



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2018-0098093
(43) 공개일자 2018년09월03일

(51) 국제특허분류(Int. Cl.)
G06F 19/22 (2017.01) G06F 19/24 (2011.01)
(52) CPC특허분류
G06F 19/22 (2018.05)
G06F 19/24 (2013.01)
(21) 출원번호 10-2017-0025113
(22) 출원일자 2017년02월24일
심사청구일자 2017년02월24일

(71) 출원인
한림대학교 산학협력단
강원도 춘천시 한림대학길 1, 한림대학교(옥천동)
(72) 발명자
공진화
경기도 양주시 화합로1327번길 27, 201호(덕정동, 대명빌라)
신재문
전라남도 광양시 광양읍 석정길 25-4
(74) 대리인
김남혁

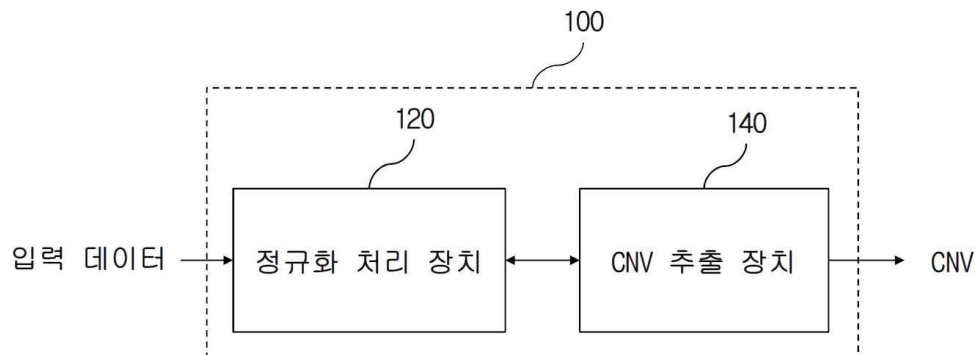
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 유전자 단위 반복 변이 검출 시스템 및 방법

(57) 요약

4단계의 정규화 변환 기법을 적용한 엑솜 시퀀싱 데이터 변환을 통해 데이터의 잡음 특성을 최소화하여 CNV 영역을 정확하게 추출하도록 한 유전자 단위 반복 변이 검출 시스템 및 방법을 제시한다. 제시된 유전자 단위 반복 변이 검출 시스템의 정규화 처리 장치는 테스트 데이터 및 컨트롤 데이터를 근거로 산출한 비율 데이터 또는 테스트 데이터를 타깃 영역에 대해 동일한 크기를 갖는 복수의 세그먼트로 분할하고, 분할된 복수의 세그먼트의 평균값의 표준화 평균값과 각 세그먼트의 비율 데이터를 근거로 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환하여 유효 비율 데이터를 추출한다.

대표도 - 도1



(72) 발명자

윤지희

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

이은주

강원도 춘천시 국사봉길 7, 104동 1403호(퇴계동, 우성아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호 NRF-2014R1A2A1A11052141

부처명 미래창조과학부

연구관리전문기관 한국연구재단

연구사업명 중견연구자지원사업(핵심연구)

연구과제명 희귀질환의 원인 유전자 발견을 위한 엑솜 시퀀싱 데이터의 가공, 분석 및 통합 프로파일링 기법에 관한 연구

기 여 율 1/1

주관기관 한림대학교 산학협력단

연구기간 2016.11.01 ~ 2017.10.31

명세서

청구범위

청구항 1

유전자 단위 반복 변이 검출 시스템의 입력 데이터를 정규화 처리하는 정규화 처리 장치에 있어서,
테스트 데이터 및 컨트롤 데이터의 리드 정렬 결과로 이루어진 두 입력 데이터를 근거로 각각 커버리지 데이터를 산출하고, 각 커버리지 데이터를 근거로 각 엑손 타깃 영역만의 리드 커버리지 데이터를 추출하고, 상기 리드 커버리지 데이터의 비율 데이터를 산출하는 커버리지 산출모듈;
상기 커버리지 산출모듈에서 산출한 비율 데이터를 근거로 타깃 영역에 대한 비율 데이터를 동일한 크기를 갖는 복수의 세그먼트로 분할하거나, 테스트 데이터를 타깃 영역에 대해 동일한 크기를 갖는 복수의 세그먼트로 분할하는 세그멘테이션 처리모듈;
상기 세그멘테이션 처리모듈에서 분할된 복수의 세그먼트의 평균값을 표준화 평균값으로 조정하는 평균값 조정모듈; 및
상기 평균값 조정모듈에서 조정한 상기 표준화 평균값과 각 세그먼트의 비율 데이터를 근거로 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환하여 유효 비율 데이터를 추출하는 정규화 변환모듈을 포함하는 정규화 처리 장치.

청구항 2

제1항에 있어서,
상기 커버리지 산출모듈은,
수학식

$$R_{T|C}^e(i, j) = \begin{cases} 0, & R_T^e(i, j) < \varepsilon \\ 1, & (R_T^e(i, j) < \varepsilon) \& (R_C^e(i, j) < \varepsilon) \\ R_T^e(i, j) / R_C^e(i, j), & others \end{cases}$$

($R_T^e(i, j)$)는 테스트 데이터의 리드 커버리지 데이터이고, ($R_C^e(i, j)$)는 컨트롤 데이터의 리드 커버리지 데이터이고, ε 는 10^{-3})을 이용하여 비율 데이터를 산출하고,

보정계수(ω)를 곱하여 상기 비율 데이터를 보정하되,

상기 보정계수(ω)는 라이브러리 사이즈의 불균형성을 보정하기 위한 방법으로, 컨트롤 데이터에 존재하는 모든 타깃 영역의 커버리지 값을 모두 더한 값을 테스트 데이터에 존재하는 모든 타깃 영역의 커버리지 값을 모두 더한 값으로 나눈 값인 정규화 처리 장치.

청구항 3

제1항에 있어서,

상기 세그멘테이션 처리모듈은 각 타깃 영역($e_i(1 \leq i \leq N_e)$)에 대한 비율 데이터($R_{T|C}^e(i, j)$, $b_f(i) \leq j \leq b_l(i)$)를 동일한 크기(b_s)를 갖는 $N_s(i) = [N_b(i)/b_s] - 1$ 개의 세그먼트로 분할하여, 각 세그먼트가 동일한 크기를 갖는 비율 데이터($R_{T|C}^{s_i, k}(j_k)$ ($1 \leq i \leq N_e$, $1 \leq k \leq n_s(i)$, $1 \leq j_k \leq b_s$))를 산출하되,

세그먼트 분할 시 $b_f(i) = [\text{mod}(n_b(i), b_s) + b_s] / 2 + 1$ 에서 타깃 영역의 첫 번째 세그먼트를 시작하고, 양끝의 남은

영역($R_{T|C}^e(i, j)$ ($1 \leq j \leq b_s(i)-1$)과 $R_{T|C}^e(i, j)$ ($b_j(i) \leq j \leq N_b(i)$))의 데이터는 제외하는 정규화 처리 장치.

청구항 4

제1항에 있어서,

상기 평균값 조정모듈은,

수학식

$$mR_s^e(i, k) = \sum_{j=b_f(i)+(k-1)*b_s}^{b_f(i)+kb_s} R_{T|C}^e(i, j)/b_s, \quad (1 \leq k \leq n_s(i))$$

을 이용하여 각 세그먼트의 평균값을 산출하고,

수학식

$$tR_s^e(i, k) = (mR_s^e(i, k) - M(mR_s^e))/(SD(mR_s^e)/\sqrt{b_s}), \quad (1 \leq k \leq n_s(i))$$

(여기서, $M(mR_s^e)$ 는 전체 세그먼트 $S_{i,k}$ ($1 \leq i \leq N_e$, $1 \leq k \leq n_s(i)$)에 걸쳐져 있는 평균값의 평균이고, $SD(mR_s^e)$ 는 전체 세그먼트 $S_{i,k}$ ($1 \leq i \leq N_e$, $1 \leq k \leq n_s(i)$)에 걸쳐져 있는 평균값의 표준 편차)

를 이용하여 상기 표준화 평균값을 산출하고,

상기 산출한 표준화 평균값이 음수이면 상기 표준화 평균값에서 표준화 평균값의 최소값을 차감하여 상기 표준화 평균값을 보정하는 정규화 처리 장치.

청구항 5

제1항에 있어서,

상기 정규화 변환모듈은

상기 표준화 평균값과 각 세그먼트의 비율 데이터의 표준 편차를 근거로 각 세그먼트 내 각 위치선의 데이터가 정규 분포를 갖도록 변환하여 유효 비율 데이터를 추출하는 정규화 처리 장치.

청구항 6

테스트 데이터 및 컨트롤 데이터를 포함하는 두 입력 데이터를 이용하여 비율 데이터를 산출한 후 정규화 처리하거나, 테스트 데이터를 포함한 입력데이터를 정규화 처리하는 정규화 처리 장치; 및

상기 정규화 처리 장치에서 정규화 처리된 입력 데이터를 근거로 유전자 단위 반복 변이를 추출하는 CNV 추출 장치를 포함하고,

상기 정규화 처리 장치는,

상기 입력 데이터에 대해 각 타깃 영역에서의 염기 단위 리드 커버리지를 산출하고, 상기 염기 단위 리드 커버리지를 근거로 세그먼테이션하여 상기 타깃 영역을 동일한 크기를 갖는 복수의 세그먼트로 분할하고, 상기 복수의 세그먼트들 각각의 리드 커버리지 데이터의 평균값을 표준화 평균값으로 조정하고, 상기 표준화 평균값 및 각 세그먼트의 비율 데이터를 근거로 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환하는 유전자 단위 반복 변이 검출 시스템.

청구항 7

유전자 단위 반복 변이 검출을 위해 정규화 처리 장치를 이용하여 입력 데이터를 정규화 처리하는 방법에 있어서,

테스트 데이터 및 컨트롤 데이터를 포함하는 입력 데이터를 근거로 타깃 단위의 커버리지를 산출하는 단계;

상기 커버리지 데이터를 근거로 각 타깃 영역의 리드 커버리지 데이터를 추출하는 단계;

상기 리드 커버리지 데이터의 비율 데이터를 산출하는 단계;

상기 비율 데이터를 근거로 타깃 영역에 대한 비율 데이터를 동일한 크기를 갖는 복수의 세그먼트로 분할하거나, 테스트 데이터를 동일한 크기를 갖는 복수의 세그먼트로 분할하는 단계;

상기 분할된 복수의 세그먼트의 평균값을 표준화 평균값으로 조정하는 단계; 및

상기 표준화 평균값과 각 세그먼트의 비율 데이터를 근거로 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환하여 유효 비율 데이터를 추출하는 단계를 포함하는 정규화 처리 방법.

청구항 8

제7항에 있어서,

상기 비율 데이터를 산출하는 단계에서는,

수학식

$$R_{T|C}^e(i, j) = \begin{cases} 0, & R_T^e(i, j) < \varepsilon \\ 1, & (R_T^e(i, j) < \varepsilon) \& (R_C^e(i, j) < \varepsilon) \\ R_T^e(i, j) / R_C^e(i, j), & others \end{cases}$$

($R_T^e(i, j)$)는 테스트 데이터의 리드 커버리지 데이터이고, ($R_C^e(i, j)$)는 컨트롤 데이터의 리드 커버리지 데이터이고, ε 는 10^{-3} 을 이용하여 비율 데이터를 산출하는 단계; 및

보정계수(ω)를 곱하여 상기 비율 데이터를 보정하는 단계를 포함하되,

상기 보정계수(ω)는 컨트롤 데이터에 존재하는 모든 타깃 영역의 커버리지 값을 모두 더한 값을 테스트 데이터에 존재하는 모든 타깃 영역의 커버리지 값을 모두 더한 값으로 나눈 값인 정규화 처리 방법.

청구항 9

제7항에 있어서,

상기 복수의 세그먼트로 분할하는 단계는,

각 타깃 영역($e_i(1 \leq i \leq N_e)$)에 대한 비율 데이터($R_{T|C}^e(i, j)$, $b_f(i) \leq j \leq b_l(i)$)를 동일한 크기(b_s)를 갖는 $N_s(i) = [N_b(i)/b_s] - 1$ 개의 세그먼트로 분할하여, 각 세그먼트가 동일한 크기를 갖는 비율 데이터($R_{T|C}^{S_i, k}(j_k)$ ($1 \leq i \leq N_e$, $1 \leq k \leq n_s(i)$, $1 \leq j_k \leq b_s$))를 산출하되,

세그먼트 분할 시 $b_f(i) = [\text{mod}(n_b(i), b_s) + b_s] / 2 + 1$ 에서 타깃 영역의 첫 번째 세그먼트를 시작하고, 양끝의 남은 영역($R_{T|C}^e(i, j)$ ($1 \leq j \leq b_s(i) - 1$))과 $R_{T|C}^e(i, j)$ ($b_j(i) \leq j \leq N_b(i)$)의 데이터를 제외하는 정규화 처리 방법.

청구항 10

제7항에 있어서,

상기 표준화 평균값으로 조정하는 단계는,

수학식

$$mR_s^e(i, k) = \sum_{j=b_f(i)+(k-1)*b_s}^{b_f(i)+kb_s} R_{T|C}^e(i, j) / b_s, \quad (1 \leq k \leq n_s(i))$$

을 이용하여 각 세그먼트의 평균값을 산출하는 단계;

수학식

$$tR_s^e(i, k) = (mR_s^e(i, k) - M(mR_s^e)) / (SD(mR_s^e) / \sqrt{b_s}), (1 \leq k \leq n_s(i))$$

(여기서, $M(mR_s^e)$ 는 전체 세그먼트 $S_{i,k}(1 \leq i \leq N_e, 1 \leq k \leq n_s(i))$ 에 걸쳐져 있는 평균값의 평균이고, $SD(mR_s^e)$ 는 전체 세그먼트 $S_{i,k}(1 \leq i \leq N_e, 1 \leq k \leq n_s(i))$ 에 걸쳐져 있는 평균값의 표준 편차를 이용하여 상기 표준화 평균값을 산출하는 단계; 및

상기 표준화 평균값이 음수이면 상기 표준화 평균값에서 표준화 평균값의 최소값을 차감하여 상기 표준화 평균값을 보정하는 단계를 포함하는 정규화 처리 방법.

청구항 11

제7항에 있어서,

상기 유효 비율 데이터를 추출하는 단계에서는,

상기 표준화 평균값과 각 세그먼트의 비율 데이터의 표준 편차를 근거로 각 세그먼트 내 각 기본 위치선의 데이터가 정규 분포를 갖도록 변환하여 유효 비율 데이터를 추출하는 정규화 처리 방법.

청구항 12

정규화 처리 장치에 의해, 테스트 데이터 및 컨트롤 데이터 또는 데이트 데이터를 포함하는 입력 데이터를 정규화 처리하는 단계; 및

CNV 추출 장치에 의해, 상기 입력 데이터를 정규화 처리하는 단계에서 정규화 처리된 입력 데이터를 근거로 유전자 단위 반복 변이를 추출하는 단계를 포함하고,

상기 입력 데이터를 정규화 처리하는 단계는,

상기 입력 데이터에 대해 각 타깃 영역에서의 염기 단위 리드 커버리지를 산출하는 단계;

상기 염기 단위 리드 커버리지를 근거로 세그먼트이션하여 상기 타깃 영역을 동일한 크기를 갖는 복수의 세그먼트로 분할하는 단계;

상기 복수의 세그먼트들 각각의 리드 커버리지 데이터의 평균값을 표준화 평균값으로 조정하는 단계; 및

상기 표준화 평균값 및 각 세그먼트의 비율 데이터를 근거로 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환하는 단계를 포함하는 유전자 단위 반복 변이 검출 방법.

발명의 설명

기술 분야

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

배경 기술

[0002] 유전자 단위 반복 변이(CNV; Copy Number Variation)는 서로 다른 두 DNA 시퀀스를 비교하여 50bp 이상의 DNA 세그먼트의 카피 수(copy number)가 서로 다른 경우의 변이로 정의된다.

[0003] 유전자 단위 반복 변이는 자폐증(autism), 지적 장애(intellectual disability), 뇌전증(epilepsy), 조현병(schizophrenia), 소아비만(obesity), 암(cancer) 등과 같은 인간의 질병과 연관성이 있는 매우 중요한 변이 유형 중 하나이다.

[0004] 특히 최근 연구에서는 엑솜 시퀀싱 데이터를 이용한 rare, de novo CNV 추출에 근거한 원인 유전자 규명에 관한 연구 결과가 지속적으로 보고되고 있다.

[0005] 그러나 WES(Whole Exome Sequencing) 데이터는 작은 엑솜(Exome) 영역의 크기, 엑솜 영역의 비연속성, 프로브

(probe) 의존성 등의 다양한 잡음 특성을 가지고 있다.

- [0006] 대표적인 WES 데이터 기반의 CNV 추출 알고리즘은 ExomeCNV, Contra, CoNIFER, XHMM, Excavator 등이 있고, 각 알고리즘은 WES 데이터의 잡음 특성을 보정하기 위하여 다음과 같은 방식을 채택하고 있다.
- [0007] ExomeCNV는 Geary-Hinkley transformation을 적용하여 리드 커버리지 데이터의 정규화 분포를 유도하고 있다.
- [0008] Contra는 base-level log-ratio와 library size 보정 등의 작업을 수행하는 정규화 방식을 채택하고 있다.
- [0009] CoNIFER와 XHMM은 singular value decomposition(SVD) 방식과 principal component analysis(PCA) 방식을 적용하여 리드 커버리지 데이터에 포함된 잡음을 제거하고 있다.
- [0010] 또한, Excavator는 GC content, mappability, exon size 등에서 발생하는 잡음을 제거하기 위하여 median 정규화 방식을 채택하고 있다.
- [0011] 그러나 CoNIFER와 XHMM은 SVD 혹은 PCA 정규화 방식을 적용하기 위하여 매우 많은 샘플 데이터를 동시에 사용하기 때문에, 많은 샘플을 갖추지 못한 제한적 환경에서는 적용이 어려운 문제점이 있다.
- [0012] 또한, Contra에서도 base control을 생성하기 위하여 많은 샘플을 이용하여야 하는 제약이 있다.
- [0013] 또한, 몇몇 알고리즘에서는 변이 영역을 구분하기 위한 분할 방식으로서 circular binary segmentation (CBS) 알고리즘을 적용하고 있다.
- [0014] 그러나, CBS 알고리즘은 일반적으로 연속 영역을 분할하는데 사용되는 방식으로서 엑솔 커버리지 데이터와 같이 비연속적인 데이터 영역에 적용하는 경우, 영역 크기가 매우 작거나 큰 경우의 CNV 영역을 놓칠 확률이 커지는 문제점이 있다.

선행기술문헌

특허문헌

- [0015] (특허문헌 0001) 한국공개특허 제10-2017-0000744호(명칭: 유전자의 복제수 변이(CNV)를 분석하는 방법 및 장치)

발명의 내용

해결하려는 과제

- [0016] 본 발명의 상술한 문제점을 해결하고자 제안된 것으로, 4단계의 정규화 변환 기법을 적용한 엑솔 시퀀싱 데이터 변환을 통해 데이터의 잡음 특성을 최소화하여 CNV 영역을 정확하게 추출하도록 한 유전자 단위 반복 변이 검출 시스템 및 방법을 제공하는 것을 목적으로 한다.

과제의 해결 수단

- [0017] 상기한 목적을 달성하기 위한 본 발명의 실시예에 따른 유전자 단위 반복 변이 검출 시스템의 입력 데이터를 정규화 처리하는 정규화 처리 장치에 있어서, 테스트 데이터 및 컨트롤 데이터의 리드 정렬 결과로 이루어진 두 입력 데이터를 근거로 각각 커버리지 데이터를 산출하고, 각 커버리지 데이터를 근거로 각 엑손 타깃 영역만의 리드 커버리지 데이터를 추출하고, 리드 커버리지 데이터의 비율 데이터를 산출하는 커버리지 산출모듈, 커버리지 산출모듈에서 산출한 비율 데이터를 근거로 타깃 영역에 대한 비율 데이터를 동일한 크기를 갖는 복수의 세그먼트로 분할하거나, 테스트 데이터를 타깃 영역에 대해 동일한 크기를 갖는 복수의 세그먼트로 분할하는 세그먼트이션 처리모듈, 세그먼트이션 처리모듈에서 분할된 복수의 세그먼트의 평균값을 표준화 평균값으로 조정하는 평균값 조정모듈 및 평균값 조정모듈에서 조정한 표준화 평균값과 각 세그먼트의 비율 데이터를 근거로 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환하여 유효 비율 데이터를 추출하는 정규화 변환모듈을 포함한다.

[0018] 커버리지 산출모듈은 수학적식

$$R_{T|C}^e(i, j) = \begin{cases} 0, & R_T^e(i, j) < \varepsilon \\ 1, & (R_T^e(i, j) < \varepsilon) \& (R_C^e(i, j) < \varepsilon) \\ R_T^e(i, j) / R_C^e(i, j), & \text{others} \end{cases}$$

[0019]

[0020] ($R_T^e(i, j)$)는 테스트 데이터의 리드 커버리지 데이터이고, ($R_C^e(i, j)$)는 컨트롤 데이터의 리드 커버리지 데이터이고, (ε 는 10^{-3})을 이용하여 비율 데이터를 산출하고, 보정계수(ω)를 곱하여 비율 데이터를 보정하되, 보정계수(ω)는 라이브러리 사이즈의 불균형성을 보정하기 위한 방법으로, 컨트롤 데이터에 존재하는 모든 타깃 영역의 커버리지 값을 모두 더한 값을 테스트 데이터에 존재하는 모든 타깃 영역의 커버리지 값을 모두 더한 값으로 나눈 값일 수 있다.

[0021] 세그먼테이션 처리모듈은 각 타깃 영역($e_i(1 \leq i \leq N_e)$)에 대한 비율 데이터($R_{T|C}^e(i, j)$, $b_f(i) \leq j \leq b_l(i)$)를 동일한 크기(b_s)를 갖는 $N_s(i) = [N_b(i)/b_s] - 1$ 개의 세그먼트로 분할하여, 각 세그먼트가 동일한 크기를 갖는 비율 데이터($R_{T|C}^{S_i, k}(j_k)$ ($1 \leq i \leq N_e$, $1 \leq k \leq n_s(i)$, $1 \leq j_k \leq b_s$))를 산출하되, 세그먼트 분할 시 $b_f(i) = [(\text{mod}(n_b(i), b_s) + b_s)/2] + 1$ 에서 타깃 영역의 첫 번째 세그먼트를 시작하고, 양끝의 남은 영역($R_{T|C}^e(i, j)$ ($1 \leq j \leq b_s(i) - 1$)과 $R_{T|C}^e(i, j)$ ($b_j(i) \leq j \leq N_b(i)$))의 데이터는 제외할 수 있다.

[0022] 평균값 조정모듈은 수학적식

$$mR_s^e(i, k) = \sum_{j=b_f(i)+(k-1)*b_s}^{b_f(i)+kb_s} R_{T|C}^e(i, j) / b_s$$

[0023]

[0024] ($1 \leq k \leq n_s(i)$)을 이용하여 각 세그먼트의 평균값을 산출하고, 수학적식

$$tR_s^e(i, k) = (mR_s^e(i, k) - M(mR_s^e)) / (SD(mR_s^e) / \sqrt{b_s}), \quad (1 \leq k \leq n_s(i))$$

[0025]

[0026] (여기서, $M(mR_s^e)$ 는 전체 세그먼트 $S_{i, k}(1 \leq i \leq N_e, 1 \leq k \leq n_s(i))$ 에 걸쳐져 있는 평균값의 평균이고, $SD(mR_s^e)$ 는 전체 세그먼트 $S_{i, k}(1 \leq i \leq N_e, 1 \leq k \leq n_s(i))$ 에 걸쳐져 있는 평균값의 표준 편차)를 이용하여 표준화 평균값을 산출하고, 산출한 표준화 평균값이 음수이면 표준화 평균값에서 표준화 평균값의 최소값을 차감하여 표준화 평균값을 보정할 수 있다.

[0027] 정규화 변환모듈은 표준화 평균값과 각 세그먼트의 비율 데이터의 표준 편차를 근거로 각 세그먼트 내 각 포지션의 데이터가 정규 분포를 갖도록 변환하여 유효 비율 데이터를 추출할 수 있다.

[0028] 상기한 목적을 달성하기 위한 본 발명의 실시예에 따른 유전자 단위 반복 변이 검출 시스템은 테스트 데이터 및 컨트롤 데이터를 포함하는 두 입력 데이터를 이용하여 비율 데이터를 산출한 후 정규화 처리하거나, 데이터 데이터를 포함한 입력데이터를 정규화 처리하는 정규화 처리 장치 및 정규화 처리 장치에서 정규화 처리된 입력 데이터를 근거로 유전자 단위 반복 변이를 추출하는 CNV 추출 장치를 포함하고, 정규화 처리 장치는 입력 데이터에 대해 각 타깃 영역에서의 염기 단위 리드 커버리지를 산출하고, 염기 단위 리드 커버리지를 근거로 세그먼테이션하여 타깃 영역을 동일한 크기를 갖는 복수의 세그먼트로 분할하고, 복수의 세그먼트들 각각의 리드 커버리지 데이터의 평균값을 표준화 평균값으로 조정하고, 표준화 평균값 및 각 세그먼트의 비율 데이터를 근거로 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환한다.

[0029] 상기한 목적을 달성하기 위한 본 발명의 실시예에 따른 정규화 처리 방법은 유전자 단위 반복 변이 검출을 위해 정규화 처리 장치를 이용하여 입력 데이터를 정규화 처리하는 방법으로, 테스트 데이터 및 컨트롤 데이터를 포함하는 입력 데이터를 근거로 타깃 단위의 커버리지를 산출하는 단계, 커버리지 데이터를 근거로 각 타깃 영역의 리드 커버리지 데이터를 추출하는 단계, 리드 커버리지 데이터의 비율 데이터를 산출하는 단계, 산출한 비율 데이터를 근거로 타깃 영역에 대한 비율 데이터를 동일한 크기를 갖는 복수의 세그먼트로 분할하거나, 테스트 데이터를 동일한 크기를 갖는 복수의 세그먼트로 분할하는 단계, 분할된 복수의 세그먼트의 평균값을 표준화 평

균값으로 조정하는 단계 및 조정한 표준화 평균값과 각 세그먼트의 비율 데이터를 근거로 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환하여 유효 비율 데이터를 추출하는 단계를 포함한다.

[0030] 비율 데이터를 산출하는 단계에서는 수학적식

$$R_{T|C}^e(i, j) = \begin{cases} 0, & R_T^e(i, j) < \varepsilon \\ 1, & (R_T^e(i, j) < \varepsilon) \& (R_C^e(i, j) < \varepsilon) \\ R_T^e(i, j) / R_C^e(i, j), & \text{others} \end{cases}$$

[0031]

($R_T^e(i, j)$)는 테스트 데이터의 리드 커버리지 데이터이고, ($R_C^e(i, j)$)는 컨트롤 데이터의 리드 커버리지 데이터이고, ε 는 10^{-3})을 이용하여 비율 데이터를 산출하는 단계 및 보정계수(ω)를 곱하여 비율 데이터를 보정하는 단계를 포함하되, 보정계수(ω)는 컨트롤 데이터에 존재하는 모든 타깃 영역의 커버리지 값을 모두 더한 값을 테스트 데이터에 존재하는 모든 타깃 영역의 커버리지 값을 모두 더한 값으로 나눈 값일 수 있다.

[0033] 복수의 세그먼트로 분할하는 단계는 각 타깃 영역($e_i(1 \leq i \leq N_e)$)에 대한 비율 데이터($R_{T|C}^e(i, j)$, $b_f(i) \leq j \leq b_f(i)$)를 동일한 크기(b_s)를 갖는 $N_s(i) = [N_b(i)/b_s] - 1$ 개의 세그먼트로 분할하여, 각 세그먼트가 동일한 크기를 갖는 비율 데이터($R_{T|C}^{e,k}(j_k)$ ($1 \leq i \leq N_e$, $1 \leq k \leq n_s(i)$, $1 \leq j_k \leq b_s$))를 산출하되, 세그먼트 분할 시 $b_f(i) = [(\text{mod}(n_b(i), b_s) + b_s)/2] + 1$ 에서 타깃 영역의 첫 번째 세그먼트를 시작하고, 양끝의 남은 영역($R_{T|C}^e(i, j)$ ($1 \leq j \leq b_s(i) - 1$))과 $R_{T|C}^e(i, j)$ ($b_j(i) \leq j \leq N_b(i)$)의 데이터를 제외할 수 있다.

[0034] 표준화 평균값으로 조정하는 단계는 수학적식

$$mR_s^e(i, k) = \sum_{j=b_f(i)+(k-1)*b_s}^{b_f(i)+kb_s} R_{T|C}^e(i, j) / b_s, \quad (1 \leq k \leq n_s(i))$$

[0035]

을 이용하여 각 세그먼트의 평균값을 산출하는 단계, 수학적식

[0036]

$$tR_s^e(i, k) = (mR_s^e(i, k) - M(mR_s^e)) / (SD(mR_s^e) / \sqrt{b_s}), \quad (1 \leq k \leq n_s(i))$$

[0037]

[0038] (여기서, $M(mR_s^e)$ 는 전체 세그먼트 $S_{i,k}(1 \leq i \leq N_e, 1 \leq k \leq n_s(i))$ 에 걸쳐져 있는 평균값의 평균이고, $SD(mR_s^e)$ 는 전체 세그먼트 $S_{i,k}(1 \leq i \leq N_e, 1 \leq k \leq n_s(i))$ 에 걸쳐져 있는 평균값의 표준 편차)를 이용하여 표준화 평균값을 산출하는 단계 및 산출한 표준화 평균값이 음수이면 표준화 평균값에서 표준화 평균값의 최소값을 차감하여 표준화 평균값을 보정하는 단계를 포함할 수 있다.

[0039] 유효 비율 데이터를 추출하는 단계에서는 표준화 평균값과 각 세그먼트의 비율 데이터의 표준 편차를 근거로 각 세그먼트 내 각 기본 포지션의 데이터가 정규 분포를 갖도록 변환하여 유효 비율 데이터를 추출할 수 있다.

[0040] 상기한 목적을 달성하기 위한 본 발명의 실시예에 따른 유전자 단위 반복 변이 검출 방법은 정규화 처리 장치에 의해, 테스트 데이터 및 컨트롤 데이터 또는 데이터 데이터를 포함하는 입력 데이터를 정규화 처리하는 단계 및 CNV 추출 장치에 의해, 입력 데이터를 정규화 처리하는 단계에서 정규화 처리된 입력 데이터를 근거로 유전자 단위 반복 변이를 추출하는 단계를 포함하고, 입력 데이터를 정규화 처리하는 단계는 입력 데이터에 대해 각 타깃 영역에서의 염기 단위 리드 커버리지를 산출하는 단계, 염기 단위 리드 커버리지를 근거로 세그먼트화하여 타깃 영역을 동일한 크기를 갖는 복수의 세그먼트로 분할하는 단계, 복수의 세그먼트들 각각의 리드 커버리지 데이터의 평균값을 표준화 평균값으로 조정하는 단계 및 표준화 평균값 및 각 세그먼트의 비율 데이터를 근거로 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환하는 단계를 포함할 수 있다.

발명의 효과

[0041] 본 발명에 의하면, 유전자 단위 반복 변이 검출 시스템 및 방법은 4단계의 정규화 변환 기법을 적용하여 엑솜 시퀀싱 데이터를 변환함으로써, 데이터의 잡음 특성을 최소화할 수 있는 효과가 있다.

[0042] 또한, 유전자 단위 반복 변이 검출 시스템 및 방법은 스케일 스페이스 필터링을 통해 다양한 모양과 크기의 CNV 영역을 정확히 추출할 수 있는 효과가 있다.

도면의 간단한 설명

[0043] 도 1은 본 발명의 실시예에 따른 유전자 단위 반복 변이 검출 시스템을 설명하기 위한 도면.

도 2는 도 1의 정규화 처리 장치를 설명하기 위한 도면.

도 3 및 도 4는 도 1의 CNV 추출 장치를 설명하기 위한 도면.

도 5는 본 발명의 실시예에 따른 유전자 단위 반복 변이 검출 방법을 설명하기 위한 흐름도.

도 6은 도 5의 정규화 처리 단계를 설명하기 위한 흐름도.

도 7은 도 5의 CNV 검출 단계를 설명하기 위한 흐름도.

발명을 실시하기 위한 구체적인 내용

[0044] 이하, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 본 발명의 기술적 사상을 용이하게 실시할 수 있을 정도로 상세히 설명하기 위하여, 본 발명의 가장 바람직한 실시예를 첨부 도면을 참조하여 설명하기로 한다. 우선 각 도면의 구성요소들에 참조부호를 부가함에 있어서, 동일한 구성요소들에 대해서는 비록 다른 도면 상에 표시되더라도 가능한 한 동일한 부호를 가지도록 하고 있음에 유의해야 한다. 또한, 본 발명을 설명함에 있어, 관련된 공지 구성 또는 기능에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명은 생략한다.

[0045] 본 발명의 실시예에 따른 유전자 단위 반복 변이 검출 시스템 및 방법은 엑솜 시퀀싱(Exome Sequencing) 데이터를 정규화 변환하여 데이터의 잡음 특성을 최소화하고, 스케일 스페이스 필터링을 통해 다양한 모양과 크기의 유전자 단위 반복 변위(CNV; Copy Number Variation) 영역을 정확하게 추출한다.

[0046] 도 1을 참조하면, 본 발명의 실시예에 따른 유전자 단위 반복 변이 검출 시스템(100)은 정규화 처리 장치(120) 및 CNV 추출 장치(140)를 포함하여 구성된다.

[0047] 정규화 처리 장치(120)는 GC content 바이어스, 라이브러리 사이즈 영향, 엑솜 경계 영역에서의 커버리지 바이어스 등으로 인한 다양한 잡음 영향을 최소화하기 위해서 입력 데이터를 정규화한다.

[0048] 이를 위해, 정규화 처리 장치(120)는 테스트(Test) 데이터 및 컨트롤(Control) 데이터 중 적어도 하나를 입력 데이터로 입력받는다.

[0049] 이때, 정규화 처리 장치(120)는 테스트 데이터만 입력되면 테스트 데이터의 리드 커버리지 데이터만을 이용하여 정규화 처리를 수행한다. 정규화 처리 장치(120)는 테스트 데이터와 컨트롤 데이터가 입력되면 테스트 데이터 및 컨트롤 데이터의 커버리지 데이터로부터 얻은 비율(ratio) 커버리지 데이터를 이용하여 정규화 처리를 수행한다. 이하에서는, 입력 데이터로서 테스트 데이터 및 컨트롤 데이터가 모두 사용하는 경우를 예로 들어 설명한다.

[0050] 정규화 처리 장치(120)는 입력 데이터에 대해 각 타깃 영역에서의 염기 단위(Base-Level) 리드 커버리지 산출하고, 산출한 염기 단위 리드 커버리지를 세그먼테이션하고, 각 세그먼트에서의 커버리지 데이터의 평균값 조정, 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환한다.

[0051] 도 2를 참조하면, 정규화 처리 장치(120)는 커버리지 산출모듈(122), 세그먼테이션 처리모듈(124), 평균값 조정 모듈(126), 정규화 변환모듈(128)을 포함하여 구성된다.

[0052] 커버리지 산출모듈(122)은 입력 데이터를 이용하여 타깃 단위로 커버리지를 산출한다. 즉, 커버리지 산출모듈(122)은 엑솜 시퀀싱 데이터인 테스트 데이터와 컨트롤 데이터의 정렬 결과를 이용하여 커버리지 데이터(R_T , R_C)를 산출한다.

[0053] 커버리지 산출모듈(122)은 기산출한 커버리지 데이터(R_T , R_C)로부터 각 타깃 영역($e_i(1 \leq i \leq N_e)$)의 모든 포지션($j(1 \leq j \leq N_b(i))$)에서의 리드 커버리지 데이터($R_T^e(i, j)$ 와 $R_C^e(i, j)$)를 추출한다. 여기서, N_e 는 전체 타깃 영역

의 개수를 의미하고, $N_b(i)$ 는 i 번째 타깃 영역의 길이를 의미한다.

[0054] 커버리지 산출모듈(122)은 추출한 리드 커버리지 데이터($R_T^e(i, j)$, $R_C^e(i, j)$)의 비율(ratio) 데이터($R_{T|C}^e(i, j)$)를 산출한다.

[0055] 이때, 커버리지 산출모듈(122)은 하기 수학식 1을 이용하여 리드 커버리지 데이터($R_T^e(i, j)$, $R_C^e(i, j)$)의 비율 데이터를 산출한다.

수학식 1

$$R_{T|C}^e(i, j) = \begin{cases} 0, & R_T^e(i, j) < \varepsilon \\ 1, & (R_T^e(i, j) < \varepsilon) \& (R_C^e(i, j) < \varepsilon) \\ R_T^e(i, j) / R_C^e(i, j), & \text{others} \end{cases}$$

[0056]

여기서, ε 는 10^{-3} 이다.

[0057]

[0058] 커버리지 산출모듈(122)은 두 리드 커버리지 데이터의 라이브러리 사이즈(library size) 보정을 위해 기산출한 비율 데이터를 보정한다. 즉, 하기 수학식 2와 같이, 커버리지 산출모듈(122)은 비율 데이터($R_{T|C}^e(i, j)$)에 보정 계수(ω)를 곱하여 비율 데이터를 보정한다

수학식 2

$$\omega R_{T|C}^e(i, j) = \omega \times R_{T|C}^e(i, j)$$

[0059]

[0060] 이때, 보정계수 $\omega = mR_C^e / mR_T^e$ 이고, R_C^e 는 컨트롤 데이터에 존재하는 모든 타깃 영역의 커버리지 값을 모두 더한 값($mR_C^e = \sum_i^{N_e} \sum_j^{N_b(i)} R_C^e(i, j)$)이고, R_T^e 는 테스트 데이터에 존재하는 모든 타깃 영역의 커버리지 값을 모두 더한 값($mR_T^e = \sum_i^{N_e} \sum_j^{N_b(i)} R_T^e(i, j)$)이다.

[0061]

세그멘테이션 처리모듈(124)은 타깃 영역에 대해 세그멘테이션을 수행한다. 즉, 실제 대부분의 엑솜 영역 경계 부분은 엑솜 영역 중앙 부분에 비해 낮은 커버리지를 갖는다. 이에, 세그멘테이션 처리모듈(124)은 각 타깃 엑솜 영역($e_i(1 \leq i \leq N_e)$)에 대하여 세그멘테이션을 수행한다.

[0062]

세그멘테이션 처리모듈(124)은 각 타깃 영역에 대한 비율 데이터를 동일한 크기를 갖는 복수의 세그먼트로 분할한다. 즉, 세그멘테이션 처리모듈(124)은 각 타깃 영역($e_i(1 \leq i \leq N_e)$)에 대한 비율 데이터($R_{T|C}^e(i, j)$, ($b_f(i) \leq j \leq b_l(i)$))를 동일한 크기(b_s)를 갖는 $N_s(i) = [N_b(i)/b_s] - 1$ 개의 세그먼트로 분할하여, 각 세그먼트가 동일한 크기를 갖는 비율 데이터($R_{T|C}^{S_i, k}(j_k)$ ($1 \leq i \leq N_e$, $1 \leq k \leq N_s(i)$, $1 \leq j_k \leq b_s$))를 산출한다.

[0063]

이때, 세그멘테이션 처리모듈(124)은 세그먼트 분할 시 $b_f(i) = [(\text{mod}(n_b(i), b_s) + b_s) / 2] + 1$ 에서 타깃 엑솜 영역의 첫 번째 세그먼트를 시작하고, 양끝의 남은 영역($R_{T|C}^e(i, j)$ ($1 \leq j \leq b_s(i) - 1$)와 $R_{T|C}^e(i, j)$ ($b_j(i) \leq j \leq N_b(i)$))의 데이터는 제외한다.

[0064]

평균값 조정모듈(126)은 세그멘테이션 처리모듈(124)에서 분할된 각 세그먼트의 평균값을 T-Score를 이용하여 전체 타깃 엑솜에 대해 표준화된 평균값(이하, 표준화 평균값)으로 조정한다.

[0065]

이때 평균값 조정모듈(126)은 하기의 수학식 3을 이용하여 각 세그먼트의 평균값을 산출한다.

수학식 3

$$mR_s^e(i, k) = \sum_{j=b_f(i)+(k-1)*b_s}^{b_f(i)+kb_s} R_{T|C}^e(i, j)/b_s$$

$$(1 \leq k \leq n_s(i))$$

평균값 조정모듈(126)은 하기의 수학식 4를 이용하여 각 세그먼트의 표준화 평균값을 산출한다.

수학식 4

$$tR_s^e(i, k) = (mR_s^e(i, k) - M(mR_s^e)) / (SD(mR_s^e) / \sqrt{b_s})$$

$$(1 \leq k \leq n_s(i))$$

여기서, $M(mR_s^e)$ 는 전체 세그먼트 $S_{i,k}(1 \leq i \leq N_e, 1 \leq k \leq n_s(i))$ 에 걸쳐져 있는 평균값의 평균이고, $SD(mR_s^e)$ 는 전체 세그먼트 $S_{i,k}(1 \leq i \leq N_e, 1 \leq k \leq n_s(i))$ 에 걸쳐져 있는 평균값의 표준 편차를 의미한다.

평균값 조정모듈(126)은 산출한 표준화 평균값이 음수인 경우 양수값을 갖도록 보정한다. 즉, 평균값 조정모듈(126)은 산출한 표준화 평균값이 음수이면, 표준화 평균값에서 표준화 평균값의 최소값을 차감하여 표준화 평균값을 보정한다.

이때, 표준화 평균값의 최소값은 $\min(tR_s^e(i, j))(1 \leq i \leq N_e, 1 \leq k \leq n_s(i))$ 을 의미한다.

정규화 변환모듈(128)은 표준화 평균값과 각 세그먼트의 비율 데이터를 근거로 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환한다.

즉, 정규화 변환모듈(128)은 표준화 평균값($tR_s^e(i, k)$)과 각 세그먼트의 비율 데이터 ($(R_{T|C}^{S_i, k}(j_k))(1 \leq i \leq N_e, 1 \leq k \leq n_s(i), 1 \leq j_k \leq b_s)$)의 표준 편차를 기반으로 각 세그먼트 $S_{i,k}(1 \leq i \leq N_e, 1 \leq k \leq n_s(i))$ 내의 각 기본 포지션 ($(j(b_f(i)+(k-1)*b_s) \leq j \leq b_f(i)+k*b_s)$)의 데이터가 정규 분포를 갖도록 변환한다.

이를 통해, 정규화 변환모듈(128)은 유효 비율 데이터($^eR_{T|C}^{S_i, k}(j)$)를 추출한다.

CNV 추출 장치(140)는 정규화 처리 장치(120)에서 추출한 유효 비율 데이터를 근거로 CNV 영역을 산출한다. 이때, CNV 추출 장치(140)는 스케일 스페이스 필터링 기법을 이용하여 CNV 영역을 산출한다. 여기서, 스케일 스페이스 필터링 기법은 입력 데이터를 다양한 해상도(multi-resolution) 데이터로 변환한 스케일 스페이스 이미지와 입력 데이터로부터 변곡점(inflexion point) 정보를 추출하여 얻어지는 지문도(finger print map)를 활용하여 신호에 포함된 이벤트를 추출하는 신호 처리 방식이다.

CNV 추출 장치(140)는 스케일 스페이스 필터링 기법을 이용하여 커버리지 데이터에 포함된 CNV 모양을 이벤트로 추출한다. 즉, CNV 추출 장치(140)는 연속 커버리지 데이터의 값 변화를 추적하여, 커버리지 값이 평균보다 낮거나 높은값이 연속되는 모양을 갖는 영역을 테스트 데이터의 서열상에 결손(loss)이나 삽입(gain)이 있는 경우를 나타낸다고 해석하고, 이 영역들을 CNV 영역으로 산출한다.

도 3은 스케일 스페이스 필터링 기법을 이용하여 이벤트 영역을 추출하는 일례로, 하단은 주어진 신호($\sigma = \sigma_0$ 에 서의 가장 아래 부분 신호)에 대한 스케일 스페이스 이미지를 나타내며, 상단은 스케일 스페이스 이미지의 지문도(Finger Print Map)를 나타낸다. 여기서, σ 는 해상도 조절을 위한 스케일 파라미터를 의미한다.

또한, 수직 점선은 $\sigma (= \sigma_0)$ 에서의 스케일 스페이스 이미지의 변곡점을 이용한 지문도 작성 과정을 나타내며, 수

평 화살표 구간은 $\sigma (= \sigma_k)$ 에서 이웃하는 두 변곡점으로 이루어지는 이벤트 정보의 추출 과정을 나타낸다.

[0081] 이를 위해, 도 4를 참조하면, CNV 추출 장치(140)는 계층 분리모듈(142), 핑거 프린트 맵 생성모듈(144), 기준 선 조정모듈(146), CNV 추출모듈(148)을 포함하여 구성된다.

[0082] 계층 분리모듈(142)은 정규화 처리 장치(120)에서 추출한 유효 비율 데이터($eR_{TIC}^{S,K}(j)$)에 가우시안 컨볼루션(Gaussian Convolution)을 적용시켜 복수의 스케일 스페이스 이미지(scale-space image) 계층으로 분리한다. 이때, 가우시안 컨볼루션은 하기 수학식 5와 같다. 이후 표기법의 간략화를 위하여 유효 비율 데이터($eR_{TIC}^{S,K}(j)$)를 $c[i]$ 로 표기하여 나타낸다.

수학식 5

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

[0083]

[0084] 여기서 $c[i, k]$ 는 스케일 스페이스 이미지이고, $g[j, \sigma_k]$ 는 가우시안 커널이고, $k(0 \leq k \leq L-1)$ 는 총 L개로 분리되는 스케일 스페이스 이미지 계층에 대한 어느 하나의 계층의 인덱스를 의미한다. m 은 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 것 일수 있다.

[0085] 다만, 시간 영역에서의 컨볼루션은 주파수 영역에서의 곱과 같으므로, 산출의 복잡성을 줄이기 위하여 이산 푸리에 변환(Discrete Fourier Transform)을 적용함으로써 각각의 스케일 스페이스 이미지 계층 $c[i, k]$ 를 구할 수 있다.

[0086] 핑거 프린트 맵 생성모듈(144)은 계층 분리모듈(142)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 수학식 6과 같이 근사화시킬 수 있다. 또한, 2차 도함수의 영교차 신호 $z[i, k]$ 는 수학식 7과 같이 정의할 수 있다.

수학식 6

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

[0087]

수학식 7

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

[0088]

[0089] 즉, $z[i, k]$ 는 영교차점 i 를 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, i 가 영교차점이 아니라면 $z[i, k]$ 는 0의 값을 갖는다.

[0090] 핑거 프린트 맵 생성모듈(144)은 복수의 계층별로 획득된 각각의 영교차점을 표시함으로써 하나의 핑거 프린트 맵(finger print map)을 생성할 수 있다. 즉, $z[i, k]$ 가 +1 또는 -1의 값을 갖는 경우, 영교차점이 핑거 프린트 맵에 표시된다. 이때, $z[i, k]$ 가 0의 값을 갖는 경우에는 핑거 프린트 맵에 아무런 표시가 되지 않는다.

[0091] 기준선 조정모듈(146)은 복수의 계층별로 기설정된 범위 내에 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 $\delta(k)$ 라고 한다면, 기설정된 범위는 $m(k) \pm w\delta(k)$ 로 놓을 수 있다. $\delta(k)$ 의 계수로 포함된 w 는 가중치(weight)로서 임의의 값으로 설정할 수 있고, $m(k)$ 및 $\delta(k)$ 의 값은 각각 수학식 8 및 수학식 9로부터 얻을 수 있다.

수학식 8

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

[0092]

수학식 9

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

[0093]

[0094] 본 발명의 일 실시예로서, 가중치 w 는 2로 설정하겠다. 기준선 조정모듈(146)은 복수의 계층별로 $m(k) \pm 2\delta(k)$ 을 벗어난 값들을 제외하고, $m(k) \pm 2\delta(k)$ 에 포함된 값들에 대한 평균값과 표준 편차값을 다시 구할 수 있다. 이때의 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.

[0095] 기준선 조정모듈(146)은 기설정된 범위 내에 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.

[0096] 여기서, $\delta^*(k)$ 의 계수로 포함된 d 는 가중치로서, 경우에 따라 다양한 값이 될 수 있다.

[0097] 나아가, 기준선 조정모듈(146)은 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 따라서는, 기준선 조정모듈(146)은 영교차점이 2개 이상 있는 계층에 대해서만 기준선을 획득할 수도 있다.

[0098] CNV 추출모듈(148)은 기준선 조정모듈(146)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 추출모듈(148)은 영교차점을 적어도 2개 이상 갖는 스케일 스페이스 이미지 계층에 한하여, d 값에 3을 적용한 기준선 $m^*(k) \pm 3\delta^*(k)$ 을 이용하여 CNV를 검출하기 위한 영역을 결정할 수 있다. 이때의 CNV를 검출하기 위한 영역은 계층별로 복수개가 될 수도 있고, 하나도 없을 수도 있다. CNV를 검출하기 위한 k 계층의 m 번째 영역은 $[l_{m,k}, u_{m,k}]$ 로 놓겠다. 여기서, $l_{m,k}$ 및 $u_{m,k}$ 는 팽거 프린트 맵에 표시된 영교차 신호 $z[i, k]$ 의 i 의 범위로서, $l_{m,k} \leq i \leq u_{m,k}$ 로 나타낼 수 있다. CNV를 검출하기 위한 영역을 결정하기 위하여 기준선의 d 값에 3을 적용한 것은 본 발명의 일 실시예로서, d 값은 3에 한정되지 않는다.

[0099] 단, CNV를 검출하기 위한 영역은 이하 3가지 조건에 기초하여 결정될 수 있다.

[0100] 첫째로, k 계층보다 높은, 즉, $k+1$ 내지 $L-1$ 계층에서 CNV를 검출하기 위한 영역으로 기지정된 영역에 대응하는 i 는 $[l_{m,k}, u_{m,k}]$ 에 포함되지 않아야 한다.

[0101] 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 i 에 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 -1이고, 둘 사이의 값들은 모두 0이어야 한다.

[0102] 같은 표현으로, 팽거 프린트 맵에서 $i=l_{m,k}$ 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.

[0103] 마지막으로, $[l_{m,k}, u_{m,k}]$ 에 포함된 스케일 스페이스 이미지의 평균값이 위에서 언급한 주어진 기준선 $m*(k) \pm 3\delta*(k)$ 를 벗어나야 한다. 여기서 $[l_{m,k}, u_{m,k}]$ 에 포함된 스케일 스페이스 이미지의 평균값은 수학식 10과 같이 표현될 수 있다.

수학식 10

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

[0104]

[0105] 도 5를 참조하면, 본 발명의 실시예에 따른 유전자 단위 반복 변이 검출 방법은 정규화 처리 단계(S100) 및 CNV 검출 단계(S200)를 포함한다.

[0106] 정규화 처리 단계(S100)에서는 테스트(Test) 데이터 및 컨트롤(Control) 데이터 중 적어도 하나를 입력 데이터로 입력받는다. 이때, 정규화 처리 장치(120)는 테스트 데이터만 입력되면 테스트 데이터의 리드 커버리지 데이터만을 이용하여 정규화 처리를 수행한다.

[0107] 정규화 처리 단계(S100)는 테스트 데이터와 컨트롤 데이터가 입력되면 테스트 데이터 및 컨트롤 데이터의 커버리지 데이터로부터 얻은 비율 (ratio) 커버리지 데이터를 이용하여 정규화 처리를 수행한다.

[0108] 정규화 처리 단계(S100)는 입력 데이터에 대해 각 타깃 영역에서의 염기 단위(Base-Level) 리드 커버리지 계산하고, 계산한 염기 단위 리드 커버리지를 세그먼테이션하고, 각 세그먼트에서의 커버리지 데이터의 평균값 조정, 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환한다.

[0109] 도 6을 참조하면, 정규화 처리 단계(S100)는 리드 커버리지 산출 단계(S120), 세그먼테이션 단계(S140), 각 세그먼트에서의 커버리지 데이터 평균값 조정 단계(S160) 및 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환하는 단계(S180)를 포함한다.

[0110] 리드 커버리지 산출 단계(S120)에서는 입력 데이터를 이용하여 타깃 단위로 커버리지를 계산한다. 즉, 리드 커버리지 산출 단계(S120)에서는 엑솔 시퀀싱 데이터인 테스트 데이터와 컨트롤 데이터의 정렬 결과를 이용하여 커버리지 데이터(R_T , R_C)를 산출한다.

[0111] 리드 커버리지 산출 단계(S120)에서는 산출한 커버리지 데이터로부터 각 타깃 영역의 모든 포지션에서의 리드 커버리지 데이터를 추출하고, 추출한 리드 커버리지 데이터의 비율 데이터를 산출한다. 여기서, 비율 데이터를 테스트 데이터 및 컨트롤 데이터의 비율을 의미한다.

[0112] 리드 커버리지 산출 단계(S120)에서는 리드 커버리지 데이터의 라이브러리 사이즈(library size) 보정을 위해 산출한 비율 데이터를 보정한다. 이때, 리드 커버리지 산출 단계(S120)에서는 비율 데이터에 보정계수를 곱하여 비율 데이터를 보정한다

[0113] 세그먼테이션 단계(S140)에서는 타깃 영역에 대해 세그먼테이션을 수행한다. 즉, 실제 대부분의 엑솔 영역 경계 부분은 엑솔 영역 중앙 부분에 비해 낮은 커버리지를 갖기 때문에, 세그먼테이션 단계(S140)에서는 각 타깃 엑솔 영역에 대하여 세그먼테이션을 수행한다.

[0114] 세그먼테이션 단계(S140)에서는 각 타깃 영역에 대한 비율 데이터를 동일한 크기를 갖는 복수의 세그먼트로 분할한다. 즉, 세그먼테이션 단계(S140)에서는 각 타깃 영역에 대한 비율 데이터를 동일한 크기를 갖는 복수개의 세그먼트로 분할하여, 각 세그먼트가 동일한 크기를 갖는 비율 데이터를 산출한다. 이때, 세그먼테이션 단계(S140)에서는 세그먼트 분할 시 타깃 엑솔 영역의 양끝에 남는 영역들의 데이터를 제외한다.

[0115] 평균값 조정 단계(S160)에서는 세그먼테이션 단계(S140)에서 분할된 각 세그먼트의 평균값을 T-Score를 이용하여 전체 타깃 엑솔에 대해 표준화된 평균값(이하, 표준화 평균값)으로 조정한다. 이때, 평균값 조정 단계(S160)에서는 표준화 평균값이 음수인 경우 양수값을 갖도록 보정한다. 즉, 평균값 조정 단계(S160)에서는 표준화 평균값이 음수이면, 표준화 평균값에서 표준화 평균값의 최소값을 차감하여 표준화 평균값이 양수값을 갖도록 보정한다.

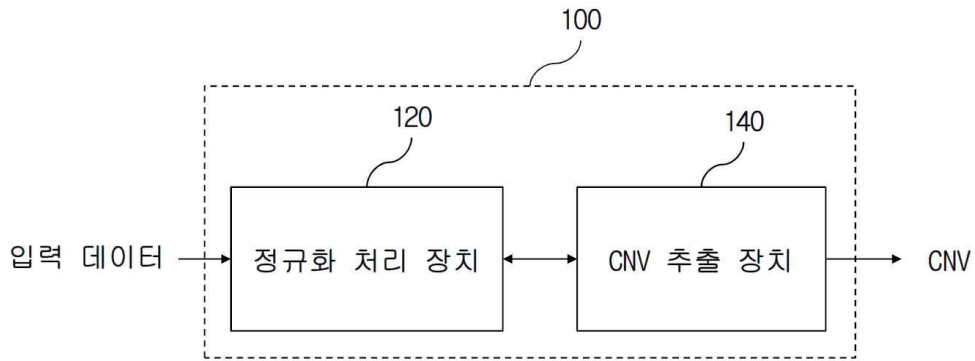
- [0116] 변환하는 단계(S180)에서는 평균값 조정 단계(S160)에서 조정된 표준화 평균값과 각 세그먼트의 비율 데이터를 근거로 각 세그먼트의 커버리지 데이터를 정규화 분포를 갖도록 변환한다. 즉, 변환하는 단계(S180)에서는 표준화 평균값과 각 세그먼트의 비율 데이터의 표준 편차를 기반으로 각 세그먼트 내의 각 기본 포지션 데이터가 정규 분포를 갖도록 변환하여 유효 비율 데이터를 추출한다.
- [0117] CNV 검출 단계(S200)는 정규화 처리 단계(S100)에서 추출한 유효 비율 데이터를 근거로 CNV 영역을 산출한다. 이때, CNV 검출 단계(S200)는 스케일 스페이스 필터링 기법을 이용하여 커버리지 데이터에 포함된 CNV 모양을 이벤트로 추출한다. 즉, CNV 검출 단계(S200)는 연속 커버리지 데이터의 값 변화를 추적하여, 커버리지 값이 평균보다 낮거나 높은값이 연속되는 모양을 갖는 영역을 테스트 데이터의 서열상에 결손(loss)이나 삽입(gain)이 있는 경우를 나타낸다고 해석하고, 이 영역들을 CNV 영역으로 산출한다.
- [0118] 도 7을 참조하면, CNV 검출 단계(S200)는 계층 분리 단계(S220), 핑거 프린트 맵 생성 단계(S240), 기준선 조정 단계(S260) 및 CNV 추출 단계(S280)를 포함한다.
- [0119] 계층 분리 단계(S220)에서는 정규화 처리 단계(S100)에서 추출한 유효 비율 데이터에 가우시안 컨볼루션(Gaussian Convolution)을 적용시켜 복수의 스케일 스페이스 이미지(scale-space image) 계층으로 분리한다.
- [0120] 핑거 프린트 맵 생성 단계(S240)에서는 계층 분리 단계(S220)에서 분리된 복수의 계층별로 2차 도함수의 영교차점을 획득할 수 있다. 핑거 프린트 맵 생성 단계(S240)에서는 복수의 계층별로 획득된 각각의 영교차점을 표시함으로써 하나의 핑거 프린트 맵(finger print map)을 생성할 수 있다. 즉, 영교차 신호가 +1 또는 -1의 값을 갖는 경우, 영교차점이 핑거 프린트 맵에 표시된다. 이때, 영교차 신호가 0의 값을 갖는 경우에는 핑거 프린트 맵에 아무런 표시가 되지 않는다.
- [0121] 기준선 조정 단계(S260)에서는 복수의 계층별로 기설정된 범위 내에 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.
- [0122] 기준선 조정 단계(S260)에서는 기설정된 범위 내에 포함된 스케일 스페이스 이미지의 값들에 대한 유효 평균값 및 유효 표준 편차값을 이용하여 복수의 계층 각각의 기준선을 획득한다.
- [0123] 기준선 조정 단계(S260)에서는 복수의 계층으로 분리된 스케일 스페이스 이미지 계층 모두에 대하여 기준선을 획득할 수도 있지만, 복수의 계층 중에서, 적어도 2개 이상의 0이 아닌 영교차 신호값을 갖는 계층에 대해서만 기준선을 획득할 수도 있다. 다시 말해, 경우에 따라서, 기준선 조정 단계(S260)에서는 영교차점이 2개 이상 있는 계층에 대해서만 기준선을 획득할 수도 있다.
- [0124] CNV 추출 단계(S280)에서는 기준선 조정 단계(S260)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, CNV 추출 단계(S280)에서는 영교차점을 적어도 2개 이상 갖는 스케일 스페이스 이미지 계층에 한하여 CNV를 검출하기 위한 영역을 결정할 수 있다. 이때의 CNV를 검출하기 위한 영역은 계층별로 복수개가 될 수도 있고, 하나도 없을 수도 있다.
- [0125] 이상에서 본 발명에 따른 바람직한 실시예에 대해 설명하였으나, 다양한 형태로 변형이 가능하며, 본 기술분야에서 통상의 지식을 가진자라면 본 발명의 특허청구범위를 벗어남이 없이 다양한 변형예 및 수정예를 실시할 수 있을 것으로 이해된다.

부호의 설명

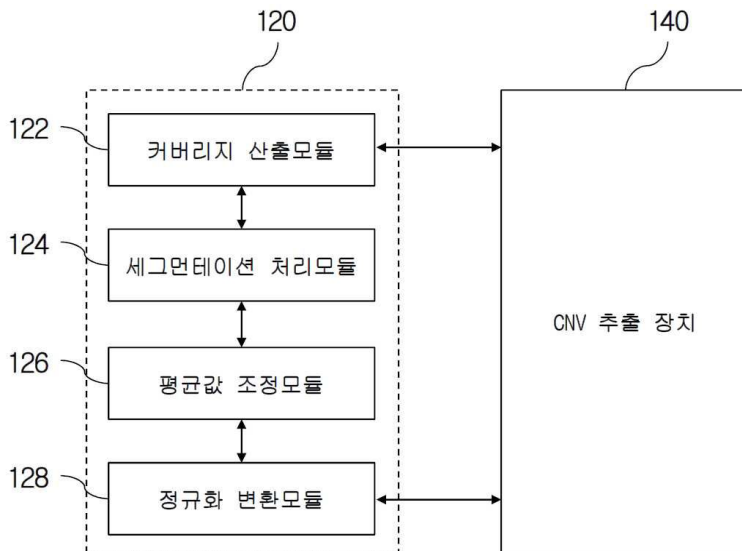
- [0126] 100: 유전자 단위 반복 변이 검출 시스템
 120: 정규화 처리 장치 122: 커버리지 산출모듈
 124: 세그먼테이션 처리모듈 126: 평균값 조정모듈
 128: 정규화 변환모듈 140: CNV 추출 장치
 142: 계층 분리모듈 144: 핑거 프린트 맵 생성모듈
 146: 기준선 조정모듈 148: CNV 추출모듈

도면

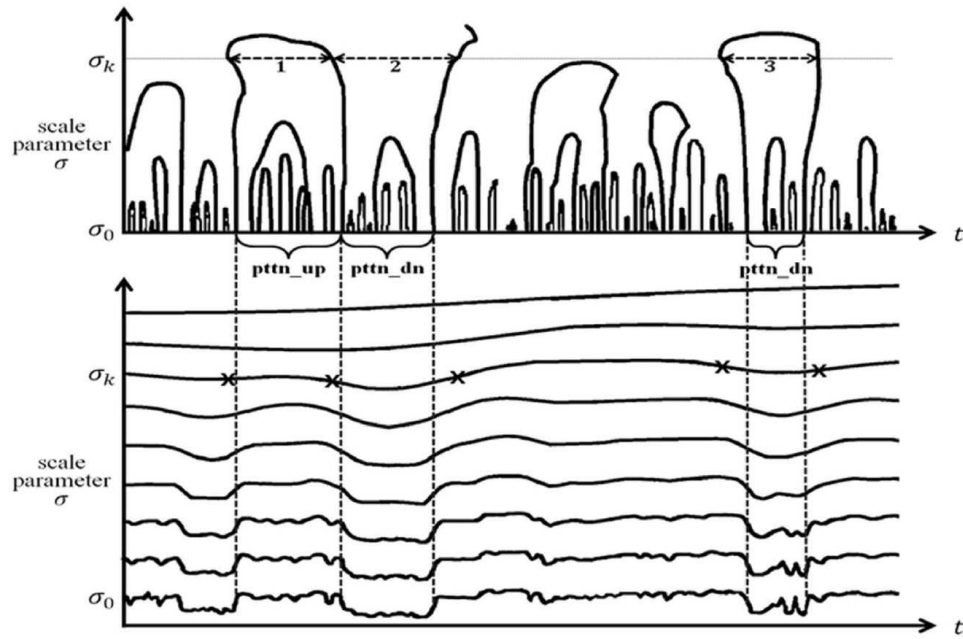
도면1



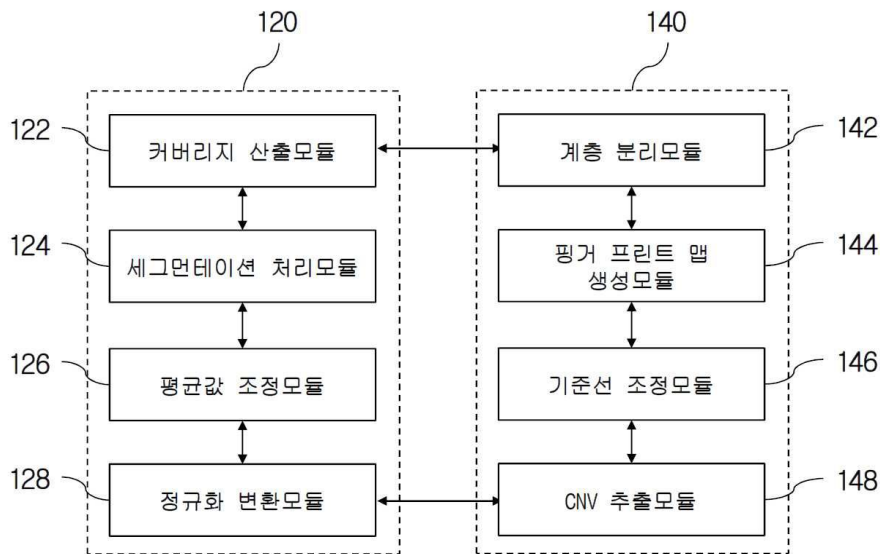
도면2



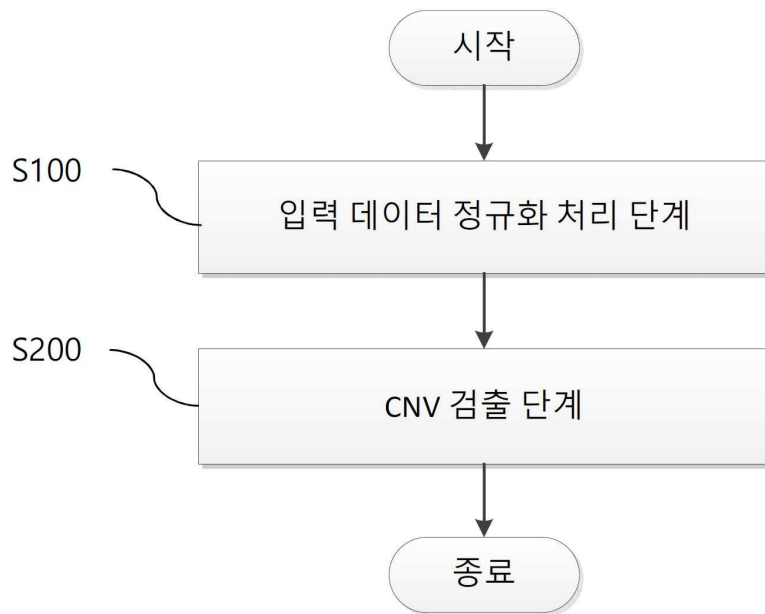
도면3



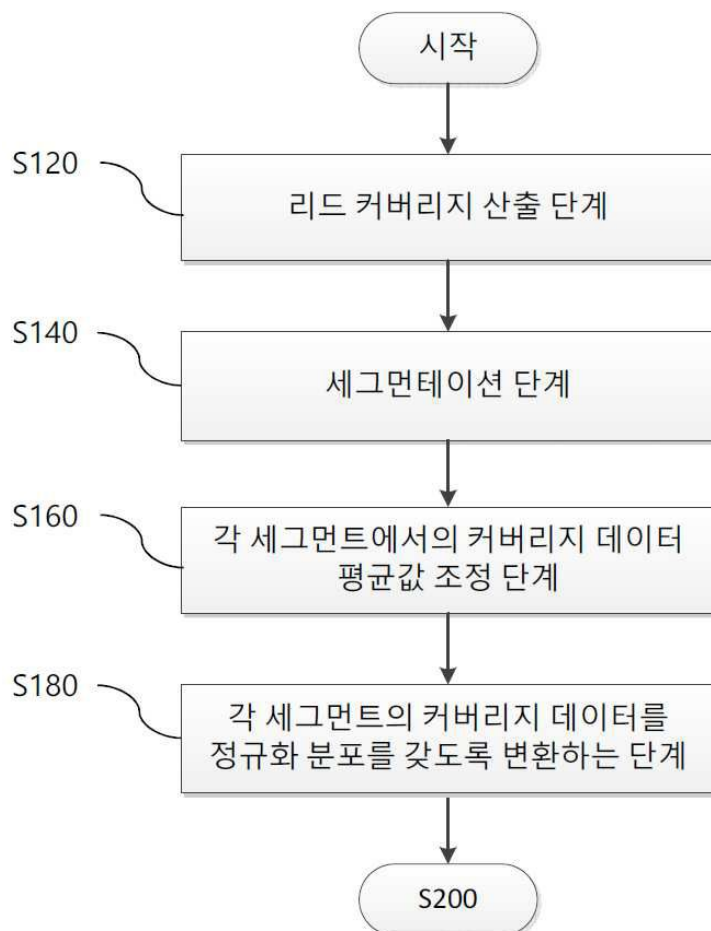
도면4



도면5



도면6



도면7

