(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년10월23일
(11) 등록번호 10-2721363
(24) 등록일자 2024년10월21일

(51) 국제특허분류(Int. Cl.)
G16B 30/00 (2019.01) *C12Q 1/68* (2018.01)
(52) CPC특허분류
G16B 30/00 (2019.02)
C12Q 1/6874 (2018.05)
(21) 출원번호 10-2017-7026167
(22) 출원일자(국제) 2016년02월17일
심사청구일자 2021년02월17일
(85) 번역문제출일자 2017년09월15일
(65) 공개번호 10-2017-0134379
(43) 공개일자 2017년12월06일
(86) 국제출원번호 PCT/US2016/018295
(87) 국제공개번호 WO 2016/134034
국제공개일자 2016년08월25일
(30) 우선권주장
62/117,256 2015년02월17일 미국(US)
62/294,208 2016년02월11일 미국(US)
(56) 선행기술조사문헌
US20140220587 A1*
A. Dayarian 외, "SOPRA: Scaffolding algorithm for paired reads via statistical optimization", BMC Bioinformatics, 11:345. (2010.06.24.)*
US20120197533 A1
M. Boetzer 외, "Scaffolding pre-assembled contigs using SSPACE", Bioinformatics, 27(4):578-579. (2010.12.12.)
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
더브테일 제노믹스 엘엘씨
미국 95066 캘리포니아주 스코츠 밸리 엔터프라이즈 웨이 100 스위트 에이101
(72) 발명자
퍼트넘 니콜라스 에이치
미국 02453 매사추세츠주 월섬 리버 스트리트 36 아파트먼트 362
스타이츠 조나단 씨
미국 95060 캘리포니아주 산타 크루즈 실바 스트리트 121 #3
라이스 브랜든 제이
미국 95060 캘리포니아주 산타 크루즈 엘름 스트리트 214 아파트먼트 1
(74) 대리인
김진희, 김태홍

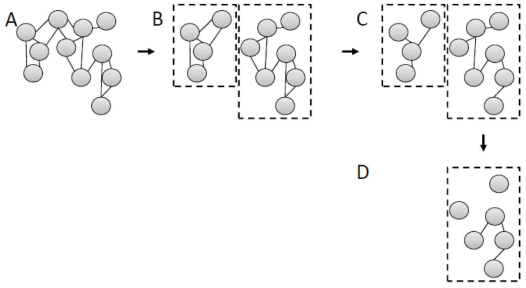
전체 청구항 수 : 총 14 항 심사관 : 김채연

(54) 발명의 명칭 **핵산 서열 어셈블리**

(57) 요 약

본원에서는, 리드 쌍 서열 정보, 예컨대 핵산 서열 상 또는 물리적 연관을 나타내는 리드 쌍 정보의 사용을 통한, 서열 어셈블리, 예컨대 단일 리드 및 콘티그의 더 큰 콘티그 및 스캐폴드로의 핵산 서열 어셈블리와 관련된 조성물, 시스템 및 방법이 개시된다.

대 표 도 - 도10



명 세 서

청구범위

청구항 1

핵산 서열 정보의 콘티그를 스캐폴딩하기 위한 컴퓨터 실행 시스템으로서,

초기 구성을 갖는 콘티그 서열의 세트를 수신하고;

분해된 샘플 핵산의 무작위 재결찰 접합부의 서열을 포함하는 페어드-엔드 리드(paired-end read)의 세트를 수신하고;

표준 페어드-엔드 리드 거리 빈도 데이터(standard paired-end read distance frequency data)를 수신하고;

하나 이상의 페어드-엔드 리드 내에 공존하는 서열을 공유하는 콘티그 쌍이 그룹화되도록 콘티그 쌍을 처리하고;

상기 그룹화된 콘티그 서열을, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드-쌍 거리 빈도 데이터가 초기 구성에서의 콘티그 서열의 리드-쌍 거리 빈도 데이터에 비해 표준 페어드-엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하도록, 스캐폴딩하고;

상기 처리된 콘티그 스캐폴드를 네트워크, 스크린, 또는 서버로 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템.

청구항 2

제1항에 있어서, 표준 페어드-엔드 리드 빈도는, 양자 리드가 공통 콘티그에 매핑되는 페어드-엔드 리드로부터 수득되는 것인 컴퓨터 실행 시스템.

청구항 3

제1항에 있어서, 표준 페어드-엔드 리드 빈도는 사전에 생성된 곡선으로부터 수득되는 것인 컴퓨터 실행 시스템.

청구항 4

제1항에 있어서, 초기 구성은 무작위 구성인 컴퓨터 실행 시스템.

청구항 5

제1항에 있어서, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드-쌍 거리 빈도 데이터는, 리드-쌍 거리 우도(likelihood)가 증가하는 경우에, 표준 페어드-엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하는 것인 컴퓨터 실행 시스템.

청구항 6

제5항에 있어서, 리드-쌍 거리 우도는 최대화되는 것인 컴퓨터 실행 시스템.

청구항 7

제1항에 있어서, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드-쌍 거리 빈도 데이터는, 상기 리드-쌍 거리 빈도 데이터와 표준 페어드-엔드 리드 거리 빈도 데이터 간의 차이에 대한 통계 측정값이 감소하는 경우에, 표준 페어드-엔드 리드 거리 빈도 데이터에 보다 더 가깝게 근접하는 것인 컴퓨터 실행 시스템.

청구항 8

제7항에 있어서, 상기 리드-쌍 거리 빈도 데이터와 표준 페어드-엔드 리드 거리 빈도 데이터 간의 거리에 대한 통계 측정값은 ANOVA, t-검정 및 X-제곱 검정 중 하나 이상을 포함하는 것인 컴퓨터 실행 시스템.

청구항 9

제1항에 있어서, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드-쌍 거리는, 표준 페어드-엔드 리드 거리 빈도와 비교하여 수득되는 순서 배열된 콘티그들 간의 리드-쌍 거리 분포의 편차가 감소하는 경우에, 표준 페어드-엔드 리드 거리 빈도 데이터에 보다 더 가깝게 매칭되는 것인 컴퓨터 실행 시스템.

청구항 10

제9항에 있어서, 표준 페어드-엔드 리드 거리 빈도와 비교하여 수득되는 순서 배열된 콘티그들 간의 리드-쌍 거리 분포의 편차는 최소화되는 것인 컴퓨터 실행 시스템.

청구항 11

제1항에 있어서, 핵산은 DNA이고 상기 DNA는 하나 이상의 DNA 결합제에 가교결합되는 것인 컴퓨터 실행 시스템.

청구항 12

제1항에 있어서, 핵산은 단리된 네이키드 DNA인 컴퓨터 실행 시스템.

청구항 13

제12항에 있어서, 단리된 DNA는 재구성된 염색질 내에 리어셈블링되는(reassembled) 것인 컴퓨터 실행 시스템.

청구항 14

제13항에 있어서, 재구성된 염색질은 가교결합되는 것인 컴퓨터 실행 시스템.

청구항 15

삭제

청구항 16

삭제

청구항 17

삭제

청구항 18

삭제

청구항 19

삭제

청구항 20

삭제

청구항 21

삭제

청구항 22

삭제

청구항 23

삭제

청구항 24

삭제

청구항 25

삭제

청구항 26

삭제

발명의 설명

기 술 분 야

[0001] **상호 참조**

[0002] 본 출원은 2015년 2월 17일에 출원된 미국 가특허 출원 제62/117,256호(이 출원은 명시적으로 그 내용 전체가 본원에 참고로 포함됨)을 우선권으로 주장하며, 본 출원은 또한 2016년 2월 11일에 출원된 미국 가특허 출원 제62/294,208호(이 출원은 명시적으로 그 내용 전체가 본원에 참고로 포함됨)도 우선권으로 주장한다.

[0003] 본 발명은 핵산 서열 어셈블리에 관한 조성물, 시스템 및 방법에 관한 것이다.

배 경 기 술

[0004] 현재 적정가격으로 이용가능한 고처리량 시퀀싱 방법은 단거리 서열 근접성 및 게놈 변이의 특성에 최적화되어 있다. 원거리 연관 및 하플로타입 페이징 (haplotype phasing)의 달성은 긴 (예컨대, 수십 킬로베이스) 서열을 직접적이고 정확하게 판독하는 능력, 또는 한 쌍의 리드 또는 그룹화된 서열 리드를 통한 연관 및 상 관계의 포착을 필요로 한다. 그러나, 원거리 연관 및 하플로타입 페이징을 달성하기 위해 필요한 시퀀싱 정보의 그룹화 및 서열 정보의 어셈블리를 생성하는 것은 계산 집약적이며 시간이 많이 걸린다. 본원에서는 한 쌍의 리드 또는 그룹화된 서열 리드에 의해 제공된 서열 정보로부터 염색체 규모의 근접성을 갖는 어셈블리를 수득하기 위한 계산 효율적인 방법 및 시스템이 개시된다.

발명의 내용

[0005] 본원에서는 핵산 데이터의 스캐폴딩에 관한 방법, 조성물, 알고리즘 및 시스템이 개시된다. 본원의 접근법은 데이터세트 내의 리드 쌍의 리드가 매핑되는 콘티그의 상 또는 물리적 연관 정보에 관한 정보를 추론하기 위해 리드 쌍을 이용한다. 핵산 데이터세트 내의 콘티그들은 순서 배열되거나, 배향되거나 또는 (한쪽 끝과 다른 한쪽 끝을 이어) 한줄로 병합되거나, 또는 일부의 경우, 그들의 상대적인 위치와 관련된 스코어 또는 파라미터에 대한 그러한 활성의 영향의 관점에서 보면 서로 삽입된다 (집합적으로 "스캐폴딩된다").

[0006] 일부의 경우, 상기 스코어 또는 파라미터는 하나 또는 다른 콘티그 구성의 리드 쌍 데이터세트에 있어서 전체 리드 쌍 분리에 대한 콘티그 재배치의 결과적인 영향의 척도이다. 리드 쌍의 데이터세트를 생성하는데 사용되는 접근법에 따라, 상기 데이터세트는 특정한 리드 쌍 분리 분포 곡선을 가질 수 있다. 빈도의 함수로서 리드 쌍 분리를 매핑하게 되면, 당업자라면 누구나 주어진 리드 쌍 데이터세트에 대해 예측된 리드 쌍 거리 분포를 측정할 수 있다. 이후, 당업자는 데이터 세트 매치에 대한 리드 쌍 거리 분포가 핵산 샘플 및 리드 쌍의 생성 방법을 고려하여 예측된 리드 쌍 거리 분포에 근접하거나 또는 보다 더 가깝게 근접하도록 하기 위하여, 상기 리드 쌍을 콘티그의 세트에 매핑하여 상기 콘티그를 (순서 배열하거나, 배향시키거나 또는 다르게) 위치시킬 수 있다.

[0007] 콘티그 위치결정은 콘티그 또는 스캐폴드를 서로에 대해 순서 배열시키는 단계, 콘티그 또는 스캐폴드를 서로에 대해 배향시키는 단계, 콘티그 또는 스캐폴드를 한줄로 접합시키는 단계, 콘티그 또는 콘티그의 스캐폴드 내의 갭에 하나 이상의 콘티그를 삽입하는 단계, 또는 데이터세트 내에 잘못 어셈블링된 콘티그 또는 스캐폴드를 분리시키는 단계를 다양하게 수반한다. 일부의 경우, 이 공정은 최적의 구성 또는 최적화된 구성이 얻어질 때까지 지속되는 한편, 다르게는 본 공정은 초기의 콘티그 또는 스캐폴드 구성에 대한 개선을 달성하기 위해서만 수행된다. 다르게는, 본 공정은 샘플 콘티그 세트의 일부 분획, 예를 들어 70%, 75%, 80%, 85%, 90%, 95%, 99% 또는 99.9% 또는 그 이상의 분획이 올바르게 스캐폴딩될 때까지 지속된다. 인간 또는 다배체 식물 샘플 게놈 또는

트랜스포존이 풍부한 게놈 샘플과 같은 훨씬 복잡한 게놈 샘플을 대표하는 서열 데이터세트의 대부분의 경우, 데이터세트 구성의 계산 평가, 및 콘티그를 순서 배열하거나, 배향하거나, 한줄로 결합시키거나, 하나의 스캐폴드를 또 다른 스캐폴드에 결합시키거나, 또는 스캐폴드 또는 콘티그를 분해하는 것 (집합적으로 "스캐폴딩")에 의한 데이터세트 개선은 8시간 이하, 7시간, 6시간, 5시간, 4시간 또는 4시간 미만 내에 완료된다.

[0008] 스코어 평가는 인접한 콘티그 또는 스캐폴드의 하위세트를 한번에 검사함으로써 전반적으로 또는 국소적으로, 또는 전반적이면서 국소적으로 이루어진다. 국소적으로 수행되는 경우에는, 예를 들어, 2, 3, 4, 5, 6개 또는 6개 이상의 하위세트를 검사하여 최적화된 스코어를 결정한 후, '윈도우'를 한 콘티그만큼 이동시키고 보통은 이전의 윈도우에 대해 결정된 최적화 구성을 감안하여 본 공정을 반복한다. 다르게는, 상기 하위세트는 한번에 0.01%, 0.1%, 1% 또는 5%와 같이 전체 핵산 서열 세트 (예컨대, 하나의 게놈 또는 복수의 게놈) 중 일부로 정의된다. 일부의 경우, '윈도우' 크기는 쉽게 어셈블링된 영역들은 큰 윈도우로 지정되는 한편, 보다 어려운 영역들, 또는 보다 큰 밀도의 리드 또는 상반되고 분석을 어렵게하는 보다 큰 밀도의 리드를 갖는 영역들은 더 작은 윈도위 크기로 지정되는 것과 같이 다양할 것이다.

[0009] 본원에서는, 초기 구성을 갖는 콘티그 서열의 세트를 수득하는 단계; 페어드 엔드 (paired end) 리드의 세트를 수득하는 단계; 표준 페어드 엔드 리드 거리 빈도 데이터(standard paired-end read distance frequency data)를 수득하는 단계; 하나 이상의 페어드 엔드 리드 내에 공존하는 서열을 공유하는 콘티그 쌍을 그룹화하는 단계; 및 상기 그룹화된 콘티그 서열을, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드 쌍 거리 빈도 데이터가 초기 구성에서 콘티그 서열의 리드 쌍 빈도 데이터에 대한 표준 페어드 엔드 리드 거리 빈도 데이터에 보다 근접하도록 스캐폴딩하는 단계를 포함하는, 핵산 서열 정보의 콘티그를 스캐폴딩하는 방법이 제공된다. 상기 스캐폴딩은 콘티그의 세트를 순서 배열하는 단계, 콘티그의 세트를 배향시키는 단계, 2종 이상의 콘티그를 한줄로 병합시키는 단계, 하나의 콘티그를 제2 콘티그에 삽입하는 단계 및 하나의 콘티그를 2개 이상의 성분 콘티그로 절단하는 단계 중 하나 이상의 단계를 포함한다. 일부 방법에서, 표준 페어드 엔드 리드 빈도는 양자 리드가 모두 공통 콘티그로 매핑되는 페어드 엔드 리드로부터 수득된다. 다르게는 또는 조합하여, 표준 페어드 엔드 리드 빈도는 사전에 생성된 곡선으로부터 수득된다. 초기의 구성은 무작위 구성이거나, 또는 사전에 구성된다. 바람직한 실시양태에서, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드 쌍 거리 빈도 데이터는 리드 쌍 거리 우도가 증가하는 경우에 페어드 엔드 리드 거리 빈도 데이터에 보다 더 가깝게 근접하게 된다. 대부분의 경우, 리드 쌍 거리 우도는 최대화된다. 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드 쌍 거리 빈도 데이터는 리드 쌍 거리 빈도 데이터와 표준 페어드 엔드 리드 거리 빈도 데이터 간의 차이에 대한 통계 측정값이 감소하는 경우에 페어드 엔드 리드 거리 빈도 데이터에 보다 더 가깝게 근접하게 된다. 다수의 통계 측정값들을 이용할 수 있다. 예를 들어, 리드 쌍 거리 빈도 데이터와 표준 페어드 엔드 리드 거리 빈도 데이터 간의 거리의 통계적 측정값은 다양한 경우에서 ANOVA, t-검정 및 X-제곱 검정 중 하나 이상을 포함한다. 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드 쌍 거리는 표준 페어드 엔드 리드 거리 빈도와 비교될 때, 수득되는 순서 배열된 콘티그들 간의 리드 쌍 거리 분포의 편차가 감소하는 경우에 페어드 엔드 리드 거리 빈도 데이터에 보다 더 가깝게 매칭된다. 다르게는 또는 조합하여, 표준 페어드 엔드 리드 거리 빈도와 비교될 때, 수득되는 순서 배열된 콘티그들 간의 리드 쌍 거리 분포의 편차는 최소화된다. 일부 스캐폴드 평가에서, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그는 아주 많은 수의 공유된 엔드 리드를 갖는 클러스터로 지정된다. 클러스터링은 흔히 유기체 내 염색체의 수보다 많거나 염색체 수와 같은 다수의 그룹에 콘티그를 배치하는 단계를 포함한다. 보통, 클러스터 중 하나의 콘티그와 단 하나의 페어드 엔드 리드만을 공유하는 콘티그는 해당 클러스터에 포함되지 않는다. 반복 서열을 포함하는 하나 이상의 페어드 엔드 리드만을 클러스터와 공유하는 콘티그는 보통 해당 클러스터에 포함되지 않는다. 이와 유사하게, 저품질의 서열을 포함하는 하나 이상의 페어드 엔드 리드만을 클러스터와 공유하는 콘티그는 보통 해당 클러스터에 포함되지 않는다. 일부 방법에서, 페어드 엔드 리드의 세트는 샘플 DNA를 분해하여 핵산 내에 내부 이중 가닥 절단부를 생성하는 단계, 상기 이중 가닥 절단부를 재결찰시켜 하나 이상의 재결찰 접합부를 형성하는 단계, 및 상기 하나 이상의 재결찰 접합부에 대해 시퀀싱하는 단계에 의해 수득된다. DNA는 하나 이상의 DNA 결합제, 예컨대 핵 단백질 또는 나노입자에 가교결합되고, 일부 접근법에서는 페어드 리드 생성에 가교결합된다. 결합 단백질을 갖는 DNA는 일부 환경 하에, 특히 DNA 분자가 서로 결합되지 않는 경우에 적절하지만, 상기 DNA는 재구축되는 염색질로 리어셈블링되는 단리된 네이키드 DNA이다. 보통, 재구성된 염색질은 가교결합된다. 상기 재구성된 염색질은 DNA 결합 단백질을 포함한다. 다르게는 또는 조합하여, 상기 재구성된 염색질은 나노입자를 포함한다. 바람직하게는 일부의 경우, 콘티그의 클러스터링은 유기체에 대한 염색체의 수와는 독립적이다. 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그는 대부분의 경우 아주 많은 수의 공유된 엔드 리드를 갖는 클러스터로 지정된다. 다르게는 또는 조합하여, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티

그는 더 큰 리드 쌍 거리 우도 값을 갖는 클러스터로 지정되거나, 또는 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그는 표준 리드 쌍 거리 분포에 대하여 그의 리드 쌍 분포에서 더 낮은 편차를 갖는 클러스터로 지정된다. 다르게는, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그는 각 클러스터로부터 배제된다. 보통, 클러스터링은 유기체 내 염색체의 수보다 많거나 염색체 수와 같은 다수의 그룹에 콘티그를 배치하는 단계를 포함한다. 일부 스캐폴딩은 상기 클러스터된 콘티그 중 인접한 것으로 추정되는 콘티그의 제1 세트를 선택하는 단계, 리드 쌍에 대한 상기 리드 쌍 거리의 총 측정값을 감소시키는, 인접한 것으로 추정되는 콘티그의 상기 제1 세트의 최소 거리 순서를 결정하는 단계, 및 상기 리드 쌍 거리의 상기 총 측정값을 감소시키기 위해 인접한 것으로 추정되는 콘티그의 상기 제1 세트를 스캐폴딩하는 단계를 포함한다. 인접한 것으로 추정되는 콘티그의 제1 세트는 2개의 콘티그로 이루어진다. 다르게는, 인접한 것으로 추정되는 콘티그의 제1 세트는 3개의 콘티그로 이루어진다. 다르게는, 인접한 것으로 추정되는 콘티그의 제1 세트는 4개의 콘티그로 이루어진다. 다르게는, 인접한 것으로 추정되는 콘티그의 제1 세트는 4개의 콘티그를 포함한다. 일부 스캐폴딩은 인접한 것으로 추정되는 콘티그의 상기 제1 세트 중 각 콘티그의 순서와 배향을 결정하는 단계를 포함한다. 최소 거리 순서의 결정은 일부의 경우, 상기 세트 중 2개의 콘티그에 매핑되는 리드를 포함하는 하나 이상의 리드 쌍에 대해 예측된 리드 쌍 거리를 가능한 모든 콘티그 구성에 대하여 비교하는 단계를 포함한다. 스캐폴딩은 보통 상기 리드 쌍 거리에 대한 최소 리드 쌍에 상응하는 콘티그 배향을 선택하는 단계를 포함한다. 일부 방법은 최대 우도 리드 쌍 거리 분포에 상응하는 콘티그 배향을 선택하는 단계를 더 포함한다. 일부 방법은 상기 콘티그 클러스터의 리드 쌍의 총 측정값에 대한 최소 리드 쌍 거리에 상응하는 콘티그 배향을 선택하는 단계를 더 포함한다. 일부 방법에서, 예측된 리드 쌍 거리는 상기 페어드 엔드 리드 거리 빈도 데이터와 비교된다. 일부 방법에서, 상기 페어드 엔드 리드 거리 빈도 데이터에 대한 상기 비교는 식 1의 이용을 포함한다. 일부 방법은 상기 클러스터된 콘티그 중 인접한 것으로 추정되는 콘티그의 제2 세트를 선택하는 단계로서, 상기 제2 세트가 상기 제1 세트 중 하나의 엔드 말단 콘티그를 제외하고 모두 포함하고, 상기 클러스터된 콘티그 중 하나의 추가 콘티그를 포함하는 것인 단계, 및 리드 쌍 거리의 상기 총 측정값을 감소시키기 위해 인접한 것으로 추정되는 콘티그의 상기 제2 세트를 스캐폴딩하는 단계를 포함한다. 일부 방법은 상기 클러스터된 콘티그 중 인접한 것으로 추정되는 콘티그의 제3 세트를 선택하는 단계로서, 상기 제3 세트가 상기 제2 세트 중 하나의 엔드 말단 콘티그를 제외하고 모두 포함하고, 상기 제1 세트에 포함되지 않고 상기 제2 세트에 포함되지 않는 상기 클러스터된 콘티그 중 하나의 추가 콘티그를 포함하는 단계, 및 리드 쌍 거리의 상기 총 측정값을 감소시키기 위해 인접한 것으로 추정되는 콘티그의 상기 제3 세트를 스캐폴딩하는 단계를 포함한다. 이는대부분의 경우 대다수의 상기 클러스터된 콘티그가 순서 배열될 때까지 하나 이상의 추가의 세트를 반복적으로 선택하는 단계가 뒤따른다. 선택은 보통 각각의 상기 클러스터된 콘티그가 순서 배열될 때까지 하나 이상의 추가의 세트를 반복적으로 선택하는 단계를 수반한다. 핵산 서열은 하나의 게놈, 또는 일부의 경우, 복수의 게놈과 같은 샘플로부터 유래된다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0010] 본원에서는 각 콘티그 쌍에 대한 로그 우도비 (log-likelihood ratio) 스코어를 지정하는 단계; 비율 스코어에 의해 관련성을 분류하는 단계; 및 어셈블리의 전체 스코어를 증가시키기 위해 비율 스코어의 내림차순으로 콘티그 관련성을 수용 또는 거부하는 단계를 포함하는, 클러스터 내의 콘티그를 스캐폴딩하는 방법이 제공된다. 일부 방법에서, 상기 스캐폴딩은 콘티그의 세트를 순서 배열하는 단계, 및/또는 콘티그의 세트를 배향시키는 단계, 및/또는 2종 이상의 콘티그를 한줄로 병합시키는 단계, 및/또는 하나의 콘티그를 제2 콘티그에 삽입하는 단계, 및/또는 하나의 콘티그를 2개 이상의 성분 콘티그로 절단하는 단계를 포함한다. 대부분의 경우, 상기 콘티그는 하나의 게놈, 또는 복수의 게놈을 포함한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0011] 본원에서는 클러스터 내의 복수의 콘티그에 대해 국소적으로 최적의 콘티그 구성을 결정하는 방법이 제공된다. 일부 이러한 방법은 a) 콘티그의 클러스터를 따라 i 위치에서 출발하는 크기 w 콘티그의 서열 윈도우를 확인하는 단계; b) 윈도우에서 각 i 위치에서 양립가능한 순서와 배향에 대한 스코어를 검사함으로써 윈도우 w 의 콘티그에 대한 $w!\ 2^w$ 순서 및 배향 옵션을 고려하는 단계; c) 상기 윈도우 내 상기 w 콘티그를 배향하고 순서 배열하여 최적의 스코어를 수득하는 단계; d) 상기 윈도우를 i+1 위치로 이동시키는 단계; 및 e) 상기 w 콘티그의 배향 및 순서 배열을 이용하여 i+1 위치에서 상기 윈도우에 대하여 단계 (a), (b) 및 (c)를 반복하여 최적의 스

코어를 결정함으로써, 상기 복수의 콘티그를 스코어에 대해 국소적으로 최적 구성으로 배향하고 순서 배열하는 단계를 포함한다. 일부 방법에서, 클러스터 내의 복수의 콘티그에 매핑된 리드 쌍 데이터가 수득되고, 표준 페어드 엔드 리드 빈도 데이터 세트가 수득되며, 상기 w 콘티그의 배향 및 순서 배열에 대한 스코어는 클러스터 내의 복수의 콘티그에 매핑된 리드 쌍 거리 데이터에 대한 리드 쌍 거리 데이터 세트가 표준 페어드 엔드 리드 빈도 데이터 세트에 얼마나 가깝게 매칭되는지에 대한 측정값이다. 일부 방법에서, 클러스터 내의 복수의 콘티그에 매핑된 리드 쌍 데이터가 수득되고, 상기 스코어는 전체 리드 쌍 거리이며, 상기 스코어는 전체 리드 쌍 거리가 최소화될 때 최적화된다. 상기 윈도우 크기 w는 3, 또는 다르게는 w는 4, 또는 다르게는 w는 5, 또는 다르게는 w는 6이다. 일부의 경우, w는 제1 클러스터에 대한 제1값을 가지며, w는 제2 클러스터에서 제2값을 가진다. w는 일부 방법에서, 상기 세트 중 1%의 콘티그, 또는 다르게는 상기 세트 중 5%의 콘티그, 또는 다르게는 상기 세트 중 10%의 콘티그를 포함하도록 선택된다. 대부분의 방법에서, 상기 스코어는 리드 쌍 거리 우도 스코어이고, 상기 스코어는 주어진 윈도우 크기에 대해 최대화될 때 최적화된다. 상기 스코어는 일부 예시적인 실시양태에서 식 1을 이용하여 계산된다. 상기 스코어는 예측된 리드 쌍 분포로부터의 편차이고, 일부의 경우, 주어진 윈도우 크기에 대해 최소화될 때 최적화된다. 상기 복수의 콘티그는 하나의 게놈, 또는 복수의 게놈, 또는 비게놈 핵산 공급원을 포함한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0012] 본원에서는 정제된 DNA를 수득하는 단계; 상기 정제된 DNA를 DNA 결합제와 결합시켜 DNA/염색질 복합체를 형성하는 단계; 상기 DNA-염색질 복합체를 제한 효소와 함께 항온배양시켜 점착성 말단을 남기는 단계; 결찰을 수행하여 DNA의 말단을 결합시키는 단계; 결찰된 DNA 접합부에 대해 시퀀싱하여 페어드 엔드 리드를 생성하는 단계; 및 상기 페어드 엔드 리드를 사용하여 상기 정제된 DNA의 서열을 나타내는 콘티그를 포함하는 핵산 데이터 세트를 스캐폴드하는 단계를 포함하는, 핵산 서열 어셈블리 방법이 제공된다. 일부 방법에서, 상기 정제된 DNA는 하나의 게놈, 또는 복수의 게놈으로부터 유래한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0013] 본원에서는 핵산 샘플에 대한 콘티그 데이터세트를 수득하는 단계; 물리적으로 연결된 비인접성 서열 정보에 상응하는 하나 이상의 리드 쌍 서열 리드를 수득하는 단계; 및 상기 리드 쌍 서열 리드 중 하나 이상의 리드가 하나의 콘티그 데이터 세트 중 2개의 별개 위치에 매핑되는 경우, 해당 리드 쌍 서열 리드를 배제시키는 단계를 포함하는, 반복성 콘티그 영역에 매핑되는 리드 쌍 서열 리드를 확인하는 방법이 제공된다. 일부 방법에서, 상기 반복성 영역은 제1 역치를 초과하는 샷건 리드 깊이 (shotgun read depth)를 갖는 서열을 포함한다. 일부 방법에서, 상기 반복성 영역은 제2 역치를 초과하는 리드 깊이를 갖는 염기 위치를 포함한다. 보통, 제1 역치 및 제2 역치는 리드 깊이의 전체적인 분포에 대하여 고정된다. 제1 역치는 대부분의 경우 리드 깊이의 전체적인 분포의 3배이다. 다르게는, 제1 역치는 2, 2.5, 3.5, 4, 4.5, 5, 5.5, 6이거나, 또는 이 세트 이내이거나 이 세트에 인접한 정수가 아닌 수치이다. 제2 역치는 보통 리드 깊이의 전체적인 분포의 3.5배이다. 다르게는, 제2 역치는 2, 2.5, 3, 4, 4.5, 5, 5.5, 6이거나, 또는 이 세트 이내이거나 이 세트에 인접한 정수가 아닌 수치이다. 일부 방법에서, 상기 정제된 DNA는 하나의 게놈, 또는 복수의 게놈으로부터 유래한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0014] 본원에서는 제1 콘티그와 제2 콘티그 간에 걸쳐있는 리드 쌍 서열의 개수 및 정황상 추정되는 분리를 관찰할 확률을 측정하는 단계로서, 상기 콘티그가 세트 [++,+-,-+,--] 내에서 o의 상대적 배향을 가지며 갭 깊이에 의해 분리되는 것인 단계를 포함하는, 콘티그 어셈블리 결정을 안내하는 방법이 제공된다. 일부 방법은 분리 거리에 대해 리드 쌍 서열 분포의 확률을 정규화하는 단계로서, 상기 정규화가 상기 리드 쌍 서열을, 핵산 샘플을 독립적으로 표본으로 삼는 노이즈 쌍과 비교하는 단계를 포함하는 것인 단계를 더 포함한다. 일부의 경우, 상기 핵산 샘플은 하나의 게놈을 포함한다. 다르게는, 상기 핵산 샘플은 복수의 게놈, 또는 비게놈 공급원을 포함한다. 보통, 노이즈 쌍의 총수는 콘티그 쌍의 샘플에 대한 링크의 밀도를 도표화함으로써 측정된다. 본원에서는 최고 밀도값 1%와 최저 밀도값 1%가 배제되는 방법이 추가로 제공된다. 이에 대한 대안에서, 최고값 0.5%, 0.6%, 0.7%, 0.8%, 0.9%, 1.1%, 1.2%, 1.3%, 1.4%, 1.5%, 1.6%, 1.7%, 1.8%, 1.9%, 2%, 3%, 4%, 5% 또는 5% 초과가 배제되며, 이와 유사하게 최저값 0.5%, 0.6%, 0.7%, 0.8%, 0.9%, 1.1%, 1.2%, 1.3%, 1.4%, 1.5%, 1.6%, 1.7%, 1.8%, 1.9%, 2%, 3%, 4%, 5% 또는 5% 초과가 배제된다. 일부 방법은 콘티그 순서를 결정하는 단계를 포함한다.

일부 방법은 콘티그 배향을 결정하는 단계를 포함한다. 일부 방법은 콘티그 순서와 콘티그 배향 모두를 결정하는 단계를 포함한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0015] 본원에서는 초기 구성을 갖는 콘티그 서열의 세트를 수득하는 단계; 페어드 엔드 리드의 세트를 수득하는 단계; 표준 페어드 엔드 리드 거리 빈도 데이터를 수득하는 단계; 하나 이상의 페어드 엔드 리드 내에 공존하는 서열을 공유하는 콘티그 쌍을 그룹화하는 단계; 상기 콘티그의 그룹에 대한 리드 쌍 빈도 데이터를 표준 페어드 엔드 리드 거리 빈도 데이터와 비교하는 단계; 상기 그룹 중 하나의 콘티그에 절단부를 도입하는 것이 상기 콘티그의 그룹에 대한 리드 쌍 빈도 데이터를 표준 페어드 엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하도록 하는지의 여부를 측정하는 단계; 및 상기 콘티그의 그룹에 대한 리드 쌍 빈도 데이터가 표준 페어드 엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하는 경우에, 상기 콘티그에 절단부를 도입하는 단계를 포함하는, 콘티그 부정합 (misjoin) 보정 방법이 제공된다. 일부 방법에서, 제1 위치는 절단부를 도입하기 전의 상기 역치 미만의 상기 로그 우도를 갖는 하나 이상의 인접한 제2 위치와 병합된다. 인접한 제2 위치는 제1 위치로부터 300개 이하의 염기 쌍만큼 떨어져 있다. 다르게는, 제2 위치는 제1 위치로부터 1000개 초과의 염기 쌍만큼 떨어진 위치를 포함하지 않는다. 다르게는, 인접한 제2 위치는 50, 100, 150, 200, 250, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800 1900 또는 2000개 이하, 또는 언급한 수치들에 걸친 범위 내의 정수값만큼 떨어져 있다. 본원에서는 로그 우도 변화를 측정하는 단계가 하나의 콘티그에 대한 평균 페어드 엔드 매핑 밀도를 확인하는 단계, 평균 페어드 엔드 매핑 밀도의 3x 이상의 페어드 엔드 매핑 밀도를 갖는 콘티그의 세그먼트를 확인하는 단계, 및 평균 페어드 엔드 매핑 밀도의 3x 이상의 페어드 엔드 매핑 밀도를 갖는 콘티그의 세그먼트를 배제시키는 단계를 포함하는 방법이 추가로 제공된다. 다르게는, 1.5x, 1.6x, 1.7x, 1.8x, 1.9x, 2.0x, 2.1x, 2.2x, 2.3x, 2.4x, 2.5x, 2.6x, 2.7x, 2.8x, 2.9x, 3.1x, 3.2x, 3.3x, 3.4x, 3.5x, 3.6x, 3.7x, 3.8x, 3.9x, 4x, 4.1x, 4.2x, 4.3x, 4.4x, 4.5x, 4.6x, 4.7, 4.8x, 4.9x, 5x 또는 5x 초과의 역치가 사용된다. 본원에서는 콘티그 서열의 세트가 하나의 게놈으로부터 유래되는 방법이 추가로 제공된다. 본원에서는 콘티그 서열의 세트가 복수의 게놈으로부터 유래되는 방법이 추가로 제공된다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0016] 본원에서는 출발 어셈블리의 절단된 콘티그를 표시하는 단계로서, 상기 절단된 콘티그가 노드이고, 상기 절단된 콘티그의 엣지가 정수의 순서쌍의 목록으로 표지되며, 상기 절단부의 엣지가 매핑된 리드 쌍 서열에 상응하는 것인 단계; 및 역치 수 미만의 매핑된 연결부를 갖는 엣지를 배제시키는 단계를 포함하는, 콘티그 어셈블리 방법이 제공된다. 일부 방법에서, 상기 역치 수는 5% 미만이다. 다르게는, 상기 역치 수는 20%, 15%, 14%, 13%, 12%, 11%, 10%, 9%, 8%, 7%, 6%, 4%, 3%, 2%, 1% 미만 또는 그 미만이다. 일부의 경우, 상기 역치 수는 t_L 링크보다 작다. 일부 방법에서, 콘티그는 해당 콘티그 길이에 대해 상응하는 노드의 그래프 내에서의 차수 (degree) 비율이 모든 분포값들 중 약 5%의 최고치를 초과하는 염기 쌍인 엣지를 포함한다. 일부 방법에서, 상기 콘티그는 하나의 게놈으로부터 유래된다. 일부 방법에서, 상기 콘티그는 복수의 게놈으로부터 유래된다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0017] 본원에서는 복수의 콘티그에 상응하는 서열 정보를 수득하는 단계, 상기 복수의 콘티그에 의해 나타난 핵산 샘플의 페어드 엔드 리드 정보를 수득하는 단계, 및 예측된 리드 쌍 거리 데이터 세트의 리드 쌍 거리 파라미터의 편차가 최소화되도록 상기 복수의 콘티그를 배치하는 단계로서, 상기 배치가 8시간 미만으로 일어나는 것인 단계를 포함하는, 콘티그 서열 정보를 하나 이상의 스캐폴드에 어셈블링하는 방법이 제공된다. 상기 예측된 리드 쌍 거리 데이터 세트는 많은 바람직한 실시양태에서 리드 쌍 거리 우도 곡선을 포함한다. 일부의 경우, 상기 리드 쌍 거리 파라미터는 리드 쌍 거리 우도 곡선에 대한 최대 거리 우도이다. 다르게는, 상기 리드 쌍 거리 파라미터는 리드 쌍 거리 우도 곡선에 대한 최소 변이이다. 국소적으로 인접한 콘티그의 세트는 2개의 콘티그를 포함한다. 다르게는, 국소적으로 인접한 콘티그의 세트는 3개의 콘티그를 포함한다. 다르게는, 국소적으로 인접한 콘티그의 세트는 4개의 콘티그를 포함한다. 다르게는, 국소적으로 인접한 콘티그의 세트는 5개의 콘티그를 포함한다. 다르게는, 국소적으로 인접한 콘티그의 세트는 6개의 콘티그를 포함한다. 바람직하게는, 상기 배치는 7시간 미만으로 일어난다. 다르게는, 상기 배치는 6시간 미만, 5시간, 4시간, 3시간, 2시간, 1시간, 또는 1시간 미

만으로 일어난다. 상기 콘티그 정보는 대부분의 경우 하나의 게놈으로부터 유래한다. 다르게는, 상기 콘티그 서열 정보는 복수의 게놈으로부터 유래한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0018] 본원에서는 핵산 샘플을 대표하는 콘티그 서열 세트를 수득하는 단계, 상기 핵산 샘플에 대한 리드 쌍 데이터를 수득하는 단계, 및 핵산 샘플에 대한 리드 쌍 데이터가 예측된 리드 쌍 분포에 보다 가깝게 근접하도록 콘티그 서열 세트를 순서 배열하여 배향시키는 단계를 포함하는, 콘티그 서열 세트를 스캐폴딩하는 방법으로서, 상기 콘티그 서열 세트 중 70%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 8시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인 방법이 제공된다. 상기 스캐폴딩은 콘티그의 세트를 순서 배열하는 단계, 콘티그의 세트를 배향시키는 단계, 2종 이상의 콘티그를 한줄로 병합시키는 단계, 하나의 콘티그를 제2 콘티그에 삽입하는 단계, 및/또는 하나의 콘티그를 2개 이상의 성분 콘티그로 절단하는 단계 중 하나 이상의 단계를 포함한다. 일부 방법에서, 콘티그 서열 세트 중 80%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 8시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향된다. 다르게는, 콘티그 서열 세트 중 90%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 8시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향된다. 다르게는, 콘티그 서열 세트 중 95%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 8시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향된다. 일부의 경우, 상기 콘티그 서열의 세트 중 70%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 4시간 이하, 또는 다르게는 2시간 이하, 또는 다르게는 1시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향된다. 상기 콘티그 정보는 대부분의 경우 하나의 게놈으로부터 유래한다. 다르게는, 상기 콘티그 서열 정보는 복수의 게놈으로부터 유래한다.

[0019] 본원에서는 스캐폴드를 포함하는 복수의 콘티그에 상응하는 서열 정보를 수득하는 단계, 페어드 엔드 리드 정보를 수득하는 단계, 및 상기 페어드 엔드 리드 정보에 대한 페어드 엔드 리드 거리 분포가 전반적으로 최적화되어 레퍼런스 페어드 엔드 리드 거리 분포에 근접하도록, 복수의 콘티그를 배치하는 단계로서, 상기 배치가 8시간 미만으로 일어나는 것인 단계를 포함하는, 핵산 서열 데이터 세트를 구성하는 방법이 제공된다. 상기 콘티그 정보는 대부분의 경우 하나의 게놈으로부터 유래한다. 다르게는, 상기 콘티그 서열 정보는 복수의 게놈으로부터 유래한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0020] 본원에서는 복수의 결합된 노드쌍을 포함하는 스캐폴드 세트를 수득하는 단계로서, 상기 노드 쌍의 각 노드가 하나 이상의 콘티그 서열을 포함하는 것인 단계, 상기 복수의 결합된 노드에 매핑된 페어드 엔드 리드 정보를 수득하는 단계, 결합된 노드쌍에 의해 공유된 리드 쌍의 개수를 계수하는 단계, 상기 개수를 역치와 비교하는 단계, 및 상기 개수가 역치 미만인 경우 노드 쌍을 결합되지 않은 노드로 절단하는 단계를 포함하는, 스캐폴드 어셈블리를 개선하는 방법이 제공된다. 일부의 경우, 고유의 콘티그 서열에 매핑된 리드 쌍만이 계수된다. 본원에서는 역치수의 별개의 리드 쌍 말단이 매핑된 콘티그 서열 세그먼트에 매핑되는 리드 쌍이 폐기되는 방법이 추가로 제공된다. 상기 역치 수는 대부분의 경우 비반복성 서열에 대한 평균 개수의 3x이다. 다르게는, 1.5x, 1.6x, 1.7x, 1.8x, 1.9x, 2x, 2.1x, 2.2x, 2.3x, 2.4x, 2.5x, 2.6x, 2.7x, 2.7x, 2.8x, 2.9x, 3,1x, 3.2x, 3.3x, 3.4x, 3.5x, 3.6x, 2.7x, 2.8x, 3.9x, 4x, 4.5x, 5x 또는 5x 초과의 역치값이 사용된다. 상기 콘티그 정보는 대부분의 경우 하나의 게놈으로부터 유래한다. 다르게는, 상기 콘티그 서열 정보는 복수의 게놈으로부터 유래한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0021] 본원에서는 복수의 결합된 노드쌍을 포함하는 스캐폴드 세트를 수득하는 단계로서, 상기 노드 쌍의 각 노드가 하나 이상의 콘티그 서열을 포함하는 것인 단계, 상기 복수의 결합된 노드에 매핑된 페어드 엔드 리드 정보를 수득하는 단계, 표준 페어드 엔드 리드 거리 빈도 데이터를 수득하는 단계; 상기 복수의 결합된 노드에 매핑된 페어드 엔드 리드 정보에 대한 페어드 엔드 리드 빈도 데이터를 표준 페어드 엔드 리드 거리 빈도 데이터와 비교하는 단계; 및 결합된 노드를 절단하는 것이 상기 복수의 결합된 노드에 매핑된 페어드 엔드 리드 정보에 대한 페어드 엔드 리드 빈도 데이터를 표준 페어드 엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하도록 유도하는 경우에, 하나 이상의 결합된 노드를 절단하는 단계를 포함하는, 스캐폴드 어셈블리를 개선하는 방법이 제공된다. 상기 콘티그 정보는 대부분의 경우 하나의 게놈으로부터 유래한다. 다르게는, 상기 콘티그 서열 정보는

복수의 게놈으로부터 유래한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0022] 본원에서는 콘티그 서열 세트를 수득하는 단계, 페어드 엔드 리드 세트를 포함하는 입력 데이터를 수득하는 단계로서, 상기 페어드 엔드 리드 중 1% 이상이 1 kb 이상의 리드 쌍 거리를 포함하고, 상기 페어드 엔드 리드의 세트가 자연적인 배향의 페어드 엔드 리드를 포함하며, 상기 리드 쌍에 대한 시퀀싱 에러율이 0.1% 이하이고, 상기 입력 데이터의 RN50이 어셈블링된 스캐폴드의 20% 이하인 것인 단계, 및 스캐폴드를 출력하는 단계로서, 스캐폴드에 대한 RN50이 입력 데이터의 RN50의 2x 이상인 것인 단계를 포함하는, 스캐폴드 어셈블리 방법이 제공된다. 임의로, 상기 에러율은 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.9%, 0.8%, 0.7%, 0.6%, 0.5%, 0.4%, 0.3%, 0.2%, 0.1%, 0.09%, 0.08%, 0.07%, 0.06%, 0.05%, 0.04%, 0.03%, 0.02%, 0.01%, 0.001%, 0.0001% 또는 0.00001% 이하 또는 그 미만이다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0023] 본원에서는 T_0 콘티그 서열을 포함하는 콘티그 서열 세트를 수득하는 단계, 페어드 엔드 리드의 세트를 수득하는 단계로서, 상기 페어드 엔드 리드 중 1% 이상이 1 kb 이상의 리드 쌍 거리를 포함하고, 상기 페어드 엔드 리드의 세트가 자연적인 배향의 페어드 엔드 리드를 포함하며, 상기 리드 쌍에 대한 시퀀싱 에러율이 0.1% 이하인 것인 단계, 및 T_1 ($T_1 < T_0$)을 포함하는 스캐폴드를 출력하는 단계를 포함하는, 스캐폴드 어셈블리 방법이 제공된다. 일부의 경우, T_1은 3 미만이다. 임의로, 상기 에러율은 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.9%, 0.8%, 0.7%, 0.6%, 0.5%, 0.4%, 0.3%, 0.2%, 0.1%, 0.09%, 0.08%, 0.07%, 0.06%, 0.05%, 0.04%, 0.03%, 0.02%, 0.01%, 0.001%, 0.0001% 또는 0.00001% 이하 또는 그 미만이다. 다르게는, T_1은 10, 9, 8, 7, 6, 5 또는 4 미만으로 선택된다. 일부의 경우, T_1은 2개의 콘티그이고, 일부의 경우, T_1은 단 하나의 콘티그이다. T_1은 50%, 40%, 30%, 20%, 10%, 5%, 3%, 2%, 1% 미만이거나, 또는 대부분의 경우 1% 미만의 T_1이다. 상기 콘티그 정보는 대부분의 경우 하나의 게놈으로부터 유래한다. 다르게는, 상기 콘티그 서열 정보는 복수의 게놈으로부터 유래한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0024] 본원에서는 리드 쌍을 포함하는 입력 데이터를 수신하는 단계로서, 상기 리드 쌍 중 1% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하고, 입력 데이터에 대한 RN50이 어셈블링된 스캐폴드의 20% 이하이며, 상기 입력 데이터에 대한 에러율이 0.1% 이하인 것인 단계; 및 스캐폴드를 포함하는 출력 데이터를 출력하는 단계로서, 상기 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 2x 이상인 것인 단계를 포함하는, 핵산 서열 데이터 처리 방법이 제공된다. 일부 방법에서, 출력 데이터에 대한 RN50은 입력 데이터의 RN50의 10x 이상, 또는 다르게는 3x, 4x, 5x, 6x, 7x, 8x, 9x, 11x, 12x, 13x, 14x, 15x, 16x, 17x, 18x, 19x, 20x, 30x, 40x, 50x, 60x, 70x, 80x, 90x, 100x, 500x, 1000x, 또는 1000x 초과이다. 본원에서는 스캐폴드가 90% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 방법이 추가로 제공된다. 본원에서는 스캐폴드가 99% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 방법이 추가로 제공된다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0025] 본원에서는 리드 쌍을 포함하는 데이터세트를 출력하는 단계로서, 상기 리드 쌍 중 1% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하고, 출력 데이터에 대한 RN50이 어셈블링된 스캐폴드의 20% 이하이며, 상기 출력 데이터에 대한 에러율이 0.1% 이하인 것인 단계; 및 스캐폴드를 포함하는 데이터세트를 수신하는 단계로서, 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 2x 이상인 것인 단계를 포함하는, 핵산 서열 데이터 처리 방법이 제공된다. 일부 방법에서, 상기 출력 데이터에 대한 RN50은 입력 데이터의 RN50의 10x 이상, 또는 다르게는 3x, 4x, 5x, 6x, 7x, 8x, 9x, 11x, 12x, 13x, 14x, 15x, 16x, 17x, 18x, 19x, 20x, 30x, 40x, 50x, 60x, 70x, 80x, 90x, 100x, 500x, 1000x, 또는 1000x 초과이다. 본원에서는 스캐폴드가 90% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 방법이 추가로 제공된다. 본원에서는

스캐폴드가 99% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 방법이 추가로 제공된다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0026] 본원에서는 리드 쌍을 포함하는 입력 데이터를 수신하는 단계로서, 상기 리드 쌍 중 1% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하고, 입력 데이터에 대한 N50이 어셈블링된 스캐폴드의 20% 이하이며, 상기 출력 데이터에 대한 에러율이 0.1% 이하인 것인 단계; 및 스캐폴드를 포함하는 출력 데이터를 출력하는 단계로서, 상기 출력 데이터에 대한 N50이 입력 데이터의 RN50의 2x 이상인 것인 단계를 포함하는, 핵산 서열 데이터 처리 방법이 제공된다. 임의로, 상기 에러율은 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.9%, 0.8%, 0.7%, 0.6%, 0.5%, 0.4%, 0.3%, 0.2%, 0.1%, 0.09%, 0.08%, 0.07%, 0.06%, 0.05%, 0.04%, 0.03%, 0.02%, 0.01%, 0.001%, 0.0001% 또는 0.00001% 이하, 또는 그 미만이다. 상기 콘티그 정보는 대부분의 경우 하나의 게놈으로부터 유래한다. 다르게는, 상기 콘티그 서열 정보는 복수의 게놈으로부터 유래한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0027] 본원에서는 리드 쌍을 포함하는 출력 데이터를 출력하는 단계로서, 상기 리드 쌍 중 1% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하고, 상기 출력 데이터에 대한 RN50이 어셈블링된 스캐폴드의 20% 이하이며, 상기 출력 데이터에 대한 에러율이 0.1% 이하인 것인 단계; 및 스캐폴드를 포함하는 입력 데이터를 수신하는 단계로서, 상기 출력 데이터에 대한 RN50이 어셈블링된 스캐폴드의 20% 이하인 단계를 포함하는, 핵산 서열 데이터 처리 방법이 제공된다. 상기 콘티그 정보는 대부분의 경우 하나의 게놈으로부터 유래한다. 임의로, 상기 에러율은 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.9%, 0.8%, 0.7%, 0.6%, 0.5%, 0.4%, 0.3%, 0.2%, 0.1%, 0.09%, 0.08%, 0.07%, 0.06%, 0.05%, 0.04%, 0.03%, 0.02%, 0.01%, 0.001%, 0.0001%, 또는 0.00001% 이하, 또는 그 미만이다. 다르게는, 상기 콘티그 서열 정보는 복수의 게놈으로부터 유래한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0028] 본원에서는 제1 콘티그에 대해 매핑된 샷건 리드의 밀도를 측정하는 단계, 제2 콘티그에 대해 매핑된 샷건 리드의 밀도를 측정하는 단계, 제1 콘티그와 제2 콘티그를 결합시키는데 있어서의 우도 스코어를 측정하는 단계, 및 제1 콘티그에 대해 매핑된 샷건 리드의 밀도가 제2 콘티그에 대해 매핑된 샷건 리드의 밀도와 크게 다른 경우에 우도 스코어를 감소시키는 단계를 포함하는, 하나 이상의 페어드 엔드 리드를 공유하는 2개의 핵산 콘티그를 결합할 확률을 평가하는 방법이 제공된다. 일부 방법에서, 우도 스코어는 로그 우도 스코어이다. 보통, 상기 스코어는 본원에 지시된 바와 같이 감소된다. 보통, 상기 스코어는 제1 콘티그에 대해 매핑된 샷건 리드의 밀도 및 제2 콘티그에 대해 매핑된 샷건 리드의 밀도 중 큰 수치에 대한 작은 수치의 비율만큼 감소된다. 상기 콘티그 정보는 대부분의 경우 하나의 게놈으로부터 유래한다. 다르게는, 상기 콘티그 서열 정보는 복수의 게놈으로부터 유래한다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0029] **참고 인용**

[0030] 본 명세서에서 언급한 모든 간행물, 특허 및 특허출원들은 각 개별 간행물, 특허 및 특허출원들이 구체적이고도 개별적으로 참고로서 인용될 것으로 지시된 것과 동일한 정도로 본원에 참고로 인용된다.

도면의 간단한 설명

[0031] 본 개시내용의 신규한 특징들을 본원에 첨부된 청구범위와 개요 및 상세한 설명에 상세히 기재한다. 본 명세서의 특징과 잇점들은 본 개시내용의 원리가 사용된 예시적인 실시양태들을 기술하고 있는 하기의 상세한 설명 및 그에 첨부된 도면을 참고하면 더 잘 이해될 것이다.

도 1A-F는 디 노보 (*de novo*) 시퀀싱용 재료를 생성하는 공정의 흐름을 나타낸 것이다.

도 2는 참조 인간 게놈 어셈블리 hg19에 매핑된 수개의 시퀀싱 라이브러리를 위한 리드 쌍 분리를 도시한 것이

다.

도 3은 다양한 리드 쌍 분리 빈 (bin)에서의 커버리지 (추정된 게놈 크기로 나눈 리드 쌍 분리의 합) 차트이다.

도 4A 및 4B는 리드 쌍의 GRCh38 참조 서열 상에 매핑된 위치가, GM12878과 참조 서열 간의 구조적 차이의 부근에 플롯팅되어 있는 것을 도시한 것이다. 도 4A는 20kb 반복성 측부 영역을 갖는 80Kb 전위에 대한 데이터를 나타낸 것이다. 도 4B는 단계적 (Phased) 이형접합 결실에 대한 데이터를 나타낸 것이다.

도 5는 디 노보 시퀀싱용 재료를 생성하는 공정의 흐름을 나타낸 것이다.

도 6은 본 발명의 일부 실시양태와 일치하는 컴퓨터 시스템의 실례를 도시한 것이다.

도 7은 본 발명의 예시적 실시양태와 일치하는 컴퓨터 시스템 2000의 제1 예시적 구성을 도시하는 블록 다이어그램이다.

도 8은 본 발명의 예시적 실시양태와 일치하는 복수의 컴퓨터 시스템, 복수의 휴대폰과 개인 정보 단말기, 및 네트워크 접속기반 파일서버 (Network Attached 저장, NAS)를 도입하도록 구성된 네트워크(800)를 보여주는 다이어그램이다.

도 9는 본 발명의 예시적 실시양태와 일치하는 공유 가상 주소 메모리 공간을 이용하는 멀티프로세서 컴퓨터 시스템(900)의 블록 다이어그램이다.

도 10A-D는 초기의 콘티그 어셈블리 (도 10A)로부터 '윈도우' 분석을 위해 선형으로 순서 배열된 콘티그의 생성 (도 10D)까지의 예시적인 작업흐름을 나타낸 것이다.

도 11은 최소 신장 트리를 나타낸 것이다. 노드 라벨은 콘티그 크기 (kb)를 나타내고, 엣지 라벨은 표시된 각 콘티그 쌍 엣지에서의 링크 수를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0032] **상세한 설명**

[0033] 짧은 리드 데이터로부터 원거리의 고도로 정확한 *디 노보* 어셈블리를 수행하는 것은 유전체학에서 가장 시급한 당면 과제이다. 우리는 본원에서 근접 결찰 라이브러리의 생성을 위한 기질로서 살아있는 염색체보다는 재구성된 염색질을 사용하여 수백 kb 이하의 DNA 연관성이 시험관 내에서 구축됨을 입증하였다. 생성된 라이브러리는, 리드 쌍 내의 거리와 리드 계수 간의 규칙적인 관계를 비롯한 원거리 게놈 어셈블리 및 페이징 (phasing)에 유용한 Hi-C 데이터의 많은 특성들을 공유하고 있다. 이 *시험관내* 원거리 메이트 페어 (mate-pair) 라이브러리와 표준 전장 게놈 샷건 및 점핑 라이브러리를 조합하여, 우리는 적은 비용과 노력으로 보다 고가의 방법과 맞먹는 원거리 정확도 및 근접성을 갖는 *디 노보* 인간 게놈 어셈블리를 제조하였다. 이 방법은 적정량의 고분자량 DNA만을 사용하고, 일반적으로 어떤 종에라도 적용가능하다. 본원에서 우리는 (예를 들어, 하나의 게놈 또는 염색체 세트를 대표하는 스캐폴드로의) *디 노보* 핵산 서열 어셈블리 또는 인간과 악어를 이용하는 스캐폴드 어셈블리를 위해서 뿐만이 아니라, 구조적 변이의 확인 및 이형접합성 변이체의 페이징을 위한 효율적인 도구로서의 이 서열 데이터의 가치를 입증하였다.

[0034] 본원에서는 *시험관내* 재구성된 염색질에 대한 예시적인 실시양태를 기반으로 한 서열 어셈블리 접근법이 개시된다. 본원의 방법, 시스템 및 조성물을 통하여, 적절히 단계적으로 그룹화되고, 순서 배열되고, 배향되며, 병합되거나 또는 분리되도록, 게놈 또는 다른 큰 서열 데이터 세트에 대한 고도로 정확한 *디 노보* 어셈블리 및 스캐폴딩이 달성된다. 이와 유사하게, 이전에 이용가능한 콘티그 및 스캐폴드 서열 정보를 리어셈블링 및 스캐폴딩함으로써 기존 어셈블리를 개선하는 유용성도 입증되었다. 일부의 경우, 단 하나의 라이브러리 및 한 레인의 일루미나 (Illumina) HiSeq 시퀀싱을 사용하여 리드 쌍을 생성하면, 스캐폴드 N50이 약 500 kbp에서 10Mbp로 증가된다. 본원에 개시된 방법은 임의의 핵산 샘플 (예컨대, 하나의 게놈 또는 복수의 게놈)을 분석하는데 사용되며, 어셈블링하기 힘든, 트랜스포존 또는 다른 반복 요소가 풍부한 반복성 또는 배수체 게놈을 포함하는 게놈 샘플, 또는 특히 계산 집약적이어서 8시간 이하, 7, 6, 5, 4, 3, 2시간 또는 2시간 미만으로 어셈블링하는 샘플 리드 데이터 세트를 유도하는 기타 샘플에 특히 적절하다. 일부 실시양태에서, 상기 방법들은 본원에 논의된 바와 같은 콘티그 및 리드 쌍 데이터를 수신하고, 상기 논의된 바와 같이 상기 데이터를 처리하여, 상기 논의된 바와 같은 개선된 파라미터를 갖는 스캐폴딩된 콘티그 데이터를 출력하도록 구성된 프로세서를 포함하는 컴퓨터 실행 시스템 상에서 수행된다.

[0035] 본원 명세서와 첨부된 청구범위에 사용된 바와 같이, 단수형 (관사와 정관사)은 명백히 달리 지시하지 않는 한

복수의 대상체도 포함한다. 따라서, 예를 들어, "하나의 콘티그"에 대한 언급은 복수개의 이러한 콘티그를 포함하며, "염색체의 물리적 배치에 대한 탐침"에 대한 언급은, 문맥상 하나의 실체를 언급하는 것을 나타내지 않는 한, 염색체의 물리적 배치에 대한 하나 이상의 탐침 방법 및 당업자에게 공지된 이의 등가물에 대한 언급을 포함한다. 또한, "및"의 사용은 달리 언급하지 않는 한 "및/또는"을 의미한다. 이와 유사하게, "포함하다," "포함하는," "비롯한" ("comprises," "comprising," "include," "includes," 및 "including")은 상호교환하여 사용할 수 있고, 이들을 구별하고자 하는 의도는 없다.

[0036] 추가로, 다양한 실시양태의 설명이 "포함하는" 이라는 용어를 사용하는 경우, 당업계의 숙련자라면 일부 특정한 경우에 "본질적으로 ...로 이루어지는" 또는 "...로 이루어지는" 이라는 말을 사용하여 대안적으로 기술되는 추가의 별개 실시양태가 포함된다는 점을 이해할 수 있을 것이다.

[0037] 본원에 사용된 "리드 (read)" 또는 "시퀀싱 리드" 라는 용어는 서열이 정해지는 DNA 세그먼트의 서열 정보를 가리킨다.

[0038] 본원에서 사용된 "콘티그 (contig)" 라는 용어는 DNA 서열의 인접한 영역을 지칭한다. "콘티그"는 당업계에 공지된 여러가지 방법에 의해, 예컨대 중복 서열에 대해 시퀀싱 리드를 비교하고/하거나, 어떤 시퀀싱 리드가 인접하고 있을 확률이 높은지를 확인하기 위해 공지된 서열의 데이터베이스에 대해 시퀀싱 리드를 비교함으로써 알아낼 수 있다. 콘티그는 개개의 서열 리드로부터 어셈블링되거나, 또는 이전에 어셈블링된 서열 정보를 중복 엔드 또는 엣지 서열을 갖는 서열 리드와 조합하여 어셈블링되는 경우가 많다. 전적으로 다 그렇다는 것은 아니지만 일반적으로는, 콘티그는 대부분의 경우 정해지지 않은 서열의 개재되는 갭 또는 영역 없이, 또는 다르게는 기지의 서열과 미지의 길이를 갖는 영역 없이 더 큰 서열 그룹으로 어셈블링되는 중복 서열 리드를 포함한다.

[0039] 본원에서 사용된 "스캐폴드" 라는 용어는 스캐폴드의 모든 서열 정보가 공통의 상을 공유하거나 해당 서열 정보가 나타내는 핵산이 물리적으로 연결되었음을 반영하도록, 단일 물리적 분자에 상응하는 하나 이상의 콘티그 또는 서열 리드의 서열 정보를 가리킨다. 일부의 경우, 스캐폴드 서열은 단 하나의 콘티그로 어셈블링되지 않고, 그의 성분 콘티그 또는 서열 리드 간에 미지의 서열, 미지의 길이, 또는 미지의 서열과 미지의 길이의 하나 이상의 갭을 가질 수 있다. 일부 이러한 경우에서, 해당 성분 서열이 상으로 존재하거나 또는 단일 물리적 분자로 매핑된다는 사실 때문에, 갭핑된 서열은 역시 단 하나의 스캐폴드를 구성한다. 일부의 경우, 스캐폴드는 단 하나의 콘티그를 포함한다 - 즉, 일부의 경우, 스캐폴드는 어떠한 갭도 없는 하나의 연속된 서열의 신장을 포함한다.

[0040] 동사로서, "스캐폴드" 라는 용어는 순서 배열하는 단계, 배향시키는 단계, 한줄로 병합시키는 단계, 하나를 다른 하나에 병합시키는 단계, 및 콘티그 또는 스캐폴드를 절단하는 단계, 그리고 순서 배열하는 단계, 배향시키는 단계, 한줄로 병합시키는 단계, 하나를 다른 하나에 병합시키는 단계, 및 콘티그 또는 스캐폴드를 절단하는 단계를 모두 포함하는 것까지 해서 이들 중 하나 이상을 가리키며, 이들은 본원에 제시된 방법을 바탕으로 수행된다. 스캐폴딩은 진핵 생물의 염색체 상에 콘티그를 매핑되는 것으로부터 생길 수 있거나, 또는 이종성 샘플 내 복수의 유기체의 게놈에 상응할 수 있는 것과 같이, 복수의 콘티그를 단일 상의 단일 분자 상에 어셈블링하거나 복수의 스캐폴드 상에 어셈블링하기 위해 수행된다.

[0041] 본원에서 사용된 바와 같이, 페어드 리드와 관련하여 "자연적 배향" 이라는 것은 페어드 서열이 이들이 유래되는 핵산 분자 세그먼트의 배향으로 또는 이들을 나타내는 배향으로 발생하는 페어드 리드를 지칭한다.

[0042] 본원에서 사용된 바와 같은 "대상체" 라는 용어는 임의의 진핵 또는 원핵 (진정세균 또는 고세균) 유기체 또는 바이러스를 지칭할 수 있다. 다르게는, 대상체는 복수의 유기체 및/또는 바이러스 유래의 핵산 물질을 포함하는 환경적 샘플과 같이 그의 유기체 기원과는 독립적인 샘플을 지칭할 수 있다. 예를 들어, 대상체는 인간과 같은 포유동물일 수 있거나, 또는 예컨대 인간과 실질적으로 인간이 아닌 핵산 서열을 모두 포함하는 것으로 예측되는 인간의 소화관으로부터 취해진 샘플일 수 있다.

[0043] 본원에서 사용된 바와 같은 "핵산" 또는 "폴리뉴클레오타이드" 라는 용어는 단일 또는 이중 가닥 형태의 데옥시리보뉴클레오타이드 (DNA) 또는 리보뉴클레오타이드 (RNA)의 중합체를 지칭할 수 있다. 구체적으로 한정되지 않는 한, 상기 용어는 참조 뉴클레오타이드와 유사한 결합 특성을 갖고/갖거나 또는 자연적으로 발생하는 뉴클레오타이드와 유사한 방식으로 대사되는 자연적으로 발생하는 뉴클레오타이드의 공지된 유사체를 함유하는 핵산 분자들을 망라한다.

[0044] 본원에서 사용된 바와 같은 "네이키드 DNA" 라는 용어는 실질적으로 복합 단백질 또는 나노입자가 없는 DNA를 지칭할 수 있다.

[0045] 본원에서 사용된 바와 같은 "재구성된 염색질" 이라는 용어는 단리된 핵 단백질 또는 기타 핵산 결합 모이어티를 네이키드 DNA에 복합체화시킴으로써 형성된 염색질을 지칭할 수 있다. 일부의 경우, 재구성된 염색질은 실제로 핵산 및 히스톤과 같은 염색질 성분을 포함하는 한편, 대안적인 실시양태에서 "재구성된 염색질"은 예를 들어, 핵산과 특이적으로 또는 비특이적으로 결합하는 하나 이상의 핵산 결합 모이어티, 예컨대 단백질, 나노입자, 또는 스퍼미딘 또는 스퍼민과 같은 비단백질성 분자와 조합된 네이키드 DNA 또는 추출된 DNA로부터 형성된 임의의 복합체를 지칭하는데 보다 비공식적으로 사용된다.

[0046] 본원에서 사용된 바와 같은 "나노입자" 라는 용어는 DNA와 결합하도록 변형될 수 있는 나노미터 크기의 구체를 지칭할 수 있다. 일부의 경우, 나노입자는 (예컨대, 아민 함유 분자를 사용한 코팅에 의해) 그 표면이 양으로 하전된다. 문헌 [Zinchenko, A. et al.(2005) "Compaction of Single-Chain DNA by Histone-Inspired Nanoparticles" *Physical Review Letters,* 95(22), 228101] (내용 전체가 본원에 참조로 포함됨)을 참고한다. 일부 실시양태에서, 재구성된 염색질은 네이키드 DNA에 나노입자를 결합시켜 합성된다.

[0047] 본원에서 사용된 바와 같은 "리드 쌍" 이라는 용어는 자연적인 핵산 분자 샘플에서는 인접하고 있지 않지만 본원 또는 다른 곳에 개시된 것과 같이 화학적 또는 효소적 조작의 결과로 인접하여 공유 결합된 2종 이상의 핵산 서열 범위를 지칭할 수 있으며, 이들은 단일 시퀀싱 리드로 시퀀싱된다. 일부의 경우, "리드 쌍"은 인공적으로 결합된 2개의 핵산 영역에 대한 시퀀싱에 의해 수득된 서열 정보를 가리킨다. 일부의 경우, 리드-쌍의 수는 매핑가능한 리드-쌍의 수를 지칭할 수 있다. 기타의 경우, 리드-쌍의 수는 생성된 리드-쌍의 총수를 가리킬 수 있다.

[0048] 본원에서 사용된 바와 같이, '샘플'은 스캐폴드 정보가 생성되거나 개선될 핵산 물질을 지칭한다. 일부 샘플은 단일 미생물계 또는 단일 다세포성 개체의 조직과 같은 균질한 공급원으로부터 유래된다. 일부의 경우, 샘플은 서열 변이, 예컨대 개체의 종양 샘플에서 나타날 수 있는 변이를 포함한다. 일부의 경우, 샘플은 그것이 복수의 유기체의 핵산을 포함하도록 이종성 공급원, 예컨대 인간 소화관 또는 배설물 샘플, 환경적 샘플 또는 유기체의 혼합물로부터 유래된다.

[0049] 본원에서 사용된 바와 같이, "약" 어느 수라는 용어는 그 어느 수를 명확하게 언급하는 것에 더하여, 그 수에 ±10%를 한 수량을 가리키는데 사용된다.

[0050] 달리 정의되지 않는 한, 본원에 사용된 모든 과학 기술 용어들은 이러한 개시내용이 속하는 당업계의 숙련자에게 통상적으로 이해되는 바와 갓은 의미를 가진다. 본원에 기술된 것들과 유사하거나 동일한 방법과 시약들이 개시된 방법과 조성물을 실시하는데 사용될 수 있지만, 이제는 예시적인 방법과 재료들에 대해 기술한다.

[0051] 리드 쌍 서열 정보, 예컨대 핵산 서열 상 또는 물리적 연관을 나타내는 리드 쌍 정보의 사용을 통한, 서열 어셈블리, 예컨대 단일 리드 및 콘티그의 더 큰 콘티그 및 스캐폴드로의 핵산 서열 어셈블리와 관련된 조성물, 시스템 및 방법이 본원에 개시된다.

[0052] 유전체학의 주요 목적은 적은 노력과 비용으로 전장 하플로타입 분석된 염색체 서열의 정확한 재구축이다. 현재 적정가격으로 이용가능한 고처리량 시퀀싱 방법은 단거리 서열 근접성 및 게놈 변이의 특성에 최적화되어 있다. 원거리 연관 및 하플로타입 페이징의 달성은 긴 (예컨대, 수십 킬로베이스) 서열을 직접적이고 정확하게 판독하는 능력, 또는 한 쌍의 리드 또는 그룹화된 서열 리드를 통한 연관 및 상 관계의 포착을 필요로 한다. 이러한 방법은 기술적으로 어렵고도 계산 집약적이어서, 게놈 샘플에 대한 전체 샘플 하플로타입 지도 정보를 생성하는데 필요한 서열 정보의 통상적이거나 시판되는 컴퓨터 분석이 불가능하다.

[0053] 고처리량 시퀀싱 방법은 유전체학의 분야에서 혁명을 불러 일으켰다. 수백만개의 짧은 DNA 단편들의 데이터를 한번에 생성함으로써, 게놈을 다시 시퀀싱하는 비용이 급감하여 인간 게놈 하나당 $1,000로 신속하게 접근하고 있으며 (Sheridan, 2014), 앞으로 더 떨어질 것으로 기대된다.

[0054] 그러나, 실질적인 장애물은 짧은 리드 서열을 길고 연속적인 게놈 어셈블리로 변환시키는데에 있다. 적은 비용의 서열 데이터로부터 참조 품질의 어셈블리를 생성하는 어려움은 오늘날의 기술로 생성된 어셈블리의 품질과 인간 참조 어셈블리를 비교해 보면 명백하다 (Alkan et al., 2011).

[0055] 38.5 Mbp의 N50 길이 및 100,000개의 염기당 1개의 에러율을 갖는 고품질의 고도로 인접한 인간 참조 표준을 생성하기 위해, BAC 클론 시퀀싱, 물리적 지도, 및 생어 시퀀싱을 비롯한 여러가지 기법들이 사용되었다 (International Human Genome Sequencing Consortium, 2004). 이와는 대조적으로, 전장 게놈 샷건 (WGS) 어셈블리 소프트웨어 파이프라인의 성능에 대한 최근의 비교에서, 복수개의 삽입 크기를 갖는 라이브러리의 매우 높

은 커버리지 데이터 세트에 대한 이들의 고안자에 의한 각 실행은, 어류 게놈에 대해서는 4.5 Mbp 이하 및 뱀 게놈에 대해서 4.0 Mbp 이하의 범위인 N50 스캐폴드 길이를 갖는 어셈블리를 생성하였다 (Bradnam et al., 2013).

[0056] 짧은 리드를 갖는 서열의 높은 커버리지는 고품질의 고도로 인접한 어셈블리를 수득하기에는 충분치 않다. 이러한 주된 이유는 동원체와 말단소체, 징크 핑거 유전자와 같은 큰 유사 유전자 부류 근처의 반복적 구조물을 비롯한 크고 작은 규모 모두에 대한 반복적 내용물, 및 LINE 및 SINE과 같은 산재성 핵 성분들의 분포에 기인한다. 이러한 어셈블링하기 어려운 내용물은 많은 진핵생물 게놈의 대부분, 예컨대 인간 게놈의 60-70%를 구성한다 (de Koning et al., 2011). 이러한 반복 단위가 입력 서열 데이터에 의해 포함되지 않는 경우에는 단편화된 부정확한 어셈블리가 생성된다. 일반적으로, *디 노보* 어셈블리를 위한 출발점은 깊은 커버리지 (최소 50X-200X)의, 단거리 (300-500 bp) 페어드 엔드 "샷건" 데이터를, 2 내지 8 kbp의 삽입 크기를 갖는 중간 정도 범위의 "메이트 페어" 라이브러리, 및 흔히 더 긴 범위의 (35 kbp) 포스미드 엔드 쌍과 조합한다 (Gnerre et al., 2011; Salzberg et al., 2012).

[0057] 저비용의 서열 데이터가 유용하지만, 상 정보가 속할 수 있는 더 큰 콘티그 또는 스캐폴드로 어셈블링하기는 어렵다. 따라서, 게놈 재배열과 관련한 가치있는 정보, 또는 하나의 유전자 자리 (독립적으로 돌연변이된 2개의 대립유전자 또는 이중 돌연변이체 대립유전자와 조합된 하나의 야생형 대립유전자에 상응하는 시스 또는 트랜스) 내에 분산된 다수의 돌연변이의 상에 관한 보다 간단한 정보는 일부 저비용의 서열 어셈블리로부터 이용가능하지 않은 경우가 많다.

[0058] *디 노보* 어셈블리의 근접성과 정확성을 증가키시기 위한 다양한 방법들이 최근에 개발되었다. 대략적으로, 이러한 방법들은 시퀀싱으로부터 생성된 리드 길이를 증가시키거나 또는 짧은 페어드 리드 간의 삽입 크기를 증가시키기 위한 시도를 한다. 예를 들어, PacBio RS II는 23 kbp (중앙값 2 kbp) 까지의 길이의 미가공 리드를 생산할 수 있다. 그러나, 이러한 접근법은 고처리량의 짧은 리드에 비해 ~15%로 에러율이 높고 ~100배 더 비싸 문제가 있는 것으로 보고된 바 있다 (Koren et al., 2012; Quail et al., 2012). 옥스포드 나노포어 (Oxford Nanopore)로부터 시판되는 긴 리드가 유망하지만, 그 조차도 높은 에러율과 낮은 처리량을 가지는 경우가 많다. 일루미나의 TruSeq 합성 고판독 (Synthetic Long-Read) 기법 (전신은 몰레큘로 (Moleculo)임)은 현재 최대 10 kbp 리드로 제한된다 (Voskoboynik et al., 2013).

[0059] 다수의 개선점에도 불구하고, 포스미드 라이브러리 생성 (Williams et al., 2012; Wu et al., 2012)은 시간이 많이 걸리고 비싸다.

[0060] 현재까지, 시퀀싱계는 수천 가지의 척추동물 종 (Haussler et al., 2009) 또는 수십만 가지의 인간 (Torjesen, 2013)의 시퀀싱과 같은 대규모 프로젝트에 필요한 규모 및 비용으로 이용가능한 큰 인서트 또는 긴 리드를 위해 일관되게 월등한 기법에 대해 확립한 것이 없다.

[0061] 핵산 서열 어셈블리를 위한 방법 및 컴퓨터 실행 시스템이 본원에 개시된다. 본원에 개시된 방법은 페어드 엔드 리드를 이용한 분석 및 서열 개선에 특히 적절하다. 페어드 엔드 리드 또는 리드 쌍은 다수의 서로 다른 접근법을 사용하여 생성된다. 일부 접근법은 수백 kbp 이하로 분리된 원거리 리드 쌍의 시험관내 제조 방법, 및 리드 쌍 내의 각 리드가 매핑되는 콘티그에 공통의 상 또는 물리적 연관 정보를 지정함에 있어서 이들의 사용을 수반한다. 본원 개시내용의 일부 실시양태에 대한 핵심은 *디 노보* 스캐폴드 어셈블리 및 페이징을 개선하는데 유용한 원거리 리드 쌍 데이터를 생성하기 위해 거리와 리드 계수간의 관계를 이용하는 Hi-C의 예기치 못한 효율적인 개선이다. 선행 기술인 Hi-C 방법과는 달리, 본원에 개시된 일부 접근법들은 고정화, 페어드 엔드 형성 및 후속 단계를 위한 기질로서 *시험관내* 재구성된 염색질을 사용한다. 생성된 데이터는 리드 쌍 계수와 리드 간 거리 사이의 관계를 비롯한 Hi-C 데이터의 많은 특성들을 공유한다. 그러나, 많은 실시양태에서, 이렇게 생성된 페어드 엔드 리드는 3차원의 염색질 또는 숙주 세포 내의 다른 핵산 구성에 관한 정보, Hi-C 기법의 특히 가치있는 목표에 대한 정보는 부족하지만, 상 또는 물리적으로 서로 연결되어 있지 않은 분자의 핵산 서열의 근접성 정보는 수반할 수 있다. 이들 새로운 *시험관내* 데이터는 단리화에서 사용될 수 있거나, 또는 매우 적은 비용과 시간으로도 페어드 엔드 샷건 리드 또는 사전에 생성된 다른 콘티그 정보와 조합하여 포스미드 기반의 어셈블리에 필적하는 정확도와 근접성을 갖는 *디 노보* 스캐폴드 어셈블리를 생성할 수 있다. 이러한 진전된 분석과 관련된 방법, 조성물 및 컴퓨터 실행 시스템 뿐만 아니라, *디 노보* 어셈블리, 하플로타입 페이징의 품질을 개선하고 구조적 변이를 확인하기 위한 이들의 유용성도 본원에 개시된다.

[0062] 본원의 개시내용은 약 10,000, 약 20,000, 약 50,000, 약 100,000, 약 200,000, 약 500,000, 약 1백만, 약 2백만, 약 5백만, 약 1천만, 약 2천만, 약 3천만, 약 4천만, 약 5천만, 약 6천만, 약 7천만, 약 8천만, 약 9천만,

약 1억, 약 2억, 약 3억, 약 4억, 약 5억, 약 6억, 약 7억, 약 8억, 약 9억, 약 10억 리드 쌍 이하, 또는 10억 초과의 염기 쌍을 사용하여, 고도로 인접하고 정확한 인간 게놈 어셈블리를 생산할 수 있는 방법 및 컴퓨터 시스템을 제공한다. 일부의 경우, 본 개시내용은 인간 게놈 내에서 약 50%, 60%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% 또는 그 이상의 이형접합성 변이체를 약 50%, 60%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 또는 그 이상의 정확도로 페이징하는 방법 및 컴퓨터 시스템을 제공한다.

[0063] 리드 쌍 엔드 생성 및 시퀀싱에 관한 조성물과 방법, 및 태깅된 엔드 리드를 생성하는 대안적인 방법은 예를 들어, 2014년 8월 7일 공개된 미국 특허공보 US20140220587호 (그 내용 전체가 본원에 참조로 포함됨)에서 찾아볼 수 있다.

[0064] 본원에 개시된 실시양태는 하기의 단계들 및 각각의 하기의 단계를 모두 포함하는 것까지 해서 이들 중 하나 이상을, 단독으로 또는 본원에 개시되거나 당업계의 숙련자에게 공지된 추가의 단계와 조합하여 포함한다:

[0065] 1. 입력 전처리

[0066] 2. 콘티그-콘티그 연관 그래프 구축

[0067] 3. 씨드 스캐폴드 구축

[0068] 4. 국소적 순서, 배향, 및 갭 크기 개선

[0069] 5. 쌍별 스캐폴드 병합을 위한 스코어의 계산

[0070] 6. 결합에 대한 과잉 합격 판정 (Greedy acceptance of joins)

[0071] 일부 실시양태에서, 일련의 상기 언급된 단계들은 조합하여 수행한다. 일부의 경우, 단계 4, 5 및 6 중에서 하나 이상의 단계는, 예를 들어 어셈블리 품질의 지속적인 개선을 가져오기 위하여 반복적으로 수행된다.

[0072] 본원에 개시된 일부 방법 및 시스템은 하기의 3가지 입력을 고려한다:

[0073] 1. 출발 어셈블리의 서열 (임의로는 FASTA 포맷으로);

[0074] 2. 출발 어셈블리에 대한 페어드 리드의 정렬 (임의로는 BAM 포맷으로, 분류하여 색인을 닮);

[0075] 3. 출발 어셈블리에 대한 샷건 리드의 정렬 (임의로는 BAM 포맷으로, 분류하여 색인을 닮).

[0076] 대안적인 데이터 포맷들도 고려되며, 방법과 시스템은 특정 데이터 포맷 또는 색인에 한정되지는 않는다. 출발 어셈블리는 서열 리드로부터 어셈블링된 하나 이상의 콘티그를 포함하며, 일부의 경우, 서열 포맷과는 독립적인 하나 이상의 콘티그를 포함하는 하나 이상의 스캐폴드를 포함한다.

[0077] 번갈아 언급되는 페어드 리드, 페어드 엔드 리드, 또는 리드 쌍은 서열 포맷과는 독립적인 비인접된 표적 샘플 서열에 상응하는 서열 정보를 포함한다. 대부분의 경우, 페어드 리드는 샘플 핵산 서열 중 긴 거리의 공통의 분자에 의해 분리된 위치에서의 단일 물리적 분자의 리드 서열에 해당한다.

[0078] 서로에 대한 콘티그의 상대적 위치, 상대적 배향 중 하나 이상을 나타내는 스캐폴드에 콘티그 데이터를 어셈블링하기 위한 방법 및 컴퓨터 시스템이 본원에 개시된다. 일부의 경우, 개개의 콘티그들은, 위치 및/또는 배향이 결정된 인접한 콘티그들을 결합함으로써, 또는 하나 이상의 콘티그를 제2 스캐폴드 또는 콘티그의 갭 또는 어셈블링되지 않은 영역으로 삽입함으로써 공정 중에 병합되어, 비슷한 크기이지만 서열 품질이 개선된 콘티그 또는 스캐폴드를 형성한다.

[0079] 샘플에 대한 페어드 엔드 리드 정보는 본원에 개시된 방법, 본원에 포함된 방법, 또는 그렇지 않으면 당업계의 숙련자에게 공지된 방법을 사용하여 생성된다. 일부의 경우, 태깅된 엔드 리드, 예컨대 태깅된 페어드 엔드 리드는 본원에 논의된 페어드 엔드 리드를 대체한다. 일부의 경우에, 페어드 엔드 리드 또는 기타 정보는, 예를 들어 동시에 수행된 이전의 시퀀싱 결과 또는 샷건 시퀀싱 결과로부터, 일부의 경우 동시에 생성되고 다른 경우에는 이와는 독립적으로 수득된 콘티그 정보, 예컨대 샷건 시퀀싱 콘티그 정보와 조합하여 사용된다. 일부의 경우, 상기 콘티그 정보는 서열 데이터베이스 또는 이전의 시퀀싱 결과로부터 수득된다.

[0080] 페어드 엔드 리드는 본원에 개시된 바와 같이, 또는 당업계의 숙련자에게 공지되거나 본원의 개시내용과 양립가능한 다른 방법들을 통해서 생성된다. 페어드 리드 생성에 있어서의 크고 작은 변형들도 고려된다. 페어드 리드는 처리 전에 샘플 재료 내에 인접하고 있지 않은 리드 서열의 쌍을 포함한다. 대부분의 경우, 페어드 리드는

하나의 물리적 분자에 매핑되지만, 서로 간에 어느 정도 거리를 둔다. 일부의 경우, 구별되는 물리적 분자의 쌍을 포함하는 페어드 리드가 생성된다. 본원에 개시된 방법에서, 이러한 페어드 엔드 리드는 비교적 드물고, 이들은 본원에 개시된 어셈블리 공정 초기에 분석에서 제외되는 경우가 많다.

[0081] 양 리드가 단일 물리적 분자로부터 생기는 페어드 엔드 리드 중에서, 일부 페어드 리드는 100, 200, 300, 400, 500개 이하, 또는 500개 이상의 염기 쌍만큼 떨어진 영역으로부터 발생하는 한편, 일부 페어드 리드는 단일 물리적 분자 상에서 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100kb, 또는 100kb 이상으로 분리된 서열로부터 발생한다. 보통, 페어드 엔드 리드의 세트는 일부 리드 쌍이 자연적 배향의 핵산 세그먼트를 나타내도록 페어드 엔드 리드 거리의 분포를 포함한다.

[0082] 일부 실시양태에서, 본원에 개시된 방법은 유기체에 대한 염색체의 수와는 독립적인 방식으로 콘티그들을 그룹화한다. 본원에 개시된 다양한 방법들로 가능한 후속 스캐폴딩 결합을 사용하여, 생성된 더 작은 콘티그 클러스터를 스캐폴드로 어셈블링하기 위한 일부 관련 기법에서보다, 단일 링크 클러스터링용 콘티그-콘티그 링크에 대해 보다 보존적인 역치가 적용된다. 이러한 실시양태에 있어서 이점은 예측된 전체 염색체 길이 또는 수를 강행하기 위한 보상 에러에 의해 어셈블리 에러가 '누적'되지 않는다는 점이다. 염색체 크기 또는 수에 의존하는 방법을 사용하면, 올바른 위치에 어셈블링되거나 스캐폴딩된, 잘못배치된 콘티그에 의해 대체된 올바른 콘티그 또는 스캐폴드가 서열 정확도를 희생하여 전체 염색체 수 또는 길이를 보존하기 위해 제2 위치에 강제로 참여해야 하기 때문에, 잘못배치된 하나의 콘티그가 복수의 에러를 초래하는 경우가 많다.

[0083] 페어드 엔드 서열은 콘티그 정보, 및 일부의 경우, 완전한 인간 게놈 프로젝트에 대해 이용가능한 서열 정보와 같은 이전에 존재하는 스캐폴드 정보에 매핑된다. 페어드 엔드는 일부의 경우, 한 쌍 중 하나 또는 두개 모두가 단일 콘티그 또는 스캐폴드의 독특한 영역에 독특하게 매핑되도록 선택된다. 한 쌍 중 양자 리드가 하나의 콘티그 또는 스캐폴드 상에서 서로에 대해 측정될 수 있는 구별되는 위치에 독특하게 매핑되는 페어드 엔드 리드의 경우, 한 쌍에 있어서 리드 간 거리가 측정된다. 리드 쌍 거리 빈도 곡선은 이 데이터로부터 계산되며, 일부의 경우, 외삽하여 광범위한 리드 쌍 거리에 대해 주어진 리드 쌍 거리에 대한 빈도 추정치를 제공한다. 결과적으로, 예를 들어 제2 콘티그에 대한 한 콘티그의 2개 또는 2개 이상의 상대적 배향에 상응하는 2개 또는 2개 이상의 구별되는 리드 쌍 거리가 가능한 단일 리드 쌍의 경우, 누구나 어떤 콘티그 배향이 보다 확률이 있는 리드 쌍 거리에 상응하는지를 결정할 수 있다. 이와 유사하게, 복수의 리드 쌍이 한 쌍의 콘티그에 매핑되는 경우, 누구나 각 리드 쌍 개개에 대한 가장 확률이 있는 리드 쌍 거리와 세트에 대한 총 거리에 대해 결정할 수 있다. 일부의 경우, 복수의 리드 쌍 거리에 대한 콘티그 배향의 상대적 영향을 평가하고, 전체적으로 좀더 또는 가장 확률이 있는 리드 쌍 거리의 세트를 유도할 배향을 선택함으로써, 주어진 콘티그 쌍의 '전체적으로 가장 확률이 있는 배향'을 결정한다.

[0084] 일반적으로, 개개의 리드 쌍에 있어서, 그리고 대부분의 리드 쌍 거리 분포에 있어서, 더 짧은 리드 쌍 거리가 더 확률이 있다. 그러나, 복수의 리드 쌍 거리를 한 쌍의 콘티그 또는 복수의 콘티그에 대해 평가하는 경우, 가장 확률이 있는 리드 쌍 분포는, 일부의 경우, 리드 쌍 거리의 분포가 총 리드 쌍 길이의 단순한 최소화보다는 예측된 리드 쌍 분포를 가장 근접하게 반영하도록 짧고 긴 리드 쌍 거리를 모두 포함한다.

[0085] 도 2에서, 본원에 기술된 예시적인 리드 쌍 거리 빈도 곡선을 볼 수 있다. 데이터는 리드 쌍 거리의 함수로써 발생 빈도로 제시되며, 리드 쌍 거리가 증가함에 따라 '기하급수적으로' 또는 '대수적으로' 감소하는 것으로 관찰된다. 다른 데이터 서술들도 본원의 개시내용과 일치한다.

[0086] 일부의 경우, 리드 쌍 거리 곡선은 리드 쌍 데이터의 세트가 한 샘플에 대해 생성되는 경우에 독립적으로 결정된다. 다르게는 또는 조합하여, 이전에 측정된 리드 쌍 데이터를 사용하여 리드 쌍 거리 곡선을 생성한다. 일부의 경우, 이전에 생성되거나 또는 독립적으로 생성된 리드 쌍 곡선(들)을 사용한다.

[0087] 순서 정보 또는 배향 정보, 또는 순서 및 배향 정보를 갖는 스캐폴드에 콘티그를 어셈블링하기 위해, 한 쌍 중의 양자 리드가 구별되는 위치에 독특하게 매핑되지만 리드 쌍의 두 리드의 구별되는 위치가 하나의 콘티그 또는 스캐폴드 상에서 서로에 대해 측정될 수 없는 페어드-리드가 선택된다. 이 상황은, 예를 들어 리드 쌍의 구별되는 리드가 공통의 스캐폴드 상에 정확히 위치되지 않은 별도의 콘티그에 매핑되는 경우, 또는 서로에 대한 콘티그의 거리, 배향, 또는 거리 및 배향이 알려지지 않도록, 개별 데이터가 한 콘티그의 다른 콘티그에 대한 위치에 대해 달리 의구심을 제기하는 경우에 발생한다.

[0088] 임의로, 리드 쌍 서열은 불량한 품질의 리드가 제외되도록 콘티그 분석 전에 선별된다. 일부의 경우, 리드 쌍 서열은 단일 콘티그에 독특하게 매핑하지 않는 하나 이상의 리드를 갖는 리드 쌍이 제외되도록 콘티그 분석 전

에 선별된다. 일부의 경우, 리드 쌍 서열은 매핑되는 콘티그마다 단일 위치에 독특하게 매핑하지 않는 하나 이상의 리드를 갖는 리드 쌍이 제외되도록, 리드 쌍 서열 콘티그 분석 전에 선별된다. 일부의 경우, 미리 스캐폴딩된 서열 데이터 세트 또는 서열 어셈블리를 겪은 데이터 세트의 서술에 전반에 걸쳐 리드 쌍 '히트 (hit)' 내에서의 스파이크에 의해 확인된 바와 같이, 리드 쌍은 이들의 서열 쌍 중 하나 이상이 균형이 맞지 않은 수의 리드 쌍 서열이 매핑되는 영역에 매핑되는 경우 선별된다. 일부 선별 접근법에서, 스텝 크기는 예컨대 1000 bp로 규정되며, 각 스텝에서, 결합된 콘티그에 대한 지지력은 결합에 대한 지지력의 대부분을 구성할 수 있는 결합된 영역에 측접된 리드 쌍의 빈을 사용하거나 사용하지 않고 계산한다. 대안적인 스텝 크기, 예컨대 100, 200, 300, 400, 500, 600, 700, 800, 900, 1,500, 2,000, 3,000, 4,000, 5,000, 10,000 또는 10,000 초과도 고려된다. 빈이 산술평균, 중앙값 또는 평균 히트 분포에 기반한 역치 초과의 다수의 리드 쌍을 갖는 경우, 이는 제외된다. 일부의 경우, 역치는 산술평균, 중앙값 또는 평균 수치의 1.5x, 1.6x, 1.7x, 1.8x, 1.9x, 2x, 2.1x, 2.2x, 2.3x, 2.4x, 2.5x, 2.6x, 2.7x, 2.8x, 2.9x, 3x, 3.1x, 3.2,x, 3.3x, 3.4x, 3.5x, 3.6x, 3.7x, 3.8x, 3.9x, 4x, 4.5x, 5x, 또는 5x 초과가 되도록 선택된다. 서열 리드 선별에 대한 파라미터와 기준은 당업계의 숙련자들에게 공지되어 있으며, 본원에서는 다수의 선별 파라미터들이 고려된다. 일부의 경우, 가장 많은 수의 리드가 매핑되는 빈과 하나 이상의 염기가 중복되는 임의의 리드는 분석에서 제외된다.

[0089] 콘티그는 초기의 콘티그 위치결정을 생성하기 위하여 서로에 대해 그룹화된다. 콘티그 위치결정을 위한 다수의 접근법들이 본원에서 고려되며, 대안적인 실시양태들은 그들의 초기의 콘티그 순서 및/또는 배향의 선택에 있어서 다르다. 예를 들어, 일부의 경우, 콘티그는 사전에 생성된 초안 또는 완성된 게놈 스캐폴드 또는 염색체 맵 상에 매핑된다. 이러한 맵은 목표한 종에 대한 이전의 시퀀싱으로부터, 유의미한 게놈 규모의 신터니 (synteny)가 기대되거나 알려진 관련 종과 같은 밀접하게 관련된 종에 대한 이전의 시퀀싱으로부터 수득되거나, 또는 대표 게놈이 시퀀싱된 종의 개별 군집에 대하여 수득된다. 예를 들어, 네안데르탈인의 게놈은 인간 게놈 스캐폴드에 대해 매핑되고, 야생형의 가지과 부류의 일원은 재배종 토마토 (Lycopersicon esculentum) 게놈에 대해 매핑되며, 또는 특정 식물 재배종은 가장 근접하게 시퀀싱된 근친의 게놈에 따라 배열된다. 초기의 콘티그 위치결정에 대한 다른 방법들, 예컨대 당업계에 공지된 대안적인 방법 또는 상 결정에 대한 컴퓨터 시스템을 사용한 순서배열이 고려되며 본원의 개시내용과 일치한다.

[0090] 일부의 경우, 콘티그는 공유된 리드 쌍의 수에 따라 그룹화된다. 즉, 일부의 경우, 많은 수의 리드 쌍을 공유하는 콘티그는 서로 밀접하게 그룹화되는 반면, 소수의 리드 상을 공유하는 콘티그는 서로 상대적으로 멀리 떨어져 위치하게 된다.

[0091] 일부 국소적인 그룹화에 있어서, 콘티그들은 선형으로 배열된다. 그러나, 초기의 그룹화는 많은 분지형 또는 원형 콘티그 배향을 나타낸 경우가 많았고, 일부의 경우, 초기의 그룹화는 예를 들어 유전적 또는 물리적 연관 그룹에 대한 공지된 유전 정보, 또는 염색체 수와 관련하여 공지된 세포학적 정보를 기초로 예측된 실질적으로 큰 그룹을 포함한다.

[0092] 일부의 경우, 초기 그룹화는 콘티그의 공유된 페어드 엔드 리드로 나타내는 콘티그들 간의 관계의 강도에 대해 평가된다. 페어드 엔드 리드의 역치 값 미만의 콘티그들은, 일부의 경우, 예를 들어 공유된 페어드 엔드 리드의 수가 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15개 이하인 경우, 또는 일부의 경우, 15개 초과의 역치 수인 경우 분리된다.

[0093] 하나 이상의 리드 쌍을 공유하는 콘티그는 단일 물리적 연관 그룹으로 확인되어 그룹화된다. 일부의 경우, 콘티그는 적어도 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20개 또는 20개 이상의 페어드 엔드 리드를 공유하는 경우에 단일 물리적 연관 그룹으로 그룹화된다. 일부의 경우, 이 역치 값은 시퀀싱되는 핵산 샘플에 따라 다르다.

[0094] 일부의 경우, 이전에 생성되거나 또는 이전에 수득된 스캐폴드 정보, 또는 콘티그들 간에 공유된 리드 쌍의 총수, 또는 콘티그 길이 또는 독특한 콘티그 서열에 의해 정규화된 콘티그들 간에 공유된 리드 쌍의 총수를 기초로, 콘티그를 서로에 대하여 예비적으로 순서를 배열한다.

[0095] 공유된 리드 쌍의 원래 수를 계수하기 위한 대안으로서, 일부의 경우, 콘티그는 각 리드 쌍에 대한 리드 쌍 거리에 대해 평가되며, 확률 수치가 역치 미만인 거리를 갖는 리드 쌍은 공통의 리드 쌍에 대한 평가에서 제외된다.

[0096] 따라서, 초기 평가에서 콘티그의 쌍이 근접성 그룹화를 보증하는지에 대한 결정 (즉, 이들을 공통의 '엣지'로 잠정적으로 지정되는 것에 대한 여부)은 공유된 페어드 리드의 총수에 의해, 또는 역치 거리 확률을 갖는 공유

된 페어드 리드의 수에 의해 결정된다. 일부의 경우, 콘티그 엣지 결정에 대한 페어드 엔드 리드의 관련성에 대한 평가를 사용한다.

[0097] 콘티그들은 국소적으로 선형 배열된다. 일부의 경우, 국소적 선형 배열은 한 말단에서 다른 콘티그와 링크되기에 불충분한 리드 쌍을 갖는 브레이크포인트와 접해 있다. 일부의 경우, 국소적 선형 배열은 한 말단에서 분지점과 접해 있는데, 이로써 하나의 콘티그가 한 말단에서 2개의 별개 콘티그에 동등히 엄격하게 결합하게 된다. 일부의 경우, 한 말단에서 분지점과 접해 있는 콘티그의 선형 배열은 분지점에서 절단되어, 별도의 선형 그룹으로 취급된다. 다르게는, 일부의 경우, 분지점 접합부는 변형되거나 제2 세트의 어셈블리 조건 하에 분지들이 더 이상 동등하게 있음직할 확률이 있지 않도록 절단되어 대안적인 어셈블리 조건 하에 리어셈블링되며, 단일 분지는 콘티그들의 선형 배열을 지속하도록 선택된다.

[0098] 일부의 경우, 특히 짧은 콘티그와 관련된 짧은 분지(들)에 있어서, 분지는 분지점 내의 다른 엣지 콘티그 사이에 매핑되는 하나의 콘티그 또는 일련의 선형 콘티그를 나타낸다. 작은 콘티그는, 우연히 또는 그렇지 않으면, 많은 수의 리드 쌍 엔드의 맵 종착지가 될 확률이 더 적다. 결과적으로, 짧은 콘티그는 그들에게 매핑되는 하나 또는 소수의 리드 쌍 엔드만을 가질 수 있어서, 단 하나의 인접한 콘티그에 대한 연관성을 예측하게 된다. 이러한 상황은 분지점의 한 일원이 단독 콘티그이거나, 또는 상대적으로 긴 서열을 전체적으로 나타내지 않는 콘티그의 선형 그룹 또는 클러스터인 경우에 관찰되는 경우가 많다. 일부의 경우, 작은 분지점 콘티그 또는 콘티그의 선형 서열은, 리드 쌍 거리에 대한 국소적 영향이 유리한 경우에 작은 분지점 콘티그 또는 콘티그의 선형 서열이 분지점에 삽입되도록, 인접한 선형 콘티그 일련의 다양한 어셈블리에 대한 리드 쌍 거리 확률 및 리드 쌍 거리에 대한 영향에 대해 평가된다. 일부의 경우, 삽입된 콘티그 또는 콘티그 일련의 리드 쌍 엣지 스코어에 대한 상대적인 영향을, 삽입 부위와는 동떨어진 위치에 삽입되는 동일한 콘티그 또는 일련의 콘티그와 비교함으로써, 리드 쌍 거리에 대한 콘티그 또는 콘티그 일련 삽입에 대한 영향을 평가한다.

[0099] 초기의 콘티그 선형 그룹화에 대한 도해를 도 10A-D에 나타낸다. 콘티그는 원으로 나타냈고, 콘티그 쌍 엣지는 선으로 나타냈다. 도 10A에서는 초기 그룹화를 볼 수 있다. 초기의 그룹화는 일부의 경우, 상대적으로 엄격성이 낮아, 모든 콘티그들이 단일 클러스터로 링크된다. 도 10B에서, 콘티그 엣지들을 평가하여 콘티그, 스캐폴드 또는 노드 간의 낮은 엄격성의 연관이 제거된다. 이후의 높은 엄격성을 갖는 그룹은 연관 그룹 (염색체 또는 염색체 섹션)에 해당되거나 매핑되는 경우가 많다. 도 10C에서, 더 약하거나 실현확률이 덜한 잔류 엣지들을 절단함으로써 그래프 내의 원들을 제거하여 분지형 트리를 형성한다. 도 10D에서, 분지형 트리를 분지점에서 절단하여 선형 스캐폴드를 생성한다. 일부의 경우, 도 10D의 선형 스캐폴드 상에서 회귀적 윈도우 분석을 수행한다.

[0100] 도 11은 예시적인 최소 신장 트리를 나타내고 있다. 원들은 절단되었지만, 다수의 분지들을 클러스터로 나타내었다. 콘티그들은 각 타원형 내에 길이를 표시한 노드로 나타내었고, 각 콘티그 쌍 엣지에서의 수는 그 엣지를 지지하는 리드 쌍의 수를 가리킨다. 이 단계에서 대다수의 분지들은 작은 콘티그 또는 콘티그 그룹의 짧은 그룹에 해당됨이 관찰된다. 흔히, 이러한 분지들은 인접한 노드들 간에 매핑하기 위해서나, 또는 노드, 예컨대 이들이 인접하여 매핑되는 노드 내에 측정되지 않은 서열의 갭 또는 영역에 매핑하기 위한 추가의 분석에서 발견된다. 작은 스캐폴드, 콘티그 또는 노드가 분지로 나타나는 경우가 많은 한 가지 이유는 이들의 길이가 짧고, 이들이 각 인접한 노드에 매핑되는 리드 쌍 서열에 의해 타격을 받을 확률이 덜하기 때문이다.

[0101] 순서배열에 대한 본원에 개시된 방법 또는 대안적인 방법을 이용하여 콘티그들을 국소적으로 선형적 순서배열로 그룹화하는 경우, 콘티그들은 리드 쌍 거리 확률에 대한 국소적으로 완전한 평가를 이용하여 재배치된다.

[0102] 콘티그의 선형 그룹이 확인되고, '윈도우' 또는 국소적으로 인접한 콘티그의 하위그룹, 예컨대 콘티그의 선형 그룹의 한 말단에서의 하위그룹이 선택된다. 다르게는, 표적 '윈도우 크기'로 분할가능한 국소 클러스터 또는 그룹으로 그룹화되는 콘티그는, 이들의 그룹이 선형이 아니더라도 하기에서 분석된다. 콘티그의 선형 그룹의 전체 길이까지를 포함하는 2, 3, 4, 5, 6, 7, 8, 9, 10, 또는 10 이상의 다수의 윈도우 크기들이 본원의 개시내용과 양립가능하다. 일부 예시적인 실시양태에서, 상기 윈도우 크기는 3 또는 4이다. 일반적으로, 더 큰 윈도우 크기는, 궁극적인 순서배열이 평가된 콘티그 리드 쌍 특징 또는 측량 (예컨대, 예측된 리드 쌍 거리의 최대 우도, 예측된 우도 거리의 최소 전체 편차, 최소 전체 거리, 또는 다른 측량)과 관련하여 최적의 상태로 되는 더 큰 계산 신뢰도를 제공한다. 그러나, 더 큰 윈도우 크기는 실질적으로 더 긴 계산 시간을 필요로 하고, 일부의 경우, 또는 일부 컴퓨터 시스템 상에서, 더 큰 윈도우 크기는 시간 또는 비용으로 볼때 계산적으로 매우 비싸다. 3-4의 윈도우 크기는 추가된 윈도우 크기를 철저하게 분석하는데 필요한 추가의 계산 시간 없이 더 큰 윈도우의 스캐폴딩 정확도에 가까운 스캐폴딩 정확도를 달성하는 것으로 밝혀졌다.

[0103] 보통, 단일 윈도우 크기는 원하는 계산 작업량 또는 계산 시간에 맞추기 위해 선택되며, 전체 스캐폴드 또는 콘

티그 데이터 세트에 대하여 반복적으로 사용된다. 그러나, 다른 대안들도 본원에 개시된 방법과 일치한다. 윈도우 크기는 일부의 경우, 윈도우에 공통된 최소 또는 최대 또는 평균 목표 수의 매핑된 리드 쌍을 포함하기 위해 윈도우가 정의되도록, 데이터세트의 서로 다른 영역들에 따라 다를 것이다. 다르게는 또는 조합하여, 윈도우 크기는 일부의 경, 계산이 어려운 영역, 예컨대 반복 영역 또는 불량한 서열 품질을 갖는 서열이 보다 간단한 서열의 영역에 대하여 다소간의 계산 능력을 제공하도록, 기저 서열의 계산 복잡도를 반영하도록 선택된다. 일부의 경우, 윈도우 크기는 최소 또는 최대 범위의 콘티그 서열 길이 단독에 의하여, 또는 본원에서 언급된 요인과 조합하여 영향을 받는다. 특성이 잘 밝혀진 게놈에 있어서는, 윈도우 크기는 트랜스포존이 풍부한 영역, 말단소체, 동원체, 리보좀 반복 영역, 및/또는 복합 부위 또는 반복 부위에 맞추기 위해 임의로 변형된다.

[0104] 윈도우 내의 콘티그 세트의 각각의 리드 쌍에 있어서, 또는 윈도우 내의 콘티그 세트의 리드 쌍의 하위세트에 있어서, 서로에 대하여 해당 쌍이 매핑되는 2개의 콘티그에 대한 각각의 4개의 배향을, 하나 이상의 리드 쌍 간의 거리 및 상기 두 콘티그 간에 공유된 각각의 리드 쌍까지를 포함하여 이에 대한 주어진 배향이 갖는 거리 영향에 대하여 평가한다. 일부의 경우, 거리 조합의 확률은, 주어진 배향에 있어서, 콘티그의 다른 배향들에 대해 해당 배향의 총 상대적 확률을 누구나 평가할 수 있도록 결정된다. 본원에 개시된 일부 방법에서, 가장 높은 총 거리 확률을 산출해내는 배향이, 서로에 대한 콘티그 서열의 물리적 위치에 대하여 가장 그럴듯한 반영으로서 선택된다. 일부의 경우, 예측된 리드 쌍 분포의 최저 편차를 나타내는 리드 쌍 거리를 산출해내는 배향이 선택된다. 다른 순서배열 기준들, 예컨대 전체 리드 쌍 길이의 최소화, 또는 역치 길이 초과의 리드 쌍의 수 최소화, 또는 다른 기준들도 대안적인 실시양태에서 고려된다.

[0105] 본원에서 논의된 바와 같이, 일부 실시양태에서, 2개 이상의 콘티그의 그룹은 이들의 순서, 배향 또는 순서 및 배향에 대해 평가된다. 일부의 경우, 2, 3, 4, 5, 6, 7, 8, 9, 10개, 또는 10개 이상의 콘티그의 '윈도우'는 서로에 대한 이들의 배향, 순서 또는 배향 및 순서 배열을 결정하기 위해 동시에 평가된다. 실제적인 문제로서, 계산 시간이 주어진 윈도우 내에서 약 4개 초과의 콘티그 그룹화 분석을 못하게끔 좌절시킨다. 그러나, 상기 본원에서 논의된 변화에서와 같이, 단일 윈도우 내에서 한번에 다수의 콘티그를 분석하는 것도 고려된다.

[0106] 상기 윈도우 분석을 보다 자세히 살펴보면, 콘티그를 연결하고 있는 리드 쌍의 분리에 대한 확률 모델을 고려하여 콘티그의 구성을 개선하기 위해 본원에 개시된 방법 및 컴퓨터 시스템을 개발하여 시험하였다. 일부 방법은 콘티그 클러스터의 초기의 순서배열에 대해 윈도우 크기 w를 변화시키는 동적 프로그래밍 알고리즘에 관한 것이다. 복수의 위치 i에서 일부의 경우, 각각의 위치 i까지, 해당 위치는 윈도우 내에서 콘티그를 순서배열하여 배향시키는 $w!\,2^w$의 모든 방법을 고려하여, 상기 윈도우 내의 콘티그의 현재 구성으로 끝나는 현재 윈도우 위치의 말단까지 모든 콘티그에 대한 최적의 순서와 배향을 나타내는 스코어를 저장한다. 그렇게 하기 위해서, i-1, i-2, ... i-w 위치에서 윈도우의 "조화된" 순서와 배향에 대한 스코어를 보고, 현재의 구성으로 그들의 순서배열에 대한 확장성을 스코어링한다. $w!\,2^w$는 매우 가파른 함수이기 때문에, 본 방법은 사실상 작은 값의 w에 한정된다. 일부 데이터에 대한 시험에서, w=3은 일부 데이터세트에 대해 구성의 정확도를 크게 개선시킬 수 있다. 대안적인 윈도우 크기들도 고려되며, 불변 및 가변 윈도우 크기 선택을 위한 대안적인 근거들도 마찬가지이다. 윈도우에 대한 최적 또는 국소적 최대 스코어를 찾을 때, 콘티그들은 상기 최적 또는 국소적 최대 스코어를 달성할 정도로 스캐폴딩된다. 이후, 상기 윈도우는 한 위치를 전진하여 분석을 반복하게 된다.

[0107] 윈도우 분석은 초기의 콘티그 순서배열 방법에 대한 일부 실시양태와는 다른데, 이는 상기 윈도우 분석이 상기 초기 순서배열과는 달리 가능한 모든 콘티그 구성, 예컨대 콘티그 순서와 콘티그 배향 모두, 및 리드 쌍 길이 및 계산되거나 이전에 측정된 리드 쌍 길이 확률 곡선에 대한 리드 쌍 길이 확률에 대한 각 구성의 영향을 국소적으로 총망라하여 다루는 경우가 많기 때문이다. 즉, 일부 윈도우 분석에서는, 윈도우 내의 가능한 모든 콘티그 구성이 평가되고, 일부의 경우, 리드 쌍 길이에 대한 콘티그 구성의 가능한 모든 영향을 측정하여, 상기 콘티그는 또는 스캐폴드 순서배열과 배향에 반영된 어떤 리드 쌍 길이의 조합이 리드 쌍 길이 확률 곡선에 비추어 볼 때 국소적으로 가장 확률이 있는지에 대해 평가한다.

[0108] 두 콘티그 간에 공유된 리드 쌍의 세트에 대한 상대적인 전체 확률의 계산을 위해, 다수의 방법과 알고리즘을 이용할 수 있다. 본원에 개시된 한 예시적인 알고리즘은 하기와 같다:

[0109] $$L(l1, l2, g, o) = \frac{N!}{(N-n)!}(1-p_o)^{N-n}\prod_{i=1}^{n} f(d_i) \quad \text{(식 1)}$$

[0110] 상기 우도 함수는, 콘티그들이 상대적 배향이 o, $o \in ++, +-, -+, --$ 를 가지며 길이 g의 갭만큼 분리

되었다고 가정할 때, 콘티그 1과 2 사이에 걸친 리드 쌍 d_i의 개수 n과 정황상 추정할 때 분리된 것을 관찰할 확률을 제공한다.

[0111] 대안적인 알고리즘도 고려되며, 이것도 본원 개시내용과 일관된다. 일부 알고리즘은, 최소 전체 리드 쌍 거리를 평가하여 상기 최소 전체 리드 쌍 거리에 상응하는 배향을 선택하도록, 상대적 콘티그 배향에 대한 전체 리드 쌍 거리를 평가하는 단계를 수반한다. 다르게는 또는 조합하여, 예측된 리드 쌍 분포 패턴으로부터 주어진 콘티그 배향 세트에 대한 리드 쌍 거리의 변화를 평가하고, 이러한 변화를 최소화하는 하나의 콘티그 배향을 선택한다.

[0112] 위에서 논의한 바와 같이, 일부 실시양태에서 2, 3, 4, 5, 6개, 또는 6개 이상의 콘티그의 윈도우는 상대적 순서, 배향 또는 그의 성분 콘티그의 구성과 같은 스캐폴딩에 대해 평가된다. 외적인 제약에 해당하는 그룹 내의 두 콘티그에 공통인 리드 쌍에 대한 총 상대 거리 빈도, 또는 예측된 리드 쌍 거리의 발생 확률을 최대화하는 순서에 비추어 볼 때 최소화하거나 감소하는 2, 3, 4, 5, 6개, 또는 6개 이상의 콘티그의 세트에 대한 순서, 배향 또는 구성을 포함하는 스캐폴드를 결정할 때, 그 그룹 내의 콘티그들은 그에 따라 순서가 배열되고 배향된다.

[0113] 주어진 윈도우 내에서 콘티그의 세트에 대해 순서배열, 배향 또는 순서배열과 배향을 할 때, 해당 순서에서 최적으로 측정된 하나의 말단 콘티그를 제외시키고, 잔류 콘티그들을 본원에서 논의된 바와 같이 콘티그들의 예비적 순서배열을 기초로 하여 인접한 콘티그를 나타낸다고 추정되는 것으로 재차 확인된 새롭게 추가한 한 콘티그와 함께 다시 분석한다.

[0114] 이 방법은 예를 들어 추정상의 콘티그 순서의 한 말단으로부터 다른 말단까지 한방향으로 반복된다. 대안적인 실시양태에서, 윈도우 배향은 배열된 콘티그 세트 내의 무작위적 위치에서 시작하여 한 콘티그 엣지를 향하여 한방향으로 진행하거나 양 엣지를 향하여 두방향으로 순차적으로 진행하거나, 또는 스캐폴드의 양 말단에서 시작하여 내부 합류점까지 진행한다.

[0115] 대안적인 실시양태에서, 윈도우는, 한 콘티그의 배제와 한 콘티그의 첨가이 동시적 수행을 통하기보다는, 2개 또는 3개의 인접한 콘티그들이 배열된 콘티그 윈도우의 한 쪽에서 제거되어 이에 상응하는 개수가 다른 쪽에서 첨가되도록, 한 스텝에 2개 또는 3개의 콘티그만큼 '이동'한다.

[0116] 윈도우 분석은, 한 그룹의 모든 콘티그들까지 대다수가 본원에서 선택된 파라미터를 최소화하거나 최대화하기 위하여 순서배열이 될 때까지 지속된다. 상기 윈도우 분석에 대한 많은 실시양태의 공통적인 특징은, 배향을 생성하는 국소적 최대 또는 최소 파라미터가 철저히 검사될 수 있도록, 그리고 일부의 경우, 리드 쌍 거리와 관련된 국소적 최대 또는 최소 파라미터가, 콘티그 그룹 중의 각각의 콘티그까지를 포함하는 윈도우 콘티그 그룹의 하위세트에 대해 확인될 수 있도록, 주어진 윈도우 내의 콘티그들에 대해 모든 배향들이 모두 철저히 분석된다는 점이다.

[0117] 부분적으로 완전하거나 완전한 윈도우 분석으로부터 생성된 상기 콘티그 배향은 일부의 경우, 단일 스캐폴드 또는 데이터 세트 크기의 '윈도우' 내의 모든 콘티그들을 철저히 분석함으로써 생성될 수 있는, 전반적으로 최적화된 콘티그 배향과 동일하거나 실질적으로 이와 유사한 콘티그 배향을 나타낸다. 그러나, 본원의 윈도우 분석을 사용함으로써, 국소적으로 최적인 콘티그 배향에 이르기 위해서 분석될 콘티그 배향 공간이 실질적으로 덜 필요하고, 국소적으로 최적인 배향에 이르기 위해서 사용할 계산 시간 또는 계산 용량이 실질적으로 덜 필요하게 된다.

[0118] 일부 실시양태에서, (i) 공지된 서열 라이브러리로부터 관찰된 리드 쌍 데이터 분포를 기초로 인서트 분포의 모델을 구축하고, (ii) 생성된 리드 쌍들을 상기 모델에 대하여 개별적으로 순위를 매긴 후, (iii) 상기 모델에 대한 비교를 바탕으로, 리드 쌍에 스코어를 부여하여, (iv) 높은 스코어를 갖는 리드 쌍을 상기 모델에 가장 근접하게 들어맞는 것으로 한다.

[0119] 일부 실시양태에서, (i) 이전에 매핑되지 않은 서열 라이브러리로부터 관찰된 리드 쌍 데이터 분포를 기초로 인서트 분포의 모델을 구축하고, (ii) 생성된 리드 쌍들을 상기 모델에 대하여 개별적으로 순위를 매긴 후, (iii) 상기 모델에 대한 비교를 바탕으로, 리드 쌍에 스코어를 부여하여, (iv) 높은 스코어를 갖는 리드 쌍을 상기 모델에 가장 근접하게 들어맞는 것으로 하고, (v) 스캐폴드 내에서 보다 확률이 있는 배열을 생성하는 스코어를 내기 위해 상기 모델을 변형시킨다.

[0120] 일부의 경우, 한 윈도우 (w)가 한번에 1, 2, 3, 4개, 또는 그 이상의 콘티그를 보고, 중복되는 윈도우만이 임의

의 한 시점에 보여지는 회귀적 알고리즘이 사용된다. 알고리즘의 회귀적 특성으로 인해, 상기 윈도우는 스캐폴드 또는 기타 그룹화된 핵산 데이터세트 상의 콘티그들을 통해 점차적으로 나아가, 상기 윈도우는 종말점으로 나아가거나 또는 스캐폴드 내의 출발 장소로 되돌아 온다.

[0121] 다수의 프로토콜을 통해 생성된 페어드 엔드 리드는 본원의 분석 및 컴퓨터 실행 시스템과 일관성이 있다. 일부 실시양태에서, 당업계에 공지된 DNA 추출 방법을 비롯한 DNA 추출 프로토콜이 사용된다. 일부 실시양태에서, 시판되는 DNA 추출 키트가 사용된다. 시판되는 예시적인 DNA 추출 키트로는 퀴아젠 블러드 앤 셀 미디 (Blood and Cell Midi) 키트를 포함한다. 일부 실시양태에서, DNA 추출을 위한 출발 조직은 인간의 체액 또는 조직이다. 일부 실시양태에서 DNA 추출을 위한 샘플은 인간이 아닌 동물, 식물 또는 진균류로부터 유래한다. 일부 실시양태에서, DNA의 공급원은 미생물 또는 바이러스로부터 유래한다.

[0122] 일부 실시양태에서, 세포 핵은 세포 용해 및 원심분리 프로토콜, 또는 당업계에 공지된 임의의 다른 방법을 사용하여 세포로부터 분리된다. 일부 실시양태에서, 핵을 효소로 처리하여 DNA가 아닌 세포성 성분들을 분해시킨다. 예시적인 효소로는 프로테이나아제 K 및 리보뉴클레아제 (RNAase) A를 포함한다. 일부 실시양태에서, 세포성 또는 바이러스성 DNA는 당업계에 잘 알려진 방법을 사용하여 정제 및 분리된다. 예시적인 키트는 DNA가 이소프로판올 중에서 세척, 용출 및 침전되고 원심분리에 의해 펠렛화되는 퀴아젠 게놈 컬럼을 들 수 있다. 건조 후, 펠렛을 200 ㎕의 TE (퀴아젠) 중에 재현탁시킨다. 대안적인 실시양태에서, 핵산은 전체 세포로부터, 또는 복수의 공급원으로부터 유래한 다수의 세포형 또는 핵산을 포함하는 샘플로부터 분리된다. 일부 실시양태에서, DNA는 *디 노보* 합성된다.

[0123] 일부의 경우, 페어드 엔드는 이중 가닥 분해가 이루어지는 동안 물리적 연관 정보를 보존하기 위하여 재어셈블링된 염색질의 쪼개진 클러스터에 대한 재결찰을 통해 생성된다. 염색질은 당업계에 공지된 임의의 방법을 사용하여 정제된 DNA를 가지고 재어셈블링된다. 예를 들어, 염색질은 액티브 모티프 (Active Motif)의 *시험관내* 염색질 어셈블리 키트를 사용하여 게놈 DNA로부터 27℃에서 밤새 어셈블링된다. 추가의 실시양태에서, 항온배양 후 성공적인 염색질 어셈블리를 확인하기 위해 시험을 수행한다. 예시적 시험은 성공적인 염색질 어셈블리를 확인하기 위해 마이크로코커스 핵산가수분해효소 (MNase) 처리를 위해 10%의 샘플을 사용하는 것이다. '재어셈블링된 염색질'은 생물학적 염색질 성분 예컨대 히스톤 또는 뉴클레오솜, 또는 기타 핵 단백질, 예컨대 전사 인자, DNA 결합 단백질, 유전자전위효소, 또는 핵산 결합과 관련된 기타 핵 단백질의 재구축과, 핵산 분자 상에 비단백질성 나노입자를 첨가하여 생성된 것과 같은 인공적으로 재구성된 염색질을 모두 지칭하는데 매우 광범위하게 사용된다.

[0124] 염색질은 일부의 경우, 분류 시약에 대한 상기 염색질의 결합을 촉진하거나 결합된 쌍의 분리를 촉진하는 시약으로 표지된다. 한 실례에서, 표지 시약은 비오틴이다. 보다 구체적인 실례에서, 염색질은 요오드아세틸-PEG-2-비오틴 (IPB)으로 비오틴화된다. 일부 실시양태에서, DNA와 염색질의 복합체를 고정화 시약을 사용하여 고정화시킨다. 일부의 경우, DNA-염색질 복합체를 실온 (RT)에서 15분간 1%의 포름알데히드 중에 고정시킨 다음, 2배 몰 초과량의 2.5M 글리신으로 냉각시킨다. 일부 실시양태에서, DNA-염색질 복합체는 DNA 단편을 생성하는 반응을 거친다. 일부 실시양태에서, DNA-염색질 복합체는 제한 효소로 처리된다. 일부의 경우, DNA는 MboI 또는 MluCI를 사용하여 처리된다.

[0125] RT에서 15분간 유리 비오틴의 존재 하에 비드를 항온배양함으로써 결합되지 않은 스트렙타비딘 부위를 임의적으로 점유한다. 일부 실시양태에서, dNTP, 예를 들어 a-S-dGTP 및 비오틴화된 dCTP와 함께 항온배양함으로써 점착성 말단을 충전시켜 넣는다. 일부 실시양태에서, 결찰 단계를 수행하여 dNTP 충전에 의해 생성된 평활 말단을 결합시킨다. 일부 실시양태에서, DNA를 엑소뉴클레아제로 처리하여 비오틴화된 유리 말단을 제거한다. 일부의 경우, 상기 엑소뉴클레아제는 엑소뉴클레아제 III (#M0206S, NEB)이다.

[0126] 일부 실시양태에서, DNA는 전단 처리를 거친다. 추가의 실시양태에서, 전단 처리된 DNA는 클레노브 폴리머라아제 및 T4 PNK를 사용하여 충전된다. 일부 실시양태에서, 충전 반응 후, 예를 들어 붕괴 반응에 의해 DNA가 풍부해진다.

[0127] 일부 실시양태에서, 제한 효소 절단 인식 부위 선택으로 생성된 접합부에 의해 서열 리드들을 규정한다. 예를 들어, MboI 또는 MluCI가 사용되는 경우, 서열 리드들 접합부가 존재할때마다 이를 절두시킨다 (GATCGATC (MboI), AATTAATT (MluCI)). 일부 실시양태에서, 리드들은 이후 SMALT [http://www.sanger.ac.uk/자원/ 소프트웨어/smalt/]를 사용하여 순방향 및 역방향 리드들을 독립적으로 정렬하기 위한 -x 옵션으로 정렬된다. 일부 실시양태에서, PCR 중복 서열을 피카르드-툴 (Picard-tool)인 마크듀플리킷 (MarkDuplicate) [http://broadinstitute.github.io/ picard/]을 사용하여 마킹한다. 일부 실시양태에서, 양자 리드가 모두 매

핑되어 10 초과의 매핑 품질을 갖는 경우, 중복되지 않은 리드 쌍을 분석에 사용한다.

[0128] **스캐폴딩 - 입력 전처리**

[0129] 예를 들어, 컴퓨터 실행 시스템에 대한 입력 전처리를 윈도우 분석 전에 또는 윈도우 분석과는 독립적으로 임의로 사용한다. 리드 쌍 군집을 청소하여 그것이 독특하게 매핑되는 서열을 포함하도록 하기 위하여, 어셈블리 또는 콘티그 세트 또는 스캐폴드 세트의 고반복 영역에 매핑되는 페어드 리드를 추가의 분석에서 제거한다. 일부의 경우, 어셈블리에 대한 전장 게놈 샷건 리드의 정렬을 사용하여 이러한 영역들을 확인한다. 다르게는 또는 추가로, 역치를 초과하는 매핑된 샷건 리드 깊이를 갖는 출발 어셈블리의 간격으로 매핑되는 리드 쌍은 제외된다. 일부 영역들이 리드 쌍에 대한 '핫 스팟'이고, 이러한 핫 스팟 영역의 쌍 데이터를 포함시키는 것은 다운스트림 분석을 편향되게 할 수 있다는 점이 관찰된다. 일부의 경우, 간극이 '트리거 (trigger)' t2를 초과하는 매핑된 리드 거리를 갖는 하나 이상의 염기를 포함하게 되면, 그 간극에 매핑되는 '컷오프 (cutoff)' t1을 초과하는 깊이를 갖는 모든 리드들을 제외시키도록, 2점 역치 방법을 사용한다. 일부의 경우, t1 또는 t2를 초과하는 매핑된 리드 깊이를 갖는 하나 이상의 염기를 포함하는 t1을 초과하는 매핑된 샷건 리드 깊이를 갖는 출발 어셈블리 모든 간극들을 확인하여 제외시키도록, 이중 역치 전략이 사용된다. 일부 예시적인 실시양태에서, t1 및 t2는 약 0.5%의 어셈블리가 차폐되도록, 또는 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 40, 50, 60, 70, 80, 또는 80% 이상의 어셈블리가 차폐되도록 선택된다. 일부 실시양태에서, t1 및 t2는 약 0.5%의 어셈블리가 차폐되도록 설정된다. 일부의 경우, 차폐될 어셈블리의 백분율은 반복적인 어셈블리의 비율에 의해 영향을 받는다. 일부 실시양태에서, 매핑된 샷건 리드의 깊이에 대한 역치 설정은 t1은 3x이고 t2는 3.5x (여기서, x는 깊이의 전체 분포에 대한 평균과 같음)이다. 예를 들어, 인간의 특정 어셈블리의 경우, 102의 역치가 달성되면 87 또는 그 이상의 깊이에 대한 히트를 갖는 영역들이 차폐되도록, t1은 87이고, t2는 102이다. 이러한 실례에서, 'x'는 29이고, 3x는 87이며, 102는 3.5x이다. 일부 실시양태에서 t1은 2x, 2.0x, 2.1x, 2.2x, 2.3x, 2.4x, 2.5x, 2.6x, 2.7x, 2.8x, 2.9x, 3.0x, 3.1x, 3.2x, 3.3x, 3.4x, 3.5x, 3.6x, 3.7x, 3.8x, 3.9x, 4.0x 미만, 또는 4.0x 초과로부터 선택된다. 일부 실시양태에서, t2는 2x, 2.0x, 2.1x, 2.2x, 2.3x, 2.4x, 2.5x, 2.6x, 2.7x, 2.8x, 2.9x, 3.0x, 3.1x, 3.2x, 3.3x, 3.4x, 3.5x, 3.6x, 3.7x, 3.8x, 3.9x, 4.0x, 4.1x, 4.2x, 4.3x, 4.4x, 4.5x, 4.6x, 4.7x, 4.8x, 4.9x, 5.0x 미만, 또는 5.0x 초과로부터 선택된다.

[0130] 일부의 경우, 리드 쌍은, 이들이 2개 이상의 리드 쌍 링크에 의해 4개 이상의 다른 입력 콘티그들에 결합되는 어셈블리 상의 1 Kbp 윈도우 내의 범위에 포함되는 경우에 제외된다. 일부의 경우, 제외되는 윈도우는 100bp 미만, 100bp, 200pb, 300pb, 400pb, 500pb, 600pb, 700pb, 800pb, 900pb, 1kb, 1.1kb, 1.2kb, 1.3kb. 1.4kb. 1.5kb, 2kb, 3kb, 4kb, 5kb 미만, 또는 5kb 초과이다. 일부의 경우, 제외되는 윈도우는 해당 영역이 2, 3, 4, 5, 6, 7개, 또는 7개 이상의 다른 입력 콘티그에 연결되는 경우에 촉발된다.

[0131] **스캐폴딩 - 확률 모델 파라미터에 대한 평가**

[0132] 일부의 경우, 입력 처리 전에 확률 모델 파라미터를 평가한다. 본원에 개시된 방법 및 시스템의 여러 단계들은 리드 쌍 데이터의 확률 모델을 사용하여 어셈블리 결정을 안내하거나 스캐폴드 내의 콘티그 구성을 최적화할 수 있다. 일부 실시양태에서, 우도를 사용하여 어셈블리 결정을 안내하거나 스캐폴드 내의 콘티그 구성을 최적화한다. 일부의 경우, 우도 함수는 아래와 같다:

[0133] $$L(l1, l2, g, o) = \frac{N!}{(N-n)!}(1-p_o)^{N-n}\prod_{i=1}^{n} f(d_i) \quad \text{(식 1)}$$

[0134] 상기 우도 함수는, 콘티그들이 상대적 배향이 o, $o \in ++, +-, -+, --$ 를 가지며 길이 g의 갭만큼 분리되었다고 가정할 때, 콘티그 1과 2 사이에 걸친 리드 쌍 d_i의 개수 n과 정황상 추정할 때 분리된 것을 관찰할 확률을 제공한다. 함수 $f(x)$는 리드 쌍의 게놈 분리 거리에 대한 정규화 확률 분포이며, 핵산 샘플을 독립적으로 표본으로 삼는 "노이즈" 쌍으로부터 기여도를 가지는 것으로 추정된다. $f(x) = p_n/G + (1 - p_n)f'(x)$이고, $f'(x)$ (식 2)는 지수적 분포의 총합으로 나타낸다. 이러한 파라미터들 중 하나 이상에 대한 평가를 포함하는 대안적인 함수도 고려되며, 이것도 본원 개시내용과 일관된다.

[0135] 일부 실시양태에서, 이용가능한 출발 어셈블리가 제한된 근접성을 가지는 경우 N, p_n, G 및 $f'(x)$에 대한 신뢰성 있는 추산치를 수득하기 위하여, 콘티그 쌍의 샘플에 대해 링크의 밀도 (n/l_1l_2로 정의됨)의 도표화하고, 최

고 밀도값 1%와 최저 밀도값 1%를 배제시키며, 하기 식을 셋팅하고

[0136] $N_n = G^2 \sum n_{ij} / \sum l_i l_j$ (식 3), G의 값으로 입력 콘티그 길이의 총합을 이용하여 생성물인 "노이즈" 쌍의 총수 N, p_n를 추정하였다. 일부의 경우, 남은 파라미터들을 분리 x에서 고르게 한 계수에 대해 $G(\sum_{i=1}^{N_c} min(0, l_i - x))^{-1}$ (식 4)의 승법적 보정계수를 적용한 후, 출발 어셈블리 콘티그에 매핑된 리드 쌍에 대해 관찰된 분리의 히스토그램에 대한 최소제곱법으로 N $f(x)$ 내에 고정시킨다. 본원의 개념에 대한 대안적인 방정식 및 수학적 표현들도 본원의 방법 및 시스템의 일부 실시양태들과 일관성이 있는 것으로 고려된다.

[0137] **스캐폴딩 - 메타게놈 확률 모델 파라미터**

[0138] 일부 예시적인 실시양태에서, 입력 데이터는 다수의 공급원 (예컨대, 메타게놈 라이브러리)으로부터 유래하는 핵산의 혼합물을 포함하는 샘플로부터 수득한다. 일부 이러한 상황에서, 본원에 개시된 수개의 방법 및 컴퓨터 시스템에 사용된 확률 모델은, 혼합된 핵산 샘플 또는 메타게놈 라이브러리로부터 유래하는 입력 데이터를 처리하기 위해 변형된다. 일부의 경우, 우도 스코어는 각각의 길이 (예컨대, l_1 및 l_2) 및 계수 (예컨대, s_1 및 s_2)를 갖는 두 단편에 대하여 계산되는 것으로 가정한다. 일부의 경우, 계수는 단편의 길이 및 입력 혼합 핵산 샘플 중의 이의 상대적 빈도를 갖는 생성물에 대략적으로 비례하는 임의의 양이다. 비제한적인 예로서, s_1 및 s_2는 적절한 시퀀싱 라이브러리의 각 콘티그에 매핑되는 리드의 수일 수 있다. 일부 실시양태에서, 우도 스코어는 예측된 수의 노이즈 리드 및 리드 쌍을 처리하기 위해 변형된다. 일부의 경우, 우도 스코어 $S = \sum_i \ln \frac{f(x_i)}{f(\infty)} - (\bar{n} - \bar{n}_0)$ (식 5)는 예측된 수의 노이즈 리드 $\bar{n}_o = Np\frac{l_1 l_2}{G^2} \rightarrow Np\frac{s_1 s_2}{S_T^2}$ (식 6) 및 리드 쌍 $\bar{n} = N\frac{1}{G}\int_A f(x)dxdy$ (식 7)을 위해 변형된다. 일부의 경우, 리드 쌍의 유효 총수가 계산된다. 일부의 경우, 리드 쌍의 유효 총수는 $N \rightarrow N\left(\frac{s_1+s_2}{l_1+l_2}\right)\left(\frac{G}{S_T}\right)$ (식 8)이다. 일부의 경우, 결합될 두 단편에 매핑된 리드 쌍의 밀도 간 차이가 증가함에 따라 우도 스코어를 감소시키는 스코어 패널티와 같은, 스코어 페널티가 적용된다. 비제한적인 예로서, 페널티는 하기와 같이 계산된다:

[0139] $$\Delta \ln \frac{L(H_1)}{L(H_0)} = \frac{Binom(s_1|s_1+s_2, \frac{l_1}{l_1+l_2})}{Poisson(s_1|\lambda = s_1)Poisson(s_2|\lambda = s_2)}$$ (식 9).

[0140] **스캐폴딩 - 입력 콘티그 내의 낮은 지지력 결합의 절단**

[0141] 입력 어셈블리 데이터는, 예를 들어 상대적으로 약한 어셈블리 결정을 다운스트림 분석에 존속하지 않도록, 임의로 처리하여 콘티그, 스캐폴드 또는 어셈블리 정보를 부분적으로 해체한다. 예를 들어, 출발 어셈블리 내의 후보 부정합을 확인하여 분해하기 위해, 출발 어셈블리 내의 각 콘티그의 각 위치 i의 왼쪽과 오른쪽을 결합함으로써 얻어진 로그 우도 변화를 계산하기 위해 우도 모델을 사용한다 (예컨대, 위치 i에서의 분해에 의해 생성될 수 있는 두 콘티그에 대한 로그 우도비 (LLR) L_i = ln $L(g = 0)$=$L(g = 8)$임). 일부 실시양태에서, 이러한 지지력이 입력 콘티그의 최대 내부 세그먼트에 대한 역치 값 t_b 미만으로 떨어지는 경우, 상기 세그먼트는 "낮은 지지력의" 세그먼트로 정의된다. 일부 실시양태에서, 예를 들어, 서로에 대해 300bp 이내에 있는 낮은 지지력의 세그먼트를 병합하고, 예를 들어 1 Kbp의 한 콘티그 엔드 내의 것들을 제외시킨 후, 상기 세그먼트 크기에 따라 추가한 변형을 수행한다. 예를 들어, 1000bp 아래의 세그먼트에 대해서는 세그먼트의 중앙 지점에 절단부를 콘티그 내에 도입하거나, 또는 세그먼트가 1000 bp보다 긴 경우에는 절단부를 세그먼트의 각 말단에 도입한다. 일부의 경우, 100bp, 100bp, 200pb, 300pb, 400pb, 500pb, 600pb, 700pb, 800pb, 900pb, 1kb, 1.1kb, 1.2kb, 1.3kb. 1.4kb. 1.5kb, 2kb, 3kb, 4kb, 5kb보다 길거나, 또는 5kb를 초과하는 경우, 절단부를 세그먼트의 각 말단에 도입한다.

[0142] **스캐폴딩 - 콘티그-콘티그 연관 그래프 구축**

[0143] 어셈블리 공정을 수행하는 동안, 생성된 연관 데이터는 임의로 그래프로 나타내는데, 이 그래프에서는 출발 어셈블리의 (절단된) 콘티그가 노드이고, 엣지는 각각 매핑된 쌍의 리드의 두 콘티그 내의 위치를 나타내는 배열

된 정수 쌍의 목록으로 표지된다. 일부 실시양태에서, 스캐폴딩의 초기 단계는, 생성된 링크 t_L의 역치 수보다 적은 엣지를 제외시킴으로써 그래프를 연결 요소들로 나누어 생성된 데이터의 하위세트에 대해서도 동시에 수행되는데, 여기서, 임의의 연결 요소를 발생시키지 못한 최저 정수 역치 값은 5% 이상의 입력 콘티그를 포함하였다. 특정 데이터 세트 또는 분석 시스템의 세부 사항들에 따라, 역치 t_L은 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20 미만 또는 대략 이 치수 미만이거나, 또는 20% 초과의 입력 콘티그를 제외하도록 선택된다.

[0144] **스캐폴딩 - 씨드 스캐폴드 구축**

[0145] 일부 실시양태에서, 스캐폴드 구축의 반복되는 상은 콘티그-콘티그 그래프의 엣지를 필터링하는 단계 및 이를 고신뢰도의 선형 서브그래프로 분해하는 단계에 의해 그 토대가 만들어진다. 일부의 경우, 상기 콘티그-콘티그 엣지는 필터링되어 상기 필터링된 그래프의 최소 신장 포리스트 (forest)가 발견된다 (하기의 "엣지 필터링" 참조). 일부의 경우, 상기 그래프는 1차수 노드를 3회 연속적으로 제거한 후, 2차수 초과의 노드를 제거함으로써 선형으로 된다. 일부의 경우, 생성된 그래프의 각 연결 요소들은 선형 위상학 (topology) 가지며, 입력 콘티그의 하위세트의 순서를 규정하였다. 일부의 경우, 초기 스캐폴드의 생성에 있어서 최고조의 단계는 각 선형 요소에 대한 콘티그 배향의 최대 우도 선택을 찾는 것이다. 일부 실시양태에서, 그래프는 1, 2, 4, 5, 6번, 또는 6번 이상의 연속적인 노드 제거에 의해 선형으로 된다. 일부 실시양태에서, 제거된 노드의 차수는 다양하다. 일부의 경우, 최대 우도 선택은 최대 우도를 결정하는 특정 방정식을 사용하여 산출된다. 일부의 경우, 최대 우도는 가장 확률이 있는 순서, 배향 또는 순서 및 배향에 대한 일반적인 평가이다.

[0146] **스캐폴딩 - 엣지 필터링**

[0147] 필터는 선형화 전에 콘티그-콘티그 그래프의 엣지에 임의로 적용된다. 예시적인 필터로는 하기를 포함한다: t_L 링크보다 작은 엣지의 제외 및 "잡다한" 콘티그 엣지의 제외. "잡다한" 콘티그들은 염기쌍 내의 콘티그 길이에 대한 상응하는 노드의 그래프에서 차수의 비율이 t_p,를 초과하거나, 또는 d_m 이상의 다른 콘티그에 대한 필터를 지나치는 링크 (1)를 갖는 것들로 확인된다. 역치 t_p 및 d_m 은 일부의 경우, 해당 수치의 분포 중 약 5%의 상단 꼬리 (upper tail) 부분을 제외하도록 선택된다. 일부의 경우, 역치 t_p 및 d_m 은 해당 수치의 분포 중 약 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20% 미만, 또는 20% 초과의 상단 꼬리 부분을 제외하도록 선택된다. 일부 실시양태에서, t_L은 7 미만, 7, 8, 9, 10, 11, 12, 13, 14, 15 까지, 또는 15 이상의 범위이다. 일부 실시양태에서 t_p는 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1 까지이다. 일부 실시양태에서 d_m은 5 미만, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 또는 15 이상의 범위이다. 일부 실시양태에서, t_L, t_p 및 d_m이 각각 11, 0.04 그리고 10인 경우에 개선된 성능이 얻어지는 것으로 밝혀졌다.

[0148] **스캐폴딩 - 콘티그 배향**

[0149] 각각의 입력 스캐폴드는 순방향 및 역방향의 염기 서열, 또는 "왓슨"과 "크릭" DNA 가닥에 해당하는 최종 어셈블리 내의 두 가지 배향 중 하나를 가진다. 본원에 개시된 바와 같이, 각 선형 스트링 내에서 스캐폴드에 대한 최적의 배향은 하기 나타낸 바와 같이 회귀적 관계를 사용하는 동적 프로그래밍과 같은 분석에 의해 찾는다. 길이 n의 스캐폴드의 차례 목록에서, 스캐폴드 $i-k$ 내지 i가 특정 배향 o_{i-k}, o_{i-k+1}, ...o_i 를 가지도록, 스캐폴드 i까지의 스캐폴드에 대한 배향 선택의 최고의 스코어링 서열의 스코어는 하기에 의해 주어진다:

[0150] $$S_m(i, o_{i-k}, o_{i-k+1}, \ldots, o_i) = \max_{o_{i-1-k} \in [+,-]} (S_m(i-1, o_{i-1-k}, o_{i-k}, \ldots, o_{i-1}) + \sum_{j=i-i-k}^{j=i-1} \log p(o_j, o_i))$$

[0151] (식 10)

[0152] 임의로, k 만큼 후퇴한 콘티그의 링크를 포함하는 것은 배향 정확도에 있어서 개선을 제공한다. 일부의 경우, 배향 정확도의 개선은, 위에서 논의한 바와 같이, 분지된 콘티그 순서를 초래하는 것과 유사한 상황에서, 이들 뛰어넘는 링크에서 비롯되는 측부 스캐폴드에 대한 중요한 배향 정보와 함께, 삽입된 작은 스캐폴드가 연결 및 그에 따른 한쪽에 대한 배향 정보만을 가질 것이기 때문에 발생한다.

[0153] **스캐폴딩 - 성분들 내에서 스캐폴드 병합**

[0154] 콘티그 엔드들은 스캐폴드 내의 그들의 위치에 대해 임의로 지정된 분류이다. 예를 들어, 콘티그 엔드는 이들이 스캐폴드의 말단에 놓인 경우에는 "유리 상태"로 분류되거나, 또는 이들이ㅡ 스캐폴드 내부에 존재하는 경우 "묻혀있는 상태"로 분류된다. 일부 실시양태에서, 각 연결 요소 내의 모든 콘티그 엔드의 쌍에 있어서, 이들의 결합에 대한 LLR 스코어는 g_0의 "표준" 갭 크기를 사용하여 계산된다. 일부 실시양태에서, 후보 결합부는 내림차순의 스코어로 분류되어 기준의 세트에 따라 평가된다. 기준의 예시적인 세트는 아래와 같다. 양 엔드가 모두 유리 상태로 서로 다른 스케폴드에서 유래하는 경우, 두 스캐폴드를 한줄로 연결하는 시험을 한다. 한쪽 말단이 묻혀있고, 다른 말단이 유리 상태이며, 상기 말단이 서로 다른 스캐폴드에서 유래하는 경우에는, 유리 말단의 스캐폴드를 묻혀있는 말단 내에 인접한 갭 내로 삽입하는 시험을 한다. 하나 또는 양 말단이 모두 묻혀있고 엔드들이 동일한 스캐폴드 상에 존재하는 경우에는, 두 말단 사이의 스캐폴드의 일부를 역전시키는 시험을 한다. 양 말단이 모두 묻혀있고, 서로 다른 스캐폴드로부터 유래하는 경우, 스캐폴드를 한줄로 결합시키는 모든 4가지의 방법을 시험한다. 일부 실시양태에서, 모든 경우에 있어서, 변화에 의해 영향을 받는 모든 콘티그의 쌍 간의 LLR 스코어를 합하여 LLR 스코어에서의 전체 변화를 계산함으로써 가능한 결합, 삽입 및 전위를 시험한다. 상기 변화가 LLR 스코어를 증가시키는 경우, 가장 잘 된 것으로 인정된다.

[0155] **스캐폴딩 - 국소적 순서, 배향 및 갭 크기 개선**

[0156] 각 스캐폴드 내의 콘티그의 국소적 순서 및 배향을 모두 개선시키기 위해, 각 스캐폴드의 배열 및 배향된 콘티그에 대해 윈도우 크기 w를 변화시키는 동적 프로그래밍 알고리즘이 임의로 적용된다. 각 위치 i에서, 윈도우 내에서 콘티그를 순서배열하여 배향시키는 $w!\,2^w$의 모든 방법을 고려하였고, 상기 윈도우 내의 콘티그의 현재 구성으로 끝나는 현재 윈도우 위치의 말단까지 모든 콘티그에 대한 최적의 순서와 배향을 나타내는 스코어를 저장하였다. i-1, i-2, ... i-w 위치에서 윈도우 내의 모든 "조화된" 순서와 배향에 대한 스코어 및 현재의 구성으로 그들의 순서배열에 대한 확장성의 스코어를 사용하였다. $w!\,2^w$는 매우 가파른 함수이기 때문에, 본 방법은 일반적으로 작은 값의 w에 한정된다. 일부 실시양태에서, w는 2, 3, 4, 5, 6, 7, 8, 9, 10 또는 그 이상이다. 일부 데이터에 대한 시험에서, w=3은 일부 데이터세트에 대한 구성의 정확도를 크게 개선시킨다. 일부 실시양태에서, 콘티그를 연결하고 있는 리드 쌍의 분리에 대한 확률 모델을 고려하여 콘티그의 구성을 개선하기 위한 방법이 제공된다. 이러한 방법은 콘티그의 초기의 순서배열에 대해 윈도우 크기 w를 변화시키는 동적 프로그래밍 알고리즘이다.

[0157] **스캐폴딩 - 반복적인 결합**

[0158] 초기의 스캐폴드가 각 연결 요소 내에서 구축된 후, 생성된 스캐폴드를 단일 풀로 복귀시키고, 한쪽 말단과 다른 한쪽 말단의 결합 및 스캐폴드 삽입 결합을 여러 회수 수행한다. 매 회수에서, 스캐폴드의 모든 쌍들을 비교하고, 한쪽 말단과 다른 한쪽 말단의 결합 및 스캐폴드 삽입 결합에 대해 우도 스코어를 동시에 계산한다. 이후, 후보 결합을 분류하여 충돌하지 않는 결합을 내림차순의 우도 스코어 증가로 수용한다.

[0159] **장점의 구분**

[0160] 고처리량 시퀀싱 데이터로부터 *디 노보* 어셈블링된 콘티그의 스캐폴딩을 크게 개선시킬 수 있는 원거리 메이트 페어 데이터의 시험관내 생성을 위한 방법 및 컴퓨터 실행 시스템이 본원에 기술된다. 이러한 접근법들은 기존의 방법들에 비해 여러 가지 장점들을 가진다.

[0161] 첫째, 데이터 라이브러리 구축은 살아있는 생체 재료, 예컨대, 1차 또는 형질전환 조직 배양물 또는 살아있는 유기체를 필요로 하지 않는다. 본원에 기술된 라이브러리는 최소 5.0 μg의 입력 DNA 또는 그 미만, 예컨대 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 또는 1 μg 미만의 입력 DNA로부터 생성된다. 또한, *시험관내* 염색질 재구성이 인간 히스톤 및 염색질 어셈블리 요소를 기반으로 하는 것이지만, 광범위하게 다양한 식물, 동물 및 미생물의 DNA가 기술된 프로토콜을 사용하여 *시험관내* 염색질 어셈블리를 대체할 수 있다.

[0162] 둘째, 데이터가 *생체내* 수득된 염색질 공급이기 보다는 오히려 *시험관내* 어셈블링된 염색질의 근접 결찰에 의해 생성되기 때문에, 어셈블리를 잠재적으로 혼동하게 할 혼재된 생물학적 신호가 없다. Hi-C 및/또는 *생체내* 염색질로부터 생성된 기타 근접 결찰 데이터는 그 안에 생물학적 연관성에 대한 원거리 근접 정보를 수반하고 있지만, 끊임없이 지속되어 게놈 또는 스캐폴드 어셈블리에 대해 잠재적인 혼동을 줄 수 있다. 일부 실시양태에서, 본원에 제공된 방법은 낮은 백그라운드 노이즈 비율과 끊임없이 지속되는 의사 리드 쌍의 사실상의 부재를 유도한다.

[0163] 셋째, *생체내* Hi-C 방법과는 대조적으로, 생성된 리드 쌍의 최대 분리는 입력 DNA의 분자량에 의해서만 제한된다. 이는 단지 짧은 단편 일루미나 서열과 생성된 라이브러리를 이용하여 척추동물의 핵산 스캐폴드 또는 게놈의 고도로 인접한 스캐폴딩의 생성을 가능하게 하였다.

[0164] 넷째, 이들 라이브러리는 원거리 "메이트 페어" 및 포스미드 라이브러리의 조합을 생성하고 시퀀싱할 필요성을 없애주며, 이러한 라이브러리를 생성하는데 통상적으로 요구되는 고분자량 DNA의 전단 또는 크기 선별을 위한 고가의 특수 장비의 사용을 필요로 하지 않는다.

[0165] 따라서, 기존의 방법에 비해 현저하게 원거리인 콘티그 어셈블리 스캐폴드를 생성하기 위해 바이오인포매틱스 방법을 사용하는 DNA 라이브러리 구축 방법 및 컴퓨터 시스템이 제공된다. 일부 실시양태에서, 본원에 제공된 DNA 라이브러리 구축 방법은 또한 핵산 샘플 또는 게놈 변이를 확인하는 단계를 제공한다. 보통, 본원에 제공된 DNA 라이브러리 구축 방법 및 컴퓨터 시스템은 적은 노력과 비용으로 전장 하플로타입 분석된 염색체 서열의 정확한 재구축을 이루어낸다.

[0166] 순서 및 배향 정확도 모두를 개선하기 위해, 상기 순서 및 배향의 문제는 순서배열과 배향 단계를 병합함으로써 해결된다. 한 실례로, 그래프 내에서 노드가 콘티그 엔드이고, 각 콘티그의 두 말단 노드가 엣지에 의해 결합되도록 초기 그래프를 구축하였다. 짧은 특정 갭 크기의 전제 하에 콘티그들 간의 엣지의 로그 우도비 스코어를 계산한 후 분류하였다. 내림차순으로 된 엣지 스코어의 목록을 훑어 내려가면서, 새로운 엣지들은 이들이 어셈블리의 전체 스코어를 증가 또는 감소시킬 것인지의 여부에 따라 수용되거나 거부된다. 포지티브 스코어를 갖는 엣지라 하더라도 어셈블리 내의 콘티그의 스코어의 총합을 감소시킬 수 있는데, 이는 콘티그(들)의 기존의 스캐폴드의 갭 내로의 삽입을 포함하는 엣지의 수용이 상기 갭의 양측 상에 연결된 콘티그의 쌍 사이의 갭 크기를 증가시킬 것이고, 이것이 잠재적으로 이들에게 낮은 스코어를 제공할 것이기 때문이다.

[0167] 추가로, 최대 우도 갭 크기에 대한 계산의 효율성을 개선시킨다. 보고된 어셈블리의 전체적인 정확도는 연속된 콘티그들 사이의 미지의 서열의 길이를 추산함으로써 증가된다. 라이브러리 리드 쌍 사이의 분리 d에 대한 확률 밀도 함수 (PDF)의 모델을 포함하는 라이브러리 생성 방법의 모델을 고려하여, 최대 우도 갭 길이는 예를 들어 갭에 걸쳐있는 쌍의 분리 d_i 의 결합 우도를 최대화함으로써 구한다. 미분가능한 모델 PDF에 있어서는, 효율적인 반복적 최적화 방법 (예컨대 뉴턴-랩슨 (Newton-Raphson) 법)이 사용된다.

[0168] **성공적인 결과를 확인하기 위한 파라미터**

[0169] 본원의 다수의 방법 및 컴퓨터 시스템은 2개 이상의 순서, 배향, 스캐폴드 연결 상태, 콘티그 절단 평가, 또는 기타 가능한 서열 위치조정 (집합적으로, '스캐폴딩')에 대한 평가를 수반한다. 출발 또는 이전의 데이터에 비해 개선을 보이는 스캐폴딩을 숫자상으로 평가하기 위한 여러가지 접근법들을 이용할 수 있다. 예시적인 실시양태에서, (예를 들어, 모두 동일한 콘티그에 매핑되는 엔드를 갖는 리드 쌍의) 하나의 콘티그 또는 스캐폴드 구성이, 독립적으로 측정되거나 동시적으로 측정된 기대 곡선에 보다 가깝게 근접하는 리드 쌍 분리 거리 분포 곡선을 만들어 내는 경우 이는 바람직하다. 곡선이, 예를 들어 상기 식 1을 사용한 평가에서 좋은 스코어를 받는 경우, 이는 독립적으로 측정된 또는 동시적으로 측정된 기대 곡선에 보다 가깝게 근접한다. 변이에 대한 분석 (ANOVA) 시험, 공변이의 평가, 또는 기타 시험과 같은 대안적인 평가들도 당업계의 숙련자들이 이용할 수 있다.

[0170] 다르게는 또는 조합하여, 전체적인 스캐폴딩 유효성에 대한 별도의 스코어 또는 측량이 사용된다. 일부 척도들은 스캐폴딩이 공지된 핵산 어셈블리에 대한 백분율 정렬 (percent alignment)을 수반하여, 상기 백분율 정렬에서 개선을 유도하는 스캐폴딩이 선호된다. 일부의 경우, 상기 개선은 50%, 60%, 70%, 75%, 80%, 85%, 90%, 95%, 99%, 99.5%, 99.9% 이상, 또는 99.9% 초과의 백분율 정렬을 유도한다.

[0171] 두번째 척도는 N50과 같은 전체적인 스캐폴드 군집에 대한 스캐폴딩의 효과이다. 즉, 일부의 경우, 스캐폴딩은 서열 데이터세트에 대해 결과로서 생긴 N50이 증가되는 경우에 바람직하다. 일부의 경우, 상기 개선은 0.5%, 1%, 2%, 3%, 4%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, 150%, 200%, 250%, 300%, 400%, 500%, 600%, 700%, 800%, 900%, 1000%, 1500%, 2000%, 2500%, 3000%, 4000%, 5000%, 6000%, 7000%, 8000%, 9000%, 10,000%, 15,000%, 20,000%, 25,000%, 30,000%, 35,000%, 40,000%, 45,000%, 50,000%, 55,000%, 60,000%, 65,000%, 70,000%, 75,000%, 80,000%, 85,000%, 90,000%, 95,000%, 100,000% 이상, 또는 100,000% 초과의 백분율 증가를 유도한다.

[0172] 다르게는 또는 조합하여, 전체적인 스캐폴드 군집에 대한 추가의 측량법들이 사용된다. 이러한 측량 중의 하나는 RN50 측정이다. RN50은 하기와 같이 이해된다. 서열 세트 S와 참조 서열 R에 있어서, RN50, 즉 R에 대한 S의 "참조 N50"은 T 중에서 가장 짧은 서열의 길이이다 (여기서, T는, T 중의 서열이 R 길이의 50% 이상을 커버하는

방식으로 R에 정렬될 수 있도록 하는 S의 가장 작은 (최소 기수) 하위세트임).

[0173] T 중의 모든 서열들이 최소한 RN50만큼 크기 때문에, 이는 무작위로 선택된 R의 염기가 최소 길이 RN50인 T 중의 (이에 따라 S 중의) 서열에 대한 정렬로 인해 신장될 확률이 50% 이상이라는 것을 의미한다.

[0174] 일부의 경우, 초기 RN50은 0.5%, 1%, 2%, 3%, 4%, 5%, 10%, 20%, 30%, 40%, 또는 50%인 값, 또는 이 범위 내의 또다른 값의 샘플 서열, 예컨대 게놈 서열을 가진다. 일부의 경우, 최종 RN50은 초기의 RN50 보다 더 큰 값을 가진다. 일부의 경우, 개선은 0.5%, 1%, 2%, 3%, 4%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% 이상, 또는 90% 초과, 100%, 2x, 5x, 10x, 20x, 50x, 100x, 200x, 500x, 1000x, 2000x, 5000x, 10000x 또는 10000x 초과의 RN50에서의 백분율 증가를 유도한다.

[0175] 본원의 개시내용과 일관된 대안적인 스캐폴딩에 대한 측량법들도 고려된다.

[0176] 도면을 참조하면, 도 1A-F는 DNA 라이브러리 생성 프로토콜에 대한 예시적인 계략도를 도시하고 있다. 도 1A에서, 염색질 (원으로 묘사된 뉴클레오솜)은 네이키드 DNA (검정색 가닥) 상에서 시험관내 재구성된다. 도 1B에서, 염색질은 포름알데히드로 고정된다 (가는 선이 가교결합임). 고정된 염색질은 스트렙타비딘 비드에 대하여 무너져 내리고, 제한 효소로 절단되어 유리 점착성 말단을 생성하게 된다 (도 1C). 도 1D에서, 점착성 말단은 비오틴화된 (더 작은 원) 뉴클레오타이드 및 티올화된 (작은 사각형) 뉴클레오타이드로 충전된다. 도 1E에서, 유리 평활 말단이 결찰된다 (별표로 나타낸 결찰). 도 1F에서, 가교결합이 뒤집어지고 단백질이 제거되어 라이브러리 단편들을 생성한다. 라이브러리 단편들은 분해되어 비결찰된 말단으로부터 비오틴화된 뉴클레오타이드를 제거한다. 라이브러리 단편들을 스트렙타비딘이 코팅된 비드를 사용하여 선택한 후, 시퀀싱에 대비하여 결찰된 어댑터를 보유시킨다.

[0177] 일부 실시양태에서, 참조 어셈블리에 매핑된 수개의 생성된 라이브러리에 대한 리드 쌍 분리를 비교하는 단계를 사용한다. 예를 들어, 도 2는 참조 인간 어셈블리, 예컨대 hg19에 매핑된 수개의 생성된 라이브러리에 대한 리드 쌍 분리에 대한 실례를 제공한다. 도면에서, 플롯의 우측에 최하위 그룹으로 표시된 데이터 추이는 50 Kbp의 입력 인간 시퀀싱 라이브러리에 해당된다. 플롯의 우측에 하부 화살표로 표시된 데이터 추이는 150 Kbp의 입력 인간 시퀀싱 라이브러리에 해당된다. 플롯의 우측에 중앙의 화살표로 표시된 데이터 추이는 150 Kbp의 입력 인간 시퀀싱 라이브러리에 해당된다. 플롯의 우측에 최상위 화살표로 표시된 데이터 추이는 인간의 Hi-C 라이브러리에 해당된다 (Kalhor et al., 2012). 진한 수직선들은 최대로 발표되거나 입증된 대안적인 메이트 페어 기법에 대한 역량을 나타낸다.

[0178] 인접한 핵산 또는 게놈 크기 추정치로 나눈 리드 쌍 분리의 총합은 대안적인 또는 추가의 측량법으로서 분리 범위에 의해 그룹화된 다양한 리드 쌍에 대해 계산된다. 이러한 비교의 예시적인 개요가 도 3에 나타나 있다.

[0179] 도 4A 및 4B를 참조하면, 참조 서열 상에 매핑된 위치에 대한 실례가 제공되는데, 예컨대, 재어셈블링된 염색질의 DNA의 근접 결찰로 생성된 리드 쌍의 GRCh38이 GM12878과 참조 서열 간의 구조적 차이 부근에 플롯팅된다. 생성된 각 리드 쌍은 대각선의 위와 아래에 모두 나타난다. 대각선 위에서, 음영은 나타낸 척도 상의 맵 품질 스코어를 나타내며; 대각선 아래에서, 음영은 페이징된 SNP와의 중첩을 기반으로 하여 생성된 리드 쌍의 추정된 하플로타입 상을 나타낸다. 일부 실시양태에서, 생성된 플롯은 도 4B에 도시된 것과 같이 측부 반복 영역을 갖는 전위를 묘사하고 있다. 일부 실시양태에서, 생성된 플롯은 도 4B에 도시된 것과 같이 단계적 이형접합 결실을 묘사하고 있다.

[0180] 도 5를 참조하면, 염색질 리어셈블링 및 처리에 대한 실례가 제공된다. 일부 실시양태에서, 정제된 고분자량 DNA는 (히스톤 및 염색질 어셈블리 요소를 이용하는) *시험관내* 염색질 어셈블리를 거친 후 비오틴화된다. 일부 실시양태에서, 생성된 DNA-염색질 복합체는 이후 고정화제를 사용하여 고정된다. 도 5에서, 상기 고정화제는 포름알데히드이다. 추가의 실시양태에서, 상기 DNA-염색질 복합체는 스트렙타비딘 비드를 사용하여 붕괴되다. 추가의 실시양태에서, 상기 DNA-염색질 복합체는 제한 효소 분해를 사용하여 처리된다. 도 5에 도시된 것과 같이, 일부의 경우, 상기 제한 효소는 MboI이다. 일부의 경우, 상기 제한 효소는 이중 가닥 DNA 내에 점착성 말단, 예컨대, 오버행을 남기는 것들이다. 일부 실시양태에서, 상기 점착성 말단은 표지된 뉴클레오타이드로 충전된다. 일부의 경우, 표지된 뉴클레오타이드는 비오틴화되거나 또는 황산화된다. 도 5에 도시된 것과 같이, 일부의 경우, 내부 충전은 황산화된 dGTP를 사용하여 수행되고, 외부 충전은 비오틴화된 dCTP를 사용하여 수행된다. 일부 실시양태에서, 평활 말단 결찰 단계가 수행되어, 충전 말단이 결합된다. 일부 실시양태에서, 상기 DNA-염색질 복합체는 효소적 분해를 거쳐 상기 복합체로부터 DNA를 방출한다. 도 5에 도시된 것과 같이, 일부 실시양태에서 상기 효소는 프로테이나아제 K이다. 일부 실시양태에서, 상기 DNA를 제한 효소로 처리하여 표지된 뉴클레오타이

드를 제거한다. 예를 들어, 도 5에 도시된 것과 같이, ExoIII 분해를 이용한 처리는 말단 상의 비오틴화된 시토신을 제거한다. 일부 실시양태에서, DNA 단편들은, 전단, 분해 및 임의의 일루미나-호환가능한 라이브러리 생성 프로토콜의 사용에 의해 분석을 위해 준비하지만, 비드 상에서의 세척 및 재현탁 단계를 대체하는 대신에 반응 사이의 통상적인 클린업 단계를 제거한다.

[0181] 다양한 실시양태에서, 본 발명의 방법 및 시스템은 컴퓨터 시스템 상의 소프트웨어 프로그램 및 이의 용도를 더 포함한다. 상기 컴퓨터 시스템은 사용자간 인터페이스로 프로그래밍될 수 있다.

[0182] 도 6에 도시된 컴퓨터 시스템(600)은 매체(611) 및/또는 임의로 고정된 매체(612)를 갖는 서버(609)에 연결될 수 있는 네트워크 포트(605)의 명령어를 읽을 수 있는 논리적 장치로 이해될 수 있다. 도 6에 나타낸 것과 같은 시스템은 CPU(601), 디스크 드라이브(603), 임의의 입력 장치, 예컨대 키보드(615) 및/또는 마우스(616) 및 임의의 모니터(607)를 포함한다. 데이터 통신은 명시된 통신 매체를 통하여 근거리 또는 원거리 위치에 있는 서버까지 이루어진다. 상기 통신 매체는 데이터를 전달 및/또는 수신하는 임의의 장치, 기구 또는 접근 수단을 포함할 수 있다. 예를 들어, 상기 통신 매체는 네트워크 연결, 무선 연결 또는 인터넷 연결일 수 있다. 이러한 연결은 월드 와이드 웹 (World Wide Web)을 통해 통신을 제공할 수 있다. 본 개시내용과 관련된 데이터는, 도 6에 도시된 것과 같이, 당사자(622)에 의한 수신 및/또는 열람을 위해 이러한 네트워크 또는 연결을 통해 전달될 수 있다.

[0183] 도 7은 본원에 기술된 예시적 실시양태에서 사용되는 컴퓨터 시스템(700)의 제1 예시 구조를 나타내는 블록 다이아그램이다. 도 7에 묘사된 것과 같이, 예시의 컴퓨터 시스템은 명령어를 처리하기 위한 프로세서(702)를 포함한다. 프로세서의 비제한적인 예로서는 하기를 포함한다: 인텔 XeonTM 프로세서, AMD OpteronTM 프로세서, 삼성 32비트 RISC ARM 1176JZ(F)-S v1.0TM 프로세서, ARM Cortex-A8 삼성 S5PC100TM 프로세서, ARM Cortex-A8 애플 A4TM 프로세서, 마벨 (Marvell) PXA 930TM 프로세서, 또는 이와 기능적으로 동등한 프로세서. 병행 처리를 위해 복수의 쓰레드 (thread)의 실행이 사용될 수 있다. 일부 실시양태에서, 복수의 프로세서 또는 복수의 코어를 갖는 프로세서가 단일 컴퓨터 시스템 중에, 클러스터로 사용되거나, 또는 복수의 컴퓨터, 휴대폰, 및/또는 개인정보단말장치를 포함하는 네트워크를 통해 시스템에 분산되어 사용된다.

[0184] 도 7에 도시된 것과 같이, 고속 캐쉬(704)가 프로세서(702)에 연결되거나 그 안에 포함되어, 상기 프로세서(702)에 의해 최근에 사용되었거나 자주 사용되었던 명령어나 데이터에 대한 고속 메모리를 제공한다. 상기 프로세서(702)는 프로세서 버스(708)에 의해 노스 브리지(706)로 연결된다. 상기 노스 브리지(706)는 메모리 버스(712)에 의해 랜덤 엑세스 메모리 (RAM) (710)로 연결되어 프로세서(702)에 의해 RAM(710)에 접근하게 된다. 상기 노스 브리지(706)는 칩셋 버스(716)에 의해 사우스 브리지(714)로도 연결된다. 사우스 브리지(714)는, 결과적으로 주변장치 버스(718)로 연결된다. 상기 주변장치 버스는, 예를 들어, PCI, PCI-X, PCI Express, 또는 기타 주변장치 버스이다. 상기 노스 브리지 및 사우스 브리지는 보통 프로세서 칩셋으로 지칭되고, 프로세서, RAM 및 주변장치 버스(718) 상의 주변장치 요소들 간에 데이터 전송을 관리한다. 일부 대안적인 구조에서, 별도의 노스 브리지 칩을 사용하는 대신에, 노스 브리지의 기능을 프로세서에 포함시킨다.

[0185] 일부 실시양태에서, 시스템(700)은 주변장치 버스(718)에 부착된 가속기 카드(722)를 포함한다. 상기 가속기는 필드 프로그래머블 게이트 어레이 (field programmable gate array, FPGA) 또는 특정 처리를 가속화하기 위한 기타 하드웨어를 포함할 수 있다. 예를 들어, 가속기는, 일부의 경우, 자체적용 데이터 재구축 (adaptive data restructuring)을 위해 사용되거나, 또는 확장된 세트 처리에 사용되는 대수식을 평가하기 위해 사용된다.

[0186] 소프트웨어 및 데이터는 외부 기억 장치(724)에 저장되어 프로세서에 의한 사용을 위해 RAM(710) 및/또는 캐쉬(704) 내에 로딩될 수 있다. 시스템(2000)은 시스템 자원을 관리하기 위한 운영 체제를 포함하며; 운영 체제의 비제한적인 예로는 하기를 포함한다: 리눅스, 윈도우즈TM, MACOSTM, 블랙베리 OSTM, iOSTM 및 기타 이와 기능적으로 동등한 운영 체제, 그리고 본 발명의 예시적 실시양태에 따른 데이터 저장과 최적화를 관리하기 위한 운영 체제 상에서 구동되는 응용 소프트웨어.

[0187] 이러한 예에서, 시스템(700)은 외부 기억 장치, 예컨대 네트워크 접속기반 파일서버 (NAS) 및 분산 병행 처리를 위해 사용될 수 있는 다른 컴퓨터 시스템에 네트워크 인터페이스를 제공하기 위해 주변장치 버스에 연결된 네트워크 인터페이스 카드(NIC) (720 및 721)도 포함한다.

[0188] 도 8은 복수의 컴퓨터 시스템 (802a 및 802b), 복수의 휴대폰과 개인 정보 단말기(802c), 및 네트워크 접속기반 파일서버 (NAS) (804a 및 804b)를 보유하는 네트워크(800)를 나타내는 다이아그램이다. 예시적 실시양태에서, 시스템(802a, 802b 및 802c)은 데이터 저장을 관리하여 네트워크 접속기반 파일서버 (NAS)(804a 및 804b) 내에

저장된 데이터에 대한 데이터 접근을 최적화할 수 있다. 수학적 모델이 일부의 경우, 데이터에 대해 사용되어, 컴퓨터 시스템(802a 및 802b) 및 휴대폰과 개인정보단말시스템(802c)에 걸친 분산 병행 처리를 이용하여 평가된다. 컴퓨터 시스템(802a 및 802b) 및 휴대폰과 개인정보단말시스템(802c)은 네트워크 접속기반 파일서버(NAS)(804a 및 804b) 내에 저장된 데이터의 자체적용 데이터 재구축을 위한 병행 처리를 제공할 수도 있다. 도 8은 한 실례만을 도시한 것이고, 광범위하게 다양한 다른 컴퓨터 구조와 시스템들을 본 발명의 다양한 실시양태와 함께 사용할 수 있다. 일부 예에서, 블레이드 서버가 병행 처리를 제공하기 위해 사용된다. 프로세서 블레이드는 병행 처리를 제공하기 위해 회로 뒤판을 통해 연결될 수 있다. 저장 장치도 회로 뒤판에 연결되거나, 또는 별도의 네트워크 인터페이스를 통한 네트워크 접속기반 파일서버 (NAS)로서 연결될 수도 있다.

[0189] 일부 예시적 실시양태에서, 프로세서는 별도의 메모리 공간을 유지하여 네트워크 인터페이스, 회로 뒤판 또는 다른 프로세서에 의한 병행 처리를 위한 기타 연결기를 통해 데이터를 전송할 수 있다. 다른 실시양태에서, 프로세서의 일부 또는 전체가 공유 가상 주소 메모리 공간을 사용할 수 있다.

[0190] 도 9는 본 발명의 예시적 실시양태에 따른 공유 가상 주소 메모리 공간을 사용하는 멀티프로세서 컴퓨터 시스템(900)의 블록 다이아그램이다. 상기 시스템은 공유된 메모리 서브시스템(904)에 접근할 수 있는 복수의 프로세서(902a-f)를 포함한다. 상기 시스템은 메모리 서브시스템(904) 내에 복수의 프로그래밍가능한 하드웨어 메모리 알고리즘 프로세서 (MAP) (906a-f)를 도입한다. 각각의 MAP(906a-f)은 메모리(908a-f) 및 하나 이상의 필드 프로그래머블 게이트 어레이 (FPGA) (910a-f)를 포함할 수 있다. 상기 MAP은 설정가능한 기능단위를 제공하여 특정 알고리즘 또는 알고리즘의 일부가 각 프로세서와의 긴밀히 협조되는 처리를 위한 (FPGA) (910a-f)에 제공된다. 예를 들어, MAP은 일부의 경우, 데이터 모델과 관련한 대수식을 평가하는데 사용되고, 예시적 실시양태에서 자체적용 데이터 재구축을 수행하기 위해 사용된다. 이러한 실례에서, 각각의 MAP은 이러한 목적을 위한 모든 프로세서에 의해 전반적으로 접근이 가능하다. 한 구성에 있어서, 각각의 MAP은 직접 메모리 접근 (DMA)을 사용하여 연관 메모리(908a-f)에 접근할 수 있어서, 각각의 마이크로프로세서(902a-f)와는 독립적으로, 그리고 이와는 비동기적으로 작업을 수행하도록 한다. 이 구성에 있어서, MAP는 알고리즘의 파이프라이닝 및 병행 실행을 위해 결과를 또 다른 MAP로 직접 제공할 수 있다.

[0191] 상기 컴퓨터 구조 및 시스템은 단지 실례로 든 것이고, 광범위하게 다양한 다른 컴퓨터, 휴대폰, 및 개인정보단말 구조 및 시스템들, 예컨대 일반 프로세서, 보조 프로세서, FPGA 및 기타 프로그래밍가능한 논리 장치, 시스템 온 칩 (system on chip, SOC), 주문형 집적회로 (application specific integrated circuit, ASIC), 및 기타 프로세싱 및 논리 소자의 임의의 조합을 이용한 시스템들도 예시적 실시양태에서 사용된다. 일부 실시양태에서, 컴퓨터 시스템의 전부 또는 일부가 소프트웨어 또는 하드웨어에서 실행된다. 다양한 임의의 데이터 저장 매체들, 예컨대 랜덤 엑세스 메모리, 하드 드라이브, 플래시 메모리, 테이프 드라이브, 디스크 어레이, 네트워크 접속기반 파일서버 (NAS) 및 기타 로컬 또는 분산 데이터 저장 장치 및 시스템이 예시적 실시양태와 관련하여 사용될 수 있다.

[0192] 예시적 실시양태에서, 상기 컴퓨터 시스템은 상기 또는 다른 컴퓨터 구조 및 시스템 상에서 실행되는 소프트웨어 모듈을 사용하여 실행된다. 다른 실시양태에서, 시스템의 기능은 펌웨어, 프로그래밍가능한 논리 장치, 예컨대 도 9에서 언급된 바와 같은 필드 프로그래머블 게이트 어레이 (FPGA), 시스템 온 칩 (SOC), 주문형 집적회로 (ASIC), 또는 기타 프로세싱 및 논리 소자에서 부분적으로 또는 완전히 실행된다.

[0193] 본 출원의 출원 당시에 사용되던 방법들에 비해, 본원에 개시된 방법 및 시스템은 여러가지 장점들을 제공한다.

[0194] 본원에 개시된 일부 방법 및 컴퓨터 시스템은 유기체에 대한 염색체의 수와는 독립적인 방식으로 콘티그들을 클러스터링한다. 본원에 개시된 다양한 방법들로 가능한 후속 스캐폴딩 결합을 이용하여, 생성된 더 작은 콘티그 클러스터를 스캐폴드 내로 어셈블링하기 위해 단일 링크 클러스터링을 위한 콘티그-콘티그 링크에 대한 보다 보존적인 역치가 적용된다.

[0195] 일부 실시양태에서, 본원에 개시된 방법들은 본질적으로 클러스터링을 수반하지 않고 신장 트리 단계로 바로 들어간 후, 위상학적 트리 가지치기 (topological tree pruning)를 수행한다. 일부 실시양태에서, 하나 이상의 클러스터링 방법, 예컨대 마르코브 (Markov) 클러스터 알고리즘 (MCL algorithm)이 사용된다. 이론에 얽매이지 않으면서, 미스어셈블리는 이들 엣지들을 각별히 신경써서 처리하여 위상학적 가지치기를 하여 어셈블리 부정합을 피함으로써 방지될 수 있다.

[0196] 스캐폴드 내의 콘티그의 순서를 고정시킨 후, 일부의 경우, 동적 프로그래밍 알고리즘을 사용하여 배향을 최적화시킨다. 이러한 접근법은 순서배열에서 인접한 콘티그의 쌍에 매핑되는 리드 쌍만이 최적화되는 스코어에 기

여하여, 좋은 단편 상의 최대 분리보다 짧은 임의의 콘티그를 제외시켜 버려 어셈블링되지 않게 한다. 배향 단계를 개선하기 위해, 최대 $w-2$ 개재 콘티그들 내에서 콘티그의 쌍에 매핑되는 모든 쌍의 데이터를 도입하는 알고리즘을 이용하여, 예컨대 순서배열에서 2 또는 그 이상의 콘티그의 값, 예컨대 2, 3, 4, 5, 6, 7, 8, 9, 10 또는 10 이상을 이용함으로써, 최근방 콘티그 스코어 상호작용 이외에 최근방 콘티그 스코어 상호작용이 아닌 콘티그들도 고려될 수 있다.

[0197] 일부 실시양태에서, 삽입 단계의 정확도가 개선된다. 임의의 이론에 얽매이지 않으면서, 트렁크의 생성 후 좋은 리드 쌍 사이의 최대 분리보다 짧은 콘티그를 갖는 어셈블리에서, 순서배열을 따라 w 콘티그의 근처 내에 있는 콘티그의 데이터는, 트렁크로부터 콘티그를 제외시키고 이를 인전합 콘티그들 간의 연관의 양을 최대화하는 부위에 재삽입하는 경우에 포함된다.

[0198] 배향 단계는 일부의 경우, 최근방 이상의 콘티그 스코어 상호작용을 고려함으로써 개선된다. 스캐폴드 내의 콘티그의 순서를 고정시킨 후, 동적 프로그래밍 알고리즘을 사용하여 배향을 최적화시킨다. 순서배열에서 인접한 콘티그의 쌍에 매핑되는 리드 쌍만이 최적화되는 스코어에 기여한다. 순서배열에서 최대 $w-2$ 개재 콘티그들 내에서 콘티그의 쌍에 매핑되는 모든 쌍의 데이터를 도입하는 알고리즘을, 좋은 단편 상의 최대 분리보다 짧은 임의의 콘티그를 갖는 어셈블리에 대하여 사용한다. 예를 들어, 순서배열에서 2 또는 그 이상의 콘티그의 값, 예컨대 2, 3, 4, 5, 6, 7, 8, 9, 10 또는 10 이상을 이용한다.

[0199] 일부 실시양태에서, 순서 및 배향 단계를 보다 기밀하게 병합함으로써 순서 및 배향 정확도 모두를 개선한다. 그래프 내에서 노드가 콘티그 엔드이고, 각 콘티그의 두 말단 노드가 엣지에 의해 결합되도록 초기 그래프를 구축한다. 짧은 특정 갭 크기의 전제 하에 콘티그들 간의 엣지의 로그 우도비 스코어를 계산한 후 분류하였다. 내림차순으로 된 엣지 스코어의 목록을 훑어 내려가면서, 새로운 엣지들은 이들이 어셈블리의 전체 스코어를 증가 또는 감소시킬 것인지의 여부에 따라 수용되거나 거부된다. 포지티브 스코어를 갖는 엣지라 하더라도 어셈블리 내의 콘티그의 스코어의 총합을 감소시킬 수 있는데, 이는 콘티그(들)의 기존의 스캐폴드의 갭 내로의 삽입을 포함하는 엣지의 수용이 상기 갭의 양측 상에 연결된 콘티그의 쌍 사이의 갭 크기를 증가시킬 것이고, 이것이 잠재적으로 이들에게 낮은 스코어를 제공할 것이기 때문이다.

[0200] 추가로, 최대 우도 갭 크기를 효율적으로 계산한다. 보고된 어셈블리의 전체적인 정확도는 연속된 콘티그들 사이의 미지의 서열의 길이를 추산함으로써 증가된다. 라이브러리 리드 쌍 사이의 분리 d에 대한 확률 밀도 함수(PDF)의 모델을 포함하는 라이브러리 생성 방법의 모델을 고려하여, 최대 우도 갭 길이는 예를 들어 갭에 걸쳐 있는 쌍의 분리 d_i의 결합 우도를 최대화함으로써 구한다. 미분가능한 모델 PDF에 있어서는, 효율적인 반복적 최적화 방법 (예컨대 뉴턴-랩슨 법)이 사용된다.

[0201] 본원에 개시된 방법 및 조성물의 원리는, 콘티그가 예컨대 8, 7, 6, 5, 4, 3, 2시간, 또는 2시간 미만의 비교적 짧은 시간 내에 실행가능하거나 수득가능하면서, 예를 들어, 2, 3, 4, 5, 6, 또는 6 이상의 콘티그의 콘티그 윈도우 중에서 콘티그 순서, 배향 또는 순서 및 배향에 대해 국소적으로 최적인 구성으로 어셈블링되는 것이다. 따라서, 일부의 경우, 본원의 방법은, 엄청난 계산 시간을 쓰지 않고 전반적으로 매우 큰 계산 공간을 살펴볼 필요없이, 계산 집약적인 문제에 고도의 계산력을 투여할 수 있게 해준다. 오히려, 국소적인 순서는 콘티그의 보통 정도의 정확한 순서를 달성하며, 이후 계산적 강도는 대부분의 경우 한번에 모든 콘티그들 전반적으로 최적화시키기 보다는 콘티그의 국소적 윈도우를 최적화하는데 소비된다. 일부의 경우, 3, 4, 5, 또는 6 범위의 윈도우 크기를 사용하여, 구성 최적화는 8, 7, 6, 5, 4, 3, 2시간, 또는 2시간 이내에 수행된다. 더 큰 윈도우 크기에 대해서, 구성 최적화는 수일 내지 일주일 내에 달성된다.

[0202] 디지털 처리 장치

[0203] 일부 실시양태에서, 본원에 기술된 콘티그 어셈블리 방법은 디지털 처리 장치, 또는 이의 사용을 포함한다. 추가의 실시양태에서, 상기 디지털 처리 장치는 장치의 기능을 수행하는 하나 이상의 하드웨어 중앙 처리 장치(CPU)를 포함한다. 추가의 실시양태에서, 상기 디지털 처리 장치는 실행가능한 명령어를 수행하도록 설정된 운영 체제를 더 포함한다. 일부 실시양태에서, 상기 디지털 처리 장치는 임의로 컴퓨터 네트워크에 연결된다. 추가의 실시양태에서, 상기 디지털 처리 장치는 월드 와이드 웹에 접근할 수 있도록 임의로 인터넷에 연결된다. 추가의 실시양태에서, 상기 디지털 처리 장치는 임의로 클라우드 컴퓨팅 인프라에 연결된다. 다른 실시양태에서, 상기 디지털 처리 장치는 임의로 인트라넷에 연결된다. 다른 실시양태에서, 상기 디지털 처리 장치는 임의로 데이터 저장 장치에 연결된다.

[0204] 본원의 설명에 따르면, 적절한 디지털 처리 장치로는 비제한적인 예로서, 서버 컴퓨터, 데스크탑 컴퓨터, 랩탑

컴퓨터, 노트북 컴퓨터, 서브노트북 컴퓨터, 넷북 컴퓨터, 넷패드 컴퓨터, 셋톱 컴퓨터, 미디어 스트리밍 장치, 포켓용 컴퓨터, 인터넷 기기, 모바일 스마트폰, 태블릿 컴퓨터, 개인 정보 단말기, 비디오 게임기 및 차량을 포함한다. 당업계의 숙련자라면 많은 스마트폰들이 본원에 기술된 시스템에 사용하기에 적절하다는 것을 알 수 있을 것이다. 또한, 당업계의 숙련자라면 임의의 컴퓨터 네트워크 연계성을 갖는 선택된 텔레비젼, 비디오 플레이어, 및 디지털 음악 플레이어가 본원에 기술된 시스템에 사용하기에 적절하다는 것을 알 수 있을 것이다. 적절한 태블릿 컴퓨터로는 당업계의 숙련자에게 공지된 북클릿 (booklet), 슬레이트 및 컨버터블 구성을 갖는 것들을 포함한다.

[0205] 일부 실시양태에서, 상기 디지털 처리 장치는 실행가능한 명령어를 수행하도록 설정된 운영 체제를 포함한다. 상기 운영 체제는, 예를 들어, 장치의 하드웨어를 관리하여 응용 프로그램의 실행에 대한 서비스를 제공하는 프로그램과 데이터를 포함하는 소프트웨어이다. 당업계의 숙련자라면, 적절한 서버 운영 체제로 비제한적인 예로서, FreeBSD, OpenBSD, NetBSD®, Linux, 애플® Mac OS X Server®, Oracle® Solaris®, 윈도우즈 Server®, 및 Novell® NetWare®를 포함한다는 것을 알 수 있을 것이다. 당업계의 숙련자라면, 적절한 개인용 컴퓨터 운영 체제로 비제한적인 예로서, Microsoft® 윈도우즈®, 애플® Mac OS X®, UNIX® 및 UNIX 유사 운영 체제, 예컨대 GNU/Linux®를 포함한다는 것을 알 수 있을 것이다. 일부 실시양태에서, 상기 운영 체제는 클라우드 컴퓨팅에 의해 제공된다. 또한, 당업계의 숙련자라면 적절한 모바일 스마트폰 운영 체제로 비제한적인 예로서, 노키아®, Symbian® OS, 애플® iOS®, Research In Motion®, BlackBerry OS®, 구글® 안드로이드®, Microsoft® 윈도우즈 폰® OS, Microsoft® Wondows 모바일® OS, Linux® 및 Palm® WebOS®를 포함한다는 것을 알 수 있을 것이다.

[0206] 일부 실시양태에서, 상기 장치는 저장 및/또는 메모리 장치를 포함한다. 상기 저장 및/또는 메모리 장치는 일시적으로 또는 영구적으로 데이터 또는 프로그램을 저장하는데 사용되는 하나 이상의 물리적 장치이다. 일부 실시양태에서, 상기 장치는 휘발성 메모리이어서 저장된 정보를 유지할 전원을 필요로 한다. 일부 실시양태에서, 상기 장치는 비휘발성 메모리이어서 상기 디지털 처리 장치가 전원이 켜있지 않은 경우에도 저장된 정보를 보유한다. 추가의 실시양태에서, 상기 비휘발성 메모리는 플래시 메모리를 포함한다. 일부 실시양태에서, 상기 비휘발성 메모리는 동적 랜덤 엑세스 메모리 (DRAM)를 포함한다. 일부 실시양태에서, 상기 비휘발성 메모리는 강유전성 랜덤 엑세스 메모리 (FRAM)를 포함한다. 일부 실시양태에서, 상기 비휘발성 메모리는 상변화 랜덤 엑세스 메모리 (PRAM)를 포함한다. 임의로, 상기 장치는 비제한적인 예로서, CD-ROM, DVD, 플래시 메모리 장치, 자기 디스크 드라이브, 자기 테이프 드라이브, 광학 디스크 드라이브 및 클라우드 컴퓨팅 기반의 저장 장치를 비롯한 저장 장치이다. 추가의 실시양태에서, 상기 저장 및/또는 메모리 장치는 본원에 개시된 것들과 같은 장치들의 조합이다.

[0207] 일부 디지털 처리 장치로는 시각 정보를 사용자에게 보내기 위한 디스플레이, 예컨대 음극선관 (CRT), 액정 디스플레이 (LCD), 박막 트랜지스터 액정 디스플레이 (TFT-LCD), 유기 발광 다이오드 (OLED) 디스플레이, 예컨대 패시브 매트릭스 OLED (PMOLED) 또는 액티브 매트릭스 OLED (AMOLED) 디스플레이, 플라즈마 디스플레이, 비디오 프로젝터, 또는 본원에 개시된 것들과 같은 장치들의 조합을 포함한다.

[0208] 보통, 상기 디지털 처리 장치는 사용자로부터 정보를 수신하기 위한 입력 장치, 예컨대 키보드, 포인팅 장치, 예컨대, 비제한적인 예로서, 마우스, 트랙볼, 트랙 패드, 조이스틱, 게임 콘트롤러, 또는 스타일러스를 포함한다. 일부 실시양태에서, 입력 장치는 터치 스크린 또는 멀티-터치 스크린, 음성을 녹음하거나 기타 소리 입력을 위한 마이크로폰, 또는 움직임을 포착하거나 시각 입력을 위한 비디오 카메라 또는 기타 센서이다. 추가의 실시양태에서, 입력 장치는 키넥트 (Kinect), 립 모션 (Leap Motion) 등이다. 흔히, 입력 장치는 본원에 개시된 것들과 같은 장치들의 조합이다.

[0209] 비일시적 컴퓨터 판독가능 저장 매체

[0210] 일부 실시양태에서, 본원에 개시된 콘티그 어셈블리 방법은 임의로 네트워킹된 디지털 처리 장치의 운영 체제에 의해 실행가능한 명령어를 포함하는 프로그램으로 암호화된, 하나 이상의 비일시적 컴퓨터 판독가능 저장 매체를 수반한다. 추가의 실시양태에서, 컴퓨터 판독가능 저장 매체는 디지털 처리 장치의 실체적 요소이다. 추가의 실시양태에서, 컴퓨터 판독가능 저장 매체는 디지털 처리 장치로부터 임의로 제거가능하다. 일부 실시양태에서, 컴퓨터 판독가능 저장 매체로는 비제한적인 예로서, CD-ROM, DVD, 플래시 메모리 장치, 반도체 메모리, 자기 디스크 드라이브, 자기 테이프 드라이브, 광학 디스크 드라이브, 클라우드 컴퓨팅 시스템 및 서비스 등을 포함한다. 일부의 경우, 프로그램과 명령어는 영구적으로, 실질적으로 영구적으로, 반영구적으로, 또는 비일시적으로 매체 상에 암호화된다.

[0211] 컴퓨터 프로그램

[0212] 일부 실시양태에서, 본원에 개시된 콘티그 어셈블리 방법은 하나 이상의 컴퓨터 프로그램, 또는 이의 사용을 포함한다. 컴퓨터 프로그램은 특정 임무를 수행하도록 기록된, 디지털 처리 장치의 CPU에서 실행가능한 연속 명령이다. 컴퓨터 판독가능 명령은 특정 임무를 수행하거나 특정 추상 데이터형을 이행하는 프로그램 모듈, 예컨대 기능, 객체, 응용 프로그래밍 인터페이스 (API), 데이터 구조 등으로 이행될 수 있다. 본원에 제공된 개시내용에 고려하면, 당업계의 숙련자라면 컴퓨터 프로그램은 다양한 언어의 다양한 버전으로 쓰여질 수 있다는 것을 알 수 있을 것이다.

[0213] 컴퓨터 판독가능 명령의 기능은 다양한 환경에서 필요에 따라 조합되거나 분산될 수 있다. 일부 실시양태에서, 컴퓨터 프로그램은 하나의 연속 명령을 포함한다. 일부 실시양태에서, 컴퓨터 프로그램은 복수의 연속 명령을 포함한다. 일부 실시양태에서, 컴퓨터 프로그램은 한 위치로부터 제공된다. 다른 실시양태에서, 컴퓨터 프로그램은 복수의 위치로부터 제공된다. 다양한 실시양태에서, 컴퓨터 프로그램은 하나 이상의 소프트웨어 모듈을 포함한다. 다양한 실시양태에서, 컴퓨터 프로그램은 부분적으로 또는 전체적으로, 하나 이상의 웹 응용프로그램, 하나 이상의 모바일 응용프로그램, 하나 이상의 독립형 응용프로그램, 하나 이상의 웹 브라우저 플러그인, 익스텐션, 애드-인 (add-in), 또는 애드-온 (add-on), 또는 이들의 조합을 포함한다.

[0214] 웹 응용프로그램

[0215] 일부 실시양태에서, 콘티그 어셈블리 방법을 실행하는 컴퓨터 프로그램 또는 컴퓨터 실행 시스템은 웹 응용프로그램을 포함한다. 본원에 제공된 개시내용에 비추어 보면, 당업계의 숙련자라면 웹 응용프로그램이, 다양한 실시양태에서, 하나 이상의 소프트웨어 프레임워크 및 하나 이상의 데이터베이스 시스템을 이용한다는 점을 알 수 있을 것이다. 일부 실시양태에서, 웹 응용프로그램은 소프트웨어 프레임워크, 예컨대 Microsoft® .NET 또는 루비 온 레일즈 (Ruby on Rails, RoR)를 바탕으로 생성된다. 일부 실시양태에서, 웹 응용프로그램은 비제한적인 예로서, 관계형, 비관계형, 객체 지향형, 결합형 및 XML 데이터베이스 시스템을 비롯한 하나 이상의 데이터베이스 시스템을 이용한다. 추가의 실시양태에서, 적절한 관계형 데이터베이스 시스템으로는, 비제한적인 예로서, Microsoft® SQL Server, mySQL™ 및 Oracle®을 포함한다. 당업계의 숙련자라면 또한 웹 응용프로그램이, 다양한 실시양태에서, 하나 이상의 언어의 하나 이상의 버전으로 기록된다는 점을 알 수 있을 것이다. 웹 응용프로그램은 하나 이상의 마크업 언어, 표현 정의 언어, 클라이언트 사이드 스크립트 언어, 서버 사이드 코팅 언어, 데이터베이스 질의 언어, 또는 이들의 조합으로 기록될 수 있다. 일부 실시양태에서, 웹 응용프로그램은 어느 정도는 마크업 언어, 예컨대 하이퍼텍스트 마크업 언어 (HTML), 확장성 하이퍼텍스트 마크업 언어 (XHTML), 또는 확장성 마크업 언어 (XML)로 기록된다. 일부 실시양태에서, 웹 응용프로그램은 어느 정도는 표현 정의 언어, 예컨대 케스케이딩 스타일 시트 (CSS)로 기록된다. 일부 실시양태에서, 웹 응용프로그램은 어느 정도는 클라이언트 사이드 스크립트 언어, 예컨대 비동기 자바스크립트 및 XML (AJAX), 플래시® 액션스크립트, 자바스크립트 또는 실버라이트®로 기록된다. 일부 실시양태에서, 웹 응용프로그램은 어느 정도는 서버 사이드 코팅 언어, 예컨대 액티브 서버 페이지 (ASP), ColdFusion®, Perl, 자바™, 자바서버 페이지 (JSP), 하이퍼텍스트 프리프로세서 (PHP), Python™, Ruby, Tcl, Smalltalk, WebDNA® 또는 Groovy로 기록된다. 일부 실시양태에서, 웹 응용프로그램은 어느 정도는 데이터베이스 질의 언어, 예컨대 구조화 질의 언어 (SQL)로 기록된다. 일부 실시양태에서, 웹 응용프로그램은 엔터프라이즈 서버 제품, 예컨대 IBM® Lotus Domino®를 병합한다. 일부 실시양태에서, 웹 응용프로그램은 미디어 플레이어 요소를 포함한다. 추가의 다양한 실시양태에서, 미디어 플레이어 요소는 비제한적인 예로서, 어도비®, 플래시®, HTML 5, 애플®, 퀵타임®, Microsoft®, 실버라이트®, 자바™ 및 Unity®를 비롯한 많은 적절한 멀티미디어기술 중 하나 이상을 이용한다.

[0216] 모바일 응용프로그램

[0217] 일부 실시양태에서, 본원에 개시된 콘티그 어셈블리 방법을 실행하는 컴퓨터 프로그램은 모바일 디지털 처리 장치에 제공된 모바일 응용프로그램을 포함한다. 일부 실시양태에서, 상기 모바일 응용프로그램은 그것이 제조될 때 모바일 디지털 처리 장치에 제공된다. 다른 실시양태에서, 상기 모바일 응용프로그램은 본원에 기술된 컴퓨터 네트워크를 통하여 모바일 디지털 처리 장치에 제공된다.

[0218] 본원에 제공된 개시내용에 비추어 보면, 모바일 응용프로그램은 당업계에 공지된 하드웨어, 언어 및 개발 환경을 이용하여 당업계의 숙련자에게 공지된 기법들에 의해 생성된다. 당업계의 숙련자라면, 모바일 응용프로그램이 수개의 언어로 작성된다는 것을 알 수 있을 것이다. 적절한 프로그래밍 언어로는, 비제한적인 예로서, C, C++, C#, 오브젝티브-C, 자바™, 자바스크립트, 파스칼, 오브젝트 파스칼, Python™, Ruby, VB.NET, WML 및

(CSS를 사용하거나 사용하지 않은) XHTML/HTML, 또는 이들의 조합을 포함한다.

[0219] 적절한 모바일 응용프로그램 개발 환경은 수개의 공급원으로부터 이용가능하다. 시판되는 개발 환경으로는, 비제한적인 예로서, AirplaySDK, alcheMo, Appcelerator®, Celsius, Bedrock, 플래시 라이트, .NET Compact 프레임워크, Rho모바일 및 워크라이트 모바일 플랫폼 (WorkLight Mobile Platform)을 포함한다. 비제한적인 예로서, Lazarus, MobiFlex, MoSync 및 Phonegap을 비롯한 기타의 개발 환경은 비용없이 이용가능하다. 또한, 모바일 장치 제조업자들은 비제한적인 예로서, 아이폰 및 아이패드 (iOS) SDK, 안드로이드™ SDK, 블랙베리® SDK, BREW SDK, Palm® OS SDK, Symbian SDK, webOS SDK 및 윈도우즈® 모바일 SDK를 비롯한 소프트웨어 개발자 키트를 배포한다.

[0220] 당업계의 숙련자라면, 비제한적인 예로서, 애플® 앱 스토어, 안드로이드™ 마켓, 블랙베리® 앱 월드, Palm 장치를 위한 앱 스토어, webOS를 위한 앱 카다로그, 모바일용 윈도우즈® 마켓플레이스, 노키아® 장치를 위한 오비 스토어, 삼성® 앱스 및 닌텐도® DSi 샵을 비롯하여 수개의 상업적인 광장들이, 모바일 응용프로그램의 배부를 위해 이용가능하다는 점을 알 수 있을 것이다.

[0221] 독립형 응용프로그램

[0222] 일부 실시양태에서, 본원에 개시된 콘티그 어셈블리 방법을 실행하는 컴퓨터 프로그램은, 기존의 공정에 대한 애드-온 (부가물), 예컨대 플러그-인이 아닌, 독립적인 컴퓨터 공정으로서 가동되는 프로그램인 독립형 응용프로그램이다. 당업계의 숙련자라면, 독립형 응용프로그램이 컴파일되는 경우가 많다는 점을 알 수 있을 것이다. 컴파일러는 프로그래밍 언어로 작성된 소스 코드를 이진 객체 코드, 예컨대 어셈블리 언어 또는 기계어로 전환하는 컴퓨터 프로그램(들)이다. 적절한 컴파일된 프로그래밍 언어로는, 비제한적인 예로서, C, C++, 오브젝티브-C, COBOL, Delphi, Eiffel, 자바™, Lisp, Python™, 비주얼 베이직 및 VB .NET, 또는 이들의 조합을 포함한다. 컴파일화는 적어도 부분적으로는 실행가능한 프로그램을 생성하기 위해 수행되는 경우가 많다. 일부 실시양태에서, 컴퓨터 프로그램은 하나 이상의 실행가능한 컴파일된 응용프로그램을 포함한다.

[0223] 웹 브라우저 플러그-인

[0224] 일부 실시양태에서, 콘티그 어셈블리 방법은 웹 브라우저 플러그인을 포함한다. 컴퓨팅에 있어서, 플러그-인은 큰 소프트웨어 응용프로그램에 특정 기능성을 부가하는 하나 이상의 소프트웨어 성분이다. 소프트웨어 응용프로그램의 제작자들은 제3자인 개발자들이 응용프로그램을 확장하는 능력을 창출하여, 새로운 특징을 쉽게 부가하는 것을 지원하고, 응용프로그램의 크기를 줄일 수 있도록 플러그-인을 지원한다. 플러그-인이 지원되면, 이는 소프트웨어 응용프로그램의 기능성을 주문 제작할 수 있도록 해준다. 예를 들어, 플러그-인은 비디오를 재생하고, 쌍방향 참가를 생성하고, 바이러스를 스캔하고, 특정 파일 유형을 디스플레이하기 위해 웹 브라우저에서 흔히 사용된다. 당업계의 숙련자라면 어도비®, 플래시® 플레이어, 마이크로소프트® 실버라이트® 및 애플® 퀵타임®을 비롯한 수개의 웹 브라우저 플러그인에 익숙할 것이다. 일부 실시양태에서, 툴바는 하나 이상의 웹 브라우저 익스텐션, 애드-인 또는 애드-온을 포함한다. 일부 실시양태에서, 툴바는 하나 이상의 익스플로러 바, 툴 밴드 또는 데스크 밴드를 포함한다.

[0225] 본원에 제공된 개시내용에 비추어 보면, 당업계의 숙련자라면, 비제한적인 예로서, C++, Delphi, 자바™, PHP, Python™ 및 VB .NET, 또는 이들의 조합을 비롯한 다양한 프로그래밍 언어에서 플러그-인의 개발을 가능하게 하는 수개의 플러그-인 프레임워크가 이용가능하다는 사실을 알 수 있을 것이다.

[0226] 웹 브라우저 (인터넷 브라우저라고도 불림)는 월드 와이드 웹 상에서 정보 자원을 검색, 제시 및 이송시키기 위해서, 네트워크 연결된 디지털 처리 장치와 함께 사용하기 위해 고안된 소프트웨어 응용프로그램이다. 적절한 웹 브라우저로는, 비제한적인 예로서, 마이크로소프트® 인터넷 익스플로러®, 모질라® 파이어폭스®, 구글® 크롬, 애플® 사파리®, 오페라 소프트웨어® 오페라® 및 KDE 캉커러 (Konqueror)를 포함한다. 일부 실시양태에서, 상기 웹 브라우저는 모바일 웹 브라우저이다. 모바일 웹 브라우저 (마이크로브라우저, 미니브라우저 및 무선 브라우저라고도 불림)는 비제한적인 예로서, 포켓용 컴퓨터, 태블릿 컴퓨터, 넷북 컴퓨터, 서브노트북 컴퓨터, 스마트폰, 음악 플레이어, 개인 정보 단말기 (PDA) 및 포켓용 비디오 게임 시스템을 비롯한 모바일 디지털 처리 장치 상에서 사용하기 위해 고안된다. 적절한 모바일 웹 브라우저로는, 비제한적인 예로서, 구글® 안드로이드® 브라우저, RIM 블랙베리® 브라우저, 애플® 사파리®, Palm® Blazer, Palm® WebOS® 브라우저, 모바일용 모질라® 파이어폭스®, 마이크로소프트® 인터넷 익스플로러® 모바일, 아마존® 킨들® 베이직 웹, 노키아® 브라우저, 오페라 소프트웨어® 오페라® 모바일 및 소니® PSP™ 브라우저를 포함한다.

[0227] 소프트웨어 모듈

[0228] 일부 실시양태에서, 본원에 개시된 콘티그 어셈블리 방법은 소프트웨어, 서버 및/또는 데이터베이스 모듈, 또는 이의 사용을 포함한다. 본원에 제공된 개시내용에 비추어 보면, 소프트웨어 모듈은 당업계에 공지된 기계, 소프트웨어 및 언어들을 사용하여 당업계의 숙련자에게 공지된 기법들에 의해 생성된다. 본원에 개시된 소프트웨어 모듈은 여러가지 방법으로 실행된다. 다양한 실시양태에서, 소프트웨어 모듈은 파일, 코드 섹션, 프로그래밍 객체, 프로그래밍 구조, 또는 이들의 조합을 포함한다. 추가의 다양한 실시양태에서, 소프트웨어 모듈은 복수의 파일, 복수의 코드 섹션, 복수의 프로그래밍 객체, 복수의 프로그래밍 구조, 또는 이들의 조합을 포함한다. 다양한 실시양태에서, 하나 이상의 소프트웨어 모듈은 비제한적인 예로서, 웹 응용프로그램, 모바일 응용프로그램 및 독립형 응용프로그램을 포함한다. 일부 실시양태에서, 소프트웨어 모듈은 하나의 컴퓨터 프로그램 또는 응용프로그램 내에 존재한다. 다른 실시양태에서, 소프트웨어 모듈은 하나 이상의 컴퓨터 프로그램 또는 응용프로그램 내에 존재한다. 일부 실시양태에서, 소프트웨어 모듈은 하나의 기계에 대해 호스트된다. 다른 실시양태에서, 소프트웨어 모듈은 하나 이상의 기계에 대하여 호스트된다. 추가의 실시양태에서, 소프트웨어 모듈은 클라우드 컴퓨팅 플랫폼에 대하여 호스트된다. 일부 실시양태에서, 소프트웨어 모듈은 한 위치의 하나 이상의 기계에 대하여 호스트된다. 다른 실시양태에서, 소프트웨어 모듈은 하나 이상의 위치의 하나 이상의 기계에 대하여 호스트된다.

[0229] 데이터베이스

[0230] 일부 실시양태에서, 본원에 개시된 콘티그 어셈블리 방법은 하나 이상의 데이터베이스, 또는 이의 사용을 포함한다. 본원에 제공된 개시내용에 비추어 보면, 당업계의 숙련자라면, 많은 데이터베이스들이 콘티그 정보의 저장과 인출에 적절하다는 사실을 알 수 있을 것이다. 다양한 실시양태에서, 적절한 데이터베이스로는, 비제한적인 예로서, 관계형 데이터베이스, 비관계형 데이터베이스, 객체 지향형 데이터베이스, 객체 데이터베이스, 개체 관계 모델 데이터베이스, 결합형 데이터베이스 및 XML 데이터베이스를 포함한다. 일부 실시양태에서, 데이터베이스는 인터넷을 기반으로 한다. 추가의 실시양태에서, 데이터베이스는 웹을 기반으로 한다. 추가의 실시양태에서, 데이터베이스는 클라우드 컴퓨팅을 기반으로 한다. 다른 실시양태에서, 데이터베이스는 하나 이상의 로컬 컴퓨터 저장 장치를 기반으로 한다.

[0231] 실시양태의 열거

[0232] 본원의 개시내용을 번호를 매긴 실시양태들의 비제한적인 리스트로서 추가로 제공한다.

[0233] 1. 초기 구성을 갖는 콘티그 서열의 세트를 수득하는 단계; 페어드 엔드 리드의 세트를 수득하는 단계; 표준 페어드 엔드 리드 거리 빈도 데이터를 수득하는 단계;

[0234] 하나 이상의 페어드 엔드 리드 내에 공존하는 서열을 공유하는 콘티그 쌍을 그룹화하는 단계; 및

[0235] 상기 그룹화된 콘티그 서열을, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드 쌍 거리 빈도 데이터가 초기 구성에서 콘티그 서열의 리드 쌍 빈도 데이터에 대한 표준 페어드 엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하도록 스캐폴딩하는 단계를 포함하는, 핵산 서열 정보의 콘티그를 스캐폴딩하는 방법.

[0236] 2. 제1 실시양태에 있어서, 상기 스캐폴딩이 콘티그의 세트를 순서 배열하는 단계를 포함하는 것인, 방법.

[0237] 3. 제1 실시양태에 있어서, 상기 스캐폴딩이 콘티그의 세트를 배향시키는 단계를 포함하는 것인, 방법.

[0238] 4. 제1 실시양태에 있어서, 상기 스캐폴딩이 2종 이상의 콘티그를 한줄로 병합시키는 단계를 포함하는 것인, 방법.

[0239] 5. 제1 실시양태에 있어서, 상기 스캐폴딩이 하나의 콘티그를 제2 콘티그에 삽입하는 단계를 포함하는 것인, 방법.

[0240] 6. 제1 실시양태에 있어서, 상기 스캐폴딩이 하나의 콘티그를 2개 이상의 성분 콘티그로 절단하는 단계를 포함하는 것인, 방법.

[0241] 7. 제1 실시양태에 있어서, 상기 표준 페어드 엔드 리드 빈도는, 양자 리드가 공통 콘티그에 매핑되는 페어드 엔드 리드로부터 수득되는 것인, 방법.

[0242] 8. 제1 실시양태에 있어서, 상기 표준 페어드 엔드 리드 빈도가 사전에 생성된 곡선으로부터 수득되는 것인, 방법.

[0243] 9. 제1 실시양태에 있어서, 상기 초기의 구성이 무작위 구성인 것인, 방법.

[0244] 10. 제1 실시양태에 있어서, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드 쌍 거리 빈도 데이터가, 리드 쌍 거리 우도가 증가하는 경우에 페어드 엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하는 것인, 방법.

[0245] 11. 제10 실시양태에 있어서, 상기 리드 쌍 거리 우도가 최대화되는 것인, 방법.

[0246] 12. 제1 실시양태에 있어서, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드 쌍 거리 빈도 데이터가, 리드 쌍 거리 빈도 데이터와 표준 페어드 엔드 리드 거리 빈도 데이터 간의 차이에 대한 통계 측정값이 감소하는 경우에 페어드 엔드 리드 거리 빈도 데이터에 보다 더 가깝게 근접하게 되는 것인, 방법.

[0247] 13. 제12 실시양태에 있어서, 리드 쌍 거리 빈도 데이터와 표준 페어드 엔드 리드 거리 빈도 데이터 간의 거리에 대한 통계 측정값이 ANOVA, t-검정 및 X-제곱 검정 중 하나 이상을 포함하는 것인, 방법.

[0248] 14. 제1 실시양태에 있어서, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드 쌍 거리가, 표준 페어드 엔드 리드 거리 빈도와 비교될 때, 수득되는 순서 배열된 콘티그들 간의 리드 쌍 거리 분포의 편차가 감소하는 경우에 페어드 엔드 리드 거리 빈도 데이터에 보다 더 가깝게 매칭되는 것인, 방법.

[0249] 15. 제14 실시양태에 있어서, 표준 페어드 엔드 리드 거리 빈도와 비교될 때, 수득되는 순서 배열된 콘티그들 간의 리드 쌍 거리 분포의 편차가 최소화되는 것인, 방법.

[0250] 16. 제1 실시양태에 있어서, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그가 아주 많은 수의 공유된 엔드 리드를 갖는 클러스터로 지정되는 것인, 방법.

[0251] 17. 실시양태 1-16 중 어느 한 실시양태에 있어서, 상기 클러스터링이 유기체 내 염색체의 수보다 많거나 염색체 수와 같은 다수의 그룹에 콘티그를 배치하는 단계를 포함하는 것인, 방법.

[0252] 18. 실시양태 1-17 중 어느 한 실시양태에 있어서, 클러스터 중 하나의 콘티그와 단 하나의 페어드 엔드 리드만을 공유하는 콘티그는 해당 클러스터에 포함되지 않는 것인, 방법.

[0253] 19. 실시양태 1-18 중 어느 한 실시양태에 있어서, 반복 서열을 포함하는 하나 이상의 페어드 엔드 리드만을 클러스터와 공유하는 콘티그는 해당 클러스터에 포함되지 않는 것인, 방법.

[0254] 20. 실시양태 1-19 중 어느 한 실시양태에 있어서, 저품질의 서열을 포함하는 하나 이상의 페어드 엔드 리드만을 클러스터와 공유하는 콘티그는 해당 클러스터에 포함되지 않는 것인, 방법.

[0255] 21. 실시양태 1-20 중 어느 한 실시양태에 있어서, 상기 페어드 엔드 리드의 세트가 샘플 DNA를 분해하여 핵산 내에 내부 이중 가닥 절단부를 생성하는 단계, 상기 이중 가닥 절단부를 재결찰시켜 하나 이상의 재결찰 접합부를 형성하는 단계, 및 상기 하나 이상의 재결찰 접합부에 대해 시퀀싱하는 단계에 의해 수득되는 것인, 방법.

[0256] 22. 제21 실시양태에 있어서, 상기 DNA가 하나 이상의 DNA 결합제에 가교결합되는 것인, 방법.

[0257] 23. 제21 실시양태에 있어서, 상기 DNA가 단리된 네이키드 DNA인 것인, 방법.

[0258] 24. 제22 실시양태에 있어서, 상기 단리된 네이키드 DNA가 재구성된 염색질 내에 리어셈블링되는 것인, 방법.

[0259] 25. 제24 실시양태에 있어서, 상기 재구성된 염색질이 가교결합되는 것인, 방법.

[0260] 26. 제23 실시양태에 있어서, 상기 재구성된 염색질이 DNA 결합 단백질을 포함하는 것인, 방법.

[0261] 27. 제23 실시양태에 있어서, 상기 재구성된 염색질이 나노입자를 포함하는 것인, 방법.

[0262] 28. 실시양태 1-27 중 어느 한 실시양태에 있어서, 상기 콘티그의 클러스터링이 유기체에 대한 염색체의 수와는 독립적인 것인, 방법.

[0263] 29. 실시양태 1-28 중 어느 한 실시양태에 있어서, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그가 아주 많은 수의 공유된 엔드 리드를 갖는 클러스터로 지정되는 것인, 방법.

[0264] 30. 실시양태 1-28 중 어느 한 실시양태에 있어서, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그가 더 큰 리드 쌍 거리 우도 값을 갖는 클러스터로 지정되는 것인, 방법.

[0265] 31. 실시양태 1-28 중 어느 한 실시양태에 있어서, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그가 표준 리드 쌍 거리 분포에 대하여 그의 리드 쌍 분포 내에 더 낮은 편차를 갖는 클러스터로 지정되는 것인, 방법.

[0266] 32. 실시양태 1-29 중 어느 한 실시양태에 있어서, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그가 각 클러스터로부터 배제되는 것인, 방법.

[0267] 33. 실시양태 1-32 중 어느 한 실시양태에 있어서, 상기 클러스터링이 유기체 내 염색체의 수보다 많거나 염색체 수와 같은 다수의 그룹에 콘티그를 배치하는 단계를 포함하는 것인, 방법.

[0268] 34. 실시양태 1-33 중 어느 한 실시양태에 있어서, 상기 스캐폴딩이 상기 클러스터된 콘티그 중 인접한 것으로 추정되는 콘티그의 제1 세트를 선택하는 단계, 리드 쌍에 대한 상기 리드 쌍 거리의 총 측정값을 감소시키는, 인접한 것으로 추정되는 콘티그의 상기 제1 세트의 최소 거리 순서를 결정하는 단계, 및 상기 리드 쌍 거리의 상기 총 측정값을 감소시키기 위해 인접한 것으로 추정되는 콘티그의 상기 제1 세트를 스캐폴딩하는 단계를 포함하는 것인, 방법.

[0269] 35. 제34 실시양태에 있어서, 상기 인접한 것으로 추정되는 콘티그의 상기 제1 세트가 2개의 콘티그로 이루어지는 것인, 방법.

[0270] 36. 제34 실시양태에 있어서, 상기 인접한 것으로 추정되는 콘티그의 상기 제1 세트가 3개의 콘티그로 이루어지는 것인, 방법.

[0271] 37. 제34 실시양태에 있어서, 상기 인접한 것으로 추정되는 콘티그의 상기 제1 세트가 4개의 콘티그로 이루어지는 것인, 방법.

[0272] 38. 제34 실시양태에 있어서, 상기 인접한 것으로 추정되는 콘티그의 상기 제1 세트가 4개의 콘티그를 포함하는 것인, 방법.

[0273] 39. 제34 실시양태에 있어서, 상기 스캐폴딩이 인접한 것으로 추정되는 콘티그의 상기 제1 세트 중 각 콘티그의 순서와 배향을 결정하는 단계를 포함하는 것인, 방법.

[0274] 40. 제34 또는 제35 실시양태에 있어서, 상기 최소 거리 순서의 결정은 상기 세트의 두 콘티그에 매핑되는 리드를 포함하는 하나 이상의 리드 쌍에 대해 예측된 리드 쌍 거리를 가능한 모든 콘티그 구성에 대해 비교하는 단계를 포함하는 것인, 방법.

[0275] 41. 제40 실시양태에 있어서, 상기 리드 쌍에 대한 최소 리드 쌍에 상응하는 콘티그 배향을 선택하는 단계를 더 포함하는 것인, 방법.

[0276] 42. 제40 실시양태에 있어서, 최대 우도 리드 쌍 거리 분포에 상응하는 콘티그 배향을 선택하는 단계를 더 포함하는 것인, 방법.

[0277] 43. 제40 또는 제41 실시양태에 있어서, 상기 콘티그 클러스터의 리드 쌍의 총 측정값에 대한 최소 리드 쌍에 상응하는 콘티그 배향을 선택하는 단계를 더 포함하는 것인, 방법.

[0278] 44. 실시양태 40-43 중 어느 한 실시양태에 있어서, 상기 예측된 리드 쌍 거리를 상기 페어드 엔드 리드 거리 빈도 데이터와 비교하는 것인, 방법.

[0279] 45. 제44 실시양태에 있어서, 상기 페어드 엔드 리드 거리 빈도 데이터에 대한 상기 비교가 식 1의 이용을 포함하는 것인, 방법.

[0280] 46. 실시양태 34-45 중 어느 한 실시양태에 있어서, 상기 클러스터된 콘티그 중 인접한 것으로 추정되는 콘티그의 제 2세트를 선택하는 단계로서, 상기 제2 세트가 상기 제1 세트 중 하나의 엔드 말단 콘티그를 제외하고 모두 포함하고, 상기 클러스터된 콘티그 중 하나의 추가 콘티그를 포함하는 것인 단계, 및 리드 쌍 거리의 상기 총 측정값을 감소시키기 위해 인접한 것으로 추정되는 콘티그의 상기 제2 세트를 스캐폴딩하는 단계를 포함하는 것인, 방법.

[0281] 47. 제46 실시양태에 있어서, 상기 클러스터된 콘티그 중 인접한 것으로 추정되는 콘티그의 제3 세트를 선택하는 단계로서, 상기 제3 세트가 상기 제2 세트 중 하나의 엔드 말단 콘티그를 제외하고 모두 포함하고, 상기 제1 세트에 포함되지 않고 상기 제2 세트에 포함되지 않는 상기 클러스터된 콘티그 중 하나의 추가 콘티그를 포함하는 단계, 및 리드 쌍 거리의 상기 총 측정값을 감소시키기 위해 인접한 것으로 추정되는 콘티그의 상기 제3 세트를 스캐폴딩하는 단계를 포함하는 것인, 방법.

[0282] 48. 제47 실시양태에 있어서, 대다수의 상기 클러스터된 콘티그가 순서 배열될 때까지 하나 이상의 추가의 세트를 반복적으로 선택하는 단계를 더 포함하는 것인, 방법.

[0283] 49. 제48 실시양태에 있어서, 각각의 상기 클러스터된 콘티그가 순서 배열될 때까지 하나 이상의 추가의 세트를 반복적으로 선택하는 단계를 더 포함하는 것인, 방법.

[0284] 50. 실시양태 1-49 중 어느 한 실시양태에 있어서, 핵산 서열이 하나의 게놈으로부터 유래되는 것인, 방법.

[0285] 51. 실시양태 1-49 중 어느 한 실시양태에 있어서, 핵산 서열이 복수의 게놈으로부터 유래되는 것인, 방법.

[0286] 52. a) 각 콘티그 쌍에 대한 로그 우도비 스코어를 지정하는 단계; b) 비율 스코어에 의해 관련성을 분류하는 단계; 및 c) 어셈블리의 전체 스코어를 증가시키기 위해 비율 스코어의 내림차순으로 콘티그 관련성을 수용 또는 거부하는 단계를 포함하는, 클러스터 내의 콘티그를 스캐폴딩하는 방법.

[0287] 53. 제52 실시양태에 있어서, 상기 스캐폴딩이 콘티그의 세트를 순서 배열하는 단계를 포함하는 것인, 방법.

[0288] 54. 제52 실시양태에 있어서, 상기 스캐폴딩이 콘티그의 세트를 배향시키는 단계를 포함하는 것인, 방법.

[0289] 55. 제52 실시양태에 있어서, 상기 스캐폴딩이 2종 이상의 콘티그를 한줄로 병합시키는 단계를 포함하는 것인, 방법.

[0290] 56. 제52 실시양태에 있어서, 상기 스캐폴딩이 하나의 콘티그를 제2 콘티그에 삽입하는 단계를 포함하는 것인, 방법.

[0291] 57. 제52 실시양태에 있어서, 상기 스캐폴딩이 하나의 콘티그를 2개 이상의 성분 콘티그로 절단하는 단계를 포함하는 것인, 방법.

[0292] 58. 제52 실시양태에 있어서, 상기 콘티그가 하나의 게놈을 포함하는 것인, 방법.

[0293] 59. 제52 실시양태에 있어서, 상기 콘티그가 복수의 게놈을 포함하는 것인, 방법.

[0294] 60. a) 콘티그의 클러스터를 따라 i 위치에서 출발하는 크기 w 콘티그의 서열 윈도우를 확인하는 단계; b) 윈도우에서 각 i 위치에서 양립가능한 순서와 배향에 대한 스코어를 검사함으로써 윈도우 w의 콘티그에 대한 $w!\ 2^w$ 순서 및 배향 옵션을 고려하는 단계; c) 상기 윈도우 내 상기 w 콘티그를 배향하고 순서 배열하여 최적의 스코어를 수득하는 단계; d) 상기 윈도우를 i+1 위치로 이동시키는 단계; 및 e) 상기 w 콘티그의 배향 및 순서 배열을 이용하여 i+1 위치에서 상기 윈도우에 대하여 단계 (a), (b) 및 (c)를 반복하여 최적의 스코어를 결정함으로써, 상기 복수의 콘티그를 스코어에 대해 국소적으로 최적 구성으로 배향하고 순서 배열하는 단계를 포함하는, 클러스터 내의 복수의 콘티그에 대한 국소적으로 최적의 콘티그 구성을 결정하는 방법.

[0295] 61. 제60 실시양태에 있어서, 클러스터 내의 복수의 콘티그에 매핑된 리드 쌍 데이터가 수득되고, 표준 페어드 엔드 리드 빈도 데이터 세트가 수득되며, 상기 w 콘티그의 배향 및 순서 배열에 대한 스코어가 클러스터 내의 복수의 콘티그에 매핑된 리드 쌍 데이터에 대한 리드 쌍 거리 데이터 세트가 표준 페어드 엔드 리드 빈도 데이터 세트에 얼마나 가깝게 매칭되는지에 대한 척도인 것인, 방법.

[0296] 62. 제60 실시양태에 있어서, 클러스터 내의 복수의 콘티그에 매핑된 리드 쌍 데이터가 수득되고, 상기 스코어가 전체 리드 쌍 거리이며, 상기 스코어가 전체 리드 쌍 거리가 최소화될 때 스코어가 최적인, 방법.

[0297] 63. 제60 실시양태에 있어서, w가 3인 것인, 방법.

[0298] 64. 제60 실시양태에 있어서, w가 4인 것인, 방법.

[0299] 65. 제60 실시양태에 있어서, w가 5인 것인, 방법.

[0300] 66. 제60 실시양태에 있어서, w가 6인 것인, 방법.

[0301] 67. 제60 실시양태에 있어서, w가 제1 클러스터에 대한 제1값을 가지며, w가 제2 클러스터에서 제2값을 갖는 것인, 방법.

[0302] 68. 제60 실시양태에 있어서, w가 상기 세트 중 1%의 콘티그를 포함하도록 선택되는 것인, 방법.

[0303] 69. 제60 실시양태에 있어서, w가 상기 세트 중 5%의 콘티그를 포함하도록 선택되는 것인, 방법.

[0304] 70. 제60 실시양태에 있어서, w가 상기 세트 중 10%의 콘티그를 포함하도록 선택되는 것인, 방법.

[0305] 71. 제60 실시양태에 있어서, 상기 스코어가 리드 쌍 거리 우도 스코어이고, 상기 스코어는 주어진 윈도우 크기에 대해 최대화될 때 최적인, 방법.

[0306] 72. 제70 실시양태에 있어서, 상기 스코어가 식 1을 이용하여 계산되는 것인, 방법.

[0307] 73. 제60 실시양태에 있어서, 상기 스코어가 예측된 리드 쌍 분포로부터의 편차이고, 해당 스코어가 주어진 윈도우 크기에 대해 최소화될 때 스코어가 최적인, 방법.

[0308] 74. 제60-73 실시양태 중 한 실시양태에 있어서, 상기 복수의 콘티그가 하나의 게놈을 포함하는 것인, 방법.

[0309] 75. 제60-73 실시양태 중 한 실시양태에 있어서, 상기 복수의 콘티그가 복수의 게놈을 포함하는 것인, 방법.

[0310] 76. a) 정제된 DNA를 수득하는 단계; b) 상기 정제된 DNA를 DNA 결합제와 결합시켜 DNA/염색질 복합체를 형성하는 단계; c) 상기 DNA-염색질 복합체를 제한 효소와 함께 항온배양시켜 점착성 말단을 남기는 단계; d) 결찰을 수행하여 DNA의 말단을 결합시키는 단계; e) 결찰된 DNA 접합부에 대해 시퀀싱하여 페어드 엔드 리드를 생성하는 단계; 및 f) 상기 페어드 엔드 리드를 사용하여 상기 정제된 DNA의 서열을 나타내는 콘티그를 포함하는 핵산 데이터 세트를 스캐폴드하는 단계를 포함하는, 핵산 서열 어셈블리 방법.

[0311] 77. 제76 실시양태에 있어서, 상기 정제된 DNA가 하나의 게놈으로부터 유래되는 것인, 방법.

[0312] 78. 제76 실시양태에 있어서, 상기 정제된 DNA가 복수의 게놈으로부터 유래되는 것인, 방법.

[0313] 79. 핵산 샘플에 대한 콘티그 데이터세트를 수득하는 단계; 물리적으로 연결된 비인접성 서열 정보에 상응하는 하나 이상의 리드 쌍 서열 리드를 수득하는 단계; 및 리드 쌍 서열 리드 중 하나 이상의 리드가 하나의 콘티그 데이터 세트 중 2개의 별개 위치에 매핑되는 경우, 해당 리드 쌍 서열 리드를 배제시키는 단계를 포함하는, 반복성 콘티그 영역에 매핑되는 리드 쌍 서열 리드를 확인하는 방법.

[0314] 80. 제79 실시양태에 있어서, 상기 반복성 영역이 제1 역치를 초과하는 샷건 리드 깊이 (shotgun read depth)를 갖는 서열을 포함하는 것인, 방법.

[0315] 81. 제80 실시양태에 있어서, 상기 반복성 영역이 제2 역치를 초과하는 리드 깊이를 갖는 염기 위치를 포함하는 것인, 방법.

[0316] 82. 제81 실시양태에 있어서, 상기 제1 역치 및 제2 역치가 리드 깊이의 전체적인 분포에 대하여 고정되는 것인, 방법.

[0317] 83. 제82 실시양태에 있어서, 상기 제1 역치가 리드 깊이의 전체적인 분포의 3배인 것인, 방법.

[0318] 84. 제82 실시양태에 있어서, 상기 제2 역치가 리드 깊이의 전체적인 분포의 3.5배인 것인, 방법.

[0319] 85. 제79-84 실시양태 중 어느 한 실시양태에 있어서, 상기 핵산 샘플이 하나의 게놈을 포함하는 것인, 방법.

[0320] 86. 제79-84 실시양태 중 어느 한 실시양태에 있어서, 상기 핵산 샘플이 복수의 게놈을 포함하는 것인, 방법.

[0321] 87. 제1 콘티그와 제2 콘티그 간에 걸쳐있는 리드 쌍 서열의 개수 및 정황상 추정되는 분리를 관찰할 확률을 측정하는 단계로서, 상기 콘티그가 세트 [++,+-,-+,--] 내에서 o의 상대적 배향을 가지며 갭 길이에 의해 분리되는 것인 단계를 포함하는, 콘티그 어셈블리 결정을 안내하는 방법.

[0322] 88. 제87 실시양태에 있어서, 분리 거리에 대해 리드 쌍 서열 분포의 확률을 정규화하는 단계로서, 상기 정규화가 상기 리드 쌍 서열을, 핵산 샘플을 독립적으로 표본으로 삼은 노이즈 쌍과 비교하는 단계를 포함하는 것인 단계를 포함하는 것인, 방법.

[0323] 89. 제88 실시양태에 있어서, 상기 핵산 샘플이 하나의 게놈을 포함하는 것인, 방법.

[0324] 90. 제88 실시양태에 있어서, 상기 핵산 샘플이 복수의 게놈을 포함하는 것인, 방법.

[0325] 91. 제88 실시양태에 있어서, 상기 노이즈 쌍의 총수가 콘티그 쌍의 샘플에 대한 링크의 밀도를 도표화함으로써 측정되는 것인, 방법.

[0326] 92. 제91 실시양태에 있어서, 최고 밀도값 1%와 최저 밀도값 1%가 배제되는 것인, 방법.

[0327] 93. 제87 실시양태에 있어서, 콘티그 순서를 결정하는 단계를 더 포함하는 것인, 방법.

[0328] 94. 제87 실시양태에 있어서, 콘티그 배향을 결정하는 단계를 더 포함하는 것인, 방법.

[0329] 95. 초기 구성을 갖는 콘티그 서열의 세트를 수득하는 단계; 페어드 엔드 리드의 세트를 수득하는 단계; 표준 페어드 엔드 리드 거리 빈도 데이터를 수득하는 단계; 하나 이상의 페어드 엔드 리드 내에 공존하는 서열을 공

유하는 콘티그 쌍을 그룹화하는 단계; 상기 콘티그의 그룹에 대한 리드 쌍 빈도 데이터를 표준 페어드 엔드 리드 거리 빈도 데이터와 비교하는 단계; 상기 그룹 중 하나의 콘티그에 절단부를 도입하는 것이 상기 콘티그의 그룹에 대한 리드 쌍 빈도 데이터를 표준 페어드 엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하도록 하는지의 여부를 측정하는 단계; 및 상기 콘티그의 그룹에 대한 리드 쌍 빈도 데이터가 표준 페어드 엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하는 경우에, 상기 콘티그에 절단부를 도입하는 단계를 포함하는, 콘티그 부정합 보정 방법.

[0330] 96. 제95 실시양태에 있어서, 제1 위치가 절단부를 도입하기 전의 상기 역치 미만의 상기 로그 우도를 갖는 하나 이상의 인접한 제2 위치와 병합되는 것인, 방법.

[0331] 97. 제95 실시양태에 있어서, 인접한 제2 위치가 제1 위치로부터 300개 이하의 염기 쌍만큼 떨어져 있는 것인, 방법.

[0332] 98. 제95 실시양태에 있어서, 제2 위치가 제1 위치로부터 1000개 초과의 염기 쌍만큼 떨어진 위치를 포함하지 않는 것인, 방법.

[0333] 99. 제95-98 실시양태 중 어느 한 실시양태에 있어서, 상기 로그 우도 변화를 측정하는 단계가 하나의 콘티그에 대한 평균 페어드 엔드 매핑 밀도를 확인하는 단계, 평균 페어드 엔드 매핑 밀도의 3x 이상의 페어드 엔드 매핑 밀도를 갖는 콘티그의 세그먼트를 확인하는 단계, 및 평균 페어드 엔드 매핑 밀도의 3x 이상의 페어드 엔드 매핑 밀도를 갖는 콘티그의 세그먼트를 배제시키는 단계를 포함하는 것인, 방법.

[0334] 100. 제95-99 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 하나의 게놈으로부터 유래되는 것인, 방법.

[0335] 101. 제95-99 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 복수의 게놈으로부터 유래되는 것인, 방법.

[0336] 102. 출발 어셈블리의 절단된 콘티그를 표시하는 단계로서, 상기 절단된 콘티그가 노드이고, 상기 절단된 콘티그의 엣지가 정수의 순서쌍의 목록으로 표지되며, 상기 절단부의 엣지가 매핑된 리드 쌍 서열에 상응하는 것인 단계; 및 역치 수 미만의 매핑된 연결부를 갖는 엣지를 배제시키는 단계를 포함하는, 콘티그 어셈블리 방법.

[0337] 103. 제102 실시양태에 있어서, 상기 역치 수가 5% 미만인 것인, 방법.

[0338] 104. 제102 실시양태에 있어서, 상기 역치 수가 t_L 링크보다 적은 것인, 방법.

[0339] 105. 제102 실시양태에 있어서, 상기 콘티그가, 해당 콘티그 길이에 대하여 상응하는 노드의 그래프 내 차수의 비율이 모든 분포값들 중 약 5%의 최고치를 초과하는 염기 쌍인 엣지를 포함하는 것인, 방법.

[0340] 106. 제102-105 실시양태에 있어서, 상기 콘티그가 하나의 게놈으로부터 유래되는 것인, 방법.

[0341] 107. 제102-105 실시양태에 있어서, 상기 콘티그가 복수의 게놈으로부터 유래되는 것인, 방법.

[0342] 108. 복수의 콘티그에 상응하는 서열 정보를 수득하는 단계, 상기 복수의 콘티그에 의해 나타난 핵산 샘플의 페어드 엔드 리드 정보를 수득하는 단계, 및 예측된 리드 쌍 거리 데이터 세트의 리드 쌍 거리 파라미터의 편차가 최소화되도록 상기 복수의 콘티그를 배치하는 단계로서, 상기 배치가 8시간 미만으로 일어나는 것인 단계를 포함하는, 콘티그 서열 정보를 하나 이상의 스캐폴드에 어셈블링하는 방법.

[0343] 109. 제108 실시양태에 있어서, 상기 예측된 리드 쌍 거리 데이터 세트가 리드 쌍 거리 우도 곡선을 포함하는 것인, 방법.

[0344] 110. 제108 또는 제109 실시양태에 있어서, 상기 리드 쌍 거리 파라미터가 리드 쌍 거리 우도 곡선에 대한 최대 거리 우도인 것인, 방법.

[0345] 111. 제108 또는 제109 실시양태에 있어서, 상기 리드 쌍 거리 파라미터가 리드 쌍 거리 우도 곡선에 대한 최소 변이인 것인, 방법.

[0346] 112. 제108-111 실시양태 중 어느 한 실시양태에 있어서, 상기 국소적으로 인접한 콘티그의 세트가 2개의 콘티그를 포함하는 것인, 방법.

[0347] 113. 제108-111 실시양태 중 어느 한 실시양태에 있어서, 상기 국소적으로 인접한 콘티그의 세트가 3개의 콘티그를 포함하는 것인, 방법.

[0348] 114. 제108-111 실시양태 중 어느 한 실시양태에 있어서, 상기 국소적으로 인접한 콘티그의 세트가 4개의 콘티그를 포함하는 것인, 방법.

[0349] 115. 제108-111 실시양태 중 어느 한 실시양태에 있어서, 상기 국소적으로 인접한 콘티그의 세트가 5개의 콘티그를 포함하는 것인, 방법.

[0350] 116. 제108-111 실시양태 중 어느 한 실시양태에 있어서, 상기 국소적으로 인접한 콘티그의 세트가 6개의 콘티그를 포함하는 것인, 방법.

[0351] 117. 제108-116 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 7시간 미만으로 일어나는 것인, 방법.

[0352] 118. 제108-116 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 6시간 미만으로 일어나는 것인, 방법.

[0353] 119. 제108-116 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 5시간 미만으로 일어나는 것인, 방법.

[0354] 120. 제108-116 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 4시간 미만으로 일어나는 것인, 방법.

[0355] 121. 제108-116 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 3시간 미만으로 일어나는 것인, 방법.

[0356] 122. 제108-116 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 2시간 미만으로 일어나는 것인, 방법.

[0357] 123. 제108-116 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 1시간 미만으로 일어나는 것인, 방법.

[0358] 124. 제108-123 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 정보가 하나의 게놈으로부터 유래되는 것인, 방법.

[0359] 125. 제108-123 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 정보가 복수의 게놈으로부터 유래되는 것인, 방법.

[0360] 126. 핵산 샘플을 대표하는 콘티그 서열 세트를 수득하는 단계, 상기 핵산 샘플에 대한 리드 쌍 데이터를 수득하는 단계, 및 핵산 샘플에 대한 리드 쌍 데이터가 예측된 리드 쌍 분포에 보다 가깝게 근접하도록 콘티그 서열 세트를 순서 배열하여 배향시키는 단계로서, 상기 콘티그 서열 세트 중 70%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 8시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인 단계를 포함하는, 콘티그 서열 세트를 스캐폴딩하는 방법.

[0361] 127. 제126 실시양태에 있어서, 상기 스캐폴딩이 콘티그의 세트를 순서 배열하는 단계를 포함하는 것인, 방법.

[0362] 128. 제126 실시양태에 있어서, 상기 스캐폴딩이 콘티그의 세트를 배향시키는 단계를 포함하는 것인, 방법.

[0363] 129. 제126 실시양태에 있어서, 상기 스캐폴딩이 2종 이상의 콘티그를 한줄로 병합시키는 단계를 포함하는 것인, 방법.

[0364] 130. 제126 실시양태에 있어서, 상기 스캐폴딩이 하나의 콘티그를 제2 콘티그에 삽입하는 단계를 포함하는 것인, 방법.

[0365] 131. 제126 실시양태에 있어서, 상기 스캐폴딩이 하나의 콘티그를 2개 이상의 성분 콘티그로 절단하는 단계를 포함하는 것인, 방법.

[0366] 132. 제126 실시양태에 있어서, 콘티그 서열 세트 중 80%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 8시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인, 방법.

[0367] 133. 제126 실시양태에 있어서, 콘티그 서열 세트 중 90%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 8시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인, 방법.

[0368] 134. 제126 실시양태에 있어서, 콘티그 서열 세트 중 95%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 8시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인, 방법.

[0369] 135. 제126 실시양태에 있어서, 콘티그 서열 세트 중 70%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 4시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인, 방법.

[0370] 136. 제126 실시양태에 있어서, 콘티그 서열 세트 중 70%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 2시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인, 방법.

[0371] 137. 제126 실시양태에 있어서, 콘티그 서열 세트 중 70%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향

을 1시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인, 방법.

[0372] 138. 제126-137 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 하나의 게놈으로부터 유래되는 것인, 방법.

[0373] 139. 제126-137 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 복수의 게놈으로부터 유래되는 것인, 방법.

[0374] 140. 복수의 콘티그에 상응하는 서열 정보를 수득하는 단계, 페어드 엔드 리드 정보를 수득하는 단계, 및 상기 페어드 엔드 리드 정보에 대한 페어드 엔드 리드 거리 분포가 전반적으로 최적화되어 참조 페어드 엔드 리드 거리 분포에 근접하도록 하는 단계로서, 상기 배치가 8시간 미만으로 일어나는 것인 단계를 포함하는, 핵산 서열 데이터 세트를 구성하는 방법.

[0375] 141. 제140 실시양태에 있어서, 상기 핵산 서열 데이터의 세트가 하나의 게놈으로부터 유래되는 것인, 방법.

[0376] 142. 제140 실시양태에 있어서, 상기 핵산 서열 데이터의 세트가 복수의 게놈으로부터 유래되는 것인, 방법.

[0377] 143. 제140 실시양태에 있어서, 상기 배치가 4시간 미만으로 일어나는 것인, 방법.

[0378] 144. 제140 실시양태에 있어서, 상기 배치가 2시간 미만으로 일어나는 것인, 방법.

[0379] 145. 복수의 결합된 노드쌍을 포함하는 스캐폴드 세트를 수득하는 단계로서, 상기 노드 쌍의 각 노드가 하나 이상의 콘티그 서열을 포함하는 것인 단계, 상기 복수의 결합된 노드에 매핑된 페어드 엔드 리드 정보를 수득하는 단계, 결합된 노드쌍에 의해 공유된 리드 쌍의 개수를 계수하는 단계, 상기 개수를 역치와 비교하는 단계, 및 상기 개수가 역치 미만인 경우 노드 쌍을 결합되지 않은 노드로 절단하는 단계를 포함하는, 스캐폴드 어셈블리를 개선하는 방법.

[0380] 146. 제145 실시양태에 있어서, 고유의 콘티그 서열에 매핑된 리드 쌍만이 계수되는 것인, 방법.

[0381] 147. 제145 실시양태에 있어서, 역치수의 별개의 리드 페어드 엔드가 매핑되는 콘티그 서열 세그먼트에 매핑된 리드 쌍이 폐기되는 것인, 방법.

[0382] 148. 제145 실시양태에 있어서, 상기 역치 수가 비반복성 서열에 대한 평균 수의 3x인 것인, 방법.

[0383] 149. 제145-148 실시양태 중 어느 한 실시양태에 있어서, 상기 스캐폴드 세트가 하나의 게놈으로부터 유래되는 것인, 방법.

[0384] 150. 제145-148 실시양태 중 어느 한 실시양태에 있어서, 상기 스캐폴드 세트가 복수의 게놈으로부터 유래되는 것인, 방법.

[0385] 151. 복수의 결합된 노드쌍을 포함하는 스캐폴드 세트를 수득하는 단계로서, 상기 노드 쌍의 각 노드가 하나 이상의 콘티그 서열을 포함하는 것인 단계, 상기 복수의 결합된 노드에 매핑된 페어드 엔드 리드 정보를 수득하는 단계, 표준 페어드 엔드 리드 거리 빈도 데이터를 수득하는 단계; 상기 복수의 결합된 노드에 매핑된 페어드 엔드 리드 정보에 대한 페어드 엔드 리드 빈도 데이터를 표준 페어드 엔드 리드 거리 빈도 데이터와 비교하는 단계; 및 결합된 노드를 절단하는 것이 상기 복수의 결합된 노드에 매핑된 페어드 엔드 리드 정보에 대한 페어드 엔드 리드 빈도 데이터를 표준 페어드 엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하도록 유도하는 경우에 하나 이상의 결합된 노드를 절단하는 단계를 포함하는, 스캐폴드 어셈블리를 개선하는 방법.

[0386] 152. 제151 실시양태에 있어서, 상기 스캐폴드 세트가 하나의 게놈을 포함하는 것인, 방법.

[0387] 153. 제151 실시양태에 있어서, 상기 스캐폴드 세트가 복수의 게놈을 포함하는 것인, 방법.

[0388] 154. 콘티그 서열 세트를 수득하는 단계, 페어드 엔드 리드 세트를 포함하는 입력 데이터를 수득하는 단계로서, 상기 페어드 엔드 리드 중 1% 이상이 1 kb 이상의 리드 쌍 거리를 포함하고, 상기 페어드 엔드 리드의 세트가 자연적인 배향의 페어드 엔드 리드를 포함하며, 상기 리드 쌍에 대한 시퀀싱 에러율이 0.1% 이하이고, 상기 입력 데이터의 RN50이 어셈블링된 스캐폴드의 20% 이하인 것인 단계, 및 스캐폴드를 출력하는 단계로서, 스캐폴드에 대한 RN50이 입력 데이터의 RN50의 2x 이상인 것인 단계를 포함하는, 스캐폴드 어셈블리 방법.

[0389] 155. 제154 실시양태에 있어서, 상기 스캐폴드에 대한 RN50이 입력 데이터에 대한 RN50의 10x 이상인 것인, 방법.

[0390] 156. T_0 콘티그 서열을 포함하는 콘티그 서열 세트를 수득하는 단계, 페어드 엔드 리드의 세트를 수득하는 단계

로서, 상기 페어드 엔드 리드 중 1% 이상이 1 kb 이상의 리드 쌍 거리를 포함하고, 상기 페어드 엔드 리드의 세트가 자연적인 배향의 페어드 엔드 리드를 포함하며, 상기 리드 쌍에 대한 시퀀싱 에러율이 0.1% 이하인 것인 단계, 및 T_1 ($T_1 < T_0$)을 포함하는 스캐폴드를 출력하는 단계를 포함하는, 스캐폴드 어셈블리 방법.

[0391] 157. 제156 실시양태에 있어서, T_1이 3 미만인 것인, 방법.

[0392] 158. 제156 실시양태에 있어서, T_1이 T_0의 10% 미만인 것인, 방법.

[0393] 159. 제156 실시양태에 있어서, T_1이 T_0의 1% 미만인 것인, 방법.

[0394] 160. 제156-159 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 하나의 게놈을 포함하는 것인, 방법.

[0395] 161. 제156-159 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 복수의 게놈을 포함하는 것인, 방법.

[0396] 162. 리드 쌍을 포함하는 입력 데이터를 수신하는 단계로서, 상기 리드 쌍 중 1% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하고, 입력 데이터에 대한 RN50이 어셈블링된 스캐폴드의 20% 이하이며, 상기 입력 데이터에 대한 에러율이 0.1% 이하인 것인 단계; 및 스캐폴드를 포함하는 출력 데이터를 출력하는 단계로서, 상기 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 2x 이상인 것인 단계를 포함하는, 핵산 서열 데이터 처리 방법.

[0397] 163. 제162 실시양태에 있어서, 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 10x 이상인 것인, 방법.

[0398] 164. 제162 실시양태에 있어서, 상기 스캐폴드가 90% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 방법.

[0399] 165. 제162 실시양태에 있어서, 상기 스캐폴드가 99% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 방법.

[0400] 166. 제162-165 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 하나의 게놈을 포함하는 것인, 방법.

[0401] 167. 제162-165 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 복수의 게놈을 포함하는 것인, 방법.

[0402] 168. 리드 쌍을 포함하는 전처리된 데이터세트를 출력하는 단계로서, 상기 리드 쌍 중 1% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하고, 상기 전처리된 데이터세트에 대한 RN50이 어셈블링된 스캐폴드의 20% 이하이며, 상기 출력 데이터에 대한 에러율이 0.1% 이하인 것인 단계; 및 스캐폴드를 포함하는 처리된 데이터세트를 수신하는 단계로서, 상기 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 2x 이상인 것인 단계를 포함하는, 핵산 서열 데이터 처리 방법.

[0403] 169. 제168 실시양태에 있어서, 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 10x 이상인 것인, 방법.

[0404] 170. 제168 실시양태에 있어서, 상기 스캐폴드가 90% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 방법.

[0405] 171. 제168 실시양태에 있어서, 상기 스캐폴드가 99% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 방법.

[0406] 172. 제168 실시양태에 있어서, 상기 리드 쌍 중 10% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하는 것인, 방법.

[0407] 173. 리드 쌍을 포함하는 입력 데이터를 수신하는 단계로서, 상기 리드 쌍 중 1% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하고, 입력 데이터에 대한 RN50이 어셈블링된 스캐폴드의 20% 이하이며, 상기 출력 데이터에 대한 에러율이 0.1% 이하인 것인 단계; 및 스캐폴드를 포함하는 출력 데이터를 출력하는 단계로서, 상기 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 2x 이상인 것인 단계를 포함하는, 핵산 서열 데이터 처리 방법.

[0408] 174. 제173 실시양태에 있어서, 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 10x 이상인 것인, 방법.

[0409] 175. 제173 실시양태에 있어서, 상기 스캐폴드가 90% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 방법.

[0410] 176. 제173 실시양태에 있어서, 상기 스캐폴드가 99% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 방법.

[0411] 177. 제173 실시양태에 있어서, 상기 리드 쌍 중 10% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하는 것인, 방법.

[0412] 178. 리드 쌍을 포함하는 전처리된 데이터세트를 출력하는 단계로서, 상기 리드 쌍 중 1% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하고, 상기 전처리된 데이터세트에 대한 RN50이 어셈블링된 스캐폴드의 20% 이하이며, 상기 출력 데이터에 대한 에러율이 0.1% 이하인 것인 단계; 및 스캐폴드를 포함하는 처리된 데이터세트를 수신하는 단계로서, 상기 처리된 데이터에 대한 RN50이 전처리된 데이터세트의 RN50의 2배 이상인 것인 단계를 포함하는, 핵산 서열 데이터 처리 방법.

[0413] 179. 제178 실시양태에 있어서, 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 10x 이상인 것인, 방법.

[0414] 180. 제178 실시양태에 있어서, 상기 스캐폴드가 90% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 방법.

[0415] 181. 제178 실시양태에 있어서, 상기 스캐폴드가 99% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 방법.

[0416] 182. 제178 실시양태에 있어서, 상기 리드 쌍 중 10% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하는 것인, 방법.

[0417] 183. 제178-182 실시양태 중 어느 한 실시양태에 있어서, 상기 핵산 서열 데이터가 하나의 게놈을 포함하는 것인, 방법.

[0418] 184. 제178-182 실시양태 중 어느 한 실시양태에 있어서, 상기 핵산 서열 데이터가 복수의 게놈을 포함하는 것인, 방법.

[0419] 185. 제1 콘티그에 대해 매핑된 샷건 리드의 밀도를 측정하는 단계, 제2 콘티그에 대해 매핑된 샷건 리드의 밀도를 측정하는 단계, 제1 콘티그와 제2 콘티그를 결합시키는데 있어서의 우도 스코어를 측정하는 단계, 및 제1 콘티그에 대해 매핑된 샷건 리드의 밀도가 제2 콘티그에 대해 매핑된 샷건 리드의 밀도와 크게 다른 경우에 우도 스코어를 감소시키는 단계를 포함하는, 하나 이상의 페어드 엔드 리드를 공유하는 2개의 핵산 콘티그의 결합할 확률을 평가하는 방법.

[0420] 186. 제185 실시양태에 있어서, 상기 우도 스코어가 로그 우도 스코어인 것인, 방법.

[0421] 187. 제185 실시양태에 있어서, 상기 스코어가 본원에 지시된 바와 같이 감소되는 것인, 방법.

[0422] 188. 제185 실시양태에 있어서, 상기 스코어가 제1 콘티그에 대해 매핑된 샷건 리드의 밀도 및 제2 콘티그에 대해 매핑된 샷건 리드의 밀도 중 큰 수치에 대한 작은 수치의 비율만큼 감소되는 것인, 방법.

[0423] 189. 제185 실시양태에 있어서, 상기 2개의 핵산 콘티그가 이종성 샘플로부터 유래되는 것인, 방법.

[0424] 190. 제185 실시양태에 있어서, 상기 2개의 핵산 콘티그가 메타게놈 샘플로부터 유래되는 것인, 방법.

[0425] 191. 제185 실시양태에 있어서, 상기 2개의 핵산 콘티그가 분리된 개별 유기체로부터 유래되는 것인, 방법.

[0426] 192. 제185 실시양태에 있어서, 상기 2개의 핵산 콘티그가 분리된 종들로부터 유래되는 것인, 방법.

[0427] 193. 초기 구성을 갖는 콘티그 서열의 세트를 수신하는 단계; 페어드 엔드 리드의 세트를 수신하는 단계; 표준 페어드 엔드 리드 거리 빈도 데이터를 수신하는 단계; 하나 이상의 페어드 엔드 리드 내에 공존하는 서열을 공유하는 콘티그 쌍이 그룹화되도록 콘티그 쌍을 처리하는 단계; 상기 그룹화된 콘티그 서열을, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드 쌍 거리 빈도 데이터가 초기 구성에서 콘티그 서열의 리드 쌍 빈도 데이터에 대한 표준 페어드 엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하도록 스캐폴딩하는 단계; 및 상기 처리된 콘티그 스캐폴드를 네트워크, 스크린, 또는 서버로 출력하는 단계를 수행하도록 설정된 프로세서를 포함하는, 핵산 서열 정보의 콘티그를 스캐폴딩하기 위한 컴퓨터 실행 시스템.

[0428] 194. 제193 실시양태에 있어서, 상기 스캐폴딩이 콘티그의 세트를 순서 배열하는 단계를 포함하는 것인, 컴퓨터

실행 시스템.

[0429] 195. 제193 실시양태에 있어서, 상기 스캐폴딩이 콘티그의 세트를 배향시키는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0430] 196. 제193 실시양태에 있어서, 상기 스캐폴딩이 2종 이상의 콘티그를 한줄로 병합시키는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0431] 197. 제193 실시양태에 있어서, 상기 스캐폴딩이 하나의 콘티그를 제2 콘티그에 삽입하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0432] 198. 제193 실시양태에 있어서, 상기 스캐폴딩이 하나의 콘티그를 2개 이상의 성분 콘티그로 절단하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0433] 199. 제193 실시양태에 있어서, 상기 표준 페어드 엔드 리드 빈도는, 양자 리드가 공통 콘티그에 매핑되는 페어드 엔드 리드로부터 수득되는 것인, 컴퓨터 실행 시스템.

[0434] 200. 제193 실시양태에 있어서, 상기 표준 페어드 엔드 리드 빈도가 사전에 생성된 곡선으로부터 수득되는 것인, 컴퓨터 실행 시스템.

[0435] 201. 제193 실시양태에 있어서, 상기 초기의 구성이 무작위 구성인 것인, 컴퓨터 실행 시스템.

[0436] 202. 제193 실시양태에 있어서, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드 쌍 거리 빈도 데이터가, 리드 쌍 거리 우도가 증가하는 경우에 페어드 엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하는 것인, 컴퓨터 실행 시스템.

[0437] 203. 제202 실시양태에 있어서, 상기 리드 쌍 거리 우도가 최대화되는 것인, 컴퓨터 실행 시스템.

[0438] 204. 제193 실시양태에 있어서, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드 쌍 거리 빈도 데이터가, 리드 쌍 거리 빈도 데이터와 표준 페어드 엔드 리드 거리 빈도 데이터 간의 차이에 대한 통계 측정값이 감소하는 경우에 페어드 엔드 리드 거리 빈도 데이터에 보다 더 가깝게 근접하게 되는 것인, 컴퓨터 실행 시스템.

[0439] 205. 제204 실시양태에 있어서, 리드 쌍 거리 빈도 데이터와 표준 페어드 엔드 리드 거리 빈도 데이터 간의 거리에 대한 통계 측정값이 ANOVA, t-검정 및 X-제곱 검정 중 하나 이상을 포함하는 것인, 컴퓨터 실행 시스템.

[0440] 206. 제193 실시양태에 있어서, 별도의 콘티그에 매핑되는 리드 쌍에 대한 리드 쌍 거리가, 표준 페어드 엔드 리드 거리 빈도와 비교될 때, 수득되는 순서 배열된 콘티그들 간의 리드 쌍 거리 분포의 편차가 감소하는 경우에 페어드 엔드 리드 거리 빈도 데이터에 보다 더 가깝게 매칭되는 것인, 컴퓨터 실행 시스템.

[0441] 207. 제206 실시양태에 있어서, 표준 페어드 엔드 리드 거리 빈도와 비교될 때, 수득되는 순서 배열된 콘티그들 간의 리드 쌍 거리 분포의 편차가 최소화되는 것인, 컴퓨터 실행 시스템.

[0442] 208. 제193 실시양태에 있어서, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그가 아주 많은 수의 공유된 엔드 리드를 갖는 클러스터로 지정되는 것인, 컴퓨터 실행 시스템.

[0443] 209. 실시양태 193-208 중 어느 한 실시양태에 있어서, 상기 클러스터링이 유기체 내 염색체의 수보다 많거나 염색체 수와 같은 다수의 그룹에 콘티그를 배치하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0444] 210. 실시양태 193-209 중 어느 한 실시양태에 있어서, 클러스터 중 하나의 콘티그와 단 하나의 페어드 엔드 리드만을 공유하는 콘티그는 해당 클러스터에 포함되지 않는 것인, 컴퓨터 실행 시스템.

[0445] 211. 실시양태 193-210 중 어느 한 실시양태에 있어서, 반복 서열을 포함하는 하나 이상의 페어드 엔드 리드만을 클러스터와 공유하는 콘티그는 해당 클러스터에 포함되지 않는 것인, 컴퓨터 실행 시스템.

[0446] 212. 실시양태 193-211 중 어느 한 실시양태에 있어서, 저품질의 서열을 포함하는 하나 이상의 페어드 엔드 리드만을 클러스터와 공유하는 콘티그는 해당 클러스터에 포함되지 않는 것인, 컴퓨터 실행 시스템.

[0447] 213. 실시양태 193-212 중 어느 한 실시양태에 있어서, 상기 페어드 엔드 리드의 세트가 샘플 DNA를 분해하여 핵산 내에 내부 이중 가닥 절단부를 생성하는 단계, 상기 이중 가닥 절단부를 재결찰시켜 하나 이상의 재결찰 접합부를 형성하는 단계, 및 하나 이상의 재결찰 접합부에 대해 시퀀싱하는 단계에 의해 수득되는 것인, 컴퓨터 실행 시스템.

[0448] 214. 제213 실시양태에 있어서, 상기 DNA가 하나 이상의 DNA 결합제에 가교결합되는 것인, 컴퓨터 실행 시스템.

[0449] 215. 제213 실시양태에 있어서, 상기 DNA가 단리된 네이키드 DNA인 것인, 컴퓨터 실행 시스템.

[0450] 216. 제214 실시양태에 있어서, 상기 단리된 네이키드 DNA가 재구성된 염색질 내에 리어셈블링되는 것인, 컴퓨터 실행 시스템.

[0451] 217. 제216 실시양태에 있어서, 상기 재구성된 염색질이 가교결합되는 것인, 컴퓨터 실행 시스템.

[0452] 218. 제216 실시양태에 있어서, 상기 재구성된 염색질이 DNA 결합 단백질을 포함하는 것인, 컴퓨터 실행 시스템.

[0453] 219. 제216 실시양태에 있어서, 상기 재구성된 염색질이 나노입자를 포함하는 것인, 컴퓨터 실행 시스템.

[0454] 220. 실시양태 193-219 중 어느 한 실시양태에 있어서, 상기 콘티그의 클러스터링이 유기체에 대한 염색체의 수와는 독립적인 것인, 컴퓨터 실행 시스템.

[0455] 221. 실시양태 193-220 중 어느 한 실시양태에 있어서, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그가 아주 많은 수의 공유된 엔드 리드를 갖는 클러스터로 지정되는 것인, 컴퓨터 실행 시스템.

[0456] 222. 실시양태 193-220 중 어느 한 실시양태에 있어서, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그가 더 큰 리드 쌍 거리 우도 값을 갖는 클러스터로 지정되는 것인, 컴퓨터 실행 시스템.

[0457] 223. 실시양태 193-220 중 어느 한 실시양태에 있어서, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그가 표준 리드 쌍 거리 분포에 대하여 그의 리드 쌍 분포 내에 더 낮은 편차를 갖는 클러스터로 지정되는 것인, 컴퓨터 실행 시스템.

[0458] 224. 실시양태 193-221 중 어느 한 실시양태에 있어서, 제1 클러스터 및 제2 클러스터와 관련된 페어드 엔드 리드 내의 서열을 공유하는 콘티그가 각 클러스터로부터 배제되는 것인, 컴퓨터 실행 시스템.

[0459] 225. 실시양태 193-224 중 어느 한 실시양태에 있어서, 상기 클러스터링이 유기체 내 염색체의 수보다 많거나 염색체 수와 같은 다수의 그룹에 콘티그를 배치하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0460] 226. 실시양태 193-225 중 어느 한 실시양태에 있어서, 상기 스캐폴딩이 상기 클러스터된 콘티그 중 인접한 것으로 추정되는 콘티그의 제1 세트를 선택하는 단계, 리드 쌍에 대한 상기 리드 쌍 거리의 총 측정값을 감소시키는, 인접한 것으로 추정되는 콘티그의 상기 제1 세트의 최소 거리 순서를 결정하는 단계, 및 상기 리드 쌍 거리의 상기 총 측정값을 감소시키기 위해 인접한 것으로 추정되는 콘티그의 상기 제1 세트를 스캐폴딩하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0461] 227. 제226 실시양태에 있어서, 상기 인접한 것으로 추정되는 콘티그의 상기 제1 세트가 2개의 콘티그로 이루어지는 것인, 컴퓨터 실행 시스템.

[0462] 228. 제226 실시양태에 있어서, 상기 인접한 것으로 추정되는 콘티그의 상기 제1 세트가 3개의 콘티그로 이루어지는 것인, 컴퓨터 실행 시스템.

[0463] 229. 제226 실시양태에 있어서, 상기 인접한 것으로 추정되는 콘티그의 상기 제1 세트가 4개의 콘티그로 이루어지는 것인, 컴퓨터 실행 시스템.

[0464] 230. 제226 실시양태에 있어서, 상기 인접한 것으로 추정되는 콘티그의 상기 제1 세트가 4개의 콘티그를 포함하는 것인, 컴퓨터 실행 시스템.

[0465] 231. 제226 실시양태에 있어서, 상기 스캐폴딩이 인접한 것으로 추정되는 콘티그의 상기 제1 세트 중 각 콘티그의 순서와 배향을 결정하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0466] 232. 제226 또는 제227 실시양태에 있어서, 상기 최소 거리 순서의 결정은 리드 매핑을 포함하는 하나 이상의 리드 쌍에 대해 예측된 리드 쌍 거리를 가능한 모든 콘티그 구성에 대한 상기 세트 중 2개의 콘티그를 비교하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0467] 233. 제232 실시양태에 있어서, 상기 리드 쌍에 대한 최소 리드 쌍에 상응하는 콘티그 배향을 선택하는 단계를

더 포함하는 것인, 컴퓨터 실행 시스템.

[0468] 234. 제232 실시양태에 있어서, 최대 우도 리드 쌍 거리 분포에 상응하는 콘티그 배향을 선택하는 단계를 더 포함하는 것인, 컴퓨터 실행 시스템.

[0469] 235. 제232 또는 제233 실시양태에 있어서, 상기 콘티그 클러스터의 리드 쌍의 총 측정값에 대한 최소 리드 쌍에 상응하는 콘티그 배향을 선택하는 단계를 더 포함하는 것인, 컴퓨터 실행 시스템.

[0470] 236. 실시양태 232-235 중 어느 한 실시양태에 있어서, 상기 예측된 리드 쌍 거리를 상기 페어드 엔드 리드 거리 빈도 데이터와 비교하는 것인, 컴퓨터 실행 시스템.

[0471] 237. 제236 실시양태에 있어서, 상기 페어드 엔드 리드 거리 빈도 데이터에 대한 상기 비교는 식 1의 이용을 포함하는 것인, 컴퓨터 실행 시스템.

[0472] 238. 실시양태 226-237 중 어느 한 실시양태에 있어서, 상기 클러스터된 콘티그 중 인접한 것으로 추정되는 콘티그의 제 2세트를 선택하는 단계로서, 상기 제2 세트가 상기 제1 세트 중 하나의 엔드 말단 콘티그를 제외하고 모두 포함하고, 상기 클러스터된 콘티그 중 하나의 추가 콘티그를 포함하는 것인 단계, 및 리드 쌍 거리의 상기 총 측정값을 감소시키기 위해 인접한 것으로 추정되는 콘티그의 상기 제2 세트를 스캐폴딩하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0473] 239. 제238 실시양태에 있어서, 상기 클러스터된 콘티그 중 인접한 것으로 추정되는 콘티그의 제3 세트를 선택하는 단계로서, 상기 제3 세트가 상기 제2 세트 중 하나의 엔드 말단 콘티그를 제외하고 모두 포함하고, 상기 제1 세트에 포함되지 않고 상기 제2 세트에 포함되지 않는 상기 클러스터된 콘티그 중 하나의 추가 콘티그를 포함하는 단계, 및 리드 쌍 거리의 상기 총 측정값을 감소시키기 위해 인접한 것으로 추정되는 콘티그의 상기 제3 세트를 스캐폴딩하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0474] 240. 제239 실시양태에 있어서, 대다수의 상기 클러스터된 콘티그가 순서 배열될 때까지 하나 이상의 추가의 세트를 반복적으로 선택하는 단계를 더 포함하는 것인, 컴퓨터 실행 시스템.

[0475] 241. 제240 실시양태에 있어서, 각각의 상기 클러스터된 콘티그가 순서 배열될 때까지 하나 이상의 추가의 세트를 반복적으로 선택하는 단계를 더 포함하는 것인, 컴퓨터 실행 시스템.

[0476] 242. 실시양태 193-241 중 어느 한 실시양태에 있어서, 핵산 서열이 하나의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0477] 243. 실시양태 193-241 중 어느 한 실시양태에 있어서, 핵산 서열이 복수의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0478] 244. a) 각 콘티그 쌍에 대한 로그 우도비 스코어를 지정하는 단계; b) 비율 스코어에 의해 관련성을 분류하는 단계; 및 c) 어셈블리의 전체 스코어를 증가시키기 위해 비율 스코어의 내림차순으로 콘티그 관련성을 수용 또는 거부하는 단계; 및 상기 처리된 콘티그 스캐폴드를 네트워크, 스크린 또는 서버로 출력하는 단계를 수행하도록 설정된 프로세서를 포함하는, 클러스터 내의 콘티그를 스캐폴딩하기 위한 컴퓨터 실행 시스템.

[0479] 245. 제244 실시양태에 있어서, 상기 스캐폴딩이 콘티그의 세트를 순서 배열하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0480] 246. 제244 실시양태에 있어서, 상기 스캐폴딩이 콘티그의 세트를 배향시키는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0481] 247. 제244 실시양태에 있어서, 상기 스캐폴딩이 2종 이상의 콘티그를 한줄로 병합시키는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0482] 248. 제244 실시양태에 있어서, 상기 스캐폴딩이 하나의 콘티그를 제2 콘티그에 삽입하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0483] 249. 제244 실시양태에 있어서, 상기 스캐폴딩이 하나의 콘티그를 2개 이상의 성분 콘티그로 절단하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0484] 250. 제244 실시양태에 있어서, 상기 콘티그가 하나의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0485] 251. 제244 실시양태에 있어서, 상기 콘티그가 복수의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0486] 252. a) 콘티그의 클러스터를 따라 i 위치에서 출발하는 크기 w 콘티그의 서열 윈도우를 확인하는 단계; b) 윈도우에서 각 i 위치에서 양립가능한 순서와 배향에 대한 스코어를 검사함으로써 윈도우 w의 콘티그에 대한 $w!\ 2^w$ 순서 및 배향 옵션을 고려하는 단계; c) 상기 윈도우 내 상기 w 콘티그를 배향하고 순서 배열하여 최적의 스코어를 수득하는 단계; d) 상기 윈도우를 i+1 위치로 이동시키는 단계; 및 e) 상기 w 콘티그의 배향 및 순서 배열을 이용하여 i+1 위치에서 상기 윈도우에 대하여 단계 (a), (b) 및 (c)를 반복하여 최적의 스코어를 결정함으로써, 상기 복수의 콘티그를 스코어에 대해 국소적으로 최적 구성으로 배향하고 순서 배열하는 단계; 및 상기 처리된 콘티그 스캐폴드를 네트워크, 스크린, 또는 서버로 출력하는 단계에 의해 콘티그의 세트를 수신하고, 상기 콘티그의 세트를 처리하도록 설정된 프로세서를 포함하는, 클러스터 내의 복수의 콘티그에 대해 국소적으로 최적의 콘티그 구성을 결정하기 위한 컴퓨터 실행 시스템.

[0487] 253. 제252 실시양태에 있어서, 클러스터 내의 복수의 콘티그에 매핑된 리드 쌍 데이터가 수득되고, 표준 페어드 엔드 리드 빈도 데이터 세트가 수득되며, 상기 w 콘티그의 배향 및 순서 배열에 대한 스코어가 클러스터 내의 복수의 콘티그에 매핑된 리드 쌍 데이터에 대한 리드 쌍 거리 데이터 세트가 표준 페어드 엔드 리드 빈도 데이터 세트에 얼마나 가깝게 매칭되는지에 대한 척도인 것인, 컴퓨터 실행 시스템.

[0488] 254. 제252 실시양태에 있어서, 클러스터 내의 복수의 콘티그에 매핑된 리드 쌍 데이터가 수득되고, 상기 스코어가 전체 리드 쌍 거리이며, 상기 스코어가 전체 리드 쌍 거리가 최소화될 때 최적화되는 것인, 컴퓨터 실행 시스템.

[0489] 255. 제252 실시양태에 있어서, w가 3인 것인, 컴퓨터 실행 시스템.

[0490] 256. 제252 실시양태에 있어서, w가 4인 것인, 컴퓨터 실행 시스템.

[0491] 257. 제252 실시양태에 있어서, w가 5인 것인, 컴퓨터 실행 시스템.

[0492] 258. 제252 실시양태에 있어서, w가 6인 것인, 컴퓨터 실행 시스템.

[0493] 259. 제252 실시양태에 있어서, w가 제1 클러스터에 대한 제1값을 가지며, w가 제2 클러스터에서 제2값을 갖는 것인, 컴퓨터 실행 시스템.

[0494] 260. 제252 실시양태에 있어서, w가 상기 세트 중 1%의 콘티그를 포함하도록 선택되는 것인, 컴퓨터 실행 시스템.

[0495] 261. 제252 실시양태에 있어서, w가 상기 세트 중 5%의 콘티그를 포함하도록 선택되는 것인, 컴퓨터 실행 시스템.

[0496] 262. 제252 실시양태에 있어서, w가 상기 세트 중 10%의 콘티그를 포함하도록 선택되는 것인, 컴퓨터 실행 시스템.

[0497] 263. 제252 실시양태에 있어서, 상기 스코어가 리드 쌍 거리 우도 스코어이고, 상기 스코어는 해당 스코어가 주어진 윈도우 크기에 대해 최대화될 때 최적인, 컴퓨터 실행 시스템.

[0498] 264. 제263 실시양태에 있어서, 상기 스코어가 식 1을 이용하여 계산되는 것인, 컴퓨터 실행 시스템.

[0499] 265. 제252 실시양태에 있어서, 상기 스코어가 예측된 리드 쌍 분포로부터의 편차이고, 해당 스코어가 주어진 윈도우 크기에 대해 최소화될 때 스코어가 최적인, 컴퓨터 실행 시스템.

[0500] 266. 제252-265 실시양태 중 한 실시양태에 있어서, 상기 복수의 콘티그가 하나의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0501] 267. 제252-265 실시양태 중 한 실시양태에 있어서, 상기 복수의 콘티그가 복수의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0502] 268. a) 정제된 DNA를 수득하는 단계; b) 상기 정제된 DNA를 DNA 결합제와 결합시켜 DNA/염색질 복합체를 형성하는 단계; c) 상기 DNA-염색질 복합체를 제한 효소와 함께 항온배양시켜 점착성 말단을 남기는 단계; d) 결찰을 수행하여 DNA의 말단을 결합시키는 단계; e) 결찰된 DNA 접합부에 대해 시퀀싱하여 페어드 엔드 리드를 생성하는 단계; 및 f) 페어드 엔드 리드를 수신하여 처리하도록 설정된 프로세서를 포함하는 컴퓨터 실행 시스템을 사용하여, 상기 정제된 DNA의 서열을 나타내는 콘티그를 포함하는 스캐폴드 핵산 데이터 세트를 출력하는 단계를 포함하는, 핵산 서열 어셈블리 방법.

[0503] 269. 제268 실시양태에 있어서, 상기 정제된 DNA가 하나의 게놈으로부터 유래되는 것인, 방법.

[0504] 270. 제268 실시양태에 있어서, 상기 정제된 DNA가 복수의 게놈으로부터 유래되는 것인, 방법.

[0505] 271. 핵산 샘플에 대한 콘티그 데이터세트를 수신하는 단계; 물리적으로 연결된 비인접성 서열 정보에 상응하는 하나 이상의 리드 쌍 서열 리드를 수신하는 단계; 리드 쌍 서열 리드 중 하나 이상의 리드가 하나의 콘티그 데이터 세트 중 2개의 별개 위치에 매핑되는 경우, 리드 쌍 서열 리드를 배제시키는 단계를 수행하도록 설정된 프로세서를 포함하는, 반복성 콘티그 영역에 대해 매핑되는 리드 쌍 서열 리드를 확인하기 위한 컴퓨터 실행 시스템.

[0506] 272. 제271 실시양태에 있어서, 상기 반복성 영역이 제1 역치를 초과하는 샷건 리드 깊이를 갖는 서열을 포함하는 것인, 컴퓨터 실행 시스템.

[0507] 273. 제272 실시양태에 있어서, 상기 반복성 영역이 제2 역치를 초과하는 리드 깊이를 갖는 염기 위치를 포함하는 것인, 컴퓨터 실행 시스템.

[0508] 274. 제273 실시양태에 있어서, 상기 제1 역치 및 제2 역치가 리드 깊이의 전체적인 분포에 대하여 고정되는 것인, 컴퓨터 실행 시스템.

[0509] 275. 제274 실시양태에 있어서, 상기 제1 역치가 리드 깊이의 전체적인 분포의 3배인 것인, 컴퓨터 실행 시스템.

[0510] 276. 제274 실시양태에 있어서, 상기 제2 역치가 리드 깊이의 전체적인 분포의 3.5배인 것인, 컴퓨터 실행 시스템.

[0511] 277. 제271-276 실시양태 중 어느 한 실시양태에 있어서, 상기 핵산 샘플이 하나의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0512] 278. 제271-276 실시양태 중 어느 한 실시양태에 있어서, 상기 핵산 샘플이 복수의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0513] 279. 콘티그 데이터 세트를 수신하는 단계, 상기 데이터 세트를 처리하여 제1 콘티그와 제2 콘티그 간에 걸쳐있는 리드 쌍 서열의 개수 및 정황상 추정되는 분리를 관찰할 확률을 측정하는 단계로서, 상기 콘티그가 세트 [++,+-,-+,--] 내에서 o의 상대적 배향을 가지며 갭 깊이에 의해 분리되는 것인 단계, 및 상기 데이터 세트 및 측정된 확률을 네트워크, 스크린 또는 서버로 출력하도록 설정된 프로세서를 포함하는 콘티그 어셈블리 결정을 안내하기 위한 컴퓨터 실행 시스템.

[0514] 280. 제279 실시양태에 있어서, 분리 거리에 대해 리드 쌍 서열 분포의 확률을 정규화하는 단계로서, 상기 정규화가 상기 리드 쌍 서열을, 핵산 샘플을 독립적으로 표본으로 삼은 노이즈 쌍과 비교하는 단계를 포함하는 것인 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0515] 281. 제280 실시양태에 있어서, 상기 핵산 샘플이 하나의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0516] 282. 제280 실시양태에 있어서, 상기 핵산 샘플이 복수의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0517] 283. 제280 실시양태에 있어서, 상기 노이즈 쌍의 총수가 콘티그 쌍의 샘플에 대한 링크의 밀도를 도표화함으로써 측정되는 것인, 컴퓨터 실행 시스템.

[0518] 284. 제283 실시양태에 있어서, 최고 밀도값 1%와 최저 밀도값 1%가 배제되는 것인, 컴퓨터 실행 시스템.

[0519] 285. 제279 실시양태에 있어서, 콘티그 순서를 결정하는 단계를 더 포함하는 것인, 컴퓨터 실행 시스템.

[0520] 286. 제279 실시양태에 있어서, 콘티그 배향을 결정하는 단계를 더 포함하는 것인, 컴퓨터 실행 시스템.

[0521] 287. 초기 구성을 갖는 콘티그 서열의 세트를 수신하는 단계; 페어드 엔드 리드의 세트를 수신하는 단계; 표준 페어드 엔드 리드 거리 빈도 데이터를 수신하는 단계; 콘티그 쌍을 하나 이상의 페어드 엔드 리드 내에 공존하는 서열을 공유하는 그룹으로 하는 처리하는 단계; 상기 콘티그의 그룹에 대한 리드 쌍 빈도 데이터를 표준 페어드 엔드 리드 거리 빈도 데이터와 비교함으로써 처리하고, 상기 그룹 중 하나의 콘티그에 절단부를 도입하는 것이 상기 콘티그의 그룹에 대한 리드 쌍 빈도 데이터를 표준 페어드 엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하도록 하는지의 여부를 측정하는 단계; 및 그러한 경우에, 상기 콘티그에 절단부를 도입하는 단계; 및 상기 처리된 콘티그 데이터를 네트워크, 스크린 또는 서버에 출력하는 단계를 수행하도록 설정된 프로세서를 포

함하는, 콘티그 부정합 보정을 위한 컴퓨터 실행 시스템.

[0522] 288. 제287 실시양태에 있어서, 제1 위치가 절단부를 도입하기 전의 상기 역치 미만의 상기 로그 우도를 갖는 하나 이상의 인접한 제2 위치와 병합되는 것인, 컴퓨터 실행 시스템.

[0523] 289. 제287 실시양태에 있어서, 인접한 제2 위치가 제1 위치로부터 300개 이하의 염기 쌍만큼 떨어져 있는 것인, 컴퓨터 실행 시스템.

[0524] 290. 제287 실시양태에 있어서, 제2 위치가 제1 위치로부터 1000개 초과의 염기 쌍만큼 떨어진 위치를 포함하지 않는 것인, 컴퓨터 실행 시스템.

[0525] 291. 제287-290 실시양태 중 어느 한 실시양태에 있어서, 상기 로그 우도 변화를 측정하는 단계가 하나의 콘티그에 대한 평균 페어드 엔드 매핑 밀도를 확인하는 단계, 평균 페어드 엔드 매핑 밀도의 3x 이상의 페어드 엔드 매핑 밀도를 갖는 콘티그의 세그먼트를 확인하는 단계, 및 평균 페어드 엔드 매핑 밀도의 3x 이상의 페어드 엔드 매핑 밀도를 갖는 콘티그의 세그먼트를 배제시키는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0526] 292. 제287-291 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 하나의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0527] 293. 제287-291 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 복수의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0528] 294. 출발 어셈블리의 절단된 콘티그를 표시하는 단계로서, 상기 절단된 콘티그가 노드이고, 상기 절단된 콘티그의 엣지가 정수의 순서쌍의 목록으로 표지되며, 상기 절단부의 엣지가 매핑된 리드 쌍 서열에 상응하는 것인 단계; 및 역치 수 미만의 매핑된 연결부를 갖는 엣지를 배제시키는 단계에 의해 콘티그의 세트를 수용하여 상기 콘티그의 세트를 처리하도록 설정된 프로세서를 포함하는, 콘티그 어셈블리를 위한 컴퓨터 실행 시스템.

[0529] 295. 제294 실시양태에 있어서, 상기 역치 수가 5% 미만인 것인, 컴퓨터 실행 시스템.

[0530] 296. 제294 실시양태에 있어서, 상기 역치 수가 t_L 링크보다 작은 것인, 컴퓨터 실행 시스템.

[0531] 297. 제294 실시양태에 있어서, 상기 콘티그가, 해당 콘티그 길이에 대하여 상응하는 노드의 그래프 내 차수의 비율이 모든 분포값들 중 약 5%의 최고치를 초과하는 염기 쌍인 엣지를 포함하는 것인, 컴퓨터 실행 시스템.

[0532] 298. 제294-297 실시양태에 있어서, 상기 콘티그가 하나의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0533] 299. 제294-297 실시양태에 있어서, 상기 콘티그가 복수의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0534] 300. 복수의 콘티그에 상응하는 서열 정보를 수신하는 단계, 상기 복수의 콘티그에 의해 나타난 핵산 샘플의 페어드 엔드 리드 정보를 수신하는 단계, 예측된 리드 쌍 거리 데이터 세트의 리드 쌍 거리 파라미터의 편차가 최소화되도록 상기 복수의 콘티그를 배치함으로써 이를 처리하는 단계로서, 상기 배치가 8시간 미만으로 일어나는 것인 단계, 및 상기 최소화된 편차를 갖는 상기 배치된 콘티그를 네트워크, 스크린 또는 서버로 출력하는 단계를 수행하도록 설정된 프로세서를 포함하는, 콘티그 서열 정보를 하나 이상의 스캐폴드에 어셈블링하기 위한 컴퓨터 실행 시스템.

[0535] 301. 제300 실시양태에 있어서, 상기 예측된 리드 쌍 거리 데이터 세트가 리드 쌍 거리 우도 곡선을 포함하는 것인, 컴퓨터 실행 시스템.

[0536] 302. 제300 또는 제301 실시양태에 있어서, 상기 리드 쌍 거리 파라미터가 리드 쌍 거리 우도 곡선에 대한 최대 거리 우도인 것인, 컴퓨터 실행 시스템.

[0537] 303. 제300 또는 제301 실시양태에 있어서, 상기 리드 쌍 거리 파라미터가 리드 쌍 거리 우도 곡선에 대한 최소 변이인 것인, 컴퓨터 실행 시스템.

[0538] 304. 제300-303 실시양태 중 어느 한 실시양태에 있어서, 상기 국소적으로 인접한 콘티그의 세트가 2개의 콘티그를 포함하는 것인, 컴퓨터 실행 시스템.

[0539] 305. 제300-303 실시양태 중 어느 한 실시양태에 있어서, 상기 국소적으로 인접한 콘티그의 세트가 3개의 콘티그를 포함하는 것인, 컴퓨터 실행 시스템.

[0540] 306. 제300-303 실시양태 중 어느 한 실시양태에 있어서, 상기 국소적으로 인접한 콘티그의 세트가 4개의 콘티

그를 포함하는 것인, 컴퓨터 실행 시스템.

[0541] 307. 제300-303 실시양태 중 어느 한 실시양태에 있어서, 상기 국소적으로 인접한 콘티그의 세트가 5개의 콘티그를 포함하는 것인, 컴퓨터 실행 시스템.

[0542] 308. 제300-303 실시양태 중 어느 한 실시양태에 있어서, 상기 국소적으로 인접한 콘티그의 세트가 6개의 콘티그를 포함하는 것인, 컴퓨터 실행 시스템.

[0543] 309. 제300-308 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 7시간 미만으로 일어나는 것인, 컴퓨터 실행 시스템.

[0544] 310. 제300-308 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 6시간 미만으로 일어나는 것인, 컴퓨터 실행 시스템.

[0545] 311. 제300-308 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 5시간 미만으로 일어나는 것인, 컴퓨터 실행 시스템.

[0546] 312. 제300-308 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 4시간 미만으로 일어나는 것인, 컴퓨터 실행 시스템.

[0547] 313. 제300-308 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 3시간 미만으로 일어나는 것인, 컴퓨터 실행 시스템.

[0548] 314. 제300-308 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 2시간 미만으로 일어나는 것인, 컴퓨터 실행 시스템.

[0549] 315. 제300-308 실시양태 중 어느 한 실시양태에 있어서, 상기 배치가 1시간 미만으로 일어나는 것인, 컴퓨터 실행 시스템.

[0550] 316. 제300-315 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 정보가 하나의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0551] 317. 제300-315 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 정보가 복수의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0552] 318. 핵산 샘플을 대표하는 콘티그 서열 세트를 수신하는 단계, 상기 핵산 샘플에 대한 리드 쌍 데이터를 수신하는 단계, 및 핵산 샘플에 대한 리드 쌍 데이터가 예측된 리드 쌍 분포에 보다 가깝게 근접하도록 콘티그 서열 세트를 순서 배열하여 배향시킴으로써 상기 수신된 데이터를 처리하는 단계로서, 상기 콘티그 서열 세트 중 70%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 8시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인 단계, 및 상기 순서 배열 및 배향된 데이터를 네트워크, 스크린 또는 서버에 출력시키는 단계를 수행하도록 설정된 프로세서를 포함하는, 콘티그 서열 세트를 스캐폴딩하기 위한 컴퓨터 실행 시스템.

[0553] 319. 제318 실시양태에 있어서, 상기 스캐폴딩이 콘티그의 세트를 순서 배열하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0554] 320. 제318 실시양태에 있어서, 상기 스캐폴딩이 콘티그의 세트를 배향시키는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0555] 321. 제318 실시양태에 있어서, 상기 스캐폴딩이 2종 이상의 콘티그를 한줄로 병합시키는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0556] 322. 제318 실시양태에 있어서, 상기 스캐폴딩이 하나의 콘티그를 제2 콘티그에 삽입하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0557] 323. 제318 실시양태에 있어서, 상기 스캐폴딩이 하나의 콘티그를 2개 이상의 성분 콘티그로 절단하는 단계를 포함하는 것인, 컴퓨터 실행 시스템.

[0558] 324. 제318 실시양태에 있어서, 콘티그 서열 세트 중 80%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 8시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인, 컴퓨터 실행 시스템.

[0559] 325. 제318 실시양태에 있어서, 콘티그 서열 세트 중 90%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 8시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인, 컴퓨터 실행 시스템.

[0560] 326. 제318 실시양태에 있어서, 콘티그 서열 세트 중 95%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 8시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인, 컴퓨터 실행 시스템.

[0561] 327. 제318 실시양태에 있어서, 콘티그 서열 세트 중 70%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 4시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인, 컴퓨터 실행 시스템.

[0562] 328. 제318 실시양태에 있어서, 콘티그 서열 세트 중 70%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 2시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인, 컴퓨터 실행 시스템.

[0563] 329. 제318 실시양태에 있어서, 콘티그 서열 세트 중 70%가 핵산 샘플 내에 이들의 서열의 상대적 순서와 배향을 1시간 이하의 시간 내에 매칭시키기 위해 순서 배열되고 배향되는 것인, 컴퓨터 실행 시스템.

[0564] 330. 제318-329 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 하나의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0565] 331. 제318-329 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 복수의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0566] 332. 복수의 콘티그에 상응하는 서열 정보를 수신하는 단계, 페어드 엔드 리드 정보를 수신하는 단계, 및 상기 페어드 엔드 리드 정보에 대한 페어드 엔드 리드 거리 분포가 전반적으로 최적화되어 레퍼런스 페어드 엔드 리드 거리 분포에 근접하도록 상기 복수의 콘티그를 배치함으로써 상기 수신된 데이터를 처리하는 단계로서, 상기 배치가 8시간 미만으로 일어나는 것인 단계, 및 상기 배치된 복수의 콘티그를 네트워크, 스크린 또는 서버에 출력하는 단계를 수행하도록 설정된 프로세서를 포함하는, 핵산 서열 데이터 세트를 구성하기 위한 컴퓨터 실행 시스템.

[0567] 333. 제332 실시양태에 있어서, 상기 핵산 서열 데이터의 세트가 하나의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0568] 334. 제332 실시양태에 있어서, 상기 핵산 서열 데이터의 세트가 복수의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0569] 335. 제332 실시양태에 있어서, 상기 배치가 4시간 미만으로 일어나는 것인, 컴퓨터 실행 시스템.

[0570] 336. 제332 실시양태에 있어서, 상기 배치가 2시간 미만으로 일어나는 것인, 컴퓨터 실행 시스템.

[0571] 337. 복수의 결합된 노드쌍을 포함하는 스캐폴드 세트를 수신하는 단계로서, 상기 노드 쌍의 각 노드가 하나 이상의 콘티그 서열을 포함하는 것인 단계, 상기 복수의 결합된 노드에 매핑된 페어드 엔드 리드 정보를 수신하는 단계, 결합된 노드쌍에 의해 공유된 리드 쌍의 개수를 계수함으로써 상기 수신된 데이터를 처리하는 단계, 상기 개수를 역치와 비교하는 단계, 및 상기 개수가 역치 미만인 경우 노드 쌍을 결합되지 않은 노드로 절단하는 단계, 및 상기 처리된 데이터를 네트워크, 스크린 또는 서버에 출력하는 단계를 수행하도록 설정된 프로세서를 포함하는, 스캐폴드 어셈블리를 개선하기 위한 컴퓨터 실행 시스템.

[0572] 338. 제337 실시양태에 있어서, 고유의 콘티그 서열에 매핑된 리드 쌍만이 계수되는 것인, 컴퓨터 실행 시스템.

[0573] 339. 제337 실시양태에 있어서, 역치수의 별개의 리드 페어드 엔드가 매핑되는 콘티그 서열 세그먼트에 매핑된 리드 쌍이 폐기되는 것인, 컴퓨터 실행 시스템.

[0574] 340. 제337 실시양태에 있어서, 상기 역치 수가 비반복성 서열에 대한 평균 개수의 3x인 것인, 컴퓨터 실행 시스템.

[0575] 341. 제337-340 실시양태 중 어느 한 실시양태에 있어서, 상기 스캐폴드 세트가 하나의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0576] 342. 제337-340 실시양태 중 어느 한 실시양태에 있어서, 상기 스캐폴드 세트가 복수의 게놈으로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0577] 343. 복수의 결합된 노드쌍을 포함하는 스캐폴드 세트를 수신하는 단계로서, 상기 노드 쌍의 각 노드가 하나 이상의 콘티그 서열을 포함하는 것인 단계, 상기 복수의 결합된 노드에 매핑된 페어드 엔드 리드 정보를 수신하는 단계, 표준 페어드 엔드 리드 거리 빈도 데이터를 수신하는 단계; 상기 복수의 결합된 노드에 매핑된 페어드 엔드 리드 정보에 대한 페어드 엔드 리드 빈도 데이터를 표준 페어드 엔드 리드 거리 빈도 데이터와 비교함으로써

상기 수신된 데이터를 처리하는 단계; 결합된 노드를 절단하는 것이 상기 복수의 결합된 노드에 매핑된 페어드 엔드 리드 정보에 대한 페어드 엔드 리드 빈도 데이터를 표준 페어드 엔드 리드 거리 빈도 데이터에 보다 가깝게 근접하도록 유도하는 경우에 하나 이상의 결합된 노드를 절단하는 단계; 및 상기 처리된 데이터를 네트워크, 스크린 또는 서버에 출력하는 단계를 수행하도록 설정된 프로세서를 포함하는, 스캐폴드 어셈블리를 개선하기 위한 컴퓨터 실행 시스템.

[0578] 344. 제343 실시양태에 있어서, 상기 스캐폴드 세트가 하나의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0579] 345. 제343 실시양태에 있어서, 상기 스캐폴드 세트가 복수의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0580] 346. 콘티그 서열 세트를 수신하는 단계, 페어드 엔드 리드 세트를 포함하는 입력 데이터를 수신하는 단계로서, 상기 페어드 엔드 리드 중 1% 이상이 1 kb 이상의 리드 쌍 거리를 포함하고, 상기 페어드 엔드 리드의 세트가 자연적인 배향의 페어드 엔드 리드를 포함하며, 상기 리드 쌍에 대한 시퀀싱 에러율이 0.1% 이하이고, 상기 입력 데이터의 RN50이 어셈블링된 스캐폴드의 20% 이하인 것인 단계, 및 스캐폴드를 출력하는 단계로서, 스캐폴드에 대한 RN50이 입력 데이터의 RN50의 2x 이상인 것인 단계를 수행하도록 설정된 프로세서를 포함하는, 스캐폴드 어셈블리를 위한 컴퓨터 실행 시스템.

[0581] 347. 제346 실시양태에 있어서, 상기 스캐폴드에 대한 RN50이 입력 데이터의 RN50의 10x 이상인 것인, 컴퓨터 실행 시스템.

[0582] 348. T_0 콘티그 서열을 포함하는 콘티그 서열 세트를 수신하는 단계, 페어드 엔드 리드의 세트를 수신하는 단계로서, 상기 페어드 엔드 리드 중 1% 이상이 1 kb 이상의 리드 쌍 거리를 포함하고, 상기 페어드 엔드 리드의 세트가 자연적인 배향의 페어드 엔드 리드를 포함하며, 상기 리드 쌍에 대한 시퀀싱 에러율이 0.1% 이하인 것인 단계, 및 T_1 ($T_1 < T_0$)을 포함하는 스캐폴드를 출력하는 단계를 수행하도록 설정된 프로세서를 포함하는, 스캐폴드 어셈블리를 위한 컴퓨터 실행 시스템.

[0583] 349. 제348 실시양태에 있어서, T_1이 3 미만인 것인, 컴퓨터 실행 시스템.

[0584] 350. 제348 실시양태에 있어서, T_1이 T_0의 10% 미만인 것인, 컴퓨터 실행 시스템.

[0585] 351. 제348 실시양태에 있어서, T_1이 T_0의 1% 미만인 것인, 컴퓨터 실행 시스템.

[0586] 352. 제348-351 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 하나의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0587] 353. 제348-351 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 복수의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0588] 354. 리드 쌍을 포함하는 입력 데이터를 수신하는 단계로서, 상기 리드 쌍 중 1% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하고, 입력 데이터에 대한 RN50이 어셈블링된 스캐폴드의 20% 이하이며, 상기 입력 데이터에 대한 에러율이 0.1% 이하인 것인 단계; 및 스캐폴드를 포함하는 출력 데이터를 출력하는 단계로서, 상기 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 2x 이상인 것인 단계를 수행하도록 설정된 프로세서를 포함하는, 핵산 서열 데이터 처리를 위한 컴퓨터 실행 시스템.

[0589] 355. 제354 실시양태에 있어서, 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 10x 이상인 것인, 컴퓨터 실행 시스템.

[0590] 356. 제354 실시양태에 있어서, 상기 스캐폴드가 90% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 컴퓨터 실행 시스템.

[0591] 357. 제354 실시양태에 있어서, 상기 스캐폴드가 99% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 컴퓨터 실행 시스템.

[0592] 358. 제354-357 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 하나의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0593] 359. 제354-357 실시양태 중 어느 한 실시양태에 있어서, 상기 콘티그 서열의 세트가 복수의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0594] 360. 리드 쌍을 포함하는 전처리된 데이터세트를 출력하는 단계로서, 상기 리드 쌍 중 1% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하고, 상기 전처리된 데이터세트에 대한 RN50이 어셈블링된 스캐폴드의 20% 이하이며, 상기 출력 데이터에 대한 에러율이 0.1% 이하인 것인 단계; 및 스캐폴드를 포함하는 처리된 데이터세트를 수신하는 단계로서, 상기 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 2x 이상인 것인 단계를 수행하도록 설정된 프로세서를 포함하는, 핵산 서열 데이터 처리를 위한 컴퓨터 실행 시스템.

[0595] 361. 제360 실시양태에 있어서, 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 10x 이상인 것인, 컴퓨터 실행 시스템.

[0596] 362. 제360 실시양태에 있어서, 상기 스캐폴드가 90% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 컴퓨터 실행 시스템.

[0597] 363. 제360 실시양태에 있어서, 상기 스캐폴드가 99% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 컴퓨터 실행 시스템.

[0598] 364. 제360 실시양태에 있어서, 상기 리드 쌍 중 10% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하는 것인, 컴퓨터 실행 시스템.

[0599] 365. 리드 쌍을 포함하는 입력 데이터를 수신하는 단계로서, 상기 리드 쌍 중 1% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하고, 입력 데이터에 대한 RN50이 어셈블링된 스캐폴드의 20% 이하이며, 상기 입력 데이터에 대한 에러율이 0.1% 이하인 것인 단계; 및 스캐폴드를 포함하는 출력 데이터를 출력하는 단계로서, 상기 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 2x 이상인 것인 단계를 수행하도록 설정된 프로세서를 포함하는, 핵산 서열 데이터 처리를 위한 컴퓨터 실행 시스템.

[0600] 366. 제365 실시양태에 있어서, 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 10x 이상인 것인, 컴퓨터 실행 시스템.

[0601] 367. 제365 실시양태에 있어서, 상기 스캐폴드가 90% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 컴퓨터 실행 시스템.

[0602] 368. 제365 실시양태에 있어서, 상기 스캐폴드가 99% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 컴퓨터 실행 시스템.

[0603] 369. 제365 실시양태에 있어서, 상기 리드 쌍 중 10% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하는 것인, 컴퓨터 실행 시스템.

[0604] 370. 리드 쌍을 포함하는 전처리된 데이터세트를 출력하는 단계로서, 상기 리드 쌍 중 1% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하고, 상기 전처리된 데이터세트에 대한 RN50이 어셈블링된 스캐폴드의 20% 이하이며, 상기 출력 데이터에 대한 에러율이 0.1% 이하인 것인 단계; 및 스캐폴드를 포함하는 처리된 데이터세트를 수신하는 단계로서, 상기 처리된 데이터에 대한 RN50이 전처리된 데이터세트의 RN50의 2배 이상인 것인 단계를 수행하도록 설정된 프로세서를 포함하는, 핵산 서열 데이터 처리를 위한 컴퓨터 실행 시스템.

[0605] 371. 제370 실시양태에 있어서, 출력 데이터에 대한 RN50이 입력 데이터의 RN50의 10x 이상인 것인, 컴퓨터 실행 시스템.

[0606] 372. 제370 실시양태에 있어서, 상기 스캐폴드가 90% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 컴퓨터 실행 시스템.

[0607] 373. 제370 실시양태에 있어서, 상기 스캐폴드가 99% 이상의 올바른 순서로 배향된 표적 게놈 샘플 서열을 포함하는 것인, 컴퓨터 실행 시스템.

[0608] 374. 제370 실시양태에 있어서, 상기 리드 쌍 중 10% 이상이 1 kb 이상 분리되고 자연적으로 배향된 2개의 핵산 세그먼트의 서열 데이터를 포함하는 것인, 컴퓨터 실행 시스템.

[0609] 375. 제370-374 실시양태 중 어느 한 실시양태에 있어서, 상기 핵산 서열 데이터가 하나의 게놈을 포함하는 것인, 컴퓨터 실행 시스템.

[0610] 376. 제370-374 실시양태 중 어느 한 실시양태에 있어서, 상기 핵산 서열 데이터가 복수의 게놈을 포함하는 것

인, 컴퓨터 실행 시스템.

[0611] 377. 콘티그 세트를 수신하는 단계; 제1 콘티그에 대해 매핑된 샷건 리드의 밀도를 측정하는 단계, 제2 콘티그에 대해 매핑된 샷건 리드의 밀도를 측정하는 단계, 제1 콘티그와 제2 콘티그를 결합시키는데 있어서의 우도 스코어를 측정하는 단계, 및 제1 콘티그에 대해 매핑된 샷건 리드의 밀도가 제2 콘티그에 대해 매핑된 샷건 리드의 밀도와 크게 다른 경우에 우도 스코어를 감소시키는 단계에 의해 상기 콘티그 세트를 처리하는 단계; 및 상기 처리된 콘티그 세트를 네트워크, 스크린 또는 서버에 출력시키는 단계를 수행하도록 설정된 프로세서를 포함하는, 하나 이상의 페어드 엔드 리드를 공유하는 2개의 핵산 콘티그의 결합할 확률을 평가하기 위한 컴퓨터 실행 시스템.

[0612] 378. 제377 실시양태에 있어서, 상기 우도 스코어가 로그 우도 스코어인 것인, 컴퓨터 실행 시스템.

[0613] 379. 제377 실시양태에 있어서, 상기 스코어가 본원에 지시된 바와 같이 감소되는 것인, 컴퓨터 실행 시스템.

[0614] 380. 제377 실시양태에 있어서, 상기 스코어가 제1 콘티그에 대해 매핑된 샷건 리드의 밀도 및 제2 콘티그에 대해 매핑된 샷건 리드의 밀도 중 큰 수치에 대한 작은 수치의 비율만큼 감소되는 것인, 컴퓨터 실행 시스템.

[0615] 381. 제377 실시양태에 있어서, 상기 2개의 핵산 콘티그가 이종성 샘플로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0616] 382. 제377 실시양태에 있어서, 상기 2개의 핵산 콘티그가 메타게놈 샘플로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0617] 383. 제377 실시양태에 있어서, 상기 2개의 핵산 콘티그가 분리된 개별 유기체로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0618] 384. 제377 실시양태에 있어서, 상기 2개의 핵산 콘티그가 분리된 종들로부터 유래되는 것인, 컴퓨터 실행 시스템.

[0619] **실시예**

[0620] **실시예 1: 재구성된 염색질로부터 생성된 페어드 리드를 사용한 게놈 피니싱 (genome finishing)**

[0621] 5.5 μg의 고분자량 DNA를 인간 세포주 GM12878 및 야생에서 포획된 미국 악어의 혈액으로부터 추출하였다. 상기 고분자량 DNA를 약 150 Kbp의 단편으로 추출하였다. 상기 DNA를 정제된 히스톤 및 염색질 어셈블리 요소와 혼합함으로써 염색질을 재구성하였다. 이후, 상기 재구성된 염색질을 포름알데히드로 고정시키고 서열 데이터 라이브러리를 생성하였다. 도 1A-F는 이러한 단계들을 도해로 나타내고 있다.

[0622] 인간 GM12878 샘플에 대해서는, 4 bp 5' 오버행을 생성하는 제한 효소인 MboI과 MluCI를 사용하여 2개의 DNA 라이브러리를 생성하였다. 이들 바코드 라이브러리를 모아 단일 일루미나 HiSeq 2500 레인 상에서 페어드된 100 bp 리드로 시퀀싱하여, 46M MboI 및 52M MluCI 라이브러리 리드 쌍을 생성하였다. 비교를 위해, 도 2에 도시된 바와 같이, 명목상 40 Kbp인 DNA를 위하여 제3 라이브러리를 제조하였다.

[0623] 미국 악어 게놈 (미시시피 악어)에 대해서, 우리는 단일 MboI 라이브러리를 구축하고, 이를 단일 레인 상에서 시퀀싱하여, 132M 리드 쌍을 유도하였다. 이들 데이터가 스캐폴드 어셈블리 및 하플로타입 페이징에 유용한지를 알아보기 위해, 우리는 GM12878 라이브러리 데이터를 참조 인간 어셈블리 (hg19)에 정렬하였다 (도 2). 상기 생성된 라이브러리는 관련이 없는 500 Kbp 게놈 윈도우 간에 약 1 (이러한 링크 0.97의 평균)의 의사 링크의 백그라운드 노이즈 비율로 게놈 상에 150 Kbp 이하의 분리에 대한 것까지 유용한 연관 정보를 제공하였다. 도 3에 도시되어 있는 바와 같이, GM12878 라이브러리 서열의 단일 레인은 0-1, 1-5, 5-10, 10-25, 25-50 및 50-200 Kbp 라이브러리에서 각각 3.8X, 8.4X, 8.6X, 18.6X, 13.5X, 6.5X 물리적 커버리지와 동일한 연관 정보를 제공하였던 반면, 악어에 있어서는 비교할 만한 커버리지 추정치는 각각 5.4X, 16.7X, 16.7X, 42.2X, 36.1X 및 16.5X 이었다.

[0624] **실시예 2: 리드 쌍 데이터에 기반한 핵산 스캐폴딩**

[0625] 라이브러리로부터 추출된 데이터의 역량과 유용성을 알아보기 위해, 일반적인 명칭으로 판매되는 300-500 bp 인서트 일루미나 샷건 라이브러리와 상기 기술된 라이브러리만을 사용하여 콘티그 어셈블리 및 스캐폴딩을 수행하였다. GM12878의 84x, 101 bp 페어드 엔드 일루미나 샷건 데이터 세트 (Chapman et al., 2011)는 MERACULOUS [pmid21876754]를 사용하여 33 Kbp의 일반적인 (N50) 크기를 갖는 스캐폴드로 초기에 어셈블링하였다. 생성된

라이브러리의 리드 쌍을 본원에 기술된 바와 같이 이 초기의 어셈블리에 매핑시켰다. 매핑된 리드 쌍 중 68.9%가 순방향 및 역방향 리드가 모두 >= 20의 맵 품질을 가지는 것이었므로, 어셈블리 내에서 독특하게 매핑된 것으로 간주되었고 중복되지 않았다. 이들 리드 쌍 중 26.8%는 서로 다른 콘티그에 매핑된 순방향 및 역방향 리드를 가졌으므로, 어셈블리의 추가의 스캐폴딩을 위해 잠재적으로 유익하였다. 또한, 동일한 라이브러리 데이터를 사용하여 페어드 250bp 리드로 50X 커버리지의 디스코바 (Discovar) 어셈블리를 스캐폴드하였다 (Sharpe et al., 2015).

[0626] 생성된 라이브러리 샘플 게놈 DNA, 및 리드 쌍을 기반으로 콘티그 절단과 재스캐폴딩을 위한 "HiRise" 라 불리는 소프트웨어 파이프라인이 어떻게 콘티그에 연결되는지를 기술하기 위하여 확률 모델을 개발하였다. 상기 모델링은 2개의 선두 WGS 어셈블러: MERACULOUS (Chapman et al., 2011) 및 ALLPATHS-LG (APLG) (Gnerre et al., 2011)에 의해 만들어진 포스미드 엔드 쌍에서의 광범위한 커버리지를 포함하는 풍부한 WGS 데이터세트의 어셈블리와 완전성, 근접성 및 정확성에 대하여 비교하는 것을 포함하였다 (표 1). 어셈블리 품질의 정렬 기반 비교를 구축하는 것과 관련된 임의적인 선택을 피하기 위하여, 배수체 NA12878 어셈블리의 각각의 하플로타입 중에서 정확히 한번 발생하는 모든 구별되는 101 bp 서열들 중 무작위로 선택된 하위세트인 2천 5백 4십만개의 101 bp 서열의 어셈블링된 위치를 기반으로 비교를 하였다 (Rozowsky et al., 2011).

표 1

[0627]

	부정합			N50	95% CI	%C	%
	5 Kbp	10 Kbp	50 Kbp	(Mbp)	50 Kbp Δ		배향오류
MERACULOUS	0.032	0.030	0.022	9.1	1.3 Kbp	94.8	0.09
APLG	0.245	0.187	0.130	12.1	6.4 Kbp	92.2	0.4
MERAC (33Kbp N50) + HiRise:	0.028	0.011	0.009	12.6	7.7 Kbp	94.1	1.3
Discovar (178Kbp N50) + HiRise:	0.102	0.097	0.076	29.9	3.2 Kbp	97.9	1.4

[0628] 표 1: 스캐폴딩 결과. 부정합을 확인하기 위한 3개의 서로 다른 역치에서 부정합을 함유하는 스캐폴드에서 각 어셈블리 중 일부. 스캐폴드 N50. 50 Kbp 분리 차이 95% 신뢰 구간 (95% CI = x 라는 말은 어셈블리 내의 독특한 101-mer 태그의 쌍이 주어지는 경우, 이들 중 95%는 참조에서 서로에 대해 50 Kbp ± x 이내임). 완전성 (%C); 101-mer의 일부가 배향이 잘못됨.

[0629] **실시예 3: 원거리 스캐폴딩 정확도 측정**

[0630] HiRise 파이프라인이 생성한 스캐폴드는 공지된 MERACULOUS 및 APLG 어셈블리 (모두 페어드 포스미드 엔드 리드 내의 깊은 커버리지에 의존함)보다 길었고 전체적인 미스어셈블리의 비율이 낮았다. 표 1은 부정합을 함유하고 있는 스캐폴드에서 발견된 전체 어셈블리 중 일부를 보여주는데, 여기서 부정합은 참조 배수체 내의 하나 이상의 염색체의 5 Kbp, 10 Kbp 또는 50 Kbp 이상의 범위인 독특한 101-mer의 구역을 갖는 것으로 정의된다. 또한, 상기 표는 NA12878의 다른 어셈블리와 비교하여 HiRise 어셈블리의 4번의 연속적인 작업에 대한 완전성과 근접성의 측정치를 보여준다.

[0631] 본원에 제공된 방법에 의해 생성된 리드 상을 생성하는 DNA 결찰의 결과가 정의된 상대적인 가닥의 리드 쌍을 생성하는데에 한정되지 않기 때문에, 스캐폴딩 도중 콘티그의 상대적 배향은 리드 밀도 정보로부터 추론되어야 한다. 그 결과, HiRise 계산을 사용하는데 이르른 스캐폴드는 대부분 작은 콘티그에서 발생하는 잘못 배향된 101-mer의 비율이 다른 어셈블리에서 발견되는 것보다 높았다 (1.3%). 잘못 배향된 101-mer를 함유하는 콘티그의 중간 크기는 2.1 Kbp였다.

[0632] **실시예 4: 개선된 악어 어셈블리**

[0633] 미국 악어 (미시시피 악어)에 대해 본원에 기술된 *디 노보* 염색질 리모델링 방법을 사용하여 구축된 단일 DNA 단편 라이브러리를 생성하여 일루미나 HiSeq 2500 상에서 2억 천 7십만 리드를 시퀀싱하였다. 리드 쌍을 공개적으로 이용가능한 데이터 (Green et al., 2014)을 사용하여 생성한 *디 노보* 어셈블리 (N50 81 Kbp)에 매핑시키고, HiRise 스캐폴딩 파이프라인을 적용시켰다. 생성된 어셈블리는 10.3 Mbp의 스캐폴드 N50을 가졌다.

이들 스캐폴드의 정확도를 평가하기 위하여, 사전에 생성된 1,485개 (Shedlock et al., 2007)의 박테리아 인공 염색체 (BAC) 엔드 서열를 어셈블리에 정렬하였다. 이들 중 1,298 쌍을 콘티그 어셈블리와 HiRise 스캐폴드 버전에 대해 90% 커버리지 및 95% 동일성을 갖는 GMAP (Wu and Watanabe, 2005)로 독특하게 정렬하였다. 입력 어셈블리에서, 12.5%의 BAC 엔드 쌍들을 예측된 배향과 이격으로 동일한 스캐폴드 내에 포획하였다. HiRise 어셈블리에서는, 96.5%의 BAC 엔드 쌍들을, 올바른 상대적 배향의 동일한 스캐폴드 상의 98.1%의 BAC 엔드 쌍과 함께 동일한 스캐폴드 내에 정렬하였다. 5개 (0.39%)의 BAC 엔드 쌍들을 동일한 스캐폴드이지만 인서트 크기 보다 현저히 멀리 떨어진 곳에 배치하였고, 14개 (1.08%)는 별도의 스캐폴드이지만 거리가 인서트 크기보다 먼 스캐폴드의 엣지로부터 충분히 멀리 떨어진 곳에 배치하였는데, 이는 8.36 Mbp의 어셈블리 당 1 미만의 부정합의 전체 밀도를 시사하는 것이다.

[0634] **실시예 5: 상 정확도의 평가**

[0635] 순방향 및 역방향 리드가 모두 이형접합 부위를 커버하는 리드 쌍들을 사용하여 직접적으로 하플로타입 상을 판독하였다. 본원에 기술된 *디 노보* 염색질 리모델링 및 단편화 방법에 의해 생성된 리드 쌍 중에 커버된 거리가 입력 DNA의 크기만큼 클 수 있기 때문에, 상 정보의 정확도 및 GM12878 샘플 내에서 하플로타입 상을 측정하는 데 있어서 그의 유용성을 평가하였다. GM12878이 3중 시퀀싱된 개체로부터 유래하기 때문에, 신뢰성 있는 하플로타입 상 정보를 사용하여 페이징 정보의 정확도를 평가하였다. 하플로타입 정보를 제공하고 10 Kbp 내지 150 Kbp 범위인 리드 쌍들은 GM12878에 대해 공지된 하플로타입 상과 99.83% 일치하였다.

[0636] **실시예 6: 구조적 변이의 확인**

[0637] 한 개체의 페어드 서열 리드를 참조에 대하여 매핑되는 것은 전위, 결실, 중복과 같은 연속된 핵산 또는 게놈 구조에서의 차이를 확인하기 위해 가장 흔히 사용되는 서열 기반 방법이다 (Tuzun et al., 2005). 도 4A 및 4B는 인간 참조 게놈 GRCh38에 매핑된 GM12878로부터 재어셈블링된 염색질의 DNA 근접 결찰에 의해 생성된 리드 쌍이 어떻게 이러한 두 구조적 차이를 드러내고 있는지를 보여준다. 구조적 차이를 확인하기 위한 리드 쌍 데이터의 민감성과 특이성을 추정하기 위해, 이형접합성 전위의 효과를 모의실험하기 위해 구축된 모의실험 데이터 세트에 대한 최대 우도 판단자를 시험하였다. GRCh38 참조 서열에 대해 생성된 NA12878 리드의 매핑으로부터 규정된 길이 L의 간격을 무작위로 선택하여, 생성된 각 리드 쌍을 독립적으로 무작위로 전위되거나 또는 참조 하플로타입으로 지정하는 단계, 및 이에 따라 상기 매핑된 좌표를 편집하는 단계에 의하여 시험 데이터를 구축하였다. 비대립형질 상동성 재조합은 인간 게놈에서 관찰되는 많은 구조적 변이에 대한 원인이 되며, 이는 반복되는 서열의 긴 블록에서 일어나는 많은 변이 브레이크포인트를 초래한다 (Kidd et al., 2008). 리드들의 거리 W 내에서 매핑된 모든 리드들을 제거함으로써 전위 브레이크포인트 주위의 다양한 길이의 반복 서열의 효과를 모의실험하였다. 전위 브레이크포인트에서 반복 서열의 부재 하에, 1 Kbp, 2 Kbp 및 5 Kbp 전위 각각에 대해, 민감도 (특이성)는 각각 0.76 (0.88), 0.89 (0.89) 및 0.97 (0.94)이었다. 전위 브레이크포인트에서 (매핑불가능한) 반복 서열의 1 Kbp 영역이 모의실험에 사용되는 경우에, 5 Kbp의 전위에 대한 민감도 (특이성)는 0.81 (0.76)이었다.

[0638] **실시예 7: DNA 준비**

[0639] 제조업자의 설명서에 따라 퀴아젠 (Qiagen)의 블러드 앤 셀 미디 키트를 사용하여 DNA를 추출하였다. 간략히 요약하면, 세포를 용해시키고 원심분리하여 핵을 분리하였다. 프로테이나아제 K와 리보뉴클레아제 (RNase) A의 혼합물을 사용하여 상기 핵을 추가로 처리하였다. DNA를 퀴아젠 게놈 컬럼에 결합, 세척, 용출시키고, 이소프로판올 중에 침전시킨 후 원심분리에 의해 펠렛화하였다. 건조 후, 펠렛을 200 $\mu\ell$의 TE (퀴아젠) 중에 재현탁시켰다.

[0640] **실시예 8: 염색질 어셈블리**

[0641] 액티브 모티프의 *시험관내* 염색질 어셈블리 키트를 사용하여 게놈 DNA로부터 염색질을 27℃에서 밤새 어셈블링하였다. 항온배양 후, 성공적인 염색질 어셈블리를 확인하기 위해 마이크로코커스 핵산가수분해효소(MNase)의 분해에 대해 10%의 샘플을 사용하였다.

[0642] **실시예 9: 비오틴화 및 제한효소 절단**

[0643] 염색질을 요오드아세틸-PEG-2-비오틴 (IPB)을 사용하여 비오틴화하였다. 비오틴화 후, 상기 염색질을 실온 (RT)에서 15분간 1%의 포름알데히드 중에 고정시킨 다음, 2배 몰 초과량의 2.5M 글리신으로 냉각시켰다. 4℃에서 최소 3시간 동안 1L의 투석 버퍼 (10mM Tris-Cl, pH8.0, 1mM EDTA)에 대하여 Slide-A-Lyzer 20 KDa MWCO 투석 카세트 (피어스, Pierce) 중에서 염색질을 투석하여 초과량의 IPB와 가교결합된 글리신을 제거하였다. 이후, 상

기 염색질을 37℃에서 4시간 동안 1X CutSmart 중에서 MboI 또는 MluCI를 사용하여 처리하였다. 상기 염색질을 4℃에서 2시간 동안 50 KDa MWCO 투석 플렉스 튜브 (IBI Scientific # IB48262) 중에서 다시 투석하고 다시 신선한 버퍼로 재차 밤새 투석한 후, 효소 뿐 아니라 짧은 유리 DNA 단편들도 제거하였다.

[0644] 염색질에 첨가하여 RT에서 1시간 동안 항온배양하기 전에, PBS + 0.1% Tween-20 중에서 Dynabead MyOne C1 스트렙타비딘 비드를 세척 및 재현탁시켜 제조하였다. 이후, 상기 비드를 세척, 재농축 및 100 ㎕의 1X NE버퍼 2 중에 재현탁시키기 전에 자성 농축기 랙 상에서 농축시켰다.

[0645] **실시예 10: dNTP 충전 (fill-in)**

[0646] 충전 반응을 수행하는 동안 표지된 dNTP (도 1A-F)가 포획되는 것을 방지하기 위해, RT에서 15분간 유리 비오틴의 존재 하에 비드를 항온배양함으로써 결합되지 않은 스트렙타비딘 부위를 점유하였다. 이후, 상기 비드를 100 ㎕의 1X NE버퍼 2 중에 재현탁시키기 전에 2회 세척하였다. 25℃에서 40분간 165 ㎕의 전체 부피 중에 25 U의 클레노브 (Klenow) (#M0210M, NEB)와 함께 a-S-dGTP 및 비오틴화된 dCTP를 비롯한 dNTP와 함께 항온배양함으로써 점착성 말단을 충전시켜 넣었다. 7 ㎕의 0.5M EDTA를 첨가하여 상기 충전 반응을 중지시켰다. 이후, 상기 비드를 100 ㎕의 예비-결찰 세척 버퍼 (PLWB: 50mM Tris 7.4; 0.4% Triton X-100; 0.1mM EDTA) 중에 재현탁시키기 전에 PLWB 중에서 2회 세척하였다.

[0647] **실시예 11: 결찰**

[0648] 16℃에서 최소 4시간 동안 1mL 이상의 T4 결찰 버퍼 중에서 결찰을 수행하였다. 서로 다른 염색질 응집물 간의 교차 결찰을 최소화하기 위해 큰 결찰 부피를 사용하였다. 40㎕의 0.5M EDTA를 첨가하여 결찰 반응을 중지시켰다. 상기 비드를 농축하고 100㎕의 추출 버퍼 (50mM Tris-cl pH 8.0, 1mM EDTA, 0.2% SDS) 중에 재현탁시켰다. 400㎍의 프로테이나아제 K (#P8102S, NEB)를 첨가한 후, 상기 비드를 55℃에서 밤새 항온배양한 다음, 55℃에서 추가의 200 ㎍의 프로테이나아제 K로 2시간 동안 처리하였다. 2:1 비율의 SPRI 비드, 컬럼 정제 키트, 또는 페놀:클로로포름 추출을 사용하여 DNA를 회수하였다. DNA를 Low TE (10 mM Tris-Cl pH 8.0, 0.5 mM EDTA) 내로 용출시켰다.

[0649] **실시예 12: 엑소뉴클레아제 처리**

[0650] DNA를 100 U의 엑소뉴클레아제 III (#M0206S, NEB)로 37℃에서 40분간 처리하여 비오틴화된 유리 말단을 제거한 다음, SPRI 클린업 및 101 ㎕의 low TE로의 용출을 시행하였다.

[0651] **실시예 13: 전단 및 라이브러리 준비**

[0652] 30초 온 (on)/ 30초 오프(off)의 60 사이클 동안 '저속' 으로 설정된 디아게노드 (Diagenode)의 바이오럽터 (Bioruptor) 세트를 사용하여 DNA를 전단시켰다. 전단 후, 클레노브 폴리머라아제 및 T4 PNK (#EK0032, 써모 사이언티픽, Thermo Scientific)를 사용하여 20℃에서 30분간 DNA를 충전시켜 넣었다. 충전 반응 후, 200 ㎕의 2X NTB (2M NaCl, 10mM Tris pH 8.0, 0.1mM EDTA pH 8.0, 0.2% Triton X-100) 중에서 재현탁시키기 전에, DNA를 준비시켜 놓은 C1 비드 상에서 Tween 세척 버퍼를 사용해 2회 세척하여 파괴하였다. 일단 샘플이 첨가되면, 상기 비드들을 흔들어주면서 실온에서 20분간 항온배양하였다. 이후, 비드를 Low TE 중에 재현탁시키기 전에 3회 세척하여 비오틴화되지 못한 DNA 단편들을 제거하였다. 확립된 프로토콜을 사용하여 시퀀싱 라이브러리를 생성하였다 (Meyer and Kircher, 2010).

[0653] **실시예 14: 리드 매핑**

[0654] 서열 리드들을 접합부가 존재할때마다 이를 절두시켰다 (서열번호 1: GATCGATC (MboI), 서열번호 2: AATTAATT (MluCI)). 이후, SMALT [http://www.sanger.ac.uk/자원/ 소프트웨어/smalt/]를 사용하여 순방향 및 역방향 리드들을 독립적으로 정렬하기 위한 -x 옵션으로 리드들을 정렬하였다. PCR 중복 서열을 피카르드-툴 (Picard-tool)인 마크듀플리킷 (MarkDuplicate) [http://broadinstitute.github.io/ picard/]을 사용하여 마킹하였다. 매핑된 양자 리드가 모두 10 초과의 매핑 품질을 갖는 경우, 중복되지 않은 리드 쌍들을 분석에 사용하였다.

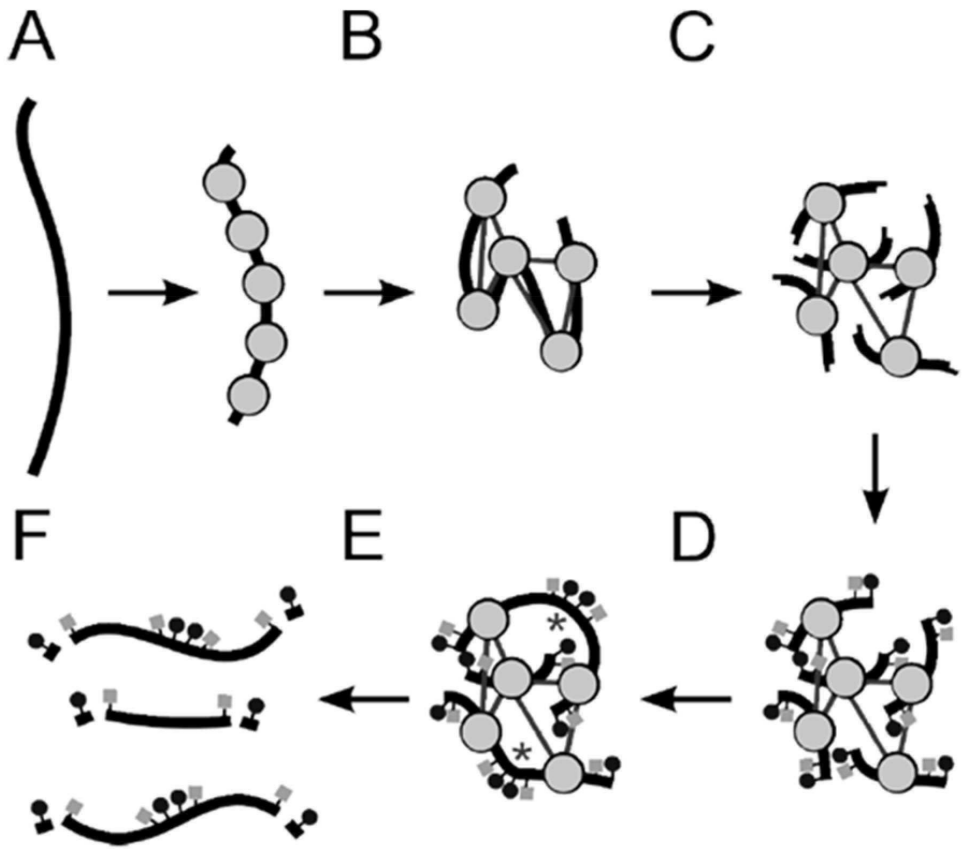
[0655] **실시예 15: *디 노보* 어셈블리**

[0656] 공개적으로 이용가능한 짧은 인서트 및 메이트 페어 리드 (Simpson and Durbin, 2012; Green et al., 2014)를 사용하는 Meraculous 2.0.3 (Chapman et al., 2011)으로 인간 및 악어의 *디 노보* 샷건 어셈블리를 생성하였다. 악어 메이트 페어 리드를 트림모매틱 (Trimmomatic)을 사용하여 어댑터 시퀀스를 트리밍 (adapter-trimmed) 하였다 (Bolger et al., 2014). 일부 중복되는 악어의 짧은 인서트 리드를 "병합"시켰다. 이들을 순방향 및 역방

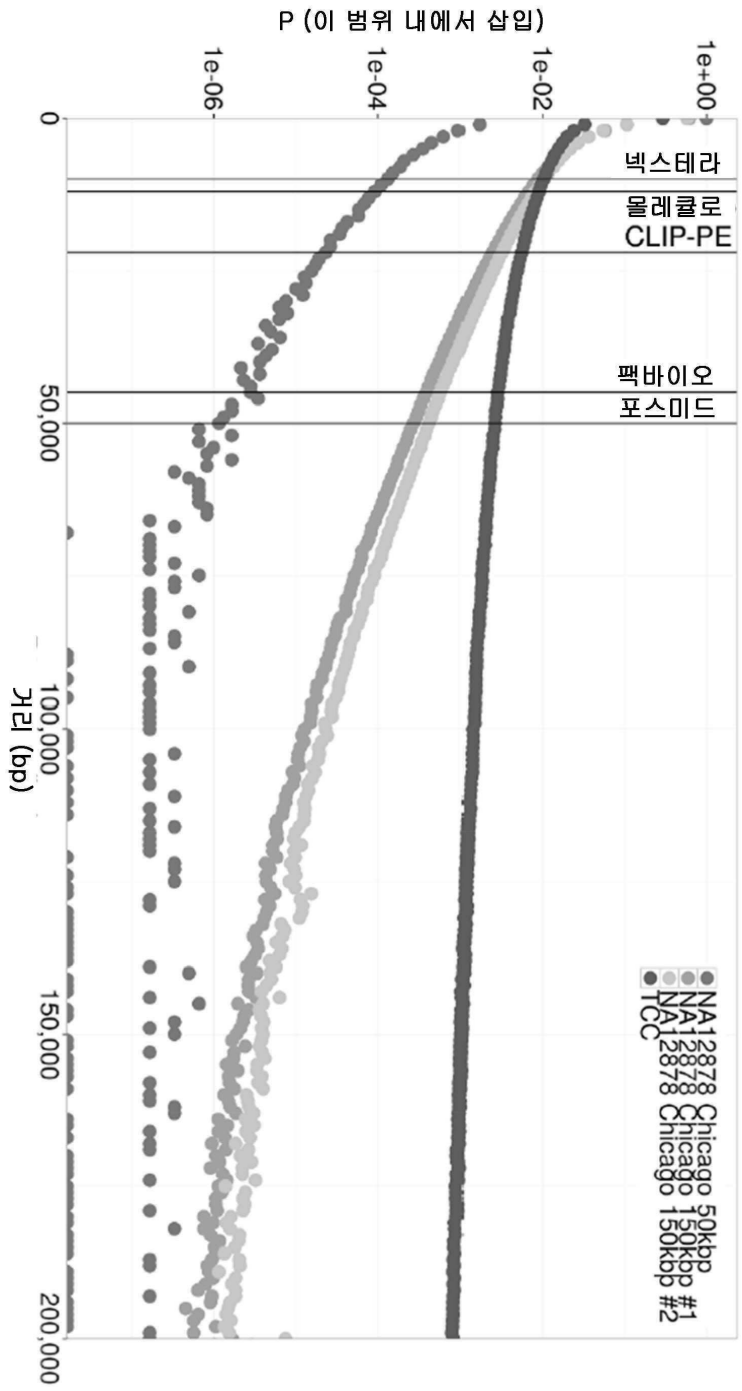
향 리드로 다시 병합 해제시켰다.

도면

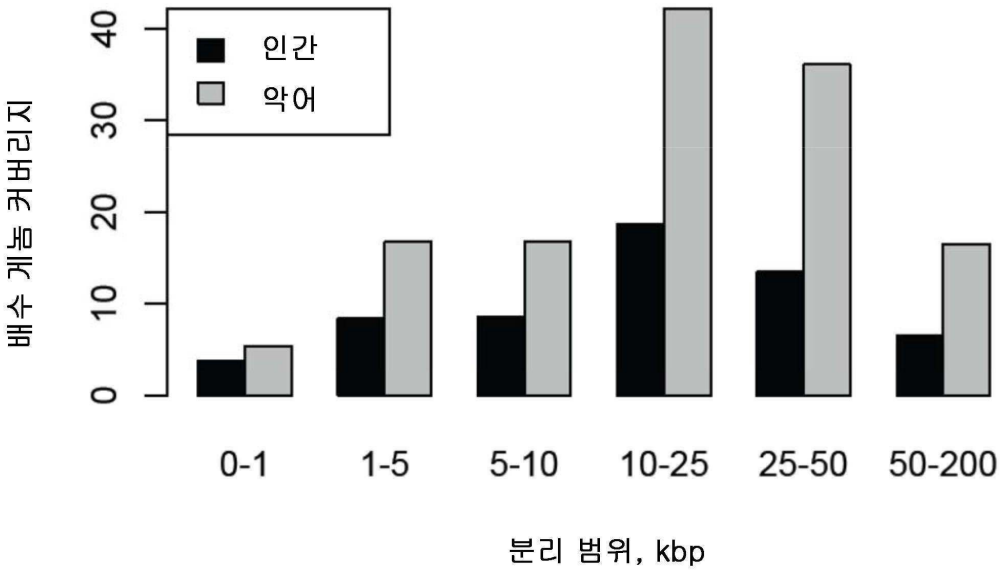
도면1



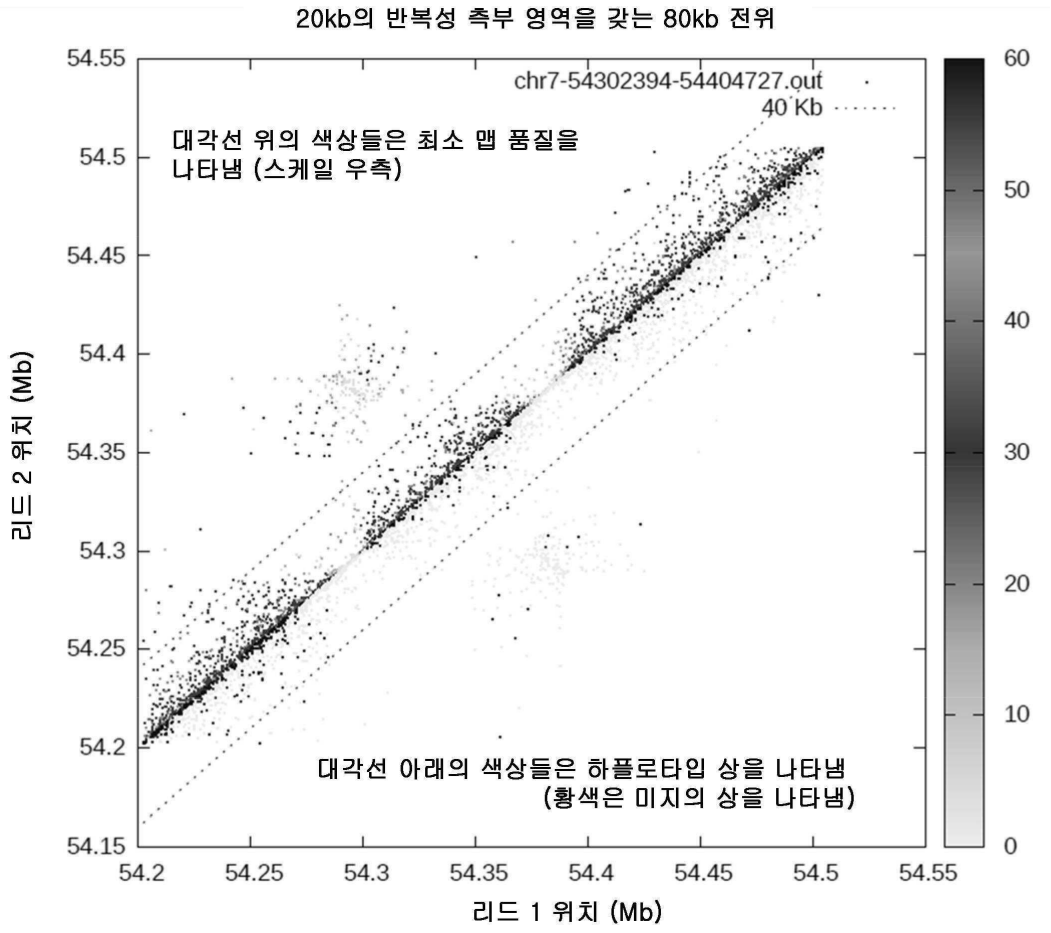
도면2


P (이 범위 내에서 삽입)
1e+00
1e-02
1e-04
1e-06
0
50,000
100,000
150,000
200,000
거리 (bp)
넥스테라
몰레큘로
CLIP-PE
팩바이오
포스미드
NA12878 Chicago 50kbp
NA12878 Chicago 150kbp #1
NA12878 Chicago 150kbp #2
TCC

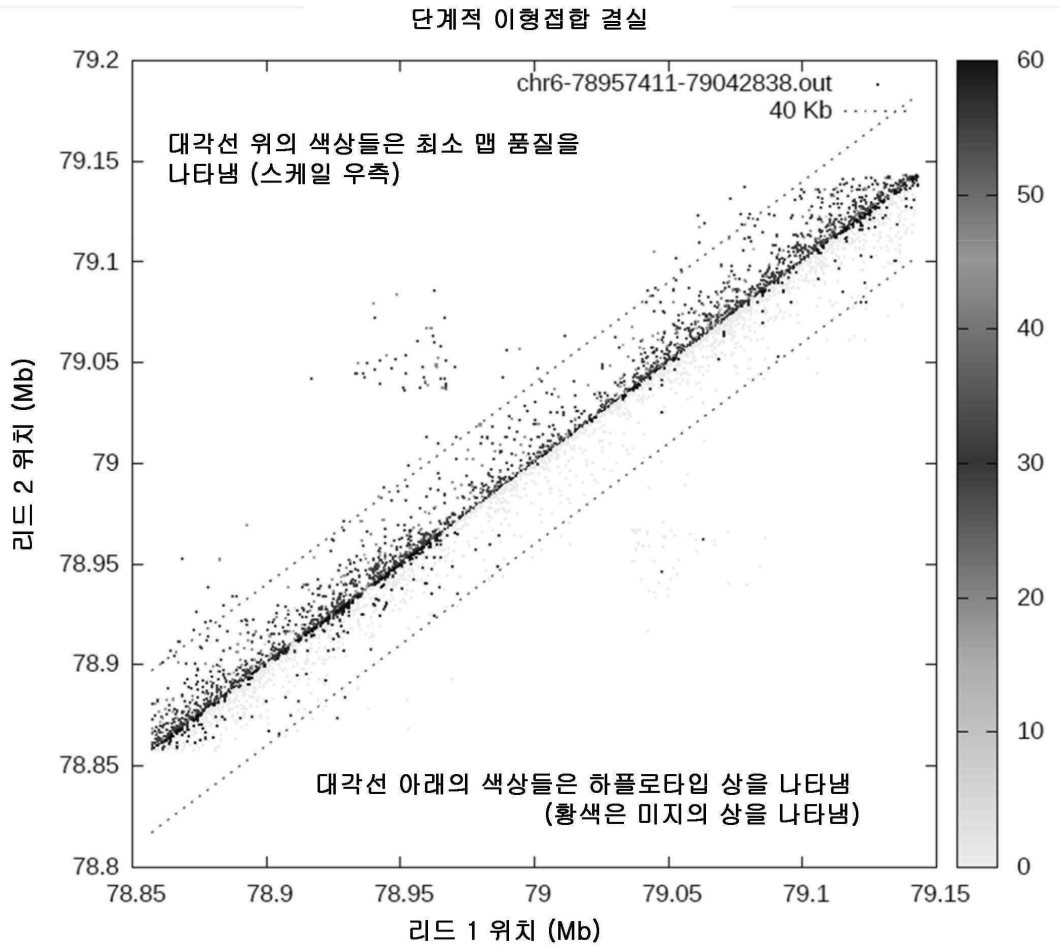
도면3


인간
악어
배수 게놈 커버리지
0
10
20
30
40
0-1
1-5
5-10
10-25
25-50
50-200
분리 범위, kbp

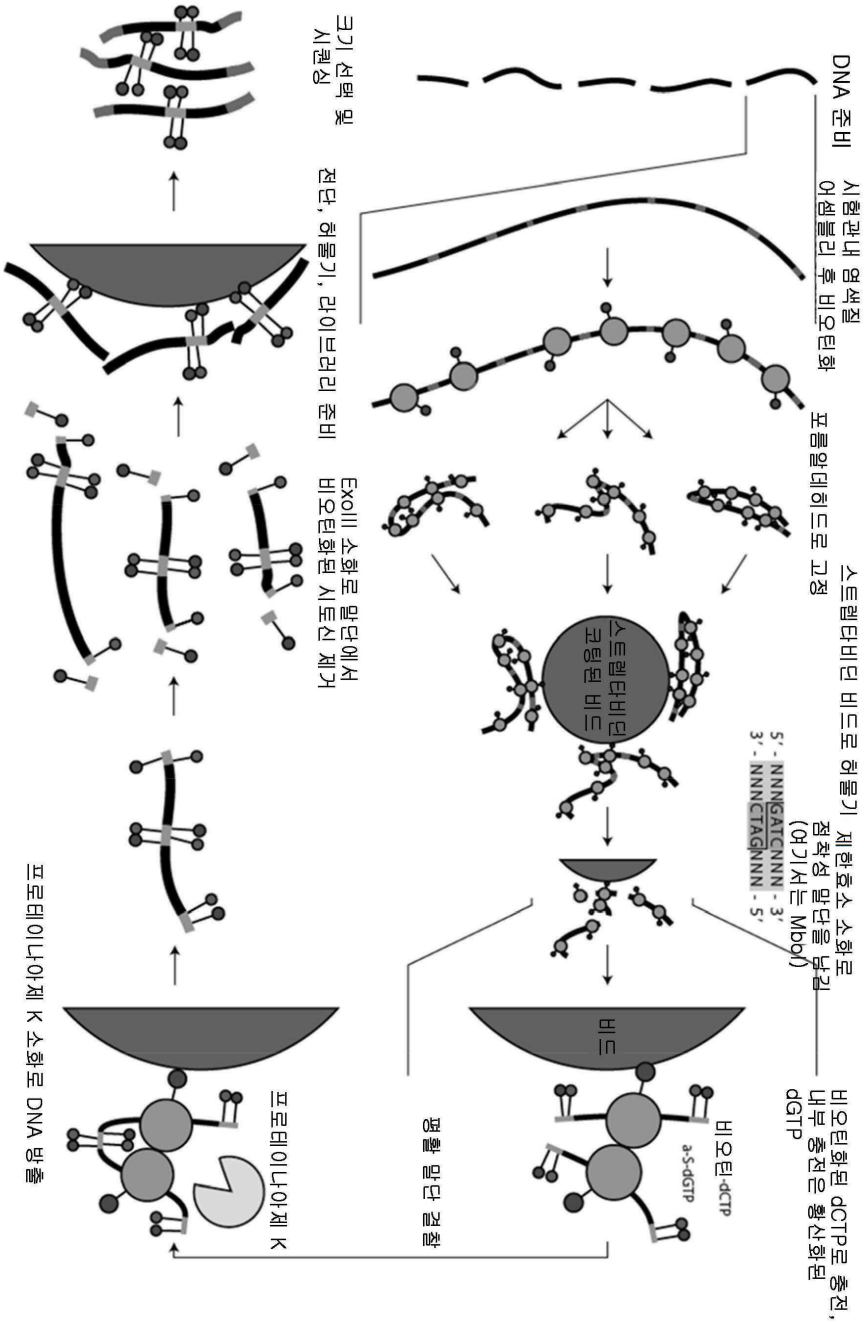
도면4a


20kb의 반복성 측부 영역을 갖는 80kb 전위
chr7-54302394-54404727.out
40 Kb
대각선 위의 색상들은 최소 맵 품질을
나타냄 (스케일 우측)
대각선 아래의 색상들은 하플로타입 상을 나타냄
(황색은 미지의 상을 나타냄)
리드 2 위치 (Mb)
리드 1 위치 (Mb)
54.15
54.2
54.25
54.3
54.35
54.4
54.45
54.5
54.55
0
10
20
30
40
50
60

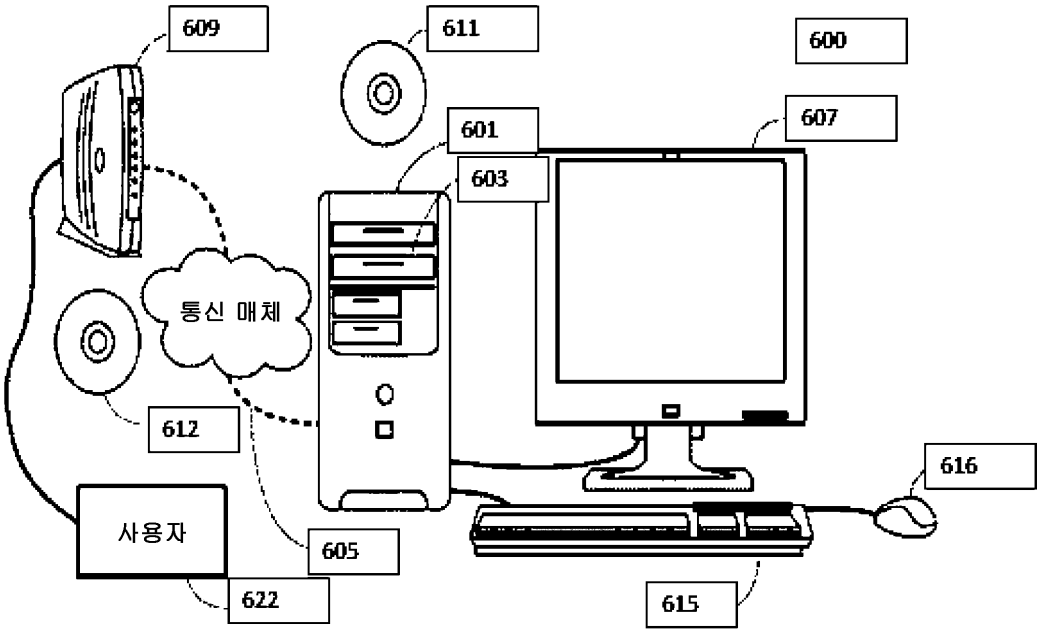
도면4b


단계적 이형접합 결실
chr6-78957411-79042838.out
40 Kb
대각선 위의 색상들은 최소 맵 품질을
나타냄 (스케일 우측)
대각선 아래의 색상들은 하플로타입 상을 나타냄
(황색은 미지의 상을 나타냄)
리드 2 위치 (Mb)
리드 1 위치 (Mb)
79.2
79.15
79.1
79.05
79
78.95
78.9
78.85
78.8
78.85
78.9
78.95
79
79.05
79.1
79.15
60
50
40
30
20
10
0

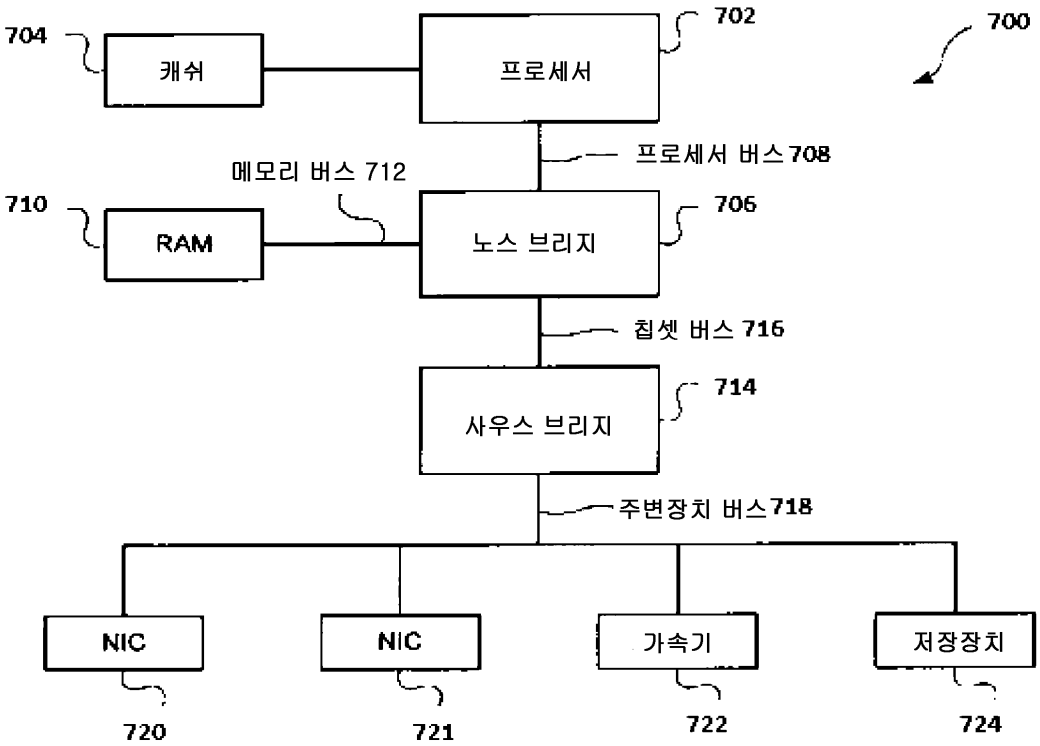
도면5


DNA 준비
시험관내 염색질 어셈블리 후 비오틴화
포름알데히드로 고정
스트렙타비딘 비드로 헤물기
제한효소 소화로 절편성 말단을 남김 (여기서는 MboI)
5' - NNNGATCNNN - 3'
3' - NNNCTAGNNN - 5'
비오틴화된 dCTP로 충전, 내부 충전은 형산화된 dGTP
스트렙타비딘 코팅된 비드
비드
비오틴-dCTP
a-S-dGTP
평활 말단 결찰
크기 선택 및 시퀀싱
전단, 헤물기, 라이브러리 준비
ExoIII 소화로 말단에서 비오틴화된 시토신 제거
프로테이나아제 K 소화로 DNA 방출
프로테이나아제 K

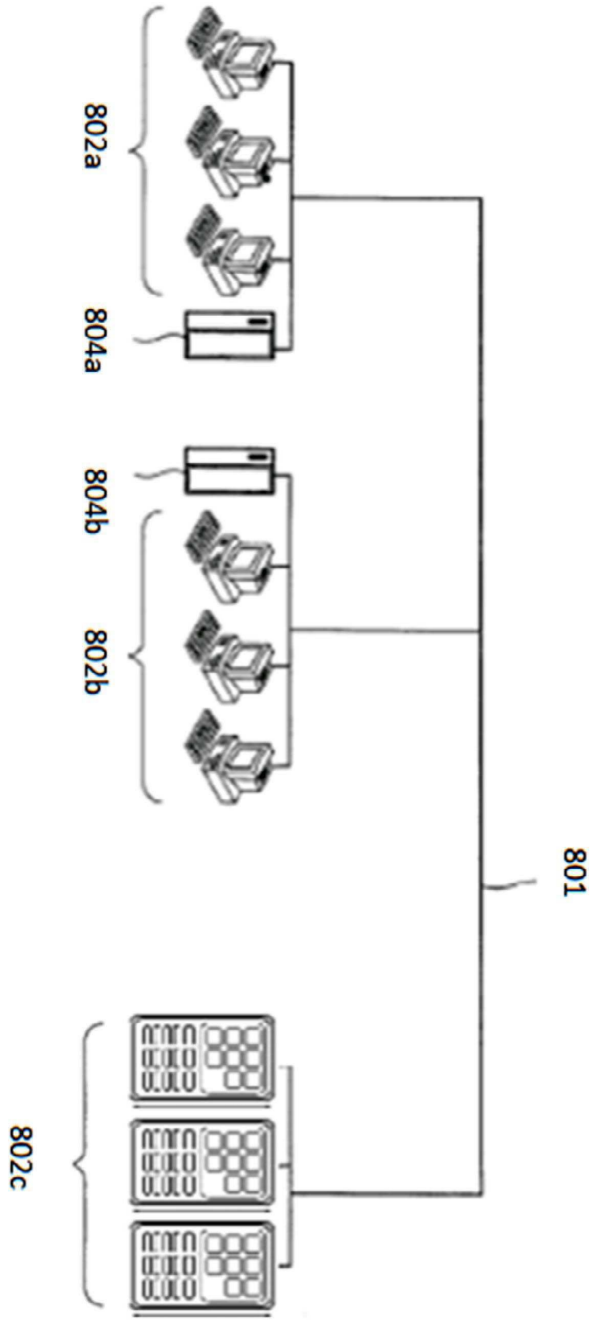
도면6


609
611
600
601
607
603
통신 매체
612
616
사용자
605
622
615

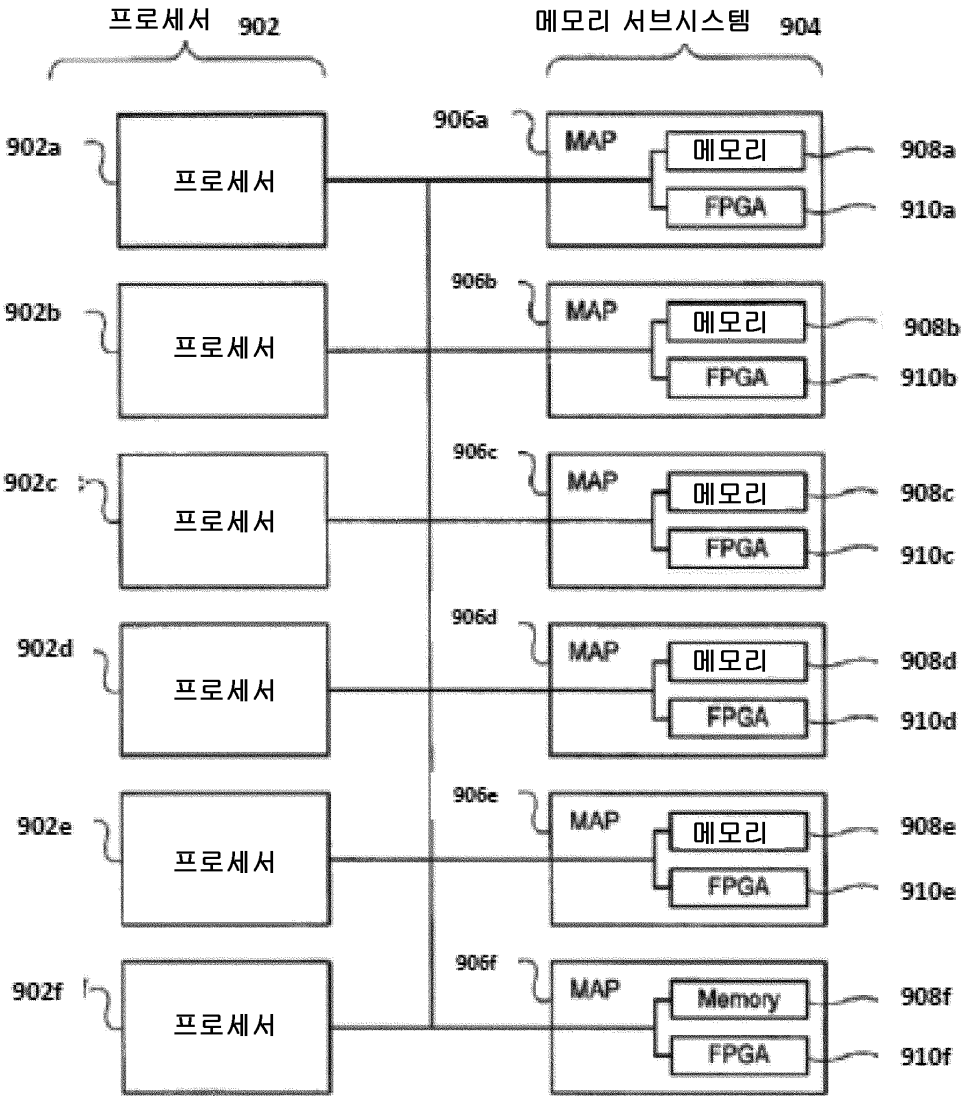
도면7


704
캐쉬
프로세서
702
700
프로세서 버스 708
메모리 버스 712
710
RAM
노스 브리지
706
칩셋 버스 716
사우스 브리지
714
주변장치 버스 718
NIC
NIC
가속기
저장장치
720
721
722
724

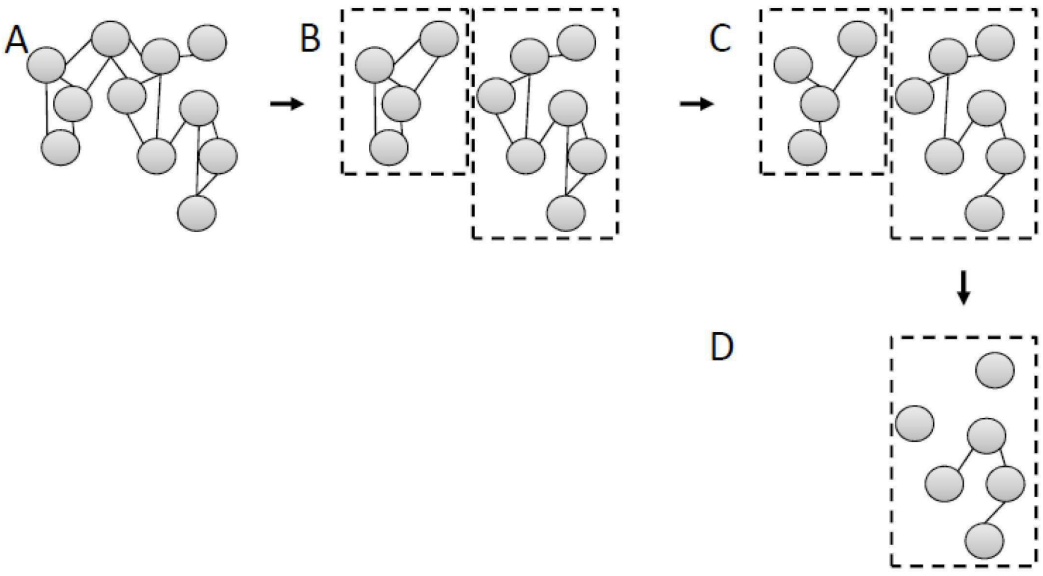
도면8



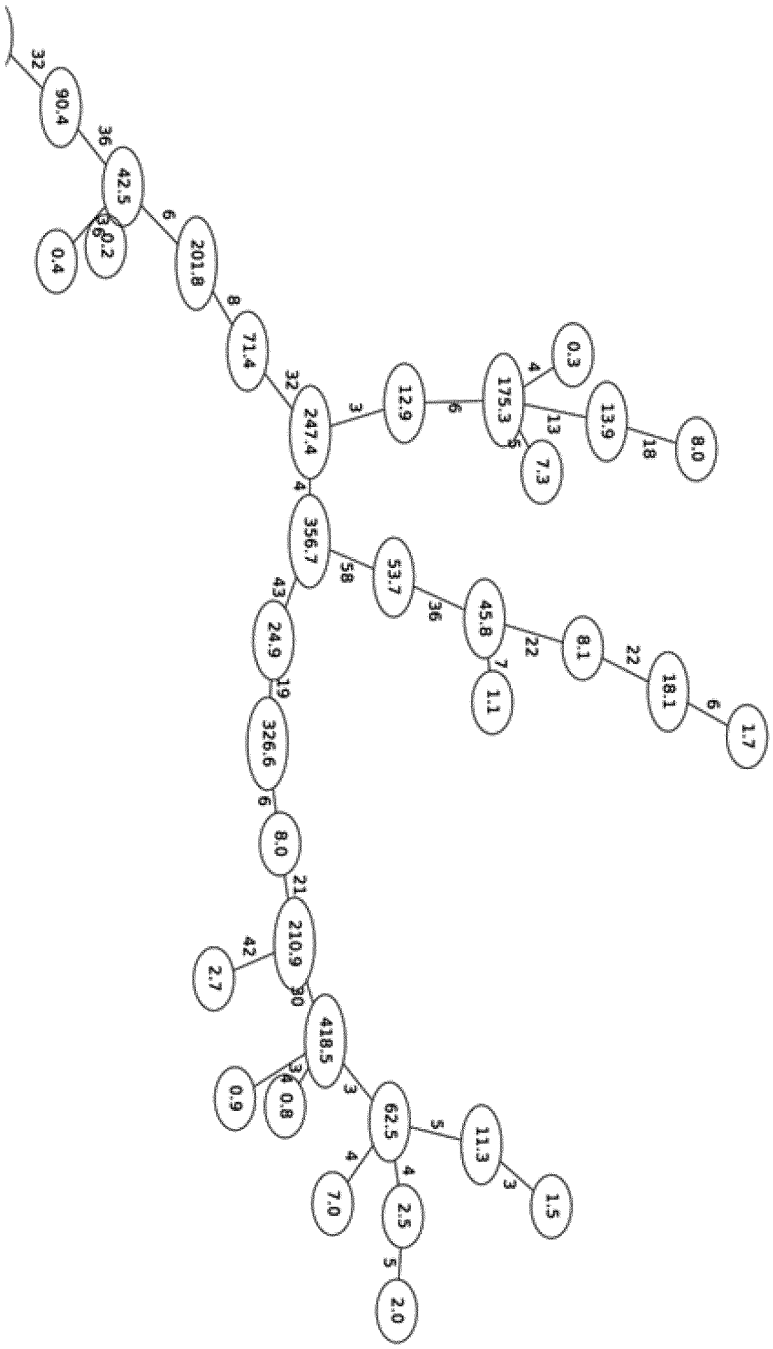
도면9


프로세서 902
메모리 서브시스템 904
902a
프로세서
906a
MAP
메모리
FPGA
908a
910a
902b
프로세서
906b
MAP
메모리
FPGA
908b
910b
902c
프로세서
906c
MAP
메모리
FPGA
908c
910c
902d
프로세서
906d
MAP
메모리
FPGA
908d
910d
902e
프로세서
906e
MAP
메모리
FPGA
908e
910e
902f
프로세서
906f
MAP
Memory
FPGA
908f
910f

도면10


A
B
C
D

도면11


32
90.4
36
42.5
6
201.8
8
71.4
32
247.4
4
356.7
58
53.7
36
45.8
22
8.1
22
18.1
6
1.7
7
1.1
43
24.9
19
326.6
6
8.0
21
210.9
42
2.7
30
418.5
3
62.5
5
11.3
3
1.5
4
2.5
5
2.0
4
7.0
3
0.8
3
0.9
3
12.9
6
175.3
4
0.3
13
13.9
18
8.0
5
7.3
6
0.2
6
0.4