(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0093202
(43) 공개일자 2020년08월05일

(51) 국제특허분류(Int. Cl.)
*C12Q 1/6876* (2018.01) *C12Q 1/6811* (2018.01)
*C12Q 1/6813* (2018.01) *C12Q 1/6869* (2018.01)
*G16B 20/00* (2019.01) *G16B 25/10* (2019.01)
*G16B 30/10* (2019.01)
(52) CPC특허분류
*C12Q 1/6876* (2018.05)
*C12Q 1/6811* (2018.05)
(21) 출원번호 10-2019-0010386
(22) 출원일자 2019년01월28일
심사청구일자 2019년01월28일

(71) 출원인
**연세대학교 산학협력단**
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
**경기대학교 산학협력단**
경기도 수원시 영통구 광교산로 154-42 (이의동, 경기대학교)
(72) 발명자
**김상우**
서울특별시 마포구 신촌로 170, 917호(대흥동)
**양인석**
경기도 화성시 10용사로 221, 105동 1004호(병점동, 성호1차아파트)
**배상원**
경기도 수원시 영통구 센트럴타운로 76, 6118동 3702호(이의동, e편한세상 광교)
(74) 대리인
**특허법인인벤싱크**

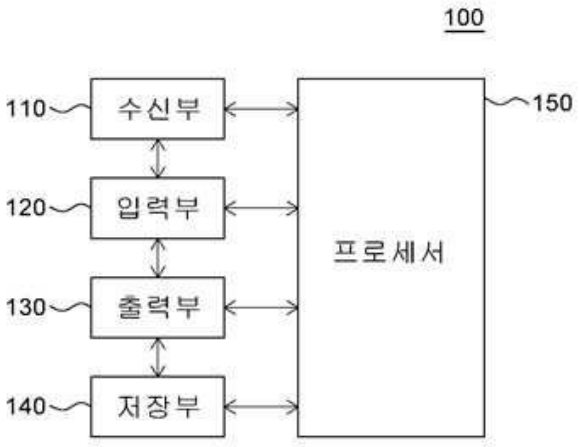
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 **폴리뉴클레오티드 바코드 세트 및 이의 제공 방법**

## (57) 요 약

본 명세서에서는, 복수개의 폴리뉴클레오티드 바코드를 포함하고, 복수개의 폴리뉴클레오티드 바코드 각각이 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 만족하고, 특징 조건이 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이, 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수, 및 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 중 적어도 하나인, 폴리뉴클레오티드 바코드 세트 및 이의 제공 방법이 제공된다.

*대 표 도* - 도1a



(52) CPC특허분류
***C12Q 1/6813*** (2018.05)
***C12Q 1/6869*** (2018.05)
***G16B 20/00*** (2019.02)
***G16B 25/10*** (2019.02)
***G16B 30/10*** (2019.02)
*C12Q 2525/204* (2013.01)
*C12Q 2535/122* (2019.08)
*C12Q 2563/185* (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 HI14C1324
부처명 보건복지부
연구관리전문기관 한국보건산업진흥원
연구사업명 연구중심병원육성 R&D
연구과제명 액체 생검 플랫폼을 이용한 임상 테스트 개발
기 여 율 1/1
주관기관 연세대학교 산학협력단
연구기간 2018.01.01 ~ 2018.12.31

# 명 세 서

## 청구범위

### 청구항 1

복수개의 폴리뉴클레오티드 바코드를 포함하고,

상기 복수개의 폴리뉴클레오티드 바코드 각각은, 상기 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 만족하고,

상기 특징 조건은,

상기 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 상기 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이, 상기 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛, 상기 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수, 및 상기 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 중 적어도 하나인, 폴리뉴클레오티드 바코드 세트.

### 청구항 2

제1항에 있어서,

상기 복수개의 폴리뉴클레오티드 바코드 각각은,

상기 특징 조건을 기초로 결정된 특징 조건 점수를 더 만족하고,

상기 특징 조건 점수는,

상기 폴리뉴클레오티드 바코드의 서열 내의 GC 함량 점수, 상기 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이 점수, 상기 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛 점수, 상기 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수 점수, 및 상기 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 점수 중 적어도 하나인, 폴리뉴클레오티드 바코드 세트.

### 청구항 3

제2항에 있어서,

상기 특징 조건 점수는, 상기 GC 함량 점수, 상기 반복 싱글 뉴클레오티드의 길이 점수, 상기 디뉴클레오티드의 반복 유닛 점수, 상기 뉴클레오티드 쌍 점수, 및 상기 상보적 뉴클레오티드 결합 쌍 점수 중 선택된 두 개 이상의 특징 조건 점수를 합산한 합산 특징 조건 점수를 포함하는, 폴리뉴클레오티드 바코드 세트.

### 청구항 4

제1항에 있어서,

상기 폴리뉴클레오티드 바코드의 서열 길이는 2 내지 25nt인, 폴리뉴클레오티드 바코드 세트.

### 청구항 5

프로세서에 의해 구현되는, 복수개의 폴리뉴클레오티드 바코드를 포함하는 폴리뉴클레오티드 바코드 세트의 제공 방법으로서,

미리 결정된 뉴클레오티드 길이에 따라 무작위로 생성된 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 수신하는 단계,

상기 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 이용하여 평가하는 단계, 및

상기 평가 결과를 기초로 추천 폴리뉴클레오티드 바코드 세트를 결정하는 단계를 포함하고,

상기 특징 조건은,

상기 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 상기 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이, 상기 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛, 상기 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수, 및 상기 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 중 적어도 하나인, 폴리뉴클레오티드 바코드 세트의 제공 방법.

**청구항 6**

제5항에 있어서,

상기 평가하는 단계는,

상기 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 상기 특징 조건을 기초로 결정된 특징 조건 점수를 산출하는 단계, 및

상기 특징 조건 점수를 기초로 상기 복수개의 폴리뉴클레오티드 바코드를 포함하는 상기 바코드 세트를 평가하는 단계를 포함하고,

상기 특징 조건 점수는,

상기 폴리뉴클레오티드 바코드의 서열 내의 GC 함량 점수, 상기 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이 점수, 상기 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛 점수, 상기 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수 점수, 및 상기 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 점수 중 적어도 하나인 폴리뉴클레오티드 바코드 세트의 제공 방법.

**청구항 7**

제6항에 있어서,

상기 평가하는 단계는,

상기 특징 조건 점수 산출하는 단계, 및

상기 복수개의 폴리뉴클레오티드 바코드 각각에 대한 특징 조건 점수가 일정한 수준에 수렴하는지 여부를 평가하는 단계를 더 포함하고,

상기 특징 조건 점수가 상기 일정한 수준에 수렴하지 않을 경우,

상기 복수개의 폴리뉴클레오티드 바코드 각각에 대한 특징 조건 점수를 기초로 폴리뉴클레오티드 바코드를 필터링하는 단계, 및

추가 폴리뉴클레오티드 바코드를 더 수신하는 단계를 더 포함하는, 폴리뉴클레오티드 바코드 세트의 제공 방법.

**청구항 8**

제7항에 있어서,

상기 폴리뉴클레오티드 바코드 세트는,

미리 결정된 개수의 복수개의 폴리뉴클레오티드를 포함하고,

상기 필터링하는 단계는,

미리 결정된 필터링률, 및 상기 복수개의 폴리뉴클레오티드 바코드 각각에 대한 특징 조건 점수를 기초로 폴리뉴클레오티드 바코드를 필터링하는 단계를 포함하고,

상기 추가 폴리뉴클레오티드 바코드를 더 수신하는 단계는,

필터링된 상기 폴리뉴클레오티드 바코드와 동일한 개수의 추가 폴리뉴클레오티드 바코드를 수신하는 단계를 포

함하는, 폴리뉴클레오티드 바코드 세트의 제공 방법.

**청구항 9**

제7항에 있어서,

상기 특징 조건 점수가 일정한 수준에 수렴하는 경우,

상기 추천 폴리뉴클레오티드 바코드 세트를 결정하는 단계는,

상기 일정한 수준에 수렴하는 복수개의 폴리뉴클레오티드 바코드를 상기 추천 폴리뉴클레오티드 바코드 세트로 결정하는 단계를 포함하는, 폴리뉴클레오티드 바코드 세트의 제공 방법.

**청구항 10**

제6항에 있어서,

상기 특징 조건 점수는,

상기 GC 함량 점수, 상기 반복 싱글 뉴클레오티드의 길이 점수, 상기 디뉴클레오티드의 반복 유닛 점수, 상기 뉴클레오티드 쌍 점수, 및 상기 상보적 뉴클레오티드 결합 쌍 점수 중 선택된 적어도 하나의 특징 조건 점수를 합산한 합산 특징 조건 점수를 포함하고,

상기 평가하는 단계는,

상기 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 상기 특징 조건 점수를 산출하는 단계,

상기 적어도 하나의 특징 조건 점수를 합산하여 합산 특징 조건 점수를 산출하는 단계, 및

상기 합산 특징 조건 점수를 기초로 상기 복수개의 폴리뉴클레오티드 바코드를 포함하는 상기 바코드 세트를 평가하는 단계를 포함하는, 폴리뉴클레오티드 바코드 세트의 제공 방법.

**청구항 11**

제10항에 있어서,

상기 추천 폴리뉴클레오티드 바코드 세트 및 상기 추천 폴리뉴클레오티드 바코드 세트에 대한 상기 합산 특징 조건 점수를 제공하는 단계를 더 포함하는, 폴리뉴클레오티드 바코드 세트 제공 방법.

**청구항 12**

제6항에 있어서,

상기 산출하는 단계는,

상기 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 상기 특징 조건 점수에 대하여 미리 결정된 가중치를 기초로 상기 특징 조건 점수를 산출하는 단계를 포함하는, 폴리뉴클레오티드 바코드 세트 제공 방법.

**청구항 13**

프로세서에 의해 구현되는, 복수개의 폴리뉴클레오티드 바코드를 포함하는 폴리뉴클레오티드 바코드 세트의 제공용 디바이스로서,

미리 결정된 뉴클레오티드 길이에 따라 무작위로 생성된 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 수신하도록 구성된 수신부, 및

상기 수신부와 통신하도록 연결된 프로세서를 포함하고,

상기 프로세서는, 상기 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 이용하여 평가하고, 상기 평가 결과를 기초로 추천 폴리뉴클레오티드 바코드 세트를 결정하도록 구성되고,

상기 특징 조건은,

상기 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 상기 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복

되는, 반복 싱글 뉴클레오티드의 길이, 상기 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛, 상기 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수, 및 상기 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 중 적어도 하나인, 폴리뉴클레오티드 바코드 세트의 제공용 디바이스. 폴리뉴클레오티드 바코드 세트의 제공용 디바이스.

**청구항 14**

제13항에 있어서,

상기 프로세서는,

상기 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 상기 특징 조건을 기초로 결정된 특징 조건 점수를 산출하고, 상기 특징 조건 점수를 기초로 상기 복수개의 폴리뉴클레오티드 바코드를 포함하는 상기 바코드 세트를 평가하도록 더 구성되고,

상기 특징 조건 점수는,

상기 폴리뉴클레오티드 바코드의 서열 내의 GC 함량 점수, 상기 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이 점수, 상기 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛 점수, 상기 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수 점수, 및 상기 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 점수 중 적어도 하나인, 폴리뉴클레오티드 바코드 세트의 제공용 디바이스

**청구항 15**

제14항에 있어서,

상기 특징 조건 점수는,

상기 GC 함량 점수, 상기 반복 싱글 뉴클레오티드의 길이 점수, 상기 디뉴클레오티드의 반복 유닛 점수, 상기 뉴클레오티드 쌍 점수, 및 상기 상보적 뉴클레오티드 결합 쌍 점수 중 선택된 적어도 하나의 상기 특징 조건 점수를 합산한 합산 특징 조건 점수를 포함하고,

상기 프로세서는,

상기 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 상기 특징 조건 점수를 산출하고, 상기 적어도 하나의 특징 조건 점수를 합산하여 합산 특징 조건 점수를 산출하고, 상기 합산 특징 조건 점수를 기초로 상기 복수개의 폴리뉴클레오티드 바코드를 포함하는 상기 바코드 세트를 평가하도록 더 구성된, 폴리뉴클레오티드 바코드 세트의 제공용 디바이스.

**청구항 16**

제15항에 있어서,

상기 프로세서는,

상기 추천 폴리뉴클레오티드 바코드 세트 및 상기 추천 폴리뉴클레오티드 바코드 세트에 대한 상기 합산 특징 조건 점수를 제공하도록 더 구성된, 폴리뉴클레오티드 바코드 세트의 제공용 디바이스.

**청구항 17**

제14항에 있어서,

상기 프로세서는,

상기 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 산출된 상기 특징 조건 점수가 일정한 수준에 수렴하는지 여부를 평가하고, 상기 특징 조건 점수가 일정한 수준에 수렴하지 않을 경우 상기 복수개의 폴리뉴클레오티드 바코드 각각에 대한 특징 조건 점수를 기초로 폴리뉴클레오티드 바코드를 필터링하도록 더 구성되고,

상기 수신부는,

상기 프로세서에 의해 상기 특징 조건 점수가 일정한 수준에 수렴하지 않을 경우, 추가 폴리뉴클레오티드 바코드를 더 수신하도록 구성된, 폴리뉴클레오티드 바코드 세트의 제공용 디바이스.

**청구항 18**

제14항에 있어서,

상기 프로세서는,

상기 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 상기 특징 조건 점수에 대하여 미리 결정된 가중치를 기초로 상기 특징 조건 점수를 산출하는 더 구성된, 폴리뉴클레오티드 바코드 세트의 제공용 디바이스.

## *발명의 설명*

### *기 술 분 야*

[0001] 본 발명은 폴리뉴클레오티드 바코드 세트 및 이의 제공 방법에 관한 것으로, 보다 구체적으로 폴리뉴클레오티드 바코드의 서열 특징을 기초로 결정된 특징 조건을 만족하는 폴리뉴클레오티드 바코드 세트 및 이의 제공 방법에 관한 것이다.

### *배 경 기 술*

[0002] 차세대 염기서열 분석 방법 (Next-generation sequencing, NGS) 은 임상적 연구를 포함하는 다양한 목적을 달성하기 위해 유전체 및 전체 분석에 적용되고 있다. 차세대 염기서열 분석에서, 표적 영역의 리드 뎁스 (read depth) 는 변이 대립 유전자 및 이들의 빈도를 확인하고, 유전자의 발현 수준을 평가하는데 이용될 수 있다. 그러나, 이러한 접근 방법은 라이브러리 준비 단계 또는 염기서열 분석 단계에서 발생하는 증폭 효율의 차이 (amplification bias), 분석 에러에 의해 정밀도 또는 정확도 높은 분석이 어려울 수 있다.

[0003] 한편, 이러한 문제점을 극복하기 위해, A, T, G 및 C의 염기로 구성되어 일정한 뉴클레오티드 길이를 갖는 분자 바코드 (molecular barcode) 와 같은 폴리뉴클레오티드 바코드 (polynucleotide barcode) 의 이용이 제안되었다. 보다 구체적으로, 폴리뉴클리오티드 바코드를 시료 내의 개별 DNA 또는 RNA에 결합시킴으로써, 중복 판독을 필터링함으로써 거짓 양성 변이들을 제거하고, 증폭 편향성을 감소시키는 것에 기여할 수 있다.

[0004] 이러한, 폴리뉴클레오티드 바코드에 기초한 염기서열 분석 방법은 1 % 이하의 저빈도 돌연변이의 검출, 전사체의 정량적 분석 및 단일 세포의 염기서열분석 등에 적용될 수 있다.

[0005] 발명의 배경이 되는 기술은 본 발명에 대한 이해를 보다 용이하게 하기 위해 작성되었다. 발명의 배경이 되는 기술에 기재된 사항들이 선행기술로 존재한다고 인정하는 것으로 이해되어서는 안 된다.

## *발명의 내용*

### *해결하려는 과제*

[0006] 한편, 본 발명의 발명자들은, 염기서열 분석에 이용되는 종래의 폴리뉴클레오티드 바코드들이 무작위로 생성된 염기서열을 이용하는 것에 주목하였다.

[0007] 특히, 본 발명의 발명자들은, 무작위로 생성된 폴리뉴클레오티드 바코드들을 이용했을 때, 바코드의 정확한 염기서열을 알 수 없어 바코드 내에 일어나는 서열 변이에 대하여 확인하기 어려울 수 있다는 점에 대하여 주목하였다.

[0008] 보다 구체적으로, 본 발명의 발명자들은, 바코드 내에서 일어난 서열 변이가 바코드들의 잘못된 클러스터링 (misclustering) 을 야기할 수 있다는 점에 주목하였다. 본 발명의 발명자들은, 이러한 잘못된 클러스터링이, 결과적으로 정밀도 또는 정확도 낮은 저빈도 돌연변이의 검출 결과와 질환 샘플의 전사체의 정량 분석 결과를 제공하게 되고, 단일 세포의 염기서열 해석에도 활용될 수 있음을 인지할 수 있었다.

[0009] 본 발명의 발명자들은, 염기서열 분석 과정에서 일어나는 상기와 같은 문제점을 해결하기 위해, 폴리뉴클레오티드 바코드의 서열 특징, 나아가 폴리뉴클레오티드 바코드들 사이의 상호 관계에 대하여 주목하였다.

[0010] 특히, 본 발명의 발명자들은, 무작위 폴리뉴클레오티드 바코드에서 관찰되는 폴리뉴클레오티드 바코드의 내부적

특징 요소들과 폴리뉴클레오티드 바코드들 각각의 상호 관계 요소들의 분포에 대하여 주목하였다.

[0011] 그 결과, 본 발명의 발명자들은, 폴리뉴클레오티드 바코드의 내부적 특징 요소들과 폴리뉴클레오티드 바코드들 각각의 상호 관계 요소들에 기초하여 바코드를 결정하여 제공할 수 있는 새로운 폴리뉴클레오티드 바코드 세트 제공 시스템을 개발하기에 이르렀다.

[0012] 보다 구체적으로, 상기 폴리뉴클레오티드 바코드 세트 제공 시스템은, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이 또는 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛과 같은 내부적 특징 요소들의 특징 조건에 만족하는 바코드 세트를 제공하도록 구성되었다.

[0013] 나아가, 상기 폴리뉴클레오티드 바코드 세트 제공 시스템은, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수, 또는 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수와 같은 상호 관계 요소들의 특징 조건에 만족하는 바코드를 제공하도록 구성되었다.

[0014] 결과적으로 본 발명의 발명자들은, 이러한 폴리뉴클레오티드 바코드 세트 제공 시스템을 개발함으로써, 차세대 염기서열 분석에 기초한 다양한 연구에 적용될 수 있는 바코드 세트를 제공할 수 있었다. 나아가, 본 발명의 발명자들은, 이러한 폴리뉴클레오티드 바코드 세트가, 무작위로 생성된 바코드 세트를 사용한 저빈도 돌연변이 검출 및 유전체 발현량 추정에서 빈번하게 발생할 수 있는 문제점들을 해결할 수 있음을 기대할 수 있었다.

[0015] 따라서, 본 발명이 해결하고자 하는 과제는, 복수개의 폴리뉴클레오티드 바코드를 포함하는 폴리뉴클레오티드 바코드 세트로서, 복수개의 폴리뉴클레오티드 바코드 각각이 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 만족하는 폴리뉴클레오티드 바코드 세트를 제공하는 것이다.

[0016] 보다 구체적으로, 본 발명이 해결하고자 하는 과제는, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이, 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수, 및 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 중 적어도 하나의 특징 조건을 만족하는 바코드를 포함하는 폴리뉴클레오티드 바코드 세트를 제공하는 것이다.

[0017] 본 발명이 해결하고자 하는 다른 과제는, 미리 결정된 뉴클레오티드 길이에 따라 무작위로 생성된 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 수신하고, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 미리 결정된 특징 조건을 이용하여 평가하고, 평가 결과를 기초로 추천 폴리뉴클레오티드 바코드 세트를 결정하도록 구성된 폴리뉴클레오티드 바코드 세트의 제공 방법을 제공하는 것이다.

[0018] 본 발명이 해결하고자 하는 다른 과제는, 무작위로 생성된 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 수신하도록 구성된 수신부, 및 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 이용하여 평가하고, 평가 결과를 기초로 추천 폴리뉴클레오티드 바코드 세트를 결정하도록 구성된 프로세서를 포함하는 폴리뉴클레오티드 바코드 세트의 제공용 디바이스를 제공하는 것이다.

[0019] 본 발명의 과제들은 이상에서 언급한 과제들로 제한되지 않으며, 언급되지 않은 또 다른 과제들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

## *과제의 해결 수단*

[0020] 전술한 바와 같은 과제를 해결하기 위해, 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트가 제공된다. 이때, 본 발명의 폴리뉴클레오티드 바코드 세트는, 복수개의 폴리뉴클레오티드 바코드를 포함한다. 한편, 복수개의 폴리뉴클레오티드 바코드 각각은, 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 만족한다. 특징 조건은, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이, 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수, 및 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 중 적어도 하나이다.

[0021] 본 발명의 특징에 따르면, 복수개의 폴리뉴클레오티드 바코드 각각은, 특징 조건을 기초로 결정된 특징 조건 점수를 더 만족할 수 있다. 이때, 특징 조건 점수는, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량 점수, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이 점수, 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛 점수, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수 점수, 및 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 점수 중 적어도 하나일 수 있다.

[0022] 본 발명이 다른 특징에 따르면 특징 조건 점수는, GC 함량 점수, 반복 싱글 뉴클레오티드의 길이 점수, 디뉴클레오티드의 반복 유닛 점수, 뉴클레오티드 쌍 점수, 및 상보적 뉴클레오티드 결합 쌍 점수 중 선택된 두 개 이상의 특징 조건 점수를 합산한 합산 특징 조건 점수를 포함할 수 있다.

[0023] 본 발명의 또 다른 특징에 따르면, 폴리뉴클레오티드 바코드는 2 개 내지 25 개의 뉴클레오티드로 이루어진 올리고뉴클레오티드 서열을 가질 수 있다. 바람직하게, 폴리뉴클레오티드 바코드는 8개 내지 12개의 뉴클레오티드로 이루어진 폴리뉴클레오티드 바코드일 수 있으나, 이에 제한되는 것이 아니다.

[0024] 전술한 바와 같은 과제를 해결하기 위해, 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드의 제공 방법이 제공된다. 본 방법은, 프로세서에 의해 구현되는 복수개의 폴리뉴클레오티드 바코드를 포함하는 폴리뉴클레오티드 바코드 세트의 제공 방법으로서, 미리 결정된 뉴클레오티드 길이에 따라 무작위로 생성된 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 수신하는 단계, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 이용하여 평가하는 단계, 및 평가 결과를 기초로 추천 폴리뉴클레오티드 바코드 세트를 결정하는 단계를 포함한다. 이때, 특징 조건은, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이, 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수, 및 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 중 적어도 하나일 수 있다.

[0025] 본 발명의 특징에 따르면, 평가하는 단계는, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 특징 조건을 기초로 결정된 특징 조건 점수를 산출하는 단계, 및 특징 조건 점수를 기초로 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 평가하는 단계를 포함할 수 있다. 이때, 특징 조건 점수는, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량 점수, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이 점수, 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛 점수, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수 점수, 및 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 점수 중 적어도 하나일 수 있다.

[0026] 본 발명의 다른 특징에 따르면, 평가하는 단계는, 특징 조건 점수 산출하는 단계, 및 복수개의 폴리뉴클레오티드 바코드 각각에 대한 특징 조건 점수가 일정한 수준에 수렴하는지 여부를 평가하는 단계를 더 포함할 수 있다. 이때, 특징 조건 점수가 일정한 수준에 수렴하지 않을 경우, 복수개의 폴리뉴클레오티드 바코드 각각에 대한 특징 점수를 기초로폴리뉴클레오티드 바코드를 필터링하는 단계, 및 추가 폴리뉴클레오티드 바코드를 더 수신하는 단계가 더 수행될 수 있다.

[0027] 본 발명의 또 다른 특징에 따르면, 폴리뉴클레오티드 바코드 세트는, 미리 결정된 개수의 복수개의 폴리뉴클레오티드를 포함할 수 있다. 이때, 필터링하는 단계는, 미리 결정된 필터링률을 기초로 복수개의 폴리뉴클레오티드 바코드 각각에 대한 특징 점수를 기초로 폴리뉴클레오티드 바코드를 필터링하는 단계를 포함할 수 있다. 나아가, 추가 폴리뉴클레오티드 바코드를 더 수신하는 단계는, 필터링된 폴리뉴클레오티드 바코드와 동일한 개수의 추가 폴리뉴클레오티드 바코드를 수신하는 단계를 포함할 수 있다.

[0028] 본 발명의 또 다른 특징에 따르면, 특징 조건 점수가 일정한 수준에 수렴하는 경우, 추천 폴리뉴클레오티드 바코드 세트를 결정하는 단계는, 일정한 수준에 수렴하는 복수개의 폴리뉴클레오티드 바코드를 추천 폴리뉴클레오티드 바코드 세트로 결정하는 단계를 포함할 수 있다.

[0029] 본 발명의 또 다른 특징에 따르면, 특징 조건 점수는, GC 함량 점수, 반복 싱글 뉴클레오티드의 길이 점수, 디뉴클레오티드의 반복 유닛 점수, 뉴클레오티드 쌍 점수, 및 상보적 뉴클레오티드 결합 쌍 점수 중 선택된 두 개 이상의 특징 조건 점수를 합산한 합산 특징 조건 점수를 포함할 수 있다. 이때, 평가하는 단계는, 복수개의 폴

리뉴클레오티드 바코드 각각에 대하여, 특징 조건 점수를 산출하는 단계, 선택된 두 개 이상의 특징 조건 점수를 합산하여 합산 특징 조건 점수를 산출하는 단계, 및 합산 특징 조건 점수를 기초로 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 평가하는 단계를 포함할 수 있다.

[0030] 본 발명의 또 다른 특징에 따르면, 본 발명의 제공 방법은, 추천 폴리뉴클레오티드 바코드 세트 및 추천 폴리뉴클레오티드 바코드 세트에 대한 합산 특징 조건 점수를 제공하는 단계를 더 포함할 수 있다.

[0031] 본 발명의 또 다른 특징에 따르면, 산출하는 단계는, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 특징 조건 점수에 대하여 미리 결정된 가중치를 기초로 특징 조건 점수를 산출하는 단계를 포함할 수 있다.

[0032] 전술한 바와 같은 과제를 해결하기 위해, 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드의 제공용 디바이스가 제공된다. 이때, 본 디바이스는, 프로세서에 의해 구현되는 복수개의 폴리뉴클레오티드 바코드를 포함하는 폴리뉴클레오티드 바코드 세트의 제공용 디바이스로서, 미리 결정된 뉴클레오티드 길이에 따라 무작위로 생성된 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 수신하도록 구성된 수신부, 및 수신부와 통신하도록 연결된 프로세서를 포함한다. 이때, 프로세서는, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 이용하여 평가하고, 평가 결과를 기초로 추천 폴리뉴클레오티드 바코드 세트를 결정하도록 구성된다. 한편, 특징 조건은, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이, 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수, 및 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 중 적어도 하나이다.

[0033] 본 발명의 특징에 따르면, 프로세서는, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 특징 조건을 기초로 결정된 특징 조건 점수를 산출하고, 특징 조건 점수를 기초로 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 평가하도록 더 구성될 수 있다. 이때, 특징 조건 점수는, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량 점수, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이 점수, 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛 점수, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수 점수, 및 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 점수 중 적어도 하나일 수 있다.

[0034] 본 발명의 다른 특징에 따르면, 특징 조건 점수는, GC 함량 점수, 반복 싱글 뉴클레오티드의 길이 점수, 디뉴클레오티드의 반복 유닛 점수, 뉴클레오티드 쌍 점수, 및 상보적 뉴클레오티드 결합 쌍 점수 중 선택된 두 개 이상의 특징 조건 점수를 합산한 합산 특징 조건 점수를 포함할 수 있다. 이때, 프로세서는, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 특징 조건 점수를 산출하고, 선택된 두 개 이상의 특징 조건 점수를 합산하여 합산 특징 조건 점수를 산출하고, 합산 특징 조건 점수를 기초로 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 평가하도록 더 구성될 수 있다.

[0035] 본 발명의 또 다른 특징에 따르면, 프로세서는, 추천 폴리뉴클레오티드 바코드 세트 및 추천 폴리뉴클레오티드 바코드 세트에 대한 합산 특징 조건 점수를 제공하도록 더 구성될 수 있다.

[0036] 본 발명의 또 다른 특징에 따르면, 프로세서는, 복수개의 폴리뉴클레오티드 바코드 각각에 대한 특징 조건 점수가 일정한 수준에 수렴하는지 여부를 평가하도록 구성될 수 있다. 나아가 프로세서는, 특징 조건 점수가 일정한 수준에 수렴하지 않을 경우 복수개의 폴리뉴클레오티드 바코드 각각에 대한 특징 점수를 기초로폴리뉴클레오티드 바코드를 필터링하도록 더 구성될 수 있다. 이때, 수신부는, 프로세서에 의해 특징 조건 점수가 일정한 수준에 수렴하지 않을 경우, 추가 폴리뉴클레오티드 바코드를 더 수신하도록 구성될 수 있다.

[0037] 본 발명의 또 다른 특징에 따르면, 프로세서는, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 특징 조건 점수에 대하여 미리 결정된 가중치를 기초로 특징 조건 점수를 산출하도록 더 구성될 수 있다.

[0038] 이하, 실시예를 통하여 본 발명을 보다 상세히 설명한다. 다만, 이들 실시예는 본 발명을 예시적으로 설명하기 위한 것에 불과하므로 본 발명의 범위가 이들 실시예에 의해 한정되는 것으로 해석되어서는 아니 된다.

## *발명의 효과*

[0039] 특히, 본 발명은, 폴리뉴클레오티드 바코드의 내부적 특징 요소들과 폴리뉴클레오티드 바코드들 각각의 상호 관계 요소들에 기초하여 바코드를 결정하여 제공할 수 있는 새로운 폴리뉴클레오티드 바코드 세트 제공 시스템을

제공할 수 있다.

[0040] 보다 구체적으로, 본 발명은, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이 또는 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛과 같은 내부적 특징 요소들의 특징 조건에 만족하는 바코드 세트를 제공할 수 있다.

[0041] 이에, 본 발명은, 이용된 바코드의 정확한 염기서열을 알 수 없는 점, 바코드 내에 일어나는 서열 변이에 대한 확인이 어려운 점과 같은 종래의 무작위로 생성된 폴리뉴클레오티드 바코드가 갖는 한계를 극복할 수 있는 효과가 있다.

[0042] 특히, 본 발명은, 차세대 염기서열 분석 과정에서 발생한 바코드 서열 내에서의 돌연변이, 삽입, 결실과 같은 변이에 의한 잘못된 클러스터링이 야기될 수 있는 종래의 분자 바코드의 이용에 따른 문제점을 개선할 수 있는 효과가 있다.

[0043] 나아가, 본 발명은, 희귀 대립 유전자 (rare allele) 의 빈도 또는 유전자의 발현량의 잘못된 평가로 인해 정밀도 및 정확도 낮은 분석 결과를 제공할 수 있는 종래의 분자 바코드의 이용에 따른 문제점을 개선할 수 있는 효과가 있다.

[0044] 따라서, 본 발명은, 정밀도 또는 정확도 높은 저빈도 돌연변이의 검출 결과와 질환 샘플의 전사체의 정량 분석 결과를 제공할 수 있고, 단일 세포에 대하여 정확도 높은 염기서열 해석을 제공할 수 있는 효과가 있다.

## *도면의 간단한 설명*

[0045] 도 1a는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공용 디바이스 구성을 예시적으로 도시한 것이다.

도 1b는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공용 디바이스의 입력부를 예시적으로 도시한 것이다.

도 1c는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공용 디바이스의 출력부를 예시적으로 도시한 것이다.

도 2는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공 방법의 절차를 예시적으로 도시한 것이다.

도 3a는 본 발명의 다른 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공 방법의 절차를 예시적으로 도시한 것이다.

도 3b는 본 발명의 다른 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공 방법의 절차 중 수렴하는 단계를 예시적으로 도시한 것이다.

도 4a는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 평가에 이용되는 GC 함량 점수의 설계를 예시적으로 도시한 것이다.

도 4b는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 평가에 이용되는 반복 싱글 뉴클레오티드의 길이 점수의 설계를 예시적으로 도시한 것이다.

도 4c는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 평가에 이용되는 디뉴클레오티드의 반복 유닛 점수의 설계를 예시적으로 도시한 것이다.

도 4d는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 평가에 이용되는 뉴클레오티드 쌍 점수의 설계를 예시적으로 도시한 것이다.

도 4e는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 평가에 이용되는 상보적 뉴클레오티드 결합 쌍 점수의 설계를 예시적으로 도시한 것이다.

도 4f는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 평가에 이용되는 다양한 특징 조건 점수의 가중치에 따른 산출 결과를 도시한 것이다.

## *발명을 실시하기 위한 구체적인 내용*

[0046] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시

예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 것이며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하며, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0047] 본 발명의 실시예를 설명하기 위한 도면에 개시된 형상, 크기, 비율, 각도, 개수 등은 예시적인 것이므로 본 발명이 도시된 사항에 한정되는 것은 아니다. 또한, 본 발명을 설명함에 있어서, 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명은 생략한다. 본 명세서 상에서 언급된 '포함한다', '갖는다', '이루어진다' 등이 사용되는 경우, '~만'이 사용되지 않는 이상 다른 부분이 추가될 수 있다. 구성요소를 단수로 표현한 경우에 특별히 명시적인 기재 사항이 없는 한 복수를 포함하는 경우를 포함한다.

[0048] 구성요소를 해석함에 있어서, 별도의 명시적 기재가 없더라도 오차 범위를 포함하는 것으로 해석한다.

[0049] 본 발명의 여러 실시예들의 각각 특징들이 부분적으로 또는 전체적으로 서로 결합 또는 조합 가능하며, 당업자가 충분히 이해할 수 있듯이 기술적으로 다양한 연동 및 구동이 가능하며, 각 실시예들이 서로에 대하여 독립적으로 실시 가능할 수도 있고 연관 관계로 함께 실시 가능할 수도 있다.

[0050] 본 명세서의 해석의 명확함을 위해, 이하에서는 본 명세서에서 사용되는 용어들을 정의하기로 한다.

[0051] 본 명세서에서 사용되는 용어, "폴리뉴클레오티드 바코드"는 A, T, G 및 C의 염기로 구성된 DNA 서열을 의미할 수 있다. 이때, 본원 명세서 내에 개시된 폴리뉴클레오티드 바코드는, 약 2 개 내지 25 개의 뉴클레오티드 바람직하게는, 8 개 내지 12 개의 뉴클레오티드로 구성된 올리고뉴클레오티드일 수 있지만 이에 제한되는 것은 아니다.

[0052] 한편, 폴리뉴클레오티드 바코드는, 하나의 시료에서 개별 DNA 혹은 RNA 절편을 구별하기 위해 사용되는 분자 바코드일 수 있으나, 이에 제한되는 것이 아니다. 예를 들어, 폴리뉴클레오티드 바코드는 여러 가지 시료에 대하여 단일의 PCR 또는 염기서열 분석을 수행하고자 할 때, 시료의 출처를 표시하기 위해 이용되는 시료 바코드(sample barcode)로 지칭될 수도 있다.

[0053] 본원 명세서에서 개시된 폴리뉴클레오티드 바코드는 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건에 만족하는 바코드일 수 있다. 미리 결정된 특징 조건은 후술한다.

[0054] 본 명세서에서 사용되는 용어, "폴리뉴클레오티드 바코드 세트"는 임의의 뉴클레오티드 길이를 갖는 복수개의 폴리뉴클레오티드 바코드로 구성된 바코드 세트를 의미할 수 있다.

[0055] 한편, 본 명세서에서 사용되는 용어, "특징 조건"은 폴리뉴클레오티드 바코드의 내부적 특징 요소들과 폴리뉴클레오티드 바코드들 각각의 상호 관계 요소들에 기초하여 설정된, 폴리뉴클레오티드 바코드로의 적합성을 평가하기 위한 조건일 수 있다.

[0056] 보다 구체적으로, 특징 조건 중 상기 내부적 특징 요소는, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이 또는 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛일 수 있다. 그러나, 특징 조건은 이에 제한되지 않고, 보다 다양한 바코드의 뉴클레오티드 서열에 대한 내부적 특징 요소를 포함할 수 있다.

[0057] 나아가, 특징 조건 중 상기 상호 관계 요소는, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수, 또는 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수일 수 있다. 그러나, 특징 조건은 이에 제한되지 않고, 바코드들 사이의 상호 관계에 관한 보다 다양한 상호 관계 요소를 포함할 수 있다.

[0058] 한편, 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트는, 상기와 같은 특징 조건에 기초하여 설정된 특징 조건 점수가 일정한 수준에 수렴하는지 평가된 복수개의 폴리뉴클레오티드 바코드로 구성될 수 있다.

[0059] 이때, 본 명세서에서 사용되는 용어, "특징 조건 점수"는 전술한 특징 조건을 기초로 폴리뉴클레오티드 바코드로의 적합성을 평가하기 위한 척도를 의미할 수 있다.

[0060] 상기 특징 조건 점수는, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량 점수 (이하, GC 함량 점수), 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는, 반복 싱글 뉴클레오티드의 길이 점수 (이하, 반복 싱글 뉴클레오티드의 길이 점수), 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛 점수 (이하, 디뉴클레

오티드의 반복 유닛 점수), 바코드 세트 내에서 선택된 두 개의 상이한 바코드에 대하여 존재하는 상이한 뉴클레오티드 쌍 점수 (이하, 뉴클레오티드 쌍 점수), 및 두 개의 상이한 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍 점수 (이하, 상보적 뉴클레오티드 결합 쌍 점수) 중 적어도 하나일 수 있다. 나아가, 상기 특징 조건 점수는, 전술한 특징 조건 점수 중 선택된 두 개 이상의 특징 조건 점수를 합산한, 합산 특징 조건 점수일 수도 있다.

[0061] 한편, 특징 조건 점수는, 전술한 것에 제한되는 것이 아니다.

[0062] 예를 들어, 특징 조건 점수는, 전술한 특징 조건에 따라 폴리뉴클레오티드 바코드로의 적합성이 낮은 바코드에 대하여 벌점이 부과되도록 설계된, 특징 조건 패널티 점수를 더 포함할 수 있다.

[0063] 보다 구체적으로, 특징 조건 패널티 점수는, GC 함량 패널티 점수, 반복 싱글 뉴클레오티드의 길이 패널티 점수, 디뉴클레오티드의 반복 유닛 패널티 점수, 뉴클레오티드 쌍 패널티 점수, 및 상보적 뉴클레오티드 결합 쌍 패널티 점수로 구성될 수 있다.

[0064] 본 명세서에서 사용되는 용어, "GC 함량 점수"는, 무작위 폴리뉴클레오티드 바코드에서 관찰되는 GC 함률 분포에 기초하여, GC 함률을 평가하도록 설계된 점수일 수 있다.

[0065] 예를 들어, GC 함량 점수는 하기 [수학식 1]에 의해 산출될 수 있다.

[0066] [수학식 1]

[0067] $$G = \frac{f(D)}{f(D=40 \text{ or } 60)} \times 10^6$$ ,

[0068] $$f(D) = 22662.54 \times e^{-\frac{(D-50)^2}{2 \times 14.74996^2}}$$

[0069] 여기서, G는 GC 함량 점수이고, D는 상기 폴리뉴클레오티드 바코드의 서열 내의 GC 함률이다. 이때, G의 값이 1000000을 초과하는 모든 경우, 상기 G는 1000000일 수 있다.

[0070] 한편, 상기 GC 함량 점수에 기초하여 GC 함률이 미리 결정된 수준 이상 또는 이하일 경우 그렇지 않은 분자 바코드보다 높은 패널티 점수가 부과되도록 설계된 GC 함량 패널티 점수가 산출될 수 있다.

[0071] 예를 들어, GC 함량 패널티 점수는 하기 [수학식 2]에 의해 산출될 수 있다.

[0072] [수학식 2]

[0073] $$P_{GCC} = \frac{10^6 - G}{10^6 - G_{min}} \times 10^6$$

[0074] 여기서, $P_{GCC}$는 GC 함량 패널티 점수이고, G는 상기 [수학식 1]에 의해 산출된 GC 함량 점수이고, $G_{min}$은 상기 D가 0 또는 100일 경우의 GC 함량 점수이다.

[0075] 이러한 GC 함량 패널티 점수의 산출 방법에 의해, 폴리뉴클레오티드 바코드 내의 GC 함량이 40 % 이하이거나, 60 % 이상인 폴리뉴클레오티드 바코드는 GC 함량이 40 내지 60 %인 바코드보다 패널티 점수가 높을 수 있다.

[0076] 그러나, GC 함량 점수 및 GC 함량 패널티 점수는 이에 제한되지 않고, 폴리뉴클레오티드 바코드의 GC 함량을 평가하는 한 보다 다양한 산출 방법으로 산출될 수 있다.

[0077] 본 명세서에서 사용되는 용어, "반복 싱글 뉴클레오티드의 길이 점수"는, 무작위 폴리뉴클레오티드 바코드에서 관찰되는 호모폴리머 (homopolymer) 의 길이 분포에 기초하여, 반복 싱글 뉴클레오티드의 길이를 평가하도록 설계된 점수일 수 있다.

[0078] 예를 들어, 상기 반복 싱글 뉴클레오티드의 길이 점수는 하기 [수학식 3]에 의해 산출될 수 있다.

[0079] [수학식 3]

[0080] $$H = \frac{f(L)_m}{f(L=2)_m} \times 10^6$$ ,

[0081] $$f(L)_m = 461428 \times e^{-1.31322 \times L}$$

[0082] 여기서, H는 반복 싱글 뉴클레오티드의 길이 점수이고, L은 반복 싱글 뉴클레오티드의 길이이다. 이때, 복수의

염기에 대하여 반복 싱글 뉴클레오티드가 존재할 경우, L은 복수의 염기 중 가장 긴 길이를 갖는 반복 싱글 뉴클레오티드의 길이이다. 한편, $f(L)_m$의 값이 0 미만인 모든 경우, $f(L)_m$은 0일 수 있다.

[0083] 한편, 반복 싱글 뉴클레오티드의 길이 점수에 기초하여, 반복 싱글 뉴클레오티드의 길이가 미리 결정된 수준 이상일 경우 그렇지 않은 분자 바코드보다 높은 패널티 점수가 부과되도록 설계된, 반복 싱글 뉴클레오티드의 길이 패널티 점수가 산출될 수 있다.

[0084] 예를 들어, 반복 싱글 뉴클레오티드의 길이 패널티 점수는 하기 [수학식 4]에 의해 산출될 수 있다.

[0085] [수학식 4]

[0086] $$P_{HP} = 10^6 - H + 2$$

[0087] 여기서, $P_{HP}$는 반복 싱글 뉴클레오티드의 길이 점수이고, H는 상기 [수학식 3]에 의해 산출된 반복 싱글 뉴클레오티드의 길이 점수이다.

[0088] 이러한 반복 싱글 뉴클레오티드의 길이 패널티 점수의 산출 방법에 의해, 폴리뉴클레오티드 바코드 내의 반복 싱글 뉴클레오티드의 길이가 2 초과인 폴리뉴클레오티드 바코드는, 반복 싱글 뉴클레오티드의 길이가 2 이하인 바코드보다 패널티 점수가 높게 산출될 수 있다.

[0089] 이때, 반복 싱글 뉴클레오티드의 길이 점수 및 반복 싱글 뉴클레오티드의 길이 패널티 점수는 이에 제한되지 않고, 폴리뉴클레오티드 바코드 내의 호모폴리머의 수를 평가하는 한 보다 다양한 산출 방법으로 산출될 수 있다.

[0090] 본 명세서에서 사용되는 용어, "디뉴클레오티드의 반복 유닛 점수"는, 무작위 폴리뉴클레오티드 바코드에서 관찰되는, 두 개의 뉴클레오티드로 구성된 디뉴클레오티드의 반복 유닛 (repeat unit) 수의 분포에 기초하여, 상기 반복 유닛의 수를 평가하도록 설계된 점수일 수 있다.

[0091] 예를 들어, 상기 디뉴클레오티드의 반복 유닛 점수는 하기 [수학식 5]에 의해 산출될 수 있다.

[0092] [수학식 5],

[0093] $$S = \frac{f(R)_m}{f(R=2)_m} \times 10^6$$ ,

[0094] $$f(R)_m = 15642980 \times e^{-3.010114 \times R}$$

[0095] 여기서, S는 디뉴클레오티드의 반복 유닛 점수이고, R은 반복 디뉴클레오티드의 반복 유닛의 개수이다. 한편, $f(R)_m$의 값이 0 미만인 모든 경우, $f(R)_m$은 0일 수 있다.

[0096] 한편, 디뉴클레오티드의 반복 유닛 점수에 기초하여, 디뉴클레오티드의 반복 유닛의 개수가 미리 결정된 수준 이상일 경우 그렇지 않은 분자 바코드보다 높은 패널티 점수가 부과되도록 설계된, 디뉴클레오티드의 반복 유닛 패널티 점수가 산출될 수 있다.

[0097] 예를 들어, 디뉴클레오티드의 반복 유닛 패널티 점수는 하기 [수학식 6]에 의해 산출될 수 있다.

[0098] [수학식 6]

[0099] $$P_{SR} = 10^6 - S + 6$$

[0100] 여기서 $P_{SR}$은 디뉴클레오티드의 반복 유닛 패널티 점수이고, S는 상기 [수학식 5]에 의해 산출된 디뉴클레오티드의 반복 유닛 점수이다.

[0101] 이러한 반복 디뉴클레오티드의 반복 유닛 패널티 점수의 산출 방법에 의해, 폴리뉴클레오티드 바코드 내의 디뉴클레오티드의 반복 유닛이 2 개 이상인 폴리뉴클레오티드 바코드는, 2 개 미만인 바코드에 비하여 높은 패널티 점수가 부과될 수 있다.

[0102] 한편, 디뉴클레오티드의 반복 유닛 점수 및 디뉴클레오티드의 반복 유닛 패널티 점수는 이에 제한되지 않고, 폴리뉴클레오티드 바코드 내의 반복되는 디뉴클레오티드의 유닛의 개수를 평가하는 한 보다 다양한 산출 방법으로 산출될 수 있다.

[0103] 본 명세서에서 사용되는 용어, "뉴클레오티드 쌍 점수"는, 동일한 뉴클레오티드 길이를 갖는 무작위 폴리뉴클레오티드 바코드들에 대하여, 동일한 위치에서 상이한 염기를 갖는 뉴클레오티드 쌍의 개수, 즉 해밍 거리(Hamming distance) 에 대한 평가 점수를 의미할 수 있다.

[0104] 한편, 뉴클레오티드 쌍의 개수가 미리 결정된 수준 이하일 경우, 그렇지 않은 바코드보다 패널티 점수가 높도록 설계된 뉴클레오티드 쌍 패널티 점수에 의해, 폴리뉴클레오티드 바코드들에 대한 평가가 수행될 수 있다.

[0105] 예를 들어, 상기 뉴클레오티드 쌍 패널티 점수는 하기 [수학식 7]에 의해 산출될 수 있다.

[0106] [수학식 7]

[0107] $$P_{HD}\begin{cases}10^{9}\times e^{-6.907755\times HD}-7.524427\times 10^{-16}, & \text{if } HD=1,2,\text{or } 3\\ 0, & \text{if } HD\geq 4\end{cases}$$

[0108] 여기서, $P_{HD}$는 뉴클레오티드 쌍 패널티 점수이고, HD는 뉴클레오티드 쌍의 개수이다. 이때, 뉴클레오티드 쌍 패널티 점수는 뉴클레오티드 쌍의 개수에 따라 상이한 방법으로 산출될 수 있다. 예를 들어, 뉴클레오티드 쌍의 개수가 1, 2 또는 3인 경우, 뉴클레오티드 쌍 패널티 점수는, 식 $P_{HD}=10^{9}\times e^{-6.907755\times HD}-7.524427\times 10^{-16}$에 의해 산출될 수 있다. 나아가, 뉴클레오티드 쌍의 개수가 4이상일 경우, 뉴클레오티드 쌍 패널티 점수는 0일 수 있다.

[0109] 이러한 뉴클레오티드 쌍 패널티 점수의 산출 방법에 의해, 폴리뉴클레오티드 바코드들 중, 이들 사이의 해밍 거리가 3 이하인 폴리뉴클레오티드 바코드들은, 그렇지 않은 폴리뉴클레오티드 바코드들 보다 높은 패널티 점수가 부과될 수 있다.

[0110] 그러나, 뉴클레오티드 쌍 패널티 점수는 이에 제한되지 않고, 복수의 폴리뉴클레오티드 바코드 사이의 해밍 거리에 대하여 평가하는 한 보다 다양한 산출 방법으로 산출될 수 있다.

[0111] 본 명세서에서 사용되는 용어, "상보적 뉴클레오티드 결합 쌍 점수"는, 동일한 뉴클레오티드 길이를 갖는 무작위 폴리뉴클레오티드 바코드들에 대하여, 상보적 결합을 갖는 뉴클레오티드 쌍의 개수의 평가 점수를 의미할 수 있다.

[0112] 한편, 상보적 결합을 갖는 뉴클레오티드 쌍의 개수가 미리 결정된 수준 이상일 경우 패널티 점수가 높도록 설계된 상보적 뉴클레오티드 결합 쌍 패널티 점수에 의해, 폴리뉴클레오티드 바코드들에 대한 평가가 수행될 수 있다.

[0113] 예를 들어, 상기 상보적 뉴클레오티드 결합 쌍 패널티 점수는 하기 [수학식 8]에 의해 산출될 수 있다.

[0114] [수학식 8]

[0115] $$P_{CP}\begin{cases}9.999999\times 10^{-17}\times e^{-4.60517\times M_{CP}}-0.0009216598, & \text{if } N_{CP}\geq 2/3\times BL\\ 0, & \text{if } N_{CP}<2/3\times BL\end{cases}$$

[0116] $$M_{CP}=\frac{N_{CP}}{BL}\times 12$$

[0117] 여기서, $P_{CP}$는 상보적 뉴클레오티드 결합 쌍 패널티 점수이고, BL은 폴리뉴클레오티드 바코드의 길이이고, $N_{CP}$는 최대 상보적 뉴클레오티드 결합 길이이다. 이때, 상보적 뉴클레오티드 결합 쌍 패널티 점수는 최대 상보적 뉴클레오티드 결합 길이와 폴리뉴클레오티드의 길이에 따라 상이한 방법으로 산출될 수 있다. 예를 들어, 최대 상보적 뉴클레오티드 결합 길이가, 폴리뉴클레오티드의 길이에 대하여 2/3을 곱한 값 이상일 경우, 상보적 뉴클레오티드 결합 쌍 패널티 점수는, 식 $P_{CP}=9.999999\times 10^{-17}\times e^{-4.60517\times M_{CP}}-0.0009216598$에 의해 산출될 수 있다. 이때, $M_{cp}$는 BL이 12nt가 아닌 바코드에 대하여 산출된 최대 상보적 뉴클레오티드 결합 길이 ($N_{CP}$) 를 이용하여 상보적 뉴클리오티드 결합 쌍 패널티 점수 ($P_{CP}$) 를 얻기위해 이용되는 변환 값을 의미할 수 있다.

[0118] 나아가, 최대 상보적 뉴클레오티드 결합 길이가, 폴리뉴클레오티드의 길이에 대하여 2/3을 곱한 값 미만일 경우, 상보적 뉴클레오티드 결합 쌍 패널티 점수는 0일 수 있다.

[0119] 이러한 상보적 뉴클레오티드 결합 쌍 패널티 점수의 산출 방법에 의해, 예를 들어, 12 개의 뉴클레오티드로 구성된 폴리뉴클레오티드 바코드들 중, 8 쌍 이상의 상보적 결합쌍을 갖는 폴리뉴클레오티드 바코드들은 그렇지

않은 폴리뉴클레오티드 바코드들보다 높은 패널티 점수가 부과될 수 있다.

[0120] 그러나, 상보적 뉴클레오티드 결합 쌍 점수는 이에 제한되지 않고, 복수의 폴리뉴클레오티드 바코드 사이의 상보적 결합에 대하여 평가하는 한 보다 다양한 산출 방법으로 산출될 수 있다.

[0121] 한편, 본 명세서에서 사용되는 용어, "합산 특징 조건 점수"는 전술한 특징 조건 점수 중 선택된 한 개의 특징 조건 점수, 또는 두 개 이상의 특징 조건 점수를 합산한 점수를 의미할 수 있다.

[0122] 이때, 합산 특징 조건 점수는, 미리 결정된 가중치가 반영되어 산출될 수도 있다.

[0123] 보다 구체적으로, 합산 특징 조건 점수는, 특징 조건 점수 각각에 대하여 미리 결정된 가중치를 곱하고, 가중치가 반영된 특징 조건 점수를 합산하여 산출된 점수일 수 있다.

[0124] 이에, "선택된 한 개의 특징 조건 점수"는, 복수의 특징 조건 점수 중 한 개의 특징 조건에 대한 가중치가 0보다 높게 설정되고, 나머지 특징 조건 점수 각각에 대한 가중치가 0으로 설정되어, 최종적으로 합산된 합산 특징 점수가, 한 개의 특징 조건 점수와 동일하게 산출된 점수를 의미할 수 있다.

[0125] 바람직하게, 합산 특징 조건 점수는, GC 함량 패널티 점수, 반복 싱글 뉴클레오티드의 길이 패널티 점수, 디뉴클레오티드의 반복 유닛 패널티 점수, 뉴클레오티드 쌍 패널티 점수 및 상보적 뉴클레오티드 결합 쌍 패널티 점수를 모두 합산한, 합산 특징 조건 패널티 점수일 수 있다.

[0126] 예를 들어, 합산 특징 조건 패널티 점수는 하기 [수학식 9]에 의해 산출될 수 있다.

[0127] [수학식 9]

[0128] $$P_{WT} = w_1 \times P_{GCCt} + w_2 \times P_{HPt} + w_3 \times P_{SRt} + w_4 \times P_{HDt} + w_5 \times P_{CPt}$$

[0129] 여기서, $P_{WP}$는 합산 특징 조건 패널티 점수이고, $W_1$은 GC 함량 점수에 대하여 미리 결정된 가중치이고, $P_{GCCt}$는 폴리뉴클레오티드 바코드 세트 내의 복수개의 바코드에 대한 GC 함량 패널티 점수의 총 합이다. 나아가, $W_2$는 반복 싱글 뉴클레오티드의 길이 점수에 대하여 미리 결정된 가중치이고, $P_{HPt}$는 폴리뉴클레오티드 바코드 세트 내의 복수개의 바코드에 대한 반복 싱글 뉴클레오티드의 길이 패널티 점수의 총 합이다. 나아가, $W_3$은 디뉴클레오티드의 반복 유닛 점수에 대하여 미리 결정된 가중치이고, $P_{SRt}$는 폴리뉴클레오티드 바코드 세트 내의 복수개의 바코드에 대한 디뉴클레오티드의 반복 유닛 패널티 점수의 총 합이다. 또한, $W_4$는 뉴클레오티드 쌍 점수에 대하여 미리 결정된 가중치이고, $P_{HDt}$는 폴리뉴클레오티드 바코드 세트 내의 복수의 폴리뉴클레오티드 바코드 쌍에 대한 뉴클레오티드 쌍 패널티 점수의 총 합이다. 나아가, $W_5$는 상보적 뉴클레오티드 결합 쌍 점수에 대하여 미리 결정된 가중치이고, $P_{CPt}$는 폴리뉴클레오티드 바코드 세트 내의 복수의 폴리뉴클레오티드 바코드 쌍에 대한 상보적 뉴클레오티드 결합 쌍 패널티 점수의 총 합이다.

[0130] 그러나, 합산 특징 조건 점수, 나아가 합산 특징 조건 패널티 점수는 상술한 것에 제한되지 않고, 특징 조건 점수의 보다 다양한 조합에 기초하여 산출될 수 있다.

[0131] 한편, 본 발명의 특징에 따르면, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 산출된 특징 조건 점수, 특징 조건 패널티 점수, 합산 특징 조건 점수, 또는 합산 특징 조건 패널티 점수에 기초하여, 폴리뉴클레오티드 바코드가 필터링될 수 있다.

[0132] 나아가, 필터링된 개수만큼의 추가 폴리뉴클레오티드 바코드가 수신될 수 있다.

[0133] 이때, 필터링 & 추가 폴리뉴클레오티드 바코드 수신의 사이클은, 상기 점수들이 일정한 수준에 수렴할 때까지 반복될 수 있다.

[0134] 이때, "일정한 수준"은, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 산출된 특징 조건 점수, 특징 조건 패널티 점수, 합산 특징 조건 점수, 또는 합산 특징 조건 패널티 점수가, 필터링 & 추가 폴리뉴클레오티드 바코드 수신의 사이클의 반복 진행에 따라 실질적으로 더 이상 떨어지지 않는 지점을 의미할 수 있다.

[0135] 다시 말해, "일정한 수준"은 일정 사이클을 반복하여도 더 이상 패널티 점수의 합이 떨어지지 않는 지점 (혹은 사이클) 으로 최소의 패널티 점수를 갖는 지점 (혹은 사이클) 을 의미할 수 있다.

[0136] 이때, 본원 명세서 내에 개시된 일정한 수준은, 고정된 값이 아니라 분자 바코드 세트 내의 폴리뉴클레오티드 바코드의 구성, 선택된 특징 조건 등에 따라 용이하게 변동될 수 있다.

[0137] 한편, 본 발명의 바코드 세트는, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 산출된 특징 조건 점수 또는, 이의 총 합이 일정한 수준에 수렴하는 것으로 평가된 폴리 뉴클레오티드 바코드들로 구성될 수 있다.

[0138] 이하에서는 도 1a 내지 도 1c을 참조하여, 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공용 디바이스에 관하여 구체적으로 설명한다.

[0139] 도 1a는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공용 디바이스 구성을 예시적으로 도시한 것이다. 도 1b는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공용 디바이스의 입력부를 예시적으로 도시한 것이다. 도 1c는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공용 디바이스의 출력부를 예시적으로 도시한 것이다.

[0140] 도 1a을 참조하면, 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공용 디바이스 (100) 는 수신부 (110), 입력부 (120), 출력부 (130), 저장부 (140) 및 프로세서 (150) 를 포함한다.

[0141] 구체적으로 수신부 (110) 는 미리 결정된 뉴클레오티드 길이에 따라 무작위로 생성된 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 수신하도록 구성될 수 있다. 나아가, 수신부 (110) 는 후술할 프로세서 (150) 에 의해 폴리뉴클레오티드 바코드가 필터링될 경우, 추가 폴리뉴클레오티드 바코드를 더 수신하도록 구성될 수 있다.

[0142] 입력부 (120) 는 키보드, 마우스, 터치 스크린 패널 등 제한되지 않는다. 입력부 (120) 는 폴리뉴클레오티드 바코드 세트의 제공용 디바이스 (100) 를 설정하고, 폴리뉴클레오티드 바코드 세트의 제공용 디바이스 (100) 의 동작을 지시할 수 있다.

[0143] 한편, 입력부 (120) 는, 폴리뉴클레오티드 서열 특징과 관련하여 수신부 (110) 로부터 수신하지 않은 추가 정보들을 직접 입력 받도록 구성될 수 있다.

[0144] 예를 들어, 도 1b를 참조하면 입력부 (120) 는, 뉴클레오티드 길이 및 바코드 개수 입력부 (121) 를 포함할 수 있다. 이때, 뉴클레오티드 길이 및 바코드 개수 입력부 (121) 는, 사용자로부터 목표하는 바코드의 뉴클레오티드의 길이, 바코드 세트를 이루는 폴리뉴클레오티드 바코드의 개수를 입력 받도록 구성될 수 있다. 나아가, 입력부 (120) 는, 후술할 프로세서 (150) 에 의해 필터링되는 폴리뉴클레오티드의 비율 즉 필터링률을 입력 받도록 구성된 필터링률 입력부 (122) 를 더 포함할 수 있다. 또한, 입력부 (120) 는, 프로세서 (150) 에 의해 산출될 수 있는 특징 조건 점수 각각을 외부로부터 입력 받도록 구성될 수 있는, GC 함량 점수 입력부 (123), 반복 싱글 뉴클레오티드의 길이 점수 입력부 (124), 디뉴클레오티드의 반복 유닛 점수 입력부 (125), 뉴클레오티드 쌍 점수 입력부 (126) 및 상보적 뉴클레오티드 결합 쌍 점수 입력부 (127) 를 더 포함할 수 있다. 또한, 입력부 (120) 는 가중치 입력부 (128) 를 포함할 수 있는데, 이때, 가중치 입력부 (128) 는 특징 조건 점수에 대하여 미리 결정된 가중치를 입력 받도록 구성될 수 있다.

[0145] 그러나, 입력부 (120) 의 구성은 이에 제한되는 것이 아니다.

[0146] 다음으로, 출력부 (130) 는 입력부 (120) 에 입력된 추가 정보들을 표시하도록 구성될 수 있다. 나아가, 출력부 (130) 는 프로세서 (150) 에 의해 결정된 추천 폴리뉴클레오티드 바코드 세트 즉, 복수개의 추천 폴리뉴클레오티드 바코드에 대한 서열을 표시하도록 구성될 수 있다.

[0147] 예를 들어, 도 1c를 참조하면, 출력부 (130) 는 복수의 바코드 세트에 대한 정보를 표시하도록 구성된 바코드 세트 정보 출력부 (131) 를 포함할 수 있다. 한편, 출력부 (130) 는, 바코드 세트 정보 출력부 (131) 에 표시된 각각의 바코드 세트에 대한 세부 정보를 표시하도록 구성된 바코드 세트 세부 정보 출력부 (133) 를 더 포함할 수 있다. 보다 구체적으로, 바코드 세트 세부 정보 출력부 (133) 는, 뉴클레오티드 길이 및 바코드 개수 입력부 (121) 로부터 입력된 뉴클레오티드 길이 및 바코드 개수를 표시하도록 구성될 수 있다. 나아가, 바코드 세트 세부 정보 출력부 (133) 는, GC 함량 점수 입력부 (123), 반복 싱글 뉴클레오티드의 길이 점수 입력부 (124), 디뉴클레오티드의 반복 유닛 점수 입력부 (125), 뉴클레오티드 쌍 점수 입력부 (126) 및 상보적 뉴클레오티드 결합 쌍 점수 입력부 (127) 를 통해 입력되거나, 후술할 프로세서 (150) 에 의해 산출된 특징 조건 점수를 표시하도록 더 구성될 수 있다. 또한, 바코드 세트 세부 정보 출력부 (133) 는, 가중치 입력부 (128) 를 통해 입력된 가중치를 표시하도록 더 구성될 수 있다. 나아가, 바코드 세트 세부 정보 출력부 (133) 는, 프로세서 (150) 에 의해 결정된 추천 폴리뉴클레오티드 바코드 세트의 복수개의 바코드들을 나열하여 표시하도록 더

구성될 수 있다.

[0148] 한편, 출력부 (130) 는 이에 제한되지 않고, 입력부 (120) 에 입력된 다양한 추가 정보들, 프로세서 (150) 에 의해 산출되거나 결정된 다양한 정보들을 디스플레이 적으로 표시하도록 구성될 수 있다.

[0149] 저장부 (140) 는 수신부 (110) 를 통해 수신한 무작위로 생성된 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 저장하고, 입력부 (120) 를 통해 설정된 폴리뉴클레오티드 바코드 세트의 제공용 디바이스 (100) 의 지시를 저장하도록 구성될 수 있다. 나아가, 저장부 (140) 는 입력부 (120) 에 입력된, 다양한 추가 정보들을 저장하도록 더 구성될 수 있다. 또한, 저장부 (140) 는 후술될 프로세서 (150) 에 의해 산출된 특징 조건 점수, 프로세서 (150) 에 의해 결정된 추천 폴리뉴클레오티드 바코드 세트를 저장하도록 구성될 수 있다.

[0150] 그러나, 전술한 것에 제한되지 않고 저장부 (140) 는, 프로세서 (150) 에 의해 결정된 다양한 정보들을 저장할 수 있다.

[0151] 프로세서 (150) 는 폴리뉴클레오티드 바코드 세트의 제공용 디바이스 (100) 의 차세대 염기서열 분석에 적합도가 높은 추천 폴리뉴클레오티드 바코드를 제공하기 위한 구성 요소일 수 있다.

[0152] 이때, 프로세서 (150) 는, 무작위로 생성된 폴리뉴클레오티드 바코드 세트에 대하여 특징 조건 점수를 기초로 반복적으로 바코드를 교체하도록 구성된 알고리즘에 기초할 수도 있다. 그러나, 이에 제한되는 것은 아니다.

[0153] 본 발명의 프로세서 (150) 는, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 이용하여 평가하고, 평가 결과를 기초로 추천 폴리뉴클레오티드 바코드 세트를 결정하도록 구성될 수 있다.

[0154] 본 발명의 일 실시예에 따르면, 프로세서 (150) 는, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 특징 조건을 기초로 결정된 특징 조건 점수를 산출하고, 특징 조건 점수를 기초로 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 평가하도록 더 구성될 수 있다.

[0155] 본 발명의 다른 실시예에 따르면 프로세서 (150) 는, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 산출된 특징 조건 점수가 일정한 수준에 수렴하는지 여부를 평가하고, 특징 조건 점수가 일정한 수준에 수렴하지 않을 경우 복수개의 폴리뉴클레오티드 바코드 각각에 대한 특징 점수를 기초로폴리뉴클레오티드 바코드를 필터링하도록 더 구성될 수 있다.

[0156] 본 발명의 또 다른 실시예에 따르면 프로세서 (150) 는, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 특징 조건 점수에 대하여 미리 결정된 가중치를 기초로 가충치 적용된 합산 특징 조건 점수를 산출하도록 더 구성될 수 있다.

[0157] 본 발명의 또 다른 실시예에 따르면 프로세서 (150) 는, 추천 폴리뉴클레오티드 바코드 세트 및 추천 폴리뉴클레오티드 바코드 세트에 대한 합산 특징 조건 점수를 제공하도록 더 구성될 수 있다.

[0158] 이에, 프로세서 (150) 는, 무작위로 생성된 폴리뉴클레오티드 바코드보다 차세대 염기서열 분석 방법에 기초한 다양한 분석 예를 들어, 1 % 이하의 저빈도 돌연변이의 검출, 전사체의 정량적 분석 및 단일 세포의 염기서열분석 등에 더욱 적합한 바코드 세트를 제공할 수 있다.

[0159] 이하에서는 도 2를 참조하여, 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공 방법을 구체적으로 설명한다. 도 2는, 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공 방법의 절차를 예시적으로 도시한 것이다.

[0160] 도 2를 참조하면, 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공 방법은 먼저, 미리 결정된 뉴클레오티드 길이에 따라 무작위로 생성된 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 수신하고 (S210), 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 이용하여 평가하고 (S220), 평가 결과를 기초로 추천 폴리뉴클레오티드 바코드 세트를 결정한다 (S230).

[0161] 보다 구체적으로, 바코드 세트를 수신하는 단계 (S210) 에서, 무작위로 생성된 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트가 수신될 수 있다. 이때, 복수개의 폴리뉴클레오티드 바코드의 수는 사용자에 의해 임의로 결정될 수 있다.

[0162] 한편, 바코드 세트를 수신하는 단계 (S210) 에서 수신된 무작위 폴리뉴클레오티드 바코드 세트는, 바코드 서열

의 내부적 특징, 바코드들 간의 상호 관계에 따라 차세대 염기서열 분석 방법의 이용 적합도가 상이한 바코드들을 포함할 수도 있다.

[0163] 이에, 평가하는 단계 (S220) 에서, 복수개의 폴리뉴클레오티드 바코드 각각은, 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 만족하는지 평가될 수 있다.

[0164] 본 발명의 일 실시예에 따르면, 평가하는 단계 (S220) 에서, 복수개의 폴리뉴클레오티드 바코드 각각은, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는 싱글 뉴클레오티드의 길이, 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수, 및 상기 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 중 적어도 하나의 조건을 만족하는지 여부가 평가될 수 있다.

[0165] 다음으로, 추천 폴리뉴클레오티드 바코드 세트를 결정하는 단계 (S230) 에서, 조건을 만족하는 추천 폴리뉴클레오티드 바코드 세트가 결정될 수 있다.

[0166] 본 발명의 일 실시예에 따르면, 추천 폴리뉴클레오티드 바코드 세트를 결정하는 단계 (S230) 에서, 평가하는 단계 (S220) 의 결과로 폴리뉴클레오티드 바코드의 서열 내의 GC 함량, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는 싱글 뉴클레오티드의 길이, 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍의 개수, 및 상기 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 중 적어도 하나의 조건을 만족하는 것으로 평가된 폴리뉴클레오티드 바코드들은 추천 폴리뉴클레오티드 바코드 세트로 결정될 수 있다.

[0167] 이상의 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트 제공 방법의 결과로, 특징 조건을 만족하는 추천 폴리뉴클레오티드 바코드 세트가 제공될 수 있다. 이때, 추천 폴리뉴클레오티드 바코드 세트는 무작위로 생성된 바코드 세트보다 차세대 염기서열 분석에 적합할 수 있다.

[0168] 이하에서는 도 3a 및 도 3b를 참조하여, 본 발명의 다른 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공 방법에 대하여 설명한다. 도 3a는 본 발명의 다른 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공 방법의 절차를 예시적으로 도시한 것이다. 도 3b는 본 발명의 다른 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공 방법의 절차 중 수렴하는 단계를 예시적으로 도시한 것이다.

[0169] 도 3a를 참조하면, 본 발명의 다른 실시예에 따른 폴리뉴클레오티드 바코드 세트의 제공 방법은 먼저, 미리 결정된 뉴클레오티드 길이에 따라 무작위로 생성된 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트를 수신하고 (S310), 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 폴리뉴클레오티드 바코드의 서열 특징을 기초로 특징 조건에 기초하여 결정된 특징 조건 점수를 산출한다 (S320). 그 다음, 특징 조건 점수가 일정한 수준에 수렴하는지 여부를 평가하고 (S330), 특징 조건 점수가 수렴하지 않을 경우 바코드 세트 내의 복수개의 폴리뉴클레오티드 바코드 각각에 대한 특징 점수를 기초로 폴리뉴클레오티드 바코드를 필터링하고 (S340), 특징 조건 점수가 수렴할 경우 추천 폴리뉴클레오티드 바코드를 결정한다 (S350).

[0170] 보다 구체적으로, 바코드 세트를 수신하는 단계 (S310) 에서, 무작위로 생성된 복수개의 폴리뉴클레오티드 바코드를 포함하는 바코드 세트가 수신될 수 있다. 이때, 복수개의 폴리뉴클레오티드 바코드의 수는 사용자에 의해 임의로 결정될 수 있다.

[0171] 한편, 바코드 세트를 수신하는 단계 (S210) 에서 수신된 무작위 폴리뉴클레오티드 바코드 세트는, 바코드 서열의 내부적 특징, 바코드들 간의 상호 관계에 따라 차세대 염기서열 분석 방법의 이용 적합도가 상이한 바코드들을 포함할 수도 있다.

[0172] 이에, 특징 조건 점수를 산출하는 단계 (S220) 에서, 복수개의 폴리뉴클레오티드 바코드 각각은, 폴리뉴클레오티드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 만족하는지 평가될 수 있다.

[0173] 본 발명의 일 실시예에 따르면, 특징 조건 점수를 산출하는 단계 (S320) 에서, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여 특징 조건을 기초로 결정된 특징 조건 점수가 산출될 수 있다. 이때, 특징 조건 점수는, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량 점수, 폴리뉴클레오티드 바코드의 서열 내의 연속으로 반복되는 싱글 뉴클레오티드의 길이 점수, 폴리뉴클레오티드 바코드의 서열 내의 디뉴클레오티드의 반복 유닛 점수, 복수개의 폴리뉴클레오티드 바코드 중에서 선택된 두 개의 상이한 폴리뉴클레오티드 바코드의 상이한 뉴클레오티드 쌍

의 개수 점수, 및 두 개의 상이한 폴리뉴클레오티드 바코드에 대하여 존재하는 상보적 뉴클레오티드 결합 쌍의 개수 점수 중 적어도 하나일 수 있다.

[0174] 본 발명의 또 다른 실시예에 따르면, 특징 조건 점수를 산출하는 단계 (S320) 에서, 복수개의 폴리뉴클레오티드 바코드 각각에 대한 특징 조건 점수가 산출되고, 이를 기초로 합산 특징 조건 점수가 산출될 수 있다.

[0175] 본 발명의 또 다른 실시예에 따르면, 특징 조건 점수를 산출하는 단계 (S320) 에서, 복수개의 폴리뉴클레오티드 바코드 각각에 대하여, 특징 조건 점수에 대하여 미리 결정된 가중치에 기초하여 특징 조건 점수가 산출될 수 있다.

[0176] 본 발명의 또 다른 실시예에 따르면, 특징 조건 점수를 산출하는 단계 (S320) 에서, GC 함량 점수는 하기 [수학식 1]에 의해 산출될 수 있다.

[0177] [수학식 1]

[0178] $$G = \frac{f(D)}{f(D=40 \text{ or } 60)} \times 10^6$$ ,

[0179] $$f(D) = 22662.54 \times e^{-\frac{(D-50)^2}{2 \times 14.74996^2}}$$

[0180] 여기서, G는 GC 함량 점수이고, D는 상기 폴리뉴클레오티드 바코드의 서열 내의 GC 함률이다. 이때, G의 값이 1000000을 초과하는 모든 경우, 상기 G는 1000000일 수 있다.

[0181] 본 발명의 또 다른 실시예에 따르면, 특징 조건 점수를 산출하는 단계 (S320) 에서, 는 GC 함량 점수에 기초하여, GC 함률이 미리 결정된 수준 이상 또는 이하일 경우 높은 패널티 점수가 부과되도록 설계된 GC 함량 패널티 점수가 [수학식 2]에 의해 산출될 수 있다.

[0182] [수학식 2]

[0183] $$P_{GCC} = \frac{10^6 - G}{10^6 - G_{min}} \times 10^6$$

[0184] 여기서, $P_{GCC}$는 GC 함량 패널티 점수이고, G는 상기 [수학식 1]에 의해 산출된 GC 함량 점수이고, $G_{min}$은 상기 D가 0 또는 100일 경우의 GC 함량 점수이다.

[0185] 이러한 GC 함량 패널티 점수의 산출 방법에 의해, 폴리뉴클레오티드 바코드 내의 GC 함량이 40 % 이하이거나, 60 % 이상인 폴리뉴클레오티드 바코드는, 그렇지 않은 폴리뉴클레오티드 바코드보다 높은 패널티 점수가 부과될 수 있다.

[0186] 본 발명의 또 다른 실시예에 따르면, 특징 조건 점수를 산출하는 단계 (S320) 에서, 반복 싱글 뉴클레오티드의 길이 점수는 하기 [수학식 3]에 의해 산출될 수 있다.

[0187] [수학식 3]

[0188] $$H = \frac{f(L)_m}{f(L=2)_m} \times 10^6$$ ,

[0189] $$f(L)_m = 461428 \times e^{-1.31322 \times L}$$

[0190] 여기서, H는 반복 싱글 뉴클레오티드의 길이 점수이고, L은 반복 싱글 뉴클레오티드의 길이이다. 이때, 복수의 염기에 대하여 반복 싱글 뉴클레오티드가 존재할 경우, L은 복수의 염기 중 가장 긴 길이를 갖는 반복 싱글 뉴클레오티드의 길이이다. 한편, $f(L)_m$의 값이 0 미만인 모든 경우, $f(L)_m$은 0일 수 있다.

[0191] 본 발명의 또 다른 실시예에 따르면, 특징 조건 점수를 산출하는 단계 (S320) 에서, 반복 싱글 뉴클레오티드의 길이가 미리 결정된 수준 이상일 경우 그렇지 않은 바코드보다 높은 패널티 점수가 부과되도록 설계된, 반복 싱글 뉴클레오티드의 길이 패널티 점수가 하기 [수학식 4]에 의해 산출될 수 있다.

[0192] [수학식 4]

[0193] $$P_{HP} = 10^6 - H + 2$$

[0194] 여기서, $P_{HP}$는 반복 싱글 뉴클레오티드의 길이 패널티 점수이고, H는 반복 싱글 뉴클레오티드의 길이 점수이다.

[0195] 이러한 반복 싱글 뉴클레오티드의 길이 패널티 점수의 산출 방법에 의해, 폴리뉴클레오티드 바코드 내의 반복 싱글 뉴클레오티드의 길이가 2 초과인 폴리뉴클레오티드 바코드는, 그렇지 않은 바코드보다 높은 패널티 점수가 부과될 수 있다.

[0196] 본 발명의 또 다른 실시예에 따르면, 특징 조건 점수를 산출하는 단계 (S320) 에서, 디뉴클레오티드의 반복 유닛 점수는 하기 [수학식 5]에 의해 산출될 수 있다.

[0197] [수학식 5],

[0198] $$S = \frac{f(R)_m}{f(R=2)_m} \times 10^6$$ ,

[0199] $$f(R)_m = 15642980 \times e^{-3.010114 \times R}$$

[0200] 여기서, S는 디뉴클레오티드의 반복 유닛 점수이고, R은 반복 디뉴클레오티드의 반복 유닛의 개수이다. 한편, $f(R)_m$의 값이 0 미만인 모든 경우, $f(R)_m$은 0일 수 있다.

[0201] 본 발명의 또 다른 실시예에 따르면, 특징 조건 점수를 산출하는 단계 (S320) 에서, 디뉴클레오티드의 반복 유닛의 개수가 미리 결정된 수준 이상일 경우 그렇지 않은 바코드보다 높은 패널티 점수가 부과되도록 설계된, 디뉴클레오티드의 반복 유닛 패널티 점수가 하기 [수학식 6]에 의해 산출될 수 있다.

[0202] [수학식 6]

[0203] $$P_{SR} = 10^6 - S + 6$$

[0204] 여기서 $P_{SR}$은 디뉴클레오티드의 반복 유닛 패널티 점수이고, S는 디뉴클레오티드의 반복 유닛 점수이다.

[0205] 이러한 반복 디뉴클레오티드의 반복 유닛 패널티 점수의 산출 방법에 의해, 폴리뉴클레오티드 바코드 내의 디뉴클레오티드의 반복 유닛이 2 개 이상인 폴리뉴클레오티드 바코드는, 그렇지 않은 바코드보다 높은 패널티 점수가 부과될 수 있다.

[0206] 본 발명의 또 다른 실시예에 따르면, 특징 조건 점수를 산출하는 단계 (S320) 에서, 즉 해밍 거리가 미리 결정된 수준 이하일 경우 그렇지 않은 바코드보다 높은 패널티 점수가 부과되도록 설계된, 뉴클레오티드 쌍 패널티 점수가 하기 [수학식 7]에 의해 산출될 수 있다.

[0207] [수학식 7]

[0208] $$P_{HD}\begin{cases} 10^9 \times e^{-6.907755 \times HD} - 7.524427 \times 10^{-16}, & \text{if } HD = 1, 2, \text{or } 3 \\ 0, & \text{if } HD \geq 4 \end{cases}$$

[0209] 여기서, $P_{HD}$는 뉴클레오티드 쌍 패널티 점수이고, HD는 뉴클레오티드 쌍의 개수이다. 이때, 뉴클레오티드 쌍 패널티 점수는 뉴클레오티드 쌍의 개수에 따라 상이한 방법으로 산출될 수 있다. 예를 들어, 뉴클레오티드 쌍의 개수가 1, 2 또는 3인 경우, 뉴클레오티드 쌍 패널티 점수는, 식 $P_{HD} = 10^9 \times e^{-6.907755 \times HD} - 7.524427 \times 10^{-16}$에 의해 산출될 수 있다. 나아가, 뉴클레오티드 쌍의 개수가 4이상일 경우, 뉴클레오티드 쌍 패널티 점수는 0일 수 있다.

[0210] 이러한 뉴클레오티드 쌍 패널티 점수의 산출 방법에 의해, 폴리뉴클레오티드 바코드들 중, 이들 사이의 해밍 거리가 3 이하인 폴리뉴클레오티드 바코드들은, 그렇지 않은 바코드들 보다 높은 패널티 점수가 부과될 수 있다.

[0211] 본 발명의 또 다른 실시예에 따르면, 특징 조건 점수를 산출하는 단계 (S320) 에서, 상보적 결합을 갖는 뉴클레오티드 쌍의 개수가 미리 결정된 수준 이상일 경우 그렇지 않은 바코드보다 높은 패널티 점수가 부과되도록 설계된, 상보적 뉴클레오티드 결합 쌍 패널티 점수가 하기 [수학식 8]에 의해 산출될 수 있다.

[0212] [수학식 8]

[0213] $$P_{CP}\begin{cases} 9.999999\times10^{-17}\times e^{-4.60517\times M_{CP}} - 0.0009216598, & \text{if } N_{CP} \geq 2/3\times BL \\ 0, & \text{if } N_{CP} < 2/3\times BL \end{cases}$$

[0214] $$M_{CP} = \frac{N_{CP}}{BL}\times 12$$

[0215] 여기서, $P_{CP}$는 상보적 뉴클레오티드 결합 쌍 패널티 점수이고, BL은 폴리뉴클레오티드 바코드의 길이이고, $N_{CP}$는 최대 상보적 뉴클레오티드 결합 길이이다. 이때, 상보적 뉴클레오티드 결합 쌍 패널티 점수는 최대 상보적 뉴클레오티드 결합 길이와 폴리뉴클레오티드의 길이에 따라 상이한 방법으로 산출될 수 있다. 예를 들어, 최대 상보적 뉴클레오티드 결합 길이가, 폴리뉴클레오티드의 길이에 대하여 2/3을 곱한 값 이상일 경우, 상보적 뉴클레오티드 결합 쌍 패널티 점수는, 식 $P_{CP} = 9.999999\times10^{-17}\times e^{-4.60517\times M_{CP}} - 0.0009216598$에 의해 산출될 수 있다. 이때, $M_{cp}$는 BL이 12nt가 아닌 바코드에 대하여 산출된 최대 상보적 뉴클레오티드 결합 길이 ($N_{CP}$) 를 이용하여 상보적 뉴클리오티드 결합 쌍 패널티 점수 ($P_{CP}$) 를 얻기위해 이용되는 변환 값을 의미할 수 있다.

[0216] 나아가, 최대 상보적 뉴클레오티드 결합 길이가, 폴리뉴클레오티드의 길이에 대하여 2/3을 곱한 값 미만일 경우, 상보적 뉴클레오티드 결합 쌍 패널티 점수는 0일 수 있다.

[0217] 이러한 상보적 뉴클레오티드 결합 쌍 패널티 점수의 산출 방법에 의해, 예를 들어, 12 개의 뉴클레오티드로 구성된 폴리뉴클레오티드 바코드들 중, 8 쌍 이상의 상보적 결합쌍을 갖는 폴리뉴클레오티드 바코드들은, 그렇지 않은 바코드들 보다 높은 패널티 점수가 부과될 수 있다.

[0218] 한편, GC 함량 점수, 반복 싱글 뉴클레오티드의 길이 점수, 디뉴클레오티드의 반복 유닛 점수, 뉴클레오티드 쌍 점수 및 상보적 뉴클레오티드 결합 쌍 점수의 특징 조건 점수의 종류 및 이의 산출 방법은 이에 제한되는 것이 아니다.

[0219] 본 발명의 또 다른 실시예에 따르면, 특징 조건 점수를 산출하는 단계 (S320) 에서, 특징 조건 점수 각각에 대하여 미리 결정된 가중치를 기초로 이들의 패널티 점수를 합산한, 합산 특징 조건 패널티 점수가 하기 [수학식 9]에 의해 산출될 수 있다.

[0220] [수학식 9]

[0221] $$P_{WT} = w_1\times P_{GCCt} + w_2\times P_{HPt} + w_3\times P_{SRt} + w_4\times P_{HDt} + w_5\times P_{CPt}$$

[0222] 여기서, $P_{WP}$는 합산 특징 조건 패널티 점수이고, $W_1$은 GC 함량 점수에 대하여 미리 결정된 가중치이고, $P_{GCCt}$는 폴리뉴클레오티드 바코드 세트 내의 복수개의 바코드에 대한 GC 함량 패널티 점수의 총 합이다. 나아가, $W_2$는 반복 싱글 뉴클레오티드의 길이 점수에 대하여 미리 결정된 가중치이고, $P_{HPt}$는 폴리뉴클레오티드 바코드 세트 내의 복수개의 바코드에 대한 반복 싱글 뉴클레오티드의 길이 패널티 점수의 총 합이다. 나아가, $W_3$은 디뉴클레오티드의 반복 유닛 점수에 대하여 미리 결정된 가중치이고, $P_{SRt}$는 폴리뉴클레오티드 바코드 세트 내의 복수개의 바코드에 대한 디뉴클레오티드의 반복 유닛 패널티 점수의 총 합이다. 또한, $W_4$는 뉴클레오티드 쌍 점수에 대하여 미리 결정된 가중치이고, $P_{HDt}$는 폴리뉴클레오티드 바코드 세트 내의 복수의 폴리뉴클레오티드 바코드 쌍에 대한 뉴클레오티드 쌍 패널티 점수의 총 합이다. 나아가, $W_5$는 상보적 뉴클레오티드 결합 쌍 점수에 대하여 미리 결정된 가중치이고, $P_{CPt}$는 폴리뉴클레오티드 바코드 세트 내의 복수의 폴리뉴클레오티드 바코드 쌍에 대한 상보적 뉴클레오티드 결합 쌍 패널티 점수의 총 합이다.

[0223] 즉, 특징 조건 점수를 산출하는 단계 (S320) 에서, 특징 조건 패널티 점수 및 가중치를 기초로, 합산 특징 조건 패널티 점수가 산출될 수 있다.

[0224] 이때, 합산 특징 조건 패널티 점수는, 하나의 특징 조건에 대한 패널티 점수만을 포함할 수도 있다.

[0225] 예를 들어, 복수의 특징 조건 점수 중 한 개의 특징 조건에 대한 가중치가 0보다 높게 설정되고, 나머지 특징 조건 점수 각각에 대한 가중치가 0으로 설정될 경우, 최종적으로 합산된 합산 특징 패널티 점수는, 한 개의 특

징 조건 점수와 동일한 값으로 산출될 수 있다.

[0226] 다음으로, 수렴하는지 여부를 평가하는 단계 (S330) 에서, 특징 조건 점수를 산출하는 단계 (S320) 의 결과로 산출된 GC 함량 점수, 반복 싱글 뉴클레오티드의 길이 점수, 디뉴클레오티드의 반복 유닛 점수, 뉴클레오티드 쌍 점수 및 상보적 뉴클레오티드 결합 쌍 점수 중 적어도 하나의 특징 조건 점수가 일정한 수준에 수렴하는지 여부가 평가될 수 있다.

[0227] 본 발명의 일 실시예에 따르면, 수렴하는지 여부를 평가하는 단계 (S330) 에서, 특징 조건 점수를 산출하는 단계 (S320) 의 결과로 산출된 GC 함량 점수, 반복 싱글 뉴클레오티드의 길이 점수, 디뉴클레오티드의 반복 유닛 점수, 뉴클레오티드 쌍 점수 및 상보적 뉴클레오티드 결합 쌍 점수 중 선택된 두 개 이상의 특징 조건 점수에 대한 합산 특징 조건 점수가 일정한 수준에 수렴하는지 여부가 평가될 수 있다.

[0228] 본 발명의 다른 실시예에 따르면, 수렴하는지 여부를 평가하는 단계 (S330) 의 결과로 특징 조건 점수 또는 합산 특징 조건 점수가 일정한 수준에 수렴하지 않는 것으로 평가된 경우, 폴리뉴클레오티드 바코드를 필터링하는 단계 (S340) 가 수행될 수 있다.

[0229] 본 발명의 또 다른 실시예에 따르면, 폴리뉴클레오티드 바코드를 필터링하는 단계 (S340) 에서, 미리 결정된 필터링률에 기초하여, 패널티 점수가 높은 바코드가 높은 확률로 필터링될 수 있다. 그 다음, 바코드 세트를 수신하는 단계 (S310) 로 돌아가, 필터링된 폴리뉴클레오티드 바코드와 동일한 개수의 추가 폴리뉴클레오티드 바코드가 수신될 수 있다.

[0230] 이때, 바코드 세트를 수신하는 단계 (S310), 특징 조건 점수를 산출하는 단계 (S320), 수렴하는지 여부를 평가하는 단계 (S330) 및 폴리뉴클레오티드 바코드를 필터링하는 단계 (S340) 는, 복수개의 바코드에 대한 특징 조건 점수, 보다 구체적으로 특정 조건에 대한 패널티 점수가 더 이상 낮아지지 않는 수준에 도달할 때 까지 반복될 수 있다.

[0231] 예를 들어, 도 3b를 참조하면, 수렴하는지 여부를 평가하는 단계 (S330) 에서, 상기 단계들 (S310, S320, S330 및 S340) 의 반복 사이클 수가 증가되어도 합산 특징 조건에 대한 패널티 점수가 더 이상 낮아지지 않는 수준에 도달하는지 여부를 평가할 수 있다.

[0232] 한편, 수렴하는지 여부를 평가하는 단계 (S330) 의 결과로 특징 조건 점수 또는 합산 특징 조건 점수가 수렴하는 것으로 평가된 경우, 추천 폴리뉴클레오티드 바코드를 결정하는 단계 (S350) 가 수행될 수 있다.

[0233] 즉, 추천 폴리뉴클레오티드 바코드를 결정하는 단계 (S350) 에서, 특징 조건 점수 또는 합산 특징 조건 점수에 대한 패널티 점수가 일정한 수준 이하로 떨어지지 않는 수준에 도달했을 때의 복수개의 폴리뉴클레오티드 바코드가, 추천 폴리뉴클레오티드 바코드 세트로 결정될 수 있다.

[0234] **본 발명의 다양한 실시예에 이용되는 특징 조건 점수의 설계**

[0235] 이하에서는 도 4a 내지 4e를 참조하여 본 발명의 다양한 실시예에 이용되는, 특징 조건 점수들의 설계 방법에 대하여 구체적으로 설명한다.

[0236] 도 4a는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 평가에 이용되는 GC 함량 점수의 설계를 예시적으로 도시한 것이다. 도 4b는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 평가에 이용되는 반복 싱글 뉴클레오티드의 길이 점수의 설계를 예시적으로 도시한 것이다. 도 4c는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 평가에 이용되는 디뉴클레오티드의 반복 유닛 점수의 설계를 예시적으로 도시한 것이다. 도 4d는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 평가에 이용되는 뉴클레오티드 쌍 점수의 설계를 예시적으로 도시한 것이다. 도 4e는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 평가에 이용되는 상보적 뉴클레오티드 결합 쌍 점수의 설계를 예시적으로 도시한 것이다.

[0237] 먼저, 도 4a의 (a) 및 하기 [표 1]을 참조하면, 무작위 폴리뉴클레오티드 바코드 세트에서 관찰되는 GC 함률 (% GCC) 분포도 및 분포 값 ($f(D)$) 이 도시된다.

[0238] [표 1]

[0239]

| % GCC | D | $f(D)$ | G | $P_{GCC}$ |
|---|---|---|---|---|
| 0.0% | 0 | 72 | 4,023 | 1,000,000 |
| 8.3% | 8.3 | 417 | 23,133 | 980,813 |
| 16.7% | 16.7 | 1,772 | 98,408 | 905,234 |
| 25.0% | 25 | 5,389 | 299,226 | 703,605 |
| 33.3% | 33.3 | 11,938 | 662,904 | 338,458 |
| 41.7% | 41.7 | 19,344 | 1,000,000 | 0 |
| 50.0% | 50 | 22,663 | 1,000,000 | 0 |
| 58.3% | 58.3 | 19,344 | 1,000,000 | 0 |
| 66.7% | 66.7 | 11,938 | 662,904 | 338,458 |
| 75.0% | 75 | 5,389 | 299,226 | 703,605 |
| 83.3% | 83.3 | 1,772 | 98,408 | 905,234 |
| 91.7% | 91.7 | 417 | 23,133 | 980,813 |
| 100.0% | 100 | 72 | 4,023 | 1,000,000 |

[0240] 보다 구체적으로, 뉴클레오티드의 길이가 12nt인 100000개의 폴리뉴클레오티드 바코드들로 구성된 무작위 폴리뉴클레오티드 바코드 세트에서, 절반 이상의 폴리뉴클레오티드 바코드들이 40 내지 60 %의 GC 함량 (D) 을 갖는 것으로 나타난다. 이에, GC 함량 점수 (G) 는 GC 함량이 40 내지 60 %일 때 상대적으로 높은 점수를 갖도록 설계될 수 있다.

[0241] 이때, 무작위 폴리뉴클레오티드 바코드 세트의 GC 함량 분포는 40 내지 60 %를 기준으로 대칭을 이루고 있음에 따라, 가우시안 (Gaussian) 분포와 핏팅 (fitting) 될 수 있다.

[0242] 이러한 GC 함량 분포에 따라, GC 함량 점수는 하기 [수학식 1]에 의해 산출될 수 있다.

[0243] [수학식 1]

[0244] $$G = \frac{f(D)}{f(D=40 \text{ or } 60)} \times 10^6$$ ,

[0245] $$f(D) = 22662.54 \times e^{-\frac{(D-50)^2}{2 \times 14.74996^2}}$$

[0246] 여기서, G는 GC 함량 점수이고, D는 상기 폴리뉴클레오티드 바코드의 서열 내의 GC 함량이다. 이때, G의 값이 1000000을 초과하는 모든 경우, 상기 G는 1000000일 수 있다.

[0247] 한편, 4a의 (b) 및 [표 1]을 참조하면, 전술한 무작위 폴리뉴클레오티드 바코드 세트의 GC 함량 분포를 기초로, 폴리뉴클레오티드 바코드 내의 GC 함량에 따라 상이한 패널티 점수를 부과하도록 설계된 GC 함량 패널티 점수 ($P_{GCC}$) 의 분포가 도시된다.

[0248] 보다 구체적으로, GC 함량 패널티 점수 ($P_{GCC}$) 는, 폴리뉴클레오티드 바코드 내의 GC 함량이 40 내지 60 % 이하이거나, 40 내지 60 % 이상인 경우 그렇지 않은 분자 바코드보다 높은 패널티 점수가 부과되도록 설계될 수 있다.

[0249] 이때, GC 함량 패널티 점수 ($P_{GCC}$) 는 하기 [수학식 2]에 의해 산출될 수 있다.

[0250] [수학식 2]

[0251] $$P_{GCC} = \frac{10^6 - G}{10^6 - G_{min}} \times 10^6$$

[0252] 여기서, $P_{GCC}$는 GC 함량 패널티 점수이고, G는 상기 [수학식 1]에 의해 산출된 GC 함량 점수이고, $G_{min}$은 상기 D가 0 또는 100일 경우의 GC 함량 점수이다.

[0254] 다음으로, 도 4b의 (a) 및 하기 [표 2]를 참조하면, 무작위 폴리뉴클레오티드 바코드 세트에서 관찰되는, 반복 싱글 뉴클레오티드 길이 (L) 분포도 및 분포 값 ($f(L)$) 이 도시된다.

[0255] [표 2]

| L | $f(L)$ | H | $P_{HP}$ |
|---|---|---|---|
| 1 | 33,378 | 1,000,000 | 2 |
| 2 | 33,378 | 1,000,000 | 2 |
| 3 | 8,977 | 268,953 | 731,049 |
| 4 | 2,414 | 72,336 | 927,666 |
| 5 | 649 | 19,455 | 980,547 |
| 6 | 175 | 5,232 | 994,770 |
| 7 | 47 | 1,407 | 998,595 |
| 8 | 13 | 378 | 999,623 |
| 9 | 3 | 102 | 999,900 |
| 10 | 1 | 27 | 999,975 |
| 11 | 0 | 7 | 999,995 |
| 12 | 0 | 2 | 1,000,000 |

[0256]

[0257] 보다 구체적으로, 무작위 폴리뉴클레오티드 바코드 세트에서, 대부분의 폴리뉴클레오티드 바코드들이 1, 나아가 2의 반복 싱글 뉴클레오티드 길이 (L) 을 갖는 것으로 나타난다. 이에, 반복 싱글 뉴클레오티드의 길이 점수 (H) 는 반복 싱글 뉴클레오티드 길이가 1 및 2일 경우, 상대적으로 높은 점수를 갖도록 설계될 수 있다.

[0258] 이때, 무작위 폴리뉴클레오티드 바코드 세트의 호모폴리머 길이 분포는 기하 급수적으로 감소되는 것으로 나타남에 따라, 지수 분포의 형태를 가질 수 있다.

[0259] 이러한 호모폴리머 길이 분포에 따라, 반복 싱글 뉴클레오티드 길이 점수는 하기 [수학식 3]에 의해 산출될 수 있다.

[0260] [수학식 3]

[0261] $$H = \frac{f(L)_m}{f(L=2)_m} \times 10^6$$ ,

[0262] $$f(L)_m = 461428 \times e^{-1.31322 \times L}$$

[0263] 여기서, H는 반복 싱글 뉴클레오티드의 길이 점수이고, L은 반복 싱글 뉴클레오티드의 길이이다.

[0264] 한편, 4b의 (b) 및 [표 2]를 참조하면, 전술한 무작위 폴리뉴클레오티드 바코드 세트의 반복 싱글 뉴클레오티드 길이 분포를 기초로, 폴리뉴클레오티드 바코드 내의 반복 싱글 뉴클레오티드 길이에 따라 상이한 패널티 점수를 부과하도록 설계된 반복 싱글 뉴클레오티드 길이 패널티 점수 ($P_{HP}$) 의 분포가 도시된다.

[0265] 보다 구체적으로, 반복 싱글 뉴클레오티드 길이 패널티 점수 ($P_{HP}$) 는, 폴리뉴클레오티드 바코드 내의 반복 싱글 뉴클레오티드의 길이가 2 초과인 경우 그렇지 않은 분자 바코드보다 높은 패널티 점수가 부과되도록 설계될 수 있다.

[0266] 이때, 반복 싱글 뉴클레오티드의 길이 패널티 점수 ($P_{HP}$) 는 하기 [수학식 4]에 의해 산출될 수 있다.

[0267] [수학식 4]

[0268] $$P_{HP} = 10^6 - H + 2$$

[0269] 여기서, $P_{HP}$는 반복 싱글 뉴클레오티드의 길이 점수이고, H는 상기 수학식 3에 의해 산출된 반복 싱글 뉴클레오티드의 길이 점수이다.

[0270] 다음으로, 도 4c의 (a) 및 하기 [표 3]을 참조하면, 무작위 폴리뉴클레오티드 바코드 세트에서 관찰되는, 디뉴클레오티드의 반복 유닛의 개수 (R) 분포도 및 분포 값 ($f(R)$) 이 도시된다.

[0271] [표 3]

| R | $f(R)$ | S | $P_{SR}$ |
|---|---|---|---|
| 1 | 37,999 | 1,000,000 | 6 |
| 2 | 37,999 | 1,000,000 | 6 |
| 3 | 1,873 | 49,286 | 950,720 |
| 4 | 92 | 2,429 | 997,577 |
| 5 | 5 | 120 | 999,886 |
| 6 | 0 | 6 | 1,000,000 |

[0272]

[0273] 보다 구체적으로, 무작위 폴리뉴클레오티드 바코드 세트에서, 대부분의 폴리뉴클레오티드 바코드들이 2의 디뉴클레오티드의 반복 유닛의 개수 (R) 를 갖는 것으로 나타난다. 이에, 디뉴클레오티드의 반복 유닛 점수 (S) 는 디뉴클레오티드의 반복 유닛 개수가 1 및 2일 경우, 상대적으로 높은 점수를 갖도록 설계될 수 있다.

[0274] 이때, 무작위 폴리뉴클레오티드 바코드 세트의 디뉴클레오티드의 반복 유닛 개수의 분포는 기하 급수적으로 감소되는 것으로 나타남에 따라, 지수 분포의 형태를 가질 수 있다.

[0275] 이러한 디뉴클레오티드의 반복 유닛의 분포에 따라, 디뉴클레오티드의 반복 유닛 점수는 하기 [수학식 5]에 의해 산출될 수 있다.

[0276] [수학식 5],

[0277] $$S = \frac{f(R)_m}{f(R=2)_m} \times 10^6$$ ,

[0278] $$f(R)_m = 15642980 \times e^{-3.010114 \times R}$$

[0279] 여기서, S는 디뉴클레오티드의 반복 유닛 점수이고, R은 반복 디뉴클레오티드의 반복 유닛의 개수이다. 한편, $f(R)_m$의 값이 0 미만인 모든 경우, $f(R)_m$은 0일 수 있다.

[0280] 한편, 4c의 (b) 및 [표 3]을 참조하면, 전술한 무작위 폴리뉴클레오티드 바코드 세트의 디뉴클레오티드의 반복 유닛의 개수 분포를 기초로, 폴리뉴클레오티드 바코드 내의 디뉴클레오티드의 반복 유닛 개수에 따라 상이한 패널티 점수를 부과하도록 설계된 디뉴클레오티드의 반복 유닛 패널티 점수 ($P_{SR}$) 의 분포가 도시된다.

[0281] 보다 구체적으로, 디뉴클레오티드의 반복 유닛 패널티 점수 ($P_{SR}$) 는, 폴리뉴클레오티드 바코드 내의 디뉴클레오티드의 반복 유닛의 개수가 2 초과인 경우 그렇지 않은 분자 바코드보다 높은 패널티 점수가 부과되도록 설계될 수 있다.

[0282] 이때, 디뉴클레오티드의 반복 유닛 패널티 점수 ($P_{SR}$) 는 하기 [수학식 6]에 의해 산출될 수 있다.

[0283] [수학식 6]

[0284] $$P_{SR} = 10^6 - S + 6$$

[0285] 여기서 $P_{SR}$은 디뉴클레오티드의 반복 유닛 패널티 점수이고, S는 상기 [수학식 5]에 의해 산출된 디뉴클레오티드의 반복 유닛 점수이다.

[0286] 다음으로, 도 4d 및 하기 [표 4]을 참조하면, 동일한 뉴클레오티드 길이를 갖는 무작위 폴리뉴클레오티드 바코드들에 대하여, 동일한 위치에서 상이한 염기를 갖는 뉴클레오티드 쌍의 개수, 즉 해밍 거리가 미리 결정된 수준 이하일 경우 그렇지 않은 분자 바코드보다 높은 패널티 점수가 부과되도록 설계된, 뉴클레오티드 쌍 패널티 점수 ($P_{HD}$) 의 분포가 도시된다.

[0287] [표 4]

[0288]

| HD | $P_{HD}$ according to the barcode length | | |
|---|---|---|---|
| | 12 | 10 | 8 |
| 1 | 1,000,000 | 1,000,000 | 1,000,000 |
| 2 | 1,000 | 1,000 | 1,000 |
| 3 | 1 | 1 | 1 |
| 4 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 |
| 9 | 0 | 0 | |
| 10 | 0 | 0 | |
| 11 | 0 | | |
| 12 | 0 | | |

[0289] 보다 구체적으로, 뉴클레오티드 쌍 패널티 점수 ($P_{HD}$) 는, 폴리뉴클레오티드 바코드들 중, 이들 사이의 해밍 거

리가 3 이하인 경우 패널티가 부과되도록 설계될 수 있다.

[0290] 이때, 뉴클레오티드 쌍 패널티 점수 ($P_{HD}$) 는 하기 [수학식 7]에 의해 산출될 수 있다.

[0291] [수학식 7]

[0292] $$P_{HD}\begin{cases}10^{9}\times e^{-6.907755\times HD}-7.524427\times 10^{-16}, & \text{if } HD=1,2,\text{or } 3\\ 0, & \text{if } HD\geq 4\end{cases}$$

[0293] 여기서, $P_{HD}$는 뉴클레오티드 쌍 패널티 점수이고, HD는 뉴클레오티드 쌍의 개수이다. 이때, 뉴클레오티드 쌍 패널티 점수는 뉴클레오티드 쌍의 개수에 따라 상이한 방법으로 산출될 수 있다.

[0294] 다음으로, 도 4e 및 하기 [표 5]을 참조하면, 동일한 뉴클레오티드 길이를 갖는 무작위 폴리뉴클레오티드 바코드들에 대하여 상보적 결합을 갖는 뉴클레오티드 쌍의 개수가 미리 결정된 수준 이상일 경우 그렇지 않은 분자 바코드보다 높은 패널티 점수가 부과되도록 설계된, 상보적 뉴클레오티드 결합 쌍 패널티 점수 ($P_{CP}$) 의 분포가 도시된다.

[0295] [표 5]

[0296]

| $N_{CP}$ | $P_{CP}$ according to the barcode length | | |
|---|---|---|---|
| | 12 | 10 | 8 |
| 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 |
| 6 | 0 | 0 | 100 |
| 7 | 0 | 6 | 100,000 |
| 8 | 1 | 1,585 | 99,999,767 |
| 9 | 100 | 398,106 | |
| 10 | 10,000 | 99,999,767 | |
| 11 | 999,998 | | |
| 12 | 99,999,767 | | |

[0297] 보다 구체적으로, 상보적 뉴클레오티드 결합 쌍 패널티 점수 ($P_{CP}$) 는, 폴리뉴클레오티드 바코드의 길이에 따른 최대 상보적 뉴클레오티드 결합 길이 ($N_{CP}$) 가 미리 결정된 수준 이상일 경우, 패널티가 부과되도록 설계될 수 있다.

[0298] 이때, 상보적 뉴클레오티드 결합 쌍 패널티 점수 ($P_{CP}$) 는 하기 [수학식 8]에 의해 산출될 수 있다.

[0299] [수학식 8]

[0300] $$P_{CP}\begin{cases}9.999999\times 10^{-17}\times e^{-4.60517\times M_{CP}}-0.0009216598, & \text{if } N_{CP}\geq 2/3\times BL\\ 0, & \text{if } N_{CP}<2/3\times BL\end{cases}$$

[0301] $$M_{CP}=\frac{N_{CP}}{BL}\times 12$$

[0302] 여기서, $P_{CP}$는 상보적 뉴클레오티드 결합 쌍 패널티 점수이고, BL은 폴리뉴클레오티드 바코드의 길이이고, $N_{CP}$는 최대 상보적 뉴클레오티드 결합 길이이다. 나아가, $M_{cp}$는 BL이 12nt가 아닌 바코드에 대하여 산출된 최대 상보적 뉴클레오티드 결합 길이 ($N_{CP}$) 를 이용하여 상보적 뉴클리오티드 결합 쌍 패널티 점수 ($P_{CP}$) 를 얻기위해 이용되는 변환 값이다. 이때, 상보적 뉴클레오티드 결합 쌍 패널티 점수는 최대 상보적 뉴클레오티드 결합 길이와 폴리뉴클레오티드의 길이에 따라 상이한 방법으로 산출될 수 있다.

[0303] 이에, 본 발명의 다양한 실시예에 따른 폴리뉴클레오티드 바코드의 제공 방법은, 전술한 방법들에 의해 무작위로 생성된 폴리뉴클레오티드 바코드에 대하여 산출된 다양한 특징 조건 점수가 일정한 수준에 수렴할 때까지 바코드 필터링 및 바코드 추가를 반복하여, 차세대 염기서열 분석에 적합한 폴리뉴클레오티드 바코드 세트를 제공할 수 있다.

[0304] 즉, 본 발명은, 이용된 바코드의 정확한 염기서열을 알 수 없는 점, 바코드 내에 일어나는 서열 변이에 대한 확

인이 어려운 점과 같은 종래의 무작위로 생성된 폴리뉴클레오티드 바코드가 갖는 한계를 극복할 수 있는 효과가 있다.

[0305] 특히, 본 발명은, 차세대 염기서열 분석 과정에서 발생한 바코드 서열 내에서의 돌연변이, 삽입, 결실과 같은 변이에 의한 잘못된 클러스터링이 야기될 수 있는 종래의 분자 바코드의 이용에 따른 문제점을 개선할 수 있는 효과가 있다.

[0306] 이하에서는, 도 4f를 참조하여, 다양한 특징 조건 점수의 가중치에 따른 산출 결과를 설명한다.

[0307] 도 4f는 본 발명의 일 실시예에 따른 폴리뉴클레오티드 바코드 세트의 평가에 이용되는 다양한 특징 조건 점수의 가중치에 따른 산출 결과를 도시한 것이다. 도 4f의 (a), (b), (c) 및 (d)를 참조하면, 바코드 세트 수신, 특징 조건 점수 산출, 수렴 확인 및 바코드 필터링 단계로 구성된 추천 바코드 세트 제공 사이클의 개시 및 수렴에 따른 특징 조건 점수의 패널티 점수가 도시된다.

[0308] 이때, 도 4f의 (a)를 참조하면, GC 함량 패널티 점수에 대한 가중치가 20.372로 설정되고, 반복 싱글 뉴클레오티드의 길이 패널티 점수에 대한 가중치가 13.653으로 설정되고, 디뉴클레오티드의 반복 유닛 패널티 점수에 대한 가중치가 14.673으로 설정되고, 뉴클레오티드 쌍 패널티 점수에 대한 가중치가 18.34로 설정되고, 상보적 뉴클레오티드 결합 쌍 패널티 점수에 대한 가중치가 10.904로 설정된다.

[0309] 이러한 가중치 설정에 따라, 특징 조건 점수가 수렴하는 수렴 사이클에서, GC 함량 패널티 점수는 62.7 %, 반복 싱글 뉴클레오티드의 길이 패널티 점수는 60.7 %, 디뉴클레오티드의 반복 유닛 패널티 점수는 73.5 %, 뉴클레오티드 쌍 패널티 점수는 64.7 %, 상보적 뉴클레오티드 결합 쌍 패널티 점수는 65.6 % 감소된 것으로 나타난다.

[0310] 도 4f의 (b)를 참조하면, GC 함량 패널티 점수에 대한 가중치가 20.372로 설정되고, 뉴클레오티드 쌍 패널티 점수에 대한 가중치가 18.34로 설정된다.

[0311] 이러한 가중치 설정에 따라, 특징 조건 점수가 수렴하는 수렴 사이클에서, 가중치가 설정된 GC 함량 패널티 점수는 93.2 %, 뉴클레오티드 쌍 패널티 점수는 86.0 % 감소된 것으로 나타난다. 이때, 반복 싱글 뉴클레오티드의 길이 패널티 점수 및 디뉴클레오티드의 반복 유닛 패널티 점수 또한, 개시 사이클에 비하여 소폭 감소한 것으로 나타난다. 그러나, 상보적 뉴클레오티드 결합 쌍 패널티 점수는 43.3 % 증가된 것으로 나타난다.

[0312] 즉, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량 및 바코드간의 해밍 거리는, 폴리뉴클레오티드 바코드의 호모폴리머의 길이, 디뉴클레오티드의 반복 유닛의 개수와 양의 상관관계가 있을 수 있다. 나아가, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량 및/또는 바코드간의 해밍 거리는, 바코드간 상보적 결합을 갖는 뉴클레오티드 쌍의 개수와 음의 상관관계에 있을 수 있다.

[0313] 도 4f의 (c)를 참조하면, GC 함량 패널티 점수에 대한 가중치가 20.372로 설정된다.

[0314] 이러한 가중치 설정에 따라, 특징 조건 점수가 수렴하는 수렴 사이클에서, 가중치가 설정된 GC 함량 패널티 점수는 100 % 감소된 것으로 나타난다. 이때, 반복 싱글 뉴클레오티드의 길이 패널티 점수 및 디뉴클레오티드의 반복 유닛 패널티 점수 또한, 개시 사이클에 비하여 소폭 감소한 것으로 나타난다. 그러나, 뉴클레오티드 쌍 패널티 점수는 34.4 % 증가되고, 상보적 뉴클레오티드 결합 쌍 패널티 점수는 60.5 % 증가된 것으로 나타난다.

[0315] 즉, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량은, 폴리뉴클레오티드 바코드의 호모폴리머의 길이, 디뉴클레오티드의 반복 유닛의 개수와 양의 상관관계가 있을 수 있다. 나아가, 폴리뉴클레오티드 바코드의 서열 내의 GC 함량은, 두 개의 폴리뉴클레오티드 바코드간의 해밍 거리, 바코드간 상보적 결합을 갖는 뉴클레오티드 쌍의 개수와 음의 상관관계에 있을 수 있다.

[0316] 도 4f의 (d)를 참조하면, 뉴클레오티드 쌍 패널티 점수에 대한 가중치가 18.314로 설정된다.

[0317] 이러한 가중치 설정에 따라, 특징 조건 점수가 수렴하는 수렴 사이클에서, 가중치가 설정된 뉴클레오티드 쌍 패널티 점수가 89.8 % 감소된 것으로 나타난다.

[0318] 이때, 나머지 특징 조건 점수에 대한 패널티 점수는 유의한 변화가 없는 것으로 나타난다. 즉, 두 개의 폴리뉴클레오티드 바코드간의 해밍 거리는, 폴리뉴클레오티드 바코드 내의 GC 함량, 폴리뉴클레오티드 바코드의 호모폴리머의 길이, 디뉴클레오티드의 반복 유닛의 개수 및 바코드간 상보적 결합을 갖는 뉴클레오티드 쌍의 개수와 상관관계에 있지 않을 수 있다.

[0319] 본 발명의 여러 실시예들의 각각 특징들이 부분적으로 또는 전체적으로 서로 결합 또는 조합 가능하며, 당업자가 충분히 이해할 수 있듯이 기술적으로 다양한 연동 및 구동이 가능하며, 각 실시예들이 서로에 대하여 독립적으로 실 시 가능할 수도 있고 연관 관계로 함께 실시 가능할 수도 있다.

[0320] 이상 첨부된 도면을 참조하여 본 발명의 실시예들을 더욱 상세하게 설명하였으나, 본 발명은 반드시 이러한 실시예로 국한되는 것은 아니고, 본 발명 의 기술사상을 벗어나지 않는 범위 내에서 다양하게 변형 실시될 수 있다. 따라서, 본 발명에 개시된 실시예들은 본 발명의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 발명의 기술 사상의 범 위가 한정되는 것은 아니다. 그러므로, 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 본 발명의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

## 부호의 설명

[0321] 100: 폴리뉴클레오티드 바코드 세트 제공용 디바이스

110: 수신부

120: 입력부

121: 뉴클레오티드 길이 및 바코드 개수 입력부

122: 필터링률 입력부

123: GC 함량 점수 입력부

124: 반복 싱글 뉴클레오티드의 길이 점수 입력부

125: 디뉴클레오티드의 반복 유닛 점수 입력부

126: 뉴클레오티드 쌍 점수 입력부

127: 상보적 뉴클레오티드 결합 쌍 점수 입력부

128: 가중치 입력부

130: 표시부

131: 바코드 세트 정보 출력부

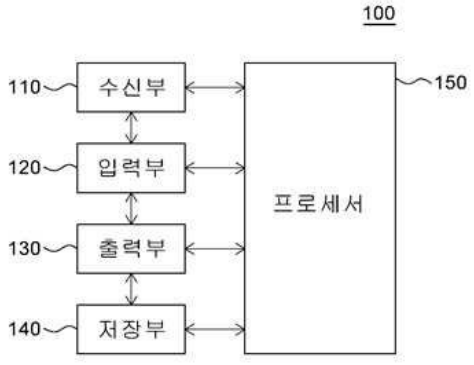
133: 바코드 세트 세부 정보 출력부

140: 저장부

150: 프로세서

## 도면

### 도면1a



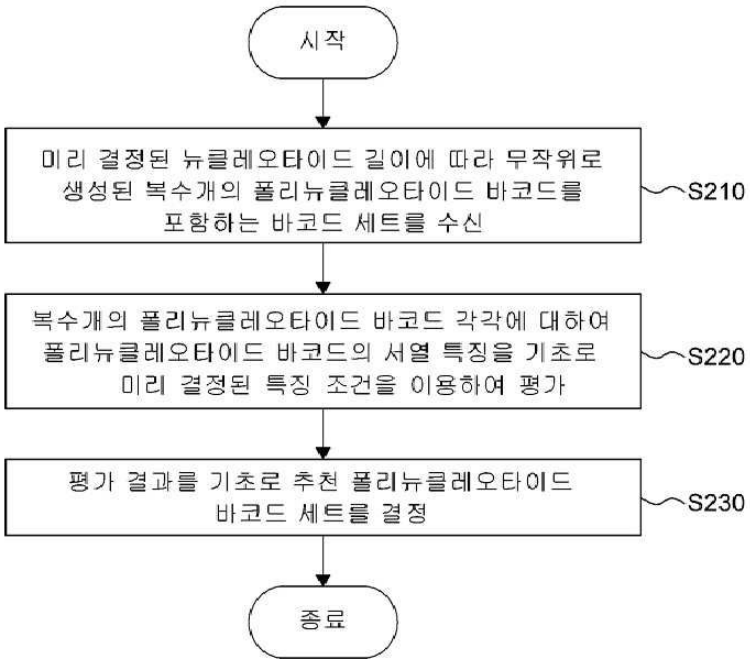
도면1b


121
122
123
124
125
126
127
128
==> param_general <==
//## General parameters
12 #--> Barcode length (BL)
100000 #--> The number of barcodes to be generated (N)
==> param_alpha <==
0.2 #--> The fraction of barcodes to be excluded in each cycle (alpha)
==> param_penalty_GCC <==
//## Penalty scores for GC contents
//## Default: barcodes with length of 12 nucleotides
//## Values from 0% to 100% must be provided in order from left to right.
//## And the values must be separated by whitespace or tab.
1000000 982116 900613 703605 353744 0 0 0 353744 703605 900613 982116 1000000
==> param_penalty_HP <==
//## Penalty scores for homopolymers (HP) of barcodes
//## Default: barcodes with length of 12 nucleotides
//## Values from HP_len=0 to HP_len=12 must provided in order from left to right.
//## HP_len=0: no HP; HP_len=12: HP with length of 12 nucletides (e.g., AAAAAAAAAAAA)
//## And the values must be separated by whitespace or tab.
2 2 2 731049 927666 980547 994770 998595 999624 999900 999975 999995 1000000
==> param_penalty_SR <==
//## Penalty scores for simple repeats with repeat unit of 2 bases (SR)
//## Default: barcodes with length of 12 nucleotides
//## Values from SR_len=0 to SR_len=12 must provided in order from left to right.
//## SR_len=0: no SR; SR_len=6: SR with repeat number of 6 (e.g., AGAGAGAGAGAG)
//## And the values must be separated by whitespace or tab.
6 6 6 950720 997577 999886 1000000
==> param_penalty_HD <==
//## Penalty scores for Hamming distances (HD)
//## Default: barcodes with length of 12 nucleotides
//## Values from HD=0 to HD=12 must provided in order from left to right.
//## HD=0: Two barcodes are 100% identical.; HD=12: Twelve different positions are found.
//## And the values must be separated by whitespace or tab.
1000000 1000000 1000 1 0 0 0 0 0 0 0 0 0
==> param_penalty_CP <==
//## Penalty scores for barcode complementation (CP)
//## Default: barcodes with length of 12 nucleotides
//## NCP = the number of reverse complement bases betwee two barcodes
//## Values from NCP=0 to NCP=12 must provided in order from left to right.
//## And the values must be separated by whitespace or tab.
0 0 0 0 0 0 0 0 0 1 100 10000 999998 99999767
==> param_weights <==
//## Weights for penalty socres of five barcode factors
20.372 #--> GC contents (GCC)
13.653 #--> Homopolymers (HP)
14.673 #--> Simple repeates with repeat unit of 2 bases (SR)
18.314 #--> Hamming distances (HD)
10.904 #--> Barcode complementation (CP)

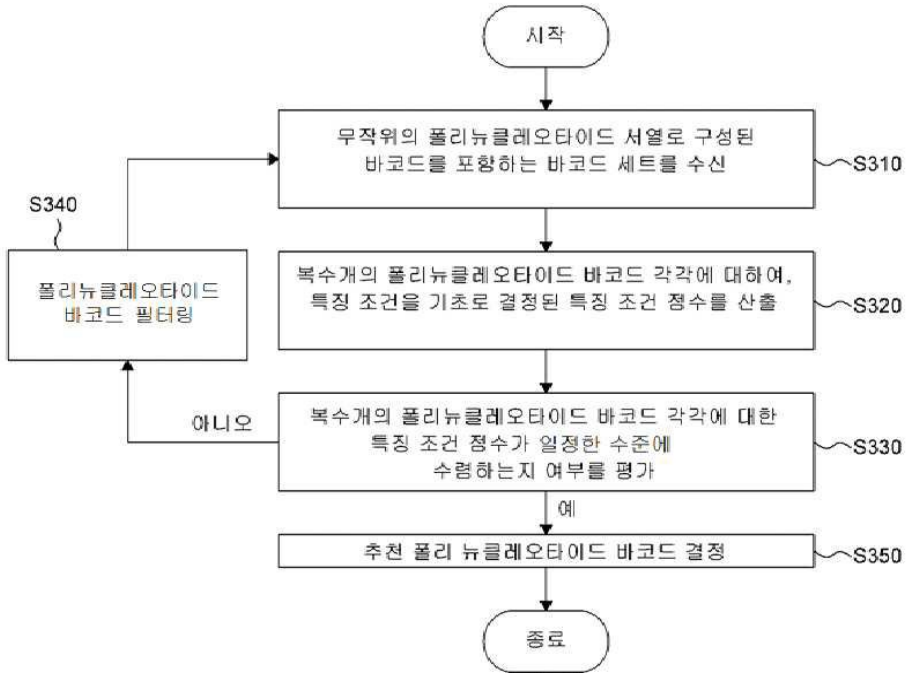
도면1c


131
133
Res_fv3_1/output_00.txt Res_fv3_1/output_05.txt
Res_fv3_1/output_01.txt Res_fv3_1/output_06.txt
Res_fv3_1/output_02.txt Res_fv3_1/output_07.txt
Res_fv3_1/output_03.txt Res_fv3_1/output_08.txt
Res_fv3_1/output_04.txt Res_fv3_1/output_09.txt
Barcode length: 12
Number of barcodes: 100000
Total GCC-Penalty: 7.72128e+09
Total HP-Penalty: 1.37472e+10
Total SR-Penalty: 5.77626e+08
Total HD-Penalty: 7.77846e+09
Total CP-Penalty: 3.14873e+10
Weighted Total Penalty: 8.39256e+11
AAAAAAACCCCG
AAAAAAAGGCCC
AAAAAAGGGCTT
AAAAAATGATTC
AAAAACAAACAC
AAAAACAGGCTT
AAAAACGCTCGG
AAAAACTAGTAG
AAAAACTCGAGC
AAAAAGAGTACC
AAAAAGAGTGAT
AAAAAGCATGCA

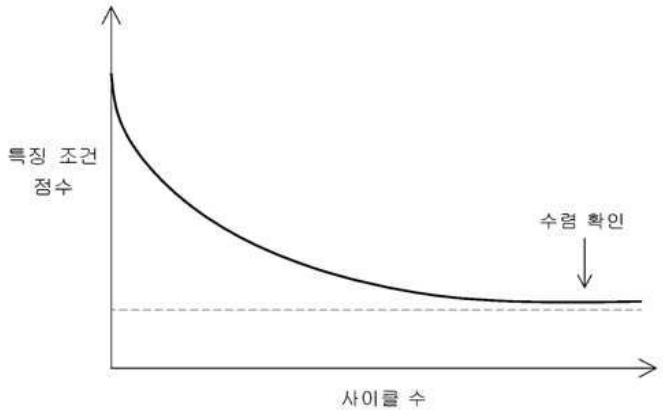
도면2


시작
미리 결정된 뉴클레오타이드 길이에 따라 무작위로 생성된 복수개의 폴리뉴클레오타이드 바코드를 포함하는 바코드 세트를 수신
S210
복수개의 폴리뉴클레오타이드 바코드 각각에 대하여 폴리뉴클레오타이드 바코드의 서열 특징을 기초로 미리 결정된 특징 조건을 이용하여 평가
S220
평가 결과를 기초로 추천 폴리뉴클레오타이드 바코드 세트를 결정
S230
종료

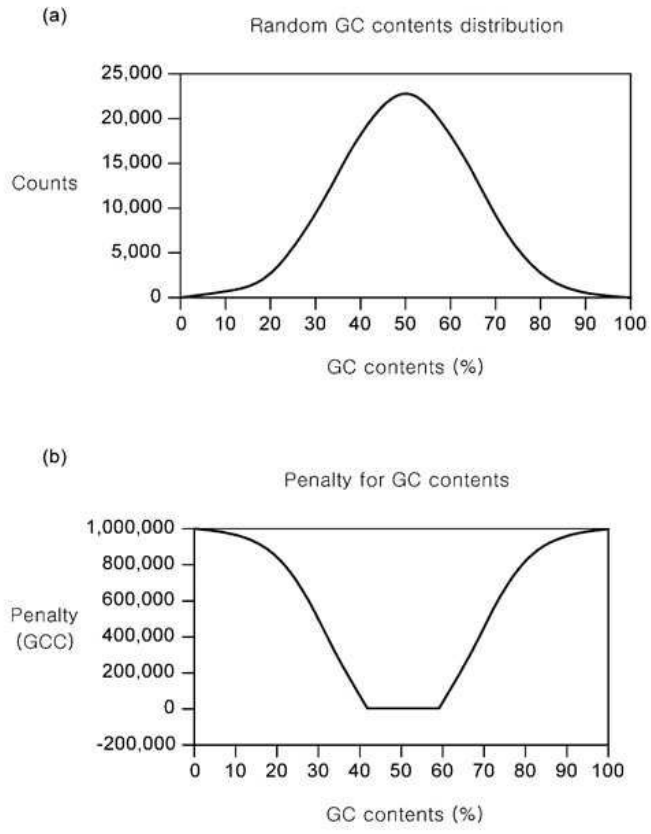
도면3a


시작
무작위의 폴리뉴클레오타이드 서열로 구성된 바코드를 포함하는 바코드 세트를 수신
S310
S340
폴리뉴클레오타이드 바코드 필터링
복수개의 폴리뉴클레오타이드 바코드 각각에 대하여, 특징 조건을 기초로 결정된 특징 조건 점수를 산출
S320
아니오
복수개의 폴리뉴클레오타이드 바코드 각각에 대한 특징 조건 점수가 일정한 수준에 수렴하는지 여부를 평가
S330
예
추천 폴리 뉴클레오타이드 바코드 결정
S350
종료

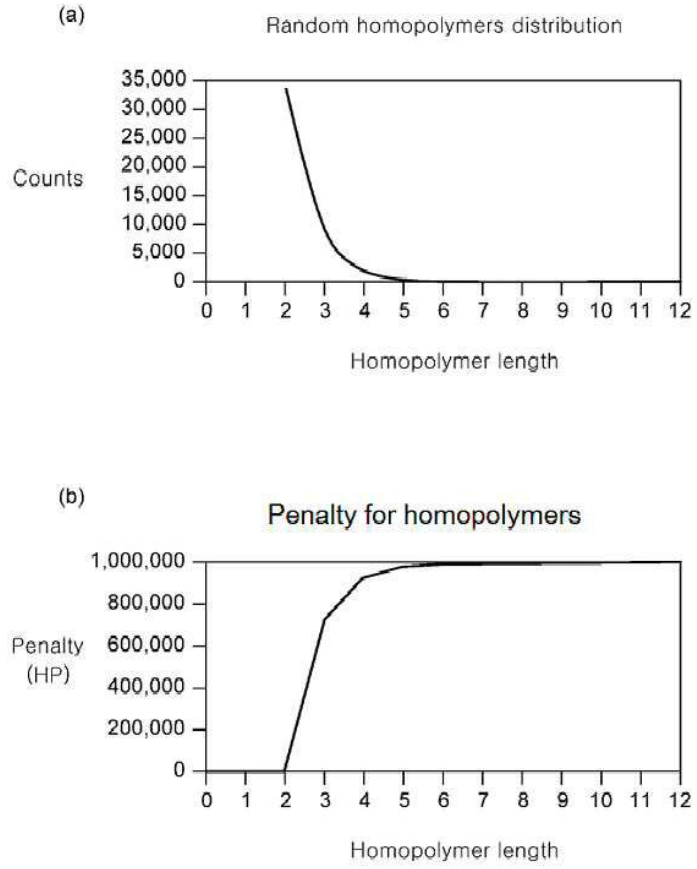
도면3b


특징 조건
점수
수렴 확인
사이클 수

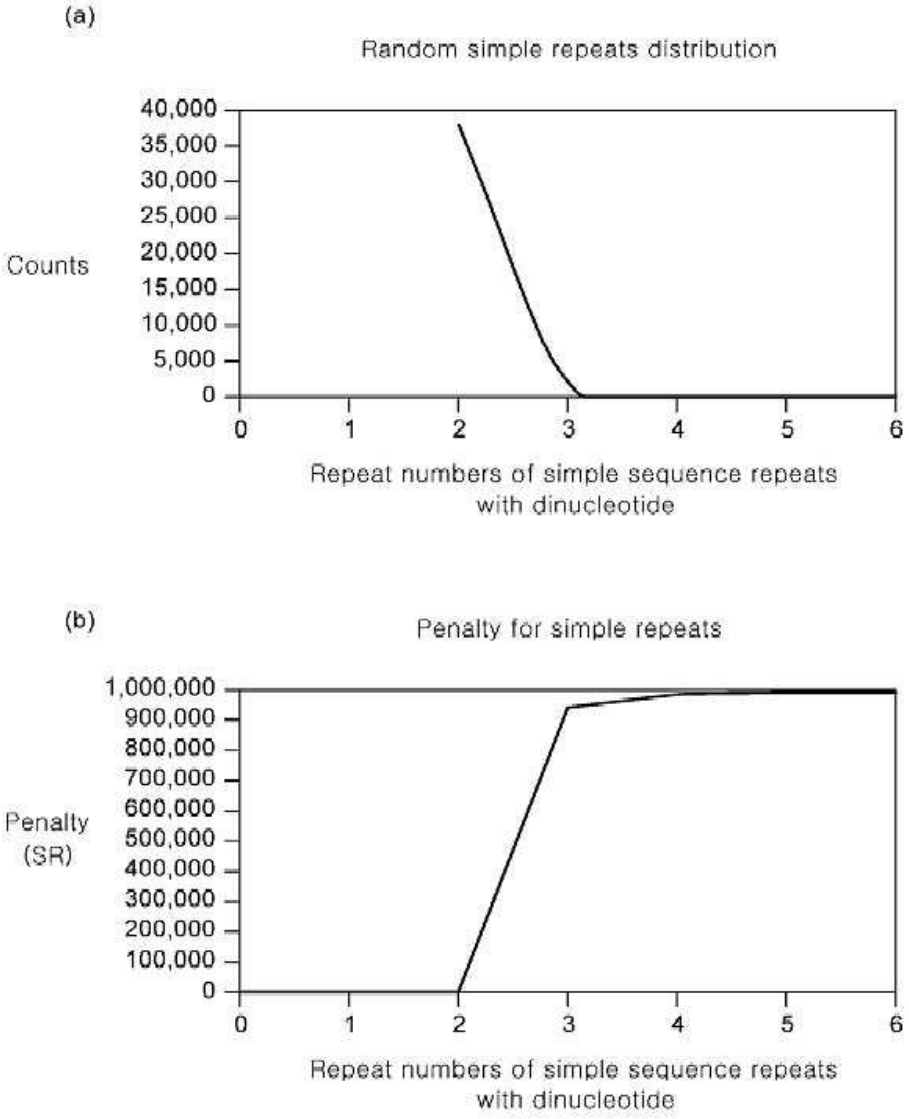
도면4a


(a)
Random GC contents distribution
Counts
25,000
20,000
15,000
10,000
5,000
0
0 10 20 30 40 50 60 70 80 90 100
GC contents (%)
(b)
Penalty for GC contents
Penalty
(GCC)
1,000,000
800,000
600,000
400,000
200,000
0
-200,000
0 10 20 30 40 50 60 70 80 90 100
GC contents (%)

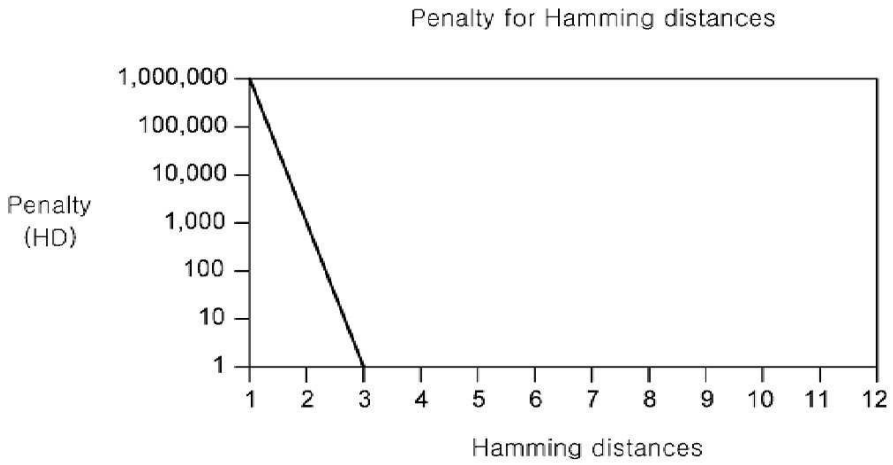
도면4b


(a)
Random homopolymers distribution
Counts
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
0 1 2 3 4 5 6 7 8 9 10 11 12
Homopolymer length
(b)
Penalty for homopolymers
Penalty (HP)
1,000,000
800,000
600,000
400,000
200,000
0
0 1 2 3 4 5 6 7 8 9 10 11 12
Homopolymer length

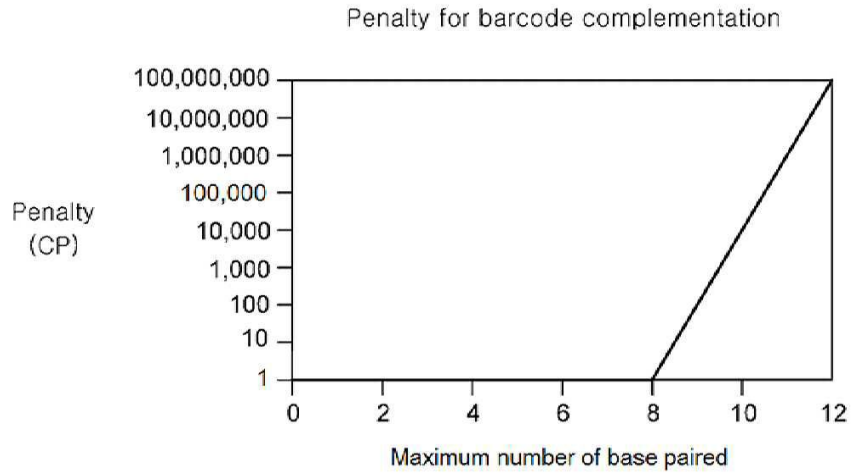
도면4c


(a)
Random simple repeats distribution
Counts
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
0
1
2
3
4
5
6
Repeat numbers of simple sequence repeats with dinucleotide
(b)
Penalty for simple repeats
Penalty (SR)
1,000,000
900,000
800,000
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
0
1
2
3
4
5
6
Repeat numbers of simple sequence repeats with dinucleotide

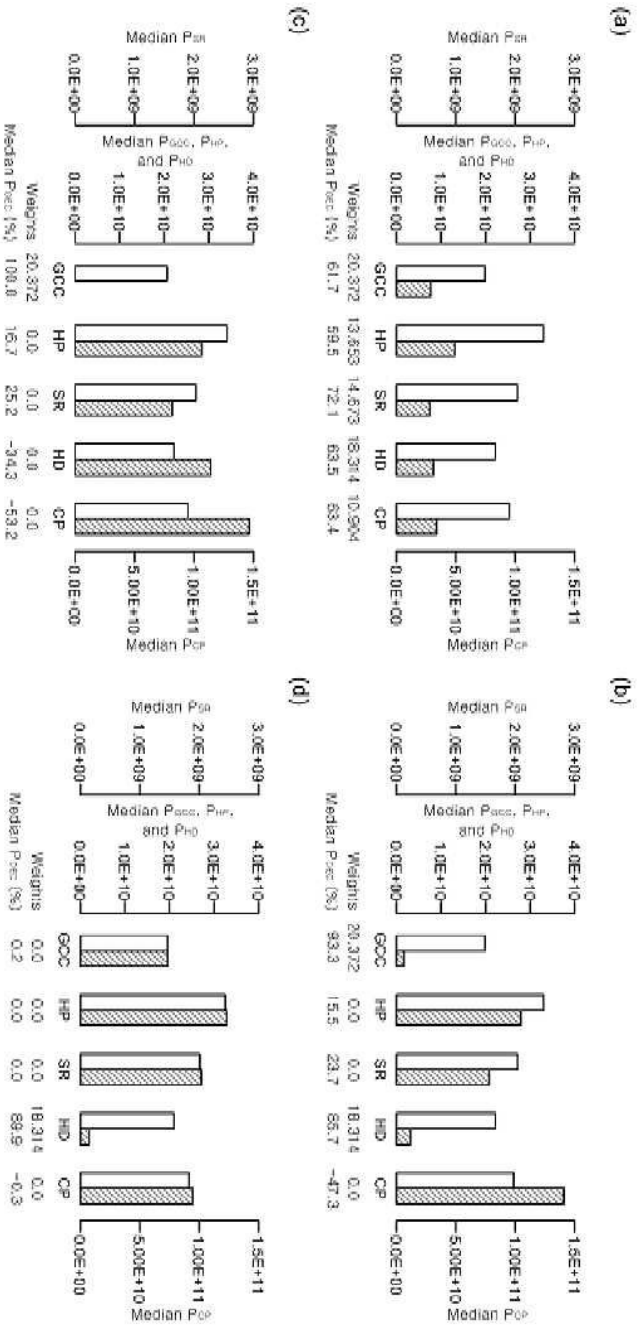
도면4d


Penalty for Hamming distances
Penalty
(HD)
1,000,000
100,000
10,000
1,000
100
10
1
1 2 3 4 5 6 7 8 9 10 11 12
Hamming distances

도면4e


Penalty for barcode complementation
Penalty
(CP)
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
0 2 4 6 8 10 12
Maximum number of base paired

도면4f


(a)
(b)
(c)
(d)
Median PSR
Median PGCC, PHP, and PHD
Median PCP
GCC
HP
SR
HD
CP
Weights
Median Prec (%)