(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0112843
(43) 공개일자 2019년10월07일

(51) 국제특허분류(Int. Cl.)
C12Q 1/6869 (2018.01) *C12Q 1/6806* (2018.01)
C12Q 1/6883 (2018.01) *C12Q 1/6886* (2018.01)
G16B 20/00 (2019.01)
(52) CPC특허분류
C12Q 1/6869 (2018.05)
C12Q 1/6806 (2018.05)
(21) 출원번호 10-2019-7028255(분할)
(22) 출원일자(국제) 2013년09월04일
심사청구일자 없음
(62) 원출원 특허 10-2015-7008319
원출원일자(국제) 2013년09월04일
심사청구일자 2018년09월03일
(85) 번역문제출일자 2019년09월26일
(86) 국제출원번호 PCT/US2013/058061
(87) 국제공개번호 WO 2014/039556
국제공개일자 2014년03월13일
(30) 우선권주장
61/696,734 2012년09월04일 미국(US)
(뒷면에 계속)

(71) 출원인
가던트 헬쓰, 인크.
미국 94063 캘리포니아주 레드우드 시티 페노브스코트 드라이브 505
(72) 발명자
타라사즈, 아미르알리
미국 94025 캘리포니아주 멘로 파크 카미노 에이로스 세로스 2181
엘토키, 헬미
미국 94027 캘리포니아주 애서튼 배리 래인 2
(74) 대리인
양영준, 김영

전체 청구항 수 : 총 1 항

(54) 발명의 명칭 **희귀 돌연변이 및 카피수 변이를 검출하기 위한 시스템 및 방법**

(57) 요 약

본 개시내용은 세포 유리 폴리뉴클레오티드 내의 희귀 돌연변이 및 카피수 변이의 검출을 위한 시스템 및 방법을 제공한다. 일반적으로, 시스템 및 방법은 샘플 제조, 또는 체액으로부터 세포 유리 폴리뉴클레오티드 서열의 추출 및 단리; 관련 기술 분야에 공지된 기술에 의한 세포 유리 폴리뉴클레오티드의 후속적인 서열분석; 및 참조물에 비교하여 희귀 돌연변이 및 카피수 변이를 검출하기 위한 생물 정보공학 도구의 적용을 포함한다. 시스템 및 방법은 또한 질환의 희귀 돌연변이, 카피수 변이 프로파일링 또는 일반적인 유전자 프로파일링의 검출을 도울 때 추가의 참조물로서 사용되는, 상이한 질환의 상이한 희귀 돌연변이 또는 카피수 변이 프로파일의 데이터베이스 또는 수집물을 함유할 수 있다.

(52) CPC특허분류
C12Q 1/6883 (2018.05)
C12Q 1/6886 (2018.05)
G16B 20/00 (2019.02)
C12Q 2537/143 (2013.01)
C12Q 2537/16 (2013.01)
C12Q 2600/112 (2013.01)
C12Q 2600/16 (2013.01)

(30) 우선권주장
61/704,400 2012년09월21일 미국(US)
61/793,997 2013년03월15일 미국(US)
61/845,987 2013년07월13일 미국(US)

명 세 서

청구범위

청구항 1

컴퓨터 프로세서에 의한 실행시에,

게놈 내의 미리 규정된 영역을 선택하고;

서열 판독체에 접근하여 미리 규정된 영역 내의 서열 판독체의 수를 계수하고;

서열 판독체의 수를 미리 규정된 영역에 걸쳐 정규화하고;

미리 규정된 영역 내의 카피수 변이의 퍼센트를 결정하는 것

을 포함하는 방법을 이행하기 위한, 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체의 용도.

발명의 설명

기 술 분 야

[0001] **관련 출원에 대한 상호 참조**

[0002] 본원은 2012년 9월 4일 출원된 미국 특허 가출원 61/696,734, 2012년 9월 21일 출원된 미국 특허 가출원 61/704,400, 2013년 3월 15일 출원된 미국 특허 가출원 61/793,997 및 2013년 7월 13일 출원된 미국 특허 가출원 61/845,987을 우선권을 주장하며, 상기 가출원 각각은 그 전문이 모든 목적을 위해 본원에 참조로 포함된다.

배 경 기 술

[0003] 폴리뉴클레오티드의 검출 및 정량은 분자 생물학 및 의료 적용, 예컨대 진단에 중요하다. 유전자 시험은 특히 많은 진단 방법에 유용하다. 예를 들어, 희귀 유전자 변경 (예를 들어, 서열 변이체) 또는 후성적 마커의 변화에 의해 야기되는 장애, 예컨대 암 및 부분적인 또는 완전한 이수성은 DNA 서열 정보를 사용하여 검출되거나 보다 정확하게 특성이 규정될 수 있다.

[0004] 유전 질환, 예컨대 암의 검출 및 모니터링은 종종 질환의 성공적인 치료 또는 관리에 유용하고 필요하다. 한 방법은 세포 유리 핵산으로부터 유래된 샘플, 즉 상이한 종류의 체액에서 발견될 수 있는 폴리뉴클레오티드의 집단의 모니터링을 포함할 수 있다. 일부 경우에, 질환은 유전자 이상, 예컨대 하나 이상의 핵산 서열의 카피수 변이 및/또는 서열 변이의 변화, 또는 다른 특정 희귀 유전자 변경의 발생을 기초로 하여 특성이 규정되거나 검출될 수 있다. 세포 유리 DNA ("cfDNA")는 관련 기술 분야에 수십 년 동안 알려져 왔고, 특정 질환과 연관된 특정 유전자 이상을 포함할 수 있다. 서열분석 및 핵산 조작 기술의 개선과 함께, 질환을 검출하고 모니터링하기 위해 세포 유리 DNA를 사용하기 위한 개선된 방법 및 시스템에 대한 필요성이 관련 기술 분야에 존재한다.

발명의 내용

[0005] 본 개시내용은 a) 대상체로부터의 신체 샘플로부터의 세포외 폴리뉴클레오티드를 서열분석하고, 여기서 각각의 세포외 폴리뉴클레오티드는 특유한 바코드에 임의로 부착되고; b) 설정된 역치를 충족하지 않는 판독체를 여과 제거하고; c) 단계 (a)로부터 얻은 서열 판독체를 참조 서열에 맵핑하고; d) 참조 서열의 2개 이상의 미리 규정된 영역 내의 맵핑된 판독체를 정량/계수하고; e) (i) 미리 규정된 영역 내의 판독체의 수를 서로 및/또는 미리 규정된 영역 내의 특유한 바코드의 수를 서로 정규화하고; (ii) 단계 (i)에서 얻은 정규화된 수를 대조 샘플로부터 얻은 정규화된 수와 비교함으로써 하나 이상의 미리 규정된 영역 내의 카피수 변이를 결정하는 것을 포함하는, 카피수 변이를 검출하는 방법을 제공한다.

[0006] 본 개시내용은 또한

[0007] a) 대상체로부터의 신체 샘플로부터의 세포외 폴리뉴클레오티드를 서열분석하고, 여기서 각각의 세포외 폴리뉴클레오티드는 다수의 서열분석 판독체를 생성하고; b) 대상체로부터의 신체 샘플로부터의 세포외 폴리뉴클레오티드를 서열분석하고, 여기서 각각의 세포외 폴리뉴클레오티드는 다수의 서열분석 판독체를 생성하고; 대상체로부터의 신체 샘플로부터의 세포외 폴리뉴클레오티드를 서열분석하고, 여기서 각각의 세포외 폴리뉴클레오티드는

다수의 서열분석 판독체를 생성하고; c) 설정된 역치를 충족하지 않는 판독체를 여과 제거하고; d) 서열분석으로부터 유래된 서열 판독체를 참조 서열 상에 맵핑하고; e) 각각의 맵핑가능한 염기 위치에서 참조 서열의 변이체와 정렬되는 맵핑된 서열 판독체의 하위세트를 확인하고; f) 각각의 맵핑가능한 염기 위치에 대해, (a) 참조 서열에 비해 변이체를 포함하는 맵핑된 서열 판독체의 수 대 (b) 각각의 맵핑가능한 염기 위치에 대한 총 서열 판독체의 수의 비를 계산하고; g) 각각의 맵핑가능한 염기 위치에 대한 변이의 비 또는 빈도를 정규화하고 잠재적인 희귀 변이체(들) 또는 돌연변이(들)를 결정하고; h) 잠재적인 희귀 변이체(들) 또는 돌연변이(들)를 갖는 각각의 영역에 대해 생성된 수를 참조 샘플로부터 유사하게 유래된 수와 비교하는 것을 포함하는, 대상체로부터 얻은 세포 유리 또는 실질적인 세포 유리 샘플에서 희귀 돌연변이를 검출하는 방법을 제공한다.

[0008] 추가로, 본 개시내용은 또한 대상체에서 세포외 폴리뉴클레오티드의 유전자 프로파일을 생성하는 것을 포함하는, 대상체에서 비정상적인 상태의 비균질성을 특성화하는 방법을 제공하고, 여기서 유전자 프로파일은 카피수 변이 및/또는 다른 희귀 돌연변이 (예를 들어, 유전자 변경) 분석에 의해 생성된 다수의 데이터를 포함한다.

[0009] 일부 실시양태에서, 대상체에서 확인된 각각의 희귀 변이체의 출현율/농도는 동시에 보고되고 정량된다. 다른 실시양태에서, 대상체 내의 희귀 변이체의 출현율/농도에 관한 신뢰도 점수가 보고된다.

[0010] 일부 실시양태에서, 세포외 폴리뉴클레오티드는 DNA를 포함한다. 다른 실시양태에서, 세포외 폴리뉴클레오티드는 RNA를 포함한다. 폴리뉴클레오티드는 단편이거나 단리 후에 단편화될 수 있다. 추가로, 본 개시내용은 핵산 단리 및 추출을 순환시키는 방법을 제공한다.

[0011] 일부 실시양태에서, 세포외 폴리뉴클레오티드는 혈액, 혈장, 혈청, 소변, 타액, 점막 분비물, 객담, 대변 및 눈물로 이루어진 군으로부터 선택될 수 있는 신체 샘플로부터 단리된다.

[0012] 일부 실시양태에서, 본 개시내용의 방법은 또한 상기 신체 샘플에서 카피수 변이 또는 다른 희귀 유전자 변경 (예를 들어, 서열 변이체)을 갖는 서열의 퍼센트를 결정하는 단계를 포함한다.

[0013] 일부 실시양태에서, 상기 신체 샘플에서 카피수 변이를 갖는 서열의 퍼센트는 미리 결정된 역치 초과 또는 미만의 폴리뉴클레오티드의 양을 갖는 미리 규정된 영역의 백분율을 계산함으로써 결정된다.

[0014] 일부 실시양태에서, 체액은 돌연변이, 희귀 돌연변이, 단일 뉴클레오티드 변이체, 삽입-결실 (indel), 카피수 변이, 염기변환 (transversion), 전위 (translocation), 역위 (inversion), 결실, 이수성, 부분적 이수성, 배수성 (polyploidy), 염색체 불안정성, 염색체 구조 변경, 유전자 융합, 염색체 융합, 유전자 말단절단 (truncation), 유전자 증폭, 유전자 중복, 염색체 병변, DNA 병변, 핵산 화학적 변형의 비정상적인 변화, 후성적 패턴의 비정상적인 변화, 핵산 메틸화 감염의 비정상적인 변화 및 암으로 이루어진 군으로부터 선택될 수 있는 비정상적인 상태를 갖는 것으로 의심되는 대상체로부터 채취된다.

[0015] 일부 실시양태에서, 대상체는 단일 뉴클레오티드 변이체, 삽입-결실, 카피수 변이, 염기변환, 전위, 역위, 결실, 이수성, 부분적 이수성, 배수성, 염색체 불안정성, 염색체 구조 변경, 유전자 융합, 염색체 융합, 유전자 말단절단, 유전자 증폭, 유전자 중복, 염색체 병변, DNA 병변, 핵산 화학적 변형의 비정상적인 변화, 후성적 패턴의 비정상적인 변화, 핵산 메틸화 감염의 비정상적인 변화 및 암으로 이루어진 군으로부터 선택된 태아 비정상일 수 있는 임신한 여성일 수 있다.

[0016] 일부 실시양태에서, 방법은 서열분석 전에 하나 이상의 바코드를 세포외 폴리뉴클레오티드 또는 그의 단편에 부착시키는 것을 포함할 수 있고, 여기서 바코드는 특유한 것이다. 다른 실시양태에서, 서열분석 전에 세포외 폴리뉴클레오티드 또는 그의 단편에 부착된 바코드는 특유한 것이 아니다.

[0017] 일부 실시양태에서, 본 개시내용의 방법은 서열분석 전에 대상체의 게놈 또는 트랜스크립톰 (transcriptome)으로부터의 영역을 선택적으로 풍부화하는 것을 포함할 수 있다. 다른 실시양태에서, 본 개시내용의 방법은 서열분석 전에 대상체의 게놈 또는 트랜스크립톰으로부터의 영역을 선택적으로 풍부화하는 것을 포함한다. 다른 실시양태에서, 본 개시내용의 방법은 서열분석 전에 대상체의 게놈 또는 트랜스크립톰으로부터의 영역을 비-선택적으로 풍부화하는 것을 포함한다.

[0018] 또한, 본 개시내용의 방법은 임의의 증폭 또는 풍부화 단계 전에 하나 이상의 바코드를 세포외 폴리뉴클레오티드 또는 그의 단편에 부착시키는 것을 포함한다.

[0019] 일부 실시양태에서, 바코드는 무작위 서열 또는 선택 영역으로부터 서열분석된 분자의 다양성과 조합되어 특유한 분자의 확인을 가능하게 하고 적어도 3, 5, 10, 15, 20, 25, 30, 35, 40, 45 또는 50량체 염기쌍 길이인 올

리고뉴클레오티드의 고정 또는 준-무작위 세트를 추가로 포함할 수 있는 폴리뉴클레오티드이다.

[0020] 일부 실시양태에서, 세포외 폴리뉴클레오티드 또는 그의 단편은 증폭될 수 있다. 일부 실시양태에서, 증폭은 포괄적 (global) 증폭 또는 전체 게놈 증폭을 포함한다.

[0021] 일부 실시양태에서, 특유한 정체 (identity)의 서열 판독체는 서열 판독체의 개시 (출발) 및 종료 (정지) 영역에서의 서열 정보 및 서열 판독체의 길이를 기초로 하여 검출될 수 있다. 다른 실시양태에서, 특유한 정체의 서열 분자는 서열 판독체의 개시 (출발) 및 종료 (정지) 영역에서의 서열 정보, 서열 판독체의 길이 및 바코드의 부착을 기초로 하여 검출된다.

[0022] 일부 실시양태에서, 증폭은 선택적 증폭, 비-선택적 증폭, 억제 증폭 또는 차감 풍부화 (subtractive enrichment)를 포함한다.

[0023] 일부 실시양태에서, 본 개시내용의 방법은 판독체의 정량 또는 계수 전에 추가의 분석으로부터의 판독체의 하위세트를 제거하는 것을 포함한다.

[0024] 일부 실시양태에서, 방법은 역치, 예를 들어 90%, 99%, 99.9% 또는 99.99% 미만의 정확도 또는 품질 점수 및/또는 역치, 예를 들어 90%, 99%, 99.9% 또는 99.99% 미만의 맵핑 점수를 갖는 판독체를 여과 제거하는 것을 포함할 수 있다. 다른 실시양태에서, 본 개시내용의 방법은 설정된 역치보다 낮은 품질 점수를 갖는 판독체를 여과 제거하는 것을 포함한다.

[0025] 일부 실시양태에서, 미리 규정된 영역은 균일한 또는 실질적으로 균일한 크기, 약 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb 또는 100 kb이다. 일부 실시양태에서, 적어도 50, 100, 200, 500, 1000, 2000, 5000, 10,000, 20,000 또는 50,000개의 영역이 분석된다.

[0026] 일부 실시양태에서, 유전자 변이체, 희귀 돌연변이 또는 카피수 변이는 유전자 융합, 유전자 중복, 유전자 결실, 유전자 전위, 미소부수체 (microsatellite) 영역, 유전자 단편 또는 이들의 조합으로 이루어진 군으로부터 선택된 게놈의 영역에서 발생한다. 다른 실시양태에서, 유전자 변이체, 희귀 돌연변이, 또는 카피수 변이는 유전자, 종양유전자, 종양 억제 (suppressor) 유전자, 프로모터, 조절 서열 요소 또는 이들의 조합으로 이루어진 군으로부터 선택된 게놈의 영역에서 발생한다. 일부 실시양태에서, 변이체는 뉴클레오티드 변이체, 단일 염기 치환, 또는 작은 삽입-결실, 염기변환, 전위, 역위, 결실, 말단절단 또는 유전자 말단절단 (약 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 또는 20개 뉴클레오티드 길이)이다.

[0027] 일부 실시양태에서, 방법은 개별 판독체의 바코드 또는 특유한 특성을 사용하여 맵핑된 판독체의 양을 보정/정규화/조정하는 것을 포함한다.

[0028] 일부 실시양태에서, 판독체의 계수는 각각의 미리 규정된 영역 내의 특유한 바코드를 계수하고 그 수를 서열분석된 미리 규정된 영역의 적어도 하나의 하위세트에 걸쳐 정규화함으로써 수행된다. 일부 실시양태에서, 동일한 대상체로부터 연속적인 시간 간격에서 채취한 샘플이 분석되고 이전의 샘플 결과와 비교된다. 본 개시내용의 방법은 바코드-부착된 세포외 폴리뉴클레오티드를 증폭한 후 부분적인 카피수 변이 빈도의 결정, 이형접합성의 상실의 결정, 유전자 발현 분석, 후성적 분석 및 과메틸화 분석을 추가로 포함할 수 있다.

[0029] 일부 실시양태에서, 카피수 변이 및 희귀 돌연변이 분석은 10,000회 초과의 서열분석 반응을 수행하거나; 적어도 10,000개의 상이한 판독체를 동시에 서열분석하거나; 또는 게놈에 걸쳐 적어도 10,000개의 상이한 판독체에 대한 데이터 분석을 수행하는 것을 포함하는 다중 서열분석을 사용하여 대상체로부터 얻은 세포 유리 또는 실질적인 세포 유리 샘플에서 결정된다. 방법은 게놈에 걸쳐 적어도 10,000개의 상이한 판독체에 대한 데이터 분석을 수행하는 것을 포함하는 다중 서열분석을 포함할 수 있다. 방법은 특유하게 확인가능한 서열분석된 판독체를 계수하는 것을 추가로 포함할 수 있다.

[0030] 일부 실시양태에서, 본 개시내용의 방법은 하나 이상의 은닉 마르코프 (hidden markov), 동적 프로그래밍, 서포트 벡터 머신, 베이지안 네트워크 (Bayesian network), 격자 해독 (trellis decoding), 비터비 해독 (Viterbi decoding), 기대값 최대화, 칼만 여과 (Kalman filtering), 또는 신경망 (neural network) 방법을 사용하여 수행되는 정규화 및 검출을 포함한다.

[0031] 일부 실시양태에서, 본 개시내용의 방법은 질환 진행의 모니터링, 잔류 질환의 모니터링, 요법의 모니터링, 상태의 진단, 상태의 예측, 또는 발견된 변이체를 기초로 한 요법의 선택을 포함한다.

[0032] 일부 실시양태에서, 요법은 가장 최근의 샘플 분석을 기초로 하여 변형된다. 또한, 본 개시내용의 방법은

종양, 감염 또는 다른 조직 비정상의 유전자 프로파일을 추정하는 것을 포함한다. 일부 실시양태에서, 종양, 감염 또는 다른 조직 비정상의 성장, 완화 또는 진행이 모니터링된다. 일부 실시양태에서, 대상체의 면역계는 한번에 또는 시간에 걸쳐 분석되고 모니터링된다.

[0033] 일부 실시양태에서, 본 개시내용의 방법은 확인된 변이체를 야기하는 것으로 의심되는 조직 비정상의 위치결정을 위한 영상화 시험 (예를 들어, CT, PET-CT, MRI, X-선, 초음파)을 통해 추적조사되는 변이체의 확인을 포함한다.

[0034] 일부 실시양태에서, 본 개시내용의 방법은 동일한 환자로부터의 조직 또는 종양 생검으로부터 얻은 유전자 데이터의 사용을 포함한다. 일부 실시양태에서, 종양, 감염 또는 다른 조직 비정상의 계통발생학이 추정된다.

[0035] 일부 실시양태에서, 본 개시내용의 방법은 집단-기반 노-콜링 (no-calling)의 수행 및 낮은-신뢰도 영역의 확인을 포함한다. 일부 실시양태에서, 서열 적용범위 (coverage)에 대한 측정 데이터를 얻는 것은 게놈의 모든 위치에서 서열 적용범위 깊이를 측정하는 것을 포함한다. 일부 실시양태에서, 서열 적용범위 편향 (bias)에 대한 측정 데이터를 보정하는 것은 윈도우-평균된 (window-averaged) 적용범위의 계산을 포함한다. 일부 실시양태에서, 서열 적용범위 편향에 대한 측정 데이터를 보정하는 것은 라이브러리 구축 및 서열분석 과정에서 GC 편향을 설명하기 위해 조정을 수행하는 것을 포함한다. 일부 실시양태에서, 서열 적용범위 편향에 대한 측정 데이터를 보정하는 것은 편향을 보상하기 위해 개별 맵핑과 연관된 추가의 가중 인자를 기초로 하여 조정을 수행하는 것을 포함한다.

[0036] 일부 실시양태에서, 본 개시내용의 방법은 이환된 세포 기원으로부터 유래된 세포외 폴리뉴클레오티드를 포함한다. 일부 실시양태에서, 세포외 폴리뉴클레오티드는 건강한 세포 기원으로부터 유래된다.

[0037] 본 개시내용은 다음 단계를 수행하기 위한 컴퓨터 판독가능 매체를 포함하는 시스템을 제공한다: 게놈 내의 미리 규정된 영역을 선택하고; 미리 규정된 영역 내의 서열 판독체의 수를 계수하고; 서열 판독체의 수를 미리 규정된 영역에 걸쳐 정규화하고; 미리 규정된 영역 내의 카피수 변이의 퍼센트를 결정하는 것. 일부 실시양태에서, 게놈 전체 또는 적어도 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% 또는 90%의 게놈이 분석된다. 일부 실시양태에서, 컴퓨터 판독가능 매체는 혈장 또는 혈청 내의 암 DNA 또는 RNA 퍼센트에 대한 데이터를 최종 사용자에게 제공한다.

[0038] 일부 실시양태에서, 유전자 변이, 예컨대 다형성 또는 원인 (causal) 변이체의 양이 분석된다. 일부 실시양태에서, 유전자 변경의 존재 또는 부재가 검출된다.

[0039] 본 개시내용은 또한 a) 대상체로부터의 신체 샘플로부터의 세포외 폴리뉴클레오티드를 서열분석하고, 여기서 각각의 세포외 폴리뉴클레오티드는 다수의 서열분석 판독체를 생성하고; b) 설정된 역치를 충족하지 않는 판독체를 여과 제거하고; c) 서열분석으로부터 유래된 서열 판독체를 참조 서열 상에 맵핑하고; d) 각각의 맵핑가능한 염기 위치에서 참조 서열의 변이체와 정렬되는 맵핑된 서열 판독체의 하위세트를 확인하고; e) 각각의 맵핑가능한 염기 위치에 대해, (a) 참조 서열에 비해 변이체를 포함하는 맵핑된 서열 판독체의 수 대 (b) 각각의 맵핑가능한 염기 위치에 대한 총 서열 판독체의 수의 비를 계산하고; f) 각각의 맵핑가능한 염기 위치에 대한 변이의 비 또는 빈도를 정규화하고 잠재적인 희귀 변이체(들) 또는 다른 유전자 변경(들)을 결정하고; g) 생성되는 수를 각각의 영역에 대해 비교하는 것을 포함하는, 대상체로부터 얻은 세포 유리 또는 실질적인 세포 유리 샘플에서 희귀 돌연변이를 검출하는 방법을 제공한다.

[0040] 본 개시내용은 또한 a. 태그부착된 (tagged) 모 폴리뉴클레오티드의 적어도 하나의 세트를 제공하고, 태그부착된 모 폴리뉴클레오티드의 각각의 세트에 대해; b. 증폭된 자손 (progeny) 폴리뉴클레오티드의 상응하는 세트를 생산하기 위해 세트 내의 태그부착된 모 폴리뉴클레오티드를 증폭하고; c. 증폭된 자손 폴리뉴클레오티드의 세트의 하위세트 (적절한 하위세트 포함)를 서열분석하여 서열분석 판독체의 세트를 생산하고; d. 컨센서스 (consensus) 서열의 세트를 생성하기 위해 서열분석 판독체의 세트를 붕괴시키는 것을 포함하는 방법을 제공하고, 여기서 각각의 컨센서스 서열은 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 상응한다. 특정 실시양태에서, 방법은 e. 태그부착된 모 분자의 각각의 세트에 대해 컨센서스 서열의 세트를 분석하는 것을 추가로 포함한다.

[0041] 일부 실시양태에서, 세트 내의 각각의 폴리뉴클레오티드는 참조 서열에 맵핑가능하다.

[0042] 일부 실시양태에서, 방법은 태그부착된 모 폴리뉴클레오티드의 다수의 세트를 제공하는 것을 포함하고, 여기서 각각의 세트는 상이한 참조 서열에 맵핑가능하다.

[0043] 일부 실시양태에서, 방법은 초기 출발 유전 물질을 태그부착된 모 폴리뉴클레오티드로 전환하는 것을 추가로 포함한다.

[0044] 일부 실시양태에서, 초기 출발 유전 물질은 100 ng 이하의 폴리뉴클레오티드를 포함한다.

[0045] 일부 실시양태에서, 방법은 전환 전에 초기 출발 유전 물질의 병목현상화 (bottlenecking)를 포함한다.

[0046] 일부 실시양태에서, 방법은 적어도 10%, 적어도 20%, 적어도 30%, 적어도 40%, 적어도 50%, 적어도 60%, 적어도 80% 또는 적어도 90%의 전환 효율로 초기 출발 유전 물질을 태그부착된 모 폴리뉴클레오티드로 전환하는 것을 포함한다.

[0047] 일부 실시양태에서, 전환은 임의의 평활 (blunt)-말단 라이게이션 (ligation), 점착성 (sticky) 말단 라이게이션, 분자 역위 프로브, PCR, 라이게이션-기반 PCR, 단일 가닥 라이게이션 및 단일 가닥 환형화 (circularization)를 포함한다.

[0048] 일부 실시양태에서, 초기 출발 유전 물질은 세포 유리 핵산이다.

[0049] 일부 실시양태에서, 다수의 참조 서열은 동일한 게놈으로부터 유래된 것이다.

[0050] 일부 실시양태에서, 세트 내의 각각의 태그부착된 모 폴리뉴클레오티드는 특유하게 태그부착된다.

[0051] 일부 실시양태에서, 태그는 비-특유한 것이다.

[0052] 일부 실시양태에서, 컨센서스 서열의 생성은 태그로부터의 정보 및/또는 서열 판독체의 개시 (출발) 및 종료 (정지) 영역에서의 적어도 하나의 서열 정보 및 서열 판독체의 길이를 기초로 한다.

[0053] 일부 실시양태에서, 방법은 태그부착된 모 폴리뉴클레오티드의 세트 내의 적어도 20%, 적어도 30%, 적어도 40%, 적어도 50%, 적어도 60%, 적어도 70%, 적어도 80%, 적어도 90%, 적어도 95%, 적어도 98%, 적어도 99%, 적어도 99.9% 또는 적어도 99.99%의 특유한 폴리뉴클레오티드 각각으로부터의 적어도 하나의 자손에 대한 서열 판독체를 생산하기에 충분한 증폭된 자손 폴리뉴클레오티드의 세트의 하위세트를 서열분석하는 것을 포함한다.

[0054] 일부 실시양태에서, 적어도 하나의 자손은 다수의 자손, 예를 들어 적어도 2, 적어도 5 또는 적어도 10개의 자손이다.

[0055] 일부 실시양태에서, 서열 판독체의 세트 내의 서열 판독체의 수는 태그부착된 모 폴리뉴클레오티드의 세트 내의 특유한 태그부착된 모 폴리뉴클레오티드의 수보다 더 크다.

[0056] 일부 실시양태에서, 서열분석된 증폭된 자손 폴리뉴클레오티드의 세트의 하위세트는 사용되는 서열분석 플랫폼의 염기당 서열분석 오류 비율 백분율과 동일한 백분율로 태그부착된 모 폴리뉴클레오티드의 세트에 나타나는 임의의 뉴클레오티드 서열이 컨센서스 서열의 세트 중에서 나타날 가능성이 적어도 50%, 적어도 60%, 적어도 70%, 적어도 80%, 적어도 90%, 적어도 95%, 적어도 98%, 적어도 99%, 적어도 99.9% 또는 적어도 99.99%가 되도록 하기에 충분한 크기를 갖는다.

[0057] 일부 실시양태에서, 방법은 (i) 태그부착된 모 폴리뉴클레오티드로 전환되는 초기 출발 유전 물질로부터의 서열의 선택적 증폭; (ii) 태그부착된 모 폴리뉴클레오티드의 선택적 증폭; (iii) 증폭된 자손 폴리뉴클레오티드의 선택적 서열 포획; 또는 (iv) 초기 출발 유전 물질의 선택적 서열 포획에 의해, 하나 이상의 선택된 참조 서열에 맵핑되는 폴리뉴클레오티드에 대한 증폭된 자손 폴리뉴클레오티드의 세트를 풍부화하는 것을 포함한다.

[0058] 일부 실시양태에서, 분석은 컨센서스 서열의 세트로부터 얻은 측정치 (예를 들어, 수)를 대조 샘플로부터의 컨센서스 서열의 세트로부터 얻은 측정치에 대해 정규화하는 것을 포함한다.

[0059] 일부 실시양태에서, 검출은 돌연변이, 희귀 돌연변이, 단일 뉴클레오티드 변이체, 삽입-결실, 카피수 변이, 염기변환, 전위, 역위, 결실, 이수성, 부분적 이수성, 배수성, 염색체 불안정성, 염색체 구조 변경, 유전자 융합, 염색체 융합, 유전자 말단절단, 유전자 증폭, 유전자 중복, 염색체 병변, DNA 병변, 핵산 화학적 변형의 비정상적인 변화, 후성적 패턴의 비정상적인 변화, 핵산 메틸화 감염의 비정상적인 변화 또는 암의 검출을 포함한다.

[0060] 일부 실시양태에서, 폴리뉴클레오티드는 DNA, RNA, 이 둘의 조합 또는 DNA + RNA-유래 cDNA를 포함한다.

[0061] 일부 실시양태에서, 폴리뉴클레오티드의 특정 하위세트는 폴리뉴클레오티드의 초기 세트로부터의 또는 증폭된 폴리뉴클레오티드로부터의 염기쌍의 폴리뉴클레오티드 길이를 기초로 하여 선택되거나 풍부화된다.

[0062] 일부 실시양태에서, 분석은 개체 내의 비정상 또는 질환, 예컨대 감염 및/또는 암의 검출 및 모니터링을 추가로

포함한다.

[0063] 일부 실시양태에서, 방법은 면역 레퍼토리 (immune repertoire) 프로파일링과 조합하여 수행된다.

[0064] 일부 실시양태에서, 폴리뉴클레오티드는 혈액, 혈장, 혈청, 소변, 타액, 점막 분비물, 객담, 대변 및 눈물로 이루어진 군으로부터 추출된다.

[0065] 일부 실시양태에서, 붕괴는 태그부착된 모 폴리뉴클레오티드 또는 증폭된 자손 폴리뉴클레오티드의 센스 또는 안티센스 가닥에 존재하는 오류, 닉 (nick) 또는 병변의 검출 및/또는 보정을 포함한다.

[0066] 본 개시내용은 또한 적어도 5%, 적어도 1%, 적어도 0.5%, 적어도 0.1% 또는 적어도 0.05%의 감도로 초기 출발 유전 물질 내의 유전자 변이를 검출하는 것을 포함하는 방법을 제공한다. 일부 실시양태에서, 초기 출발 유전 물질은 100 ng 미만 양의 핵산으로 제공되고, 유전자 변이는 카피수/이형접합성 변이이고, 검출은 하위-염색체 해상도; 예를 들어, 적어도 100 메가염기 해상도, 적어도 10 메가염기 해상도, 적어도 1 메가염기 해상도, 적어도 100 킬로염기 해상도, 적어도 10 킬로염기 해상도 또는 적어도 1 킬로염기 해상도로 수행된다. 또 다른 실시양태에서, 방법은 태그부착된 모 폴리뉴클레오티드의 다수의 세트를 제공하는 것을 포함하고, 여기서 각각의 세트는 상이한 참조 서열에 맵핑가능하다. 또 다른 실시양태에서, 참조 서열은 종양 마커의 유전자좌이고, 분석은 컨센서스 서열의 세트 내의 종양 마커를 검출하는 것을 포함한다. 또 다른 실시양태에서, 종양 마커는 증폭 단계에서 도입되는 오류 비율보다 낮은 빈도로 컨센서스 서열의 세트에 존재한다. 또 다른 실시양태에서, 적어도 하나의 세트는 다수의 세트이고, 참조 서열은 각각이 종양 마커의 유전자좌인 다수의 참조 서열을 포함한다. 또 다른 실시양태에서, 모 폴리뉴클레오티드의 적어도 2개의 세트 사이의 컨센서스 서열의 카피수 변이를 검출하는 것을 포함한다. 또 다른 실시양태에서, 분석은 참조 서열에 비해 서열 변이의 존재를 검출하는 것을 포함한다. 또 다른 실시양태에서, 분석은 참조 서열에 비해 서열 변이의 존재를 검출하고, 모 폴리뉴클레오티드의 적어도 2개의 세트 사이의 컨센서스 서열의 카피수 변이를 검출하는 것을 포함한다. 또 다른 실시양태에서, 붕괴는 i. 증폭된 자손 폴리뉴클레오티드로부터 서열분석된 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 동일한 태그부착된 모 폴리뉴클레오티드로부터 증폭된 것이고; ii. 패밀리 내의 서열 판독체를 기초로 하여 컨센서스 서열을 결정하는 것을 포함한다.

[0067] 본 개시내용은 또한 다음 단계를 수행하기 위한 컴퓨터 판독가능 매체를 포함하는 시스템을 제공한다: a. 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트를 제공하고, 태그부착된 모 폴리뉴클레오티드의 각각의 세트에 대해; b. 세트 내의 태그부착된 모 폴리뉴클레오티드를 증폭시켜 상응하는 증폭된 자손 폴리뉴클레오티드의 세트를 생산하고; c. 증폭된 자손 폴리뉴클레오티드의 세트의 하위세트 (적절한 하위세트 포함)를 서열분석하여 서열분석 판독체의 세트를 생산하고; d. 서열분석 판독체의 세트를 붕괴시켜, 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성하고, 임의로, e. 태그부착된 모 분자의 각각의 세트에 대해 컨센서스 서열의 세트를 분석하는 단계.

[0068] 본 개시내용은 a. 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트를 제공하고, 태그부착된 모 폴리뉴클레오티드의 각각의 세트에 대해; b. 세트 내의 태그부착된 모 폴리뉴클레오티드를 증폭시켜 상응하는 증폭된 자손 폴리뉴클레오티드의 세트를 생산하고; c. 증폭된 자손 폴리뉴클레오티드의 세트의 하위세트 (적절한 하위세트 포함)를 서열분석하여 서열분석 판독체의 세트를 생산하고; d. 서열분석 판독체의 세트를 붕괴시켜, 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성하고; e. 컨센서스 서열 중에서 품질 역치를 충족하지 않는 것을 여과 제거하는 단계를 포함하는 방법을 제공한다. 한 실시양태에서, 품질 역치는 컨센서스 서열로 붕괴되는 증폭된 자손 폴리뉴클레오티드로부터의 서열 판독체의 수를 고려한다. 또 다른 실시양태에서, 품질 역치는 컨센서스 서열로 붕괴되는 증폭된 자손 폴리뉴클레오티드로부터의 서열 판독체의 수를 고려한다. 본 개시내용은 또한 상기 방법을 수행하기 위한 컴퓨터 판독가능 매체를 포함하는 시스템을 제공한다.

[0069] 본 개시내용은 또한 a. 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트를 제공하고, 여기서 각각의 세트는 하나 이상의 게놈 내의 상이한 참조 서열에 맵핑되고, 태그부착된 모 폴리뉴클레오티드의 각각의 세트에 대해; i. 제1 폴리뉴클레오티드를 증폭시켜 증폭된 폴리뉴클레오티드의 세트를 생산하고; ii. 증폭된 폴리뉴클레오티드의 세트의 하위세트를 서열분석하여 서열분석 판독체의 세트를 생산하고; iii. 1. 증폭된 자손 폴리뉴클레오티드로부터 서열분석된 서열 판독체를, 각각의 패밀리가 동일한 태그부착된 모 폴리뉴클레오티드로부터 증폭된 것인 패밀리로 분류하여 서열 판독체를 붕괴시키는 것을 포함하는 방법을 제공한다. 한 실시양태에서, 붕괴는 2. 각각의 패밀리 내의 서열 판독체의 정량적 척도를 결정하는 것을 추가로 포함한다. 또 다른 실시양태에서, 방법은 (a를 포함하면서): b. 특유한 패밀리의 정량적 척도를 결정하고; c. (1) 특유한 패밀리의 정량적

척도 및 (2) 각각의 군 내의 서열 판독체의 정량적 척도를 기초로 하여, 세트 내의 특유한 태그부착된 모 폴리뉴클레오티드의 척도를 추정하는 것을 추가로 포함한다. 또 다른 실시양태에서, 추정은 통계적 또는 확률적 모델을 이용하여 수행된다. 또 다른 실시양태에서, 적어도 하나의 세트는 다수의 세트이다. 또 다른 실시양태에서, 방법은 2개의 세트 사이의 증폭 또는 표상적 편향 (representational bias)에 대한 보정을 추가로 포함한다. 또 다른 실시양태에서, 방법은 2개의 세트 사이의 증폭 또는 표상적 편향을 보정하기 위해 대조군 또는 대조 샘플의 세트를 사용하는 것을 추가로 포함한다. 또 다른 실시양태에서, 방법은 세트 사이의 카피수 변이를 결정하는 것을 추가로 포함한다. 또 다른 실시양태에서, 방법은 (a, b, c를 포함하면서): d. 패밀리 중에서 다형체 형태의 정량적 척도를 결정하고; e. 결정된 다형체 형태의 정량적 척도를 기초로 하여, 다형체 형태의 정량적 척도를 추정된 특유한 태그부착된 모 폴리뉴클레오티드의 수로 추정하는 것을 추가로 포함한다. 또 다른 실시양태에서, 다형체 형태는 치환, 삽입, 결실, 역위, 미소부수체 변화, 염기변환, 전위, 융합, 메틸화, 과메틸화, 히드록시메틸화, 아세틸화, 후성적 변이체, 조절-연관 변이체 또는 단백질 결합 부위를 포함하나 이에 제한되지는 않는다. 또 다른 실시양태에서, 세트는 공통 샘플로부터 유래하고, 방법은 a. 각각의 다수의 참조 서열에 맵핑되는 각각의 세트 내의 태그부착된 모 폴리뉴클레오티드의 추정된 수의 비교를 기초로 하여 다수의 세트에 대한 카피수 변이를 추정하는 것을 추가로 포함한다. 또 다른 실시양태에서, 각각의 세트 내의 폴리뉴클레오티드의 본래의 수가 추가로 추정된다. 본 개시내용은 또한 상기 방법을 수행하기 위한 컴퓨터 판독가능 매체를 포함하는 시스템을 제공한다.

[0070] 본 개시내용은 또한 폴리뉴클레오티드를 포함하는 샘플 내의 카피수 변이를 결정하는 방법을 제공하고, 이 방법은 a. 제1 폴리뉴클레오티드의 적어도 2개의 세트를 제공하고, 여기서 각각의 세트는 게놈 내의 상이한 참조 서열에 맵핑되고, 제1 폴리뉴클레오티드의 각각의 세트에 대해; i. 폴리뉴클레오티드를 증폭시켜 증폭된 폴리뉴클레오티드의 세트를 생산하고; ii. 증폭된 폴리뉴클레오티드의 세트의 하위세트를 서열분석하여 서열분석 판독체의 세트를 생산하고; iii. 증폭된 폴리뉴클레오티드로부터 서열분석된 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 세트 내의 동일한 제1 폴리뉴클레오티드로부터 증폭된 것이고; iv. 세트 내의 패밀리의 정량적 척도를 추정하고; b. 각각의 세트 내의 패밀리의 정량적 척도를 비교함으로써 카피수 변이를 결정하는 것을 포함한다. 본 개시내용은 또한 상기 방법을 수행하기 위한 컴퓨터 판독가능 매체를 포함하는 시스템을 제공한다.

[0071] 본 개시내용은 또한 a. 제1 폴리뉴클레오티드의 적어도 하나의 세트를 제공하고, 여기서 각각의 세트는 하나 이상의 게놈 내의 상이한 참조 서열에 맵핑되고, 제1 폴리뉴클레오티드의 각각의 세트에 대해; i. 제1 폴리뉴클레오티드를 증폭시켜 증폭된 폴리뉴클레오티드의 세트를 생산하고; ii. 증폭된 폴리뉴클레오티드의 세트의 하위세트를 서열분석하여 서열분석 판독체의 세트를 생산하고; iii. 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 제1 폴리뉴클레오티드로부터 증폭된 폴리뉴클레오티드의 서열 판독체를 포함하고; b. 제1 폴리뉴클레오티드의 각각의 세트에 대해, 제1 폴리뉴클레오티드의 세트 내의 하나 이상의 염기에 대한 콜 (call) 빈도를 추정하고, 여기서 추정은 i. 각각의 패밀리에 대해, 각각의 다수의 콜에 대한 신뢰도 점수를 배정하고, 신뢰도 점수는 패밀리의 구성원 중에서 콜의 빈도를 고려한 것이고; ii. 각각의 패밀리에 배정된 하나 이상의 콜의 신뢰도 점수를 고려하여 하나 이상의 콜의 빈도를 평가하는 것을 포함하는 것임을 포함하는, 폴리뉴클레오티드의 샘플 내의 서열 콜의 빈도를 추정하는 방법을 제공한다. 본 개시내용은 또한 상기 방법을 수행하기 위한 컴퓨터 판독가능 매체를 포함하는 시스템을 제공한다.

[0072] 본 개시내용은 또한 a. 적어도 하나의 개별 폴리뉴클레오티드 분자를 제공하고; b. 신호를 생성하기 위해 적어도 하나의 개별 폴리뉴클레오티드 분자 내의 서열 정보를 암호화(encoding)하고; c. 적어도 일부의 신호를 채널에 통과시켜 적어도 하나의 개별 폴리뉴클레오티드분자에 대한 뉴클레오티드 서열 정보를 포함하는 수신된 신호를 생산하고, 여기서 수신된 신호는 노이즈 (noise) 및/또는 왜곡을 포함하고; d. 수신된 신호를 해독하여 적어도 하나의 개별 폴리뉴클레오티드 분자에 대한 서열 정보를 포함하는 메세지 (message)를 생산하고, 여기서 해독은 메세지 내의 노이즈 및/또는 왜곡을 감소시키고; e. 메세지를 수신자에게 제공하는 것을 포함하는, 적어도 하나의 개별 폴리뉴클레오티드 분자에 대한 서열 정보를 통신하는 방법을 제공한다. 한 실시양태에서, 노이즈는 부정확한 뉴클레오티드 콜을 포함한다. 또 다른 실시양태에서, 왜곡은 다른 개별 폴리뉴클레오티드 분자에 비해 개별 폴리뉴클레오티드 분자의 불균등한 증폭을 포함한다. 또 다른 실시양태에서, 왜곡은 증폭 또는 서열분석 편향에 의해 발생한다. 또 다른 실시양태에서, 적어도 하나의 개별 폴리뉴클레오티드 분자는 다수의 개별 폴리뉴클레오티드 분자이고, 해독은 다수의 각각의 분자에 대한 메세지를 생산한다. 또 다른 실시양태에서, 암호화는 임의로 태그부착된 적어도 개별 폴리뉴클레오티드 분자를 증폭시키는 것을 포함하고, 신호는 증폭된 분자의 수집물을 포함한다. 또 다른 실시양태에서, 채널은 폴리뉴클레오티드 서열분석기를 포함하고, 수신된 신호는 적어도 하나의 개별 폴리뉴클레오티드 분자로부터 증폭된 다수의 폴리뉴클레오티드의 서열 판독체를 포함한다. 또 다른 실시양태에서, 해독은 적어도 하나의 개별 폴리뉴클레오티드 분자 각각으로부터 증폭된 분자의

서열 판독체를 분류하는 것을 포함한다. 또 다른 실시양태에서, 해독은 생성된 서열 신호를 여과하는 확률적 또는 통계적 방법으로 이루어진다. 본 개시내용은 또한 상기 방법을 수행하기 위한 컴퓨터 판독가능 매체를 포함하는 시스템을 제공한다.

[0073] 또 다른 실시양태에서, 폴리뉴클레오티드는 종양 게놈 DNA 또는 RNA로부터 유래된다. 또 다른 실시양태에서, 폴리뉴클레오티드는 세포 유리 폴리뉴클레오티드, 엑소솜 폴리뉴클레오티드, 박테리아 폴리뉴클레오티드 또는 바이러스 폴리뉴클레오티드로부터 유래된다. 또 다른 실시양태에서, 방법은 영향받는 분자 경로의 검출 및/또는 연관분석을 추가로 포함한다. 또 다른 실시양태에서, 방법은 개체의 건강 또는 질환 상태의 순차적인 모니터링을 추가로 포함한다. 또 다른 실시양태에서, 개체 내의 질환과 연관된 게놈의 계통발생이 추정된다. 또 다른 실시양태에서, 방법은 질환의 진단, 모니터링 또는 치료를 추가로 포함한다. 또 다른 실시양태에서, 치료 요법은 다형체 형태 또는 CNV 또는 연관 경로를 기초로 하여 선택되거나 변형된다. 또 다른 실시양태에서, 치료는 조합 요법을 포함한다.

[0074] 본 개시내용은 또한 다음 단계를 수행하도록 설정된 실행가능 코드를 포함하는 비-일시적인, 유형 (tangible) 형태의 컴퓨터 판독가능 매체를 제공한다: 게놈 내의 미리 규정된 영역을 선택하고; 서열 판독체에 접근하여 미리 규정된 영역 내의 서열 판독체의 수를 계수하고; 서열 판독체의 수를 미리 규정된 영역에 걸쳐 정규화하고; 미리 규정된 영역 내의 카피수 변이의 퍼센트를 결정하는 것.

[0075] 본 개시내용은 또한 다음 단계를 수행하도록 설정된 실행가능 코드를 포함하는 비-일시적인, 유형 형태의 컴퓨터 판독가능 매체를 제공한다: a. 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고; b. 설정된 역치를 충족하지 않는 판독체를 여과 제거하고; c. 서열분석으로부터 유래된 서열 판독체를 참조 서열 상에 맵핑하고; d. 각각의 맵핑가능한 염기 위치에서 참조 서열의 변이체와 정렬되는 맵핑된 서열 판독체의 하위세트를 확인하고; e. 각각의 맵핑가능한 염기 위치에 대해, (a) 참조 서열에 비해 변이체를 포함하는 맵핑된 서열 판독체의 수 대 (b) 각각의 맵핑가능한 염기 위치에 대한 총 서열 판독체의 수의 비를 계산하고; f. 각각의 맵핑가능한 염기 위치에 대한 변이의 비 또는 빈도를 정규화하고 잠재적인 희귀 변이체(들) 또는 다른 유전자 변경(들)을 결정하고; g. 잠재적인 희귀 변이체(들) 또는 돌연변이(들)를 갖는 각각의 영역에 대해 생성된 수를 참조 샘플로부터 유사하게 유래된 수와 비교하는 단계.

[0076] 본 개시내용은 또한 다음 단계를 수행하도록 설정된 실행가능 코드를 포함하는 비-일시적인 유형 형태의 컴퓨터 판독가능 매체를 제공한다: a. 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고; b. 서열분석 판독체의 세트를 붕괴시켜, 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성하는 단계.

[0077] 본 개시내용은 또한 다음 단계를 수행하도록 설정된 실행가능 코드를 포함하는 비-일시적인, 유형 형태의 컴퓨터 판독가능 매체를 제공한다: a. 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고; b. 서열분석 판독체의 세트를 붕괴시켜, 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성하고; c. 컨센서스 서열 중에서 품질 역치를 충족하지 않는 것을 여과 제거하는 단계.

[0078] 본 개시내용은 또한 다음 단계를 수행하도록 설정된 실행가능 코드를 포함하는 비-일시적인, 유형 형태의 컴퓨터 판독가능 매체를 제공한다: a. 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고; i. 1. 증폭된 자손 폴리뉴클레오티드로부터 서열분석된 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 동일한 태그부착된 모 폴리뉴클레오티드로부터 증폭된 것이고, 임의로, 2. 각각의 패밀리 내의 서열 판독체의 정량적 척도를 결정함으로써 서열 판독체를 붕괴시키는 단계. 특정 실시양태에서, 실행가능 코드는 b. 특유한 패밀리의 정량적 척도를 결정하고; c. (1) 특유한 패밀리의 정량적 척도 및 (2) 각각의 군 내의 서열 판독체의 정량적 척도를 기초로 하여, 세트 내의 특유한 태그부착된 모 폴리뉴클레오티드의 척도를 추정하는 단계를 추가로 수행한다. 특정 실시양태에서, 실행가능 코드는 d. 패밀리 중에서 다형체 형태의 정량적 척도를 결정하고; e. 결정된 다형체 형태의 정량적 척도를 기초로 하여, 다형체 형태의 정량적 척도를 추정된 특유한 태그부착된 모 폴리뉴클레오티드의 수로 추정하는 단계를 추가로 수행한다.

[0079] 본 개시내용은 또한 다음 단계를 수행하도록 설정된 실행가능 코드를 포함하는 비-일시적인, 유형 형태의 컴퓨터 판독가능 매체를 제공한다: a. 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판

독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고, 증폭된 폴리뉴클레오티드로부터 서열분석된 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 세트 내의 동일한 제1 폴리뉴클레오티드로부터 증폭된 것이고; b. 세트 내의 패밀리의 정량적 척도를 추정하고; c. 각각의 세트 내의 패밀리의 정량적 척도를 비교함으로써 카피수 변이를 결정하는 단계.

[0080] 본 개시내용은 또한 다음 단계를 수행하도록 설정된 실행가능 코드를 포함하는 비-일시적인, 유형 형태의 컴퓨터 판독가능 매체를 제공한다: a. 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고, 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 동일한 제1 폴리뉴클레오티드로부터 증폭된 폴리뉴클레오티드의 서열 판독체를 포함하고; b. 제1 폴리뉴클레오티드의 각각의 세트에 대해, 제1 폴리뉴클레오티드의 세트 내의 하나 이상의 염기에 대한 콜 빈도를 추정하고, 여기서 추정은 c. 각각의 패밀리에 대해, 각각의 다수의 콜에 대한 신뢰도 점수를 배정하고, 신뢰도 점수는 패밀리의 구성원 중에서 콜의 빈도를 고려한 것이고; d. 각각의 패밀리에 배정된 하나 이상의 콜의 신뢰도 점수를 고려하여 하나 이상의 콜의 빈도를 평가하는 것을 포함하는 단계.

[0081] 본 개시내용은 또한 다음 단계를 수행하도록 설정된 실행가능 코드를 포함하는 비-일시적인, 유형 형태의 컴퓨터 판독가능 매체를 제공한다: a. 적어도 하나의 개별 폴리뉴클레오티드 분자로부터의 암호화된 서열 정보를 포함하는 수신된 신호를 포함하는 데이터 파일에 접근하고, 여기서 수신된 신호는 노이즈 및/또는 왜곡을 포함하고; b. 수신된 신호를 해독하여 적어도 하나의 개별 폴리뉴클레오티드 분자에 대한 서열 정보를 포함하는 메세지를 생산하고, 여기서 해독은 메세지 내에서 각각의 개별 폴리뉴클레오티드에 대한 노이즈 및/또는 왜곡을 감소시키고; c. 적어도 하나의 개별 폴리뉴클레오티드 분자에 대한 서열 정보를 포함하는 메세지를 컴퓨터 파일에 기록하는 단계.

[0082] 본 개시내용은 또한 다음 단계를 수행하도록 설정된 실행가능 코드를 포함하는 비-일시적인, 유형 형태의 컴퓨터 판독가능 매체를 제공한다: a. 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고; b. 서열분석 판독체의 세트를 붕괴시켜, 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성하고; c. 컨센서스 서열 중에서 품질 역치를 충족하지 않는 것을 여과 제거하는 단계.

[0083] 본 개시내용은 또한 다음 단계를 수행하도록 설정된 실행가능 코드를 포함하는 비-일시적인, 유형 형태의 컴퓨터 판독가능 매체를 제공한다: a. 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고; b. i. 증폭된 자손 폴리뉴클레오티드로부터 서열분석된 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 동일한 태그부착된 모 폴리뉴클레오티드로부터 증폭된 것이고; ii. 임의로, 각각의 패밀리 내의 서열 판독체의 정량적 척도를 결정함으로써 서열 판독체를 붕괴시키는 단계. 특정 실시양태에서, 실행가능 코드는 c. 특유한 패밀리의 정량적 척도를 결정하고; d. (1) 특유한 패밀리의 정량적 척도 및 (2) 각각의 군 내의 서열 판독체의 정량적 척도를 기초로 하여, 세트 내의 특유한 태그부착된 모 폴리뉴클레오티드의 척도를 추정하는 단계를 추가로 수행한다. 특정 실시양태에서, 실행가능 코드는 e. 패밀리 중에서 다형체 형태의 정량적 척도를 결정하고; f. 결정된 다형체 형태의 정량적 척도를 기초로 하여, 다형체 형태의 정량적 척도를 추정된 특유한 태그부착된 모 폴리뉴클레오티드의 수로 추정하는 단계를 추가로 수행한다. 특정 실시양태에서, 실행가능 코드는 e. 각각의 다수의 참조 서열에 맵핑되는 각각의 세트 내의 태그부착된 모 폴리뉴클레오티드의 추정된 수의 비교를 기초로 하여 다수의 세트에 대한 카피수 변이를 추정하는 단계를 추가로 수행한다.

[0084] 본 개시내용은 또한 다음 단계를 수행하도록 설정된 실행가능 코드를 포함하는 비-일시적인, 유형 형태의 컴퓨터 판독가능 매체를 제공한다: a. 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고; b. 증폭된 폴리뉴클레오티드로부터 서열분석된 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 세트 내의 동일한 제1 폴리뉴클레오티드로부터 증폭된 것이고; c. 세트 내의 패밀리의 정량적 척도를 추정하고; d. 각각의 세트 내의 패밀리의 정량적 척도를 비교함으로써 카피수 변이를 결정하는 단계.

[0085] 본 개시내용은 또한 다음 단계를 수행하도록 설정된 실행가능 코드를 포함하는 비-일시적인, 유형 형태의 컴퓨터 판독가능 매체를 제공한다: a. 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부

터 유래하고, 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 동일한 제1 폴리뉴클레오티드로부터 증폭된 폴리뉴클레오티드의 서열 판독체를 포함하고; b. 제1 폴리뉴클레오티드의 각각의 세트에 대해, 제1 폴리뉴클레오티드의 세트 내의 하나 이상의 염기에 대한 콜 빈도를 추정하고, 여기서 추정은 i. 각각의 패밀리에 대해, 각각의 다수의 콜에 대한 신뢰도 점수를 배정하고, 신뢰도 점수는 패밀리의 구성원 중에서 콜의 빈도를 고려한 것이고; ii. 각각의 패밀리에 배정된 하나 이상의 콜의 신뢰도 점수를 고려하여 하나 이상의 콜의 빈도를 평가하는 것을 포함하는 단계.

[0086] 본 개시내용은 또한 a. 세포 유리 DNA (cfDNA) 폴리뉴클레오티드의 100 내지 100,000개의 반수체 인간 게놈 등가물을 포함하는 샘플을 제공하고; b. 폴리뉴클레오티드를 2 내지 1,000,000개의 특유한 식별자 (identifier)로 태그부착하는 것을 포함하는 방법을 제공한다. 특정 실시양태에서, 특유한 식별자의 수는 적어도 3, 적어도 5, 적어도 10, 적어도 15 또는 적어도 25 및 최대 100, 최대 1000 또는 최대 10,000이다. 특정 실시양태에서, 특유한 식별자의 수는 최대 100, 최대 1000, 최대 10,000, 최대 100,000이다.

[0087] 본 개시내용은 또한 a. 단편화된 폴리뉴클레오티드의 다수의 인간 반수체 게놈 등가물을 포함하는 샘플을 제공하고; b. z를 결정하고, 여기서 z는 게놈 내의 임의의 위치에서 출발하는 예상된 수의 중복 폴리뉴클레오티드의 중심 경향도 (central tendency) (예를 들어, 평균, 중간값 또는 최빈값)의 척도이고, 중복 폴리뉴클레오티드는 동일한 출발 및 정지 위치를 갖고; c. 샘플 내의 폴리뉴클레오티드를 n개의 특유한 식별자로 태그부착하고, 여기서 n은 2 내지 100,000*z, 2 내지 10,000*z, 2 내지 1,000*z 또는 2 내지 100*z임을 포함하는 방법을 제공한다.

[0088] 본 개시내용은 또한 a. 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트를 제공하고, 태그부착된 모 폴리뉴클레오티드의 각각의 세트에 대해; b. 세트 내의 각각의 태그부착된 모 폴리뉴클레오티드에 대한 다수의 서열 판독체를 생산하여 서열분석 판독체의 세트를 생산하고; c. 서열분석 판독체의 세트를 붕괴시켜, 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성하는 것을 포함하는 방법을 제공한다.

[0089] 본 개시내용은 a) 대상체로부터의 신체 샘플로부터의 세포외 폴리뉴클레오티드를 서열분석하고, 여기서 각각의 세포외 폴리뉴클레오티드는 다수의 서열분석 판독체를 생성하고; b) 설정된 역치를 충족하지 않는 판독체를 여과 제거하고; c) 판독체를 여과 제거한 후에, 단계 (a)로부터 얻은 서열 판독체를 참조 서열에 맵핑하고; d) 참조 서열의 2개 이상의 미리 규정된 영역에서 맵핑된 판독체를 정량 또는 계수하고; e) (i) 미리 규정된 영역 내의 판독체의 수를 서로 및/또는 미리 규정된 영역 내의 특유한 서열 판독체의 수를 서로 정규화하고; (ii) 단계 (i)에서 얻은 정규화된 수를 대조 샘플로부터 얻은 정규화된 수와 비교함으로써 하나 이상의 미리 규정된 영역에서 카피수 변이를 결정하는 것을 포함하는, 카피수 변이를 검출하는 방법을 제공한다.

[0090] 본 개시내용은 또한 a) 대상체로부터의 신체 샘플로부터의 세포외 폴리뉴클레오티드를 서열분석하고, 여기서 각각의 세포외 폴리뉴클레오티드는 다수의 서열분석 판독체를 생성하고; b) 풍부화가 수행되지 않을 경우 영역에 대한 다중 서열분석 또는 전체-게놈 서열분석을 수행하고; c) 설정된 역치를 충족하지 않는 판독체를 여과 제거하고; d) 서열분석으로부터 유래된 서열 판독체를 참조 서열 상에 맵핑하고; e) 각각의 맵핑가능한 염기 위치에서 참조 서열의 변이체와 정렬되는 맵핑된 서열 판독체의 하위세트를 확인하고; f) 각각의 맵핑가능한 염기 위치에 대해, (a) 참조 서열에 비해 변이체를 포함하는 맵핑된 서열 판독체의 수 대 (b) 각각의 맵핑가능한 염기 위치에 대한 총 서열 판독체의 수의 비를 계산하고; g) 각각의 맵핑가능한 염기 위치에 대한 변이의 비 또는 빈도를 정규화하고 잠재적인 희귀 변이체(들) 또는 돌연변이(들)를 결정하고; h) 잠재적인 희귀 변이체(들) 또는 돌연변이(들)를 갖는 각각의 영역에 대해 생성된 수를 참조 샘플로부터 유사하게 유래된 수와 비교하는 것을 포함하는, 대상체로부터 얻은 세포 유리 또는 실질적인 세포 유리 샘플에서 희귀 돌연변이를 검출하는 방법을 제공한다.

[0091] 본 개시내용은 또한 대상체에서 세포외 폴리뉴클레오티드의 유전자 프로파일을 생성하는 것을 포함하는, 대상체에서 비정상적인 상태의 비균질성을 특성화하는 방법을 제공하고, 여기서 유전자 프로파일은 카피수 변이 및 희귀 돌연변이 분석에 의해 생성된 다수의 데이터를 포함한다.

[0092] 일부 실시양태에서, 대상체에서 확인된 각각의 희귀 변이체의 출현율/농도는 동시에 보고되고 정량된다. 일부 실시양태에서, 대상체 내의 희귀 변이체의 출현율/농도에 관한 신뢰도 점수가 보고된다.

[0093] 일부 실시양태에서, 세포외 폴리뉴클레오티드는 DNA를 포함한다. 일부 실시양태에서, 세포외 폴리뉴클레오티드는 RNA를 포함한다.

[0094] 일부 실시양태에서, 방법은 신체 샘플로부터 세포외 폴리뉴클레오티드의 단리를 추가로 포함한다. 일부 실시양태에서, 단리는 핵산 단리 및 추출을 순환시키는 방법을 포함한다. 일부 실시양태에서, 방법은 상기 단리된 세포외 폴리뉴클레오티드를 단편화하는 것을 추가로 포함한다. 일부 실시양태에서, 신체 샘플은 혈액, 혈장, 혈청, 소변, 타액, 점막 분비물, 객담, 대변 및 눈물로 이루어진 군으로부터 선택된다.

[0095] 일부 실시양태에서, 방법은 상기 신체 샘플에서 카피수 변이 또는 희귀 돌연변이 또는 변이체를 갖는 서열의 퍼센트를 결정하는 단계를 추가로 포함한다. 일부 실시양태에서, 결정은 미리 결정된 역치 초과 또는 미만의 폴리뉴클레오티드의 양을 갖는 미리 규정된 영역의 백분율을 계산하는 것을 포함한다.

[0096] 일부 실시양태에서, 대상체는 비정상적인 상태를 갖는 것으로 의심받는다. 일부 실시양태에서, 비정상적인 상태는 돌연변이, 희귀 돌연변이, 삽입-결실, 카피수 변이, 염기변환, 전위, 역위, 결실, 이수성, 부분적 이수성, 배수성, 염색체 불안정성, 염색체 구조 변경, 유전자 융합, 염색체 융합, 유전자 말단절단, 유전자 증폭, 유전자 중복, 염색체 병변, DNA 병변, 핵산 화학적 변형의 비정상적인 변화, 후성적 패턴의 비정상적인 변화, 핵산 메틸화 감염의 비정상적인 변화 및 암으로 이루어진 군으로부터 선택된다.

[0097] 일부 실시양태에서, 대상체는 임신한 여성이다. 일부 실시양태에서, 카피수 변이 또는 희귀 돌연변이 또는 유전자 변이체는 태아 비정상을 나타낸다. 일부 실시양태에서, 태아 비정상은 돌연변이, 희귀 돌연변이, 삽입-결실, 카피수 변이, 염기변환, 전위, 역위, 결실, 이수성, 부분적 이수성, 배수성, 염색체 불안정성, 염색체 구조 변경, 유전자 융합, 염색체 융합, 유전자 말단절단, 유전자 증폭, 유전자 중복, 염색체 병변, DNA 병변, 핵산 화학적 변형의 비정상적인 변화, 후성적 패턴의 비정상적인 변화, 핵산 메틸화 감염의 비정상적인 변화 및 암으로 이루어진 군으로부터 선택된다.

[0098] 일부 실시양태에서, 방법은 서열분석 전에 하나 이상의 바코드를 세포외 폴리뉴클레오티드 또는 그의 단편에 부착하는 것을 추가로 포함한다. 일부 실시양태에서, 서열분석 전에 세포외 폴리뉴클레오티드 또는 그의 단편에 부착된 각각의 바코드는 특유한 것이다. 일부 실시양태에서, 서열분석 전에 세포외 폴리뉴클레오티드 또는 그의 단편에 부착된 각각의 바코드는 특유한 것이 아니다.

[0099] 일부 실시양태에서, 방법은 서열분석 전에 대상체의 게놈 또는 트랜스크립톰으로부터의 영역을 선택적으로 풍부화하는 것을 추가로 포함한다. 일부 실시양태에서, 방법은 서열분석 전에 대상체의 게놈 또는 트랜스크립톰으로부터의 영역을 비-선택적으로 풍부화하는 것을 포함한다.

[0100] 일부 실시양태에서, 방법은 임의의 증폭 또는 풍부화 단계 전에 하나 이상의 바코드를 세포외 폴리뉴클레오티드 또는 그의 단편에 부착시키는 것을 추가로 포함한다. 일부 실시양태에서, 바코드는 폴리뉴클레오티드이다. 일부 실시양태에서, 바코드는 무작위 서열을 포함한다. 일부 실시양태에서, 바코드는 선택 영역으로부터 서열분석된 분자의 다양성과 조합되어 특유한 분자의 확인을 가능하게 하는 올리고뉴클레오티드의 고정 또는 준-무작위 세트를 포함한다. 일부 실시양태에서, 바코드는 적어도 3, 5, 10, 15, 20, 25, 30, 35, 40, 45 또는 50량체 염기쌍 길이의 올리고뉴클레오티드를 포함한다.

[0101] 일부 실시양태에서, 방법은 세포외 폴리뉴클레오티드 또는 그의 단편을 증폭시키는 것을 추가로 포함한다. 일부 실시양태에서, 증폭은 포괄적 증폭 또는 전체 게놈 증폭을 포함한다. 일부 실시양태에서, 증폭은 선택적 증폭을 포함한다. 일부 실시양태에서, 증폭은 비-선택적 증폭을 포함한다. 일부 실시양태에서, 억제 증폭 또는 차감 풍부화가 수행된다.

[0102] 일부 실시양태에서, 특유한 정체의 서열 판독체는 서열 판독체의 개시 (출발) 및 종료 (정지) 영역에서의 서열 정보 및 서열 판독체의 길이를 기초로 하여 검출된다. 일부 실시양태에서, 특유한 정체의 서열 분자는 서열 판독체의 개시 (출발) 및 종료 (정지) 영역에서의 서열 정보, 서열 판독체의 길이 및 바코드의 부착을 기초로 하여 검출된다.

[0103] 일부 실시양태에서, 방법은 판독체의 정량 또는 계수 전에 추가의 분석으로부터의 판독체의 하위세트를 제거하는 것을 추가로 포함한다. 일부 실시양태에서, 제거는 역치, 예를 들어 90%, 99%, 99.9% 또는 99.99% 미만의 정확도 또는 품질 점수 및/또는 역치, 예를 들어 90%, 99%, 99.9% 또는 99.99% 미만의 맵핑 점수를 갖는 판독체를 여과 제거하는 것을 포함한다. 일부 실시양태에서, 방법은 설정된 역치보다 낮은 품질 점수를 갖는 판독체를 여과 제거하는 것을 추가로 포함한다.

[0104] 일부 실시양태에서, 미리 규정된 영역은 균일한 또는 실질적으로 균일한 크기를 갖는다. 일부 실시양태에서, 미리 규정된 영역은 적어도 약 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb 또는 100 kb

의 크기이다.

[0105] 일부 실시양태에서, 적어도 50, 100, 200, 500, 1000, 2000, 5000, 10,000, 20,000 또는 50,000개의 영역이 분석된다.

[0106] 일부 실시양태에서, 변이체는 유전자 융합, 유전자 중복, 유전자 결실, 유전자 전위, 미소부수체 영역, 유전자 단편 또는 이들의 조합으로 이루어진 군으로부터 선택된 게놈의 영역에서 발생한다. 일부 실시양태에서, 변이체는 유전자, 종양유전자, 종양 억제 유전자, 프로모터, 조절 서열 요소 또는 이들의 조합으로 이루어진 군으로부터 선택된 게놈의 영역에서 발생한다. 일부 실시양태에서, 변이체는 뉴클레오티드 변이체, 단일 염기 치환, 작은 삽입-결실, 염기변환, 전위, 역위, 결실, 말단절단 또는 유전자 말단절단 (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 또는 20개 뉴클레오티드 길이)이다.

[0107] 일부 실시양태에서, 방법은 개별 판독체의 바코드 또는 특유한 특성을 사용하여 맵핑된 판독체의 양을 보정/정규화/조정하는 것을 추가로 포함한다. 일부 실시양태에서, 판독체의 계수는 각각의 미리 규정된 영역 내의 특유한 바코드를 계수하고 그 수를 서열분석된 미리 규정된 영역의 적어도 하나의 하위세트에 걸쳐 정규화함으로써 수행된다.

[0108] 일부 실시양태에서, 동일한 대상체로부터 연속적인 시간 간격에서 채취한 샘플이 분석되고 이전의 샘플 결과와 비교된다. 일부 실시양태에서, 방법은 바코드-부착된 세포외 폴리뉴클레오티드를 증폭시키는 것을 추가로 포함한다. 일부 실시양태에서, 방법은 부분적인 카피수 변이 빈도의 결정, 이형접합성의 상실의 결정, 유전자 발현 분석의 수행, 후성적 분석의 수행 및/또는 과메틸화 분석의 수행을 추가로 포함한다.

[0109] 본 개시내용은 또한 대상체로부터 얻은 세포 유리 또는 실질적인 세포 유리 샘플에서 다중 서열분석을 사용하여 카피수 변이를 결정하거나 희귀 돌연변이 분석을 수행하는 것을 포함하는 방법을 제공한다.

[0110] 일부 실시양태에서, 다중 서열분석은 10,000회 초과의 서열분석 반응을 수행하는 것을 포함한다. 일부 실시양태에서, 다중 서열분석은 적어도 10,000개의 상이한 판독체를 동시에 서열분석하는 것을 포함한다. 일부 실시양태에서, 다중 서열분석은 게놈에 걸쳐 적어도 10,000개의 상이한 판독체에 대한 데이터 분석을 수행하는 것을 포함한다. 일부 실시양태에서, 정규화 및 검출은 하나 이상의 은닉 마르코프, 동적 프로그래밍, 서포트 벡터 머신, 베이지안 또는 확률적 모델링, 격자 해독, 비터비 해독, 기대값 최대화, 칼만 여과, 또는 신경망 방법을 사용하여 수행된다. 일부 실시양태에서, 방법은 질환 진행의 모니터링, 잔류 질환의 모니터링, 요법의 모니터링, 상태의 진단, 상태의 예측, 또는 대상체에 대해 발견된 변이체를 기초로 한 요법의 선택을 추가로 포함한다. 일부 실시양태에서, 요법은 가장 최근의 샘플 분석을 기초로 하여 변형된다. 일부 실시양태에서, 종양, 감염 또는 다른 조직 비정상의 유전자 프로파일이 추정된다.

[0111] 일부 실시양태에서, 종양, 감염 또는 다른 조직 비정상의 성장, 완화 또는 진행이 모니터링된다. 일부 실시양태에서, 대상체의 면역계에 관련된 서열은 한번에 또는 시간에 걸쳐 분석되고 모니터링된다. 몇몇 실시양태에서, 변이체의 확인은 확인된 변이체를 야기하는 것으로 의심되는 조직 비정상의 위치결정을 위한 영상화 시험 (예를 들어, CT, PET-CT, MRI, X-선, 초음파)을 통해 추적조사된다. 일부 실시양태에서, 분석은 동일한 환자로부터의 조직 또는 종양 생검으로부터 얻은 유전자 데이터의 사용을 추가로 포함한다. 일부 실시양태에서, 종양, 감염 또는 다른 조직 비정상의 계통발생학이 추정된다. 일부 실시양태에서, 방법은 집단-기반 노-콜링의 수행 및 낮은-신뢰도 영역의 확인을 추가로 포함한다. 일부 실시양태에서, 서열 적용범위에 대한 측정 데이터를 얻는 것은 게놈의 모든 위치에서 서열 적용범위 깊이를 측정하는 것을 포함한다. 일부 실시양태에서, 서열 적용범위 편향에 대한 측정 데이터를 보정하는 것은 윈도우-평균된 적용범위의 계산을 포함한다. 일부 실시양태에서, 서열 적용범위 편향에 대한 측정 데이터를 보정하는 것은 라이브러리 구축 및 서열분석 과정에서 GC 편향을 설명하기 위해 조정을 수행하는 것을 포함한다. 일부 실시양태에서, 서열 적용범위 편향에 대한 측정 데이터를 보정하는 것은 편향을 보상하기 위해 개별 맵핑과 연관된 추가의 가중 인자를 기초로 하여 조정을 수행하는 것을 포함한다.

[0112] 일부 실시양태에서, 세포외 폴리뉴클레오티드는 이환된 세포 기원으로부터 유래된다. 일부 실시양태에서, 세포외 폴리뉴클레오티드는 건강한 세포 기원으로부터 유래된다.

[0113] 본 개시내용은 또한 다음 단계를 수행하기 위한 컴퓨터 판독가능 매체를 포함하는 시스템을 제공한다: 게놈 내의 미리 규정된 영역을 선택하고; 미리 규정된 영역 내의 서열 판독체의 수를 계수하고; 서열 판독체의 수를 미리 규정된 영역에 걸쳐 정규화하고; 미리 규정된 영역 내의 카피수 변이의 퍼센트를 결정하는 것.

[0114] 일부 실시양태에서, 전체 게놈 또는 적어도 85%의 게놈이 분석된다. 일부 실시양태에서, 컴퓨터 판독가능 매체

는 혈장 또는 혈청 내의 암 DNA 또는 RNA 퍼센트에 대한 데이터를 최종 사용자에게 제공한다. 일부 실시양태에서, 확인된 카피수 변이체는 샘플 내의 비균질성 때문에 분수 (즉, 비-정수 수준)이다. 일부 실시양태에서, 선택된 영역의 풍부화가 수행된다. 일부 실시양태에서, 카피수 변이 정보는 본원에서 설명되는 방법을 기초로 하여 동시에 추출된다. 일부 실시양태에서, 방법은 샘플 내의 폴리뉴클레오티드의 출발 초기 카피의 수 또는 다양성을 제한하기 위해 폴리뉴클레오티드 병목현상화의 초기 단계를 포함한다.

[0115] 본 개시내용은 또한 a) 대상체의 신체 샘플로부터의 세포외 폴리뉴클레오티드를 서열분석하고, 여기서 각각의 세포외 폴리뉴클레오티드는 다수의 서열분석 판독체를 생성하고; b) 세트 품질 역치를 충족하지 않는 판독체를 여과 제거하고; c) 서열분석으로부터 유래된 서열 판독체를 참조 서열 상에 맵핑하고; d) 각각의 맵핑가능한 염기 위치에서 참조 서열의 변이체와 정렬되는 맵핑된 서열 판독체의 하위세트를 확인하고; e) 각각의 맵핑가능한 염기 위치에 대해, (a) 참조 서열에 비해 변이체를 포함하는 맵핑된 서열 판독체의 수 대 (b) 각각의 맵핑가능한 염기 위치에 대한 총 서열 판독체의 수의 비를 계산하고; f) 각각의 맵핑가능한 염기 위치에 대한 변이의 비 또는 빈도를 정규화하고 잠재적인 희귀 변이체(들) 또는 다른 유전자 변경(들)을 결정하고; g) 잠재적인 희귀 변이체(들) 또는 돌연변이(들)를 갖는 각각의 영역에 대해 생성된 수를 참조 샘플로부터 유사하게 유래된 수와 비교하는 것을 포함하는, 대상체로부터 얻은 세포 유리 또는 실질적인 세포 유리 샘플에서 희귀 돌연변이를 검출하는 방법을 제공한다.

[0116] 본 개시내용은 또한 a) 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트를 제공하고, 태그부착된 모 폴리뉴클레오티드의 각각의 세트에 대해; b) 세트 내의 태그부착된 모 폴리뉴클레오티드를 증폭시켜 상응하는 증폭된 자손 폴리뉴클레오티드의 세트를 생산하고; c) 증폭된 자손 폴리뉴클레오티드의 세트의 하위세트 (적절한 하위세트 포함)를 서열분석하여 서열분석 판독체의 세트를 생산하고; d) 서열분석 판독체의 세트를 붕괴시켜, 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성한다.

[0117] 일부 실시양태에서, 세트 내의 각각의 폴리뉴클레오티드는 참조 서열에 맵핑가능하다. 일부 실시양태에서, 방법은 태그부착된 모 폴리뉴클레오티드의 다수의 세트를 제공하는 것을 포함하고, 여기서 각각의 세트는 참조 서열 내의 상이한 맵핑가능 위치에 맵핑가능하다. 일부 실시양태에서, 방법은 e) 별개로 또는 조합으로 태그부착된 모 분자의 각각의 세트에 대해 컨센서스 서열의 세트를 분석하는 것을 추가로 포함한다. 일부 실시양태에서, 방법은 초기 출발 유전 물질을 태그부착된 모 폴리뉴클레오티드로 전환하는 것을 추가로 포함한다. 일부 실시양태에서, 초기 출발 유전 물질은 100 ng 이하의 폴리뉴클레오티드를 포함한다. 일부 실시양태에서, 이 방법은 전환 전에 초기 출발 유전 물질의 병목현상화를 포함한다. 일부 실시양태에서, 방법은 적어도 10%, 적어도 20%, 적어도 30%, 적어도 40%, 적어도 50%, 적어도 60%, 적어도 80% 또는 적어도 90%의 전환 효율로 초기 출발 유전 물질을 태그부착된 모 폴리뉴클레오티드로 전환하는 것을 포함한다. 일부 실시양태에서, 전환은 임의의 평활-말단 라이게이션, 점착성 말단 라이게이션, 분자 역위 프로브, PCR, 라이게이션-기반 PCR, 단일 가닥 라이게이션 및 단일 가닥 환형화를 포함한다. 일부 실시양태에서, 초기 출발 유전 물질은 세포 유리 핵산이다. 일부 실시양태에서, 다수의 세트는 동일한 게놈으로부터의 참조 서열 내의 상이한 맵핑가능 위치에 맵핑된다.

[0118] 일부 실시양태에서, 세트 내의 각각의 태그부착된 모 폴리뉴클레오티드는 특유하게 태그부착된다. 일부 실시양태에서, 모 폴리뉴클레오티드의 각각의 세트는 참조 서열 내의 위치에 맵핑가능하고, 각각의 세트 내의 폴리뉴클레오티드는 특유하게 태그부착되지 않은 것이다. 일부 실시양태에서, 컨센서스 서열의 생성은 태그로부터의 정보, 및/또는 (i) 서열 판독체의 개시 (출발) 영역에서의 서열 정보, (ii) 종료 (정지) 영역에서의 서열 정보 및 (iii) 서열 판독체의 길이 중 적어도 하나를 기초로 한다.

[0119] 일부 실시양태에서, 방법은 태그부착된 모 폴리뉴클레오티드의 세트 내의 적어도 20%, 적어도 30%, 적어도 40%, 적어도 50%, 적어도 60%, 적어도 70%, 적어도 80%, 적어도 90%, 적어도 95%, 적어도 98%, 적어도 99%, 적어도 99.9% 또는 적어도 99.99%의 특유한 폴리뉴클레오티드 각각으로부터의 적어도 하나의 자손에 대한 서열 판독체를 생산하기에 충분한 증폭된 자손 폴리뉴클레오티드의 세트의 하위세트를 서열분석하는 것을 포함한다. 일부 실시양태에서, 적어도 하나의 자손은 다수의 자손, 예를 들어 적어도 2, 적어도 5 또는 적어도 10개의 자손이다. 일부 실시양태에서, 서열 판독체의 세트 내의 서열 판독체의 수는 태그부착된 모 폴리뉴클레오티드의 세트 내의 특유한 태그부착된 모 폴리뉴클레오티드의 수보다 더 크다. 일부 실시양태에서, 서열분석된 증폭된 자손 폴리뉴클레오티드의 세트의 하위세트는 사용되는 서열분석 플랫폼의 염기당 서열분석 오류 비율 백분율과 동일한 백분율로 태그부착된 모 폴리뉴클레오티드의 세트에 나타나는 임의의 뉴클레오티드 서열이 컨센서스 서열의 세트 중에서 나타날 가능성이 적어도 50%, 적어도 60%, 적어도 70%, 적어도 80%, 적어도 90%, 적어도

95%, 적어도 98%, 적어도 99%, 적어도 99.9% 또는 적어도 99.99%가 되도록 하기에 충분한 크기를 갖는다.

[0120] 일부 실시양태에서, 방법은 (i) 태그부착된 모 폴리뉴클레오티드로 전환되는 초기 출발 유전 물질로부터의 서열의 선택적 증폭; (ii) 태그부착된 모 폴리뉴클레오티드의 선택적 증폭; (iii) 증폭된 자손 폴리뉴클레오티드의 선택적 서열 포획; 또는 (iv) 초기 출발 유전 물질의 선택적 서열 포획에 의해, 참조 서열 내의 하나 이상의 선택된 맵핑가능 위치에 맵핑되는 폴리뉴클레오티드에 대한 증폭된 자손 폴리뉴클레오티드의 세트를 풍부화하는 것을 포함한다.

[0121] 일부 실시양태에서, 분석은 컨센서스 서열의 세트로부터 얻은 측정치 (예를 들어, 수)를 대조 샘플로부터의 컨센서스 서열의 세트로부터 얻은 측정치에 대해 정규화하는 것을 포함한다. 일부 실시양태에서, 분석은 돌연변이, 희귀 돌연변이, 삽입-결실, 카피수 변이, 염기변환, 전위, 역위, 결실, 이수성, 부분적 이수성, 배수성, 염색체 불안정성, 염색체 구조 변경, 유전자 융합, 염색체 융합, 유전자 말단절단, 유전자 증폭, 유전자 중복, 염색체 병변, DNA 병변, 핵산 화학적 변형의 비정상적인 변화, 후성적 패턴의 비정상적인 변화, 핵산 메틸화 감염의 비정상적인 변화 또는 암의 검출을 포함한다.

[0122] 일부 실시양태에서, 폴리뉴클레오티드는 DNA, RNA, 이 둘의 조합, 또는 DNA + RNA-유래 cDNA를 포함한다. 일부 실시양태에서, 폴리뉴클레오티드의 특정 하위세트는 폴리뉴클레오티드의 초기 세트로부터의 또는 증폭된 폴리뉴클레오티드로부터의 염기쌍의 폴리뉴클레오티드 길이를 기초로 하여 선택되거나 풍부화된다. 일부 실시양태에서, 분석은 개체 내의 비정상 또는 질환, 예컨대 감염 및/또는 암의 검출 및 모니터링을 추가로 포함한다. 일부 실시양태에서, 방법은 면역 레퍼토리 프로파일링과 조합하여 수행된다. 일부 실시양태에서, 폴리뉴클레오티드는 혈액, 혈장, 혈청, 소변, 타액, 점막 분비물, 객담, 대변 및 눈물로 이루어진 군으로부터 선택된 샘플로부터 추출된다. 일부 실시양태에서, 붕괴는 태그부착된 모 폴리뉴클레오티드 또는 증폭된 자손 폴리뉴클레오티드의 센스 또는 안티센스 가닥에 존재하는 오류, 닉 또는 병변의 검출 및/또는 보정을 포함한다.

[0123] 본 개시내용은 또한 적어도 5%, 적어도 1%, 적어도 0.5%, 적어도 0.1% 또는 적어도 0.05%의 감도로 비-특유하게 태그부착된 초기 출발 유전 물질 내의 유전자 변이를 검출하는 것을 포함하는 방법을 제공한다.

[0124] 일부 실시양태에서, 초기 출발 유전 물질은 100 ng 미만 양의 핵산으로 제공되고, 유전자 변이는 카피수/이형접합성 변이이고, 검출은 하위-염색체 해상도; 예를 들어, 적어도 100 메가염기 해상도, 적어도 10 메가염기 해상도, 적어도 1 메가염기 해상도, 적어도 100 킬로염기 해상도, 적어도 10 킬로염기 해상도 또는 적어도 1 킬로염기 해상도로 수행된다. 일부 실시양태에서, 방법은 태그부착된 모 폴리뉴클레오티드의 다수의 세트를 제공하는 것을 포함하고, 여기서 각각의 세트는 참조 서열 내의 상이한 맵핑가능 위치에 맵핑가능하다. 일부 실시양태에서, 참조 서열 내의 맵핑가능 위치는 종양 마커의 유전자좌이고, 분석은 컨센서스 서열의 세트 내의 종양 마커를 검출하는 것을 포함한다.

[0125] 일부 실시양태에서, 종양 마커는 증폭 단계에서 도입되는 오류 비율보다 낮은 빈도로 컨센서스 서열의 세트에 존재한다. 일부 실시양태에서, 적어도 하나의 세트는 다수의 세트이고, 참조 서열의 맵핑가능 위치는 참조 서열 내의 다수의 맵핑가능 위치를 포함하고, 각각의 맵핑가능 위치는 종양 마커의 유전자좌이다. 일부 실시양태에서, 분석은 모 폴리뉴클레오티드의 적어도 2개의 세트 사이의 컨센서스 서열의 카피수 변이를 검출하는 것을 포함한다. 일부 실시양태에서, 분석은 참조 서열에 비해 서열 변이의 존재를 검출하는 것을 포함한다.

[0126] 일부 실시양태에서, 분석은 참조 서열에 비해 서열 변이의 존재를 검출하고 모 폴리뉴클레오티드의 적어도 2개의 세트 사이의 컨센서스 서열의 카피수 변이를 검출하는 것을 포함한다. 일부 실시양태에서, 붕괴는 (i) 증폭된 자손 폴리뉴클레오티드로부터 서열분석된 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 동일한 태그부착된 모 폴리뉴클레오티드로부터 증폭된 것이고; (ii) 패밀리 내의 서열 판독체를 기초로 하여 컨센서스 서열을 결정하는 것을 포함한다.

[0127] 본 개시내용은 또한 다음 단계를 수행하기 위한 컴퓨터 판독가능 매체를 포함하는 시스템을 제공한다: a) 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트를 수용하고, 태그부착된 모 폴리뉴클레오티드의 각각의 세트에 대해; b) 세트 내의 태그부착된 모 폴리뉴클레오티드를 증폭시켜 상응하는 증폭된 자손 폴리뉴클레오티드의 세트를 생산하고; c) 증폭된 자손 폴리뉴클레오티드의 세트의 하위세트 (적절한 하위세트 포함)를 서열분석하여 서열분석 판독체의 세트를 생산하고; d) 서열분석 판독체의 세트를 붕괴시켜, 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성하고, 임의로, e) 태그부착된 모 분자의 각각의 세트에 대해 컨센서스 서열의 세트를 분석하는 단계.

[0128] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 또는 유전자 변이의 양을 검출하는 것을 포함하는

방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 10%의 개체의 게놈이 서열분석된다.

[0129] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 또는 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 20%의 개체의 게놈이 서열분석된다.

[0130] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 또는 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 30%의 개체의 게놈이 서열분석된다.

[0131] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 또는 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 40%의 개체의 게놈이 서열분석된다.

[0132] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 또는 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 50%의 개체의 게놈이 서열분석된다.

[0133] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 또는 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 60%의 개체의 게놈이 서열분석된다.

[0134] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 또는 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 70%의 개체의 게놈이 서열분석된다.

[0135] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 또는 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 80%의 개체의 게놈이 서열분석된다.

[0136] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 또는 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 90%의 개체의 게놈이 서열분석된다.

[0137] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 및 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 10%의 개체의 게놈이 서열분석된다.

[0138] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 및 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 20%의 개체의 게놈이 서열분석된다.

[0139] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 및 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 30%의 개체의 게놈이 서열분석된다.

[0140] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 및 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 40%의 개체의 게놈이 서열분석된다.

[0141] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 및 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 50%의 개체의 게놈이 서열분석된다.

[0142] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 및 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 60%의 개체의 게놈이 서열분석된다.

[0143] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 및 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 70%의 개체의 게놈이 서열분석된다.

[0144] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 및 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 80%의 개체의 게놈이 서열분석된다.

[0145] 본 개시내용은 또한 개체에서 유전자 변경의 존재 또는 부재 및 유전자 변이의 양을 검출하는 것을 포함하는 방법을 제공하고, 여기서 검출은 세포 유리 핵산의 서열분석의 도움을 받아 수행되고, 적어도 90%의 개체의 게놈이 서열분석된다.

[0146] 일부 실시양태에서, 유전자 변경은 카피수 변이 또는 하나 이상의 희귀 돌연변이이다. 일부 실시양태에서, 유전자 변이는 하나 이상의 원인 변이체 및 하나 이상의 다형성을 포함한다. 일부 실시양태에서, 개체에서 유전자 변경 및/또는 유전자 변이의 양은 알려진 질환을 가지는 하나 이상의 개체에서의 유전자 변경 및/또는 유전자 변이의 양과 비교될 수 있다. 일부 실시양태에서, 개체에서 유전자 변경 및/또는 유전자 변이의 양은 질환이 없는 하나 이상의 개체에서의 유전자 변경 및/또는 유전자 변이의 양과 비교될 수 있다. 일부 실시양태에서, 세포 유리 핵산은 DNA이다. 일부 실시양태에서, 세포 유리 핵산은 RNA이다. 일부 실시양태에서, 세포 유리 핵산은 DNA 및 RNA이다. 일부 실시양태에서, 질환은 암 또는 전암이다. 일부 실시양태에서, 방법은 질환의 진단 또는 치료를 추가로 포함한다.

[0147] 본 개시내용은 또한 a) 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트를 제공하고, 태그부착된 모 폴리뉴클레오티드의 각각의 세트에 대해; b) 세트 내의 태그부착된 모 폴리뉴클레오티드를 증폭시켜 상응하는 증폭된 자손 폴리뉴클레오티드의 세트를 생산하고; c) 증폭된 자손 폴리뉴클레오티드의 세트의 하위세트 (적절한 하위세트 포함)를 서열분석하여 서열분석 판독체의 세트를 생산하고; d) 서열분석 판독체의 세트를 붕괴시켜, 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성하고; e) 컨센서스 서열 중에서 품질 역치를 충족하지 않는 것을 여과 제거하는 것을 포함하는 방법을 제공한다.

[0148] 일부 실시양태에서, 품질 역치는 컨센서스 서열로 붕괴되는 증폭된 자손 폴리뉴클레오티드로부터의 서열 판독체의 수를 고려한다. 일부 실시양태에서, 품질 역치는 컨센서스 서열로 붕괴되는 증폭된 자손 폴리뉴클레오티드로부터의 서열 판독체의 수를 고려한다.

[0149] 본 개시내용은 또한 본원에서 설명되는 방법을 수행하기 위한 컴퓨터 판독가능 매체를 포함하는 시스템을 제공한다.

[0150] 본 개시내용은 또한 a) 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트를 제공하고, 여기서 각각의 세트는 하나 이상의 게놈 내의 상이한 참조 서열에 맵핑되고, 태그부착된 모 폴리뉴클레오티드의 각각의 세트에 대해; i) 제1 폴리뉴클레오티드를 증폭시켜 증폭된 폴리뉴클레오티드의 세트를 생산하고; ii) 증폭된 폴리뉴클레오티드의 세트의 하위세트를 서열분석하여 서열분석 판독체의 세트를 생산하고; iii) (1) 증폭된 자손 폴리뉴클레오티드로부터 서열분석된 서열 판독체를, 각각의 패밀리가 동일한 태그부착된 모 폴리뉴클레오티드로부터 증폭된 것인 패밀리로 분류하여 서열 판독체를 붕괴시키는 것을 포함하는 방법을 제공한다.

[0151] 일부 실시양태에서, 붕괴는 각각의 패밀리 내의 서열 판독체의 정량적 척도를 결정하는 것을 추가로 포함한다. 일부 실시양태에서, 방법은 a) 특유한 패밀리의 정량적 척도를 결정하고; b) (1) 특유한 패밀리의 정량적 척도 및 (2) 각각의 군 내의 서열 판독체의 정량적 척도를 기초로 하여, 세트 내의 특유한 태그부착된 모 폴리뉴클레오티드의 척도를 추정하는 것을 추가로 포함한다. 일부 실시양태에서, 추정은 통계적 또는 확률적 모델을 사용하여 수행된다. 일부 실시양태에서, 적어도 하나의 세트는 다수의 세트이다. 일부 실시양태에서, 방법은 2개의 세트 사이의 증폭 또는 표상적 편향에 대한 보정을 추가로 포함한다. 일부 실시양태에서, 방법은 2개의 세트 사이의 증폭 또는 표상적 편향을 보정하기 위해 대조군 또는 대조 샘플의 세트를 사용하는 것을 추가로 포함한다. 일부 실시양태에서, 방법은 세트 사이의 카피수 변이를 결정하는 것을 추가로 포함한다.

[0152] 일부 실시양태에서, 방법은 d) 패밀리 중에서 다형체 형태의 정량적 척도를 결정하고; e) 결정된 다형체 형태의 정량적 척도를 기초로 하여, 다형체 형태의 정량적 척도를 추정된 특유한 태그부착된 모 폴리뉴클레오티드의 수로 추정하는 것을 추가로 포함한다. 일부 실시양태에서, 다형체 형태는 치환, 삽입, 결실, 역위, 미소부수체 변화, 염기변환, 전위, 융합, 메틸화, 과메틸화, 히드록시메틸화, 아세틸화, 후성적 변이체, 조절-연관 변이체

또는 단백질 결합 부위를 포함하나 이에 제한되지는 않는다.

[0153] 일부 실시양태에서, 세트는 공통 샘플로부터 유래하고, 방법은 d) 참조 서열 내의 각각의 다수의 맵핑가능 위치에 맵핑되는 각각의 세트 내의 태그부착된 모 폴리뉴클레오티드의 추정된 수의 비교를 기초로 하여 다수의 세트에 대한 카피수 변이를 추정하는 것을 추가로 포함한다. 일부 실시양태에서, 각각의 세트 내의 폴리뉴클레오티드의 본래의 수가 추가로 추정된다. 일부 실시양태에서, 각각의 세트 내의 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 하위세트는 비-특유하게 태그부착된다.

[0154] 본 개시내용은 또한 폴리뉴클레오티드를 포함하는 샘플 내의 카피수 변이를 결정하는 방법을 제공하고, 이 방법은 a) 제1 폴리뉴클레오티드의 적어도 2개의 세트를 제공하고, 여기서 각각의 세트는 게놈 내의 참조 서열 내의 상이한 맵핑가능 위치에 맵핑되고, 제1 폴리뉴클레오티드의 각각의 세트에 대해; (i) 폴리뉴클레오티드를 증폭시켜 증폭된 폴리뉴클레오티드의 세트를 생산하고; (ii) 증폭된 폴리뉴클레오티드의 세트의 하위세트를 서열분석하여 서열분석 판독체의 세트를 생산하고; (iii) 증폭된 폴리뉴클레오티드로부터 서열분석된 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 세트 내의 동일한 제1 폴리뉴클레오티드로부터 증폭된 것이고; (iv) 세트 내의 패밀리의 정량적 척도를 추정하고; b) 각각의 세트 내의 패밀리의 정량적 척도를 비교함으로써 카피수 변이를 결정하는 것을 포함한다.

[0155] 본 개시내용은 또한 a) 제1 폴리뉴클레오티드의 적어도 하나의 세트를 제공하고, 여기서 각각의 세트는 하나 이상의 게놈 내의 참조 서열 내의 상이한 맵핑가능 위치에 맵핑되고, 제1 폴리뉴클레오티드의 각각의 세트에 대해; (i) 제1 폴리뉴클레오티드를 증폭시켜 증폭된 폴리뉴클레오티드의 세트를 생산하고; (ii) 증폭된 폴리뉴클레오티드의 세트의 하위세트를 서열분석하여 서열분석 판독체의 세트를 생산하고; (iii) 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 동일한 제1 폴리뉴클레오티드로부터 증폭된 폴리뉴클레오티드의 서열 판독체를 포함하고; b) 제1 폴리뉴클레오티드의 각각의 세트에 대해, 제1 폴리뉴클레오티드의 세트 내의 하나 이상의 염기에 대한 콜 빈도를 추정하고, 여기서 추정은 (i) 각각의 패밀리에 대해, 각각의 다수의 콜에 대한 신뢰도 점수를 배정하고, 신뢰도 점수는 패밀리의 구성원 중에서 콜의 빈도를 고려한 것이고; (ii) 각각의 패밀리에 배정된 하나 이상의 콜의 신뢰도 점수를 고려하여 하나 이상의 콜의 빈도를 평가하는 것을 포함하는 것임을 포함하는, 폴리뉴클레오티드의 샘플 내의 서열 콜의 빈도를 추정하는 방법을 제공한다.

[0156] 본 개시내용은 또한 a) 적어도 하나의 개별 폴리뉴클레오티드 분자를 제공하고; b) 적어도 하나의 개별 폴리뉴클레오티드 분자 내의 서열 정보를 암호화하여 신호를 생산하고; c) 적어도 일부의 신호를 채널에 통과시켜 적어도 하나의 개별 폴리뉴클레오티드 분자에 대한 뉴클레오티드 서열 정보를 포함하는 수신된 신호를 생산하고, 여기서 수신된 신호는 노이즈 및/또는 왜곡을 포함하고; d) 수신된 신호를 해독하여 적어도 하나의 개별 폴리뉴클레오티드 분자에 대한 서열 정보를 포함하는 메세지를 생산하고, 여기서 해독은 메세지 내의 각각의 개별 폴리뉴클레오티드에 대한 노이즈 및/또는 왜곡을 감소시키고; e) 적어도 하나의 개별 폴리뉴클레오티드 분자에 대한 서열 정보를 포함하는 메세지를 수신자에게 제공하는 것을 포함하는, 적어도 하나의 개별 폴리뉴클레오티드 분자에 대한 서열 정보를 통신하는 방법을 제공한다.

[0157] 일부 실시양태에서, 노이즈는 부정확한 뉴클레오티드 콜을 포함한다. 일부 실시양태에서, 왜곡은 다른 개별 폴리뉴클레오티드 분자에 비해 개별 폴리뉴클레오티드 분자의 불균등한 증폭을 포함한다. 일부 실시양태에서, 왜곡은 증폭 또는 서열분석 편향에 의해 발생한다. 일부 실시양태에서, 적어도 하나의 개별 폴리뉴클레오티드 분자는 다수의 개별 폴리뉴클레오티드 분자이고, 해독은 다수의 각각의 분자에 대한 메세지를 생산한다. 일부 실시양태에서, 암호화는 임의로 태그부착된 적어도 하나의 개별 폴리뉴클레오티드 분자를 증폭시키는 것을 포함하고, 신호는 증폭된 분자의 수집물을 포함한다. 일부 실시양태에서, 채널은 폴리뉴클레오티드 서열분석기를 포함하고, 수신된 신호는 적어도 하나의 개별 폴리뉴클레오티드 분자로부터 증폭된 다수의 폴리뉴클레오티드의 서열 판독체를 포함한다. 일부 실시양태에서, 해독은 적어도 하나의 개별 폴리뉴클레오티드 분자 각각으로부터 증폭된 분자의 서열 판독체를 분류하는 것을 포함한다. 일부 실시양태에서, 해독은 생성된 서열 신호를 여과하는 확률적 또는 통계적 방법으로 이루어진다.

[0158] 일부 실시양태에서, 폴리뉴클레오티드는 종양 게놈 DNA 또는 RNA로부터 유래된다. 일부 실시양태에서, 폴리뉴클레오티드는 세포 유리 폴리뉴클레오티드, 엑소솜 폴리뉴클레오티드, 박테리아 폴리뉴클레오티드 또는 바이러스 폴리뉴클레오티드로부터 유래된다. 본원의 임의의 방법의 일부 실시양태에서, 방법은 영향받는 분자 경로의 검출 및/또는 연관분석을 추가로 포함한다. 본원의 임의의 방법의 일부 실시양태에서, 방법은 개체의 건강 또는 질환 상태의 순차적인 모니터링을 추가로 포함한다. 일부 실시양태에서, 개체 내의 질환과 연관된 게놈의 계통발생이 추정된다. 일부 실시양태에서, 본원에서 설명되는 임의의 방법은 질환의 진단, 모니터링 또는 치료

를 추가로 포함한다. 일부 실시양태에서, 치료 요법은 다형체 형태 또는 CNV 또는 연관 경로를 기초로 하여 선택되거나 변형된다. 일부 실시양태에서, 치료는 조합 요법을 포함한다. 일부 실시양태에서, 진단은 방사선사진 기술, 예컨대 CT-스캔, PET-CT, MRI, 초음파, 미세기포를 사용한 초음파 등을 사용하여 질환의 위치를 결정하는 것을 추가로 포함한다.

[0159] 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에, 게놈 내의 미리 규정된 영역을 선택하고; 서열 판독체에 접근하여 미리 규정된 영역 내의 서열 판독체의 수를 계수하고; 서열 판독체의 수를 미리 규정된 영역에 걸쳐 정규화하고; 미리 규정된 영역 내의 카피수 변이의 퍼센트를 결정하는 단계를 포함하는 방법을 이행하는 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 제공한다.

[0160] 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에, 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고; 설정된 역치를 충족하지 않는 판독체를 여과 제거하고; 서열분석으로부터 유래된 서열 판독체를 참조 서열 상에 맵핑하고; 각각의 맵핑가능한 염기 위치에서 참조 서열의 변이체와 정렬되는 맵핑된 서열 판독체의 하위세트를 확인하고; 각각의 맵핑가능한 염기 위치에 대해, (a) 참조 서열에 비해 변이체를 포함하는 맵핑된 서열 판독체의 수 대 (b) 각각의 맵핑가능한 염기 위치에 대한 총 서열 판독체의 수의 비를 계산하고; 각각의 맵핑가능한 염기 위치에 대한 변이의 비 또는 빈도를 정규화하고 잠재적인 희귀 변이체(들) 또는 다른 유전자 변경(들)을 결정하고; 잠재적인 희귀 변이체(들) 또는 돌연변이(들)를 갖는 각각의 영역에 대해 생성된 수를 참조 샘플로부터 유사하게 유래된 수와 비교하는 것을 포함하는 방법을 이행하는 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 제공한다.

[0161] 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에, a) 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고; b) 서열분석 판독체의 세트를 붕괴시켜, 서열은 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성하는 것을 포함하는 방법을 이행하는 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 제공한다.

[0162] 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에, a) 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고; b) 서열분석 판독체의 세트를 붕괴시켜, 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성하고; c) 컨센서스 서열 중에서 품질 역치를 충족하지 않는 것을 여과 제거하는 것을 포함하는 방법을 이행하는 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 제공한다.

[0163] 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에, a) 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고; i) (1) 증폭된 자손 폴리뉴클레오티드로부터 서열분석된 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 동일한 태그부착된 모 폴리뉴클레오티드로부터 증폭된 것이고, 임의로, (2) 각각의 패밀리 내의 서열 판독체의 정량적 척도를 결정함으로써 서열 판독체를 붕괴시키는 것을 포함하는 방법을 이행하는 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 제공한다.

[0164] 일부 실시양태에서, 실행가능 코드는 컴퓨터 프로세서에 의한 실행시에, b) 특유한 패밀리의 정량적 척도를 결정하고; c) (1) 특유한 패밀리의 정량적 척도 및 (2) 각각의 군 내의 서열 판독체의 정량적 척도를 기초로 하여, 세트 내의 특유한 태그부착된 모 폴리뉴클레오티드의 척도를 추정하는 단계를 추가로 수행한다.

[0165] 일부 실시양태에서, 실행가능 코드는 컴퓨터 프로세서에 의한 실행시에,

[0166] d) 패밀리 중에서 다형체 형태의 정량적 척도를 결정하고; e) 결정된 다형체 형태의 정량적 척도를 기초로 하여, 다형체 형태의 정량적 척도를 추정된 특유한 태그부착된 모 폴리뉴클레오티드의 수로 추정하는 단계를 추가로 수행한다.

[0167] 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에, a) 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고, 증폭된 폴리뉴클레오티드로부터 서열분석된 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 세트 내의 동일한 제1 폴리뉴클레오티드로부터 증폭된 것이고; b) 세트 내의 패밀리의 정량적 척도를 추정하고; c) 각각의 세트 내의 패밀리의 정량적 척도를 비교함으로써 카피수 변이를 결정하는 것을 포함하는 방법을 이행하는 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 제공한다.

[0168] 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에, a) 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고, 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 동일한 제1 폴리뉴클레오티드로부터 증폭된 폴리뉴클레오티드의 서열 판독체를 포함하고; b) 제1 폴리뉴클레오티드의 각각의 세트에 대해, 제1 폴리뉴클레오티드의 세트 내의 하나 이상의 염기에 대한 콜 빈도를 추정하고, 여기서 추정은 c) 각각의 패밀리에 대해, 각각의 다수의 콜에 대한 신뢰도 점수를 배정하고, 신뢰도 점수는 패밀리의 구성원 중에서 콜의 빈도를 고려한 것이고; d) 각각의 패밀리에 배정된 하나 이상의 콜의 신뢰도 점수를 고려하여 하나 이상의 콜의 빈도를 평가하는 것을 포함하는 것임을 포함하는 방법을 이행하는 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 제공한다.

[0169] 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에, a) 적어도 하나의 개별 폴리뉴클레오티드 분자로부터의 암호화된 서열 정보를 포함하는 수신된 신호를 포함하는 데이터 파일에 접근하고, 여기서 수신된 신호는 노이즈 및/또는 왜곡을 포함하고; b) 수신된 신호를 해독하여 적어도 하나의 개별 폴리뉴클레오티드 분자에 대한 서열 정보를 포함하는 메세지를 생산하고, 여기서 해독은 메세지 내에서 각각의 개별 폴리뉴클레오티드에 대한 노이즈 및/또는 왜곡을 감소시키고; c) 적어도 하나의 개별 폴리뉴클레오티드 분자에 대한 서열 정보를 포함하는 메세지를 컴퓨터 파일에 기록하는 단계를 포함하는 방법을 이행하는 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 제공한다.

[0170] 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에, a) 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고; b) 서열분석 판독체의 세트를 붕괴시켜, 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성하고; c) 컨센서스 서열 중에서 품질 역치를 충족하지 않는 것을 여과 제거하는 것을 포함하는 방법을 이행하는 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 제공한다.

[0171] 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에, a) 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고; b) (i) 증폭된 자손 폴리뉴클레오티드로부터 서열분석된 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 동일한 태그부착된 모 폴리뉴클레오티드로부터 증폭된 것이고; (ii) 임의로, 각각의 패밀리 내의 서열 판독체의 정량적 척도를 결정함으로써 서열 판독체를 붕괴시키는 것을 포함하는 방법을 이행하는 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 제공한다.

[0172] 일부 실시양태에서, 실행가능 코드는 컴퓨터 프로세서에 의한 실행시에, d) 특유한 패밀리의 정량적 척도를 결정하고; e) (1) 특유한 패밀리의 정량적 척도 및 (2) 각각의 군 내의 서열 판독체의 정량적 척도를 기초로 하여, 세트 내의 특유한 태그부착된 모 폴리뉴클레오티드의 척도를 추정하는 단계를 추가로 수행한다.

[0173] 일부 실시양태에서, 실행가능 코드는 컴퓨터 프로세서에 의한 실행시에, e) 패밀리 중에서 다형체 형태의 정량적 척도를 결정하고; f) 결정된 다형체 형태의 정량적 척도를 기초로 하여, 다형체 형태의 정량적 척도를 추정된 특유한 태그부착된 모 폴리뉴클레오티드의 수로 추정하는 단계를 추가로 수행한다.

[0174] 일부 실시양태에서, 실행가능 코드는 컴퓨터 프로세서에 의한 실행시에, e) 각각의 다수의 참조 서열에 맵핑되는 각각의 세트 내의 태그부착된 모 폴리뉴클레오티드의 추정된 수의 비교를 기초로 하여 다수의 세트에 대한 카피수 변이를 추정하는 단계를 추가로 수행한다.

[0175] 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에, a) 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고; b) 증폭된 폴리뉴클레오티드로부터 서열분석된 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 세트 내의 동일한 제1 폴리뉴클레오티드로부터 증폭된 것이고; c) 세트 내의 패밀리의 정량적 척도를 추정하고; d) 각각의 세트 내의 패밀리의 정량적 척도를 비교함으로써 카피수 변이를 결정하는 것을 포함하는 방법을 이행하는 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 제공한다.

[0176] 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에, 다수의 서열분석 판독체를 포함하는 데이터 파일에 접근하고, 여기서 서열 판독체는 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트로부터 증폭된 자손 폴리뉴클레오티드의 세트로부터 유래하고, 서열 판독체를 패밀리로 분류하고, 각각의 패밀리는 동일한 제1 폴리뉴클레

오티드로부터 증폭된 폴리뉴클레오티드의 서열 판독체를 포함하고; 제1 폴리뉴클레오티드의 각각의 세트에 대해, 제1 폴리뉴클레오티드의 세트 내의 하나 이상의 염기에 대한 콜 빈도를 추정하고, 여기서 추정은 (i) 각각의 패밀리에 대해, 각각의 다수의 콜에 대한 신뢰도 점수를 배정하고, 신뢰도 점수는 패밀리의 구성원 중에서 콜의 빈도를 고려한 것이고; (ii) 각각의 패밀리에 배정된 하나 이상의 콜의 신뢰도 점수를 고려하여 하나 이상의 콜의 빈도를 평가하는 것을 포함하는 것임을 포함하는 방법을 이행하는 비-일시적인 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 제공한다.

[0177] 본 개시내용은 또한 cfDNA 폴리뉴클레오티드의 100 내지 100,000개의 인간 반수체 게놈 등가물을 포함하는 조성물을 제공하고, 여기서 폴리뉴클레오티드는 2 내지 1,000,000개의 특유한 식별자로 태그부착된다.

[0178] 일부 실시양태에서, 조성물은 cfDNA 폴리뉴클레오티드의 1000 내지 50,000개의 반수체 인간 게놈 등가물을 포함하고, 여기서 폴리뉴클레오티드는 2 내지 1,000개의 특유한 식별자로 태그부착된다. 일부 실시양태에서, 특유한 식별자는 뉴클레오티드 바코드를 포함한다. 본 개시내용은 또한 a) cfDNA 폴리뉴클레오티드의 100 내지 100,000개의 반수체 인간 게놈 등가물을 포함하는 샘플을 제공하고; b) 폴리뉴클레오티드를 2 내지 1,000,000개의 특유한 식별자로 태그부착하는 것을 포함하는 방법을 제공한다.

[0179] 본 개시내용은 또한 a) 단편화된 폴리뉴클레오티드의 다수의 인간 반수체 게놈 등가물을 포함하는 샘플을 제공하고; b) z를 결정하고, 여기서 z는 게놈 내의 임의의 위치에서 출발하는 예상된 수의 중복 폴리뉴클레오티드의 중심 경향도 (예를 들어, 평균, 중간값 또는 최빈값)의 척도이고, 중복 폴리뉴클레오티드는 동일한 출발 및 정지 위치를 갖고; c) 샘플 내의 폴리뉴클레오티드를 n개의 특유한 식별자로 태그부착하고, 여기서 n은 2 내지 100,000*z, 2 내지 10,000*z, 2 내지 1,000*z 또는 2 내지 100*z임을 포함하는 방법을 제공한다. 본 개시내용은 또한 a) 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 세트를 제공하고, 태그부착된 모 폴리뉴클레오티드의 각각의 세트에 대해; b) 세트 내의 각각의 태그부착된 모 폴리뉴클레오티드에 대한 다수의 서열 판독체를 생산하여 서열분석 판독체의 세트를 생산하고; c) 서열분석 판독체의 세트를 붕괴시켜, 태그부착된 모 폴리뉴클레오티드의 세트 중의 특유한 폴리뉴클레오티드에 각각 상응하는 컨센서스 서열의 세트를 생성하는 것을 포함하는 방법을 제공한다.

[0180] 본 개시내용은 또한 본원에서 설명되는 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 포함하는 시스템을 제공한다. 본 개시내용은 또한 컴퓨터 프로세서에 의한 실행시에 본원에서 설명되는 방법을 이행하는 기계-실행가능 코드를 포함하는 컴퓨터 판독가능 매체를 포함하는 시스템을 제공한다.

[0181] 본원의 추가의 측면 및 이점은 다음 상세한 설명으로부터 통상의 기술자에게 쉽게 명백해질 것이고, 여기서 단지 본원의 예시적인 실시양태이 제시되고 설명된다. 알 수 있는 바와 같이, 본원은 다른 및 상이한 실시양태로 실시될 수 있고, 그의 여러 상세한 내용은 모두 본 개시내용으로부터 벗어나지 않으면서 다양한 명백한 측면에서 변형될 수 있다. 따라서, 도면 및 상세한 설명은 제한하는 의미가 아니라, 단지 예시적인 것으로서 간주되어야 한다.

[0182] **문헌 원용**

[0183] 본 명세서에서 언급된 모든 공개문, 특허, 및 특허 출원은 각각의 개별 공개문, 특허, 또는 특허 출원이 구체적으로 및 개별적으로 참조로 포함되는 것으로 지시되는 것과 동일한 정도로 본원에 참조로 포함된다.

도면의 간단한 설명

[0184] 본 개시내용의 시스템 및 방법의 신규한 특징은 첨부된 청구항에 상세하게 제시된다. 본 개시내용의 시스템 및 방법의 원리가 이용되는 예시적 실시양태를 제시하는 다음 상세한 설명 및 첨부하는 도면을 참조하여 본 개시내용의 특징 및 이점에 대한 보다 양호한 이해가 이루어질 것이다:

도 1은 단일 샘플을 사용하는 카피수 변이의 검출의 방법의 흐름도 (flow chart representation)이다.

도 2는 짝을 이룬 (paired) 샘플을 사용하는 카피수 변이의 검출의 방법의 흐름도이다.

도 3은 희귀 돌연변이 (예를 들어, 단일 뉴클레오티드 변이체)의 검출 방법의 흐름도이다.

도 4a는 정상적인 비-암성 대상체로부터 생성된 그래프형 카피수 변이 검출 보고서이다.

도 4b는 전립선암에 걸린 대상체로부터 생성된 그래프형 카피수 변이 검출 보고서이다.

도 4c는 전립선암에 걸린 대상체의 카피수 변이 분석으로부터 생성된 보고서의 인터넷을 통한 접속의 모식도이

다.

도 5a는 전립선암 완화를 보이는 대상체로부터 생성된 그래프형 카피수 변이 검출 보고서이다.

도 5b는 재발 전립선암에 걸린 대상체로부터 생성된 그래프형 카피수 변이 검출 보고서이다.

도 6a는 MET 및 TP53의 야생형 및 돌연변이체 카피를 모두 함유하는 DNA 샘플을 사용하는 다양한 혼합 실험으로부터 생성된 그래프형 검출 보고서 (예를 들어, 단일 뉴클레오티드 변이체에 대한)이다.

도 6b는 (예를 들어, 단일 뉴클레오티드 변이체) 검출 결과의 로그 (logarithmic) 그래프이다. 관찰된 대 예상된 암 비율 측정치를, MET, HRAS 및 TP53의 야생형 및 돌연변이체 카피를 모두 함유하는 DNA 샘플을 사용하는 다양한 혼합 실험에 대해 보여준다.

도 7a는 참조물 (대조군)에 비해 전립선암에 걸린 대상체에서 2개의 유전자, 즉 PIK3CA 및 TP53에서 2개 (예를 들어, 단일 뉴클레오티드 변이체)의 비율에 대한 그래프형 보고서이다.

도 7b는 전립선암에 걸린 대상체의 (예를 들어, 단일 뉴클레오티드 변이체) 분석으로부터 생성된 보고서의 인터넷을 통한 접속의 모식도이다.

도 8은 유전 물질의 분석 방법의 흐름도이다.

도 9는 태그부착된 모 폴리뉴클레오티드의 세트에서 노이즈 및/또는 왜곡이 감소된 정보를 제시하기 위해 서열 판독체의 세트에서 정보를 해독하는 방법에 대한 흐름도이다.

도 10은 서열 판독체의 세트로부터 CNV의 결정에서 왜곡을 감소시키는 방법의 흐름도이다.

도 11은 서열 판독체의 세트로부터 태그부착된 모 폴리뉴클레오티드 집단 내의 유전자좌에서 염기 또는 염기 서열의 빈도를 평가하는 방법을 보여주는 흐름도이다.

도 12는 서열 정보를 통신하는 방법을 보여준다.

도 13은 표준 서열분석 및 디지털 서열분석 (Digital Sequencing) 작업흐름을 이용하여 0.3% LNCaP cfDNA 적정에서 전체 70 kb 패널 (panel)에 걸쳐 검출된 작은 대립유전자 빈도를 보여준다. 표준 "아날로그 (analog)" 서열분석 (도 13a)은 Q30 여과에도 불구하고 PCR 및 서열분석 오류에 의한 매우 큰 노이즈로 모든 진-양성 희귀 변이체를 차단한다. 디지털 서열분석 (도 13b)은 모든 PCR 및 서열분석 노이즈를 제거하고, 위-양성을 갖지 않는 진정한 돌연변이를 보여주고: 초록색 원은 정상 cfDNA에서 SNP 지점이고, 적색 원은 검출된 LNCaP 돌연변이이다.

도 14는 LNCap cfDNA의 적정을 보여준다.

도 15는 본원의 다양한 방법을 실행하기 위해 프로그래밍되거나 달리 설정된 컴퓨터 시스템을 보여준다.

발명을 실시하기 위한 구체적인 내용

[0185] **I. 전반적인 개요**

[0186] 본원은 세포 유리 폴리뉴클레오티드 내의 희귀 돌연변이 (예를 들어, 단일 또는 다중 뉴클레오티드 변이) 및 카피수 변이의 검출을 위한 시스템 및 방법을 제공한다. 일반적으로, 시스템 및 방법은 샘플 제조, 또는 체액으로부터 세포 유리 폴리뉴클레오티드 서열의 추출 및 단리; 관련 기술 분야에 공지된 기술에 의한 세포 유리 폴리뉴클레오티드의 후속적인 서열분석; 및 참조물에 비교하여 희귀 돌연변이 및 카피수 변이를 검출하기 위한 생물 정보공학 도구의 적용을 포함한다. 시스템 및 방법은 또한 질환의 희귀 돌연변이 (예를 들어, 단일 뉴클레오티드 변이 프로파일링), 카피수 변이 프로파일링 또는 일반적인 유전자 프로파일링의 검출을 도울 때 추가의 참조물로서 사용되는, 상이한 질환의 상이한 희귀 돌연변이 또는 카피수 변이 프로파일의 데이터베이스 또는 수집물을 포함할 수 있다.

[0187] 시스템 및 방법은 세포 유리 DNA의 분석에서 특히 유용할 수 있다. 일부 경우에, 세포 유리 DNA는 쉽게 접근가능한 체액, 예컨대 혈액으로부터 추출하고 단리한다. 예를 들어, 세포 유리 DNA는 이소프로판올 침전 및/또는 실리카 기반 정제를 포함하나 이에 제한되지는 않는 관련 기술 분야에 공지된 다양한 방법을 사용하여 추출될 수 있다. 세포 유리 DNA는 임의의 많은 대상체, 예컨대 암이 없는 대상체, 암 위험이 있는 대상체, 또는 암이 있는 것으로 알려진 (예를 들어 다른 수단을 통해) 대상체로부터 추출될 수 있다.

[0188] 단리/추출 단계 이후에, 임의의 많은 상이한 서열분석 작업을 세포 유리 폴리뉴클레오티드 샘플에 대해 수행할 수 있다. 샘플은 하나 이상의 시약 (예를 들어, 효소, 특유한 식별자 (예를 들어, 바코드), 프로브 등)으로 서열분석 전에 처리될 수 있다. 일부 경우에, 샘플이 특유한 식별자, 예컨대 바코드로 처리되면, 샘플 또는 샘플의 일부는 특유한 식별자로 개별적으로 또는 하위군에서 태그부착될 수 있다. 이어서, 태그부착된 샘플은 개별 분자가 모 분자에 추적될 수 있는 하류 적용, 예컨대 서열분석 반응에서 사용될 수 있다.

[0189] 세포 유리 폴리뉴클레오티드 서열의 서열분석 데이터가 수집된 후에, 메틸화 프로파일을 포함하나 이에 제한되지는 않는, 유전자 특징 또는 이상, 예컨대 카피수 변이, 희귀 돌연변이 (예를 들어, 단일 또는 다중 뉴클레오티드 변이) 또는 후성적 마커의 변화를 검출하기 위해 하나 이상의 생물 정보공학 과정을 서열 데이터에 적용할 수 있다. 카피수 변이 분석이 요구되는 일부 경우에, 서열 데이터는 1) 참조 게놈과 정렬되고; 2) 여과 및 맵핑되고; 3) 서열의 윈도우 또는 빈 (bin)으로 분할되고; 4) 적용범위 판독체가 각각의 윈도우에 대해 계수되고; 5) 이어서, 적용범위 판독체는 확률적 또는 통계적 모델링 알고리즘 이용하여 정규화될 수 있고; 6) 게놈 내의 다양한 위치에서 별개의 카피수 상태를 반영하는 출력 파일이 생성될 수 있다. 희귀 돌연변이 분석이 요구되는 다른 경우에, 서열 데이터는 1) 참조 게놈과 정렬되고; 2) 여과 및 맵핑되고; 3) 변이체 염기의 빈도를 그 특정한 염기에 대한 적용범위 판독체를 기초로 하여 계산하고; 4) 변이체 염기 빈도를 확률적, 통계적 또는 확률적 모델링 알고리즘을 이용하여 정규화하고; 5) 게놈 내의 다양한 위치에서 돌연변이 상태를 반영하는 출력 파일이 생성될 수 있다.

[0190] 핵산 서열분석, 핵산 정량, 서열분석 최적화, 유전자 발현의 검출, 유전자 발현의 정량, 게놈 프로파일링, 암 프로파일링, 또는 발현된 마커의 분석을 포함하나 이에 제한되지는 않는 다양한 상이한 반응 및/또는 작업이 본원에 개시된 시스템 및 방법 내에서 일어날 수 있다. 또한, 시스템 및 방법은 많은 의학적 용도를 갖는다. 예를 들어, 이것은 암을 비롯한 다양한 유전 및 비-유전 질환 및 장애의 확인, 검출, 진단, 치료, 병기 결정, 또는 위험 예측을 위해 사용될 수 있다. 이것은 상기 유전 및 비-유전 질환의 상이한 치료에 대한 대상체 반응을 평가하거나, 또는 질환 진행 및 예측에 대한 정보를 제공하기 위해 사용될 수 있다.

[0191] 폴리뉴클레오티드 서열분석은 통신 이론에서의 문제와 비교될 수 있다. 초기 개별 폴리뉴클레오티드 또는 폴리뉴클레오티드의 집단 (ensemble)은 원래의 메세지로 생각된다. 태그부착 및/또는 증폭은 원래의 메세지를 신호로 암호화하는 것으로 생각될 수 있다. 서열분석은 통신 채널로서 생각될 수 있다. 서열분석기의 결과, 예를 들어 서열 판독체는 수신된 신호로서 생각될 수 있다. 생물 정보공학 처리는 수신된 신호를 해독하여 전송된 메세지, 예를 들어 뉴클레오티드 서열 또는 서열들을 생산하는 수신기로 생각될 수 있다. 수신된 신호는 아티팩트 (artifact), 예컨대 노이즈 및 왜곡을 포함할 수 있다. 노이즈는 신호에 대한 원치 않는 무작위 부가물로 생각될 수 있다. 왜곡은 신호 또는 신호의 일부의 크기의 변경으로 생각될 수 있다.

[0192] 노이즈는 폴리뉴클레오티드의 복사 및/또는 판독의 오류를 통해 도입될 수 있다. 예를 들어, 서열분석 프로세스에서, 단일 폴리뉴클레오티드는 먼저 증폭에 적용될 수 있다. 증폭은 증폭된 폴리뉴클레오티드의 하위세트가 특정 유전자좌에서의 원래의 염기와 동일하지 않은 염기를 상기 유전자좌에 포함할 수 있도록 오류를 도입할 수 있다. 또한, 판독 과정에서, 임의의 특정 유전자좌에서의 염기는 부정확하게 판독될 수 있다. 그 결과, 서열 판독체의 수집물은 원래의 염기와 동일하지 않은 유전자좌에서의 염기 콜의 특정 백분율을 포함할 수 있다. 전형적인 서열분석 기술에서, 상기 오류 비율은 10% 미만, 예를 들어 2%-3%일 수 있다. 무도 동일한 서열을 갖는 것으로 추정되는 분자의 수집물이 서열분석될 때, 상기 노이즈는 높은 신뢰도로 원래의 염기를 확인할 수 있을 정도로 충분히 작다.

[0193] 그러나, 모 폴리뉴클레오티드의 수집물이 특정 유전자좌에 서열 변이체를 갖는 폴리뉴클레오티드의 하위세트를 포함할 경우, 노이즈는 유의한 문제일 수 있다. 이것은 예를 들어 세포 유리 DNA가 생식계열 DNA뿐만 아니라, 또 다른 공급원으로부터의 DNA, 예컨대 태아 DNA 또는 암 세포로부터의 DNA를 포함할 때 그러하다. 이 경우, 서열 변이체를 갖는 분자의 빈도가 서열분석 과정에 의해 도입되는 오류의 빈도와 동일한 범위 내에 존재하면, 진정한 서열 변이체는 노이즈로부터 구별가능하지 않을 수 있다. 이것은 예를 들어 샘플 내의 서열 변이체의 검출을 방해할 수 있다.

[0194] 왜곡은 신호 강도, 예를 들어 동일한 빈도로 모 집단 내의 분자에 의해 생성된 서열 판독체의 총수의 차이로서 서열분석 과정에서 나타날 수 있다. 왜곡은 예를 들어 증폭 편향, GC 편향, 또는 서열분석 편향을 통해 도입될 수 있다. 이것은 샘플 내의 카피수 변이의 검출을 방해할 수 있다. GC 편향은 서열 판독에서 GC 함량이 풍부하거나 빈약한 영역의 불균등한 제시를 유발한다.

[0195] 본 발명은 폴리뉴클레오티드 서열분석 과정에서 서열분석 아티팩트, 예컨대 노이즈 및/또는 왜곡을 감소시키는

방법을 제공한다. 서열 판독체를 원래의 개별 분자로부터 유래된 패밀리로 분류하는 것은 단일 개별 분자로부터의 또는 분자의 집단으로부터의 노이즈 및/또는 왜곡을 감소시킬 수 있다. 단일 분자에 관하여, 판독체를 패밀리로 분류하면, 예를 들어 많은 서열 판독체가 많은 상이한 분자보다 단일 분자를 실제로 제시함을 나타냄으로써 왜곡을 감소시킬 수 있다. 컨센서스 서열로의 서열 판독체의 붕괴는 하나의 분자로부터 수신된 메세지 내의 노이즈를 감소시키는 한 방법이다. 수신된 빈도를 전환하는 확률 함수를 이용하는 것은 또 다른 방식이다. 분자의 집단에 관하여, 판독체를 패밀리로 분류하고, 패밀리의 정량적 척도를 결정하면, 각각의 다수의 상이한 유전자좌에서, 예를 들어 분자의 양에서의 왜곡을 감소시킬 수 있다. 다시, 컨센서스 서열로의 상이한 패밀리의 서열 판독체의 붕괴는 증폭에 의해 도입된 오류 및/또는 서열분석 오류를 제거한다. 또한, 패밀리 정보로부터 유래된 확률을 기초로 하여 염기 콜의 빈도를 결정하는 것도 분자의 집단로부터 수신된 메세지 내의 노이즈를 감소시킨다.

[0196] 서열분석 과정으로부터 노이즈 및/또는 왜곡을 감소시키는 방법은 공지되어 있다. 예를 들어, 이들은 예를 들어 서열이 품질 역치를 충족할 것을 필요로 하는 서열의 여과, 또는 GC 편향의 감소를 포함한다. 상기 방법은 대체로 서열분석기의 출력인 서열 판독체의 수집물에 대해 수행되고, 패밀리 구조 (원래의 단일 모 분자로부터 유래된 서열의 하위 수집물)를 고려하지 않고 서열 판독체별로 수행될 수 있다. 본 발명의 특정 방법은 서열 판독체의 패밀리 내에서 노이즈 및/또는 왜곡을 감소시킴으로써, 즉 단일 모 폴리뉴클레오티드 분자로부터 유래된 패밀리로 분류된 서열 판독체에 대해 실행함으로써 노이즈 및 왜곡을 감소시킨다. 패밀리 수준에서 신호 아티팩트 감소는 서열 판독체별 수준에서 또는 전체로서의 서열분석기 결과에 대해 수행된 아티팩트 감소보다 유의하게 더 작은, 궁극적인 메세지 내의 노이즈 및 왜곡을 생성할 수 있다.

[0197] 본원은 초기 유전 물질의 샘플에서 유전자 변이를 높은 감도로 검출하기 위한 방법 및 시스템을 추가로 제공한다. 방법은 다음 도구 중의 하나 또는 둘 모두의 사용을 수반한다: 먼저, 초기 유전 물질의 샘플 내의 개별 폴리뉴클레오티드가 서열-준비된 (sequence-ready) 샘플에서 제시될 확률을 증가시키기 위해, 초기 유전 물질의 샘플 내의 개별 폴리뉴클레오티드의 서열-준비된 태그부착된 모 폴리뉴클레오티드로의 효율적인 전환. 이것은 초기 샘플 내에 보다 많은 폴리뉴클레오티드에 관한 서열 정보를 생산할 수 있다. 두 번째로, 태그부착된 모 폴리뉴클레오티드로부터 증폭된 자손 폴리뉴클레오티드의 고속 샘플링, 및 생성된 서열 판독체의 모 태그부착된 폴리뉴클레오티드의 서열을 제시하는 컨센서스 서열로의 붕괴에 의한, 태그부착된 모 폴리뉴클레오티드에 대한 컨센서스 서열의 고수율 생성. 이것은 증폭 편향 및/또는 서열분석 오류에 의해 도입된 노이즈를 감소시킬 수 있고, 검출 감도를 증가시킬 수 있다. 붕괴는 증폭된 분자의 판독체로부터 생성된 다수의 서열 판독체 또는 단일 분자의 다수의 판독체에 대해 수행된다.

[0198] 서열분석 방법은 일반적으로 샘플 제조, 서열 판독체를 생산하기 위해 제조된 샘플 내의 폴리뉴클레오티드의 서열분석 및 샘플에 대한 정량적 및/또는 정성적 유전 정보를 생산하기 위한 서열 판독체의 생물 정보공학 조작을 수반한다. 샘플 제조는 일반적으로 샘플 내의 폴리뉴클레오티드를 사용되는 서열분석 플랫폼에 적합한 형태로 전환하는 것을 수반한다. 상기 전환은 폴리뉴클레오티드를 태그부착하는 것을 포함할 수 있다. 본 발명의 특정 실시양태에서, 태그는 폴리뉴클레오티드 서열 태그를 포함한다. 서열분석에 사용되는 전환 방법은 100% 효율적이지는 않을 수 있다. 예를 들어, 약 1-5%의 전환 효율로 샘플 내의 폴리뉴클레오티드를 전환하는 것은 드물지 않고, 즉 샘플 내의 약 1-5%의 폴리뉴클레오티드가 태그부착된 폴리뉴클레오티드로 전환된다. 태그부착된 분자로 전환되지 않은 폴리뉴클레오티드는 서열분석을 위한 태그부착된 라이브러리에 제시되지 않는다. 따라서, 초기 유전 물질 내에서 낮은 빈도로 나타난 유전자 변이체를 갖는 폴리뉴클레오티드는 태그부착된 라이브러리에 제시되지 않을 수 있고, 따라서 서열분석되거나 검출될 수 없다. 전환 효율을 증가시킴으로써, 초기 유전 물질 내의 희귀한 폴리뉴클레오티드가 태그부착된 라이브러리에 제시되고 따라서 서열분석에 의해 검출될 확률이 증가한다. 또한, 라이브러리 제조의 낮은 전환 효율 문제를 직접 해결하는 것보다, 현재 대부분의 프로토콜은 투입 물질로서 1 마이크로그램 초과의 DNA를 필요로 한다. 그러나, 투입 샘플 물질이 제한되거나 또는 낮은 제시 수준의 폴리뉴클레오티드의 검출이 요구될 때, 높은 전환 효율은 샘플을 효율적으로 서열분석하고/하거나 상기 폴리뉴클레오티드를 적절하게 검출할 수 있다.

[0199] 본 개시내용은 적어도 10%, 적어도 20%, 적어도 30%, 적어도 40%, 적어도 50%, 적어도 60%, 적어도 80% 또는 적어도 90%의 전환 효율로 초기 폴리뉴클레오티드를 태그부착된 폴리뉴클레오티드로 전환하는 방법을 제공한다. 이 방법은 예를 들어, 임의의 평활-말단 라이게이션, 점착성 말단 라이게이션, 분자 역위 프로브, PCR, 라이게이션-기반 PCR, 다중 (multiplex) PCR, 단일 가닥 라이게이션 및 단일 가닥 환형화의 사용을 수반한다. 방법은 또한 초기 유전 물질의 양을 제한하는 것을 수반할 수 있다. 예를 들어, 초기 유전 물질의 양은 1 μg 미만, 100 ng 미만 또는 10 ng 미만일 수 있다. 이들 방법은 본원에서 보다 상세하게 설명된다.

[0200] 태그부착된 라이브러리 내의 폴리뉴클레오티드에 대한 정확한 정량적 및 정성적 정보를 획득하면, 초기 유전 물질에 대한 보다 상세한 특성화가 가능하다. 대체로, 태그부착된 라이브러리 내의 폴리뉴클레오티드는 증폭되고, 생성되는 증폭된 분자는 서열분석된다. 사용되는 서열분석 플랫폼의 처리량에 따라, 증폭된 라이브러리 내의 분자의 하위세트만 서열 판독체를 생성한다. 따라서, 예를 들어 서열분석을 위해 샘플링된 증폭된 분자의 수는 태그부착된 라이브러리 내의 특유한 폴리뉴클레오티드의 약 50%에 불과할 수 있다. 또한, 증폭은 태그부착된 라이브러리의 특정 서열 또는 특정 구성원에 유리하게 또는 불리하게 편향될 수 있다. 이것은 태그부착된 라이브러리 내의 서열의 정량적 척도를 왜곡할 수 있다. 또한, 서열분석 플랫폼은 서열분석에 오류를 도입할 수 있다. 예를 들어, 서열의 염기당 오류 비율은 0.5-1%일 수 있다. 증폭 편향 및 서열분석 오류는 최종 서열분석 생성물에 노이즈를 도입한다. 상기 노이즈는 검출 감도를 저하시킬 수 있다. 예를 들어, 그의 태그부착된 집단 내의 빈도가 서열분석 오류 비율보다 작은 서열 변이체는 노이즈로 오인될 수 있다. 또한, 집단 내의 그의 실제 수보다 더 많거나 더 적은 양으로 서열 판독체를 제공함으로써, 증폭 편향은 카피수 변이의 측정치를 왜곡할 수 있다. 별법으로, 단일 폴리뉴클레오티드로부터의 다수의 서열 판독체는 증폭 없이 생산될 수 있다. 이것은 예를 들어 나노포어 방법으로 수행될 수 있다.

[0201] 본 개시내용은 태그부착된 풀 (pool) 내의 특유한 폴리뉴클레오티드를 정확하게 검출하고 판독하는 방법을 제공한다. 특정 실시양태에서, 본 개시내용은 증폭되고 서열분석될 때, 또는 다수의 서열 판독체를 생산하기 위해 다수회 서열분석될 때 자손 폴리뉴클레오티드의 특유한 태그 모 폴리뉴클레오티드 분자로의 역추적, 또는 붕괴를 허용한 정보를 제공하는 서열-태그부착된 폴리뉴클레오티드를 제공한다. 증폭된 자손 폴리뉴클레오티드의 패밀리의 붕괴는 원래의 특유한 모 분자에 대한 정보를 제공함으로써 증폭 편향을 감소시킨다. 붕괴는 또한 서열분석 데이터로부터 자손 분자의 돌연변이체 서열을 제거함으로써 서열분석 오류를 감소시킨다.

[0202] 태그부착된 라이브러리 내의 특유한 폴리뉴클레오티드의 검출 및 판독은 2개의 전략을 포함할 수 있다. 한 전략에서, 태그부착된 모 폴리뉴클레오티드의 세트 내의 특유한 태그부착된 모 폴리뉴클레오티드의 큰 백분율에 대해, 특유한 태그부착된 모 폴리뉴클레오티드로부터 생산된 패밀리 내의 적어도 하나의 증폭된 자손 폴리뉴클레오티드에 대해 생산된 서열 판독체가 존재하도록 증폭된 자손 폴리뉴클레오티드 풀의 충분히 큰 하위세트가 서열분석된다. 제2 전략에서, 증폭된 자손 폴리뉴클레오티드 세트는 특유한 모 폴리뉴클레오티드로부터 유래된 패밀리의 다중 자손 구성원으로부터 서열 판독체를 생산하기 위한 수준에서 서열분석을 위해 샘플링된다. 패밀리의 다중 자손 구성원으로부터 서열 판독체의 생성은 서열의 컨센서스 모 서열로의 붕괴를 허용한다.

[0203] 따라서, 예를 들어 태그부착된 모 폴리뉴클레오티드의 세트 내의 특유한 태그부착된 모 폴리뉴클레오티드의 수와 동일한 (특히 수가 적어도 10,000개일 때) 증폭된 자손 폴리뉴클레오티드의 세트로부터 많은 증폭된 자손 폴리뉴클레오티드의 샘플링은 통계적으로 세트 내의 약 68%의 태그부착된 모 폴리뉴클레오티드의 적어도 하나의 자손에 대한 서열 판독체를 생산할 것이고, 원래의 세트 내의 약 40%의 특유한 태그부착된 모 폴리뉴클레오티드가 적어도 2개의 자손 서열 판독체에 의해 제시될 것이다. 특정 실시양태에서, 증폭된 자손 폴리뉴클레오티드 세트는 각각의 패밀리에 대해 평균 5 내지 10개의 서열 판독체를 생산하도록 충분히 샘플링된다. 특유한 태그부착된 모 폴리뉴클레오티드의 수만큼 많은 분자의 10배의 증폭된 자손 세트로부터의 샘플링은 통계적으로 99.995%의 패밀리에 대한 서열 정보를 생성할 것이고, 그의 99.95%의 총 패밀리가 다수의 서열 판독체에 의해 포함될 것이다. 컨센서스 서열은 가능하게는 10의 수 제곱 더 낮은 비율로 명목상 염기당 서열분석 오류 비율로부터 오류 비율을 극적으로 감소시키기 위해 각각의 패밀리 내의 자손 폴리뉴클레오티드로부터 만들어질 수 있다. 예를 들어, 서열분석기가 1%의 무작위 염기당 오류 비율을 갖고 선택된 패밀리가 10개의 판독체를 가질 경우, 상기 10개의 판독체로부터 만들어진 컨센서스 서열은 0.0001% 미만의 오류 비율을 가질 것이다. 따라서, 서열분석되는 증폭된 자손의 샘플링 크기는, 사용되는 서열분석 플랫폼의 비율까지 명목상 염기당 서열분석 오류 비율보다 크지 않은 샘플 내 빈도를 갖는 서열이 적어도 하나의 판독체에 의해 제시될 적어도 99% 가능성을 갖는 것을 보장하도록 선택될 수 있다.

[0204] 또 다른 실시양태에서, 증폭된 자손 폴리뉴클레오티드의 세트는 사용되는 서열분석 플랫폼의 염기당 서열분석 오류 비율과 거의 동일한 빈도로 태그부착된 모 폴리뉴클레오티드의 세트에 제시되는 서열이 적어도 하나의 서열 판독체 및 바람직하게는 다수의 서열 판독체에 의해 포함될 높은 확률, 예를 들어 적어도 90%를 생성하기 위한 수준으로 샘플링된다. 따라서, 예를 들어 서열 또는 서열의 세트에 0.2%의 염기당 오류 비율을 갖는 서열분석 플랫폼이 약 0.2%의 빈도로 태그부착된 모 폴리뉴클레오티드의 세트에 제시될 경우, 서열분석된 증폭된 자손 풀 내의 폴리뉴클레오티드의 수는 대략 태그부착된 모 폴리뉴클레오티드의 세트 내의 특유한 분자의 수의 X배일 수 있다.

[0205] 이들 방법은 예를 들어 컨센서스 서열을 생성하기 위해 사용되는 서열의 풀에 포함되기 위한 서열 판독체를 평

가하는 것을 포함하는 설명된 임의의 노이즈 감소 방법과 조합될 수 있다.

[0206] 상기 정보는 이제 정성적 및 정량적 분석 모두를 위해 사용될 수 있다. 예를 들어, 정량적 분석을 위해, 참조 서열에 맵핑되는 태그부착된 모 분자의 양의 척도, 예를 들어 계수를 결정한다. 상기 척도를 상이한 게놈 영역에 맵핑되는 태그부착된 모 분자의 척도와 비교할 수 있다. 즉, 참조 서열, 예컨대 인간 게놈 내의 제1 위치 또는 맵핑가능 위치에 맵핑되는 태그부착된 모 분자의 양은 참조 서열 내의 제2 위치 또는 맵핑가능 위치에 맵핑되는 태그부착된 모 분자의 척도와 비교될 수 있다. 상기 비교는 예를 들어 각각의 영역에 맵핑되는 모 분자의 상대적인 양을 보여줄 수 있다. 이것은 다시 특정 영역에 맵핑되는 분자에 대한 카피수 변이의 표지를 제공한다. 예를 들어, 제1 참조 서열에 맵핑되는 폴리뉴클레오티드의 척도가 제2 참조 서열에 맵핑되는 폴리뉴클레오티드의 척도보다 크면, 이것은 모 집단, 및 나아가 원래의 샘플이 이수성을 보이는 세포로부터의 폴리뉴클레오티드를 포함함을 나타낼 수 있다. 척도는 다양한 편향을 제거하기 위해 대조 샘플에 대해 정규화될 수 있다. 정량적 척도는 예를 들어, 수, 계수, 빈도 (상대적, 추정된 또는 절대적)를 포함할 수 있다.

[0207] 참조 게놈은 관심 있는 임의의 종의 게놈을 포함할 수 있다. 참조물로서 유용한 인간 게놈 서열은 hg19 조립체 (assembly) 또는 임의의 이전의 또는 이용가능한 hg 조립체를 포함할 수 있다. 상기 서열은 genome.ucsc.edu/index.html에서 이용가능한 게놈 브라우저를 사용하여 조사할 수 있다. 다른 종의 게놈은 예를 들어 PanTro2 (챔팬지) 및 mm9 (마우스)를 포함한다.

[0208] 정성 분석을 위해, 참조 서열에 맵핑되는 태그부착된 폴리뉴클레오티드의 세트로부터의 서열은 변이체 서열에 대해 분석될 수 있고, 태그부착된 모 폴리뉴클레오티드의 집단 내의 그의 빈도가 측정될 수 있다.

[0209] **II. 샘플 제조**

[0210] **A. 폴리뉴클레오티드 단리 및 추출**

[0211] 본 개시내용의 시스템 및 방법은 세포 유리 폴리뉴클레오티드의 조작, 제제, 확인 및/또는 정량에서 매우 다양한 용도를 가질 수 있다. 폴리뉴클레오티드의 예는 DNA, RNA, 앰플리콘, cDNA, dsDNA, ssDNA, 플라스미드 DNA, 코스미드 DNA, 고 분자량 (MW) DNA, 염색체 DNA, 게놈 DNA, 바이러스 DNA, 박테리아 DNA, mtDNA (미토콘드리아 DNA), mRNA, rRNA, tRNA, nRNA, siRNA, snRNA, snoRNA, scaRNA, 마이크로RNA, dsRNA, 리보자임, 리보스위치 (riboswitch) 및 바이러스 RNA (예를 들어, 레트로바이러스 RNA)를 포함하나 이에 제한되지는 않는다.

[0212] 세포 유리 폴리뉴클레오티드는 인간, 포유동물, 비-인간 포유동물, 유인원, 원숭이, 침팬지, 파충류, 양서류, 또는 조류 공급원을 포함하는 다양한 공급원으로부터 유래될 수 있다. 또한, 샘플은 혈액, 혈청, 혈장, 유리체액, 객담, 소변, 눈물, 땀, 타액, 정액, 점막 분비물, 점액, 척수액, 양수, 림프액 등을 포함하나 이에 제한되지는 않는 세포 유리 서열을 포함하는 다양한 동물 체액으로부터 추출될 수 있다. 세포 유리 폴리뉴클레오티드는 태아에서 기원하거나 (임신한 대상체로부터 채취한 유체를 통해), 또는 대상체 자체의 조직으로부터 유래될 수 있다.

[0213] 세포 유리 폴리뉴클레오티드의 단리 및 추출은 다양한 기술을 사용하여 체액의 수집을 통해 수행될 수 있다. 일부 경우에, 수집은 주사기를 사용한, 대상체로부터 체액의 흡입을 포함할 수 있다. 다른 경우에, 수집은 체액을 피펫을 사용한 수집 또는 수집 용기 내로의 직접 수집을 포함할 수 있다.

[0214] 체액 수집 후에, 관련 기술 분야에 공지된 다양한 기술을 이용하여 세포 유리 폴리뉴클레오티드를 단리하고 추출할 수 있다. 일부 경우에, 상업상 이용가능한 키트, 예컨대 퀴아젠 (Qiagen) Qiamp® 서큘레이팅 뉴클레익 애시드 (Circulating Nucleic Acid) 키트 프로토콜을 사용하여 세포 유리 DNA를 단리하고, 추출하고, 조제할 수 있다. 다른 예에서, 퀴아젠 큐빗(Qubit)™ dsDNA HS 분석 키트 프로토콜, 애질런트(Agilent)™ DNA 1000 키트, 또는 TruSeq™ 서열분석 라이브러리 제제; 저-처리량 (LT) 프로토콜을 이용할 수 있다.

[0215] 일반적으로, 세포 유리 폴리뉴클레오티드는 용액에서 발견되는 세포 유리 DNA가 세포 및 체액의 다른 비가용성 성분으로부터 분리되는 분할 단계를 통해 체액으로부터 추출하고 단리된다. 분할은 원심분리 또는 여과와 같은 기술을 포함할 수 있고 이로 제한되지 않는다. 다른 경우에, 세포는 먼저 세포 유리 DNA로부터 분할되지 않고, 오히려 용해된다. 상기 예에서, 무손상 세포의 게놈 DNA는 선택적 침전을 통해 분할된다. DNA를 비롯한 세포 유리 폴리뉴클레오티드는 가용형으로 남을 수 있고, 불용형 게놈 DNA로부터 분리되고 추출될 수 있다. 일반적으로, 완충제 및 상이한 키트에 특이적인 다른 세척 단계 후에, 이소프로판올 침전을 이용하여 DNA를 침전시킬 수 있다. 오염물 또는 염을 제거하기 위해 실리카 기반 컬럼과 같은 추가의 세정 단계를 이용할 수 있다. 일반적인 단계는 특수한 용도를 위해 최적화될 수 있다. 예를 들어, 절차의 특정 측면, 예컨대 수율을 최적화하

기 위해 비-특이적 벌크 담체 폴리뉴클레오티드가 반응을 통해 첨가될 수 있다.

[0216] 세포 유리 DNA의 단리 및 정제는 시그마 알드리치 (Sigma Aldrich), 라이프 테크놀로지스 (Life Technologies), 프로메가 (Promega), 아피메트릭스 (Affymetrix), IBI 등과 같은 회사에 의해 공급되는 시판 키트 및 프로토콜의 사용을 포함하나 이에 제한되지는 않는 임의의 수단을 이용하여 달성할 수 있다. 키트 및 프로토콜은 또한 비-상업적으로 이용가능할 수 있다.

[0217] 단리 후에, 일부 경우에, 세포 유리 폴리뉴클레오티드는 서열분석 전에 하나 이상의 추가의 물질, 예컨대 하나 이상의 시약 (예를 들어, 리가제, 프로테아제, 폴리머라제)과 미리 혼합된다.

[0218] 전환 효율을 증가시키기 위한 한 방법은 단일-가닥 DNA에 대한 최적 반응성을 위해 조작된 리가제, 예컨대 ThermoPhage ssDNA 리가제 유도체의 사용을 수반한다. 상기 리가제는 중간체 세정 단계에 의한 불량한 효율 및/또는 누적된 소실을 가질 수 있는 말단-복구 및 A-테일링 (tailing)의 라이브러리 제조의 전통적인 단계를 우회하고, 센스 또는 안티센스 출발 폴리뉴클레오티드가 적절하게 태그부착된 폴리뉴클레오티드로 2배의 확률로 전환되는 것을 허용한다. 이것은 또한 전형적인 말단 복구 반응에 의해 충분하게 평활-말단화될 수 없는 오버행 (overhang)을 가질 수 있는 이중 가닥 폴리뉴클레오티드를 전환한다. 상기 ssDNA 반응을 위한 최적 반응 조건은 다음과 같다: 1x 반응 완충제 (50 mM MOPS (pH 7.5), 1 mM DTT, 5 mM $MgCl_2$, 10 mM KCl). 50 mM ATP, 25 mg/ml BSA, 2.5 mM $MnCl_2$, 200 pmol 85 nt ssDNA 올리고머 및 5 U ssDNA 리가제와 함께 65℃에서 1시간 동안 인큐베이팅한다. PCR을 사용한 후속적인 증폭은 태그부착된 단일-가닥 라이브러리를 이중 가닥 라이브러리로 전환하고, 20% 초과의 웰의 총 전환 효율을 생성한다. 전환율을 예를 들어 10% 초과로 증가시키는 다른 방법은 예를 들어 다음 중의 임의의 하나 또는 조합을 포함한다: 어닐링-최적화 분자-역위 프로브, 잘 제어된 폴리뉴클레오티드 크기 범위를 사용한 평활-말단 라이게이션, 점착성-말단 라이게이션 또는 융합 프라이머를 사용하거나 사용하지 않는 사전 다중 (upfront multiplex) 증폭 단계.

[0219] **B. 세포 유리 폴리뉴클레오티드의 분자 바코드 부착**

[0220] 본 개시내용의 시스템 및 방법은 또한 특정 폴리뉴클레오티드의 후속적인 확인 및 발생을 허용하기 위해 세포 유리 폴리뉴클레오티드가 태그부착되거나 추적되는 것을 가능하게 할 수 있다. 상기 특징은, 함께 모은 또는 다중 반응을 이용하고 측정 또는 분석을 단지 다중 샘플의 평균으로서 제공하는 다른 방법과 대조적이다. 여기서, 폴리뉴클레오티드의 개체 또는 하위군에 대한 식별자의 배정은 특유한 정체가 개별 서열 또는 서열의 단편에 배정되는 것을 허용할 수 있다. 이것은 개별 샘플로부터 데이터의 획득을 허용할 수 있고, 샘플의 평균으로 제한되지 않는다.

[0221] 일부 예에서, 단일 가닥으로부터 유래된 핵산 또는 다른 분자는 공통적인 태그 또는 식별자를 공유할 수 있고, 따라서 그 가닥으로부터 유래되는 것으로 나중에 확인될 수 있다. 유사하게, 핵산의 단일 가닥으로부터의 모든 단편은 동일한 식별자 또는 태그로 태그부착될 수 있고, 이에 의해 모 가닥으로부터의 단편의 후속적인 확인이 가능하다. 다른 경우에, 유전자 발현 산물 (예를 들어, mRNA)은 발현을 정량하기 위해 태그부착될 수 있고, 이에 의해 바코드, 또는 부착되는 서열과 조합된 바코드가 계수될 수 있다. 또 다른 경우에, 시스템 및 방법은 PCR 증폭 대조군으로서 사용될 수 있다. 그 경우에, PCR 반응으로부터의 다중 증폭 산물은 동일한 태그 또는 식별자로 태그부착될 수 있다. 산물이 후에 서열분석되고 서열 차이를 보이면, 동일한 식별자를 갖는 산물 사이의 차이는 PCR 오류에 기여할 수 있다.

[0222] 추가로, 개별 서열은 판독체 자체에 대한 서열 데이터의 특징을 기초로 하여 확인될 수 있다. 예를 들어, 개별 서열분석 판독체의 개시 (출발) 및 종료 (정지) 부분에서 특유한 서열 데이터의 검출은 단독으로, 또는 특유한 정체를 개별 분자에 배정하기 위한 각각의 서열 판독체 특유 서열의 염기쌍의 길이, 또는 수와 조합하여 사용될 수 있다. 특유한 정체가 배정된 핵산의 단일 가닥으로부터의 단편은 이에 의해 모 가닥으로부터의 단편의 후속적인 확인이 가능할 수 있다. 이것은 다양성을 제한하기 위한 초기 출발 유전 물질의 병목현상화와 함께 사용될 수 있다.

[0223] 또한, 개별 서열분석 판독체의 개시 (출발) 및 종료 (정지) 부분에서의 특유한 서열 데이터 및 서열분석 판독체 길이의 사용은 단독으로 또는 바코드의 사용과 조합하여 사용될 수 있다. 일부 경우에, 바코드는 본원에서 설명되는 바와 같이 특유할 수 있다. 다른 경우에, 바코드 자체는 특유하지 않을 수 있다. 이 경우에, 개별 서열분석 판독체의 개시 (출발) 및 종료 (정지) 부분에서의 서열 데이터 및 서열분석 판독체 길이와 조합한 비-특유한 바코드의 사용은 개별 서열에 대한 특유한 정체의 배정을 허용할 수 있다. 이와 유사하게, 특유한 정체가 배정된 핵산의 단일 가닥으로부터의 단편은 이에 의해 모 가닥으로부터의 단편의 후속적인 확인이 가능할 수 있

다.

[0224] 일반적으로, 본원에 제공되는 방법 및 시스템은 하류 적용 서열분석 반응의 세포 유리 폴리뉴클레오티드 서열의 제조에 유용하다. 종종, 서열분석 방법은 전통적인 생거 (Sanger) 서열분석이다. 서열분석 방법은 다음을 포함할 수 있고 이로 제한되지 않는다: 고-처리량 서열분석, 피로서열분석, 합성에 의한 서열분석, 단일-분자 서열분석, 나노포어 서열분석, 반도체 서열분석, 라이게이션에 의한 서열분석, 혼성화에 의한 서열분석, RNA-Seq (일루미나), 디지털 유전자 발현 (헬리코스 (Helicos)), 차세대 (Next generation) 서열분석, 합성에 의한 단일 분자 서열분석 (SMSS)(헬리코스), 대규모 병렬형 (massively-parallel) 서열분석, 클로날 단일 분자 어레이 (Clonal Single Molecule Array) (솔렉사 (Solexa)), 샷건 (shotgun) 서열분석, 맥심-길버트 (Maxim-Gilbert) 서열분석, 프라이머 워킹 (walking), 및 관련 기술 분야에 공지된 임의의 다른 서열분석 방법.

[0225] **C. 세포 유리 폴리뉴클레오티드 서열에 대한 바코드의 배정**

[0226] 본원에 개시된 시스템 및 방법은 세포 유리 폴리뉴클레오티드에 특유한 또는 비-특유한 식별자, 또는 분자 바코드의 배정을 수반하는 용도에서 사용될 수 있다. 종종, 식별자는 폴리뉴클레오티드의 태그부착에 사용되는 바코드 올리고뉴클레오티드이지만; 일부 경우에, 상이한 특유한 식별자가 사용된다. 예를 들어, 일부 경우에, 특유한 식별자는 혼성화 프로브이다. 다른 경우에, 특유한 식별자는 염료이고, 이 경우에 부착은 염료의 분석물 분자 내로의 삽입 (intercalation) (예컨대 DNA 또는 RNA 내로의 삽입) 또는 염료로 표지된 프로브에 대한 결합을 포함할 수 있다. 또 다른 경우에, 특유한 식별자는 핵산 올리고뉴클레오티드일 수 있고, 이 경우에 폴리뉴클레오티드 서열에 대한 부착은 올리고뉴클레오티드와 서열 사이의 라이게이션 반응 또는 PCR을 통한 도입을 포함할 수 있다. 다른 경우에, 반응은 금속 동원원소의 분석물에 대한 직접 부가 또는 동위원소로 표지된 프로브에 의한 부가를 포함할 수 있다. 일반적으로, 본 개시내용의 반응에서 특유한 또는 비-특유한 식별자, 또는 분자 바코드의 배정은 예를 들어, US 특허 출원 20010053519, 20030152490, 20110160078 및 US 특허 US 6,582,908에 기재된 방법 및 시스템을 따를 수 있다.

[0227] 종종, 방법은 라이게이션 반응을 포함하나 이에 제한되지는 않는 효소 반응을 통해 올리고뉴클레오티드 바코드를 핵산 분석물에 부착하는 것을 포함한다. 예를 들어, 리가제 효소는 DNA 바코드를 단편화된 DNA (예를 들어, 고 분자량 DNA)에 공유 부착시킬 수 있다. 바코드의 부착 후에, 분자는 서열분석 반응에 적용될 수 있다.

[0228] 그러나, 다른 반응이 사용될 수도 있다. 예를 들어, 바코드 서열을 포함하는 올리고뉴클레오티드 프라이머는 DNA 주형 분석물의 증폭 반응 (예를 들어, PCR, qPCR, 역전사효소 PCR, 디지털 PCR 등)에 사용되어, 태그부착된 분석물을 생산할 수 있다. 개별 세포 유리 폴리뉴클레오티드 서열에 대한 바코드 배정 후에, 분자의 풀이 서열분석될 수 있다.

[0229] 일부 경우에, 세포 유리 폴리뉴클레오티드 서열의 포괄적 증폭을 위해 PCR을 이용할 수 있다. 이것은 먼저 상이한 분자에 라이게이션될 수 있는 어댑터 (adapter) 서열의 사용, 이어서 범용 프라이머를 사용하는 PCR 증폭을 포함할 수 있다. 서열분석을 위한 PCR은 누겐 (Nugen) (WGA 키트), 라이프 테크놀로지스, 아피메트릭스, 프로메가, 퀴아젠 등에 의해 제공되는 시판 키트의 사용을 포함하나 이에 제한되지는 않는 임의의 수단을 이용하여 수행할 수 있다. 다른 경우에, 세포 유리 폴리뉴클레오티드 분자의 집단 내에서 단지 특정 표적 분자만이 증폭될 수 있다. 특이적 프라이머는 어댑터 라이게이션과 함께, 하류 서열분석을 위해 특정 표적을 선택적으로 증폭하기 위해 사용될 수 있다.

[0230] 특유한 식별자 (예를 들어, 올리고뉴클레오티드 바코드, 항체, 프로브 등)이 세포 유리 폴리뉴클레오티드 서열에 무작위로 또는 비-무작위로 도입될 수 있다. 일부 경우에, 이들은 특유한 식별자 대 마이크로웰 (microwell)의 예상된 비율로 도입된다. 예를 들어, 특유한 식별자는 약 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50,000, 100,000, 500,000, 1,000,000, 10,000,000, 50,000,000 또는 1,000,000,000개 초과의 특유한 식별자가 게놈 샘플당 로딩되도록 로딩될 수 있다. 일부 경우에, 특유한 식별자는 약 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50,000, 100,000, 500,000, 1,000,000, 10,000,000, 50,000,000 또는 1,000,000,000개 미만의 특유한 식별자가 게놈 샘플당 로딩되도록 로딩될 수 있다. 일부 경우에, 샘플 게놈당 로딩되는 특유한 식별자의 평균 수는 게놈 샘플당 약 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50,000, 100,000, 500,000, 1,000,000, 10,000,000, 50,000,000 또는 1,000,000,000개 미만 또는 초과의 특유한 식별자이다.

[0231] 일부 경우에, 특유한 식별자는 각각의 바코드가 적어도 약 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000개의 염기쌍이 되도록 하는 다양한 길이일 수 있다. 다른 경우에, 바코드는 1, 2, 3, 4, 5, 6, 7, 8, 9,

10, 20, 50, 100, 500, 1000개 미만의 염기쌍을 포함할 수 있다.

[0232] 일부 경우에, 특유한 식별자는 미리 결정된 또는 무작위 또는 준-무작위 서열 올리고뉴클레오티드일 수 있다. 다른 경우에, 바코드가 반드시 복수로 서로 특유히지는 않도록 다수의 바코드가 사용될 수 있다. 상기 예에서, 바코드는 바코드가 라이게이션되는 서열과 바코드의 조합물이 개별적으로 추적될 수 있는 특유한 서열을 생성하도록 개별 분자에 라이게이션될 수 있다. 본원에서 설명되는 바와 같이, 서열 판독체의 개시 (출발) 및 종료 (정지) 부분의 서열 데이터와 조합하여 비 특유한 바코드의 검출은 특정 분자에 특유한 정체의 배정을 허용한다. 개체 서열 판독체의 길이, 또는 염기쌍의 수도 특유한 정체를 그러한 분자에 배정하기 위해 사용될 수 있다. 본원에서 설명되는 바와 같이, 이에 의해 특유한 정체가 배정된 핵산의 단일 가닥으로부터의 단편은 모 가닥으로부터 단편의 후속적인 확인을 허용할 수 있다. 상기 방식으로, 샘플 내의 폴리뉴클레오티드는 특유하게 또는 실질적으로 특유하게 태그부착될 수 있다.

[0233] 특유한 식별자는 RNA 또는 DNA 분자를 포함하나 이에 제한되지는 않는 광범위한 분석물을 태그부착하기 위해 사용될 수 있다. 예를 들어, 특유한 식별자 (예를 들어, 바코드 올리고뉴클레오티드)는 핵산의 전체 가닥 또는 핵산의 단편 (예를 들어, 단편화된 게놈 DNA, 단편화된 RNA)에 부착될 수 있다. 특유한 식별자 (예를 들어, 올리고뉴클레오티드)는 또한 유전자 발현 산물, 게놈 DNA, 미토콘드리아 DNA, RNA, mRNA 등에 결합할 수 있다.

[0234] 많은 용도에서, 개별 세포 유리 폴리뉴클레오티드 서열이 각각 상이한 특유한 식별자 (예를 들어, 올리고뉴클레오티드 바코드)를 수용하는지 결정하는 것은 중요할 수 있다. 시스템 및 방법 내로 도입된 특유한 식별자의 집단이 유의하게 다양하지 않으면, 상이한 분석물은 가능하게는 동일한 식별자로 태그부착될 수 있다. 본원에 개시된 시스템 및 방법은 동일한 식별자로 태그부착된 세포 유리 폴리뉴클레오티드 서열의 검출을 가능하게 할 수 있다. 일부 경우에, 참조 서열은 분석할 세포 유리 폴리뉴클레오티드 서열의 집단과 함께 포함될 수 있다. 참조 서열은 예를 들어, 공지의 서열 및 공지의 양을 갖는 핵산일 수 있다. 특유한 식별자가 올리고뉴클레오티드 바코드이고 분석물이 핵산일 경우, 태그부착된 분석물은 후속적으로 서열분석되고 정량될 수 있다. 이들 방법은 하나 이상의 단편 및/또는 분석물에게 동일한 바코드가 배정될 수 있음을 나타낼 수 있다.

[0235] 본원에 개시된 방법은 분석물에 대한 바코드의 배정을 위해 필요한 시약을 이용하는 것을 포함할 수 있다. 라이게이션 반응의 경우에, 리가제 효소, 완충제, 어댑터 올리고뉴클레오티드, 다수의 특유한 식별자 DNA 바코드 등을 포함하나 이에 제한되지는 않는 시약이 시스템 및 방법 내로 로딩될 수 있다. 풍부화의 경우에, 다수의 PCR 프라이머, 특유한 확인 서열 함유 올리고뉴클레오티드, 또는 바코드 서열, DNA 폴리머라제, DNTP, 및 완충제 등을 포함하나 이에 제한되지는 않는 시약이 서열분석을 위한 제제에 사용될 수 있다.

[0236] 일반적으로, 본 개시내용의 방법 및 시스템은 분자 또는 분석물을 계수하기 위해 분자 바코드를 사용할 때 US 특허 US 7,537,897의 방법을 이용할 수 있다.

[0237] 다수의 게놈으로부터 단편화된 게놈 DNA, 예를 들어 세포 유리 DNA (cfDNA)를 포함하는 샘플 내에, 상이한 게놈으로부터의 하나 초과의 폴리뉴클레오티드가 동일한 출발 및 정지 위치 ("중복체" 또는 "동족체 (cognate)")를 가질 일부 가능성이 존재한다. 임의의 위치에서 시작하는 중복체의 가능한 수는 샘플 내의 반수체 게놈 등가물의 수 및 단편 크기의 분포의 함수이다. 예를 들어, cfDNA는 약 160개 뉴클레오티드에서 단편의 피크를 갖고, 상기 피크에서 대부분의 단편은 약 140개 뉴클레오티드 내지 180개 뉴클레오티드 범위이다. 따라서, 약 3십억 개 염기의 게놈 (예를 들어, 인간 게놈)으로부터의 cfDNA는 거의 2천만 ($2x10^7$) 개의 폴리뉴클레오티드 단편으로 이루어질 수 있다. 약 30 ng DNA의 샘플은 약 10,000개의 반수체 인간 게놈 등가물을 함유할 수 있다 (유사하게, 약 100 ng의 DNA의 샘플은 약 30,000개의 반수체 인간 게놈 등가물을 함유할 수 있다). 그러한 DNA의 약 10,000 (10^4)개 반수체 게놈 등가물을 함유하는 샘플은 약 2천억 ($2x10^{11}$) 개의 개별 폴리뉴클레오티드 분자를 가질 수 있다. 인간 DNA의 약 10,000개의 반수체 게놈 등가물의 샘플 내에, 임의의 주어진 위치에서 시작하는 약 3개의 중복 폴리뉴클레오티드가 존재함이 경험적으로 결정되었다. 따라서, 그러한 수집물은 약 $6x10^{10}$-$8x10^{10}$ (약 6백억-8백억, 예를 들어 약 7백억 ($7x10^{10}$)) 개의 상이하게 서열분석된 폴리뉴클레오티드 분자의 다양성을 포함할 수 있다.

[0238] 분자를 정확하게 확인할 확률은 게놈 등가물의 초기 수, 서열분석된 분자의 길이 분포, 서열 균일성 및 태그의 수에 의해 결정된다. 태그 계수가 1일 때, 등가물은 특유한 태그를 갖지 않거나 태그부착되지 않은 것이다. 아래 표는 상기한 바와 같은 전형적인 세포 유리 크기 분포를 가정하면서 분자를 특유한 것으로 정확하게 확인할 확률을 제시한다.

태그 계수	정확하게 특유한 것으로 확인된 태그 %
1000개의 인간 반수체 게놈 등가물	
1	96.9643
4	99.2290
9	99.6539
16	99.8064
25	99.8741
100	99.9685
3000개의 인간 반수체 게놈 등가물	
1	91.7233
4	97.8178
9	99.0198
16	99.4424
25	99.6412
100	99.9107

[0239]

[0240] 이 경우에, 게놈 DNA의 서열분석시에, 어떤 서열 판독체가 모 분자로부터 유래되는지 결정하는 것이 가능하지 않을 수 있다. 이 문제는 2개의 중복 분자, 즉, 동일한 출발 및 정지 위치를 갖는 분자가 상이한 특유한 식별자를 보유하여 서열 판독체가 특정 모 분자까지 역추적할 수 있을 가능성이 존재하도록, 모 분자를 충분한 수의 특유한 식별자 (예를 들어, 태그 계수)로 태그부착함으로써 감소될 수 있다. 상기 문제에 대한 하나의 방안은 샘플 내의 모든 또는 거의 모든 상이한 모 분자를 특유하게 태그부착하는 것이다. 그러나, 샘플 내의 반수체 유전자 등가물의 수 및 단편 크기의 분포에 따라, 이것은 수십억 개의 상이한 특유한 식별자를 요구할 수 있다.

[0241] 상기 방법은 번거롭고 비용이 많이 소용될 수 있다. 본 발명은 단편화된 게놈 DNA의 샘플 내의 폴리뉴클레오티드의 집단이 n개의 상이한 특유한 식별자로 태그부착되는 방법 및 조성물을 제공하고, 여기서 n은 적어도 2 내지 100,000*z 이하이고, 여기서 z는 동일한 출발 및 정지 위치를 갖는 예상된 수의 중복 분자의 중심 경향도 (예를 들어, 평균, 중간값 또는 최빈값)의 척도이다. 특정 실시양태에서, n은 적어도 2*z, 3*z, 4*z, 5*z, 6*z, 7*z, 8*z, 9*z, 10*z, 11*z, 12*z, 13*z, 14*z, 15*z, 16*z, 17*z, 18*z, 19*z, 또는 20*z (예를 들어, 하한)의 임의의 값이다. 다른 실시양태에서, n은 100,000*z, 10,000*z, 1000*z 또는 100*z 이하 (예를 들어, 상한)이다. 따라서, n은 상기 하한 내지 상한의 임의의 조합의 범위일 수 있다. 특정 실시양태에서, n은 5*z 내지 15*z, 8*z 내지 12*z, 또는 약 10*z이다. 예를 들어, 반수체 인간 게놈 등가물은 약 3 피코그램의 DNA를 갖는다. 약 1 마이크로그램의 DNA의 샘플은 약 300,000개의 반수체 인간 게놈 등가물을 함유한다. 숫자 n은 15 내지 45, 24 내지 36 또는 약 30일 수 있다. 서열분석의 개선은 적어도 일부의 중복 또는 동족체 폴리뉴클레오티드가 특유한 식별자를 보유하면, 즉 상이한 태그를 보유하면 달성될 수 있다. 그러나, 특정 실시양태에서, 사용되는 태그의 수는 임의의 하나의 위치에서 출발하는 모든 중복 분자가 특유한 식별자를 보유할 적어도 95%의 가능성이 존재하도록 선택된다. 예를 들어, cfDNA의 약 10,000개의 반수체 인간 게놈 등가물을 포함하는 샘플은 약 36개의 특유한 식별자로 태그부착될 수 있다. 특유한 식별자는 6개의 특유한 DNA 바코드를 포함할 수 있다. 폴리뉴클레오티드의 양 말단에 부착된 36개의 가능한 특유한 식별자가 생산된다. 상기 방식으로 태그부착된 샘플은 약 10 ng 내지 약 100 ng, 약 1 ㎍, 약 10 ㎍의 단편화된 폴리뉴클레오티드, 예를 들어 게놈 DNA, 예를 들어 cfDNA를 갖는 것일 수 있다.

[0242] 따라서, 본 발명은 또한 태그부착된 폴리뉴클레오티드의 조성물을 제공한다. 폴리뉴클레오티드는 단편화된 DNA, 예를 들어 cfDNA를 포함할 수 있다. 게놈 내의 맵핑가능한 염기 위치에 맵핑되는 조성물 내의 폴리뉴클레오티드의 세트는 비-특유하게 태그부착될 수 있고, 즉, 상이한 식별자의 수는 적어도 2 및 맵핑가능한 염기 위치에 맵핑되는 폴리뉴클레오티드의 수 미만일 수 있다. 약 10 ng 내지 약 10 ㎍ (예를 들어, 임의의 약 10 ng-

1 μg, 약 10 ng-100 ng, 약 100 ng-10 μg, 약 100 ng-1 μg, 약 1 μg-10 μg)의 조성물은 임의의 2, 5, 10, 50 또는 100개 내지 임의의 100, 1000, 10,000 또는 100,000개의 상이한 식별자를 보유할 수 있다. 예를 들어, 5 내지 100개의 상이한 식별자가 상기 조성물 내에서 폴리뉴클레오티드를 태그부착하기 위해 사용될 수 있다.

[0243] **III. 핵산 서열분석 플랫폼**

[0244] 체액으로부터 세포 유리 폴리뉴클레오티드의 추출 및 단리 후에, 세포 유리 서열은 서열분석될 수 있다. 종종, 서열분석 방법은 전통적인 생거 서열분석이다. 서열분석 방법은 다음을 포함할 수 있고 이로 제한되지 않는다: 고-처리량 서열분석, 피로서열분석, 합성에 의한 서열분석, 단일-분자 서열분석, 나노포어 서열분석, 반도체 서열분석, 라이게이션에 의한 서열분석, 혼성화에 의한 서열분석, RNA-Seq (일루미나), 디지털 유전자 발현 (헬리코스), 차세대 서열분석, 합성에 의한 단일 분자 서열분석 (SMSS) (헬리코스), 대규모 병렬형 서열분석, 클로날 단일 분자 어레이 (솔렉사), 샷건 서열분석, 맥심-길버트 서열분석, 프라이머 워킹, PacBio, SOLiD, 이온 토렌트, 또는 나노포어 플랫폼을 사용한 서열분석 및 관련 기술 분야에 공지된 임의의 다른 서열분석 방법.

[0245] 일부 경우에, 본원에서 설명되는 다양한 종류의 서열분석 반응은 다양한 샘플 처리 유닛을 포함할 수 있다. 샘플 처리 유닛은 다중 레인 (lane), 다중 채널, 다중 웰, 또는 다수의 샘플 세트를 실질적으로 동시에 처리하는 다른 수단을 포함할 수 있고 이로 제한되지 않는다. 추가로, 샘플 처리 유닛은 다중 실행 처리를 동시에 시행할 수 있도록 하는 다중 샘플 챔버를 포함할 수 있다.

[0246] 일부 예에서, 동시 서열분석 반응은 다중 서열분석을 이용하여 수행할 수 있다. 일부 경우에, 세포 유리 폴리뉴클레오티드는 적어도 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100,000회의 서열분석 반응으로 서열분석될 수 있다. 다른 경우에, 세포 유리 폴리뉴클레오티드는 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100,000회 미만의 서열분석 반응으로 서열분석될 수 있다. 서열분석 반응은 순차적으로 또는 동시에 수행할 수 있다. 후속적인 데이터 분석은 모든 또는 일부의 서열분석 반응에 대해 수행할 수 있다. 일부 경우에, 데이터 분석은 적어도 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100,000회의 서열분석 반응에 대해 수행할 수 있다. 다른 경우에, 데이터 분석은 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100,000회 미만의 서열분석 반응에 대해 수행할 수 있다.

[0247] 다른 예에서, 서열 반응의 수는 상이한 양의 게놈에 대한 적용범위를 제공할 수 있다. 일부 경우에, 게놈의 서열 적용범위는 적어도 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99%, 99.9% 또는 100%일 수 있다. 다른 경우에, 게놈의 서열 적용범위는 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99%, 99.9% 또는 100% 미만일 수 있다.

[0248] 일부 예에서, 서열분석은 다양한 상이한 종류의 핵산을 포함할 수 있는 세포 유리 폴리뉴클레오티드에 대해 수행할 수 있다. 핵산은 폴리뉴클레오티드 또는 올리고뉴클레오티드일 수 있다. 핵산은 DNA 또는 RNA, 단일 가닥 또는 이중 가닥 또는 RNA/cDNA 쌍을 포함하나 이에 제한되지는 않는다.

[0249] **IV. 폴리뉴클레오티드 분석 전략**

[0250] 도 8은 초기 유전 물질의 샘플 내에서 폴리뉴클레오티드를 분석하기 위한 전략을 보여주는 도표 (800)이다. 단계 802에서, 초기 유전 물질을 함유하는 샘플을 제공한다. 샘플은 표적 핵산을 낮은 풍부도로 포함할 수 있다. 예를 들어, 정상 또는 야생형 게놈 (예를 들어, 생식계열 게놈)으로부터의 핵산은 유전자 변이를 함유하는 적어도 하나의 다른 게놈, 예를 들어 암 게놈 또는 태아 게놈, 또는 또 다른 종으로부터의 게놈으로부터의 20% 이하, 10% 이하, 5% 이하, 1% 이하, 0.5% 이하 또는 0.1% 이하의 핵산을 또한 포함하는 샘플 내에서 우세할 수 있다. 샘플은 예를 들어, 세포 유리 핵산 또는 핵산을 포함하는 세포를 포함할 수 있다. 초기 유전 물질은 100 ng 이하의 핵산으로 이루어질 수 있다. 이것은 서열분석 또는 유전자 분석 과정에 의한 원래의 폴리뉴클레오티드의 적절한 과다샘플링 (oversampling)에 기여할 수 있다. 별법으로, 샘플은 핵산의 양을 100 ng 이하로 감소시키거나 관심 있는 서열만을 분석하기 위해 선택적으로 풍부화하기 위해서 인공적으로 캐핑 (capping)하거나 또는 병목현상화될 수 있다. 샘플은 참조 서열 내의 각각의 하나 이상의 선택된 위치에 맵핑되는 분자의 서열 판독체를 선택적으로 생산하도록 변형될 수 있다. 100 ng의 핵산의 샘플은 약 30,000개의 인간 반수체 게놈 등가물, 즉, 함께 인간 게놈의 30,000배의 적용범위를 제공하는 분자를 함유할 수 있다.

[0251] 단계 804에서, 초기 유전 물질을 태그부착된 모 폴리뉴클레오티드의 세트로 전환시킨다. 태그부착은 서열분석된 태그를 초기 유전 물질 내의 분자에 부착하는 것을 포함할 수 있다. 서열분석된 태그는 참조 서열 내의 동일한 위치에 맵핑되는 모든 특유한 폴리뉴클레오티드가 특유한 확인 태그를 갖도록 선택될 수 있다. 전환은 높

은 효율, 예를 들어 적어도 50%의 효율로 수행될 수 있다.

[0252] 단계 806에서, 태그부착된 모 폴리뉴클레오티드의 세트가 증폭되어 증폭된 자손 폴리뉴클레오티드의 세트를 생산한다. 증폭은 예를 들어 1,000배일 수 있다.

[0253] 단계 808에서, 증폭된 자손 폴리뉴클레오티드의 세트를 서열분석을 위해 샘플링한다. 샘플링 속도는 생산된 서열 판독체가 (1) 태그부착된 모 폴리뉴클레오티드의 세트 내의 목표하는 수의 특유한 분자를 포함하고 (2) 모 폴리뉴클레오티드의 표적 적용범위 배수, 예를 들어 5 내지 10배의 적용범위에서 태그부착된 모 폴리뉴클레오티드의 세트 내의 특유한 분자를 포함하도록 선택된다.

[0254] 단계 810에서, 서열 판독체의 세트는 붕괴되어, 특유한 태그부착된 모 폴리뉴클레오티드에 상응하는 컨센서스 서열의 세트를 생산한다. 서열 판독체는 분석에 포함되기 위해 품질이 평가될 수 있다. 예를 들어, 품질 대조군 점수를 만족하지 못하는 서열 판독체는 풀로부터 제거할 수 있다. 서열 판독체는 특정 특유한 모 분자로부터 유래된 자손 분자의 판독체를 나타내는 패밀리로 분류될 수 있다. 예를 들어, 증폭된 자손 폴리뉴클레오티드의 패밀리는 단일 모 폴리뉴클레오티드로부터 유래된 이들 증폭된 분자를 구성할 수 있다. 패밀리 내의 자손의 서열을 비교함으로써, 원래의 모 폴리뉴클레오티드의 컨센서스 서열이 추정될 수 있다. 이것은 태그부착된 풀 내에 특유한 모 폴리뉴클레오티드를 나타내는 컨센서스 서열의 세트를 생산한다.

[0255] 단계 812에서, 컨센서스 서열의 세트는 본원에서 설명되는 임의의 분석 방법을 이용하여 분석된다. 예를 들어, 특정 참조 서열 위치에 맵핑되는 컨센서스 서열은 유전자 변이의 경우를 검출하기 위해 분석될 수 있다. 특정 참조 서열에 맵핑되는 컨센서스 서열은 측정되고 대조 샘플에 대해 정규화될 수 있다. 참조 서열에 맵핑되는 분자의 척도는 카피수가 상이하거나 이형접합성이 상실된 게놈 내의 영역을 확인하기 위해 게놈에 걸쳐 비교될 수 있다.

[0256] 도 9는 서열 판독체의 수집물에 의해 나타난 신호로부터 정보를 추출하는 보다 일반적인 방법을 제시하는 도표이다. 상기 방법에서, 증폭된 자손 폴리뉴클레오티드를 서열분석 한 후에, 서열 판독체를 특유한 정체의 분자로부터 증폭된 분자의 패밀리로 분류한다 (910). 상기 분류는 보다 고충실도, 예를 들어 보다 낮은 노이즈 및/또는 왜곡으로 태그부착된 모 폴리뉴클레오티드의 함량을 결정하기 위해 서열 내의 정보를 해석하는 방법에 대한 출발점일 수 있다.

[0257] 서열 판독체의 수집물의 분석을 통해, 서열 판독체가 그로부터 생성된 모 폴리뉴클레오티드 집단에 대해 추정할 수 있다. 상기 추정은 서열분석이 대체로 포괄적증폭된 총 폴리뉴클레오티드의 부분적인 하위세트에 대한 판독만을 수반하기 때문에 유용할 수 있다. 따라서, 모든 모 폴리뉴클레오티드가 서열 판독체의 수집물 내의 적어도 하나의 서열 판독체에 의해 제시될 것이라고 확신할 수 없다.

[0258] 하나의 상기 추정은 원래의 풀 내의 특유한 모 폴리뉴클레오티드의 수이다. 상기 추정은 서열 판독체가 분류될 수 있는 특유한 패밀리의 수 및 각각의 패밀리 내의 서열 판독체의 수를 기초로 하여 이루어질 수 있다. 상기 경우에, 패밀리는 원래의 모 폴리뉴클레오티드까지 역추적가능한 서열 판독체의 수집물을 의미한다. 추정은 잘 공지된 통계적 방법을 이용하여 이루어질 수 있다. 예를 들어, 분류가 각각 하나 또는 몇 개의 자손에 의해 제시되는 많은 패밀리를 생성하는 경우에, 원래의 집단이 서열분석되지 않은 보다 많은 특유한 모 폴리뉴클레오티드를 포함한 것으로 추정할 수 있다. 다른 한편으로, 분류가 각각 많은 자손에 의해 제시되는 몇 개의 패밀리만을 생성하는 경우에, 모 집단 내의 특유한 폴리뉴클레오티드의 대부분은 그 패밀리 내의 적어도 하나의 서열 판독체 군에 의해 제시됨을 추정할 수 있다.

[0259] 또 다른 상기 추정은 폴리뉴클레오티드의 원래의 풀 내의 특정 유전자좌에서의 염기 또는 염기의 서열의 빈도이다. 그러한 추정은 서열 판독체가 분류될 수 있는 특유한 패밀리의 수 및 각각의 패밀리 내의 서열 판독체의 수를 기초로 하여 이루어질 수 있다. 서열 판독체의 패밀리 내의 유전자좌에서 염기 콜을 분석할 때, 신뢰도 점수가 각각의 특정 염기 콜 또는 서열에 배정된다. 이어서, 다수의 패밀리 내의 각각의 염기 콜에 대한 신뢰도 점수를 고려하여, 유전자좌에서 각각의 염기 또는 서열의 빈도를 결정한다.

[0260] **V. 카피수 변이 검출**

[0261] **A. 단일 샘플을 사용하는 카피수 변이 검출**

[0262] 도 1은 단일 대상체에서 카피수 변이의 검출을 위한 전략을 보여주는 도표 (100)이다. 여기에 제시된 바와 같이, 카피수 변이 검출 방법은 다음과 같이 실행할 수 있다. 단계 102에서 세포 유리 폴리뉴클레오티드의 추출 및 단리 후에, 단계 104에서 특유한 단일 샘플을 관련 기술 분야에 공지된 핵산 서열분석 플랫폼에 의해 서열분

석할 수 있다. 상기 단계는 다수의 게놈 단편 서열 판독체를 생성한다. 일부 경우에, 이들 서열 판독체는 바코드 정보를 함유할 수 있다. 다른 예에서, 바코드는 이용되지 않는다. 서열분석 후에, 판독체에 품질 점수를 배정한다. 품질 점수는 이들 판독체가 역치에 기반한 후속적인 분석에 유용할 수 있는지를 나타내는 판독체의 표시일 수 있다. 일부 경우에, 일부 판독체는 후속적인 맵핑 단계를 수행하기 위해 충분한 품질 또는 길이를 갖지 않는다. 적어도 90%, 95%, 99%, 99.9%, 99.99% 또는 99.999%의 품질 점수를 갖는 서열분석 판독체를 데이터로부터 여과 제거할 수 있다. 다른 경우에, 90%, 95%, 99%, 99.9%, 99.99% 또는 99.999% 미만의 품질 점수가 배정된 서열분석 판독체를 데이터 세트로부터 여과 제거할 수 있다. 단계 106에서, 명시된 품질 점수 역치를 충족하는 게놈 단편 판독체를 참조 게놈, 또는 카피수 변이를 함유하지 않은 것으로 알려진 주형 서열에 맵핑한다. 맵핑 정렬 후에, 서열 판독체에 맵핑 점수를 배정한다. 맵핑 점수는 각각의 위치가 특유하게 맵핑 가능한지 또는 그렇지 않은지 나타내는, 참조 서열에 다시 맵핑된 표시 또는 판독체일 수 있다. 이 경우에, 판독체는 카피수 변이 분석에 비관련된 서열일 수 있다. 예를 들어, 일부 서열 판독체는 오염물 폴리뉴클레오티드로부터 기원할 수 있다. 적어도 90%, 95%, 99%, 99.9%, 99.99% 또는 99.999%의 맵핑 점수를 갖는 서열분석 판독체는 데이터 세트로부터 여과 제거할 수 있다. 다른 경우에, 90%, 95%, 99%, 99.9%, 99.99% 또는 99.999% 미만의 품질 점수가 배정된 서열분석 판독체를 데이터 세트로부터 여과 제거할 수 있다.

[0263] 데이터 여과 및 맵핑 후에, 다수의 서열 판독체는 적용범위의 염색체 영역을 생성한다. 단계 108에서, 이들 염색체 영역을 가변 길이 윈도우 또는 빈으로 나눌 수 있다. 윈도우 또는 빈은 적어도 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb 또는 1000 kb일 수 있다. 윈도우 또는 빈은 또한 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb 또는 1000 kb까지의 염기를 가질 수 있다. 윈도우 또는 빈은 또한 약 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb 또는 1000 kb일 수 있다.

[0264] 단계 110에서 적용범위 정규화를 위해, 각각의 윈도우 또는 빈은 대략 동일한 수의 맵핑가능한 염기를 함유하도록 선택된다. 일부 경우에, 염색체 영역 내의 각각의 윈도우 또는 빈은 정확한 수의 맵핑가능한 염기를 함유할 수 있다. 다른 경우에, 각각의 윈도우 또는 빈은 상이한 수의 맵핑가능한 염기를 함유할 수 있다. 추가로, 각각의 윈도우 또는 빈은 인접한 윈도우 또는 빈과 비-겹침일 수 있다. 다른 경우에, 윈도우 또는 빈은 또 다른 인접한 윈도우 또는 빈과 겹칠 수 있다. 일부 경우에, 윈도우 또는 빈은 적어도 1 bp, 2 bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, 또는 1000 bp 겹칠 수 있다. 다른 경우에, 윈도우 또는 빈은 1 bp, 2 bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, 또는 1000 bp까지 겹칠 수 있다. 일부 경우에, 윈도우 또는 빈은 약 1 bp, 2 bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, 또는 1000 bp 겹칠 수 있다.

[0265] 일부 경우에, 각각의 윈도우 영역은 대략 동일한 수의 특유하게 맵핑가능한 염기를 함유하도록 하는 크기일 수 있다. 윈도우 영역을 포함하는 각각의 염기의 맵핑가능성이 결정되고, 각각의 파일에 대한 참조물에 다시 맵핑된 참조물로부터의 판독체의 표시를 함유하는 맵핑가능성 파일을 생성하기 위해 사용한다. 맵핑가능성 파일은 모든 위치마다 1개의 열 (row)을 함유하고, 이것은 각각의 위치가 특유하게 맵핑가능한지 또는 그렇지 않은지 나타낸다.

[0266] 추가로, 서열분석하기 어려운, 또는 실질적으로 높은 GC 편향을 함유하는 것으로 게놈 전체에 걸쳐 알려진 미리 규정된 윈도우는 데이터 세트로부터 여과될 수 있다. 예를 들어, 염색체의 동원체 주위에 놓인 것으로 알려진 영역 (즉, 동원체 DNA)은 위-양성 결과를 생성할 수 있는 고도로 반복적인 서열을 함유하는 것으로 알려져 있다. 이들 영역을 여과 제거할 수 있다. 게놈의 다른 영역, 예컨대 특이하게 높은 농도의 다른 고도로 반복적인 서열을 함유하는 영역, 예컨대 미소부수체 DNA를 데이터 세트로부터 여과 제거할 수 있다.

[0267] 분석된 윈도우의 수는 또한 다양할 수 있다. 일부 경우에, 적어도 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5,000, 10,000, 20,000, 50,000 또는 100,000개의 윈도우가 분석된다. 다른 경우에, 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5,000, 10,000, 20,000, 50,000 또는 100,000개까지의 윈도우가 분석된다.

[0268] 세포 유리 폴리뉴클레오티드 서열로부터 유래된 예시적인 게놈에 대해, 다음 단계는 각각의 윈도우 영역에 대한 판독체 적용범위를 결정하는 것을 포함한다. 이것은 바코드를 갖는 판독체 또는 바코드를 갖지 않는 판독체를 사용하여 수행할 수 있다. 바코드를 갖지 않는 경우에, 선행 맵핑 단계는 상이한 염기 위치의 적용범위를 제공할 것이다. 충분한 맵핑 및 품질 점수를 갖고 여과되지 않는 염색체 윈도우 내에 있는 서열 판독체가 계수될 수 있다. 적용범위 판독체의 수에는 각각의 맵핑가능 위치마다 점수가 배정될 수 있다. 바코드를 포함하는 경우에는, 이들은 모두 샘플 모 분자로부터 유래되므로, 동일한 바코드, 물리적 특성 또는 이 2가지의 조합을 갖

는 모든 서열이 1개의 판독체로 붕괴될 수 있다. 상기 단계는 임의의 선행하는 단계, 예컨대 증폭을 수반한 단계 동안 도입될 수 있는 편향을 감소시킨다. 예를 들어, 하나의 분자가 10배 증폭되지만 다른 분자는 1000배 증폭될 경우, 각각의 분자는 붕괴 후에 1회만 제시되고, 따라서 불균등한 증폭의 효과를 무효화한다. 특유한 바코드를 갖는 판독체만이 각각의 맵핑가능 위치에 대해 계수될 수 있고, 배정된 점수에 영향을 미친다.

[0269] 컨센서스 서열은 관련 기술 분야에 공지된 임의의 방법에 의해 서열 판독체의 패밀리로부터 생성될 수 있다. 상기 방법은 예를 들어, 디지털 통신 이론, 정보 이론, 또는 생물 정보공학으로부터 유래된 선형 또는 비-선형 컨센서스 서열 구축 방법 (예컨대 보팅 (voting), 평균화 (averaging), 통계적, 최대 사후 (posteriori) 또는 최대 가능도 검출, 동적 프로그래밍, 베이지안, 은닉 마르코프 또는 서포트 벡터 머신 방법 등)을 포함한다.

[0270] 서열 판독체 적용범위가 결정된 후에, 확률적 모델링 알고리즘을 적용하여, 각각의 윈도우 영역에 대한 정규화된 핵산 서열 판독체 적용범위를 별개의 카피수 상태로 전환시킨다. 일부 경우에, 상기 알고리즘은 은닉 마르코프 모델, 동적 프로그래밍, 서포트 벡터 머신, 베이지안 네트워크, 격자 해독, 비터비 해독, 기대값 최대화, 칼만 여과 방법 및 신경망 중 하나 이상을 포함할 수 있다.

[0271] 단계 112에서, 각각의 윈도우 영역의 별개의 카피수 상태는 염색체 영역 내에서 카피수 변이를 확인하기 위해 이용될 수 있다. 일부 경우에, 카피수 변이 상태의 존재 또는 부재를 보고하기 위해 동일한 카피수를 갖는 모든 인접한 윈도우 영역은 하나의 절편으로 병합될 수 있다. 일부 경우에, 다양한 윈도우는 다른 절편과 병합되기 전에 여과될 수 있다.

[0272] 단계 114에서, 카피수 변이는 게놈 내의 다양한 위치, 및 각각의 개별 위치에서 카피수 변이의 상응하는 증가 또는 감소 또는 유지를 나타내는 그래프로서 보고될 수 있다. 추가로, 카피수 변이는 얼마나 많은 질환 물질 (또는 카피수 변이를 갖는 핵산)이 세포 유리 폴리뉴클레오티드 샘플 내에 존재하는지 나타내는 백분율 점수를 보고하기 위해 사용될 수 있다.

[0273] 카피수 변이를 결정하는 하나의 방법을 도 10에 제시한다. 상기 방법에서, 서열 판독체를 단일 모 폴리뉴클레오티드로부터 생성된 패밀리로 분류한 후 (1010), 예를 들어 각각의 다수의 상이한 참조 서열 위치에 맵핑되는 패밀리의 수를 결정함으로써 패밀리를 정량한다. CNV는 각각의 다수의 상이한 유전자좌에서 패밀리의 정량적 척도를 비교함으로써 직접 결정될 수 있다 (1016b). 별법으로, 예를 들어 상기 논의한 바와 같이 패밀리의 정량적 척도 및 각각의 패밀리 내의 패밀리 구성원의 정량적 척도 둘 모두를 이용하여 태그부착된 모 폴리뉴클레오티드의 집단 내의 패밀리의 정량적 척도를 추정할 수 있다. 이어서, CNV는 다수의 유전자좌에서 추정된 양의 척도를 비교함으로써 결정될 수 있다. 다른 실시양태에서, 원래의 양의 유사한 추정이 이루어진 후, 서열분석 과정 동안 표상적 편향, 예컨대 GC 편향 등에 대한 정규화가 수행될 수 있는 혼성 (hybrid) 방안을 실시할 수 있다.

[0274] **B. 짝을 이룬 샘플을 사용한 카피수 변이 검출**

[0275] 짝을 이룬 샘플 카피수 변이 검출은 본원에서 설명되는 단일 샘플 방안과 많은 단계 및 파라미터를 공유한다. 그러나, 도 2의 200에 도시된 바와 같이, 짝을 이룬 샘플을 이용하는 카피수 변이 검출은 서열 적용범위를 게놈의 예측된 맵핑가능성에 비교하기보다는 대조 샘플에 비교할 것을 필요로 한다. 상기 방안은 윈도우에 걸친 정규화를 도울 수 있다.

[0276] 도 2는 짝을 이룬 대상체에서 카피수 변이의 검출을 위한 전략을 보여주는 도표 (200)이다. 여기에 제시된 바와 같이, 카피수 변이 검출 방법은 다음과 같이 실행할 수 있다. 단계 204에서, 특유한 단일 샘플은 단계 202에서 샘플의 추출 및 단리 후에 관련 기술 분야에 공지된 핵산 서열분석 플랫폼에 의해 서열분석할 수 있다. 상기 단계는 다수의 게놈 단편 서열 판독체를 생성한다. 추가로, 샘플 또는 대조 샘플을 또 다른 대상체로부터 채취한다. 일부 경우에, 대조군 대상체는 질환이 있는 것으로 알려지지 않은 대상체일 수 있는 반면, 다른 대상체는 특정 질환이 있거나 그의 위험이 있을 수 있다. 일부 경우에, 이들 서열 판독체는 바코드 정보를 함유할 수 있다. 다른 예에서, 바코드는 이용되지 않는다. 서열분석 후에, 판독체에 품질 점수를 배정한다. 일부 경우에, 일부 판독체는 후속적인 맵핑 단계를 수행하기에 충분한 품질 또는 길이를 갖지 않는다. 적어도 90%, 95%, 99%, 99.9%, 99.99% 또는 99.999%의 품질 점수를 갖는 서열분석 판독체를 데이터 세트로부터 여과 제거할 수 있다. 다른 경우에, 90%, 95%, 99%, 99.9%, 99.99% 또는 99.999% 미만의 품질 점수가 배정된 서열분석 판독체를 데이터 세트로부터 여과 제거할 수 있다. 단계 206에서, 명시된 품질 점수 역치를 충족하는 게놈 단편 판독체를 참조 게놈, 또는 카피수 변이를 함유하지 않은 것으로 알려진 주형 서열에 맵핑한다. 맵핑 정렬 후에, 서열 판독체에 맵핑 점수를 배정한다. 이 경우에, 판독체는 카피수 변이 분석에 비관련된 서열일 수 있

다. 예를 들어, 일부 서열 판독체는 오염물 폴리뉴클레오티드로부터 기원할 수 있다. 적어도 90%, 95%, 99%, 99.9%, 99.99% 또는 99.999%의 맵핑 점수를 갖는 서열분석 판독체를 데이터 세트로부터 여과 제거할 수 있다. 다른 경우에, 90%, 95%, 99%, 99.9%, 99.99% 또는 99.999% 미만의 맵핑 점수가 배정된 서열분석 판독체를 데이터 세트로부터 여과 제거할 수 있다.

[0277] 데이터 여과 및 맵핑 후에, 다수의 서열 판독체는 각각의 시험 및 대조군 대상체에 대한 적용범위의 염색체 영역을 생성한다. 단계 208에서, 이들 염색체 영역은 가변 길이 윈도우 또는 빈으로 나누어질 수 있다. 윈도우 또는 빈은 적어도 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb 또는 1000 kb일 수 있다. 윈도우 또는 빈은 또한 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb 또는 1000 kb 미만일 수 있다.

[0278] 단계 210에서 적용범위 정규화를 위해, 각각의 윈도우 또는 빈은 각각의 시험 및 대조군 대상체에 대해 대략 동일한 수의 맵핑가능한 염기를 함유하도록 선택된다. 일부 경우에, 염색체 영역 내의 각각의 윈도우 또는 빈은 정확한 수의 맵핑가능한 염기를 함유할 수 있다. 다른 경우에, 각각의 윈도우 또는 빈은 상이한 수의 맵핑가능한 염기를 함유할 수 있다. 추가로, 각각의 윈도우 또는 빈은 인접한 윈도우 또는 빈과 비-겹침일 수 있다. 다른 경우에, 윈도우 또는 빈은 또 다른 인접한 윈도우 또는 빈과 겹칠 수 있다. 일부 경우에, 윈도우 또는 빈은 적어도 1 bp, 2 bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, 또는 1000 bp 겹칠 수 있다. 다른 경우에, 윈도우 또는 빈은 1 bp, 2 bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, 또는 1000 bp 미만으로 겹칠 수 있다.

[0279] 일부 경우에, 각각의 윈도우 영역은 각각의 시험 및 대조군 대상체에 대해 대략 동일한 수의 특유하게 맵핑가능한 염기를 함유하도록 하는 크기이다. 윈도우 영역을 포함하는 각각의 염기의 맵핑가능성을 결정하고, 각각의 파일에 대한 참조물에 다시 맵핑된 참조물로부터의 판독체의 표시를 함유하는 맵핑가능성 파일을 생성하기 위해 사용한다. 맵핑가능성 파일은 모든 위치마다 1개의 열을 함유하고, 이것은 각각의 위치가 특유하게 맵핑가능한지 또는 그렇지 않은지 나타낸다.

[0280] 추가로, 서열분석하기 어려운, 또는 실질적으로 높은 GC 편향을 함유하는 것으로 게놈 전체에 걸쳐 알려진 미리 규정된 윈도우는 데이터 세트로부터 여과될 수 있다. 예를 들어, 염색체의 동원체 주위에 놓인 것으로 알려진 영역 (즉, 동원체 DNA)은 위-양성 결과를 생성할 수 있는 고도로 반복적인 서열을 함유하는 것으로 알려져 있다. 이들 영역을 여과 제거할 수 있다. 게놈의 다른 영역, 예컨대 특이하게 높은 농도의 다른 고도로 반복적인 서열을 함유하는 영역, 예컨대 미소부수체 DNA를 데이터 세트로부터 여과 제거할 수 있다.

[0281] 분석된 윈도우의 수는 또한 다양할 수 있다. 일부 경우에, 적어도 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5,000, 10,000, 20,000, 50,000 또는 100,000개의 윈도우가 분석된다. 다른 경우에, 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5,000, 10,000, 20,000, 50,000 또는 100,000개 미만의 윈도우가 분석된다.

[0282] 세포 유리 폴리뉴클레오티드 서열로부터 유래된 예시적인 게놈에 대해, 다음 단계는 각각의 시험 및 대조군 대상체에 대해 각각의 윈도우 영역에 대한 판독체 적용범위를 결정하는 것을 포함한다. 이것은 바코드를 갖는 판독체 또는 바코드를 갖지 않는 판독체를 사용하여 수행할 수 있다. 바코드를 갖지 않는 경우에, 선행 맵핑 단계는 상이한 염기 위치의 적용범위를 제공할 것이다. 충분한 맵핑 및 품질 점수를 갖고, 여과되지 않는 염색체 윈도우 내에 있는 서열 판독체가 계수될 수 있다. 적용범위 판독체의 수에는 각각의 맵핑가능 위치마다 점수가 배정될 수 있다. 바코드를 포함한 경우에는, 이들은 모두 샘플 모 분자로부터 유래되므로, 동일한 바코드를 갖는 모든 서열이 1개의 판독체로 붕괴될 수 있다. 상기 단계는 임의의 선행하는 단계, 예컨대 증폭을 수반한 단계 동안 도입될 수도 있는 편향을 감소시킨다. 특유한 바코드를 갖는 판독체만이 각각의 맵핑가능 위치에 대해 계수될 수 있고, 배정된 점수에 영향을 미친다. 상기 이유 때문에, 바코드 라이게이션 단계가 최소량의 편향을 생성하기 위해 최적화된 방식으로 수행되는 것이 중요하다.

[0283] 각각의 윈도우에 대한 핵산 판독체 적용범위를 결정하는데 있어서, 각각의 윈도우의 적용범위는 그 샘플의 평균 적용범위에 의해 정규화될 수 있다. 그러한 방안을 이용하여, 유사한 조건 하에 시험 대상체 및 대조군을 모두 서열분석하는 것이 바람직할 수 있다. 이어서, 각각의 윈도우에 대한 판독체 적용범위는 유사한 윈도우를 가로질러 비율로서 표현될 수 있다.

[0284] 시험 대상체의 각각의 윈도우에 대한 핵산 판독체 적용범위 비율은 시험 샘플의 각각의 윈도우 영역의 판독체 적용범위를 대조 샘플의 상응하는 윈도우 영역의 판독체 적용범위로 나누어 결정할 수 있다.

[0285] 서열 판독체 적용범위 비율이 결정된 후에, 확률적 모델링 알고리즘을 적용하여, 각각의 윈도우 영역에 대한 정

규화된 비율을 별개의 카피수 상태로 전환시킨다. 일부 경우에, 상기 알고리즘은 은닉 마르코프 모델을 포함할 수 있다. 다른 경우에, 확률적 모델은 동적 프로그래밍, 서포트 벡터 머신, 베이지안 모델링, 확률적 모델링, 격자 해독, 비터비 해독, 기대값 최대화, 칼만 여과 방법, 또는 신경망을 포함할 수 있다.

[0286] 단계 212에서, 각각의 윈도우 영역의 별개의 카피수 상태는 염색체 영역 내에서 카피수 변이를 확인하기 위해 이용될 수 있다. 일부 경우에, 카피수 변이 상태의 존재 또는 부재를 보고하기 위해 동일한 카피수를 갖는 모든 인접한 윈도우 영역은 하나의 절편으로 병합될 수 있다. 일부 경우에, 다양한 윈도우는 다른 절편과 병합되기 전에 여과될 수 있다.

[0287] 단계 214에서, 카피수 변이는 게놈 내의 다양한 위치, 및 각각의 개별 위치에서 카피수 변이의 상응하는 증가 또는 감소 또는 유지를 나타내는 그래프로서 보고될 수 있다. 추가로, 카피수 변이는 얼마나 많은 질환 물질이 세포 유리 폴리뉴클레오티드 샘플 내에 존재하는지 나타내는 백분율 점수를 보고하기 위해 사용될 수 있다.

[0288] **VI. 희귀 돌연변이 검출**

[0289] 희귀 돌연변이 검출은 두 카피수 변이 방안과 유사한 특징을 공유한다. 그러나, 도 3의 300에 도시된 바와 같이, 희귀 돌연변이 검출에서는 서열 적용범위를 게놈의 상대적인 맵핑가능성에 비교하기보다는 대조 샘플 또는 참조 서열에 대한 비교를 이용한다. 이 방안은 윈도우에 걸친 정규화를 도울 수 있다.

[0290] 일반적으로, 단계 302에서 정제되고 단리된 게놈 또는 트랜스크립톰의 선택적으로 풍부화된 영역 상에서 희귀 돌연변이 검출을 수행할 수 있다. 본원에서 설명된 바와 같이, 유전자, 종양유전자, 종양 억제 유전자, 프로모터, 조절 서열 요소, 비-코딩 영역, miRNA, snRNA 등을 포함할 수 있고 이로 제한되지 않는 특수한 영역은 세포 유리 폴리뉴클레오티드의 총 집단으로부터 선택적으로 증폭할 수 있다. 이것은 본원에 설명된 바와 같이 수행할 수 있다. 한 예에서, 개별 폴리뉴클레오티드 서열에 대해 바코드 표지를 사용하거나 사용하지 않는 다중 서열분석을 이용할 수 있다. 다른 예에서, 서열분석은 관련 기술 분야에 공지된 임의의 핵산 서열분석 플랫폼을 이용하여 수행할 수 있다. 이 단계는 단계 304에서와같이 다수의 게놈 단편 서열 판독체를 생성한다. 추가로, 참조 서열은 또 다른 대상체로부터 채취한 대조 샘플로부터 얻는다. 일부 경우에, 대조군 대상체는 공지의 유전자 이상 또는 질환을 갖지 않은 것으로 알려진 대상체일 수 있다. 일부 경우에, 이들 서열 판독체는 바코드 정보를 함유할 수 있다. 다른 예에서, 바코드가 이용되지 않는다. 서열분석 후에, 판독체에 품질 점수를 배정한다. 품질 점수는 이들 판독체가 역치에 기반한 후속적인 분석에서 유용할 수 있는지 여부를 지시하는 판독체의 표시일 수 있다. 일부 경우에, 일부 판독체는 후속적인 맵핑 단계를 수행하기 위해 충분한 품질 또는 길이의 것이 아니다. 적어도 90%, 95%, 99%, 99.9%, 99.99% 또는 99.999%의 품질 점수를 갖는 서열분석 판독체를 데이터 세트로부터 여과 제거할 수 있다. 다른 경우에, 적어도 90%, 95%, 99%, 99.9%, 99.99% 또는 99.999%의 품질 점수가 배정된 서열분석 판독체를 데이터 세트로부터 여과 제거할 수 있다. 단계 306에서, 명시된 품질 점수 역치를 충족하는 게놈 단편 판독체를 참조 게놈, 또는 희귀 돌연변이를 함유하지 않은 것으로 알려진 참조 서열에 맵핑한다. 맵핑 정렬 후에, 서열 판독체에 맵핑 점수를 배정한다. 맵핑 점수는 각각의 위치가 특유하게 맵핑가능한지 또는 그렇지 않은지 나타내는, 참조 서열에 다시 맵핑된 표시 또는 판독체일 수 있다. 이 경우에, 판독체는 희귀 돌연변이 분석에 비관련된 서열일 수 있다. 예를 들어, 일부 서열 판독체는 오염물 폴리뉴클레오티드로부터 기원할 수 있다. 적어도 90%, 95%, 99%, 99.9%, 99.99% 또는 99.999%의 맵핑 점수를 갖는 서열분석 판독체를 데이터 세트로부터 여과 제거할 수 있다. 다른 경우에, 90%, 95%, 99%, 99.9%, 99.99% 또는 99.999% 미만의 맵핑 점수가 배정된 서열분석 판독체를 데이터 세트로부터 여과 제거할 수 있다.

[0291] 각각의 맵핑가능한 염기에 대해, 맵핑가능성에 대한 최소 역치를 충족하지 못하는 염기, 또는 낮은 품질 염기는 참조 서열에서 발견되는 상응하는 염기로 교체될 수 있다.

[0292] 데이터 여과 및 맵핑 후에, 대상체 및 참조 서열로부터 얻어진 서열 판독체 사이에서 발견된 변이체 염기를 분석한다.

[0293] 세포 유리 폴리뉴클레오티드 서열로부터 유래된 예시적인 게놈에 대해, 다음 단계는 각각의 맵핑가능한 염기 위치에 대한 판독체 적용범위를 결정하는 것을 포함한다. 이것은 바코드를 갖는 판독체 또는 바코드를 갖지 않는 판독체를 사용하여 수행할 수 있다. 바코드를 갖지 않는 경우에, 선행 맵핑 단계는 상이한 염기 위치의 적용범위를 제공할 것이다. 충분한 맵핑 및 품질 점수를 갖는 서열 판독체가 계수될 수 있다. 적용범위 판독체의 수에는 각각의 맵핑가능 위치마다 점수가 배정될 수 있다. 바코드를 포함한 경우에는, 이들은 모두 샘플 모 분자로부터 유래되므로 동일한 바코드를 갖는 모든 서열이 1개의 컨센서스 판독체로 붕괴될 수 있다. 각각의 염기에 대한 서열은 그 특이적 위치에 대한 가장 우세한 뉴클레오티드 판독체로서 정렬된다. 또한, 각각의 위치에

서 동시 정량을 유도하기 위해 각각의 위치에서 특유한 분자의 수를 계수될 수 있다. 상기 단계는 임의의 선행하는 단계, 예컨대 증폭을 수반한 단계 동안 도입될 수도 있는 편향을 감소시킨다. 특유한 바코드를 갖는 판독체만이 각각의 맵핑가능 위치에 대해 계수될 수 있고, 배정된 점수에 영향을 미친다.

[0294] 일단 판독체 적용범위가 확정될 수 있고, 각각의 판독체 내에서 대조군 서열에 비해 변이체 염기가 확인되면, 변이체를 함유하는 판독체의 수를 판독체의 총수로 나누어 변이체 염기의 빈도를 계산할 수 있다. 이것은 게놈 내의 각각의 맵핑가능 위치에 대한 비율로서 표현될 수 있다.

[0295] 각각의 염기 위치에 대해, 4개의 모든 뉴클레오티드, 즉, 시토신, 구아닌, 티민, 아데닌의 빈도를 참조 서열에 비해 분석한다. 확률적 또는 통계적 모델링 알고리즘을 적용하여, 각각의 염기 변이체에 대한 빈도 상태를 반영하도록 각각의 맵핑가능 위치에 대한 정규화된 비율을 전환시켰다. 일부 경우에, 상기 알고리즘은 은닉 마르코프 모델, 동적 프로그래밍, 서포트 벡터 머신, 베이지안 또는 확률적 모델링, 격자 해독, 비터비 해독, 기대값 최대화, 칼만 여과 방법, 및 신경망 중 하나 이상을 포함할 수 있다.

[0296] 단계 312에서, 각각의 염기 위치의 별개의 희귀 돌연변이 상태를 이용하여, 참조 서열의 기준선에 비교하여 높은 빈도의 변이를 갖는 염기 변이체를 확인할 수 있다. 일부 경우에, 기준선은 적어도 0.0001%, 0.001%, 0.01%, 0.1%, 1.0%, 2.0%, 3.0%, 4.0%, 5.0%, 10%, 또는 25%의 빈도를 나타낼 수 있다. 다른 경우에, 기준선은 적어도 0.0001%, 0.001%, 0.01%, 0.1%, 1.0%, 2.0%, 3.0%, 4.0%, 5.0%. 10%, 또는 25%의 빈도를 나타낼 수 있다. 일부 경우에, 희귀 돌연변이의 존재 또는 부재를 보고하기 위해 염기 변이체 또는 돌연변이를 갖는 모든 인접한 염기 위치는 절편으로 병합될 수 있다. 일부 경우에, 다양한 위치는 다른 절편과 병합되기 전에 여과될 수 있다.

[0297] 각각의 염기 위치에 대한 변이의 빈도를 계산한 후에, 참조 서열에 비교할 때 대상체로부터 유래된 서열 내에 특이적 위치에 대해 최대 편차를 갖는 변이체가 희귀 돌연변이로서 확인된다. 일부 경우에, 희귀 돌연변이는 암 돌연변이일 수 있다. 다른 경우에, 희귀 돌연변이는 질환 상태와 상관될 수 있다.

[0298] 희귀 돌연변이 또는 변이체는 단일 염기 치환, 또는 작은 삽입-결실, 염기변환, 전위, 역위, 결실, 말단절단 또는 유전자 말단절단을 포함하지만 이로 제한되지 않는 유전자 이상을 포함할 수 있다. 일부 경우에, 희귀 돌연변이의 길이는 최대 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 또는 20개 뉴클레오티드일 수 있다. 다른 경우에, 희귀 돌연변이의 길이는 적어도 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 또는 20개 뉴클레오티드일 수 있다.

[0299] 단계 314에서, 돌연변이의 존재 또는 부재는 게놈 내의 다양한 위치, 및 각각의 개별 위치에서 돌연변이의 빈도의 상응하는 증가 또는 감소 또는 유지를 나타내는 그래프 형태로 반영될 수 있다. 추가로, 희귀 돌연변이는 얼마나 많은 질환 물질이 세포 유리 폴리뉴클레오티드 샘플 내에 존재하는지 나타내는 백분율 점수를 보고하기 위해 사용될 수 있다. 신뢰도 점수는 비-질환 참조 서열 내의 보고된 위치에서 전형적인 변이의 공지의 통계학을 고려하여 각각의 검출된 돌연변이를 동반할 수 있다. 돌연변이는 대상체 내의 풍부도 순서로 순위 결정되거나, 임상적으로 작용가능한 중요성에 의해 순위 결정될 수 있다.

[0300] 도 11은 집단 폴리뉴클레오티드 내에서 특정 유전자좌에서 염기 또는 염기의 서열의 빈도의 추정 방법을 보여준다. 서열 판독체는 원래의 태그부착된 폴리뉴클레오티드로부터 생성된 패밀리로 분류된다 (1110). 각각의 패밀리에 대해, 유전자좌에서 하나 이상의 염기에는 각각 신뢰도 점수가 배정된다. 신뢰도 점수는 임의의 많은 공지된 통계적 방법에 의해 배정될 수 있고, 적어도 부분적으로, 패밀리에 속하는 서열 판독체 사이에서 염기가 나타나는 빈도에 기반할 수 있다 (1112). 예를 들어, 신뢰도 점수는 서열 판독체 사이에서 염기가 나타나는 빈도일 수 있다. 또 다른 예로서, 각각의 패밀리에 대해, 은닉 마르코프 모델을 작성할 수 있어서, 단일 패밀리 내에 특정 염기의 발생의 빈도에 기반하여 최대 가능성 또는 최대 사후 결정이 이루어질 수 있다. 상기 모델의 일부로서, 특정 결정에 대한 오류의 확률 및 결과의 신뢰도 점수가 또한 출력될 수 있다. 이어서, 원래의 집단 내의 염기의 빈도가 패밀리 사이에서 신뢰도 점수에 기반하여 배정될 수 있다 (1114).

[0301] **VII. 용도**

[0302] **A. 암의 조기 검출**

[0303] 본원에서 설명되는 방법 및 시스템을 이용하여 수많은 암을 검출할 수 있다. 대부분의 세포와 마찬가지로, 암 세포는 오래된 세포가 죽고, 보다 새로운 세포로 교체되는 순환 (turnover) 속도를 특징으로 할 수 있다. 일반적으로, 제시된 대상체의 혈관계와 접촉하는 죽은 세포는 DNA 또는 DNA의 단편을 혈류 내로 방출할 수 있다. 이것은 또한 질환의 다양한 병기 동안 암 세포의 경우에도 사실이다. 암 세포는 또한 질환의 병기에 따라, 다양한 유전자 이상, 예컨대 카피수 변이 및 희귀 돌연변이를 특징으로 할 수 있다. 이 현상은 본원에서 설명하

는 방법 및 시스템을 이용하여 개체에서 암의 존재 또는 부재를 검출하기 위해 이용될 수 있다.

[0304] 예를 들어, 암의 위험이 있는 대상체로부터의 혈액을 채취하고, 세포 유리 폴리뉴클레오티드의 집단을 생성하기 위해 본원에서 설명되는 바와 같이 제조할 수 있다. 한 예에서, 이것은 세포 유리 DNA일 수 있다. 본 개시내용의 시스템 및 방법은 존재하는 특정 암에 존재할 수 있는 희귀 돌연변이 또는 카피수 변이를 검출하기 위해 사용될 수 있다. 방법은 질환의 증상 또는 다른 징표가 없더라도 신체에서 암성 세포의 존재를 검출하는 것을 도울 수 있다.

[0305] 검출할 수 있는 암의 종류 및 수는 혈액암, 뇌암, 폐암, 피부암, 코암, 인후암, 간암, 뼈암, 림프종, 췌장암, 피부암, 대장암, 직장암, 갑상선암, 방광암, 신장암, 구강암, 위암, 고체 상태 종양, 비균질 종양, 균질 종양 등을 포함할 수 있고 이로 제한되지 않는다.

[0306] 암의 조기 검출에서, 암을 검출하기 위해 희귀 돌연변이 검출 또는 카피수 변이 검출을 비롯한 본원에서 설명되는 임의의 시스템 또는 방법을 이용할 수 있다. 이들 시스템 및 방법은 암을 일으키거나 암으로부터 생성될 수 있는 임의의 많은 유전자 이상을 검출하기 위해 사용될 수 있다. 이들은 돌연변이, 희귀 돌연변이, 삽입-결실, 카피수 변이, 염기변환, 전위, 역위, 결실, 이수성, 부분적 이수성, 배수성, 염색체 불안정성, 염색체 구조 변경, 유전자 융합, 염색체 융합, 유전자 말단절단, 유전자 증폭, 유전자 중복, 염색체 병변, DNA 병변, 핵산 화학적 변형의 비정상적인 변화, 후성적 패턴의 비정상적인 변화, 핵산 메틸화 감염의 비정상적인 변화 및 암을 포함할 수 있고 이로 제한되지 않는다.

[0307] 추가로, 본원에서 설명되는 시스템 및 방법은 또한 특정 암을 특성화하는 것을 돕기 위해 사용될 수 있다. 본 개시내용의 시스템 및 방법으로부터 생성된 유전자 데이터는 의사가 특수한 형태의 암을 보다 잘 특성화하도록 도울 수 있다. 종종, 암은 조성 및 병기 모두에서 비균질이다. 유전자 프로파일 데이터는 특수한 하위형의 암의 특성화를 허용할 수 있고, 이것은 그 특수한 하위형의 진단 또는 치료에 중요할 수 있다. 상기 정보는 또한 대상체 또는 의사에게 특수한 종류의 암의 예측에 관한 실마리를 제공할 수 있다.

[0308] **B. 암 모니터링 및 예측**

[0309] 본원에 제공되는 시스템 및 방법은 특정 대상체에서 이미 공지된 암 또는 다른 질환을 모니터링하기 위해 사용될 수 있다. 이것은 대상체 또는 의사가 질환의 진행에 따라 치료 선택권을 채택하도록 허용할 수 있다. 본 예에서, 본원에서 설명되는 시스템 및 방법은 질환의 과정의 특정 대상체의 유전자 프로파일을 구성하기 위해 이용될 수 있다. 몇몇 예에서, 암은 진행하고 보다 침습성이고 유전학적으로 불안정해질 수 있다. 다른 예에서, 암은 양성으로 남거나, 불활성이거나, 잠재성이거나, 완화될 수 있다. 본 개시내용의 시스템 및 방법은 질환 진행, 완화 또는 재발을 결정하는데 유용할 수 있다.

[0310] 또한, 본원에서 설명되는 시스템 및 방법은 특정 치료 선택권의 효능을 결정하는데 유용할 수 있다. 한 예에서, 치료가 성공적이면 보다 많은 암이 죽고 DNA을 방출할 수 있으므로, 성공적인 치료 선택권은 대상체의 혈액 내에서 검출된 카피수 변이 또는 희귀 돌연변이의 양을 실제로 증가시킬 수 있다. 다른 예에서, 이것은 일어나지 않을 수 있다. 또 다른 예에서, 아마도 특정 치료 선택권은 시간 경과에 따른 암의 유전자 프로파일과 상관될 수 있다. 이 상관성은 요법을 선택하는데 유용할 수 있다. 추가로, 암이 치료 후에 완화되는 것으로 관찰되면, 본원에서 설명되는 시스템 및 방법은 잔류 질환 또는 질환의 재발을 모니터링하는데 유용할 수 있다.

[0311] 예를 들어, 역치 수준에서 시작하는 빈도의 범위로 발생하는 돌연변이는 대상체, 예를 들어 환자로부터의 샘플 내에서 DNA로부터 결정될 수 있다. 돌연변이는 예를 들어, 암 관련 돌연변이일 수 있다. 빈도는 예를 들어, 적어도 0.1%, 적어도 1%, 또는 적어도 5% 내지 100% 범위일 수 있다. 샘플은 예를 들어, 세포 유리 DNA 또는 종양 샘플일 수 있다. 치료의 과정은 예를 들어, 그들의 빈도를 비롯한 빈도 범위 내에서 발생하는 임의의 또는 모든 돌연변이에 기반하여 처방될 수 있다. 샘플은 대상체로부터 임의의 후속 시간에 채취할 수 있다. 원래의 범위의 빈도 또는 상이한 범위의 빈도 내에서 발생하는 돌연변이가 결정될 수 있다. 치료의 과정은 후속적인 측정에 기반하여 조정될 수 있다.

[0312] **C. 다른 질환 또는 질환 상태의 조기 검출 및 모니터링**

[0313] 본원에서 설명되는 방법 및 시스템은 단지 암과 연관된 희귀 돌연변이 및 카피수 변이의 검출에만 제한되지는 않을 수 있다. 다양한 다른 질환 및 감염이 조기 검출 및 모니터링을 위해 적합할 수 있는 다른 종류의 상태를 일으킬 수 있다. 예를 들어, 특정 경우에, 유전 장애 또는 감염성 질환은 대상체 내에서 특정 유전적 모자이크 현상 (mosaicism) 유발할 수 있다. 상기 유전적 모자이크 현상은 관찰할 수 있는 카피수 변이 및 희귀 돌연변

이를 유발할 수 있다. 또 다른 예에서, 본 개시내용의 시스템 및 방법은 또한 체내에서 면역 세포의 게놈을 모니터링하기 위해 사용될 수 있다. B 세포와 같은 면역 세포는 특정 질환의 존재 시에 신속한 클론 확장(clonal expansion)을 겪을 수 있다. 카피수 변이 검출을 이용하여 클론 확장을 모니터링할 수 있고, 특정 면역 상태를 모니터링할 수 있다. 상기 예에서, 카피수 변이 분석은 특정 질환이 진행할 수 있는 방식의 프로파일을 생성하기 위해 시간 경과에 따라 수행할 수 있다.

[0314] 또한, 본 개시내용의 시스템 및 방법은 또한, 박테리아 또는 바이러스와 같은 병원체에 의해 유발될 수 있는 전신 감염 자체를 모니터링하기 위해 사용될 수 있다. 카피수 변이 또는 심지어 희귀 돌연변이 검출은 병원체의 집단이 감염 과정 동안 변화하는 방식을 결정하기 위해 이용될 수 있다. 이것은 바이러스가 감염 과정 동안 생활주기 상태를 변화시키고/시키거나 보다 독성 형태로 돌연변이될 수 있는 만성 감염, 예컨대 HIV/AIDs 또는 간염 감염 동안 특히 중요할 수 있다.

[0315] 본 개시내용의 시스템 및 방법을 이용할 수 있는 또 다른 예는 이식 대상체의 모니터링이다. 일반적으로, 이식된 조직은 이식 시에 신체에 의한 특정 정도의 거부반응을 겪게 된다. 면역 세포는 이식된 조직을 파괴하려고 시도하므로, 본 개시내용의 방법은 숙주 신체의 거부 활성을 결정하거나 프로파일링하기 위해 사용될 수 있다. 이것은 이식된 조직의 상태를 모니터링하고 또한 거부반응의 치료 또는 예방의 과정을 변경시키는데 유용할 수 있다.

[0316] 또한, 본 개시내용의 방법은 대상체에서 비정상적인 상태의 비균질성을 특성화하기 위해 사용될 수 있고, 이 방법은 대상체에서 세포외 폴리뉴클레오티드의 유전자 프로파일을 생성하는 것을 포함하고, 여기서 유전자 프로파일은 카피수 변이 및 희귀 돌연변이 분석에 의해 생성된 다수의 데이터를 포함한다. 일부 경우에, 암을 포함하나 이에 제한되지는 않는 질환은 비균질성일 수 있다. 질환 세포는 동일하지 않을 수 있다. 암의 예에서, 일부 종양은 상이한 종류의 종양 세포를 포함하는 것으로 알려져 있고, 일부 세포는 암의 상이한 병기에 있다. 다른 예에서, 비균질성은 다중 초점 (foci)의 질환을 포함할 수 있다. 다시, 암의 예에서, 다중 종양 초점이 존재할 수 있고, 아마도 하나 이상의 초점은 원발 부위로부터 확산된 전이의 결과이다.

[0317] 본 개시내용의 방법은 비균일한 질환에서 상이한 세포로부터 유래된 유전 정보의 요약인 데이터의 지문(fingerprint) 또는 세트를 생성하거나 프로파일링하기 위해 사용될 수 있다. 상기 데이터 세트는 카피수 변이 및 희귀 돌연변이 분석을 단독으로 또는 조합으로 포함할 수 있다.

[0318] **D. 태아 기원의 다른 질환 또는 질환 상태의 조기 검출 및 모니터링**

[0319] 추가로, 본 개시내용의 시스템 및 방법은 태아 기원의 암 또는 다른 질환을 진단하거나, 예측하거나, 모니터링하거나, 관찰하기 위해 사용될 수 있다. 즉, 이들 방법은 그의 DNA 및 다른 폴리뉴클레오티드가 모계 분자와 동시-순환할 수 있는 태중 대상체 내에서 암 또는 다른 질환을 진단하거나, 예측하거나, 모니터링하거나, 관찰하기 위해 임신한 대상체에서 사용될 수 있다.

[0320] **VIII. 용어**

[0321] 본원에서 사용되는 용어는 단지 특정 실시양태를 설명하기 위한 것이고, 본원의 시스템 및 방법을 제한하려고 의도되지 않는다. 본원에 사용될 때, 단수형 관사 ("a", "an" 및 "the")는 문맥상 명백하게 달리 지시되지 않으면 복수 형태를 또한 포함하도록 의도된다. 또한, 용어 "포함하는", "포함하다", "갖는", "갖다", "을 갖는", 또는 그의 변이형이 상세한 설명 및/또는 청구항에 사용되는 경우에, 그러한 용어는 용어 "구성되는"과 유사한 방식으로 포괄적인 것으로 의도된다.

[0322] 본원의 시스템 및 방법의 몇몇 측면은 예시를 위한 실례를 참조로 상기 설명된다. 수많은 구체적인 상세한 내용, 상관관계, 및 방법은 시스템 및 방법의 충분한 이해를 위해 설명됨을 이해해야 한다. 그러나, 관련 기술분야의 통상의 기술자는 시스템 및 방법이 하나 이상의 구체적인 상세한 설명 없이도 또는 다른 방법을 사용하여 실시될 수 있음을 쉽게 알 것이다. 본 개시내용은 행동 또는 사건의 예시된 순서에 의해 제한되지 않고, 이것은 일부 행동은 다른 행동 또는 사건과 상이한 순서로 및/또는 동시에 일어날 수 있기 때문이다. 또한, 본원에 따른 방법론을 실행하기 위해 예시된 행동 또는 사건이 모두 요구되는 것은 아니다.

[0323] 범위는 본원에서 "약" 하나의 특정 값으로부터, 및/또는 "약" 또 다른 특정 값까지로서 표현할 수 있다. 그러한 범위가 표현될 때, 또 다른 실시양태는 하나의 특정 값으로부터 및/또는 다른 특정 값까지를 포함한다. 유사하게, 값이 선행사 "약"을 사용하여 근사치로 표현될 때, 특정 값이 또 다른 실시양태를 형성함이 이해될 것이다. 각각의 범위의 종점은 다른 종점에 관련하여 및 다른 종점에 독립적으로 둘 모두 유의함이 추가로 이해될 것이다. 본원에서 사용될 때 용어 "약"은 특정 용도의 문맥 내에서 진술된 수치로부터 플러스 또는 마이너

스 15% 인 범위를 나타낸다. 예를 들어, 약 10은 8.5 내지 11.5 범위를 포함할 것이다.

[0324] **컴퓨터 시스템**

[0325] 본원의 방법은 컴퓨터 시스템을 사용하여 또는 그의 도움으로 실행할 수 있다. 도 15는 본원의 방법을 실행하기 위해 프로그래밍되거나 달리 구성된 컴퓨터 시스템 (1501)을 보여준다. 컴퓨터 시스템 (1501)은 샘플 제조, 서열분석 및/또는 분석의 다양한 측면을 조절할 수 있다. 일부 예에서, 컴퓨터 시스템 (1501)은 핵산 서열분석을 비롯하여, 샘플 제조 및 샘플 분석을 수행하기 위해 구성된다.

[0326] 컴퓨터 시스템 (1501)은 중앙 처리 유닛 (CPU 또한 본원에서 "프로세서" 및 "컴퓨터 프로세서") (1505)를 포함하고, 이것은 단일 코어 또는 멀티 코어 프로세서, 또는 병렬 처리를 위한 다수의 프로세서일 수 있다. 컴퓨터 시스템 (1501)은 또한 메모리 또는 메모리 위치 (1510) (예를 들어, 랜덤 액세스 (random-access) 메모리, 읽기-전용 (read-only) 메모리, 플래시 (flash) 메모리), 전자 저장 유닛 (1515) (예를 들어, 하드 디스크), 하나 이상의 다른 시스템과 통신하기 위한 통신 인터페이스 (1520) (예를 들어, 네트워크 어댑터), 및 주변 장치 (1525), 예컨대 캐시 (cache), 다른 메모리, 데이터 저장 및/또는 전자 디스플레이 어댑터를 포함한다. 메모리 (1510), 저장 유닛 (1515), 인터페이스 (1520) 및 주변 장치 (1525)는 마더보드 (motherboard)와 같은 통신 버스 (bus) (실선)를 통해 CPU (1505)와 통신한다. 저장 유닛 (1515)는 데이터를 저장하기 위한 데이터 저장 유닛 (또는 데이터 저장소)일 수 있다. 컴퓨터 시스템 (1501)은 통신 인터페이스 (1520)의 도움으로 컴퓨터 네트워크 ("네트워크") (1530)에 작동가능하게 연결될 수 있다. 네트워크 (1530)은 인터넷, 인터넷 및/또는 엑스트라넷 (extranet), 또는 인터넷과 통신하는 인트라넷 (intranet) 및/또는 엑스트라넷일 수 있다. 네트워크 (1530)는 일부 경우에 전자통신 및/또는 데이터 네트워크이다. 네트워크 (1530)은 하나 이상의 컴퓨터 서버 (server)를 포함할 수 있고, 이것은 클라우드 컴퓨팅 (cloud computing)과 같은 분산 컴퓨팅을 가능하게 할 수 있다. 네트워크 (1530)는 일부 경우에 컴퓨터 시스템 (1501)의 도움으로, 동등계층 (peer-to-peer) 네트워크를 실행할 수 있고, 이것은 컴퓨터 시스템 (1501)에 연결된 장치가 클라이언트 (client) 또는 서버로서 행동하도록 할 수 있다.

[0327] CPU (1505)는 프로그램 또는 소프트웨어 내에 구현될 수 있는 기계 판독가능 명령어의 순서를 실행할 수 있다. 명령어는 메모리 (1510)와 같은 메모리 위치 내에 저장될 수 있다. CPU (1505)에 의해 수행된 작업의 예는 호출 (fetch), 해독, 실행, 및 라이트백 (writeback)을 포함할 수 있다.

[0328] 저장 유닛 (1515)은 드라이버 (driver), 라이브러리 (library) 및 저장된 프로그램과 같은 파일을 저장할 수 있다. 저장 유닛 (1515)은 사용자에 의해 생성된 프로그램 및 기록된 세션 (recorded session), 및 프로그램과 연관된 출력(들)을 저장할 수 있다. 저장 유닛 (1515)은 사용자 데이터, 예를 들어 사용자 선호물 (preference) 및 사용자 프로그램을 저장할 수 있다. 몇몇 경우에, 컴퓨터 시스템 (1501)은 컴퓨터 시스템 (1501) 외부에 있는, 예컨대 인트라넷 또는 인터넷을 통해 컴퓨터 시스템 (1501)과 통신하는 원격 서버 상에 위치하는 하나 이상의 추가의 데이터 저장 유닛을 포함할 수 있다.

[0329] 컴퓨터 시스템 (1501)은 네트워크 (1530)를 통해 하나 이상의 원격 컴퓨터 시스템과 통신할 수 있다. 예를 들어, 컴퓨터 시스템 (1501)은 사용자의 원격 컴퓨터 시스템과 통신할 수 있다 (예를 들어, 오퍼레이터 (operator)). 원격 컴퓨터 시스템의 예는 개인용 컴퓨터 (예를 들어, 휴대용 PC), 슬레이트 (slate) PC 또는 태블릿 PC (예를 들어, 애플(Apple)® 아이패드 (iPad), 삼성(Samsung)® 갤럭시 탭 (Galaxy Tab)), 전화기, 스마트폰 (예를 들어, 애플® 아이폰 (iPhone), 안드로이드 이용가능 (Android-enabled) 장치, 블랙베리 (Blackberry)®), 또는 개인 휴대정보 단말기 (digital assistant)를 포함한다. 사용자는 네트워크 (1530)를 통해 컴퓨터 시스템 (1501)에 접근할 수 있다.

[0330] 본원에서 설명되는 방법은 컴퓨터 시스템 (1501)의 전자 저장 위치, 예컨대 예를 들어 메모리 (1510) 또는 전자 저장 유닛 (1515) 상에 저장된 기계 (예를 들어, 컴퓨터 프로세서) 실행가능 코드에 의해 실행할 수 있다. 기계-실행가능 코드 또는 기계 판독가능 코드는 소프트웨어의 형태로 제공될 수 있다. 사용 동안, 코드는 프로세서 (1505)에 의해 실행될 수 있다. 일부 경우에, 코드는 저장 유닛 (1515)로부터 검색되고, 프로세서 (1505)에 의한 빠른 접근을 위해 메모리 (1510) 상에 저장될 수 있다. 일부 상황에서, 전자 저장 유닛 (1515)은 배제될 수 있고, 기계-실행가능 명령어는 메모리 (1510) 상에 저장된다.

[0331] 코드는 예비-컴파일링 (pre-compiled)되고 기계와 함께 사용되도록 구성될 수 있거나, 코드를 실행하도록 채택된 프로세서 (processer)를 갖거나, 실행 시간 동안 컴파일링될 수 있다. 코드는 코드가 예비컴파일링된 또는 컴파일링된 방식으로 실행되도록 하기 위해 선택될 수 있는 프로그래밍 언어 내에 제공될 수 있다.

[0332] 본원에 제공되는 시스템 및 방법, 예컨대 컴퓨터 시스템 (1501)의 측면은 프로그래밍으로 구현될 수 있다. 기술의 다양한 측면은 전형적으로 기계 판독가능 매체의 형태 상에서 운반되거나 그러한 형태 내에서 구현되는 기계 (또는 프로세서) 실행가능 코드 및/또는 연관된 데이터의 형태의 "생산품 (product)" 또는 "제조품 (article of manufacture)"으로서 생각될 수 있다. 기계-실행가능 코드는 전자 저장 유닛, 그러한 메모리 (예를 들어, 읽기 전용 메모리, 랜덤 액세스 메모리, 플래시 메모리) 또는 하드 디스크 상에 저장될 수 있다. "저장"형 매체는 컴퓨터, 프로세서 등의 임의의 또는 모든 유형의 메모리, 또는 그의 연관된 모듈, 예컨대 다양한 반도체 메모리, 테이프 드라이브, 디스크 드라이브 등을 포함할 수 있고, 이것은 소프트웨어 프로그래밍을 위해 언제든지 비-일시적인 저장을 제공할 수 있다. 소프트웨어의 전부 또는 일부는 때때로 인터넷 또는 다양한 다른 전기통신 네트워크를 통해 통신할 수 있다. 그러한 통신은 예를 들어, 하나의 컴퓨터 또는 프로세서로부터 다른 것으로, 예를 들어 관리 서버 또는 호스트 컴퓨터로부터 어플리케이션 서버의 컴퓨터 플랫폼으로 소프트웨어의 로딩을 가능하게 할 수 있다. 따라서, 소프트웨어 요소를 보유할 수 있는 또 다른 종류의 매체는 지역 (local) 장치 사이의 물리적 인터페이스를 가로질러, 유선 및 광학 지상통신망 (landline) 네트워크를 통해 및 다양한 에어링크 (air-link)를 넘어 사용된 것과 같은 광학, 전기 및 전자기파를 포함한다. 그러한 파장을 운반하는 물리적 요소, 예컨대 유선 또는 무선 링크, 광학 링크 등은 또한 소프트웨어를 보유하는 매체로서 간주될 수 있다. 본원에서 사용될 때, 비-일시적인 유형의 "저장" 매체에 제한되지 않으면, 컴퓨터 또는 기계 "판독가능 매체"와 같은 용어는 실행을 위해 프로세서에 명령어를 제공하는데 참여하는 임의의 매체를 나타낸다.

[0333] 따라서, 기계 판독가능 매체, 예컨대 컴퓨터-실행가능 코드는 유형의 저장 매체, 반송파 (carrier wave) 매체 또는 물리적 전송 매체를 포함하나 이에 제한되지는 않는 많은 형태를 취할 수 있다. 비-휘발성 (volatile) 저장 매체는 예를 들어, 광학 또는 자기 디스크, 예컨대 임의의 컴퓨터(들) 내의 임의의 저장 장치 등을 포함하고, 이것은 도면에 제시된 데이터베이스 등을 실행하기 위해 사용될 수 있다. 휘발성 저장 매체는 동적 메모리, 예컨대 그러한 컴퓨터 플랫폼의 주 메모리를 포함한다. 유형의 전송 매체는 컴퓨터 시스템 내에 버스를 포함하는 전선을 포함한 동축 (coaxial) 케이블; 구리 전선 및 광섬유를 포함한다. 반송파 전송 매체는 전기 또는 전자기 신호, 또는 음파 또는 광파, 예컨대 고주파 (RF) 및 적외선 (IR) 데이터 통신 동안 생성된 것의 형태를 취할 수 있다. 따라서, 일반적인 형태의 컴퓨터-판독가능 매체는 예를 들어: 플로피 디스크, 플렉시블 (flexible) 디스크, 하드 디스크, 자기 테이프, 임의의 다른 자기 매체, CD-ROM, DVD 또는 DVD-ROM, 임의의 다른 광학 매체, 천공 카드 (punch cards) 종이 테이프, 홀 (hole)의 패턴을 갖는 임의의 다른 물리적 저장 매체, RAM, ROM, PROM 및 EPROM, FLASH-EPROM, 임의의 다른 메모리 칩 또는 카트리지 (cartridge), 데이터 또는 명령어를 수송하는 반송파, 그러한 반송파를 수송하는 케이블 또는 링크, 또는 그로부터 컴퓨터가 프로그래밍 코드 및/또는 데이터를 판독할 수 있는 임의의 다른 매체를 포함한다. 이들 형태의 많은 컴퓨터 판독가능 매체는 실행을 위해 프로세서에 하나 이상의 명령어의 하나 이상의 순서를 운반하는데 관여할 수 있다.

[0334] 컴퓨터 시스템 (1501)은 예를 들어, 샘플 분석의 하나 이상의 결과를 제공하기 위한 사용자 인터페이스 (UI)를 포함하는 전자 디스플레이를 포함하거나 그와 통신 상태일 수 있다. UI의 예는 비제한적으로 그래픽 사용자 인터페이스 (GUI) 및 웹-기반 사용자 인터페이스를 포함한다.

[0335] **실시예**

[0336] **실시예 1 - 전립선암 예측 및 치료**

[0337] 혈액 샘플을 전립선암 대상체로부터 채취하였다. 이전에, 종양의는 대상체가 II기 전립선암에 걸렸음을 결정하고 치료를 권유하였다. 초기 진단 후에 6개월 마다 세포 유리 DNA를 추출하고, 단리하고, 서열분석하고, 분석하였다.

[0338] 퀴아젠 큐빗 키트 프로토콜을 이용하여 세포 유리 DNA를 혈액으로부터 추출하고 단리하였다. 수득률을 증가시키기 위해 담체 DNA를 첨가하였다. PCR 및 범용 프라이머를 이용하여 DNA를 증폭하였다. 일루미나 MiSeq 개인용 서열분석기를 사용하는 대규모 병렬형 서열분석 방안을 이용하여 10 ng의 DNA를 서열분석하였다. 대상체의 게놈의 90%가 세포 유리 DNA의 서열분석을 통해 포함된다.

[0339] 서열 데이터를 모으고 카피수 변이에 대해 분석하였다. 서열 판독체를 맵핑하고, 건강한 개체 (대조군)에 비교하였다. 서열 판독체의 수에 기반하여, 염색체 영역을 50 kb 비-겹침 영역으로 나누었다. 서열 판독체를 서로 비교하고, 각각의 맵핑가능 위치에 대해 비율을 결정하였다.

[0340] 카피수를 각각의 윈도우에 대해 별개의 상태로 전환시키기 위해 은닉 마르코프 모델을 적용하였다.

[0341] 보고서가 생성되었고, 맵핑 게놈 위치 및 카피수 변이를 도 4a (건강한 개체에 대해) 및 암이 있는 대상체에 대

해 도 4b에 제시한다.

[0342] 공지의 성과를 갖는 대상체의 다른 프로파일에 비해, 이들 보고서는 상기 특정 암이 침습성이고 치료에 저항성임을 나타냈다. 세포 유리 종양 부담 (tumor burden)은 21%이다. 대상체를 18개월 동안 모니터링하였다. 제 18월에, 카피수 변이 프로파일은 21%의 세포 유리 종양 부담에서 30%로 극적으로 증가하기 시작하였다. 다른 전립선암 대상체의 유전자 프로파일과 비교하였다. 카피수 변이의 상기 증가가 전립선암이 II기에서 III기로 진행하고 있음을 나타내는 것으로 결정되었다. 처방된 원래의 치료 계획은 더 이상 암을 치료하지 않았다. 새로운 치료가 처방되었다.

[0343] 또한, 이들 보고서를 제출하였고, 인터넷을 통해 전자적으로 접근된다. 서열 데이터의 분석은 대상체의 위치 이외의 장소에서 일어났다. 보고서가 생성되고, 대상체의 위치로 전송되었다. 인터넷 접속이 가능한 컴퓨터를 통해, 대상체는 그의 종양 부담을 반영하는 보고서에 접근한다 (도 4c).

[0344] **실시예 2 - 전립선암 완화 및 재발**

[0345] 혈액 샘플을 전립선암 생존자로부터 채취하였다. 대상체는 이전에 수많은 라운드의 화학요법 및 방사선 치료를 겪었다. 시험 당시 대상체는 암에 관련된 증상 또는 건강 문제를 나타내지 않았다. 표준 스캔 및 검정에서 대상체는 암이 없는 상태로 밝혀졌다.

[0346] 퀴아젠 TruSeq 키트 프로토콜을 이용하여 세포 유리 DNA를 혈액으로부터 추출하고 단리하였다. 수득률을 증가시키기 위해 담체 DNA를 첨가하였다. PCR 및 범용 프라이머를 이용하여 DNA를 증폭하였다. 일루미나 MiSeq 개인용 서열분석기를 사용하는 대규모 병렬형 서열분석 방안을 이용하여 10 ng의 DNA를 서열분석하였다. 라이게이션 방법을 이용하여 12량체 바코드를 개별 분자에 첨가하였다.

[0347] 서열 데이터를 모으고 카피수 변이에 대해 분석하였다. 서열 판독체를 맵핑하고, 건강한 개체 (대조군)에 비교하였다. 서열 판독체의 수에 기반하여, 염색체 영역을 40 kb 비-겹침 영역으로 나누었다. 서열 판독체를 서로 비교하고, 각각의 맵핑가능 위치에 대해 비율을 결정하였다.

[0348] 증폭으로부터 편향을 정규화하는 것을 돕기 위해 비-특유한 바코드 연결 서열은 단일 판독체로 붕괴시켰다.

[0349] 카피수를 각각의 윈도우에 대해 별개의 상태로 전환시키기 위해 은닉 마르코프 모델을 적용하였다.

[0350] 보고서가 생성되었고, 맵핑 게놈 위치 및 카피수 변이를 완화 상태의 암이 있는 대상체에 대해 도 5a, 및 재발 상태의 암이 있는 대상체에 대해 도 5b에 제시하였다.

[0351] 공지의 성과를 갖는 대상체의 다른 프로파일에 비해, 상기 보고서는 제18월에, 카피수 변이의 희귀 돌연변이 분석이 5%의 세포 유리 종양 부담에서 검출되었음을 나타낸다. 종양의가 다시 치료를 처방하였다.

[0352] **실시예 3 - 갑상선암 및 치료**

[0353] 대상체는 IV기 갑상선암에 걸려있고, I-131을 사용하는 방사선 요법을 포함한 표준 치료를 받는 것으로 알려져 있었다. CT 스캔은 방사선 요법이 암성 덩어리를 파괴하는지 여부에 대해 결정을 내리지 못하였다. 혈액을 마지막 방사선 시기 전후에 채취하였다.

[0354] 퀴아젠 큐빗 키트 프로토콜을 이용하여 세포 유리 DNA를 혈액으로부터 추출하고 단리하였다. 수득률을 증가시키기 위해 비-특이적 벌크 DNA의 샘플을 샘플 제조 반응액에 첨가하였다.

[0355] BRAF 유전자는 상기 갑상선암에서 아미노산 위치 600에서 돌연변이될 수 있음이 공지되어 있다. 세포 유리 DNA의 집단으로부터, 유전자에 특이적인 프라이머를 이용하여 BRAF DNA를 선택적으로 증폭하였다. 판독체를 계수하기 위한 대조군으로서 20량체 바코드를 모 분자에 첨가하였다.

[0356] 일루미나 MiSeq 개인용 서열분석기를 사용하는 대규모 병렬형 서열분석 방안을 이용하여 10 ng의 DNA를 서열분석하였다

[0357] 서열 데이터를 모으고 카피수 변이 검출에 대해 분석하였다. 서열 판독체를 맵핑하고, 건강한 개체 (대조군)에 비교하였다. 바코드 서열을 계수함으로써 결정한 서열 판독체의 수에 기반하여, 염색체 영역을 50 kb 비-겹침 영역으로 나누었다. 서열 판독체를 서로 비교하고, 각각의 맵핑가능 위치에 대해 비율을 결정하였다.

[0358] 카피수를 각각의 윈도우에 대해 별개의 상태로 전환시키기 위해 은닉 마르코프 모델을 적용하였다

[0359] 맵핑 게놈 위치 및 카피수 변이를 포함하는 보고서가 생성되었다.

[0360] 치료 전후에 생성된 보고서를 비교하였다. 종양 세포 부담 백분율은 30%에서 방사선 시기 후에 60%로 급증하였다. 종양 부담의 상기 급증은 치료의 결과로서 정상적인 조직에 비한 암 조직의 괴사의 증가인 것으로 결정되었다. 종양의는 대상체가 처방된 치료를 지속하도록 권유하였다.

[0361] **실시예 4 - 희귀 돌연변이 검출의 감도**

[0362] DNA의 집단 내에 존재하는 희귀 돌연변이의 검출 범위를 결정하기 위해, 혼합 실험을 수행하였다. DNA의 서열(일부는 유전자 TP53, HRAS 및 MET의 야생형 카피를 함유하고, 일부는 동일한 유전자에서 희귀 돌연변이를 갖는 카피를 함유한다)을 상이한 비율로 함께 혼합하였다. 돌연변이체 DNA 대 야생형 DNA의 비율 또는 백분율이 100% 내지 0.01% 범위가 되도록 DNA 혼합물을 제조하였다.

[0363] 일루미나 MiSeq 개인용 서열분석기를 사용하는 대규모 병렬형 서열분석 방안을 이용하여, 각각의 혼합 실험에 대해 10 ng의 DNA를 서열분석하였다

[0364] 서열 데이터를 모으고 희귀 돌연변이 검출에 대해 분석하였다. 서열 판독체를 맵핑하고, 참조 서열 (대조군)에 비교하였다. 서열 판독체의 수에 기반하여, 각각의 맵핑가능 위치에 대한 변이의 빈도가 결정되었다.

[0365] 각각의 맵핑가능 위치에 대한 변이의 빈도를 염기 위치에 대한 별개의 상태로 전환시키기 위해 은닉 마르코프 모델을 적용하였다.

[0366] 맵핑 게놈 염기 위치 및 참조 서열에 의해 결정된 기준선에 비해 희귀 돌연변이의 검출 백분율을 포함한 보고서가 생성되었다 (도 6a).

[0367] 0.1% 내지 100% 범위의 다양한 혼합 실험의 결과를 로그 규모 그래프에 제시하였고, 여기서 희귀 돌연변이를 갖는 DNA의 측정된 백분율을 희귀 돌연변이를 갖는 DNA의 실제 백분율의 함수로서 그래프로 그렸다 (도 6b). 3개의 유전자, 즉, TP53, HRAS 및 MET를 제시하였다. 측정된 및 예상된 희귀 돌연변이 집단 사이에 강한 선형 상관관계가 발견되었다. 추가로, 비-돌연변이된 DNA의 집단 내에서 희귀 돌연변이를 갖는 DNA의 약 0.1%의 보다 낮은 감도 역치가 이들 실험으로 밝혀졌다 (도 6b).

[0368] **실시예 5 - 전립선암 대상체에서 희귀 돌연변이 검출**

[0369] 대상체는 초기 전립선암에 걸린 것으로 생각되었다. 다른 임상 시험에서는 결론을 내리지 못하는 결과를 제공하였다. 혈액을 대상체로부터 채취하고, 세포 유리 DNA를 추출하고, 단리하고, 제조하고, 서열분석하였다.

[0370] 유전자 특이적 프라이머를 사용하는 택맨(TaqMan)ⓒ PCR 키트 (인비트로겐 (Invitrogen))을 이용한 선택적 증폭을 위해 다양한 종양유전자 및 종양 억제 유전자의 패널을 선택하였다. 증폭된 DNA 영역은 PIK3CA 및 TP53 유전자를 함유하는 DNA를 포함하였다.

[0371] 일루미나 MiSeq 개인용 서열분석기를 사용하는 대규모 병렬형 서열분석 방안을 이용하여 10 ng의 DNA를 서열분석하였다

[0372] 서열 데이터를 모으고 희귀 돌연변이 검출에 대해 분석하였다. 서열 판독체를 맵핑하고, 참조 서열 (대조군)에 비교하였다. 서열 판독체의 수에 기반하여, 각각의 맵핑가능 위치에 대한 변이의 빈도를 결정하였다.

[0373] 각각의 맵핑가능 위치에 대한 변이의 빈도를 각각의 염기 위치에 대한 별개의 상태로 전환시키기 위해 은닉 마르코프 모델을 적용하였다.

[0374] 맵핑 게놈 염기 위치 및 참조 서열에 의해 결정된 기준선에 비교한 희귀 돌연변이의 검출 백분율을 포함하는 보고서가 생성되었다 (도 7a). 희귀 돌연변이는 2개의 유전자, 즉 PIK3CA 및 TP53에 대해 각각 5%의 발생률로 발견되었고, 이것은 대상체가 초기 암에 걸렸음을 나타낸다. 치료를 개시하였다.

[0375] 또한, 이들 보고서를 제출하였고, 인터넷을 통해 전자적으로 접근된다. 서열 데이터의 분석은 대상체의 위치 이외의 장소에서 일어났다. 보고서가 생성되고, 대상체의 위치로 전송되었다. 인터넷 접속이 가능한 컴퓨터를 통해, 대상체는 그의 종양 부담을 반영하는 보고서에 접근한다 (도 7b).

[0376] **실시예 6 - 결장직장암 대상체에서 희귀 돌연변이 검출**

[0377] 대상체는 중기 결장직장암에 걸린 것으로 생각되었다. 다른 임상 시험에서는 결정을 내리지 못하는 결과를 제공하였다. 혈액을 대상체로부터 채취하고, 세포 유리 DNA를 추출하였다.

[0378] 단일 튜브의 혈장으로부터 추출한 10 ng의 세포 유리 유전 물질을 사용하였다. 초기 유전 물질을 태그부착된

모 폴리뉴클레오티드의 세트로 전환시켰다. 태그부착은 서열분석을 위해 요구되는 태그, 및 자손 분자를 추적하기 위한 비-특유한 식별자를 모 핵산에 부착시키는 것을 포함한다. 전환은 상기 설명된 최적화된 라이게이션 반응을 통해 수행되고, 전환 수득률은 라이게이션 후 분자의 크기 프로파일을 검토함으로써 확인하였다. 전환 수득률은 두 단부가 태그로 라이게이팅된 출발 초기 분자의 백분율로서 측정되었다. 상기 방안을 이용하는 전환은 예를 들어, 적어도 50%의 높은 효율로 수행된다.

[0379] 태그부착된 라이브러리를 PCR-증폭하고, 결장직장암과 가장 연관된 유전자 (예를 들어, KRAS, APC, TP53 등)에 대해 풍부화시키고, 생성되는 DNA를 일루미나 MiSeq 개인용 서열분석기를 사용하는 대규모 병렬형 서열분석 방안을 이용하여 서열분석하였다.

[0380] 서열 데이터를 모으고 희귀 돌연변이 검출에 대해 분석하였다. 서열 판독체를 모 분자에 속하는 패밀리 그룹으로 붕괴시키고 (또한 붕괴 시에 오류-교정하고), 참조 서열 (대조군)을 이용하여 맵핑하였다. 서열 판독체의 수에 기반하여, 각각의 맵핑가능 위치에 대한 희귀 변이 (치환, 삽입, 결실 등), 및 카피수 및 이형접합성의 변이 (적절한 경우에)의 빈도를 결정하였다.

[0381] 맵핑 게놈 염기 위치, 및 참조 서열에 의해 결정된 기준선에 비한 희귀 돌연변이의 검출 백분율을 포함하는 보고서가 생성되었다. 희귀 돌연변이는 2개의 유전자, KRAS 및 FBXW7에 대해 각각 0.3-0.4%의 발생률로 발견되었고, 이것은 대상체가 잔류 암에 걸려 있음을 나타낸다. 치료를 개시하였다.

[0382] 또한, 이들 보고서를 제출하였고, 인터넷을 통해 전자적으로 접근된다. 서열 데이터의 분석은 대상체의 위치 이외의 장소에서 일어났다. 보고서가 생성되고, 대상체의 위치로 전송되었다. 인터넷 접속이 가능한 컴퓨터를 통해, 대상체는 그의 종양 부담을 반영하는 보고서에 접근한다.

[0383] **실시예 7 - 디지털 서열분석 기술**

[0384] 종양-방출 핵산의 농도는 대체로 낮아서, 현재의 차세대 서열분석 기술은 단지 그러한 신호를 산발적으로 또는 마지막으로 높은 종양 부담을 가진 환자에서만 검출할 수 있다. 그러한 기술이 오류 비율 및 편향에 의해 피해를 입는 주요 이유는 순환 DNA에서 암과 연관된 드노보 (de novo) 유전자 변경을 신뢰가능하게 검출하기 위해 요구되는 것보다 더 높은 자릿수일 수 있다. 본원에서 새로운 서열분석 방법론, 즉, 디지털 서열분석 기술 (DST)을 제시하고, 이것은 생식계열 단편 사이에서 희귀한 종양-유래 핵산의 검출 및 정량의 감도와 특이성을 적어도 1-2 자릿수 증가시킨다.

[0385] DST 아키텍처 (architecture)는 근대 통신 채널에 의해 유발된 높은 노이즈 및 왜곡을 퇴치하고, 극도로 높은 데이터 속도로 무결하게 디지털 정보를 전송할 수 있는 최신 디지털 통신 시스템에 의해 고무되었다. 이와 유사하게, 현재의 차세대 작업흐름은 극도로 높은 노이즈 및 왜곡 (샘플-제조, PCR-기반 증폭 및 서열분석으로 인한)에 의해 피해를 입는다. 디지털 서열분석은 이들 프로세스에 의해 생성된 오류 및 왜곡을 제거하고, 모든 희귀 변이체 (CNV 포함)의 거의 완벽한 제시물을 생성할 수 있다.

[0386] <u>높은 다양성 라이브러리 제조</u>

[0387] 비효율적인 라이브러리 전환 때문에 다수의 추출된 순환 DNA 단편이 소실되는 통상적인 서열분석 라이브러리 제조 프로토콜과 달리, 본 발명자들의 디지털 서열분석 기술 작업흐름에서는 대다수의 출발 분자가 전환되고 서열분석될 수 있다. 이것은 전체 10 mL 튜브의 혈액 내에 소량의 체세포 돌연변이된 분자만 존재할 수 있으므로 희귀 변이체의 검출에 매우 중요하다. 개발된 효율적 분자 생물학 전환 프로세스에 의해 희귀 변이체의 검출을 위한 최고의 가능한 감도가 가능해졌다.

[0388] <u>포괄적 작용가능한 종양유전자 패널</u>

[0389] 표적화된 영역은 단일 엑손만큼 작거나 또는 전체 엑손 (또는 심지어 전체 게놈)만큼 넓을 수 있으므로, DST 플랫폼 둘레에 작업되는 작업흐름은 탄력적이고 고도로 조율가능하다. 표준 패널은 15개의 작용가능한 암-관련 유전자의 모든 엑손 염기, 및 추가의 36개의 종양유전자/종양-억제 유전자의 "핫 (hot)" 엑손 (예를 들어, COSMIC 내에 적어도 하나의 또는 그 이상의 보고된 체세포 돌연변이를 함유하는 엑손)의 적용범위로 이루어진다.

[0390] **실시예 8: 분석적 연구**

[0391] 본 발명의 기술의 성능을 연구하기 위해, 분석적 샘플에서 그의 감도를 평가하였다. 본 발명자들은 다양한 양의 LNCaP 암 세포주 DNA를 정상 cfDNA의 배경 내로 스파이킹(spiking)하였고, 0.1% 감도에 이르기까지 체세포

돌연변이를 성공적으로 검출할 수 있었다 (도 13 참조).

[0392] 전임상 연구

[0393] 마우스 내에서 인간 이종이식편 모델에서 종양 gDNA와 순환 DNA의 일치성을 조사하였다. 7마리의 CTC-음성 마우스 (각각 2가지 상이한 인간 유방암 종양 중 하나를 갖는)에서, 종양 gDNA에서 검출된 모든 체세포 돌연변이가 또한 DST를 이용하여 마우스 혈액 cfDNA에서 검출되었고, 이것은 비-침습 종양 유전자 프로파일링을 위한 cfDNA의 효용을 추가로 입증하였다.

[0394] 파일럿 (pilot) 임상 연구

[0395] 종양 생검 대 순환 DNA 체세포 돌연변이의 상관성

[0396] 상이한 암 종류에 걸쳐 인간 샘플에 대해 파일럿 연구를 개시하였다. 매칭된 종양 생검 샘플로부터 유래된 것과 순환 세포 유리 DNA로부터 유래된 종양 돌연변이 프로파일의 일치성을 조사하였다. 14명의 환자에 걸쳐 결장직장암 및 흑색종 암 모두에서 종양 및 cfDNA 체세포 돌연변이 프로파일 사이에 93% 초과의 높은 일치성이 밝혀졌다 (표 1).

표 1

[0397]

환자 ID	병기	매칭된 종양에서 돌연변이체 유전자	돌연변이체 cfDNA의 백분율
CRC #1	II-B	TP53	0.2%
CRC #2	II-C	KRAS	0.6%
		SMAD4	1.5%
		GNAS	1.4%
		FBXW7	0.8%
CRC #3	III-B	KRAS	1.1%
		TP53	1.4%
		PIK3CA	1.7%
		APC	0.7%
CRC #4	III-B	KRAS	0.3%
		TP53	0.4%
CRC #5	III-B	KRAS	0.04%
CRC #6	III-C	KRAS	0.03%
CRC #7	IV	PIK3CA	1.3%
		KRAS	0.6%
		TP53	0.8%
CRC #8	IV	APC	0.3%
		SMO	0.6%

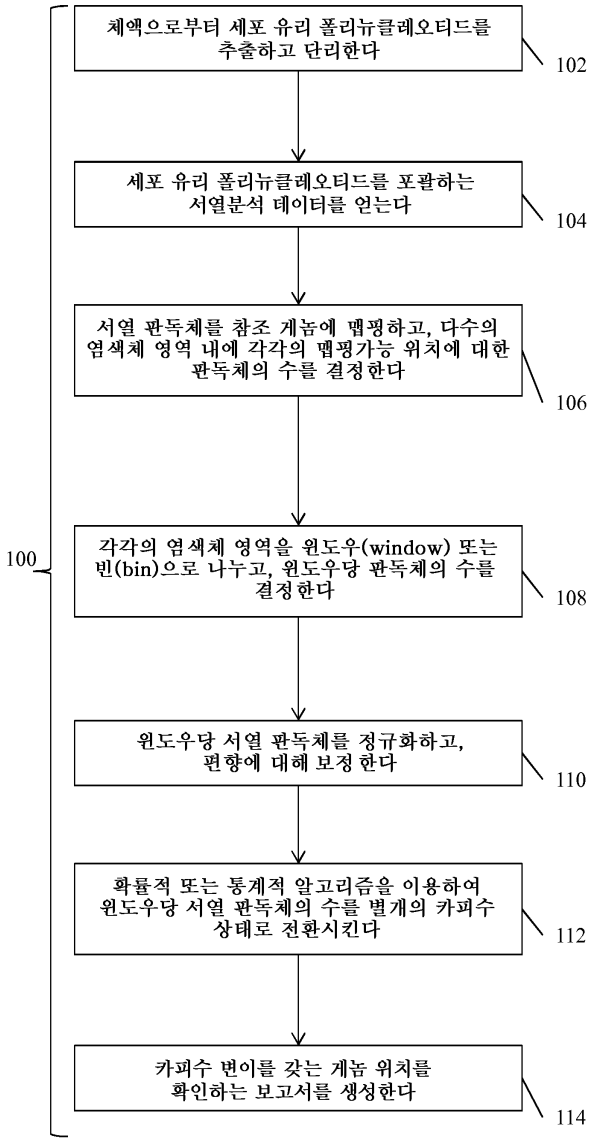
		TP53	0.4%
		KRAS	0.0%
CRC #9	IV	APC	47.3%
		APC	40.2%
		KRAS	37.7%
		PTEN	0.0%
		TP53	12.9%
CRC #10	IV	TP53	0.9%
흑색종 #1	IV	BRAF	0.2%
흑색종 #2	IV	APC	0.3%
		EGFR	0.9%
		MYC	10.5%
흑색종 #3	IV	BRAF	3.3%
흑색종 #4	IV	BRAF	0.7%

[0398]

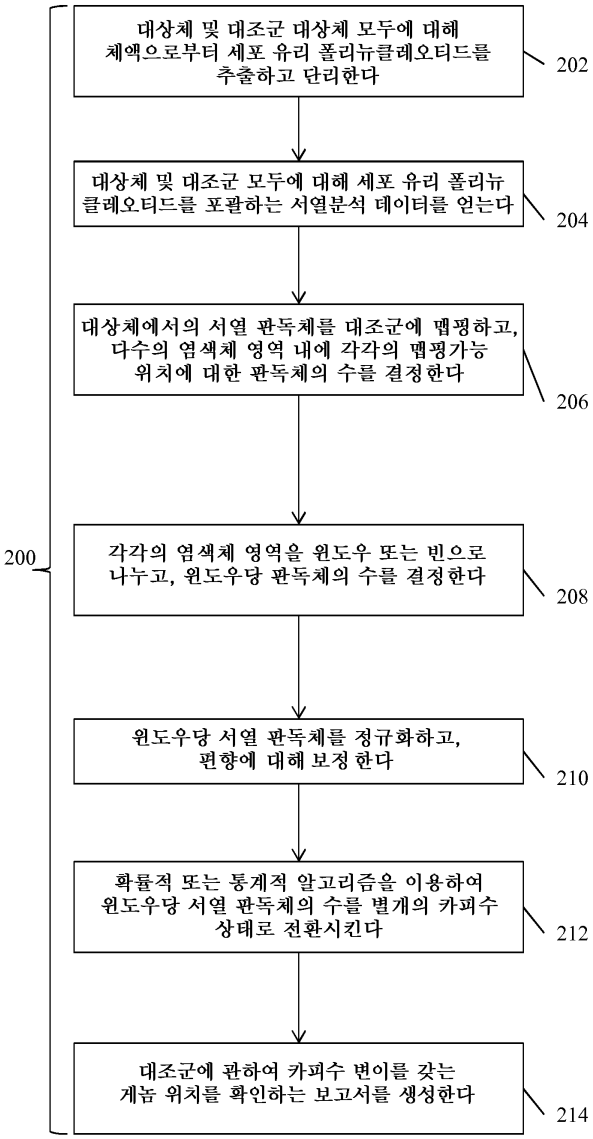
[0399] 상기한 내용으로부터 특정 실행이 예시되고 설명되지만, 그에 대한 다양한 변형이 이루어질 수 있고 본원에서 고려됨을 이해해야 한다. 또한, 본 발명은 명세서 내에서 제공된 구체적인 예에 의해 제한되는 것으로 의도되지 않는다. 본 발명은 상기한 명세서를 참조하여 설명되었지만, 상세한 설명 및 본원에서 바람직한 실시양태의 예시가 제한하는 의미로 해석되는 것을 의미하지는 않는다. 또한, 본 발명의 모든 측면은 다양한 조건 및 변수에 의존적인 본원에 제시된 구체적인 서술, 배열형태 또는 상대적인 비율에 제한되지 않음을 이해해야 한다. 본 발명의 실시양태의 형태의 다양한 변형 및 상세한 내용은 관련 기술 분야의 통상의 기술자에게 자명해질 것이다. 따라서, 본 발명은 임의의 그러한 변형, 변이 및 등가물을 또한 포함할 것으로 고려된다.

도면

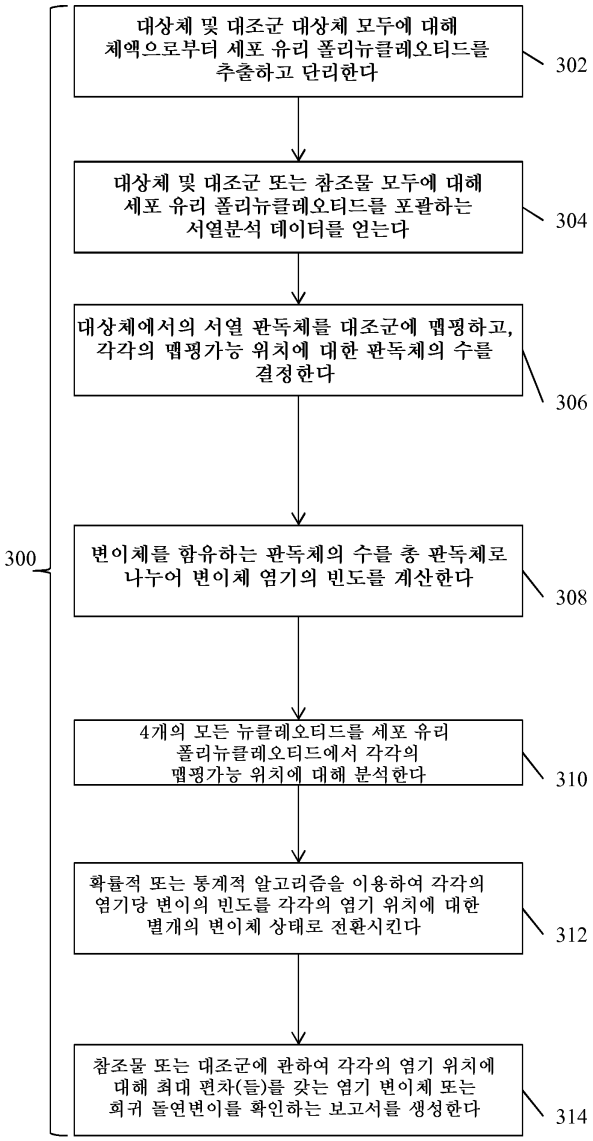
도면1


체액으로부터 세포 유리 폴리뉴클레오티드를 추출하고 단리한다
102
세포 유리 폴리뉴클레오티드를 포괄하는 서열분석 데이터를 얻는다
104
서열 판독체를 참조 게놈에 맵핑하고, 다수의 염색체 영역 내에 각각의 맵핑가능 위치에 대한 판독체의 수를 결정한다
106
각각의 염색체 영역을 윈도우(window) 또는 빈(bin)으로 나누고, 윈도우당 판독체의 수를 결정한다
108
윈도우당 서열 판독체를 정규화하고, 편향에 대해 보정 한다
110
확률적 또는 통계적 알고리즘을 이용하여 윈도우당 서열 판독체의 수를 별개의 카피수 상태로 전환시킨다
112
카피수 변이를 갖는 게놈 위치를 확인하는 보고서를 생성한다
114
100

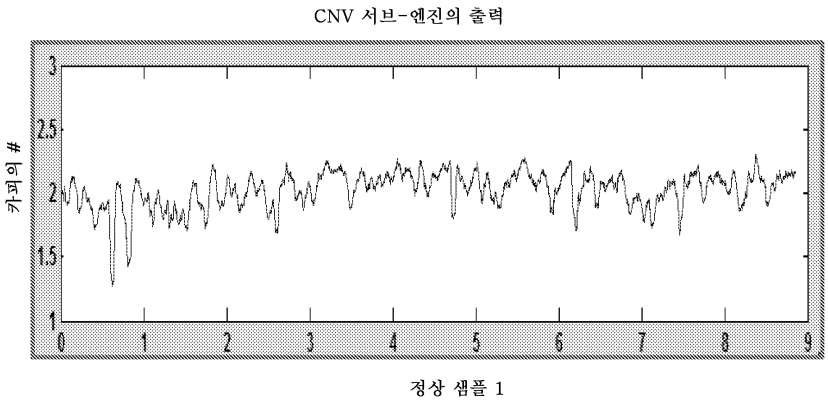
도면2


대상체 및 대조군 대상체 모두에 대해 체액으로부터 세포 유리 폴리뉴클레오티드를 추출하고 단리한다
202
대상체 및 대조군 모두에 대해 세포 유리 폴리뉴클레오티드를 포괄하는 서열분석 데이터를 얻는다
204
대상체에서의 서열 판독체를 대조군에 맵핑하고, 다수의 염색체 영역 내에 각각의 맵핑가능 위치에 대한 판독체의 수를 결정한다
206
200
각각의 염색체 영역을 윈도우 또는 빈으로 나누고, 윈도우당 판독체의 수를 결정한다
208
윈도우당 서열 판독체를 정규화하고, 편향에 대해 보정 한다
210
확률적 또는 통계적 알고리즘을 이용하여 윈도우당 서열 판독체의 수를 별개의 카피수 상태로 전환시킨다
212
대조군에 관하여 카피수 변이를 갖는 게놈 위치를 확인하는 보고서를 생성한다
214

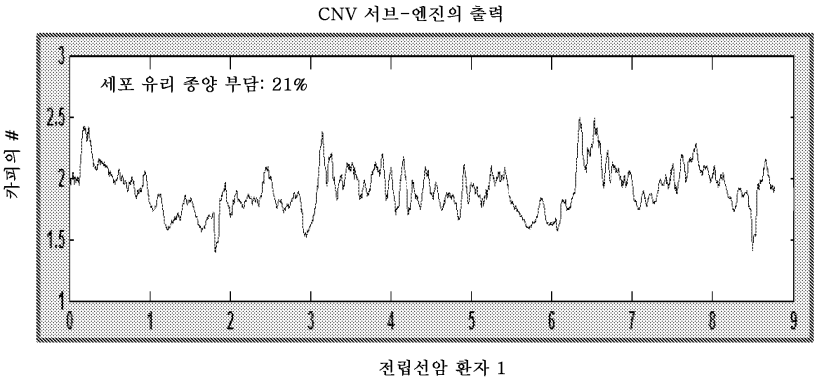
도면3


대상체 및 대조군 대상체 모두에 대해 체액으로부터 세포 유리 폴리뉴클레오티드를 추출하고 단리한다
302
대상체 및 대조군 또는 참조물 모두에 대해 세포 유리 폴리뉴클레오티드를 포괄하는 서열분석 데이터를 얻는다
304
대상체에서의 서열 판독체를 대조군에 맵핑하고, 각각의 맵핑가능 위치에 대한 판독체의 수를 결정한다
306
300
변이체를 함유하는 판독체의 수를 총 판독체로 나누어 변이체 염기의 빈도를 계산한다
308
4개의 모든 뉴클레오티드를 세포 유리 폴리뉴클레오티드에서 각각의 맵핑가능 위치에 대해 분석한다
310
확률적 또는 통계적 알고리즘을 이용하여 각각의 염기당 변이의 빈도를 각각의 염기 위치에 대한 별개의 변이체 상태로 전환시킨다
312
참조물 또는 대조군에 관하여 각각의 염기 위치에 대해 최대 편차(들)를 갖는 염기 변이체 또는 희귀 돌연변이를 확인하는 보고서를 생성한다
314

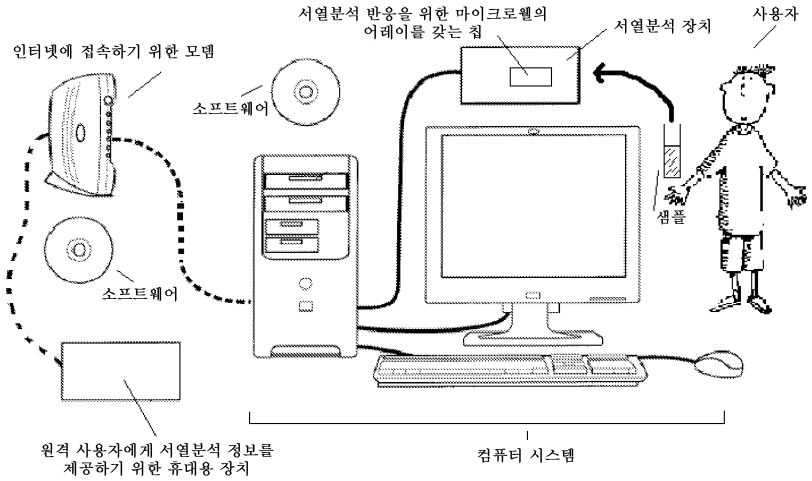
도면4a


CNV 서브-엔진의 출력
카피의 #
3
2.5
2
1.5
1
0
1
2
3
4
5
6
7
8
9
정상 샘플 1

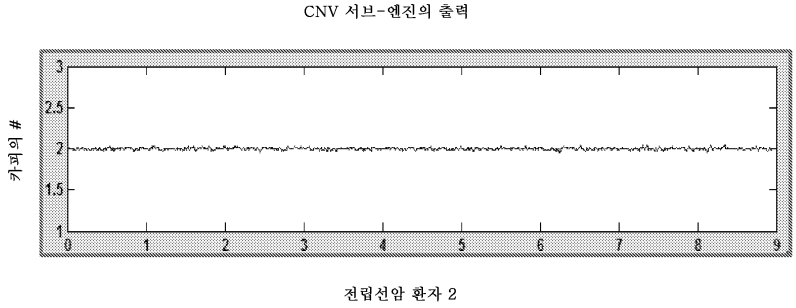
도면4b



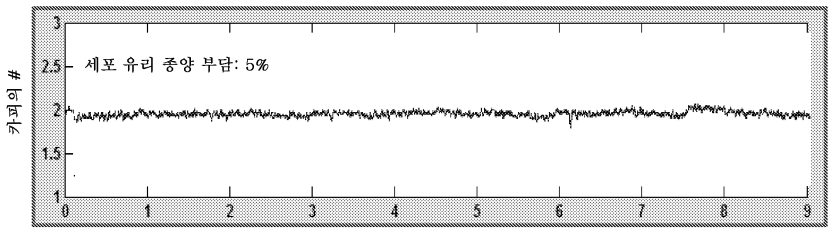
도면4c



도면5a

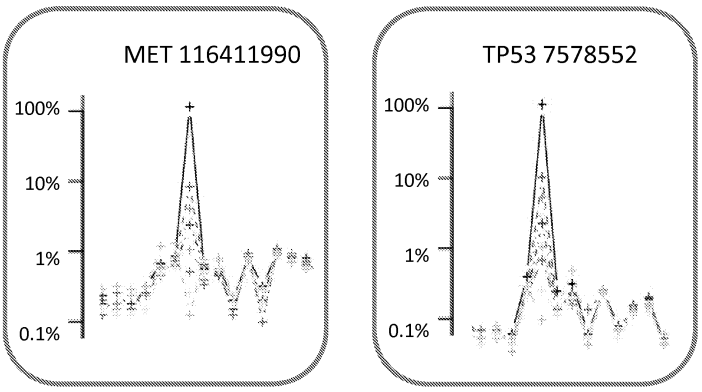


도면5b

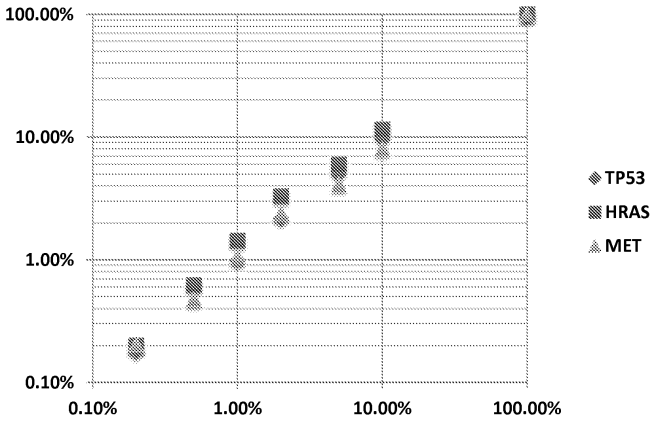


전립선암 환자 3

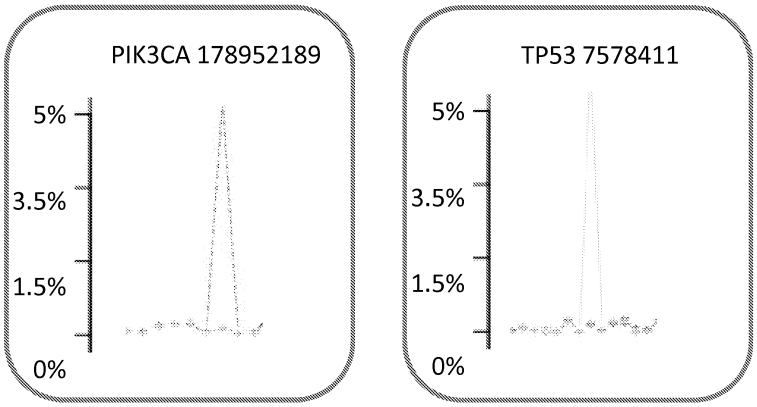
도면6a



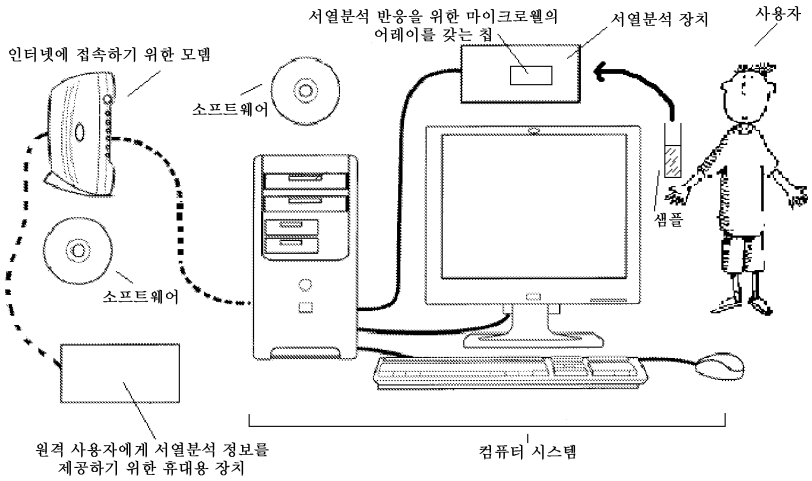
도면6b



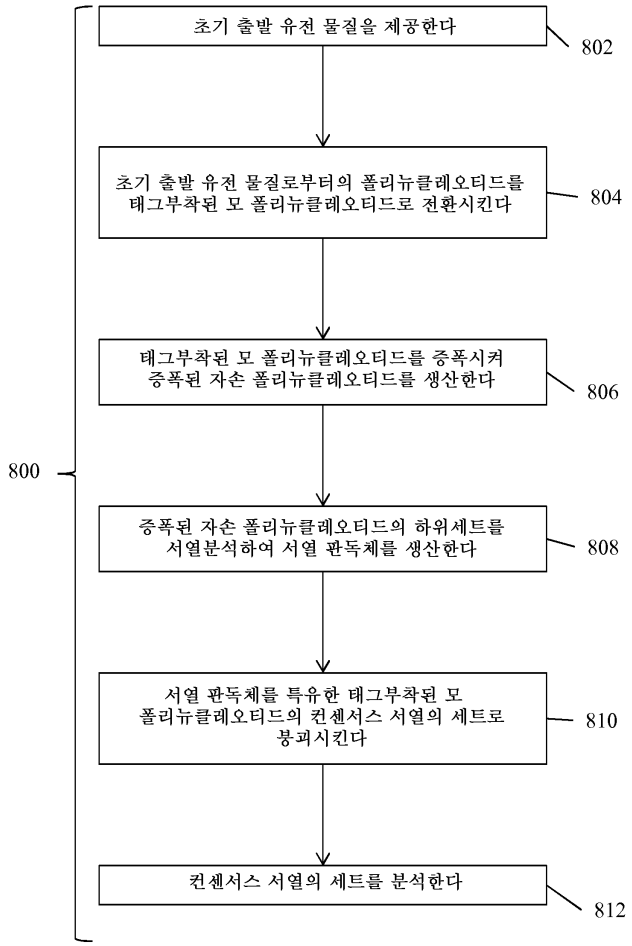
도면7a


PIK3CA 178952189
5%
3.5%
1.5%
0%
TP53 7578411
5%
3.5%
1.5%
0%

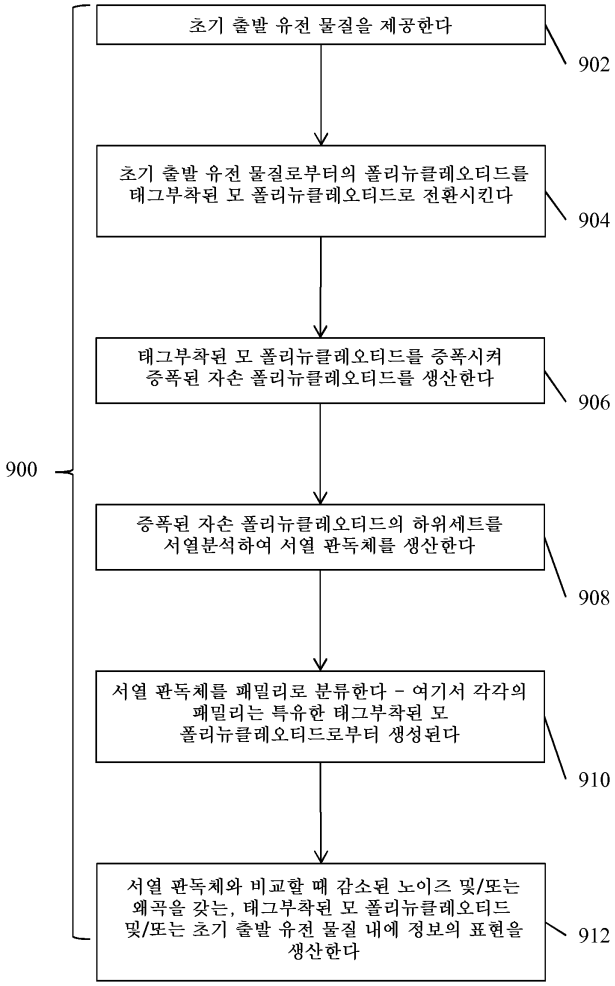
도면7b


서열분석 반응을 위한 마이크로웰의 어레이를 갖는 칩
서열분석 장치
사용자
인터넷에 접속하기 위한 모뎀
소프트웨어
샘플
소프트웨어
원격 사용자에게 서열분석 정보를 제공하기 위한 휴대용 장치
컴퓨터 시스템

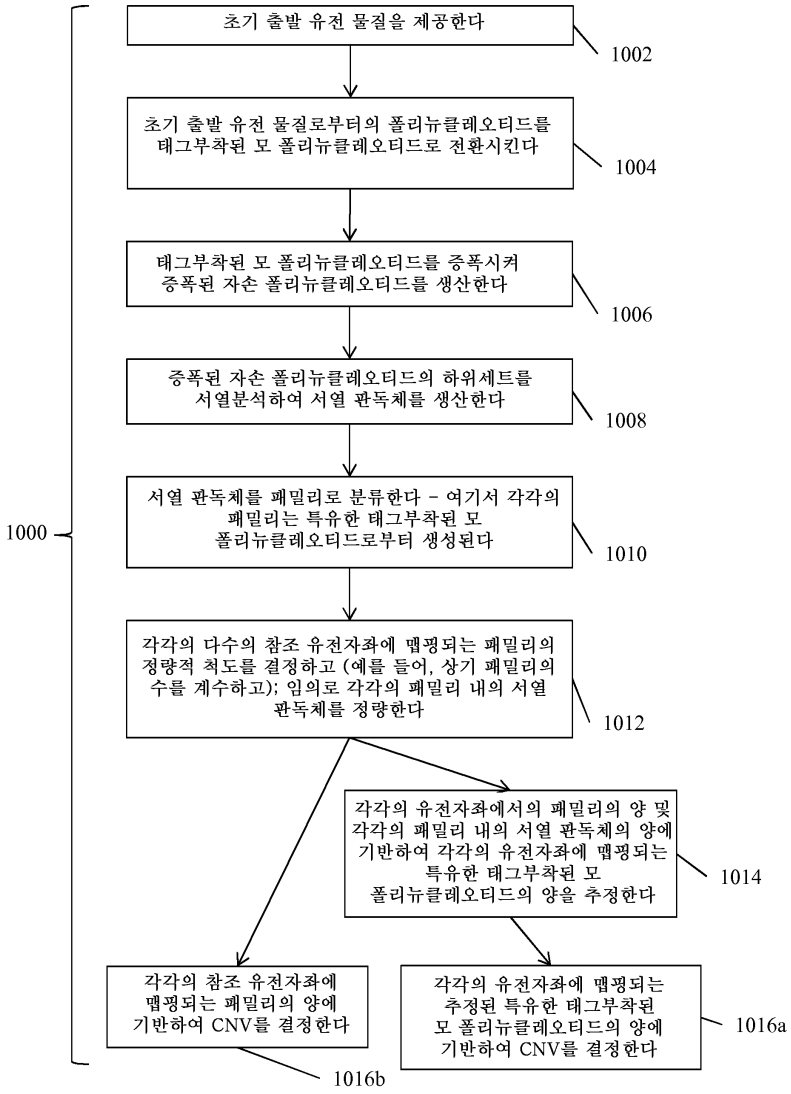
도면8


초기 출발 유전 물질을 제공한다
802
초기 출발 유전 물질로부터의 폴리뉴클레오티드를 태그부착된 모 폴리뉴클레오티드로 전환시킨다
804
태그부착된 모 폴리뉴클레오티드를 증폭시켜 증폭된 자손 폴리뉴클레오티드를 생산한다
806
800
증폭된 자손 폴리뉴클레오티드의 하위세트를 서열분석하여 서열 판독체를 생산한다
808
서열 판독체를 특유한 태그부착된 모 폴리뉴클레오티드의 컨센서스 서열의 세트로 붕괴시킨다
810
컨센서스 서열의 세트를 분석한다
812

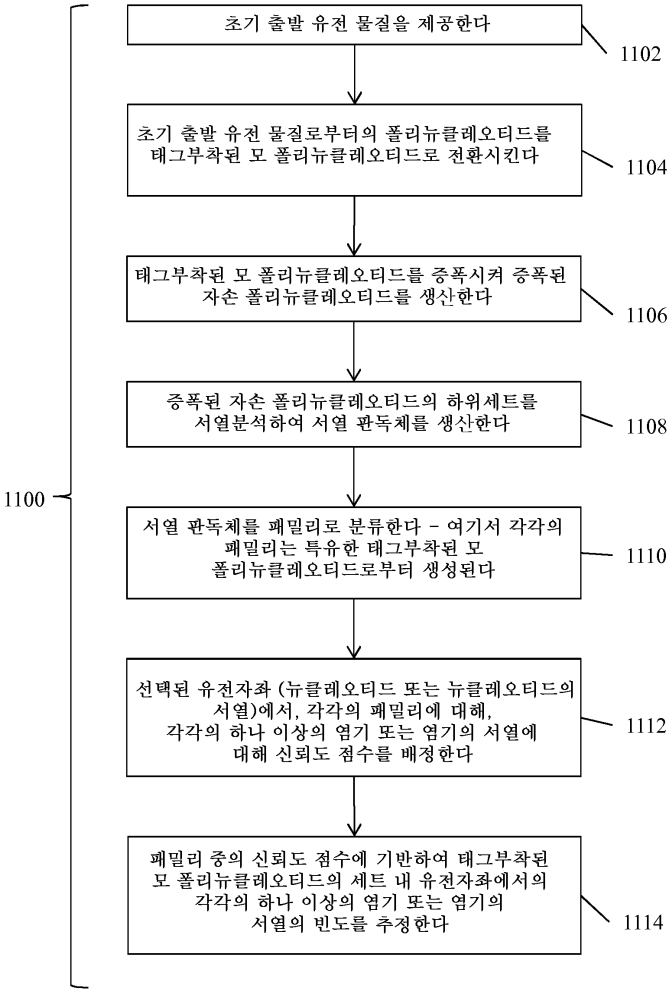
도면9


초기 출발 유전 물질을 제공한다
902
초기 출발 유전 물질로부터의 폴리뉴클레오티드를 태그부착된 모 폴리뉴클레오티드로 전환시킨다
904
태그부착된 모 폴리뉴클레오티드를 증폭시켜 증폭된 자손 폴리뉴클레오티드를 생산한다
906
900
증폭된 자손 폴리뉴클레오티드의 하위세트를 서열분석하여 서열 판독체를 생산한다
908
서열 판독체를 패밀리로 분류한다 - 여기서 각각의 패밀리는 특유한 태그부착된 모 폴리뉴클레오티드로부터 생성된다
910
서열 판독체와 비교할 때 감소된 노이즈 및/또는 왜곡을 갖는, 태그부착된 모 폴리뉴클레오티드 및/또는 초기 출발 유전 물질 내에 정보의 표현을 생산한다
912

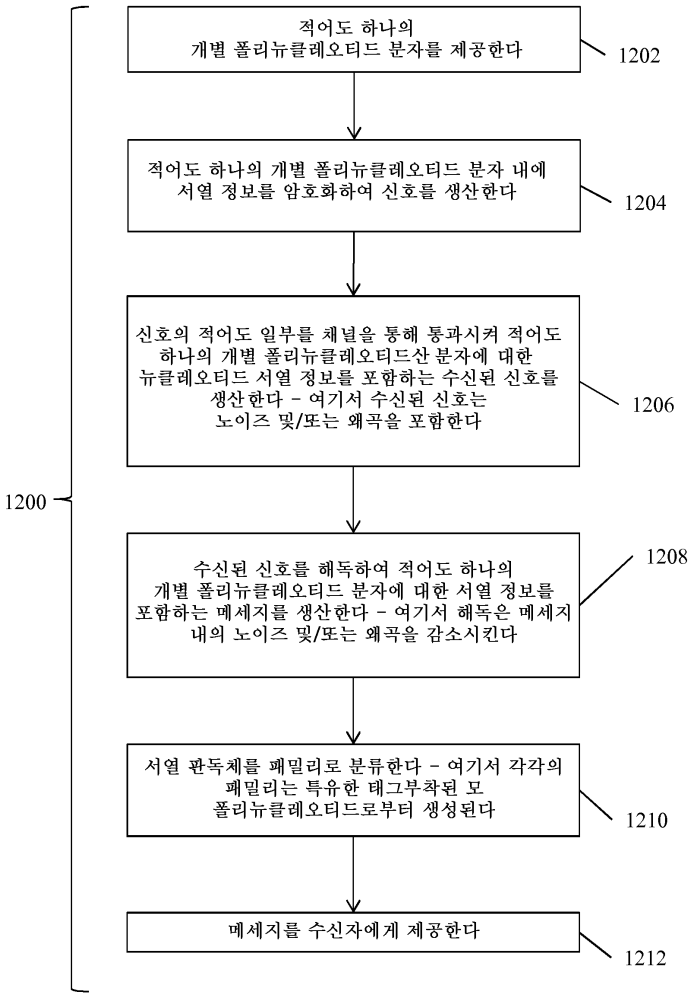
도면10


초기 출발 유전 물질을 제공한다
1002
초기 출발 유전 물질로부터의 폴리뉴클레오티드를 태그부착된 모 폴리뉴클레오티드로 전환시킨다
1004
태그부착된 모 폴리뉴클레오티드를 증폭시켜 증폭된 자손 폴리뉴클레오티드를 생산한다
1006
증폭된 자손 폴리뉴클레오티드의 하위세트를 서열분석하여 서열 판독체를 생산한다
1008
서열 판독체를 패밀리로 분류한다 - 여기서 각각의 패밀리는 특유한 태그부착된 모 폴리뉴클레오티드로부터 생성된다
1010
1000
각각의 다수의 참조 유전자좌에 맵핑되는 패밀리의 정량적 척도를 결정하고 (예를 들어, 상기 패밀리의 수를 계수하고); 임의로 각각의 패밀리 내의 서열 판독체를 정량한다
1012
각각의 유전자좌에서의 패밀리의 양 및 각각의 패밀리 내의 서열 판독체의 양에 기반하여 각각의 유전자좌에 맵핑되는 특유한 태그부착된 모 폴리뉴클레오티드의 양을 추정한다
1014
각각의 참조 유전자좌에 맵핑되는 패밀리의 양에 기반하여 CNV를 결정한다
1016b
각각의 유전자좌에 맵핑되는 추정된 특유한 태그부착된 모 폴리뉴클레오티드의 양에 기반하여 CNV를 결정한다
1016a

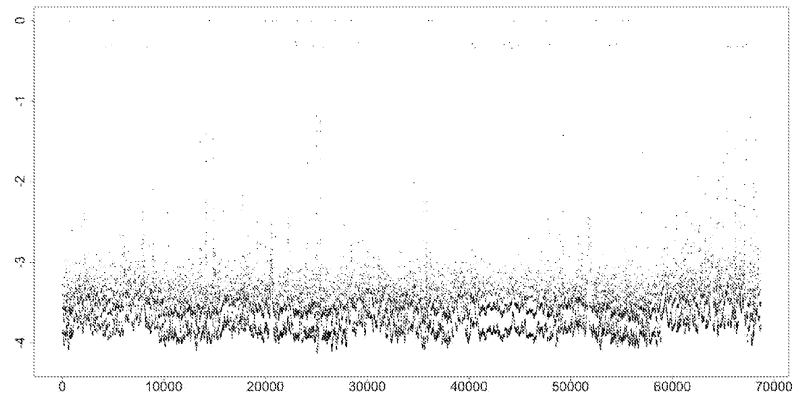
도면11


초기 출발 유전 물질을 제공한다
1102
초기 출발 유전 물질로부터의 폴리뉴클레오티드를 태그부착된 모 폴리뉴클레오티드로 전환시킨다
1104
태그부착된 모 폴리뉴클레오티드를 증폭시켜 증폭된 자손 폴리뉴클레오티드를 생산한다
1106
증폭된 자손 폴리뉴클레오티드의 하위세트를 서열분석하여 서열 판독체를 생산한다
1108
1100
서열 판독체를 패밀리로 분류한다 - 여기서 각각의 패밀리는 특유한 태그부착된 모 폴리뉴클레오티드로부터 생성된다
1110
선택된 유전자좌 (뉴클레오티드 또는 뉴클레오티드의 서열)에서, 각각의 패밀리에 대해, 각각의 하나 이상의 염기 또는 염기의 서열에 대해 신뢰도 점수를 배정한다
1112
패밀리 중의 신뢰도 점수에 기반하여 태그부착된 모 폴리뉴클레오티드의 세트 내 유전자좌에서의 각각의 하나 이상의 염기 또는 염기의 서열의 빈도를 추정한다
1114

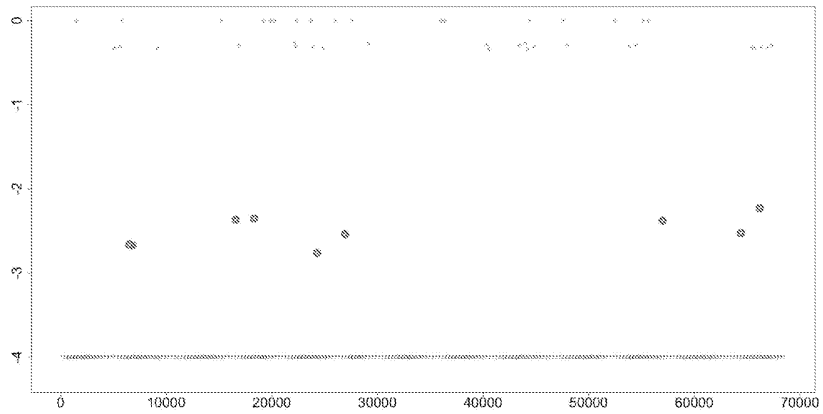
도면12


적어도 하나의
개별 폴리뉴클레오티드 분자를 제공한다
1202
적어도 하나의 개별 폴리뉴클레오티드 분자 내에
서열 정보를 암호화하여 신호를 생산한다
1204
신호의 적어도 일부를 채널을 통해 통과시켜 적어도
하나의 개별 폴리뉴클레오티드산 분자에 대한
뉴클레오티드 서열 정보를 포함하는 수신된 신호를
생산한다 - 여기서 수신된 신호는
노이즈 및/또는 왜곡을 포함한다
1206
1200
수신된 신호를 해독하여 적어도 하나의
개별 폴리뉴클레오티드 분자에 대한 서열 정보를
포함하는 메세지를 생산한다 - 여기서 해독은 메세지
내의 노이즈 및/또는 왜곡을 감소시킨다
1208
서열 판독체를 패밀리로 분류한다 - 여기서 각각의
패밀리는 특유한 태그부착된 모
폴리뉴클레오티드로부터 생성된다
1210
메세지를 수신자에게 제공한다
1212

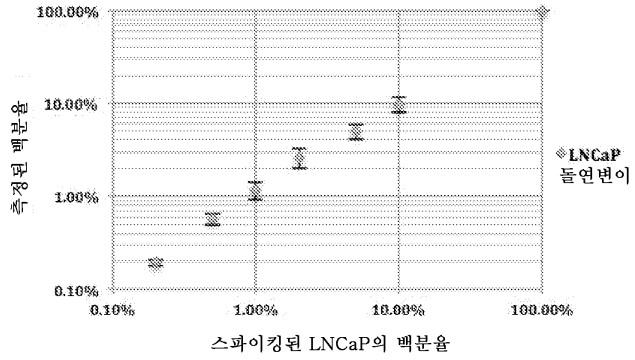
도면13a


0
-1
-2
-3
-4
0
10000
20000
30000
40000
50000
60000
70000

도면13b


0
-1
-2
-3
-4
0
10000
20000
30000
40000
50000
60000
70000

도면14


100.00%
10.00%
1.00%
0.10%
측정된 백분율
0.10%
1.00%
10.00%
100.00%
LNCaP
돌연변이
스파이킹된 LNCaP의 백분율

도면15