(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0117036
(43) 공개일자 2023년08월07일

(51) 국제특허분류(Int. Cl.)
G16B 45/00 (2019.01) *G16B 20/10* (2019.01)
G16B 20/20 (2019.01) *G16B 30/10* (2019.01)
G16B 30/20 (2019.01) *G16B 40/20* (2019.01)
(52) CPC특허분류
G16B 45/00 (2019.02)
G16B 20/10 (2019.02)
(21) 출원번호 10-2022-7045307
(22) 출원일자(국제) 2021년12월10일
심사청구일자 없음
(85) 번역문제출일자 2022년12월22일
(86) 국제출원번호 PCT/US2021/062963
(87) 국제공개번호 WO 2022/125995
국제공개일자 2022년06월16일
(30) 우선권주장
63/124,622 2020년12월11일 미국(US)

(71) 출원인
일루미나, 인코포레이티드
미국 캘리포니아 92122 샌디에고 일루미나 웨이 5200
(72) 발명자
돌젠코 이고르
미국 캘리포니아주 92122 샌디에고 일루미나 웨이 5200
에버럴 마이클 에이.
미국 캘리포니아주 92122 샌디에고 일루미나 웨이 5200
(74) 대리인
특허법인아주김장리

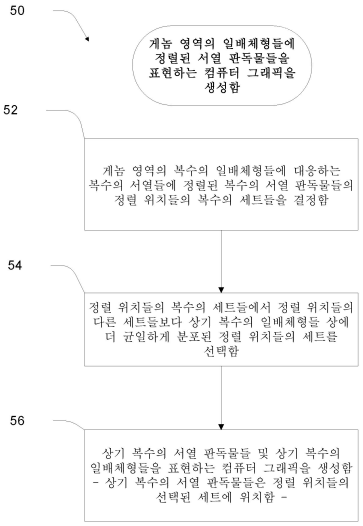
전체 청구항 수 : 총 77 항

(54) 발명의 명칭 **게놈의 반복 영역들에서의 짧은 판독물들을 시각화하기 위한 방법들 및 시스템들**

(57) 요 약

개시된 실시형태들은 의학적으로 중요한 짧은 탠덤 반복체(STR)들과 같은 반복 서열들을 유전자형분석하고 시각화하기 위한 방법들, 장치들, 시스템들 및 컴퓨터 프로그램 제품들에 관한 것이다. 일부 구현예들은 2개 이상의 반복 하위서열들을 각각 포함하는 반복 서열들을 유전자형분석하고 시각화하는 데 사용될 수 있다. 일부 구현예들은 반복 서열의 상이한 유전자형들을 갖는 샘플들에 대한 반복 서열들을 시각화하기 위한 서열 판독물 파일업들을 생성하는 컴퓨터 도구를 제공하며, 각각의 서열 파일업은 2개 이상의 상이한 일배체형들에 정렬된 판독물들을 포함한다.

대 표 도 - 도5



(52) CPC특허분류
G16B 20/20 (2019.02)
G16B 30/10 (2019.02)
G16B 30/20 (2019.02)
G16B 40/20 (2019.02)

명 세 서

청구범위

청구항 1

게놈 영역의 일배체형들에 정렬된 서열 판독물들을 표현하는 컴퓨터 그래픽을 생성하기 위한 시스템으로서, 상기 시스템은 하나 이상의 프로세서들 및 시스템 메모리를 포함하고, 상기 하나 이상의 프로세서들은,

(a) 게놈 영역의 복수의 일배체형들에 대응하는 복수의 일배체형 서열들 상의 정렬 위치들의 세트에 복수의 서열 판독물들을 정렬시키고 - 상기 복수의 서열 판독물들은 핵산 샘플의 게놈 영역으로부터 획득됨 -;

(b) 정렬 위치들의 세트에 대한 정렬 점수를 추정하고;

(c) 복수의 반복 동안 (a) 내지 (b)를 반복하여 정렬 위치들의 복수의 상이한 세트들에 대한 복수의 정렬 점수들을 획득하고;

(d) 상기 복수의 정렬 점수들에 기초하여 정렬 위치들의 복수의 상이한 세트들로부터 정렬 위치들의 세트를 선택하고;

(e) 상기 복수의 서열 판독물들 및 상기 복수의 일배체형들을 표현하는 컴퓨터 그래픽을 생성하도록 구성되는 - 상기 복수의 서열 판독물들은 (d)에서 선택된 정렬 위치들의 세트에서 상기 복수의 일배체형들에 정렬됨 -, 게놈 영역의 일배체형들에 정렬된 서열 판독물들을 표현하는 컴퓨터 그래픽을 생성하기 위한 시스템.

청구항 2

제1항에 있어서, 상기 정렬 점수는 상기 복수의 서열 판독물들이 상기 복수의 일배체형 서열들 상에 얼마나 균일하게 분포되어 있는지를 나타내는, 시스템.

청구항 3

제1항 또는 제2항에 있어서, 상기 게놈 영역은 하나 이상의 탠덤 반복체들을 포함하는, 시스템.

청구항 4

제1항 내지 제3항 중 어느 한 항에 있어서, 상기 복수의 일배체형들 중 적어도 하나의 일배체형은 반복 확장을 포함하는, 시스템.

청구항 5

제1항 내지 제4항 중 어느 한 항에 있어서, 각각의 일체배형은 대립유전자를 포함하는, 시스템.

청구항 6

제1항 내지 제5항 중 어느 한 항에 있어서, 상기 복수의 일배체형들은 2개의 일배체형들을 포함하는, 시스템.

청구항 7

제1항 내지 제6항 중 어느 한 항에 있어서, 정렬 위치들의 선택된 세트는 상이한 정렬 위치들의 복수의 세트들 중에서 최상의 정렬 점수를 갖는, 시스템.

청구항 8

제1항 내지 제7항 중 어느 한 항에 있어서, 정렬 위치들의 선택된 세트는 선택 기준을 초과하는 정렬 점수를 갖는, 시스템.

청구항 9

제1항 내지 제8항 중 어느 한 항에 있어서, 상기 복수의 일배체형들 중 적어도 하나의 일배체형은 구조적 변이체를 포함하는, 시스템.

청구항 10

제9항에 있어서, 상기 구조적 변이체는 50 bp보다 길고, 결실, 중복, 복제수 변이체, 삽입, 역위, 전좌 및 이들의 임의의 조합으로 구성된 군으로부터 선택되는, 시스템.

청구항 11

제9항에 있어서, 상기 구조적 변이체는 50 bp보다 짧은 변이체를 포함하는, 시스템.

청구항 12

제11항에 있어서, 50 bp보다 짧은 변이체는 단일 뉴클레오티드 다형성(SNP)을 포함하는, 시스템.

청구항 13

제1항 내지 제12항 중 어느 한 항에 있어서, (a)는,

각각의 일배체형에 대한 각각의 판독물의 가능한 정렬 위치들을 결정하는 것 - 상기 복수의 서열 판독물들은 쌍형성된 말단 서열분석(paired-end sequencing)에 의해 획득된 판독물 쌍들을 포함함 -;

(ii) (A) 판독물 쌍의 두 판독물이 동일한 일배체형에 정렬되고 (B) 판독물 쌍의 대응하는 단편 길이가 평균 단편 길이와 최대한 가깝도록, 구성 판독물들의 정렬 위치들로부터 각각의 판독물 쌍에 대한 제한된 정렬 위치들을 생성하는 것; 및

(iii) 상기 제한된 정렬 위치들로부터 각각의 판독물 쌍에 대한 정렬 위치를 무작위로 선택하는 것을 포함하는, 시스템.

청구항 14

제1항 내지 제13항 중 어느 한 항에 있어서, 상기 정렬 점수는 2개의 연속적인 판독물들의 시작 위치들 간의 거리의 평균으로부터의 제곱 평균 제곱근 차이를 포함하는, 시스템.

청구항 15

제1항 내지 제14항 중 어느 한 항에 있어서, 상기 정렬 점수는 판독물 쌍들이 상기 복수의 일배체형 서열들 상에 균일하게 분포된다고 가정하는 확률 모델을 사용하여 추정되는, 시스템.

청구항 16

제15항에 있어서, 상기 정렬 점수는, 상기 확률 모델을 고려하여 상기 복수의 서열 판독물들이 정렬 위치들의 세트로부터 도출되는 확률을 포함하는, 시스템.

청구항 17

제16항에 있어서, 상기 복수의 서열 판독물들은 핵산 단편들로부터 획득된 쌍-말단 판독물들을 포함하고, 상기 확률 모델은 평균 단편 길이를 입력으로서 수신하도록 구성되는, 시스템.

청구항 18

제16항에 있어서, 상기 확률 모델은 일배체형의 길이를 입력으로서 수신하도록 구성되는, 시스템.

청구항 19

제16항에 있어서, $p_{(k)}$로 표시되는, 일배체형의 시작으로부터의 k번째 판독물 쌍의 개별 정렬 위치 x의 확률은 하기와 같이 모델링되며:

$$\mathcal{P}(p_{(k)} = x) = \sum_{j=0}^{n_i-k} \binom{n_i}{j} \Big(\big(1 - F(x)\big)^j \big(F(x)\big)^{n_i-j} - \big(1 - F(x) + f(x)\big)^j \big(F(x) - f(x)\big)^{n_i-j} \Big)$$

상기 식에서,

i는 판독물 쌍이 정렬된 일배체형이고,

$F(x) = \frac{x}{H_i - L - 1}$,

$f(x) = \frac{1}{H_i - L - 1}$,

H_i는 일배체형 i의 길이이고,

L은 평균 단편 길이이고,

n_i는 일배체형 i에 정렬된 판독물 쌍들의 개수인, 시스템.

청구항 20

제19항에 있어서, 정렬 위치들의 세트에 대한 정렬 점수는 개별 정렬 위치들의 확률의 곱으로서 추정되는, 시스템.

청구항 21

제1항 또는 제2항에 있어서, 상기 하나 이상의 프로세서들은 정렬 위치들의 선택된 세트에서 상기 복수의 일배체형 서열들에 정렬된 복수의 서열 판독물들에 대한 하나 이상의 서열분석 메트릭들을 추정하도록 추가로 구성되는, 시스템.

청구항 22

제21항에 있어서, 상기 하나 이상의 서열분석 메트릭들은 서열 커버리지(sequence coverage)를 포함하는, 시스템.

청구항 23

제22항에 있어서, 상기 하나 이상의 서열분석 메트릭들은 각각의 정렬 위치에 대한 서열 커버리지를 포함하는, 시스템.

청구항 24

제21항에 있어서, 상기 하나 이상의 서열분석 메트릭들은 정렬 품질 점수를 포함하는, 시스템.

청구항 25

제24항에 있어서, 상기 하나 이상의 서열분석 메트릭들은 각각의 정렬 위치에 대한 정렬 품질 점수를 포함하는, 시스템.

청구항 26

제21항에 있어서, 상기 하나 이상의 서열분석 메트릭들은 맵핑 품질 점수를 포함하는, 시스템.

청구항 27

제1항 내지 제26항 중 어느 한 항에 있어서, 상기 복수의 서열 판독물들은 적어도 100개의 서열 판독물들을 포함하는, 시스템.

청구항 28

제1항 내지 제27항 중 어느 한 항에 있어서, 상기 하나 이상의 프로세서들은 서열 판독물들의 상이한 세트들을 사용하여 상이한 게놈 영역들에 대해 동작 (a)를 수행하도록 구성되는, 시스템.

청구항 29

제28항에 있어서, 상기 상이한 게놈 영역들은 적어도 100개의 상이한 게놈 영역들을 포함하는, 시스템.

청구항 30

제1항 내지 제29항 중 어느 한 항에 있어서, 상기 하나 이상의 프로세서들은, 동작 (a) 전에, 게놈 영역에 대응하는 하나 이상의 서열 그래프들에 제1 개수의 서열 판독물들을 정렬시켜 상기 복수의 서열 판독물들 및/또는 상기 복수의 일배체형들을 획득하도록 구성되는, 시스템.

청구항 31

제30항에 있어서, 상기 서열 그래프에 상기 제1 개수의 서열 판독물들을 정렬시키는 것은,

핵산 샘플의 제1 개수의 서열 판독물들을 제공하는 것;

서열 그래프에 의해 각각 표현된 하나 이상의 반복 서열들에 상기 제1 개수의 서열 판독물들을 정렬시키는 것 - 상기 서열 그래프는 핵산 서열들을 표현하는 정점들 및 상기 정점들을 연결하는 방향성 에지(edge)들을 갖는 방향성 그래프의 데이터 구조를 갖고, 상기 서열 그래프는 하나 이상의 자가-루프(self-loop)들을 포함하며, 각각의 자가-루프는 반복 하위서열을 표현하고, 각각의 반복 하위서열은 하나 이상의 뉴클레오티드들의 반복 단위의 반복체들을 포함함 -;

상기 하나 이상의 반복 서열들의 하나 이상의 유전자형들을 결정하는 것; 및

(a)의 복수의 서열 판독물들로서의 제1 개수의 서열 판독물들 및/또는 상기 하나 이상의 반복 서열들의 하나 이상의 유전자형들을 제공하는 것을 포함하는, 시스템.

청구항 32

제31항에 있어서, 상기 하나 이상의 프로세서들은 상기 하나 이상의 반복 서열들의 하나 이상의 유전자형들을 단계적으로 분석하여 (b)의 복수의 일배체형들을 결정하도록 구성되는, 시스템.

청구항 33

제31항에 있어서, 상기 하나 이상의 프로세서들은 초기에 게놈에 제2 개수의 서열 판독물들을 정렬시켜 상기 제1 개수의 서열 판독물들을 제공하도록 구성되고, 상기 제2 개수의 서열 판독물들은 적어도 10,000개의 서열 판독물들을 포함하는, 시스템.

청구항 34

컴퓨터 그래픽을 생성하기 위한 방법으로서, 하나 이상의 프로세서들 및 시스템 메모리를 포함하는 컴퓨터를 사용하여 구현되고, 상기 방법은,

(a) 상기 하나 이상의 프로세서들을 사용하여, 게놈 영역의 복수의 일배체형들에 대응하는 복수의 일배체형 서열들 상의 정렬 위치들의 세트에 복수의 서열 판독물들을 정렬시키는 단계 - 상기 복수의 서열 판독물들은 핵산 샘플의 게놈 영역으로부터 획득됨 -;

(b) 상기 하나 이상의 프로세서들에 의해, 정렬 위치들의 세트에 대한 정렬 점수를 추정하는 단계;

(c) 복수의 반복 동안 (a) 내지 (b)를 반복하여 정렬 위치들의 복수의 상이한 세트들에 대한 복수의 정렬 점수들을 획득하는 단계;

(d) 상기 하나 이상의 프로세서들에 의해, 상기 복수의 정렬 점수들에 기초하여 정렬 위치들의 복수의 상이한 세트들로부터 정렬 위치들의 세트를 선택하는 단계; 및

(e) 상기 하나 이상의 프로세서들을 사용하여, 상기 복수의 서열 판독물들 및 상기 복수의 일배체형들을 표현하는 컴퓨터 그래픽을 생성하는 단계 - 상기 복수의 서열 판독물들은 (d)에서 선택된 정렬 위치들의 세트에서 상기 복수의 일배체형들에 정렬됨 -를 포함하는, 컴퓨터 그래픽을 생성하기 위한 방법.

청구항 35

제34항에 있어서, 상기 정렬 점수는 상기 복수의 서열 판독물들이 상기 복수의 일배체형 서열들 상에 얼마나 균일하게 분포되어 있는지를 나타내는, 방법.

청구항 36

제34항에 있어서, 상기 게놈 영역은 하나 이상의 탠덤 반복체들을 포함하는, 방법.

청구항 37

제34항 내지 제36항 중 어느 한 항에 있어서, 상기 복수의 일배체형들 중 적어도 하나의 일배체형은 반복 확장을 포함하는, 방법.

청구항 38

제34항 내지 제37항 중 어느 한 항에 있어서, 각각의 일체배형은 대립유전자를 포함하는, 방법.

청구항 39

제34항 내지 제39항 중 어느 한 항에 있어서, 상기 복수의 일배체형들은 2개의 일배체형들을 포함하는, 방법.

청구항 40

제34항 내지 제39항 중 어느 한 항에 있어서, 정렬 위치들의 선택된 세트는 상이한 정렬 위치들의 복수의 세트들 중에서 최상의 정렬 점수를 갖는, 방법.

청구항 41

제34항 내지 제40항 중 어느 한 항에 있어서, 정렬 위치들의 선택된 세트는 선택 기준을 초과하는 정렬 점수를 갖는, 방법.

청구항 42

제34항 내지 제41항 중 어느 한 항에 있어서, 상기 복수의 일배체형들 중 적어도 하나의 일배체형은 구조적 변이체를 포함하는, 방법.

청구항 43

제42항에 있어서, 상기 구조적 변이체는 50 bp보다 길고, 결실, 중복, 복제수 변이체, 삽입, 역위, 전좌 및 이들의 임의의 조합으로 구성된 군으로부터 선택되는, 방법.

청구항 44

제42항에 있어서, 상기 구조적 변이체는 50 bp보다 짧은 변이체를 포함하는, 방법.

청구항 45

제44항에 있어서, 50 bp보다 짧은 변이체는 단일 뉴클레오티드 다형성(SNP)을 포함하는, 방법.

청구항 46

제34항 내지 제45항 중 어느 한 항에 있어서, (a)는,

각각의 일배체형에 대한 각각의 판독물의 가능한 정렬 위치들을 결정하는 단계 - 상기 복수의 서열 판독물들은 쌍형성된 말단 서열분석에 의해 획득된 판독물 쌍들을 포함함 -;

(ii) (A) 판독물 쌍의 두 판독물이 동일한 일배체형에 정렬되고 (B) 판독물 쌍의 대응하는 단편 길이가 평균 단편 길이와 최대한 가깝도록, 구성 판독물들의 정렬 위치들로부터 각각의 판독물 쌍에 대한 제한된 정렬 위치들을 생성하는 단계; 및

(iii) 상기 제한된 정렬 위치들로부터 각각의 판독물 쌍에 대한 정렬 위치를 무작위로 선택하는 단계를 포함하는, 방법.

청구항 47

제34항 내지 제47항 중 어느 한 항에 있어서, 상기 정렬 점수는 2개의 연속적인 판독물들의 시작 위치들 간의

거리의 평균으로부터의 제곱 평균 제곱근 차이를 포함하는, 방법.

청구항 48

제34항 내지 제47항 중 어느 한 항에 있어서, 상기 정렬 점수는 판독물 쌍들이 상기 복수의 일배체형 서열들 상에 균일하게 분포된다고 가정하는 확률 모델을 사용하여 추정되는, 방법.

청구항 49

제48항에 있어서, 상기 정렬 점수는, 상기 확률 모델을 고려하여 상기 복수의 서열 판독물들이 정렬 위치들의 세트로부터 도출되는 확률을 포함하는, 방법.

청구항 50

제49항에 있어서, 상기 복수의 서열 판독물들은 핵산 단편들로부터 획득된 쌍-말단 판독물들을 포함하고, 상기 확률 모델은 평균 단편 길이를 입력으로서 수신하도록 구성되는, 방법.

청구항 51

제49항에 있어서, 상기 확률 모델은 일배체형의 길이를 입력으로서 수신하도록 구성되는, 방법.

청구항 52

제49항에 있어서, $p_{(k)}$로 표시되는, 일배체형의 시작으로부터의 k번째 판독물 쌍의 개별 정렬 위치 x의 확률은 하기와 같이 모델링되며:

$$\mathcal{P}\left(p_{(k)}=x\right)=\sum_{j=0}^{n_i-k}\binom{n_i}{j}\Big(\big(1-F(x)\big)^{j}\big(F(x)\big)^{n_i-j} - \big(1-F(x)+f(x)\big)^{j}\big(F(x)-f(x)\big)^{n_i-j}\Big)$$

상기 식에서,

i는 판독물 쌍이 정렬된 일배체형이고,

$F(x)=\frac{x}{H_i-L-1}$,

$f(x)=\frac{1}{H_i-L-1}$,

H_i는 일배체형 i의 길이이고,

L은 평균 단편 길이이고,

n_i는 일배체형 i에 정렬된 판독물 쌍들의 개수인, 방법.

청구항 53

제52항에 있어서, 정렬 위치들의 세트에 대한 정렬 점수는 개별 정렬 위치들의 확률의 곱으로서 추정되는, 방법.

청구항 54

제34항 내지 제53항 중 어느 한 항에 있어서, 정렬 위치들의 선택된 세트에서 상기 복수의 일배체형 서열들에 정렬된 복수의 서열 판독물들에 대한 하나 이상의 서열분석 메트릭들을 추정하는 단계를 추가로 포함하는, 방법.

청구항 55

제54항에 있어서, 상기 하나 이상의 서열분석 메트릭들은 서열 커버리지를 포함하는, 방법.

청구항 56

제55항에 있어서, 상기 하나 이상의 서열분석 메트릭들은 각각의 정렬 위치에 대한 서열 커버리지를 포함하는, 방법.

청구항 57

제54항에 있어서, 상기 하나 이상의 서열분석 메트릭들은 정렬 품질 점수를 포함하는, 방법.

청구항 58

제57항에 있어서, 상기 하나 이상의 서열분석 메트릭들은 각각의 정렬 위치에 대한 정렬 품질 점수를 포함하는, 방법.

청구항 59

제54항에 있어서, 상기 하나 이상의 서열분석 메트릭들은 맵핑 품질 점수를 포함하는, 방법.

청구항 60

제34항에 있어서, 상기 복수의 서열 판독물들은 적어도 100개의 서열 판독물들을 포함하는, 방법.

청구항 61

제34항 내지 제61항 중 어느 한 항에 있어서, 서열 판독물들의 상이한 세트들을 사용하여 상이한 게놈 영역들에 대해 동작 (a)를 수행하는 단계를 추가로 포함하는, 방법.

청구항 62

제61항에 있어서, 상기 상이한 게놈 영역들은 적어도 100개의 상이한 게놈 영역들을 포함하는, 방법.

청구항 63

제34항 내지 제63항 중 어느 한 항에 있어서, 동작 (a) 전에, 게놈 영역에 대응하는 하나 이상의 서열 그래프들에 제1 개수의 서열 판독물들을 정렬시켜 상기 복수의 서열 판독물들 및/또는 상기 복수의 일배체형들을 획득하는 단계를 추가로 포함하는, 방법.

청구항 64

제63항에 있어서, 상기 서열 그래프에 상기 제1 개수의 서열 판독물들을 정렬시키는 단계는,

핵산 샘플의 제1 개수의 서열 판독물들을 제공하는 단계;

서열 그래프에 의해 각각 표현된 하나 이상의 반복 서열들에 상기 제1 개수의 서열 판독물들을 정렬시키는 단계 - 상기 서열 그래프는 핵산 서열들을 표현하는 정점들 및 상기 정점들을 연결하는 방향성 에지들을 갖는 방향성 그래프의 데이터 구조를 갖고, 상기 서열 그래프는 하나 이상의 자가-루프들을 포함하며, 각각의 자가-루프는 반복 하위서열을 표현하고, 각각의 반복 하위서열은 하나 이상의 뉴클레오티드들의 반복 단위의 반복체들을 포함함 -;

상기 하나 이상의 반복 서열들의 하나 이상의 유전자형들을 결정하는 단계; 및

(a) 의 복수의 서열 판독물들로서의 제1 개수의 서열 판독물들 및/또는 상기 하나 이상의 반복 서열들의 하나 이상의 유전자형들을 제공하는 단계를 포함하는, 방법.

청구항 65

제64항에 있어서, 상기 하나 이상의 반복 서열들의 하나 이상의 유전자형들을 단계적으로 분석하여 (b)의 복수의 일배체형들을 결정하는 단계를 추가로 포함하는, 방법.

청구항 66

제64항에 있어서, 초기에 게놈에 제2 개수의 서열 판독물들을 정렬시켜 상기 제1 개수의 서열 판독물들을 제공하는 단계 - 상기 제2 개수의 서열 판독물들은 적어도 10,000개의 서열 판독물들을 포함함 -를 추가로 포함하는, 방법.

청구항 67

컴퓨터 제품으로서, 컴퓨터 시스템의 하나 이상의 프로세서들에 의해 실행될 때, 컴퓨터 시스템으로 하여금 하기를 수행하게 하는 컴퓨터-실행가능 명령어들이 저장되어 있는 하나 이상의 컴퓨터 판독가능 비일시적 저장 매체를 포함하는, 컴퓨터 제품:

(a) 게놈 영역의 복수의 일배체형들에 대응하는 복수의 일배체형 서열들 상의 정렬 위치들의 세트에 복수의 서열 판독물들을 정렬시키고 - 상기 복수의 서열 판독물들은 핵산 샘플의 게놈 영역으로부터 획득됨 -;

(b) 정렬 위치들의 세트에 대한 정렬 점수를 추정하고;

(c) 복수의 반복 동안 (a) 내지 (b)를 반복하여 정렬 위치들의 복수의 상이한 세트들에 대한 복수의 정렬 점수들을 획득하고;

(d) 상기 복수의 정렬 점수들에 기초하여 정렬 위치들의 복수의 상이한 세트들로부터 정렬 위치들의 세트를 선택하고;

(e) 상기 복수의 서열 판독물들 및 상기 복수의 일배체형들을 표현하는 컴퓨터 그래픽을 생성함 - 상기 복수의 서열 판독물들은 (d)에서 선택된 정렬 위치들의 세트에서 상기 복수의 일배체형들에 정렬됨 -.

청구항 68

컴퓨터 그래픽을 생성하기 위한 방법으로서, 하나 이상의 프로세서들 및 시스템 메모리를 포함하는 시스템을 사용하여 구현되고, 상기 방법은,

(a) 게놈 영역의 복수의 일배체형들에 대응하는 복수의 일배체형 서열들에 정렬된 복수의 서열 판독물들의 정렬 위치들의 복수의 세트들을 결정하는 단계 - 상기 복수의 서열 판독물들은 핵산 샘플의 게놈 영역으로부터 획득됨 -;

(b) 정렬 위치들의 복수의 세트들에서 정렬 위치들의 다른 세트들보다 상기 복수의 일배체형들 상에 더 균일하게 분포된 정렬 위치들의 세트를 선택하는 단계; 및

(c) 상기 복수의 서열 판독물들 및 상기 복수의 일배체형들을 표현하는 컴퓨터 그래픽을 생성하는 단계 - 상기 복수의 서열 판독물들은 (b)에서 선택된 정렬 위치들의 세트에서 상기 복수의 일배체형들에 정렬됨 -를 포함하는, 컴퓨터 그래픽을 생성하기 위한 방법.

청구항 69

제68항에 있어서, 상기 복수의 서열 판독물들은 핵산 단편들로부터 획득된 쌍-말단 판독물들을 포함하는, 방법.

청구항 70

제68항 또는 제69항에 있어서, 상기 복수의 일배체형들은 2개의 일배체형들인, 방법.

청구항 71

제68항 내지 제70항 중 어느 한 항에 있어서, 상기 게놈 영역은 하나 이상의 탠덤 반복체들을 포함하는, 방법.

청구항 72

제71항에 있어서, 상기 복수의 일배체형들 중 적어도 하나의 일배체형은 반복 확장을 포함하는, 방법.

청구항 73

제68항 내지 제72항 중 어느 한 항에 있어서, 상기 복수의 일배체형들 중 적어도 하나의 일배체형은 구조적 변이체를 포함하는, 방법.

청구항 74

제73항에 있어서, 상기 구조적 변이체는 50 bp보다 길고, 결실, 중복, 복제수 변이체, 삽입, 역위, 전좌 및 이들의 임의의 조합으로 구성된 군으로부터 선택되는, 방법.

청구항 75

게놈 영역에 서열 판독물들을 정렬시키는 방법으로서, 하나 이상의 프로세서들 및 시스템 메모리를 포함하는 시스템을 사용하여 구현되고, 상기 방법은,

(a) 게놈 영역의 복수의 일배체형들에 대응하는 복수의 일배체형 서열들 상의 정렬 위치들의 세트에 복수의 서열 판독물들을 정렬시키는 단계 - 상기 복수의 서열 판독물들은 핵산 샘플의 게놈 영역으로부터 획득됨 -;

(b) 정렬 위치들의 세트에 대한 정렬 점수를 추정하는 단계;

(c) 복수의 반복 동안 (a) 내지 (b)를 반복하여 정렬 위치들의 복수의 상이한 세트들에 대한 복수의 정렬 점수들을 획득하는 단계;

(d) 상기 복수의 정렬 점수들에 기초하여 정렬 위치들의 복수의 상이한 세트들로부터 정렬 위치들의 세트를 선택하는 단계; 및

(e) 상기 복수의 서열 판독물들의 최종 정렬 위치가 (d)에서 선택된 정렬 위치들의 세트인 것으로 결정하는 단계를 포함하는, 게놈 영역에 서열 판독물들을 정렬시키는 방법.

청구항 76

시스템으로서, 하나 이상의 프로세서들 및 시스템 메모리를 포함하고, 상기 하나 이상의 프로세서들은 제68항 내지 제75항의 방법들 중 어느 하나를 수행하도록 구성되는, 시스템.

청구항 77

컴퓨터 제품으로서, 컴퓨터 시스템의 하나 이상의 프로세서들에 의해 실행될 때, 컴퓨터 시스템으로 하여금 제68항 내지 제75항의 방법들 중 어느 하나를 수행하게 하는 컴퓨터-실행가능 명령어들이 저장되어 있는 하나 이상의 컴퓨터 판독가능 비일시적 저장 매체를 포함하는, 컴퓨터 제품.

발명의 설명

기 술 분 야

[0001] 참조에 의한 통합

[0002] PCT 요청 양식이 본원의 일부로서 본 명세서와 동시에 제출된다. 동시 제출된 PCT 요청 양식에서 식별된 바와 같은, 본원이 이득 또는 우선권을 주장하는 각각의 출원은 그 전체가 모든 목적을 위해 본원에 참조로 포함된다.

배 경 기 술

[0003] 탠덤 반복체(TR: tandem repeat)는 특정 DNA 모티프들이 반복되는 반복 DNA의 영역이다. 특정 관례에서, 본원에서 반복 단위로도 지칭되는 반복된 모티프들이 10개 미만의 염기쌍들을 포함하는 경우, TR은 짧은 탠덤 반복체(STR: short-tandem repeat) 또는 마이크로부수체(microsatellite)로 간주된다. 반복된 모티프들이 10 내지 60개의 염기쌍 범위인 경우, TR은 미니부수체(minisatellite)로 간주된다.

[0004] 짧은 탠덤 반복체(STR)는 인간 게놈 전체에 걸쳐 어디에나 존재한다. STR 생물학에 대한 우리의 이해는 완전한 것과는 거리가 멀지만, 최근의 증거는 STR들이 기본 세포 과정에서 중요한 역할을 한다는 것을 시사한다.

[0005] 반복 확장(repeat expansion)은 유기체의 TR이 참조 서열보다 더 많은 수의 반복된 모티프들을 갖는 상태이다. 반복 확장은 또한, STR들이 특정 크기를 넘어 확장되는 경우의 이의 불안정성으로 인해 동적 돌연변이로 알려져 있다. STR 확장은 근위축성 측삭경화증(ALS: amyotrophic lateral sclerosis), 프리드라이히 운동실조증(FRDA: Friedreich ataxia), 헌팅턴 병(HD: Huntington's disease), 및 취약 X 증후군(FXS: fragile X syndrome)을 포함하는 많은 중증의 신경학적 장애들의 주요 원인이다.

[0006] 반복 확장들을 식별하는 것은 이들 유전적 장애의 진단 및 치료에 중요하다. 그러나 TR들, 특히 STR들을 결정하는 데에는 많은 기술적 어려움이 존재한다. 일부 어려움은 대규모 병렬 서열분석 기술들에서 일반적으로 사용되는, 반복 서열을 완전히 횡단하지 않는 짧은 서열 판독물과 관련이 있다. 따라서, 의학적으로 관련된 반복 확장을 검출하고 유전자형분석하기 위해 짧은 서열 판독물을 사용하는 방법들을 개발하는 것이 바람직하다.

[0007] 반복 확장들을 검출하고 유전자형분석하는 데 있어서의 수많은 기술적 어려움으로 인해, STR들의 유전자형들 및 상기 유전자형들을 결정하는 데 사용되는 서열 판독 데이터를 시각화하기 위한 컴퓨터 도구들이 매우 필요한 실정이다. 이러한 도구들은 유전자형 호출들을 검증하고 STR들과 관련된 임상적으로 및 생물학적으로 중요한 유전적 특징들을 이해하는 데 도움이 될 수 있다. 본원에 개시된 다양한 구현예들은, 반복 확장들을 검출하고 유전자형분석하고 STR들의 유전자형들 및 상기 유전자형들을 결정하는 데 사용되는 서열 데이터를 시각화하도록 설계되었다.

발명의 내용

[0008] 개시된 구현예들은 유전적 장애들과 상관될 수 있는 짧은 탠덤 반복 서열들과 같은 반복 서열들을 포함하는 게놈 좌위들을 서열분석하고 도표로 시각화하기 위한 방법들, 장치들, 시스템들, 및 컴퓨터 프로그램 제품들에 관한 것이다. 상기 시각화 방법들은 복수의 일배체형들, 특히 반복 서열들을 포함하는 일배체형들에 정렬된 서열 판독물들의 시각적인 표현을 포함하는 서열 파일업(pileup)들을 생성한다.

[0009] 본 개시내용의 제1 양태는 컴퓨터 그래픽을 생성하기 위한 컴퓨터-구현 방법들을 제공하며, 각각의 그래픽은, 예를 들어 탠덤 반복체 또는 구조적 변이체를 포함하는 게놈 영역의 복수의 일배체형들에 정렬된 서열 판독물들을 표현한다. 상기 방법들은 하나 이상의 프로세서들 및 시스템 메모리를 포함하는 컴퓨터를 사용하여 구현된다. 상기 방법들은 하기를 포함한다: (a) 상기 하나 이상의 프로세서들을 사용하여, 게놈 영역의 복수의 일배체형들에 대응하는 복수의 일배체형 서열들 상의 정렬 위치들의 세트에 복수의 서열 판독물들을 정렬시키는 단계 - 상기 복수의 서열 판독물들은 핵산 샘플의 게놈 영역으로부터 획득됨 -; (b) 상기 하나 이상의 프로세서들에 의해, 정렬 위치들의 세트에 대한 정렬 점수를 추정하는 단계; (c) 복수의 반복 동안 (a) 내지 (b)를 반복하여 정렬 위치들의 복수의 상이한 세트들에 대한 복수의 정렬 점수들을 획득하는 단계; (d) 상기 하나 이상의 프로세서들에 의해, 상기 복수의 정렬 점수들에 기초하여 정렬 위치들의 복수의 상이한 세트들로부터 정렬 위치들의 세트를 선택하는 단계; 및 (e) 상기 하나 이상의 프로세서들을 사용하여, 상기 복수의 서열 판독물들 및 상기 복수의 일배체형들을 표현하는 컴퓨터 그래픽을 생성하는 단계 - 상기 복수의 서열 판독물들은 (d)에서 선택된 정렬 위치들의 세트에서 상기 복수의 일배체형들에 정렬됨.

[0010] 일부 구현예들에서, 상기 정렬 점수는 상기 복수의 서열 판독물들이 상기 복수의 일배체형 서열들 상에 얼마나 균일하게 분포되어 있는지를 나타낸다.

[0011] 일부 구현예들에서, 상기 게놈 영역은 하나 이상의 탠덤 반복체들을 포함한다.

[0012] 일부 구현예들에서, 상기 복수의 일배체형들 중 적어도 하나의 일배체형은 반복 확장을 포함한다. 일부 구현예들에서, 각각의 일배체형은 대립유전자를 포함한다. 일부 구현예들에서, 상기 복수의 일배체형들은 2개의 일배체형들을 포함한다.

[0013] 일부 구현예들에서, 정렬 위치들의 선택된 세트는 상이한 정렬 위치들의 복수의 세트들 중에서 최상의 정렬 점수를 갖는다. 일부 구현예들에서, 정렬 위치들의 선택된 세트는 선택 기준을 초과하는 정렬 점수를 갖는다.

[0014] 일부 구현예들에서, 상기 복수의 일배체형들 중 적어도 하나의 일배체형은 구조적 변이체를 포함한다. 일부 구현예들에서, 상기 구조적 변이체는 50 bp보다 길고, 결실, 중복, 복제수 변이체, 삽입, 역위, 전좌 및 이들의 임의의 조합으로 구성된 군으로부터 선택된다. 일부 구현예들에서, 상기 구조적 변이체는 50 bp보다 짧은 변이체를 포함한다. 일부 구현예들에서, 50 bp보다 짧은 변이체는 단일 뉴클레오티드 다형성(SNP: single nucleotide polymorphism)을 포함한다.

[0015] 일부 구현예들에서, (a)는 하기를 포함한다: (i) 각각의 일배체형에 대한 각각의 판독물의 가능한 정렬 위치들을 결정하는 단계 - 상기 복수의 서열 판독물들은 쌍형성된 말단 서열분석(paired-end sequencing)에 의해 획득된 판독물 쌍들을 포함함 -; (ii) (A) 판독물 쌍의 두 판독물이 동일한 일배체형에 정렬되고 (B) 판독물 쌍의 대응하는 단편 길이가 평균 단편 길이와 최대한 가깝도록, 구성 판독물들의 정렬 위치들로부터 각각의 판독물 쌍에 대한 제한된 정렬 위치들을 생성하는 단계; 및 (iii) 상기 제한된 정렬 위치들로부터 각각의 판독물 쌍에 대한 정렬 위치들을 무작위로 선택하는 단계.

[0016] 일부 구현예들에서, 상기 정렬 점수는 2개의 연속적인 판독물들의 시작 위치들 간의 거리의 평균으로부터의 제곱 평균 제곱근 차이를 포함한다.

[0017] 일부 구현예들에서, 상기 정렬 점수는 판독물 쌍들이 상기 복수의 일배체형 서열들 상에 균일하게 분포된다고 가정하는 확률 모델을 사용하여 추정된다. 일부 구현예들에서, 상기 정렬 점수는, 상기 확률 모델을 고려하여 상기 복수의 서열 판독물들이 정렬 위치들의 세트로부터 도출되는 확률을 포함한다. 일부 구현예들에서, 상기 복수의 서열 판독물들은 핵산 단편들로부터 획득된 쌍-말단 판독물들을 포함하고, 상기 확률 모델은 평균 단편 길이를 입력으로서 수신하도록 구성된다. 일부 구현예들에서, 상기 확률 모델은 일배체형의 길이를 입력으로서 수신하도록 구성된다.

[0018] 일부 구현예들에서, $p_{(k)}$로 표시되는, 일배체형의 시작으로부터의 k번째 판독물 쌍의 개별 정렬 위치 x의 확률은 하기와 같이 모델링되며:

$$\mathcal{P}\left(p_{(k)}=x\right)=\sum_{j=0}^{n_i-k}\binom{n_i}{j}\Big(\big(1-F(x)\big)^{j}\big(F(x)\big)^{n_i-j} - \big(1-F(x)+f(x)\big)^{j}\big(F(x)-f(x)\big)^{n_i-j}\Big)$$

[0019]

[0020] 상기 식에서,

[0021] i는 판독물 쌍이 정렬된 일배체형이고,

[0022] $F(x)=\frac{x}{H_i-L-1}$,

[0023] $f(x)=\frac{1}{H_i-L-1}$,

[0024] H_i는 일배체형 i의 길이이고,

[0025] L은 평균 단편 길이이고,

[0026] n_i는 일배체형 i에 정렬된 판독물 쌍들의 개수이다.

[0027] 일부 구현예들에서, 정렬 위치들의 세트에 대한 정렬 점수는 개별 정렬 위치들의 확률의 곱으로서 추정된다.

[0028] 일부 구현예들에서, 상기 방법들은, 정렬 위치들의 선택된 세트에서 상기 복수의 일배체형 서열들에 정렬된 복수의 서열 판독물들에 대한 하나 이상의 서열분석 메트릭들을 추정하는 단계를 추가로 포함한다. 일부 구현예들에서, 상기 하나 이상의 서열분석 메트릭들은 서열 커버리지(sequence coverage)를 포함한다. 일부 구현예들에서, 상기 하나 이상의 서열분석 메트릭들은 각각의 정렬 위치에 대한 서열 커버리지를 포함한다. 일부 구현예들에서, 상기 하나 이상의 서열분석 메트릭들은 정렬 품질 점수를 포함한다. 일부 구현예들에서, 상기 하나 이상의 서열분석 메트릭들은 각각의 정렬 위치에 대한 정렬 품질 점수를 포함한다. 일부 구현예들에서, 상기 하나 이상의 서열분석 메트릭들은 맵핑 품질 점수를 포함한다.

[0029] 일부 구현예들에서, 상기 복수의 서열 판독물들은 적어도 100개의 서열 판독물들을 포함한다.

[0030] 일부 구현예들에서, 상기 방법들은 서열 판독물들의 상이한 세트들을 사용하여 상이한 게놈 영역들에 대해 동작 (a)를 수행하는 단계를 추가로 포함한다. 일부 구현예들에서, 상기 상이한 게놈 영역들은 적어도 100개의 상이한 게놈 영역들을 포함한다.

[0031] 일부 구현예들에서, 상기 방법들은, 동작 (a) 전에, 게놈 영역에 대응하는 하나 이상의 서열 그래프들에 제1 개수의 서열 판독물들을 정렬시켜 상기 복수의 서열 판독물들 및/또는 상기 복수의 일배체형들을 획득하는 단계를 추가로 포함한다. 일부 구현예들에서, 상기 서열 그래프에 상기 제1 개수의 서열 판독물들을 정렬시키는 단계는 하기를 포함한다: (i) 핵산 샘플의 제1 개수의 서열 판독물들을 제공하는 단계; (ii) 서열 그래프에 의해 각각 표현된 하나 이상의 반복 서열들에 상기 제1 개수의 서열 판독물들을 정렬시키는 단계 - 상기 서열 그래프는

핵산 서열들을 표현하는 정점들 및 상기 정점들을 연결하는 방향성 에지(edge)들을 갖는 방향성 그래프의 데이터 구조를 갖고, 상기 서열 그래프는 하나 이상의 자가-루프(self-loop)들을 포함하며, 각각의 자가-루프는 반복 하위서열을 표현하고, 각각의 반복 하위서열은 하나 이상의 뉴클레오티드들의 반복 단위의 반복체들을 포함함 -; (iii) 상기 하나 이상의 반복 서열들의 하나 이상의 유전자형들을 결정하는 단계; 및 (iv) (a)의 복수의 서열 판독물들로서의 제1 개수의 서열 판독물들 및/또는 상기 하나 이상의 반복 서열들의 하나 이상의 유전자형들을 제공하는 단계.

[0032] 일부 구현예들에서, 상기 방법들은 상기 하나 이상의 반복 서열들의 하나 이상의 유전자형들을 단계적으로 분석하여 (b)의 복수의 일배체형들을 결정하는 단계를 추가로 포함한다. 일부 구현예들에서, 상기 방법들은, 초기에 게놈에 제2 개수의 서열 판독물들을 정렬시켜 상기 제1 개수의 서열 판독물들을 제공하는 단계 - 상기 제2 개수의 서열 판독물들은 적어도 10,000개의 서열 판독물들을 포함함 -를 추가로 포함한다.

[0033] 본 개시내용의 또 다른 양태는 컴퓨터 그래픽을 생성하기 위한 시스템들을 제공하며, 각각의 그래픽은 게놈 영역의 복수의 일배체형들에 정렬된 서열 판독물들을 표현한다.

[0034] 일부 구현예들에서, 상기 시스템은 또한 테스트 샘플의 핵산들을 서열분석하기 위한 서열분석기를 포함한다.

[0035] 일부 구현예들에서, 상기 하나 이상의 프로세서들은 본원에 기술된 다양한 방법들을 수행하도록 구성된다.

[0036] 본 개시내용의 다른 양태는 프로그램 코드를 저장하는 비일시적 기계 판독가능 매체를 포함하는 컴퓨터 프로그램 제품을 제공하며, 상기 프로그램 코드는, 컴퓨터 시스템의 하나 이상의 프로세서들에 의해 실행될 때, 컴퓨터 시스템으로 하여금 컴퓨터 그래픽을 생성하기 위한 상기 방법을 구현하게 하며, 각각의 그래픽은 게놈 영역의 복수의 일배체형들에 정렬된 서열 판독물들을 표현한다.

[0037] 일부 구현예들에서, 상기 프로그램 코드는 본원에 기술된 방법들의 동작들을 수행하기 위한 코드를 포함한다.

도면의 간단한 설명

[0038] 도 1a는 참조 서열 상의 반복 서열에 대한 서열 판독물들의 정렬에서의 어려움을 예시하는 개략도이다.

도 1b는 도 1a에 도시된 어려움을 극복하기 위한 특정한 개시된 구현예들에 따른 쌍형성된 말단 판독물들을 사용한 서열 판독물들의 정렬을 예시하는 개략도이다.

도 1c는 CAG 모티프를 갖는 탠덤 반복체의 예시를 도시한다.

도 1d는 판독물 길이보다 더 긴 탠덤 반복체를 서열분석함으로써 생성된 쌍형성된 판독물들의 예시를 도시한다.

도 2a 및 도 2b는 쌍형성된 말단 판독물들을 사용하더라도 TR 영역에 판독물들을 정렬시키는 것이 어려운 시나리오를 예시한다.

도 3a는 통상적인 판독물 파일업을 개략적으로 예시한다.

도 3b는 일부 구현예들에 따른 판독물 파일업을 개략적으로 예시한다.

도 4는 일부 구현예들에 따라 판독물 파일업들을 생성하기 위한 개략적인 작업 흐름을 도시한다.

도 5는 게놈 영역의 일배체형들에 정렬된 서열 판독물들을 표현하는 컴퓨터 그래픽을 생성하기 위한 방법(50)에 대한 흐름도를 도시한다.

도 6은 복수의 일배체형들을 포함하는 서열 판독물 파일업을 표현하는 컴퓨터 그래픽을 생성하기 위한 방법(600)에 대한 흐름도를 도시한다.

도 7은 정렬 위치들의 세트에 서열 판독물들을 정렬시키기 위한 방법(700)의 흐름도를 도시한다.

도 8은 일부 구현예들에 따른 반복 서열을 포함하는 게놈 좌위를 유전자형분석하기 위한 방법을 예시하는 흐름도를 도시한다.

도 9는 제1 게놈 좌위를 표현하는 제1 서열 그래프를 도시한다.

도 10은 제2 게놈 좌위를 표현하는 제2 서열 그래프를 도시한다.

도 11은 제3 게놈 좌위를 표현하는 제3 서열 그래프를 도시한다.

도 12는 일부 구현예들에 따른 2개의 STR 서열들을 포함하는 HTT 좌위에서의 변이체들의 유전자형들을 결정하기

위한 방법의 개략도를 도시한다.

도 13은 일부 구현예들에 따른, SNV 및 STR을 포함하는 린치 I(Lynch I) 좌위에서의 변이체들의 유전자형들을 결정하기 위한 방법의 개략도를 도시한다. 도 12의 좌측 패널은 표적화된 유전자형분석을 위한 일반적인 방법의 개략도를 도시하고; 우측 패널은 린치 I 증후군과 연관된 좌위에서의 변이체들의 유전자형분석에 이 방법을 적용하는 것을 도시한다.

도 14는 샘플 내의 반복 서열의 확장의 존재 또는 부재를 결정하기 위한 방법의 일례의 고수준 묘사를 제공하는 흐름도이다.

도 15 및 도 16은 쌍형성된 말단 판독물들을 사용하여 반복 확장을 검출하기 위한 방법의 예를 예시하는 흐름도이다.

도 17은 반복 확장을 결정하기 위해 관심있는 임의의 반복 서열과 연관되지 않은 비정렬된 판독물들을 사용하는 방법의 흐름도이다.

도 18은 테스트 샘플을 프로세싱하기 위한 분산 시스템의 블록도이다.

도 19는 일부 구현예들에 따라 구현된 *ATXN3* 반복체에 대한 판독물 파일업을 도시한다.

도 20은 일부 구현예들에 따라 구현된 *DMPK* 반복체에 대한 판독물 파일업을 도시한다.

도 21a는 일부 구현예들에 따라 구현된 *HTT* 좌위에 대한 판독물 파일업을 도시한다.

도 21b는 통상적인 방법에 의해 생성된 *HTT* 좌위에 대한 판독물 파일업을 도시한다.

도 22는 일부 구현예들에 따른 *C9ORF72* 반복체의 잘못 호출된 확장을 포함하는 판독물 파일업을 도시한다.

도 23은 일부 구현예들에 따른 *FMR1* 반복체의 잘못 호출된 확장을 포함하는 판독물 파일업을 도시한다.

발명을 실시하기 위한 구체적인 내용

[0039] 본 개시내용은 의학적으로 중요한 반복 서열들의 확장과 같은, 관심있는 반복 확장들을 식별하고 시각화하기 위한 방법들, 장치들, 시스템들, 및 컴퓨터 프로그램 제품들에 관한 것이다. 반복 확장들의 예는 취약 X 증후군, ALS, 헌팅턴 병, 프리드라이히 운동실조증, 척수소뇌성 운동실조증, 척수구근 근위축증, 근긴장성 이영양증, 마카도-조셉 병, 및 치상핵적핵 담창구시상하부 위축증과 같은 유전적 장애들과 연관된 확장들을 포함하지만, 이들로 제한되지 않는다.

[0040] 달리 지시되지 않는 한, 본원에 개시된 방법들 및 시스템들의 실시는 당해 분야의 기술 내에 있는 분자 생물학, 미생물학, 단백질 정제, 단백질 공학, 단백질 및 DNA 서열분석, 및 재조합 DNA 분야에서 일반적으로 사용되는 통상적인 기술들 및 장치들을 수반한다. 이러한 기술들 및 장치들은 당업자에게 공지되어 있으며, 다수의 문헌 및 참조 문헌에 기술되어 있다(예를 들어, 문헌[Sambrook et al., "Molecular Cloning: A Laboratory Manual," Third Edition (Cold Spring Harbor),]); 및 문헌[Ausubel et al., "Current Protocols in Molecular Biology"] 참조).

[0041] 수치 범위는 범위를 정의하는 숫자들을 포함한다. 본 명세서 전반에 걸쳐 주어진 모든 최대 수치 제한은, 모든 더 낮은 수치 제한이 마치 본원에 명시적으로 기재된 것처럼, 이러한 더 낮은 수치 제한을 포함하는 것으로 의도된다. 본 명세서 전반에 걸쳐 주어진 모든 최소 수치 제한은, 모든 더 높은 수치 제한이 마치 본원에 명시적으로 기재된 것처럼, 이러한 더 높은 수치 제한을 포함할 것이다. 본 명세서 전반에 걸쳐 주어진 모든 수치 범위는, 이러한 더 넓은 수치 범위 내에 속하는 모든 더 좁은 수치 범위가 모두 본원에 명시적으로 기재된 것처럼, 이러한 더 좁은 수치 범위들을 포함할 것이다.

[0042] 본원에 제공된 제목들은 본 개시내용을 제한하고자 하는 것은 아니다.

[0043] 본원의 예들은 인간에 관한 것이고, 그 언어는 주로 인간 관심사에 관한 것이지만, 본원에 기술된 개념들은 임의의 식물 또는 동물로부터의 게놈들에 적용 가능하다. 본 개시내용의 이들 및 다른 목적들 및 특징들은 하기의 설명 및 첨부된 청구범위로부터 더욱 완전히 명백해지거나, 또는 이하에서 기술되는 바와 같은 본 개시내용의 실시에 의해 학습될 수 있다.

[0044] 본원에서 달리 정의되지 않는 한, 본원에 사용되는 모든 기술 및 과학 용어들은 기술 분야의 당업자에 의해 통상적으로 이해되는 것과 동일한 의미를 갖는다. 본원에 포함된 용어들을 포함하는 다양한 과학 용어들은 통상

의 기술자에게 공지되어 있고 이용 가능하다. 본원에 기술된 것들과 유사한 또는 동등한 임의의 방법들 및 재료들이 본원에 개시된 실시예의 실시 또는 테스트에 사용되지만, 일부 방법들 및 재료들이 기술된다.

[0045] 바로 아래에 정의된 용어들은 전체적으로 본 명세서를 참조하여 더욱 완전히 설명된다. 본 개시내용은 기술된 특정 방법론, 프로토콜들, 및 시약들로 제한되지 않는다는 것이 이해되어야 하는데, 그 이유는 이들이 당업자에 의해 사용되는 맥락에 따라 달라질 수 있기 때문이다.

[0046] ***정의***

[0047] 본원에 사용되는 바와 같이, 단수 용어("a", "an" 및 "the")는 문맥상 달리 명확하게 지시하지 않는 한, 복수의 지시대상을 포함한다.

[0048] 달리 지시되지 않는 한, 핵산은 5'에서 3' 배향으로 좌측에서 우측으로 쓰이고 아미노산 서열은 아미노에서 카르복시 배향으로 좌측에서 우측으로 각각 쓰여진다.

[0049] 용어 "복수"는 하나 초과의 요소를 지칭한다. 예를 들어, 이 용어는 본원에 개시된 방법들을 사용하여 테스트 샘플들 및 대조군 샘플들에서 반복 확장들의 유의차를 식별하기에 충분한 개수의 핵산 분자들 또는 서열 판독물들과 관련하여 본원에서 사용된다.

[0050] 용어 "일배체형"은 본원에서 염색체 상의 연결된 유전자의들 클러스터 내의 대립유전자들의 세트를 지칭하기 위해 사용된다. 본원의 다양한 구현예들에서, 일배체형은 TR들의 대립유전자들을 포함한다.

[0051] 용어 "일배체형 서열"은 염색체 상의 대립유전자들의 세트를 포함하는 연속된 유전자 서열을 지칭한다. 예를 들어, 일배체형 서열은 2개의 플랭킹 영역(flanking region)들 및 STR 서열(예를 들어, 도 20)을 포함하거나, 또는 2개의 플랭킹 영역들, 개재 서열을 사이에 둔 2개의 인접 STR 서열들을 포함할 수 있다(예를 들어, 도 21a 및 도 21b).

[0052] 용어 "반복 서열"은 더 짧은 서열의 반복적인 발생을 포함하는 핵산 서열을 지칭한다. 상기 더 짧은 서열은 본원에서 "반복 단위" 또는 "반복 모티프" 또는 간단히 "모티프"로 지칭된다. 반복 단위의 반복적인 발생은 반복 단위의 "반복체들" 또는 "복제물들"로 지칭된다. 많은 맥락에서, 반복 서열의 위치는 단백질을 인코딩하는 유전자와 연관된다. 다른 상황에서, 반복 서열은 비-코딩 영역 내에 있을 수 있다. 반복 단위들은 반복 단위들 사이에 절단부(break)들을 갖거나 갖지 않으면서 반복 서열에서 발생할 수 있다. 예를 들어, 정상 샘플들에서, *FMR1* 유전자는 CGG 반복체들 내에 AGG 절단부를 포함하는 경향이 있다(예를 들어, (CGG)10 + (AGG) + (CGG)9). 절단부가 없는 샘플들뿐만 아니라, 절단부가 거의 없는 긴 반복 서열들은 연관된 유전자의 확장을 반복하기 쉬우며, 이는, 반복체들이 특정 개수를 초과하여 확장됨에 따라 유전병을 유발할 수 있다. 본 개시내용의 다양한 실시형태들에서, 반복체들의 개수는 절단부에 상관없이 프레임-내 반복체(in-frame repeat)들로서 카운팅된다. 프레임-내 반복체들을 추정하기 위한 방법들은 이하에서 추가로 기술된다.

[0053] 다양한 실시형태들에서, 반복 단위 각각은 1 내지 100개의 뉴클레오티드들을 포함한다. 널리 연구된 다수의 반복 단위들은 트리뉴클레오티드 또는 헥사뉴클레오티드 단위들이다. 잘 연구되어 왔으며 본원에 개시된 실시형태들에 적용 가능한 일부 다른 반복 단위들은 4, 5, 6, 8, 12, 33, 또는 42개 뉴클레오티드들의 단위를 포함하지만, 이들로 제한되지 않는다. 예를 들어, 문헌[Richards (2001) *Human Molecular Genetics, Vol. 10, No. 20*, 2187-2194]을 참조한다. 본 개시내용의 응용 분야는, 반복 단위들의 복수의 반복체들 또는 복제물들을 갖는 반복 서열과 비교하여 비교적 짧은 한, 전술된 특정 개수의 뉴클레오티드 염기들로 제한되지 않는다. 예를 들어, 반복 단위는 적어도 3, 6, 8, 10, 15, 20, 30, 40, 50개의 뉴클레오티드들을 포함할 수 있다. 대안적으로 또는 추가적으로, 반복 단위는 최대 약 100, 90, 80, 70, 60, 50, 40, 30, 20, 10, 6 또는 3개의 뉴클레오티드들을 포함할 수 있다.

[0054] 반복 서열은 진화, 발달, 및 돌연변이 유발 조건에서 확장되어, 동일한 반복 단위의 더 많은 복제물들을 생성할 수 있다. 이것은 해당 분야에서 "반복 확장"으로 지칭된다. 이러한 과정은 또한, 반복 단위의 확장의 불안정한 성질로 인해, "동적 돌연변이"로 지칭된다. 일부 반복 확장들은 유전적 장애 및 병리학적 증상과 연관된 것으로 나타났다. 다른 반복 확장들은 잘 이해되거나 연구되어 있지 않다. 본원에 개시된 방법들은 이전에 알려진 반복 확장들 및 새로운 반복 확장들 둘 모두를 식별하는 데 사용될 수 있다. 일부 실시형태들에서, 반복 확장을 갖는 반복 서열은 약 100, 150, 300 또는 500개의 염기쌍(bp)보다 더 길다. 일부 실시형태들에서, 반복 확장을 갖는 반복 서열은 약 1000 bp, 2000 bp, 3000 bp, 4000 bp, 5000 bp, 또는 10000 bp 등보다 더 길다.

[0055] 그래프 이론에서, 정점 및 에지는 그래프를 구성하는 2개의 기본 단위들이다. 정점 또는 노드(node)는, 그래프

를 정의하고 에지들에 의해 연결될 수 있는 지점들 중 하나이다. 그래프의 다이어그램에서, 정점은 라벨을 갖는 형상으로 표현될 수 있고, 에지는 하나의 정점으로부터 다른 정점까지 연장되는 선(무방향성 에지) 또는 화살표(방향성 에지)에 의해 표현된다.

[0056] 에지에 의해 연결된 2개의 정점은 에지의 종점으로 지칭된다. 그래프가 에지 (x, y)를 포함하는 경우 정점 x는 다른 정점 y에 인접한 것으로 지칭된다.

[0057] 무방향성 그래프는 정점들의 세트 및 (정점의 비순서화된 쌍을 연결하는) 무방향성 에지들의 세트로 구성되는 반면, 방향성 그래프는 정점들의 세트 및 (정점의 순서화된 쌍을 연결하는) 방향성 에지들의 세트로 구성된다.

[0058] 그래프 이론에서, 각각의 에지는 그것이 부착되는 2개의(또는 하이퍼그래프들에서는, 더 많은) 정점들을 가지며, 이는 이의 종점으로 불린다. 에지들은 방향성이거나 무방향성일 수 있고; 무방향성 에지들은 선으로도 불리며, 방향성 에지들은 호(arc) 또는 화살표로도 불린다.

[0059] 방향성 에지는 상류측 정점 및 하류측 정점을 연결하는 에지이며, 여기서 상류측 정점은 방향성 에지 앞에 나타나고 하류측 정점은 방향성 에지 뒤에 나타난다.

[0060] 무방향성 에지는 2개의 정점들을 연결하는 에지이며, 여기서 어느 하나의 정점이 그래프 경로에서 다른 정점 앞에 나타날 수 있다.

[0061] 루프, 자가-루프, 및 단일-노드 루프는 본원에서 상호교환 가능하게 사용된다. 루프는 하나의 노드, 및 2개의 말단들 모두가 상기 하나의 노드에 연결된 에지를 갖는다.

[0062] 사이클은 2개 이상의 정점들을 포함하는 경로이며, 여기서 상기 사이클의 경로는 동일한 정점에서 시작하고 종료한다. 단순 사이클은 시작 및 종료 정점 이외에 반복된 정점들 또는 에지들을 갖지 않는 사이클이다.

[0063] 순환 그래프는 적어도 하나의 사이클을 포함하는 그래프이다.

[0064] 비순환 그래프는 임의의 사이클 또는 자가-루프를 포함하지 않는 그래프이다.

[0065] 방향성 비순환 그래프(DAG: directed acyclic graph)는 임의의 사이클 또는 자가-루프가 없는 방향성 그래프이다.

[0066] 그래프 경로는 정점들 및 에지들의 서열이며, 여기서 에지의 양쪽 종점은 서열에서 에지에 인접하게 나타난다. 방향성 그래프의 그래프 경로는 방향성 에지(또는 호 또는 화살표) 앞에 나타나는 상류측 정점 및 방향성 에지 뒤에 나타나는 하류측 정점을 갖는다.

[0067] 푸아송 분포(Poisson distribution)는, 특정 개수의 이벤트들이 마지막 이벤트 이후의 시간과 독립적으로 그리고 알려진 일정한 레이트로 발생하는 경우, 이들 이벤트들이 고정된 시간 간격 또는 공간에서 발생하는 확률을 나타내는 이산 확률 분포이다.

[0068] 완전히 특정된 염기 기호들은 각각 구아닌, 아데닌, 티민, 및 시토신에 대한 G, A, T, C를 포함한다.

[0069] 불완전하게 특정된 핵산 명명법은 특히 하기와 같은 것을 포함한다.

[0070] 퓨린(아데닌 또는 구아닌): R

[0071] 피리미딘(티민 또는 시토신): Y

[0072] 아데닌 또는 티민: W

[0073] 구아닌 또는 시토신: S

[0074] 아데닌 또는 시토신: M

[0075] 구아닌 또는 티민: K

[0076] 아데닌 또는 티민 또는 시토신: H

[0077] 구아닌 또는 시토신 또는 티민: B

[0078] 구아닌 또는 아데닌 또는 시토신: V

[0079] 구아닌 또는 아데닌 또는 티민: D

[0080] 구아닌 또는 아데닌 또는 티민 또는 시토신: N

[0081] 용어 "쌍형성된 말단 판독물"은 핵 단편의 각각의 말단으로부터 하나의 판독물을 획득하는 쌍형성된 말단 서열분석으로부터 획득된 판독물을 지칭한다. 쌍형성된 말단 서열분석은 DNA를 삽입물로 불리는 서열들로 단편화하는 것을 수반한다. Illumina에 의해 사용되는 일부 프로토콜들과 같은 일부 프로토콜들에서, 더 짧은 삽입물로부터의(예를 들어, 대략 수십 내지 수백 bp) 판독물은 짧은-삽입물 쌍형성된 말단 판독물 또는 간단히 쌍형성된 말단 판독물로 지칭된다. 대조적으로, 더 긴 삽입물로부터의(예를 들어, 대략 수천 bp) 판독물은 메이트 쌍 판독물(mate pair read)로 지칭된다. 본 개시내용에서, 짧은-삽입물 쌍형성된 말단 판독물 및 긴-삽입물 메이트 쌍 판독물 둘 모두가 사용될 수 있고, 반복 확장을 분석하기 위한 방법과 관련하여 구별되지 않는다. 따라서, 용어 "쌍형성된 말단 판독물"은 짧은-삽입물 쌍형성된 말단 판독물 및 긴-삽입물 메이트 쌍 판독물 둘 모두를 지칭할 수 있으며, 이는 본원에서 이후에 추가로 기술된다. 일부 실시형태들에서, 쌍형성된 말단 판독물들은 약 20 bp 내지 1000 bp의 판독물들을 포함한다. 일부 실시형태들에서, 쌍형성된 말단 판독물들은 약 50 bp 내지 500 bp, 약 80 bp 내지 150 bp, 또는 약 100 bp의 판독물들을 포함한다. 쌍형성된 말단에서의 2개의 판독물들은 서열분석되는 단편의 최극단에 위치될 필요가 없다는 것을 이해할 것이다. 오히려, 하나 또는 둘 모두의 판독물이 단편의 말단에 근접하게 있을 수 있다. 또한, 쌍형성된 말단 판독물의 맥락에서의 본원에 예시된 방법들은, 판독물들이 단편의 말단 또는 단편의 다른 부분으로부터 유도되는지 여부에 관계없이 다양한 쌍형성된 판독물들 중 임의의 것으로 수행될 수 있다.

[0082] 본원에 사용되는 바와 같이, 용어 "정렬" 및 "정렬시키는"은 판독물을 참조 서열과 비교하고 이에 의해 참조 서열이 판독된 서열을 포함하는지 여부를 결정하는 과정을 지칭한다. 정렬 과정은 판독물이 참조 서열에 맵핑될 수 있는지 여부를 결정하려고 시도하지만, 항상 참조 서열에 정렬된 판독물을 생성하는 것은 아니다. 참조 서열이 판독물을 포함하는 경우, 판독물은 참조 서열에 맵핑될 수 있거나, 또는 특정 실시형태에서, 참조 서열 내의 특정 위치에 맵핑될 수 있다. 일부 경우에 있어서, 정렬은 판독물이 특정 참조 서열의 구성원인지의 여부(즉, 판독물이 참조 서열에 존재하는지 아니면 존재하지 않는지)를 간단히 알려준다. 예를 들어, 인간 염색체 13에 대한 참조 서열에 대한 판독물의 정렬은 판독물이 염색체 13에 대한 참조 서열에 존재하는지 여부를 알려줄 것이다. 이러한 정보를 제공하는 도구는 세트 멤버십 테스터(set membership tester)로 불릴 수 있다. 일부 경우에 있어서, 정렬은, 추가로, 판독물이 맵핑되는 참조 서열 내의 위치를 나타낸다. 예를 들어, 참조 서열이 전체 인간 게놈 서열인 경우, 정렬은 판독물이 염색체 13 상에 존재함을 나타낼 수 있고, 판독물이 염색체 13의 특정 가닥 및/또는 부위 상에 있음을 추가로 나타낼 수 있다.

[0083] 정렬된 판독물들은 참조 게놈과 같은 알려진 참조 서열에 대해 이들의 핵산 분자들의 순서와 관련하여 매치인 것으로서 식별되는 하나 이상의 서열들이다. 정렬된 판독물 및 참조 서열 상에서의 이의 결정된 위치는 서열 태그를 구성한다. 정렬은 전형적으로 컴퓨터 알고리즘에 의해 구현되지만, 본원에 개시된 방법들을 구현하기 위해 합리적인 기간에 판독물들을 정렬시키는 것이 불가능한 경우, 정렬은 수동으로 행해질 수 있다. 서열들을 정렬시키는 것으로부터의 알고리즘의 일례는 Illumina Genomics Analysis 파이프라인의 일부로서 배포된 Efficient Local Alignment of Nucleotide Data(ELAND) 컴퓨터 프로그램이다. 대안적으로, 블룸 필터(Bloom filter) 또는 유사한 세트 멤버십 테스터가, 판독물들을 참조 게놈들에 정렬시키기 위해 사용될 수 있다. 2014년 4월 25일자로 출원된 미국 특허 출원 제14/354,528호를 참조하며, 이는 그 전체가 본원에 참고로 포함된다. 정렬 시 서열 판독물의 매칭은 100% 서열 매치 또는 100% 미만(즉, 완벽하지 않은 매치)일 수 있다.

[0084] 본원에 사용되는 용어 "맵핑"은 정렬에 의해, 더 큰 서열, 예를 들어, 참조 게놈에 판독된 서열을 할당하는 것을 지칭한다.

[0085] 일부 경우에 있어서, 2개의 쌍형성된 말단 판독물들 중 하나의 말단 판독물은 참조 서열의 반복 서열에 정렬되는 반면, 2개의 쌍형성된 말단 판독물들 중 다른 말단 판독물은 정렬되지 않는다. 이러한 경우에서, 참조 서열의 반복 서열에 정렬되는 쌍형성된 판독물은 "앵커 판독물(anchor read)"로 지칭된다. 반복 서열에 정렬되지 않지만 앵커 판독물과 쌍을 이루는, 쌍형성된 말단 판독물은 앵커링된 판독물(anchored read)로 지칭된다. 이와 같이, 비정렬된 판독물은 반복 서열에 앵커링되고 이와 연관될 수 있다. 일부 실시형태들에서, 상기 비정렬된 판독물들은 참조 서열에 정렬될 수 없는 판독물들 및 참조 서열에 불량하게 정렬되는 판독물들 둘 모두를 포함한다. 판독물이 특정 기준보다 많은 개수의 미스매칭된 염기들을 가지면서 참조 서열에 정렬되는 경우, 판독물은 불량하게 정렬된 것으로 간주된다. 예를 들어, 다양한 실시형태들에서, 판독물은 그것이 적어도 약 1, 2, 3, 4, 5, 6, 7, 8, 9 또는 10개의 미스매치들과 정렬되는 경우 불량하게 정렬된 것으로 간주된다. 일부 경우에 있어서, 한 쌍의 판독물 중 둘 모두는 참조 서열에 정렬된다. 이러한 경우에 있어서, 판독물 둘 모두는 다양한

구현예들에서 "앵커 판독물"로서 분석될 수 있다.

[0086] 용어 "폴리뉴클레오티드", "핵산" 및 "핵산 분자"는 상호교환 가능하게 사용되고 공유 결합된 뉴클레오티드 서열(즉, RNA에 대한 리보뉴클레오티드 및 DNA에 대한 데옥시리보뉴클레오티드)을 지칭하며, 여기서 하나의 뉴클레오티드의 펜토스의 3' 위치는 포스포다이에스테르기에 의해 다음의 펜토스의 5' 위치에 결합된다. 뉴클레오티드들은 RNA 및 DNA 분자들, 예를 들어 무세포 DNA(cfDNA) 분자들을 포함하지만 이들로 제한되지 않는, 임의의 형태의 핵산의 서열을 포함한다. 용어 "폴리뉴클레오티드"는 단일- 및 이중-가닥 폴리뉴클레오티드들을 포함하지만, 이들로 제한되지 않는다.

[0087] 본원에서 용어 "테스트 샘플"은, 전형적으로, 복제수 변이에 대해 스크리닝될 적어도 하나의 핵산 서열을 갖는 핵산 또는 핵산들의 혼합물을 포함하는 생물학적 유체, 세포, 조직, 장기, 또는 유기체로부터 유도된 샘플을 지칭한다. 특정 실시형태에서, 샘플은, 복제수가 변이를 겪은 것으로 의심되는 적어도 하나의 핵산 서열을 갖는다. 이러한 샘플들은 가래/구강액, 양수, 혈액, 혈액 분획물, 또는 미세 바늘 생검 샘플들, 소변, 복막액, 흉수 등을 포함하지만, 이들로 제한되지 않는다. 샘플은 종종 인간 대상체(예를 들어, 환자)로부터 취해지지만, 검정은 개, 고양이, 말, 염소, 양, 소, 돼지 등을 포함하지만 이들로 제한되지 않는, 임의의 포유류로부터의 샘플들에서의 복제수 변이(CNV)에 사용될 수 있다. 샘플은 생물학적 공급원에서 획득된 대로 직접 사용하거나, 샘플의 특성을 수정하기 위한 전처리 후에 사용할 수 있다. 예를 들어, 이러한 전처리는 혈액으로부터 혈장을 제조하는 것, 점성 유체를 희석시키는 것 등을 포함할 수 있다. 전처리 방법은 또한 여과, 침전, 희석, 증류, 혼합, 원심분리, 동결, 동결건조, 농축, 증폭, 핵산 단편화, 간섭 구성요소의 불활성화, 시약의 첨가, 용해 등을 수반할 수 있지만, 이들로 제한되지 않는다. 이러한 전처리 방법들이 샘플에 대해 사용되는 경우, 이러한 전처리 방법들은 전형적으로, 관심 있는 핵산(들)이, 때때로 미처리된 테스트 샘플(예를 들어, 즉, 임의의 이러한 전처리 방법(들)을 거치지 않은 샘플)의 농도에 비례하는 농도로, 테스트 샘플에 잔류하도록 하는 것이다. 이러한 "처리된" 또는 "프로세싱된" 샘플은 여전히 본원에 기술된 방법들에 대한 생물학적 "테스트" 샘플로 간주된다.

[0088] 대조군 샘플은 음성 또는 양성 대조군 샘플일 수 있다. "음성 대조군 샘플" 또는 "영향받지 않은 샘플"은 병원성이 아닌 범위 내에서 다수의 반복체들을 갖는 반복 서열을 가지는 것으로 알려지거나 예상되는 핵산들을 포함하는 샘플을 지칭한다. "양성 대조군 샘플" 또는 "영향받은 샘플"은 병원성인 범위 내에서 다수의 반복체들을 갖는 반복 서열을 가지는 것으로 알려지거나 예상된다. 음성 대조군 샘플 내의 반복 서열의 반복체들은 전형적으로 정상 범위를 넘어서 확장되지 않은 반면, 양성 대조군 샘플 내의 반복 서열의 반복체들은 전형적으로 정상 범위를 넘어서 확장되었다. 이와 같이, 테스트 샘플 내의 핵산들은 하나 이상의 대조군 샘플들과 비교될 수 있다.

[0089] 본원에서 용어 "관심있는 서열"은 건강한 개체 대 질병에 걸린 개체에서의 서열 표현의 차이와 연관된 핵산 서열을 지칭한다. 관심있는 서열은 질병 또는 유전적 병태에서 확장되는 염색체 상의 반복 서열일 수 있다. 관심있는 서열은 염색체, 유전자, 코딩 또는 비-코딩 서열의 일부일 수 있다.

[0090] 본원에서 용어 "차세대 서열분석(NGS: next generation sequencing)"은 클론 증폭된 분자의 및 단일 핵산 분자의 대규모 병렬 서열분석을 허용하는 서열분석 방법을 지칭한다. NGS의 비제한적인 예는, 가역적 염료 종결자를 사용하는 합성에 의한 서열분석(sequencing-by-synthesis), 및 결찰에 의한 서열분석(sequencing-by-ligation)을 포함한다.

[0091] 본원에서 용어 "파라미터"는 물리적 특성을 특성화하는 수치 값을 지칭한다. 종종, 파라미터는 정량적 데이터 세트 및/또는 정량적 데이터 세트들 사이의 수치 관계를 수치적으로 특성화한다. 예를 들어, 염색체로 맵핑된 서열 태그들의 개수와 상기 태그들이 맵핑되는 염색체의 길이 사이의 비율(또는 비율의 함수)이 파라미터이다.

[0092] 본원에서 용어 "호출 기준"은 의학적 병태를 갖는 것으로 의심되는 유기체로부터의 핵산을 포함하는 테스트 샘플과 같은 샘플을 특성화하기 위한 컷오프로서 사용되는 임의의 숫자 또는 양을 지칭한다. 임계값은 파라미터 값과 비교되어서, 이러한 파라미터 값을 생성하는 샘플이 해당 유기체가 의학적 병태를 가지고 있음을 시사하는지 여부를 결정할 수 있다. 특정 실시형태에서, 임계값은 대조군 데이터 세트를 사용하여 계산되고, 유기체에서 반복 확장의 진단의 한계로서의 역할을 한다. 일부 구현예들에서, 임계값이 본원에 개시된 방법들로부터 획득된 결과에 의해 초과되는 경우, 대상체는 반복 확장으로 진단될 수 있다. 본원에 기술된 방법들에 대한 적절한 임계값은 대조군 샘플들 또는 샘플들의 트레이닝 세트에 대해 계산된 값을 분석함으로써 식별될 수 있다. 서열분석 깊이, 판독물 길이, 반복 서열 길이 등과 같은 경험적 파라미터들로부터 임계값을 계산할 수도 있다. 대안적으로, 반복 확장을 갖는 것으로 알려진 영향받은 샘플들을 사용하여, 선택된 임계값이 테스트 세트 내의

영향받은 샘플들과 영향받지 않은 샘플들을 구별하는 데 유용하다는 것을 확인할 수도 있다. 임계값의 선택은 사용자가 분류를 행할 때 원하는 신뢰도의 수준에 따라 결정된다. 일부 실시형태들에서, 적절한 임계값을 식별하는 데 사용되는 트레이닝 세트는 적어도 10, 적어도 20, 적어도 30, 적어도 40, 적어도 50, 적어도 60, 적어도 70, 적어도 80, 적어도 90, 적어도 100, 적어도 200, 적어도 300, 적어도 400, 적어도 500, 적어도 600, 적어도 700, 적어도 800, 적어도 900, 적어도 1000, 적어도 2000, 적어도 3000, 적어도 4000개, 또는 그 초과의 적격 샘플들을 포함한다. 임계값의 진단 유용성을 개선하기 위해 보다 큰 세트의 적격 샘플들을 사용하는 것이 유리할 수 있다.

[0093] 용어 "판독물"은 핵산 샘플의 일부로부터의 서열 판독물을 지칭한다. 필수적인 것은 아니지만, 전형적으로, 판독물은 샘플 내의 연속된 염기쌍들의 짧은 서열을 표현한다. 판독물은 샘플 부분의 (ATCG에서의) 염기쌍 서열에 의해 상징적으로 표현될 수 있다. 이것은 메모리 디바이스에 저장될 수 있으며, 이것이 참조 서열과 매칭하는지 또는 다른 기준을 충족시키는지 여부를 결정하기 위해 경우에 따라 프로세싱될 수 있다. 판독물은 서열분석 장치로부터 직접 획득되거나 샘플에 관하여 저장된 서열 정보로부터 간접적으로 획득될 수 있다. 일부 경우에 있어서, 판독물은, 예를 들어 염색체 또는 게놈 영역 또는 유전자에 정렬되고 맵핑될 수 있는 더 큰 서열 또는 영역을 식별하는 데 사용될 수 있는 충분한 길이(예를 들어, 적어도 약 25 bp)의 DNA 서열이다.

[0094] 용어 "게놈 판독물"은 개체의 전체 게놈에서의 임의의 세그먼트의 판독물과 관련하여 사용된다.

[0095] 용어 "부위"는 참조 게놈 상의 고유 위치(즉, 염색체 ID, 염색체 위치 및 배향)를 지칭한다. 일부 실시형태들에서, 부위는 서열 상의 잔기, 서열 태그, 또는 세그먼트의 위치일 수 있다.

[0096] 본원에 사용되는 바와 같이, 용어 "참조 게놈" 또는 "참조 서열"은 대상체로부터 식별된 서열을 참조하기 위해 사용될 수 있는 임의의 유기체 또는 바이러스의, 부분적이든 완전하든, 임의의 특정한 알려진 게놈 서열을 지칭한다. 예를 들어, 인간 대상체에 사용되는 참조 게놈뿐만 아니라 많은 다른 유기체들이 ncbi.nlm.nih.gov에서의 국립 생물공학 정보 센터(National Center for Biotechnology Information)에서 찾아질 수 있다. "게놈"은 핵산 서열에서 발현되는, 유기체 또는 바이러스의 완전한 유전자 정보를 지칭한다.

[0097] 다양한 실시형태들에서, 참조 서열은 이에 정렬된 판독물들보다 상당히 크다. 예를 들어, 이것은 적어도 약 100배 더 크거나, 또는 적어도 약 1000배 더 크거나, 또는 적어도 약 10,000배 더 크거나, 또는 적어도 약 10^5배 더 크거나, 또는 적어도 약 10^6배 더 크거나, 또는 적어도 약 10^7배 더 클 수 있다.

[0098] 일례에서, 참조 서열은 전장 인간 게놈의 것이다. 이러한 서열은 게놈 참조 서열로 지칭될 수 있다. 다른 예에서, 참조 서열은 염색체 13과 같은 특정 인간 염색체로 제한된다. 일부 실시형태들에서, 참조 Y 염색체는 인간 게놈 버전 hg19로부터의 Y 염색체 서열이다. 이러한 서열은 염색체 참조 서열로 지칭될 수 있다. 참조 서열의 다른 예는 임의의 종의 염색체, 하위염색체 영역(예를 들어, 가닥) 등뿐만 아니라 다른 종의 게놈들을 포함한다.

[0099] 일부 실시형태들에서, 정렬을 위한 참조 서열은 판독물의 길이의 약 1 내지 약 100배의 서열 길이를 가질 수 있다. 이러한 실시형태에서, 상기 정렬 및 서열분석은 전체 게놈 정렬 또는 서열분석 대신에, 표적화된 정렬 또는 서열분석으로 간주된다. 이들 실시형태에서, 참조 서열은 전형적으로, 관심있는 유전자 및/또는 반복 서열을 포함한다.

[0100] 다양한 실시형태들에서, 참조 서열은 복수의 개체들로부터 유도된 공통 서열 또는 다른 조합이다. 그러나, 특정 응용 분야에서, 참조 서열은 특정 개체로부터 취해질 수 있다.

[0101] 본원에서 용어 "임상적으로 관련된 서열"은 유전적 또는 질병 병태와 연관되거나 연루된 것으로 알려지거나 의심되는 핵산 서열을 지칭한다. 임상적으로 관련된 서열의 부재 또는 존재를 결정하는 것은 진단을 결정하거나 의학적 상태의 진단을 확인하거나, 또는 질병의 발병에 대한 예후를 제공하는 데 유용할 수 있다.

[0102] 핵산 또는 핵산들의 혼합물의 맥락에서 사용될 때, 용어 "유도된"은 본원에서, 핵산(들)이 이들이 유래하는 공급원으로부터 획득되는 수단을 지칭한다. 예를 들어, 일 실시형태에서, 2개의 상이한 게놈들로부터 유도된 핵산들의 혼합물은, 상기 핵산, 예를 들어, cfDNA가 괴사 또는 사멸과 같은 자연 발생 과정을 통해 세포들에 의해 자연적으로 방출되었음을 의미한다. 다른 실시형태에서, 2개의 상이한 게놈들로부터 유도된 핵산들의 혼합물은, 상기 핵산들이 대상체로부터의 2개의 상이한 유형의 세포들로부터 추출되었음을 의미한다.

[0103] 특정 정량적 값을 획득하는 맥락에서 사용될 때, 용어 "~에 기초한"은 본원에서, 다른 양을 입력으로 사용하여

특정 정량적 값을 출력으로서 계산하는 것을 지칭한다.

[0104] 용어 "환자 샘플"은 본원에서 환자, 즉, 의학적 관심, 케어 또는 치료의 수용자로부터 획득된 생물학적 샘플을 지칭한다. 상기 환자 샘플은 본원에 기술된 샘플들 중 임의의 것일 수 있다. 특정 실시형태에서, 상기 환자 샘플은 비-침습 시술에 의해 획득되는 것, 예를 들어, 말초 혈액 샘플 또는 분변 샘플이다. 본원에 기술된 방법들은 인간에 제한될 필요는 없다. 따라서, 다양한 수의학적 응용 분야가 고려되고, 이 경우에서, 상기 환자 샘플은 비-인간 포유류(예를 들어, 고양이, 돼지, 말, 소 등)로부터의 샘플일 수 있다.

[0105] 용어 "생물학적 유체"는 본원에서 생물학적 공급원으로부터 취해진 액체를 지칭하며 예를 들어, 혈액, 혈청, 혈장, 가래, 세척액, 뇌척수액, 소변, 정액, 땀, 눈물, 타액 등을 포함한다. 본원에 사용되는 바와 같이, 용어 "혈액", "혈장" 및 "혈청"은 이들의 분획물들 또는 프로세싱된 부분들을 명확히 포함한다. 유사하게, 샘플이 생검, 면봉, 스미어(smear) 등으로부터 취해지는 경우, 상기 "샘플"은 생검, 면봉, 스미어 등으로부터 유도된 프로세싱된 분획물들 또는 부분들을 명확히 포함한다.

[0106] 본원에 사용되는 바와 같이, 용어 "~에 대응하는"은 때때로, 상이한 대상체들의 게놈에 존재하고, 모든 게놈들에서 반드시 동일한 서열을 갖는 것은 아니지만, 관심있는 서열, 예를 들어, 유전자 또는 염색체의 유전자 정보보다는 아이덴티티를 제공하는 역할을 하는 핵산 서열, 예를 들어, 유전자 또는 염색체를 지칭한다.

[0107] 본원에 사용되는 바와 같이, 용어 "염색체"는 DNA 및 단백질 구성요소들(특히, 히스톤)을 포함하는 염색질 가닥으로부터 유도된, 살아있는 세포의 유전-보유 유전자 운반체(heredity-bearing gene carrier)를 지칭한다. 통상적인 국제적으로 인식되는 개별 인간 게놈 염색체 넘버링 시스템이 본 발명에 사용된다.

[0108] 본원에 사용되는 바와 같이, 용어 "폴리뉴클레오티드 길이"는 참조 게놈의 서열 또는 영역 내의 핵산 단량체 하위단위(뉴클레오티드)의 절대 개수를 지칭한다. 용어 "염색체 길이"는, 예를 들어, 월드와이드웹 상에서 |genome|.|ucsc|.|edu/cgi- bin/hgTracks?hgsid=167155613&chromInfoPage=에서 발견되는 인간 염색체의 NCBI36/hg18 어셈블리에서 제공되는, 염기쌍들로 주어진, 염색체의 알려진 길이를 지칭한다.

[0109] 용어 "대상체"는 본원에서 인간 대상체뿐만 아니라 포유류, 무척추동물, 척추동물, 진균, 효모, 박테리아, 및 바이러스와 같은 비-인간 대상체를 지칭한다. 본원의 예는 인간에 관한 것이고 그 언어는 주로 인간 관심사에 관한 것이지만, 본원에 개시된 개념들은 임의의 식물 또는 동물로부터의 게놈들에 적용 가능하고, 수의학, 동물과학, 연구 실험실 등의 분야에서 유용하다.

[0110] 본원에 사용되는 바와 같이, 용어 "프라이머"는 연장 생성물의 합성을 유도하는 조건(예를 들어, 상기 조건은 뉴클레오티드들, DNA 중합효소와 같은 유도제, 및 적합한 온도 및 pH를 포함함) 하에 배치되는 경우 합성의 개시 지점으로서 작용할 수 있는 단리된 올리고뉴클레오티드를 지칭한다. 상기 프라이머는 바람직하게는, 증폭에서의 최대 효율을 위해 단일 가닥일 수 있지만, 대안적으로 이중 가닥일 수 있다. 이중 가닥인 경우, 연장 생성물을 제조하는 데 사용되기 전에, 먼저 이의 가닥을 분리시키도록 처리된다. 상기 프라이머는 올리고데옥시리보뉴클레오티드일 수 있다. 상기 프라이머는 상기 유도제의 존재 하에 연장 생성물의 합성을 프라이밍하기에 충분히 길다. 상기 프라이머의 정확한 길이는 온도, 프라이머의 공급원, 상기 방법의 용도, 및 프라이머 설계에 사용되는 파라미터들을 포함하는 많은 인자들에 따라 결정될 것이다.

[0111] ***서론***

[0112] 탠덤 반복체(TR)들로 알려진 비교적 짧은 DNA 조각들의 반복체들로 구성된 서열들은 게놈 전체에서 발생한다(예를 들어 도 1c). TR 돌연변이 비율은 다른 게놈 영역들보다 10배에서 1000배 더 높을 수 있어서, TR들은 인간의 유전적 변이에 큰 기여를 한다. TR들은, 세대 사이에 반복체들의 개수가 증가하거나 감소하는 "미끄러짐(slippage)"을 통해 크게 돌연변이된다. 축적된 증거에 따르면, TR들은 기본 세포 과정에서 역할을 하며 탠덤 반복체들의 큰 확장들은 근위축성 측삭경화증(ALS), 취약 X 증후군 및 다양한 형태의 운동 실조증을 포함하는 다양한 신경학적 장애들과 관련이 있다.

[0113] TR을 포함하는 영역을 서열분석하면, 반복 서열과 부분적으로 또는 완전히 중첩되는 판독물들의 집합이 생성된다(도 1d). 이러한 판독물들의 정렬들을 함께 연결함으로써 각각의 일배체형 상의 반복체의 길이를 결정할 수 있다. 본 발명자들은 표적화된 및 게놈-전체 TR 분석을 위한 여러 가지 방법들을 개발하였다. 이하에서, 출원인은 판독물 길이보다 더 짧거나 더 긴 반복체들의 크기를 추정할 수 있는 하나 또는 복수의 인접한 TR들을 포함하는 영역들에 대한 표적화된 분석을 위한 방법인 ExpansionHunter를 설명한다. 또한, ExpansionHunter에 의해 생성된 판독물들의 정렬을 시각화하고 표시하기 위한 방법들, 시스템들 및 컴퓨터 프로그램 제품들에 대해서

도 설명한다.

[0114] TR 유전자형분석은 매우 어려운 문제이며, 최고의 방법들조차도 때때로 잘못된 유전자형 호출을 할 수 있다. 이 때문에, 해당 반복체를 유전자형분석하는 데 사용되는 판독물들의 정렬들을 검사하기 위한 강력한 시각화 방법을 갖는 것이 중요하다. 또한, 이러한 시각화 방법은 임상적으로 중요한 효과를 생성할 수 있는 반복 모티프의 변화들(예를 들어 중단들)을 검출하는 것을 가능하게 한다. 표준 데이터 시각화 파이프라인은 일반적으로, 참조 게놈에 대한 판독물들의 정렬들을 표시하는 것으로 제한되므로, 참조물에 비해 확장된 반복체들 또는 상이한 길이의 대립유전자들을 갖는 반복체들에는 적합하지 않다. 이러한 문제를 해결하기 위해, 본 발명의 발명자들은 ExpansionHunter에 의해 출력된 그래프 재정렬된 판독물들을 시각화하기 위한 도구인 Repeat Expansion Viewer(REViewer)를 개발하였다. REViewer는 인접한 반복체들을 단계적으로 분석함으로써 일배체형 서열을 결정한 다음, 이들 일배체형에 판독물 정렬들을 분포시킨다. 생성된 정적 이미지들은 특정 유전자형 호출의 정확성을 시각적으로 평가하고, 반복 서열이 임의의 중단을 포함하고 있는지 여부를 식별하는 것을 가능하게 한다.

[0115] 반복 확장들은 생물학적 및 의학적 중요성을 갖지만, 하기에 설명된 바와 같이 짧은 판독물들을 사용하여 STR들을 유전자형분석하고 반복 확장들을 검출하는 것은 어렵다. 반복 확장들을 검출하고 유전자형분석하는 기술을 개발할 필요성이 클 뿐만 아니라, 서열 판독 데이터 및 상기 서열 판독 데이터로부터 결정된 유전자형들을 시각화하기 위한 컴퓨터-구현 도구들에 대한 필요성 또한 크다. 이러한 도구들은 유전자형 호출들을 검증하고 STR들과 관련된 임상적으로 및 생물학적으로 중요한 유전적 특징들을 이해하는 데 도움이 될 수 있다.

[0116] STR 확장들은 수많은 중증의 신경학적 장애들의 주요 원인이다. 표 1은 정상 샘플들 내의 반복 서열들과 상이한, 소수의 병원성 반복 확장들을 예시한다. 열들은 반복 서열들과 연관된 유전자들, 반복 단위들의 핵산 서열들, 정상 및 병원성 서열들에 대한 반복 단위들의 반복체들의 예시적인 개수(상이한 응용 분야들에서 반복체들의 상이한 컷오프들이 사용될 수 있음), 및 반복 확장들과 연관된 질병들을 보여준다.

[0117] [표 1]

병원성 반복 확장들의 예

유전자	반복	정상	병원성	질병
FMR1	CGG	6-60	200-900	취약 X
AR	CAG	9-36	38-62	척수구근 근위축증
GHTT	CAG	11-34	40-121	헌팅턴 병
FXN	GAA	6-32	200-1700	프리드라이히 운동실조증
ATXN1	CAG	6-39	40-82	척수소뇌성 운동실조증
ATXN10	ATTCT	10-20	500-4500	척수소뇌성 운동실조증
ATXN2	CAG	15-24	32-200	척수소뇌성 운동실조증
ATXN3	CAG	13-36	61-84	척수소뇌성 운동실조증
ATXN7	CAG	4-35	37-306	척수소뇌성 운동실조증
C9ofr72	GGGGCC	< 30	100's	ALS

[0118]

[0119] 반복 확장들을 수반하는 유전적 장애들은 많은 양태에서 불균질하다. 반복 단위의 크기, 확장의 정도, 영향받은 유전자에 대한 위치, 및 병원성 메커니즘은 장애마다 다를 수 있다. 예를 들어, ALS는 염색체 9 개방 판독 프레임(open reading frame) 72의 짧은 아암 상에 위치된 *C9orf72* 유전자 내의 뉴클레오티드 GGGGCC의 헥사뉴클레오티드 반복 확장을 수반한다. 대조적으로, 취약 X 증후군은 X 염색체 상의 취약 X 정신 지체 1(*FMR1*) 유전자에 영향을 미치는 CGG 트리뉴클레오티드 반복체(삼중 반복체)의 확장과 연관된다. CGG 반복체의 확장은 정상적인 신경 발달에 필요한 취약 X 정신 지체 단백질(FMRP: fragile X mental retardation protein)을 발현하지 못하도록 할 수 있다. CGG 반복체의 길이에 따라, 대립유전자는 정상(상기 증후군에 의해 영향받지 않음), 돌연변이-전(취약 X 연관 장애의 위험에 있음), 또는 완전 돌연변이(일반적으로 상기 증후군에 의해 영향받음)로 분류될 수 있다. 다양한 추정치에 따르면, 운동실조증에 걸리기 쉬운 보균자들에서 60 내지 230개의 반복체들, 및 영향받지 않은 개체들에서 5 내지 54개의 반복체들이 존재하는 것과 비교하여, 영향받은 환자들에서 취약 X 증후군을 유발하는 돌연변이된 *FMR1* 유전자들 내에는 230 내지 4000개의 CGG 반복체들이 존재한다. *FMR1*

유전자의 반복 확장은 자폐증의 원인인데, 그 이유는 자폐증 개체들의 약 5%가 *FMR1* 반복 확장을 갖는 것으로 밝혀졌기 때문이다. 문헌[McLennan, et al. (2011), Fragile X Syndrome, Current Genomics 12 (3): 216-224]. 취약 X 증후군의 확정적인 진단은 CGG 반복체들의 개수를 결정하기 위한 유전자 검사를 수반한다.

[0120] 반복 확장 관련 질병들의 다양한 일반적인 특성들이 복수의 연구에서 식별되었다. 반복 확장 또는 동적 돌연변이는 일반적으로, 반복체 개수의 증가로서 나타나며, 돌연변이 비율은 반복체들의 개수와 관련된다. 반복 중단의 손실과 같은 드문 이벤트들은 대립유전자들이 확장될 가능성이 증가되도록 유도할 수 있으며, 이러한 이벤트들은 창시자 이벤트(founder event)로 알려져 있다. 반복 서열에서의 반복체들의 개수와, 반복 확장에 의해 유발되는 질병의 중증도 및/또는 발병 사이에 관계가 있을 수 있다.

[0121] 따라서, 반복 확장들을 식별하고 호출하는 것은 다양한 질병들의 진단 및 치료에 중요하다. 그러나, 특히 반복 서열을 완전히 횡단하지 않는 판독물들을 사용하여, 반복 서열들을 식별하는 것은, 다양한 난제를 갖는다. 첫째, 반복체들을 참조 서열에 정렬시키는 것은 어려운데, 그 이유는 판독물과 참조 게놈 사이에 명확한 일대일 맵핑이 없기 때문이다. 추가적으로, 판독물이 참조 서열에 정렬되더라도, 판독물은 종종 너무 짧아서 의학적으로 관련된 반복 서열을 완전히 커버할 수 없다. 예를 들어, 판독물들은 약 100 bp일 수 있다. 비교하면, 반복 확장은 수백 내지 수천개의 염기쌍들에 걸쳐 있을 수 있다. 예를 들어, 취약 X 증후군에서, *FMR1* 유전자는 1000 bp를 훨씬 초과하여, 3000 bp를 초과하여 걸쳐 있는 반복체들을 가질 수 있다. 따라서 100-bp 판독물은 반복 확장의 전체 길이에 맵핑될 수 없다. 더욱이, 짧은 판독물들을 더 긴 서열로 어셈블하는 것은 짧은 판독물 대 긴 반복체 문제를 극복하지 못할 수 있는데, 그 이유는 하나의 판독물 내의 반복체들과 다른 판독물 상의 반복체들의 모호한 정렬로 인해 짧은 판독물들을 더 긴 서열로 어셈블하는 것이 어렵기 때문이다.

[0122] 정렬은, 참조 서열의 불완전성, 판독물과 참조 서열 상의 부위들 사이의 비-고유 대응, 또는 참조 서열로부터의 상당한 편차로 인해, 정보 손실의 주요 원인이다. 시스템적인 서열분석 오차들, 및 판독 정확도에 영향을 주는 다른 문제들은 반복 서열들을 검출하는 데 있어서의 실패에 대한 2차 인자이다. 일부 실험 프로토콜들에서, 약 7% 판독물들은 정렬되지 않거나 또는 0의 MAPQ 점수를 갖는다. 연구자들이 서열분석 기술 및 분석 도구들을 개선하기 위해 노력하더라도, 항상 상당한 양의 정렬 불가능한 판독물들 및 불량하게 정렬된 판독물들이 존재할 것이다. 본원의 방법들의 구현예들은 반복 확장들을 식별하기 위해, 정렬 불가능한 판독물들 또는 불량하게 정렬된 판독물들에 의존한다.

[0123] 반복 확장들을 검출하기 위해 긴 판독물들을 사용하는 방법들은 그 자신의 난제를 갖는다. 차세대 서열분석에서, 더 긴 판독물들을 사용하는 현재 이용 가능한 기술들은 더 짧은 판독물들을 사용하는 기술들보다 더 느리고 더 오차가 발생하기 쉽다. 더욱이, 긴 판독물들은 무세포 DNA의 서열분석과 같은 일부 응용 분야들에서는 실현 가능하지 않다. 모체 혈액에서 획득된 무세포 DNA는 산전 유전자 진단을 위해 사용될 수 있다. 무세포 DNA는 전형적으로 200 bp보다 짧은 단편들로서 존재한다. 따라서, 긴 판독물들을 사용하는 방법들은 무세포 DNA를 사용하는 산전 유전자 진단에 적합하지 않다. 본원에 기술된 방법들의 구현예들은 의학적으로 관련된 반복 확장들을 식별하기 위해 짧은 판독물들을 사용한다.

[0124] 또한, 통상적인 방법들은 복수의 반복체들을 보유하는 복잡한 좌위들을 취급하도록 설계되지 않는다. 이러한 좌위들의 중요한 예는 CCG 반복체에 의해 플랭킹된 HD를 유발하는 CAG 반복체, 아데노신 동종중합체에 의해 플랭킹된 FRDA를 유발하는 GAA 반복체, 및 ACT 반복체에 의해 플랭킹된 척수소뇌성 운동실조증 8형(SCA8: Spinocerebellar ataxia type 8)을 유발하는 CAG 반복체를 포함한다. 더욱 더 극단적인 예는, 확장이 근긴장성 이영양증 2형(DM2: Myotonic Dystrophy type 2)을 유발하는, *CNBP* 유전자 내의 CCTG 반복체이다. 이러한 반복체는 다형성 TG 및 TCTG 반복체들에 인접해 있어서(문헌[J. E. Lee and Cooper 2009]), 이 좌위에 판독물들을 정확하게 정렬시키는 것을 특히 어렵게 만든다. 다른 유형의 복잡한 반복체는 지금까지 적어도 9개의 장애들과 연관된 폴리알라닌 반복체이다(문헌[Shoubridge and Gecz 2012]). 폴리알라닌 반복체는 α-아미노산 코돈 GCA, GCC, GCG, 또는 GCT의 반복으로 구성된다.

[0125] 변이체들의 클러스터는 정렬 및 유전자형분석 정확도에 영향을 미칠 수 있다(문헌[Lincoln et al. 2019]). 낮은 복잡도 다형성 서열에 인접한 변이체들은 추가적으로 문제가 될 수 있는데, 그 이유는 변이체 발견을 위한 방법들이 이러한 게놈 영역들에서 일관성 없게 표현되거나 가짜인 변이체 호출들의 클러스터들을 출력할 수 있기 때문이다. 이는, 부분적으로, 서열분석 데이터에서의 이러한 영역의 상승된 오차율에 기인한다(문헌[Benjamini and Speed 2012; Dolzhenko et al. 2017]). 일례는 린치 증후군 I(Lynch syndrome I)을 유발하는 *MSH2* 내의 아데노신 동종중합체에 인접한 단일-뉴클레오티드 변이체(SNV: single-nucleotide variant)이다(문헌[Froggatt et al. 1999]).

[0126] 본원에 개시된 구현예들은 전술된 바와 같이 복잡한 좌위들을 취급할 수 있다. 이들은 각각의 표적 좌위의 일반적이고 유연한 모델로서 서열 그래프를 사용한다.

[0127] 일부 구현예들에서, 개시된 방법들은 쌍형성된 말단 서열분석을 이용함으로써 반복 확장들을 식별하고 호출하는 데 있어서 전술한 난제들을 해결한다. 쌍형성된 말단 서열분석은 DNA를 삽입물로 불리는 서열들로 단편화하는 것을 수반한다. Illumina에 의해 사용되는 일부 프로토콜과 같은 일부 프로토콜에서, 더 짧은 삽입물로부터의 (예를 들어, 대략 수십 내지 수백 bp) 판독물은 짧은-삽입물 쌍형성된 말단 판독물 또는 간단히 쌍형성된 말단 판독물로 지칭된다. 대조적으로, 더 긴 삽입물로부터의(예를 들어, 대략 수천 bp) 판독물은 메이트 쌍 판독물(mate pair read)로 지칭된다. 상기에 언급된 바와 같이, 짧은-삽입물 쌍형성된 말단 판독물 및 긴-삽입물 메이트 쌍 판독물은 둘 모두 본원에 개시된 방법들의 다양한 구현예들에서 사용될 수 있다.

[0128] 도 1a는, 특히 반복 확장을 갖는 긴 반복 서열의 샘플로부터 획득된 서열 판독물들을 정렬시킬 때, 참조 서열 상의 반복 서열에 서열 판독물들을 정렬시키는 데 있어서의 특정한 어려움을 보여주는 개략도이다. 도 1a의 하부에는 수직 해치 선들(hatch lines)에 의해 예시된 비교적 짧은 반복 서열(103)을 갖는 참조 서열(101)이 있다. 도면의 중간에는, 또한 수직 해치 선들에 의해 예시된, 반복 확장을 보유하는 긴 반복 서열(107)을 갖는 환자 샘플의 가상 서열(105)이 있다. 도면의 상부에는 샘플 서열(105)의 대응하는 부위들의 위치들에 보여진 서열 판독물들(109, 111)이 예시되어 있다. 이들 서열 판독물 중 일부, 예를 들어, 판독물들(111)에서, 일부 염기쌍은, 수직 해치 선들에 의해 또한 예시되고 원 내에서 하이라이트되는 바와 같이, 긴 반복 서열(107)로부터 유래한다. 이들 반복체들을 갖는 판독물들(111)은 참조 서열(101)에 정렬시키기가 잠재적으로 어려운데, 그 이유는 반복체들이 참조 서열(101) 상에서 명확한 대응하는 위치를 갖지 않기 때문이다. 이러한 잠재적으로 비정렬된 판독물들은 참조 서열(101) 내의 반복 서열(103)과 명확히 연관될 수 없기 때문에, 이들 잠재적으로 비정렬된 판독물들(111)로부터 반복 서열의 확장 및 반복 서열에 관한 정보를 획득하는 것은 어렵다. 더욱이, 이들 판독물은 반복 확장을 보유하는 긴 반복 서열(107)보다 더 짧은 경향이 있기 때문에, 이들은 반복 서열(107)의 아이덴티티 또는 위치에 관한 확정적인 정보를 직접 제공할 수 없다. 추가적으로, 판독물들(111) 내의 반복체는, 참조 서열(101) 상의 이의 모호한 대응 위치 및 판독물들(111) 사이의 모호한 관계로 인해, 어셈블하기가 어렵다. 샘플 내의 긴 반복 서열(107)로부터 부분적으로 유래하는 판독물, 반은 해칭되고 반은 검은색-실선인 것으로 예시된 것들은, 반복 서열(107)의 외부로부터 유래하는 염기에 의해 정렬될 수 있다. 판독물이 반복 서열(107)의 밖에서 너무 적은 수의 염기쌍들을 갖는 경우, 판독물은 불량하게 정렬될 수 있거나 정렬되지 않을 수 있다. 따라서, 부분 반복체들을 갖는 이들 판독물 중 일부는 앵커 판독물들로서 분석될 수 있고, 다른 것들은 하기에 추가로 기술되는 바와 같이 앵커링된 판독물들로서 분석될 수 있다.

[0129] 도 1b는 도 1a에 도시된 어려움을 극복하기 위해 일부 개시된 실시형태들에서 쌍형성된 말단 판독물들이 어떻게 이용될 수 있는지를 예시하는 개략도이다. 쌍형성된 말단 서열분석에서, 서열분석은 테스트 샘플 내의 핵산들의 단편들의 양 말단들로부터 발생한다. 도 1b의 하부에는, 참조 서열(101) 및 샘플 서열(105)뿐만 아니라, 도 1a에 도시된 것들과 동등한 판독물들(109, 111)이 예시되어 있다. 도 1b의 상부에는, 테스트 샘플 서열(105)로부터 유도된 단편(125), 및 쌍형성된 말단 판독물들의 2개의 판독물들(135, 137)을 획득하기 위한 판독물 1 프라이머 영역(131) 및 판독물 2 프라이머 영역(133)이 예시되어 있다. 단편(125)은 또한 쌍형성된 말단 판독물들을 위한 삽입물로 지칭된다. 일부 실시형태들에서, 삽입물들은 PCR을 사용하여 또는 PCR 없이 증폭될 수 있다. 다수의 GC 또는 GCC 반복체들을 포함하는 것들과 같은 일부 반복 서열은 PCR 증폭을 포함하는 전통적인 방법으로 양호하게 서열분석될 수 없다. 이러한 서열들의 경우, 증폭은 PCR이 없을 수 있다. 다른 서열들의 경우, 증폭은 PCR을 사용하여 수행될 수 있다.

[0130] 도 1b에 예시된 삽입물(125)은 도면의 하부 절반부에 예시된 2개의 수직 화살표에 의해 플랭킹된 샘플 서열(105)의 섹션에 대응하거나 또는 이로부터 유도된다. 구체적으로, 삽입물(125)은 샘플 서열(105) 내의 긴 반복체(107)의 일부에 대응하는 반복 섹션(127)을 보유한다. 삽입물의 길이는 다양한 응용 분야를 위해 조정될 수 있다. 일부 실시형태들에서, 삽입물들은 관심있는 반복 서열 또는 반복 확장을 갖는 반복 서열보다 다소 더 짧을 수 있다. 다른 실시형태들에서, 삽입물들은 반복 서열 또는 반복 확장을 갖는 반복 서열과 유사한 길이를 가질 수 있다. 또 다른 실시형태들에서, 삽입물들은 반복 서열 또는 반복 확장을 갖는 반복 서열보다 다소 더 길 수도 있다. 이러한 삽입물들은 이하에서 추가로 기술되는 일부 실시형태들에서 메이트 쌍 서열분석을 위한 긴 삽입물들일 수 있다. 전형적으로, 삽입물들로부터 획득된 판독물들은 반복 서열보다 짧다. 삽입물들은 판독물들보다 길기 때문에, 쌍형성된 말단 판독물들은 단일-말단 판독물들보다 샘플 내의 반복 서열의 더 긴 신장부로부터 신호를 더 잘 캡처할 수 있다.

[0131] 예시된 삽입물(125)은 삽입물의 2개의 말단들에서 2개의 판독물 프라이머 영역들(131, 133)을 갖는다. 일부 실

시형태들에서, 판독물 프라이머 영역은 삽입물에 내재한다. 다른 실시형태들에서, 프라이머 영역은 결찰 또는 연장에 의해 삽입물에 도입된다. 삽입물의 좌측 말단에는 판독물 1 프라이머 영역(131)이 예시되어 있으며, 이는 삽입물(125)에 대한 판독물 1 프라이머(132)의 혼성화를 허용한다. 판독물 1 프라이머(132)의 연장은 135로 라벨링된, 제1 판독물 또는 판독물 1을 생성한다. 삽입물(125)의 우측 말단에는 판독물 2 프라이머 영역(133)이 예시되어 있으며, 이는 삽입물(125)에 대한 판독물 2 프라이머(134)의 혼성화를 허용하여, 137로 라벨링된, 제2 판독물 또는 판독물 2를 개시한다. 일부 실시형태들에서, 삽입물(125)은 또한 인덱스 바코드 영역(여기서는 도면에 도시되지 않음)을 포함하여, 멀티플렉스 서열분석 과정에서 상이한 샘플들을 식별하기 위한 메커니즘을 제공할 수 있다. 일부 실시형태들에서, 쌍형성된 말단 판독물들(135, 137)은 Illumina의 합성 플랫폼에 의한 서열분석에 의해 획득될 수 있다. 이러한 플랫폼 상에서 구현된 서열분석 방법의 일례가 이후에 서열분석 방법 섹션에 추가로 기술되며, 이 방법은 2개의 쌍형성된 말단 판독물들 및 2개의 인덱스 판독물들을 생성한다.

[0132] 이어서, 도 1b에 예시된 바와 같이 획득된 쌍형성된 말단 판독물들은 비교적 짧은 반복 서열(103)을 갖는 참조 서열(101)에 정렬될 수 있다. 이와 같이, 한 쌍의 판독물의 상대적 위치 및 방향은 알려져 있다. 이는, 원(111) 내에 도시된 것들과 같은 정렬 불가능하거나 불량하게 정렬된 판독물이 도 1b의 하부에서 볼 수 있는 바와 같이 판독물의 대응하는 쌍형성된 판독물(109)을 통해 샘플 서열(105) 내의 비교적 긴 반복 서열(107)과 간접적으로 연관될 수 있게 한다. 예시적인 예에서, 쌍형성된 말단 서열분석으로부터 획득된 판독물들은 약 100 bp이며, 삽입물들은 약 500 bp이다. 이러한 예시적인 구성에서, 2개의 쌍형성된 말단 판독물들의 상대적 위치는 이들의 3' 말단으로부터 약 300개의 염기쌍들로 떨어져 있고, 이들은 반대 방향을 갖는다. 판독물 쌍들 사이의 관계는 판독물들을 반복 영역들에 더 잘 연관시킬 수 있게 한다. 일부 경우에 있어서, 한 쌍 내의 제1 판독물은 참조 서열 상의 반복 영역을 플랭킹하는 비-반복 서열과 정렬되고, 해당 쌍 내의 제2 판독물은 참조물에 적절하게 정렬되지 않는다. 예를 들어, 도 1b의 하부 절반부에 예시된 한 쌍의 판독물들(109a, 111a)을 참조하며, 해당 쌍의 좌측의 것(109a)은 제1 판독물이고, 우측의 것(111a)은 제2 판독물이다. 2개의 판독물들(109a, 111a)의 쌍형성을 고려하면, 제2 판독물(111a)이 참조 서열(101)에 정렬될 수 없다는 사실에도 불구하고, 제2 판독물(111a)은 샘플 서열(105) 내의 반복 영역(107)과 연관될 수 있다. 제1 판독물(109a)에 대한 제2 판독물(111a)의 거리 및 방향을 알면, 추가로, 긴 반복 영역(107) 내에서의 제2 판독물(111a)의 위치를 결정할 수 있다. 제2 판독물(111a) 내의 반복체들 사이에 절단부가 존재하는 경우, 참조 서열(101)에 대한 절단부의 위치가 또한 결정될 수 있다. 참조물에 정렬되는 좌측 판독물(109a)과 같은 판독물은 본 개시내용에서 앵커 판독물로 지칭된다. 참조 서열에 정렬되지 않지만 앵커 판독물과 쌍을 이루는 우측의 것(111a)과 같은 판독물은 앵커링된 판독물로 지칭된다. 이와 같이, 비정렬된 서열은 반복 확장에 앵커링되고 이와 연관될 수 있다. 이러한 방식으로, 긴 반복 확장들을 검출하기 위해 짧은 판독물들을 사용할 수 있다. 반복 확장들을 검출하는 과제는 전형적으로, 서열분석의 증가된 어려움으로 인해 확장의 길이에 따라 증가하지만, 본원에 개시된 방법들은 더 짧은 반복 확장 서열들로부터의 것보다 더 높은 신호를 더 긴 반복 확장 서열들로부터 검출할 수 있다. 그 이유는, 반복 서열 또는 반복 확장이 더 길어짐에 따라, 더 많은 판독물이 확장 영역에 앵커링될 것이고, 더 많은 판독물이 반복 영역 내에 완전히 속할 수 있고, 판독물당 더 많은 반복체들이 발생할 수 있기 때문이다.

[0133] 도 2a 및 도 2b는 쌍형성된 말단 판독물들을 사용하더라도 TR 영역에 판독물들을 정렬시키는 것이 어려운 시나리오를 예시한다. 그 이유는, TR 영역으로부터 유도된 서열 판독물들이 TR 영역의 상이한 게놈 위치들에 정렬되거나 두 대립유전자 중 하나에 정렬될 수 있기 때문이다.

[0134] 도 2a는 빗금 패턴 및 2개의 플랭킹 영역으로 도시된 반복 서열을 포함하는, 반복 영역의 2개의 대립유전자들을 나타낸다. 대립유전자 1은 상단에, 대립유전자 2는 하단에 도시되었다. 대립유전자 1은 대립유전자 2보다 짧은 TR 서열을 갖는다. 한 쌍의 서열 판독물들(20)은 두 대립유전자 각각의 하나의 위치에 고유하게 정렬될 수 있다. 도 2b는 2개의 대립유전자들과 TR 서열로부터 유도된 한 쌍의 판독물들(22)을 도시한다. 상기 쌍의 두 판독물 모두는 반복 서열 상의 상이한 위치들에 정렬될 수 있다. 두 판독물의 상대 위치를 제한하더라도, 반복 서열 상의 복수의 위치에 여전히 정렬할 수 있다. 이들은 또한 대립유전자 중 하나에 정렬될 수 있다. 판독물 쌍의 정렬 위치의 모호성을 감안할 때, 판독물 쌍이 실제로 유도되는 게놈 영역의 위치를 결정하는 것은 어렵거나 불가능한다. 이것은 또한 대립유전자들에 대한 판독물들의 정렬을 시각화하는 것을 어렵게 만든다.

[0135] 상기에 설명된 바와 같이 짧은 판독물을 사용하여 TR(특히 STR)들을 유전자형분석하는 데 있어서의 기술적인 어려움으로 인해, 서열 판독 데이터 및 상기 서열 판독 데이터로부터 결정된 유전자형들을 시각화하기 위한 컴퓨터-구현 도구들을 개발하는 것이 바람직하다. 이러한 도구들은 유전자형 호출들을 검증하고 STR들과 관련된 임상적으로 및 생물학적으로 중요한 유전적 특징들을 이해하는 데 도움이 될 수 있다. 예를 들어, 이러한 시각화 도구들은 임상적으로 중요한 효과를 가질 수 있는 반복 모티프의 변화들(예를 들어 중단들)을 검출하는 것을 가

능하게 한다.

[0136] ***STR에 대한 서열 판독물 파일업의 시각화***

[0137] 반복 영역들에 서열 판독물들을 정렬시키는 것이 어렵기 때문에, 정렬의 품질을 검사하고 반복 영역들의 유전자형들을 검증하기 위해 탠덤 반복 영역들에 정렬된 서열 판독물들을 시각화하기 위한 컴퓨터-구현 도구들을 개발하는 것이 중요하다. 그러나, 통상적인 시각화 도구들은 표준 참조 서열에 서열 판독물들을 정렬시킨다. 도 3a는 STR 서열을 포함하는 참조 서열에 정렬된 서열 판독물들의 통상적인 시각적인 표현을 개략적으로 예시한다. 참조물 및 참조 서열에 정렬된 서열 판독물들의 시각적인 표현은 서열 판독 "파일업"으로 지칭된다.

[0138] 도 3a에 도시된 바와 같은 통상적인 시각화 도구들은 개별 샘플 또는 대상체에 대해 맞춤화되지 않은 표준 참조 서열을 사용한다. 이러한 접근법은 반복 확장들을 갖는 탠덤 반복 영역들을 시각화하는 데 있어 다양한 제한을 갖는다. 이것은 개별 샘플의 탠덤 반복 서열의 실제 길이와 세부 사항을 효과적으로 반영할 수 없다. 참조 서열에 없는 반복 모티프들을 포함하는 서열 판독물들은 절단될 수 있다. 예를 들어 서열 판독물(32)을 참조한다. 개별 샘플의 반복 서열이 참조 서열의 반복 서열보다 짧은 경우(이 예에서는 표시되지 않음), TR 서열로부터 유도된 서열 판독물들은 균일하지 않은 커버리지를 생성할 수 있다.

[0139] 본 개시내용의 일부 구현예들은 탠덤 반복 영역들을 시각화하기 위한 컴퓨터 그래픽을 생성하기 위한 컴퓨터-구현 도구들을 제공한다. 상기 도구들은 서열 판독물 파일업을 생성하며, 각각의 파일업은 샘플에 특이한 복수의 일배체형들을 포함한다. 도 3b에 도시된 예에서, 샘플은 2개의 상이한 일배체형들을 갖는다. 제1 일배체형(34)은 상단에 도시되어 있으며, 이는 하단에 도시된 제2 일배체형(36)보다 더 짧은 탠덤 반복 영역을 갖는다. 서열 판독물들은 두 일배체형 각각에 정렬된다. 서열 판독물이 일배체형 상의 복수의 위치들에 정렬될 수 있는 경우, 빗금 패턴으로 도시된 탠덤 반복 영역 내에서 종종, 서열 판독물들은 일배체형 상에 균일하게 분포되어 일배체형 전체에 걸쳐 균일한 커버리지를 만든다.

[0140] 일부 구현예들에서, 일배체형은 여기에 제시된 바와 같은 하나의 반복 서열을 포함할 수 있다. 다른 구현예에서, 일배체형은 복수의 반복 서열들을 포함할 수 있다. 반복체의 유전자형들을 결정하기 위한 유전자형분석 도구들이 이 변형체 유형을 효과적으로 검출하지 못하는 경우에도, 이들은 짧은 삽입-결실을 시각화하는 데 사용될 수 있다. 본원에 기술된 다양한 구현예들은 TR 영역들을 시각화하지만, 이들은 다양한 일배체형 상에서 다양한 유전자형을 갖는 다른 유형의 변이체들을 시각화하는 데에도 사용될 수 있다.

[0141] 다양한 구현예들에서, 각각의 서열 파일업은 샘플에 대해 맞춤화된 개별화된 일배체형들을 포함한다. 이것은 반복 영역의 길이와 서열의 더 양호한 시각화를 허용한다. 반복 서열의 및 반복체를 바로 둘러싼 서열의 중단들을 검출하기 위해 이들 플롯을 사용할 수 있다. 이것은 또한 일배체형들에 대한 정렬의 특성을 검사하도록 허용하여, 게놈 영역 내의 반복 서열들의 유전자형들을 검증하는 수단을 제공한다. 이하의 실험 데이터에서 예시된 바와 같이, 제공된 일배체형들이 정확한 경우 서열 판독물들은 일배체형들 상에 균일하게 분포되는 경향이 있고, 상이한 게놈 위치들이 유사한 커버리지를 갖는 경향이 있다.

[0142] 또한, 일부 구현예들은 생물학적 또는 임상적 중요성을 갖는 모티프들 및 뉴클레오티드들의 시각화를 허용한다. 일부 구현예들에서, 일배체형은 복수의 TR 서열들을 포함할 수 있다. 이러한 응용 분야들에서, 서열 데이터는 단계적으로 분석되고 2개 이상의 염색체들에 대한 일배체형들로 조합될 필요가 있을 수 있다. TR 서열들의 유전형들은, 후술되는 서열 그래프 기술들 및 쌍형성된 말단 판독물 앵커링 기술들과 같은 다양한 기술들을 사용하여 결정될 수 있다. 일부 구현예들에서, 전체 게놈으로부터의 서열 판독물 데이터는 서열 판독물들의 서브세트를 제공하기 위해 본원에 기술된 기술들을 사용하여 사전-프로세싱될 수 있다.

[0143] 도 4는, 반복 영역을 시각화하는 데 사용되는 일배체형들 및 서열 판독물들을 획득하기 위해 서열 그래프 정렬 기술들을 사용하는 일부 구현예들에 따른 개략적인 작업 흐름을 도시한다. 도 4의 패널 1은 반복 서열들을 포함하는 관심 있는 표적 영역으로부터 획득된 서열 판독물들을 예시한다. 상기 판독물들은 쌍형성된 말단 판독물들이다. 이들은, 예를 들어, 통상적인 정렬 방법들을 사용하여 게놈에 전체 게놈 판독물들을 정렬시키고, 표적 영역에 또는 그 근처에 정렬된 판독물들을 선택함으로써, 획득될 수 있다.

[0144] 도 4의 패널 2는, 표적 영역에 대해 서열 판독물들이 획득된 후, 서열 판독물들이 표적 영역을 표현하는 서열 그래프에 정렬되는 것을 예시한다. 이 서열 그래프에 의해 좌측에서 우측으로 표현되는 반복 영역은 좌측 플랭킹 영역, CAG 탠덤 반복 서열, CAACAG 개재 서열, CCG 탠덤 반복 서열 및 우측 플랭킹 영역을 포함한다.

[0145] 상기 서열 그래프에 대한 판독물 정렬은 패널 3에 도시된 재정렬 서열 판독물들을 제공한다. 상기 서열 그래프에 서열 판독물들을 정렬시켜 재정렬 판독물들을 획득하는 것에 대한 추가적인 세부 사항은 도 8 내지 도 13을

참조하여 이후에 설명된다.

[0146] 상기 서열 그래프에 대한 판독물 정렬은 또한 반복 영역 내의 STR 서열들의 유전자형들을 결정한다. 상기 서열 그래프에 대한 서열 판독물들의 정렬은, CAG STR의 하나의 대립유전자가 4개의 반복 단위들을 포함하고 다른 대립유전자는 78개의 반복 단위들을 포함하는 것으로 결정한다. 서열 그래프 정렬은 또한 CCG STR이 하나의 대립유전자에서 7개의 반복 단위들을 포함하고 다른 대립유전자에서 10개의 반복 단위들을 포함하는 것으로 결정한다. 결정된 유전자형들을 고려하여, 두 쌍의 가능한 일배체형 서열들을 도 4의 패널 5에 도시하였다. 일부 구현예들은, 재정렬된 판독물들과 가장 매칭하는 일배체형 쌍을 결정하기 위해 유전자형들을 단계적으로 분석하는 것을 수반한다. 패널 6에 도시된 바와 같이, 최상의 일배체형 쌍은 4개의 CAG 반복 단위들 및 7개의 CCG 반복체들을 포함하는 제1 일배체형 및 78개의 CAG 반복체들 및 10개의 CCG 반복체들을 포함하는 제2 일배체형을 갖는다.

[0147] 그런 다음, 예시된 구현예는 재정렬된 판독물들 및 최상의 일배체형 쌍을 사용하여 반복 영역의 파일업을 시각화한다. 일부 구현예들에서, 상이한 서열 판독물 쌍들은 상이한 가능한 정렬 시나리오들을 가질 수 있다. 예를 들어, 서열 판독물 쌍 "a"는 일배체형 1 상의 한 위치 및 일배체형 2 상의 동일한 위치에 정렬된다. 서열 판독물 쌍 "b"는 일배체형 2 상의 복수의 위치들에 정렬된다. 서열 판독물 쌍 "c"는 일배체형 1 상의 단일 위치에 정렬된다. 서열 판독물 쌍 "d"는 일배체형 1 상의 한 위치 및 일배체형 2 상의 대응 위치에 정렬된다.

[0148] 일부 구현예들은 각각의 판독물 쌍에 대한 모든 가능한 정렬 위치들을 결정한다. 두 판독물 쌍 모두가 동일한 일배체형에 정렬된다. 그런 다음, 정렬 위치들의 세트를 결정하기 위해 모든 판독물 쌍에 대해 각각의 판독물 쌍의 무작위 위치가 선택된다. 정렬 위치들의 복수의 세트들을 획득하기 위해 동일한 무작위 선택을 반복한다. 다양한 구현예들에서, 정렬 위치들의 적어도 1,000, 5,000, 10,000, 50,000, 또는 100,000개의 세트들이 획득된다. 2개의 일배체형들 상에서 가장 균일한 분포를 갖는 정렬 위치들의 세트를 선택하여, 패널 8에 도시된 바와 같이 2개의 일배체형들 및 2개의 일배체형들에 정렬된 서열 판독물 쌍들을 포함하는 파일업을 생성한다.

[0149] 도 5는 게놈 영역의 일배체형들에 정렬된 서열 판독물들을 표현하는 컴퓨터 그래픽을 생성하기 위한 방법(50)에 대한 흐름도를 도시한다. 상기 그래픽은 전술된 바와 같은 서열 판독물 파일업을 포함한다. 방법(50)은, 게놈 영역의 복수의 일배체형들에 대응하는 복수의 일배체형 서열들에 정렬된 복수의 서열 판독물들의 정렬 위치들의 복수의 세트들을 결정하는 단계를 수반한다. 박스(52)를 참조한다. 상기 복수의 서열 판독물들은 핵산 샘플의 게놈 영역으로부터 획득된다. 방법(50)은, 정렬 위치들의 복수의 세트들에서 정렬 위치들의 다른 세트들보다 상기 복수의 일배체형들 상에 더 균일하게 분포된 정렬 위치들의 세트를 선택하는 단계를 추가로 수반한다. 블록(54)을 참조한다. 방법(50)은 상기 복수의 서열 판독물들 및 상기 복수의 일배체형들을 표현하는 컴퓨터 그래픽을 생성하는 단계를 추가로 수반한다. 상기 복수의 서열 판독물들은 정렬 위치들의 선택된 세트에 위치한다. 일부 구현예들에서, 방법(50)은 도 6에 도시된 방법(600)에 기술된 특징들을 포함할 수 있다.

[0150] 도 6은 복수의 일배체형들을 포함하는 서열 판독물 파일업을 표현하는 컴퓨터 그래픽을 생성하기 위한 방법(600)에 대한 흐름도를 도시한다. 방법(600)은 게놈 영역의 복수의 일배체형들에 대응하는 복수의 일배체형 서열들 상의 정렬 위치들의 세트에 복수의 서열 판독물들을 정렬시키는 단계를 수반한다. 박스(602)를 참조한다. 일부 구현예들에서, 상기 복수의 서열 판독물들은 적어도 100, 500, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 또는 10,000개의 서열 판독물들을 포함한다.

[0151] 일부 구현예들에서, 상기 복수의 일배체형들 중 적어도 하나의 일배체형은 반복 확장을 포함한다. 일부 구현예들에서, 상기 복수의 일배체형들은 염색체 쌍 상의 게놈 영역의 2개의 일배체형들을 포함한다. 다양한 구현예들에서, 상기 복수의 일배체형들은 적어도 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90 또는 100개의 일배체형들을 포함한다. 다양한 구현예들에서, 상기 게놈 영역은 적어도 20, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 또는 10,000 bp를 포함한다.

[0152] 일부 구현예들에서, 상기 복수의 일배체형들 중 적어도 하나의 일배체형은 구조적 변이체를 포함한다. 일부 구현예들에서, 상기 구조적 변이체는 50 bp보다 길다. 상기 구조적 변이체는 결실, 중복, 복제수 변이체, 삽입, 역위, 전좌 등일 수 있다.

[0153] 일부 구현예들에서, 상기 구조적 변이체는 50 bp보다 짧다. 일부 구현예들에서, 50 bp보다 짧은 구조적 변이체는 단일 뉴클레오티드 다형성(SNP)을 포함한다.

[0154] 도 7은 정렬 위치들의 세트에 서열 판독물들을 정렬시키기 위한 방법(700)의 흐름도를 도시한다. 일부 구현예

들에서, 방법(600)의 동작(602)은 방법(700)에 따라 구현될 수 있다. 방법(700)은 각각의 일배체형에 대한 각각의 판독물의 가능한 정렬 위치들을 결정하는 단계를 수반하며, 여기서 상기 복수의 서열 판독물들은 쌍형성된 말단 서열분석에 의해 획득된 판독물 쌍들을 포함한다.

[0155] 방법(700)은, (A) 판독물 쌍의 두 판독물이 동일한 일배체형에 정렬되고 (B) 판독물 쌍의 대응하는 단편 길이가 평균 단편 길이와 최대한 가깝도록, 구성 판독물들의 정렬 위치들로부터 각각의 판독물 쌍에 대한 제한된 정렬 위치들을 생성하는 단계를 추가로 수반한다.

[0156] 방법(700)은 또한 제한된 정렬 위치로부터 각각의 판독물 쌍에 대한 정렬 위치를 무작위로 선택하는 단계를 수반한다. 일부 구현예들에서, 교체 기술(replacement technique)을 사용하지 않는 무작위 샘플링을 사용하여, 제한된 정렬 위치들로부터 판독물을 선택한다. 이러한 기술은 모든 위치 공간을 보다 빠르게 다룰 수 있다. 모든 위치를 샘플링한 후, 모든 샘플을 교체할 수 있다. 일부 구현예들에서, 교체 기술을 사용한 무작위 샘플링이 사용되며, 이는 마지막에 교체를 필요로 하지 않으며, 때때로, 교체하지 않는 경우보다 더 빨리 위치들의 원하는 조합을 획득할 수 있다. 이러한 후자의 접근법은, 고정된 반복 횟수 대신 미리 설정된 수렴 기준(예를 들어 원하는 정렬 점수)을 사용하여 정렬 위치들의 검색을 중지하는 경우, 시간을 절약할 수 있다.

[0157] 도 6으로 돌아가면, 일부 구현예들에서, 방법(600)은 상이한 게놈 영역들에 서열 판독물들의 상이한 세트들을 정렬시키는 단계를 수반한다. 일부 다양한 구현예들에서, 상기 상이한 게놈 영역들은 적어도 100, 200, 300, 500, 600, 700, 800, 900, 1,000, 5,000, 또는 10,000개의 영역들을 포함한다.

[0158] 일부 구현예들에서, 상기 복수의 일배체형들은 본원에 기술된 서열 그래프 정렬 기술들을 사용하여 획득될 수 있다. 다른 구현예들에서, 상기 복수의 서열 판독물들 및/또는 상기 복수의 일배체형들은 후술되는 쌍형성된 말단 판독물 앵커링 기술들을 사용하여 획득될 수 있다.

[0159] 일부 구현예들에서, 방법(600)은 게놈 영역에 대응하는 하나 이상의 서열 그래프들에 제1 개수의 서열 판독물들을 정렬시켜 상기 복수의 서열 판독물들 및/또는 상기 복수의 일배체형들을 획득하는 단계를 수반한다. 일부 구현예들에서, 상기 서열 그래프에 상기 제1 개수의 서열 판독물들을 정렬시키는 단계는 핵산 샘플의 제1 개수의 서열 판독물들을 제공하는 단계 및 서열 그래프에 의해 각각 표현된 하나 이상의 반복 서열들에 상기 제1 개수의 서열 판독물들을 정렬시키는 단계를 포함한다. 상기 서열 그래프는 핵산 서열들을 표현하는 정점들 및 상기 정점들을 연결하는 방향성 에지들을 갖는 방향성 그래프의 데이터 구조를 갖는다. 상기 서열 그래프는 하나 이상의 자가-루프들을 포함하며, 각각의 자가-루프는 반복 하위서열을 표현하고, 각각의 반복 하위서열은 하나 이상의 뉴클레오티드들의 반복 단위의 반복체들을 포함한다. 상기 서열 그래프에 상기 제1 개수의 서열 판독물들을 정렬시키는 단계는 또한, 상기 하나 이상의 반복 서열들의 하나 이상의 유전자형들을 결정하는 단계, 및 (a)의 복수의 서열 판독물들로서의 제1 개수의 서열 판독물들 및/또는 상기 하나 이상의 반복 서열들의 하나 이상의 유전자형들을 제공하는 단계를 포함한다.

[0160] 일부 구현예들에서, 방법(600)은 상기 하나 이상의 유전자형들을 단계적으로 분석하여 상기 복수의 일배체형들을 결정하는 단계를 추가로 포함한다. 일부 구현예들에서, 상기 방법은 초기에 게놈에 제2 개수의 서열 판독물들을 정렬시켜 상기 제1 개수의 서열 판독물들을 제공하는 단계를 추가로 수반한다. 상기 제2 개수의 서열 판독물들은 전체 게놈 판독물이고 적어도 10,000, 100,000, 1백만개의 서열 판독물들을 포함할 수 있다.

[0161] 방법(600)은 정렬 위치들의 세트에 대한 정렬 점수를 추정하는 단계를 추가로 수반한다. 블록(604)을 참조한다. 그런 다음, 방법(600)은 정렬 위치들의 복수의 상이한 세트들에 대해 반복하기 위해 동작(602)으로 되돌아간다. 일부 구현예들에서, 상기 방법은 제한된 반복 횟수 동안 루프를 돌 수 있다. 다양한 구현예들에서, 상기 방법은 상이한 정렬 위치들의 적어도 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 10,000, 20,000, 50,000, 100,000 또는 500,000개의 세트들을 획득한다. 다른 구현예들에서, 상기 방법은 정렬 점수가 기준을 충족할 때까지 반복을 반복한다. 다른 구현예들에서, 정렬 위치들에 대한 다른 정렬 메트릭들을 사용하여, 루프를 중지하기 위한 기준을 설정할 수 있다. 예를 들어, 정렬 품질 점수, 맵핑 품질 점수 또는 커버리지를 사용하여, 루프를 종료하는 기준을 설정할 수 있다.

[0162] 일부 구현예들에서, 상기 정렬 점수는 상기 복수의 서열 판독물들이 상기 복수의 일배체형들에 대응하는 복수의 일배체형 서열들 상에 얼마나 균일하게 분포되어 있는지를 나타낸다. 판독물들이 더 균일하게 분포되는 경우, 커버리지 수준이 일배체형 전체에 걸쳐 더 균일해진다. 개념적으로, 동일한 길이의 DNA 단편들을 사용하여 판독물들을 생성하고 DNA 단편들이 게놈 상에 균일하게 분포되어 있다고 가정하면, 판독물들의 가장 균일한 분포는 임의의 2개의 연속적이고 중첩되지 않는 판독물들 사이의 거리가 정확히 동일할 것이다. 판독물들이 덜 균

일하게 분포되기 때문에, 개별적인 연속적인 판독물들은 상기 거리의 평균에서 더 멀리 벗어난다. 따라서, 일부 구현예들에서, 상기 정렬 점수는 2개의 연속적인 판독물들의 시작 위치들 간의 거리의 평균으로부터의 제곱 평균 제곱근 차이를 포함한다. 상기 정렬 점수가 작을수록, 서열 판독물들이 일배체형들 상에 더 균일하게 분포되고 정렬 점수가 더 양호하다.

[0163] 일부 구현예들에서, 상기 정렬 점수는 판독물 쌍들이 상기 복수의 일배체형 서열들 상에 균일하게 분포된다고 가정하는 확률 모델을 사용하여 추정된다. 일부 구현예들에서, 상기 정렬 점수는, 상기 확률 모델을 고려하여 상기 복수의 서열 판독물들이 정렬 위치들의 세트로부터 도출되는 확률이다. 일부 구현예들에서, 상기 복수의 서열 판독물들은 핵산 단편들로부터 획득된 쌍-말단 판독물들을 포함하고, 상기 확률 모델은 평균 단편 길이를 입력으로서 수신하도록 구성된다. 일부 구현예들에서, 상기 확률 모델은 일배체형의 길이를 입력으로서 수신하도록 구성된다.

[0164] 일부 구현예들에서, $p_{(k)}$로 표시되는, 일배체형의 시작으로부터의 k번째 판독물 쌍의 개별 정렬 위치 x의 확률은 하기와 같이 모델링되며:

[0165]
$$\mathcal{P}\left(p_{(k)}=x\right)=\sum_{j=0}^{n_i-k}\binom{n_i}{j}\Big(\left(1-F(x)\right)^{j}\left(F(x)\right)^{n_i-j} - \left(1-F(x)+f(x)\right)^{j}\left(F(x)-f(x)\right)^{n_i-j}\Big)$$

[0166]
$$F(x)=\frac{x}{H_i-L-1},$$

[0167]
$$f(x)=\frac{1}{H_i-L-1},$$

[0168] 상기 식에서 i는 판독물 쌍이 정렬된 일배체형이고, H_i는 일배체형 i의 길이이고, L은 평균 단편 길이이고, n_i는 일배체형 i에 정렬된 판독물 쌍들의 개수이다. 정렬 위치들의 세트에 대한 정렬 점수는 개별 정렬 위치들의 확률의 곱으로서 추정된다.

[0169] 방법(600)은 상기 복수의 정렬 점수들에 기초하여 정렬 위치들의 복수의 상이한 세트들로부터 정렬 위치들의 세트를 선택하는 단계를 수반한다. 일부 구현예들에서, 정렬 위치들의 선택된 세트는 상이한 정렬 위치들의 복수의 세트들 중에서 최상의 정렬 점수를 갖는다. 일부 구현예들에서, 정렬 위치들의 선택된 세트는 선택 기준을 초과하는 정렬 점수를 갖는다. 일부 구현예들에서, 상기 선택 기준은 정렬 점수의 상위 1, 2, 3, 4, 5, 10, 20 백분위수일 수 있다. 이것은 정렬 위치들의 최종 세트를 선택함에 있어 고려될 하나 이상의 다른 메트릭들(예를 들어 커버리지, 맵핑 품질, 정렬 품질)과 상기 정렬 점수의 조합을 허용할 수 있다.

[0170] 일부 구현예들에서, 방법(600)은 임의로, 상기 복수의 서열 판독물들 및 상기 복수의 일배체형들을 표현하는 컴퓨터 그래픽을 생성하는 단계를 수반하며, 상기 복수의 서열 판독물들은 정렬 위치들의 선택된 세트에 위치한다. 블록(608)을 참조한다.

[0171] 일부 구현예들에서, 방법(600)은 동작(608)을 필요로 하지 않는다. 대신에, 이것은 게놈 영역의 위치들에 서열 판독물들을 할당할 수 있으며, 이러한 할당된 위치들은 컴퓨터 그래픽을 생성하거나 생성하지 않으면서 다른 다운스트림 프로세싱에 사용될 수 있다.

[0172] 일부 구현예들은, 정렬 위치들의 선택된 세트에서 상기 복수의 일배체형 서열들에 정렬된 복수의 서열 판독물들에 대한 하나 이상의 서열분석 메트릭들을 추정하는 단계를 수반한다. 일부 구현예들에서, 상기 하나 이상의 서열분석 메트릭들은 서열 커버리지를 포함한다. 일부 구현예들에서, 상기 하나 이상의 서열분석 메트릭들은 각각의 정렬 위치에 대한 서열 커버리지를 포함한다. 일부 구현예들에서, 상기 하나 이상의 서열분석 메트릭들은 정렬 품질 점수를 포함하며, 이는 판독 서열과 참조 서열 사이의 매칭 품질을 나타낸다. 일부 구현예들에서, 상기 하나 이상의 서열분석 메트릭들은 각각의 정렬 위치에 대한 정렬 품질 점수를 포함한다. 일부 구현예들에서, 상기 하나 이상의 서열분석 메트릭들은 맵핑 품질 점수를 포함하며, 이는 판독물이 게놈 좌표들에 올바르게 맵핑된다는 신뢰도를 나타낸다. 예를 들어, 하나의 판독물이, 모든 위치에서 거의 완벽하게 매칭하면서 여러 게놈 위치에 맵핑될 수 있다. 이 경우, 정렬 점수는 높지만 맵핑 품질은 낮을 것이다.

[0173] 서열분석 품질 메트릭들은 라이브러리 제조, 염기 호출, 판독물 정렬 및 변이체 호출을 포함한, 이러한 방법의 각각의 단계의 정확도에 대한 중요한 정보를 제공할 수 있다. 프레드(Phred) 품질 점수(Q 점수)에 의해 측정되는 염기 호출 정확도는 서열분석 플랫폼의 정확도를 평가하는 데 사용되는 일반적인 메트릭이다. 이것은, 특정 염기가 서열분석기에 의해 잘못 호출될 확률을 나타낸다. 도 24는 *C9ORF72* 반복체를 포함하는 게놈 영역에 대한 일부 구현예들에 따른 판독물들의 맵핑 품질 점수를 도시한다. 상단 패널은 짧은 반복체를 갖는 일배체형을 도시하고 하단 패널은 긴 반복체를 갖는 일배체형을 도시한다. 가로축은 일배체형 상의 빈(bin)들을 나타낸다. 수직 막대는, 히스토그램과 유사하게, 빈들에서의 판독물들의 커버리지를 나타낸다. 일부 구현예들에 따라 일배체형들의 빈들에 할당된 판독물들에 대해 Q 점수가 결정된다. Q 점수가 11보다 높은 판독물들은 각각의 막대의 하단에 반영되고, Q 점수가 11 이하인 판독물들은 각각의 막대의 상단에 반영된다. 상단 패널의 짧은 일배체형에 정렬된 98%의 판독물들은 Q 점수가 11 초과이다. 하단 패널의 긴 일배체형에 정렬된 97%의 판독물들은 Q 점수가 11 초과이다. 일부 구현예들에 따라 각각의 빈에 대한 커버리지가 결정된다. 커버리지의 분산(variance)이 결정될 수 있으며, 이는 판독 분포의 균일성의 척도를 제공한다. 긴 반복체 일배체형의 평균 커버리지는 26이고 짧은 반복체의 경우는 18이다. 전반적으로, 판독물들은 일배체형들 내에서 및 일배체형들에 걸쳐 비교적 균일하게 분포된다. 이러한 서열 메트릭들 및 파생적 척도를 사용하여, 판독물 정렬의 품질을 검사하고, 후술되는 실시예 1 내지 5의 것들과 같은 서열들 내의 대립유전자들의 유전자형들의 타당성을 추론할 수 있다.

[0174] ***서열 그래프를 사용한 반복 서열 좌위에서의 변이체들의 유전자형분석***

[0175] 도 8은 일부 구현예들에 따른 반복 서열을 포함하는 게놈 좌위를 유전자형분석하는 방법(140)을 예시하는 흐름도를 도시한다. 일부 구현예들은, 판독물 길이보다 더 짧거나 더 긴 반복체들의 크기를 추정할 수 있는 하나 또는 복수의 인접한 TR들을 포함하는 영역들에 대한 표적화된 분석을 위한 방법들을 제공한다. 일부 구현예들에서, 유전자 좌위는 게놈 위치들 및 게놈 위치들에서의 좌위들의 구조를 포함하는 변이체 카탈로그에서 사전정의된다. 도 9, 10 및 11은 일부 구현예들에 따른 3개의 상이한 서열 그래프들을 도시한다.

[0176] 도 12는 일부 구현예들에 따른 2개의 STR 서열들을 포함하는 HTT 좌위에서의 변이체들의 유전자형들을 결정하기 위한 방법의 개략도를 도시한다. 도 12의 패널 (a)는 게놈 좌위들 및 이들의 구조를 좌위 사양으로서 포함하는 변이체 카탈로그의 일부를 예시한다. 예를 들어, 반복체들을 무시하면, 좌위 HTT에서의 서열은 CAGCAACAGCGG (서열 번호 2)이고; 좌위 CNBP에서의 서열은 CAGGCAGACA(서열 번호 3)이다.

[0177] 도 13은 일부 구현예들에 따른, SNV 및 STR을 포함하는 린치 I 좌위에서의 변이체들의 유전자형들을 결정하기 위한 방법의 개략도를 도시한다. 도 13의 박스(162)는 좌위 사양의 일반적인 구조를 도시하고, 박스(163)는 린치 I(MSH2)의 좌위 사양의 특정 예를 도시한다.

[0178] 변이체 카탈로그에서, 좌위 구조는 정규 표현식 신택스의 제한된 서브세트를 사용하여 특정된다. 예를 들어, HD에 연결된 반복 영역은 CAACAG 중단에 의해 분리된 가변 개수의 CAG 및 CCG 반복체들을 보유한다는 것을 나타내는 표현식 (CAG)*CAACAG(CGG)* 또는 서열 번호 2(반복체들을 무시함)에 의해 정의될 수 있고; FRDA 영역에 연결된 영역은 표현식 (A)*(GAA)*에 대응하고; SCA8에 연결된 영역은 (CTA)*(CTG)*에 대응하고; 3개의 인접한 반복체들로 구성된 DM2 반복 영역은 (CAGG)*(CAGA)*(CA)* 또는 서열 번호 3(반복체들을 무시함)에 의해 정의되고; 린치 증후군 I을 유발하는 A 동종중합체에 인접한 *MSH2* SNV는 (A|T)(A)*에 대응한다.

[0179] 추가적으로, 정규 표현식은 국제 순수 응용 화학 연합(International Union of Pure and Applied Chemistry, IUPAC) 표기법을 사용하여 특정될 수 있는 다중-대립유전자 또는 "축퇴" 염기 기호를 포함하도록 허용된다(문헌["Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB)" 1986]).

[0180] 축퇴 코돈 내의 염기에 대응하는 불완전하게 특정된 염기는 본원에서 축퇴 염기로 지칭된다. 축퇴 염기는, 예를 들어, 상이한 염기들이 동일한 위치에서 발생할 수 있는 특정 부류의 불완전한 DNA 반복체를 표현하는 것을 가능하게 한다. 이러한 표기법을 사용하여, 폴리알라닌 반복체들은 표현식 (GCN)*에 의해 인코딩될 수 있고, 폴리글루타민 반복체들은 표현식 (CAR)*에 의해 인코딩될 수 있다.

[0181] 일부 구현예들에서, 게놈 좌위에 포함된 반복 서열은 짧은 탠덤 반복(STR) 서열을 포함한다. 일부 구현예들에서, FTR의 연장은 취약 X 증후군, 근위축성 측삭경화증(ALS), 헌팅턴 병, 프리드라이히 운동실조증, 척수소뇌성 운동실조증, 척수구근 근위축증, 근긴장성 이영양증, 마카도-조셉 병, 또는 치상핵적핵 담창구시상하부 위축증과 연관된다.

[0182] 방법(140)은 데이터베이스로부터 테스트 샘플의 핵산 서열 판독물들을 수집하는 단계를 수반한다. 블록(142)을 참조한다. 일부 구현예들에서, 핵산 서열 판독물은 초기에 참조 게놈에 정렬되었지만, 여기서 상기 방법은 하기에 설명되는 바와 같이 서열 판독물들을 관심있는 게놈 좌위에 재정렬시킨다. 대안적인 구현예들에서, 판독물들은 초기에 참조 게놈에 정렬되지 않고 상기 서열 그래프에 직접 정렬될 수 있다.

[0183] 방법(140)은 하나 이상의 반복 서열들을 포함하는 게놈 좌위에 대한 서열에 서열 판독물들을 정렬시키는 단계를 수반한다. 블록(144)을 참조한다. 게놈 좌위의 서열은 서열 그래프의 데이터 구조를 갖는 시스템 메모리에 저장된 데이터로 표현된다. 상기 서열 그래프는 핵산 서열들을 표현하는 정점들 및 상기 정점들을 연결하는 방향성 에지들을 갖는 방향성 그래프를 포함한다. 정점에서의 핵산 서열은 하나 이상의 핵산 염기들을 포함한다. 상기 서열 그래프는 하나 이상의 자가-루프들을 포함한다. 각각의 자가-루프는 하나 이상의 반복 서열들의 반복 서열을 표현한다. 각각의 반복 서열은 하나 이상의 뉴클레오티드들의 반복 단위의 반복체들을 포함한다.

[0184] 일부 구현예들에서, 서열 판독물들은 초기에 참조 게놈에 정렬되어, 초기에 정렬된 판독물들의 서브세트가 하나 이상의 관심있는 서열을 표현하는 하나 이상의 서열 그래프들에 정렬되기 전에, 판독물들의 게놈 좌표들을 결정하도록 한다. 일부 구현예들에서, 초기에 정렬된 판독물들은, 수십 내지 수천개의 영역들에서의 반복 확장들을 결정하기 위해 서열 그래프들에 정렬된다(각각의 영역은 서열 그래프에 대응함). 구현예의 각각의 호출 동안 서열 그래프들에 재정렬되는 초기에 정렬된 판독물들의 총 개수는 수천 내지 수백만개의 판독물들의 범위일 수 있다.

[0185] 일부 구현예들에서, 관심있는 서열 또는 좌위에 또는 그 근처에 초기에 정렬되는 판독물들은 판독물들의 서브세트로서 선택되고, 이 서브세트는 이어서 서열 그래프에 의해 각각 표현된 반복 서열들에 정렬되며, 상기 서열 그래프는 하나 이상의 반복 서열들을 표현하는 하나 이상의 자가-루프들을 갖는다. 다양한 구현예들에서, 관심있는 서열 또는 좌위로부터의 약 10, 50, 100, 500, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 10,000, 50,000, 100,000개의 염기들 내의 판독물이 관심있는 서열 또는 좌위 근처에 있는 것으로 고려된다. 일부 구현예들에서, 관심있는 좌위로부터의 약 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 또는 10,000개의 염기들 내의 판독물이 관심있는 좌위 근처에 있다. 원시 판독물들 중 일부는 불량한 초기 정렬을 가질 수 있는데, 그 이유는 예를 들어, 이들이 명확하게 정렬하기 어려운 반복 서열들을 포함하기 때문이다. 일부 구현예들에서, 불량한 초기 정렬을 갖지만(예를 들어, 정렬 점수에 의해 측정됨) 관심있는 좌위에 또는 그 근처에 정렬된 판독물과 각각 쌍을 이루는(쌍-말단 판독물 쌍에서) 판독물들이 상기 서열 그래프에 정렬된다. 일부 구현예들에서, 판독물 오정렬에 대한 알려진 핫스팟인 비-표적 영역들에 초기에 정렬된 판독물들이 상기 서열 그래프에 정렬된다.

[0186] 도 9, 10 및 11은 일부 구현예들에 따른 3개의 상이한 서열 그래프들을 도시한다. 도 9는 트리뉴클레오티드 반복 단위 CAG를 갖는 반복 서열을 포함하는 제1 게놈 좌위를 표현하는 제1 서열 그래프(1100)를 도시한다. 제1 서열 그래프(1100)는 2개의 플랭킹 서열들을 각각 표현하는 정점들(1102, 1112)을 포함한다. 제1 서열 그래프는 또한 트리뉴클레오티드 반복 단위 CAG를 포함하는 반복 서열을 표현하는 정점(1106)을 포함한다. 제1 서열 그래프는 정점(1102)(플랭킹 서열)과 정점(1106)(CAG 반복 서열)을 연결하는 방향성 에지(1104)를 포함하며, 이 방향은 정점(1102)으로부터 정점(1106)으로 간다. 에지의 방향은 2개의 핵산 서열들의 상대적 위치를 나타낸다. 제1 서열 그래프는 또한 정점(1102)(플랭킹 서열)과 정점(1106)(CAG 반복 서열)을 연결하는 방향성 에지(1104)를 포함하며, 이 방향은 정점(1102)으로부터 정점(1106)으로 간다. 제1 서열 그래프는 또한 정점(1106)(CAG 반복 서열)과 정점(1112)(플랭킹 서열)을 연결하는 방향성 에지(1110)를 포함하며, 이 방향은 정점(1106)으로부터 정점(1112)으로 간다. 제1 서열 그래프는 또한 자가-루프(1108)를 포함하는데, 이는 반복 서열이 1회 이상 반복되는 반복 단위 CAG(정점(1106)에 도시됨)를 포함한다는 것을 표현한다. 서열 그래프의 시작 정점으로부터 종료 정점으로 가는 경로는 게놈 좌위의 서열을 표현하며, 이는 플랭킹 서열들과 같은 반복 서열들 근처의 뉴클레오티드들을 포함할 수 있다.

[0187] 도 10은 제2 게놈 좌위를 표현하는 제2 서열 그래프(1200)를 도시한다. 제2 서열 그래프(1200)는 2개의 플랭킹 서열들을 각각 표현하는 정점들(1202, 1224)을 포함한다. 제2 서열 그래프는 또한 트리뉴클레오티드 반복 단위 CAG를 포함하는 반복 서열 및 트리뉴클레오티드 반복 단위 CCG를 포함하는 반복 서열을 각각 표현하는 정점(1206) 및 정점(1216)을 각각 포함한다. 제2 서열 그래프는 비-반복 서열 CAACAG를 표현하는 정점(1212)을 추가로 포함한다. 제2 서열 그래프는 방향성 에지들(1204, 1210, 1214, 1220)을 포함한다. 이러한 방향성 에지들은 예시된 바와 같이 정점들(1202, 1206, 1212, 1216, 1224)을 방향적으로 연결한다. 제2 서열 그래프는 또한 자가-루프(1208)를 포함하는데, 이는 반복 서열이 1회 이상 반복되는 반복 단위 CAG(정점(1206)에 도시됨)를 포함한다는 것을 표현한다. 제2 서열 그래프는 또한 자가-루프(1218)를 포함하는데, 이는 반복 서열이 1회 이

상 반복되는 반복 단위 CCG(정점(1216)에 도시됨)를 포함한다는 것을 표현한다.

[0188] 도 11은 제3 게놈 좌위를 표현하는 제3 서열 그래프(1300)를 도시한다. 제3 서열 그래프(1300)는 제2 서열 그래프(1200)와 유사하지만, 2개의 대립유전자 CAC 및 CAT를 표현하는 2개의 대안적인 경로들을 포함한다. 2개의 대립유전자들은 SNV 또는 SNP의 대립유전자들일 수 있다. 방향성 에지(1310), 정점(1312), 및 방향성 에지(1314)는 CAC의 제1 대립유전자를 표현한다. 방향성 에지(1316), 정점(1318), 및 방향성 에지(1320)는 CAT의 제2 대립유전자를 표현한다. 제3 서열 그래프는 정점들(1302, 1306, 1322, 1328)을 포함한, 그 외에는 제2 서열 그래프에서의 요소들과 유사한 요소들을 포함한다. 이것은 또한 반복 서열들인 CAG 반복체들 및 CCG 반복체들을 나타내는 자가-루프들(1308, 1324)을 포함한다. 이것은 방향성 에지들(1304, 1326)을 추가로 포함한다.

[0189] 일부 구현예들에서, 서열 판독물들은 하기와 같이 기술된 기술을 사용하여 서열 그래프에 정렬된다.

[0190] 1. 서열로부터의 kmer을 고려하여, 이러한 kmer가 시작되거나 종료되는 모든 그래프 노드들을 열거할 수 있도록, kmer 인덱스가 전체 그래프 상에 구축된다. 일부 경우에 있어서, kmer은 하나의 노드 상에서 시작하고 다른 노드 상에서 종료될 수 있다.

[0191] 2. 각각의 그래프 히트에 대해, 2개의 하위그래프들을 추출하며, 하나는 kmer의 순방향으로 그리고 하나는 역방향으로 추출한다. 상기 하위그래프들은 나머지 판독물 길이까지 반복 확장들을 펼치며(unroll), 반복체들이 확장되지 않은 것으로 가정하면, 나머지 판독물 길이보다 kmer 히트로부터 더 멀리 떨어진 어떠한 노드도 포함하지 않는다. 상기 절차는 너비 우선 탐색(breadth first search)이고, 하기를 포함하는 데이터 구조를 생성한다:

[0192] -상기 하위그래프에서의 모든 노드 서열들(확장된 반복체들을 포함함)의 연결(concatenation)

[0193] -스미스-워터만 방법(smith-waterman process) 상에서 역추적할 때 서열에서의 오프셋으로부터 노드 id를 획득하는 것이 용이하도록 하는, 노드들에 대한 인덱스

[0194] -각각의 노드 시작 오프셋에 대해, 들어오는 에지들을 갖는 노드들의 말단들의 오프셋들의 서열

[0195] -염기가 노드의 시작에 있는지 또는 노드의 시작에 있지 않은지를 알아내고 선행 노드들의 모든 말단 오프셋들을 열거하는 것이 쉽도록 하는, 각각의 노드에 대한 인덱스.

[0196] 3. 정렬:

[0197] -아핀 갭(affine gap)을 지원한다.

[0198] -상기 정보 및 페널티 매트릭스를 고려하여 서열에 대한 최상의-점수의 정렬(들)을 찾는다.

[0199] 2개의 차이 인터페이스들이 이용 가능하다:

[0200] -최상의 정렬 및 두 번째 최상의 정렬 점수가 보고된다.

[0201] -최상의 정렬들 및 두 번째 최상의 정렬 점수의 전체 어레이.

[0202] 상기 정렬들은 후보 kmer과 정렬된 서열의 시작 사이의 갭에 대해 페널티를 주는 전역 정렬들이다. 일부 구현예들은 컴파일-시간 파라미터들을 수정한다.

[0203] 매트릭스 채우기를 위한 현재의 알고리즘이 2개의 구현예들에서 이용 가능하다:

[0204] - N*M 복잡도를 갖는 순차적 루프들.

[0205] - gcc가 자동으로 인식하고 CPU 상에서 SSE 또는 AVX 벡터 명령어들로 변환하는, 16으로 디폴트되는 고정 길이 파일-시간 파라미터의 고정-크기 루프들의 순차적 루프들.

[0206] 일부 구현예들에서, 상기 하나 이상의 반복 서열들의 반복 서열의 특정 반복 단위는 적어도 하나의 불완전하게 특정된 뉴클레오티드를 포함한다. 일부 구현예들에서, 상기 특정 반복 단위는 축퇴 코돈들을 포함한다.

[0207] 일부 구현예들에서, 상기 하나 이상의 자가-루프들은 2개 이상의 반복 서열들을 표현하는 2개 이상의 자가-루프들을 포함한다. 예를 들어, 도 10, 도 11, 및 도 12의 패널 (b)를 참조한다.

[0208] 일부 구현예들에서, 상기 서열 그래프는 2개 이상의 대립유전자들에 대한 2개 이상의 대안적인 경로들을 추가로 포함한다. 예를 들어, 도 11의 도면 부호들(1312, 1318)을 참조한다. 또한 도 13의 좌위 린치 I(MSH2)에 대한 도면 부호들(165, 167a)을 참조하며, 여기서 상부 경로는 핵산 염기 A에 대한 정점을 포함하고, 하부 경로는 핵

산 염기 T에 대한 정점을 포함한다.

[0209] 일부 구현예들에서, 상기 2개 이상의 대립유전자들은 삽입-결실 또는 치환을 포함한다. 일부 구현예들에서, 상기 치환은 단일 뉴클레오티드 변이체(SNV) 또는 단일 뉴클레오티드 다형성(SNP)을 포함한다. 예를 들어, 도 11의 도면 부호들(1312, 1318)을 참조한다.

[0210] 일부 구현예들에서, 상기 서열 그래프에 서열 판독물을 정렬시키는 단계는 하기를 포함한다: 서열 판독물과 상기 서열 그래프의 경로 사이의 kmer 매치를 찾고, 이어서 이 경로를 전체 정렬로 연장시키는 단계. 일부 구현예들에서, 상기 정렬은 하기를 포함한다: 경로 주위에서 하위그래프를 추출하는 단계; 상기 하위그래프 내의 임의의 루프를 펼쳐 방향성 비순환 그래프를 획득하는 단계; 및 상기 방향성 비순환 그래프에 대한 서열 판독물의 스미스-워터만 정렬을 수행하는 단계.

[0211] 일부 구현예들에서, 상기 서열 그래프에 서열 판독물을 정렬시키는 단계는 상기 정렬들의 낮은 신뢰도 말단들을 제거하는 것에 의한 그래프 축소를 포함한다. 판독물이 그래프에 정렬된 후, 상기 방법은 다른 유사한 대안적인 정렬들에 대해 검색한다. 이것은, 원래의 정렬의 경로와 중첩되는 그래프를 통해 경로들에 원래의 판독물을 재정렬시킴으로써 행해진다. 이것은 초기 정렬의 일 말단 또는 양 말단이 낮은 신뢰도를 갖는지를 검출하는 것을 가능하게 하는데, 이러한 신뢰도는 이들이 상이한 방식으로 정렬될 수 있었음을 나타낸다. 정렬의 높은 신뢰도 부분 및 낮은 신뢰도 부분을 검출할 수 있음으로써, 판독물이 어느 유전적 변이체들을 지원하는지를 정확하게 결정할 수 있게 한다.

[0212] 일부 구현예들에서, 상기 서열 그래프에 서열 판독물을 정렬시키는 단계는, 서열 그래프에 판독물의 하위서열들을 정렬시키고; 상기 하위서열들의 정렬들을 병합하여 서열 판독물의 전체 정렬을 형성하는 것에 의한, 정렬 병합을 포함한다.

[0213] 일부 구현예들에서, 상기 방법은 또한 게놈 좌위의 좌위 구조를 포함하는 좌위 사양에 기초하여 상기 서열 그래프를 생성하는 단계를 수반한다. 일부 구현예들에서, 상기 좌위 사양은 상기에 설명된 바와 같이 변이체 카탈로그에 정의된다.

[0214] HTT 좌위에 대한 서열 그래프에 대한 판독물들의 정렬의 개략도에 대해서는 도 12의 패널 (b) 내지 (d)를 또한 참조한다. 도 13의 참조는 좌위 린치 I에 대한 것(165)과 같은, 서열 그래프에 대한 판독물들의 정렬을 수행하기 위한 좌위 분석기(164)를 개략적으로 예시한다.

[0215] 방법(140)은 상기 서열 그래프에 정렬된 서열 판독물들을 사용하여 상기 하나 이상의 반복 서열들에 대한 하나 이상의 유전자형들을 결정하는 단계를 추가로 수반한다. 블록(140)을 참조한다. HTT 좌위에서의 2개의 STR들(CAG 및 CCG)을 결정하는 것을 예시하는 도 12의 패널 (e)를 또한 참조한다. CAG의 반복체들을 포함하는 좌측의 서열은 CAGCAGCAGCAGCAG(서열 번호 4)이다. CCG의 반복체들을 포함하는 좌측의 서열은 CCGCCGCCGCCGCCG(서열 번호 5)이다.

[0216] 도 13은 A/T 대립유전자들(169a)을 갖는 SNV 및 A 단량체 반복체(169b)를 포함하는 린치 I 좌위에서의 변이체들을 결정하기 위한 변이체 유전자형분석기 모듈(168)을 예시한다. 도 13은 또한 서열 정렬 데이터를 큐레이팅하고 이를 변이체 유전자형분석기(168)에 제공하기 위한 변이체 분석기 모듈(166), 및 A/T 대립유전자들(167a)을 갖는 SNV 및 A 단량체 반복체(167b)에 대한 변이체 분석기의 구현예들을 예시한다. 유전자형분석기로부터의 좌위 결과는 도 13의 박스(170)에서, 구체적으로 A/T 대립유전자들(171a)을 갖는 SNV 및 A 단량체 반복체(171b)의 유전자형으로서 도시된다.

[0217] 일부 구현예들에서, 상기 서열 그래프는 2개의 대립유전자들에 대한 2개의 대안적인 경로들을 포함하고, 상기 방법은 상기 2개 이상의 대안적인 경로들에 정렬된 서열 판독물들을 사용하여 상기 2개 이상의 대립유전자들을 유전자형분석하는 단계를 추가로 수반한다. 일부 구현예들에서, 상기 2개 이상의 대립유전자들을 유전자형분석하는 단계는, 상기 2개 이상의 대립유전자들의 확률을 결정하기 위해 확률 모델에 상기 2개 이상의 대안적인 경로들의 커버리지를 제공하는 단계를 수반한다. 일부 구현예들에서, 상기 확률 모델은 대립유전자의 커버리지의 함수로서 대립유전자의 확률을 시뮬레이션하며, 상기 함수는 푸아송 분포, 음-이항 분포(negative-binomial distribution), 이항 분포, 또는 베타-이항 분포로부터 선택된다.

[0218] 일부 구현예들에서, 상기 확률 함수는 푸아송 분포이고, 이의 레이트 파라미터는 게놈 좌위에서 관찰된 판독물 길이 및 평균 깊이로부터 추정된다.

[0219] 푸아송-기반 모델에서, 대립유전자의 확률은 하기와 같이 표현된다.

[0220] $P(Y=y) = (C^{y} \times e^{-C})/y!$

[0221] · y는 염기의 판독물 커버리지이고

[0222] · C는 게놈 좌위에서의 평균 깊이이다.

[0223] 일부 구현예들에서, 평균 깊이(C)는 하기와 같이 추정된다.

[0224] $C = LN / G$

[0225] · G는 게놈 좌위의 길이이고

[0226] · L은 판독물 길이이고

[0227] · N은 모든 판독물들의 개수이다.

[0228] ***그래프툴 라이브러리***

[0229] 일부 구현예들에서, 염기 서열 그래프 기능은 그래프툴 라이브러리를 적용한다. 상기 라이브러리는 코어 그래프 추출(그래프 자체, 그래프 경로들, 및 그래프 정렬들), 이들에 대한 동작들, 및 그래프들에 선형 서열들을 정렬시키기 위한 알고리즘들을 구현한다.

[0230] 일부 구현예들에서, 서열 그래프는 노드들 및 방향성 에지들로 구성된다. 상기 그래프들은 자가-루프들(노드를 그 자체에 연결하는 에지)을 포함하도록 허용되지만, 다른 사이클은 포함하지 않도록 한다. 노드들은 코어 염기들 및 IUPAC 축퇴 염기 코드들로 구성된 서열들을 포함한다.

[0231] 그래프 경로는, 제1 노드 상의 경로의 시작 위치 및 마지막 노드 상의 종료 위치와 함께 경로가 통과하는 노드들의 서열에 의해 정의된다. 위치들은 제로-기반 반-개방 좌표계를 사용하여 특정된다. 상기 라이브러리는 경로 연장 및 축소, 중첩 검사, 및 경로 병합을 포함하는 경로들에 대한 복수의 동작들을 정의한다.

[0232] 그래프 정렬들은 선형 질의 서열들(통상적으로 서열분석되는 판독물들)이 상기 그래프들에 정렬되는 방식을 인코딩한다. 일부 구현예들에서, 그래프 정렬은 그래프 경로 및 상기 그래프 경로의 노드들에 대한 상기 질의 서열의 정렬을 정의하는 선형 정렬의 서열을 포함한다. 경로들에 대한 대응하는 동작들을 사용하여, 그래프 정렬들은 축소되거나 다른 그래프 정렬들과 병합될 수 있다. 경로 축소는 정렬들의 낮은 신뢰도 말단들을 제거하기 위한 메커니즘을 제공하는 반면, 정렬 병합은 하위서열들(예를 들어, kmer들)의 정렬들로부터 상기 질의 서열의 전체 정렬을 함께 스티칭하기 위한 그래프 정렬 알고리즘들에 의해 사용된다. 일부 구현예들에서, 정렬 알고리즘은, 상기 질의 서열과 상기 그래프 사이의 kmer 매치를 찾은 다음 이러한 매치를 전체 정렬로 연장시킴으로써 동작한다. 일부 구현예들에서, 상기 정렬은 kmer 매치에 대응하는 경로 주위에서 하위그래프를 추출하는 것(해당 과정에서 임의의 루프들을 펼치는 것)을 포함한다. 그런 다음, 이것은 생성된 방향성 비순환 그래프에 대해 스미스-워터만 정렬을 수행한다. 일부 구현예들에서, 상기 알고리즘은 아핀 갭 페널티를 지원하고, 컴파일러들이 SIMD 코드를 생성할 수 있게 하기 위해 일정한-길이의 루프들을 사용하여 작성된다.

[0233] 일부 구현예들에서, 그래프 경로는 검색 알고리즘으로 획득될 수 있는데, 이는 상기 정렬이 검색 기준 또는 수렴에 도달할 때까지(예를 들어, 정렬 점수가 최대화됨) 자가-루프에 의해 표현된 반복 단위의 반복체들의 개수를 증가 또는 감소시킴으로써 경로를 연장시키거나 축소하는 것을 수반한다.

[0234] 일부 구현예들에서, 복수의 그래프 경로들이 서열 그래프로부터 생성되고, 각각의 그래프 경로는 자가-루프에 의해 표현된 반복 단위의 반복체들의 특정 개수를 표현한다. 질의 서열이 상기 복수의 그래프 경로들에 정렬되고, 이어서 정렬 기준을 충족시키는 경로가 그래프 정렬을 위해 선택된다.

[0235] ***응용 아키텍처***

[0236] 일부 구현예는 표적화된 변이체 유전자형분석을 위한 일반적인 도구로서 설계된다(도 13). 각각의 런 동안, 프로그램은 변이체 카탈로그 파일에 기술된 변이체들의 세트를 유전자형분석하려고 시도한다.

[0237] 서로 매우 근접하게 위치된 변이체들은 동일한 좌위로 그룹화된다. 좌위 구조는 정규 표현식(RE) 신택스의 제한된 서브세트를 사용하여 특정된다. RE는 코어 염기 기호들 및 IUPAC 축퇴 염기 코드들로 구성된 알파벳들에 걸친 서열들을 포함하며, 가능하게는 서열 중단들에 의해 분리된 표현식들 (<sequence>)?, (<sequence a>|<sequence b>), (<sequence>)*, (<sequence>)+ 중 하나 이상을 포함해야 한다. 이들 표현식은 삽입/결실, 치환, 0회 이상 반복되는 서열, 및 적어도 1회 반복되는 서열에 각각 대응한다. 추가적으로, 각각의 좌위의 설

명은 그 좌위에 대한 참조 영역들의 세트 및 각각의 구성 변이체의 참조 좌표들을 포함한다.

[0238] 대부분의 작업은 초기화 동안 대응하는 RE로부터의 좌위를 표현하는 서열 그래프를 합성하는 좌위분석기(LocusAnalyzer) 클래스의 객체들에 의해 조직된다. 초기화 후, 좌위 분석기는, 상기 그래프에 관련 판독물들을 정렬시키고 이어서 생성된 정렬들을 좌위 내에 포함된 각각의 변이체에 대해 정의되는 변이체분석기(VariantAnalyzer)에 전달함으로써, 관련 판독물들을 처리한다. 변이체분석기는 연관된 변이체를 유전자형분석하기 위한 관련된 정보를 추출하고, 실제 유전자형분석을 수행하는 유전자형분석기(Genotyper)에 이를 전달한다. 이어서, 각각의 유전자형분석기에 의해 출력된 결과는 출력 VCF 파일을 생성하는 데 사용된다.

[0239] 예를 들어, 린치 I 증후군과 연관된 병원성 변이체를 갖는 좌위를 프로세싱하는 것을 담당하는 좌위분석기는 SNV 분석기 및 STR 분석기를 이용한다(도 S1, 우측 패널).

[0240] ***삽입-결실 유전자형분석기***

[0241] 일부 STR들은 인근에 작은 삽입 또는 결실(삽입-결실)을 가질 수 있다. 이러한 삽입-결실은 STR의 플랭킹 서열에서의 추가적인 하위그래프로서 모델링된다. 각각의 대립유전자(또는 그래프 경로)에 맵핑되는 판독물들의 개수는, 좌위에서 관찰된 평균 깊이 및 판독물 길이로부터 레이트 파라미터가 추정되는, 푸아송 분포로 모델링된다. 유전자형 가능성은 베이지안 프레임워크(Bayesian framework) 하에서 계산된다.

[0242] ***반복 확장들의 식별***

[0243] 본원에 개시된 실시형태들을 사용하여, 통상적인 방법들에 비해 높은 효율, 감도, 및/또는 선택성으로 반복 확장과 관련된 다양한 유전적 병태들을 결정할 수 있다. 본 발명의 일부 실시형태들은, 반복 서열을 완전히 횡단하지 않는 서열 판독물들을 사용하여 취약 X 증후군에서 정신 지체를 유발하는 CGG 반복 확장과 같은 의학적으로 관련된 반복 확장들을 식별하고 호출하기 위한 방법들을 제공한다. 100 bp 판독물들과 같은 짧은 판독물들은 많은 반복 확장들을 통해 서열분석할 만큼 충분히 길지 않다. 그러나, 개시된 방법들로 분석될 때, 반복 확장을 갖는 샘플들은 다수의 반복 서열을 포함하는 통계학적으로 유의한 과량의 판독물들을 나타낸다. 추가적으로, 극도로 큰 반복 확장들은 비정렬된 판독물 쌍들을 포함하며, 여기서 판독물 둘 모두는 전적으로 또는 거의 전적으로 반복 서열로 구성된다. 정상 샘플들은 배경 기대치들을 식별하는 데 사용된다.

[0244] 통상적인 믿음은, 전체 반복체에 걸쳐 있는 판독물들 없이는 반복 확장이 검출될 수 없다는 것이다. 반복 확장들을 검출하기 위한 이전의 접근법들은 긴 판독물들을 이용한 표적화된 서열분석을 사용하며, 이것은 일부 경우에 있어서, 반복 서열에 걸쳐 있을 만큼 충분히 길지 않은 판독물들로 인해 성공적이지 못하였다. 일부 개시된 실시형태들의 결과는 놀랍게도 충족되었는데, 그 이유는 부분적으로, 이들이 정상(비-표적화된) 서열 데이터 및 단지 약 100 bp의 판독물 길이를 사용하지만, 반복 확장들을 검출하기 위한 매우 높은 감도를 야기하기 때문이다. 본원에 기재된 방법들은 전체 반복 서열의 길이보다 짧은 삽입물 길이(즉, 2개의 서열 판독물들 및 개재 서열)를 갖는 쌍형성된 판독물들을 사용하여 반복 확장 내의 반복 단위들의 개수를 검출할 수 있다.

[0245] 일부 실시형태들에 따른 반복 확장의 존재를 결정하기 위한 방법들의 상세 사항을 참조하면, 도 14는 샘플 내의 반복 서열의 반복 확장의 존재 또는 부재를 결정하기 위한 실시형태들의 고수준 묘사를 제공하는 흐름도를 도시한다. 상기 반복 서열은 반복 단위로 지칭되는 짧은 서열의 반복적인 출현을 포함하는 핵산 서열이다. 상기 표 1은 반복 단위들, 정상 및 병원성 서열들에 대한 반복 서열들 내의 반복 단위들의 반복체들의 개수, 반복 서열들과 연관된 유전자들, 및 반복 확장과 연관된 질병들의 예를 제공한다. 도 14의 방법(200)은 테스트 샘플의 쌍형성된 말단 판독물들을 획득함으로써 시작된다. 블록(202)을 참조한다. 쌍형성된 말단 판독물들은 관심있는 반복 서열을 포함하는 참조 서열에 정렬되도록 프로세싱되었다. 일부 문맥에서, 상기 정렬 방법은 맵핑 방법으로도 지칭된다. 테스트 샘플은 핵산을 포함하며, 하기의 샘플 섹션에 추가로 기술된 바와 같은, 체액, 조직 등의 형태일 수 있다. 서열 판독물들은 참조 서열에 맵핑되기 위한 정렬 과정을 거쳤다. 다양한 정렬 도구들 및 알고리즘들이, 본 개시내용의 다른 곳에 기술된 바와 같이 참조 서열에 판독물들을 정렬시키려고 시도하는 데 사용될 수 있다. 평소와 같이, 정렬 알고리즘들에서, 일부 판독물들은 참조 서열에 성공적으로 정렬되지만, 다른 것들은 참조 서열에 성공적으로 정렬되지 않을 수 있거나, 또는 불량하게 정렬될 수 있다. 참조 서열에 연속적으로 정렬된 판독물들은 참조 서열 상의 부위들과 연관된다. 정렬된 판독물 및 이들의 연관된 부위들은 서열 태그로도 지칭된다. 상기에 설명된 바와 같이, 다수의 반복체들을 포함하는 일부 서열 판독물들은 참조 서열에 정렬시키기에 더 어려운 경향이 있다. 판독물이 특정 기준보다 많은 개수의 미스매칭된 염기들을 가지면서 참조 서열에 정렬되는 경우, 판독물은 불량하게 정렬된 것으로 간주된다. 다양한 실시형태들에서, 판독물들은 이들이 적어도 약 1, 2, 3, 4, 5, 6, 7, 8, 9, 또는 10개의 미스매치들을 가지면서 정렬되는 경우 불량

하게 정렬된 것으로 간주된다. 다른 실시형태들에서, 판독물들은 이들이 적어도 약 5%의 미스매치들을 가지면서 정렬되는 경우 불량하게 정렬된 것으로 간주된다. 다른 실시형태들에서, 판독물들은 이들이 적어도 약 10%, 15%, 또는 20%의 미스매칭된 염기들을 가지면서 정렬되는 경우 불량하게 정렬된 것으로 간주된다.

[0246] 도 2에 예시된 바와 같이, 방법(200)은 쌍형성된 말단 판독물들에서 앵커 판독물들 및 앵커링된 판독물들을 식별하도록 진행한다. 블록(204)을 참조한다. 앵커 판독물들은 관심있는 반복 서열에 또는 그 근처에 정렬되는 쌍형성된 말단 판독물들 중의 판독물들이다. 예를 들어, 앵커 판독물은 삽입물의 서열 길이보다 작은 서열 길이만큼 반복 서열로부터 분리되는 참조 서열 상의 위치에 정렬될 수 있다. 분리 길이는 더 짧을 수 있다. 예를 들어, 앵커 판독물은 앵커 판독물의 서열 길이보다 작은, 또는 앵커 판독물 및 앵커 판독물을 앵커링된 판독물에 연결하는 서열의 조합된 서열 길이(즉, 삽입물의 길이에서 앵커링된 판독물의 길이를 뺀 값)보다 작은 서열 길이만큼 반복 서열로부터 분리되는 참조 서열 상의 위치에 정렬될 수 있다. 일부 실시형태들에서, 관심있는 반복 서열은 반복 단위 CGG의 반복체들을 포함하는 *FMR1* 유전자 내의 반복 서열일 수 있다. 정상적인 참조 서열에서, *FMR1* 유전자 내의 반복 서열은 반복 단위 CGG의 약 6 내지 32개의 반복체들을 포함한다. 반복체들이 200개의 복제물들을 초과하도록 확장됨에 따라, 반복 확장은 병원성이 되어, 취약 X 증후군을 유발하는 경향이 있다. 일부 실시형태들에서, 판독물들은 관심있는 반복 서열의 1000 bp 내에서 정렬되는 경우 관심있는 서열 근처에 정렬된 것으로 간주된다. 다른 실시형태들에서, 이 파라미터는 약 100 bp, 200 bp, 300 bp, 400 bp, 500 bp, 600 bp, 700 bp, 800 bp, 900 bp, 1500 bp, 2000 bp, 3000 bp, 5000 bp 등 내에서 조정될 수 있다. 또한, 상기 방법은 앵커링된 판독물들을 식별하며, 이러한 판독물들은 앵커 판독물들과 쌍을 이루지만 참조 서열에 불량하게 정렬되거나 정렬될 수 없는 판독물들이다. 불량하게 정렬된 판독물들의 추가적인 상세 사항은 전술되어 있다.

[0247] 방법(200)은 식별된 앵커링된 판독물들에 적어도 부분적으로 기초하여 반복 서열의 반복 확장이 테스트 샘플에 존재할 가능성이 있는지 여부를 결정하는 단계를 추가로 수반한다. 블록(206)을 참조한다. 이러한 결정 단계는 하기에 추가로 기술되는 바와 같은 다양한 적합한 분석들 및 계산들을 수반할 수 있다. 일부 실시형태들에서, 상기 방법은 반복 확장이 존재할 가능성이 있는지 여부를 결정하기 위해, 식별된 앵커 판독물들뿐만 아니라 앵커링된 판독물들을 사용한다. 일부 실시형태들에서, 식별된 앵커 판독물들 및 앵커링된 판독물들 내의 반복체들의 개수가 분석되고, 영향받은 대조군 샘플들의 경험적 데이터로부터 도출되거나 이론적으로 도출된 하나 이상의 기준과 비교된다.

[0248] 본원에 기술된 다양한 실시형태들에서, 반복체들은 프레임-내 반복체들로서 획득되며, 여기서 동일한 반복 단위의 2개의 반복체들은 동일한 판독 프레임에 속한다. 판독 프레임은 핵산(DNA 또는 RNA) 분자 내의 뉴클레오티드들의 서열을 연속적인 비-중첩 트리플릿들의 세트로 분할하는 방식이다. 변환 동안, 트리플릿은 아미노산을 인코딩하며, 코돈으로 지칭된다. 따라서, 임의의 특정 서열은 3개의 가능한 판독 프레임들을 갖는다. 일부 실시형태들에서, 반복체들은 3개의 상이한 판독 프레임들에 따라 카운팅되고, 3개의 카운트들 중 가장 큰 카운트는 판독물에 대한 대응하는 반복체들의 개수인 것으로 결정된다.

[0249] 추가적인 동작 및 분석들을 수반하는 방법의 예가 도 3에 예시되어 있다. 도 15는 다수의 반복체들을 갖는 쌍형성된 말단 판독물들을 사용하여 반복 확장을 검출하기 위한 방법(300)를 예시하는 흐름도를 도시한다. 방법(300)은 테스트 샘플을 처리하기 위한 추가적인 상류측 동작들을 포함한다. 상기 방법은 핵산들을 포함하는 테스트 샘플을 서열분석하여 쌍형성된 말단 판독물들을 획득함으로써 시작된다. 블록(302)을 참조한다. 일부 실시형태들에서, 테스트 샘플은 하기 샘플 섹션에 추가로 기술된 바와 같이 다양한 방식으로 획득되고 제조될 수 있다. 예를 들어, 테스트 샘플은 생물학적 유체, 예를 들어, 혈장, 또는 하기에 기술된 바와 같은 임의의 적합한 샘플일 수 있다. 샘플은 간단한 혈액 채취와 같은 비-침습성 절차를 사용하여 획득될 수 있다. 일부 실시형태들에서, 테스트 샘플은 핵산 분자들, 예를 들어, cfDNA 분자들의 혼합물을 포함한다. 일부 실시형태들에서, 테스트 샘플은 태아 및 모체 cfDNA 분자들의 혼합물을 포함하는 모체 혈장 샘플이다.

[0250] 서열분석 전에, 핵산들이 샘플로부터 추출된다. 적합한 추출 방법들 및 장치가 본원의 다른 곳에 기술되어 있다. 일부 구현예들에서, 상기 장치는 다수의 샘플로부터의 DNA를 프로세싱하여 멀티플렉스된 라이브러리들 및 서열 데이터를 제공한다. 일부 실시형태들에서, 상기 장치는 8개 이상의 테스트 샘플들로부터의 DNA를 병렬로 프로세싱한다. 후술되는 바와 같이, 서열분석 시스템은 추출된 DNA를 프로세싱하여 코딩된(예를 들어, 바코딩된) DNA 단편들의 라이브러리를 생성할 수 있다.

[0251] 일부 실시형태들에서, 테스트 샘플 내의 핵산들은, 하기의 서열분석 라이브러리 제조 섹션에 추가로 기술된 바와 같이, 멀티플렉스 또는 싱글플렉스 서열분석을 위한 서열분석 라이브러리들을 제조하도록 추가로 프로세싱될

수 있다. 샘플들이 프로세싱 및 제조된 후, 다양한 방법들에 의해 핵산들의 서열분석이 수행될 수 있다. 일부 실시형태들에서, 다양한 차세대 서열분석 플랫폼들 및 프로토콜들이 이용될 수 있으며, 이는 하기의 서열분석 방법 섹션에서 추가로 기술된다.

[0252] 특정 서열분석 플랫폼 및 프로토콜에 관계없이, 블록(302)에서, 샘플 내에 포함된 핵산들의 적어도 일부가 서열분석되어 수만, 수십만, 또는 수백만 개의 서열 판독물들, 예를 들어, 100 bp 판독물들을 생성한다. 일부 실시형태들에서, 판독물들은 쌍형성된 말단 판독물들을 포함한다. 다른 실시형태들에서, 도 5를 참조하여 후술된 것들과 같이, 쌍형성된 말단 판독물들에 더하여, 수백, 수천, 또는 수만개의 염기들을 포함하는 단일-말단 긴 판독물들이 반복 서열을 결정하는 데 사용될 수 있다. 일부 실시형태들에서, 서열 판독물들은 약 20 bp, 약 25 bp, 약 30 bp, 약 35 bp, 약 36 bp, 약 40 bp, 약 45 bp, 약 50 bp, 약 55 bp, 약 60 bp, 약 65 bp, 약 70 bp, 약 75 bp, 약 80 bp, 약 85 bp, 약 90 bp, 약 95 bp, 약 100 bp, 약 110 bp, 약 120 bp, 약 130, 약 140 bp, 약 150 bp, 약 200 bp, 약 250 bp, 약 300 bp, 약 350 bp, 약 400 bp, 약 450 bp, 또는 약 500 bp를 포함한다. 기술적 진보는 500 bp 초과의 단일-말단 판독물들을 가능하게 하며, 쌍형성된 말단 판독물들이 생성되는 경우 약 1000 bp 초과의 판독물들을 가능하게 할 것으로 예상된다.

[0253] 방법(300)은 반복 서열을 포함하는 참조 서열에 블록(302)으로부터 획득된 쌍형성된 말단 판독물들을 정렬시키도록 진행한다. 블록(304)을 참조한다. 일부 실시형태들에서, 반복 서열은 확장하는 경향이 있다. 일부 실시형태들에서, 반복 확장은 유전적 장애와 연관되는 것으로 알려져 있다. 다른 실시형태들에서, 반복 서열의 반복 확장은 유전적 장애와의 연관을 확립하기 위해 이전에 연구된 바가 없다. 본원에 개시된 방법들은 임의의 연관된 병리에 관계없이 반복 서열 및 반복 확장의 검출을 허용한다. 일부 실시형태들에서, 판독물들은 참조 게놈, 예를 들어, hg18에 정렬된다. 다른 실시형태들에서, 판독물들은 참조 게놈의 일부, 예를 들어, 염색체 또는 염색체 세그먼트에 정렬된다. 참조 게놈에 고유하게 맵핑되는 판독물들은 서열 태그로 알려져 있다. 일 실시형태에서, 적어도 약 3 x 10^6개의 적격의 서열 태그들, 적어도 약 5 x 10^6개의 적격의 서열 태그들, 적어도 약 8 x 10^6개의 적격의 서열 태그들, 적어도 약 10 x 10^6개의 적격의 서열 태그들, 적어도 약 15 x 10^6개의 적격의 서열 태그들, 적어도 약 20 x 10^6개의 적격의 서열 태그들, 적어도 약 30 x 10^6개의 적격의 서열 태그들, 적어도 약 40 x 10^6개의 적격의 서열 태그들, 또는 적어도 약 50 x 10^6개의 적격의 서열 태그들이, 참조 게놈에 고유하게 맵핑되는 판독물들로부터 획득된다.

[0254] 일부 실시형태들에서, 상기 방법은 정렬 전에 서열 판독물들을 필터링할 수 있다. 일부 실시형태들에서, 판독물 필터링은 에러가 있고 저품질인 판독물들을 필터링하기 위해 서열분석기 내에 구현되는 소프트웨어 프로그램에 의해 인에이블되는 품질-필터링 과정이다. 예를 들어, Illumina의 Sequencing Control Software(SCS) 및 Consensus Assessment of Sequence and Variation 소프트웨어 프로그램은 서열분석 반응에 의해 생성된 원시 이미지 데이터를 강도 점수들, 염기 호출들, 품질 점수 정렬들 및 추가적인 포맷들로 변환하여 에러가 있고 저품질인 판독물들을 필터링하여 다운스트림 분석을 위한 생물학적 관련 정보를 제공한다.

[0255] 특정 실시형태들에서, 서열분석 장치에 의해 생성된 판독물들은 전자 포맷으로 제공된다. 정렬은 후술되는 바와 같이 연산 장치를 이용하여 달성된다. 개별 판독물들이 참조 게놈에 비교되며, 이러한 게놈은 판독물들이 참조 게놈에 고유하게 대응하는 부위들을 식별하기 위해 종종 방대하다(수백만개의 염기쌍들). 일부 실시형태들에서, 정렬 절차는 판독물들과 참조 게놈 사이의 제한된 미스매치를 허용한다. 일부 경우에 있어서, 판독물에서의 1개, 2개, 3개, 또는 그 초과의 염기쌍들이 참조 게놈에서의 대응하는 염기쌍들과 미스매칭하는 것이 허용되며, 그래도 맵핑은 여전히 구성된다. 일부 실시형태들에서, 판독물들은, 판독물들이 참조 서열에 대해 1개, 2개, 3개, 또는 4개 이하의 염기쌍들과 정렬되는 경우 정렬된 판독물들로 간주된다. 이에 대응하여, 비정렬된 판독물들은 정렬될 수 없거나 불량하게 정렬된 판독물들이다. 불량하게 정렬된 판독물들은 정렬된 판독물들보다 더 많은 미스매치들을 갖는 판독물들이다. 일부 실시형태들에서, 판독물들은, 판독물들이 참조 서열에 대해 1%, 2%, 3%, 4%, 5%, 또는 10% 이하의 염기쌍들과 정렬되는 경우 정렬된 판독물들로 간주된다.

[0256] 관심있는 반복 서열을 포함하는 참조 서열에 쌍형성된 말단 판독물들을 정렬시킨 후에, 방법(300)은 쌍형성된 말단 판독물들 중에서 앵커 판독물들 및 앵커링된 판독물들을 식별하도록 진행한다. 블록(306)을 참조한다. 전술된 바와 같이, 앵커 판독물들은 반복 서열에 또는 그 근처에 정렬된 쌍형성된 말단 판독물들이다. 일부 실시형태들에서, 앵커 판독물들은 반복 서열의 1 kb 내에 정렬되는 쌍형성된 말단 판독물들이다. 앵커링된 판독물들은 앵커 판독물들과 쌍을 이루지만, 상기에 설명된 바와 같이 참조 서열에 정렬될 수 없거나 이에 불량하게 정렬된다.

[0257] 방법(300)은 식별된 앵커 판독물들 및/또는 앵커링된 판독물들에서의 반복 단위들의 반복체들의 개수를 분석하여 반복 서열의 확장의 존재 또는 부재를 결정한다. 보다 구체적으로, 방법(300)은 판독물들 내의 반복체들의 개수를 사용하여 앵커 판독물들 및/또는 앵커링된 판독물들에서의 고 카운트(high-count) 판독물들의 개수를 획득하는 단계를 수반한다. 고 카운트 판독물들은 임계값보다 더 많은 반복체들을 갖는 판독물들이다. 일부 실시형태들에서, 고 카운트 판독물들은 앵커링된 판독물들로부터만 획득된다. 다른 실시형태들에서, 고 카운트 판독물은 앵커 판독물들 및 앵커링된 판독물들 둘 모두로부터 획득된다. 일부 실시형태들에서, 반복체들의 개수가 판독물에 대한 가능한 반복체들의 최대 개수에 가까운 경우, 상기 판독물은 고 카운트 판독물로 간주된다. 예를 들어, 판독물이 100 bp이고, 고려 중인 반복 단위가 3 bp인 경우, 반복체들의 최대 개수는 33일 것이다. 다시 말하면, 최대값은 쌍형성된 말단 판독물의 길이 및 반복 단위의 길이로부터 계산된다. 구체적으로, 반복체들의 최대 개수는 판독물 길이를 반복 단위의 길이로 나누고, 숫자를 반내림함으로써 획득할 수 있다. 이 예에서, 다양한 구현예들은 적어도 약 28, 29, 30, 31, 32, 또는 33개의 반복체들을 갖는 100 bp 판독물들을 고 카운트 판독물들로서 식별할 수 있다. 반복체들의 개수는 경험적 인자들 및 고려사항에 기초하여 고 카운트 판독물들에 대해 상향 또는 하향 조정될 수 있다. 다양한 실시형태들에서, 고 카운트 판독물들에 대한 임계값은 반복체들의 최대 개수의 적어도 약 80%, 85%, 90%, 또는 95%이다.

[0258] 이어서, 방법(300)은 고 카운트 판독물들의 개수에 기초하여 반복 서열의 반복 확장이 존재할 가능성이 있는지 여부를 결정한다. 블록(310)을 참조한다. 일부 실시형태들에서, 상기 분석은 획득된 고 카운트 판독물들을 호출 기준에 비교하고, 기준을 초과하는 경우 반복 확장이 존재할 가능성이 있는 것으로 결정한다. 일부 실시형태들에서, 상기 호출 기준은 대조군 샘플들의 고 카운트 판독물들의 분포로부터 획득된다. 예를 들어, 정상 반복 서열을 갖거나 갖는 것으로 의심되는 것으로 알려진 복수의 대조군 샘플들이 분석되고, 전술된 것과 동일한 방식으로 대조군 샘플들에 대해 고 카운트 판독물들이 획득된다. 대조군 샘플들에 대한 고 카운트 판독물들의 분포가 획득될 수 있고, 특정 값보다 많은 고 카운트 판독물들을 갖는, 영향받지 않은 샘플의 확률을 추정할 수 있다. 이러한 확률은, 이러한 특정 값에서 호출 기준이 설정된 경우, 감도 및 선택성을 결정할 수 있게 한다. 일부 실시형태들에서, 상기 호출 기준은, 임계값 초과의 고 카운트 판독물들을 갖는 영향받지 않은 샘플의 확률이 5% 미만이 되도록 하는 임계값으로 설정된다. 다시 말하면, p-값은 0.05 보다 작다. 이들 실시형태들에서, 반복체들이 확장됨에 따라, 반복 서열이 더 길어지고, 더 많은 판독물들이 전적으로 반복 서열 내로부터 유래할 수 있고, 샘플에 대해 더 많은 고 카운트 판독물들이 획득될 수 있다. 다양한 대안적인 구현예들에서, 임계값 초과의 고 카운트 판독물들을 갖는 영향받지 않은 샘플의 확률이 약 1%, 0.1%, 0.01%, 0.001%, 0.0001% 등 미만이 되도록, 더 보수적인 호출 기준이 선택될 수 있다. 상기 호출 기준은 다양한 인자들 및 테스트의 감도 또는 선택성을 증가시켜야 하는 필요성에 따라 상향 또는 하향 조정될 수 있음을 이해할 것이다.

[0259] 일부 실시형태들에서, 대조군 샘플들로부터의 고 카운트 판독물들의 개수의 호출 기준을 경험적으로 획득하는 것 대신에 또는 그에 더하여, 반복 확장을 결정하기 위해 호출 기준이 이론적으로 획득될 수 있다. 쌍형성된 말단 판독물의 길이, 반복 확장을 갖는 서열의 길이, 및 서열분석 깊이를 포함하는 다수의 파라미터들이 주어지면, 완전히 반복체 내에 있는 판독물들의 예상 개수를 계산하는 것이 가능하다. 예를 들어, 정렬된 게놈 내의 판독물들 사이의 평균 간격을 계산하기 위해 서열분석 깊이를 사용할 수 있다. 개별 샘플을 30배 깊이로 서열분석한 경우, 서열분석된 총 염기는 게놈의 크기에 깊이를 곱한 것과 동일하다. 인간의 경우, 이것은 약 $3x10^9x30 = 9x10^{10}$에 이를 것이다. 각각의 판독물이 100 bp 길이인 경우, 이 깊이를 달성하는 데 총 $9x10^8$개의 판독물들이 필요하다. 게놈은 이배체이기 때문에, 이들 판독물의 절반이 하나의 염색체/일배체형을 서열분석하고 나머지는 다른 염색체/일배체형을 서열분석한다. 일배체형당 $4.5x10^8$개의 판독물들이 있고, 전체 게놈 크기를 이러한 개수로 나누는 것은 각각의 판독물의 시작 위치들 사이의 평균 간격 즉, $3x10^9/4.5x10^8 = 1$ 판독물마다 평균 6.7 bp를 산출한다. 이러한 숫자를 사용하여, 특정 개체에서 반복 서열의 크기에 기초하여 그 반복 서열 내에 완전히 있을 판독물들의 개수를 추정할 수 있다. 총 반복 서열 크기가 300 bp이면, 그 반복 서열의 처음 200 bp 내에서 시작하는 임의의 판독물은 반복 서열 내에 완전히 있을 것이다(마지막 100 bp 내에서 시작하는 임의의 판독물은 100 bp 판독물 길이를 기준으로 적어도 부분적으로 반복 서열을 벗어날 것이다). 판독물이 매 6.7 bp마다 정렬될 것으로 예상되므로, 200 bp/(6.7 bp/판독물) = 30개의 판독물들이 반복 서열 내에서 완전히 정렬될 것으로 예상된다. 이러한 개수에 대해 가변성이 있을 것이지만, 이는 임의의 확장 크기에 대해 반복 서열 내에 완전히 있을 전체 판독물들을 추정할 수 있게 한다. 이 방법에 따라 계산되는 반복 서열 길이 및 반복 서열 내에 완전히 정렬되는 판독물들의 대응하는 예상 개수가 하기의 실시예 1의 표 2에 제공된다.

[0260] 일부 실시형태들에서, 호출 기준이 판독물들 내의 반복 서열의 첫 번째 관찰과 마지막 관찰 사이의 거리로부터

계산되고, 따라서 반복 서열에서의 돌연변이들 및 서열분석 오차들을 허용한다.

[0261] 일부 구현예들에서, 상기 방법은 검사 샘플이 획득되는 개체가 취약 X 증후군, ALS, 헌팅턴 병, 프리드라이히 운동실조증, 척수소뇌성 운동실조증, 척수구근 근위축증, 근긴장성 이영양증, 마카도-조셉 병, 및 치상핵적핵담창구시상하부 위축증 등과 같은 유전적 장애들에 대한 위험이 증가된 것으로 진단하는 단계를 추가로 포함할 수 있다. 이러한 진단은 반복 확장이 테스트 샘플에 존재할 가능성이 있다는 결정과 반복 확장과 관련된 유전자 및 반복 서열에 기초할 수 있다. 다른 실시형태들에서, 유전적 장애가 알려져 있지 않은 경우, 일부 실시형태들은 비정상적으로 고 카운트인 반복체들을 검출하여 질병의 유전적 원인들을 새로 식별할 수 있다.

[0262] 도 16는 일부 실시형태들에 따른 반복 확장을 검출하기 위한 다른 방법을 도시하는 흐름도이다. 방법(400)은 고 카운트 판독물들 대신에 테스트 샘플의 쌍형성된 말단 판독물들에서의 반복체들의 개수를 사용하여 반복 확장의 존재를 결정한다. 방법(400)은 핵산을 포함하는 테스트 샘플을 서열분석하여 쌍형성된 말단 판독물들을 획득함으로써 시작된다. 방법(300)의 블록(302)과 동등한 블록(402)을 참조한다. 방법(400)은 반복 서열을 포함하는 참조 서열에 쌍형성된 말단 판독물들을 정렬시킴으로써 계속된다. 방법(300)에서의 블록(304)과 동등한 블록(404)을 참조한다. 상기 방법은 쌍형성된 말단 판독물들 내의 앵커 판독물들 및 앵커링된 판독물들을 식별함으로써 진행하며, 여기서 앵커 판독물들은 반복 서열에 또는 그 근처에 정렬된 판독물들이고, 앵커링된 판독물들은 앵커 판독물들과 쌍을 이루는 비정렬된 판독물들이다. 일부 실시형태들에서, 비정렬된 판독물들은 참조 서열에 정렬될 수 없는 판독물들 및 이에 불량하게 정렬되는 판독물들 둘 모두를 포함한다.

[0263] 앵커 판독물들 및 앵커링된 판독물들을 식별한 후에, 방법(400)은 테스트 샘플로부터의 앵커 판독물들 및/또는 앵커링된 판독물들의 반복체들의 개수를 획득한다. 블록(408)을 참조한다. 이어서, 상기 방법은 테스트 샘플로부터 획득된 모든 앵커 판독물들 및/또는 앵커링된 판독물들에 대한 반복체들의 개수의 분포를 획득한다. 일부 실시형태들에서, 앵커 판독물들로부터의 반복체들의 개수만이 분석된다. 다른 실시형태들에서, 앵커링된 판독물들 및 앵커 판독물들 둘 모두의 반복체들이 분석된다. 이어서, 테스트 샘플의 반복체들의 개수의 분포를 하나 이상의 대조군 샘플들의 분포와 비교한다. 블록(410)을 참조한다. 일부 실시형태들에서, 테스트 샘플의 분포가 대조군 샘플들의 분포와 통계적으로 유의하게 상이한 경우, 상기 방법은 반복 서열의 반복 확장이 테스트 샘플에 존재하는 것으로 결정한다. 블록(412)을 참조한다. 방법(400)은 방법(300)에 관하여 전술된 것과 같은 고 카운트 판독물들만을 분석하는 방법과 상이하게, 고 카운트 판독물들뿐만 아니라 저 카운트 판독물들도 포함하는 판독물들에 대한 반복체들의 개수를 분석한다.

[0264] 일부 실시형태들에서, 테스트 샘플의 분포 및 대조군 샘플들의 분포의 비교는 만-휘트니 순위 검정(Mann-Whitney rank test)을 사용하여 두 분포들이 유의하게 상이한지 결정하는 것을 수반한다. 일부 실시형태들에서, 테스트 샘플의 분포가 대조군 샘플들에 비해 더 높은 반복체들의 개수로 치우치고, 만-휘트니 순위 검정에 대한 p-값이 약 0.0001 또는 0.00001 보다 작은 경우, 분석은 테스트 샘플에 반복 확장이 존재할 가능성이 있는 것으로 결정한다. p-값은 시험의 선택성 또는 감도를 개선하기 위해 필요에 따라 조정될 수 있다.

[0265] 도 2 내지 도 4와 관련하여 전술된 반복 확장을 검출하기 위한 방법은 관심있는 반복 서열에 대해 정렬된 판독물들과 쌍을 이루는 비정렬된 판독물들인 앵커링된 판독물들을 사용한다. 이들 방법에 대한 변형들은, 의학적으로 관련될 수 있는 이전에 식별되지 않은 새로운 반복 확장들을 발견하기 위해 둘 모두 거의 전적으로 일부 유형의 반복 서열로 구성된 판독물 쌍에 대한 비정렬된 판독물들을 검색하는 단계를 포함할 수 있다. 이 방법은 반복체들의 정확한 개수를 정량화하지는 않지만, 추가적인 정량화를 위해 플래깅(flag)되어야 하는 극단적인 반복체 확장들 또는 그 이상점(outlier)들을 식별하는 데 강력하다. 더 긴 판독물들과 조합되어, 이 방법은 총 길이가 최대 200 bp 이상인 반복체들을 식별하고 정량화할 수 있다.

[0266] 도 17은 반복 확장들을 식별하기 위해 관심있는 임의의 반복체 서열과 연관되지 않은 비정렬된 판독물들을 사용하는 방법(500)의 흐름도를 예시한다. 방법(500)은 반복 확장을 검출하기 위해 전체 게놈 비정렬된 판독물들을 사용할 수 있다. 상기 방법은 핵산들을 포함하는 테스트 샘플을 서열분석하여 쌍형성된 말단 판독물들을 획득함으로써 시작된다. 블록(502)을 참조한다. 방법(500)은 참조 게놈에 쌍형성된 말단 판독물들을 정렬시킴으로써 진행된다. 블록(504)을 참조한다. 이어서, 상기 방법은 전체 게놈에 대한 비정렬된 판독물들을 식별한다. 비정렬된 판독물들은 참조 서열에 대해 정렬될 수 없거나 또는 이에 불량하게 정렬된 쌍형성된 말단 판독물들을 포함한다. 일부 구현예들에서, 불량하게 정렬된 판독물들은 기준 미만의 정렬 품질 점수 또는 맵핑 점수로 참조 서열에 정렬된 판독물들을 포함한다. 일부 구현예들에서, 불량하게 정렬된 판독물들은 다수의 미스매칭된, 삽입된, 결실된 염기들을 가지면서 정렬된 판독물들을 포함한다. 블록(506)을 참조한다. 이어서, 상기 방법은

비정렬된 판독물들 내의 반복 단위의 반복체들의 개수를 분석하여, 반복 확장이 테스트 샘플에 존재할 가능성이 있는지 여부를 결정한다. 이러한 분석은 임의의 특정 반복 서열에 무관할 수 있다. 상기 분석은 다양한 잠재적인 반복 단위들에 적용될 수 있으며, 테스트 샘플로부터의 상이한 반복 단위들에 대한 반복체들의 개수는 복수의 대조군 샘플들의 것들과 비교될 수 있다. 전술된 테스트 샘플과 대조군 샘플들 사이의 비교 기술들이 이러한 분석에 적용될 수 있다. 상기 비교가 테스트 샘플이 반복 단위의 반복체들의 비정상적으로 높은 개수를 가짐을 나타내는 경우, 테스트 샘플이 관심있는 특정 반복 서열의 반복 확장을 포함하는지 여부를 결정하기 위해 추가적인 분석이 수행될 수 있다. 블록(510)을 참조한다.

[0267] 일부 실시형태들에서, 상기 추가적인 분석은 의학적으로 관련된 반복 확장들을 갖는 긴 반복 서열들에 잠재적으로 걸쳐 있을 수 있는 매우 긴 서열 판독물들을 수반한다. 이러한 추가적인 분석의 판독물들은 쌍형성된 말단 판독물들보다 더 길다. 일부 실시형태들에서, 단일 분자 서열분석 또는 합성 긴-판독물 서열분석을 사용하여 긴 판독물들을 획득한다. 일부 실시형태들에서, 반복 확장과 유전적 장애 사이의 관계가 당업계에 공지되어 있다. 그러나, 다른 실시형태들에서, 반복 확장과 유전적 장애 사이의 관계가 당업계에 확립될 필요는 없다.

[0268] 일부 실시형태들에서, 동작(510)의 비정렬된 판독물들 내의 반복 단위의 반복체들의 개수를 분석하는 단계는 도 3의 동작(308)의 것에 필적하는 고 카운트 분석을 수반한다. 상기 분석은 고 카운트 판독물들의 개수를 획득하는 단계 - 고 카운트 판독물들은 임계값보다 더 많은 반복체들을 갖는 비정렬된 판독물들임 -, 및 테스트 샘플 내의 고 카운트 판독물들의 개수를 호출 기준과 비교하는 단계를 포함한다. 일부 실시형태들에서, 고 카운트 판독물들에 대한 임계값은 반복체들의 최대 개수의 적어도 약 80% 이며, 이 최대값은 반복 단위의 길이에 대한 쌍형성된 말단 판독물들의 길이의 비율로서 계산된다. 일부 실시형태들에서, 고 카운트 판독물들은 또한 비정렬된 판독물들과 쌍을 이루고 임계값보다 더 많은 반복체들을 갖는 판독물들을 포함한다.

[0269] 일부 실시형태들에서, 동작(510)의 추가적인 분석 이전에, 상기 방법은 (a) 비정렬된 판독물들과 쌍을 이루고 참조 게놈 상의 반복 서열에 또는 그 근처에 정렬되는 쌍형성된 말단 판독물들을 식별하는 단계; 및 (b) 동작(510)에 대한 관심있는 특정 반복 서열로서 반복 서열을 제공하는 단계를 추가로 수반한다. 이어서, 관심있는 반복 서열의 추가적인 분석은 도 2 내지 도 4와 연관하여 전술된 방법들 중 임의의 것을 사용할 수 있다.

[0270] ***샘플***

[0271] 반복 확장을 결정하는 데 사용되는 샘플들은 관심있는 하나 이상의 반복 서열들에 대한 반복 확장이 결정될 핵산들을 포함하는 임의의 세포, 유체, 조직, 또는 장기로부터 취해진 샘플들을 포함할 수 있다. 태아(fetus)의 진단을 수반하는 일부 실시형태들에서, 무세포 핵산, 예를 들어, 무세포 DNA(cfDNA)를 모체 체액으로부터 획득하는 것이 유리하다. 무세포 DNA를 포함하는 무세포 핵산은, 혈장, 혈청 및 소변을 포함하지만 이들로 제한되지 않는 생물학적 샘플들로부터 당업계에 공지된 다양한 방법들에 의해 획득될 수 있다(예를 들어, 문헌[Fan et al., Proc Natl Acad Sci 105:16266-16271]; 문헌[Koide et al., Prenatal Diagnosis 25:604-607]; 문헌[Chen et al., Nature Med. 2: 1033-1035]; 문헌[Lo et al., Lancet 350: 485-487]; 문헌[Botezatu et al., Clin Chem. 46: 1078-1084, 2000; and Su et al., J Mol. Diagn. 6: 101-107] 참조).

[0272] 다양한 실시형태들에서, 샘플에 존재하는 핵산들(예를 들어, DNA 또는 RNA)은 특이적으로 또는 비특이적으로 사용 전(예를 들어, 서열분석 라이브러리를 제조하기 전에)에 풍부화될 수 있다. DNA는 하기의 예시적인 실시예에서 핵산들의 예로서 사용된다. 샘플 DNA의 비특이적 풍부화는 cfDNA 서열분석 라이브러리를 제조하기 전에 샘플 DNA의 수준을 증가시키는 데 사용될 수 있는, 샘플의 게놈 DNA 단편들의 전체 게놈 증폭을 지칭한다. 전체 게놈 증폭 방법은 당업계에 공지되어 있다. 축퇴 올리고뉴클레오티드-프라이밍된 PCR(DOP: degenerate oligonucleotide-primed PCR), 프라이머 연장 PCR 기술(PEP: primer extension PCR technique) 및 다중 치환 증폭(MDA: multiple displacement amplification)은 전체 게놈 증폭 방법의 예이다. 일부 실시형태들에서, 샘플은 DNA에 대해 비-풍부화된다.

[0273] 본원에 기술된 방법들이 적용되는 핵산들을 포함하는 샘플은 전형적으로, 전술된 바와 같은 생물학적 샘플("테스트 샘플")을 포함한다. 일부 실시형태들에서, 반복 확장에 대해 스크리닝될 핵산들은 다수의 널리 공지된 방법 중 임의의 것에 의해 정제되거나 단리된다.

[0274] 따라서, 특정 실시형태들에서, 샘플은 정제된 또는 단리된 폴리뉴클레오티드를 포함하거나 이로 본질적으로 이루어지거나, 또는 이는 조직 샘플, 생물학적 유체 샘플, 세포 샘플 등과 같은 샘플들을 포함할 수 있다. 적합한 생물학적 유체 샘플들은 혈액, 혈장, 혈청, 땀, 눈물, 가래, 소변, 가래, 귀 유출액(ear flow), 림프액, 타액, 뇌척수액, 황폐액(ravages), 골수 현탁액, 질 유출액, 자궁경부내 세척액(trans-cervical lavage), 뇌액,

복수, 모유, 호흡관, 장관 및 비뇨생식관의 분비액, 양수, 모유, 및 류코포레시스(leukophoresis) 샘플들을 포함하지만. 이들로 제한되지 않는다. 일부 실시형태들에서, 샘플은 비-침습 시술에 의해 용이하게 획득 가능한 샘플, 예를 들어, 혈액, 혈장, 혈청, 땀, 눈물, 가래, 소변, 가래, 귀 유출액, 타액 또는 대변이다. 특정 실시형태에서, 샘플은 말초 혈액 샘플, 또는 말초 혈액 샘플의 혈장 및/또는 혈청 분획물들이다. 다른 실시형태들에서, 생물학적 샘플은 면봉 또는 스미어, 생검 시료, 또는 세포 배양물이다. 다른 실시형태에서, 샘플은 둘 이상의 생물학적 샘플들의 혼합물이며, 예를 들어, 생물학적 샘플은 생물학적 유체 샘플, 조직 샘플, 및 세포 배양 샘플 중 2개 이상을 포함할 수 있다. 본원에 사용되는 바와 같이, 용어 "혈액", "혈장" 및 "혈청"은 이들의 분획물들 또는 프로세싱된 부분들을 명확히 포함한다. 유사하게, 샘플이 생검, 면봉, 스미어 등으로부터 취해지는 경우, 상기 "샘플"은 생검, 면봉, 스미어 등으로부터 유도된 프로세싱된 분획물들 또는 부분들을 명확히 포함한다.

[0275] 특정 실시형태에서, 샘플들은, 상이한 개체들로부터의 샘플, 동일하거나 상이한 개체들의 상이한 발달 단계로부터의 샘플, 상이한 병든 개체들(예를 들어, 유전적 장애를 갖는 것으로 의심되는 개체들)로부터의 샘플, 정상 개체들, 개체의 질병의 상이한 단계들에서 획득한 샘플, 질병에 대한 상이한 치료들을 받은 개체로부터 획득한 샘플, 상이한 환경적 인자들을 받은 개체들로부터의 샘플, 병리에 대한 소인을 갖는 개체들로부터의 샘플, 감염성 질병 제제에 대한 노출을 갖는 개체들로부터의 샘플을 포함하지만 이들로 제한되지 않는 공급원으로부터 획득될 수 있다.

[0276] 예시적이지만 비제한적인 일 실시형태에서, 샘플은 임신한 여성, 예를 들어 임부로부터 획득된 모체 샘플이다. 이 경우에 있어서, 샘플은 태아의 잠재적인 염색체 이상의 산전 진단을 제공하기 위해 본원에 기술된 방법들을 이용하여 분석될 수 있다. 모체 샘플은 조직 샘플, 생물학적 유체 샘플, 또는 세포 샘플일 수 있다. 생물학적 유체는, 비제한적인 예로서, 혈액, 혈장, 혈청, 땀, 눈물, 가래, 소변, 가래, 귀 유출액, 림프액, 타액, 뇌척수액, 황폐액, 골수 현탁액, 질 유출액, 자궁경부내 세척액, 뇌액, 복수, 모유, 호흡관, 장관 및 비뇨생식관의 분비액, 및 류코포레시스 샘플들을 포함한다.

[0277] 특정 실시형태에서, 샘플들은 또한 시험관내 배양된 조직, 세포, 또는 다른 폴리뉴클레오티드-함유 공급원으로부터 획득될 수 있다. 배양된 샘플들은 상이한 배지 및 조건(예를 들어, pH, 압력, 또는 온도)에 유지된 배양물(예를 들어, 조직 또는 세포), 상이한 기간 동안 유지된 배양물(예를 들어, 조직 또는 세포), 상이한 인자들 또는 시약들(예를 들어, 약물 후보 또는 조절제)로 처리된 배양물(예를 들어, 조직 또는 세포), 또는 다양한 유형의 조직 및/또는 세포의 배양물을 포함하지만 이들로 제한되지 않는 공급원으로부터 취할 수 있다.

[0278] 생물학적 공급원으로부터 핵산들을 단리하는 방법은 널리 공지되어 있으며, 공급원의 성질에 따라 상이할 것이다. 당업자는 본원에 기술된 방법들에 필요한 공급원으로부터 핵산들을 용이하게 단리할 수 있다. 일부 경우에 있어서, 핵산 샘플 내의 핵산 분자들을 단편화하는 것이 유리할 수 있다. 단편화는 무작위적일 수 있거나, 예를 들어 제한 엔도뉴클레아제 분해를 사용하여 달성되는 바와 같이, 특이적일 수 있다. 무작위 단편화를 위한 방법들은 당업계에 널리 공지되어 있으며, 예를 들어, 제한된 DNAse 분해, 알칼리 처리 및 물리적 전단을 포함한다.

[0279] ***서열분석 라이브러리 제조***

[0280] 다양한 실시형태들에서, 서열분석은 서열분석 라이브러리의 제조를 필요로 하는 다양한 서열분석 플랫폼들 상에서 수행될 수 있다. 제조는 전형적으로, DNA를 단편화하고(초음파 분해, 분무 또는 전단), 이어서 DNA 수복 및 말단 폴리싱(무딘 말단 또는 A 오버행), 및 플랫폼-특이적 어댑터 결찰을 수반한다. 일 실시형태에서, 본원에 기술된 방법들은 다수의 샘플들이 개별적으로 게놈 분자들로서 서열분석될 수 있게 하거나(즉, 싱글플렉스 서열분석) 또는 단일 서열분석 런 상의 인덱싱된 게놈 분자들을 포함하는 풀링된 샘플들로서 서열분석될 수 있게 하는(예를 들어, 멀티플렉스 서열분석) 차세대 서열분석 기술(NGS)을 이용할 수 있다. 이들 방법은 최대 수억개의 DNA 서열들의 판독물들을 생성할 수 있다. 다양한 실시형태들에서 게놈 핵산의 서열, 및/또는 인덱싱된 게놈 핵산의 서열은, 예를 들어, 본원에 기술된 차세대 서열분석 기술(NGS)들을 이용하여 결정될 수 있다. 다양한 실시형태들에서, NGS를 이용하여 획득된 대량의 서열 데이터의 분석은 본원에 기술된 바와 같이 하나 이상의 프로세서들을 이용하여 수행될 수 있다.

[0281] 다양한 실시형태들에서, 이러한 서열분석 기술들의 사용은 서열분석 라이브러리들의 제조를 수반하지 않는다.

[0282] 그러나, 특정 실시형태에서, 본원에서 고려된 서열분석 방법은 서열분석 라이브러리의 제조를 수반한다. 일 예시적인 접근법에서, 서열분석 라이브러리 제조는 서열분석될 준비가 된 어댑터-변형된 DNA 단편들(예를 들어,

폴리뉴클레오티드들)의 무작위 집합의 생성을 수반한다. 폴리뉴클레오티드들의 서열분석 라이브러리들은 DNA 또는 RNA, 예를 들어, DNA 또는 cDNA의 동등물, 유사체, 예를 들어, DNA 또는 역전사효소의 작용에 의한 RNA 템플릿으로부터 생성된 상보적인 cDNA 또는 복제 DNA로부터 제조될 수 있다. 폴리뉴클레오티드들은 이중-가닥 형태(예를 들어, dsDNA, 예를 들어, 게놈 DNA 단편들, cDNA, PCR 증폭 생성물 등)로 유래할 수 있거나, 또는 특정 실시형태에서, 폴리뉴클레오티드들은 단일-가닥 형태(예를 들어, ssDNA, RNA 등)로 유래되어 dsDNA 형태로 전환되었을 수 있다. 예시로서, 특정 실시형태에서, 단일-가닥 mRNA 분자들은 서열분석 라이브러리 제조 시 이용에 적합한 이중-가닥 cDNA들로 복제될 수 있다. 1차 폴리뉴클레오티드 분자들의 정확한 서열은 일반적으로 라이브러리 제조의 방법에 중요하지 않고, 알려져 있거나 또는 알려져 있지 않을 수 있다. 일 실시형태에서, 폴리뉴클레오티드 분자들은 DNA 분자들이다. 더욱 구체적으로, 특정 실시형태에서, 폴리뉴클레오티드 분자들은 유기체의 전체 유전적 보체 또는 실질적으로 유기체의 전체 유전적 보체를 나타내며, 일반적으로 인트론 서열 및 엑손 서열(코딩 서열)뿐만 아니라 프로모터 및 인핸서 서열과 같은 비-코딩 조절 서열을 포함하는 게놈 DNA 분자들(예를 들어, 세포 DNA, 무세포 DNA(cfDNA) 등)이다. 특정 실시형태에서, 1차 폴리뉴클레오티드 분자들은 인간 게놈 DNA 분자들, 예를 들어, 임신한 대상체의 말초 혈액 내에 존재하는 cfDNA 분자들을 포함한다.

[0283] 일부 NGS 서열분석 플랫폼들에 대한 서열분석 라이브러리들의 제조는 특정 범위의 단편 크기를 포함하는 폴리뉴클레오티드들의 사용에 의해 용이하게 된다. 이러한 라이브러리들의 제조는 전형적으로, 원하는 크기 범위의 폴리뉴클레오티드들을 획득하기 위해 거대 폴리뉴클레오티드들(예를 들어 세포 게놈 DNA)의 단편화를 수반한다.

[0284] 반복 확장을 결정하기 위한 본원에 개시된 방법들 및 시스템을 위해 쌍형성된 말단 판독물들이 사용된다. 단편 또는 삽입물 길이는 판독물 길이보다 길고, 전형적으로 2개의 판독물의 길이의 합보다 더 길다.

[0285] 일부 예시적인 실시형태에서, 샘플 핵산(들)은 게놈 DNA로서 획득되며, 이는 대략적으로 100개 이상, 대략 200개 이상, 대략 300개 이상, 대략 400개 이상, 또는 대략 500개 이상의 염기쌍들의 단편들로 단편화되고, 여기에는 NGS 방법들이 용이하게 적용될 수 있다. 일부 실시형태들에서, 쌍형성된 말단 판독물들은 약 100 내지 5000 bp의 삽입물들로부터 획득된다. 일부 실시형태들에서, 삽입물들은 길이가 약 100 내지 1000 bp이다. 이들은 때때로 규칙적인 짧은-삽입물 쌍형성된 말단 판독물들로서 구현된다. 일부 실시형태들에서, 삽입물들은 길이가 약 1000 내지 5000 bp이다. 이들은 때때로, 전술된 바와 같이 긴-삽입물 메이트 쌍형성된 판독물들로서 구현된다.

[0286] 일부 구현예들에서, 긴 삽입물들은 매우 길고 확장된 반복 서열들을 평가하도록 설계된다. 일부 구현예들에서, 메이트 쌍 판독물들은 수천개의 염기쌍들에 의해 이격된 판독물들을 획득하기 위해 적용될 수 있다. 이들 구현예에서, 삽입물들 또는 단편들은 수백 내지 수천개의 염기쌍 범위이며, 여기서 2개의 비오틴 접합 어댑터들이 삽입물의 2개의 말단들 상에 있다. 이어서, 비오틴 접합 어댑터들은 삽입물의 2개의 말단들을 결합하여 원형화된 분자를 형성하며, 이는 이어서 추가로 단편화된다. 비오틴 접합 어댑터들 및 원래의 삽입물의 2개의 말단들을 포함하는 하위단편은, 더 짧은 단편들을 서열분석하도록 설계된 플랫폼 상에서의 서열분석을 위해 선택된다.

[0287] 단편화는 통상의 기술자에 공지된 다수의 방법들 중 임의의 것에 의해 달성될 수 있다. 예를 들어, 단편화는 분무화, 초음파처리 및 하이드로시어(hydroshear)를 포함하지만 이들로 제한되지 않는 기계적 수단들에 의해 달성될 수 있다. 그러나 기계적 단편화는 전형적으로 C-O, P-O 및 C-C 결합에서 DNA 백본을 절단하여, C-O, P-O 및/ C-C 결합이 끊어진 3'- 및 5'-오버행 말단들과 무딘 말단들의 불균질한 혼합물을 초래하며(예를 들어, 문헌[Alnemri and Liwack, J Biol. Chem 265:17323-17333]; 문헌[Richards and Boyer, J Mol Biol 11:327-240] 참조), 이는 수복될 필요가 있을 수 있는데, 그 이유는 서열분석을 위한 DNA 제조에 필요한 후속 효소 반응, 예를 들어 서열분석 어댑터들의 결찰에 필요한 5'-포스페이트가 부족할 수 있기 때문이다.

[0288] 대조적으로, cfDNA는 전형적으로, 약 300개 미만의 염기쌍들의 단편들로 존재하고, 결과적으로, 단편화는 전형적으로, cfDNA 샘플들을 이용하여 서열분석 라이브러리를 생성하는 데 필요하지 않다.

[0289] 전형적으로, 폴리뉴클레오티드들이 강제로 단편화되는지(예를 들어, 체외 단편화됨) 또는 자연적으로 단편으로 존재하는지 여부에 관계없이, 이들은 5'-포스페이트 및 3'-하이드록실을 갖는 무딘-말단 DNA로 전환된다. 표준 프로토콜들, 예를 들어, 예를 들어, 본원의 다른 곳에 기술된 바와 같은 일루미나 플랫폼을 사용하는 서열분석을 위한 프로토콜들은, 사용자에게 샘플 DNA를 말단-수복하고, dA-테일링(dA-tailing) 전에 말단-수복된 생성물을 정제하고, 라이브러리 제조의 어댑터-결찰 단계 전에 dA-테일링 생성물을 정제하도록 지시한다.

[0290] 본원에 기술된 서열 라이브러리 제조 방법의 다양한 실시형태들은 NGS에 의해 서열분석될 수 있는 변형된 DNA 생성물을 획득하기 위해 표준 프로토콜들에 의해 전형적으로 명령받는 단계들 중 하나 이상을 수행할 필요성을

제거한다. 약칭된 방법(ABB 방법), 1-단계 방법, 및 2-단계 방법은 2012년 7월 20일자로 출원된 특허 출원 제 13/555,037호에서 찾아볼 수 있는, 서열분석 라이브러리를 제조하기 위한 방법들의 예이며, 이 문헌은 그 전체가 본원에 참고로 포함된다.

[0291] ***서열분석 방법***

[0292] 상기에 나타낸 바와 같이, 제조된 샘플들(예를 들어, 서열분석 라이브러리들)은 복제수 변이(들)를 식별하기 위한 절차의 일부로서 서열분석된다. 다수의 서열분석 기술들 중 임의의 것이 사용될 수 있다.

[0293] 후술되는 바와 같이, Affymetrix Inc.(캘리포니아주 써니베일 소재)의 혼성화에 의한 서열분석 플랫폼, 및 454 Life Sciences(코네티컷주 브래드포드 소재), Illumina/Solexa(캘리포니아주 샌디에고 소재) 및 Helicos Biosciences(매사추세츠주 캠브리지 소재)의 합성에 의한 서열분석 플랫폼, 및 Applied Biosystems(캘리포니아주 포스터 시티 소재)의 결찰에 의한 서열분석 플랫폼과 같은 일부 서열분석 기술들이 상업적으로 이용 가능하다. Helicos Biosciences의 합성에 의한 서열분석을 이용하여 수행된 단일 분자 서열분석 이외에, 다른 단일 분자 서열분석 기술들은 Pacific Biosciences의 SMRT™ 기술, ION TORRENT™ 기술, 및 예를 들어 Oxford Nanopore Technologies에 의해 개발된, 나노기공 서열분석을 포함하지만, 이들로 제한되지 않는다.

[0294] 자동화된 생어(Sanger) 방법은 '제1 세대' 기술로 고려되지만, 자동화된 생어 서열분석을 포함하는 생어 서열분석 또한 본원에 기술된 방법들에 사용될 수 있다. 추가적인 적합한 서열분석 방법은 핵산 이미징 기술, 예를 들어, 원자힘 현미경(AFM: atomic force microscopy) 또는 투과전자 현미경(TEM: transmission electron microscopy)을 포함하지만, 이들로 제한되지 않는다. 예시적인 서열분석 기술들은 하기에 더 상세하게 기술된다.

[0295] 일부 실시형태들에서, 개시된 방법들은 Illumina의 합성에 의한 서열분석 및 가역적 종결자 기반 서열분석 화학반응을 사용하여 수백만개의 DNA 단편들을 대규모로 병렬로 서열분석하여 테스트 샘플 내의 핵산들에 대한 서열 정보를 획득하는 단계를 수반한다(예를 들어, 문헌[Bentley et al., Nature 6:53-59]에 기술된 바와 같음). 템플릿 DNA는 게놈 DNA, 예를 들어, 세포 DNA 또는 cfDNA일 수 있다. 일부 실시형태들에서, 단리된 세포들로부터의 게놈 DNA가 템플릿으로서 사용되고, 수백개의 염기쌍들의 길이들로 단편화된다. 다른 실시형태들에서, cfDNA가 템플릿으로서 사용되고, cfDNA가 짧은 단편들로 존재하는 경우에 단편화는 필요하지 않다. 예를 들어 태아 cfDNA는 길이가 대략 170개의 염기쌍(bp)인 단편들로서 혈류 내에서 순환하고(문헌[Fan et al., Clin Chem 56:1279-1286]), 서열분석 이전에 DNA의 단편화가 필요하지 않다. Illumina의 서열분석 기술은 올리고뉴클레오티드 앵커들이 결합되는 편평하고, 광학적으로 투명한 표면에 단편화된 게놈 DNA를 부착하는 것에 의존한다. 템플릿 DNA는 말단-수복되어 5'-과인산화된 무딘 말단들을 생성하고, 클레노우 단편(Klenow fragment)의 중합효소 활성을 사용하여 단일 A 염기를 무딘 과인산화된 DNA 단편들의 3' 말단에 첨가한다. 이 첨가는 올리고뉴클레오티드 어댑터들에 결찰하기 위한 DNA 단편들을 제조하고, 이들은 결찰 효율을 증가시키기 위해 이들의 3' 말단에 단일 T 염기의 오버행을 갖는다. 어댑터 올리고뉴클레오티드들은 플로우-셀 앵커 올리고들에 상보적이다(반복 확장의 분석에서 앵커/앵커링된 판독물들과 혼동되지 않음). 제한-희석 조건 하에서, 어댑터-변형된, 단일-가닥 템플릿 DNA는 플로우 셀에 추가되고 혼성화에 의해 앵커 올리고들에 고정된다. 부착된 DNA 단편들은 연장되고 브리지 증폭되어 수억개의 클러스터들을 갖는 초고밀도 서열분석 플로우 셀을 생성하며, 각각은 동일한 템플릿의 약 1,000개의 복제물들을 포함한다. 일 실시형태에서, 무작위로 단편화된 게놈 DNA는 클러스터 증폭을 거치기 전에 PCR을 이용하여 증폭된다. 대안적으로, 증폭이 없는 게놈 라이브러리 제조가 사용되고, 무작위로 단편화된 게놈 DNA는 클러스터 증폭만을 사용하여 풍부화된다(문헌[Kozarewa et al., Nature Methods 6:291-295]). 템플릿들은 제거 가능한 형광 염료들을 구비한 가역적 종결자를 사용하는 견고한 4색 DNA 합성에 의한 서열분석 기술을 이용하여 서열분석된다. 고감도 형광 검출은 레이저 여기 및 내부전반사 광학계를 이용하여 달성된다. 약 수십 내지 수백개의 염기쌍들의 짧은 서열 판독물들이 참조 게놈에 정렬되고, 참조 게놈에 대한 짧은 서열 판독물들의 고유한 맵핑은 특별히 개발된 데이터 분석 파이프라인 소프트웨어를 이용하여 식별된다. 제1 판독물의 완료 후에, 템플릿들은 단편들의 반대편 말단으로부터의 제2 판독물을 가능하게 하도록 원위치(in situ)에서 재생성될 수 있다. 따라서, DNA 단편들의 단일-말단 또는 쌍형성된 말단 서열분석이 이용될 수 있다.

[0296] 개시내용의 다양한 실시형태들은 쌍형성된 말단 서열분석을 허용하는 합성에 의한 서열분석을 이용할 수 있다. 일부 실시형태들에서, Illumina에 의한 합성 플랫폼에 의한 서열분석은 단편을 클러스터링하는 것을 수반한다. 클러스터링은 각각의 단편 분자가 등온적으로 증폭되는 방법이다. 일부 실시형태들에서, 여기에 설명된 예로서, 단편은 단편의 두 말단에 부착되는 2개의 상이한 어댑터들을 갖고, 상기 어댑터들은 단편이 플로우 셀

레인(lane)의 표면 상에서 2개의 상이한 올리고들과 혼성화되게 한다. 단편은 추가로 단편의 두 말단에서 2개의 인덱스 서열들을 포함하거나 또는 2개의 인덱스 서열들에 연결되고, 이 인덱스 서열들은 멀티플렉스 서열분석에서 상이한 샘플들을 식별하기 위한 라벨들을 제공한다. 일부 서열분석 플랫폼들에서, 서열분석될 단편은 또한 삽입물로 지칭된다.

[0297] 일부 구현예들에서, 일루미나 플랫폼 내의 클러스터링을 위한 플로우 셀은 레인을 구비한 유리 슬라이드이다. 각각의 레인은 2가지 유형의 올리고들의 론(lawn)으로 코팅된 유리 채널이다. 혼성화는 표면 상의 2가지 유형의 올리고 중 제1 올리고에 의해 가능해진다. 이 올리고는 단편의 일 말단 상의 제1 어댑터에 상보적이다. 중합효소는 혼성화된 단편의 보체 가닥을 생성한다. 이중-가닥 분자는 변성되고, 원래의 템플릿 가닥은 세척 제거된다. 남은 가닥은, 많은 다른 남은 가닥과 병렬로, 브리지 적용을 통해 클론 증폭된다.

[0298] 브리지 증폭에서, 가닥의 제2 말단 상의 제2 어댑터 영역이 플로우 셀 표면 상의 제2 유형의 올리고들과 혼성화된다. 중합효소는 상보적 가닥을 생성하여, 이중-가닥 브리지 분자를 형성한다. 이 이중-가닥 분자는 변성되어, 2개의 상이한 올리고들을 통해 플로우 셀에 테더링되는 2개의 단일-가닥 분자를 생성한다. 이어서, 상기 방법은 계속 반복되고, 수백만개의 클러스터에 대하여 동시에 발생하여 모든 단편의 클론 증폭을 야기한다. 브리지 증폭 후에, 역방향 가닥은 절단되고 세척 제거되어, 순방향 가닥만 남게 된다. 3' 말단들은 차단되어 원치 않는 프라이밍을 방지한다.

[0299] 클러스터링 후에, 제1 서열분석 프라이머를 연장시켜 제1 판독물을 생성함에 따라 서열분석이 시작된다. 각각의 사이클에서, 형광 표지된 뉴클레오티드들은 성장하는 사슬에 첨가되기 위해 경쟁한다. 템플릿의 서열에 기초하여 단지 하나만이 통합된다. 각각의 뉴클레오티드의 첨가 이후에, 클러스터는 광원에 의해 여기되고, 특성 형광 신호가 방출된다. 사이클의 개수는 판독물의 길이를 결정한다. 방출 파장 및 신호 세기는 염기 호출을 결정한다. 특정 클러스터에 대하여, 모든 동일한 가닥이 동시에 판독된다. 수억개의 클러스터가 대규모 병렬 방식으로 서열분석된다. 제1 판독물의 완료 시, 판독물 생성물은 세척 제거된다.

[0300] 2개의 인덱스 프라이머들을 수반하는 프로토콜들의 다음 단계에서, 인덱스 1 프라이머가 도입되고 템플릿 상의 인덱스 1 영역으로 혼성화된다. 인덱스 영역은 단편들의 식별을 제공하고, 이는 멀티플렉스 서열분석 과정에서 샘플들을 디-멀티플렉스(de-multiplex)하는데 유용하다. 인덱스 1 판독물은 제1 판독물과 유사하게 생성된다. 인덱스 1 판독물의 완료 후에, 판독물 생성물은 세척 제거되고 가닥의 3' 말단은 탈보호된다. 이어서 템플릿 가닥은 접히고 플로우 셀 상의 제2 올리고에 결합된다. 인덱스 2 서열은 인덱스 1과 동일한 방식으로 판독된다. 이어서 인덱스 2 판독물 생성물은 단계의 완료 시 세척 제거된다.

[0301] 두 인덱스를 판독한 후에, 중합 효소를 사용하여 제2 플로우 셀 올리고들을 연장시켜 이중-가닥 브리지를 형성함으로써 판독물 2가 개시된다. 이 이중 가닥 DNA은 변성되고, 3' 말단은 차단된다. 원래의 순방향 가닥은 절단되고 세척 제거되어, 역방향 가닥을 남긴다. 판독물 2는 판독물 2 서열분석 프라이머의 도입으로 시작된다. 판독물 1과 같이, 서열분석 단계들이 원하는 길이가 달성될 때까지 반복된다. 판독물 2 생성물이 세척 제거된다. 이 전체 방법은 수백만개의 판독물을 생성하여, 모든 단편을 표현한다. 풀링된 샘플 라이브러리들로부터의 서열들은 샘플 제조 동안 도입된 고유 인덱스들에 기초하여 분리된다. 각각의 샘플에 대하여, 염기 호출들의 유사한 신장부들의 판독물들은 국부적으로 클러스터링된다. 순방향 및 역방향 판독물은 쌍을 이루어 연속된 서열들을 생성한다. 이들 연속된 서열들은 변이체 식별을 위해 참조 게놈에 정렬된다.

[0302] 전술된 합성에 의한 서열분석의 예는 쌍형성된 말단 판독물들을 수반하고, 이는 개시된 방법의 실시형태들 중 다수에 사용된다. 쌍형성된 말단 서열분석은 단편의 두 말단으로부터의 2개의 판독물들을 수반한다. 쌍형성된 말단 판독물들은 모호한 정렬을 해결하기 위해 사용된다. 쌍형성된 말단 서열분석은 사용자가 삽입물(또는 서열분석될 단편)의 길이를 선택하고 삽입물의 양 말단을 서열분석하여, 고품질의 정렬가능 서열 데이터를 생성할 수 있게 한다. 각각의 쌍형성된 판독물 사이의 거리가 알려져 있기 때문에, 정렬 알고리즘들은 이러한 정보를 사용하여 반복 영역들 위에 판독물들을 보다 정밀하게 맵핑할 수 있다. 이는, 특히 게놈의 서열분석하기 어려운 반복 영역들에 걸쳐, 판독물들의 더 양호한 정렬을 야기한다. 쌍형성된 말단 서열분석은 삽입 및 결실(삽입-결실) 및 역위를 포함하는 재배열을 검출할 수 있다.

[0303] 쌍형성된 말단 판독물들은 상이한 길이(즉, 서열분석될 상이한 단편 크기)의 삽입물을 사용할 수 있다. 본 개시내용에서 디폴트 의미로서, 쌍형성된 말단 판독물은, 다양한 삽입물 길이로부터 획득된 판독물을 지칭하는 데 사용된다. 일부 경우에 있어서, 짧은-삽입물 쌍형성된 말단 판독물을 긴-삽입물 쌍형성된 말단 판독물로부터 구별하기 위해, 후자는 구체적으로 메이트 쌍 판독물로 지칭된다. 메이트 쌍 판독물들을 수반하는 일부 실시형태들에서, 두 비오틴 접합 어댑터가 비교적 긴 삽입물(예를 들어, 수 kb)의 두 말단에 먼저 부착된다. 이어서

비오틴 접합 어댑터들은 삽입물의 두 말단을 연결하여 원형화된 분자를 형성한다. 이어서, 원형화된 분자를 추가로 단편화함으로써 비오틴 접합 어댑터들을 포함하는 하위단편들이 획득될 수 있다. 이어서, 반대 서열 순서의 원래 단편의 두 말단을 포함하는 하위단편은 전술된 짧은-삽입물 쌍형성된 말단 서열분석과 동일한 절차에 의해 서열분석될 수 있다. 일루미나 플랫폼을 사용한 메이트 쌍 서열분석의 추가적인 상세 사항은 하기의 주소에 있는 온라인 간행물에 나타나 있으며, 이는 그 전체가 본원에 참고로 포함된다: res.illumina.com/documents/products/technotes/technote_nextera_matepair_data_processing.pdf

[0304] DNA 단편의 서열분석 후에, 사전결정된 길이, 예를 들어, 100 bp의 서열 판독물들이 공지된 참조 게놈에 맵핑 또는 정렬된다. 맵핑된 또는 정렬된 판독물 및 이들의 참조 서열 상의 대응하는 위치는 또한 태그로 지칭된다. 반복 확장을 결정하기 위한 본원에 개시된 많은 실시형태들의 분석들은 불량하게 정렬되거나 정렬될 수 없는 판독물뿐만 아니라 정렬된 판독물들(태그들)도 이용한다. 일 실시형태에서, 참조 게놈 서열은 NCBI36/hg18 서열이며, 이는 월드와이드웹 상에서 genome.ucsc.edu/cgi-bin/hgGateway?org=Human&db=hg18&hgsid=166260105)에서 이용 가능하다. 대안적으로, 참조 게놈 서열은 GRCh37/hg19이며, 이는 월드와이드웹 상에서 genome.ucsc.edu/cgi-bin/hgGateway에서 이용 가능하다. 공개 서열 정보의 다른 공급원은 GenBank, dbEST, dbSTS, EMBL(유럽 분자 생물학 실험실(the European Molecular Biology Laboratory)), 및 DDBJ(일본 DNA 데이터뱅크(DNA Databank of Japan))를 포함한다. 다수의 컴퓨터 알고리즘들이 서열을 정렬하는 데 이용 가능하며, 이는 BLAST(문헌[Altschul et al., 1990]), BLITZ(MPsrch)(문헌[Sturrock & Collins, 1993]), FASTA(문헌[Person & Lipman, 1988]), BOWTIE(문헌[Langmead et al., Genome Biology 10:R25.1-R25.10]), 또는 ELAND(미국 캘리포니아주, 샌디에고 소재의 Illumina, Inc.)를 포함하지만, 이들로 제한되지 않는다. 일 실시형태에서, 혈장 cfDNA 분자들의 클론 확장된 복제물들의 일 말단은, Efficient Large-Scale Alignment of Nucleotide Databases(ELAND) 소프트웨어를 사용하는 Illumina Genome Analyzer에 대한 생물정보학 정렬 분석에 의해 서열분석되고 프로세싱된다.

[0305] 예시적이지만 비제한적인 일 실시형태에서, 본원에 기술된 방법들은 Helicos True Single Molecule Sequencing(tSMS) 기술(예를 들어, 문헌[Harris T.D. et al., Science 320:106-109]에 기술된 바와 같음)의 단일 분자 서열분석 기술을 사용하여 테스트 샘플 내의 핵산들에 대한 서열 정보를 획득하는 단계를 포함한다. tSMS 기술에서, DNA 샘플은 대략 100 내지 200개의 뉴클레오티드들의 가닥으로 절단되고, 폴리 A 서열이 각각의 DNA 가닥의 3' 말단에 첨가된다. 각각의 가닥은 형광 표지된 아데노신 뉴클레오티드의 첨가에 의해 표지된다. 이어서, DNA 가닥은 플로우 셀로 혼성화되며, 이는 플로우 셀 표면에 고정된 수백만개의 올리고-T 포획 부위들을 포함한다. 특정 실시형태에서, 템플릿들은 약 1억개 템플릿/㎠의 밀도일 수 있다. 이어서, 플로우 셀은 기기, 예를 들어, HeliScope™ 서열분석기에 로딩되고, 레이저는 플로우 셀의 표면을 조명하여, 각각의 템플릿의 위치를 드러낸다. CCD 카메라는 플로우 셀 표면 상의 템플릿들의 위치를 맵핑할 수 있다. 이어서, 템플릿 형광 라벨은 절단되고 세척 제거된다. 서열분석 반응은 DNA 중합효소 및 형광 표지된 뉴클레오티드를 도입함으로써 시작된다. 올리고-T 핵산은 프라이머로서의 역할을 한다. 중합효소는 표지된 뉴클레오티드를 템플릿 지정 방식으로 프라이머에 통합한다. 중합효소 및 통합되지 않은 뉴클레오티드들이 제거된다. 형광 표지된 뉴클레오티드의 통합을 지정한 템플릿들은 플로우 셀 표면을 이미징함으로써 구별된다. 이미징 후, 절단 단계는 형광 라벨을 제거하고, 상기 방법은 원하는 판독물 길이가 달성될 때까지 다른 형광 표지된 뉴클레오티드들로 반복된다. 각각의 뉴클레오티드 첨가 단계를 이용하여 서열 정보가 수집된다. 단일 분자 서열분석 기술에 의한 전체 게놈 서열분석은 서열 라이브러리의 제조에서 PCR-기반 증폭을 배제하거나 전형적으로 제거하며, 상기 방법은 해당 샘플의 복제물들의 측정보다는 샘플의 직접적인 측정을 가능하게 한다.

[0306] 다른 예시적이지만 비제한적인 실시형태에서, 본원에 기술된 방법들은 454 서열분석(Roche)(예를 들어, 문헌[Margulies, M. et al. Nature 437:376-380]에 기술된 바와 같음)을 사용하여 테스트 샘플 내의 핵산들에 대한 서열 정보를 획득하는 단계를 포함한다. 454 서열분석은 전형적으로 2개의 단계를 수반한다. 제1 단계에서, DNA는 대략적으로 300 내지 800개의 염기쌍들의 단편들로 전단되고, 단편들의 말단은 무디게 처리된다. 이어서, 올리고뉴클레오티드 어댑터들이 단편들의 말단들에 결찰된다. 어댑터들은 단편들의 증폭 및 서열분석을 위한 프라이머들로서 기능한다. 단편들은 예를 들어 5'-비오틴 태그를 포함하는 어댑터 B를 사용하여 DNA 포획 비드들, 예를 들어, 스트렙타비딘-코팅된 비드들에 부착될 수 있다. 비드들에 부착된 단편들은 오일-물 에멀젼의 액적 내에서 PCR 증폭된다. 결과는 각각의 비드 상의 클론 증폭된 DNA 단편들의 복수의 복제물들이다. 제2 단계에서, 비드들은 웰들(예를 들어, 피코리터(picoliter)-크기의 웰들)에 포획된다. 파이로서열분석이 각각의 DNA 단편에 대해 병렬로 수행된다. 하나 이상의 뉴클레오티드들의 첨가는 서열분석 기기에서 CCD 카메라에 의해 기록되는 광 신호를 생성한다. 신호 강도는 통합된 뉴클레오티드들의 개수에 비례한다. 파이로서열분석은 뉴클레오티드 첨가 시에 방출되는 파이로포스페이트(PPi)를 이용한다. PPi는 아데노신 5' 포스포설페이트의 존

재 하에 ATP 설퍼릴라제에 의해 ATP로 전환된다. 루시퍼라아제는 ATP를 사용하여 루시페린을 옥시루시페린으로 전환시키고, 이 반응은 측정되고 분석되는 광을 생성한다.

[0307] 다른 예시적이지만 비제한적인 실시형태에서, 본원에 기술된 방법들은 SOLiD™ 기술(Applied Biosystems)을 사용하여 테스트 샘플 내의 핵산들에 대한 서열 정보를 획득하는 단계를 포함한다. SOLiD™ 결찰에 의한 서열분석에서, 게놈 DNA는 단편들로 전단되고, 어댑터들이 단편들의 5' 말단 및 3' 말단에 부착되어 단편 라이브러리를 생성한다. 대안적으로, 내부 어댑터들은, 단편의 5' 말단 및 3' 말단에 어댑터를 결찰하고, 단편들을 원형화하고, 원형화된 단편을 분해하여 내부 어댑터를 생성하고, 생성된 단편들의 5' 말단 및 3' 말단에 어댑터를 부착하여 메이트 쌍형성된 라이브러리를 생성하는 것에 의해 도입될 수 있다. 다음으로, 클론 비드 집단들이, 비드들, 프라이머들, 템플릿들 및 PCR 구성요소들을 포함하는 미세반응계(microreactor)에서 제조된다. PCR에 이어서, 템플릿들이 변성되고 비드들이 풍부화되어, 연장된 템플릿들이 있는 비드들을 분리시킨다. 선택된 비드들 상의 템플릿들은 유리 슬라이드에 대한 결합을 허용하는 3' 변형을 거친다. 서열은 특정 형광단에 의해 식별되는 중심 결정된 염기(또는 염기들의 쌍)와 부분 무작위 올리고뉴클레오티드들의 순차적 혼성화 및 결찰에 의해 결정될 수 있다. 색상이 기록된 후, 결찰된 올리고뉴클레오티드들을 절단 및 제거하고, 이어서 상기 방법을 반복한다.

[0308] 다른 예시적이지만 비제한적인 실시형태에서, 본원에 기술된 방법들은 Pacific Biosciences의 single molecule, real-time(SMRT™) 서열분석 기술을 사용하여 테스트 샘플 내의 핵산들에 대한 서열 정보를 획득하는 단계를 포함한다. SMRT 서열분석에서, 염료-표지된 뉴클레오티드들의 연속적 통합은 DNA 합성 동안 이미징된다. 단일 DNA 중합효소 분자들은 서열 정보를 획득하는 개별 제로-모드 파장 검출기(ZMW 검출기)의 바닥 표면에 부착되는 한편, 포스포링크된(phospholinked) 뉴클레오티드들은 성장하는 프라이머 가닥에 통합된다. ZMW 검출기는 ZMW의 내외부로 신속하게(예를 들어, 마이크로초 내에) 확산되는 형광 뉴클레오티드들의 배경에 대해, DNA 중합효소에 의한 단일 뉴클레오티드의 통합의 관찰을 가능하게 하는 격리 구조를 포함한다. 성장 가닥에 뉴클레오티드를 통합시키기 위해서는 전형적으로 수 밀리초가 소요된다. 이 시간 동안, 형광 라벨이 여기되어 형광 신호를 생성하고, 형광 태그가 절단된다. 염료의 대응하는 형광의 측정은 어느 염기가 통합되었는지를 나타낸다. 상기 방법을 반복하여 서열을 제공한다.

[0309] 다른 예시적이지만 비제한적인 실시형태에서, 본원에 기술된 방법들은 나노기공 서열분석(예를 들어, 문헌[Soni GV and Meller A. Clin Chem 53: 1996-2001]에 기술된 바와 같음)을 사용하여 테스트 샘플 내의 핵산들에 대한 서열 정보를 획득하는 단계를 포함한다. 나노기공 서열분석 DNA 분석 기술은, 예를 들어, Oxford Nanopore Technologies(영국 옥스포드 소재), Sequenom, NABsys 등을 포함하는 다수의 회사에 의해 개발된다. 나노기공 서열분석은 DNA의 단일 분자가 나노기공을 통과할 때 직접 서열분석되는 단일 분자 서열분석 기술이다. 나노기공은 전형적으로 직경이 1 나노미터 정도인 작은 구멍이다. 전도성 유체 중의 나노기공의 침지 및 이것에 걸친 전위(전압)의 인가는 나노기공을 통한 이온의 전도로 인해 약간의 전류를 야기한다. 유동하는 전류의 양은 나노기공의 크기 및 형상에 민감하다. DNA 분자가 나노기공을 통과함에 따라, DNA 분자 상의 각각의 뉴클레오티드는 나노기공을 상이한 정도로 차단하여, 나노기공을 통한 전류의 크기를 상이한 정도로 변화시킨다. 따라서, DNA 분자가 나노기공을 통과할 때의 전류의 이러한 변화는 DNA 서열의 판독물을 제공한다.

[0310] 다른 예시적이지만 비제한적인 실시형태에서, 본원에 기술된 방법들은 화학적 감응 전계 효과 트랜지스터(chemFET: chemical-sensitive field effect transistor) 어레이(예를 들어, 미국 특허출원공개 제2009/0026082호에 기술된 바와 같음)를 사용하여 테스트 샘플 내의 핵산들에 대한 서열 정보를 획득하는 단계를 포함한다. 이 기술의 일례에서, DNA 분자들은 반응 챔버 내에 배치될 수 있고, 템플릿 분자들은 중합효소에 결합된 서열분석 프라이머에 혼성화될 수 있다. 서열분석 프라이머의 3' 말단에서 새로운 핵산 가닥 내로의 하나 이상의 트리포스페이트의 통합은 chemFET에 의한 전류의 변화로서 식별될 수 있다. 상기 어레이는 다수의 chemFET 센서를 가질 수 있다. 다른 예에서, 단일 핵산들이 비드들에 부착될 수 있고, 핵산들은 비드 상에서 증폭될 수 있고, 개별 비드는 chemFET 어레이 상의 개별 반응 챔버로 전달될 수 있으며, 각각의 챔버는 chemFET 센서를 갖고, 핵산들은 서열분석될 수 있다.

[0311] 다른 실시형태에서, DNA 서열분석 기술은 Ion Torrent 단일 분자 서열분석이며, 이는 반도체 기술과 단순한 서열분석 화학반응을 결합하여 화학적으로 암호화된 정보(A, C, G, T)를 반도체 칩 상의 디지털 정보(0, 1)로 직접 변환한다. 사실상, 뉴클레오티드가 중합효소에 의해 DNA의 가닥 내로 통합될 때, 수소 이온이 부산물로서 방출된다. Ion Torrent는 미세기계가공된 웰들의 고밀도 어레이를 사용하여, 이러한 생화학적 과정을 대규모 병렬 방식으로 수행한다. 각각의 웰은 상이한 DNA 분자를 보유한다. 웰들 아래에는 이온 감응 층이 있고, 그 아래에는 이온 센서가 있다. 뉴클레오티드, 예를 들어, C가 DNA 템플릿에 첨가되고 이어서 DNA의 가닥 내로 통

합될 때, 수소 이온이 방출될 것이다. 이러한 이온으로부터의 전하는 용액의 pH를 변화시킬 것이며, 이는 Ion Torrent의 이온 센서에 의해 검출될 수 있다. 서열분석기(본질적으로 세계 최소의 고체 pH 측정기)는 염기를 호출하여, 화학적 정보로부터 디지털 정보로 직접 진행한다. 이어서, Ion personal Genome Machine(PGM™) 서열분석기는 연속적으로 뉴클레오티드로 칩을 순차적으로 채운다. 칩을 채우는 다음 뉴클레오티드가 매치가 아닌 경우. 어떠한 전압 변화도 기록되지 않을 것이고, 어떠한 염기도 호출되지 않을 것이다. DNA 가닥 상에 2개의 동일한 염기가 있는 경우, 전압은 두 배일 것이고, 칩은 호출되는 2개의 동일한 염기를 기록할 것이다. 직접적인 검출은 초 단위의 뉴클레오티드 통합의 기록을 허용한다.

[0312] 다른 실시형태에서, 본 방법은 혼성화에 의한 서열분석을 사용하여 테스트 샘플 내의 핵산들에 대한 서열 정보를 획득하는 단계를 포함한다. 혼성화에 의한 서열분석은 복수의 폴리뉴클레오티드 서열들을 복수의 폴리뉴클레오티드 프로브들과 접촉시키는 단계를 포함하며, 여기서 상기 복수의 폴리뉴클레오티드 프로브 각각은 임의로 기재에 부착될 수 있다. 기재는 공지된 뉴클레오티드 서열들의 어레이를 포함하는 평평한 표면일 수 있다. 어레이에 대한 혼성화의 패턴은 샘플에 존재하는 폴리뉴클레오티드 서열들을 결정하는 데 사용될 수 있다. 다른 실시형태들에서, 각각의 프로브는 비드, 예를 들어, 자성 비드 등에 부착된다. 샘플 내의 복수의 폴리뉴클레오티드 서열들을 식별하기 위해 비드들에 대한 혼성화가 결정되고 사용될 수 있다.

[0313] 본원에 기술된 방법들의 일부 실시형태들에서, 서열 판독물들은 약 20 bp, 약 25 bp, 약 30 bp, 약 35 bp, 약 40 bp, 약 45 bp, 약 50 bp, 약 55 bp, 약 60 bp, 약 65 bp, 약 70 bp, 약 75 bp, 약 80 bp, 약 85 bp, 약 90 bp, 약 95 bp, 약 100 bp, 약 110 bp, 약 120 bp, 약 130, 약 140 bp, 약 150 bp, 약 200 bp, 약 250 bp, 약 300 bp, 약 350 bp, 약 400 bp, 약 450 bp, 또는 약 500 bp이다. 기술적 진보는 500 bp 초과의 단일-말단 판독물들을 가능하게 하며, 쌍형성된 말단 판독물들이 생성되는 경우 약 1000 bp 초과의 판독물들을 가능하게 할 것으로 예상된다. 일부 실시형태들에서, 쌍형성된 말단 판독물들이 반복 확장을 결정하는 데 사용되며, 이는 약 20 bp 내지 1000 bp, 약 50 bp 내지 500 bp, 또는 80 bp 내지 150 bp인 서열 판독물들을 포함한다. 다양한 실시형태들에서, 쌍형성된 말단 판독물들이 반복 확장을 갖는 서열을 평가하는 데 사용된다. 반복 확장을 갖는 서열은 상기 판독물들보다 더 길다. 일부 실시형태들에서, 반복 확장을 갖는 서열은 약 100 bp, 500 bp, 1000 bp, 또는 4000 bp보다 길다. 서열 판독물들의 맵핑은 판독물들의 서열을 참조 서열과 비교하여 서열분석된 핵산 분자의 염색체 기원을 결정함으로써 달성되며, 특정 유전자 서열 정보는 필요하지 않다. 혼합된 샘플 내의 참조 게놈과 게놈들 사이에 존재할 수 있는 소수의 다형성을 고려하기 위해 작은 정도의 미스매치(판독물당 0 내지 2개의 미스매치들)가 허용될 수 있다. 일부 실시형태들에서, 참조 서열에 정렬된 판독물들은 앵커 판독물들로서 사용되며, 앵커 판독물들에 대해 쌍을 이루지만 참조물에 정렬될 수 없거나 불량하게 정렬된 판독물들은 앵커링된 판독물들로서 사용된다. 일부 실시형태들에서, 불량하게 정렬된 판독물들은 판독물당 비교적 큰 수의 미스매치 백분율, 예를 들어, 판독물당 적어도 약 5%, 적어도 약 10%, 적어도 약 15%, 또는 적어도 약 20%의 미스매치들을 가질 수 있다.

[0314] 복수의 서열 태그들(즉, 참조 서열에 정렬된 판독물들)이 전형적으로 샘플당 획득된다. 일부 실시형태들에서, 예를 들어 100 bp의, 적어도 약 3 x 10^6개의 서열 태그들, 적어도 약 5 x 10^6개의 서열 태그들, 적어도 약 8 x 10^6개의 서열 태그들, 적어도 약 10 x 10^6개의 서열 태그들, 적어도 약 15 x 10^6개의 서열 태그들, 적어도 약 20 x 10^6개의 서열 태그들, 적어도 약 30 x 10^6개의 서열 태그들, 적어도 약 40 x 10^6개의 서열 태그들, 또는 적어도 약 50 x 10^6개의 서열 태그들이, 샘플당 참조 게놈에 대한 판독물들을 맵핑함으로써 획득된다. 일부 실시형태들에서, 모든 서열 판독물들은 참조 게놈의 모든 영역들에 맵핑되어, 게놈-전체 판독물들을 제공한다. 다른 실시형태들에서, 판독물들은 관심있는 서열, 예를 들어, 염색체, 염색체의 세그먼트, 또는 관심있는 반복 서열에 맵핑된다.

[0315] ***반복 확장을 결정하기 위한 장치 및 시스템***

[0316] 서열분석 데이터의 분석 및 이로부터 도출된 진단은 전형적으로, 다양한 컴퓨터 실행 알고리즘들 및 프로그램들을 이용하여 수행된다. 따라서, 특정 실시형태들은 하나 이상의 컴퓨터 시스템들 또는 기타 프로세싱 시스템들에 저장되거나 또는 이를 통해 전달되는 데이터를 수반하는 방법들을 사용한다. 본원에 개시된 실시형태들은 또한 이들 동작을 수행하기 위한 장치에 관한 것이다. 이 장치는 요구되는 목적을 위해 특별히 구성될 수 있거나, 또는 컴퓨터에 저장된 컴퓨터 프로그램 및/또는 데이터 구조에 의해 선택적으로 활성화 또는 재구성되는 범용 컴퓨터(또는 컴퓨터들의 그룹)일 수 있다. 일부 실시형태들에서, 프로세서들의 그룹은, (예를 들어, 네트워크 또는 클라우드 컴퓨팅을 통해) 협력적으로 및/또는 병렬로, 언급된 분석 동작들의 일부 또는 전부를 수행한

다. 본원에 기술된 방법들을 수행하기 위한 프로세서 또는 프로세서들의 그룹은 프로그램가능 디바이스들(예를 들어, CPLD들 및 FPGA들) 및 게이트 어레이 ASIC들 또는 범용 마이크로프로세서들과 같은 비-프로그램가능 디바이스들과 같은 마이크로제어기들 및 마이크로프로세서들을 포함하는 다양한 유형의 것일 수 있다.

[0317] 일 실시형태는 반복 서열들을 포함하는 게놈 좌위들에서 변이체들의 유전자형들을 결정하는 데 사용하기 위한 시스템을 제공하며, 상기 시스템은 핵산 샘플을 수용하고 샘플로부터 핵산 서열 정보를 제공하기 위한 서열분석기; 프로세서; 및 명령어들을 저장하는 기계 판독가능 저장 매체를 포함하고, 상기 명령어들은 상기 프로세서 상에서 실행되어, (a) 데이터베이스로부터 테스트 샘플의 핵산 서열 판독물들을 수집하고;(b) 서열 그래프에 의해 각각 표현된 하나 이상의 반복 서열들에 서열 판독물들을 정렬시키고 - 상기 서열 그래프는 핵산 서열들을 표현하는 정점들 및 상기 정점들을 연결하는 방향성 에지들을 갖는 방향성 그래프의 데이터 구조를 갖고, 상기 서열 그래프는 하나 이상의 자가-루프들을 포함하며, 각각의 자가-루프는 반복 하위서열을 표현하고, 각각의 반복 하위서열은 하나 이상의 뉴클레오티드들의 반복 단위의 반복체들을 포함함 -; (c) 상기 하나 이상의 반복 서열들에 정렬된 서열 판독물을 사용하여 상기 하나 이상의 반복 서열들에 대한 하나 이상의 유전자형들을 결정함으로써, 변이체들을 유전자형분석한다.

[0318] 본원에 제공된 시스템들 중 임의의 것의 일부 실시형태들에서, 서열분석기는 차세대 서열분석(NGS)을 수행하도록 구성된다. 일부 실시형태들에서, 서열분석기는 가역적 염료 종결자를 이용한 합성에 의한 서열분석을 이용하여 대규모 병렬 서열분석을 수행하도록 구성된다. 다른 실시형태들에서, 서열분석기는 결찰에 의한 서열분석을 수행하도록 구성된다. 또 다른 실시형태들에서, 서열분석기는 단일 분자 서열분석을 수행하도록 구성된다.

[0319] 또한, 특정 실시형태는 다양한 컴퓨터-구현 동작들을 수행하기 위한 프로그램 명령어들 및/또는 데이터(데이터 구조를 포함함)를 포함하는 유형의 및/또는 비일시적 컴퓨터 판독가능 매체 또는 컴퓨터 프로그램 제품들에 관한 것이다. 컴퓨터 판독가능 매체의 예는 반도체 메모리 디바이스들, 디스크 드라이브들, 자기 테이프들과 같은 자기 매체, CD와 같은 광학 매체, 자기-광학 매체, 및 판독전용 메모리 디바이스(ROM: read-only memory device)들 및 랜덤 액세스 메모리(RAM: random access memory)와 같은, 프로그램 명령어들을 저장 및 수행하도록 특별히 구성된 하드웨어 디바이스들을 포함하지만, 이들로 제한되지 않는다. 컴퓨터 판독가능 매체는 최종 사용자에 의해 직접 제어될 수 있거나 또는 상기 매체는 최종 사용자에 의해 간접적으로 제어될 수 있다. 직접 제어 매체의 예는 사용자 시설에 위치한 매체 및/또는 다른 엔티티와 공유되지 않는 매체를 포함한다. 간접 제어 매체의 예는 외부 네트워크를 통해 및/또는 "클라우드"와 같은 공유 리소스를 제공하는 서비스를 통해 사용자에 간접적으로 액세스가능한 매체를 포함한다. 프로그램 명령어들의 예는 컴파일러에 의해 생성된 것과 같은 기계 코드, 및 인터프리터를 이용하여 컴퓨터에 의해 실행될 수 있는 더 높은 수준의 코드를 포함하는 파일들 둘 모두를 포함한다.

[0320] 다양한 실시형태들에서, 개시된 방법들 및 장치들에 사용된 데이터 또는 정보는 전자 포맷으로 제공된다. 이러한 데이터 또는 정보는 핵산 샘플로부터 유도된 판독물들 및 태그들, 참조 서열들(오로지 또는 주로 다형성을 제공하는 참조 서열들을 포함함), 반복 확장 호출들과 같은 호출들, 카운슬링 추천, 진단 등을 포함할 수 있다. 본원에 사용되는 바와 같이, 전자 포맷으로 제공되는 데이터 또는 기타 정보는 기계 상의 저장 및 기계들 간의 전송에 이용 가능하다. 통상적으로, 전자 포맷의 데이터는 디지털로 제공되고, 비트 및/또는 바이트로 다양한 데이터 구조, 리스트, 데이터베이스 등에 저장될 수 있다. 상기 데이터는 전자적으로, 광학적으로 등으로 구현될 수 있다.

[0321] 일 실시형태는 테스트 샘플 내의 반복 확장의 존재 또는 부재를 나타내는 출력을 생성하기 위한 컴퓨터 프로그램 제품을 제공한다. 상기 컴퓨터 제품은 반복 확장을 결정하기 위한 전술된 방법들 중 임의의 하나 이상을 수행하기 위한 명령어들을 포함할 수 있다. 설명된 바와 같이, 상기 컴퓨터 제품은, 프로세서가 앵커링된 판독물 및 앵커링된 판독물 내의 반복체들을 결정하고 반복 확장이 존재하는지 또는 부재하는지 여부를 결정할 수 있게 하기 위한 컴퓨터 실행가능 또는 컴파일가능한 로직(예를 들어, 명령어들)이 기록된 비일시적 및/또는 유형의 컴퓨터 판독가능 매체를 포함할 수 있다. 일례에서, 상기 컴퓨터 제품은 프로세서가 반복 확장을 진단할 수 있게 하기 위한 컴퓨터 실행가능 또는 컴파일가능한 로직(예를 들어, 명령어들)이 기록된 컴퓨터 판독가능 매체를 포함하며, 상기 로직은, 생물학적 샘플로부터의 핵산 분자들의 적어도 일부로부터 서열분석 데이터를 수신하기 위한 수신 절차 - 상기 서열분석 데이터는 반복 서열에 대한 정렬을 거친 쌍형성된 말단 판독물들을 포함함 -; 상기 수신된 데이터로부터 반복 확장을 분석하기 위한 컴퓨터 보조 로직; 및 상기 반복 확장의 존재, 부재 또는 종류를 나타내는 출력을 생성하기 위한 출력 절차를 포함한다.

[0322] 고려되는 샘플로부터의 서열 정보는 염색체 참조 서열들에 맵핑되어, 관심있는 반복 서열에 정렬되거나 이에 앵

커링된 쌍형성된 말단 판독물들을 식별하고 반복 서열의 반복 확장들을 식별할 수 있다. 다양한 실시형태들에서, 참조 서열들은 관계형 또는 객체 데이터베이스와 같은 데이터베이스에 저장된다.

[0323] 인간이 도움을 받지 않고 본원에 개시된 방법들의 연산 동작들을 수행하는 것은 실용적이지 않거나, 또는 대부분의 경우에서 심지어 불가능하다는 것이 이해되어야 한다. 예를 들어, 샘플로부터의 단일 30 bp 판독물을 인간 염색체 중 임의의 것에 맵핑하는 것은 연산 장치의 도움이 없으면 몇 년의 노력을 필요로 할 수 있다. 물론, 신뢰성 있는 반복 확장 호출들은 일반적으로 수천개(예를 들어, 적어도 약 10,000개) 또는 심지어 수백만개의 판독물들을 하나 이상의 염색체들에 맵핑하는 것을 필요로 하기 때문에, 이 문제는 가중된다.

[0324] 다양한 구현예들에서, 원시 서열 판독물들이, 하나 이상의 관심있는 서열을 표현하는 하나 이상의 서열 그래프들에 정렬된다. 다양한 구현예들에서, 적어도 10,000, 100,000, 500,000, 1,000,000, 5,000,000 또는 10,000,000개의 판독물들이 하나 이상의 서열 그래프들에 정렬된다. 다양한 구현예들에서, 상기 하나 이상의 서열 그래프들은 적어도 1, 2, 5, 10, 50, 100, 500, 1000, 5,000, 10,000, 또는 50,000개의 서열 그래프들을 포함한다.

[0325] 일부 구현예들에서, 원시 서열 판독물들은 초기에 참조 게놈에 정렬되어, 초기에 정렬된 판독물들의 서브세트가 하나 이상의 관심있는 서열을 표현하는 하나 이상의 서열 그래프들에 정렬되기 전에, 판독물들의 게놈 좌표들을 결정하도록 한다. 다양한 구현예들에서, 적어도 10,000, 100,000, 500,000, 1,000,000, 5,000,000, 10,000,000, 또는 100,000,000개의 판독물들이 초기에 참조 게놈에 정렬된다. 일부 구현예들에서, 초기에 정렬된 판독물들은, 다수의 영역에서 반복 확장들을 결정하기 위해 서열 그래프들에 재정렬된다(각각의 영역은 서열 그래프에 대응함). 구현예의 각각의 호출 동안 서열 그래프들에 재정렬되는 판독물들의 총 개수는 수천 내지 수백만개의 판독물의 범위일 수 있다. 다양한 구현예들에서, 적어도 100, 500, 1,000, 5,000, 10,000, 50,000, 100,000, 500,000, 1,000,000, 5,000,000 또는 10,000,000개의 판독물들이 각각의 서열 그래프에 재정렬된다. 다양한 구현예들에서, 상기 하나 이상의 서열 그래프들은 적어도 1, 2, 5, 10, 50, 100, 500, 1000, 5,000, 10,000, 또는 50,000개의 서열 그래프들을 포함한다.

[0326] 본원에 개시된 방법들은 반복 서열을 포함하는 게놈 좌위에서 변이체들의 유전자형들을 결정하기 위한 시스템을 사용하여 수행될 수 있다. 상기 시스템은: (a) 샘플로부터 핵산 서열 정보를 제공하는 테스트 샘플로부터 핵산들을 수용하기 위한 서열분석기; (b) 프로세서; 및 (c) 명령어들이 저장된 하나 이상의 컴퓨터 판독가능 저장 매체를 포함하고, 상기 명령어들은 상기 프로세서 상에서 실행되어 반복 서열을 포함하는 게놈 좌위들에서의 변이체들을 유전자형분석한다. 일부 실시형태들에서, 상기 방법은 임의의 반복 확장들을 식별하기 위한 방법을 수행하기 위한 컴퓨터 판독가능 명령어들이 저장된 컴퓨터 판독가능 매체에 의해 지시된다. 따라서, 일 실시형태는 프로그램 코드를 저장하는 비일시적 기계 판독가능 매체를 포함하는 컴퓨터 프로그램 제품을 제공하며, 상기 프로그램 코드는, 컴퓨터 시스템의 하나 이상의 프로세서들에 의해 실행될 때, 컴퓨터 시스템으로 하여금 핵산들을 포함하는 테스트 샘플에서 반복 서열의 반복 확장들을 식별하기 위한 방법을 구현하게 하며, 여기서 반복 서열은 뉴클레오티드들의 반복 단위의 반복체들을 포함한다. 상기 프로그램 코드는, (a) 데이터베이스로부터 테스트 샘플의 서열 판독물들을 수집하기 위한 코드; (b) 서열 그래프에 의해 각각 표현된 하나 이상의 반복 서열들에 서열 판독물들을 정렬시키기 위한 코드 - 상기 서열 그래프는 핵산 서열들을 표현하는 정점들 및 상기 정점들을 연결하는 방향성 에지들을 갖는 방향성 그래프의 데이터 구조를 갖고, 상기 서열 그래프는 하나 이상의 자가-루프들을 포함하며, 각각의 자가-루프는 반복 하위서열을 표현하고, 각각의 반복 하위서열은 하나 이상의 뉴클레오티드들의 반복 단위의 반복체들을 포함함 - 및; (c) 상기 하나 이상의 반복 서열들에 정렬된 서열 판독물을 사용하여 상기 하나 이상의 반복 서열들에 대한 하나 이상의 유전자형들을 결정하기 위한 코드를 포함할 수 있다.

[0327] 일부 실시형태들에서, 상기 명령어들은 상기 방법에 관련된 정보, 예를 들어 반복체들 및 앵커 판독물들, 및 테스트 샘플을 제공하는 인간 대상체에 대한 환자 의료 기록에서의 반복 확장의 존재 또는 부재를 자동으로 기록하는 것을 추가로 포함할 수 있다. 환자 의료 기록은, 예를 들어, 실험실, 의사의 사무실, 병원, 보건기관, 보험회사, 또는 개인 의료 기록 웹사이트에 의해 유지될 수 있다. 추가로, 프로세서-구현 분석의 결과에 기초하여, 상기 방법은 테스트 샘플이 취해진 인간 대상체의 치료를 처방, 개시, 및/또는 변경하는 단계를 추가로 수반할 수 있다. 이는 대상체로부터 취해진 추가적인 샘플들에 대해 하나 이상의 추가적인 테스트 또는 분석을 수행하는 단계를 포함할 수 있다.

[0328] 개시된 방법들은 또한 임의의 반복 확장들을 식별하기 위한 방법을 수행하도록 적응되거나 구성된 컴퓨터 프로세싱 시스템을 사용하여 수행될 수 있다. 일 실시형태는 본원에 개시된 바와 같은 방법을 수행하도록 적응되거

나 구성된 컴퓨터 프로세싱 시스템을 제공한다. 일 실시형태에서, 상기 장치는 샘플 내의 핵산 분자들 중 적어도 일부를 서열분석하여 본원의 어딘가 다른 곳에 기술된 서열 정보의 유형을 획득하도록 적응되거나 구성된 서열분석 디바이스를 포함한다. 상기 장치는 또한 샘플을 프로세싱하기 위한 구성요소들을 포함할 수 있다. 이러한 구성요소들은 본원에서 다른 곳에 기술되어 있다.

[0329] 서열 또는 기타 데이터는 직접 또는 간접적으로 컴퓨터에 입력되거나 또는 컴퓨터 판독가능 매체 상에 저장될 수 있다. 일 실시형태에서, 컴퓨터 시스템은 샘플들로부터의 핵산들의 서열들을 판독 및/또는 분석하는 서열분석 디바이스에 직접 결합된다. 이러한 도구들로부터의 서열 또는 기타 정보는 컴퓨터 시스템 내의 인터페이스를 통해 제공된다. 대안적으로, 시스템에 의해 프로세싱되는 서열들은 데이터베이스 또는 기타 저장소와 같은 서열 저장 소스로부터 제공된다. 프로세싱 장치에 이용 가능하면, 메모리 디바이스 또는 대량 저장 디바이스는, 적어도 일시적으로, 핵산들의 서열을 버퍼링 또는 저장한다. 또한, 메모리 디바이스는 다양한 염색체들 또는 게놈들 등에 대한 태그 카운트들을 저장할 수 있다. 메모리는 또한 제시된 서열 또는 맵핑된 데이터를 분석하기 위한 다양한 루틴들 및/또는 프로그램들을 저장할 수 있다. 이러한 프로그램들/루틴들은 통계적 분석 등을 수행하기 위한 프로그램들을 포함할 수 있다.

[0330] 일례에서, 사용자는 샘플을 서열분석 장치 내에 제공한다. 컴퓨터에 연결된 서열분석 장치에 의해 데이터가 수집 및/또는 분석된다. 컴퓨터 상의 소프트웨어는 데이터 수집 및/또는 분석을 가능하게 한다. 데이터는 저장되고/거나, (모니터 또는 기타 유사한 디바이스를 통해) 표시되고/거나, 다른 위치로 송신될 수 있다. 컴퓨터는 원격 사용자(예를 들어, 의사, 과학자 또는 분석가)에 의해 사용되는 핸드헬드 디바이스에 데이터를 전송하는 데 사용되는 인터넷에 연결될 수 있다. 데이터는 송신 이전에 저장 및/또는 분석될 수 있음이 이해된다. 일부 실시형태들에서, 미가공 데이터는 수집되고, 데이터를 분석 및/또는 저장할 원격 사용자 또는 장치에 송신된다. 송신은 인터넷을 통해 일어날 수 있지만, 또한 위성 또는 기타 연결을 통해 일어날 수 있다. 대안적으로, 데이터는 컴퓨터 판독가능 매체 상에 저장될 수 있고, 매체는 최종 사용자에 (예를 들어, 우편을 통해) 배송될 수 있다. 원격 사용자는 빌딩, 도시, 주, 국가 또는 대륙을 포함하지만 이들로 제한되지 않는 동일한 또는 상이한 지리적 위치에 있을 수 있다.

[0331] 일부 실시형태들에서, 상기 방법들은 또한 복수의 폴리뉴클레오티드 서열들에 관한 데이터(예를 들어, 판독물들, 태그들 및/또는 참조 염색체 서열들)를 수집하고 데이터를 컴퓨터 또는 다른 연산 시스템에 송신하는 단계를 포함한다. 예를 들어, 컴퓨터는 실험실 장비, 예를 들어, 샘플 수집 장치, 뉴클레오티드 증폭 장치, 뉴클레오티드 서열분석 장치, 또는 혼성화 장치에 연결될 수 있다. 이어서 컴퓨터는 실험실 디바이스에 의해 채집된 적용 가능한 데이터를 수집할 수 있다. 데이터는 임의의 단계에서, 예를 들어, 실시간으로 수집되는 동안, 송신 이전에, 송신 동안 또는 송신과 함께, 또는 송신 후에 컴퓨터 상에 저장될 수 있다. 데이터는 컴퓨터로부터 추출될 수 있는 컴퓨터 판독가능 매체 상에 저장될 수 있다. 수집 또는 저장된 데이터는 컴퓨터로부터 원격 위치로, 예를 들어, 로컬 네트워크 또는 인터넷과 같은 광역 네트워크를 통해 전송될 수 있다. 원격 위치에서 다양한 동작들이 후술된 바와 같이 전송된 데이터 상에서 수행될 수 있다.

[0332] 본원에 개시된 시스템들, 장치들, 및 방법들에서 저장, 송신, 분석, 및/또는 조작될 수 있는 전자적으로 포맷화된 데이터의 유형 중에는 하기가 있다:

[0333] 테스트 샘플 내의 핵산들을 서열분석함으로써 획득된 판독물들

[0334] 판독물들을 참조 게놈 또는 기타 참조 서열 또는 서열들에 정렬시킴으로써 획득된 태그들

[0335] 참조 게놈 또는 서열

[0336] 좌위 아이덴티티, 위치, 및 구조를 나타내는 좌위 사양

[0337] 판독물 커버리지

[0338] 변이체들의 유전자형

[0339] 서열 그래프

[0340] 그래프 경로들

[0341] 그래프 정렬 정보

[0342] 반복 확장의 실제 호출들

[0343] 진단들(호출들과 연관된 임상 조건)

[0344] 호출들 및/또는 진단들로부터 도출된 추가적인 테스트에 대한 추천

[0345] 호출들 및/또는 진단들로부터 도출된 치료 및/또는 모니터링 계획

[0346] 이들 다양한 유형의 데이터는 별개의 장치를 이용하여 하나 이상의 위치들에서 획득, 저장, 전송, 분석, 및/또는 조작될 수 있다. 프로세싱 옵션들은 넓은 스펙트럼에 펼쳐 있다. 스펙트럼의 일 단부에서, 이 정보의 전부 또는 대부분이 테스트 샘플이 프로세싱되는 위치, 예를 들어, 의사의 사무실 또는 기타 임상 설정에 저장되고 사용된다. 다른 극단에서, 샘플은 한 위치에서 획득되고, 상이한 위치에서 프로세싱되고 임의로 서열분석되고, 판독물들은 정렬되고 하나 이상의 상이한 위치들에서 호출이 이루어지고, 진단, 추천, 및/또는 계획이 또 다른 위치(이는 샘플이 획득되었던 위치일 수 있음)에서 준비된다.

[0347] 다양한 실시형태들에서, 판독물은 서열분석 장치로 생성되고, 이어서 반복 확장 호출들을 생성하기 위해 프로세싱되는 원격 장소로 송신된다. 이 원격 위치에서, 일례로서, 판독물들은 참조 서열에 정렬되어 앵커 판독물들 및 앵커링된 판독물들을 생성한다. 별개의 위치들에서 사용될 수 있는 프로세싱 동작들 중에는 하기가 있다:

[0348] 샘플 수집

[0349] 서열분석을 준비하는 샘플 프로세싱

[0350] 서열분석

[0351] 서열 데이터 분석 및 반복 확장 호출 도출

[0352] 진단

[0353] 진단 및/또는 호출을 환자 또는 건강관리 제공자에 보고

[0354] 추가적인 치료, 테스트, 및/또는 모니터링을 위한 계획 개발

[0355] 이 동작들 중 임의의 하나 이상은 본원의 다른 곳에 기재된 바와 같이 자동화될 수 있다. 전형적으로, 서열분석 및 서열 데이터의 분석, 및 반복 확장 호출들을 도출하는 것은 컴퓨터를 사용하여 수행될 것이다. 다른 동작들은 수동으로 또는 자동으로 수행될 수 있다.

[0356] 도 18은 테스트 샘플로부터 호출 또는 진단을 생성하기 위한 분산 시스템의 일 구현예를 도시한다. 환자로부터 테스트 샘플을 획득하기 위해 샘플 수집 위치(01)가 사용된다. 이어서, 샘플들은 테스트 샘플이 전술된 바와 같이 프로세싱 및 서열분석될 수 있는 프로세싱 및 서열분석 위치(03)에 제공된다. 위치(03)는 샘플을 프로세싱하기 위한 장치뿐만 아니라 프로세싱된 샘플을 서열분석하기 위한 장치를 포함한다. 서열분석의 결과는, 본원의 다른 곳에 기술된 바와 같이, 전형적으로 전자 포맷으로 제공되고 인터넷과 같은 네트워크에 제공되고, 도 18의 도면 부호 05에 의해 표시되는 판독물들의 집합이다.

[0357] 서열 데이터는 분석 및 호출 생성이 수행되는 원격 위치(07)에 제공된다. 이 위치는 컴퓨터들 또는 프로세서들과 같은 하나 이상의 강력한 연산 디바이스들을 포함할 수 있다. 위치(07)의 연산 리소스가 이의 분석을 완료하고, 수신된 서열 정보로부터 호출을 생성한 후에, 호출은 네트워크(05)로 다시 전달된다. 일부 구현예들에서, 위치(07)에서는 호출이 생성될뿐만 아니라 연관된 진단도 또한 생성된다. 이어서, 호출 및 또는 진단은 도 18에 예시된 바와 같이 네트워크를 거쳐 다시 샘플 수집 위치(01)로 전송된다. 설명된 바와 같이, 이것은 단순히, 호출 또는 진단을 생성하는 것과 연관된 다양한 동작들이 어떻게 다양한 위치들 중에 분배될 수 있는지에 대한 많은 변형들 중 하나이다. 하나의 일반적인 변형은 단일 위치에서 샘플 수집 및 프로세싱 및 서열분석을 제공하는 것을 포함한다. 다른 변형은 분석 및 호출 생성과 동일한 위치에서 프로세싱 및 서열분석을 제공하는 것을 포함한다.

[0358] **실험**

[0359] ***실시예 1: 짧은 반복체***

[0360] 실시예 1 내지 3은 일부 구현예에 따른 올바르게 유전자형이 지정된 반복 영역들을 시각화한다. 대립유전자들이 도 19에 도시된 판독물 길이보다 짧은 *ATXN3* 반복체에 대한 판독물 파일업을 고려한다. 도 19는 두 일배체형 모두에서 반복 영역(1902)에 20개의 모티프들을 갖는, 유전자형 20/20을 갖는 *ATXN3* 반복체에 대한 판독물 파일업을 도시한다. 서열 중단들은 대부분의 판독물 정렬들에서 미스매칭된 위치들에 대응한다.

[0361] 이 플롯의 각각의 패널은 일배체형에 대응한다. 일배체형 서열들과 판독물들은 반복체(1902)(주황색) 또는 둘러싼 플랭킹 서열(청색)과의 중첩에 따라 색상이 지정된다. 판독물들에서의 미스매칭된 모든 염기들이 도시되며, 정렬이 잘린 위치는 들쭉날쭉한 에지로 표시된다.

[0362] 상기 파일업 플롯은 유전자형 호출이 판독물들에 의해 잘 지원된다는 것을 보여주는데, 그 이유는, 각각의 대립유전자가 많은 스패닝 판독물(spanning read)들(반복체 전체에 걸쳐 있는 판독물들)에 의해 지원되며, 일치하지 않는 정렬들을 갖는 판독물들이 없기 때문이다. (일치하지 않는 정렬은 판독물이 두 일배체형 중 하나와 일치하지 않음을 의미하며, 예를 들어, 40개의 반복체들을 갖는 판독물은 유전자형 20/20과 일치하지 않을 것이다.) 반복 서열 내의 중단들에 대한 분명한 증거가 존재한다. 예를 들어 마지막에서 세 번째 모티프의 시토신은 티민으로 돌연변이된다.

[0363] ***실시예 2: 확장된 반복체***

[0364] 도 20은 정규 크기 대립유전자(2002) 및 확장된 대립유전자(2204)를 갖는 *DMPK* 반복체를 도시한다. 상기 확장된 반복체는 판독물들에 의해 잘 지원되는데, 그 이유는, 구현예들이 반복체 전체에 판독물들을 분포시켜 전체 일배체형에 걸쳐 유사한 판독물 커버리지를 달성하기 때문이다. 반복체 내에서의 판독물들의 정렬 위치들은 무작위로 선택됨을 주목한다. 짧은 대립유전자는 또한 다수의 스패닝 판독물들에 의해 잘 지원된다.

[0365] ***실시예 3: 2개의 인접한 반복체들을 갖는 좌위***

[0366] 일부 구현예의 더 복잡한 적용을 설명하기 위해, 이들을 사용하여 2개의 인접한 반복체들, 즉, 병원성 CAG 반복체 및 근처의 "방해" CCG 반복체를 포함하는 *HTT* 반복 영역을 시각화한다. 전자의 반복체는 14/17로 유전자형이 지정되고 후자의 반복체는 9/12로 유전자형이 지정된다. 도 21a는 2개의 인접한 반복체들, 즉 CAG 반복체(2104) 및 2108 CCG 반복체를 포함하는 *HTT* 좌위에 대한 판독물 파일업을 도시한다. 상기 파일업은 또한 좌측 플랭크(2102), 개재 서열(2106) CAACAG 및 우측 플랭크(2110)를 포함한다.

[0367] 결과적으로, 도 21a에 나타낸 일배체형들 중 하나는 각각 크기 14 및 12의 반복체들을 포함하는 반면, 다른 일배체형은 크기 17 및 9의 반복체들을 포함한다. 두 일배체형 모두 판독물들에 의해 잘 지원된다는 것이 분명한다. 또한, 상기 파일업 플롯은 두 일배체형 모두에서 CCG 반복체 모티프의 제2 복제물 내에 G에서 A로의 돌연변이가 있음을 보여준다. 특히 커버리지 수준은 두 일배체형 모두에서 위치 전체에 걸쳐 비교적 균일하다.

[0368] 비교를 위해, 도 21b는 통상적인 도구 및 동일한 서열 판독 데이터를 사용한 *HTT* 영역의 서열 파일업을 도시한다. 파일업은 2개의 개별화된 일배체형들 대신에 단일 가닥 참조 서열만 포함한다. 반복 영역은 2개의 인접한 반복체, 즉 CAG 반복체(2124) 및 CCG 반복체(2128)를 포함한다. 파일업은 또한 좌측 플랭크(2122), 개재 서열 CAACAG(2126) 및 우측 플랭크(2130)를 포함한다. 서열 판독물들은 참조 서열 전체에 균일하게 분포되지 않았음을 주목한다. 반복 영역(2128)의 커버리지는 낮으며, 다수의 판독물들이 나뉘어서, 커버리지가 낮거나 없는 반복 영역의 섹션에 걸쳐 신장된다. 이것은 데이터가 이 영역에서 참조물의 유전자형과 매칭되지 않는다는 신호이지만, 상기 파일업은 샘플의 진정한 유전자형을 명확하게 나타내지 않는다.

[0369] ***실시예 4: 과대추정된 반복체 크기***

[0370] 실시예 4 및 5는 잘못 유전자형이 지정된 반복 영역들을 시각화한다. 위양성(false-positive) 반복 확장 호출에 대응하는 파일업의 예를 제공하기 위해, 실시예 4는 동형접합 유전자형 10/10을 갖는 *C9ORF72* 반복 영역으로부터의 시뮬레이션된 판독물을 보여준다. 실무자들은 반복 서열과 다소 유사한 C 동종중합체 판독물에 스파이크하였고, 반복체 유전자형이 10/10 대신에 10/30이 되도록 일부 구현예들을 실행하였다. **도 22**는 시뮬레이션된 데이터를 사용하여 하나의 일배체형 상의 2204 반복 영역에서의 *C9ORF72* 반복체의 잘못 호출된 확장을 포함하는 판독물 파일업을 도시한다. 다른 일배체형 상의 반복 영역 2202는 확장되지 않았다. 예상대로, 상기 파일업은 더 긴 반복체를 갖는 일배체형 상에 배치된 판독물들 중 하나를 제외한 모두가 더 짧은 일배체형과도 일치함을 보여준다. 상기 확장을 지원하는 불량하게 정렬된 판독물은 하나만 존재한다. 실제로 이것은 단일한 저품질 판독물로 인해 발생하는 위양성 호출일 가능성이 있는 것으로 간주될 것이다.

[0371] ***실시예 5: 과소추정된 반복체 크기***

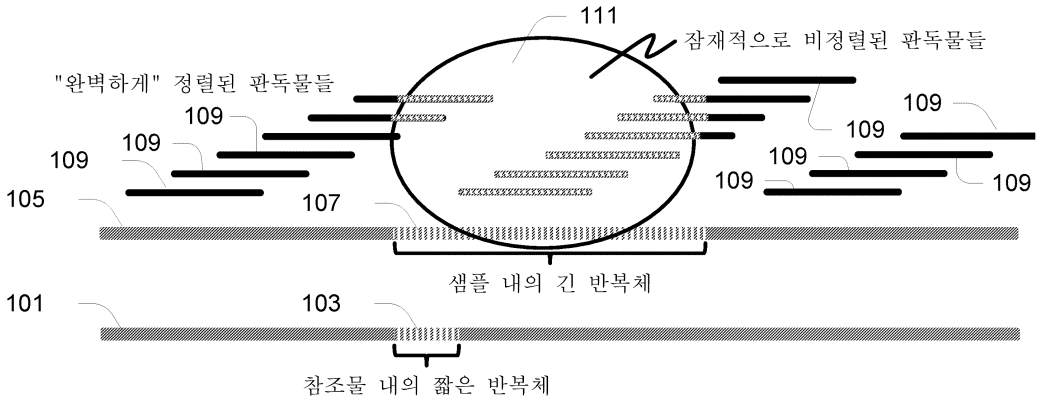
[0372] 위음성 반복 확장 호출의 실시예를 생성하기 위해, 실무자들은 유전자형 15/55를 갖는 *FMR1* 반복체를 시뮬레이션한 다음 (잘못된) 유전자형 15/30에 대응하는 판독물 파일업을 강제로 생성하였다. 도 23은 2304에서의 가장 긴 대립유전자의 크기가 과소추정되고 가장 짧은 대립유전자(2302)의 크기가 정확한 유전자형에 대응하는 *FMR1* 반복체 파일업을 도시한다. 크기 55의 반복체 내에서 발생하는 판독물들을 조정하기 위해, 일부 구현예들은 정

렬들의 말단들을 가장 긴 대립유전자의 크기로 잘랐음을 주목한다. 30개의 모티프들을 갖는 반복체와 중첩되는 과량의 판독물이 존재하고 이러한 모든 판독물들이 반복 서열로 구성되기 때문에, 실무자들은 반복체의 크기가 과소추정될 가능성이 있다고 결론을 내릴 수 있다.

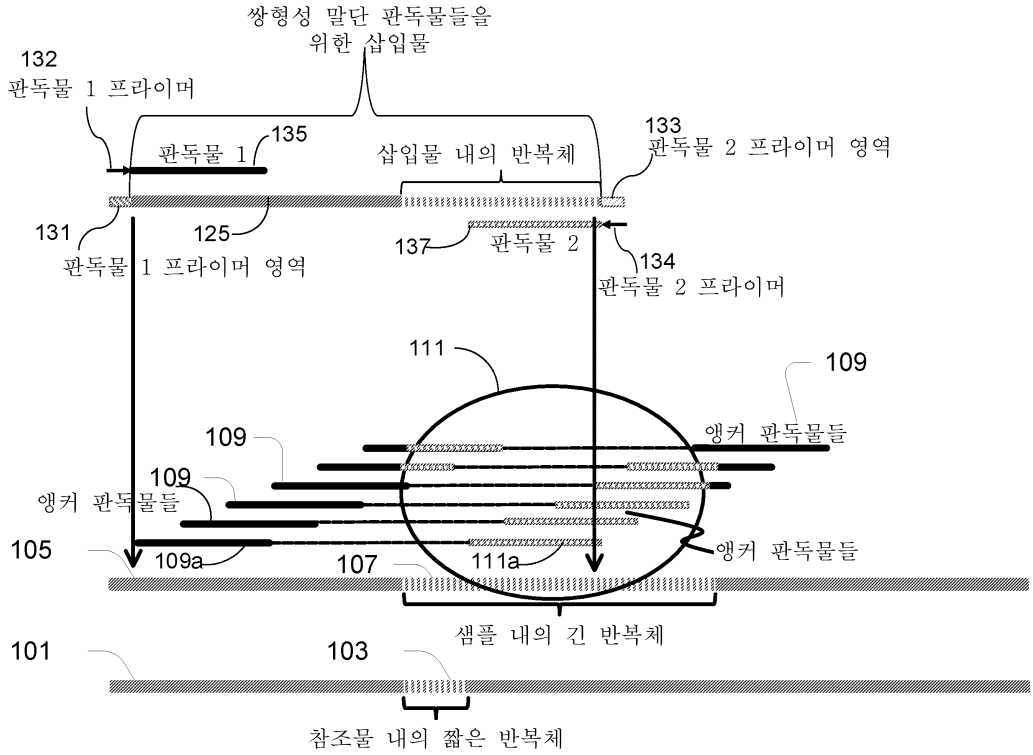
[0373] 본 개시내용은 이의 사상 또는 본질적인 특성들로부터 벗어나지 않으면서 다른 특정 형태들로 구현될 수 있다. 기술된 실시형태들은 모든 점에서 단지 예시적인 것으로 그리고 제한적이지 않은 것으로 고려되어야 한다. 따라서, 본 발명의 범위는 전술한 설명에 의해서라기보다는 첨부된 청구범위에 의해 설명된다. 청구범위의 등가물의 의미 및 범위 내에 있는 모든 변화들은 이의 범위 내에 포함되는 것으로 간주된다.

도면

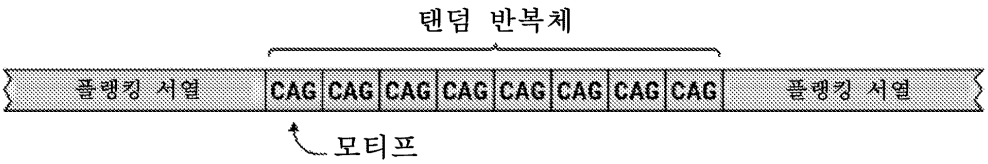
도면1a



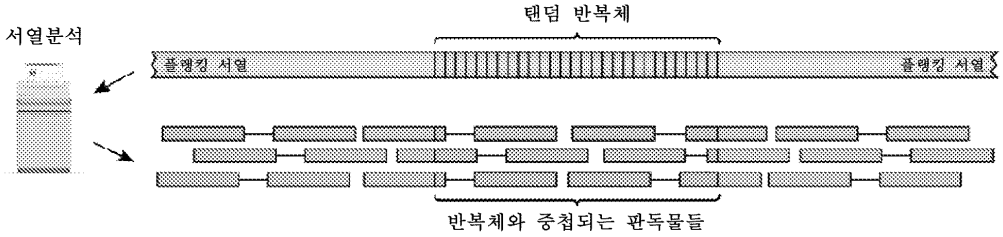
도면1b



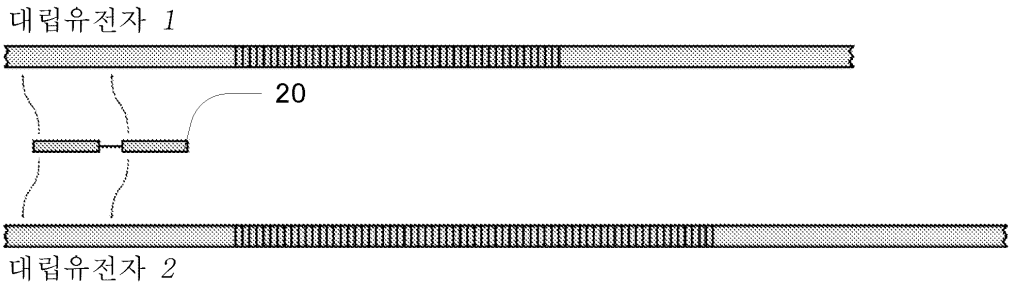
도면1c



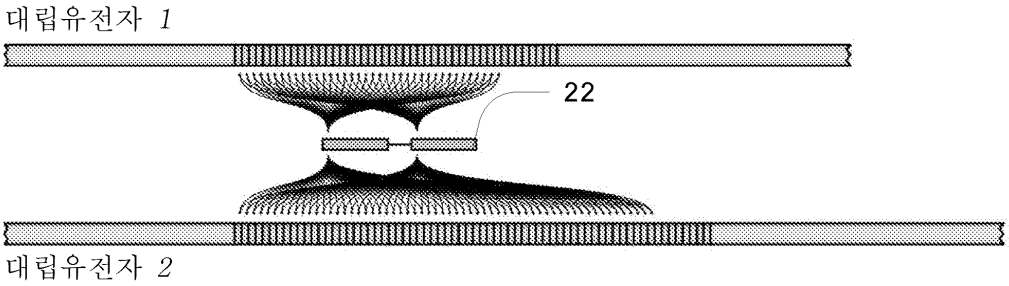
도면1d



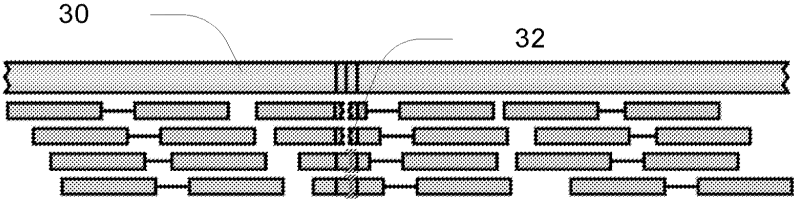
도면2a



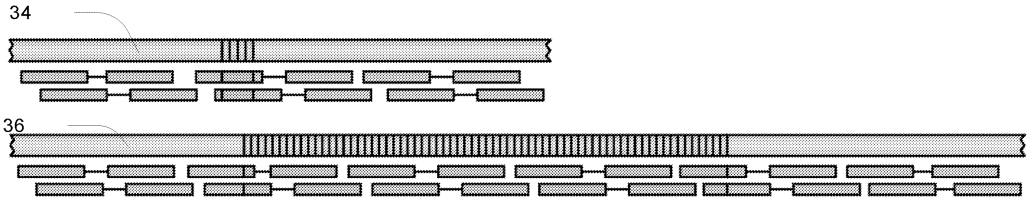
도면2b



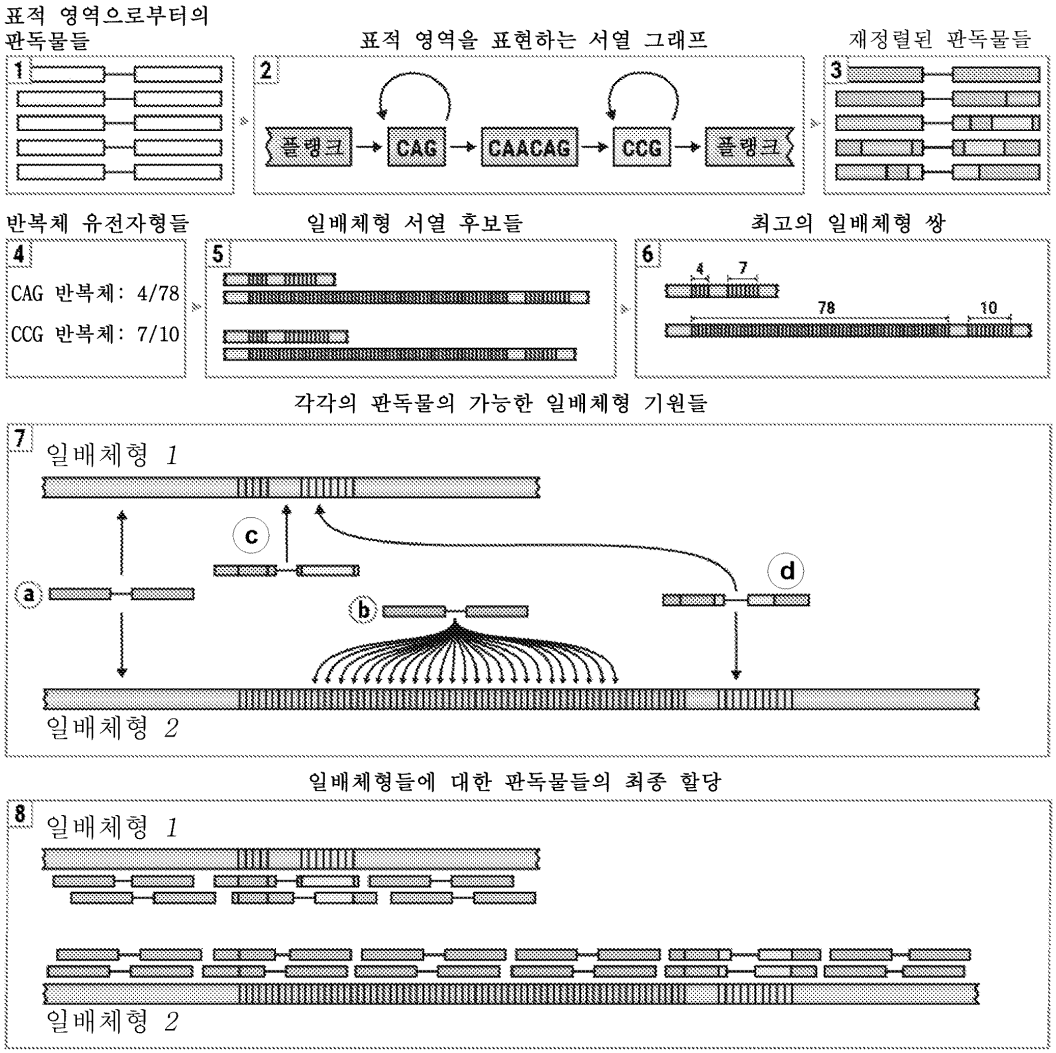
도면3a



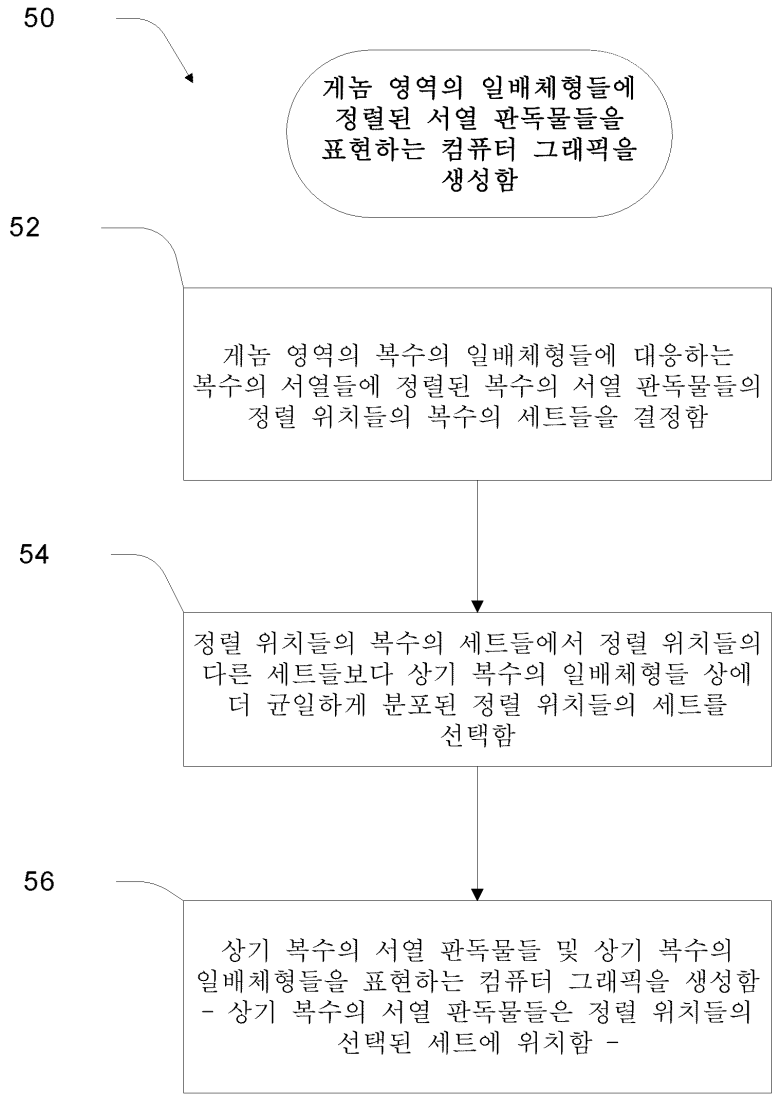
도면3b


34
36

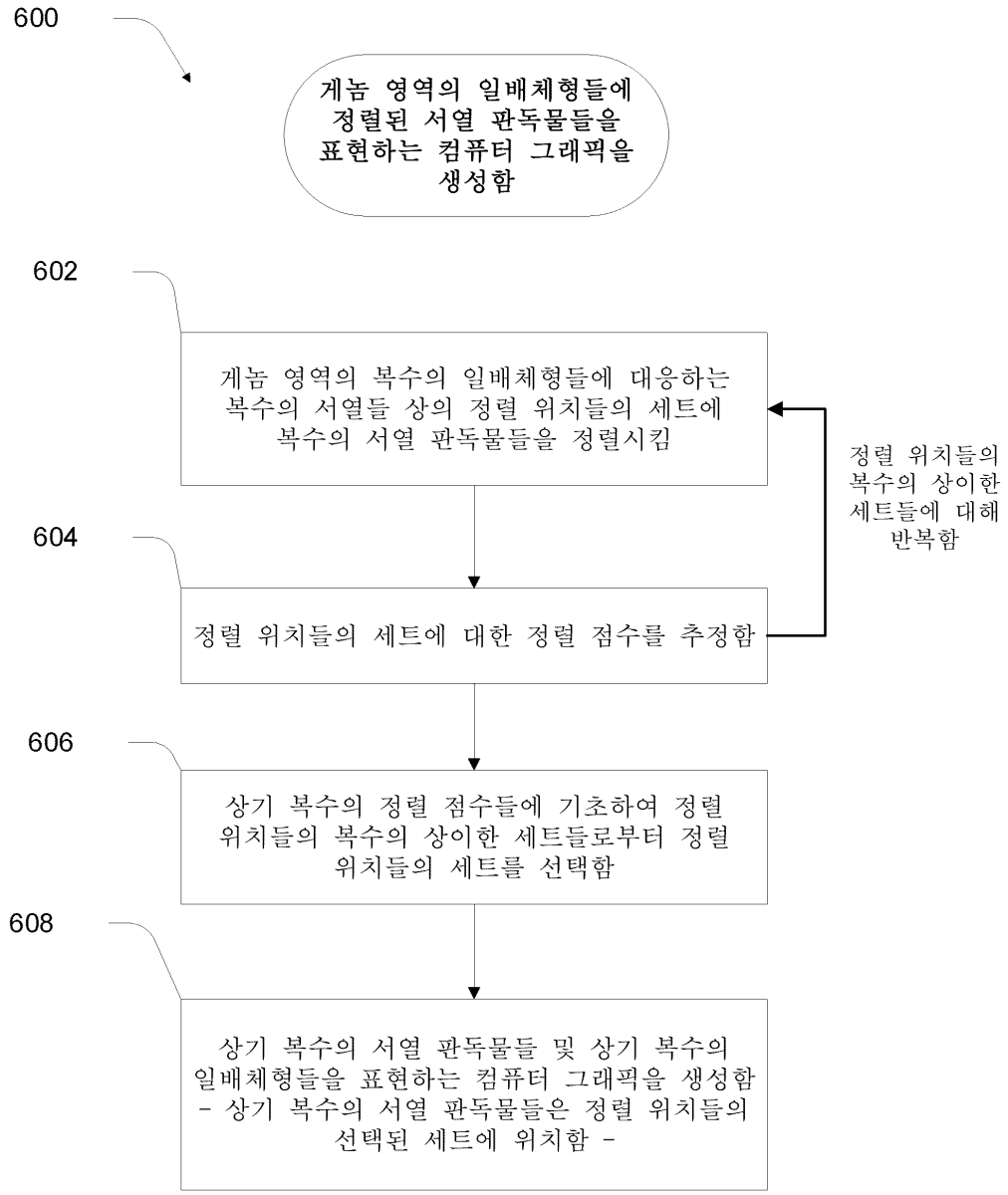
도면4


표적 영역으로부터의 판독물들
표적 영역을 표현하는 서열 그래프
재정렬된 판독물들
1
2
3
플랭크
CAG
CAACAG
CCG
플랭크
반복체 유전자형들
일배체형 서열 후보들
최고의 일배체형 쌍
4
CAG 반복체: 4/78
CCG 반복체: 7/10
5
6
4
7
78
10
각각의 판독물의 가능한 일배체형 기원들
7
일배체형 1
a
b
c
d
일배체형 2
일배체형들에 대한 판독물들의 최종 할당
8
일배체형 1
일배체형 2

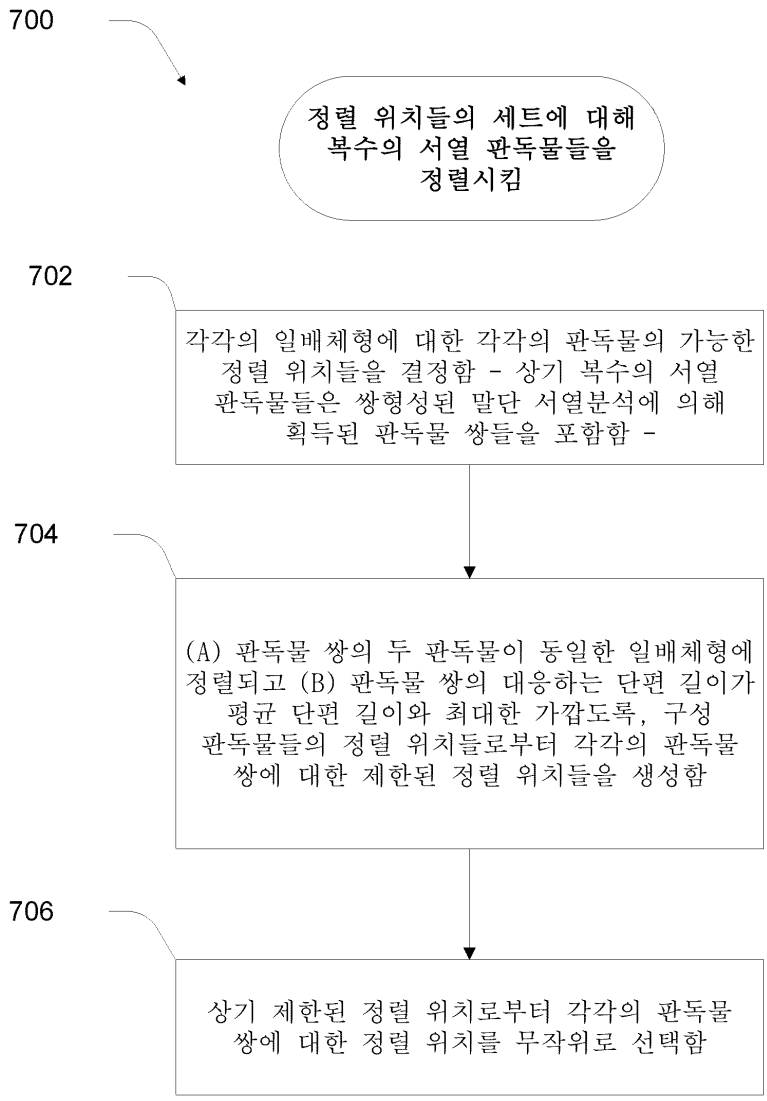
도면5


50
게놈 영역의 일배체형들에 정렬된 서열 판독물들을 표현하는 컴퓨터 그래픽을 생성함
52
게놈 영역의 복수의 일배체형들에 대응하는 복수의 서열들에 정렬된 복수의 서열 판독물들의 정렬 위치들의 복수의 세트들을 결정함
54
정렬 위치들의 복수의 세트들에서 정렬 위치들의 다른 세트들보다 상기 복수의 일배체형들 상에 더 균일하게 분포된 정렬 위치들의 세트를 선택함
56
상기 복수의 서열 판독물들 및 상기 복수의 일배체형들을 표현하는 컴퓨터 그래픽을 생성함 - 상기 복수의 서열 판독물들은 정렬 위치들의 선택된 세트에 위치함 -

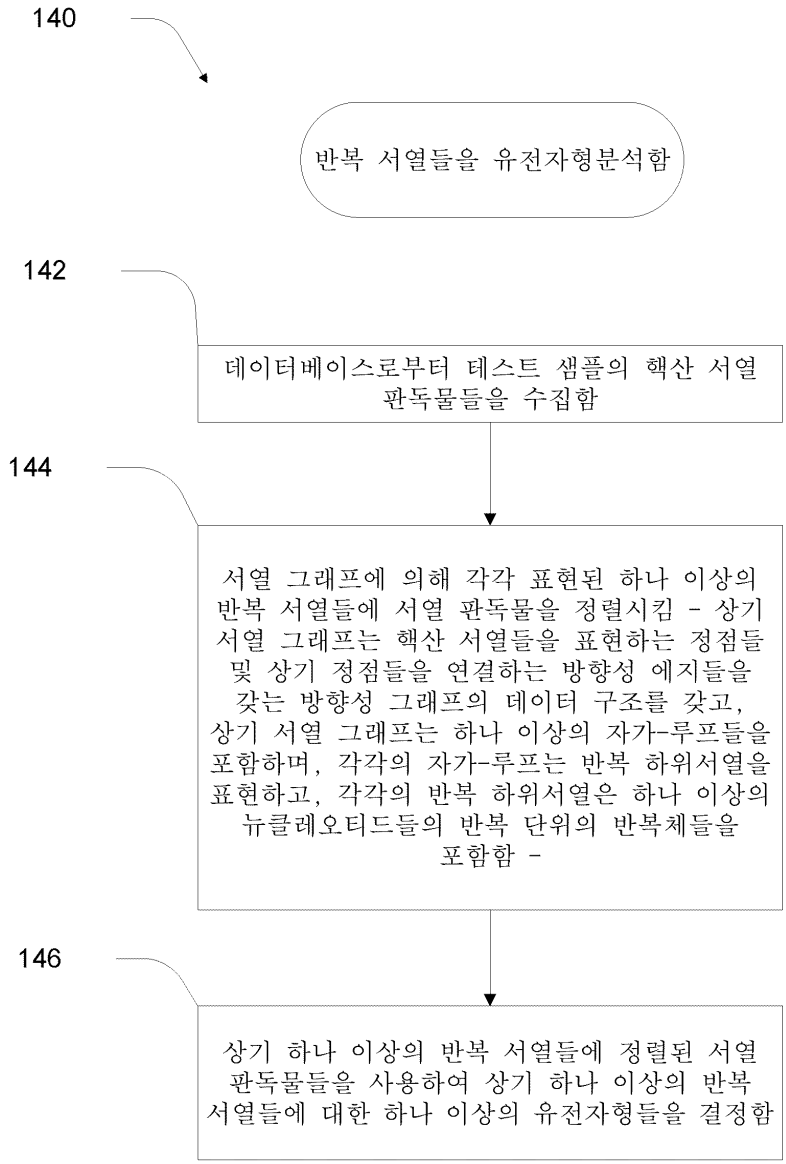
도면6


600
게놈 영역의 일배체형들에 정렬된 서열 판독물들을 표현하는 컴퓨터 그래픽을 생성함
602
게놈 영역의 복수의 일배체형들에 대응하는 복수의 서열들 상의 정렬 위치들의 세트에 복수의 서열 판독물들을 정렬시킴
604
정렬 위치들의 세트에 대한 정렬 점수를 추정함
정렬 위치들의 복수의 상이한 세트들에 대해 반복함
606
상기 복수의 정렬 점수들에 기초하여 정렬 위치들의 복수의 상이한 세트들로부터 정렬 위치들의 세트를 선택함
608
상기 복수의 서열 판독물들 및 상기 복수의 일배체형들을 표현하는 컴퓨터 그래픽을 생성함 - 상기 복수의 서열 판독물들은 정렬 위치들의 선택된 세트에 위치함 -

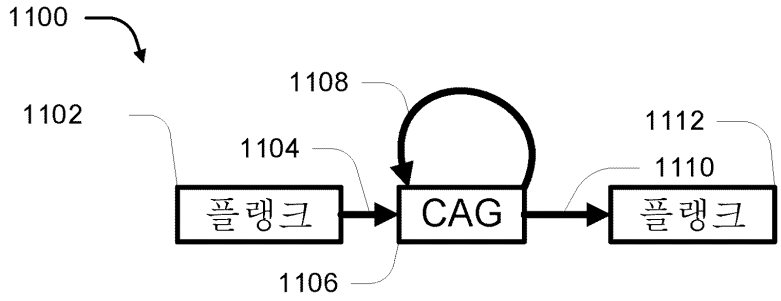
도면7


700
정렬 위치들의 세트에 대해 복수의 서열 판독물들을 정렬시킴
702
각각의 일배체형에 대한 각각의 판독물의 가능한 정렬 위치들을 결정함 - 상기 복수의 서열 판독물들은 쌍형성된 말단 서열분석에 의해 획득된 판독물 쌍들을 포함함 -
704
(A) 판독물 쌍의 두 판독물이 동일한 일배체형에 정렬되고 (B) 판독물 쌍의 대응하는 단편 길이가 평균 단편 길이와 최대한 가깝도록, 구성 판독물들의 정렬 위치들로부터 각각의 판독물 쌍에 대한 제한된 정렬 위치들을 생성함
706
상기 제한된 정렬 위치로부터 각각의 판독물 쌍에 대한 정렬 위치를 무작위로 선택함

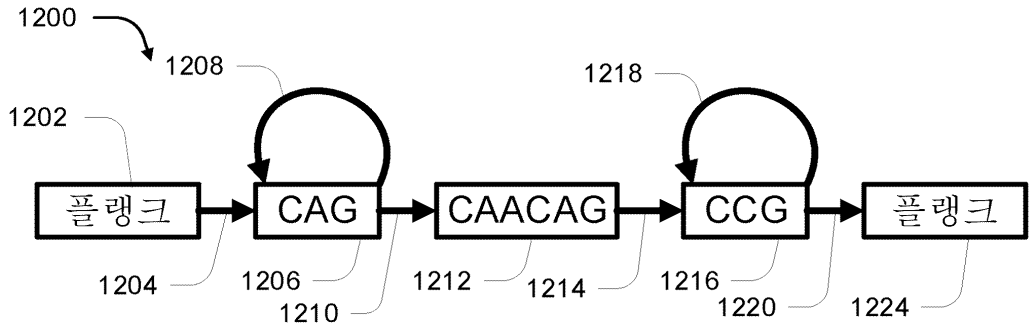
도면8


140
반복 서열들을 유전자형분석함
142
데이터베이스로부터 테스트 샘플의 핵산 서열 판독물들을 수집함
144
서열 그래프에 의해 각각 표현된 하나 이상의 반복 서열들에 서열 판독물을 정렬시킴 - 상기 서열 그래프는 핵산 서열들을 표현하는 정점들 및 상기 정점들을 연결하는 방향성 에지들을 갖는 방향성 그래프의 데이터 구조를 갖고, 상기 서열 그래프는 하나 이상의 자가-루프들을 포함하며, 각각의 자가-루프는 반복 하위서열을 표현하고, 각각의 반복 하위서열은 하나 이상의 뉴클레오티드들의 반복 단위의 반복체들을 포함함 -
146
상기 하나 이상의 반복 서열들에 정렬된 서열 판독물들을 사용하여 상기 하나 이상의 반복 서열들에 대한 하나 이상의 유전자형들을 결정함

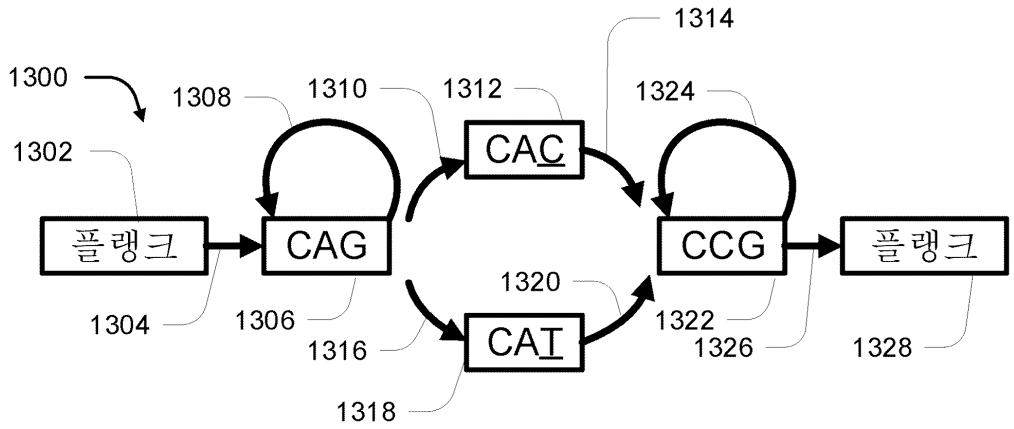
도면9


1100
1102
1104
1106
1108
1110
1112
플랭크
CAG
플랭크

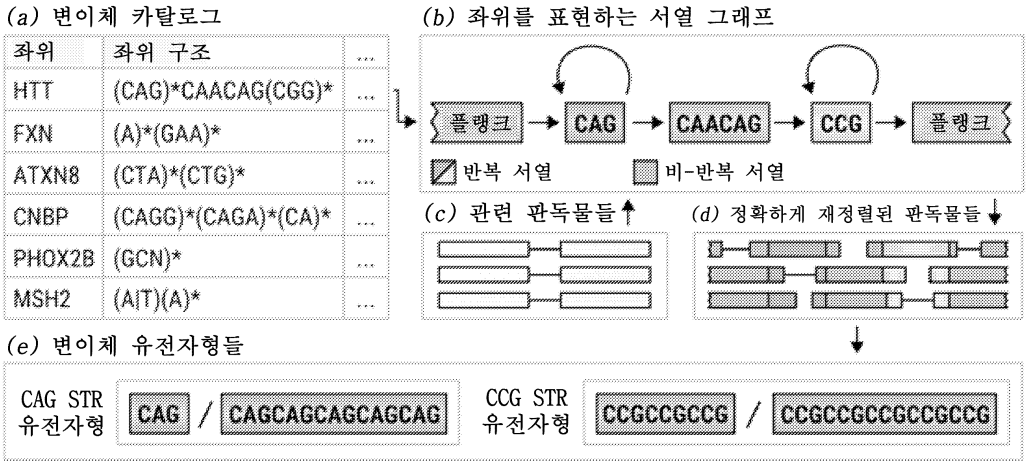
도면10



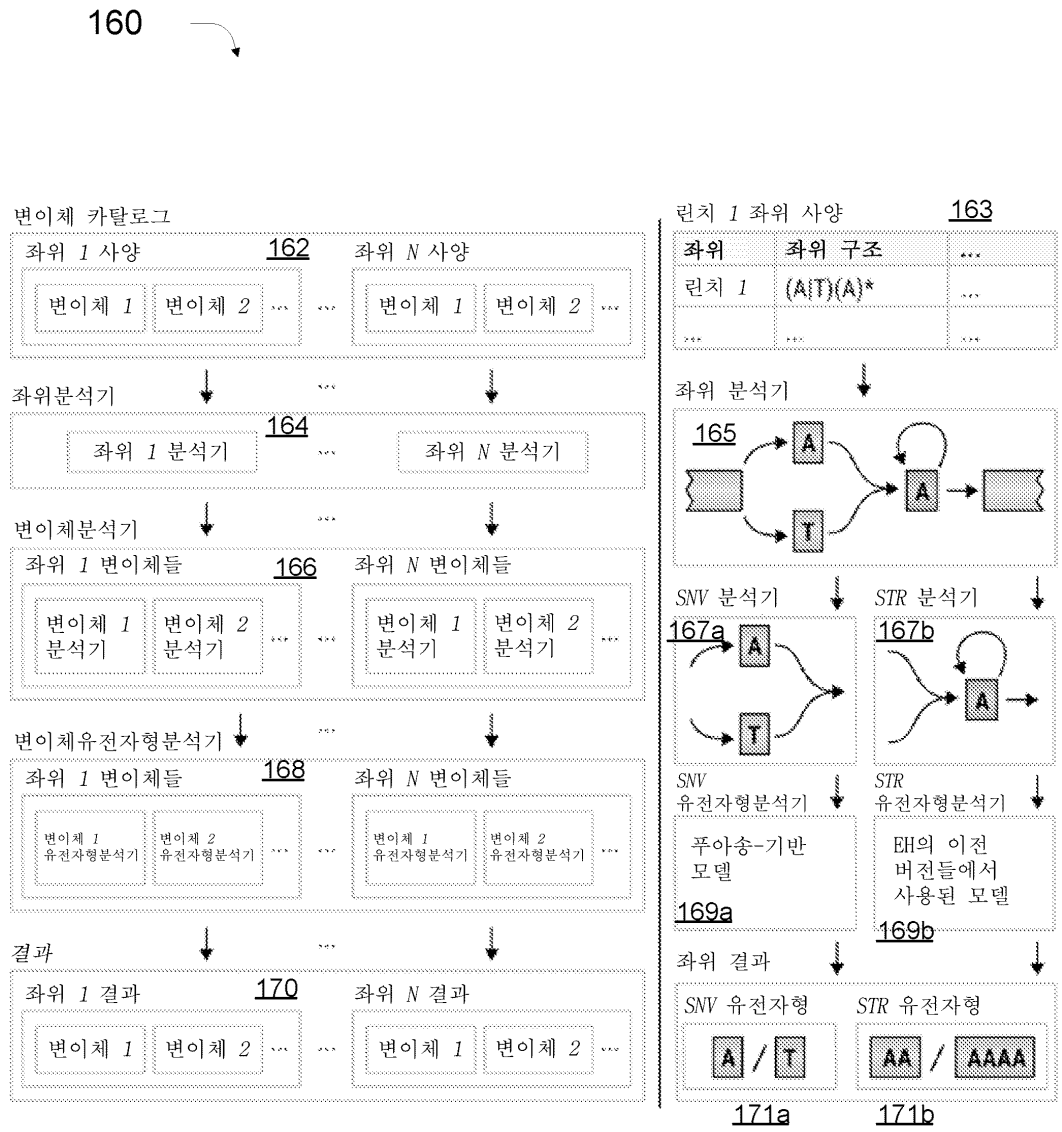
도면11



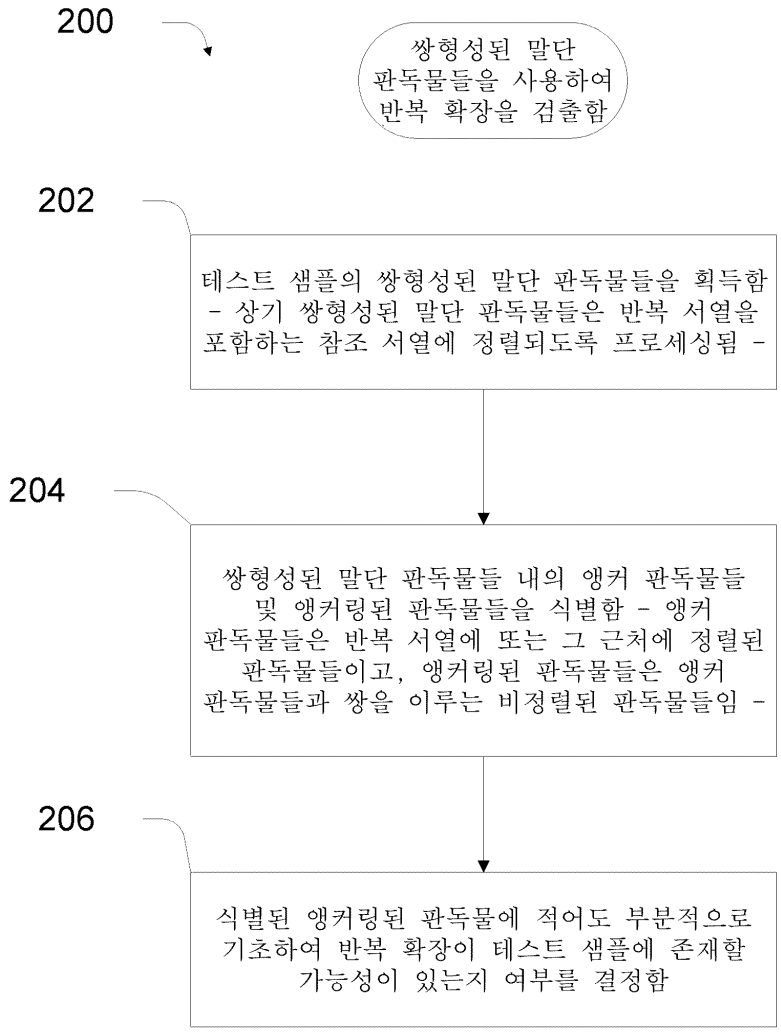
도면12



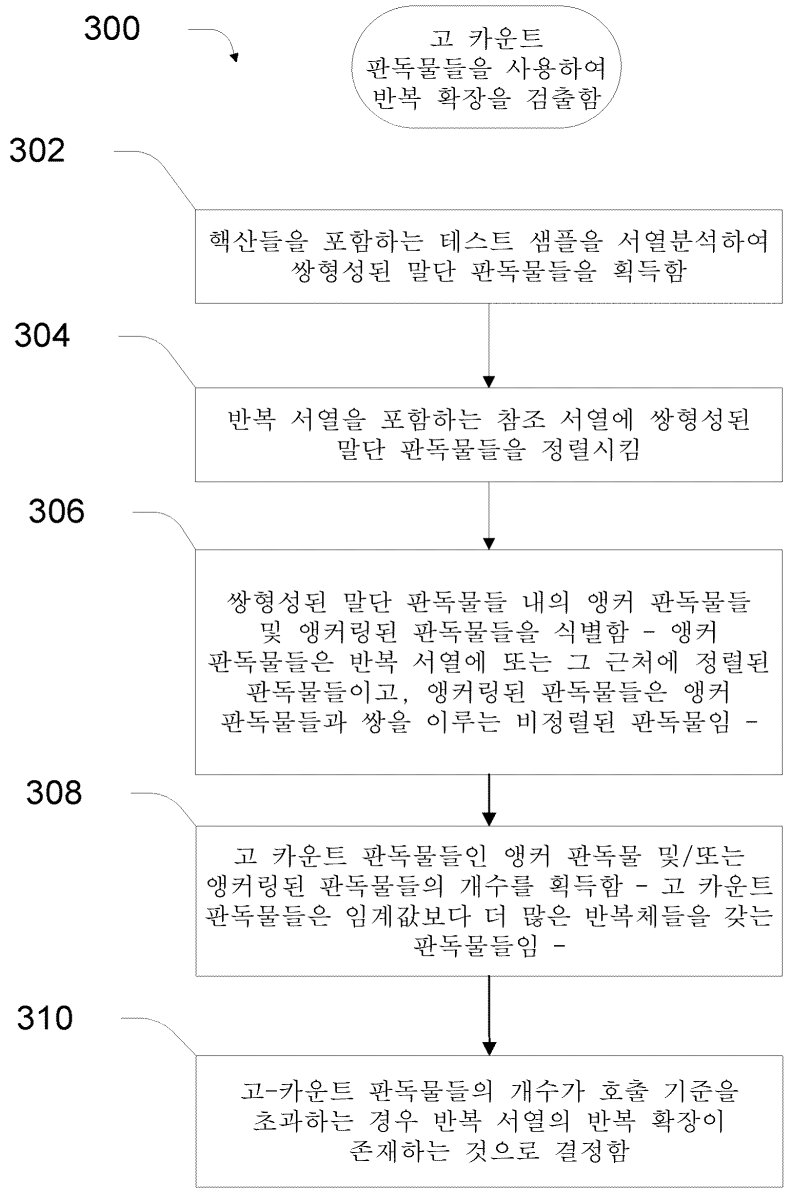
도면13


160
변이체 카탈로그
좌위 1 사양
162
변이체 1
변이체 2
좌위 N 사양
변이체 1
변이체 2
좌위분석기
164
좌위 1 분석기
좌위 N 분석기
변이체분석기
좌위 1 변이체들
166
변이체 1 분석기
변이체 2 분석기
좌위 N 변이체들
변이체 1 분석기
변이체 2 분석기
변이체유전자형분석기
좌위 1 변이체들
168
변이체 1 유전자형분석기
변이체 2 유전자형분석기
좌위 N 변이체들
변이체 1 유전자형분석기
변이체 2 유전자형분석기
결과
좌위 1 결과
170
변이체 1
변이체 2
좌위 N 결과
변이체 1
변이체 2
린치 1 좌위 사양
163
좌위
좌위 구조
린치 1
(A|T)(A)*
좌위 분석기
165
A
T
A
SNV 분석기
STR 분석기
167a
167b
A
T
A
SNV 유전자형분석기
STR 유전자형분석기
푸아송-기반 모델
169a
EH의 이전 버전들에서 사용된 모델
169b
좌위 결과
SNV 유전자형
STR 유전자형
A / T
AA / AAAA
171a
171b

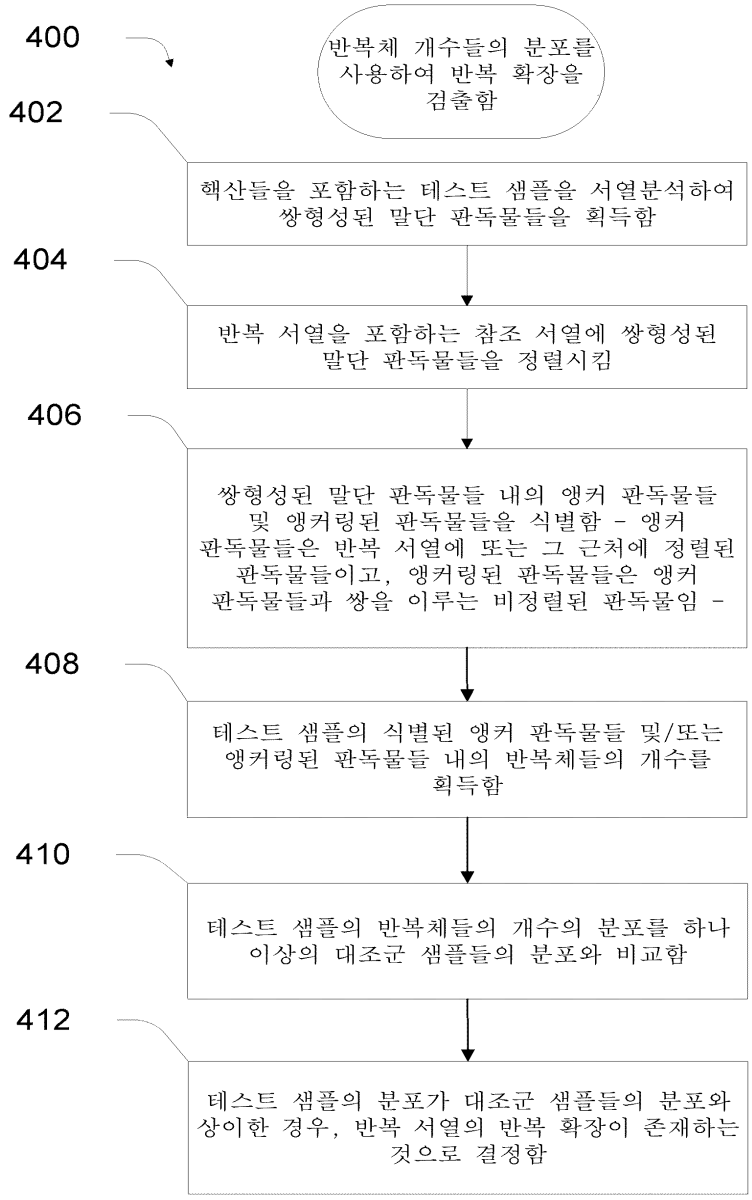
도면14


200
쌍형성된 말단 판독물들을 사용하여 반복 확장을 검출함
202
테스트 샘플의 쌍형성된 말단 판독물들을 획득함 - 상기 쌍형성된 말단 판독물들은 반복 서열을 포함하는 참조 서열에 정렬되도록 프로세싱됨 -
204
쌍형성된 말단 판독물들 내의 앵커 판독물들 및 앵커링된 판독물들을 식별함 - 앵커 판독물들은 반복 서열에 또는 그 근처에 정렬된 판독물들이고, 앵커링된 판독물들은 앵커 판독물들과 쌍을 이루는 비정렬된 판독물들임 -
206
식별된 앵커링된 판독물에 적어도 부분적으로 기초하여 반복 확장이 테스트 샘플에 존재할 가능성이 있는지 여부를 결정함

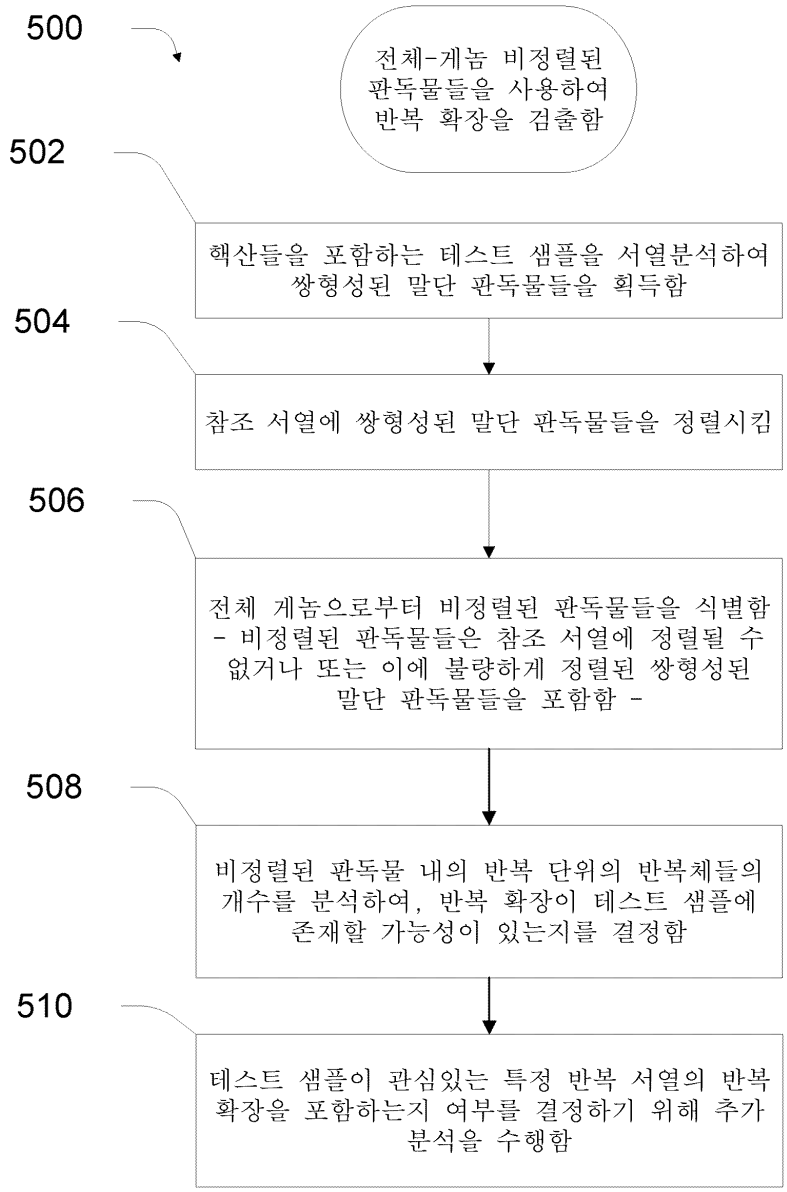
도면15


300
고 카운트 판독물들을 사용하여 반복 확장을 검출함
302
핵산들을 포함하는 테스트 샘플을 서열분석하여 쌍형성된 말단 판독물들을 획득함
304
반복 서열을 포함하는 참조 서열에 쌍형성된 말단 판독물들을 정렬시킴
306
쌍형성된 말단 판독물들 내의 앵커 판독물들 및 앵커링된 판독물들을 식별함 - 앵커 판독물들은 반복 서열에 또는 그 근처에 정렬된 판독물들이고, 앵커링된 판독물들은 앵커 판독물들과 쌍을 이루는 비정렬된 판독물임 -
308
고 카운트 판독물들인 앵커 판독물 및/또는 앵커링된 판독물들의 개수를 획득함 - 고 카운트 판독물들은 임계값보다 더 많은 반복체들을 갖는 판독물들임 -
310
고-카운트 판독물들의 개수가 호출 기준을 초과하는 경우 반복 서열의 반복 확장이 존재하는 것으로 결정함

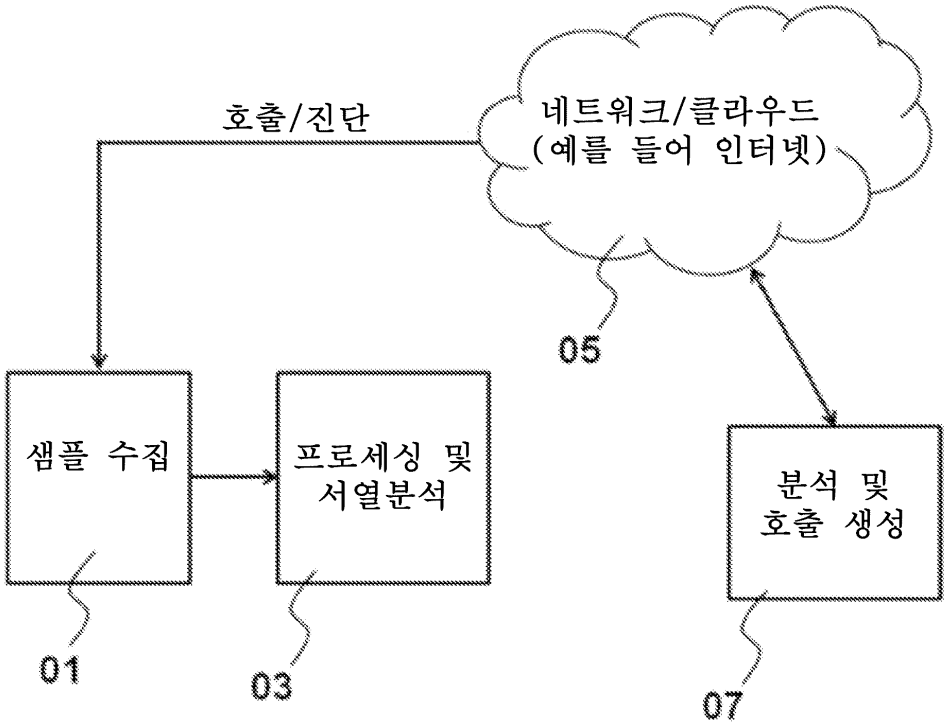
도면16


400
반복체 개수들의 분포를 사용하여 반복 확장을 검출함
402
핵산들을 포함하는 테스트 샘플을 서열분석하여 쌍형성된 말단 판독물들을 획득함
404
반복 서열을 포함하는 참조 서열에 쌍형성된 말단 판독물들을 정렬시킴
406
쌍형성된 말단 판독물들 내의 앵커 판독물들 및 앵커링된 판독물들을 식별함 - 앵커 판독물들은 반복 서열에 또는 그 근처에 정렬된 판독물들이고, 앵커링된 판독물들은 앵커 판독물들과 쌍을 이루는 비정렬된 판독물임 -
408
테스트 샘플의 식별된 앵커 판독물들 및/또는 앵커링된 판독물들 내의 반복체들의 개수를 획득함
410
테스트 샘플의 반복체들의 개수의 분포를 하나 이상의 대조군 샘플들의 분포와 비교함
412
테스트 샘플의 분포가 대조군 샘플들의 분포와 상이한 경우, 반복 서열의 반복 확장이 존재하는 것으로 결정함

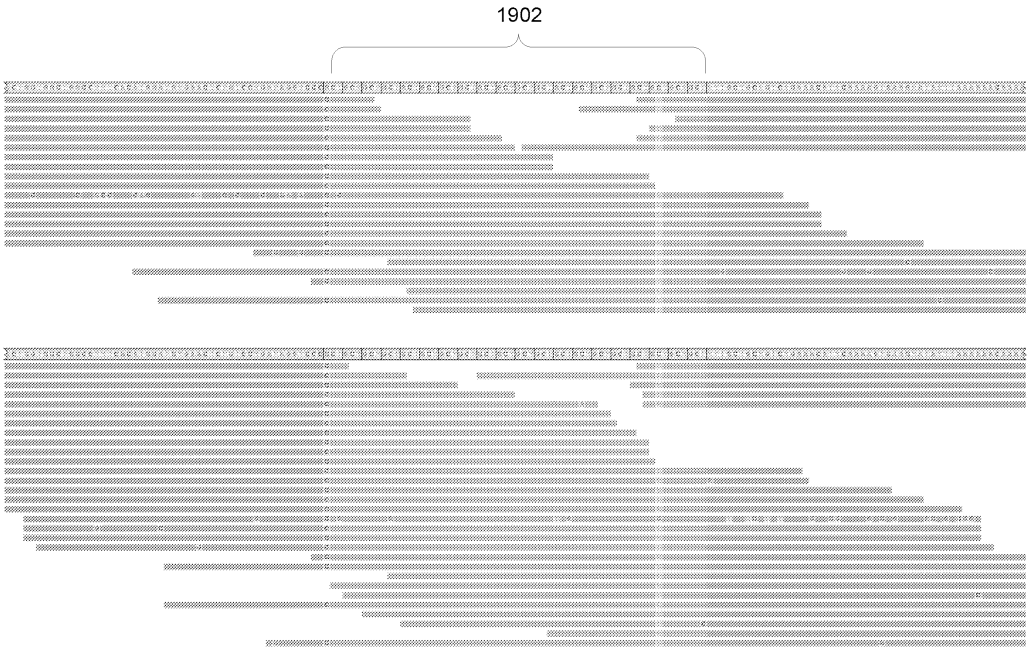
도면17


500
전체-게놈 비정렬된 판독물들을 사용하여 반복 확장을 검출함
502
핵산들을 포함하는 테스트 샘플을 서열분석하여 쌍형성된 말단 판독물들을 획득함
504
참조 서열에 쌍형성된 말단 판독물들을 정렬시킴
506
전체 게놈으로부터 비정렬된 판독물들을 식별함 - 비정렬된 판독물들은 참조 서열에 정렬될 수 없거나 또는 이에 불량하게 정렬된 쌍형성된 말단 판독물들을 포함함 -
508
비정렬된 판독물 내의 반복 단위의 반복체들의 개수를 분석하여, 반복 확장이 테스트 샘플에 존재할 가능성이 있는지를 결정함
510
테스트 샘플이 관심있는 특정 반복 서열의 반복 확장을 포함하는지 여부를 결정하기 위해 추가 분석을 수행함

도면18


호출/진단
네트워크/클라우드
(예를 들어 인터넷)
05
샘플 수집
프로세싱 및
서열분석
분석 및
호출 생성
01
03
07

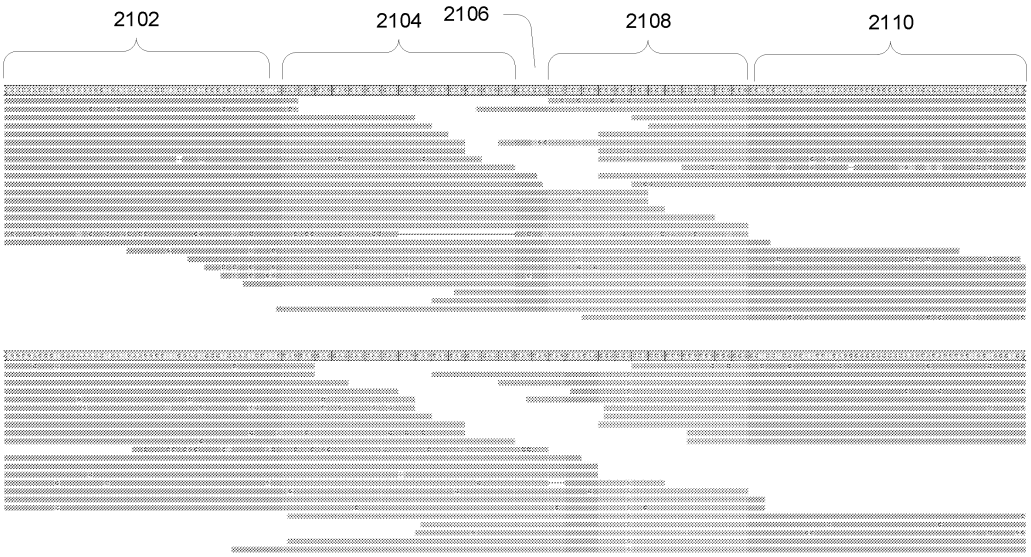
도면19


1902

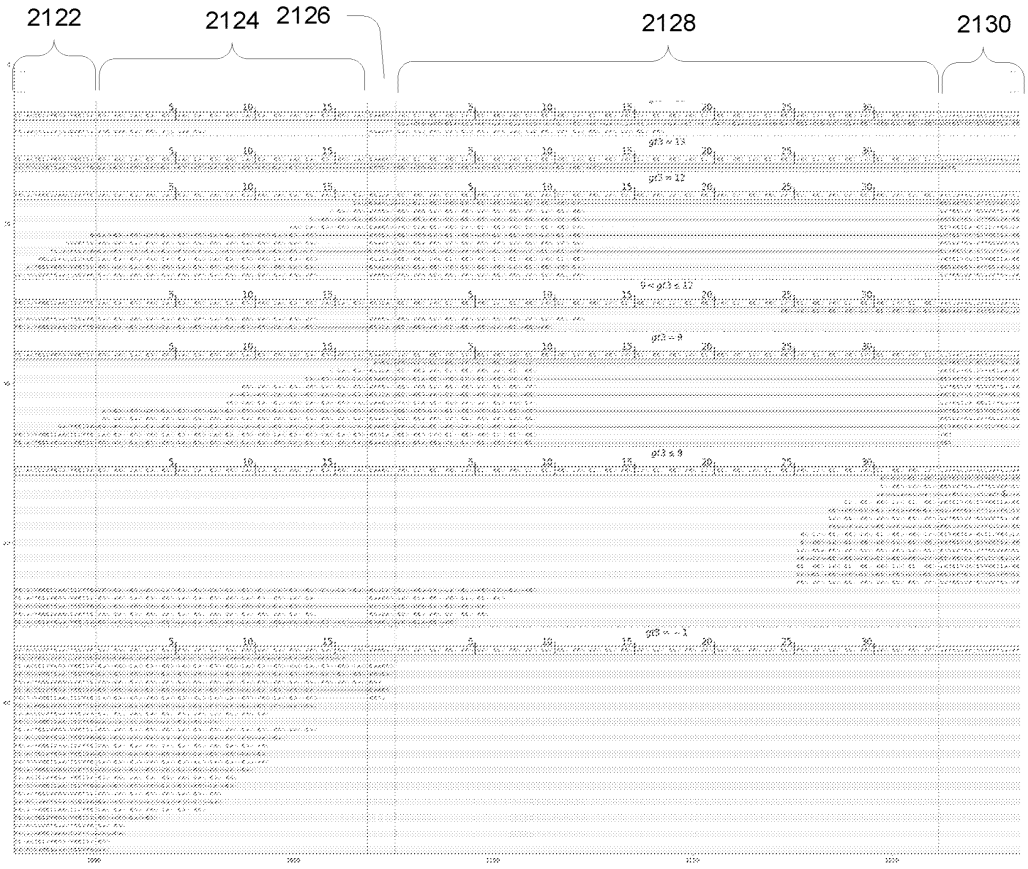
도면20



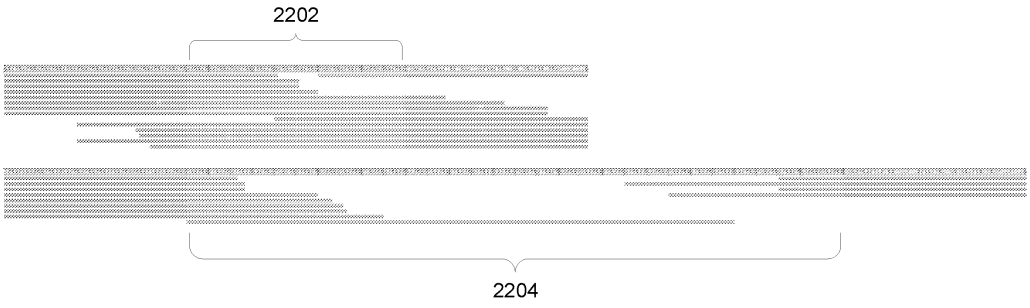
도면21a



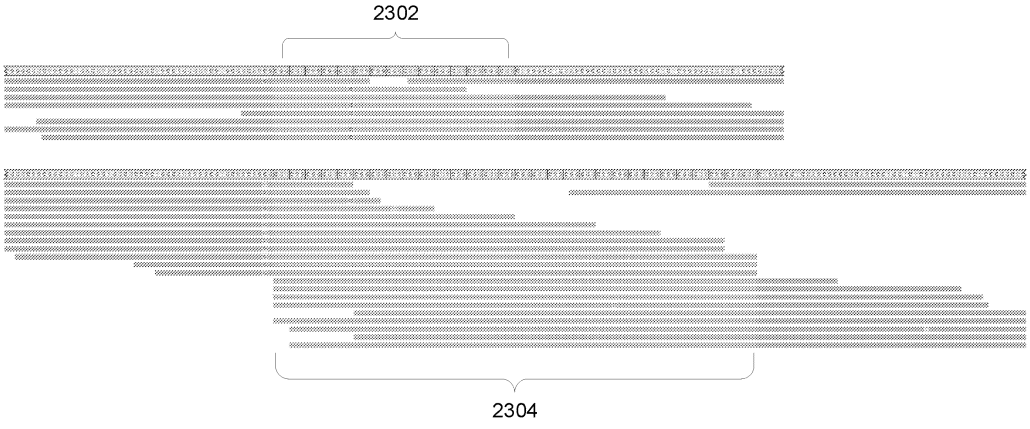
도면21b



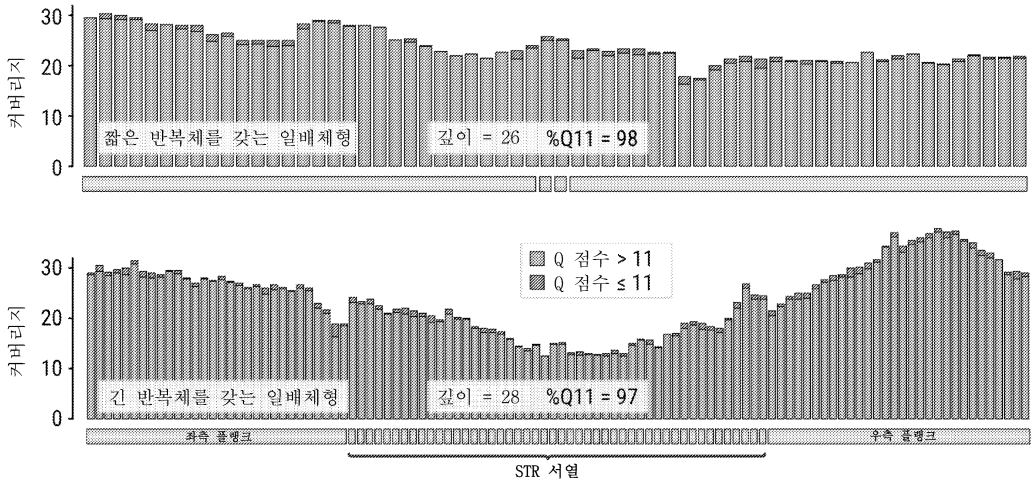
도면22



도면23



도면24


커버리지
30
20
10
0
짧은 반복체를 갖는 일배체형
깊이 = 26
%Q11 = 98
Q 점수 > 11
Q 점수 ≤ 11
긴 반복체를 갖는 일배체형
깊이 = 28
%Q11 = 97
좌측 플랭크
우측 플랭크
STR 서열