명 세 서

청구범위

청구항 1

(1) 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 염색체상 위치(Chromosomal Position)별 표준 참조 염기 서열 데이터에 리드 매핑(Read Mapping)한 리드 카운트(Read Count)에 기반하여, 실험군 벡터 및 대조군 벡터를 생성하는 단계;

(2) 다음의 단계 (i) 내지 (v)를 포함하는 1차 바이어스 (Bias) 제거 단계:

(i) 실험군 벡터와 대조군 벡터를 결합하여 결합 행렬을 생성하는 단계;

(ii) 상기 생성된 결합 행렬을 복수개의 영역으로 나누는 단계;

(iii) 상기 복수개의 영역별로 NMF를 수행하는 단계;

(iv) 상기 NMF 수행 결과로부터 바이어스 요소를 선별하는 단계; 및

(v) 바이어스 제거 후 영역별 결합 행렬을 재결합하는 단계;

(3) 상기 1차 바이어스 제거 단계 이후에, 다음의 단계 (vi) 내지 (viii)를 포함하는 2차 바이어스 제거 단계:

(vi) 실험군 벡터와 대조군 벡터 간 비특이적 영역을 선별하여 무차별 영역으로 설정하는 단계;

(vii) 상기 무차별 영역에 기초하여 바이어스를 제거하는 단계; 및

(viii) 상기 바이어스가 제거된 실험군 벡터 및 대조군 벡터의 영역별 가중치를 계산하는 단계;

(4) 상기 바이어스가 제거된 영역별 TRR(Target Region Ratio) 벡터를 영역별로 취합하여 출력하는 단계; 및

(5) 상기 TRR 벡터를 이용하여, TRR벡터의 값이 특정 임계값보다 높은 영역을 타겟 염기 서열에서의 체세포 복제수 변이 영역으로 확인하는 단계

를 포함하고,

상기 단계 (1)의 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터는 차세대 서열분석(Next Generation Sequencing: NGS)에 의하여 얻어진 것이고,

상기 실험군 벡터 및 대조군 벡터의 무차별 영역은 동일한 영역인,

타겟 염기 서열에서의 체세포 복제수 변이 확인 방법.

청구항 2

제1항에 있어서, 단계 (1) 이전에, 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 수신하는 단계를 추가로 포함하는, 타겟 염기 서열에서의 체세포 복제수 변이 확인 방법.

청구항 3

제1항에 있어서, 상기 단계 (vi)의 실험군 벡터와 대조군 벡터 간 비특이 영역은 다음의 수학식 8에서 얻어진 W_{ratio}값이 특정 임계값 (Θ)보다 작은 위치인, 타겟 염기 서열에서의 체세포 복제수 변이 확인 방법:

[수학식 8]

$$W_{ratio} = \frac{W_{,2}}{W_{,1}} \times \frac{\sum W_{,1}}{\sum W_{,2}}$$

상기 식에서, W,1 벡터는 실험군 벡터 및 대조군 벡터의 공통 요소 벡터이고, W,2 벡터는 실험군 벡터 또는 대조군 벡터의 특이 요소 벡터임.

청구항 4

제1항에 있어서, 상기 단계 (vii)은 무차별 영역에 대응하는 요소 벡터의 값을 -1로 변환하여 해당 요소 벡터에 대응하는 바이어스를 제거하는 것인, 타겟 염기 서열에서의 체세포 복제수 변이 확인 방법.

청구항 5

제1항에 있어서, 상기 단계 (ii)는 수학식 2의 결합행렬을 영역별로 분리하여 수학식 3으로 전개되는 것인, 타겟 염기 서열에서의 체세포 복제수 변이 확인 방법:

[수학식 2]

$$V = [T, N]$$

[수학식 3]

$$V = \begin{bmatrix} T_{b1} & N_{b1} \\ T_{b2} & N_{b2} \\ T_{b3} & N_{b3} \\ \cdots & \cdots \\ T_{bl-1} & N_{bl-1} \\ T_{bl} & N_{bl} \end{bmatrix} = \begin{bmatrix} V_{b1} \\ V_{b2} \\ V_{b3} \\ \cdots \\ V_{bl-1} \\ V_{bl} \end{bmatrix}$$

상기 식에서, T는 실험군 벡터이고, N은 대조군 벡터이며, l은 영역의 개수이고, b는 영역(Boundary)를 의미함.

청구항 6

제1항에 있어서, 상기 단계 (iii)은 수학식 3의 행렬 Vb에 NMF를 적용하여 다음의 수학식 4로 전개되는 것인, 타겟 염기 서열에서의 체세포 복제수 변이 확인 방법:

[수학식 4]

$$V_b = [T_b, N_b] = WH$$

상기 식에서, V=n*p이고, W는 n*r, H는 r*p이며, n은 타겟 영역의 수, r은 NMF시 사용되는 rank, p는 각 영역별로 구분된 l개 영역 내 존재하는 타겟의 수를 의미함.

청구항 7

제1항에 있어서,

상기 리드 카운트는, 상기 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터에 위치한 적어도 하나의 타겟 영역에서 계산되는 것인, 타겟 염기서열에서의 체세포 복제수 변이 확인 방법.

청구항 8

제1항에 있어서,

상기 TRR(Target Region Ratio) 벡터는, 상기 실험 시료 염기 서열 데이터 또는 실험군 벡터와, 상기 대조 시료 염기 서열 데이터 또는 대조군 벡터에 위치한 적어도 하나의 타겟의 수에 기초하여 생성되는 것인, 타겟 염기 서열에서의 체세포 복제수 변이 확인 방법.

청구항 9

삭제

청구항 10

하드웨어에 결합되어 제1항 내지 제8항 중 어느 한 항의 타겟 염기 서열에서의 체세포 복제수 변이 확인 방법의

단계를 실행시키기 위하여 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램.

청구항 11

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 타겟 염기 서열 분석에서의 바이어스 제거 방법에 관한 것으로, 암 시료에 대한 서열 해독에서 발생하는 바이어스를 영역별로 제거하여 체세포 복제수 변이를 정확하게 판별할 수 있도록 정보를 제공하는 방법에 관한 것이다.

배 경 기 술

[0002] 암의 진단 및 치료를 위해서는 암에 특이적으로 존재하는 구조 변이를 발굴하는 것이 중요하다. 이에 따라, 다수의 암에서 발생되는 점 돌연변이와 같은 체세포 변이(somatic mutation), DNA 복제수(copy number) 및 염색체 재배열(rearrangement) 등을 밝혀 암의 원인 유전자를 규명하는 분야에 대한 연구가 활발하게 진행되고 있다. 이 중에서도 체세포 복제수 변이(somatic copy number variation)는 정상 세포에서는 존재하지 않는 유전자의 복제수의 변이를 의미하는 것으로서 암의 발병과 높은 연관성이 제기되고 있다.

[0003] 전장 유전체 서열 해독(whole genome sequencing, WGS)은 한 생명체가 가지는 전체 DNA 서열을 분석하는 방법이다. 따라서, 상기 방법을 통해 분석된 전체 DNA 서열 데이터는 한 개체의 모든 염기 서열을 포함하며, WGS 데이터를 이용하여 복제수 변이를 찾아내기 위해서는 주변 지역보다 복제수가 유의미하게 차이나는 지역을 찾아내야 한다. 그러나, WGS는 데이터 생산 비용이 높아서 WGS 데이터는 일반적으로 쉽게 생산할 수 없다.

[0004] 한편, WGS의 대안으로 타겟 엑솜 서열 해독(targeted exome sequencing)을 들 수 있다. 타겟 엑솜 서열 해독은 단백질을 번역하는 엑솜 영역 중에서도 관심 영역의 염기 서열만 포착(capture)하여 데이터를 생산하는 방식이다. 타겟 엑솜 서열 해독은 WGS 데이터 생산 방식에 비해서는 비용이 상대적으로 저렴하여 일반적으로 많이 사용되지만, 포착 효율(capture efficiency)이나 G-C 함량(구아닌-사이토신 함량) 등의 영향에 의해 바이어스(bias)가 발생하기 때문에 반드시 바이어스를 제거하여야 정확한 체세포 복제수 변이를 찾아낼 수 있다(Benjamini Y et al., Nucleic Acids Research DOI:10.1093/nar/gks001; 국내출원공개 제2014-0023847호; 국제공개 제WO2014/0044724호 참조).

[0005] 다만, 바이어스를 제거하기 위한 방법들이 일부 개발되어 왔으나, 기존의 방법들은 전체 타겟 영역을 대상으로 한 번에 바이어스를 제거하여 각 염색체 및 유전자 단위에 존재하는 소규모의 체세포 복제수 변이를 민감하게 검출할 수 없다는 한계가 있다. 또한, 기존의 특이값 분해(single value decomposition; SVD) 기반의 방법은 컷오프(cutoff)를 결정하는 기준이 모호하여, 샘플별 바이어스 제거시 재현성이 낮으며, 올바른 신호(true signal)까지 제거할 수 있다는 위험이 있다.

[0006] 따라서, 체세포 복제수 변이의 검출 민감도를 향상시키기 위해 새로운 바이어스 제거 방법이 요구된다.

발명의 내용

해결하려는 과제

[0007] 일 실시예는 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 표준 참조 염기 서열 데이터에 리드 매핑하여 계산된 리드 카운트에 기초하여 실험군 벡터 및 대조군 벡터를 생성하며, 생성된 실험군 벡터 및 대조군 벡터에 영역별로, 예컨대 비음수 행렬 인수분해법 (Non-negative Matrix Factorization; "NMF")를 통하여, 바이어스를 1차적으로 제거하고, 무차별 영역을 선정함으로써 노이즈를 2차적으로 제거함으로써, 체세포 복제수 변이 발굴의 민감도를 증가시킬 수 있는, 타겟 염기 서열 분석에서의 바이어스 제거 기술 및 이를 이용한 타겟 염기 서열 분석 기술을 제공한다.

[0008] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0009] 일 실시예는, NMF(Non-negative Matrix Factorization)를 이용하는 타겟 염기 서열 분석에서의 바이어스 제거 방법을 제공한다.

[0010] 구체예에서, 상기 타겟 염기 서열 분석에서의 바이어스 제거 방법은,

[0011] (1) 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 염색체상 위치(Chromosomal Position)별 표준 참조 염기 서열 데이터에 리드 매핑(Read Mapping)한 리드 카운트(Read Count)에 기반하여, 실험군 벡터 및 대조군 벡터를 생성하는 단계;

[0012] (2) 상기 생성된 실험군 벡터 및 대조군 벡터를 결합한 결합 행렬을 생성하고, 상기 생성된 결합 행렬을 영역별로 나누어 바이어스(Bias)를 제거하는 단계 (1차 바이어스 제거 단계);

[0013] (3) 상기 바이어스가 제거된 결합 행렬을 재결합하는 단계; 및

[0014] (4) 상기 바이어스가 제거된 영역별 TRR(Target Region Ratio) 벡터를 영역별로 취합하여 출력하는 단계

[0015] 를 포함하는 것일 수 있다.

[0016] 상기 바이어스 제거 방법은 타겟 염기 서열 분석에서의 바이어스 제거 장치에서 실행되는 타겟 염기 서열 분석에서의 바이어스 제거 방법일 수 있다.

[0017] 상기 단계 (1)의 실험 시료 염기 서열 데이터 및 대조 시료 염기 서열 데이터는 각각 독립적으로 유전체 서열 분석기(Sequencer)에서 생성된 서열 데이터를 직접 또는 간접적으로 수신하거나, 이미 생성된 서열 데이터가 저장된 컴퓨터 판독 가능한 저장 매체를 통하여 수득(준비)할 수 있다. 따라서, 상기 타겟 염기 서열 분석에서의 바이어스 제거 방법은, 단계 (1) 이전에, 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 준비하는 단계를 추가로 포함할 수 있으며, 상기 실험 시료 염기 서열 데이터 및 대조 시료 염기 서열 데이터는, 각각 독립적으로, 유전체 서열 분석기(Sequencer)에서 생성된 서열 데이터를 직접 또는 간접적으로 수신하거나, 이미 생성된 서열 데이터가 저장된 컴퓨터 판독 가능한 저장 매체를 적용함으로써 준비할 수 있다.

[0018] 상기 제1 바이어스 제거 단계는 NMF(Non-negative Matrix Factorization)를 이용하여 수행되는 것일 수 있다.

[0019] 일 예에서, 상기 타겟 염기 서열 분석에서의 바이어스 제거 방법은, 상기 1차 바이어스 제거 단계 이후, 예컨대, 상기 단계 (3)과 (4) 사이에, 다음의 단계 (2차 바이어스 제거 단계)를 추가로 포함할 수 있다:

[0020] (a) 상기 실험군 벡터와 대조군 벡터 간 비특이 영역을 선별 후 무차별 영역으로 설정하여 바이어스를 제거하는 단계; 및

[0021] (b) 상기 설정된 무차별 영역으로 바이어스가 제거된 상기 실험군 벡터 및 대조군 벡터의 영역별 가중치를 계산하는 단계

[0022] 다른 예는 상기 바이어스 제거 방법을 포함하는 타겟 염기 서열 분석을 위한 컴퓨터 판독 방법을 제공한다.

[0023] 다른 예는 상기 바이어스 제거 방법의 단계를 실행시키기 위하여 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램을 제공한다.

[0024] 다른 예는 상기 바이어스 제거 방법의 단계를 실행하기 위한 시스템을 제공한다.

[0025] 다른 예는 상기 바이어스 제거 방법을 포함하는 타겟 염기 서열의 컴퓨터 판독 방법을 실행시키기 위하여 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램을 제공한다.

[0026] 다른 예는 상기 바이어스 제거 방법의 단계를 실행시키기 위한 컴퓨터에서 실행 가능한 프로그램(computer executable instruction)이 수록된 컴퓨터 판독 가능한 저장 매체 (또는 기록 매체)를 제공한다.

[0027] 다른 예는 상기 바이어스 제거 방법을 포함하는 적 염기 서열의 컴퓨터 판독 방법을 실행시키기 위한 컴퓨터에서 실행 가능한 프로그램(computer executable instruction)이 수록된 컴퓨터 판독 가능한 저장 매체 (또는 기록 매체)를 제공한다.

발명의 효과

[0028] 전술한 기술적 해결 방법에서 제공된 수단 중 어느 하나에 의하면, 타겟 서열 해독에서 영역별로 발생하는 특이

적인 바이어스 및 비특이적인 노이즈를 제거할 수 있고, 리드 카운트의 바이어스를 제거하여, 체세포 복제수 변이 발굴의 정확성을 향상시킬 수 있다.

도면의 간단한 설명

[0029] 도 1은 일 실시예에 따른 타겟 염기 서열 분석에서의 바이어스 제거 시스템을 설명하기 위한 구성도이다.

도 2는 일 실시예에 따른 바이어스 제거 방법이 수행되는 장치를 설명하기 위한 블록 구성도이다.

도 3은 일 실시예에 따른 타겟 염기 서열 분석에서의 바이어스 제거 방법을 설명하기 위한 블록 구성도이다.

도 4는 일 실시예에 따른 바이어스 제거 방법에서 실험 시료 염기 서열 데이터에 기반한 실험군 벡터를 생성하는 과정을 설명하기 위한 도면이다.

도 5는 일 실시예에 따른 바이어스 제거 방법에서 실험군 벡터와 대조군 벡터를 생성하는 과정을 설명하기 위한 도면이다.

도 6은 일 실시예에 따른 바이어스 제거 방법에서 영역별로 실험군 벡터와 대조군 벡터를 나누는 과정을 설명하기 위한 도면이다.

도 7은 일 실시예에 따른 바이어스 제거 방법에서 바이어스를 제거하기 전과 후의 타겟 영역수에 대한 TRR 벡터를 도시한 그래프이다.

도 8은 다양한 방법으로 바이어스를 제거한 후의 타겟 영역 수에 대한 TRR을 도시한 그래프이다.

도 9는 본 발명의 일 실시예에 따른 타겟 염기 서열 분석에서의 바이어스 제거 방법을 설명하기 위한 동작 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0030] 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.

[0031] 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 배제하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미하며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0032] 본 명세서에서, 다른 정의가 없는 한, 실험 시료는 시험하고자 하는 염기서열을 포함하는 시료로서, 암 시료(즉, 암세포로부터 추출된 유전체 (DNA 및/또는 RNA) 시료)일 수 있고, 대조 시료는 정상 시료 (즉, 정상 세포로부터 추출된 유전체 (DNA 및/또는 RNA) 시료)일 수 있다. 상기 실험 시료 및 대조 시료는 동물, 예컨대 인간을 포함하는 포유동물로부터 얻어진 (분리된) 세포, 조직, 또는 이들로부터 추출된 유전체(DNA 및/또는 RNA) 시료일 수 있다. 상기 유전체는 게놈 또는 염색체의 전부 또는 일부의 DNA 및/또는 RNA를 의미한다.

[0033] 본 명세서에서, 다른 정의가 없는 한, 타겟 염기 서열 분석은 체세포 복제수 변이를 확인하기 위한 것으로, 타겟 영역에서의 유전체 복제수 변이를 확인하기 위한 타겟 영역의 염기 서열 분석일 수 있다.

[0034] 본 명세서에서, 다른 정의가 없는 한, 타겟 영역 및 타겟 염기 서열은 게놈 또는 염색체의 전부 또는 일부 내의 분석하고자 하는 영역 (타겟 영역) 및 상기 영역의 염기 서열 (타겟 염기 서열)을 각각 의미한다. 상기 타겟 영역 및 타겟 염기 서열은 하나의 시료에 대하여 하나 이상 존재할 수 있다.

[0035] 본 명세서에서 수치 앞에 기재된 "약"은, 다른 정의가 없는 한, 기재된 수치의 10%, 5%, 또는 3%의 변동폭(증감분)을 포함하기 위하여 사용된 것일 수 있다.

[0036] 이하 첨부된 도면을 참고하여 본 발명을 상세히 설명하기로 한다.

[0037] 도 1은 본 발명의 일 실시예에 따른 타겟 염기 서열 분석에서의 바이어스 제거 시스템을 설명하기 위한 구성도

이다. 도 1을 참조하면, 바이어스 제거 시스템(1)은, 유전체 서열 분석기(100)와 타겟 염기 서열 분석에서의 바이어스 제거 장치(300)를 포함할 수 있다. 다만, 이러한 도 1의 실험 시료 바이어스 제거 시스템(1)은 본 발명의 일 실시예에 불과하므로 도 1을 통해 본 발명이 한정 해석되는 것은 아니다.

[0038] 도 1의 각 구성요소들은 네트워크(network, 200)를 통해 연결될 수 있다. 예를 들어, 도 1에 도시된 바와 같이, 네트워크(200)를 통하여 유전체 서열 분석기(100)와 타겟 염기 서열 분석에서의 바이어스 제거 장치(300)가 연결될 수 있다. 이때, 유전체 서열 분석기(100)에서 생성된 대조 시료 염기 서열 데이터 및/또는 실험 시료 염기 서열 데이터만을 타겟 염기 서열 분석에서의 바이어스 제거 장치(300)가 수신하면 되므로, 직접 또는 간접적인 연결을 모두 포함할 수 있다. 예컨대, 유전체 서열 분석기(100)와 타겟 염기 서열 분석에서의 바이어스 제거 장치(300)가 네트워크(200)을 통하여 직접 연결되거나, 웹하드와 같은 인터넷 상 저장 공간을 통하여 연결될 수 있다. 다른 예에서, 상기 유전제 해독기에서 생성된 대조 시료 염기 서열 데이터 및/또는 실험 시료 염기 서열 데이터는 컴퓨터 판독 가능한 저장 매체에 저장되어 바이어스 제거 장치에 적용될 수 있다.

[0039] 여기서, 네트워크(200)는 단말들 및 서버들과 같은 각각의 노드 상호 간에 정보 교환이 가능한 연결 구조를 의미하는 것으로, 이러한 네트워크(200)의 일 예는, WCDMA, 인터넷(Internet), LAN(Local Area Network), Wireless LAN(Wireless Local Area Network), WAN(Wide Area Network), PAN(Personal Area Network), ATM 방식을 활용한 E1 망, 3G, 4G, LTE, Wi-Fi 등이 포함되나 이에 한정되지는 않는다. 또한, 도 1에 개시된 유전체 서열 분석기(100)와 타겟 염기 서열 분석에서의 바이어스 제거 장치(300)는 도 1에 도시된 것들로 한정 해석되는 것은 아니다.

[0040] 유전체 서열 분석기(100)는 DNA 서열을 증폭시킨 후 형광 표식 등을 촬영 수단으로 촬영하고, 이미지 처리를 수행함으로써 DNA 유전 정보를 병렬 데이터화할 수 있는 모든 장치를 의미할 수 있다. 예컨대, 상기 유전체 서열 분석기는 차세대 서열분석(Next Generation Sequencing: NGS)과 같은 대규모 병렬형 서열분석(massively parallel sequencing) 기술의 수행이 가능한 장치일 수 있으나, 이에 제한되는 것은 아니다. 일 예에서, 상기 대규모 병렬 염기서열분석은, 예컨대 454 플랫폼(platform) (Margulies, 등, Nature (2005) 437:376~380), Illumina Genome Analyzer (또는 Solexa™ platform), Illumina HiSeq2000, HisSeq2500, MiSeq, NextSeq500, Life Tech Ion PGM, Ion Proton, Ion S5, Ion S5XL, 또는 SOLiD (Applied Biosystems) 또는 Helicos True Single Molecule DNA 서열분석 기술 (Harris, 등, Science (2008) 320:106~109), Pacific Biosciences의 단일 분자, 및/또는 실시간(SMRT™) 기술 등에 의하여 수행될 수 있다. 이 외에도 상업적으로 입수 가능한 서열분석 기기를 사용하여 폴리뉴클레오타이드 단편들의 서열정보를 수득할 수 있다.

[0041] 유전체 서열 분석기(100)는 유전자 변이, DNA 복제수(Copy Number) 및 염색체 재배열을 파악하는 분야에도 적용될 수 있으며, 이를 위하여 유전체 서열 분석기(100)는 하나의 DNA를 여러 번 읽을 수 있는데, 여기서 읽은 횟수를 리드 카운트(Read Count)라 정의하고, 리드 카운트는 깊이(Depth)라고도 정의될 수 있다.

[0042] 본 명세서에서, 리드(read)는 유전체 서열 분석기가 한번에 읽는 DNA 단편 길이를 의미하는 것으로, 약 10 내지 약 2000 bp, 약 10 내지 약 1000bp, 약 10 내지 약 500bp, 약 10 내지 약 300bp, 약 10 내지 약 200 bp, 약 25 내지 약 2000 bp, 약 25 내지 약 1000 bp, 약 25 내지 약 500bp, 약 25 내지 약 300bp, 약 25 내지 약 200bp, 약 25 내지 약 100bp, 약 50 내지 약 2000 bp, 약 50 내지 약 1000 bp, 약 50 내지 약 500bp, 약 50 내지 약 300bp, 약 50 내지 약 200bp, 약 50 내지 약 100bp, 약 100 내지 약 2000 bp, 약 100 내지 약 1000 bp, 약 100 내지 약 500bp, 약 100 내지 약 300bp, 약 100 내지 약 200bp, 약 150 내지 약 2000 bp, 약 150 내지 약 1000 bp, 약 150 내지 약 500bp, 또는 약 150 내지 약 300bp 길이를 갖는 것일 수 있다. 타겟 염기 서열 분석에서의 바이어스 제거 장치(300)는, 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 표준 참조 염기 서열 데이터에 리드 매핑(Read Mapping)하여 리드 카운트(Read Count)를 계산하고, 리드 카운트에 기반하여 실험 시료 염기 서열 데이터로부터 실험군 벡터를, 대조 시료 염기 서열 데이터로부터 대조군 벡터를 생성할 수 있다. 그리고 나서, 바이어스 제거 장치(300)는 1차적으로 바이어스를 제거하는데, 실험군 벡터와 대조군 벡터를 결합한 결합 행렬을 생성하고, 결합 행렬을 영역별로 나누어 NMF(Non-negative Matrix Factorization)를 통하여 바이어스를 제거할 수 있다.

[0043] 또한, 바이어스 제거 장치(300)는, 체세포 복제수 변이 발굴의 민감도 향상을 위하여, 2차적으로 바이어스 제거를 실시할 수 있는데, 실험군 벡터와 대조군 벡터 간 비특이 영역을 선별하여 무차별 영역으로 설정하고, 무차별 영역에 기초하여 바이어스를 제거할 수 있다. 이때, 타겟 염기 서열 분석에서의 바이어스 제거 장치(300)는, 네트워크(200)를 통하여 원격지의 서버나 단말에 접속할 수 있는 컴퓨터로 구현될 수 있다. 여기서, 컴퓨터는 예를 들어, 노트북, 데스크톱(Desktop), 랩톱(Laptop) 등을 포함할 수 있다.

[0044] 도 2는 일 실시예에 따른 바이어스 제거 방법이 수행되는 장치 (시스템)를 설명하기 위한 블록 구성도이고, 도 3은 일 실시예에 따른 타겟 염기 서열 분석에서의 바이어스 제거 방법을 설명하기 위한 블록 구성도이고, 도 4는 일 실시예에 따른 바이어스 제거 방법에서 실험 시료 염기 서열 데이터에 기반한 실험군 벡터를 생성하는 과정을 설명하기 위한 도면이고, 도 5는 일 실시예에 따른 바이어스 제거 방법에서 실험군 벡터와 대조군 벡터를 생성하는 과정을 설명하기 위한 도면이고, 도 6은 일 실시예에 따른 바이어스 제거 방법에서 영역별로 실험군 벡터와 대조군 벡터를 나누는 과정을 설명하기 위한 도면이고, 도 7은 일 실시예에 따른 바이어스 제거 방법에서 바이어스를 제거하기 전과 후의 타겟 영역수에 대한 TRR 벡터를 도시한 그래프이고, 도 8은 다양한 방법으로 바이어스를 제거한 후의 타겟 영역 수에 대한 TRR을 도시한 그래프이다.

[0045] 우선, 일 실시예에 따른 바이어스 제거 장치(300)는, 1차적으로 NMF를 통하여 바이어스를 제거하고, 2차적으로 비특이 영역을 선별함으로써 바이어스를 제거하는데, 이를 순서대로 설명하기로 한다.

[0046] 도 2를 참조하면, 일 실시예에 따른 바이어스 제거 장치(300)는, 수신부(310), 생성부(330), 제 1 제거부(350) 및 출력부(370)를 포함할 수 있고, 임의로 제 2 제거부(390)를 추가로 포함할 수 있다.

[0047] 또한, 일 실시예에 따른 타겟 염기 서열 분석에서의 바이어스 제거 방법은,

[0048] (1) 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 염색체상 위치(Chromosomal Position)별 표준 참조 염기 서열 데이터에 리드 매핑(Read Mapping)한 리드 카운트(Read Count)에 기반하여, 실험군 벡터 (실험 시료로부터 생성) 및 대조군 벡터 (대조군 시료로부터 생성)를 생성하는 단계;

[0049] (2) 상기 생성된 실험군 벡터 및 대조군 벡터를 결합한 결합 행렬을 생성하고, 상기 생성된 결합 행렬을 영역별로 나누어 바이어스(Bias)를 제거하는 단계;

[0050] (3) 상기 바이어스가 제거된 결합 행렬을 재결합하는 단계; 및

[0051] (4) 상기 바이어스가 제거된 영역별 TRR(Target Region Ratio) 벡터를 영역별로 취합하여 출력하는 단계

[0052] 를 포함하는 것일 수 있다.

[0053] 상기 바이어스 제거 방법은 타겟 염기 서열 분석에서의 바이어스 제거 장치에서 실행되는 타겟 염기 서열 분석에서의 바이어스 제거 방법일 수 있다.

[0054] 상기 단계 (1)의 실험 시료 염기 서열 데이터 및 대조 시료 염기 서열 데이터는 각각 독립적으로 유전체 서열 분석기(Sequencer)에서 생성된 서열 데이터를 직접 또는 간접적으로 수신하거나, 이미 생성된 서열 데이터가 저장된 컴퓨터 판독 가능한 저장 매체를 통하여 수득(준비)할 수 있다. 따라서, 상기 타겟 염기 서열 분석에서의 바이어스 제거 방법은, 단계 (1) 이전에, 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 준비(수신 또는 수득)하는 단계를 추가로 포함할 수 있다. 상기 실험 시료 염기 서열 데이터 및 대조 시료 염기 서열 데이터는, 각각 독립적으로, 유전체 서열 분석기(Sequencer)에서 생성된 서열 데이터를 직접 또는 간접적으로 수신하거나, 이미 생성된 서열 데이터가 저장된 컴퓨터 판독 가능한 저장 매체를 적용함으로써 준비할 수 있다.

[0055] 상기 제1 바이어스 제거 단계는 NMF(Non-negative Matrix Factorization)를 이용하여 수행되는 것일 수 있다.

[0056] 일 예에서, 상기 타겟 염기 서열 분석에서의 바이어스 제거 방법은, 상기 1차 바이어스 제거 단계 이후, 예컨대 상기 단계 (3)과 (4) 사이에, 다음의 단계를 포함하는 2차 바이어스 제거 단계를 추가로 포함할 수 있다:

[0057] (a) 상기 실험군 벡터와 대조군 벡터 간 비특이 영역을 선별 후 무차별 영역으로 설정 후 바이어스를 제거하는 단계; 및

[0058] (b) 상기 설정된 무차별 영역으로 바이어스가 제거된 상기 실험군 벡터 및 대조군 벡터의 영역별 가중치를 계산하는 단계.

[0059] 상기 2차 바이어스 제거 단계는 상기 1차 바이어스 제거 단계에서 바이어스가 제거된 결과물에 대하여 수행된다.

[0060] 도 3을 참조하면, 상기 1차 바이어스 제거 단계 (단계 (2) 및 (3) 해당)는 다음의 (i) 내지 (v)를 포함할 수 있다:

[0061] (i) 실험군 벡터와 대조군 벡터를 결합하여 결합 행렬을 생성하는 단계 (S3300);

[0062] (ii) 상기 생성된 결합 행렬을 복수개의 영역으로 나누는 단계 (S3400);

[0063] (iii) 상기 복수개의 영역별로 NMF를 수행하는 단계 (S3500);

[0064] (iv) 상기 NMF 수행 결과로부터 바이어스 요소를 선별하는 단계 (S3600); 및

[0065] (v) 바이어스 제거 후 영역별 결합 행렬을 재결합하는 단계 (S3610).

[0066] 또한, 상기 2차 바이어스 제거 단계는 다음의 (vi) 내지 (viii)를 포함할 수 있다:

[0067] (vi) 실험군 벡터와 대조군 벡터 간의 비특이적 영역을 선별하는 단계 (S3700);

[0068] (vii) 비특이적 영역을 제거하는 단계 (S3710); 및

[0069] (viii) 상기 바이어스가 제거된 실험군 벡터 및 대조군 벡터의 영역별 가중치를 계산하는 단계 (S3720).

[0070] 도 2를 이용하여 예시적으로 설명하면, 상기 단계 (1)은 생성부 (330), 단계 (2) 또는 단계 (i) 내지 (iii)은 제1 제거부 (350), 단계 (a) 및 (b) 또는 단계 (vi)내지 (viii)는 제2 제거부, 및 단계 (3) 및 (4) 또는 단계 (iv)는 출력부 (370)에서 각각 수행될 수 있으며, 임의로 단계 (1) 이전에 추가 가능한 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 준비하는 단계는 수신부 (310)에서 수행될 수 있다.

[0071] 수신부(310)는, 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 준비하는 부분으로, 예컨대, 유전체 서열 분석기(100)에서 생성된 실험 시료 염기 서열 데이터 및/또는 대조 시료 염기 서열 데이터를 수신하거나, 또는 컴퓨터 판독 가능한 저장 매체에 저장된 실험 시료 염기 서열 데이터 및/또는 대조 시료 염기 서열 데이터를 판독한다. 이때, 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터는 도 4 및 도 5와 같이, 유전체 서열 분석기(100)에서 실험 시료와 대조 시료를 각각 복수회 읽어들여 복수회의 리드 카운트(Read Count)를 가진 데이터일 수 있다.

[0072] 생성부(330)는, 준비된 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 염색체상 위치(Chromosomal Position)별 표준 참조 염기 서열 데이터에 리드 매핑(Read Mapping)한 리드 카운트(Read Count)에 기반하여, 실험군 벡터 및 대조군 벡터를 생성할 수 있다(S3100, S3200). 상기 리드 카운트는, 상기 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터에 위치한 적어도 하나의 타겟 영역에서 계산되는 것일 수 있다.

[0073] 본 명세서에 사용된 바로서, 용어 "표준 참조 염기 서열 데이터"는 한 종을 대표하는 게놈 염기 서열 데이터베이스 또는 상기 데이터베이스로부터 구축된 특정 염색체 또는 특정 염색체상 위치 (또는 영역)의 염기 서열 데이터를 지칭한다. 현재 인간의 표준 참조 염기 서열 데이터는 빌드 37(build 37: GRCh37), hg18, hg19, hg38과 같은 간행된(예컨대, UCSC, NCBI 등) 기준 게놈 서열에 근거하여 구축된 것일 수 있다.

[0074] 예를 들어, 유전체 서열 분석기(100)에서 250회의 리드 카운트를 가졌다고 가정하면, 250회 실험 시료와 대조 시료의 서열 데이터를 각각 읽어 들이면서 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터의 타겟 영역별 리드 카운트의 수를 계산할 수 있다. 이때, 리드 카운트는, 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터에 위치한 적어도 하나의 타겟 영역에서 계산될 수 있다. 또한, 대조 시료 염기 서열 데이터가 존재하지 않더라도, 즉 이미 생성(준비)한 표준 대조군 벡터가 존재하는 경우에는, 실험 시료 염기 서열 데이터와 표준 대조군 벡터에 위치한 적어도 하나의 타겟 영역에서 계산될 수 있다. 그리고, 실험군 벡터 및 대조군 벡터는 하기 수학식 1과 같다.

수학식 1

[0075]
$$T = (t_1, t_2, t_3, \ldots t_{k-1}, t_k)$$
$$N = (n_1, n_2, n_3, \ldots n_{k-1}, n_k)$$

[0076] 여기서, T는 실험군 벡터이고, N은 대조군 벡터이며, t_k는 실험군 벡터를 이루는 타겟 영역에서의 리드 카운트, 즉 깊이(Depth)이며, n_k는 대조군 벡터를 이루는 타겟 영역에서의 리드 카운트, 즉 깊이이고, k는 타겟 영역의 개수를 의미한다. 이 때, 타겟 영역의 개수 k는 시험 목적에 따라서 설정된 타겟 염기 서열의 영역의 개수를 의미하는 것일 수 있으며, 하나 이상일 수 있다.

[0077] 제 1 제거부(350)는 1차 바이어스 제거 단계를 수행하는 부분으로, 생성된 실험군 벡터 및 대조군 벡터를 결합한 결합 행렬을 생성하고, 생성된 결합 행렬을 영역별로 나누어 바이어스(Bias)를 제거할 수 있다. 여기서, 바이어스는, NMF(Non-negative Matrix Factorization)를 통하여 제거될 수 있다.

[0078] 우선, (i) 실험군 벡터와 대조군 벡터를 결합하여 결합 행렬을 생성하는 단계 (S3300)가 수행될 수 있다. 이때, 결합 행렬에 따른 행렬은 하기 수학식 2와 같다(S3300):

수학식 2

[0079] $$V = [T, N]$$

[0080] 그 다음으로, (ii) 상기 생성된 결합 행렬을 복수개의 영역으로 나누는 단계 (S3400)가 수행될 수 있다. 이 때, 상기 수학식 2의 결합 행렬에 따른 행렬은 도 6과 같이, 영역별로 분리되어 하기 수학식 3과 같이 전개될 수 있다(S3400):

수학식 3

[0081] $$V = \begin{bmatrix} T_{b1} & N_{b1} \\ T_{b2} & N_{b2} \\ T_{b3} & N_{b3} \\ \cdots & \cdots \\ T_{b,l-1} & N_{b,l-1} \\ T_{bl} & N_{bl} \end{bmatrix} = \begin{bmatrix} V_{b1} \\ V_{b2} \\ V_{b3} \\ \cdots \\ V_{b,l-1} \\ V_{bl} \end{bmatrix}$$

[0082] 여기서, l은 영역의 개수이고, k는 타겟 영역의 수이고, b는 영역(Boundary)를 의미한다. 이때, l 개의 영역에 p개의 요소가 포함되어 있다고 가정하면, k=l*p가 성립될 수 있다. 즉, p는 각 영역 (boundary) 별로 구분된 l 개 영역 내 존재하는 타겟의 수 이고, p 값의 범위는 50~200 중에서 선택된 임의의 수일 수 있다. 각 boundary 별 타겟 요소가 동일하게 존재한다면, 영역의 개수 l은 l=k/p로 자연적으로 결정된다.

[0083] 그 다음으로, (iii) 복수개의 영역별로 NMF를 수행하는 단계 (S3500)가 수행될 수 있다.

[0084] NMF(Non-negative Matrix Factorization)는 하나의 행렬을 비음수 (양수 + 0)으로 구성된 두 개의 행렬, 즉, W (특이 요소 행렬) 및 H (가중치 행렬)로 인수분해 하는 방법을 의미하며, 주로 데이터 내 독립된 특성을 추출하는데 사용된다.

[0085] NMF를 적용하면, 수학식 3은 하기 수학식 4와 같이 정리될 수 있다. 즉, 각 영역별로 구분된 수학식 3의 행렬 Vb에 NMF를 적용하면 하기 수학식 4와 같다(S3500).

수학식 4

[0086] $$V_b = [T_b, N_b] = WH$$

[0087] 여기서, V=n*p이고, W는 n*r, H는 r*p가 되므로 (이 때, n은 target 영역의 수, r은 NMF시 사용되는 rank를 의미함), NMF 적용시 Rank를 2로 고정하면, 수학식 4는 수학식 5와 같이 전개될 수 있다.

수학식 5

[0088]

$$\begin{bmatrix} V_{1,1} & V_{1,2} \\ V_{2,1} & V_{2,2} \\ V_{3,1} & V_{3,2} \\ \cdots & \cdots \\ V_{p-1,1} & V_{p-1,2} \\ V_{p,1} & V_{p,2} \end{bmatrix} = \begin{bmatrix} t_1 & n_1 \\ t_2 & n_2 \\ t_3 & n_3 \\ \cdots & \cdots \\ t_{p-1} & n_{p-1} \\ t_p & n \end{bmatrix} = \begin{bmatrix} W_{1,1} & W_{1,2} \\ W_{2,1} & W_{2,2} \\ W_{3,1} & W_{3,2} \\ \cdots & \cdots \\ W_{p-1,1} & W_{p-1,2} \\ W_{p,1} & W_{p,2} \end{bmatrix} \times \begin{bmatrix} H_{1,1} & H_{1,2} \\ H_{2,1} & H_{2,2} \end{bmatrix}$$

[0089] 수학식 5를 보면, T_b는 $W_{,1}xH_{1,1}+W_{,2}xH_{2,1}$이고, Nb는 $W_{,1}xH_{1,2}+W_{,2}xH_{2,2}$로 구성된다는 것을 알 수 있다. 또한, $W_{,1}$ 벡터는 실험군 벡터 및 대조군 벡터의 공통 요소 벡터이고, $W_{,2}$ 벡터는 실험군 벡터 또는 대조군 벡터의 특이 요소 벡터이다. 그리고, $H_{2,1}$과 $H_{2,2}$는 실험군 벡터 및 대조군 벡터에만 각각 곱해지는 가중치이다. 또한, $H_{1,1}$와 $H_{1,2}$는 대부분 0보다 매우 큰 값을 가지나, $H_{2,1}$과 $H_{2,2}$는 경우에 따라 0에 가까운 값을 가질 수 있다. NMF 수행 후 분해된 $H_{2,1}$과 $H_{2,2}$의 요소의 값을 비교하여 바이어스 요소를 선별할 수 있다. 따라서, 상기 수학식 5는 NMF에서 rank를 2로 고정한 경우의 바이어스 요소 선별 단계 (S3600)를 예시하는 것일 수 있다.

[0090] 상술한 바와 같이 NMF를 수행하고 난 후,

[0091] (3) 바이어스 제거 후 영역별 결합 행렬을 재결합하는 단계 (S3610)), 및

[0092] (4) 상기 바이어스가 제거된 영역별 TRR(Target Region Ratio) 벡터를 영역별로 취합하여 출력하는 단계를 수행할 수 있다.

[0093] 상기 단계 (3) 및 (4)는 출력부(370)에서 수행되는 것일 수 있다.

[0094] 상기 단계 (3) 및 (4)에 있어서, 아래의 수학식 6과 같이, $H_{2,1}$가 $H_{2,2}$보다 큰 경우, 예를 들어 $H_{2,2}$가 0에 가까운 경우에는, $W_{,2}$가 실험군에 특이 요소 벡터임을 의미하므로, $H_{2,2}$를 제외하고 행렬을 재결합할 수 있다(S3610).

수학식 6

[0095]

$$H_{2,1} > H_{2,2} \quad \text{인 경우,}$$

$$T_b^* = W_{,1} \times H_{1,1} - W_{,2} \times H_{2,1}$$

$$N_b^* = W_{,1} \times H_{1,2}$$

[0096] 즉, $H_{2,2}$가 0에 가까운 경우에는, $H_{2,2}$가 실험군에 특이 요소 벡터, 즉 노이즈라는 의미이므로, 이를 포함하는 항을 삭제함으로써 바이어스를 제거할 수 있다.

[0097] 반대로, 수학식 7과 같이, $H_{2,1}$이 $H_{2,2}$보다 작은 경우, 예를 들어 $H_{2,1}$이 0에 가까운 경우에는, $W_{,2}$가 대조군에 특이 요소 벡터임을 의미하므로, $H_{2,1}$를 제외하고 행렬을 재결합할 수 있다(S3610).

수학식 7

[0098]

$$H_{2,1} < H_{2,2} \quad \text{인 경우,}$$

$$T_2^* = W_{,1} \times H_{1,1}$$

$$N_2^* = W_{,1} \times H_{1,2} + W_{,2} \times H_{2,2}$$

[0099] 즉, $H_{2,1}$가 0에 가까운 경우에는, $H_{2,1}$가 대조군에 특이 요소 벡터, 즉 노이즈라는 의미이므로, 이를 포함하는 항을 삭제함으로써 바이어스를 1차적으로 제거할 수 있다.

[0100] 상술한 바와 같이, 본 발명의 일 실시예에 따른 바이어스 제거 방법은, 실험군 벡터와 대조군 벡터에 대한 결합 벡터 행렬을 영역별로 나누고, 영역별로 노이즈인 바이어스를 제거한 후, 이를 다시 취합함으로써 모든 영역에서 일괄적으로 바이어스를 제거함에 따라 발생했던 민감도 하락의 문제를 없앨 수 있고, 영역에 특이적으로 발생할 수 있는 바이어스를 영역별로 제거함에 따라 체세포 복제수 변이를 파악하는데 정확도를 높일 수 있다.

[0101] 한편, 본 발명의 일 실시예에 따른 바이어스 제거 방법은, 바이어스를 NMF를 이용하여 1차적으로 제거한 후, 2차적으로 비특이 영역을 선별함에 따라 바이어스를 제거하는 단계 (2차 바이어스 제거 단계)를 추가로 수행할 수 있다. 이하에서는 비특이 영역을 선별하는 바이어스 제거 방법을 설명하기로 한다. 상기 2차 바이어스 제거 단계는 제 2 제거부(390)에서 수행될 수 있다.

[0102] 구체적으로, 상기 2차 바이어스 제거 단계는,

[0103] (a) 실험군 벡터 및 대조군 벡터의 비특이 영역을 각각 선별하여 무차별 영역으로 설정하여 바이어스를 제거하는 단계(S3700, S3710); 및

[0104] (b) 상기 설정된 무차별 영역으로 바이어스가 제거된 상기 실험군 벡터 및 대조군 벡터의 영역별 가중치를 계산하는 단계(S3720)

[0105] 를 통하여 수행될 수 있다.

[0106] 우선,(a) 실험군 벡터와 대조군 벡터 간 비특이 영역을 선별 후 무차별 영역으로 설정하여 바이어스를 제거하는 단계(S3700, S3710)에 있어서, 실험군 벡터와 대조군 벡터 간 비특이 영역을 하기 수학식 8 및 수학식 9를 통하여 선별할 수 있다.

수학식 8

[0107]

$$W_{ratio} = \frac{W_{,2}}{W_{,1}} \times \frac{\sum W_{,1}}{\sum W_{,2}}$$

수학식 9

[0108]

$$W_{ratio} < \theta$$

[0109] 즉, $W_{,1}$ 벡터에 비하여, $W_{,2}$ 벡터의 값이 매우 작은 위치, 즉 공통 요소 벡터에 비하여 특이 요소 벡터의 값이 매우 작은 위치를 선별하는데, 해당 위치는 실험군과 대조군 간의 비특이 영역을 의미하게 된다. 수학식 9와 같이, 특정 임계값(Θ)보다 작은 위치를 선별할 수 있다. 즉, 특이값의 비율이 낮으면 실험군과 대조군이 유사하다는 것이므로, 이는 곧 무차별 영역으로 선택될 수 있다. 이렇게, 수학식 8 및 수학식 9를 통하여 무차별 영역이 선택된 경우, 하기 수학식 10 및 수학식 11과 같이, 무차별 영역에 대응하는 요소 벡터를 -1로 변환함으로써, 해당 요소 벡터에 대응하는 바이어스를 2차적으로 제거하도록 한다. 여기서, 실험군 벡터 및 대

조군 벡터의 무차별 영역은 동일하다.

수학식 10

[0110] $$T_b^{*} = \begin{bmatrix} T_1^{*} \\ T_2^{*} \\ T_3^{*} \\ \cdots \\ T_{p-1}^{*} \\ T_p^{*} \end{bmatrix} \rightarrow T_b^{**} = \begin{bmatrix} T_1^{*} \\ -1 \\ -1 \\ \cdots \\ T_{p-1}^{*} \\ T_p^{*} \end{bmatrix}$$

수학식 11

[0111] $$N_b^{*} = \begin{bmatrix} N_1^{*} \\ N_2^{*} \\ N_3^{*} \\ \cdots \\ N_{p-1}^{*} \\ N_p^{*} \end{bmatrix} \rightarrow N_b^{**} = \begin{bmatrix} N_1^{*} \\ -1 \\ -1 \\ \cdots \\ N_{p-1}^{*} \\ N_p^{*} \end{bmatrix}$$

[0112] (b) 상기 설정된 무차별 영역으로 바이어스가 제거된 상기 실험군 벡터 및 대조군 벡터의 영역별 가중치를 계산하는 단계에 있어서, 설정된 무차별 영역으로 바이어스가 제거된 상기 실험군 벡터 및 대조군 벡터의 영역별 가중치를 계산하고(S3720), 바이어스가 제거된 영역별 TRR(Target Region Ratio) 벡터를 영역별로 취합하여 출력할 수 있다 (상기 수학식 10 및 11에서, p는 1 개의 영역의 요소 개수를 의미한다. T_b^{*}와 N_b^{*}는 바이어스를 1차로 제거한 후의 영역별 대조군 벡터와 실험군 벡터를 각각 의미한다. T_b^{**}와 N_b^{**}는 무차별 영역을 -1로 치환한 후의 영역별 대조군 벡터와 실험군 벡터를 각각 의미한다. 다시 말하면, 수학식 10 및 수학식 11에서와 같이, 영역별 가중치를 재계산하여 수학식 12와 같이 영역별 TRR 벡터를 계산할 수 있다.

수학식 12

[0113] $$TRR_b^{*} = \frac{T_b^{**}}{N_b^{**}} \times \frac{\sum N}{\sum T} \times \frac{\sum T_b^{**}}{\sum N_b^{**}}$$

[0114] 여기서, TRR(Target Region Ratio) 벡터는, 실험 시료 염기 서열 데이터 또는 실험군 벡터와, 대조 시료 염기 서열 데이터 또는 대조군 벡터에 위치한 적어도 하나의 타겟의 수에 기초하여 생성될 수 있다. 이렇게 계산된 영역별 TRR 벡터는 영역별로 취합되어 하기 수학식 13과 같이 생성될 수 있다(S3800)(상기 수학식 12 및 13에서, b는 영역(Boundary)을 의미하고, 1은 영역의 개수를 의미하고, TRR_b^{*}는 바이어스를 1차와 2차로 제거한 후의 영역별 TRR 벡터를 의미한다.):

수학식 13

[0115] $$TRR^{\cdot} = \begin{bmatrix} TRR_{1}^{\cdot} \\ TRR_{2}^{\cdot} \\ TRR_{3}^{\cdot} \\ \cdots \\ TRR_{l-1}^{\cdot} \\ TRR_{l}^{\cdot} \end{bmatrix}$$

[0116] 상술한 바와 같이, NMF를 통하여 1차적으로 바이어스가 제거되고, 무차별 영역을 제거함으로써 2차적으로 바이어스가 제거된 TRR을 통한 체세포 복제수 변이 판단은, 영역별로 특이적인 바이어스를 제거하거나 노이즈를 없앤 후 취합함으로써, 영역별 정확도 및 민감도를 증가시킬 수 있고, 이는 도 7 및 도 8을 참조하여 설명한다.

[0117] 도 7은 인간 게놈 염색체(시험 시료: HCC1143 Cell line (ATCC), 대조 시료: HCC1143 BL (ATCC))에 있어서 바이어스의 제거 전 후의 타겟 영역 수에 대한 TRR 벡터를 도시한 그래프로서, (a)는 바이어스를 제거하기 전의 타겟 영역 수에 대한 TRR 벡터를 도시하고, (b)는 1차 바이어스 제거 및 2차 바이어스 제거를 수행한 후의 타겟 영역 수에 대한 TRR 벡터를 도시한다. 여기서, (a)와 (b)를 비교하여 보면, 각각의 영역별로 TRR 벡터의 구분이 본 발명의 따른 (b)가 더 잘되어 있는 것을 알 수 있다. 즉, (a)는 바이어스나 노이즈가 많아 체세포 복제수 변이를 식별하기 어렵지만, (b)는 바이어스 및 노이즈가 제거된 상태이므로, 체세포 복제수 변이의 식별이 보다 용이해짐을 알 수 있다.

[0118] 도 8은 인간 게놈 염색체(시험 시료: HCC1143 Cell line (ATCC), 대조 시료: HCC1143 BL (ATCC))에 있어서 다양한 방법으로 바이어스를 제거한 후의 타겟 영역 수에 대한 TRR을 도시한 그래프로서,

[0119] (a)는 바이어스를 제거하기 전의 타겟 영역 수에 대한 TRR 벡터를 도시한 것이고,

[0120] (b)는 SVD (singular Value Deposition; 컷 오프 값으로 l=1, 즉 1개의 Singular value를 제거하였음) 방법으로 전체 영역에 대해서 한꺼번에 바이어스를 제거한 후(즉, 앞서 제시된 방법에서 단계 (2)의 결합 행렬을 영역별로 나누는 단계 (단계 (ii) (S3400))를 제외하고 NMF 대신에 SVD를 수행하여 바이어스 제거 단계를 수행하고, 2차 바이어스 제거 단계(S3700)는 수행하지 않음)의 타겟 영역 수에 대한 TRR 벡터를 도시한 것이고,

[0121] (c)는 NMF 방법으로 전체 영역에 대해서 한꺼번에 바이어스를 제거한 후 (즉, 앞서 제시된 방법에서 단계 (2)의 결합 행렬을 영역별로 나누는 단계 (단계 (ii) (S3400))를 제외하고 바이어스 제거 단계를 수행하고 2차 바이어스 제거 단계 (S3700)를 수행함)의 타겟 영역 수에 대한 TRR 벡터를 도시한 것이고,

[0122] (d)는 앞서 제시된 바와 같이 NMF 방법으로 각 영역별로 1차 바이어스 (S3600)와 2차 바이어스 (S3700)를 제거한 후의 타겟 영역 수에 대한 TRR 벡터를 도시한 것이다.

[0123] 상기 도 8의 (b)의 SVD는 아래의 참고식 1-8에 의하여 수행하였다:

[0124] <참고식 1>

[0125] $$\overrightarrow{n_1} = (n_{11}, \quad n_{12}, \quad \ldots, \quad n_{1k})$$
$$\overrightarrow{n_l} = (n_{l1}, \quad n_{l2}, \quad \ldots, \quad n_{lk})$$

[0126] <참고식 2>

[0127] $$\overrightarrow{m_l} = \frac{1}{\sqrt[k]{\prod_{i=1}^{k} n_{li}}} \overrightarrow{n_l}$$

[0128] <참고식 3>

$$\vec{s} = \frac{1}{N}\left(\sum_{l=1}^{N} m_{l1}, \sum_{l=1}^{N} m_{l2}, \cdots, \sum_{l=1}^{N} m_{lk-1}, \sum_{l=1}^{N} m_{lk}\right)$$

[0129]

[0130] (상기 참고식 1-3에서, l은 l 번째 대조 시료를 의미하고, m은 l 번째 대조 시료의 정규 벡터를 의미하고, s는 대조 시료의 표준 대조군 벡터를 의미하고, N은 대조 시료의 개수를 의미함)

[0131] <참고식 4>

$$\vec{t_1} = \left(\frac{n_1}{r_1}, \quad \frac{n_2}{r_2}, \quad \cdots, \quad \frac{n_k}{r_k}\right)$$

[0132]

[0133] (참고식 4에서, t_1은 TRR 벡터이고, n_i는 실험 시료 서열 데이터의 i 위치에서의 리드 카운트의 수이고, r_i는 표준 대조군 벡터의 i 위치에서의 리드 카운트의 수이고, 참고식 4-8의 k는 타겟의 수임)

[0134] <참고식 5>

$$T = [\vec{t_1} \quad \vec{t_2} \quad \ldots \quad \vec{t_N}]$$

[0135]

[0136] (참고식 5에서, T는 벡터 어레이이고, t_1, t_2,...t_N은 TRR 벡터이며, N은 실험 시료 서열 데이터의 개수임)

[0137] <참고식 6>

$$T = U\Sigma V^T, \Sigma = \begin{bmatrix} sv_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & sv_k \end{bmatrix}$$

$$set\ sv_i = 0, where\ i = 1 \sim l$$

[0138]

[0139] (참고식 6에서, T는 $U\sum V^T$와 같이 인수분해될 수 있고, l은 특이값 컷 오프(Singular Value Cutoff)로 정의되므로, 컷 오프는 하기 참고식 7로 결정됨)

[0140] <참고식 7>

$$\Sigma = \begin{bmatrix} sv_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & sv_k \end{bmatrix}$$

$$r = \frac{\sum_{j=1}^{l} sv_j}{\sum_{i=1}^{k} sv_i}$$

[0141]

[0142] (참고식 7에서, $0.1 \leq r \leq 0.6$이 되도록 컷 오프인 l을 결정함 (도 8의 경우 l=1)

[0143] <참고식 8>

$$T' = U * \begin{pmatrix} 0 & & & \cdots & 0 \\ & \ddots & & & \\ & & 0 & & \\ \vdots & & & sv_{l+1} & \\ & & & \ddots & \vdots \\ 0 & & & \cdots & sv_k \end{pmatrix} * V^T$$

[0144]

[0145] (참고식 8에서, T'는 바이어스 제거 후의 TRR 벡터를 의미함)

[0146] 도 8에서 사전에 알려진 복제수 변이가 있는 영역에 대해서는 "T"로 표기하였다. "T"로 표기되지 않은 영역에서 threshold인 기준선을 넘어 존재하는 TRR 값은 False positive일 가능성이 높다.

[0147] 도 8의 (a)는 T로 표시된 영역 이외에도 기준선을 넘은 TRR 값이 많은 것을 볼 수 있어 바이어스로 인해 체세포 복제수 변이를 식별하기 어렵다. (b)는 T로 표시된 영역에서 TRR 값이 증가해 복제수 변이 발굴의 민감성은 증가하지만, 기준선을 넘은 영역이 더 많아져 바이어스 제거 효과는 크지 않다. (c)는 (a)나 (b)에 비해 T로 표시된 영역의 TRR 값이 증가하고, T로 표시되지 않은 영역의 TRR 값은 감소하여 바이어스가 일부 제거되는 것을 볼 수 있다. (d)는 (c)에서 일부 남아있는 T로 표시되지 않은 영역의 TRR 값은 더욱 감소하고, T로 표시된 영역의 TRR 값이 더욱 증가하는 것을 볼 수 있다.

[0148] 따라서 NMF를 이용하여 각 영역별로 바이어스를 제거 시, 다른 방법에 비해 체세포 복제수 변이의 식별이 보다 용이해짐을 알 수 있다.

[0149] 염기 서열 분석염기 서열 분석도 9는 일 실시예에 따른 타겟 염기 서열 분석에서의 바이어스 제거 방법을 예시적으로 설명하기 위한 동작 흐름도이다. 도 9를 참조하면, 바이어스 제거 장치는, 유전체 서열 분석기(Sequencer)에서 생성된 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 수신한다(S8100).

[0150] 그리고 나서, 바이어스 제거 장치는, 수신된 실험 시료 염기 서열 데이터와 대조 시료 염기 서열 데이터를 염색체상 위치(Chromosomal Position)별 표준 참조 염기 서열 데이터에 리드 매핑(Read Mapping)한 리드 카운트(Read Count)에 기반하여, 실험군 벡터 및 대조군 벡터를 생성한다(S8200).

[0151] 또한, 바이어스 제거 장치는, 생성된 실험군 벡터 및 대조군 벡터를 결합한 결합 행렬을 생성하고, 생성된 결합 행렬을 영역별로 나누어 바이어스(Bias)를 제거하고(S8300), 바이어스가 제거된 결합 행렬을 재결합하고, 바이어스가 제거된 영역별 TRR(Target Region Ratio) 벡터를 영역별로 취합하여 출력한다(S8400).

[0152] 이와 같은 도 9의 타겟 염기 서열 분석에서의 바이어스 제거 방법에 대해서 설명되지 아니한 사항은 앞서 도 1 내지 도 8을 통해 타겟 염기 서열 분석에서의 바이어스 제거 방법에 대하여 설명된 내용과 동일하거나 설명된 내용으로부터 용이하게 유추 가능하므로 이하 설명을 생략하도록 한다.

[0153] 또한, 다른 예는 상기 바이어스 제거 방법의 단계를 수행하기 위한 시스템을 제공한다. 일 예에서, 상기 시스템은, 상기한 바와 같은 단계를 수행하는, 수신부(310), 생성부(330), 제 1 제거부(350) 및 출력부(370)를 포함할 수 있고, 임의로 제 2 제거부(390)를 추가로 포함하는 컴퓨터 시스템일 수 있다.

[0154] 다른 예는 통상적인 염기 서열 분석에 있어서, 상기한 바와 같은 바이어스 제거 방법을 수행하는 단계를 포함하는, 타겟 염기 서열 분석을 위한 컴퓨터 판독 방법 방법을 제공한다.

[0155] 상기 바이어스 제거 방법 또는 이를 포함하는 컴퓨터 판독 방법은 컴퓨터에 의해 실행 가능한 프로그램(computer executable instruction)으로서, 공지된 컴퓨터 판독 가능한 매체 상에서 전체적 또는 부분적으로 구현 및/또는 처리될 수 있다. 예컨대, 본 명세서에 기재된 방법은 하드웨어에 결합되어 구현될 수 있다. 상기 하드웨어는 컴퓨터, 표준 다목적(multi-purpose) CPU, ASIC(application-specific integrated circuit) 또는 다른 하드-와이어드 장치(hard-wired device)와 같은 특수하게 설계된 하드웨어 또는 펌웨어를 의미하는 것일 수 있으며, 이하 사용되는 용어 '컴퓨터'는 이들을 총칭하기 위한 것일 수 있다.

[0156] 다른 예는 상기 바이어스 제거 방법 또는 이를 포함하는 컴퓨터 판독 방법의 단계를 실행시키기 위하여 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램을 제공한다. 상기 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램은 하드웨어와 결합된 것일 수 있다. 상기 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램은 상기한 바와 같은 바이어스 제거 방법 또는 이를 포함하는 컴퓨터 판독 방법의 각 단계를 컴퓨터에서 실행시키기 위한 프로그램이며, 이 때 상기한 모든 단계가 하나의 프로그램에 의하여 실행되거나, 하나 이상의 단계를 실행하는 두 개 이상의 프로그램에 의하여 실행될 수 있다.

[0157] 다른 예는 상기 바이어스 제거 방법 또는 이를 포함하는 컴퓨터 판독 방법의 단계를 실행시키기 위한 컴퓨터에서 실행 가능한 프로그램(computer executable instruction)이 수록된 컴퓨터 판독 가능한 저장 매체 (또는 기록 매체)를 제공한다.

[0158] 상기 컴퓨터에서 실행 가능한 프로그램은 컴퓨터 판독 가능한 저장 매체 (예컨대, 메모리 등)에 저장되고, 하나 이상의 프로세서 상에 구현된 소프트웨어로 구현될 수 있다. 일반적으로 알려진 바와 같이, 프로세서는 하나 이

상의 컨트롤러(controller), 연산 유닛(calculation unit) 및/또는 컴퓨터 시스템의 다른 유닛과 결합되거나, 적절한 펌웨어(firmware)에 이식될 수 있다. 상기 프로그램이 소프트웨어에 이식되는 경우, RAM (Random Access Memory), ROM (Read Only Memory), EEPROM (Electrically Erasable Programmable Read-Only Memory), 플래쉬 메모리 (e.g., USB(Universal Serial Bus) 메모리, SD(Secure Digital) 메모리, SSD(Soli State Drive), CF (Compact Flash) 메모리, xD 메모리 등), 자기 디스크, 레이저 디스크, 또는 기타 저장 매체와 같은 컴퓨터 판독가능한 저장 매체에 저장될 수 있다. 상기 컴퓨터 판독 가능한 저장 매체에 저장된 프로그램 또는 소프트웨어는, 예컨대, 전화선, 인터넷, 무선 접속 등과 같은 통신 채널 상에서, 또는 컴퓨터 판독가능한 디스크, 플래쉬 드라이브 등과 같은, 휴대용 매체(transportable medium)를 통한 것을 포함하는 모든 공지된 전달 방법을 통하여 컴퓨터 장치에 전달될 수 있다.

[0159] 상기한 바와 같은 다양한 단계들이 통상적으로 알려진 다양한 블록, 작업(operation), 툴, 모듈, 및 하드웨어, 펌웨어, 소프트웨어, 또는 하드웨어, 펌웨어 및/또는 소프트웨어의 조합에서 구현될 수 있는 기법으로서 구현될 수 있다. 하드웨어에서 구현되는 경우, 블록, 작업, 기법 등의 일부 또는 전부가, 예컨대, 맞춤화 집적 회로(custom IC), ASIC(application specific integrated circuit), FPGA(field programmable logic array), PLA(programmable logic array) 등에서 구현될 수 있다. 소프트웨어에서 구현되는 경우, 소프트웨어는 자기 디스크, 광 디스크, 또는 다른 저장 매체와 같은 공지된 컴퓨터 판독가능한 매체, 컴퓨터의 RAM, 또는 ROM 또는 플래쉬 메모리, 프로세서, 하드 디스크 드라이브, 광 디스크 드라이브, 테이프 드라이브 등에 저장될 수 있다. 또한, 소프트웨어는, 예컨대, 컴퓨터 판독가능한 디스크 또는 다른 휴대용 컴퓨터 저장 메카니즘을 포함한 공지된 전달 방법을 통해 사용자 또는 컴퓨터 시스템에 전달될 수 있다.

[0160] 상기 바이어스 제거 방법, 컴퓨터 판독 방법, 프로그램, 및 저장매체는 다수의 다른 범용(general purpose) 또는 특수 목적 컴퓨팅 시스템 환경 또는 구조에서 운영될 수 있다. 상기 바이어스 제거 방법, 컴퓨터 판독 방법, 프로그램, 및 저장매체를 실행하기에 적합한 컴퓨팅 시스템, 환경, 및/또는 구조는 예컨대, 퍼스널 컴퓨터(PC), 서버 컴퓨터, 휴대용 또는 랩탑(laptop) 장치, 멀티프로세서 시스템, 마이크로프로세서-기반 시스템, 셋탑 박스, 프로그램가능한(programmable) 가전(consumer electronics), 네트워크 PC, 미니컴퓨터, 메인프레임 컴퓨터, 및/또는 상기한 시스템 또는 장치를 포함하고 통신 네트워크를 통해 연결된 원격 처리 장치들에 의해 수행되는 분산 컴퓨팅(distributed computing) 환경 등을 포함할 수 있으나, 이에 제한되지 않는다. 통합 컴퓨팅 환경 및 분산 컴퓨팅 환경 모두에서, 프로그램 모듈은 메모리 저장 장치를 포함한, 로컬 및 원격 컴퓨터 저장 매체에 위치될 수 있다.

[0161] 컴퓨터는 통상적으로 다양한 컴퓨터 판독가능한 매체를 포함할 수 있다. 컴퓨터 판독가능한 매체는 컴퓨터에 의해 접근 가능하고 이용 가능한 매체일 수 있고 휘발성 매체 및 비휘발성 매체, 이동성(removable) 매체 및 비이동성 매체를 포함할 수 있다. 예컨대, 컴퓨터 판독가능한 매체는 컴퓨터 저장 매체 및/또는 통신 매체(communication media)를 포함할 수 있다.

[0162] 상기 컴퓨터 판독 가능한 저장 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 매체 비분리형 매체, 이동성(removable) 매체 및/또는 비이동성 매체 등 통상적인 모든 매체를 의미하는 것일 수 있다. 또한, 컴퓨터 판독 가능한 저장 매체는 컴퓨터 저장 매체 및 통신 매체를 모두 포함할 수 있다.

[0163] 상기 컴퓨터 저장 매체는 컴퓨터 판독가능한 명령어, 데이터 구조, 프로그램 모듈 및/또는 기타 데이터와 같은 정보의 저장을 위한 방법 또는 기술에서 구현된, 휘발성 또는 비휘발성, 및/또는 이동성 또는 비이동성 매체를 포함할 수 있다. 컴퓨터 저장 매체는 RAM, ROM, EEPROM, 플래쉬 메모리(e.g., USB 메모리, SD 메모리, SSD, CF 메모리, xD 메모리 등), 자기 디스크, 레이저디스크, 또는 기타 메모리, CD-ROM, DVD(digital versatile disk) 또는 기타 광학적 디스크, 자기 카세트(magnetic cassette), 자기테이프, 자기 디스크 저장 또는 기타 자기 저장 장치, 또는 원하는 정보를 저장하기 위해 이용될 수 있고 컴퓨터에 의해 접근 가능한 모든 매체들 중에서 하나 이상 선택될 수 있으나, 이에 제한되지 않는다.

[0164] 상기 통신 매체는 통상적으로 컴퓨터 판독가능한 명령어, 데이터 구조, 프로그램 모듈, 또는 반송파(carrier wave)와 같은 모듈화 데이터 신호 (modulated data signal) 중 데이터 전송 또는 기타 전송(transport) 메카니즘을 구현하는 정보 전달 매체(information delivery media)를 포함할 수 있다. 용어 "모듈화 데이터 신호(modulated data signal)"는 신호에 정보를 코딩하는 방식으로 설정되거나 변경된 하나 이상의 특징을 갖는 신호를 의미한다. 예컨대, 상기 통신 매체는 유선 네트워크 또는 직접-유선 연결(direct-wired connection)과 같은 유선 매체, 및 음향(acoustic) 매체, RF, 적외선 및 기타 무선 매체와 같은 무선 매체를 포함한다.

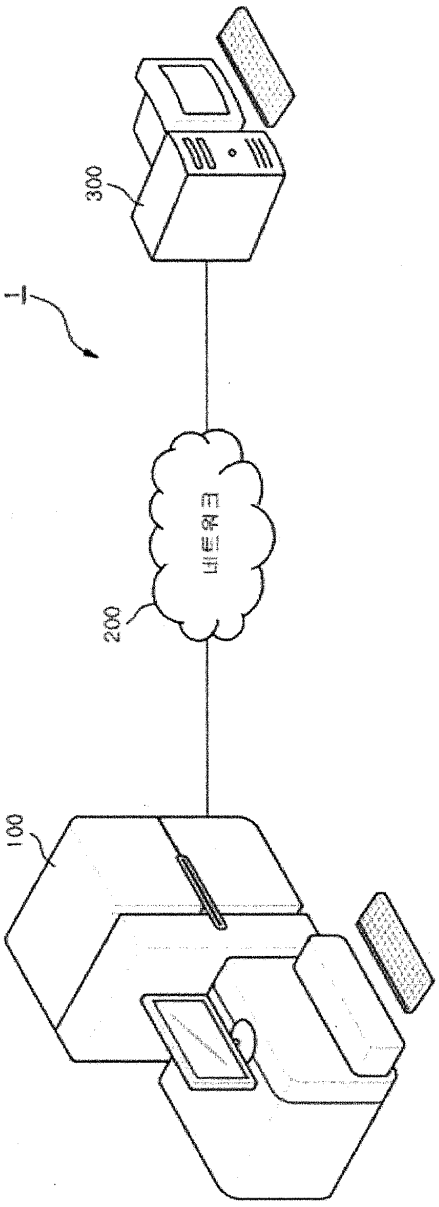
[0165] 상기한 매체들 중 하나 이상의 조합도 컴퓨터 판독 가능한 매체의 범위 내에 포함될 수 있다.

[0166] 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.

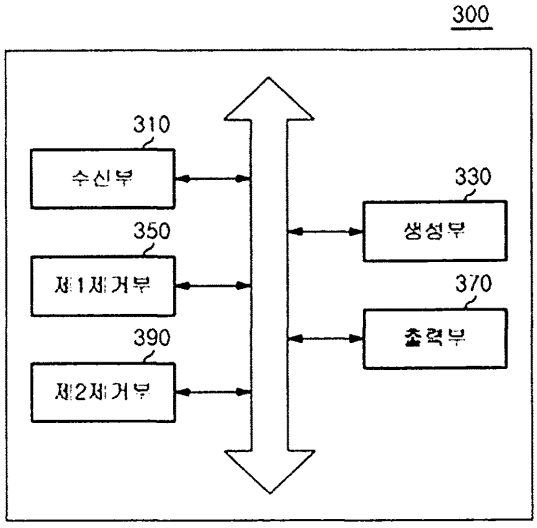
[0167] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

도면

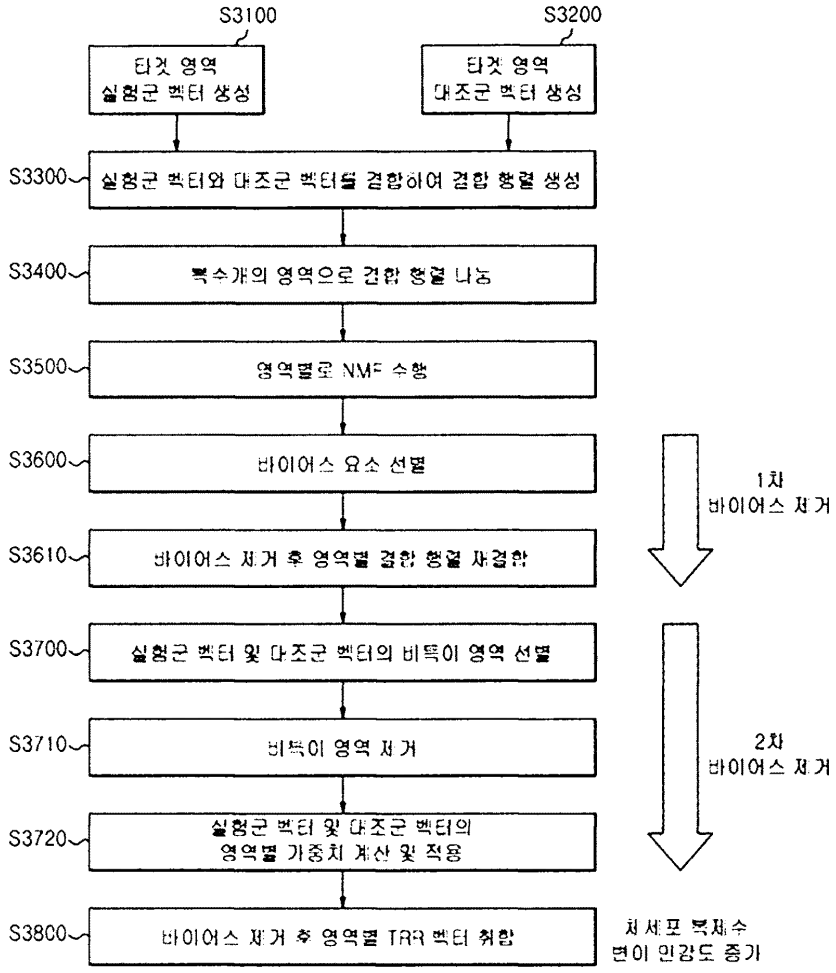
도면1



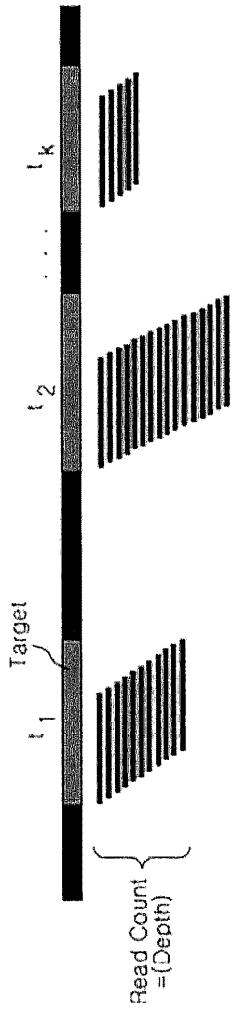
도면2


300
310
수신부
330
생성부
350
제1제거부
370
출력부
390
제2제거부

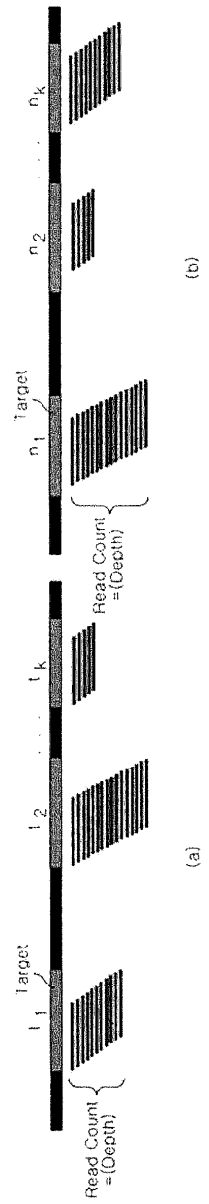
도면3


S3100
타겟 영역 실험군 벡터 생성
S3200
타겟 영역 대조군 벡터 생성
S3300
실험군 벡터와 대조군 벡터를 결합하여 결합 행렬 생성
S3400
복수개의 영역으로 결합 행렬 나눔
S3500
영역별로 NMF 수행
S3600
바이어스 요소 선별
S3610
바이어스 제거 후 영역별 결합 행렬 재결합
1차 바이어스 제거
S3700
실험군 벡터 및 대조군 벡터의 비특이 영역 선별
S3710
비특이 영역 제거
S3720
실험군 벡터 및 대조군 벡터의 영역별 가중치 계산 및 적용
2차 바이어스 제거
S3800
바이어스 제거 후 영역별 TRR 벡터 취합
차세포 복제수 변이 민감도 증가

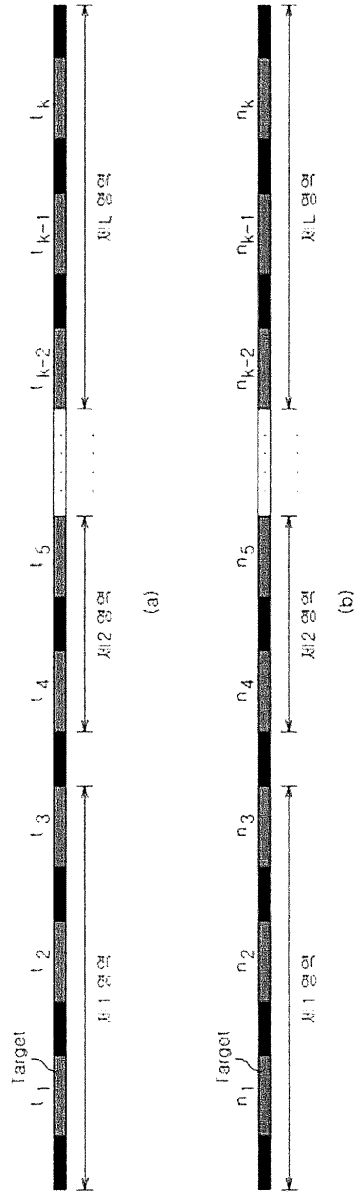
도면4



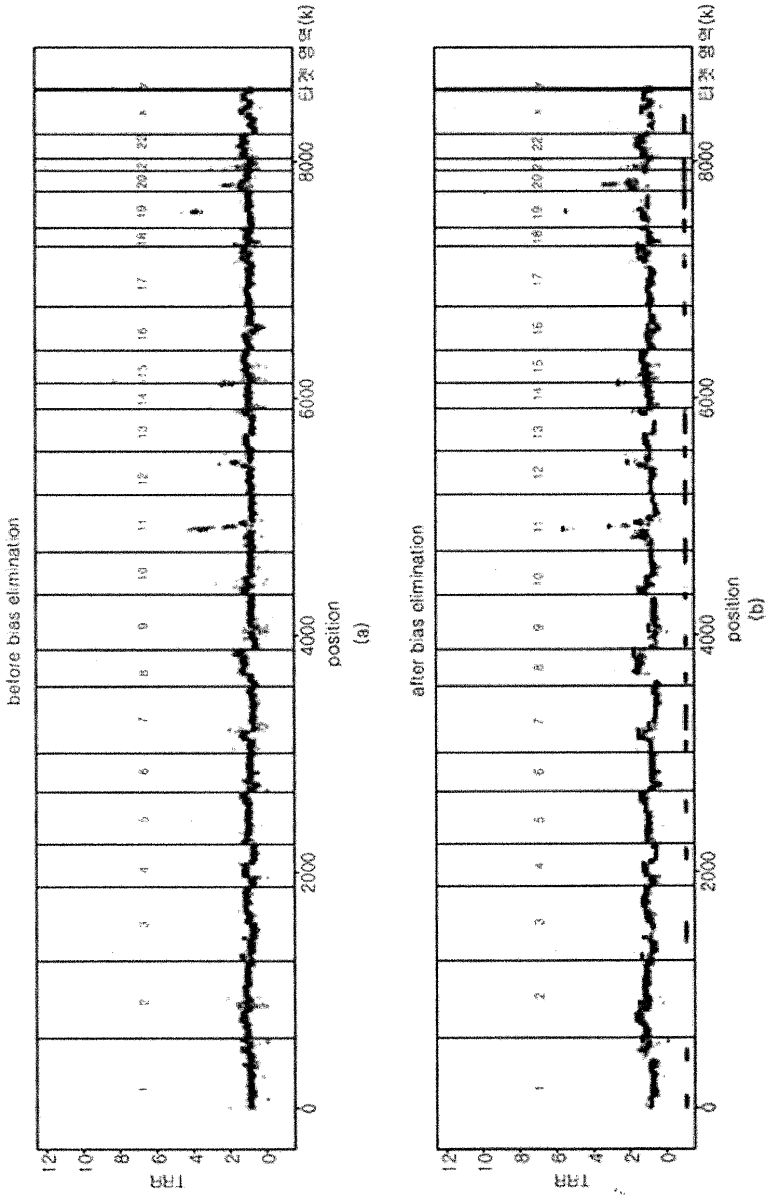
도면5


t1
Target
t2
tk
Read Count
=(Depth)
(a)
n1
Target
n2
nk
Read Count
=(Depth)
(b)

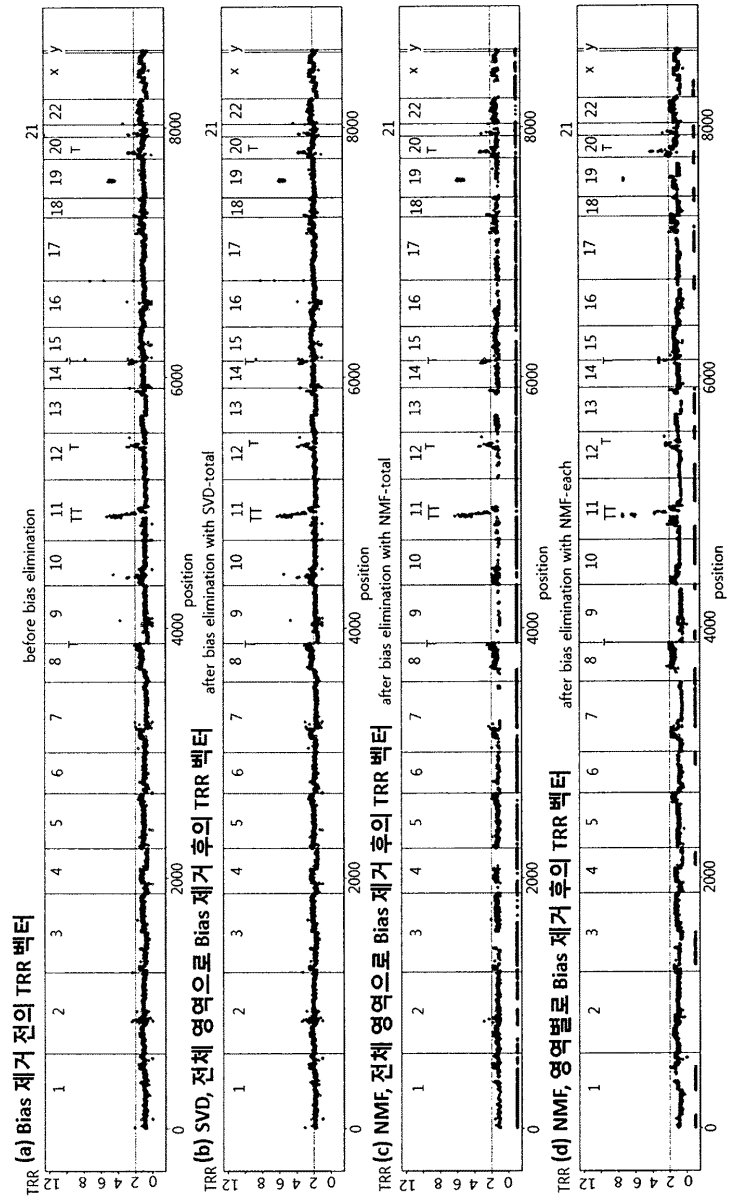
도면6


target
t_1
t_2
t_3
t_4
t_5
t_{k-2}
t_{k-1}
t_k
제1 영역
제2 영역
제L 영역
(a)
target
n_1
n_2
n_3
n_4
n_5
n_{k-2}
n_{k-1}
n_k
제1 영역
제2 영역
제L 영역
(b)

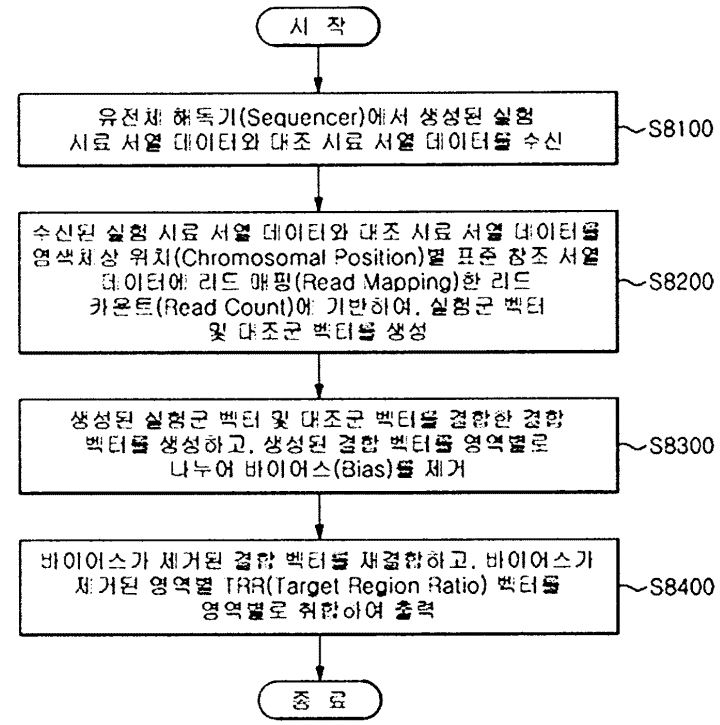
도면7


before bias elimination
after bias elimination
position
(a)
(b)
TAR
타겟 영역(k)
0
2000
4000
6000
8000
0
2
4
6
8
10
12

도면8


(a) Bias 제거 전의 TRR 벡터
before bias elimination
(b) SVD, 전체 영역으로 Bias 제거 후의 TRR 벡터
after bias elimination with SVD-total
(c) NMF, 전체 영역으로 Bias 제거 후의 TRR 벡터
after bias elimination with NMF-total
(d) NMF, 영역별로 Bias 제거 후의 TRR 벡터
after bias elimination with NMF-each
TRR
position
0
2000
4000
6000
8000
21

도면9


시 작
유전체 해독기(Sequencer)에서 생성된 실험 시료 서열 데이터와 대조 시료 서열 데이터들 수신
S8100
수신된 실험 시료 서열 데이터와 대조 시료 서열 데이터들 염색체상 위치(Chromosomal Position)별 표준 참조 서열 데이터에 리드 매핑(Read Mapping)한 리드 카운트(Read Count)에 기반하여, 실험군 벡터 및 대조군 벡터들 생성
S8200
생성된 실험군 벡터 및 대조군 벡터들 결합한 결합 벡터들 생성하고, 생성된 결합 벡터들 영역별로 나누어 바이어스(Bias)들 제거
S8300
바이어스가 제거된 결합 벡터들 재결합하고, 바이어스가 제거된 영역별 TRR(Target Region Ratio) 벡터들 영역별로 취합하여 출력
S8400
종 료