(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2014-0137560
(43) 공개일자 2014년12월03일

(51) 국제특허분류(Int. Cl.)
C12Q 1/68 (2006.01) *G06F 19/00* (2011.01)
C12N 15/11 (2006.01)
(21) 출원번호 10-2013-0058152
(22) 출원일자 2013년05월23일
심사청구일자 2013년05월23일

(71) 출원인
한림대학교 산학협력단
강원도 춘천시 한림대학길 1, 한림대학교(옥천동)
(72) 발명자
이종근
경기도 고양시 일산동구 일산로 323 국립암센터 암유전체연구과
윤지희
서울 성동구 상원12길 15, 102동 208호 (성수동1가, 동아그린아파트)
(뒷면에 계속)
(74) 대리인
특허법인엠에이피에스

전체 청구항 수 : 총 10 항

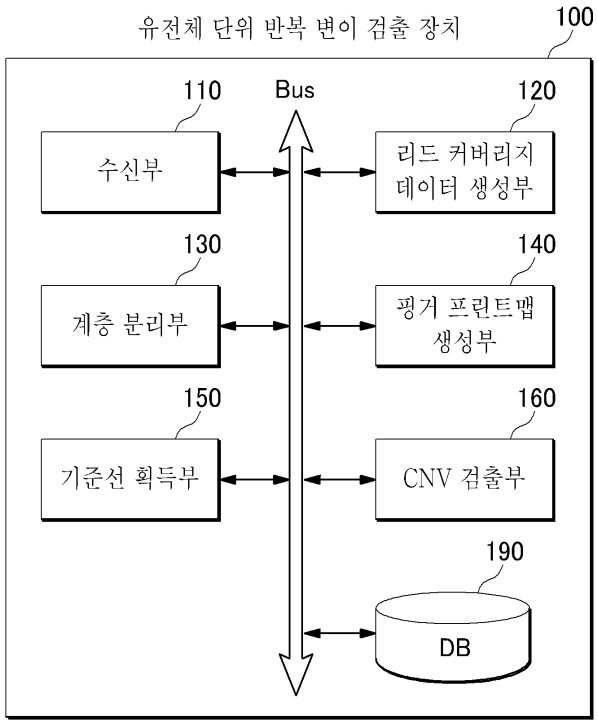
(54) 발명의 명칭 **유전체 단위 반복 변이를 검출하는 장치 및 방법**

(57) 요 약

유전체 단위 반복 변이(CNV: Copy Number Variation) 검출 장치는 복수의 테스트 시퀀스 리드를 수신하는 수신부, 상기 수신된 복수의 테스트 시퀀스 리드(read)를 레퍼런스 시퀀스에 매핑함으로써 리드 커버리지 데이터를 생성하는 리드 커버리지 데이터 생성부, 상기 생성된 리드 커버리지 데이터에 가우시안 컨볼루션을 적용시킴

(뒷면에 계속)

대 표 도 - 도1



으로써 상기 리드 커버리지 데이터를 복수의 스케일 스페이스 이미지(scale-space image) 계층으로 분리시키는 계층 분리부, 상기 분리된 복수의 계층별로 2차 도함수의 영교차점을 획득하고, 획득된 각각의 영교차점이 표시된 하나의 핑거 프린트맵(finger print map)을 생성하는 핑거 프린트맵 생성부, 상기 복수의 계층별로 기설정된 범위내에 포함된 상기 스케일 스페이스 이미지의 값들을 이용하여 상기 복수의 계층 각각의 기준선을 획득하는 기준선 획득부, 및 상기 획득된 각각의 기준선들로부터 CNV를 검출하기 위한 영역을 결정하고, 상기 결정된 영역에 기초하여 CNV를 검출하는 CNV 검출부를 포함한다.

(72) 발명자

이은주

강원 춘천시 퇴계로 168, 212동 706호 (퇴계동, 퇴계2주공아파트)

김백섭

강원 춘천시 동산면 원무동길 214,

특허청구의 범위

청구항 1

유전체 단위 반복 변이(CNV: Copy Number Variation) 검출 장치에 있어서,

복수의 테스트 시퀀스 리드를 수신하는 수신부;

상기 수신된 복수의 테스트 시퀀스 리드(read)를 레퍼런스 시퀀스에 매핑함으로써 리드 커버리지 데이터를 생성하는 리드 커버리지 데이터 생성부;

상기 생성된 리드 커버리지 데이터에 가우시안 컨볼루션을 적용시킴으로써 상기 리드 커버리지 데이터를 복수의 스케일 스페이스 이미지(scale-space image) 계층으로 분리시키는 계층 분리부;

상기 분리된 복수의 계층별로 2차 도함수의 영교차점을 획득하고, 획득된 각각의 영교차점이 표시된 하나의 핑거 프린트맵(finger print map)을 생성하는 핑거 프린트맵 생성부;

상기 복수의 계층별로 기설정된 범위내에 포함된 상기 스케일 스페이스 이미지의 값들을 이용하여 상기 복수의 계층 각각의 기준선을 획득하는 기준선 획득부; 및

상기 획득된 각각의 기준선들로부터 CNV를 검출하기 위한 영역을 결정하고, 상기 결정된 영역에 기초하여 CNV를 검출하는 CNV 검출부를 포함하는 유전체 단위 반복 변이 검출 장치.

청구항 2

제 1 항에 있어서,

상기 리드 커버리지 데이터 생성부는 상기 수신된 복수의 테스트 시퀀스 리드를 상기 레퍼런스 시퀀스에 정렬시킴으로써 상기 리드 커버리지 데이터를 생성하고, 상기 생성된 리드 커버리지 데이터에 미디안 필터를 적용시키는 것인, 유전체 단위 반복 변이 검출 장치.

청구항 3

제 1 항에 있어서,

상기 계층 분리부는 상기 검출하려는 CNV의 크기를 고려하여 상기 가우시안 컨볼루션을 적용시키는 것인, 유전체 단위 반복 변이 검출 장치.

청구항 4

제 1 항에 있어서,

상기 기설정된 범위는 상기 스케일 스페이스 이미지의 평균값 및 표준 편차값을 이용하여 결정된 것인, 유전체 단위 반복 변이 검출 장치.

청구항 5

제 1 항에 있어서,

상기 기준선은 상기 기설정된 범위내에 포함된 상기 스케일 스페이스 이미지의 값들에 대한 유효 평균값 및 유효 표준 편차값을 이용하여 획득된 것인, 것인, 유전체 단위 반복 변이 검출 장치.

청구항 6

제 5 항에 있어서,

상기 CNV를 검출하기 위한 영역은 상기 유효 평균값 및 유효 표준 편차값을 이용하여 결정되는 것인, 유전체 단위 반복 변이 검출 장치.

청구항 7

제 1 항에 있어서,

상기 CNV 검출부는 상기 결정된 영역에 대응하는 가장 낮은 계층의 영역 내에서 상기 CNV를 검출하는 것인, 유전체 단위 반복 변이 검출 장치.

청구항 8

제 1 항에 있어서,

상기 획득된 영교차점은 상기 2차 도함수에서의 전후 부호 변화에 따라서, 서로 다르게 하나의 핑거 프린트맵에 표시되어 이용되는 것인, 유전체 단위 반복 변이 검출 장치.

청구항 9

유전체 단위 반복 변이(CNV: Copy Number Variation) 검출 방법에 있어서,

복수의 테스트 시퀀스 리드를 수신하는 단계;

상기 수신된 복수의 테스트 시퀀스 리드(read)를 레퍼런스 시퀀스에 매핑함으로써 리드 커버리지 데이터를 생성하는 단계;

상기 생성된 리드 커버리지 데이터에 가우시안 컨볼루션을 적용시킴으로써 상기 리드 커버리지 데이터를 복수의 스케일 스페이스 이미지(scale-space image) 계층으로 분리시키는 단계;

상기 분리된 복수의 계층별로 2차 도함수의 영교차점을 획득하고, 획득된 각각의 영교차점이 표시된 하나의 핑거 프린트맵(finger print map)을 생성하는 단계;

상기 복수의 계층별로 기설정된 범위내에 포함된 상기 스케일 스페이스 이미지의 값들을 이용하여 상기 복수의 계층 각각의 기준선을 획득하는 단계; 및

상기 획득된 각각의 기준선들로부터 CNV를 검출하기 위한 영역을 결정하고, 상기 결정된 영역에 기초하여 CNV를 검출하는 단계를 포함하는 유전체 단위 반복 변이 검출 방법.

청구항 10

제 9 항에 있어서,

상기 CNV는 상기 결정된 영역에 대응하는 가장 낮은 계층의 영역 내에서 검출되는 것인, 유전체 단위 반복 변이 검출 방법.

명 세 서

기 술 분 야

[0001] 인간 유전체의 구조적 변이(SV: Structural Variation) 중 하나인 유전체 단위 반복 변이(CNV: Copy Number Variation)를 검출하는 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 인간 유전체의 구조적 변이(SV: Structural Variation) 중 하나인 유전체 단위 반복 변이(CNV: Copy Number Variation)는 1Kbps(Kilo base pairs) 이상의 크기를 가지는 염기 서열이 중복 (duplication) 혹은 결실(deletion)에 의해 그 반복 횟수의 차이가 개인 간에 존재하는 것으로서 차세대 염기 서열 분석법이 대중화되면서 질병 및 표현형질과 유관한 CNV의 발굴에 대한 연구의 중요성이 부각되고 있다.

[0003] 인간 유전체 서열에 존재하는 CNV 영역을 검출하기 위한 기존의 대표적인 방법으로서 마이크로어레이 기술을 이용한 방법과 서열 비교 방법이 있다. 마이크로어레이 기술을 이용한 방법은 oligonucleotide array 혹은 BAC array 등을 이용한 실험적 방법으로 주로 수백 kbp 이상의 큰 사이즈(large size)의 CNV 발견에 유용하다. 한편, 서열 비교 방법은 서로 다른 서열 데이터를 상호 비교하는 계산적 접근법(computational approach)을 사용하는 방법으로 마이크로어레이 기술을 이용하는 방법에 비하여 CNV 영역을 보다 정확하게 밝힐 수 있어, 더 작은 사이즈(small/ intermediate size)의 CNV 발견에 유용하다.

발명의 내용

해결하려는 과제

[0004] 인간 유전체의 구조적 변이(SV: Structural Variation) 중 하나인 유전체 단위 반복 변이(CNV: Copy Number Variation)를 검출하는 장치 및 방법을 제공하고자 한다. 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0005] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 발명의 제 1 측면에 따른 유전체 단위 반복 변이(CNV: Copy Number Variation) 검출 장치는 복수의 테스트 시퀀스 리드를 수신하는 수신부, 상기 수신된 복수의 테스트 시퀀스 리드(read)를 레퍼런스 시퀀스에 매핑함으로써 리드 커버리지 데이터를 생성하는 리드 커버리지 데이터 생성부, 상기 생성된 리드 커버리지 데이터에 가우시안 컨볼루션을 적용시킴으로써 상기 리드 커버리지 데이터를 복수의 스케일 스페이스 이미지(scale-space image) 계층으로 분리시키는 계층 분리부, 상기 분리된 복수의 계층별로 2차 도함수의 영교차점을 획득하고, 획득된 각각의 영교차점이 표시된 하나의 핑거 프린트맵(finger print map)을 생성하는 핑거 프린트맵 생성부, 상기 복수의 계층별로 기설정된 범위내에 포함된 상기 스케일 스페이스 이미지의 값들을 이용하여 상기 복수의 계층 각각의 기준선을 획득하는 기준선 획득부, 및 상기 획득된 각각의 기준선들로부터 CNV를 검출하기 위한 영역을 결정하고, 상기 결정된 영역에 기초하여 CNV를 검출하는 CNV 검출부를 포함하는 것을 특징으로 한다.

[0006] 한편, 본 발명의 제 2 측면에 따른 유전체 단위 반복 변이(CNV: Copy Number Variation) 검출 방법은 복수의 테스트 시퀀스 리드를 수신하는 단계, 상기 수신된 복수의 테스트 시퀀스 리드(read)를 레퍼런스 시퀀스에 매핑함으로써 리드 커버리지 데이터를 생성하는 단계, 상기 생성된 리드 커버리지 데이터에 가우시안 컨볼루션을 적용시킴으로써 상기 리드 커버리지 데이터를 복수의 스케일 스페이스 이미지(scale-space image) 계층으로 분리시키는 단계, 상기 분리된 복수의 계층별로 2차 도함수의 영교차점을 획득하고, 획득된 각각의 영교차점이 표시된 하나의 핑거 프린트맵(finger print map)을 생성하는 단계, 상기 복수의 계층별로 기설정된 범위내에 포함된 상기 스케일 스페이스 이미지의 값들을 이용하여 상기 복수의 계층 각각의 기준선을 획득하는 단계, 및 상기 획득된 각각의 기준선들로부터 CNV를 검출하기 위한 영역을 결정하고, 상기 결정된 영역에 기초하여 CNV를 검출하는 단계를 포함하는 것을 특징으로 한다.

발명의 효과

[0007] 전술한 본 발명의 과제 해결 수단 중 어느 하나에 의하면, 유전체 단위 반복 변이 검출 장치는 커버리지 배수가 상대적으로 낮은 경우에도 다양한 형태의 CNV 영역을 효율적으로 검출할 수 있다.

도면의 간단한 설명

[0008] 도 1은 본 발명의 일 실시예에 따른, 유전체 단위 반복 변이 검출 장치의 구성도이다.

도 2는 본 발명의 일 실시예에 따른, 미디안 필터의 적용 전후 각각에 대한 리드 커버리지 데이터값의 변화를 도시한 그래프이다.

도 3은 본 발명의 일 실시예에 따른, 리드 커버리지 데이터로부터 분리된 복수의 스케일 스페이스 이미지 계층을 도시한 그래프이다.

도 4는 본 발명의 일 실시예에 따른, 핑거 프린트맵과 복수의 계층별 스케일 스페이스 이미지의 2차 도함수를 도시한 그래프이다.

도 5는 본 발명의 일 실시예에 따른, 핑거 프린트맵과 복수의 계층별 스케일 스페이스 이미지, 그리고 복수의 계층별 스케일 스페이스 이미지의 2차 도함수를 도시한 그래프이다.

도 6은 본 발명의 일 실시예에 따른, 업스테이지(up stage)단과 다운스테이지(down stage)단 각각에서 CNV를 검출하기 위한 과정을 나타낸 블록도 이다.

도 7은 본 발명의 일 실시예에 따른, 유전체 단위 반복 변이 검출 장치가 CNV를 검출하는 방법을 나타낸 동작 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0009] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0010] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 연결"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0011] 이하, 첨부된 구성도 또는 처리 흐름도를 참고하여, 본 발명의 실시를 위한 구체적인 내용을 설명하도록 한다.

[0012] 도 1은 본 발명의 일 실시예에 따른, 유전체 단위 반복 변이 검출 장치의 구성도이다. 도 1을 참조하면, 유전체 단위 반복 변이 검출 장치(100)는 수신부(110), 리드 커버리지 데이터 생성부(120), 계층 분리부(130), 핑거 프린트맵 생성부(140), 기준선 획득부(150), CNV 검출부(160) 및 DB(190)를 포함할 수 있다. 다만, 도 1에 도시된 유전체 단위 반복 변이 검출 장치(100)는 본 발명의 하나의 구현 예에 불과하며, 도 1에 도시된 구성요소들을 기초로 하여 여러 가지 변형이 가능하다. 예를 들어, 유전체 단위 반복 변이 검출 장치(100)는 이용자로부터 어떤 명령 내지 정보를 입력받기 위한 유저 인터페이스가 더 포함될 수 있다. 이 경우, 유저 인터페이스는 일반적으로 키보드, 마우스 등과 같은 입력 장치가 될 수도 있으나, 영상 표시 장치에 표현되는 그래픽 유저 인터페이스(GUI, Graphical User interface)가 될 수도 있다.

[0013] 수신부(110)는 입력 데이터로서 복수의 테스트 시퀀스(sequence) 리드(read)를 수신할 수 있다. 또한, 수신부(110)는 복수의 테스트 시퀀스 리드와 함께 또는 시간을 달리하여 리드 커버리지 데이터 생성부(120)에서 이용될 레퍼런스 시퀀스(reference sequence)를 수신할 수 있다. 이 때, 복수의 테스트 시퀀스 리드 또는 레퍼런스 시퀀스는 직접 유전체 단위 반복 변이 검출 장치(100)를 통해 수신될 수도 있고, 네트워크를 통해 유전체 단위 반복 변이 검출 장치(100)의 외부로부터 수신될 수도 있다. 여기서, 수신부(110)에서 수신되는 복수의 테스트 시퀀스 리드는 차세대 시퀀싱(Next-Generation Sequencing) 기술에 테스트 시퀀스를 적용함으로써 생성된 것일 수 있다.

[0014] 이와 달리, 수신부(110)는 차세대 시퀀싱 기술에 적용되는 테스트 시퀀스 자체를 직접 수신할 수 있고, 수신된 테스트 시퀀스는 시퀀싱부(미도시)에서 차세대 시퀀싱 기술에 적용됨으로써 복수의 테스트 시퀀스 리드가 생성될 수 있고, 생성된 복수의 테스트 시퀀스 리드는 이하 설명할 리드 커버리지 데이터 생성부(120)에서 리드 커버리지 데이터를 생성하는데 이용될 수 있다.

[0015] 본 발명의 일 실시예에 따르면, 네트워크는 단말들 및 서버들과 같은 각각의 노드 상호 간에 정보 교환이 가능한 연결 구조를 의미하는 것으로, 이러한 네트워크(network)의 일 예에는 인터넷(Internet), Wireless LAN(Wireless Local Area Network), WAN(Wide Area Network), PAN(Personal Area Network), 3G, LTE(Long Term Evolution), WiFi(Wireless Fidelity), WiMAX(World Interoperability for Microwave Access), WiGig(Wireless Gigabit) 등이 포함되나 이에 한정되지는 않는다.

[0016] 리드 커버리지 데이터 생성부(120)는 수신부(110)에서 수신된 테스트 시퀀스에 포함된 리드(read)들을 레퍼런스

시퀀스에 매핑함으로써 매핑 정보를 획득하고, 획득된 매핑 정보를 이용하여 리드 커버리지 데이터를 생성할 수 있다. 구체적으로 리드 커버리지 데이터 생성부(120)는 수신부(110)에서 수신된 테스트 시퀀스의 리드를 상기 레퍼런스 시퀀스에 정렬시킴으로써 리드 커버리지 데이터를 생성할 수 있다.

[0017] 나아가, 리드 커버리지 데이터 생성부(120)는 앞서 생성된 리드 커버리지 데이터에 미디안 필터를 적용시킬 수 있다. 미디안 필터는 신호의 선명한 엣지(edge)를 희릿하게 만들지 않으면서 노이즈를 줄일 수 있는 효과적인 필터이다. 도 2는 본 발명의 일 실시예에 따른, 미디안 필터의 적용 전후 각각에 대한 리드 커버리지 데이터값의 변화를 도시한 그래프로서, 도 2를 참조하면, 미디안 필터를 적용하기 전, 후의 리드 커버리지 데이터의 차이를 확인할 수 있다. 도 2의 (a)는 미디안 필터를 적용하기 이전의 리드 커버리지 데이터이고, 도 2의 (b)는 미디안 필터를 적용시킨 후의 리드 커버리지 데이터이다. 도 2의 (a) 및 (b) 각각의 x축은 포지션 i를 나타내고, y축은 리드 커버리지 데이터값을 나타낸다. 미디안 필터링을 통하여 노이즈가 감소된 것을 확인할 수 있다.

[0018] 구체적으로, 리드 커버리지 데이터는 아래 수식 1을 참조하여 설명하겠다. c[i]는 게놈(genome) 포지션 i에 정렬된 리드 커버리지 데이터값들의 배열을 의미하는바, 예를 들면, c_1은 1번 포지션의 리드 커버리지 데이터값을 c_2는 2번 포지션의 리드 커버리지 데이터값을 나타낸다. 이 때, n은 시퀀스의 길이(최대 포지션)를 의미한다.

[0019] [수식 1]

[0020] $$c[i] = c_1 c_2 \ldots c_n,\ i\ (1 < i < n)$$

[0021] 나아가, 미디안 필터는 각각의 c_i를 $c_{i-w/2} \cdots c_i \cdots c_{i+w/2}$로 교체시킬 수 있고, 여기서 w는 미디안 필터의 슬라이딩 윈도우(sliding window)의 크기에 대응하는 값이다. 본 발명의 일실시는 w를 150으로 지정하나, w값은 이에 한정되지 않는다.

[0022] 계층 분리부(130)는 리드 커버리지 데이터 생성부(120)에서 생성된 리드 커버리지 데이터에 가우시안 컨볼루션을 적용시킴으로써 리드 커버리지 데이터를 복수의 스케일 스페이스 이미지(scale-space image) 계층으로 분리시킬 수 있다. 가우시안 컨볼루션을 적용하는 과정은 수식 2와 같이 나타낼 수 있다.

[0023] [수식 2]

[0024] $$c[i,k] = c[i] * g[j,\sigma_k] = \sum_{j=-m}^{m} c[i-j] \frac{1}{\sigma_k \sqrt{2\pi}} e^{-\frac{j^2}{2\sigma_k^2}}$$

[0025] 여기서 c[i, k]는 스케일 스페이스 이미지이고, $g[j,\sigma_k]$는 가우시안 커널이고, k($0 \le k \le L-1$)는 총 L개로 분리되는 스케일 스페이스 이미지 계층에 대한 어느 하나의 계층의 인덱스를 의미한다. m은 가우시안 커널의 윈도우(window) 크기이고, σ_k 는 스케일 파라미터로서 실험적으로 $m = 3\sigma_k$, $\sigma_k = 10^3 \times (1.1)^k$ 로 지정될 수 있다. $\sigma_k = 10^3 \times (1.1)^k$ 값은 검출하려는 CNV의 크기와 시간 복잡도(time complexity)를 고려하여 결정된 것일 수 있다.

[0026] 다만, 시간 영역에서의 컨볼루션은 주파수 영역에서의 곱과 같으므로, 계산의 복잡성을 줄이기 위하여 이산 푸리에 변환(Discrete Fourier Transform)을 적용함으로써 각각의 스케일 스페이스 이미지 계층 c[i, k]를 구할 수 있다.

[0027] 분리된 스케일 스페이스 이미지 계층은 도 3에 도시하였다. 도 3은 본 발명의 일 실시예에 따른, 리드 커버리지 데이터로부터 분리된 복수의 스케일 스페이스 이미지 계층을 도시한 그래프로서, k=0, 1, 40, 46, 52, 및 59일 때를 예시적으로 도시하였다. 다만, 각각의 그래프는 포지션 i에 대한 스케일 스페이스 이미지의 값을 의미하나, k값에 따른 복수의 그래프를 보기 좋게 하나의 도면으로 도시한 것일 뿐, k값이 커질수록 스케일 스페이스 이미지값이 커지는 것은 아니다. 오히려 수식 2를 참조하면, k값이 커질수록 σ_k 값이 커지므로, 스케일 스페이스 이미지값은 점차 smoothing 되어 작아짐을 알 수 있다.

[0028] 핑거 프린트맵 생성부(140)는 계층 분리부(130)에서 분리된 복수의 계층별로 2차 도함수의 영교차점을 획득할

수 있다. k 계층의 스케일 스페이스 이미지 c[i, k]의 2차 도함수 c"[i, k]는 수식 3과 같이 근사화시킬 수 있다. 또한, 2차 도함수의 영교차 신호 z[i, k]는 수식 4와 같이 정의할 수 있다.

[0029] [수식 3]

[0030] $$c''[i,k] \approx c[i+1,k] - 2c[i,k] + c[i-1,k]$$

[0031] [수식 4]

[0032] $$z[i,k] = \begin{cases} +1, & c''[i+1,k] > 0 \text{ and } c''[i-1,k] < 0 \\ -1, & c''[i+1,k] < 0 \text{ and } c''[i-1,k] > 0 \\ 0, & else \end{cases}$$

[0033] 즉, z[i, k]는 영교차점 i를 전후 기점으로, 2차 도함수의 값이 음(minus)에서 양(plus)으로 변하는 경우 1의 값을 갖고, 반대로 2차 도함수의 값이 양에서 음으로 변하는 경우 -1의 값을 갖게 된다. 물론, i가 영교차점이 아니라면 z[i, k]는 0의 값을 갖는다.

[0034] 핑거 프린트맵 생성부(140)는 복수의 계층별로 획득된 각각의 영교차점을 표시함으로써 하나의 핑거 프린트맵(finger print map)을 생성할 수 있다. 즉, z[i, k]가 +1 또는 -1의 값을 갖는 경우, 영교차점이 핑거 프린트맵에 표시되고, 핑거 프린트맵은 도 4에 도시하였다. 이 때, z[i, k]가 0의 값을 갖는 경우에는 핑거 프린트맵에 아무런 표시가 되지 않는다.

[0035] 도 4는 본 발명의 일 실시예에 따른, 핑거 프린트맵과 복수의 계층별 스케일 스페이스 이미지의 2차 도함수를 도시한 그래프이다. 도 4를 참조하면, 아랫쪽의 도면은 복수의 계층별 스케일 스페이스 이미지의 2차 도함수이고, 윗쪽의 도면은 복수의 계층별 스케일 스페이스 이미지의 2차 도함수 각각에서의 영교차점을 표시한 핑거 프린트맵이다. 핑거 프린트맵에서 +1의 z[i, k] 값을 갖는 영교차점은 k계층의 스케일 스페이스 이미지상에 ' o '로, -1의 z[i, k] 값을 갖는 영교차점은 k계층의 스케일 스페이스 이미지상에 ' + '로 표시되었다.

[0036] 좀 더 자세히 설명하면, k=40 일 때의 2차 도함수 c"[i, 40]는 401a, 401b를 포함하는 복수개의 영교차점을 갖는다. 이 때, 영교차점 401a에 대한 z[i(401a), 40] 값은 -1이고, 영교차점 401b에 대한 z[i(401b), 40] 값은 +1이다. 따라서, 영교차점 401a 및 401b에 대응하는 핑거 프린트맵 위의 점(dot)은 k=40인 스케일 스페이스 이미지 계층에 각각 ' + ' 및 ' o '으로 표시되었다. 마찬가지로, k=46 일 때의 2차 도함수 c"[i, 46]는 402a, 402b를 포함하는 복수개의 영교차점을 갖고, 영교차점 402a에 대응하는 z[i(402a), 46] 값은 +1이고, 영교차점 402b에 대응하는 z[i(402b), 46] 값은 -1 이므로 영교차점 402a 및 402b에 대응하는 핑거 프린트맵 위의 점(dot)은 k=46인 스케일 스페이스 이미지 계층에 각각 ' o ' 및 ' + '으로 표시되었다.

[0037] 기준선 획득부(150)는 복수의 계층별로 기설정된 범위내에 포함된 스케일 스페이스 이미지의 값들을 이용하여 복수의 계층 각각의 기준선을 획득할 수 있고, 기설정된 범위는 스케일 스페이스 이미지의 평균값 및 표준 편차값을 이용하여 결정된 것일 수 있다. 여기서, 스케일 스페이스 이미지의 평균값을 m(k)라고 하고, 스케일 스페이스 이미지의 표준 편차값을 δ(k)라고 한다면, 기설정된 범위는 m(k)±wδ(k)로 놓을 수 있다. δ(k)의 계수로 포함된 w는 가중치(weight)로서 임의의 값으로 설정할 수 있고, m(k) 및 δ(k)의 값은 각각 수식 5 및 수식 6으로부터 얻을 수 있다.

[0038] [수식 5]

[0039] $$m(k) = \sum_i c[i,k] / \sum_i$$

[0040] [수식 6]

[0041] $$\delta(k) = \sqrt{\sum_i (c[i,k] - m[k])^2 / \sum_i}$$

[0042] 본 발명의 일 실시예로서, 가중치 w는 2로 설정하겠다. 기준선 획득부(150)는 복수의 계층별로 $m(k)\pm2\delta(k)$을 벗어난 값들을 제외하고, $m(k)\pm2\delta(k)$에 포함된 값들에 대한 평균값과 표준 편차 값을 다시 구할 수 있다. 이 때의 평균값과 표준 편차값은 앞서 언급한 평균값 $m(k)$과 표준 편차값 $\delta(k)$과 구별하기 위하여 유효 평균값 $m^*(k)$과 유효 표준 편차값 $\delta^*(k)$이라 하겠다.

[0043] 기준선 획득부(150)는 기설정된 범위내에 포함된 스케일 스페이스 이미지의 값들에 대한 유효 평균값 및 유효 표준 편차값을 이용하여 복수의 계층 각각의 기준선 $m^*(k)$, $m^*(k)+d\delta^*(k)$, $m^*(k)-d\delta^*(k)$를 획득할 수 있다. 여기서, $\delta^*(k)$의 계수로서 포함된 d는 가중치로서, 경우에 따라 다양한 값이 될 수 있다.

[0044] 나아가, 기준선 획득부(150)는 복수의 계층으로 분리된 스케일 스페이스 이미지 계층 모두에 대하여 기준선을 획득할 수도 있지만, 복수의 계층 중에서, 적어도 2개 이상의 0이 아닌 영교차 신호 z[i, k] 값을 갖는 계층에 대해서만 기준선을 획득할 수도 있다. 다시 말해, 경우에 따라서, 기준선 획득부(150)는 영교차점이 2개 이상 있는 계층에 대해서만 기준선을 획득할 수도 있다.

[0045] CNV 검출부(160)는 기준선 획득부(150)에서 획득된 각각의 기준선들로부터 CNV를 검출하기 위한 영역을 결정할 수 있다. CNV를 검출하기 위한 영역은 유효 평균값 및 유효 표준 편차값을 이용하여 결정될 수 있고, 본 발명의 일 실시예에 따르면, CNV 검출부(160)는 영교차점을 적어도 2개 이상 갖는 스케일 스페이스 이미지 계층에 한하여, d값에 3을 적용한 기준선 $m^*(k)\pm3\delta^*(k)$을 이용하여 CNV를 검출하기 위한 영역을 결정할 수 있다. 이 때의 CNV를 검출하기 위한 영역은 계층별로 복수개가 될 수도 있고, 하나도 없을 수도 있다. CNV를 검출하기 위한 k계층의 m번째 영역은 $[l_{m,k}, u_{m,k}]$로 놓겠다. 여기서, $l_{m,k}$ 및 $u_{m,k}$는 핑거 프린트맵에 표시된 영교차 신호 z[i, k]의 i의 범위로서, $l_{m,k}\leq i\leq u_{m,k}$로 나타낼 수 있다. CNV를 검출하기 위한 영역을 결정하기 위하여 기준선의 d값에 3을 적용한 것은 본 발명의 일 실시예로서, d값은 3에 한정되지 않는다.

[0046] 단, CNV를 검출하기 위한 영역은 이하 3가지 조건에 기초하여 결정될 수 있다. 첫째로, k계층보다 높은, 즉, k+1 내지 L-1 계층에서 CNV를 검출하기 위한 영역으로 기지정된 영역에 대응하는 i는 $[l_{m,k}, u_{m,k}]$에 포함되지 않아야 한다. 둘째로, $z[l_{m,k}, k]$ 값과 $z[u_{m,k}, k]$ 값의 곱은 0보다 작아야 하고, $[l_{m,k}, u_{m,k}]$에 포함된 i에 대응하는 z[i, k] 값은 모두 0이어야 한다. 즉, $z[l_{m,k}, k]$ 값과 $z[u_{m,k}, k]$ 값 중 어느 하나는 +1이고, 나머지 하나는 -1 이고, 둘 사이의 값들은 모두 0이어야 한다. 같은 표현으로, 핑거 프린트맵에서 $i=l_{m,k}$ 일 때의 값과 $i=u_{m,k}$ 일 때의 값 중 어느 하나는 ' o '로 표시되고, 나머지 하나는 ' + '로 표시되고, 둘 사이에는 아무런 표시도 되지 않아야 한다. 마지막으로, $[l_{m,k}, u_{m,k}]$에 포함된 스케일 스페이스 이미지의 평균값이 위에서 언급한 주어진 기준선 $m^*(k)\pm3\delta^*(k)$를 벗어나야 한다. 여기서 $[l_{m,k}, u_{m,k}]$에 포함된 스케일 스페이스 이미지의 평균값은 수식 7과 같이 표현될 수 있다.

[0047] [수식 7]

[0048]
$$\sum_{i=l_{m,k}}^{u_{m,k}} c[i,k]/(u_{m,k}-l_{m,k}+1)$$

[0049] 이하에서, CNV를 검출하기 위한 영역을 결정하는 방법을 도 5를 참조하여 설명하겠다.

[0050] 도 5는 본 발명의 일 실시예에 따른, 핑거 프린트맵과 복수의 계층별 스케일 스페이스 이미지, 그리고 복수의 계층별 스케일 스페이스 이미지의 2차 도함수를 도시한 그래프이다. 아랫쪽의 도면은 복수의 계층별 스케일 스페이스 이미지 및 이들의 2차 도함수를 함께 도시한 도면이다. 구체적으로 실선으로 도시된 그래프는 스케일 스페이스 이미지값을 나타내고, 점선으로 도시된 그래프는 스케일 스페이스 이미지의 2차 도함수값을 나타낸다. 또한, 윗쪽의 도면은 복수의 계층별 스케일 스페이스 이미지의 2차 도함수 각각에서의 영교차점을 표시한 핑거 프린트맵이다.

[0051] k=52인 경우를 예를 들겠다. 세번째 조건을 만족하기 위해서, $[l_{m,52}, u_{m,52}]$에 포함된 스케일 스페이스 이미지값들 중 일부 또는 전부가 주어진 기준선 $m^*(52)\pm3\delta^*(52)$를 벗어나야 한다. 그래야만 평균값이 기준선 $m^*(52)\pm$

$3\delta^*(52)$를 벗어나서 세번째 조건을 만족할 수 있다. 따라서, k=52 일 때, 아래쪽 도면에서 기준선 $m^*(52)\pm3\delta^*(52)$를 벗어나는 스케일 스페이스 이미지값들을 찾아보면, 501a 와 501b 사이의 값들이 여기에 해당됨을 알 수 있다. 다만, 세번째 조건은 영역내의 스케일 스페이스 이미지의 평균값을 이용하기 때문에, 501a 와 501b 사이의 값들만 세번째 조건을 만족하는 것은 아니고, 501a 이전 부분과 501b 이후 부분 역시 세번재 조건을 만족할 수 있다. 다음, 두번째 조건을 살펴보면, 앞서 언급한바와 같이 $z[l_{m,52}, 52]$ 값과 $z[u_{m,52}, 52]$ 값 중 어느 하나는 +1이고, 나머지 하나는 -1 이고, 둘 사이의 값들은 모두 0이어야 한다. 이 때, c"[i, k] 값이 0 일 때, z[i, k] 값이 +1 또는 -1 이라는 것을 의미하므로, 두번째 조건을 만족하기 위해서는 $c"[l_{m,52}, 52]$ 값과 $c"[u_{m,52}, 52]$ 값은 0이고, 핑거 프린트맵에서 하나는 ' o ' , 다른 하나는 ' + ' 임을 알 수 있다. 따라서, 위쪽의 도면에서 k=52 인 계층을 살펴보았을 때, 501c와 501d 사이의 영역이 두번째 조건을 만족함을 알 수 있다. 첫번째 조건에 대해서는, k가 53 내지 59 일 때 두번째 조건 또는 세번째 조건을 만족하지 못하는 것으로 가정하겠다. 따라서, k=52일 때, 501c와 501d 사이의 영역 $[501c_{,52}, 501d_{,52}]$이 CNV를 검출하기 위한 영역으로 결정될 수 있다.

[0052] 이어서 k=46인 경우를 예를 들겠다. k=52인 경우와 마찬가지로, 세번째 조건을 만족하는 영역을 찾기 위해, 아래쪽 도면에서 기준선 $m^*(46)\pm3\delta^*(46)$를 벗어나는 스케일 스페이스 이미지값들을 찾아보면, 502a 와 502b 사이의 값들, 503a 와 503b 사이의 값들이 여기에 해당됨을 알 수 있다. 우선, 502a 와 502b 사이의 영역에 대해서, 첫번째 조건을 살펴보면, k=46 보다 높은 계층인 k=52 계층에서 이미 $[501c_{,52}, 501d_{,52}]$이 CNV 검출하기 위한 영역으로 결정되었기 때문에, 502a 와 502b 사이의 영역은 첫번째 조건을 만족하지 못하여 CNV 검출하기 위한 영역이 될 수 없다. 다음으로, 503a 와 503b 사이의 영역은 첫번째 조건을 만족한다는 전제하에, 위쪽 도면을 살펴보면 k=46일 때의, 503c와 503d 사이의 영역이 두번째 조건을 만족함을 알 수 있다. 따라서, k=46일 때, 503c와 503d 사이의 영역 $[503c_{,46}, 503d_{,46}]$이 CNV를 검출하기 위한 영역으로 결정될 수 있다.

[0053] 이러한 방식으로 k=40, 46, 52일 때 결정된 CNV 검출하기 위한 영역을 위쪽 도면에 두껍게 표시하였다. (각각의 k=40, 46, 52 일 때의 아래쪽 도면과 점선으로 연결된 부분) 이러한 CNV를 검출하기 위한 영역은, 통계학적으로 중요한 변화가 발생하는 영역으로서, CNV의 이득 또는 손실이 발생될 수 있다.

[0054] 나아가, CNV 검출부(160)는 결정된 영역에 기초하여 CNV를 검출할 수 있다. CNV 검출부(160)는 결정된 영역에 대응하는 가장 낮은 계층(k=0)의 영역 내에서 CNV를 검출할 수 있다. 도 5를 참조하면, k=52일 때, CNV를 검출하기 위한 영역은 $[501c_{,52}, 501d_{,52}]$이고, 이에 대응하는 k=0일 때의 영역은 $[l'_{m,52}, u'_{m,52}]$이다. 즉, CNV 검출부(160)는 k=52 이 때, CNV를 검출하기 위해 $[501c_{,52}, 501d_{,52}]$에 대응하는 $[l'_{m,52}, u'_{m,52}]$ 영역을 탐색할 수 있다.

[0055] DB(190)는 데이터를 저장한다. 이 때, 데이터는 유전체 단위 반복 변이 검출 장치(100) 내부의 각 구성요소들 간에 입력 및 출력되는 데이터를 포함하고, 유전체 단위 반복 변이 검출 장치(100)와 유전체 단위 반복 변이 검출 장치(100) 외부의 구성요소들간에 입력 및 출력되는 데이터를 포함한다. 예를 들어, DB(190)는 리드 커버리지 데이터 생성부(120)에서 생성된 리드 커버리지 데이터를 저장할 수 있고, 계층 분리부(130)에서 생성된 복수의 계층별 스케일 스페이스 이미지값들을 저장할 수 있다. 이러한 DB(190)의 일 예에는 유전체 단위 반복 변이 검출 장치(100) 내부 또는 외부에 존재하는 하드디스크드라이브, ROM(Read Only Memory), RAM(Random Access Memory), 플래쉬메모리 및 메모리카드 등이 포함된다.

[0056] 당업자라면, 수신부(110), 리드 커버리지 데이터 생성부(120), 계층 분리부(130), 핑거 프린트맵 생성부(140), 기준선 획득부(150), CNV 검출부(160) 및 DB(190) 각각이 분리되어 구현되거나, 이 중 하나 이상이 통합되어 구현될 수 있음을 충분히 이해할 것이다.

[0057] 도 6은 본 발명의 일 실시예에 따른, 업스테이지(up stage)단과 다운스테이지(down stage)단 각각에서 CNV를 검출하기 위한 과정을 나타낸 블록도 이다.

[0058] 도 6을 참조하면, 업스테이지단(왼쪽 도면)은 입력 데이터로서 입력받은 리드들 $\{R_j\}$을 매핑하고 미디안 필터를 적용시켜 리드 커버리지 데이터 c[i]를 생성하고, 가우시안 컨볼루션을 통해 k계층의 스케일 스페이스 이미지값 c[i, k]를 생성하고, 생성된 스케일 스페이스 이미지를 이용하여 (2차 도함수) 핑거 프린트맵에 표시되는 z[i, k] 값을 획득할 수 있다. 다운스테이지단(오른쪽 도면)은 핑거 프린트맵에 표시되는 z[i, k] 값과 스케일 스페

이스 이미지값 c[i, k]을 이용하여 기준선 $m^*(k)$, $m^*(k)+d\delta^*(k)$, $m^*(k)-d\delta^*(k)$를 획득하고, 핑거 프린트맵에 표시되는 z[i, k] 값과 스케일 스페이스 이미지값 c[i, k], 그리고 기준선을 이용하여 CNV를 검출하기 위한 영역 $[l_{m,k}, u_{m,k}]$을 획득하고, 획득된 $[l_{m,k}, u_{m,k}]$ 영역을 이용하여 CNV를 검출할 수 있다.

[0059] 도 7은 본 발명의 일 실시예에 따른, 유전체 단위 반복 변이 검출 장치가 CNV를 검출하는 방법을 나타낸 동작 흐름도이다. 도 7에 도시된 실시예에 따른 CNV를 검출하는 방법은 도 1 내지 도 6에 도시된 실시예에 따른 유전체 단위 반복 변이 검출 장치(100)에서 시계열적으로 처리되는 단계들을 포함한다. 따라서, 이하 생략된 내용이라고 하더라도 도 1 내지 도 6의 유전체 단위 반복 변이 검출 장치(100)에 관하여 기술된 내용은 도 7에 도시된 실시예에 따른 CNV를 검출하는 방법에도 적용된다.

[0060] 단계 S701에서 유전체 단위 반복 변이 검출 장치(100)는 복수의 테스트 시퀀스 리드를 수신할 수 있다.

[0061] 단계 S702에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S701에서 수신된 복수의 테스트 시퀀스 리드를 레퍼런스 시퀀스에 매핑함으로써 리드 커버리지 데이터를 생성할 수 있다.

[0062] 단계 S703에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S702에서 생성된 리드 커버리지 데이터에 가우시안 컨볼루션을 적용시킴으로써 리드 커버리지 데이터를 복수의 스케일 스페이스 이미지 계층으로 분리할 수 있다.

[0063] 단계 S704에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S703에서 분리된 복수의 계층별로 2차 도함수의 영교차점을 획득하고, 획득된 각각의 영교차점이 표시된 하나의 핑거 프린트맵을 생성할 수 있다.

[0064] 단계 S705에서 유전체 단위 반복 변이 검출 장치(100)는 복수의 계층별로 기설정된 범위내에 포함된 스케일 스페이스 이미지의 값들을 이용하여 복수의 계층 각각의 기준선을 획득할 수 있다.

[0065] 단계 S706에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S705에서 획득된 각각의 기준선들로부터 CNV를 검출하기 위한 영역을 결정하고, 결정된 영역에 기초하여 CNV를 검출할 수 있다.

[0066] 본 발명의 일 실시예는 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행 가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체 및 통신 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다. 통신 매체는 전형적으로 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈, 또는 반송파와 같은 변조된 데이터 신호의 기타 데이터, 또는 기타 전송 메커니즘을 포함하며, 임의의 정보 전달 매체를 포함한다.

[0067] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0068] 본 발명의 범위는 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0069] 100: 유전체 단위 반복 변이 검출 장치

110: 수신부

120: 리드 커버리지 데이터 생성부

130: 계층 분리부

140: 핑거 프린트맵 생성부

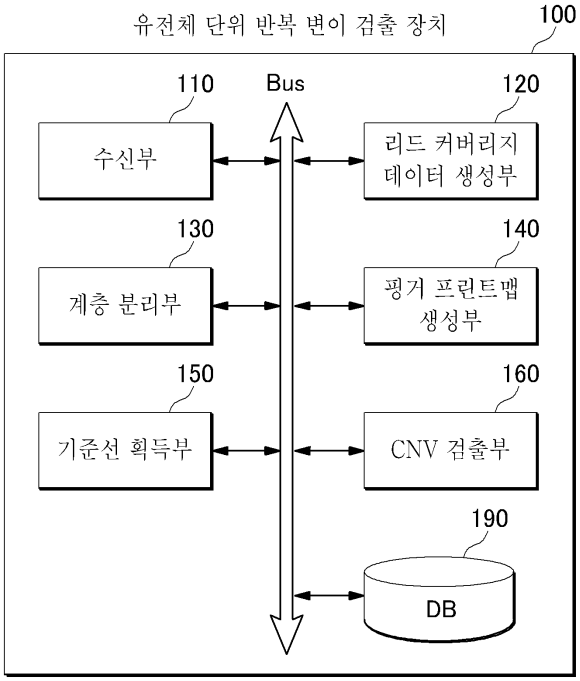
150: 기준선 획득부

160: CNV 검출부

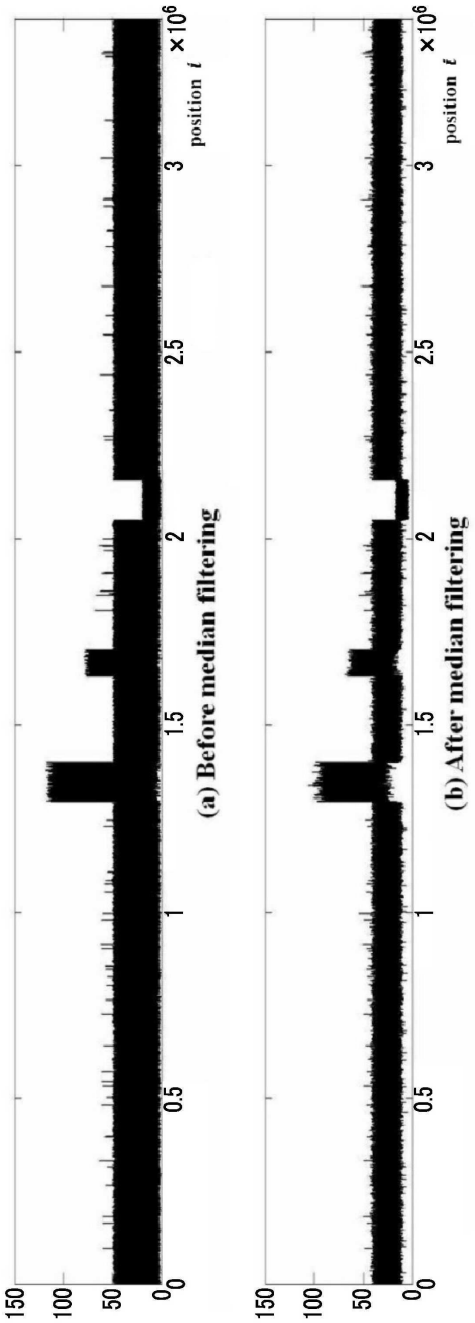
190: DB

도면

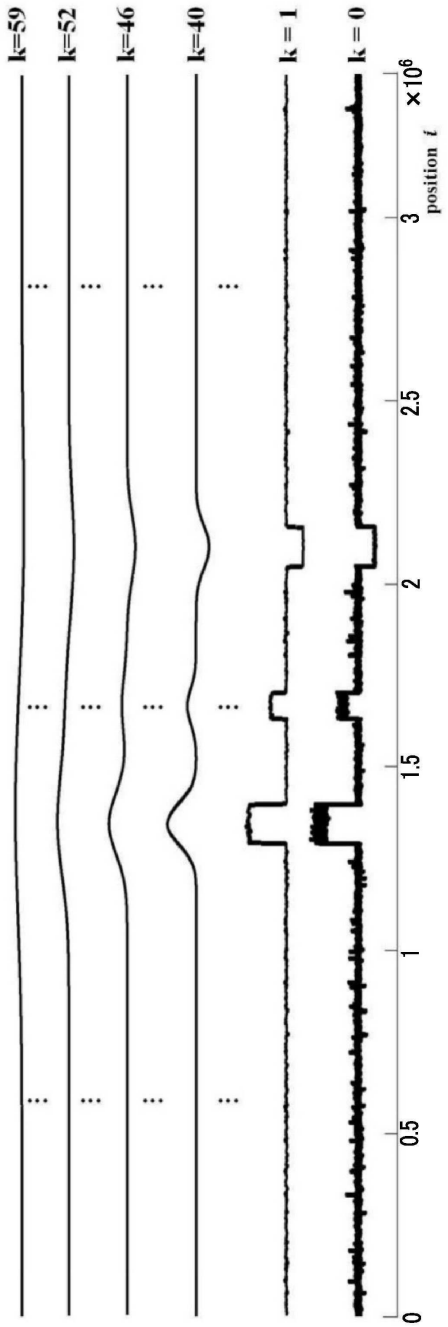
도면1



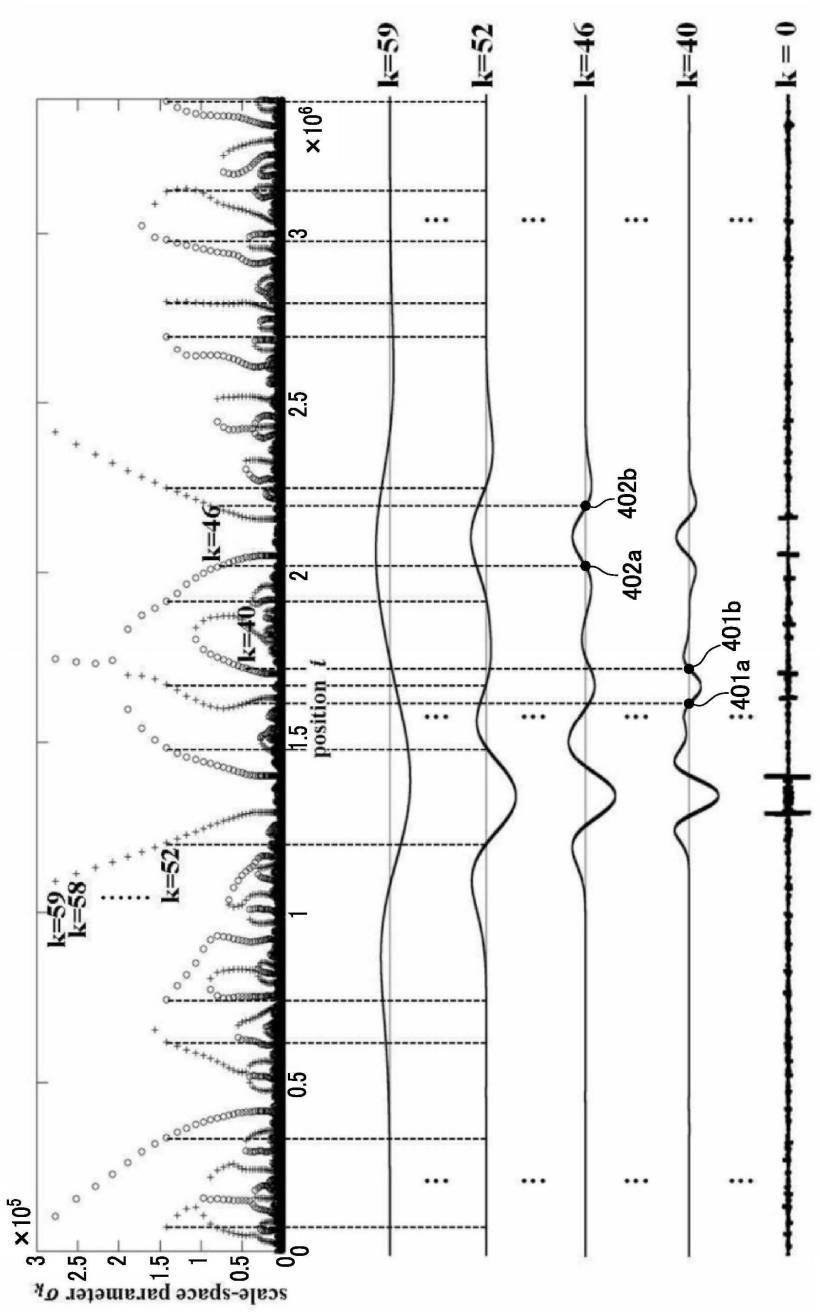
도면2


150
100
50
0
0
0.5
1
1.5
2
2.5
3
×10^6
position i
(a) Before median filtering
150
100
50
0
0
0.5
1
1.5
2
2.5
3
×10^6
position i
(b) After median filtering

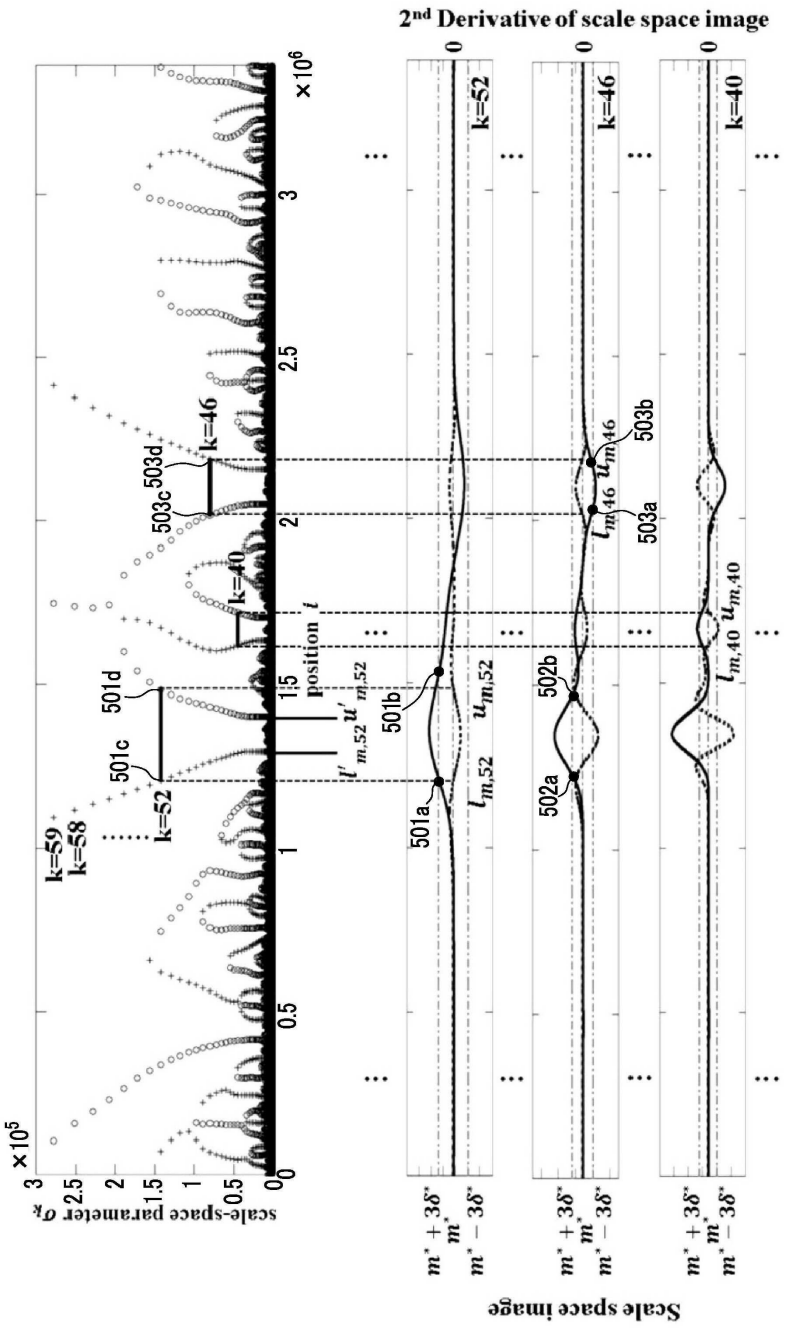
도면3



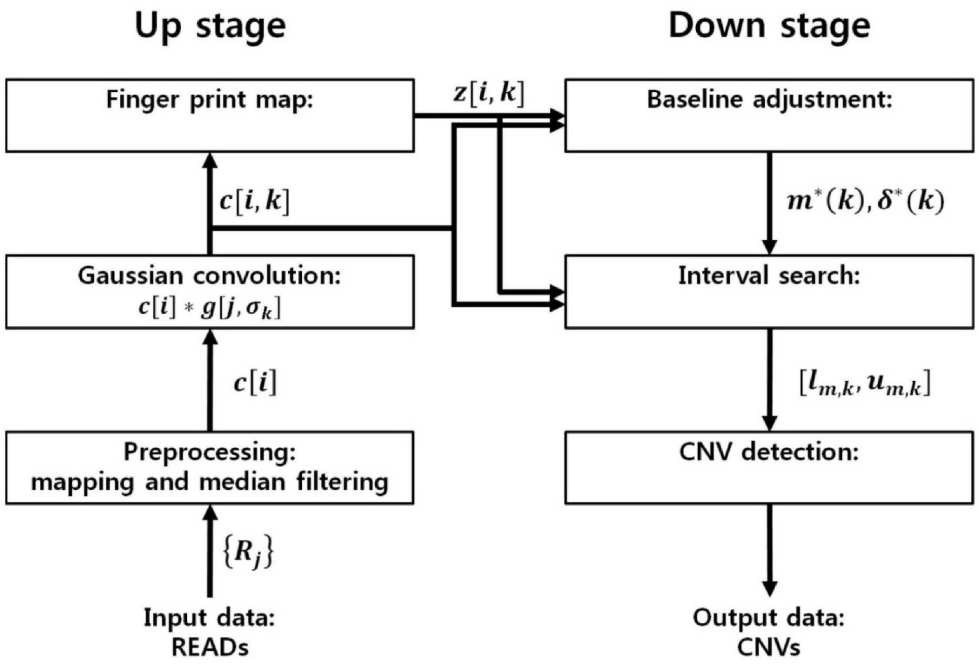
도면4


k=59
k=52
k=46
k=40
k = 0
402b
402a
401b
401a
position t
scale-space parameter σk
×10^5
×10^6
k=59
k=58
k=52
k=46
k=40

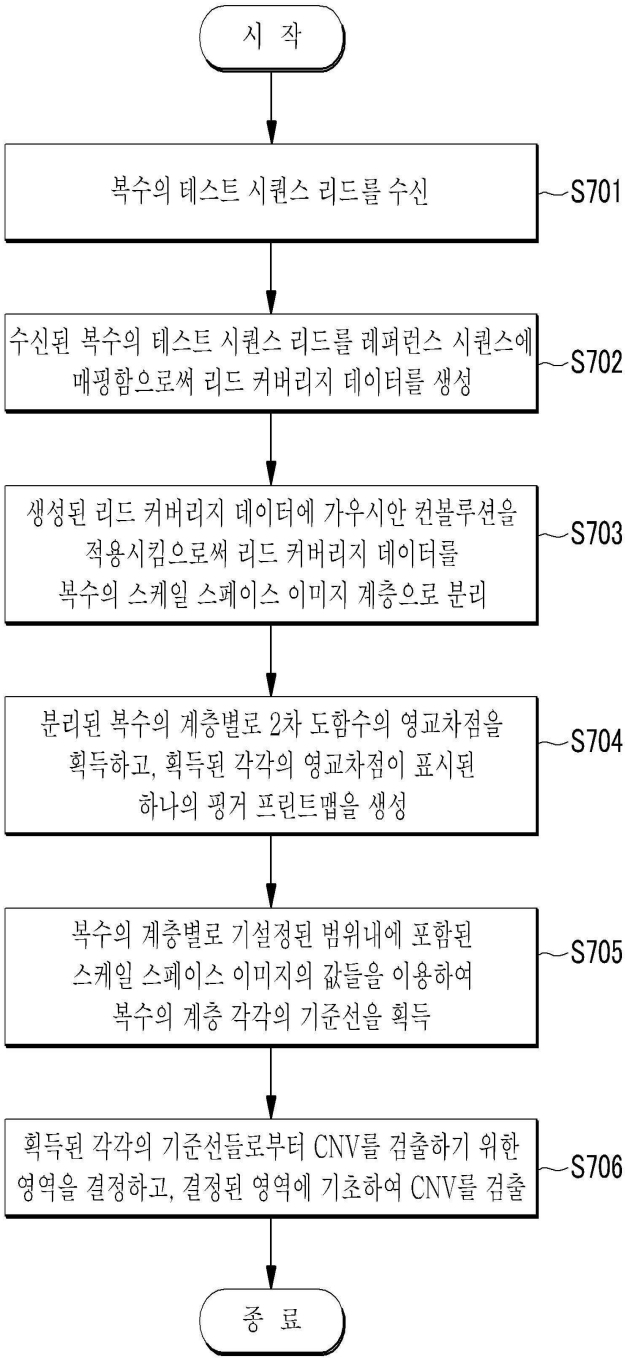
도면5


2nd Derivative of scale space image
scale-space parameter σk
position t
Scale space image
k=59
k=58
k=52
k=46
k=40
501a
501b
501c
501d
502a
502b
503a
503b
503c
503d

도면6


Up stage
Down stage
Finger print map:
$z[i,k]$
Baseline adjustment:
$c[i,k]$
$m^*(k), \delta^*(k)$
Gaussian convolution:
$c[i] * g[j,\sigma_k]$
Interval search:
$c[i]$
$[l_{m,k}, u_{m,k}]$
Preprocessing:
mapping and median filtering
CNV detection:
$\{R_j\}$
Input data:
READs
Output data:
CNVs

도면7


시 작
복수의 테스트 시퀀스 리드를 수신
S701
수신된 복수의 테스트 시퀀스 리드를 레퍼런스 시퀀스에 매핑함으로써 리드 커버리지 데이터를 생성
S702
생성된 리드 커버리지 데이터에 가우시안 컨볼루션을 적용시킴으로써 리드 커버리지 데이터를 복수의 스케일 스페이스 이미지 계층으로 분리
S703
분리된 복수의 계층별로 2차 도함수의 영교차점을 획득하고, 획득된 각각의 영교차점이 표시된 하나의 핑거 프린트맵을 생성
S704
복수의 계층별로 기설정된 범위내에 포함된 스케일 스페이스 이미지의 값들을 이용하여 복수의 계층 각각의 기준선을 획득
S705
획득된 각각의 기준선들로부터 CNV를 검출하기 위한 영역을 결정하고, 결정된 영역에 기초하여 CNV를 검출
S706
종 료