



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2017-0026383
(43) 공개일자 2017년03월08일

(51) 국제특허분류(Int. Cl.)
C12Q 1/68 (2006.01) G06F 19/22 (2017.01)
(52) CPC특허분류
C12Q 1/6827 (2013.01)
C12Q 1/6883 (2013.01)
(21) 출원번호 10-2016-7036255
(22) 출원일자(국제) 2015년06월26일
심사청구일자 없음
(85) 번역문제출일자 2016년12월26일
(86) 국제출원번호 PCT/US2015/038141
(87) 국제공개번호 WO 2015/200869
국제공개일자 2015년12월30일
(30) 우선권주장
62/017,808 2014년06월26일 미국(US)
62/072,214 2014년10월29일 미국(US)

(71) 출원인
10엑스 제노믹스, 인크.
미국, 캘리포니아 94566, 플레젠티, 스위트 401,
7068 콜 센터 파크웨이
(72) 발명자
슈날-레빈 마이클
미국 94306 캘리포니아주 팔로 알토 리키스 웨이
유닛 티 4238
차로스 미르나
미국 94041 캘리포니아주 마운틴 뷰 팔로 알토 에
비뉴 272
(뒷면에 계속)
(74) 대리인
김진희, 김태홍

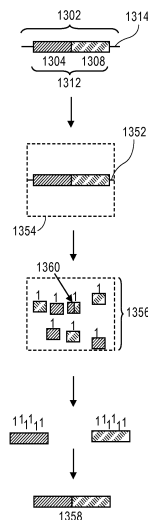
전체 청구항 수 : 총 163 항

(54) 발명의 명칭 핵산 서열의 분석

(57) 요약

본 발명은 일배체형 위상 및 카피 수 변형 검정을 위한 방법, 조성물 및 시스템에 관한 것이다. 비드를 포함하는 바코드를 복수의 별개 구획에서 샘플과 배합하기 위한 방법 및 시스템 뿐만 아니라 바코드화 샘플을 처리하고, 서열 분석하고 분석하는 방법이 본 발명에 포함된다

대표도 - 도13b



(52) CPC특허분류

G06F 19/22 (2013.01)
C12Q 2535/122 (2013.01)
C12Q 2537/16 (2013.01)
C12Q 2563/159 (2013.01)
C12Q 2565/629 (2013.01)
C12Q 2600/156 (2013.01)

(72) 발명자

힌슨 크리스토퍼

미국 94566 캘리포니아주 플레젠튼 버널 에비뉴
4868 아파트먼트 비

네스 케빈

미국 94566 캘리포니아주 플레젠튼 캠퍼니아 플레
이스 4306

삭소노프 서지

미국 94619 캘리포니아주 오클랜드 발모랄 드라이
브 5799

힌슨 벤자민

미국 94566 캘리포니아주 플레젠튼 테라 코트 1789

정 그레이스 엑스 와이

미국 94041 캘리포니아주 마운틴 뷰 부시 스트리트
245 아파트먼트 18

막스 패트릭

미국 94110 캘리포니아주 샌 프란시스코 21번 스트
리트 3025 아파트먼트 5

스텔프나겔 존

미국 94566 캘리포니아주 플레젠튼 스위트 401 콜
센터 파크웨이 7068

명세서

청구범위

청구항 1

다음을 포함하는, 핵산 중의 하나 이상의 변형을 동정하는 방법:

- (a) 상기 핵산 중의 제1 단편을 제공하는 단계로서, 상기 제1 단편이 10 킬로베이스(kb)를 초과하는 길이를 갖는, 단계;
- (b) 상기 제1 단편의 복수의 제2 단편을 서열 분석하여 복수의 단편 서열을 제공하는 단계로서, 이 복수의 단편 서열이 공통 바코드 서열을 공유하는, 단계;
- (c) 상기 복수의 단편 서열을 상기 공통 바코드 서열의 존재로 상기 제1 단편에 귀속시키는 단계;
- (d) 상기 복수의 단편 서열을 사용하여 상기 제1 단편의 핵산 서열을 결정하는 단계로서, 상기 핵산서열이 1% 미만의 오차율에서 결정되는 단계; 및
- (e) 단계 (d)에서 결정된 상기 제1 단계의 상기 핵산 서열 중의 상기 하나 이상의 변형을 동정하여 핵산 내에서 상기 하나 이상의 변형을 동정하는 단계.

청구항 2

청구항 1에 있어서, 상기 제1 단편이 복수의 별개 구획 중의 별개의 구획에 존재하는, 방법.

청구항 3

청구항 2에 있어서, 상기 별개의 구획이 에멀전 중의 액적인, 방법.

청구항 4

청구항 1에 있어서, 상기 동정 단계가 상기 제1 단편의 상기 핵산 서열로부터 상기 핵산 중의 위상 변이체를 동정하는 단계를 포함하는, 방법.

청구항 5

청구항 1에 있어서, 상기 동정 단계가 상기 제1 단편의 상기 핵산 서열로부터 상기 핵산 중의 하나 이상의 구조적 변형을 동정하는 단계를 포함하는, 방법.

청구항 6

청구항 1에 있어서, 상기 제1 단편이 15 kb를 초과하는 길이를 갖는, 방법.

청구항 7

청구항 1에 있어서, 상기 제1 단편이 20 kb를 초과하는 길이를 갖는, 방법.

청구항 8

청구항 1에 있어서, 상기 결정 단계가 상기 복수의 단편 서열을 참조 서열에 맵핑하는 단계를 포함하는, 방법.

청구항 9

청구항 1에 있어서, 상기 결정 단계가 상기 복수의 단편 서열을 상기 공통 바코드 서열로 조립하는 단계를 포함하는, 방법.

청구항 10

청구항 1에 있어서, 길이가 적어도 10 kb인 상기 핵산의 복수의 제1 단편을 제공하는 단계를 추가로 포함하고, 상기 동정하는 단계가 상기 복수의 제1 단편 각각으로부터 핵산 서열을 결정하는 단계 및 상기 복수의 제1 단편

각각의 상기 핵산 서열로부터 상기 핵산 중의 하나 이상의 변형을 동정하는 단계를 포함하는, 방법.

청구항 11

청구항 10에 있어서, 상기 둘 이상의 핵산 서열의 중복 핵산 서열을 기초로 하는 추정 콘택에 상기 복수의 제1 단편의 둘 이상의 핵산 서열을 연결하는 단계를 추가로 포함하고, 상기 최대 추정 콘택 길이가 적어도 10 kb인, 방법.

청구항 12

청구항 11에 있어서, 상기 최대 추정 콘택 길이가 적어도 20 kb인, 방법.

청구항 13

청구항 12에 있어서, 상기 최대 추정 콘택 길이가 적어도 40 kb인, 방법.

청구항 14

청구항 13에 있어서, 상기 최대 추정 콘택 길이가 적어도 50 kb인, 방법.

청구항 15

청구항 14에 있어서, 상기 최대 추정 콘택 길이가 적어도 100 kb인 방법.

청구항 16

청구항 15에 있어서, 상기 최대 추정 콘택 길이가 적어도 200 kb인, 방법.

청구항 17

청구항 16에 있어서, 상기 최대 추정 콘택 길이가 적어도 500 kb인, 방법.

청구항 18

청구항 17에 있어서, 상기 최대 추정 콘택 길이가 적어도 750 kb인, 방법.

청구항 19

청구항 18에 있어서, 상기 최대 추정 콘택 길이가 적어도 1 메가베이스(Mb)인, 방법.

청구항 20

청구항 19에 있어서, 상기 최대 추정 콘택 길이가 적어도 1.75 Mb인, 방법.

청구항 21

청구항 20에 있어서, 상기 최대 추정 콘택 길이가 적어도 2.5 Mb인, 방법.

청구항 22

청구항 10에 있어서, 상기 복수의 제1 단편의 상기 둘 이상의 핵산 서열 내의 중복 위상 변이체를 기초로 하는 위상 블록에서 상기 복수의 제1 단편의 둘 이상의 핵산 서열을 연결하는 단계를 추가로 포함하고, 상기 최대 위상 블록 길이가 적어도 10 kb인, 방법.

청구항 23

청구항 22에 있어서, 상기 최대 위상 블록 길이가 적어도 20 kb인, 방법.

청구항 24

청구항 23에 있어서, 상기 최대 위상 블록 길이가 적어도 40kb인, 방법.

청구항 25

청구항 24에 있어서, 상기 최대 위상 블록 길이가 적어도 50kb인, 방법.

청구항 26

청구항 25에 있어서, 상기 최대 위상 블록 길이가 적어도 100kb인, 방법.

청구항 27

청구항 26에 있어서, 상기 최대 위상 블록 길이가 적어도 200 kb인, 방법.

청구항 28

청구항 27에 있어서, 상기 최대 위상 블록 길이가 적어도 500 kb인, 방법.

청구항 29

청구항 28에 있어서, 상기 최대 위상 블록 길이가 적어도 750 kb인, 방법.

청구항 30

청구항 29에 있어서, 상기 최대 위상 블록 길이가 적어도 1 Mb인, 방법.

청구항 31

청구항 30에 있어서, 상기 최대 위상 블록 길이가 적어도 1.75 Mb인, 방법.

청구항 32

청구항 31에 있어서, 상기 최대 위상 블록 길이가 적어도 2.5 Mb인, 방법.

청구항 33

청구항 10에 있어서, 상기 둘 이상의 핵산 서열의 중복 핵산 서열에 기초하는 추정 콘택에서 상기 복수의 제1 단편의 둘 이상의 핵산 서열을 연결하여 추정 콘택의 집단을 생성하는 단계를 추가로 포함하고, 상기 추정 콘택 집단의 N50이 적어도 10 kb인, 방법.

청구항 34

청구항 33에 있어서, 상기 추정 콘택 집단의 N50이 적어도 20 kb인, 방법.

청구항 35

청구항 34에 있어서, 상기 추정 콘택 집단의 N50이 적어도 40 kb인, 방법.

청구항 36

청구항 35에 있어서, 상기 추정 콘택 집단의 N50이 적어도 50 kb인, 방법.

청구항 37

청구항 36에 있어서, 상기 추정 콘택 집단의 N50이 적어도 100 kb인, 방법.

청구항 38

청구항 37에 있어서, 상기 추정 콘택 집단의 N50이 적어도 200 kb인, 방법.

청구항 39

청구항 38에 있어서, 상기 추정 콘택 집단의 N50이 적어도 500 kb인, 방법.

청구항 40

청구항 39에 있어서, 상기 추정 콘택 집단의 N50이 적어도 750 kb인, 방법.

청구항 41

청구항 40에 있어서, 상기 추정 콘택 집단의 N50이 적어도 1 Mb인, 방법.

청구항 42

청구항 41에 있어서, 상기 추정 콘택 집단의 N50이 적어도 1.75 Mb인, 방법.

청구항 43

청구항 42에 있어서, 상기 추정 콘택 집단의 N50이 적어도 2.5 Mb인, 방법.

청구항 44

청구항 10에 있어서, 상기 복수의 제1 단편의 상기 둘 이상의 핵산 서열 내의 중복 위상 변이체에 기초하는 위상 블록에서 상기 복수의 제1 단편의 둘 이상의 핵산 서열을 연결하여 위상 블록의 집단을 생성하는 단계를 추가 포함하고, 상기 위상 블록 집단의 N50은 적어도 10 kb인, 방법.

청구항 45

청구항 44에 있어서, 상기 위상 블록 집단의 N50이 적어도 20 kb인, 방법.

청구항 46

청구항 45에 있어서, 상기 위상 블록 집단의 N50이 적어도 40kb인, 방법.

청구항 47

청구항 46에 있어서, 상기 위상 블록 집단의 N50이 적어도 50kb인, 방법.

청구항 48

청구항 47에 있어서, 상기 위상 블록 집단의 N50이 적어도 100kb인, 방법.

청구항 49

청구항 48에 있어서, 상기 위상 블록 집단의 N50이 적어도 200 kb인, 방법.

청구항 50

청구항 49에 있어서, 상기 위상 블록 집단의 N50이 적어도 500 kb인, 방법.

청구항 51

청구항 50에 있어서, 상기 위상 블록 집단의 N50이 적어도 750 kb인, 방법.

청구항 52

청구항 51에 있어서, 상기 위상 블록 집단의 N50이 적어도 1 Mb인, 방법.

청구항 53

청구항 52에 있어서, 상기 위상 블록 집단의 N50이 적어도 1.75 Mb인, 방법.

청구항 54

청구항 53에 있어서, 상기 위상 블록 집단의 N50이 적어도 2.5 Mb인, 방법.

청구항 55

다음을 포함하는, 핵산의 구조적 변형의 존재를 결정하는 방법:

(a) 상기 핵산의 복수의 제1 단편 분자를 제공하는 단계로서, 상기 복수의 제1 단편 분자의 소정의 제1 단편 분자가 상기 구조적 변형을 포함하는, 단계;

- (b) 상기 복수의 제1 단편 분자의 복수의 제2 단편 분자를 서열 분석하여 복수의 단편 서열을 제공하는 단계로서, 소정의 제1 단편 분자에 상응하는 상기 복수의 단편 서열 각각이 공통 바코드 서열을 공유하는, 단계; 및
- (c) (i) 상기 복수의 단편 서열을 참조 서열에 맵핑하는 단계, (ii) 상기 공통 바코드 서열을 공유하는 상기 복수의 단편 서열을 동정하는 단계, 및 (iii) 상기 소정의 제1 단편 분자의 길이보다 더 떨어져 있는 위치에서 상기 참조 서열에 맵핑하는 상기 공통 바코드 서열을 공유하는 상기 복수의 단편 서열의 상승량의 존재에 기초하는 구조적 변형을 동정하는 단계로서, 이 상승량이 상기 구조적 변형이 결여된 서열에 대한 것인, 단계에 의해 상기 구조적 변형의 존재를 결정하는 단계.

청구항 56

청구항 55에 있어서, 상기 상승량이 상기 구조적 변형을 갖는 상기 핵산의 영역으로부터 유래되는 상기 제1 단편 분자의 총 수에 대하여 1% 이상인, 방법.

청구항 57

청구항 56에 있어서, 상기 상승량이 상기 구조적 변형을 갖는 상기 핵산의 영역으로부터 유래되는 상기 제1 단편 분자의 총 수에 대하여 2% 이상인, 방법.

청구항 58

청구항 55에 있어서, 상기 위치가 적어도 약 100 베이스 떨어져 있는, 방법.

청구항 59

청구항 58에 있어서, 상기 위치가 적어도 약 500 베이스 떨어져 있는, 방법.

청구항 60

청구항 59에 있어서, 상기 위치가 적어도 약 1 킬로베이스(kb) 떨어져 있는, 방법.

청구항 61

청구항 60에 있어서, 상기 위치가 적어도 약 10 kb 떨어져 있는, 방법.

청구항 62

청구항 55에 있어서, 상기 복수의 단편 서열로부터 상기 소정의 단편 분자의 조립체를 생성함으로써 구조적 변형을 동정하는 단계를 추가로 포함하고, 상기 복수의 단편 서열이 상기 공통 바코드 서열에 기초하는 상기 조립체용 입력으로서 선택되는, 방법.

청구항 63

청구항 62에 있어서, 상기 조립체가 상기 복수의 단편 서열로부터 공통 서열을 생성함으로써 생성되는, 방법.

청구항 64

청구항 55에 있어서, 상기 구조적 변형이 전좌를 포함하는, 방법.

청구항 65

다음을 포함하는, 변이체 핵산 서열의 특성화 방법:

- (a) 변이체 핵산을 단편화하여 10 킬로베이스(kb)를 초과하는 길이를 갖는 복수의 제1 단편을 제공하는 단계;
- (b) 상기 복수의 제1 단편을 별개의 구획으로 분리하는 단계;
- (c) 복수의 제2 단편을 이의 각각의 구획 내의 각각의 제1 단편으로부터 생성하는 단계로서, 상기 복수의 제2 단편이 이에 부착된 바코드 서열을 갖고, 소정의 구획 내의 이 바코드 서열이 공통 바코드 서열인, 단계;
- (d) 상기 복수의 제2 단편 및 이에 부착된 상기 바코드 서열을 서열 분석하여 복수의 제2 단편 서열을 제공하는 단계;

(e) 상기 제2 단편 서열을 상기 공통 바코드 서열의 존재에 적어도 부분적으로 기초하는 원래 제1 단편으로 귀속시켜 상기 제2 단편 서열을 위한 제1 단편 서열 맥락을 제공하는 단계; 및

(f) 상기 제1 단편 서열 맥락으로부터 상기 변이체 핵산의 변이체 부분을 동정하여 상기 변이체 핵산서열을 특성화하는 단계.

청구항 66

청구항 65에 있어서, 상기 귀속화 단계가 적어도 부분적으로 상기 공통 바코드 서열의 존재에 기초하는 상기 복수의 제2 단편 서열로부터 상기 복수의 제1 단편으로부터 개별 단편을 위한 서열의 적어도 일부를 조립하는 단계를 포함하는, 방법.

청구항 67

청구항 65에 있어서, 상기 귀속화 단계가 상기 복수의 제2 단편 서열을 적어도 부분적으로 상기 공통 바코드 서열에 기초하는 상기 복수의 제1 단편으로부터 개별 제1 단편에 맵핑하는 단계를 포함하는, 방법.

청구항 68

청구항 65에 있어서, 둘 이상의 상기 복수의 제1 단편을, 상기 둘 이상의 복수의 제1 단편 사이의 중복 서열에 기초하여, 추정 콘텍에 연결하는 단계를 추가로 포함하는, 방법.

청구항 69

청구항 65에 있어서, 상기 동정 단계가 상기 제1 단편 서열 맥락으로부터 하나 이상의 위상 변이체를 동정하는 단계를 포함하는, 방법.

청구항 70

청구항 65에 있어서, 둘 이상의 상기 복수의 제1 단편을, 상기 둘 이상의 복수의 제1 단편 사이의 중복 위상 변이체에 기초하여, 위상 블록으로 연결하는 단계를 추가로 포함하는, 방법.

청구항 71

청구항 70에 있어서, 상기 동정 단계가 상기 제1 단편 서열 맥락으로부터 하나 이상의 구조적 변형을 동정하는 단계를 포함하는, 방법.

청구항 72

청구항 71에 있어서, 상기 하나 이상의 구조적 변형이 독립적으로 삽입, 결실, 전좌, 레트로트랜스포존, 역전 및 중복으로부터 선택되는, 방법.

청구항 73

청구항 72에 있어서, 상기 구조적 변형이 삽입 또는 전좌를 포함하고, 상기 제1 단편 서열 맥락이 삽입 또는 전좌의 존재를 나타내는, 방법.

청구항 74

다음을 포함하는, 핵산 서열 중의 변이체를 동정하는 방법:

상기 핵산의 복수의 개별 단편 분자의 핵산 서열을 획득하는 단계로서, 상기 복수의 개별 단편 분자의 상기 핵산 서열이 각각 적어도 1 킬로베이스(kb)의 길이를 갖는, 단계;

하나 이상의 추정 콘텍에서 상기 복수의 개별 단편 분자 중의 하나 이상의 서열을 연결하는 단계; 및

상기 하나 이상의 추정 콘텍으로부터 하나 이상의 변이체를 동정하는 단계.

청구항 75

청구항 74에 있어서, 상기 획득 단계가 길이가 10 kb를 초과하는 복수의 단편 분자의 상기 핵산 서열을 획득하

는 단계를 포함하는, 방법.

청구항 76

청구항 75에 있어서, 상기 수득 단계가 다음을 포함하는, 방법:

상기 복수의 개별 단편 분자의 각 개별 단편 분자의 복수의 바코드화 단편을 제공하는 단계로서, 소정의 개별 단편 분자의 상기 바코드화 단편이 공통 바코드를 갖는, 단계;

상기 복수의 개별 단편 분자의 상기 복수의 바코드화 단편을 서열 분석하는 단계로서, 상기 서열분석 단계가 1% 미만의 서열 분석 오차율을 제공하는, 단계; 및

상기 복수의 바코드화 단편 및 그들의 관련 바코드의 서열로부터 상기 복수의 개별 단편 분자의 서열을 결정하는 단계.

청구항 77

청구항 76에 있어서, 상기 연결 단계가 둘 이상의 개별 단편 분자 사이의 하나 이상의 중복 서열을 동정하여 상기 둘 이상의 개별 단편 분자를 상기 하나 이상의 추정 콘텍에 연결하는 단계를 포함하는, 방법.

청구항 78

청구항 77에 있어서, 상기 연결 단계가 둘 이상의 개별 단편 분자 사이의 하나 이상의 공통 변이체를 동정하여 상기 둘 이상의 개별 단편 분자를 상기 하나 이상의 추정 콘텍에 연결하는 단계를 포함하는, 방법.

청구항 79

청구항 78에 있어서, 상기 하나 이상의 공통 변이체가 위상 변이체이고, 상기 하나 이상의 추정 콘텍이 적어도 100kb의 최대 위상 블록 길이를 포함하는, 방법.

청구항 80

청구항 74에 있어서, 상기 동정 단계에서 동정된 상기 하나 이상의 변이체가 구조적 변형을 포함하는, 방법.

청구항 81

청구항 80에 있어서, 상기 구조적 변형이 삽입, 결실, 전좌, 레트로트랜스포존, 역전 및 중복으로부터 선택되는, 방법.

청구항 82

다음을 포함하는, 핵산의 특성화 방법:

적어도 10 킬로베이스(kb)의 길이를 갖는 복수의 단편 분자의 핵산 서열을 수득하는 단계;

상기 복수의 단편 분자의 상기 핵산 서열에서 하나 이상의 위상 변이체 위치를 동정하는 단계;

적어도 제1 단편 분자의 상기 핵산 서열을 상기 제1 및 제2 단편 분자 내의 하나 이상의 공통 위상 변이체 위치의 존재에 기초하는 적어도 제2 단편 분자에 연결하여 적어도 10 kb의 최대 위상 블록 길이를 갖는 위상 블록을 제공하는 단계; 및

적어도 10 kb의 최대 위상 블록 길이를 갖는 상기 위상 블록으로부터 하나 이상의 위상 변이체를 동정하는 단계.

청구항 83

청구항 82에 있어서, 상기 위상 블록으로부터 하나 이상의 추가의 위상 변이체를 동정하는 단계를 추가로 포함하는, 방법.

청구항 84

청구항 82에 있어서, 상기 복수의 단편 분자가 별개의 구획에 존재하는, 방법.

청구항 85

청구항 84에 있어서, 상기 별개의 구획이 에멀전 중의 액적인, 방법.

청구항 86

청구항 82에 있어서, 상기 복수의 단편 분자의 길이가 적어도 50 kb인, 방법.

청구항 87

청구항 86에 있어서, 상기 복수의 단편 분자의 길이가 적어도 100 kb인, 방법.

청구항 88

청구항 82에 있어서, 상기 최대 위상 블록 길이가 적어도 50 kb인, 방법.

청구항 89

청구항 88에 있어서, 상기 최대 위상 블록 길이가 적어도 100 kb인, 방법.

청구항 90

청구항 89에 있어서, 상기 최대 위상 블록 길이가 적어도 1 Mb인, 방법.

청구항 91

청구항 90에 있어서, 상기 최대 위상 블록 길이가 적어도 2 Mb인, 방법.

청구항 92

청구항 91에 있어서, 상기 최대 위상 블록 길이가 적어도 2.5 Mb인, 방법.

청구항 93

다음을 포함하는, 방법:

- a) 제1 핵산을 제1 구획으로 구획화하는 단계로서, 상기 제1 핵산이 유기체의 제1 염색체로부터 유래되는 표적 서열을 포함하는, 단계;
- b) 제2 핵산을 제2 구획으로 구획화하는 단계로서, 상기 제2 핵산이 상기 유기체의 제2 염색체로부터 유래되는 표적 서열을 포함하는, 단계;
- c) 상기 제1 구획에서, 제1 바코드 서열을 상기 제1 핵산의 단편에 또는 상기 제1 핵산의 일부의 카피에 부착하여 제1 바코드화 단편을 제공하는 단계;
- d) 상기 제2 구획에서, 제2 바코드 서열을 상기 제2 핵산의 단편에 또는 상기 제2 핵산의 일부의 카피에 부착하여 제2 바코드화 단편을 제공하는 단계로서, 상기 제2 바코드 서열이 상기 제1 바코드 서열과 상이한, 단계;
- e) 상기 제1 및 제2 바코드화 단편의 상기 핵산 서열을 결정하고, 상기 제1 및 제2 핵산의 핵산서열을 조립하는 단계; 및
- f) 상기 제1 및 제2 핵산의 상기 핵산 서열을 비교하여 각각 제1 및 제2 염색체로부터 유래되는 상기 제1 및 제2 핵산을 특성화하는 단계.

청구항 94

청구항 93에 있어서, 상기 제1 바코드 서열을 포함하는 올리고뉴클레오티드가 상기 제1 핵산과 공동-구획화되고, 상기 제2 바코드화 서열을 포함하는 올리고뉴클레오티드가 상기 제2 핵산과 공동-구획화되는, 방법.

청구항 95

청구항 94에 있어서, 상기 제1 바코드 서열을 포함하는 상기 올리고뉴클레오티드가 제1 비드에 박리가능하게 부

착되고, 상기 제2 바코드 서열을 포함하는 상기 올리고뉴클레오티드가 제2 비드에 박리가능하게 부착되고, 상기 공동-구획화 단계가 상기 제1 및 제2 비드를 각각 상기 제1 및 제2 구획으로 공동-구획화하는 단계를 포함하는, 방법.

청구항 96

청구항 93에 있어서, 상기 제1 및 제2 구획이 에멀전 중의 액적을 포함하는, 방법.

청구항 97

청구항 93에 있어서, 상기 제1 염색체가 부계 염색체이고, 상기 제2 염색체가 모계 염색체인, 방법.

청구항 98

청구항 93에 있어서, 상기 제1 염색체 및 상기 제2 염색체가 상동 염색체인, 방법.

청구항 99

청구항 93에 있어서, 상기 제1 핵산 및 상기 제2 핵산이 하나 이상의 변형을 포함하는, 방법.

청구항 100

다음을 포함하는, 방법:

- a) 제1 핵산을 제1 구획으로 구획화하는 단계로서, 상기 제1 핵산이 유기체의 제1 염색체로부터 유래되는 표적 서열을 포함하는, 단계;
- b) 제2 핵산을 제2 구획으로 구획화하는 단계로서, 상기 제2 핵산이 상기 유기체의 제2 염색체로부터 유래되는 표적 서열을 포함하는, 단계;
- c) 상기 제1 구획에서, 제1 바코드 서열을 상기 제1 핵산의 단편에 또는 상기 제1 핵산의 일부의 카피에 부착하여 제1 바코드화 단편을 제공하는 단계;
- d) 상기 제2 구획에서, 제2 바코드 서열을 상기 제2 핵산의 단편에 또는 상기 제2 핵산의 일부의 카피에 부착하여 제2 바코드화 단편을 제공하는 단계로서, 상기 제2 바코드 서열이 상기 제1 바코드 서열과 상이한, 단계;
- e) 상기 제1 및 제2 바코드화 단편의 상기 핵산 서열을 결정하고, 상기 제1 및 제2 핵산의 핵산서열을 조립하는 단계; 및
- f) 상기 제1 및 제2 핵산의 상기 핵산 서열을 비교하여 상기 제1 및 제2 핵산의 상기 핵산 서열사이의 임의의 변형을 동정하는 단계.

청구항 101

청구항 100에 있어서, 상기 제1 바코드 서열을 포함하는 올리고뉴클레오티드가 상기 제1 핵산과공동-구획화되고, 상기 제2 바코드 서열을 포함하는 올리고뉴클레오티드가 상기 제2 핵산과 공동-구획화되는, 방법.

청구항 102

청구항 101에 있어서, 상기 제1 바코드 서열을 포함하는 상기 올리고뉴클레오티드가 제1 비드에박리가능하게 부착되고, 상기 제2 바코드 서열을 포함하는 상기 올리고뉴클레오티드가 제2 비드에 박리가능하게 부착되고, 상기 공동-구획화 단계가 상기 제1 및 제2 비드를 각각 상기 제1 및 제2 구획으로 공동-구획화하는 단계를 포함하는, 방법.

청구항 103

청구항 100에 있어서, 상기 제1 및 제2 구획이 에멀전 중의 액적을 포함하는, 방법.

청구항 104

청구항 100에 있어서, 상기 제1 염색체가 부계 염색체이고, 상기 제2 염색체가 모계 염색체인, 방법.

청구항 105

청구항 100에 있어서, 상기 제1 염색체 및 상기 제2 염색체가 상동 염색체인, 방법.

청구항 106

청구항 100에 있어서, 상기 제1 핵산 및 상기 제2 핵산이 하나 이상의 변형을 포함하는, 방법.

청구항 107

청구항 93 또는 청구항 100에 있어서, 상기 제1 및 제2 염색체가 태아로부터 유래되는, 방법.

청구항 108

청구항 93 또는 청구항 100에 있어서, 상기 제1 및 제2 핵산이 임신부로부터 채취된 샘플로부터 획득되는, 방법.

청구항 109

청구항 93 또는 청구항 100에 있어서, 상기 제1 염색체가 염색체 21, 18 또는 13인, 방법.

청구항 110

청구항 93 또는 청구항 100에 있어서, 상기 제2 염색체가 염색체 21, 18 또는 13인, 방법.

청구항 111

청구항 93 또는 청구항 100에 있어서, 상기 제1 또는 제2 염색체의 상대량을 결정하는 단계를 추가로 포함하는, 방법.

청구항 112

청구항 93 또는 청구항 100에 있어서, 참조 염색체에 대한 상기 제1 또는 제2 염색체의 양을 결정하는 단계를 추가로 포함하는, 방법.

청구항 113

청구항 93 또는 청구항 100에 있어서, 상기 제1 염색체 또는 제2 염색체 또는 둘 다가 카피 수의 증가를 갖는, 방법.

청구항 114

청구항 113에 있어서, 상기 카피 수의 증가가 암 또는 이수성의 결과인, 방법.

청구항 115

청구항 93 또는 청구항 100에 있어서, 상기 제1 염색체 또는 제2 염색체 또는 둘 다가 카피 수의 감소를 갖는, 방법.

청구항 116

청구항 115에 있어서, 상기 카피 수의 감소가 암 또는 이수성의 결과인, 방법.

청구항 117

다음을 포함하는, 태아 핵산 서열의 특성화 방법:

(a) 모계 핵산 서열을 결정하는 단계로서, 상기 모계 핵산이 다음에 의해 태아의 임신한 어머니로부터 유래되는, 단계:

(i) 모계 핵산을 단편화하여 복수의 제1 모계 단편을 제공하는 단계;

(ii) 상기 복수의 제1 모계 단편을 모계 구획으로 분리하는 단계;

- (iii) 복수의 제2 모계 단편을 그들 각각의 모계 구획 내의 상기 제1 모계 단편 각각으로부터 생성하는 단계로서, 상기 복수의 제2 모계 단편이 이에 부착된 제1 바코드 서열을 갖고, 상기 모계 구획의 소정의 모계 구획 내에서 상기 제2 모계 단편이 이에 부착된 제1 공통 바코드 서열을 포함하는, 단계;
- (iv) 상기 복수의 제2 모계 단편을 서열 분석하여 복수의 모계 단편 서열을 제공하는 단계;
- (v) 상기 모계 단편 서열을 적어도 부분적으로 상기 제1 공통 바코드 서열의 존재에 기초하는 원래 제1 모계 단편으로 귀속시켜 상기 모계 핵산 서열을 결정하는 단계;
- (b) 부계 핵산 서열을 결정하는 단계로서, 상기 부계 핵산이 다음에 의해 태아의 아버지로부터 유래되는, 단계:
 - (i) 부계 핵산을 단편화하여 복수의 제1 부계 단편을 제공하는 단계;
 - (ii) 상기 복수의 제1 부계 단편을 부계 별개 구획으로 분리하는 단계;
 - (iii) 복수의 제2 부계 단편을 이의 각각의 구획 내의 각각의 제1 부계 단편으로부터 생성하는 단계로서, 상기 복수의 제2 부계 단편이 이에 부착된 제2 바코드 서열을 갖고, 소정의 부계 구획 내에서, 상기 제2 부계 단편이 이에 부착된 제2 공통 바코드 서열을 포함하는, 단계;
 - (iv) 상기 복수의 제2 부계 단편 및 이에 부착된 상기 제2 바코드 서열을 서열 분석하여 복수의 부계 단편 서열을 제공하는 단계;
 - (v) 상기 부계 단편 서열을 적어도 부분적으로 상기 제2 공통 바코드 서열의 존재에 기초하는 원래 제1 부계 단편으로 귀속시켜 상기 부계 핵산 서열을 결정하는 단계;
- (c) 상기 임신한 어머니로부터 태아 핵산을 수득하고, 상기 모계 핵산 서열 및 상기 부계 핵산 서열을 사용하여 상기 태아 핵산의 서열 및/또는 상기 태아 핵산의 서열의 하나 이상의 유전적 변형을 결정하는 단계.

청구항 118

청구항 117에 있어서, 상기 부계 단편 서열 및 상기 모계 단편 서열이 각각 서열을 하나 이상의 추정 콘텍에 연결하는데 사용되는, 방법.

청구항 119

청구항 118에 있어서, 상기 추정 콘텍이 모계 및 부계 위상 블록을 작제하는데 사용되는, 방법.

청구항 120

청구항 119에 있어서, 상기 태아 핵산의 서열이 태아 위상 블록을 작제하기 위한 상기 모계 및 부계 위상 블록과 비교되는, 방법.

청구항 121

청구항 120에 있어서, 상기 부계 단편 서열이 상기 복수의 제1 부계 단편을 위한 서열의 적어도 일부를 생성하기 위해 조립되고, 이에 의해 상기 부계 핵산 서열을 결정하고, 상기 모계 단편 서열이 상기 복수의 제1 모계 단편을 위한 서열의 적어도 일부를 생성하기 위해 조립되고, 이에 의해 상기 모계 핵산 서열을 결정하는, 방법.

청구항 122

청구항 117에 있어서, 상기 부계 핵산 서열을 결정하는 단계가 상기 부계 단편 서열을 부계 참조서열에 맵핑시키는 단계를 포함하고, 상기 모계 핵산 서열을 결정하는 단계가 상기 모계 단편 서열을 모계 참조 서열에 맵핑하는 단계를 포함하는, 방법.

청구항 123

청구항 117에 있어서, 상기 태아 핵산의 서열이 적어도 99%의 정확성으로 결정되는, 방법.

청구항 124

청구항 117에 있어서, 상기 태아 핵산의 서열의 상기 하나 이상의 유전적 변형이 적어도 99%의 정확성으로 결정되는, 방법.

청구항 125

청구항 117에 있어서, 상기 하나 이상의 유전적 변형이 구조적 변형 및 단일 뉴클레오타이드 다형성(SNP)으로 이루어진 그룹으로부터 선택되는, 방법.

청구항 126

청구항 117에 있어서, 상기 하나 이상의 유전적 변형이 카피 수 변형, 삽입, 결실, 전좌, 레트로트랜스포존, 역전, 재배열, 반복 확대 및 중복으로 이루어진 그룹으로부터 선택된 구조적 변형인, 방법.

청구항 127

청구항 117에 있어서, 단계 (c)에서, 상기 태아 핵산의 서열의 상기 하나 이상의 유전적 변형을 상기 모계 핵산 서열 및 상기 부계 핵산 서열에 대해 결정된 하나 이상의 유전적 변형을 사용하여 결정하는 단계를 추가로 포함하는, 방법.

청구항 128

청구항 117에 있어서, 단계 (c)에서, 상기 태아 핵산의 하나 이상의 신생 돌연변이를 결정하는 단계를 추가로 포함하는, 방법.

청구항 129

청구항 117에 있어서, 단계 (c) 동안 또는 이후에 상기 태아 핵산과 관련되는 이수성을 결정하는 단계를 추가로 포함하는, 방법.

청구항 130

청구항 117에 있어서, 단계 (a)의 단계 (v) 동안 또는 이후에 상기 모계 핵산 서열을 일배체형화하여 일배체형-분해 모계 핵산 서열을 제공하는 단계 및 단계 (b)의 단계 (v) 동안 또는 이후에 상기 부계 핵산 서열을 일배체형화하여 일배체형-분해 부계 핵산 서열을 제공하는 단계를 추가로 포함하는, 방법.

청구항 131

청구항 130에 있어서, 단계 (c)에서, 상기 일배체형-분해 모계 핵산 서열 및 상기 일배체형-분해 부계 핵산 서열을 사용하여 상기 태아 핵산의 서열 및/또는 상기 하나 이상의 유전적 변형을 결정하는 단계를 추가로 포함하는, 방법.

청구항 132

청구항 117에 있어서, 하나 이상의 상기 모계 핵산 및 상기 부계 핵산이 게놈 데옥시리보핵산(DNA)인, 방법.

청구항 133

청구항 117에 있어서, 단계 (c)에서, 상기 태아 핵산이 무세포 핵산을 포함하는, 방법.

청구항 134

청구항 117에 있어서, 단계 (a)에서, 상기 모계 핵산 서열을 적어도 99%의 정확성으로 결정하는 단계를 추가로 포함하는, 방법.

청구항 135

청구항 117에 있어서, 단계 (b)에서, 상기 부계 핵산 서열을 적어도 99%의 정확성으로 결정하는 단계를 추가로 포함하는, 방법.

청구항 136

청구항 117에 있어서, 상기 모계 핵산 서열 및/또는 상기 부계 핵산 서열이 10 킬로베이스(kb)를 초과하는 길이를 갖는, 방법.

청구항 137

청구항 117에 있어서, 상기 모계 및 부계 구획이 에멀전 중의 액적을 포함하는, 방법.

청구항 138

청구항 117에 있어서, 단계 (a)에서, 상기 제1 바코드 서열이 제1 입자에 박리가능하게 부착된 상기 소정의 모계 구획에 제공되는, 방법.

청구항 139

청구항 138에 있어서, 단계 (b)에서, 상기 제2 바코드 서열이 제2 입자에 박리가능하게 부착된 상기 소정의 부계 구획에 제공되는, 방법.

청구항 140

다음을 포함하는, 샘플 핵산의 특성화 방법:

- (a) 생물학적 샘플을 대상체로부터 수득하는 단계로서, 이 생물학적 샘플이 무세포 샘플 핵산을 포함하는, 단계;
- (b) 액적에서, 바코드 서열을 상기 무세포 샘플 핵산의 단편에 또는 상기 샘플 핵산의 일부의 카피에 부착하여 바코드화 샘플 단편을 제공하는 단계;
- (c) 상기 바코드화 샘플 단편의 핵산 서열을 결정하고 상기 바코드화 샘플 단편의 상기 핵산 서열에 기초하는 샘플 핵산 서열을 제공하는 단계;
- (d) 프로그램된 컴퓨터 프로세서를 사용하여 샘플 핵산 서열과 참조 핵산 서열의 비교를 생성하는 단계로서, 이 참조 핵산 서열이 10 킬로베이스(kb)보다 긴 길이 및 적어도 99%의 정확성을 갖는, 단계; 및
- (e) 상기 샘플 핵산 서열 중의 하나 이상의 유전적 변형을 동정하기 위해 상기 비교를 사용하여 상기 핵산을 질 환과 관련시키는 단계.

청구항 141

청구항 140에 있어서, 상기 샘플 핵산 서열 중의 상기 하나 이상의 유전적 변형이 구조적 변형 및 단일 뉴클레오티드 다형성(SNP)으로 이루어진 그룹으로부터 선택되는, 방법.

청구항 142

청구항 140에 있어서, 상기 샘플 핵산 서열의 상기 하나 이상의 유전적 변형이 카피 수 변형, 삽입, 결실, 레트로트랜스포존, 전좌, 역전, 재배열, 반복 확대 및 중복으로 이루어진 그룹으로부터 선택되는 구조적 변형인, 방법.

청구항 143

청구항 140에 있어서, 단계 (c)에서, 상기 샘플 핵산 서열이 적어도 99%의 정확성으로 제공되는, 방법.

청구항 144

청구항 140에 있어서, 단계 (b)에서, 상기 바코드 서열이 입자에 박리가능하게 부착된 상기 액적에 제공되고, 단계 (b)가 추가로 상기 바코드 서열을 부착시키기 전에 상기 입자로부터 상기 바코드 서열을 상기 액적으로 박리시키는 단계를 추가로 포함하는, 방법.

청구항 145

청구항 144에 있어서, 단계 (b)에서, 상기 바코드 서열이 상기 입자에 박리가능하게 부착된 프라이머 서열의 일부로서 제공되고, 상기 프라이머 서열이 또한 랜덤 N-량체 서열을 포함하고, 단계 (b)가 추가로 상기 바코드 서열을 부착시키기 전에 상기 입자로부터 상기 프라이머 서열을 상기 액적으로 방출시키는 단계를 추가로 포함하는, 방법.

청구항 146

청구항 145에 있어서, 단계 (b)에서, 상기 바코드 서열을 상기 무세포 샘플 핵산의 상기 단편에 또는 상기 무세포 샘플 핵산의 일부의 상기 카피에 상기 프라이머를 사용하는 증폭 반응으로 부착시키는 단계를 추가로 포함하는, 방법.

청구항 147

청구항 140에 있어서, (i) 추가의 액적에서, 추가의 바코드 서열을 참조 핵산의 단편에 또는 상기 참조 핵산의 일부의 카피에 부착시켜 바코드화 참조 단편을 제공하는 단계; 및 (ii) 상기 바코드화 참조 단편의 핵산 서열을 결정하고, 상기 바코드화 참조 단편의 상기 핵산 서열에 기초하는 상기 참조 핵산 서열을 결정하는 단계를 추가로 포함하는, 방법.

청구항 148

청구항 147에 있어서, 상기 참조 핵산 서열의 결정 단계가 상기 바코드화 참조 단편의 상기 핵산서열을 조립하는 단계를 포함하는, 방법.

청구항 149

청구항 147에 있어서, 입자에 박리가능하게 부착된 상기 추가의 액적에 상기 추가의 바코드 서열을 제공하고, 상기 입자로부터 상기 추가의 바코드 서열을 상기 추가의 바코드 서열을 부착시키기 전에 상기 추가의 구획으로 박리시키는 단계를 추가로 포함하는, 방법.

청구항 150

청구항 149에 있어서, 상기 추가의 바코드 서열을 상기 입자에 박리가능하게 부착된 프라이머 서열의 일부로서 제공하는 단계로서, 상기 프라이머 서열이 또한 랜덤 N-량체 서열을 포함하는, 단계 및 상기 입자로부터 상기 프라이머를 상기 추가의 바코드 서열을 부착시키기 전에 상기 추가의 액적으로 박리시키는 단계를 추가로 포함하는, 방법.

청구항 151

청구항 150에 있어서, 상기 추가의 바코드 서열을 상기 참조 핵산의 단편에 또는 상기 참조 핵산의 일부의 상기 카피에 상기 프라이머를 사용하는 증폭 반응으로 부착시키는 단계를 추가로 포함하는, 방법.

청구항 152

청구항 147에 있어서, 상기 참조 핵산 서열에서 하나 이상의 유전적 변형을 결정하는 단계를 추가로 포함하는, 방법.

청구항 153

청구항 152에 있어서, 상기 참조 핵산 서열 중의 상기 하나 이상의 유전적 변형이 구조적 변형 및 단일 뉴클레오타이드 다형성(SNP)으로 이루어진 그룹으로부터 선택되는, 방법.

청구항 154

청구항 152에 있어서, 상기 참조 핵산 서열 중의 상기 하나 이상의 유전적 변형이 카피 수 변형, 삽입, 결실, 레트로트랜스포존, 전좌, 역전, 재배열, 반복 확대 및 중복으로 이루어진 그룹으로부터 선택되는 구조적 변형인, 방법.

청구항 155

청구항 147에 있어서, 상기 참조 핵산이 생식 세포 핵산 서열을 포함하는, 방법.

청구항 156

청구항 147에 있어서, 상기 참조 핵산이 암 핵산 서열을 포함하는, 방법.

청구항 157

청구항 140에 있어서, 상기 샘플 핵산 서열이 10 kb를 초과하는 길이를 갖는, 방법.

청구항 158

청구항 147에 있어서, 상기 참조 핵산이 질환 상태의 부재를 나타내는 게놈으로부터 유래되는, 방법.

청구항 159

청구항 147에 있어서, 상기 참조 핵산이 질환 상태를 나타내는 게놈으로부터 유래되는, 방법.

청구항 160

청구항 158 또는 159 중의 어느 한 항에 있어서, 상기 질환 상태가 암을 포함하는, 방법.

청구항 161

청구항 158 또는 159 중의 어느 한 항에 있어서, 상기 질환 상태가 이수성을 포함하는, 방법.

청구항 162

청구항 140에 있어서, 상기 무세포 샘플 핵산이 종양 핵산을 포함하는, 방법.

청구항 163

청구항 162에 있어서, 상기 종양 핵산이 순환성 종양 핵산을 포함하는, 방법.

발명의 설명

기술 분야

[0001] 본원은 하기에 대한 우선권을 주장한다: 미국 가출원 제 62/017,808호(2014년 6월 26일에 출원), 및 미국 가출원 제 62/072,214호(2014년 10월 29일에 출원), 이들 각각은 모든 목적을 위해 그 전체가 본원에 참조로 인용된다.

배경 기술

[0002] 특정 인간 게놈의 기본적인 이해는 단순히 돌연변이와 같은 특정 유전자 변형의 유무를 동정하는 것 이상을 필요로 할 수 있다. 특정 유전자 변형이 (또한 위상으로서 공지된) 동일하거나 상이한 염색체 상에서 나타나는지의 여부를 결정하는 것도 또한 중요하다. 유전자의 카피 수에 대한 정보처럼, 일배체형과 같은 유전자 변형 패턴에 대한 정보도 또한 중요하다.

[0003] 용어 "일배체형"은 연속 블록에서 함께 상속된 DNA 서열 변이체의 세트(대립 유전자)를 의미한다. 일반적으로, 인간 게놈은 각 유전자의 두 카피 - 모계 카피 및 부계 카피를 함유한다. 각각 두 개의 가능한 대립 유전자, 예를 들면, 유전자 대립 유전자 "A" 및 "a", 및 유전자 대립 유전자 "B" 및 "b"를 갖는 유전자 쌍의 경우, 소정의 개체의 게놈은 두 일배체형 "AB/ab"(여기서, A 및 B 대립 유전자는 동일한 염색체 상에 거주한다("시스" 구성)), 또는 "Ab/aB"(여기서, A 및 B 대립 유전자는 상이한 염색체 상에 거주한다("트랜스" 구성)) 중의 하나를 포함한다. 위상 방법 또는 검정은 대립 유전자의 특정 세트가 동일하거나 상이한 염색체 상에 거주하는지의 여부를 결정하기 위해 사용될 수 있다. 일부 경우에, 일배체형을 규정하는 다수의 연결된 대립 유전자가 특정 질환 표현형과 상관될 수 있거나 관련될 수 있고; 그러한 경우에, 임의의 하나의 특정 유전자 변형보다 일배체형이 환자가 질환을 나타내는지의 여부에 가장 결정적인 요인일 수 있다.

[0004] 유전자 카피 수는 또한 일부 질환 표현형에서 역할을 할 수 있다. 대부분의 유전자는 일반적으로 두 카피로 존재하지만, 증폭된 유전자는 두 개 이상의 기능적 카피로 존재하는 유전자이다. 일부 경우에, 유전자는 또한 기능적 카피의 손실을 겪을 수 있다. 유전자 카피 수의 손실 또는 증가는 잠재적으로 암 상태 또는 다른 장애로 이어지는 mRNA 및 단백질 발현의 비정상적 수준의 생산으로 이어질 수 있다. 암 및 다른 유전 장애는 종종 비정상적(증가되거나 감소된) 염색체 수와 상관된다("이수성"). 동일 반응계 혼성화 또는 비교용 게놈 혼성화에서 형광과 같은 세포 유전학적 기술은 비정상적 유전자 또는 염색체 카피 수의 존재를 검출하기 위해 사용될 수 있

다. 유전 위상 정보, 일배체형 또는 카피 수 변형을 검출하는 개선된 방법이 당해 기술 분야에서 필요하다.

발명의 내용

요약

본 발명은 유전 물질의 특성화에 상당한 진보를 제공하는데 유용할 수 있는 방법 및 시스템을 제공한다. 이러한 방법 및 시스템은, 예를 들면, 일배체형 위상, 구조적 변형, 예를 들면, 결실, 중복, 카피-수 변이체, 삽입, 역전, 전좌, 긴 탠덤 반복체 (LTR), 짧은 탠덤 반복체 (STR)의 동정을 포함하는 일반적으로 이용가능한 기술을 사용하는 것이 실질적으로 어려울 수 있는 유전적 특성화 및 다양한 다른 유용한 특성화를 제공하는데 유용할 수 있다.

본 발명의 한 국면은 다음 단계를 포함하여, 핵산에서 하나 이상의 변형을 동정하는 방법을 제공한다: a) 핵산의 제1 단편을 제공하는 단계로서, 상기 제1 단편이 10 킬로베이스(kb)보다 큰 길이를 갖는, 단계; (b) 복수의 단편 서열을 제공하기 위해 상기 제1 단편의 복수의 제2 단편을 서열 분석하는 단계로서, 이 복수의 단편 서열이 공통 바코드 서열을 공유하는, 단계; (c) 복수의 단편 서열을 공통 바코드 서열의 존재에 의해 제1 단편으로 귀속시키는 단계; (d) 복수의 단편 서열을 사용하여 제1 단편의 핵산 서열을 결정하는 단계로서, 상기 핵산 서열이 1% 미만의 오차율로 결정되는, 단계; 및 (e) 단계 (d)에서 결정된 제1 단편의 핵산 서열에서 하나 이상의 변동을 동정함으로써 핵산 내의 하나 이상의 서열을 동정하는 단계.

일부 경우에, 제1 단편은 복수의 별개 구획 중의 별개 구획으로 존재한다. 일부 경우에, 별개 구획은 에멀전 중의 액적이다. 일부 경우에, 동정하는 단계는 제1 단편의 핵산 서열의 위상화 변이체를 동정하는 단계를 포함한다. 일부 경우에, 동정하는 단계는 제1 단편의 핵산 서열로부터의 핵산에서 하나 이상의 구조적 변형을 동정하는 단계를 포함한다. 일부 경우에, 제1 단편은 15 kb보다 큰 길이를 갖는다. 일부 경우에, 제1 단편은 20 kb보다 큰 길이를 갖는다. 일부 경우에, 결정하는 단계는 참조에 복수의 단편 서열을 맵핑하는 단계를 포함한다. 일부 경우에, 결정하는 단계는 공통 바코드 서열을 갖는 복수의 단편 서열을 조립하는 단계를 포함한다.

일부 경우에, 하나 이상의 변동을 동정하는 방법은 추가로 길이가 적어도 10 kb인 핵산의 복수의 제1 단편을 제공하는 단계를 포함하고, 동정하는 단계는 복수의 제1 단편 각각으로부터 핵산 서열을 결정하는 단계 및 복수의 제1 단편 각각으로부터의 핵산 서열의 핵산에서 하나 이상의 변동을 동정하는 단계를 포함한다.

일부 경우에, 하나 이상의 변동을 동정하는 방법은 복수의 제1 단편의 둘 이상의 핵산 서열을 둘 이상의 핵산 서열의 중첩 핵산 서열에 기초하는 추정된 콘텍(contig)으로 연결하는 단계를 추가로 포함하고, 여기서 최대 추정 콘텍 길이는 적어도 10 kb이다. 일부 경우에, 최대 추정 콘텍 길이는 적어도 20 kb이다. 일부 경우에, 최대 추정 콘텍 길이는 적어도 40 kb이다. 일부 경우에, 최대 추정 콘텍 길이는 적어도 50 kb이다. 일부 경우에, 최대 추정 콘텍 길이는 적어도 100 kb이다. 일부 경우에, 최대 추정 콘텍 길이는 적어도 200 kb이다. 일부 경우에, 최대 추정 콘텍 길이는 적어도 500 kb이다. 일부 경우에, 최대 추정 콘텍 길이는 적어도 750 kb이다. 일부 경우에, 최대 추정 콘텍 길이는 적어도 1 메가베이스(Mb)이다. 일부 경우에, 최대 추정 콘텍 길이는 적어도 1.75 Mb이다. 일부 경우에, 최대 추정 콘텍 길이는 적어도 2.5 Mb이다.

일부 경우에, 하나 이상의 변동을 동정하는 방법은 복수의 제1 단편의 둘 이상의 핵산 서열을 복수의 제1 단편의 둘 이상의 핵산 서열 내의 중첩 위상 변이체에 기초하는 위상 블록으로 연결하는 단계를 추가로 포함하고, 여기서 최대 위상 블록 길이는 적어도 10 kb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 20 kb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 40 kb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 50 kb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 100 kb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 200 kb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 500 kb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 750 kb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 1 Mb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 1.75 Mb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 2.5 Mb이다.

일부 경우에, 하나 이상의 변동을 동정하는 방법은 복수의 제1 단편의 둘 이상의 핵산 서열을 둘 이상의 핵산 서열의 중복 핵산 서열에 기초하는 추정 콘텍으로 연결하여 추정 콘텍의 집단을 생성하는 단계를 추가로 포함하고, 여기서 추정 콘텍 집단의 N50은 적어도 10 kb이다. 일부 경우에, 추정 콘텍 집단의 N50은 적어도 20 kb이다. 일부 경우에, 추정 콘텍 집단의 N50은 적어도 40 kb이다. 일부 경우에, 추정 콘텍 집단의 N50은 적어도 50 kb이다. 일부 경우에, 추정 콘텍 집단의 N50은 적어도 100 kb이다. 일부 경우에, 추정 콘텍 집단의 N50은 적어도 200 kb이다. 일부 경우에, 추정 콘텍 집단의 N50은 적어도 500 kb이다. 일부 경우에, 추정 콘텍 집단의 N50은 적어도 750 kb이다. 일부 경우에, 추정 콘텍 집단의 N50은 적어도 1 Mb이다. 일부 경우에, 추정 콘텍 집

단의 N50은 적어도 1.75 Mb이다. 일부 경우에, 추정 콘택 집단의 N50은 적어도 2.5 Mb이다.

[0013] 일부 경우에, 하나 이상의 변동을 동정하는 방법은 복수의 제1 단편의 둘 이상의 핵산 서열을 복수의 제1 단편의 둘 이상의 핵산 서열 내의 중복 위상 변이체에 기초하는 위상 블록으로 연결하여 위상 블록의 집단을 생성하는 단계를 추가로 포함하고, 여기서 위상 블록 집단의 N50은 적어도 10 kb이다. 일부 경우에, 위상 블록 집단의 N50은 적어도 20 kb이다. 일부 경우에, 위상 블록 집단의 N50은 적어도 40kb이다. 일부 경우에, 위상 블록 집단의 N50은 적어도 50kb이다. 일부 경우에, 위상 블록 집단의 N50은 적어도 100kb이다. 일부 경우에, 위상 블록 집단의 N50은 적어도 200 kb이다. 일부 경우에, 위상 블록 집단의 N50은 적어도 500 kb이다. 일부 경우에, 위상 블록 집단의 N50은 적어도 750 kb이다. 일부 경우에, 위상 블록 집단의 N50은 적어도 1 Mb이다. 일부 경우에, 위상 블록 집단의 N50은 적어도 1.75 Mb이다. 일부 경우에, 위상 블록 집단의 N50은 적어도 2.5 Mb이다.

[0014] 본 발명의 추가의 국면은 핵산의 구조적 변형의 존재를 결정하는 방법을 제공한다. 상기 방법은 다음을 포함할 수 있다: (a) 핵산의 복수의 제1 단편 분자를 제공하는 단계로서, 복수의 제1 단편 분자의 소정의 제1 단편 분자가 구조적 변형을 포함하는, 단계; (b) 복수의 단편 서열을 제공하기 위해 복수의 제1 단편 분자 각각의 복수의 제2 단편 분자를 서열 분석하는 단계로서, 소정의 제1 단편 분자에 상응하는 복수의 단편 서열의 각각이 공통 바코드 서열을 공유하는 단계; 및 (c) (i) 복수의 단편 서열을 참조 서열에 맵핑하는 단계, (ii) 공통 바코드 서열을 공유하는 복수의 단편 서열을 동정하는 단계, (iii) 소정의 제1 단편 분자의 길이보다 더 떨어져 있는 위치에서 참조 서열에 맵핑되는 공통 바코드 서열을 공유하는 복수의 단편 서열의 상승된 양의 존재에 기초하는 구조적 변형을 동정하는 단계로서, 이 상승된 양이 구조적 변형이 결여된 서열에 대한 것인, 단계에 의해 구조적 변형의 존재를 결정하는 단계.

[0015] 일부 경우에, 상승된 양은 구조적 변형을 갖는 핵산의 영역으로부터 유래된 제1 단편 분자의 총수에 대하여 1% 이상이다. 일부 경우에, 상승된 양은 구조적 변형을 갖는 핵산의 영역으로부터 유래된 제1 단편 분자의 총수에 대하여 2% 이상이다. 일부 경우에, 위치는 적어도 약 100 베이스 떨어져 있다. 일부 경우에, 위치는 적어도 약 500 베이스 떨어져 있다. 일부 경우에, 위치는 적어도 약 1 킬로베이스(kb) 떨어져 있다. 일부 경우에, 위치는 적어도 약 10 kb 떨어져 있다.

[0016] 일부 경우에, 핵산의 구조적 변형의 존재를 결정하는 방법은 복수의 단편 서열로부터 소정의 제1 단편 분자의 조립체를 생성함으로써 구조적 변형을 동정하는 단계를 추가로 포함하고, 여기서 복수의 단편 서열은 공통 바코드 서열의 존재에 기초하는 조립체를 위한 입력으로서 선택된다. 일부 경우에, 조립체는 복수의 단편 서열로부터 공통 서열을 생성함으로써 생성된다. 일부 경우에, 구조적 변형은 전좌를 포함한다.

[0017] 본 발명의 추가의 국면은 변형된 핵산 서열을 특성화하는 방법을 제공한다. 일부 경우에, 상기 방법은 다음을 포함할 수 있다: (a) 10 킬로베이스(kb)보다 큰 길이를 갖는 복수의 제1 단편을 제공하기 위해 변형된 핵산을 단편화하는 단계; (b) 복수의 제1 단편을 별개의 구획으로 분리하는 단계; (c) 이의 각각의 구획 내의 각각의 제1 단편으로부터 복수의 제2 단편을 생성하는 단계로서, 복수의 제2 단편이 이에 부착된 바코드 서열을 갖고, 소정의 구획 내의 이 바코드 서열이 공통 바코드 서열인, 단계; (d) 복수의 제2 단편 서열을 제공하기 위해 복수의 제2 단편 및 이에 부착된 바코드 서열을 서열 분석하는 단계; (e) 제2 단편 서열을, 제2 단편 서열을 위한 제1 단편 서열 맥락을 제공하기 위해 공통 바코드 서열의 존재에 적어도 부분적으로 기초하는 원래 제1 단편으로 귀속시키는 단계; 및 (f) 제1 단편 서열 맥락으로부터 변형된 핵산의 변형 부분을 동정하여 변동된 핵산 서열을 특성화하는 단계. 일부 경우에, 귀속시키는 단계는 적어도 부분적으로 공통 바코드 서열의 존재에 기초하는 복수의 제2 단편 서열로부터의 복수의 제1 단편으로부터의 개별 단편을 위한 서열의 적어도 일부를 조립하는 단계를 포함한다. 일부 경우에, 귀속시키는 단계는 복수의 제2 단편 서열을 적어도 부분적으로 공통 바코드 서열에 기초하는 복수의 제1 단편으로부터의 개별적 제1 단편으로 맵핑하는 단계를 포함한다.

[0018] 일부 경우에, 변형된 핵산 서열을 특성화하는 방법은 복수의 제1 단편 중의 둘 이상을 복수의 제1 단편 중 둘 이상 사이의 중복 서열에 기초하여, 추정 콘택으로 연결하는 단계를 추가로 포함한다. 일부 경우에, 동정하는 단계는 제1 단편 서열 맥락으로부터 하나 이상의 위상 변이체를 동정하는 단계를 포함한다. 일부 경우에, 변형된 핵산 서열을 특성화하는 방법은 복수의 제1 단편 중 둘 이상을 복수의 제1 단편 중 둘 이상 사이의 중복 위상 변이체에 기초하여, 위상 블록으로 연결하는 단계를 추가로 포함한다. 일부 경우에, 동정하는 단계는 제1 단편 서열 맥락으로부터 하나 이상의 구조적 변형을 동정하는 단계를 포함한다. 일부 경우에, 하나 이상의 구조적 변형은 독립적으로 삽입, 결실, 전좌, 레트로트랜스포존, 역전 및 중복으로부터 선택된다. 일부 경우에, 구조적 변형은 삽입 또는 전좌를 포함하고, 제1 단편 서열 맥락은 삽입 또는 전좌의 존재를 나타낸다.

[0019] 본 발명의 추가의 국면은 핵산의 서열에서 변이체를 동정하는 방법을 제공한다. 일부 경우에, 상기 방법은 다음

을 포함한다: 핵산의 복수의 개별적 단편 분자의 핵산 서열을 수득하는 단계로서, 복수의 개별적 단편 분자의 핵산 서열은 각각 적어도 1 킬로베이스(kb)의 길이를 갖는, 단계; 복수의 개별적 단편 분자 중 하나 이상의 서열을 하나 이상의 추정 콘텍으로 연결하는 단계; 및 하나 이상의 추정 콘텍으로부터 하나 이상의 변이체를 동정하는 단계. 일부 경우에, 수득하는 단계는 길이가 10 kb를 초과하는 복수의 단편 분자의 핵산 서열을 수득하는 단계를 포함한다. 일부 경우에, 수득하는 단계는 다음을 포함한다: 복수의 개별적 단편 분자의 각각의 개별적 단편 분자의 복수의 바코드화 단편을 제공하는 단계로서, 소정의 개별적 단편 분자의 바코드화 단편이 공통 바코드를 갖는, 단계; 복수의 개별적 단편 분자의 복수의 바코드화 단편을 서열 분석하는 단계로서, 상기 서열 분석은 1% 미만의 서열분석 오차율을 제공하는, 단계; 및 복수의 바코드화 단편 및 이들의 관련 바코드의 서열로부터 복수의 개별적 단편 분자의 서열을 결정하는 단계.

[0020] 일부 경우에, 연결하는 단계는 둘 이상의 개별적 단편 분자를 하나 이상의 추정 콘텍으로 연결시키기 위해 둘 이상의 개별적 단편 분자 사이의 하나 이상의 중복 서열을 동정하는 단계를 포함한다. 일부 경우에, 연결하는 단계는 둘 이상의 개별적 단편 분자를 하나 이상의 추정 콘텍으로 연결하기 위해 둘 이상의 개별적 단편 분자 사이의 하나 이상의 공통 변이체를 동정하는 단계를 포함한다. 일부 경우에, 하나 이상의 공통 변이체는 위상 변이체이고, 하나 이상의 추정 콘텍은 적어도 100kb의 최대 위상 블록 길이를 포함한다. 일부 경우에, 동정 단계에서 동정된 하나 이상의 변이체는 구조적 변형을 포함한다. 일부 경우에, 구조적 변형은 삽입, 결실, 전좌, 레트로트랜스포존, 역전 및 중복으로부터 선택된다.

[0021] 본 발명의 추가의 국면은 핵산을 특성화하는 방법을 제공한다. 일부 경우에, 상기 방법은 다음을 포함한다: 적어도 10 킬로베이스(kb)의 길이를 갖는 복수의 단편 분자의 핵산 서열을 수득하는 단계; 복수의 단편 분자의 핵산 서열 중의 하나 이상의 위상 변이체 위치를 동정하는 단계; 적어도 제1 단편 분자의 핵산 서열을 적어도 10kb의 최대 위상 블록 길이를 갖는 위상 블록을 제공하기 위해 제1 및 제2 단편 분자 내의 하나 이상의 공통 위상 변이체 위치의 존재에 기초하는 적어도 제2 단편 분자에 연결하는 단계; 및 적어도 10 kb의 최대 위상 블록 길이를 갖는 위상 블록으로부터 하나 이상의 위상 변이체를 동정하는 단계. 일부 경우에, 핵산을 특성화하는 방법은 위상 블록으로부터 하나 이상의 추가의 위상 변이체를 동정하는 단계를 포함한다. 일부 경우에, 복수의 단편 분자는 별개의 구획으로 존재한다. 일부 경우에, 별개의 구획은 에멀전 중의 액적이다. 일부 경우에, 복수의 단편 분자의 길이는 적어도 50 kb이다. 일부 경우에, 복수의 단편 분자의 길이는 적어도 100 kb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 50 kb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 100 kb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 1 Mb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 2 Mb이다. 일부 경우에, 최대 위상 블록 길이는 적어도 2.5 Mb이다.

[0022] 발명의 추가의 국면은 다음을 포함하는 방법을 제공한다: (a) 제1 핵산을 제1 구획으로 구획화하는 단계로서, 상기 제1 핵산이 유기체의 제1 염색체로부터 유래된 표적 서열을 포함하는, 단계; (b) 제2 핵산을 제2 구획으로 구획화하는 단계로서, 상기 제2 핵산이 유기체의 제2 염색체로부터 유래된 표적 서열을 포함하는, 단계; (c) 제1 구획에서, 제1 바코드 서열을 제1 바코드화 단편을 제공하기 위해 제1 핵산의 단편에 또는 제1 핵산의 부분의 카피에 부착하는 단계; (d) 제2 구획에서, 제2 바코드 서열을 제2 바코드화 단편을 제공하기 위해 제2 핵산의 단편에 또는 제2 핵산의 부분의 카피에 부착하는 단계로서, 상기 제2 바코드 서열이 상기 제1 바코드 서열과 상이한, 단계; (e) 제1 및 제2 바코드화 단편의 핵산 서열을 결정하고, 제1 및 제2 핵산의 핵산 서열을 조립하는 단계; 및 (f) 각각 제1 및 제2 염색체로부터 유래되는 제1 및 제2 핵산을 특성화하기 위해 제1 및 제2 핵산의 핵산 서열을 비교하는 단계. 일부 경우에, 제1 바코드 서열을 포함하는 올리고뉴클레오티드는 제1 핵산과 공동-구획화되고, 제2 바코드 서열을 포함하는 올리고뉴클레오티드는 제2 핵산과 공동-구획화된다. 일부 경우에, 제1 바코드 서열을 포함하는 올리고뉴클레오티드는 박리가능하게 제1 비드에 부착되고, 제2 바코드 서열을 포함하는 올리고뉴클레오티드는 박리가능하게 제2 비드에 부착되고, 공동-구획화 단계는 제1 및 제2 비드를 각각 제1 및 제2 구획으로 공동-구획화하는 단계를 포함한다. 일부 경우에, 제1 및 제2 구획은 에멀전 중의 액적을 포함한다. 일부 경우에, 제1 염색체는 부계 염색체이고, 제2 염색체는 모계 염색체이다. 일부 경우에, 제1 염색체 및 제2 염색체는 상동 염색체이다. 일부 경우에, 제1 핵산 및 제2 핵산은 하나 이상의 변형을 포함한다.

[0023] 일부 경우에, 제1 및 제2 염색체는 태아로부터 유도된다. 일부 경우에, 제1 및 제2 핵산은 임신부로부터 취한 샘플로부터 수득된다. 일부 경우에, 제1 염색체는 염색체 21, 18 또는 13이다. 일부 경우에, 제2 염색체는 염색체 21, 18 또는 13이다. 일부 경우에, 상기 방법은 제1 또는 제2 염색체의 상대적 양을 결정하는 단계를 추가로 포함한다. 일부 경우에, 상기 방법은 참조 염색체에 대한 제1 또는 제2 염색체의 양을 결정하는 단계를 추가로 포함한다. 일부 경우에, 제1 염색체 또는 제2 염색체, 또는 둘 다는 카피 수의 증가를 갖는다. 일부 경우에, 카피 수의 증가는 암 또는 이수성의 결과이다. 일부 경우에, 제1 염색체 또는 제2 염색체, 또는 둘 다는 카피 수

의 감소를 갖는다. 일부 경우에, 카피 수의 감소는 암 또는 이수성의 결과이다.

[0024]

발명의 추가의 국면은 다음을 포함하는 방법을 제공한다: (a) 제1 핵산을 제1 구획으로 구획화하는 단계로서, 상기 제1 핵산이 유기체의 제1 염색체로부터 유래된 표적 서열을 포함하는, 단계; (b) 제2 핵산을 제2 구획으로 구획화하는 단계로서, 상기 제2 핵산이 유기체의 제2 염색체로부터 유래된 표적 서열을 포함하는, 단계; (c) 제1 구획에서, 제1 바코드 서열을 제1 바코드화 단편을 제공하기 위해 제1 핵산의 단편에 또는 제1 핵산의 부분의 카피에 부착하는 단계; (d) 제2 구획에서, 제2 바코드 서열을 제2 바코드화 단편을 제공하기 위해 제2 핵산의 단편에 또는 제2 핵산의 부분의 카피에 부착하는 단계로서, 상기 제2 바코드 서열이 상기 제1 바코드 서열과 상이한, 단계; (e) 제1 및 제2 바코드화 단편의 핵산 서열을 결정하고, 제1 및 제2 핵산의 핵산 서열을 조립하는 단계; 및 (f) 제1 및 제2 핵산의 핵산 서열 사이의 임의의 변형을 동정하기 위해 제1 및 제2 핵산의 핵산 서열을 비교하는 단계. 일부 경우에, 제1 바코드 서열을 포함하는 올리고뉴클레오타이드는 제1 핵산과 공동-구획화되고, 제2 바코드 서열을 포함하는 올리고뉴클레오타이드는 제2 핵산과 공동-구획화된다. 일부 경우에, 제1 바코드 서열을 포함하는 올리고뉴클레오타이드는 박리가 가능하게 제1 비드에 부착되고, 제2 바코드 서열을 포함하는 올리고뉴클레오타이드는 박리가 가능하게 제2 비드에 부착되고, 공동-구획화 단계는 제1 및 제2 비드를 각각 제1 및 제2 구획으로 공동-구획화하는 단계를 포함한다. 일부 경우에, 제1 및 제2 구획은 에멀전 중의 액적을 포함한다. 일부 경우에, 제1 염색체는 부계 염색체이고, 제2 염색체는 모계 염색체이다. 일부 경우에, 제1 염색체 및 제2 염색체는 상동 염색체이다. 일부 경우에, 제1 핵산 및 제2 핵산은 하나 이상의 변형을 포함한다. 일부 경우에, 제1 및 제2 염색체는 태아로부터 유도된다. 일부 경우에, 제1 및 제2 핵산은 임신부로부터 취한 샘플로부터 수득된다. 일부 경우에, 제1 염색체는 염색체 21, 18 또는 13이다. 일부 경우에, 제2 염색체는 염색체 21, 18 또는 13이다. 일부 경우에, 상기 방법은 제1 또는 제2 염색체의 상대적 양을 결정하는 단계를 추가로 포함한다. 일부 경우에, 상기 방법은 참조 염색체에 대한 제1 또는 제2 염색체의 양을 결정하는 단계를 추가로 포함한다. 일부 경우에, 제1 염색체 또는 제2 염색체, 또는 둘 다는 카피 수의 증가를 갖는다. 일부 경우에, 카피 수의 증가는 암 또는 이수성의 결과이다. 일부 경우에, 제1 염색체 또는 제2 염색체, 또는 둘 다는 카피 수의 감소를 갖는다. 일부 경우에, 카피 수의 감소는 암 또는 이수성의 결과이다.

[0025]

본 발명의 추가의 국면은 태아 핵산 서열을 특성화하는 방법을 제공한다. 일부 경우에, 상기 방법은 다음을 포함한다: (a) 모계 핵산 서열을 결정하는 단계로서, 상기 모계 핵산이 다음에 의해 태아의 임신한 어머니로부터 유래되는 단계: (i) 복수의 제1 모계 단편을 제공하기 위해 모계 핵산을 단편화하는 단계; (ii) 복수의 제1 모계 단편을 모계 구획으로 분리하는 단계; (iii) 복수의 제2 모계 단편을 그들의 각각의 모계 구획 내의 제1 모계 단편 각각으로부터 생성하는 단계로서, 복수의 제2 모계 단편이 이에 부착된 제1 바코드 서열을 갖고, 이때 모계 구획들의 소정의 모계 구획 내에서 제2 모계 단편이 이에 부착된 제1 공통 바코드 서열을 포함하는, 단계; (iv) 복수의 모계 단편 서열을 제공하기 위해 복수의 제2 모계 단편을 서열 분석하는 단계; (v) 모계 단편 서열을 모계 핵산 서열을 결정하기 위해 제1 공통 바코드 서열의 존재에 적어도 부분적으로 기초하는 원래 제1 모계 단편으로 귀속시키는 단계; (b) 부계 핵산 서열을 결정하는 단계로서, 상기 부계 핵산이 다음에 의해 태아의 아버지로부터 유래되는, 단계: (i) 복수의 제1 부계 단편을 제공하기 위해 부계 핵산을 단편화하는 단계; (ii) 복수의 제1 부계 단편을 부계 별개의 구획으로 분리하는 단계; (iii) 복수의 제2 부계 단편을 그의 각각의 구획 내의 제1 부계 구획 각각으로부터 생성하는 단계로서, 복수의 제2 부계 단편이 이에 부착된 제2 바코드 서열을 갖고, 이때 소정의 부계 구획 내에서, 제2 부계 단편은 이에 부착된 제2 공통 바코드 서열을 포함하는, 단계; (iv) 복수의 부계 단편 서열을 제공하기 위해 복수의 제2 부계 단편 및 이에 부착된 제2 바코드 서열을 서열 분석하는 단계; (v) 부계 단편 서열을 부계 핵산 서열을 결정하기 위해 적어도 부분적으로 제2 공통 바코드 서열의 존재에 기초하는 원래 제1 부계 단편으로 귀속시키는 단계; (c) 태아 핵산을 임신한 어머니로부터 수득하고, 태아 핵산 서열 및/또는 태아 핵산 서열의 하나 이상의 유전적 변형을 모계 핵산 서열 및 부계 핵산 서열을 사용하여 결정하는 단계.

[0026]

일부 경우에, 부계 단편 서열 및 모계 단편 서열은 각각 서열을 하나 이상의 추정 콘텍에 연결하기 위해 사용된다. 일부 경우에, 추정 콘텍은 모계 및 부계 위상 블록을 작제하기 위해 사용된다. 일부 경우에, 태아 핵산의 서열은 태아 위상 블록을 작제하기 위해 모계 및 부계 위상 블록과 비교된다. 일부 경우에, 부계 단편 서열은 복수의 제1 부계 단편을 위한 서열의 적어도 일부를 생산하기 위해 조립되고, 이에 의해 부계 핵산 서열을 결정하고, 여기서 모계 단편 서열은 복수의 제1 모계 단편을 위한 서열의 적어도 일부를 생산하기 위해 조립되고, 이에 의해 모계 핵산 서열을 결정한다. 일부 경우에, 부계 핵산 서열을 결정하는 단계는 부계 단편 서열을 부계 참조에 맵핑하는 단계를 포함하고, 여기서 모계 핵산 서열을 결정하는 단계는 모계 단편 서열을 모계 참조에 맵핑하는 단계를 포함한다.

- [0027] 일부 경우에, 태아 핵산 서열은 적어도 99%의 정확성으로 결정된다. 일부 경우에, 태아 핵산 서열의 하나 이상의 유전적 변형은 적어도 99%의 정확성으로 결정된다. 일부 경우에, 하나 이상의 유전적 변형은 구조적 변형 및 단일 뉴클레오타이드 다형성(SNP)으로 이루어진 그룹으로부터 선택된다. 일부 경우에, 하나 이상의 유전적 변형은 카피 수 변형, 삽입, 결실, 전좌, 레트로트랜스포존, 역전, 재배열, 반복 확대 및 중복으로 이루어진 그룹으로부터 선택된 구조적 변형이다.
- [0028] 일부 경우에, 태아 핵산 서열을 특성화하는 방법은 단계 (c)에서 모계 핵산 서열 및 부계 핵산 서열을 위해 결정된 하나 이상의 유전적 변형을 사용하여 태아 핵산 서열의 하나 이상의 유전적 변형을 결정하는 단계를 추가로 포함한다. 일부 경우에, 태아 핵산 서열을 특성화하는 방법은 단계 (c)에서, 태아 핵산의 하나 이상의 신생 돌연변이를 결정하는 단계를 추가로 포함한다. 일부 경우에, 태아 핵산 서열을 특성화하는 방법은 단계 (c) 동안 또는 후에 태아 핵산과 관련된 이수성을 결정하는 단계를 추가로 포함한다.
- [0029] 일부 경우에, 태아 핵산 서열을 특성화하는 방법은 단계 (a)의 단계 (v) 동안 또는 이후에 일배체형-분해 모계 핵산 서열을 제공하기 위해 모계 핵산 서열을 일배체화하는 단계 및, 단계 (b)의 단계 (v) 동안 또는 이후에, 일배체형-분해 부계 핵산 서열을 제공하기 위해 부계 핵산 서열을 일배체화하는 단계를 추가로 포함한다. 일부 경우에, 태아 핵산 서열을 특성화하는 방법은 단계 (c)에서 일배체형-분해 모계 핵산 서열 및 일배체형-분해 부계 핵산 서열을 사용하여 태아 핵산의 서열 및/또는 하나 이상의 유전적 변형을 결정하는 단계를 추가로 포함한다. 일부 경우에, 하나 이상의 모계 핵산 및 부계 핵산은 게놈 데옥시리보핵산(DNA)이다. 일부 경우에, 단계 (c)에서, 태아 핵산은 무세포 핵산을 포함한다. 일부 경우에, 태아 핵산 서열을 특성화하는 방법은 단계 (a)에서 모계 핵산 서열을 적어도 99%의 정확성으로 결정하는 단계를 추가로 포함한다. 일부 경우에, 태아 핵산 서열을 특성화하는 방법은 단계 (b)에서 부계 핵산 서열을 적어도 99%의 정확성으로 결정하는 단계를 추가로 포함한다.
- [0030] 일부 경우에, 모계 핵산 서열 및/또는 부계 핵산 서열은 10 킬로베이스(kb)를 초과하는 길이를 갖는다. 일부 경우에, 모계 및 부계 구획은 에멀전 중의 액적을 포함한다. 일부 경우에, 단계 (a)에서, 제1 바코드 서열은 제1 입자에 박리가능하게 부착된 소정의 모계 구획에 제공된다. 일부 경우에, 단계 (b)에서, 제2 바코드 서열은 제2 입자에 박리가능하게 부착된 소정의 부계 구획에 제공된다.
- [0031] 본 발명의 추가의 국면은 샘플 핵산을 특성화하는 방법을 제공한다. 일부 경우에, 상기 방법은 다음을 포함한다: (a) 대상체로부터 생물학적 샘플을 수득하는 단계로서, 이 생물학적 샘플이 무세포 샘플 핵산을 포함하는, 단계; (b) 액적에서, 바코드 서열을 바코드화 샘플 단편을 제공하기 위해 무세포 샘플 핵산의 단편에 또는 샘플 핵산의 부분의 카피에 부착시키는 단계; (c) 바코드화 샘플 단편의 핵산 서열을 결정하고, 바코드화 샘플 단편의 핵산 서열에 기초하는 샘플 핵산 서열을 제공하는 단계; (d) 샘플 핵산 서열과 참조 핵산 서열의 비교를 생성하기 위해 프로그램된 컴퓨터 프로세서를 사용하는 단계로서, 이 참조 핵산 서열이 10 킬로베이스(kb)를 초과하는 길이 및 적어도 99%의 정확성을 갖는, 단계; 및 (e) 샘플 핵산 서열 중의 하나 이상의 유전적 변형을 동정하기 위해 비교를 사용하여 샘플 핵산을 질환과 관련시키는 단계. 일부 경우에, 샘플 핵산 서열 중의 하나 이상의 유전적 변형은 구조적 변형 및 단일 뉴클레오타이드 다형성(SNP)으로 이루어진 그룹으로부터 선택된다. 일부 경우에, 샘플 핵산 서열의 하나 이상의 유전적 변형은 카피 수 변형, 삽입, 결실, 레트로트랜스포존, 전좌, 역전, 재배열, 반복 확대 및 중복으로 이루어진 그룹으로부터 선택된다. 일부 경우에, 단계 (c)에서, 샘플 핵산 서열은 적어도 99%의 정확성으로 제공된다. 일부 경우에, 단계 (b)에서, 바코드 서열은 입자에 박리가능하게 부착된 액적에 제공되고, 단계 (b)는 바코드 서열을 부착시키기 전에 입자로부터 바코드 서열을 액적으로 박리시키는 단계를 추가로 포함한다. 일부 경우에, 단계 (b)에서, 바코드 서열은 입자에 박리가능하게 부착된 프라이머 서열의 일부로서 제공되고, 여기서 프라이머 서열은 또한 랜덤 N-량체 서열을 포함하고, 여기서 단계 (b)는 바코드 서열을 부착시키기 전에 입자로부터 프라이머 서열을 액적으로 박리시키는 단계를 추가로 포함한다. 일부 경우에, 단계 (b)에서, 바코드 서열은 프라이머를 사용하는 증폭 반응으로 무세포 샘플 핵산의 단편에 또는 무세포 샘플 핵산의 일부의 카피에 부착시킨다.
- [0032] 일부 경우에, 샘플 핵산을 특성화하는 방법은 다음을 추가로 포함한다: (i) 추가의 액적에서, 추가의 바코드 서열을 바코드화 참조 단편을 제공하기 위해 참조 핵산의 단편에 또는 참조 핵산의 일부의 카피에 부착시키는 단계; 및 (ii) 바코드화 참조 단편의 핵산 서열을 결정하고, 바코드화 참조 단편의 핵산 서열에 기초하는 참조 핵산 서열을 결정하는 단계. 일부 경우에, 참조 핵산 서열의 결정화 단계는 바코드화 참조 단편의 핵산 서열을 조립하는 단계를 포함한다. 일부 경우에, 샘플 핵산을 특성화하는 방법은 입자에 박리가능하게 부착된 추가의 액적에 추가의 바코드 서열을 제공하는 단계 및 추가의 바코드 서열을 부착시키기 전에 입자로부터 추가의 바코드 서열을 추가의 구획으로 박리시키는 단계를 추가로 포함한다. 일부 경우에, 샘플 핵산을 특성화하는 방법은 입

자에 박리가능하게 부착된 프라이머 서열의 일부로서 추가의 바코드 서열을 제공하는 단계로서, 여기서 프라이머 서열이 또한 랜덤 N-량체 서열을 포함하는, 단계 및 추가의 바코드 서열을 부착시키기 전에 입자로부터 프라이머를 추가의 액적으로 박리시키는 단계를 추가로 포함한다. 일부 경우에, 샘플 핵산을 특성화하는 방법은 추가의 바코드 서열을 프라이머를 사용하는 증폭 반응으로 참조 핵산의 단편에 또는 참조 핵산의 일부의 카피에 부착시키는 단계를 추가로 포함한다. 일부 경우에, 샘플 핵산을 특성화하는 방법은 참조 핵산 서열에서 하나 이상의 유전적 변형을 결정하는 단계를 추가로 포함한다.

[0033] 일부 경우에, 참조 핵산 서열에서 하나 이상의 유전적 변형은 구조적 변형 및 단일 뉴클레오타이드다형성(SNP)으로 이루어진 그룹으로부터 선택된다. 일부 경우에, 참조 핵산 서열에서 하나 이상의 유전적 변형은 카피 수 변형, 삽입, 결실, 레트로트랜스포존, 전좌, 역전, 재배열, 반복 확대 및 중복으로 이루어진 그룹으로부터 선택된다. 일부 경우에, 참조 핵산은 생식 세포 핵산 서열을 포함한다. 일부 경우에, 참조 핵산은 암 핵산 서열을 포함한다. 일부 경우에, 샘플 핵산 서열은 10 kb를 초과하는 길이를 갖는다. 일부 경우에, 참조 핵산은 질환 상태의 부재를 나타내는 게놈으로부터 유도된다. 일부 경우에, 참조 핵산은 질환 상태를 나타내는 게놈으로부터 유도된다. 일부 경우에, 질환 상태는 암을 포함한다. 일부 경우에, 질환 상태는 이수성을 포함한다. 일부 경우에, 무세포 샘플 핵산은 종양 핵산을 포함한다. 일부 경우에, 종양 핵산은 순환성 종양 핵산을 포함한다.

[0034] 본 발명의 추가의 국면 및 이점은 다음 상세한 설명으로부터 당업자에게 용이하게 자명해질 것이고, 단지 본 발명의 예시적 구현예가 제시되고 기재된다. 실현되는 바와 같이, 본 발명은 다른 상이한 구현예도 가능하고, 이의 다수의 상세함은 모두가 본발명으로부터 벗어나지 않고 다양한 명백한 관점에서 변경될 수 있다. 따라서, 도면 및 상세한 설명은 특성상 예시로서 제한적인 것으로 간주되지 않아야 한다.

[0035] 참조 인용

[0036] 본 명세서에서 언급된 모든 공보, 특허 및 특허 출원은 각각의 개별적 공보, 특허 또는 특허 출원이 구체적이고 개별적으로 참조로 인용되도록 지시되는 바와 같이 동일한 정도로 그 전체가 본원에 참조로 인용된다.

도면의 간단한 설명

[0037] 본 발명의 신규한 특징은 첨부된 특허청구범위에 상세하게 기재되어 있다. 본 발명의 특징 및 장점의 우수한 이해는 본 발명의 원리를 이용하는 예시적 구현예를 나타내는 다음 상세한 설명 및 다음과 같은 동반되는 도면에 의해 수득될 것이다:

도 1은 종래의 공정 대 실시예 공정 및 본원에 기재된 시스템을 사용하는 위상 변이체의 동정 및 분석의 개략도를 제공한다.

도 2는 종래의 공정 대 실시예 공정 및 본원에 기재된 시스템을 사용하는 구조적 변형의 동정 및 분석의 개략도를 제공한다.

도 3은 본원에 개시된 방법 및 조성물을 사용하여 카피 수 및 일배체형을 검출하기 위한 검정을 수행하는 예시적 작업 흐름을 도시한다.

도 4는 핵산 샘플과 비드를 배합하고 핵산 및 비드를 별개의 액적으로 구획화하기 위한 예시적 공정의 개략도를 제공한다.

도 5는 염색체 핵산 단편의 바코딩 및 증폭을 위한 예시적 공정의 개략도를 제공한다.

도 6은 서열 데이터를 개별 염색체로 귀속시키는데 있어서 염색체 핵산 단편의 바코딩의 예시적 사용의 개략도를 제공한다.

도 7은 위상 서열 분석 공정의 예의 개략도를 제공한다.

도 8은 건강한 환자(상부 패널) 및 일배체형 카피 수의 증가(중양 패널) 또는 일배체형 카피 수의 손실(하부 패널)을 갖는 암 환자의 게놈의 예시적 부분집합의 개략도를 제공한다.

도 9a-b는 다음을 제공한다: (a) 종양 DNA의 상대적 기여도를 나타내는 개략도 및 (b) 통상적인 서열분석 방법에 의해 상기 카피 증가 및 손실을 검출하는 대표도.

도 10은 단일 변이체 위치(좌측 패널) 및 배합된 변이체 위치(우측 패널)를 사용하여 카피 증가 및 손실을 검출하는 예의 개략도를 제공한다.

도 11은 카피 수의 증가 및 손실을 동정하기 위한 기재된 방법 및 시스템의 전위의개략도를 제공한다.

도 12는 본원에 기재된 방법 및 조성물을 사용하는 염색체 수 및 카피 수 변형의 결정에 기초하여 이수성 시험을 수행하기 위한 예시적 작업 흐름을 도시한다.

도 13a-b는 유전 샘플 중의 전좌 및 유전자 융합과 같은 구조적 변형을 동정하는 공정의 예시적 개요를 도시한다.

도 14는 본원에 기재된 방법 및 조성물을 사용하여 카피 수 변동의 결정에 기초하여 암 진단 시험을 수행하기 위한 예시적 작업 흐름을 도시한다.

도 15는 NCI-H2228 암 세포주로부터 EML-4-ALK 구조적 변형의 개략도를 제공한다.

도 16a 및 16b는 음성 대조군 세포주(도 16b)와 비교하여 암 세포주(도 16a)에서 도 15에 도시된 EML-4-ALK 변이체 구조의 존재를 동정하기 위한 본원에 기재된 시스템을 사용하는 바코드 맵핑 데이터를 제공한다.

도 17은 본원에 기재된 부계 핵산 서열을 분석하는 예시적 작업 흐름을 개략적으로 도시한다.

도 18은 본원에 기재된 모계 핵산 서열을 분석하는 예시적 작업 흐름을 개략적으로 도시한다.

도 19는 본원에 기재된 태아 핵산 서열을 분석하는 예시적 작업 흐름을 개략적으로 도시한다.

도 20은 본원에 기재된 참조 핵산 서열을 분석하는 예시적 작업 흐름을 개략적으로 도시한다.

도 21은 본원에 기재된 샘플 핵산 서열을 분석하는 예시적 작업 흐름을 개략적으로 도시한다.

도 22는 예시적 컴퓨터 대조 시스템을 개략적으로 도시한다.

발명을 실시하기 위한 구체적인 내용

상세한 설명

본 발명의 다양한 구현예가 본원에 제시되고 기재되었지만, 이러한 구현예가 단지 예로써 제공된다는 것이 당업자에게 명백할 것이다. 다수의 변형, 변화 및 치환이 본 발명으로부터 벗어나지 않고 당업자에게 발생할 수 있다. 본원에 기재된 본 발명의 구현예에 대한 다양한 대안들이 사용될 수 있다는 것이 이해되어야 한다.

본원에 사용된 용어 "유기체"는 일반적으로 연속 생명 시스템을 의미한다. 유기체의 비제한적인 예는 동물(예: 인간, 다른 종류의 포유류, 조류, 파충류, 곤충, 본원의 다른 곳에 기재된 동물의 다른 예시적 형태), 식물, 진균 및 세균을 포함한다.

본원에 사용된 용어 "콘택"은 일반적으로 소정의 길이의 연속 핵산 서열을 의미한다. 연속 서열은 짧거나 긴 관독 서열을 포함하는 개별 서열 관독치로부터 또는 관독치 내의 중복 서열에 기초하여 정렬되고 조립되거나, 또는 다른 공지된 결합 데이터, 예를 들면, 본원의 다른 곳에서 기재된 바와 같은 공통 바코드를 갖는 태깅에 기초하는 단편 내에 연결된 것으로 정의되는 서열 관독치의 조립체로부터 유도될 수 있다. 이러한 중복 서열 관독치는 또한 짧은 관독치, 예를 들면, 500 미만 염기, 예를 들면, 일부 경우에, 약 100 내지 500 염기, 일부 경우에 100 내지 250 염기를 포함할 수 있거나, 보다 긴 서열 관독치, 예를 들면, 500 초과 염기, 1000 염기 또는 심지어 10,000 초과 염기에 기초한다.

I. 개요

본 발명은 유전 물질의 특성화에 상당한 진보를 제공하는데 유용한 방법 및 시스템을 제공한다. 일부 경우에, 상기 방법 및 시스템은, 예를 들면, 일배체형 위상, 구조적 변형, 예를 들면, 결실, 중복, 카피-수 변형, 삽입, 역전, 레트로트랜스포존, 전좌, LTR, STR의 동정을 포함하는 일반적으로 이용가능한 기술을 사용하는 매우 어렵거나 심지어 불가능한 유전적 특성화 및 다양한 다른 유용한 특성화를 제공하는데 유용할 수 있다.

일반적으로, 본원에 기재된 방법 및 시스템은, 예를 들면, 원래 서열에서 상당한 거리로 분리되어있는, 예를 들면, 짧은 관독 서열 분석 기술에 의해 제공된 것보다 더 긴 다른 변형을 함유하는 서열 세그먼트를 포함하는 상이한 서열 세그먼트에 대한 변형에 관련되는 긴 범위 변이체 정보의 동정 및 사용을 가능하게 하는 긴 개별 핵산 분자의 서열 분석을 위해 제공함으로써 상기 목적을 달성한다. 그러나, 이러한 방법 및 시스템은 보고된 긴 관독 길이 서열 분석 기술, 예를 들면, SMRT 서열 분석 및 나노 기공 서열 분석 기술과 같은 단일 분자 서열 분석보다 훨씬 낮은, 짧은 관독 서열 분석 기술의 매우 낮은 서열 분석 오차율의 이점으로 이러한 목적을 달성한

다.

- [0045] 일반적으로, 본원에 기재된 방법 및 시스템은 긴 핵산 분자를 고처리량, 보다 높은 정확성 짧은 판독 서열 분석 기술을 사용하여 서열 분석가능한 짧은 단편으로 분할하지만, 이러한 분할은 서열 정보가 원래의 보다 긴 개별 핵산 분자에 귀속되는 짧은 단편으로부터 유래되도록 하는 방식으로 수행한다. 서열 판독치를 원래 긴 핵산 분자에 귀속시킴으로써, 일반적으로 짧은 서열 판독치 단독으로부터 수득할 수 없는, 그 긴 핵산 서열에 대한 상당한 특성화 정보를 획득할 수 있다. 언급한 바와 같이, 이러한 특성화 정보는 일배체형 위상, 구조적 변형의 동정 및 카피 수 변형의 동정을 포함할 수 있다.
- [0046] 본원에 기재된 방법 및 시스템의 장점은 다수의 일반적인 예와 관련하여 기재된다. 제1 예에서, 위상 서열 변이체를 동정하고, 본원에 기재된 방법 및 시스템을 사용하여 특성화한다. 도 1은 위상 변이체 소집 문제 및 본원에 기재된 방법에 의해 제공된 해결책을 도시한다. 도시된 바와 같이, 패널 I 중 핵산(102 및 104)은 상이한 염색체, 예를 들면, 모계 및 부계 상속 염색체의 동일 영역의 두 개의 반수체 서열을 나타낸다. 각 서열은 각각의 반수체 서열을 특성화하는 상이한 대립 유전자에서 일련의 변이체, 예를 들면, 핵산(102) 상의 변이체(106-114) 및 핵산(104) 상의 변이체(116-122)를 포함한다. 그들의 매우 짧은 서열 판독치 때문에, 대부분의 서열 분석 기술은 동일한 반수체 서열 상의 다른 변이체에 대한 개별 변이체의 맥락을 제공하기 위해 사용할 수 없다. 추가로, 그들은 개별 분자 성분, 예를 들면, 각각의 반수체 서열을 분리하지 않는 샘플 제조 기술에 의존하기 때문에, 하나는 다양한 변이체의 위상, 예를 들면, 변이체가 유래되는 반수체 서열을 동정할 수 없다. 결과적으로, 이들 짧은 판독 기술은 이러한 변이체를 이들의 원래 분자로 분해할 수 없다. 이러한 접근법의 어려움은 패널 IIa 및 IIIa에 개략적으로 도시된다. 간단히, 패널 IIa에 도시된 두 반수체 서열로부터 풀링된 단편은 서열 분석되어 다수의 짧은 서열판독치(124)를 초래하고, 생성된 서열(126)은 조립된다(패널 IIIa에 도시됨). 도시된 바와 같이, 하나가 패널 IIa 중 임의의 보다 짧은 서열 판독치의 상대적 위상 맥락을 갖지 않기 때문에, 하나는 조립 공정에서 두 개의 상이한 반수체 서열 사이에서처럼 변이체를 분해할 수 없을 것이다. 따라서, 패널 IIIa에 도시된 생성되는 조립체는 변이체(106-122) 모두를 포함하여 단일 공통 서열조립체(126)를 초래한다.
- [0047] 대조적으로, 도 1의 패널 IIb에 도시된 바와 같이, 본원에 기재된 방법 및 시스템은 상기된 접근법의 경우와 같이 긴 핵산(102 및 104)을 보다 짧은 서열 분석가능한 단편으로 분해하거나 분할하지만, 이러한 단편에 그들의 원래 분자 맥락으로 귀속시키는 능력은 유지시킨다. 이는 패널 IIb에 개략적으로 도시되고, 여기서, 상이한 단편은 이들의 원래 분자 맥락에 따라서 그룹화되거나 "구획화된다". 본 발명의 맥락에서, 이 그룹화는 단편을 분자 맥락을 유지하는 그룹으로 물리적으로 구획화 뿐만 아니라 후속적으로 그 맥락을 설명할 수 있기 위해 이들 단편을 태깅함 중의 하나 또는 둘 다를 통해 달성될 수 있다.
- [0048] 이 그룹화는 각각 핵산 (102 및 104)으로부터 짧은 서열 판독치를 나타내는 그룹(128 및 130) 사이에서와 같이 보다 짧은 서열 판독치의 할당으로서 개략적으로 도시된다. 원래 서열 맥락이 서열 분석 공정을 통해 유지되기 때문에, 각각 서열 (102 및 104) 사이에서와같이 다양한 변이체 (106-114 및 116-122)의 원래 분자 맥락, 예를 들면, 위상을 분해하는데 그 맥락을 사용할 수 있다.
- [0049] 또 다른 예시적인 유리한 적용에서, 방법 및 시스템은 짧은 판독 서열 기술을 사용하여 일반적으로 동정가능하지 않거나 적어도 동정하기 어려운 구조적 변이체를 특성화하는데 유용하다.
- [0050] 이는 도 2에 간단한 전좌 이벤트를 참조하여 개략적으로 도시된다. 도시된 바와 같이, 게놈 샘플은 전좌 이벤트, 예를 들면, 서열(202)로부터 유전 요소(206)의 서열(204)로의 전좌를 포함하는 핵산을 포함할 수 있다. 이러한 전좌는, 예를 들면, 동일한 염색체의 상이한 영역 사이의 동일한 영역 또는 상이한 영역으로든 상이한 염색체 사이의 전좌를 포함하는 다양한 상이한 전좌 형태 중의 어느 하나일 수 있다.
- [0051] 다시, 상기 도 1에 도시된 예의 경우와 같이, 종래의 서열 분석은 패널 I에서 서열(202 및 204)을 작은 단편으로 분해하고, 패널 IIa에 도시된 바와 같이, 이들 단편으로부터 짧은 서열 판독치(208)를 생성함으로써 개시한다. 이러한 서열 단편(208)은 비교적 짧기 때문에, 전좌된 서열(206), 즉, 동일하거나 상이한 서열 상의 변이체 위치로부터 기원하는 맥락은 조립 공정 동안 용이하게 손실된다. 또한, 그들의 짧은 판독 길이 때문에, 서열 조립은 거의 정의에 의해 구조적 변형을 반영하지 않는 참조 서열의 사용을 전제로 한다. 이와 같이, 짧은 서열 판독치(208)은 변함없이 전좌된 서열(206)의 적절한 위치를 무시하도록 조립되고, 대신, 패널 IIIa에 도시된 바와 같이, 비변형 서열(210 및 212)을 조립한다.
- [0052] 대조적으로, 본원에 기재된 방법 및 시스템을 사용하여, 서열 (202 및 204)로부터 유도된 짧은 서열 판독치가 패널 IIb에 도시된 구획화에 보다 적은 서열 단편의 원래 분자 그룹화를 유지하는 그룹(214 및 216)으로 제공되

어 패널 IIIb에 도시된 서열(218 및 220)로서 이들의 조립을 허용하고, 원래 서열(202 및 204)로의 역 귀속 및 전좌 변동, 예를 들면, 패널 IIIb에 도시된 바와 같이, 올바른 서열 조립체(218 및 220) 중의 전좌된 서열 세그먼트(206a)의 동정을 가능하게 한다.

[0053] 상기 언급된 바와 같이, 본원에 기재된 방법 및 시스템은 보다 긴 핵산의 짧은 서열 판독치를 위한 개별적 분자 맥락을 제공한다. 본원에 사용된 바와 같이, 개별적 분자 맥락은, 예를 들면, 서열 판독치 자체 내에 포함되지 않는 인접 또는 근위 서열에 관련되는 특이적 서열 판독치를 초과하는 서열 맥락을 의미하고, 그 자체로, 일반적으로, 그들이 짧은 서열 판독치, 예를 들면, 쌍을 이룬 판독치에 대해 약 150 염기 또는 약 300 염기의 판독치에 전체로 또는 부분적으로 포함되지 않을 정도일 것이다. 일부 국면에서, 방법 및 시스템은 짧은 서열 판독치를 위한 긴 범위 서열 맥락을 제공한다. 이러한 긴 범위 맥락은 소정의 서열 판독치와 서로 1 킬로베이스 (kb) 이상, 5 kb 이상, 10 kb 이상, 15 kb 이상, 20 kb 이상, 30 kb 이상, 40 kb 이상, 50 kb 이상, 60 kb 이상, 70 kb 이상, 80 kb 이상, 90 kb 이상 또는 심지어 100 kb 이상 또는 그 이상의 거리 내에 있는 서열 판독치의 관계 또는 결합을 포함한다. 보다 긴 범위 개별 분자 맥락을 제공함으로써, 본원에 기재된 방법 및 시스템은 또한 더욱 더 긴 추정 분자 맥락을 제공한다. 본원에 기재된 서열 맥락은, 예를 들면, 짧은 서열 판독치의 개별적인 더 긴 분자 또는 연결된 분자의 콘텍트로부터 저해상도 맥락 뿐만 아니라, 예를 들면, 예를 들면, 개별 분자의 연속 결정 서열을 갖는 보다 긴 개별 분자의 많은 부분의 긴 범위 서열 분석으로부터 고해상도 서열 맥락을 포함할 수 있고, 여기서, 상기 결정 서열은 1 kb 이상, 5 kb 이상, 10 kb 이상, 15 kb 이상, 20 kb 이상, 30 kb 이상, 40 kb 이상, 50 kb 이상, 60 kb 이상, 70 kb 이상, 80 kb 이상, 90 kb 또는 심지어 100 kb 이상이다. 서열 맥락의 경우와 같이, 짧은 서열의 긴 핵산, 예를 들면, 개별적인 긴 핵산 분자 또는 연결된 핵산 분자 또는 콘텍트의 수집 모두로의 귀속은 고 수준의 서열 맥락을 제공하기 위해 긴 핵산 연신에 대한 짧은 서열의 맵핑 뿐만 아니라 이러한 긴 핵산을 통해 짧은 서열로부터 조립된 서열을 제공함을 모두 포함할 수 있다.

[0054] 또한, 긴 개별 분자와 관련된 긴 범위 서열 맥락을 사용할 수 있는 반면, 이러한 긴 범위 서열 맥락을 갖는 것은 또한 훨씬 더 긴 범위 서열 맥락을 추정할 수 있도록 한다. 일례로서, 상기한 긴 범위 분자 맥락을 제공함으로써, 상이한 유래 분자로부터의 긴 서열 중에서 중복 변이체 부분, 예를 들면, 위상 변이체, 전좌된 서열 등을 동정하여 이들 분자 사이의 추정 결합을 가능하게 할 수 있다. 이러한 추정 결합 또는 분자 맥락은 본원에서 "추정 콘텍트"로서 칭명된다. 위상 서열의 맥락에서 논의된 일부 경우에, 추정 콘텍트는 일반적으로, 예를 들면, 중복 위상 변이체에 의해, 개별 유래 분자보다 실질적으로 더 긴 길이의 위상 콘텍트를 추정할 수 있는 위상 서열을 나타낼 수 있다. 이러한 위상 콘텍트는 본원에서 "위상 블록"으로서 칭명된다.

[0055] 긴 단일 분자 판독치로 개시함으로써, 다르게 짧은 판독 서열 분석 기술 또는 위상 서열 분석에의 다른 접근법을 사용하여 달성가능한 것보다 긴 추정 콘텍트 또는 위상 블록을 유도할 수 있다. 예를 들면, 다음을 참조한다: 공개된 미국 특허 공보 제 2013/0157870호, 이의 전체 개시 내용은 전체가 본원에 참조로 인용된다. 특히, 본원에 기재된 방법 및 시스템을 사용하여, N50 (그 길이 이상의 모든 위상 블록 또는 콘텍트의 수집이 모든 콘텍트 또는 위상 블록의 길이의 합의 적어도 절반을 함유하고, 그 길이 이하의 모든 콘텍트 또는 위상 블록의 수집이 또한 모든 콘텍트 또는 위상 블록의 길이의 합의 적어도 절반을 함유하는 콘텍트 또는 위상 블록 길이), 방식, 평균, 또는 적어도 약 10 킬로베이스(kb), 적어도 약 20kb, 적어도 약 50kb의 중앙값을 갖는 추정 콘텍트 또는 위상 블록 길이를 수득할 수 있다. 일부 국면에서, 추정 콘텍트 또는 위상 블록 길이는 N50, 방식, 평균, 또는 적어도 약 100kb, 적어도 약 150kb, 적어도 약 200kb의 중앙값을 갖고, 일부 경우에, 적어도 약 250kb, 적어도 약 300 kb, 적어도 약 350 kb, 적어도 약 400 kb, 일부 경우에, 적어도 약 500 kb, 적어도 약 750 kb, 적어도 약 1 Mb, 적어도 약 1.75 Mb, 적어도 약 2.5 Mb 이상이 달성된다. 또 다른 경우에, 20 kb, 40 kb, 50 kb, 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 750 kb, 1 메가베이스(Mb), 1.75 Mb, 2 Mb 또는 2.5 Mb 이상의 최대 추정 콘텍트 또는 위상 블록 길이가 수득될 수 있다. 또 다른 경우에, 추정 콘텍트 또는 위상 블록 길이는 적어도 약 20 kb, 적어도 약 40 kb, 적어도 약 50 kb, 적어도 약 100 kb, 적어도 약 200 kb, 일부 경우에, 적어도 약 500 kb, 적어도 약 750 kb, 적어도 약 1 Mb, 일부 경우에, 적어도 약 1.75 Mb, 적어도 약 2.5 Mb 이상일 수 있다.

[0056] 하나의 국면에서, 본원에 기재된 방법 및 시스템은 샘플 핵산 또는 이의 단편의 별개의 구획실 또는 구획(본원에서 구획으로 상호교환적으로 칭명됨)으로의 구획화, 증착 또는 분할을 제공하고, 여기서, 각 구획은 다른 구획의 내용물로부터 이의 내용물의 분리를 유지시킨다. 고유 식별자, 예를 들면, 바코드는 이전에, 이후에 또는 동시에 특별한 구획 내에 포함된 샘플 핵산, 특히 원래 구획에 증착될 수 있는 연속 샘플 핵산의 비교적 긴 스트레치에 대한 특성, 예를 들면, 핵산 서열 정보의 귀속을 허용하기 위해 구획화되거나 분할된 샘플 핵산을 유지하는 구획으로 전달될 수 있다.

[0057] 샘플 핵산은 핵산이 구획에서 연속 핵산 분자의 비교적 긴 단편 또는 스트레치로 존재하도록 구획화될 수 있다.

이러한 단편은, 예를 들면, 전체 염색체, 엑솜, 또는 다른 대형 게놈 단편과 같은 분석될 전체 샘플 핵산의 다수의 중첩 단편을 나타낼 수 있다. 이러한 샘플 핵산은 전체 게놈, 개별 염색체, 엑솜, 앰플리콘, 또는 목적하는 임의의 다양한 상이한 핵산을 포함할 수 있다. 일부 경우에, 샘플 핵산의 이러한 단편은 100 베이스 이상, 500 베이스 이상, 1 kb 이상, 5 kb 이상, 10 kb 이상, 15 kb 이상, 20 kb 이상, 30 kb 이상, 40 kb 이상, 50 kb 이상, 60 kb 이상, 70 kb 이상, 80 kb 이상, 90 kb 이상 또는 심지어 100 kb 이상일 수 있고, 이는 상기한 더 긴 범위 분자 맥락을 허용한다.

[0058] 샘플 핵산은 또한 소정의 구획이 출발 샘플 핵산의 두 개의 중복 단편을 포함할 매우 낮은 가능성을 갖는 수준으로 구획화될 수 있다. 이는 샘플 핵산을 구획화 공정 동안 낮은 유입량 및/또는 농도로 제공함으로써 달성될 수 있다. 그 결과, 일부 경우에, 소정의 구획은 출발 샘플 핵산의 다수의 길지만 비중복성 단편을 포함할 수 있다. 이어서, 다른 구획 중의 샘플 핵산을 고유 식별자와 관련시키고, 여기서, 임의의 소정의 구획을 경우, 내부에 함유된 핵산은 동일한 고유 식별자를 포함하지만, 상이한 구획은 상이한 고유 식별자를 포함할 수 있다. 또한, 구획화가 샘플 성분을 매우 작은 용적 구획 또는 액적으로 할당하기 때문에, 상기 나타낸 바와 같은 할당을 달성하기 위해, 고용적 공정, 예를 들면, 튜브 또는 다중웰 플레이트의 웰에서 필요할 수 있는 샘플의 실질적인 희석은 수행할 필요가 없다는 것이 이해될 것이다. 또한, 본원에 기재된 시스템이 이러한 높은 수준의 바코드 다양성을 사용하기 때문에, 상기 제공된 바와 같이, 높은 수의 게놈 등가물 중에서 다양한 바코드를 할당할 수 있다. 특히, 상기한 다중웰 플레이트(참조: 예를 들면, 미국 특허 공보 제 2013/0079231호 및 제2013/0157870호, 이의 전체 개시 내용은 그 전체가 본원에 참조로 인용된다)는 단지 백개 내지 수 백개의 상이한 바코드 서열로 작동할 수 있고, 바코드를 상이한 세포/핵산으로 귀속시킬 수 있도록 하기 위해 이들 샘플의 제한적인 희석 공정을 사용한다. 이와 같이, 그들은 일반적으로 100개보다 훨씬 적은 세포로 작동하고, 이는 1:10 정도의 게놈:(바코드 형태)의 비, 확실히 충분히 1:100 이상의 비를 제공할 수 있다. 본원에 기재된 시스템은, 한편으로, 높은 수준의 바코드 다양성, 예를 들면, 10,000, 100,000, 500,000개 등을 초과하는 다양한 바코드 형태가 1:50 이하, 1:100 이하, 1:1000 이하 정도의 게놈:(바코드 형태) 비 또는 훨씬 더 작은 비에서 작동할 수 있는 반면, 또한 게놈당 훨씬 개선된 바코드 다양성을 제공하면서 높은 수의 게놈(예: 검정당 100개 초과 게놈, 검정당 500개 초과 게놈, 검정당 1000개 게놈 또는 그 이상의 정도)의 적재를 가능하게 한다.

[0059] 흔히, 샘플은 구획화 전에 비드에 박리가능하게 부착된 올리고뉴클레오타이드 태그의 세트와 배합된다. 올리고뉴클레오타이드는 적어도 제1 및 제2 영역을 포함할 수 있다. 제1 영역은, 소정의 구획 내의 올리고뉴클레오타이드 사이에서와 같이, 실질적으로 동일한 바코드 서열일 수 있지만, 상이한 구획 사이에서와 같이, 상이한 바코드 서열일 수 있고, 대부분의 경우, 상이한 바코드 서열인 바코드 영역일 수 있다. 제2 영역은 구획 내의 샘플 내 핵산을 프라이밍하는데 사용될 수 있는 N-량체(예: 랜덤 N-량체 또는 특정 서열을 표적화하도록 설계된 N-량체)일 수 있다. 일부 경우에, N-량체가 특정 서열을 표적화하도록 설계된 경우, 이는 특정 염색체(예: 염색체 1, 13, 18 또는 21), 또는 염색체의 영역, 예를 들면, 엑솜 또는 다른 표적화 영역을 표적화하도록 설계될 수 있다. 일부 경우에, N-량체는 특정 유전자 또는 유전 영역, 예를 들면, 질환 또는 장애(예: 암)와 관련된 유전자 또는 영역을 표적화하도록 설계될 수 있다. 구획 내에서, 증폭 반응은 핵산의 길이에 따라 상이한 장소에서 핵산 샘플을 프라이밍하기 위해 제2 N-량체를 사용하여 수행될 수 있다. 증폭의 결과로서, 각 구획은 동일하거나 거의 동일한 바코드에 부착되고, 각 구획에서 핵산의 중복성 작은 단편을 나타낼 수 있는 핵산의 증폭된 제품을 함유할 수 있다. 바-코드는 핵산의 세트가 동일 구획으로부터 기원되고, 따라서 잠재적으로 또한 핵산의 동일한 가닥으로부터 기원된다는 것을 의미하는 마커로서 기능할 수 있다. 증폭 후, 핵산은 서열 분석 알고리즘을 사용하여 풀링되고, 서열 분석되고, 정렬될 수 있다. 짧은 서열 판독치가 그들의 관련 바코드 서열에 의해 정렬되고, 샘플 핵산의 단일 긴 단편에 귀속되기 때문에, 그 서열 상의 동정된 변이체 모두는 단일 기원 단편 및 단일 기원 염색체로 귀속될 수 있다. 또한, 복수의 긴 단편에서 다수의 동일 위치 변이체를 정렬함으로써, 그 염색체 기여를 추가로 특성화할 수 있다. 따라서, 특정 유전적 변이체의 위상에 관한 결론이 이어서 도출될 수 있다. 이러한 정보는 일반적으로 동일한 핵산 가닥 상 또는 상이한 핵산 가닥 상에 존재하는 유전적 변이체의 특정 세트인 일배체형을 동정하는데 유용할 수 있다. 카피 수 변형도 또한 이러한 방식으로 동정될 수 있다.

[0060] 기재된 방법 및 시스템은 현재의 핵산 서열 분석 기술 및 이들의 관련된 샘플 제조 방법에 비해상당한 이점을 제공한다. 일배체형 위상 및 카피 수 변형 데이터는 일반적으로 생물학적 샘플(예: 혈액, 세포 또는 조직 샘플)이 세포의 앙상블로부터 유전 물질을 추출하기 위해 한꺼번에 처리되고, 이를 소정의 서열 분석 기술을 위해 구체적으로 구성된 서열 분석 라이브러리로 전환시키기 때문에 게놈 DNA의 서열 분석에 의해 이용되지 않는다. 이 앙상블 샘플 처리 접근법의 결과로서, 서열 분석 데이터는 일반적으로 비위상 유전자형을 제공하고, 여기서, 유전 정보가 동일한 또는 상이한 염색체 상에 존재하는지의 여부를 결정하는 것이 가능하지 않다.

- [0061] 유전 특성을 특정 염색체에 귀속시키는 무능력 이외에, 이러한 양상불 샘플 제조 및 서열 분석 방법은 또한 주로 샘플 중의 대부분의 성분을 동정하고 특성화하는 경향이 있고, 소수 성분, 예를 들면, 하나의 염색체에 의해 또는 추출된 샘플에서 전체 DNA의 작은 비율을 구성하는 하나 또는 소수의 세포, 또는 혈류에서 순환하는 단편화된 종양 세포 DNA 분자에 의해 기여된 유전 물질을 동정하고 특성화하도록 설계되지 않는다. 기재된 방법 및 시스템은 또한 더 큰 샘플에 존재하는 최소 집단을 검출하는 상당한 이점을 제공한다. 이와 같이, 그들은, 종종 임상 샘플의 적은 부분만이 카피 수 변형을 갖는 조직을 함유하기 때문에, 샘플 중 카피 수 변형을 평가하는데 유용할 수 있다. 예를 들면, 샘플이 임산부의 혈액 샘플인 경우, 샘플의 작은 분획만이 순환성 무세포 태아 DNA를 함유한다.
- [0062] 본원에 개시된 바코딩 기술의 사용은 유전 마커의 소정의 세트에 개별 분자 맥락을 제공하는, 즉(단일 마커와 대조적으로) 유전 마커의 소정의 세트를 개별 샘플 핵산 분자로 귀속시키는 고유한 능력을 부여하고, 변이체 배워된 조립체를 통해 및/또는 특정 염색체에 복수의 단일 핵산 분자 중에서 광범위하거나 심지어 더 긴 범위의 추정된 개별 분자 맥락을 제공한다. 이러한 유전 마커는 특이적 유전자좌, 예를 들면, SNP와 같은 변이체를 포함할 수 있거나, 그들은 짧은 서열을 포함할 수 있다. 또한, 바코딩의 사용은, 예를 들면, 혈류 중 순환성 종양 DNA의 검출 및 특성화를 위한 샘플로부터 추출된 전체 핵산 집단의 소수 성분 및 주요 성분을 구별하는 능력을 촉진시키는 추가의 이점을 부여하고, 또한 임의의 증폭 동안 증폭 바이어스를 감소시키거나 제거한다. 또한, 미세 유체 포맷의 구현은 DNA의 매우 작은 샘플 용적 및 낮은 유입량으로 작동하는 능력 뿐만 아니라 게놈 전체 태깅을 촉진시키기 위해 다수의 샘플 구획(예: 액적)을 신속하게 처리하는 능력을 부여한다.
- [0063] 이미 기재된 바와 같이, 본원에 기재된 방법 및 시스템의 이점은, 그들이 보편적으로 이용가능한 짧은 판독 서열 분석 기술의 사용을 통해 결과를 달성할 수 있다는 것이다. 이러한 기술은 충분히 특성화되고 매우 효과적인 프로토콜 및 시약 시스템과 함께 쉽게 이용가능하고 연구 단체에 널리 분산되는 이점을 갖는다. 이러한 짧은 판독 서열 분석 기술은 다음으로부터 이용가능한 것들을 포함한다: 예를 들면, 일루미나, 인코포레이티드(Illumina, Inc.) (예: GXII, NextSeq, MiSeq, HiSeq, X10), 열-피셔의 이온 토렌트 분열 (예: 이온 양성자 및 이온 PGM), 피로시퀀싱(pyrosequencing) 방법 뿐만 아니라 기타.
- [0064] 특별한 장점은, 본원에 기재된 방법 및 시스템이 이러한 짧은 판독 서열 분석 기술을 이용하고, 이들의 관련된 낮은 오차율로 그렇게 수행한다는 점이다. 특히, 본원에 기재된 방법 및 시스템은 상기한 바와 같이, 그러나 짝짓기 쌍 확장을 제외하고, 1000 bp 이하, 500 bp 이하, 300 bp 이하, 200 bp 이하, 150 bp 이하 또는 훨씬 짧은 개별 서열 분석 판독치; 및 상기 개별 분자 판독 길이에 대해 5% 미만, 1% 미만, 0.5% 미만, 0.1% 미만, 0.05% 미만, 0.01% 미만, 0.005% 미만, 또는 심지어 0.001% 미만의 서열 분석 오차율을 사용하여 개별 분자 판독 길이 또는 맥락을 달성한다.
- [0065] **II. 작업 흐름 개요**
- [0066] 하나의 예시적 국면에서, 본 발명에 기재된 방법 및 시스템은 개별 샘플(예: 핵산)을 별개의 구획으로 증착 또는 구획화를 제공하고, 여기서 각 구획은 다른 구획의 내용물로부터 이의 자체 내용물의 분리를 유지시킨다. 본원에 사용된 바와 같이, 구획은 다양한 상이한 형태, 예를 들면, 웰, 튜브, 마이크로 또는 나노웰, 관통 홀 등을 포함할 수 있는 용기 또는 용기를 의미한다. 그러나, 일부 국면에서, 구획은 유체 스트림 내에서 유동가능하다. 이러한 용기는, 예를 들면, 내부 유체 중심 또는 코어를 둘러싸는 외부 장벽을 갖는 마이크로캡슐 또는 마이크로-소포로 구성될 수 있거나, 그들은 이의 매트릭스 내부에 물질을 비말 동반하고/하거나 보유할 수 있는 매트릭스일 수 있다. 그러나, 일부 국면에서, 이러한 구획은 비수성 연속 상, 예를 들면, 오일 상 내에 수성 유체의 액적을 포함할 수 있다. 다양한 상이한 용기는, 예를 들면, 다음에 기재되어 있다: 미국 특허 출원 제 13/966,150호(2013년 8월 13일에 출원). 또한, 비수성 또는 오일 연속 상 중의 안정한 액적을 생성하기 위한 에멀전 시스템은 다음에 상세히 기재되어 있다: 예를 들면, 미국 특허 공보 제 2010/0105112호, 이의 전체 개시 내용은 그 전체가 본원에 참조로 인용된다. 특정 경우에, 미세 유체 채널 네트워크는 본원에 기재된 구획을 생성하기에 적합할 수 있다. 이러한 미세 유체 장치의 예는 다음에 상세히 기재된 것들을 포함한다: 미국 가출원 제 61/977,804호(2014년 4월 10일에 출원), 이의 전체 개시 내용은 이의 전체가 모든 목적을 위해 본원에 참조로 인용된다. 대체 메커니즘은 또한 세포의 수성 혼합물이 비수성 유체에서 압출되는 다공성 막을 포함하는 개별 세포의 구획화에 사용될 수 있다. 이러한 시스템은 일반적으로 다음으로부터 이용가능하다: 예를 들면, 나노미, 인코포레이티드(Nanomi, Inc.).
- [0067] 에멀전 중 액적의 경우에, 샘플 물질, 예를 들면, 핵산의 별개의 구획으로의 구획화는 일반적으로 수성 샘플 함유 스트림을, 또한 수성 액적이 유동 스트림 구획화 유체 내에서 생성되도록 구획화 유체, 예를 들면, 불소화

오일의 비수성 스트림이 유동하는 접합부로 유동시킴으로써 달성될 수 있고, 여기서 상기 액적은 샘플 물질을 포함한다. 이하 기재된 바와 같이, 구획, 예를 들면, 액적은 또한 공동-구획화된 바코드 올리고뉴클레오타이드를 포함할 수 있다. 임의의 특별한 구획 내의 샘플 물질의 상대적인 양은, 예를 들면, 수성 스트림 중의 샘플의 농도, 수성 스트림 및/또는 비수성 스트림의 유동 속도 등을 포함하는 시스템의 다양한 상이한 파라미터를 조절함으로써 조정될 수 있다. 본원에 기재된 구획은 종종 매우 적은 용적을 가짐을 특징으로 한다. 예를 들면, 액적 기반 구획의 경우에, 액적은 1000 피코리터(pL) 미만, 900 pL 미만, 800 pL 미만, 700 pL 미만, 600 pL 미만, 500 pL 미만, 400pL 미만, 300 pL 미만, 200 pL 미만, 100pL 미만, 50 pL 미만, 20 pL 미만, 10 pL 미만 또는 심지어 1 pL 미만인 전체 용적을 가질 수 있다. 비드와 공동-구획화되는 경우, 구획 내의 샘플 유체 용적은 상기한 용적의 90% 미만, 상기한 용적의 80% 미만, 70% 미만, 60% 미만, 50% 미만, 40% 미만, 30% 미만, 20% 미만 또는 심지어 10% 미만일 수 있다는 것이 이해될 것이다. 일부 경우에, 낮은 반응 용적 구획의 사용은 매우 소량의 출발 시약, 예를 들면, 유입 핵산으로 반응을 수행하는데 유리할 수 있다. 낮은 유입 핵산을 사용하여 샘플을 분석하기 위한 방법 및 시스템은 다음에 제시된다: 미국 가출원 제 62/017,580호(2014년 6월 26일에 출원), 이의 전체 개시 내용은 그 전체가 본원에 참조로 인용된다.

[0068] 샘플이 그들의 각각의 구획으로 도입되면, 본원에 기재된 방법 및 시스템에 따라서, 구획 내의 샘플 핵산은 일반적으로 그러한 핵산의 특성화시, 그들이 그들 각각의 기원으로부터 유도되었다고 간주될 수 있도록 고유 식별자가 제공된다. 따라서, 샘플 핵산은 고유 식별자(예: 바코드 서열)와 공동-구획화될 수 있다. 일부 국면에서, 고유 식별자는 그러한 샘플에 부착될 수 있는 핵산 바코드 서열을 포함하는 올리고뉴클레오타이드 형태로 제공된다. 올리고뉴클레오타이드는 소정의 구획내의 올리고뉴클레오타이드 사이에서와 같이, 내부에 함유된 핵산 바코드 서열은 동일하지만, 상이한 구획 사이에서와 같이, 올리고뉴클레오타이드는 상이한 바코드 서열을 가질 수 있도록 구획화된다. 일부 경우에는, 둘 이상의 상이한 바코드 서열이 존재할 수 있지만, 일부 국면에서, 하나의 핵산 바코드 서열만이 소정의 구획과 관련될 수 있다.

[0069] 핵산 바코드 서열은 올리고뉴클레오타이드 서열 내에 6 내지 약 20개 이상의 뉴클레오타이드를 포함할 수 있다. 이러한 뉴클레오타이드는, 즉, 인접 뉴클레오타이드의 단일 스트레치로 완전히 연속적일 수 있거나, 그들은 하나 이상의 뉴클레오타이드에 의해 분리되는 둘 이상의 별도의 서열로 분리될 수 있다. 일부 경우에, 분리된 서열은 길이가 약 4 내지 약 16개 뉴클레오타이드일 수 있다.

[0070] 공동-구획화된 올리고뉴클레오타이드는 또한 구획화된 핵산의 처리에 유용한 다른 기능적 서열을 포함할 수 있다. 이러한 서열은, 예를 들면, 관련된 바코드 서열을 부착하고 프라이머를 서열 분석하면서 구획 내의 개별 핵산으로부터 게놈 DNA를 증폭시키기 위한 표적화된 또는 랜덤/범용 증폭 프라이머 서열, 예를 들면, 서열의 존재를 동정하거나 바코드화 핵산을 풀 다운(pulling down)시키기 위한 혼성화 또는 프로빙 서열, 또는 임의의 다수의 다른 잠재적인 기능적 서열을 포함한다. 다시, 샘플 물질과 함께, 올리고뉴클레오타이드 및 관련 바코드 및 다른 기능적 서열의 공동-구획화는, 예를 들면, 다음에 기재되어 있다: 미국 가출원 제 61/940,318호(2014년 2월 7일에 출원) 제61/991,018호(2014년 5월 9일에 출원) 및 미국 특허 출원 제 14/316,383호(2014년 6월 26일에 출원) 뿐만 아니라 미국 특허 출원 제 14/175,935호(2014년 2월 7일에 출원), 이의 전체 개시 내용은 그 전체가 본원에 참조로 인용된다.

[0071] 간단히, 하나의 예시적 공정에서, 각각 비드에 박리가가능하게 부착된 상기한 올리고뉴클레오타이드의 다수를 포함할 수 있는 비드가 제공되고, 여기서, 특정 비드에 부착된 올리고뉴클레오타이드는 모두 동일한 핵산 바코드 서열을 포함할 수 있지만, 다양한 바코드 서열의 다수가 사용된 비드 집단에 걸쳐 나타날 수 있다. 일부 경우에, 비드의 집단은 적어도 1000개의 상이한 바코드 서열, 적어도 10,000개의 상이한 바코드 서열, 적어도 100,000개의 상이한 바코드 서열, 또는 일부 경우에, 적어도 1,000,000개의 상이한 바코드 서열을 포함할 수 있는 다양한 바코드 서열 라이브러리를 제공할 수 있다. 추가로, 각 비드는 부착된 다수의 올리고뉴클레오타이드 분자가 제공될 수 있다. 특히, 개별 비드 상에 바코드 서열을 포함하는 올리고뉴클레오타이드의 분자의 수는 적어도 약 10,000개 올리고뉴클레오타이드, 적어도 100,000개 올리고뉴클레오타이드 분자, 적어도 1,000,000개 올리고뉴클레오타이드 분자, 적어도 100,000,000개 올리고뉴클레오타이드 분자, 및 일부 경우에, 적어도 10억개 올리고뉴클레오타이드 분자일 수 있다.

[0072] 올리고뉴클레오타이드는 특정 자극을 비드에 적용시 비드로부터 박리가가능할 수 있다. 일부 경우에, 자극은, 예를 들면, 올리고뉴클레오타이드를 박리시킬 수 있는 광-불안정한 결합의 절단을 통한 광-자극일 수 있다. 일부 경우에, 비드 환경의 온도 상승이 결합의 절단 또는 비드로부터 올리고뉴클레오타이드의 다른 박리를 유도할 수 있는 경우, 열 자극이 사용될 수 있다. 일부 경우에, 비드에 대한 올리고뉴클레오타이드의 결합을 절단하거나 다르게는

비드로부터 올리고뉴클레오타이드의 박리를 유도할 수 있는 화학적 자극이 사용될 수 있다.

[0073] 본원에 기재된 방법 및 시스템에 따라서, 부착된 올리고뉴클레오타이드를 포함하는 비드는 단일 비드 및 단일 샘플이 개별 구획 내에 함유되도록 개별 샘플과 공동-구획화될 수 있다. 일부 경우에, 단일 비드 구획이 목적시되는 경우, 유체의 상대적 유동 속도는, 평균적으로, 구획이 점유된 이들 구획이 주로 단독으로 점유됨을 보장하기 위해 구획당 하나 미만의 비드를 함유하도록 조절될 수 있다. 마찬가지로, 높은 비율의 구획이 점유되어, 예를 들면, 점유되지 않은 구획의 적은 비율만을 허용함을 제공하기 위해 유동 속도를 조절하는 것을 원할 수 있다. 일부 국면에서, 유동 및 채널 구조는 점유되지 않은 구획의 특정 수준 미만 및 다중 점유된 구획의 특정 수준 미만의 단독으로 점유된 구획의 목적하는 수를 보장하도록 조절된다.

[0074] 도 3은 바코딩에 이어, 예를 들면, 카피 수 변형 또는 일배체형 검정을 위해 사용하기 위한 샘플 핵산을 서열 분석하기 위한 예시적 방법을 도시한다. 먼저, 핵산을 포함하는 샘플은 공급원(300)으로부터 수득할 수 있고, 바코드화 비드의 세트도 또한 수득될 수 있다(310). 비드는 하나 이상의 바코드 서열 뿐만 아니라 랜덤 N-량체 또는 다른 프라이머와 같은 프라이머를 함유하는 올리고뉴클레오타이드에 연결될 수 있다. 일부 경우에, 바코드 서열은 바코드화 비드로부터, 예를 들면, 바코드와 비드 사이의 결합의 절단을 통해 또는 바코드를 박리시키는 기초 비드의 분해를 통해 또는 둘의 조합으로 박리가능하다. 예를 들면, 일부 국면에서, 바코드화 비드는 바코드 서열을 박리시키는 환원제와 같은 제제에 의해 분해 또는 용해될 수 있다. 이 예에서, 핵산(305)을 포함하는 적은 양의 샘플, 바코드화 비드(315), 및, 일부 경우에, 다른 제제, 예를 들면, 환원제(320)를 배합하고, 구획화에 적용한다. 예로써, 이러한 구획화는 액적 생성 시스템, 예를 들면, 미세 유체 장치(325)에 성분을 도입함을 포함할 수 있다. 미세 유체 장치(325)의 보조로, 유중수 에멀전(330)이 형성되고, 여기서, 상기 에멀전은 샘플 핵산(305), 환원제(320) 및 바코드화 비드(315)를 함유하는 수성 액적을 함유한다. 환원제는 바코드화 비드를 용해시키거나 분해하여 액적(335) 내의 비드로부터 바코드 및 랜덤 N-량체를 포함하는 올리고뉴클레오타이드를 박리시킬 수 있다. 이어서, 랜덤 N-량체는 샘플 핵산의 상이한 영역을 프라이밍하여 증폭 후 샘플의 증폭된 카피를 유도할 수 있고, 여기서 각 카피는 바코드 서열(340)로 태깅된다. 일부 경우에, 각 액적은 동일한 바코드 서열 및 상이한 랜덤 N-량체 서열을 함유하는 올리고뉴클레오타이드의 세트를 함유한다. 후속적으로, 에멀전은 파괴되고(345), 추가의 서열(예: 특정 서열 분석 방법, 추가의 바코드 등에 보조하는 서열)을, 예를 들면, 증폭 방법(350)을 통해 첨가할 수 있다(예: PCR). 이어서, 서열 분석을 수행할 수 있고(355), 알고리즘을 서열 분석 데이터를 해석하기 위해 적용한다(360). 서열 분석 알고리즘은 일반적으로, 예를 들면, 서열 분석 판독치를 정렬 시키고/시키거나 특정 서열 판독치가 속하는 샘플을 동정하기 위해 바코드의 분석을 실행할 수 있다.

[0075] 상기 언급된 바와 같이, 단일 비드 점유도가 목적시될 수 있지만, 다중 점유된 구획 또는 점유되지 않은 구획이 종종 존재할 수 있음이 이해될 것이다. 샘플 및 바코드 올리고뉴클레오타이드를 포함하는 비드를 공동-구획화하기 위한 미세 유체 채널 구조의 예는 도 4에 개략적으로 도시된다. 도시된 바와 같이, 채널 세그먼트(402, 404, 406, 408 및 410)가 채널 접합부(412)에서 유체 통신으로 제공된다. 개별 샘플(414)을 포함하는 수성 스트림은 채널 접합부(412)를 향해 채널 세그먼트(402)를 통해 유동한다. 본원의 다른 곳에서 기재된 바와 같이, 이들 샘플은 구획화 공정 전에 수성 유체 내에 현탁될 수 있다.

[0076] 동시에, 바코드 운반 비드(416)를 포함하는 수성 스트림은 채널 접합부(412)를 향해 채널 세그먼트(404)를 통해 유동한다. 비수성 구획화 유체는 측면 채널(406 및 408) 각각으로부터 채널 접합부(412)로 도입되고, 배합된 스트림은 유출 채널(410)로 유동한다. 채널 접합부(412) 내에서, 채널 세그먼트(402 및 404)로부터의 두 개의 배합된 수성 스트림이 배합되고, 공동-구획화된 샘플(414) 및 비드(416)를 포함하는 액적(418)으로 구획화된다. 이미 언급한 바와 같이, 채널 접합부(412)에서 배합하는 유체 각각의 유동 특성을 조절할 뿐만 아니라 채널 접합부의 기하학을 조절함으로써, 생성되는 구획(418) 내에서 비드, 샘플 또는 둘 다의 목적하는 점유도 수준을 달성하기 위해 조합 및 구획화를 최적화할 수 있다.

[0077] 알 수 있는 바와 같이, 예를 들면, 화학적 자극, 핵산 확대, 전사 및/또는 증폭 시약, 예를 들면, 폴리머라제, 역전사 효소, 뉴클레오사이드 트리포스페이트 또는 NTP 유사체, 프라이머 서열 및 추가의 공동 인자, 예를 들면, 이러한 반응에 사용된 2가 금속 이온, 결합 반응 시약, 예를 들면, 리가제 효소 및 결합 서열, 염료, 표지, 또는 다른 태깅 시약을 포함하는 다수의 다른 시약은 샘플 및 비드와 함께 공동-구획화될 수 있다.

[0078] 공동-구획화되면, 비드 상에 배치된 올리고뉴클레오타이드는 구획화된 샘플을 바코드화하고 증폭시키기 위해 사용될 수 있다. 샘플을 증폭시키고 바코딩하는데 이러한 바코드 올리고뉴클레오타이드를 사용하는 예시적인 공정은 다음에 상세히 기재된다: 미국 특허 출원 제 61/940,318호(2014년 2월 7일에 출원) 제61/991,018호(2014년 5월 9일에 출원) 및 미국 특허 출원 제 14/316,383호(2014년 6월 26일에 출원), 이의 전체 개시 내용은 그 전체가 본

원에 참조로 인용된다. 간단히, 하나의 국면에서, 올리고뉴클레오티드는 샘플과 함께 공동-구획화되고, 샘플을 갖는 구획에서 이들의 비드로부터 박리되는 비드 상에 존재한다. 올리고뉴클레오티드는, 바코드 서열과 함께 이의 5' 말단에 프라이머 서열을 포함할 수 있다. 이 프라이머 서열은 샘플의 다수의 상이한 영역을 무작위로 프라이밍하도록 의도된 랜덤 올리고뉴클레오티드 서열일 수 있거나, 이는 샘플의 특정 표적화 영역의 상류를 프라이밍하도록 표적화된 특이적 프라이머 서열일 수 있다.

[0079] 박리되면, 올리고뉴클레오티드의 프라이머 부분은 샘플의 상보성 영역으로 어닐링할 수 있다. 확대 반응 시약, 예를 들면, DNA 폴리머라제, 뉴클레오시드 트리포스페이트, 공동 인자(예: 또한 샘플 및 비드와 공동-구획화된 Mg^{2+} 또는 Mn^{2+} 등)는 이어서, 프라이머가 어닐링되는 주형의 가닥에 대한 상보성 단편을 생성하기 위해 주형으로 샘플을 사용하여 프라이머 서열을 확장시키고, 상보성 단편은 올리고뉴클레오티드 및 이의 관련 바코드 서열을 포함한다. 샘플의 상이한 부분으로 다중 프라이머의 어닐링 및 확대는 샘플의 중복성 상보성 단편의 거대한 풀을 초래할 수 있고, 각각은 그것이 생성된 구획을 나타내는 자체 바코드 서열을 포함한다. 일부 경우에, 이러한 상보성 단편은 자체가 다시 바코드 서열을 포함하는 보충물의 보충물을 생성하기 위해 구획에 존재하는 올리고뉴클레오티드에 의해 프라이밍되는 주형으로서 사용될 수 있다. 일부 경우에, 이 복제 공정은 제1 보충물이 복제되는 경우, 그것이 이의 말단 또는 부근에 두 개의 상보성 서열을 생성하여 반복 카피를 추가로 생성하기 위한 기초가 되는 분자의 능력을 감소시키는 헤어핀 구조 또는 부분적 헤어핀 구조의 형성을 가능하게 하도록 구성된다. 이의 일례의 개략도는 도 5에 도시된다.

[0080] 도면이 도시하는 바와 같이, 바코드 서열을 포함하는 올리고뉴클레오티드는, 예를 들면, 샘플 핵산(504)과 함께 에멀전 중의 액적(502)으로 공동-구획화된다. 본원의 다른 곳에서 언급된 바와 같이, 올리고뉴클레오티드(508)는 샘플 핵산(504)과 공동-구획화되는 비드(506) 상에 제공될 수 있고, 이 올리고뉴클레오티드는 패널 A에 도시된 바와 같이, 비드(506)로부터 박리가 가능할 수 있다. 올리고뉴클레오티드(508)는 하나 이상의 기능적 서열, 예를 들면, 서열 510, 514 및 516) 이외에 바코드 서열(512)을 포함한다. 예를 들어, 올리고뉴클레오티드(508)는 바코드 서열(512), 뿐만 아니라 소정의 서열 시스템을 위한 부착 또는 고정화 서열로서 기능할 수 있는 서열(510), 예를 들면, Illumina HiSeq 또는 Miseq 시스템의 유동 세포에서 부착용으로 사용되는 P5 서열을 포함하는 것으로 도시된다. 도시된 바와 같이, 올리고뉴클레오티드는 또한 샘플 핵산(504)의 부분의 프라이밍 복제를 위한 랜덤 또는 표적화 N-량체를 포함할 수 있는 프라이머 서열(516)을 포함한다. 또한, 올리고뉴클레오티드(508) 내에 서열 분석 시스템에서 합성 반응에 의해 폴리머라제 매개된, 주형 지시된 서열 분석을 프라이밍하는데 사용되는 서열 분석 프라이밍 영역, 예를 들면, "read1" 또는 R1 프라이밍 영역을 제공할 수 있는 서열(514)이 포함된다. 일부 경우에, 바코드 서열(512), 고정화 서열(510) 및 R1 서열(514)은 소정의 비드에 부착된 모든 올리고뉴클레오티드에 공통일 수 있다. 프라이머 서열(516)은 랜덤 N-량체 프라이머에 따라 가변적일 수 있거나, 특정 표적화 적용을 위한 소정의 비드 상의 올리고뉴클레오티드에 공통일 수 있다.

[0081] 프라이머 서열(516)의 존재에 기초하여, 올리고뉴클레오티드는 폴리머라제 효소 및 또한 비드(506) 및 샘플 핵산(504)과 공동-부분화된 다른 확대 시약을 사용하여 올리고뉴클레오티드(508 및 508a)의 확대를 가능하게 하는 패널 B에 도시된 바와 같은 샘플 핵산을 프라이밍할 수 있다. 패널 C에 도시된 바와 같이, 랜덤 N-량체 프라이머의 경우, 샘플 핵산(504)의 복수의 상이한 영역에 어닐링하는 올리고뉴클레오티드의 확대 후, 핵산의 다중 중복 보충물 또는 단편, 예를 들면, 단편(518 및 520)이 생성된다. 샘플 핵산의 부분에 상보성인 서열 부분, 예를 들면, 서열(522 및 524)을 포함하지만, 이들 작제물은 본원에서 일반적으로 부착된 바코드 서열을 갖는 샘플 핵산(504)의 단편을 포함하는 것으로 칭명된다. 알 수 있는 바와 같이, 상기한 바와 같은 주형 서열의 복제 부분은 종종 본원에서 그 주형 서열의 "단편"으로 칭명된다. 그러나, 전술한 내용에도 불구하고, 용어 "단편"은 주형 서열의 일부를 제공하는 다른 메커니즘, 예를 들면, 소정의 서열 분자의 실제 단편화에 의해, 예를 들면, 효소, 화학적 또는 기계적 간편화를 통해 생성된 것들을 포함하여 기원 핵산 서열, 예를 들면, 주형 또는 샘플 핵산의 일부의 임의의 대표를 포함한다. 그러나, 일부 국면에서, 주형 또는 샘플 핵산 서열의 단편은 기본 서열 또는 이의 보충물의 복제부분을 나타낼 수 있다.

[0082] 이어서, 바코드화 핵산 단편은, 예를 들면, 서열 분석을 통해 특성화에 적용할 수 있거나, 그들은 패널 D에 도시된 바와 같이, 공정에서 추가로 증폭될 수 있다. 예를 들면, 또한 비드(306)로부터 박리된 추가의 올리고뉴클레오티드, 예를 들면, 올리고뉴클레오티드(508b)는 단편(518 및 520)을 프라이밍할 수 있다. 특히, 다시, (일부 경우에, 소정의 구획 중의 다른 랜덤 N-량체, 예를 들면, 프라이머 서열(516)과 상이할 수 있는) 올리고뉴클레오티드(508b) 중의 랜덤 N-량체 프라이머(516b)의 존재에 기초하여, 올리고뉴클레오티드는 단편(518)으로 어닐링하고, 샘플 핵산 서열의 일부의 복제본을 포함하는 서열(528)을 포함하는 단편(518)의 적어도 일부에 대한 보충물(526)을 생성하기 위해 확대된다. 올리고뉴클레오티드(508b)의 확대는 그것이 단편(518)의 올리고뉴클레오

티드 부분(508)을 통해 복제될 때까지 계속된다. 본원의 다른 곳에서 언급된 바와 같이, 패널 D에 도시된 바와 같이, 올리고뉴클레오타이드는 목적하는 지점에서, 예를 들면, 단편(518) 내에 포함되는 올리고뉴클레오타이드(508)의 서열(516 및 514)을 통한 복제 후, 폴리머라제에 의한 복제의 중지를 유발하도록 구성될 수 있다. 본원에 기재된 바와 같이, 이는, 예를 들면, 사용된 폴리머라제 효소에 의해 처리될 수 없는 상이한 뉴클레오타이드 및/또는 뉴클레오타이드 유사체의 도입을 포함하는 상이한 방법에 의해 달성될 수 있다. 예를 들면, 이는 비우라실 내성 폴리머라제가 그 영역의 복제를 중지하는 것을 방지하기 위해 서열 영역(512) 내에 뉴클레오타이드를 함유하는 우라실의 포함을 포함할 수 있다. 그 결과, 한쪽 말단에 바코드 서열(512), 부착 서열(510), R1 프라이머 영역(514), 및 랜덤 N-량체 서열(516b)을 포함하는 전장 올리고뉴클레오타이드(508b)를 포함하는 단편(526)이 생성된다. 서열의 나머지 말단에는 제1 올리고뉴클레오타이드(508)의 랜덤 N-량체에 대한 보충물(516') 뿐만 아니라 서열(514')으로서 나타낸, R1 서열의 모두 또는 일부에 대한 보충물이 포함될 수 있다. 이어서, R1 서열(514) 및 이의 보충물(514')은 부분 헤어핀 구조(528)를 형성하기 위해 함께 혼성화할 수 있다. 랜덤 N-량체가 상이한 올리고뉴클레오타이드 중에서 상이하기 때문에 알 수 있는 바와 같이, 이러한 서열 및 이들의 보충물은 랜덤 N-량체 서열(516b)에 상보적인 것으로 기대되지 않는 랜덤 N-량체(516)에 대한 보충물인 헤어핀 형성에, 예를 들면, 서열(516')에 참여하는 것으로 기대되지 않는다. 이는 N-량체가 소정의 구획 내의 올리고뉴클레오타이드 사이에 공통인 다른 적용, 예를 들면, 표적화 프라이머에 대한 경우는 아닐 것이다.

[0083] 이러한 부분 헤어핀 구조를 형성함으로써, 추가의 복제로부터 샘플 서열의 제1 수준 복제물의 제거를 가능하게 하고, 예를 들면, 카피의 반복 카피를 방지한다. 부분 헤어핀 구조는 또한 생성된 단편, 예를 들면, 단편(526)의 후속 처리에 유용한 구조를 제공한다.

[0084] 이어서, 복수의 상이한 구획으로부터의 모든 단편은 본원에 기재된 바와 같은 고처리량 서열 분석기 상의 서열 분석을 위해 풀링될 수 있다. 각 단편이 원래의 이의 구획에 대해 코딩되기 때문에, 그 단편의 서열은 바코드의 존재에 기초하여 이의 기원으로 다시 귀속시킬 수 있다. 이는 도 6에 개략적으로 도시되어 있다. 일례로 도시된 바와 같이, 제1 공급원(600)(예: 개별 염색체, 핵산의 가닥 등)으로부터 유래된 핵산(604) 및 상이한 염색체(602) 또는 핵산의 가닥으로부터 유래된 핵산(606)은 각각 상기한 바와 같은 바코드 올리고뉴클레오타이드의 자체 세트와 함께 구획화된다.

[0085] 각 구획 내에서, 각 핵산(604 및 606)은 이어서 별도로 제1 단편(들)의 제2 단편의 중복 세트, 예를 들면, 제2 단편 세트(608 및 610)를 제공하기 위해 처리한다. 이 처리는 또한 특정 제1 단편으로부터 유래된 제2 단편 각각에 대해 동일한 바코드 서열을 갖는 제2 단편을 제공한다. 도시된 바와 같이, 제2 단편 세트(608)의 바코드 서열은 "1"로 표시되는 반면, 단편 세트(610)의 바코드 서열은 "2"로 표시된다. 바코드의 다양한 라이브러리는 상이한 단편 세트의 다수를 시차적으로 바코드화하는데 사용될 수 있다. 그러나, 상이한 제1 단편으로부터 모든 제2 단편 세트가 상이한 바코드 서열로 바코드화될 필요는 없다. 일부 경우에, 복수의 상이한 제1 단편은 동일한 바코드 서열을 포함하기 위해 동시에 처리될 수 있다. 다양한 바코드 라이브러리는 본원의 다른 곳에서 상세히 기재된다.

[0086] 이어서, 예를 들면, 단편 세트(608 및 610)로부터의 바코드화 단편을, 예를 들면, 일루미나(Illumina)로부터 이용가능한 합성 기술 또는 써모 피셔, 인코포레이티드(Thermo Fisher, Inc.)의 이온 토렌트 분열에 의한 서열을 사용하는 서열 분석을 위해 풀링시킬 수 있다. 서열 분석되면, 서열 판독치(612)는, 예를 들면, 응집된 판독치(614 및 616)에 나타낸 바와 같이, 포함된 바코드에 적어도 부분적으로 기초하고, 일부 경우에, 단편 자체 서열에 부분적으로 기초하는 그들의 각각의 단편 세트에 귀속시킬 수 있다. 이어서, 각 단편 세트에 대한 귀속된 서열 판독치는 각 샘플 단편을 위한 조립된 서열, 예를 들면, 서열(618 및 620)을 제공하기 위해 조립하고, 이는 또한 이들 각각의 기원 염색체(600 및 602)로 추가로 역 귀속될 수 있다. 게놈 서열을 조립하는 방법 및 시스템은, 예를 들면, 다음에 기재된다: 미국 가출원 제 62/017,589호(2014년 6월 26일에 출원), 이의 전체 개시 내용은 그 전체가 본원에 참조로 인용된다. 일부 예에서, 게놈 서열은 신규 조립체 및/또는 참조 기반 조립체(예: 참조에 대한 맵핑)에 의해 조립된다.

[0087] III. 위상 및 카피 수 결정에 방법 및 시스템의 적용

[0088] 본원에 기재된 시스템 및 방법의 하나의 국면에서, 서열 판독치를 긴 기원 분자로 귀속시키는 능력은 서열에 대한 위상 정보를 결정하는데 사용된다. 하나의 예에서, 둘 이상의 특이적 유전자 변이체 서열(예: 대립 유전자, 유전 마커)을 나타내는 서열과 관련된 바코드는 유전 마커의 세트가 샘플 중의 동일 염색체 또는 상이한 염색체 상에 존재하는지의 여부를 결정하기 위해 비교된다. 이러한 위상 정보는 샘플 중의 특정 표적 염색체 또는 유전자의 상대적 카피 수를 결정하기 위해 사용될 수 있다. 기재된 방법 및 증상의 이점은 다수의 위치, 유전자좌,

변이체 등이 위상 및 카피 수 정보를 결정하기 위해 그들이 유래되는 개별 염색체 또는 핵산 가닥을 동정하기 위해 사용될 수 있다는 것이다. 종종, 염색체에 따라 다수의 위치(예: 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 또는 500000 초과)가 본원에 기재된 위상, 일배체형 및 카피 수 변형을 결정하기 위해 사용된다.

[0089] 예로써, 상기 언급된 바와 같이, 본원에 기재된 방법 및 시스템은, 상기한 구체화 및 귀속 국면에의해, 비교적 더 짧은 서열 판독치를 제공할 수 있는 서열 분석 기술을 사용함에도 불구하고, 개별 핵산 단편, 예를 들면, 개별 핵산 분자로부터 효과적인 긴 서열 판독치를 제공하는데 유용할 수 있다. 이러한 긴 서열 판독치가 단일 출발 단편 또는 분자로 귀속될 수 있기 때문에, 서열 중의 변이체 위치는, 마찬가지로, 단일 분자로, 그리고 외삽에 의해 단일 염색체로 귀속될 수 있다. 또한, 동일 염색체로부터 유래되는 것으로 추정될 수 있는 정렬된 서열을 제공하기 위해 인접 단편의 정렬 특징으로서 임의의 소정의 단편 상의 다수의 위치를 사용할 수 있다. 예로써, 제1 단편은 서열 분석될 수 있고, 상기 귀속 방법 및 시스템에 의해, 그 서열 상에 존재하는 변이체는 모두 단일 염색체로 귀속될 수 있다. 이어서, 단지 하나의 염색체 상에 존재하는 것으로 결정되는 이러한 변이체의 복수를 공유하는 제2 단편은 동일 염색체로부터 유래되고 따라서 두 단편의 위상 정렬을 생성하기 위해 제1 단편과 함께 정렬되는 것으로 가정될 수 있다. 이를 반복하면, 긴 범위 위상 정보의 동정을 가능하게 한다. 단일 염색체 상의 변이체의 동정은 다음과 같은 공지된 참조물로부터 획득될 수 있다: 예를 들면, HapMap, 또는, 예를 들면, 다르게 동일한 서열 스트레치 상의 상이한 변이체를 나타내는 서열 분석 데이터의 응집으로부터.

[0090] 도 7은 예시적인 위상 서열 분석 공정의 개략도를 제공한다. 도시된 바와 같이, 예를 들면, 염색체, 염색체 단편, 엑솜 또는 다른 거대한 단일 핵산 분자와 같은 유래되는 핵산(702)은 복수의 거대한 단편(704, 706, 708)으로 단편화될 수 있다. 유래되는 핵산(702)은 특정 핵산 분자, 예를 들면, 염색체에 특이적인 다수의 서열 변이체(A, B, C, D, E, F 및 G)를 포함할 수 있다. 본원에 기재된 공정에 따라서, 유래되는 핵산은 관련 서열 변이체의 서브세트를 포함하는 다수의 거대한 중복 단편(704, 706 및 708)으로 단편화될 수 있다. 이어서, 각 단편은 구체화되고, 추가로 서브단편으로 단편화되고, 거대한 단편의 복수의 중복 바코드화 서브단편을 제공하기 위해 본원에 기재된 바와 같이, 바코드화될 수 있고, 여기서, 소정의 거대한 단편의 서브단편은 동일한 바코드 서열을 포함한다. 예를 들면, 바코드 서열 "1" 및 바코드 서열 "2"와 관련된 서브단편은 각각 구체화(710 및 712)에 도시되고, 이어서 바코드화 서브단편은 풀링되고, 서열 분석될 수 있고, 서열 분석된 서브단편은 긴 단편 서열(714, 716, 및 717)을 제공하기 위해 조립된다. 하나 이상의 긴 단편 서열(714, 716, 및 717)은 다수의 변이체를 포함할 수 있다. 이어서, 긴 단편 서열은 위상 위치가 결정될 수 있는 위상 서열(718)을 제공하기 위해 서열(714, 716, 및 717)로부터 중복성 위상 변이체 정보에 기초하여 추가로 조립될 수 있다.

[0091] 위상 위치가 결정되면, 다양한 방식으로 그 정보를 추가로 활용할 수 있다. 예를 들면, 특정 장애에 대한 유전적 위험을 평가하는데 위상 변이체의 지식을 사용할 수 있고, 부계 때 모계 특성을 동정하고, 이소성을 동정하거나 일배체형 정보를 동정할 수 있다.

[0092] 본원에 개시된 시스템 및 방법의 일부 국면에서, 카피 수 변형 검정은 카피 수 계수의 정확성을 향상시키기 위해 둘 이상의 위상 유전 마커의 동시 검출을 사용하여 수행된다. 위상 정보를 이용하면 다중 유전자좌 및 교차 일배체형에 대한 계수 판독치에만 기초하는 천연 방법하의 분산과 비교하여 신호의 상대적 강도를 증가시킬 수 있다. 추가로, 위상 정보를 이용하면, 신호를 실질적으로 추가로 증폭시키는 위치 특이적 바이어스의 정상화를 가능하게 한다. 카피 수 변형(CNV) 정확성은 서열 분석 깊이, CNV의 길이, 카피 수 등)을 포함하는 수많은 요인에 의존할 수 있다. 본원에 제공된 방법 및 시스템은 적어도 70%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 99%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9%, 99.95%, 99.99%, 99.995% 또는 99.999%의 정확성으로 CNV를 결정할 수 있다. 일부 경우에, 본원에 제공된 방법 및 시스템은 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.1%, 0.05%, 0.01%, 0.005%, 0.001%, 0.0005%, 0.0001%, 0.00005%, 0.00001% 또는 0.000005% 미만의 오차율로 CNV를 결정한다. 유사하게, 본원에 제공된 방법 및 시스템은 적어도 70%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 99%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9%, 99.95%, 99.99%, 99.995% 또는 99.999%의 정확성으로 둘 이상의 유전적 변이체의 위상/일배체형 정보를 검출할 수 있다. 일부 경우에, 본원에 제공된 방법 및 시스템은 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.1%, 0.05%, 0.01%, 0.005%, 0.001%, 0.0005%, 0.0001%, 0.00005%, 0.00001% 또는 0.000005% 미만의 오차율로 위상 또는 일배체형 정보를 결정한다. 본 발명은 또한 유전자좌-특이적 바이어스를 제거하는 방법을 제공하고, 여기서 유전자좌-특이적분산은 적어도 2배, 3배, 4배, 5배, 10배, 20배, 30배, 40배, 50배, 60배, 70배, 80배, 90배, 100배, 200배, 500배, 1000배, 5000배 또는 10000배까지 감소된다. 본원에 제공된 방법 및 시스템은, 예를 들면, 카피 수의 변화가 염색체의 수 또는 염색체의 일부의 변화를 반영하는 경우와 같은 카피 수의 변동을 검

출하기 위해 사용될 수 있다. 일부 경우에, 본원에 제공된 방법 및 시스템은 동일한 염색체 상에 존재하는 유전자의 카피 수의 변동을 검출하기 위해 사용될 수 있다.

[0093] **도 8** (상부 패널)은 건강한 환자의 계놈의 서브셋을 도시하는 개략도이다. 이 환자는 지시된 유전자좌에 이형 유전자형과 별도의 염색체 가닥 위에 위치한 두 개의 별개의 일배체형(1 및 2)(805, 810)을 갖는다. 환자의 자연 발생적 변동(예: SNP 또는 indel)은 원으로 묘사된다. **도 8**은 또한 암 환자의 계놈(815)을 도시한다. 특정 암은 일배체형 카피 수의 증가와 관련된다. 중간 패널은 일배체형(2)의 증가(810)를 도시한다. 암은 또한 일배체형(2)의 손실(820)을 나타내는 **도 8**의 하부 패널에 묘사된 바와 같이, 일배체형 수의 손실과 관련될 수 있다. 공통 서열 분석 기술은 이러한 일배체형 카피의 손실 또는 증가를 정확하게 결정할 수 없다. **도 9a**에 도시된 바와 같이, 이는 부분적으로 환자 혈액 중 종양 기어 DNA(910)는 단지 전체 DNA의 작은 분획이고, 이의 대부분이 정상 조직(905)에 의해 기어 DNA이라는 사실에 기인한다. 이러한 종양 DNA의 저농도는 정상적인 서열 분석 기술에 의한 카피 수의 부정확한 검출을 유도한다(참조: **도 9b**). 평균 깊이(D)(935)에서 카피 변동이 없는 피크(920) 및 카피 손실 피크(925)(940) 및 카피 증가 피크(930)(945)의 예상된 계수의 차이는 검출하기 어렵다. 임의의 소정의 개별 마커의 경우, 복제 시험에서 카피 수 검정 결과의 분포는 포아송 분포(Poisson distribution)에 근접하는 방식으로 올바른 답 주위에 분포될 수 있고, 여기서, 분포의 폭은 검정에서 랜덤 오차의 다양한 공급원에 의존한다. 소정의 샘플의 경우, 카피 수의 변화가 샘플의 비교적 작은 부분일 수 있기 때문에, 단일 유전 마커의 모니터링을 위한 광범위한 가능성 분포는 올바른 결과를 차폐할 수 있다. 이러한 어려움은 정상적인 서열 분석 기술이 **도 10** (좌측 패널)에 도시된 바와 같이, 단지 한번에 일배체형의 하나의 단일 변이체 위치를 관찰한다는 사실에 기인한다. 이러한 기술을 사용하면, 카피 손실(1025), 정상 카피(1020) 및 카피 증가(1030)를 나타내는 피크 사이에 상당한 중복이 존재할 수 있다. 본원에 개시된 기술은 전체 (또는 부분적) 일배체형의 검출을 허용하여 해상도를 증가시키고 카피 증가 및 손실의 검출을 향상시킨다(**도 10**)(우측 패널). 이러한 개선은 **도 11**에 개략적으로 도시되고, 여기서 정상 검출(1100)은 중복 피크를 전개시키는 반면, 본원의 기술(1110)은 카피 증가 또는 손실의 미세한 피크 및 향상된 해상도를 허용한다. 둘 이상의 위상 유전 마커, 특히 단일 염색 상에 공동-위치되는 것으로 공지되고, 따라서 대부분 항상 동기화되지 않은 비-랜덤 방식으로 크거나 작은 수로 나타날 수 있는 마커의 동시 모니터링을 사용하면 예상된 결과 분포의 폭을 좁히는 동시에 계수의 정확성을 향상시키는 효과를 갖는다.

[0094] 암을 검출하고 진단하는 이점 이외에, 본원에 제공된 방법 및 시스템은 또한 태아 이수성을 검출하기에 보다 정확하고 민감한 공정을 제공한다.

[0095] 태아 이수성은 태아 염색체 수의 이상이다. 이수성은 통상적으로 중요한 물리적 및 신경학적 손상을 초래한다. 예를 들면, X 염색체 수의 감소는 터너 증후군의 원인이다. 염색체 번호 21의 카피 수의 증가는 다운 증후군을 유발한다. 양수천자 또는 용모막 용모 샘플링(CVS)과 같은 침습적 시험은 유산 위험을 초래할 수 있고, 모체혈액을 시험하는 덜 침습적 방법이 여기에 사용된다.

[0096] 본원에 기재된 방법은 태아 이수성을 비침습적으로 검출하는데 유용할 수 있다. 예시적인 공정은 **도 12**에 도시되어 있다. 이수성 계놈을 갖는 태아를 가질 위험이 있는 임신부가 검사된다(1200). 태아 유전 물질을 함유하는 모체 혈액 샘플을 수집한다(1205). 이어서, 유전 물질(예: 무세포 핵산)을 혈액 샘플로부터 추출한다(1210). 바코드화 비드 세트도 또한 획득될 수 있다(1215). 비드는 하나 이상의 바코드 서열 뿐만 아니라 랜덤 N-량체 또는 다른 프라이머와 같은 프라이머를 함유하는 올리고뉴클레오타이드에 연결될 수 있다. 일부 경우에, 바코드 서열은 바코드화 비드로부터, 예를 들면, 바코드와 비드 사이의 결합의 절단을 통해 또는 바코드를 박리시키는 기초 비드의 분해를 통해 또는 둘의 조합으로 박리가능하다. 예를 들면, 일부 국면에서, 바코드화 비드는 바코드 서열을 박리시키는 환원제와 같은 제제에 의해 분해 또는 용해될 수 있다. 이러한 예에서, 샘플(1210), 바코드화 비드(1220), 및, 일부 경우에, 다른 시약, 예를 들면, 환원제를 배합하고 구획화에 적용한다. 예로써, 이러한 구획화는 액적 생성 시스템, 예를 들면, 미세 유체 장치(1225)에 성분을 도입함을 포함할 수 있다. 미세 유체 장치(1225)의 보조로, 유중수 에멀전(1230)이 형성될 수 있고, 여기서, 에멀전은 샘플 핵산(1210), 바코드화 비드(1215), 및, 일부 경우에, 환원제를 함유하는 수성 액적을 함유한다. 환원제는 바코드화 비드를 용해시키거나 분해하여 액적(1235) 내의 비드로부터 바코드 및 랜덤 N-량체를 갖는 올리고뉴클레오타이드를 박리할 수 있다. 이어서, 랜덤 N-량체는 샘플 핵산의 상이한 영역을 프라이밍하여 증폭 후 샘플의 증폭된 카피를 유도할 수 있고, 여기서 각 카피는 바코드 서열(1240)로 태깅된다. 일부 경우에, 각 액적은 동일한 바코드 서열 및 상이한 랜덤 N-량체 서열을 함유하는 올리고뉴클레오타이드의 세트를 함유한다. 다른 구현예에서, 개별 액적은 고유 바코드 서열을 포함하고; 또는, 일부 경우에, 액적의 전체 집단의 특정 비율은 고유 서열을 갖는다. 후속적으로, 에멀전은 파괴되고(1245), 추가의 서열(예: 특히 서열 분석 방법, 추가의 바코드 등을 보조하는 서열)을, 예를

들면, 증폭 방법을 통해 첨가할 수 있다(예: PCR). 이어서, 서열 분석은 임의의 적합한 형태의 서열 분석 플랫폼을 통해 수행될 수 있다(예: Illumina, Ion Torrent, Pacific Biosciences SMRT, Roche 454 sequencing SOLiD sequencing 등), (1250), 및 서열 분석 데이터를 해석하기 위해 적용된 알고리즘(1255). 서열 분석 알고리즘은 일반적으로, 예를 들면, 서열 분석 판독치를 정렬시키고/시키거나 특정 서열 판독치가 속하는 샘플을 동정하기 위해 바코드의 분석을 실행할 수 있다. 정렬된 서열은 부착된 고유 바코드에 기초하여 그들 각각의 유전적 기원(예: 염색체)으로 추가로 귀속될 수 있다. 이어서, 염색체 카피 수를 정상 이배체 염색체의 카피 수와 비교한다(1260). 환자는 상이한 염색체의 임의의 카피 수 이상 및 관련 위험/질환을 통보 받는다(1265).

[0097] 위상, 예를 들면, 유전적 변이체가 상이한 염색체 상에 연결되거나 존재하는지의 여부를 결정하는것은 다양한 적용에 대한 유용한 정보를 제공할 수 있다. 예로써, 위상은 질환과 관련되는 게놈의 특정 전좌가 존재할 경우를 결정하는데 유용하다. 이러한 전좌의 검출은 또한 감별 진단 및 수정 치료를 가능하게 할 수 있다. 게놈 중의 어떤 대립 유전자가 연결되는지를 결정하는 것은 유전자가 어떻게 유전되는지를 고려하는데 유용할 수 있다.

[0098] 종종 염색체 쌍의 각 개별 염색체에 대한 대립 유전자의 패턴, 일배체형을 아는 것이 유용할 수 있다. 예를 들면, 하나의 염색체 상에 존재하는 돌연변이를 불활성화하는 두 카피는 제한된 결과일 수 있지만, 두 염색체 사이에 분포되는 경우, 예를 들면, 어떤 염색체도 활성 유전자 산물을 공급하지 않는 경우 현저한 효과를 가질 수 있다. 이러한 효과는, 예를 들면, 질환 위험의 증가 또는 특정 약물에 대한 반응의 부족으로 발현될 수 있다.

[0099] IV. 구조적 변형의 동정/특성화에 방법 및 시스템의 적용

[0100] 다른 적용에서, 본원에서 기재된 방법 및 시스템은 광범위한 범위의 상이한 유전적 구조적 변형의 동정 및 특성화에 대한 긴 범위 분자 서열 정보를 수득하는데 매우 유용하다. 상기 언급된 바와 같이, 이러한 변형은 삽입, 결실, 중복, 레트로트랜스포존, 전좌, 역전 짧은 및 긴 tandem 반복체 등을 포함하는 다양한 종류의 상이한 변형 이벤트를 포함한다. 이러한 구조적 변형은, 그들이 다양한 유전적 질환의 범위와 관련되는 것으로 간주되기 때문에, 상당한 과학적 관심이다.

[0101] 이러한 변형에서의 관심에도 불구하고, 이러한 구조적 변형을 동정하고 특성화하는 몇몇 효과적이고 효율적인 방법이 있다. 부분적으로, 이는 이러한 변형이 비정상적 서열 세그먼트의 존재를 특징으로 하지 않지만, 대신 정상적인 서열 세그먼트로 간주될 수 있는 비정상적인 서열 맥락 또는 단순히 누락 서열 정보를 포함하기 때문이다. 이들의 비교적 짧은 판독 길이 때문에, 대부분의 서열 분석 기술은 상당한 맥락, 특히 그들이 생산하는 서열 판독치의 경우, 예를 들면, 그들의 판독 길이를 초과하는 긴 범위 서열 맥락을 제공할 수 없고, 따라서 조립 공정에서 이러한 변형의 동정을 손실한다. 이러한 변형을 동정하는데 있어서 어려움은 많은 분자, 예를 들면, 복수의 염색체를 배합하여 모두 변형을 포함하고 포함하지 않는 게놈 물질을 포함할 수 있는 공통 서열을 수득하는 이러한 기술의 앙상블 접근법에 의해 더 복잡하게 된다.

[0102] 그러나, 현재 기재된 방법 및 시스템의 맥락에서, 개별 유래 핵산 분자에 귀속될 수 있는 긴 범위서열 정보를 유도하는 짧은 판독 서열 분석 기술을 이용할 수 있고, 따라서 그들 개별 분자에 전체로 또는 부분적으로 함유된 변이체 영역의 긴 범위 서열 맥락을 유지시킬 수 있다.

[0103] 상기한 바와 같이, 본원에 기재된 방법 및 시스템은 개별 유래 핵산 분자에 귀속될 수 있는 긴 범위 서열 정보를 제공할 수 있고, 추가로, 이 긴 범위 서열 정보를 포함하는데 있어서, 이러한 긴 서열 정보의 비교 및 중복을 통해 훨씬 더 긴 범위 서열 맥락을 추정할 수 있다. 이러한 긴 범위 서열 정보 및/또는 추정 서열 맥락은 이용가능한 기술을 사용하여 용이하게 동정되지 않은 다수의 구조적 변형의 동정 및 특성화를 가능하게 한다.

[0104] 도 2에서 단순화된 방식으로 도시되지만, 도 13a 및 13b는 본원에 기재된 방법 및 시스템을 사용하는 구조적 변형의 특정 형태를 동정하기 위한 보다 상세한 예시적인 공정을 제공한다. 도시된 바와 같이, 유기체의 게놈, 또는 유기체의 조직은 일반적으로 도 13a에 도시된 제1 유전자형을 포함할 수 있고, 여기서 제1 유전자(1304)를 포함하는 제1 유전자 영역(1302)은 제2 유전자(1208)을 포함하는 제2 유전자 영역(1306)으로부터 분리된다. 이러한 분리는, 예를 들면, 동일한 엑손 중의 상이한 영역, 동일한 염색체 상의 상이한 엑손, 상이한 염색체 등을 포함하는 유전자 사이의 거리의 범위를 반영할 수 있다. 그러나, 도 13b에 도시된 바와 같이, 유전자(1308)가 그것이 변이체 서열(1314) 중의 유전자 융합(1312)으로서 유전자(1304 및 1308) 사이에 유전자 융합을 생성하도록 유전자 영역(1304)에 삽입되는 경우에 발생하는 전좌 이벤트를 반영하는 유전자형이 도시된다.

[0105] 거대한 게놈 구조 변이체(예: 거대한 역전 또는 전좌)를 검출하기 위한 현재의 방법은 변이체의 중단점(예를 들면, 전좌된 부분이 함께 융합되는 게놈 유전자좌)에 걸친 판독 쌍에 의존한다. 이러한 판독 쌍이 서열 분석 실험 동안 관찰되는 것을 보장하기 위해, 매우 깊은 서열 분석이 필요할 수 있다. 표적화 서열 분석(예: 엑솜 서

열 분석)에서, 중단점이 표적화 영역 내(예: 엑손 내)에 존재하지 않는 한 현재 서열 분석 기술로 구조적 변이체를 검출하는 것은 불가능한데, 이는 매우 불가능하다.

[0106] 그러나 본원에 기재된 바코드 방법 및 시스템에 의해 제공된 정보는 구조적 변이체를 검출하는 능력을 크게 향상시킬 수 있다. 직관적으로, 중단점의 왼쪽 및 오른쪽 유전자좌는 게놈 DNA의 공통 단편 상에 존재하는 경향이 있을 수 있고, 따라서 단일 구획 내에 유지되고, 따라서 공통 또는 공유된 바코드 서열로 바코드화될 수 있다. 전단의 확률적 특성으로 인해, 이러한 바코드의 공유는 서열이 중단점으로부터 더 멀어질수록 감소한다. 통계적 방법을 사용하여 두 게놈 유전자좌 사이에 바코드 중복이 우연히 예상되는 것보다 상당히 더 큰지의 여부를 결정할 수 있다. 이러한 중복은 중단점의 존재를 시사한다. 중요하게는, 바코드 정보는 이러한 정보가 이용가능한 경우, 전통적인 서열 분석에 의해 제공된 정보(예: 중단점에 걸친 판독치로부터의 정보)를 보충한다.

[0107] 본원에 기재된 방법의 맥락에서, 관련 유전자 영역을 포함하는 유기체로부터의 게놈 물질은, 상기한 바와 같이, 그것이 비교적 긴 단편을 포함하도록 단편화된다. 이는 도 13a에 비전좌 유전자형에 대해 도시한다. 도시된 바와 같이, 각각 유전자 영역(1302 및 1306)을 포함하는 두 개의 긴 개별적 제1 분자 단편(1316 및 1318)이 생성된다. 이러한 단편은 각각 구획(1320 및 1322)로 별도로 구획화되고, 제1 단편 각각은 구획 내에서 각각 다수의 제2 단편(1324 및 126)으로 단편화되고, 이 단편화 공정은 고유 식별자 태그 또는 바코드 서열을 소정의 구획 내의 제2 단편 모두에 공통인 제2 단편에 부착한다. 태그 또는 바코드는 각각의 구획(1320 및 1322)에 대해 각각 "1" 또는 "2"로 표시된다. 결과로서, 완전히 별개의 유전자(1304 및 1308)는 제2 단편의 상이하게 구획화되고 상이하게 바코드화된 그룹을 초래할 수 있다.

[0108] 바코드화되면, 제2 단편을 풀링시킨 다음, 제2 단편의 서열 뿐만 아니라 그 단편의 바코드 서열을 모두 제공할 수 있는 핵산 서열 분석 공정에 적용한다. 다음과 같은 특정 바코드의 존재에 기초한다: 예를 들면, 1 또는 2, 이어서, 제2 단편 서열은 바코드를 각 서열에 귀속시킴으로써 도시된 바와 같이, 특정 유래 서열, 예를 들면, 유전자(1304 또는 1308)에 귀속시킬 수 있다. 일부 경우에, 별개의 유래된 제1 단편 서열에 관한 바코드화 제2 단편 서열의 맵핑은 어떤 전좌도 발생하지 않았음을 결정하기에 충분히 확정적일 수 있다. 그러나, 일부 경우에, 조립된 서열(1330 및 1332)에 의해 도시된 바와 같이 유래하는 제1 단편 서열 모두 또는 일부에 대해 조립된 서열을 제공하기 위해 제2 단편 서열을 조립할 수 있다.

[0109] 도 13a에 도시된 비-전좌 유전자형 예와 대조적으로, 도 13b는 전좌 함유 유전자형에 적용된 동일한 공정의 개략도를 도시한다. 도시된 바와 같이, 제1의 긴 핵산 단편(1352)은 변이체 서열(1314)로부터 생성되고, 전좌 변이체, 예를 들면, 유전자 융합(1312)의 적어도 일부를 포함한다. 이어서, 제1 단편(1352)은 별개의 구획(1354)로 구획화된다. 구획(1354) 내에서, 제1 단편(1352)은 다시, 구획(1354)(바코드 "1"로서 나타냄) 내에 모든 제2 단편(1356)에 대해 동일한 고유 바코드를 포함하는 제2 단편(1356)으로 추가로 단편화된다. 상기한 바와 같이, 제2 단편의 풀링 및 서열 분석은 제2 단편의 기본 서열 뿐만 아니라 그들의 관련 바코드를 제공한다. 이어서, 이러한 바코드화 서열은 그들의 각각의 유전자 서열에 귀속시킬 수 있다. 그러나, 도시된 바와 같이, 두 유전자는 동일 바코드 서열을 포함하는 귀속된 제2 단편 서열을 반영하여 그들이 동일한 구획, 및 잠재적으로 유전자 융합을 나타내는 동일한 유래 분자로부터 유래됨을 나타낼 수 있다. 이는 또한 유전자 융합의 적어도 일부를 포함하지만, 상이한 바코드를 갖는 상이한 구획에서 처리되는 다수의 중복 제1 단편을 제공함으로써 추가로 검증될 수 있다.

[0110] 일부 경우에, 원래 분리된 유전자 각각에 귀속되는 복수의 상이한 바코드 (및 그들의 기본 단편 서열)의 존재는 유전자 융합 또는 다른 전좌 이벤트의 존재를 나타낼 수 있다. 일부 경우에, 참조 서열에 기초하여 분리된 것으로 예상된 두 유전자 영역으로 적어도 2개의 바코드, 적어도 3개의 상이한 바코드, 적어도 4개의 상이한 바코드, 적어도 5개의 상이한 바코드, 적어도 10개의 상이한 바코드, 적어도 20개의 상이한 바코드 또는 그 이상의 귀속은 그러한 영역을 근위에, 입접하게 배치하거나 또는 다르게는 서로 통합하는 전좌 이벤트의 표시를 제공할 수 있다. 일부 경우에, 구획화되는 단편의 크기는 변이체 결합을 동정할 수 있는 감도를 나타낼 수 있다. 특히, 소정의 액적 중의 단편이 길이 10 kb인 경우, 그 10 kb 크기 범위 내에 있는 결합이 검출될 것으로 예상된다.

[0111] 마찬가지로, 두 변이체 및 야생형 구조가 동일한 10kb 단편 내에 속하는 경우, 둘 다가 공통 또는 공유된 바코드를 통해 결합을 나타내기 때문에, 그 변이체의 동정이 더욱 어려울 것으로 예상된다. 이와 같이, 단편 크기 선택은 야생형이든 변이체이든 검출된 연결 서열의 상대적 근접성을 조정하기 위해 사용될 수 있다. 일반적으로, 그러나, 일반적으로, 100 베이스 이상, 500 베이스 이상, 1 kb 이상, 10 kb, 20 kb 이상, 30 kb 이상, 40 kb 이상, 50 kb 이상, 60 kb 이상, 70 kb 이상, 80 kb 이상, 90 kb 이상, 100 kb 이상, 200 kb 이상 또는 심지어 그 이상에 의해 분리되는 근위 서열을 초래하는 구조적 변이체는 본원에서 변이체 게놈 중의 그러

한 연결되지 않은 서열 세그먼트 사이의 결합을 동정함으로써 용이하게 동정될 수 있고, 이 결합은 공유된 또는 공통 바코드에 의해 및/또는, 언급된 바와 같이, 중지점에 걸친 서열 데이터에 의해 지시된다. 이러한 결합은 일반적으로 그러한 연결된 서열이 게놈 서열 내에서 50 kb 미만, 40 kb 미만, 30 kb 미만, 20 kb 미만, 10 kb 미만, 5 kb 미만, 4 kb 미만, 3 kb 미만, 2 kb 미만, 1 kb 미만, 500개 미만 염기, 200개 미만 염기 또는 심지어 더 미만으로 분리될 때 동정가능하다.

[0112] 일부 경우에, 서로 근위로 배치되거나 연결되는 두 서열을 초래하는 구조적 변형은, 그들이 일반적으로, 예를 들면, 10 kb 이상, 20 kb 이상, 30 kb 이상, 40 kb 이상 또는 50 kb 이상 또는 그 이상으로 분리되는 경우, 두 서열 부분에 공통인 바코드를 포함하는 맵핑가능한 바코드화 서열의 총 수의 비율에 의해 동정될 수 있다.

[0113] 알 수 있는 바와 같이, 일부 경우에, 본원에 기재된 공정은 특정 서열 거리 내에 있는 서열이 야생형으로이든 또는 변이체 서열로서이든 단일 구획 내에, 예를 들면, 단일 핵산 단편으로서 포함될 것임을 보장할 수 있다. 예를 들면, 공통 또는 중복 바코드 서열이 두 서열에 맵핑된 바코드의 전체 수의 1%보다 큰 경우, 이는 두 서열 세그먼트 사이에서와 같이, 특히 구조적 변형과 같은 정상적으로 연결되지 않는 두 서열 세그먼트 사이에서와 같이 결합을 동정하기 위해 사용될 수 있다. 일부 경우에, 공유된 또는 공통 바코드는 게놈 내의 구조적 변형을 구성하는 구조적 결합을 동정하기 위해 두 개의 정상적으로 분리된 서열에 전체 맵핑가능한 바코드의 2% 이상, 3% 이상, 4% 이상, 5% 이상, 6% 이상, 7% 이상, 8% 이상, 일부 경우에, 9% 이상 또는 심지어 10% 이상일 수 있다. 일부 경우에, 공유된 또는 공통 바코드는 구조적 변형을 갖지 않는 것으로 공지된 대조군 게놈보다 통계적으로 상당히 큰 비율 또는 수로 검출될 수 있다. 추가로, 제2 서열 단편이, 변이체 서열이, 예를 들면, 제2 단편(1358)에서와 같이, "정상" 서열, 또는 "중지점"을 충족시키는 지점에 걸치는 경우, 유전자 융합의 추가의 증거로서 이 정보를 사용할 수 있다.

[0114] 다시, 상기한 바와 같이, 조립된 서열(1360)로서 도시된 유전자 융합(1312)의 조립된 서열을 수득하기 위해, 제 2 단편 서열을 조립함으로써 유전자 융합(1312)의 구조를 추가로 설명할 수 있다.

[0115] 또한, 바코드 서열의 존재가 긴 기원 단편을 위한 서열로 짧은 서열의 조립을 가능하게 하는 반면, 이러한 긴 단편은 또한 상이한 중복성 기원의 긴 단편으로부터 조립된 중복성 긴 단편으로부터 긴 범위 서열 정보를 추정하게 한다. 이러한 생성되는 조립체는 유전자 융합(1312)의 긴 범위 서열 수준 동정 및 특성화를 가능하게 한다.

[0116] 일부 경우에, 상기한 방법은 레트로트랜스포존의 존재를 동정하는데 유용하다. 레트로트랜스포존은 전사에 이어, 스플라이싱된 메신저 RNA (mRNA)의 역전사 및 게놈 중의 새로운 위치에서의 삽입에 의해 생성될 수 있다. 따라서, 이러한 구조적 변이체는 인트론이 부족하고, 종종 염색체간에 존재하지만, 다르게는 다양한 특징을 갖는다. 레트로트랜스포존이 유전자의 기능적 카피를 도입할 때, 그들은 레트로유전자로서 칭명되고, 이는 인간 및 *드로소필라(Drosophila)* 게놈에 보고되어 있다. 다른 경우, 레트로카피는 전체 전사물, 특이적 전사 아이소폼 또는 불완전한 전사물을 함유할 수 있다. 또한, 대안의 전사 개시 부위 및 프로모터 서열은 때로는 전사물 내에 존재하고, 따라서 레트로트랜스포존은 때로는 프로모터 서열을 다운스트림 서열의 발현을 구동할 수 있는 게놈의 재삽입된 영역 내에 도입한다.

[0117] 탠덤 중복과는 달리, 레트로트랜스포존은 엑손 또는 인트론 내의 부계 유전자로부터 멀리 삽입한다. 유전자 근처에 삽입되면, 레트로트랜스포존은 발현을 위해 국소 조절 서열을 이용할 수 있다. 유전자 근처에 삽입은 또한 수용 유전자를 불활성화하거나 새로운 키메라 전사물을 생성할 수 있다. 레트로트랜스포존 매개 키메라 유전자 전사물은 인간 샘플의 RNA-Seq 데이터에서 보고되었다.

[0118] 레트로트랜스포존의 중요성에도 불구하고, 그들의 검출은 메이트 쌍 라이브러리로부터 쌍을 이룬판독 지원, 전체 게놈 서열 분석(WGS)에서 엑손-엑손 접합부 발견 또는 레트로트랜스포존 키메라의 RNA-Seq 인식에 의존하는 지시된 접근법에 제한될 수 있다. 이러한 방법 모두는 분석을 복잡하게 하는 위양성(false positive)을 가질 수 있다.

[0119] 레트로트랜스포존은 본원에 기재된 시스템 및 방법을 사용하여 전체 게놈 라이브러리로부터 동정될 수 있고, 이들의 삽입 부위는 상기 논의된 바코드 맵핑을 사용하여 맵핑할 수 있다. 예를 들면, Ceph NA12878 게놈은 SKA3-DDX10 키메라 레트로트랜스포존을 갖는다. SKA3 인트론 없는 전사물은 DDX10의 엑손 10 및 11 사이에 삽입된다. 또한, CBX3-C15ORF17 레트로트랜스포존은 또한 본원에 기재된 방법을 사용하여 NA12878에서 검출될 수 있다. CBX3의 아이소폼 2는 C15ORF17의 엑손 2와 3 사이에 삽입된다. 이러한 키메라 전사물은 HapMap 프로젝트로부터의 유럽 RNA-Seq 샘플의 20%에서 관찰되었다(참조: D.R. Schrider 등 PLoS Genetics 2013).

- [0120] 레트로트랜스포존은 또한 본원에 기재된 방법 및 시스템을 사용하여 제조된 전체 엑솜 라이브러리에서 검출될 수 있다. 레트로트랜스포존은 엑솜 표적화로 용이하게 농축되지만, 인트론이 포획 동안 제거되기 때문에 전좌 이벤트와 레트로트랜스포존 사이를 구별하는 것이 어렵거나 불가능할 수 있다. 그러나, 본원에 기재된 시스템 및 방법을 사용하여, 의심되는 레트로트랜스포존에 대한 인트론바이트(intronic bait)를 도입함으로써 전체 엑손 서열 분석(WES) 라이브러리에서 레트로트랜스포존을 동정할 수 있다(참조: 미국 가출원 제 62/072,164호 (2014년 10월 29일에 출원), 모든 목적을 위해 그 전체가 본원에 참조로 인용된다). 인트론 신호의 부족은 레트로트랜스포존 구조적 변이체를 나타낼 수 있는 반면, 인트론 신호는 전좌를 나타낼 수 있다.
- [0121] 알 수 있는 바와 같이, 상기한 변형을 동정하고 특성화하는데 긴 범위 서열 맥락을 사용하는 능력은 바코드를 변형 내의 영역에 맵핑하고/하거나 변형을 스페닝함으로써 삽입, 결실, 레트로트랜스포존, 역전 등을 포함하는 다른 구조적 변형의 범위를 동정하는데 동등하게 적용가능하다.
- [0122] **V. 카피 수 변형으로 인한 질환 & 장애**
- [0123] 본 방법 및 시스템은 광범위한 범위의 질환 및 장애를 진단하고/하거나 검출하는 매우 정확하고민감한 접근법을 제공한다. 카피 수 변형과 관련된 질환은, 예를 들면, 디조지/입천장심장얼굴 증후군(22q11.2 결실), 프라더-윌리 증후군(15q11-q13 결실), 윌리엄스-보이렌 증후군(7q11.23 결실), 밀러-디커 증후군(MDLS)(17p13.3 미세 결실), 스미스-마제니스 증후군(SMS)(17p11.2 미세 결실), 신경섬유종증 제1형(NF1)(17q11.2 미세 결실), 펠란-맥에르미드 증후군(22q13 결실), 레트 증후군(염색체 Xq28 상의 MECP2 중 기능 손실 돌연변이), 메르츠바하병(PLP1의 CNV), 척수성 근위축증(SMA)(염색체 5q13 상의 텔로머릭 SMN1의 동형 접합 부재), 포톡키-낭창키 증후군(PTLS, 염색체 17p.11.2의 중복)을 포함할 수 있다. PMP22 유전자의 추가의 카피는 샤르코-마리-투스 신경병증 유형 IA(CMT1A) 및 압박 마비에 대한 책임이 있는 유전성 신경병증(HNPP)과 관련될 수 있다. 질환은 다음 문헌에 기재된 질환일 수 있다: Lupski J. (2007) Nature Genetics 39: S43-S47.
- [0124] 본원에 제공된 방법 및 시스템은 또한 광범위한 범위의 태아 이수성을 정확하게 검출하거나 진단할 수 있다. 종종, 본원에 제공된 방법은 샘플 내의 태아 핵산을 평가하기 위해 임신부로부터 취한 샘플(예: 혈액 샘플)을 분석함을 포함한다. 태아 이수성은, 예를 들면, 삼염색체성 13(파타우 증후군), 삼염색체성 18(에드워드 증후군), 삼염색체성 21(다운 증후군), 클라인펠터 증후군(XXY), 하나 이상의 염색체의 일염색체성(X 염색체 일염색체성, 터너 증후군), 삼염색체성 X, 하나 이상의 염색체의 삼염색체성, 다음과 같은 하나 이상의 염색체의 사염색체성 또는 오염색체성을 포함할 수 있다: (예: XXXX, XXYY, XXXY, XYYY, XXXX, XXXY, XXXY, XYYY 및 XYYY), 삼배체성 (모든 염색체 중 3개, 예를 들면, 인간의 69 염색체), 사배체성 (모든 염색체 중 4개, 예를 들면, 인간의 92 염색체), 및 배수성. 일부 구현예에서, 이수성은 세그먼트 이수성일 수 있다. 세그먼트 이수성은 다음을 포함할 수 있다: 예를 들면, 1p36 중복, dup(17)(p11.2p11.2) 증후군, 다운 증후군, 펠리제우스-메르츠바하병, dup(22)(q11.2q11.2) 증후군, 및 묘안 증후군. 일부 경우에, 비정상적 유전형, 예를 들면, 태아 유전형은 묘성 증후군(Cri-du-chat syndrome), 윌프-허쉬호른, 윌리엄스-보이렌 증후군, 샤르코-마리-투스 질환, 압박 마비에 책임이 있는 유전성 신경병증, 스미스-마제니스 증후군, 신경섬유종증, 알라질 증후군, 입천장심장얼굴 증후군, 디조지 증후군, 스테로이드 설파타제 결핍, 칼만 증후군, 선형 피부 결함을 갖는 소안구증, 부신 저형성증, 글리세롤 키나제 결핍, 펠리제우스-메르츠바하병, Y 상의 고환-결정 인자, 무정자증(인자 a), 무정자증(인자 b), 무정자증(인자 c), 또는 1p36 결실과 같은 상태를 초래할 수 있는 성 또는 삼염색체성 염색체의 하나 이상의 결실에 기인한다. 일부 구현예에서, 염색체 수의 감소는 XO 증후군을 초래한다
- [0125] 과도한 게놈 DNA 카피 수 변형은 또한 리-프라우메니 암 소인 증후군과 관련된다(참조: Shlien 등 (2008) PNAS 105:11264-9). CNV는 CHARGE (결손증, 심장 이상, 후비공 폐쇄증, 지체, 생식기, 및 귀 기형), 피터스-폴러스, 피트-홉킨스 및 혈소판 감소증 부재 반경 증후군을 포함하는 기형 증후군과 관련된다(참조: 예를 들면, Ropers HH (2007) Am J of Hum Genetics 81: 199-207). 카피 수 변형 및 암 사이의 관계는, 예를 들면, 다음 문헌에 기재되어 있다: Shlien A. 및 Malkin D. (2009) Genome Med. 1(6): 62. 카피 수 변형은, 예를 들면, 자폐증, 정신 분열증 및 특발성 학습 장애와 관련된다. 예를 들면, 다음을 참조한다: Sebat J., 등 (2007) Science 316: 445-9; Pinto J. 등
- [0126] 본원에 기재된 바와 같이, 본원에 제공된 방법 및 시스템은 또한 상이한 종류의 암과 관련된 CNV를 검출하는데 유용하다. 예를 들면, 방법 및 시스템은 비-소세포 폐암에서 증가될 수 있는 EGFR 카피 수를 검출하는데 사용될 수 있다.
- [0127] 본원에 제공된 방법 및 시스템은 또한 병원체(예: 바이러스, 세균, 미생물, 진균 등)로부터의 감염에 대한 민감성을 포함하는, 특정 질환 또는 장애에 대한 대상체의 민감성 수준을 결정하기 위해 사용될 수 있다. 예를

들면, 상기 방법은 CCL3L1의 카피 수를 분석함으로써 HIV 감염에 대한 대상체의 민감성을 결정하는데 사용될 수 있고, 단 비교적 높은 수준의 CCL3L1은 HIV 감염에 대한 낮은 대상체의 낮은 민감성에 관련된다(참조: Gonzalez E. 등 (2005) Science 307: 1434-1440). 또 다른 예에서, 상기 방법은 전신 홍반성 낭창에 대한 대상체의 민감성을 결정하기 위해 사용될 수 있다. 이러한 경우에, 예를 들면, 상기 방법은, 이 분자의 낮은 카피 수가 전신 홍반성 낭창에 대한 증가된 민감성과 관련되기 때문에, FCGR3B(CD16 세포 표면 면역글로불린 수용체)의 카피 수를 검출하기 위해 사용될 수 있다(참조: Aitman T.J. 등 (2006) Nature 439: 851-855). 본원에 제공된 방법 및 시스템은 또한 다른 질환 또는 장애와 관련된 CNV, 예를 들면, 자폐증, 정신 분열증 또는 특발성 학습 장애와 관련된 CNV를 검출하는데 사용될 수 있다(참조: Kinght 등, (1999) The Lancet 354 (9191): 1676-81.). 유사하게, 상기 방법 및 시스템은 염색체 4p16에서 카피-수-변형가능한 영역의 5개 탠덤 카피에 연결된 상염색체성-우성 소이증을 검출하는데 사용될 수 있다(참조: Balikova I. (2008) Am J. Hum Genet. 82: 181-187).

[0128] VI. 질환 및 장애의 검출, 진단 및 치료

[0129] 본원에 제공된 방법 및 시스템은 또한 질환 또는 장애의 검출, 진단 및 치료를 보조할 수 있다. 일부 경우에, 본원에 기재된 시스템 또는 방법을 사용하여 질환 또는 장애를 검출하고, 질환의 검출에 기초하여 치료를 대상체에게 추가로 제공함을 포함한다. 예를 들면, 암이 검출되는 경우, 대상체는 외과적 개입, 이러한 암을 치료하도록 설계된 약물을 투여하고, 호르몬 요법을 제공하고/하거나 방사선 또는 더 일반화된 화학요법을 투여함으로써 치료될 수 있다.

[0130] 종종, 상기 방법 및 시스템은 또한 감별 진단을 가능하게 하고, 환자를 표적 요법으로 치료함을 추가로 포함할 수 있다. 일반적으로, 질환 또는 장애 (또는 이의 부재)의 감별 진단은 질환 또는 장애를 갖는 것으로 의심 받는 대상체로부터 취득된 샘플 핵산의 서열을 결정하여 특성화하고, 이를 장애 또는 질환 상태의 존재 (또는 부재)를 나타내는 참조 핵산의 서열 및/또는 서열 특성화와 비교함으로써 샘플 핵산을 장애 또는 질환 상태 (또는 이의 부재)의 지표로서 추가로 특성화함으로써 달성될 수 있다.

[0131] 참조 핵산 서열은 질환 또는 장애 상태를 나타내는 게놈(예: 생식 세포 핵산)으로부터 유도될 수 있거나 질환 또는 장애 상태를 나타내는 게놈(예: 암 핵산, 이수성을 나타내는 핵산 등)으로부터 유도될 수 있다. 또한, (예를 들면, 1 kb 이상, 5 kb 이상, 10 kb 이상, 15 kb 이상, 20 kb 이상, 30 kb 이상, 40 kb 이상, 50 kb 이상, 60 kb 이상, 70 kb 이상, 80 kb 이상, 90 kb 이상 또는 심지어 100 kb 이상의 길이를 갖는) 참조 핵산 서열은 특정 서열의 존재 (또는 부재)를 결정하는 단계, 특정 일배체형의 존재 (또는 부재)를 결정하는 단계, 하나 이상의 유전적 변형(예: 구조적 변형(예: 카피 수 변형, 삽입, 결실, 전좌, 역전, 레트로트랜스포존, 재배열, 반복 확장, 중복 등), 단일 뉴클레오티드 다형성(SNP) 등)의 존재 (또는 부재)를 결정하는 단계 및 이의 조합을 포함하는 비제한적 예로 하나 이상의 국면에서 특성화될 수 있다. 또한, 참조 서열의 임의의 적합한 형태 및 서열 특성의 수가 샘플 핵산의 서열을 특성화하는데 사용될 수 있다. 예를 들면, 참조 핵산 서열의 하나 이상의 유전적 변형 (또는 이의 결여) 또는 구조적 변형 (또는 이의 결여)는 장애 또는 질환 상태의 존재 (또는 부재)의 지표로서 참조 핵산을 동정하는 서열 서명으로서 사용될 수 있다. 이용된 참조 핵산 서열의 특성화에 기초하여, 샘플 핵산 서열은 유사한 방식으로 특성화될 수 있고, 그것이 참조 핵산 서열과 유사한 특성을 나타내는지의 여부에 기초하여 장애 또는 질환을 나타내는 핵산으로부터 유래되는 (또는 유래되지 않는) 것으로서 추가로 특성화/동정될 수 있다. 일부 경우에, 샘플 핵산 서열 및/또는 참조 핵산 서열의 특성화 및 이들의 비교는 프로그램된 컴퓨터 프로세서의 보조로 완성될 수 있다. 일부 경우에, 이러한 프로그램된 컴퓨터 프로세서는 본원의 다른 곳에서 기재된 예시적인 컴퓨터제어 시스템과 같은 컴퓨터 제어 시스템에 포함될 수 있다.

[0132] 샘플 핵산은 본원의 다른 곳에 기재된 샘플 공급원 및 생물학적 샘플 공급원을 포함하는 임의의 적합한 공급원로부터 취득될 수 있다. 일부 경우에, 샘플 핵산은 무세포 핵산을 포함할 수 있다. 일부 경우에, 샘플 핵산은 종양 핵산(예: 종양 DNA)을 포함할 수 있다. 일부 경우에, 샘플 핵산은 순환성 종양 핵산(예: 순환성 종양 DNA(ctDNA))을 포함할 수 있다. 순환성 종양 핵산은 순환성 종양 세포(CTC)로부터 유래될 수 있고/있거나, 예를 들면, 대상체의 혈액, 혈장, 다른 체액 또는 조직으로부터 취득될 수 있다.

[0133] 도 20-21은 질환 검출 및 진단 맥락에서 샘플 핵산을 특성화하기 위한 예시적 방법을 도시한다. 도 20은 긴 범위 서열 맥락이 참조 핵산(예: 생식 세포 핵산(예: 생식 세포 게놈 DNA), 특정 장애 또는 질환 상태와 관련된 핵산)에 대해 짧은 바코드화 단편으로부터, 예를 들면, 도 6에 도시된 것과 유사한 방식으로 결정될 수 있는 예시적 방법을 입증한다. 도 20과 관련하여, 참조 핵산이 취득될 수 있고(2000), 바코드화 비드의 세트도 또한 취득될 수 있다(2010). 비드는 하나 이상의 바코드 서열 뿐만 아니라 랜덤 N-량체 또는 다른 프라이머와 같은 프라이머를 함유하는 올리고뉴클레오티드에 연결될 수 있다. 일부 경우에, 바코드 서열은 바코드화 비드로부터,

예를 들면, 바코드와 비드 사이의 결합의 절단을 통해 또는 바코드를 박리시키는 기초 비드의 분해를 통해 또는 둘의 조합으로 박리가 가능하다. 예를 들면, 일부 국면에서, 바코드화 비드는 바코드 서열을 박리시키는 환원제와 같은 제제에 의해 분해 또는 용해될 수 있다. 이러한 예에서, 참조 핵산(2005), 바코드화 비드(2015), 및, 일부 경우에, 다른 시약, 예를 들면, 환원제(2020)가 배합되고, 구획화에 적용된다. 일부 경우에, 참조 핵산(2000)은 구획화 전에 단편화될 수 있고, 생성되는 단편의 적어도 일부는 바코딩을 위해 2005로서 구획화된다. 예로써, 이러한 구획화 단계는 성분을 액적 생성 시스템, 예를 들면, 미세 유체 장치(2025)에 도입하는 단계를 포함할 수 있다. 미세 유체 장치(2025)의 보조로, 유중수 에멀전(2030)이 형성될 수 있고, 여기서 상기 에멀전은 참조 핵산(2005), 환원제(2020) 및 바코드화 비드(2015)를 함유하는 수성 액적을 함유한다. 환원제는 바코드화 비드를 용해 또는 분해시켜 액적 내의 비드로부터 바코드 및 랜덤 N-량체를 갖는 올리고뉴클레오타이드를 박리시킬 수 있다(2035). 이어서, 랜덤 N-량체는 참조 핵산의 상이한 영역을 프라이밍하여 증폭 후 참조 핵산의 증폭된 카피를 초래하고, 여기서 각 카피는 바코드 서열로 태깅된다(2040) 일부 경우에, 증폭(2040)은 본원의 다른 곳에서 기재되고 도 5에 개략적으로 도시된 것과 유사한 방법에 의해 달성될 수 있다. 일부 경우에, 각 액적은 동일한 바코드 서열 및 상이한 랜덤 N-량체 서열을 함유하는 올리고뉴클레오타이드의 세트를 함유한다. 이어서, 에멀전은 파괴되고(2045), 추가의 서열(예: 특히 서열 분석 방법, 추가의 바코드 등을 보조하는 서열)이, 예를 들면, 증폭 방법을 통해 첨가될 수 있다(2050)(예: PCR). 이어서, 서열 분석이 수행될 수 있고(2055), 서열 분석 데이터를 해석하기 위해 알고리즘을 적용한다(2060). 일부 경우에, 서열 분석 데이터의 해석(2060)은 참조 핵산의 적어도 일부에 대한 서열을 제공하는 단계를 포함할 수 있다. 일부 경우에, 참조 핵산에 대한 긴 범위 서열 맥락이 수득되고, 예를 들면, 참조 핵산이 질환 상태(예: 본원의 다른 곳에 기재된 바와 같은 하나 이상의 일배체형의 결정, 하나 이상의 구조적 변형(예: 카피 수 변형, 삽입, 결실, 전좌, 역전, 재배열, 반복 확대, 중복, 레트로트랜스포존, 유전자 융합 등)의 결정, 하나 이상의 SNP의 호출 등)로부터 유도되는 경우와 같이 특성화된다. 일부 경우에, 변이체는 도 7과 관련하여 본원의 다른 곳에 기재된 것과 같이, 공급원 및 보다 긴 범위 서열 맥락을 제공하기 위해 생성된 추정 콘텍스트로부터 수득된 다양한 참조 핵산에 요구될 수 있다.

[0134]

도 21은 도 20에 도시된 바와 같이 수득된 참조 2060 특성화로부터 샘플 핵산 서열을 특성화하는 예를 입증한다. 긴 범위 서열 맥락은, 예를 들면, 도 6에 개략적으로 도시된 방법을 통해서와 같은, 본원의 다른 곳에 기재된 바와 같이 더 짧은 바코드화 단편의 서열 분석으로부터 수득될 수 있다. 도 21에 도시된 바와 같이, 핵산 샘플(예: 순환성 종양 핵산을 포함하는 샘플)은 장애 또는 질환(예: 암)을 갖는 것으로 의심되는 대상체로부터 수득될 수 있고(2100), 바코드화 비드의 세트도 또한 수득될 수 있다(2110). 비드는 하나 이상의 바코드 서열 뿐만 아니라 랜덤 N-량체 또는 다른 프라이머와 같은 프라이머를 함유하는 올리고뉴클레오타이드에 연결될 수 있다. 일부 경우에, 바코드 서열은 바코드화 비드로부터, 예를 들면, 바코드와 비드 사이의 결합의 절단을 통해 또는 바코드를 박리시키는 기초 비드의 분해를 통해 또는 둘의 조합으로 박리가 가능하다. 예를 들면, 일부 국면에서, 바코드화 비드는 바코드 서열을 박리시키는 환원제와 같은 제제에 의해 분해 또는 용해될 수 있다. 이러한 예에서, 샘플 핵산(2105), 바코드화 비드(2115), 및 일부 경우에, 다른 제제, 예를 들면, 환원제(2120)가 배합되고, 구획화에 적용된다. 일부 경우에, 태아 샘플(2100)은 구획화 전에 단편화되고, 생성되는 단편의 적어도 일부는 바코딩을 위해 2105로서 구획화된다. 예로써, 이러한 구획화는 성분을 액적 생성 시스템, 예를 들면, 미세 유체 장치(2125)에 도입하는 단계를 포함할 수 있다. 미세 유체 장치(2125)의 보조로, 유중수 에멀전(2130)이 형성될 수 있고, 여기서, 상기 에멀전은 샘플 핵산(2105), 환원제(2120) 및 바코드화 비드(2115)를 함유하는 수성 액적을 함유한다. 환원제는 바코드화 비드를 용해 또는 분해하여 액적 내의 비드로부터 바코드 및 랜덤 N-량체를 갖는 올리고뉴클레오타이드를 박리시킬 수 있다(2135). 이어서, 랜덤 N-량체는 샘플 핵산의 상이한 영역을 프라이밍하여 증폭 후 샘플 핵산의 증폭된 카피를 초래할 수 있고, 여기서 각 카피는 바코드 서열로 태깅된다(2140). 일부 경우에, 증폭(2140)은 본원의 다른 곳에서 기재되고 도 5에 개략적으로 도시된 것과 유사한 방법에 의해 달성될 수 있다. 일부 경우에, 각 액적은 동일한 바코드 서열 및 상이한 랜덤 N-량체 서열을 함유하는 올리고뉴클레오타이드의 세트를 함유한다. 이어서, 에멀전은 파괴되고(2145), 추가의 서열(예: 특히 서열 분석 방법, 추가의 바코드 등을 보조하는 서열)이, 예를 들면, 증폭 방법을 통해 첨가될 수 있다(2150)(예: PCR). 이어서, 서열 분석을 수행할 수 있고(2155), 서열 분석 데이터를 해석하기 위해 알고리즘을 적용한다(2160). 일부 경우에, 서열 분석 데이터의 해석(2160)은 샘플 핵산의 서열을 제공하는 단계를 포함할 수 있다. 일부 경우에, 핵산 샘플을 위한 긴 범위 서열 맥락이 수득된다. 샘플 핵산 서열은 참조 핵산 서열의 특성화(2060)를 사용하여 특성화될 수 있다(2160)(예: 본원의 다른 곳에 기재된 바와 같은 하나 이상의 일배체형의 결정, 하나 이상의 구조적 변형(예: 카피 수 변형, 삽입, 결실, 전좌, 역전, 재배열, 반복 확대, 중복, 레트로트랜스포존, 유전자 융합 등). 샘플 핵산 서열 및 이의 특성화와 참조 핵산의 서열 및 특성화의 비교에 기초하여, 장애 또는 질환 상태의 존재 (또는 부재)에 관한 감별 진단(2170)이 수행될 수 있다.

- [0135] 알 수 있는 바와 같이, 참조 핵산 및 샘플 핵산의 분석은 별도의 구획화 분석으로서 완료될 수 있거나, 단일 구획화 분석의 일부로서 완료될 수 있다. 예를 들면, 샘플 및 참조 핵산은 동일한 장치에 첨가할 수 있고, 바코드화 샘플 및 참조 단편은 도 20 및 21에 따르는 액적으로 생성되고, 여기서 에멀전은 두 종류의 핵산을 위한 액적을 포함한다. 이어서, 에멀전은 파괴시키고, 액적의 내용물은 풀링하고, 추가로 처리하고(예: PCR을 통해 추가의 서열의 벌크 첨가), 본원에서 다른 곳에 기재된 바와 같이 서열 분석했다. 바코드화 단편으로부터 개별 서열 분석 판독치는 바코드 서열을 통해 그들 각각의 샘플 서열에 귀속될 수 있다. 샘플 핵산으로부터 수득된 서열은 참조 핵산 서열의 특성화에 기초하여 특성화될 수 있다.
- [0136] 본원의 방법 및 시스템을 이용하면 본원에 기재된 바와 같은 참조 및 샘플 핵산 서열의 긴-범위서열 맥락을 포함하여 핵산의 긴 범위 핵산 서열을 결정하는데 있어서 정확성을 향상시킬 수 있다. 본원에 제공된 방법 및 시스템은 적어도 70%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 99%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9%, 99.95%, 99.99%, 99.995% 또는 99.999%의 정확성으로 참조 및/또는 샘플 핵산의 긴-범위 서열 맥락을 결정할 수 있다. 일부 경우에, 본원에 제공된 방법 및 시스템은 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.1%, 0.05%, 0.01%, 0.005%, 0.001%, 0.0005%, 0.0001%, 0.00005%, 0.00001% 또는 0.000005% 미만의 오차율로 참조 및/또는 샘플 핵산의 긴-범위 서열 맥락을 결정할 수 있다.
- [0137] 또한, 본원의 방법 및 시스템은 하나 이상의 국면(예: 서열의 결정, 하나 이상의 유전적 변형의 결정 일배체형의 결정 등)에서 참조 핵산 서열 및/또는 샘플 핵산 서열을 특성화하는데 있어서 정확성을 또한 향상시킬 수 있다. 따라서, 본원에 제공된 방법 및 시스템은 하나 이상의 국면에서 적어도 70%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 99%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9%, 99.95%, 99.99%, 99.995% 또는 99.999%의 정확성으로 참조 핵산 서열 및/또는 샘플 핵산 서열을 특성화할 수 있다. 일부 경우에, 본원에 제공된 방법 및 시스템은 하나 이상의 국면에서 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.1%, 0.05%, 0.01%, 0.005%, 0.001%, 0.0005%, 0.0001%, 0.00005%, 0.00001% 또는 0.000005% 미만의 오차율로 참조 핵산 서열 및/또는 샘플 핵산 서열을 특성화할 수 있다.
- [0138] 또한, 상기 논의된 바와 같이, 참조 핵산의 긴-범위 서열 맥락을 결정하고 이들의 특성화하는데 있어서 향상된 정확성은 샘플 핵산의 서열 분석 및 특성화 및 장애 또는 질환의 감별 진단에서의 후속적 사용에 있어서 향상된 정확성을 초래할 수 있다. 따라서, 샘플 핵산 서열(긴-범위 서열 맥락 포함)은 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.1%, 0.05%, 0.01%, 0.005%, 0.001%, 0.0005%, 0.0001%, 0.00005%, 0.00001% 또는 0.000005% 미만의 오차율로 참조 핵산 서열의 분석으로부터 제공될 수 있다. 일부 경우에, 샘플 핵산 서열은 적어도 70%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 99%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9%, 99.95%, 99.99%, 99.995% 또는 99.999%의 정확성으로 참조 핵산의 서열 및/또는 서열의 특성화와의 비교로 장애 또는 질환 (또는 이의 부재)의 감별 진단용으로 사용될 수 있다. 일부 경우에, 샘플 핵산 서열은 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.1%, 0.05%, 0.01%, 0.005%, 0.001%, 0.0005%, 0.0001%, 0.00005%, 0.00001% 또는 0.000005% 미만의 오차율로 참조 핵산의 서열 및/또는 서열의 특성화와의 비교로 장애 또는 질환 (또는 이의 부재)의 감별 진단용으로 사용될 수 있다.
- [0139] 일례에서, 방법 및 시스템은 폐암이 EGFR 유전자의 변형과 관련되는 비-소세포 폐암인지의 여부를 결정하기 위해 폐암 환자의 카피 수 변형을 검출하는데 사용될 수 있다. 이러한 진단 후, 환자의 치료 선택은 감별 진단과 관련시키기 위해 개선시킬 수 있다. 표적 요법 또는 분자 표적 요법은 암의 치료 (약물 요법)의 주요 양식 중의 하나이고, 다른 것은 호르몬 요법 및 세포독성 화학요법이다. 표적 요법은 단순히 (예를 들면, 전통적인 화학요법으로) 모두 빠르게 분열하는 세포를 방해하기보다는 암 발생 및 종양 성장에 필요한 특정 표적 분자를 방해하여 암 세포의 성장을 차단한다.
- [0140] 도 14는 비-소세포 폐암을 감별 진단하는 예시적인 공정을 도시한다. 만성 기침, 체중 감소 및 호흡 곤란이 있는 환자는 폐암에 대해 시험한다(1400). 혈액은 환자로부터 채취하고(1405), 샘플(예: 순환성 종양 세포, 무세포 DNA, 순환성 핵산(예: 순환성 종양 핵산) 등)은 혈액으로부터 유래된다(1410). 바코드화 비드의 세트도 또한 수득될 수 있다(1415). 비드는 하나 이상의 바코드 서열 뿐만 아니라 랜덤 N-량체 또는 다른 프라이머와 같은 프라이머를 함유하는 올리고뉴클레오타이드에 연결될 수 있다. 일부 경우에, 바코드 서열은 바코드화 비드로부터, 예를 들면, 바코드와 비드 사이의 결합의 절단을 통해 또는 바코드를 박리시키는 기초 비드의 분해를 통해 또는 둘의 조합으로 박리가능하다. 예를 들면, 일부 국면에서, 바코드화 비드는 바코드 서열을 박리시키는 환원제와 같은 제제에 의해 분해 또는 용해될 수 있다. 이러한 예에서, 샘플(1410), 바코드화 비드(1420), 및, 일부 경우에, 다른 시약, 예를 들면, 환원제가 배합되고, 구획화에 적용된다. 예로써, 이러한 구획화는 성분을 액적 생성 시스템, 예를 들면, 미세 유체 장치(1425)에 도입하는 단계를 포함할 수 있다. 미세 유체 장치(1425)의 보조로,

유중수 에멀전(1430)이 형성될 수 있고, 여기서 에멀전은 샘플 핵산(1410), 바코드화 비드(1415), 및, 일부 경우에, 환원제를 함유하는 수성 액적을 함유한다. 환원제는 바코드화 비드를 용해 또는 분해시켜 액적 내의 비드로부터 바코드 및 랜덤 N-량체를 갖는 올리고뉴클레오타이드를 박리시킬 수 있다(1435). 이어서, 랜덤 N-량체는 샘플 핵산의 상이한 영역을 프라이밍하여 증폭 후 샘플의 증폭된 카피를 초래할 수 있고, 여기서 각 카피는 바코드화 서열로 태깅된다(1440). 일부 경우에, 각 액적은 동일한 바코드 서열 및 상이한 랜덤 N-량체 서열을 함유하는 올리고뉴클레오타이드의 세트를 함유한다. 이어서, 에멀전은 파괴되고(1445), 추가의 서열(예: 특히 서열 분석 방법, 추가의 바코드 등을 보조하는 서열)이, 예를 들면, 증폭 방법을 통해 첨가될 수 있다(예: PCR). 이어서, 서열 분석을 수행할 수 있고(1450), 서열 분석 데이터를 해석하기 위해 알고리즘을 적용한다(1455). 서열 분석 알고리즘은 일반적으로, 예를 들면, 서열 분석 판독치를 정렬시키고/시키거나 특정 서열 판독치가 속하는 샘플을 동정하기 위해 바코드의 분석을 실행할 수 있다.

[0141] 이어서, 분석된 서열은 상이한 유전자의 CNV를 결정하기 위해 공지된 게놈 참조 서열과 비교한다(1460). DNA 중의 EGFR 카피 수가 정상보다 높을 경우, 환자는 소세포 폐암 대신 비-소세포 폐암(NSCLC)을 감별 진단될 수 있다(1465). 비-소세포 폐암의 CTC는 또한 이를 소세포 폐암과 추가로 구별할 수 있는 다른 카피 수 변형을 가질 수 있다. 암의 단계에 따라, 수술, 화학요법 또는 방사선 요법이 처방된다(1470). 일부 경우에, NSCLC가 진단된 환자는 이러한 암을 위해 표적화된 약물, 예를 들면, ALK 억제제를 투여한다(예: 크리조티닙). EGFR 변형의 일부 경우에, 환자는 세툽시맙, 파니투무맙, 라파티닙 및/또는 카페시타빈을 투여한다. 상이한 상황에서, 표적은 상이한 유전자, 예를 들면, ERBB2일 수 있고, 요법은 트라스투주맙(헤르셉틴)을 포함한다. (참조: (2010) Nature 466: 368-72; Cook E.H. 및 Scherer S.W. (2008) Nature 455: 919-923).

[0142] 표적 요법의 주요 카테고리는 소분자, 소분자 약물 집합체 및 모노클로날 항체이다. 소분자는 티로신-키나제, 예를 들면, 이마티닙 메실레이트(글리벡, 또한 STI-571로서 공지됨)(이는 만성 골수성 백혈병, 위장관 간질 종양 및 임부 다른 유형의 암에 승인된다); 게피티닙(이레사, 또한 ZD1839로서 공지됨)(이는 표피 성장 인자 수용체 (EGFR) 티로신 키나제를 표적화하고, 비소세포 폐암용으로 미국에서 승인된다); 에를로티닙(타세바로서 시판됨); 보르테오미브(벨카데)(이는 단백질을 방해하여 암 세포가 세포사를 겪게 하는 아포토시스-유발 프로테아솜 억제제 약물이다); 타목시펜; JAK 억제제(예: 토파티닙), ALK 억제제(예: 크리조티닙)를 포함할 수 있다; Bcl-2 억제제(예: 임상 시험 중인 오바토클락, ABT-263 및 고시폴); PARP 억제제(예: 이니파리브, 임상 시험 중인 올라파리브); PI3K 억제제(예: 위상 III 시험 중 페리포신). 아파티닙(이는 선택적 VEGF 수용체 2 억제제); AN-152, [D-Lys(6)]-LHRH에 연결된 (AEZS-108) 독소루비신; Braf 억제제(베무라페닙, 다브라페닙, LGX818)(BRAF V600E 돌연변이를 품는 전이성 흑색종을 치료하기 위해 사용됨); MEK 억제제(트라메티닙, MEK162); 다음과 같은 CDK 억제제: 예를 들면, PD-0332991, 임상 시험 중인 LEE011; Hsp90 억제제; 및 살리노마이신.

[0143] 다른 요법은 엽산 수용체를 표적화하는 소분자로 이루어진 소분자 약물 집합체인 빈타폴리드와 같은 소분자 약물 집합체를 포함한다.

[0144] 모노클로날 항체는 본원에 제공된 방법의 일부로서 투여될 수 있는 요법의 또 다른 형태이다. 모노클로날 항체도 또한 투여될 수 있다. 예시적인 모노클로날 항체를 다음을 포함한다: 리툭시맙(맵테라 또는 리툭산으로 시판됨)(이는 B 세포 상에서 발견된 CD20을 표적화하고 비호지킨 림프종을 표적화한다); 트라스투주맙(헤르셉틴)(이는 일부 형태의 유방암에서 발현된 Her2/neu(또한 ErbB2로서 공지됨) 수용체를 표적화한다); 세툽시맙(에르비투스로서 시판됨) 및 파니투무맙 베바시주맙(아바스틴으로서 시판됨)(이는 VEGF 리간드를 표적화한다).

[0145] VII. 부계 핵산으로부터 태아 핵산의 특성화

[0146] 본원에 다른 곳에서 언급된 바와 같이, 본원에 기재된 방법 및 시스템은 또한 대상체의 혈액 또는 혈장 내의 순환성 핵산을 특성화하는데 사용될 수 있다. 이러한 분석은 환자의 잠재적 질환 상태의 동정에 사용하기 위한 순환성 종양 DNA, 또는 태아 DNA를 비침습적 방식으로, 예를 들면, 양수 천자 또는 다른 침습적 절차를 통해 직접 샘플링에 재분류하지 않고 특성화하기 위해 임신한 여성의 혈액 또는 혈장 내의 순환성 태아 DNA의 분석을 포함한다.

[0147] 일부 경우에, 상기 방법은 부계 핵산 서열의 분석에 적어도 부분적으로 기초하여 태아 핵산 서열, 예를 들면, 순환성 태아 DNA를 특성화하는데 사용될 수 있다. 예를 들면, 긴 범위 서열 맥락은 본원에 기재된 방법 및 시스템을 사용하여 짧은 바코드화 단편으로부터 (예를 들면, 1 kb 이상, 5 kb 이상, 10 kb 이상, 15 kb 이상, 20 kb 이상, 30 kb 이상, 40 kb 이상, 50 kb 이상, 60 kb 이상, 70 kb 이상, 80 kb 이상, 90 kb 이상 또는 심지어 100 kb 이상의 길이를 갖는) 부계 및 모계 핵산에 대해 결정될 수 있다. 긴 범위 서열 맥락은 부계 및 모계 핵산 서열 둘 다에서 하나 이상의 일배체형 및 단일 뉴클레오타이드 다형성(SNP), 구조적 변형(예: 카피 수 변형,

삽입, 결실, 전좌, 역전, 재배열, 반복 확대, 레트로트랜스포존, 중복, 유전자 융합 등)을 포함하는 하나 이상의 유전적 변형을 결정하는데 사용될 수 있다. 또한, 부계 및 모계 핵산 및 임의의 결정된 SNP, 일배체형 및/또는 구조적 변형 정보의 긴 범위 서열 맥락은 임신한 어머니로부터 수득된 태아 핵산(예: 예를 들면, 무세포 태아 핵산과 같은 순환성 태아 핵산)의 서열을 특성화하는데 사용될 수 있다. 일부 경우에, 모계 및 부계 서열 및 특성화와의 비교를 통한 태아 핵산의 특성화는 프로그램된 컴퓨터 프로세서의 보조로 완료될 수 있다. 일부 경우에, 이러한 프로그램된 컴퓨터 프로세서는 본원의 다른 곳에서 기재된 예시적인 컴퓨터제어 시스템과 같은 컴퓨터 제어 시스템에 포함될 수 있다.

[0148] 예를 들면, 부계 및/또는 모계 핵산의 서열 및/또는 긴 범위 서열 맥락은 태아 핵산 서열을 포함하는 태아 핵산을 특성화하기 위한 참조로서 사용될 수 있다. 실제로, 본원에 제공된 방법 및 시스템에 의해 수득된 긴 범위 서열은 태아 핵산 서열이 특성화될 수 있는 부계 및 모계 핵산에 대한 향상된 긴 범위 서열 맥락 정보를 제공할 수 있다. 일부 경우에, 참조로서 부계 핵산으로부터 태아 핵산 서열의 특성화는 태아 핵산의 적어도 일부에 대한 서열을 결정하고/하거나 태아 핵산의 하나 이상의 SNP를 호출하고, 태아 핵산 서열의 하나 이상의 신규 돌연변이를 결정하고, 태아 핵산 서열의 하나 이상의 일배체형을 결정하고/하거나 태아 핵산의 서열에서 하나 이상의 구조적 변형 등을 결정하고 특성화하는 단계를 포함할 수 있다.

[0149] 도 17 내지 19는 짧은 바코드화 단편의 서열 분석을 통해 부계 및 모계 핵산에 대해 수득된 보다 긴 범위 서열 맥락으로부터 태아 핵산을 특성화하는 예시적 방법을 도시한다. 도 17은 보다 긴 범위 서열 맥락이 보다 짧은 바코드화 단편으로부터, 예를 들면, 도 6에 도시된 것과 유사한 방식으로 부계 핵산 샘플(예: 부계 게놈 DNA)에 대해 결정될 수 있는 예시적 방법을 입증한다. 도 17과 관련하여, 부계 핵산을 포함하는 샘플은 태아의 아버지로부터 수득될 수 있고(1700), 바코드화 비드의 세트도 또한 수득될 수 있다(1710). 비드는 하나 이상의 바코드 서열 뿐만 아니라 랜덤 N-량체 또는 다른 프라이머와 같은 프라이머를 함유하는 올리고뉴클레오타이드에 연결될 수 있다. 일부 경우에, 바코드 서열은 바코드화 비드로부터, 예를 들면, 바코드와 비드 사이의 결합의 절단을 통해 또는 바코드를 박리시키는 기초 비드의 분해를 통해 또는 둘의 조합으로 박리가 가능하다. 예를 들면, 일부 국면에서, 바코드화 비드는 바코드 서열을 박리시키는 환원제와 같은 제제에 의해 분해 또는 용해될 수 있다. 이러한 예에서, 핵산을 포함하는 부계 샘플(1705), 바코드화 비드(1715), 및, 일부 경우에, 다른 시약, 예를 들면, 환원제(1720)를 배합하고, 구획화에 적용한다. 일부 경우에, 부계 샘플(1700)을 구획화 전에 단편화하고, 생성되는 단편의 적어도 일부를 바코딩을 위해 1705로서 구획화된다. 예로써, 이러한 구획화는 성분을 액적 생성 시스템, 예를 들면, 미세 유체 장치(1725)에 도입하는 단계를 포함할 수 있다. 미세 유체 장치(1725)의 보조로, 유중수 에멀전(1730)이 형성될 수 있고, 여기서 상기 에멀전은 부계 샘플 핵산(1705), 환원제(1720) 및 바코드화 비드(1715)를 함유하는 수성 액적을 함유한다. 환원제는 바코드화 비드를 용해 또는 분해시켜 액적 내의 비드로부터 바코드 및 랜덤 N-량체를 갖는 올리고뉴클레오타이드를 방출시킬 수 있다(1735). 이어서, 랜덤 N-량체는 부계 샘플 핵산의 상이한 영역을 프라이밍하여 증폭 후 부계 샘플의 증폭된 카피를 초래할 수 있고, 여기서 각 카피는 바코드 서열로 태깅된다(1740). 일부 경우에, 증폭(1740)은 본원의 다른 곳에 기재되고 도 5에 개략적으로 도시된 것과 유사한 방법에 의해 달성될 수 있다. 일부 경우에, 각 액적은 동일한 바코드 서열 및 상이한 랜덤 N-량체 서열을 함유하는 올리고뉴클레오타이드의 세트를 함유한다. 이어서, 에멀전을 파괴되고(1745), 추가의 서열(예: 특히 서열 분석 방법, 추가의 바코드 등을 보조하는 서열)이, 예를 들면 증폭 방법을 통해 첨가될 수 있다(1750)(예: PCR). 이어서, 서열 분석이 수행될 수 있고(1755), 서열 분석 데이터를 해석하기 위해 알고리즘이 적용된다(1760). 일부 경우에, 예를 들면, 서열 분석 데이터의 해석(1760)은 부계 핵산의 적어도 일부에 대한 서열을 제공함을 포함할 수 있다. 일부 경우에, 부계 핵산 샘플에 대한 긴 범위 서열 맥락이 수득되고 특성화될 수 있다(예: 본원의 다른 곳에 기재된 바와 같은 하나 이상의 일배체형의 결정, 하나 이상의 구조적 변형(예: 카피 수 변형, 삽입, 삭제, 전좌, 역전, 재배열, 반복 확대, 중복, 레트로트랜스포존, 유전자 융합 등)의 결정, 하나 이상의 SNP의 호출, 하나 이상의 다른 유전적 변형의 결정 등). 일부 경우에, 변이체는 도 7과 관련하여 본원의 다른 곳에서 기재된 바와 같은 보다 긴 범위 서열 맥락을 제공하기 위해 생성된 추정 콘텍 및 다양한 부계 핵산에 대해 요구될 수 있다.

[0150] 도 18은 긴 범위 서열 맥락이, 예를 들면, 도 6에 도시된 것과 유사한 방식으로와 같이, 더 짧은 바코드화 단편으로부터 모계 핵산 샘플(예: 모계 게놈 DNA)에 대해 결정될 수 있는 예시적 방법을 입증한다. 도 18과 관련하여, 모계 핵산을 포함하는 샘플은 태아의 임신한 어머니로부터 수득될 수 있고(1800), 바코드화 비드의 세트도 또한 수득될 수 있다(1810). 비드는 하나 이상의 바코드 서열 뿐만 아니라 랜덤 N-량체 또는 다른 프라이머와 같은 프라이머를 함유하는 올리고뉴클레오타이드에 연결될 수 있다. 일부 경우에, 바코드 서열은 바코드화 비드로부터, 예를 들면, 바코드와 비드 사이의 결합의 절단을 통해 또는 바코드를 박리시키는 기초 비드의 분해를 통해 또는 둘의 조합으로 박리가 가능하다. 예를 들면, 일부 국면에서, 바코드화 비드는 바코드 서열을 박리시키는

환원제와 같은 제제에 의해 분해 또는 용해될 수 있다. 이 예에서, 핵산을 포함하는 모계 샘플(1805), 바코드화 비드(1815), 및, 일부 경우에, 다른 시약, 예를 들면, 환원제(1820)가 배합되고, 구획화에 적용된다. 일부 경우에, 모계 샘플(1800)은 구획화 전에 단편화되고, 생성되는 단편의 적어도 일부는 바코딩을 위해 1805로서 구획화된다. 예로써, 이러한 구획화는 성분을 액적 생성 시스템, 예를 들면, 미세 유체 장치(1825)에 도입하는 단계를 포함할 수 있다. 미세 유체 장치(1825)의 보조로, 유중수 에멀전(1830)이 형성될 수 있고, 여기서 에멀전은 모계 샘플 핵산(1805), 환원제(1820) 및 바코드화 비드(1815)를 함유하는 수성 액적을 함유한다. 환원제는 바코드화 비드를 용해 또는 분해시켜 액적 내의 비드로부터 바코드 및 랜덤 N-량체를 갖는 올리고뉴클레오티드를 박리시킬 수 있다(1835). 이어서, 랜덤 N-량체는 모계 샘플 핵산의 상이한 영역을 프라이밍하여 증폭 후 모계 샘플의 증폭된 카피를 초래할 수 있고, 여기서 각 카피는 바코드 서열로 태깅된다(1840). 일부 경우에, 증폭(1840)은 본원의 다른 곳에 기재되고 도 5에 개략적으로 도시된 바와 유사한 방법에 의해 달성될 수 있다. 일부 경우에, 각 액적은 동일한 바코드 서열 및 상이한 랜덤 N-량체 서열을 함유하는 올리고뉴클레오티드의 세트를 함유한다. 이어서, 에멀전은 파괴되고(1845), 추가의 서열(예: 특히 서열 분석 방법, 추가의 바코드 등을 보조하는 서열)이, 예를 들면, 증폭 방법을 통해 첨가될 수 있다(1850)(예: PCR). 이어서, 서열 분석이 수행될 수 있고(1855), 서열 분석 데이터를 해석하기 위해 알고리즘이 적용된다(1860). 일부 경우에, 예를 들면, 서열 분석 데이터의 해석(1860)은 모계 핵산의 적어도 일부에 대한 서열을 제공함을 포함할 수 있다. 일부 경우에, 모계 핵산 샘플에 대한 긴 범위 서열 맥락이 수득되고 특성화될 수 있다(예: 본원의 다른 곳에 기재된 하나 이상의 일배체형의 결정, 하나 이상의 구조적 변형(예: 카피 수 변형, 삽입, 결실, 전좌, 역전, 재배열, 반복 확대, 중복, 레트로트랜스포존, 유전자 융합 등)의 결정, 하나 이상의 SNP의 호출, 하나 이상의 다른 유전적 변형의 결정 등. 일부 경우에, 변이체가 도 7과 관련하여 본원의 다른 곳에 기재된 바와 같은, 더 긴 범위 서열 맥락을 제공하기 위해 생성된 추정 콘텍 및 샘플로부터 수득된 다양한 모계 핵산에 요구될 수 있다.

[0151]

도 19는 각각 도 17 및 도 18에 도시된 바와 같이 수득된 부계(1760) 및 모계(1860) 특성화로부터 태아 샘플 서열을 특성화하는 예를 입증한다. 도 19에 도시된 바와 같이, 태아 핵산 샘플은 임신한 어머니로부터 수득될 수 있다(1900). 긴 범위 서열 맥락은, 예를 들면, 도 6에 개략적으로 도시된 방법을 통해서와 같이, 본원의 다른 곳에 기재된 바와 같이 더 짧은 바코드화 단편의 서열 분석으로부터 태아 핵산에 대해 수득될 수 있다. 일부 경우에, 태아 핵산 샘플은, 예를 들면, 임신한 어머니의 혈액, 혈장, 다른 체액 또는 조직으로부터 수득될 수 있는 순환성 태아 DNA 및/또는 무세포 DNA일 수 있다. 바코드화 비드의 세트도 또한 수득될 수 있다(1910). 비드는 하나 이상의 바코드 서열 뿐만 아니라 프라이머, 예를 들면, 랜덤 N-량체 또는 다른 프라이머를 함유하는 올리고뉴클레오티드에 연결될 수 있다. 일부 경우에, 바코드 서열은 바코드화 비드로부터, 예를 들면, 바코드화 비드 사이의 결합의 절단을 통해 또는 바코드를 박리시키는 기초 비드의 분해를 통해 또는 둘의 조합으로 박리 가능하다. 예를 들면, 일부 국면에서, 바코드화 비드는 바코드 서열을 박리시키는 환원제와 같은 제제에 의해 분해 또는 용해될 수 있다. 이러한 예에서, 핵산을 함유하는 태아 샘플(1905), 바코드화 비드(1915), 및, 일부 경우에, 다른 시약, 예를 들면, 환원제(1920)를 배합하고, 1905로서 구획화에 적용한다. 일부 경우에, 태아 샘플(1900)은 구획화 전에 단편화되고, 생성되는 단편의 적어도 일부는 바코딩을 위해 1905로서 구획화된다. 예로써, 이러한 구획화는 성분을 액적 생성 시스템, 예를 들면, 미세 유체 장치(1925)에 도입하는 단계를 포함할 수 있다. 미세 유체 장치(1925)의 보조로, 유중수 에멀전(1930)이 형성될 수 있고, 여기서 에멀전은 모계 샘플 핵산(1905), 환원제(1920) 및 바코드화 비드(1915)를 함유하는 수성 액적을 함유한다. 환원제는 바코드화 비드를 용해 또는 분해시켜 바코드 및 랜덤 N-량체를 갖는 올리고뉴클레오티드를 액적 내의 비드로부터 박리시킬 수 있다(1935). 이어서, 랜덤 N-량체는 태아 샘플 핵산의 상이한 영역을 프라이밍하여 증폭 후 태아 샘플의 증폭된 카피를 초래할 수 있고, 여기서 각 카피는 바코드 서열로 태깅된다(1940). 일부 경우에, 증폭(1940)은 본원에 다른 곳에 기재되고 도 5에 개략적으로 도시된 것과 유사한 방법에 의해 달성될 수 있다. 일부 경우에, 각 액적은 동일한 바코드 서열 및 상이한 랜덤 N-량체 서열을 함유하는 올리고뉴클레오티드의 세트를 함유한다. 이어서, 에멀전은 파괴되고(1945), 추가의 서열(예: 특히 서열 분석 방법, 추가의 바코드 등을 보조하는 서열)이, 예를 들면, 증폭 방법을 통해 첨가될 수 있다(1950)(예: PCR). 이어서, 서열 분석이 수행될 수 있고(1955), 알고리즘이 서열 분석 데이터를 해석하기 위해 적용된다(1960). 일반적으로, 태아 핵산 샘플에 대한 더 긴 서열 맥락은 서열 분석된 더 짧은 바코드화 단편으로부터 수득될 수 있다. 일부 경우에, 예를 들면, 서열 분석 데이터의 해석(1960)은 태아 핵산의 적어도 일부를 위한 서열을 제공하는 단계를 포함할 수 있다. 태아 핵산 서열은 긴-범위 서열 맥락 및/또는 부계(1760) 및 모계(1860) 샘플의 특성화를 사용하여 특성화될 수 있다(1960)(예: 본원의 다른 것에 기재된 바와 같은 하나 이상의 일배체형의 결정, 하나 이상의 구조적 변형(예: 카피 수 변형, 삽입, 결실, 전좌, 역전, 재배열, 반복 확대, 중복, 레트로트랜스포존, 유전자 융합 등)의 결정, 하나 이상의 신규 돌연변이의 결정, 하나 이상의 SNP의 호출 등). 일부 경우에, 태아 핵산의 위상 블록은 태아 핵산 서열과

모계 및 부계 위상 블록의 비교로 결정될 수 있다.

[0152] 알 수 있는 바와 같이, 부계 핵산, 모계 핵산 및/또는 태아 핵산의 분석은 별도의 구획화 분석의 일부로서 완료될 수 있거나 하나 이상의 배합된 구획화 분석의 일부로서 완료될 수 있다. 예를 들면, 부계, 모계 및 태아 핵산이 동일한 장치에 첨가될 수 있고, 바코드화 모계, 부계 및 태아 단편이 도 17 내지 19에 따라서 액적으로 생성되고, 여기서 에멀전은 세 종류의 핵산을 위한 액적을 포함한다. 이어서, 에멀전은 파괴시키고, 액적의 내용물은 풀링하고, 추가로 처리하고(예: PCR을 통해 추가의 서열의 벌크 첨가), 본원에서 다른 곳에 기재된 바와 같이 서열 분석했다. 바코드화 단편으로부터 개별 서열 분석 판독치는 바코드 서열을 통해 그들 각각의 샘플 서열에 귀속될 수 있다.

[0153] 일부 경우에, 태아 게놈의 서열을 포함하는 태아 핵산 서열 및/또는 태아 핵산 서열의 유전적 변형은 본원에 기재된 방법 및 시스템을 사용하여 수득된 긴 범위 부계 및 모계 서열 맥락 및 특성화로부터 결정될 수 있다. 예를 들면, 순환성 태아 핵산의 서열 분석과 함께 부계 및 모계 게놈의 게놈 서열 분석을 사용하여 상응하는 태아 게놈 서열을 결정하는데 사용될 수 있다. 부계 게놈 및 무세포 태아 핵산의 서열 분석으로부터 게놈 태아 핵산의 서열을 결정하는 예는 다음 문헌에서 찾을 수 있다: Kitzman 등 (2012 June 6) Sci Transl. Med. 4(137): 137ra76, 이는 전체로 본원에 참조로 인용된다. 태아 게놈의 결정은, 예를 들면, 태아 이수성을 포함하는 태아의 유전적 장애의 산전 결정 및 진단에 유용할 수 있다. 본원에 다른 곳에서 논의된 바와 같이, 본원에 제공된 방법 및 시스템은 핵산 서열에서 일배체형을 분해하는데 유용할 수 있다. 일배체형-분해된 부계 및 모계 서열은 각각 태아 게놈의 서열을 보다 정확하게 결정하고/하거나 그들을 특성화하는데 보조할 수 있는 부계 및 모계 샘플 핵산 서열에 대해 결정될 수 있다.

[0154] 본원의 방법 및 시스템을 이용하면 부모 핵산 서열(예: 모계 핵산 서열, 부계 핵산 서열)의 긴-범위 서열 맥락을 포함하여 긴 범위 서열 맥락을 결정하는데 있어서 정확성을 향상시킬 수 있다. 본원에 제공된 방법 및 시스템은 적어도 70%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 99%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9%, 99.95%, 99.99%, 99.995% 또는 99.999%의 정확성으로 부모 핵산의 긴-범위 서열 맥락을 결정할 수 있다. 일부 경우에, 본원에 제공된 방법 및 시스템은 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.1%, 0.05%, 0.01%, 0.005%, 0.001%, 0.0005%, 0.0001%, 0.00005%, 0.00001% 또는 0.000005% 미만의 오차율로 부모 핵산의 긴-범위 서열 맥락을 결정할 수 있다. 또한, 본원의 방법 및 시스템은 또한 하나 이상의 국면에서 부계 핵산 서열을 특성화하는데(예: 서열의 결정, 하나 이상의 유전적 변형의 결정, 하나 이상의 구조적 변이의 결정, 일배체형의 결정 등) 있어서 정확성을 향상시킬 수 있다. 따라서, 본원에 제공된 방법 및 시스템은 하나 이상의 국면에서 적어도 70%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 99%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9%, 99.95%, 99.99%, 99.995% 또는 99.999%의 정확성으로 부계 핵산 서열을 특성화할 수 있다. 일부 경우에, 본원에 제공된 방법 및 시스템은 하나 이상의 국면에서 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.1%, 0.05%, 0.01%, 0.005%, 0.001%, 0.0005%, 0.0001%, 0.00005%, 0.00001% 또는 0.000005% 미만의 오차율로 부모 핵산 서열을 특성화할 수 있다.

[0155] 또한, 상기 논의된 바와 같이, 부모 핵산 서열의 긴-범위 서열 맥락을 결정하고, 이들의 특성화하는데 있어서 향상된 정확성은 태아 핵산의 서열 분석 및 특성화하는데 있어서 향상된 정확성을 초래할 수 있다. 따라서, 일부 경우에, 태아 핵산 서열 (긴-범위 서열 맥락 포함)은 적어도 70%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 99%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9%, 99.95%, 99.99%, 99.995% 또는 99.999%의 정확성으로 부모 핵산 서열의 분석으로부터 제공될 수 있다. 일부 경우에, 태아 핵산 서열 (긴-범위 서열 맥락을 포함)은 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.1%, 0.05%, 0.01%, 0.005%, 0.001%, 0.0005%, 0.0001%, 0.00005%, 0.00001% 또는 0.000005% 미만의 오차율로 부모 핵산 서열의 분석으로부터 제공될 수 있다. 일부 경우에, 태아 핵산 서열은 하나 이상의 국면에서 적어도 70%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 99%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9%, 99.95%, 99.99%, 99.995% 또는 99.999%의 정확성으로 본원에 기재된 바와 같은 부모 핵산 서열의 분석(예: 서열의 결정, 하나 이상의 유전적 변형의 결정, 하나 이상의 구조적 변형의 결정, 일배체형의 결정 등)을 통해 특성화될 수 있다. 일부 경우에, 태아 핵산 서열은 하나 이상의 국면에서 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.1%, 0.05%, 0.01%, 0.005%, 0.001%, 0.0005%, 0.0001%, 0.00005%, 0.00001% 또는 0.000005% 미만의 오차율로 본원에 기재된 바와 같은 부모 핵산 서열의 분석(예: 서열의 결정, 하나 이상의 유전적 변형의 결정, 일배체형의 결정, 하나 이상의 구조적 변형의 결정 등)을 통해 특성화될 수 있다.

[0156] VIII. 샘플

- [0157] 질환 또는 장애의 검출은 환자로부터 샘플을 수득하는 것으로 시작할 수 있다. 본원에 사용된 용어 "샘플"은 일반적으로 생물학적 샘플을 의미한다. 생물학적 샘플의 예는 핵산 분자, 아미노산, 폴리펩티드, 단백질, 탄수화물, 지방 또는 바이러스를 포함한다. 예에서, 생물학적 샘플은 하나 이상의 핵산 분자를 포함하는 핵산 샘플이다. 예시적인 샘플은 폴리뉴클레오티드, 핵산, 올리고뉴클레오티드, 무세포 핵산(예: 무세포 DNA (cfDNA)), 순환성 무세포 핵산, 순환성 종양 핵산(예: 순환성 종양 DNA (ctDNA)), 순환성 종양 세포 (CTC) 핵산, 핵산 단편, 뉴클레오티드, DNA, RNA, 펩티드 폴리뉴클레오티드, 상보성 DNA (cDNA), 이중 가닥 DNA (dsDNA), 단일 가닥 DNA (ssDNA), 플라스미드 DNA, 코스미드 DNA, 염색체 DNA, 게놈 DNA (gDNA), 바이러스 DNA, 세균 DNA, mtDNA (미토콘드리아 DNA), 리보솜 RNA, 무세포 DNA, 무세포 태아 DNA (cffDNA), mRNA, rRNA, tRNA, nRNA, siRNA, snRNA, snoRNA, scaRNA, 마이크로RNA, dsRNA, 바이러스 RNA 등을 포함할 수 있다. 요약하면, 사용되는 샘플은 특정 처리 요구 사항에 따라 달라질 수 있다.
- [0158] 핵산을 포함하는 임의의 물질은 샘플의 공급원일 수 있다. 물질은 유체, 예를 들면, 생물학적 유체일 수 있다. 유체 물질은, 혈액, 체대혈, 타액, 소변, 땀, 혈청, 정액, 질 유체, 위 및 소화액, 척수액, 태반액, 공동 유체, 안구액, 혈청, 모유, 림프액 또는 이의 조합을 포함할 수 있지만, 이에 한정되지 않는다. 물질은 고체, 예를 들면, 생물학적 조직일 수 있다. 물질은 정상적인 건강한 조직, 병든 조직, 또는 건강한 조직 및 병든 조직의 혼합물을 포함할 수 있다. 일부 경우에, 물질은 종양을 포함할 수 있다. 종양은 양성(비-암) 또는 악성(암)일 수 있다. 종양의 비제한적인 예는 다음을 포함할 수 있다: 섬유육종, 점액육종, 지방육종, 연골육종, 골육종, 척색종, 혈관육종, 내피육종, 림프관육종, 림프관내피육종, 활막종, 중피종, 유잉(Ewing's), 평활근육종, 횡문근육종, 위장관계 암종, 결장 암종, 췌장암, 유방암, 비뇨생식기계 암종, 난소암, 전립선암, 편평 세포 암종, 기저 세포 암종, 선암종, 한선 암종, 피지선 암종, 유두상 암종, 유두상 선암종, 낭종암, 수질 암종, 기관지 암종, 신장 세포 암종, 간세포암, 담관 암종, 융모막 암종, 정상 피종, 배아종, 윌름 종양, 자궁암, 내분비계 암종, 고환 종양, 폐 암종, 소세포 폐 암종, 비소세포 폐 암종, 방광 암종, 상피성 암종, 신경교종, 성상 세포종, 수모 세포종, 두개인두종, 상의세포종, 송과체종, 혈관모 세포종, 청신경종, 핍지교종, 수막종, 흑색종, 신경아세포종, 망막아종 또는 이들의 조합. 물질은 다양한 종류의 기관과 관련될 수 있다. 기관의 비제한적인 예는 뇌, 간, 폐, 신장, 전립선, 난소, 비장, 림프절(편도선 포함), 갑상선, 췌장, 심장, 골격근, 장, 후두, 식도, 위 또는 이들의 조합을 포함할 수 있다. 일부 경우에, 물질은 다음을 포함하지만, 이에 한정되지 않는 다양한 세포를 포함할 수 있다: 진핵 세포, 원핵 세포, 진균 세포, 심장 세포, 폐 세포, 신장 세포, 간 세포, 췌장 세포, 생식 세포, 줄기 세포, 유도된 다능성 줄기 세포, 위장 세포, 혈액 세포, 암 세포, 세균 세포, 인간 미생물군집 샘플로부터 분리된 세균 세포 등. 일부 경우에, 물질은, 예를 들면, 단일 세포의 내용물 또는 다중 세포의 내용물과 같은 세포의 내용물을 포함할 수 있다. 개별 세포를 분석하기 위한 방법 및 시스템은 다음에 제공된다: 예를 들면, 미국 가출원 제 62/017,558호(2014년 6월 26일에 출원), 이의 전체 개시 내용은 그 전체가 본원에 참조로 인용된다.
- [0159] 샘플은 다양한 대상체로부터 수득될 수 있다. 대상체는 살아 있는 대상체 또는 죽은 대상체일 수 있다. 대상체의 예는 인간, 포유동물, 비인간 포유동물, 설치류, 양서류, 파충류, 개파, 고양이과, 소과, 염소, 양류, 암탉, 아빈(avines), 마우스, 토끼, 곤충, 민달팽이, 미생물, 세균, 기생충 또는 물고기를 포함할 수 있지만, 이에 한정되지 않는다. 일부 경우에, 대상체는 질환 또는 장애를 갖는, 갖는 것으로 의심되거나 발생시킬 위험이 있는 환자일 수 있다. 일부 경우에, 대상체는 임신부일 수 있다. 일부 경우에, 대상체는 정상적인 건강한 임신부일 수 있다. 일부 경우에, 대상체는 특정 선천성 결함이 있는 아기를 임신할 위험이 있는 임신부일 수 있다.
- [0160] 샘플은 다양한 접근법에 의해 대상체로부터 수득할 수 있다. 예를 들면, 샘플은 순환계 접근(예: 주사기 또는 다른 장치를 통해 정맥내 또는 동맥내), 분비된 생물학적 샘플(예: 타액, 객담, 소변, 대변 등)의 수집, 생물학적 샘플(예: 수술 중 샘플, 수술 후 샘플 등)의 외과적(예: 생검) 획득, 면봉 표본 채집 (예: 구강 면봉, 인두 면봉) 또는 피켓팅을 통해 대상체로부터 수득될 수 있다.
- [0161] CNV는 치료의 효율과 관련될 수 있다. 예를 들면, 증가된 HER2 유전자 카피 수는 진행된 비소세포 폐암에서 게피티니브 요법에 대한 반응을 향상시킬 수 있다. 다음을 참조한다: Cappuzzo F. 등 (2005) J. Clin. Oncol. 23: 5007-5018. 높은 EGFR 유전자 카피 수는 라파티니브 및 카페시타빈에 대한 증가된 민감성을 예측할 수 있다. 다음을 참조한다: Fabi 등 (2010) J. Clin. Oncol. 28:15s (2010 ASCO Annual Meeting). 높은 EGFR 유전자 카피 수는 세툽시맙 및 파니투무맙에 대한 증가된 민감성과 관련된다.
- [0162] 카피 수 변형은 특정 치료제에 대한 암 환자의 내성과 관련될 수 있다. 예를 들면, 티미딜레이트 신타제의 증폭은 전이성 결장직장암 환자에서 5-플루로로우라실 치료에 대한 내성을 초래할 수 있다. 다음을 참조한다: Wang

등 (2002) PNAS USA vol. 99, pp. 16156-61.

[0163] **IX. 컴퓨터 제어 시스템**

[0164] 본 발명은 본원에 제공된 방법, 예를 들면, 본원에 기재된 바와 같이 핵산 서열 분석 및 유전적 변형의 결정화 하고, 참조 핵산 서열을 저장하고, 서열 분석을 수행하고/하거나 샘플 및 참조 핵산 서열을 비교하는 방법을 구현하도록 프로그래밍되거나 다르게는 구성된 컴퓨터 시스템을 제공한다. 이러한 컴퓨터 시스템의 일례는 도 22에 도시된다. 도 22에 도시된 바와 같이, 컴퓨터 시스템(2201)은 단일 코어 또는 다중 코어 프로세서 또는 병렬 처리를 위한 복수의 프로세서일 수 있는 중앙 처리 장치(CPU, 또한, 본원에서 "프로세서" 또는 "컴퓨터 프로세서")(2205)를 포함한다. 컴퓨터 시스템(2201)은 또한 메모리 또는 메모리 위치(2210)(예: 랜덤-액세스 메모리, 읽기 전용 메모리, 플래시 메모리), 전자 저장 장치(2215)(예: 하드 디스크), 하나 이상의 다른 시스템과 통신하기 위한 통신 인터페이스(2220)(예: 네트워크 어댑터), 및 주변 장치(2225), 예를 들면, 캐시, 다른 메모리, 데이터 저장 및/또는 전자 디스플레이 어댑터를 포함한다. 메모리(2210), 저장 장치(2215), 인터페이스(2220) 및 주변 장치(2225)는 송신 버스(실선), 예를 들면, 마더보드를 통해 CPU(2205)와 통신한다. 저장 장치(2215)는 데이터를 저장하기 위한 데이터 저장 장치 (또는 데이터 저장소)일 수 있다. 컴퓨터 시스템(2201)은 통신 인터페이스(2220)의 보조로 컴퓨터 네트워크("네트워크")(2230)에 작동적으로 연결될 수 있다. 네트워크(2230)는 인터넷, 인터넷 및/또는 엑스트라넷, 또는 인터넷과 통신되는 인트라넷 및/또는 엑스트라넷일 수 있다. 일부 경우에, 네트워크(2230)은 전기통신 및/또는 데이터 네트워크이다. 네트워크(2230)는 클라우드 컴퓨팅과 같은 분산된 컴퓨팅을 가능하게 할 수 있는 하나 이상의 컴퓨터 서버를 포함할 수 있다. 네트워크(2230)는, 일부 경우에, 컴퓨터 시스템(2201)의 보조로, 장치가 클라이언트 또는 서버로서 동작하는 컴퓨터 시스템(2201)에 연결되도록 할 수 있는 피어-투-피어 네트워크(peer-to-peer network)를 구현할 수 있다.

[0165] CPU(2205)는 프로그램 또는 소프트웨어로 구현될 수 있는 기계-판독가능한 명령어의 서열을 실행할 수 있다. 명령어는 메모리(2210)와 같은 메모리 위치에 저장될 수 있다. CPU(2205)에 의해 실행된 작업의 예는 페치, 디코드, 실행 및 후기입을 포함할 수 있다.

[0166] 저장 장치(2215)는 드라이버, 라이브러리 및 저장 프로그램과 같은 파일을 저장할 수 있다. 저장 장치(2215)는 사용자 데이터, 예를 들면, 사용자 선호도 및 사용자 프로그램을 저장할 수 있다. 컴퓨터 시스템(2201)은, 일부 경우에, 인트라넷 또는 인터넷을 통해 컴퓨터 시스템(2201)과 통신하는 원격 서버에 위치되는 것과 같은 컴퓨터 시스템(2201)의 외부에 있는 하나 이상의 부가적인 데이터 저장 단위를 포함할 수 있다.

[0167] 컴퓨터 시스템(2201)은 네트워크(2230)를 통해 하나 이상의 원격 컴퓨터 시스템과 통신할 수 있다. 예를 들면, 컴퓨터 시스템(2201)은 사용자의 원격 컴퓨터 시스템(예: 오퍼레이터)과 통신할 수 있다. 원격 컴퓨터 시스템의 예는 개인용 컴퓨터(예: 휴대용 PC), 슬레이트 또는 태블릿 PC를 포함한다: (예: Apple® 아이패드, Samsung® 갤럭시 탭), 전화, 스마트 폰(예: Apple® 아이폰, 안드로이드 지원 장치, Blackberry®), 또는 개인 디지털 또는 개인 휴대 정보 단말기. 사용자는 네트워크(2230)를 통해 컴퓨터 시스템(2201)에 액세스할 수 있다.

[0168] 본원에 기재된 방법은, 예를 들면, 메모리(2210) 또는 전자 저장 장치(2215) 상과 같은 컴퓨터 시스템(2201)의 전자 장치 위치에 저장된 기계(예: 컴퓨터 프로세서) 실행가능한 코드에 의해 구현될 수 있다. 기계 실행가능하거나 기계 판독가능한 코드는 소프트웨어의 형태로 제공될 수 있다. 사용 동안, 코드는 프로세서(2205)에 의해 실행될 수 있다. 일부 경우에, 코드는 저장 단위(2215)로부터 검색될 수 있고, 프로세서(2205)에 의한 용이한 액세스를 위해 메모리(2210)에 저장될 수 있다. 일부 상황에서, 전자 저장 장치(2215)는 배제될 수 있고, 기계 실행가능한 명령어는 메모리(2210)에 저장된다.

[0169] 코드는 사전 컴파일되어 코드를 실행하도록 개조된 프로세서를 갖는 기계로 사용하기 위해 구성될 수 있거나 런타임 동안 컴파일될 수 있다. 코드는 사전 컴파일된 또는 컴파일된 방식으로 코드를 실행할 수 있도록 선택될 수 있는 프로그래밍 언어로 제공될 수 있다.

[0170] 본원에 제공된 시스템 및 방법, 예를 들면, 컴퓨터 시스템(2201)의 국면은 프로그래밍으로 구현될 수 있다. 기술의 다양한 국면은 전형적으로 기계 판독가능한 매체의 형태 상에서 수행되거나 형태로 구현되는 기계 (또는 프로세서) 실행가능한 코드 및/또는 관련 데이터의 형태의 "제품" 또는 "제조품"으로 간주될 수 있다. 기계-실행가능한 코드는 전자 저장 장치, 예를 들면, 메모리(예: 읽기 전용 메모리, 랜덤-액세스 메모리, 플래시 메모리) 또는 하드 디스크에 저장될 수 있다. "저장"형 매체는 소프트웨어 프로그래밍을 위해 언제든지 비일시적 저장을 제공할 수 있는 다양한 반도체 메모리, 테이프 드라이브, 디스크 드라이브 등과 같은 컴퓨터, 프로세서 등의 유형의 메모리, 또는 이의 관련 모듈의 임의의 하나 또는 모두를 포함할 수 있다. 소프트웨어의 전체 또는

일부는 때로 인터넷 또는 다양한 다른 전기통신 네트워크를 통해 통신할 수 있다. 이러한 통신은, 예를 들면, 하나의 컴퓨터 또는 프로세서로부터 다른 것, 예를 들면, 관리 서버 또는 호스트 컴퓨터로부터 응용 서버의 컴퓨터 플랫폼으로 소프트웨어의 로딩을 가능하게 할 수 있다. 따라서, 소프트웨어 요소를 포함할 수 있는 다른 종류의 매체는 유선 및 광학 지상 통신선 네트워크를 통해 각종 무선-링크를 통해 로컬 장치들 사이의 물리적 인터페이스를 통해 사용되는 것과 같은 광학적, 전기적 및 전자기파를 포함한다. 유선 또는 무선 링크, 광학 링크 등과 같은 이러한 전파를 휴대하는 물리적 요소는 또한 소프트웨어를 포함하는 매체로서 간주될 수 있다.

본원에서 사용된 바와 같이, 비밀시적인 유형의 "저장" 매체로 제한되지 않는 한, 컴퓨터 또는 기계 "판독가능한 매체"와 같은 용어는 명령어를 실행하기 위한 프로세서로 제공하는데 참여하는 임의의 매체를 의미한다.

[0171] 따라서, 기계 판독가능한 매체, 예를 들면, 컴퓨터-실행가능한 코드는 유형의 저장 매체, 반송파 매체 또는 물리적 전송 매체를 포함하지만, 이에 한정되지 않는 다양한 형태를 취할 수 있다. 비휘발성 저장 매체는 도면에 도시된 데이터베이스 등을 구현하는데 사용될 수 있는 것과 같이, 예를 들면, 임의의 컴퓨터(들) 등의 임의의 저장 장치와 같은 광학 또는 자기 디스크를 포함한다. 휘발성 저장 매체는 이러한 컴퓨터 플랫폼의 메인 메모리와 같은 동적 메모리를 포함한다. 유형의 전송 매체는 동축 케이블; 구리선 및 컴퓨터 시스템 내에 버스를 포함하는 와이어를 포함하는 광섬유를 포함한다. 반송파 전송 매체는 전기 또는 전자기 신호, 또는 무선 주파수(RF) 및 적외선(IR) 데이터 통신 중에 발생하는 것과 같은 음파 또는 광파의 형태를 취할 수 있다. 따라서, 컴퓨터-판독가능한 매체의 일반적인 형태는, 예를 들면, 다음을 포함한다: 플로피 디스크, 플렉시블 디스크, 하드 디스크, 자기 테이프, 임의의 다른 자기 매체, CD-ROM, DVD 또는 DVD-ROM, 임의의 다른 광학 매체, 펀치 카드 종이 테이프, 홀의 패턴을 갖는 임의의 다른 물리적 저장 매체, RAM, ROM, PROM 및 EPROM, FLASH-EPROM, 임의의 다른 메모리 칩 또는 카트리지, 반송파 수송 데이터 또는 명령어, 이러한 반송파를 수송하는 케이블 또는 링크, 또는 컴퓨터가 프로그래밍 코드 및/또는 데이터를 판독할 수 있는 임의의 다른 매체. 이러한 형태의 컴퓨터 판독가능한 매체 중의 다수는 실행을 위한 프로세서로 하나 이상의 명령어의 하나 이상의 서열을 전달하는데 관련될 수 있다.

[0172] 컴퓨터 시스템(2201)은, 예를 들면, 컴퓨터 시스템(2201)에 연결된 핵산 서열 분석 기기의 출력 또는 판독치를 제공하기 위한 사용자 인터페이스(UI)를 포함하는 전자 디스플레이(2235)를 포함하거나 통신할 수 있다. 이러한 판독치는 핵산 서열 분석 판독치, 예를 들면, 소정의 핵산 샘플을 포함하는 핵산 염기 서열을 포함할 수 있다. UI는 또한 이러한 판독치를 사용하는 분석 결과를 나타내기 위해 사용될 수 있다. UI의 예는, 제한 없이, 그래픽 사용자 인터페이스(GUI) 및 웹 기반 사용자 인터페이스를 포함한다. 전자 디스플레이(2235)는 컴퓨터 모니터 또는 용량성 또는 저항성 터치스크린일 수 있다.

[0173] 실시예

[0174] 실시예 1: 위상 변이체의 동정

[0175] NA12878 인간 세포주로부터 게놈 DNA는 길이가 약 10kb인 단편을 회수하기 위해 블루 피핀DNA 사이징 시스템을 사용하여 단편의 크기 기초 분리에 적용하였다. 이어서, 크기 선택된 샘플 핵산을 미세 유체 구획화 시스템을 사용하여 불소화 오일 연속 상 내의 수성 액적 중의 바코드 비드와 함께 공동구획화하였다(참조: 예를 들면, 미국 가출원 제 61/977,804호(2014년 4월 10일에 출원), 모든 목적을 위해 이의 전체가 본원에 참조로 인용된다), 여기서, 수성 액적은 또한 dNTP, 열경화성 DNA 폴리머라제 및 액적 내의 증폭을 수행하기 위한 다른 시약, 뿐만 아니라 비드로부터 바코드 올리고뉴클레오타이드를 박리시키기 위한 화학적 활성제를 포함한다. 이는 1 ng의 전체 입력 DNA 및 2 ng의 전체 입력 DNA 모두에 대해 반복했다. 바코드 비드는 700,000개 이상의 상이한 바코드 서열의 바코드 다양성을 나타내는 스톡 라이브러리의 서브세트로서 획득되었다. 올리고뉴클레오타이드를 함유하는 바코드는 추가의 서열 성분을 포함하고, 다음과 같은 일반적인 구조를 가졌다:

[0176] 비드-P5-BC-R1-N량체

[0177] 여기서, P5 및 R1은 각각 Illumina 부착 및 Read1 프라이머 서열을 의미하고, BC는 올리고뉴클레오타이드의 바코드 부분을 나타내고, N-량체는 주형 핵산을 프라이밍하는데 사용된 랜덤 10 염기 N-량체 프라이밍 서열을 나타낸다. 다음을 참조한다: 예를 들면, 미국 특허 출원 제 14/316,383호(2014년 6월 26일이 출원), 이의 전체 개시 내용은 그 전체가 모든 목적을 위해 본원에 참조로 인용된다.

[0178] 비드 용해 후, 액적은 각 액적 내의 샘플 핵산의 주형에 대해 바코드 올리고의 프라이머 신장가능하도록 열순환되었다. 이는 상기 제시된 다른 포함 서열 이외에, 기원 구획을 나타내는 바코드 서열을 포함한 샘플 핵산의 카피된 단편을 초래했다.

[0179] 카피 단편의 바코드 표지 후, 증폭된 카피 단편을 포함하는 액적의 에멀전이 파괴되었고, 추가의 서열 분석기 필요 성분, 예를 들면, Illumina 서열 분석기를 위한 read2 프라이머 서열 및 P7 부착 서열을 추가의 증폭을 통해 카피 단편에 첨가했고, 이는 이러한 서열을 카피 단편의 다른 말단에 부착시켰다.

[0180] 이어서, 서열 분석 라이브러리를 Illumina HiSeq 시스템에서 10X 적용 범위, 20X 적용 범위 및 30X 적용 범위로 서열 분석한 다음, 생성되는 서열 판독치 및 이의 관련 바코드 서열을 분석했다. 이어서, 공통 바코드를 공유한 근위 맵핑 서열을 큰 콘텍으로 조립하고, 단일 뉴클레오타이드 다형성을 동정하고, 위상 SNP를 동정하기 위해 그들의 관련 바코드 및 서열 맵핑을 기준으로 하는 개별 출발 분자와 관련시켰다. 이어서, 중복 위상 SNP를 포함한 서열은 중복 위상 SNP에 기초하는 위상 서열 데이터의 위상 블록 또는 추정 콘텍으로 조립했다. 생성되는 데이터는 비교용 세포주를 위한 공지된 일배체형 맵과 비교하였다.

[0181] 적어도 하나의 접근법에서, 일련의 이형 변이체의 각 대립 유전자는 두 개 중의 하나 내지 두 개의 일배체형에 할당된다. 관찰된 판독치 및 바코드 데이터, 소정의 변이체의 세트 및 이형 변이체의 위상 할당의 로그-가능성을 반환하는 로그-가능성 함수 $\log P(\text{바코드화 판독치} \mid \text{위상 할당, 변이체})$ 가 정의된다. 로그-가능성 함수의 형태는 바코드화 서열 판독 데이터에 대한 두 개의 주요 관찰로부터 유래한다: (1) 하나의 바코드로부터 판독치는 반수체 게놈의 작은 분획을 포함하고, 따라서 게놈의 소정의 영역 중의 두 일배체형에 대한 판독치를 함유하는 하나의 바코드의 가능성은 작다. 반대로, 게놈의 로컬 영역에서 하나의 바코드에 대한 판독치는 매우 단일 일배체형으로부터 유래될 가능성이 있고; (2) 관찰된 염기가 그것이 유래된 일배체형 중의 진정한 염기와 상이할 가능성은 서열 분석기에 의해 할당된 관찰된 염기의 Phred QV에 의해 기재된다.

[0182] 이어서, 바코드화 판독치 및 변이체의 소정의 세트에 대한 로그-가능성 함수를 최대화하는 위상 구성이 보고된다. 이어서, 최대-추정 점수 일배체형 구성이 구조화된 검색 절차에 의해 발견된다. 먼저, 빔 탐색을 사용하여 인접 변이체의 작은 블록의 최적 위상 구성을 찾는다(예: 약 50개 변이체). 둘째, 블록의 상대적 위상은 블록 접합부 위의 스윕으로 결정된다. 이 시점에, 전체적인 거의 최적 위상 구성이 발견되고, 추가의 최적화를 위한 출발점으로 사용된다. 이어서, 개별 변이체의 일배체형 할당은 위상에 대한 국소적 개선을 발견하기 위해 반전시키고, 교체된 구성 사이의 로그-가능성의 차이는 그 위상 할당의 신뢰도 추정치를 제공한다. 마지막으로, 위상 구성은 내부적으로 정확할 확률이 높은 위상 블록으로 나뉜다. 이어서, 최적 구성의 로그-가능성을 현재 SNP의 모든 SNP 권리가 그들 일배체형 할당이 반전되는 구성과 비교함으로써 각 SNP에서 위상 블록을 파괴하는지의 여부를 시험한다.

[0183] 이하 표는 NA 12878 게놈에 대해 수득된 위상 매트릭스를 제공한다. 명백한 바와 같이, 매우 긴 위상 블록은 짧은 판독 서열 데이터로부터 수득되어 매우 낮은 짧거나 긴 스위치 오차로 위상 SNP의 상당한 비율을 올바르게 동정한다.

	10X 적용 범위	20X 적용 범위	30X 적용 범위	30X 빔 탐색
N50 위상 블록	193kb	385kb	428kb	489kb
가장 긴 위상 블록	2121kb	2514kb	2514kb	3027kb
긴 스위치 오차	0.0053	0.0021	0.0018	0.0015
짧은 스위치 오차	0.004	0.0017	0.0014	0.0012
위상 SNP	83%	94%	95%	95.2%

[0184]

[0185] 추가의 실험은 NA12878 트리오 (NA12878, NA12882 및 NA12877), 구자라티 (NA20847), 멕시코 (NA19662) 및 아프리카 (NA19701) 세포주 샘플을 포함하는 다수의 추가 샘플로부터 SNP를 위상화한다. 약 1MB의 N50 위상 블록 길이는 0.3% 미만의 스위치 오차로 위상화된 SNP의 95% 초과로 달성되었다. 예를 들면, 표적 풀 다운이 바코딩에 이어지는 동일 샘플의 전체 엑솔 서열 분석은 0.3% 미만의 스위치 오차로 다시 약 90%의 유전자 SNP 위상을 나타냈다.

[0186] **실시예 2: EML-4/ALK 유전자 역전/전좌의 동정**

[0187] 본원에 기재된 방법 및 공정은 특성화된 암 세포주로부터 구조적 변형을 검출하기 위해 사용되었다. 특히, NCI-H2228 폐암 세포주는 이의 게놈 내에 EML4-ALK 융합 전좌를 갖는 것으로 공지된다. 야생형과 비교된 변형 구조는 도 15에 도시되어 있다. 상부 패널에 도시된 바와 같이, 변이체 구조에서, 동일한 염색체 상에서 ALK 유전자로부터 상대적으로 분리되거나 멀리 떨어져 있는 EML-4 유전자는 대신 전좌되고 ALK 유전자에 융합된다(참조: 예를 들면, Choi, 등, *Identification of Novel Isoforms of the EML4-LK Transforming Gene in Non-Small Cell Lung Cancer*, J. Cancer Res., 68:4971 (July 2008)). 전좌와 함께, EML4 유전자는 또한 역전된다. 전좌

는, 전좌가 ALK의 엑손 20-29(백색 박스로서 도시됨)에 대한 EML-4의 엑손 1-6(블랙 박스로서 도시됨)의 융합 뿐만 아니라 EML-4의 엑손 1-19에 융합된 ALK의 엑손 7-23의 융합을 초래하는 야생형 구조와 비교하여, 추가로 패널 II에 도시된다.

[0188] 이러한 변형을 동정하기 위해, NCI-H2228 세포주로부터 게놈 DNA는 길이 약 10 kb의 단편을 선택하기 블루 Pippin® 시스템(Sage Sciences, Inc.)을 사용하는 크기 분리에 적용했다.

[0189] 이어서, 크기 선택된 샘플 핵산은, DNA를 바코딩 후 서열 분석 전에 아질런트 슈어셀렉트 엑솜 포획 키트(Agilent SureSelect Exome capture kit)를 사용하는 하이브리드 포획에 적용한 것 이외에는, 실시예 1에 대해 상기한 바와 같이, 바코드 비드와 함께 공동구획화하고, 증폭시키고, 서열 분석 라이브러리로 처리했다. 이어서, 서열 분석 라이브러리는 Illumina HiSeq 시스템 상에서 약 80X 적용범위로 서열 분석한 다음, 생성되는 서열 판독치 및 그들의 관련 바코드 서열을 분석했다. 전좌 이벤트에 걸친 게놈의 일부 사이에 공유된 바코드의 보다 많은 수가 야생형과 비교하여 명백하게 입증되고 야생형에 존재하지 않는 융합된 성분 사이의 구조적 근접성을 나타낸다. 특히, 도 16a에 도시된 바와 같이, 융합 구조는, 이형 세포주를 위한 야생형 작제물의 중복 바코드에 필적하는 12개 바코드의 EML-4 엑손 1-6과 ALK 엑손 20-29 사이 및 20개 바코드의 EML-4 엑손 7-23 및 ALK 엑손 1-19 사이에 바코드 중복을 나타냈다.

[0190] 대조적으로, 비변이체 세포주(NA12878)를 사용하여 작동된 음성 대조군은, 도 16b에 도시된 바와 같이, 실질적으로 약 140X의 서열 적용범위로 3ng의 출발 DNA를 사용하여 야생형 대 변이체 작제물에 대한 바코드 중복만을 나타냈다.

[0191] 특히, 다양한 서열 세그먼트에서 다수의 전체 맵핑된 바코드를 나타내지만, 단지 매우 작은 비율의 중복 바코드, 예를 들면, 0.5% 미만의 전체 맵핑된 바코드화 서열이 매우 많은 수의 공통 또는 중복 바코드를 나타내는 야생형 구조와 비교하여 융합 구조에 대해 관찰되었다. 그 결과, 융합 또는 전좌 중지점에 걸쳐 공통적으로 맵핑되는 바코드는 이러한 전좌 이벤트를 동정하기 위한 강력한 근거를 제공한다.

[0192] 이 탐색을 효율적인 희귀한 매트릭스-곱셈으로서 코딩하는 중요한 바코드 교차/중복을 갖는 게놈유전자 좌의 모든 상을 먼저 탐색하는 SV 검출용 알고리즘이 또한 사용되었다. 이어서, 이 첫 번째 단계의 지원자는 판독-쌍, 분할-판독 및 바코드 데이터를 도입하는 확률론적모델을 사용하여 필터링된다. NA12878 및 NA20847 상의 SV-호출은 복수의 대규모 결실 및 반전을 호출하고, 그들을 인접한 위상 블록에 대하여 위상화하고, 상기한 핵 트리오의 상속 패턴으로 위상의 일관성을 나타냄을 초래한다.

[0193] 실시예 3: CNV 스크리닝을 통해 낭창에 대한 증가된 민감성의 검출

[0194] 환자는 낭창에 대한 민감성에 대해 시험한다. 혈액은 환자로부터 채취한다. 무세포 DNA 샘플은 본원에 인용된 기술을 사용하여 서열 분석한다. 이어서, 서열은 상이한 유전자의 CNV를 결정하기 위해 공지된 게놈 참조 서열과 비교한다. 낮은 카피 수의 FCGR3B (CD16 세포 표면 면역글로불린 수용체)는 전신 홍반성 낭창에 대한 증가된 민감성을 나타낸다. 환자는 임의의 카피 수 이상 및 관련 위험/질환을 통보 받는다.

[0195] 실시예 4: CNV 스크리닝을 통해 신경아세포종에 대한 증가된 소인 검출

[0196] 환자는 신경아세포종에 대한 소인에 대해 시험한다. 혈액은 환자로부터 채취한다. 무세포 DNA 샘플은 본원에 인용된 기술을 사용하여 서열 분석한다. 이어서, 서열은 상이한 유전자의 CNV를 결정하기 위해 공지된 게놈 참조 서열과 비교한다. 1q21.1에서 CNV는 신경아세포종에 대한 증가된 소인을 나타낸다. 환자는 임의의 카피 수 이상 및 관련 위험/질환을 통보 받는다.

[0197] 실시예 5: CNV 스크리닝을 통해 폐암의 감별 진단

[0198] 만성 기침, 체중 감소 및 호흡 곤란이 있는 환자는 폐암에 대해 시험한다. 혈액은 환자로부터 채취한다. 순환성 종양 세포(CTC) 또는 무세포 DNA 샘플은 본원에 인용된 기술을 사용하여 서열 분석한다. 이어서, CTC 서열은 상이한 유전자의 CNV를 결정하기 위해 공지된 게놈 참조 서열과 비교한다. DNA 중의 EGFR 카피 수가 정상보다 높은 경우, 환자는 소세포 폐암 대신 비-소세포 폐암(NSCLC)을 감별 진단받을 수 있다. 비-소세포 폐암의 CTC는 또한 이를 소세포 폐암과 추가로 구별할 수 있는 다른 카피 수 변형을 가질 수 있다. 암의 단계에 따라, 수술, 화학요법 또는 방사선 요법이 처방된다.

[0199] 소세포 폐암은 대부분 자주 비-소세포 폐 암종보다 더 빠르고 넓게 전이된다(따라서 상이하게 단계화된다). NSCLC는 일반적으로 화학요법 및/또는 방사선에 매우 민감하지 않고, 따라서 수술은 초기 단계에 진단되는 경우, 종종 시스플라틴을 포함하는 보조 (부수적) 화학요법과 함께 선택되는 치료이다. 표적 요법은 또한 비-소

세포 폐암(NSCLC), 예를 들면, 크리조티니브와 같은 ALK 억제제를 갖는 환자에게 이용가능할 수 있다. 표적 요법은 단순히 (예를 들면, 전통적인 화학요법으로) 모두 빠르게 분열하는 세포를 방해하기보다는 암 발생 및 종양 성장에 필요한 특정 표적 분자를 방해하여 암 세포의 성장을 차단한다.

[0200] **실시예 6: 위상을 통해 태아 이수성의 감별 진단**

[0201] 태아 이수성은 염색체 수의 이상이다. 이수성은 통상적으로 중요한 물리적 및 신경학적 손상을 초래한다. X 염색체 수의 감소는 터너 증후군의 원인이다. 염색체 수 21의 카피 수의 증가는 다운 증후군을 초래한다. 양수천자 또는 용모막 용모 샘플링(CVS)과 같은 침습적 시험은 유산 위험을 초래할 수 있고, 모계혈액을 시험하는 덜 침습적 방법이 여기에 사용된다.

[0202] 다운 증후군 또는 터너 증후군의 가족력이 있는 임신한 환자를 시험한다. 태아 유전 물질을 함유하는 모계 혈액 샘플을 수집한다. 이어서, 상이한 염색체로부터 핵산은 본원에 기재된 바와 같은 바코드화 태그 분자와 함께 상이한 구획으로 분리한다. 이어서, 샘플을 서열 분석하고, 각 염색체 카피의 수를 정상 이배체 염색체 상 서열과 비교한다. 환자는 상이한 염색체에 대한 임의의 카피 수 이상 및 관련 위험/질환을 통보 받는다.

[0203] **실시예 7: 버킷 림프종의 감별 진단에 대한 위상을 통해 염색체 전좌 검출**

[0204] 버킷 림프종은 염색체의 t(8;14) 전좌를 특징으로 한다. 일반적으로 림프종을 진단받은 환자는 버킷 림프종에 대해 시험한다. 중앙-생검 시험편은 림프절로부터 수집된다. 상이한 염색체로부터 핵산은 본원에 기재된 바와 같은 바코드화 태그 분자와 함께 상이한 구획으로 분리된다. 이어서, 샘플은 서열 분석하고, 염색체 전좌를 검출하기 위해 대조군 DNA 샘플과 비교한다. 환자가 버킷 림프종을 갖는 것으로 진단되면, CHOP 또는 R-CHOP 섭생을 포함하여 다른 종류의 림프종의 경우보다 더욱 집중적인 화학요법 섭생이 필요할 수 있다. CHOP는 다음과 같이 구성된다: 이에 결합시키고, 가교결합의 형성을 유도함으로써 DNA를 손상시키는 알킬화제, 사이클로포스파미드; DNA 염기 사이에 그 자체를 삽입함으로써 DNA를 손상시키는 삽입제, 하이드록시다우노루비신 (또한 독소루비신 또는 아드리아마이신이라 칭명됨); 단백질 투블린에 결합함으로써 세포가 복제되는 것을 방지하는 온코빈 (빈크리스틴); 코르티코스테로이드인 프레드니손 또는 프레드니솔론. 이 요법은 또한 버킷 림프종이 B 세포 기원이기 때문에 모노클로날 항체 리툭시맙과 조합될 수 있고, 이 조합은 R-CHOP라 칭명된다.

[0205] **실시예 8: 부모 게놈과의 비교로 무세포 DNA로부터 유래된 태아 게놈 서열의 위상화**

[0206] 임신한 환자의 모계 DNA를 포함하는 샘플 및 태아의 아버지의 부계 DNA를 포함하는 샘플을 수집한다. 각 샘플의 핵산은 본원에 기재된 분자 바코드화 태그와 함께 상이한 구획으로 분리된다. 이어서, 샘플을 서열 분석하고, 서열은 구획화된 모계 및 부계 단편 각각에 대한 추정 콘텍을 생성하는데 사용한다. 추정 콘텍은 모계 및 부계 염색체 각각의 일부를 위한 일배체형 블록을 작제하는데 사용한다.

[0207] 태아 유전 물질을 함유하는 모계 혈액 샘플을 수집한다. 무세포 DNA는 모계 순환성 DNA 및 태아 순환성 DNA 둘다의 서열을 생성하기 위해 서열 분석한다. 판독치는 상기 생성된 부계 및 모계 위상 블록과 비교한다. 일부 위상 블록은 감수분열 동안 재조합에 적용한다. 부계 위상 블록이 아니라 모계 위상 블록과 일치하는 태아 물질이 동정된다. 일부 경우에, 태아 물질은 부계 위상 블록의 전체와 일치하고, 태아가 부계 상속 염색체 중의 부계 위상 블록을 갖는다는 것이 결정된다. 다른 경우에, 태아 물질은 위상 블록의 일부와 일치하고, 이어서 제2 위상 블록과 일치하고, 여기서, 두 위상 블록은 부계 게놈의 상동 염색체 영역에 존재한다. 감수성 재조합 이벤트가 이 영역에서 발생한다고 결정되고, 재조합의 가능성이 높은 지점이 결정되고, 두 부계 위상 블록의 조합인 신규한 태아 위상 블록이 생성된다.

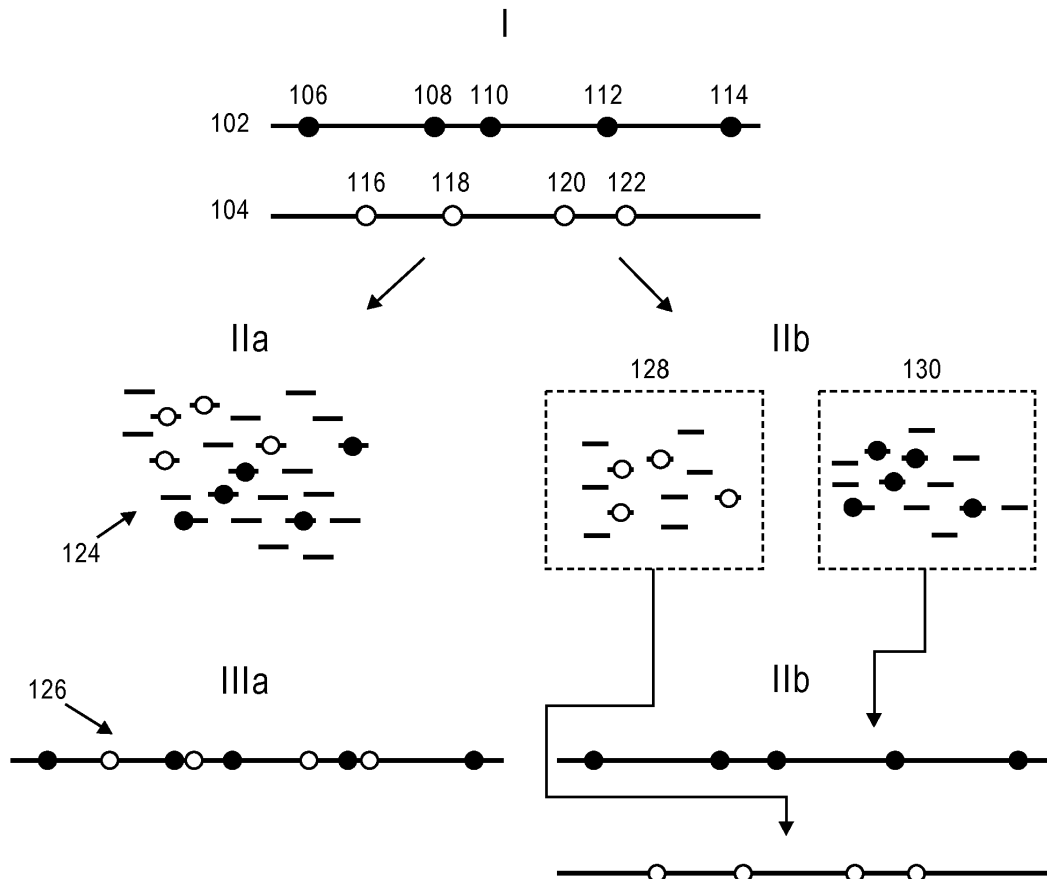
[0208] 순환성 DNA의 서열은 모계 위상 블록과 비교한다. 모계 위상 블록 중의 이형 접합성 부위는 모계 유래 태아 염색체의 가능성이 가장 높은 위상을 결정하는데 사용된다. 순환성 DNA 서열은 모계 게놈의 이형 접합성 부위에서 카피 수를 결정하는데 사용된다. 특정 모계 위상 블록의 상승된 카피 수는 태아의 모계 유래 염색체가 상승된 위상 블록의 서열을함유한다는 것을 나타낸다. 일부 경우에, 부계 경우에 기재된 바와 유사하게, 상동성 영역의 처음 하나의 위상 블록은 상승된것으로 나타나고, 이어서 동일 영역의 또 다른 위상 블록의 일부가 상승된 것으로 나타나서 감수성 재조합이 발생했음을 나타낸다. 이러한 경우에, 재조합의 가능성이 가장 높은 영역이 결정되고, 신규 태아 위상 블록은 두 개의 모계 위상 블록으로부터 작제된다.

[0209] 본 발명의 바람직한 구현예가 본원에 도시되고 기재되었지만, 이러한 구현예가 단지 예로써 제공된다는 것이 당업자에게 명백할 것이다. 본 발명이 명세서 내에 제공된 특정 실시예에 의해 제한되는 것이 의도되지 않는다. 본 발명이 상기 명세서를 참조하여 기재되었지만, 본원의 구현예의 상세한 설명 및 예시는 제한적인 의미로 해석되는 것을 의미하지 않는다. 다수의 변형, 변경 및 대체가 본 발명으로부터 벗어나지 않고 당업자에게 발생할

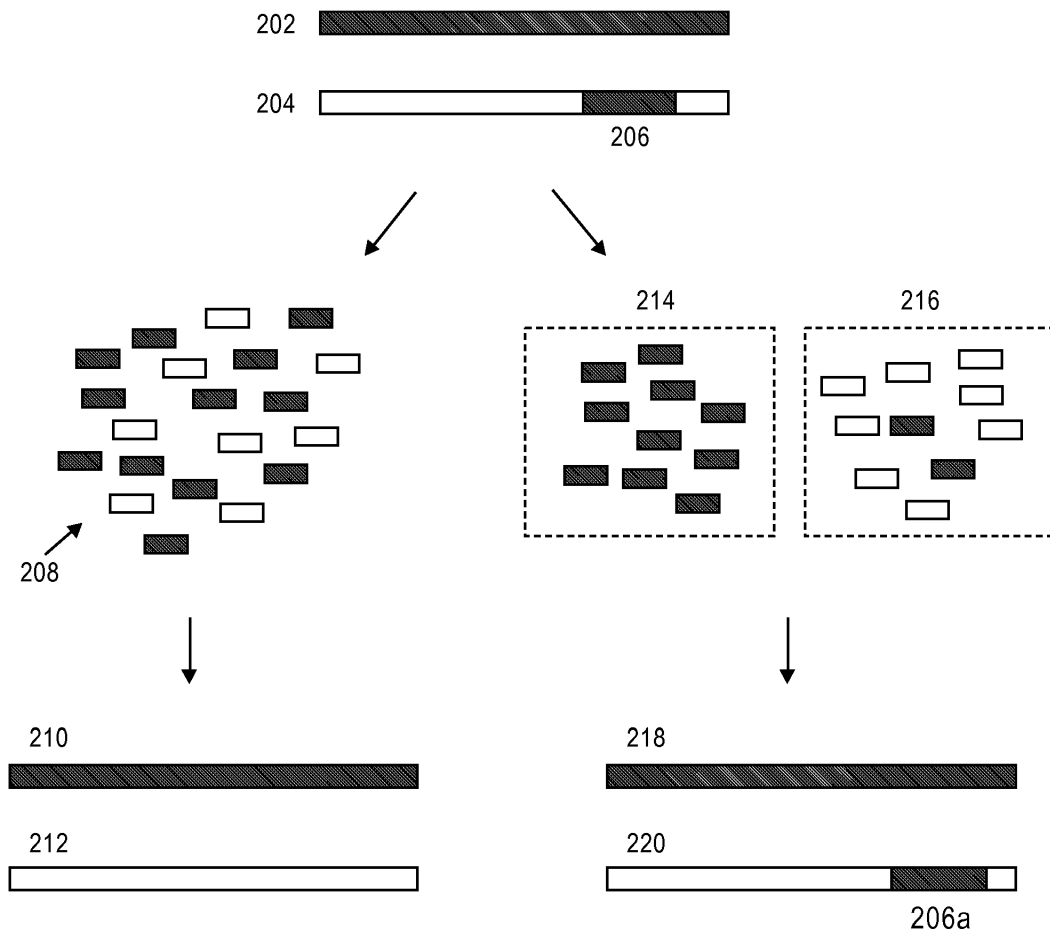
것이다. 나아가, 본 발명의 모든 측면은 본원에 명시되고, 다양한 조건 및 변수에 좌우되는 특정한 서술, 구성 또는 상대적 비율에 제한되지 않는다는 것이 이해되어야 한다. 본원에 기재된 본 발명의 구현예에 대한 다양한 대안이 본 발명을 실행하는데 사용될 수 있음이 이해되어야 한다. 따라서, 본 발명은 또한 이러한 대안, 변경, 변형 또는 등가물을 포함한다는 것이 고려된다. 다음 특허청구범위는 본 발명의 범위를 규정하고, 이들 특허청구범위 내의 방법 및 구조 및 이들의 등가물은 그것으로 포함되는 것이 의도된다.

도면

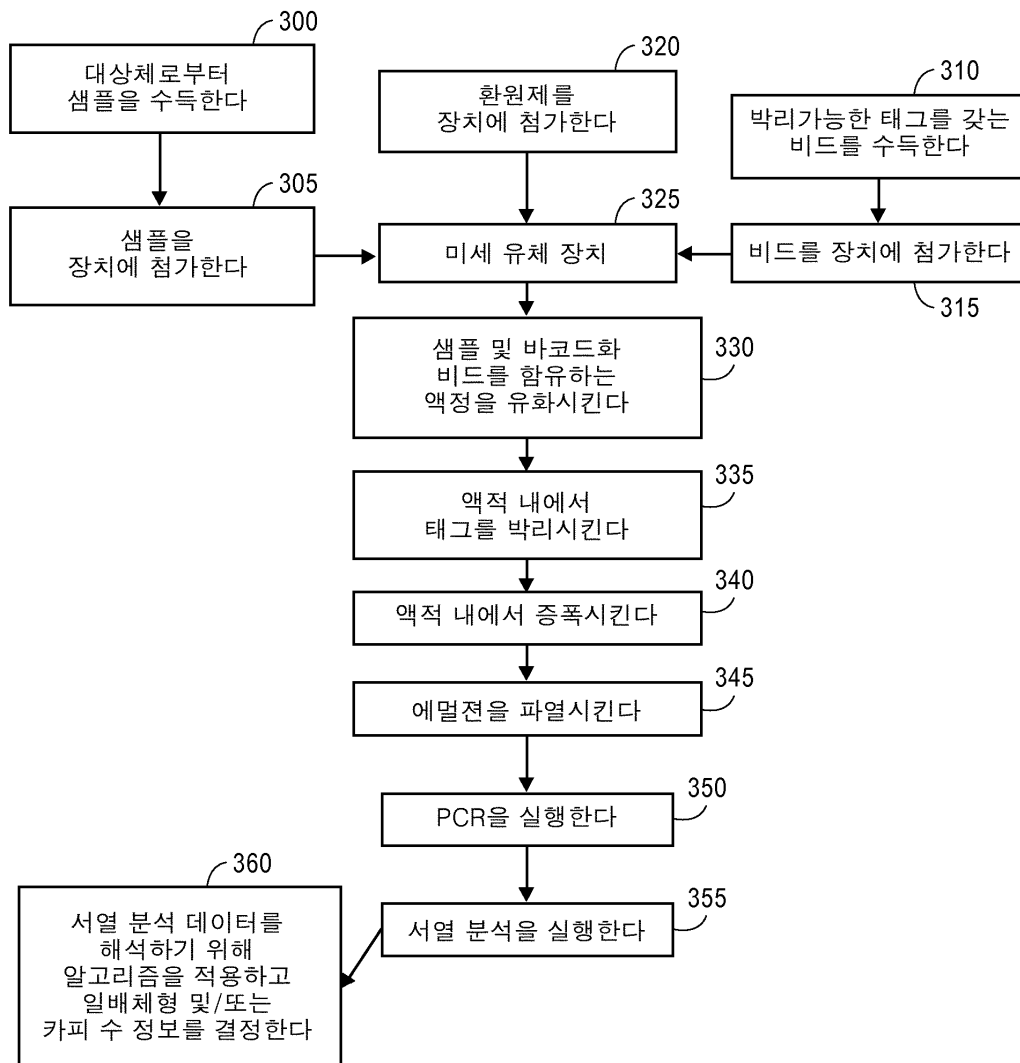
도면1



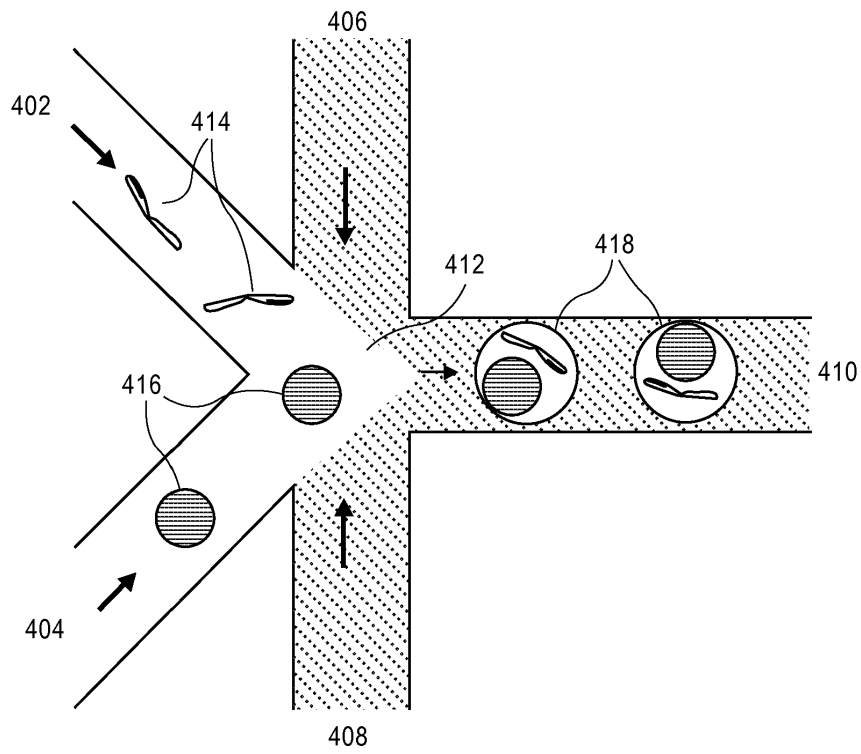
도면2



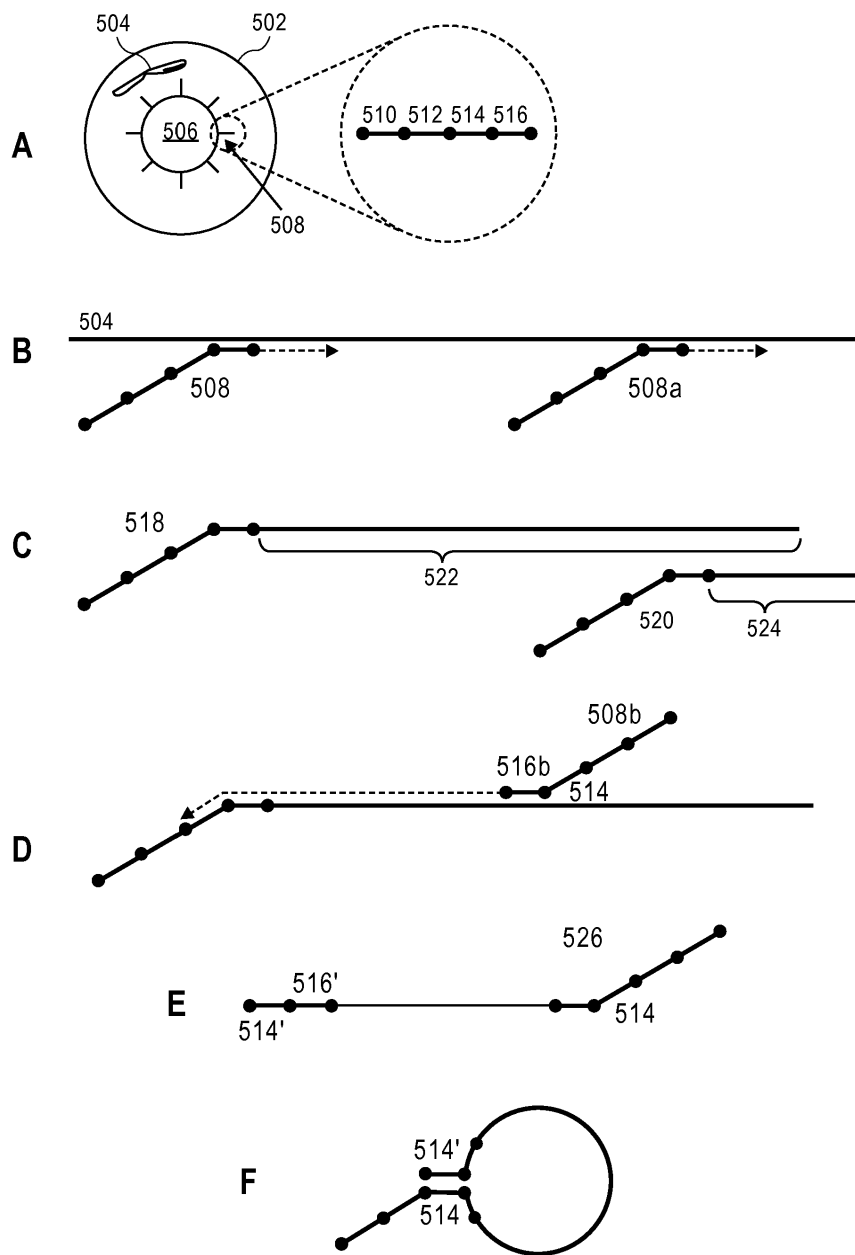
도면3



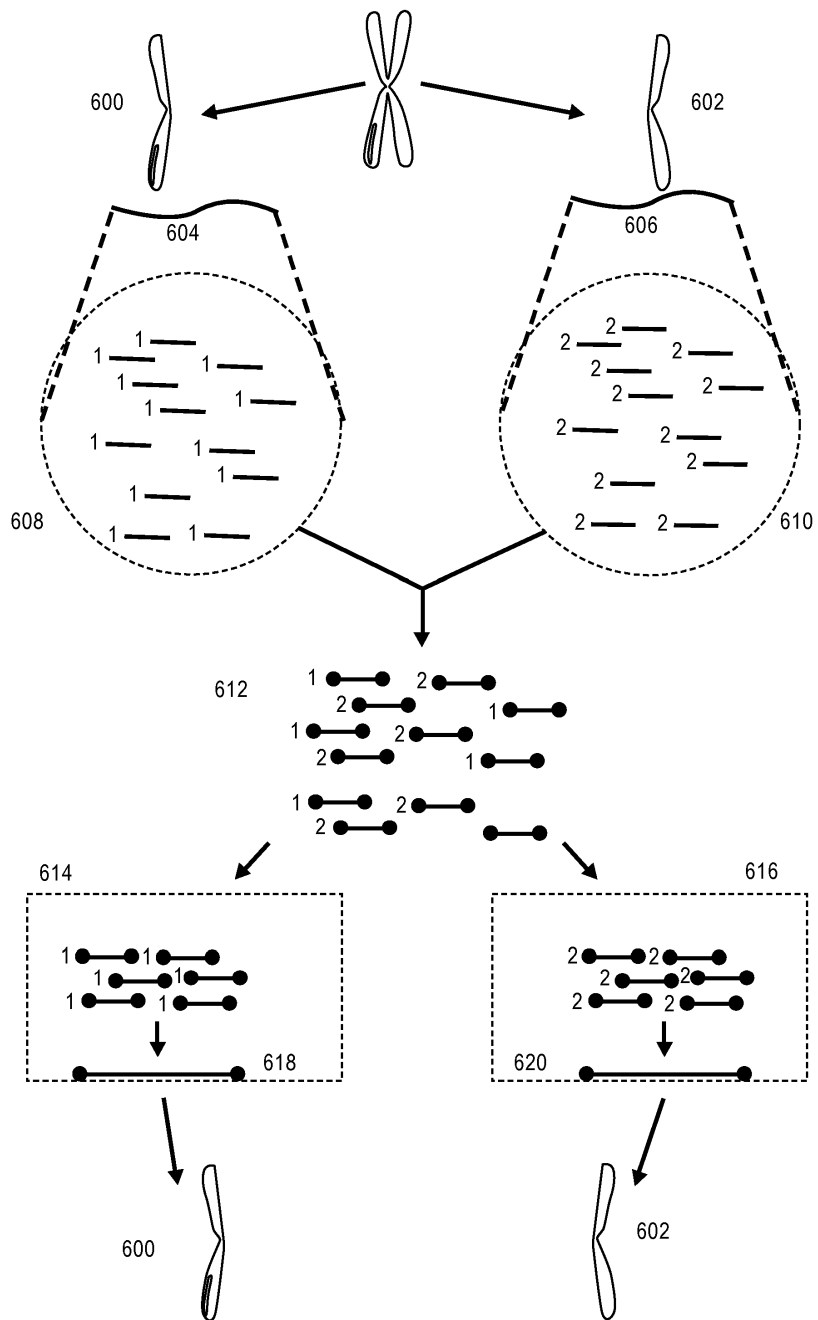
도면4



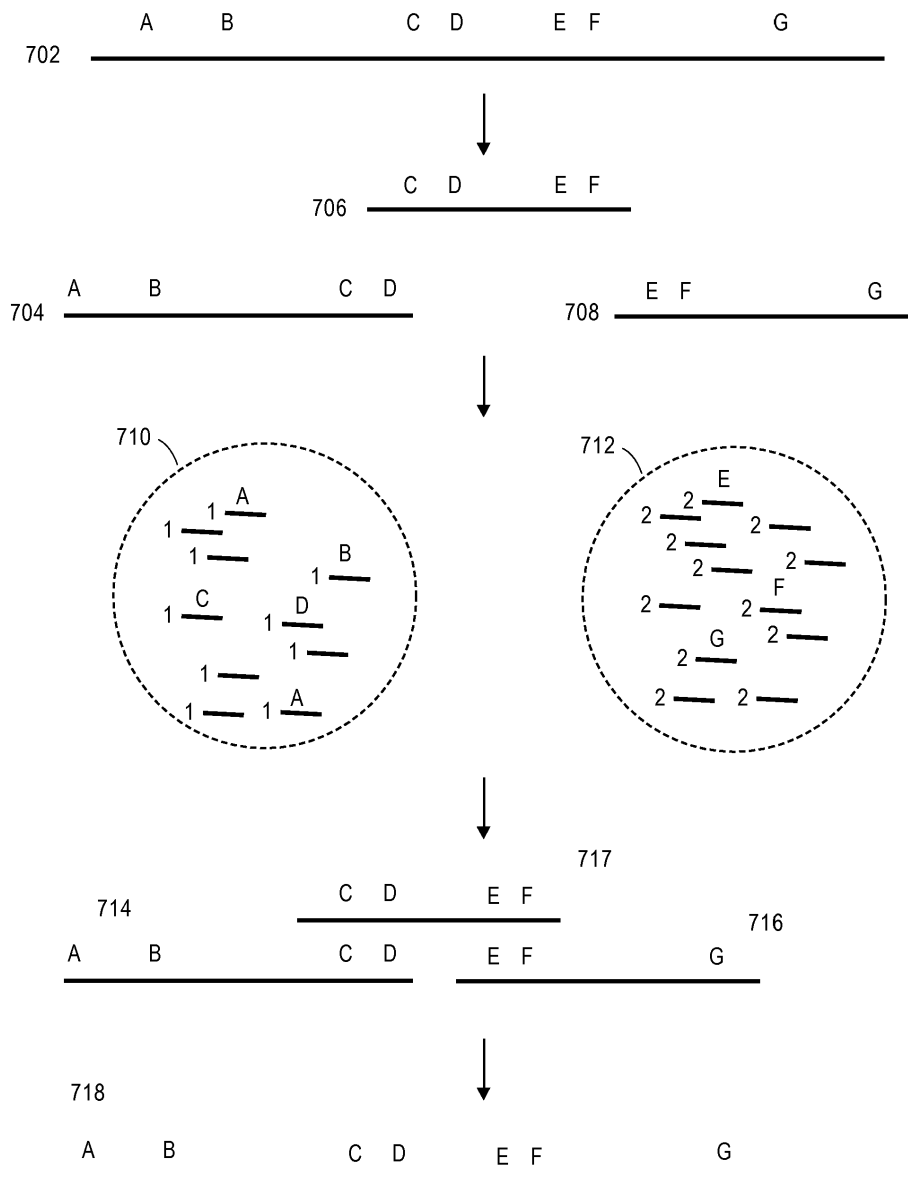
도면5



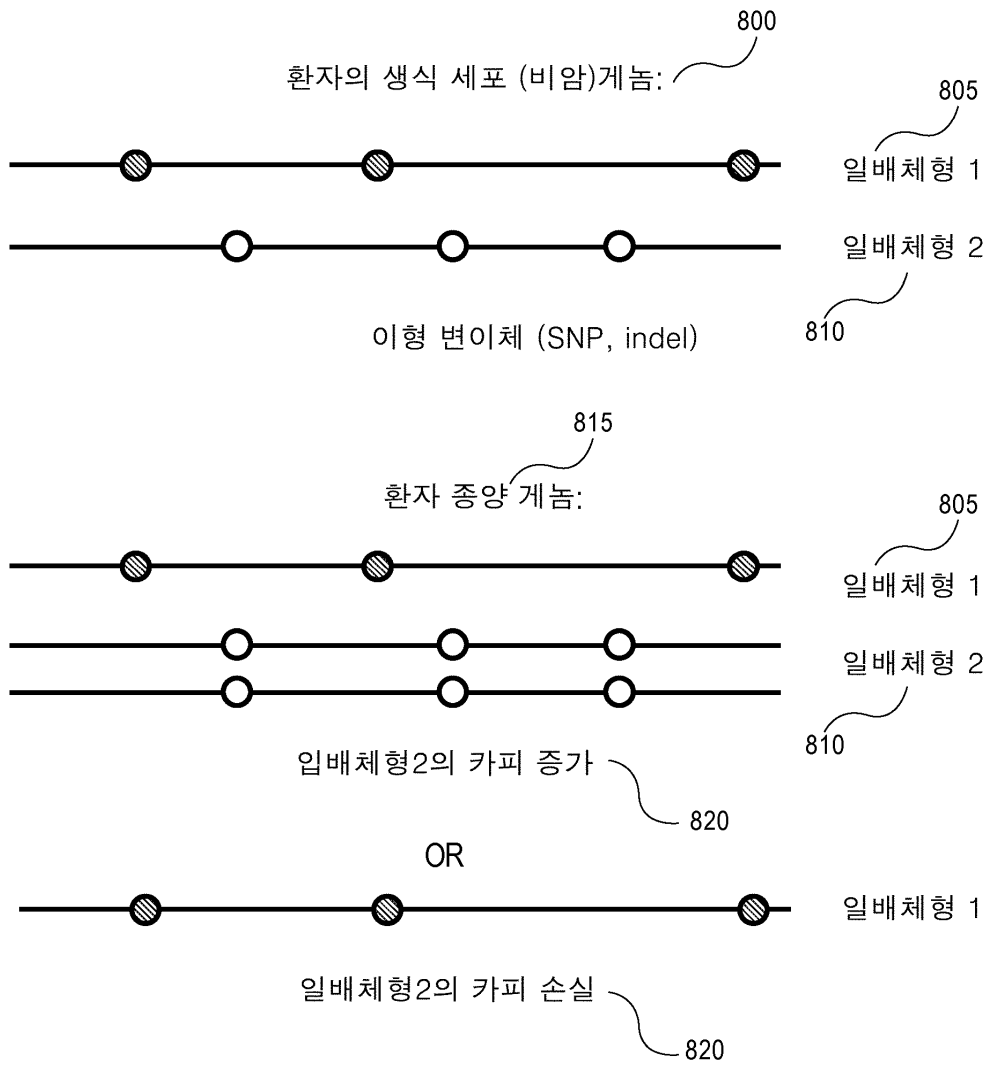
도면6



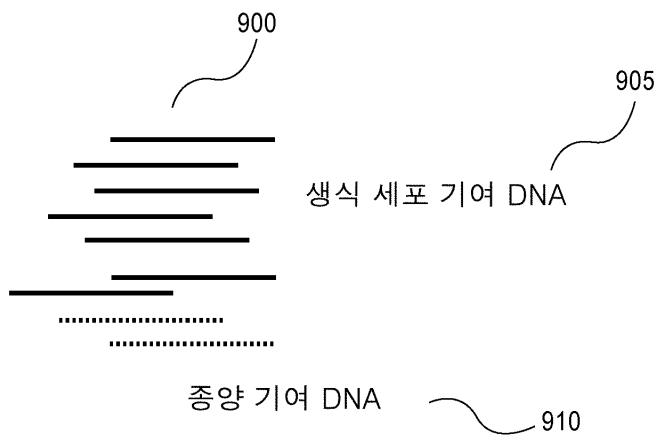
도면7



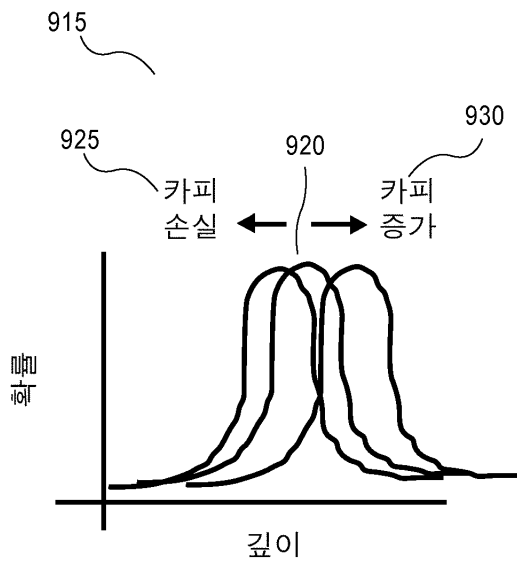
도면8



도면9a



도면9b



예상된 계수 (평균 깊이 D에서)

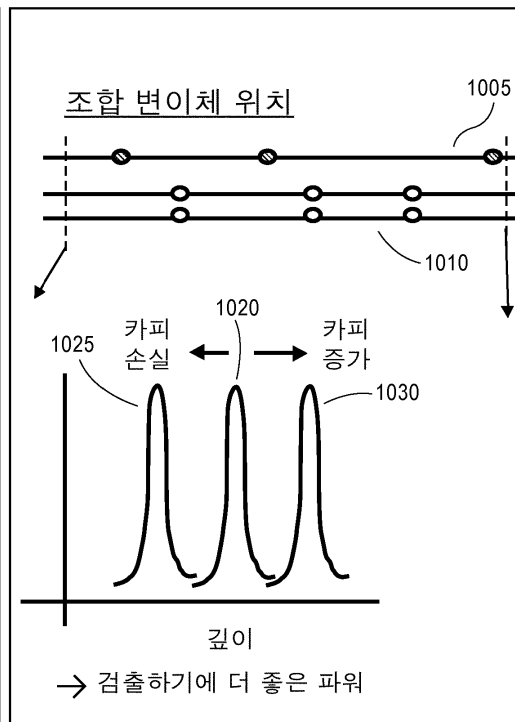
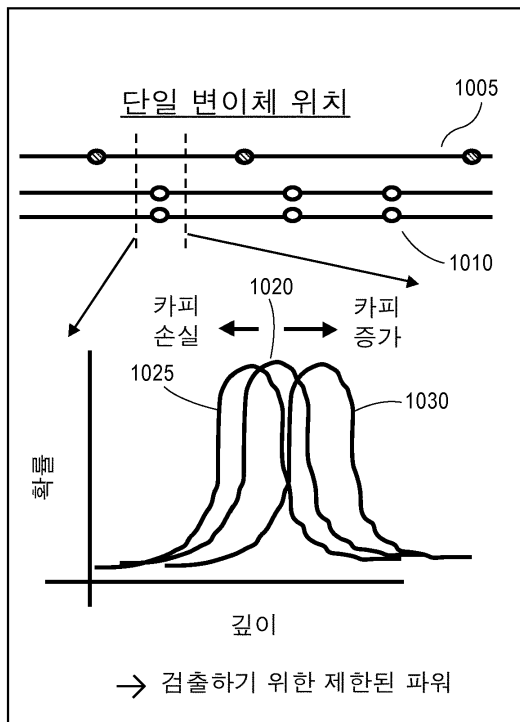
CNV 없음: D 935

손실: $D * (1 - F/2)$ 940

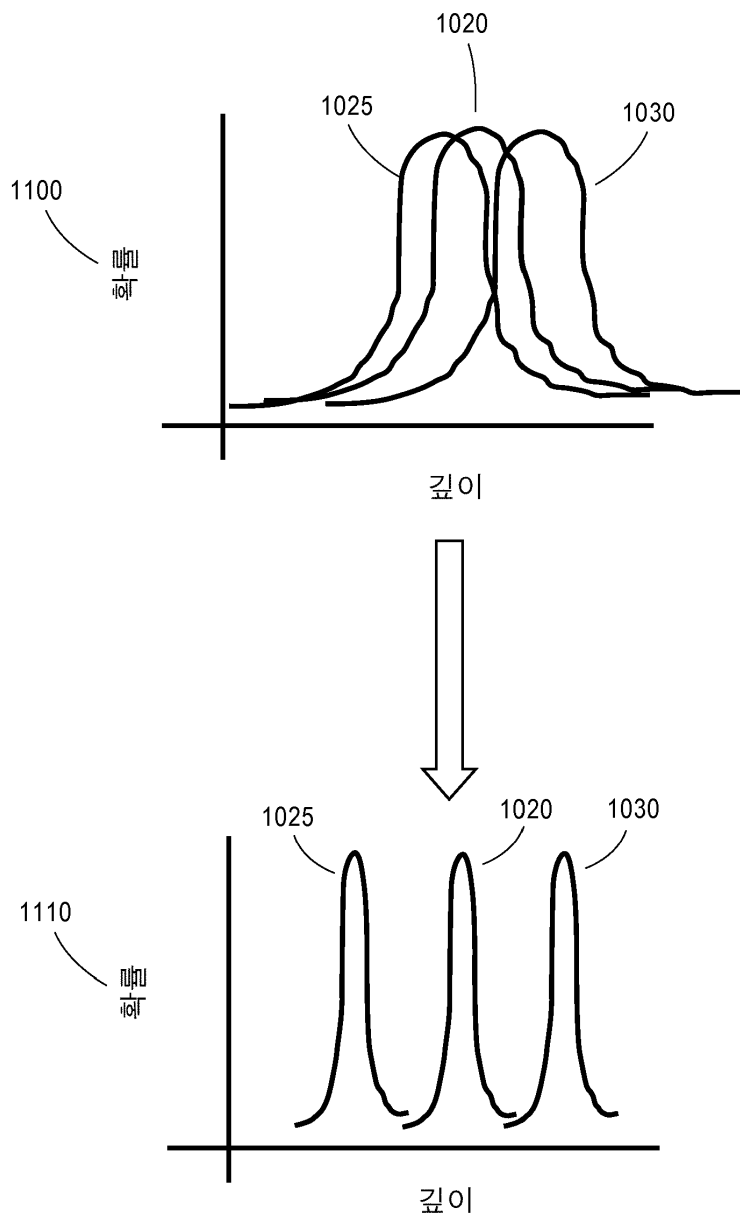
증가: $D * (1 + F/2)$ 945

→ 검출하기 위한 제한된 파워

도면10

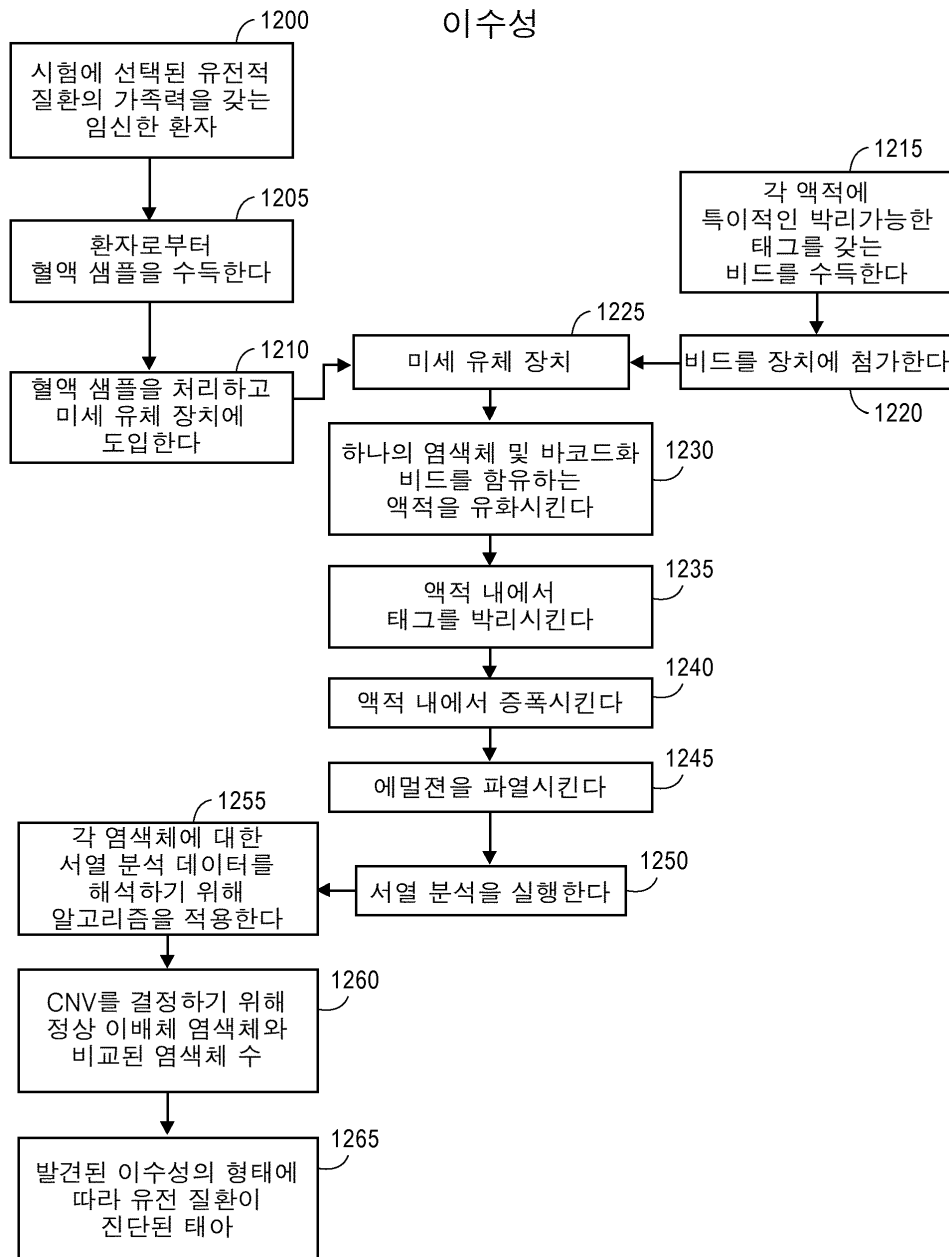


도면11

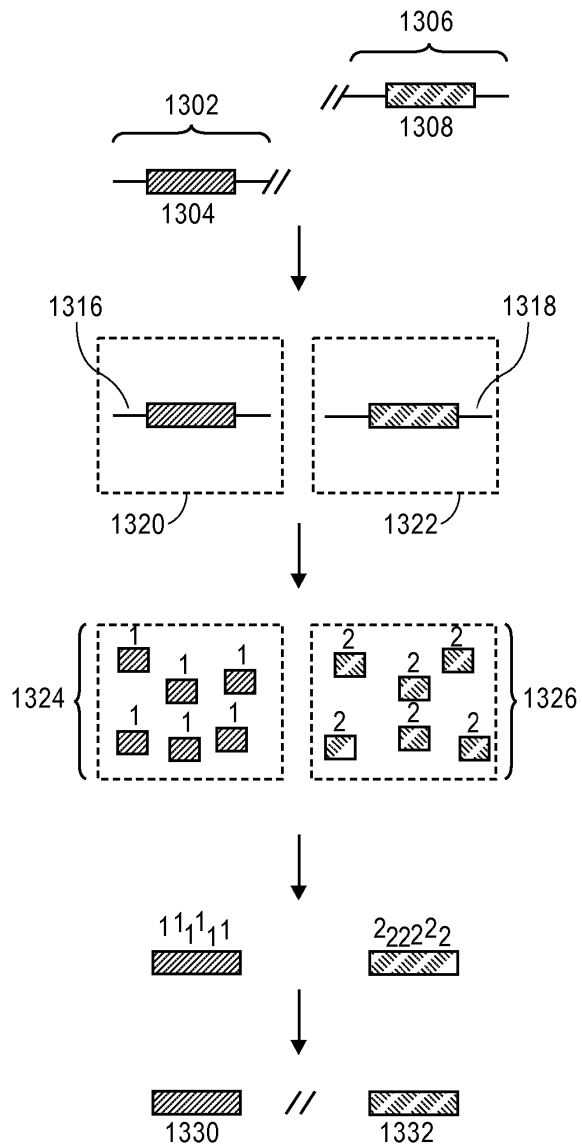


도면12

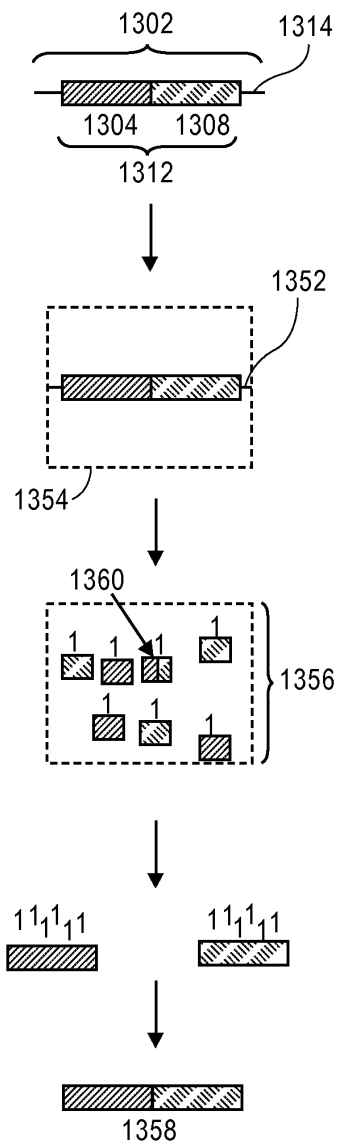
이수성



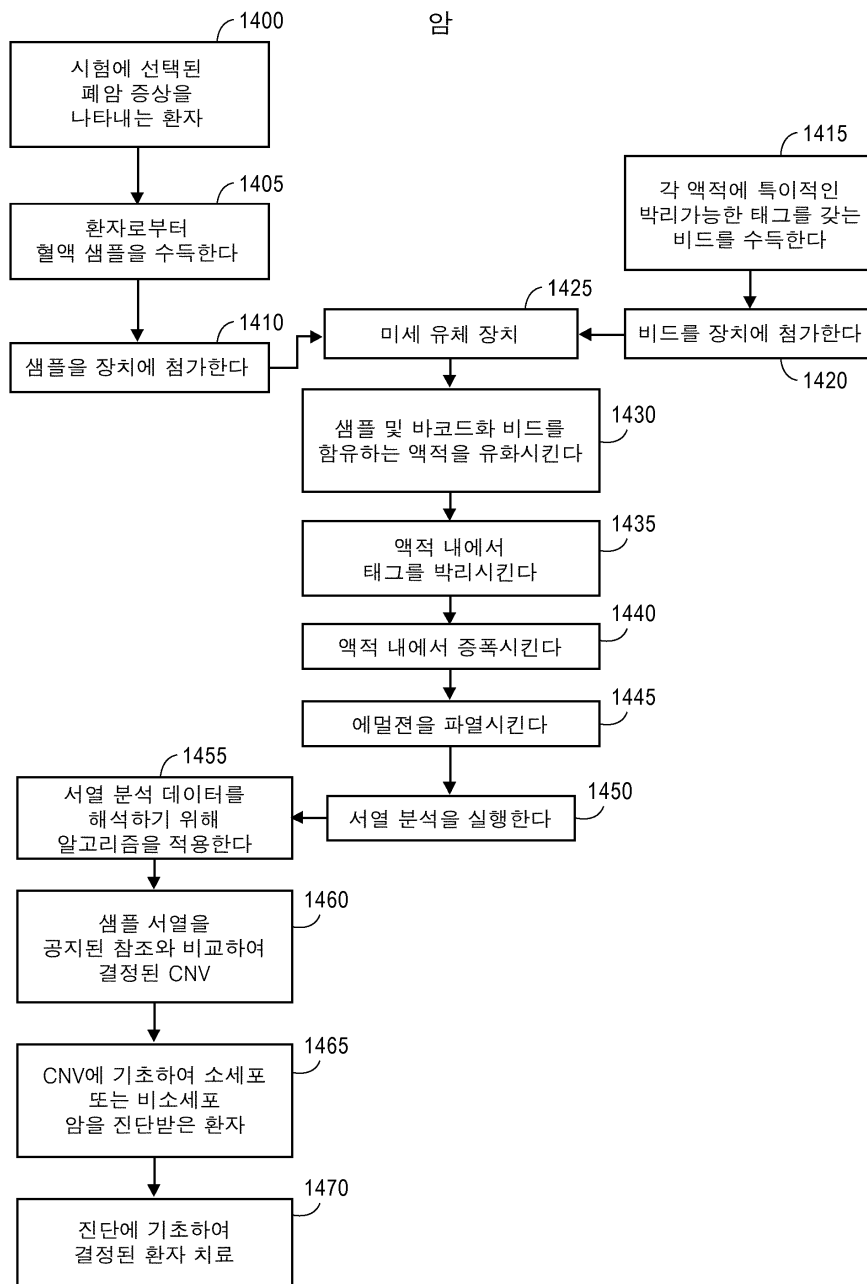
도면13a



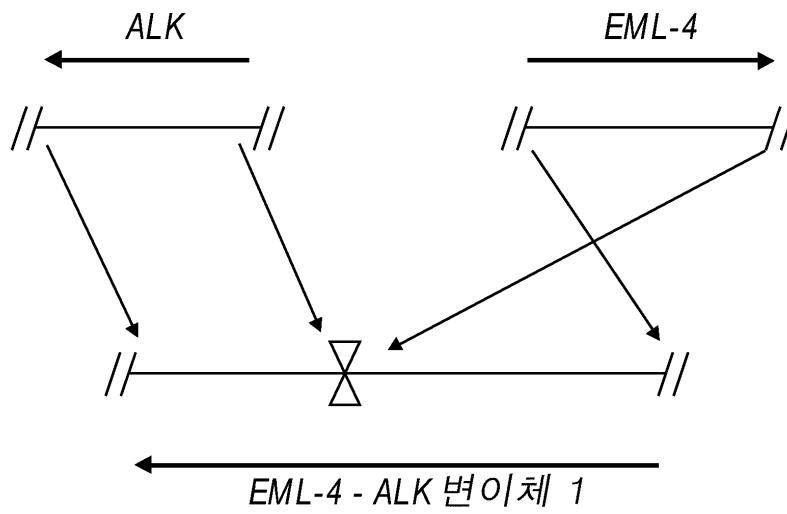
도면13b



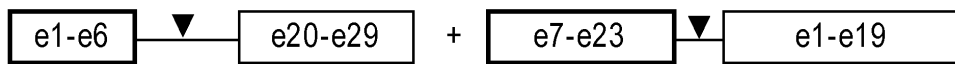
도면14



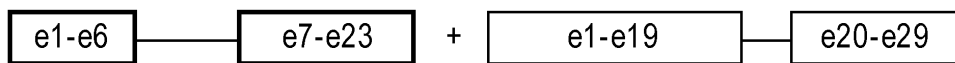
도면15



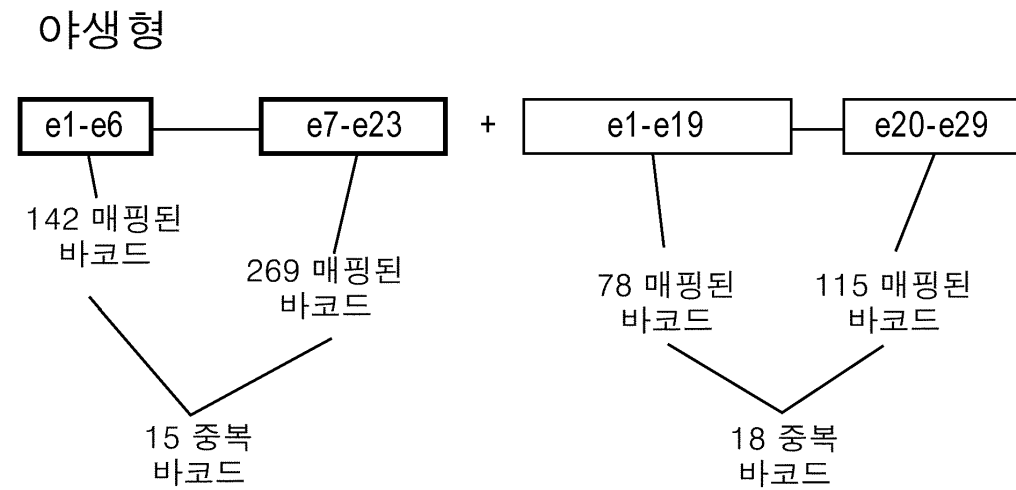
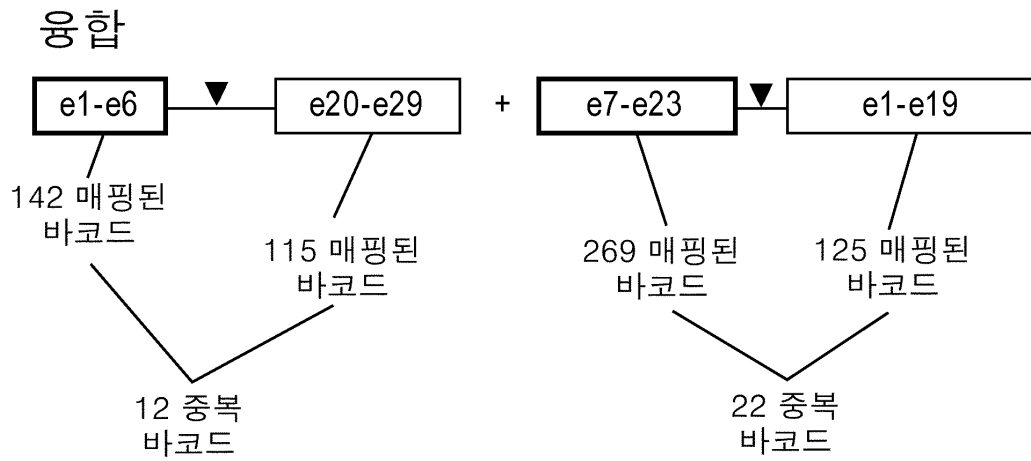
융합



야생형

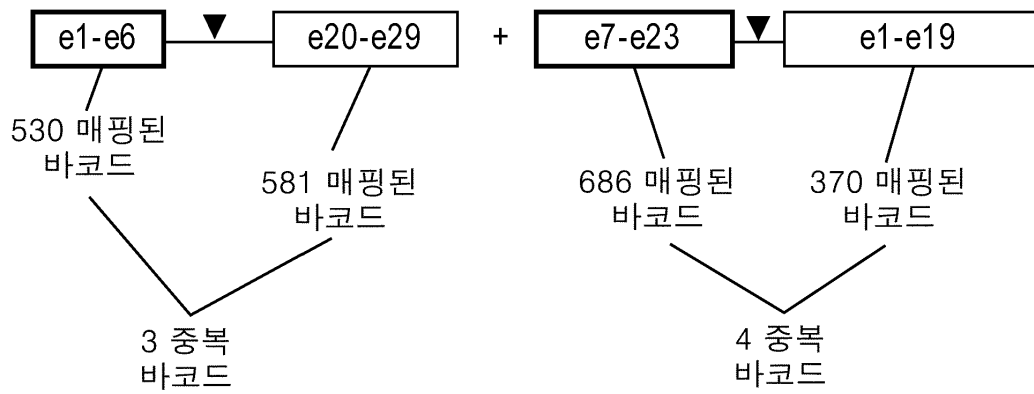


도면16a

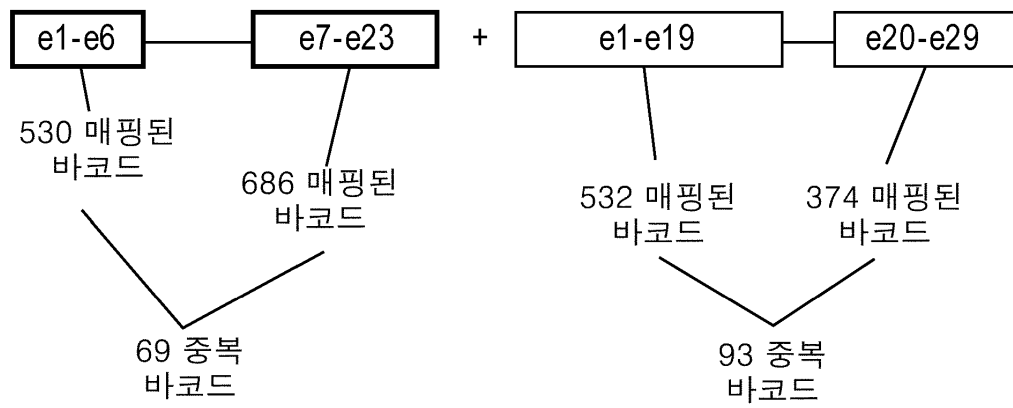


도면16b

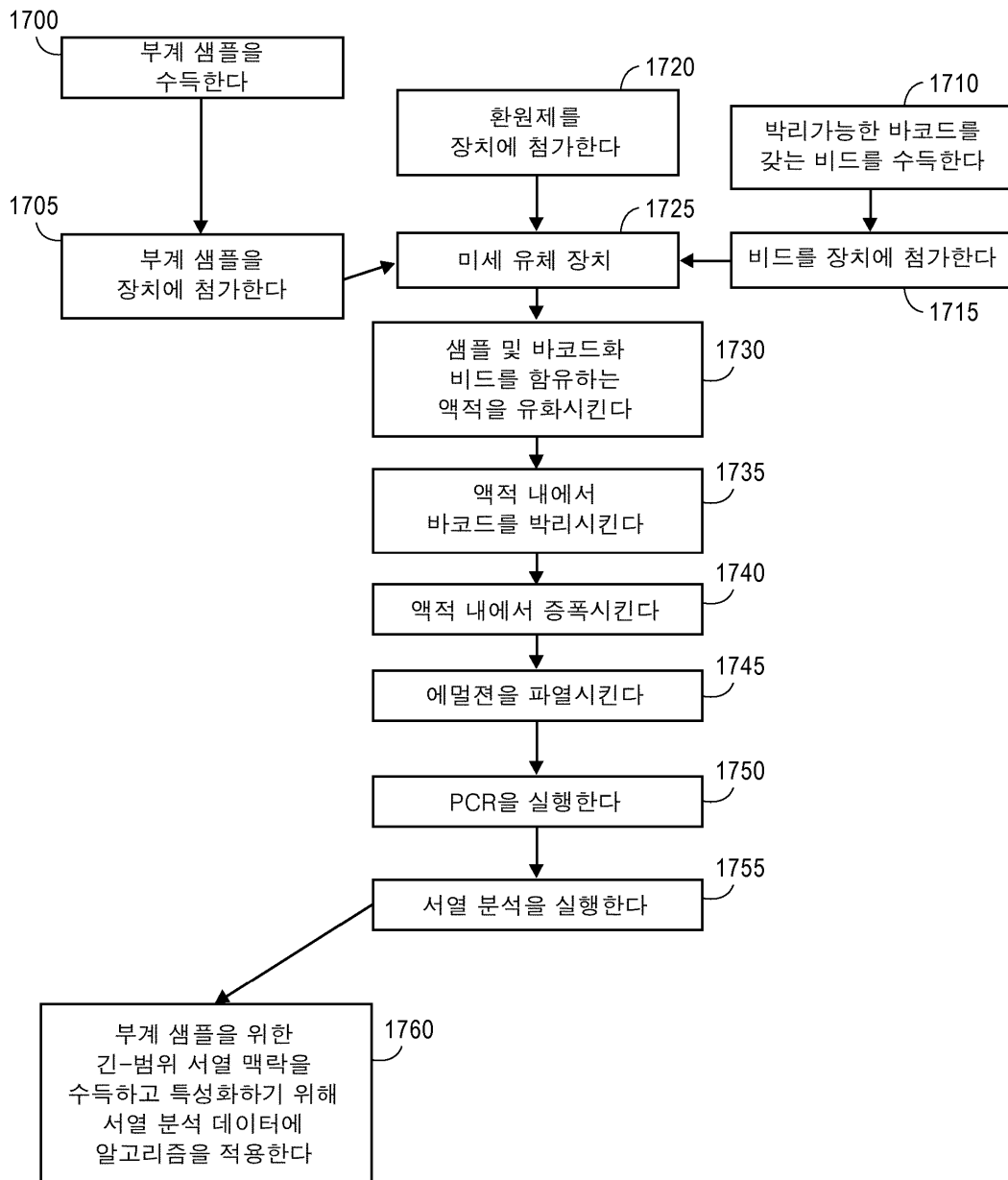
융합



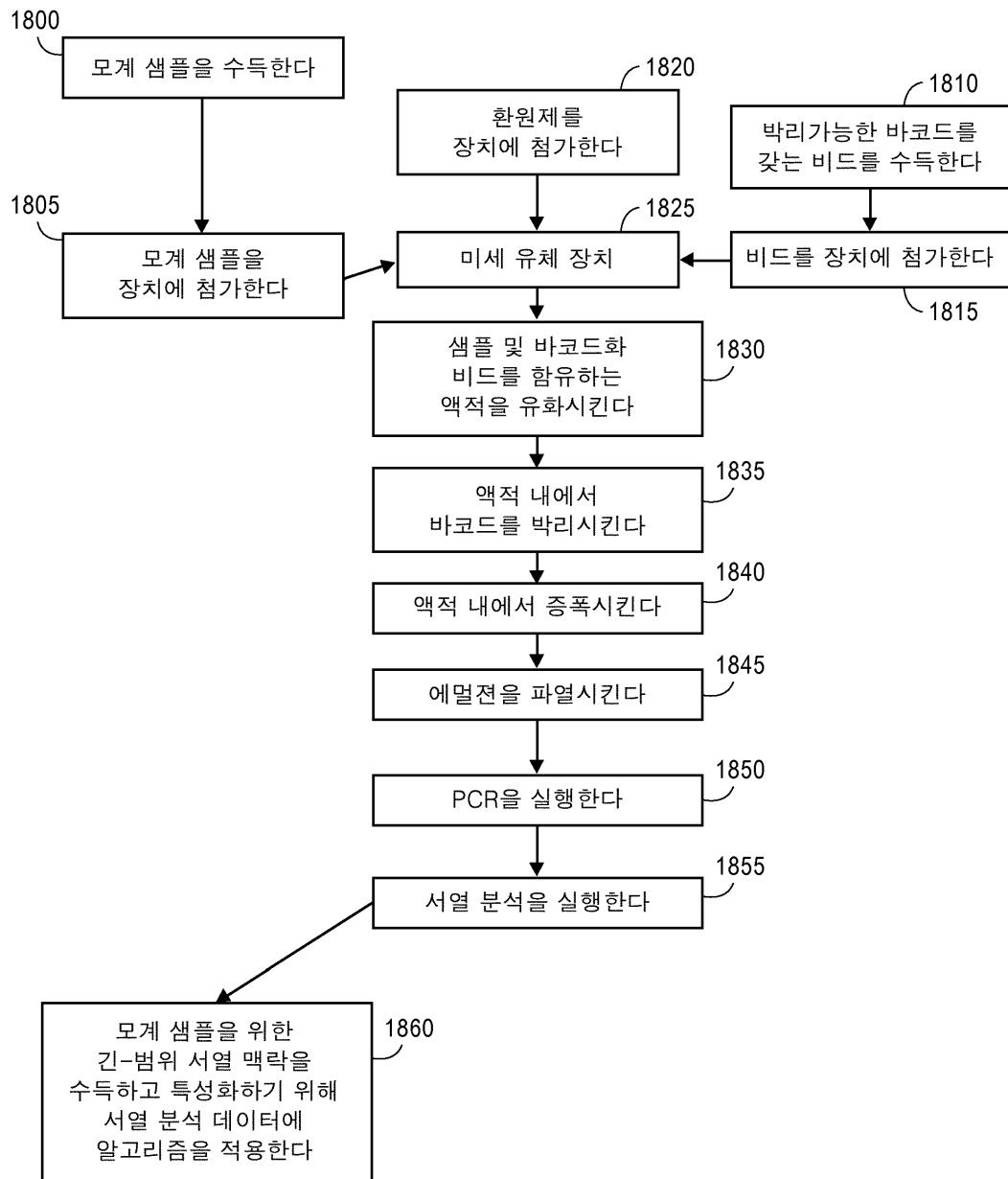
야생형



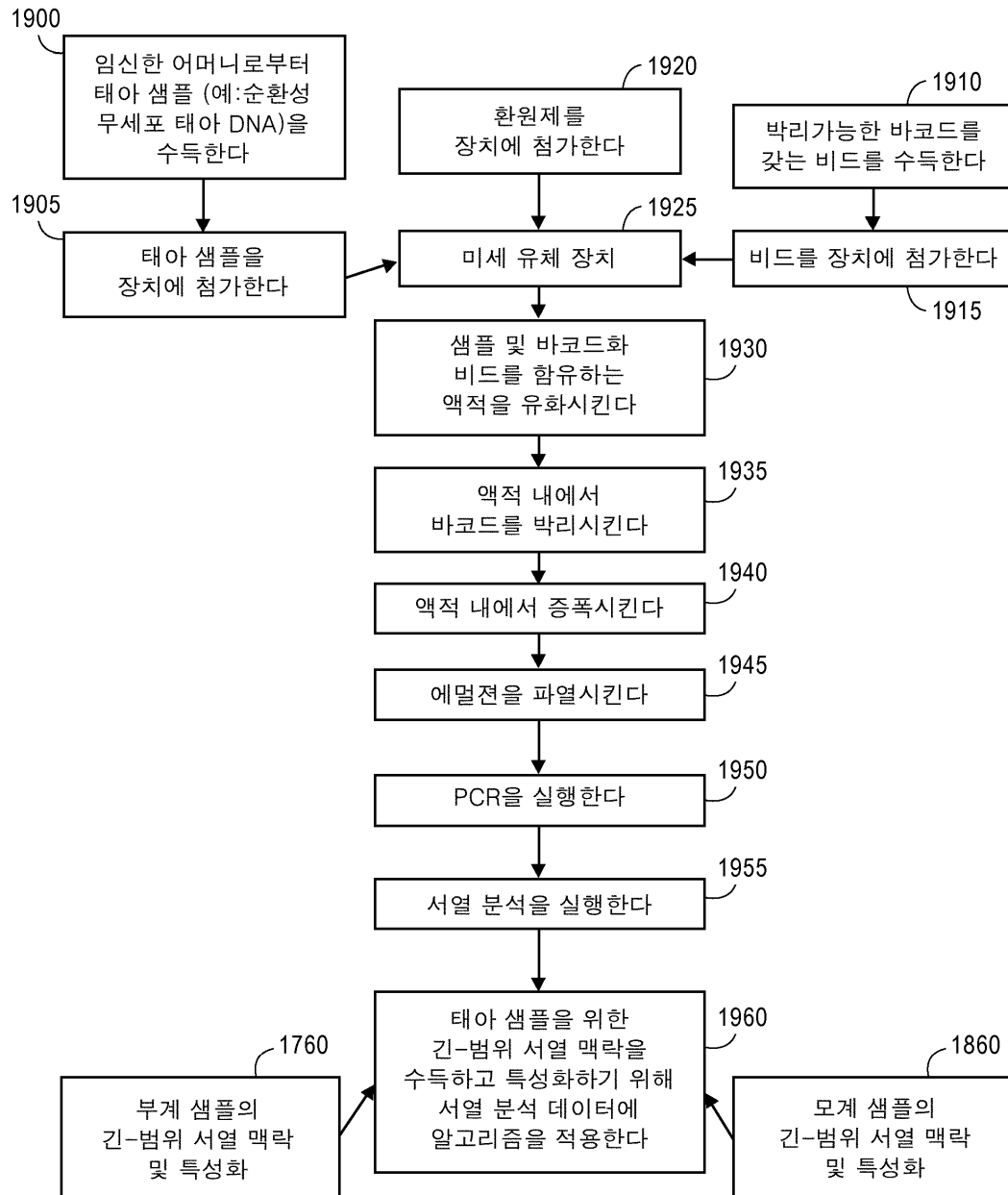
도면17



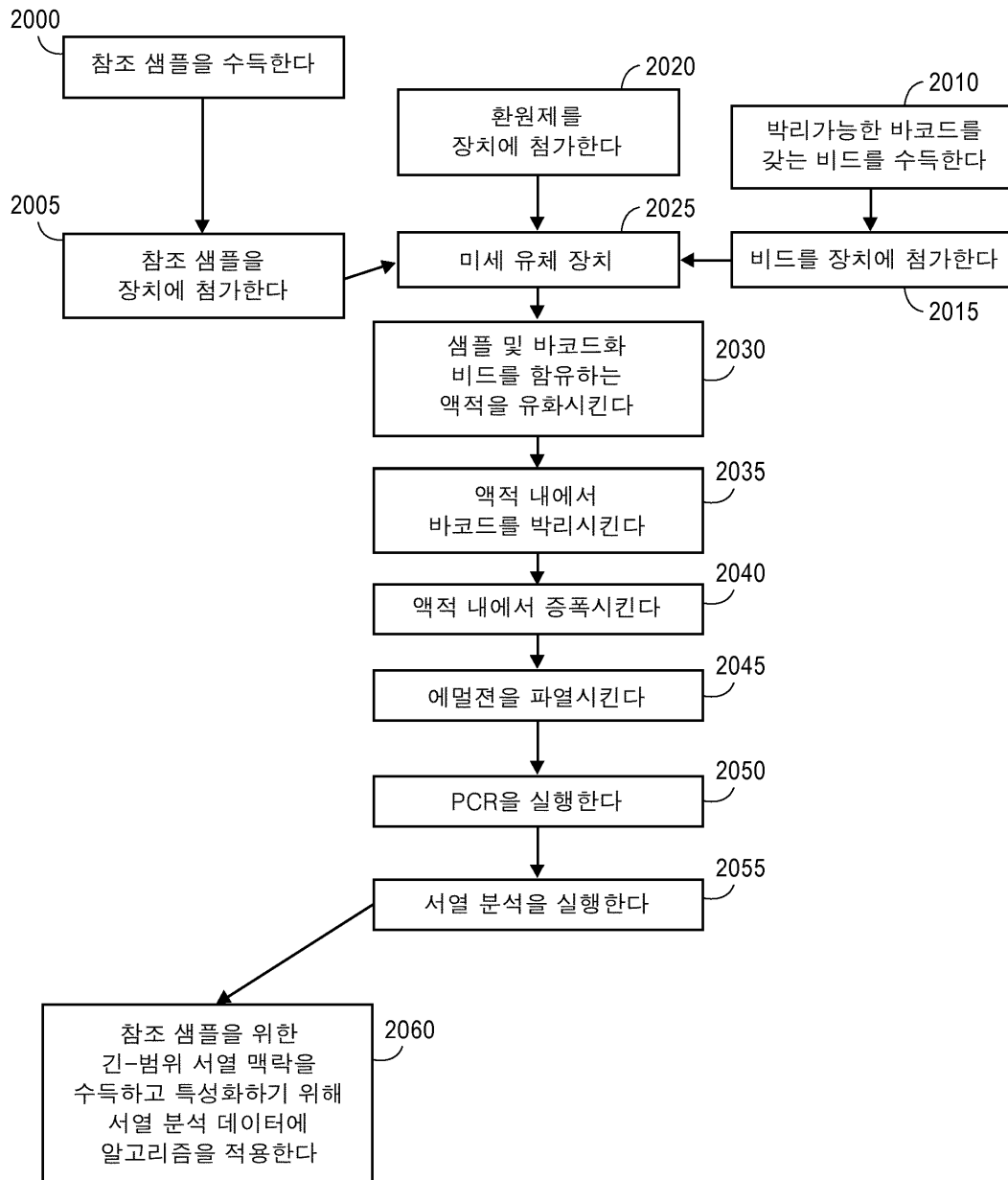
도면18



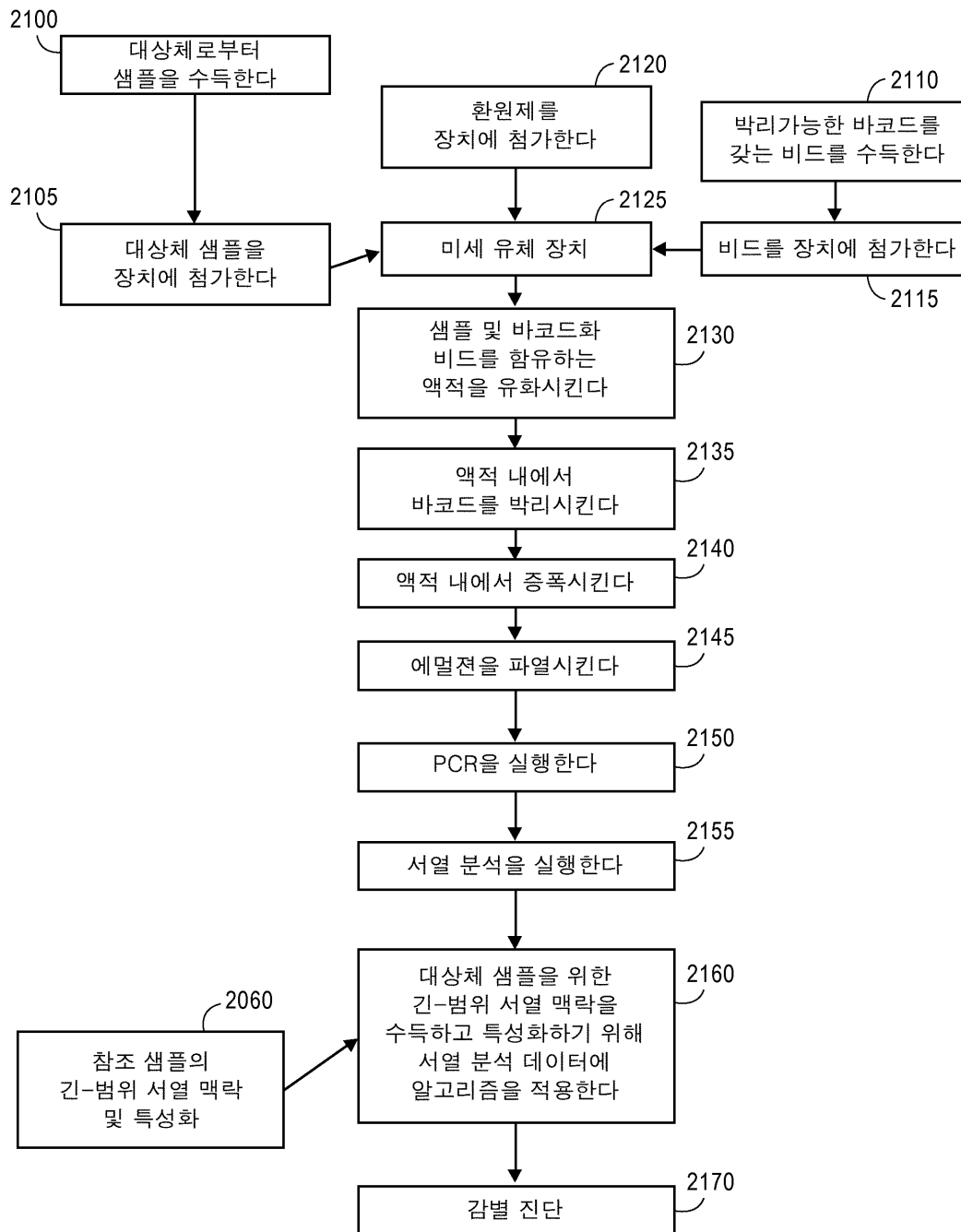
도면19



도면20



도면21



도면22

