



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2015-0084571

(43) 공개일자 2015년07월22일

(51) 국제특허분류(Int. Cl.)

G06F 19/10 (2011.01) C12Q 1/68 (2006.01)

(21) 출원번호 10-2014-0004706

(22) 출원일자 2014년01월14일

심사청구일자 2014년01월14일

(71) 출원인

한림대학교 산학협력단

강원도 춘천시 한림대학길 1, 한림대학교(옥천동)

(72) 발명자

홍상균

경기도 양평군 강상면 지례1길 36

윤지희

서울특별시 성동구 상원12길 15, 102동 208호(성수동1가, 성수동동아그린아파트)

(뒷면에 계속)

(74) 대리인

특허법인엠에이피에스

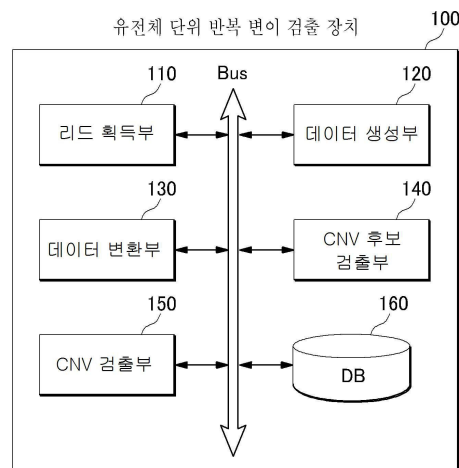
전체 청구항 수 : 총 10 항

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

(57) 요약

유전체 단위 반복 변이(CNV: Copy Number Variation) 검출 장치는 테스트 시퀀스에 대응하는 복수의 리드를 획득하는 리드 획득부; 상기 획득된 복수의 리드(read)를 레퍼런스 시퀀스에 정렬시킴으로써 리드 커버리지 데이터를 생성하는 데이터 생성부; 상기 생성된 리드 커버리지 데이터에 제 1 변환 및 제 2 변환을 각각 적용하여 제 1 변환 데이터 및 제 2 변환 데이터를 생성하는 데이터 변환부; 상기 생성된 제 1 변환 데이터 및 제 2 변환 데이터를 이용하여 적어도 하나 이상의 CNV 후보를 검출하는 CNV 후보 검출부; 및 상기 적어도 하나 이상의 CNV 후보 중 적어도 하나 이상을 CNV로 검출하는 CNV 검출부를 포함할 수 있다.

대표도 - 도1



(72) 발명자

홍동완

경기도 고양시 일산동구 일산로 323 (마두동) 국립
암센터 암유전체연구과

이은주

강원도 춘천시 퇴계로 168, 212동 706호 (퇴계동,
퇴계(2)주공아파트)

김백섭

강원도 춘천시 동산면 원무동길 214

박상현

서울특별시 강남구 개포로 310, 120동 503호 (개포
동, 개포주공1차아파트)

명세서

청구범위

청구항 1

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

테스트 시퀀스에 대응하는 복수의 리드를 획득하는 리드 획득부;

상기 획득된 복수의 리드(read)를 레퍼런스 시퀀스에 정렬시킴으로써 리드 커버리지 데이터를 생성하는 데이터 생성부;

상기 생성된 리드 커버리지 데이터에 제 1 변환 및 제 2 변환을 각각 적용하여 제 1 변환 데이터 및 제 2 변환 데이터를 생성하는 데이터 변환부;

상기 생성된 제 1 변환 데이터 및 제 2 변환 데이터를 이용하여 적어도 하나 이상의 CNV 후보를 검출하는 CNV 후보 검출부; 및

상기 적어도 하나 이상의 CNV 후보 중 적어도 하나 이상을 CNV로 검출하는 CNV 검출부를 포함하는 검출 장치.

청구항 2

제 1 항에 있어서,

상기 제 1 변환은 상기 리드 커버리지 데이터의 평균값을 이용하고, 상기 제 2 변환은 상기 리드 커버리지 데이터의 평균 기울기값을 이용하는 것인, 검출 장치.

청구항 3

제 1 항에 있어서,

상기 CNV 후보 검출부는 상기 검출된 적어도 하나 이상의 CNV 후보 중 연이은 CNV 후보를 병합하여 하나의 CNV 후보로 결정하는 것인, 검출 장치.

청구항 4

제 1 항에 있어서,

상기 CNV 후보 검출부는 상기 검출된 적어도 하나 이상의 CNV 후보 중 연이은 양의 값을 갖는 CNV 후보 또는 연이은 음의 값을 갖는 CNV 후보를 병합하여 하나의 CNV 후보로 결정하는 것인, 검출 장치.

청구항 5

제 1 항에 있어서,

상기 CNV는 기설정된 길이(kbp) 이상이고, 상기 테스트 시퀀스 또는 상기 레퍼런스 시퀀스에 적어도 둘 이상 포함되는 것인, 검출 장치.

청구항 6

제 1 항에 있어서,

상기 데이터 생성부는 상기 복수의 리드가 상기 레퍼런스 시퀀스에 완전 매칭되는 경우 및 기설정된 범위 내의 교체(replacing) 또는 삽입(inserting)을 허용하여 매칭되는 경우를 상기 복수의 리드가 상기 레퍼런스 시퀀스에 정렬되는 것으로 결정하는 것인, 검출 장치.

청구항 7

제 1 항에 있어서,

상기 데이터 변환부는 중심 극한 정리(central limit theorem) 및 통계적 추론(statistical estimation)에 기초하여 결정된 문턱값 범위(threshold boundary)을 이용하여 상기 제 1 변환 데이터 및 상기 제 2 변환 데이터를 생성하는 것인, 검출 장치.

청구항 8

제 1 항에 있어서,

상기 데이터 변환부는 상기 리드 커버리지 데이터를 샘플링하는 윈도우의 크기 및 이동값을 결정하여 시퀀싱 에러 및/또는 충분치 않은 커버리지에 의한 효과를 감쇄시키는 것인, 검출 장치.

청구항 9

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

테스트 시퀀스에 대응하는 복수의 리드를 획득하는 단계;

상기 획득된 복수의 리드(read)를 레퍼런스 시퀀스에 정렬시킴으로써 리드 커버리지 데이터를 생성하는 단계;

상기 생성된 리드 커버리지 데이터에 제 1 변환 및 제 2 변환을 각각 적용하여 제 1 변환 데이터 및 제 2 변환 데이터를 생성하는 단계;

상기 생성된 제 1 변환 데이터 및 제 2 변환 데이터를 이용하여 적어도 하나 이상의 CNV 후보를 검출하는 단계; 및

상기 적어도 하나 이상의 CNV 후보 중 적어도 하나 이상을 CNV로 검출하는 단계를 포함하는 검출 방법.

청구항 10

제 9 항에 있어서,

상기 검출된 적어도 하나 이상의 CNV 후보 중 연이은 CNV 후보를 병합하여 하나의 CNV 후보로 결정하는 단계를 더 포함하는 검출 방법.

발명의 설명

기술 분야

[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.

배경 기술

[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 테스트 시퀀스에 대응하는 복수의 리드를 획득하는 리드 획득부; 상기 획득된 복수의 리드(read)를 레퍼런스 시퀀스에 정렬시킴으로써 리드 커버리지 데이터를 생성하는 데이터 생성부; 상기 생성된 리드 커버리지 데이터에 제 1 변환 및 제 2 변환을 각각 적용하여 제 1 변환 데이터 및 제 2 변환 데이터를 생성하는 데이터 변환부; 상기 생성된 제 1 변환 데이터 및 제 2 변환 데이터를 이용하여 적어도 하나 이상의 CNV 후보를 검출하는 CNV 후보 검출부; 및 상기 적어도 하나 이상의 CNV 후보 중 적어도 하나 이상을 CNV로 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 테스트 시퀀스에 대응하는 복수의 리드를 획득하는 단계; 상기 획득된 복수의 리드(read)를 레퍼런스 시퀀스에 정렬시킴으로써 리드 커버리지 데이터를 생성하는 단계; 상기 생성된 리드 커버리지 데이터에 제 1 변환 및 제 2 변환을 각각 적용하여 제 1 변환 데이터 및 제 2 변환 데이터를 생성하는 단계; 상기 생성된 제 1 변환 데이터 및 제 2 변환 데이터를 이용하여 적어도 하나 이상의 CNV 후보를 검출하는 단계; 및 상기 적어도 하나 이상의 CNV 후보 중 적어도 하나 이상을 CNV로 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 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, 리드 커버리지 데이터를 생성하는 방법을 나타낸 도면이다.

도 3은 본 발명의 일 실시예에 따른, 이상적인 CNV의 형태를 나타낸 리드 커버리지 데이터의 그래프이다.

도 4는 본 발명의 일 실시예에 따른, CNV를 검출하기 위한 CNV의 조건을 나타낸 도면이다.

도 5는 본 발명의 일 실시예에 따른, 유전체 단위 반복 변이 검출 장치가 CNV를 검출하는 방법을 나타낸 동작 흐름도이다.

도 6은 본 발명의 다른 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(CNV: Copy Number Variation) 검출 장치의 구성도이다. 도 1을 참조하면, 유전체 단위 반복 변이 검출 장치(100)는 리드 획득부(110), 데이터 생성부(120), 데이터 변환부(130), CNV 후보 검출부(140), CNV 검출부(150) 및 DB(160)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 이용자로 부터 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] 리드 획득부(110)는 입력 데이터로서 테스트 시퀀스(sequence)에 대응하는 복수의 리드(read)를 획득할 수 있다. 또한, 리드 획득부(110)는 복수의 리드와 함께 또는 시간을 달리하여 데이터 생성부(120)에서 이용될 레퍼런스 시퀀스(reference sequence)를 획득할 수 있다. 이 때, 복수의 리드 또는 레퍼런스 시퀀스는 직접 유전체 단위 반복 변이 검출 장치(100)에 연결된 저장 매체를 통해 획득될 수도 있고, 네트워크를 통해 유전체 단위 반복 변이 검출 장치(100)의 외부로부터 수신하여 획득될 수도 있다. 여기서, 리드 획득부(110)에서 획득되는 복수의 리드는 차세대 시퀀싱(Next-Generation Sequencing) 기술에 테스트 시퀀스를 적용함으로써 생성된 것일 수 있다.

[0014] 이와 달리, 리드 획득부(110)는 차세대 시퀀싱 기술에 적용되는 테스트 시퀀스 자체를 직접 획득할 수 있고, 획득된 테스트 시퀀스는 시퀀싱부(미도시)에서 차세대 시퀀싱 기술에 적용됨으로써 복수의 리드가 생성될 수 있고, 생성된 복수의 테스트 시퀀스 리드는 이하 설명할 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] 데이터 생성부(120)는 리드 획득부(110)에서 획득된 리드(read)를 레퍼런스 시퀀스에 매핑함으로써 매핑 정보를 획득하고, 획득된 매핑 정보를 이용하여 리드 커버리지 데이터를 생성할 수 있다. 즉, 데이터 생성부(120)는 리드 획득부(110)에서 획득된 리드를 상기 레퍼런스 시퀀스에 정렬시킴으로써 리드 커버리지 데이터를 생성할 수 있다.

[0017] 나아가, 데이터 생성부(120)는 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.

[0018] 데이터 생성부(120)에서 리드 커버리지 데이터를 생성하는 방법을 이하 도 2를 참조하여 구체적으로 설명하겠다.

[0019] 도 2는 본 발명의 일 실시예에 따른, 리드 커버리지 데이터를 생성하는 방법을 나타낸 도면이다. 도 2의 상단은 레퍼런스 시퀀스 S에 복수의 리드 Rj를 정렬시킨 모습을 나타낸 도면이다.

[0020] 레퍼런스 시퀀스 S의 서브시퀀스 Sp에 대하여 리드 Rj가 매칭될 때, 서브시퀀스 Sp에 리드 Rj가 정렬되었다고 볼 수 있다. 다만, 레퍼런스 시퀀스 또는 테스트 시퀀스에는 서로 다른 유전적 변이(예를 들어, indel, CNV, SNP 등)가 포함될 수 있기 때문에, 서브시퀀스 Sp에 포함된 일부 유전자의 교체(replacing) 또는 삽입(inserting)이 이루어진 Sp'가 리드 Rj와 매칭되는 경우까지 서브시퀀스 Sp에 리드 Rj가 정렬된 것으로 볼 수

있다. 이는 아래 수식1과 같이 나타낼 수 있다.

[수식 1]

$$c_i = \sum_{j=1}^N SC_j, \text{ where } SC_j = \begin{cases} 1 & , R_j \cong S_p (p \leq i \leq p+l-1) \\ 0 & , \text{ otherwise} \end{cases}$$

여기서 c_i 는 길이 n 의 리드 커버리지 데이터 $C=c_0 \cdots c_{n-1}$ 의 i 번째 값을 나타낸다. $R_j=r_0^j \cdots r_{d-1}^j$ 는 길이 d 의 리드이고, $1 \leq j \leq N$ 이고, N 은 복수의 리드의 개수이다. $S_p=s_p \cdots s_{p+l-1}$ 는 길이 l 의 서브시퀀스이고, $0 \leq p, l, p+l \leq n-1$ 이고, n 은 리드 커버리지 데이터 C 의 길이이다.

도 2의 하단은 레퍼런스 시퀀스 $S(201)$ 의 포지션 i 별로 정렬된 복수의 리드 R_j 의 개수를 카운팅하여 생성된 리드 커버리지 데이터를 나타낸 도면이다. 그래프의 x 축은 레퍼런스 시퀀스의 포지션 i 를 나타내고, 그래프의 y 축은 리드 커버리지 데이터의 값을 나타낸다.

도 3은 본 발명의 일 실시예에 따른, 이상적인 CNV의 형태를 나타낸 리드 커버리지 데이터의 그래프이다. 도 3(a)는 카피 삽입(CNV gain)이 발생한 경우이고, 도 3(b)는 카피 손실(CNV loss)이 발생한 경우를 나타낸 그래프이다. 리드 커버리지 데이터의 평균값을 기준으로 상대적으로 높은 값을 갖는 영역이 CNV gain이고, 상대적으로 낮은 값을 갖는 영역이 CNV loss이다.

다만, 실제의 리드 커버리지 데이터의 그래프는 도 3의 이상적인 경우와는 달리, 시퀀싱 에러 또는 충분치 않은 커버리지로 인하여 0값이 많이 나타나고, 무작위로 특정 포지션 i 에서의 값이 과도하게 높게 나올 수 있다. 따라서, 이러한 오차를 보정하기 위해 본 발명은 생성된 리드 커버리지 데이터를 변환한 데이터를 이용할 수 있다.

다시 도 1을 참조하면, 데이터 변환부(130)는 생성된 리드 커버리지 데이터에 제 1 변환 및 제 2 변환을 각각 적용하여 제 1 변환 데이터 및 제 2 변환 데이터를 생성할 수 있다. 여기서, 제 1 변환은 리드 커버리지 데이터의 평균값을 이용하고, 제 2 변환은 리드 커버리지 데이터의 평균 기울기값을 이용하는 것일 수 있다.

제 1 변환은 평균 이동 변환(mean shift transform) $mSH_{w,k}(C) = h_0 \cdots h_{s-1}$ 으로 각각의 h_i 값은 이하 수식 2와 같이 표현할 수 있다.

[수식 2]

$$h_i = \begin{cases} +1 & , E(c_i) > c_{th}^+ \\ 0 & , c_{th}^- \leq E(c_i) \leq c_{th}^+ \\ -1 & , E(c_i) < c_{th}^- \end{cases}$$

여기서, $E(c_i)$ 는 i 번째 윈도우의 이동 평균이고, c_{th}^+ 와 c_{th}^- 는 각각 $E(c_i)$ 의 상방 문턱값(upper threshold)과 하방 문턱값(lower threshold)을 나타낸다. 이 경우, 각각의 문턱값은 중심 극한 정리(central limit theorem)로부터 유도된 p -value $p_h(<0.05, \text{ two-tailed})$ 와 통계적 추론(statistical estimation)에 기초하여 결정된 값일 수 있다.

$E(c_i)$ 는 수식 3과 같이 표현할 수 있다.

[0033] [수식 3]

$$E(c_i) = \frac{1}{w} \sum_{j=ik}^{ik+w-1} c_j$$

[0034]

[0035] 여기서, w 는 해당 윈도우의 크기를 나타내고, k 는 해당 윈도우의 이동값(shift value)을 나타낸다. 또한, s 는 윈도우의 개수로서 $s = \lceil \frac{n-w+1}{k} \rceil$ 로 나타낼 수 있다.

[0036] 제 2 변환은 평균 기울기 변환(mean slope transform) $mSL_{w,k}(C) = l_0 \cdots l_{s-1}$ 으로 각각의 l_i 값은 이하 수식 4와 같이 표현할 수 있다.

[0037] [수식 4]

$$l_i = \begin{cases} +1 & , \quad \hat{c}_{th}^- \leq dE_i \leq \hat{c}_{th}^+ \\ 0 & , \quad otherwise \end{cases}$$

[0038]

[0039] 여기서, dE_i 는 i 번째 윈도우에 포함된 리드 커버리지 데이터의 평탄화된 차이(smoothed difference)를 나타낸다. $\hat{c}_{th}^+$ 와 $\hat{c}_{th}^-$ 는 각각 dE_i 의 상방 문턱값(upper threshold)과 하방 문턱값(lower threshold)을 나타낸다. 이 경우, 각각의 문턱값은 중심 극한 정리(central limit theorem)로부터 유도된 p-value $p_i (< 0.01, \text{two-tailed})$ 와 통계적 추론(statistical estimation)에 기초하여 결정된 값일 수 있다.

[0040] dE_i 는 수식 5와 같이 표현할 수 있다.

[0041] [수식 5]

$$dE_i = \frac{1}{w'} \left(\sum_{j=ik+w-w'}^{ik+w-1} c_j - \sum_{j=ik}^{ik+w'-1} c_j \right)$$

[0042]

[0043] 여기서, w 는 해당 윈도우의 크기를 나타내고, k 는 해당 윈도우의 이동값(shift value)을 나타내고, w' 는 0보다 크고 $w/2$ 보다 작은 값을 갖는다. 또한, s 는 윈도우의 개수로서 $s = \lceil \frac{n-w+1}{k} \rceil$ 로 나타낼 수 있다.

[0044] 앞서 언급한 제 1 변환 및 제 2 변환을 통해, 리드 커버리지 데이터는 크기값 w 와 이동값 k 를 갖는 윈도우를 통해 표본화될(sampled) 수 있다. 나아가, 크기값 w 와 이동값 k 는 시퀀싱 에러 및/또는 충분치 않은 커버리지에 의한 효과를 감쇄시키기 위해 결정될 수 있다.

[0045] CNV 후보 검출부(140)는 데이터 변환부(130)에서 생성된 제 1 변환 데이터 및 제 2 변환 데이터를 이용하여 적어도 하나 이상의 CNV 후보를 검출할 수 있다. 예를 들어, CNV 후보 검출부(140)는 제 1 변환 데이터로부터 리드 커버리지 데이터의 값들 중 통계적으로 유의한 차이를 보이는 위치(primary CNV)를 찾고, 제 2 변환 데이터로부터 안정된(stationary) 위치를 찾을 수 있다. 나아가, CNV 후보 검출부(140)는 통계적으로 유의한 차이를 보이는 연속되는 안정된 위치를 CNV 후보로 결정할 수 있다.

[0046] 본 발명의 일 실시예에서, CNV 후보 검출부(140)는 데이터 변환부(130)에서 생성된 제 1 변환 데이터 및 제 2 변환 데이터를 이용하여 CCNV gain 및 CCNV loss를 결정할 수 있다. CCNV gain는 gain으로 검출된 후보 CNV(양의 값을 갖는 CNV 후보)를 의미하고, CCNV loss는 loss로 검출된 후보 CNV(음의 값을 갖는 CNV 후보)를 의미한다.

[0047] CNV 후보 검출부(140)는 CCNV gain 및 CCNV loss를 결정하기 위하여 제 1 변환 데이터 $mSH_{w,k}(C) = h_0 \cdots h_{s-1}$ 및

제 2 변환 데이터 $mSL_{w,k}(C) = l_0 \cdots l_{s-1}$ 각각에 있어서, 동일한 포지션 i 에 대한 값들을 서로 곱할 수 있다.

[0048] 이 경우, $i_{\min} \leq i \leq i_{\max}$ 의 전제 하에, h_i 와 l_i 를 곱한 값이 0이 아닌 양수일 때, CCNV gain를 $CCNV_j^+ = [p_j, p_j+m_j-1]$ 와 같이 나타낼 수 있다. 여기서 $p_j = i_{\min} * k$ 이고, $p_j+m_j-1 = i_{\max} * k + w - 1$ 이고, $l_{cnv} \leq m_j$ 일 수 있다. 여기서, l_{cnv} 는 CNV를 식별하기 위한 최소 크기이고, p_j 와 p_j+m_j-1 각각은 j 번째 CCNV gain의 시작 위치와 끝 위치일 수 있다.

[0049] 이와 유사하게, $i_{\min} \leq i \leq i_{\max}$ 의 전제 하에, h_i 와 l_i 를 곱한 값이 0이 아닌 음수일 때, CCNV loss를 $CCNV_j^- = [p_j, p_j+m_j-1]$ 와 같이 나타낼 수 있다. 여기서 $p_j = i_{\min} * k$ 이고, $p_j+m_j-1 = i_{\max} * k + w - 1$ 이고, $l_{cnv} \leq m_j$ 일 수 있다. 여기서, l_{cnv} 는 CNV를 식별하기 위한 최소 크기이고, p_j 와 p_j+m_j-1 각각은 j 번째 CCNV loss의 시작 위치와 끝 위치일 수 있다.

[0050] CNV 후보 검출부(140)는 검출된 적어도 하나 이상의 CNV 후보 중 연이은 CNV 후보를 병합하여 하나의 CNV 후보로 결정할 수 있다. 이 경우, CNV 후보 검출부(140)는 연이은 CNV 후보를 합쳐 하나의 CNV 후보로 결정하기 위하여, 연이은 CNV 후보간 CNV 후보 병합 스코어(CCNV merging score) 값을 이용할 수 있다. 구체적으로, CNV 후보 검출부(140)는 검출된 적어도 하나 이상의 CNV 후보 중 연이은 양의 값을 갖는 CNV 후보(CCNV gain) 또는 연이은 음의 값을 갖는 CNV 후보(CCNV loss)를 병합하여 하나의 CNV 후보로 결정할 수 있다.

[0051] CNV 후보 검출부(140)는 연이은 CCNV gain에 있어서, j 번째 CCNV gain으로서 $CCNV_j^+ = [p_j, p_j+m_j-1]$ 와 $j+1$ 번째 CCNV gain으로서 $CCNV_{j+1}^+ = [p_{j+1}, p_{j+1}+m_{j+1}-1]$ 를 이용하여, CCNV gain에 대한 병합 스코어 CMS_j^+ 를 이하 수식 6과 같이 구할 수 있다.

[0052] [수식 6]

$$CMS_j^+ = \frac{\sum_{i=i_s}^{i_e} h_i}{p_{j+1} + m_{j+1} - p_j}$$

[0053]

[0054] 여기서 $p_j = i_s * k$ 와 $p_{j+1}+m_{j+1}-1 = i_e * k + w - 1$ 각각은 j 번째 CCNV gain의 시작 위치, $j+1$ 번째 CCNV gain의 끝 위치를 나타낸다. 또한, i_s 는 j 번째 CCNV gain의 $i_{\min}$, i_e 는 $j+1$ 번째 CCNV gain의 $i_{\max}$ 를 나타낸다.

[0055] j 번째 CCNV gain와 $j+1$ 번째 CCNV gain에 대한 병합 스코어 CMS_j^+ 에 대하여, CMS_j^+ 값이 기설정된 값보다 크고, $CCNV_j^+$ 와 $CCNV_{j+1}^+$ 사이에 CCNV loss가 포함되지 않는 경우, CNV 후보 검출부(140)는 연이은 $CCNV_j^+$ 와 $CCNV_{j+1}^+$ 를 병합하여 CNV 후보로 결정할 수 있다.

[0056] CCNV gain의 경우와 유사하게, CNV 후보 검출부(140)는 연이은 CCNV loss에 있어서, j 번째 CCNV loss로서 $CCNV_j^- = [p_j, p_j+m_j-1]$ 와 $j+1$ 번째 CCNV loss로서 $CCNV_{j+1}^- = [p_{j+1}, p_{j+1}+m_{j+1}-1]$ 를 이용하여, CCNV loss에 대한 병합 스코어 CMS_j^- 를 이하 수식 7을 이용하여 구할 수 있다.

[0057] [수식 7]

$$CMS_j^- = \frac{\sum_{i=i_s}^{i_e} h_i}{p_{j+1} + m_{j+1} - p_j}$$

[0058]

[0059] $p_j = i_s * k$ 와 $p_{j+1}+m_{j+1}-1 = i_e * k + w - 1$ 각각은 j 번째 CCNV loss의 시작 위치, $j+1$ 번째 CCNV loss의 끝 위치를 나타낸다. 또한, i_s 는 j 번째 CCNV loss의 $i_{\min}$, i_e 는 $j+1$ 번째 CCNV loss의 $i_{\max}$ 를 나타낸다.

- [0060] j 번째 CCNV loss와 j+1 번째 CCNV loss에 대한 병합 스코어 CMS_j^- 에 대하여, CMS_j^- 값이 기설정된 값보다 작고, $CCNV_j^-$ 와 $CCNV_{j+1}^-$ 사이에 CCNV gain이 포함되지 않는 경우, CNV 후보 검출부(140)는 연이은 $CCNV_j^-$ 와 $CCNV_{j+1}^-$ 를 병합하여 CNV 후보로 결정할 수 있다.
- [0061] 이 때, CCNV gain 및 CCNV loss, 또는 병합된 CCNV gain 및 CCNV loss를 포함하는 CNV 후보는 CNV의 조건으로서 기설정된 길이(kbp) 이상이고, 테스트 시퀀스 또는 레퍼런스 시퀀스에 CCNV가 적어도 둘 이상 포함되는 것일 수 있다. 즉, 기설정된 길이(kbp) 미만의 연이은 CCNV gain 또는 연이은 CCNV loss가 병합되어 기설정된 길이(kbp) 이상이 되었다면 CNV의 조건을 만족하는 것으로 볼 수 있다.
- [0062] 이하 도 4를 참조하여, CNV의 조건을 설명하겠다.
- [0063] 도 4는 본 발명의 일 실시예에 따른, CNV를 검출하기 위한 CNV의 조건을 나타낸 도면이다. 도 4를 참조하면, 레퍼런스 시퀀스 S에 서브 시퀀스 X가 1개 있고, 테스트 시퀀스 T에 서브 시퀀스 X가 2개 있음을 알 수 있다. 즉, 테스트 시퀀스 T에는 레퍼런스 시퀀스 S에 있는 서브 시퀀스 X의 포지션 i와는 다른 포지션에 서브 시퀀스 X가 존재하고, 이는 CNV의 첫번째 조건을 만족한다. 또한, 서브 시퀀스 X의 길이는 기설정된 길이(예를 들어, 1kbp)보다 긴 2kbp이므로 CNV의 두번째 조건을 만족한다.
- [0064] 다시, 도 1을 참조하면, CNV 검출부(150)는 앞서 언급한 CNV 조건으로서 기설정된 길이(kbp) 이상이고, 테스트 시퀀스 또는 레퍼런스 시퀀스에 해당 CNV가 적어도 둘 이상 포함되는 경우의 CNV 후보를 CNV로서 검출할 수 있다. 예를 들어, CNV 검출부(150)는 CNV 후보 검출부(140)에서 검출된 적어도 하나 이상의 CNV 후보 중 적어도 하나 이상을 CNV로 검출할 수 있다. 즉, CNV 후보가 모두 CNV가 될 수도 있지만, 항상 그런 것은 아니고, CNV 후보들 중 일부의 CNV 후보만이 CNV로 검출될 수 있다.
- [0065] DB(160)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 구성요소들 간에 입력 및 출력되는 데이터를 포함한다. 예를 들어, DB(160)는 데이터 생성부(120)에서 생성된 리드 커버리지 데이터를 저장할 수 있고, 데이터 변환부(130)에서 생성된 제 1 변환 데이터 및/또는 제 2 변환 데이터를 저장할 수 있다. 이러한 DB(160)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.
- [0066] 당업자라면, 리드 획득부(110), 데이터 생성부(120), 데이터 변환부(130), CNV 후보 검출부(140), CNV 검출부(150) 및 DB(160) 각각이 분리되어 구현되거나, 이 중 하나 이상이 통합되어 구현될 수 있음을 충분히 이해할 것이다.
- [0067] 도 5는 본 발명의 일 실시예에 따른, 유전체 단위 반복 변이 검출 장치가 CNV를 검출하는 방법을 나타낸 동작 흐름도이다. 도 5에 도시된 실시예에 따른 CNV를 검출하는 방법은 도 1 내지 도 4에 도시된 실시예에 따른 유전체 단위 반복 변이 검출 장치(100)에서 시계열적으로 처리되는 단계들을 포함한다. 따라서, 이하 생략된 내용이라고 하더라도 도 1 내지 도 4의 유전체 단위 반복 변이 검출 장치(100)에 관하여 기술된 내용은 도 5에 도시된 실시예에 따른 CNV를 검출하는 방법에도 적용될 수 있다.
- [0068] 단계 S501에서 유전체 단위 반복 변이 검출 장치(100)는 테스트 시퀀스에 대응하는 복수의 리드를 획득할 수 있다.
- [0069] 단계 S502에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S501에서 획득된 복수의 리드(read)를 레퍼런스 시퀀스에 정렬시킴으로써 리드 커버리지 데이터를 생성할 수 있다.
- [0070] 단계 S503에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S502에서 생성된 리드 커버리지 데이터에 제 1 변환 및 제 2 변환을 각각 적용하여 제 1 변환 데이터 및 제 2 변환 데이터를 생성할 수 있다.
- [0071] 단계 S504에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S503에서 생성된 제 1 변환 데이터 및 제 2 변환 데이터를 이용하여 적어도 하나 이상의 CNV 후보를 검출할 수 있다.
- [0072] 단계 S505에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S504에서 검출된 적어도 하나 이상의 CNV 후보 중 적어도 하나 이상을 CNV로 검출할 수 있다.
- [0073] 상술한 설명에서, 단계 S501 내지 S505는 본 발명의 구현예에 따라서, 추가적인 단계들로 더 분할되거나, 더 적

은 단계들로 조합될 수 있다. 또한, 일부 단계는 필요에 따라 생략될 수도 있고, 단계 간의 순서가 변경될 수도 있다.

- [0074] 도 6은 본 발명의 다른 실시예에 따른, 유전체 단위 반복 변이 검출 장치(100)가 CNV를 검출하는 방법을 나타낸 동작 흐름도이다. 도 6에 도시된 실시예에 따른 CNV를 검출하는 방법은 도 1 내지 도 5에 도시된 실시예에 따른 유전체 단위 반복 변이 검출 장치(100)에서 시계열적으로 처리되는 단계들을 포함한다. 따라서, 이하 생략된 내용이라고 하더라도 도 1 내지 도 5의 유전체 단위 반복 변이 검출 장치(100)에 관하여 기술된 내용은 도 6에 도시된 실시예에 따른 CNV를 검출하는 방법에도 적용될 수 있다.
- [0075] 유전체 단위 반복 변이 검출 장치(100)가 CNV를 검출하는 방법은 크게 4단계로 나눌 수 있다. 리드 커버리지 데이터를 생성하는 제 1 단계는 단계 S601 및 단계 S602를 포함할 수 있고, 변환 데이터를 생성하는 제 2단계는 단계 S603 및 단계 S604를 포함할 수 있다. 또한, CNV 후보를 검출하는 제 3단계는 단계 S605를 포함할 수 있고, 후처리 단계로서 CNV를 검출하는 제 4단계는 단계 S606 및 단계 S607을 포함할 수 있다.
- [0076] 단계 S601에서 유전체 단위 반복 변이 검출 장치(100)는 레퍼런스 시퀀스에 복수의 리드를 정렬시킬 수 있고, 단계 S602에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S601에서 레퍼런스 시퀀스에 정렬된 복수의 리드들의 개수를 레퍼런스 시퀀스의 포지션 i 별로 카운트하여 테스트 시퀀스의 리드 커버리지 데이터를 생성할 수 있다.
- [0077] 단계 S603에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S602에서 생성된 리드 커버리지 데이터에 제 1 변환 Mean Shift Transform을 적용하여 제 1 변환 데이터를 생성할 수 있다. 또한, 단계 S604에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S602에서 생성된 리드 커버리지 데이터에 제 2 변환 Mean Slope Transform을 적용하여 제 2 변환 데이터를 생성할 수 있다.
- [0078] 단계 S605에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S603 및 단계 S604에서 생성된 제 1 변환 데이터 및 제 2 변환 데이터를 이용하여 CNV 후보 영역을 검출할 수 있다.
- [0079] 단계 S606에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S605에서 검출된 CNV 후보 영역 중 인접한 2개의 CNV 후보에 대하여 병합 스코어(merging score)를 획득하고, 획득된 병합 스코어에 기초하여 인접한 2개의 CNV 후보를 병합할지 결정할 수 있다.
- [0080] 단계 S607에서 유전체 단위 반복 변이 검출 장치(100)는 단계 S605에서 검출된 CNV 후보 및 단계 S606에서 병합된 CNV 후보 중 적어도 하나 이상을 CNV로서 검출할 수 있다.
- [0081] 상술한 설명에서, 단계 S601 내지 S607는 본 발명의 구현예에 따라서, 추가적인 단계들로 더 분할되거나, 더 적은 단계들로 조합될 수 있다. 또한, 일부 단계는 필요에 따라 생략될 수도 있고, 단계 간의 순서가 변경될 수도 있다.
- [0082] 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.
- [0083] 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.
- [0084] 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.

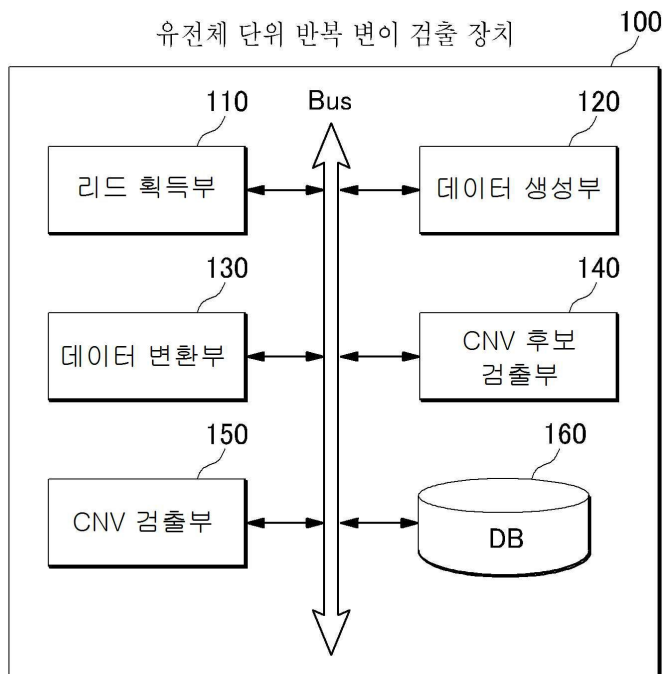
부호의 설명

[0085]

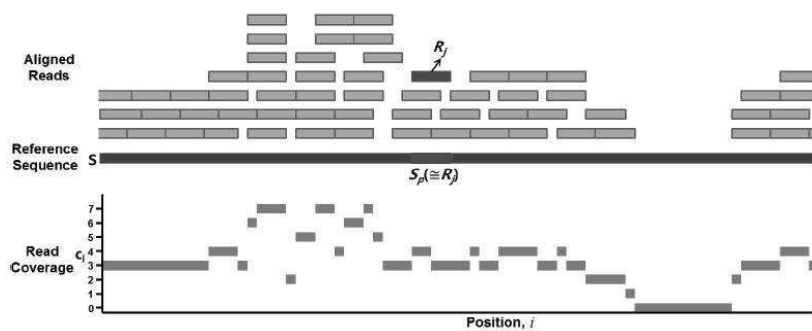
- 100: 장치
- 110: 획득부
- 120: 데이터 생성부
- 130: 데이터 변환부
- 140: CNV 후보 검출부
- 150: CNV 검출부
- 160: DB

도면

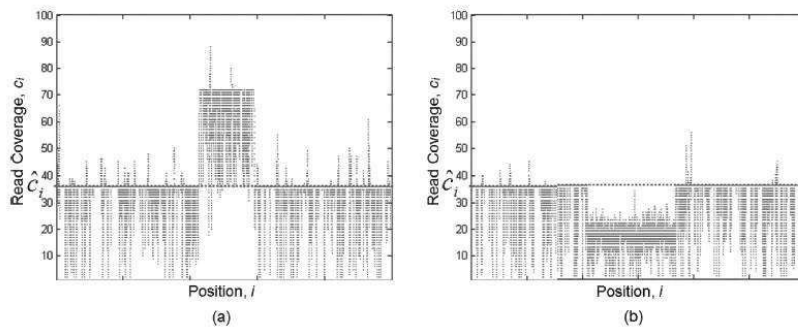
도면1



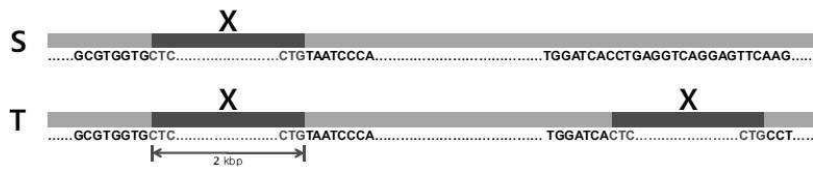
도면2



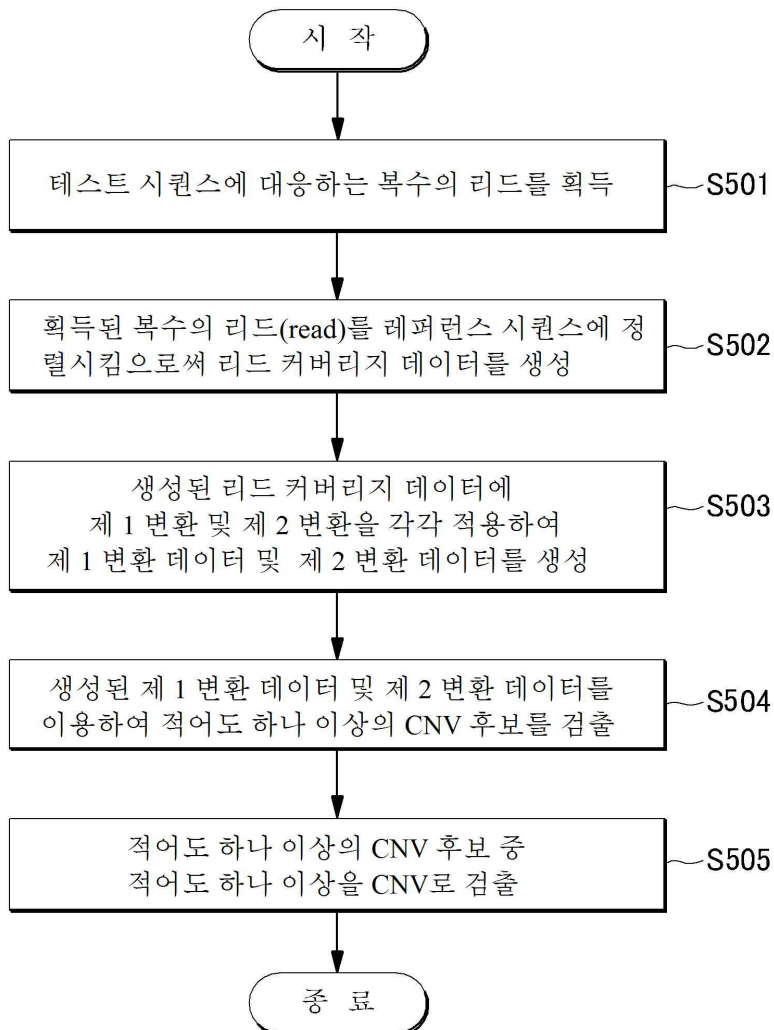
도면3



도면4



도면5



도면6

