.
- 도 22는 리드 중복 그래프이다.
- 도 23은 어댑터 함량 그래프이다.
- 도 24는 K-mer 함량 그래프이다.
- 도 25는 C. cucurbitarum 유전체 리드의 염기당 퀄리티 분포를 나타낸다.



도 26은 C. cucurbitarum 유전체 리드의 리드당 평균 켈리티 분포를 나타낸다.

도 27은 C. cucurbitarum 유전체 리드의 염기서열 위치당 ATGC 함량 그래프이다.

도 28은 C. cucurbitarum 유전체 리드 품질보정 그래프이다.

도 29는 K-mer histogram 그래프이다.

도 30은 C. cucurbitarum 최적 어셈블리 컨티그 누적길이 및 GC함량 그래프이다.

도 31은 C. cucurbitarum 어셈블리의 외래 DNA 오염 확인. GC 함량과 각 스캐폴드의 coverage를 나타낸 분산형 그래프이다.

도 32는 C. cucurbitarum 어셈블리의 28S rDNA 서열이다.

### 발명을 실시하기 위한 구체적인 내용

[0064] 이하, 본 발명을 구체적으로 설명한다.

[0066] 본 발명은 균류 유전체 서열 조립 및 평가 시스템을 제공한다.

[0067] 상기 시스템은 수집된 염기서열 조각을 페이드-엔드(paired-end) 시퀀싱하여 하나의 조각 당 두 개의 리드(read) 서열을 수집 및 저장하는 수집부;

[0068] 하나 이상의 유전체 조립 소프트웨어를 사용하여 하나 이상의 후보 유전체를 조립하는 유전체 조립부; 및

[0069] 후보 유전체를 평가하여 최적 유전체를 선별하는 유전체 조립 평가부를 포함한다.

[0070] 또한, 본 발명의 시스템은 수집된 리드 서열을 전처리하는 리드 전처리부 및 선정된 최적 유전체를 후처리하는 유전체 조립 후처리부를 추가로 포함할 수 있다.

[0071] 일 구체예에서 상기 시스템은 수집부; 리드 전처리부; 유전체 조립부; 유전체 조립 평가부 및 유전체 조립 후처리부로 구성될 수 있으며, 각 구성에서 사용되는 소프트웨어는 하나의 파이프라인으로 연결될 수 있다.

[0072] 수집부는 수집된 염기서열 조각(인서트)을 페이드-엔드(paired-end) 시퀀싱하여 하나의 조각 당 두 개의 리드(read) 서열을 수집 및 저장한다. 이때, 염기서열은 균류로부터 추출한 것일 수 있다.

[0073] 리드 전처리부는 상기 수집부에서 수집된 리드 서열을 전처리 한다. 상기 전처리부에서는 리드 품질요약, 어댑터 서열제거, 리드 품질보정 및 k-mer 빈도 계산으로 이루어진 그룹 중에서 하나 이상을 수행할 수 있다.

[0074] 유전체 조립부는 하나 이상의 유전체 조립 소프트웨어를 사용하여 하나 이상의 후보 유전체를 조립할 수 있다.

[0075] 상기 유전체 조립부는 ALLPATHS 파이프라인부 및 Non-ALLPATHS 파이프라인부를 포함할 수 있다. 하나의 조각에서 수집된 두 개의 리드 서열의 합이 조각의 서열보다 길 경우 ALLPATHS 파이프라인부에서 조립이 수행되며, 하나의 조각에서 수집된 두 개의 리드 서열의 합이 조각의 서열보다 짧을 경우 Non-ALLPATHS 파이프라인부에서 조립이 수행될 수 있다.

[0076] 유전체 조립 평가부는 상기 유전체 조립부에서 조립된 후보 유전체를 평가하여 최적 유전체를 선별한다. 상기 유전체 조립 평가부에서는 우도 평가, N50값 평가, 스캐폴드 개수 평가 또는 최대 스캐폴드 길이 평가로 이루어진 그룹 중에서 하나 이상을 수행하여 후보 유전체를 평가하며, 각각의 평가 별로 순위를 부여하여 순위의 합이 가장 적은 유전체를 최적 유전체로 선별할 수 있다.

[0077] 유전체 조립 후처리부는 유전체 조립 평가부에서 선별된 최적 유전체를 후처리할 수 있다. 상기 후처리부에서는 외래 DNA에 의한 오염 확인, 28S rDNA 서열 확인, 미토콘드리아 서열 조립 및 웹서버 업로드로 이루어진 그룹 중에서 하나 이상을 수행할 수 있다.

[0079] 또한, 본 발명은 균류 유전체 서열 조립 및 평가 방법에 관한 것이다.

[0080] 상기 방법은 (A) 수집된 염기서열 조각을 페이드-엔드(paired-end) 시퀀싱하여 하나의 조각 당 두 개의 리드(read) 서열을 수집하는 단계;

- [0081] (B) 하나 이상의 유전체 조립 소프트웨어를 사용하여 하나 이상의 후보 유전체를 조립하는 단계; 및
- [0082] (C) 후보 유전체를 평가하여 최적 유전체를 선별하는 단계를 포함한다.
- [0083] 상기 방법에서는 수집된 리드 서열을 전처리하는 단계; 및 최적 유전체를 후처리하는 단계를 추가로 포함할 수 있다.
- [0084] 이하, 첨부된 도면을 참조하여 발명을 실시하기 위한 구체적인 내용을 상세히 설명한다.
- [0085] 일 구체예에서 균류 유전체 서열 조립 및 평가 방법은 도 7에 나타난 바와 같이, (a) 수집된 염기서열 조각을 페이드-엔드(paired-end) 시퀀싱하여 하나의 조각 당 두 개의 리드(read) 서열을 수집하는 단계(이하, DNA 시퀀서 단계, 100);
- [0086] (b) 수집된 리드 서열을 전처리하는 단계(이하, 리드 전처리 단계, 300);
- [0087] (c) 하나 이상의 유전체 조립 소프트웨어를 사용하여 하나 이상의 후보 유전체를 조립하는 단계(이하, 유전체 조립 단계, 400);
- [0088] (d) 조립된 후보 유전체를 평가하여 최적 유전체를 선별하는 단계(이하, 유전체 조립 평가 단계, 500); 및
- [0089] (e) 최적 유전체를 후처리하는 단계(이하, 유전체 조립 후처리 단계, 600)로 구성될 수 있다.
- [0090] 본 발명에서는 도 8에 나타난 바와 같이, 각각의 단계마다 사용되는 소프트웨어를 하나의 파이프라인으로 연결하여, DNA 시퀀서에서 생성된 염기서열 리드의 전처리, 유전체 조립, 평가 및 후처리를 일련의 과정으로 자동수행할 수 있다. 즉, 본 발명에서는 유전체의 리드 서열에 대하여 여러 가지의 알고리즘을 동시에 적용하고 이를 평가하여 최적의 어셈블리를 생산하는 소프트웨어적 파이프라인을 제공할 수 있다. 이에 따라 각 소프트웨어를 독립적으로 시행할 때보다 인력을 최소한으로 할 수 있으며, 다양한 정보를 한 번의 파이프라인 실행으로 얻을 수 있다.
- [0091] 본 발명에서 단계 (a)는 DNA 시퀀서 단계(100)로, 염기서열을 결정하고 수집하여 리드를 생성하는 역할을 수행할 수 있다. 상기 단계는 수집부에서 수행되며, 염기서열은 균류로부터 추출한 것일 수 있다.
- [0092] 단계 (a)는 플랫폼에 따라 서로 다른 형태의 리드를 생성할 수 있으며, 본 발명에서는 일루미나 플랫폼의 리드를 사용할 수 있다.
- [0093] 상기 일루미나 플랫폼은 주어진 서열에 대해 양 쪽으로 최대 300 bp를 읽어 한 서열에 대해 두 개의 리드 서열을 제공할 수 있다. 이러한 작업을 페어드-엔드(paired-end) 시퀀싱이라 한다. 상기 시퀀싱에 의해 읽힌 서열은 인서트(insert)라고 하며 DNA 시퀀서에 의해 양쪽으로 읽혀 실제로 유전체 조립에 사용하는 서열을 리드(read)라고 한다.
- [0094] 본 발명에서 도 13은 페어드-엔드(paired-end) 시퀀싱을 나타내는 모식도이다. 유전체 DNA를 쪼갠 조각을 인서트(insert)라 하며, 일루미나 플랫폼에서는 이 인서트의 양 끝으로 일정 길이의 염기서열을 읽을 수 있다. 이때 읽힌 염기서열을 리드(read)라고 하며, 유전체 분석시 얻게 되는 데이터는 1쌍의 리드 서열과 리드 1과 리드 2가 같은 인서트로부터 생성되었다는 정보(pair information)이다.
- [0095] 본 발명에서 단계 (b)는 리드 전처리 단계(300)로, 상기 단계에서는 전술한 단계 (a)에서 수집된 리드 서열을 전처리할 수 있다. 상기 단계는 리드 전처리부에서 수행될 수 있다.
- [0096] 리드 전처리 단계(300)는 도 9에 나타난 바와 같이, 리드 품질요약 단계(301), 어댑터 서열제거 단계(302), 리드 품질보정 단계(303) 및 k-mer 빈도 계산 단계(304)으로 이루어진 그룹 중에서 하나 이상을 수행할 수 있다. 상기 단계에서는 FastQC(201), Cutadapt(202), HTQC(203) 및 Jellyfish(204)로 이루어진 그룹 중에서 하나 이상의 소프트웨어가 사용되며, 상기 소프트웨어는 파이프라인(이하, 리드 전처리 파이프라인)으로 구축되어 있을 수 있다.
- [0097] 리드 품질요약 단계(301)에서는 FastQC(201) 소프트웨어가 사용된다. 상기 단계에서는 리드의 분포를 분석하고 라이브러리의 다양한 통계치를 확인할 수 있는데, 구체적으로 DNA 시퀀서로부터 생성된 리드들의 특성에 대해 그래프와 표를 통해 시각화함으로써 리드의 퀄리티나 리드의 분포에 대한 개략적인 정보를 얻을 수 있다.
- [0098] 본 발명에서는 FastQC(201) 소프트웨어를 통해 퀄리티 분포 그래프, 타일당 퀄리티 히트맵, 리드당 퀄리티 평균 그래프, 염기서열 위치당 A, T, G, C 분포 그래프, 리드당 GC 함량 그래프, 염기서열 위치당 'N' 함량

그래프, 리드 길이 분포 그래프, 리드 중복 그래프, 어댑터 함량 그래프 및 K-mer 함량 그래프를 확인할 수 있다.

- [0099] 퀄리티 분포 그래프는 리드의 염기서열 위치 당 퀄리티의 분포를 나타내는 그래프로, x축은 리드에서의 염기의 위치, y축은 퀄리티 점수를 나타낸다. 일루미나 플랫폼에서 생성된 초기 리드는 모두 같은 길이를 가지며 각 염기서열은 고유한 퀄리티 점수를 가진다. 시퀀서에서 생산한 퀄리티 점수의 경우 0-40 사이의 값을 가지며, 높은 값을 가질수록 신뢰할 수 있는 염기서열임을 의미한다. 통상적으로 퀄리티 점수 20 미만의 염기서열은 신뢰할 수 없는 염기서열로 분류되어 후에 품질보정 과정에서 제거될 수 있다.
- [0100] 본 발명에서 도 15는 FastQC에 의해 생성된 퀄리티 분포 그래프로, 위 그래프는 좋은 퀄리티의 리드의 예이며, 아래는 퀄리티가 낮은 리드의 예이다.
- [0101] 타일당 퀄리티 히트맵(heatmap) 그래프는 타일에서 생산된 리드들의 퀄리티 평균을 나타내는데, 해당 타일의 퀄리티가 낮을수록 붉은색으로, 높을수록 파란색으로 나타낸다. 일루미나 플랫폼은 타일이 구분되어 있어 타일마다 일정한 양의 리드를 생산하는데, 특정 타일에 문제가 있을 경우 해당 타일에서 생산되는 리드의 퀄리티는 전반적으로 낮을 것이다. 따라서 상기 그래프를 통해 오류가 있는 타일을 확인하여 유전체 조립에서 배제할 수 있다.
- [0102] 본 발명에서 도 16은 타일당 퀄리티 히트맵(heatmap) 그래프로, 위 그래프는 모든 타일에서 높은 퀄리티를 보이고 있으며 아래 그래프는 몇몇 타일에서 낮은 퀄리티를 보이는 것을 확인할 수 있다.
- [0103] 리드 당 퀄리티 평균 그래프는 리드의 모든 염기서열에 해당하는 퀄리티 점수의 평균을 계산하고 그 빈도를 그래프로 나타낸 것으로, 상기 그래프를 통해 전체 리드의 퀄리티를 가늠할 수 있다.
- [0104] 본 발명에서 도 17은 리드당 퀄리티 평균 그래프로, 위 그래프는 거의 모든 리드가 36에서 38사이의 높은 퀄리티를 가지고 있음을 보여주고 있는 반면, 아래 그래프는 많은 리드가 20 미만의 평균 퀄리티를 가지고 있음을 보여주는 것을 확인할 수 있다.
- [0105] 염기서열 위치당 A, T, G, C의 분포 그래프는 염기서열 위치당 A, T, G, C의 빈도를 퍼센트로 나타낸다. DNA 염기서열은 아데닌(A), 구아닌(G), 시토닌(C), 티민(T)의 네 가지 염기로 구성되어 있으며, 읽힌 서열이 유전체에서 고르게 샘플링이 된 것을 가정하면 리드의 염기서열 위치당 A, T, G, C의 빈도는 전 위치에 걸쳐 일정할 것이다. 만일 유전체의 특정 자리에서 편향적으로 시퀀싱이 이루어지거나 어댑터 시퀀스가 대량으로 존재할 경우 네 염기의 빈도가 염기서열 위치마다 일정하지 않은 그래프를 얻어질 수 있다.
- [0106] 본 발명에서 도 18은 염기서열 위치당 A, T, G, C의 분포 그래프를 나타낸다. 유전체 서열에 편향된 샘플링이 일어나지 않고 시퀀싱이 균등하게 이루어졌다면 위 그래프처럼 모든 염기서열이 리드 자리마다 일정한 비율을 차지하며, 그렇게 않은 경우에는 아래 그래프처럼 자리마다 일정하지 않게 나타난다.
- [0107] 리드당 GC content 그래프는 각 리드의 GC 함량을 계산하여 그 빈도를 나타낸다. 모든 생물종은 종마다 특정 GC 함량을 가지는데 이론적인 GC 함량의 분포를 함께 제시하여 해당 유전체의 서열분석이 적절하게 이루어졌는지 판단할 수 있다.
- [0108] 본 발명에서 도 19는 리드당 GC 함량 그래프 및 리드의 GC 함량에 대한 빈도 그래프로, x축은 평균 GC 함량, y축은 빈도를 나타낸다. 상기 도 19에 나타난 바와 같이, 파란 곡선은 이론적인 빈도 곡선을 나타내며 이에서 벗어난 형태를 가질 경우 시퀀싱이 적절하게 이루어지지 않았음을 확인할 수 있다.
- [0109] 염기서열 위치당 'N'의 함량 그래프는 염기서열 위치 당 'N'의 개수를 세어 퍼센트 함량으로 계산한 값을 나타낸다. DNA 시퀀서는 A, T, G, C중 어떤 염기서열로도 결정할 수 없을 경우 'N'을 생성하는데, 'N'의 함량이 많을수록 이용할 수 있는 정보의 양도 줄어드는 것을 의미한다.
- [0110] 본 발명에서 도 20은 염기서열 위치당 'N' 함량 그래프로, 염기서열이 'N'으로 된 것은 시퀀서가 해당 염기를 결정할 수 없을 때이며 이의 함량을 그래프로 나타낸 것이다. 아래 그래프에서 약 5%의 리드에서 리드 중간에 'N'이 포함되어 있는 것을 확인할 수 있다.
- [0111] 리드 길이 분포 그래프는 리드 길이의 빈도를 계산하여 나타낸다. 일루미나 플랫폼의 경우 길이가 동일한 리드를 생성하기 때문에, 하나의 피크로 나타날 수 있다.
- [0112] 본 발명에서 도 21은 리드 길이 분포 그래프로, 일루미나 플랫폼의 경우 일정한 길이를 생산하기 때문에 하나의 피크로 나타나는 것을 확인할 수 있다.

- [0113] 리드 중복 그래프는 같은 서열의 리드가 존재할 경우 그 빈도를 계산한 값을 나타낸 그래프로, 특정 서열의 리드가 비정상적으로 많은지 확인할 수 있다. 편향된 서열의 샘플링 또는 어댑터 다이머(adaptor dimer)가 다량으로 존재할 경우 리드의 중복이 높은 빈도로 나타나게 된다. 기대값 이상 중복된 리드의 목록도 표로 요약되어 있다.
- [0114] 본 발명에서 도 22는 리드 중복 그래프로, 아래 그래프는 5,000개 이상의 리드가 동일한 서열을 가지고 있음을 확인할 수 있다.
- [0115] 어댑터 함량 그래프는 어댑터가 리드에 포함되어 있을 경우 그 위치별 퍼센트 함량을 계산한 값을 나타낸다.
- [0116] 본 발명에서 도 23은 어댑터 함량 그래프로, 어댑터 서열이 리드에 포함되어 있을 때 어댑터 서열의 리드상 위치에 따른 함량을 보여줄 수 있다. 어댑터의 종류에 따라 다른 색상으로 표기된다.
- [0117] 또한, K-mer 함량 그래프는 모든 리드를 7개의 k-mer로 나눈 뒤 그 빈도를 계산하여 기대값이상 존재하는 K-mer의 목록을 표로 나타낸 것이다. 그래프를 통해 다수 존재하는 K-mer의 리드의 어떤 위치에 존재하는지도 파악할 수 있다.
- [0118] 본 발명에서 도 24는 K-mer 함량 그래프로, 특정 k-mer의 양이 기대값 이상보다 많은 경우 해당 k-mer의 위치분포 그래프와 함께 보여줄 수 있다.
- [0119] 어댑터 서열제거 단계(302)에서는 Cutadapt(202) 소프트웨어가 사용될 수 있다. 상기 단계에서는 리드에 포함되어 있는 서열 중 일루미나의 Truseq 라이브러리 준비 키트에 해당하는 어댑터 서열을 인식하여 잘라내는 역할을 수행할 수 있다. 어댑터 서열은 리드의 길이가 읽고자 하는 서열의 길이보다 길 때 리드의 끝 쪽에 포함된다. 다른 종류에 해당하는 어댑터 서열 또한 필요시 적용할 수 있다.
- [0120] 리드 품질보정 단계(303)에서는 HTQC(203) 소프트웨어가 사용될 수 있다. 상기 단계에서는 퀄리티가 일정 수준 미만인 염기서열을 잘라내고(trimming) 잘라낸 리드의 길이가 짧을 경우 이를 전체 리드에서 제거(filtering)하는 과정을 수행할 수 있다. 즉 상기 단계를 통해 퀄리티가 좋은 부분만을 유전체 조립에 사용할 수 있도록 할 수 있다.
- [0121] 또한 k-mer 빈도 계산 단계(304)에서는 Jellyfish(204) 소프트웨어가 사용될 수 있다. 상기 단계에서는 리드 품질보정 단계(303)에서 품질보정된 리드에 대하여 k-mer의 개수를 세어 ploidy 또는 예측 유전체의 크기를 계산할 수 있다.
- [0122] 구체적으로, 모든 리드로부터 17 및 19의 k-mer를 각각 추출하여 각 k-mer에 대한 빈도를 계산한다. 결과는 표의 형태로 출력되며 파이프라인의 자체 스크립트를 통해 그래프로의 시각화가 이루어진다. 빈도의 분포로부터 오픈소스 스크립트인 estimate\_genome\_size.pl를 사용하여 유전체의 크기와 시퀀스 커버리지(sequence coverage)를 추정한다. 추정된 유전체의 크기는 분석대상인 균류와 계통학적으로 가까운 종들 중 유전체 조립이 이미 이루어진 것들과 비교하여 서열분석의 퀄리티를 실제 유전체 조립에 앞서 평가해볼 수 있다.
- [0123] 즉, 상기 단계에서는 유전체 조립을 시작하기 앞서 유전체에 대한 정보를 예측함으로써 유전체 조립 결과물의 품질을 가늠해볼 수 있다.
- [0124] 본 발명에서 단계 (c)는 유전체 조립 단계(400)로, 하나 이상의 유전체 조립 소프트웨어를 사용하여 하나 이상의 후보 유전체를 조립하는 단계이다.
- [0125] 페어드-엔드 라이브러리를 나타내는 모식도인 도 14와 같이, 단계 (a)에서 수집된 인서트(조각) 또는 단계 (b)에서 전처리된 인서트(조각)은 상기 인서트의 길이에 따라 두 종류로 구분할 수 있다. 구체적으로 길이가 충분히 짧아 양 쪽 두 개의 리드 서열에 겹치는 부분이 생기는 라이브러리와 인서트의 길이가 길어 두 개의 리드 서열 사이에 읽히지 않는 부분이 생기는 라이브러리로 구분할 수 있다.
- [0126] 본 발명에서는 상기 구분된 인서트에 따라 두 종류의 조립 파이프라인을 선택해서 사용할 수 있다.
- [0127] 구체적으로 하나의 인서트(조각)에서 수집된 두 개의 리드 서열의 합이 인서트(조각)의 서열보다 길 경우 ALLPATHS 파이프라인의 소프트웨어를 사용할 수 있으며, 상기 하나의 인서트(조각)에서 수집된 두 개의 리드 서열의 합이 인서트(조각)의 서열보다 짧을 경우 Non-ALLPATHS 파이프라인의 소프트웨어를 사용할 수 있다.
- [0128] 본 발명에서 도 10은 본 발명에 따른 유전체 조립 단계를 나타내는 모식도로, 유전체 조립 파이프라인은 여러 알고리즘을 이용하는 유전체 조립 오픈소스 소프트웨어를 실행시키는 역할을 수행하며, 라이브러리의 종류에 따



라 ALLPATHS 파이프라인(401) 또는 Non-ALLPATHS 파이프라인(402) 중 적합한 것을 선택하게 할 수 있다.

- [0129] 본 발명에서 ALLPATHS 파이프라인(401)의 소프트웨어는 ALLPATHS 소프트웨어(205)를 사용할 수 있다.
- [0130] 또한, Non-ALLPATHS 파이프라인(402)은 오픈소스 유전체 조립 소프트웨어인 ABySS 소프트웨어(206), SOAPdenovo 소프트웨어(207), Velvet 소프트웨어(208) 또는 Celera 소프트웨어(209)를 사용할 수 있다.
- [0131] ALLPATHS 파이프라인(401)은 필요한 파일을 주어진 파라미터에 따라 자동으로 작성할 수 있다. 이는 실제 ALLPATHS를 실행시켜 보지 않은 연구자들도 특별한 학습 과정 없이 손쉽게 유전체 조립을 할 수 있도록 도와줄 수 있다.
- [0132] ALLPATHS(205)는 주어진 리드로부터 유전체를 조립하는 역할을 수행한다. 상기 ALLPATHS(205)는 조각 라이브러리(fragment library)와 점핑 라이브러리(jumping library)의 두 종류의 라이브러리를 필수로 요구한다.
- [0133] 상기 조각 라이브러리는 읽힌 서열의 길이가 양쪽 리드의 길이의 합보다 작아 겹치는 부분이 발생해야 한다. 예를 들면 리드의 길이가 100 bp라면 읽히는 서열의 길이는 180 bp정도가 되어야 한다. 점핑 라이브러리는 EcoP15I 등 점핑 라이브러리 구축 방법에 따라 3 kbp 내지 10 kbp 사이에 해당하는 길이를 가지는 것이 요구된다.
- [0134] ALLPATHS(205)를 실행시키기 위해서는 in\_groups.csv, in\_libs.csv 두 가지의 파일을 작성하여야 한다. in\_groups.csv는 각 라이브러리의 이름과 해당하는 리드 파일의 위치를 csv 포맷에 맞추어 작성한다. in\_libs.csv는 각 라이브러리에 대해 군류 이름, 라이브러리 종류(조각 라이브러리, 점핑 라이브러리 등), 리드의 페어드-엔드 여부, 읽힌 서열의 크기와 표준편차, 두 리드의 방향성 등 유전체 조립에 필요한 기술적인 정보를 작성한다.
- [0135] ALLPATHS(205)에 의해 유전체 조립이 완료되면 final.assembly.fasta로 최종 어셈블리 파일을 FASTA 포맷으로 얻을 수 있다.
- [0136] 추가로 유전체 조립의 여러 통계적 수치들을 얻을 수 있는데, 이는 동일 디렉토리에 생성되는 assembly.report 파일과 library\_coverage.report 파일에서 확인할 수 있다. assembly.report 파일은 사용한 리드의 정보와 함께 결과물인 어셈블리의 통계치들을 보여준다. 컨티그와 스캐폴드의 개수, N50 값, 1 Mbp에 해당하는 평균 스캐폴드의 수, 전체 어셈블리 중 공백서열(gap)의 비율, 유전체 크기 등 유전체 조립의 평가에 이용되는 다양한 수치들이 포함되어 있다. library\_coverage.report 파일은 전체 유전체 조립을 위해 유전체 크기의 몇 배의 염기 서열을 사용하였는지를 나타낸다. 예를 들면 40 Mbp의 유전체를 조립하는데 120 Mbp의 염기서열에 해당하는 리드를 가지고 조립을 했다면 3배의 시퀀스 커버리지(sequence coverage) 값을 얻게 된다. 이 수치가 크면 클수록 더 많은 데이터를 사용했다는 것을 의미하기 때문에 유전체 서열에 대한 신뢰도가 더 커지게 된다.
- [0137] Non-ALLPATHS 유전체 조립 파이프라인(402)은 ALLPATHS 유전체 조립 파이프라인(401)에 비해 라이브러리의 사용에 있어 유연성이 있으며 페어드-엔드 라이브러리를 기본 골격으로 한다. 이 파이프라인에서는 읽힌 리드들에 대하여 유전체 조립 소프트웨어인 ABySS(206), SOAPdenovo(207), Velvet(208), Celera assembler(209)을 독립적으로 실행시킬 수 있다.
- [0138] ABySS(206)는 짧은 리드들을 조립하는데 적합하며 페어드-엔드 리드와 멀티 스프레드 기능을 지원한다. 구체적으로, ABySS(206)은 조각 라이브러리를 기본적으로 이용하며 점핑 라이브러리가 추가 스캐폴딩을 위해 이용될 수 있다. De Bruijn 그래프를 그리기 위해 모든 리드에서 k-mer를 추출하는데, 이 k-mer의 길이는 유전체 조립을 위한 필수 파라미터이다.
- [0139] 일반적으로 작은 k-mer를 사용하면 반복서열을 조립하는데 어려움이 있으며, 큰 k-mer의 경우 긴 컨티그 조각을 생성하는데 어려움이 있다. 어떤 k-mer가 적합한지에 대해서는 유전체의 반복서열의 분포에 따라 달라질 수 있기 때문에 여러 k-mer를 이용하여 유전체를 조립한 뒤 유전체 평가를 통해 최적 어셈블리를 선택하는 전략을 택한다.
- [0140] Non-ALLPATHS 유전체 조립 파이프라인(402)은 리드의 길이 평균과 최소 k-mer로부터 주어진 리드에 적합한 k-mer의 범위 값을 자동으로 제공할 수 있다.
- [0141] 각 k-mer에 대해 ABySS를 이용한 어셈블리가 진행되고, 유니티그, 컨티그, 스캐폴드가 차례로 생성이 된다. 각 파일들은 FASTA 포맷으로 작성된다. 생성 디렉토리내에는 생성된 어셈블리의 통계적 수치가 계산되어 <prefix>-stats 파일에 작성된다. 이 파일에는 어셈블리 스캐폴드(유니티그, 컨티그)의 개수, 500 bp 이상에 해당하는 스

캐폴드의 개수, 최소 스캐폴드의 길이, N80, N50, N20의 값, 평균 스캐폴드의 길이, 최대 스캐폴드의 길이, 유전체의 크기에 대한 수치를 제공한다.

- [0142] SOAPdenovo(207)는 크기가 큰 유전체를 조립하는데 유리하며 짧은 리드들을 다루는데 최적화 된 소프트웨어이다. 상기 SOAPdenovo(207)는 ABySS와 마찬가지로 k-mer를 필수로 요구하며, 전술한 바와 같이 Non-ALLPATHS 유전체 조립 파이프라인(402)에서 적합한 범위의 값을 선택할 수 있다.
- [0143] 상기 SOAPdenovo(207)는 추가로 제공하는 라이브러리의 정보를 config 파일에 작성하여야 한다. 읽힌 서열의 평균 길이, 리드의 방향성, 조립 종류 (컨티그 조립, 스캐폴드 조립, 모두 조립), 유전체 조립에 사용하는 최대 리드의 길이, 라이브러리의 우선순위, 스캐폴딩을 위한 최소 페어의 개수, 최소 정렬 길이의 작성이 요구된다. Non-ALLPATHS 유전체 조립 파이프라인(402)은 주어진 파라미터와 간단한 계산 과정을 거쳐 config 파일을 자동으로 작성한다.
- [0144] SOAPdenovo(207)의 각 k-mer의 유전체 조립 결과 파일은 <prefix>.contig, <prefix>.scafSeq 파일에 각각 FASTA 포맷으로 저장된다. 어셈블리에 대한 통계적 수치는 <prefix>.scafStatistics 파일에 작성되어 있으며, 스캐폴드 수, 평균 스캐폴드 길이, 최소/최대 스캐폴드 길이, A, T, G, C의 함량, N10 - N90 값이 계산되어 제공된다.
- [0145] Velvet(208)은 리드로부터 de Bruijn 그래프를 구축하여 유전체 조립을 수행하며 시퀀싱 에러를 제거하거나 반복서열을 단순화시키는 프로세스를 포함한다. 상기 Velvet(208)은 velveth와 velvetg의 두 개의 프로그램을 순차적으로 실행하여 유전체를 조립한다. Velveth는 리드로부터 주어진 k-mer를 추출하여 데이터 구조 중 하나인 사전에 저장하는 역할을 하며 velvetg는 구축된 사전을 이용하여 de Bruijn 그래프를 만들고 시퀀싱 에러를 제거하고 그래프를 단순화하여 반복서열을 그래프상에서 해결하여 최종 유전체 조립을 완성한다.
- [0146] 전술한 두 소프트웨어와 마찬가지로 k-mer를 필수로 요구하며 Non-ALLPATHS 유전체 조립 파이프라인(402)에서 적합한 범위의 값을 선택할 수 있다.
- [0147] Velvetg(208)의 경우 결과 어셈블리에 큰 영향을 끼치는 두 종류의 파라미터를 추가로 요구하는데, exp\_cov 옵션과 cov\_cutoff 옵션이다. exp\_cov 옵션은 예상되는 시퀀스 커버리지(sequence coverage)를 유전체 조립에 직접 이용하기 위해 필요하며, cov\_cutoff 옵션은 유전체 조립에 사용할 최소 시퀀스 커버리지(sequence coverage)를 지정하는데 이 두 값에 따라 상이한 결과 어셈블리를 얻게 된다.
- [0148] 두 옵션에 해당하는 값을 최적으로 구하기 위해 추가로 VelvetOptimiser 소프트웨어를 이용할 수 있다. VelvetOptimiser는 k-mer의 범위만 사용자에게 의해 지정되면 나머지 파라미터들은 여러 차례의 유전체 조립 시도와 자체 평가를 통해 최적으로 결정된다. VelvetOptimiser는 유전체 조립 결과 어셈블리를 contigs.fa이라는 파일에 FASTA 포맷으로 저장한다. 결과 어셈블리에 대한 통계적 수치는 로그 파일을 통해 제공하며, 유전체 크기, N50 값, 최대 스캐폴드 길이, 1 Kbp 길이 이상의 스캐폴드 개수가 포함된다.
- [0149] Celera assembler(209)는 반복서열 검색(find repeats), 오버랩(overlap), 에러 수정(error correction), 유니티그 생성(unitigging), 스캐폴딩(scaffolding)의 과정을 순차적으로 수행하여 유전체를 조립한다. 상기 Celera assembler(209)는 앞서 세 소프트웨어와 다르게 Overlap-layout-consensus 알고리즘을 사용하기 때문에 리드를 k-mer로 쪼개지 않고 리드 자체를 그래프를 그리는데 이용한다. 따라서 K-mer를 제공하지 않아도 된다.
- [0150] Celera assembler(209)는 일반적으로 이용되는 리드의 포맷인 FASTQ를 직접 이용할 수 없기 때문에 적당한 포맷으로 변경해주어야 한다. 이는 Celera assembler 패키지에 포함되어 있는fastqToCA 스크립트를 통해서 얻을 수 있다.
- [0151] Celera assembler(209)는 유전체 조립 결과 어셈블리를 <prefix>.scf.fasta 파일에 FASTA 포맷으로 저장한다. 유전체의 통계적 수치는 <prefix>.qc 파일에 저장되며 스캐폴드 개수, 컨티그 개수, 스캐폴드를 이루는 평균 컨티그 개수, 조립 유전체 크기, 최소/최대 스캐폴드 길이, N25, N50, N75 값, 2 Kbp 이상 스캐폴드 개수, 공백서열의 수와 평균 길이가 포함된다.
- [0152] 본 발명에서 단계 (d)는 유전체 조립 평가 단계(500)로 전술한 단계 (c)에서 조립된 후보 유전체를 평가하여 최적 유전체를 선별하는 단계이다.
- [0153] 본 발명에서 도 11은 유전체 조립 평가 단계를 나타내는 모식도로, 상기 단계에서는 여러 후보 어셈블리(후보 유전체)에 대한 평가가 수행될 수 있다. 이러한 조립된 후보 유전체의 평가는 우도 평가 단계(501), N50값 평가 단계(502), 스캐폴드 개수 평가 단계(503) 및 최대 스캐폴드 길이 평가 단계(504)로 이루어진 그룹 중에서 하나

이상을 수행할 수 있다. 상기 단계들은 유전체 조립 평가 파이프라인으로 구축되어 수행할 수 있다.

- [0154] 우도(likelihood) 평가 단계(501)는 조립한 어셈블리가 원래의 유전체 서열과 얼마나 유사한지를 간접적으로 측정하는 단계로, CGAL(210)을 실행시킴으로써 이루어진다. 상기 우도 평가에서는 각 스캐폴드를 조립하기 위해 사용된 리드 수가 얼마나 균일한지, 시퀀싱 에러가 얼마나 존재하는지, 읽힌 서열의 길이가 얼마나 균일한지, 유전체 조립에 사용되지 않은 리드 수가 얼마나 많은지를 종합적으로 평가하여 하나의 값인 우도로 계산해준다.
- [0155] CGAL(210)은 기존 다른 유전체 조립 평가 프로그램들이 이미 조립이 잘 이루어진 유전체와 비교하여 평가를 하였던 반면, 비교할 유전체를 요구하지 않는다는 점에서 처음으로 유전체를 조립하는 경우에 유리하다. CGAL(210)은 유전체 조립 결과물인 어셈블리에 조립에 이용했던 리드를 정렬한 BAM 포맷의 파일을 필요로 한다. 따라서 Bowtie2를 이용하여 리드를 어셈블리로 정렬하여 BAM 포맷의 파일을 얻는다. 유전체 조립 평가 파이프라인(500)은 생성된 어셈블리마다 BAM 파일을 생성한다.
- [0156] CGAL(210)에 의한 계산의 결과물은 로그 값을 취한 음의 정수의 값으로 얻게 되며, 절대 값이 작을수록 신뢰할 수 있는 어셈블리임을 의미한다.
- [0157] N50값 평가 단계(502)는 각 유전체 조립 소프트웨어의 로그 파일에서 추출된 N50 값을 비교하여 어떤 소프트웨어가 전체적으로 긴 스캐폴드를 조립하였는지를 비교한다. 상기 N50값은 스캐폴드의 평균 길이 또는 중앙값과 유사하지만 소수의 긴 스캐폴드가 조립될 때 큰 값을 가진다. 따라서 N50값이 크면 길이가 긴 스캐폴드가 다수 조립되었음을 의미한다.
- [0158] 스캐폴드 개수 평가 단계(503)는 유전체 조립 소프트웨어에서 생성된 어셈블리가 가지는 스캐폴드의 개수를 비교하여 어떤 어셈블리가 더 적은 스캐폴드를 생성했는지를 비교한다. 유사한 어셈블리 크기에 대하여 스캐폴드의 수가 적을수록 더 조립이 많이 이루어졌음을 의미한다. 스캐폴드 개수는 각 유전체 조립 소프트웨어의 로그 파일에서 추출할 수 있다.
- [0159] 최대 스캐폴드 길이 평가 단계(504)는 조립된 스캐폴드 중 가장 길게 조립이 된 스캐폴드의 길이를 비교한다. 최대 스캐폴드 길이가 길수록 유전체 조립이 더 많이 이루어졌음을 의미한다. 최대 스캐폴드 길이는 각 소프트웨어의 로그 파일에서 추출할 수 있다.
- [0160] 또한, 사용된 리드 수 평가 단계(505)는 주어진 전체 리드에서 얼마만큼의 리드를 사용하여 유전체를 조립하였는지를 비교한다. 사용된 리드 수가 많을수록, 즉 생성되는 유전체 조립의 시퀀스 커버리지(sequence coverage)가 높을수록 유전체 조립이 더 우수하게 이루어졌음을 의미한다. 사용된 리드 수는 ABySS(206), Velvet(208)의 로그 파일에서 추출 가능하며, SOAPdenovo(207)와 Cellear Assembler(209)은 로그 파일에서 따로 해당 정보를 제공하지 않기 때문에 전체 리드를 다시 어셈블리에 정렬하여 정렬이 된 리드의 개수를 셀 수 있다.
- [0161] 본 발명에서는 전술한 우도 평가(501), N50값 평가(502), 스캐폴드 개수 평가(503) 및 최대 스캐폴드 길이 평가(504)의 후보 유전체 평가를 수행한 후, 각각의 평가 별로 순위를 부여하여 순위의 합이 가장 적은 유전체를 최적 유전체로 선별할 수 있다.
- [0162] 본 발명에서 단계 (e)는 유전체 조립 후처리 단계(600)로, 최적 유전체를 후처리하는 단계이다. 상기 단계(600)에서는 어셈블리, 즉 유전체의 수치적 정보, 외래 DNA에 의한 오염 확인, 미토콘드리아 서열 조립, 계통학적 연관 종과의 비교 등을 차례로 수행하며 웹서버와의 통신을 통해 연구자가 웹 브라우저를 통해 어셈블리에 관한 정보를 종합적으로 확인할 수 있게 한다.
- [0163] 본 발명에서 도 12는 본 발명에 따른 유전체 조립 후처리 단계를 나타내는 모식도로, 상기 유전체 조립 후처리 단계(600)는 외래 DNA에 의한 오염 확인 단계(601), 28S rDNA 서열 확인 단계(602), 미토콘드리아 서열 조립 단계(603) 및 웹서버 업로드 단계(604)로 이루어진 그룹 중에서 하나 이상을 수행할 수 있다.
- [0164] 일 구체예에서, 상기 단계에서는 오픈소스 소프트웨어인 BLAST+(211)를 사용할 수 있다. 상기 BLAST+(211)는 주어진 DB 서열에 대해 조립된 어셈블리를 정렬하여 유사한 서열이 존재하는 지 검색할 수 있다.
- [0165] 외래 DNA에 의한 오염 확인 단계(601)는 DNA 추출 도중 오염 등의 원인에 의해 분석한 균류의 DNA 서열 이외에 다른 서열이 유전체 조립 결과에 포함되어 있는 경우 이를 찾아내고 제거할 수 있도록 한다.
- [0166] 상기 외래 DNA를 확인하기 위해 GC 함량과 시퀀스 커버리지(sequence coverage)의 두 종류의 통계수치를 이용할 수 있다.
- [0167] GC 함량은 균류의 유전체로부터 세균의 유전체를 확인하는데 유용하다. 모든 생물종은 고유의 GC 함량을 가지며

세균이 균류에 비해 더 높은 평균 GC 함량을 가지는 것이 보고되고 있다. 따라서 GC 함량이 높은 스캐폴드의 경우 배양 시 세균의 DNA가 유입되어 서열 분석이 되었을 가능성이 있다.

- [0168] 각 스캐폴드의 시퀀스 커버리지(Sequeunce coverage)는 외래 DNA를 확인하는데 유용하다. 서열분석을 하기 전 세포를 배양하는데, 이 때 증식된 세포 수만큼의 유전체 DNA를 가지게 된다. 이것을 임의로 조각내고 서열분석을 하므로 스캐폴드당 시퀀스 커버리지 분포는 일정 값을 평균으로 하는 정규곡선을 그릴 것이다. 따라서 이 정규곡선에서 벗어나는 시퀀스 커버리지를 가지는 스캐폴드, 즉 기대값보다 낮은 시퀀스 커버리지를 가지는 스캐폴드나 높은 시퀀스 커버리지를 가지는 스캐폴드의 경우 외부에서 유입된 DNA 조각이거나 미토콘드리아의 DNA 서열일 가능성이 있다.
- [0169] 각 스캐폴드는 NCBI에 등록된 모든 서열에 정렬하여 유사한 서열이 어떤 종에 속하는지를 확인할 수 있다.
- [0170] 본 발명에서는 GC 함량 및 시퀀스 커버리지를 이용하여 2차원의 그래프를 그리고, NCBI에서 정렬된 서열의 종을 계(kingdom) 또는 문(phylum)에 따라 색을 다르게 하여 세균 등 외래 DNA의 존재를 한 눈에 확인할 수 있도록 할 수 있다.
- [0171] 28S rDNA 서열 확인 단계(602)는 유전체 어셈블리 중 28S rDNA에 해당하는 서열을 추출하여 균류 28S rDNA 서열 데이터베이스인 RDP에 정렬 후, 어떤 종에 속하는지를 확인한다.
- [0172] 28S rDNA 서열은 PCR 과정을 통해 전체 유전체 분석 없이도 쉽게 얻을 수 있기 때문에 많은 균류 종에 대하여 데이터베이스화 되어 있으며, 조립된 어셈블리가 실제 분석하고자 했던 종의 서열과 동일한지 확인하기 위해 사용할 수 있다.
- [0173] 이때, BLAST+(211)는 주어진 쿼리(query)서열을 데이터베이스 서열에 정렬한다. 서열이 정렬된 종이 분석하고자 했던 종과 일치하거나 근연종과 높은 유사성을 가지는 경우 분석하고자 했던 종을 알맞게 분석했음을 확인할 수 있다.
- [0174] 미토콘드리아 서열 조립 단계(603)는 주어진 리드로부터 해당 균류의 미토콘드리아의 염색체 서열을 조립한다.
- [0175] 미토콘드리아는 진핵생물의 소기관 중 하나이며 독립적으로 작동하는 염색체를 가진다. 미토콘드리아의 수는 세포당 100 내지 1000 개이기 때문에 일정 수의 균류의 세포를 서열분석을 했을 때 균류의 세포 수에 비해 100배에서 1000배에 달하는 미토콘드리아의 염색체 수를 얻을 수 있다. 따라서 조립 후 시퀀스 커버리지도 균류의 염색체에 비해 높은 값을 가지게 될 것이다. 미토콘드리아 염색체 서열의 조립은 이런 특징을 이용하여 이루어진다.
- [0176] 상기 미토콘드리아 서열 조립 단계(603)에서는 Velvet(208)의 exp\_cov와 cov\_cutoff 옵션을 이용한다. 먼저 Velvet(208)을 이용하여 처음 리드들을 조립을 하면 Velvet이 제공해주는 정보에 따라 각 스캐폴드에 해당하는 시퀀스 커버리지(sequence coverage)를 얻을 수 있다. 이 스캐폴드들을 NCBI의 미토콘드리아 데이터베이스에 정렬하여 조립한 스캐폴드 중 미토콘드리아 서열에 해당하는 것을 추출한다. 미토콘드리아의 서열에 해당하는 스캐폴드의 시퀀스 커버리지(sequence coverage)를 추출하여 평균값과 표준편차를 구하고, 이를 미토콘드리아 염색체를 조립하기 위해 실행하는 Velvet의 exp\_cov와 cov\_cutoff의 파라미터로 넣어준다. exp\_cov는 평균값이 되고, cov\_cutoff는 (평균값 - 표준편차) 값이 된다.
- [0177] 웹서버 업로드 단계(604)는 유전체 조립의 전과정동안 생성된 결과물을 웹서버로 업로드하여 사용자가 쉽게 결과를 확인하고 판단할 수 있도록 할 수 있다.
- [0178] 상기 웹서버 업로드(604)에서 보여지는 결과물은 FastQC 결과(605), Read QC 결과(606), K-mer frequency 결과(607), 28S rDNA 서열 정렬 결과(608), GC-cov 결과(609) 및 어셈블리 통계치(610)로 구성될 수 있다.
- [0179] FastQC 결과(605)는 리드의 전처리(300) 파이프라인 중 수행된 FastQC(201)의 결과물을 각각의 라이브러리에 대해 보여준다.
- [0180] Read QC 결과(606)는 리드의 어댑터를 어댑터 서열 제거(302), 리드 품질보정(303)의 결과를 수치적으로 보여준다.
- [0181] K-mer frequency 결과(607)은 주어진 각 라이브러리에 대해 k-mer 빈도 계산(304)에 의해 나온 값인 추정 시퀀스 커버리지(sequence coverage)와 추정 유전체 크기를 보여주며, k-mer 중첩 정도에 따른 빈도를 그래프로 나타낸다.



- [0182] 28S rDNA 서열 정렬 결과(608)는 어셈블리 후처리(600) 중 28S rDNA 서열 확인(602)의 결과를 보여준다. 스캐폴드 이름, 정렬된 길이의 coverage, 유사성, 정렬된 서열의 종이 포함된다.
- [0183] GC-cov 결과(609)는 어셈블리 후처리(600) 중 외래 DNA에 의한 오염 확인(601)의 결과를 보여준다. 각 스캐폴드의 GC 함량과 coverage를 분산형 그래프로 나타내고 BLAST+(211)의 결과를 각 분류학의 계층에 따라 각 점의 색으로 표시한다.
- [0184] 어셈블리 통계치(610)는 각 유전체 조립 결과물로부터 N50 값, 유전체 크기, 최대 스캐폴드 길이, 스캐폴드 개수, 사용된 리드 수, CGAL 값을 표로 나타낸다.
- [0186] [실시예]
- [0187] 유전체 조립(10) 파이프라인을 실제 데이터에 적용하여 그 결과를 살펴보았다.
- [0188] 사용한 컴퓨팅 자원은 Intel<sup>®</sup> Xeon E5-2640 (15M Cache, 2.50 GHz, 7.20 GT/s Intel<sup>®</sup> QPI), 256 GB RAM, Ubuntu 14.04 64 bit 운영체제이다. 하드드라이브는 약 200 GB의 여유공간을 확보하였다.
- [0189] 사용한 데이터는 *Choanephora cucurbitarum* 중에서 DNA를 추출하여 일루미나 플랫폼에서 생성한 조각 라이브러리(10,572,329개 리드)였다.
- [0190] *Choanephora cucurbitarum*은 조롱박(cucurbits)을 포함하여 호박, 주키니(zucchini), 박류에 식물 병원균으로 작용한다. 또한 접합균류(Zygomycota)에 속하는 균으로서 접합균류는 유전체 연구가 많이 되어 있지 않기 때문에 이 균주의 유전체를 분석함으로써 접합균류의 유전체 특성을 파악하는 데도 도움을 줄 수 있다.
- [0191] 현재 *C. cucurbitarum*의 유전체 서열은 어떤 데이터베이스에도 공개되어 있지 않다.
- [0192] 유전체 조립 파이프라인(10)을 실행하기 앞서 필요한 오픈소스 소프트웨어(201 내지 211)를 각 소프트웨어의 설치 가이드에 따라 설치하였다.
- [0193] 리드 전처리 파이프라인(300)을 이용하여 리드 품질요약 단계(301), 어댑터 서열 제거 단계(302), 리드 품질보정 단계(303), k-mer 빈도 계산 단계(304)를 진행하였다.
- [0194] 리드 품질요약은 FastQC(201)의 결과물로서 확인하였다.
- [0195] 본 발명에서 도 25는 *C. cucurbitarum* 유전체 리드의 염기당 퀄리티 분포를 나타낸다.
- [0196] 상기 도 25에 나타난 바와 같이, 염기당 퀄리티 분포 그래프는 리드의 퀄리티가 5 ‘upstream은 30 이상으로 신뢰할 수 있는 퀄리티를 보여주는 반면 3’ 방향으로 갈수록 점차 퀄리티가 낮아지는 양상을 보여주었다. 특히 읽혀진 양쪽 리드 중 한 쪽인 리드 1의 260 염기부터는 평균 퀄리티가 20보다 낮아지기 때문에, 상당수의 리드가 이 근처염기에서 품질보정 과정 동안 잘려나갈 것으로 예측할 수 있다(위 그래프). 리드 2는 리드 1보다 상대적으로 낮은 퀄리티 양상을 보여주고 있으며 220 염기부터 평균 퀄리티가 20 보다 낮아지고 있는 것을 확인할 수 있다(아래 그래프).
- [0197] 도 26은 *C. cucurbitarum* 유전체 리드의 리드당 평균 퀄리티 분포를 나타낸다.
- [0198] 상기 도 26에 나타난 바와 같이, 리드당 평균 퀄리티 분포는 리드 1과 리드 2에서 상당한 차이를 보여주고 있는데, 리드 1은 리드당 평균 퀄리티가 33과 39에서 가장 많은 빈도수를 보여주고 있고(위 그래프), 리드 2는 14와 28 근처에서 가장 많은 빈도수를 보여주고 있다(아래 그래프). 평균 퀄리티의 경우 그 편차를 고려하지 않았기 때문에 해당 리드를 사용할지 안할지를 결정할 수 있는 직접적인 요인은 아니지만 개략적인 퀄리티를 확인하는 좋은 방법이 된다.
- [0199] 도 27은 *C. cucurbitarum* 유전체 리드의 염기서열 위치당 ATGC 함량 그래프이다.
- [0200] 상기 도 27에 나타난 바와 같이, ATGC 함량 및 GC 함량 그래프는 주어진 리드가 일반적인 유전체 DNA 서열 리드의 분포와 유사함을 보여주고 있다. 리드 2의 경우 염기서열 뒤편으로 갈수록 ATGC의 함량이 변하는 것을 볼 수 있는데(아래 그래프) 이는 퀄리티가 낮은 염기서열이 전체 ATGC함량에 영향을 미친 결과이다.
- [0201] 리드 중 특정 서열이 비정상적으로 많이 존재할 경우 해당 서열을 FastQC 결과를 통해 확인할 수 있는데, 본 발명의 실시예에서 사용한 데이터에서는 1개의 서열이 확인되었으며, 이는 어댑터 서열과 일치하였다. 이 서열은

전체 리드의 약 1%에서 발견되었다. 따라서 이 어댑터 서열을 인식하여 자르는 과정이 필요하다는 것을 알 수 있다.

- [0202] 어댑터 서열 제거(302) 과정을 통해 전체 리드 1, 리드 2 중 386,490(3.7%), 495,106(4.7%)의 리드서 53,501,276(1.88%), 909,916,970(0.52%)의 염기서열이 어댑터로 각각 인식되어 제거되었다.
- [0203] 도 28은 *C. cucurbitarum* 유전체 리드 품질보정 그래프이다. 상기 도 28에 나타난 바와 같이, 염기서열당 켈리 티 분포를 품질보정 전(위) 및 품질보정 후(아래)로 나타내었으며, 리드의 오른쪽에 해당하는 부분의 평균 켈리 티가 높아진 것을 확인할 수 있다.
- [0204] 구체적으로, 리드 품질보정(303) 단계에서는 10,572,329의 전체 리드 중 1,402,442개(13.26%)의 리드가 낮은 켈 리티로 분류돼 전체 리드에서 제거되었다. 따라서 9,169,887의 리드가 유전체 조립을 위해 사용되었다. 이는 리 드 하나의 길이를 250 bp로 가정했을 때 4,584,943,500개의 염기서열에 해당하며 균류의 유전체를 35 Mbp로 가 정했을 때 균류 유전체 크기의 131배에 해당하는 염기서열 수이다. 균류 유전체를 조립하기위한 충분한 수의 염 기서열을 확보하였음을 간접적으로 확인할 수 있었다.
- [0205] K-mer 빈도 계산(304) 단계에서는 품질보정이 완료된 리드에 대해 k-mer 빈도를 계산하여 시퀀스 커버리지 (sequence coverage) 및 유전체 크기를 예측하였다.
- [0206] 도 29는 K-mer histogram 그래프로, x축은 해당 k-mer의 중복 회수이며, y축은 해당 중복 회수를 가진 서로 다 른 k-mer의 수를 나타낸다. 상기 그래프에 나타나듯이, 두 종류의 k-mer 17과 19에 대해 독립적으로 수행하였고 일반적인 k-mer frequency 곡선을 얻었다. 예측 유전체 예측 유전체 크기는 29.25 Mbp와 29.58 Mbp로 나타났다.
- [0207] ALLPATHS 유전체 조립 파이프라인(401)의 실행을 위해서는 두 종류의 라이브러리가 필요한데, 실시예에서 사용 한 데이터는 조각 라이브러리(조각 서열의 길이는 평균 400 bp)이므로 점핑 라이브러리를 시뮬레이션을 통해 생 산하였다.
- [0208] 먼저 ABySS(206)를 이용하여 1차 유전체 조립을 실시한 뒤 이 어셈블리로부터 오픈소스 소프트웨어인 wgsim으로 시뮬레이션하여 4,557,436개의 리드를 얻었다.
- [0209] 실제 데이터인 조각 라이브러리와 시뮬레이션 데이터인 점핑라이브러리를 이용하여 ALLPATHS 유전체 조립 파이프라인을 진행하였다. 그 결과 29.0 Mbp의 유전체 크기를 가진 2,814개의 스캐폴드를 조립하였다. N50값은 28 Kbp, 시퀀스 커버리지(sequence coverage)는 86.2x로 우수한 유전체 조립을 보여주었다. 이전 단계에서 예측한 유전체 크기가 일치함을 확인하였다.
- [0210] ALLPATHS 유전체 조립 파이프라인(401)에서 생성된 최종 어셈블리는 어셈블리 후처리(600) 파이프라인에서 최종 어셈블리를 다양한 측면에서 분석하였다.
- [0211] 도 30은 *C. cucurbitarum* 최적 어셈블리 컨티그 누적길이 및 GC함량 그래프로, ALLPATHS 유전체 조립 파이프라인으로부터 최종 선별된 어셈블리의 컨티그 누적길이(위)와 주어진 윈도우 크기의 GC함량 빈도 그래프(아래)를 보여주고 있다.
- [0212] 또한, 도 31은 *C. cucurbitarum* 어셈블리의 외래 DNA 오염 확인을 위한 GC 함량과 각 스캐폴드의 coverage를 나타낸 분산형 그래프로, 외래 DNA에 의한 오염확인(601) 과정에서 GC 함량과 스캐폴드당 sequeunce coverage를 계산하여 분산형 그래프로 확인하였고, 세균에 의한 오염이나 외래 DNA에 의한 뚜렷한 오염은 발견되지 않았다. NCBI nt 데이터베이스에 정렬되어 색이 표기된 스캐폴드의 경우 대다수가 Fungi로 밝혀졌다.
- [0213] 많은 스캐폴드가 NCBI nt 데이터베이스에 정렬이 안 된 것은 스캐폴드의 길이가 짧거나 데이터베이스에 집합균 류에 해당하는 균류의 서열정보가 충분하지 않기 때문인 것으로 보인다.
- [0214] 따라서 최종 어셈블리가 외래 DNA에 의한 오염없이 진행되었음을 확인하였다.
- [0215] 28S rDNA 서열 확인(602)은 RDP (Ribosomal Database Project) 데이터베이스에 정렬된 서열의 정보를 정리된 표의 형태로 얻었다.
- [0216] 도 32는 *C. cucurbitarum*의 28S rDNA 서열 정렬한 표로, 조립한 어셈블리의 분자생물학적 분류를 위해 어셈블리 를 균류의 28S rDNA 서열을 데이터베이스화한 RDP에 BLASP로 정렬하여 유사한 서열을 검색한 결과이다. 총 2개 의 스캐폴드에서 정렬이 되었으며 두 종 모두 알려진 근연종으로서 서열분석한 종이 *C. cucurbitarum*임을 확인

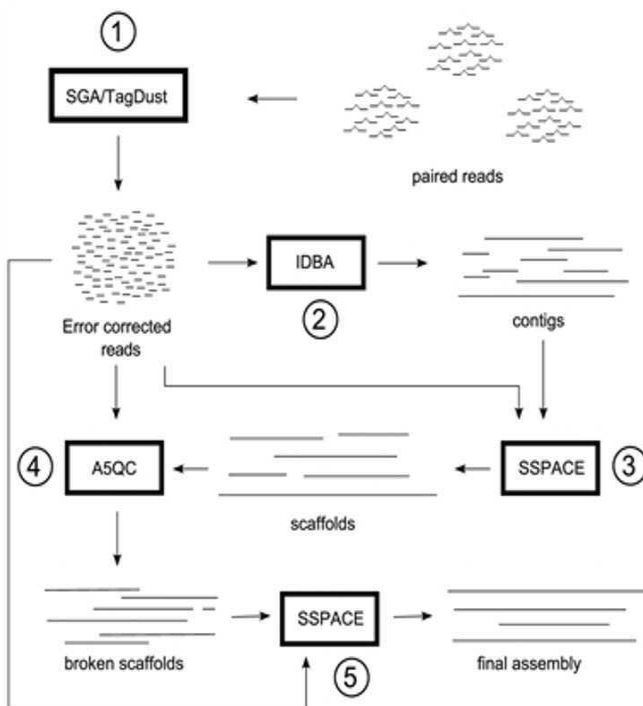
할 수 있다.

[0217] 구체적으로, scaffold\_1308과 scaffold\_1520 두 개의 스캐폴드에서 rDNA 서열이 발견되었으며 각 스캐폴드는 *C. cucurbitarum*과 근연종인 *Rhizopus*와 *Mucor* 속에 정렬되었다. 따라서 서열분석하고자 했던 균류의 유전체를 올바르게 분석하였음을 확인하였다.

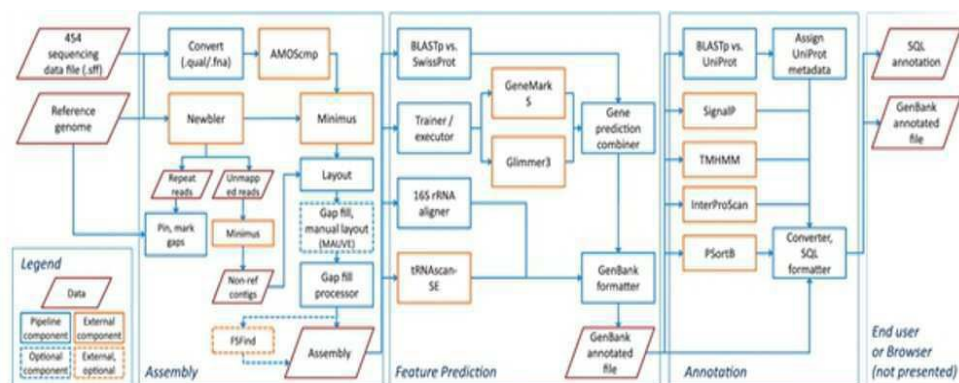
[0218] 미토콘드리아 서열 조립(603) 과정은 미토콘드리아 서열 데이터베이스에 최종 어셈블리를 정렬하였고 167개의 스캐폴드가 미토콘드리아의 서열에 정렬이 되었다. 이 중 150개는 500 bp 미만의 짧은 서열이고 제일 긴 스캐폴드의 길이는 47 Kbp였다. 시퀀스 커버리지(sequence coverage)를 이용하여 Velvet을 이용하는 미토콘드리아 서열 조립 파이프라인(603)에 의해 조립한 결과 203개의 스캐폴드로 이루어진 미토콘드리아 유전체 어셈블리를 얻었다. 이 중 5 Kbp이상인 스캐폴드는 11개였다.

## 도면

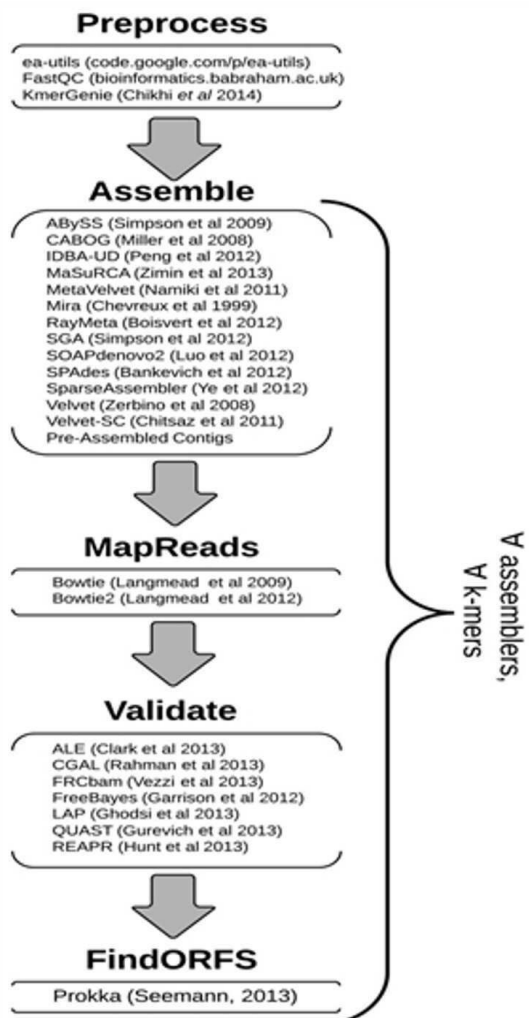
### 도면1



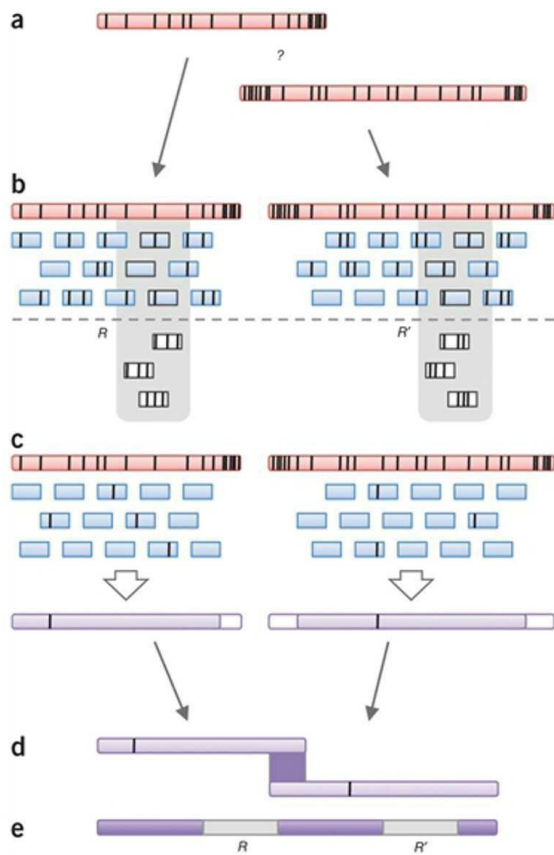
### 도면2



도면3



도면4



도면5

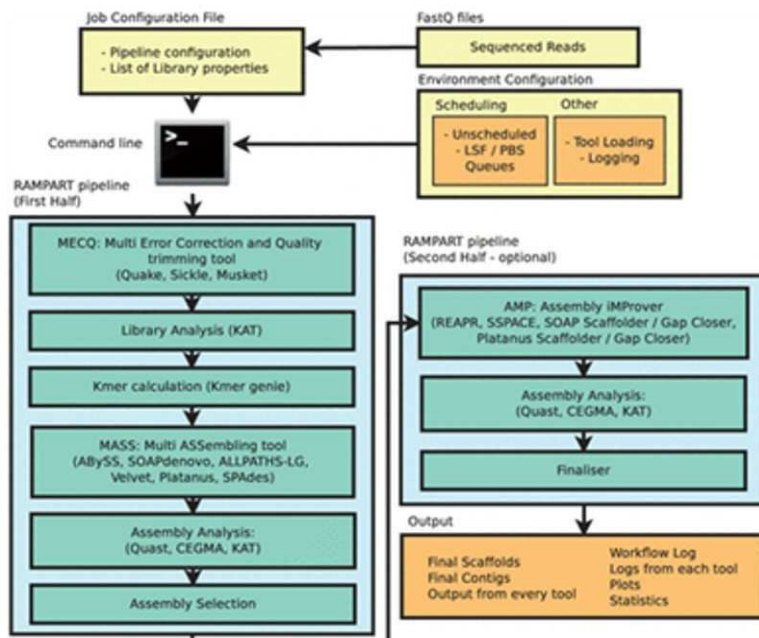
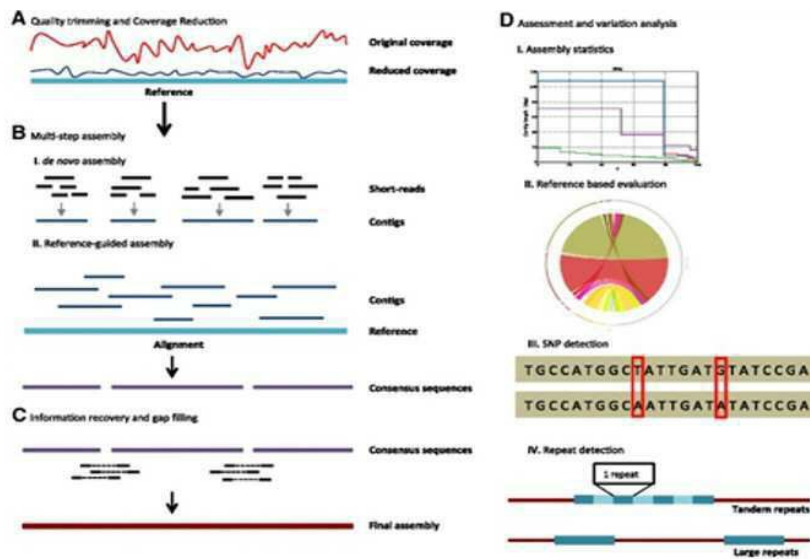
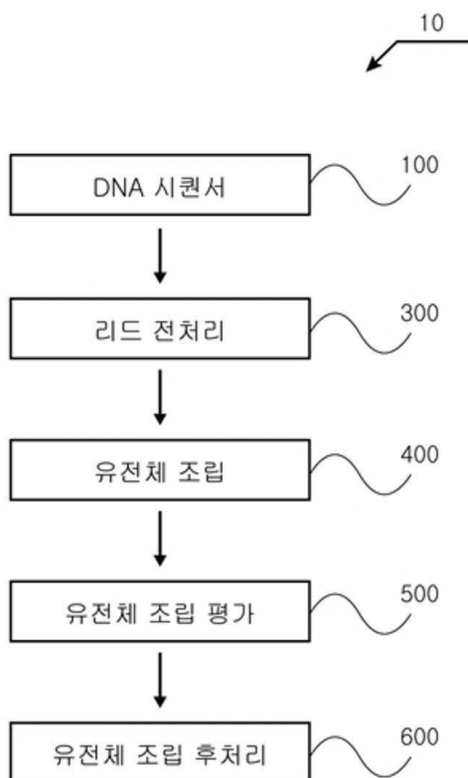


그림 6 RAMPART 모식도

도면6

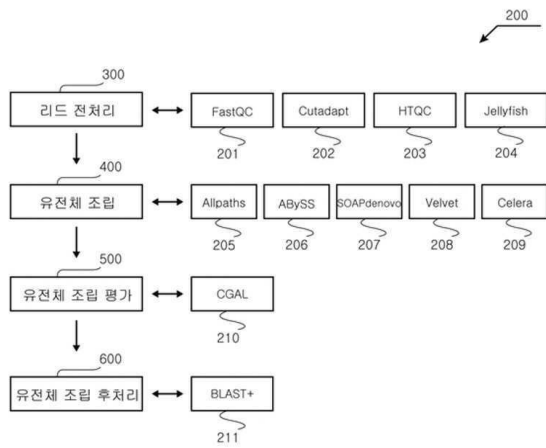


도면7

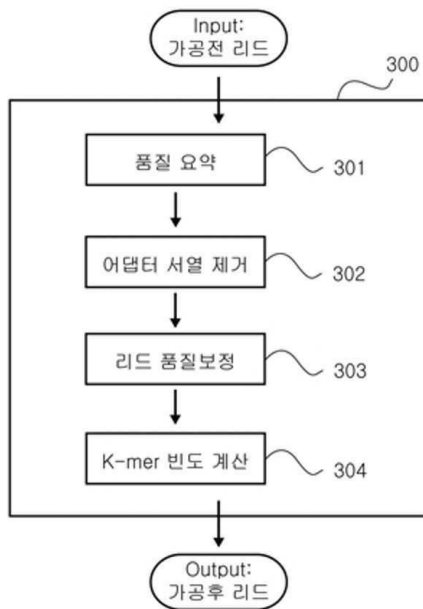




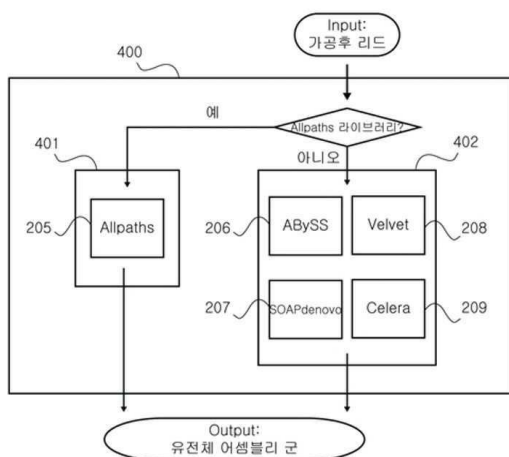
도면8



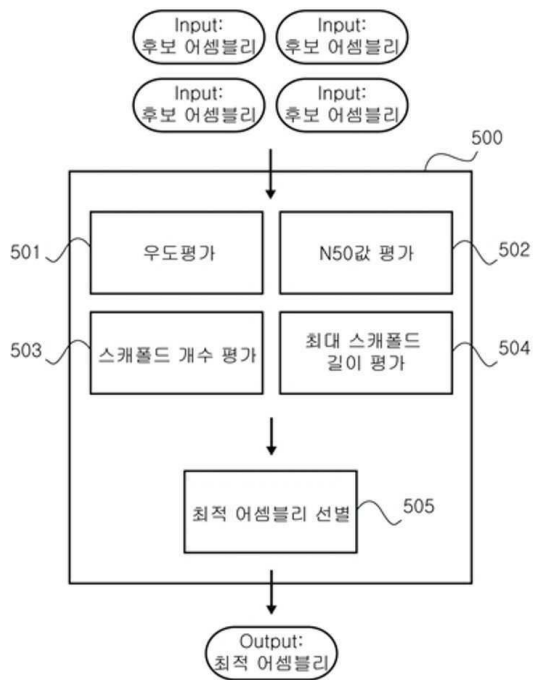
도면9



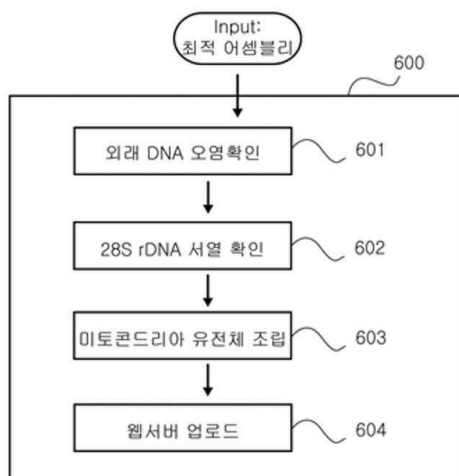
도면10



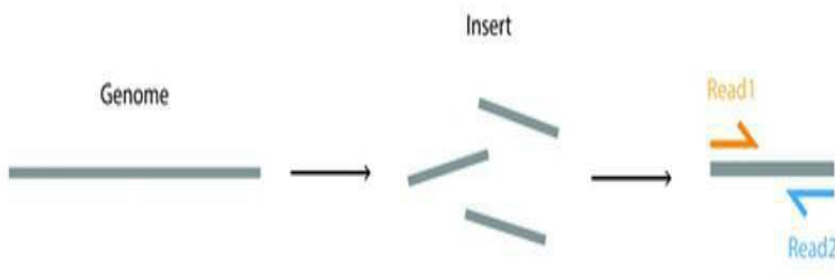
도면11



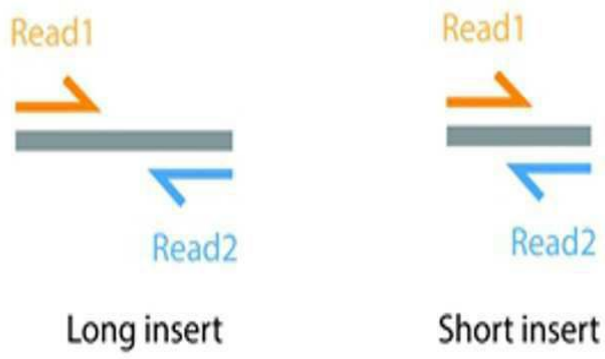
도면12



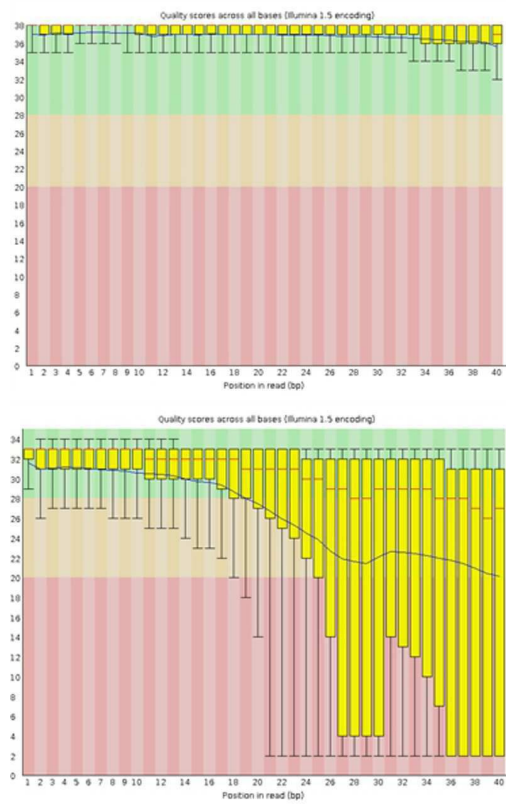
도면13



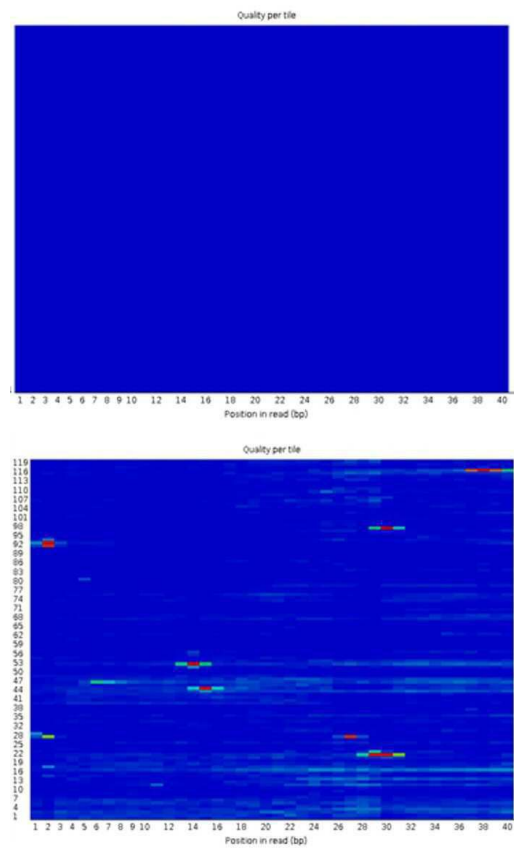
도면14



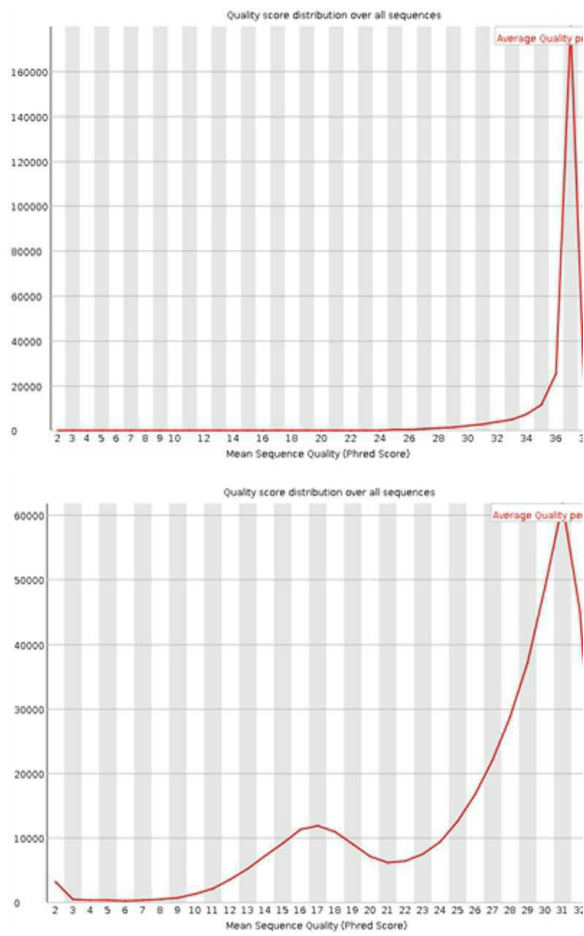
도면15



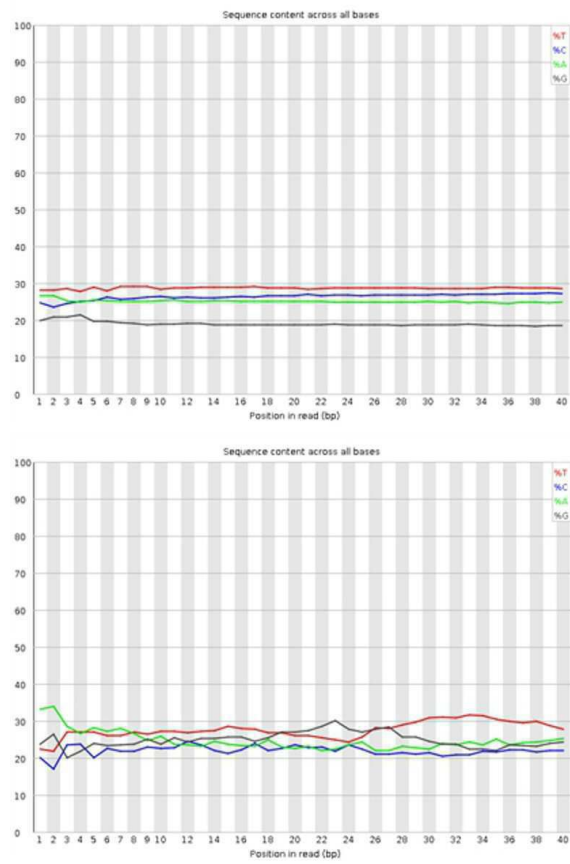
도면16



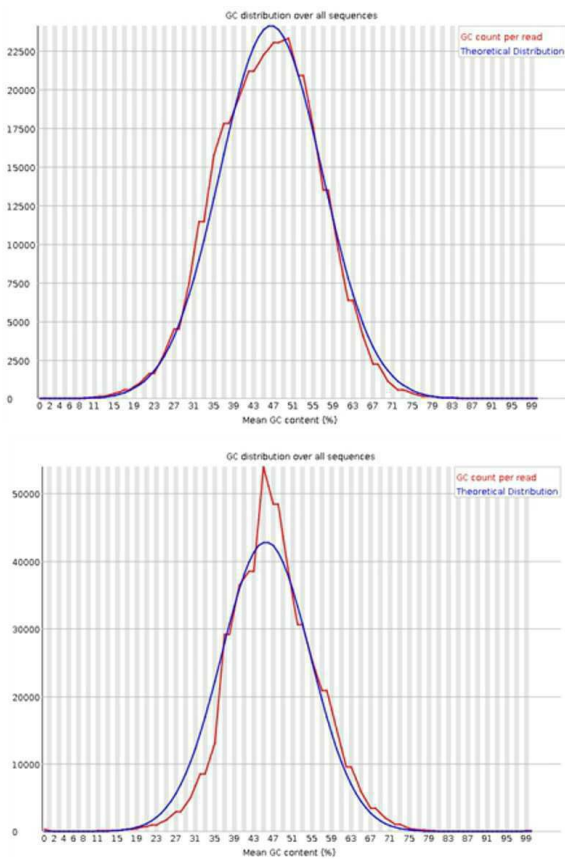
도면17



도면18

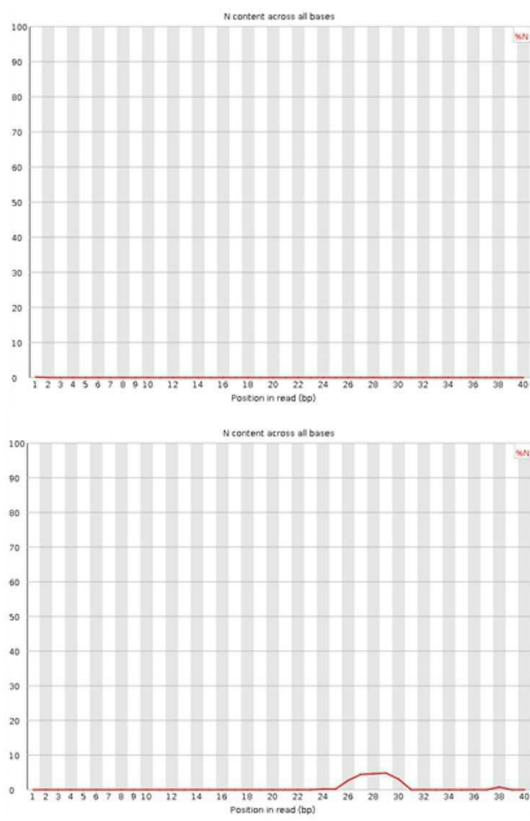


도면19

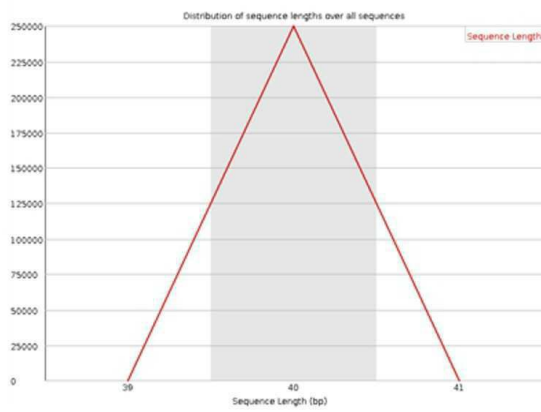




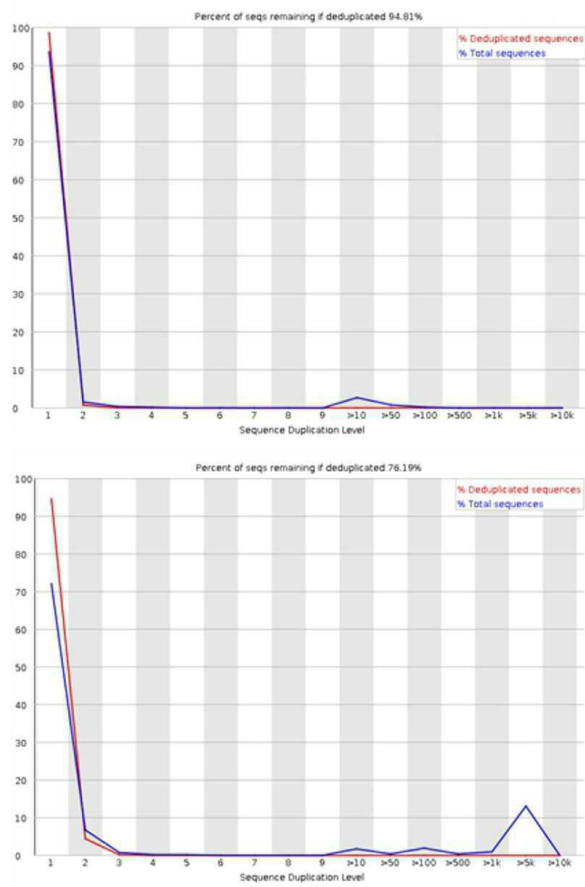
도면20



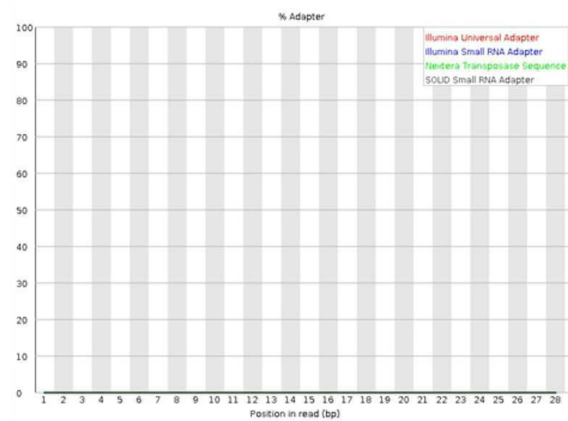
도면21



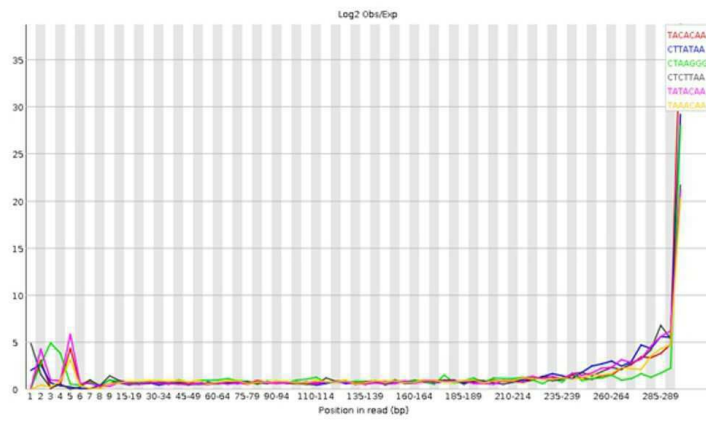
도면22



도면23



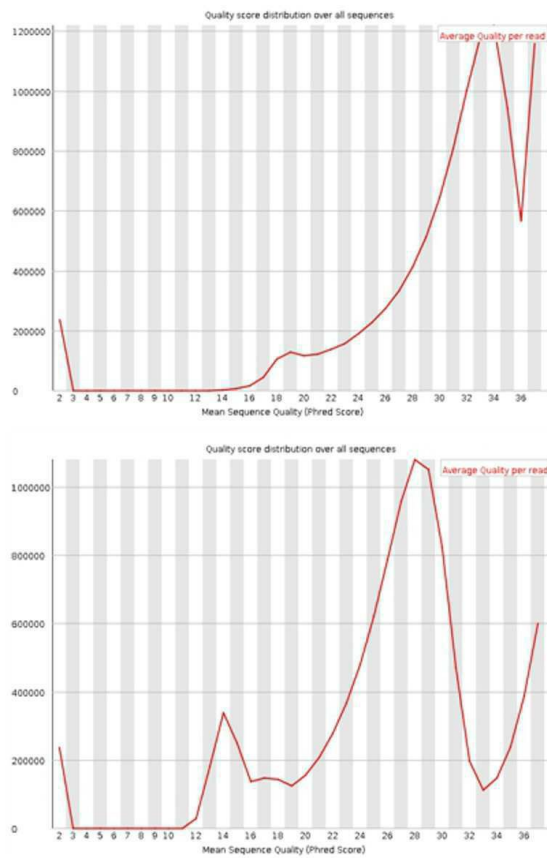
도면24



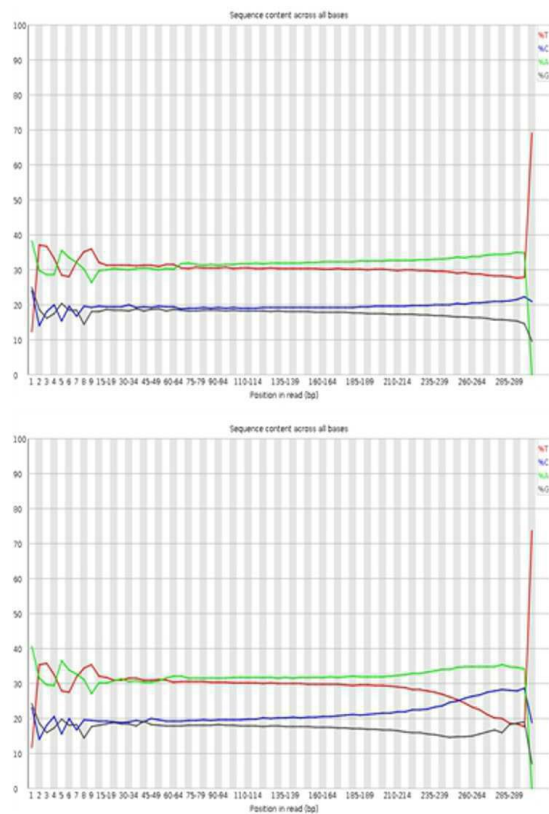
도면25



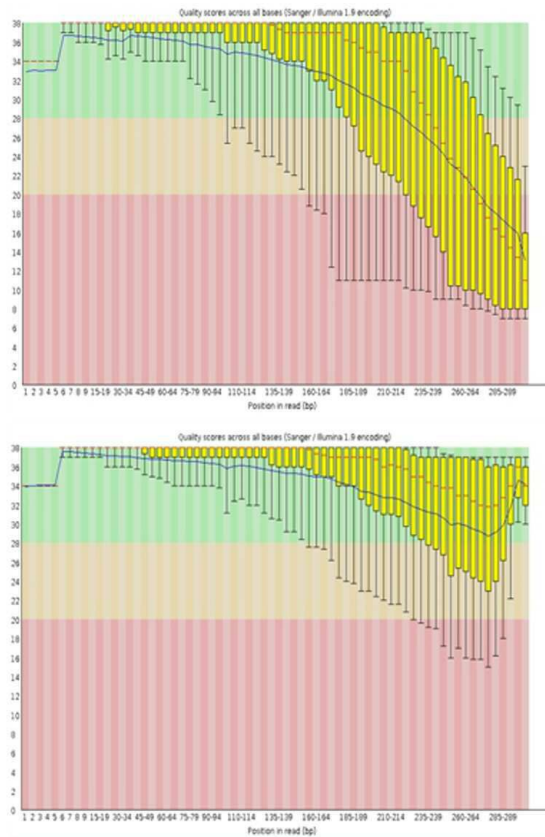
도면26



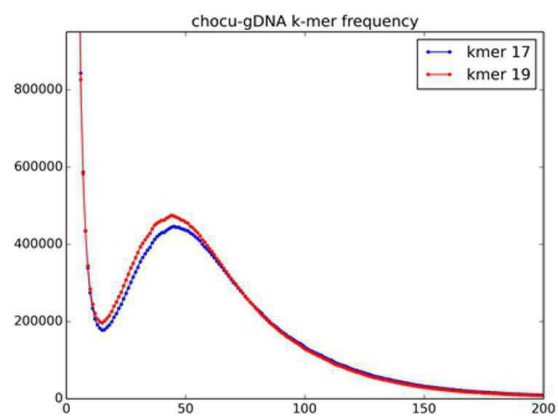
도면27



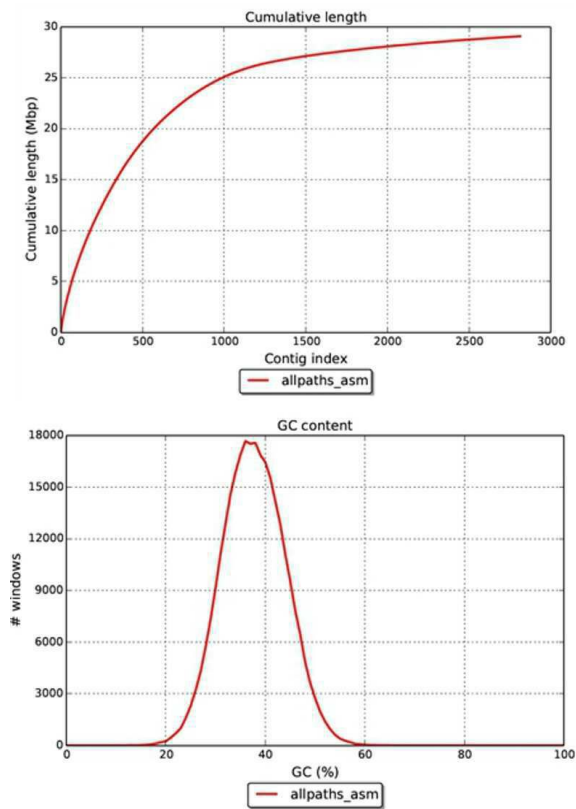
도면28



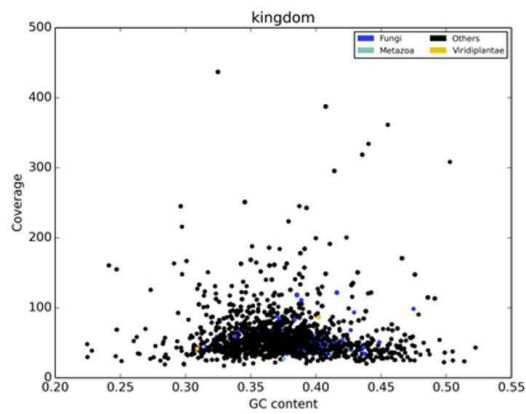
도면29



도면30



도면31



도면32

#### Blast to RDP result

| id | scaffold_name | coverage | identity | matched_bases | phylum                    | species                |
|----|---------------|----------|----------|---------------|---------------------------|------------------------|
| 1  | scaffold_1308 | 2.83%    | 89%      | 81/91         | Zygomycota incertae sedis | Rhizopus stolonifer    |
| 2  | scaffold_1308 | 4.0%     | 83%      | 107/129       | Zygomycota incertae sedis | Cokeromyces recurvatus |
| 3  | scaffold_1308 | 3.72%    | 83%      | 107/129       | Zygomycota incertae sedis | Mucor racemosus        |
| 4  | scaffold_1308 | 3.84%    | 82%      | 106/129       | Archaeosporomycetes       | Geosiphon pyriformis   |
| 5  | scaffold_1520 | 10.12%   | 93%      | 327/351       | Zygomycota incertae sedis | Mucor racemosus        |
| 6  | scaffold_1520 | 26.69%   | 92%      | 295/320       | Zygomycota incertae sedis | Mucor racemosus        |
| 7  | scaffold_1520 | 4.03%    | 97%      | 126/130       | Zygomycota incertae sedis | Cokeromyces recurvatus |
| 8  | scaffold_1520 | 4.01%    | 96%      | 124/129       | Zygomycota incertae sedis | Rhizopus stolonifer    |