(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2011-0095717
(43) 공개일자 2011년08월25일

(51) Int. Cl.
C12Q 1/68 (2006.01) *G06F 19/00* (2011.01)
(21) 출원번호 10-2010-0015333
(22) 출원일자 2010년02월19일
심사청구일자 2010년02월19일

(71) 출원인
연세대학교 산학협력단
서울 서대문구 신촌동 134 연세대학교
(72) 발명자
박상현
서울특별시 송파구 잠실동 레이크팰리스 127동 2501호
박치현
서울특별시 은평구 갈현동 522-50
(뒷면에 계속)
(74) 대리인
특허법인우인

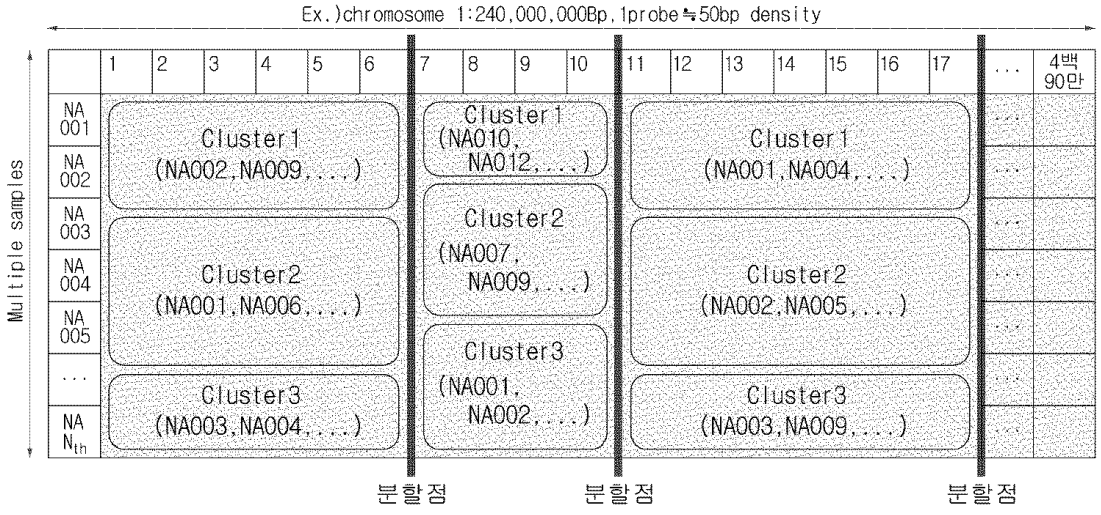
전체 청구항 수 : 총 19 항

(54) 유전체단위반복변이 검출장치 및 방법

(57) 요 약

본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출장치 및 방법은 어레이 발현값 데이터(aCGH data)상의 서로 인접한 열벡터들을 대비하고, 대비 결과에 따라 그 어레이 발현값 데이터를 복수의 세그먼트들로 구획하고, 세그먼트마다 그 세그먼트 내의 행벡터들을 서로 대비하고 대비 결과에 따라 그 세그먼트를 기 설정된 개수의 클러스터들로 재구성하고, 세그먼트마다 그 클러스터들의 분포 형태에 상응하여 그 세그먼트를 후보적 유전체단위반복변이구역으로서 선택적으로 결정하고, 그 후보적 유전체단위반복변이구역 내에서 샘플별로 유전체단위반복변이를 검출하고, 그 후보적 유전체단위반복변이구역(들)에 대해 머징(merging)과 프루닝(pruning)을 수행하여 최종적 유전체단위반복변이구역(들)을 획득한다.

대 표 도



(72) 발명자

안재균

서울특별시 금천구 독산4동 181-4

윤영미

서울특별시 양천구 신정1동 311 목동아파트 1021동 1102호

특허청구의 범위

청구항 1

유전체의 프로브들 각각 및 복수의 샘플들 각각마다 발현값을 나타내는 어레이 발현값 데이터상의 서로 인접한 열벡터들을 대비하고, 대비 결과에 따라 상기 어레이 발현값 데이터를 복수의 세그먼트들로 구획하는 구획부;

상기 세그먼트마다 상기 세그먼트 내의 행벡터들을 서로 대비하고 대비 결과에 따라 상기 세그먼트를 기 설정된 개수의 클러스터들로 재구성하는 클러스터링부; 및

상기 세그먼트마다, 상기 세그먼트 내의 상기 클러스터들의 분포형태에 상응하여 상기 세그먼트를 유전체단위반복변이구역으로 선택적으로 결정하는 결정부를 포함하는 것을 특징으로 하는 유전체단위반복변이 검출장치.

청구항 2

제1 항에 있어서,

상기 유전체단위반복변이 검출장치는 상기 유전체단위반복변이구역에서 상기 샘플마다 유전체단위반복변이를 검출하는 유전체단위반복변이 검출장치.

청구항 3

제1 항에 있어서, 상기 구획부는

상기 인접한 열벡터들간의 상관도와 거리를 고려하여 상기 인접한 열벡터들간을 선택적으로 분할하여, 상기 어레이 발현값 데이터를 상기 세그먼트들로 구획하는 것을 특징으로 하는 유전체단위반복변이 검출장치.

청구항 4

제1 항에 있어서, 상기 클러스터링부는

상기 세그먼트마다, 서로 인접한 값을 갖는 상기 행벡터끼리 그룹핑하여 상기 기 설정된 개수의 클러스터들을 생성하는 유전체단위반복변이 검출장치.

청구항 5

제4 항에 있어서, 상기 클러스터링부는

상기 세그먼트마다, 상기 행벡터들 각각의 대표값을 서로 대비하고 대표값이 일정 범위내에서 서로 유사한 상기 행벡터끼리 그룹핑하여 상기 기 설정된 개수의 클러스터들을 생성하는 것을 특징으로 하는 유전체단위반복변이 검출장치.

청구항 6

제1 항에 있어서,

상기 유전체단위반복변이 검출장치는 상기 어레이 발현값 데이터상의 노이즈를 제거하는 스무딩부를 더 포함하고, 상기 구획부에 주어진 상기 어레이 발현값 데이터는 상기 노이즈가 제거된 어레이 발현값 데이터인 유전체단위반복변이 검출장치.

청구항 7

제6 항에 있어서, 상기 스무딩부는

상기 샘플들 각각마다, 상기 프로브의 발현값을 상기 프로브를 포함한 기 설정된 개수의 프로브들의 발현값들의 대표값으로 대체함으로써 상기 노이즈를 제거하는 것을 특징으로 하는 유전체단위반복변이 검출장치.

청구항 8

제1 항에 있어서, 상기 결정부는

상기 세그먼트마다, 상기 세그먼트내의 상기 클러스터들 각각의 중심값 간의 차이의 절대값의 합을 고려하여 상기 세그먼트를 후보적 상기 유도체단위반복변이구역으로 결정하는 것을 특징으로 하는 유도체단위반복변이 검출장치.

청구항 9

제8 항에 있어서,

상기 결정부는 상기 후보적 유전체단위반복변이구역에 대해 머징(merging)과 프루닝(pruning)을 수행하여 최종적 유전체단위반복변이구역을 획득하는 것을 특징으로 하는 유전체단위반복변이 검출장치.

청구항 10

유전체의 프로브들 각각 및 복수의 샘플들 각각마다 발현값을 나타내는 어레이 발현값 데이터상의 서로 인접한 열벡터들을 대비하고, 대비 결과에 따라 상기 어레이 발현값 데이터를 복수의 세그먼트들로 구획하는 단계;

상기 세그먼트마다 상기 세그먼트 내의 행벡터들을 서로 대비하고 대비 결과에 따라 상기 세그먼트를 기 설정된 개수의 클러스터들로 재구성하는 단계; 및

상기 세그먼트마다, 상기 세그먼트 내의 상기 클러스터들의 분포형태에 상응하여 상기 세그먼트를 유전체단위반복변이구역으로 선택적으로 결정하는 단계를 포함하는 것을 특징으로 하는 유전체단위반복변이 검출방법.

청구항 11

제10 항에 있어서,

상기 유전체단위반복변이 검출방법은 상기 유전체단위반복변이구역에서 상기 샘플마다 유전체단위반복변이를 검출하는 단계를 더 포함하는 것을 특징으로 하는 유전체단위반복변이 검출방법.

청구항 12

제10 항에 있어서, 상기 구획하는 단계는

상기 인접한 열벡터들간의 상관도와 거리를 고려하여 상기 인접한 열벡터들간을 선택적으로 분할하여, 상기 어레이 발현값 데이터를 상기 세그먼트들로 구획하는 것을 특징으로 하는 유전체단위반복변이 검출방법.

청구항 13

제10 항에 있어서, 상기 재구성하는 단계는

상기 세그먼트마다, 서로 인접한 값을 갖는 상기 행벡터끼리 그룹핑하여 상기 기 설정된 개수의 클러스터들을 생성하는 유전체단위반복변이 검출방법.

청구항 14

제13 항에 있어서, 상기 재구성하는 단계는

상기 세그먼트마다, 상기 행벡터들 각각의 대표값을 서로 대비하고 대표값이 일정 범위내에서 서로 유사한 상기 행벡터끼리 그룹핑하여 상기 기 설정된 개수의 클러스터들을 생성하는 것을 특징으로 하는 유전체단위반복변이 검출방법.

청구항 15

제10 항에 있어서,

상기 유전체단위반복변이 검출방법은 상기 어레이 발현값 데이터상의 노이즈를 제거하는 단계를 더 포함하고, 상기 구획하는 단계에 주어진 상기 어레이 발현값 데이터는 상기 노이즈가 제거된 어레이 발현값 데이터인 유전체단위반복변이 검출방법.

청구항 16

제15 항에 있어서, 상기 제거하는 단계는

상기 샘플들 각각마다, 상기 프로브의 발현값을 상기 프로브를 포함한 기 설정된 개수의 프로브들의 발현값들의 대표값으로 대체함으로써 상기 노이즈를 제거하는 것을 특징으로 하는 유전체단위반복변이 검출방법.

청구항 17

제10 항에 있어서, 상기 결정하는 단계는

상기 세그먼트마다, 상기 세그먼트내의 상기 클러스터들 각각의 중심값 간의 차이의 절대값의 합을 고려하여 상기 세그먼트를 후보적 상기 유도체단위반복변이구역으로 결정하는 것을 특징으로 하는 유도체단위반복변이 검출방법.

청구항 18

제17 항에 있어서, 상기 결정하는 단계는 상기 후보적 유전체단위반복변이구역에 대해 머징(merging)과 프루닝(pruning)을 수행하여 최종적 유전체단위반복변이구역을 획득하는 것을 특징으로 하는 유전체단위반복변이 검출방법.

청구항 19

제10 항 내지 제18 항 중 어느 한 항의 방법을 컴퓨터에서 실행시키기 위한 컴퓨터 프로그램을 저장한 컴퓨터로 읽을 수 있는 기록매체.

명 세 서

기 술 분 야

[0001] 본 발명은 유전체에 관한 것으로, 보다 상세하게는, 어레이 발현값 데이터(소위, aCGH 데이터)상에서의 유전체단위반복변이 검출 방안에 관한 것이다.

배 경 기 술

[0002] 어레이 발현값 데이터(aCGH (array Comparative Genomic Hybridization) 데이터)는 유전체의 프로브들 각각 및 복수의 샘플들 각각마다 발현값을 나타내는 어레이 형태의 데이터를 의미한다.

[0003] 이러한 발현값들 중 그 값이 임계치를 초과하는 발현값을 유전체단위반복변이(CNV:Copy Number Variation)라 하며, 어레이 발현값 데이터상에서 유전체단위반복변이들(CNVs)을 신속 정확히 검출하는 것은, 염색체의 발현정도를 측정함에 있어 매우 중요한 사항이나, 현재의 검출방안은 고정밀의 어레이 발현값 데이터상에서의 유전체단위반복변이들의 검출, 특히 작은 크기의 유전체단위반복변이들의 검출에 많은 한계를 갖고 있다.

발명의 내용

해결하려는 과제

[0004] 본 발명의 적어도 일 실시예가 이루고자 하는 기술적 과제는, 고정밀의 어레이 발현값 데이터에서의 작은 크기의 유전체단위반복변이들의 검출도 신속 정확히 수행할 수 있는 유전체단위반복변이 검출장치를 제공하는 데 있다.

[0005] 본 발명의 적어도 일 실시예가 이루고자 하는 다른 기술적 과제는, 고정밀의 어레이 발현값 데이터에서의 작은 크기의 유전체단위반복변이들의 검출도 신속 정확히 수행할 수 있는 유전체단위반복변이 검출방법을 제공하는 데 있다.

[0006] 본 발명의 적어도 일 실시예가 이루고자 하는 또 다른 기술적 과제는 고정밀의 어레이 발현값 데이터에서의 작은 크기의 유전체단위반복변이들의 검출도 신속 정확히 수행할 수 있도록 하는 검출방법을 컴퓨터에서 실행시키기 위한 컴퓨터 프로그램을 저장한 컴퓨터로 읽을 수 있는 기록매체를 제공하는 데 있다.

과제의 해결 수단

[0007] 상기 과제를 이루기 위해, 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출장치는 유전체의 프로브

들 각각 및 복수의 샘플들 각각마다 발현값을 나타내는 어레이 발현값 데이터상에서 서로 인접한 열벡터들을 대비하고, 대비 결과에 따라 상기 어레이 발현값 데이터를 복수의 세그먼트들로 구획하는 구획부; 상기 세그먼트마다 상기 세그먼트 내의 행벡터들을 서로 대비하고 대비 결과에 따라 상기 세그먼트를 기 설정된 개수의 클러스터들로 재구성하는 클러스터링부; 및 상기 세그먼트마다, 상기 클러스터들의 분포 형태에 상응하여 상기 세그먼트를 유전체단위반복변이구역으로서 선택적으로 결정하는 결정부를 포함한다.

[0008] 여기서, 상기 유전체단위반복변이 검출장치는 상기 유전체단위반복변이구역에서 상기 샘플마다 유전체단위반복변이를 검출할 수 있다.

[0009] 여기서, 상기 구획부는 상기 인접한 열벡터들간의 상관도와 거리를 고려하여 상기 인접한 열벡터들간을 선택적으로 분할하여, 상기 어레이 발현값 데이터를 상기 세그먼트들로 구획할 수 있다.

[0010] 여기서, 상기 클러스터링부는 상기 세그먼트마다, 서로 인접한 값을 갖는 상기 행벡터끼리 그룹핑하여 상기 기 설정된 개수의 클러스터들을 생성할 수 있다. 이 때, 상기 클러스터링부는 상기 세그먼트마다, 상기 행벡터들 각각의 대표값을 서로 대비하고 대표값이 일정 범위내에서 서로 유사한 상기 행벡터끼리 그룹핑하여 상기 기 설정된 개수의 클러스터들을 생성할 수 있다.

[0011] 여기서, 상기 유전체단위반복변이 검출장치는 상기 어레이 발현값 데이터상의 노이즈를 제거하는 스무딩부를 더 포함하고, 상기 구획부에 주어진 상기 어레이 발현값 데이터는 상기 노이즈가 제거된 어레이 발현값 데이터일 수 있다. 이 때, 상기 스무딩부는 상기 샘플들 각각마다, 상기 프로브의 발현값을 상기 프로브를 포함한 기 설정된 개수의 프로브들의 발현값들의 대표값으로 대체함으로써 상기 노이즈를 제거할 수 있다.

[0012] 여기서, 상기 결정부는 상기 세그먼트마다, 상기 세그먼트내의 상기 클러스터들 각각의 중심값 간의 차이의 절대값의 합을 고려하여 상기 세그먼트를 후보적 상기 유도체단위반복변이구역으로 결정할 수 있다. 이 때, 상기 결정부는 상기 후보적 유전체단위반복변이구역에 대해 머징(merging)과 프루닝(pruning)을 수행하여 최종적 유전체단위반복변이구역을 획득할 수 있다.

[0013] 상기 다른 기술적 과제를 해결하기 위해, 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출방법은 유전체의 프로브들 각각 및 복수의 샘플들 각각마다 발현값을 나타내는 어레이 발현값 데이터상에서 서로 인접한 열벡터들을 대비하고, 대비 결과에 따라 상기 어레이 발현값 데이터를 복수의 세그먼트들로 구획하는 단계; 상기 세그먼트마다 상기 세그먼트 내의 행벡터들을 서로 대비하고 대비 결과에 따라 상기 세그먼트를 기 설정된 개수의 클러스터들로 재구성하는 단계; 및 상기 세그먼트마다, 상기 클러스터들의 분포 형태에 상응하여 상기 세그먼트를 유전체단위반복변이구역으로서 선택적으로 결정하는 단계를 포함한다.

[0014] 여기서, 상기 유전체단위반복변이 검출방법은 상기 유전체단위반복변이구역에서 상기 샘플마다 유전체단위반복변이를 검출하는 단계를 더 포함할 수 있다.

[0015] 여기서 상기 구획하는 단계는 상기 인접한 열벡터들간의 상관도와 거리를 고려하여 상기 인접한 열벡터들간을 선택적으로 분할하여, 상기 어레이 발현값 데이터를 상기 세그먼트들로 구획할 수 있다.

[0016] 여기서, 상기 재구성하는 단계는 상기 세그먼트마다, 서로 인접한 값을 갖는 상기 행벡터끼리 그룹핑하여 상기 기 설정된 개수의 클러스터들을 생성할 수 있다. 이 때, 상기 재구성하는 단계는 상기 세그먼트마다, 상기 행벡터들 각각의 대표값을 서로 대비하고 대표값이 일정 범위내에서 서로 유사한 상기 행벡터끼리 그룹핑하여 상기 기 설정된 개수의 클러스터들을 생성할 수 있다.

[0017] 여기서, 상기 유전체단위반복변이 검출방법은 상기 어레이 발현값 데이터상의 노이즈를 제거하는 단계를 더 포함하고, 상기 구획하는 단계에 주어진 상기 어레이 발현값 데이터는 상기 노이즈가 제거된 어레이 발현값 데이터일 수 있다. 이 때, 상기 제거하는 단계는 상기 샘플들 각각마다, 상기 프로브의 발현값을 상기 프로브를 포함한 기 설정된 개수의 프로브들의 발현값들의 대표값으로 대체함으로써 상기 노이즈를 제거할 수 있다.

[0018] 여기서, 상기 결정하는 단계는 상기 세그먼트마다, 상기 세그먼트내의 상기 클러스터들 각각의 중심값 간의 차이의 절대값의 합을 고려하여 상기 세그먼트를 후보적 상기 유도체단위반복변이구역으로 결정할 수 있다. 이 때, 상기 결정하는 단계는 상기 후보적 유전체단위반복변이구역에 대해 머징(merging), 프루닝(pruning)을 수행하여 최종적 유전체단위반복변이구역을 획득할 수 있다.

[0019] 상기 또 다른 기술적 과제를 해결하기 위해, 본 발명의 적어도 일 실시예에 따른 컴퓨터로 읽을 수 있는 기록매체는 유전체의 프로브들 각각 및 복수의 샘플들 각각마다 발현값을 나타내는 어레이 발현값 데이터상에서 서로 인접한 열벡터들을 대비하고, 대비 결과에 따라 상기 어레이 발현값 데이터를 복수의 세그먼트들로 구획하는 단

계; 상기 세그먼트마다 상기 세그먼트 내의 행벡터들을 서로 대비하고 대비 결과에 따라 상기 세그먼트를 기 설정된 개수의 클러스터들로 재구성하는 단계; 및 상기 세그먼트마다, 상기 클러스터들의 분포 형태에 상응하여 상기 세그먼트를 유전체단위반복변이구역으로서 선택적으로 결정하는 단계를 포함하는 유전체단위반복변이 검출 방법을 컴퓨터에서 실행시키기 위한 컴퓨터 프로그램을 저장할 수 있다.

발명의 효과

[0020] 본 발명의 적어도 일 실시예에 따르면, 고정밀의 어레이 발현값 데이터에서의 작은 크기의 유전체단위반복변이들의 검출이라 하더라도 신속 정확히 수행할 수 있고, 이로써, 본 발명의 적어도 일 실시예는 유전체의 발현 정도를 신속 정확히 측정할 수 있다.

도면의 간단한 설명

[0021] 도 1은 어레이 발현값 데이터를 설명하기 위한 도면이다.

도 2는 CNV, CNVR, CNVE, CNVZ를 설명하기 위한 도면이다.

도 3은 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출장치를 나타내는 블록도이다.

도 4는 어레이 발현값 데이터의 raw data를 설명하기 위한 도면이다.

도 5 및 도 6은 도 3에 도시된 스무딩부의 동작을 상세히 설명하기 위한 도면들이다.

도 7 및 도 8은 도 3에 도시된 구획부의 동작을 설명하기 위한 도면들이다.

도 9 및 도 10은 도 3에 도시된 클러스터링부의 동작을 설명하기 위한 도면들이다.

도 11은 도 3에 도시된 결정부의 동작을 설명하기 위한 도면이다.

도 12는 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출방법을 나타내는 플로우챠트이다.

발명을 실시하기 위한 구체적인 내용

[0022] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시에 의하여 달성되는 목적을 충분히 이해하기 위해서는 본 발명의 바람직한 실시예를 예시하는 첨부 도면 및 그 첨부 도면을 설명하는 내용을 참조하여야만 한다.

[0023] 이하, 본 발명의 적어도 일 실시예에 의한 유전체단위반복변이 검출장치 및 방법을 첨부한 도면들을 참조하여 다음과 같이 설명한다.

[0024] 도 1은 어레이 발현값 데이터를 설명하기 위한 도면이다.

[0025] 앞서 설명한 바와 같이, 어레이 발현값 데이터는 aCGH (array Comparative Genomic Hybridization) 데이터라 명명 가능하며, 이는 '유전체의 프로브(probe)들 각각' 및 '복수의 샘플들 각각' 마다 "발현값"을 나타내는 어레이(array) 형태의 데이터를 의미한다. 본 명세서에서 '프로브'란 DNA 칩(chip) 위에 올려지는 유전체 조각으로서 칩에 심겨지는 기본 단위를 의미하고, '샘플'은 어떤 생물체(예를 들어, 인체)의 유전체를 의미하고 이러한 샘플은 여러 개의 프로브들로 나뉘며 그 프로브들 각각이 칩에 심겨진다.

[0026] 도 1에 도시된 바와 같이, 어레이 발현값 데이터에서 각 행은 개개의 샘플을 의미하고, 각 열은 개개의 프로브를 의미한다. 도 1의 경우 하나의 유전체(P)는 m개(단 m은 2이상의 정수)의 프로브들로 나뉘고, 어레이 발현값 데이터는 총 n(단, n은 2이상의 정수)개의 샘플들에 대한 데이터이고, 각각의 샘플마다 각각의 프로브에 대한 발현값을 나타낸다. 도 1에 도시된 바에서 (단, p는 1≤p≤n 인 정수)는 p번째 샘플의 1번째 프로브에서의 발현값을 의미하고, 는 p번째 샘플의 4번째 프로브에서의 발현값을 의미하고, 는 p번째 샘플의 g번째(단, g는 1≤g≤m인 정수) 프로브에서의 발현값을 의미한다.

[0027] 도 2는 CNVs, CNVR, CNVE, 및 CNVZ를 설명하기 위한 도면이다. 도 2에서는 설명의 편의상, 3개의 샘플에 대해서만 설명하겠으나 도 1에서와 같이 40개의 샘플에 대해서도 동일하게 설명될 수 있음은 물론이다.

[0028] CNVs는 유전체단위반복변이 '들'을 의미하고, CNVR은 샘플들 중 한 샘플이라도 CNVs가 존재하는 구간을 의미하며, CNVE는 샘플들 각각의 CNVs 간에 51%이상 겹쳐지는(overlapped) 구간을 의미하고, CNVZ는 본 발명의 적어도 일 실시예에 따른 '유전체단위반복변이구역'을 의미하며, 이러한 유전체단위반복변이구역을 결정하는 방안을 이하 도 3 내지 도 12를 통해 설명할 것이다.

[0029] 본 발명의 적어도 일 실시예는 '어레이 발현값 데이터' 상의 '유전체단위반복변이' 를 '후보적 유전체단위반복변이구역' (candidate CNVZ)을 결정한 뒤 그 결정된 후보적 유전체단위반복변이구역내에서 검출하며, 그 결정된 '후보적 유전체단위반복변이구역' 에 대해 후술할 머징(merging)과 프루닝(pruning)을 수행하여 '최종적 유전체단위반복변이구역' (final CNVZ)을 결정한다.

[0030] 도 3은 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출장치를 나타내는 블록도이고, 도 4는 어레이 발현값 데이터의 raw data를 설명하기 위한 도면이고, 도 5 및 도 6은 도 3에 도시된 스무딩부의 동작을 상세히 설명하기 위한 도면들이고, 도 7 및 도 8은 도 3에 도시된 구획부의 동작을 설명하기 위한 도면들이고, 도 9 및 도 10은 도 3에 도시된 클러스터링부의 동작을 설명하기 위한 도면들이고, 도 11은 도 3에 도시된 결정부의 동작을 설명하기 위한 도면이다.

[0031] 도 3에 도시된 바에 따르면, 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출장치는 스무딩부(310), 구획부(320), 클러스터링부(330), 결정부(340), 검출부(350)를 포함할 수 있으며, 이하 도 3을 도 4 내지 도 11을 참조하며 구체적으로 설명한다.

[0032] 스무딩부(310)는 어레이 발현값 데이터상에 존재하는 노이즈(noise)를 제거한다. 어레이 발현값 데이터의 raw data는 도 4를 참조할 것이며, 도 5 내지 도 11은 도 4의 어레이 발현값 데이터에 대한 설명을 위한 도면들이다. 도4는 40명의 샘플들 각각의 유전체에 대한 발현값들을 나타내고, 각각의 유전체가 4백90만개의 프로브들로 구성되어 있는 경우로서 그 프로브들 각각마다 발현값을 나타낸다. 도 4에서 'size of chr1: 240,000,000 bp' 란 염색체(chromosome)의 크기가 240,000,000 base pair 임을 의미하고, '1probe ≒ 50bp density' 라 함은 하나의 프로브의 길이가 대략 50 base pair를 커버하는 길이임을 의미한다.

[0033] 구체적으로, 스무딩부(310)는 샘플들 각각마다 '어떤 한 프로브' 의 발현값을 그 어떤 한 프로브를 포함한 기 설정된 개수의 프로브들의 발현값들의 대표값으로 대체하는 과정을 모든 프로브들에 대해 수행함으로써, 어레이 발현값 데이터상의 노이즈를 제거한다. 여기서 어떤 한 프로브를 포함한 기 설정된 개수의 프로브들은 그 어떤 한 프로브와 이웃한 기 설정된 개수의 프로브들을 의미하고, '대표값' 이란 설명의 편의상 '평균값' 이라 가정한다. 이를 도 5 및 도 6을 참조하여 설명하면, 6*40의 행렬 형태의 윈도우인 슬라이딩 윈도우(sliding window)가 도5에 도시된 바와 같이 위치한 상태에서 스무딩부(310)는 각 샘플마다 '1번째 프로브' 에 해당하는 발현값을 '1번째 프로브 내지 6번째 프로브에 해당하는 6개의 발현값들의 평균값' 으로 대체하고(즉, 1번째 샘플의 1번째 프로브의 발현값을 1번째 샘플의 1번째 내지 6번째 프로브들의 발현값들의 평균값으로 대체하고, 2번째 샘플의 1번째 프로브의 발현값을 1번째 샘플의 1번째 내지 6번째 프로브들의 발현값들의 평균값으로 대체하는 등), 그 슬라이딩 윈도우를 우측으로 1 프로브만큼 이동시킨 뒤, 각 샘플마다 '2번째 프로브' 에 해당하는 발현값을 '2번째 프로브 내지 7번째 프로브에 해당하는 6개의 발현값들의 평균값' 으로 대체하고, 이와 같은 일련의 과정을 어레이 발현값 데이터상의 모든 프로브들에 대해 수행하여 어레이 발현값 데이터상의 노이즈를 모두 제거할 수 있다. 도 5, 도 6에 도시된 바와 같은 크기의 슬라이딩 윈도우는 설명의 편의상 설정된 크기의 슬라이딩 윈도우이며 이의 다양한 변형실시가 가능함은 물론이다.

[0034] 스무딩부(310)는 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출장치에 도 3에 도시된 바와 같이 포함될 수도 있고, 도 3에 도시된 바와 달리 포함되지 않을 수도 있다.

[0035] 구획부(320)는 어레이 발현값 데이터상에서 서로 인접한 열벡터들을 대비하고, 대비 결과에 따라 어레이 발현값 데이터를 복수의 세그먼트(segment)들로 구획한다. 본 명세서에서 열벡터란 어레이 발현값 데이터상의 열벡터 즉, 동일한 프로브에 대한 모든 샘플들 각각에서의 발현값을 나타내는 벡터를 의미한다. 같은 원리로, 후술할 행벡터란 어레이 발현값 데이터상의 행벡터 즉, 동일한 샘플에 대한 모든 프로브들 각각에서의 발현값을 나타내는 벡터를 의미한다.

[0036] 즉, 구획부(320)는 q(단, q는 1≤q<4백 90만)번째 열벡터와 (q+1)번째 열벡터를 서로 대비하고 대비 결과를 고려하여, q번째 열벡터와 (q+1)번째 열벡터 사이를 분할(break)할지의 여부를 결정한다. 이와 같은 결정에 따라 구획부(320)가 분할을 수행한 경우 그 분할된 영역들 각각이 바로 '세그먼트' 가 되는 것이다.

[0037] 구체적으로, 구획부(320)는 '어레이 발현값 데이터상에서의 인접한 열벡터들마다' 그 인접한 열벡터들간의 상관도 및 거리를 고려하여 그 인접한 열벡터들간을 선택적으로 분할하여 그 어레이 발현값 데이터를 복수의 세그먼트들로 구획한다. 여기서, 상관도란 그 인접한 열벡터들간의 상관 계수를 의미하며, 1로 갈수록 양자는 양의 상관관계를 가지며 -1로 갈수록 양자는 음의 상관관계를 가지며 0은 양자간에 아무런 상관관계가 없음을 의미한다. Pearson's Correlation Coefficient(PCC)는 이러한 '상관도' 의 일 례이다. 또한 인접한 열벡터들간의 '

거리' 란 인접한 열벡터들간의 상대적 거리를 의미하며 '유클리드 거리(euclidean distance)' 는 이러한 '거리' 의 일례이다.

[0038] 보다 구체적으로, 구획부(320)는 인접한 열벡터들간의 거리가 (기 설정된) 임계 거리 미만이면서 그 인접한 열벡터들간의 상관도가 임계 상관도 이상인 경우, 그 인접한 열벡터들 사이를 분할(break)하지 않는다. 반면, 그 밖의 경우 즉, 인접한 열벡터들간의 거리가 임계 거리 이상이면서 그 인접한 열벡터들간의 거리가 임계상관도 이상인 경우, 그 인접한 열벡터들간의 거리가 임계거리 미만이면서 그 인접한 열벡터들간의 상관도가 임계 상관도 미만인 경우, 그 인접한 열벡터들간의 거리가 임계 거리 이상이면서 그 인접한 열벡터들간의 상관도가 임계 상관도 미만인 경우, 구획부(320)는 그 인접한 열벡터들 사이를 분할한다. 도 7 에서 '그 인접한 열벡터들' 은 '1번째 열벡터와 2번째 열벡터(도 7에서 직사각형으로 묶인 부분)', '2번째 열벡터와 3번째 열벡터', '3번째 열벡터와 4번째 열벡터',... 각각을 의미한다. 도 8은 구획부(320)에 의해 생성된 세그먼트들의 일 례를 나타내며, 도 8의 경우, 6번째 열벡터와 7번째 열벡터 사이에서 분할된 것이고, 10번째 열벡터와 11번재 열벡터 사이에서 분할된 것이다.

[0039] 클러스터링부(330)는 '세그먼트' 마다, 세그먼트 내의 행벡터들을 서로 대비하고 대비 결과에 따라 그 세그먼트를 기 설정된 개수의 클러스터들로 재구성한다.

[0040] 구체적으로, 클러스터링부(330)는 '세그먼트' 마다, 서로 인접한 값을 갖는 행벡터끼리 그룹핑하여 기 설정된 개수의 클러스터들을 생성한다. 보다 구체적으로 클러스터링부(330)는 세그먼트마다, 행벡터들 각각의 대표값을 서로 대비하고 대표값이 일정 범위내에서 서로 유사한 행벡터끼리 그룹핑함으로써, 기 설정된 개수의 클러스터들을 생성한다. 도9를 이용하여 '세그먼트 1' 에 대한 클러스터링부(330)의 동작에 대해 설명하면, 클러스터링부(330)는 '1번째 샘플의 1번째 내지 6번째 프로브들에 대한 발현값들의 평균치', '2 번째 샘플의 1번째 내지 6 번째 프로브들에 대한 발현값들의 평균치', '3 번째 샘플의 1번째 내지 6번째 프로브들에 대한 발현값들의 평균치', ..., '40번째 샘플의 1번째 내지 6번째 프로브들에 대한 발현값들의 평균치' 를 서로 대비하여 서로 유사한 행벡터끼리 그룹핑함으로써 도 10에 도시된 바와 같은 클러스터들을 생성할 수 있다. 도 10의 경우 세그먼트 1이 클러스터 0, 클러스터 1, 클러스터 2로 재구성된 경우를 나타내며, 이 때의 클러스터 0은 2번째 샘플, 9번째 샘플 기타 등등의 행벡터들의 조합을 의미하고, 클러스터 1은 1번째 샘플, 6번째 샘플 기타 등등의 행벡터들의 조합을 의미하고, 클러스터 2는 3번째 샘플, 4번째 샘플 기타 등등의 행벡터들의 조합을 의미한다.

[0041] 클러스터링부(330)는 소위 'K-평균 군집화 기법(K-means clustering 방식)' 에 따라 동작할 수 있다(도 9, 도 10의 경우 K=3).

[0042] 결정부(340)는 '세그먼트' 마다, 세그먼트 내의 클러스터들의 분포 형태에 상응하여 그 세그먼트를 유전체단위반복변이구역으로서 선택적으로 결정한다. 즉 결정부(340)는 세그먼트 내의 클러스터들의 분포 형태를 고려하여, 세그먼트를 유전체단위반복변이구역이라 결정할 수도 있고, 그 세그먼트를 유전체단위반복변이구역이라 결정하지 않을 수도 있다.

[0043] 구체적으로, 결정부(340)는 '세그먼트' 마다, 세그먼트내의 클러스터들 각각의 중심값 간의 차이의 절대값의 합을 고려하여 그 세그먼트를 후보적 유도체단위반복변이구역(candidate CNVZ)으로 결정할 수 있다. 여기서, 클러스터의 중심값이라 함은 클러스터 내 발현값들의 평균값을 의미한다. 이러한 '합' 은 다음의 수학식 1로 표현될 수 있다.

[0044] [수학식 1]

$$SC(seg_g) = \alpha \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \left| C_i - C_j \right| \quad , i \neq j \; and \; i, j \leq k$$

[0045]

[0046] 여기서, k는 K-평균 군집화 기법에서의 K를 의미하고, i, j 각각은 클러스터 각각을 의미하고, Ci, Cj 각각은 i번째 클러스터의 중심값, j번째 클러스터의 중심값 각각을 의미하고, α는 비례 계수를 의미한다. 도 10과 수학식 1을 이용하여 세그먼트 1에 대한 결정부(340)의 동작을 설명하면, 세그먼트 1의 경우 수학식 1의 우측항에서의 α를 제외한 나머지는 세그먼트 1에서의 '클러스터 0의 중심값과 클러스터 1의 중심값간의 차이', '클러스터 1의 중심값과 클러스터 2의 중심값간의 차이', '클러스터 0의 중심값과 클러스터 2의 중심값간의 차이' 의 모든 합을 의미하며, 이러한 '합' 이 크면 SC(즉, score)도 크고 SC가 크면 클수록 클러스터들은 서로 멀리 떨어져 있는 것이고 그렇다면 클러스터에 포함되는 샘플들이 highly positive 혹은 highly negative 한 발현값

들을 가질 가능성이 크기 때문에 세그먼트 1에 대한 SC가 임계치를 초과한다면 결정부(340)는 세그먼트 1을 후보적 유전체단위반복변이구역이라 결정한다. 만일 세그먼트 1 내의 샘플들 모두가 highly positive 혹은 highly negative값인 경우에도 α값을 통한 보정에 의해 SC값은 여전히 높게 나올 수 있고, 이에 따라 세그먼트 1에 대한 SC가 임계치를 초과한다면 결정부(340)가 세그먼트1을 후보적 유전체단위반복변이구역으로 결정할 수 있음은 물론이다.

[0047] 결정부(340)는 후보적 유전체단위반복변이구역에 대해 머징(merging)과 프루닝(pruning)을 수행하여 최종적 유전체단위반복변이구역(final CNVZ)을 획득한다. 여기서, 머징(merging)이란 인접한 후보적 유전체단위반복변이구역들간의 공백이 일정길이 이하(예를 들어 500Bp(base pair))인 경우 그 후보적 유전체단위반복변이구역들을 합쳐 그 후보적 유전체단위반복변이구역들의 시작부터 종료까지 모두를 최종적 유전체단위반복변이구역으로 결정하는 것을 의미한다. 이는 어레이 발현값 데이터에 실험적 오차가 존재할 수 있다는 점과 염색체를 고르게 잘라 hybridization 실험을 한다해도 중간 중간 실험이 잘 안 되는 부분이 있는데 이러한 부분은 본래 유전체단위반복변이가 매우 높은 양 혹은 음의 값을 보여야 함에도 불구하고 0에 가깝게 나올 수 있다는 점을 감안하여 수행되는 작업이다. 한편, 프루닝(pruning)이란 후보적 유전체단위반복변이구역의 길이가 일정 길이(예를 들어, 500 base pair) 이하인 경우 이를 유전체변이로 인식하지 않고 실험적 오차로 간주하여 제거함으로써 그 후보적 유전체단위반복변이구역을 최종적 유전체단위반복변이구역으로 취하지 않는 것을 의미한다. 이는 현재까지 알려진 가장 작은 유전체단위반복변이의 단위가 대략 500Bp 정도의 길이를 가지고 있음에 기인하여 수행되는 과정이다. 물론, 프루닝(pruning)의 수행 여부의 기준이 되는 그 '일정 길이'는 사용자에 의해 설정되기 나름이다.

[0048] 검출부(350)는 각 샘플마다, 후보적 유전체단위반복변이구역에서 유전체단위반복변이(CNV)를 검출한다.

[0049] 도 12는 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출방법을 나타내는 플로우챠트이다.

[0050] 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출장치는 어레이 발현값 데이터상에 존재하는 노이즈를 제거한다(제1210 단계). 다만 제1210 단계는 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출방법에 포함되지 않을 수도 있다.

[0051] 제1210 단계 후에 혹은 제1210 단계를 거치지 않고, 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출장치는 어레이 발현값 데이터상의 서로 인접한 열벡터들을 대비하고, 대비 결과에 따라 그 어레이 발현값 데이터를 복수의 세그먼트들로 구획한다(제1220 단계).

[0052] 제1220 단계 후에, 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출장치는 세그먼트마다 그 세그먼트 내의 행벡터들을 서로 대비하고 대비 결과에 따라 그 세그먼트를 기 설정된 개수의 클러스터들로 재구성한다(제1230 단계).

[0053] 제1230 단계 후에, 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출장치는 세그먼트마다, 그 세그먼트 내의 클러스터들의 분포 형태에 상응하여 그 세그먼트를 후보적 유전체단위반복변이구역으로 선택적으로 결정한다(제1240 단계).

[0054] 제1240 단계에서 후보적 유전체단위반복변이구역으로서 결정되면, 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출장치는 제1240 단계에서 결정된 후보적 유전체단위반복변이구역에서 각 샘플별로 유전체단위반복변이를 검출한다(제1250 단계).

[0055] 제1240 단계 후에, 본 발명의 적어도 일 실시예에 따른 유전체단위반복변이 검출장치는 제1240 단계에서 결정된 후보적 유전체단위반복변이구역(들)에 대해 머징(merging)과 프루닝(pruning)을 수행하여 최종적 유전체단위반복변이구역(들)을 획득한다(제1260 단계).

[0056] 이상에서 언급된 본 발명에 의한 유전체단위반복변이 검출방법을 컴퓨터에서 실행시키기 위한 프로그램은 컴퓨터로 읽을 수 있는 기록매체에 저장될 수 있다.

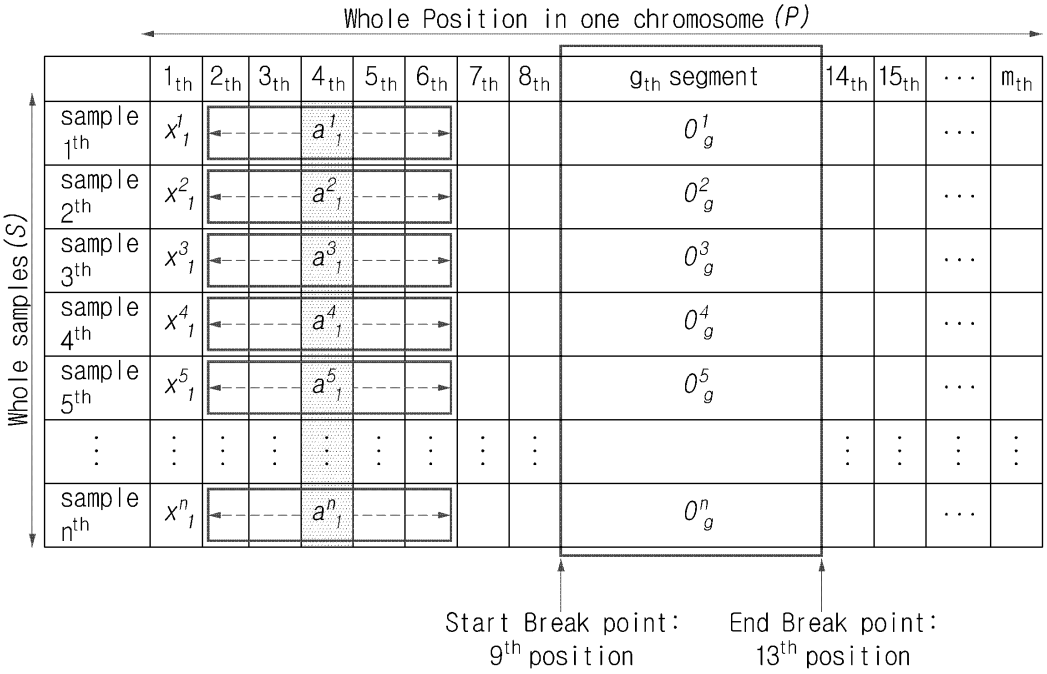
[0057] 여기서, 컴퓨터로 읽을 수 있는 기록매체는 마그네틱 저장매체(예를 들면, 롬(ROM), 플로피 디스크, 하드 디스크 등), 및 광학적 판독 매체(예를 들면, 시디롬(CD-ROM), 디브이디(DVD: Digital Versatile Disc))와 같은 저장매체를 포함한다.

[0058] 이제까지 본 발명을 바람직한 실시예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로, 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든

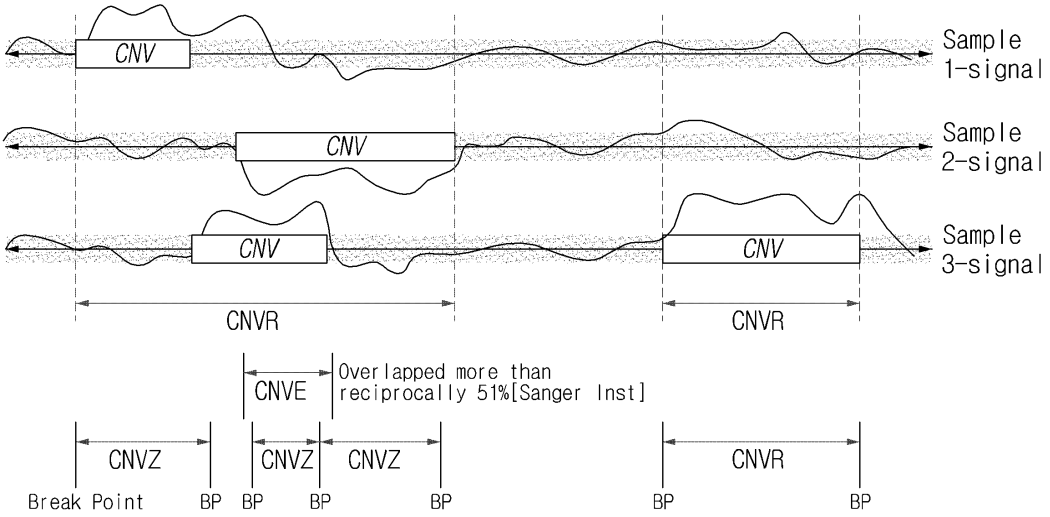
차이점들은 본 발명에 포함된 것으로 해석되어야 할 것이다.

도면

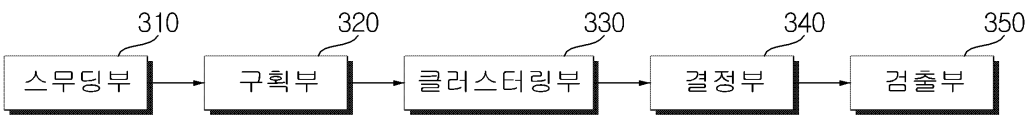
도면1



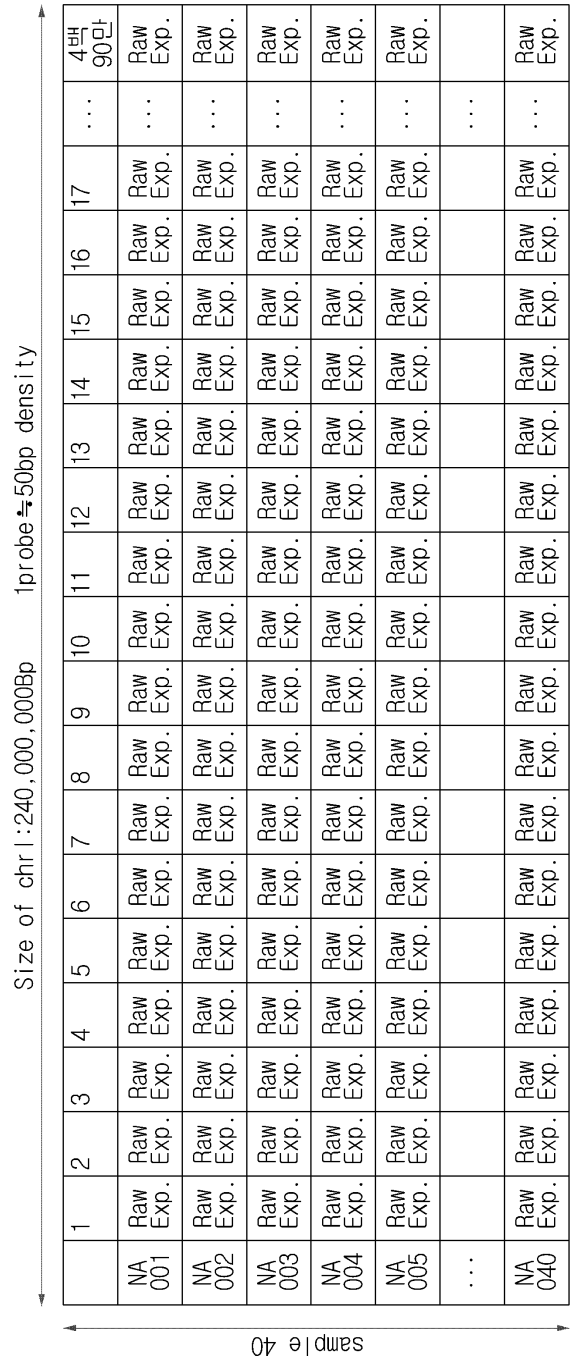
도면2



도면3

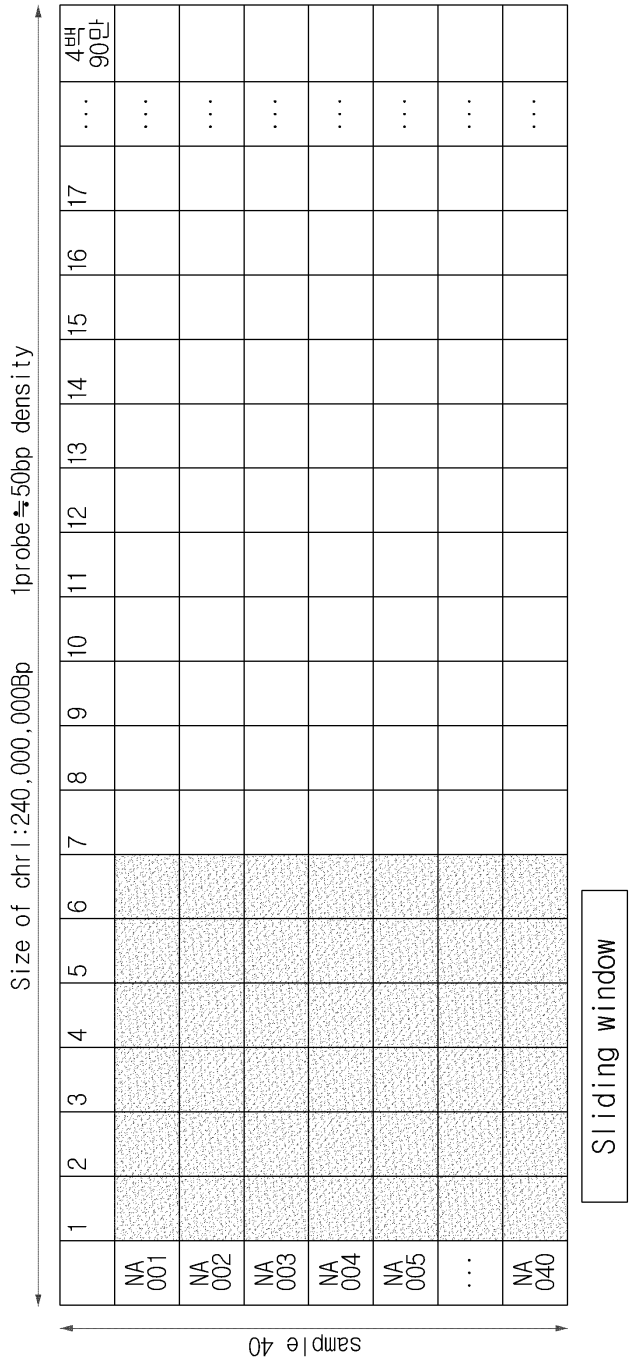


도면4

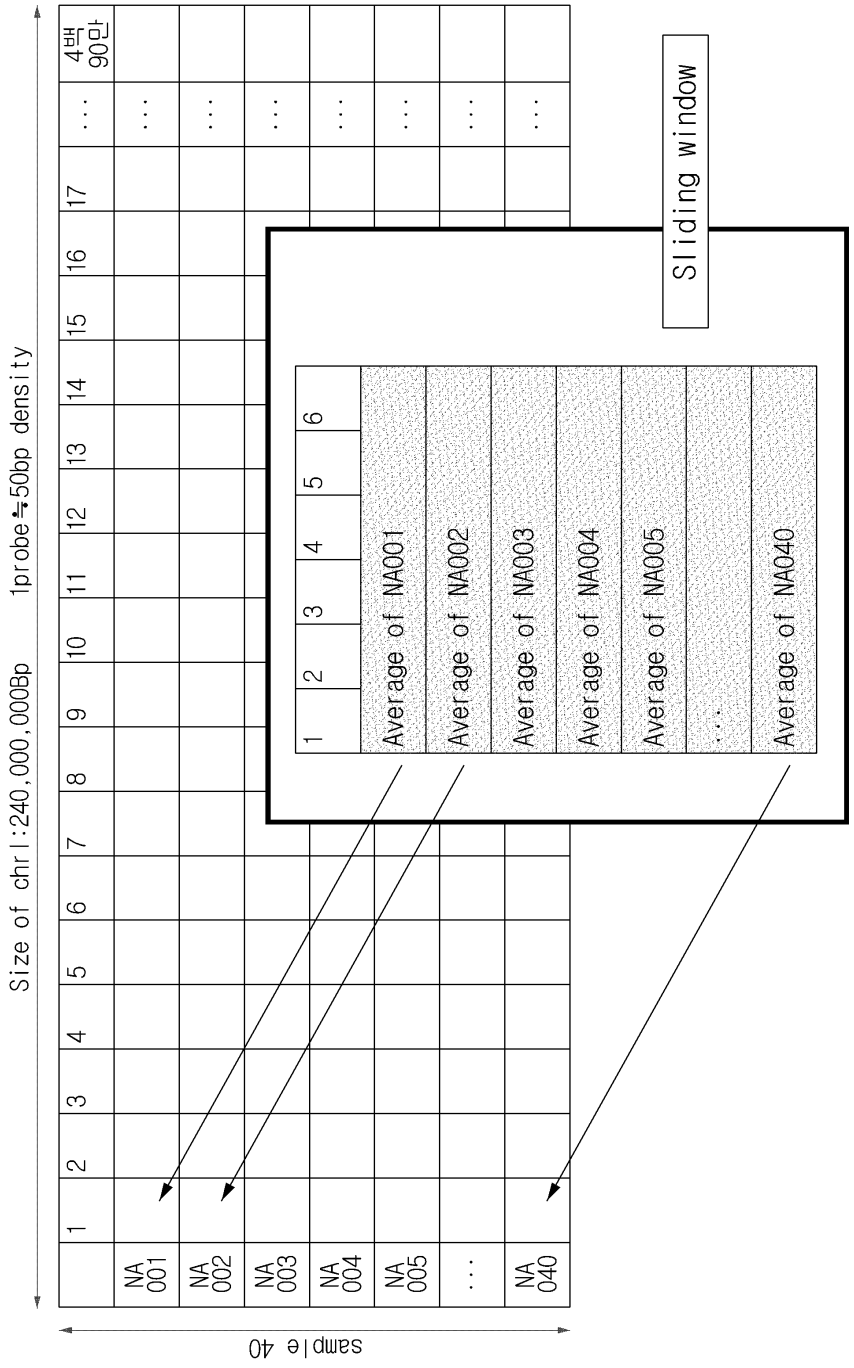


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	...	4백 90만
NA 001	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	...	Raw Exp.
NA 002	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	...	Raw Exp.
NA 003	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	...	Raw Exp.
NA 004	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	...	Raw Exp.
NA 005	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	...	Raw Exp.
...																		...	
NA 040	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	Raw Exp.	...	Raw Exp.

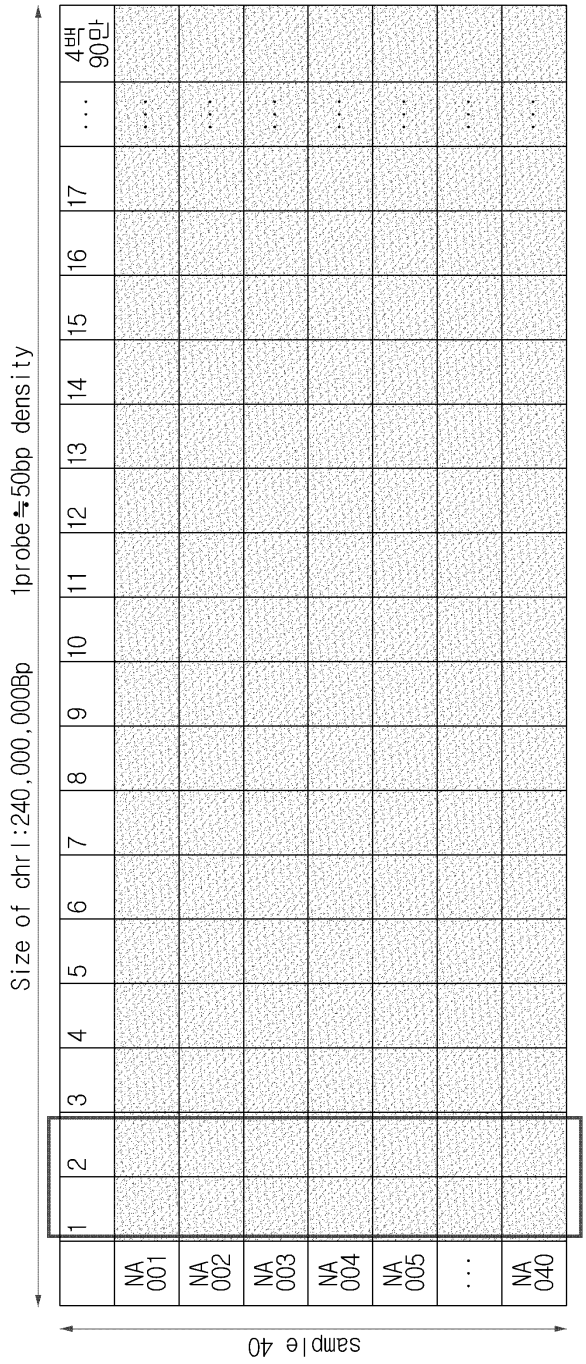
도면5


Size of chr1:240,000,000Bp
1probe≒50bp density
sample 40
Sliding window
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
...
4백 90만
NA 001
NA 002
NA 003
NA 004
NA 005
...
NA 040

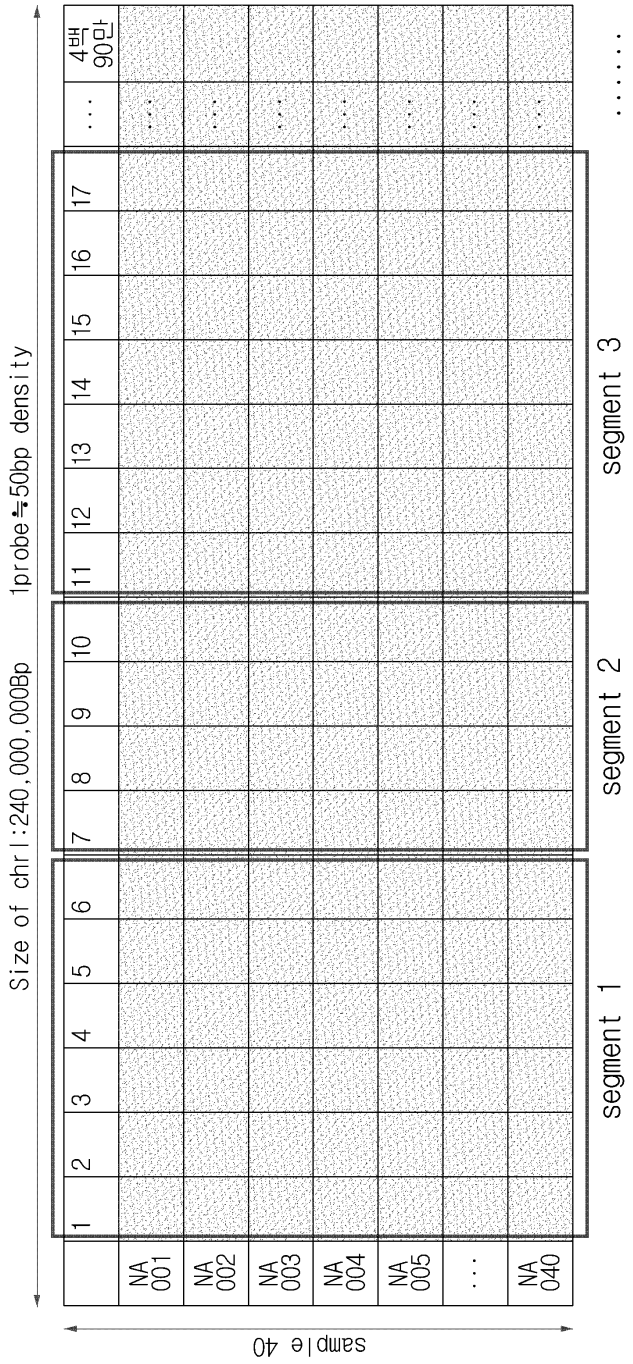
도면6


Size of chr1:240,000,000Bp
1probe≒50bp density
sample 40
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
...
4백 90만
NA 001
NA 002
NA 003
NA 004
NA 005
...
NA 040
Sliding window
1
2
3
4
5
6
Average of NA001
Average of NA002
Average of NA003
Average of NA004
Average of NA005
...
Average of NA040

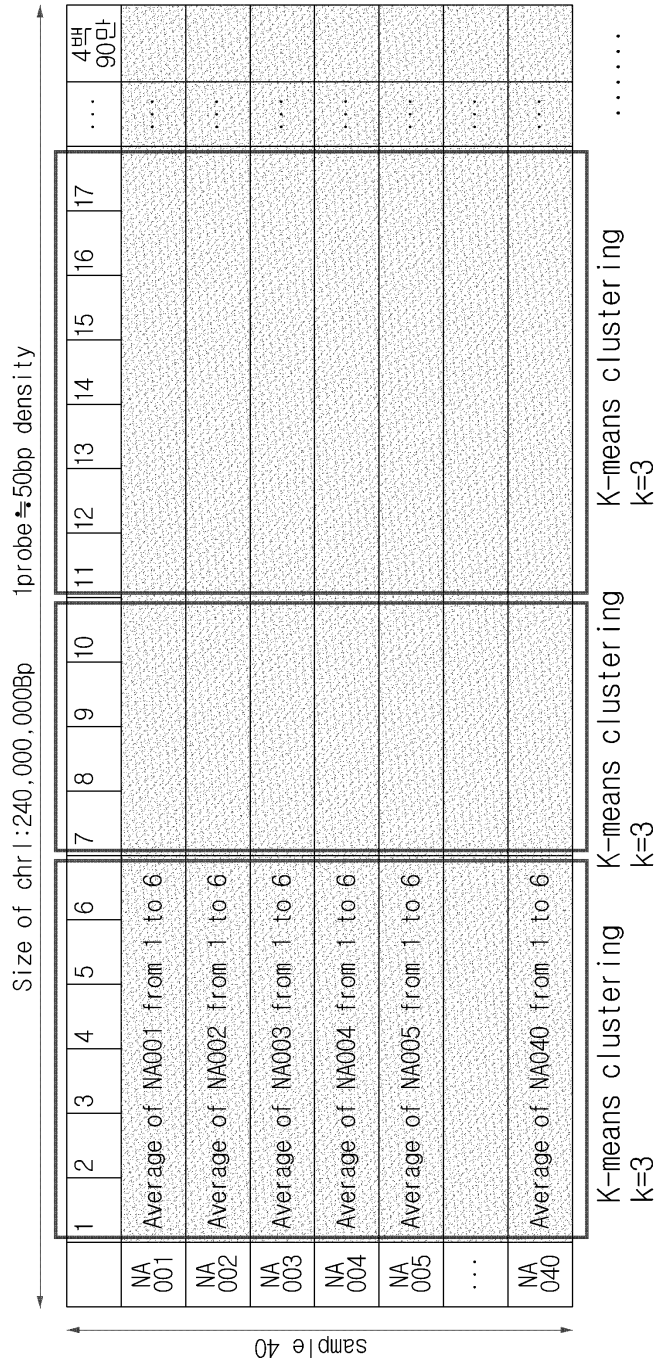
도면7


Size of chr1:240,000,000Bp
1probe≒50bp density
sample 40
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
...
4백 90만
NA 001
NA 002
NA 003
NA 004
NA 005
...
NA 040

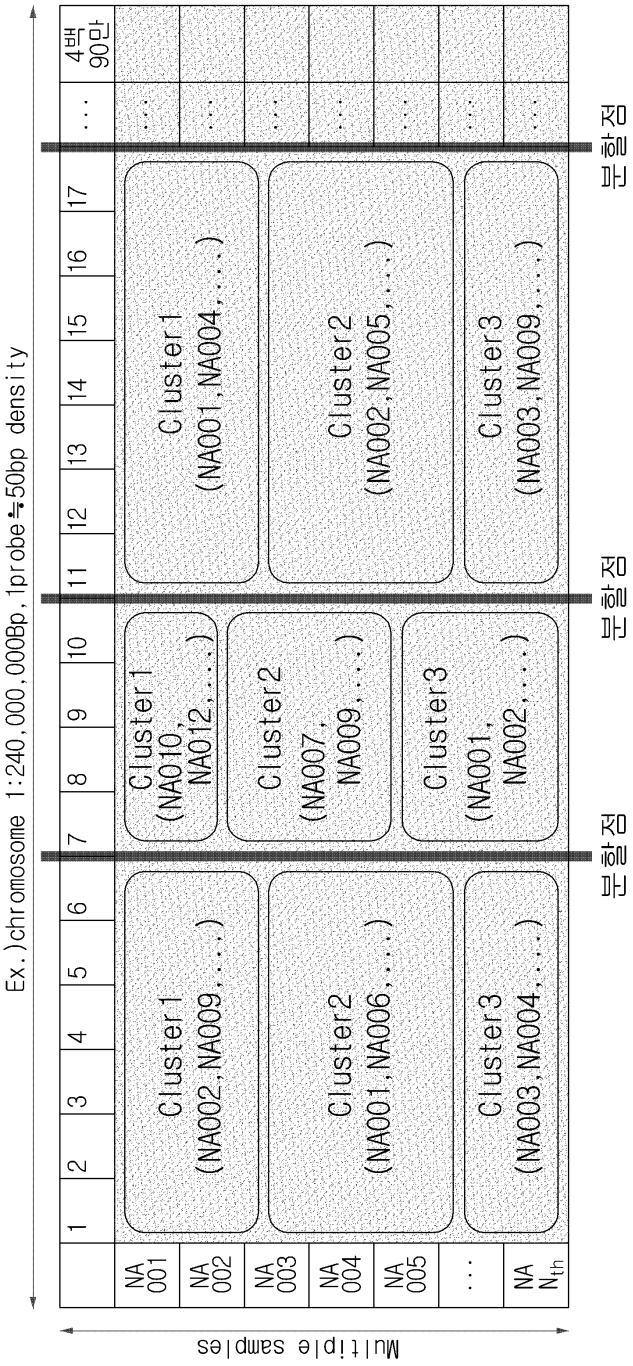
도면8


Size of chrI:240,000,000Bp
1probe≒50bp density
sample 40
segment 1
segment 2
segment 3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
...
4백 90만
NA 001
NA 002
NA 003
NA 004
NA 005
...
NA 040

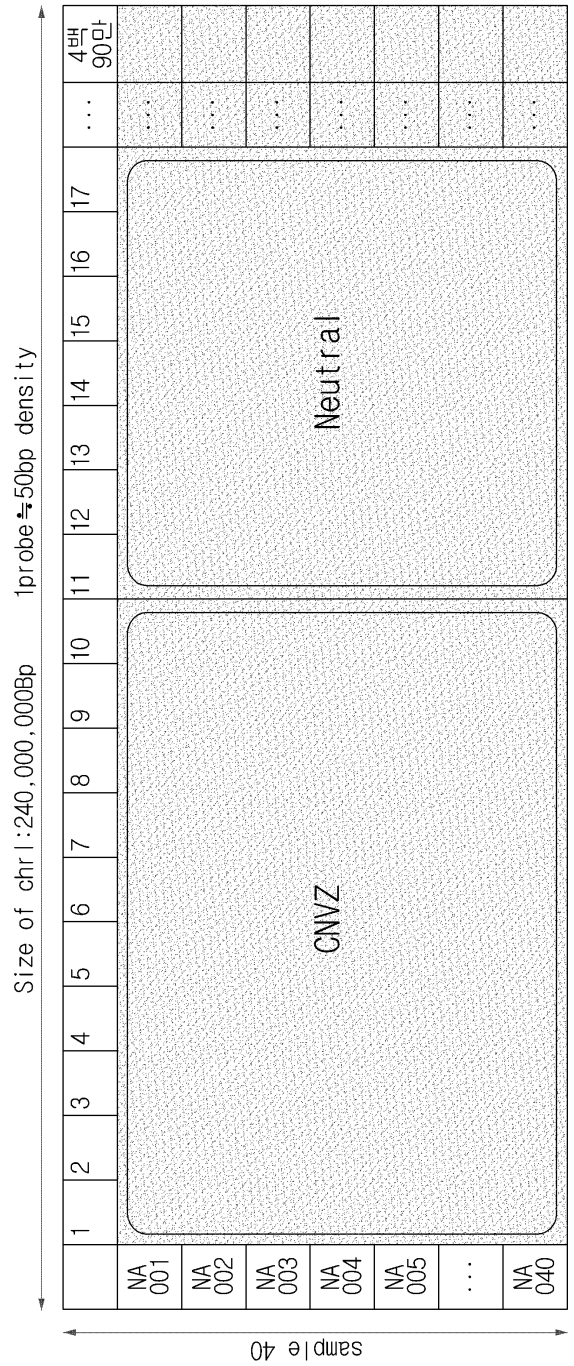
도면9


Size of chr1:240,000,000Bp
1probe≒50bp density
sample 40
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
...
4백 90만
NA 001
NA 002
NA 003
NA 004
NA 005
...
NA 040
Average of NA001 from 1 to 6
Average of NA002 from 1 to 6
Average of NA003 from 1 to 6
Average of NA004 from 1 to 6
Average of NA005 from 1 to 6
Average of NA040 from 1 to 6
K-means clustering k=3
K-means clustering k=3
K-means clustering k=3
.....

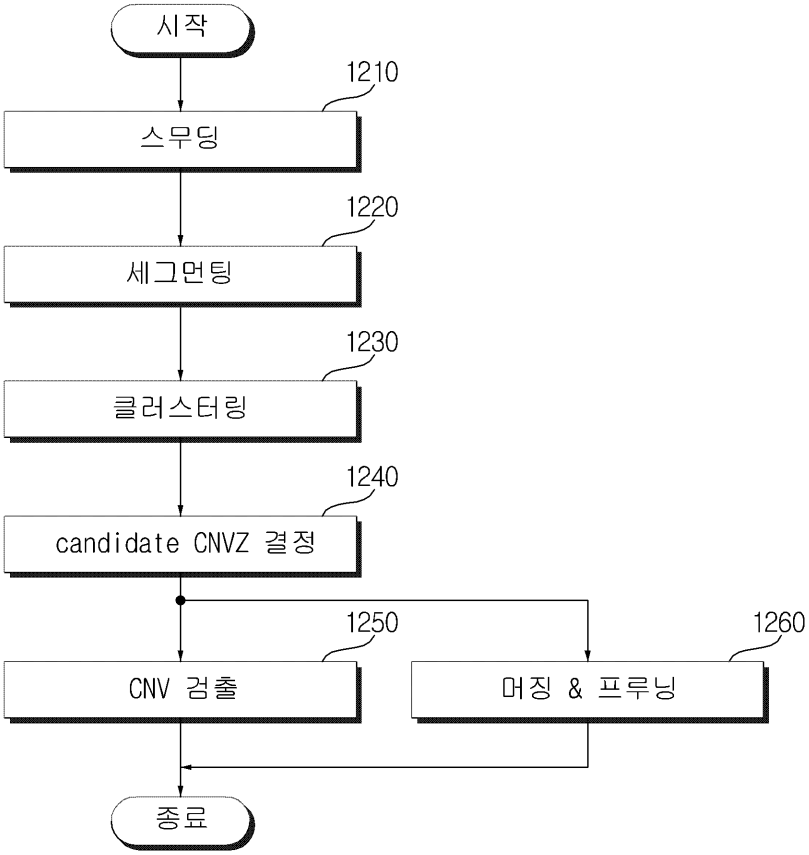
도면10


Ex.)chromosome 1:240,000,000Bp, 1probe≒50bp density
Multiple samples
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
...
4백 90만
NA 001
NA 002
NA 003
NA 004
NA 005
...
NA Nth
Cluster1 (NA002,NA009,...)
Cluster2 (NA001,NA006,...)
Cluster3 (NA003,NA004,...)
Cluster1 (NA010, NA012,...)
Cluster2 (NA007, NA009,...)
Cluster3 (NA001, NA002,...)
Cluster1 (NA001,NA004,...)
Cluster2 (NA002,NA005,...)
Cluster3 (NA003,NA009,...)
분할점
분할점
분할점

도면11


Size of chr1:240,000,000Bp
1probe≒50bp density
sample 40
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
...
4백 90만
NA 001
NA 002
NA 003
NA 004
NA 005
...
NA 040
CNVZ
Neutral

도면12


시작
1210
스무딩
1220
세그먼팅
1230
클러스터링
1240
candidate CNVZ 결정
1250
CNV 검출
1260
머징 & 프루닝
종료