(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0001450
(43) 공개일자 2024년01월03일

(51) 국제특허분류(Int. Cl.)
G16B 30/10 (2019.01) *C12Q 1/6869* (2018.01)
G16B 20/00 (2019.01)
(52) CPC특허분류
G16B 30/10 (2019.02)
C12Q 1/6869 (2018.05)
(21) 출원번호 10-2022-0078105
(22) 출원일자 2022년06월27일
심사청구일자 2022년06월27일
기술이전 희망 : 기술양도

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
한림대학교 산학협력단
강원도 춘천시 한림대학길 1, 한림대학교(옥천동)
(72) 발명자
송진원
서울특별시 서초구 서초중앙로 220, 102동 1901호
박경민
서울특별시 동대문구 안암로16길 18, 306호
(뒷면에 계속)
(74) 대리인
위병갑

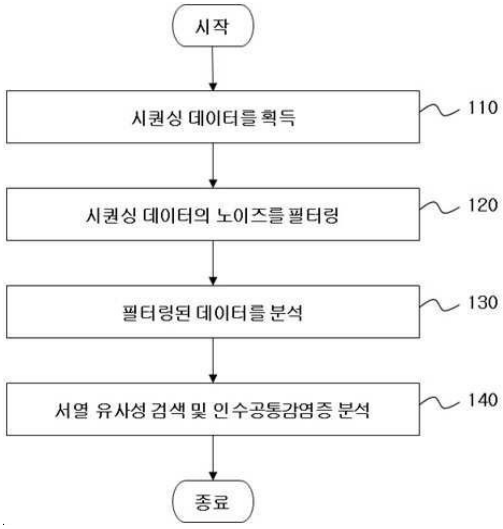
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 **나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석**

(57) 요 약

본 개시의 일 측면으로서, 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법으로서:시퀀싱 데이터를 획득하는 단계;상기 획득된 시퀀싱 데이터에 대해 노이즈 필터링을 수행하는 단계;상기 노이즈 필터링된시퀀싱 데이터에 대해 분석을 수행하는 단계;및 상기 분석된 데이터에 대해 서열 유사성 검색 및 인수공통감염증 분석을 수행하는 단계를 포함하는, 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법이 제공될 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G16B 20/00 (2019.02)
(72) 발명자
김원근
강원도 춘천시 지석로 97, 103동 101호

김기진
부산광역시 금정구 오륜대로 46, B동 201호

이 발명을 지원한 국가연구개발사업
과제고유번호 1711130148
과제번호 2017M3A9E4061992
부처명 한국연구재단
과제관리(전문)기관명 한국연구재단
연구사업명 차세대바이오사회밀착형지원
연구과제명 통합생물정보기반의설치류매개감염병대응국가방역체계연계원천기술확보
기 여 율 50/100
과제수행기관명 고려대학교
연구기간 2017.06.30 ~ 2022.06.29
이 발명을 지원한 국가연구개발사업
과제고유번호 1525012283
과제번호 202104662
부처명 해양수산과학기술진흥원
과제관리(전문)기관명 해양수산과학기술진흥원
연구사업명 해양환경내감염성바이러스빅데이터기반진단, 치료및제어기술개발
연구과제명 해양환경내감염성바이러스빅데이터기반진단, 치료및제어기술개발
기 여 율 25/100
과제수행기관명 한국해양과학기술원
연구기간 2021.04.01 ~ 2026.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1345352346
과제번호 2021R1I1A2049607
부처명 한국연구재단
과제관리(전문)기관명 한국연구재단
연구사업명 보호연구지원사업
연구과제명 신출현인수공통 RNA 바이러스에대한원헬스모니터링및능동감시기술개발
기 여 율 25/100
과제수행기관명 한림대학교
연구기간 2021.06.01 ~ 2025.05.31

명 세 서

청구범위

청구항 1

나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법으로서:

시퀀싱 데이터를 획득하는 단계;

상기 획득된 시퀀싱 데이터에 대해 노이즈 필터링을 수행하는 단계;

상기 노이즈 필터링된시퀀싱 데이터에 대해 분석을 수행하는 단계;및

상기 분석된 데이터에 대해 서열 유사성 검색 및 인수공통감염증 분석을 수행하는 단계를 포함하는,나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법.

청구항 2

제1 항에 있어서,

상기 노이즈 필터링을 수행하는 단계는:

상기 획득된 시퀀싱 데이터에서,미리 결정된 리드 품질 및 리드 길이의 서열,어댑터 서열,및 숙주 유전체 유래 서열을 제거하는 단계를 포함하는,나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법.

청구항 3

제2 항에 있어서,

상기 미리 결정된 리드 품질 및 리드 길이의 서열은 8 이하의 품질 점수를 갖는 서열 및 중복 염기 서열을 포함하는, 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법.

청구항 4

제2 항에 있어서,

상기 어댑터 서열은, 미리 결정된 품질 이하의 리딩(leading) 서열 및 트레일링(trailing)서열을 포함하는, 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법.

청구항 5

제2 항에 있어서,

상기 숙주 유전체 유래 서열을 제거하는 단계는, 리드를 숙주 유전체 서열에 매핑하는 단계를 더 포함하는, 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법.

청구항 6

제1 항에 있어서,

상기 노이즈 필터링된 시퀀싱 데이터에 대해 분석을 수행하는 단계는:

참조 유전체 매핑(reference genome mapping) 기반 분석, 리드 분류군 분류(read taxonomy classification) 기반 분석,및 드노보 어셈블리 (De novo assembly)기반 분석을 수행하는 단계를 포함하는,나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법.

청구항 7

제6 항에 있어서,

참조 유전체 매핑 기반 분석 및 리드 분류군 분류 기반 분석은, 바이러스 서열이 미리 입력된 경우 수행되는,

나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법.

청구항 8

제6 항에 있어서,

드노보 어셈블리 기반 분석을 수행하는 단계는, 파편화된 리드를 재배열하여 복원된 콘티그(contig) 를 생성하는 단계를 포함하는, 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법.

청구항 9

제8 항에 있어서,

상기 서열 유사성 검색 및 상기 인수공통감염증 분석은, 상기 콘티그에 대해 수행되는, 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법.

청구항 10

제8 항에 있어서,

상기 콘티그가 인수공통감염을 유발하는 바이러스 서열인지 여부를 결정하는 단계를 더 포함하는, 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법.

청구항 11

나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 장치로서:

인스트럭션을 저장하는 메모리;및

상기 인스트럭션을 실행하여:

시퀀싱 데이터를 획득하고,

상기 획득된 시퀀싱 데이터에 대해 노이즈 필터링을 수행하고,

상기 노이즈 필터링된시퀀싱 데이터에 대해 분석을 수행하고,

상기 분석된 데이터에 대해 서열 유사성 검색 및 인수공통감염증 분석을 수행하도록 구성된 프로세서를 포함하는, 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬분석 장치.

발명의 설명

기 술 분 야

[0001] 개시된 실시예는 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석에 관하며, 구체적으로, 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석을 위한 전산 플랫폼에 관한 것이다.

배 경 기 술

[0002] 고처리량시퀀싱 (High-Throughput Sequencing)기술은, 시료 내에 존재하는 핵산 염기 서열을 동시다발적으로 서열화하여 대량의 서열 데이터를 획득하기 위해 사용될 수 있다. 고처리량시퀀싱 기술과 관련된 공학 기술의 지속적인 발전을 통해, 획득된 데이터의 정확성과 확장성이 대폭 개선됨에 따라, 고처리량시퀀싱 기술은 바이러스학을 비롯한 생물학 전반 분야에서 사용될 수 있다.

[0003] 대량 서열화 데이터를 분석하기 위해 전산 프로그램이 이용될 수 있다.하지만, 연구자들에게는 전산 프로그램의 사용이 요원한 경우가 많다. 사용하고자 하는 프로그램의 작동 원리에 대한 연구자의 이해가 부족하기 때문이다. 전산학 이론에 대한 교육이 부족한 연구자 입장에서, 최신 전산 이론을 반영한 프로그램의 작동 원리를 이해하기 어려울 수 있다.

[0004] 전산 프로그램의 사용 장벽이 높으므로,전산 프로그램에 충분히 숙련되지 않은 사용자가 프로그램을 손쉽게 사용할 수 있는 방안의 개발이 요구되고 있다.

발명의 내용

해결하려는 과제

[0005] 본 개시는, 고처리량 시퀀싱 데이터로부터 나노포어 시퀀싱을 이용하여 새로운 바이러스를 탐지하고바이롬(Virome)을 분석하기 위한 플랫폼을 제공하고자 한다.

과제의 해결 수단

[0006] 본 개시의 일 측면으로서, 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법으로서:시퀀싱 데이터를 획득하는 단계;상기 획득된 시퀀싱 데이터에 대해 노이즈 필터링을 수행하는 단계;상기 노이즈 필터링된시퀀싱 데이터에 대해 분석을 수행하는 단계;및 상기 분석된 데이터에 대해 서열 유사성 검색 및 인수공통감염증 분석을 수행하는 단계를 포함하는, 나노포어시퀀싱을 이용한 바이러스 탐색 및 바이롬 분석 방법이 제공될 수 있다.

발명의 효과

[0007] 본 개시에 따르면, 고처리량 시퀀싱 데이터로부터 나노포어시퀀싱을 이용하여 새로운 바이러스를 탐지하고 바이롬 (Virome)을 분석하기 위한 플랫폼을 제공할 수 있다.

도면의 간단한 설명

[0008] 도 1은 나노포어 시퀀싱을 이용한 바이러스의 탐지 및 바이롬 분석 방법의 개략적인 흐름도이다.

도 2는 시퀀싱 데이터의 노이즈 필터링 방법의 흐름도이다.

도 3은 시퀀싱 데이터의 분석 방법의 흐름도이다.

도 4는 서열 유사성 검색 및 인수공통감염증 분석을 수행하는 방법의 흐름도이다.

도 5는 노이즈 필터링 후 남은 총 리드 수를 나타내는 그래프이다.

도 6, 7 및 8은 각 참조 유전체 서열을 대상으로 매핑한 결과를 나타내는 그래프들이다.

도 9는 실행 결과를 리드 분류군 분석결과 시각화 프로그램인 Krona를 이용해 표현한 그래프이다.

도 10은 도 9와 같은 실행 결과를 리드 분류군 분석 결과 시각화 프로그램인 MetaComp를 이용해 표현한 히트 맵(Heatmap)이다.

발명을 실시하기 위한 구체적인 내용

[0009] 이하, 첨부된 도면들에 기재된 내용들을 참조하여 본 개시에 따른 예시적실시예를 상세하게 설명한다. 또한, 첨부된 도면들에 기재된 내용들을 참조하여 본 개시의 실시예에 따른 전자 장치를 구성하고 사용하는 방법을 상세히 설명한다. 각 도면에서 제시된 동일한 참조번호 또는 부호는 실질적으로 동일한 기능을 수행하는 부품 또는 구성요소를 나타낸다.

[0010] 제1, 제2 등과 같이 서수를 포함하는 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 구성 요소들은 용어들에 의해 한정되지는 않는다. 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 개시의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 항목들의 조합 또는 복수의 관련된 항목들 중의 어느 하나의 항목을 포함한다.

[0011] 본 개시에서 사용한 용어는 실시예를 설명하기 위해 사용된 것으로, 본 개시를 제한 및/또는 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 개시에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0012] 본 개시에서 어떤 부분이 다른 부분과 연결되어 있다고 할 때, 이는 직접적으로 연결되어 있는 경우뿐 아니라, 그 중간에 다른 요소를 사이에 두고 전기적으로 연결되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구

성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "...부", "모듈", "...기" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.

[0013] 전산 프로그램에 충분히 숙련되지 않은 사용자가 프로그램을 손쉽게 사용할 수 있도록, 분석 파이프라인 프로그램이 이용될 수 있다.분석 파이프라인 프로그램에서 파이프라인이란, 각자 구별되는 기능을 가진 여러 개의 데이터 처리 단계가 연속적으로 일어나도록 구성된 작업 흐름을 가리키며, 분석 파이프라인 프로그램에서 이전 단계의 결과가 다음 단계의 입력으로 사용될 수 있다.분석 파이프라인 프로그램은, 개별적인 목적을 가진 복수의 전산 프로그램들을 분석 목적에 맞게 조합 및 구성한 것으로서, 사용자는 분석 파이프라인 프로그램을 이용하여, 초기 입력에서 최종 결과까지 자동화된 방식으로 분석을 수행할 수 있다.

[0014] 바이러스 연구에서도 분석 파이프라인 프로그램이 이용될 수 있다.예를 들어, 차세대 염기서열 분석 (Next-generation sequencing, NGS) 데이터를 분석하여 시료 내에 존재하는 바이러스의 분류와 그 서열을 알아내는 '바이러스 확인 파이프라인'(Virus identification pipeline) 프로그램이 사용될 수 있다. 파이프라인 프로그램을 사용하면, 사용자가 파이프라인의 구성 프로그램에 대한 자세한 사용 방법 혹은 작동 원리를 모르더라도, 최적화된 분석 결과를 얻을 수 있다는 장점이 있다. 하지만, 개별 구성 프로그램에 대한 어느 정도의 이해를 가진 사용자의 입장에서는 분석의 구성 혹은 매개변수 등을 조정할 수 없다는 것은 단점이다.

[0015] 플랫폼 프로그램은 사용자가 개인의 수요에 맞춰 파이프라인을 구성할 수 있도록 파이프라인 프로그램에 일정 정도의 유연성을 부여한 것이다. 파이프라인을 구성할 프로그램과 핵심 매개변수는 사용자가 선택하되, 분석의 선후 관계에 제약을 주거나 모범 사례 (Best practice)를 제시하여 적절한 분석 구성에 대한 가이드라인을 줄 수 있다. 플랫폼 프로그램을 이용하면, 사용자가 직접 분석 프로그램을 수행할 때 얻을 수 있는 유연성과 파이프라인 프로그램을 통해 얻을 수 있는 편리함 사이의 균형을 이룰 수 있다. 플랫폼 프로그램의 예시로 갤럭시(galaxy)가있다. 갤럭시 상에서는 다양한 분야의 분석 프로그램을 웹 인터페이스 상에서 사용자가 자유롭게 구성하여 파이프라인을 구현할 수 있다. 또한 커뮤니티를 운영하여 사용자-사용자 혹은 사용자-개발자 간의 의견 교환을 장려하고 별도의 교육을 주기적으로 제공함으로써 사용자가 검증된 가이드라인을 따라갈 수 있도록 한다.

[0016] 현재 바이러스를 위한 분석 플랫폼 프로그램의 수는 매우 제한적이다. 바이러스 유전체는 ①세균 유전체를 비롯한 타 미생물 유전체에 비해 평균적으로 그 길이가 짧고, ②시료 내에 존재하는 핵산의 양이 상대적으로 적으며, ③유전체의 진화 속도가 빨라서, 타 유전체와는 다른 분석을 필요로 한다. 따라서 바이러스 연구를 위해서는 바이러스 유전체 분석에 특화된 프로그램이 필요하지만, 현존하는 플랫폼 프로그램은 대체로 그러한 프로그램을 충분히 갖추지 못하고 있다. 또한, Illumina sequencing 으로 대표되는 차세대 염기서열 분석법 (NGS)의 뒤를 이어, 서열화된 파편화 서열, 즉 리드 (read)의 길이가 훨씬 더 길어진 3세대 염기서열 분석법 (Third-generation sequencing) 관련 기술이 부상하고 있고, 3세대 염기서열 분석법에서 생산된 데이터는 차세대 염기서열 분석법에 의해 생성되는 데이터와 특성이 다르므로, 차세대 염기서열 분석에 최적화되어 발전해 온 전산학적 방법론으로는 분석하기에 적합하지 않을 수 있다. 따라서, 3세대 염기서열 분석법에서 생성된 데이터 분석을 위한 새로운 플랫폼의 개발이 필요하다.

[0017] 이에 본 개시에서는 최신 서열 분석 프로그램을 갖춘 바이러스 유전체 서열 발견 및 분석용 통합 플랫폼을 제공하고자 한다. 기존의 차세대 염기서열 분석법에 의해 생성된 데이터뿐만 아니라 3세대 염기서열 분석법에서 생성된 데이터도 분석이 가능케 함으로써 플랫폼의 활용 가능성을 더 높였다. 본 플랫폼은 차세대 염기서열분석법의 대표 플랫폼인 Illumina sequencing, 3세대 염기서열분석법의 대표 플랫폼인 Nanopore sequencing에서 생성된 데이터에 대한 분석을 지원한다. 본 플랫폼을 사용하면 바이러스학 분야에서 대표적으로 사용되고 있는 여러 가지 분석 방법을 동시에 수행하여 단시간에 상호보완적인 분석 결과를 얻을 수 있다. 특히, 사용자 수요에 맞게 파이프라인을 손쉽게 맞춤 구성할 수 있으므로, 특정 바이러스에 국한되지 않고 사용자가 원하는 바이러스 유전체를 발견 및 분석할 수 있다. 따라서,메타지놈 (Metagenome) 분석에도 본 플랫폼이 적용될 수 있다.

[0018] 개시된 실시예에 따르면, 플랫폼은 리눅스 운영체제에서 동작하도록 설계될 수 있으나,이에 제한되지 않으며 다양한 운영체제에서 동작할 수 있다.본 개시에서 플랫폼은 소프트웨어로 제한되지 않으며,플랫폼 프로그램이 구동되는 하드웨어,예를 들어,컴퓨팅 장치를 가리킬 수 있다.

[0019] 개시된 실시예에 따르면,플랫폼은 명령어 인터페이스 (Command Line Interface)로 작동할 수 있으나, 이에 제한되지 않으며 그래픽 인터페이스로 작동하도록 구현될 수 있다. 사용자는 각 분석 단계를 독립적으로 실행할 수 있으며, 여러 단계를 파이프라인으로 구성할 수 있다.

[0020] 개시된 실시예에 따르면, 플랫폼은 프로그래밍 언어인 Python과 Nextflow를 이용하여 구현될 수 있으나, 이에 제한되지 않으며,다양한 프로그래밍 언어를 이용하여 구현될 수 있다. 플랫폼의 각 단계는 상이한 프로그래밍 언어를 이용하여 구현될 수 있다. 예를 들어, 사용자로부터 분석의 구성, 입력 데이터 파일, 기타 옵션값을 입력 받고 각 분석을 수행하는 구성 프로그램에 맞게 매개 변수를 조정하는 1단 wrapper 프로그램은Python으로 구현되고, wrapper로부터 전해진 매개 변수를 이용해 운영체제에 분석 프로그램 실행을 명령하는 2단 wrapper 프로그램은 Nextflow로 구현될 수 있다.

[0021] 개시된 실시예에 따르면, 플랫폼에서는 전반적인 분석의 신뢰도를 높이기 위해 서열화된 리드 중 노이즈로 판단되는 것을 필터링하는 필터링프로그램이 사용될 수 있다.나아가,플랫폼에서는필터링된 리드 집합을 사용자의 목적에 맞게 분석하는 분석 프로그램이 사용될 수 있다. 분석 프로그램의 분석 방법으로서, 참조 유전체 매핑, 리드 분류군 분류법, 드노보 어셈블리 (De novo Assembly)기반 분석법 등이 사용될 수 있으나,이에 제한되지 않는다.각 분석 프로그램은 순서를 가지고 수행될 수 있다. 각 분석 프로그램의 분석 방법의 장점과 단점은 표 1을 참조한다.

표 1

[0022]

	참조 유전체 매핑 기반 분석	리드 분류군 분류기반 분석	드노보 어셈블리 기반 분석
장점	높은 탐지력을 보임. 바이러스 변이체 분석에 유리함	속도가 빠름. 바이러스 군집 수준에서 비교분석 가능	참조 유전체 서열 없이 새로운 바이러스 유전자 서열의 발견이 가능. 다양한 후속 분석.
단점	기존에 알려진 서열만 탐지가 가능.	기존에 알려진 서열만 탐지가 가능. 잘못된 분류 가능성이 상대적으로 높음.	속도가 느림. 상대적으로 더 많은 바이러스 리드가 필요.

[0023] 도 1은 나노포어시퀀싱을 이용한 바이러스의 탐지 및 바이롬 분석 방법의 개략적인 흐름도이다.

[0024] 도 1을 참조하면,동작 110에서 플랫폼은 시퀀싱 데이터를 획득할 수 있다. 시퀀싱 데이터는 사용자에 의해 플랫폼에 입력될 수 있으나, 이에 제한되지 않는다.예를 들어,사용자가 플랫폼에 입력한 조건에 따라, 플랫폼은 시퀀싱 데이터를 수집하거나 획득할 수 있다.

[0025] 시퀀싱 데이터는 나노포어시퀀싱을 통해 획득될 수 있다, 이때, 1D technique을 적용한 cell로 시퀀싱될 수 있다.

[0026] 예를 들어, 사전에 MinKnow 등의 프로그램을 통해 basecalling과 리드 품질 필터링, 어댑터 서열 제거가 적용된 데이터가, 1 개의 fastq 파일로서 입력 데이터를 구성할 수 있다.

[0027] 시퀀싱으로부터 생성된 리드의 평균적인 품질을 확인하는 보고서를 생성하기 위해, 아래와 같은 명령어가 플랫폼에 입력될 수 있으나, 이에 제한되지 않으며,동일한 목적을 달성하기 위한 다른 명령어(다른 프로그램)가 사용되거나, 케이스에 따른 최적화를 위해 동일한 명령어(동일한 프로그램) 사용 시에도 상이한 옵션이 사용될 수 있다.

[0028] $ NanoPlot -fastainput.fq --minlength 200

[0029] 상기 명령어에 의해, Nanoplot 을 이용해 입력 fastq 파일에 대해 보고서가 생성될 수 있다. 이때, 향후 리드 길이 필터링이 이뤄질 것을 고려하여, minlength 옵션을 이용해 최소 길이가 200 이상인 리드만을 사용해 수치를 계산할 수 있다.

[0030] 동작 120에서 플랫폼은 노이즈 필터링을 수행할 수 있고, 이는 도 2를 참조하여 후술할 것이다.

[0031] 동작 130에서 플랫폼은 필터링된 데이터를 분석할 수 있고, 이는 도 3을 참조하여 후술할 것이다.

[0032] 동작 140에서 플랫폼은 분석된 데이터에서 서열 유사성을 검색할 수 있고, 이는 도 4를 참조하여 후술할 것이다.

[0033] 동작 150에서 플랫폼은 분석된 데이터에서 인수공통감염증 분석을 수행할 수 있고, 이는 도 5를 참조하여 후술할 것이다.

[0034] 도 2는 시퀀싱 데이터의 노이즈 필터링 방법의 흐름도이다.

[0035] 도면에서 알파벳을 포함하는 평행사변형은 그 위치에 따라 입력 데이터 및 출력 데이터를 가리키며, 이전 단계의 출력 데이터는 다음 단계의 입력 데이터와 동일할 수 있다.

[0036] 동작 222에서 플랫폼은 리드 품질 및 리드 길이에 기초하여 시퀀싱 데이터를 필터링할 수 있다. 동작 222에서 필터링이 수행되는 시퀀싱 데이터는 도 1의 동작 110에서 획득된 입력 데이터일 수 있다. 예를 들어, 동작 222에서 필터링이 수행되는 시퀀싱 데이터는 input.fq일 수 있다.

[0037] 서열화된 리드는 개별 염기서열에 대해서 서열화 장비 (sequencer)로부터 계산된 품질 점수(quality score) 정보를 갖고 있다. 이 점수를 계산하는 방법은 서열화 플랫폼에 따라 다르지만, 공통적으로 낮은 신뢰도를 가진 염기서열에 대해서 낮은 점수가 부여된다. 따라서 각 리드에 대해 리드의 전체 염기서열에 배정된 품질 점수의 평균을 계산하여, 일정 점수에 미치지 못하는 점수를 가진 리드를 제거함으로써, 후속 분석의 신뢰성을 향상시킬 수 있다.

[0038] 서열이 짧으면 짧을수록 우연성에 의해 생성되었을 가능성이 높으므로, 서열이 긴, 즉, 길이가 긴 리드는 추후 분석에서바이러스를 특정하는데 방해가 될 수 있다. 따라서, 길이가 짧은 리드를 제거함으로써,후속 분석의 신뢰성을 향상시킬 수 있다.

[0039] 예를 들어, 리드 품질 및 리드 길이에 기초하여 입력 데이터를 필터링하기 위해, 아래의 명령어가 플랫폼에 입력될 수 있다.

[0040] $ NanoFilt --length 200 --readtype 1D --quality 8 input.fq>output.fq

[0041] 플랫폼은 NanoFilt을 이용해 평균 품질 8 이하의 리드를 제거하고, 평균 길이 200 이하의 리드를 제거할 수 있다. 필터링된 리드는 새로운 fastq 파일에 저장된다.

[0042] 동작 226에서, 플랫폼은 동작 222에서 필터링된 데이터를 숙주 유전체 유래 리드에 기초하여 필터링할 수 있다.

[0043] 바이러스 유전체를 대상으로 하는 분석에서는 숙주 유전체에서 유래된 리드가 노이즈로 작용한다. 일반적으로, 숙주 유전체에 비해 바이러스 유전체는 시료 내에 극히 소량 존재하므로, 숙주 유전체 유래 리드를 제거함으로써, 상당량의노이즈를 제거함과 동시에 분석 시간을 단축할 수 있다.

[0044] 예를 들어, 숙주 유전체 유래 리드를 필터링 하기 위해 아래 명령어가 플랫폼에 입력될 수 있다.

[0045] $ minimap2 -ax map-onthost.fainput.fq | samtools view -Sb - | samtools sort -o output.bam

[0046] 플랫폼은 minimap2를 이용해 리드를 숙주 유전체 서열에 매핑시킨 후, samtools를 이용해 매핑된 리드만 포함된 bam 파일을 생성할 수 있다. 이때,입력 리드 형태가 ont(oxford nanopore tech) 이고, 매핑을 하는 상황이므로 "map-ont" 옵션이 사용될 수 있다.

[0047] 나아가, samtools의 fastq 명령어를 통해 bam파일로부터 리드를 추출해 fastq 파일로 다시 생성하기 위해, 아래 명령어가 플랫폼에 입력될 수 있다.

[0048] $ samtoolsfastq -f 4 output.bam>output.fq

[0049] 도 3은 시퀀싱 데이터의 분석 방법의 흐름도이다.

[0050] 본 플랫폼은 총 3가지의 후속 분석을 지원할 수 있다.

[0051] 동작 332에서, 플랫폼은 참조 유전체 매핑(reference genome mapping)기반 분석을 수행할 수 있다.

[0052] 분석자가 시료 내에서 찾고자 하는 바이러스 서열을 이미 알고 있는 경우에는 리드 서열을 해당 바이러스 유전체 서열과의 염기서열 유사성을 계산하여 바이러스 유전체 상에 리드를 매핑할 수 있다. 단 한 개의 리드라도 참조 유전체와 유사하다면 탐지해낼 수 있어, 타깃 바이러스 서열이 확보된 상태라면 가장 유리한 분석 방법이다. 만약 해당 바이러스가 실제로 존재하는 경우에는, 다량의 리드가 전체 유전체 위치에 고르게 매핑이 되게 되므로, 바이러스의 존재 여부를 판단할 수 있다. 하지만, 참조 유전체 매핑은, 유전체 서열을 모르는 바이러스에 대해서는 실행할 수 없다는 단점이 있다.

[0053] 예를 들어, minimap2를 이용해 리드를 참조 유전체 서열에 매핑시켜, samtools를 이용해 bam 파일을 생성하기 위해, 아래 명령어가 플랫폼에 입력될 수 있다.

[0054] $ minimap2 -ax map-ontreference.fainput.fq | samtools view -Sb - | samtools sort -o output.bam

[0055] 이때, minimap2 명령어에는 디폴트 옵션이 사용될 수 있다.

[0056] 나아가,BamCov를 이용해 매핑 결과를 요약적인 수치(매핑된 리드 수, 평균 리드 깊이 등)로 출력하기 위해 아래 명령어가 플랫폼에 입력될 수 있다.

[0057] $ bamcov -H output.bam> output.txt

[0058] 이때,bamcov명령어에는 "H" 옵션을 사용하여, 플랫폼은 결과에 헤더를 덧붙여 출력할 수 있다.

[0059] 예를 들어, Qualimap을 이용해 매핑 bam 파일에 대한 보고서를 생성하기 위해, 아래 명령어가 플랫폼에 입력될 수 있다.

[0060] $ qualimapbamqc -bam output.bam -outfile output.pdf

[0061] 동작 334에서, 플랫폼은 리드 분류군 분류 (read taxonomy classification)기반 분석을 수행할 수 있다.

[0062] 메타지놈 분석에 대한 수요가 발생하면서, 리드를 분류군에 직접 분류하는 전산학적 방법론이 개발됐다. 이 분석 방법의 특징은 각 분류군에 분류된 리드의 개수를 통하여, 대략적인 핵산의 양을 추정할 수 있다는 것이다. 또한, 서열 간의 직접 비교가 아닌 서열을 대표하는 값을 비교함으로써 분석 시간을 대폭 줄인 분석 기술이 개발되었다. 리드 뷥류군 분류는, 리드의 길이가 상대적으로 짧으면담고 있는 정보의 양이 적어지고, 따라서 부정확한 분류가 일어날 수 있다는 단점이 있다. 또한 참조 유전체 매핑과 마찬가지로 이미 유전체서열을 알고 있는 바이러스만 탐지해낼 수 있다.

[0063] 예를 들어, 각 리드에 대해 분류군을 예측함으로써, 샘플 내 바이롬구성을 알아내기 위해,아래 명령어가 플랫폼에 입력될 수 있다.

[0064] $ kraken2 --dbViruses_db --report report.csv --confidence 0.1 input.fq

[0065] 상기 명령어에 따르면,플랫폼은 주어진 입력 fastq 파일을 대상으로 Kraken2를 이용해 분류군 분류를 실행한다. 이때, 사전에 생성된 kraken 데이터베이스 "Viruses_db" 가 이용될 수 있다. confidence 옵션에 0.1의 값을 부여하여, 플랫폼은 분류의 신뢰도를 수치화한 confidence 값이 0.1 이상인 결과만을 남길 수 있고, 분류 결과는 report.csv에 저장될 수 있다.

[0066] Kraken2 분류 결과를 시각화하기 위해 결과 파일을 Krakentools의 kreport2krona.py 스크립트로 변환한 뒤, KronaTools의 ktImportText 프로그램으로 시각화한다. 결과가 html 형태로 저장하기 위해,아래 명령어가 플랫폼에 입력될 수 있다.

[0067] $ kreport2krona.py -r report.csv -o input.krona

[0068] $ ktImportTextinput.krona -o output.html

[0069] 동작 336에서 플랫폼은 드노보 어셈블리(De novo assembly)기반 분석을 수행할 수 있다.

[0070] 대량 서열화 방법에서는 동일한 핵산 내 다른 위치에서 복사된 여러 개의 리드가 생성된다. 이러한 리드는 끝단의 서열을 다른 리드와 공유하게 되며, 이를 오버랩(overlap)이라고 한다. 모든 리드를 대상으로 오버랩을 알아낸 뒤, 신뢰할 수 있는 오버랩을 골라 연결된 리드를 연결하면, 파편화되기 이전의 상대적으로 긴 서열을 얻을 수 있다. 이 긴 서열을 콘티그(contig), 전체 서열 조립 과정을 De novo Assembly라고 한다. De novo Assembly를 이용하면 기존에 보고되지 않은 유전체 서열을 발견할 수 있는 장점이 있다. 바이러스 유전체는 진화 속도가 빨라, 기존에 알려진 유전체에 비해 다른 서열로 구성된 유전체가 발견될 확률이 높으므로, 신규 바이러스 유전체를 발견하고자 하는 목적으로 사용된다. 하지만, De novo Assembly는 많은 계산과정을 필요로 하여 분석이 느리고, 전체 파이프라인에서 병목 현상을 일으켜 분석 시간을 늘린다. 또한, 오버랩의 신뢰도를 판단하기 위해서 바이러스 유전체로부터 유래된 리드가 다른 분석 방식에 비해 더 많이 확보되어야 한다.

[0071] 예를 들어, 파편화된 서열인 리드를 재배열하여 파편화 이전의 유전체로 복원하기 위해, 아래 명령어가 플랫폼에 입력될 수 있다.

[0072] $ flye --nano-raw input.fq --out-dir result -meta

[0073] 플랫폼은,Flye를 이용해 입력 fastq를 대상으로 드노보 어셈블리를 실행할 수 있다.이때, 리드 커버리지(coverage)가 낮은 메타지노믹(metagenomics)시퀀싱 상황에 대응해서 만들어진 옵션인 "meta"옵션이 사용될 수

있다 (이른 바,metaFlye). 그 외에는 디폴트 옵션이 사용될 수 있다.

[0074] 이렇게 생성된 콘티그(contig)는 다양한 후속 분석의 대상이 될 수 있고, 이는 도 4를 참조하여 설명한다.

[0075] 도 4는 서열 유사성 검색(sequence homology search) 및 인수공통감염증 분석(zoonotic analysis)을 수행하는 방법의 흐름도이다.

[0076] 동작 442에서 플랫폼은 서열 유사성 검색을 수행하여,콘티그의 서열을 기존에 알려져 있는 바이러스 유전체 서열과 비교하여, 유사도가 높은 결과를 토대로 시료 내에 존재할 것이라고 예상되는 바이러스의 분류를 추측할 수 있다.

[0077] 예를 들어, 해당 분석을 수행하는 대표적인 프로그램으로 BLAST가 있다. 드노보 어셈블리로 생성된 콘티그서열들의 분류군을 알아내기 위해, 서열 유사성 검색을 실행하기 위해 아래 명령어가 플랫폼에 입력될 수 있다.

[0078] $ blastn -query input.fa -dbViruses_db -task blastn -evalue 1.0e-5 -max_target_seqs 1 -outfmt "6 qseqidsseqidstaxidsstitlepidentqlen length mismatch gapopenqstartqendsstart send evaluebitscore"

[0079] 이때, Blast의 명령어 프로그램인 Blast+의 blastn 프로그램이 이용될 수 있다. 콘티그서열이 포함된 입력 파일에"query" 옵션이 사용될 수 있다. 이때, 사전에 생성된 Blast 데이터베이스"Viruses_db" 가 사용될 수 있다. 얼마나 정확한 비교를 수행하는 알고리즘을 이용할지 정하는 "task"옵션에는 "blastn"값이 사용될 수 있다. 일정 값보다 낮은 유사도를 가진 결과를 제거하는 "evalue" 옵션에는 1.0e-5 값이 사용될 수 있다. 각 콘티그에 대해 가장 유사한 바이러스 서열 1개만 표현하기 위해 "max_target_seqs"옵션에"1"이 사용될 수 있다. 각 비교 결과에 대해 어떤 수치를 표현할지 지정하는 "outfmt" 옵션에는 "6"을 사용하고, 표현할 수치를 입력할 수 있다.

[0080] 동작 444에서 플랫폼은 인수공통감염증 분석을 수행하여,알려진 인수공통감염 바이러스의 서열상 특징을 분석해, 새롭게 만들어진 콘티그 서열이 해당 특징을 갖고 있는지 여부를 판단하면 콘티그서열의인수공통감염성 잠재성을 판단할 수 있다.

[0081] 드노보 어셈블리로 생성된 콘티그 서열들이 인수공통감염을 유발하는 바이러스 서열인지 알아내기 위해, zoonotic rank 프로그램이 사용될 수 있다.

[0082] 먼저, Zoonotic rank 에 이용될 metadata 파일을 생성하기 위해 Prodigal 을 이용해 서열내 유전자를 예측할 수 있고, 예를 들어, 아래 명령어가 플랫폼에 입력될 수 있다.

[0083] $ prodigal -iinput.fa -f sco -o result.sco -p meta

[0084] 이때,유전자 좌표만을 알아내기 위해 "f" 옵션에 "sco" 값을 사용할 수 있다. 유전자 예측시 사전에 계산된 트레이닝 파일을 사용하기 위해 "p" 옵션에 "meta" 값을 사용할 수 있다.이로써,유전자 예측 결과가 좌표와 함께 "o" 옵션에 사용된 result.sco파일이저장될 수 있다.

[0085] prodigal_sco_to_zoonotic_rank_metadata.py 스크립트를 이용해 Prodigal을 이용한 유전자 예측 결과 파일을 zoonotic rank에서 요구하는 metadata 형식으로 전환하기 위해,아래 명령어가 플랫폼에 입력될 수 있다.

[0086] $ python prodigal_sco_to_zoonotic_rank_metadata.py --input result.sco --output metadata.txt

[0087] 나아가, Zoonotic rank 의 PredictNovel.R 스크립트를 통해 각 콘티그가 인수공통감염 바이러스 서열일 가능성을 수치화하고 점수에 따라 결과를 분류한 결과 파일을 생성하기 위해, 아래 명령어가 입력될 수 있다.

[0088] $ RscriptPredictNovel.Rfastainput.fa metadata.txt result

[0089] 도 1 내지 도 4의 각 동작 단계에서 사용되는 프로그램은 표 2를 참조한다.

표 2

[0090]

리드 품질 관리	NanoPlot
리드 품질 및 리드 길이 필터링	NanoFilt
어댑터 서열 제거	Guppy
숙주 유전체 유래 리드 필터링	Minimap2
참조 유전체 매핑	Minimap2, Bamcov, Qualimap, Samtools, Bedtools
리드 분류군 분류법	Kraken2, Kaiju, Krona, KronaTools, MetaComp
De novo Assembly	Megahit, Canu, Racon, Medaka, Flye

서열 유사성 검색	BLAST, Diamond, taxonomizr
인수공통감염증 분석	Prodigal, Zoonotic rank

[0092] <실시예>

[0093] 예제로 사용한 데이터는 신증후군출혈열의 원인병원체 중 하나인 한타바이러스과(Hantavridae family)에 속하는 한탄바이러스(Hantaanorthohantavirus, HTNV)에 감염된 환자로부터 시료를 추출해, MinION서열화 플랫폼으로 서열화한Nanopore 리드 데이터이다. 플랫폼 상의 분석 프로그램을 사용하여 예제 데이터를 분석한 결과는 다음과 같다.

[0095] 1) 노이즈 필터링 예시 결과

[0096] 해당 예시 데이터는 염기서열호출(Basecalling) 이후 어댑터 서열 제거 단계가 이미 수행되었으므로, 본 예시 분석에서는 해당 단계가 제외되었다. 리드품질 및 리드 길이 필터링 단계에서는 리드 품질 8 미만, 리드 길이 200 이하의 리드가 제거되었으며, 숙주 유전체 유래 리드필터링 단계에서는 인간 유전체 서열 GRCh38.p13에 매핑된 리드가 제거되었다. 도 5는 예시 데이터의 각 노이즈 필터링 세부 단계 이후 남은 총 리드 수를 나타낸다.

[0098] 2) 참조 유전체 매핑 예시 결과

[0099] 표 3은 예시 데이터를 한탄바이러스 76-118주의 각 분절 서열을 참조 유전체로 삼아 참조 유전체 매핑한 결과를 나타낸다. 도 6, 7, 및 8은 각 참조 유전체 서열을 대상으로 매핑한 결과를 도시하며, 사용자는 이를 통해 참조 유전체 상 위치에 따라 매핑 결과를 파악할 수 있다.

[0100] 도 6, 7, 및 8은 예시 데이터의 참조 유전체 매핑 결과로서, 각각 한탄바이러스 76-118 주 L분절, M분절, 및 S 분절을 나타낸다.

표 3

[0101]

참조 유전체 이름	참조 유전체 서열 길이 (A)	매핑된 리드 수	커버된참조유전체 서열 길이 (B)*	커버된 비율 (A / B x 100)	평균 리드 깊이**	평균 리드 서열 품질	평균 리드 매핑 품질
HTNV 76-118 L 분절	6,533	34,819	6,533	100	2,942.62	20.4	58.8
HTNV 76-118 M 분절	3,616	9,129	3,614	99.9447	1,529.76	20.3	58
HTNV 76-118 S 분절	1,696	14,307	1,696	100	4,224.72	19.9	57.4

[0102] *최소한 1개 이상의 리드가 참조유전체 서열 상에 매핑된 경우 커버되었다고 정의한다.

[0103] ** 리드 깊이는 참조 유전체 상 각 좌표에 매핑된 리드의 개수로 정의한다.

[0105] 3) 리드 분류군 분류법 예시 결과

[0106] 예시 데이터를 대상으로 리드 분류군 분석 프로그램인 Kraken2를 kraken2-viral데이터베이스를 이용해 실행하였다. 도 9는 실행 결과를 리드 분류군 분석결과 시각화 프로그램인 Krona를 이용해 표현한 것이며,도 10은 같은 결과를 리드 분류군 분석 결과 시각화 프로그램인 MetaComp를 이용해 히트 맵(Heatmap)으로 표현한 것이다.

[0108] 4) De novo Assembly 기반 분석법 예시 결과

[0109] (1) 서열 유사성 검색 예시 결과

[0110] 표 4는 예시 데이터를 대상으로 드노보 어셈블리를 실행하여 생성된 콘티그들을 대상으로 서열 유사성 검색 분석 프로그램인 BLAST를 실행한 결과의 일부를 나타낸 것이다.

표 4

[0111]

superkingdom	Viruses
order	Bunyavirales
family	Hantaviridae
genus	Orthohantavirus
species	Hantaanorthohantavirus
Accession ID	ref|NC_005222.1|
Contig ID	tig847
EVALUE	0
BITSCORE	876

[0113] (2) 인수공통감염증 분석 결과

[0114] 표 5는 예시 데이터를 대상으로 드노보 어셈블리를 실행하여 생성된 콘티그들을 대상으로 인수공통감염증 분석 프로그램인 Zoonotic rank를 실행한 결과의 일부를 나타낸 것이다.

표 5

[0115]

Name	priority_category
tig847	Medium

[0116] 소프트웨어 모듈 또는 알고리즘으로 구현되는 방법들은 프로세서상에서 실행 가능한 컴퓨터가 읽을 수 있는 코드들 또는 프로그램 명령들로서 컴퓨터가 읽을 수 있는 기록 매체 상에 저장될 수 있다. 여기서 컴퓨터가 읽을 수 있는 기록 매체로 마그네틱 저장 매체(예컨대, ROM(read-only memory), RAM(random-access memory), 플로피 디스크, 하드 디스크 등) 및 광학적 판독 매체(예컨대, 시디롬(CD-ROM), 디브이디(DVD: Digital Versatile Disc)) 등이 있다. 컴퓨터가 읽을 수 있는 기록 매체는 네트워크로 연결된 컴퓨터 시스템들에 분산되어, 분산 방식으로 컴퓨터가 판독 가능한 코드가 저장되고 실행될 수 있다. 매체는 컴퓨터에 의해 판독가능하며, 메모리에 저장되고, 프로세서에서 실행될 수 있다.

[0117] 본 개시에서 인용하는 공개 문헌, 특허 출원, 특허 등을 포함하는 모든 문헌들은 각 인용 문헌이 개별적으로 및 구체적으로 병합하여 나타내는 것 또는 본 개시에서 전체적으로 병합하여 나타낸 것과 동일하게 본 개시에 병합될 수 있다.

[0118] 본 개시의 이해를 위하여, 도면에 도시된 바람직한 실시 예들에서 참조 부호를 기재하였으며, 본 개시의 실시 예들을 설명하기 위하여 특정 용어들을 사용하였으나, 특정 용어에 의해 본 개시가 한정되는 것은 아니며, 본 개시는 당업자에 있어서 통상적으로 생각할 수 있는 모든 구성 요소들을 포함할 수 있다.

[0119] 본 개시는 기능적인 블록 구성들 및 다양한 처리 단계들로 나타내어질 수 있다. 이러한 기능 블록들은 특정 기능들을 실행하는 다양한 개수의 하드웨어 또는/및 소프트웨어 구성들로 구현될 수 있다. 예를 들어, 본 개시는 하나 이상의 마이크로프로세서들의 제어 또는 다른 제어 장치들에 의해서 다양한 기능들을 실행할 수 있는, 메모리, 프로세싱, 로직(logic), 룩업 테이블(look-up table) 등과 같은 직접 회로 구성들을 채용할 수 있다.

[0120] 본 개시에의 구성 요소들이 소프트웨어 프로그래밍 또는 소프트웨어 요소들로 실행될 수 있는 것과 유사하게, 본 개시는 데이터 구조, 프로세스들, 루틴들 또는 다른 프로그래밍 구성들의 조합으로 구현되는 다양한 알고리즘을 포함하여, C, C++, 자바(Java), 어셈블러(assembler) 등과 같은 프로그래밍 또는 스크립팅 언어로 구현될 수 있다. 기능적인 측면들은 하나 이상의 프로세서들에서 실행되는 알고리즘으로 구현될 수 있다. 또한, 본 개시는 전자적인 환경 설정, 신호 처리, 및/또는 데이터 처리 등을 위하여 종래 기술을 채용할 수 있다. "매커니

즘", "요소", "수단", "구성"과 같은 용어는 넓게 사용될 수 있으며, 기계적이고 물리적인 구성들로서 한정되는 것은 아니다. 상기 용어는 프로세서 등과 연계하여 소프트웨어의 일련의 처리들(routines)의 의미를 포함할 수 있다.

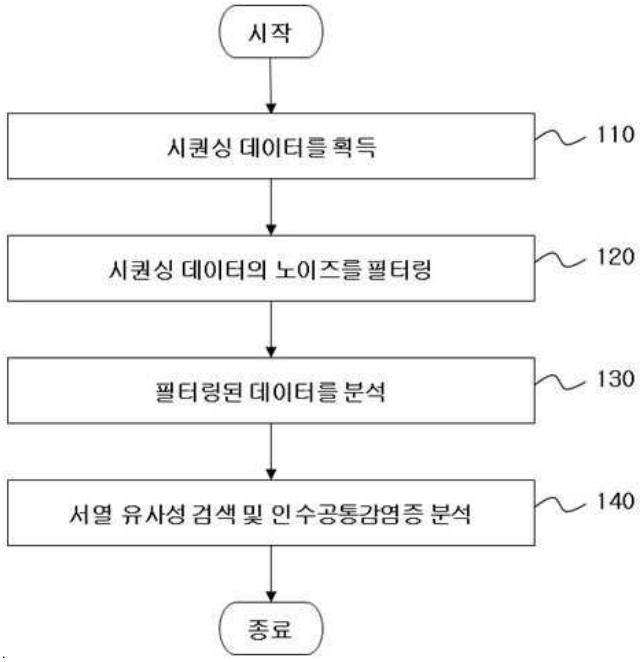
[0121] 본 개시에서 설명하는 특정 실행들은 일 실시 예들로서, 어떠한 방법으로도 본 개시의 범위를 한정하는 것은 아니다. 명세서의 간결함을 위하여, 종래 전자적인 구성들, 제어 시스템들, 소프트웨어, 상기 시스템들의 다른 기능적인 측면들의 기재는 생략될 수 있다. 또한, 도면에 도시된 구성 요소들 간의 선들의 연결 또는 연결 부재들은 기능적인 연결 및/또는 물리적 또는 회로적 연결들을 예시적으로 나타낸 것으로서, 실제 장치에서는 대체 가능하거나 추가의 다양한 기능적인 연결, 물리적인 연결, 또는 회로 연결들로서 나타내어질 수 있다. 또한, "필수적인", "중요하게" 등과 같이 구체적인 언급이 없다면 본 개시의 적용을 위하여 반드시 필요한 구성 요소가 아닐 수 있다.

[0122] 본 개시의 명세서(특히 특허청구범위에서)에서 "상기"의 용어 및 이와 유사한 지시 용어의 사용은 단수 및 복수 모두에 해당하는 것일 수 있다. 또한, 본 개시에서 범위(range)를 기재한 경우 상기 범위에 속하는 개별적인 값을 적용한 발명을 포함하는 것으로서(이에 반하는 기재가 없다면), 발명의 상세한 설명에 상기 범위를 구성하는 각 개별적인 값을 기재한 것과 같다. 마지막으로, 본 개시에 따른 방법을 구성하는 단계들에 대하여 명백하게 순서를 기재하거나 반하는 기재가 없다면, 상기 단계들은 적당한 순서로 행해질 수 있다. 반드시 상기 단계들의 기재 순서에 따라 본 개시가 한정되는 것은 아니다. 본 개시에서 모든 예들 또는 예시적인 용어(예들 들어, 등등)의 사용은 단순히 본 개시를 상세히 설명하기 위한 것으로서 특허청구범위에 의해 한정되지 않는 이상 상기 예들 또는 예시적인 용어로 인해 본 개시의 범위가 한정되는 것은 아니다. 또한, 당업자는 다양한 수정, 조합 및 변경이 부가된 특허청구범위 또는 그 균등물의 범주 내에서 설계 조건 및 팩터에 따라 구성될 수 있음을 알 수 있다.

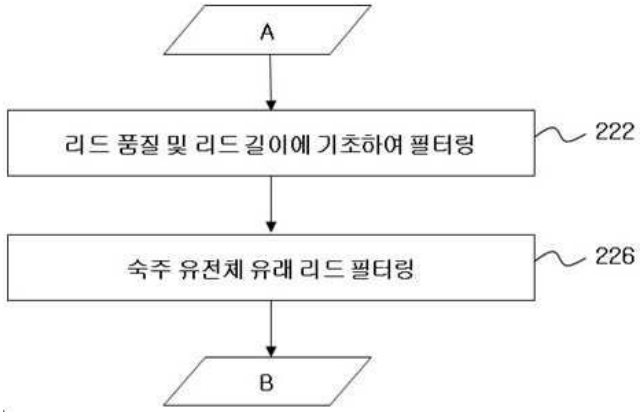
[0123] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(operating system: OS) 및 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 어플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

도면

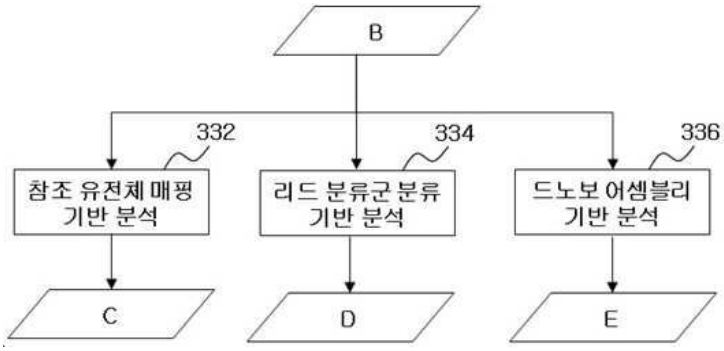
도면1



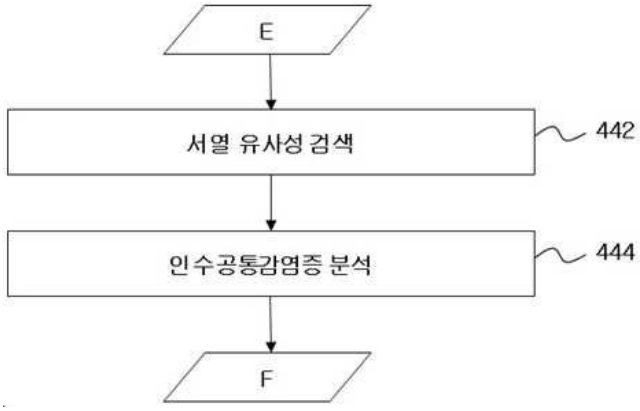
도면2



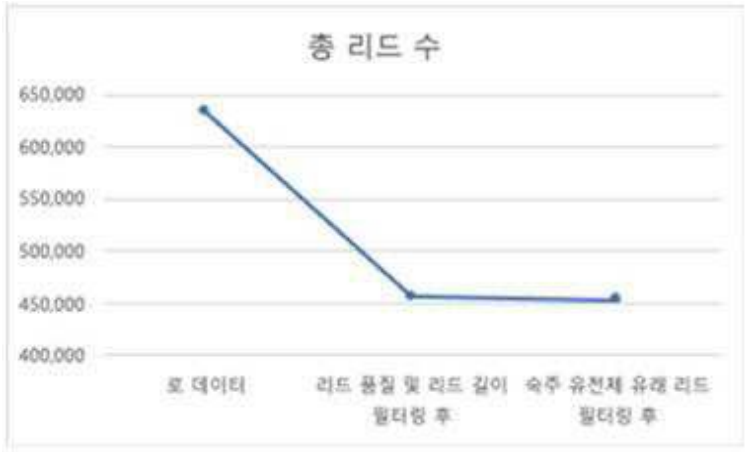
도면3



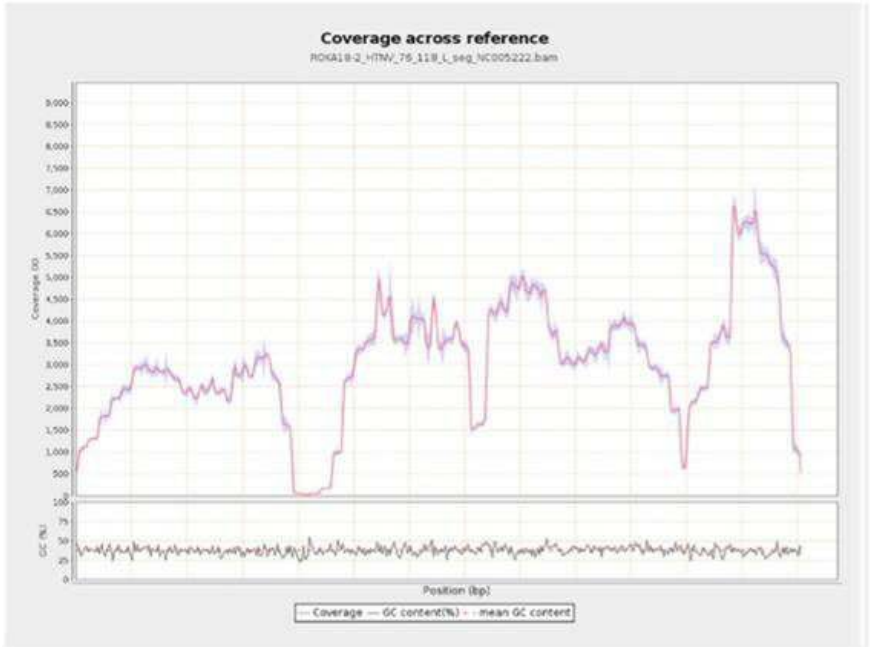
도면4


E
서열 유사성 검색
442
인수공통감염증 분석
444
F

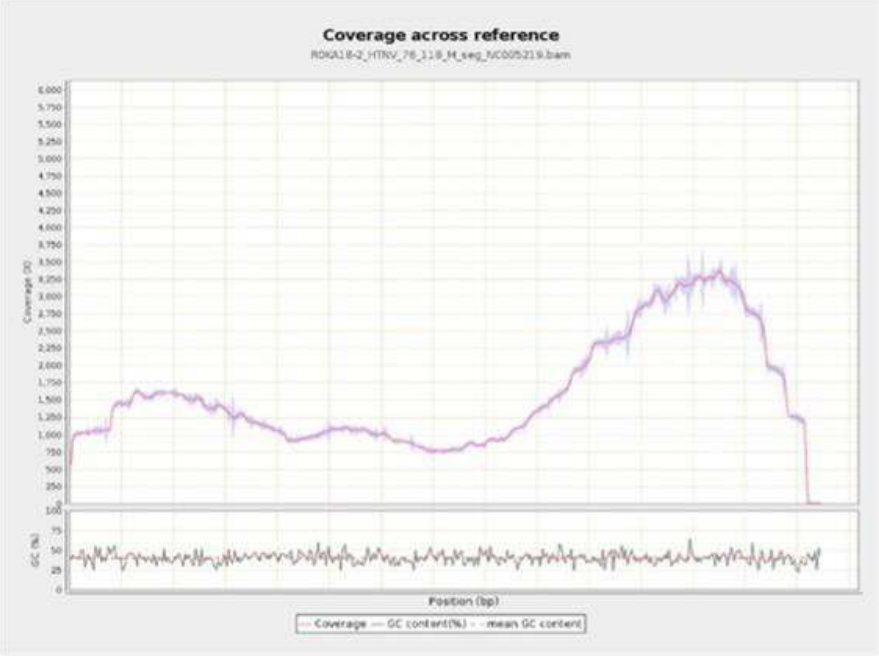
도면5


총 리드 수
650,000
600,000
550,000
500,000
450,000
400,000
로 데이터
리드 품질 및 리드 길이 필터링 후
숙주 유전체 유래 리드 필터링 후

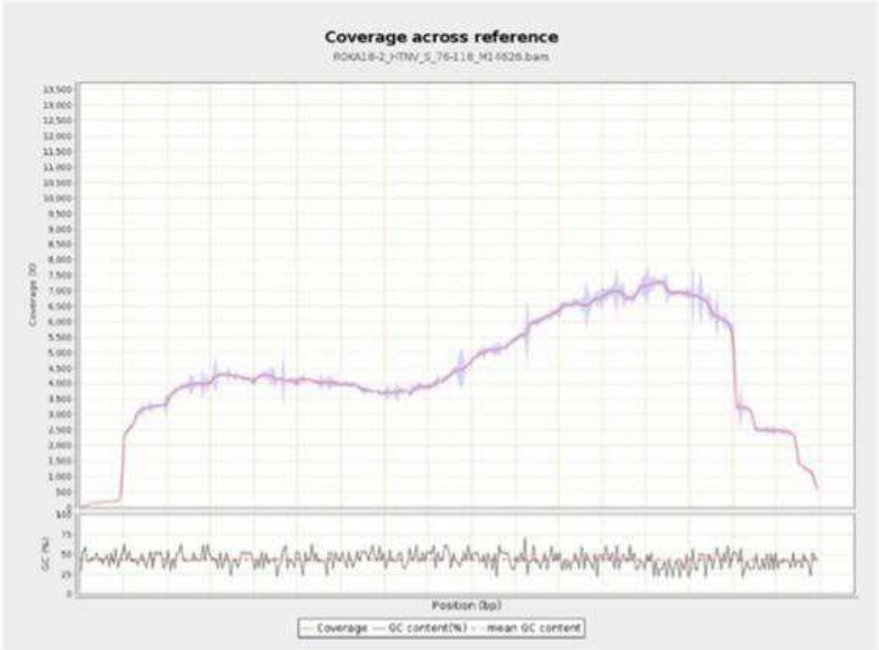
도면6


Coverage across reference
Position (bp)
Coverage
GC content(%)
mean GC content

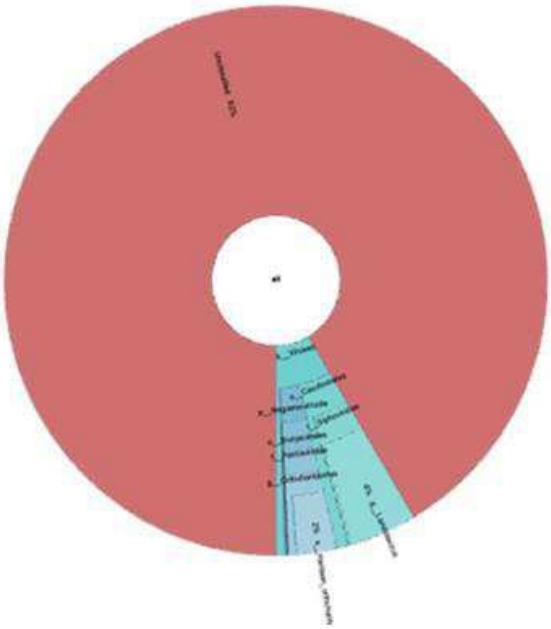
도면7


Coverage across reference
Position (bp)
Coverage
GC content(%)
mean GC content

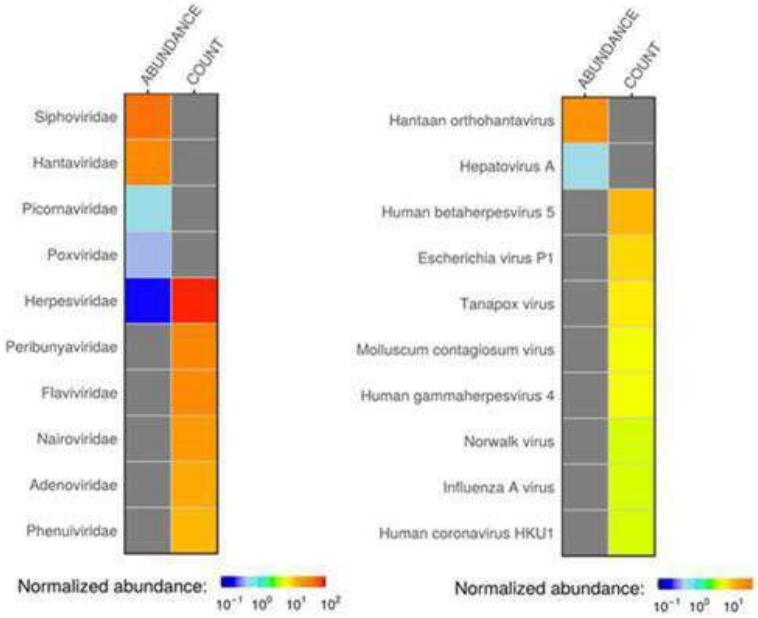
도면8


Coverage across reference
Position (bp)
Coverage
GC content(%)
mean GC content

도면9



도면10


ABUNDANCE
COUNT
Siphoviridae
Hantaviridae
Picornaviridae
Poxviridae
Herpesviridae
Peribunyaviridae
Flaviviridae
Nairoviridae
Adenoviridae
Phenuiviridae
Normalized abundance:
10^{-1} 10^{0} 10^{1} 10^{2}
ABUNDANCE
COUNT
Hantaan orthohantavirus
Hepatovirus A
Human betaherpesvirus 5
Escherichia virus P1
Tanapox virus
Molluscum contagiosum virus
Human gammaherpesvirus 4
Norwalk virus
Influenza A virus
Human coronavirus HKU1
Normalized abundance:
10^{-1} 10^{0} 10^{1}