(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년06월22일
(11) 등록번호 10-2412442
(24) 등록일자 2022년06월20일

(51) 국제특허분류(Int. Cl.)
C12Q 1/6806 (2018.01) *C12Q 1/6816* (2018.01)
C40B 40/06 (2006.01)
(52) CPC특허분류
C12Q 1/6806 (2018.05)
C12Q 1/6816 (2018.05)
(21) 출원번호 10-2018-7036302
(22) 출원일자(국제) 2017년05월12일
심사청구일자 2020년05월11일
(85) 번역문제출일자 2018년12월13일
(65) 공개번호 10-2019-0037201
(43) 공개일자 2019년04월05일
(86) 국제출원번호 PCT/US2017/032466
(87) 국제공개번호 WO 2017/197300
국제공개일자 2017년11월16일
(30) 우선권주장
62/336,252 2016년05월13일 미국(US)
62/410,599 2016년10월20일 미국(US)
(56) 선행기술조사문헌
WO2016044313 A1*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
더브테일 제노믹스 엘엘씨
미국 95066 캘리포니아주 스코츠 밸리 엔터프라이즈 웨이 100 스위트 에이101
(72) 발명자
트롤 크리스토퍼 존
미국 95065 캘리포니아주 산타크루즈 루니 스트리트 63
파워스 마틴 피
미국 94116 캘리포니아주 샌프란시스코 22번 애비뉴 2654
(뒷면에 계속)
(74) 대리인
김진희, 김태홍

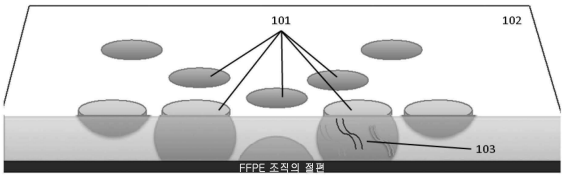
전체 청구항 수 : 총 18 항 심사관 : 김현태

(54) 발명의 명칭 **보존된 샘플로부터의 장범위 링키지 정보의 회수**

(57) 요 약

본 개시내용은 보존된 샘플로부터 게놈 또는 염색체 수준의 구조적 정보를 단리하는 방법을 제공한다. 일부 경우에, 장범위 핵산 정보가 회복 불가능하게 상실되는 것으로 여겨지는 조건 하에서 보존된 샘플, 예컨대 FFPE 샘플이, 샘플 보존 과정의 일부로서 안정화된 핵산-단백질 복합체를 회수하기 위해 처리된다. 상기 복합체는 어느 핵산이 공통의 복합체에 결합되어 있는지에 관한 정보를 회수하도록 처리되고, 상기 정보는 게놈의 구조적 정보를 회수하는 데 사용된다.

대 표 도 - 도1a



(52) CPC특허분류
C40B 40/06 (2013.01)
C12Q 2521/501 (2013.01)
C12Q 2523/101 (2013.01)
C12Q 2525/161 (2013.01)
C12Q 2563/179 (2013.01)
C12Q 2565/514 (2013.01)
(72) 발명자
퍼트넘 니콜라스 에이치
미국 02453 매사추세츠주 월섬 리버 스트리트 36 아파트먼트 362
블란첫 마르코
미국 95060 캘리포니아주 산타크루즈 펠톤 애비뉴 707 아파트먼트 307
하틀리 폴
미국 95136 캘리포니아주 새너제이 마운트캐슬 웨이 4265

명 세 서

청구범위

청구항 1

대상체로부터 단백질-DNA 복합체를 포함하는 보존된 샘플을 수득하는 단계;

상기 보존된 샘플을 40℃ 이하의 온도에서 프로테이나제로 처리하여, 단백질-DNA 복합체가 파괴되지 않고, 제1 세그먼트와 제2 세그먼트가 포스포디에스테르 백본과는 관계없이 함께 유지되고, 제1 세그먼트와 제2 세그먼트가 각각 하나 이상의 노출된 핵산 말단을 갖도록 핵산을 단리하는 단계; 및

상기 샘플 내의 상기 단백질-DNA 복합체의 핵산을 분석함으로써 게놈의 구조적 정보를 도출하는 단계

를 포함하는 게놈의 구조적 정보를 수득하는 방법으로서, 상기 보존된 샘플은 포르말린 고정 파라핀 포매된(formalin fixed paraffin-embedded, FFPE) 샘플인 방법.

청구항 2

제1항에 있어서, 보존된 샘플이 가교결합되는 것인 방법.

청구항 3

제2항에 있어서, 보존된 샘플이 포름알데히드, 포르말린, UV 광, 미토마이신 C, 질소 머스타드, 멜팔란, 1,3-부타디엔 디에폭시드, 시스 디아민디클로로백금(II) 및 사이클로포스파미드 중 적어도 하나를 사용하여 가교결합되는 것인 방법.

청구항 4

제1항에 있어서, 보존된 샘플이 그 내부의 핵산에 대한 위치 정보를 유지하는 것인 방법.

청구항 5

제1항에 있어서, 게놈의 구조적 정보가 참조 게놈에 비해 역위, 삽입, 결실 및 전좌 중 적어도 하나를 나타내는 것인 방법.

청구항 6

제1항에 있어서, 핵산의 제1 세그먼트 및 제2 세그먼트에 대한 페이즈(phase) 상태를 나타내는 정보를 도출하는 단계를 포함하는 방법.

청구항 7

제1항에 있어서, 물리적 링키지 정보를 전달하기 위해 제1 세그먼트의 노출된 핵산 말단과 제2 세그먼트의 노출된 핵산 말단에 태그를 부착하는 단계를 포함하는 방법.

청구항 8

제7항에 있어서, 태그 부착이, 올리고뉴클레오티드가 구조적 정보를 나타내는 정보를 전달하도록, 보존된 샘플의 제1 세그먼트의 노출된 말단에 올리고뉴클레오티드를 라이게이션하는 것을 포함하는 것인 방법.

청구항 9

제7항에 있어서, 태그 부착이, 쌍을 이룬 말단 분자를 형성하도록, 제1 세그먼트의 노출된 말단을 제2 세그먼트의 노출된 말단에 라이게이션하는 것을 포함하는 것인 방법.

청구항 10

제1항에 있어서, 보존된 샘플이, 보존된 조직 샘플을 크실렌 및 에탄올 중 적어도 하나에 접촉시킴으로써 처리

되는 것인 방법.

청구항 11

제1항에 있어서, 보존된 샘플이, 보존된 조직 샘플을 안트라닐레이트 및 포스파닐레이트 중 적어도 하나에 접촉시킴으로써 처리되는 것인 방법.

청구항 12

제1항에 있어서, 보존된 샘플이 조직 내의 그의 입체배열(configuration)을 반영하는 위치 정보를 보존하는 것인 방법.

청구항 13

제1항에 있어서, 보존된 샘플이, 핵산을 단리하기 전에 균질화되지 않는 것인 방법.

청구항 14

제1항에 있어서, 보존된 샘플이, 핵산을 단리하기 전에 적어도 1주일 동안 보관되는 것인 방법.

청구항 15

제1항에 있어서, 보존된 샘플이, 핵산을 단리하기 전에 적어도 6개월 동안 보관되는 것인 방법.

청구항 16

제1항에 있어서, 보존된 샘플이, 핵산을 단리하기 전에 수집 지점으로부터 수송되는 것인 방법.

청구항 17

제1항에 있어서, 프로테이나제가 프로테이나제 K를 포함하는 것인 방법.

청구항 18

제1항에 있어서, 보존된 샘플을 1 시간 이하 동안 프로테이나제로 처리하는 것인 방법.

청구항 19

삭제

청구항 20

삭제

청구항 21

삭제

청구항 22

삭제

청구항 23

삭제

청구항 24

삭제

청구항 25

삭제

청구항 26

삭제

청구항 27

삭제

청구항 28

삭제

청구항 29

삭제

청구항 30

삭제

청구항 31

삭제

청구항 32

삭제

청구항 33

삭제

청구항 34

삭제

청구항 35

삭제

청구항 36

삭제

청구항 37

삭제

청구항 38

삭제

청구항 39

삭제

청구항 40

삭제

청구항 41

삭제

청구항 42

삭제

청구항 43

삭제

청구항 44

삭제

청구항 45

삭제

청구항 46

삭제

발명의 설명

배 경 기 술

[0001] [상호 참조]

[0002] 본원은 그 전부가 본원에 참고로 포함되는, 2016년 5월 13일 출원된 미국 특허 가출원 제62/336,252호의 이익을 주장하고, 그 전부가 본원에 참고로 포함되는, 2016년 10월 20일 출원된 미국 특허 가출원 제62/410,599호의 이익을 주장한다.

[0003] [배경기술]

[0004] 이론적으로 및 실제적으로 고품질의 매우 인접한 게놈 서열을 생산하는 것은 여전히 어렵다. 이 문제는 포르말린 고정 파라핀 포매된(FFPE: formalin-fixed, paraffin-embedded) 샘플과 같은 보존된 샘플로부터 게놈 서열, 페이징(phasing) 정보 또는 다른 유전 정보를 회수하려고 시도할 때 복잡해진다. FFPE 샘플은 가장 일반적인 보관된 임상 샘플 및 암 샘플 유형이다. 그러나, 고정 및 포매 단계, 및 탈수 및 장기간 보관과 같은 추가의 요인이 DNA 손상을 일으키는 것으로 생각된다. 추가의 DNA 손상 및 단편화는 종종 하룻밤 동안의 프로테이나제 K 처리 및 가교결합을 파괴하기 위한 비등을 포함하는 DNA 추출 과정 동안 발생할 수 있다. 추출 후 전형적인 DNA 단편 길이는 500개 염기쌍 미만이며, 종종 300개 염기쌍 미만이다.

발명의 내용

[0005] 외과적 절제에 따라 보관되거나 약물 시험에 따라 보관된 샘플과 같은 보존된 샘플로부터 게놈의 구조적 정보를 얻는 방법이 본원에서 제공된다. 이러한 방법 중 일부는 핵산을 포함하는, 개체로부터 보존된 샘플을 얻는 단계; 및 샘플 내의 핵산을 분석함으로써 게놈의 구조적 정보를 도출하는 단계를 포함한다. 일부 경우에, 보존된 샘플은 예를 들어 포름알데히드, 포르말린, UV 광, 미토마이신 C, 질소 머스타드, 멜팔란, 1,3-부타디엔 디에폭시드, 시스 디아민디클로로 백금(II) 및 사이클로포스파미드 중 적어도 하나를 사용하여 가교결합된다. 대안으로, 보존된 샘플은 포르말린을 사용하여 가교결합된다. 종종, 보존된 샘플은 그 안에 있는 핵산에 대한 위치 정보를 유지한다. 선택적으로, 보존된 샘플은 포르말린 고정 파라핀 포매된(FFPE) 샘플과 같은 포매된 샘플이다. 게놈의 구조적 정보는 샘플 게놈에 존재하는 경우, 참조 게놈에 비해 역위, 삽입, 결실, 및 전좌(translocation) 중 적어도 하나를 나타내기 위해 충분하다. 다수의 참조 게놈, 예컨대 대상체에 공통인 종의 야생형 게놈 또는 대상체의 참조 조직으로부터 수득된 게놈은 본원의 개시내용과 일치한다. 방법은 종종 핵산의 제1 세그먼트 및 제2 세그먼트에 대한 페이즈(phase) 상태를 나타내는 정보를 도출하는 단계를 포함한다. 선택적으로, 상기 방법은 물리적 링키지 정보(physical linkage information)를 전달하기 위해 샘플의 노출된 핵산 말단에 태그를 부착하는 단계를 포함한다. 일부 경우에, 태그 부착은 보존된 샘플로부터 방출된 DNA 단백질 복합체에 올리고뉴클레오티드를 라이게이션하여 올리고뉴클레오티드가 공통의 복합체를 나타내는 정보를 전달하도록 하는 단계를 포함한다. 올리고뉴클레오티드는 복합체에 특이적인 또는 복합체에 특유한 염기 서열을 포함한다. 다르게는, 바람직한 실시양태에서, 태그 부착은 쌍을 이룬 말단 분자를 형성하기 위해 복합체의 제1 핵산

세그먼트를 복합체의 제2 세그먼트에 라이게이션시키는 것을 포함한다. 이 경우에, 일부 방법은 제1 핵산 세그먼트의 일부 및 제2 핵산 세그먼트의 일부를 시퀀싱하는 단계를 포함한다. 제1 핵산 세그먼트의 일부에 공통적인 특유한 서열을 갖는 콘티그 및 제2 핵산 세그먼트의 일부에 공통적인 특유한 서열을 갖는 콘티그가 핵산 어셈블리의 공통의 스캐폴드에 할당된다. 일부 방법은 쌍을 이룬 말단 핵산 분자를 프로브 세트, 예컨대 형광 프로브이거나 증폭을 지지할 수 있고, 게놈 구조적 재배열에 관여하는 제1 유전자좌 및 제2 유전자좌에 어닐링하는 항체 또는 핵산 프로브에 접촉시키는 단계를 포함한다. 종종, 제1 유전자좌 및 제2 유전자좌는 게놈 구조적 재배열에 의해 영향을 받지 않는 게놈에서 인접하지 않는다. 대안으로, 제1 유전자좌 및 제2 유전자좌는 게놈 구조적 재배열에 의해 영향을 받지 않는 게놈에서 인접한다. 선택적으로, 상기 방법은 프로브 세트에 대한 접촉이 재배열을 나타낼 경우 샘플의 핵산을 시퀀싱하는 단계를 포함한다. 일부 방법은 쌍을 이룬 말단 핵산 분자를 핵산 프라이머를 포함하는 프로브의 세트에 접촉시키는 단계를 포함한다. 일부 경우에, 핵산 프라이머의 세트는 게놈 구조적 재배열에 관여하는 제1 유전자좌 및 제2 유전자좌에 어닐링한다. 이 경우, 핵산 프라이머의 세트는 제1 유전자좌 및 제2 유전자좌가 라이게이션된 쌍을 이룬 말단 분자를 형성할 경우, 핵산 증폭 반응에서 앰플리콘을 생성한다. 이와 유사하게, 일부 경우에, 핵산 프라이머의 세트는 제1 유전자좌 및 제2 유전자좌가 라이게이션된 쌍을 이룬 말단 분자를 형성하지 않을 경우 핵산 증폭 반응에서 앰플리콘을 생성하지 않는다. 일부 경우에, 제1 유전자좌 및 제2 유전자좌는 게놈 구조적 재배열에 의해 영향을 받지 않는 게놈에서 인접하지 않는다. 대안으로, 제1 유전자좌 및 제2 유전자좌는 게놈 구조적 재배열에 의해 영향을 받지 않는 게놈에서 인접한다. 일부 실시양태는 앰플리콘이 쌍을 이룬 말단 핵산 분자에 접촉된 핵산 프라이머의 세트로부터 생성될 경우 샘플의 핵산을 시퀀싱하는 것을 선택적으로 포함한다. 바람직하게는, 보존된 조직 샘플은 단백질 DNA 복합체가 파괴되지 않도록 핵산을 단리하기 위해 처리된다. 일부 경우에, 제1 핵산 세그먼트 및 제2 핵산 세그먼트가 포스포디에스테르 백본과는 관계없이 함께 유지되도록 단백질 DNA 복합체가 단리된다. 일부 경우에, 보존된 조직 샘플은 보존된 조직 샘플을 크실렌에 접촉시킴으로써 처리된다. 일부 경우에, 보존된 조직 샘플은 보존된 조직 샘플을 에탄올에 접촉시켜 처리된다. 일부 경우에, 보존된 조직 샘플은 비등 조건으로부터 샘플을 보호함으로써 처리된다. 일부 경우에, 보존된 조직 샘플은 보존된 조직 샘플을 안트라닐레이트 및 포스파닐레이트 중 적어도 하나에 접촉시킴으로써 처리된다. 일부 경우에, 보존된 조직 샘플은 40℃ 이하의 온도에서 처리된다. 선택적으로, DNA 단백질 복합체는 염색질을 포함한다. 일부 경우에, 보존된 조직 샘플은 조직 내의 그의 입체배열(configuration)을 반영하는 위치 정보를 보존한다. 종종, 보존된 조직 샘플은 보존 동안 또는 핵산을 단리하기 전에 균질화되지 않아, 샘플로부터 절제된 DNA 단백질 복합체의 위치 정보가 보존되고 게놈 구조적 분석의 일부로서 이용 가능하다. 일부 경우에, 보존된 조직 샘플은 핵산을 단리하기 전에 적어도 1주일 동안 보관한다. 일부 경우에, 보존된 조직 샘플은 핵산을 단리하기 전에 적어도 6개월 동안 보관된다. 일부 경우에, 보존된 조직 샘플은 핵산을 단리하기 전에 수집 지점으로부터 수송된다. 일부 경우에, 보존된 조직 샘플은 멸균 환경에서 수집한다. 일부 경우에, 보존된 조직 샘플은 핵산을 단리하기 전에 비멸균 환경에 위치한다.

[0006] 보존된 샘플, 예컨대 가교결합된 파라핀 포매된 조직 샘플로부터 긴 거리 서열 정보, 예컨대 게놈의 구조적 정보를 얻는 방법이 본원에서 제공된다. 이러한 방법 중 일부는 단백질 DNA 복합체가 파괴되거나 붕괴되지 않도록 가교결합된 파라핀 포매된 조직 샘플로부터 핵산을 단리하는 단계; 제1 DNA 세그먼트 및 제2 DNA 세그먼트가 공통적인 단백질 DNA 복합체로부터 발생하는 것으로 확인되도록 단백질 DNA 복합체에 태그를 부착하는 단계; 공통적인 DNA 복합체로부터 제1 DNA 세그먼트 및 제2 DNA 세그먼트를 분리하는 단계; 제1 DNA 세그먼트 및 제2 DNA 세그먼트로부터 서열 정보를 생성하는 단계; 및 공통적인 단백질 DNA 복합체를 나타내는 태그 서열을 공유하는 서열 정보를 공통적인 게놈 구조에 할당하는 단계를 포함한다. 일부 경우에, 가교결합된 파라핀 포매된 조직은 핵산을 단리하기 전에 균질화되지 않는다. 일부 경우에, 태그 서열은 복합체를 확인하는 올리고 태그를 포함한다. 일부 경우에, 태그 서열은 제1 세그멘트를 제2 세그멘트에 라이게이션함으로써 생성된다. 일부 경우에, 단백질 DNA 복합체가 파괴되거나 붕괴되지 않도록 보존된 샘플, 예컨대 가교결합된 파라핀 포매된 조직 샘플로부터 핵산을 단리하는 것은 가교결합된 파라핀 포매된 조직 샘플을 크실렌에 접촉시키는 것을 포함한다. 일부 경우에, 단백질 DNA 복합체가 파괴되거나 붕괴되지 않도록 보존된 샘플, 예컨대 가교결합된 파라핀 포매된 조직 샘플로부터 핵산을 단리하는 것은 가교결합된 파라핀 포매된 조직 샘플을 에탄올에 접촉시키는 것을 포함한다. 일부 경우에, 단백질 DNA 복합체가 파괴되거나 붕괴되지 않도록 보존된 샘플, 예컨대 가교결합된 파라핀 포매된 조직 샘플로부터 핵산을 단리하는 것은 가교결합된 파라핀 포매된 조직 샘플을 비등 조건으로부터 보호하는 단계를 포함한다. 일부 경우에, 공통적인 DNA 복합체로부터 제1 DNA 세그먼트 및 제2 DNA 세그먼트를 분리하는 것은 프로테이나제 K 처리를 포함한다. 추출 과정은 선택적으로 추출 과정 동안 임의의 가교결합제의 첨가를 수반하지 않는다. 오히려, 샘플 보전에 따라

생성되는 복합체는 보존된 샘플 내의 핵산에 잠재적으로 해를 주는 가교결합에 대한 노출 횟수를 최소화하도록 결정된다. 대안으로, 핵산은 단리되고, 가교결합제는 핵산 단리 및 염색질 재어셈블리 후에만 첨가된다.

[0007] 보존된 샘플, 예컨대 가교결합된 파라핀 포매된 조직 샘플로부터 긴 거리 서열 정보, 예컨대 게놈의 구조적 정보를 얻는 방법이 본원에서 제공된다. 그러한 방법의 일부는 50 kb 초과의 핵산 단편이 회수되도록 가교결합된 파라핀 포매된 조직 샘플로부터 핵산을 단리하는 단계; 핵산 분자의 제1 DNA 세그먼트 및 제2 DNA 세그먼트가 그들의 공통적인 포스포디에스테르 백본과는 관계없이 함께 유지되도록 하는 적어도 하나의 복합체를 형성시키기 위해 핵산을 다수의 핵산 결합 모이어티와 접촉시키는 단계; 적어도 하나의 복합체의 적어도 하나의 포스포디에스테르 백본을 절단하는 단계; 제1 DNA 세그먼트 및 제2 DNA 세그먼트가 공통의 복합체로부터 발생하는 것으로 확인되도록 적어도 하나의 복합체에 태그를 부착하는 단계; 공통의 복합체로부터 제1 DNA 세그먼트 및 제2 DNA 세그먼트를 분리하는 단계; 제1 DNA 세그먼트 및 제2 DNA 세그먼트로부터 서열 정보를 생성하는 단계; 및 공통적인 단백질 DNA 복합체를 나타내는 태그 서열을 공유하는 서열 정보를 공통적인 게놈 구조에 할당하는 단계를 포함한다. 일부 경우에, 가교결합된 파라핀 포매된 조직 샘플은 핵산을 단리하기 전에 균질화되지 않는다. 일부 경우에, 태그 서열은 복합체를 확인하는 올리고 태그를 포함한다. 일부 경우에, 태그 서열은 제1 DNA 세그먼트를 제2 DNA 세그먼트에 라이게이션함으로써 생성된다. 일부 경우에, 50 kb 초과의 핵산 단편이 회수되도록 보존된 샘플, 예컨대 가교결합된 파라핀 포매된 조직 샘플로부터 핵산을 단리하는 것은 보존된 샘플, 예컨대 가교결합된 파라핀 포매된 조직 샘플을 안트라닐레이트 및 포스파닐레이트 중 적어도 하나에 접촉시키는 단계를 포함한다. 일부 경우에, 단리는 40℃ 이하의 온도에서 수행된다. 일부 경우에, 단리는 40℃ 이하의 온도에서 수행된다. 일부 경우에, 공통적인 DNA 복합체로부터 제1 DNA 세그먼트 및 제2 DNA 세그먼트를 분리하는 것은 프로테이나제 K 처리를 포함한다. 일부 경우에, 다수의 핵산 결합 모이어티는 핵 단백질을 포함한다. 일부 경우에, 다수의 핵산 결합 모이어티는 트랜스 포사제를 포함한다. 일부 경우에, 다수의 핵산 결합 모이어티는 히스톤을 포함한다. 일부 경우에, 다수의 핵산 결합 모이어티는 핵산 결합 단백질을 포함한다. 일부 경우에, 다수의 핵산 결합 모이어티는 나노 입자를 포함한다. 일부 경우에, 적어도 하나의 복합체의 적어도 하나의 포스포디에스테르 백본을 절단하는 단계는 제한 엔도뉴클레아제에 접촉시키는 것을 포함한다. 일부 경우에, 적어도 하나의 복합체의 적어도 하나의 포스포디에스테르 백본을 절단하는 단계는 비특이적 엔도뉴클레아제에 접촉시키는 것을 포함한다. 일부 경우에, 적어도 하나의 복합체의 적어도 하나의 포스포디에스테르 백본을 절단하는 단계는 DNA를 전단하는 것을 포함한다. 일부 경우에, 적어도 하나의 복합체의 적어도 하나의 포스포디에스테르 백본을 절단하는 단계는 트랜스포사제에 접촉시키는 것을 포함한다. 일부 경우에, 적어도 하나의 복합체의 적어도 하나의 포스포디에스테르 백본을 절단하는 단계는 토포이소머라제에 접촉시키는 것을 포함한다.

[0008] 보존된 조직 샘플로부터 공간적으로 분포된 게놈의 구조적 정보를 회수하는 방법이 본원에서 제공된다. 이러한 방법 중 일부는 조직 샘플을 얻는 단계; 상기 보존된 조직 샘플, 예컨대 고정된 3차원 파라핀 포매된 조직 샘플의 제1 위치로부터 일부를 추출하는 단계; 단백질 DNA 복합체가 파괴되거나 붕괴되지 않도록 제1 위치로부터의 일부로부터 핵산을 단리하는 단계; 제1 DNA 세그먼트 및 제2 DNA 세그먼트가 공통적인 단백질 DNA 복합체로부터 생성되는 것으로 확인되도록 단백질 DNA 복합체에 태그를 부착하는 단계; 공통적인 DNA 복합체로부터 제1 DNA 세그먼트 및 제2 DNA 세그먼트를 분리하는 단계; 제1 DNA 세그먼트 및 제2 DNA 세그먼트로부터 서열 정보를 생성하는 단계; 공통적인 단백질 DNA 복합체를 나타내는 태그 서열을 공유하는 서열 정보를 공통적인 게놈 구조에 할당하는 단계; 및 공통적인 게놈 구조를 보존된 조직 샘플의 제1 위치에 할당하는 단계를 포함한다. 일부 경우에, 보존된 조직 샘플은 핵산을 단리하기 전에 균질화되지 않는다. 일부 경우에, 조직 샘플은 고정된 3차원 파라핀 포매된 조직 샘플을 포함한다. 일부 경우에, 태그 서열은 복합체를 확인하는 올리고 태그를 포함한다. 일부 경우에, 태그 서열은 제1 세그멘트를 제2 세그멘트에 라이게이션함으로써 생성된다. 일부 경우에, 단백질 DNA 복합체가 파괴되거나 붕괴되지 않도록 가교결합된 파라핀 포매된 조직 샘플로부터 핵산을 단리하는 단계는 가교결합된 파라핀 포매된 조직 샘플을 크실렌에 접촉시키는 것을 포함한다. 일부 경우에, 단백질 DNA 복합체가 파괴되거나 붕괴되지 않도록 가교결합된 파라핀 포매된 조직 샘플로부터 핵산을 단리하는 단계는 가교결합된 파라핀 포매된 조직 샘플을 에탄올에 접촉시키는 것을 포함한다. 일부 경우에, 단백질 DNA 복합체가 파괴되거나 붕괴되지 않도록 가교결합된 파라핀 포매된 조직 샘플로부터 핵산을 단리하는 단계는 샘플을 비등 조건으로부터 보호하는 것을 포함한다. 일부 경우에, 제1 DNA 세그먼트 및 제2 DNA 세그먼트를 공통적인 DNA 복합체로부터 분리하는 단계는 프로테이나제 K 처리를 포함한다. 일부 경우에, 조직 샘플은 고정된 3차원 파라핀 포매된 조직 샘플을 포함한다.

[0009] 치료 요법 시험 결과를 재평가하는 방법이 본원에서 제공된다. 이러한 방법 중 일부는 환자 집단에서 치료 요법 결과에 관한 데이터를 얻는 단계; 상기 환자 집단의 다수의 환자로부터 보존된 조직 샘플, 예컨대 고정된 조직 샘플을 얻는 단계; 상기 고정된 조직 샘플로부터 핵산 복합체를 추출하는 단계; 다수의 상기 고정된 조직 샘플

에 대해 상기 핵산 복합체를 사용하여 게놈의 구조적 정보를 결정하는 단계; 및 치료 요법 결과에 관련된 게놈의 구조적 정보를 확인하기 위해 치료 요법 결과에 관한 데이터를 게놈의 구조적 정보에 서로 관련시키는 단계를 포함한다. 일부 경우에, 보존된 조직 샘플은 핵산을 추출하기 전에 균질화되지 않는다. 일부 경우에, 상기 고정된 조직 샘플로부터 핵산 복합체를 추출하는 단계; 및 다수의 상기 고정된 조직 샘플에 대해 상기 핵산 복합체를 사용하여 게놈의 구조적 정보를 결정하는 단계는 본원에서 개시되는 임의의 방법을 포함한다.

[0010] 뉴클레오티드 서열 어셈블리 방법이 본원에서 제공된다. 이러한 방법 중 일부는 고정된 조직 샘플을 제공하는 단계; 상기 고정된 조직 샘플로부터 가교결합된 DNA:단백질 복합체를 회수하는 단계; 상기 가교결합된 DNA:단백질 복합체로부터의 DNA의 제1 섹션을 상기 가교결합된 DNA:단백질 복합체로부터의 DNA의 제2 섹션에 라이게이션하여 라이게이션된 DNA를 형성하는 단계; 상기 가교결합된 DNA:단백질 복합체로부터 상기 라이게이션된 DNA를 추출하는 단계; 상기 라이게이션된 DNA를 시퀀싱하는 단계; 및 상기 시퀀싱으로부터의 정보를 이용하여 뉴클레오티드 서열을 어셈블리하는 단계를 포함한다. 일부 경우에, 상기 고정된 조직 샘플은 포르말린에 의해 고정된다. 일부 경우에, 고정된 조직 샘플은 핵산을 단리하기 전에 균질화되지 않는다. 일부 경우에, 상기 고정된 조직은 포르말린 고정, 파라핀 포매된(FFPE) 것이다. 일부 경우에, 상기 가교결합된 DNA:단백질 복합체는 염색질을 포함한다. 일부 경우에, 상기 라이게이션은 평활(blunt) 말단 라이게이션을 포함한다. 일부 경우에, 본원에서 개시되는 방법은 상기 라이게이션 전에 상기 가교결합된 DNA:단백질 복합체로부터 DNA를 소화시키는 단계를 추가로 포함한다. 일부 경우에, 상기 소화는 제한 효소 소화를 포함한다. 일부 경우에, 본원에서 개시되는 방법은 상기 소화 후, 상기 소화에 의해 생성된 점착성(sticky) 말단을 충전하여 평활 말단을 생성하는 단계를 추가로 포함한다. 일부 경우에, 상기 충전은 비오티닐화된 뉴클레오티드를 사용하여 수행된다. 일부 경우에, 상기 회수는 상기 가교결합된 DNA:단백질 복합체로부터의 DNA를 고체 지지체에 결합시키는 것을 포함한다. 일부 경우에, 상기 추출은 상기 가교결합된 DNA:단백질 복합체로부터 단백질을 소화시키는 것을 포함한다. 일부 경우에, 상기 정보는 2000개 염기쌍(bp) 초과의 거리에 걸친 장범위 정보(long-range information)를 포함한다. 일부 경우에, 상기 거리는 10,000 bp 초과이다. 일부 경우에, 상기 거리는 100,000 bp 초과이다. 일부 경우에, 상기 거리는 200,000 bp 초과이다. 일부 경우에, 본원에서 개시되는 방법은 상기 회수 전에 상기 고정된 조직 샘플의 포매 물질을 용해시키는 단계를 추가로 포함한다. 일부 경우에, 상기 포매 물질은 파라핀을 포함한다.

[0011] 조직 샘플 분석 방법이 본원에서 제공된다. 이러한 방법 중 일부는 고정된 조직 샘플을 제공하는 단계; 상기 고정된 조직 샘플의 제1 부분 및 상기 고정된 조직 샘플의 제2 부분을 수집하는 단계로서, 상기 제1 부분 및 상기 제2 부분은 상기 고정된 조직 샘플의 상이한 영역으로부터 유래되는 것인 단계; 상기 제1 부분으로부터 제1 가교결합된 DNA:단백질 복합체를, 및 상기 제2 부분으로부터 제2 가교결합된 DNA:단백질 복합체를 회수하는 단계; (i) 상기 제1 가교결합된 DNA:단백질 복합체로부터의 DNA의 제1 섹션을 상기 제1 가교결합된 DNA:단백질 복합체로부터의 DNA의 제2 섹션에 라이게이션하여 제1 라이게이션된 DNA를 형성하는 단계, 및 (ii) 상기 제2 가교결합된 DNA:단백질 복합체로부터의 DNA의 제2 섹션을 상기 제2 가교결합된 DNA:단백질 복합체로부터의 DNA의 제2 섹션에 라이게이션하여 제2 라이게이션된 DNA를 형성하는 단계; 상기 제1 가교결합된 DNA:단백질 복합체로부터 상기 제1 라이게이션된 DNA를 및 상기 제2 가교결합된 DNA:단백질 복합체로부터 상기 제2 라이게이션된 DNA를 추출하는 단계; 상기 제1 라이게이션된 DNA 및 상기 제2 라이게이션 DNA를 시퀀싱하는 단계; 및 상기 시퀀싱으로부터의 정보를 이용하여 제1 뉴클레오티드 서열 및 제2 뉴클레오티드 서열을 어셈블리하는 단계를 포함한다. 일부 경우에, 고정된 조직 샘플은 핵산을 단리하기 전에 균질화되지 않는다. 일부 경우에, 상기 고정된 조직 샘플은 포르말린에 의해 고정된다. 일부 경우에, 상기 고정된 조직은 포르말린 고정, 파라핀 포매된(FFPE) 조직이다. 일부 경우에, 상기 제1 가교결합된 DNA:단백질 복합체 및 상기 제2 가교결합된 DNA:단백질 복합체는 각각 염색질을 포함한다. 일부 경우에, (d)(i) 및 (d)(ii)에서의 상기 라이게이션은 평활 말단 라이게이션을 포함한다. 일부 경우에, 본원에서 개시되는 방법은 (d)(i) 및 (d)(ii)에서의 상기 라이게이션 전에, 상기 제1 가교결합된 DNA:단백질 복합체로부터의 DNA 및 상기 제2 가교결합된 DNA:단백질 복합체로부터의 DNA를 소화시키는 단계를 추가로 포함한다. 일부 경우에, 상기 소화는 제한 효소 소화를 포함한다. 일부 경우에, 본원에서 개시되는 방법은 상기 소화 후, 상기 소화에 의해 생성된 점착성 말단을 충전하여 평활 말단을 생성하는 단계를 추가로 포함한다. 일부 경우에, 상기 충전은 비오티닐화된 뉴클레오티드를 사용하여 수행된다. 일부 경우에, 상기 회수는 상기 제1 가교결합된 DNA:단백질 복합체로부터의 DNA 및 상기 제2 가교결합된 DNA:단백질 복합체로부터의 DNA를 고체 지지체에 결합시키는 것을 포함한다. 일부 경우에, 상기 추출은 상기 제1 가교결합된 DNA:단백질 복합체로부터 및 상기 제2 가교결합된 DNA:단백질 복합체로부터 단백질을 소화시키는 것을 포함한다. 일부 경우에, 상기 정보는 2000개 초과의 염기쌍(bp)의 거리에 걸친 장범위 정보를 포함한다. 일부 경우에, 상기 거리는 10,000 bp 초과이다. 일부 경우에, 상기 거리는 100,000 bp 초과이다. 일부 경우에, 상기 거리는 200,000 bp 초과이다. 일부 경우에, 본원에서 개시되는 방법은 상기 회수 전에 상기 고정된 조직 샘플의 포매 물질을 용해

시키는 단계를 추가로 포함한다. 일부 경우에, 상기 포매 물질은 파라핀을 포함한다.

[0012] 또한, 보존된 샘플로부터 게놈의 구조적 정보를 얻기 위한 키트가 본원에서 제공된다. 이러한 키트 중 일부는 완충제, DNA 결합제, 친화성 태그 결합제, 데옥시뉴클레오티드, 태그 부착된 데옥시뉴클레오티드, DNA 단편화제, 말단 수복 효소, 리가제, 단백질 제거제 및 보존된 샘플로부터 게놈의 구조적 정보를 수득할 때 사용하기 위한 사용 설명서를 포함한다. 선택적으로, 키트는 PCR용 시약 또는 PCR 시약과 조합하여 키트를 사용하기 위한 사용 설명서를 추가로 포함한다. 일부 경우에, PCR용 시약은 완충제, 뉴클레오티드, 정방향 프라이머, 역방향 프라이머 및 열안정성 DNA 폴리머라제를 포함한다. 다양한 완충제는 제한 소화 완충제, 말단 수복 완충제, 라이게이션 완충제, TE 완충제, 세척 완충제, TWB 용액, NTB 용액, LWB 용액, NWB 용액 및 가교결합 반전(crosslink reversal) 완충제 중 적어도 하나를 포함한다. 일부 경우에, 제한 소화 완충제는 DpnII 완충제를 포함한다. 예를 들어, 말단 수복 완충제는 종종 NEB 완충제 2를 포함한다. 라이게이션 완충제는 종종 T4 DNA 리가제 완충제, BSA 및 트리톤 X-100을 포함한다. TE 완충제는 종종 트리스 및 EDTA를 포함한다. 일부 경우에, 세척 완충제는 트리스 및 염화나트륨을 포함한다. 일부 경우에, TWB 용액은 트리스, EDTA 및 트윈(Tween) 20을 포함한다. 일부 경우에, NTB 용액은 트리스, EDTA 및 염화나트륨을 포함한다. 일부 경우에, LWB 용액은 트리스, 염화리튬, EDTA 및 트윈 20을 포함한다. 일부 경우에, NWB 용액은 트리스, 염화나트륨, EDTA 및 트윈 20을 포함한다. 일부 경우에, 가교결합 반전 완충제는 트리스, SDS, 및 염화칼슘을 포함한다. 일부 경우에, DNA 결합제는 염색질 포획 비드를 포함한다. 일부 경우에, 염색질 포획 비드는 PEG-800 분말, 트리스 완충제, 염화나트륨, EDTA, 계면활성제, TE 완충제 및 세라-맥(sera-mag) 비드를 포함한다. 일부 경우에, 친화성 태그 결합제는 스트렙타비딘 비드를 포함한다. 일부 경우에, 스트렙타비딘 비드는 디나비드를 포함한다. 일부 경우에, 데옥시뉴클레오티드는 dATP, dTTP, dGTP 및 dCTP 중 적어도 3개를 포함한다. 일부 경우에, 비오티닐화된 데옥시뉴클레오티드는 비오티닐화된 dCTP, 비오티닐화된 dATP, 비오티닐화 dTTP 및 비오티닐화된 dGTP 중 적어도 하나를 포함한다. 일부 경우에, DNA 단편화제는 제한 효소, 트랜스포사제, 뉴클레아제, 초음파 처리 장치, 유체역학적 전단 장치 및 2가 금속 양이온 중 적어도 하나이다. 일부 경우에, 제한 효소는 DpnII를 포함한다. 일부 경우에, 말단 수복 효소는 T4 DNA 폴리머라제, 클레나우(klenow) DNA 폴리머라제 및 T4 폴리뉴클레오티드 키나제 중 적어도 하나를 포함한다. 일부 경우에, 리가제는 T4 DNA 리가제를 포함한다. 일부 경우에, 단백질 제거제는 프로테아제 및 페놀 중 적어도 하나를 포함한다. 일부 경우에, 프로테아제는 프로테이나제 K, 스트렙토마이세스 그리세우스(*Streptomyces griseus*) 프로테아제, 세린 프로테아제, 시스테인 프로테아제, 트레오닌 프로테아제, 아스파르트산 프로테아제, 글루탐산 프로테아제, 메탈로프로테아제 및 아스파라긴 펩티드 리아제 중 적어도 하나를 포함한다. 일부 경우에, 키트는 선택적으로 포매 물질을 제거하기 위한 용매를 포함한다. 일부 경우에, 용매는 크실렌, 벤젠 및 톨루엔 중 적어도 하나이다. 본원에서 열거되는 키트 성분 및 실질적으로 동등한 그의 변이체를 감안하여, 적어도 하나의 상업적으로 이용 가능한 키트 성분이 배제되고 독립적으로 수득되는 시약과 조합하여 나머지 성분을 성공적으로 사용하기 위한 사용 설명서에 의해 대체되는 대체 키트가 고려된다.

[0013] [참조에 의한 통합]

[0014] 본 명세서에서 언급되는 모든 간행물, 특허 및 특허 출원은 마치 각각의 개별 간행물, 특허 또는 특허 출원이 구체적이고 개별적으로 참고로 포함된다고 지시된 것과 동일한 정도로 본원에 참고로 포함된다. 본 명세서에서 언급되는 모든 간행물, 특허 및 특허 출원 및 이들에서 인용되는 임의의 참고문헌은 그 전부가 본원에 참고로 포함된다.

도면의 간단한 설명

[0015] 도 1a는 포르말린 고정, 파라핀 포매된(FFPE) 조직 샘플의 예시적인 개략도를 도시한 것이다.

도 1b는 염색질 기반 차세대 시퀀싱(NGS) 라이브러리 제조를 위한 프로토콜의 예시적인 개략도를 도시한 것이다.

도 2a 및 도 2b는 상호 전좌(reciprocal translocation)를 발견하기 위해 사용될 수 있는 예시적인 단순 커널(simple kernel)을 도시한 것이다.

도 3은 ETV6과 NTRK3 사이의 상호 전좌의 신호를 갖는 이미지를 도시한 것이다.

도 4a, 도 4b, 및 도 4c는 3개의 상이한 샘플에서 비교된 동일한 염색체 쌍에서의 이미지 분석 기반 결과를 도시한 것이다.

도 5a, 도 5b, 및 도 5c는 1번 염색체 대 7번 염색체(도 5a), 2번 염색체 대 5번 염색체(도 5b), 및 1번 염색체

대 1번 염색체(도 5c)에 대한 중간의 표준화된 리드 밀도(10개의 샘플에 걸친)를 도시한 것이다.

도 6a 및 도 6b는 다양한 빈(bin) 취급 방법을 도시한 것이다. 도 6a는 동일한 빈 크기를 보여주고, 도 6b는 빈 내삽(interpolation)을 보여준다.

도 7은 전체 게놈 스캐닝 분석 파이프라인에 의한 분석을 도시한 것이다.

도 8a 및 도 8b는 FFPE 기반 '시카고(Chicago)' 리드쌍 라이브러리(도 8a) 및 전통적인 '시카고' 기반 리드쌍 라이브러리(도 8b)로부터 유도된 리드쌍 거리 빈도 데이터를 도시한 것이다.

도 9a 및 도 9b는 리드쌍의 GRCh38 참조 서열 상의 매핑된 위치가 GM12878과 참조 서열 사이의 구조적 차이 근처에 플로팅됨을 보여준다. 도 9a는 인접하는 20 kb의 반복 영역을 갖는 80 kb 역위에 대한 데이터를 도시한 것이다. 도 9b는 페이징된 이형접합성 결실에 대한 데이터를 도시한 것이다.

도 10은 본원에서 제공되는 방법을 구현하도록 프로그램되거나 달리 구성되는 예시적인 컴퓨터 시스템을 보여준다.

도 11a는 본 개시내용의 방법에 의한 FFPE 조직 및 FFPE 세포 배양 샘플의 분석 결과를, Hi-C에 의해 분석된 세포 배양과 비교하여 보여준다.

도 11b, 도 11c, 및 도 11d는 장범위 게놈 링키지 데이터를 생성하는 아슈케나지(Ashkenazi) 부계(GM24149) 세포 배양 FFPE 샘플의 분석 결과를 보여준다.

발명을 실시하기 위한 구체적인 내용

[0016] 보존된 샘플, 예컨대 포르말린 고정 파라핀 포매(FFPE) 조직 샘플에 많은 생물학적 정보의 저장소가 보관되며, 이러한 샘플은 환자로부터 이환된 또는 손상된 조직을 절제하기 위한 수술과 같은 수술 동안 통상적으로 얻어진다. 그러나, 그러한 샘플을 보존하는 동안 발생하는 가교결합은 이들 샘플로부터의 DNA 추출을 방해하는 것으로 생각되었다. 보존 및 보관은 기술적으로 간단하고 경제적이며, 결과적으로 매우 많은 환자 샘플이 상기 방법을 사용하여 보관되었다. 그 결과, 예를 들어 암 치료 시험을 받고 있는 환자의 종양 조직으로부터 샘플을 얻고 보존하는 것은 오랫동안 일상적으로 시행되고 있다.

[0017] 최근까지 이들 샘플은 구조적 정보에 접근하는 데만 유용하였다. 3차원 조직 절편은 잘 보존되어 형태학적 분석에 이용 가능하지만, 조직 보존 과정은 보존된 샘플로부터 게놈 수준의 정보에 접근하는 것을 방해하였다. 예를 들어, 도 1a는 보존된 샘플(예를 들어, FFPE 샘플)의 예시적인 개략도를 도시한 것이다. 세포(101)는 그의 3차원 분포가 보존되도록 고정된 샘플의 조직(102) 내에 공간적으로 분포된 상태로 도시되어 있다. 핵산(103)은 세포 내에 존재한다.

[0018] 이들 샘플로부터 핵산 정보를 얻기 위한 노력이 이루어졌지만, 수득된 핵산은 짧고 고도로 분해되어, 단지 국소 서열 정보만이 얻어질 수 있다. 따라서, 재배열에 관한 게놈 수준의 정보는 쉽게 얻을 수 없다. 재배열은 결실, 중복, 삽입, 역위 또는 반전(reversal), 전좌, 연결, 융합 및 분열을 포함할 수 있고, 이로 제한되지 않는다.

[0019] 많은 공지된 장애에서, 질환에 관련된 것은 이러한 게놈 규모의 재배열이다. 유전자 융합, 특히 게놈 재배열로 인한 융합은 일부 암에서 특히 흔하며, 종종 요법에 대한 반응으로 질환의 결과를 나타낸다. 일반적으로, 이러한 재배열 패턴은 보존된 샘플에서 하나의 또는 또 다른 형태학적 구조와 신뢰할 수 있는 상관관계가 없다. 오히려, 이들은 직접 유전자형이 결정되어야 한다. 그 결과, 이 정보는 종양 샘플 자체가 보존되고 화학요법 또는 다른 요법에 대한 종양의 반응에 관한 데이터가 쉽게 이용 가능함에도 불구하고 이용할 수 없었다.

[0020] 본원에서의 방법 및 조성물은 보존된 샘플, 예컨대 상기 고려된 샘플로부터 게놈의 구조적 정보를 결정하는 것에 관련된다. 본원에서 일부 방법은 보존된 샘플에 함유된 게놈의 구조적 정보에 접근하기 위해 추출 방법을 이용하는 방법에 의존한다. 복합체가 파괴되거나 붕괴되지 않도록 단백질 DNA 복합체는 샘플로부터 추출되고, 핵산의 제1 세그먼트 및 제2 세그먼트가 그의 포스포디에스테르 백본과는 관계없이 함께 유지된다는 사실을 이용한다. 세그먼트는 올리고뉴클레오티드를 사용하거나 세그먼트를 서로 라이게이션하여 태그가 부착되고, 서열 정보가 매핑되는 콘티그를 공통의 스캐폴드에 할당할 수 있도록 하는 서열 정보가 얻어진다. 라이게이션된 세그먼트를 평가함으로써 생성된 리드쌍의 빈도 및 유형을 평가함으로써, 물리적 링키지 또는 페이즈 정보를 모두 추론하고, 질환에 관련된 구조적 재배열과 같은 특정 게놈의 구조적 재배열의 존재를 결정할 수 있다.

[0021] 또한, 이들 샘플에는 보존된 조직의 3차원 입체배열이 보존되어 있다. 암성 종양은 일반적으로 그들의 게놈 구조에 대해 이질적이다. 종양은 DNA 복구 결함, 세포 사멸 억제, 종양 성장 및 전이와 관련된 별개의 돌연변이로

특징지어진다. 종양은 일반적으로 돌연변이의 다양한 조합을 갖고 다양한 정도의 건강에 대한 위험을 갖는 다중 세포 하위 집단을 포함한다. 종종, 이러한 위험은 국소 형태와 상관관계가 있다. 종양 세포 집단은 휴지 상태로부터 양성의 국소적으로 복제하는 세포 집단, 상대적으로 높은 건강 위험을 나타내는 전이 세포 모집단에 이르기까지 다양하다. 따라서, 종양에서 일반적으로 제시된 게놈 구조의 존재뿐만 아니라 종양 샘플 내의 공간적으로 분리된 하위 집단의 국소 게놈 구조의 존재를 확인하는 것은 이전의 약물 치료의 상대적인 효능을 평가하려고 시도하거나 또는 위험이 알려지지 않은 종양을 제시하는 환자에 대한 적절한 약물을 선택하려고 시도하는 연구자 및 의사에게 가치있는 것이다. 특히, 게놈 구조를 종양 내 위치 및 종양 내의 알려진 세포 형태에 서로 관련시키는 것은 어떤 게놈 구조가 가장 위험한 종양 위치 및 국소 세포 형태와 가장 밀접하게 상응하는지를 결정하는 데 중요하다.

[0022] 관련 기술 분야의 방법을 사용하여 보존된 샘플, 예컨대 FFPE 샘플로부터 추출된 DNA는 종종 길이가 300개 염기쌍 미만인 것으로 생각된다. 보존(예를 들어, FFPE) 과정 및 후속 탈수 및 장기간의 보관 동안 일부 닉 형성(nicking) 또는 손상이 발생할 수 있다. 전형적으로 하룻밤 동안의 프로테이나제 K 처리 및 가교결합을 파괴하고 DNA를 방출시키기 위한 후속적인 비등을 수반하는 추출 과정 동안 상당한 양의 단편화가 또한 발생할 수 있다. 그럼에도 불구하고, 본원의 방법을 통해, 그러한 핵산 분자는 DNA 단백질 복합체의 파괴 또는 붕괴 없이 절제된 DNA 단백질 복합체에 보존된 구조적 정보와 함께 게놈의 구조적 재배열에 관한 정보를 생성한다.

[0023] **천연 및 재구성된 염색질**

[0024] 보존된 샘플은 종종 천연 또는 재구성된 염색질을 포함하거나, 가교결합제와 접촉하기 직전에 제1 세그먼트 및 제2 세그먼트가 이들의 공통적인 포스포디에스테르 백본과는 관계없이 함께 유지되도록 단백질 또는 비단백질 스캐폴드에 다수의 지점에서 결합된 핵산을 갖는다. 진핵생물에서, 게놈 DNA가 핵 내에서 염색체로서 염색질에 채워진다. 진핵생물의 천연 염색질의 기본적인 구조 단위는 뉴클레오솜이이고, 이것은 히스톤 팔량체 주위를 감싸는 146개 염기쌍(bp)의 DNA로 이루어진다. 히스톤 팔량체는 각각 코어 히스톤 H2A-H2B 이량체 및 H3-H4 이량체의 2개의 카피로 이루어진다. 뉴클레오솜은 DNA에 따라 규칙적으로 간격을 두고 위치하고, 일반적으로 "염주 모양(beads on a string)"으로 언급된다.

[0025] 코어 히스톤 및 DNA의 뉴클레오솜으로의 어셈블리는 샤페론(chaperone) 단백질 및 관련 어셈블리 인자에 의해 매개된다. 거의 모든 이들 인자가 코어 결합 단백질이다. 뉴클레오솜 어셈블리 단백질-1(NAP-1)과 같은 히스톤 샤페론의 일부는 히스톤 H3 및 H4에 대한 결합의 선호를 나타낸다. 또한, 새로 합성된 히스톤은 아세틸화된 후, 염색질로 어셈블리된 후 탈아세틸화되는 것으로 관찰되었다. 따라서, 히스톤 아세틸화 또는 탈아세틸화를 매개하는 인자는 염색질 어셈블리 과정에서 중요한 역할을 수행한다.

[0026] 일반적으로, 염색질을 재구성하거나 어셈블리하기 위한 2가지의 시험관 내 방법이 개발되었다. 한 가지 방법은 ATP 독립적인 반면, 다른 방법은 ATP 의존적이다. 염색질을 재구성하기 위한 ATP 독립적 방법은 DNA 및 코어 히스톤 + NAP-1과 같은 단백질 또는 히스톤 샤페론으로 기능하는 염을 포함한다. 이 방법은 세포의 천연 코어 뉴클레오솜 입자를 정확하게 모방하지 않는 DNA 상의 히스톤의 무작위 배열을 초래한다. 이들 입자는 규칙적으로 배열되지 않고 연장된 뉴클레오솜 어레이가 아니고 사용된 DNA 서열이 대체로 250 bp보다 길지 않기 때문에, 종종 모노뉴클레오솜으로 언급된다(Kundu, T. K. *et al.*, Mol. Cell 6: 551-561, 2000). 보다 긴 DNA 서열 상에 규칙적으로 배열된 뉴클레오솜의 연장된 어레이를 생성하기 위해, 염기쌍은 ATP 의존적 과정을 통해 어셈블리되어야 한다.

[0027] 천연 염색질에서 관찰되는 것과 유사한 주기적인 뉴클레오솜 어레이의 ATP 의존적 어셈블리는 DNA 서열, 코어 히스톤 입자, 샤페론 단백질 및 ATP 이용 염색질 어셈블리 인자를 필요로 한다. ACF(ATP 이용 염색질 어셈블리 및 리모델링 인자) 또는 RSF(리모델링 및 스페이싱 인자)는 시험관 내에서 염색질 내로의 뉴클레오솜의 연장된 규칙적으로 배열된 어레이를 생성하기 위해 사용되는 2개의 광범위하게 연구된 어셈블리 인자이다([Fyodorov, D.V., and Kadonaga, J.T. Method Enzymol. 371: 499-515, 2003]; [Kundu, T. K. *et al.* Mol. Cell 6: 551-561, 2000]).

[0028] 특정 실시양태에서, 본 개시내용의 방법은 예를 들어 혈장, 혈청 및/또는 소변으로부터 단리된 자유 DNA; 세포 및/또는 조직으로부터의 아폽토시스 DNA; 시험관 내에서 효소에 의해(예를 들어, DNase I, 트랜스포사제 및/또는 제한 엔도뉴클레아제에 의해) 단편화된 DNA; 및/또는 기계적 힘(유체-전단(hydro-shear), 초음파 처리, 분무 등)에 의해 단편화된 DNA를 포함하고 이로 제한되지 않는 임의의 유형의 단편화된 이중 가닥 DNA에 용이하게 적용될 수 있다.

[0029] 재구성된 염색질은 뉴클레오솜 또는 심지어 단백질을 포함할 필요가 없다. 오히려, 광범위하게 규정된 재구성된 염색질은 제1 세그먼트 및 제2 세그먼트가 이들의 포스포디에스테르 백본과는 관계없이 함께 유지되도록 결합된 적어도 하나의 핵산을 포함한다. 많은 핵산 결합 모이어티가 염색질 재구성에 적합하다. 그 예는 개별적으로 또는 뉴클레오솜 내에 어셈블리된 히스톤과 같은 핵 단백질뿐만 아니라, 전사 인자, 트랜스포존 또는 핵산 결합 활성을 갖는 임의의 다른 단백질과 같은 다른 핵산 결합 단백질을 포함한다. 비핵 단백질, 예컨대 소기관(organellar) 핵산 결합 단백질이 또한 고려된다. 비단백질 모이어티, 예컨대 나노 입자 또는 핵산 결합 표면도 고려된다.

[0030] **보존된 추출 핵산의 DNA 연결성 정보의 보존**

[0031] 포르말린 고정, 파라핀 포매된 샘플과 같은 보존된 샘플은 종종 고정 및/또는 포매 물질에 의해 야기되는 손상과 같은 손상을 갖는 핵산을 포함한다. DNA를 이용할 때 관련 성분은 DNA 손상제에 적용된 단리된 DNA의 DNA 물리적 링키지 정보의 완전성을 보존한다. DNA는 비교적 안정한 분자이지만, DNA의 완전성은 환경적 요소 및 특히 시간의 영향을 받는다. 뉴클레아제 오염, 가수분해, 산화, 화학적, 물리적 및 기계적 손상의 존재는 DNA 보존에 대한 주요 위협 중 일부를 나타낸다. 수송 중에 DNA가 직면하는 기계적, 환경적 및 물리적 인자는 종종 그 단편을 남기고, 게놈 분석에 중요한 장범위 정보를 상실할 가능성이 있다. DNA 정보를 보존하는 기존의 방법은 대부분 DNA의 붕괴를 지연시키지만, 시간이 지남에 따라, 특히 단편화가 발생하는 경우 DNA 손상에 대한 보호를 거의 제공하지 않는다. 많은 경우에, 이러한 DNA 손상은 장기간의 보관이 의도되는 샘플을 고정하고 포매함으로써 완화할 수 있다. 예를 들어, FFPE(포르말린 고정, 파라핀 포매된) 샘플은 오랫동안 보존될 수 있다. 그러나, 보존 과정은 DNA 손상을 초래할 수 있다. 또한, 후속 DNA 추출 방법은 종종 거칠고, 추가의 DNA 손상 및 단편화를 초래한다.

[0032] DNA 복합체 또는 염색질 응집체, 예컨대 보존된(예를 들어, FFPE) 샘플(조직 기반 보존 샘플 및 세포 배양 기반 보존 샘플 포함)에서 보관된 가교결합된 염색질 내의 핵산 분자와 같은 보존 및/또는 보관된 핵산 분자로부터 장거리 게놈 정보를 회수하는 것에 관련된 방법, 조성물 및 키트가 본원에서 개시된다. 특히, 방법, 조성물, 시스템 및 키트는 핵산의 물리적 링키지 정보가 보존되도록 상기 보존된 샘플로부터 핵산 샘플의 회수에 관련된다. 물리적 링키지 정보는 FFPE 추출 과정에서 핵산 자체를 보존함으로써, 또는 추출 과정에서 핵산 자체에 발생할 수 있는 임의의 손상과 상관없이 물리적 링키지 정보가 보존되도록 핵산 복합체를 보존함으로써 보존된다.

[0033] 종종, DNA 보관 중에 또는 FFPE 샘플과 같은 보존된 샘플로부터 DNA를 추출하는 동안 이중 가닥 파단이 발생하여 물리적 링키지 정보의 상실을 초래한다. 물리적 링키지 정보의 상실은 특히 해로운데, 그 이유는 그 상실이, 2배체 유기체 샘플에서 공통적인 유전자좌에 매핑되는 돌연변이가 실제로 동일한 대립유전자에 존재하거나 또는 2배체 게놈의 상이한 가닥 상에 위치하는 2개의 별개의 상동성 대립유전자 상에 존재하는지의 여부를 서열 어셈블러가 결정하는 것을 방해하기 때문이다. 게놈 정보는 개인 맞춤형 의약 또는 보다 많은 의학적 또는 치료 목적으로 사용되기 때문에, 어셈블리된 콘티그 서열에 물리적 링키지 정보를 할당하는 것이 점차 중요해지고 있다.

[0034] 전세계적인 장기간의 역사적 또는 대규모의 게놈 연구에 대한 프로그램의 확대와 함께 게놈학 기술이 발전함에 따라, 상기 DNA의 완전성 요건은 문제가 되고 있다. 그러한 연구는 현재 인간 집단 및 개인의 게놈 및 인간 건강에 미치는 그의 영향을 이해하고 훨씬 더 강력한 기술로 미래의 연구를 위해 현재의 게놈을 보존하는 데 절대적으로 필요하다. 후자의 관심은 또한 법의학적 이익과도 겹치기 때문에, 추후 분석 및 확인을 위해 DNA 샘플을 무기한 보관할 필요가 있다.

[0035] **물리적 링키지의 보존**

[0036] 포르말린 고정, 파라핀 포매된 샘플과 같은 보존된 샘플은 종종 보존된 샘플로부터의 핵산의 물리적 링키지 정보를 결정하는 데 어려움을 제시한다. 많은 하류 분석이 샘플로부터 물리적인 링키지 정보를 얻기 위해 사용될 수 있으며, 따라서 이들 분석은 FFPE 샘플 DNA 추출 동안 상기 정보의 상실로 인해 손상되거나 복잡해진다. 핵산 샘플은 관심 영역에 인접하여 어닐링하는 것으로 알려진 프라이머를 사용하는 폴리머라제 연쇄 반응("PCR")을 통한 큰 단편의 증폭을 위한 주형으로서 종종 의도된다. PCR은 그로부터 다수의 앰플리콘 핵산 분자가 생성되는 주형의 존재에 의존한다. 증폭은 단일 분자 상에서 서로 물리적으로 연결된 2개의 어닐링 부위(또는 한 어닐링 부위 및 제2 어닐링 부위의 역보체)에 의존한다. 따라서, 프라이머 어닐링 부위 사이의 물리적 링키지의 상실은 PCR 증폭을 포함하는 분석을 복잡하게 만든다.

[0037] 유사하게, 복제, 증폭, 발현 또는 유전자 도입에 의해 조작될 수 있도록 단편을 세포 숙주에 클로닝하는 것은 출발 물질로서 단일 분자를 보유함으로써 크게 용이해진다. 단편에 대한 물리적 링키지의 상실(즉, 단편의 절단)은 클로닝을 복잡하게 만들고, 단편 어셈블리에 다수의 단계를 추가로 필요하게 한다.

[0038] 대안으로, 일부 분석 방법은 물리적 근접성의 보존을 필요로 하지만, 핵산의 제1 세그먼트 및 제2 세그먼트가 그들의 포스포디에스테르 백본에 의해 물리적으로 연결된 채로 유지될 것을 필요로 하지 않는다. 예를 들어, 프로브가 비분해 샘플의 공통적인 분자 상에 존재하는지의 여부를 결정하기 위해 프로브가 제1 핵산 세그먼트 및 제2 핵산 세그먼트의 공존을 분석할 수 있다. 물리적 링키지의 보존은 이러한 분석을 용이하게 하지만, 상기 분석에 필요한 것은 아니다. 예를 들어, 제1 세그먼트 및 세그멘트가 이들의 공통적인 포스포디에스테르 백본과는 관계없이 결합되도록 재구성된 염색질 복합체로 분자를 어셈블리하면, 상기 분석이 유사하게 용이해진다. 이들의 공통적인 포스포디에스테르 백본이 절단되는 경우에도, 제1 세그먼트 및 제2 세그먼트에 대한 물리적 근접성 정보는 제1 및 제2 프로브를 사용한 복합체의 프로빙이 제1 세그먼트 및 제2 세그먼트가 본래의 샘플 내의 공통적인 분자 상에 존재하는지의 여부를 나타내도록 보존된다.

[0039] 시퀀싱은 물리적 링키지 정보의 보존으로부터 이익을 얻지만 물리적 링키지 또는 심지어 물리적 근접성의 보존을 필요로 하지 않는 또 다른 분석이다. 물리적 링키지의 보존은 시퀀싱을 용이하게 하지만, 본원에 개시되고 관련 기술 분야의 통상의 기술자에게 알려진 다른 방법도 시퀀싱을 용이하게 한다. 예를 들어, 물리적 근접성을 보존하면, 근접하여 유지되는 단편이 물리적 링키지 정보를 전달하기 위해 쉽게 말단 표지되기 때문에 시퀀싱이 용이해진다. 노출된 내부 말단은 인접한 단편 서열을 공통적인 분자에 매핑될 수 있도록 하는 올리고뉴클레오티드 태그를 사용하여 표지된다. 대안으로 또는 조합하여, 노출된 말단은 표시된 라이게이션 이벤트의 측면 상의 서열이 공통적인 분자에 매핑되는 리드쌍을 생성하기 위해 무작위로 서로 라이게이션된다. 물리적 근접성이 없는 경우에도, 물리적 근접성 정보가 상실되기 전에 물리적 근접성 마커를 추가하도록 핵산 샘플을 처리하면, 서열 분석이 용이해진다. 상기 처리는 분자의 세그먼트 사이의 물리적 링키지를 손상시키거나 상기 링키지의 상실을 초래할 수 있는 분해에 샘플을 적용하기 전에 수행되는 경우, 핵산 분자 상에 재구성된 염색질의 어셈블리, 내부 이중 가닥 말단의 노출 및 통상적인 뉴클레오티드를 사용한 교차 라이게이션 또는 태그 부착을 통한 상기 노출된 말단의 표지이다.

[0040] 이러한 모든 이유 때문에, 보존된(예를 들어, FFPE) 샘플로부터 DNA에 의해 코딩되는 물리적 링키지 정보를 추출하기 위한 간단하고 적당한 기술이 이 분야에서 매우 필요하게 되었다. 본원에서 개시되는 방법은 법의학, 농업, 환경 연구, 재생 가능 에너지, 역학 또는 질환 발생 반응, 및 종 보존을 포함하는 많은 분야에서 유용하다. 본 개시내용의 기술은 종양 샘플과 같은 조직 샘플의 이질성을 매핑하기 위해 사용된다. 예를 들어, 조직 블록은 그의 부피 전체에 걸쳐 샘플이 추출될 수 있고, 본 개시내용의 기술은 샘플을 분석하기 위해 사용될 수 있어, 조직 부피 전체에 걸친 변화를 비교할 수 있다. 감염은 또한 조직 부피 전체에서 분석될 수 있다. 본 개시내용의 기술은 임상적으로 중요한 영역의 페이징, 구조적 변이체의 분석, 카피 수 변이체의 분석, 위유전자(예를 들어, STRC)의 분석, 암에서 약물 적합성(druggable) 구조적 변이체에 대한 표적화된 패널, 및 기타 용도에 사용될 수 있다.

[0041] 본원에서 개시되는 방법의 일부 실시양태에서, 샘플 채취(예를 들어, FFPE 샘플로부터의 추출) 동안 물리적 링키지 정보 및/또는 물리적 링키지 정보의 상실은 핵산 파손을 물리적으로 방지하거나 줄임으로써 방지되거나 감소된다. 페이즈 정보 및/또는 물리적 링키지 정보의 상실은 그들의 포스포디에스테르 백본과는 관계없이 제1 세그먼트 및 제2 세그먼트를 물리적으로 근접하게 유지함으로써 방지되거나 감소된다. 대안으로 또는 조합하여, 페이즈 정보 및/또는 물리적 링키지 정보의 상실은 물리적 근접성 정보의 상실 및 공통적인 포스포디에스테르 백본 테더(tether)의 상실이있을 때 제1 세그먼트 및 제2 세그먼트에 부착되는 시퀀싱 태그 정보가 2개의 서열을 원래의 비분해된 샘플에서 공통적인 페이즈 또는 공통적인 분자를 공유하는 것으로 확인하기에 충분하도록 공통적인 또는 상호 보완적인 태그를 사용하여 제1 세그먼트 및 제2 세그먼트에 표지함으로써 방지되거나 감소된다. 추가로 또는 대안으로, 표지는 제1 세그먼트를 제2 세그먼트에 라이게이션함으로써 이루어지고, 여기서 제1 세그먼트 및 제2 세그먼트는 동일한 원래의 DNA 분자 상에서 물리적으로 연결되어 있지만, 제2 세그먼트는 제1 세그먼트에 인접하여 존재하는 것은 아니다.

[0042] 핵산 분해는 많은 다양한 원인으로부터 발생한다. 본원에서 고려되는 것은 다수의 유형의 DNA 분해, 특히 핵산 샘플 내의 원래의 공급원 분자 상의 제1 세그먼트와 제2 세그먼트 사이의 물리적 링키지의 상실을 초래하는 것과 같은 이중 가닥 파단의 도입을 유발하는 DNA 분해로부터의 보호이다. 특히 중요한 것은 보관된 핵산 샘플에서 시간이 지남에 따라 발생하거나 실온에서 보관된 샘플에서 발생하는 것과 같은 비효소적 DNA 분해이다. 비효소적 핵산 분해는 비등, 프로테이나제 처리, UV 조사, 산화, 가수분해, 물리적 스트레스, 예컨대 전단 또는 꼬

임, 또는 분자가 절단되거나 또는 올가미(lariat)가 형성되도록 핵산 분자의 내부 결합 상에 자유 3' 히드록실기에 의한 친핵성 부착을 포함한다. 또한, 효소 활성, 예컨대 비특이적 엔도뉴클레아제 활성, 단일 가닥의 닉 형성 또는 이중 가닥 파단을 수반하는 토포이소머라제 활성, 제한 엔도뉴클레아제 활성, 트랜스포사제 활성, DNA 미스매치 수복 또는 염기 절제, 또는 핵산 손상, 예컨대 페이즈 정보의 상실 및/또는 물리적 링키지 정보의 상실을 초래하는 다른 효소 활성이 본원에서 고려된다. 효소 분해는 불완전한 핵산 단리, 또는 먼 위치 또는 예를 들어 과학적 자원에 대한 전염병 또는 다른 부담으로 인해 멸균 상태가 쉽게 또는 일정하게 얻어지지 않는 위치와 같은 '현장에서의" 수집 동안 마주치게 될 수 있는 것과 같은 비멸균 환경에서의 초기 단리로 인한 경우와 같이 일부 경우에 외인성이다.

[0043] 본원에서 일부 실시양태는 핵산 분자의 제1 세그먼트를 핵산 분자의 제2 세그먼트에 관련시키는 물리적 링키지 정보가 제1 핵산 분자와 제2 핵산 분자 사이에서 이중 가닥 파단이 발생하는 경우에 상실되지 않도록, 보존된 (예를 들어, FFPE) 샘플로부터 추출된 핵산과 같은 부분적으로 또는 완전히 단리된 핵산 상에 염색질을 시험관내에서 어셈블리하는 것에 관한 것이다. 재어셈블리된 염색질은 일부 경우에 또 다른 공급원으로부터 제공된 핵산 결합 단백질을 포함한다. 대안으로, 일부 경우에, 불완전하게 단리된 핵산 샘플, 예컨대 그의 천연 염색질 입체배열을 파괴 또는 붕괴시키기 위해, 천연 뉴클레아제 활성을 불활성화하기 위해, 또는 천연 염색질을 파괴 또는 붕괴시키고 천연 뉴클레아제 활성을 불활성화하기 위해 처리된 핵산 샘플은 샘플 내의 핵산을 안정화하기 위해 가교결합제와 접촉된다. 다른 경우에, 보존된 샘플로부터의 핵산은 샘플에 보존된 천연 염색질 구조를 사용하여 분석된다.

[0044] 이중 가닥 파단은 시간이 지남에 따라 DNA 보관 동안 종종 발생한다. 따라서, 변이체가 긴 거리에 걸쳐 하플로타입과 확실하게 연결될 수 없기 때문에 DNA 분자의 페이징 정보를 얻기가 어렵다. 또한, 긴 반복 영역에 의해 분리된 핵산 세그먼트는 공통의 스캐폴드로 연결되거나 어셈블리될 수 없다. 이러한 문제는 FFPE 추출 방법, 비등, 프로테이나제 처리, 장기간 보관, 실온 보관, 효소 또는 비효소적 분해, 또는 뉴클레아제 활성을 갖는 조성물로 단리하는 동안 또는 후의 오염에 의한 이중 가닥 파단의 도입에 의해서만 증폭된다.

[0045] 샘플 분해는 디 노보(de novo) 어셈블리에 상당한 영향을 미친다. 본 개시내용은 시간 경과에 따른 이중 가닥 파단을 통한 DNA 손상을 방지함으로써, 및 선택적으로 추가로 이중 가닥 파단의 페이즈 결정에 대한 영향을 감소시킴으로써 일부 실시양태에서 상기 문제를 동시에 해결한다. 보존된 높은 DNA 완전성은 수백 킬로 염기 단위의 게놈 거리 및 적절한 투입 DNA를 갖는 최대 메가 염기에 이르는 극장범위 리드쌍 데이터(XLRP)를 생성하는 방법을 가능하게 한다.

[0046] 이러한 데이터는 이중 가닥 파단, DNA 단편화, 및 중심체를 포함하는 게놈 내의 큰 반복 영역으로 인한 물리적 링키지 정보의 상실에 의한 물리적 링키지 정보의 상실에 의해 나타나는 실질적인 장벽을 극복하기 위해; 비용 효율적인 디 노보 어셈블리을 가능하게 하기 위해; 및 게놈 분석 및 개인 맞춤형 의약을 위한 충분한 완전성 및 정확성을 갖는 재시퀀싱 데이터를 생성하기 위해 매우 중요하다.

[0047] 본원의 개시내용은 통상적인 추출(예를 들어, FFPE 추출) 방법에서 대체로 발생하는 페이즈 및/또는 물리적 링키지 정보의 상실을 방지하거나, 또는 대안으로, 물리적 링키지 정보가 하류 처리, 예컨대 프로테이나제 처리의 비등시에도 보존되도록, 이중 가닥 파단과는 독립적으로 페이즈 및/또는 물리적 링키지 정보를 보존함으로써 상기 문제를 해결한다. 물리적 링키지 정보는 핵산 분자의 제1 세그먼트 및 제2 세그먼트를 결합시켜 이들이 이들의 공통적인 포스포디에스테르 백본과는 관계없이 유지되도록 함으로써 물리적으로 보존될 수 있다. 대안으로 또는 조합하여, 물리적 링키지 정보는 세그먼트 사이에 이중 가닥 파단이 도입되는 경우 제1 세그먼트 및 인접 서열 및 제2 세그먼트 및 인접 서열의 시퀀싱을 통해 얻은 태그 또는 다른 표지 정보가 제1 세그먼트 및 제2 세그먼트를 공통적인 핵산 분자의 공통적인 페이즈에 매핑하기에 충분하도록, 공통적인 핵산 분자의 제1 세그먼트 및 제2 세그먼트의 태그 부착 또는 상호 표지를 통해 보존될 수 있다. 태그 부착은 대안으로 제1 세그먼트를 제2 세그먼트에 라이게이션함으로써 이루어질 수 있고, 여기서 제1 세그먼트 및 제2 세그먼트는 동일한 원래의 DNA 분자 상에서 물리적으로 연결되어 있지만, 제2 세그먼트는 제1 세그먼트에 인접하여 존재하는 것은 아니다. 예를 들어, 제1 세그먼트 및 제2 세그먼트는 DNA 분자 서열을 따라 인접하지 않을 수 있지만, 서로 물리적으로 근접하거나 또는 염색질과 같은 구조에서 폴딩으로 인해 공통의 복합체에서 적어도 구성 성분으로 존재할 수 있다. 이러한 세그먼트의 노출된 말단은 함께 라이게이션될 수 있다. 또 다른 예에서, 태그 부착은 제1 세그먼트 및 제2 세그먼트가 공통의 복합체 또는 공통적인 분자에 인식 가능하게 매핑되도록 제1 및 제2 세그먼트 둘 모두에 바코드(예를 들어, 올리고뉴클레오티드 바코드) 또는 다른 태그를 라이게이션함으로써 이루어진다. 염색질 재어셈블리 또는 핵산 표지 또는 태그 부착을 통한 물리적 링키지 정보를 보존하는 방법은 이미 문헌에 기재되

어 있다(그 전부가 본 원에 포함된 PCT 특허 출원 PCT/US2016/024225).

[0048] 본원에서 일부 실시양태의 유의하게 중요한 것은 염색질이 단백질 또는 비단백질 핵산 결합 모이어티를 사용하여 재구성될 수 있도록, 보존된 샘플, 예컨대 FFPE 포매된 샘플로부터의 긴 핵산을 보존하는 것이다. 재구성된 염색질의 사용은 DNA의 매우 멀리 떨어져 있지만 분자적으로 연결된 세그먼트 사이의 회합을 형성하는 데 유리하다. 본 개시내용은 먼 세그먼트가 공통적인 포스포디에스테르 백본과는 관계없이 함께 회합되고 서로 물리적으로 결합되어, 공통적인 DNA 분자의 이전에 멀리 떨어져 있는 부분을 물리적으로 연결하는 것을 가능하게 한다. 결과적으로, 이들 상이한 핵산 세그먼트 사이의 이중 가닥 연결의 파괴는 페이즈 및/또는 물리적 링키지 정보의 상실을 초래하지 않는다. 바람직하게는, 염색질 재구성은 개개의 재구성된 염색질 단위당 하나 초과의 핵산 분자의 포함을 최소화하거나 방지하는 조건 하에서 발생하도록 유의해야 한다. 후속 처리는 관련 세그먼트의 서열이 확인되어, 그의 게놈에서의 분리가 투입 DNA 분자의 전체 길이까지 연장되는 리드쌍을 생성할 수 있도록 한다.

[0049] **샘플**

[0050] 본원에서 샘플은 예를 들어 포르말린 고정 파라핀 포매된 샘플로서 보존되고, 일부 경우에 분석하기 전에 상당한 기간 동안 보관된다. 샘플은 약물 시험을 통해 얻을 수 있으며, 양성 약물 치료 결과와 관련이 있거나 이 결과를 예측할 수 있는 게놈의 구조적 재배열을 확인하기 위해 수년 후에 검사될 수 있다. 이러한 샘플은 게놈의 구조적 정보와 같은 긴 거리의 서열 정보를 결정하는 데 사용될 수 있다. 본원에서 개시되는 방법에 의해 생성된 장범위 정보는 역위, 결실 및 중복과 같은 구조적 변화를 검출하기 위해 사용될 수 있다. 구조적 변이 검출은 또한 활성 증강 인자가 암 유전자에 근접하게 될 때 또는 억제성 시스 작용 요소가 종양 억제 인자에 근접하게 될 때를 확인하기 위해 사용될 수 있다. 이러한 추진(driver) 사건의 확인은 암 연구, 특히 연구가 완료된 후 종양 조직이 오래 보존되고 종양의 다양한 세포 하위 집단이 상이한 게놈 재구조화 사건을 포함하는 연구에 적용 가능하다. 예를 들어, 새로운 구조적 변이체가 검출되고, 암 유형의 원인 인자로 결정될 수 있다.

[0051] 본원의 방법은 보존된 샘플, 예컨대 환자, 연구 동물 또는 환경 샘플로부터 수득된 샘플로부터 게놈의 구조적 정보를 수득하기 위해 사용된다. 일부 샘플에는 생검 샘플, 수술 샘플, 종양 샘플, 전체 장기 및 다른 샘플이 포함된다. 이러한 샘플은 종종 고정제, 예컨대 포름알데히드, 포르말린, UV 광, 미토마이신 C, 질소 머스타드, 멜팔란, 1,3-부타디엔 디에폭시드, 시스 디아민디클로로백금(II) 또는 사이클로포스파미드 내에서 보존된다. 보존된 샘플은 직접 고정되고, 일부 경우에 고정액 내에 샘플을 떨어뜨려 균질화하지 않고 고정된다. 보존된 샘플은 수개월 또는 수년 동안 보관할 수 있다. 또한, 샘플의 무손상 특성은 샘플의 위치 정보를 보존하여, 게놈의 구조적 정보를 샘플 전체에 걸쳐 공간적으로 분석할 수 있다. 예를 들어, 생검 샘플의 가장자리에서 게놈의 구조적 정보는 생검 샘플의 중앙의 게놈의 구조적 정보와 비교될 수 있다.

[0052] 본원에서 개시되는 방법에 기초한 구조적 변이 검출은 또한 유전자 융합체의 DNA 구조를 결정하기 위해 사용될 수 있다. 일반적으로 사용되는 FISH 방법 또는 RNA-seq는 DNA 재배열이 일어났는지를 결정할 수 있지만, 재배열의 실제 서열은 이러한 방법에 의해 제공되지 않는다. 다른 한편으로, 관심 유전자 융합체를 생성하는 구조적 변이체를 결정하기 위한 방법이 본원에서 제공된다.

[0053] 3차원 DNA 구조 정보를 결정하는 방법이 본원에서 제공된다. 일부 경우에, 염색질의 열림 또는 닫힘 상태가 이러한 방법으로 검출된다. 본원에서 개시되는 방법에 의해 수집된 구조 정보는 또한 격리제(insulator) 또는 루프의 존재 또는 부재를 결정하거나, 신규한 루프 또는 다른 새로운 염색체 내부 또는 염색체 사이의 회합을 검출하기 위해 사용될 수 있다.

[0054] 조직 매핑을 위한 방법이 본원에서 제공된다. 조직 매핑은 종양과 같은 조직의 상이한 영역의 펀치 생검 및 구조적 또는 페이징 정보가 상이한 영역의 게놈 이질성을 결정하기 위해 각각의 생검으로부터 결정되는 과정이다.

[0055] 본원에서 개시되는 방법은 보존된(예를 들어, FFPE) 샘플로부터 장범위 정보를 포함하는 리드쌍 라이브러리를 생성하기 위해 사용될 수 있다. 이 라이브러리는 무기한 보존된 샘플, 예를 들어 FFPE 조직에서 회수될 수 있다.

[0056] 림프구의 구조적 및 페이즈 정보를 결정하는 방법이 본원에서 제공된다. 일부 경우에, 이러한 방법은 상이한 세포 또는 수용체 아형을 구별하기 위해 사용된다.

[0057] 본원에서 제공되는 방법은 장범위 데이터 및 데이터를 포함하는 페이즈 정보를 사용하여 구조적 변이체 또는 게놈 재배열의 검출을 위해 일부 실시양태에서 사용된다. 이 방법의 출발 물질은 대부분의 임상 샘플 보존에서 통상적으로 사용되는 바와 같이, 포르말린에 고정되고 파라핀에 포매된 샘플이다. 본원에서 제공되는 방법을 사용

하여, 구조적 및 장범위 정보가 샘플로부터 얻어지고; 상기 정보는 높은 수준의 DNA 단편화로 인해 현재의 방법을 사용하여 얻을 수 없다. 따라서, 본원에서 제공되는 방법을 사용하면, 임상 연구 및 약물 발견의 많은 분야에서 이 새로운 데이터를 사용할 수 있는 기회가 제공된다.

[0058] 본원에서 제공되는 방법의 임상 연구 적용은 환자 샘플을 사용하여 요법 반응 또는 저항을 추적하는 것을 포함한다. 라이브러리 준비 또는 시퀀싱 변이를 완화하기 위해, 샘플을 동시에 처리하는 것이 유리하다. 이를 위해서는 FFPE와 같은 초기 시점 샘플을 보존해야 한다. 본원에서 제공되는 방법은 다수의 시점으로부터의 샘플을 동시에 처리하고 분석할 수 있도록 사용 가능한 게놈 물질을 상기 보존된 샘플로부터 효율적으로 추출하는 방법을 제공한다.

[0059] 한 예에서, 샘플(예를 들어, 생검)은 환자로부터 채취되고, 의료 시술 중에 고정제(예를 들어, 포르말린)에 위치한다. 이 고정된 샘플은 본 개시내용의 기술을 사용하여 후속적으로 분석된다. 예를 들어, 암과 관련된 재배열과 같은 게놈 특징을 확인할 수 있다. 암의 게놈 정보를 체세포의 게놈 정보와 구별하기 위해 종양/비종양 페이징이 분석될 수 있다.

[0060] 또한, 본원에서 제공되는 방법을 사용하여, 유용한 장범위 게놈 정보는 또한 이러한 추출 방법의 발명 이전에 보존된 보다 오래된 샘플로부터도 얻을 수 있다. 예를 들어, 종양 샘플 은행은 본원에서 제공되는 방법을 사용하여 처리할 수 있고, 임상 관련 정보를 얻기 위해 상기 정보를 수집하기 위해 환자의 알려진 결과와 서로 연결시킬 수 있다. 이러한 방식으로, 본원에서 제공되는 방법은 예측 및 진단의 상관관계 판단을 허용한다.

[0061] 본원에서 제공되는 방법 및 조성물은 보존된 조직의 구조적 변이 프로파일을 결정하기 위해 사용될 수 있다. 이러한 구조적 변이 프로파일은 별개의 아형 또는 다른 클러스터를 정의하기 위해 다른 데이터 세트, 예를 들어 유전자 발현 프로파일, 돌연변이 프로파일, 메틸화 프로파일 등과 함께 사용될 수 있다.

[0062] 본원에서 제공되는 방법에 의해 결정된 구조적 변이 프로파일은 또한 시간 경과에 따른 돌연변이의 구조적 변화를 결정하기 위해 사용된다. 예를 들어, 일부 경우에 진행 또는 회귀를 통해 시작부터 종양 게놈 구조의 구조적 변이체의 변화를 모니터링할 수 있다. 이런 식으로, 종양 악성 및 전이가 더 잘 이해될 수 있다. 모니터링은 샘플의 가용성에 따라, 3차원 샘플의 다양한 하위 집단을 조사함으로써 공간적으로, 및 보존된 샘플의 시간 경과를 조사하여 시간적으로 수행할 수 있다.

[0063] 본원에서 제공되는 방법은 예탁, 보관 또는 다른 장기 보존된 유전자 샘플에서 수행될 수도 있다. 예를 들어, 희귀하거나 알려지지 않은 질환으로 고통받고 현재 사망한 환자의 보존된 조직 샘플의 보관물은 본원에서 제공되는 방법에 의해 분석할 수 있고, 따라서 표준 방법을 사용하여 얻을 수 없는 통찰력을 제공한다.

[0064] 본원에서 개시되는 기술에 의해 분석되는 샘플은 DNA의 보존 또는 구조적 정보를 포함하는 장범위 DNA 정보의 보존에 해로운 조건을 포함하는 다양한 조건에 적용되었거나 분해될 수 있다. 일부 경우에, 샘플은 산 처리에 적용되어야 한다. 일부 경우에, 샘플은 가교결합제, 예컨대 포름알데히드 또는 포르말린에 적용되었다. 일부 경우에, 샘플에 대해 포매, 예컨대 파라핀 포매가 수행되었다. 일부 경우에, 샘플에 대해 포매, 예컨대 파라핀 포매가 수행되지 않았다. 일부 경우에, 샘플은 열 처리(예를 들어, 포매 물질을 녹이기 위해)에 적용되었다. 일부 경우에, 샘플은 용매, 예컨대 크실렌(예를 들어, 접착제를 용해시키기 위해)에 적용되었다.

[0065] 고정된 샘플은 고정 후에, 그러나 후속적인 처리 또는 분석 전에 다양한 조건에 적용될 수 있다. 예를 들어, 고정 후에, 적어도 약 10분, 20분, 30분, 40분, 50분, 1시간, 1.5시간, 2시간, 3시간, 4시간, 5시간, 6시간, 7시간, 8시간, 9시간, 10시간, 11시간, 12시간, 18시간, 1일, 2일, 3일, 4일, 5일, 6일, 1주, 2주, 3주, 1달, 2달, 3달, 4달, 5달, 6달, 7달, 8달, 9달, 10달, 11달, 1년, 2년, 3년, 4년, 5년, 6년, 7년, 8년, 9년, 10년, 15년, 20년, 25년, 30년, 35년, 40년, 45년, 50년, 55년, 60년, 65년, 70년, 75년, 80년, 85년, 90년, 95년, 100년 또는 그 초과의 시간이 경과할 수 있다. 고정 후에, 샘플은 적어도 약 5℃, 10℃, 15℃, 20℃, 25℃, 30℃, 35℃, 40℃, 45℃, 50℃, 55℃, 60℃, 65℃, 70℃, 75℃, 80℃, 85℃, 90℃, 95℃, 100℃ 또는 그 초과의 온도 상승을 겪을 수 있다. 고정 후에, 샘플은 적어도 약 5℃, 10℃, 15℃, 20℃, 25℃, 30℃, 35℃, 40℃, 45℃, 50℃, 55℃, 60℃, 65℃, 70℃, 75℃, 80℃, 85℃, 90℃, 95℃, 100℃ 또는 그 초과의 온도 감소를 겪을 수 있다. 고정 후에, 샘플은 적어도 약 10 파스칼(Pa), 20 Pa, 30 Pa, 40 Pa, 50 Pa, 60 Pa, 70 Pa, 80 Pa, 90 Pa, 100 Pa, 110 Pa, 120 Pa, 130 Pa, 140 Pa, 150 Pa, 160 Pa, 170 Pa, 180 Pa, 190 Pa, 200 Pa, 210 Pa, 220 Pa, 230 Pa, 240 Pa, 250 Pa, 260 Pa, 270 Pa, 280 Pa, 290 Pa, 300 Pa, 310 Pa, 320 Pa, 330 Pa, 340 Pa, 350 Pa, 360 Pa, 370 Pa, 380 Pa, 390 Pa, 400 Pa, 410 Pa, 420 Pa, 430 Pa, 440 Pa, 450 Pa, 460 Pa, 470 Pa, 480 Pa, 490 Pa, 500 Pa, 550 Pa, 600 Pa, 650 Pa, 700 Pa, 750 Pa, 800 Pa, 850 Pa, 900 Pa, 950

Pa, 1000 Pa, 2000 Pa, 3000 Pa, 4000 Pa, 5000 Pa, 6000 Pa, 7000 Pa, 8000 Pa, 9000 Pa, 10000 Pa, 20000 Pa, 30000 Pa, 40000 Pa, 50000 Pa, 60000 Pa, 70000 Pa, 80000 Pa, 90000 Pa, 100000 Pa, 101325 Pa, 또는 그 초과의 압력(예를 들어, 주위 압력) 감소를 겪을 수 있다. 고정 후, 샘플은 적어도 약 10 파스칼(Pa), 20 Pa, 30 Pa, 40 Pa, 50 Pa, 60 Pa, 70 Pa, 80 Pa, 90 Pa, 100 Pa, 110 Pa, 120 Pa, 130 Pa, 140 Pa, 150 Pa, 160 Pa, 170 Pa, 180 Pa, 190 Pa, 200 Pa, 210 Pa, 220 Pa, 230 Pa, 240 Pa, 250 Pa, 260 Pa, 270 Pa, 280 Pa, 290 Pa, 300 Pa, 310 Pa, 320 Pa, 330 Pa, 340 Pa, 350 Pa, 360 Pa, 370 Pa, 380 Pa, 390 Pa, 400 Pa, 410 Pa, 420 Pa, 430 Pa, 440 Pa, 450 Pa, 460 Pa, 470 Pa, 480 Pa, 490 Pa, 500 Pa, 550 Pa, 600 Pa, 650 Pa, 700 Pa, 750 Pa, 800 Pa, 850 Pa, 900 Pa, 950 Pa, 1000 Pa, 2000 Pa, 3000 Pa, 4000 Pa, 5000 Pa, 6000 Pa, 7000 Pa, 8000 Pa, 9000 Pa, 10000 Pa, 20000 Pa, 30000 Pa, 40000 Pa, 50000 Pa, 60000 Pa, 70000 Pa, 80000 Pa, 90000 Pa, 100000 Pa, 101325 Pa, 또는 그 초과의 압력(예를 들어, 주위 압력) 증가를 겪을 수 있다. 고정 후에, 샘플은 적어도 약 0.1 미터(m), 0.2 m, 0.3 m, 0.4 m, 0.5 m, 0.6 m, 0.7 m, 0.8 m, 0.9 m, 1 m, 2 m, 3 m, 4 m, 5 m, 6 m, 7 m, 8 m, 9 m, 10 m, 11 m, 12 m, 13 m, 14 m, 15 m, 16 m, 17 m, 18 m, 19 m, 20 m, 또는 그 초과의 높이 변화를 겪을 수 있다.

[0066] 고정된 샘플은 적어도 약 10분, 20분, 30분, 40분, 50분, 1시간, 1.5시간, 2시간, 3시간, 4시간, 5시간, 6시간, 7시간, 8시간, 9시간, 10시간, 11시간, 12시간, 18시간, 24시간 또는 그 초과의 시간 동안 지속되는 고정 반응에서 고정될 수 있다. 일부 경우에, 고정된 샘플은 적어도 약 30분 동안 지속되는 고정 반응에서 고정된다. 일부 경우에, 고정 반응 시간은 고정 반응이 켄칭되기 전에 경과된 시간일 수 있다. 일부 경우에, 고정된 샘플은 켄칭되지 않는 고정 반응에서 고정된다.

[0067] 본원에서 개시되는 방법은 관심있는 선택적인 게놈 영역뿐만 아니라 관심있는 선택적인 영역과 상호작용할 수 있는 게놈 영역의 유전 정보의 분석에 사용될 수 있다. 본원에서 개시되는 바와 같은 증폭 방법은 유전 분석을 위해 관련 기술 분야에 공지된 장치, 키트 및 방법, 예컨대 미국 특허 제6,449,562호, 제6,287,766호, 제7,361,468호, 제7,414,117호, 제6,225,109호 및 제6,110,709호에 기재된 것에 사용될 수 있다. 일부 경우에, 본 개시내용의 증폭 방법은 다형성의 존재 또는 부재를 결정하기 위한 DNA 혼성화 연구를 위해 표적 핵산을 증폭시키기 위해 사용될 수 있다. 다형성 또는 대립유전자는 유전 질환과 같은 질환 또는 병태와 관련될 수 있다. 다른 경우에, 다형성은 질환 또는 병태, 예를 들어 중독, 퇴행성 및 연령 관련 병태, 암 등과 관련된 다형성에 대한 감수성과 관련될 수 있다. 다른 경우에, 다형성은 관상동맥 건강의 증가, 또는 HIV 또는 말라리아와 같은 질병에 대한 저항성, 또는 골다공증, 알츠하이머병 또는 치매와 같은 퇴행성 질환에 대한 저항성과 같은 유익한 특징과 관련될 수 있다.

[0068] 본 개시내용의 조성물 및 방법은 진단, 예측, 치료, 환자 계층화, 약물 개발, 치료 선택 및 스크리닝 목적을 위해 사용될 수 있다. 본 개시내용은 많은 상이한 표적 분자가 본 개시내용의 방법을 사용하여 단일 생체분자 샘플로부터 한 번에 분석될 수 있다는 이점을 제공한다. 이것은 예를 들어 하나의 샘플에 대해 여러 진단 시험을 수행할 수 있도록 한다.

[0069] 본 개시내용의 조성물 및 방법은 게놈학에서 사용될 수 있다. 본원에서 설명되는 방법은 본원에서 매우 바람직한 응답을 신속하게 제공할 수 있다. 본원에서 설명되는 방법 및 조성물은 진단 또는 예측을 위해 및 건강 및 질환의 지표로서 사용될 수 있는 바이오마커를 찾는 과정에서 사용될 수 있다. 본원에서 설명되는 방법 및 조성물은 약물의 스크리닝, 예를 들어, 약물 개발, 치료법 선택, 치료 효능의 결정 및/또는 제약 개발을 위한 표적의 확인을 위해 사용될 수 있다. 단백질이 체내에서 최종 유전자 산물이기 때문에, 약물을 포함한 스크리닝 검정에서 유전자 발현을 시험하는 능력은 매우 중요하다. 일부 실시양태에서, 본원에서 설명되는 방법 및 조성물은 수행되는 특정 스크리닝에 관한 대부분의 정보를 제공할 단백질 및 유전자 발현을 동시에 측정할 것이다.

[0070] 본 개시내용의 조성물 및 방법은 유전자 발현 분석에 사용될 수 있다. 본원에서 설명되는 방법은 뉴클레오티드 서열을 구별한다. 표적 뉴클레오티드 서열 사이의 차이는 예를 들어 단일 핵산 염기 차이, 핵산 결실, 핵산 삽입 또는 재배열일 수 있다. 하나 초과의 염기를 포함하는 이러한 서열 차이가 또한 검출될 수 있다. 본 개시내용의 방법은 감염성 질환, 유전 질환 및 암을 검출할 수 있다.

[0071] 본 발명의 방법은 샘플 내에 이환된 세포 유형이 존재하는지의 여부, 질환의 단계, 환자의 예후, 환자가 특정 치료에 반응하는 능력, 또는 환자를 위한 최상의 치료를 결정하기 위해 환자로부터 얻거나 환자로부터 유래된 생체분자 샘플의 분석에 적용될 수 있다. 본 발명의 방법은 또한 특정 질환에 대한 바이오마커를 확인하기 위해서도 적용될 수 있다.

[0072] 일부 실시양태에서, 본원에서 설명되는 방법은 병태의 진단에 사용된다. 본원에서 사용되는 바와 같이, 병태를

"진단하다" 또는 병태의 "진단"은 병태의 예측 또는 진단, 병태에 대한 소인의 결정, 병태의 치료에 대한 모니터링, 질환의 치료 반응, 또는 병태의 예후, 병태 진행, 또는 병태의 특정 치료에 대한 반응의 진단을 포함할 수 있다. 예를 들어, 보존된(예를 들어, FFPE) 임상 샘플은 샘플에서 질환 또는 악성 세포 유형의 마커의 존재 및/또는 양을 결정하여 질환 또는 암을 진단하거나 병기를 결정하기 위해 본원에서 설명되는 임의의 방법에 따라 검정될 수 있다.

[0073] 일부 실시양태에서, 본원에서 설명되는 방법 및 조성물은 병태의 진단 및 예측을 위해 사용된다. 많은 면역학, 증식성 및 악성 질환 및 장애는 특히 본원에서 설명되는 방법에 적용될 수 있다. 면역학적 질환 및 장애는 알레르기 질환 및 장애, 면역 기능 장애 및 자가면역 질환 및 병태를 포함한다. 알레르기 질환 및 장애는 알레르기성 비염, 알레르기성 결막염, 알레르기성 천식, 아토피성 습진, 아토피성 피부염 및 음식 알레르기를 포함하고, 이로 제한되지 않는다. 면역 결핍은 다음을 포함하고, 이로 제한되지 않는다: 중증 복합 면역결핍증(SCID), 호산구 과다 증후군, 만성 육아종성 질환, 백혈구 부착 결핍 I 및 II, 고 IgE 증후군, 체디악 히가시(Chediak Higashi), 호중구 증가증, 호중구 감소증, 무형성증, 무감마글로불린혈증, 고 IgM 증후군, 디죠지/입천장-심장-얼굴(DiGeorge/Velocardial-facial) 증후군, 및 인터페론 감마-TH1 경로 결함. 자가면역 및 면역 조절장애는 다음을 포함하고, 이로 제한되지 않는다: 류마티스 관절염, 당뇨병, 전신성 홍반성 루푸스, 그레이브스(Graves) 병, 그레이브스 안병증, 크론(Crohn) 병, 다발성 경화증, 건선, 전신 경화증, 갑상선종 및 림프성 갑상선종(하시모토(Hashimoto) 갑상선염, 림프종성 갑상선종), 특발성 혈소판 감소성 자반병, 원형 탈모증, 자가면역성 심근염, 경화 태선, 자가면역성 포도막염, 애디슨(Addison) 병, 위축성 위염, 중증 근무력증, 특발성 혈소판 감소 자반병, 용혈성 빈혈, 원발성 담즙성 경화증, 베게너(Wegener) 육아종증, 결절성 다발성 동맥염, 및 염증성 장 질환, 동종이식편 거부 반응 및 감염성 미생물 또는 환경 항원에 대한 알레르기 반응에 의한 조직 파괴.

[0074] 본 개시내용의 방법에 의해 평가될 수 있는 증식성 질환 및 장애는 다음을 포함하고, 이로 제한되지 않는다: 신생아에서의 혈관종증; 속발성 진행성 다발성 경화증; 만성 진행성 골수변성 질환; 신경섬유종증; 신경절 신경종증; 켈로이드 형성; 뼈의 파제트(Paget) 병; 섬유낭종 질환(예를 들어, 유방 또는 자궁의); 유육종증; 페로니(Peronie) 및 뒤피트랑(Duputren)의 섬유증, 간경화, 아테롬성 경화증 및 혈관 재협착.

[0075] 본 개시내용의 방법에 의해 평가될 수 있는 악성 질환 및 장애는 혈액학적 악성 종양 및 고형 종양을 모두 포함한다.

[0076] 혈액학적 악성 종양은 특히 샘플이 혈액 샘플일 때 본 개시내용의 방법에 적용될 수 있고, 그 이유는 상기 악성 종양이 혈액 내 세포의 변화를 수반하기 때문이다. 이러한 악성 종양은 비호지킨 림프종, 호지킨 림프종, 비-B 세포 림프종 및 다른 림프종, 급성 또는 만성 백혈병, 적혈구 증가증, 혈소판 증가증, 다발성 골수종, 골수이형성 장애, 골수증식성 장애, 골수섬유증, 비정형 면역 림프증식성 및 형질 세포 질환을 포함한다.

[0077] 본 개시내용의 방법에 의해 평가될 수 있는 형질 세포 질환은 다발성 골수종, 아밀로이드증 및 발덴스트롬(Waldenstrom)의 거대글로불린혈증을 포함한다.

[0078] 고형 종양의 예는 대장암, 유방암, 폐암, 전립선암, 뇌종양, 중추신경계 종양, 방광 종양, 흑색종, 간암, 골육종 및 다른 골암, 정소 및 난소 암종, 두경부 종양 및 자궁 경부 신생물을 포함하고 이로 제한되지 않는다.

[0079] 유전 질환은 또한 본 개시내용의 방법에 의해 검출될 수 있다. 이것은 염색체 및 유전자 이상 또는 유전 질환에 대한 출생전 또는 출생후 스크리닝에 의해 수행될 수 있다. 검출될 수 있는 유전 질환의 예는 다음을 포함한다: 21 히드록실라제 결핍증, 낭포성 섬유증, 취약 X 증후군, 터너(Turner) 증후군, 뒤시엔느(Duchenne) 근이영양증, 다운 증후군 또는 다른 삼염색체증, 심장 질환, 단일 유전자 질환, HLA 유형 분류, 페닐케톤뇨증, 겸상 적혈구 빈혈, 테이-삭스(Tay-Sachs) 병, 지중해 빈혈, 클라인펠터(Klinefelter) 증후군, 헌팅턴(Huntington) 병, 자가면역 질환, 지방증, 비만 결함, 혈우병, 선천성 대사 이상 및 당뇨병.

[0080] 본원에서 설명되는 방법은 샘플 내의 박테리아 또는 바이러스 각각의 마커의 존재 및/또는 양을 결정함으로써 병원체 감염, 예를 들어 세포내 박테리아 및 바이러스에 의한 감염을 진단하기 위해 사용될 수 있다.

[0081] 매우 광범한 감염성 질환이 본 개시내용의 방법에 의해 검출될 수 있다. 감염성 질환은 박테리아, 바이러스, 기생충 및 진균 감염 인자에 의해 유발될 수 있다. 약물에 대한 다양한 감염 인자의 저항성이 또한 본 개시내용을 사용하여 결정될 수 있다.

[0082] 본 개시내용에 의해 검출될 수 있는 박테리아 감염 인자는 다음을 포함한다: 에셔리키아 콜라이(*Escherichia coli*), 살모넬라(*Salmonella*), 시겔라(*Shigella*), 클렙시엘라(*Klesbiella*), 슈도모나스(*Pseudomonas*), 리스테리아(*Listeria*), 모노시토게네스(*monocytogenes*), 미코박테리움 튜베르큘로시스(*Mycobacterium tuberculosis*),

미코박테리움 아비우민트라셀룰라레(*Mycobacterium aviumintracellulare*), 예르시니아(*Yersinia*), 프란시셀라(*Francisella*), 파스테우렐라(*Pasteurella*), 브루셀라(*Brucella*), 클로스트리디아(*Clostridia*), 보르데텔라 페르투시스(*Bordetellapertussis*), 박테로이데스(*Bacteroides*), 스타필로코쿠스 아우레우스(*Staphylococcus aureus*), 스트렙토코쿠스 뉴모니아(*Streptococcus pneumonia*), 비-헤몰리틱 종(*B-Hemolytic strep.*), 코리네박테리아(*Corynebacteria*), 레지오넬라(*Legionella*), 미코플라스마(*Mycoplasma*), 우레아플라스마(*Ureaplasma*), 클라미디아(*Chlamydia*), 네이세리아 고노레아(*Neisseria gonorrhea*), 네이세리아 메닝기티데스(*Neisseria meningitides*), 헤모필루스 인플루엔자(*Hemophilus influenza*), 엔테로코쿠스 파에칼리스(*Enterococcus faecalis*), 프로테우스 불가리스(*Proteus vulgaris*), 프로테우스 미라빌리스(*Proteus mirabilis*), 헬리코박터 필로리(*Helicobacter pylori*), 트레포네마 팔라듐(*Treponema palladium*), 보렐리아 부르그도르페리(*Borrelia burgdorferi*), 보렐리아 레쿠렌티스(*Borrelia recurrentis*), 리케치알 파토겐스(*Rickettsial pathogens*), 노카르디아(*Nocardia*), 및 악티노미세테스(*Acitnomycetes*).

[0083] 본 개시내용에 의해 검출될 수 있는 진균 감염 인자는 다음을 포함한다: 크립토코쿠스 네오포르만스(*Cryptococcus neoformans*), 블라스토마이세스 더마티티디스(*Blastomyces dermatitidis*), 히스토플라스마 캡술라툼(*Histoplasma capsulatum*), 콕시디오이데스 이미티스(*Coccidioides immitis*), 파라콕시디오이데스 브라실리엔시스(*Paracoccidioides brasiliensis*), 칸디다 알비칸스(*Candida albicans*), 아스페르길루스 푸미가우투스(*Aspergillus fumigautus*), 키코마이세테스(리조푸스)(*Phycomycetes* (*Rhizopus*)), 스포로트릭스 셴키(*Sporothrix schenckii*), 크로모미코시스(*Chromomycosis*), 및 마두로미코시스(*Maduromycosis*).

[0084] 본 개시내용에 의해 검출될 수 있는 바이러스 감염 인자는 다음을 포함한다: 인간 면역 결핍 바이러스, 인간 T 세포 림프구 영양성(lymphocytotrophic) 바이러스, 간염 바이러스(예를 들어, B형 간염 바이러스 및 C형 간염 바이러스), 엡스타인-바 바이러스, 사이토메갈로바이러스, 인간 유두종 바이러스, 오르토믹소 바이러스, 파라믹소 바이러스, 아데노바이러스, 코로나 바이러스, 랍도 바이러스, 폴리오 바이러스, 토가 바이러스, 분야 바이러스, 아레나 바이러스, 루벨라 바이러스 및 레오 바이러스.

[0085] 본 개시내용에 의해 검출될 수 있는 기생충 인자는 다음을 포함한다: 플라스모듐 팔시파룸(*Plasmodium falciparum*), 플라스모듐 말라리아(*Plasmodium malaria*), 플라스모듐 비박스(*Plasmodium vivax*), 플라스모듐 오발레(*Plasmodium ovale*), 온초베르바 볼불루스(*Onchoverva volvulus*), 레이슈마니아(*Leishmania*), 트리파노소마 종(*Trypanosoma spp.*), 스키스토소마 종(*Schistosoma spp.*), 엔카모에바 히스톨리티카(*Entamoeba histolytica*), 크립토스포리둠(*Cryptosporidum*), 지아르디아 종(*Giardia spp.*), 트리키모나스 종(*Trichimonas spp.*), 발라티듐 콜라이(*Balatidium coli*), 부케레리아 반크프티(*Wuchereria bancrofti*), 톡소플라즈마 종(*Toxoplasma spp.*), 엔테로비우스 베르미큘라리스(*Enterobius vermicularis*), 아스카리스 룸브리코이데스(*Ascaris lumbricoides*), 트리쿠리스 트리키우라(*Trichuris trichiura*), 드라쿤쿨루스 메디네시스(*Dracunculus medinesis*), 흡충류, 디필로보트리움 라툼(*Diphyllobothrium latum*), 타에니아 종(*Taenia spp.*), 뉴모시스티스 카리니이(*Pneumocystis carinii*), 및 네카토르 아메리카니스(*Necator americanis*).

[0086] 본 개시내용은 또한 감염 인자에 의한 약물 내성의 검출에 유용하다. 예를 들어, 반코마이신 내성 엔테로코쿠스 파에시움(*Enterococcus faecium*), 메티실린 내성 스타필로코쿠스 아우레우스, 페니실린 내성 스트렙토코쿠스스 뉴모니아에, 다제 내성 미코박테리움 튜베르큘로시스, 및 AZT 내성 인간 면역 결핍 바이러스는 모두 본 개시내용으로 확인될 수 있다,

[0087] 따라서, 본 개시내용의 조성물 및 방법을 사용하여 검출된 표적 분자는 환자 마커(예컨대, 암 마커) 또는 박테리아 또는 바이러스 마커와 같은 외래 물질 감염의 마커일 수 있다.

[0088] 본 개시내용의 조성물 및 방법은 그의 과다한 양이 생물학적 상태 또는 질환 상태를 나타내는 표적 분자, 예를 들어, 질환 상태의 결과로서 상향조절되거나 하향조절되는 혈액 마커를 확인 및/또는 정량하기 위해 사용될 수 있다.

[0089] 일부 실시양태에서, 본 개시내용의 방법 및 조성물은 사이토카인 발현에 사용될 수 있다. 본원에서 설명되는 방법의 낮은 민감도는 예를 들어, 병태, 암과 같은 질환의 진단 또는 예측, 및 준임상 병태의 확인의 바이오마커로서의 사이토카인의 조기 검출에 도움이 될 것이다.

[0090] 표적 폴리뉴클레오티드가 그로부터 유래되는 상이한 샘플은 동일한 개체로부터의 여러 샘플, 상이한 개체로부터의 샘플 또는 이들의 조합을 포함할 수 있다. 일부 실시양태에서, 샘플은 단일 개체로부터의 다수의 폴리뉴클레오티드를 포함한다. 일부 실시양태에서, 샘플은 둘 이상의 개체로부터의 다수의 폴리뉴클레오티드를 포함한다.

개체는 표적 폴리뉴클레오티드가 그의 유래될 수 있는 임의의 유기체 또는 그의 일부이며, 그의 비제한적인 예는 식물, 동물, 진균, 원생생물, 모네란(moneran), 바이러스, 미토콘드리아 및 엽록체를 포함한다. 샘플 폴리뉴클레오티드는 예를 들어 조직 또는 종양 생검을 포함하는 보존된(예를 들어, FFPE) 세포 샘플, 보존된(예를 들어, FFPE) 조직 샘플, 또는 이로부터 유래된 장기 샘플과 같은 대상체로부터 단리될 수 있다. 대상체는 소, 돼지, 마우스, 래트, 닭, 고양이, 개 등과 같은 동물을 포함하고 이로 제한되지 않는 동물일 수 있고, 일부 경우에 포유동물, 예컨대 인간이다. 샘플은 또한 화학적 합성과 같이 인공적으로 유도될 수 있다. 일부 실시양태에서, 샘플은 DNA를 포함한다. 일부 실시양태에서, 샘플은 게놈 DNA를 포함한다. 일부 실시양태에서, 샘플은 미토콘드리아 DNA, 엽록체 DNA, 플라스미드 DNA, 박테리아 인공 염색체, 효모 인공 염색체, 올리고뉴클레오티드 태그 또는 이들의 조합을 포함한다. 일부 실시양태에서, 샘플은 폴리머라제 연쇄 반응(PCR), 역전사 및 이들의 조합을 포함하고 이로 제한되지 않는, 프라이머와 DNA 폴리머라제의 임의의 적합한 조합을 사용하는 프라이머 연장 반응에 의해 생성된 DNA를 포함한다. 프라이머 연장 반응을 위한 주형이 RNA인 경우, 역전사 산물은 상보성 DNA(cDNA)로 언급된다. 프라이머 연장 반응에 유용한 프라이머는 하나 이상의 표적에 특이적인 서열, 무작위 서열, 부분적 무작위 서열, 및 이들의 조합을 포함할 수 있다. 프라이머 연장 반응에 적합한 반응 조건은 관련 기술 분야에 공지되어 있다. 일반적으로, 샘플 폴리뉴클레오티드는 표적 폴리뉴클레오티드를 포함할 수도 있고 포함하지 않을 수도 있는, 샘플에 존재하는 임의의 폴리뉴클레오티드를 포함한다.

[0091] 핵산의 추출 및 정제 방법은 관련 기술 분야에 잘 알려져 있다. 예를 들어, 핵산은 페놀, 페놀/클로로포름/이소아밀 알콜, 또는 트리졸(TRIzol) 및 트리리에이전트(TriReagent)를 비롯한 유사한 제제를 이용한 유기 추출에 의해 정제될 수 있다. 추출 기술의 다른 비제한적인 예는 다음을 포함한다: (1) 자동화된 핵산 추출 장치, 예를 들어 어플라이드 바이오시스템스(Applied Biosystems, 미국 캘리포니아주 포스터 시티 소재)에서 입수할 수 있는 모델 341 DNA 추출 장치를 사용하거나 사용하지 않으면서, 예를 들어 페놀/클로로포름 유기 시약을 사용한 유기 추출, 및 후속적인 에탄올 침전(Ausubel *et al.*, 1993); (2) 정지상 흡착 방법(미국 특허 제5,234,809호; [Walsh *et al.*, 1991]); 및 (3) 염 유도 핵산 침전 방법([Miller *et al.*, (1988)], 이러한 침전 방법은 일반적으로 "염석" 방법으로 언급됨). 핵산 단리 및/또는 정제의 또 다른 예는 자성 입자의 사용을 포함하고, 여기서 핵산은 이 자성 입자에 특이적으로 또는 비특이적으로 결합하고, 이어서 자석을 사용하여 비드가 단리되고, 세척되고, 비드로부터 핵산이 용리된다(예를 들어, US 특허 제5,705,628호 참조). 일부 실시양태에서, 상기 단리 방법에서 샘플로부터 원하지 않는 단백질을 제거하는 것을 돕는 효소 소화 단계, 예를 들어 프로테이나제 K, 또는 다른 유사 프로테아제를 사용한 소화가 선행될 수 있다. 예를 들어, 미국 특허 제7,001,724호를 참조한다. 바람직한 경우, RNase 억제제가 용해 완충제에 첨가될 수 있다. 특정 세포 또는 샘플 유형에 대해, 단백질 변성/소화 단계를 상기 프로토콜에 추가하는 것이 바람직할 수 있다. 정제 방법은 DNA, RNA, 또는 둘 모두를 단리하도록 지시될 수 있다. DNA 및 RNA 둘 모두가 추출 과정 동안 또는 후에 함께 단리될 때, 추가의 단계를 이용하여 하나 또는 둘 모두를 나머지로부터 별개로 정제할 수 있다. 예를 들어 크기, 서열, 또는 다른 물리적 또는 화학적 특성에 기반한 정제에 의해 추출된 핵산의 하위 분획을 또한 생성할 수 있다. 초기 핵산 단리 단계 이외에도, 예를 들어 과량의 또는 원하지 않는 시약, 반응물 또는 생성물을 제거하기 위해서, 본 개시내용의 방법의 임의의 단계 후에 핵산의 정제를 수행할 수 있다. 핵산 주형 분자는 2003년 10월 9일 공개된 미국 특허 출원 공개 제US2002/0190663 A1호에 기재된 바와 같이 수득할 수 있다. 일반적으로, 핵산은 다양한 기술, 예를 들어 문헌 [Maniatis, *et al.*, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor, N.Y., pp. 280-281 (1982)]에 기재된 기술에 의해 생물학적 샘플로부터 추출될 수 있다. 일부 경우에, 핵산은 먼저 생물학적 샘플로부터 추출된 후, 시험관 내에서 가교결합될 수 있다. 일부 경우에, 천연 회합 단백질(예를 들어, 히스톤)이 핵산으로부터 추가로 제거될 수 있다.

[0092] **<u>천연 염색질의 추출 및 회수</u>**

[0093] 긴 단편 길이 및/또는 페이즈 정보 함유 단편을 보존된 샘플(예를 들어, FFPE 샘플)로부터 추출하는 방법이 본원에서 제공된다. 일부 경우에, 이 방법은 보존된 샘플(예를 들어, FFPE 샘플)에 존재하는 염색질 구조를 보존하기 위해 보존된 세포(예를 들어, FFPE 세포)의 핵을 부드럽게 처리하는 것을 수반한다.

[0094] 장범위 DNA 단편 및/또는 페이즈 정보 함유 단편의 보존을 위한 추출 및 제자리(*in situ*) 라이브러리 제조를 수행하는 방법이 개시된다. 방출된 DNA는 이어서 리드쌍 라이브러리를 생성하기 위해 사용되는 것과 같이 분석을 위해 추가로 처리될 수 있다.

[0095] 보존된 샘플(예컨대, FFPE 샘플)을 용해제로 처리하여 포매 물질(예를 들어, 파라핀)을 용해시킬 수 있다. 일부 경우에, 용해제는 크실렌과 같은 용매이다. 적합한 용매제의 다른 예는 크실렌, 톨루엔 및 벤젠과 같은 유기 용매뿐만 아니라, 각각의 적합한 이성질체를 포함하고 이로 제한되지 않는다. 상기 조성물은 포매 물질이 용해제

에 용해되도록 혼합될 수 있다. 일부 경우에, 혼합은 볼텍싱(vortexing) 또는 고속 진탕 또는 교반을 포함한다. 대안으로, 부드러운 교반이 일부 경우에 사용된다. 샘플을 용매 및 용해된 포매 물질로부터 분리하기 위해 샘플은 예를 들어 샘플을 펠릿화하기에 충분한 속도로 원심분리함으로써 처리된다. 충분한 속도는 분당 14,000 회전과 같은 탁상용 원심분리기의 최대 속도를 포함하고 이로 제한되지 않는다. 용해된 포매 물질을 포함하는 용해제는 펠릿을 붕괴시키지 않도록 종종 부드럽게 제거될 수 있다. 이어서, 과량의 용해제는 세척 시약으로 제거할 수 있다. 일부 예에서, 세척제는 에탄올, 예를 들어 100% 에탄올이다. 샘플을 혼합하거나, 볼텍싱하거나, 교반하여 샘플 펠릿을 보관 용기의 내벽으로부터 제거한다. 샘플을 선택적으로 다시 원심분리하여 재펠릿화할 수 있다. 이어서, 임의의 남은 액체는 보관 용기로부터 제거되고, 샘플은 건조된다. 대표적인 건조 기술은 공기 건조, 진공 건조 또는 관련 기술 분야에 널리 알려진 다른 건조 기술을 포함한다. 건조 후, 완충제, 예컨대 용해 완충제를 샘플에 첨가한다. 용해 완충제는 트리스와 같은 완충제, 염화나트륨과 같은 염, 하나 이상의 세제, 예컨대 소듐 도데실 술페이트(SDS), 트리톤, 킬레이팅제, 예컨대 EDTA 및 이들의 임의의 조합물을 포함할 수 있다. 대표적인 용해 완충제는 50 mM 트리스 pH 8, 50 mM NaCl, 1% SDS, 0.15% 트리톤, 1 mM EDTA를 포함하지만, 관련 기술 분야의 통상의 기술자는 이 조성물에 대한 변형물이 쉽게 생성될 수 있음을 이해할 수 있다. 적절한 프로토콜이 다른 포매 물질을 제거하기 위해 사용될 수 있다.

[0096] 샘플은 선택적으로 진탕하거나 교반하면서, 예를 들어 충분한 시간 동안 인큐베이팅(예컨대, 37℃에서)함으로써 재수화할 수 있다. 이어서, 샘플을 교반하거나, 피펫팅하거나, 다른 방식으로 혼합하여 펠릿을 파괴하고 용해 완충제에 재현탁할 수 있다. 잔류하는 비가용성 잔해는 이어서 예를 들어 충분한 속도의 원심분리에 의해 용해 완충제로부터 분리될 수 있다. DNA-단백질 복합체는 핵산 단편에 태그를 부착하는 기술과 같은 하류 기술을 사용하여 회수하고 평가할 수 있다.

[0097] 천연 DNA:단백질 복합체(예를 들어, 염색질)는 핵산이 아니라 복합체가 그대로 보존되도록 보존된 샘플(예를 들어, FFPE 샘플)로부터 단리될 수 있다. 이러한 방법에서, 핵산의 물리적 링키지 정보는 복합체의 통상적으로 태그 부착된 단편이 원래의 샘플 내의 구조적 또는 물리적 링키지 배열을 갖는 것으로 추정될 수 있도록, 반드시 핵산의 포스포디에스테르 백본을 보존하는 것이 아니라 포스포디에스테르 백본 상태와는 독립적으로 링키지 정보를 보존함으로써 보존될 수 있다.

[0098] 염색질의 가용화는 천연 DNA:단백질 복합체를 단리하고 FFPE 샘플과 같은 보존된 샘플로부터 장범위 링키지 정보를 추출할 때 중요한 단계일 수 있다. 염색질 복합체는 프로테이나제 소화 및 초음파 처리를 포함하고 이로 제한되지 않는 다양한 방법을 통해 가용화될 수 있다. 이러한 가용화 방법은 조직 및 염색질을 붕괴시켜 가용성 염색질을 방출할 수 있다.

[0099] 프로테이나제 소화를 통한 가용화는 프로테이나제 K, 엔도프로테이나제 트립신, 키모트립신, 엔도프로테이나제 Asp-N, 엔도프로티아제 Arg-C, 엔도프로테이나제 Glu-C, 엔도프로테이나제 Lys-C, 써몰리신, 파파인, 섭틸리신, 클로스트리파인, 카르복시펩티다제 B, 카르복시펩티다제 P, 카르복시펩티다제 Y, 카텝신 C, 아실아미노산 방출 효소, 및 피로글루타메이트 아미노펩티다제 중 하나 이상을 포함하고 이로 제한되지 않는 다양한 프로테이나제 효소(펩티다제 또는 프로테아제 효소로도 알려짐)를 사용할 수 있다. 프로테이나제 효소는 세린 프로테아제, 시스테인 프로테아제, 트레오닌 프로테아제, 아스파르트산 프로테아제, 글루탐산 프로테아제, 메탈로프로테아제 또는 아스파라긴 펩티드 리아제일 수 있다.

[0100] 프로테이나제 소화를 통한 가용화의 예시적인 프로토콜은 포매 물질(예를 들어, 파라핀)의 제거, 프로테이나제 소화, 가용화된 염색질의 회수(예를 들어, SPRI 비드와 같은 카르복실화된 비드를 사용한) 및 시퀀싱 라이브러리 제조를 포함할 수 있다. 예를 들어, 먼저 조직 물질을 튜브(예를 들어, 1.5 mL 에펜도르프(Eppendorf) 튜브)에 넣을 수 있다. 이어서, 용매, 예컨대 크실렌, Hemo-De 또는 리모넨을 사용하여 포매 물질(예를 들어, 파라핀)을 용해시킬 수 있다. 에탄올(예를 들어, 100% EtOH)을 사용하여 용매를 제거할 수 있고, 샘플을 건조하여 에탄올을 제거할 수 있다. 그런 다음, 샘플을 프로테이나제 효소(예를 들어, 프로테이나제 K)로 소화시킬 수 있다. 이것은 대부분의 또는 모든 조직 샘플을 가용화시킬 수 있다. 이론에 의해 제한되지 않으면서, 프로테이나제 처리는 프로테이나제 처리 조건(예를 들어, 37℃에서 1시간) 동안 단백질-DNA 메틸렌 가교결합 역전이 매우 약하기 때문에 효과적일 수 있다.

[0101] 초음파 처리를 통한 가용화의 예시적인 프로토콜은 포매 물질(예를 들어, 파라핀)의 제거, 용해, 균질화, 초음파 처리, 가용화된 염색질의 회수(예를 들어, SPRI 비드와 같은 카르복실화된 비드를 사용한) 및 시퀀싱 라이브러리 제조를 포함할 수 있다. 예를 들어, 먼저 용매, 예컨대 크실렌, Hemo-De 또는 리모넨을 사용하여 포매 물질(예를 들어, 파라핀)을 용해시킬 수 있다. 이어서, 조직 시편은 예를 들어 100% 에탄올에서 순수한 물에 이르

기까지 상이한 농도의 에탄올로 연속적으로 세척할 때 재수화할 수 있다. 이어서, 조직 물질을 튜브에 넣고, 용해 완충제에서 인큐베이팅할 수 있다(예를 들어, 1시간 동안). 그런 다음, 조직을 소화 완충제(예를 들어, MNase 소화 완충제)과 같은 완충제에 재현탁할 수 있다. 이어서, 샘플은 다운스(Dounce) 균질화를 포함하고 이로 제한되지 않는 방법에 의해 균질화될 수 있다. 샘플을 이어서 초음파 처리하고, 초음파 처리 완충제에 다시 재현탁할 수 있다. 그런 다음, 초음파 사이클(예를 들어, 최고 출력에서 30초)을 충분한 가용화된 염색질을 얻기 위해 필요한 만큼의 많은 사이클(예를 들어, 10 사이클, 20 사이클, 30 사이클, 40 사이클)을 반복할 수 있다. 이어서, 가용성 분획은 회수될 수 있다.

[0102] 가용화 후, 샘플은 가용화된 염색질의 회수(예를 들어, 고상 가역적 고정(SPRI) 비드에 대한 결합에 의한), 시퀀싱 라이브러리의 제조, 예컨대 본원에서 설명되는 시카고 라이브러리(예를 들어, 핵산의 절단, 태그 부착 및 라이게이션), 시퀀싱(예를 들어, 장범위 정보 포함) 및 서열 어셈블리와 같은, 본원에서 설명되는 방법에 따라 추가로 처리될 수 있다.

[0103] **크기 선택**

[0104] 보존된(예를 들어, FFPE) 생물학적 샘플로부터 수득된 핵산은 분석에 적합한 단편을 생성하기 위해 단편화될 수 있다. 주형 핵산은 다양한 기계적, 화학적 및/또는 효소적 방법을 사용하여 원하는 길이로 단편화되거나 전단될 수 있다. DNA는 초음파 처리, 예를 들어 코바리스(Covaris) 방법, DNase에 대한 짧은 노출, 또는 하나 이상의 제한 효소의 혼합물, 또는 트랜스포사제 또는 닉 형성 효소의 사용을 통해 무작위로 전단될 수 있다. RNA는 RNase, 열 + 마그네슘에 대한 짧은 노출, 또는 전단에 의해 단편화될 수 있다. RNA는 cDNA로 전환될 수 있다. 단편화가 사용되는 경우, RNA는 단편화 전 또는 후에 cDNA로 전환될 수 있다. 일부 실시양태에서, 생물학적 샘플로부터의 핵산은 초음파 처리에 의해 단편화된다. 다른 실시양태에서, 핵산은 유체 전단 기기에 의해 단편화된다. 일반적으로, 개개의 핵산 주형 분자는 약 2 kb 염기 내지 약 40 kb일 수 있다. 다양한 실시양태에서, 핵산은 약 6 kb-10 kb 단편일 수 있다. 핵산 분자는 단일 가닥, 이중 가닥 또는 단일 가닥 영역이 있는 이중 가닥(예를 들어, 줄기 및 루프 구조)일 수 있다.

[0105] 일부 실시양태에서, 가교결합된 DNA 분자는 크기 선택 단계를 거칠 수 있다. 핵산의 크기 선택은 특정 크기 미만 또는 초과의 가교결합된 DNA 분자로 수행될 수 있다. 또한, 크기 선택은 가교결합의 빈도 및/또는 단편화 방법에 의해, 예를 들어 빈번하거나 드문 절단 제한 효소의 선택에 의해 영향을 받을 수 있다. 일부 실시양태에서, 약 1 kb 내지 5 Mb, 약 5 kb 내지 5 Mb, 약 5 kb 내지 2 Mb, 약 10 kb 내지 2 Mb, 약 10 kb 내지 1 Mb, 20 kb 내지 1 Mb, 약 20 kb 내지 500 kb, 약 50 kb 내지 500 kb, 약 50 kb 내지 200 kb, 약 60 kb 내지 200 kb, 약 60 kb 내지 150 kb, 약 80 kb 내지 150 kb, 약 80 kb 내지 120 kb, 또는 약 100 kb 내지 120 kb 범위 또는 상기 임의의 값에 의해 한정되는 임의의 범위(예를 들어, 약 150 kb 내지 1 Mb)의 가교결합된 DNA 분자를 포함하여 조성물을 제조할 수 있다.

[0106] 일부 실시양태에서, 샘플 폴리뉴클레오티드는 하나 이상의 특정 크기 범위(들)의 단편화된 DNA 분자의 집단으로 단편화된다. 일부 실시양태에서, 단편은 적어도 약 1개, 약 2개, 약 5개, 약 10개, 약 20개, 약 50개, 약 100개, 약 200개, 약 500개, 약 1000개, 약 2000개, 약 5000개, 약 10,000개, 약 20,000개, 약 50,000개, 약 100,000개, 약 200,000개, 약 500,000개, 약 1,000,000개, 약 2,000,000개, 약 5,000,000개, 약 10,000,000개 또는 그 초과의 출발 DNA의 게놈 등가물로부터 생성될 수 있다. 단편화는 화학적, 효소적 및 기계적 단편화를 포함하여 관련 기술 분야에 공지된 방법에 의해 수행될 수 있다. 일부 실시양태에서, 단편의 평균 길이는 약 10 내지 약 10,000개, 약 20,000개, 약 30,000개, 약 40,000개, 약 50,000개, 약 60,000개, 약 70,000개, 약 80,000개, 약 90,000개, 약 100,000개, 약 150,000개, 약 200,000개, 약 300,000개, 약 400,000개, 약 500,000개, 약 600,000개, 약 700,000개, 약 800,000개, 약 900,000개, 약 1,000,000개, 약 2,000,000개, 약 5,000,000개, 약 10,000,000개 또는 그 초과의 뉴클레오티드이다. 일부 실시양태에서, 단편의 평균 길이는 약 1 kb 내지 약 10 Mb이다. 일부 실시양태에서, 단편의 평균 길이는 약 1 kb 내지 5 Mb, 약 5 kb 내지 5 Mb, 약 5 kb 내지 2 Mb, 약 10 kb 내지 2 Mb, 약 10 kb 내지 1 Mb, 약 20 kb 내지 1 Mb, 약 20 kb 내지 500 kb, 약 50 kb 내지 500 kb, 약 50 kb 내지 200 kb, 약 60 kb 내지 200 kb, 약 60 kb 내지 150 kb, 약 80 kb 내지 150 kb, 약 80 kb 내지 120 kb, 또는 약 100 kb 내지 120 kb, 또는 상기 임의의 값에 의해 한정되는 임의의 범위(예를 들어, 약 60 내지 120 kb)이다. 일부 실시양태에서, 단편의 평균 길이는 약 10 Mb 미만, 약 5 Mb 미만, 약 1 Mb 미만, 약 500 kb 미만, 약 200 kb 미만, 약 100 kb 미만, 또는 약 50 kb 미만이다. 다른 실시양태에서, 단편의 평균 길이는 약 5 kb 초과, 약 10 kb 초과, 약 50 kb 초과, 약 100 kb 초과, 약 200 kb 초과, 약 500 kb 초과, 약 1 Mb 초과, 약 5 Mb 초과, 또는 약 10 Mb 초과이다. 일부 실시양태에서, 단편화는 샘플 DNA 분자를 음향 초음파 처리를 거치는 단계를 포함하는 기계적 방식으로 수행된다. 일부 실시양태에서, 단편화는 하나 이상의 효

소가 이중 가닥 핵산의 파단을 생성하기에 적합한 조건 하에서 샘플 DNA 분자를 하나 이상의 효소로 처리하는 단계를 포함한다. DNA 단편을 생성하는 데 유용한 효소의 예는 서열 특이적 및 비특이적 뉴클레아제를 포함한다. 뉴클레아제의 비제한적인 예는 DNase I, 단편화 효소, 제한 엔도뉴클레아제, 이들의 변이체, 및 이들의 조합물을 포함한다. 예를 들어, DNase I을 사용한 소화는 Mg^{++}가 없을 경우와 Mn^{++}가 존재할 경우 DNA에 무작위적인 이중 가닥 파단을 유도할 수 있다. 일부 실시양태에서, 단편화는 샘플 DNA 분자를 하나 이상의 제한 엔도뉴클레아제로 처리하는 단계를 포함한다. 단편화는 5' 오버행, 3' 오버행, 평활 말단, 또는 이들의 조합물을 갖는 단편을 생성할 수 있다. 일부 실시양태에서, 단편화가 하나 이상의 제한 엔도뉴클레아제의 사용을 포함할 때와 같이, 샘플 DNA 분자의 절단은 예측 가능한 서열을 갖는 오버행을 남긴다. 일부 실시양태에서, 상기 방법은 컬럼 정제 또는 아가로스 겔로부터의 단리와 같은 표준 방법을 통해서 단편의 크기를 선택하는 단계를 포함한다.

[0107] **시퀀싱 라이브러리 제조**

[0108] 도 1b는 염색질 기반 차세대 시퀀싱(NGS) 라이브러리 제조(예를 들어, "시카고")의 예시적인 개략도를 보여준다. 제1 단계(111)에서, 염색질 뉴클레아제(파란색 원)는 가교결합(적색 라인)되어 염색질 응집체를 형성한다. 제2 단계(112)에서, 염색질 응집체는 제한 엔도뉴클레아제로 절단된다. 제3 단계(113)에서, 절단된 말단은 평활 말단이고, 라이게이션되고, 표시된다(예를 들어, 바이오틴으로)(작은 녹색 원). 제4 단계(114)에서, 평활 말단은 무작위로 라이게이션되어, 단범위, 중범위 및 장범위 회합을 형성한다(붉은 별표는 라이게이션 사건을 나타낸다). 제5 단계(115)에서, 가교결합은 파괴되고, DNA를 정제하고, 마커 풀다운(pulldown)을 위해 유용한 라이게이션 함유 단편을 선택한다. 이어서, 종래의 시퀀싱 라이브러리 제조를 수행할 수 있다. 생성되는 리드쌍은 투입 DNA의 최대 크기까지 게놈 거리를 확장할 수 있다. 이러한 라이브러리는 염색체 규모의 초대형 스캐폴드와 매우 인접한 게놈 어셈블리를 구축하기 위해 사용할 수 있다.

[0109] 도 1c는 보존된 샘플(예를 들어, FFPE 샘플)으로부터 염색질 추출 및 라이브러리 제조(예를 들어, "시카고" 라이브러리 제조)를 위한 작업 흐름의 예시적인 개략도를 보여준다. 보존된 샘플은 고정된 염색질을 추출하기 위해 처리될 수 있고, 이어서 추출된 염색질은 장범위 게놈 링키지 정보를 생성하고 시퀀싱하는 방법에 적용될 수 있다. 예를 들어, 보존된 샘플(121)은 추출되고(122) 단편화된(예를 들어, DpnII와 같은 제한 효소를 사용하여) 염색질을 가질 수 있다. 염색질은 가교결합(123)을 포함할 수 있다. 오버행(예를 들어, 4 bp 5' 오버행)은 비오티닐화된 뉴클레오티드(124)를 포함한 뉴클레오티드 혼합물로 채워질 수 있다. 이어서, 평활 말단을 라이게이션될 수 있고(125), 마커(예를 들어, 비오틴)를 풀다운할 수 있다(예를 들어, 스트렙타비딘을 사용하여)(126). 이어서, 비표지된(예를 들어, 비-비오티닐화된) 평활 말단을 제거할 수 있고, 시퀀싱 어댑터(예를 들어, Illumina 시퀀싱 어댑터, Pacific Biosciences 시퀀싱 어댑터, 나노포어(nanopore) 시퀀싱 어댑터)를 부착할 수 있으며, 시퀀싱 라이브러리(127)를 제조할 수 있다. 라이브러리는 비오티닐화된 라이게이션된 연결부를 포함하는 분자에 대해 풍부화되고, 증폭(예를 들어, PCR에 의해)되고, 시퀀싱된다(예를 들어, Illumina 시퀀서, 예컨대 MiSeq 또는 HiSeq를 사용하여, Pacific Biosciences 긴 리드 시퀀서를 사용하여, 또는 나노포어 시퀀서, 예컨대 Oxford 나노포어 또는 Genia를 사용하여). 일부 경우에, 퍼시픽 바이오사이언시스(Pacific Biosciences) 또는 나노포어 시퀀서와 같은 긴 리드 시퀀서를 사용하는 경우와 같이, 다수의 분자를 시퀀싱하기 전에 보다 긴 분자로 연결(예를 들어, 라이게이션)할 수 있다.

[0110] 풍부화는 관심 유전자 영역에 대해 표지된 뉴클레오티드(예를 들어, 비오티닐화된 뉴클레오티드, 후성적으로 변형된 뉴클레오티드)에 대한 풍부화에 대한 대안으로 또는 그에 추가하여 수행될 수 있다. 예를 들어, 융합 유전자의 알려진 관련 절반을 표적화에 의한 것과 같이, 샘플 또는 라이브러리를 융합 유전자에 대해 풍부하게 할 수 있다. 본원에서 논의되는 다른 유전적 및 게놈 특징이 또한 풍부화를 위해 표적화될 수 있다.

[0111] 많은 경우, 정화 과정의 일부로서 이전에 수득된 샘플(예컨대, FFPE 샘플)에 고정제는 첨가되지 않는다. 오히려, 본원에서 단리된 DNA-단백질(예를 들어, 염색질) 복합체를 안정화하기 위해, 원래의 샘플 보존 과정에 따라 이전에 생성된 가교결합에 의존할 수 있고, 추출 과정은 상당한 양의 새로운 복합체를 생성하기보다는 연결된 복합체를 보존한다. 이어서, 용해 완충제 중에 가용화된 샘플의 분획은 본원에서 개시되는 임의의 방법으로 처리된다.

[0112] 대안으로, 일부 실시양태에서, DNA를 포함하는 보존된 샘플(예를 들어, FFPE 보존된 샘플)로부터 추출된 고품질 핵산으로부터 생성된 재구성된 염색질로부터의 리드쌍 라이브러리를 생성하기 위해 시험관내 근접 라이게이션(예를 들어, 시카고 시험관내 근접 라이게이션) 또는 다른 단백질-DNA 복합 태그 부착 방법이 사용된다. 예를 들어, 보존된 샘플(예를 들어, FFPE 샘플)을 추출 과정에서 DNA 손상을 최소화하기 위해 DNA와 같은 핵산을 추

출하도록 처리할 수 있다. 일부 경우에, 단리된 네이키드 DNA 손상을 감소시키기 위해, 볼텍싱, 전단, 비등, 고온 인큐베이션 또는 DNase 관련 효소 처리 중 하나 이상이 핵산 추출 프로토콜로부터 제외된다. 회수된 단리된 DNA는 물리적 링키지, 페이즈, 또는 게놈의 구조적 정보를 보존하기에 충분한 품질을 갖는 것일 수 있다. 추출된 핵산은 희석되고, DNA/단백질 복합체가 단일 DNA 분자 및 적어도 하나의 DNA 결합 모이어티를 포함하도록 재구성된 염색질을 생성하기 위해 사용될 수 있다(예를 들어, 그 전부가 본원에 참고로 포함된, 2014년 8월 7일 공개된 PCT 공개 WO2014/121091에, 또는 그 전부가 본원에 참고로 포함된, 2016년 2월 4일 공개된 PCT 공개 WO2016/019360에 기재된 바와 같은 방법을 사용하여). 재구성된 염색질은 그들의 공통적인 포스포디에스테르 백본과는 관계없이 동일한 DNA 분자 내의 DNA 서열의 근접 정보를 보존하기 위해, 예컨대 포름알데히드를 사용하여 가교결합될 수 있다. 중요하게는, 가교결합은 보존된 샘플로부터 단리된 후, 보존된 샘플(예컨대, FFPE 샘플)로부터 추출된 DNA에 대해 수행될 수 있다. DNA-단백질 복합체의 단리와 관련하여 상기 논의된 바와 같이, 많은 경우에, 가교결합제는 단리 과정 동안 첨가되지 않는다. 이들 가교결합된 재구성된 복합체는 예컨대 비오틴, 메틸화, 술페닐화, 아세틸화 또는 다른 염기 변형으로 표지될 수 있고, 이어서 예컨대 비오틴 표지의 경우 스트렙타비딘 비드를 사용하여 단리될 수 있다. 이어서, 단리된 복합체는 제한 효소로 소화하여 자유로운 점착성 말단을 생성한 후, 이 말단은 예컨대 비오티닐화된 뉴클레오티드 또는 언급한 바와 같은 다른 뉴클레오티드로 표지된 뉴클레오티드로 채워진다.

[0113] 기존의 것이든(예를 들어, 보존된 샘플의 분해에 의한) 또는 본원에서 개시되는 프로토콜(예를 들어, 효소적 또는 물리적 절단)의 결과이든 관계 없이, DNA:단백질 복합체 내의 노출된 DNA 말단은 라이게이션되어, 동일한 DNA 분자 내의 DNA 서열 사이에 쌍을 이룬 말단을 생성할 수 있다. 이러한 라이게이션된 쌍을 이룬 말단은 DNA 분자에서 원래 서로 인접하지 않을 수 있다. 쌍을 이룬 말단은 일부 경우에 점착성 말단을 채운 결과로서 평활 말단이 될 수 있다.

[0114] 대안으로 또는 추가로, 노출된 핵산 복합체 말단은 본원에서 논의되는 펑츄에이션(punctuation) 올리고뉴클레오티드를 통해 서로 라이게이션될 수 있거나, 또는 핵산 단편이 공통적인 DNA 단백질 복합체에 확인 가능하게 매핑되도록 올리고뉴클레오티드 태그의 집단을 사용하여 태그 부착될 수 있다. 일부 경우에, 쌍을 이룬 말단 리드는 직접 라이게이션되는 DNA-복합체의 절단된 말단이 아니라 공통적인 펑츄에이션 올리고뉴클레오티드에 연결된 절단된 말단으로부터 생성된다. 펑츄에이션 올리고뉴클레오티드는 페이즈 보존 재배열을 겪는 샘플 분자의 2개의 절단된 내부 말단을 가교시키도록 표적 폴리뉴클레오티드에 연결될 수 있는 임의의 올리고뉴클레오티드를 포함한다. 펑츄에이션 올리고뉴클레오티드는 DNA, RNA, 뉴클레오티드 유사체, 비표준 뉴클레오티드, 표지된 뉴클레오티드, 변형된 뉴클레오티드, 또는 이들의 조합물을 포함할 수 있다. 많은 예에서, 이중 가닥 펑츄에이션 올리고뉴클레오티드는 서로 혼성화된 2개의 별개의 올리고뉴클레오티드("올리고뉴클레오티드 이중체(duplex)"로도 언급됨)를 포함하고, 혼성화는 하나 이상의 평활 말단, 하나 이상의 3' 오버행, 하나 이상의 5' 오버행, 미스매치된 및/또는 쌍을 이루지 않은 뉴클레오티드에 기인한 하나 이상의 돌출부(bulge), 또는 이들의 임의의 조합을 생성할 수 있다. 일부 예에서, 상이한 펑츄에이션 올리고뉴클레오티드는 순차적인 반응으로 또는 동시에 표적 폴리뉴클레오티드에 연결된다. 예를 들어, 제1 및 제2 펑츄에이션 올리고뉴클레오티드는 동일한 반응에 첨가될 수 있다. 대안으로, 펑츄에이션 올리고 집단은 일부 경우에 균일하다. 펑츄에이션 분자 및 게놈의 구조적 및 근접성 정보의 보존 및 결정에 사용되는 방법은 이전에 문헌에 기재된 바 있다(모두 그 전부가 본원에 참고로 포함된 미국 특허 가출원 제62/298906호, 제62/298966호 및 제62/305957호). 일부 펑츄에이션 올리고뉴클레오티드는 펑츄에이션 올리고뉴클레오티드를 포함하는 라이브러리의 단편이 쉽게 단리될 수 있도록, 단리를 용이하게 하는 태그 또는 표지, 예컨대 비오틴 태그를 포함한다. 대체 태그는 메틸화, 아세틸화 또는 다른 염기 변형을 포함하고 이로 제한되지 않는다. 일반적으로, 펑츄에이션 올리고뉴클레오티드는 노출된 핵산 말단에 라이게이션되지만, 펑츄에이션 올리고뉴클레오티드를 라이브러리 내로 도입하는 대안적인 방법이 또한 고려된다.

[0115] 점착성 말단을 채우기 위해 사용되는 것과 같은 뉴클레오티드를 표지할 수 있다. 표지된 뉴클레오티드는 비오티닐화되거나, 황산화되거나, 형광단에 부착되거나, 탈인산화되거나 또는 임의의 다른 수의 뉴클레오티드 변형을 포함할 수 있다. 뉴클레오티드 변형은 또한 후성적 변형, 예컨대 메틸화(예를 들어, 5-mC, 5-hmC, 5-fC, 5-caC, 4-mC, 6-mA, 8-oxoG, 8-oxoA)를 포함할 수 있다. 표지 또는 변형은 시퀀싱 동안 검출 가능한 것, 예컨대 나노포어 시퀀싱에 의해 검출 가능한 후성적 변형으로부터 선택될 수 있고; 이러한 방식으로 라이게이션 접합부의 위치를 시퀀싱 동안 검출할 수 있다. 이러한 표지 또는 변형은 또한 결합 또는 풍부화를 위해 표적화될 수 있고; 예를 들어, 메틸-시토신을 표적화하는 항체는 메틸-시토신으로 채워진 평활 말단을 포획, 표적화, 결합 또는 표지하기 위해 사용될 수 있다. 비천연 뉴클레오티드, 비표준 또는 변형된 뉴클레오티드 및 핵산 유사체는 또한 평활 말단의 채워진 위치를 표지하기 위해 사용될 수 있다. 비표준 또는 변형된 뉴클레오티드는 슈도우리딘(Ψ), 디히드로우리딘(D), 이노신(I), 7-메틸구아노신(m7G), 잔틴, 히포잔틴, 퓨린, 2,6-디아미노퓨린, 및 6,8-

디아미노퓨린을 포함할 수 있다. 핵산 유사체는 펩티드 핵산(PNA), 모르폴리노 및 잠금 핵산(LNA: locked nucleic acid), 글리콜 핵산(GNA) 및 트레오스 핵산(TNA)을 포함할 수 있다. 일부 경우에, 오버행은 비표지된 dNTP, 예컨대 비오틴이 없는 dNTP로 채워진다. 트랜스포존에 의한 절단과 같은 일부 경우에, 채우기를 필요로 하지 않는 평활 말단이 생성된다. 이러한 자유로운 평활 말단은 트랜스포사제가 2개의 연결되지 않은 펑츄에이션 올리고뉴클레오티드를 삽입할 때 생성된다. 그러나, 펑츄에이션 올리고뉴클레오티드는 원하는 바와 같은 점착성 또는 평활 말단을 갖도록 합성될 수 있다. 히스톤과 같은 샘플 핵산과 회합된 단백질도 변형될 수 있다. 예를 들어, 히스톤은 아세틸화(예를 들어, 리신 잔기에서의) 및/또는 메틸화(예를 들어, 리신 및 아르기닌 잔기에서의)될 수 있다.

[0116] 일부 실시양태에서, Hi-C 또는 다른 라이게이션 또는 태그 부착 매개 방법을 사용하여 가교결합된 천연 생성 염색질, 예를 들어 샘플 보존에 따라 가교결합된 염색질으로부터 리드쌍 라이브러리를 생성할 수 있다. DNA는 보존 과정 동안 천연 염색질 구조를 보존하기 위해, 예를 들어 포름알데히드를 사용하여 가교결합될 수 있다. 추출은 가교결합된 DNA-단백질 구조를 파괴하지 않으면서 임의의 샘플 보존제 또는 고정제, 예컨대 파라핀으로부터 상기 DNA-단백질 구조를 분리함으로써, 포스포디에스테르 백본과는 관계없이 DNA 분자 사이의 근접성 정보를 보존하기 위해 상기한 바와 같이 수행될 수 있다. 이러한 가교결합된 구조는 제한 효소로 소화하여 자유로운 점착성 말단을 생성한 후, 이 말단은 예컨대 비오틴 표지된 뉴클레오티드와 같은 태그 부착된 뉴클레오티드로 채워진다. 생성된 평활 말단을 함께 라이게이션하여 DNA 단편의 쌍을 이룬 말단을 생성할 수 있다. 이 쌍을 이룬 말단은 염색질 구조에서 서로 근접하여 위치하는 DNA 분자를 나타낸다. Hi-C 방법 및 변형은 관련 기술 분야에 알려져 있다(그 전부가 본원에 참고로 포함된 [Liberman-Aiden et al., 2009, Science 326, 289]; 그 전부가 본원에 참고로 포함된 US20130096009).

[0117] 쌍을 이룬 말단은 예컨대 효소 소화(예를 들어, 프로테이나제, 예컨대 프로테이나제 K로)에 의해 염색질 단백질로부터 방출될 수 있다. 방출된 쌍을 이룬 말단은 표지된 뉴클레오티드만이 라이게이션된 쌍을 이룬 말단 사이에 존재하도록 엑소뉴클레아제로 처리하여 남아있는 자유 말단으로부터 표지된 뉴클레오티드를 제거할 수 있다. 이어서, 상기 쌍을 이룬 말단은 예를 들어 비오틴 표지의 경우 스트렙타비딘 비드를 사용하여 정제될 수 있다. 정제는 또한 다른 수단에 의해, 예컨대 SPRI 비드(예를 들어, 카르복실화된 비드)를 사용하여 또는 전기영동(예를 들어, 겔 전기영동, 모세관 전기영동)을 통해 수행될 수 있다. 이어서, 쌍을 이룬 말단은 시퀀싱을 위해 준비될 수 있다. 예를 들어, 쌍을 이룬 말단을 시퀀싱 어댑터에 부착한 후, 리드쌍 라이브러리를 생성하기 위해 시퀀싱할 수 있다. 시카고 시험관내 근접 라이게이션 방법은 이전에 문헌에 기재된 바 있다(예를 들어, 그 전부가 본원에 참고로 포함된 미국 특허 출원 공개 제20140220587호; 그 전부가 본원에 참고로 포함된 미국 특허 출원 공개 제20150363550호 참조).

[0118] 예시적인 실시양태에서, 섹션당 약 3 x 10^5개의 세포를 갖는, 15-20 마이크로미터 두께의 섹션에 FFPE에 이전에 포매된 세포로부터 라이브러리가 생성된다. 대안으로, FFPE에 포매된 세포는 섹션당 약 10^3, 10^4, 10^5, 10^5, 또는 10^7개의 세포를 갖는, 1-5, 5-10, 10-15, 15-20, 25-30, 35-40, 또는 45-50 마이크로미터 두께의 섹션에 제공된다. 일부 경우에, 샘플은 AJ GIAB('Gonome In A Bottle') 샘플 GM24149(부계) 및 GM24385(아들)이다. 섹션을 용매, 예를 들어 크실렌, 톨루엔 또는 벤젠으로 세척하여 포매 물질을 제거한다. 에탄올 용액으로 섹션을 세척하여 용매를 제거하고, 일부 경우에는 100% 에탄올을 사용하여 섹션을 세척한다. 이어서, 파라핀이 없는 조직 샘플을 완충제, 예를 들어 세제 완충제 내에 가용화한다. 샘플 내의 핵산은 엔도뉴클레아제, 예를 들어 제한 효소, 예컨대 Mbo1로 소화된다. 평활 말단은 제한 효소 소화에 의해 생성되는 오버행을 DNA 폴리머라제 및 뉴클레오티드, 예컨대 비오티닐화된 dNTP를 사용하여 채움으로써 소화된 핵산에서 생성된다. 평활 말단은 평활 말단 라이게이션에 유리한 반응에서 DNA 리가제, 예를 들어 T4 DNA 리가제를 사용하여 함께 라이게이션되어 DNA의 비오티닐화된 단편을 생성한다. 이들 단편은 시퀀싱 반응에 사용하기 위해 제조된다.

[0119] **시퀀싱**

[0120] 또한, 물리적 링키지 정보와 같은 게놈의 구조적 정보를 갖는 핵산 시퀀싱 라이브러리를 생성하는 방법 및 조성물이 본원에서 개시된다. DNA 복합체는 FFPE 유래 핵산 샘플과 같은 보존된 샘플로부터 생성된다. 쌍을 이룬 말단, 라이게이션 접합부, 펑츄에이션 말단 또는 통상적으로 태그 부착된 말단은 제1 세그먼트 및 제2 세그먼트가 임의의 포스포디에스테르 백본 결합과는 독립적으로 함께 유지되고, 노출된 말단은 태그 부착되고, 태그 접합부는 단리되도록 핵산 복합체의 단리를 통해 생성된다. 태그 부착은 접합부의 어느 한쪽 상의 서열이 게놈 스캐폴드 상의 먼 위 위치에 상응하는 콘티그에 매핑되거나, 비스캐폴딩되거나, 또는 비재배열된 게놈에서 상이한 염색체에 매핑된다는 사실로부터 접합부가 확인될 수 있도록, 제2의 노출된 말단을 이용하여 하나의 노출된 말단

에 직접 태그를 부착하는 단계를 포함한다. 대안으로, 태그 부착은 펑츄에이션 올리고를 사용하여 노출된 말단을 연결하거나, 또는 태그 부착된 말단에 인접한 서열이 공통적인 DNA 복합체, 및 이에 따라 DNA 복합체가 그로부터 생성된 공급원 핵산의 공통적인 페이즈에 확실하게 매핑되도록 공통적인 올리고 태그를 복합체의 노출된 말단에 부가하는 단계를 포함한다.

[0121] 쌍을 이룬 말단, 콘카테머화된(concatamerized) 쌍을 이룬 말단 또는 펑츄에이션된 분자는 적절한 짧은 리드 또는 긴 리드 시퀀싱 기술 플랫폼을 사용하여 시퀀싱되고, 이어서 서열 리드가 분석된다.

[0122] 일부 경우에, 다수의 쌍을 이룬 말단 분자는 본원에서 설명되는 바와 같이 생성되고, 이어서 짧은 리드 시퀀싱 기술을 사용하여 시퀀싱된다. 이러한 경우, 쌍을 이룬 말단 라이게이션 접합부를 가로지르는 짧은 서열 리드가 생성되거나, 쌍을 이룬 말단 단편의 각각의 말단으로부터의 끝에서 짧은 리드가 생성되어 리드 쌍을 만든다. 제1 및 제2 핵산 세그먼트로부터의 서열이 단일 서열 리드 또는 리드쌍에서 검출되는 경우, 제1 및 제2 핵산 세그먼트는 투입 DNA 샘플에서 동일한 DNA 분자 상에 같은 페이즈로 존재하는 것으로 결정된다. 이 경우, 생성되는 서열 라이브러리는 DNA 세그먼트에 대한 페이즈 및 구조적 정보를 생성한다.

[0123] 제시된 펑츄에이션된 분자 서열 리드 또는 리드쌍에 대해, 펑츄에이션 요소에 의해 국소적으로 방해되지 않는 서열 세그먼트가 관찰된다. 이 세그먼트 내의 서열은 같은 페이즈로 존재하고 국소적으로 올바르게 정렬되고 배향된 것으로 추정된다. 세그먼트는 펑츄에이션 올리고에 의해 분리된 것으로 관찰된다. 펑츄에이션 올리고의 어느 한쪽 상의 세그먼트는 통상적인 샘플 핵산 분자에서 서로 같은 페이즈로 존재하지만 펑츄에이션 분자에서 서로에 대해 올바르게 정렬되고 배향되지 않은 것으로 추정된다. 재배열의 이점은 서로 멀리 떨어진 위치에 있는 세그멘트가 종종 근접하게 되어, 샘플 분자에서 이들이 서열 페이징이 어려운 동일한 거리의 먼 거리로 분리되어 있는 경우에도 이들이 공통의 리드로 판독되고, 공통 페이즈에 확실하게 할당된다는 것이다. 또 다른 이점은 페이즈 정보 이외에, 일부 경우에 새로운 서열 어셈블리를 수행하기에 충분한 콘티그 정보가 일부 경우에 결정되도록, 세그먼트 서열 자체가 원래의 샘플 서열의 대부분, 실질적으로 전부 또는 전부를 포함한다는 것이다. 이 새로운 서열은 신규한 스캐폴드 또는 콘티그 세트를 생성하거나, 이전에 또는 독립적으로 생성된 콘티그 또는 스캐폴드 서열 세트를 증가시키기 위해 선택적으로 사용된다.

[0124] 일부 경우에, 다수의 펑츄에이션된 DNA 분자가 본원에서 개시되는 바와 같이 생성되고, 하나의 긴 핵산 분자로 콘카테머화되거나 또는 하나의 재배열된 긴 분자로서 전단 또는 절단없이 보존된 후, 긴 리드 시퀀싱 기술을 사용하여 시퀀싱된다. 각각의 펑츄에이션된 분자를 시퀀싱하고, 서열 리드를 분석한다. 바람직한 예에서, 서열 리드는 서열 반응에 대해 평균적으로 10 kb이다. 다른 예에서, 서열 리드는 평균적으로 약 5 kb, 6 kb, 7 kb, 8 kb, 9 kb, 10 kb, 11 kb, 12 kb, 13 kb, 14 kb, 15 kb, 16 kb, 17 kb, 18 kb, 19 kb, 20 kb, 21 kb, 22 kb, 25 kb, 30 kb, 35 kb, 40 kb, 또는 그 초과이다. 바람직한 예에서, 펑츄에이션 올리고 서열에 의해 연결된, 제1 세그멘트의 적어도 500개 염기 및 제2 세그멘트의 500개 염기를 포함하는 서열 리드가 확인된다. 다른 예에서, 서열 리드는 제1 DNA 세그먼트의 적어도 약 100개 염기, 200개 염기, 300개 염기, 400개 염기, 500개 염기, 600개 염기, 700개 염기, 800개 염기, 900개 염기, 1000개 염기 또는 그 초과의 염기, 및 제2 DNA 세그먼트의 적어도 약 100개 염기, 200개 염기, 300개 염기, 400개 염기, 500개 염기, 600개 염기, 700개 염기, 800개 염기, 900개 염기, 1000개 염기 또는 그 초과의 염기를 포함한다. 일부 예에서, 제1 및 제2 세그먼트 서열은 스캐폴드 게놈에 매핑되고, 적어도 100 kb만큼 분리된 콘티그에 매핑되는 것으로 밝혀졌다. 다른 예에서, 분리 거리는 8 kb, 9 kb, 10 kb, 12.5 kb, 15 kb, 17.5 kb, 20 kb, 25 kb, 30 kb, 35 kb, 40 kb, 45 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb, 100 kb, 125 kb, 150 kb, 200 kb, 300 kb, 400 kb, 500 kb, 600 kb, 700 kb, 800 kb, 900 kb, 1 Mb 또는 그 초과이다. 대부분의 경우에, 제1 콘티그 및 제2 콘티그는 각각 하나의 이형접합성 위치를 포함하고, 그의 페이즈는 스캐폴드에서 결정되지 않는다. 바람직한 예에서, 제1 콘티그의 이형접합성 위치는 긴 리드의 제1 세그먼트에 의해 걸쳐지고, 제2 콘티그의 이형접합성 위치는 긴 리드의 제2 세그먼트에 의해 걸쳐진다. 이러한 경우에, 리드는 각각 그의 콘티그의 각각의 이형접합성 영역에 걸쳐 있고, 리드 세그먼트의 서열은 제1 콘티그의 제1 대립유전자 및 제2 콘티그의 제1 대립유전자가 같은 페이즈로 존재함을 나타낸다. 제1 및 제2 핵산 세그먼트로부터의 서열이 하나의 긴 서열 리드에서 검출되는 경우, 제1 및 제2 핵산 세그먼트는 투입 DNA 샘플에서 동일한 DNA 분자 상에 포함되는 것으로 결정된다. 이들 실시양태에서, 본원에서 개시되는 방법 및 조성물에 의해 생성된 핵산 서열 라이브러리는 게놈 스캐폴드 상에서 서로 멀리 떨어져 위치하는 콘티그에 대한 페이즈 정보를 제공한다.

[0125] 대안으로, 다수의 쌍을 이룬 말단 분자가 본원에서 설명되는 바와 같이 생성되고, 이어서 긴 리드 시퀀싱 기술을 사용하여 시퀀싱된다. 일부 경우에, 라이브러리에 대한 평균 리드 길이는 약 1 kb로 결정된다. 다른 경우에, 라이브러리에 대한 평균 리드 길이는 약 100 bp, 200 bp, 300 bp, 400 bp, 500 bp, 600 bp, 700 bp, 800 bp,

900 bp, 1 kb, 1.1 kb, 1.2 kb, 1.3 kb, 1.4 kb, 1.5 kb, 또는 그 초과이다. 대부분의 예에서, 쌍을 이룬 말단 분자는 투입 DNA 샘플 내에서 같은 페이즈로 존재하고 10 kb 초과의 거리만큼 분리된 제1 DNA 세그먼트 및 제2 세그먼트를 포함한다. 일부 예에서, 상기 2개의 DNA 세그먼트 사이의 분리 거리는 약 5 kb, 6 kb, 7 kb, 8 kb, 9 kb, 10 kb, 11 kb, 12 kb, 13 kb, 14 kb, 15 kb, 20 kb, 23 kb, 25 kb, 30 kb, 32 kb, 35 kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 750 kb, 1 Mb, 또는 그 초과이다. 대부분의 경우, 서열 리드는 쌍을 이룬 말단 분자로부터 생성되며, 이들 중 일부는 제1 핵산 세그먼트로부터의 서열의 적어도 300개 염기 및 제2 핵산 세그먼트로부터의 서열의 적어도 300개 염기를 포함한다. 다른 예에서, 서열 리드는 제1 DNA 세그먼트의 적어도 약 50개 염기, 100개 염기, 150개 염기, 200개 염기, 250개 염기, 300개 염기, 350개 염기, 400개 염기, 450개 염기, 500개 염기, 550개 염기, 600개 염기, 650개 염기, 700개 염기, 750개 염기, 800개 염기 또는 그 초과의 염기, 및 제2 DNA 세그먼트의 적어도 약 50개 염기, 100개 염기, 150개 염기, 200개 염기, 250개 염기, 300개 염기, 350개 염기, 400개 염기, 450개 염기, 500개 염기, 550개 염기, 600개 염기, 650개 염기, 700개 염기, 750개 염기, 800개 염기 또는 그 초과의 염기를 포함한다. 제1 및 제2 핵산 세그먼트로부터의 서열이 단일 서열 리드 또는 리드쌍에서 검출되는 경우, 제1 및 제2 핵산 세그먼트는 투입 DNA 샘플에서 동일한 DNA 분자 상에 같은 페이즈로 존재하는 것으로 결정된다. 이 경우, 생성된 서열 라이브러리는 이를 시퀀싱하기 위해 사용된 시퀀싱 기술의 리드 길이보다 긴 거리만큼 핵산 샘플에서 분리된 DNA 세그먼트에 대한 페이즈 정보를 생성한다.

[0126] 다양한 실시양태에서, 본원에서 설명되는 또는 관련 기술 분야에 공지된 적합한 시퀀싱 방법을 사용하여 샘플 내의 핵산 분자로부터 서열 정보를 얻는다. 시퀀싱은 관련 기술 분야에 잘 알려진 고전적인 생어(Sanger) 시퀀싱 방법을 통해 수행될 수 있다. 또한, 시퀀싱은 시퀀싱되는 뉴클레오티드가 성장하는 가닥에 도입될 때 또는 도입 직후에 검출되도록 하는, 예컨대 실시간으로 또는 실질적으로 실시간으로 서열이 검출되도록 하는 고처리량 시스템을 이용하여 수행될 수 있다. 일부 경우에, 고처리량 시퀀싱은 시간당 적어도 1,000개, 적어도 5,000개, 적어도 10,000개, 적어도 20,000개, 적어도 30,000개, 적어도 40,000개, 적어도 50,000개, 적어도 100,000개 또는 적어도 500,000개의 서열 리드를 생성하며, 시퀀싱 리드는 리드당 약 50개, 약 60개, 약 70개, 약 80개, 약 90개, 약 100개, 약 120개, 약 150개, 약 180개, 약 210개, 약 240개, 약 270개, 약 300개, 약 350개, 약 400개, 약 450개, 약 500개, 약 600개, 약 700개, 약 800개, 약 900개 또는 약 1000개의 염기일 수 있다.

[0127] 일부 실시양태에서, 고처리량 시퀀싱은 일루미나(Illumina)의 게놈 분석기(Genome Analyzer) IIX, MiSeq 개인용 시퀀서, 또는 HiSeq 시스템, 예컨대 HiSeq 2500, HiSeq 1500, HiSeq 2000, 또는 HiSeq 1000 기기를 사용하는 시스템으로 이용 가능한 기술의 사용을 수반한다. 이들 기기는 합성 화학에 의한 가역적 터미네이터 기반 시퀀싱을 사용한다. 이들 기기는 8일 내에 2천억 개 이상의 DNA 리드를 처리할 수 있다. 3일, 2일, 1일 이하의 시간 내에 작업을 수행하기 위해서는 더 작은 시스템을 사용할 수 있다.

[0128] 일부 실시양태에서, 고처리량 시퀀싱은 ABI 솔리드 시스템(ABI Solid System)에서 이용 가능한 기술의 사용을 수반한다. 이 유전자 분석 플랫폼은 비드에 연결된 클론으로 증폭된 DNA 단편의 초병렬 시퀀싱을 가능하게 한다. 시퀀싱 방법은 염료 표지된 올리고뉴클레오티드의 순차적 라이게이션을 기초로 한다.

[0129] 차세대 시퀀싱은 이온 반도체 시퀀싱(예를 들어, Life Technologies (Ion Torrent)의 기술을 사용)을 포함할 수 있다. 이온 반도체 시퀀싱은 뉴클레오티드가 DNA 가닥에 도입될 때 이온이 방출될 수 있다는 사실을 이용할 수 있다. 이온 반도체 시퀀싱을 수행하기 위해서, 미세 기계 가공된 웰의 고밀도 어레이를 형성할 수 있다. 각각의 웰은 단일 DNA 주형을 담을 수 있다. 웰 아래에는 이온 감응층이 존재할 수 있고, 이온 감응층 아래에는 이온 센서가 존재할 수 있다. 뉴클레오티드가 DNA에 첨가될 때, H+가 방출될 수 있고, 이것은 pH의 변화로서 측정될 수 있다. H+ 이온은 전압으로 전환되고, 반도체 센서에 의해 기록될 수 있다. 어레이 칩은 한 뉴클레오티드에서 다른 뉴클레오티드로 순차적으로 잠길 수 있다. 스캐닝, 광원 또는 카메라가 필요하지 않을 수 있다. 일부 경우에, 이온프로톤(IONPROTON)™ 시퀀서를 사용하여 핵산의 서열을 결정한다. 일부 경우에, IONPGM™ 시퀀서가 사용된다. 이온 토렌트 개인용 게놈 기기(PGM: Ion Torrent Personal Genome Machine)는 2시간 안에 1,000만 개의 리드를 처리할 수 있다.

[0130] 일부 실시양태에서, 고처리량 시퀀싱은 헬리코스 바이오사이언시스 코포레이션(Helicos BioSciences Corporation, 미국 매사추세츠주 캠브리지 소재)에 의해 이용 가능한 기술, 예컨대 합성에 의한 단일 분자 시퀀싱(SMSS: Single Molecule Sequencing by Synthesis) 방법의 사용을 수반한다. SMSS는 최대 24시간 내에 전체 인간 게놈의 시퀀싱을 허용하기 때문에 특유한 것이다. 마지막으로, SMSS는 미국 특허 출원 공개 제20060024711

호; 제20060024678호; 제20060012793호; 제20060012784호; 및 제20050100932호에 부분적으로 기재되어 있다.

[0131] 일부 실시양태에서, 고처리량 시퀀싱은 454 라이프사이언시스, 인크.(Lifesciences, Inc., 미국 코네티컷주 브랜포드 소재)에 의해 이용 가능한 기술, 예컨대 기기 내의 CCD 카메라로 기록되는 시퀀싱 반응에 의해 생성되는 화학발광 신호를 전송하는 광섬유 플레이트를 포함하는 피코타이터플레이트(PicoTiterPlate) 장치의 사용을 수반한다. 이 광섬유를 사용하면, 4.5시간 내에 최소한 2천만 개의 염기쌍을 검출할 수 있다.

[0132] 비드 증폭에 이어 광섬유 검출을 이용하는 방법은 문헌 [Marguiles, M., *et al.* "Genome sequencing in microfabricated high-density picolitre reactors", Nature, doi:10.1038/nature03959]; 및 미국 특허 출원 공개 제20020012930호; 제20030068629호; 제20030100102호; 제20030148344호; 제20040248161호; 제20050079510호; 제20050124022; 및 제20060078909호에 기재되어 있다.

[0133] 일부 실시양태에서, 고처리량 시퀀싱은 가역적 터미네이터 화학을 이용하는 클론 단일 분자 어레이(Clonal Single Molecule Array)(Solexa, Inc) 또는 합성에 의한 시퀀싱(SBS: sequencing-by-synthesis)을 이용하여 수행된다. 이 기술은 미국 특허 제6,969,488호; 제6,897,023호; 제6,833,246호; 제6,787,308호; 및 미국 특허 출원 공개 제20040106110호; 제20030064398호; 제20030022207호; 및 문헌 [Constans, A, The Scientist 2003, 17(13):36]에 부분적으로 기재되어 있다.

[0134] 차세대 시퀀싱 기술은 퍼시틱 바이오사이언스(Pacific Biosciences)의 실시간(SMRT™) 기술을 포함할 수 있다. SMRT에서, 4개의 DNA 염기는 각각 4개의 상이한 형광 염료 중 하나에 부착될 수 있다. 이 염료는 인에 연결될 수 있다. 단일 DNA 폴리머라제는 제로 모드 도파관(ZMW: zero-mode waveguide)의 바닥에 주형 단일 가닥 DNA의 단일 분자로 고정될 수 있다. ZMW는 ZMW 밖으로 신속하게(마이크로초 내에) 확산할 수 있는 형광 뉴클레오티드의 배경에 대해 DNA 폴리머라제에 의한 단일 뉴클레오티드의 도입을 관찰할 수 있게 하는 구속(confinement) 구조일 수 있다. 성장하는 가닥에 뉴클레오티드를 도입하는 데 수 밀리초가 걸릴 수 있다. 이 시간 동안, 형광 표지가 여기되어 형광 신호를 생성할 수 있으며, 형광 태그는 절단될 수 있다. ZMW는 아래쪽으로부터 조명될 수 있다. 여기 빔으로부터 감쇠된 빛은 각각의 ZMW의 하부 20-30 nm를 통과할 수 있다. 검출 한계가 20 젭토 리터(10" 리터)인 현미경을 만들 수 있다. 작은 검출 부피는 배경 노이즈 감소에 대한 1000배의 개선 효과를 제공할 수 있다. 염료의 상응하는 형광 검출은 염기가 도입되었음을 나타낼 수 있다. 과정이 반복될 수 있다,

[0135] 일부 경우에, 차세대 시퀀싱은 나노포어 시퀀싱이다(예를 들어, 문헌 [Soni GV and Meller A. (2007) Clin Chem 53: 1996-2001] 참조). 나노포어는 약 1 나노미터 직경의 작은 구멍일 수 있다. 전도성 유체에 나노포어를 담그고 여기에 전위를 인가하면, 나노포어를 통한 이온의 전도로 인해 약간의 전류가 발생할 수 있다. 흐르는 전류의 양은 나노포어의 크기에 민감할 수 있다. DNA 분자가 나노포어를 통과하면서, DNA 분자 상의 각각의 뉴클레오티드는 나노포어를 상이한 정도로 방해할 수 있다. 따라서, DNA 분자가 나노포어를 통과할 때 나노포어를 통한 전류의 변화는 DNA 서열의 리드를 나타낼 수 있다. 나노포어 시퀀싱 기술은 옥스포드 나노포어 테크놀로지스(Oxford Nanopore Technologies)로부터 입수할 수 있고, 그 예는 그리들온(GridlON) 시스템이다. 단일 나노포어는 마이크로웰 상부에 걸친 중합체 막에 삽입될 수 있다. 각각의 마이크로웰은 개별 감지를 위한 전극을 가질 수 있다. 마이크로웰은 칩당 100,000개 이상의 마이크로웰(예를 들어, 200,000, 300,000, 400,000, 500,000, 600,000, 700,000, 800,000, 900,000 또는 1,000,000개 초과)을 갖는 어레이 칩으로 제작될 수 있다. 기기(또는 노드)를 사용하여 칩을 분석할 수 있다. 데이터는 실시간으로 분석될 수 있다. 한 번에 하나 이상의 기기가 작동될 수 있다. 나노포어는 단백질 나노포어, 예를 들어 단백질 알파-헤몰리신, 칠량체 단백질 세공일 수 있다. 나노포어는 제조된 고체 상태 나노포어, 예를 들어 합성 막(예를 들어, SiN_x 또는 SiO_2)에 형성된 나노미터 크기의 구멍일 수 있다. 나노포어는 혼성 세공(예를 들어, 고체 상ㅌ 막 내의 단백질 세공의 통합)일 수 있다. 나노포어는 집적 센서(예를 들어, 터널링 전극 검출기, 용량 검출기, 또는 그래핀 기반의 나노 갭 또는 에지 상태 검출기를 가진 나노포어일 수 있다(예를 들어, 문헌 [Garaj *et al.* (2010) Nature vol. 67, doi: 10.1038/nature09379] 참조). 나노포어는 특정 유형의 분자(예를 들어, DNA, RNA 또는 단백질)를 분석하기 위해 관능화될 수 있다. 나노포어 시퀀싱은 무손상 DNA 중합체가 단백질 나노포어를 통해 통과되면서 DNA가 세공을 전위함에 따라 실시간으로 시퀀싱될 수 있는 "가닥 시퀀싱"을 포함할 수 있다. 효소는 이중 가닥 DNA의 가닥을 분리하고, 나노포어를 통해 한 가닥을 공급할 수 있다. DNA는 한쪽 말단에서 헤어핀을 가질 수 있고, 시스템은 두 가닥을 판독할 수 있다. 일부 경우에, 나노포어 시퀀싱은 개개의 뉴클레오티드가 진행성(processive) 엑소뉴클레아제에 의해 DNA 가닥으로부터 절단될 수 있고, 뉴클레오티드가 단백질 나노포어를 통해 통과될 수 있는 "엑소뉴클레아제 시퀀싱"이다. 뉴클레오티드는 세공 내의 분자(예를 들어, 사이클로덱스트란)에 일시적으로 결합할 수 있다. 전류의 특징적인 파괴를 사용하여 염기를 확인할 수 있다.

[0136] 제니아(GENIA)의 나노포어 시퀀싱 기술이 사용될 수 있다. 조작된 단백질 세공은 지질 이중층 막에 매몰될 수 있다. "능동 제어(Active Control)" 기술을 사용하여 효율적인 나노포어-막 어레이 및 채널을 통한 DNA 이동의 제어를 가능하게 할 수 있다. 일부 경우에, 나노포어 시퀀싱 기술은 NABsys에서 제공된다. 게놈 DNA는 평균 길이가 약 100 kb인 가닥으로 단편화될 수 있다. 100 kb 단편은 단일 가닥으로 만들어진 후, 6-mer 프로브와 혼성화될 수 있다. 프로브가 있는 게놈 단편은 나노포어를 통해 유도될 수 있으며, 이는 전류 대 시간 추적 정보를 생성할 수 있다. 전류 추적 정보는 각각의 게놈 단편 상의 프로브의 위치를 제공할 수 있다. 게놈 단편을 정렬하여 게놈에 대해 프로브 맵을 생성할 수 있다. 이 과정은 프로브 라이브러리에 대해 병렬로 수행될 수 있다. 각각의 프로브에 대한 게놈 길이 프로브 맵을 생성할 수 있다. 오류는 "혼성화에 의한 이동 창 시퀀싱(mwSBH: moving window Sequencing by Hybridization)"으로 불리는 과정으로 수정될 수 있다. 일부 경우에, 나노포어 시퀀싱 기술은 IBM/로슈(Roche)에서 제공된다. 전자빔을 사용하여 마이크로칩에 나노포어 크기의 개구부를 만들 수 있다. 전기장을 이용하여 DNA를 나노포어를 통해 당기거나 꿰어넣을 수 있다. 나노포어 내의 DNA 트랜지스터 장치는 금속과 유전체가 교대하는 나노미터 크기의 층을 포함할 수 있다. DNA 백본에서 이산 전하는 DNA 나노포어 내부의 전기장에 의해 포집될 수 있다. 게이트 전압을 끄고 켜면 DNA 서열을 판독할 수 있다.

[0137] 차세대 시퀀싱은 DNA 나노볼 시퀀싱(예를 들어 컴플리트 게노믹스(Complete Genomics)에 의해 실시되는 바와 같은; 예를 들어, 문헌 [Drmanac *et al.* (2010) Science 327: 78-81] 참조)을 포함할 수 있다. DNA는 단리, 단편화되고, 크기가 선택될 수 있다. 예를 들어, DNA는 약 500 bp의 평균 길이로 단편화될 수 있다(예를 들어, 초음파 처리에 의해). 어댑터(Ad1)는 단편의 말단에 부착될 수 있다. 어댑터는 시퀀싱 반응을 위해 앵커와 혼성화하는 데 사용될 수 있다. 각각의 말단에 결합된 어댑터를 갖는 DNA는 PCR에 의해 증폭될 수 있다. 상보성 단일 가닥 말단이 서로 결합하여 원형 DNA를 형성하도록 어댑터 서열을 변형시킬 수 있다. DNA는 후속 단계에서 사용되는 IIS형 제한 효소에 의한 절단으로부터 보호하기 위해 메틸화될 수 있다. 어댑터(예를 들어, 우측 어댑터)는 제한 인식 부위를 가질 수 있고, 제한 인식 부위는 메틸화되지 않은 상태로 유지될 수 있다. 어댑터 내의 비메틸화된 제한 인식 부위는 제한 효소(예를 들어, AcuI)에 의해 인식될 수 있고, DNA는 우측 어댑터의 우측 13 bp가 AcuI에 의해 절단되어 선형 이중 가닥 DNA를 형성할 수 있다. 두 번째 라운드의 우측 및 좌측 어댑터(Ad2)는 선형 DNA의 한쪽 말단에 라이게이션될 수 있으며, 두 어댑터가 결합된 모든 DNA는 PCR에 의해 증폭될 수 있다(예를 들어 PCR에 의해). Ad2 서열은 서로 결합하여 원형 DNA를 형성할 수 있도록 변형될 수 있다. DNA는 메틸화될 수 있지만, 제한 효소 인식 부위는 좌측 Ad1 어댑터에 메틸화되지 않은 상태로 유지될 수 있다. 제한 효소(예를 들어, AcuI)가 적용될 수 있고, DNA는 Ad1의 좌측 13 bp가 절단되어 선형 DNA 단편을 형성할 수 있다. 세 번째 라운드의 우측 및 좌측 어댑터(Ad3)는 선형 DNA의 우측 및 좌측 측면에 라이게이션될 수 있으며, 생성된 단편은 PCR에 의해 증폭될 수 있다. 어댑터는 서로 결합하여 원형 DNA를 형성할 수 있도록 변형될 수 있다. III형 제한 효소(예를 들어, EcoP15)가 첨가될 수 있고; EcoP15는 Ad3의 좌측 26 bp, Ad2의 우측 26 bp DNA를 절단할 수 있다. 이 절단은 DNA의 큰 세그먼트를 제거하고, DNA를 다시 한번 선형화할 수 있다. 네 번째 라운드의 우측 및 좌측 어댑터(Ad4)는 DNA에 라이게이션될 수 있고, DNA는 증폭될 수 있고(예를 들어, PCR에 의해), 서로 결합하여 완전한 원형 DNA 주형을 형성하도록 변형될 수 있다.

[0138] 회전 환형 복제(예를 들어, Phi 29 DNA 폴리머라제 사용)는 작은 DNA 단편을 증폭하기 위해 사용될 수 있다. 4개의 어댑터 서열은 혼성화할 수 있는 팔린드롬 서열을 포함할 수 있으며, 단일 가닥은 자체적으로 폴딩되어 평균 직경이 대략 200-300 나노미터일 수 있는 DNA 나노볼(DNB™)을 형성할 수 있다. DNA 나노볼은 마이크로어레이(시퀀싱 유동 셀)에 부착(흡착에 의해)될 수 있다. 유동 셀은 이산화규소, 티타늄 및 헥사메틸디실라잔(HMDS)으로 코팅된 실리콘 웨이퍼 및 포토레지스트 재료일 수 있다. 시퀀싱은 형광 프로브를 DNA에 라이게이션함으로써 비연쇄 시퀀싱에 의해 수행될 수 있다. 조사되는 위치의 형광 색상은 고해상도 카메라에 의해 가시화될 수 있다. 어댑터 서열 사이의 뉴클레오티드 서열의 종류가 결정될 수 있다.

[0139] 일부 실시양태에서, 고처리량 시퀀싱은 애니도트.칩스(AnyDotchips) (Genovoxx, 독일 소재)을 사용하여 수행할 수 있다. 특히, 애니도트.칩스는 뉴클레오티드 형광 신호 검출을 10배 - 50배 향상시킬 수 있다. 애니도트.칩스 및 이의 사용 방법은 국제 출원 공개 WO 02088382, WO 03020968, WO 03031947, WO 2005044836, PCT/EP 05/05657, PCT/EP 05/05655호; 및 독일 특허 출원 DE 101 49 786, DE 102 14 395, DE 103 56 837, DE 10 2004 009 704, DE 10 2004 025 696, DE 10 2004 025 746, DE 10 2004 025 694, DE 10 2004 025 695, DE 10 2004 025 744, DE 10 2004 025 745, 및 DE 10 2005 012 301에 부분적으로 기재되어 있다.

[0140] 다른 고처리량 시퀀싱 시스템은 문헌 [Venter, J., *et al.* Science 16 February 2001]; [Adams, M. *et al.* Science 24 March 2000]; 및 [M. J. Levene, *et al.* Science 299:682-686, January 2003]; 및 미국 특허 출원 공개 제20030044781호 및 제2006/0078937호에 개시된 것들을 포함한다. 이러한 모든 시스템은 핵산 분자 상에서

측정되는, 즉, 시퀀싱되는 주형 핵산 분자 상의 핵산 중합 효소의 활성이 실시간으로 추적되는 중합 반응을 통해 일시적인 염기 부가에 의해 다수의 염기를 갖는 표적 핵산 분자를 시퀀싱하는 단계를 수반한다. 이어서, 염기 부가 서열에 각각의 단계에서 핵산 중합 효소의 촉매 활성에 의해 표적 핵산의 성장하는 상보성 가닥에 도입되고 있는 염기를 확인함으로써 서열을 추정할 수 있다. 표적 핵산 분자 복합체 상의 폴리머라제는 표적 핵산 분자를 따라 이동하고 활성 부위에서 올리고뉴클레오티드 프라이머를 연장하는 데 적합한 위치에 제공된다. 다수의 표지된 유형의 뉴클레오티드 유사체는 활성 부위에 근접하게 제공되고, 여기서 각각의 구별 가능한 유형의 뉴클레오티드 유사체는 표적 핵산 서열의 상이한 뉴클레오티드에 상보성이다. 성장하는 핵산 가닥은 활성 부위에서 핵산 가닥에 뉴클레오티드 유사체를 부가하는 폴리머라제를 사용하여 연장되며, 여기서 부가되는 뉴클레오티드 유사체는 활성 부위에서 표적 핵산의 뉴클레오티드에 상보성이다. 중합 단계의 결과로서 올리고뉴클레오티드 프라이머에 부가된 뉴클레오티드 유사체가 확인된다. 표지된 뉴클레오티드 유사체를 제공하는 단계, 성장하는 핵산 가닥을 중합하는 단계 및 부가된 뉴클레오티드 유사체를 확인하는 단계를 반복하여, 핵산 가닥이 추가로 연장되고 표적 핵산의 서열이 결정된다.

[0141] 시퀀싱 전에, 핵산 분자는 바코드화(barcoding)되거나 다른 방식으로 표지될 수 있다. 바코드화를 사용하면 서열 리드를 쉽게 분류할 수 있다. 예를 들어, 바코드를 사용하여 동일한 핵산 분자 또는 DNA 단백질 복합체로부터 유래된 서열을 확인할 수 있다. 또한, 바코드는 개별 접합부를 특유하게 확인하기 위해 사용될 수도 있다. 예를 들어, 각각의 접합부는 접합부를 특유하게 확인할 수 있는 특유한(예를 들어, 무작위로 생성된) 바코드로 표시될 수 있다. 동일한 핵산 분자 또는 DNA 단백질 복합체로부터 유래된 서열을 확인하기 위한 제1 바코드 및 개별 접합부를 특유하게 확인하는 제2 바코드와 같은 다수의 바코드가 함께 사용될 수 있다.

[0142] 바코드화는 다수의 기술을 통해 달성될 수 있다. 일부 경우에, 바코드는 펑츄에이션 올리고뉴클레오티드 내의 서열로서 포함될 수 있다. 다른 경우에, 핵산 분자는 적어도 2개의 세그먼트를 포함하는 올리고뉴클레오티드에 접촉될 수 있고, 여기서 하나의 세그먼트는 바코드를 함유하고, 제2 세그먼트는 펑츄에이션 서열에 상보성인 서열을 함유한다. 펑츄에이션 서열에 대한 어닐링 후에, 바코드화된 올리고뉴클레오티드는 동일한 펑츄에이션된 핵산 분자로부터 바코드화된 분자를 생성하기 위해 폴리머라제로 연장될 수 있다. 펑츄에이션된 핵산 분자는 페이즈 정보가 보존되는 투입 핵산 분자의 재배열된 버전이기 때문에, 생성된 바코드화된 분자는 또한 동일한 투입 핵산 분자로부터 유래한다. 이러한 바코드화된 분자는 바코드 서열, 펑츄에이션 상보성 서열 및 게놈 서열을 포함한다.

[0143] 펑츄에이션 서열을 갖거나 갖지 않는 핵산 분자(예를 들어, DNA 단백질 복합체의 일부이거나 또는 상기 복합체로부터 회수된 핵산)에 대해, 분자는 다른 수단에 의해 바코드화될 수 있다. 예를 들어, 핵산 분자는 핵산 분자로부터의 서열을 통합하기 위해 연장될 수 있는 바코드화된 올리고뉴클레오티드에 접촉될 수 있다. 바코드는 펑츄에이션 서열, 제한 효소 인식 부위, 관심 부위(예를 들어, 관심 게놈 영역) 또는 무작위 부위(예를 들어, 바코드 올리고뉴클레오티드 상의 무작위 n-mer 서열을 통해)에 혼성화될 수 있다. 핵산 분자는 다수의 핵산 분자가 동일한 바코드 서열을 갖지 않도록 샘플 내의 다른 핵산 분자로부터 적절한 농도 및/또는 분리(예를 들어, 공간적 또는 시간적 분리)를 사용하여 바코드에 접촉될 수 있다. 예를 들어, 핵산 분자를 포함하는 용액은 단지 하나의 핵산 분자만 또는 단지 하나의 DNA 단백질 복합체만이 제시된 바코드 서열을 갖는 바코드 또는 바코드의 군에 접촉되는 농도로 희석될 수 있다. 바코드는 자유 용액 내의, 유체 구획(fluidic partition)(예를 들어, 소적 또는 웰) 내의, 또는 어레이(예를 들어, 특정 어레이 스폿) 상의 산 분자에 접촉할 수 있다.

[0144] 바코드화된 핵산 분자(예를 들어, 연장 산물)는 예를 들어 짧은 리드 시퀀싱 기기 상에서 시퀀싱될 수 있고, 서열 정보는 동일한 바코드를 갖는 서열 리드를 공통적인 정렬, 스캐폴드, 페이즈 또는 다른 군으로 분류함으로써 결정된다. 이러한 방식으로, 긴 합성 리드는 짧은 리드 시퀀싱을 통해 달성될 수 있다. 대안으로, 시퀀싱 전에, 바코드화된 산물은 예를 들어 긴 리드 시퀀싱 기술을 사용하여 시퀀싱된 긴 분자를 생성하기 위해, 예를 들어 벌크 라이게이션을 통해 함께 연결될 수 있다. 이 경우에, 삽입된 리드쌍은 증폭 어댑터 및 펑츄에이션 서열을 통해 확인할 수 있다. 추가의 정보는 리드쌍의 바코드 서열로부터 얻어진다.

[0145] 대안으로, 일부 경우에 본원에서 설명되는 바와 같이 생성된 라이브러리 분자는 펑츄에이션 올리고 삽입 없이 연결된다. 그럼에도 불구하고, 이들 분자는 5 kb, 10 kb, 20 kb 또는 그 초과의 긴 리드를 생성하기 위해 상업적으로 이용 가능한 긴 리드 화학을 사용하는 시퀀싱에 적합하다. 이러한 경우, 연결 접합부는 서열 분석을 통해 쉽게 확인된다.

[0146] 그렇지 않으면 짧은 리드로부터 결정하기가 어렵거나 불가능할 수 있는, 페이징 정보와 같은 정보를 얻기 위해 긴 리드(예를 들어, 합성 또는 실제의 긴 리드)가 사용될 수 있다. 페이징 정보는 모계/부계 페이징뿐만 아니라

종양/비종양 페이징 정보를 포함한다. 종양/비종양 페이징은 암 게놈 정보를 체세포 게놈 정보와 구별하기 위해 사용할 수 있다.

[0147] 한 예에서, 상기 설명된 바와 같이, FFPE 샘플로부터 생성된 라이브러리와 같은 라이브러리로부터의 단편은 말단 서열이 결정된다. 리드쌍이 관찰되는데, 이것은 각각의 말단이 매핑된 콘티그가 샘플 내의 공통적인 핵산 분자 상에 물리적으로 연결되어 있음을 나타낸다. 생성되는 라이브러리는 단리된 서열의 위치를 게놈 어셈블리와 비교함으로써 회수된 단편의 쌍을 이룬 말단 사이의 거리를 결정하기 위해 시퀀싱에 의해 추가로 분석된다. FFPE 샘플 내의 긴 거리의 리드쌍 빈도는 비-FFPE 샘플의 긴 거리의 리드쌍 빈도와 비교된다. 상기 라이브러리와 같은 예시적인 라이브러리에서, 시퀀싱은 FFPE-시카고 방법이 비-FFPE 샘플에서 수행된 시카고 방법(100 kbp - 200 kbp 삽입체)과 대등하거나(>200 kbp 삽입체) 이보다 큰 긴 거리의 리드쌍 빈도를 유발함을 제시한다. FFPE-시카고 라이브러리의 복잡성 및 원시 시퀀싱 커버리지(raw sequencing coverage)도 결정된다. 라이브러리의 복잡성은 라이브러리 내의 상이한 분자의 다양성을 나타낸다.

[0148] **유전 정보**

[0149] 페이징 정보, 염색체 입체형태(conformation), 서열 어셈블리, 및 구조적 변이(SV), 카피수 변이체(CNV), 이형접합성의 상실(LOH), 단일 뉴클레오티드 변이체(SNV), 단일 뉴클레오티드 다형성(SNP), 염색체 전좌, 유전자 융합, 및 삽입 및 결실(INDEL)을 포함하고 이로 제한되지 않는 유전적 특징은 본원에서 개시되는 방법에 의해 생성된 서열 리드 데이터의 분석에 의해 결정될 수 있다. 유전적 특징의 분석을 위한 다른 입력 사항은 참조 게놈(예를 들어, 주석과 함께), 게놈 마스킹 정보, 및 후보 유전자, 유전자 쌍 및/또는 관심 좌표의 목록을 포함할 수 있다. 입체배열 파라미터(configuration parameter) 및 게놈 마스킹 정보는 사용자 지정되거나, 디폴트 파라미터 및 게놈 마스킹을 사용할 수 있다. 한 예에서, 리드쌍은 게놈에 매핑된 후, 각각의 쌍은 각각 리드쌍의 리드 1 및 리드 2의 연결된 참조 염색체 상의 매핑된 위치와 동일한 x 및 y 좌표를 갖는 평면 내의 지점으로 표시된다. x-y 평면은 겹치지 않는 정사각형 빈으로 나눌 수 있으며, 각 빈에 매핑되는 리드쌍의 수를 도표화할 수 있다. 빈 수는 픽셀에 상응하도록 만들어진 빈을 갖는 이미지(예를 들어, 히트 맵)로서 가시화될 수 있다. 이미지 처리 기술과 같은 다양한 분석 기술을 사용하여 상이한 재배열과 같은 유전적 특징의 특성을 확인할 수 있다. 예를 들어, 커널 컨볼루션 필터링(kernel convolution filtering)은 융합된 게놈 유전자좌의 쌍에 상응하는 이미지 내의 지점을 찾기 위해 사용될 수 있다. 도 2a 및 도 2b는 도 3에 도시된 것과 같은 상호 전좌를 발견하기 위해 사용될 수 있는 예시적인 단순 커널을 보여준다. 도 3은 ETV6과 NTRK3 사이의 상호 전위의 신호를 갖는 이미지를 보여준다. 오른쪽 위 및 왼쪽 아래 사분면에 있는 "나비 넥타이" 모양의 특징은 상호 전좌의 특징인 게놈의 두 영역 사이의 상호작용을 나타낸다.

[0150] 서열 리드 데이터와 같은 입력은 적절한 파일 포맷으로 포맷될 수 있다. 예를 들어, 서열 리드 데이터는 FASTA 파일, FASTQ 파일, BAM 파일, SAM 파일 또는 다른 파일 포맷에 포함될 수 있다. 입력 서열 리드 데이터는 정렬되지 않을 수 있다. 입력 서열 리드 데이터는 정렬될 수 있다.

[0151] 서열 리드 데이터는 분석을 위해 준비될 수 있다. 예를 들어, 리드는 품질을 위해 트리밍될 수 있다. 또한, 필요한 경우, 시퀀싱 어댑터를 제거하기 위해 리드는 트리밍될 수 있다.

[0152] 서열 리드 데이터는 정렬될 수 있다. 예를 들어, 리드쌍은 특정된 참조 게놈에 정렬될 수 있다. 일부 경우에, 참조 게놈은 CRCh38이다. 정렬은 SNAP, 버로우즈-휠러(Burrows-Wheeler) 정렬기(예를 들어, bwa-sw, bwa-mem, bwa-aln), Bowtie2, 노보얼라인(Novoalign) 및 이들의 변형 또는 변이를 포함하고 이로 제한되지 않는 다양한 알고리즘 또는 도구로 수행될 수 있다.

[0153] 분석에 대한 품질 관리(QC) 보고서가 또한 생성될 수 있다. QC 보고서를 사용하여 더 복잡한 시퀀싱을 수행하기 전에 실패한 라이브러리를 확인할 수 있다. 이러한 품질 관리 보고서에는 다양한 측정 기준이 포함될 수 있다. QC 측정 기준은 전체 리드쌍, 중복 비율(예를 들어, PCR 중복), 매핑되지 않은 리드의 비율, 낮은 맵 품질(예를 들어, Q<20)의 리드 비율, 상이한 염색체에 매핑된 리드쌍 비율, 0 내지 1 kbp의 리드쌍 삽입체(예컨대, 매핑 위치 사이의 거리)의 비율, 1 kbp 내지 100 kbp의 리드쌍 삽입체의 비율, 100 kbp 내지 1 Mbp의 리드쌍 삽입체의 비율, 1 Mbp 초과의 리드쌍 삽입체의 비율, 라이게이션 접합부를 포함하는 리드쌍의 비율, 제한 단편 말단에 대한 근접성, 리드쌍 분리 플롯, 및 라이브러리 복잡성의 추정치를 포함할 수 있고 이로 제한되지 않는다. QC 측정 기준을 사용하여 분석을 최적화하고, 시약, 샘플 및 사용자의 품질 문제를 확인할 수 있다. 하나 이상의 QC 측정 기준을 기반으로 서열 정렬을 필터링할 수 있다. 밀접하게 상응하는 위치에서의 리드의 비교에 기초하여 중복 리드가 또한 필터링될 수 있다.

[0154] 서열 리드 분석 결과는 링크 밀도 결과를 포함할 수 있다. 링크 밀도 결과는 전체 게놈, 하나의 유전자좌, 및 링크 밀도 결과의 2개의 유전자좌 뷰(view)를 포함할 수 있다. 링크 밀도 결과는 데이터 세트로 출력될 수 있다. 링크 밀도 결과는 염색체 또는 게놈의 영역 사이의 상호작용(예를 들어, 접촉)의 히트 맵과 같은 링키지 밀도 플롯(LDP: linkage density plot)으로 나타낼 수 있다. 링크 밀도 결과는 품질 점수와 같은 점수와 연결될 수 있다. 일부 경우에, 점수 임계값을 초과하는 결과에 대한 링크 밀도 가시화물이 출력된다. 한 예에서, 가시화물은 전체 게놈, 점수 임계값을 초과하는 디 노보 콜(*de novo* call), 점수 임계값을 초과하는 단측 후보 콜(single-sided candidate call), 및 음성으로 분류된 것을 포함하는 모든 양측(double-sided) 후보에 대한 가시화물을 포함한다. 링크 밀도 가시화는 스케일(예를 들어, 컬러 스케일), 길이 스케일 바, 유전자 이름 라벨, 유전자에 대한 엑손/인트론 구조 글리프(glyph) 및 검출된 재배열의 강조 표시를 포함할 수 있다.

[0155] 링키지 정보는 커버리지, 단편 매핑 가능성, 단편 GC 함량 및 단편 길이와 같은 영향 및 바이어스를 제어하기 위해 정규화될 수 있다. 정규화는 매트릭스 밸런싱 또는 다른 요인에 제한되지 않는(factor-agnostic) 방법에 의해 수행될 수 있다. 매트릭스 밸런싱은 싱크혼-놉(Sinkhorn-Knopp) 알고리즘 또는 나이트-루이즈(Knight-Ruiz) 정규화와 같은 알고리즘을 사용할 수 있다. 위양성을 초래할 수 있는 배경 신호를 보정하기 위해 정규화를 수행할 수도 있다. 예를 들어, 도 4a, 도 4b, 및 도 4c는 3개의 상이한 샘플에서 비교된 동일한 염색체 쌍에서의 이미지 분석 기반 결과를 보여준다. 여러 개의 "히트"(도면에서 원으로 표시)는 여러 샘플에 걸쳐 같은 위치에서 발견되어, 이들이 위양성이라는 의혹을 야기한다. 샘플의 풀(예를 들어, 10개의 샘플)에 걸친 중간의 정규화된 리드 밀도와 같은 정규화는 예를 들어 샘플 픽셀을 중간 픽셀로 나누어 개별 샘플 데이터를 수정하기 위해 사용할 수 있다. 도 5a, 5b, 및 도 5c는 1번 염색체 대 7번 염색체(도 5a), 2번 염색체 대 5번 염색체(도 5b), 및 1번 염색체 대 1번 염색체(도 5c)에 대한 중간의 정규화된 리드 밀도(10개의 샘플에 걸친)를 나타낸다. 정규화는 도 6a에 도시된 바와 같이 동일한 빈 크기를 포함하는 다양한 빈 취급 방법 및 도 6b에 도시된 바와 같이 빈 내삽을 사용하여 수행될 수 있다. 일부 경우에, 빈 내삽은 동일한 빈 크기와 비교할 때 배경 노이즈를 감소시키고, 보다 급격하게 분해된 특징을 제공할 수 있다.

[0156] 정렬된 서열 데이터는 전체 게놈을 통한 재배열 및 특정한 2개의 유전자좌(또는 양측) 후보 유전자에서의 재배열을 포함하는 재배열에 대해 분석될 수 있다. 분석은 또한 접촉, 융합 및 연결의 확인도 포함할 수 있다. 서열 리드 데이터의 정렬(예를 들어, BAM 파일 또는 다른 적절한 포맷에서)이 분석에 입력될 수 있다. 게놈 마스킹 정보를 또한 입력하거나, 디폴트 게놈 마스킹 정보를 분석에 사용할 수 있다. 분석은 전체 게놈에 걸쳐 수행될 수 있다. 추가로 또는 대안으로, 양측 후보 융합체의 목록에 대한 분석이 수행될 수 있다. 일부 경우에, 후보 융합체의 목록에 대해 수행된 분석은 전체 게놈에 대해 수행된 분석보다 더 민감하다. 양측 후보 융합체의 분석은 게놈에 걸친 스캐닝에 의해 상실될 수 있는 DNA의 비교적 짧은 세그먼트의 전좌를 포함하는 융합체를 검출할 수 있다.

[0157] 접촉 및 재배열(결실, 중복, 삽입, 역위 또는 반전, 전좌, 연결, 융합 및 분열을 포함하고 이로 제한되지 않음), 및 다른 상호작용과 같은 특징을 확인하는 분석은 다양한 기술로 수행될 수 있다. 분석 기술은 통계 및 확률 분석, 푸리에(Fourier) 분석, 컴퓨터 비전 및 다른 이미지 처리, 언어 처리(예를 들어, 자연 언어 처리) 및 기계 학습을 포함한 신호 처리를 포함할 수 있다. 예를 들어, 접촉 매트릭스와 같은 상호작용 플롯은 특징을 나타내는 특징에 대해 분석될 수 있다. 일부 경우에, 필터를 플롯 또는 다른 데이터에 적용할 수 있다. 필터는 스무딩(smoothing) 필터(예를 들어, 커널 스무딩 또는 사비츠키-골레이(Savitzky-Golay) 필터, 가우스 흐림(Gaussian blur)를 포함하고 이로 제한되지 않는 컨볼루션 필터일 수 있다.

[0158] 일부 실시양태는 게놈 구조 결정의 성분으로서 기계 학습을 포함하고, 따라서 일부 컴퓨터 시스템은 기계 학습 능력을 갖는 모듈을 포함하도록 구성된다. 기계 학습 모듈은 기계 학습 기능을 구성하기 위해 다음에 나열된 방식 중 적어도 하나를 포함한다.

[0159] 기계 학습을 구성하는 방식은 자동화된 질량 분광 데이터 스팟 검출 및 콜링(calling)을 수행할 수 있도록 데이터 필터링 용량을 다양하게 입증한다. 이 방식은 일부 경우에 역위, 삽입, 결실 또는 전좌와 같은 다양한 게놈의 구조적 변화를 나타내는 예측된 패턴의 존재에 의해 촉진된다.

[0160] 기계 학습을 구성하는 방식은 리드쌍 빈도를 하류 분석에 도움이 되는 형태로 만들도록 데이터 처리 또는 데이터 처리 용량을 다양하게 입증한다. 데이터 처리의 예는 로그 변환, 배율 비율(scaling ratio)의 지정, 또는 데이터를 하류 분석에 도움이 되는 형태로 만들기 위해 데이터를 작성된 특징에 매핑하는 것을 포함하지만, 반드시 이로 제한되는 것은 아니다.

[0161] 본원에서 개시되는 기계 학습 데이터 분석 성분은 1 내지 10,000개의 특징 또는 2 내지 300,000개의 특징과 같

은 리드쌍 데이터 세트 내의 광범위한 특징 또는 이들 범위 내의 또는 이들 범위를 초과하는 많은 특징을 규칙적으로 처리한다. 일부 경우에, 데이터 분석은 적어도 1k, 2k, 3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 20k, 30k, 40k, 50k, 60k, 70k, 80k, 90k, 100k, 120k, 140k, 160k, 180k, 200k, 220k, 2240k, 260k, 280k, 300k 또는 300k 초과의 특징을 포함한다.

[0162] 리드쌍 분배 패턴은 본원의 개시내용과 일치하는 임의의 수의 방법을 사용하여 확인된다. 일부 경우에, 리드쌍 분배 패턴의 선택은 본원의 개시내용과 일치하고 관련 기술 분야의 통상의 기술자에게 친숙한 탄성 네트, 정보 획득, 랜덤 포레스트 대치(random forest imputing) 또는 다른 특징 선택 방법을 포함한다.

[0163] 선택된 리드쌍 분배 패턴은 본원의 개시내용과 일치하는 임의의 수의 방법을 다시 사용하여, 게놈의 구조적 변화를 나타내는 예측된 패턴에 대해 매칭된다. 일부 경우에, 리드쌍 패턴 검출은 본원의 개시내용과 일치하고 관련 기술 분야의 통상의 기술자에게 친숙한 로지스틱 회귀, SVM, 랜덤 포레스트, KNN 또는 다른 분류 방법을 포함한다.

[0164] 본원에서 개시되는 분석을 위해 구성된 컴퓨터에서 기계 학습을 적용하거나 기계 학습 모듈을 제공하는 것은 진행 중인 모니터링 절차의 일부로서 무증상 질환 검출 또는 조기 검출에 대한 관련 게놈의 구조적 변화의 검출을 가능하게 하여, 증상 발달 전에 질병 또는 장애를 확인하거나 개입이 보다 쉽게 달성되거나 성공적인 결과를 낳을 확률이 더 높아진다.

[0165] 본원에서 개시되는 분석을 위해 구성된 컴퓨터에서 기계 학습을 적용하거나 기계 학습 모듈을 제공하는 것은 또한 예를 들어 약물 시험의 일부로서 약물 치료를받은 개체에서 구조적 재배열의 확인을 가능하게 하여, 개체 또는 집단에 대한 시험 결과는 약물 효능에 긍정적으로 또는 부정적으로 상응하는 게놈의 특정 구조적 사건을 확인하기 위해 동시에 또는 후향적으로(retrospectively) 상호 연결될 수 있다.

[0166] 본원에서 개시되는 분석을 위해 구성된 컴퓨터에서 기계 학습을 적용하거나 기계 학습 모듈을 제공하는 것은 또한 균질화 없이 수집된 종양 조직 샘플과 같은 유전적으로 이질적인 샘플의 특정 영역에 상응하는 구조적 재배열의 확인을 가능하게 하여, 샘플 내의 위치 정보를 보존할 수 있다. 일부 종양 영역은 특히 전이 또는 종양 전파에 숙달된 세포 집단에 상응하는 것으로 알려져 있기 때문에, 이러한 세포 집단과 상호관련되는 게놈 재배열 또는 다른 페이즈 정보를 확인하면, 이러한 특히 위험한 세포 집단을 표적화하는 치료 요법을 선택하는 데 도움을 줄 수 있다.

[0167] 모니터링은 그에 대한 발병 또는 진행의 특징이 모니터링되는 장애에 대한 유전적 소인을 나타내는 유전적 평가와 조합하여 또는 상기 평가를 지지하면서 종종 수행되지만, 반드시 수행될 필요가 있는 것은 아니다. 이와 유사하게, 일부 경우에 기계 학습은 치료 요법이 진행 중인 단백질체학 매개 모니터링에 의해 지시되는 바와 같이 시간이 지남에 따라 수정되거나, 지속되거나, 해결될 수 있도록 치료 요법에 대한 치료 효능의 모니터링 또는 치료 효능의 평가를 용이하게 하기 위해 사용된다.

[0168] 기계 학습 알고리즘을 실행하도록 구성된 모듈을 갖는 기계 학습 방법 및 컴퓨터 시스템은 복잡성이 다양한 데이터세트에서 페이즈 정보 또는 게놈 재배열의 확인을 용이하게 한다. 일부 경우에, 페이즈 정보 또는 게놈 재배열은 다량의 질량 분광 데이터, 예컨대 다수의 시점에서 단일 개체로부터 수득된 데이터, 다수의 개체, 예컨대 관심 병태에 대한 알려진 상태 또는 알려진 치료 결과 또는 반응의 다수의 개체로부터, 또는 다수의 시점 및 다수의 개체로부터 채취한 샘플을 포함하는 비표적화된 데이터베이스로부터 확인된다.

[0169] 대안으로, 일부 경우에, 기계 학습은 예를 들어 개체의 건강 상태가 시점에 대해 알려져 있을 때 다수의 시점에 걸쳐 단일 개체로부터 게놈 재배열 또는 페이즈 정보를 수집함으로써, 또는 관심 병태에 대한 알려진 상태의 다수의 개체로부터 서열 정보를 수집함으로써, 또는 다수의 시점에서 다수의 개체로부터 서열 정보를 수집함으로써 게놈 재배열 또는 페이즈 정보를 표적화하는 데이터베이스의 분석을 통해 게놈 재배열 또는 페이즈 정보의 개선을 용이하게 한다. 쉽게 알 수 있듯이, 일부 경우에, 수술에 따라 수집된 가교결합된 샘플 또는 약물 시험에 따라 수집된 FFPE 샘플과 같은 보존된 샘플의 사용을 통해 서열 정보의 수집이 용이해진다.

[0170] 따라서, 서열 정보는 단독으로 또는 약물 시험 결과 또는 외과적 개입 결과의 정보와 함께 수집된다. 서열 데이터는 단독으로 또는 하나 이상의 추가의 마커와 조합하여 건강 상태 신호를 설명하는 게놈 재배열에 상응하는 패턴을 나타내는 리드쌍의 하위세트를 확인하기 위해, 예를 들어, 본원에서 개시되는 바와 같이 구성된 컴퓨터 시스템에서 기계 학습에 적용된다. 따라서, 기계 학습은 일부 경우에 DNA 또는 RNA 서열과 같은 서열의 확인, 또는 개체의 건강 상태를 개별적으로 알려주는 게놈 재배열을 용이하게 한다.

[0171] 검출 가능한 재배열에 대한 중단점 사이의 최소 거리는 2 bp, 3 bp, 4 bp, 5 bp, 6 bp, 7 bp, 8 bp, 9 bp, 10

bp, 20 bp, 30 bp, 40 bp, 50 bp, 60 bp, 70 bp, 80 bp, 90 bp, 100 bp, 200 bp, 300 bp, 400 bp, 500 bp, 600 bp, 700 bp, 800 bp, 900 bp, 1 kb, 2 kb, 3 kb, 4 kb, 5 kb, 6 kb, 7 kb, 8 kb, 9 kb, 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb, 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 600 kb, 700 kb, 800 kb, 900 kb, 1 Mb, 2 Mb, 3 Mb, 4 Mb, 5 Mb, 6 Mb, 7 Mb, 8 Mb, 9 Mb, 10 Mb, 20 Mb, 30 Mb, 40 Mb, 50 Mb, 60 Mb, 70 Mb, 80 Mb, 90 Mb, 100 Mb, 200 Mb, 300 Mb, 400 Mb, 500 Mb, 600 Mb, 700 Mb, 800 Mb, 900 Mb, 또는 1 Gb 미만이거나, 이와 비슷하거나, 또는 이를 포함하는 핵산 길이의 목록으로부터 선택된 2개의 숫자에 의해 정의된 범위 내의 숫자일 수 있다.

[0172] 재배열 분석은 대상 게놈에서 연결된 것으로 간주되는 중단점 쌍의 목록을 생성할 수 있다. 중단점 좌표 쌍의 목록에는 중단점 좌표 쌍에 대한 통계적 유의성 또는 신뢰도 측정 기준(예를 들어, p-값)도 포함될 수 있다. 이러한 중단점 쌍은 브라우저 확장 가능 데이터(BED) 또는 BED-PE와 같은 적절한 형식으로 출력할 수 있다.

[0173] 염색체 입체형태의 분석은 또한 본원에서 개시되는 기술을 사용하여 수행될 수 있다. 예를 들어, 위상 결합 도메인(TAD: topologically associating domain) 및 TAD 경계를 결정할 수 있다. 또한, 라미나 결합 도메인(LAD: lamina-associated domain), 복제 시간 대역, 및 대형 조직화된 염색질 K9 변형(LOCK) 도메인을 포함하고 이로 제한되지 않는 다른 위상 도메인 및 경계를 결정할 수도 있다.

[0174] 도 7은 전체 게놈 스캐닝 분석 파이프라인에 의한 분석을 도시한 것이다. 분석 파이프라인에 의해 만들어진 샘플 콜은 흰색 원으로 표시된다. 도 7은 250k 빈을 갖는 3번 염색체 대 6번 염색체의 플롯을 보여준다.

[0175] 예시적인 실시양태에서, 시퀀싱 데이터는 출발 FFPE 샘플에 존재하는 것으로 알려진 다형성에 대한 페이징 정보를 결정하기 위해 사용된다. 예를 들어, 시퀀싱 데이터는 SNP와 같은 특정 다형성이 동일하거나 상이한 DNA 분자 상에 존재하는지의 여부를 결정하기 위해 사용된다. 이 방법을 사용하여 결정된 페이징의 정확도는 GIAB 샘플의 서열과 같은 알려진 서열과 비교함으로써 측정된다. 예를 들어, 일부 경우에 0-10,000 사이에 132,796개의 SNP가 발견되었고, 99.059%가 올바른 페이즈로 존재하였다. 높은 일치도(>95%)는 약 1.5 MB까지 나타난다(13개 중 1개가 상실된 70-80 kb 빈 및 15개 중 2개가 상실된 1.1 - 1.3 MB 빈 제외). 1.7 - 1.9 MB 범위에서, 7개의 SNP 쌍 페이즈 중 7개가 적절하게 콜링되었다. 이러한 데이터로부터, 낮은 수준의 허위(spurious) 링키지에도 불구하고, FFPE-시카고 방법을 사용하여, 심지어 메가베이스 범위까지 적절한 장범위 정보가 결정된다고 결론지었다. 중요한 것은 이러한 '일치도' 예측 비율이 95% 이상이고, 이것은 무작위로 예상되는 50% 성공률보다 유의하게 더 높다는 것이다.

[0176] *구조적 페이징 정보*

[0177] 현재, 구조적 및 페이징 분석(예를 들어, 의학적 목적을 위한)은 여전히 어려운 도전 과제이다. 예를 들어, 암, 동일한 유형의 암이 존재하는 개체 중에, 또는 심지어 동일한 종양 내에서도 놀라운 이질성이 있다. 결과로 발생하는 효과로부터 원인을 파악하는 것은 낮은 샘플당 비용에서의 매우 높은 정밀도 및 처리량을 필요로 할 수 있다. 개인 맞춤형 의약 분야에서, 게놈 치료의 최적 표준 중 하나는 크고 작은 구조적 재배열 및 새로운 돌연변이를 포함하는, 철저히 특성화되고 페이징된 모든 변이체를 갖는 시퀀싱된 게놈이다. 종래의 기술을 사용하여 이를 달성하려면, 통상적인 의료 절차가 되기에는 현재 너무 비싸고 번거로운 디 노보 어셈블리에 필요한 것과 비슷한 노력이 필요하다.

[0178] 페이징 정보는 모계/부계 페이징뿐만 아니라 종양/비종양 페이징 정보를 포함한다. 종양/비종양 페이징은 암 게놈 정보를 체세포 게놈 정보와 구별하기 위해 사용할 수 있다.

[0179] 본 개시내용의 일부 실시양태에서, 대상체로부터의 보존된 조직(예를 들어, FFPE 조직)이 제공될 수 있고, 상기 방법은 어셈블리된 게놈, 콜링된 변이체(큰 구조적 변이체 및 카피수 변이체 포함)와의 정렬, 페이징된 변이체 콜 또는 임의의 추가의 분석을 다시 수행할 수 있다. 다른 실시양태에서, 본원에서 개시되는 방법은 개체에 대해 긴 거리의 리드쌍 라이브러리를 직접 제공할 수 있다.

[0180] 본 개시내용의 다양한 실시양태에서, 본원에서 개시되는 방법은 긴 거리에 의해 분리된 장범위 리드쌍을 생성할 수 있다. 이 거리의 상한은 크기가 큰 DNA 샘플을 수집하는 능력에 의해 개선될 수 있다. 일부 경우에, 리드쌍은 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, 225, 250, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000, 4000, 5000 kbp 또는 그 초과의 게놈 거리까지 걸칠 수 있다. 일부 예에서, 리드쌍은 500 kbp까지의 게놈 거리에 걸칠 수 있다. 다른 예에서, 리드쌍은 2000 kbp까지의 게놈 거리에 걸칠 수 있다. 본원에서 개시되는 방법은 분자 생물학의 표준 기술을 통합하고 이를 바탕으로 할 수 있으며, 효율, 특이성 및 게놈 커버리지의 증가를 위해 더욱 적합하다.

[0181] 다른 실시양태에서, 본원에서 개시되는 방법은 현재 사용되는 시퀀싱 기술과 함께 사용될 수 있다. 예를 들어, 이 방법은 잘 검증된 및/또는 널리 사용되는 시퀀싱 장비와 함께 사용할 수 있다. 추가의 실시양태에서, 본원에서 개시되는 방법은 현재 사용되는 시퀀싱 기술로부터 유도된 기술 및 방법과 함께 사용될 수 있다.

[0182] 다양한 실시양태에서, 본 개시내용은 보존된(예를 들어, FFPE) 샘플 또는 세포 내의 염색체의 물리적 배치를 조사하는 단계를 포함하는, 본원에서 개시되는 하나 이상의 방법을 제공한다. 시퀀싱을 통해 염색체의 물리적 배치를 조사하는 기술의 예는 "C" 기술 패밀리, 예컨대 염색체 입체형태 포획("3C"), 환형 염색체 입체형태 포획("4C"), 탄소-카피 염색체 포획("5C"), 및 Hi-C 기반 방법; 및 ChIP 기반 방법, 예컨대 ChIP-루프, ChIP-PET를 포함한다. 이러한 기술은 핵에서 공간적 관계를 강화하기 위해 살아있는 세포 내의 염색질의 고정을 이용한다. 후속 처리 및 생성물의 시퀀싱을 통해, 연구자는 게놈 영역 간의 근접한 회합 매트릭스를 복구할 수 있다. 추가의 분석을 통해, 이러한 회합은 보존된(예를 들어, FFPE) 샘플에 물리적으로 배열된 바와 같이 염색체의 3차원 기하학적 맵을 생성하는 데 사용될 수 있다. 이러한 기술은 염색체의 별개의 공간적 구성을 설명하며, 염색체 유전자좌 중의 기능적 상호작용에 대한 정확한 설명을 제공한다.

[0183] 일부 실시양태에서, 염색체 내 상호작용은 염색체 연결성과 상관관계가 있다. 일부 경우에, 염색체 내 데이터가 게놈 어셈블리를 도울 수 있다. 일부 경우에, 염색질은 시험관 내에서 재구축된다. 이것은 염색질, 특히 염색질의 주요 단백질 성분 인 히스톤이 시퀀싱을 통해 염색질 입체형태 및 구조를 검출하는 가장 일반적인 "C" 기술 패밀리, 즉 3C, 4C, 5C 및 Hi-C 하에서 고정을 위해 중요하기 때문에 유리할 수 있다. 염색질은 서열 측면에서 매우 비특이적이며, 일반적으로 게놈에 걸쳐 균일하게 어셈블리될 것이다. 일부 경우에, 염색질을 사용하지 않는 종의 게놈은 재구축된 염색질에 어셈블리될 수 있고, 따라서 생명의 모든 영역에 대한 공개를 위한 지평을 확장할 수 있다.

[0184] 리드쌍 데이터는 염색질 입체형태 포획 기술로부터 얻을 수 있다. 일부 예에서, 라이게이션 또는 다른 태그 부착은 물리적으로 근접한 게놈 영역을 표시하기 위해 수행된다. 단백질(예를 들어, 히스톤)이 염색질 내에서 DNA 분자, 예를 들어 게놈 DNA와 복합체로 안정하게 결합되도록 상기 복합체를 가교결합시키는 것은 본원의 다른 곳에서 보다 상세히 설명되거나 관련 기술 분야에 공지된 적절한 방법에 따라 수행될 수 있다. 일부 경우에, 샘플 보존(예를 들어, 고정)으로 인해 발생하는 가교결합은, 상기 복합체가 예컨대 프로테이나제 K 처리의 배제를 통해 분해되지 않도록 하는 조건 하에 DNA-단백질 복합체를 추출함으로써 이용된다. 예를 들어, 게놈 서열을 따라 근접하지 않은 뉴클레오티드 세그먼트는 염색질과 같은 구조의 일부일 때 물리적으로 근접해질 수 있다. 이러한 뉴클레오티드 세그먼트는 본 개시내용의 방법에 따라 함께 라이게이션된 후, 분석될 수 있다. 예를 들어, 라이게이션된 뉴클레오티드 세그먼트를 시퀀싱할 수 있고, 2개의 라이게이션된 세그먼트의 시퀀싱된 말단 사이의 거리(삽입 거리)를 분석할 수 있다. 도 8a는 본 개시내용의 기술에 의해 분석된 보존된 샘플(예를 들어, FFPE 샘플)에 대한 염기쌍(bp)의 삽입 거리의 함수로서 특정 범위의 삽입 확률 그래프를 보여준다. 도 8b는 시카고 방법을 사용하여 분석된 샘플에 대한 유사한 그래프를 보여준다. 두 그래프에서, x축은 0 내지 300,000의 삽입 거리(bp)를 보여주고, y축은 축의 상부의 10^{0}으로부터 축의 하부의 10^{-8}까지의 상기 거리의 삽입 확률(로그 형식)을 보여준다.

[0185] 일부 경우에, 2개 이상의 뉴클레오티드 서열은 하나 이상의 뉴클레오티드 서열에 결합된 단백질을 통해 가교결합될 수 있다. 한 가지 방법은 염색질을 자외선 조사에 노출시키는 것이다(Gilmour *et al.*, Proc. Nat'l. Acad. Sci. USA 81:4275-4279, 1984). 폴리뉴클레오티드 세그먼트의 가교결합은 화학적 또는 물리적(예를 들어, 광학적) 가교결합과 같은 다른 방법을 이용하여 수행될 수도 있다. 적합한 화학적 가교결합제는 포름알데히드 및 소랄렌을 포함하고 이로 제한되지 않는다([Solomon *et al.*, Proc. Natl. Acad. Sci. USA 82:6470-6474, 1985]; [Solomon *et al.*, Cell 53:937-947, 1988]). 예를 들어, 가교결합은 DNA 분자 및 염색질 단백질을 포함하는 혼합물에 2% 포름알데히드를 첨가함으로써 수행될 수 있다. DNA의 가교결합에 사용될 수 있는 작용제의 다른 예는 UV 광, 미토마이신 C, 질소 머스타드, 멜팔란, 1,3-부타디엔 디에폭시드, 시스 디아민디클로로백금(II) 및 사이클로포스파미드를 포함하고 이로 제한되지 않는다. 적절하게는, 가교결합제는 비교적 짧은 거리, 예컨대 약 2Å을 가교시켜 역전될 수 있는 긴밀한 상호작용을 선택하는 가교결합을 형성할 것이다.

[0186] 일반적으로, 염색체의 물리적 배치를 조사하기 위한 절차, 예컨대 Hi-C 기반 기술은 배양된 세포 또는 1차 조직으로부터 단리된 염색질과 같은 세포/유기체 내에서 형성되는 염색질을 이용한다. 시카고 기반 방법은 세포/유기체로부터 단리된 염색질을 사용하는 것뿐만 아니라 재구성된 염색질을 사용하는 상기 기술의 사용을 모두 제공한다. 재구성된 염색질은 다양한 특징을 통해 세포/유기체 내에 형성된 염색질과 구분된다. 첫째, 많은 샘플의 경우, 체액 수집, 협측 또는 직장 부위 샘플의 면봉 채취, 상피 샘플 채취 등과 같은 다양한 비침습적 및 침

습적 방법을 사용하여 네이키드 DNA 샘플을 수집할 수 있다. 둘째, 염색질 재구성은 게놈 어셈블리 및 하플로타입 페이징에 대한 인공물을 생성하는, 염색체간 및 다른 장범위 상호작용의 형성을 실질적으로 방지한다. 일부 경우에, 샘플은 본 개시내용의 방법 및 조성물에 따른 약 20, 15, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0.5, 0.4, 0.3, 0.2, 0.1% 미만의 또는 이보다 더 작은 값 미만의 염색체간 또는 분자간 가교결합에 관한 것이다. 일부 예에서, 샘플은 약 5% 미만의 염색체간 또는 분자간 가교결합을 가질 수 있다. 일부 예에서, 샘플은 약 3% 미만의 염색체간 또는 분자간 가교결합을 가질 수 있다. 추가의 예에서, 약 1% 미만의 염색체간 또는 분자간 가교결합을 가질 수 있다. 셋째로, 가교결합이 가능한 부위의 빈도 및 이에 따른 폴리뉴클레오티드 내의 분자내 가교결합의 빈도가 조절될 수 있다. 예를 들어, 히스톤에 대한 DNA의 비율은 뉴클레오솜 밀도가 원하는 값으로 조절될 수 있도록 다양할 수 있다. 일부 경우에, 뉴클레오솜 밀도는 생리학적 수준 미만으로 감소한다. 따라서, 가교결합의 분포는 장범위 상호작용을 선호하도록 변경될 수 있다. 일부 실시양태에서, 다양한 가교결합 밀도를 갖는 하위 샘플은 단범위 및 장범위 결합을 모두 포함하도록 제조될 수 있다. 예를 들어, 가교결합 조건은 적어도 약 1%, 약 2%, 약 3%, 약 4%, 약 5%, 약 6%, 약 7%, 약 8%, 약 9%, 약 10%, 약 11%, 약 12%, 약 13%, 약 14%, 약 15%, 약 16%, 약 17%, 약 18%, 약 19%, 약 20%, 약 25%, 약 30%, 약 40%, 약 45%, 약 50%, 약 60%, 약 70%, 약 80%, 약 90%, 약 95%, 또는 약 100%의 가교결합이, 샘플 DNA 분자 상에서 적어도 약 50 kb, 약 60 kb, 약 70 kb, 약 80 kb, 약 90 kb, 약 100 kb, 약 110 kb, 약 120 kb, 약 130 kb, 약 140 kb, 약 150 kb, 약 160 kb, 약 180 kb, 약 200 kb, 약 250 kb, 약 300 kb, 약 350 kb, 약 400 kb, 약 450 kb, 또는 약 500 kb 떨어진 DNA 세그먼트 사이에서 발생하도록 조절될 수 있다.

[0187] 암 게놈 시퀀싱에 의해 요구되는 고도의 정확도는 본원에서 설명되는 방법 및 시스템을 사용하여 달성될 수 있다. 부정확한 참조 게놈은 암 게놈을 시퀀싱할 때 염기 콜링을 어렵게 만든다. 이질적 샘플 및 작은 출발 물질, 예를 들어 생검으로 얻은 샘플은 추가의 어려움을 야기한다. 또한, 대규모의 구조적 변이체 및/또는 이형접합성의 상실의 검출은 암 게놈 시퀀싱뿐만 아니라, 체세포 변이체 및 염기 콜링의 오류를 구별하는 능력에 종종 중요하다.

[0188] 본원에서 설명되는 시스템 및 방법은 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20개 또는 그 초과의 다양한 게놈을 함유하는 복합 샘플로부터 정확한 긴 서열을 생성할 수 있다. 정상, 양성 및/또는 종양 기원의 혼합된 샘플을 분석할 수 있으며, 경우에 따라 정상 대조군을 필요로 하지 않는다. 일부 실시양태에서, 100 ng의 적은 또는 심지어 수백 개의 극히 적은 게놈 등가물의 출발 샘플을 사용하여 정확한 긴 서열을 생성한다. 본원에서 설명되는 시스템 및 방법은 카피수 변이체, 대규모의 구조적 변이체 및 재배열의 검출을 허용할 수 있고, 페이징된 변이체 콜은 약 1 kbp, 약 2 kbp, 약 5 kbp, 약 10 kbp, 약 20 kbp, 약 50 kbp, 약 100 kbp, 약 200 kbp, 약 500 kbp, 약 1 Mbp, 약 2 Mbp, 약 5 Mbp, 약 10 Mbp, 약 20 Mbp, 약 50 Mbp 또는 약 100 Mbp 또는 그 초과의 뉴클레오티드에 걸친 긴 서열에 대해 얻을 수 있다. 예를 들어, 약 1 Mbp 또는 약 2 Mbp에 걸친 긴 서열에 대해 페이즈 변이체 콜을 얻을 수 있다.

[0189] 샘플은 다양한 부피 및 표면적의 조직 절편을 포함할 수 있다. 일부 경우에, 샘플은 두께가 약 5 ㎛ 내지 10 ㎛인 조직 절편을 포함한다. 일부 경우에, 샘플은 두께가 약 1 ㎛, 2 ㎛, 3 ㎛, 4 ㎛, 5 ㎛, 6 ㎛, 7 ㎛, 8 ㎛, 9 ㎛, 10 ㎛, 11 ㎛, 12 ㎛, 13 ㎛, 14 ㎛, 15 ㎛, 16 ㎛, 17 ㎛, 18 ㎛, 19 ㎛, 20 ㎛, 25 ㎛, 30 ㎛, 35 ㎛, 40 ㎛, 45 ㎛, 50 ㎛, 55 ㎛, 60 ㎛, 65 ㎛, 70 ㎛, 75 ㎛, 80 ㎛, 85 ㎛, 90 ㎛, 95 ㎛, 100 ㎛, 150 ㎛, 200 ㎛, 250 ㎛, 300 ㎛, 350 ㎛, 400 ㎛, 450 ㎛, 500 ㎛, 550 ㎛, 600 ㎛, 650 ㎛, 700 ㎛, 750 ㎛, 800 ㎛, 850 ㎛, 900 ㎛, 950 ㎛, 1000 ㎛ 또는 그 초과인 조직 절편을 포함한다. 일부 경우에, 샘플은 두께가 적어도 약 1 ㎛, 2 ㎛, 3 ㎛, 4 ㎛, 5 ㎛, 6 ㎛, 7 ㎛, 8 ㎛, 9 ㎛, 10 ㎛, 11 ㎛, 12 ㎛, 13 ㎛, 14 ㎛, 15 ㎛, 16 ㎛, 17 ㎛, 18 ㎛, 19 ㎛, 20 ㎛, 25 ㎛, 30 ㎛, 35 ㎛, 40 ㎛, 45 ㎛, 50 ㎛, 55 ㎛, 60 ㎛, 65 ㎛, 70 ㎛, 75 ㎛, 80 ㎛, 85 ㎛, 90 ㎛, 95 ㎛, 100 ㎛, 150 ㎛, 200 ㎛, 250 ㎛, 300 ㎛, 350 ㎛, 400 ㎛, 450 ㎛, 500 ㎛, 550 ㎛, 600 ㎛, 650 ㎛, 700 ㎛, 750 ㎛, 800 ㎛, 850 ㎛, 900 ㎛, 950 ㎛, 1000 ㎛ 또는 그 초과인 조직 절편을 포함한다. 일부 경우에, 샘플은 두께가 최대 약 1 ㎛, 2 ㎛, 3 ㎛, 4 ㎛, 5 ㎛, 6 ㎛, 7 ㎛, 8 ㎛, 9 ㎛, 10 ㎛, 11 ㎛, 12 ㎛, 13 ㎛, 14 ㎛, 15 ㎛, 16 ㎛, 17 ㎛, 18 ㎛, 19 ㎛, 20 ㎛, 25 ㎛, 30 ㎛, 35 ㎛, 40 ㎛, 45 ㎛, 50 ㎛, 55 ㎛, 60 ㎛, 65 ㎛, 70 ㎛, 75 ㎛, 80 ㎛, 85 ㎛, 90 ㎛, 95 ㎛, 100 ㎛, 150 ㎛, 200 ㎛, 250 ㎛, 300 ㎛, 350 ㎛, 400 ㎛, 450 ㎛, 500 ㎛, 550 ㎛, 600 ㎛, 650 ㎛, 700 ㎛, 750 ㎛, 800 ㎛, 850 ㎛, 900 ㎛, 950 ㎛, 1000 ㎛ 또는 그 초과인 조직 절편을 포함한다. 일부 경우에, 샘플은 표면적이 약 100 내지 300 ㎟인 조직 절편을 포함한다. 일부 경우에, 샘플은 표면적이 약 10 ㎟, 20 ㎟, 30 ㎟, 40 ㎟, 50 ㎟, 60 ㎟, 70 ㎟, 80 ㎟, 90 ㎟, 100 ㎟, 200 ㎟, 300 ㎟, 400 ㎟, 500 ㎟, 600 ㎟, 700 ㎟, 800 ㎟, 900 ㎟

^{2}, 1000 ㎟ 또는 그 초과인 조직 절편을 포함한다. 일부 경우에, 샘플은 표면적이 적어도 약 10 ㎟, 20 ㎟, 30 ㎟, 40 ㎟, 50 ㎟, 60 ㎟, 70 ㎟, 80 ㎟, 90 ㎟, 100 ㎟, 200 ㎟, 300 ㎟, 400 ㎟, 500 ㎟, 600 ㎟, 700 ㎟, 800 ㎟, 900 ㎟, 1000 ㎟ 또는 그 초과인 조직 절편을 포함한다. 일부 경우에, 샘플은 표면적이 최대 약 10 ㎟, 20 ㎟, 30 ㎟, 40 ㎟, 50 ㎟, 60 ㎟, 70 ㎟, 80 ㎟, 90 ㎟, 100 ㎟, 200 ㎟, 300 ㎟, 400 ㎟, 500 ㎟, 600 ㎟, 700 ㎟, 800 ㎟, 900 ㎟, 1000 ㎟ 또는 그 초과인 조직 절편을 포함한다.

[0190] 본원에서 설명되는 방법 및 시스템을 사용하여 결정된 하플로타입은 컴퓨터 리소스, 예를 들어 클라우드 시스템과 같은 네트워크를 통한 컴퓨터 리소스에 할당될 수 있다. 필요한 경우, 컴퓨터 리소스에 저장된 관련 정보를 사용하여 짧은 변이체 콜을 수정할 수 있다. 구조 변이체는 짧은 변이체 콜 및 컴퓨터 리소스에 저장된 정보로부터 조합된 정보를 기초로 하여 검출될 수 있다. 게놈의 다루기 어려운 부분, 예컨대 세그먼트 중복, 구조적 변이가 쉬운 영역, 매우 가변적이고 의학적으로 관련이 있는 MHC 영역, 동원체 및 말단소체 영역, 및 반복 영역, 낮은 서열 정확도, 높은 변이율, ALU 반복, 세그먼트 중복, 또는 관련 기술 분야에 공지된 임의의 다른 관련된 다루기 어려운 부분을 포함하고 이로 제한되지 않는 기타 이질 염색질 영역은 정확도를 높이기 위해 재어셈블리될 수 있다.

[0191] 샘플 유형은 지역적으로 또는 클라우드와 같은 네트워크화된 컴퓨터 리소스에 있는 서열 정보에 할당될 수 있다. 정보의 공급원이 공지된 경우, 예를 들어 정보 공급원이 암 또는 정상 조직에서 유래할 경우, 공급원은 샘플 유형의 일부로서 샘플에 할당될 수 있다. 다른 샘플 유형의 예는 일반적으로 조직 유형, 샘플 수집 방법, 감염의 존재, 감염 유형, 처리 방법, 샘플의 크기 등을 포함하고 이로 제한되지 않는다. 암 게놈과 비교하는 정상 게놈과 같이 완전 또는 부분 비교 게놈 서열이 이용 가능한 경우, 샘플 데이터와 비교 게놈 서열 간의 차이를 결정하고, 선택적으로 출력할 수 있다.

[0192] *하플로타입 페이징 방법*

[0193] 본원에서 개시되는 방법에 의해 생성된 리드쌍은 일반적으로 염색체 내 접촉으로부터 유래하기 때문에, 이형접합성 부위를 함유하는 임의의 리드쌍은 또한 이들의 페이징에 대한 정보를 보유할 것이다. 이 정보를 이용하여, 짧은, 중간, 및 심지어 먼(메가염기) 거리에 걸쳐 신뢰할 만한 페이징을 신속하고 정확하게 수행할 수 있다. 1000개 게놈 트리오(모계/부계/자손 게놈의 세트) 중 하나로부터의 페이즈 데이터에 설계된 실험은 페이징을 신뢰할 수 있게 추론한다. 추가로, 문헌 [Selvaraj *et al.* (*Nature Biotechnology* 31:1111-1118 (2013)]과 유사한 근접 라이게이션을 이용한 하플로타입 재구축은 또한 본원에서 개시되는 하플로타입 페이징 방법과 함께 사용될 수 있다.

[0194] 예를 들어, 근접 라이게이션 기반 방법을 사용하는 하플로타입 재구축은 게놈을 페이징하는 데 있어 본원에서 개시되는 방법에 또한 사용될 수 있다. 근접 라이게이션 기반 방법을 이용한 하플로타입 재구축은 근접 라이게이션 및 DNA 시퀀싱을 하플로타입 어셈블리를 위한 확률론적인 알고리즘과 조합한다. 첫째, 근접 라이게이션 시퀀싱은 염색체 포획 프로토콜, 예컨대 Hi-C 프로토콜을 사용하여 수행한다. 이 방법은 3차원 공간에서 함께 고리를 형성한 2개의 먼 게놈 유전자좌로부터 DNA 단편을 포획할 수 있다. 생성된 DNA 라이브러리의 샷건 DNA 시퀀싱 후, 쌍을 이룬 말단 시퀀싱 리드는 수백 개의 염기쌍에서 수천만 개의 염기쌍에 이르는 범위의 '삽입체 크기'를 갖는다. 따라서, Hi-C 실험에서 생성된 짧은 DNA 단편은 작은 하플로타입 블록을 생성할 수 있고, 긴 단편은 궁극적으로 이들 작은 블록을 함께 연결할 수 있다. 시퀀싱 커버리지가 충분하면, 이 방법은 불연속적인 블록의 변이체를 연결하고 이러한 모든 블록을 단일 하플로타입으로 어셈블리할 가능성을 갖는다. 그 후, 이 데이터는 하플로타입 어셈블리를 위한 확률론적인 알고리즘과 조합된다. 확률론적인 알고리즘은 노드(node)가 이형접합성 변이체에 상응하고 에지(edge)가 변이체를 연결할 수 있는 중첩 서열에 상응하는 그래프를 이용한다. 이 그래프는 시퀀싱 오류 또는 트랜스 상호작용으로 인한 허위 에지를 포함할 수 있다. 이어서, 최대 절단(max-cut) 알고리즘은 입력 시퀀싱 리드 세트에 의해 제공되는 하플로타입 정보와 최대로 일치하는 인색한(parsimonious) 해법을 예측하기 위해 사용된다. 근접 라이게이션은 종래의 게놈 시퀀싱 또는 메이트-쌍 시퀀싱보다 더 큰 그래프를 생성하기 때문에, 계산 시간 및 반복 횟수를 조정하여 하플로타입을 합리적인 속도 및 높은 정확도로 예측할 수 있다. 이어서, 생성된 데이터는 비글(Beagle) 소프트웨어 및 게놈 프로젝트의 시퀀싱 데이터를 사용하여 국소 페이징을 유도하여 높은 해상도 및 정확도로 염색체 스패닝 하플로타입을 생성하는 데 사용될 수 있다.

[0195] *쌍을 이룬 말단을 이용한 페이즈 정보의 결정*

[0196] FFPE-샘플로부터 유도된 쌍을 이룬 말단으로부터 페이즈 정보를 결정하기 위한 방법 및 조성물이 본원에서 추가로 제공된다. 쌍을 이룬 말단은 개시된 임의의 방법 또는 제공된 실시예에서 추가로 설명되는 방법에 의해 생성될 수 있다. 예를 들어, 후속적으로 절단되는 고체 표면에 결합된 DNA 분자의 경우, 자유 말단의 재라이게이션 후에, 재라이게이션된 DNA 세그먼트는 예를 들어 제한 소화에 의해 고상 부착된 DNA 분자로부터 방출된다. 이 방출은 다수의 쌍을 이룬 말단 단편을 생성한다. 일부 경우에는, 쌍을 이룬 말단은 증폭 어댑터에 라이게이션되고, 증폭되고, 짧은 리드 기술로 시퀀싱된다. 이러한 경우, 다수의 상이한 고상 결합 결합된 DNA 분자로부터의 쌍을 이룬 말단은 시퀀싱된 샘플 내에 존재한다. 그러나, 쌍을 이룬 말단 접합부의 어느 쪽에 대해서도, 접합부 인접 서열이 공통적인 분자의 공통 페이즈로부터 유도된다고 자신있게 결론지을 수 있다. 쌍을 이룬 말단이 펑츄에이션 올리고뉴클레오티드와 연결되는 경우, 시퀀싱 리드에서 쌍을 이룬 말단 접합부는 펑츄에이션 올리고뉴클레오티드 서열에 의해 확인된다. 다른 경우에, 쌍을 이룬 말단은 사용된 변형된 뉴클레오티드의 서열에 기초하여 확인될 수 있는 변형된 뉴클레오티드에 의해 연결되었다.

[0197] 대안으로, 쌍을 이룬 말단의 방출 후에, 자유로운 쌍을 이룬 말단을 증폭 어댑터에 라이게이션하고, 증폭할 수 있다. 이러한 경우에, 다수의 쌍을 이룬 말단은 이어서 긴 리드 시퀀싱 기술을 사용하여 판독되는 긴 분자를 생성하기 위해 함께 벌크 라이게이션된다. 다른 예에서, 방출된 쌍을 이룬 말단은 중간의 증폭 단계 없이 서로 벌크 라이게이션된다. 두 경우 모두, 삽입된 리드쌍은 펑츄에이션 서열 또는 변형된 뉴클레오티드와 같은 연결 서열에 인접한 천연 DNA 서열을 통해 확인 가능하다. 연결된 쌍을 이룬 말단은 긴 서열 장치에서 판독되고, 다수의 접합부에 대한 서열 정보가 얻어진다. 쌍을 이룬 말단은 다수의 상이한 고상 결합된 DNA 분자로부터 유래되었기 때문에, 2개의 개개의 쌍을 이룬 말단에 걸치는 서열, 예컨대 증폭 어댑터 서열의 측면에 인접하는 서열은 다수의 상이한 DNA 분자에 매핑되는 것으로 밝혀졌다. 그러나, 쌍을 이룬 말단 접합부의 어느 쪽에 대해서도, 접합부 인접 서열은 공통적인 분자의 공통 페이즈로부터 유래된다고 자신있게 결론지을 수 있다. 예를 들어, 펑츄에이션 분자로부터 유래되는 쌍을 이룬 말단의 경우, 펑츄에이션 서열의 측면에 인접하는 서열은 공통적인 DNA 분자에 자신있게 할당된다. 바람직한 경우에, 개개의 쌍을 이룬 말단은 본원에서 개시되는 방법 및 조성물을 사용하여 연결되기 때문에, 다수의 쌍을 이룬 말단을 단일 리드로 시퀀싱할 수 있다.

[0198] 본원에서 설명되는 방법 및 조성물을 사용하여 생성된 시퀀싱 데이터는 바람직한 실시양태에서 페이징된 디 노보 서열 어셈블리를 생성하고, 페이즈 정보를 결정하고/하거나, 구조적 변이를 확인하기 위해 사용된다.

[0199] *구조적 변이 및 다른 유전적 특징의 결정*

[0200] 도 9a 및 도 9b에서, 재어셈블리된 염색질로부터 DNA의 근접 라이게이션으로부터 생성된 리드쌍의 참조 서열, 예를 들어, GRCh38 상의 매핑된 위치가 GM12878과 참조 서열 사이의 구조적 차이 근처에 도시되어 있는 예가 제시된다. 생성된 각각의 리드쌍은 대각선 위와 아래에 표시된다. 대각선 위에서, 음영은 표시된 눈금의 맵 품질 점수를 나타내고, 대각선 아래에서, 음영은 페이징된 SNP와의 중첩에 기초하여 생성된 리드쌍의 유추된 하플로타입 페이즈를 나타낸다. 일부 실시양태에서, 생성된 플롯은 도 9b에 도시된 바와 같이 인접하는 반복 영역을 갖는 역위를 제시한다. 일부 실시양태에서, 생성된 플롯은 도 9b에 도시된 바와 같이 페이징된 이형접합성 결실에 대한 데이터를 제시한다.

[0201] 참조 서열에 대해 한 개체로부터의 쌍을 이룬 서열 리드를 매핑하는 것은 역위, 결실 및 중복과 같은 연속적인 핵산 또는 게놈 구조의 차이를 확인하기 위해 가장 일반적으로 사용되는 서열 기반 방법이다(Tuzun et al., 2005). 도 9a 및 도 9b는 인간 참조 게놈 GRCh38에 매핑된 GM12878로부터의 재어셈블리된 염색질로부터의 DNA의 근접 라이게이션에 의해 생성된 리드쌍이 2개의 상기 구조적 차이를 나타내는 방법을 보여준다. 구조적 차이를 확인하기 위해 리드쌍 데이터의 민감도와 특이성을 평가하기 위해, 이형접합성 역위의 효과를 시뮬레이션하기 위해 구축된 시뮬레이션된 데이터 세트의 최대 우도 판별기(maximum likelihood discriminator)가 시험되었다. 시험 데이터는 생성된 NA12878 리드의 GRCh38 참조 서열에 대한 매핑으로부터 정의된 길이 L의 간격을 무작위로 선택하고, 각각의 생성된 리드쌍을 무작위로 독립적으로 역위 또는 참조 하플로타입에 할당하고, 이에 따라 매핑된 좌표를 편집함으로써 구축되었다. 비대립유전자 상동성 재조합은 인간 게놈에서 관찰되는 구조적 변이의 대부분을 담당하여, 반복 서열의 긴 블록에서 발생하는 많은 변이 중단점을 초래한다(Kidd et al., 2008). 역위 중단점을 둘러싼 다양한 길이의 반복 서열이 미치는 영향은 그의 W의 거리 내에 매핑된 모든 리드를 제거하여 시뮬레이션되었다. 역위 중단점에서의 반복 서열이 없는 경우, 각각 1 Kbp, 2 Kbp 및 5 Kbp 역위에 대해 민감도(특이도)는 각각 0.76(0.88), 0.89(0.89) 및 0.97(0.94)이었다. 역위 중단점에서 반복(매핑할 수 없는) 서열의 1 Kbp 영역이 시뮬레이션에 사용된 경우, 5 Kbp 역위에 대한 민감도(특이성)는 0.81(0.76)이었다.

[0202] *수행*

[0203] 본원에서 개시되는 기술로 수행된 분석은 높은 정확도로 수행될 수 있다. 분석은 적어도 약 50%, 60%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99.9% 99.99%, 99.999% 또는 그 초과의 정확도로 수행될 수 있다. 분석은 적어도 70%의 정확도로 수행될 수 있다. 분석은 적어도 80%의 정확도로 수행될 수 있다. 분석은 적어도 90%의 정확도로 수행될 수 있다.

[0204] 본원에서 개시되는 기술로 수행된 분석은 높은 특이성으로 수행될 수 있다. 분석은 적어도 약 50%, 60%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99.9%, 99.99%, 99.999% 또는 그 초과의 특이성으로 수행될 수 있다. 분석은 적어도 70%의 특이성으로 수행될 수 있다. 분석은 적어도 80%의 특이성으로 수행될 수 있다. 분석은 적어도 90%의 특이성으로 수행될 수 있다.

[0205] 본원에서 개시되는 기술로 수행된 분석은 높은 민감도로 수행될 수 있다. 분석은 적어도 약 50%, 60%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99.9%, 99.99%, 99.999% 또는 그 초과의 민감도로 수행될 수 있다. 분석은 적어도 70%의 민감도로 수행될 수 있다. 분석은 적어도 80%의 민감도로 수행될 수 있다. 분석은 적어도 90%의 민감도로 수행될 수 있다.

[0206] 본 개시내용의 기술의 사용은 이들이 구현되는 컴퓨터 시스템의 기능을 향상시킬 수 있다. 예를 들어, 본 기술은 주어진 분석의 처리 시간을 적어도 약 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95% 또는 그 초과로 감소시킬 수 있다. 본 기술은 주어진 분석에 대한 메모리 요건을 적어도 약 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95% 또는 그 초과로 감소시킬 수 있다.

[0207] 본 개시내용의 기술의 사용은 이전에 가능하지 않은 분석을 수행할 수 있게 한다. 예를 들어, 본 개시내용의 방법이 없으면 서열 정보로부터 검출될 수 없는 특정 유전적 특징이 상기 서열 정보로부터 검출될 수 있다.

[0208] **컴퓨터 시스템**

[0209] 도 10은 본원에서 제시되는 방법을 실행하도록 프로그래밍되거나 구성된 컴퓨터 시스템(1001)을 보여준다. 컴퓨터 시스템(1001)은 사용자의 전자 장치 또는 전자 장치에 대해 원격으로 위치하는 컴퓨터 시스템일 수 있다. 전자 장치는 이동식 전자 장치일 수 있다.

[0210] 컴퓨터 시스템(1001)은 단일 코어 또는 멀티 코어 프로세서, 또는 병렬 처리를 위한 다수의 프로세서일 수 있는 중앙 처리 장치(CPU, 또한 본원에서 "프로세서" 및 "컴퓨터 프로세서")(1005)를 포함한다. 컴퓨터 시스템(1001)은 또한 메모리 또는 메모리 위치(1010)(예를 들어, 무작위 액세스 메모리, 판독 전용 메모리, 플래시 메모리); 전자 저장 장치(1015)(예를 들어, 하드 디스크), 하나 이상의 다른 시스템과 통신하기 위한 통신 인터페이스(1020)(예를 들어, 네트워크 어댑터), 및 주변 장치(1025), 예컨대 캐시, 다른 메모리, 데이터 저장 장치 및/또는 전자 디스플레이 어댑터를 포함한다. 메모리(1010), 저장 장치(1015), 인터페이스(1020) 및 주변 장치(1025)는 마더보드와 같은 통신 버스(실선)를 통해 CPU(1005)와 연결된다. 저장 장치(1015)는 데이터를 저장하기 위한 데이터 저장 장치(또는 데이터 저장소)일 수 있다. 컴퓨터 시스템(1001)은 통신 인터페이스(1020)의 도움으로 컴퓨터 네트워크("네트워크")(1030)에 동작 가능하게 연결된다. 네트워크(1030)는 인터넷, 인터넷 및/또는 엑스트라넷, 또는 인터넷과 연결된 인트라넷 및/또는 엑스트라넷일 수 있다. 네트워크(1030)는 일부 경우에 원격 통신 및/또는 데이터 네트워크일 수 있다. 네트워크(1030)는 클라우드 컴퓨팅과 같은 분산 컴퓨팅을 가능하게 할 수 있는 적어도 하나의 컴퓨터 서버를 포함할 수 있다. 일부 경우에, 컴퓨터 시스템(1001)의 도움으로 네트워크(1030)는 피어 투 피어(peer-to-peer) 네트워크를 실행할 수 있고, 이는 컴퓨터 시스템(1001)에 연결된 장치가 클라이언트 또는 서버로서 작동할 수 있게 한다.

[0211] CPU(1005)는 프로그램 또는 소프트웨어로 구현될 수 있는 일련의 기계 판독 가능 명령어를 실행할 수 있다. 명령어는 메모리(1010)와 같은 메모리 위치에 저장될 수 있다. 명령어는 CPU(1005)로 유도될 수 있으며, 여기서 이후에 본 개시내용의 방법을 구현하기 위해 CPU(1005)를 프로그래밍하거나 구성할 수 있다. CPU(1005)에 의해 수행되는 명령어의 예는 인출(fetch), 해독(decode), 실행 및 회신(writeback)을 포함할 수 있다.

[0212] CPU(1005)는 집적 회로와 같은 회로의 일부일 수 있다. 시스템(1001)의 하나 이상의 다른 구성 요소가 회로에 포함될 수 있다. 일부 경우에, 회로는 주문형 집적 회로(ASIC)이다.

[0213] 저장 장치(1015)는 드라이버, 라이브러리 및 저장된 프로그램과 같은 파일을 저장할 수 있다. 저장 장치(1015)는 사용자 데이터, 예를 들어 사용자 선호도 및 사용자 프로그램을 저장할 수 있다. 컴퓨터 시스템(1001)은 인트라넷 또는 인터넷을 통해 컴퓨터 시스템(1001)과 통신하는 원격 서버 상에 위치하는 것과 같이 컴퓨터 시스템

(1001)의 외부에 있는 하나 이상의 추가의 데이터 저장 장치를 포함할 수 있다.

[0214] 컴퓨터 시스템(1001)은 네트워크(1030)를 통해 하나 이상의 원격 컴퓨터 시스템과 통신할 수 있다. 예를 들어, 컴퓨터 시스템(1001)은 사용자(예를 들어, 서비스 제공자)의 원격 컴퓨터 시스템과 통신할 수 있다. 원격 컴퓨터 시스템의 예는 개인용 컴퓨터(예를 들어, 휴대용 PC), 슬레이트 또는 태블릿 PC(예를 들어, Apple® iPad, Samsung® Galaxy Tab), 전화기, 스마트폰(예를 들어, Apple® iPhone, Android 지원 기기, Blackberry®) 또는 개인용 정보 단말기를 포함한다. 사용자는 네트워크(1030)를 통해 컴퓨터 시스템(1001)에 액세스할 수 있다.

[0215] 본원에서 설명되는 방법은 예를 들어 메모리(1010) 또는 전자 저장 장치(1015)와 같은 컴퓨터 시스템(1001)의 전자 저장 위치에 저장된 기계(예를 들어, 컴퓨터 프로세서) 실행 가능 코드에 의해 구현될 수 있다. 기계 실행 가능 또는 기계 판독 가능 코드는 소프트웨어의 형태로 제공될 수 있다.

[0216] 사용 동안, 코드는 프로세서(1005)에 의해 실행될 수 있다. 일부 경우에, 코드는 저장 장치(1015)으로부터 검색되고, 프로세서(1005)에 의한 액세스 준비를 위해 메모리(1010) 상에 저장될 수 있다. 일부 상황에서, 전자 저장 장치(1015)는 배제될 수 있고, 기계 실행 가능 명령어가 메모리(1010)에 저장된다.

[0217] 코드는 코드를 실행하도록 조정된 프로세서를 갖는 기계로 사용하기 위해 사전에 컴파일링 및 구성될 수 있거나, 또는 런타임 동안 컴파일링될 수 있다. 코드는 사전 컴파일링 또는 동시 컴파일링(as-compiled) 방식으로 코드를 실행할 수 있도록 선택될 수 있는 프로그래밍 언어로 제공될 수 있다.

[0218] 컴퓨터 시스템(1001)과 같은 본원에서 제공되는 시스템 및 방법의 측면은 프로그래밍으로 구체화될 수 있다. 기술의 다양한 측면은 전형적으로 기계 판독 가능 매체(machine readable medium)의 형태로 수행되거나 구현되는 기계(또는 프로세서) 실행 가능 코드 및/또는 관련 데이터의 형태인 "제품" 또는 "제조품"으로 생각될 수 있다. 기계 실행 가능 코드는 메모리(예를 들어, 판독 전용 메모리, 무작위 액세스 메모리, 플래시 메모리) 또는 하드 디스크와 같은 전자 저장 장치에 저장될 수 있다. "저장" 유형의 매체는 컴퓨터, 프로세서 등의 임의의 또는 모든 유형의 메모리 또는 이들의 관련 모듈, 예컨대 다양한 반도체 메모리, 테이프 드라이브, 디스크 드라이브 등을 포함할 수 있고, 이는 소프트웨어 프로그래밍을 위해 언제든지 비일시적인 저장을 제공할 수 있다. 소프트웨어의 전부 또는 일부는 때때로 인터넷 또는 다양한 다른 원격 통신 네트워크를 통해 전달될 수 있다. 예를 들어, 이러한 전달을 통해, 하나의 컴퓨터 또는 프로세서로부터 또 다른 컴퓨터 또는 프로세서로, 예를 들어 관리 서버 또는 호스트 컴퓨터로부터 응용 프로그램 서버의 컴퓨터 플랫폼으로 소프트웨어를 로딩할 수 있다. 따라서, 소프트웨어 요소를 보유할 수 있는 또 다른 유형의 매체는 유선 및 광 지상통신(optical landline) 네트워크 및 다양한 무선 링크를 통해 로컬 장치 사이의 물리적 인터페이스에 걸쳐 사용되는 것과 같은 광파, 전기파 및 전자기파를 포함한다. 유선 또는 무선 링크, 광 링크 등과 같은 상기 파를 운반하는 물리적 요소가 또한 소프트웨어를 탑재한 매체로 간주될 수 있다. 본원에서 사용되는 바와 같이, 비일시적인 유형의 "저장" 매체에 한정되지 않는 한, 컴퓨터 또는 기계 "판독 가능 매체"와 같은 용어는 실행을 위해 프로세서에 명령어를 제공하는 데 참여하는 임의의 매체를 지칭할 수 있다.

[0219] 따라서, 컴퓨터 실행 가능 코드와 같은 기계 판독 가능 매체는 유형의 저장 매체, 반송파(carrier wave) 매체 또는 물리적 전송 매체를 포함하고 이로 제한되지 않는 많은 형태를 취할 수 있다. 비휘발성 저장 매체는 예를 들어 임의의 컴퓨터(들) 등의 임의의 저장 장치와 같은 광 또는 자기 디스크를 포함할 수 있고, 이는 시스템을 실행하는 데 사용될 수 있다. 휘발성 저장 매체는 그러한 컴퓨터 플랫폼의 메인 메모리와 같은 동적 메모리를 포함한다. 유형의 전송 매체는 동축 케이블, 구리선 및 광섬유(컴퓨터 시스템 내의 버스(bus)를 포함하는 와이어 포함)를 포함할 수 있다. 반송파 전송 매체는 전기 또는 전자기 신호, 또는 음파 또는 광파, 예컨대 무선 주파수(RF) 및 적외선(IR) 데이터 통신 동안 생성되는 것의 형태를 취할 수 있다. 따라서, 컴퓨터 판독 가능 매체의 일반적인 형태는 예를 들어 다음을 포함한다: 플로피 디스크, 가요성 디스크, 하드 디스크, 자기 테이프, 임의의 다른 자기 매체, CD-ROM, DVD, DVD-ROM, 임의의 다른 광 매체, 펀치 카드, 종이 테이프, 구멍 패턴을 갖는 임의의 다른 물리적 저장 매체, RAM, ROM, PROM 및 EPROM, FLASH-EPROM, 임의의 다른 메모리 칩 또는 카트리지, 데이터 또는 명령어를 전송하는 반송파, 상기 반송파를 전송하는 케이블 또는 링크, 또는 그로부터 컴퓨터가 프로그래밍 코드 및/또는 데이터를 판독할 수 있는 임의의 다른 매체. 이러한 형태의 많은 컴퓨터 판독 가능 매체는 실행을 위해 하나 이상의 명령어의 하나 이상의 시퀀스를 프로세서에 전달하는 것과 관련될 수 있다.

[0220] 컴퓨터 시스템(1001)은 예를 들어 훈련된(trained) 알고리즘의 출력 또는 리드를 제공하기 위한 사용자 인터페이스(UI)(1040)를 포함하는 전자 디스플레이(1035)를 포함할 수 있거나 이와 통신할 수 있다. UI의 예는 비제한적으로, 그래픽 사용자 인터페이스(GUI) 및 웹 기반 사용자 인터페이스를 포함한다.

[0221] 본 개시내용의 방법 및 시스템은 하나 이상의 알고리즘에 의해 구현될 수 있다. 알고리즘은 중앙 처리 장치(1005)에 의한 실행시에 소프트웨어에 의해 구현될 수 있다.

[0222] 본원에서 컴퓨터 시스템은 일부 경우에 본원 명세서에 개시되거나 관련 기술 분야의 통상의 기술자에게 공지된 것과 같은 기계 학습 작동을 실행하도록 구성된다.

[0223] **비시퀀싱 기반 검정**

[0224] 비시퀀싱 기반 검정, 예컨대 혼성화(예를 들어, 표지화, 어레이 혼성화, 형광 프로브 혼성화, 예컨대 FISH, 항체 혼성화) 또는 증폭(예를 들어, PCR)을 이용하여 DNA-단백질 복합체(예를 들어, 염색질) 또는 다른 결합된 DNA 복합체(예를 들어, 비드 또는 기타 기재와 복합체화된 DNA)의 유전적 특징(예를 들어, 유전자 재배열)을 검출할 수 있다.

[0225] DNA 복합체(예를 들어, DNA-단백질 복합체, 예컨대 염색질 또는 다른 결합된 DNA 복합체)는 본원에서 논의된 기술을 사용하여 수집될 수 있다. 예를 들어, DNA 복합체는 보존된 샘플(예를 들어, FFPE 샘플)로부터 회수되거나, 또는 단리된 DNA로부터 재구성될 수 있다. 한 예에서, 염색질은 열처리 및 단백질 분해에 의해 보존된 샘플(예를 들어, FFPE 샘플)로부터 방출될 수 있다.

[0226] DNA 복합체는 포획 또는 정제될 수 있다. 예를 들어 DNA 복합체(예를 들어, 염색질)는 고상에 포획될 수 있다. 일부 경우에, 고상은 카르복실화된 기재, 예컨대 카르복실화된 상자성 비드를 포함한다.

[0227] DNA 복합체는 효소적(예를 들어, 제한 효소, 단편화 효소, 트랜스포사제), 열적 및 물리적 단편화를 포함하고 이로 제한되지 않는, 본원에서 개시되는 방법에 의해 단편화되고 라이게이션될 수 있다. 라이게이션 전에 평활 말단화가 수행될 수 있다.

[0228] DNA 복합체는 추가 분석을 위해 분할될 수 있다. 예를 들어, DNA 복합체(예를 들어, 염색질)는 소적(예를 들어, 미세유체 소적), 웰, 어레이 스폿 또는 다른 파티션으로 분할될 수 있다.

[0229] DNA 복합체는 다양한 수단에 의해 분석될 수 있다. 변이체 중단점을 표적화하는(예를 들어, 프라이머쌍으로 표적화하는) 증폭(예를 들어, PCR)을 수행할 수 있다(예를 들어, 소적 PCR과 같은 파티션에서). 형광 올리고뉴클레오티드 프로브를 사용하는 것과 같은 혼성화 검정을 사용하여 변이체 중단점을 표적화할 수 있다. 재배열는 가까운 유전자좌의 근접 라이게이션 확률의 변경에 의한 신호의 변화로 검출될 수 있다. 일부 경우에, Taq-Man 프로브를 사용할 수 있다. 일부 경우에, SYBR 프로브를 사용할 수 있다. 이러한 분석은 예를 들어 소적, 웰, 어레이 스폿 또는 다른 파티션에서 다중화될 수 있다.

[0230] 한 예에서, 염색질은 약한 열처리 및 단백질 분해에 의해 보존된 샘플(예를 들어, FFPE)로부터 방출된다. 방출된 염색질은 상자성 카르복실화된 폴리스티렌 비드를 포함하는 고상에 포획된다. 포획된 염색질에 결합된 DNA는 단편화되고(예를 들어, 효소적으로), 단편화된 말단은 평활화된다. 염색질과 회합된 평활 말단 DNA는 다른 근처의 DNA에 라이게이션된다. 염색체간 변이체의 존재는 예컨대 소적 기반 PCR 또는 형광 올리고뉴클레오티드 프로브 혼성화에 의해 정량된다. 결실 및 역위는 근처의 유전자좌의 근접 라이게이션 가능성의 변화(예를 들어, 증가)로 인해 신호를 변경(예를 들어, 증가)시킨다.

[0231] 재배열 검정은 시퀀싱에 기초한 재배열의 검정을 포함하여, 본원에서 설명되는 바와 같은 시퀀싱 기반 검정과 조합될 수 있다. 예를 들어, PCR 또는 혼성화 검정 후에, 염색질은 본원에서 개시되는 바와 같이 시퀀싱되고, 분석될 수 있다.

[0232] **키트**

[0233] 본원에서 개시된 기술을 수행하기 위한 키트가 본원에서 개시된다. 키트는 상자와 같은 포장재에 담을 수 있으며, 각각의 포장 단위에는 특정한 수의 반응물질이 들어 있다. 일부 경우에, 키트는 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50개 또는 그 초과의 반응물질을 함유한다.

[0234] 본원에서 개시되는 키트는 방법을 실시하고 본원에서 개시되는 조성물을 생성 또는 분석하는 데 필요한 일부의 또는 모든 시약을 포함한다. 일부 경우에, 키트는 본 방법을 실시하고 본원에서 개시되는 조성물을 생성 또는 분석하는 데 필요한 시약의 하위세트를 포함하고, 키트에 포함되지 않지만 종종 시약 판매자로부터 용이하게 입수할 수 있는 시약에 대한 사용 설명서를 선택적으로 포함한다.

[0235] 본원에서 개시되는 일부 키트는 완충제, DNA 결합제, 친화성 태그 결합제, 데옥시뉴클레오티드, 태그 부착된 데

옥시뉴클레오티드, DNA 단편화제, 말단 수복 효소, 리가제, 단백질 제거제, 및 보존된 샘플로부터 게놈의 구조적 정보를 얻을 때 사용하기 위한 사용 설명서를 포함한다. 키트는 완충제, 뉴클레오티드, 정방향 프라이머, 역방향 프라이머 및 열 안정성 DNA 폴리머라제와 같은 PCR용 시약을 선택적으로 포함한다.

[0236] 일부 키트의 완충제는 제한 소화 완충제, 말단 수복 완충제, 라이게이션 완충제, TE 완충제, 세척 완충제, TWB 용액, NTB 용액, LWB 용액, NWB 용액 및 가교결합 반전 완충제 중 적어도 하나를 포함한다. 대표적인 소화 완충제는 DpnII 완충제, 또는 NEB 완충제 2와 기능적으로 유사한 시판되는 완충제이다. 예시적인 라이게이션 완충제는 T4 DNA 리가제 완충제, BSA, 및 트리톤 X-100을 포함한다.

[0237] 키트에 포함되거나 또는 키트 시약과 함께 사용하기 위한 사용 설명서에 언급된 다른 예시적인 시약은 트리스 및 EDTA를 포함하는 TE 완충제, 트리스 및 염화나트륨을 포함하는 세척 완충제, 트리스, EDTA 및 트윈 20 중 하나 이상을 포함하는 TWB 용액, 트리스, EDTA 및 염화나트륨 중 하나 이상을 포함하는 NTB 용액, 트리스, 염화리튬, EDTA 및 트윈 20 중 하나 이상을 포함하는 LWB 용액, 트리스, 염화나트륨, EDTA 및 트윈 20 중 적어도 하나를 포함하는 NWB 용액, 및 트리스, SDS, 및 염화칼슘 중 하나 이상을 포함하는 가교결합 반전 완충제를 포함한다.

[0238] 일부 키트는 스트렙타비딘 비드, 예를 들어 다이나비드와 같은 친화성 태그 결합제를 포함하거나 또는 이와 상용성이도록 구성된다.

[0239] 키트는 dATP, dCTP, dGTP 및 dTTP와 같은 뉴클레오티드, 및 일부 경우에 뉴클레오티드의 비오티닐화된 버전을 포함하거나 또는 이와 상용성이다.

[0240] 본원의 키트에 포함되거나 이와 상용성인 DNA 단편화제는 DpnI과 같은 제한 효소, 트랜스포사제, 뉴클레아제, 초음파 처리 장치, 유체역학적 전단 장치 및 2가 금속 양이온 중 적어도 하나를 포함한다.

[0241] 본원의 키트에 포함되거나 키트와 상용성인 말단 수복 효소는 T4 DNA 폴리머 라제, 클레나우 DNA 폴리머라제 및 T4 폴리뉴클레오티드 키나제 중 적어도 하나를 포함한다.

[0242] 본 발명의 키트 내의 또는 키트와 상용성인 예시적인 리가제는 T4 리가제를 포함한다.

[0243] 본 발명의 키트에 포함되거나 또는 키트와 함께 사용되는 단백질 제거 시약은 페놀 및 프로테이나제, 예컨대 프로테이나제 K, 스트렙토마이세스 그리세우스 프로테아제, 세린 프로테아제, 시스테인 프로테아제, 트레오닌 프로테아제, 아스파르트산 프로테아제, 글루탐산 프로테아제, 메탈로프로테아제 및 아스파라긴 펩티드 리아제를 포함한다.

[0244] 키트는 용매, 예컨대 파라핀과 같은 포매 물질을 제거하기 위해 사용되는 용매를 선택적으로 포함하거나 이 용매와 상용성이다.

[0245] **정의**

[0246] 본원 명세서 및 첨부된 청구범위에서 사용되는 바와 같이, 단수 형태 "하나의"("a", "an" 및 "the")는 문맥에 따라 달리 명백하게 지시하지 않는 한 복수 대상을 포함한다. 따라서, 예를 들어, "콘티그"에 대한 언급은 이러한 콘티그의 복수를 포함하고, "염색체의 물리적 배치의 조사"라는 언급은 염색체의 물리적 배치를 조사하기 위한 하나 이상의 방법 및 관련 기술 분야의 통상의 기술자에게 알려진 그의 등가물 등의 언급을 포함한다.

[0247] 또한, "및"의 사용은 달리 언급되지 않는 한, "및/또는"을 의미한다. 이와 유사하게, "포함한다", "포함하다", "포함하는", "포괄한다", "포괄하다", "포괄하는"은 교환가능하게 사용되고, 제한하려는 의미를 의도하지 않는다.

[0248] 또한, 다양한 실시양태에 대한 설명에서 용어 "포함하는"을 사용하는 경우, 관련 기술 분야의 통상의 기술자는 몇몇의 특정 경우에서, 그 실시양태가 용어 "~로 본질적으로 이루어진" 또는 "~로 이루어진"을 사용하여 대안적으로 설명될 수 있음을 이해할 것이라고 추가로 이해하여야 한다.

[0249] 본원에서 사용되는 용어 "시퀀싱 리드(sequencing read)"는 서열이 결정된 DNA의 단편을 의미한다.

[0250] 본원에 사용되는 용어 "콘티그(contig)"는 DNA 서열의 인접한 영역을 나타낸다. "콘 티그"는 어느 시퀀싱 리드가 인접할 가능성이 큰지를 확인하기 위해서, 관련 기술 분야에 공지된 임의의 수의 방법, 예를 들어 중첩 서열에 대해 시퀀싱 리드를 비교하고/하거나, 시퀀싱 리드를 공지된 서열의 데이터 베이스에 대해 비교함으로써 결정될 수 있다.

[0251] 본원에서 사용되는 용어 "대상체"는 임의의 진핵 또는 원핵 유기체를 지칭할 수 있다.

[0252] 본원에서 사용되는 용어 "네이키드 DNA"는 실질적으로 복합체화된 단백질이 없는 DNA를 지칭할 수 있다. 예를 들어, 이것은 세포핵에서 발견되는 내인성 단백질의 약 50%, 약 40%, 약 30%, 약 20%, 약 10%, 약 5% 또는 약 1% 미만과 복합체화된 DNA를 지칭할 수 있다.

[0253] 본원에서 사용되는 용어 "재구성된 염색질"은 핵산 결합 모이어티를 핵산, 예컨대 네이키드 DNA에 복합체화함으로써 형성된 염색질을 의미할 수 있다. 일부 경우에, 이들 모이어티는 핵산 단백질, 예컨대 핵 단백질 또는 히스톤이지만, 나노 입자와 같은 다른 모이어티도 고려된다.

[0254] 본원에서 사용되는 "리드쌍(read pair)" 또는 "리드-쌍(read-pair)"이라는 용어는 서열 정보를 제공하도록 연결된 2개 이상의 요소를 지칭할 수 있다. 일부 경우에, 리드쌍의 수는 매핑 가능한 리드쌍의 수를 지칭할 수 있다. 다른 경우에, 리드쌍의 수는 생성된 리드쌍의 총수를 지칭할 수 있다.

[0255] 본원에서 사용되는 "조직 샘플"은 개체 또는 잠재적으로 핵산을 포함하는 환경으로부터의 생물학적 샘플을 지칭한다. 예를 들어, 종양은 조직으로 간주되며, 종양에서 채취한 샘플은 조직 샘플을 구성하지만, 일부 경우에 이 용어는 이질적 환경, 예컨대 위 또는 장 부분으로부터 채취한 샘플 또는 서로에 대해 공간적으로 분포된 다수의 공급원으로부터의 핵산을 포함하는 환경 샘플을 지칭한다.

[0256] 숫자에 관하여 본원에서 사용되는 "약"은 해당 숫자의 +/- 10%를 의미한다. 범위와 관련하여 사용될 때, '약'은 범위의 표시된 하한보다 10% 더 낮은 하한 및 범위의 표시된 상한보다 10% 더 큰 상한을 갖는 범위를 나타낸다.

[0257] 본원 명세서에서 사용된 "프로브"는 표적에 대한 결합을 통해 정보를 전달하는 분자를 지칭한다. 예시적인 프로브는 올리고뉴클레오티드 분자 및 항체를 포함한다. 올리고뉴클레오티드 분자는 표적에 어닐링하고 형광 특징을 변화시킴으로써 정보를 전달하거나, 또는 표적에 어닐링하고 표적의 존재를 나타내는 앰플리콘과 같은 생성물의 합성을 촉진함으로써 프로브로서 작용할 수 있다. 즉, 본원 명세서에서 사용되는 용어 프로브는 항체 프로브 및 다른 소분자 프로브뿐만 아니라, 예를 들어 형광 상태의 변화를 유도하는 표적에 대한 혼성화를 통해 신호를 직접 생성함으로써 작용하거나, 또는 표적 존재를 나타내는 앰플리콘의 합성을 촉진함으로써 작용하는 올리고뉴클레오티드 분자를 다양하게 고려한다.

[0258] 본원에서 사용되는 바와 같이, DNA 단백질 복합체는 단백질 및 핵산이 복합체를 형성하기 위해 더 이상 어셈블리되지 않을 때 파괴되거나 붕괴된다. 일부 경우에, 복합체는 완전히 변성되거나 비어셈블리되어, 단백질 DNA 결합이 남아 있지 않게 된다. 대안으로, 일부 경우에 제1 핵산 세그먼트 및 제2 핵산 세그먼트가 포스포디에스테르 결합과는 독립적으로 더 이상 함께 유지되지 않을 때, DNA 단백질 복합체는 실질적으로 파괴된다.

[0259] 달리 정의되지 않는 한, 본원에서 사용되는 모든 기술 및 과학 용어는 본 개시내용이 속하는 기술 분야의 통상의 기술자에게 일반적으로 이해되는 것과 동일한 의미를 갖는다. 본원에서 설명되는 것과 유사한 또는 동등한 임의의 방법 및 시약이 개시된 방법 및 조성물의 실시에 사용될 수 있지만, 예시적인 방법 및 물질이 이제 설명된다.

[0260] 본원의 개시내용은 다음과 같이 번호가 매겨진 실시양태의 부분적인 목록을 참조하여 더욱 명확해진다. 1. 보존된 조직 샘플로부터 게놈의 구조적 정보를 얻는 방법으로서, 단백질 DNA 복합체가 붕괴되지 않도록 보존된 조직 샘플로부터 핵산을 단리하는 단계; 제1 DNA 세그먼트 및 제2 DNA 세그먼트가 공통적인 단백질 DNA 복합체로부터 발생하는 것으로 확인되도록 단백질 DNA 복합체에 태그를 부착하는 단계; 공통적인 DNA 복합체로부터 제1 DNA 세그먼트 및 제2 DNA 세그먼트를 분리하는 단계; 제1 DNA 세그먼트 및 제2 DNA 세그먼트로부터 서열 정보를 생성하는 단계; 및 공통적인 단백질 DNA 복합체를 나타내는 태그 서열을 공유하는 서열 정보를 공통적인 게놈 구조에 할당하는 단계를 포함하는 것인 방법. 2. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 보존된 조직 샘플이 가교결합된 파라핀 포매된 조직 샘플인 방법. 3. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 태그 서열이 복합체를 확인하는 올리고 태그를 포함하는 것인 방법. 4. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 태그 서열이 제1 세그멘트를 제2 세그멘트에 라이게이션함으로써 생성되는 것인 방법. 5. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 단백질 DNA 복합체가 붕괴되지 않도록 보존된 조직 샘플로부터 핵산을 단리하는 단계가 가교결합된 파라핀 포매된 조직 샘플을 크실렌에 접촉시키는 단계를 포함하는 것인 방법. 6. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 단백질 DNA 복합체가 붕괴되지 않도록 보존된 조직 샘플로부터 핵산을 단리하는 단계가 보존된 조직 샘플을 에탄올에 접촉시키는 단계를 포함하는 것인 방법. 7. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 단백질 DNA 복합체가 붕괴되지 않도록 보존된 조직 샘플로부

터 핵산을 단리하는 단계가 샘플을 비등 조건으로부터 보호하는 단계를 포함하는 것인 방법. 8. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 공통적인 DNA 복합체로부터 제1 DNA 세그먼트 및 제2 DNA 세그먼트를 분리하는 단계가 프로테이나제 K 처리를 포함하는 것인 방법. 9. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 보존된 조직 샘플이 조직 내의 그의 입체배열을 반영하는 위치 정보를 보존하는 것인 방법. 10. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 균질화되지 않는 것인 방법. 11. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 적어도 1주일 동안 보관되는 것인 방법. 12. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 적어도 6개월 동안 보관되는 것인 방법. 13. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 수집 지점으로부터 수송되는 것인 방법. 14. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 보존된 조직 샘플이 멸균 환경에서 수집되는 것인 방법. 15. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 비멸균 환경에 위치하는 것인 방법. 16. 보존된 조직 샘플로부터 게놈의 구조적 정보를 얻는 방법으로서, 50 kb 초과의 핵산 단편이 회수되도록 조직 샘플로부터 핵산을 단리하는 단계; 핵산 분자의 제1 DNA 세그먼트 및 제2 DNA 세그먼트가 그들의 공통적인 포스포디에스테르 백본과는 관계없이 함께 유지되도록 하는 적어도 하나의 복합체를 형성시키기 위해 핵산을 다수의 핵산 결합 모이어티와 접촉시키는 단계; 적어도 하나의 복합체의 적어도 하나의 포스포디에스테르 백본을 절단하는 단계; 제1 DNA 세그먼트 및 제2 DNA 세그먼트가 공통의 복합체로부터 발생하는 것으로 확인되도록 적어도 하나의 복합체에 태그를 부착하는 단계; 공통의 복합체로부터 제1 DNA 세그먼트 및 제2 DNA 세그먼트를 분리하는 단계; 제1 DNA 세그먼트 및 제2 DNA 세그먼트로부터 서열 정보를 생성하는 단계; 및 공통적인 단백질 DNA 복합체를 나타내는 태그 서열을 공유하는 서열 정보를 공통적인 게놈 구조에 할당하는 단계를 포함하는 것인 방법. 17. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 보존된 조직 샘플이 가교결합된 파라핀 포매된 조직 샘플인 방법. 18. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 태그 서열이 복합체를 확인하는 올리고 태그를 포함하는 것인 방법. 19. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 태그 서열이 제1 DNA 세그먼트를 제2 DNA 세그먼트에 라이게이션함으로써 생성되는 것인 방법. 20. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 50 kb 초과의 핵산 단편이 회수되도록 보존된 조직 샘플로부터 핵산을 단리하는 단계가 보존된 조직 샘플을 안트라닐레이트 및 포스파닐레이트 중 적어도 하나에 접촉시키는 단계를 포함하는 것인 방법. 21. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 단리가 40℃ 이하의 온도에서 수행되는 것인 방법. 22. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 단리가 40℃ 이하의 온도에서 수행되는 것인 방법. 23. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 공통적인 DNA 복합체로부터 제1 DNA 세그먼트 및 제2 DNA 세그먼트를 분리하는 단계가 프로테이나제 K 처리를 포함하는 것인 방법. 24. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 다수의 핵산 결합 모이어티가 핵 단백질을 포함하는 것인 방법. 25. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 다수의 핵산 결합 모이어티가 트랜스포사제를 포함하는 것인 방법. 26. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 다수의 핵산 결합 모이어티가 히스톤을 포함하는 것인 방법. 27. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 다수의 핵산 결합 모이어티가 핵산 결합 단백질을 포함하는 것인 방법. 28. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 다수의 핵산 결합 모이어티가 나노 입자를 포함하는 것인 방법. 29. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 적어도 하나의 복합체의 적어도 하나의 포스포디에스테르 백본을 절단하는 단계가 제한 엔도뉴클레아제에 접촉시키는 단계를 포함하는 것인 방법. 30. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 적어도 하나의 복합체의 적어도 하나의 포스포디에스테르 백본을 절단하는 단계가 비특이적 엔도뉴클레아제에 접촉시키는 단계를 포함하는 것인 방법. 31. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 적어도 하나의 복합체의 적어도 하나의 포스포디에스테르 백본을 절단하는 단계가 DNA를 전단하는 단계를 포함하는 것인 방법. 32. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 적어도 하나의 복합체의 적어도 하나의 포스포디에스테르 백본을 절단하는 단계가 트랜스포사제에 접촉시키는 단계를 포함하는 것인 방법. 33. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 적어도 하나의 복합체의 적어도 하나의 포스포디에스테르 백본을 절단하는 단계가 토포이소머라제에 접촉시키는 단계를 포함하는 것인 방법. 34. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 보존된 조직 샘플이 조직 내의 그의 입체배열을 반영하는 위치 정보를 보존하는 것인 방법. 35. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 균질화되지 않는 것인 방법. 36. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태

16에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 적어도 1주일 동안 보관되는 것인 방법. 37. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 적어도 6개월 동안 보관되는 것인 방법. 38. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 수집 지점으로부터 수송되는 것인 방법. 39. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 보존된 조직 샘플이 멸균 환경에서 수집되는 것인 방법. 40. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 비멸균 환경에 위치하는 것인 방법. 41. 조직 샘플로부터 공간적으로 분포된 게놈의 구조적 정보를 회수하는 방법으로서, 조직 샘플을 얻는 단계; 고정된 3차원 파라핀 포매된 조직 샘플의 제1 위치로부터 일부를 추출하는 단계; 단백질 DNA 복합체가 붕괴되지 않도록 제1 위치로부터의 일부로부터 핵산을 단리하는 단계; 제1 DNA 세그먼트 및 제2 DNA 세그먼트가 공통적인 단백질 DNA 복합체로부터 생성되는 것으로 확인되도록 단백질 DNA 복합체에 태그를 부착하는 단계; 공통적인 DNA 복합체로부터 제1 DNA 세그먼트 및 제2 DNA 세그먼트를 분리하는 단계; 제1 DNA 세그먼트 및 제2 DNA 세그먼트로부터 서열 정보를 생성하는 단계; 공통적인 단백질 DNA 복합체를 나타내는 태그 서열을 공유하는 서열 정보를 공통적인 게놈 구조에 할당하는 단계; 및 공통적인 게놈 구조를 조직 샘플의 제1 위치에 할당하는 단계를 포함하는 것인 방법. 42. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 조직 샘플이 고정된 3차원 파라핀 포매된 조직 샘플을 포함하는 것인 방법. 43. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 조직 내의 그의 입체배열을 반영하는 위치 정보를 보존하는 것인 방법. 44. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 핵산을 단리하기 전에 균질화되지 않는 것인 방법. 45. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 핵산을 단리하기 전에 적어도 1주일 동안 보관되는 것인 방법. 46. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 핵산을 단리하기 전에 적어도 6개월 동안 보관되는 것인 방법. 47. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 핵산을 단리하기 전에 수집 지점으로부터 수송되는 것인 방법. 48. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 멸균 환경에서 수집되는 것인 방법. 49. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 핵산을 단리하기 전에 비멸균 환경에 위치하는 것인 방법. 50. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 태그 서열이 복합체를 확인하는 올리고 태그를 포함하는 것인 방법. 51. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 태그 서열이 제1 세그멘트를 제2 세그멘트에 라이게이션함으로써 생성되는 것인 방법. 52. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 단백질 DNA 복합체가 붕괴되지 않도록 가교결합된 파라핀 포매된 조직 샘플로부터 핵산을 단리하는 단계가 가교결합된 파라핀 포매된 조직 샘플을 크실렌에 접촉시키는 단계를 포함하는 것인 방법. 53. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 단백질 DNA 복합체가 붕괴되지 않도록 가교결합된 파라핀 포매된 조직 샘플로부터 핵산을 단리하는 단계가 가교결합된 파라핀 포매된 조직 샘플을 에탄올에 접촉시키는 단계를 포함하는 것인 방법. 54. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 단백질 DNA 복합체가 붕괴되지 않도록 가교결합된 파라핀 포매된 조직 샘플로부터 핵산을 단리하는 단계가 샘플을 비등 조건으로부터 보호하는 단계를 포함하는 것인 방법. 55. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 공통적인 DNA 복합체로부터 제1 DNA 세그먼트 및 제2 DNA 세그먼트를 분리하는 단계가 프로테이나제 K 처리를 포함하는 것인 방법. 56. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41에 있어서, 조직 샘플이 고정된 3차원 파라핀 포매된 조직 샘플을 포함하는 것인 방법. 57. 치료 요법 시험 결과를 재평가하는 방법으로서, 환자 집단에서 치료 요법 결과에 관한 데이터를 얻는 단계; 상기 환자 집단의 다수의 환자로부터 고정된 조직 샘플을 얻는 단계; 상기 고정된 조직 샘플로부터 핵산 복합체를 추출하는 단계; 다수의 상기 고정된 조직 샘플에 대해 상기 핵산 복합체를 사용하여 게놈의 구조적 정보를 결정하는 단계; 및 치료 요법 결과에 관련된 게놈의 구조적 정보를 확인하기 위해 치료 요법 결과에 관한 데이터를 게놈의 구조적 정보에 서로 관련시키는 단계를 포함하는 것인 방법. 58. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 57에 있어서, 상기 고정된 조직 샘플로부터 핵산 복합체를 추출하는 단계; 및 다수의 상기 고정된 조직 샘플에 대해 상기 핵산 복합체를 사용하여 게놈의 구조적 정보를 결정하는 단계가 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 1의 방법을 포함하는 것인 방법. 59. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 57에 있어서, 상기 고정된 조직 샘플로부터 핵산 복합체를 추출하는 단계; 및 다수의 상기 고정된 조직 샘플에 대해 상기 핵산 복합체를 사용하여 게놈의 구조적 정보를 결정하는 단계가 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 16의 방법을 포함하는 것인 방법. 60. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 57에 있어서, 상기 고정된 조직 샘플로부터

핵산 복합체를 추출하는 단계; 및 다수의 상기 고정된 조직 샘플에 대해 상기 핵산 복합체를 사용하여 게놈의 구조적 정보를 결정하는 단계가 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 41의 방법을 포함하는 것인 방법. 61. 뉴클레오티드 서열 어셈블리 방법으로서, (a) 고정된 조직 샘플을 제공하는 단계; (b) 상기 고정된 조직 샘플로부터 가교결합된 DNA:단백질 복합체를 회수하는 단계; (c) 상기 가교결합된 DNA:단백질 복합체로부터의 DNA의 제1 섹션을 상기 가교결합된 DNA:단백질 복합체로부터의 DNA의 제2 섹션에 라이게이션하여 라이게이션된 DNA를 형성하는 단계; (d) 상기 가교결합된 DNA:단백질 복합체로부터 상기 라이게이션된 DNA를 추출하는 단계; (e) 상기 라이게이션된 DNA의 라이게이션 접합부의 어느 한 측면 상의 적어도 일부분을 시퀀싱하는 단계; 및 (f) 상기 시퀀싱으로부터의 정보를 이용하여 뉴클레오티드 서열을 어셈블리하는 단계를 포함하는 것인 방법. 62. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 상기 고정된 조직 샘플이 포르말린 고정된 샘플인 방법. 63. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 62에 있어서, 상기 고정된 조직이 포르말린 고정 파라핀 포매된(FFPE) 샘플인 방법. 64. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 상기 가교결합된 DNA:단백질 복합체가 염색질을 포함하는 것인 방법. 65. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 상기 라이게이션이 평활 말단 라이게이션을 포함하는 것인 방법. 66. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 상기 라이게이션 전에, 상기 가교결합된 DNA:단백질 복합체로부터 DNA를 소화하는 단계를 추가로 포함하는 방법. 67. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 66에 있어서, 상기 소화가 제한 효소 소화를 포함하는 것인 방법. 68. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 66에 있어서, 상기 소화 후에, 상기 소화에 의해 생성된 점착성 말단을 충전하여 평활 말단을 생성하는 단계를 추가로 포함하는 방법. 69. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 68에 있어서, 상기 충전이 비오티닐화된 뉴클레오티드를 사용하여 수행되는 것인 방법. 70. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 상기 회수가 상기 가교결합된 DNA:단백질 복합체로부터의 DNA를 고체 지지체에 결합시키는 단계를 포함하는 것인 방법. 71. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 상기 추출이 상기 가교결합된 DNA:단백질 복합체로부터 단백질을 소화시키는 단계를 포함하는 것인 방법. 72. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 상기 정보가 2000개 염기쌍(bp) 초과의 거리에 걸친 장범위 정보를 포함하는 것인 방법. 73. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 72에 있어서, 상기 거리가 10,000 bp 초과인 방법. 74. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 73에 있어서, 상기 거리가 100,000 bp 초과인 방법. 75. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 74에 있어서, 상기 거리가 200,000 bp 초과인 방법. 76. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 상기 회수 전에, 상기 고정된 조직 샘플의 포매 물질을 용해시키는 단계를 추가로 포함하는 방법. 77. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 76에 있어서, 상기 포매 물질이 파라핀을 포함하는 것인 방법. 78. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 조직 내의 그의 입체배열을 반영하는 위치 정보를 보존하는 것인 방법. 79. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 핵산을 단리하기 전에 균질화되지 않는 것인 방법. 80. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 핵산을 단리하기 전에 적어도 1주일 동안 보관되는 것인 방법. 81. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 핵산을 단리하기 전에 적어도 6개월 동안 보관되는 것인 방법. 82. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 핵산을 단리하기 전에 수집 지점으로부터 수송되는 것인 방법. 83. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 가교결합된 파라핀 포매 조직 샘플이 멸균 환경에서 수집되는 것인 방법. 84. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 61에 있어서, 가교결합된 파라핀 포매된 조직 샘플이 핵산을 단리하기 전에 비멸균 환경에 위치하는 것인 방법. 85. 조직 샘플 분석 방법으로서, (a) 고정된 조직 샘플을 제공하는 단계; (b) 상기 고정된 조직 샘플의 제1 부분 및 상기 고정된 조직 샘플의 제2 부분을 수집하는 단계로서, 상기 제1 부분 및 상기 제2 부분은 상기 고정된 조직 샘플의 상이한 영역으로부터 유래되는 것인 단계; (c) 상기 제1 부분으로부터 제1 가교결합된 DNA:단백질 복합체를, 및 상기 제2 부분으로부터 제2 가교결합된 DNA:단백질 복합체를 회수하는 단계; (d) (i) 상기 제1 가교결합된 DNA:단백질 복합체로부터의 DNA의 제1 섹션을 상기 제1 가교결합된 DNA:단백질 복합체로부터의 DNA의 제2 섹션에 라이게이션하여 제1 라이게이션된 DNA를 형성하는 단계, 및 (ii) 상기 제2 가교결합된 DNA:단백질 복합체로부터의 DNA의 제2 섹션을 상기 제2 가교결합된 DNA:단백질 복합체로부터의 DNA의 제2 섹션에 라이게이션하여 제2 라이게이션된 DNA를 형성하는 단계; (e) 상기 제1 가교결합된 DNA:단백질 복합체로부터 상기 제1 라이게이션된 DNA를 및 상기 제2 가교결합된 DNA:단백질 복합체로부터 상기 제2 라이게이션된 DNA를 추출하는 단계; (f) 상기 제1 라이게이션된 DNA 및 상기 제2 라이게이션 DNA를 시퀀싱하는 단계; 및 (g) 상기

시퀀싱으로부터의 정보를 이용하여 제1 뉴클레오티드 서열 및 제2 뉴클레오티드 서열을 어셈블리하는 단계를 포함하는 것인 방법. 86. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 상기 고정된 조직 샘플이 포르말린 고정된 샘플인 방법. 87. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 86에 있어서, 상기 고정된 조직이 포르말린 고정 파라핀 포매된(FFPE) 샘플인 방법. 88. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 상기 제1 가교결합된 DNA:단백질 복합체 및 상기 제2 가교결합된 DNA:단백질 복합체가 각각 염색질을 포함하는 것인 방법. 89. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 상기 (d)(i) 및 (d)(ii)에서의 라이게이션이 평활 말단 라이게이션을 포함하는 것인 방법. 90. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, (d)(i) 및 (d)(ii)에서의 상기 라이게이션 전에, 상기 제1 가교결합된 DNA:단백질 복합체로부터 DNA 및 상기 제2 가교결합된 DNA:단백질 복합체로부터의 DNA를 소화시키는 단계를 추가로 포함하는 방법. 91. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 90에 있어서, 상기 소화가 제한 효소 소화를 포함하는 것인 방법. 92. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 90에 있어서, 상기 소화 후에, 상기 소화에 의해 생성된 점착성 말단을 충전하여 평활 말단을 생성하는 단계를 추가로 포함하는 방법. 93. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 92에 있어서, 상기 충전이 비오티닐화된 뉴클레오티드를 사용하여 수행되는 것인 방법. 94. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 상기 회수가 상기 제1 가교결합된 DNA:단백질 복합체로부터의 DNA 및 상기 제2 가교결합된 DNA:단백질 복합체로부터의 DNA를 고체 지지체에 결합시키는 단계를 포함하는 것인 방법. 95. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 상기 추출이 상기 제1 가교결합된 DNA:단백질 복합체로부터 및 상기 제2 가교결합된 DNA:단백질 복합체로부터 단백질을 소화시키는 단계를 포함하는 것인 방법. 96. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 상기 정보가 2000개 염기쌍(bp) 초과의 거리에 걸친 장범위 정보를 포함하는 것인 방법. 97. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 96에 있어서, 상기 거리가 10,000 bp 초과인 방법. 98. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 93에 있어서, 상기 거리가 100,000 bp 초과인 방법. 99. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 98에 있어서, 상기 거리가 200,000 bp 초과인 방법. 100. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 상기 회수 전에, 상기 고정된 조직 샘플의 포매 물질을 용해시키는 단계를 추가로 포함하는 방법. 101. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 100에 있어서, 상기 포매 물질이 파라핀을 포함하는 것인 방법. 102. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 고정된 조직 샘플이 조직 내의 그의 입체배열을 반영하는 위치 정보를 보존하는 것인 방법. 103. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 고정된 조직 샘플이 핵산을 단리하기 전에 균질화되지 않는 것인 방법. 104. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 고정된 조직 샘플이 핵산을 단리하기 전에 적어도 1주일 동안 보관되는 것인 방법. 105. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 고정된 조직 샘플이 핵산을 단리하기 전에 적어도 6개월 동안 보관되는 것인 방법. 106. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 고정된 조직 샘플이 핵산을 단리하기 전에 수집 지점으로부터 수송되는 것인 방법. 107. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 고정된 조직 샘플이 멸균 환경에서 수집되는 것인 방법. 108. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 85에 있어서, 고정된 조직 샘플이 핵산을 단리하기 전에 비멸균 환경에 위치하는 것인 방법. 109. 보존된 조직 샘플로부터 게놈 재배열을 검출하는 방법으로서, 단백질 DNA 복합체가 파괴되지 않도록 보존된 조직 샘플로부터 단백질 DNA 복합체를 단리하는 단계; 복합체의 노출된 DNA 말단을 라이게이션하여 적어도 하나의 쌍을 이룬 말단 라이게이션 생성물을 형성하는 단계; 적어도 하나의 쌍을 이룬 말단 라이게이션 생성물을 한 쌍의 프로브에 접촉시키는 단계를 포함하고, 여기서 한 쌍의 프로브는 세포 유형에서 재배열된 제1 영역 및 제2 영역에 결합하는 것인 방법. 110. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 단백질 DNA 복합체가, 제1 세그먼트 및 제2 세그먼트가 포스포디에스테르 백본과는 관계없이 함께 유지되도록 단리되는 것인 방법. 111. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 보존된 샘플이 가교결합된 샘플인 방법. 112. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 한 쌍의 프로브가 표지되는 것인 방법. 113. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 한 쌍의 프로브가 형광단을 포함하는 것인 방법. 114. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 한 쌍의 프로브가 올리고뉴클레오티드 프로브를 포함하는 것인 방법. 115. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 110에 있어서, 공통적인 쌍을 이룬 말단 라이게이션 생성물에 대한 한 쌍의 올리고핵산의 어닐링을 검정하는 단계를 추가로 포함하는 방법. 116. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 115에 있어서, 단리된 핵산의 적어도 일부를 시퀀싱하는 단계를 추가로 포함하는 방법. 117. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 한 쌍의 프로브가 정방향 프라이머 및 역방향 프라이머를 포함하고, 상기

정방향 프라이머 및 역방향 프라이머 중 적어도 하나는 재배열에 관련된 DNA 세그먼트에 어닐링하는 것인 방법. 118. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 117에 있어서, 정방향 프라이머 및 역방향 프라이머를 사용하여 핵산 증폭을 수행하는 단계를 추가로 포함하는 방법. 119. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 118에 있어서, 단리된 핵산의 적어도 일부를 시퀀싱하는 단계를 포함하는 방법. 120. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 게놈 재배열이 역위, 삽입, 결실 및 전좌로부터 선택되는 것인 방법. 121. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 보존된 조직 샘플이 포르말린 고정된 샘플인 방법. 122. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 보존된 조직이 포르말린 고정 파라핀 포매된(FFPE) 조직인 방법. 123. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 단리하기 전에, 고정된 조직 샘플의 포매 물질을 제거하는 단계를 추가로 포함하는 방법. 124. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 123에 있어서, 포매 물질이 파라핀을 포함하는 것인 방법. 125. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 단리가 보존된 조직 샘플을 크실렌에 접촉시키는 단계를 포함하는 것인 방법. 126. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 단리가 보존된 조직 샘플을 에탄올에 접촉시키는 단계를 포함하는 것인 방법. 127. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 단리가 샘플을 비등 조건으로부터 보호하는 단계를 포함하는 것인 방법. 128. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 단리가 가교결합된 조직 샘플을 안트라닐레이트 및 포스파닐레이트 중 적어도 하나에 접촉시키는 단계를 포함하는 것인 방법. 129. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 단리가 40℃ 이하의 온도에서 수행되는 것인 방법. 130. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 가교결합된 DNA:단백질 복합체가 염색질을 포함하는 것인 방법. 131. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 109에 있어서, 단리가 가교결합된 DNA:단백질 복합체로부터의 DNA를 고체 지지체에 결합시키는 단계를 포함하는 것인 방법. 132. DNA 세그먼트에서 게놈 재배열을 검출하는 방법으로서, DNA 세그먼트에 대한 게놈 유전자좌 상호작용 정보를 얻는 단계; 및 게놈 유전자좌 상호작용 정보의 관찰된 분포를 게놈 유전자좌 상호작용 정보의 예상된 분포와 비교하는 단계를 포함하는 것인 방법. 133. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 132에 있어서, 관찰된 분포와 예상된 분포 사이의 차이가 DNA 세그먼트의 재배열을 나타내는 것인 방법. 134. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 132에 있어서, 게놈 유전자좌 상호작용 정보가 DNA 세그먼트의 라이게이션된 하위세트에 대한 쌍을 이룬 말단 리드쌍 정보를 포함하는 것인 방법. 135. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 132에 있어서, 게놈 재배열이 역위, 삽입, 결실 및 전좌로부터 선택되는 것인 방법. 136. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 132에 있어서, 관찰된 분포의 상호작용 빈도가 예상된 분포의 상호작용 빈도보다 크고, 게놈 재배열이 역위를 포함하는 것인 방법. 137. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 132에 있어서, 관찰된 분포의 상호작용 빈도가 예상된 분포의 상호작용 빈도보다 작고, 게놈 재배열이 결실을 포함하는 것인 방법. 138. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 132에 있어서, DNA 세그먼트가 가교결합된 조직 샘플로부터 수득되는 것인 방법. 139. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 138에 있어서, 가교결합된 조직 샘플이 포르말린 고정된 샘플인 방법. 140. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 138에 있어서, 가교결합된 조직 샘플이 포르말린 고정 파라핀 포매된(FFPE) 샘플인 방법. 141. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 138에 있어서, 단백질 DNA 복합체가 파괴되지 않도록 가교결합된 조직 샘플로부터 핵산을 단리하기 위해 가교결합된 조직 샘플이 처리되는 것인 방법. 142. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 141에 있어서, 단백질 DNA 복합체가, 제1 세그먼트 및 제2 세그먼트가 포스포디에스테르 백본과는 관계없이 함께 유지되도록 단리되는 것인 방법. 143. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 141에 있어서, 처리 전에, 고정된 조직 샘플의 포매 물질이 용해되는 것인 방법. 144. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 142에 있어서, 포매 물질이 파라핀을 포함하는 것인 방법. 145. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 141에 있어서, 처리가 가교결합된 파라핀 포매된 조직 샘플을 크실렌에 접촉시키는 단계를 포함하는 것인 방법. 146. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 141에 있어서, 처리가 가교결합된 파라핀 포매된 조직 샘플을 에탄올에 접촉시키는 단계를 포함하는 것인 방법. 147. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 141에 있어서, 처리가 샘플을 비등 조건으로부터 보호하는 단계를 포함하는 것인 방법. 148. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 141에 있어서, 처리가 가교결합된 조직 샘플을 안트라닐레이트 및 포스파닐레이트 중 적어도 하나에 접촉시키는 단계를 포함하는 것인 방법. 149. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 141에 있어서, 처리가 40℃ 이하의 온도에서 수행되는 것인 방법. 150. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 141에 있어서, DNA 단백질 복합체가 염색질을 포함하는 것인 방법. 151. 공통적인 보존된 샘플로부터 유래된 제1 DNA 단백질 복합체 및 제2

DNA 단백질 복합체를 포함하는 조성물로서, 상기 제1 DNA 단백질 복합체는 DNA 세그먼트가 공통의 복합체로부터 생성되는 것으로 확인되도록 태그 부착된 DNA 세그먼트를 포함하고, 제1 DNA 단백질 복합체는 공통적인 보존된 샘플의 제1 위치에 할당 가능하고 제2 DNA 단백질 복합체는 공통적인 보존된 샘플의 제2 위치에 할당 가능한 것인 조성물. 152. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 태그 부착된 DNA 세그먼트가 공통의 복합체를 나타내는 서열을 갖는 올리고뉴클레오티드를 사용하여 태그 부착되는 것인 조성물. 153. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 태그 부착된 DNA 세그먼트가, 라이게이션 접합부의 어느 한 측면 상의 특유한 서열이 공통의 복합체에 할당되도록 쌍을 이룬 말단을 형성하는 라이게이션에 의해 태그 부착되는 것인 조성물. 154. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 공통적인 보존된 샘플이 가교결합제와 접촉되는 것인 조성물. 155. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 가교결합제가 포름알데히드 또는 포르말린 중 적어도 하나를 포함하는 것인 조성물. 156. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 가교결합제가 UV 광, 미토마이신 C, 질소 머스타드, 멜팔란, 1,3-부타디엔 디에폭시드, 시스 디아민디클로로백금(II) 및 사이클로포스파미드 중 적어도 하나를 포함하는 것인 조성물. 157. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 보존된 샘플이 포르말린 고정 파라핀 포매된(FFPE) 샘플인 조성물. 158. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 단백질 DNA 복합체가 파괴되지 않도록 보존된 조직 샘플로부터 핵산을 단리하기 위해 보존된 조직 샘플이 처리되는 것인 방법. 159. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 158에 있어서, 단백질 DNA 복합체가, 제1 세그먼트 및 제2 세그먼트가 포스포디에스테르 백본과는 관계없이 함께 유지되도록 단리되는 것인 방법. 160. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 158에 있어서, 처리 전에, 보존된 조직 샘플의 포매 물질을 용해시키는 단계를 추가로 포함하는 방법. 161. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 159에 있어서, 포매 물질이 파라핀을 포함하는 것인 조성물. 162. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 처리가 가교결합된 파라핀 포매된 조직 샘플을 크실렌에 접촉시키는 단계를 포함하는 것인 조성물. 163. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 처리가 가교결합된 파라핀 포매된 조직 샘플을 에탄올에 접촉시키는 단계를 포함하는 것인 조성물. 164. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 처리가 샘플을 비등 조건으로부터 보호하는 단계를 포함하는 것인 조성물. 165. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 처리가 가교결합된 조직 샘플을 안트라닐레이트 및 포스파닐레이트 중 적어도 하나에 접촉시키는 단계를 포함하는 것인 조성물. 166. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 처리가 40℃ 이하의 온도에서 수행되는 것인 방법. 167. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 151에 있어서, 제1 DNA 단백질 복합체 또는 제2 DNA 단백질 복합체가 염색질을 포함하는 것인 방법. 168. 대상체로부터 핵산을 포함하는 보존된 샘플을 얻는 단계; 및 샘플 내의 핵산을 분석함으로써 게놈의 구조적 정보를 도출하는 단계를 포함하는 방법. 제169. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 보존된 샘플이 가교결합되는 것인 방법. 170. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 169에 있어서, 보존된 샘플이 포름알데히드, 포르말린, UV 광, 미토마이신 C, 질소 머스타드, 멜팔란, 1,3-부타디엔 디에폭시드, 시스 디아민디클로로백금(II) 및 사이클로포스파미드 중 적어도 하나를 사용하여 가교결합되는 것인 방법. 171. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 169에 있어서, 보존된 샘플이 포르말린을 사용하여 가교결합되는 것인 방법. 172. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 보존된 샘플이 그 안에 있는 핵산에 대한 위치 정보를 유지하는 것인 방법. 173. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 보존된 샘플이 포매된 샘플인 방법. 174. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 보존된 샘플이 포르말린 고정 파라핀 포매된(FFPE) 샘플 인 방법. 175. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 게놈의 구조적 정보가 참조 게놈에 비해 역위, 삽입, 결실 및 전좌 중 적어도 하나를 나타내는 것인 방법. 176. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 175에 있어서, 참조 게놈이 대상체에 공통적인 종의 야생형 게놈인 방법. 177. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 175에 있어서, 참조 게놈이 대상체의 참조 조직으로부터 수득되는 것인 방법. 178. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 핵산의 제1 세그먼트 및 제2 세그먼트에 대한 페이즈 상태를 나타내는 정보를 도출하는 단계를 포함하는 방법. 179. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 물리적 링키지 정보를 전달하기 위해 샘플의 노출된 핵산 말단에 태그를 부착하는 단계를 포함하는 방법. 180. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 179에 있어서, 태그 부착이 보존된 샘플로부터 방출된 DNA 단백질 복합체에 올리고뉴클레오티드를 라이게이션하여, 올리고뉴클레오티드가 공통의 복합체를 나타내는 정보를 전달하도록 하는 단계를 포함하는 것인 방법. 181. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 180에 있어서, 올리고뉴클레오티드가 복합체에 특이적인 염기 서열

을 포함하는 것인 방법. 182. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 180에 있어서, 올리고뉴클레오티드가 복합체에 특유한 염기 서열을 포함하는 것인 방법. 183. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 179에 있어서, 태그 부착이 복합체의 제1 핵산 세그먼트를 복합체의 제2 세그먼트에 라이게이션하여 쌍을 이룬 말단 분자를 형성하는 단계를 포함하는 것인 방법. 184. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 183에 있어서, 제1 핵산 세그먼트의 일부 및 제2 핵산 세그먼트의 일부를 시퀀싱하는 단계를 포함하는 방법. 185. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 184에 있어서, 제1 핵산 세그먼트의 일부에 공통적인 특유한 서열을 갖는 콘티그 및 제2 핵산 세그먼트의 일부에 공통적인 특유한 서열을 갖는 콘티그를 핵산 어셈블리의 공통의 스캐폴드에 할당하는 단계를 포함하는 방법. 186. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 183에 있어서, 쌍을 이룬 말단 핵산 분자를 핵산 프로브의 세트에 접촉시키는 단계를 포함하는 방법. 187. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 186에 있어서, 핵산 프로브의 세트가 형광 프로브인 방법. 188. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 186에 있어서, 핵산 프로브의 세트가 게놈 구조적 재배열에 관여하는 제1 유전자좌 및 제2 유전자좌에 어닐링하는 것인 방법. 189. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 188에 있어서, 제1 유전자좌와 제2 유전자좌가 게놈 구조적 재배열에 의해 영향을 받지 않는 게놈에서 인접하지 않는 것인 방법. 190. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 188에 있어서, 제1 유전자좌 및 제2 유전자좌가 게놈 구조적 재배열에 의해 영향을 받지 않는 게놈에서 인접하는 것인 방법. 191. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 186-190에 있어서, 핵산 프로브의 세트에 대한 접촉이 재배열을 나타낼 경우 샘플의 핵산을 시퀀싱하는 단계를 포함하는 방법. 192. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 183에 있어서, 쌍을 이룬 말단 핵산 분자를 핵산 프라이머의 세트에 접촉시키는 단계를 포함하는 방법. 193. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 192에 있어서, 핵산 프라이머의 세트가 게놈 구조적 재배열에 관여하는 제1 유전자좌 및 제2 유전자좌에 어닐링하는 것인 방법. 194. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 193에 있어서, 핵산 프라이머의 세트가, 제1 유전자좌 및 제2 유전자좌가 라이게이션된 쌍을 이룬 말단 분자를 형성할 경우 핵산 증폭 반응에서 앰플리콘을 생성하는 것인 방법. 195. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 193에 있어서, 핵산 프라이머의 세트가, 제1 유전자좌 및 제2 유전자좌가 라이게이션된 쌍을 이룬 말단 분자를 형성하지 않을 경우 핵산 증폭 반응에서 앰플리콘을 생성하지 않는 것인 방법. 196. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 188에 있어서, 제1 유전자좌 및 제2 유전자좌가 게놈 구조적 재배열에 의해 영향을 받지 않는 게놈에서 인접하지 않는 것인 방법. 197. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 188에 있어서, 제1 유전자좌 및 제2 유전자좌가 게놈 구조적 재배열에 의해 영향을 받지 않는 게놈에서 인접하는 것인 방법. 198. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 192-197에 있어서, 앰플리콘이 쌍을 이룬 말단 핵산 분자에 접촉된 핵산 프라이머의 세트로부터 생성될 경우 샘플의 핵산을 시퀀싱하는 단계를 포함하는 방법. 199. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 169에 있어서, 단백질 DNA 복합체가 파괴되지 않도록 핵산을 단리하기 위해 보존된 조직 샘플이 처리되는 것인 방법. 200. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 199에 있어서, 단백질 DNA 복합체가, 제1 세그먼트 및 제2 세그먼트가 포스포디에스테르 백본과는 관계없이 함께 유지되도록 단리되는 것인 방법. 201. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 199에 있어서, 보존된 조직 샘플이 보존된 조직 샘플을 크실렌에 접촉시킴으로써 처리되는 것인 방법. 202. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 199에 있어서, 보존된 조직 샘플이 보존된 조직 샘플을 에탄올에 접촉시킴으로써 처리되는 것인 방법. 203. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 199에 있어서, 보존된 조직 샘플이 샘플을 비등 조건으로부터 보호함으로써 처리되는 것인 방법. 204. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 199에 있어서, 보존된 조직 샘플이 보존된 조직 샘플을 안트라닐레이트 및 포스파닐레이트 중 적어도 하나에 접촉시킴으로써 처리되는 것인 방법. 205. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 199에 있어서, 보존된 조직 샘플이 40℃ 이하의 온도에서 처리되는 것인 방법. 206. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 199에 있어서, DNA 단백질 복합체가 염색질을 포함하는 것인 방법. 207. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 보존된 조직 샘플이 조직 내의 그의 입체배열을 반영하는 위치 정보를 보존하는 것인 방법. 208. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 균질화되지 않는 것인 방법. 209. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 보존된 조직 샘플이 핵산을 단리하기 적어도 1주일 동안 보관되는 것인 방법. 210. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 적어도 6개월 동안 보관되는 것인 방법. 211. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 수집 지점으로부터 수송되는 것인 방법. 212. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 보존된 조직 샘플

이 멸균 환경에서 수집되는 것인 방법. 213. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 168에 있어서, 보존된 조직 샘플이 핵산을 단리하기 전에 비멸균 환경에 위치하는 것인 방법. 214. 보존된 샘플로부터 게놈의 구조적 정보를 얻기 위한 키트로서, 완충제, DNA 결합제, 친화성 태그 결합제, 데옥시뉴클레오티드, 태그 부착된 데옥시뉴클레오티드, DNA 단편화제, 말단 수복 효소, 리가제, 단백질 제거제, 및 보존된 샘플로부터 게놈의 구조적 정보를 획득할 때 사용하기 위한 사용 설명서를 포함하는 것인 키트. 215. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 214에 있어서, PCR용 시약을 추가로 포함하는 키트. 216. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 215에 있어서, PCR용 시약이 완충제, 뉴클레오티드, 정방향 프라이머, 역방향 프라이머 및 열 안정성 DNA 폴리머라제를 포함하는 것인 키트. 217. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 214에 있어서, 완충제가 제한 소화 완충제, 말단 수복 완충제, 라이게이션 완충제, TE 완충제, 세척 완충제, TWB 용액, NTB 용액, LWB 용액, NWB 용액 및 가교결합 반전 완충제 중 적어도 하나를 포함하는 것인 키트. 218. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 217에 있어서, 제한 소화 완충제가 DpnII 완충제를 포함하는 것인 키트. 219. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 217에 있어서, 말단 수복 완충제가 NEB 완충제 2를 포함하는 것인 키트. 220. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 217에 있어서, 라이게이션 완충제가 T4 DNA 리가제 완충제, BSA 및 트리톤 X-100을 포함하는 것인 키트. 221. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 217에 있어서, TE 완충제가 트리스 및 EDTA를 포함하는 것인 키트. 222. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 217에 있어서, 세척 완충제가 트리스 및 염화나트륨을 포함하는 것인 키트. 223. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 217에 있어서, TWB 용액이 트리스, EDTA 및 트윈 20을 포함하는 것인 키트. 224. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 217에 있어서, NTB 용액이 트리스, EDTA 및 염화나트륨을 포함하는 것인 키트. 225. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 217에 있어서, LWB 용액이 트리스, 염화리튬, EDTA 및 트윈 20을 포함하는 것인 키트. 226. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 217에 있어서, NWB 용액이 트리스, 염화나트륨, EDTA 및 트윈 20을 포함하는 것인 키트. 227. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 217에 있어서, 가교결합 반전 완충제가 트리스, SDS 및 염화칼슘을 포함하는 것인 키트. 228. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 214에 있어서, DNA 결합제가 염색질 포획 비드를 포함하는 것인 키트. 229. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 228에 있어서, 염색질 포획 비드가 PEG-800 분말, 트리스 완충제, 염화나트륨, EDTA, 계면활성제, TE 완충제 및 세라-맥 비드를 포함하는 것인 키트. 230. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 214에 있어서, 친화성 태그 결합제가 스트렙타비딘 비드를 포함하는 것인 키트. 231. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 230에 있어서, 스트렙타비딘 비드가 디나비드를 포함하는 것인 키트. 232. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 214에 있어서, 데옥시뉴클레오티드가 dATP, dTTP, dGTP 및 dCTP 중 적어도 3개를 포함하는 것인 키트. 233. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 214에 있어서, 비오티닐화된 데옥시뉴클레오티드가 비오티닐화된 dCTP, 비오티닐화된 dATP, 비오티닐화된 dTTP, 및 비오티닐화된 dGTP 중 적어도 하나를 포함하는 것인 키트. 234. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 214에 있어서, DNA 단편화제가 제한 효소, 트랜스포사제, 뉴클레아제, 초음파 처리 장치, 유체역학적 전단 장치 및 2가 금속 양이온 중 적어도 하나인 키트. 235. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 234에 있어서, 제한 효소가 DpnII를 포함하는 것인 키트. 236. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 214에 있어서, 말단 수복 효소가 T4 DNA 폴리머라제, 클레나우 DNA 폴리머라제 및 T4 폴리뉴클레오티드 키나제 중 적어도 하나를 포함하는 것인 키트. 237. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 214에 있어서, 리가제가 T4 DNA 리가제를 포함하는 것인 키트. 238. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 214에 있어서, 단백질 제거제가 프로테아제 및 페놀 중 적어도 하나를 포함하는 것인 방법. 239. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 238에 있어서, 프로테아제가 프로테이나제 K, 스트렙토마이세스 그리세우스 프로테아제, 세린 프로테아제, 시스테인 프로테아제, 트레오닌 프로테아제, 아스파르트산 프로테아제, 글루탐산 프로테아제, 메탈로프로테아제 및 아스파라긴 펩티드 리아제 중 적어도 하나를 포함하는 것인 키트. 240. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 214에 있어서, 포매 물질을 제거하기 위한 용매를 추가로 포함하는 키트. 241. 상기 실시양태 중 어느 하나의 실시양태, 예컨대 실시양태 240에 있어서, 용매가 크실렌, 벤젠 및 톨루엔 중 적어도 하나인 키트.

[0261] 하기 실시예는 본 개시내용을 설명하기 위한 것으로서, 본 개시내용을 제한하고자 의도되지 않는다. 이들은 사용될 수 있는 것들의 전형적인 것이지만, 관련 기술 분야의 통상의 기술자에게 알려진 다른 절차가 대안으로 사용될 수 있다.

[0262] **실시예**

[0263] 실시예 1. FFPE 샘플에서 리드쌍 라이브러리 생성

[0264] AJ GIAB('Gonome In A Bottle') 샘플 GM24149(부계) 및 GM24385(아들)를 호라이즌 디스커버리(Horizon Discovery)로부터 입수하였다. 세포주는 이전에 FFPE에 포매되어 있었다. 섹션당 약 $3x10^5$ 세포를 함유하는, 대략 15-20 마이크로미터 두께의 섹션이 본 실험에 사용되었다. 섹션을 크실렌으로 세척하여 파라핀 왁스를 제거하였다. 에탄올로 섹션을 세척하여 크실렌을 제거하였다. 이어서, 방출된 조직 샘플을 세제 완충제에 재현탁하였다. 이어서, 핵산을 함유하는 샘플을 제한 효소(본 실시예에서는 MboI)로 DNA를 소화시킨 다음, 생성된 오버행을 비오티닐화된 뉴클레오티드로 충전하는 것을 포함하는 말단 라이게이션을 수행하였다. 평활 말단을 함께 라이게이션한 후, 라이게이션된 말단을 방출시켰다. 비오티닐화된 단편을 수득하고, 말단을 시퀀싱하고, 각각 매핑된 콘티그가 샘플 내의 공통적인 핵산 분자 상에 물리적으로 연결되어 있음을 나타내는 리드쌍을 채취하였다.

[0265] 단리된 서열의 위치를 게놈 어셈블리와 비교함으로써 회수된 단편의 쌍을 이룬 말단 사이의 거리를 결정하기 위해 시퀀싱을 수행하였다. 결과는 FFPE-시카고 방법(표 1 - GIAB 칼럼)이 비-FFPE 샘플(표 1 - GIAB 칼럼)에서 수행된 시카고 방법(100 kbp - 200 kbp 삽입체)과 대등하거나(>200 kbp 삽입체) 이보다 큰 긴 거리의 리드쌍 빈도를 유발함을 제시한다. 이들 데이터는 또한 FFPE-시카고 라이브러리의 복잡성 및 원시 시퀀싱 커버리지를 결정하기 위해 분석되었다(표 2). 라이브러리의 복잡성은 라이브러리 내의 상이한 분자의 다양성을 나타낸다.

표 1

삽입체 길이 빈도

[0266]

	GIAB	시카고	시카고
0 < 삽입체 <= 2kbp	48.078%	20.731%	9.92%
2kbp < 삽입체 <= 10kbp	0.458%	6.045%	1.811%
10 kbp < 삽입체 <= 100kbp	0.553%	5.356%	1.884%
100kbp < 삽입체 <= 200kbp	0.171%	0.022%	0.044%
200kbp < 삽입체	1.49%	1.828%	1.499%

표 2

복잡성 및 원시 물리적 커버리지

[0267]

	FFPE	시카고
라이브러리 복잡성 (푸아송(Poisson))	229,196,982	1,013,303,912
150M 리드쌍으로 확대된 원시 물리적 커버리지	5.622 X	66.343 X

[0268] 실시예 2. FFPE-시카고 라이브러리로부터 페이즈 결정

[0269] 실시예 1에서 생성된 시퀀싱 데이터를 사용하여 출발 GIAB 샘플에 존재하는 것으로 알려진 SNP 세트의 페이즈 정보를 결정하였다. 달리 설명하면, 시퀀싱 데이터는 SNP 세트가 동일하거나 상이한 DNA 분자 상에 존재하는지를 결정하기 위해 사용되었다. 이어서, 이들 데이터는 페이즈 콜링의 정확성을 결정하기 위해 GIAB 샘플의 알려진 서열과 비교되었다.

[0270] 표 3의 각각의 빈은 발견되고 다음 빈의 크기까지 일치하는 SNP의 수를 나타낸다. 예를 들어, 첫 번째 줄에는 0-10,000 사이에 132,796개의 SNP가 발견되었고, 99.059%가 올바른 페이즈로 존재하였다. 높은 일치도(>95%)는 약 1.5 MB까지 나타난다(13개 중 1개가 상실된 70-80 kb 빈 및 15개 중 2개가 상실된 1.1 - 1.3 MB 빈 제외). 1.7 - 1.9 MB 범위에서, 7개의 SNP 쌍 페이즈 중 7개가 적절하게 콜링되었다.

[0271] 이러한 데이터로부터, 낮은 수준의 허위 연결에도 불구하고, FFPE-시카고 방법을 사용하여, 심지어 메가베이스 범위까지 적절한 장범위 정보가 결정된다고 결론지었다. 중요한 것은 이러한 '일치도' 예측 비율이 95% 이상이고, 이것은 무작위로 예상되는 50% 성공률보다 유의하게 더 높다는 것이다.

표 3

각각의 빈 내의 SNP

빈	일치도	일치하는 n	불일치하는 n	총 리드쌍
0	99.059	131547	1249	132796
10000	99.346	152	1	153
20000	100	60	0	60
30000	97.619	41	1	42
40000	97.222	35	1	36
50000	100	26	0	26
60000	100	26	0	26
70000	92.308	12	1	13
80000	100	18	0	18
90000	100	8	0	8
100000	98.148	159	3	162
300000	95.238	80	4	84
500000	98	49	1	50
700000	100	28	0	28
900000	96.552	28	1	29
1100000	86.667	13	2	15
1300000	100	16	0	16
1500000	78.571	11	3	14
1700000	100	7	0	7
1900000	85.714	6	1	7
2000000	87.097	27	4	31
3000000	72.222	26	10	36
4000000	84	21	4	25
5000000	69.565	16	7	23
6000000	52.941	9	8	17
7000000	77.778	7	2	9
8000000	61.111	11	7	18
10000000	64.183	267	149	416

[0272]

[0273] 실시예 3. DNA 추출의 개선

[0274] 세제 완충제를 SDS 함유 완충제로부터 트리톤 X 함유 완충제로 변경하고, 실시예 1에 기재된 펠릿을 가시화하면, DNA 추출이 증가되었다. 후속적인 라이브러리 분석은 실시예 1 및 2에서 설명된 라이브러리와 비교할 때 높은 수준의 긴 리드를 유지하면서 상기 라이브러리가 증가된 복잡성을 갖는다는 것을 제시하였. 그 결과를 표 4에 나타내었다.

[0275] 인간 샘플 1 데이터는 실시예 1(FFPE 샘플에서 수행된 평활 말단 라이게이션)에 기재된 바와 같이 처리된 GIAB 샘플로부터 수집하였다. 샘플의 모든 DNA를 라이브러리 제조에 사용하였다.

[0276] 인간 샘플 2 데이터는 실시예 1(FFPE 샘플에서 수행된 평활 말단 라이게이션)에 기재된 바와 같이 처리된 제2

GIAB 샘플로부터 수집하였다. 샘플의 모든 DNA를 라이브러리 제조에 사용하였다.

[0277] 인간 샘플 3 데이터는 실시예 1(FFPE 샘플에서 수행된 평활 말단 라이게이션)에 기재된 바와 같이 처리된 제3 GIAB 샘플로부터 수집하였다. 샘플의 약 500 ng의 DNA를 라이브러리 제조에 사용하였다.

[0278] 인간 샘플 4 데이터는 실시예 1(FFPE 샘플에서 수행된 평활 말단 라이게이션)에 기재된 바와 같이 처리된 제3 GIAB 샘플(인간 샘플 3과 동일한 샘플)로부터 수집하였다. 샘플의 약 50 ng의 DNA를 라이브러리 제조에 사용하였다.

[0279] 인간 샘플 5 데이터는 실시예 1(FFPE 샘플에서 수행된 평활 말단 라이게이션)에 기재된 바와 같이 처리된 제3 GIAB 샘플(인간 샘플 3 및 4와 동일한 샘플)로부터 수집하였다. 샘플의 약 10 ng의 DNA를 라이브러리 제조에 사용하였다.

표 4

개선된 DNA 추출을 갖는 결과

[0280]

프로젝트	인간 (1)	인간 (2)	인간 (3)	인간 (4)	인간 (5)
라이브러리 ID	DPH593_ chicago_ miseq	DPH594_ chicago_ miseq	DPH595_ chicago_ miseq	DPH596_ chicago_ miseq	DPH597_ chicago_ miseq
PCR/광학적 중복	0.166%	0.17%	0.179%	0.546%	1.717%
비매핑	8.157%	8.364%	8.263%	8.559%	8.358%
낮은 맵 품질	10.12%	10.134%	9.99%	9.809%	9.628%
상이한 스캐폴드	16.481%	16.374%	13.779%	10.576%	10.557%
0 < 삽입체 <= 2kbp	57.001%	56.383%	60.844%	65.109%	64.924%
2kbp < 삽입체 <= 10kbp	1.661%	1.794%	1.456%	1.154%	1.001%
10 kbp < 삽입체 <=100kbp	1.438%	1.57%	1.245%	0.979%	0.859%
100kbp < 삽입체 <=200kbp	0.44%	0.476%	0.382%	0.303%	0.266%
200kbp < 삽입체	4.536%	4.735%	3.861%	2.965%	2.69%
라이브러리 복잡성 (푸아송)	1,295,157,213	1,144,409,461	1,321,625,959	497,115,139	107,132,825
150M리드쌍으로 확대된 원시 물리적 커버리지	15.426 X	16.808 X	13.372 X	10.616 X	9.447 X

[0281]

[0282] 실시예 4. FFPE 샘플로부터의 성공적이지 않은 DNA 추출

[0283] BA 종양 샘플은 암 환자로부터 생검되고, 파라핀으로 포매하기 전에 포르말린으로 고정하였다. 그런 다음, FFPE 샘플을 보관하였다. 6개월 후, 환자에 대해 새로운 화합물로 치료를 받는 동안 종양의 진행을 추적하기 위해 임상 연구에 착수하였다. 치료 동안, FFPE 종양 생검 샘플은 몇 주마다 준비되어 보관하였다. 환자는 치료에 매우 잘 반응하였고, 임상 팀은 환자의 특정 암 아형에 대해 더 많은 것을 배우는 데 관심을 가졌다. 연구의 각 단계에서 종양에 존재하는 구조적 변이를 결정하기 위해, 임상 팀은 FFPE 종양 샘플에서 DNA를 추출하기 위해 시도하였다. 유감스럽게도, 회수된 DNA는 고도로 단편화되었고, 단지 짧은 단편만 회수되었다. 이들 짧은 단편 리드는 구조적 변이 결정에 부적합하고, 따라서 중요한 임상 정보는 상실되었다.

[0284] 실시예 5. FFPE 샘플에서 천연 염색질로부터의 성공적인 장거리 데이터

[0285] 실시예 4의 FFPE 종양 샘플을 천연 DNA-단백질 복합체를 보존하기 위해 부드러운 방식으로 처리하였다. DNA 추출은 파라핀 왁스를 제거하기 위해 크실렌으로 FFPE 샘플을 세척하여 수행하였다. 에탄올로 세척하여 크실렌을

제거하였다. 이어서, 샘플을 Hi-C 처리를 실시하기 전에 세제 완충제에 재현탁하였다. FFPR 샘플로부터 단리된 고정된 DNA 단백질 복합체는 비오틴 표지된 뉴클레오티드로 채워진 점착성 오버행을 생성하기 위해 소화하였다. 생성된 평활 말단은 함께 라이게이션되어 동일한 DNA 단백질 복합체로부터 유래된 DNA 서열의 쌍을 이룬 단부를 생성하였다. 쌍을 이룬 말단은 DNA 전단에 의해 DNA 단백질 복합체로부터 방출되고, 스트렙타비딘 비드를 사용하여 단리되었다. 회수된 쌍을 이룬 말단을 시퀀싱 어댑터에 라이게이션시키고, 리드쌍 라이브러리를 생성하기 위해 시퀀싱하였다.

[0286] 임상 팀은 리드쌍 라이브러리를 분석하여, 연구 6개월 전에 채취한 샘플을 포함하여 환자의 종양의 시간에 따른 구조적 변이를 결정할 수 있다. 이 데이터는 암의 아형을 결정하고 동일한 암 아형이 존재하는 다른 환자의 치료 예후를 알리기 위해 사용된다.

[0287] 실시예 6. FFPE 샘플에서 재구성된 염색질로부터의 성공적인 장거리 데이터

[0288] DNA를 실시예 5에 설명된 바와 같이 FFPE 샘플로부터 추출하였다. 네이키드 DNA를 단리하고, 길이가 50 kb를 초과하는 단편을 선택하였다. 재구성된 염색질은 각각의 DNA 단백질 복합체가 단일 DNA 분자를 포함하도록 크기 선택된 DNA를 정제된 염색질 단백질에 결합시킴으로써 생성되었다. 이어서, 이들 DNA 단백질은 포름알데히드를 사용하여 가교결합하였다. 이어서, 가교결합된 복합체를 소화시키고 처리하여, 동일한 DNA 분자로부터 유래된 DNA 서열로부터 쌍을 이룬 말단을 생성하였다. 쌍을 이룬 말단은 리드쌍 라이브러리를 생성하기 위해 시퀀싱하였다. 리드쌍 라이브러리로부터의 데이터는 상기 설명한 환자의 종양 샘플을 특성화하는 데 유용한 페이징 및 구조적 변이 정보를 결정하기 위해 사용되는 긴 거리의 서열 정보를 제시한다.

[0289] 실시예 7. FFPE 샘플로부터 게놈 이질성의 결정

[0290] 실시예 4에서의 FFPE 샘플은 종양의 상이한 영역에서 게놈 이질성을 결정하기 위한 연구에 사용된다. 펀치 생검은 FFPE 종양 샘플의 상이한 세그먼트로부터 채취한 후, 실시예 5에서 설명된 바와 같이 처리하였다. 생성된 데이터는 종양의 성장하는 가장자리를 결정하고, 실시예 5에서 설명된 신규한 화합물 처리에 의한 퇴행 또는 종양 성장 동안 돌연변이 및 구조적 변이가 어떻게 진행되고 축적되는지 또는 사라지는지를 알기 위해 사용된다.

[0291] 실시예 8. FFPE의 가용화 및 샘플 용해

[0292] 1 밀리리터의 크실렌을 FFPE 샘플에 첨가하고, 파라핀이 용해될 때까지 볼텍싱하였다. 샘플을 분당 14,000번의 회전으로 2분 동안 원심분리하였다. 크실렌은 부드럽게 제거되었다. 1 밀리리터의 100% 에탄올을 첨가하고, 샘플을 볼텍싱하여 튜브의 내벽으로부터 세포 펠릿을 분리시켰다. 샘플을 최고 속도로 2분 동안 다시 원심분리한 다음, 에탄올을 제거하였다. 펠릿을 공기 건조하였다. 펠릿이 완전히 건조되면, 50 마이크로리터의 용해 완충제(50 mM 트리스 pH 8, 50 mM NaCl, 1% SDS, 0.15% 트리톤, 1 mM EDTA)을 샘플에 첨가하였다. 샘플을 부드럽게 진탕하면서 37℃에서 15분 동안 인큐베이션하였다. 전체 샘플을 1.5 mL 튜브로 옮겼다. 샘플을 반복적으로 피펫팅하여 세포 펠릿을 붕괴시켰다. 샘플에 100 ㎕의 SPRI(고상 가역적 고정) 비드를 2:1의 SPRI 비드 대 가용성 염색질 비율로 첨가한 후, 실온에서 10분 동안 인큐베이션하였다. SPRI 비드를 2회 세척하였다. SPRI-비드 단리된 샘플은 시카고 또는 Hi-C와 같은 하류 기술에 사용된다.

[0293] 실시예 9: FFPE 샘플은 장범위 게놈 링키지 정보를 보존한다

[0294] 게놈 링키지 데이터를 추출하기 위해 본 개시내용의 방법에 따라 FFPE 샘플을 수득하고 처리하였다. 도 11a는 3개의 샘플의 분석 결과를 보여준다. 인간 세포 배양물(적색, 1103) 및 비장 조직(녹색, 1102) FFPE 샘플을 수득하고, 본원의 개시내용의 방법에 따라 처리하여 게놈 링키지 데이터를 추출하였다. 쌍을 이룬 말단은 hg19 참조물에 매핑되고, 각각의 리드쌍의 리드 사이의 물리적 거리가 계산되었다. 이 데이터는 Hi-C 방법(청색, 1101)으로 세포 배양물 샘플을 사용하여 준비된 데이터와 비교되었다. x축은 리드 사이 물리적 거리(Mb)로 비닝된 리드쌍을 보여준다(왼쪽에서 오른쪽으로 0.01, 0.1, 1, 10 및 100의 축 수치). y축은 리드쌍의 비율을 보여준다(위에서 아래로 0.01, 0001, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}, 10^{-10}, 10^{-11} 및 10^{-12}의 축 수치).

[0295] 실시예 10: 장범위 게놈 링키지 정보를 추출하기 위해 처리된 FFPE 샘플에서의 SNP 일치도

[0296] 도 11b는 본 개시내용의 방법에 따라 장범위 게놈 링키지 데이터를 생성하기 위해 처리된 아슈케나지 부계(GM24149) 세포 배양 FFPE 샘플의 분석 결과를 보여준다. 이 데이터는 쌍을 이룬 말단 리드 둘 모두에 존재하는 높은 신뢰도의 SNP에 대해 필터링되었다. 이 필터링된 데이터세트는 2개의 리드 사이의 물리적 거리(x축)에 기초하여 빈으로 구성되었고, 일치하는 SNP 쌍의 비율은 각각의 빈에 대해 계산되었다(y축). 상부의 적색 선(1111)은 일치하는 SNP를 보여주고, 하부의 청색 선(1112)은 참조를 위한 무작위 일치도를 보여준다.

[0297] 실시예 11: FFPE 샘플은 구조적 변이(SV) 확인을 허용하는 장범위 게놈 링키지 정보를 보존한다

[0298] 데이터를 또한 아슈케나지 부계(GM24149) 세포 배양물 FFPE 샘플로부터 추출하고, hg19 참조물에 대해 리드쌍을 매핑함으로써 구조적 변이체의 존재를 분석하였다. 쌍을 이룬 리드의 중간점은 도 11c 및 도 11d의 x축 상에 도시되고, 상응하는 물리적인 분리는 y축 상에 도시된다. 맵 품질 점수는 범례에 표시된 바와 같이 각각의 데이터 점의 회색조로 표시된다.

[0299] 도 11c는 상실된 게놈 세그먼트에 상응하는 중간점을 갖는 리드쌍의 낮은 밀도를 기초로 하여 ~100 Kb 1번 염색체 결실이 명백하다는 것을 보여준다.

[0300] 도 11d는 별표 아래의 리드의 예상보다 더 높은 밀도를 기초로 하여 ~4 Mb 8번 염색체 역위가 명백하다는 것을 보여준다. 역위는 일반적으로 낮은 맵 품질 점수를 유도하는 반복 영역에 통상적으로 인접하여 위치한다.

[0301] 실시예 12. 샘플 수집, 후속 분석 및 처리 선택

[0302] 환자는 조직을 제거하기 위해 수술을 받는다. 조직을 멸균 환경에서 절제하고, 포르말린 내에 침적시킨다. 수집에 따라 조직의 균질화는 일어나지 않는다.

[0303] 조직을 보존하고, 환자를 모니터링한다. 환자는 절제 부위에서 재성장이 일어나는 것으로 관찰된다. 조직은 보존된 조직의 내부 및 둘레를 포함하는 위치로부터 핵산 단백질 복합체를 절제하는 것을 비롯한 실험실 환경에서의 분석을 거친다.

[0304] 게놈 정보는 보존된 조직으로부터 수득된 핵산 단백질 복합체로부터 수득된다. 종양 전이와 관련된 특정 게놈 입체배열(genomic configuration)을 나타내는 주변 조직으로부터 게놈 재배열이 확인된다.

[0305] 화학요법에 의한 치료는 종양 전이에 관여하는 게놈 입체배열과 관련하여 공지된 효능을 기초로 하여 선택된다. 환자에게 화학요법에 의한 치료를 실시하고, 종양 재성장의 중지가 관찰된다.

[0306] 실시예 13. 약물 시험 재평가

[0307] 약물 시험은 공통 종양 유형을 갖는 개체에 대해 수행된다. 종양 샘플은 약물 시험과 함께 채취된다. 치료받는 개체의 하위 세트가 치료에 긍정적으로 반응하지만, 전체적으로 치료는 약물 개발을 보증하기에 충분한 효능을 갖는 것으로 관찰되지 않는다.

[0308] 치료된 집단의 샘플을 샷건 게놈 시퀀싱에 적용하였다. 짧은 리드 서열 정보가 수득되지만, 실질적인 게놈의 구조적 정보는 수득되지 않는다. 단일 뉴클레오티드 다형성 정보와 같은 개별 서열 정보는 치료 효능과 상관관계가 있는 것으로 관찰되지 않는다.

[0309] 상당한 시간이 경과한 후에, 샘플을 재평가한다. 복합체 완전성이 보존되도록 샘플에 대해 핵산 단백질 복합체 절제를 실시하고, 본원에서 개시되는 바와 같은 분석을 수행한다.

[0310] 복합체를 단리하고, 노출된 핵산 말단을 라이게이션하여 쌍을 이룬 말단 단편을 형성한다. 쌍을 이룬 말단 단편은 라이게이션 부위에 도입된 비오티닐화된 염기를 사용하여 단리된다.

[0311] 리드쌍은 라이게이션 접합부의 어느 한 측면 상의 서열 정보를 얻기 위해 시퀀싱된다. 리드쌍 정보는 분석되고, 샘플의 하위세트는 샷건 시퀀싱 분석으로부터 명백하지 않은 게놈 재배열을 포함하는 것으로 관찰된다.

[0312] 약물 반응은 게놈의 구조적 정보에 비추어 재평가되고, 특정 재배열은 치료 효능과 상관관계가 있는 것으로 관찰된다. 치료 효능과 상관관계가 있는 게놈 재배열은 반응자를 확인하기 위한 마커로서 개발되고, 약물은 장애를 치료하기 위한 마커에 대한 시험과 함께 사용된다.

[0313] 실시예 14. 무서열 재배열 검출

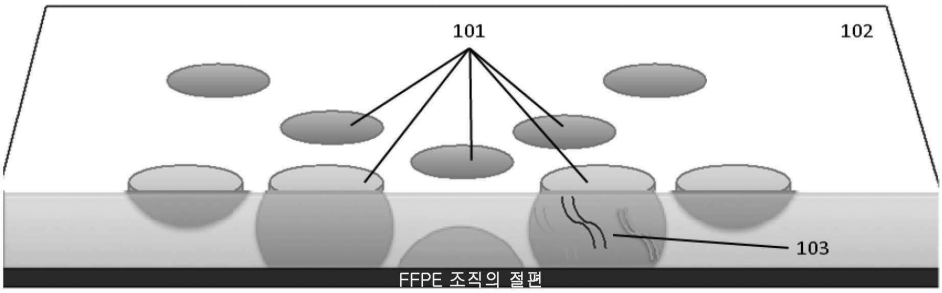
[0314] 쌍을 이룬 말단 라이브러리는 다수의 보존된 샘플로부터 생성된다. 라이브러리는 암과 관련된 게놈 전좌 동안 같은 페이즈로 존재하는 것으로 알려진 게놈 영역에 어닐링하는 프라이머를 사용하여 프로빙된다.

[0315] 라이브러리는 샘플의 하위세트에 대해 보다 높은 빈도를 갖는 전좌된 세그먼트 사이의 물리적 링키지를 나타내는 앰플리콘을 생성하는 것으로 관찰된다. 앰플리콘을 생성하는 라이브러리는 시퀀싱 및 쌍을 이룬 말단 분석을 거치고, 암에 관련된 것으로 의심되는 전좌를 독립적으로 보유하는 것으로 밝혀졌다. 전좌는 동일하지 않고 전좌된 세그먼트의 배향 위치 및 근접성이 다르기 때문에, 게놈의 직접적인 PCR 분석을 통해 대부분의 전좌가 검출되지 않을 수 있다. 그러나, 라이게이션된 쌍을 이룬 말단 라이브러리 생성을 통해, 올리고뉴클레오티드 프라

이머는 전좌의 존재에 대해 샘플을 프로빙할 때 효과적이다. 이것은 하류 서열 분석을 위해 샘플의 하위세트로부터 라이브러리를 선택함으로써 자원을 보존할 수 있도록 한다.

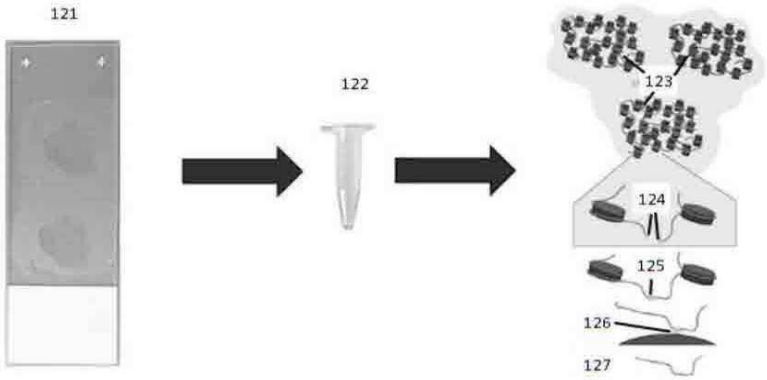
도면

도면1a

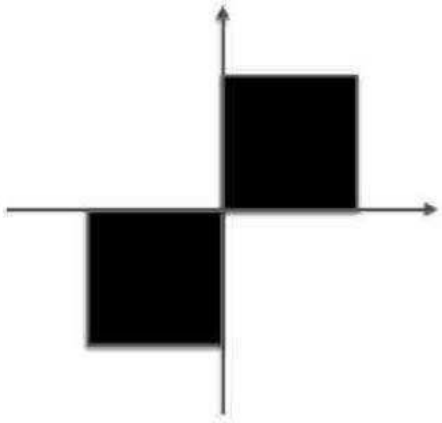


도면1b

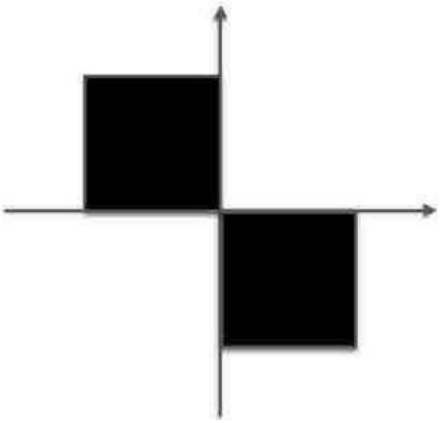
도면1c



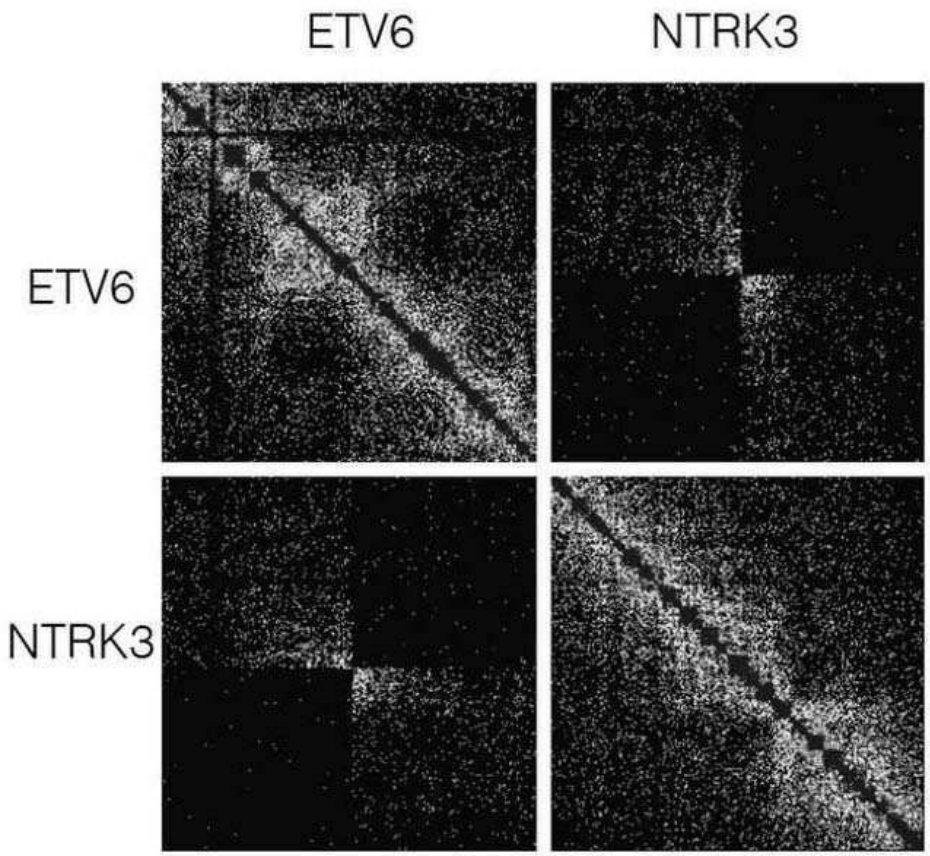
도면2a



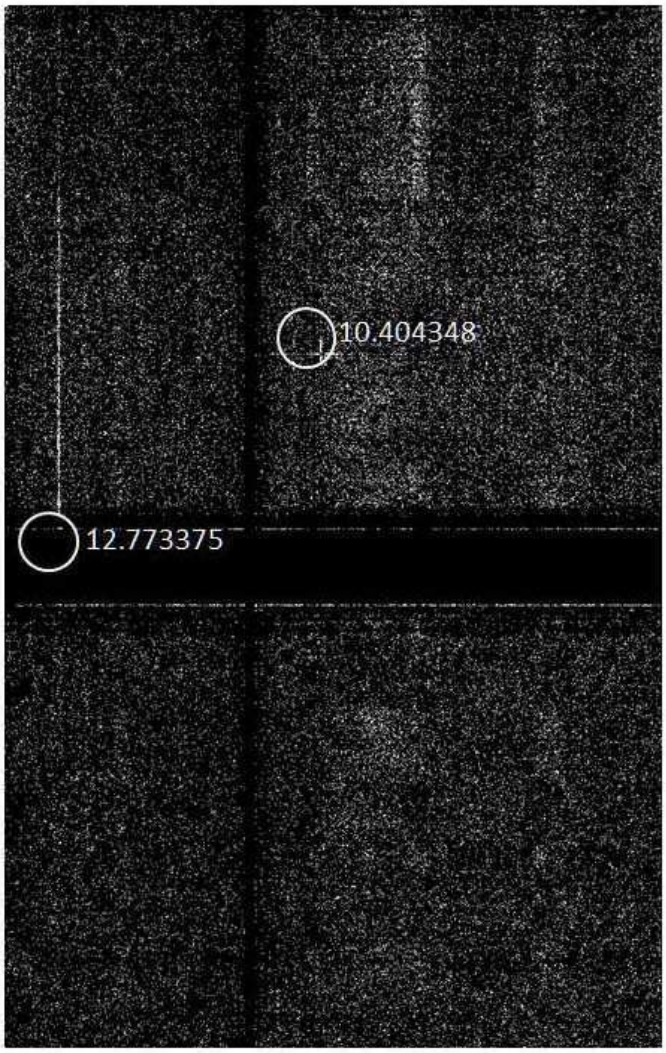
도면2b



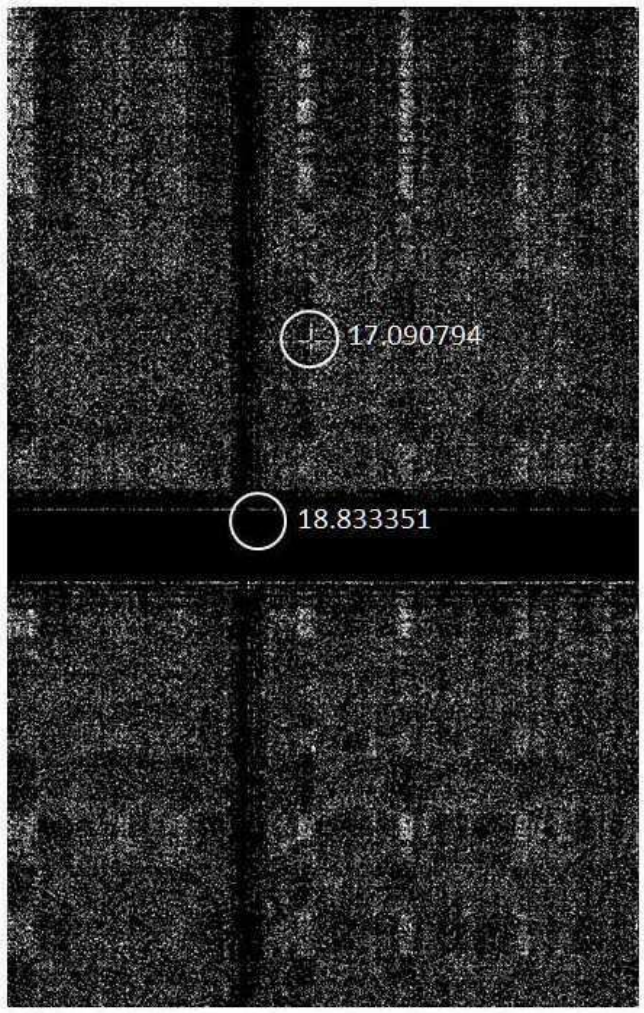
도면3



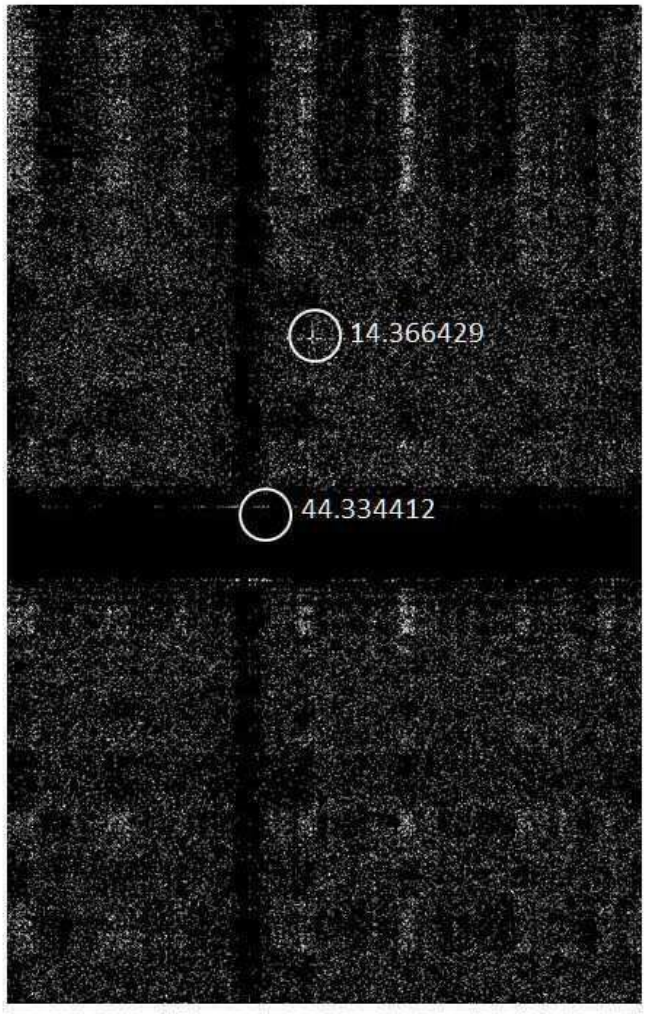
도면4a



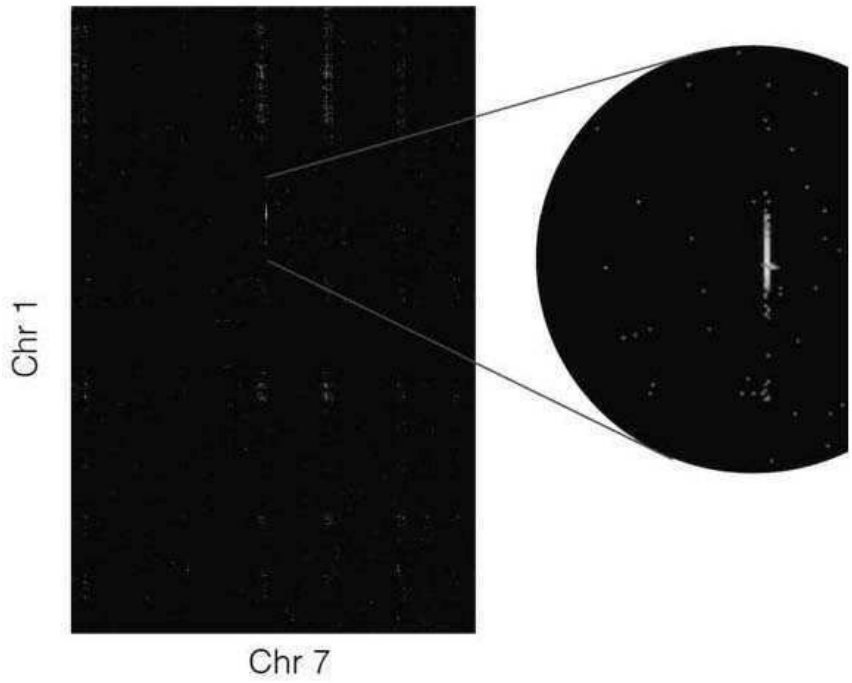
도면4b



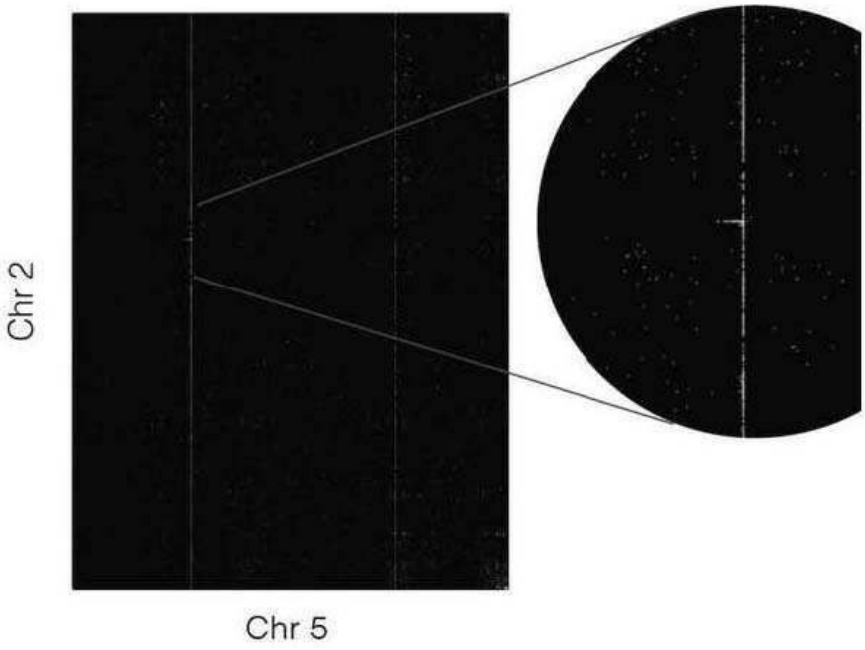
도면4c



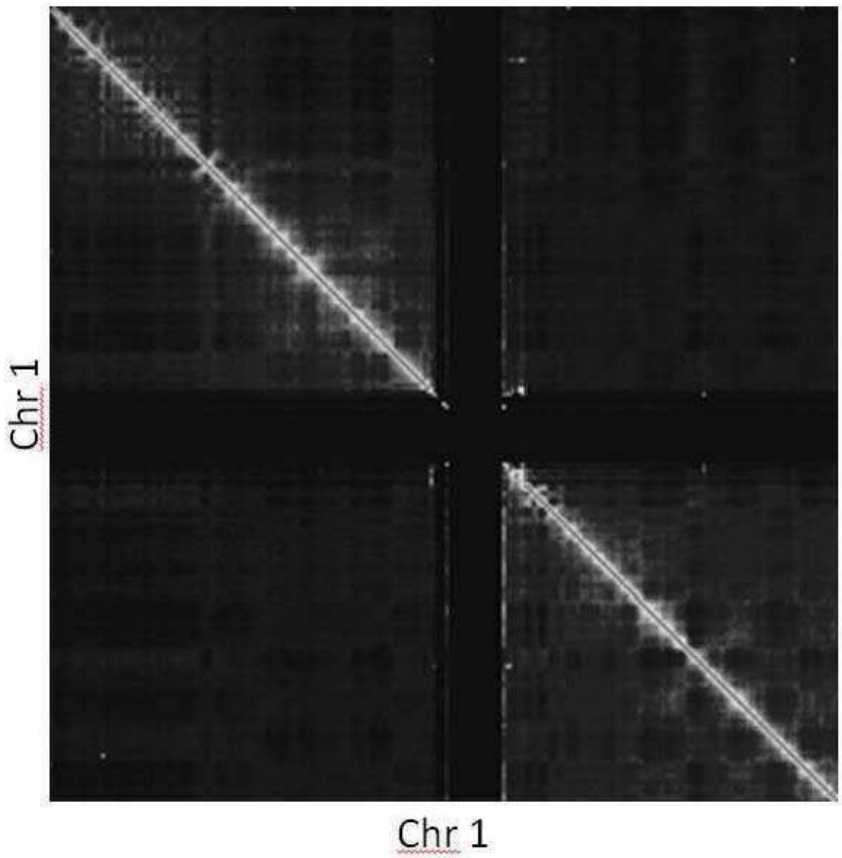
도면5a



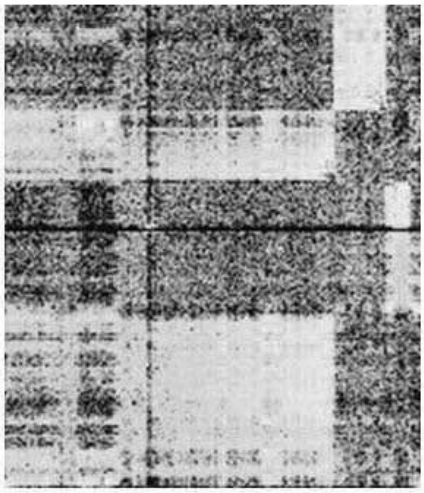
도면5b



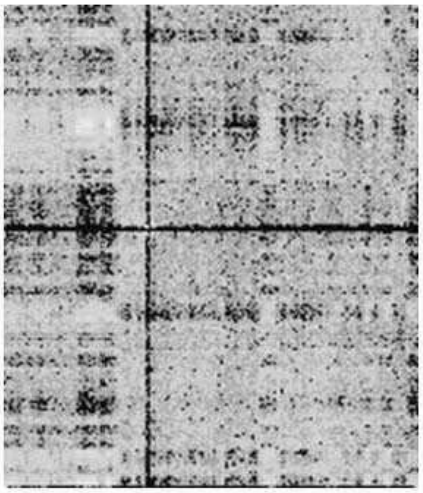
도면5c



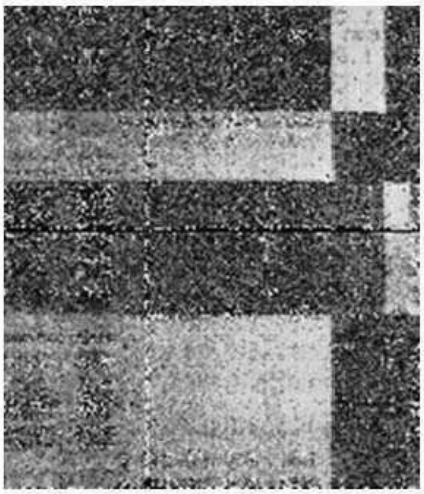
도면6a



샘플

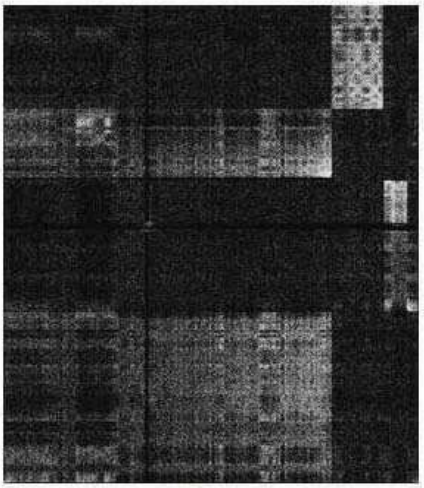


중간

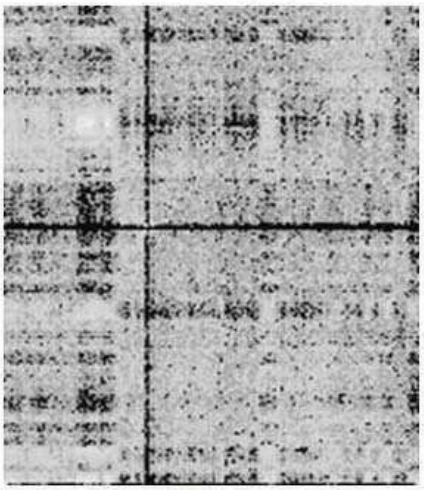


샘플/중간

도면6b



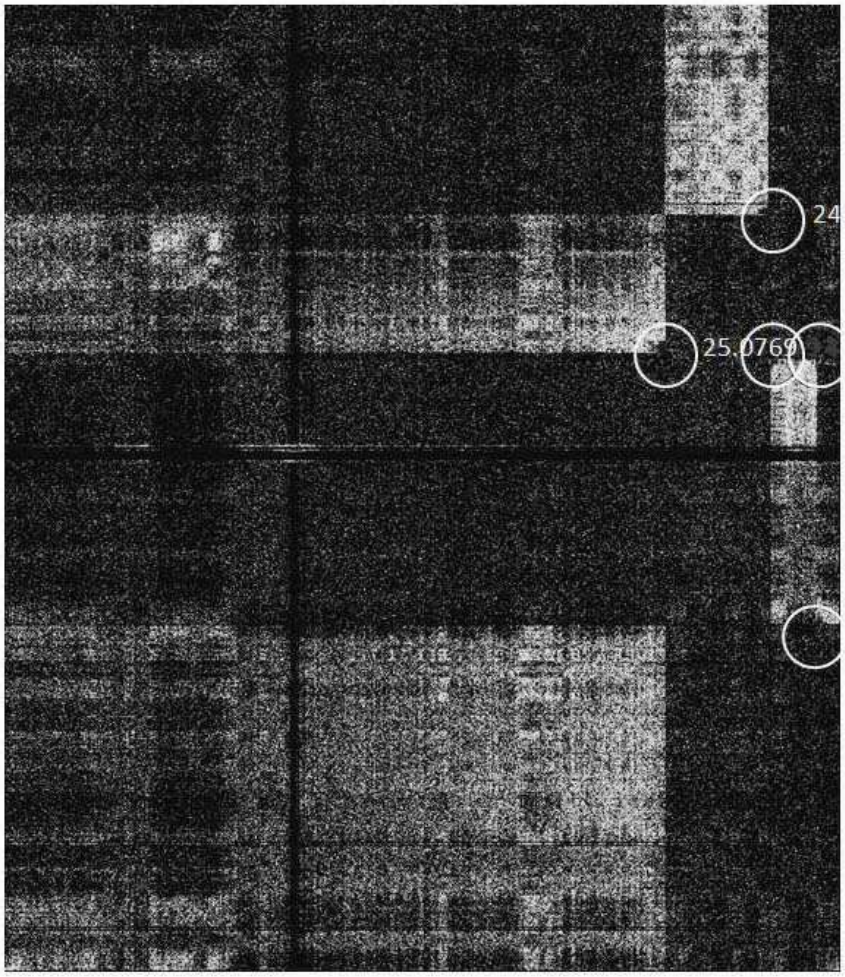
샘플



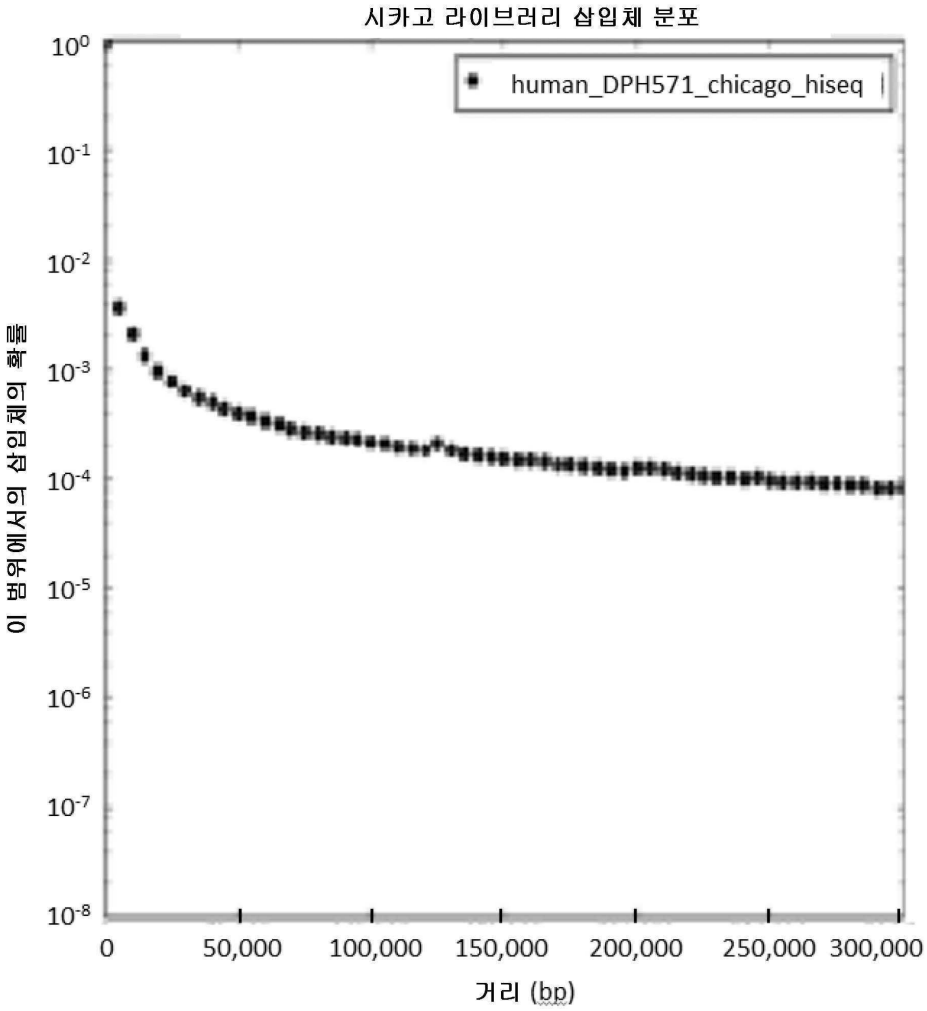
중간

샘플/중간

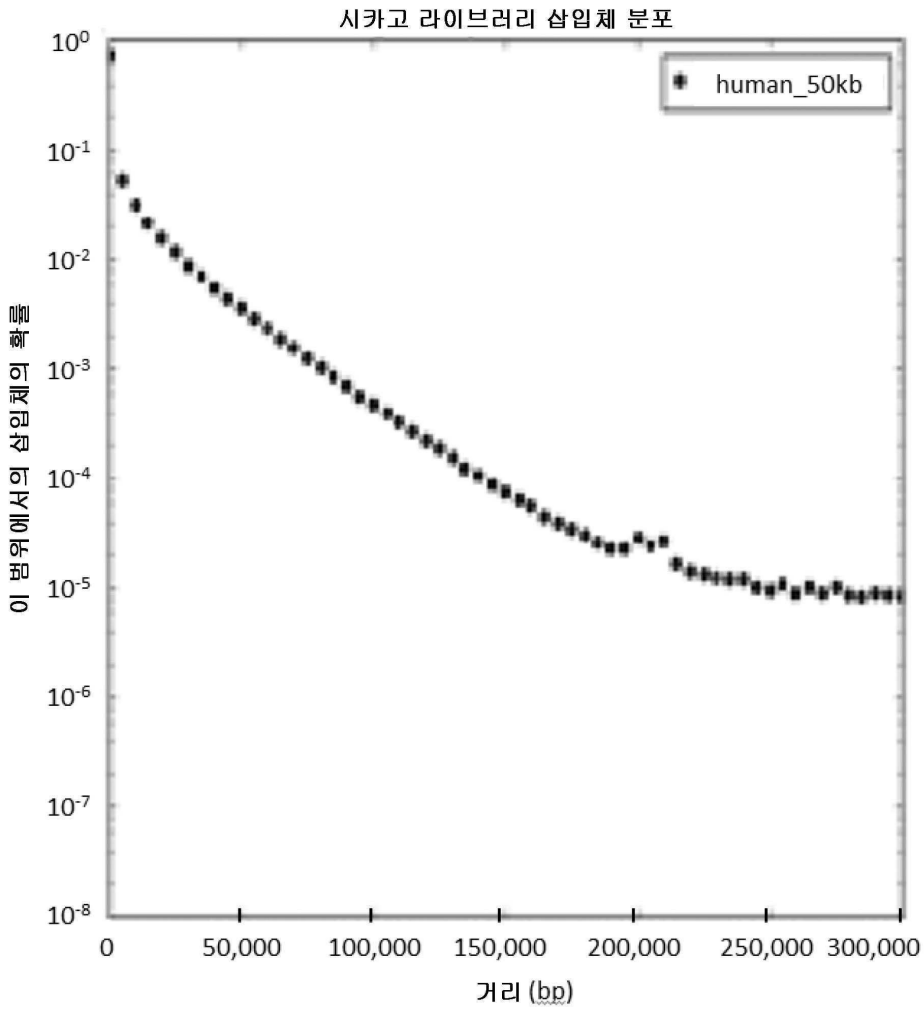
도면7



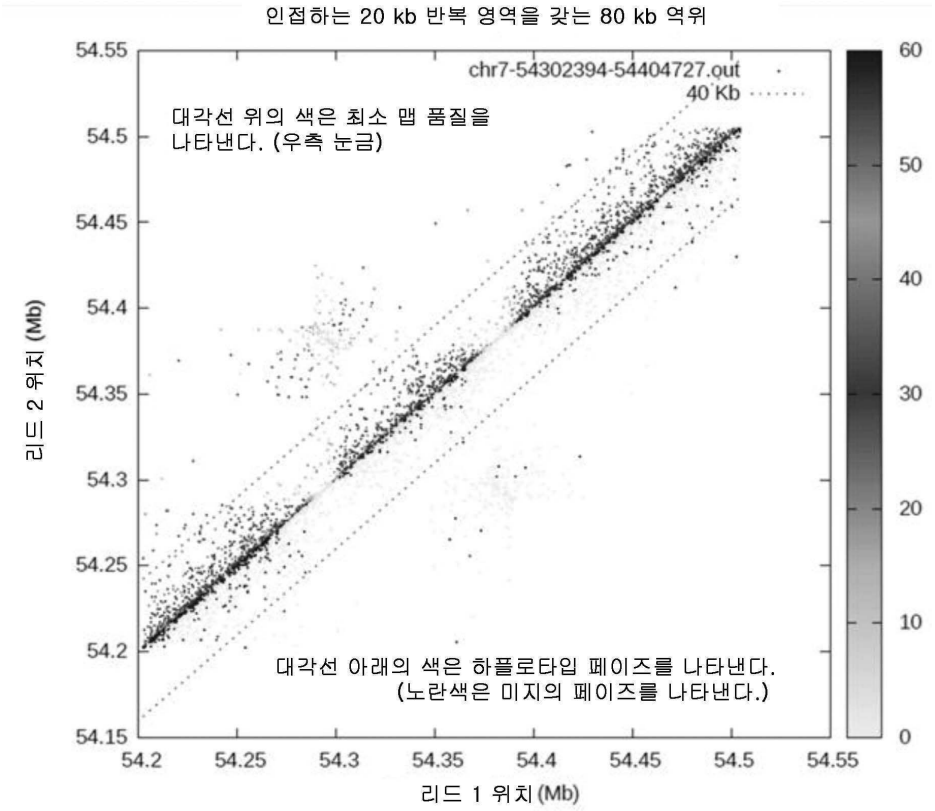
도면8a


시카고 라이브러리 삽입체 분포
human_DPH571_chicago_hiseq
이 범위에서의 삽입체의 확률
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^-8
0
50,000
100,000
150,000
200,000
250,000
300,000
거리 (bp)

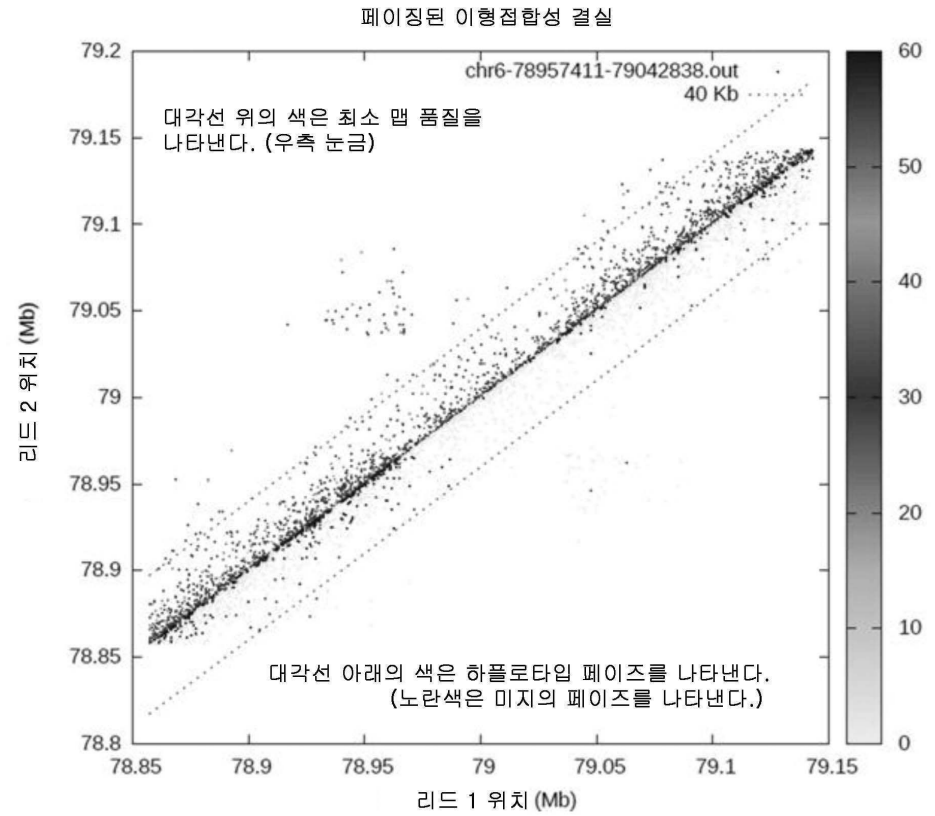
도면8b


시카고 라이브러리 삽입체 분포
human_50kb
이 범위에서의 삽입체의 확률
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^-8
0
50,000
100,000
150,000
200,000
250,000
300,000
거리 (bp)

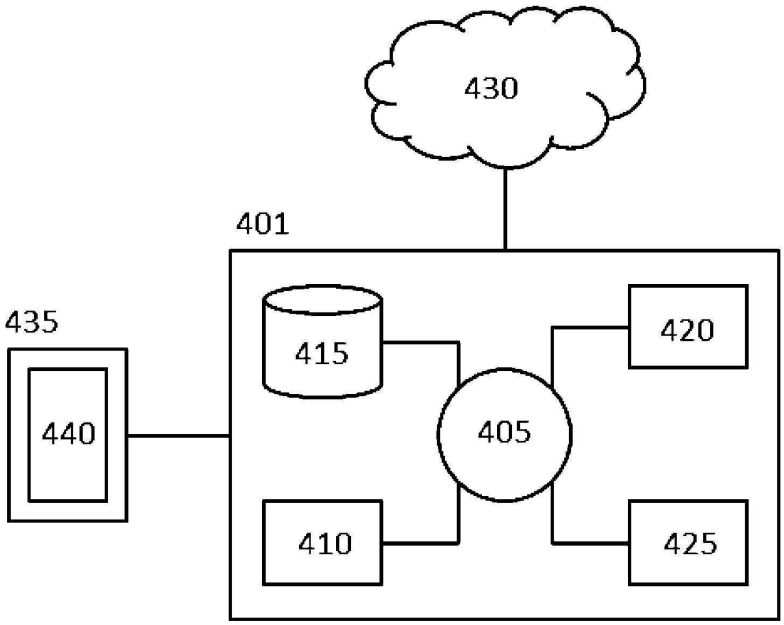
도면9a


인접하는 20 kb 반복 영역을 갖는 80 kb 역위
chr7-54302394-54404727.out
40 Kb
대각선 위의 색은 최소 맵 품질을
나타낸다. (우측 눈금)
대각선 아래의 색은 하플로타입 페이즈를 나타낸다.
(노란색은 미지의 페이즈를 나타낸다.)
리드 2 위치 (Mb)
리드 1 위치 (Mb)
54.55
54.5
54.45
54.4
54.35
54.3
54.25
54.2
54.15
54.2
54.25
54.3
54.35
54.4
54.45
54.5
54.55
60
50
40
30
20
10
0

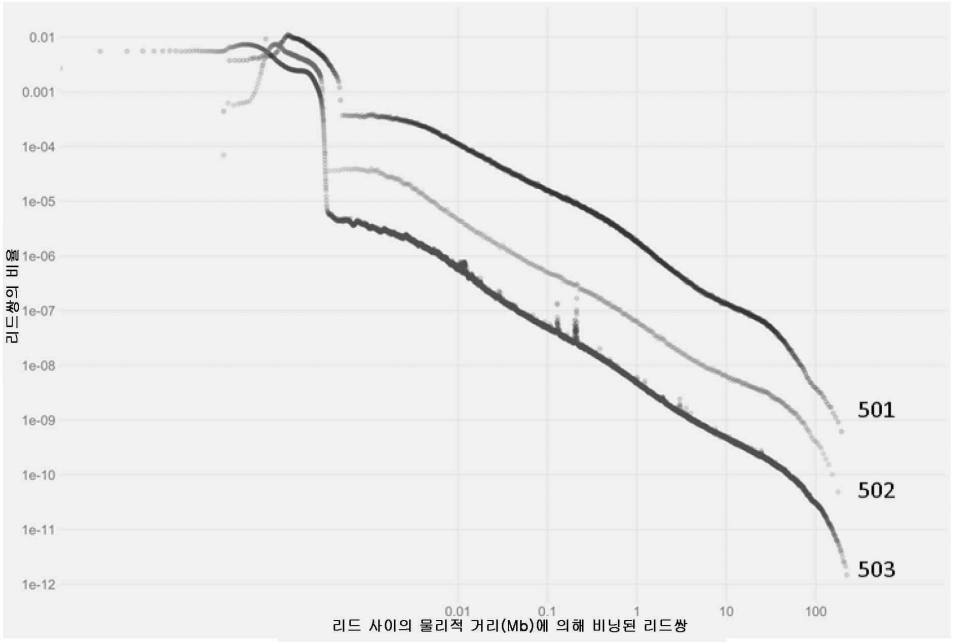
도면9b


페이징된 이형접합성 결실
chr6-78957411-79042838.out
40 Kb
대각선 위의 색은 최소 맵 품질을
나타낸다. (우측 눈금)
대각선 아래의 색은 하플로타입 페이즈를 나타낸다.
(노란색은 미지의 페이즈를 나타낸다.)
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

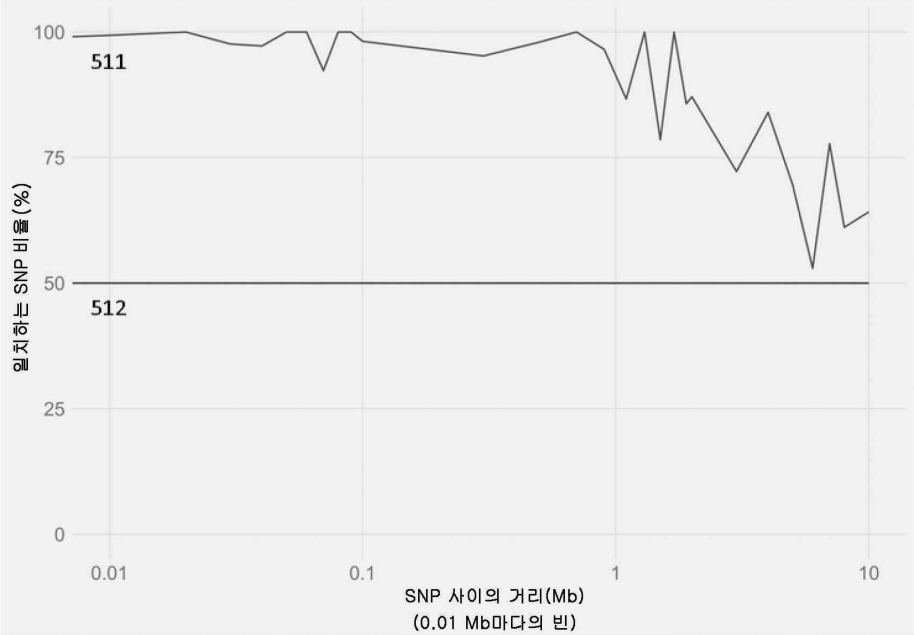
도면10


430
401
435
440
415
420
405
410
425

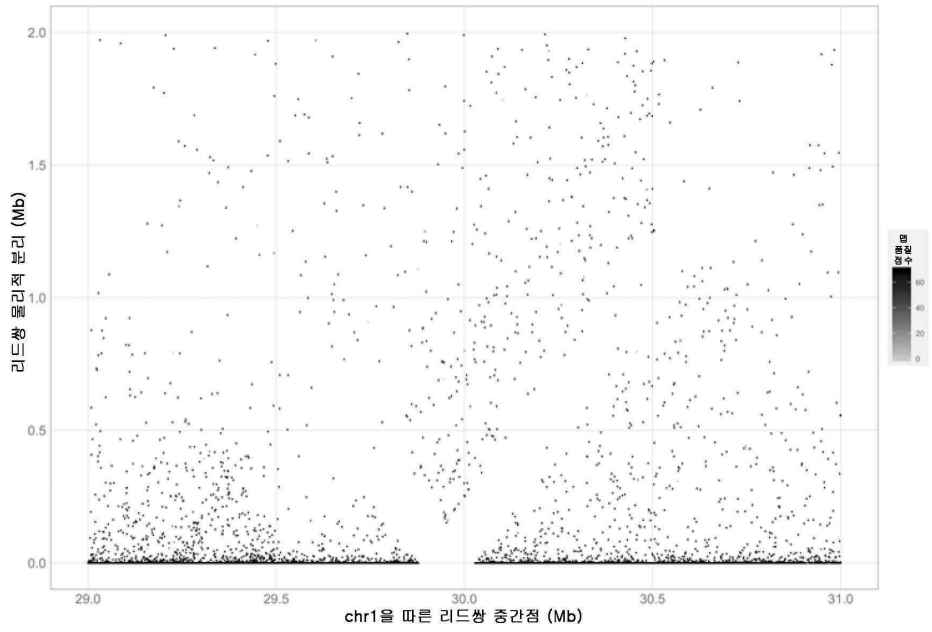
도면11a


0.01
0.001
1e-04
1e-05
1e-06
1e-07
1e-08
1e-09
1e-10
1e-11
1e-12
리드쌍의 비율
0.01
0.1
1
10
100
리드 사이의 물리적 거리(Mb)에 의해 비닝된 리드쌍
501
502
503

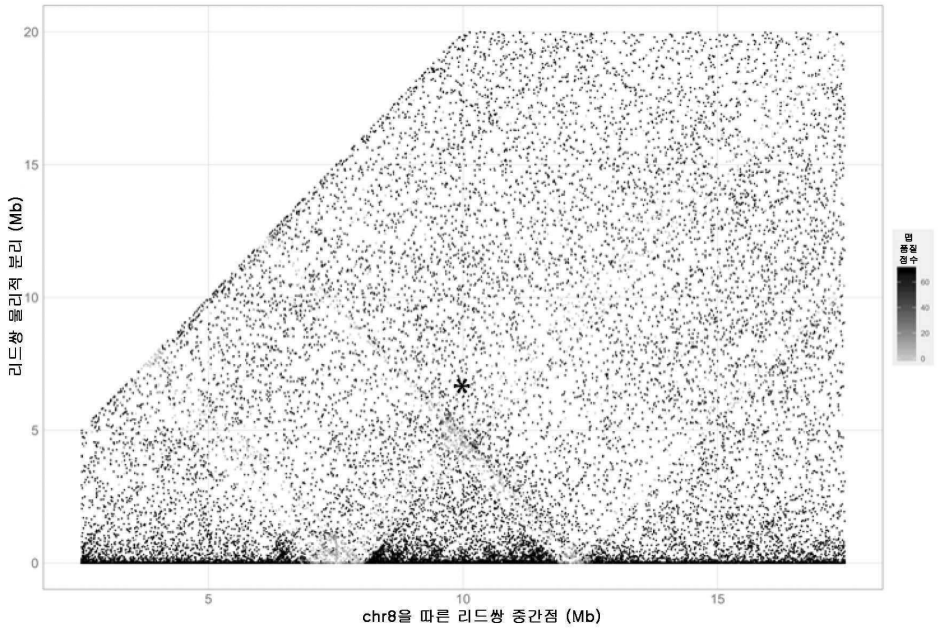
도면11b


100
75
50
25
0
511
512
일치하는 SNP 비율 (%)
0.01
0.1
1
10
SNP 사이의 거리(Mb)
(0.01 Mb마다의 빈)

도면11c


2.0
1.5
1.0
0.5
0.0
리드쌍 물리적 분리 (Mb)
29.0
29.5
30.0
30.5
31.0
chr1을 따른 리드쌍 중간점 (Mb)

도면11d


20
15
10
5
0
리드쌍 물리적 분리 (Mb)
5
10
15
chr8을 따른 리드쌍 중간점 (Mb)
맵 품질 점수
60
40
20
0