(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2017-0073667
(43) 공개일자 2017년06월28일

(51) 국제특허분류(Int. Cl.)
C12Q 1/68 (2006.01)
(52) CPC특허분류
C12Q 1/6806 (2013.01)
C12Q 1/6837 (2013.01)
(21) 출원번호 10-2017-7013904
(22) 출원일자(국제) 2015년10월29일
심사청구일자 없음
(85) 번역문제출일자 2017년05월23일
(86) 국제출원번호 PCT/US2015/058142
(87) 국제공개번호 WO 2016/069939
국제공개일자 2016년05월06일
(30) 우선권주장
62/072,164 2014년10월29일 미국(US)

(71) 출원인
10엑스 제노믹스, 인크.
미국, 캘리포니아 94566, 플레젠튼, 스위트 401, 7068 콜 센터 파크웨이
(72) 발명자
자로즈 미르나
미국 94566 캘리포니아주 플레전턴 스위트 401 콜 센터 파크웨이 7068
슈날-레빈 마이클
미국 94566 캘리포니아주 플레전턴 스위트 401 콜 센터 파크웨이 7068
(뒷면에 계속)
(74) 대리인
유미특허법인

전체 청구항 수 : 총 95 항

(54) 발명의 명칭 **표적화 핵산 서열 분석을 위한 방법 및 조성물**

(57) 요 약

본 발명은 게놈의 표적화 영역에 함유된 서열 정보를 포획하고 분석하기 위한 방법, 조성물 및 시스템에 관한 것이다. 상기 표적화된 영역은 엑솜, 부분적 엑솜, 인트론, 엑손 및 인트론 영역의 조합, 유전자, 유전자 패널 및 목적할 수 있는 전체 게놈의 임의의 다른 서브세트를 포함할 수 있다.

(52) CPC특허분류
C12Q 1/6874 (2013.01)
C12Q 2535/122 (2013.01)
C12Q 2537/159 (2013.01)
C12Q 2563/179 (2013.01)
C12Q 2565/514 (2013.01)
(72) 발명자
삭소노프 세르게
미국 94566 캘리포니아주 플레전턴 스위트 401 콜 센터 파크웨이 7068
힌드슨 벤자민
미국 94566 캘리포니아주 플레전턴 스위트 401 콜 센터 파크웨이 7068
정 신잉
미국 94566 캘리포니아주 플레전턴 스위트 401 콜 센터 파크웨이 7068

명 세 서

청구범위

청구항 1

게놈의 하나 이상의 선택된 부분을 서열 분석하기 위한 방법으로서,

(a) 출발 게놈 물질을 제공하는 단계;

(b) 상기 출발 게놈 물질 기원의 개별 핵산 분자를 별개의 파티션에 분포시켜 각각의 별개의 파티션이 제1 개별 핵산 분자를 함유하도록 하는 단계;

(c) 상기 개별 파티션에서 개별 핵산 분자를 단편화시켜 다수의 단편을 형성하고(여기서, 상기 단편 각각은 바코드를 추가로 포함하고, 각각의 소정의 별개의 파티션 내 단편들은 통상의 바코드를 포함한다) 각각의 단편을 이것이 유래된 개별 핵산 분자와 관련시키는 단계;

(d) 게놈의 하나 이상의 선택된 부분의 적어도 일부를 포함하는 단편에 대해 집단을 집적시키는 단계;

(e) 상기 집단으로부터 서열 정보를 수득하고 게놈의 하나 이상의 선택된 부분을서열 분석하는 단계를 포함하는, 방법.

청구항 2

청구항 1에 있어서, 상기 제공 단계 (d)가

(i) 게놈의 하나 이상의 선택된 부분 내 또는 근처에서의 영역에 상보적인 프로브를 상기 단편에 하이브리드화하여 프로브-단편 복합체를 형성하는 단계;

(ii) 상기 프로브-단편 복합체를 고형 지지체의 표면에 포획시켜

상기 게놈의 하나이상의 선택된 부분의 적어도 일부를 포함하는 단편을 갖는 집단을 집적시키는 단계를 포함하는, 방법.

청구항 3

청구항 2에 있어서, 상기 고형 지지체가 비드를 포함하는, 방법.

청구항 4

청구항 2 또는 3에 있어서, 상기 프로브가 결합 모이어티를 포함하고 상기 표면이 포획 모이어티를 포함하고, 상기 프로브-단편 복합체가 상기 결합 모이어티와 상기 포획 모이어티 간의 반응을 통해 표면 상으로 포획되는, 방법.

청구항 5

청구항 4에 있어서, 상기 포획 모이어티가 스트렙타비딘을 포함하고 상기 결합 모이어티가 비오틴을 포함하는, 방법.

청구항 6

청구항 4에 있어서, 상기 포획 모이어티가 스트렙타비딘 자기 비드를 포함하고 상기 결합 모이어티가 비오티닐화된 RNA 라이브러리 베이트를 포함하는, 방법.

청구항 7

청구항 4에 있어서, 상기 포획 모이어티가 하기 포획으로 이루어진 그룹으로부터 선택된 구성원으로 지시되는, 방법: 전체 또는 부분적 엑솜 포획, 패널 포획, 표적화 엑손 포획, 앵커된 엑솜 포획 및 타일화된 게놈 영역 포획.

청구항 8

청구항 1 내지 7 중 어느 한 청구항에 있어서, 상기 수득 단계 (e) 전에, 상기 단편이증폭 생성물을 형성하도록 증폭되는, 방법.

청구항 9

청구항 8에 있어서, 상기 증폭 생성물이 부분적 또는 완전한 헤어핀 구조를 형성할 수 있는, 방법.

청구항 10

청구항 1 내지 9 중 어느 한 청구항에 있어서, 상기 수득 단계 (e)가 하기 반응으로 이루어진 그룹으로부터 선택되는 서열 분석 반응을 포함하는, 방법: 상기 서열 분석 반응이 짧은 판독-길이 서열 분석 반응 및 긴 판독-길이 서열 분석 반응.

청구항 11

청구항 10에 있어서, 상기 서열 분석 반응이 짧은 판독, 높은 정확도의 서열 분석 반응인, 방법.

청구항 12

청구항 1 내지 11 중 어느 한 청구항에 있어서, 상기 수득 단계 (e)가 90% 미만의 출발 게놈 물질에 대한 서열 정보를 제공하는, 방법.

청구항 13

청구항 1 내지 11 중 어느 한 청구항에 있어서, 상기 수득 단계 (e)가 75% 미만의 출발 게놈 물질에 대한 서열 정보를 제공하는, 방법.

청구항 14

청구항 1 내지 11 중 어느 한 청구항에 있어서, 상기 수득 단계 (e)가 50% 미만의 출발 게놈 물질에 대한 서열 정보를 제공하는, 방법.

청구항 15

청구항 1 내지 14 중 어느 한 청구항에 있어서, 상기 방법이 단리된 단편의 중첩서열을 기준으로 추론된 콘티그에서 2개 이상의 개별 핵산 분자를 연결함을 추가로 포함하고, 상기 추론된 콘티그가 적어도 10 kb의 길이 N50을 포함하는, 방법.

청구항 16

청구항 15에 있어서, 상기 추론된 콘티그가 적어도 20 kb의 길이 N50을 포함하는, 방법.

청구항 17

청구항 15에 있어서, 상기 추론된 콘티그가 적어도 40 kb의 길이 N50을 포함하는, 방법.

청구항 18

청구항 15에 있어서, 상기 추론된 콘티그가 적어도 50 kb의 길이 N50을 포함하는, 방법.

청구항 19

청구항 15에 있어서, 상기 추론된 콘티그가 적어도 100 kb의 길이 N50을 포함하는, 방법.

청구항 20

청구항 15에 있어서, 상기 추론된 콘티그가 적어도 200 kb의 길이 N50을 포함하는, 방법.

청구항 21

청구항 1 내지 14 중 어느 한 청구항에 있어서, 상기 방법이 상기 단리된 단편들의 서열 내 중첩 단계화된 변이체를 기준으로 상 블록에서 2개 이상의 개별 핵산 분자를 연결함을 추가로 포함하고 상기 상 블록이 적어도 10

kb의 길이 N50을 포함하는, 방법.

청구항 22

청구항 21에 있어서, 상기 단계 블록이 적어도 20 kb의 길이 N50을 포함하는, 방법.

청구항 23

청구항 21에 있어서, 상기 단계 블록이 적어도 40 kb의 길이 N50을 포함하는, 방법.

청구항 24

청구항 18에 있어서, 상기 단계 블록이 적어도 50 kb의 길이 N50을 포함하는, 방법.

청구항 25

청구항 21에 있어서, 상기 단계 블록이 적어도 100 kb의 길이 N50을 포함하는, 방법.

청구항 26

청구항 21에 있어서, 상기 단계 블록이 적어도 200 kb의 길이 N50을 포함하는, 방법.

청구항 27

청구항 1 내지 26 중 어느 한 청구항에 있어서, 상기 게놈의 선택된 부분이 엑솜을 포함하는, 방법.

청구항 28

청구항 1 내지 26 중 어느 한 청구항에 있어서, 각각의 별개의 파티션 내 상기 개별핵산 분자가 단일 세포 기원의 게놈 DNA를 포함하는, 방법.

청구항 29

청구항 1 내지 28 중 어느 한 청구항에 있어서, 각각의 별개의 파티션이 상이한 염색체 기원의 게놈 DNA를 포함하는, 방법.

청구항 30

청구항 1 내지 29 중 어느 한 청구항에 있어서, 상기 별개의 파티션이 에멀젼 내소적을 포함하는, 방법.

청구항 31

청구항 1 내지 30 중 어느 한 청구항에 있어서, 상기 단편에 부착된 바코드가 적어도 700,000개 바코드 라이브러리로부터 기원하는, 방법.

청구항 32

청구항 1 내지 31 중 어느 한 청구항에 있어서, 상기 바코드가 추가의 서열 분절을 추가로 포함하는, 방법.

청구항 33

청구항 32에 있어서, 상기 추가의 서열 분절이 하기 요소로 이루어진 그룹으로부터 선택되는 하나 이상의 구성원을 포함하는, 방법: 프라이머, 부착 서열, 무작위 n-량체 올리고뉴클레오타이드, 우라실 뉴클레오염기.

청구항 34

게놈 샘플의 하나 이상의 표적화 부분으로부터 서열 정보를 수득하는 방법으로서, 상기 방법은,

(a) 별개의 파티션에 게놈 샘플의 개별 제1 핵산 단편 분자를 제공하는 단계;

(b) 상기 별개의 파티션 내 개별 제1 핵산 단편 분자를 단편화시켜 상기 개별 제1 핵산 단편 분자 각각으로부터 다수의 제2 단편을 생성시키는 단계;

(c) 통상의 바코드 서열을 별개의 파티션 내 다수의 제2 단편에 부착시켜 다수의제2 단편 각각이 이들이 함유된

별개의 파티션에 기인할 수 있도록 하는 단계;

(d) 상기 게놈 샘플의 하나 이상의 표적화 부분으로 지시된 프로브 라이브러리를 제2 단편에 적용하는 단계;

(e) 프로브 라이브러리에 하이브리드화된 다수의 제2 단편의 서열을 동정하기 위해 서열 분석 반응을 수행하여 상기 게놈 샘플의 하나 이상의 표적화 부분으로부터 서열 정보를 수득하는 단계를 포함하는, 방법.

청구항 35

청구항 34에 있어서, 상기 프로브 라이브러리가 결합 모이어티에 부착되고 상기수행 단계 (e)전에, 상기 제2 단편이 결합 모이어티와 포획 분자 간의 반응을 통해 포획 모이어티를 포함하는 표면 상에 포획되는, 방법.

청구항 36

청구항 35에 있어서, 상기 수행 단계 (e) 전에, 상기 제2 단편이 표면상에 포획되기 전 또는 후에 상기 제2 단편이 증폭되는, 방법.

청구항 37

청구항 35 또는 36에 있어서, 상기 결합 모이어티가 비오틴을 포함하고 상기 포획 모이어티가 스트렙타비딘을 포함하는, 방법.

청구항 38

청구항 34 내지 37 중 어느 한 청구항에 있어서, 하기 반응으로 이루어진 그룹으로부터 선택되는 구성원인, 방법: 상기 서열 분석 반응이 짧은 판독-길이 서열 분석 반응 및 긴 판독-길이 서열 분석 반응.

청구항 39

청구항 34 내지 38 중 어느 한 청구항에 있어서, 상기 서열 분석 반응이 짧은 판독의 높은 정확도의 서열 분석 반응인, 방법.

청구항 40

청구항 34 내지 39 중 어느 한 청구항에 있어서, 상기 방법이 다수의 제2 단편의중첩 서열을 기준으로 추론된 콘티그 내 2개 이상의 개별 단편 분자를 연결시킴을 추가로 포함하고, 상기 추론된 콘티그가 적어도 10 kb의 길이 N50을 포함하는, 방법.

청구항 41

청구항 40에 있어서, 상기 추론된 콘티그가 적어도 20 kb의 길이 N50을 포함하는, 방법.

청구항 42

청구항 40에 있어서, 상기 추론된 콘티그가 적어도 40 kb의 길이 N50을 포함하는, 방법.

청구항 43

청구항 40에 있어서, 상기 추론된 콘티그가 적어도 50 kb의 길이 N50을 포함하는, 방법.

청구항 44

청구항 40에 있어서, 상기 추론된 콘티그가 적어도 100 kb의 길이 N50을 포함하는, 방법.

청구항 45

청구항 40에 있어서, 상기 추론된 콘티그가 적어도 200 kb의 길이 N50을 포함하는, 방법.

청구항 46

청구항 34 내지 39 중 어느 한 청구항에 있어서, 상기 방법이 다수의 제2 단편의서열 내 중첩 단계화된 변이체를 기준으로 단계 블록으로 2개 이상의 다수의 개별 핵산 단편 분자를 연결시킴을 추가로 포함하고, 상기 단계

블록이 적어도 10 kb의 길이 N50을 포함하는, 방법.

청구항 47

청구항 46에 있어서, 상기 단계 블록이 적어도 20 kb의 길이 N50을 포함하는, 방법.

청구항 48

청구항 46에 있어서, 상기 단계 블록이 적어도 40 kb의 길이 N50을 포함하는, 방법.

청구항 49

청구항 46에 있어서, 상기 단계 블록이 적어도 50 kb의 길이 N50을 포함하는, 방법.

청구항 50

청구항 46에 있어서, 상기 단계 블록이 적어도 100 kb의 길이 N50을 포함하는, 방법.

청구항 51

청구항 46에 있어서, 상기 단계 블록이 적어도 200 kb의 길이 N50을 포함하는, 방법.

청구항 52

청구항 34 내지 51 중 어느 한 청구항에 있어서, 상기 게놈 샘플의 표적화 부분이 엑솜을 포함하는, 방법.

청구항 53

청구항 34 내지 51 중 어느 한 청구항에 있어서, 각각의 별개의 파티션 내 게놈샘플이 단일 세포 기원의 게놈 DNA를 포함하는, 방법.

청구항 54

청구항 34 내지 51 중 어느 한 청구항에 있어서, 각각의 별개의 파티션이 상이한 염색체 기원의 게놈 DNA를 포함하는, 방법.

청구항 55

청구항 34 내지 54 중 어느 한 청구항에 있어서, 상기 별개의 파티션이 에멀젼 중 소적을 포함하는, 방법.

청구항 56

청구항 34 내지 55 중 어느 한 청구항에 있어서, 상기 제2 단편에 부착된 상기 바코드 서열이 적어도 700,000개의 바코드 라이브러리로부터 기원하는, 방법.

청구항 57

청구항 34 내지 56 중 어느 한 청구항에 있어서, 상기 바코드가 추가의 서열 분절을 추가로 포함하는, 방법.

청구항 58

청구항 57에 있어서, 상기 추가의 서열 분절이 하기 요소를 포함하는 올리고뉴클레오타이드로 이루어진 그룹으로부터 선택되는 하나 이상의 구성원을 포함하는, 방법: 프라이머, 부착 서열, 무작위 n-량체 올리고뉴클레오타이드, 우라실 뉴클레오염기.

청구항 59

청구항 36 내지 58 중 어느 한 청구항에 있어서, 상기 제2 단편이, 수득한 증폭 생성물이 부분적이거나 완전한 헤어핀 구조를 형성할 수 있도록 증폭되는, 방법.

청구항 60

분자 형태를 보유하면서 게놈 샘플의 하나 이상의 표적화 영역으로부터 서열 정보를 수득하기 위한 방법으로서,

상기 방법은,

(a) 출발 게놈 물질을 제공하는 단계;

(b) 상기 출발 게놈 물질 기원의 개별 핵산 분자를 별개의 파티션에 분포시켜 각각의 별개의 파티션이 제1 개별 핵산 분자를 함유하도록 하는 단계;

(c) 별개의 파티션 중 제1 개별 핵산 분자를 단편화시켜 다수의 단편을 형성하는단계;

(d) 게놈의 하나 이상의 선택된 부분의 적어도 일부를 포함하는 단편에 대해 집단을 집적시키는 단계;

(e) 상기 집단으로부터 서열 정보를 수득하여 분자 형태를 보유하면서 게놈 샘플의 하나 이상의 표적화 부분을 서열 분석하는 단계를 포함하는, 방법.

청구항 61

청구항 60에 있어서, 상기 수득 단계 (e) 전에, 다수의 단편이 바코드로 태깅되어각각의 단편을 이것이 형성되는 별개의 파티션과 관련시키는, 방법.

청구항 62

청구항 60에 있어서, 단계 (b)에서 상기 개별 핵산 분자가 분포되어 각각의 제1 개별 핵산 분자의 분자 형태가 유지되도록 하는, 방법.

청구항 63

게놈 샘플의 하나 이상의 표적화 부분으로부터 서열 정보를 수득하는 방법으로서, 상기 방법은

(a) 별개의 파티션에 상기 게놈 샘플의 개별 핵산 분자를 제공하는 단계;

(b) 별개의 파티션에서 개별 핵산 분자를 단편화시켜 다수의 단편을 형성하여(여기서, 단편 각각은 바코드를 추가로 포함하고 각각의 소정의 별개의 파티션 내 단편들은 통상의 바코드를 포함한다) 각각의 단편을 이것이 유래된 개별 핵산 분자와 관련시키는 단계;

(c) 게놈 샘플의 하나 이상의 표적화 부분으로 지시된 프로브 라이브러리를 다수의 단편에 적용하는 단계(여기서, 프로브 라이브러리 중 적어도 다수의 프로브는 정통한 단일 뉴클레오타이드 다형태(SNP)에 하이브리드화하도록 디자인된다);

(d) 프로브 라이브러리에 하이브리드화하는 다수의 단편의 서열을 동정하기 위해서열 분석 반응을 수행하여 상기 게놈 샘플의 하나 이상의 표적화 부분으로부터 서열 정보를 수득하는 단계를 포함하는, 방법.

청구항 64

청구항 63에 있어서, 상기 프로브 라이브러리 중 약 80% 내지 99%의 프로브가 정통한 SNP에 하이브리드화하도록 디자인된, 방법.

청구항 65

청구항 63에 있어서, 상기 프로브 라이브러리 중 약 65% 내지 85%의 프로브가 정통한 SNP에 하이브리드화하도록 디자인된, 방법.

청구항 66

청구항 63에 있어서, 상기 프로브 라이브러리 중 약 70% 내지 80%의 프로브가 정통한 SNP에 하이브리드화하도록 디자인된, 방법.

청구항 67

청구항 63에 있어서, 상기 프로브 라이브러리 중 적어도 65%의 프로브가 정통한SNP에 하이브리드화하도록 디자인된, 방법.

청구항 68

청구항 63에 있어서, 상기 프로브 라이브러리 중 적어도 75%의 프로브가 정통한 SNP에 하이브리드화하도록 디자인된, 방법.

청구항 69

청구항 63에 있어서, 상기 프로브 라이브러리 중 적어도 85%의 프로브가 정통한 SNP에 하이브리드화하도록 디자인된, 방법.

청구항 70

청구항 63에 있어서, 상기 프로브 라이브러리 중 적어도 90%의 프로브가 정통한 SNP에 하이브리드화하도록 디자인된, 방법.

청구항 71

청구항 63 내지 70 중 어느 한 청구항에 있어서, 상기 정통한 SNP가 상기 게놈 샘플의 표적화 영역에서 엑손 및 인트론 둘다 내에 위치하는, 방법.

청구항 72

청구항 63 내지 70 중 어느 한 청구항에 있어서, 상기 프로브 라이브러리 중 대다수의 프로브가 약 1 킬로베이스 내지 약 15 킬로베이스 (kb)에 의해 격리되어 있는 정통한 SNP에 하이브리드화하도록 추가로 디자인된, 방법.

청구항 73

청구항 63 내지 70 중 어느 한 청구항에 있어서, 상기 프로브 라이브러리 중 대다수의프로브가 약 5kb 내지 약 10 kb로 격리된 정통한 SNP에 하이브리드화하도록 디자인된, 방법.

청구항 74

청구항 63 내지 70 중 어느 한 청구항에 있어서, 상기 프로브 라이브러리 중 대다수의프로브가 약 3kb 내지 약 6kb로 격리된 정통한 SNP에 하이브리드화하도록 추가로 디자인된, 방법.

청구항 75

청구항 63 내지 70 중 어느 한 청구항에 있어서, 상기 프로브 라이브러리 중 대다수의프로브가 약 1kb로 격리된 정통한 SNP에 하이브리드화하도록 추가로 디자인된, 방법.

청구항 76

청구항 63 내지 70 중 어느 한 청구항에 있어서, 상기 프로브 라이브러리 중 대다수의프로브가 약 3kb로 격리된 정통한 SNP에 하이브리드화하도록 추가로 디자인된, 방법.

청구항 77

청구항 63 내지 70 중 어느 한 청구항에 있어서, 상기 프로브 라이브러리 중 대다수의프로브가 약 10 kb로 격리된 정통한 SNP에 하이브리드화하도록 추가로 디자인된, 방법.

청구항 78

청구항 63 내지 77 중 어느 한 청구항에 있어서, 상기 프로브 라이브러리 내 다수의 프가 임의의 조합으로 하기의 조건 중 하나 이상에 충족하도록 추가로 디자인된, 방법:

(i) 엑손과 인트론 간의 경계선 10 내지 50 kb 내 정통한 SNP가 없는 게놈 샘플의 표적화 부분에 대해, 다수의 프로브가 상기 경계선 기원의 인트론 내 정통한 SNP에 하이브리드화하도록 디자인되는 것;

(ii) 엑손 내 제1 정통한 SNP이고, 상기 제1 정통한 SNP가 인접한 인트론과의 경계선으로부터 10 내지 50 kb에 위치하고 제2 정통한 SNP가 인접한 인트론 내에 있고 상기 제2 정통한 SNP가 경계선으로부터 10 내지 50 kb에 위치한 게놈 샘플의 표적화 부분에 대해, 다수의 프로브가 제1 및 제2 정통한 SNP 간의 게놈 샘플 영역에 하이

브리드화하도록 디자인되는 것;

(iii) 적어도 10 내지 50 kb에 대해 어떠한 정통한 SNP를 포함하지 않는 게놈 샘플의 표적화 부분에 대해, 다수의 프로브가 게놈 샘플의 표적화 부분으로 0.5, 1, 3, 또는 5 kb 마다 하이브리드화하도록 다지인되는 것;

(iv) 엑손과 인트론 간의 경계선 10 내지 50 kb 내에 어떠한 정통한 SNP가 없는게놈 샘플의 표적화 부분에 대해, 다수의 프로브가 엑손-인트론 경계선 다음에 가장 인접한 정통한 SNP에 하이브리드화하도록 디자인되는 것.

청구항 79

청구항 63 내지 78 중 어느 한 청구항에 있어서, 상기 프로브 라이브러리가 바코드를 거쳐 연결 정보를 제공하는 밀도로 엑손을 플랭킹하는 게놈 샘플의 영역에 하이브리드화하도록 디자인된 프로브를 포함하는, 방법.

청구항 80

청구항 63 내지 79 중 어느 한 청구항에 있어서, 상기 프로브 라이브러리에 의해 나타낸 커버 범위가 별개의 파티션 내 게놈 샘플의 개별 핵산 단편 분자의 길이 분포에 역비례하여 보다 높은 비율의 보다 긴 개별 핵산 단편 분자를 함유하는 방법이 보다 작은 범위의 커버와 함께 프로브 라이브러리를 사용하는, 방법.

청구항 81

청구항 63 내지 80 중 어느 한 청구항에 있어서, 상기 프로브 라이브러리가 게놈 샘플의 표적화 부분의 커버를 위해 최적화되고 상기 게놈 샘플의 표적화 부분이 높은 맵 품질 영역을 포함하는, 방법.

청구항 82

청구항 63 내지 80 중 어느 한 청구항에 있어서, 상기 프로브 라이브러리가 게놈샘플의 하나 이상의 표적화 부분을 특징으로 하는 알려진 특징을 가져:

(i) 높은 맵 품질을 갖는 표적화 부분에 대해, 프로브 라이브러리가 엑손과 인트론의 경계선의 1kb-1메가베이스(Mb) 내 정통한 SNP에 하이브리드화하는 프로브를 포함하고;

(ii) 바코드화된 단편의 길이의 분포가 약 250 kb 보다 긴 높은 비율의 단편을 갖는 표적화 부분에 대해, 프로브 라이브러리는 적어도 50 kb에 의해 분리된 정통한 SNP에 하이브리드화하는 프로브를 포함하고;

(iii) 낮은 맵 품질을 갖는 표적화 부분에 대해, 프로브 라이브러리는 1kb의 엑손-인트론 경계선 내 정통한 SNP에 하이브리드화하는 프로브 및 엑손 내 및 인트론 내 정통한 SNP에 하이브리드화하는 프로브를 포함하고;

(iv) 유전자 사이 영역을 포함하는 표적화 부분에 대해, 프로브 라이브러리는 적어도 2kb의 거리로 격리된 정통한 SNP에 하이브리드화하는 프로브를 포함하는, 방법.

청구항 83

청구항 63 내지 82 중 어느 한 청구항에 있어서, 단계 (d)에서 수득된 서열 정보가 하기 요소로 이루어진 그룹의 하나 이상의 구성에 대한 정보를 포함하는, 방법: 유전자 융합, 카피수 변화, 삽입및 결실.

청구항 84

게놈 샘플의 하나 이상의 표적화 부분으로부터 서열 정보를 수득하는 방법으로서, 상기 방법은,

(a) 상기 게놈 샘플의 개별 핵산 분자를 제공하는 단계;

(b) 개별 핵산 분자를 단편화하여 다수의 단편을 형성하여 (여기서, 단편 각각은바코드를 추가로 포함하고, 동일한 개별 핵산 분자 기원의 단편은 통상의 바코드를 포함한다) 각각의 단편을 이것이 유래된 개별 핵산 분자와 관련시키는 단계;

(c) 상기 게놈 샘플의 하나 이상의 표적화 부분을 함유하는 단편에 대해 다수의단편을 집적시키는 단계;

(d) 집적된 다수의 단편을 동정하기 위해 서열 분석 반응을 수행하여 상기 게놈샘플의 하나 이상의 표적화 부분으로부터 서열 정보를 수득하는 단계를 포함하는, 방법.

청구항 85

청구항 84에 있어서, 상기 바코드가 상기 단편화 단계 (b) 전에 개별 핵산 분자에 부가되는, 방법.

청구항 86

청구항 85에 있어서, 상기 바코드가 트랜스포존을 사용하여 개별 핵산 분자에 부가되는, 방법.

청구항 87

청구항 84에 있어서, 바코드가 단편화와 동시에 부가되는, 방법.

청구항 88

청구항 87에 있어서, 상기 단편화가 증폭 단계를 포함하는, 방법.

청구항 89

청구항 84 내지 88 중 어느 한 청구항에 있어서, 상기 집적 단계 (c)가 상기 게놈 샘플의 하나 이상의 표적화 부분으로 지시된 프로브 라이브러리를 적용함을 포함하는, 방법.

청구항 90

청구항 89에 있어서, 상기 프로브 라이브러리가 결합 모이어티에 부착되고 상기수행 단계 (d) 전에, 상기 단편이 결합 모이어티와 포획 모이어티 간의 반응을 통해 포획되는, 방법.

청구항 91

청구항 90에 있어서, 상기 결합 모이어티와 포획 모이어티 간의 반응이 표면 상에 단편을 고정화시키는, 방법.

청구항 92

청구항 90 또는 91에 있어서, 상기 결합 모이어티가 비오틴을 포함하고 상기 포획 모이어티가 스트렙타비딘을 포함하는, 방법.

청구항 93

청구항 84 내지 92 중 어느 한 청구항에 있어서, 하기 반응으로 이루어진 그룹으로부터 선택되는 구성원인, 방법: 상기 서열 분석 반응이 짧은 판독-길이 서열 분석 반응 및 긴 판독-길이 서열 분석 반응.

청구항 94

청구항 84 내지 92 중 어느 한 청구항에 있어서, 상기 서열 분석 반응이 짧은 판독의 높은 정확도의 서열 분석 반응인, 방법.

청구항 95

청구항 84 내지 94 중 어느 한 청구항에 있어서, 상기 단편이, 수득한 증폭 생성물이 부분적이거나 완전한 헤어핀 구조를 형성할 수 있도록 증폭되는, 방법.

발명의 설명

기 술 분 야

[0001] 본 출원은 하기 미국 우선권을 주장한다. 가출원 번호 제62/072,164호, 2014년 10월 29일자로 출원되었고 이는 모든 목적을 위해 이의 전문이 본원에 참조로 인용된다.

배 경 기 술

[0002] 게놈을 정확하고 신속하게 서열 분석하는 능력은 혁신적인 생물학 및 의료학이다. 복잡한 게놈의 연구, 및 특히 인간에서 질환의 유전학적 근거에 대한 연구는 대규모의 유전학적 분석을 포함한다. 전체 게놈 수준에 대한 상기 유전학적 분석은 금전상으로 고가일 뿐만 아니라 시간 및 노동 소모적이다. 이들 비용은 별도의 개별 DNA 샘

플의 분석을 포함하는 프로토콜과 함께 증가한다. 질환 발병과 연계된 게놈에서 다형성 영역의 서열 분석(및 재-서열분석)은 암 및 치료제 개발과 같은 질환의 이해에 크게 기여하고 약물 반응에서 변동성과 관련된 유전자 및 기능적 다형성을 동정하기 위한 약리게놈학적 과제에 충족하도록 도와준다. 통계학적으로 유의적인 데이터를 산출하기에 충분한 대형 집단에 대해 수행된 수많은 유전학적 마커에 대한 스크리닝은 소정의 유전자형과 특정 질환 간의 관련성을 수행하기 전에 요구된다.

[0003] 대규모의 게놈 분석의 이득을 보유하면서 게놈 서열 분석과 관련된 비용을 감소시키기 위한 한가지 방법은 게놈의 표적화 영역에 대한 고속 처리, 높은 정확도의 서열 분석을 수행하는 것이다. 광범위하게 사용된 방법은 인간 게놈의 약 1%를 차지하고 임상 및 기본 연구에서 통상의 기술이 된 게놈의 전체 단백질 암호화 영역(엑솜)의 많은 부분을 점유한다. 엑솜 서열 분석은 전체 게놈 서열 분석에 대해 이점을 제공한다: 이것은 유의적으로 덜 고비용이고, 기능 해석을 위해 보다 용이하게 이해되고, 분석을 유의적으로 신속하게 하고 매우 깊숙한 서열 분석이 가용해지도록 하고 관리하기가 보다 용이한 데이터 세트가 수득된다. 높은 정확도 및 고속 처리 서열 분석 및 유전학적 분석을 위해 목적하는 표적 영역의 집적을 위한 방법, 시스템 및 조성물에 대한 필요성이 존재한다.

발명의 내용

[0004] **발명의 요약**

[0005] 따라서, 본 발명은 게놈의 표적 영역에 대한 서열 정보를 수득하기 위한 방법, 시스템 및 조성물을 제공한다.

[0006] 일부 측면에서, 본 발명의 기재는 게놈의 하나 이상의 선택된 부분을 서열 분석하기 위한 방법을 제공하고, 상기 방법은 일반적으로 하기의 단계를 포함한다: (a) 출발 게놈 물질을 제공하는 단계, (b) 개별 핵산 분자를 출발 게놈 물질로부터 별개의 파티션(partition)으로 분포시켜 각각의 개별 파티션이 제1 개별 핵산 분자를 함유하도록 하는 단계; (c) 별개의 파티션에서 개별 핵산 분자를 단편화시켜 다수의 단편을 형성하는 단계로서 여기서 단편 각각이 바코드를 추가로 포함하고 소정의 별개의 파티션 내 단편 각각이 공통의 바코드를 포함하여 각각의 단편을 이로부터 유래된 개별 핵산 분자와 관련시키는 단계; (d) 상기 게놈의 하나 이상의 선택된 부분의 적어도 일부를 포함하는 단편을 위해 집적된 집단을 제공하는 단계; (e) 상기 집단으로부터 서열 정보를 수득하여 게놈의 하나 이상의 선택된 부분을 서열 분석하는 단계.

[0007] 추가의 구현예에서 그리고 상기에 따라, 상기 게놈의 하나 이상의 선택된 부분의 적어도 일부를 포함하는 단편에 대해 집적된 집단을 제공하는 것이 (i) 상기 게놈의 하나 이상의 선택된 부분에서 또는 이의 부근의 영역에 상보적인 프로브를 상기 단편에 하이브리드화시켜 프로브-단편 복합체를 형성하는 단계; 및 (ii) 프로브-단편 복합체를 고형 지지체의 표면에 포획하여 상기 게놈의 하나 이상의 선택된 부분의 적어도 일부를 포함하는 단편을 갖는 집단을 집적시키는 단계를 포함한다. 여전히 추가의 구현예에서, 고형 지지체는 비드를 포함한다. 여전히, 추가의 구현예에서, 상기 프로브는 결합 모이어티를 포함하고 상기 표면은 포획 모이어티를 포함하고 프로브-단편 복합체는 결합 모이어티와 포획 모이어티 간의 반응을 통해 표면상에 포획된다. 추가의 예에서, 포획 모이어티는 스트렙타비딘을 포함하고 결합 모이어티는 비오틴을 포함한다. 여전히 추가의 예에서, 포획 모이어티는 스트렙타비딘 자기 비드를 포함하고 결합 모이어티는 비오티닐화된 RNA 라이브러리 베이트(bait)를 포함한다.

[0008] 일부 구현예에서 그리고 상기 중 어느 하나에 따라, 본 발명의 방법은 전체 또는 부분적 엑솜 포획, 패널 포획, 표적화 엑손 포획, 앵커된 엑솜 포획 또는 타일화된 게놈 영역 포획에 지시되는 포획 모이어티의 용도를 포함한다.

[0009] 여전히 추가의 구현예에서 그리고 상기 중 어느 하나에 따라, 본원에 기재된 방법은 서열 분석 반응을 포함하는 수득 단계를 포함한다. 추가의 예에서, 서열 분석 반응은 짧은 판독-길이 서열 분석 반응 또는 긴 판독-길이 서열 분석 반응이다. 여전히 추가의 예에서, 서열 분석 반응은 90% 미만, 75% 미만, 또는 50% 미만의 출발 게놈 물질에 대한 서열 정보를 제공한다.

[0010] 여전히 추가의 구현예에서, 본원에 기재된 방법은 단리된 단편들의 중첩 서열을 기준으로 추론된 콘티그에서 개별 핵산 분자 중 2개 이상을 연결시킴을 추가로 포함하고, 여기서, 상기 추론된 콘티그는 적어도 10 kb, 20 kb, 40 kb, 50 kb, 100 kb, 또는 200 kb의 길이 N50을 포함한다.

[0011] 여전히 추가의 예에서 그리고 상기 중 어느 하나에 따라, 본원에 기재된 방법은 단리된 단편의 서열내 중첩 단계화된 변이체를 기준으로 단계(phase) 블록에서 개별 핵산 분자의 2개 이상을 연결시킴을 추가로 포함하고, 여

기서, 상기 단계 블록은 적어도 10 kb, 적어도 20 kb, 적어도 40 kb, 적어도 50 kb, 적어도 100 kb 또는 적어도 200 kb의 길이 N50을 포함한다.

[0012] 여전히 추가의 구현예에서 그리고 상기 중 어느 하나에 따라, 본원에 기재된 방법은 엑솜을 함께 커버하는 게놈의 선택된 부분으로부터의 서열 정보를 제공한다. 여전히 추가의 구현예에서, 별도의 파티션에서 개별 핵산 분자는 단일 세포 기원의 게놈 DNA를 포함한다. 여전히 추가의 구현예에서, 별개의 파티션 각각은 상이한 염색체 기원의 게놈 DNA를 포함한다.

[0013] 추가의 측면에서, 본원의 기재는 게놈 샘플의 하나 이상의 표적화 부분으로부터 서열 정보를 수득하는 방법을 제공한다. 상기 방법은 제한 없이 하기의 단계를 포함한다: (a) 별개의 파티션에서 게놈 샘플의 개별 제1 핵산 단편 분자를 제공하는 단계; (b) 별개의 파티션 내 개별의 제1 핵산 단편 분자를 단편화시켜 개별 제1 핵산 단편 분자들 각각으로부터 복수의 제2 단편을 생성하는 단계; (c) 통상의 바코드 서열을 별개의 파티션 내 복수의 제2 단편에 부착시켜 복수의 제2 단편들 각각이 이들이 함유된 별개의 파티션 때문일 수 있도록 하는 단계; (d) 게놈 샘플의 하나 이상의 표적화 부분으로 지시된 프로브 라이브러리를 제2 단편에 적용하는 단계; (e) 서열 분석 반응을 수행하여 프로브 라이브러리에 하이브리드화된 복수의 제2 단편의 서열을 동정하여 게놈 샘플의 하나 이상의 표적화 부분으로부터 서열 정보를 수득하는 단계. 추가의 구현예에서, 프로브 라이브러리는 결합 모이어티에 부착시키고 상기 수행 단계(e) 전에, 제2 단편은 결합 모이어티와 포획 모이어티 간의 반응을 통해 포획 모이어티를 포함하는 표면 상에 포획된다. 여전히 추가의 구현예에서 그리고 수행 단계 (e) 전에, 제2 단편은 제2 단편이 상기 표면 상에 포획되기 전후에 증폭된다. 여전히 추가의 구현예에서, 결합 모이어티는 비오틴을 포함하고 포획 모이어티는 스트렙타비딘을 포함한다. 여전히 추가의 구현예에서, 서열 분석 반응은 짧은 판독의 높은 정확도의 서열 분석 반응이다. 여전히 추가의 구현예에서, 제2 단편은 수득한 증폭 생성물이 부분적 또는 완전한 헤어핀 구조를 형성할 수 있도록 증폭된다.

[0014] 추가의 양상에서 그리고 상기 중 어느 하나에 따라, 본원 기재내용은 분자 형태를 보유하면서 게놈 샘플의 하나 이상의 표적화 부분으로부터 서열 정보를 수득하기 위한 방법을 제공한다. 상기 방법은 하기의 단계를 포함한다: (a) 출발 게놈 물질을 제공하는 단계; (b) 출발 게놈 물질 기원의 개별 핵산 분자를 별개의 파티션으로 분포시켜 각각의 별개의 파티션이 제1 개별 핵산 분자를 함유하도록 하는 단계; (c) 별개의 파티션에서 제1 개별 핵산 분자를 단편화시켜 복수의 단편을 형성하는 단계; (d) 게놈의 하나 이상의 선택된 부분의 적어도 일부를 포함하는 단편들에 대해 집적된 집단을 제공하는 단계; (e) 집단으로부터 서열 정보를 수득하여 분자 형태를 보유하면서 게놈 샘플의 하나 이상의 표적화 부분을 서열 분석하는 단계. 추가의 구현예에서, 수득 단계 (e) 전에, 복수의 단편들은 이것이 형성되는 별개의 파티션과 각각의 단편을 관련시키기 위해 바코드를 태그시킨다. 여전히 추가의 구현예에서, 단계 (b)에서 개별 핵산 분자는 각각의 제1 개별 핵산 분자의 분자 형태가 유지되도록 분포된다.

[0015] 일부 양상에서, 본원의 개시내용은 게놈 샘플의 하나 이상의 표적화 부분으로부터 서열 정보를 수득하는 방법을 제공한다. 상기 방법은 제한 없이 (a) 게놈 샘플의 개별 핵산 분자를 제공하는 단계; (b) 개별 핵산 분자를 단편화하여 복수의 단편을 형성하는 단계로서 단편 각각은 바코드를 추가로 포함하고, 상기의 동일한 개별 핵산 분자 기원의 단편들이 공통된 바코드를 가져 각각의 단편을 이것이 유래된 개별 핵산 분자와 관련시키는 단계; (c) 게놈 샘플의 하나 이상의 표적화 부분을 함유하는 단편에 대해 복수의 단편을 집적시키는 단계; 및 (d) 집적된 복수의 단편의 서열을 동정하기 위한 서열 분석 반응을 수행하여 게놈 샘플의 하나 이상의 표적화 부분으로부터 서열 정보를 수득하는 단계. 추가의 구현예에서, 집적 단계는 게놈 샘플의 하나 이상의 표적화 부분으로 지시된 프로브 라이브러리를 적용함을 포함한다. 여전히 추가의 구현예에서, 프로브 라이브러리를 결합 모이어티에 부착시키고 수행 단계 전에 단편은 결합 모이어티와 포획 모이어티 사이의 반응을 통해 포획된다. 예시적 구현예에서, 결합 모이어티와 포획 모이어티 간의 반응은 단편을 표면상에 고정화시킨다.

도면의 간단한 설명

[0016] 도 1은 본원에 기재된 공정 및 시스템에 대해 통상의 공정을 사용한 표적화 게놈 영역의 동정 및 분석의 도해를 제공한다.

도 2a 및 도 2b는 본원에 기재된 공정 및 시스템을 사용한 표적화 게놈 영역의 동정 및 분석의 도해를 제공한다.

도 3은 본원에 기재된 방법 및 조성물을 사용한 서열 정보를 검출하는 검정을 수행하기 위한 전형적 작업 흐름을 설명한다.

도 4는 핵산 샘플을 비드와 배합하고 상기 핵산 및 비드를 별개의 소적으로 분할시키기 위한 공정의 도해를 제공한다.

도 5는 염색체 핵산 단편의 바코드화 및 증폭을 위한 공정의 도해를 제공한다.

도 6은 개별 염색체에 대한 서열 데이터를 기여하는데 있어서 염색체 핵산 단편의 바코드화 용도의 도해를 제공한다.

도 7은 본 발명의 방법의 일반 구현예를 설명한다.

도 8은 본 발명의 방법의 일반 구현예를 설명한다.

발명을 실시하기 위한 구체적인 내용

[0017] **발명의 상세한 설명**

[0018] 본 발명의 수행은 달리 지적되지 않는 경우 당업계의 기술 범위내에 있는 통상의 기술 및 유기 화학 기재내용, 중합체 기술, 분자 생물학(재조합 기술을 포함하는), 세포 생물학, 생화학 및 면역학을 사용할 수 있다. 상기 통상의 기술은 중합체 어레이 합성, 하이브리드화, 연결, 파아지 디스플레이, 및 표지를 사용한 하이브리드화의 검출을 포함한다. 적합한 기술의 특정 설명은 하기 본원의 실시예를 참조할 수 있다. 그러나, 물론, 다른 균등의 통상적 과정이 또한 사용될 수 있다. 상기 통상의 기술 및 기재내용은 *게놈 분석과 같은 표준 실험 매뉴얼에서 발견될 수 있다: 문헌참조[A Laboratory Manual Series* (Vols.I-IV), *Using Antibodies: A Laboratory Manual, Cells: A Laboratory Manual, PCR Primer: A Laboratory Manual, and Molecular Cloning: A Laboratory Manual* (all from Cold Spring Harbor Laboratory Press), Stryer, L.(1995) *Biochemistry* (4th Ed.)Freeman, New York, Gait, *"Oligonucleotide Synthesis: A Practical Approach"* 1984, IRL Press, London, Nelson and Cox (2000), *Lehninger, Principles of Biochemistry* 3rd Ed.,W.H.Freeman Pub., New York, N.Y. and Berg 등(2002) *Biochemistry,* 5th Ed.,W.H.Freeman Pub., New York, N.Y.], 상기 문헌 모두는 모든 목적을 위해 이들의 전문이 본원에 참조로 인용된다.

[0019] 본원에 사용된 바와 같이 그리고 첨부된 청구항에서 단수 형태 "a," "an," 및 "the"은 달리 명백하게 지정되지 않는 경우 복수에 대한 언급을 포함한다. 따라서, 예를 들어, "폴리머라제"에 대한 언급은 하나의 제제 또는 상기 제제들의혼합물을 언급하고 "상기 방법"에 대한 언급은 균등의 단계 및 통상의 기술자에게 공지된 방법 등에 대한 언급을 포함한다.

[0020] 달리 정의되지 않는 경우, 본원에 사용된 모든 기술적 및 과학적 용어는 본 발명이 속하는 통상의 기술자에 의해 통상적으로 이해되는 바와 동일한 의미를 갖는다. 본원에 언급된 모든 공보는 공보에 기재되고 본원에 기재된 발명과 관련하여 사용될 수 있는 장치, 조성물, 제형 및 방법을 기술하고 기재할 목적을 위해 본원에 참조로 인용된다.

[0021] 값의 범위가 제공되는 경우, 달리 명백히 지정되지 않는 경우 하한치 유니트의 10분의 1에 대한 각각의 사이 값, 상기 범위의 상한치에서 하한치 및 상기 진술된 범위에서 임의의 다른 진술된 값 또는 사이 값은 본 발명내에 포괄되는 것으로 이해된다. 이들 보다 작은 범위의 상한치 및 하한치는 독립적으로 보다 작은 범위에 포함될 수 있고 이는 또한 본 발명에 포괄되고 상기 진술된 범위에서 임의의 특이적으로 배제된 치수에 적용된다. 상기 진술된 범위가 수치 중 하나 또는 둘다를 포함하는 경우, 이들 포함된 수치둘다를 배제하는 범위가 또한 본 발명에 포함된다.

[0022] 하기의 기재에서, 수많은 구체적 세부 사항은 본 발명의 보다 완전한 이해를 제공하기 위해 제시된다. 그러나, 본 발명이 하나 이상의 이들 특정 세부 사항 없이 수행될 수 있음은 통상의 기술자에게 자명할 것이다. 다른 경우에, 통상의 기술자에게 널리 공지된 특징 및 과정은 본 발명을 모호하게 함을 회피하기 위해 기재되지 않았다.

[0023] 본원에 사용된 바와 같이, 용어 "포함하는"은 조성물 및 방법이 다른 것들을 배제하지 않으면서 언급된 요소들을 포함함을 의미하는 것으로 의도된다. 조성물 및 방법을 한정하기 위해 사용되는 경우의 "필수적으로 이루어진"은 조성물 또는 방법에 대한 임의의 필수 유의성의 다른 요소들을 배제함을 의미한다. "로 이루어진"은 청구된 조성물 및 실질적 방법 단계에 대한 다른 성분들의 미량 이상의 요소들을 배제함을 의미한다. 이들 이행 용어 각각에 의해 정의되는 구현예는 본 발명의 범위 내에 있다. 따라서, 방법 및 조성물은 어떠한 유의성이 없는 단계 및 조성물 (필수적으로 이루어진)을 (포함하는) 또는 대안적으로 함유하는(including) 추가의 단계 및

성분들을 포함할 수 있는 것으로 의도되거나 또는 단지 진술된 방법 단계 또는 조성물(로 이루어진)을 의도한다..

[0024] 범위를 포함하는 모든 수치적 지정값, 예를 들어, pH, 온도, 시간, 농도 및 분자량은 0.1의 증가분으로 (+) 또는 (-)로 변화되는 대략적인 값이다. 항상 명백하게 진술되지는 않더라도, 모든 수치적 지정 값 앞에 "약"이 있는 것으로 이해되어야만 한다. 용어 "약"은 또한 "X + 0.1" 또는 "X - 0.1"과 같은 "X"의 최소 증가분 뿐만 아니라 정확한 값 "X"를 포함한다. 또한, 항상 명백하게 진술되지 않더라도 본원에 기재된 시약은 단지 예시적이고 상기의 균등물이 당업계에 공지되어 있는 것으로 이해되어야만 한다.

[0025] *I. 개요*

[0026] 본원의 개시내용은 유전학적 물질의 특징 분석을 위해 유용한 방법, 조성물 및 시스템을 제공한다. 특히, 본원에 기재된 방법, 조성물 및 시스템은 제한 없이 특정 염색체, 염색체 영역, 모든 엑손(엑솜), 엑솜 부분, 특이적 유전자, 유전자 패널 (*예를 들어,*키놈 또는 다른 표적화 유전자 패널), 인트론 영역, 게놈의 타일화된 부분 또는 게놈의 임의의 다른 선택된 부분을 포함하는 게놈의 표적화 영역의 유전학적 특징 분석을 제공한다.

[0027] 일반적으로, 본원에 기재된 방법 및 시스템은 긴 개별 핵산 분자의 서열의 결정및/또는 동정 및 긴 범위의 서열 정보의 사용을 가능하게 하는 긴 서열 스트레치에 의해 분리된 2개의 서열 분절 사이와 같이 직접적인 분자 연결체의 동정을 위해 제공함에 의한 표적화 게놈 서열 분석을 성취하지만, 상기 서열 분석 정보는 극히 낮은 서열 분석 오류율 및 짧은 판독 서열 분석 기술의 고속 처리의 이점을 갖는 방법을 사용하여 수득된다. 본원에 기재된 방법 및 시스템은 고속 처리, 보다 높은 정확도의 짧은 판독 서열 분석 기술을 사용하여 서열 분석될 수 있는 보다 작은 단편으로 긴 핵산 분자를 분절시키고 상기 분절은 본래의 긴 범위의 분자량 서열 형태를 보유하기 위해 보다 작은 단편으로부터 유래된 서열 정보를 가능하게 하는, 즉, 보다 짧은 판독이 보다 긴 개별 핵산 분자의 기원을 밝히는 방식으로 성취된다. 서열 판독을 기원하는 보다 긴 핵산 분자에 기인함에 의해, 통상의 기술자는 통상의 기술자가 일반적으로 단독의 짧은 서열 판독으로부터 수득할 수 없는 보다 긴 핵산 서열에 대해 유의적인 특징 분석 정보를 수득할 수 있다. 상기 긴 범위의 분자 형태는 서열 분석 공정 동안 보존될 뿐만 아니라 본원에 기재된 표적화 서열 분석 방법에 사용되는 표적화 집적 공정 동안 보존되고 여기서, 어떠한 다른 서열 분석 방법도 이러한 능력을 나타내지 않았다.

[0028] 일반적으로, 보다 작은 단편으로부터의 서열 정보는 본원에 기재되고 당업계에 공지된 바와 같은 바코드의 첨가를 포함하는, 태그 과정의 사용 동안 본래의 긴 범위의 분자 서열 형태를 보유한다. 특이적 예에서, 동일한 본래의 보다 긴 개별 핵산 분자로부터 기원하는 단편은 통상의 바코드로 태그하여 상기 단편 기원의 임의의 이후 서열 판독이 보다 긴 개별 핵산 분자에 기인될 수 있도록 한다. 상기 바코드는 트랜스포존을 사용한 바코드의 본래의 개별 핵산 분자로의 삽입뿐만 아니라 개별 핵산 분자의 분절을 증폭시키는 증폭 방법 동안에 바코드 서열의 부가를 포함하는 당업계에 공지된 임의의 방법을 사용하여 부가될 수 있고, 이는 하기 문헌에 기재된 것들과 같은 방법을 포함하고[참조: Amini 등,Nature Genetics 46: 1343-1349 (2014) (advance online publication on October 29, 2014)] 이는 모든 목적을 위해 및 특히 어댑터, 및 트랜스포존을 사용한 다른 올리고뉴클레오타이드를 부가하는 것과 관련된 모든 교시를 위해 이의 전문이 본원에 참조로 인용된다. 일단, 핵산이 상기 방법을 사용하여 태그되는 경우, 수득한 태그된 단편은 본원에 기재된 방법을 사용하여 집적될 수 있어 단편의 집단은 게놈의 표적화 영역을 나타낸다. 이와 같이, 상기 집단으로부터 서열 판독은 게놈의 선택 영역의 표적화 서열 분석을 가능하게 하고 상기 서열 판독은 또한 기원하는 핵산 분자에 기인될 수 있고 따라서 본래의 긴 범위의 분자 서열 형태를 보존한다. 상기 서열 판독은 당업계에 공지되고 본원에 기재된 임의의 서열 분석 방법 및 플랫폼을 사용하여 수득될 수 있다.

[0029] 게놈의 표적화 영역으로부터 서열 정보를 수득하는 능력을 제공하는 것 뿐만 아니라, 본원에 기재된 방법 및 시스템은 또한 제한 없이 반수체형 단계화, 구조적 변화의 동정, 및 모든 목적을 위해서 및 특히 게놈 물질의 특징 분석에 지시된 모든 기재 내용, 도면 및 실시예에 대해 이들의 전문이 참조로 인용되는 공계류 출원 USSN 14/752,589 및 14/752,602(이 둘다는 2015년 6월 26일자로 출원됨)에 기재된 바와 같이 카피수 변화를 동정하는 게놈 물질의 다른 특징 분석을 제공할 수 있다.

[0030] 본원에 기재된 방법 및 시스템에 따라 핵산을 가공하고 서열 분석하는 방법은 또한 문헌[참조: USSNs 14/316,383; 14/316,398; 14/316,416; 14/316,431; 14/316,447; 및 14/316,463]에서 추가로 상세히 기재되어 있고 이는 또한 모든 목적을 위해서 및 특히 핵산을 가공하고 특히 게놈 물질의 서열 분석 및 다른 특징 분석에 지시된 모든 기재 내용, 도면 및 실시예를 위해 이들의 전문이 본원에 참조로 인용된다.

[0031] 일반적으로, 도 1에 도시된 바와 같이, 본원에 기재된 방법 및 시스템을 사용하여 핵산을 특징 분석할 수 있다. 특히, 나타낸 바와 같이 2개의 별개의 개별 핵산 102 및 104가 설명되고 각각은 목적하는 다수의 영역, 예를 들어, 핵산 102에서 영역 106 및 108 및 핵산 104에서 영역 110 및 112를 갖는다. 각각의 핵산에서 목적하는 영역들은 동일한 핵산 분자 내에 연결되어 있지만 서로간에 상대적으로 분리되어 있을 수 있고, 예를 들어, 1kb 초과 격리, 5 kb 초과 격리, 10 kb 초과 격리, 20 kb 초과 격리, 30 kb 초과 격리, 40 kb 초과 격리, 50 kb 초과 격리이고 일부 경우에는 100 kb 정도로 큰 격리이다. 상기 영역은 게놈의 개별 유전자, 유전자 그룹, 엑손 또는 단순히 별개 및 별도의 부분을 지칭할 수 있다. 유일하게 논의의 용이함을 위해, 도 1에 보여지는 영역은 엑손 106, 108, 110 및 112로 언급된다. 나타낸 바와 같이, 각각의 핵산 102 및 104는 각각 이 자체의 파티션 114 및 116으로 분리되어 있다. 본원에서 그 밖의 다른 곳에서 주지된 바와 같이, 이들 파티션은 많은 경우에 유중수 에멀젼에서 수성 소적이다. 각각의 소적내에서, 각각의 단편 부분은 *예를 들어*, 동일 분자로부터 기원하는 바와 같은 상기 단편의 본래의 분자 형태를 보존하는 방식으로 카피된다. 나타낸 바와 같이, 이것은 설명된 바와 같이 바코드 서열, *예를 들어*, 바코드 서열 “1” 또는 “2”의 각각의 카피된 단편에서의 내포를 통해 성취되고, 이는 기원하는 단편이 분할되는 소적을 대표한다. 전체 게놈 서열 분석 적용을 위해, 통상의 기술자는 단순히, 기원하는 핵산 102 및 104의 각각으로부터 기원하는 서열을 분석하고 완전한 서열 정보를 리어셈블리하기 위해 모든 카피된 단편 및 이들의 관련된 바코드를 수집할 수 있다. 그러나, 많은 경우에, 전체 게놈의 특이적 표적화 부분, *예를 들어*, 엑솜, 특이적 유전자 등을 분석하는 것이 보다 바람직할 수 있고 이는 게놈의 과학적 관련 부분에 대한 보다 큰 집중을 제공하고 게놈의 덜 관련되거나 관련없는 부분에 대해 서열 분석을 수행하는 시간 및 비용을 최소화하는 것이다.

[0032] 본원에 기재된 방법에 따라, 표적 집적 단계는 목적하는 표적과 관련된 서열을 ” 풀 다운” 하기 위해 바코드화된 서열 단편의 라이브러리에 적용될 수 있다. 이들은 엑손 표적화 풀 다운, 유전자 패널 특이적 표적화 풀 다운 등을 포함할 수 있다. 게놈의 특이적 표적화 영역의 집적된 분리를 가능하게 하는 대다수의 표적화 풀 다운 키트는 시판되고 있고, 예를 들어, Agilent SureSelect 엑솜 풀 다운 키트 등이 있다. 도 1에 보여지는 바와 같은 표적화 집적의 적용은 집적된 바코드화된 서열 라이브러리 118을 수득한다. 추가로, 라이브러리 118 내의 풀 다운된 단편은 본래의 분자 형태를 예를 들어, 바코드 정보의 보유를 통해 보유하기 때문에, 이들은 매립된 긴 범위 연결 정보와 함께 , 예를 들어, 목적하는 어셈블리된 영역 106: 108 및 110: 112 각각 사이에서와 같이 추론된 연결과 함께 본래의 분자 형태로 재어셈블리될 수 있다. 예를 들어, 통상의 기술자는 게놈의 2개의 이질적인 표적화 부분, 예를 들어, 2개 이상의 엑손들 사이의 직접적인 분자 연결을 동정할 수 있고, 직접적인 분자 연결을 사용하여 구조적 변형 및 다른 게놈 특징을 동정할 수 있을 뿐만 아니라 2개 이상의 엑손에 관한 단계 정보를 동정할 수 있고, 예를 들어, 게놈의 전체 단계적 엑솜 또는 다른 단계적 표적화 부분을 잠재적으로 포함하는 단계적 엑손을 제공할 수 있다.

[0033] 일반적으로, 본 발명의 방법은 도 7에 도해된 단계를 포함하고, 이는 본원에서 추가로 상세히 논의되는 본 발명의 도식적 개요를 제공한다. 인식된 바와 같이, 도 9에 도시된 방법은 요구되는 바와 같이 및 본원에 기재된 바와 같이 변화되거나 변형될 수 있는 예시적 구현예이다.

[0034] 도 7에 나타낸 바와 같이, 본원에 기재된 방법은 대부분의 예에서 목적하는 표적화 영역을 함유하는 샘플 핵산이 분할되는 단계를 포함한다(701). 일반적으로, 각각의 파티션은 일반적으로 이들이 함유된 파티션에 특이적인 단편을 바코드화 함에 의해 단편의 본래의 분자 형태를 보존하는 것에 관한 방식으로 단편화되거나 카피된 특정 유전자좌 기원의 단일 개별 핵산 분자를 포함한다(702). 각각의 파티션은 일부 예에서 하나 이상의 핵산을 포함할 수 있고 일부 경우에 다중 핵산이 파티션 내에 있는 상황에서 수백 개의 핵산 분자를 함유하고, 게놈의 임의의 특정 유전자좌는 일반적으로 바코드화하기 전에 단일 개별 핵산에 의해 대표된다. 단계 (702)의 바코드화된 단편은 당업계에 공지된 임의의 방법을 사용하여 생성될 수 있고-일부 예에서, 올리고뉴클레오타이드는 두드러진 파티션 내의 샘플이다. 상기 올리고뉴클레오타이드는 샘플의 수많은 상이한 영역을 무작위로 프라이밍하는 것으로 의도된 무작위 서열을 포함할 수 있거나, 이들은 샘플의 표적화 영역의 업스트림을 프라이밍하기 위해 표적화 특이적 프라이머 서열을 포함할 수 있다. 추가의 예에서, 이들 올리고뉴클레오타이드는 또한 바코드 서열을 함유하여 상기 복제 공정은 또한 본래의 샘플 핵산의 수득한 복제된 단편을 바코드화한다. 샘플을 증폭하고 바코드화하는데 있어서 이들 바코드 올리고뉴클레오타이드의 사용을 위한 특히 정교한 공정은 하기 문헌에 상세히 기재되어 있다: 미국특허 출원 번호14/316,383, 14/316,398, 14/316,416, 14/316,431, 14/316,447, 14/316,463[이 모두는 2014년 6월 26일에 출원되었고 이들 각각은 모든 목적을 위해 이의 전문이 본원에 참조로 인용된다].연장 반응 시약, 예를 들어,DNA 폴리머라제, 뉴클레오사이드 트리포스페이트, 보조 인자 (예를 들어,Mg^{2+} 또는 Mn^{2+} 등)는 또한 파티션 내에 함유되고, 이어서 주형으로서 샘플을 사용하여 프라이머 서열을 연장

하여 프라이머가 어닐링되는 주형의 가닥에 상보적인 단편을 제조하고 상기 상보적 단편은 올리고뉴클레오타이드 및 이의 관련된 바코드 서열을 포함한다. 다수의 프라이머의 샘플의 상이한 부분으로의 어닐링 및 연장은 샘플의 중첩 상보적 단편의 대형 풀을 수득할 수 있고 각각은 이것이 생성된 파티션을 지적하는 이 자신의 바코드 서열을 프로세싱한다. 일부 경우에, 이들 상보적 단편 자체는 바코드 서열을 다시 포함하는 상보체의 상보체를 생성하도록 파티션에 존재하는 올리고뉴클레오타이드에 의해 프라이밍된 주형으로서 사용될 수 있다. 추가의 예에서, 상기 복제 공정은 제1 상보체가 복제되는 경우 이것이 추가의 반복적 카피를 생성하기 위해 기본이 되는 분자의 능력을 감소시키는, 헤어핀 구조 또는 부분적 헤어핀 구조의 형성을 가능하게 하기 위한 말단에서 또는 말단 부근에서 2개의 상보적 서열을 생성하도록 구성된다.

[0035] 도 7에 예시된 방법을 다시 참조하면, 파티션-특이적 바코드가 카피된 단편에 부착되면, 이어서 바코드화된 단편이 수집된다(703). 이어서, 표적 집적 기술은 목적하는 표적화 영역을 “풀 다운”시키기 위해 적용(704)될 수 있다. 이어서 목적하는 상기 표적화 영역은 서열 분석되고(705) 단편의 서열은 이들의 기원하는 분자 형태(706)에 기인하여 상기 목적하는 표적화 영역 둘다가 동정되고 또한 상기 기원하는 분자 형태와 연결된다. 본원에 기재되고 도 7에 설명된 방법 및 시스템의 유일한 특징은 바코드가 표적화 집적 단계(704) 전에 단편에 부착되는 것이다(702). 본원에 기재된 방법 및 시스템의 이점은 표적화 게놈 영역에 대해 단편을 집적시키기 전에 카피된 단편에 파티션- 또는 샘플-특이적 바코드를 부착시키는 것은 상기 표적화 영역의 본래의 분자 형태를 보존하여 이들이 이들의 본래의 파티션 및 따라서 이들의 본래의 샘플 핵산에 기인하도록 한다.

[0036] 일반적으로, 표적화 게놈 영역은 집적되고, 단리되고 분리되고, 즉,추가의 분석, 특히 서열 분석을 위해 “풀 다운”되고, 칩-기반 및 용액-기반 포획 방법 둘다를 포함하는 방법을 사용한다. 상기 방법은 목적하는 게놈 영역 또는 목적하는 게놈 영역의 근처 또는 이에 인접한 영역에 상보적인 프로브를 사용한다. 예를 들어, 하이브리드(또는 칩-기반) 포획에서, 목적하는 영역을 함께 커버하는 서열과 함께 포획 프로브(일반적으로 단일-가닥의 올리고뉴클레오타이드)를 함유하는 마이크로어레이를 표면에 고정시킨다. 게놈 DNA는 단편화되고 평활 말단을 생성하기 위한 말단-복구 및/또는 범용 프라이밍 서열과 같은 추가 특징의 부가와 같은 프로세싱을 추가로 받을 수 있다. 이들 단편은 마이크로어레이 상에서 프로브에 하이브리드화된다. 하이브리드화되지 않은 단편은 세척 제거하고 목적하는 단편은 용출시키거나 다르게는 서열 분석 또는 다른 분석을 위해 표면 상에 프로세싱되고 따라서 표면 상에 잔류하는 단편의 집단은 목적하는 표적화 영역(예를 들어, 포획 프로브에 함유되는 것들에 상보적인 서열을 포함하는 영역)을 함유하는 단편들에 대해 집적된다. 단편의 집적된 집단은 추가로 당업계에 공지된 임의의 증폭 기술을 사용하여 증폭될 수 있다.

[0037] 표적화 게놈 영역 포획의 추가의 방법은 용액-기반 방법을 포함하고, 여기서, 게놈 DNA 단편은 올리고뉴클레오타이드 프로브에 하이브리드화된다. 올리고뉴클레오타이드 프로브는 흔히 “베이트”로서 언급된다. 이들 베이트는 일반적으로 포획 분자에 부착되고 제한 없이 비오틴 분자를 포함한다. 베이트는 게놈의 표적화 영역(또는 목적하는 표적화 영역 근처 또는 인접한 영역)에 상보적이어서 게놈 DNA 단편에 적용시, 베이트는 단편에 하이브리드화하고 이어서 포획 분자 (예를 들어, 비오틴)는 목적하는 표적화 영역(예를 들어, 자기 스트렙타비딘 비드와 함께)을 선택적으로 풀 다운하기 위해 사용됨에 따라 목적하는 표적화 영역을 포함하는 것들과 함께 수득한 단편 집단을 집적시킨다.

[0038] 전체 엑솜을 커버하는 표적화 영역이 요구되는 예에서, 전체 엑솜을 함께 커버하는 베이트 라이브러리는 상기 표적화 서열을 포획하기 위해 사용된다. 상기 예에서, 포획 프로토콜은 당업계에 공지된 임의의 것들을 포함할 수 있고, 제한 없이 엑솜 포획 프로토콜 및 키트[제조원(Roche/NimbleGen, Illumina, and Agilent)에 의해 제조됨]를 포함한다.

[0039] 본원에 기재된 방법 및 시스템에 사용하기 위한 표적화 게놈 영역의 포획은 전체엑솜으로 제한되지 않고 부분적 엑솜, 유전자, 유전자 패널, 인트론 및 인트론 및 엑손의 조합 중 임의의 하나 또는 조합을 포함할 수 있다. 이들 상이한 유형의 표적화 영역의 포획을 위한 과정은 목적하는 표적화 영역을 함유하는 단편들을 풀 다운시키기 위해 베이트를 사용하는 일반적인 방법에 따른다. 베이트, 특히, 목적하는 표적화 영역에 또는 이의 근처에 하이브리드화하는 베이트의 올리고뉴클레오타이드 프로브의 디자인은 부분적으로 포획될 표적화 영역의 유형에 의존한다.

[0040] 추가의 분석을 위해 단지 부분적 엑솜이 요구되는 예에서, 베이트는 엑솜의 상기 부분을 포획하도록 디자인될 수 있다. 특정 예에서, 요구되는 엑솜 부분의 특정 동정체는 공지되어 있고 베이트 라이브러리는 상기 동정된 부분에 상보적이거나 상기 부분 근처이거나 인접한 영역에 상보적인 올리고뉴클레오타이드를 포함한다. 상기 예는 제한 없이 특정 유전자 및/또는 유전자 패널, 또는 장애 또는 질환과 같은 특정 표현형과 관련된 것으로 공

지된 엑솜의 동정된 부분을 추가로 포함할 수 있다. 일부 예에서, 엑솜 또는 전체 게놈 (인트론 및 엑손 영역 둘다를 포함하는)의 특정 부분이 추가의 분석을 위해 요구될 수 있지만 포획될 게놈 부분에 대한 특정 서열은 공지되어 있지 않다. 상기 예에서, 사용된 베이트는 전체 게놈으로 지시된 라이브러리 서브세트일 수 있고, 상기 서브세트는 무작위로, 또는 베이트 라이브러리가 게놈 또는 엑솜의 표적화 서브섹션에 상보적인 프로브에 대해 선택되거나 집적되는 임의의 유형의 이상적 디자인을 통해 선택될 수 있다.

[0041] 본원에 기재된 임의의 방법에 대해, 표적화 영역은 전체 또는 표적화 영역의 일부에 상보적인 올리고뉴클레오타이드 프로브를 포함하는 베이트를 사용하여 포획될 수 있거나 상기 올리고뉴클레오타이드 프로브는 또 다른 영역, *예를 들어,* 표적화 영역 근처 또는 표적화 영역에 인접한 인트론 영역에 상보적일 수 있다. 예를 들어, 도 2a에 도식적으로 설명된 바와 같이, 게놈 서열 201은 엑손 영역 202 및 203을 포함한다. 이들 엑손 영역은 인근 인트론 서열 (예를 들어, 엑손 영역 202를 포획하기 위한 인트론 영역 204 및/또는 205 및 엑손 영역 203의 포획을 위한 영역 206) 중 하나 이상으로 베이트를 지시함에 의해 포획될 수 있다. 다른 말로, 엑손 영역 202 또는 203을 포함하는 단편 집단은 인트론 영역 204 및/또는 205 및 206에 상보적인 베이트의 사용을 통해 포획될 수 있다. 도 2a에 보여지는 바와 같이, 인근 엑손 영역에 대한 인트론 베이트로서 사용되는 인트론 영역은 목적하는 엑손 영역에 인접할 수 있고, 즉, 인트론 영역과 표적화 엑손 영역 간에 갭이 없다. 다른 예에서, 인근 엑손 영역을 포획하기 위해 사용되는 인트론 영역은 양쪽 영역이 동일한 단편에 있을 가능성이 있지만 엑손 영역과 인트론 영역 (예를 들어, 도 2a에서 202 및 205) 간에 하나 이상의 뉴클레오타이드의 갭이 있다.

[0042] 일부 예에서, 게놈의 특정 영역을 표적화하기 위해 베이트를 디자인하기 보다는 타일화 방법이 사용된다. 상기 방법에서, 특정 엑손 또는 인트론 영역을 표적화하기 보다는, 베이트는 특정범위 또는 거리에서 게놈 부분에 상보적이도록 디자인된다. 예를 들어, 베이트 라이브러리는 게놈을 따라 5킬로베이스(kb) 마다 서열을 커버하도록 디자인되어 상기 베이트 라이브러리를 단편화된 게놈 샘플에 적용하는 것은 단지 게놈의 특정 서브세트 - *즉,* 베이트에 대해 상보적인 서열을 함유하는 단편에 함유된 영역을 포획하도록 한다. 평가된 바와 같이, 베이트는 인간 게놈 참조 서열과 같은 참조 서열을 기반으로 디자인될 수 있다. 추가의 예에서, 베이트의 타일화된 라이브러리는 게놈의 1, 2, 5, 10, 15, 20, 25, 50, 100, 200, 250, 500, 750, 1000, 또는 10000 킬로베이스 마다 영역을 포획하도록 디자인된다. 여전히 추가의 예에서, 베이트의 타일화된 라이브러리는 거리상의 혼합물을 포획하도록 디자인되고 - 상기 혼합물은 거리상의 무작위 혼합물일 수 있고 게놈의 특정 부분 또는 %가 포획되도록 디자인될 수 있다. 평가된 바와 같이, 상기 포획의 타일화 방법은 서열 분석과 같은 추가의 분석을 위해 게놈의 인트론 및 엑손 영역 둘다를 포획한다. 본원에 기재된 임의의 타일화 또는 다른 인트론 베이트화 방법은 긴 삽입 인트론영역에 의해 광범위하게 분리된 엑손으로부터의 서열 정보를 연결시키는 방식을 제공한다.

[0043] 추가의 예에서, 본원에 기재된 타일화 또는 다른 포획 방법은 전체 게놈의 약 5%, 10%, 15%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%를 포획한다. 여전히 추가의 예에서, 본원에 기재된 포획 방법은 전체 게놈의 약 1-10%, 5-20%, 10-30%, 15-40%, 20-50%, 25-60%, 30-70%, 35-80%, 40-90%, 또는 45-95%를 포획한다.

[0044] 일부 예에서, 게놈 DNA를 단편화하고, 증폭시키고, 분할시키고 다르게는 프로세싱하는 방법을 포함하는 샘플 제조 방법은 게놈의 특정 영역의 편향 또는 보다 낮은 커버를 유도할 수 있다. 상기 편향 또는 보다 낮은 커버 범위는 게놈의 표적화 게놈을 포획하기 위해 사용되는 베이트의 농도 또는 게놈 위치를 변화시킴에 의해 본원에 기재된 방법 및 시스템에서 보상될 수 있다. 일부 예에서, 높은 GC 함량 또는 다른 구조적 변형을 함유하는 게놈의 특정 영역은 낮은 커버 범위를 유도할 것이라고 공지된 것일 수 있고 - 상기 상황하에서, 베이트의 라이브러리는 낮은 커버 범위의 상기 영역으로 지시된 베이트의 농도를 증가시키도록 변화될 수 있고 - 다른 말로, 사용되는 베이트 집단은 상기 낮은 커버 범위 영역에서 게놈의 표적화 영역을 함유하는 충분한 수의 단편이 서열분석될 최종 단편 집단에서 수득되는 것을 확실히 하기 위해 "스파이킹" 될 수 있다. 상기 베이트의 스파이킹은 보다 낮은 커버 범위 영역 쪽으로 지시된 베이트의 커스텀 라이브러리가 재고품의 엑솜 포획 키트에 부가되도록 시판되는 전체 엑솜 키트에서 수행될 수 있다. 추가로, 베이트는 목적하는 영역에 매우 근접해 있지만 보다 선호될 수 있는 커버 범위를 갖는 게놈 영역을 표적화하기 위한 디자인일 수 있고 이는 또한 본원에서 추가로 상세하게 논의되고 이의 양태는 도 2에 도식적으로 설명된다.

[0045] 추가의 예에서, 본 발명의 방법에 사용되는 베이트 라이브러리는 본원에서 추가로 기재된 바와 같이 하나 이상의 특성을 충족하는 주지된 디자인을 갖는 제품이다. 상기 정통한 디자인은 베이트 라이브러리는 정통한 단일 뉴클레오타이드 다형태(SNP)에 지시된다. 본원에 사용된 바와 같은 용어 "유익한 SNP"는 이종접합성인 SNP를 언급한다. 일부 예에서 베이트 라이브러리는 정통한 SNP를 함유하는 게놈 샘플 영역으로 지시된 다수의 프로브를 함유하도록 디자인된다. 본원에 사용된 바와 같은 "으로 지시된"이란 프로브가 SNP를 포함하는 서열에 상보적인 서열을 함유함을 의미한다. 추가의 예에서, 베이트의 라이브러리는 엑손과 인트론의 경계선으로부터 미

리 결정된 거리 상에 있는 SNP로 지시된 프로브를 함유하도록 디자인된다. 게놈의 표적화 영역이 SNP가 없거나 매우 소수의 SNP를 함유하는 영역을 포함하는 상황에서, 베이트의 라이브러리는 미리 결정된 거리에서 상기 영역에 걸쳐 타일화되어 있고/있거나 다음에 가장 근접한 인트론 또는 엑손 내에 제1의 유익한 SNP에 하이브리드화하는 프로브를 포함한다.

[0046] 본원에 기재된 방법 및 시스템의 이점은 포획된 표적화 영역이, 표적화 영역을 포획하고 서열 분석을 수행하는 단계 후에도 이들 표적화 영역의 본래의 분자 형태가 보유되는 방식으로 포획 전에 프로세싱된다. 본원에서 추가로 상세하게 논의된 바와 같이, 특정 표적화 영역이 이들의 본래의 분자 형태(이들이 유래된 본래의 염색체 또는 염색체 영역 및/또는 완전한 게놈 내 서로 관련하여 특정 표적화 영역의 위치를 포함할 수 있다)에 기인하는 능력은 달리 불량하게 맵핑되거나 통상의 서열 분석 기술을 사용한 불량한 커버를 갖는 게놈의 영역으로부터 서열 정보를 수득하는 방식을 제공한다.

[0047] 예를 들어, 일부 유전자는 일반적으로 가용한 서열 분석 기술을 사용하여, 특히, 긴-판독 기술과 비교하여 보다 우수한 정확도를 갖는 짧은 판독 기술을 사용하여 포괄하기에는 너무 긴 인트론을 갖는다. 그러나, 본원에 기재된 방법 및 시스템에서, 표적화 영역의 분자 형태는 일반적으로 도 1에서 설명되고 본원에서 추가로 상세히 기재된 태깅 과정을 통해 보유된다. 이와 같이, 연결은 게놈의 확대된 영역을 거쳐 만들어질 수 있다. 예를 들어, 도 2b에 도식적으로 설명된 바와 같이, 핵산 분자 207은 긴 인트론 영역(208)에 의해 차단된 2개의 엑손(차양막대)을 함유한다. 일반적으로 사용된 서열 분석 기술은 2개의 엑손의 관계에 대한 정보를 제공하도록 인트론에 걸친 거리 상을 포괄할 수 없다. 본원에 기재된 방법에서, 개별 핵산 분자 207은 이 자체의 별개의 파티션 209로 분포되고 이어서 단편화되어 상이한 단편은 엑손 및 인트론의 상이한 부분을 함유한다. 상기 단편 각각은 단편으로부터 수득된 임의의 서열 정보가 이어서 이것이 생성되는 별개의 파티션에 기인될 수 있도록 태깅되기 때문에, 각각의 단편은 또한 이것이 유래된 개별 핵산 분자 207에 기인할 수 있다. 일반적으로 그리고 본원에서 추가로 상세히 기재된 바와 같이, 단편화 및 태깅 후, 상이한 파티션으로부터의 단편은 함께 조합된다. 이어서 표적화 포획 방법은 목적하는 표적화 영역을 함유하는 단편과 함께 , 추가의 서열 분석과 같은 추가의 분석을 받는 단편 집단을 집적시키기 위해 사용될 수 있다. 도 2b에 설명된 예에서, 사용되는 베이트는 2개의 엑손 중 한 부분 및/또는 삽입 인트론 부분을 함하는 것들만을 포획하지만 상기 엑손 및 인트론의 외부 영역(209 및 210과 같은)은 포획하지 않는 단편 집단을 집적시킨다. 따라서, 서열 분석에 적용되는 최종 단편 집단은 목적하는 2개의 엑손 부분을 함유하는 단편에 대해 집적된다. 이어서 짧은 판독의 고 정확도의 서열 분석 기술을 사용하여 집적된 단편 집단의 서열을 동정할 수 있고, 단편 각각은 태깅되고 따라서 이의 본래의 분자 형태, 즉 이의 본래의 개별 핵산 분자에 기인할 수 있기 때문에, 상기 짧은 판독 서열은 2개의 엑손 간의 관계에 대한 정보를 제공하도록 삽입 인트론의 긴 길이를 포괄하는 정보를 제공할 수 있다.

[0048] 상기된 바와 같이, 본원에 기재된 방법 및 시스템은 보다 긴 핵산의 짧은 서열 판독에 대한 개별 분자 형태를 제공한다. 본원에 사용된 바와 같은 개별 분자 형태는 예를 들어, 서열 판독 자체 내에 포함되지 않는 인접하거나 근접한 서열과 관련하여, 특정 서열 판독에 넘어서는 서열 형태를 언급하고 이와 같이 이들이 짧은 서열 판독, 예를 들어, 쌍을 이룬 판독에 대해 약 150개 염기 또는 약 300개 염기의 판독에 전체적으로 또는 부분적으로 포함되지 않도록 할 것이다. 특히 바람직한 양상에서, 방법 및 시스템은 짧은 서열 판독에 대한 긴 범위 서열형태를 제공한다. 상기 긴 범위 형태는 1 kb 보다 긴, 5 kb 보다 긴, 10 kb 보다 긴, 15 kb 보다 긴, 20 kb 보다 긴, 30 kb 보다 긴, 40 kb 보다 긴, 50 kb 보다 긴, 60 kb 보다 긴, 70 kb 보다 긴, 80 kb 보다 긴, 90 kb 보다 긴 또는 심지어 100 kb 보다 긴 서로의 거리 이내에 있는 서열 판독에 대한 소정의 서열 판독의 관계 또는 연결을 포함한다. 평가된 바와 같이, 긴 범위의 개별 분자 형태를 제공함에 의해, 당업자는 또한 상기 개별 분자 형태내에 변이체의 단계적 정보를 도출할 수 있고, 예를 들어, 특정 긴 분자 상의 변이체는 정의에 의해 통상적으로 단계화된다.

[0049] 보다 긴 범위의 개별 분자 형태를 제공함에 의해, 본 발명의 방법 및 시스템은 또한 훨씬 보다 긴 추론된 분자 형태(또한 본원에서 “긴 실제 단일 분자 판독”)를 제공한다. 본원에 기재된 바와 같은 서열 형태는 완전한 게놈 서열의 상이한 (일반적으로 킬로베이스 규모상에서) 범위에 걸친 단편의 연결체를 맵핑하거나 제공함을 포함할 수 있다. 이들 방법은 예를 들어, 개별 분자의 결정된 연속 서열을 갖는 보다 긴 개별 분자의 대형 부분의 긴 범위 서열 분석 뿐만 아니라 개별의 보다 긴 분자 또는 연결된 분자의 콘티그에 대해 짧은 서열 판독을 맵핑함을 포함하고, 여기서, 상기 결정된 서열은 1 kb 초과, 5 kb 초과, 10 kb 초과, 15 kb 초과, 20 kb 초과, 30 kb 초과, 40 kb 초과, 50 kb 초과, 60 kb 초과, 70 kb 초과, 80 kb 초과, 90 kb 초과 또는 심지어 100 kb 초과이다. 서열 형태와 관련하여, 보다 긴 핵산, 예를 들어, 연결된 핵산 분자 또는 콘티그의 개별적 긴 핵산 분자 또는 수집물 둘다로의 짧은 서열의 기인은 이들 보다 긴 핵산을 통한 짧은 서열로부터 어셈블리된 서열을 제공

하는 것 뿐만 아니라 높은 수준의 서열 형태를 제공하기 위해 보다 긴 핵산 스트레치에 대한 짧은 서열의 맵핑 둘다를 포함할 수 있다.

[0050] 추가로, 통상의 기술자는 긴 개별 분자들과 관련된 긴 범위의 서열 형태를 사용할 수 있지만, 상기 긴 범위의 서열 형태를 갖는 것은 또한 통상의 기술자가 보다 긴 범위의 서열 형태를 추론할 수 있게 해준다. 하나의 예로서, 상기된 긴 범위의 분자 형태를 제공함에 의해, 통상의 기술자는 상이한 본래의 분자 기원의 긴 서열들 중에서 중첩 변이체 부분, 예를 들어, 단계적 변이체, 전좌 서열 등을 동정할 수 있어 상기 분자 간의 추론된 연결을 가능하게 한다. 상기 추론된 연결체 또는 분자 형태는 본원에서 "추론된 콘티그"로서 언급된다. 단계적 서열의 형태에서 논의되는 일부 경우에, 추론된 콘티그는 통상적으로 단계적 서열을 나타낼 수 있고, 예를 들어, 중첩 단계화된 변이체에 의해, 통상의 기술자는 개별 기원의 분자 보다 실질적으로 큰 길이의 단계적 콘티그를 추론할 수 있다. 이들 단계적 콘티그는 "단계 블록"으로서 본원에 언급된다.

[0051] 보다 긴 단일 분자 판독 (예를 들어, 상기 논의된 "긴 실제 단일 분자 판독")으로 개시함에 의해, 통상의 기술자는 단계적 서열 분석에 대해 짧은 판독 서열 분석 기술 또는 다른 방법을 사용하여 달성될 수 있는 것 보다 긴 추론된 콘티그 또는 단계 블록을 도출할 수 있다. 문헌참조: 예를 들어, 공개된 U.S. 특허 출원 번호. 2013-0157870. 특히, 본원에 기재된 방법 및 시스템을 사용하여, 통상의 기술자는 적어도 약 10 kb, 적어도 약 20 kb, 적어도 약 50 kb의 N50(여기서, 진술된 N50 수 보다 큰 블록 길이의 합은 모든 블록 길이의 합의 50%이다)을 갖는 추론된 콘티그 또는 단계 블록 길이를 수득할 수 있다. 보다 바람직한 양상에서, 적어도 약 100 kb, 적어도 약 150 kb, 적어도 약 200 kb, 및 많은 경우에, 적어도 약 250 kb, 적어도 약 300 kb, 적어도 약 350 kb, 적어도 약 400 kb, 및 일부 경우에, 적어도 약 500 kb 이상의 N50을 갖는 추론된 콘티그 또는 단계 블록 길이가 달성된다. 여전히 다른 경우에, 200 kb 초과, 300 kb 초과, 400 kb 초과, 500 kb 초과, 1 Mb 초과, 또는 심지어 2 Mb 초과의 최대 단계 블록 길이가 수득될 수 있다.

[0052] 하나의 양상에서, 그리고 상기된 및 본원에서 이후에 기재된 임의의 포획 방법과 연계하여, 본원에 기재된 방법 및 시스템은 샘플 핵산 또는 이의 단편의 별개의 격실 또는 파티션 내로의 격실화, 침적 또는 분할(분할로서 본원에서 상호교환적으로 언급됨)을 제공하고, 여기서, 각각의 파티션은 다른 파티션의 내용물로부터 이 자신의 내용물의 분리를 유지한다. 고유 식별자, 예를 들어, 바코드는 이전에, 후속적으로 또는 동시에 격실화되거나 분할된 샘플 핵산을 유지하는 파티션으로 전달될 수 있고, 이는 특징, 예를 들어, 핵산 서열 정보의 특정 격실 내에 포함되는 샘플 핵산 및 특히 본래 파티션 내로 침적될 수 있는 연속 샘플 핵산의 상대적으로 긴 스트레치로의 이후 기인을 가능하게 하기 위한 것이다.

[0053] 본원에 기재된 방법에 사용되는 샘플 핵산은 전형적으로 분석될 전체 샘플, 예를 들어, 전체 염색체, 엑솜 또는 다른 대형 게놈 부분의 다수의 중첩 부분을 나타낸다. 이들 샘플 핵산은 전체 게놈, 개별 염색체, 엑솜, 앰플리콘 또는 목적하는 임의의 다양한 상이한 핵산을 포함할 수 있다. 샘플 핵산은 전형적으로 핵산이 연속적 핵산 분자의 상대적으로 긴 단편 또는 스트레치로 파티션 내 존재하도록 분할된다. 전형적으로, 샘플 핵산의 이들 단편은 1 kb 초과, 5 kb 초과, 10 kb 초과, 15 kb 초과, 20 kb 초과, 30 kb 초과, 40 kb 초과, 50 kb 초과, 60 kb 초과, 70 kb 초과, 80 kb 초과, 90 kb 초과 또는 심지어 100 kb 초과의 길이일 수 있고, 이는 상기된 보다 긴 범위의 분자 형태를 허용한다.

[0054] 샘플 핵산은 또한 전형적으로 소정의 파티션이 출발 샘플 핵산의 2개의 중첩 단편을 포함할 매우 낮은 가능성을 갖도록 하는 수준으로 분할된다. 이것은 전형적으로 분할 과정 동안에 낮은 투입량 및/또는 농도로 샘플 핵산을 제공함에 성취된다. 결과로서, 바람직한 경우에, 소정의 파티션은 출발 샘플 핵산의 다수의 길지만 비-중첩 단편을 포함할 수 있다. 상이한 파티션에서 샘플 핵산은 이어서 고유 식별자와 관련되고, 여기서, 임의의 소정의 파티션에 대해 여기에 함유된 핵산은 동일한 고유 식별자를 갖지만 상이한 파티션은 상이한 고유 식별자를 포함할 수 있다. 더욱이, 분할 단계는 샘플 성분을 매우 작은 용량 파티션 또는 소적으로 할당하기 때문에, 상기된 바와 같이 목적하는 할당을 성취하기 위해, 통상의 기술자는 예를 들어, 튜브, 또는 다중웰 플레이트의 웰에서 보다 높은 용적 공정으로 요구되는 바와 같이 샘플의 상당한 희석을 수행할 필요는 없다. 추가로, 본원에 기재된 시스템은 상기 높은 수준의 바코드 다양성을 사용하기 때문에, 통상의 기술자는 상기 제공된 바와 같은 보다 높은 수의 게놈 균등물 중에서 다양한 바코드를 할당할 수 있다. 특히, 이전에 기재된 다중웰 플레이트 방법(문헌참조: 예를 들어, 미국공개 출원 번호 제2013-0079231호 및 제2013-0157870호)은 단지 전형적으로 백개 내지 수백개의 상이한 바코드 서열을 사용하여 작동하고 바코드가 상이한 세포/핵산에 기인할 수 있도록 하기 위해 이들의 샘플의 제한 희석 과정을 사용한다. 이와 같이, 이들은 일반적으로 100개 훨씬 미만의 세포로 작동하고, 이는 전형적으로 1: 10의 정도, 및 특히 훨씬 1: 100 초과의 게놈: (바코드 유형)의 비율을 제공한다. 한편, 본원에 기재된 시스템은 예를 들어, 10,000, 100,000, 500,000 초과 등의 고수준의 바코드 다양성, 다양한 바코드

유형 때문에 1: 50 이하, 1: 100 이하, 1: 1000 이하, 또는 심지어 보다 작은 비율 정도의 게놈: (바코드 유형)에서 작동할 수 있고, 또한 보다 높은 수의 게놈(예를 들어, 검정 당 100개 초과의 게놈 정도, 검정 당 500개 초과의 게놈, 검정 당 1000개 초과의 게놈, 또는 심지어 그 이상)을 가능하게 하고 게놈 당 훨씬 개선된 바코드 다양성을 여전히 제공한다.

[0055] 흔히, 샘플은 분할 단계 전에 방출가능하게 비드에 부착된 올리고뉴클레오타이드 태그 세트와 조합된다. 상기 조합은 이어서 당업계에 공지되고 본원에 기재된 방법을 사용하여 샘플 중핵산의 바코드화를 유도할 수 있다. 일부 예에서, 증폭 방법은 일부 경우에 이들이 유래되는 완전한 본래의 핵산 분자의 보다 작은 분절(단편)으로 바코드를 부가하기 위해 사용되고 이는 일부 예에서 완전한 본래의 핵산 분자의 보다 작은 분절(단편)을 함유한다. 일부 예에서, 트랜스포존을 사용한 방법은 하기 문헌에 기재된 바와 같이 사용되고[문헌참조: Amini 등, Nature Genetics 46: 1343-1349 (2014) (advance online publication on October 29, 2014)], 상기 문헌은 모든 목적을 위해 및 특히 바코드 또는 다른 올리고뉴클레오타이드를 핵산으로 부착시키는 것과 관련된 모든 교시를 위해 이의 전문이 본원에 참조로 인용된다. 추가의 예에서, 바코드를 부착시키는 방법은 이중 가닥의 샘플 핵산을 따라 갭을을 제조하기 위한 recA와 같은 닉킹 효소 또는 폴리머라제 및/또는 침습적 프로브의 용도를 포함할 수 있고 - 이어서 바코드는 상기 갭으로 삽입될 수 있다.

[0056] 증폭이 핵산 단편을 태그하기 위해 사용되는 예에서, 올리고뉴클레오타이드 태그는 적어도 제1 및 제2 영역을 포함할 수 있다. 제1 영역은 소정의 파티션 내 올리고뉴클레오타이드들 사이에서와 같이 실질적으로 동일한 바코드 서열일 수 있지만 상이한 파티션 사이에서와 같이 대부분의 경우에 상이한 바코드 서열일 수 있는 바코드 영역일 수 있다. 제2 영역은 파티션 내의 샘플 내 핵산을 프라이밍하기 위해 사용될 수 있는 N-량체(무작위 N-량체 또는 특정 서열을 표적화하기 위해 디자인된 N-량체)일 수 있다. 일부 경우에, N-량체가 특정 서열을 표적화하기 위해 디자인된 경우, 특정 염색체(또는 염색체 1, 13, 18, 또는 21), 또는 염색체 영역, 예를 들어, 엑솜 또는 다른 표적화 영역을 표적화하도록 디자인될 수 있다. 일부 경우에, N-량체는 질환 또는 장애(예를 들어, 암)과 관련된 유전자 또는 영역과 같은 특정 유전자 또는 유전학적 영역을 표적화하도록 디자인될 수 있다. 파티션 내에 증폭 반응은 핵산 길이에 따른 상이한 위치에서 핵산 샘플을 프라이밍하기 위해 제2 N-량체를 사용하여 수행될 수 있다. 증폭 결과로서, 각각의 파티션은 동일하거나 거의 동일한 바코드에 부착되고 각각의 파티션 내 중첩된 보다 작은 핵산 단편을 나타낼 수 있는 핵산의 증폭된 생성물을 함유할 수 있다. 바코드는 동일한 파티션으로부터 기원하고 따라서 잠재적으로 핵산의 동일한 가닥으로부터 기원하는 핵산 세트를 의미하는 마커로서 작용할 수 있다. 증폭 후, 핵산은 서열 분석 알고리듬을 사용하여 수집되고, 서열 분석되고 정렬될 수 있다. 보다 짧은 서열 판독은 이들의 관련된 바코드 서열에 의해 할당될 수 있고 샘플핵산의 단일의 긴 단편에 기인될 수 있기 때문에, 상기 서열 상에서 동정된 모든 변이체는 단일 기원의 단편 및 단일 기원의 염색체에 기인될 수 있다. 추가로, 다수의 긴 단편에 걸쳐 다수의 동시에 위치된 변이체를 할당함에 의해 통상의 기술자는 상기 염색체 기여를 추가로 특징 분석할 수 있다. 따라서, 특정 유전학적 변이체의 단계화에 관한 결론은 게놈 서열의 긴 범위에 걸친 분석- 예를 들어, 게놈의 불량하게 특징 분석된 영역의 스트레치에 걸친 서열 정보의 동정에서 있을 수 있는 바와 같이 이끌어질 수 있다. 상기 정보는 또한 동일한 핵산 가닥 또는 상이한 핵산 가닥 상에 존재하는, 일반적으로 특정유전학적 변이체 세트인 반수체를 동정하기 위해 유용할 수 있다. 카피수 변화는 또한 상기 방식으로 동정될 수 있다.

[0057] 기재된 방법 및 시스템은 현재 핵산 서열 분석 기술 및 이들과 관련된 샘플 제조방법 보다 상당한 이점을 제공한다. 총체적 샘플 제조 및 서열 분석 방법은 주로 샘플 내 주요 성분들을 동정하고 특징분석하는 쪽으로의 경향이 있고 소수 성분들, 예를 들어, 하나의 염색체 또는 하나 또는 소수의 세포에 의해 기여되는 유전학적 물질, 또는 추출된 샘플 내 작은 %의 총 DNA를 구성하는 혈류 중에 순환하는 단편화된 종양 세포 DNA 분자를 동정하고 특징 분석하도록 디자인되지 않는다. 기재된 방법 및 시스템은 또한 보다 큰 샘플 내 존재하는 집단을 검출하기 위해 상당한 이점을 제공한다. 이와 같이, 이들은 특히 반수체 및 카피수 변화를 평가하기 위해 유용하고 - 본원에 기재된 방법은 또한 샘플 제조 동안에 도입된 편향으로 인해 핵산 표적의 집단 내에서 불량하게 특징 분석되거나 불량하게 제공되는 게놈 영역에 대한 서열 정보를 제공하기 위해 유용하다.

[0058] 본원에 기재된 바코드화 기술의 사용은 소정의 유전학적 마커 세트에 대한 개별 분이자 형태를 제공하는, 즉, 소정의 유전학적 마커(단일 마커와는 반대되는)가 개별 샘플 핵산 분자에 기인하도록 하는 고유 능력을 부여하고 변이체 공동작용 어셈블리를 통해 다수의 샘플 핵산 분자 들 중에서 및/또는 특적 염색체에 대해 보다 광범위하거나 심지어 보다 긴 범위의 추론된 개별 분자 형태를 제공한다. 이들 유전학적 마커는 특정 유전학적 유전자좌, 예를 들어, SNP와 같은 변이체를 포함할 수 있거나 이들은 짧은 서열을 포함할 수 있다. 추가로, 바코드화의 사용은 예를 들어, 혈류에 순환하는 종양 DNA의 검출 및 특징 분석을 위해 샘플로부터 추출된 총 핵산 집

단의 소수 성분과 주요 성분들을 식별하는 능력을 촉진시키는 추가의 이점을 제공하고 또한 임의의 증폭 단계 동안에 증폭 편향을 감소시키거나 제거한다. 추가로, 미세유체 포맷으로의 수행은 극히 작은 샘플 용적 및 DNA의 낮은 투입양을 사용한 작업 능력 및 게놈-광범위 태깅을 촉진시키기 위한 다수의 샘플 파티션(소적)을 신속히 처리하는 능력을 부여한다.

[0059] 이전에 기재된 바와 같이, 본원에 기재된 방법 및 시스템의 이점은 이들이 널리 가용한 짧은 판독 서열 분석 기술의 사용을 통해 목적하는 결과는 성취할 수 있다는 것이다. 상기 기술은 널리 특징 분석되고 고도로 효과적인 프로토콜 및 시약 시스템과 함께 연구 분야 내에서 용이하게 가용하고 널리 보급되어 있다는 이점을 갖는다. 이들 짧은 판독 서열 분석 기술은 예를 들어 하기 제조원으로부터 시판되는 것들 제조원: Illumina, inc.(GXII, NextSeq, MiSeq, HiSeq, X10), Ion Torrent division of Thermo-Fisher (Ion Proton and Ion PGM), 피로서열분석 방법 등을 포함한다.

[0060] 특히, 본원에 기재된 방법 및 시스템이 이들 짧은 판독 서열 분석 기술을 사용하고 이와 관련된 낮은 오류율로 수행한다는 것이 이점이다. 특히, 본원에 기재된 방법 및 시스템은 상기된 바와 같은 목적하는 개별 분자 판독 길이 또는 형태를 성취하지만 개별 서열 분석 판독과 함께 1000 bp 미만, 500 bp 미만, 300 bp 미만, 200 bp 미만, 150 bp 미만 또는 그 이하인 메이트 쌍(mate pair) 연장을 배제하고, 상기 개별 분자 판독 길이에 대한 서열 분석 오류율은 5% 미만, 1% 미만, 0.5% 미만, 0.1% 미만, 0.05% 미만, 0.01% 미만, 0.005% 미만, 또는 심지어 0.001% 미만이다.

[0061] *II. 작업흐름 개요*

[0062] 하나의 예시적 양상에서, 기재된 방법 및 시스템은 별개의 파티션에서 개별 샘플(예를 들어, 핵산)을 침적시키거나 분할시킴을 제공하고, 여기서, 각각의 파티션은 다른 파티션 내 내용물과 자신의 내용물의 분리를 유지한다. 본원에 사용된 바와 같은 파티션은 다양한 상이한 형태, 예를 들어, 웰, 튜브, 마이크로 또는 나노웰, 쓰루홀 등을 포함할 수 있는 컨테이너 또는 용기를 언급한다. 그러나, 바람직한 양상에서, 파티션은 유체 스트림 내에서 유동성이다. 이들 용기는 예를 들어, 내부 유체 센터 또는 코어를 둘러싸는 외곽 벽을 갖는 미세캡슐 또는 미세-소체로 이루어질 수 있거나, 이들은 매트릭스 내에 물질을 동반하고/하거나 보유할 수 있는 다공성 매트릭스일 수 있다. 그러나, 바람직한 양상에서, 이들 파티션은 비-수성 연속상, 예를 들어, 오일 상 내의 수성 유체의 소적을 포함할 수 있다. 다양한 상이한 용기가 예를 들어, 다음 문헌에 기재되어 있다: 미국.특허 출원 번호 제13/966,150호, 2013년 8월 13일자에 출원됨.또한, 비-수성 또는 오일 연속 상에서 안정한 소적을 생성하기 위한 에멀젼 시스템은 예를 들어, 하기 문헌에 상세히 기재되어 있다[문헌참조:.공개된 미국특허 출원 번호 제2010-0105112호].특정 경우에, 미세유동 채널 네트워크는 특히 본원에 기재된 바와 같은 입자를 생성시키기 위해 적합하다. 상기 미세유동 장치의 예는 하기 문헌에 상세히 기재된 것들을 포함한다: 미국특허 출원 번호 제14/682,952호, 2015년 4월 9일자로 출원됨, 이의 전체 기재내용은 모든 목적을 위해 전문이 본원에 참조로 인용됨. 이를 통해 세포의 수성 혼합물이 비-수성 유체로 압출되는 다공성 막을 포함하는 대안적 장치가 또한 개별 세포의 분할에 사용될 수 있다. 상기 시스템은 일반적으로 예를 들어, 하기 제조원으로부터 가용하다: Nanomi, Inc.

[0063] 에멀젼 중 소적의 경우에, 샘플 물질, 예를 들어, 핵산의 별개의 파티션으로의 분할은 일반적으로 유체, 예를 들어, 불소화된 오일의 비-수성 스트림이 또한 유동하여 수성 소적(상기 소적은 샘플 물질을 포함한다)이 상기 유동 스트림 분할 유체내에 생성되도록 접합부로 수성 샘플 함유 스트림을 유동시킴에 의해 성취될 수 있다. 하기된 바와 같이, 파티션, 예를 들어, 소적은 또한 전형적으로 동시-분할된 바코드 올리고뉴클레오타이드를 포함한다. 임의의 특정 파티션 내 샘플 물질의 상대적 양은 예를 들어, 수성 스트림 내 샘플의 농도, 수성 스트림 및/또는 비-수성 스트림의 유속 등에서 샘플의 농도를 포함하는 시스템의 다양한 상이한 파라미터를 조절함에 의해 조정될 수 있다. 본원에 기재된 파티션은 흔히 극히 작은 용적을 가짐을 특징으로 한다. 예를 들어, 소적 기반 파티션의 경우에, 소적은 1000 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 함께 동시 분할되는 경우, 파티션 내 샘플 유체 용적이 상기된 용적의 90% 미만, 80% 미만, 70% 미만, 60% 미만, 50% 미만, 40% 미만, 30% 미만, 20% 미만, 또는 심지어 상기된 용적의 10% 미만일 수 있는 것으로 평가된다. 일부 경우에, 낮은 반응 용적 파티션의 사용은 특히, 매우 소량의 출발 시약, 예를 들어, 투입 핵산과의 반응을 수행하는데 유리하다. 낮은 투입 핵산을 갖는 샘플을 분석하기 위한 방법 및 시스템은 하기 문헌참조에 제공된다: 미국특허 출원 번호 제14/752,602호, 2015년 6월 26일자로 출원됨, 이의 전체 기재내용은 이의 전문이 본원에 참조로 인용된다.

[0064] 본원에 기재된 방법 및 시스템에 따라 샘플이 이들 각각의 파티션으로 도입되면, 파티션 내 샘플 핵산은 일반적으로 고유 식별자와 함께 제공됨에 따라 상기 핵산의 특징 분석시 이들은 이들 각각의 오리진으로부터 유래된 것으로서 기인될 수 있다. 따라서, 샘플 핵산은 전형적으로 고유 식별자(예를 들어, 바코드 서열)와 동시-분할된다. 특히 바람직한 양상에서, 고유 식별자는 이들 샘플에 부착될 수 있는 핵산 바코드 서열을 포함하는 올리고뉴클레오타이드 형태로 제공된다. 올리고뉴클레오타이드는 소정의 파티션 내 올리고뉴클레오타이드 간에서와 같이 여기에 함유된 핵산 바코드 서열이 동일하지만 상이한 파티션 간에서는 올리고뉴클레오타이드가 상이한 바코드 서열을 가질 수 있지만 바람직하게 갖도록 분할된다. 바람직한 양상에서, 단지 하나의 핵산 바코드 서열은 소정의 파티션과 관련되지만 일부 경우에, 2개 이상의 상이한 바코드 서열은 존재할 수 있다.

[0065] 핵산 바코드 서열은 전형적으로 올리고뉴클레오타이드 서열 내 6 내지 약 20개이상의 뉴클레오타이드를 포함한다. 이들 뉴클레오타이드는 즉 인접한 뉴클레오타이드의 단일 스트레치 내에서 완전히연속적일 수 있거나 이들은 하나 이상의 뉴클레오타이드에 의해 분리된 2개 이상의 별도의 서브서열로 분리될 수 있다. 전형적으로, 분리된 서브서열은 전형적으로 약 4 내지 약 16개 뉴클레오타이드 길이일 수 있다.

[0066] 동시-분할된 올리고뉴클레오타이드는 또한 전형적으로 분할된 핵산의 프로세싱에유용한 다른 기능적 서열을 포함한다. 이들 서열은 예를 들어, 서열의 존재 확인을 위해 또는 바코드 핵산 또는 임의의 다수의 다른 잠재적 기능성 서열을 풀 다운하기 위한, 관련 바코드 서열, 서열 분석 프라이머, 하이브리드화 또는 프로브 서열을 부착시키면서, 파티션 내 개별 핵산으로부터 게놈 DNA를 증폭시키기 위한 무작위/범용 증폭 프라이머 서열을 포함한다. 다시, 샘플 물질과 함께 올리고뉴클레오타이드 및 관련 바코드 및 다른 기능성 서열의 동시-분할은 예를 들어, 하기 문헌에 기재되어 있다: 미국특허 출원 번호미국특허 출원 번호 제14/316,383호, 제14/316,398호, 제14/316,416호, 제14/316,431호, 제14/316,447호, 제14/316,463호, 모두 2014년 6월 26일자로 출원됨, 및 미국특허 출원 번호 제14/175,935호, 2014년 2월 7일자로 출원됨, 이의 전체 기재 내용은 이들의 전문이 참조로 인용된다.

[0067] 간략하게, 하나의 예시적 공정에서, 비드가 제공되고 각각은 비드에 방출가능하게부착된 다수의 상기된 올리고뉴클레오타이드를 포함할 수 있고 여기서 특정 비드에 부착된 모든 올리고뉴클레오타이드는 동일한 핵산 바코드 서열을 포함할 수 있지만 다수의 다양한 바코드 서열은 사용된 비드 집단을 가로질러 제공될 수 있다. 전형적으로, 비드 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, 각각의 비드는 전형적으로 다수의 올리고뉴클레오타이드 분자가 부착되어 제공된다. 특히, 개별 비드 상에 바코드 서열을 포함하는 다수의 올리고뉴클레오타이드 분자는 적어도 약 10,000개의 올리고뉴클레오타이드, 적어도 100,000개의 올리고뉴클레오타이드 분자, 적어도 1,000,000개의 올리고뉴클레오타이드 분자, 적어도 100,000,000개의 올리고뉴클레오타이드 분자 및 일부 경우에 적어도 10억개의 올리고뉴클레오타이드 분자일 수 있다.

[0068] 올리고뉴클레오타이드는 특정 자극을 비드에 적용시 비드로부터 방출될 수 있다. 일부 경우에, 상기 자극은 광-자극, 예를 들어, 올리고뉴클레오타이드를 방출시킬수 있는 광-불안정한 연결체의 절단을 통한 것일 수 있다. 일부 경우에, 열 자극이 사용될 수 있고, 여기서, 비드 환경의 온도의 상승은 비드로부터 연결체의 절단 또는 올리고뉴클레오타이드의 다른 방출을 유도할 수 있다. 일부 경우에, 비드에 대한 올리고뉴클레오타이드의 연결체를 절단하거나 비드로부터 올리고뉴클레오타이드의 방출을 유도할 수 있는 화학적 자극이 사용될 수 있다.

[0069] 본원에 기재된 방법 및 시스템에 따라, 부착된 올리고뉴클레오타이드를 포함하는 비드는 개별 샘플과 동시 분할되어 단일 비드 및 단일 샘플은 개별 파티션 내에 함유된다. 일부 경우에, 단일 비드 파티션이 요구되는 경우, 유체의 상대적 유속을 조절하여 평균적으로 파티션이 파티션 당 1개 미만의 비드를 함유하도록 함으로써 점유된 파티션이 주로 단일로 점유되도록 보장할 수 있다. 마찬가지로, 통상의 기술자는 보다 높은 %의 파티션이 점유되어 예를 들어, 단지 작은 %의 비점유된 파티션이 가능하도록 유속을 조절하고 싶을 수 있다. 바람직한 양상에서, 흐름 및 채널 구조는 목적하는 수의 단일로 점유된 파티션, 특정 수준 미만의 비점유된 파티션 및 특정 수준 미만의 다중 점유된 파티션을 보장하기 위해 조절된다.

[0070] 도 3은 샘플 핵산을 바코드화하고 후속적으로 서열 분석 하기 위해, 특히 카피수 변화 또는 반수체 검정용에 사용하기 위한 하나의 특정 예시적 방법을 설명한다. 먼저, 핵산을 포함하는 샘플은 공급원 300으로부터 수득될 수 있고 바코드화된 비드 세트 310이 또한 수득될 수 있다. 상기 비드는 바람직하게 무작위 N-량체와 같은 프라이머 또는 다른 프라이머와 같은 프라이머 뿐만 아니라 하나 이상의 바코드 서열을 함유하는 올리고뉴클레오타이드에 연결된다. 바람직하게, 바코드 서열은 바코드화된 비드로부터 방출될 수 있고, 예를 들어, 바코드와 비

드 간의 연결체의 절단을 통해 또는 바코드를 방출하도록 하부 비드의 분해를 통해 또는 상기 2개의 조합을 통해 방출될 수 있다. 예를 들어, 특정 바람직한 양상에서, 바코드화된 비드는 바코드 서열을 방출시키기 위한 환원제와 같은 제제에 의해 분해되거나 용해될 수 있다. 상기 예에서, 핵산 305, 바코드화된 비드 315, 및 임의로 다른 시약, 예를 들어, 환원제 320을 포함하는 낮은 양의 샘플이 조합되고 분할에 적용된다. 예를 들어, 상기 분할은 성분을 미세유동 장치 325와 같은 소적 생성 시스템에 도입함을 포함할 수 있다. 미세유동 장치325의 도움으로, 유중수 에멀젼 330이 형성될 수 있고, 여기서, 에멀젼은 샘플 핵산 305, 환원제 320, 및 바코드화된 비드 315를 함유하는 수성 소적을 함유한다. 상기 환원제는 바코드화된 비드를 용해시키거나 분해시켜 소적 335 내 비드로부터 바코드 및 무작위 N-량체를 갖는 올리고뉴클레오타이드를 방출시킬 수 있다. 이어서 무작위 N-량체는 샘플 핵산의 상이한 영역을 프라이밍하여 증폭 후 샘플의 증폭된 복제물을 유도하고, 각각의 카피는 바코드 서열 340으로 태그한다. 바람직하게, 각각의 소적은 동일한 바코드 서열 및 상이한 무작위 N-량체 서열을 함유하는 올리고뉴클레오타이드 세트를 함유한다. 후속적으로, 에멀젼 345는 부숴지고 추가의 서열(예를 들어, 특정 서열 분석 방법을 원조하는 서열, 추가의 바코드 등)은 예를 들어, 증폭 방법 350(예를 들어, PCR)을 통해 부가될 수 있다. 이어서, 서열 분석 355가 수행될 수 있고 알고리듬은 서열 분석 데이터 360을 해석하기 위해 적용된다. 서열 분석 알고리듬은 일반적으로 예를 들어, 바코드의 분석을 수행하여 서열 분석판독을 정렬하고/하거나 특정 서열 판독이 속하는 샘플을 동정할 수 있다. 추가로, 및 본원에 기재된 바와 같이, 이들 알고리듬을 또한 추가로 사용하여 복제물의 서열이 이들의 기원하는 분자 형태에 기인할 수 있도록 한다.

[0071] 상기된 바와 같이, 단일 비드 점유는 가장 요망되는 상태일 수 있지만, 다중 점유된 파티션 또는 비점유된 파티션이 흔히 존재할 수 있는 것으로 평가된다. 바코드 올리고뉴클레오타이드를 포함하는 샘플 및 비드를 동시 분할하기 위한 미세유동 채널 구조의 예는 도 4에서 도식적으로 설명된다. 나타낸 바와 같이, 채널 분절 402, 404, 406, 408 및 410은 채널 접합부 412에서 유체 소통하도록 제공된다. 개별 샘플 414를 포함하는 수성 스트림은 채널 접합부 412를 향해 채널 분절 402를 통해 유동한다. 본원의 다른 곳에서 기재된 바와 같이, 이들 샘플은 분할 공정 전에 수성 유체 내에 현탁될 수 있다.

[0072] 동시에, 비드 416을 갖는 바코드를 포함하는 수성 스트림은 채널 접합부 412를향해 채널 분절 404를 통해 유동한다. 비-수성 분할 유체는 측면 채널 406 및 408 각각으로부터 채널 접합부 412로 도입되고 조합된 스트림은 출구 채널 410으로 유동한다. 채널 접합부 412 내에, 채널 분절 402 및 404로부터 2개의 조합된 수성 스트림은 조합되고 동시 분할된 샘플 414 및 비드 416을 포함하는 소적 418로 분할된다. 이전에 주지된 바와 같이, 채널 접합부의 기하학을 조절하는 것 뿐만 아니라 채널 접합부 412에서 유체 조합 각각의 유동 특징을 조절함에 의해, 통상의 기술자는 생성된 파티션 418 내에 비드, 샘플 또는 둘다의 목적하는 점유 수준을 성취하기 위해 조합 및 분할을 최적화할 수 있다.

[0073] 평가된 바와 같이, 다수의 다른 시약은 예를 들어, 화학적 자극, 핵산 연장, 전사 및/또는 증폭 시약, 예를 들어, 폴리머라제, 역전사효소, 뉴클레오사이드 트리포스페이트 또는 NTP 유사체, 프라이머 서열 및 추가의 조인자, 예를 들어, 상기 반응에 사용되는 2가 금속 이온, 연결 반응 시약, 예를 들어, 리가제 효소 및 연결 서열, 염료, 표지 또는 다른 태깅 시약을 포함하는 샘플 및 비드와 함께 동시 분할될 수 있다.

[0074] 일단 동시 분할되면, 비드 상에 분배된 올리고뉴클레오타이드는 분할된 샘플을 바코드화하고 증폭시키기 위해 사용될 수 있다. 샘플을 증폭하고 바코드화하는데 있어서 이들 바코드 올리고뉴클레오타이드의 사용을 위한 특히 정교한 공정은 하기 문헌에 상세히 기재되어 있다: 미국특허 출원 번호 제14/316,383호, 제14/316,398호, 제14/316,416호, 제14/316,431호, 제14/316,447호, 제14/316,463호, 이 모두는 2014년 6월 28일자로 출원됨, 이의 완전한 기재내용은 이들의 전문이 참조로 본원에 인용된다. 간략하게, 하나의 양상에서, 올리고뉴클레오타이드는 샘플과 함께 동시 분할되고 샘플과 함께 이들의 비드로부터 파티션으로 방출되는 비드 상에 존재한다. 올리고뉴클레오타이드는 전형적으로 바코드 서열과 함께 이의 5’ 말단에 프라이머 서열을 포함한다. 상기 프라이머 서열은 무작위 또는 구조화된 것일 수 있다. 무작위 프라이머 서열은 일반적으로 샘플의 많은 상이한 영역을 무작위로 프라이밍하기 위해 의도된다. 구조화된 프라이머 서열은 일부 종류의 부분적으로 한정된 구조를 갖는 프라이머뿐만 아니라 샘플의 특이적 표적화 영역의 업스트림을 프라이밍하기 위해 표적화 한정된 서열을 포함하는 상이한 구조 범위를 포함할 수 있고, 이는 제한 없이 특정 %의 염기 (예를 들어, GC N-량체 %)를 함유하는 프라이머, 부분적으로 또는 전반적으로 퇴행성 서열을 함유하는 프라이머, 및/또는 본원에 임의의 기재에 따라 부분적으로 무작위 및 부분적으로 구조화된 서열을 함유하는 프라이머를 포함한다. 평가된 바와 같이, 상기 유형의 무작위 및 구조화된 프라이머의 임의의 하나 이상은 임의의 조합으로 올리고뉴클레오타이드에 포함될 수 있다.

[0075] 일단 방출되면, 올리고뉴클레오타이드의 프라이머 부분은 샘플의 상보적 영역으로 어닐링할 수 있다. 연장 반응

시약, 예를 들어,DNA 폴리머라제, 뉴클레오사이드 트리포스페이트, 보조 인자 (예를 들어,Mg2+ 또는 Mn2+ 등)은 또한 샘플 및 비드와 동시 분할되고 이어서 프라이머가 어닐링된 주형의 가닥에 상보적인 단편을 생성시키기 위해 주형으로서 샘플을 사용하는 프라이머 서열을 연장시키고 상보적 단편과 함께 올리고뉴클레오타이드 및 이의 관련된 바코드 서열을 포함한다. 샘플의 상이한 부분으로 다중 프라이머의 어닐링 및 연장은 샘플의 중첩 상보적 단편의 대형 풀을 수득할 수 있고, 상기 단편 각각은 이것이 생성된 파티션을 지적하는 자신의 바코드 서열을 소유한다. 일부 경우에, 이들 상보적 단편 자체는 바코드 서열을 다시 포함하는 상보체의 상보체를 생성하도록 파티션에 존재하는 올리고뉴클레오타이드에 의해 프라이밍된 주형으로서 사용될 수 있다. 일부 경우에, 상기 복제 공정은 제1 상보체가 복제되는 경우 이것이 이의 말단에서 또는 이의 근처에서 2개의 상보적 서열을 생성하도록 하여 추가의 반복적 복제물을 제조하기 위한 기반이 되는 분자의 능력을 감소시키는 헤어핀 구조 또는 부분적 헤어핀 구조의 형성을 허용하도록 구성된다. 이의 하나의 예의 도식적 도해는 도 5에 나타낸다.

[0076] 도면이 보여주는 바와 같이, 바코드 서열을 포함하는 올리고뉴클레오타이드는 샘플 핵산 504와 함께 에멀젼 중의 소적 502에서 동시 분할된다. 본원의 다른 곳에서 주지된 바와 같이, 올리고뉴클레오타이드 508은 샘플 핵산 504와 동시 분할되는 비드 506 상에 제공될 수 있고, 상기 올리고뉴클레오타이드는 바람직하게 패널 A에 보여지는 바와 같이 비드 506으로부터 방출될 수 있다. 올리고뉴클레오타이드 508은 하나 이상의 기능성 서열, 예를 들어, 서열 510, 514 및 516 뿐만 아니라 바코드 서열 512를 포함한다. 예를 들어, 올리고뉴클레오타이드 508은 소정의 서열 분석 시스템에 대한 부착 또는 고정화 서열로서 기능할 수 있는 서열 510, 예를 들어, Illumina Hiseq 또는 Miseq 시스템의 유동 세포에서 부착을 위해 사용되는 P5 서열 뿐만 아니라 바코드 서열 512를 포함하는 것으로서 보여진다. 보여지는 바와 같이, 올리고뉴클레오타이드는 또한 샘플 핵산 504의 일부의 프라이밍 복제를 위해 무작위 또는 표적화 N-량체를 포함할 수 있는 프라이머 서열 516을 포함한다. 또한 올리고뉴클레오타이드 508 내에 포함되는 것은 서열 분석 프라이밍 영역, 예를 들어, 서열 분석 시스템에서 합성 반응에 의해 폴리머라제 매개된, 주형 지시된 서열 분석을 프라이밍하기 위해 사용되는 “판독1” 또는 R1 프라이밍 영역을 제공할 수 있는 서열 514이다. 많은 경우에, 바코드 서열 512, 고정화 서열 510 및 R1 서열 514는 소정의 비드에 부착된 모든 올리고뉴클레오타이드에 통상적일 수 있다. 프라이머 서열 516은 무작위 N-량체 프라이머를 위해 다양할 수 있거나 특정 표적화 적용을 위한 소정의 비드 상의 올리고뉴클레오타이드에 통상적일 수 있다.

[0077] 프라이머 서열 516의 존재를 기준으로, 올리고뉴클레오타이드는 패널 B에서 나타난 바와 같은 샘플 핵산을 프라이밍할 수 있고, 이는 비드 506 및 샘플 핵산 504와 또한 동시 분할되는 폴리머라제 효소 및 다른 연장 시약을 사용한 올리고뉴클레오타이드 508 및 508a의 연장을 허용한다. 패널 C에 보여지는 바와 같이, 무작위 N-량체 프라이머에 대해 샘플 핵산 504의 다중 상이한 영역에 어닐링하는 올리고뉴클레오타이드의 연장 후; 핵산의 다중 중첩 상보체 또는 단편, 예를 들어, 단편 518 및 520이 생성된다. 샘플 핵산, 예를 들어, 서열 522 및 524의 부분에 상보적인 서열 부분을 포함하지만, 이들 작제물은 일반적으로 부착된 바코드 서열을 갖는, 샘플 핵산 504의 단편을 포함하는 것으로서 본원에 언급된다. 평가된 바와 같이, 상기된 바와 같은 주형 서열의 복제된 부분은 흔히 본원에서 상기 주형 서열의 “단편”으로서 언급된다. 그러나, 이전의 기재에도 불구하고, 용어 “단편”은 예를 들어, 효소적, 화학적 또는 기계적 단편화를 통해, 소정의 서열 분자의 실제 단편화와 같은 주형 서열의 부분을 제공하는 다른 장치에 의해 생성된 것들을 포함하는, 본래 핵산 서열의 일부, 예를 들어, 주형 또는 샘플 핵산의 임의의 표현을 포괄한다. 그러나, 바람직한 양상에서, 주형 또는 샘플 핵산 서열의 단편은 기본 서열 또는 이의 상보체의 복제된 부분을 지칭한다.

[0078] 이어서 바코드화된 핵산 단편은 예를 들어, 서열 분석을 통해 특징 분석에 적용될 수 있거나 이들은 패널 D에 나타낸 바와 같이 상기 공정에서 추가로 증폭될 수 있다. 예를 들어, 추가의 올리고뉴클레오타이드, 예를 들어, 비드 306으로부터 또한 방출되는 올리고뉴클레오타이드 508b는 단편 518 및 520을 프라이밍할 수 있다. 특히, 다시, 올리고뉴클레오타이드 508b의 무작위 N-량체 프라이머 516b(이는 많은 경우에 소정의 파티션에서 다른 무작위 N-량체와 상이한, 예를 들어, 프라이머 서열 516)의 존재를 기준으로, 올리고뉴클레오타이드는 단편 518과 어닐링하고 샘플 핵산 서열의 일부의 복제물을 포함하는 서열 528을 포함하는 단편 518의 적어도 일부에 상보체 526을 생성하기 위해 연장된다. 올리고뉴클레오타이드 508b의 연장은 이것이 단편 518의 올리고뉴클레오타이드 부분 508을 통해 복제될 때까지 계속한다. 본원의 다른 곳에 주지된 바와 같이, 그리고 패널 D에서 설명된 바와 같이, 올리고뉴클레오타이드는 예를 들어, 단편 518 내에 포함되는 올리고뉴클레오타이드 508의 완전한 서열 516 및 514를 복제한 후 목적하는 지점에서 폴리머라제에 의한 복제에서 중단을 촉진하도록 구성될 수 있다. 본원에 기재된 바와 같이, 이것은 예를 들어, 사용되는 폴리머라제 효소에 의해 가공될 수 없는 상이한 뉴클레오타이드 및/또는 뉴클레오타이드 유사체의 혼입을 포함하는 상이한 방법에 의해 성취될 수 있다. 예를 들어, 이것은 상기 영역의 복제를 중단시키는 비-우라실 내성 폴리머라제를 차단하기 위해 서열 영역 512 내에 우라실 함유 뉴클레오타이드의 내포를 포함할 수 있다. 결과로서, 하나의 말단에 전장 올리고뉴클레오타이드 508b를 포

함하는 단편 526이 생성되고 이는 바코드 서열 512, 부착 서열 510, R1 프라이머 영역 514 및 무작위 N-량체 서열 516b를 포함한다. 서열의 다른 말단에, 서열 514'로서 나타낸, R1 서열 모두 또는 일부에 대한 상보체 뿐만 아니라 제1 올리고뉴클레오타이드 508의 무작위 N-량체에 대한 상보체 516'가 포함된다. 이어서 R1 서열 514 및 이의 상보체 514'는 부분적 헤어핀 구조 528을 형성하도록 함께 하이브리드화할 수 있다. 평가된 바와 같이, 무작위 N-량체는 상이한 올리고뉴클레오타이드 간에 상이하기 때문에, 이들 서열 및 이들의 상보체는 헤어핀 형성에 참여할 것으로 예상되지 않고, 예를 들어, 무작위 N-량체 516에 대한 상보체인 서열 516'은 무작위 N-량체 서열 516b에 상보적인 것으로 예상되지 않는다. 이것은 다른 적용, 예를 들어, 표적화 프라이머에 대한 경우가 아니고, 여기서 N-량체는 소정의 파티션 내에서 올리고뉴클레오타이드 간에 통상적이다.

[0079] 이들 부분적 헤어핀 구조를 형성함에 의해, 이것은 복제물의 반복적 카피를 차단하면서, 추가의 복제로부터 샘플 서열의 제1 수준의 복제물의 제거를 허용한다. 부분적 헤어핀 구조는 또한 생성된 단편, 예를 들어, 단편 526의 후속적 프로세싱을 위해 유용한 구조를 제공한다.

[0080] 이어서 다수의 상이한 파티션으로부터의 모든 단편은 본원에 기재된 바와 같이 고속 처리 서열 분석이기 상에서의 서열 분석을 위해 수집될 수 있다. 각각의 단편은 오리진의 파티션에 관하여 코드화되기 때문에, 상기 단편의 서열은 바코드의 존재를 기준으로 다시 이의 오리진에 기인할 수 있다. 이것은 도 6에서 도식적으로 설명된다. 하나의 예에서 보여지는 바와 같이, 제1 공급원 600(예를 들어, 개별 염색체, 핵산 가닥 등)으로부터 기원하는 핵산 604 및 상이한 염색체 602 또는 핵산 가닥으로부터 유래된 핵산 606은 각각 상기된 바와 같이 이들 자신의 바코드 올리고뉴클레오타이드 세트와 함께 분할된다.

[0081] 각각의 파티션 내에서, 각각의 핵산 604 및 606은 이어서 제1 단편(들)의 제2 단편의 중첩 세트, 예를 들어, 제2 단편 세트 608 및 610을 별도로 제공하도록 프로세싱된다. 상기 프로세싱은 또한 특정 제1 단편으로부터 유래된 제2 단편 각각에 대해 동일한 바코드 서열과 함께 제2 단편을 제공한다. 나타낸 바와 같이, 제2 단편 세트 608에 대한 바코드 서열은 "1"로 지칭되고, 단편 세트 610에 대한 바코드 서열은 "2"로 지칭된다. 다양한 바코드 라이브러리는 다수의 상이한 단편 세트를 차등적으로 바코드화하기 위해 사용될 수 있다. 그러나, 상이한 제1 단편으로부터 모든 제2 단편 세트를 상이한 바코드 서열로 바코드화할 필요는 없다. 실제로, 많은 경우에, 다수의 상이한 제1 단편은 동일한 바코드 서열을 포함하도록 동시에 프로세싱될 수 있다. 다양한 바코드 라이브러리는 본원에서 그밖의 다른 곳에서 상세히 기재되어 있다.

[0082] 예를 들어, 단편 세트 608 및 610으로부터의 바코드화된 단편은 이어서 예를 들어, 제조원[Illumina or Ion Torrent division of Thermo Fisher, Inc.]으로부터 가용한 합성 기술에 의한 서열을 사용하는 서열 분석을 위해 수집될 수 있다. 일단 서열 분석되면, 서열 판독 612는 예를 들어, 적어도 부분적으로 포함된 바코드를 기준으로 및 임의로 및 바람직하게 부분적으로 단편 자체의 서열을 기준으로 집계된 판독 614 및 616에서 보여지는 바와 같이 이들 각각의 단편 세트에 기인할 수 있다. 각각의 단편 세트에 대해 기인된 서열 판독은 이어서 각각의 샘플 단편, 예를 들어, 618 및 620(이는 이어서 다시 이들 각각이 기원하는 염색체(600 및 602)에 추가로 기인될 수 있다)에 대해 어셈블리된 서열을 제공하기 위해 어셈블리된다. 게놈 서열을 어셈블리하기 위한 방법 및 시스템은 예를 들어, 하기 문헌에 기재되어 있다: 미국특허 출원 번호 제14/752,773호, 2015년 6월 26일자로 출원됨, 이의 개시내용은 이의 전문이 참조로 본원에 인용됨.

[0083] *III. 표적화 서열 분석으로 방법 및 시스템의 적용*

[0084] 하나의 양상에서, 본원에 기재된 시스템 및 방법은 표적화 게놈 영역으로부터 서열 정보를 수득하기 위해 사용된다.

[0085] 게놈의 "표적화" 영역(및 이의 임의의 문법적 균등물)이란 목적하는 바와 같이 동정되고/되거나 본원에 기재된 하나 이상의 방법을 통해 선택된 전체 게놈 또는 상기 게놈의 임의의 하나 이상의 영역을 의미한다. 본원에 기재된 방법 및 시스템에 의해 서열 분석된 게놈의 표적화 영역은 제한없이 인트론, 엑손, 유전자 사이 영역 또는 이의 임의의 조합을 포함한다. 특정 예에서, 본원에 기재된 방법 및 시스템은 전체 엑솜, 엑솜 일부, 하나 이상의 선택된 유전자(선택된 유전자 패널을 포함하는), 하나 이상의 인트론 및 인트론 및 엑손 서열의 조합에 대한 서열 정보를 제공한다.

[0086] 게놈의 표적화 영역은 또한 서열에 의해 동정된 영역 보다는 게놈의 특정 부분 또는 %를 포함할 수 있다. 특정 구현예에서, 본원에 기재된 방법에 따라 포획되고 분석된 게놈의 표적화 영역은 게놈의 1, 2, 5, 10, 15, 20, 25, 50, 100, 200, 250, 500, 750, 1000, 또는 10000 킬로베이스 마다 위치한 게놈의 부분들을 포함한다. 추가의 구현예에서, 게놈의 표적화 영역은 전체 게놈의 5%, 10%, 15%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%,

95%를 포함한다. 여전히 추가의 구현예에서, 표적화 영역은 전체 게놈의 1-10%, 5-20%, 10-30%, 15-40%, 20-50%, 25-60%, 30-70%, 35-80%, 40-90%, 또는 45-95%를 포함한다.

[0087] 일반적으로, 게놈의 표적화 영역은 당업계에 공지되고 본원에 기재된 임의의 서열 분석 방법에 사용하기 위해 포획된다. 본원에 사용된 바와 같은 "포획된"이란 핵산 및/또는 핵산 단편 집단을 집적시켜 상기 수득된 집단이 목적하지 않은 게놈 영역과 비교하여 목적하는 표적화 영역의 증가된 %를 함유하도록 함을 의미한다. 추가의 구현예에서, 집적된 집단은 표적화 영역을 포함하는 적어도 50%, 55%, 60%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99%, 또는 100%의 핵산/핵산 단편을 함유한다.

[0088] 포획 방법은 일반적으로 칩-기반 방법(여기서, 상기 표적화 영역은 표면 상의 포획 분자와의 하이브리드화 또는 다른 연합을 통해 포획된다) 및 용액 기반 방법(여기서, 표적화 영역(또는 표적화 영역에 가까운 부분)에 상보적인 올리고뉴클레오타이드 프로브(베이트)가 게놈 단편 라이브러리에 하이브리드화한다)을 통해 포획된다. 본원에 기재된 포획 방법에 사용되는 프로브는 일반적으로 이들이 하이브리드화하는 프로브 및 단편을 "풀 다운"하기 위해 사용될 수 있는 비오틴과 같은 포획 분자에 부착되고, 이들 풀 다운 방법은 목적하는 표적화 영역을 함유하는 핵산 또는 핵산 단편에 하이브리드화된 베이트가 목적하는 영역을 함유하지 않는 단편으로부터 분리되는 임의의 방법을 포함한다. 프로브가 비오티닐화된 구현예에서, 자기 스트렙타비딘 비드는 결합된 표적화 영역을 갖는 베이트를 선택적으로 풀 다운하고 집적시키기 위해 사용된다.

[0089] 추가의 양상에서, 추가의 연구를 위해 요구되는 모든 표적화 영역을 커버하는 베이트 라이브러리가 사용된다. 따라서, 전체 엑솜 분석의 경우에, 상기 베이트 라이브러리는 완전한 엑솜을 함께 커버하는 올리고뉴클레오타이드를 포함한다. 특정 구현예에서, 엑솜의 단지 일부만이 추가의 분석을 위해 요구된다. 상기 구현예에서, 상기 베이트는 상기 엑솜 서브세트를 표적화하기 위해 디자인된다. 상기 디자인은 당업계에 공지된 방법 및 알고리듬을 사용하여 성취될 수 있고 일반적으로 인간 게놈과 같은 참조 서열을 기준으로 한다.

[0090] 일부 예에서, 본원에 기재된 방법 및 시스템에 따라 프로세싱되고 서열 분석된 표적화 게놈 영역은 완전하거나 부분적인 엑솜이다. 이들 완전하거나 부분적인 엑솜은 당업계에 공지된 임의의 방법을 사용한 서열 분석을 위해 포획될 수 있고, 제한 없이 임의의 Roche/NimbleGen 엑솜 프로토콜을 포함하고, 이는 NimbleGen 2.1M 인간 엑솜 어레이 및 NimbleGen SeqCap EZ 엑솜 라이브러리, 임의의 Agilent SureSelect 제품, 임의의 Illumina 엑솜 포획 제품을 포함하고, 이는 TruSeq 및 Nextera 엑솜 제품, 및 당업계에 공지된 임의의 다른 제품, 방법, 시스템 및 프토토콜을 포함한다.

[0091] 추가의 구현예에서, 목적하는 표적화 영역이 엑솜 전체 또는 일부를 포함하는 경우, 상기 표적화 영역을 포획하기 위해 사용되는 베이트는 상기 엑손 서열에 상보적이도록 디자인될 수 있다. 다른 구현예에서, 상기 베이트는 엑손 서열 자체에는 상보적이지 않지만 대신 엑손 서열 또는 2개 엑손 사이의 인트론 서열 부근의 서열에 상보적이다. 상기 디자인은 또한 본원에서 "앵커된 엑솜 포획" 또는 "인트론 베이팅"으로 언급되고, 이는 본원에 논의된 바와 같이, 하나 이상의 엑솜 부분이 목적하는 하나 이상의 엑솜 부분 근처 또는 인접한 하나 이상의 인트론 서열에 상보적인 베이트의 사용을 통해 포획된다. 예를 들어, 도 2에 도식적으로 설명된 바와 같이, 게놈 서열 201은 엑손 영역 202 및 203을 포함한다. 상기 엑손 영역은 인근의 하나 이상의 인트론 서열(예를 들어, 엑손 영역 202를 포획하기 위한 인트론 영역 204 및/또는 205 및 엑손 영역 203의 포획을 위한 인트론 영역 206)에 지시된 베이트를 사용함에 의해 포획될 수 있다. 다른 말로, 엑손 영역 202 또는 203을 포함하는 단편 집단은 인트론 영역 204 및/또는 205 및 206에 상보적인 베이트의 사용을 통해 포획된다. 일부 구현예에서, 인트론 베이팅을 사용하여 보다 긴 인트론을 드문드문 베이팅에 의해 긴 인트론 영역에 의해 분리된 엑손을 브릿징시킨다. 상기 구현예에서, 상기 베이트는 필연적으로 목적하는 엑손 영역에 인접한 표적화 인트론 영역이 아니지만 상기 베이트는 대신 특정 거리(또는 거리 세트)에 의해 분리된 영역을 표적화하기 위해 디자인되거나 특정 수의 염기 또는 다수의 염기의 조합에 의해 인트론 영역에 걸쳐 타일화하도록 디자인된다. 상기 구현예는 하기에 추가로 상세히 기재된다.

[0092] 일부 구현예에서, 본 발명의 앵커된 엑솜 포획/인트론 베이팅 기술에 대해 사용되는 인트론 영역은 포획될 엑손 영역에 인접해 있다. 추가의 구현예에서, 인트론 영역들은 약 1-50, 2-45, 3-40, 4-35, 5-30, 6-25, 7-20, 8-15, 9-10, 2-20, 3-15, 4-10, 5-30, 10-40, 15-50, 20-75, 25-100개 뉴클레오타이드에 의해 포획될 엑손 영역으로부터 분리되어 있다. 여전히 추가의 구현예에서, 인트론 영역은 약 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, 300, 400, 또는 500개 뉴클레오타이드에 의해 포획될 엑손 영역으로부터 분리되어 있다. 추가의 구현에에서, 특히, 인트론 영역의 희박한 베이팅이 사용 중인 상황에 대해(예를 들어, 대형 인트론 길이를 따라 연결된 엑손 영역들의 단계 변이체 검출 또는 동정에 대해), 인트론 영역은 예를 들어, 하기

킬로베이스 정도의 거리로 포획될 엑손 영역으로부터 분리되어 있다. 1-20, 2-18, 3-16, 4-14, 5-12, 6-10 킬로베이스. 집적된 올리고뉴클레오타이드 집단의 본래의 분자 형태가 보유되기 때문에, 인트론 영역의 상기 희박한 베이팅은 긴 인트론에 의해 분리된 엑손 영역 사이의 서열의 연결을 허용한다.

[0093] 추가의 양상에서, 게놈의 특정 영역을 표적화하는 베이트를 디자인하기 보다는 타일링 방법이 사용된다. 상기 방법에서, 특정 엑손 또는 인트론 영역을 표적화하기 보다는, 베이트는 대신특정 범위의 거리에서 게놈 부분에 상보적이도록 디자인된다. 예를 들어, 베이트 라이브러리는 게놈을 따라 5킬로베이스(kb) 마다 위치된 서열에 하이브리드화하도록 디자인되어 이러한 베이트 라이브러리의 단편화된 게놈 샘플로의 적용은 상기 게놈의 특정 서브세트 - 즉, 베이트에 상보적인 서열을 함유하는 단편에 함유된 영역들만을 포획할 수 있도록 할 것이다. 평가된 바와 같이, 베이트는 인간 게놈 참조 서열과 같은 참조 서열을 기반으로 디자인될 수 있다. 추가의 구현예에서, 타일화된 베이트 라이브러리는 게놈의 1, 2, 5, 10, 15, 20, 25, 50, 100, 200, 250, 500, 750, 1000, 또는 10000 킬로베이스 마다 영역을 포획하도록 디자인된다. 일부 예에서, 타일화 방법은 드문 인트론 영역을 포획하는 효과를 가져 긴 인트론 영역에 의해 분리된 엑손 영역들의 서열 정보를 연결하기 위한 방법을 제공할 수 있는데 그 이유는 인트론 영역의 희박한 포획을 통해 포획된 상기 엑손 영역들의 본래의 분자 형태가 보유되기 때문이다.

[0094] 여전히 추가의 구현예에서, 상기 베이트는 무작위 또는 조합 방식으로 게놈을 타일화하도록 디자인되고, 예를 들어, 타일화된 라이브러리의 혼합물이 사용될 수 있고, 여기서, 상기 라이브러리의 일부는 1kb 마다 영역을 포획하는 반면 혼합물 중 다른 라이브러리는 100 kb 마다 영역을 포획한다. 여전히 추가의 구현예에서, 타일화된 라이브러리는 베이트가 게놈 내 특정 범위의 위치내 표적화하도록 디자인되고, 예를 들어, 상기 베이트는 게놈의 1-10, 2-5, 5-200, 10-175, 15-150, 20-125, 30-100, 40-75, 50-60 kb 모두의 영역을 표적화할 수 있다. 추가의 예에서, 본원에 기재된 타일화되거나 다른 포획 방법은 전체 게놈의 약 5%, 10%, 15%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%를 포획한다. 평가된 바와 같이, 상기 포획의 타일화 방법은 서열 분석과 같은 추가의 분석을 위해 게놈의 인트론 및 엑손 영역 둘다를 포획한다.

[0095] 여전히 추가의 구현예에서 그리고 본원에 기재된 임의의 방법에 따라, 본 발명의 방법에 사용되는 베이트 라이브러리는 본원에서 추가로 기재된 바와 같이 하나 이상의 특징을 충족하는 정통한 디자인의 생성물이다. 상기 정통한 디자인은 베이트 라이브러리는 정통한 단일 뉴클레오타이드 다형태(SNP)에 지시된다. 상기 논의된 바와 같이, 본원에 사용된 바와 같은 용어 “정통한 SNP”는 이종접합성인 SNP를 언급한다. 일부 예에서 베이트 라이브러리는 정통한 SNP를 함유하는 게놈 샘플 영역으로 지시된 다수의 프로브를 함유하도록 디자인된다. 본원에 사용된 바와 같은 “으로 지시된”이란 프로브가 게놈 서열의 상기 영역에 상보적인 서열을 함유함을 의미한다. 정통한 베이트 디자인은 완전한 커버의 표적화 집적을 허용하여 표적화 서열 분석 방법을 최적화하는 능력을 제공하여 동시에 요구되는 프로브의 수를 감소시킨다(따라서, 비용을 감소시키고 작업흐름을 간소화한다)..

[0096] 일반적으로, 정통한 베이트 디자인을 사용하는 방법에 대해, 베이트 라이브러리는 상기 정통한 SNP의 영역 및/또는 위치(들)에서 정통한 SNP의 존재 또는 부재를 기반으로 게놈의 표적화 영역에서 특정 서열로 지시되는 베이트를 포함하도록 디자인된다. 정통한 베이트 다지인에 대한 일반적 고려에 대한 예시적 설명은 도 8에 제공된다. 게놈 801의 영역은 엑손(802 및 803)을 포함할 수 있다. 일부 예에서, 정통한 SNP 804는 엑손(802)와 인접한 인트론 사이의 경계선에 위치한다. 상기 상황에서, 베이트 라이브러리는 경계선으로부터 이탈한 특정 거리에서 하나이상의 뉴클레오타이드(805)로 지시된 프로브를 포함하도록 디자인될 수 있다. 엑손과 인접한 인트론(806) 사이의 경계선에서 정통한 SNP가 없는 추가의 예에서, 베이트 라이브러리는 경계선 근처의 인트론 (807 및 808)에 하나 이상의 위치에 지시된 프로브를 포함하도록 디자인될 수 있다. 상기 위치는 아마도 정통한 SNP를 포함하지만 또한 다른 SNP 및/또는 경우에 따라 다른 서열을 포함할 수 있다. 엑손 803이 엑손의 내부에 정통한 SNP 809를 함유하지만 경계선에는 정통한 SNP가 없는 여전히 추가의 예에서, 베이트 라이브러리는 정통하고 정통하지 않은 SNP(경우에 따라 임의의 다른 서열 뿐만 아니라)의 혼합물을 포함하는 인접한 인트론 내 여러 위치 810, 811, 및 812로 지시된 프로브를 포함하도록 디자인될 수 있다.

[0097] 일부 양상에서, 하나 이상의 투입 특징을 사용하여 다양한 영역에서 맵 품질 뿐만 아니라 상기 투입 특징을 기반으로 게놈을 따라 위치를 쉬프팅하도록 지시된 프로브 베이트 라이브러리를 디자인한다. 상기 디자인은 일반적으로 인트론 및 엑손 위치 상에서 보다는 정통한 SNP 사이의 공간을 기준으로 한다. 그러나, 인트론 및 엑손 위치를 기준으로 베이트 디자인과 관련하여 본원에 제공된 임의의 기재내용은 또한 정통한 SNP를 기준으로 하는 정통한 베이트 디자인 방법과 조합하여 사용될 수 있다. 정통한 베이트 디자인에 사용되는 투입 특징은 제한 없이 그리고 임의의 조합으로 엑손, 인트론, 유전자 사이 영역, 정통한 SNP, 및 반복 서열 영역(예를 들어, GC-풍

부 영역), 센트로미어 및 샘플 핵산 길이를 포함한다.

[0098] 논의의 편의를 위해, 정통한 디자인 프로브 라이브러리의 상이한 특징은 상이한잠재적 구현예 측면에서 하기에 기재된다. 평가된 바와 같이, 본원에 논의된 임의의 프로브 라이브러리는 상기 논의된 임의의 정통한 디자인 요소 또는 임의의 다른 유형의 디자인을 사용하든 상관 없이 단일로 또는 임의의 조합으로 사용될 수 있다. 사용된 디자인 요소는 목적하는 영역에 대한 샘플 투입 및 맵핑의 품질 뿐만 아니라 목적하는 표적화 게놈 영역을 기반으로 선택된다.

[0099] 일부 구현예에서, 프로브 베이트 라이브러리는 소정의 샘플에서 정통한 SNP를 함유하는 높은 가능성을 갖는 영역으로 지시된 프로브를 포함하도록 디자인된다. 상기 표적은 정통한 SNP에 근접하거나 인접한 개별 염기(정통한 SNP 자체) 또는 하나 이상의 염기를 포함할 수 있다. 여전히 추가의 구현예에서, 프로브 베이트에 대한 표적은 직접적으로 정통한 SNP에 인접할 수 있거나 정통한 SNP로부터 약 1-200, 10-190, 20-180, 30-170, 40-160, 50-150, 60-140, 70-130, 80-120, 90-100 염기의 거리에 의해 분리될 수 있다.

[0100] 추가의 구현예에서, 프로브 베이트 라이브러리는 핵산 분자의 평균 길이와 관련된 특정 밀도 영역으로 지시된 프로브를 포함한다. 예를 들어, 프로브는 프로브가 하이브리드화하는 핵산 분자/단편의 평균 길이 보다 x배 더 밀도가 높은 표적 서열 밀도로 프로브를 포함하도록 디자인될 수 있고, 여기서, x는 제한 없이 1, 5, 10, 20, 50, 75, 100, 125, 150, 또는 200일 수 있다. 핵산의 길이에 상대적으로 프로브 표적의 밀도를 증가시키는 것은 동일한 물리적분자 상의 유전자좌를 거쳐 프로브를 연결시키는 능력을 증가시킨다. 상기 방법은 또한 연결된 영역이 정통한 SNP를 포함할 가능성을 개선시킴에 따라서 프로브 베이트 라이브러리를 게놈의 표적화 영역으로 부착시키는 능력을 추가로 개선시킨다.

[0101] 프로브 표적의 밀도는 또한 목적하는 소정의 영역에서 정통한 SNP의 높은 가능성이 없는 상황(집단 수준에서)에서 증가될 수 있다. 상기 영역에서, 본원에 기재된 것들과 같은 타일화 방법은 영역을 따라 주기적 공간에서 프로브를 지시하기 위해 사용될 수 있다. 특정 구현예에서, 공간 밀도는 차등적으로 편향되어, 정통한 SNP가 없는 이들 영역에서 프로브 공간 밀도가 정통한 SNP를 함유하는 영역에서 프로브 공간 보다 1, 2, 5, 10, 25, 50-배 짧은 거리에 있도록 한다.

[0102] 추가의 구현예에서, 프로브 베이트 라이브러리는 유전자(엑손 및 인트론을 포함하는) 내 단지 정통한 SNP 분포를 고려하도록 디자인된다. 상기 디자인 방법은 유전자의 하나의 말단으로부터 다른 하나로 연결/단계화시키기 위한 주요 위치에서 충분한 수의 이종접합성 SNP를 포획하도록 지시된다. 상기 디자인 방법은 엑손 정통한 SNP와 하나 이상의 비-엑손 SNP를 조합하는 표적 세트에 지시된 베이트를 포함하여 유전자 내 정통한 SNP 사이의 거리는 상기된 공간 밀도 미만이도록 한다.

[0103] 상기 정통한 디자인 방법은 게놈의 일반적인 표적화 영역의 검출을 가능하게 할뿐만 아니라 전위 및 유전자 융합과 같은 게놈 구조적 변화의 검출 및 단계화를 가능하게 한다. 임의의 개별 유전자가 단계화될 수 있도록 보장함에 의해, 막대한 다수의 유전자융합 반응이 본원에 기재된 방법을 사용하여 검출되고 단계화될 수 있도록 한다.

[0104] 특정 구현예에서 그리고 상기 임의의 방법에 따라, 베이트 라이브러리는 약 1 kb 내지 약 2 Mb의 거리에서 프로브를 표적화하도록 디자인된다. 추가의 구현예에서, 거리는 약 1-50, 5-45, 10-40, 15-35, 20-30, 10-50 kb이다.

[0105] 추가의 구현예에서, 프로브 베이트에 의해 표적화되는 핵산 단편은 약 2 kb 내지 약 250 Mb이다. 여전히 추가의 구현예에서, 단편은 약 10-1000, 20-900, 30-800, 40-700, 50-600, 60-500, 70-400, 80-300, 90-200, 100-150, 50-500, 25-300 kb이다.

[0106] 일부 구현예에서, 프로브 베이트 라이브러리는 프로브의 약 60 내지 95%가 정통한 SNP를 함유하는 서열에 하이브리드화하도록 디자인된다. 추가의 구현예에서, 프로브 베이트 라이브러리는 프로브 라이브러리에서 프로브의 약 65% - 85%, 70% - 80%, 60-90%, 80-90%, 90-95%, 95%-99%가 정통한 SNP에 하이브리드화하도록 디자인된다. 여전히 추가의 구현예에서, 프로브 라이브러리에서 프로브의 적어도 65%, 75%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%가 정통한 SNP에 하이브리드화하도록 디자인된다. 평가된 바와 같이, 정통한 SNP에 “하이브리드화 하도록” 디자인된 프로브에 대해서는 상기 프로브가 정통한 SNP를 포함하는 서열 영역에 하이브리드화하도록 함을 의미한다.

[0107] 추가의 구현예에서, 프로브 베이트 라이브러리는 게놈 샘플의 표적화 부분에서 엑손 및 인트론 둘다 내에 위치

하는 정통한 SNP로 지시된 다수의 프로브를 포함하도록 디자인된다.

[0108] 여전히 추가의 구현예에서, 라이브러리는 라이브러리 내 다수의 프로브가 약 1-15, 5-10, 3-6 kb에 의해 이격된 공간의 정통한 SNP에 하이브리드화하도록 디자인된다. 여전히 추가의 구현예에서, 프로브 라이브러리 내 다수의 프로브는 약 1, 3, 5, 10, 20, 30, 50 kb에 의해 이격된 정통한 SNP에 하이브리드화하도록 추가로 디자인된다.

[0109] 추가의 구현예에서, 프로브 라이브러리내 다수의 프로브가 디자인되고 엑손과 인트론의 경계선의 5-300, 10-50, 20-100, 30-150, 또는 40-200 kb 내 어떠한 정통한 SNP가 없는 게놈 샘플의 표적화 부분에 대해 다수의 프로브가 상기 경계선내 인트론 내 정통한 SNP에서 하이브리드화하도록 디자인된다.

[0110] 추가의 구현예에서, 프로브 라이브러리 내 다수의 프로브가 디자인되고 엑손 내 제1 정통한 SNP가 있고 제1 정통한 SNP가 인접한 인트론 및 상기 인접한 인트론내 제2 정통한 SNP와의 경계선으로부터 5-300, 10-50, 20-100, 30-150, 또는 40-200 kb에 위치하고 제2 정통한 SNP가 경계선으로부터 10 내지 50 kb에 위치하는 게놈 샘플의 표적화 부분에 대해, 다수의 프로브는 제1 및 제2 정통한 SNP 사이에서 게놈 샘플의 영역으로 하이브리드화하도록 디자인된다;

[0111] 추가의 구현예에서, 프로브 라이브러리내 다수의 프로브가 디자인되고, 적어도 5-300, 10-50, 20-100, 30-150, 또는 40-200 kb에 대해 어떠한 정통한 SNP를 포함하지 않는 게놈 샘플의 표적화 부분에 대해, 다수의 프로브는 게놈 샘플의 상기 표적화 부분에 대해 0.5, 1, 3, 또는 5 kb 마다 하이브리드화하도록 디자인된다. 추가의 구현예에서, 다수의 프로브는 게놈 샘플의 표적화 부분을 따라 0.1, 0.5, 1, 1.5, 3, 5, 10, 15, 20, 30, 35, 40, 45, 50 kb 마다 하이브리드화하도록 디자인된다.

[0112] 추가의 구현에에서, 프로브 라이브러리 내 다수의 프로브가 디자인되고, 엑손과 인트론 사이의 경계선의 5-300, 10-50, 20-100, 30-150, 또는 40-200 내 어떠한 정통한 SNP가 없는 게놈 샘플의 표적화 부분에 대해, 다수의 프로브는 엑손-인트론 경계선 다음에 가장 인접한 정통한 SNP에 하이브리드화하도록 디자인된다.

[0113] 추가의 구현예에서, 프로브 라이브러리는 바코드를 거쳐 연결 정보를 제공하는 밀도로 엑손을 플랭킹하는 게놈 샘플의 영역에 하이브리드화하도록 디자인된 프로브를 포함한다.

[0114] 여전히 추가의 구현예에서, 프로브 라이브러리에 의해 나타나는 커버 범위는 별개의 파티션에서 게놈 샘플의 개별 핵산 단편 분자의 길이 분포에 역비례하여 보다 높은 비율의 보다 긴 개별 핵산 단편 분자들을 함유하는 방법은 보다 작은 커버 범위로 프로브 라이브러리를 사용한다.

[0115] 여전히 추가의 구현예에서, 프로브 라이브러리는 게놈 샘플의 표적화 영역의 커버를 위해 최적화된다. 여전히 추가의 구현예에서, 커버 밀도는 높은 맵 품질의 영역, 특히 정통한 SNP를 함유하는 영역에 대해 보다 낮을 수 있고 상기 밀도는 추가로 낮은 맵 품질의 영역에 대해 추가로 보다 높을 수 있어 연결 정보가 표적화 영역을 따라 확실히 제공되도록 한다.

[0116] 여전히 추가의 구현예에서, 프로브 라이브러리는 게놈 샘플의 하나 이상의 표적화 부분을 특징으로 하는 것으로 알려진 특징을 가져, 높은 맵 품질과 함께 표적화 부분에 대해, 프로브 라이브러리는 엑손과 인트론의 1kb-1Mb의 경계선 내 정통한 SNP에 하이브리드화하는 프로브를 포함한다. 프로브 라이브러리는 상기 상황에서 엑손과 인트론의 경계선의 10-500, 20-450, 30-400, 40-350, 50-300, 60-250, 70-200, 80-150, 90-100 kb 내 정통한 SNP에 하이브리드화하는 프로브를 추가로 포함한다.

[0117] 여전히 추가의 구현예에서, 프로브 라이브러리는 게놈 샘플의 하나 이상의 표적화 부분이 특징인 것으로 알려진 특징을 가져, 바코드화된 단편의 길이의 분포가 약 100, 150, 200, 250 kb 보다 긴 단편의 높은 비율을 갖는 표적화 부분에 대해, 프로브 라이브러리는 적어도 50 kb에 의해 분리된 정통한 SNP에 하이브리드화하는 프로브를 포함한다. 프로브 라이브러리는 상기 상황에서 적어도 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 75, 100, 125, 150, 175, 200 kb에 의해 분리된 정통한 SNP에 하이브리드화하는 프로브를 추가로 포함한다.

[0118] 여전히 추가의 구현예에서, 프로브 라이브러리는 게놈 샘플의 하나 이상의 표적화 부분이 특징인 것으로 알려진 특징을 가져, 낮은 맵 품질을 갖는 표적화 부분에 대해, 프로브 라이브러리는 엑손-인트론 경계선 1kb 내에 정통한 SNP에 하이브리드화하는 프로브를 포함한다. 프로브 라이브러리는 상기 상황에서 엑손-인트론 경계선의 2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 75, 100, 125, 150, 175, 200 kb내에 정통한 SNP에 하이브리드화하는 프로브를 추가로 포함한다. 상기 상황에서, 라이브러리는 하이브리드화하는 프로브 및 액손 내, 인트론 내 둘다 내에 정통한 SNP에 하이브리드화하는 프로브를 추가로 포함한다.

[0119] 여전히 추가의 구현예에서, 프로브 라이브러리는 게놈 샘플의 하나 이상의 표적화 부분이 특징인 것으로 알려

진 특징을 가져, 유전자 사이 영역을 포함하는 표적화 부분에 대해, 프로브 라이브러리는 적어도 1, 2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 75, 100 kb의 거리로 이격되어 있는 정통한 SNP에 하이브리드화하는 프로브를 포함한다.

[0120] 본원에 기재된 포획 방법에 사용되는 베이트는 게놈의 표적화 영역을 함유하는 단편 집단을 집적시키기 위해 유용한 임의의 크기 또는 구조일 수 있다. 상기 논의된 바와 같이, 일반적으로 본 발명에 사용되는 베이트는 비오틴과 같은 같은 포획 분자에 부착된 올리고뉴클레오타이드 프로브를 포함한다. 올리고뉴클레오타이드 프로브는 목적하는 표적화 영역 내 서열에 상보적일 수 있거나 이들은 표적화 영역 외부이지만 “앵커링” 영역 및 표적화 영역 둘다가 동일한 단편 내에 있는 상기 표적화 영역에 충분히 가까운 영역에 상보적일 수 있어 베이트는 상기 인근 영역(예를 들어, 플랭킹 인트론)에 하이브리드화함에 의해 표적화 영역을 풀 다운시킬 수 있다.

[0121] 상기 베이트에 부착된 포획 분자는 집단 중 다른 단편으로부터 베이트 및 이의 하이브리드화 파트너를 단리시키기 위해 사용될 수 있는 임의의 포획 분자일 수 있다. 일반적으로, 본원에 사용된 베이트는 비오틴에 부착되고 이어서 스트렙타비딘(제한 없이 자기 스트렙타비딘 비드를 포함하는)을 포함하는 고형 지지체를 사용하여 이들이 하이브리드화하는 베이트 및 단편을 포획할 수 있다. 다른 포획 분자 쌍은 제한 없이 비오틴/뉴트라비딘, 항원/항체 또는 상보적 올리고뉴클레오타이드 서열을 포함할 수 있다.

[0122] 추가의 구현예에서, 베이트의 올리고뉴클레오타이드 프로브 부분은 표적화 영역또는 표적화 영역 근처 영역에 하이브리드화하기 위해 적합한 임의의 길이일 수 있다. 일부 구현예에서, 본원에 기재된 방법에 따라 사용되는 베이트의 올리고뉴클레오타이드 프로브 부분, 즉 게놈의 표적화 영역 또는 표적화 영역 근처의 영역에 하이브리드화하는 부분은 일반적으로 약 10 내지 약 150개 뉴클레오타이드 길이(예를 들어, 35개 뉴클레오타이드, 50개 뉴클레오타이드, 100개 뉴클레오타이드)를 갖고 목적하는 표적 서열에 특이적으로 하이브리드화하도록 선택된다. 추가의 구현예에서, 올리고뉴클레오타이드 프로브 부분은 약 5-10, 10-50, 20-100, 30-90, 40-80, 50-70개 뉴클레오타이드 길이를 포함한다. 평가된 바와 같이, 본원에 기재된 임의의 올리고뉴클레오타이드 프로브 부분은RNA, DNA, 비-천연 뉴클레오타이드, 예를 들어, PNA, LNA 등 또는 이의 조합체를 포함할 수 있다.

[0123] 본원에 기재된 방법 및 시스템의 이점은 포획된 표적화 영역이, 표적화 영역을 포획하고 서열 분석을 수행하는 단계 후에도 이들 표적화 영역의 본래의 분자 형태가 보유되는 방식으로 포획 전에 프로세싱된다. 특정 표적화 영역이 이들의 본래의 분자 형태(이들이 유래된 본래의 염색체 또는염색체 영역 및/또는 완전한 게놈 내 서로 관련하여 특정 표적화 영역의 위치를 포함할 수 있다)에 기인하는 능력은 달리 불량하게 맵핑되거나 통상의 서열 분석 기술을 사용한 불량한 커버를 갖는 게놈의 영역으로부터 서열 정보를 수득하는 방식을 제공한다.

[0124] 예를 들어, 일부 유전자들은 일반적으로 가용한 서열 분석 기술을 사용하여, 특히 짧은 판독 기술을 사용하여 포괄하기에는 너무 긴 인트론을 갖는다. 짧은 판독 기술은 흔히 바람직한 서열 분석 기술인데 그 이유는 이들이 긴-판독기술과 비교하여 보다 우수한 정확도를 갖기 때문이다. 그러나, 일반적으로 사용되는 짧은 판독 기술은 게놈의 긴 영역을 포괄할 수 없고, 따라서 정보는 탠덤 반복 서열, 높은 GC 함량 및 긴 인트론을 함유하는 엑손과 같은 구조적 특징으로 인해 특징 분석하기 어려운 게놈 영역에서 이들 통상의 기술을 사용하여 수득할 수 없다. 그러나, 본원에 기재된 방법 및 시스템에서, 표적화 영역의 분자 형태는 일반적으로 도 1에서 설명되고 본원에서 추가로 상세히 기재된 태깅 과정을 통해 보유된다. 이와 같이, 연결은 게놈의 확대된 영역을 거쳐 만들어질 수 있다. 예를 들어, 도 2b에서 도식적으로 설명된 바와 같이, 핵산 분자 207은 긴 인트론 영역(208)과 함께 2개의 엑손(음영 막대)을 함유한다. 본원에 기재된 방법에서, 개별 핵산 분자 207은 이 자신의 별개의 파티션 211로 분포시키고 이어서 상이한 단편이 엑손 및 인트론의 상이한 부분을 함유하도록 단편화된다. 상기 단편 각각은 단편으로부터 수득된 임의의 서열 정보가 이어서 이것이 생성되는 별개의 파티션에 기인될 수 있도록 태깅되기 때문에, 각각의 단편은 또한 이것이 유래된 개별 핵산 분자 207에 기인할 수 있다.

[0125] 일반적으로 그리고 본원에서 추가로 상세히 기재된 바와 같이, 단편화 및 태깅 후, 상이한 파티션으로부터의 단편은 함께 조합된다. 이어서 표적화 포획 방법은 목적하는 표적화 영역을 함유하는 단편과 함께 , 추가의 서열 분석과 같은 추가의 분석을 받는 단편 집단을 집적시키기 위해 사용될 수 있다. 도 2b에 도시된 예에서, 사용된 베이트는 엑손 부분을 함유하는 것들만을 포획하기 위해 단편 집단을 집적시키지만 엑손 및 인트론 (예를 들어, 209 및 210)의 외부 영역은 포획되지 않는다. 따라서, 서열 분석 받는 최종 단편 집단은 엑손이 긴 인트론 영역에 의해 분리되어 있더라도 엑손 부분을 함유하는 단편에 대해 집적된다. 이어서, 짧은 판독, 높은 정확도 서열 분석 기술을 사용하여 상기 단편의 집적된집단의 서열을 동정할 수 있고, 단편 각각은 태깅되고 따라서 이의 본래의 분자 형태, 즉 이의 본래의 개별 핵산 분자에 기인할 수 있도록 하기 때문에, 짧은 판독 서열은 함께 구성되어 엑손 간의 관계에 대한 정보를 제공한다. 일부 구현예에서, 하나 이상의 엑손 모두 또는 일부를 함유하는

단편을 포획하기위해 사용되는 베이트는 하나 이상의 엑손 자체의 하나 이상의 부분에 상보적이다. 다른 구현예에서, 베이트는 엑손 영역의 3’ 또는5’ 측면 상의 엑손에 인접하거나 근처의 서열 또는 삽입 인트론의 하나 이상의 부분에 상보적이다(상기 베이트는 또한 본원에서 “인트론 베이트”로서 언급된다). 추가의 구현예에서, 엑손 모두 또는 일부를 함유하는 단편을 포획하기 위해 사용되는 베이트는 엑손 자체에 상보적인 베이트 및 인트론 베이트를 포함한다.

[0126] 서열 분석을 위해 포획된 표적 영역의 분자 형태를 보유하는 능력은 또한 게놈의불량하게 특징 분석된 영역에 걸친 서열 분석을 허용하는 이점을 제공한다. 평가된 바와 같이, 인간 게놈의 유의적 %(예를 들어, 하기 문헌에 따라 적어도 5-10% Altemose 등,*PLOS* Computational Biology, May 15, 2014, Vol.10, Issue 5)는 어셈블리되지 않고, 맵핑되지 않고 불량하게 특징 분석된 상태이다. 상기 참조 어셈블리는 일반적으로 이들 소실 영역에 다중-메가염기 이종염색질 갭이라는 주석을 달고, 상기 갭은 주로 말단 동원체형 염색체의 센트로미어 근처 및 짧은 판독상에서 주로 발견된다. 게놈의 상기 소실 분획은 일반적으로 사용된 서열 분석 기술을 사용하여 정확 특징 분석에 내성인 채로 남아있는 구조적 특징을 포함한다. 게놈의 확대된 영역에 대한 정보를 연결하는 능력을 제공함에 의해, 본원에 기재된 방법은 이들 불량하게 특징 분석된 영역에 대한 서열 분석을 허용하는 방식을 제공한다.

[0127] 일부 예에서, 게놈 DNA를 단편화하고, 증폭시키고, 분할시키고 다르게는 프로세싱하는 방법을 포함하는 샘플 제조 방법은 게놈의 특정 영역의 편향 또는 보다 낮은 커버를 유도할 수 있다. 상기 편향 또는 보다 낮아진 커버는 본원에 기재된 방법 및 시스템에서 게놈의 표적화 영역을 포획하기 위해 사용되는 베이트의 농도를 변화시킴에 보상될 수 있다. 예를 들어, 일부 상황에서, 게놈의 특정 영역이 단편 라이브러리가 프로세싱된 후 낮은 커버를 갖는 것으로 공지되어 있고, 예를 들어, 상기 영역은 무엇보다 게놈의 특정 영역에 대한 편향을 유도하는 GC 함량 또는 다른 구조적 변화를 함유하는 영역이다. 상기 상황에서, 베이트 라이브러리는 낮은 커버의 영역으로 지시된 베이트의 농도를 증가시키기 위해 변형될 수 있고, 다른 말로, 사용되는 베이트 집단은 “스파이킹”되어 상기 낮은 커버 영역에서 게놈의 표적화 영역을 함유하는 충분한 수의 단편이 서열 분석될 최종 단편 집단에서 수득되도록 보장할 수 있다. 베이트의 상기 스파이킹은 일부 구현예에서 커스텀 라이브러리의 디자인을 통해 수행될 수 있다. 추가의 구현예에서, 베이트의 스파이킹은 시판되는 전체 엑솜 키트에서 수행될 수 있어 보다 낮은 커버 영역 쪽으로 지시된 베이트의 커스텀 라이브러리는 재고품 엑솜 포획 키트에 부가된다.

[0128] 본원에 기재된 방법 및 시스템의 이점은 포획된 표적화 영역이, 표적화 영역을 포획하고 서열 분석을 수행하는 단계 후에도 이들 표적화 영역의 본래의 분자 형태가 보유되는 방식으로 포획 전에 프로세싱된다. 본원에서 추가로 상세하게 논의된 바와 같이, 특정 표적화 영역이 이들의 본래의 분자 형태(이들이 유래된 본래의 염색체 또는 염색체 영역 및/또는 완전한 게놈 내 서로 관련하여 특정 표적화 영역의 위치를 포함할 수 있다)에 기인하는 능력은 달리 불량하게 맵핑되거나 통상의 서열 분석 기술을 사용한 불량한 커버를 갖는 게놈의 영역으로부터 서열 정보를 수득하는 방식을 제공한다.

[0129] 예를 들어, 일부 유전자는 일반적으로 가용한 서열 분석 기술을 사용하여, 특히, 긴-판독 기술과 비교하여 보다 우수한 정확도를 갖는 짧은 판독 기술을 사용하여 포괄하기에는 너무 긴 인트론을 갖는다. 그러나, 본원에 기재된 방법 및 시스템에서, 표적화 영역의 분자 형태는 일반적으로 도 1에서 설명되고 본원에서 추가로 상세히 기재된 태깅 과정을 통해 보유된다. 이와 같이, 연결은 게놈의 확대된 영역을 거쳐 만들어질 수 있다. 예를 들어, 도 2b에 도식적으로 설명된 바와 같이, 핵산 분자 207은 긴 인트론 영역에 의해 차단된 엑손(음영 막대)을 함유한다. 일반적으로 사용되는 서열 분석 기술은 2개의 엑손 간의 관계에 대한 정보를 제공하기 위해 인트론에 걸친 거리를 포괄할 수 없다. 본원에 기재된 방법에서, 개별 핵산 분자 207은 이 자체의 별개의 파티션 209로 분포되고 이어서 단편화되어 상이한 단편은 엑손 및 인트론의 상이한 부분을 함유한다. 상기 단편 각각은 단편으로부터 수득된 임의의 서열 정보가 이어서 이것이 생성되는 별개의 파티션에 기인될 수 있도록 태깅되기 때문에, 각각의 단편은 또한 이것이 유래된 개별 핵산 분자 207에 기인할 수 있다. 일반적으로 그리고 본원에서 추가로 상세히 기재된 바와 같이, 단편화 및 태깅 후, 상이한 파티션으로부터의 단편은 함께 조합된다. 이어서 표적화 포획 방법은 목적하는 표적화 영역을 함유하는 단편과 함께 , 추가의 서열 분석과 같은 추가의 분석을 받는 단편 집단을 집적시키기 위해 사용될 수 있다. 도 2b에 설명된 예에서, 사용된 베이트는 엑손 중 한 부분을 함유하는 것들만을 포획하기 위해 단편 집단을 집적시키지만 엑손(209 및 210)의 외부 영역은 포획되지 않는다. 따라서, 서열 분석 받은 단편의 최종 집단은 목적하는 엑손을 함유하는 단편에 대해 집적된다. 이어서 짧은 판독, 높은 정확도의 서열 분석 기술을 사용하여 집적된 단편 집단의 서열을 동정할 수 있고, 단편 각각은 태깅되고 따라서 이의 본래의 분자 형태, 즉, 이의 본래의 개별 핵산 분자에 기인할 수 있기 때문에, 짧은 판독 서열은 함께 구성되어 2개의 엑손에 대한 연결된 서열 정보를 제공하기 위해 삽입 인트론 길이를 포괄할 수 있다

(이것은 일부 예에서 길이가 1, 2, 5, 10개 이상의 킬로베이스의 정도일 수 있다).

[0130] 상기된 바와 같이, 본원에 기재된 방법 및 시스템은 보다 긴 핵산의 짧은 서열 판독에 대한 개별 분자 형태를 제공한다. 본원에 사용된 바와 같은 개별 분자 형태는 예를 들어, 서열 판독 자체 내에 포함되지 않는 인접하거나 근접한 서열과 관련하여, 특정 서열 판독에 넘어서는 서열 형태를 언급하고 이와 같이 이들이 짧은 서열 판독, 예를 들어, 쌍을 이룬 판독에 대해 약 150개 염기 또는 약 300개 염기의 판독에 전체적으로 또는 부분적으로 포함되지 않도록 할 것이다. 특히 바람직한 양상에서, 방법 및 시스템은 짧은 서열 판독에 대한 긴 범위 서열형태를 제공한다. 상기 긴 범위 형태는 1 kb 보다 긴, 5 kb 보다 긴, 10 kb 보다 긴, 15 kb 보다 긴, 20 kb 보다 긴, 30 kb 보다 긴, 40 kb 보다 긴, 50 kb 보다 긴, 60 kb 보다 긴, 70 kb 보다 긴, 80 kb 보다 긴, 90 kb 보다 긴 또는 심지어 100 kb 보다 긴 서로의 거리 이내에 있는 서열 판독에 대한 소정의 서열 판독의 관계 또는 연결을 포함한다. 보다 긴 범위의 개별 분자 형태를 제공함에 의해, 본 발명의 방법 및 시스템은 또한 보다 긴 추론된 분자 형태를 제공한다. 본원에 기재된 바와 같은 서열 형태는 예를 들어, 연결된 분자의 보다 긴 개별 분자 또는 콘티그에 대해 짧은 서열을 맵핑하는 것으로부터 보다 낮은 해상 형태, 및 예를 들어, 개별 분자의 연속 결정된 서열을 갖는 보다 긴 개별 분자의 대형 부분의 긴 범위 서열 분석으로부터 보다 높은 해상 서열 형태를 포함할 수 있고, 여기서, 상기 결정된 서열은 1 kb 보다 길거나, 5 kb 보다 길거나, 10 kb 보다 길거나, 15 kb 보다 길거나, 20 kb 보다 길거나, 30 kb 보다 길거나, 40 kb 보다 길거나, 50 kb 보다 길거나, 60 kb 보다 길거나, 70 kb 보다 길거나, 80 kb 보다 길거나, 90 kb 보다 길거나 심지어 100 kb 보다 길다. 서열 형태와 관련하여, 보다 긴 핵산, 예를 들어, 연결된 핵산 분자 또는 콘티그의 개별적 긴 핵산 분자 또는 수집물 둘다로의 짧은 서열의 기인은 이들 보다 긴 핵산을 통한 짧은 서열로부터 어셈블리된 서열을 제공하는 것 뿐만 아니라 높은 수준의 서열 형태를 제공하기 위해 보다 긴 핵산 스트레치에 대한 짧은 서열의 맵핑 둘다를 포함할 수 있다.

[0131] *IV. 샘플*

[0132] 평가된 바와 같이, 본원에 논의된 방법 및 시스템은 임의의 유형의 게놈 물질로부터 표적화 서열 정보를 수득하기 위해 사용될 수 있다. 상기 게놈 물질은 환자로부터 채취한 샘플로부터 수득될 수 있다. 본원에서 논의된 방법 및 시스템에서 사용하는 게놈 물질의 예시적 샘플 및 게놈 물질의 유형은 제한 없이 폴리뉴클레오타이드, 핵산, 올리고뉴클레오타이드, 순환 세포 부재 핵산, 순환 종양 세포 (CTC), 핵산 단편, 뉴클레오타이드, DNA, RNA, 펩타이드 폴리뉴클레오타이드, 상보적 DNA (cDNA), 이중 나선 DNA (dsDNA), 단일 가닥 DNA (ssDNA), 플라스미드 DNA, 코스미드 DNA, 염색체 DNA, 게놈 DNA (gDNA), 바이러스 DNA, 세균 DNA, mtDNA (미토콘드리아 DNA), 리보솜 RNA, 세포 부재 DNA, 세포 부재 태아 DNA (cffDNA), mRNA, rRNA, tRNA, nRNA, siRNA, snRNA, snoRNA, scaRNA, 마이크로RNA, dsRNA, 바이러스 RNA 등을 포함한다. 요약하면, 사용되는 샘플은 특정 프로세싱 필요에 따라 다양할 수 있다.

[0133] 핵산을 포함하는 임의의 물질은 샘플의 공급원일 수 있다. 상기 물질은 유체, 예를 들어, 생물학적 유체일 수 있다. 유체 물질은 혈액, 제대혈, 침샘, 뇨, 땀, 혈청, 정액, 질 유체, 위액 및 소화액, 척수액, 태반액, 와동 유체, 안구 유체, 혈청, 모유, 림프 유체 또는 이의 조합물을 포함할 수 있지만 이에 제한되지 않는다. 물질은 고체, 예를 들어, 생물학적 조직일 수 있다. 물질은 정상의 건강한 조직, 환부 조직 또는 건강한 조직과 환부 조직의 혼합물을 포함할 수 있다. 일부 경우에, 상기 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, 척삭종, 혈관육종, 내피육종, 림프관육종, 림프관 내피육종, 활막종, 중피종, 유잉 종양, 평활근육종, 횡문근육종, 위장계 암종, 결장 암종, 췌장암, 유방암, 비뇨생식기 암종, 난소암, 전립선암, 편평 세포 암종, 기저 세포 암종, 선암종, 땀샘 암종, 피부기름샘 암종, 유두상 암종, 유두상 선암종, 낭선암종, 수질 암종, 기관지 암종, 신장 세포 암종, 간암, 담관 암종, 융모암, 정상피종, 배 암종, 빌름스 종양(Wilms' tumor), 자궁암, 내분비계 암종, 고환 종양, 폐 암종, 소 세포 폐 암종, 비-소세포 폐 암종, 유방 암종, 상피 암종, 신경교종, 성상세포종, 수모세포종, 두개인두종, 상의세포종, 송과체부종양, 혈관아세포종, 청신경종, 핍지교종, 뇌수막종, 흑색종, 신경아세포종, 망막아세포종, 또는 이의 조합. 물질은 다양한 유형의 기관과 관련될 수 있다. 기관의 비제한적인 예는 뇌, 간, 폐, 신장, 전립선, 난소, 비장, 림프절(편도를 포함하는), 갑상선, 췌장, 심장, 골격근, 장, 후두, 식도, 위 또는 이의 조합을 포함할 수 있다. 일부 경우에, 물질은 다음을 포함하지만 이에 제한되지 않는 다양한 세포를 포함할 수 있다: 진핵 세포, 원핵 세포, 진균 세포, 심장 세포, 폐 세포, 신장 세포, 간 세포, 췌장세포, 생식 세포, 줄기 세포, 유도 만능 줄기 세포, 위장 세포, 혈액 세포, 암 세포, 세균 세포, 인간 마이크로바이옴 샘플 등으로부터 단리된 세균 세포. 일부 경우에, 물질은 세포의 내용물, 예를 들어, 단일 세포의 내용물 또는 다중세포의 내용물을 포함할 수 있다. 개별 세포를 분석하기 위한

방법 및 시스템은 예를 들어, 다음 문헌에서 제공된다: 미국특허 출원 번호 제14/752,641호, 2015년 6월 26일자로 출원됨, 이의 완전한 기재내용은 이의 전문, 특히 개별 세포로부터 핵산을 분석하는 것과 관련된 모든 교시로 본원에 참조로 인용됨.

[0134] 샘플은 다양한 대상체로부터 수득될 수 있다. 대상체는 생존 대상체 또는 죽은 대상체일 수 있다. 대상체의 예는 인간, 포유동물, 비-인간 포유동물, 설치류, 양서류, 파충류, 개, 고양이, 소, 말, 염소, 양, 닭, 아빈(avine), 마우스, 토끼, 곤충, 민달팽이, 미생물, 세균, 기생충 또는 어류를 포함할 수 있지만 이에 제한되지 않는다. 일부 경우에, 대상체는 질환 또는 장애를 갖거나, 이를 발병할 위험에 처한 환자일 수 있다. 일부 경우에, 대상체는 임신 여성일 수 있다. 일부 경우에, 대상체는 정상의 건강한 임신 여성일 수 있다. 일부 경우에, 대상체는 특정 태생 결함을 갖는 아기를 가질 위험에 처한 임신 여성일 수 있다.

[0135] 샘플은 당업계에 공지된 임의의 수단에 의해 대상체로부터 수득될 수 있다. 예를 들어, 샘플은 순환계로의 접근(예를 들어, 시린지 또는 다른 장치를 통한 정맥내 또는 동맥내), 분비된 생물학적 샘플(예를 들어, 침샘, 객담, 뇨, 배변 등)의 수집, 생물학적 샘플(예를 들어, 작동내 샘플, 수술 후 샘플 등)의 수술적 획득(예를 들어, 생검), 면봉 채취(예를 들어, 협측 면봉, 입 인두 면봉) 또는 피펫팅을 통해 대상체로부터 수득될 수 있다.

[0136] 본 발명의 바람직한 양태가 본원에 나타내고 기재되었지만, 상기 양태가 단지 예를 통해 제공된다는 것은 통상의 기술자에게 자명할 것이다. 수많은 변형, 변화 및 치환은 본 발명을 벗어나는 것 없이 당업자에게 자명할 것이다. 본원에 기재된 본 발명의 양태에 대한 다양한 대안은 본 발명을 수행하는데 사용될 수 있는 것으로 이해되어야만 한다. 하기의 특허청구범위는 본 발명의 범위를 한정하고 이들 특허청구범위 내 방법 및 구조 및 이의 등가물이 보호되는 것으로 의도된다.

[0137] **실시예**

[0138] *실시예 1: 전체 엑솜 포획 및 서열 분석: NA12878*

[0139] NA12878 인간 세포주로부터의 게놈 DNA는 대략 10 kb 길이 이상인 단편을 회수하기 위해 블루 피핀(Blue Pippin) DNA 크기 측정 시스템을 사용하여 단편의 크기 기반 분리에 적용하였다. 크기 선택된 샘플 핵산은 이어서 미세유동 분할 시스템을 사용한 불소화 오일 연속 상 내 수성 소적 중에 바코드 비드와 함께 동시 분할시키고(문헌참조: 예를 들어, 미국특허 출원 번호 제14/682,952호, 2015년 4월 9일자로 출원됨, 및 모든 목적을 위해 이의 전문이 본원에 참조로 인용됨), 여기서, 상기 수성 소적은 또한 비드로부터 바코드 올리고뉴클레오타이드를 방출시키기 위한 DTT 뿐만 아니라 소적 내 dNTP, 열안정성 DNA 폴리머라제 및 증폭을 수행하기 위한 다른 시약을 포함한다. 이것은 총 투입 DNA 1ng 및 총 투입 DNA 2ng 둘다에 대해 반복하였다. 상기 바코드 비드는 700,000개의 상이한 바코드 서열의 바코드 다양성을 나타내는 스톡 라이브러리 서브세트로 수득되었다. 올리고뉴클레오타이드를 함유하는 바코드는 추가의 서열 성분을 포함하고 하기 일반 구조를 가졌다:

[0140] 비드-P5-BC-R1-N량체

[0141] P5 및 R1이 각각 Illumina 부착 및 판독1 프라이머 서열을 언급하는 경우, BC는올리고뉴클레오타이드의 바코드 부분을 지칭하고 N량체는 주형 핵산을 프라이밍하기 위해 사용되는 무작위 10개 염기 N-량체를 지칭한다. 문헌참조: 예를 들어, 미국특허 출원 번호 제14/316,383호, 2014년 6월 26일자로 출원됨, 이의 전반적인 내용은 모든 목적을 위해 이의 전문이 본원에 참조로 인용된다.

[0142] 비드 용해 후, 소적은 각각의 소적 내 샘플 핵산의 주형에 대한 바코드 올리고스의 프라이머 연장을 허용하기 위해 열순환시켰다. 이것은 상기 제시된 다른 포함된 서열 뿐만 아니라 본래의 파티션을 대표하는 바코드 서열을 포함하는 샘플 핵산의 증폭된 카피 단편을 유도하였다.

[0143] 카피 단편의 바코드 표지 후, 증폭된 카피 단편을 포함하는 소적의 에멀젼을 부수고 추가의 서열 분석기 요구되는 성분들, 예를 들어, 판독 2 프라이머 서열 및 P7 부착 서열은 추가의 증폭 단계를 통해 카파 단편에 부가하였고, 상기 단계는 이들 서열을 카피 단편의 다른 말단에 부착시켰다. 이어서 바코드화된 DNA는 Agilent SureSelect 엑솜 포획 키트를 사용하여 하이브리드 포획에 적용하였다.

[0144] 하기 표는 NA 12878 게놈에 대한 표적화 통계를 제공한다:

샘플	중앙 삽입물 크기	표적 상의 단편 %	표적 상의 염기 %
버전 1.A	258	81%	51%
버전 1.B	224	81%	55%
버전 1.C	165	81%	63%

[0145]

[0146] 상기된 3개의 상이한 버전은 제2 어댑터 부착 단계 전에 바코드화된 단편에 대한 상이한 전단 길이를 나타낸다.

[0147] *실시예 2: 전체 엑솜 포획 및 서열 분석: NA19701 및 NA19661*

[0148] NA19701 및 NA19661 세포주로부터 기원하는 게놈 DNA는 실시예 1에서 상기된방법에 따라 제조하였다. 상기 2개의 세포주로부터 단계화 데이터를 포함하는 데이터는 하기 표에 제공된다:

	NA19661	NA19701
N50_phase_block	29,535	83,953
N90_phase_block	8,595	25,684
mean_phase_block	5,968	21,128
median_phase_block	0	76.5
longest_phase_block	209,323	504,140
fract_genes_phased	0.719	0.841
fract_genes_completely_phased	0.679	0.778
fract_snps_phased	0.869	0.832
fract_snps_barcode_both_alleles	0.328	0.351
prob_snp_correct_in_gene	0.906	0.927
prob_snp_phased_in_gene	0.807	0.889
snp_short_switch_error	0.013	0.013
snp_long_switch_error	0.012	0.013

[0149]

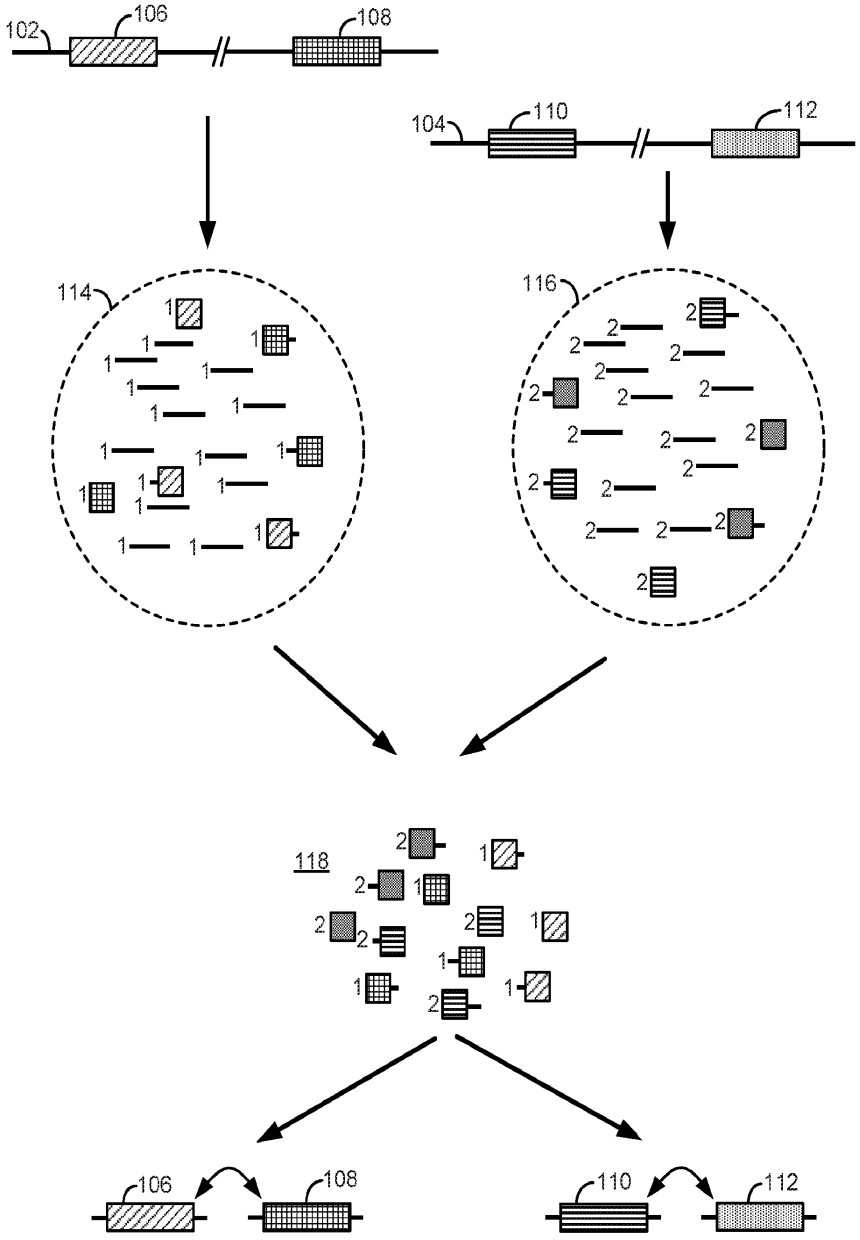
[0150] 표 3. 단계적 메트릭스. 도 1에서 보여진 바와 같이, NA19701의 단편 길이는 NA19661보다 훨씬 더 길고, 이로써 단계적 성능이 훨씬 더 좋게 된다.

[0151] 본 명세서는 현재-기재된 기술의 예시적 양상에서 방법, 시스템 및/또는 구조 및 이의 용도의 완전한 기재를 제공한다. 상기 기술의 다양한 양상이 특정 정도의 특이성으로 또는 하나 이상의 개별적인 양상과 관련하여 상기되었지만, 통상의 기술자는 이의 기술 취지 또는 범위를 벗어나는 것 없이 기재된 양상에 다양한 변형을 만들 수 있다. 많은 양상은 현재 기재된 기술의 취지 및 범위로부터 벗어나는 것 없이 만들어질 수 있기 때문에, 적절한 범위는 이후 첨부된 특허청구범위에 있다. 따라서 다른 양상이 고려된다. 추가로, 달리 명백히 청구되지 않는 경우 임의의 작동이 임의의 순서로 수행될 수 있거나 특정 순서는 고유하게 청구항 기재 내용에 의해 요구되는 것으로 이해되어야 한다. 상기에 포함되고 첨부된 도면에 나타낸 모든 매터는 단지 특정 양상을 설명하는 것으로 이해되고 나타낸 구현예에 제한되지 않는 것으로 의도된다. 문단에서 달리 명백하지 않거나 명백히 진술되지 않는 경우, 본원에 제공된 임의의 농도 값은 일반적으로 혼합물의 특정 성분의 첨가 즉시 또는 첨가 후 발생하는 임의의 전환과 관련하여 혼합 값 또는 % 측면에서 주어진다. 본원에서 이미 명백하게 인용되지 않는 정도로, 본원에 언급된 모든 공개된 참조문헌 및 특허문헌은 모든 목적을 위해 이의 전문이 본원에 참조로 인용된

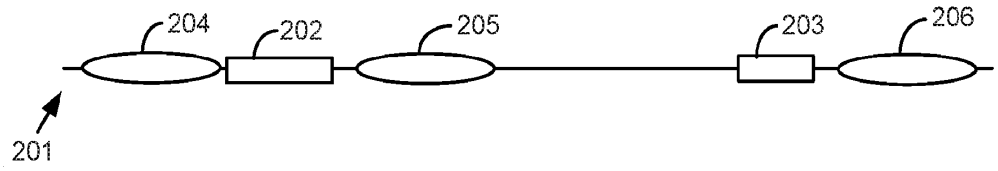
다. 세부적인 변화 또는 구조는 하기 청구항에 정의된 바와 같은 본 기술의 기본요소로부터 벗어나는 것 없이 수행될 수 있다.

도면

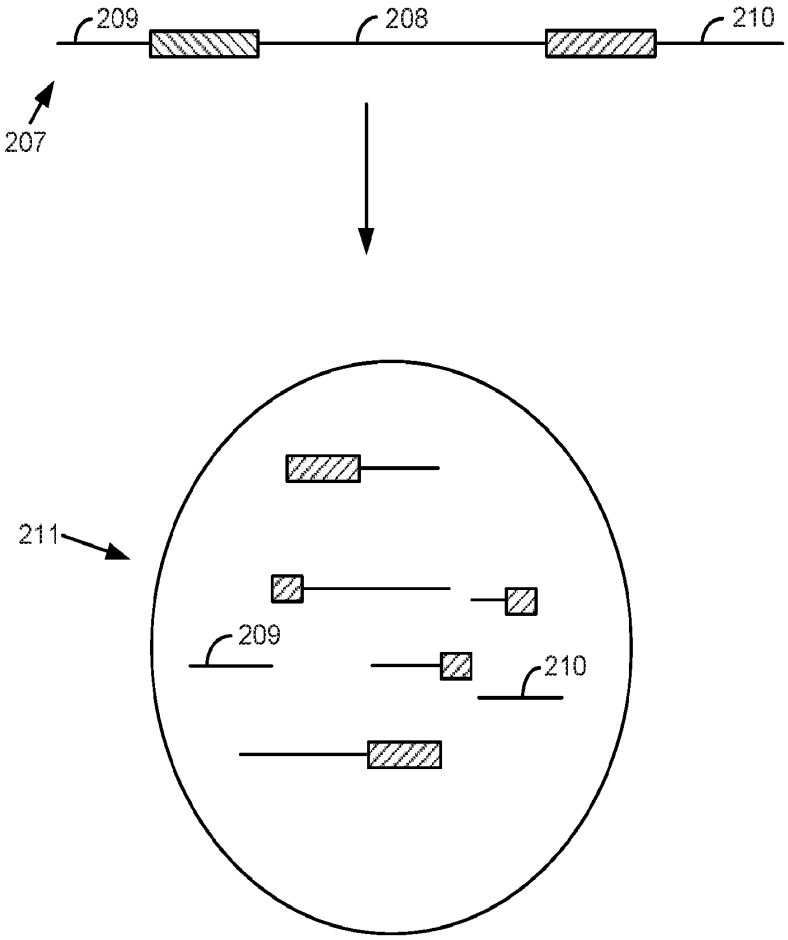
도면1



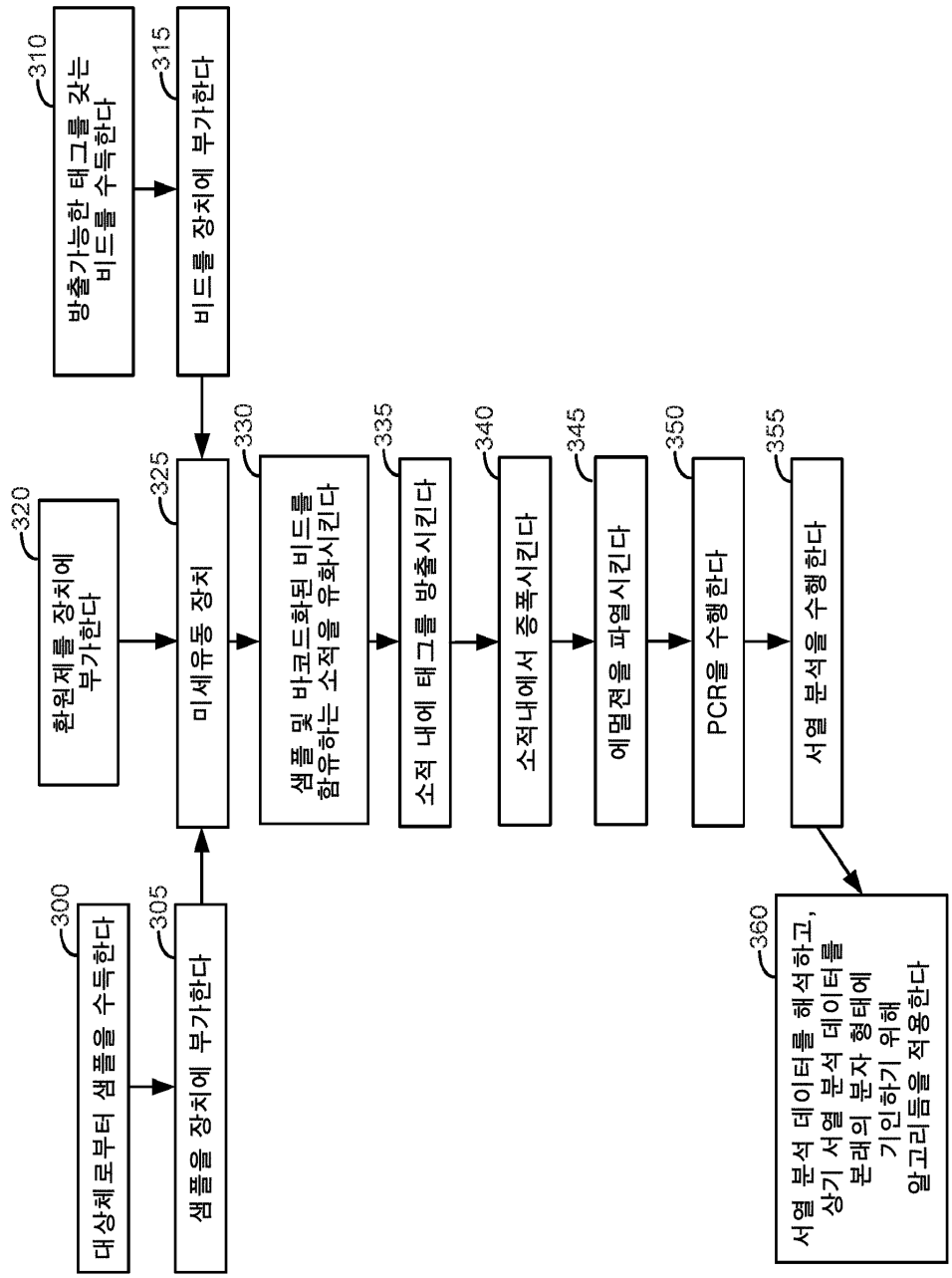
도면2a



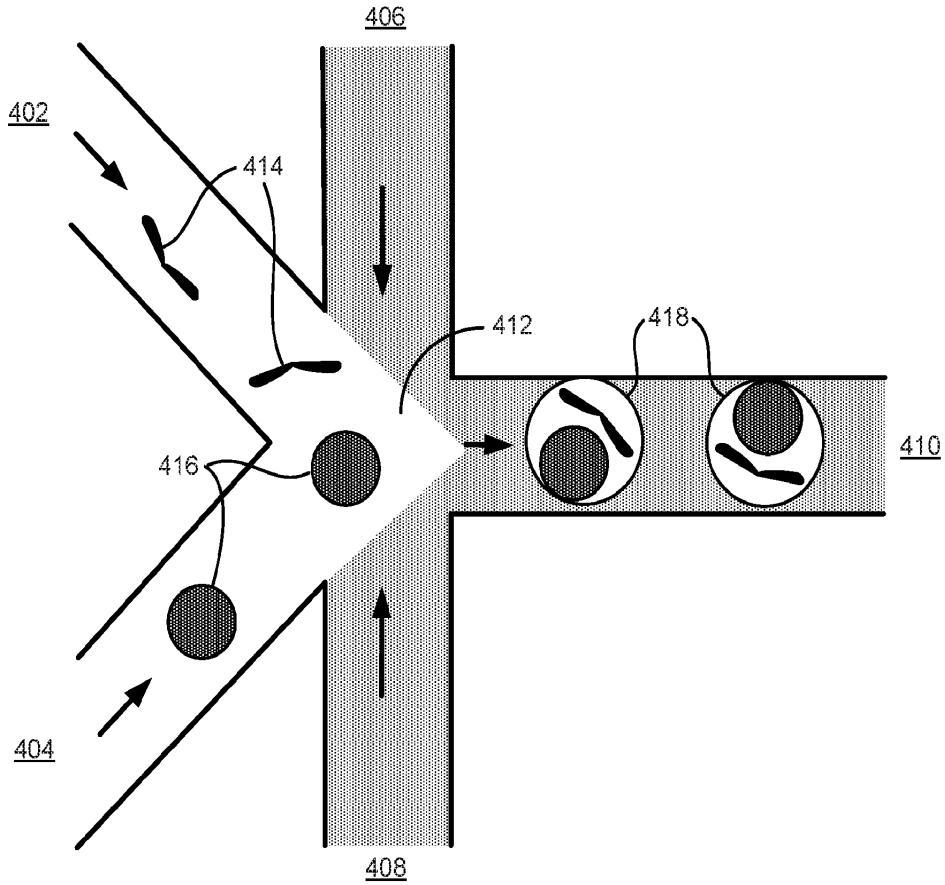
도면2b



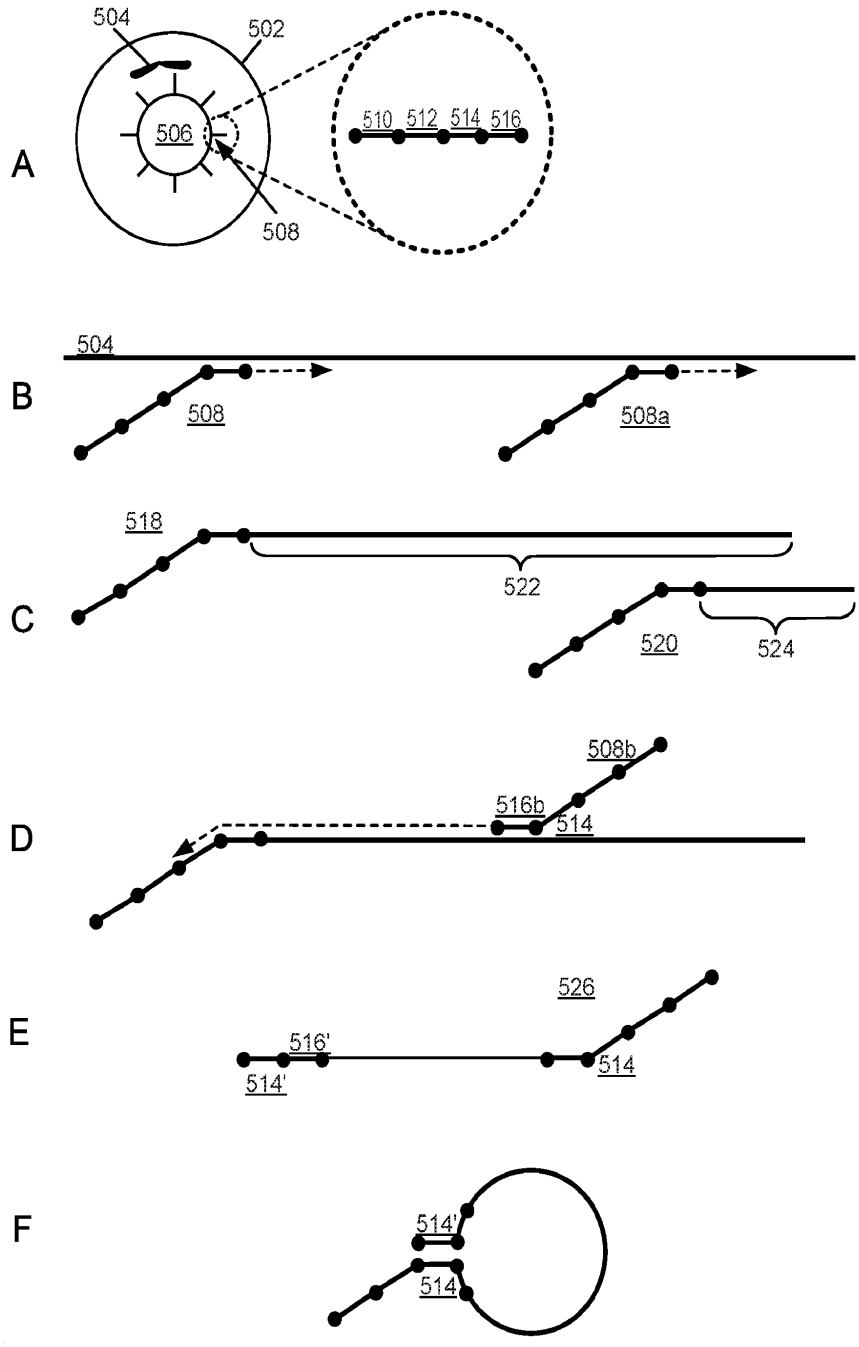
도면3


300
대상체로부터 샘플을 수득한다
305
샘플을 장치에 부가한다
310
방출가능한 태그를 갖는 비드를 수득한다
315
비드를 장치에 부가한다
320
환원제를 장치에 부가한다
325
미세유동 장치
330
샘플 및 바코드화된 비드를 함유하는 소적을 유화시킨다
335
소적 내에 태그를 방출시킨다
340
소적내에서 증폭시킨다
345
에멀젼을 파열시킨다
350
PCR을 수행한다
355
서열 분석을 수행한다
360
서열 분석 데이터를 해석하고, 상기 서열 분석 데이터를 본래의 분자 형태에 기인하기 위해 알고리듬을 적용한다

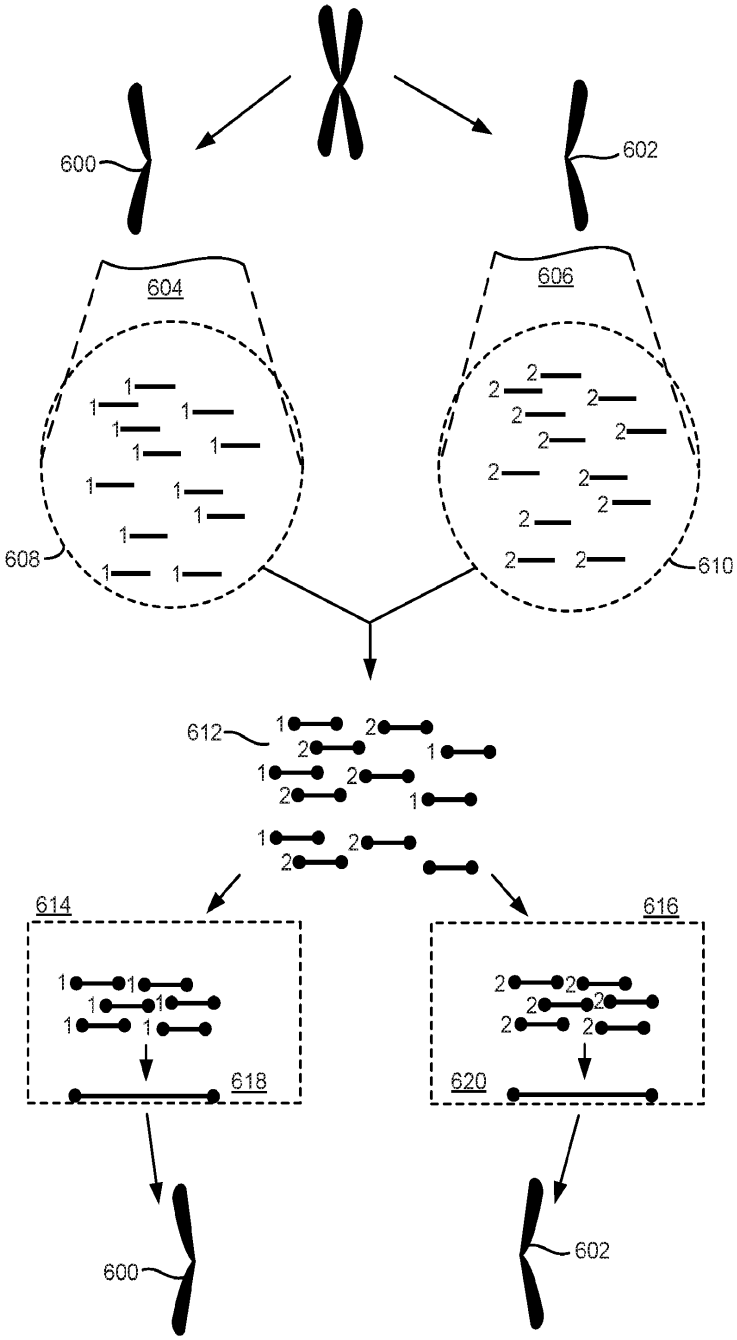
도면4



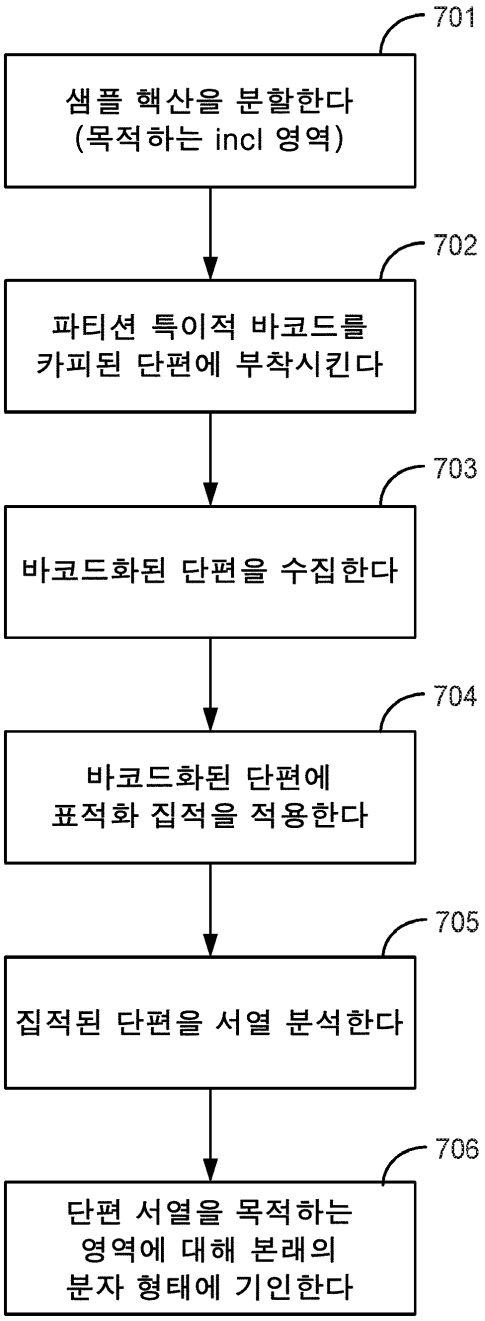
도면5


A
504
502
506
510 512 514 516
508
B
504
508
508a
C
518
522
520
524
D
508b
516b
514
E
526
516'
514
514'
F
514'
514

도면6



도면7


701
샘플 핵산을 분할한다
(목적하는 incl 영역)
702
파티션 특이적 바코드를
카피된 단편에 부착시킨다
703
바코드화된 단편을 수집한다
704
바코드화된 단편에
표적화 집적을 적용한다
705
집적된 단편을 서열 분석한다
706
단편 서열을 목적하는
영역에 대해 본래의
분자 형태에 기인한다

도면8