



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년10월29일
(11) 등록번호 10-2722820
(24) 등록일자 2024년10월23일

(51) 국제특허분류(Int. Cl.)
G16B 40/00 (2019.01) G06F 16/22 (2019.01)
G06F 16/28 (2019.01) G16B 25/00 (2019.01)
G16B 99/00 (2019.01)
(52) CPC특허분류
G16B 40/00 (2019.02)
G06F 16/2246 (2019.01)
(21) 출원번호 10-2019-7015857
(22) 출원일자(국제) 2017년11월07일
심사청구일자 2020년11월06일
(85) 번역문제출일자 2019년06월03일
(65) 공개번호 10-2019-0077062
(43) 공개일자 2019년07월02일
(86) 국제출원번호 PCT/US2017/060451
(87) 국제공개번호 WO 2018/089378
국제공개일자 2018년05월17일
(30) 우선권주장
62/419,291 2016년11월08일 미국(US)
62/446,227 2017년01월13일 미국(US)
(56) 선행기술조사문헌
P. D' haeseleer 외, "How does gene
expression clustering work?", Nature
Biotechnology, 23권, 12호, pp.1499-1501,
2005.12.*
T. Sorlie 외, "Gene expression patterns of
breast carcinomas distinguish tumor
subclasses with clinical implication", PNAS,
98권, 19호, pp.10869-10874, 2001.09.11.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
백톤 디킨슨 앤드 컴퍼니
미국, 뉴저지 07417-1880, 프랭클린 레이크스, 1
백톤 드라이브
(72) 발명자
판, 주
미국 94025 캘리포니아주 덴로 파크 스위트 110
캠벨 애비뉴 4040 셀룰러 리서치, 인크. 내
장, 제시
미국 94025 캘리포니아주 덴로 파크 스위트 110
캠벨 애비뉴 4040 셀룰러 리서치, 인크. 내
후, 정
미국 94025 캘리포니아주 덴로 파크 스위트 110
캠벨 애비뉴 4040 셀룰러 리서치, 인크. 내
(74) 대리인
양영준

전체 청구항 수 : 총 17 항

심사관 : 성경아

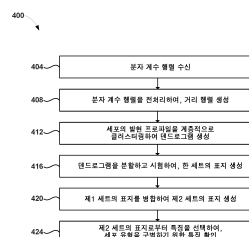
(54) 발명의 명칭 발현 프로파일 분류 방법

(57) 요약

세포 유형을 구별하기 위한 표적의 확인을 위한 방법 및 시스템이 본원에 개시된다. 일부 실시형태에서, 상기 방법은 세포의 발현 프로파일을 계층적으로 클러스터링하여, 각각의 리프(leaf)가 상이한 개별 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하는 단계를 포함한다. 덴드로그램은 무효한 분할(및 그들의 자식)을 제거함으

(뒷면에 계속)

대표도 - 도4



로써 프루닝(pruned)될 수 있다. 텐드로그램의 잔류 리프는 서로에 대한 그들의 거리에 기초하여, 텐드로그램 내의 그들의 위치에 독립적으로 병합되어, 발현 프로파일의 클러스터를 생성할 수 있다. 상기 방법은 클러스터에서 차등적으로 발현되는 표적에 기초하여 세포 유형을 구별하기 위한 표적을 확인할 수 있다.

(52) CPC특허분류

G06F 16/283 (2019.01)

G16B 25/00 (2019.02)

G16B 99/00 (2019.02)

명세서

청구범위

청구항 1

(a) 복수의 세포의 발현 프로파일을 수신하는 단계로서, 상기 발현 프로파일이 복수의 세포 중 각각의 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함하는 단계;

(b) 상기 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계로서,

각각의 클러스터가 (1) 부모 클러스터 및 (2) 2개 이상의 자식 클러스터 중 하나 또는 둘 모두와 하나 이상의 연관성을 가지며, 상기 부모 클러스터가 상기 클러스터에 의해 나타나는 복수의 세포 중 하나 이상의 세포의 발현 프로파일을 나타내며, 상기 클러스터가 2개 이상의 자식 클러스터에 의해 나타나는 발현 프로파일을 나타내고,

상기 복수의 세포의 발현 프로파일을 클러스터링하는 것은, 상기 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 상기 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여 상기 복수의 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하는 것을 포함하며, 상기 덴드로그램이 복수의 클러스터를 포함하며, 상기 복수의 클러스터가 루트 클러스터, 복수의 리프 클러스터 및 복수의 비-루트, 비-리프 클러스터를 포함하고,

상기 복수의 리프 클러스터 및 상기 복수의 비-루트, 비-리프 클러스터의 각각이 부모 클러스터와 연관성을 가지고,

상기 루트 클러스터 및 상기 복수의 비-루트, 비-리프 클러스터의 각각이 좌측 자식 클러스터 및 우측 자식 클러스터와 연관성을 가지며, 클러스터의 좌측 자식 클러스터 및 우측 자식 클러스터에 의해 나타나는 발현 프로파일을 나타내고,

상기 루트 클러스터가 복수의 세포의 발현 프로파일을 나타내는 단계;

(c) 상기 덴드로그램의 루트 클러스터로부터 상기 덴드로그램의 복수의 리프 클러스터로 상기 덴드로그램의 각각의 클러스터를 통해 트레이싱하는 동안, 2개 이상의 자식 클러스터를 갖는 각각의 클러스터에 있어서,

상기 클러스터와 상기 클러스터의 자식 클러스터 사이의 연관성이 유효한지 무효한지를 결정하는 단계로서, 상기 좌측 자식 클러스터와 상기 우측 자식 클러스터 사이의 거리가 연관 임계값 초과이면, 연관성을 유효한 것으로 결정하고, 그렇지 않으면, 무효한 것으로 결정하는 것을 포함하고,

상기 좌측 자식 클러스터와 상기 우측 자식 클러스터 사이의 거리는 상기 좌측 자식 클러스터와 상기 우측 자식 클러스터에 의해 나타나는 발현 프로파일 사이의 복수의 표적 중 각 표적에 대해 수행되는 독립적인 통계 테스트로부터 도출된 p-값으로 표시되고,

상기 클러스터와 2개 이상의 자식 클러스터 사이의 연관성이 무효하면, 상기 클러스터를 병합 클러스터 세트에 추가하는 단계;

(d) 반복적으로, 상기 병합 클러스터 세트 내의 각각의 제1 클러스터에 있어서, 상기 병합 클러스터 세트 내의 제1 클러스터와 제1 클러스터에 가장 가까운 병합 클러스터 세트 내의 제2 클러스터 사이의 거리가 병합 거리 임계값 이내이면, 상기 제1 클러스터와 상기 제2 클러스터를 병합하여, 병합된 클러스터를 생성하는 단계로서, 상기 병합된 클러스터가 상기 제1 클러스터와 상기 제2 클러스터의 발현 프로파일을 포함하는 단계; 및

(e) 병합 클러스터 세트 내의 각각의 클러스터에 있어서, 상기 클러스터에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계를 포함하는 세포 유형을 구별하기 위한 표적의 확인 방법.

청구항 2

제1항에 있어서, 상기 복수의 세포의 발현 프로파일을 수신하는 단계가 표적 계수 데이터 구조를 수신하는 것을 포함하며, 상기 표적 계수 데이터 구조가 표적 계수 행렬을 포함하며, 상기 표적 계수 행렬의 각각의 행 또는

각각의 열이 복수의 세포 중 상이한 개별 세포의 발현 프로파일을 포함하는 방법.

청구항 3

제1항에 있어서, 상기 덴드로그램의 복수의 클러스터를 트래버싱하는 경우 각각의 클러스터에서:

상기 연관이 유효하면, 상기 클러스터로부터 상기 클러스터의 좌측 자식 클러스터 및 우측 자식 클러스터로의 트래버싱을 계속하고;

상기 연관이 무효하면, 상기 클러스터로부터 상기 클러스터의 좌측 자식 클러스터 및 우측 자식 클러스터로의 트래버싱을 중단하는 단계를 포함하는 방법.

청구항 4

제1항에 있어서, 상기 좌측 자식 클러스터와 상기 우측 자식 클러스터 사이의 거리가

상기 좌측 자식 클러스터에 의해 나타나는 발현 프로파일과 상기 우측 자식 클러스터에 의해 나타나는 각각의 발현 프로파일 사이의 복수의 표적 중 각각의 표적에서 수행되는 통계적 검정의 최대 p-값에 기초하여 결정되는 방법.

청구항 5

제1항에 있어서, 상기 덴드로그램의 복수의 클러스터를 트래버싱하는 경우 각 클러스터에서: (3) 상기 클러스터가 단일 세포의 발현 프로파일을 나타낸다면, 상기 클러스터를 상기 병합 클러스터 세트에 추가하는 단계를 포함하는 방법.

청구항 6

제1항에 있어서, 상기 덴드로그램의 복수의 클러스터를 트래버싱하는 경우, 각 클러스터에서: 클러스터 표지를 상기 클러스터에 할당하는 단계를 포함하는 방법.

청구항 7

제6항에 있어서,

상기 클러스터가 단일 세포의 발현 프로파일을 나타내면, 상기 클러스터의 클러스터 표지가 단일 세포 지정을 포함하며,

그렇지 않고, 상기 클러스터가 상기 부모 클러스터의 좌측 자식 클러스터이면, 상기 클러스터의 클러스터 표지가 상기 부모 클러스터의 클러스터 표지 및 좌측 지정을 포함하고,

그렇지 않으면, 상기 클러스터의 클러스터 표지는 상기 부모 클러스터의 클러스터 표지 및 우측 지정을 포함하는 방법.

청구항 8

제1항 내지 제7항 중 어느 한 항에 있어서, 상기 병합 클러스터 세트 내의 각각의 클러스터에 있어서, 상기 클러스터에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계가

상기 클러스터에 의해 나타나는 발현 프로파일과 상기 병합 클러스터 세트 내의 또 다른 클러스터에 의해 나타나는 발현 프로파일 사이의, 세포 유형을 구별하기 위한 표적과 연관된 별개의 서열을 갖는 분자 표지의 수의 차이가 유의성 임계값보다 더 큰 것을 결정하는 것을 포함하는 방법.

청구항 9

제1항 내지 제7항 중 어느 한 항에 있어서, (b)에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계 이전에:

(i) 상기 복수의 세포의 발현 프로파일의 거리 데이터 구조를 결정하는 단계를 포함하며, 상기 거리 데이터 구조가 상기 복수의 세포의 발현 프로파일의 거리 행렬을 포함하며,

(b)에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계가

상기 거리 행렬에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 것을 포함하는 방법.

청구항 10

제9항에 있어서, (i)에서 거리 데이터 구조를 결정하는 단계 이전에, 상기 표적 계수 데이터 구조를 로그-변환된 표적 계수 데이터 구조로 로그-변환하는 단계를 포함하며,

상기 표적 계수 데이터 구조의 요소의 거리 데이터 구조를 결정하는 단계가 상기 로그-변환된 표적 계수 데이터 구조의 거리 데이터 구조를 결정하는 것을 포함하며,

(b)에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계가 상기 로그-변환된 표적 계수 데이터 구조 및 상기 거리 데이터 구조에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 클러스터를 생성하는 것을 포함하는 방법.

청구항 11

제1항 내지 제7항 중 어느 한 항에 있어서, (b)에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하는 단계가

상기 복수의 세포의 각각의 발현 프로파일을 복수의 클러스터 내의 상이한 리프 클러스터에 할당하고;

제2 클러스터가 복수의 클러스터 중 제1 클러스터에 가장 가까운 클러스터이면, 복수의 클러스터 중 제1 클러스터 및 제2 클러스터를 반복적으로 조합하여, 제1 클러스터 및 제2 클러스터의 부모 클러스터를 생성하는 것을 포함하는 방법.

청구항 12

제11항에 있어서, 상기 제1 클러스터와 상기 제2 클러스터 사이의 거리가 상기 제1 클러스터에 의해 나타나는 임의의 발현 프로파일과, 상기 제2 클러스터에 의해 나타나는 임의의 발현 프로파일 사이의 최대 거리인 방법.

청구항 13

제11항에 있어서, 상기 제1 클러스터의 클러스터-내 상관관계 및 상기 제2 클러스터의 클러스터-내 상관관계가 상기 제1 클러스터와 상기 제2 클러스터의 클러스터-간 상관관계보다 더 큰 방법.

청구항 14

제11항에 있어서, 상기 제1 클러스터의 클러스터-내 상관관계 및 상기 제2 클러스터의 클러스터-내 상관관계의 표시가 상기 제1 클러스터 및 상기 제2 클러스터의 클러스터-간 상관관계보다 더 큰 방법.

청구항 15

제14항에 있어서, 상기 제1 클러스터의 클러스터-내 상관관계 및 상기 제2 클러스터의 클러스터-내 상관관계의 표시가

상기 제1 클러스터 및 상기 제2 클러스터의 클러스터-내 최대 상관관계,

상기 제1 클러스터 및 상기 제2 클러스터의 클러스터-내 평균 상관관계,

상기 제1 클러스터 및 상기 제2 클러스터의 클러스터-내 중간값 상관관계,

상기 제1 클러스터 및 상기 제2 클러스터의 클러스터-내 최소 상관관계 및

그의 임의의 조합

중 적어도 하나에 기초하며, 상기 제1 클러스터의 클러스터-내 상관관계가

상기 제1 클러스터의 클러스터-내 최대 상관관계,

상기 제1 클러스터의 클러스터-내 평균 상관관계,

상기 제1 클러스터의 클러스터-내 중간값 상관관계,

상기 제1 클러스터의 클러스터-내 최소 상관관계 및

그의 임의의 조합

중 적어도 하나에 기초하며, 상기 제2 클러스터의 클러스터-내 상관관계가

상기 제2 클러스터의 클러스터-내 최대 상관관계,

상기 제2 클러스터의 클러스터-내 평균 상관관계,

상기 제2 클러스터의 클러스터-내 중간값 상관관계,

상기 제2 클러스터의 클러스터-내 최소 상관관계 및

그의 임의의 조합

중 적어도 하나에 기초하며, 상기 제1 클러스터 및 상기 제2 클러스터의 클러스터-간 상관관계가

상기 제1 클러스터 및 상기 제2 클러스터의 클러스터-간 최대 상관관계,

상기 제1 클러스터 및 상기 제2 클러스터의 클러스터-간 평균 상관관계,

상기 제1 클러스터 및 상기 제2 클러스터의 클러스터-간 중간값 상관관계,

상기 제1 클러스터 및 상기 제2 클러스터의 클러스터-간 최소 상관관계 및

그의 임의의 조합

중 적어도 하나에 기초하는 방법.

청구항 16

하드웨어 프로세서; 및

명령어가 저장된 비-일시적 메모리로서, 상기 하드웨어 프로세서에 의해 실행되는 경우 상기 프로세서가 제1항 내지 제7항 중 어느 한 항의 방법을 수행하게 하는 비-일시적 메모리를 포함하는,

표적의 수를 결정하기 위한 컴퓨터 시스템.

청구항 17

제1항 내지 제7항 중 어느 한 항의 방법을 수행하기 위한 코드가 저장된 컴퓨터 판독 가능한 유형 저장 매체.

청구항 18

삭제

청구항 19

삭제

청구항 20

삭제

청구항 21

삭제

청구항 22

삭제

청구항 23

삭제

청구항 24

삭제

청구항 25

삭제

청구항 26

삭제

청구항 27

삭제

청구항 28

삭제

청구항 29

삭제

청구항 30

삭제

청구항 31

삭제

청구항 32

삭제

청구항 33

삭제

청구항 34

삭제

청구항 35

삭제

청구항 36

삭제

청구항 37

삭제

청구항 38

삭제

청구항 39

삭제

청구항 40

삭제

청구항 41

삭제

청구항 42

삭제

청구항 43

삭제

청구항 44

삭제

청구항 45

삭제

청구항 46

삭제

청구항 47

삭제

청구항 48

삭제

청구항 49

삭제

청구항 50

삭제

청구항 51

삭제

청구항 52

삭제

청구항 53

삭제

청구항 54

삭제

청구항 55

삭제

청구항 56

삭제

청구항 57

삭제

청구항 58

삭제

청구항 59

삭제

청구항 60

삭제

청구항 61

삭제

청구항 62

삭제

청구항 63

삭제

청구항 64

삭제

청구항 65

삭제

청구항 66

삭제

청구항 67

삭제

청구항 68

삭제

청구항 69

삭제

청구항 70

삭제

청구항 71

삭제

청구항 72

삭제

청구항 73

삭제

청구항 74

삭제

청구항 75

삭제

발명의 설명

기술 분야

- [0001] 관련 출원
- [0002] 본 출원은 2016년 11월 8일 출원된 미국 가출원 제62/419291호; 및 2017년 1월 13일 출원된 미국 가출원 제 62/446227호에 대한 우선권을 주장한다. 이들 관련 출원의 각각의 내용은 본원에 그의 전문이 명시적으로 참조로 포함된다.
- [0003] 저작권 및 상표 공지
- [0004] 본 특허 문서의 개시내용의 일부는 저작권 보호를 받는 자료를 포함하고 있다. 저작권자는 특허청 및 상표청의 특허 파일 또는 기록물에 나타난 바와 같이 특허 문서 또는 특허 개시내용을 누구라도 팩시밀리로 복사하는 것에 이의가 없지만, 그 외에는 무엇이든 모든 저작권을 보유한다.
- [0005] 기술분야
- [0006] 본 발명은 일반적으로 발현 프로파일의 분류, 특히 세포 유형을 구별하기 위한 표적의 확인의 분야에 관한 것이다.

배경 기술

- [0007] 바코딩(예를 들어, 확률적 바코딩)과 같은 방법 및 기법이 세포 분석에 유용하다. 예를 들어, 바코딩은 예를 들어, 역전사, 중합효소 연쇄 반응(PCR) 증폭 및 차세대 시퀀싱(NGS)을 사용하여, 단일 세포의 유전자 발현 프로파일을 판독하여 그들의 상태를 결정하기 위하여 사용될 수 있다. 그러나, 세포 유형을 구별하기 위한 마커를 확인하고, 분석되는 세포의 유형을 결정하기 위하여, 이들 방법 및 기법에 의해 생성되는 다량의 데이터를 추가로 분석할 필요가 있다.

발명의 내용

- [0008] 세포 유형을 구별하기 위한 표적의 확인 방법이 본원에 개시된다. 일부 실시형태에서, 상기 방법은 (a) 표적 계수 데이터 구조를 수신하는 단계로서, 표적 계수 데이터 구조가 복수의 세포의 발현 프로파일을 포함하며, 복수의 세포의 발현 프로파일이 복수의 세포 중 각각의 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함하는 단계; (b) 표적 계수 데이터 구조 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 복수의 세포의 발현 프로파일을 나타내는 덴드로그램(dendrogram)을 생성하는 단계로서, 덴드로그램이 복수의 노드(node)를 포함하며, 복수의 노드가 루트(root) 노드, 복수의 리프(leaf) 노드 및 복수의 비-루트, 비-리프 노드를 포함하며, 복수의 리프 노드 중 각각의 리프 노드가 복수의 세포 중 상이한 세포의 발현 프로파일을 나타내며, 루트 노드가 복수의 세포의 발현 프로파일을 나타내는 단계; (c) 덴드로그램의 루트 노드로부터 덴드로그램의 복수의 리프 노드로 덴드로그램의 복수의 노드 중 각각의 노드

를 통해 트래버싱하는 동안: (1) 노드의 자식 노드로의 노드의 분할이 유효한지 무효한지(예를 들어, 자식 노드 사이의 차이가 유의미하지 않음)를 결정하고; (2) 노드의 자식 노드로의 노드의 분할이 무효하면, 노드를 병합 클러스터 세트에 추가하는 단계; (d) 반복적으로, 병합 클러스터 세트 내의 각각의 제1 노드에 있어서, 병합 클러스터 세트 내의 제1 노드와 제1 노드에 가장 가까운 병합 클러스터 세트 내의 제2 노드 사이의 거리가 병합 거리 임계값 내에 있으면, 제1 노드와 제2 노드를 병합하여, 제1 노드 및 제2 노드에 의해 나타나는 발현 프로파일을 포함하는 병합된 노드를 생성하는 단계; 및 (e) 병합 클러스터 세트 내의 각각의 노드에 있어서, 노드에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계를 포함한다.

[0009] 일부 실시형태에서, 표적 계수 데이터 구조는 표적 계수 행렬을 포함한다. 표적 계수 행렬의 각각의 행 또는 각각의 열은 복수의 세포 중 상이한 개별 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함할 수 있다.

[0010] 일부 실시형태에서, 복수의 리프 노드 및 복수의 비-루트, 비-리프 노드의 각각은 부모 노드와 연관될 수 있으며, 루트 노드 및 복수의 비-루트, 비-리프 노드의 각각은 좌측 자식 노드 및 우측 자식 노드와 연관될 수 있으며, 노드의 좌측 자식 노드 및 우측 자식 노드에 의해 나타나는 발현 프로파일을 나타낸다.

[0011] 일부 실시형태에서, 상기 방법은 (a)에서 표적 계수 데이터 구조를 수신하는 단계 이전에: (f) 복수의 바코드를 사용하여 복수의 세포 내의 복수의 표적을 바코딩하여, 복수의 바코딩된 표적을 생성하는 단계로서, 복수의 바코드의 각각이 세포 표지 및 분자 표지를 포함하며, 상이한 세포의 표적으로부터 생성된 바코딩된 표적이 상이한 세포 표지를 가지며, 복수의 세포 중 하나의 세포의 표적으로부터 생성된 바코딩된 표적이 상이한 분자 표지를 갖는 단계; (g) 복수의 바코딩된 표적의 시퀀싱 데이터를 획득하는 단계; 및 (h) 복수의 세포의 각각에 있어서: (1) 세포에 대한 시퀀싱 데이터에서 복수의 표적 중 각각의 표적과 연관된 별개의 서열을 갖는 분자 표지의 수를 계수하는 단계; 및 (2) (h)(1)에서 계수된 시퀀싱 데이터에서의 표적과 연관된 별개의 서열을 갖는 분자 표지의 수에 기초하여 세포에 대한 복수의 표적 중 각각의 표적의 수를 추정하는 단계를 포함한다. 예를 들어, 상기 방법은 (a)에서 표적 계수 데이터 구조를 수신하는 단계 이전에: 단계(들) (f) 복수의 확률적 바코드를 사용하여 복수의 세포 내의 복수의 표적을 확률적으로 바코딩하여, 복수의 확률적으로 바코딩된 표적을 생성하는 단계로서, 복수의 확률적 바코드의 각각이 세포 표지 및 분자 표지를 포함하며, 상이한 세포의 표적으로부터 생성된 확률적으로 바코딩된 표적이 상이한 세포 표지를 가지며, 복수의 세포 중 하나의 세포의 표적으로부터 생성된 확률적으로 바코딩된 표적이 상이한 분자 표지를 갖는 단계; 및/또는 (g) 복수의 확률적으로 바코딩된 표적의 시퀀싱 데이터를 획득하는 단계를 포함할 수 있다. 표적 계수 데이터 구조를 수신하는 단계는 (h)(2)에서 추정된 세포에 대한 복수의 표적 중 각각의 표적의 수로부터 표적 계수 데이터 구조를 생성하는 것을 포함할 수 있으며, 복수의 세포 중 세포의 발현 프로파일은 (h)(2)에서 추정된 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함한다.

[0012] 일부 실시형태에서, 상기 방법은 (b)에서 표적 계수 데이터 구조 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 복수의 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하는 단계 이전에: (i) 표적 계수 행렬의 요소의 거리 데이터 구조를 결정하는 단계로서, 거리 데이터 구조가 복수의 세포의 발현 프로파일 사이의 거리를 포함하는 단계를 포함한다. 거리 데이터 구조는 거리 행렬을 포함할 수 있다. 거리 행렬의 각각의 대각선 요소는 0의 값을 가질 수 있다. (b)에서 표적 계수 데이터 구조 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 복수의 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하는 단계는 표적 계수 데이터 구조 및 거리 데이터 구조에 기초하여 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하는 것을 포함할 수 있다. 복수의 세포의 발현 프로파일 사이의 거리는 복수의 세포의 발현 프로파일 사이의 쌍별 상관관계 거리를 포함할 수 있다.

[0013] 일부 실시형태에서, 상기 방법은 (i)에서 표적 계수 데이터 구조의 요소의 거리 데이터 구조를 결정하는 단계 이전에, 표적 계수 데이터 구조를 로그-변환된 표적 계수 데이터 구조로 로그-변환하는 단계를 포함하며, 표적 계수 데이터 구조의 요소의 거리 데이터 구조를 결정하는 단계가 로그-변환된 표적 계수 데이터 구조의 거리 데이터 구조를 결정하는 것을 포함하며, (b)에서 표적 계수 데이터 구조 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하는 단계는 로그-변환된 표적 계수 데이터 구조 및 거리 데이터 구조에 기초하여 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 덴드로그램을 생성하는 것을 포함한다. 로그-변환된 표적 계수 데이터 구조로 표적 계수 데이터 구조를 로그-변환하는 단계는 표적 계수 데이터 구조의 각각의 요소의 값을 증분만큼(예컨대 1) 증가시키는 것을 포함할 수 있다.

- [0014] 일부 실시형태에서, (b)에서 표적 계수 데이터 구조 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하는 단계는 복수의 세포의 각각의 발현 프로파일을 상이한 리프 노드에 할당하고; 제2 노드가 복수의 노드에서 제1 노드에 가장 가까운 노드이면, 복수의 노드 중 제1 노드 및 제2 노드를 반복적으로 조합하여, 제1 노드 및 제2 노드의 부모 노드를 생성하는 것을 포함한다. 제1 노드와 제2 노드 사이의 거리는 제1 노드에 의해 나타나는 발현 프로파일을 갖는 임의의 세포와 제2 노드에 의해 나타나는 발현 프로파일을 갖는 임의의 세포 사이의 최대 거리이다.
- [0015] 일부 실시형태에서, 상기 방법은 덴드로그램의 복수의 노드를 트래버싱하는 경우에, 각 노드에서: 분할이 유효하면, 노드로부터 노드의 좌측 자식 노드 및 우측 자식 노드로의 트래버싱을 계속하는 단계; 및 분할이 무효하면, 노드로부터 노드의 좌측 자식 노드 및 우측 자식 노드로의 트래버싱을 중단하는 단계를 포함한다. 제1 노드의 노드-내 상관관계 및 제2 노드의 노드-내 상관관계 중 적어도 하나는 제1 노드 및 제2 노드의 노드-간 상관관계보다 더 클 수 있다. 제1 노드의 노드-내 상관관계 및 제2 노드의 노드-내 상관관계의 척도 또는 표시는 제1 노드 및 제2 노드의 노드-간 상관관계보다 더 클 수 있다. 제1 노드의 노드-내 상관관계 및 제2 노드의 노드-내 상관관계의 척도는 다음 중 적어도 하나에 기초할 수 있다: 제1 노드 및 제2 노드의 노드-내 최대 상관관계, 제1 노드 및 제2 노드의 노드-내 평균 상관관계, 제1 노드 및 제2 노드의 노드-내 중간값 상관관계, 제1 노드 및 제2 노드의 노드-내 최소 상관관계 및 그의 임의의 조합. 제1 노드의 노드-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 노드의 노드-내 최대 상관관계, 제1 노드의 노드-내 평균 상관관계, 제1 노드의 노드-내 중간값 상관관계, 제1 노드의 노드-내 최소 상관관계 및 그의 임의의 조합. 제2 노드의 노드-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제2 노드의 노드-내 최대 상관관계, 제2 노드의 노드-내 평균 상관관계, 제2 노드의 노드-내 중간값 상관관계, 제2 노드의 노드-내 최소 상관관계 및 그의 임의의 조합. 제1 노드 및 제2 노드의 노드-간 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 노드 및 제2 노드의 노드-간 최대 상관관계, 제1 노드 및 제2 노드의 노드-간 평균 상관관계, 제1 노드 및 제2 노드의 노드-간 중간값 상관관계, 제1 노드 및 제2 노드의 노드-간 최소 상관관계 및 그의 임의의 조합.
- [0016] 일부 실시형태에서, 노드와 노드의 자식 노드의 분할이 유효한지 무효한지를 결정하는 단계는 좌측 자식 노드와 우측 자식 노드 사이의 거리가 분할 임계값 초과이면, 분할을 유효한 것으로 결정하고, 그렇지 않으면 무효한 것으로 결정하는 것을 포함한다. 좌측 자식 노드와 우측 자식 노드 사이의 거리는 좌측 자식 노드와 우측 자식 노드에 의해 나타나는 발현 프로파일 사이의 복수의 표적 중 각각의 표적에서 수행되는 통계적 검정에 기초하여 결정될 수 있다. 통계적 검정은 웰치의 t-검정(Welch's *t*-test)을 포함할 수 있다. 좌측 자식 노드와 우측 자식 노드 사이의 거리는 좌측 자식 노드에 의해 나타나는 각각의 발현 프로파일과 우측 자식 노드에 의해 나타나는 각각의 발현 프로파일 사이에서 복수의 표적 중 각각의 표적에서 수행되는 통계적 검정의 최대 p-값에 기초하여 결정될 수 있다.
- [0017] 일부 실시형태에서, 노드와 노드의 자식 노드의 분할이 유효한지 무효한지를 결정하는 단계는 제1 노드의 노드-내 상관관계 및 제2 노드의 노드-내 상관관계 중 적어도 하나가 제1 노드 및 제2 노드의 노드-간 상관관계보다 더 크면, 분할을 유효한 것으로 결정하고, 그렇지 않으면 무효한 것으로 결정하는 것을 포함한다. 일부 실시형태에서, 노드와 노드의 자식 노드의 분할이 유효한지 무효한지를 결정하는 단계는 제1 노드의 노드-내 상관관계 및 제2 노드의 노드-내 상관관계의 척도 또는 표시가 제1 노드 및 제2 노드의 노드-간 상관관계보다 더 크면, 분할을 유효한 것으로 결정하고, 그렇지 않으면 무효한 것으로 결정하는 것을 포함한다. 제1 노드의 노드-내 상관관계 및 제2 노드의 노드-내 상관관계의 척도는 다음 중 적어도 하나에 기초할 수 있다: 제1 노드 및 제2 노드의 노드-내 최대 상관관계, 제1 노드 및 제2 노드의 노드-내 평균 상관관계, 제1 노드 및 제2 노드의 노드-내 중간값 상관관계, 제1 노드 및 제2 노드의 노드-내 최소 상관관계 및 그의 임의의 조합. 제1 노드의 노드-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 노드의 노드-내 최대 상관관계, 제1 노드의 노드-내 평균 상관관계, 제1 노드의 노드-내 중간값 상관관계, 제1 노드의 노드-내 최소 상관관계 및 그의 임의의 조합. 제2 노드의 노드-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제2 노드의 노드-내 최대 상관관계, 제2 노드의 노드-내 평균 상관관계, 제2 노드의 노드-내 중간값 상관관계, 제2 노드의 노드-내 최소 상관관계 및 그의 임의의 조합. 제1 노드 및 제2 노드의 노드-간 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 노드 및 제2 노드의 노드-간 최대 상관관계, 제1 노드 및 제2 노드의 노드-간 평균 상관관계, 제1 노드 및 제2 노드의 노드-간 중간값 상관관계, 제1 노드 및 제2 노드의 노드-간 최소 상관관계 및 그의 임의의 조합.
- [0018] 일부 실시형태에서, 상기 방법은 덴드로그램의 복수의 노드를 트래버싱하는 경우 각 노드에서: (3) 노드가 단일 세포의 발현 프로파일을 나타낸다면, 노드를 병합 클러스터 세트에 추가하는 단계를 포함한다. 일부 실시형태에서, 상기 방법은 덴드로그램의 복수의 노드를 트래버싱하는 경우, 각 노드에서: 노드 표지를 노드에 할당하는

단계를 포함할 수 있다. 노드가 단일 세포의 발현 프로파일을 나타낸다면, 노드의 노드 표지는 단일 세포 지정을 포함하며, 그렇지 않고 노드가 부모 노드의 좌측 자식 노드이면, 노드의 노드 표지는 부모 노드의 노드 표지 및 좌측 지정을 포함하고, 그렇지 않으면, 노드의 노드 표지는 부모 노드의 노드 표지 및 우측 지정을 포함한다.

[0019] 일부 실시형태에서, 병합 클러스터 세트 내의 각각의 노드에 있어서, 노드에 의해 나타나는 세포의 복수의 표지의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계는 노드에 의해 나타나는 발현 프로파일과 병합 클러스터 세트 내의 또 다른 노드에 의해 나타나는 발현 프로파일 사이의, 세포 유형을 구별하기 위한 표적과 연관된 별개의 서열을 갖는 분자 표지의 수의 차이가 유의성 임계값보다 더 큰 것을 결정하는 것을 포함한다.

[0020] 일부 실시형태에서, 상기 방법은 (d)에서 제1 노드와 제2 노드를 병합하여 병합된 노드를 생성하기 이전에: 제3 노드와 제4 노드 사이의 거리가 노드 거리 임계값 이내이면, 단일 세포의 발현 프로파일을 나타내는 병합 클러스터 세트 내의 각각의 제3 노드를 병합 클러스터 세트 내의 제4 노드와 병합하는 단계를 포함한다. 일부 실시형태에서, 상기 방법은 세포의 발현 프로파일을 나타내는 병합 클러스터 세트 내의 노드에 기초하여 복수의 세포를 분류하는 단계를 포함한다. 상기 방법은 확인된 세포 유형을 구별하기 위한 표적에 기초하여, 전체 전사체 검정을 설계하는 단계를 포함할 수 있다. 일부 실시형태에서, 상기 방법은 확인된 세포 유형을 구별하기 위한 표적에 기초하여, 표적화된 전사체 검정을 설계하는 단계를 포함할 수 있다.

[0021] 세포 유형을 구별하기 위한 표적의 확인 방법이 본원에 개시된다. 일부 실시형태에서, 상기 방법은 (a) 복수의 세포의 발현 프로파일을 수신하는 단계로서, 발현 프로파일이 복수의 세포 중 각각의 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함하는 단계; (b) 표적 계수 행렬 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 복수의 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하는 단계로서, 덴드로그램이 복수의 노드를 포함하며, 복수의 노드가 루트 노드, 복수의 리프 노드 및 복수의 비-루트, 비-리프 노드를 포함하며, 복수의 리프 노드 중 각각의 리프 노드가 복수의 세포 중 상이한 세포의 발현 프로파일을 나타내며, 루트 노드가 복수의 세포의 발현 프로파일을 나타내는 단계; (c) 덴드로그램의 루트 노드로부터 덴드로그램의 복수의 리프 노드로 덴드로그램의 복수의 노드 중 각각의 노드를 통해 트레이싱하는 동안: (1) 노드의 (예를 들어, 노드의 자식 노드에 의해 나타나는) 2개의 하위-분지가 유의미하게 상이한지를 결정하고; (2) 노드의 2개의 하위-분지가 유의미하게 상이하면, (예를 들어, 노드의 2개의 하위-분지로 트레이싱함으로써) 노드를 2개의 클러스터 세트로 분할하는 단계를 포함한다. 일부 실시형태에서, 상기 방법은 (3) 노드의 자식 노드로의 노드의 분할이 무효하면, 노드를 병합 클러스터 세트에 추가하는 단계를 포함한다. 일부 실시형태에서, 상기 방법은 (d) 반복적으로, 병합 클러스터 세트 내의 각각의 제1 노드에 있어서, 병합 클러스터 세트 내의 제1 노드와 제1 노드에 가장 가까운 병합 클러스터 세트 내의 제2 노드 사이의 거리가 병합 거리 임계값 이내이면, 제1 노드를 제2 노드와 병합하여, 병합 클러스터 세트 내에 병합된 노드를 생성하는 단계; 및 (e) 병합 클러스터 세트 내의 각각의 노드에 있어서, 노드에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계를 포함한다.

[0022] 세포 유형을 구별하기 위한 표적의 확인 방법이 본원에 개시된다. 일부 실시형태에서, 상기 방법은 (a) 복수의 세포의 발현 프로파일을 수신하는 단계로서, 발현 프로파일이 복수의 세포 중 각각의 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함하는 단계; (b) 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계로서, 각각의 클러스터가 (1) 부모 클러스터 및 (2) 2개 이상의 자식 클러스터 중 하나 또는 둘 모두와 하나 이상의 연관을 가지며, 부모 클러스터가 클러스터에 의해 나타나는 복수의 세포 중 하나 이상의 세포의 발현 프로파일을 나타내며, 클러스터가 2개 이상의 자식 클러스터에 의해 나타나는 발현 프로파일을 나타내는 단계; (c) 2개 이상의 자식 클러스터를 갖는 각각의 클러스터에 있어서, 클러스터와 2개 이상의 자식 클러스터 사이의 연관이 무효하면(예를 들어, 2개 이상의 자식 클러스터 사이의 차이가 유의미하지 않으면), 클러스터를 병합 클러스터 세트에 추가하는 단계; (d) 반복적으로, 병합 클러스터 세트 내의 각각의 제1 클러스터에 있어서, 병합 클러스터 세트 내의 제1 클러스터와 제1 클러스터와 가장 가까운 병합 클러스터 세트 내의 제2 클러스터 사이의 거리가 병합 거리 임계값 이내이면, 제1 클러스터와 제2 클러스터를 병합하여, 병합된 클러스터를 생성하는 단계로서, 병합된 클러스터가 제1 클러스터와 제2 클러스터의 발현 프로파일을 포함하는 단계; 및 (e) 병합 클러스터 세트 내의 각각의 클러스터에 있어서, 클러스터에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계를 포함한다.

[0023] 일부 실시형태에서, 복수의 세포의 발현 프로파일을 수신하는 단계는 표적 계수 데이터 구조를 수신하는 것을

포함한다. 표적 계수 데이터 구조는 표적 계수 행렬을 포함할 수 있다. 표적 계수 행렬의 각각의 행 또는 각각의 열은 복수의 세포 중 상이한 개별 세포의 발현 프로파일을 포함할 수 있다. 복수의 세포의 발현 프로파일을 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 발현 프로파일의 클러스터로 클러스터링하는 단계는 복수의 세포의 발현 프로파일을 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 계층적으로 클러스터링하여 복수의 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하는 것을 포함할 수 있다. 덴드로그램은 복수의 클러스터를 포함할 수 있다. 복수의 클러스터는 루트 클러스터, 복수의 리프 클러스터 및 복수의 비-루트, 비-리프 클러스터를 포함할 수 있다. 리프 클러스터는 하나의 세포의 발현 프로파일을 나타낼 수 있다. 비-루트, 비-리프 클러스터는 비-루트, 비-리프 클러스터의 자식 클러스터에 의해 나타나는 세포의 발현 프로파일을 나타낼 수 있다. 루트 클러스터는 그의 자식 클러스터의 발현 프로파일을 나타낼 수 있다. 복수의 리프 클러스터, 및 복수의 비-루트, 비-리프 클러스터의 각각은 부모 클러스터와 연관될 수 있다. 루트 클러스터 및 복수의 비-루트, 비-리프 클러스터의 각각은 좌측 자식 클러스터와 우측 자식 클러스터와의 연관성을 가질 수 있으며, 클러스터의 좌측 자식 클러스터 및 우측 자식 클러스터에 의해 나타나는 발현 프로파일을 나타낸다. 루트 클러스터는 복수의 세포의 발현 프로파일을 나타낼 수 있다.

[0024] 일부 실시형태에서, 2개 이상의 자식 클러스터를 갖는 각각의 클러스터에 있어서, 클러스터와 2개 이상의 자식 클러스터 사이의 연관성이 무효하면, 클러스터를 병합 클러스터 세트에 추가하는 단계는 덴드로그램의 루트 클러스터로부터 덴드로그램의 복수의 리프 클러스터로 덴드로그램의 각각의 클러스터를 통해 트래버싱하는 동안 (1) 클러스터와 클러스터의 자식 클러스터의 연관성이 유효한지 무효한지를 결정하고; (2) 연관성이 무효하면, 클러스터를 병합 클러스터 세트에 추가하는 것을 포함한다.

[0025] 일부 실시형태에서, 상기 방법은 (a)에서 복수의 세포의 발현 프로파일을 수신하는 단계 이전에: (f) 복수의 바코드를 사용하여 복수의 세포 내의 복수의 표적을 바코딩하여, 복수의 바코딩된 표적을 생성하는 단계로서, 복수의 바코드의 각각이 세포 표지 및 분자 표지를 포함하며, 상이한 세포의 표적으로부터 생성된 바코딩된 표적이 상이한 세포 표지를 가지며, 복수의 세포 중 하나의 세포의 표적으로부터 생성된 바코딩된 표적이 상이한 분자 표지를 갖는 단계; (g) 복수의 바코딩된 표적의 시퀀싱 데이터를 획득하는 단계; 및 (h) 복수의 세포의 각각에 있어서: (1) 세포에 대한 시퀀싱 데이터에서 복수의 표적 중 각각의 표적과 연관된 별개의 서열을 갖는 분자 표지의 수를 계수하고; (2) (h)(1)에서 계수된 시퀀싱 데이터에서의 표적과 연관된 별개의 서열을 갖는 분자 표지의 수에 기초하여 세포에 대한 복수의 표적 중 각각의 표적의 수를 추정하는 단계를 포함한다. 예를 들어, 상기 방법은 (a)에서 복수의 세포의 발현 프로파일을 수신하는 단계 이전에: 단계(들) (f) 복수의 확률적 바코드를 사용하여 복수의 세포 내의 복수의 표적을 확률적으로 바코딩하여 복수의 확률적으로 바코딩된 표적을 생성하는 단계로서, 복수의 확률적 바코드의 각각이 세포 표지 및 분자 표지를 포함하며, 상이한 세포의 표적으로부터 생성된 확률적으로 바코딩된 표적이 상이한 세포 표지를 가지며, 복수의 세포 중 하나의 세포의 표적으로부터 생성된 확률적으로 바코딩된 표적이 상이한 분자 표지를 갖는 단계; 및/또는 (g) 복수의 확률적으로 바코딩된 표적의 시퀀싱 데이터를 획득하는 단계를 포함할 수 있다.

[0026] 일부 실시형태에서, 복수의 세포 중 세포의 발현 프로파일은 (h)(2)에서 추정된 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함한다. 일부 실시형태에서, 상기 방법은 (b)에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계 이전에: (i) 복수의 세포의 발현 프로파일의 거리 데이터 구조를 결정하는 단계를 포함한다. 거리 데이터 구조는 복수의 세포의 발현 프로파일의 거리 행렬을 포함할 수 있다. 거리 행렬의 각각의 대각선 요소는 0의 값을 갖는다. (b)에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계는: 거리 행렬에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 것을 포함할 수 있다. 복수의 세포의 발현 프로파일 사이의 거리는 복수의 세포의 발현 프로파일 사이의 쌍별 상관관계 거리일 수 있다.

[0027] 일부 실시형태에서, 상기 방법은 (i)에서 거리 데이터 구조를 결정하는 단계 이전에, 표적 계수 데이터 구조를 로그-변환된 표적 계수 데이터 구조로 로그-변환하는 단계를 포함하며, 표적 계수 데이터 구조의 요소의 거리 데이터 구조를 결정하는 단계는 로그-변환된 표적 계수 데이터 구조의 거리 데이터 구조를 결정하는 것을 포함하며, (b)에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계는 로그-변환된 표적 계수 데이터 구조 및 거리 데이터 구조에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 클러스터를 생성하는 것을 포함한다. 로그-변환된 표적 계수 데이터 구조로 표적 계수 데이터 구조를 로그-변환하는 단계는 표적 계수 데이터 구조의 각각의 요소의 값을 충분한만큼 증가시키는 것을 포함할 수 있다. 증분은 1일 수 있다.

- [0028] 일부 실시형태에서, (b)에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하는 단계는 복수의 세포의 각각의 발현 프로파일을 복수의 클러스터 내의 상이한 리프 클러스터에 할당하고; 제2 클러스터가 복수의 클러스터에서 제1 클러스터에 가장 가까운 클러스터이면, 복수의 클러스터 중 제1 클러스터 및 제2 클러스터를 반복적으로 조합하여, 제1 클러스터 및 제2 클러스터의 부모 클러스터를 생성하는 것을 포함한다. 제1 클러스터와 제2 클러스터 사이의 거리는 제1 클러스터에 의해 나타나는 임의의 발현 프로파일과, 제2 클러스터에 의해 나타나는 임의의 발현 프로파일 사이의 최대 거리일 수 있다.
- [0029] 일부 실시형태에서, 제1 클러스터의 클러스터-내 상관관계 및 제2 클러스터의 클러스터-내 상관관계는 제1 클러스터와 제2 클러스터의 클러스터-간 상관관계보다 더 크다. 제1 클러스터의 클러스터-내 상관관계 및 제2 클러스터의 클러스터-내 상관관계의 척도 또는 표시는 제1 클러스터 및 제2 클러스터의 클러스터-간 상관관계보다 더 크다. 제1 클러스터의 클러스터-내 상관관계 및 제2 클러스터의 클러스터-내 상관관계의 척도는 다음 중 적어도 하나에 기초할 수 있다: 제1 클러스터 및 제2 클러스터의 클러스터-내 최대 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-내 평균 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-내 중간값 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 제1 클러스터의 클러스터-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 클러스터의 클러스터-내 최대 상관관계, 제1 클러스터의 클러스터-내 평균 상관관계, 제1 클러스터의 클러스터-내 중간값 상관관계, 제1 클러스터의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 제2 클러스터의 클러스터-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제2 클러스터의 클러스터-내 최대 상관관계, 제2 클러스터의 클러스터-내 평균 상관관계, 제2 클러스터의 클러스터-내 중간값 상관관계, 제2 클러스터의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 제1 클러스터 및 제2 클러스터의 클러스터-간 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 클러스터 및 제2 클러스터의 클러스터-간 최대 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-간 평균 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-간 중간값 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-간 최소 상관관계 및 그의 임의의 조합.
- [0030] 일부 실시형태에서, 상기 방법은 덴드로그램의 복수의 클러스터를 트레이싱하는 경우 각 클러스터에서: 연관이 유효하면, 클러스터로부터 클러스터의 좌측 자식 클러스터 및 우측 자식 클러스터로의 트레이싱을 계속하고; 연관이 무효하면, 클러스터로부터 클러스터의 좌측 자식 클러스터 및 우측 자식 클러스터로의 트레이싱을 중단하는 단계를 포함한다. 클러스터와 클러스터의 자식 클러스터의 연관이 유효한지 무효한지를 결정하는 것은 좌측 자식 클러스터와 우측 자식 클러스터 사이의 거리가 연관 임계값 초과이면, 연관을 유효한 것으로 결정하고, 그렇지 않으면, 무효한 것으로 결정하는 것을 포함할 수 있다.
- [0031] 일부 실시형태에서, 좌측 자식 클러스터와 우측 자식 클러스터 사이의 거리는 좌측 자식 클러스터 및 우측 자식 클러스터에 의해 나타나는 발현 프로파일 사이에서 복수의 표적 중 각각의 표적에서 수행되는 통계적 검정에 기초하여 결정될 수 있다. 통계적 검정은 웰치의 t-검정을 포함할 수 있다. 좌측 자식 클러스터와 우측 자식 클러스터 사이의 거리는 좌측 자식 클러스터에 의해 나타나는 발현 프로파일과 우측 자식 클러스터에 의해 나타나는 각각의 발현 프로파일 사이에서 복수의 표적 중 각각의 표적에서 수행되는 통계적 검정의 최대 p-값에 기초하여 결정될 수 있다.
- [0032] 일부 실시형태에서, 클러스터와 클러스터의 자식 클러스터의 연관이 유효한지 무효한지를 결정하는 단계는: 제1 클러스터의 클러스터-내 상관관계 및 제2 클러스터의 클러스터-내 상관관계 중 적어도 하나가 제1 클러스터 및 제2 클러스터의 클러스터-간 상관관계보다 더 크면, 연관을 유효한 것으로 결정하고, 그렇지 않으면 무효한 것으로 결정하는 것을 포함할 수 있다. 일부 실시형태에서, 클러스터와 클러스터의 자식 클러스터의 연관이 유효한지 무효한지를 결정하는 단계는 제1 클러스터의 클러스터-내 상관관계 및 제2 클러스터의 클러스터-내 상관관계의 척도 또는 표시가 제1 클러스터 및 제2 클러스터의 클러스터-간 상관관계보다 더 크면, 연관을 유효한 것으로 결정하는 것을 포함할 수 있다. 제1 클러스터의 클러스터-내 상관관계 및 제2 클러스터의 클러스터-내 상관관계의 척도는 다음 중 적어도 하나에 기초할 수 있다: 제1 클러스터 및 제2 클러스터의 클러스터-내 최대 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-내 평균 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-내 중간값 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 제1 클러스터의 클러스터-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 클러스터의 클러스터-내 최대 상관관계, 제1 클러스터의 클러스터-내 평균 상관관계, 제1 클러스터의 클러스터-내 중간값 상관관계, 제1 클러스터의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 제2 클러스터의 클러스터-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제2 클러스터의 클러스터-내 최대 상관관계, 제2 클러스터의 클러스터-내 평균 상관관계, 제2 클러스터의 클러스터-내 중간값 상관관계, 제2 클러스터의 클러스터-내 최소 상관관계 및 그의 임의의 조합.

의 조합. 제1 클러스터 및 제2 클러스터의 클러스터-간 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 클러스터 및 제2 클러스터의 클러스터-간 최대 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-간 평균 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-간 중간값 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-간 최소 상관관계 및 그의 임의의 조합.

[0033] 일부 실시형태에서, 상기 방법은 덴드로그램의 복수의 클러스터를 트레이싱하는 경우 각 클러스터에서: (3) 클러스터가 단일 세포의 발현 프로파일을 나타내면, 클러스터를 병합 클러스터 세트에 추가하는 단계를 포함한다. 상기 방법은 덴드로그램의 복수의 클러스터를 트레이싱하는 경우, 각 클러스터에서: 클러스터 표지를 클러스터에 할당하는 단계를 포함할 수 있다. 일부 실시형태에서, 클러스터가 단일 세포의 발현 프로파일을 나타낸다면, 클러스터의 클러스터 표지는 단일 세포 지정을 포함하며, 그렇지 않고 클러스터가 부모 클러스터의 좌측 자식 클러스터이면, 클러스터의 클러스터 표지는 부모 클러스터의 클러스터 표지 및 좌측 지정을 포함하고, 그렇지 않으면, 클러스터의 클러스터 표지는 부모 클러스터의 클러스터 표지 및 우측 지정을 포함한다.

[0034] 일부 실시형태에서, 병합 클러스터 세트 내의 각각의 클러스터에 있어서, 클러스터에 의해 나타나는 세포의 복수의 표지의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계는 클러스터에 의해 나타난 발현 프로파일과 병합 클러스터 세트 내의 또 다른 클러스터에 의해 나타난 발현 프로파일 사이의, 세포 유형을 구별하기 위한 표적과 연관된 별개의 서열을 갖는 분자 표지의 수의 차이가 유의성 임계값보다 더 큰 것을 결정하는 것을 포함한다.

[0035] 일부 실시형태에서, 상기 방법은 (d)에서 제1 클러스터와 제2 클러스터를 병합하여 병합된 클러스터를 생성하는 단계 이전에: 제3 클러스터와 제4 클러스터 사이의 거리가 클러스터 거리 임계값 이내이면, 단일 세포의 발현 프로파일을 나타내는 병합 클러스터 세트 내의 각각의 제3 클러스터를 병합 클러스터 세트 내의 제4 클러스터와 병합하는 단계를 포함한다. 상기 방법은 세포의 발현 프로파일을 나타내는 병합 클러스터 세트 내의 클러스터에 기초하여 복수의 세포를 분류하는 단계를 포함할 수 있다. 상기 방법은 확인된 세포 유형을 구별하기 위한 표적에 기초하여, 전체 전사체 검정을 설계하는 단계 또는 확인된 세포 유형을 구별하기 위한 표적에 기초하여, 표적화된 전사체 검정을 설계하는 단계를 포함할 수 있다.

[0036] 세포 유형을 구별하기 위한 표적의 확인을 위한 시스템이 본원에 개시된다. 일부 실시형태에서, 시스템은 하드웨어 프로세서; 및 명령어가 저장된 비-일시적 메모리로서, 하드웨어 프로세서에 의해 실행되는 경우 프로세서가 본원에 개시된 임의의 방법을 수행하게 하는 비-일시적 메모리를 포함한다. 세포 유형을 구별하기 위한 표적의 확인을 위한 컴퓨터 판독 가능한 매체가 본원에 개시된다. 일부 실시형태에서, 컴퓨터 판독 가능한 매체는 본원에 개시된 방법 중 임의의 것을 수행하기 위한 코드를 포함한다.

[0037] 세포 유형을 구별하기 위한 표적의 확인을 위한 시스템의 실시형태가 본원에 개시된다. 일부 실시형태에서, 시스템은 실행 가능한 명령어를 저장하도록 구성된 비-일시적 메모리, 및 비-일시적 메모리와 통신하는 하드웨어 프로세서를 포함하며, 이러한 하드웨어 프로세서는, (a) 표적 계수 데이터 구조를 수신하고, 표적 계수 데이터 구조가 복수의 세포의 발현 프로파일을 포함하며, 복수의 세포의 발현 프로파일이 복수의 세포 중 각각의 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함하며; (b) 표적 계수 데이터 구조 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 복수의 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하고, 덴드로그램이 복수의 노드를 포함하며, 복수의 노드가 루트 노드, 복수의 리프 노드 및 복수의 비-루트, 비-리프 노드를 포함하며, 복수의 리프 노드 중 각각의 리프 노드가 복수의 세포 중 상이한 세포의 발현 프로파일을 나타내며, 루트 노드가 복수의 세포의 발현 프로파일을 나타내며; (c) 덴드로그램의 루트 노드로부터 덴드로그램의 복수의 리프 노드로 덴드로그램의 복수의 노드 중 각각의 노드를 통해 트레이싱하는 동안: (1) 노드의 자식 노드로의 노드의 분할이 유효한지 무효한지를 결정하고; (2) 노드의 자식 노드로의 노드의 분할이 무효하다면, 노드를 병합 클러스터 세트에 추가하며; (d) 반복적으로, 병합 클러스터 세트 내의 각각의 제1 노드에 있어서, 병합 클러스터 세트 내의 제1 노드와 제1 노드에 가장 가까운 병합 클러스터 세트 내의 제2 노드 사이의 거리가 병합 거리 임계값 이내이면, 제1 노드와 제2 노드를 병합하여, 제1 노드 및 제2 노드에 의해 나타나는 발현 프로파일을 포함하는 병합된 노드를 생성하며; (e) 병합 클러스터 세트 내의 각각의 노드에 있어서, 노드에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하도록, 실행 가능한 명령어에 의해 프로그래밍된다.

[0038] 일부 실시형태에서, 표적 계수 데이터 구조는 표적 계수 행렬을 포함한다. 표적 계수 행렬의 각각의 행 또는 각각의 열은 복수의 세포 중 상이한 개별 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함할 수 있다. 복수의 리프 노드 및 복수의 비-루트, 비-리프 노드의 각각은 부모 노드와 연관될 수 있으며, 루트 노드 및 복수의

비-루트, 비-리프 노드의 각각은 좌측 자식 노드 및 우측 자식 노드와 연관될 수 있으며, 노드의 좌측 자식 노드 및 우측 자식 노드에 의해 나타나는 발현 프로파일을 나타낸다.

[0039] 일부 실시형태에서, 하드웨어 프로세서는 (a)에서 표적 계수 데이터 구조를 수신하기 이전에: (f) 복수의 바코드를 사용하여 복수의 세포 내의 복수의 표적이 바코딩되게 하여, 복수의 바코딩된 표적을 생성하고, 복수의 바코드의 각각이 세포 표지 및 분자 표지를 포함하며, 상이한 세포의 표적으로부터 생성된 바코딩된 표적이 상이한 세포 표지를 가지며, 복수의 세포 중 하나의 세포의 표적으로부터 생성된 바코딩된 표적이 상이한 분자 표지를 가지며; (g) 복수의 바코딩된 표적의 시퀀싱 데이터를 취득하며; (h) 복수의 세포의 각각에 있어서: (1) 세포에 대한 시퀀싱 데이터에서 복수의 표적 중 각각의 표적과 연관된 별개의 서열을 갖는 분자 표지의 수를 계수하며; (2) (h)(1)에서 계수된 시퀀싱 데이터에서의 표적과 연관된 별개의 서열을 갖는 분자 표지의 수에 기초하여 세포에 대한 복수의 표적 중 각각의 표적의 수를 추정하도록 프로그래밍될 수 있다. 표적 계수 데이터 구조를 수신하기 위하여, 하드웨어 프로세서는 (h)(2)에서 추정된 세포에 대한 복수의 표적 중 각각의 표적의 수로부터 표적 계수 데이터 구조를 생성하며, 복수의 세포 중 세포의 발현 프로파일이 (h)(2)에서 추정된 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함하도록 프로그래밍될 수 있다.

[0040] 일부 실시형태에서, 하드웨어 프로세서는 (b)에서 표적 계수 데이터 구조 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 복수의 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하기 이전에: (i) 표적 계수 데이터 구조의 요소의 거리 데이터 구조를 결정하고, 거리 데이터 구조가 복수의 세포의 발현 프로파일 사이의 거리를 포함하도록 프로그래밍될 수 있다. 거리 데이터 구조는 거리 행렬을 포함한다. 거리 행렬의 각각의 대각선 요소는 0의 값을 갖는다.

[0041] 일부 실시형태에서, (b)에서 표적 계수 데이터 구조 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 복수의 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하기 위하여, 하드웨어 프로세서는 표적 계수 데이터 구조 및 거리 데이터 구조에 기초하여 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하도록 프로그래밍될 수 있다. 복수의 세포의 발현 프로파일 사이의 거리는 복수의 세포의 발현 프로파일 사이의 쌍별 상관관계 거리를 포함할 수 있다.

[0042] 일부 실시형태에서, 하드웨어 프로세서는 (i)에서 표적 계수 데이터 구조의 요소의 거리 데이터 구조를 결정하기 이전에, 표적 계수 데이터 구조를 로그-변환된 표적 계수 데이터 구조로 로그-변환하도록 프로그래밍될 수 있다. 표적 계수 데이터 구조의 요소의 거리 데이터 구조를 결정하기 위하여, 하드웨어 프로세서는 로그-변환된 표적 계수 데이터 구조의 거리 데이터 구조를 결정하도록 프로그래밍될 수 있다. (b)에서 표적 계수 데이터 구조 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하기 위하여, 하드웨어 프로세서는 로그-변환된 표적 계수 데이터 구조 및 거리 데이터 구조에 기초하여 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 덴드로그램을 생성하도록 프로그래밍될 수 있다. 로그-변환된 표적 계수 데이터 구조로 표적 계수 데이터 구조를 로그-변환하기 위하여, 하드웨어 프로세서는 표적 계수 데이터 구조의 각 요소의 값을 증분만큼 증가시키도록 프로그래밍될 수 있다. 증분은 1일 수 있다.

[0043] (b)에서 표적 계수 데이터 구조 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하기 위하여, 하드웨어 프로세서는 복수의 세포의 각각의 발현 프로파일을 상이한 리프 노드에 할당하고; 제2 노드가 복수의 노드 내에서 제1 노드에 가장 가까운 노드이면, 복수의 노드 중 제1 노드 및 제2 노드를 반복적으로 조합하여, 제1 노드 및 제2 노드의 부모 노드를 생성하도록 프로그래밍될 수 있다. 제1 노드와 제2 노드 사이의 거리는 제1 노드에 의해 나타나는 발현 프로파일을 갖는 임의의 세포와 제2 노드에 의해 나타나는 발현 프로파일을 갖는 임의의 세포 사이의 최대 거리일 수 있다.

[0044] 일부 실시형태에서, 제1 노드의 노드-내 상관관계 및 제2 노드의 노드-내 상관관계 중 적어도 하나는 제1 노드 및 제2 노드의 노드-간 상관관계보다 더 클 수 있다. 제1 노드의 노드-내 상관관계 및 제2 노드의 노드-내 상관관계의 척도 또는 표시는 제1 노드 및 제2 노드의 노드-간 상관관계보다 더 클 수 있다. 제1 노드의 노드-내 상관관계 및 제2 노드의 노드-내 상관관계의 척도는 다음 중 적어도 하나에 기초할 수 있다: 제1 노드 및 제2 노드의 노드-내 최대 상관관계, 제1 노드 및 제2 노드의 노드-내 평균 상관관계, 제1 노드 및 제2 노드의 노드-내 중간값 상관관계, 제1 노드 및 제2 노드의 노드-내 최소 상관관계 및 그의 임의의 조합. 제1 노드의 노드-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 노드의 노드-내 최대 상관관계, 제1 노드의 노드-내 평균 상관관계, 제1 노드의 노드-내 중간값 상관관계, 제1 노드의 노드-내 최소 상관관계 및 그의 임의의 조합. 제2 노드의 노드-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제2 노드의 노드-내 최대 상관관계, 제2 노드의 노드-내 평균 상관관계, 제2 노드의 노드-내 중간값 상관관계, 제2 노드의 노드-내 최소 상관관계 및 그의

임의의 조합. 제1 노드 및 제2 노드의 노드-간 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 노드 및 제2 노드의 노드-간 최대 상관관계, 제1 노드 및 제2 노드의 노드-간 평균 상관관계, 제1 노드 및 제2 노드의 노드-간 중간값 상관관계, 제1 노드 및 제2 노드의 노드-간 최소 상관관계 및 그의 임의의 조합.

[0045] 일부 실시형태에서, 하드웨어 프로세서는 덴드로그램의 복수의 노드를 트래버싱하는 경우 각 노드에서: 분할이 유효하면, 노드로부터 노드의 좌측 자식 노드 및 우측 자식 노드로의 트래버싱을 계속하고; 분할이 무효하면, 노드로부터 노드의 좌측 자식 노드 및 우측 자식 노드로의 트래버싱을 중단하도록 프로그래밍될 수 있다. 노드와 노드의 자식 노드의 분할이 유효한지 무효한지를 결정하기 위하여, 하드웨어 프로세서는 좌측 자식 노드와 우측 자식 노드 사이의 거리가 분할 임계값 초과이면, 분할을 유효한 것으로 결정하고, 그렇지 않으면, 무효한 것으로 결정하도록 프로그래밍될 수 있다. 좌측 자식 노드와 우측 자식 노드 사이의 거리는 좌측 자식 노드 및 우측 자식 노드에 의해 나타나는 발현 프로파일 사이의 복수의 표적 중 각각의 표적에서 수행되는 통계적 검정에 기초하여 결정될 수 있다. 통계적 검정은 웰치의 t-검정을 포함할 수 있다. 좌측 자식 노드와 우측 자식 노드 사이의 거리는 좌측 자식 노드에 의해 나타나는 각각의 발현 프로파일과 우측 자식 노드에 의해 나타나는 각각의 발현 프로파일 사이의 복수의 표적 중 각각의 표적에서 수행되는 통계적 검정의 최대 p-값에 기초하여 결정될 수 있다.

[0046] 일부 실시형태에서, 하드웨어 프로세서는 덴드로그램의 복수의 노드를 트래버싱하는 경우 각 노드에서: (3) 노드가 단일 세포의 발현 프로파일을 나타낸다면, 노드를 병합 클러스터 세트에 추가하도록 프로그래밍될 수 있다. 일부 실시형태에서, 덴드로그램의 복수의 노드를 트래버싱하는 경우, 각 노드에서, 하드웨어 프로세서는 노드 표지를 노드에 할당하도록 프로그래밍될 수 있다. 노드가 단일 세포의 발현 프로파일을 나타낸다면, 노드의 노드 표지는 단일 세포 지정을 포함할 수 있으며, 그렇지 않고 노드가 부모 노드의 좌측 자식 노드이면, 노드의 노드 표지는 부모 노드의 노드 표지 및 좌측 지정을 포함할 수 있으며, 그렇지 않으면, 노드의 노드 표지는 부모 노드의 노드 표지 및 우측 지정을 포함할 수 있다.

[0047] 일부 실시형태에서, 병합 클러스터 세트 내의 각각의 노드에 있어서, 노드에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하기 위하여, 하드웨어 프로세서는 노드에 의해 나타나는 발현 프로파일과 병합 클러스터 세트 내의 또 다른 노드에 의해 나타나는 발현 프로파일 사이의, 세포 유형을 구별하기 위한 표적과 연관된 별개의 서열을 갖는 분자 표지의 수의 차이가 유의성 임계값보다 더 큰 것을 결정하도록 프로그래밍될 수 있다.

[0048] 일부 실시형태에서, 하드웨어 프로세서는 (d)에서 제1 노드와 제2 노드를 병합하여 병합된 노드를 생성하기 이전에: 제3 노드와 제4 노드 사이의 거리가 노드 거리 임계값 이내이면, 단일 세포의 발현 프로파일을 나타내는 병합 클러스터 세트 내의 각각의 제3 노드를 병합 클러스터 세트 내의 제4 노드와 병합하도록 프로그래밍될 수 있다. 하드웨어 프로세서는 세포의 발현 프로파일을 나타내는 병합 클러스터 세트 내의 노드에 기초하여 복수의 세포를 분류하도록 프로그래밍될 수 있다. 하드웨어 프로세서는 확인된 세포 유형을 구별하기 위한 표적에 기초하여, 전체 전사체 검정을 설계하도록 프로그래밍될 수 있다. 하드웨어 프로세서는 확인된 세포 유형을 구별하기 위한 표적에 기초하여, 표적화된 전사체 검정을 설계하도록 프로그래밍될 수 있다.

[0049] 세포 유형을 구별하기 위한 표적의 확인을 위한 시스템의 실시형태가 본원에 개시된다. 일부 실시형태에서, 시스템은 실행 가능한 명령어를 저장하도록 구성된 비-일시적 메모리, 및 비-일시적 메모리와 통신하는 하드웨어 프로세서를 포함하며, 하드웨어 프로세서는 (a) 복수의 세포의 발현 프로파일을 수신하며, 발현 프로파일이 복수의 세포 중 각각의 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함하며; (b) 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하며, 각각의 클러스터가 (1) 부모 클러스터 및 (2) 2개 이상의 자식 클러스터 중 하나 또는 둘 모두와 하나 이상의 연관을 가지며, 부모 클러스터가 클러스터에 의해 나타나는 복수의 세포 중 하나 이상의 세포의 발현 프로파일을 나타내며, 클러스터가 2개 이상의 자식 클러스터에 의해 나타나는 발현 프로파일을 나타내며; (c) 2개 이상의 자식 클러스터를 갖는 각각의 클러스터에 있어서, 클러스터와 2개 이상의 자식 클러스터 사이의 연관이 무효하면, 클러스터를 병합 클러스터 세트에 추가하며; (d) 반복적으로, 병합 클러스터 세트 내의 각각의 제1 클러스터에 있어서, 병합 클러스터 세트 내의 제1 클러스터와 제1 클러스터에 가장 가까운 병합 클러스터 세트 내의 제2 클러스터 사이의 거리가 병합 거리 임계값 이내이면, 제1 클러스터와 제2 클러스터를 병합하여, 병합된 클러스터를 생성하며, 병합된 클러스터가 제1 클러스터 및 제2 클러스터의 발현 프로파일을 포함하며; (e) 병합 클러스터 세트 내의 각각의 클러스터에 있어서, 클러스터에 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하도록 실행 가능한 명령어에 의해

프로그래밍된다.

- [0050] 일부 실시형태에서, 하드웨어 프로세서는 표적 계수 데이터 구조를 수신하는 것을 포함하여 복수의 세포의 발현 프로파일을 수신하도록 프로그래밍될 수 있다. 표적 계수 데이터 구조는 표적 계수 행렬을 포함할 수 있다. 표적 계수 행렬의 각각의 행 또는 각각의 열은 복수의 세포 중 상이한 개별 세포의 발현 프로파일을 포함할 수 있다.
- [0051] 일부 실시형태에서, 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 복수의 발현 프로파일의 클러스터로 클러스터링하기 위하여, 하드웨어 프로세서는 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여 복수의 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하도록 프로그래밍될 수 있으며, 덴드로그램은 복수의 클러스터를 포함하며, 복수의 클러스터는 루트 클러스터, 복수의 리프 클러스터 및 복수의 비-루트, 비-리프 클러스터를 포함한다. 복수의 리프 클러스터, 및 복수의 비-루트, 비-리프 클러스터의 각각은 부모 클러스터와 연관될 수 있다. 루트 클러스터 및 복수의 비-루트, 비-리프 클러스터의 각각은 좌측 자식 클러스터 및 우측 자식 클러스터와 연관될 수 있으며, 클러스터의 좌측 자식 클러스터 및 우측 자식 클러스터에 의해 나타나는 발현 프로파일을 나타낸다. 루트 클러스터는 복수의 세포의 발현 프로파일을 나타낼 수 있다. 2개 이상의 자식 클러스터를 갖는 각각의 클러스터에 있어서, 클러스터와 2개 이상의 자식 클러스터 사이의 연관이 무효하면, 클러스터를 병합 클러스터 세트에 추가하기 위하여 하드웨어 프로세서는 덴드로그램의 루트 클러스터로부터 덴드로그램의 복수의 리프 클러스터로 덴드로그램의 각각의 클러스터를 통해 트래버싱하는 동안 (1) 클러스터와 클러스터의 자식 클러스터 사이의 연관이 유효한지 무효한지를 결정하고; (2) 연관이 무효하면, 클러스터를 병합 클러스터 세트에 추가하도록 프로그래밍될 수 있다.
- [0052] 일부 실시형태에서, 하드웨어 프로세서는 (a)에서 복수의 세포의 발현 프로파일을 수신하기 이전에: (f) 복수의 바코드를 사용하여 복수의 세포 내의 복수의 표적을 바코딩하여, 복수의 바코딩된 표적을 생성하며, 복수의 바코드의 각각이 세포 표지 및 분자 표지를 포함하며, 상이한 세포의 표적으로부터 생성된 바코딩된 표적이 상이한 세포 표지를 가지며, 복수의 세포 중 하나의 세포의 표적으로부터 생성된 바코딩된 표적이 상이한 분자 표지를 가지며; (g) 복수의 바코딩된 표적의 시퀀싱 데이터를 수득하며; (h) 복수의 세포의 각각에 있어서: (1) 세포에 대한 시퀀싱 데이터에서 복수의 표적 중 각각의 표적과 연관된 별개의 서열을 갖는 분자 표지의 수를 계수하며; (2) (h)(1)에서 계수된 시퀀싱 데이터에서의 표적과 연관된 별개의 서열을 갖는 분자 표지의 수에 기초하여 세포에 대한 복수의 표적 중 각각의 표적의 수를 추정하도록 프로그래밍될 수 있다. 복수의 세포 중 세포의 발현 프로파일은 (h)(2)에서 추정된 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함할 수 있다.
- [0053] 일부 실시형태에서, 하드웨어 프로세서는 (b)에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하기 이전에: (i) 복수의 세포의 발현 프로파일의 거리 데이터 구조를 결정하도록 프로그래밍될 수 있다. 거리 데이터 구조는 복수의 세포의 발현 프로파일의 거리 행렬을 포함할 수 있다. 거리 행렬의 각각의 대각선 요소는 0의 값을 가질 수 있다. (b)에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하기 위하여, 하드웨어 프로세서는 거리 행렬에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하도록 프로그래밍될 수 있다. 복수의 세포의 발현 프로파일 사이의 거리는 복수의 세포의 발현 프로파일 사이의 쌍별 상관관계 거리일 수 있다.
- [0054] 일부 실시형태에서, 하드웨어 프로세서는 (i)에서 거리 데이터 구조를 결정하기 이전에, 표적 계수 데이터 구조를 로그-변환된 표적 계수 데이터 구조로 로그-변환하도록 프로그래밍될 수 있다. 표적 계수 데이터 구조의 요소의 거리 데이터 구조를 결정하기 위하여, 하드웨어 프로세서는 로그-변환된 표적 계수 데이터 구조의 거리 데이터 구조를 결정하도록 프로그래밍될 수 있다. (b)에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하기 위하여, 하드웨어 프로세서는 로그-변환된 표적 계수 데이터 구조 및 거리 데이터 구조에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 클러스터를 생성하도록 프로그래밍될 수 있다. 표적 계수 데이터 구조를 로그-변환된 표적 계수 데이터 구조로 로그-변환하기 위하여, 하드웨어 프로세서는 표적 계수 데이터 구조의 각 요소의 값을 충분한만큼 증가시키도록 프로그래밍될 수 있다. 증분은 1일 수 있다.
- [0055] 일부 실시형태에서, (b)에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하기 위하여, 하드웨어 프로세서는 복수의 세포의 각각의 발현 프로파일을 복수의 클러스터 내

의 상이한 리프 클러스터에 할당하고; 제2 클러스터가 복수의 클러스터 중 제1 클러스터에 가장 가까운 클러스터이면, 복수의 클러스터 중 제1 클러스터 및 제2 클러스터를 반복적으로 조합하여, 제1 클러스터 및 제2 클러스터의 부모 클러스터를 생성하도록 프로그래밍될 수 있다. 1 클러스터와 제2 클러스터 사이의 거리는 제1 클러스터에 의해 나타난 임의의 발현 프로파일과, 제2 클러스터에 의해 나타난 임의의 발현 프로파일 사이의 최대 거리일 수 있다.

[0056] 일부 실시형태에서, 제1 클러스터의 클러스터-내 상관관계 및 제2 클러스터의 클러스터-내 상관관계는 제1 클러스터와 제2 클러스터의 클러스터-간 상관관계보다 더 클 수 있다. 제1 클러스터의 클러스터-내 상관관계 및 제2 클러스터의 클러스터-내 상관관계의 척도 또는 표시는 제1 클러스터 및 제2 클러스터의 클러스터-간 상관관계보다 더 크다. 제1 클러스터의 클러스터-내 상관관계 및 제2 클러스터의 클러스터-내 상관관계의 척도는 다음 중 적어도 하나에 기초할 수 있다: 제1 클러스터 및 제2 클러스터의 클러스터-내 최대 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-내 평균 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-내 중간값 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 제1 클러스터의 클러스터-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 클러스터의 클러스터-내 최대 상관관계, 제1 클러스터의 클러스터-내 평균 상관관계, 제1 클러스터의 클러스터-내 중간값 상관관계, 제1 클러스터의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 제2 클러스터의 클러스터-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제2 클러스터의 클러스터-내 최대 상관관계, 제2 클러스터의 클러스터-내 평균 상관관계, 제2 클러스터의 클러스터-내 중간값 상관관계, 제2 클러스터의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 제1 클러스터 및 제2 클러스터의 클러스터-간 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 클러스터 및 제2 클러스터의 클러스터-간 최대 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-간 평균 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-간 중간값 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-간 최소 상관관계 및 그의 임의의 조합.

[0057] 일부 실시형태에서, 하드웨어 프로세서는 덴드로그램의 복수의 클러스터를 트레이싱하는 경우 각 클러스터에서: 연관이 유효하면, 클러스터로부터 클러스터의 좌측 자식 클러스터 및 우측 자식 클러스터로의 트레이싱을 계속하고; 연관이 무효하면, 클러스터로부터 클러스터의 좌측 자식 클러스터 및 우측 자식 클러스터로의 트레이싱을 중단하도록 프로그래밍될 수 있다. 클러스터와 클러스터의 자식 클러스터의 연관이 유효한지 무효한지를 결정하기 위하여, 하드웨어 프로세서는: 좌측 자식 클러스터와 우측 자식 클러스터 사이의 거리가 연관 임계값 초과이면, 연관을 유효한 것으로 결정하고, 그렇지 않으면, 무효한 것으로 결정하도록 프로그래밍될 수 있다. 좌측 자식 클러스터와 우측 자식 클러스터 사이의 거리는 좌측 자식 클러스터 및 우측 자식 클러스터에 의해 나타난 발현 프로파일 사이의 복수의 표적 중 각각의 표적에서 수행되는 통계적 검정에 기초하여 결정될 수 있다. 통계적 검정은 웰치의 t-검정을 포함할 수 있다. 좌측 자식 클러스터와 우측 자식 클러스터 사이의 거리는 좌측 자식 클러스터에 의해 나타난 발현 프로파일과 우측 자식 클러스터에 의해 나타난 각각의 발현 프로파일 사이의 복수의 표적 중 각각의 표적에서 수행되는 통계적 검정의 최대 p-값에 기초하여 결정될 수 있다.

[0058] 일부 실시형태에서, 하드웨어 프로세서는 덴드로그램의 복수의 클러스터를 트레이싱하는 경우 각 클러스터에서: (3) 클러스터가 단일 세포의 발현 프로파일을 나타내면, 클러스터를 병합 클러스터 세트에 추가하도록 프로그래밍될 수 있다. 하드웨어 프로세서는 덴드로그램의 복수의 클러스터를 트레이싱하는 경우, 각 클러스터에서: 클러스터 표지를 클러스터에 할당하도록 프로그래밍될 수 있다. 클러스터가 단일 세포의 발현 프로파일을 나타내면, 클러스터의 클러스터 표지는 단일 세포 지정을 포함하며, 그렇지 않고 클러스터가 부모 클러스터의 좌측 자식 클러스터이면, 클러스터의 클러스터 표지는 부모 클러스터의 클러스터 표지 및 좌측 지정을 포함하고, 그렇지 않으면, 클러스터의 클러스터 표지는 부모 클러스터의 클러스터 표지 및 우측 지정을 포함한다.

[0059] 일부 실시형태에서, 병합 클러스터 세트 내의 각각의 클러스터에 있어서, 클러스터에 의해 나타난 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하기 위하여, 하드웨어 프로세서는 클러스터에 의해 나타난 발현 프로파일과 병합 클러스터 세트 내의 또 다른 클러스터에 의해 나타난 발현 프로파일 사이의, 세포 유형을 구별하기 위한 표적과 연관된 별개의 서열을 갖는 분자 표지의 수의 차이가 유의성 임계값보다 더 큰 것을 결정하도록 프로그래밍될 수 있다. 하드웨어 프로세서는 (d)에서 제1 클러스터와 제2 클러스터를 병합하여 병합된 클러스터를 생성하기 이전에: 제3 클러스터와 제4 클러스터 사이의 거리가 클러스터 거리 임계값 이내이면, 단일 세포의 발현 프로파일을 나타내는 병합 클러스터 세트 내의 각각의 제3 클러스터를 병합 클러스터 세트 내의 제4 클러스터와 병합하도록 프로그래밍될 수 있다.

[0060] 일부 실시형태에서, 하드웨어 프로세서는 세포의 발현 프로파일을 나타내는 병합 클러스터 세트 내의 클러스터에 기초하여 복수의 세포를 분류하도록 프로그래밍될 수 있다. 하드웨어 프로세서는 확인된 세포 유형을 구별하

기 위한 표적에 기초하여, 전체 전사체 검정을 설계하도록 프로그래밍될 수 있다. 하드웨어 프로세서는 확인된 세포 유형을 구별하기 위한 표적에 기초하여, 표적화된 전사체 검정을 설계하도록 프로그래밍될 수 있다.

도면의 간단한 설명

[0061]

도 1은 비제한적인 예시적인 바코드(예를 들어, 확률적 바코드)를 예시한 것이다.

도 2는 바코딩 및 디지털 계수(예를 들어, 확률적 바코딩 및 디지털 계수)의 **비제한적인** 예시적인 작업흐름을 보여준다.

도 3은 복수의 표적으로부터 바코딩된 표적(예를 들어, 확률적으로 바코딩된 표적)의 인덱싱된 라이브러리를 생성하기 위한 비제한적인 예시적인 과정을 보여주는 개략도이다.

도 4는 덴드로그램을 사용하여 세포의 발현 프로파일을 클러스터링함으로써 세포 유형을 구별하기 위한 표적을 확인하는 비제한적인 예시적인 방법을 보여주는 흐름도이다.

도 5는 예시적인 덴드로그램의 개략도이다.

도 6은 세포의 발현 프로파일을 클러스터링함으로써 세포 유형을 구별하기 위한 표적을 확인하는 비제한적인 예시적인 방법을 보여주는 흐름도이다.

도 7은 본 발명의 방법을 구현하도록 구성된 예시적인 컴퓨팅 시스템의 블록도이다.

도 8의 패널 (a) 내지 (d)는 단일 세포의 발현 프로파일의 분할 및 병합 이후의 2차원 공간에서의 발현 프로파일의 비제한적인 예시적인 플롯을 보여준다.

도 9a 내지 도 9h의 패널 (a) 내지 (x)는 분할이 어떻게 결정될 수 있는지를 보여주는 2차원 공간에서의 발현 프로파일의 비제한적인 예시적인 플롯이다.

도 10은 제5 사이클의 분할 이후 2차원 공간에서의 발현 프로파일의 비제한적인 예시적인 플롯을 보여준다.

도 11의 패널 (a) 내지 (l)은 덴드로그램에서 특정 분할이 도 10에 나타난 제5 사이클의 분할 동안 유지되는 이유를 보여주는 2차원 공간에서의 발현 프로파일의 비제한적인 예시적인 플롯이다.

도 12a 내지 도 12c의 패널 (a) 내지 (i)는 병합이 어떻게 결정될 수 있는지를 보여주는 2차원 공간에서의 발현 프로파일의 비제한적인 예시적인 플롯이다.

도 13은 제2 사이클의 병합 이후 2차원 공간에서의 발현 프로파일의 비제한적인 예시적인 플롯을 보여준다.

도 14의 패널 (a) 내지 (d)는 도 13에 나타난 제2 사이클의 병합이 어떻게 결정되었는지를 보여주는 2차원 공간에서의 발현 프로파일의 비제한적인 예시적인 플롯이다.

도 15a 및 도 15b의 패널 (a) 내지 (f)는 비제한적인 예시적인 하나의 유형의 차등적 발현 분석을 보여주는 플롯이다.

도 16a 내지 도 16d의 패널 (a) 내지 (o)는 비제한적인 예시적인 또 다른 유형의 차등적 발현 분석을 보여주는 플롯이다.

도 17의 패널 (a) 내지 (g)는 클러스터 사이의 거리를 시각화시키는 비제한적인 예시적인 플롯이다.

도 18a 내지 도 18f의 패널 (a) 내지 (e)는 비제한적인 예시적인 덴드로그램을 보여준다.

도 19a 내지 도 19j의 패널 (a) 내지 (s)는 파라미터 스위핑(sweeping)을 보여주는 비제한적인 예시적인 플롯이다.

도 20은 임계값을 확인하기 위해 파라미터 스위핑이 어떻게 사용될 수 있는 지를 보여주는 비제한적인 예시적인 플롯이다.

도 21a 내지 도 21e의 패널 (a) 내지 (j)는 제1 분할의 결과를 보여주는 비제한적인 예시적인 플롯이다.

도 22는 2차원 공간에서 발현 프로파일의 분할 결과를 보여주는 비제한적인 예시적인 플롯이다.

도 23a 내지 도 23f의 패널 (a) 내지 (d)는 2개의 클러스터로 분류된 발현 프로파일을 보여주는 비제한적인 예시적인 덴드로그램을 보여준다.

도 24는 파라미터 스위핑을 보여주는 또 다른 비제한적인 예시적인 플롯이다.

발명을 실시하기 위한 구체적인 내용

- [0062] 하기의 상세한 설명에서, 본 명세서의 일부를 형성하는 첨부된 도면을 참조한다. 도면에서, 문맥상 달리 지시하지 않는 한, 유사한 기호는 전형적으로 유사한 구성요소를 나타낸다. 상세한 설명, 도면, 및 청구범위에 기재된 예시적인 실시형태는 한정하고자 하는 것이 아니다. 본 명세서에 제시된 주제의 사상 또는 범주를 벗어나지 않고 다른 실시형태가 이용될 수 있으며 다른 변화가 이루어질 수 있다. 본 명세서에 일반적으로 기재되고 도면에 도시된 바와 같은 본 발명의 양태가 매우 다양한 여러 가지 구성으로 배열되고, 치환되고, 조합되고, 분리되고, 설계될 수 있고, 이들 모두는 명백하게 본 명세서에서 고려되며, 본 명세서의 일부를 이룬다는 것이 용이하게 이해될 것이다.
- [0063] 모든 특허, 공개된 특허 출원, 다른 간행물 및 진뱅크(GenBank)로부터의 서열 및 본원에 언급된 다른 데이터베이스는 관련 기술에 관하여 그들 전체가 참조로 포함된다.
- [0064] 소량의 핵산 또는 표적, 예를 들어, 메신저 리보핵산(mRNA) 분자의 정량화는 예를 들어, 상이한 발생 단계에 또는 상이한 환경 조건 하에 세포 내에서 발현되는 유전자를 결정하는데 임상적으로 중요하다. 그러나, 특히 분자의 수가 매우 적은 경우, 핵산 분자(예를 들어, mRNA 분자)의 절대 개수를 결정하는 것은 매우 어려울 수 있다. 샘플 내의 분자의 절대 개수를 결정하기 위한 하나의 방법은 디지털 중합효소 연쇄 반응(PCR)이다. 독특한 분자 표지(ML, 분자 인덱스(MI)로도 지칭됨)를 갖는 바코드(예를 들어, 확률적 바코드)를 사용하여 분자의 수를 계수할 수 있다. 각 세포 표지에 대하여 독특한 분자 표지를 갖는 바코드를 사용하여 각 세포 내의 분자의 수를 계수할 수 있다. 비제한적인 예시적인 바코딩(예를 들어, 확률적 바코딩) 검정은 프리사이스(Precise)TM 검정(셀룰러 리서치, 인크.(Cellular Research, Inc.)(미국 캘리포니아주 팔로 알토 소재)), 리졸브(Resolve)TM 검정(셀룰러 리서치, 인크.(미국 캘리포니아주 팔로 알토 소재)) 또는 랩소디(Rhapsody)TM 검정(셀룰러 리서치, 인크.(미국 캘리포니아주 팔로 알토 소재))을 포함한다.
- [0065] 랩소디TM 검정은 RT 단계 동안 샘플 내의 모든 폴리(A)-mRNA에 혼성화하기 위하여 다수의, 예를 들어, 6561 내지 65536개의 폴리(T) 올리고뉴클레오타이드 상의 독특한 분자 표지가 있는 바코드(예를 들어, 확률적 바코드)의 비-소모성 풀을 사용할 수 있다. 분자 표지에 더하여, 바코드의 세포 표지를 사용하여 마이크로웰 플레이트의 각 웰 내의 각각의 단일 세포를 확인할 수 있다. 바코드(예를 들어, 확률적 바코드)는 범용 PCR 프라이밍 부위를 포함할 수 있다. RT 동안 표적 유전자 분자는 무작위로 바코드와 반응한다. 각각의 표적 분자는 바코드에 혼성화하여, 바코딩된 상보적 리보핵산(cDNA) 분자(예를 들어, 확률적으로 바코딩된 cDNA 분자)의 생성을 초래할 수 있다. 표지 후에, 마이크로웰 플레이트의 마이크로웰로부터 바코딩된 cDNA 분자를 PCR 증폭 및 시퀀싱을 위하여 단일의 튜브 내로 풀링할 수 있다. 미가공 시퀀싱 데이터를 분석하여, 독특한 분자 표지를 갖는 바코드(예를 들어, 확률적 바코드)의 수를 생성할 수 있다.
- [0066] 세포 유형을 구별하기 위한 표적의 확인 방법이 본원에 개시된다. 일부 실시형태에서, 상기 방법은 (a) 표적 계수 데이터 구조를 수신하는 단계로서, 표적 계수 데이터 구조가 복수의 세포의 발현 프로파일을 포함하며, 복수의 세포의 발현 프로파일은 복수의 세포 중 각각의 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함하는 단계; (b) 표적 계수 데이터 구조 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 복수의 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하는 단계로서, 덴드로그램이 복수의 노드를 포함하며, 복수의 노드가 루트 노드, 복수의 리프 노드 및 복수의 비-루트, 비-리프 노드를 포함하며, 복수의 리프 노드 중 각각의 리프 노드가 복수의 세포 중 상이한 세포의 발현 프로파일을 나타내며, 루트 노드가 복수의 세포의 발현 프로파일을 나타내는 단계; (c) 덴드로그램의 루트 노드로부터 덴드로그램의 복수의 리프 노드로 덴드로그램의 복수의 노드 중 각각의 노드를 통해 트레이싱하는 동안: (1) 노드의 자식 노드로의 노드의 분할이 유효한지 무효한(예를 들어, 자식 노드 사이의 차이가 유의미하지 않은)지를 결정하고; (2) 노드의 자식 노드로의 노드의 분할이 무효하면, 노드를 병합 클러스터 세트에 부가하는 단계; (d) 반복적으로, 병합 클러스터 세트 내의 각각의 제1 노드에 있어서, 병합 클러스터 세트 내의 제1 노드와 제1 노드에 가장 가까운 병합 클러스터 세트 내의 제2 노드 사이의 거리가 병합 거리 임계값 이내이면, 제1 노드와 제2 노드를 병합하여, 제1 노드 및 제2 노드에 의해 나타나는 발현 프로파일을 포함하는 병합된 노드를 생성하는 단계; 및 (e) 병합 클러스터 세트 내의 각각의 노드에 있어서, 노드에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계를 포함한다.
- [0067] 세포 유형을 구별하기 위한 표적의 확인을 위한 시스템이 본원에 개시된다. 일부 실시형태에서, 시스템은 하드

웨어 프로세서; 및 명령어가 저장된 비-일시적 메모리로서, 하드웨어 프로세서에 의해 실행되는 경우 프로세서가 본원에 개시된 임의의 방법을 수행하게 하는 비-일시적 메모리를 포함한다. 세포 유형을 구별하기 위한 표적의 확인을 위한 컴퓨터 판독 가능한 매체가 본원에 개시된다. 일부 실시형태에서, 컴퓨터 판독 가능한 매체는 본원에 개시된 방법 중 임의의 것을 수행하기 위한 코드를 포함한다.

[0068] 정의

[0069] 달리 정의되지 않는 한, 본원에 사용된 기술 및 과학 용어는 본 발명이 속하는 분야의 숙련자에 의해 일반적으로 이해되는 것과 동일한 의미를 갖는다. 예를 들어, 문헌[Singleton et al., Dictionary of Microbiology and Molecular Biology 2nd ed., J. Wiley & Sons (New York, NY 1994)]; 문헌[Sambrook et al., Molecular Cloning, A Laboratory Manual, Cold Springs Harbor Press (Cold Springs Harbor, NY 1989)]을 참조한다. 본 발명의 목적을 위하여, 하기의 용어가 하기에 정의된다.

[0070] 본원에 사용되는 용어 "어댑터"는 연관된 핵산의 증폭 또는 시퀀싱을 용이하게 하기 위한 서열을 의미할 수 있다. 연관된 핵산은 표적 핵산을 포함할 수 있다. 연관된 핵산은 공간 표지, 표적 표지, 샘플 표지, 인덱싱 표지, 바코드, 확률적 바코드 또는 분자 표지 중 하나 이상을 포함할 수 있다. 어댑터는 선형일 수 있다. 어댑터는 사전-아데닐화된 어댑터일 수 있다. 어댑터는 이중- 또는 단일-가닥일 수 있다. 하나 이상의 어댑터는 핵산의 5' 또는 3' 말단에 위치할 수 있다. 어댑터가 5' 및 3' 말단에 공지된 서열을 포함하는 경우, 공지된 서열은 동일한 또는 상이한 서열일 수 있다. 폴리뉴클레오타이드의 5' 및/또는 3' 말단에 위치한 어댑터는 표면 상에 고정된 하나 이상의 올리고뉴클레오타이드에 혼성화할 수 있다. 어댑터는 일부 실시형태에서, 범용 서열을 포함할 수 있다. 범용 서열은 둘 이상의 핵산 분자에 공통인 뉴클레오타이드 서열의 영역일 수 있다. 둘 이상의 핵산 분자는 상이한 서열의 영역을 가질 수 있다. 따라서, 예를 들어, 5' 어댑터는 동일한 및/또는 범용 핵산 서열을 포함할 수 있으며, 3' 어댑터는 동일한 및/또는 범용 서열을 포함할 수 있다. 복수의 핵산 분자의 상이한 구성원에 존재할 수 있는 범용 서열은 범용 서열에 상보적인 단일의 범용 프라이머를 사용하여 다수의 상이한 서열의 복제 또는 증폭을 가능하게 할 수 있다. 유사하게, 핵산 분자의 집합물의 상이한 구성원에 존재할 수 있는 적어도 1, 2개(예를 들어, 한 쌍) 이상의 범용 서열은 범용 서열에 상보적인 적어도 1, 2개(예를 들어, 한쌍) 이상의 단일의 범용 프라이머를 사용하여 다수의 상이한 서열의 복제 또는 증폭을 가능하게 할 수 있다. 따라서, 범용 프라이머는 이러한 범용 서열에 혼성화할 수 있는 서열을 포함한다. 표적 핵산 서열-함유 분자는 범용 어댑터(예를 들어, 비-표적 핵산 서열)를 상이한 표적 핵산 서열 중 하나의 또는 양 말단에 부착시키도록 변형될 수 있다. 표적 핵산에 부착되는 하나 이상의 범용 프라이머는 범용 프라이머의 혼성화를 위한 부위를 제공할 수 있다. 표적 핵산에 부착되는 하나 이상의 범용 프라이머는 서로 동일하거나 상이할 수 있다.

[0071] 본 명세서에 사용되는 바와 같이, 용어 "연관된" 또는 "~와 연관된"은 2개 이상의 종이 소정의 시점에 동시-위치하는 것으로 확인될 수 있음을 의미할 수 있다. 연관된 2개 이상의 종이 유사한 용기 내에 존재하거나, 존재하였음을 의미할 수 있다. 연관은, 예를 들어 2개 이상의 종에 관한 디지털 정보가 저장되어 있으며, 1개 이상의 종이 소정의 시점에 동시-위치한 것을 결정하기 위해 사용될 수 있는 정보과학 연관일 수 있다. 연관은 또한 물리적 연관일 수 있다. 일부 실시형태에서, 2개 이상의 연관된 종은 서로 또는 공통의 고체 또는 반고체 표면에 "테더링되거나", "부착되거나", "고정된다". 연관은 비드와 같은 고체 또는 반고체 지지체에 표지를 부착하기 위한 공유 또는 비공유 수단을 지칭할 수 있다. 연관은 표적과 표지 사이의 공유 결합일 수 있다.

[0072] 본 명세서에 사용되는 바와 같이, 용어 "상보성"은 2개의 뉴클레오타이드 사이의 정확한 쌍형성을 위한 능력을 지칭할 수 있다. 예를 들어, 핵산의 주어진 위치의 뉴클레오타이드가 또 다른 핵산의 뉴클레오타이드와 수소 결합할 수 있는 경우, 2개의 핵산은 그 위치에서 서로 상보성인 것으로 간주된다. 2개의 단일 가닥 핵산 분자 사이의 상보성은, 단지 일부의 뉴클레오타이드만이 결합하는 경우 "부분적"일 수 있거나, 단일 가닥 분자 사이에 완전한 상보성이 존재하는 경우, 완전할 수 있다. 제1 뉴클레오타이드 서열이 제2 뉴클레오타이드 서열에 상보성인 경우, 제1 뉴클레오타이드 서열은 제2 서열의 "상보체"라고 할 수 있다. 제1 뉴클레오타이드 서열이 제2 서열의 역방향(즉, 뉴클레오타이드의 순서가 역방향인) 서열에 상보성인 경우, 제1 뉴클레오타이드 서열은 제2 서열의 "역 상보체"라고 할 수 있다. 본 명세서에 사용되는 바와 같이, 용어 "상보체", "상보성" 및 "역 상보체"는 상호 교환 가능하게 사용될 수 있다. 분자가 또 다른 분자와 혼성화할 수 있는 경우, 이는 혼성화되는 분자의 상보체일 수 있는 것이 본 발명으로부터 이해된다.

[0073] 본 명세서에 사용되는 바와 같이, 용어 "디지털 계수"는 샘플에서 표적 분자의 수를 추정하는 방법을 지칭할 수 있다. 디지털 계수는 샘플에서 표적과 연관된 고유 표지의 수를 결정하는 단계를 포함할 수 있다. 이러한 확률적 방법론은, 분자 계수 문제를 동일한 분자의 위치지정 및 확인 중 하나로부터 소정의 표지 세트의 검출에 관

한 일련의 예/아니오 디지털 질문으로 변환한다.

[0074] 본 명세서에 사용되는 바와 같이, 용어 "표지" 또는 "표지들"은 샘플 내의 표적과 연관된 핵산 코드를 지칭할 수 있다. 표지는, 예를 들어, 핵산 표지일 수 있다. 표지는 전체적으로 또는 부분적으로 증폭 가능한 표지일 수 있다. 표지는 전적으로 또는 부분적으로 시퀀싱 가능한 표지일 수 있다. 표지는 별개의 것으로 확인 가능한 고유 핵산의 일부일 수 있다. 표지는 공지 서열일 수 있다. 표지는 핵산 서열의 접합부, 예를 들어 고유 및 비-고유 서열의 접합부를 포함할 수 있다. 본 명세서에 사용되는 바와 같이, 용어 "표지"는 용어 "인덱스", "태그" 또는 "표지-태그"와 상호 교환 가능하게 사용할 수 있다. 표지는 정보를 전달할 수 있다. 예를 들어, 다양한 실시형태에서, 표지는 샘플의 정체성, 샘플의 공급원, 세포의 정체성 및/또는 표적을 결정하기 위해 사용될 수 있다.

[0075] 본 명세서에 사용되는 바와 같이, 용어 "비-소모성 저장소"는 많은 상이한 표지로 이루어진 확률적 바코드의 풀(pool)을 지칭할 수 있다. 비-소모성 저장소는 다수의 상이한 확률적 바코드를 포함하여, 비-소모성 저장소가 표적의 풀과 연관되는 경우, 각각의 표적은 독특한 확률적 바코드와 연관될 수 있다. 각각의 표지된 표적 분자의 독특성은 무작위 선택 통계에 의해 결정될 수 있으며, 표지의 다양성과 비교하여, 집합물 내의 동일한 표적 분자의 카피 수에 좌우된다. 표지된 표적 분자의 생성 세트의 크기는 바코드화 과정의 확률적 특성에 의해 결정될 수 있으며, 그 후, 검출된 확률적 바코드 수의 분석에 의해, 원래의 집합물 또는 샘플에 존재하는 표적 분자의 수를 계산할 수 있다. 존재하는 표적 분자의 카피 수 대 독특한 확률적 바코드 수의 비가 낮은 경우, 표지된 표적 분자는 매우 독특하다(즉, 주어진 표지로 하나 초과와 표적 분자가 표지될 확률이 매우 낮음).

[0076] 본 명세서에 사용되는 바와 같이, 용어 "핵산"은 폴리뉴클레오티드 서열 또는 그의 단편을 지칭한다. 핵산은 뉴클레오티드를 포함할 수 있다. 핵산은 세포에 대해 외인성이거나 내인성일 수 있다. 핵산은 무 세포 환경에서 존재할 수 있다. 핵산은 유전자 또는 그의 단편일 수 있다. 핵산은 DNA일 수 있다. 핵산은 RNA일 수 있다. 핵산은 하나 이상의 유사체(예를 들어, 변경된 백본, 당 또는 핵염기)를 포함할 수 있다. 유사체의 일부 비제한적 예는 다음을 포함한다: 5-브로모우라실, 펩티드 핵산, 이중 핵산, 모르폴리노, 잠금 핵산(locked nucleic acid), 글리콜 핵산, 트레오스 핵산, 디데옥시뉴클레오티드, 코디세핀, 7-데자아-GTP, 형광단(예를 들어, 당에 연결된 로다민 또는 플루오레세인), 티올 함유 뉴클레오티드, 비오틴 연결 뉴클레오티드, 형광 염기 유사체, CpG 섬(island), 메틸-7-구아노신, 메틸화된 뉴클레오티드, 이노신, 티오우리딘, 슈도우리딘, 디하이드로우리딘, 케오신(queuosine) 및 와이오신(wyosine). "핵산", "폴리뉴클레오티드", "표적 폴리뉴클레오티드" 및 "표적 핵산"은 상호 교환 가능하게 사용될 수 있다.

[0077] 핵산은, 핵산에 새로운 또는 증진된 특징(예를 들어, 개선된 안정성)을 제공하는 하나 이상의 변형(예를 들어, 염기 변형, 백본 변형)을 포함할 수 있다. 핵산은 핵산 친화도 태그를 포함할 수 있다. 뉴클레오시드는 염기-당 조합일 수 있다. 뉴클레오시드의 염기 부분은 헤테로사이클릭 염기일 수 있다. 이러한 헤테로사이클릭 염기의 가장 흔한 2가지 부류는 퓨린 및 피리미딘이다. 뉴클레오티드는 뉴클레오시드의 당 부분에 공유적으로 연결된 포스페이트 기를 추가로 포함하는 뉴클레오시드일 수 있다. 펜토피라노실 당을 포함하는 뉴클레오시드에 있어서, 포스페이트 기는 당의 2', 3' 또는 5' 하이드록실 모이어티에 연결될 수 있다. 핵산을 형성하는 경우, 포스페이트 기는 인접한 뉴클레오시드를 서로 공유적으로 연결시켜, 선형 중합체 화합물을 형성할 수 있다. 차례로, 이러한 선형 중합체 화합물의 각각의 말단을 추가로 연결하여, 원형 화합물을 형성할 수 있으나; 선형 화합물이 일반적으로 적합하다. 또한, 선형 화합물은 내부 뉴클레오티드 염기 상보성을 가질 수 있고, 따라서 완전한 또는 부분적인 이중 가닥 화합물을 생성하는 방식으로 폴딩될 수 있다. 핵산 내에서, 포스페이트 기는 통상적으로 핵산의 뉴클레오시드 간 백본을 형성하는 것으로 지칭될 수 있다. 연결 또는 백본은 3'에서 5' 포스포 디에스테르 연결일 수 있다.

[0078] 핵산은 변형된 백본 및/또는 변형된 뉴클레오시드 간 연결을 포함할 수 있다. 변형된 백본은, 백본에 인 원자를 보유하는 것 및 백본에 인 원자를 갖지 않는 것을 포함할 수 있다. 내부에 인 원자를 함유하는 적합한 변형된 핵산 백본은, 예를 들어, 포스포로티오에이트, 키랄 포스포로티오에이트, 포스포로디티오에이트, 포스포트리에스테르, 아미노알킬 포스포트리에스테르, 메틸 및 다른 알킬 포스포네이트, 예컨대 3'-알킬렌 포스포네이트, 5'-알킬렌 포스포네이트, 키랄 포스포네이트, 포스포네이트, 3'-아미노 포스포르아미데이트 및 아미노알킬 포스포르아미데이트를 포함하는 포스포르아미데이트, 포스포로디아미데이트, 티오노포스포르아미데이트, 티오노알킬 포스포네이트, 티오노알킬포스포트리에스테르, 셀레노포스페이트 및 일반적인 3'-5' 연결을 갖는 보라노포스페이트, 2'-5' 연결된 유사체, 반대 극성을 갖는 것을 포함할 수 있으며, 여기서, 하나 이상의 뉴클레오티드 간 연결은 3'에서 3', 5'에서 5' 또는 2'에서 2' 연결이다.

- [0079] 핵산은 단쇄 알킬 또는 사이클로알킬 뉴클레오시드 간 연결, 혼합 헤테로원자 및 알킬 또는 사이클로알킬 뉴클레오시드 간 연결, 또는 하나 이상의 단쇄 헤테로원자 또는 헤테로사이클릭 뉴클레오시드 간 연결에 의해 형성된 폴리뉴클레오티드 백본을 포함할 수 있다. 이들은 모르폴리노 연결(뉴클레오시드의 당 부분으로부터 부분적으로 형성됨)을 갖는 것; 실록산 백본; 술폰, 술폰사이드 및 술폰 백본; 포름아세틸 및 티오포름아세틸 백본; 메틸렌 포름아세틸 및 티오포름아세틸 백본; 리보아세틸 백본; 알켄 함유 백본; 술폰메이트 백본; 메틸렌이미노 및 메틸렌하이드라지노 백본; 술폰메이트 및 술폰아미드 백본; 아미드 백본; 및 N, O, S 및 CH₂ 성분 부분이 혼합된 다른 것을 포함할 수 있다.
- [0080] 핵산은 핵산 모방체를 포함할 수 있다. 용어 "모방체"는, 단지 푸라노스 고리만이 또는 푸라노스 고리 및 뉴클레오티드 간 연결 둘 모두가 비-푸라노스 기로 대체된 폴리뉴클레오티드를 포함하는 것으로 의도될 수 있으며, 푸라노스 고리만의 대체는 당 대용물로 지칭될 수 있다. 헤테로사이클릭 염기 모이어티 또는 변형된 헤테로사이클릭 염기 모이어티는 적절한 표적 핵산과의 혼성화를 위해 유지될 수 있다. 이러한 핵산 중 하나는 펩티드 핵산(PNA)일 수 있다. PNA에서, 폴리뉴클레오티드의 당-백본은 아미드 함유 백본, 특히 아미노에틸글리신 백본으로 대체될 수 있다. 뉴클레오티드는 유지될 수 있으며, 백본의 아미드 부분의 아자 질소 원자에 직접적으로 또는 간접적으로 결합된다. PNA 화합물 내의 백본은 2개 이상의 연결된 아미노에틸글리신 단위를 포함할 수 있으며, 이는 PNA에 아미드 함유 백본을 제공한다. 헤테로사이클릭 염기 모이어티는 백본의 아미드 부분의 아자 질소 원자에 직접적으로 또는 간접적으로 결합될 수 있다.
- [0081] 핵산은 모르폴리노 백본 구조를 포함할 수 있다. 예를 들어, 핵산은 리보스 고리 대신 6-원 모르폴리노 고리를 포함할 수 있다. 이들 실시형태의 일부에서, 포스포로디아미데이트 또는 다른 비-포스포디에스테르 뉴클레오시드 간 연결은 포스포디에스테르 연결을 대체할 수 있다.
- [0082] 핵산은 모르폴리노 고리에 부착된 헤테로사이클릭 염기를 갖는 연결된 모르폴리노 단위(즉, 모르폴리노 핵산)를 포함할 수 있다. 연결 기는 모르폴리노 핵산에서 모르폴리노 단량체 단위를 연결할 수 있다. 비이온성 모르폴리노계 올리고머 화합물은 세포 단백질과의 바람직하지 않은 상호작용이 더 적을 수 있다. 모르폴리노계 폴리뉴클레오티드는 핵산의 비이온성 모방체일 수 있다. 모르폴리노 부류 내의 다양한 화합물은 상이한 연결 기를 사용하여 연결될 수 있다. 폴리뉴클레오티드 모방체의 추가의 부류는 사이클로헥세닐 핵산(CeNA)으로 지칭될 수 있다. 핵산 분자에 보통 존재하는 푸라노스 고리는 사이클로헥세닐 고리로 대체될 수 있다. CeNA DMT 보호된 포스포르아미드 단량체가 제조될 수 있으며, 포스포르아미드 화학을 사용하는 올리고머 화합물 합성에 사용될 수 있다. 핵산 사슬로의 CeNA 단량체의 혼입은 DNA/RNA 혼성체의 안정성을 증가시킬 수 있다. CeNA 올리고아데닐레이트는 고유 복합체와 유사한 안정성으로 핵산 상보체와 복합체를 형성할 수 있다. 추가의 변형은, 2'-하이드록실 기가 당 고리의 4' 탄소 원자에 연결되어, 2'-C, 4'-C-옥시메틸렌 연결을 형성함으로써, 바이사이클릭 당 모이어티를 형성하는 잠금 핵산(LNA)을 포함할 수 있다. 연결은 2' 산소 원자 및 4' 탄소 원자를 가교 연결하는 메틸렌(-CH₂-) 기일 수 있으며, 여기서 n은 1 또는 2이다. LNA 및 LNA 유사체는 상보성 핵산과 매우 높은 듀플렉스 열 안정성(T_m = +3 내지 +10°C), 3'-엑소핵산절단(exonucleolytic) 분해에 대한 안정성 및 우수한 용해도 특성을 나타낼 수 있다.
- [0083] 핵산은 또한 핵염기(종종 단순히 "염기"로 지칭됨) 변형 또는 치환을 포함할 수 있다. 본 명세서에 사용되는 바와 같이, "비변형된" 또는 "천연" 핵염기는 퓨린 염기(예를 들어, 아데닌(A) 및 구아닌(G)) 및 피리미딘 염기(예를 들어, 티민(T), 시토신(C) 및 우라실(U))을 포함할 수 있다. 변형된 핵염기는 다른 합성 및 천연 핵염기, 예컨대 5-메틸시토신(5-me-C), 5-하이드록시메틸 시토신, 잔틴, 하이포잔틴, 2-아미노아데닌, 아데닌 및 구아닌의 6-메틸 및 다른 알킬 유도체, 아데닌 및 구아닌의 2-프로필 및 다른 알킬 유도체, 2-티오우라실, 2-티오티민 및 2-티오시토신, 5-할로우라실 및 시토신, 5-프로피닐(-C≡C-CH₃) 우라실 및 시토신 및 피리미딘 염기의 다른 알킬 유도체, 6-아조 우라실, 시토신 및 티민, 5-우라실(슈도우라실), 4-티오우라실, 8-할로, 8-아미노, 8-티올, 8-티오알킬, 8-하이드록실 및 다른 8-치환된 아데닌 및 구아닌, 5-할로, 특히 5-브로모, 5-트리플루오로메틸 및 다른 5-치환된 우라실 및 시토신, 7-메틸구아닌 및 7-메틸아데닌, 2-F-아데닌, 2-아미노아데닌, 8-아자구아닌 및 8-아자아데닌, 7-데아자구아닌 및 7-데아자아데닌 및 3-데아자구아닌 및 3-데아자아데닌을 포함할 수 있다. 변형된 핵 염기는 트리사이클릭 피리미딘, 예컨대 페녹사진 시티딘(1H-피리미도(5,4-b)(1,4)벤조자진-2(3H)-온), 페노티아진 시티딘(1H-피리미도(5,4-b)(1,4)벤조티아진-2(3H)-온), G-클램프, 예컨대 치환된 페녹사진 시티딘(예를 들어, 9-(2-아미노에톡시)-H-피리미도(5,4-(b)(1,4)벤조자진-2(3H)-온), 페노티아진 시티딘(1H-피리미도(5,4-b)(1,4)벤조티아진-2(3H)-온), G-클램프, 예컨대 치환된 페녹사진 시티딘(예를 들어, 9-(2-아미노에톡시)-H-피리미도(5,4-(b)(1,4)벤조자진-2(3H)-온), 카바졸 시티딘(2H-피리미도(4,5-b)인돌-2-온), 피리도인돌 시티딘(H-피리도(3',2':4,5)피롤로[2,3-d]피리미딘-2-온)을 포함할 수 있다.

- [0084] 본 명세서에 사용되는 바와 같이, 용어 "샘플"은 표적을 포함하는 조성물을 지칭할 수 있다. 개시된 방법, 장치 및 시스템에 의한 분석에 적합한 샘플은 세포, 조직, 기관 또는 유기체를 포함한다.
- [0085] 본 명세서에 사용되는 바와 같이, 용어 "샘플링 장치" 또는 "장치"는, 샘플의 섹션을 취할 수 있고/거나, 기관 상에 섹션을 배치할 수 있는 장치를 지칭할 수 있다. 샘플 장치는, 예를 들어 형광 활성화된 세포 분류(FACS)기, 세포 분류기, 생검 바늘, 생검 장치, 조직 섹션화 장치, 미세유체 장치, 블레이드 그리드(blade grid) 및/또는 마이크로톰(microtome)을 지칭할 수 있다.
- [0086] 본 명세서에 사용되는 바와 같이, 용어 "고체 지지체"는 복수의 확률적 바코드가 부착될 수 있는 별개의 고체 또는 반고체 표면을 지칭할 수 있다. 고체 지지체는, 핵산이 (예를 들어, 공유적으로 또는 비공유적으로) 고정될 수 있는, 플라스틱, 세라믹, 금속 또는 중합체 재료(예를 들어, 하이드로겔)로 구성된 임의의 유형의 고체, 다공성 또는 중공 구, 볼, 베어링, 실린더 또는 다른 유사한 형태를 포괄할 수 있다. 고체 지지체는, 구형(예를 들어, 미소구형)일 수 있거나, 비-구형 또는 불규칙한 형상, 예컨대 입체, 입방체, 피라미드형, 원통형, 원추형, 직사각형 또는 디스크형 등을 가질 수 있는 별개의 입자를 포함할 수 있다. 어레이에서 이격된 복수의 고체 지지체는 기관을 포함하지 않을 수 있다. 고체 지지체는 용어 "비드"와 상호 교환 가능하게 사용될 수 있다.
- [0087] 고체 지지체는 "기관"을 지칭할 수 있다. 기관은 고체 지지체의 하나의 유형일 수 있다. 기관은 본 발명의 방법이 수행될 수 있는 연속 고체 또는 반-고체 표면을 지칭할 수 있다. 기관은 예를 들어, 어레이, 카트리지, 칩, 장치 및 슬라이드를 지칭할 수 있다.
- [0088] 본 명세서에 사용되는 바와 같이, 용어 "공간 표지"는 공간 내의 소정의 위치와 연관될 수 있는 표지를 지칭할 수 있다.
- [0089] 본 명세서에 사용되는 바와 같이, 용어 "확률적 바코드"는 표지를 포함하는 폴리뉴클레오타이드 서열을 지칭할 수 있다. 확률적 바코드는, 확률적 바코딩에 사용될 수 있는 폴리뉴클레오타이드 서열일 수 있다. 확률적 바코드는 샘플 내의 표적을 정량화하기 위해 사용될 수 있다. 확률적 바코드는, 표지가 표적과 연관된 후 발생할 수 있는 오류를 제어하기 위해 사용될 수 있다. 예를 들어, 확률적 바코드는 증폭 또는 시퀀싱 오류를 평가하기 위해 사용할 수 있다. 표적과 연관된 확률적 바코드는 확률적 바코드-표적 또는 확률적 바코드-태그-표적으로 일컬어질 수 있다.
- [0090] 본 명세서에 사용되는 바와 같이, 용어 "유전자-특이적 확률적 바코드"는 표지 및 유전자-특이적인 표적-결합 영역을 포함하는 폴리뉴클레오타이드 서열을 지칭할 수 있다. 확률적 바코드는, 확률적 바코딩에 사용될 수 있는 폴리뉴클레오타이드 서열일 수 있다. 확률적 바코드는 샘플 내의 표적을 정량화하기 위해 사용될 수 있다. 확률적 바코드는, 표지가 표적과 연관된 후 발생할 수 있는 오류를 제어하기 위해 사용될 수 있다. 예를 들어, 확률적 바코드는 증폭 또는 시퀀싱 오류를 평가하기 위해 사용할 수 있다. 표적과 연관된 확률적 바코드는 확률적 바코드-표적 또는 확률적 바코드-태그-표적으로 일컬어질 수 있다.
- [0091] 본 명세서에 사용되는 바와 같이, 용어 "확률적 바코딩"은 핵산의 무작위 표지화(예를 들어, 바코딩)를 지칭할 수 있다. 확률적 바코딩은 재귀 푸아송 전략(recursive Poisson strategy)을 사용하여, 표적과 연관시키고, 표적과 연관된 표지를 정량화할 수 있다. 본 명세서에 사용되는 바와 같이, 용어 "확률적 바코딩"은 "유전자-특이적 확률적 바코딩"과 상호 교환 가능하게 사용될 수 있다.
- [0092] 본 명세서에 사용되는 바와 같이, 용어 "표적"은 확률적 바코드와 연관될 수 있는 조성물을 지칭할 수 있다. 개시된 방법, 장치 및 시스템에 의한 분석에 적합한 예시적인 표적은 올리고뉴클레오타이드, DNA, RNA, mRNA, 마이크로RNA, tRNA 등을 포함한다. 표적은 단일 또는 이중 가닥일 수 있다. 일부 실시형태에서, 표적은 단백질일 수 있다. 일부 실시형태에서, 표적은 지질이다.
- [0093] 본 명세서에 사용되는 바와 같이, 용어 "역전사효소"는 역전사효소 활성을 갖는(즉, RNA 주형으로부터의 DNA 합성을 촉매작용시키는) 효소 군을 지칭할 수 있다. 일반적으로, 이러한 효소는 레트로바이러스 역전사효소, 레트로트랜스포존 역전사효소, 레트로플라스미드 역전사효소, 레트론(retrotron) 역전사효소, 박테리아 역전사효소, II 군 인트론-유래 역전사효소, 및 그의 돌연변이체, 변이체 또는 유도체를 포함하지만, 이에 제한되지 않는다. 비-레트로바이러스 역전사효소는 비-LTR 레트로트랜스포존 역전사효소, 레트로플라스미드 역전사효소, 레트론 역전사효소 및 II군 인트론 역전사효소를 포함한다. II군 인트론 역전사효소의 예는 락토코커스 락티스 LI.LtrB 인트론 역전사효소, 싸모사이네코코커스 엘롱가투스(*Thermosynechococcus elongatus*) TeI4c 인트론 역전사효소 또는 게오바실러스 스테아로써모필러스(*Geobacillus stearothermophilus*) GsI-IIc 인트론 역전사효소를 포함한

다. 다른 부류의 역전사효소는 많은 부류의 비-레트로바이러스 역전사효소(즉, 특히 레트로, II군 인트론 및 다양성 생성 레트로요소)를 포함할 수 있다.

[0094] 세포 유형을 구별하기 위한 표적의 확인을 위한 시스템 및 방법이 본원에 개시된다. 일부 실시형태에서, 상기 방법은 (a) 발현 프로파일을 포함하는 표적 계수 데이터 구조(예를 들어, 표적 계수 행렬)를 수신하는 단계; (b) 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 발현 프로파일을 나타내는 덴드로그램을 생성하는 단계; (c) 덴드로그램의 루트 노드로부터 덴드로그램의 리프 노드로 덴드로그램의 각각의 노드를 통해 트래버싱하는 동안: (1) 노드의 자식 노드로의 노드의 분할이 유효한지 무효한지(예를 들어, 자식 노드 사이의 차이가 유의미하지 않은지)를 결정하고; (2) 노드의 자식 노드로의 노드의 분할이 무효하면, 노드를 병합 클러스터 세트에 추가하는 단계; (d) 반복적으로, 병합 클러스터 세트 내의 각각의 제1 노드에 있어서, 병합 클러스터 세트 내의 제1 노드와 제1 노드에 가장 가까운 병합 클러스터 세트 내의 제2 노드 사이의 거리가 병합 거리 임계값 이내이면, 제1 노드와 제2 노드를 병합하여, 제1 노드 및 제2 노드에 의해 나타나는 발현 프로파일을 포함하는 병합된 노드를 생성하는 단계; 및 (e) 병합 클러스터 세트 내의 각각의 노드에 있어서, 노드에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계를 포함한다.

[0095] 바코드

[0096] 바코딩, 예컨대 확률적 바코딩은 예를 들어, US20150299784호, W02015031691호 및 문헌[Fu et al., Proc Natl Acad Sci U.S.A. 2011 May 31;108(22):9026-31] 및 문헌[Fan et al., *Science* (2015) 347(6222):1258367]에 기재되어 있으며; 이들 공보의 내용은 그의 전문이 본원에 포함된다. 일부 실시형태에서, 본원에 개시된 바코드는 표적을 확률적으로 표지(예를 들어, 바코딩, 태깅)하기 위하여 사용될 수 있는 폴리뉴클레오타이드 서열일 수 있는 확률적 바코드일 수 있다. 바코드는 확률적 바코드의 상이한 바코드 서열의 수 대 표지될 표적 중 임의의 것의 출현 횟수의 비가 1:1, 2:1, 3:1, 4:1, 5:1, 6:1, 7:1, 8:1, 9:1, 10:1, 11:1, 12:1, 13:1, 14:1, 15:1, 16:1, 17:1, 18:1, 19:1, 20:1, 30:1, 40:1, 50:1, 60:1, 70:1, 80:1, 90:1, 100:1, 또는 약 1:1, 2:1, 3:1, 4:1, 5:1, 6:1, 7:1, 8:1, 9:1, 10:1, 11:1, 12:1, 13:1, 14:1, 15:1, 16:1, 17:1, 18:1, 19:1, 20:1, 30:1, 40:1, 50:1, 60:1, 70:1, 80:1, 90:1, 100:1 또는 이들 값 중 임의의 2개 사이의 수 또는 범위일 수 있으면, 확률적 바코드로 지칭될 수 있다. 표적은 예를 들어, 동일한 또는 거의 동일한 서열을 갖는 mRNA 분자를 포함하는 mRNA 중일 수 있다. 바코드는 확률적 바코드의 상이한 바코드 서열의 수 대 표지될 표적 중 임의의 것의 출현 횟수의 비가 적어도 또는 최대 1:1, 2:1, 3:1, 4:1, 5:1, 6:1, 7:1, 8:1, 9:1, 10:1, 11:1, 12:1, 13:1, 14:1, 15:1, 16:1, 17:1, 18:1, 19:1, 20:1, 30:1, 40:1, 50:1, 60:1, 70:1, 80:1, 90:1 또는 100:1이면, 확률적 바코드로 지칭될 수 있다. 확률적 바코드의 바코드 서열은 분자 표지로 지칭될 수 있다.

[0097] 바코드, 예를 들어, 확률적 바코드는 하나 이상의 표지를 포함할 수 있다. 예시적인 표지는 범용 표지, 세포 표지, 바코드 서열(예를 들어, 분자 표지), 샘플 표지, 플레이트 표지, 공간 표지 및/또는 사전-공간 표지를 포함할 수 있다. 도 1은 공간 표지를 갖는 예시적인 바코드(104)를 예시한 것이다. 바코드(104)는 바코드를 고체 지지체(105)에 연결시킬 수 있는 5' 아민을 포함할 수 있다. 바코드는 범용 표지, 차원 표지, 공간 표지, 세포 표지 및/또는 분자 표지를 포함할 수 있다. 바코드 내의 상이한 표지(범용 표지, 차원 표지, 공간 표지, 세포 표지 및 분자 표지를 포함하나 이들에 한정되지 않음)의 순서는 달라질 수 있다. 예를 들어, 도 1에 나타난 바와 같이, 범용 표지는 가장 5'의 표지일 수 있으며, 분자 표지는 가장 3'인 표지일 수 있다. 공간 표지, 차원 표지 및 세포 표지는 임의의 순서로 존재할 수 있다. 일부 실시형태에서, 범용 표지, 공간 표지, 차원 표지, 세포 표지 및 분자 표지는 임의의 순서로 존재한다. 바코드는 표적-결합 영역을 포함할 수 있다. 표적-결합 영역은 샘플 내의 표적(예를 들어, 표적 핵산, RNA, mRNA, DNA)과 상호작용할 수 있다. 예를 들어, 표적-결합 영역은 mRNA의 폴리(A) 테일과 상호작용할 수 있는 올리고(dT) 서열을 포함할 수 있다. 일부 경우에, 바코드의 표지(예를 들어, 범용 표지, 차원 표지, 공간 표지, 세포 표지 및 바코드 서열)는 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 또는 20개 이상의 뉴클레오타이드에 의해 분리될 수 있다.

[0098] 표지, 예를 들어, 세포 표지는 오류 정정 능력을 제공하도록 설계될 수 있는, 한정된 길이, 예를 들어 각각 7개 뉴클레오타이드(일부 해밍 오류 정정 코드(Hamming error correction code)에 사용된 비트의 수와 동등함)의 독특한 핵산 하위서열 세트를 포함할 수 있다. 오류 정정 하위서열 세트는, 세트 내의 서열의 임의의 쌍별 조합이 정의된 "유전적 거리"(또는 미스매칭된 염기의 수)를 나타내도록 설계될 수 있는 7개의 뉴클레오타이드 서열을 포함하며, 예를 들어, 오류 정정 하위서열 세트는 3개의 뉴클레오타이드의 유전적 거리를 나타내도록 설계될 수 있다. 이러한 경우에, 표지된 표적 핵산 분자에 대한 서열 데이터의 세트 내의 오류 정정 서열의 검토(하기에 더욱 충분히 설명됨)는 증폭 또는 시퀀싱 오류를 검출하거나 정정하게 할 수 있다. 일부 실시형태에서, 오류 정정

코드의 생성을 위해 사용되는 핵산 하위서열의 길이는 다양할 수 있으며, 예를 들어 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 31, 40, 50개 또는 약 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 31, 40, 50개 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 뉴클레오타이드 길이일 수 있다. 일부 실시형태에서, 다른 길이의 핵산 하위서열이 오류 정정 코드를 생성하기 위하여 사용될 수 있다.

[0099] 바코드는 표적-결합 영역을 포함할 수 있다. 표적-결합 영역은 샘플 내의 표적과 상호작용할 수 있다. 표적은 리보핵산(RNA), 메신저 RNA(mRNA), 마이크로RNA, 작은 간섭 RNA(siRNA), RNA 분해 산물, 각각 폴리(A) 테일을 포함하는 RNA, 또는 이들의 임의의 조합일 수 있거나, 이를 포함할 수 있다. 일부 실시형태에서, 복수의 표적은 데옥시리보핵산(DNA)을 포함할 수 있다.

[0100] 일부 실시형태에서, 표적-결합 영역은 mRNA의 폴리(A) 테일과 상호작용할 수 있는 올리고(dT) 서열을 포함할 수 있다. 바코드의 표지(예를 들어, 범용 표지, 차원 표지, 공간 표지, 세포 표지 및 바코드 서열(예를 들어, 분자 표지)) 중 하나 이상은 바코드의 나머지 표지 중 또 다른 하나 또는 둘로부터 스페이스에 의해 분리될 수 있다. 스페이스는 예를 들어, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 또는 20개 이상의 뉴클레오타이드일 수 있다. 일부 실시형태에서, 바코드의 표지 중 어느 것도 스페이스에 의해 분리되지 않는다.

[0101] 범용 표지

[0102] 바코드는 하나 이상의 범용 표지를 포함할 수 있다. 일부 실시형태에서, 하나 이상의 범용 표지는 제공된 고체 지지체에 부착된 바코드의 세트 내의 모든 바코드에 대하여 동일할 수 있다. 일부 실시형태에서, 하나 이상의 범용 표지는 복수의 비드에 부착된 모든 바코드에 대해 동일할 수 있다. 일부 실시형태에서, 범용 표지는, 시퀀싱 프라이머에 혼성화할 수 있는 핵산 서열을 포함할 수 있다. 시퀀싱 프라이머는 범용 표지를 포함하는 바코드를 시퀀싱하기 위해 사용될 수 있다. 시퀀싱 프라이머(예를 들어, 범용 시퀀싱 프라이머)는 고처리량 시퀀싱 플랫폼과 연관된 시퀀싱 프라이머를 포함할 수 있다. 일부 실시형태에서, 범용 표지는, PCR 프라이머에 혼성화할 수 있는 핵산 서열을 포함할 수 있다. 일부 실시형태에서, 범용 표지는, 시퀀싱 프라이머 및 PCR 프라이머에 혼성화할 수 있는 핵산 서열을 포함할 수 있다. 시퀀싱 프라이머 또는 PCR 프라이머에 혼성화할 수 있는 범용 표지의 핵산 서열은 프라이머 결합 부위로서 지칭될 수 있다. 범용 표지는 바코드의 전사를 개시시키기 위해 사용될 수 있는 서열을 포함할 수 있다. 범용 표지는, 바코드 또는 바코드 내의 영역의 연장을 위해 사용될 수 있는 서열을 포함할 수 있다. 범용 표지는 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 약 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 뉴클레오타이드 길이일 수 있다. 예를 들어, 범용 표지는 적어도 약 10개의 뉴클레오타이드를 포함할 수 있다. 범용 표지는 적어도 또는 최대 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200 또는 300개의 뉴클레오타이드 길이일 수 있다. 일부 실시형태에서, 바코드가 지지체로부터 절단될 수 있게 하기 위해, 절단 가능한 링커 또는 변형된 뉴클레오타이드가 범용 표지 서열의 일부일 수 있다.

[0103] 차원 표지

[0104] 바코드는 하나 이상의 차원 표지를 포함할 수 있다. 일부 실시형태에서, 차원 표지는, 표지(예를 들어, 확률적 표지)가 발생하는 차원에 관한 정보를 제공하는 핵산 서열을 포함할 수 있다. 예를 들어, 차원 표지는, 표적이 확률적으로 바코딩된 시간에 대한 정보를 제공할 수 있다. 차원 표지는 샘플의 바코딩(예를 들어, 확률적 바코딩) 시간과 연관될 수 있다. 차원 표지는 표지 시간에 활성화될 수 있다. 상이한 차원 표지는 상이한 시간에 활성화될 수 있다. 차원 표지는, 표적, 표적의 군 및/또는 샘플이 확률적으로 바코딩되는 순서에 관한 정보를 제공한다. 예를 들어, 세포 집단은 세포 주기의 G0기에 확률적으로 바코딩될 수 있다. 세포는 세포 주기의 G1기에 바코드(예를 들어, 확률적 바코드)에 의해 다시 펄싱(pulsing)될 수 있다. 마찬가지로, 세포는 세포 주기의 S기에 바코드에 의해 다시 펄싱될 수 있다. 각각의 펄스(예를 들어, 세포 주기의 각 기)에서의 바코드는 상이한 차원 표지를 포함할 수 있다. 이러한 방식으로, 차원 표지는, 어느 단계의 세포 주기에 어떤 표적이 표지되었는지에 관한 정보를 제공한다. 차원 표지는 많은 상이한 생물학적 시간의 정보를 얻을 수 있다. 예시적인 생물학적 시간은 세포 주기, 전사(예를 들어, 전사 개시) 및 전사물 분해를 포함할 수 있지만, 이에 제한되지 않는다. 또 다른 예에서, 샘플(예를 들어, 세포, 세포 집단)은 약물 및/또는 치료법에 의한 처리 이전 및/또는 이후에 확률적으로 표지될 수 있다. 별개의 표적의 카피 수의 변화는 약물 및/또는 치료법에 대한 샘플의 반응을 나타낼 수 있다.

[0105] 차원 표지는 활성화 가능할 수 있다. 활성화 가능한 차원 표지는 특정 시점에 활성화될 수 있다. 활성화 가능한 표지는 예를 들어, 구성적으로 활성화될 수 있다(예를 들어, 턴 오프되지 않음). 활성화 가능한 차원 표지는 예

를 들어, 가역적으로 활성화될 수 있다(예를 들어, 활성화 가능한 차원 표지는 턴 온되고, 턴 오프될 수 있다). 차원 표지는 예를 들어, 적어도 1, 2, 3, 4, 5, 6, 7, 8, 9, 또는 10회 이상 가역적으로 활성화 가능할 수 있다. 차원 표지는 예를 들어, 적어도 1, 2, 3, 4, 5, 6, 7, 8, 9, 또는 10회 이상 가역적으로 활성화 가능할 수 있다. 일부 실시형태에서, 차원 표지는 형광, 광, 화학적 사건(예를 들어, 절단, 또 다른 분자의 라이게이션, 변형의 부가(예를 들어, 폐길화, 수모화(sumoylation), 아세틸화, 메틸화, 탈아세틸화, 탈메틸화)), 광화학적 사건(예를 들어, 포토케이징(photocaging)), 및 비천연 뉴클레오타이드의 도입에 의해 활성화될 수 있다.

[0106] 일부 실시형태에서, 차원 표지는 주어진 고체 지지체(예를 들어, 비드)에 부착된 모든 바코드(예를 들어, 확률적 바코드)에 대해 동일할 수 있지만, 상이한 고체 지지체(예를 들어, 비드)에 대해 상이할 수 있다. 일부 실시형태에서, 동일한 고체 지지체 상의 바코드 중 적어도 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99% 또는 100%가 동일한 차원 표지를 포함할 수 있다. 일부 실시형태에서, 동일한 고체 지지체 상의 바코드 중 적어도 60%가 동일한 차원 표지를 포함할 수 있다. 일부 실시형태에서, 동일한 고체 지지체 상의 바코드 중 적어도 95%가 동일한 차원 표지를 포함할 수 있다.

[0107] 복수의 고체 지지체(예를 들어, 비드)에 10^6 개 이상의 독특한 차원 표지 서열이 나타날 수 있다. 차원 표지는 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 약 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 뉴클레오타이드 길이일 수 있다. 차원 표지는 적어도 또는 최대 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200 또는 300개 뉴클레오타이드 길이일 수 있다. 차원 표지는 약 5 내지 약 200개의 뉴클레오타이드를 포함할 수 있다. 차원 표지는 약 10 내지 약 150개의 뉴클레오타이드를 포함할 수 있다. 차원 표지는 약 20 내지 약 125개의 뉴클레오타이드 길이를 포함할 수 있다.

[0108] 공간 표지

[0109] 바코드는 하나 이상의 공간 표지를 포함할 수 있다. 일부 실시형태에서, 공간 표지는, 바코드와 연관된 표적 분자의 공간 배향에 관한 정보를 제공하는 핵산 서열을 포함할 수 있다. 공간 표지는 샘플의 좌표와 연관될 수 있다. 좌표는 고정 좌표일 수 있다. 예를 들어, 좌표는 기관을 기준으로 고정될 수 있다. 공간 표지는 2차원 또는 3차원 그리드를 기준으로 존재할 수 있다. 좌표는 랜드마크를 기준으로 고정될 수 있다. 랜드마크는 공간에서 확인 가능할 수 있다. 랜드마크는 이미지화될 수 있는 구조일 수 있다. 랜드마크는 생물학적 구조, 예를 들어 해부학적 랜드마크일 수 있다. 랜드마크는 세포 랜드마크, 예를 들어 세포소기관일 수 있다. 랜드마크는 비천연 랜드마크, 예컨대 색상 코드, 바코드, 자성, 형광, 방사능 또는 독특한 크기 또는 형상과 같은 확인 가능한 확인자를 갖는 구조일 수 있다. 공간 표지는 물리적 구획(예를 들어, 웰, 용기 또는 액적)과 연관될 수 있다. 일부 실시형태에서, 다수의 공간 표지를 함께 사용하여, 공간 내의 하나 이상의 위치를 인코딩한다.

[0110] 공간 표지는 제공된 고체 지지체(예를 들어, 비드)에 부착된 모든 바코드에 대해 동일할 수 있지만, 상이한 고체 지지체(예를 들어, 비드)에 대해 상이할 수 있다. 일부 실시형태에서, 동일한 공간 표지를 포함하는 동일한 고체 지지체 상의 바코드의 백분율은 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99%, 100% 또는 약 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99%, 100% 또는 이들 값 중 임의의 2개 사이의 수 또는 범위일 수 있다. 일부 실시형태에서, 동일한 공간 표지를 포함하는 동일한 고체 지지체 상의 바코드의 백분율은 적어도 또는 최대 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99% 또는 100%일 수 있다. 일부 실시형태에서, 동일한 고체 지지체 상의 바코드 중 적어도 60%가 동일한 공간 표지를 포함할 수 있다. 일부 실시형태에서, 동일한 고체 지지체 상의 바코드 중 적어도 95%가 동일한 공간 표지를 포함할 수 있다.

[0111] 복수의 고체 지지체(예를 들어, 비드)에 10^6 개 이상의 독특한 공간 표지 서열이 나타날 수 있다. 공간 표지는 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 약 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 뉴클레오타이드 길이일 수 있다. 공간 표지는 적어도 또는 최대 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200 또는 300개 뉴클레오타이드 길이일 수 있다. 공간 표지는 약 5 내지 약 200개의 뉴클레오타이드를 포함할 수 있다. 공간 표지는 약 10 내지 약 150개의 뉴클레오타이드를 포함할 수 있다. 공간 표지는 약 20 내지 약 125개의 뉴클레오타이드 길이를 포함할 수 있다.

[0112] 세포 표지

[0113] 바코드는 하나 이상의 세포 표지를 포함할 수 있다. 일부 실시형태에서, 세포 표지는 어떤 표적 핵산이 어떤 세포로부터 유래된 것인지를 결정하기 위한 정보를 제공하는 핵산 서열을 포함할 수 있다. 일부 실시형태에서, 세포 표지는 제공된 고체 지지체(예를 들어, 비드)에 부착된 모든 바코드에 대해 동일하지만, 상이한 고체 지지체

(예를 들어, 비드)에 대해 상이하다. 일부 실시형태에서, 동일한 세포 표지를 포함하는 동일한 고체 지지체 상의 바코드의 백분율은 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99%, 100% 또는 약 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99%, 100% 또는 이들 값 중 임의의 2개 사이의 수 또는 범위일 수 있다. 일부 실시형태에서, 동일한 세포 표지를 포함하는 동일한 고체 지지체 상의 바코드의 백분율은 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99% 또는 100% 또는 약 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99% 또는 100%일 수 있다. 예를 들어, 동일한 고체 지지체 상의 바코드 중 적어도 60%는 동일한 세포 표지를 포함할 수 있다. 또 다른 예로서, 동일한 고체 지지체 상의 바코드 중 적어도 95%는 동일한 세포 표지를 포함할 수 있다.

[0114] 복수의 고체 지지체(예를 들어, 비드)에 10^6 개 이상의 독특한 세포 표지 서열이 나타날 수 있다. 세포 표지는 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 약 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 뉴클레오티드 길이일 수 있다. 세포 표지는 적어도 또는 최대 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200 또는 300개 뉴클레오티드 길이일 수 있다. 예를 들어, 세포 표지는 약 5 내지 약 200개의 뉴클레오티드를 포함할 수 있다. 또 다른 예로서, 세포 표지는 약 10 내지 약 150개의 뉴클레오티드를 포함할 수 있다. 또 다른 예로서, 세포 표지는 약 20 내지 약 125개의 뉴클레오티드 길이를 포함할 수 있다.

[0115] 바코드 서열

[0116] 바코드는 하나 이상의 바코드 서열을 포함할 수 있다. 일부 실시형태에서, 바코드 서열은 바코드에 혼성화되는 표적 핵산 종의 구체적인 유형에 대한 확인 정보를 제공하는 핵산 서열을 포함할 수 있다. 바코드 서열은 바코드(예를 들어, 표적-결합 영역)에 혼성화되는 표적 핵산 종의 구체적인 발생에 대한 계수기를 제공하는(예를 들어, 대략적인 근사치를 제공하는) 핵산 서열을 포함할 수 있다.

[0117] 일부 실시형태에서, 다양한 바코드 서열의 세트가 제공된 고체 지지체(예를 들어, 비드)에 부착된다. 일부 실시형태에서, 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 개 또는 약 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 개 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 독특한 분자 표지 서열이 존재할 수 있다. 예를 들어, 복수의 바코드는 별개의 서열을 갖는 약 6561개의 바코드 서열을 포함할 수 있다. 또 다른 예로서, 복수의 바코드는 별개의 서열을 갖는 약 65536개의 바코드 서열을 포함할 수 있다. 일부 실시형태에서, 적어도 또는 최대 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 또는 10^9 개의 독특한 바코드 서열이 존재할 수 있다. 독특한 분자 표지 서열은 제공된 고체 지지체(예를 들어, 비드)에 부착될 수 있다.

[0118] 바코드는 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개, 또는 약 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 뉴클레오티드 길이일 수 있다. 바코드는 적어도 또는 최대 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200 또는 300개 뉴클레오티드 길이일 수 있다.

[0119] 분자 표지

[0120] 확률적 바코드는 하나 이상의 분자 표지를 포함할 수 있다. 분자 표지는 바코드 서열을 포함할 수 있다. 일부 실시형태에서, 분자 표지는 확률적 바코드에 혼성화되는 표적 핵산 종의 구체적인 유형에 관한 확인 정보를 제공하는 핵산 서열을 포함할 수 있다. 분자 표지는 확률적 바코드(예를 들어, 표적-결합 영역)에 혼성화되는 표적 핵산 종의 구체적인 발생에 대한 계수기를 제공하는 핵산 서열을 포함할 수 있다.

[0121] 일부 실시형태에서, 다양한 분자 표지의 세트가 제공된 고체 지지체(예를 들어, 비드)에 부착된다. 일부 실시형태에서, 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 개 또는 약 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 개 또는 수 또는 범위의 독특한 분자 표지 서열이 존재할 수 있다. 예를 들어, 복수의 확률적 바코드는 별개의 서열을 갖는 약 6561개의 분자 표지를 포함할 수 있다. 또 다른 예로서, 복수의 확률적 바코드는 별개의 서열을 갖는 약 65536개의 분자 표지를 포함할 수 있다. 일부 실시형태에서, 적어도 또는 최대 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 또는 10^9 개의 독특한 분자 표지 서열이 존재할 수 있다. 독특한 분자 표지 서열을 갖는 확률적 바코드가 제공된 고체 지지체(예를 들어, 비드)에 부착될 수 있다.

[0122] 복수의 확률적 바코드를 사용한 확률적 바코딩을 위하여, 상이한 분자 표지 서열의 수 대 표적 중 임의의 것의

출현 횟수의 비는 1:1, 2:1, 3:1, 4:1, 5:1, 6:1, 7:1, 8:1, 9:1, 10:1, 11:1, 12:1, 13:1, 14:1, 15:1, 16:1, 17:1, 18:1, 19:1, 20:1, 30:1, 40:1, 50:1, 60:1, 70:1, 80:1, 90:1, 100:1 또는 약 1:1, 2:1, 3:1, 4:1, 5:1, 6:1, 7:1, 8:1, 9:1, 10:1, 11:1, 12:1, 13:1, 14:1, 15:1, 16:1, 17:1, 18:1, 19:1, 20:1, 30:1, 40:1, 50:1, 60:1, 70:1, 80:1, 90:1, 100:1, 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 수 또는 범위일 수 있다. 표적은 동일한 또는 거의 동일한 서열을 갖는 mRNA 분자를 포함하는 mRNA 중일 수 있다. 일부 실시형태에서, 상이한 분자 표지 서열의 수 대 표적 중 임의의 것의 출현의 수의 비는 적어도 또는 최대 1:1, 2:1, 3:1, 4:1, 5:1, 6:1, 7:1, 8:1, 9:1, 10:1, 11:1, 12:1, 13:1, 14:1, 15:1, 16:1, 17:1, 18:1, 19:1, 20:1, 30:1, 40:1, 50:1, 60:1, 70:1, 80:1, 90:1 또는 100:1이다.

[0123] 분자 표지는 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 약 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 뉴클레오타이드 길이일 수 있다. 분자 표지는 적어도 또는 최대 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200 또는 300개 뉴클레오타이드 길이일 수 있다.

[0124] 표적-결합 영역

[0125] 바코드는 하나 이상의 표적 결합 영역, 예컨대 포획 프로브를 포함할 수 있다. 일부 실시형태에서, 표적-결합 영역은 관심 표적과 혼성화할 수 있다. 일부 실시형태에서, 표적 결합 영역은 표적(예를 들어, 분석될 표적 핵산, 표적 분자, 예를 들어, 세포 핵산)에, 예를 들어, 특정 유전자 서열에 특이적으로 혼성화하는 핵산 서열을 포함할 수 있다. 일부 실시형태에서, 표적 결합 영역은 특정 표적 핵산의 특정 위치에 부착(예를 들어, 혼성화)할 수 있는 핵산 서열을 포함할 수 있다. 일부 실시형태에서, 표적 결합 영역은 제한 효소 부위 오버행(예를 들어, EcoRI 점착성-말단 오버행)에 특이적으로 혼성화할 수 있는 핵산 서열을 포함할 수 있다. 이어서, 바코드는 제한 효소 오버행에 상보적인 서열을 포함하는 임의의 핵산 분자에 라이게이션될 수 있다.

[0126] 일부 실시형태에서, 표적 결합 영역은 비-특이적인 표적 핵산 서열을 포함할 수 있다. 비-특이적인 표적 핵산 서열은 표적 핵산의 특정 서열과 관계 없이, 다수의 표적 핵산에 결합할 수 있는 서열을 지칭할 수 있다. 예를 들어, 표적 결합 영역은 무작위 다량체 서열 또는 mRNA 분자 상의 폴리(A) 테일에 혼성화하는 올리고(dT) 서열을 포함할 수 있다. 무작위 다량체 서열은 예를 들어, 임의의 길이의 무작위 2량체, 3량체, 4량체, 5량체, 6량체, 7량체, 8량체, 9량체, 10량체 또는 그이상의 다량체 서열일 수 있다. 일부 실시형태에서, 표적 결합 영역은 제공된 비드에 부착된 모든 바코드에 대해 동일하다. 일부 실시형태에서, 주어진 비드에 부착된 복수의 바코드에 대한 표적 결합 영역은 2개 이상의 상이한 표적 결합 서열을 포함할 수 있다. 표적 결합 영역은 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 약 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 뉴클레오타이드 길이일 수 있다. 표적 결합 영역은 최대 약 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 이상의 뉴클레오타이드 길이일 수 있다.

[0127] 일부 실시형태에서, 표적-결합 영역은 폴리아데닐화 말단을 포함하는 mRNA와 혼성화할 수 있는 올리고(dT)를 포함할 수 있다. 표적-결합 영역은 유전자-특이적일 수 있다. 예를 들어, 표적-결합 영역은 표적의 특정 영역에 혼성화하도록 구성될 수 있다. 표적-결합 영역은 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30개 또는 약 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30개 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 뉴클레오타이드 길이일 수 있다. 표적-결합 영역은 적어도 또는 최대 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 또는 30개 뉴클레오타이드 길이일 수 있다. 표적-결합 영역은 약 5 내지 30개 뉴클레오타이드 길이일 수 있다. 바코드가 유전자-특이적 표적-결합 영역을 포함하는 경우, 바코드는 본원에 유전자-특이적 바코드로 지칭될 수 있다.

[0128] 배향 특성

[0129] 바코드는 바코드를 배향(예를 들어, 정렬)하는 데 사용될 수 있는 하나 이상의 배향 특성을 포함할 수 있다. 바코드는 등전점 포커싱(isoelectric focusing)을 위한 모이어티를 포함할 수 있다. 상이한 바코드는 상이한 등전점 포커싱 점을 포함할 수 있다. 이들 바코드가 샘플에 도입되는 경우, 샘플을 등전점 포커싱에 적용하여, 바코드를 공지된 방식으로 배향할 수 있다. 이러한 방식으로, 배향 특성은 샘플 내의 공지된 바코드 맵(map)을 개발하기 위해 사용될 수 있다. 예시적인 배향 특성은 (예를 들어, 바코드의 크기에 기초한) 전기영동 이동성, 등전점, 스핀, 전도성 및/또는 자가-조립을 포함할 수 있다. 예를 들어, 자가-조립의 배향 특성을 갖는 바코드는 활성화시 특정 배향(예를 들어, 핵산 나노구조)으로 자가-조립될 수 있다.

[0130] 친화도 특성

[0131] 바코드는 하나 이상의 친화도 특성을 포함할 수 있다. 예를 들어, 공간 표지는 친화도 특성을 포함할 수 있다. 친화도 특성은, 또 다른 엔티티(예를 들어, 세포 수용체)로의 바코드의 결합을 용이하게 할 수 있는 화학적 및/또는 생물학적 모이어티를 포함할 수 있다. 예를 들어, 친화도 특성은 항체, 예를 들어 샘플 상의 특정 모이어티(예를 들어, 수용체)에 특이적인 항체를 포함할 수 있다. 일부 실시형태에서, 항체는 바코드를 특정 세포 유형 또는 분자에 안내할 수 있다. 특정 세포 유형 또는 분자에서 및/또는 그 근처에서 표적이 확률적으로 표지될 수 있다. 친화도 특성은 일부 실시형태에서, 항체가 바코드를 특정 위치로 안내할 수 있기 때문에, 공간 표지의 뉴클레오타이드 서열에 더하여 공간 정보를 제공할 수 있다. 항체는 치료용 항체, 예를 들어, 모노클로날 항체 또는 폴리클로날 항체일 수 있다. 항체는 인간화 또는 키메라일 수 있다. 항체는 네이키드 항체 또는 융합 항체일 수 있다.

[0132] 항체는 전장(즉, 천연 발생이거나, 보통 면역글로불린 유전자 단편 재조합 과정에 의해 형성된) 면역글로불린 분자(예를 들어, IgG 항체) 또는 항체 단편과 같은 면역글로불린 분자의 면역학적으로 활성인(즉, 특이적으로 결합하는) 부분일 수 있다.

[0133] 항체 단편은 예를 들어, F(ab')₂, Fab', Fab, Fv, sFv 등과 같은 항체의 부분일 수 있다. 일부 실시형태에서, 항체 단편은 전장 항체에 의해 인식되는 동일한 항원과 결합할 수 있다. 항체 단편은, 항체의 가변 영역으로 이루어진 단리된 단편, 예컨대, 중쇄 및 경쇄의 가변 영역으로 이루어진 "Fv" 단편 및, 경쇄 및 중쇄 가변 영역이 펩티드 링커에 의해 연결된 재조합 단일 사슬 폴리펩티드 분자("scFv 단백질")를 포함할 수 있다. 예시적인 항체는 암 세포에 대한 항체, 바이러스에 대한 항체, 세포 표면 수용체(CD8, CD34, CD45)에 결합하는 항체 및 치료용 항체를 포함할 수 있지만, 이에 제한되지 않는다.

[0134] 범용 어댑터 프라이머

[0135] 바코드는 하나 이상의 범용 어댑터 프라이머를 포함할 수 있다. 예를 들어, 유전자-특이적 바코드, 예컨대 유전자-특이적 확률적 바코드는 범용 어댑터 프라이머를 포함할 수 있다. 범용 어댑터 프라이머는 모든 바코드에 걸쳐 범용인 뉴클레오타이드 서열을 지칭할 수 있다. 범용 어댑터 프라이머는 유전자-특이적 바코드를 구축하는데 사용될 수 있다. 범용 어댑터 프라이머는 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30개 또는 약 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30개 또는 이들 뉴클레오타이드 길이 중 임의의 2개 사이의 수 또는 범위일 수 있다. 범용 어댑터 프라이머는 적어도 또는 최대 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 또는 30개 뉴클레오타이드 길이일 수 있다. 범용 어댑터 프라이머는 5 내지 30개 뉴클레오타이드 길이일 수 있다.

[0136] 링커

[0137] 바코드가 1가지 유형의 표지를 1개 초과로(예를 들어, 1개 초과의 세포 표지 또는 1개 초과의 바코드 서열, 예컨대 1개의 분자 표지) 포함하는 경우, 표지에는 링커 표지 서열이 산재될 수 있다. 링커 표지 서열은 적어도 약 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 이상의 뉴클레오타이드 길이일 수 있다. 링커 표지 서열은 최대 약 5, 10, 15, 20, 25, 30, 35, 40, 45, 50개 이상의 뉴클레오타이드 길이일 수 있다. 일부 경우에, 링커 표지 서열은 12개 뉴클레오타이드 길이이다. 링커 표지 서열은 바코드의 합성을 용이하게 하기 위해 사용될 수 있다. 링커 표지는 오류-정정(예를 들어, 해밍) 코드를 포함할 수 있다.

[0138] 고체 지지체

[0139] 본원에 개시된 바코드, 예컨대 확률적 바코드는 일부 실시형태에서, 고체 지지체와 연관될 수 있다. 고체 지지체는 예를 들어, 합성 입자일 수 있다. 일부 실시형태에서, 고체 지지체 상의 복수의 바코드(예를 들어, 제1 복수의 바코드)의 확률적 바코드(예를 들어, 제1 바코드 서열)를 위한 바코드 서열, 예컨대 분자 표지의 일부 또는 전부는 적어도 하나의 뉴클레오타이드가 상이하다. 동일한 고체 지지체 상의 바코드의 세포 표지는 동일할 수 있다. 상이한 고체 지지체 상의 바코드의 세포 표지는 적어도 하나의 뉴클레오타이드가 상이할 수 있다. 예를 들어, 제1 고체 지지체 상의 제1의 복수의 바코드의 제1 세포 표지는 동일한 서열을 가질 수 있으며, 제2 고체 지지체 상의 제2의 복수의 바코드의 제2 세포 표지는 동일한 서열을 가질 수 있다. 제1 고체 지지체 상의 제1의 복수의 바코드의 제1 세포 표지 및 제2 고체 지지체 상의 제2의 복수의 바코드의 제2 세포 표지는 적어도 하나의 뉴클레오타이드가 상이할 수 있다. 세포 표지는 예를 들어, 약 5 내지 20개 뉴클레오타이드 길이일 수 있다. 바코드 서열은 예를 들어, 약 5 내지 20개 뉴클레오타이드 길이일 수 있다. 합성 입자는 예를 들어, 비드일 수

있다.

- [0140] 비드는 예를 들어, 실리카 겔 비드, 조절 포어 유리 비드, 자성 비드, 다이나비드(Dynabead), 세파덱스(Sephadex)/세파로스(Sepharose) 비드, 셀룰로스 비드, 폴리스티렌 비드 또는 그의 임의의 조합일 수 있다. 비드는 폴리디메틸실록산(PDMS), 폴리스티렌, 유리, 폴리프로필렌, 아가로스, 젤라틴, 하이드로겔, 상자성, 세라믹, 플라스틱, 유리, 메틸스티렌, 아크릴계 중합체, 티타늄, 라텍스, 세파로스, 셀룰로스, 나일론, 실리콘 또는 그의 임의의 조합과 같은 물질을 포함할 수 있다.
- [0141] 일부 실시형태에서, 비드는 중합체 비드, 예를 들어, 바코드 또는 확률적 바코드로 작용화된 변형 가능한 비드 또는 겔 비드(예컨대 10X 게노믹스(10X Genomics)(미국 캘리포니아주 샌 프란시스코 소재)의 겔 비드)일 수 있다. 일부 구현예에서, 겔 비드는 중합체 기반의 겔을 포함할 수 있다. 겔 비드는 예를 들어, 하나 이상의 중합체 전구체를 액적으로 캡슐화시킴으로써 생성될 수 있다. 가속제(예를 들어, 테트라메틸에틸렌디아민(TEMED))에 중합체 전구체가 노출되면 겔 비드가 생성될 수 있다.
- [0142] 일부 실시형태에서, 입자는 분해 가능할 수 있다. 예를 들어, 중합체 비드는 예를 들어, 요망되는 조건 하에서 용해, 용융 또는 분해될 수 있다. 요망되는 조건은 환경 조건을 포함할 수 있다. 요망되는 조건은 제어된 방식으로 중합체 비드 용해, 용융 또는 분해를 초래할 수 있다. 겔 비드는 화학적 자극, 물리적 자극, 생물학적 자극, 열 자극, 자성 자극, 전기 자극, 광 자극 또는 그의 임의의 조합으로 인하여 용해, 용융 또는 분해될 수 있다.
- [0143] 분석물 및/또는 시약, 예컨대 올리고뉴클레오타이드 바코드는 예를 들어, 겔 비드의 내면(예를 들어, 올리고뉴클레오타이드 바코드 및/또는 올리고뉴클레오타이드 바코드를 생성하기 위해 사용되는 물질의 확산을 통해 접근 가능한 내부) 및/또는 겔 비드의 외면 또는 본원에 기재된 임의의 다른 마이크로캡슐에 커플링/고정될 수 있다. 커플링/고정은 임의의 형태의 화학적 결합(예를 들어, 공유 결합, 이온 결합) 또는 물리적 현상(예를 들어, 반데르발스 힘, 쌍극자-쌍극자 상호작용 등)을 통해 이루어질 수 있다. 일부 실시형태에서, 겔 비드 또는 본원에 기재된 임의의 다른 마이크로캡슐로의 시약의 커플링/고정은, 예컨대 불안정성 모이어티를 통해(예를 들어, 본원에 기재된 화학적 가교제를 포함하는 화학적 가교제를 통해) 가역적일 수 있다. 자극의 적용시에, 불안정성 모이어티는 절단될 수 있고, 고정된 시약은 유리된다. 일부 실시형태에서, 불안정성 모이어티는 이황화 결합이다. 예를 들어, 올리고뉴클레오타이드 바코드가 이황화 결합을 통해 겔 비드에 고정되는 경우에, 이황화 결합이 환원제에 노출됨으로써, 이황화 결합이 절단되고, 비드로부터 올리고뉴클레오타이드 바코드가 유리될 수 있다. 불안정성 모이어티는 겔 비드 또는 마이크로캡슐의 부분으로서, 시약 또는 분석물을 겔 비드 또는 마이크로캡슐에 연결하는 화학적 링커의 부분으로서 및/또는 시약 또는 분석물의 부분으로서 포함될 수 있다. 일부 실시형태에서, 복수의 바코드 중 적어도 하나의 바코드는 입자 상에 고정되거나, 입자 상에 부분적으로 고정되거나, 입자 내에 봉입되거나, 입자 내에 부분적으로 봉입되거나, 그의 임의의 조합일 수 있다.
- [0144] 일부 실시형태에서, 겔 비드는 중합체, 감열성 중합체, 감광성 중합체, 자성 중합체, pH 감수성 중합체, 염 감수성 중합체, 화학적 감수성 중합체, 다당 전해질, 다당류, 펩티드, 단백질 및/또는 플라스틱을 포함하나 이에 제한되지 않는 매우 다양한 상이한 중합체를 포함할 수 있다. 중합체는 폴리(N-이소프로필아크릴아미드)(PNIPAAm), 폴리(스티렌 술포네이트)(PSS), 폴리(알릴 아민)(PAAm), 폴리(아크릴산)(PAA), 폴리(에틸렌 이민)(PEI), 폴리(디알릴디메틸-암모늄 클로라이드)(PDADMAC), 폴리(피롤)(PPy), 폴리(비닐피롤리돈)(PVPON), 폴리(비닐 피리딘)(PVP), 폴리(메타크릴산)(PMAA), 폴리(메틸 메타크릴레이트)(PMMA), 폴리스티렌(PS), 폴리(테트라하이드로푸란)(PTHF), 폴리(프탈알데히드)(PTHF), 폴리(헥실 비올로젠)(PHV), 폴리(L-라이신)(PLL), 폴리(L-아르기닌)(PARG), 폴리(락틱-코-글리콜산)(PLGA)과 같은 물질을 포함할 수 있지만 이에 제한되지 않는다.
- [0145] 수많은 화학적 자극을 사용하여 비드의 파괴, 용해 또는 분해를 촉발시킬 수 있다. 이들 화학적 변화의 예는 비드 벽에 대한 pH-매개의 변화, 가교 결합의 화학적 절단을 통한 비드 벽의 붕괴, 비드 벽의 촉발된 해중합 및 비드 벽 전환 반응을 포함할 수 있지만, 이에 제한되지 않는다. 또한, 벌크 변화를 사용하여, 비드의 파괴를 촉발시킬 수 있다.
- [0146] 다양한 자극을 통한 마이크로캡슐에 대한 벌크 또는 물리적 변화는 또한 시약을 방출하기 위한 캡슐을 설계하는데 많은 이점을 제공한다. 벌크 또는 물리적 변화는 거시적인 규모로 발생하며, 비드 파괴는 자극에 의해 유도되는 기계-물리적 힘의 결과이다. 이들 과정은 압력 유도 파괴, 비드 벽 용융 또는 비드 벽의 다공성의 변화를 포함할 수 있지만, 이에 제한되지 않는다.
- [0147] 또한, 생물학적 자극을 사용하여 비드의 파괴, 용해 또는 분해를 촉발시킬 수 있다. 일반적으로, 생물학적 트리

거는 화학적 트리거와 비슷하지만, 많은 예는 생분자 또는 생명계에서 흔히 관찰되는 분자, 예컨대 효소, 펩티드, 당류, 지방산, 핵산 등을 사용한다. 예를 들어, 비드는 특이적인 프로테아제에 의한 절단에 감수성인 펩티드 가교 결합을 갖는 중합체를 포함할 수 있다. 더욱 구체적으로, 하나의 예는 GFLGK 펩티드 가교 결합을 포함하는 마이크로캡슐을 포함할 수 있다. 생물학적 트리거, 예컨대 프로테아제 카텝신 B의 첨가 시에, 셸(shell)의 펩티드 가교 결합이 절단되며, 비드의 내용물이 방출된다. 다른 경우에, 프로테아제는 열-활성화될 수 있다. 또 다른 예에서, 비드는 셀룰로스를 포함하는 셸 벽을 포함한다. 가수분해 효소 키토산의 첨가는 셀룰로스 결합의 절단, 셸 벽의 해중합 및 그의 내부 내용물의 방출을 위한 생물학적 트리거로서의 역할을 한다.

[0148] 또한, 비드는 열 자극의 적용시에 그들의 내용물을 방출하도록 유도될 수 있다. 온도의 변화는 비드에 다양한 변화를 야기할 수 있다. 열의 변화는 비드의 용융을 야기하여, 비드 벽이 붕괴되게 할 수 있다. 다른 경우에, 열은 비드의 내부 성분의 내압을 증가시켜, 비드가 파열되거나 폭발되게 할 수 있다. 또 다른 경우에, 열은 비드를 수축되고 탈수된 상태로 변환시킬 수 있다. 또한, 열은 비드의 벽 내의 감열성 중합체에 작용하여 비드의 파괴를 야기할 수 있다.

[0149] 마이크로캡슐의 비드 벽에 자성 나노입자를 포함시키면 비드의 파열이 촉발되게 할 뿐 아니라 비드를 어레이로 안내할 수 있다. 본 발명의 장치는 어느 하나의 목적을 위하여 자성 비드를 포함할 수 있다. 일 예에서, 다가 전해질 함유 비드로의 Fe_3O_4 나노입자를 혼입시킴으로써, 진동 자기장 자극의 존재 하에 파열이 촉발된다.

[0150] 또한, 비드는 전기 자극의 결과로서 파괴되거나, 용해되거나, 분해될 수 있다. 이전 섹션에 기재된 자성 입자와 유사하게, 전기 민감성 비드는 비드의 파열 촉발, 및 다른 기능, 예컨대 전기장에서의 정렬, 전기 전도도 또는 산화환원 반응을 가능하게 할 수 있다. 일 예에서, 전기 민감성 물질을 함유하는 비드는 내부 시약의 방출이 제어될 수 있도록 전기장에서 정렬된다. 다른 예에서, 전기장은 비드 벽 그 자체 내에서 산화환원 반응을 유도할 수 있으며, 이는 다공성을 증가시킬 수 있다.

[0151] 또한, 광 자극을 사용하여 비드를 파괴할 수 있다. 수많은 광 트리거가 가능하며, 특정 범위의 파장의 광자를 흡수할 수 있는 다양한 분자, 예컨대 나노입자 및 발색단을 사용하는 시스템을 포함할 수 있다. 예를 들어, 산화금속 코팅은 캡슐 트리거로서 사용될 수 있다. SiO_2 로 코팅된 다가 전해질 캡슐의 UV 조사는 비드 벽의 붕괴를 초래할 수 있다. 또 다른 예에서, 광 전환 가능(photo switchable) 물질, 예컨대 아조벤젠기가 비드 벽에 혼입될 수 있다. UV 또는 가시광의 적용 시에, 이들과 같은 화학물질은 광자의 흡수시 가역적인시스-트랜스 이성질화를 겪는다. 이러한 양태에서, 광자 스위치의 혼입은 비드 벽이 광 트리거의 적용 시에 붕괴되거나 더욱 다공성이 될 수 있게 한다.

[0152] 예를 들어, 도 2에 예시된 바코딩(예를 들어, 확률적 바코딩)의 비제한적인 예에서, 블록 208에서 마이크로웰 어레이의 복수의 마이크로웰 상으로의 세포, 예컨대 단일 세포의 도입 후에, 블록 212에서 비드는 마이크로웰 어레이의 복수의 마이크로웰 상으로 도입될 수 있다. 각각의 마이크로웰은 하나의 비드를 포함할 수 있다. 비드는 복수의 바코드를 포함할 수 있다. 바코드는 비드에 부착된 5' 아민 영역을 포함할 수 있다. 바코드는 범용 표지, 바코드 서열(예를 들어, 분자 표지), 표적-결합 영역 또는 그의 임의의 조합을 포함할 수 있다.

[0153] 본원에 개시된 바코드는 고체 지지체(예를 들어, 비드)와 연관(예를 들어, 이에 부착)될 수 있다. 고체 지지체와 연관된 바코드는 각각 독특한 서열을 갖는 적어도 100 또는 1000개의 바코드 서열을 포함하는 군으로부터 선택되는 바코드 서열을 포함할 수 있다. 일부 실시형태에서, 고체 지지체와 연관된 상이한 바코드는 상이한 서열의 바코드 서열을 포함할 수 있다. 일부 실시형태에서, 고체 지지체와 연관된 소정의 백분율의 바코드는 동일한 세포 표지를 포함한다. 예를 들어, 백분율은 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99%, 100% 또는 약 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99%, 100%, 또는 이들 값 중 임의의 2개 사이의 수 또는 범위일 수 있다. 또 다른 예로서, 백분율은 적어도 또는 최대 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99% 또는 100%일 수 있다. 일부 실시형태에서, 고체 지지체와 연관된 바코드는 동일한 세포 표지를 가질 수 있다. 상이한 고체 지지체와 연관된 바코드는 독특한 서열을 갖는 적어도 100 또는 1000개의 세포 표지를 포함하는 군으로부터 선택되는 상이한 세포 표지를 가질 수 있다.

[0154] 본원에 개시된 바코드는 고체 지지체(예를 들어, 비드)와 연관(예를 들어, 이에 부착)될 수 있다. 일부 실시형태에서, 샘플 내의 복수의 표적의 확률적 바코딩은 복수의 바코드와 연관된 복수의 합성 입자를 포함하는 고체 지지체를 사용하여 수행될 수 있다. 일부 실시형태에서, 고체 지지체는 복수의 바코드와 연관된 복수의 합성 입자를 포함할 수 있다. 상이한 고체 지지체 상의 복수의 바코드의 공간 표지는 적어도 하나의 뉴클레오타이드가 상이할 수 있다. 고체 지지체는 예를 들어, 2차원 또는 3차원에서 복수의 바코드를 포함할 수 있다. 합성 입자는

비드일 수 있다. 비드는 실리카 겔 비드, 조절 포어 유리 비드, 자성 비드, 다이나비드, 세파텍스/세파로스 비드, 셀룰로스 비드, 폴리스티렌 비드 또는 그의 임의의 조합일 수 있다. 고체 지지체는 중합체, 매트릭스, 하이드로겔, 니들 어레이 디바이스, 항체 또는 그의 임의의 조합을 포함할 수 있다. 일부 실시형태에서, 고체 지지체는 자유 부유할 수 있다. 일부 실시형태에서, 고체 지지체는 반고체 또는 고체 어레이에 매립될 수 있다. 바코드는 고체 지지체와 연관되지 않을 수 있다. 바코드는 개별 뉴클레오티드일 수 있다. 바코드는 기관과 연관될 수 있다.

[0155] 본 명세서에 사용되는 바와 같이, 용어 "테더링된", "부착된" 및 "고정된"은 상호 교환 가능하게 사용되며, 바코드를 고체 지지체에 부착시키기 위한 공유적 또는 비-공유적 수단을 지칭할 수 있다. 임의의 다양한 여러 가지 고체 지지체는 바코드의 동소 고체상 합성을 위해 또는 사전 합성된 바코드의 부착을 위해 고체 지지체로서 사용될 수 있다.

[0156] 일부 실시형태에서, 고체 지지체는 비드이다. 비드는, 핵산이 (예를 들어, 공유적으로 또는 비-공유적으로) 고정될 수 있는, 중실, 다공성 또는 중공 구, 볼, 베어링, 실린더 또는 다른 유사한 형상 중 하나 이상의 유형을 포함할 수 있다. 비드는 예를 들어, 플라스틱, 세라믹, 금속, 중합체 물질 또는 그의 임의의 조합으로 구성될 수 있다. 비드는 구형(예를 들어, 미소구)이거나, 정육면체, 직육면체, 피라미드형, 원통형, 원추형, 직사각형 또는 디스크형 등과 같이, 비-구형 또는 불규칙한 형상을 갖는 개별 입자일 수 있거나, 이를 포함할 수 있다. 일부 실시형태에서, 비드는 비-구형 형상일 수 있다.

[0157] 비드는 상자성 물질(예를 들어, 마그네슘, 폴리브덴, 리튬 및 탄탈), 초상자성 물질(예를 들어, 페라이트(Fe_3O_4 ; 자철석) 나노입자), 강자성 물질(예를 들어, 철, 니켈, 코발트, 그의 일부 합금 및 일부 희토류 금속 화합물), 세라믹, 플라스틱, 유리, 폴리스티렌, 실리카, 메틸스티렌, 아크릴 중합체, 티타늄, 라텍스, 세파로스, 아가로스, 하이드로겔, 중합체, 셀룰로스, 나일론 또는 그의 임의의 조합을 포함하지만, 이에 제한되지 않는 다양한 물질을 포함할 수 있다.

[0158] 일부 실시형태에서, 비드(예를 들어, 표지가 부착된 비드)는 하이드로겔 비드이다. 일부 실시형태에서, 비드는 하이드로겔을 포함한다.

[0159] 본원에 개시된 일부 실시형태는 하나 이상의 입자(예를 들어, 비드)를 포함한다. 입자의 각각은 복수의 올리고뉴클레오티드(예를 들어, 바코드)를 포함할 수 있다. 복수의 올리고뉴클레오티드의 각각은 바코드 서열(예를 들어, 분자 표지), 세포 표지 및 표적-결합 영역(예를 들어, 올리고(dT) 서열, 유전자-특이적 서열, 무작위 다량체 또는 그의 조합)을 포함할 수 있다. 복수의 올리고뉴클레오티드의 각각의 세포 표지 서열은 동일할 수 있다. 상이한 입자 상의 올리고뉴클레오티드의 세포 표지 서열은 상이하여, 상이한 입자 상의 올리고뉴클레오티드가 확인될 수 있도록 할 수 있다. 상이한 세포 표지 서열의 수는 상이한 구현예에서 상이할 수 있다. 일부 실시형태에서, 세포 표지 서열의 수는 10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 10^6 , 10^7 , 10^8 , 10^9 개 또는 약 10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 10^6 , 10^7 , 10^8 , 10^9 개 또는 이들 값 중 임의의 2개 사이의 수 또는 범위 이상일 수 있다. 일부 실시형태에서, 세포 표지 서열의 수는 적어도 또는 최대 10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 10^6 , 10^7 , 10^8 또는 10^9 개일 수 있다. 일부 실시형태에서, 복수의 입자 중 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000개 이하, 또는 그 이상이 동일한 세포 서열을 갖는 올리고뉴클레오티드를 포함한다. 일부 실시형태에서, 동일한 세포 서열을 갖는 올리고뉴클레오티드를 포함하는 복수의 입자는 최대 0.1%, 0.2%, 0.3%, 0.4%, 0.5%, 0.6%, 0.7%, 0.8%, 0.9%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10% 이상일 수 있다. 일부 실시형태에서, 복수의 입자 중 어느 것도 동일한 세포 표지 서열을 갖지 않는다.

[0160] 각각의 입자 상의 복수의 올리고뉴클레오티드는 상이한 바코드 서열(예를 들어, 분자 표지)을 포함할 수 있다. 일부 실시형태에서, 바코드 서열의 수는 10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 10^6 , 10^7 , 10^8 , 10^9 개 또는 약 10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000,

90000, 100000, 10^6 , 10^7 , 10^8 , 10^9 개, 또는 이들 값 중 임의의 2개 사이의 수 또는 범위일 수 있다. 일부 실시 형태에서, 바코드 서열의 수는 적어도 또는 최대 10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 10^6 , 10^7 , 10^8 또는 10^9 개일 수 있다. 예를 들어, 복수의 올리고뉴클레오타이드 중 적어도 100개는 상이한 바코드 서열을 포함한다. 또 다른 예로서, 단일의 입자에서, 복수의 올리고뉴클레오타이드 중 적어도 100, 500, 1000, 5000, 10000, 15000, 20000, 50000개, 이들 값 중 임의의 2개 사이의 수 또는 범위는 상이한 바코드 서열을 포함한다. 일부 실시형태는 바코드를 포함하는 복수의 입자를 제공한다. 일부 실시형태에서, 표지될 표적의 존재(또는 카피 또는 개수)에 상이한 바코드 서열의 비는 적어도 1:1, 1:2, 1:3, 1:4, 1:5, 1:6, 1:7, 1:8, 1:9, 1:10, 1:11, 1:12, 1:13, 1:14, 1:15, 1:16, 1:17, 1:18, 1:19, 1:20, 1:30, 1:40, 1:50, 1:60, 1:70, 1:80, 1:90 이상일 수 있다. 일부 실시형태에서, 복수의 올리고뉴클레오타이드의 각각은 샘플 표지, 범용 표지 또는 둘 모두를 추가로 포함한다. 입자는 예를 들어, 나노입자 또는 마이크로입자일 수 있다.

[0161] 비드의 크기는 달라질 수 있다. 예를 들어, 비드의 직경은 0.1 마이크로미터 내지 50 마이크로미터의 범위일 수 있다. 일부 실시형태에서, 비드의 직경은 0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50 마이크로미터 또는 약 0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50 마이크로미터, 또는 이들 값 중 임의의 2개 사이의 수 또는 범위일 수 있다.

[0162] 비드의 직경은 기관의 웰의 직경과 관련될 수 있다. 일부 실시형태에서, 비드의 직경은 웰의 직경보다 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100% 또는 약 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100% 또는 이들 값 중 임의의 2개 사이의 수 또는 범위 만큼 더 길거나 더 짧을 수 있다. 비드의 직경은 세포(예를 들어, 기관의 웰에 의해 포획되는 단일 세포)의 직경과 관련될 수 있다. 일부 실시형태에서, 비드의 직경은 웰의 직경보다 적어도 또는 최대 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% 또는 100% 더 길거나 더 짧을 수 있다. 비드의 직경은 세포(예를 들어, 기관의 웰에 의해 포획되는 단일 세포)의 직경과 관련될 수 있다. 일부 실시형태에서, 비드의 직경은 세포의 직경보다 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, 150%, 200%, 250%, 300% 또는 약 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, 150%, 200%, 250%, 300%, 또는 이들 값 중 임의의 2개 사이의 수 또는 범위 만큼 더 길거나 더 짧을 수 있다. 일부 실시형태에서, 비드의 직경은 세포의 직경보다 적어도 또는 최대 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, 150%, 200%, 250% 또는 300% 더 길거나 더 짧을 수 있다.

[0163] 비드는 기관에 부착되고/거나, 매립될 수 있다. 비드는 겔, 하이드로겔, 중합체 및/또는 매트릭스에 부착되고/거나, 매립될 수 있다. 기관(예를 들어, 겔, 매트릭스, 스캐폴드 또는 중합체) 내의 비드의 공간적 위치는, 위치 주소로서 기능할 수 있는 비드 상의 바코드 상에 존재하는 공간 표지를 사용하여 확인될 수 있다.

[0164] 비드의 예는 스트렙트아비딘 비드, 아가로스 비드, 자성 비드, 다이아비드®, MACS® 마이크로비드, 항체 컨쥬게이트된 비드(예를 들어, 항-면역글로불린 마이크로비드), 단백질 A 컨쥬게이트된 비드, 단백질 G 컨쥬게이트된 비드, 단백질 A/G 컨쥬게이트된 비드, 단백질 L 컨쥬게이트된 비드, 올리고(dT) 컨쥬게이트된 비드, 실리카 비드, 실리카-유사 비드, 항-비오틴 마이크로비드, 항-형광색소 마이크로비드 및 BcMag™ 카복실-말단 자성 비드를 포함할 수 있지만, 이에 제한되지 않는다.

[0165] 비드는 하나의 형광 광학 채널 또는 다중 광학 채널에서 형광이 되도록 양자점 또는 형광 염료와 연관(예를 들어, 이로 함침)될 수 있다. 비드는 산화철 또는 산화크롬과 연관되어, 상자성 또는 강자성이 될 수 있다. 비드는 확인 가능할 수 있다. 예를 들어, 비드는 카메라를 사용하여 이미지화할 수 있다. 비드는 비드와 연관된 검출 가능한 코드를 가질 수 있다. 예를 들어, 비드는 바코드를 포함할 수 있다. 비드는, 예를 들어 유기 또는 무기 용액에서의 팽윤으로 인해, 크기가 변할 수 있다. 비드는 소수성일 수 있다. 비드는 친수성일 수 있다. 비드는 생체적합성일 수 있다.

[0166] 고체 지지체(예를 들어, 비드)는 시각화될 수 있다. 고체 지지체는 시각화 태그(예를 들어, 형광 염료)를 포함할 수 있다. 고체 지지체(예를 들어, 비드)는 확인자(예컨대, 숫자)로 예칭될 수 있다. 확인자는 비드를 이미지화함으로써 시각화될 수 있다.

[0167] 고체 지지체는 불용성, 반-용해성 또는 불용성 물질을 포함할 수 있다. 고체 지지체는, 그에 부착된 링커, 스캐폴드, 빌딩 블록 또는 다른 반응성 모이어티를 포함하는 경우, "작용화된(functionalized)" 것으로 지칭될 수 있는 한편, 고체 지지체는, 그에 부착된 이러한 반응성 모이어티가 결여된 경우, "비작용화된" 것일 수 있다. 고체 지지체는 용액, 예컨대 마이크로타이터 웰 형식; 관류 형식, 예컨대 컬럼; 또는 딥스틱(dipstick)에서 자

유롭게 사용될 수 있다.

[0168] 고체 지지체는 막, 종이, 플라스틱, 코팅 표면, 평평한 표면, 유리, 슬라이드, 칩 또는 그의 임의의 조합을 포함할 수 있다. 고체 지지체는 수지, 겔, 미소구 또는 다른 기하학적 형상의 형태를 취할 수 있다. 고체 지지체는 실리카 칩, 마이크로입자, 나노입자, 플레이트, 어레이, 모세관, 평평한 지지체, 예컨대 유리 섬유 필터, 유리 표면, 금속 표면(강철, 금은, 알루미늄, 실리콘 및 구리), 유리 지지체, 플라스틱 지지체, 실리콘 지지체, 칩, 필터, 막, 마이크로웰 플레이트, 슬라이드, 멀티웰 플레이트 또는 막을 포함한 플라스틱 물질(예를 들어, 폴리에틸렌, 폴리프로필렌, 폴리아미드, 폴리비닐리덴디플루오라이드로 형성됨) 및/또는 웨이퍼, 빔, 핀 또는 니들(예를 들어, 조합 합성 또는 분석에 적합한 핀의 어레이) 또는 웨이퍼(예를 들어, 실리콘 웨이퍼), 필터 바닥이 존재하거나, 부재하는 피트가 구비된 웨이퍼와 같은 평평한 표면의 피트 또는 나노리터 웰의 어레이 내의 비드를 포함할 수 있다.

[0169] 고체 지지체는 중합체 매트릭스(예를 들어, 겔, 하이드로겔)를 포함할 수 있다. 중합체 매트릭스는 세포 내 공간(예를 들면, 세포소기관 주위)을 투과할 수 있다. 중합체 매트릭스는 순환계를 통해 펌핑될 수 있다.

[0170] 고체 지지체는 생물학적 분자일 수 있다. 예를 들어, 고체 지지체는 핵산, 단백질, 항체, 히스톤, 세포 구획, 지질, 탄수화물 등일 수 있다. 생물학적 분자인 고체 지지체는 증폭, 번역, 전사, 분해, 및/또는 변형(예를 들어, 폐길화, 수모화, 아세틸화, 메틸화)될 수 있다. 생물학적 분자인 고체 지지체는, 생물학적 분자에 부착된 공간 표지에 더하여, 공간 및 시간 정보를 제공할 수 있다. 예를 들어, 생물학적 분자는, 변형되지 않은 경우, 제1 확인을 포함할 수 있지만, 변형된 경우, 제2 확인으로 변할 수 있다. 상이한 형태는 본 발명의 바코드(예를 들어, 확률적 바코드)를 표적에 노출시킬 수 있다. 예를 들어, 생물학적 분자는, 생물학적 분자의 폴딩으로 인해 접근 가능하지 않은 바코드를 포함할 수 있다. 생물학적 분자의 변형(예를 들어, 아세틸화) 시에, 생물학적 분자는 형태를 변화시켜, 바코드를 노출시킬 수 있다. 변형의 시기는 본 발명의 바코딩 방법에 또 다른 시간 차원을 제공할 수 있다.

[0171] 일부 실시형태에서, 본 발명의 바코드 시약을 포함하는 생물학적 분자는 세포의 세포질에 위치할 수 있다. 활성화 시에, 생물학적 분자는 핵으로 이동할 수 있으며, 이때 바코딩이 이루어질 수 있다. 이러한 방식으로, 생물학적 분자의 변형은 바코드에 의해 확인되는 표적에 대한 추가적인 시공간 정보를 인코딩할 수 있다.

[0172] 기관 및 마이크로웰 어레이

[0173] 본원에 사용되는 바와 같이, 기관은 하나의 유형의 고체 지지체를 지칭할 수 있다. 기관은 본 발명의 바코드 및 확률적 바코드를 포함할 수 있는 고체 지지체를 지칭할 수 있다. 기관은 예를 들어, 복수의 마이크로웰을 포함할 수 있다. 예를 들어, 기관은 2개 이상의 마이크로웰을 포함하는 웰 어레이일 수 있다. 일부 실시형태에서, 마이크로웰은 정의된 부피의 작은 반응 챔버를 포함할 수 있다. 일부 실시형태에서, 마이크로웰은 하나 이상의 세포를 포획할 수 있다. 일부 실시형태에서, 마이크로웰은 오직 하나의 세포만을 포획할 수 있다. 일부 실시형태에서, 마이크로웰은 하나 이상의 고체 지지체를 포획할 수 있다. 일부 실시형태에서, 마이크로웰은 오직 하나의 고체 지지체만을 포획할 수 있다. 일부 실시형태에서, 마이크로웰은 단일 세포 및 단일 고체 지지체(예를 들어, 비드)를 포획한다. 마이크로웰은 본 발명의 조합적 바코드 시약을 포함할 수 있다.

[0174] 바코딩 방법

[0175] 본 발명은 신체 샘플(예를 들어, 조직, 기관, 종양, 세포) 내의 별개의 위치에서의 별개의 표적의 수를 추정하기 위한 방법을 제공한다. 상기 방법은 바코드(예를 들어, 확률적 바코드)를 샘플과 매우 근접하게 배치하는 단계, 샘플을 용해시키는 단계, 별개의 표적을 바코드와 연관시키는 단계, 표적을 증폭시키는 단계 및/또는 표적을 디지털 계수하는 단계를 포함할 수 있다. 상기 방법은 바코드 상의 공간 표지로부터 획득되는 정보를 분석 및/또는 시각화하는 단계를 추가로 포함할 수 있다. 일부 실시형태에서, 방법은 샘플 내의 복수의 표적을 시각화시키는 단계를 포함한다. 샘플의 맵 상으로 복수의 표적을 맵핑하는 것은 샘플의 2차원 맵 또는 3차원 맵을 생성하는 것을 포함할 수 있다. 2차원 맵 및 3차원 맵은 샘플 내의 복수의 표적의 바코딩(예를 들어, 확률적 바코딩) 이전에 또는 그 이후에 생성될 수 있다. 샘플 내의 복수의 표적의 시각화는 샘플의 맵 상으로 복수의 표적을 맵핑하는 것을 포함할 수 있다. 샘플의 맵 상으로 복수의 표적을 맵핑하는 것은 샘플의 2차원 맵 또는 3차원 맵을 생성하는 것을 포함할 수 있다. 2차원 맵 및 3차원 맵은 샘플 내의 복수의 표적의 바코딩 이전에 또는 그 이후에 생성될 수 있다. 일부 실시형태에서, 2차원 맵 및 3차원 맵은 샘플을 용해시키기 이전 또는 이후에 생성될 수 있다. 2차원 맵 또는 3차원 맵을 생성하기 이전 또는 이후의 샘플의 용해는 샘플의 가열, 샘플과 세제의 접촉, 샘플의 pH의 변화 또는 그의 임의의 조합을 포함할 수 있다.

- [0176] 일부 실시형태에서, 복수의 표적의 바코딩은 복수의 바코드를 복수의 표적과 혼성화시켜, 바코딩된 표적(예를 들어, 확률적으로 바코딩된 표적)을 생성하는 것을 포함한다. 복수의 표적의 바코딩은 바코딩된 표적의 인덱싱된 라이브러리의 생성을 포함할 수 있다. 바코딩된 표적의 인덱싱된 라이브러리의 생성은 복수의 바코드(예를 들어, 확률적 바코드)를 포함하는 고체 지지체로 수행될 수 있다.
- [0177] 샘플과 바코드의 접촉
- [0178] 본 발명은 샘플(예를 들어, 세포)을 본 발명의 기관과 접촉시키기 위한 방법을 제공한다. 예를 들어, 세포, 기관 또는 조직의 얇은 섹션을 포함하는 샘플을 바코드(예를 들어, 확률적 바코드)와 접촉시킬 수 있다. 세포를 예를 들어, 중력 유동에 의해 접촉시킬 수 있으며, 여기서, 세포는 침강하고 단층을 생성할 수 있다. 샘플은 조직 얇은 섹션일 수 있다. 얇은 섹션은 기관 상에 배치될 수 있다. 샘플은 1-차원일 수 있다(예를 들어, 평면을 형성한다). 샘플(예를 들어, 세포)은 예를 들어, 기관 상에서 세포를 성장/배양함으로써 기관의 도처에 확산될 수 있다.
- [0179] 바코드가 표적과 매우 근접한 경우, 표적은 바코드에 혼성화할 수 있다. 바코드는 각각의 별개의 표적이 본 발명의 별개의 바코드와 연관될 수 있도록 비-고갈 가능한 비로 접촉될 수 있다. 표적과 바코드 사이의 효율적인 연관을 보장하기 위하여, 표적은 바코드에 가교결합될 수 있다.
- [0180] 세포 용해
- [0181] 세포 및 바코드의 분포 후에, 세포를 용해시켜, 표적 분자를 유리시킬 수 있다. 세포 용해는 다양한 수단 중 임의의 것에 의해, 예를 들어 화학적 또는 생화학적 수단에 의해, 삼투압 충격에 의해, 또는 열 용해, 기계적 용해 또는 광학적 용해에 의해 달성될 수 있다. 세포는 세제(예를 들어, SDS, Li 도데실 설페이트, 트리톤 X-100, 트윈-20 또는 NP-40), 유기 용매(예를 들어, 메탄올 또는 아세톤), 또는 분해 효소(예를 들어, 프로테아제 K, 펩신, 또는 트립신), 또는 그의 임의의 조합을 포함하는 세포 용해 완충액의 첨가에 의해 용해될 수 있다. 표적과 바코드의 연관을 증가시키기 위해, 표적 분자의 확산 속도는, 예를 들어, 온도를 감소시키고/거나, 용해물의 점도를 증가시킴으로써, 변경될 수 있다.
- [0182] 일부 실시형태에서, 샘플은 여과지를 사용하여 용해될 수 있다. 여과지는 여과지의 상측에서 용해 완충액으로 함침될 수 있다. 여과지는 샘플의 용해 및 기관에 대한 샘플의 표적의 혼성화를 용이하게 할 수 있는 압력으로 샘플에 적용될 수 있다.
- [0183] 일부 실시형태에서, 용해는 기계적 용해, 열 용해, 광학적 용해 및/또는 화학적 용해에 의해 수행될 수 있다. 화학적 용해는 소화 효소, 예컨대 프로테아제 K, 펩신 및 트립신의 이용을 포함할 수 있다. 용해는 기관에 용해 완충액을 첨가함으로써 수행될 수 있다. 용해 완충액은 트리스 HCl을 포함할 수 있다. 용해 완충액은 적어도 약 0.01, 0.05, 0.1, 0.5 또는 1M 이상의 트리스 HCl을 포함할 수 있다. 용해 완충액은 최대 약 0.01, 0.05, 0.1, 0.5 또는 1M 이상의 트리스 HCL을 포함할 수 있다. 용해 완충액은 약 0.1M의 트리스 HCl을 포함할 수 있다. 용해 완충액의 pH는 적어도 약 1, 2, 3, 4, 5, 6, 7, 8, 9 또는 10 이상일 수 있다. 용해 완충액의 pH는 최대 약 1, 2, 3, 4, 5, 6, 7, 8, 9 또는 10 이상일 수 있다. 일부 실시형태에서, 용해 완충액의 pH는 약 7.5이다. 용해 완충액은 염(예를 들어, LiCl)을 포함할 수 있다. 용해 완충액 중 염의 농도는 적어도 약 0.1, 0.5 또는 1M 이상일 수 있다. 용해 완충액 중 염의 농도는 최대 약 0.1, 0.5 또는 1 M 이상일 수 있다. 일부 실시형태에서, 용해 완충액 중 염의 농도는 약 0.5M이다. 용해 완충액은 세제(예를 들어, SDS, Li 도데실 설페이트, 트리톤 X, 트윈(tween), NP-40)를 포함할 수 있다. 용해 완충액 중 세제의 농도는 적어도 약 0.0001%, 0.0005%, 0.001%, 0.005%, 0.01%, 0.05%, 0.1%, 0.5%, 1%, 2%, 3%, 4%, 5%, 6% 또는 7% 이상일 수 있다. 용해 완충액 중 세제의 농도는 최대 약 0.0001%, 0.0005%, 0.001%, 0.005%, 0.01%, 0.05%, 0.1%, 0.5%, 1%, 2%, 3%, 4%, 5%, 6% 또는 7% 이상일 수 있다. 일부 실시형태에서, 용해 완충액 중 세제의 농도는 약 1% Li 도데실 설페이트이다. 용해 방법에 사용되는 시간은 사용되는 세제의 양에 좌우될 수 있다. 일부 실시형태에서, 사용되는 세제가 더 많을수록, 용해에 필요한 시간은 더 적다. 용해 완충액은 킬레이트화제(예를 들어, EDTA, EGTA)를 포함할 수 있다. 용해 완충액 중의 킬레이트화제의 농도는 적어도 약 1, 5, 10, 15, 20, 25 또는 30 mM 이상일 수 있다. 용해 완충액 중 킬레이트화제의 농도는 최대 약 1, 5, 10, 15, 20, 25 또는 30 mM 이상일 수 있다. 일부 실시형태에서, 용해 완충액 중 킬레이트화제의 농도는 약 10 mM이다. 용해 완충액은 환원제(예를 들어, 베타-머캅토에탄올, DTT)를 포함할 수 있다. 용해 완충액 중 환원제의 농도는 적어도 약 1, 5, 10, 15 또는 20 mM 이상일 수 있다. 용해 완충액 중 환원제의 농도는 최대 약 1, 5, 10, 15 또는 20 mM 이상일 수 있다. 일부 실시형태에서, 용해 완충액 중 환원제의 농도는 약 5 mM이다. 일부 실시형태에서, 용해 완충액은 약 0.1 M 트리스 HCl,

약 pH 7.5, 약 0.5 M LiCl, 약 1% 리튬 도데실 설페이트, 약 10 mM EDTA 및 약 5 mM DTT를 포함할 수 있다.

[0184] 용해는 약 4, 10, 15, 20, 25 또는 30℃의 온도에서 수행될 수 있다. 용해는 약 1, 5, 10, 15 또는 20분 이상 동안 수행될 수 있다. 용해된 세포는 적어도 약 100000, 200000, 300000, 400000, 500000, 600000 또는 700000 개 이상의 표적 핵산 분자를 포함할 수 있다. 용해된 세포는 최대 약 100000, 200000, 300000, 400000, 500000, 600000 또는 700000개 이상의 표적 핵산 분자를 포함할 수 있다.

[0185] 표적 핵산 분자의 바코드의 부착

[0186] 세포의 용해 및 그로부터의 핵산 분자의 방출 후에, 핵산 분자는 공동-위치된 고체 지지체의 바코드와 무작위로 연관될 수 있다. 연관된 표적 핵산 분자의 상보성 부분으로의 바코드의 표적 인식 영역의 혼성화를 포함할 수 있다(예를 들어, 바코드의 올리고(dT)가 표적의 폴리(A) 테일과 상호작용할 수 있다). 혼성화에 사용되는 검정 조건(예를 들어, 완충액 pH, 이온 강도, 온도 등)은 특이적인 안정한 혼성물의 형성을 촉진하도록 선택될 수 있다. 일부 실시형태에서, 용해된 세포로부터 방출되는 핵산 분자는 기관 상의 복수의 프로브와 연관(예를 들어, 기관 상의 프로브와 혼성화)될 수 있다. 프로브가 올리고(dT)를 포함하는 경우, mRNA 분자는 프로브에 혼성화되고 역전사될 수 있다. 올리고뉴클레오타이드의 올리고(dT) 부분은 cDNA 분자의 제1 가닥 합성을 위한 프라이머로서 작용할 수 있다. 예를 들어, 도 2에 예시된 비제한적인 바코딩의 예에서, 블록 216에서, mRNA 분자는 비드 상의 바코드에 혼성화할 수 있다. 예를 들어, 단일-가닥 뉴클레오타이드 단편은 바코드의 표적-결합 영역에 혼성화할 수 있다.

[0187] 부착은 바코드의 표적 인식 영역과 표적 핵산 분자의 일부의 라이게이션을 추가로 포함할 수 있다. 예를 들어, 표적 결합 영역은, 제한 부위 오버행(예를 들어, EcoRI 점착성-말단 오버행)에 특이적으로 혼성화할 수 있는 핵산 서열을 포함할 수 있다. 검정 절차는 제한 효소(예를 들어, EcoRI)로 표적 핵산을 처리하여, 제한 부위 오버행을 생성하는 단계를 추가로 포함할 수 있다. 이어서, 바코드는 제한 부위 오버행에 상보적인 서열을 포함하는 임의의 핵산 분자에 라이게이션될 수 있다. 리가아제(예를 들어, T4 DNA 리가아제)를 사용하여, 2개의 단편을 결합시킬 수 있다.

[0188] 예를 들어, 도 2에 예시된 비제한적인 바코딩의 예에서, 블록 220에서, 복수의 세포(또는 복수의 샘플)로부터의 표지된 표적(예를 들어, 표적-바코드 분자)이 이후에 예를 들어, 튜브 내로 풀링될 수 있다. 표지된 표적은, 예를 들어, 표적-바코드 분자가 부착된 비드 및/또는 바코드를 회수함으로써, 풀링될 수 있다.

[0189] 부착된 표적-바코드 분자의 고체 지지체-기반의 집합물의 회수는 자성 비드 및 외부 인가 자기장의 사용에 의해 구현될 수 있다. 표적-바코드 분자가 풀링되면, 모든 추가의 처리는 단일 반응 용기에서 진행될 수 있다. 추가 처리는, 예를 들어 역전사 반응, 증폭 반응, 절단 반응, 해리 반응 및/또는 핵산 연장 반응을 포함할 수 있다. 추가 처리 반응은 마이크로웰 내에서, 즉, 복수의 세포로부터 표지된 표적 핵산 분자를 먼저 풀링하지 않고, 수행될 수 있다.

[0190] 역전사

[0191] 본 발명은 역전사를 사용하여 표적-바코드 컨주게이트를 생성하는 방법을 제공한다(예를 들어, 도 2의 블록 224에서). 표적-바코드 컨주게이트는 바코드 및 표적 핵산의 전부 또는 일부의 상보성 서열을 포함할 수 있다(즉, 바코딩된 cDNA 분자, 예컨대 확률적으로 바코딩된 cDNA 분자). 연관된 RNA 분자의 역전사는 역전사 효소와 함께 역전사 프라이머의 첨가에 의해 일어날 수 있다. 역전사 프라이머는 올리고(dT) 프라이머, 무작위 핵사뉴클레오타이드 프라이머 또는 표적-특이적 올리고뉴클레오타이드 프라이머일 수 있다. 올리고(dT) 프라이머는 12 내지 18개 또는 약 12 내지 18개의 뉴클레오타이드 길이일 수 있으며, 포유류 mRNA의 3' 말단에서 내인성 폴리(A) 테일에 결합할 수 있다. 무작위 핵사뉴클레오타이드 프라이머는 다양한 상보성 부위에서 mRNA에 결합할 수 있다. 표적-특이적 올리고뉴클레오타이드 프라이머는 전형적으로 관심 mRNA를 선택적으로 프라이밍한다.

[0192] 일부 실시형태에서, 표지된-RNA 분자의 역전사는 역전사 프라이머의 첨가에 의해 발생할 수 있다. 일부 실시형태에서, 역전사 프라이머는 올리고(dT) 프라이머, 무작위 핵사뉴클레오타이드 프라이머 또는 표적-특이적 올리고뉴클레오타이드 프라이머이다. 일반적으로, 올리고(dT) 프라이머는 12 내지 18개 뉴클레오타이드 길이이며, 포유동물 mRNA의 3' 말단의 내인성 폴리(A)+ 테일에 결합한다. 무작위 핵사뉴클레오타이드 프라이머는 다양한 상보성 부위에서 mRNA에 결합할 수 있다. 표적-특이적 올리고뉴클레오타이드 프라이머는 전형적으로 관심 mRNA를 선택적으로 프라이밍한다.

[0193] 역전사는 반복적으로 발생하여, 다중의 표지된-cDNA 분자를 생성할 수 있다. 본원에 개시된 방법은 적어도 약 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 또는 20회의 역전사 반응을 행하는 것을

포함할 수 있다. 상기 방법은 적어도 약 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 또는 100회의 역전사 반응을 행하는 것을 포함할 수 있다.

[0194] 증폭

[0195] (예를 들어, 도 2의 블록 228에서) 1회 이상의 핵산 증폭 반응을 수행하여, 표지된 표적 핵산 분자의 다수의 카피를 생성할 수 있다. 증폭은 다중화된 방식으로 수행될 수 있으며, 다중 표적 핵산 서열은 동시에 증폭된다. 증폭 반응을 사용하여 핵산 분자에 시퀀싱 어댑터를 부가할 수 있다. 증폭 반응은 존재한다면 샘플 표지의 적어도 일부를 증폭시키는 것을 포함할 수 있다. 증폭 반응은 세포 표지 및/또는 바코드 서열(예를 들어, 분자 표지)의 적어도 일부를 증폭시키는 것을 포함할 수 있다. 증폭 반응은 샘플 태그, 세포 표지, 공간 표지, 바코드(예를 들어, 분자 표지), 표적 핵산 또는 그의 조합의 적어도 일부를 증폭시키는 것을 포함할 수 있다. 증폭 반응은 복수의 핵산의 0.5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 100% 또는 이들 값 중 임의의 2개 사이의 범위 또는 수를 증폭시키는 것을 포함할 수 있다. 상기 방법은 하나 이상의 cDNA 합성 반응을 행하여, 샘플 표지, 세포 표지, 공간 표지 및/또는 바코드 서열(예를 들어, 분자 표지)을 포함하는 표적-바코드 분자의 하나 이상의 cDNA 카피를 생성하는 것을 추가로 포함할 수 있다.

[0196] 일부 실시형태에서, 증폭은 증합효소 연쇄 반응(PCR)을 사용하여 수행될 수 있다. 본원에 사용되는 바와 같이, PCR은 DNA의 상보성 가닥의 동시의 프라이머 연장에 의한 특이적인 DNA 서열의 시험관내 증폭을 위한 반응을 지칭할 수 있다. 본원에 사용되는 바와 같이, PCR은 RT-PCR, 리얼-타임(real-time) PCR, 네스티드 PCR, 정량적 PCR, 다중화 PCR, 디지털 PCR 및 어셈블리 PCR을 포함하나 이에 제한되지 않는 유도 형태의 반응을 포함할 수 있다.

[0197] 표지된 핵산의 증폭은 비-PCR 기반의 방법을 포함할 수 있다. 비-PCR 기반의 방법의 예에는 다중 변위 증폭(MDA), 전사-매개의 증폭(TMA), 핵산 서열-기반의 증폭(NASBA), 가닥 변위 증폭(SDA), 리얼-타임 SDA, 회전환 증폭 또는 환-대-환(circle-to-circle) 증폭이 포함되나 이에 제한되지 않는다. 다른 비-PCR-기반의 증폭 방법은 DNA 또는 RNA 표적을 증폭시키기 위한 다수의 사이클의 DNA-의존성 RNA 증합효소-유도 RNA 전사 증폭 또는 RNA-유도된 DNA 합성 및 전사, 리가제 연쇄 반응(LCR) 및 Q β 복제효소(Q β) 방법, 회문식 프로브의 이용, 가닥 변위 증폭, 제한 엔도뉴클레아제를 사용한 올리고뉴클레오티드-유도 증폭, 프라이머가 핵산 서열에 혼성화되고, 생성된 듀플렉스가 연장 반응 및 증폭 이전에 절단되는 증폭 방법, 5' 엑소뉴클레아제 활성이 결여된 핵산 증합효소를 사용한 가닥 변위 증폭, 회전환 증폭 및 분기 연장 증폭(ramification extension amplification; RAM)을 포함한다. 일부 실시형태에서, 증폭은 고리화된 전사물을 생성하지 않는다.

[0198] 일부 실시형태에서, 본원에 개시된 방법은 표지된 핵산(예를 들어, 표지된-RNA, 표지된-DNA, 표지된-cDNA) 상에서 증합효소 연쇄 반응을 행하여, 표지된-앰플리콘(예를 들어, 확률적으로 표지된-앰플리콘)을 생성하는 것을 추가로 포함한다. 표지된-앰플리콘은 이중-가닥 분자일 수 있다. 이중-가닥 분자는 이중-가닥 RNA 분자, 이중-가닥 DNA 분자 또는 DNA 분자에 혼성화된 RNA 분자를 포함할 수 있다. 이중-가닥 분자의 하나의 또는 둘 모두의 가닥은 샘플 표지, 공간 표지, 세포 표지 및/또는 바코드 서열(예를 들어, 분자 표지)을 포함할 수 있다. 표지된-앰플리콘은 단일-가닥 분자일 수 있다. 단일-가닥 분자는 DNA, RNA 또는 그의 조합을 포함할 수 있다. 본 발명의 핵산은 합성 또는 변경된 핵산을 포함할 수 있다.

[0199] 증폭은 하나 이상의 비-천연 뉴클레오티드의 이용을 포함할 수 있다. 비-천연 뉴클레오티드는 광불안정성 또는 촉발성 뉴클레오티드를 포함할 수 있다. 비-천연 뉴클레오티드의 예는 펩티드 핵산(PNA), 모르폴리노 및 잠금 핵산(LNA), 및 글리콜 핵산(GNA) 및 트레오스 핵산(TNA)을 포함할 수 있지만, 이에 제한되지 않는다. 비-천연 뉴클레오티드는 증폭 반응의 하나 이상의 사이클에 첨가될 수 있다. 비-천연 뉴클레오티드의 첨가는 증폭 반응에서 특정 사이클 또는 시점으로서, 산물을 확인하기 위해, 사용될 수 있다.

[0200] 하나 이상의 증폭 반응의 수행은 하나 이상의 프라이머의 사용을 포함할 수 있다. 하나 이상의 프라이머는 예를 들어, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 또는 15개 이상의 뉴클레오티드를 포함할 수 있다. 하나 이상의 프라이머는 적어도 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 또는 15개 이상의 뉴클레오티드를 포함할 수 있다. 하나 이상의 프라이머는 12 내지 15개 미만의 뉴클레오티드를 포함할 수 있다. 하나 이상의 프라이머는 복수의 표지된 표적(예를 들어, 확률적으로 표지된 표적)의 적어도 일부에 어닐링할 수 있다. 하나 이상의 프라이머는 복수의 표지된 표적의 3' 말단 또는 5' 말단에 어닐링할 수 있다. 하나 이상의 프라이머는 복수의 표지된 표적의 내부 영역에 어닐링할 수 있다. 내부 영역은 복수의 표지된 표적의 3' 말단으로부터 적어도 약 50, 100, 150, 200, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350,

360, 370, 380, 390, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 650, 700, 750, 800, 850, 900 또는 1000개의 뉴클레오타이드일 수 있다. 하나 이상의 프라이머는 고정된 프라이머 패널을 포함할 수 있다. 하나 이상의 프라이머는 적어도 하나 이상의 맞춤형 프라이머를 포함할 수 있다. 하나 이상의 프라이머는 적어도 하나 이상의 제어 프라이머를 포함할 수 있다. 하나 이상의 프라이머는 적어도 하나 이상의 유전자-특이적 프라이머를 포함할 수 있다.

[0201] 하나 이상의 프라이머는 범용 프라이머를 포함할 수 있다. 범용 프라이머는 범용 프라이머 결합 부위에 어닐링할 수 있다. 하나 이상의 맞춤형 프라이머는 제1 샘플 표지, 제2 샘플 표지, 공간 표지, 세포 표지, 바코드 서열(예를 들어, 분자 표지), 표적 또는 그의 임의의 조합에 어닐링할 수 있다. 하나 이상의 프라이머는 범용 프라이머 및 맞춤형 프라이머를 포함할 수 있다. 맞춤형 프라이머는 하나 이상의 표적을 증폭시키도록 설계될 수 있다. 표적은 하나 이상의 샘플 내의 전체 핵산의 하위세트를 포함할 수 있다. 표적은 하나 이상의 샘플 내의 전체 표지된 표적의 하위세트를 포함할 수 있다. 하나 이상의 프라이머는 적어도 96개 이상의 맞춤형 프라이머를 포함할 수 있다. 하나 이상의 프라이머는 적어도 960개 이상의 맞춤형 프라이머를 포함할 수 있다. 하나 이상의 프라이머는 적어도 9600개 이상의 맞춤형 프라이머를 포함할 수 있다. 하나 이상의 맞춤형 프라이머는 2개 이상의 상이한 표지된 핵산에 어닐링할 수 있다. 2개 이상의 상이한 표지된 핵산은 하나 이상의 유전자에 상응할 수 있다.

[0202] 임의의 증폭 체계가 본 발명의 방법에 사용될 수 있다. 예를 들어, 하나의 체계에서, 제1차 PCR은 유전자 특이적 프라이머 및 범용 일루미나(Illumina) 시퀀싱 프라이머 1 서열에 대한 프라이머를 사용하여, 비드에 부착된 분자를 증폭시킬 수 있다. 제2차 PCR은 일루미나 시퀀싱 프라이머 2 서열이 플랭킹된 네스티드 유전자 특이적 프라이머 및 범용 일루미나 시퀀싱 프라이머 1 서열에 대한 프라이머를 사용하여, 제1 PCR 산물을 증폭시킬 수 있다. 제3차 PCR은 P5 및 P7과 샘플 인덱스를 첨가함으로써, PCR 산물을 일루미나 시퀀싱 라이브러리로 전환한다. 150 bp x 2 시퀀싱을 사용한 시퀀싱에 의해, 판독 1에서 세포 표지 및 바코드 서열(예를 들어, 분자 표지), 판독 2에서 유전자 및 인덱스 1 판독에서 샘플 인덱스가 드러날 수 있다.

[0203] 일부 실시형태에서, 핵산은 화학적 절단을 사용하여 기관으로부터 제거될 수 있다. 예를 들어, 핵산에 존재하는 화학 기 또는 변형된 염기를 사용하여 고체 지지체로부터의 그의 제거를 용이하게 할 수 있다. 예를 들어, 효소를 사용하여 기관으로부터 핵산을 제거할 수 있다. 예를 들어, 핵산은 제한 엔도뉴클레아제 분해를 통해 기관으로부터 제거될 수 있다. 예를 들어, dUTP 또는 ddUTP를 함유하는 핵산의 우라실-d-글리코실라제(UDG)로의 처리를 사용하여, 핵산을 기관으로부터 제거할 수 있다. 예를 들어, 핵산은 뉴클레오타이드 절제를 수행하는 효소, 예컨대 염기 절제 수복 효소, 예컨대 무푸린/무피리미딘(apurinic/aprimidinic; AP) 엔도뉴클레아제를 사용하여 기관으로부터 제거될 수 있다. 일부 실시형태에서, 핵산은 광 절단 가능한 기 및 광을 사용하여 기관으로부터 제거될 수 있다. 일부 실시형태에서, 절단 가능한 링커를 사용하여 핵산을 기관으로부터 제거할 수 있다. 예를 들어, 절단 가능한 링커는 비오틴/아비딘, 비오틴/스트렙타아비딘, 비오틴/뉴트라비딘, Ig-단백질 A, 광-불안정성 링커, 산 또는 염기 불안정성 링커 기 또는 압타머 중 적어도 하나를 포함할 수 있다.

[0204] 프로브가 유전자-특이적인 경우, 분자는 프로브에 혼성화하고, 역전사되고/거나 증폭될 수 있다. 일부 실시형태에서, 핵산이 합성된(예를 들어, 역전사된) 후에, 증폭될 수 있다. 증폭은 다중 방식으로 수행될 수 있으며, 다중의 표적 핵산 서열이 동시에 증폭된다. 증폭은 시퀀싱 어댑터를 핵산에 부가할 수 있다.

[0205] 일부 실시형태에서, 증폭은 예를 들어, 브릿지(bridge) 증폭을 사용하여 기관 상에서 수행될 수 있다. cDNA는 기관 상의 올리고(dT) 프로브를 사용하는 브릿지 증폭을 위한 양립성 말단을 생성하기 위해 맞춤형 동종중합체일 수 있다. 브릿지 증폭에서, 주형 핵산의 3' 말단에 상보적인 프라이머는 고체 입자에 공유적으로 부착된 각각의 쌍의 제1 프라이머일 수 있다. 주형 핵산을 함유하는 샘플이 입자와 접촉하고, 단일의 열 사이클이 수행되는 경우에, 주형 분자는 제1 프라이머에 어닐링될 수 있으며, 제1 프라이머는 뉴클레오타이드의 첨가에 의해 정방향으로 신장되어, 주형 분자 및 주형에 상보적인 새로 형성된 DNA 가닥으로 이루어진 듀플렉스 분자를 형성한다. 다음 사이클의 가열 단계에서, 듀플렉스 분자는 변성되어, 입자로부터 주형 분자를 방출시키고, 제1 프라이머를 통해 입자에 부착된 상보적 DNA 가닥을 남길 수 있다. 이어지는 어닐링 및 신장 단계 중 어닐링 단계에서, 상보성 가닥은 제1 프라이머로부터 제거된 위치에서 상보성 가닥의 세그먼트에 상보적인 제2 프라이머에 혼성화할 수 있다. 이러한 혼성화는 상보성 가닥이 제1 및 제2 프라이머 사이에 브릿지를 형성하게 할 수 있으며, 이는 공유 결합에 의해 제1 프라이머에 고정되고 혼성화에 의해 제2 프라이머에 고정된다. 신장 단계에서, 제2 프라이머는 동일한 반응 혼합물에서 뉴클레오타이드의 첨가에 의해 역방향으로 신장되어, 그에 의해, 브릿지가 이중-가닥 브릿지로 전환될 수 있다. 이어서, 다음 사이클이 시작되고, 이중-가닥 브릿지를 변성시켜, 2개의 단일-가닥 핵산 분자를 제공할 수 있으며, 각각은 하나의 말단이 제1 및 제2 프라이머를 통해 입자 표면에 부착되고,

각각의 다른 말단은 부착되지 않는다. 이러한 제2 사이클의 어닐링 및 신장 단계에서, 각각의 가닥은 동일한 입자 상에서 이전에 사용되지 않은 추가의 상보성 프라이머에 혼성화하여, 새로운 단일-가닥 브릿지를 형성할 수 있다. 이제 혼성화된, 이전에 사용되지 않은 2개의 프라이머가 신장되어, 2개의 새로운 브릿지가 이중-가닥 브릿지로 전환된다.

- [0206] 증폭 반응은 복수의 핵산 중 적어도 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97% 또는 100%를 증폭시키는 것을 포함할 수 있다.
- [0207] 표지된 핵산의 증폭은 PCR-기반의 방법 또는 비-PCR 기반의 방법을 포함할 수 있다. 표지된 핵산의 증폭은 표지된 핵산의 지수적 증폭을 포함할 수 있다. 표지된 핵산의 증폭은 표지된 핵산의 선형 증폭을 포함할 수 있다. 증폭은 중합효소 연쇄 반응(PCR)에 의해 수행될 수 있다. PCR은 DNA의 상보성 가닥의 동시의 프라이머 연장에 의한 특이적인 DNA 서열의 시험관내 증폭을 위한 반응을 지칭할 수 있다. PCR은 RT-PCR, 리얼-타임 PCR, 네스티드 PCR, 정량적 PCR, 다중화 PCR, 디지털 PCR, 억제 PCR, 반-억제 PCR 및 어셈블리 PCR를 포함하나 이에 제한되지 않는 유도 형태의 반응을 포함할 수 있다.
- [0208] 일부 실시형태에서, 표지된 핵산의 증폭은 비-PCR 기반의 방법을 포함한다. 비-PCR 기반의 방법의 예는 다중 변위 증폭(MDA), 전사-매개의 증폭(TMA), 핵산 서열-기반의 증폭(NASBA), 가닥 변위 증폭(SDA), 리얼-타임 SDA, 회전환 증폭 또는 환-대-환 증폭을 포함하나 이에 제한되지 않는다. 다른 비-PCR-기반의 증폭 방법은 DNA 또는 RNA 표적을 증폭시키기 위한 다수의 사이클의 DNA-의존성 RNA 중합효소-유도 RNA 전사 증폭 또는 RNA-유도 DNA 합성 및 전사, 리가제 연쇄 반응(LCR) 및 Q β 복제효소(Q β), 회문식 프로브의 이용, 가닥 변위 증폭, 제한 엔도뉴클레아제를 사용한 올리고뉴클레오티드-유도 증폭, 프라이머가 핵산 서열에 혼성화되고, 생성된 듀플렉스가 연장 반응 및 증폭 이전에 절단되는 증폭 방법, 5' 엑소뉴클레아제 활성이 결여된 핵산 중합효소를 사용한 가닥 변위 증폭, 회전환 증폭 및/또는 분기 연장 증폭(RAM)을 포함한다.
- [0209] 일부 실시형태에서, 본원에 개시된 방법은 증폭된 앰플리콘(예를 들어, 표적) 상에서 네스티드 중합효소 연쇄 반응을 행하는 것을 추가로 포함한다. 앰플리콘은 이중-가닥 분자일 수 있다. 이중-가닥 분자는 이중-가닥 RNA 분자, 이중-가닥 DNA 분자 또는 DNA 분자에 혼성화된 RNA 분자를 포함할 수 있다. 이중-가닥 분자의 하나의 또는 둘 모두의 가닥은 샘플 태그 또는 분자 확인자 표지를 포함할 수 있다. 대안적으로, 앰플리콘은 단일-가닥 분자일 수 있다. 단일-가닥 분자는 DNA, RNA 또는 그의 조합을 포함할 수 있다. 본 발명의 핵산은 합성 또는 변경된 핵산을 포함할 수 있다.
- [0210] 일부 실시형태에서, 상기 방법은 표지된 핵산을 반복적으로 증폭시켜, 다중 앰플리콘을 생성하는 단계를 포함한다. 본원에 개시된 방법은 적어도 약 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 또는 20회의 증폭 반응을 행하는 것을 포함할 수 있다. 대안적으로, 상기 방법은 적어도 약 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 또는 100회의 증폭 반응을 행하는 것을 포함한다.
- [0211] 증폭은 하나 이상의 대조군 핵산을 복수의 핵산을 포함하는 하나 이상의 샘플에 첨가하는 것을 추가로 포함할 수 있다. 증폭은 하나 이상의 대조군 핵산을 복수의 핵산에 첨가하는 것을 추가로 포함할 수 있다. 대조군 핵산은 대조군 표지를 포함할 수 있다.
- [0212] 증폭은 하나 이상의 비-천연 뉴클레오티드의 이용을 포함할 수 있다. 비-천연 뉴클레오티드는 광불안정성 및/또는 촉발성 뉴클레오티드를 포함할 수 있다. 비-천연 뉴클레오티드의 예는 펩티드 핵산(PNA), 모르폴리노 및 잠금 핵산(LNA), 및 글리콜 핵산(GNA) 및 트레오스 핵산(TNA)을 포함하지만, 이에 제한되지 않는다. 비-천연 뉴클레오티드는 증폭 반응의 하나 이상의 사이클에 첨가될 수 있다. 비-천연 뉴클레오티드의 첨가는 증폭 반응에서 특정 사이클 또는 시점으로서, 산물을 확인하기 위해, 사용될 수 있다.
- [0213] 하나 이상의 증폭 반응의 수행은 하나 이상의 프라이머의 사용을 포함할 수 있다. 하나 이상의 프라이머는 하나 이상의 올리고뉴클레오티드를 포함할 수 있다. 하나 이상의 올리고뉴클레오티드는 적어도 약 7 내지 9개의 뉴클레오티드를 포함할 수 있다. 하나 이상의 올리고뉴클레오티드는 12 내지 15개 미만의 뉴클레오티드를 포함할 수 있다. 하나 이상의 프라이머는 복수의 표지된 핵산의 적어도 일부에 어닐링될 수 있다. 하나 이상의 프라이머는 복수의 표지된 핵산의 3' 말단 및/또는 5' 말단에 어닐링될 수 있다. 하나 이상의 프라이머는 복수의 표지된 핵산의 내부 영역에 어닐링될 수 있다. 내부 영역은 복수의 표지된 핵산의 3' 말단으로부터 적어도 약 50, 100, 150, 200, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590,

600, 650, 700, 750, 800, 850, 900 또는 1000개 뉴클레오타이드일 수 있다. 하나 이상의 프라이머는 고정된 패널의 프라이머를 포함할 수 있다. 하나 이상의 프라이머는 적어도 하나 이상의 맞춤형 프라이머를 포함할 수 있다. 하나 이상의 프라이머는 적어도 하나 이상의 대조군 프라이머를 포함할 수 있다. 하나 이상의 프라이머는 적어도 하나 이상의 하우스키핑 유전자 프라이머를 포함할 수 있다. 하나 이상의 프라이머는 범용 프라이머를 포함할 수 있다. 범용 프라이머는 범용 프라이머 결합 부위에 어닐링될 수 있다. 하나 이상의 맞춤형 프라이머는 제1 샘플 태그, 제2 샘플 태그, 분자 확인자 표지, 핵산 또는 그의 산물에 어닐링될 수 있다. 하나 이상의 프라이머는 범용 프라이머 및 맞춤형 프라이머를 포함할 수 있다. 맞춤형 프라이머는 하나 이상의 표적 핵산을 증폭시키도록 설계될 수 있다. 표적 핵산은 하나 이상의 샘플 내의 전체 핵산의 하위세트를 포함할 수 있다. 일부 실시형태에서, 프라이머는 본 발명의 어레이에 부착된 프로브이다.

[0214] 일부 실시형태에서, 샘플 내의 복수의 표적의 바코딩(예를 들어, 확률적 바코딩)은 바코딩된 단편의 인덱싱된 라이브러리를 생성하는 것을 추가로 포함한다. 상이한 바코드의 바코드 서열(예를 들어, 상이한 확률적 바코드의 분자 표지)은 서로 상이할 수 있다. 바코딩된 표적(예를 들어, 확률적으로 바코딩된 표적)의 인덱싱된 라이브러리를 생성하는 것은 샘플 내의 복수의 표적으로부터 복수의 인덱싱된 폴리뉴클레오타이드를 생성하는 것을 포함한다. 예를 들어, 제1 인덱싱된 표적 및 제2 인덱싱된 표적을 포함하는 바코딩된 표적의 인덱싱된 라이브러리에 있어서, 제1 인덱싱된 폴리뉴클레오타이드의 표지 영역은 제2 인덱싱된 폴리뉴클레오타이드의 표지 영역과 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50개, 약, 적어도 또는 최대 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50개 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 뉴클레오타이드가 상이할 수 있다. 일부 실시형태에서, 바코딩된 표적의 인덱싱된 라이브러리를 생성하는 것은 복수의 표적, 예를 들어, mRNA 분자와 폴리(T) 영역 및 표지 영역을 포함하는 복수의 올리고뉴클레오타이드를 접촉시키고; 제1 가닥 합성을 역전사효소를 사용하여 행하여, 각각이 cDNA 영역 및 표지 영역을 포함하는 단일-가닥의 표지된 cDNA 분자를 생성하는 것을 포함하며, 여기서, 복수의 표적은 상이한 서열의 적어도 2개의 mRNA 분자를 포함하며, 복수의 올리고뉴클레오타이드는 상이한 서열의 적어도 2개의 올리고뉴클레오타이드를 포함한다. 바코딩된 표적의 인덱싱된 라이브러리를 생성하는 것은 단일-가닥의 표지된 cDNA 분자를 증폭시켜, 이중-가닥의 표지된 cDNA 분자를 생성하고; 이중-가닥의 표지된 cDNA 분자 상에서 네스티드 PCR을 행하여, 표지된 앰플리콘을 생성하는 것을 추가로 포함할 수 있다. 일부 실시형태에서, 상기 방법은 어댑터-표지된 앰플리콘을 생성하는 것을 포함할 수 있다.

[0215] 확률적 바코딩은 핵산 바코드 또는 태그를 사용하여 개별 핵산(예를 들어, DNA 또는 RNA) 분자를 표지할 수 있다. 일부 실시형태에서, 그것은 그들이 mRNA로부터 생성되기 때문에, DNA 바코드 또는 태그를 cDNA 분자에 부가하는 것을 포함한다. 네스티드 PCR을 수행하여, PCR 증폭 편향을 최소화시킬 수 있다. 어댑터는 예를 들어, 차세대 시퀀싱(NGS)을 사용한 시퀀싱을 위해 부가될 수 있다. 시퀀싱 결과를 사용하여, 예를 들어, 도 2의 블록 232에서 세포 표지, 바코드 서열(예를 들어, 분자 표지) 및 표적의 하나 이상의 카피의 뉴클레오타이드 단편의 서열을 결정할 수 있다.

[0216] 도 3은 바코딩된 표적(예를 들어, 확률적으로 바코딩된 표적), 예를 들어, mRNA의 인덱싱된 라이브러리를 생성하는 비제한적인 예시적인 과정을 보여주는 개략도이다. 단계 1에 나타난 바와 같이, 역전사 과정은 독특한 바코드 서열(예를 들어, 분자 표지), 세포 표지 및 범용 PCR 부위를 갖는 각각의 mRNA 분자를 인코딩할 수 있다. 예를 들어, RNA 분자(302)를 역전사시켜, RNA 분자(302)의 폴리(A) 테일 영역(308)으로의 바코드(예를 들어, 확률적 바코드)(310)의 세트의 혼성화(예를 들어, 확률적 혼성화)에 의해 cDNA 영역(306)을 포함하는 표지된 cDNA 분자(304)를 생성할 수 있다. 바코드(310)의 각각은 표적-결합 영역, 예를 들어, 폴리(dT) 영역(312), 바코드 서열 또는 분자 표지(314) 및 범용 PCR 영역(316)을 포함할 수 있다.

[0217] 일부 실시형태에서, 세포 표지는 3 내지 20개의 뉴클레오타이드를 포함할 수 있다. 일부 실시형태에서, 바코드 서열(예를 들어, 분자 표지)은 3 내지 20개의 뉴클레오타이드를 포함할 수 있다. 일부 실시형태에서, 복수의 확률적 바코드의 각각은 범용 표지 및 세포 표지 중 하나 이상을 추가로 포함하며, 범용 표지는 고체 지지체 상의 복수의 확률적 바코드에 대하여 동일하며, 세포 표지는 고체 지지체 상의 복수의 확률적 바코드에 대하여 동일하다. 일부 실시형태에서, 범용 표지는 3 내지 20개의 뉴클레오타이드를 포함할 수 있다. 일부 실시형태에서, 세포 표지는 3 내지 20개의 뉴클레오타이드를 포함한다.

[0218] 일부 실시형태에서, 표지 영역(314)은 바코드 서열 또는 분자 표지(318) 및 세포 표지(320)를 포함할 수 있다. 일부 실시형태에서, 표지 영역(314)은 범용 표지, 차원 표지 및 세포 표지 중 하나 이상을 포함할 수 있다. 바코드 서열 또는 분자 표지(318)는 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100개, 약, 적어도 또는 최대 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100개 또는 이들 값 중 임의의 것 사이의 수 또는 범위의 뉴클레오타이드 길이일 수 있다. 세포 표지(320)는 1, 2, 3, 4, 5, 6, 7, 8,

9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100개, 약, 적어도 또는 최대 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100개, 또는 이들 값 중 임의의 것 사이의 수 또는 범위의 뉴클레오타이드 길이일 수 있다. 범용 표지는 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100개, 약, 적어도 또는 최대 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100개 또는 이들 값 중 임의의 것 사이의 수 또는 범위의 뉴클레오타이드 길이일 수 있다. 범용 표지는 고체 지지체 상의 복수의 확률적 바코드에 대하여 동일할 수 있으며, 세포 표지는 고체 지지체 상의 복수의 확률적 바코드에 대하여 동일하다. 차원 표지는 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100개, 약, 적어도 또는 최대 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100개, 또는 이들 값 중 임의의 것 사이의 수 또는 범위의 뉴클레오타이드 길이일 수 있다.

[0219] 일부 실시형태에서, 표지 영역(314)은 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000개, 약, 적어도 또는 최대 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000개, 또는 이들 값 중 임의의 것 사이의 수 또는 범위의 상이한 표지, 예컨대 바코드 서열 또는 분자 표지(318) 및 세포 표지(320)를 포함할 수 있다. 각각의 표지는 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100개, 약, 적어도 또는 최대 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100개, 또는 이들 값 중 임의의 것 사이의 수 또는 범위의 뉴클레오타이드 길이일 수 있다. 바코드 또는 확률적 바코드(310)의 세트는 10, 20, 40, 50, 70, 80, 90, 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 , 10^{10} , 10^{11} , 10^{12} , 10^{13} , 10^{14} , 10^{15} , 10^{20} 개, 약, 적어도 또는 최대 10, 20, 40, 50, 70, 80, 90, 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 , 10^{10} , 10^{11} , 10^{12} , 10^{13} , 10^{14} , 10^{15} , 10^{20} 개, 또는 이들 값 중 임의의 것 사이의 수 또는 범위의 바코드 또는 확률적 바코드(310)를 함유할 수 있다. 그리고 바코드 또는 확률적 바코드(310)의 세트는 예를 들어, 각각 독특한 표지 영역(314)을 함유할 수 있다. 표지된 cDNA 분자(304)를 정제하여, 과잉의 바코드 또는 확률적 바코드(310)를 제거할 수 있다. 정제는 앰퓨어(Ampure) 비드 정제를 포함할 수 있다.

[0220] 단계 2에서 나타낸 바와 같이, 단계 1에서 역전사 과정으로부터의 산물은 1개 튜브 내로 풀링될 수 있으며, 제1 PCR 프라이머 풀 및 제1 범용 PCR 프라이머로 PCR 증폭될 수 있다. 풀링은 독특한 표지 영역(314) 때문에 가능하다. 특히, 표지된 cDNA 분자(304)를 증폭시켜, 네스티드 PCR 표지된 앰플리콘(322)을 생성할 수 있다. 증폭은 다중 PCR 증폭을 포함할 수 있다. 증폭은 단일의 반응 부피에서 96가지의 다중 프라이머를 사용한 다중 PCR 증폭을 포함할 수 있다. 일부 실시형태에서, 다중 PCR 증폭은 단일 반응 부피에서 10, 20, 40, 50, 70, 80, 90, 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 , 10^{10} , 10^{11} , 10^{12} , 10^{13} , 10^{14} , 10^{15} , 10^{20} 가지, 약, 적어도 또는 최대 10, 20, 40, 50, 70, 80, 90, 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 , 10^{10} , 10^{11} , 10^{12} , 10^{13} , 10^{14} , 10^{15} , 10^{20} 가지 또는 이들 값 중 임의의 것 사이의 수 또는 범위의 다중 프라이머를 사용할 수 있다. 증폭은 특이적인 유전자를 표적화하는 맞춤형 프라이머(326A-C)의 제1 PCR 프라이머 풀(324) 및 범용 프라이머(328)를 포함할 수 있다. 맞춤형 프라이머(326)는 표지된 cDNA 분자(304)의 cDNA 부분(306') 내의 영역에 혼성화할 수 있다. 범용 프라이머(328)는 표지된 cDNA 분자(304)의 범용 PCR 영역(316)에 혼성화할 수 있다.

[0221] 도 3의 단계 3에 나타낸 바와 같이, 단계 2에서 PCR 증폭으로부터의 산물을 네스티드 PCR 프라이머 풀 및 제2 범용 PCR 프라이머로 증폭시킬 수 있다. 네스티드 PCR은 PCR 증폭 편향을 최소화시킬 수 있다. 예를 들어, 네스티드 PCR 표지된 앰플리콘(322)을 네스티드 PCR에 의해 추가로 증폭시킬 수 있다. 네스티드 PCR은 단일의 반응 부피에서 네스티드 PCR 프라이머(332a-c)의 네스티드 PCR 프라이머 풀(330) 및 제2 범용 PCR 프라이머(328')를 사용하는 다중 PCR을 포함할 수 있다. 네스티드 PCR 프라이머 풀(328)은 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000가지, 약, 적어도 또는 최대 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000가지 또는 이들 값 중 임의의 것 사이의 수 또는 범위의 상이한 네스티드 PCR 프라이머(330)를 함유할 수 있다. 네스티드 PCR 프라이머(332)는 어댑터(334)를 함유하고, 표지된 앰플리콘(322)의 cDNA 부분(306') 내의 영역에 혼성화할 수 있다. 범용 프라이머(328')는 어댑터(336)를 함유하고, 표지된 앰플리콘(322)의 범용 PCR 영역(316)에 혼성화할 수 있다. 따라서, 단계 3은 어댑터-표지된 앰플리콘(338)을 생성한다. 일부 실시형태에서, 네스티드 PCR 프라이머(332) 및 제2 범용 PCR 프라이머(328')는 어댑터(334 및 336)를 함유하지 않을 수 있다. 대신에 어댑터(334 및 336)를 네스티드 PCR의 산물에 라이게이션시켜, 어댑터-표지된 앰플리콘(338)을 생성할 수 있다.

- [0222] 단계 4에 나타난 바와 같이, 단계 3으로부터의 PCR 산물은 라이브러리 증폭 프라이머를 사용하여 시퀀싱을 위해 PCR 증폭될 수 있다. 특히, 어댑터(334 및 336)를 사용하여, 어댑터-표지된 앰플리콘(338) 상에서 하나 이상의 추가의 검정을 행할 수 있다. 어댑터(334 및 336)는 프라이머(340 및 342)에 혼성화될 수 있다. 하나 이상의 프라이머(340 및 342)는 PCR 증폭 프라이머일 수 있다. 하나 이상의 프라이머(340 및 342)는 시퀀싱 프라이머일 수 있다. 하나 이상의 어댑터(334 및 336)는 어댑터-표지된 앰플리콘(338)의 추가의 증폭을 위해 사용될 수 있다. 하나 이상의 어댑터(334 및 336)는 어댑터-표지된 앰플리콘(338)의 시퀀싱을 위해 사용될 수 있다. 프라이머(342)는 플레이트 인덱스(344)를 포함하여, 동일한 세트의 바코드 또는 확률적 바코드(310)를 사용하여 생성되는 앰플리콘이 하나의 시퀀싱 반응에서 차세대 시퀀싱(NGS)을 사용하여 시퀀싱될 수 있게 할 수 있다.
- [0223] 덴드로그램을 사용한 발현 프로파일의 클러스터링
- [0224] 세포 유형을 구별하기 위한 표적의 확인을 위한 방법이 본원에 개시된다. 일부 실시형태에서, 상기 방법은 (a) 표적 계수 데이터 구조를 수신하는 단계로서, 표적 계수 데이터 구조가 복수의 세포의 발현 프로파일을 포함하며, 복수의 세포의 발현 프로파일이 복수의 세포 중 각각의 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함하는 단계; (b) 표적 계수 데이터 구조 및 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 복수의 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하는 단계로서, 덴드로그램이 복수의 노드를 포함하며, 복수의 노드가 루트 노드, 복수의 리프 노드 및 복수의 비-루트, 비-리프 노드를 포함하며, 복수의 리프 노드 중 각각의 리프 노드가 복수의 세포 중 상이한 세포의 발현 프로파일을 나타내며, 루트 노드가 복수의 세포의 발현 프로파일을 나타내는 단계; (c) 덴드로그램의 루트 노드로부터 덴드로그램의 복수의 리프 노드로 덴드로그램의 복수의 노드 중 각각의 노드를 통해 트래버싱하는 동안: (1) 노드의 자식 노드로의 노드의 분할이 유효한지 무효한(예를 들어, 자식 노드 사이의 차이가 유의미하지 않은)지를 결정하고; (2) 노드의 자식 노드로의 노드의 분할이 무효하면, 노드를 병합 클러스터 세트에 부가하는 단계; (d) 반복적으로, 병합 클러스터 세트 내의 각각의 제1 노드에 있어서, 병합 클러스터 세트 내의 제1 노드와 제1 노드에 가장 가까운 병합 클러스터 세트 내의 제2 노드 사이의 거리가 병합 거리 임계값 이내이면, 제1 노드와 제2 노드를 병합하여, 제1 노드 및 제2 노드에 의해 나타나는 발현 프로파일을 포함하는 병합된 노드를 생성하는 단계; 및 (e) 병합 클러스터 세트 내의 각각의 노드에 있어서, 노드에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계를 포함한다.
- [0225] 도 4는 세포 유형을 구별하기 위한 표적의 비제한적인 예시적 확인 방법(400)을 보여주는 흐름도이다. 상기 방법(400)은 분자 계수 데이터 구조(예를 들어, 분자 계수 행렬)를 클러스터 표지의 세트 및 중요한 유전자의 세트로 맵핑한다. 일부 실시형태에서, 입력은 분자 계수의 $N \times M$ 행렬 또는 i, j 차 엔트리가 세포 i 로부터의 판독을 사용하여 관찰되는 유전자 j 에 대한 분자의 수를 설명하는 어레이일 수 있다. 알고리즘은 2가지 출력을 생성할 수 있다. 제1 출력은 N 개의 표지의 세트일 수 있으며, 하나는 각 세포에 대한 것이다(동일한 표지의 세포는 "유사한" 것으로 추론될 수 있다). 제2 출력은 클러스터를 서로 구별하기 위해 사용될 수 있는 유전자의 세트일 수 있다.
- [0226] 상기 방법(400)은 덴드로그램 분할, 시험 및 병합 방법을 사용하여 2가지 출력을 생성한다. 데이터의 전처리 및 거리 데이터 구조(예를 들어, 거리 행렬) D 의 생성 후에, 알고리즘은 D 를 계층적으로 클러스터링하여, 덴드로그램을 생성할 수 있다. 알고리즘은 2가지 단계를 포함할 수 있다. 상기 방법(400)은 분할 및 시험 단계 동안, 덴드로그램의 상측(예를 들어, 도 5에서 트리 또는 덴드로그램(500)의 루트 노드(504))에서 시작한다. 덴드로그램(500)은 루트 노드(504), 복수의 리프 노드(520a-520h) 및 복수의 비-루트, 비-리프 노드(508a-508b, 512a-512b 및 516a-516b)를 포함한다. 덴드로그램의 각 노드(리프 노드, 예컨대 리프 노드(520a-520h) 제외)에서, 트리를 2개의 서브트리로 분할한다(예를 들어, 루트 노드(504)는 2개의 서브트리(508a, 508b)로 분할된다). 분할은 2개의 후보 서브클러스터(예를 들어, 각각은 적어도 하나의 세포의 발현 프로파일을 함유)로 분할되는(예를 들어, 2개 이상의 세포의 발현 프로파일을 함유하는) 클러스터에 상응한다. 분할의 질은 점수화될 수 있다. 서브클러스터가 충분히 상이한 것으로 간주되면, 각 서브트리에서 알고리즘을 계속 시행한다. 그렇지 않으면, 덴드로그램의 이러한 부분에 대하여 알고리즘을 종료한다. 이 단계는 데이터세트에 대하여 표지의 세트를 생성한다. 병합 단계 동안, 방법(400)은 분할 및 시험 단계 동안 생성된 표지를 사용하여, 이들 클러스터 중 임의의 것을 조합하여 하나의 클러스터를 형성하여야 하는지를 결정한다. 일부 실시형태에서, 분할 및 시험 단계는 각각 소수의 샘플의 작은 클러스터를 생성하는 경향이 있다. 병합 단계는 그들을 더 큰 클러스터와 병합함으로써 더 작은 클러스터를 "클린-업(clean-up)"할 수 있다.
- [0227] 블록 404에서, 방법(400)은 분자 계수 데이터 구조(예를 들어, 분자 계수 행렬)를 수신하는 단계를 포함할 수 있다. 행렬은 오직 전체, 비부정(nonnegative) 엔트리를 포함할 수 있으며, 크고 희소한 경향이 있다. 일부 실

시형태에서, 입력은 분자 계수의 $N \times M$ 행렬 또는 i, j 차 엔트리가 세포 i 로부터 관독을 사용하여 관찰되는 유전자 j 에 대한 분자의 수를 설명하는 어레이일 수 있다.

[0228] 블록 408에서, 방법(400)은 분자 계수 데이터 구조를 전처리하여, 거리 데이터 구조(예를 들어, 거리 행렬)를 생성하는 단계를 포함할 수 있다. 일부 실시형태에서, 입력 데이터 구조는 로그-변환된다. 자연 로그를 취하기 전에 값 1을 각 엔트리에 더한다. 상관관계 거리를 사용하여, 각 쌍의 세포 사이의 쌍별 비유사성 d 를 설명할 수 있다. 세포 c_i 및 c_j 에 있어서, 2개의 세포 사이의 상관관계 거리는 식 [1]을 사용하여 결정될 수 있다.

$$d(c_i, c_j) = 1 - \frac{(c_i - \bar{c}_i) \cdot (c_j - \bar{c}_j)}{\|c_i - \bar{c}_i\|_2 \|c_j - \bar{c}_j\|_2}, \quad \text{식 [1]}$$

[0230] 상기 식에서, $\bar{c}_i$ 는 c_i 의 모든 요소의 평균을 나타낸다. 전처리 단계의 출력은 대각선을 따라 0인 거리의 정사각형, 대칭 행렬 D 일 수 있다.

[0231] 블록 412에서, 방법(400)은 세포의 발현 프로파일을 계층적으로 클러스터링하여, 덴드로그램을 생성하는 단계를 포함할 수 있다. 세포의 발현 프로파일을 계층적으로 클러스터링하여, 덴드로그램을 생성하는 단계는 덴드로그램의 가장 가까운 2개의 클러스터를 반복적으로 병합하는 것을 포함할 수 있다. 모든 클러스터는상기와 같이 기재된 쌍별 거리를 갖는 개별 점으로서 개시될 수 있다. 클러스터 사이의 거리 D 의 계산은 완전 연결을 사용하여 행하였다. 클러스터 A 및 B에 있어서, 2개의 클러스터 사이의 거리는 식 [2]를 사용하여 결정될 수 있다:

$$D(A, B) = \max_{a \in A, b \in B} d(a, b). \quad \text{식 [2]}$$

[0233] 완전한 덴드로그램이 이러한 블록에서 수득될 수 있다. 일부 실시형태에서, 클러스터 A의 클러스터-내 상관관계 및 클러스터 B의 클러스터-내 상관관계는 클러스터 A 및 클러스터 B의 클러스터-간 상관관계보다 더 크다. 클러스터 A의 클러스터-내 상관관계 및 클러스터 B의 클러스터-내 상관관계의 척도 또는 표시는 클러스터 A 및 클러스터 B의 클러스터-간 상관관계보다 더 크다. 클러스터 A의 클러스터-내 상관관계 및 클러스터 B의 클러스터-내 상관관계의 척도는 다음 중 적어도 하나에 기초할 수 있다: 클러스터 A 및 클러스터 B의 클러스터-내 최대 상관관계, 클러스터 A 및 클러스터 B의 클러스터-내 평균 상관관계, 클러스터 A 및 클러스터 B의 클러스터-내 중간값 상관관계, 클러스터 A 및 클러스터 B의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 클러스터 A의 클러스터-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 클러스터 A의 클러스터-내 최대 상관관계, 클러스터 A의 클러스터-내 평균 상관관계, 클러스터 A의 클러스터-내 중간값 상관관계, 클러스터 A의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 클러스터 B의 클러스터-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 클러스터 B의 클러스터-내 최대 상관관계, 클러스터 B의 클러스터-내 평균 상관관계, 클러스터 B의 클러스터-내 중간값 상관관계, 클러스터 B의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 클러스터 A 및 클러스터 B의 클러스터-간 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 클러스터 A 및 클러스터 B의 클러스터-간 최대 상관관계, 클러스터 A 및 클러스터 B의 클러스터-간 평균 상관관계, 클러스터 A 및 클러스터 B의 클러스터-간 중간값 상관관계, 클러스터 A 및 클러스터 B의 클러스터-간 최소 상관관계 및 그의 임의의 조합. 예를 들어, 2개의 서브-클러스터의 클러스터-내 중간값 상관관계는 클러스터-간 중간값 상관관계보다 더 클 수 있다.

[0234] 블록 416에서, 방법(400)은 덴드로그램을 분할하고 시험하여, 표지의 세트를 생성하는 단계를 포함할 수 있다. 분할 및 시험은 덴드로그램의 상측에서 개시될 수 있다. 덴드로그램 서브트리 T_0 을 고려해 볼 때, 트리는 정확하게 2개의 서브트리 TL 및 TR 로 분할될 수 있다. 통계적 검정을 수행하여, 좌측 서브트리 TL 내의 세포가 우측 서브트리 TR 내의 세포와 충분히 상이한지를 결정할 수 있다. 일부 실시형태에서, 통계적 검정은 2개의 모집단에 대하여 각 유전자에서 웰치의 t -검정을 수행하는 것을 포함한다. 두 모집단 모두에서 분산이 0인 것으로 추정되면, 무한의 t -통계가 발생할 수 있으며; 이러한 경우는 무시할 수 있다. 모든 검정 중에 최소 p -값이 소정의 임계값(오검출률에 대하여 보수적으로 보정됨)보다 더 낮으면, 분할은 유효한 것으로 간주될 수 있으며, 알고리즘은 2개의 서브트리에서 다시 수행된다. 최소 p -값이 임계값 미만이면, 방법(400)은 서브트리 T_0 에 대하여 종료된다. TL 이 정확히 1가지의 샘플을 함유하면(즉, TL 이 싱글톤(singleton)이면), TL 은 무시할 수 있고, 알고리즘은 TR 로 절차를 반복한다. TR 이 정확히 1가지의 샘플을 함유하면, TR 은 무시할 수 있고, 알고리즘은 TL 로 절차를 반복한다. TL 및 TR 둘 모두가 각각 정확히 1가지의 샘플을 함유하면, 알고리즘은 서브트리 T_0 에 대하여 종료된다.

[0235] 블록 416에서, 방법(400)은 하기와 같이 클러스터 표지를 결정하는 단계를 포함할 수 있다. 먼저, 모든 서브트

리는 'r'로 표지될 수 있다. 분할이 일어나고 p-값 이슈로 인하여 거부되지 않을 때마다, TL에서 세포의 모든 표지는 'L'이 첨부되고, TR에서 세포의 모든 표지는 'R'이 첨부될 수 있다. 이는 싱글톤을 스킵핑하는 경우, 표지가 여전히 영향을 받는 것을 의미한다. 싱글톤은 임의의 다른 데이터 점과 공유되지 않는 독특한 표지를 자동으로 얻는다.

[0236] 블록 416에서, 방법(400)은 각각의 최종 클러스터의 응집성을 결정하는 단계를 포함할 수 있다. 최종 클러스터 내의 모든 샘플이 서로 멀리 떨어져 있으면(즉, 클러스터 내의 쌍별 거리가 하위, 예를 들어, 모든 거리의 50 백분위수에 없으면), 클러스터는 해체될 수 있다. 이어서, 각각의 샘플은 싱글톤으로 표시될 수 있다.

[0237] 블록 420에서, 방법(400)은 블록 416에서 생성된 표지의 세트를 병합하여, 또 다른 세트의 표지를 생성하는 단계를 포함할 수 있다. 일부 실시형태에서, 병합은 2-단계 과정일 수 있다. 제1 단계에서, 각각의 싱글톤은 전처리 블록 408로부터 거리 데이터 구조(예를 들어, 거리 행렬)를 사용하여 결정되는 그의 가장 가까운 이웃과 동일한 클러스터에 배치될 수 있다. 싱글톤의 가장 가까운 이웃에 대한 싱글톤의 거리가 거리의 상위 10% 이내이면(즉, 그것이 모든 다른 세포로부터 멀리 있으면), 싱글톤은 이상점으로 표시될 수 있으며, 그의 자체의 클러스터에 유지된다. 이러한 제1 단계는 모든 클러스터가 적어도 2개의 비-이상점 구성원을 함유하는 것을 보장한다. 제2 단계에서, 이상점을 제거한 후에, 클러스터 사이의 쌍별 거리를 몇몇의 통계적 검정을 사용하여 계산하여, 쌍별 클러스터 거리 행렬 D_c 를 초래한다. 2개의 클러스터 사이의 거리를 모든 유전자에 대한 독립적인 웰치의 t-검정에서 수득된 가장 작은 p-값의 음의 로그로서 추정한다. 전반적인 가장 작은 거리에서 시작하여, 2개의 상응하는 클러스터를 병합한다. 모든 기존의 클러스터에 대한 새로운 클러스터의 거리를 계산하고, 모든 쌍별 거리가 소정의 거리 초과일 때까지 과정을 반복한다. 또한, 패키지는 D_c 상에서 루벤(Louvain) 알고리즘을 시행함으로써 병합하기 위한 커뮤니티-검출-기반의 접근법을 제공한다.

[0238] 블록 424에서, 방법(400)은 블록 420에서 결정된 표지의 세트로부터 특징을 선택하여, 세포 유형을 구별하기 위한 특징을 확인하는 단계를 포함할 수 있다. 일부 실시형태에서, 방법(400)은 병합 블록 420으로부터 생성된 표지를 사용하여 2가지 유형의 특징 선택을 수행할 수 있다. 제1 유형의 특징 선택을 위하여, 블록의 분할 및 시험 동안, 분할이 유지될 때마다, K개의 가장 작은 p-값을 갖는 K개의 유전자가 저장된다. 더 많은 유전자가 분할로부터 덴드로그램의 상측에 더 가깝게 유지될 수 있다. 궁극적으로, 모든 분할의 연합으로부터 독특한 유전자의 목록이 출력된다. 제2 유형의 특징 선택을 위하여, 각각의 클러스터에 있어서, 관심 클러스터에서 더 높은 평균을 갖는 유전자만을 사용하여 몇몇의 1-대-나머지(one-vs-rest) 검정을 행한다. 각각의 클러스터에 대한 중요한 유전자의 표가 각각의 유전자에 관한 몇몇의 추가의 정보(예를 들어, p-값, 배수 변화, 클러스터 내의 평균 발현 수준)와 함께 출력될 수 있다.

[0239] 방법(400)은 탐색적 분석을 수행하는 단계를 포함할 수 있다. 일부 실시형태에서, 방법(400)은 사용될 수 있는 분할 및 병합 단계에서 소정의 단계를 시각화시키기 위하여 몇몇의 함수를 사용할 수 있다. 예를 들어, 이들 함수는 분할(또는 병합)에 관련된 세포, 각 서브트리(또는 조합된 클러스터)에서 끝나는 세포, 및 이러한 분할(또는 병합)을 지시하는 유전자를 보여준다. 또 다른 예로서, 방법(400)은 (예를 들어, 어떤 유전자가 각 쌍의 클러스터를 구별하는지를 결정하기 위하여) 모든 클러스터 중에 쌍별 비교를 수행할 수 있으며, 덴드로그램을 작도하기 위한 기능을 한다. 방법(400)은 클러스터 내의 쌍별 거리의 분포에 기초할 수 있다. 일부 실시형태에서, 방법(400)은 파라미터 스위핑을 수행하는 단계를 포함할 수 있다.

[0240] 발현 프로파일의 클러스터링

[0241] 세포 유형을 구별하기 위한 표적을 확인하기 위한 방법이 본원에 개시된다. 일부 실시형태에서, 상기 방법은 (a) 복수의 세포의 발현 프로파일을 수신하는 단계로서, 발현 프로파일은 복수의 세포 중 각각의 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함하는 단계; (b) 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계로서, 각각의 클러스터가 (1) 부모 클러스터 및 (2) 2개 이상의 자식 클러스터 중 하나 또는 둘 모두와 하나 이상의 연관을 가지며, 부모 클러스터가 클러스터에 의해 나타나는 복수의 세포 중 하나 이상의 세포의 발현 프로파일을 나타내며, 클러스터가 2개 이상의 자식 클러스터에 의해 나타나는 발현 프로파일을 나타내는 단계; (c) 2개 이상의 자식 클러스터를 갖는 각각의 클러스터에 있어서, 클러스터와 2개 이상의 자식 클러스터 사이의 연관이 무효하면(예를 들어, 2개 이상의 자식 클러스터 사이의 차이가 유의미하지 않으면), 클러스터를 병합 클러스터 세트에 추가하는 단계; (d) 반복적으로, 병합 클러스터 세트 내의 각각의 제1 클러스터에 있어서, 병합 클러스터 세트 내의 제1 클러스터와 제1 클러스터에 가장 가까운 병합 클러스터 세트 내의 제2 클러스터 사이의 거리가 병합 거리 임계값 이내이면, 제1 클러스터와 제2 클러스터를 병합하여, 병합된 클러스터를 생성하는 단계로서, 병

합된 클러스터가 제1 클러스터와 제2 클러스터의 발현 프로파일을 포함하는 단계; 및 (e) 병합 클러스터 세트 내의 각각의 클러스터에 있어서, 클러스터에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계를 포함한다.

[0242] 도 6은 세포의 발현 프로파일을 클러스터링함으로써 세포 유형을 구별하기 위한 표적을 확인하는 비제한적인 예시적 방법을 보여주는 흐름도이다. 블록 604에서, 방법(600)은 복수의 세포의 발현 프로파일을 수신한다. 각각의 발현 프로파일은 복수의 세포 중 상이한 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함할 수 있다. 일부 실시형태에서, 복수의 세포의 발현 프로파일을 수신하는 단계는 표적 계수 데이터 구조(예를 들어, 표적 계수 행렬)를 수신하는 것을 포함한다. 표적 계수 행렬의 각각의 행은 복수의 세포 중 하나의 세포의 발현 프로파일을 포함할 수 있다.

[0243] 수신되는 발현 프로파일의 수는 상이한 구현예에서 상이할 수 있다. 일부 실시형태에서, 수신되는 발현 프로파일의 수는 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 또는 약 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 또는 이들 값 중 임의의 2개 사이의 수 또는 범위일 수 있다. 일부 실시형태에서, 수신되는 발현 프로파일의 수는 적어도 또는 최대 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000 또는 10000일 수 있다.

[0244] 일부 실시형태에서, 방법(600)은 블록 604에서 복수의 세포의 발현 프로파일을 수신하는 단계 이전에: 복수의 확률적 바코드를 사용하여 복수의 세포 내의 복수의 표적을 확률적으로 바코딩하여, 복수의 확률적으로 바코딩된 표적을 생성하는 단계로서, 복수의 확률적 바코드의 각각이 세포 표지 및 분자 표지를 포함하며, 상이한 세포의 표적으로부터 생성된 확률적으로 바코딩된 표적이 상이한 세포 표지를 가지며, 복수의 세포 중 하나의 세포의 표적으로부터 생성된 확률적으로 바코딩된 표적이 상이한 분자 표지를 갖는 단계; 복수의 확률적으로 바코딩된 표적의 시퀀싱 데이터를 획득하는 단계; 및 복수의 세포의 각각에 있어서: (1) 세포에 대한 시퀀싱 데이터에서 복수의 표적 중 각각의 표적과 연관된 별개의 서열을 갖는 분자 표지의 수를 계수하는 단계; 및 (2) (1)에서 계수된 시퀀싱 데이터에서의 표적과 연관된 별개의 서열을 갖는 분자 표지의 수에 기초하여 세포에 대한 복수의 표적 중 각각의 표적의 수를 추정하는 단계를 포함한다. 일부 실시형태에서, 복수의 세포 중 세포의 발현 프로파일은 (2)에서 추정된 세포에 대한 복수의 표적 중 각각의 표적의 수를 포함한다.

[0245] 블록 608에서, 방법(600)은 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계를 포함할 수 있다. 방법(600)은 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 발현 프로파일의 클러스터를 생성할 수 있다. 각각의 클러스터에 의해 나타나는 발현 프로파일의 수는 상이한 구현예에서 상이할 수 있다. 일부 실시형태에서, 각각의 클러스터에 의해 나타나는 발현 프로파일의 수는 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000개 또는 약 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000개, 또는 이들 값 중 임의의 2개 사이의 수 또는 범위일 수 있다. 일부 실시형태에서, 각각의 클러스터에 의해 나타나는 발현 프로파일의 수는 적어도 또는 최대 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000 또는 10000개일 수 있다. 일부 실시형태에서, 각각의 클러스터에 의해 나타나는 발현 프로파일은 블록 604에서 수신된 발현 프로파일의 수의 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100% 또는 약 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, 또는 이들 값 중 임의의 2개 사이의 수 또는 범위일 수 있다. 일부 실시형태에서, 각각의 클러스터에 의해 나타나는 발현 프로파일은 블록 604에서 수신되는 발현 프로파일의 수의 적어도 또는 최대 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 또는 100%일 수 있다.

[0246] 각각의 클러스터는 (1) 부모 클러스터 및 (2) 2개 이상의 자식 클러스터(예컨대 3, 4, 5, 6, 7, 8, 9, 10개 이상의 자식 클러스터) 중 하나 또는 둘 모두와 연관을 가질 수 있다. 부모 클러스터는 클러스터에 의해 나타나는 복수의 세포 중 하나 이상의 세포의 발현 프로파일을 나타낸다. 클러스터는 그의 2개 이상의 자식 클러스터에 의해 나타나는 발현 프로파일을 나타낸다.

[0247] 일부 실시형태에서, 발현 프로파일은 도 4, 예컨대 도 4의 블록 412를 참조하여 설명되는 바와 같이 클러스터링

될 수 있다. 예를 들어, 방법(600)은 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 계층적으로 클러스터링하여, 복수의 세포의 발현 프로파일을 나타내는 덴드로그램을 생성하는 단계를 포함할 수 있다. 덴드로그램은 복수의 클러스터를 포함할 수 있다. 복수의 클러스터는 루트 클러스터, 복수의 리프 클러스터 및 복수의 비-루트, 비-리프 클러스터를 포함할 수 있다. 리프 클러스터의 수는 예를 들어, 발현 프로파일의 수 n 과 동일할 수 있다. 비-루트, 비-리프 클러스터의 수는 예를 들어, $n-2$ 일 수 있다.

[0248] 복수의 리프 클러스터 및 복수의 비-루트, 비-리프 클러스터의 각각은 부모 클러스터와 연관될 수 있다. 루트 클러스터 및 복수의 비-루트, 비-리프 클러스터의 각각은 좌측 자식 클러스터 및 우측 자식 클러스터와 연관될 수 있으며, 클러스터의 좌측 자식 클러스터 및 우측 자식 클러스터에 의해 나타나는 발현 프로파일을 나타낸다. 루트 클러스터는 복수의 세포의 발현 프로파일을 나타낼 수 있다. 일부 구현예에서, 리프 클러스터는 하나의 세포의 발현 프로파일을 나타낼 수 있다. 비-루트, 비-리프 클러스터는 비-루트, 비-리프 클러스터의 자식 클러스터에 의해 나타나는 세포의 발현 프로파일을 나타낼 수 있다. 루트 클러스터는 그의 자식 클러스터의 발현 프로파일을 나타낼 수 있다.

[0249] 일부 실시형태에서, 블록 608에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 클러스터링하는 단계는 복수의 세포의 각각의 발현 프로파일을 복수의 클러스터 내의 상이한 리프 클러스터에 할당하고; 제2 클러스터가 복수의 클러스터 중 제1 클러스터에 가장 가까운 클러스터이면, 복수의 클러스터 중 제1 클러스터 및 제2 클러스터를 반복적으로 조합하여, 제1 클러스터 및 제2 클러스터의 부모 클러스터를 생성하는 것을 포함한다. 제1 클러스터와 제2 클러스터 사이의 거리는 제1 클러스터에 의해 나타나는 발현 프로파일을 갖는 임의의 세포와 제2 클러스터에 의해 나타나는 발현 프로파일을 갖는 임의의 세포 사이의 최대 거리일 수 있다.

[0250] 일부 실시형태에서, 제1 클러스터의 클러스터-내 상관관계 및 제2 클러스터의 클러스터-내 상관관계는 제1 클러스터와 제2 클러스터의 클러스터-간 상관관계보다 더 크다. 제1 클러스터의 클러스터-내 상관관계 및 제2 클러스터의 클러스터-내 상관관계의 척도 또는 표시는 제1 클러스터 및 제2 클러스터의 클러스터-간 상관관계보다 더 크다. 제1 클러스터의 클러스터-내 상관관계 및 제2 클러스터의 클러스터-내 상관관계의 척도는 다음 중 적어도 하나에 기초할 수 있다: 제1 클러스터 및 제2 클러스터의 클러스터-내 최대 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-내 평균 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-내 중간값 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 제1 클러스터의 클러스터-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 클러스터의 클러스터-내 최대 상관관계, 제1 클러스터의 클러스터-내 평균 상관관계, 제1 클러스터의 클러스터-내 중간값 상관관계, 제1 클러스터의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 제2 클러스터의 클러스터-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제2 클러스터의 클러스터-내 최대 상관관계, 제2 클러스터의 클러스터-내 평균 상관관계, 제2 클러스터의 클러스터-내 중간값 상관관계, 제2 클러스터의 클러스터-내 최소 상관관계 및 그의 임의의 조합. 제1 클러스터 및 제2 클러스터의 클러스터-간 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 클러스터 및 제2 클러스터의 클러스터-간 최대 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-간 평균 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-간 중간값 상관관계, 제1 클러스터 및 제2 클러스터의 클러스터-간 최소 상관관계 및 그의 임의의 조합.

[0251] 일부 실시형태에서, 방법(600)은 블록 608에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계 이전에, 복수의 세포의 발현 프로파일의 거리 데이터 구조(예를 들어, 거리 행렬)를 결정하는 단계를 포함할 수 있다. 행렬은 $n \times n$ 의 차원을 가질 수 있으며, n 은 블록 604에서 수신되는 발현 프로파일의 수를 나타낸다. 거리 행렬의 각각의 대각선 요소는 0의 값을 갖는다. 블록 608에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여, 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계는 거리 데이터 구조에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 것을 포함할 수 있다. 복수의 세포의 발현 프로파일 사이의 거리는 복수의 세포의 발현 프로파일 사이의 쌍별 상관관계 거리일 수 있다.

[0252] 일부 실시형태에서, 방법(600)은 (i)에서 거리 데이터 구조를 결정하는 단계 이전에, 표적 계수 데이터 구조를 로그-변환된 표적 계수 데이터 구조(예를 들어, 로그-변환된 표적 계수 행렬)로 로그-변환하는 단계를 포함할 수 있다. 표적 계수 데이터 구조의 요소의 거리 데이터 구조를 결정하는 단계는 로그-변환된 표적 계수 데이터 구조의 거리 데이터 구조를 결정하는 것을 포함한다. 블록 608에서 복수의 세포의 발현 프로파일 사이의 거리에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 발현 프로파일의 클러스터를 생성하는 단계는

로그-변환된 표적 계수 데이터 구조 및 거리 데이터 구조에 기초하여 복수의 세포의 발현 프로파일을 클러스터링하여, 복수의 클러스터를 생성하는 것을 포함할 수 있다. 표적 계수 데이터 구조를 로그-변환된 표적 계수 데이터 구조로 로그-변환시키는 단계는 표적 계수 데이터 구조의 각각의 요소의 값을 충분만큼 증가시키는 것을 포함할 수 있다. 증분은 예를 들어, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 이상일 수 있다.

[0253] 블록 612에서, 방법(600)은 클러스터와 그의 자식 클러스터 사이의 연관이 무효하면(예를 들어, 자식 클러스터 사이의 차이가 유의미하지 않으면), 2개 이상의 자식 클러스터를 갖는 발현 프로파일의 클러스터 중 각각의 클러스터를 병합 클러스터 세트에 부가하는 단계를 포함할 수 있다. 일부 실시형태에서, 블록 608에서 발현 프로파일을 클러스터링하여, 덴드로그램을 생성한다면, 방법(600)은 도 4, 예컨대 도 4의 블록 416을 참조하여 기재된 바와 같이 덴드로그램을 분할하고 시험하여, 표지의 세트를 생성함으로써 2개 이상의 자식 클러스터를 갖는 각각의 클러스터를 병합 클러스터 세트에 부가할 수 있다.

[0254] 일부 실시형태에서, 2개 이상의 자식 클러스터를 갖는 각각의 클러스터에 있어서, 클러스터와 2개 이상의 자식 클러스터 사이의 연관이 무효하면, 방법(600)은 덴드로그램의 루트 클러스터로부터 덴드로그램의 복수의 리프 클러스터로 덴드로그램의 각각의 클러스터를 통해 트래버싱하는 동안 (1) 클러스터와 클러스터의 자식 클러스터 사이의 연관이 유효한지 무효한지를 결정하고; (2) 연관이 무효하면, 클러스터를 병합 클러스터 세트에 부가함으로써 클러스터를 병합 클러스터 세트에 부가할 수 있다.

[0255] 블록 616에서, 방법(600)은 2개의 클러스터 사이의 거리가 병합 거리 임계값 이내이면, 병합 클러스터 세트 내의 각각의 클러스터를 그의 가장 가까운 병합 클러스터 세트 내의 클러스터와 병합하는 단계를 포함할 수 있다. 병합된 클러스터는 제1 클러스터 및 제2 클러스터의 발현 프로파일을 포함한다. 방법(600)은 도 4, 예컨대 도 4의 블록 420을 참조하여 기재된 바와 같이, 병합 클러스터 세트 내의 각각의 클러스터를 그의 가장 가까운 클러스터와 병합할 수 있다.

[0256] 일부 실시형태에서, 방법(600)은 덴드로그램의 복수의 클러스터를 트래버싱하는 경우에, 각각의 클러스터에서: 연관이 유효하면, 클러스터로부터 클러스터의 좌측 자식 클러스터 및 우측 자식 클러스터로의 트래버싱을 계속하는 단계; 및 연관이 무효하면, 클러스터로부터 클러스터의 좌측 자식 클러스터 및 우측 자식 클러스터로의 트래버싱을 중단하는 단계를 포함할 수 있다. 클러스터와 클러스터의 자식 클러스터의 연관이 유효한지 무효한지를 결정하는 단계는 좌측 자식 클러스터와 우측 자식 클러스터 사이의 거리가 연관 임계값 초과이면 연관을 유효한 것으로 결정하고, 그렇지 않으면, 무효한 것으로 결정하는 것을 포함할 수 있다.

[0257] 일부 실시형태에서, 제1 노드의 노드-내 상관관계 및 제2 노드의 노드-내 상관관계 중 적어도 하나가 제1 노드 및 제2 노드의 노드-간 상관관계보다 더 클 수 있다. 제1 노드의 노드-내 상관관계 및 제2 노드의 노드-내 상관관계의 척도 또는 표시는 제1 노드 및 제2 노드의 노드-간 상관관계보다 더 클 수 있다. 제1 노드의 노드-내 상관관계 및 제2 노드의 노드-내 상관관계의 척도는 다음 중 적어도 하나에 기초할 수 있다: 제1 노드 및 제2 노드의 노드-내 최대 상관관계, 제1 노드 및 제2 노드의 노드-내 평균 상관관계, 제1 노드 및 제2 노드의 노드-내 중간값 상관관계, 제1 노드 및 제2 노드의 노드-내 최소 상관관계 및 그의 임의의 조합. 제1 노드의 노드-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 노드의 노드-내 최대 상관관계, 제1 노드의 노드-내 평균 상관관계, 제1 노드의 노드-내 중간값 상관관계, 제1 노드의 노드-내 최소 상관관계 및 그의 임의의 조합. 제2 노드의 노드-내 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제2 노드의 노드-내 최대 상관관계, 제2 노드의 노드-내 평균 상관관계, 제2 노드의 노드-내 중간값 상관관계, 제2 노드의 노드-내 최소 상관관계 및 그의 임의의 조합. 제1 노드 및 제2 노드의 노드-간 상관관계는 다음 중 적어도 하나에 기초할 수 있다: 제1 노드 및 제2 노드의 노드-간 최대 상관관계, 제1 노드 및 제2 노드의 노드-간 평균 상관관계, 제1 노드 및 제2 노드의 노드-간 중간값 상관관계, 제1 노드 및 제2 노드의 노드-간 최소 상관관계 및 그의 임의의 조합.

[0258] 일부 실시형태에서, 좌측 자식 클러스터와 우측 자식 클러스터 사이의 거리는 좌측 자식 클러스터와 우측 자식 클러스터에 의해 나타나는 발현 프로파일 사이의 복수의 표적 중 각각의 표적에서 수행되는 통계적 검정에 기초하여 결정될 수 있다. 통계적 검정은 예를 들어, 웰치의 t-검정일 수 있다. 좌측 자식 클러스터와 우측 자식 클러스터 사이의 거리는 좌측 자식 클러스터에 의해 나타나는 발현 프로파일과 우측 자식 클러스터에 의해 나타나는 각각의 발현 프로파일 사이의 복수의 표적 중 각각의 표적에서 수행되는 통계적 검정의 최대 p-값에 기초하여 결정될 수 있다.

[0259] 일부 실시형태에서, 방법(600)은 블록 616에서 제1 클러스터와 제2 클러스터를 병합하여, 병합된 클러스터를 생성하는 단계 이전에: 제3 클러스터와 제4 클러스터 사이의 거리가 클러스터 거리 임계값 이내이면, 단일 세포의 발현 프로파일을 나타내는 병합 클러스터 세트 내의 각각의 제3 클러스터를 병합 클러스터 세트 내의 제4 클러

스터와 병합하는 단계를 포함한다. 상기 방법은 세포의 발현 프로파일을 나타내는 병합 클러스터 세트 내의 클러스터에 기초하여 복수의 세포를 분류하는 단계를 포함할 수 있다. 상기 방법은 확인된 세포 유형을 구별하기 위한 표적에 기초하여, 전체 전사체 검정을 설계하는 단계 또는 확인된 세포 유형을 구별하기 위한 표적에 기초하여, 표적화된 전사체 검정을 설계하는 단계를 포함할 수 있다.

[0260] 일부 실시형태에서, 방법(600)은 덴드로그램의 복수의 클러스터를 트래버싱하는 경우 각각의 클러스터에서: (3) 클러스터가 단일 세포의 발현 프로파일을 나타내면, 클러스터를 병합 클러스터 세트에 추가하는 단계를 포함한다. 상기 방법은 덴드로그램의 복수의 클러스터를 트래버싱하는 경우, 각각의 클러스터에서: 클러스터 표지를 클러스터에 할당하는 단계를 포함할 수 있다. 일부 실시형태에서, 클러스터가 단일 세포의 발현 프로파일을 나타내면, 클러스터의 클러스터 표지는 단일 세포 지정을 포함하며, 그렇지 않고 클러스터가 부모 클러스터의 좌측 자식 클러스터이면, 클러스터의 클러스터 표지는 부모 클러스터의 클러스터 표지 및 좌측 지정을 포함하고, 그렇지 않으면, 클러스터의 클러스터 표지는 부모 클러스터의 클러스터 표지 및 우측 지정을 포함한다.

[0261] 블록 620에서, 방법(600)은 병합 클러스터 세트 내의 각각의 클러스터에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계를 포함할 수 있다. 방법(600)은 도 4, 예컨대 도 4의 블록 424를 참조로 기재된 바와 같이, 세포 유형을 구별하기 위한 표적을 확인할 수 있다. 일부 실시형태에서, 병합 클러스터 세트 내의 각각의 클러스터에 있어서, 클러스터에 의해 나타나는 세포의 복수의 표적의 발현 프로파일에 기초하여 세포 유형을 구별하기 위한 표적을 확인하는 단계는 클러스터에 의해 나타나는 발현 프로파일과 병합 클러스터 세트 내의 또 다른 클러스터에 의해 나타나는 발현 프로파일 사이의, 세포 유형을 구별하기 위한 표적과 연관된 별개의 서열을 갖는 분자 표지의 수의 차이가 유의성 임계값보다 더 큰 것을 결정하는 것을 포함한다.

[0262] 시퀀싱

[0263] 일부 실시형태에서, 상이한 바코딩된 표적(예를 들어, 확률적으로 바코딩된 표적)의 수를 추정하는 단계는 표지된 표적, 공간 표지, 분자 표지, 샘플 표지, 세포 표지 또는 그의 임의의 산물(예를 들어, 표지된-애플리곤 또는 표지된-cDNA 분자)의 서열을 결정하는 것을 포함할 수 있다. 증폭된 표적은 시퀀싱으로 처리될 수 있다. 바코딩된 표적(예를 들어, 확률적으로 바코딩된 표적) 또는 그의 임의의 산물의 서열을 결정하는 것은 시퀀싱 반응을 행하여, 샘플 표지, 공간 표지, 세포 표지, 분자 표지의 적어도 일부, 표지된 표적(예를 들어, 확률적으로 표지된 표적)의 적어도 일부, 그의 상보물, 그의 역 상보물 또는 그의 임의의 조합의 서열을 결정하는 것을 포함할 수 있다.

[0264] 바코딩된 표적 또는 확률적으로 바코딩된 표적(예를 들어, 증폭된 핵산, 표지된 핵산, 표지된 핵산의 cDNA 카피 등)의 서열의 결정은 혼성화에 의한 시퀀싱(SBH), 라이게이션에 의한 시퀀싱(SBL), 정량적 증분 형광 뉴클레오타이드 첨가 시퀀싱(QIFNAS), 단계적 라이게이션 및 절단, 형광 공명 에너지 전달(FRET), 분자 비콘, TaqMan 리포터 프로브 분해, 파이로시퀀싱(pyrosequencing), 형광 동소 시퀀싱(fluorescent in situ sequencing)(FISSEQ), FISSEQ 비드, 동요 시퀀싱(wobble sequencing), 다중 시퀀싱, 중합 콜로니(POLONY) 시퀀싱; 나노그리드 회전환 시퀀싱(nanogrid rolling circle sequencing)(ROLONY), 대립유전자-특이적 올리고 라이게이션 검정(예를 들어, 올리고 라이게이션 검정(OLA), 라이게이션된 선형 프로브 및 회전환 증폭(RCA) 관독물, 라이게이션된 패드락 프로브(ligated padlock probe)를 사용하는 단일 주형 분자 OLA, 또는 라이게이션된 환형 패드락 프로브 및 회전환 증폭(RCA) 관독물을 사용하는 단일 주형 분자 OLA) 등을 포함하나 이에 제한되지 않는 다양한 시퀀싱 방법을 사용하여 수행될 수 있다.

[0265] 일부 실시형태에서, 바코딩된 표적(예를 들어, 확률적으로 바코딩된 표적) 또는 그의 임의의 산물의 서열의 결정은 쌍형성-말단 시퀀싱, 나노포어 시퀀싱(nanopore sequencing), 고처리량 시퀀싱, 샷건 시퀀싱(shotgun sequencing), 염료 종결인자 시퀀싱, 다중 프라이머 DNA 시퀀싱, 프라이머 이동(walking), 생어 디데옥시 시퀀싱(Sanger dideoxy sequencing), 맥심-길버트(Maxim-Gilbert) 시퀀싱, 파이로시퀀싱, 실제 단일 분자 시퀀싱(true single molecule sequencing) 또는 그의 임의의 조합을 포함한다. 대안적으로, 바코딩된 표적 또는 그의 임의의 산물의 서열은 전자 현미경 또는 화학-민감성 전계 효과 트랜지스터(chemical-sensitive field effect transistor)(chemFET) 어레이에 의해 결정될 수 있다.

[0266] 로슈(Roche) 454, 일루미나 솔렉사(Illumina Solexa), ABI-SOLiD, 이온 토렌트(ION Torrent), 컴플리트 게노믹스(Complete Genomics), 퍼시픽 바이오사이언스(Pacific Bioscience), 헬리코스(Helicos) 또는 폴로나토르(Polonator) 플랫폼과 같은 플랫폼을 사용하는 환형 어레이 시퀀싱과 같은 고처리량 시퀀싱 방법이 사용될 수 있다. 일부 실시형태에서, 시퀀싱은 MiSeq 시퀀싱을 포함할 수 있다. 일부 실시형태에서, 시퀀싱은 HiSeq 시퀀

싱을 포함할 수 있다.

[0267] 표지된 표적(예를 들어, 확률적으로 표지된 표적)은 유기체 계놈의 약 0.01%의 유전자 내지 유기체 계놈의 약 100%의 유전자를 나타내는 핵산을 포함할 수 있다. 예를 들어, 유기체 계놈의 약 0.01%의 유전자 내지 유기체 계놈의 약 100%의 유전자가, 샘플로부터의 상보성 서열을 함유하는 유전자의 포획에 의해, 복수의 다량체를 포함하는 표적 상보성 영역을 사용하여 시퀀싱될 수 있다. 일부 실시형태에서, 바코딩된 표적은 유기체의 전사체의 약 0.01%의 전사물 내지 유기체의 전사체의 약 100%의 전사물을 나타내는 핵산을 포함한다. 예를 들어, 유기체의 전사체의 약 0.501%의 전사물 내지 유기체의 전사체의 약 100%의 전사물은 샘플로부터의 mRNA의 포획에 의해, 폴리(T) 테일을 포함하는 표적 상보성 영역을 사용하여 시퀀싱될 수 있다.

[0268] 복수의 바코드(예를 들어, 확률적 바코드)의 공간 표지 및 분자 표지의 서열의 결정은 0.00001%, 0.0001%, 0.001%, 0.01%, 0.1%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 99%, 100%, 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 복수의 바코드를 시퀀싱하는 것을 포함할 수 있다. 복수의 바코드의 표지, 예를 들어, 샘플 표지, 공간 표지 및 분자 표지의 서열의 결정은 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 , 10^{10} , 10^{11} , 10^{12} , 10^{13} , 10^{14} , 10^{15} , 10^{16} , 10^{17} , 10^{18} , 10^{19} , 10^{20} , 또는 이들 값 중 임의의 2개 사이의 수 또는 범위의 복수의 바코드를 시퀀싱하는 것을 포함할 수 있다. 복수의 바코드 중 일부 또는 전부의 시퀀싱은 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000개, 약, 적어도 또는 최대 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000개, 이들 값 중 임의의 2개 사이의 수 또는 범위의 뉴클레오타이드 또는 염기의 관독물 길이를 갖는 서열을 생성하는 것을 포함할 수 있다.

[0269] 시퀀싱은 바코딩된 표적의 적어도 또는 적어도 약 10, 20, 30, 40, 50, 60, 70, 80, 90, 100개 이상의 뉴클레오타이드 또는 염기쌍을 시퀀싱하는 것을 포함할 수 있다. 예를 들어, 시퀀싱은 복수의 바코딩된 표적에서 중합효소 연쇄 반응(PCR) 증폭을 수행함으로써 50, 75 또는 100개 이상의 뉴클레오타이드의 관독물 길이를 갖는 서열로 시퀀싱 데이터를 생성하는 것을 포함할 수 있다. 시퀀싱은 바코딩된 표적의 적어도 또는 적어도 약 200, 300, 400, 500, 600, 700, 800, 900, 1,000개 이상의 뉴클레오타이드 또는 염기쌍을 시퀀싱하는 것을 포함할 수 있다. 시퀀싱은 바코딩된 표적의 적어도 또는 적어도 약 1500, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000 또는 10000개 이상의 뉴클레오타이드 또는 염기쌍을 시퀀싱하는 것을 포함할 수 있다.

[0270] 시퀀싱은 시행당 적어도 약 200, 300, 400, 500, 600, 700, 800, 900, 1,000개 이상의 시퀀싱 관독물을 포함할 수 있다. 일부 실시형태에서, 시퀀싱은 시행당 적어도 또는 적어도 약 1500, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000 또는 10000개 이상의 시퀀싱 관독물을 시퀀싱하는 것을 포함한다. 시퀀싱은 시행당 약 1,600,000,000개 이하의 시퀀싱 관독물을 포함할 수 있다. 시퀀싱은 시행당 약 200,000,000개 이하의 관독물을 포함할 수 있다.

[0271] 샘플

[0272] 일부 실시형태에서, 복수의 표적은 하나 이상의 샘플에 포함될 수 있다. 샘플은 하나 이상의 세포 또는 하나 이상의 세포로부터의 핵산을 포함할 수 있다. 샘플은 단일의 세포 또는 단일의 세포로부터의 핵산일 수 있다. 하나 이상의 세포는 하나 이상의 세포 유형의 것일 수 있다. 하나 이상의 세포 유형 중 적어도 하나는 뇌 세포, 심장 세포, 암 세포, 순환 종양 세포, 기관 세포, 상피 세포, 전이 세포, 양성 세포, 일차 세포, 순환 세포 또는 그의 임의의 조합일 수 있다.

[0273] 본 발명의 방법에 사용하기 위한 샘플은 하나 이상의 세포를 포함할 수 있다. 샘플은 하나 이상의 세포를 지칭할 수 있다. 일부 실시형태에서, 복수의 세포는 하나 이상의 세포 유형을 포함할 수 있다. 하나 이상의 세포 유형 중 적어도 하나는 뇌 세포, 심장 세포, 암 세포, 순환 종양 세포, 기관 세포, 상피 세포, 전이 세포, 양성 세포, 일차 세포, 순환 세포 또는 그의 임의의 조합일 수 있다. 일부 실시형태에서, 세포는 암 조직, 예를 들어, 유방암, 폐암, 결장암, 전립선암, 난소암, 췌장암, 뇌암, 흑색종 및 비-흑색종 피부암 등으로부터 절제된 암 세포이다. 일부 실시형태에서, 세포는 암으로부터 유래되지만 체액(예를 들어, 순환 종양 세포)으로부터 수집된다. 암의 비제한적인 예는 선종, 선암종, 편평 세포 암종, 기저 세포 암종, 소세포 암종, 거대 세포 미분화 암종, 연골육종 및 섬유육종을 포함할 수 있다. 샘플은 조직, 세포 단층, 고정된 세포, 조직 섹션 또는 그의 임의의 조합을 포함할 수 있다. 샘플은 생물학적 샘플, 임상적 샘플, 환경 샘플, 대상체로부터의 생물학적 유체, 조직 또는 세포를 포함할 수 있다. 샘플은 인간, 포유동물, 개, 랫트, 마우스, 어류, 파리, 벌레, 식물, 진균,

박테리아, 바이러스, 척추동물 또는 무척추동물로부터 수득될 수 있다.

- [0274] 일부 실시형태에서, 세포는 바이러스로 감염되고, 바이러스 올리고뉴클레오티드를 함유하는 세포이다. 일부 실시형태에서, 바이러스 감염은 바이러스, 예컨대 단일-가닥(+ 가닥 또는 "센스") DNA 바이러스(예를 들어, 파코 바이러스(parvovirus)) 또는 이중-가닥 RNA 바이러스(예를 들어, 레오바이러스(reovirus))에 의해 야기될 수 있다. 일부 실시형태에서, 세포는 박테리아이다. 이들은 그람-양성 또는 그람-음성 박테리아 중 어느 하나를 포함할 수 있다. 일부 실시형태에서, 세포는 진균이다. 일부 실시형태에서, 세포는 원생동물 또는 다른 기생충이다.
- [0275] 본원에 사용되는 용어 "세포"는 하나 이상의 세포를 지칭할 수 있다. 일부 실시형태에서, 세포는 정상 세포, 예를 들어 상이한 발생 단계의 인간 세포 또는 상이한 기관 또는 조직 유형으로부터의 인간 세포이다. 일부 실시형태에서, 세포는 비-인간 세포, 예를 들어, 다른 유형의 포유동물 세포(예를 들어, 마우스, 랫트, 돼지, 개, 소 또는 말)이다. 일부 실시형태에서, 세포는 다른 유형의 동물 또는 식물 세포이다. 다른 실시형태에서, 세포는 임의의 원핵 또는 진핵 세포일 수 있다.
- [0276] 일부 실시형태에서, 세포는 세포를 비드와 연관시키기 전에 분류된다. 예를 들어, 세포는 형광-활성화된 세포 분류 또는 자성-활성화된 세포 분류, 또는, 더욱 일반적으로 유세포계수법에 의해 분류될 수 있다. 세포는 크기 별로 여과될 수 있다. 일부 실시형태에서, 잔류물은 비드와 연관된 세포를 함유한다. 일부 실시형태에서, 통과액은 비드와 연관된 세포를 함유한다.
- [0277] 샘플은 복수의 세포를 지칭할 수 있다. 샘플은 단층의 세포를 지칭할 수 있다. 샘플은 박편(예를 들어, 조직 박편)을 지칭할 수 있다. 샘플은 어레이 상에 1차원으로 배치될 수 있는 세포의 고체 또는 반-고체 집합물을 지칭할 수 있다.
- [0278] 실행 환경
- [0279] 본 발명은 본 발명의 방법(예를 들어, 방법(400) 또는 방법(600))을 구현하기 위해 프로그래밍되는 컴퓨터 시스템을 제공한다. 도 7은 프로그래밍되거나, 그렇지 않으면 본원에 개시된 방법 중 임의의 것을 구현하도록 구성된 컴퓨터 시스템(700)을 보여준다. 컴퓨터 시스템(700)은 사용자의 전자 장치이거나 또는 전자 장치에 대하여 원위에 위치한 컴퓨터 시스템일 수 있다. 전자 장치는 모바일 전자 장치일 수 있다.
- [0280] 컴퓨터 시스템(700)은 단일 코어 또는 다중 코어 프로세서, 또는 병렬 처리를 위한 복수의 프로세서일 수 있는 중앙 처리 장치(CPU, 본원에서 또한 "프로세서" 및 "컴퓨터 프로세서")(705)를 포함한다. 컴퓨터 시스템(700)은 또한 메모리 또는 메모리 위치(710)(예를 들어, 랜덤-액세스 메모리, 읽기-전용 메모리, 플래시 메모리), 전자 저장 장치(715)(예를 들어, 하드 디스크), 하나 이상의 다른 시스템과 통신하기 위한 통신 인터페이스(720)(예를 들어, 네트워크 어댑터), 및 주변 장치(725), 예를 들어, 캐시, 기타 메모리, 데이터 저장 및/또는 전자 디스플레이 어댑터를 포함한다. 메모리(710), 저장 장치(715), 인터페이스(720) 및 주변 장치(725)는 마더보드와 같은 통신 버스(실선)를 통해 CPU(705)와 통신한다. 저장 장치(715)는 데이터를 저장하기 위한 데이터 저장 장치(또는 데이터 저장소)일 수 있다. 컴퓨터 시스템(700)은 통신 인터페이스(720)의 도움으로 컴퓨터 네트워크("네트워크")(730)에 작동 가능하게 연결될 수 있다. 네트워크(730)는 인터넷, 인터넷 및/또는 엑스트라넷, 또는 인터넷과 통신하는 인트라넷 및/또는 엑스트라넷일 수 있다. 네트워크(730)는 일부 경우에 원거리통신 및/또는 데이터 네트워크이다. 네트워크(730)는 클라우드 컴퓨팅과 같은 분산 컴퓨팅을 가능하게 할 수 있는 하나 이상의 컴퓨터 서버를 포함할 수 있다. 일부 경우에 컴퓨터 시스템(700)의 도움으로 네트워크(730)는 피어-투-피어 네트워크를 구현할 수 있으며, 이는 컴퓨터 시스템(700)에 결합된 장치가 클라이언트 또는 서버로 거동하는 것을 가능하게 할 수 있다.
- [0281] CPU(705)는 프로그램 또는 소프트웨어로 구현될 수 있는 일련의 기계-판독 가능한 명령어를 실행할 수 있다. 명령어는 메모리(710)와 같은 메모리 위치에 저장될 수 있다. 명령어는 CPU(705)에 지시될 수 있으며, 이는 이후에 본 발명의 방법을 구현하도록 CPU(705)를 프로그래밍하거나 달리 구성할 수 있다. CPU(705)에 의해 수행되는 작업의 예는 페치(fetch), 디코드(decode), 실행 및 라이트백(writeback)을 포함할 수 있다. CPU(705)는 집적 회로와 같은 회로의 일부일 수 있다. 시스템(700)의 하나 이상의 다른 구성요소가 회로에 포함될 수 있다. 일부 경우에, 회로는 주문형 집적 회로(ASIC)이다.
- [0282] 저장 장치(715)는 드라이버, 라이브러리 및 저장된 프로그램과 같은 파일을 저장할 수 있다. 저장 장치(715)는, 예를 들어, 사용자 선호도 및 사용자 프로그램과 같은 사용자 데이터를 저장할 수 있다. 컴퓨터 시스템(700)은 일부 경우에 인트라넷 또는 인터넷을 통해 컴퓨터 시스템(700)과 통신하는 원격 서버에 위치하는 것과 같이 컴퓨터 시스템(700)의 외부에 있는 하나 이상의 추가의 데이터 저장 장치를 포함할 수 있다.

- [0283] 컴퓨터 시스템(700)은 네트워크(730)를 통해 하나 이상의 원격 컴퓨터 시스템과 통신할 수 있다. 예를 들어, 컴퓨터 시스템(700)은 사용자(예를 들어, 미생물학자)의 원격 컴퓨터 시스템과 통신할 수 있다. 원격 컴퓨터 시스템의 예는 개인 컴퓨터(예를 들어, 휴대용 PC), 슬레이트 또는 태블릿 PC(예를 들어, 애플(Apple)® 아이패드(iPad), 삼성(Samsung)® 갤럭시 탭(Galaxy Tab)), 전화기, 스마트폰(예를 들어, 애플® 아이폰(iPhone), 안드로이드(Android)-지원 장치, 블랙베리(Blackberry)®), 또는 개인 디지털 보조장치를 포함한다. 사용자는 네트워크(730)를 통해 컴퓨터 시스템(700)에 액세스할 수 있다.
- [0284] 컴퓨터 시스템(700)은, 예를 들어, 스트링으로 나타낸 바와 같은 미생물의 복수의 분류군의 상호작용 또는 스트링 동시-발생을 나타내는 출력을 제공하기 위한 사용자 인터페이스(UI)(740)를 포함하는 전자 디스플레이(735)를 포함하거나, 이와 통신할 수 있다. UI의 예는 그래픽 사용자 인터페이스(GUI) 및 웹-기반 사용자 인터페이스를 포함하나, 이에 제한되지는 않는다.
- [0285] 본원에 기재된 바와 같은 방법은, 예를 들어, 메모리(710) 또는 전자 저장 장치(715)와 같은 컴퓨터 시스템(700)의 전자 저장 위치에 저장된 기계(예를 들어, 컴퓨터 프로세서) 실행 가능한 코드에 의해 구현될 수 있다. 기계 실행 가능하거나 기계 판독 가능한 코드는 소프트웨어의 형태로 제공될 수 있다. 사용 동안, 코드는 프로세서(705)에 의해 실행될 수 있다. 일부 경우에, 코드는 저장 장치(715)로부터 검색될 수 있고, 프로세서(705)에 의한 액세스 준비를 위해 메모리(710)에 저장될 수 있다. 일부 상황에서, 전자 저장 장치(715)는 배제될 수 있고, 기계-실행 가능한 명령어는 메모리(710)에 저장된다.
- [0286] 코드는 사전 컴파일링되고, 코드를 실행하도록 이루어진 프로세서를 갖는 기계와 함께 사용하도록 구성될 수 있거나, 런타임 동안 컴파일링될 수 있다. 코드는 사전 컴파일링된 또는 컴파일링된 방식으로 코드를 실행할 수 있도록 선택될 수 있는 프로그래밍 언어로 제공될 수 있다.
- [0287] 컴퓨터 시스템(700)과 같은 본원에 제공된 시스템 및 방법의 양태는 프로그래밍으로 구현될 수 있다. 상기 기술의 다양한 양태는 통상적으로 기계 판독 가능한 매체의 유형으로 수행되거나 구현되는 기계(또는 프로세서) 실행 가능한 코드 및/또는 관련 데이터의 형태인 "제품" 또는 "제조 물품"으로 생각될 수 있다. 기계-실행 가능한 코드는 메모리(예를 들어, 읽기-전용 메모리, 랜덤-액세스 메모리, 플래시 메모리) 또는 하드디스크와 같은 전자 저장 장치에 저장될 수 있다. "저장" 유형 매체는 소프트웨어 프로그래밍을 위해 언제든지 비-일시적인 저장을 제공할 수 있는 컴퓨터, 프로세서 등, 또는 그의 관련 모듈의 유형 메모리 일부 또는 전부, 예컨대, 다양한 반도체 메모리, 테이프 드라이브, 디스크 드라이브 등을 포함할 수 있다. 소프트웨어의 전부 또는 일부는 때때로 인터넷 또는 다양한 다른 원격통신 네트워크를 통해 통신할 수 있다. 예를 들어, 이러한 통신은 하나의 컴퓨터 또는 프로세서로부터 또 다른 컴퓨터 또는 프로세서로, 예를 들어, 관리 서버 또는 호스트 컴퓨터로부터 애플리케이션 서버의 컴퓨터 플랫폼으로 소프트웨어의 로딩을 가능하게 할 수 있다. 따라서, 소프트웨어 요소를 가질 수 있는 또 다른 유형의 매체는 유선 및 광학 육상 통신선 네트워크 및 다양한 에어-링크(air-link)를 통해 로컬 장치 사이의 물리적 인터페이스에 걸쳐 사용되는 것과 같은 광학, 전기 및 전자기파를 포함한다. 유선 또는 무선 링크, 광학 링크 등과 같은 상기 파를 운반하는 물리적 요소는 또한 소프트웨어를 갖는 매체로 간주될 수 있다. 비-일시적인 유형 "저장" 매체로 제한되지 않는 한, 본원에서 사용되는 컴퓨터 또는 기계 "판독 가능한 매체"와 같은 용어는 실행을 위해 프로세서에 명령어를 제공하는데 참여하는 임의의 매체를 나타낸다.
- [0288] 그러므로, 컴퓨터-실행 가능한 코드와 같은 기계 판독 가능한 매체는 유형 저장 매체, 반송파 매체 또는 물리적 전송 매체를 포함하나 이에 제한되지는 않는 많은 형태를 취할 수 있다. 비-휘발성 저장 매체는, 예를 들어, 도면에 제시된 바와 같은 데이터베이스 등을 구현하는데 사용될 수 있는 것과 같은 임의의 컴퓨터(들) 등의 저장 장치 중 임의의 저장 장치와 같은 광학 또는 자기 디스크를 포함한다. 휘발성 저장 매체는 이러한 컴퓨터 플랫폼의 메인 메모리와 같은 동적 메모리를 포함한다. 유형 전송 매체는 동축 케이블; 컴퓨터 시스템 내의 버스를 포함하는 와이어를 포함하는 구리 와이어 및 광섬유를 포함한다. 반송파 전송 매체는 전기 또는 전자기 신호, 또는 무선 주파수(RF) 및 적외선(IR) 데이터 통신 동안 생성되는 것과 같은 음파 또는 광파의 형태를 취할 수 있다. 따라서, 컴퓨터-판독 가능한 매체의 일반적인 형태는, 예를 들어, 플로피 디스크, 플렉시블 디스크, 하드 디스크, 자기 테이프, 임의의 다른 자기 매체, CD-ROM, DVD 또는 DVD-ROM, 임의의 다른 광학 매체, 펀치 카드 종이 테이프, 구멍의 패턴을 갖는 임의의 다른 물리적 저장 매체, RAM, ROM, PROM 및 EPROM, FLASH-EPROM, 임의의 다른 메모리 칩 또는 카트리지, 데이터 또는 명령어를 전송하는 반송파, 이러한 반송파를 전송하는 케이블 또는 링크, 또는 컴퓨터가 프로그래밍 코드 및/또는 데이터를 판독할 수 있는 임의의 다른 매체를 포함한다. 이들 형태의 컴퓨터 판독 가능한 매체 중 많은 것이 실행을 위해 프로세서에 하나 이상의 명령어의 하나 이상의 시퀀스를 전달하는 것에 수반될 수 있다.

- [0289] 일부 실시형태에서, 컴퓨터 시스템(700)의 분석 기능의 일부 또는 전부는 단일의 소프트웨어 패키지에 패키징될 수 있다. 일부 실시형태에서, 데이터 분석 능력의 완전한 세트는 소프트웨어 패키지의 모음을 포함할 수 있다. 일부 실시형태에서, 데이터 분석 소프트웨어는 검정 기기 시스템과 독립적으로, 사용자가 이용할 수 있게 만들어진 자립형 패키지일 수 있다. 일부 실시형태에서, 소프트웨어는 웹-기반일 수 있으며, 사용자가 데이터를 공유하게 할 수 있다. 일부 실시형태에서, 상업적으로 입수 가능한 소프트웨어를 사용하여 데이터 분석의 전부 또는 일부를 수행할 수 있으며, 예를 들어, 세븐 브릿지즈(Seven Bridges)(<https://www.sbgenomics.com/>) 소프트웨어를 사용하여 전체 세포의 수집물에 대하여 각 세포 내에서 발생하는 하나 이상의 유전자의 카피의 수의 표를 컴파일링할 수 있다.
- [0290] 본 발명의 방법 및 시스템은 하나 이상의 알고리즘 또는 방법에 의해 구현될 수 있다. 방법은 중앙 처리 장치(705)에 의한 실행 시에 소프트웨어에 의해 구현될 수 있다. 소프트웨어에 의해 구현되는 알고리즘 또는 방법의 예시적인 응용은 서열 판독 처리(예를 들어, 병합, 필터링, 트리밍, 클러스터링), 정렬 및 호출 및 스트링 데이터 및 광학 밀도 데이터(예를 들어, 가장 유망한 수 및 배양 가능한 존재비 결정)의 처리를 위한 바이오인포매틱스 방법을 포함한다.
- [0291] 예시적인 실시형태에서, 컴퓨터 시스템(700)은 단일 세포, 확률적 바코딩 검정을 수행함으로써 생성되는 서열 데이터세트에서 데이터 분석을 수행할 수 있다. 데이터 분석 기능의 예는 (i) 샘플 표지, 세포 표지, 공간 표지 및 분자 표지, 및 검정의 시행에서 생성된 확률적 바코드 라이브러리를 시퀀싱함으로써 제공되는 표적 서열 데이터의 디코딩/역다중화를 위한 알고리즘, (ii) 세포당 유전자당 관독물의 수 및 데이터에 기초하여 세포당 유전자당 독특한 전사물 분자의 수를 결정하고, 요약 표를 생성하기 위한 알고리즘, (iii) 예를 들어, 유전자 발현 데이터에 의한 세포의 클러스터링을 위한, 또는 세포당 유전자당 전사물 분자의 수의 결정을 위하여 신뢰성 구간의 예측 등을 위한 서열 데이터의 통계적 분석, (iv) 예를 들어, 주성분 분석, 계층적 클러스터링, k-평균 클러스터링, 자기-조직화 맵, 신경망 등을 사용하여 회귀 세포의 하위-집단을 확인하기 위한 알고리즘, (v) 유전자 서열 데이터를 알려져 있는 참조 서열과 정렬하고, 돌연변이, 다형성 마커 및 스플라이스 변이체를 검출하기 위한 서열 정렬 능력, 및 (vi) 증폭 또는 시퀀싱 오류를 보상하기 위한 분자 표지의 자동화된 클러스터링을 포함하나 이들에 제한되지 않는다. 일부 실시형태에서, 컴퓨터 시스템(700)은 세포의 집합물 중 각각의 세포에서 발생하는 하나 이상의 유전자의 카피의 수를 나타내는 유용한 그래픽 형식, 예를 들어, 열지도로 시퀀싱 결과를 출력할 수 있다. 일부 실시형태에서, 컴퓨터 시스템(700)은 예를 들어, 세포의 집합물 중 각각의 세포에서 발생하는 하나 이상의 유전자의 카피의 수를 세포의 유형, 회귀 세포의 유형 또는 특정 질환 또는 병증을 갖는 대상체로부터 유래된 세포와 상호관련시킴으로써 시퀀싱 결과로부터 생물학적 의미를 추출하기 위한 알고리즘을 실행할 수 있다. 일부 실시형태에서, 컴퓨터 시스템(700)은 상이한 생물학적 샘플에 걸쳐 세포의 모집단을 비교하기 위한 알고리즘을 실행할 수 있다.
- [0292] 실시예
- [0293] 상기 논의된 실시형태의 일부 양태는 어떠한 방식이든 본 발명의 범주를 제한하고자 하지 않는 하기의 실시예에 추가로 상세히 개시되어 있다.
- [0294] 실시예 1
- [0295] 재귀적 덴드로그램 분할 및 시험에 이어서 병합에 의한 클러스터링
- [0296] 이러한 실시예는 회귀적 분할(예를 들어, 회귀적 덴드로그램 분할) 및 시험에 이어서 병합에 의한 클러스터링 방법을 기술한다.
- [0297] 주의
- [0298] 이러한 실시예에 예시된 방법에서, 덴드로그램 분할 단계 동안, 알고리즘이 충분히 낮은 p -값(또는 충분히 높은 $-\log_{10}(p\text{-값})$)을 달성하는 적어도 하나의 유전자를 찾을 수 있으면, (예를 들어, 디폴트에 의해) 분할은 생물학적으로 적절한 것으로 간주된다. 다시 말하면, 튜닝되는 하이퍼파라미터만이 일부 실시형태에서, 점수 임계값 파라미터이다. 더 높은 점수 임계값(예를 들어 100)은 더 낮은 p -값($10e-100$)에 상응하며, 이는 분할을 유효한 것으로 간주하기 위하여 더욱 유의미한 유전자를 찾아야 함을 의미한다. 더 높은 점수 임계값은 더 적은 수의 클러스터를 초래한다.
- [0299] 너무 많은 클러스터가 분할 단계 후에 생성되면, 사용자는 점수 임계값을 증가시키는 시도를 할 수 있다. 분할 단계에서 너무 적은 클러스터가 생성되면, 사용자는 점수 임계값을 감소시키는 시도를 할 수 있다. 다수의 점수

임계값을 동일한 거리 행렬에서 시험할 수 있다. 거리 행렬을 사전 계산함으로써, 많은 계산 시간을 절약할 수 있다.

- [0300] 상이한 점수 임계값을 통한 스위핑을 계속하여, 무의미한 결과를 생성하면, 제1 장소(즉, 거리 행렬)에서 생성된 덴드로그램에 문제가 있을 수 있다. 하기 셀 [3]에 나타난 바와 같이, 알고리즘의 제1 단계는 분자 계수의 행렬로부터 거리 행렬로 가는 것을 필요로 한다(사전처리 단계). 상이한 종류의 사전처리를 시도하는 것이 바람직할 수 있다. 사용자는 아마도 또 다른 거리 미터법을 시도하거나, 로그를 취하지 않거나, 세포 및/또는 유전자를 사전 필터링하는 시도를 할 수 있으며, 이는 응용을 위하여 더욱 정확한 거리 미터법을 생성할 수 있다.
- [0301] 분할 단계가 관련 없는 것으로 보이는 많은 작은 클러스터를 생성하면, 해체 백분위수 파라미터가 감소될 수 있다. 이러한 파라미터는 얼마나 많은 그의 쌍별 거리가 전체 쌍별 거리의 하위 해체 백분위수 내에 있는지에 기초하여, 최종 클러스터를 유지할지 아닐지를 결정한다. 20의 해체 백분위수를 사용하여 알고리즘을 시행하는 것은 예를 들어, 적어도 하나의 쌍별 거리가 전체 거리의 하위 20% 내에 있는 경우에만 클러스터를 유지할 것이다.
- [0302] 클러스터가 2개의 클러스터로 분리되는 이유를 결정하기 위하여, 2개의 클러스터를 확인하고, 쌍별 t-검정 분석을 행한다. 이것은 클러스터의 모든 쌍에 있어서 실시예 2의 셀 [13]에서 행해질 수 있다. 이러한 함수는 2개의 클러스터를 구별하는 마커를 보여줄 것이다. 또한, 분할을 초래하는 분할 알고리즘의 정확한 단계를 확인하기 위해 분할이 결정되는 방법의 탐색 특징을 확인한다.
- [0303] 부정확한 클러스터가 함께 병합되는지를 결정하기 위하여, 병합 단계에서 점수 임계값 파라미터를 감소시킨다. 여기서 점수 임계값이 더 클수록, 2개의 상이한 클러스터가 함께 병합될 가능성이 더 크다. 또한, 병합을 초래하는 병합 알고리즘의 정확한 단계를 확인하기 위하여 병합이 결정되는 방법의 탐색 특징을 확인한다.
- [0304] 더 많은 이상점을 확인하기 위하여, 병합 단계에서 이상점_임계값_백분위수_파라미터를 감소시키는 시도를 한다.
- [0305] 종속성
- [0306] 모듈은 하기의 종속성을 가졌다: - numpy (1.10.4) - scipy (0.17.0) - matplotlib (1.5.1) - sklearn (0.17.1) - networkx (1.11) - community - rpy2 (2.8.2)
- [0307] networkx, community 및 rpy2는 디폴트에 의해 요구되지 않는다. networkx 및 community는 커뮤니티 검출을 위해 사용된다. 또한, networkx는 최대-가중치-매칭(maxed-weight-matching)을 위해 사용된다(2개의 세트의 표지가 얼마나 가까운지의 미터법으로서). rpy2는 2개의 모집단이 실제로 하나의 모집단일지 아닐지에 대한 통계적 검정인 sigclust를 시행하기 위해 사용된다. sigclust를 시행하기 위하여, 사용자는 또한 sigclust 패키지와 함께 R을 설치할 필요가 있을 수 있다.
- [0308] [1]에서: # 관련 모듈 및 라이브러리를 로드한다
- [0309] %load_ext autoreload
- [0310] %autoreload 2
- [0311] %matplotlib inline
- [0312] from dendroplit import split,merge import pickle
- [0313] import numpy as np
- [0314] import matplotlib.pyplot as plt np.set_printoptions(precision=2,suppress=True)
- [0315] 파이프라인 시행
- [0316] 파이프라인으로의 입력은 'X'로 지칭되는 분자 계수(자연수)의 $N \times M$ 행렬이다. '유전자'는 유전자 명칭의 길이-M 목록이다. 'x1' 및 'x2'는 사용자가 선택한 방법을 사용하는 데이터의 2-D 임베딩을 나타낸다. 'x1' 및 'x2'는 단지 중간 단계와 함께 파이프라인의 결과를 시각화시키기 위하여 사용된다. 알고리즘은 합이 0인 'X'의 모든 열이 제거되는 것을 필요로 하며, 이러한 코드의 셀은 그를 고려한다.
- [0317] [2]에서: # 데이터를 로드한다
- [0318] dataset = 'Resolve4'

[0319] pickledir = '/Users/user1/Desktop/datasets/'

[0320] X,genes = pickle.load(file(pickledir+dataset+'.pickle'))

[0321] x1,x2 = pickle.load(file(pickledir+dataset+'tsne.pickle'))

[0322] # 함이 0인 X의 열을 제거한다

[0323] X,genes = split.filter_genes(X,genes)

[0324] 모든 세포에 걸쳐 0개 초과와 계수를 갖는 19307개의 유전자를 유지한다.

[0325] 거리 행렬을 먼저 계수 행렬로부터 생성할 수 있다. 하기의 셀은 로그-변환된 샘플 사이의 쌍별 상관관계 거리 ($\log(X+1)$)를 계산함으로써 이를 달성한다. 알고리즘의 분할 부분은 아래 나타낸 바와 같이 사용자가 거리 행렬을 피드할 수 있지만, 오직 계수 행렬을 입력으로서 필요로 하였다. 이러한 알고리즘의 부분은 알고리즘에 의해 생성된 모든 중간 정보를 추적하는 데이터 구조인 'history'와 함께 샘플에 대한 표지의 길이-N 세트(스트링)를 반환한다. 'history'는 알고리즘이 이러한 표지를 어떻게 생성하는지(그리고 이러한 표지를 생성하기 위하여 어떤 특징이 가장 중요하였는지)를 분석하기 위해 사용되는 이후의 함수에 유용하였다. 표지는 거리 행렬을 사용하여 생성되는 덴드로그램에 따라 클러스터가 위치한 곳을 나타내는 스트링이었다. 예를 들어, 'rLLR'는 이러한 점이 루트의 좌측 서브트리의 좌측 서브트리의 우측 서브트리에 속하는 것을 의미한다.

[0326] [3]에서: # 제1 세트의 표지를 얻는다. 알고리즘의 외측에서 거리 행렬을 계산하는 것은 매우 권고된다

[0327] D = split.log_correlation(X)

[0328] ys,shistory = split.dendroplit((D,X), preprocessing='precomputed', score_threshold=10, verbose=True, disband_percentile=50)

[0329] 잠재적인 분할 결과: 883 및 3

[0330] dendroplit/feature selection.py:106: RuntimeWarning: divide by zero encountered in log10

[0331] gene_scores = np.nan_to_num(-np.log10(p[keep_inds]))

[0332] 분할 점수 1.8E+308

[0333] 잠재적인 분할 결과: 1 및 882

[0334] 잠재적인 분할 결과: 484 및 398

[0335] /Users/user1/anaconda2/lib/python2.7/site-packages/scipy/stats/

[0336] _distn_infrastructure.py:1748: Runtime

[0337] cond1 = (scale > 0) & (x > self.a) & (x < self.b)

[0338] /Users/user1/anaconda2/lib/python2.7/site-packages/scipy/stats/_distn_infrastructure.py:1748: Runtime

[0339] cond1 = (scale > 0) & (x > self.a) & (x < self.b)

[0340] /Users/user1/anaconda2/lib/python2.7/site-packages/scipy/stats/_distn_infrastructure.py:1749: Runtime

[0341] cond2 = cond0 & (x <= self.a)

[0342] 분할 점수 182.26

[0343] 잠재적인 분할 결과: 481 및 3 분할 점수 1.8E+308

[0344] 잠재적인 분할 결과: 1 및 480

[0345] 잠재적인 분할 결과: 1 및 479

[0346] 잠재적인 분할 결과: 195 및 284

[0347] 분할 점수 125.49

[0348] 잠재적인 분할 결과: 177 및 18

[0349]	분할 점수 15.35
[0350]	잠재적인 분할 결과: 1 및 176
[0351]	잠재적인 분할 결과: 1 및 175
[0352]	잠재적인 분할 결과: 1 및 174
[0353]	잠재적인 분할 결과: 12 및 162
[0354]	분할 점수 18.88
[0355]	잠재적인 분할 결과: 1 및 11
[0356]	잠재적인 분할 결과: 1 및 10
[0357]	잠재적인 분할 결과: 2 및 8
[0358]	분할 점수 6.11
[0359]	잠재적인 분할 결과: 1 및 161
[0360]	잠재적인 분할 결과: 1 및 160
[0361]	잠재적인 분할 결과: 28 및 132
[0362]	분할 점수 12.32
[0363]	잠재적인 분할 결과: 25 및 3
[0364]	분할 점수 13.94
[0365]	잠재적인 분할 결과: 1 및 24
[0366]	잠재적인 분할 결과: 11 및 13
[0367]	분할 점수 4.77
[0368]	잠재적인 분할 결과: 1 및 2
[0369]	잠재적인 분할 결과: 1 및 1
[0370]	해체(서로 너무 먼 클러스터 내의 점) 잠재적인 분할 결과: 122 및 10
[0371]	분할 점수 18.52
[0372]	잠재적인 분할 결과: 13 및 109
[0373]	분할 점수 24.92
[0374]	잠재적인 분할 결과: 6 및 7
[0375]	분할 점수 3.77
[0376]	잠재적인 분할 결과: 105 및 4
[0377]	분할 점수 31.72
[0378]	잠재적인 분할 결과: 15 및 90
[0379]	분할 점수 11.31
[0380]	잠재적인 분할 결과: 3 및 12
[0381]	분할 점수 6.55
[0382]	잠재적인 분할 결과: 17 및 73
[0383]	분할 점수 8.91
[0384]	잠재적인 분할 결과: 2 및 2

[0385]	분할 점수 1.58
[0386]	해체(서로 너무 먼 클러스터 내의 점) 잠재적인 분할 결과: 2 및 8
[0387]	분할 점수 5.79
[0388]	잠재적인 분할 결과: 1 및 17
[0389]	잠재적인 분할 결과: 1 및 16
[0390]	잠재적인 분할 결과: 1 및 15
[0391]	잠재적인 분할 결과: 4 및 11
[0392]	분할 점수 4.57
[0393]	잠재적인 분할 결과: 1 및 283
[0394]	잠재적인 분할 결과: 1 및 282
[0395]	잠재적인 분할 결과: 1 및 281
[0396]	잠재적인 분할 결과: 271 및 10
[0397]	분할 점수 38.04
[0398]	잠재적인 분할 결과: 2 및 269
[0399]	분할 점수 233.23
[0400]	잠재적인 분할 결과: 1 및 1
[0401]	해체(서로 너무 먼 클러스터 내의 점) 잠재적인 분할 결과: 1 및 268
[0402]	잠재적인 분할 결과: 265 및 3
[0403]	분할 점수 80.24
[0404]	잠재적인 분할 결과: 4 및 261
[0405]	분할 점수 100.26
[0406]	잠재적인 분할 결과: 1 및 3
[0407]	잠재적인 분할 결과: 1 및 2
[0408]	잠재적인 분할 결과: 1 및 1
[0409]	해체(서로 너무 먼 클러스터 내의 점) 잠재적인 분할 결과: 192 및 69
[0410]	분할 점수 9.66
[0411]	잠재적인 분할 결과: 1 및 2
[0412]	잠재적인 분할 결과: 1 및 1
[0413]	해체(서로 너무 먼 클러스터 내의 점) 잠재적인 분할 결과: 2 및 8
[0414]	분할 점수 5.12
[0415]	잠재적인 분할 결과: 1 및 2
[0416]	잠재적인 분할 결과: 1 및 1
[0417]	해체(서로 너무 먼 클러스터 내의 점) 잠재적인 분할 결과: 1 및 397
[0418]	잠재적인 분할 결과: 1 및 396
[0419]	잠재적인 분할 결과: 1 및 395
[0420]	잠재적인 분할 결과: 392 및 3

[0421]	분할 점수 228.58
[0422]	잠재적인 분할 결과: 1 및 391
[0423]	잠재적인 분할 결과: 1 및 390
[0424]	잠재적인 분할 결과: 1 및 389
[0425]	잠재적인 분할 결과: 1 및 388
[0426]	잠재적인 분할 결과: 1 및 387
[0427]	잠재적인 분할 결과: 1 및 386
[0428]	잠재적인 분할 결과: 32 및 354
[0429]	분할 점수 33.24
[0430]	잠재적인 분할 결과: 1 및 31
[0431]	잠재적인 분할 결과: 1 및 30
[0432]	잠재적인 분할 결과: 21 및 9
[0433]	분할 점수 7.20
[0434]	잠재적인 분할 결과: 1 및 353
[0435]	잠재적인 분할 결과: 1 및 352
[0436]	잠재적인 분할 결과: 1 및 351
[0437]	잠재적인 분할 결과: 19 및 332
[0438]	분할 점수 32.86
[0439]	잠재적인 분할 결과: 1 및 18
[0440]	잠재적인 분할 결과: 3 및 15
[0441]	분할 점수 8.90
[0442]	잠재적인 분할 결과: 6 및 326
[0443]	분할 점수 83.57
[0444]	잠재적인 분할 결과: 1 및 5
[0445]	잠재적인 분할 결과: 1 및 4
[0446]	잠재적인 분할 결과: 2 및 2
[0447]	분할 점수 1.32
[0448]	해체(서로 너무 먼 클러스터 내의 점) 잠재적인 분할 결과: 3 및 323
[0449]	분할 점수 148.25
[0450]	잠재적인 분할 결과: 1 및 2
[0451]	잠재적인 분할 결과: 1 및 1
[0452]	해체(서로 너무 먼 클러스터 내의 점) 잠재적인 분할 결과: 314 및 9
[0453]	분할 점수 71.43
[0454]	잠재적인 분할 결과: 221 및 93
[0455]	분할 점수 48.70
[0456]	잠재적인 분할 결과: 1 및 220

- [0457] 잠재적인 분할 결과: 1 및 219
- [0458] 잠재적인 분할 결과: 1 및 218
- [0459] 잠재적인 분할 결과: 1 및 217
- [0460] 잠재적인 분할 결과: 215 및 2
- [0461] 분할 점수 133.42
- [0462] 잠재적인 분할 결과: 166 및 49
- [0463] 분할 점수 7.64
- [0464] 잠재적인 분할 결과: 1 및 1
- [0465] 잠재적인 분할 결과: 40 및 53
- [0466] 분할 점수 9.31
- [0467] 잠재적인 분할 결과: 4 및 5
- [0468] 분할 점수 3.20
- [0469] 해체(서로 너무 먼 클러스터 내의 점) 잠재적인 분할 결과: 1 및 2
- [0470] 잠재적인 분할 결과: 1 및 1
- [0471] 해체(서로 너무 먼 클러스터 내의 점) 잠재적인 분할 결과: 1 및 2
- [0472] 잠재적인 분할 결과: 1 및 1
- [0473] 해체(서로 너무 먼 클러스터 내의 점)
- [0474] 점수 함수가 호출되는 횟수: 40
- [0475] 총 계산 시간은 9.532초였다
- [0476] 병합 단계는 상기 분할 절차에 의해 생성되는 모든 클러스터의 쌍별 비교를 행하는 것을 포함하였다. 충분히 상이하지 않았던 클러스터를 2개의 가장 유사한 클러스터로 시작하여, 함께 병합하였다. 분할 단계와 같이, 병합 단계는 둘 모두의 표지(길이-N) 및 중간 단계의 history를 반환하였다. 표지는 정수였다. 이상점은 '-1'로 표지된다. 커뮤니티 검출에 기초한 병합 단계에 대한 대안적인 방법에 대해서는 하기를 참조한다.
- [0477] [4]에서: # 클러스터 표지를 병합한다
- [0478] `ym,mhistory = merge.dendromerge((D,X),ys,score_threshold=10, preprocessing='precomputed',verbose=True,outlier_threshold_percentile=90)`
- [0479] 886개의 샘플 중 0개가 싱글톤이다
- [0480] 이상점 임계값은 0.51이다.
- [0481] 821의 가장 가까운 이웃: 클러스터 76 내의 72(D = 0.375)
- [0482] 661의 가장 가까운 이웃: 클러스터 76 내의 29(D = 0.379)
- [0483] 729의 가장 가까운 이웃: 클러스터 76 내의 281(D = 0.381)
- [0484] 559의 가장 가까운 이웃: 클러스터 76 내의 79(D = 0.381)
- [0485] 690의 가장 가까운 이웃: 클러스터 76 내의 171(D = 0.381)
- [0486] 564의 가장 가까운 이웃: 클러스터 76 내의 79(D = 0.381)
- [0487] 776의 가장 가까운 이웃: 클러스터 38 내의 474(D = 0.387)
- [0488] 860의 가장 가까운 이웃: 클러스터 38 내의 340(D = 0.390)
- [0489] 816의 가장 가까운 이웃: 클러스터 78 내의 379(D = 0.390)

[0490]	787의 가장 가까운 이웃: 클러스터 38 내의 63(D = 0.391)
[0491]	737의 가장 가까운 이웃: 클러스터 76 내의 72(D = 0.392)
[0492]	874의 가장 가까운 이웃: 클러스터 76 내의 220(D = 0.392)
[0493]	743의 가장 가까운 이웃: 클러스터 76 내의 72(D = 0.394)
[0494]	877의 가장 가까운 이웃: 클러스터 76 내의 174(D = 0.394)
[0495]	753의 가장 가까운 이웃: 클러스터 76 내의 190(D = 0.397)
[0496]	774의 가장 가까운 이웃: 클러스터 38 내의 158(D = 0.398)
[0497]	565의 가장 가까운 이웃: 클러스터 76 내의 190(D = 0.399)
[0498]	785의 가장 가까운 이웃: 클러스터 76 내의 79(D = 0.401)
[0499]	706의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.403)
[0500]	829의 가장 가까운 이웃: 클러스터 38 내의 213(D = 0.404)
[0501]	701의 가장 가까운 이웃: 클러스터 76 내의 179(D = 0.404)
[0502]	770의 가장 가까운 이웃: 클러스터 38 내의 453(D = 0.404)
[0503]	630의 가장 가까운 이웃: 클러스터 76 내의 79(D = 0.406)
[0504]	866의 가장 가까운 이웃: 클러스터 38 내의 87(D = 0.407)
[0505]	795의 가장 가까운 이웃: 클러스터 76 내의 159(D = 0.407)
[0506]	865의 가장 가까운 이웃: 클러스터 76 내의 179(D = 0.407)
[0507]	869의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.409)
[0508]	830의 가장 가까운 이웃: 클러스터 38 내의 165(D = 0.412)
[0509]	851의 가장 가까운 이웃: 클러스터 76 내의 29(D = 0.412)
[0510]	782의 가장 가까운 이웃: 클러스터 76 내의 50(D = 0.412)
[0511]	627의 가장 가까운 이웃: 클러스터 76 내의 72(D = 0.412)
[0512]	848의 가장 가까운 이웃: 클러스터 76 내의 83(D = 0.413)
[0513]	883의 가장 가까운 이웃: 클러스터 12 내의 687(D = 0.413)
[0514]	793의 가장 가까운 이웃: 클러스터 76 내의 107(D = 0.414)
[0515]	631의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.416)
[0516]	720의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.418)
[0517]	885의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.418)
[0518]	813의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.419)
[0519]	788의 가장 가까운 이웃: 클러스터 38 내의 278(D = 0.420)
[0520]	748의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.422)
[0521]	762의 가장 가까운 이웃: 클러스터 38 내의 158(D = 0.423)
[0522]	804의 가장 가까운 이웃: 클러스터 18 내의 177(D = 0.425)
[0523]	854의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.426)
[0524]	605의 가장 가까운 이웃: 클러스터 76 내의 159(D = 0.437)
[0525]	849의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.437)

[0526]	835의 가장 가까운 이웃: 클러스터 18 내의 101($D = 0.438$)
[0527]	790의 가장 가까운 이웃: 클러스터 76 내의 32($D = 0.442$)
[0528]	744의 가장 가까운 이웃: 클러스터 38 내의 188($D = 0.448$)
[0529]	822의 가장 가까운 이웃: 클러스터 38 내의 282($D = 0.449$)
[0530]	723의 가장 가까운 이웃: 클러스터 76 내의 170($D = 0.456$)
[0531]	884의 가장 가까운 이웃: 클러스터 18 내의 101($D = 0.459$)
[0532]	563의 가장 가까운 이웃: 클러스터 76 내의 34($D = 0.463$)
[0533]	867의 가장 가까운 이웃: 클러스터 18 내의 160($D = 0.463$)
[0534]	771의 가장 가까운 이웃: 클러스터 76 내의 34($D = 0.473$)
[0535]	826의 가장 가까운 이웃: 클러스터 38 내의 165($D = 0.475$)
[0536]	777의 가장 가까운 이웃: 클러스터 76 내의 174($D = 0.478$)
[0537]	759의 가장 가까운 이웃: 클러스터 18 내의 101($D = 0.483$)
[0538]	855의 가장 가까운 이웃: 클러스터 18 내의 101($D = 0.485$)
[0539]	702의 가장 가까운 이웃: 클러스터 18 내의 160($D = 0.492$)
[0540]	750의 가장 가까운 이웃: 클러스터 76 내의 230($D = 0.495$)
[0541]	704의 가장 가까운 이웃: 클러스터 78 내의 216($D = 0.497$)
[0542]	711의 가장 가까운 이웃: 클러스터 76 내의 55($D = 0.502$)
[0543]	708의 가장 가까운 이웃: 클러스터 78 내의 537($D = 0.510$)
[0544]	791의 가장 가까운 이웃: 클러스터 76 내의 115($D = 0.534$)
[0545]	722의 가장 가까운 이웃: 클러스터 76 내의 15($D = 0.547$)
[0546]	700의 가장 가까운 이웃: 클러스터 76 내의 107($D = 0.549$)
[0547]	846의 가장 가까운 이웃: 클러스터 76 내의 72($D = 0.552$)
[0548]	876의 가장 가까운 이웃: 클러스터 76 내의 85($D = 0.560$)
[0549]	868의 가장 가까운 이웃: 클러스터 78 내의 740($D = 0.562$)
[0550]	569의 가장 가까운 이웃: 클러스터 76 내의 68($D = 0.572$)
[0551]	817의 가장 가까운 이웃: 클러스터 76 내의 56($D = 0.582$)
[0552]	798의 가장 가까운 이웃: 클러스터 38 내의 310($D = 0.585$)
[0553]	717의 가장 가까운 이웃: 클러스터 78 내의 216($D = 0.597$)
[0554]	879의 가장 가까운 이웃: 클러스터 76 내의 209($D = 0.612$)
[0555]	727의 가장 가까운 이웃: 클러스터 76 내의 96($D = 0.616$)
[0556]	828의 가장 가까운 이웃: 클러스터 38 내의 142($D = 0.618$)
[0557]	840의 가장 가까운 이웃: 클러스터 78 내의 632($D = 0.640$)
[0558]	747의 가장 가까운 이웃: 클러스터 76 내의 202($D = 0.698$)
[0559]	842의 가장 가까운 이웃: 클러스터 38 내의 797($D = 0.703$)
[0560]	442의 가장 가까운 이웃: 클러스터 78 내의 336($D = 0.735$)
[0561]	총 이상점의 수: 18

- [0562] 싱글톤 할당(0.052 s)
- [0563] Dc 생성(13.181 s)
- [0564] 병합 이전: 14개 클러스터
- [0565] 거리 3.60을 갖는 표지 0(N = 10) 및 6(N = 15)의 병합
- [0566] 병합 이전: 13개 클러스터
- [0567] 거리 4.31을 갖는 표지 2(N = 15) 및 4(N = 10)의 병합 병합 이전: 12개 클러스터
- [0568] 거리 4.37을 갖는 표지 1(N = 13) 및 11(N = 25)의 병합 병합 이전: 11개 클러스터
- [0569] 거리 5.23을 갖는 표지 0(N = 25) 및 10(N = 38)의 병합 병합 이전: 10개 클러스터
- [0570] 거리 6.04를 갖는 표지 3(N = 30) 및 7(N = 95)의 병합 병합 이전: 9개 클러스터
- [0571] 거리 6.81을 갖는 표지 2(N = 10) 및 5(N = 2)의 병합 병합 이전: 8개 클러스터
- [0572] 거리 7.19를 갖는 표지 4(N = 25) 및 5(N = 63)의 병합 병합 이전: 7개 클러스터
- [0573] 거리 7.23을 갖는 표지 2(N = 18) 및 5(N = 12)의 병합 병합 이전: 6개 클러스터
- [0574] 거리 9.76을 갖는 표지 3(N = 125) 및 5(N = 30)의 병합 클러스터의 병합에 25.977초가 걸림
- [0575] 커뮤니티 검출에 기초한 병합은 networkx 및 community 파이썬 모듈을 사용하였다. 여기서 반환된 히스토리 데이터 구조는 오직 입력 표지 및 싱글톤-처리-후 표지를 함유한다.
- [0576] [5]에서: `ym_community = merge.dendromerge((D,X),ys, preprocessing='precomputed', verbose=True,outlier_threshold_percentile=90, perform_community_detection=True)`
- [0577] 886개의 샘플 중 80개는 싱글톤이다 이상점 임계값은 0.51이다
- [0578] 821의 가장 가까운 이웃: 클러스터 76 내의 72(D = 0.375)
- [0579] 661의 가장 가까운 이웃: 클러스터 76 내의 29(D = 0.379)
- [0580] 729의 가장 가까운 이웃: 클러스터 76 내의 281(D = 0.381)
- [0581] 559의 가장 가까운 이웃: 클러스터 76 내의 79(D = 0.381)
- [0582] 690의 가장 가까운 이웃: 클러스터 76 내의 171(D = 0.381)
- [0583] 564의 가장 가까운 이웃: 클러스터 76 내의 79(D = 0.381)
- [0584] 776의 가장 가까운 이웃: 클러스터 38 내의 474(D = 0.387)
- [0585] 860의 가장 가까운 이웃: 클러스터 38 내의 340(D = 0.390)
- [0586] 816의 가장 가까운 이웃: 클러스터 78 내의 379(D = 0.390)
- [0587] 787의 가장 가까운 이웃: 클러스터 38 내의 63(D = 0.391)
- [0588] 737의 가장 가까운 이웃: 클러스터 76 내의 72(D = 0.392)
- [0589] 874의 가장 가까운 이웃: 클러스터 76 내의 220(D = 0.392)
- [0590] 743의 가장 가까운 이웃: 클러스터 76 내의 72(D = 0.394)
- [0591] 877의 가장 가까운 이웃: 클러스터 76 내의 174(D = 0.394)
- [0592] 753의 가장 가까운 이웃: 클러스터 76 내의 190(D = 0.397)
- [0593] 774의 가장 가까운 이웃: 클러스터 38 내의 158(D = 0.398)
- [0594] 565의 가장 가까운 이웃: 클러스터 76 내의 190(D = 0.399)
- [0595] 785의 가장 가까운 이웃: 클러스터 76 내의 79(D = 0.401)

[0596]	706의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.403)
[0597]	829의 가장 가까운 이웃: 클러스터 38 내의 213(D = 0.404)
[0598]	701의 가장 가까운 이웃: 클러스터 76 내의 179(D = 0.404)
[0599]	770의 가장 가까운 이웃: 클러스터 38 내의 453(D = 0.404)
[0600]	630의 가장 가까운 이웃: 클러스터 76 내의 79(D = 0.406)
[0601]	866의 가장 가까운 이웃: 클러스터 38 내의 87(D = 0.407)
[0602]	795의 가장 가까운 이웃: 클러스터 76 내의 159(D = 0.407)
[0603]	865의 가장 가까운 이웃: 클러스터 76 내의 179(D = 0.407)
[0604]	869의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.409)
[0605]	830의 가장 가까운 이웃: 클러스터 38 내의 165(D = 0.412)
[0606]	851의 가장 가까운 이웃: 클러스터 76 내의 29(D = 0.412)
[0607]	782의 가장 가까운 이웃: 클러스터 76 내의 50(D = 0.412)
[0608]	627의 가장 가까운 이웃: 클러스터 76 내의 72(D = 0.412)
[0609]	848의 가장 가까운 이웃: 클러스터 76 내의 83(D = 0.413)
[0610]	883의 가장 가까운 이웃: 클러스터 12 내의 687(D = 0.413)
[0611]	793의 가장 가까운 이웃: 클러스터 76 내의 107(D = 0.414)
[0612]	631의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.416)
[0613]	720의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.418)
[0614]	885의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.418)
[0615]	813의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.419)
[0616]	788의 가장 가까운 이웃: 클러스터 38 내의 278(D = 0.420)
[0617]	748의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.422)
[0618]	762의 가장 가까운 이웃: 클러스터 38 내의 158(D = 0.423)
[0619]	804의 가장 가까운 이웃: 클러스터 18 내의 177(D = 0.425)
[0620]	854의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.426)
[0621]	605의 가장 가까운 이웃: 클러스터 76 내의 159(D = 0.437)
[0622]	849의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.437)
[0623]	835의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.438)
[0624]	790의 가장 가까운 이웃: 클러스터 76 내의 32(D = 0.442)
[0625]	744의 가장 가까운 이웃: 클러스터 38 내의 188(D = 0.448)
[0626]	822의 가장 가까운 이웃: 클러스터 38 내의 282(D = 0.449)
[0627]	723의 가장 가까운 이웃: 클러스터 76 내의 170(D = 0.456)
[0628]	884의 가장 가까운 이웃: 클러스터 18 내의 101(D = 0.459)
[0629]	563의 가장 가까운 이웃: 클러스터 76 내의 34(D = 0.463)
[0630]	867의 가장 가까운 이웃: 클러스터 18 내의 160(D = 0.463)
[0631]	771의 가장 가까운 이웃: 클러스터 76 내의 34(D = 0.473)

[0632]	826의 가장 가까운 이웃: 클러스터 38 내의 165($D = 0.475$)
[0633]	777의 가장 가까운 이웃: 클러스터 76 내의 174($D = 0.478$)
[0634]	759의 가장 가까운 이웃: 클러스터 18 내의 101($D = 0.483$)
[0635]	855의 가장 가까운 이웃: 클러스터 18 내의 101($D = 0.485$)
[0636]	702의 가장 가까운 이웃: 클러스터 18 내의 160($D = 0.492$)
[0637]	750의 가장 가까운 이웃: 클러스터 76 내의 230($D = 0.495$)
[0638]	704의 가장 가까운 이웃: 클러스터 78 내의 216($D = 0.497$)
[0639]	711의 가장 가까운 이웃: 클러스터 76 내의 55($D = 0.502$)
[0640]	708의 가장 가까운 이웃: 클러스터 78 내의 537($D = 0.510$)
[0641]	791의 가장 가까운 이웃: 클러스터 76 내의 115($D = 0.534$)
[0642]	722의 가장 가까운 이웃: 클러스터 76 내의 15($D = 0.547$)
[0643]	700의 가장 가까운 이웃: 클러스터 76 내의 107($D = 0.549$)
[0644]	846의 가장 가까운 이웃: 클러스터 76 내의 72($D = 0.552$)
[0645]	876의 가장 가까운 이웃: 클러스터 76 내의 85($D = 0.560$)
[0646]	868의 가장 가까운 이웃: 클러스터 78 내의 740($D = 0.562$)
[0647]	569의 가장 가까운 이웃: 클러스터 76 내의 68($D = 0.572$)
[0648]	817의 가장 가까운 이웃: 클러스터 76 내의 56($D = 0.582$)
[0649]	798의 가장 가까운 이웃: 클러스터 38 내의 310($D = 0.585$)
[0650]	717의 가장 가까운 이웃: 클러스터 78 내의 216($D = 0.597$)
[0651]	879의 가장 가까운 이웃: 클러스터 76 내의 209($D = 0.612$)
[0652]	727의 가장 가까운 이웃: 클러스터 76 내의 96($D = 0.616$)
[0653]	828의 가장 가까운 이웃: 클러스터 38 내의 142($D = 0.618$)
[0654]	840의 가장 가까운 이웃: 클러스터 78 내의 632($D = 0.640$)
[0655]	747의 가장 가까운 이웃: 클러스터 76 내의 202($D = 0.698$)
[0656]	842의 가장 가까운 이웃: 클러스터 38 내의 797($D = 0.703$)
[0657]	442의 가장 가까운 이웃: 클러스터 78 내의 336($D = 0.735$)
[0658]	총 이상점의 수: 18
[0659]	싱글톤 할당(0.054 s)
[0660]	Dc 생성(12.773 s)
[0661]	그래프를 14개의 노드 및 24개의 간선으로 구축(12.774 s)
[0662]	클러스터의 병합에는 12.775초가 걸림
[0663]	전체적으로, 이들 데이터는 재귀적 분할 및 시험에 이어서 병합에 의한 클러스터링을 보여준다.
[0664]	<u>실시예 2</u>
[0665]	<u>재귀적 덴드로그램 분할 및 시험에 이어서 병합에 의한 클러스터링의 결과의 시각화</u>
[0666]	본 실시예는 실시예 1에 예시된 재귀적 덴드로그램 분할 및 시험에 이어서 병합에 의한 클러스터링의 결과의 시각화를 설명한다.

[0667] 분할 및 병합 단계 후에 생성된 클러스터 표지를 시험하였다.

[0668] [6]에서: `plt.scatter(x1,x2,edgecolors='none')`

[0669] `_ = plt.axis('off')`

[0670] `plt.title('Pre-clustering')`

[0671] # 사전-병합된 표지를 사용한 클러스터링 결과(싱글톤 표지)

[0672] `plt.figure()`

[0673] `split.plot_labels_legend(x1,x2,split.str_labels_to_ints(ys))`

[0674] `plt.title('After splitting step')`

[0675] # 이후-병합된 표지를 사용한 클러스터링 결과

[0676] `plt.figure()`

[0677] `split.plot_labels_legend(x1,x2,ym)`

[0678] `plt.title('After merging step')`

[0679] # 이후-병합된 표지를 사용한 클러스터링 결과

[0680] `plt.figure()`

[0681] `split.plot_labels_legend(x1,x2,ym_community)`

[0682] `plt.title('After merging step using community detection')`

[0683] Out[6]: <matplotlib.text.Text at 0x112674510>

[0684] 도 8의 패널 (a) 내지 (d)는 단일 세포의 발현 프로파일의 분할 및 병합 이후 2차원 공간에서의 발현 프로파일의 비제한적인 예시적 플롯을 보여준다.

[0685] 분할이 어떻게 결정되었는지의 탐색

[0686] "print_history" 함수를 사용하여, 방법의 분할 단계가 "print_history" 함수를 사용하여 데이터셋을 어떻게 다루는지를 탐색할 수 있다. 줄 i 는 i 번째의 유효한 분할을 설명한다. 생성된 둘 모두의 클러스터가 "min_clust_size"를 초과하며, 분할로부터 생성된 가장 낮은 p-값이 임계값 미만이면, 분할은 유효한 것으로 간주된다.

[0687] [7]에서: `split.print_history(genes,shistory)`

[0688] 사전-분할: 886 L: 883 R: 3 점수: 1.8E+308 상위 유전자: RPL31 상위 유전자 점수: 1.8E+308

[0689] 사전-분할: 882 L: 484 R: 398 점수: 182.26 상위 유전자: FTL 상위 유전자 점수: 182.26

[0690] 사전-분할: 484 L: 481 R: 3 점수: 1.8E+308 상위 유전자: RPL23 상위 유전자 점수: 1.8E+308

[0691] 사전-분할: 479 L: 195 R: 284 점수: 125.49 상위 유전자: IGHM 상위 유전자 점수: 125.49

[0692] 사전-분할: 195 L: 177 R: 18 점수: 15.35 상위 유전자: RRP7A 상위 유전자 점수: 15.35

[0693] 사전-분할: 174 L: 12 R: 162 점수: 18.88 상위 유전자: ANXA11 상위 유전자 점수: 18.88

[0694] 사전-분할: 160 L: 28 R: 132 점수: 12.32 상위 유전자: TTF1 상위 유전자 점수: 12.32

[0695] 사전-분할: 28 L: 25 R: 3 점수: 13.94 상위 유전자: SRPK1 상위 유전자 점수: 13.94

[0696] 사전-분할: 132 L: 122 R: 10 점수: 18.52 상위 유전자: TOP2A 상위 유전자 점수: 18.52

[0697] 사전-분할: 122 L: 13 R: 109 점수: 24.92 상위 유전자: CACYBP 상위 유전자 점수: 24.92

[0698] 사전-분할: 109 L: 105 R: 4 점수: 31.72 상위 유전자: RPSA 상위 유전자 점수: 31.72

[0699] 사전-분할: 105 L: 15 R: 90 점수: 11.31 상위 유전자: PSMD14 상위 유전자 점수: 11.31

- [0700] 사전-분할: 281 L: 271 R: 10 점수: 38.04 상위 유전자: RNASEH2B 상위 유전자 점수: 38.04
- [0701] 사전-분할: 271 L: 2 R: 269 점수: 233.23 상위 유전자: GAS8 상위 유전자 점수: 233.23
- [0702] 사전-분할: 268 L: 265 R: 3 점수: 80.24 상위 유전자: CNPY3 상위 유전자 점수: 80.24
- [0703] 사전-분할: 265 L: 4 R: 261 점수: 100.26 상위 유전자: MZB1 상위 유전자 점수: 100.26
- [0704] 사전-분할: 395 L: 392 R: 3 점수: 228.58 상위 유전자: CREB3L1 상위 유전자 점수: 228.58
- [0705] 사전-분할: 386 L: 32 R: 354 점수: 33.24 상위 유전자: VMP1 상위 유전자 점수: 33.24
- [0706] 사전-분할: 351 L: 19 R: 332 점수: 32.86 상위 유전자: EIF2B1 상위 유전자 점수: 32.86
- [0707] 사전-분할: 332 L: 6 R: 326 점수: 83.57 상위 유전자: NUDT5 상위 유전자 점수: 83.57
- [0708] 사전-분할: 326 L: 3 R: 323 점수: 148.25 상위 유전자: TMSB4X 상위 유전자 점수: 148.25
- [0709] 사전-분할: 323 L: 314 R: 9 점수: 71.43 상위 유전자: C12orf57 상위 유전자 점수: 71.43
- [0710] 사전-분할: 314 L: 221 R: 93 점수: 48.70 상위 유전자: RPL27A 상위 유전자 점수: 48.70
- [0711] 사전-분할: 217 L: 215 R: 2 점수: 133.42 상위 유전자: JUN 상위 유전자 점수: 133.42
- [0712] 각각의 분할에 수반되었던 점을 시각화하였다. 각각의 행은 2개의 도면을 갖는다. 행 i 는 저장된 i 번째 분할을 설명한다. 도 9a 내지 도 9h의 패널 (a) 내지 (x)는 분할이 어떻게 결정되었는지를 보여주는 2차원 공간에서의 발현 프로파일의 비제한적인 예시적 플롯이다. 각각의 패널에서, 좌측 도면은 분할이 어떻게 수행되었는지를 나타낸다. 청색 점은 분할에 전혀 관련되지 않았다. 적색 및 녹색 점은 동일한 클러스터에 존재한 다음, 분리되었다. 좌측 도면의 제목은 분할 번호 및 (절대 값을 취한 후에) 가장 큰 t 통계량을 달성한 3개의 유전자를 나타낸다. 각각의 유전자와 연관된 수는 상응하는 p 값의 $-\log_{10}$ 이다. 각각의 유전자 옆의 괄호 안의 '0' 또는 '1'은 그 유전자의 더 큰 평균 발현을 가졌던 클러스터를 나타낸다. 우측 도면은 가장 큰 t 통계량을 달성한 유전자의 로그 발현을 보여준다.
- [0713] [8]에서: `split.visualize_history(np.log(1+X),x1,x2,genes,shistory)`
- [0714] /Users/user1/anaconda2/lib/python2.7/site-packages/matplotlib/pyplot.py:516: RuntimeWarning: More max open warning, RuntimeWarning)
- [0715] "analyze_split" 함수를 사용하여, 특정 분할이 유지되었던 이유를 나타내는 유전자를 더 면밀히 볼 수 있다. 분할에 관련되지 않은 세포도 나타내기 위해 "show_background" 키워드를 사용한다. 특정 클러스터에서 더 고도로 발현되는 유전자만 보기 위하여 "clust"를 사용한다. "num_genes"를 사용하여 맞춤 유전자의 수를 나타낼 수 있다.
- [0716] [9]에서: # 분할 5를 관찰한다
- [0717] `split_num = 5`
- [0718] `cluster_of_interest = None`
- [0719] `show_background = False`
- [0720] `split.analyze_split(X,x1,x2,genes,shistory, split_num, num_genes=12,`
- [0721] `show_background=show_background, clust=cluster_of_interest)`
- [0722] 도 10은 제5 사이클의 분할 후의 2차원 공간에서의 발현 프로파일의 비제한적인 예시적 플롯을 보여준다. 도 11의 패널 (a) 내지 (1)은 덴드로그램에서 특정 분할이 도 10에 나타낸 제5 사이클의 분할에 대하여 유지되었던 이유를 보여주는 2차원 공간에서의 발현 프로파일의 비제한적인 예시적 플롯이다.
- [0723] 병합이 어떻게 결정되었는지의 탐색
- [0724] 분할이 어떻게 결정되었는지 탐색하기 위한 함수는 병합이 어떻게 수행되었는지를 탐색하기 위해 사용될 수 있다.
- [0725] [10]에서: `split.print_history(genes,mhistory)`

- [0726] `split.visualize_history(np.log(1+X),x1,x2,genes,mhistory)`
- [0727] 886개의 샘플 중 80개는 싱글톤이다
- [0728] 싱글톤(들) 442, 569, 700, 708, 717, 722, 727, 747, 791, 798, 817, 828, 840, 842, 846, 868, 876, 879 m
- [0729] 싱글톤(들) 15는 클러스터 12(N = 24)와 병합하여 클러스터 1(N = 25)을 형성한다
- [0730] 싱글톤(들) 3, 4, 6, 7, 9, 10, 11, 13, 14, 19, 20, 21, 22, 24, 25, 26은 클러스터 18(N = 90)과 병합한다
- [0731] 싱글톤(들) 5, 29, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41은 클러스터 38(N = 261)과 병합하여
- [0732] 싱글톤(들) 28, 53, 55, 56, 58, 59, 60, 61, 63, 64, 65, 67, 68, 69, 70, 71, 72, 73, 74, 75, 79, 80, 81을 형성한다
- [0733] 싱글톤(들) 54, 66은 클러스터 78(N = 93)과 병합하여 클러스터 13(N = 95)을 형성한다
- [0734] 병합-후: 25 L: 10 R: 15 점수: 3.60 상위 유전자: ENOSF1 상위 유전자 점수: 3.60
- [0735] 병합-후: 25 L: 15 R: 10 점수: 4.31 상위 유전자: MAGED1 상위 유전자 점수: 4.31
- [0736] 병합-후: 38 L: 13 R: 25 점수: 4.37 상위 유전자: PRPF40A 상위 유전자 점수: 4.37
- [0737] 병합-후: 63 L: 25 R: 38 점수: 5.23 상위 유전자: ALDOC 상위 유전자 점수: 5.23
- [0738] 병합-후: 125 L: 30 R: 95 점수: 6.04 상위 유전자: PARP1 상위 유전자 점수: 6.04
- [0739] 병합-후: 12 L: 10 R: 2 점수: 6.81 상위 유전자: IGLC3 상위 유전자 점수: 6.81
- [0740] 병합-후: 88 L: 25 R: 63 점수: 7.19 상위 유전자: HMGB2 상위 유전자 점수: 7.19
- [0741] 병합-후: 30 L: 18 R: 12 점수: 7.23 상위 유전자: VIM 상위 유전자 점수: 7.23
- [0742] 병합-후: 155 L: 125 R: 30 점수: 9.76 상위 유전자: HMGN5 상위 유전자 점수: 9.76
- [0743] 도 12a 내지 도 12c의 패널 (a) 내지 (i)는 병합이 어떻게 결정되었는지를 보여주는 2차원 공간의 발현 프로파일의 비제한적인 예시적 플롯이다.
- [0744] [11]에서: # 병합 2를 관찰한다
- [0745] `merge_num = 2`
- [0746] `cluster_of_interest = None`
- [0747] `show_background = False`
- [0748] `split.analyze_split(X,x1,x2,genes,mhistory, merge_num, num_genes=4,`
- [0749] `show_background=show_background, clust=cluster_of_interest)`
- [0750] 도 13은 제2 사이클의 병합 후에 2차원 공간에서의 발현 프로파일의 비제한적인 예시적 플롯을 보여준다. 도 14의 패널 (a) 내지 (d)는 도 13에 나타난 제2 사이클의 병합이 어떻게 결정되었는지를 보여주는 2차원 공간에서의 발현 프로파일의 비제한적인 예시적 플롯이다.
- [0751] 차등 발현
- [0752] 모듈은 2가지 종류의 간단한 차등 발현 분석을 가능하게 한다. 제1의 것은 각각의 클러스터에 대하여 1-대-나머지 비교를 행하여, 각각의 유전자에 대한 t-검정에 따라 각각의 클러스터에 가장 중요한 유전자를 시각화한다. 제2의 것은 모든 2개의 쌍의 클러스터에 대한 쌍별 비교를 행한다.
- [0753] [12]에서:
- [0754] `split.save_more_highly_expressed_genes_in_one_clust(X,genes,ym,x1,x2,num_genes=3, show_plots=True)`
- [0755] 도 15a 및 도 15b의 패널 (a) 내지 (f)는 하나의 비제한적인 예시적 유형의 차등 발현 분석을 보여주는 플롯이다.
- [0756] [13]에서: `split.pairwise_cluster_comparison(X,genes,ym,x1=x1,x2=x2, num_genes=3,`

show_plots=True, verbose=F

- [0757] dendroplit/utlis.py:39: FutureWarning: elementwise comparison failed; returning scalar instead, but in plt.plot(x1[y==i], x2[y==i], '.', c=RGBs[j], label=str(i)+' ('+str(np.sum(y==i))+')')
- [0758] dendroplit/feature selection.py:221: RuntimeWarning: divide by zero encountered in double scalars
fold = g mean j/g mean i
- [0759] 도 16a 내지 도 16d의 패널 (a) 내지 (o)는 또 다른 비제한적인 예시적 유형의 차등 발현 분석을 보여주는 플롯이다.
- [0760] 거리 분포
- [0761] 또한, 모듈은 사용자가 각각의 클러스터 내의 거리의 분포를 시각화하게 한다. 주어진 클러스터에 있어서, 이러한 함수는 (모든 클러스터의 모든 점에 대한) 전체 세트의 쌍별 거리의 각각의 백분위수 빈(bin)에 대한 쌍별 거리(클러스터 내의 점 사이)의 비를 플롯팅한다. 예를 들어, 1에서 0.3은 쌍별 거리의 30%가 전체 거리의 5 백분위수와 10 백분위수 사이에 속하는 것을 나타낸다. 이러한 함수를 사용하여, 원래의 거리 행렬에 따라 클러스터가 얼마나 응집성인지의 느낌을 얻는다. 직감적으로, 우수한 클러스터는 모두 서로 근접해 있는 점을 가져야 한다. 예를 들어, 하위 10 빈(즉, 하위 50 백분위수)에 거리가 없는 클러스터는 불량으로 간주될 것이다. 예상되는 바와 같이, 이것은 이상점을 포함하는 하기 클러스터 "-1"에 대한 경우인 것을 주목한다.
- [0762] [14]에서: merge.visualize_within_cluster_distance_distributions(D, ym, show_D_dist=True)
- [0763] 도 17의 패널 (a) 내지 (g)는 클러스터 사이의 거리를 시각화하는 비제한적인 예시적 플롯이다.
- [0764] 덴드로그램
- [0765] 모듈은 또한, 사용자가 덴드로그램을 생성하고, 덴드로그램에 따라 셀의 순서를 얻게 한다. 덴드로그램은 아이파이썬(iPython) 노트북에서 보기 어려울 수 있다. 일부 실시형태에서, 덴드로그램은 하기 나타낸 바와 같이 저장될 수 있다. 사용자는 클러스터 표지("labels" 키워드)를 피딩할 수 있다. 필요하다면, 함수는 클러스터 내의 모든 샘플의 명칭을 동일한 색으로 표시할 수 있다.
- [0766] [15]에서: cell_order = split.plot_dendro(D, return_cell_order=True, labels=ym, save_name='/Users/user1/Desktop/dendrogram')
- [0767] dendroplit/split.py:233: FutureWarning: comparison to 'None' will result in an elementwise object comp if labels != None:
- [0768] 도 18a 내지 도 18f는 비제한적인 예시적 덴드로그램을 보여준다.
- [0769] 전체적으로, 이들 데이터는 회귀적 분할 및 병합에 이어서 병합의 다양한 단계 및 결과를 시각화하기 위한 본 발명의 다양한 툴을 보여준다.
- [0770] 실시예 3
- [0771] 회귀적 덴드로그램 분할 및 시험에 이어서 병합에 의한 클러스터링을 위한 파라미터 스위핑
- [0772] 이러한 실시예는 회귀적 분할 및 시험에 이어서 병합을 위하여 파라미터를 최적화하기 위한 파라미터 스위핑을 설명한다.
- [0773] 방법의 분할 단계 동안, 2개의 하이퍼파라미터를 튜닝할 수 있다: 점수 임계값 및 해체 백분위수. 어떤 상이한 클러스터가 상이한 하이퍼파라미터로 생성될 수 있는지를 탐색할 수 있다. 더 작은 점수 임계값으로 생성된 클러스터(더 작은 임계값은 더 많은 클러스터를 초래함)가 더 큰 점수 임계값으로 생성된 클러스터를 분할한다는 사실을 탐색함으로써 몇몇의(분할-후 병합-전) 클러스터링 결과가 신속하게 수득될 수 있다. 먼저, 매우 낮은 임계값으로 분할 단계를 시행한다. 두번째로, history() 함수로부터 얻은 클러스터를 사용한다.
- [0774] 다양한 점수 임계값을 통한 스위핑에 대한 일 예는 하기에 나타나 있다. 이는 해체 백분위수 값으로 행해질 수 있다.
- [0775] [16]에서: ys, shistory = split.dendroplit((D,X), preprocessing='precomputed',
- [0776] score_threshold=2, verbose=False, disband_percentile=50)

[0777] `ys_sweep = []`

[0778] `thresholds = range(5,100,5)`

[0779] `for threshold in thresholds:`

[0780] `ys_sweep.append(split.get_clusters_from_history(D,shistory,threshold,50))`

[0781] `plt.figure()`

[0782] `split.plot_labels_legend(x1,x2,split.str_labels_to_ints(ys_sweep[-1]))`

[0783] `plt.title('Clustering result using a threshold of %.3f'%(threshold))`

[0784] 도 19a 내지 도 19j의 패널 (a) 내지 (s)는 파라미터 스위핑을 보여주는 비제한적인 예시적 플롯이다.

[0785] 파라미터를 스위핑함으로써, 임계값의 함수로서 클러스터의 수가 어떻게 변하는지를 조사할 수 있다. 이는 사용자에게 특정 응용을 위하여 최적의 임계값을 선택하는 방법에 대한 통찰력을 제공할 수 있다.

[0786] [17]에서: `def count_nonsingleton_clusters(y):`

[0787] `return sum([1 for i in np.unique(y) if np.sum(y==i) != 1])`

[0788] `plt.plot(thresholds,[count_nonsingleton_clusters(i) for i in ys_sweep])`

[0789] `plt.grid()`

[0790] `plt.xlabel('thresholds (-log10(p-value))')`

[0791] `plt.ylabel('number of nonsingleton clusters')`

[0792] Out[17]: <matplotlib.text.Text at 0x117fb3290>

[0793] 도 20은 임계값을 확인하기 위하여 어떻게 파라미터 스위핑을 사용할 수 있는지를 보여주는 비제한적인 예시적 플롯이다. 다수의 발현 프로파일의 클러스터가 오직 5의 임계값으로 확인되었기 때문에, 더 적은 발현 프로파일의 클러스터가 더 큰 임계값으로 확인되었다(예를 들어, 도 19e의 패널 (h)에 나타낸 40).

[0794] 전체적으로, 이들 데이터는 회귀적 분할 및 시험에 이어서 병합을 위하여 파라미터 스위핑에 의해 하이퍼파라미터를 최적화시키는 것을 보여준다.

[0795] 실시예 4

[0796] 회귀적 덴드로그램 분할 및 시험에 이어서 병합에 의한 클러스터링

[0797] 이러한 실시예는 회귀적 분할(예를 들어, 회귀적 덴드로그램 분할) 및 시험에 이어서 병합에 의한 클러스터링 방법을 설명한다. 이러한 실시예에서 덴드로그램의 각각의 클러스터 또는 노드(리프 노드 제외)에서, 2개의 서브-클러스터의 클러스터-내 중간값 상관관계는 클러스터-간 중간값 상관관계보다 더 컸다.

[0798] 357가지 세포의 발현 프로파일의 분할 및 시험 단계 동안, 덴드로그램의 상측에서 시작하여, 트리를 2개의 후보 서브트리로 분할하였다. 분할은 2개의 서브-클러스터의 클러스터-내 중간값 상관관계가 클러스터-간 중간값 상관관계보다 더 커야 한다는 제약 하에, 2개의 후보 서브-클러스터로 분할되는 클러스터에 상응하였다. 분할을 질을 점수화하였다. 서브클러스터가 충분히 상이한 것으로 간주되면, 각각의 서브트리로 단계를 계속하였다. 그렇지 않으면, 이러한 덴드로그램의 부분에 대하여 상기 방법을 종료하였다. 이러한 단계는 데이터세트를 위한 표지의 세트를 생성하였다.

[0799] 도 21a 내지 도 21e의 패널 (a) 내지 (j)는 제1 분할의 결과를 보여주는 비제한적인 예시적 플롯이다. 제1 분할 동안, 20가지의 유전자(표 1에 나타냄)가 357가지의 세포에서 상이하게 발현되는 것으로 결정되었다.

표 1

20가지의 유전자가 제1 분할 동안 357가지의 세포에서 상이하게 발현되는 것으로 결정되었다.

분할	유전자	p-값	더 큰 클러스터
1	IGLC3 ENST00000390325.2 Reference_end	201.35	0
1	JCHAIN NM_144646.3 Reference_end	105.57	0
1	ADA NM_000022.3 Reference_end	89.27	1
1	TCL1A NM_021966.2 Reference_end	81.19	0
1	CD74 NM_004355.3 Reference_end	62.65	0
1	CD3D NM_000732.4 Reference_end	50.32	1
1	POU2AF1 NM_006235.2 Reference_end	39.94	0
1	CD52 NM_001803.2 Reference_end	39.1	0
1	QPCT NM_012413.3 Reference_end	38.87	0
1	HLA-DRA NM_019111.4 Reference_end	26.64	0
1	CD22 NM_001771.3 Reference_end	25.96	0
1	IRF8 NM_002163.2 Reference_end	21.25	0
1	MS4A1 NM_021950.3 PolyA_1	19.99	0
1	CD37 NM_001774.2 Reference_end	18.49	0
1	LEF1 NM_016269.4 Reference_end	17.63	1
1	MME NM_000902.3 Reference_end	15.59	0
1	BCL6 NM_001706.4 Reference_end	13.39	0
1	CD27 NM_001242.4 Reference_end	11.02	0
1	IL32 NM_004221.4 Reference_end	10.86	1
1	CD38 NM_001775.3 Reference_end	10.65	0

[0800]

[0801]

도 22는 357가지의 세포의 발현 프로파일의 분할 결과를 예시하는 비제한적인 예시적 t-분포 확률적 이웃 임베딩(t-SNE) 플롯이며, 이는 357가지의 세포가 10의 임계값을 사용하여 2개의 클러스터로 분류되었음을 보여준다. 도 23a 내지 도 23f는 표 2에 나타난 특징에 기초하여, 2개의 클러스터로 분류된 발현 프로파일을 보여주는 비제한적인 예시적 텐드로그램을 보여준다(표 1의 클러스터 0은 표 2의 클러스터 1에 상응하며, 표 2의 클러스터 1은 표 2의 클러스터 2에 상응한다). 도 24는 파라미터 스위핑을 보여주는 비제한적인 예시적 플롯이다. 2개의 발현 프로파일의 클러스터가 오직 10의 임계값을 사용하여 확인되었기 때문에, 동일한 2개의 발현 프로파일의 클러스터가 더 큰 임계값을 사용하여 확인되었다(도 24를 도 20과 비교).

표 2

p-값에 의해 순서화된 2개의 클러스터의 쌍별 클러스터 특징.

비교	유전자	p-값	더 큰 클러스터	더 큰 클러스터에 대한 발현의 배수 변화
클러스터1 대 클러스터2	IGLC3 ENST00000390325.2 Reference_end	201.35	1	183.947
클러스터1 대 클러스터2	JCHAIN NM_144646.3 Reference_end	105.572	1	50.085
클러스터1 대 클러스터2	ADA NM_000022.3 Reference_end	89.274	2	11.82
클러스터1 대 클러스터2	TCL1A NM_021966.2 Reference_end	81.191	1	134.689
클러스터1 대 클러스터2	CD74 NM_004355.3 Reference_end	62.653	1	173.268
클러스터1 대 클러스터2	CD3D NM_000732.4 Reference_end	50.32	2	109.703
클러스터1 대 클러스터2	POU2AF1 NM_006235.2 Reference_end	39.943	1	19.778
클러스터1 대 클러스터2	CD52 NM_001803.2 Reference_end	39.105	1	50.988
클러스터1 대 클러스터2	QPCT NM_012413.3 Reference_end	38.87	1	INF
클러스터1 대 클러스터2	HLA-DRA NM_019111.4 Reference_end	26.642	1	71.744
클러스터1 대 클러스터2	CD22 NM_001771.3 Reference_end	25.96	1	64.976
클러스터1 대 클러스터2	IRF8 NM_002163.2 Reference_end	21.245	1	INF
클러스터1 대 클러스터2	MS4A1 NM_021950.3 PolyA_1	19.993	1	INF
클러스터1 대 클러스터2	CD37 NM_001774.2 Reference_end	18.492	1	INF
클러스터1 대 클러스터2	LEF1 NM_016269.4 Reference_end	17.63	2	4.807
클러스터1 대 클러스터2	MME NM_000902.3 Reference_end	15.585	1	37.451
클러스터1 대 클러스터2	BCL6 NM_001706.4 Reference_end	13.393	1	INF
클러스터1 대 클러스터2	CD27 NM_001242.4 Reference_end	11.018	1	INF
클러스터1 대 클러스터2	IL32 NM_004221.4 Reference_end	10.862	2	INF
클러스터1 대 클러스터2	CD38 NM_001775.3 Reference_end	10.649	1	INF

[0802]

[0803]

전체적으로, 이들 데이터는 회귀적 분할 및 시험에 이어서 병합에 의한 클러스터링을 보여준다. 이러한 예에서, 덴드로그램의 각각의 클러스터 또는 노드(리프 노드 제외)에서, 2개의 서브-클러스터의 클러스터-내 중간값 상관관계는 클러스터-간 중간값 상관관계보다 더 컸다.

[0804]

전술된 실시형태 중 적어도 일부에서, 일 실시형태에서 사용되는 하나 이상의 요소는, 이러한 대체가 기술적으로 실현 불가능한 것이 아닌 한, 또 다른 실시형태에서 상호 교환 가능하게 사용될 수 있다. 당업자는, 청구되는 본 발명의 대상의 범위를 벗어남이 없이, 상기 기재된 방법 및 구조에 다양한 다른 생략, 첨가 및 변형이 이루어질 수 있음을 이해할 것이다. 이러한 모든 변형 및 변화는, 첨부된 청구범위에 의해 정의된 바와 같은 본 발명의 대상의 범위 내에 속하는 것으로 본다.

[0805]

본원에서 실질적으로 임의의 복수형 및/또는 단수형 용어의 사용과 관련하여, 당업자는 맥락 및/또는 적용상 적절한 것으로서, 복수형을 단수형으로 번역하고/거나, 단수형을 복수형으로 번역할 수 있다. 다양한 단수형/복수형 교체는 명료성을 위해, 본원에 명백하게 제시될 수 있다. 본 명세서 및 첨부된 청구범위에 사용되는 바와 같이, 단수형("a", "an" 및 "the")은 문맥에서 명백하게 달리 나타내지 않는 한, 복수의 언급대상을 포함한다. 본원에서 "또는"에 대한 임의의 언급은 달리 나타내지 않는 한, "및/또는"을 포함하는 것으로 의도된다.

[0806]

일반적으로, 본원, 특히 첨부된 청구범위(예를 들어, 첨부된 청구범위의 바디)에 사용되는 용어가 일반적으로

"개방형" 용어(예를 들어, 용어 "~를 포함하는"은 "~를 포함하지만, 이에 제한되지 않는"으로 해석되어야 하고, 용어 "~를 갖는"은 "적어도 ~를 갖는"으로 해석되어야 하며, 용어 "~를 포함한다"는 "~를 포함하지만, 이에 제한되지 않는다"로 해석되는 등의 방식)로 의도됨이 당업자에 의해 이해될 것이다. 특정 수의 도입된 청구 범위 열거대상이 의도된 경우, 이러한 의도는 청구범위에 명시적으로 열거될 것이며, 이러한 열거가 존재하지 않는 경우, 이러한 의도는 존재하지 않는다는 것이 당업자에 의해 추가로 이해될 것이다. 예를 들어, 이해를 돕고자 하는 것으로서, 하기에 첨부된 청구범위는 청구되는 열거대상을 도입하기 위해, 도입구 "적어도 하나의" 및 "하나 이상의"의 사용을 포함할 수 있다. 그러나, 이러한 어구의 사용은, 청구되는 열거대상에 부정관사 단수형("a" 또는 "an")이 도입된 것이, 심지어, 동일한 청구항이 도입구 "하나 이상의" 또는 "적어도 하나의" 및 부정관사, 예컨대 단수형("a" 또는 "an")을 포함하는 경우에도, 청구되는 이러한 도입 열거대상을 함유하는 임의의 특정 청구항을, 이러한 열거대상을 단지 하나만 함유하는 실시형태로 제한하는 것을 함축하는 것으로 이해되어서는 안되며(예를 들어, 단수형("a" 및/또는 "an")은 "적어도 하나의" 또는 "하나 이상의"를 의미하는 것으로 해석되어야 함); 이는 청구되는 열거대상을 도입하기 위해 정관사가 사용된 경우에도 마찬가지이다. 이에 더하여, 심지어, 특정 수의 청구되는 도입 열거대상이 분명하게 열거된 경우에도, 당업자는, 이러한 열거대상이 적어도 열거된 수만큼임을 의미하는 것으로 해석되어야 함을 인식할 것이다(예를 들어, 다른 수식어가 없이, "2개의 열거대상"이 다른 어구 없이 열거된 경우, 이는 적어도 2개의 열거대상 또는 2개 이상의 열거대상을 의미함). 추가로, "A, B, 및 C 중 적어도 하나 등"과 유사한 항목이 사용된 예에서, 일반적으로 이러한 구성은, 당업자가 관계를 이해하는 개념으로 의도된다(예를 들어, A, B, 및 C 중 적어도 하나를 갖는 시스템"은, A만, B만, C만, A 및 B 모두, A 및 C 모두, B 및 C 모두 및/또는 A, B 및 C 모두를 갖는 시스템 등을 포함하지만, 이에 제한되지 않음). "A, B, 또는 C 중 적어도 하나 등"과 유사한 관계가 사용된 예에서, 일반적으로 이러한 구성은, 당업자가 이러한 관계를 이해하는 개념으로 의도된다(예를 들어, A, B, 또는 C 중 적어도 하나를 갖는 시스템"은, A만, B만, C만, A 및 B 모두, A 및 C 모두, B 및 C 모두 및/또는 A, B 및 C 모두를 갖는 시스템 등을 포함하지만, 이에 제한되지 않음). 사실상, 2개 이상의 대체 용어를 제시하는 임의의 이원적 단어 및/또는 어구는, 본 기재내용, 청구범위 또는 도면의 존재 여부에 상관없이, 이러한 용어 중 하나, 이러한 용어 중 어느 하나 또는 두 용어 모두를 포함할 가능성을 고려하는 것으로 이해되어야 함이 당업자에 의해 추가로 이해될 것이다. 예를 들어, 어구 "A 또는 B"는 "A" 또는 "B" 또는 "A 및 B"의 가능성을 포함하는 것으로 이해될 것이다.

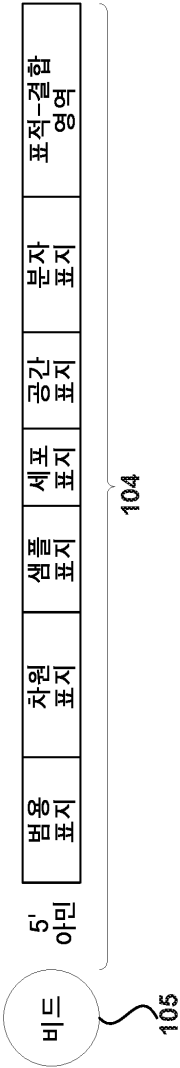
[0807] 이에 더하여, 본 발명의 특성 또는 양태가 마쿠쉬 그룹(Markush group)의 관점으로 기재된 경우, 당업자는, 본 발명이 또한 이에 의해, 마쿠쉬 그룹의 임의의 개별 구성원 또는 구성원의 하위그룹의 관점으로 기재됨을 인식할 것이다.

[0808] 당업자가 이해하는 바와 같이, 예를 들어, 작성된 기재내용을 제공하는 관점에서, 모든 목적을 위해, 본원에 개시된 모든 범위는 또한, 모든 가능한 하위범위 및 그의 하위범위의 조합을 포함한다. 열거된 임의의 범위는, 적어도 동일 범위의 2분의 1, 3분의 1, 4분의 1, 5분의 1, 10분의 1 등으로의 동일 범위의 세분을 충분히 기재하고, 이를 가능하게 하는 것으로서, 용이하게 인식될 수 있다. 비제한적인 예로서, 본원에 논의된 각각의 범위는 하위 3분의 1, 중위 3분의 1 및 상위 3분의 1 등으로 용이하게 세분될 수 있다. 또한, 당업자가 이해하는 바와 같이, "최대", "적어도", "초과", "미만" 등과 같은 모든 용어는 열거된 수를 포함하며, 이어서, 상기에서 논의된 바와 같이, 하위범위로 세분될 수 있는 범위를 지칭한다. 마지막으로, 당업자가 이해하는 바와 같이, 범위는 각각의 개별 구성원을 포함한다. 따라서, 예를 들어, 1 내지 3개의 항목을 갖는 그룹은 1, 2 또는 3개의 항목을 갖는 그룹을 지칭한다. 마찬가지로, 1 내지 5개의 항목을 갖는 그룹은 1, 2, 3, 4, 또는 5개의 항목 등을 갖는 그룹을 지칭한다.

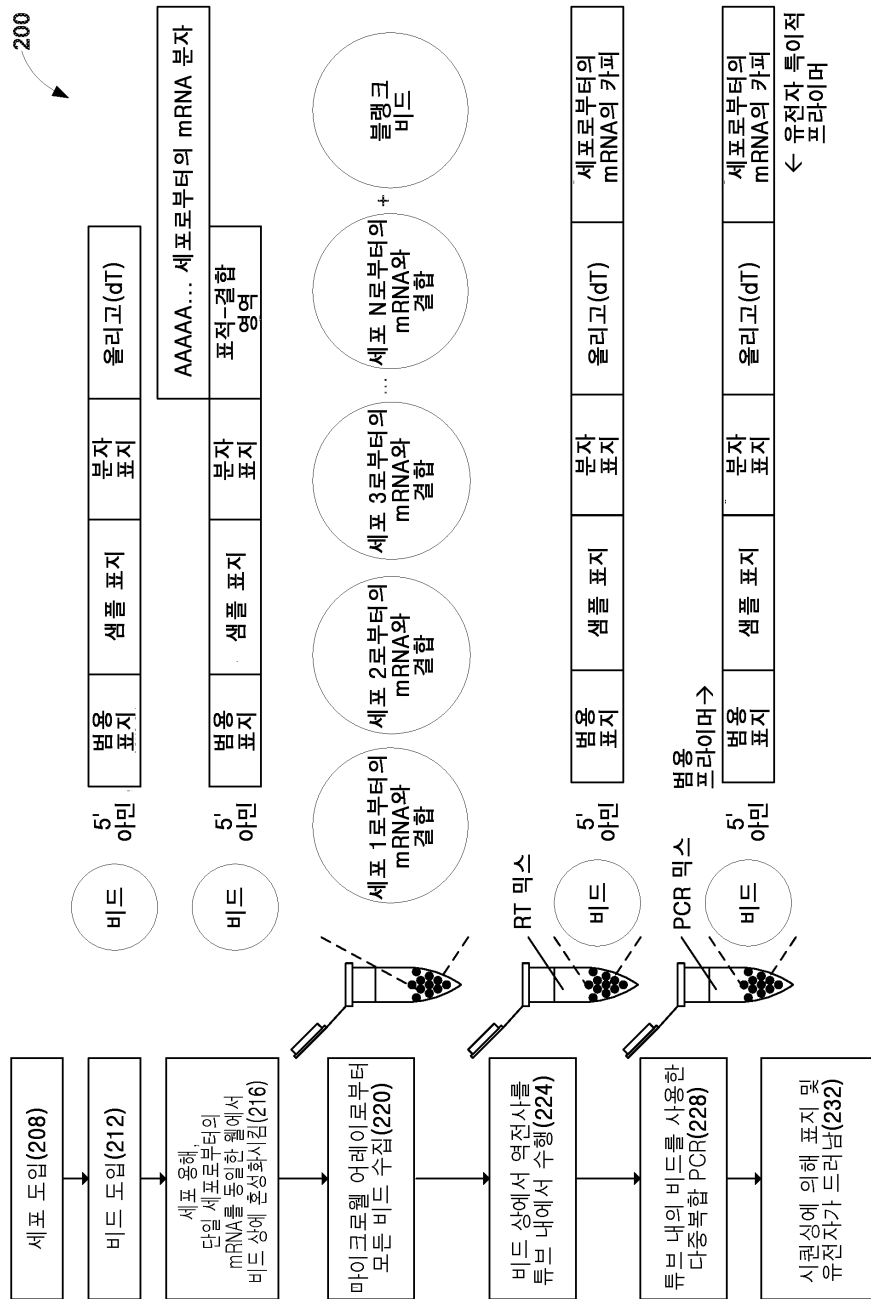
[0809] 다양한 양태 및 실시형태가 본원에 개시되지만, 다른 양태 및 실시형태가 당업자에게 명백할 것이다. 본원에 개시된 다양한 양태 및 실시형태는 예시의 목적을 위한 것이며, 제한하는 것으로 의도되지 않으며, 실제 범주와 목적은 하기의 청구범위에 의해 나타난다.

도면

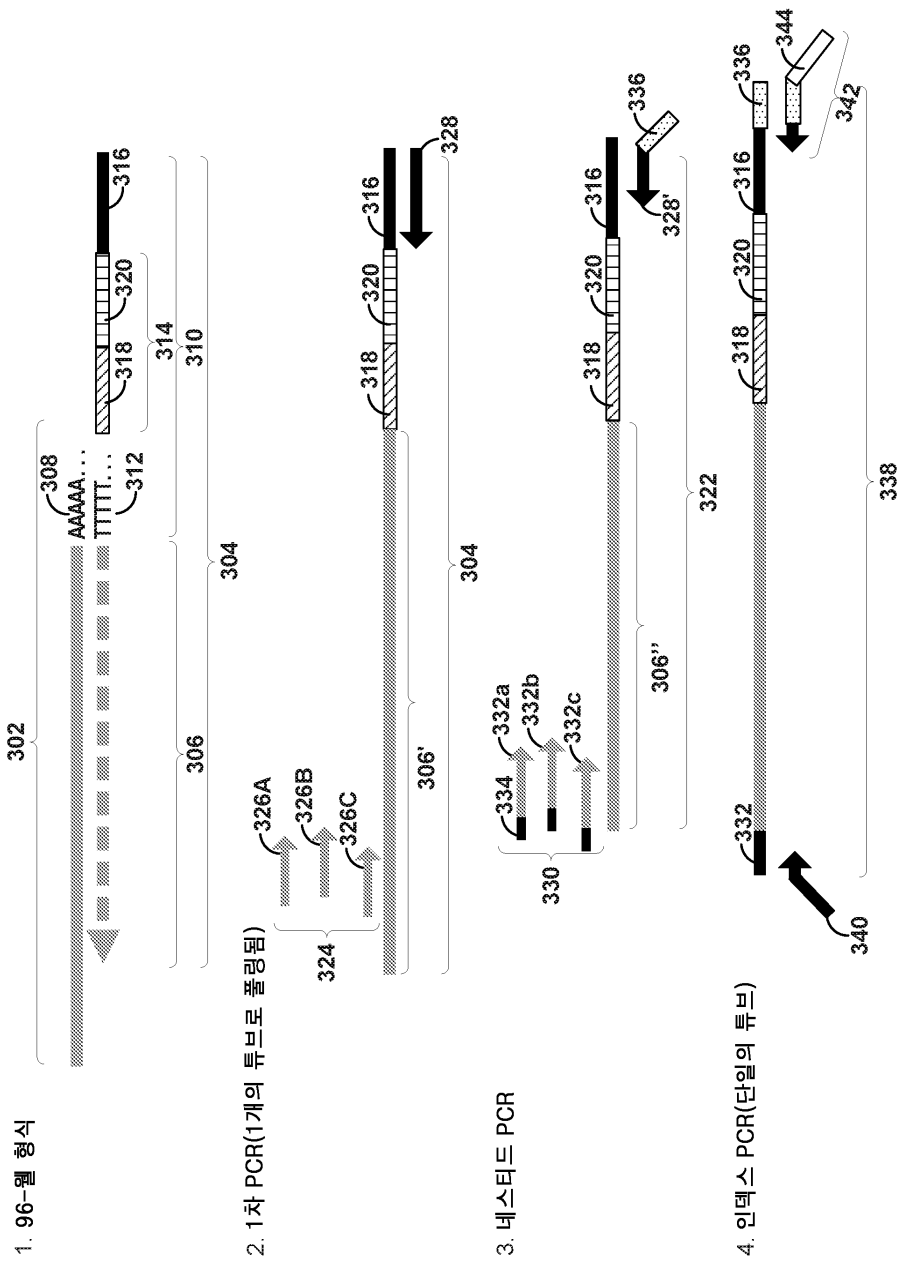
도면1



도면2

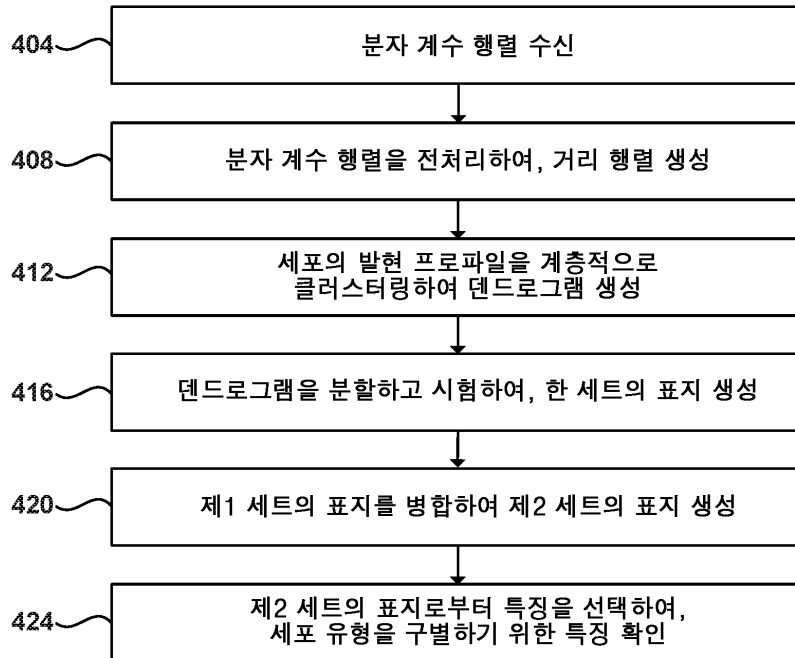


도면3

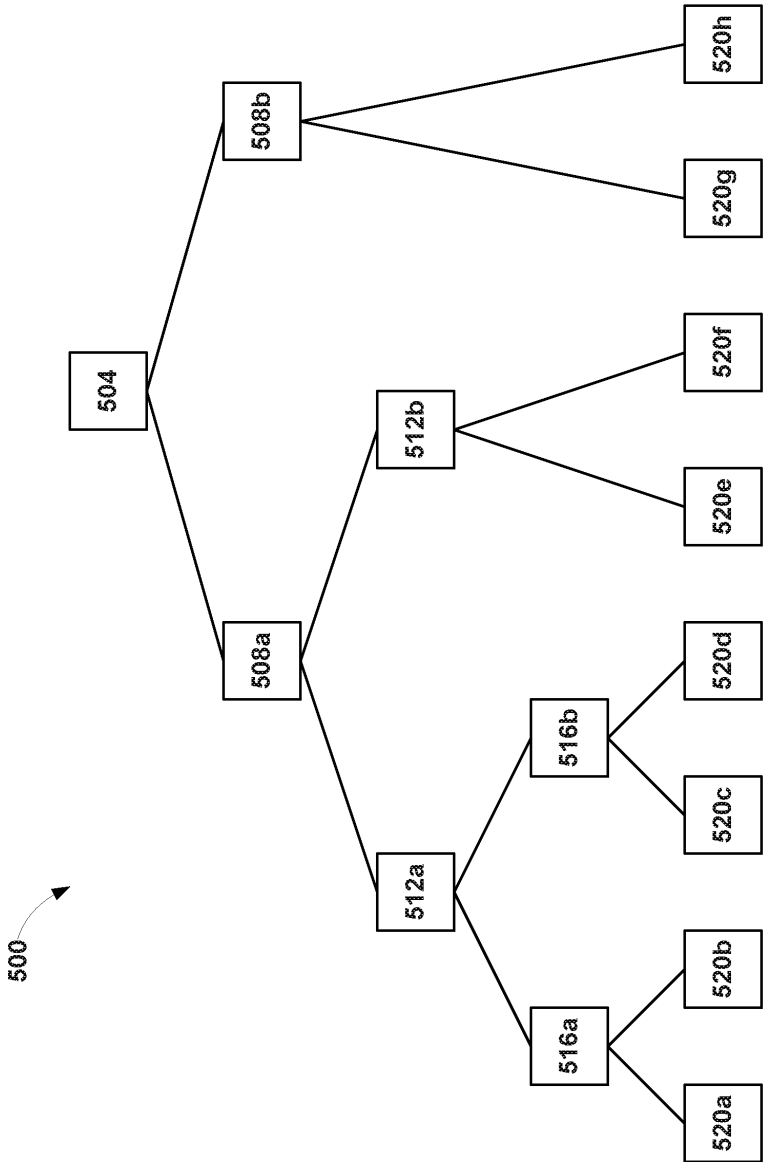


도면4

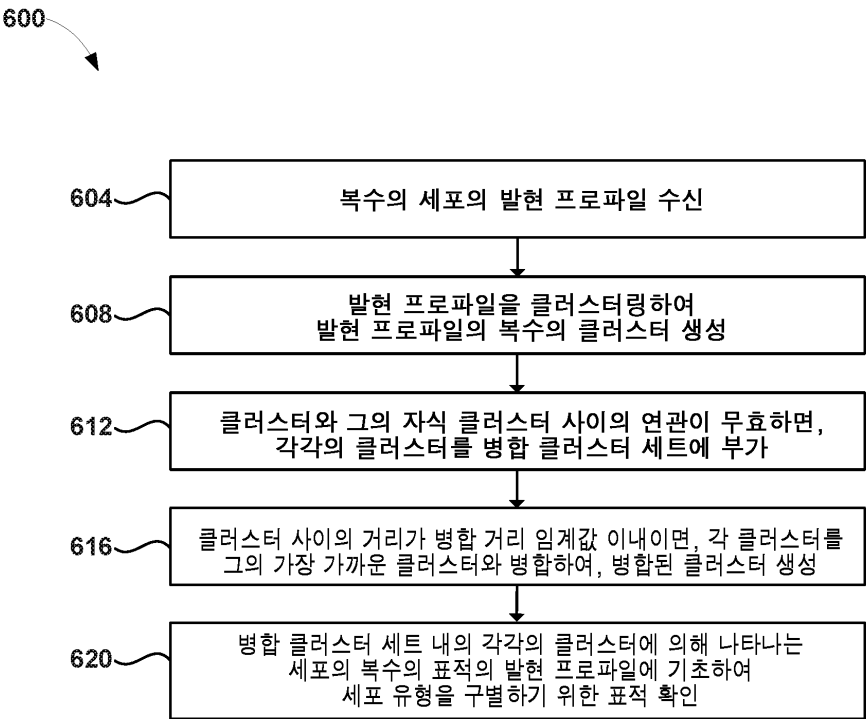
400



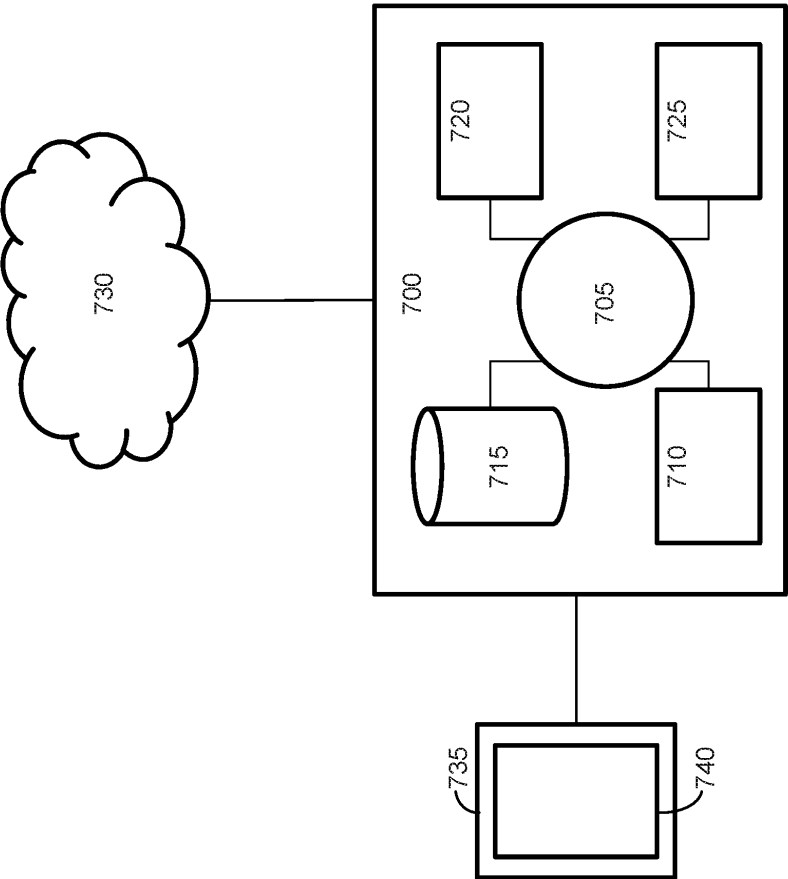
도면5



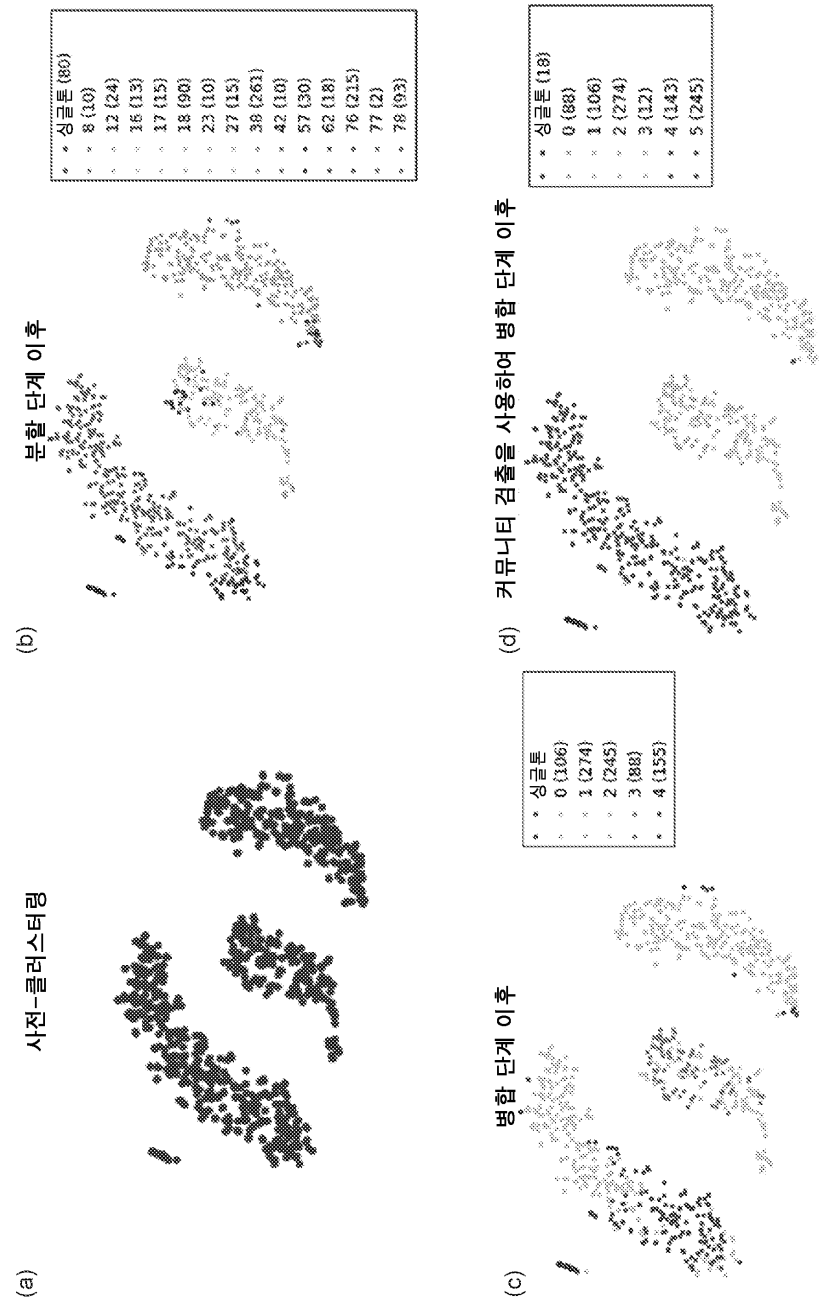
도면6



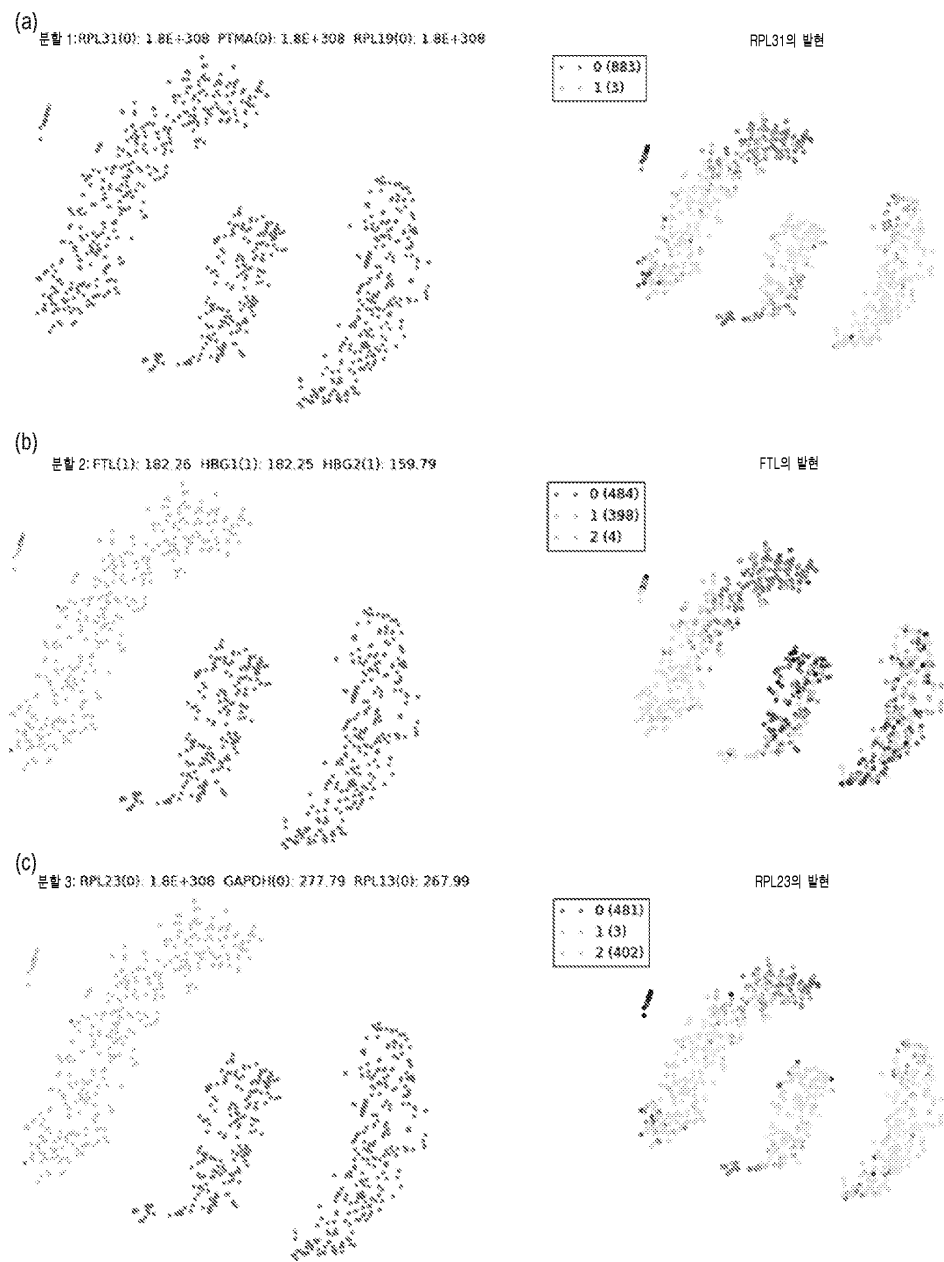
도면7



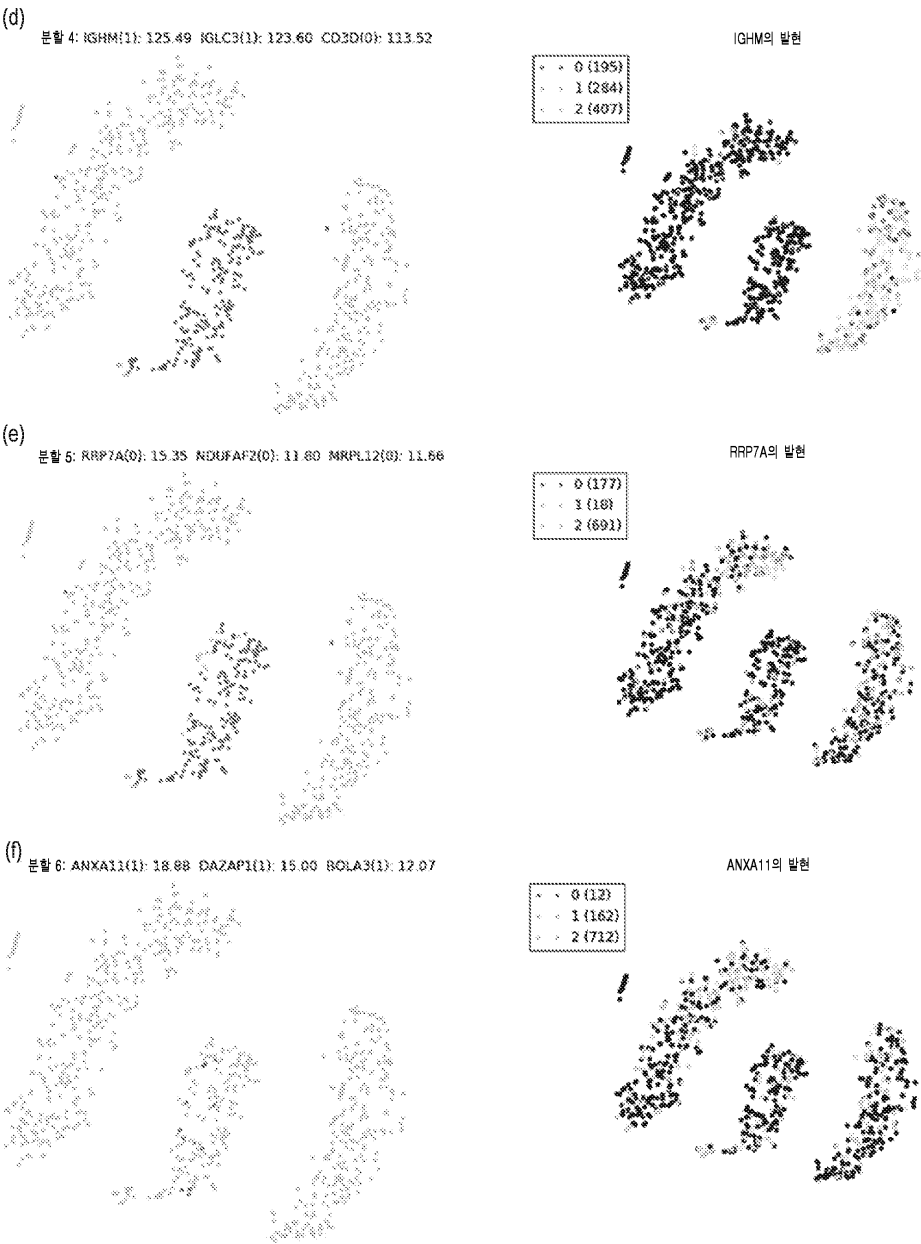
도면8



도면9a



도면9b

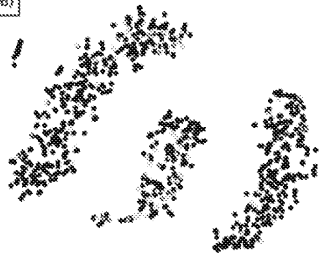
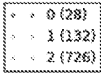


도면9c

(g) 분할 7: TTF1(1): 12.32 SPTBN1(1): 11.17 SLC3A2(1): 9.50



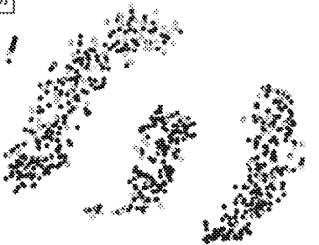
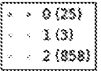
TTF1의 발현



(h) 분할 8: SRPK1(1): 13.94 SMNDC1(1): 10.31 C1orf35(1): 10.31



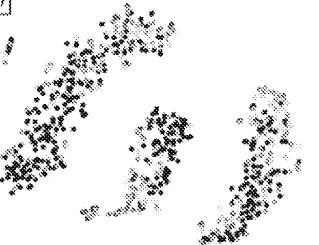
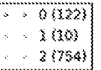
SRPK1의 발현



(i) 분할 9: TOP2A(0): 18.52 TJ884B(0): 18.48 EC12(0): 18.19



TOP2A의 발현



도면9d

(j) 분할 10: CACYBP(1): 24.92 PHPTI(1): 15.44 ATIC(1): 13.68



CACYBP의 발현

• 0 (13)
• 1 (109)
• 2 (764)



(k) 분할 11: RPSA(0): 31.72 HNRNPOL(0): 29.92 NUCKS1(0): 29.68



RPSA의 발현

• 0 (105)
• 1 (4)
• 2 (777)

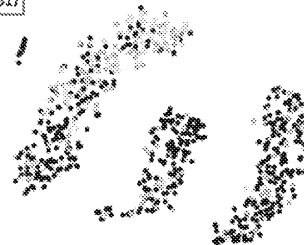


(l) 분할 12: PSMD14(1): 11.31 DBF4(1): 11.19 MDCS2(1): 11.05



PSMD14의 발현

• 0 (15)
• 1 (90)
• 2 (781)

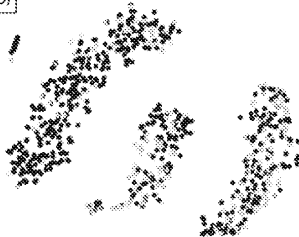


도면9e

(m) 분할 13: RNASEH2B(0): 38.04 EIF3F(0): 34.36 PPP2CA(0): 34.00



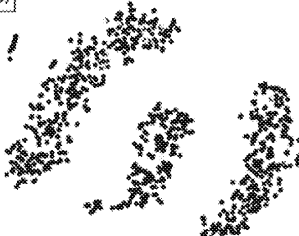
RNASEH2B의 발현



(n) 분할 14: GAS8(0): 233.23 ALKBH3(0): 177.62 RPL19(1): 169.72



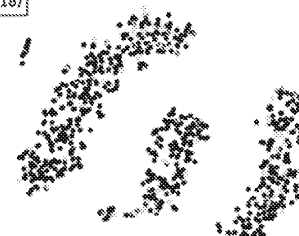
GAS8의 발현



(o) 분할 15: CNPY3(1): 80.24 NDUFB2(0): 79.70 IFT57(1): 76.61



CNPY3의 발현

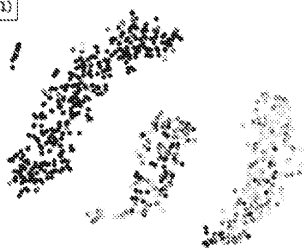


도면9f

(p) 분할 16: MZB1(1): 100.26 RPL1-5106.1(1): 76.45 MYL128(1): 79.36



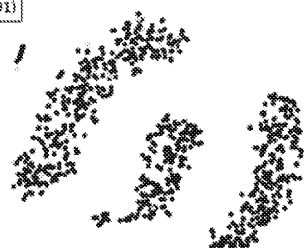
MZB1의 발현



(q) 분할 17: CREB3L1(1): 228.58 RPL32(0): 224.99 RPS20(0): 212.30



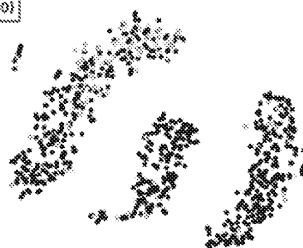
CREB3L1의 발현



(r) 분할 18: VMP1(1): 33.24 TFR2(1): 22.67 STT3A(1): 19.27



VMP1의 발현



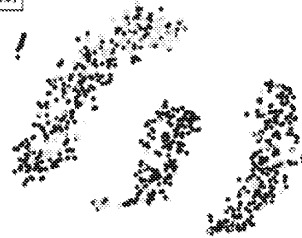
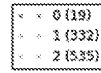
도면9g

(s)

분할 19: EIF2B1(1): 32.86 ARPC1A(1): 31.45 HMGNS(1): 30.06



EIF2B1의 발현

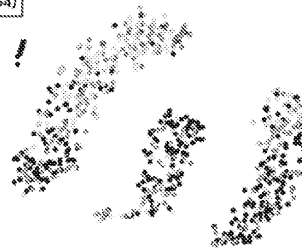
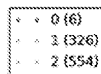


(t)

분할 20: NUDT5(1): 83.57 MSN(1): 81.87 ANAPCS(1): 80.59

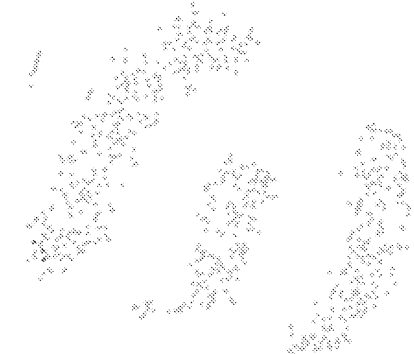


NUDT5의 발현

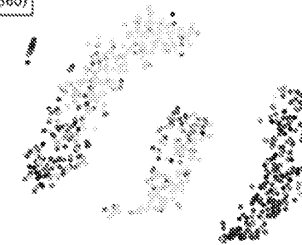
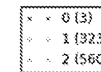


(u)

분할 21: TMSB4X(1): 148.25 PPIA(1): 126.76 SNRPB(1): 122.61



TMSB4X의 발현

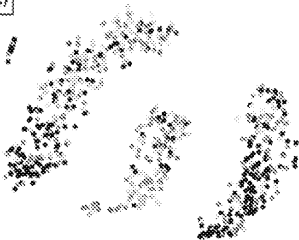


도면9h

(v) 분할 22: C12orf57(0): 71.43 RPI1-234A1.1(0): 65.08 RNPS1(0): 62.22



C12orf57의 발현



(w) 분할 23: RPL27A(0): 48.70 UQCRLH(0): 44.94 UBA52(0): 44.85



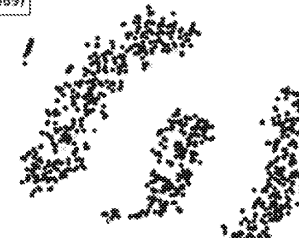
RPL27A의 발현



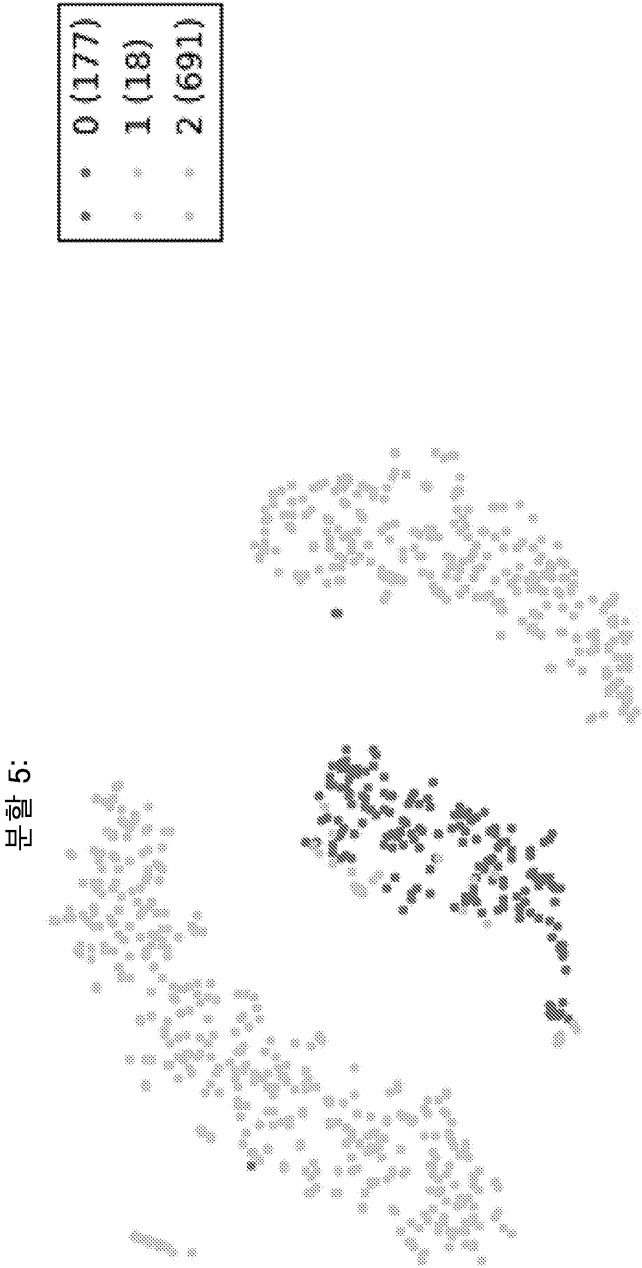
(x) 분할 24: JUN(1): 132.42 NDUFS5(0): 133.15 PEBP1(0): 126.54



JUN의 발현



도면10



도면11

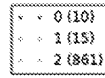


도면 12a

(a) 병합 1: ENOSF1(0): 3.60 HINT1(0): 3.51 RPS24(0): 3.42



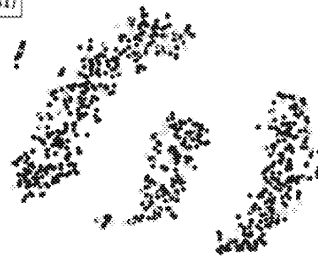
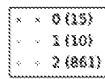
ENOSF1의 발현



(b) 병합 2: MAGED1(1): 4.31 ALDH1A2(1): 3.82 TUBA1B(0): 3.67



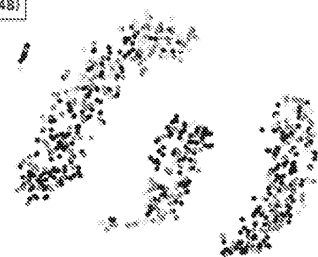
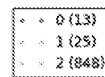
MAGED1의 발현



(c) 병합 3: PRPF40A(1): 4.37 RBM39(1): 4.36 CYFIP2(1): 3.87



PRPF40A의 발현

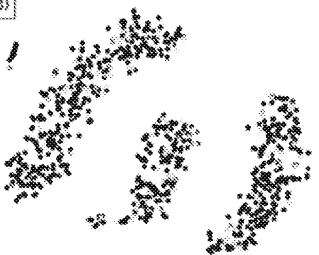
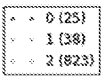


도면12b

(d) 병합 4: ALDOC(0): 5.23 RPL31(0): 4.29 RPS11(0): 4.24



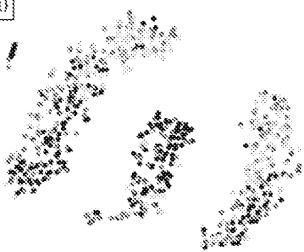
ALDOC의 발현



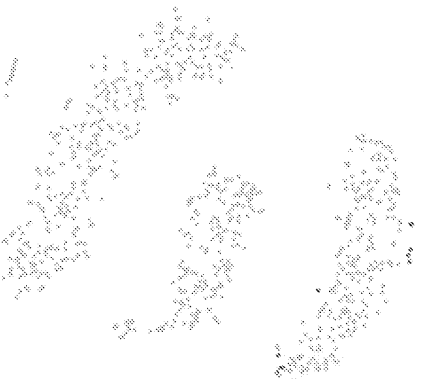
(e) 병합 5: PARP1(0): 6.04 SET(0): 5.67 CALM2(0): 5.60



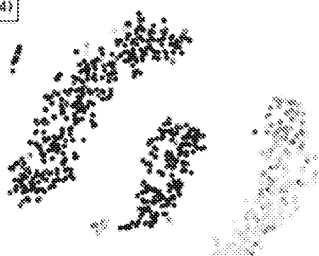
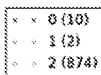
PARP1의 발현



(f) 병합 6: IGLC3(0): 6.81 LCP1(0): 6.59 IGMM(0): 5.56



IGLC3의 발현

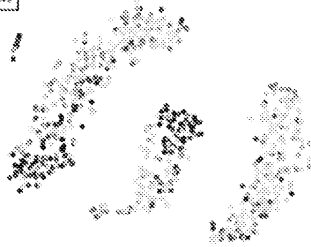
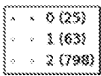


도면12c

(g) 병합 7: HMGB2(0): 7.19 HSP90AA1(0): 6.71 TOP2A(0): 5.49



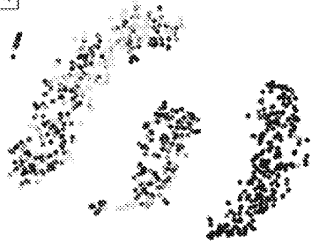
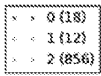
HMGB2의 발현



(h) 병합 8: VIM(0): 7.23 GLUL(0): 6.55 DLX1(0): 6.00



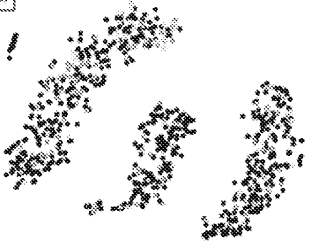
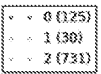
VIM의 발현



(i) 병합 9: HMGN5(0): 9.76 YWHAZ(1): 6.37 TCHP(0): 6.18



HMGN5의 발현



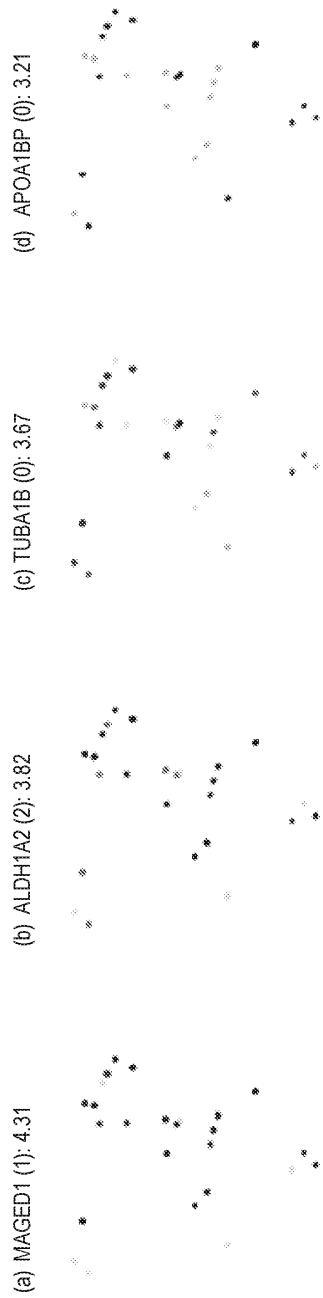
도면13

•	•	0 (15)
•	•	1 (10)
•	•	2 (861)

도면 2:



도면14



도면15a



도면15b

(d)

클러스터: 2, 점수: 137.51

FTL 137.51

HBG1 120.99

UQCRL 114.51



(e)

클러스터: 3, 점수: 32.53

CD3D 32.53

AOA 28.54

CHI3L2 28.50



(f)

클러스터: 4, 점수: 18.15

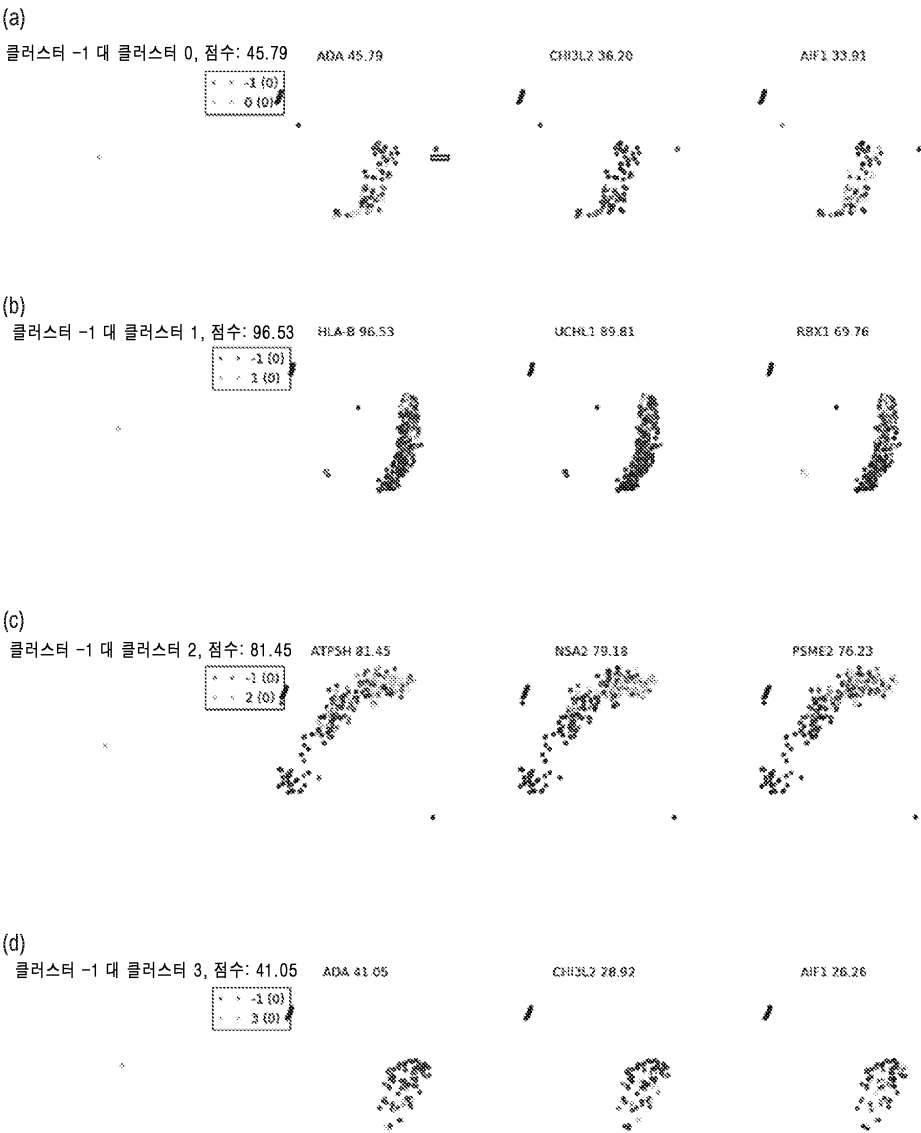
FTL 18.15

PRAME 14.63

KRT18 10.43



도면16a



도면 16b

(e)

클러스터 -1 대 클러스터 4, 점수: 33.87

> > -1 {0}
 > > 4 {0}

ATP5H 33.87

SNRPC 30.90

NSA2 30.75

(f)

클러스터 0 대 클러스터 1, 점수: 62.52

> > 0 {0}
 > > 1 {0}

CD3D 62.52

IGHM 53.49

CH3L2 48.75

(g)

클러스터 0 대 클러스터 2, 점수: 119.88

> > 0 {0}
 > > 2 {0}

APOC1 119.88

HBG1 109.45

HBG2 101.04

(h)

클러스터 0 대 클러스터 3, 점수: 10.81

> > 0 {0}
 > > 3 {0}

H2AFZ 10.81

RAN 10.20

UBE2D3 8.96

도면16c

(i)



(j)



(k)

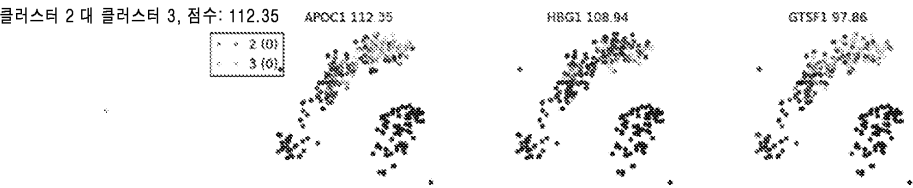


(l)



도면16d

(m)



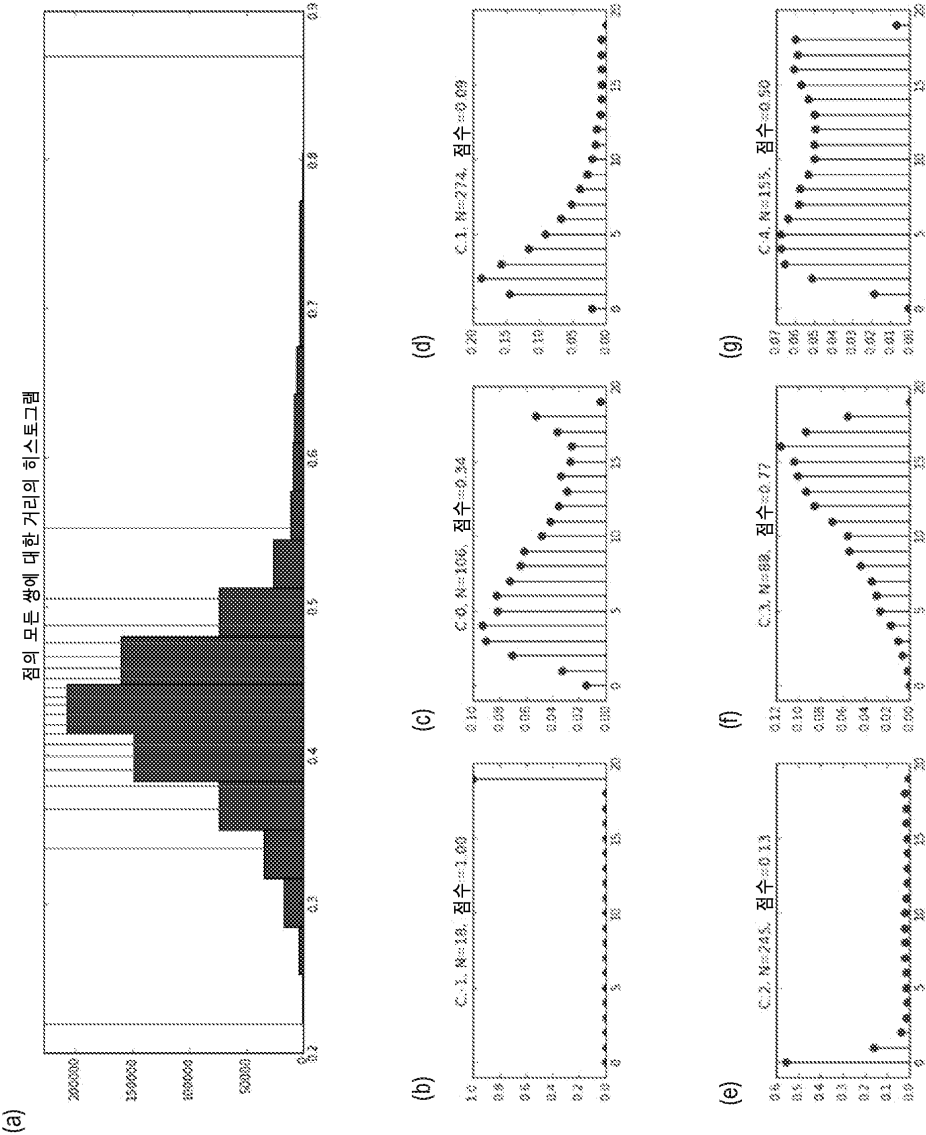
(n)



(o)



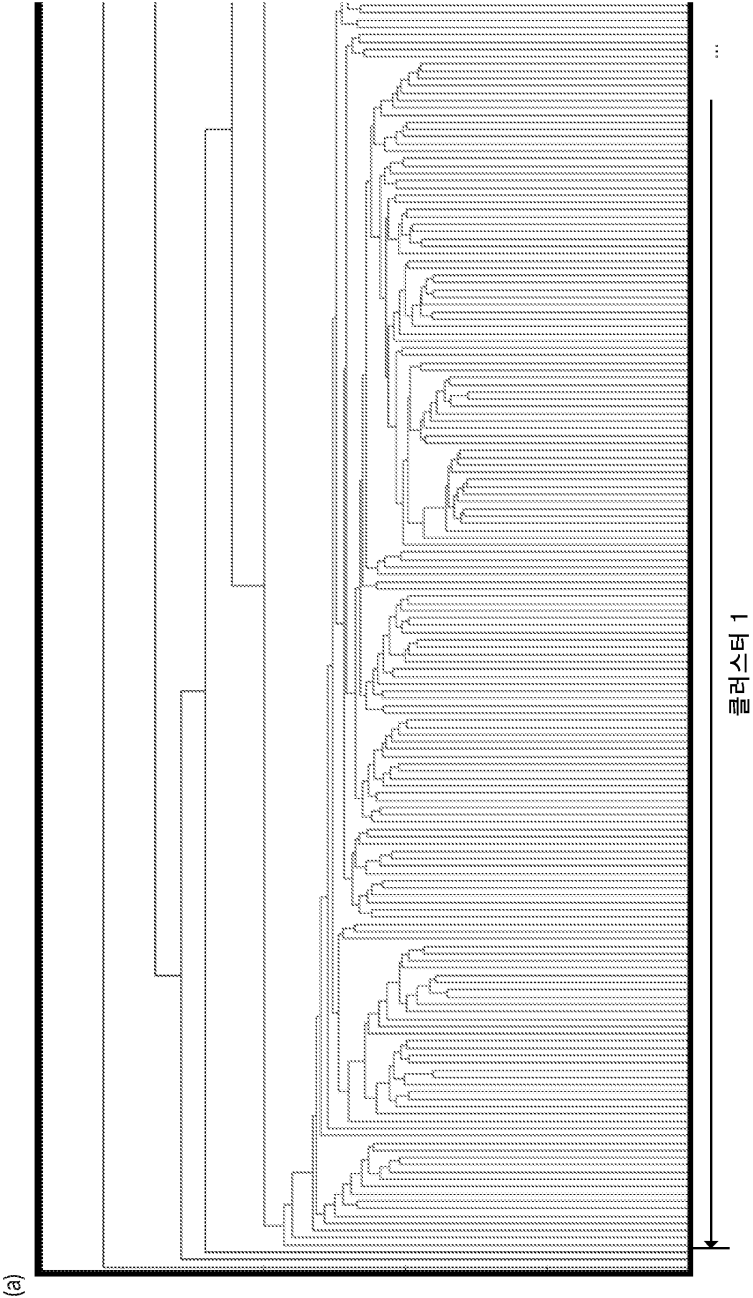
도면17



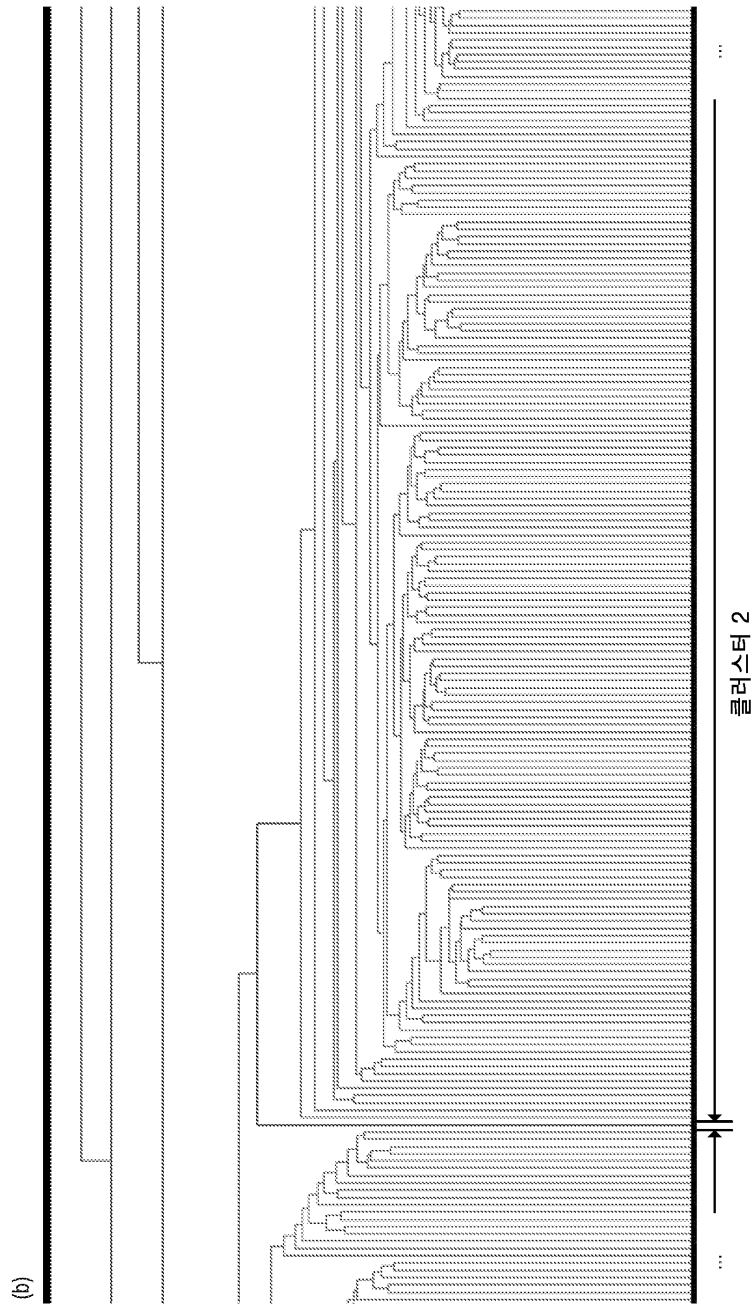
도면18a

패널 (a)	패널 (b)	패널 (c)	패널 (d)	패널 (e)
--------	--------	--------	--------	--------

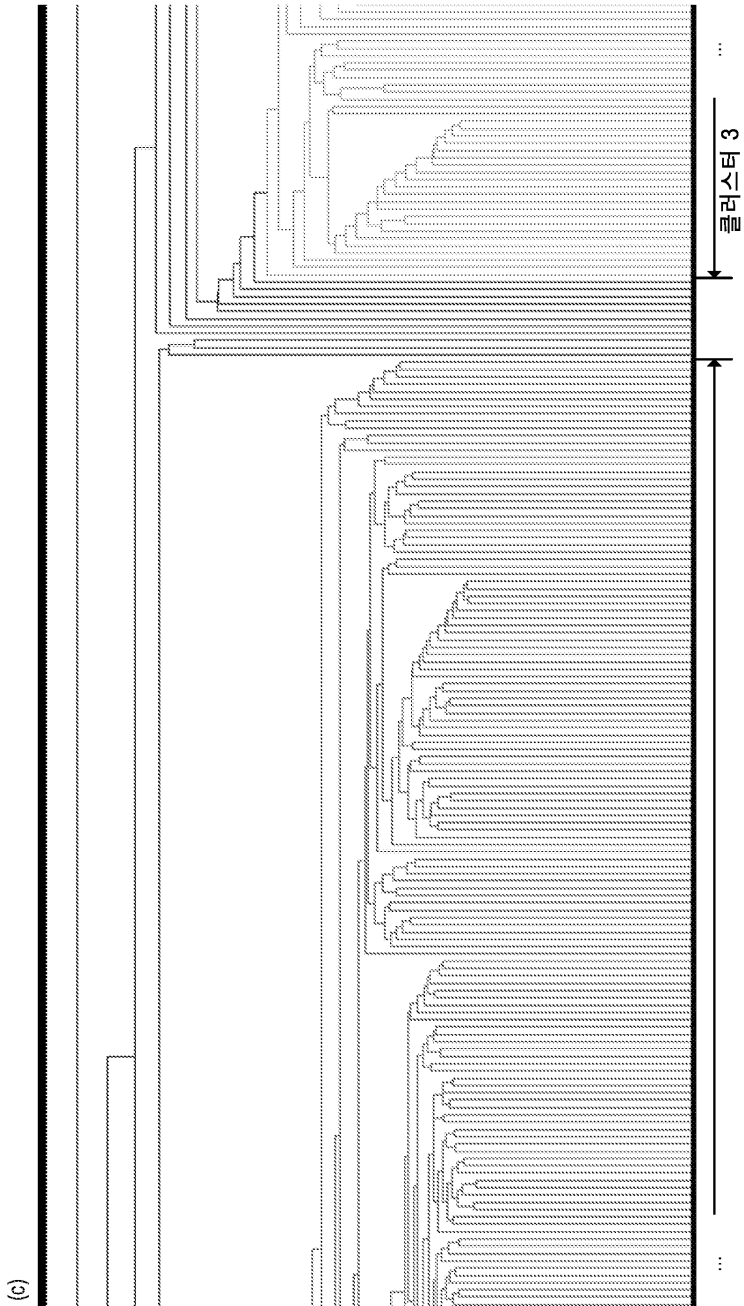
도면18b



도면18c



도면18d



도면18e



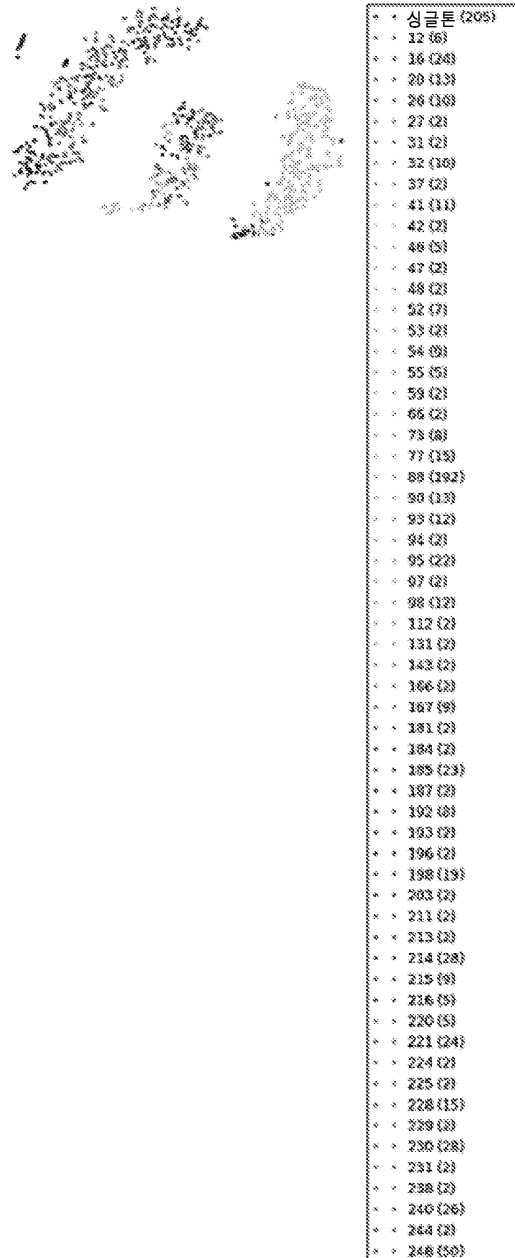
도면18f



도면 19a

(a)

5.00의 임계값을 사용한 클러스터링 결과



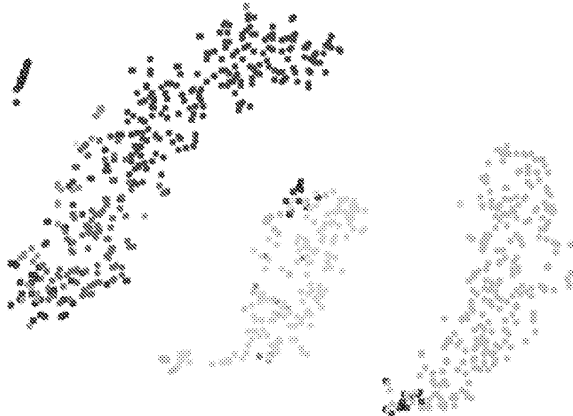
도면 19b

(b) 10.000의 임계값을 사용한 클러스터링 결과



* * 싱글톤 (80)
* * 8 (10)
* * 12 (24)
* * 16 (13)
* * 17 (15)
* * 18 (90)
* * 23 (10)
* * 27 (15)
* * 38 (261)
* * 42 (10)
* * 57 (30)
* * 62 (18)
* * 76 (215)
* * 77 (2)
* * 78 (93)

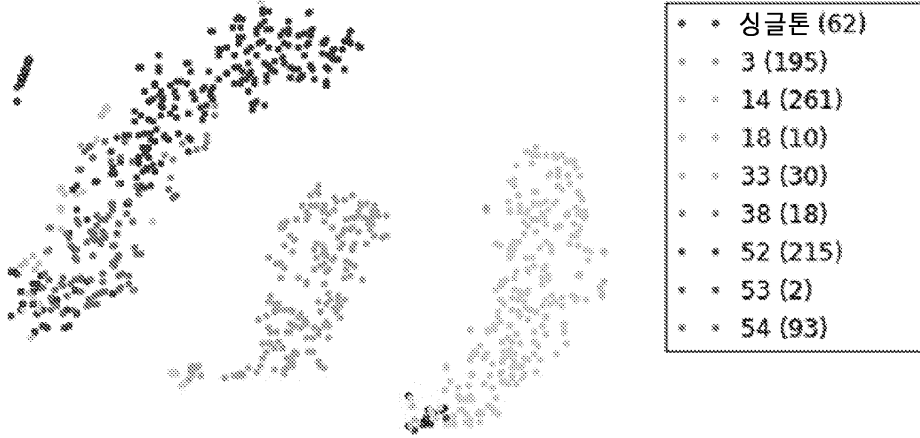
(c) 15.000의 임계값을 사용한 클러스터링 결과



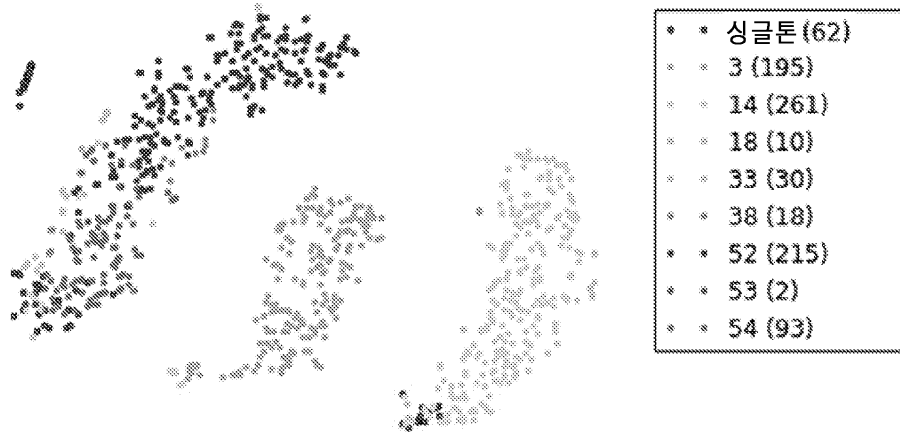
* * 싱글톤 (72)
* * 8 (10)
* * 11 (160)
* * 15 (15)
* * 26 (261)
* * 30 (10)
* * 45 (30)
* * 50 (18)
* * 64 (215)
* * 65 (2)
* * 66 (93)

도면19c

(d) 20.000의 임계값을 사용한 클러스터링 결과

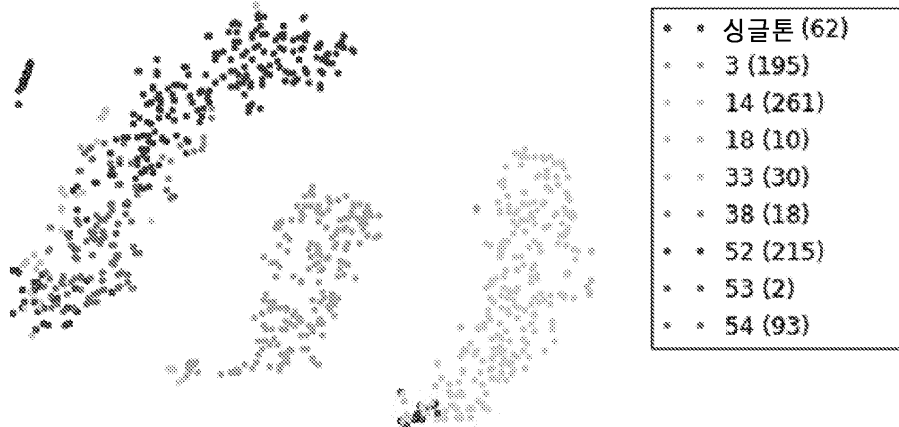


(e) 25.000의 임계값을 사용한 클러스터링 결과

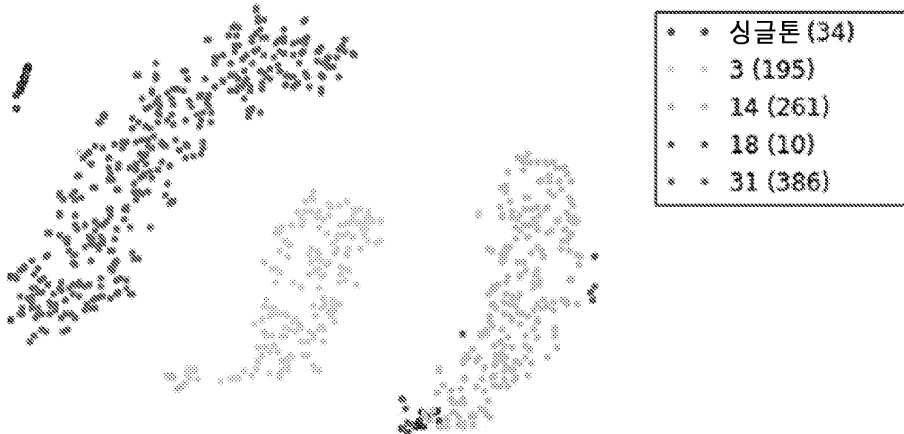


도면 19d

(f) 30.000의 임계값을 사용한 클러스터링 결과

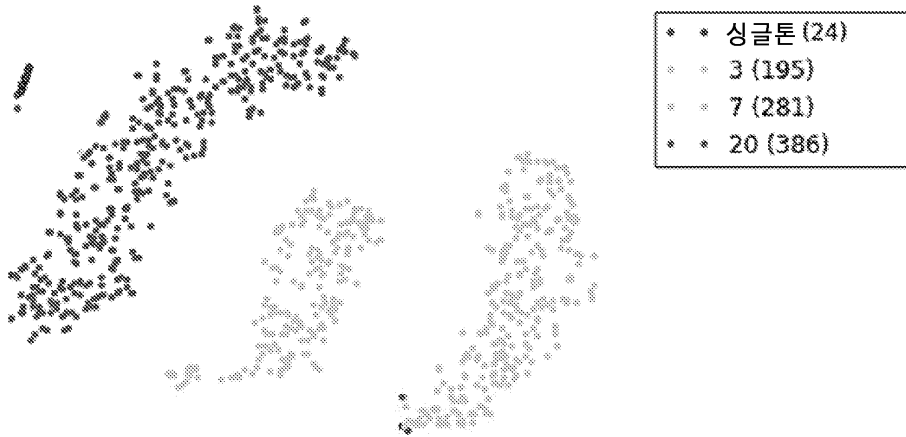


(g) 35.000의 임계값을 사용한 클러스터링 결과

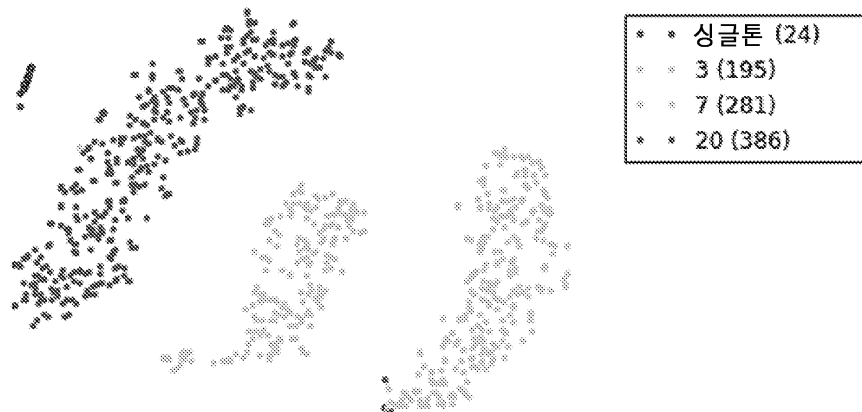


도면 19e

(h) 40.000의 임계값을 사용한 클러스터링 결과

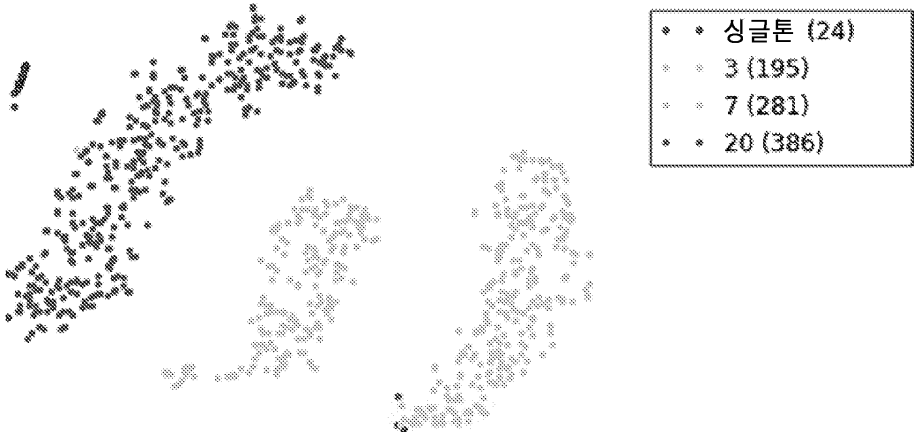


(i) 45.000의 임계값을 사용한 클러스터링 결과

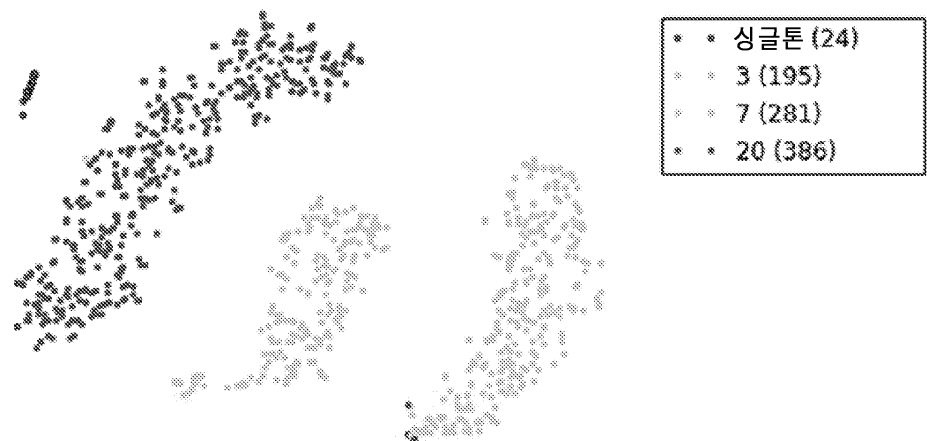


도면19f

(j) 50.000의 임계값을 사용한 클러스터링 결과

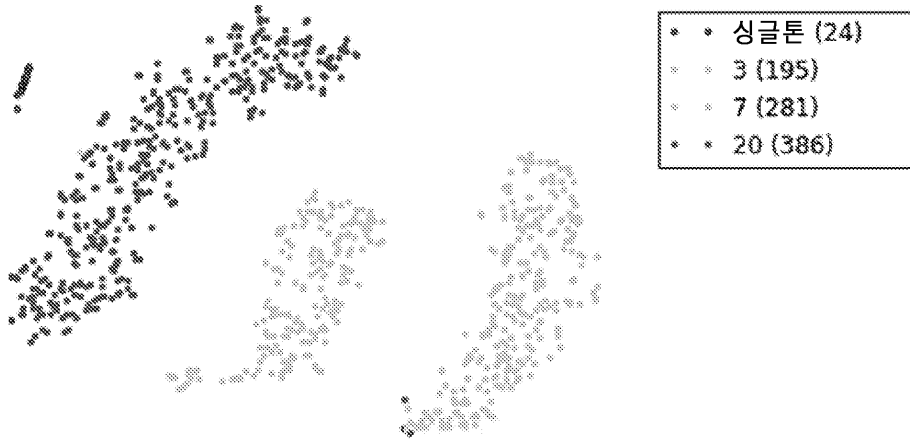


(k) 55.000의 임계값을 사용한 클러스터링 결과

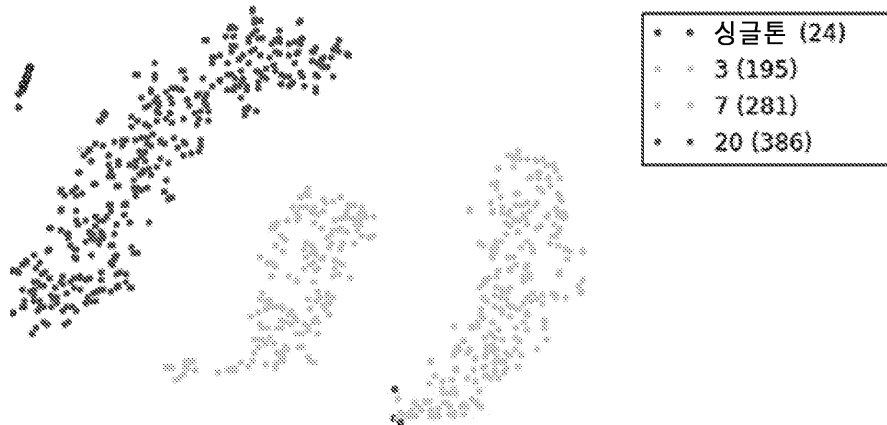


도면 19g

(l) 60.000의 임계값을 사용한 클러스터링 결과

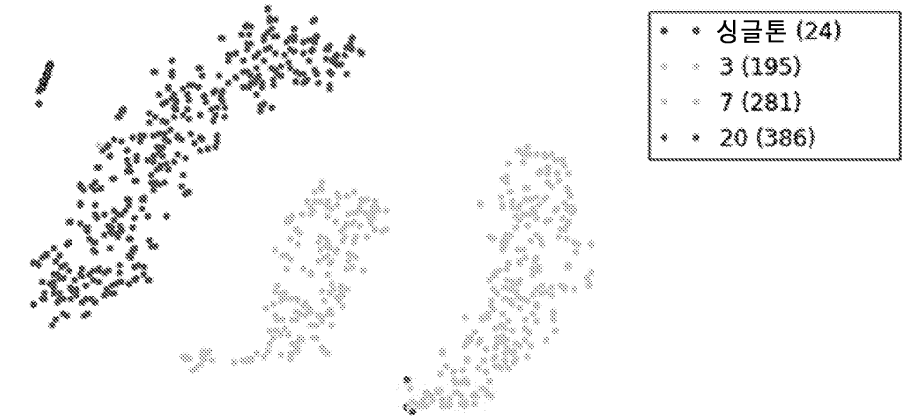


(m) 65.000의 임계값을 사용한 클러스터링 결과

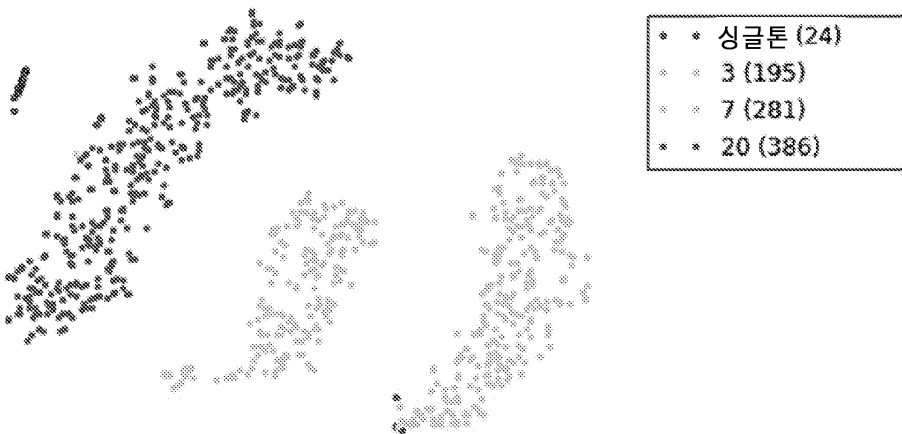


도면 19h

(n) 70.000의 임계값을 사용한 클러스터링 결과

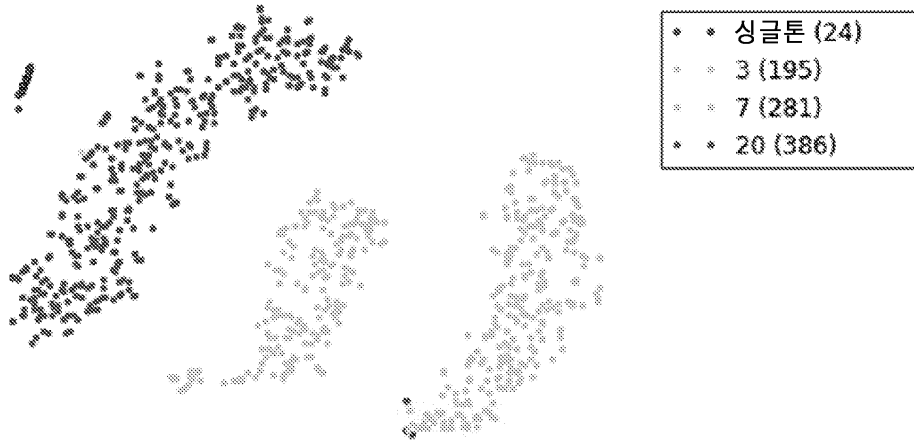


(o) 75.000의 임계값을 사용한 클러스터링 결과

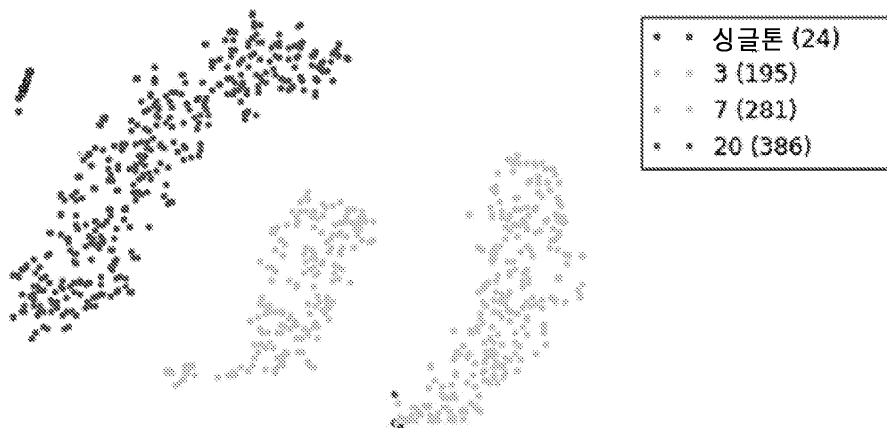


도면19i

(p) 80.000의 임계값을 사용한 클러스터링 결과

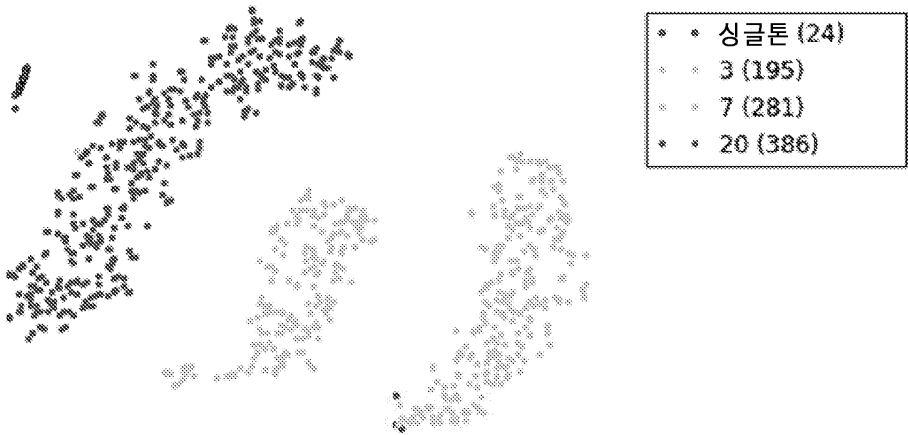


(q) 85.000의 임계값을 사용한 클러스터링 결과

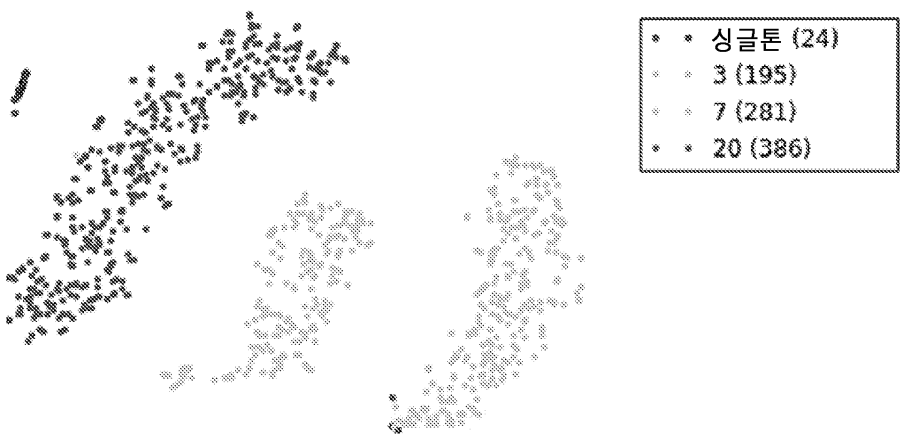


도면19j

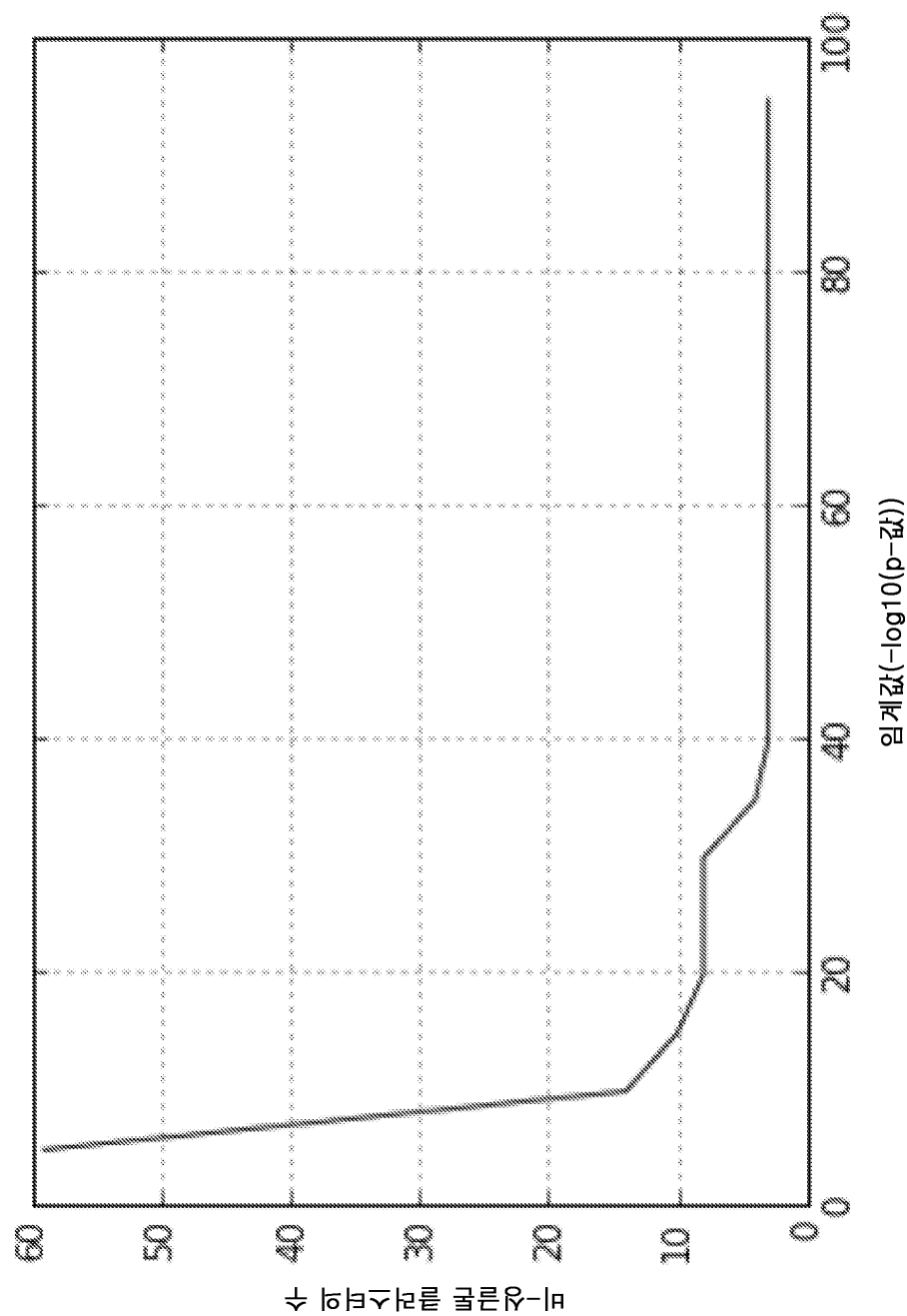
(r) 90.000의 임계값을 사용한 클러스터링 결과



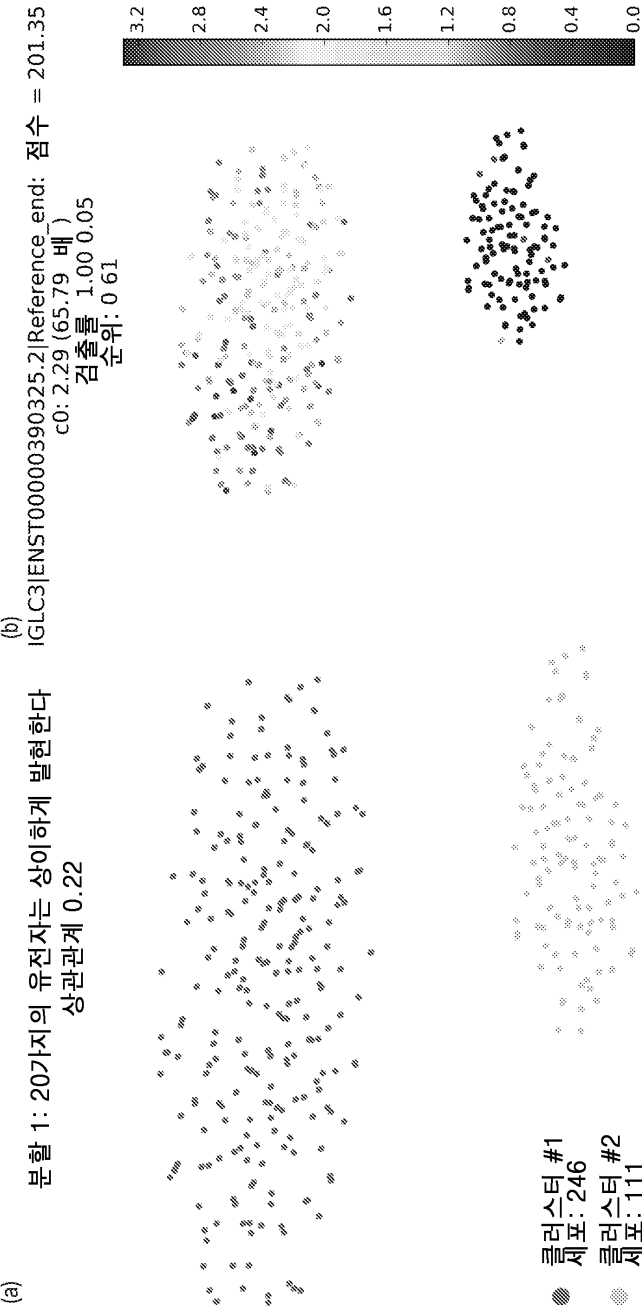
(s) 95.000의 임계값을 사용한 클러스터링 결과



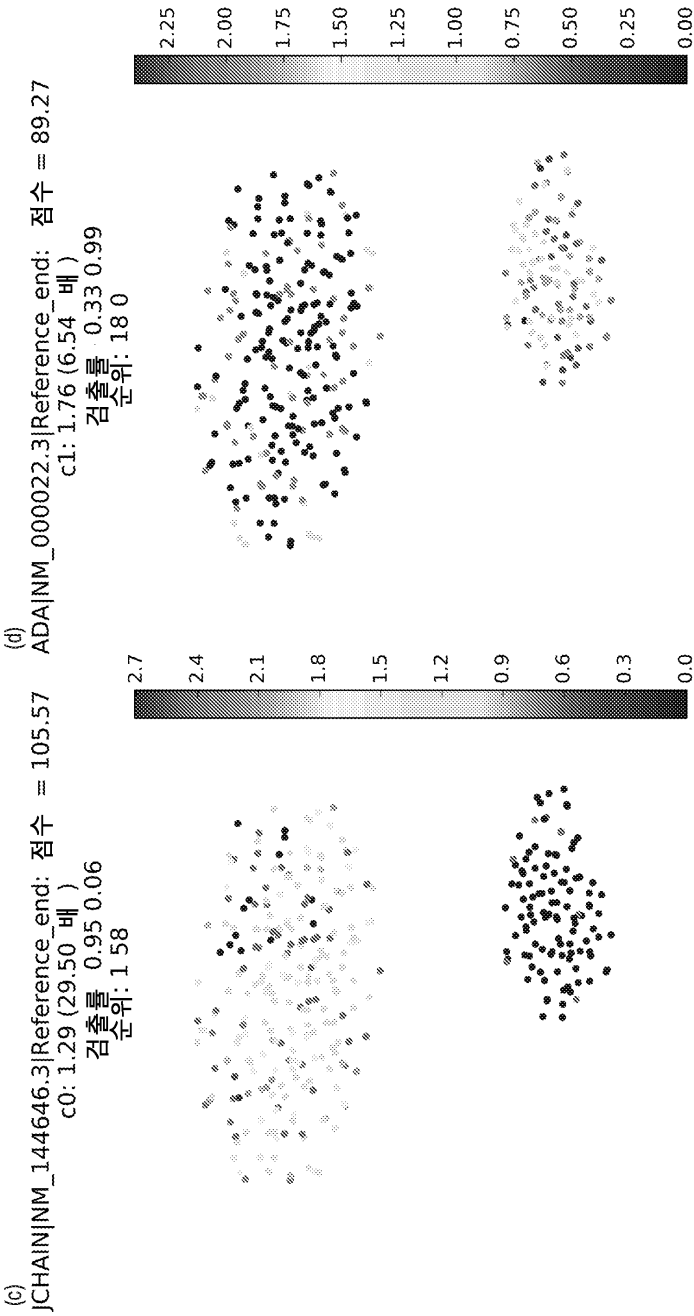
도면20



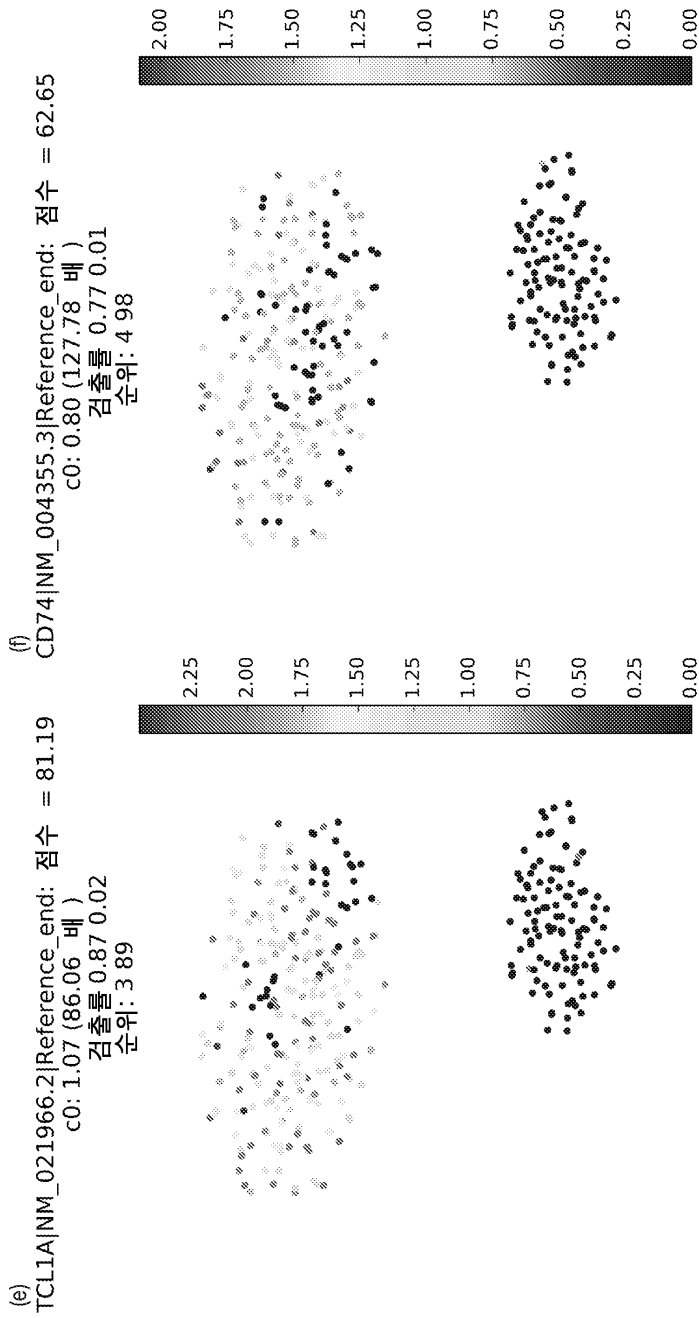
도면21a



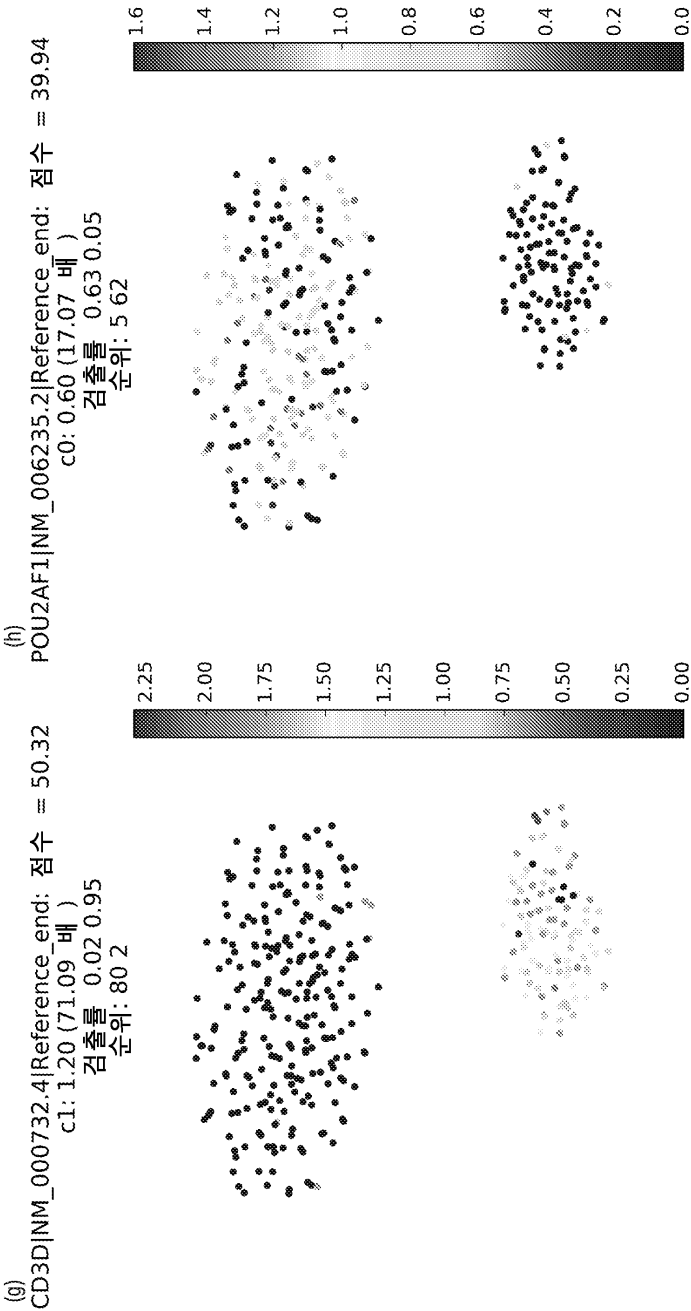
도면21b



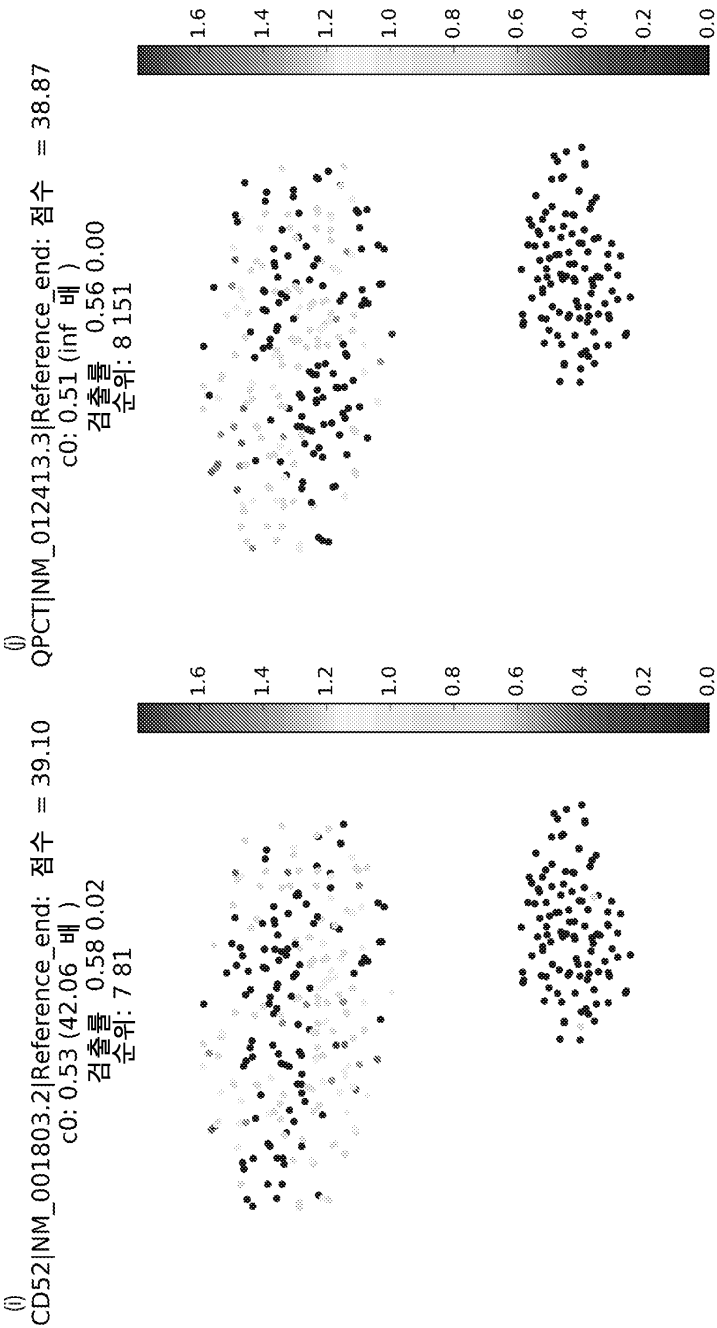
도면21c



도면21d

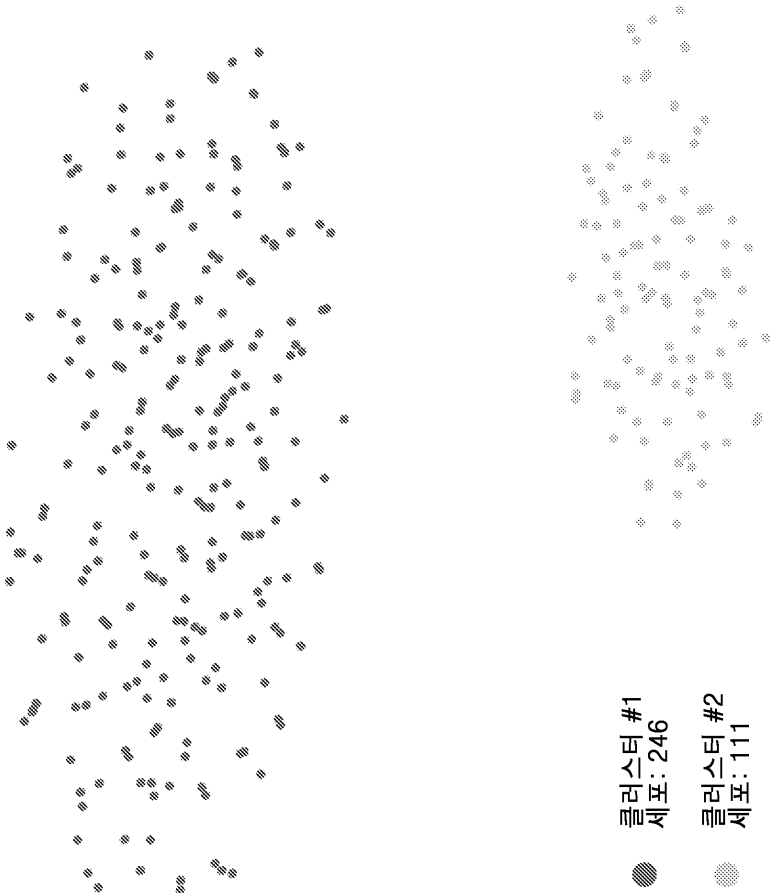


도면21e



도면22

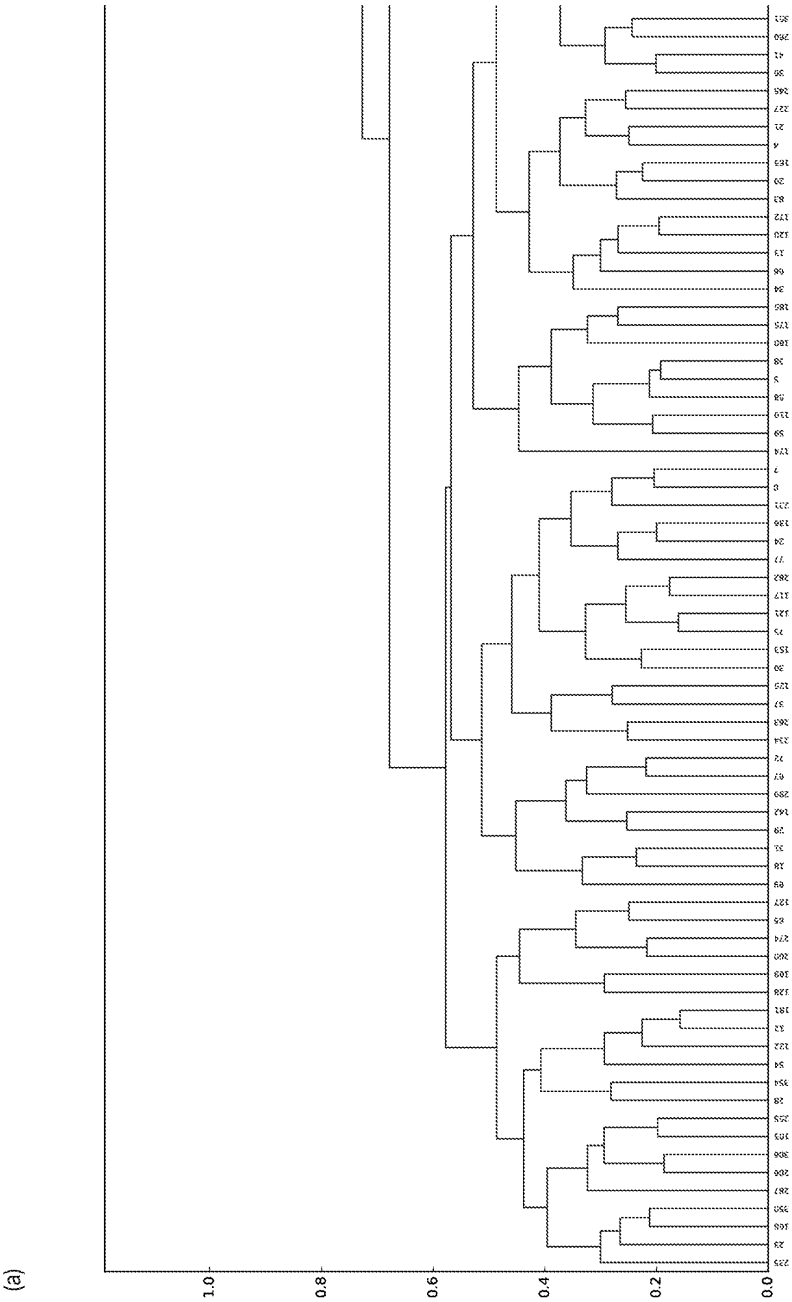
10의 임계값을 사용한 2개 클러스터 분할 결과에서의
t-SNE 차원 감소



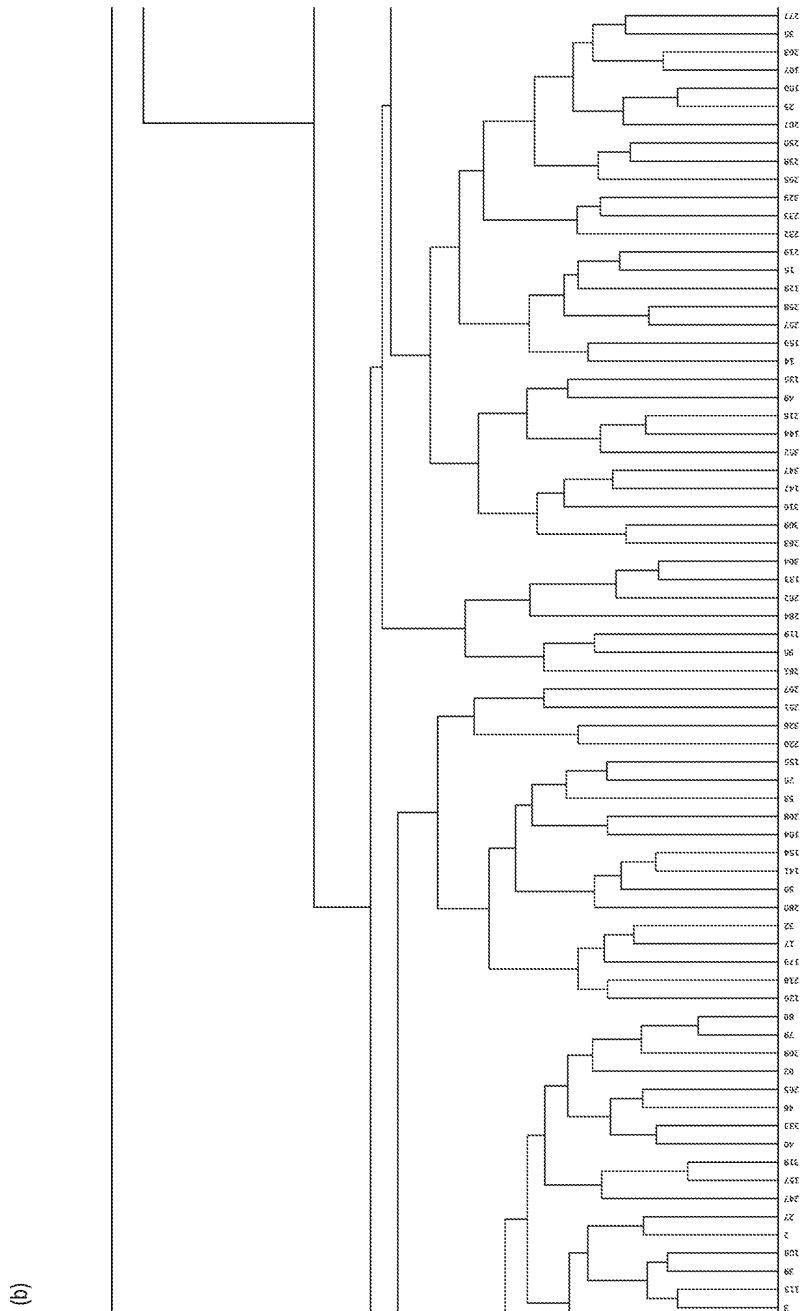
도면23a

패널 (a)	패널 (b)	패널 (c)	패널 (d)	패널 (e)
--------	--------	--------	--------	--------

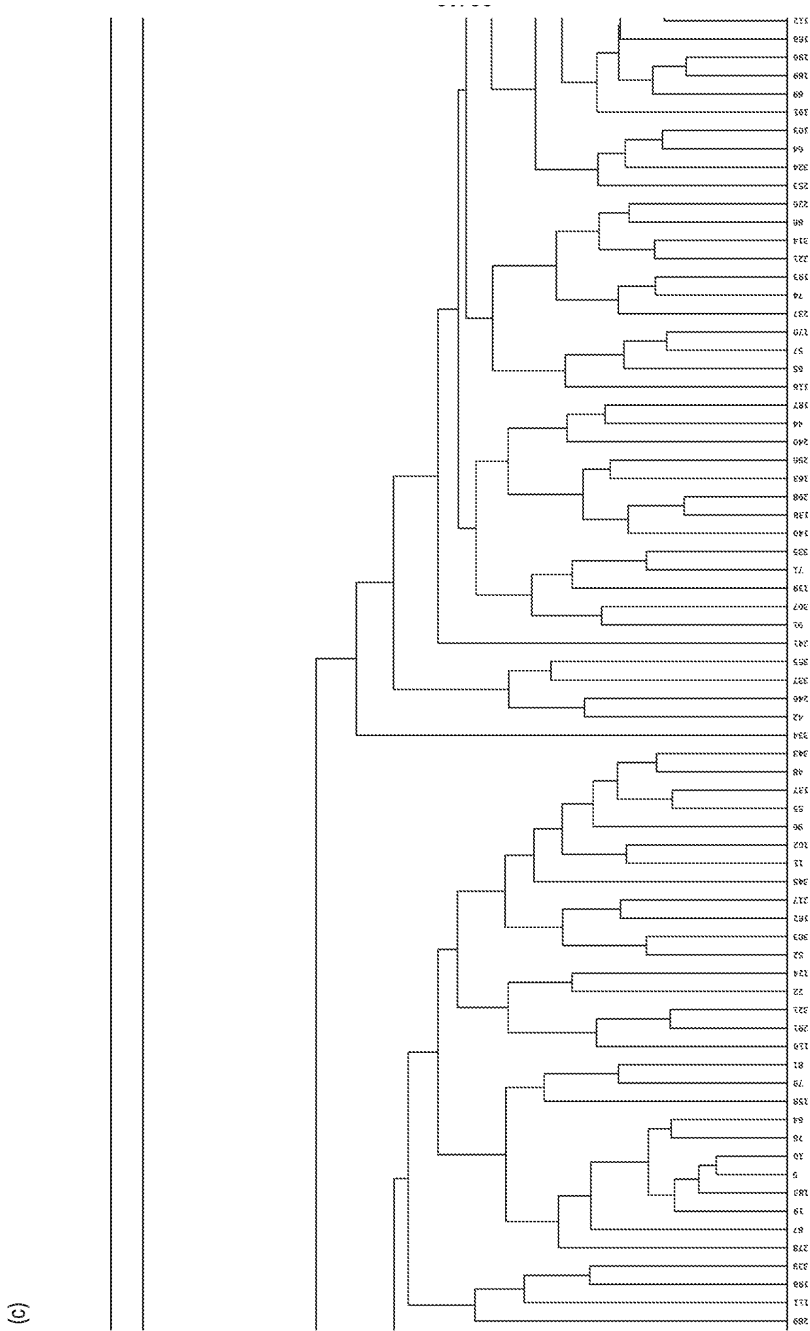
도면23b



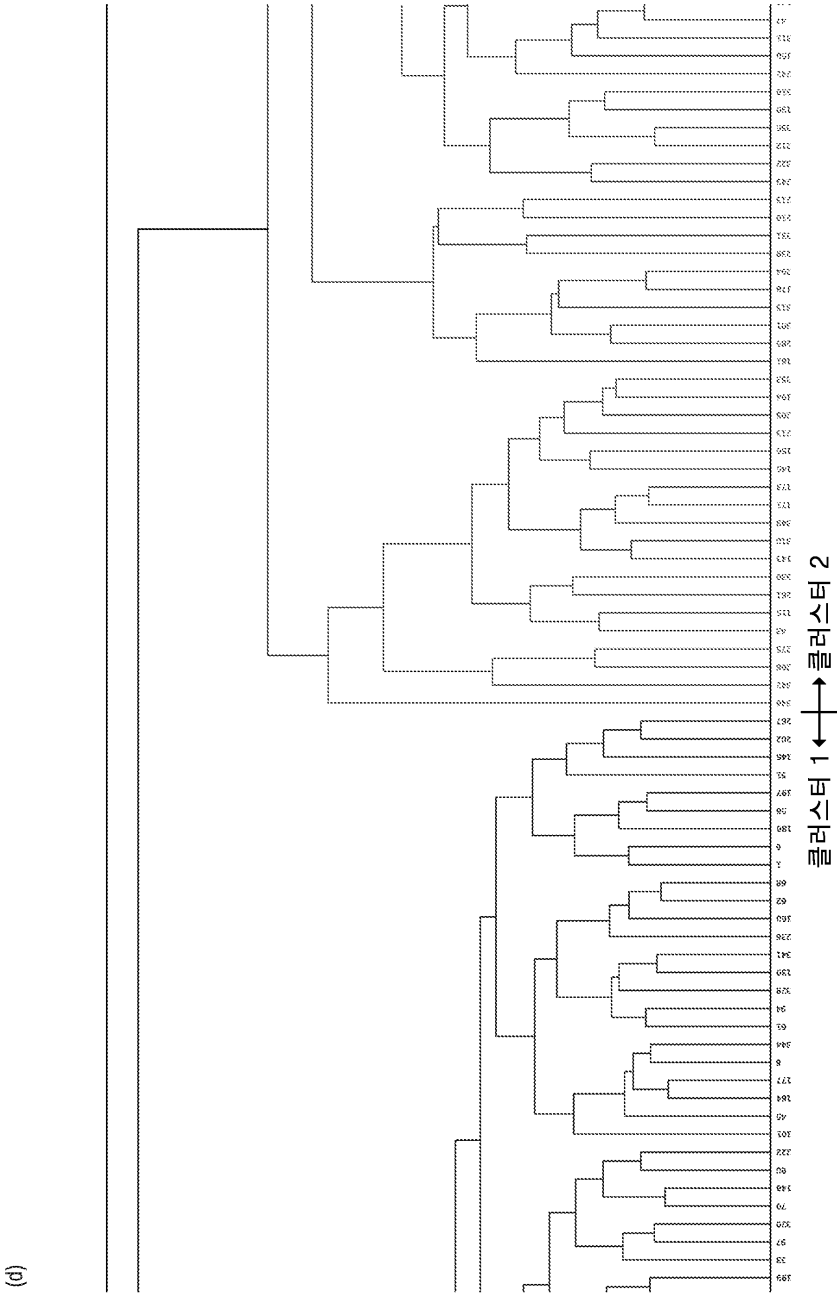
도면23c



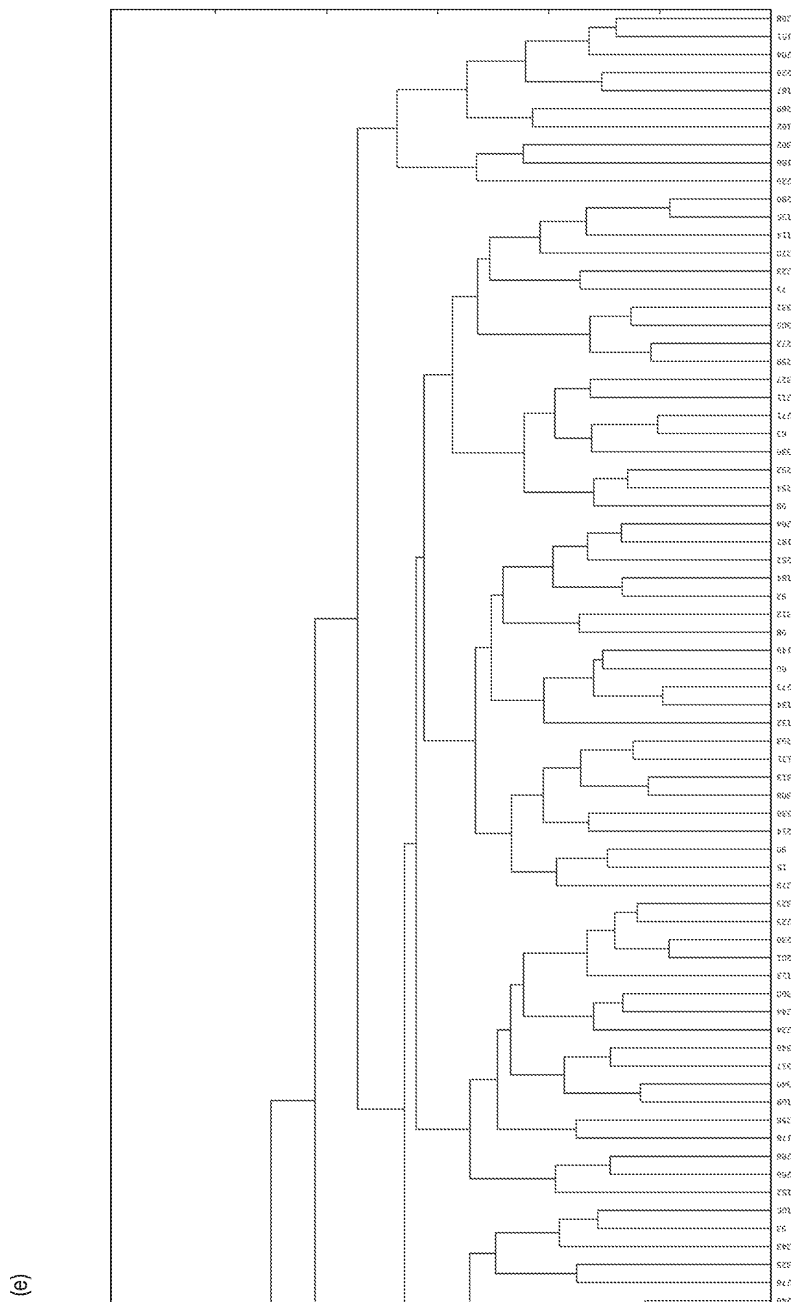
도면23d



도면23e



도면23f



도면24

