(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0044170
(43) 공개일자 2025년03월31일

(51) 국제특허분류(Int. Cl.)
G16B 40/00 (2019.01) *C12Q 1/6827* (2018.01)
C12Q 1/6883 (2018.01) *G16B 25/30* (2019.01)
G16B 45/00 (2019.01) *G16B 50/00* (2019.01)
(52) CPC특허분류
G16B 40/00 (2019.02)
C12Q 1/6827 (2018.05)
(21) 출원번호 10-2024-0128488
(22) 출원일자 2024년09월23일
심사청구일자 2024년09월23일
(30) 우선권주장
1020230127012 2023년09월22일 대한민국(KR)

(71) 출원인
인천대학교 산학협력단
인천광역시 연수구 갯벌로 27, 인천대학교 이노베이션센터(송도동)
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
한미령
인천광역시 연수구 아카데미로 119, 인천대학교(송도동)
황태원
서울특별시 서대문구 연세로 50, 연세대학교(신촌동)
양현호
인천광역시 연수구 아카데미로 119, 인천대학교(송도동)
(74) 대리인
특허법인지평

전체 청구항 수 : 총 12 항

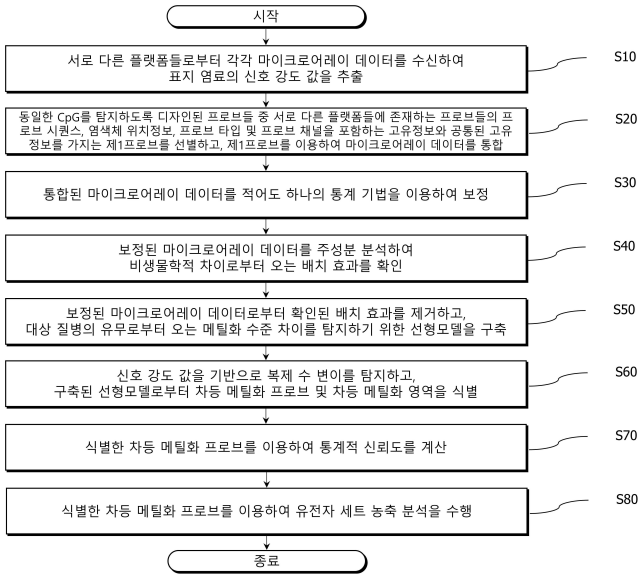
(54) 발명의 명칭 **DNA 메틸레이션 마이크로어레이 데이터의 분석을 수행하는 전자장치 및 분석 방법**

(57) 요 약

본 발명의 일 실시예에 따른 전자장치에 의해 수행되는 DNA 메틸레이션 마이크로어레이 데이터의 분석 방법에 있어서, 서로 다른 플랫폼들로부터 각각 마이크로어레이 데이터를 수신하여 표지 염료의 신호 강도 값을 추출하는 단계; 동일한 시토신-구아닌 디뉴클레오타이드(cytosine-guanine dinucleotides; CpG)를 탐지하도록 디자인된

(뒷면에 계속)

대 표 도 - 도2



프로브들 중 상기 서로 다른 플랫폼들에 존재하는 프로브들의 프로브 시퀀스, 염색체 위치정보, 프로브 타입 및 프로브 채널을 포함하는 고유정보와 공통된 고유정보를 가지는 제1프로브를 선별하는 단계;상기 제1프로브를 이용하여 상기 마이크로어레이 데이터를 통합하는 단계; 상기 통합된 마이크로어레이 데이터를 적어도 하나의 통계기법을 이용하여 보정하는 단계; 상기 보정된 마이크로어레이 데이터를 주성분 분석하여 비생물학적 차이로부터 오는 배치 효과를 확인하는 단계; 상기 보정된 마이크로어레이 데이터로부터 확인된 배치 효과를 제거하고, 대상 질병의 유무로부터 오는 메틸화 수준 차이를 탐지하기 위한 선형모델을 구축하는 단계; 상기 신호 강도 값을 기반으로 복제 수 변이(Copy-number variation, CNV)를 탐지하고, 상기 구축된 선형모델로부터 차등 메틸화 프로브(differentially methylated probe, DMP) 및 차등 메틸화 영역(differentially methylated region, DMR)을 식별하는 단계; 상기 식별한 차등 메틸화 프로브를 이용하여 통계적 신뢰도를 계산하는 단계; 상기 식별한 차등 메틸화 프로브를 이용하여 유전자 세트 농축 분석을 수행하는 단계;를 포함할 수 있다.

(52) CPC특허분류
C12Q 1/6883 (2022.01)
G16B 25/30 (2019.02)
G16B 45/00 (2019.02)
G16B 50/00 (2019.02)
C12Q 2600/154 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711182889
과제번호	2020R1C1C1012288
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	영상 유전체 통합 빅데이터 분석을 활용한 여성암 정밀의료 기반 구축
과제수행기관명	인천대학교
연구기간	2020.03.01 ~ 2025.02.28

명 세 서

청구범위

청구항 1

전자장치에 의해 수행되는 DNA 메틸레이션 마이크로어레이 데이터의 분석 방법에 있어서,

서로 다른 플랫폼들로부터 각각 마이크로어레이 데이터를 수신하여 표지 염료의 신호 강도 값을 추출하는 단계;

동일한 시토신-구아닌 디뉴클레오타이드(cytosine-guanine dinucleotides; CpG)를 탐지하도록 디자인된 프로브들 중 상기 서로 다른 플랫폼들에 존재하는 프로브들의 프로브 시퀀스, 염색체 위치정보, 프로브 타입 및 프로브 채널을 포함하는 고유정보와 공통된 고유정보를 가지는 제1프로브를 선별하는 단계;

상기 제1프로브를 이용하여 상기 마이크로어레이 데이터를 통합하는 단계;

상기 통합된 마이크로어레이 데이터를 적어도 하나의 통계 기법을 이용하여 보정하는 단계;

상기 보정된 마이크로어레이 데이터를 주성분 분석하여 비생물학적 차이로부터 오는 배치 효과를 확인하는 단계;

상기 보정된 마이크로어레이 데이터로부터 확인된 배치 효과를 제거하고, 대상 질병의 유무로부터 오는 메틸화 수준 차이를 탐지하기 위한 선형모델을 구축하는 단계;

상기 신호 강도 값을 기반으로 복제 수 변이(Copy-number variation, CNV)를 탐지하고, 상기 구축된 선형모델로부터 차등 메틸화 프로브(differentially methylated probe, DMP) 및 차등 메틸화 영역(differentially methylated region, DMR)을 식별하는 단계;

상기 식별한 차등 메틸화 프로브를 이용하여 통계적 신뢰도를 계산하는 단계;

상기 식별한 차등 메틸화 프로브를 이용하여 유전자 세트 농축 분석을 수행하는 단계;를 포함하는 분석 방법.

청구항 2

제1항에 있어서,

상기 보정하는 단계는,

상기 통합된 마이크로어레이 데이터 내 복수의 샘플 그룹 중 일정 비율의 하위 샘플 그룹들을 기 정의된 횟수만큼 무작위 복원 반복 추출하는 단계;

상기 하위 샘플 그룹들 내 각 샘플에 대해 측정된 마할라노비스 거리(Mahalanobis distance)를 기반으로 이상치를 탐지하는 단계;를 포함하는 분석 방법.

청구항 3

제1항에 있어서,

상기 배치 효과를 확인하는 단계는,

상기 보정된 마이크로어레이 데이터에서 실험군 데이터와 대조군 데이터의 설정 상태를 디스플레이하는 단계;를 포함하는 분석 방법.

청구항 4

제1항에 있어서,

상기 배치 효과를 제거하는 단계는,

상기 확인된 배치 효과가 알려진 경우인지 및 동일한 배치 효과를 가지는 샘플이 복수 개 존재하는지 여부를 판단하는 단계;

상기 판단 결과에 따라 상기 배치 효과를 제거할 알고리즘을 선택하는 단계;를 포함하는 분석 방법.

청구항 5

제1항에 있어서,

상기 마이크로어레이 데이터를 통합하는 단계 이후에,

상기 통합된 마이크로어레이 데이터에 포함된 프로브들을 신호 강도 값, 비드 개수, p-값, CpG 사이트, 성 염색체 연관 여부 및 탐지 재현률 중 적어도 하나에 기초하여 필터링하는 단계;를 포함하는 분석 방법.

청구항 6

제5항에 있어서,

상기 필터링하는 단계는,

상기 프로브들에 대한 분석 여부를 설정하는 사용자 입력을 수신하는 단계를 포함하는 분석 방법.

청구항 7

DNA 메틸레이션 마이크로어레이 데이터의 분석을 수행하는 전자장치에 있어서,

서로 다른 플랫폼들로부터 각각 마이크로어레이 데이터를 수신하여 표지 염료의 신호 강도 값을 추출하고,

동일한 시토신-구아닌 디뉴클레오타이드(cytosine-guanine dinucleotides; CpG)를 탐지하도록 디자인된 프로브들 중 상기 서로 다른 플랫폼들에 존재하는 프로브들의 프로브 시퀀스, 염색체 위치정보, 프로브 타입 및 프로브 채널을 포함하는 고유정보와 공통된 고유정보를 가지는 제1프로브를 선별하고,

상기 제1프로브를 이용하여 상기 마이크로어레이 데이터를 통합하고,

상기 통합된 마이크로어레이 데이터를 적어도 하나의 통계 기법을 이용하여 보정하고,

상기 보정된 마이크로어레이 데이터를 주성분 분석하여 비생물학적 차이로부터 오는 배치 효과를 확인하고,

상기 보정된 마이크로어레이 데이터로부터 확인된 배치 효과를 제거하고, 대상 질병의 유무로부터 오는 메틸화 수준 차이를 탐지하기 위한 선형모델을 구축하고,

상기 신호 강도 값을 기반으로 복제 수 변이(Copy-number variation, CNV)를 탐지하고, 상기 구축된 선형모델로부터 차등 메틸화 프로브(differentially methylated probe, DMP) 및 차등 메틸화 영역(differentially methylated region, DMR)을 식별하고,

상기 식별한 차등 메틸화 프로브를 이용하여 통계적 신뢰도를 계산하고,

상기 식별한 차등 메틸화 프로브를 이용하여 유전자 세트 농축 분석을 수행하는 프로세서를 포함하는 전자장치.

청구항 8

제7항에 있어서,

상기 프로세서는,

상기 통합된 마이크로어레이 데이터 내 복수의 샘플 그룹 중 일정 비율의 하위 샘플 그룹들을 기 정의된 횟수만큼 무작위 복원 반복 추출하고,

상기 하위 샘플 그룹들 내 각 샘플에 대해 측정된 마할라노비스 거리(Mahalanobis distance)를 기반으로 이상치를 탐지하는 전자장치.

청구항 9

제7항에 있어서,

상기 프로세서는,

상기 보정된 마이크로어레이 데이터에서 실험군 데이터와 대조군 데이터의 설정 상태를 디스플레이하는 전자장

치.

청구항 10

제7항에 있어서,

상기 프로세서는,

상기 확인된 배치 효과가 알려진 경우인지 및 동일한 배치 효과를 가지는 샘플이 복수 개 존재하는지 여부를 판단하고,

상기 판단 결과에 따라 상기 배치 효과를 제거할 알고리즘을 선택하는 전자장치.

청구항 11

제7항에 있어서,

상기 프로세서는,

상기 통합된 마이크로어레이 데이터에 포함된 프로브들을 신호 강도 값, 비드 개수, p-값, CpG 사이트, 성 염색체 연관 여부 및 탐지 재현률 중 적어도 하나에 기초하여 필터링하는 전자장치.

청구항 12

제11항에 있어서,

상기 프로세서는,

상기 프로브들에 대한 분석 여부를 설정하는 사용자 입력을 수신하는 전자장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 DNA 메틸레이션 마이크로어레이 데이터의 분석을 수행하는 전자장치 및 분석 방법에 관한 것이다.

배 경 기 술

[0002] DNA 메틸레이션(DNA methylation; DNAm)이란 DNA 메틸 전이 효소가 DNA 염기에 메틸기를 추가하는 현상을 뜻하고 인간을 포함한 포유류에서는 시토신-구아닌 디뉴클레오타이드(cytosine-guanine dinucleotides; CpG)의 시토신의 5번 탄소에 주로 전이된다. 이는 세포의 분화, X염색체 비활성화, 배아 발달 등에 중요한 역할을 하는 것으로 알려져 있다. 메틸화된 DNA는 RNA 중합효소가 인식할 수 없는 구조가 되어 유전자 발현이 조절된다. 즉, 질병과 연관된 유전자의 메틸화 수준에 변화가 생겨 유전자 발현이 바뀌면 발병의 한 원인으로 작용할 수 있다.

[0003] DNA 메틸화 수준의 변이는 여러 암, 제2형 당뇨, 그리고 신경 정신질환에서 잘 알려진 특성 중 하나이다. DNA 메틸레이션은 특히 개인습관이나 주변 환경적인 요소에 큰 영향을 받을 수 있다. 이를 통해 특정 유전자 좌위의 메틸화 수준을 측정하여 유전자 서열의 변형 없이 해당 유전자의 발현이 어떠한 영향을 받았는지 예측할 수 있다. 특정 유전자 좌위의 메틸화 수준은 유전자 좌위를 포함하는 서열을 가진 프로브(프로브)에 결합하는 뉴클레오타이드가 방출하는 표지 염료, 예를 들어 cy3 및 cy5 염료의 신호 강도를 측정함으로써 계산할 수 있다.

[0004] 현재까지 DNA 메틸레이션 마이크로어레이 데이터를 분석하는 다양하고 새로운 방법들이 제시되어왔다. 2022년 11월 일루미나 사(Illumina)에서 약 93만개의 유전자 좌위에서 메틸화 수준을 측정할 수 있는 마이크로어레이 분석 플랫폼(Infinium Methylation EPIC v2.0)을 발표하였다. 따라서, 새롭게 발표된 알고리즘들과 분석 플랫폼을 쉽게 다루기 위한 최신 파이프라인의 개발이 필요한 실정이다.

발명의 내용

해결하려는 과제

[0005] 본 발명의 목적은 DNA 메틸레이션 마이크로어레이 데이터를 보다 빠르고 정확하게 분석할 수 있는 전자장치 및 그 분석 방법을 제공하는 것이다.

과제의 해결 수단

[0006] 본 발명의 일 실시예에 따른 전자장치에 의해 수행되는 DNA 메틸레이션 마이크로어레이 데이터의 분석 방법에 있어서, 서로 다른 플랫폼들로부터 각각 마이크로어레이 데이터를 수신하여 표지 염료의 신호 강도 값을 추출하는 단계; 동일한 시토신-구아닌 디뉴클레오타이드(cytosine-guanine dinucleotides; CpG)를 탐지하도록 디자인된 프로브들 중 상기 서로 다른 플랫폼들에 존재하는 프로브들의 프로브 시퀀스, 염색체 위치정보, 프로브 타입 및 프로브 채널을 포함하는 고유정보와 공통된 고유정보를 가지는 제1프로브를 선별하는 단계; 상기 제1프로브를 이용하여 상기 마이크로어레이 데이터를 통합하는 단계; 상기 통합된 마이크로어레이 데이터를 적어도 하나의 통계 기법을 이용하여 보정하는 단계; 상기 보정된 마이크로어레이 데이터를 주성분 분석하여 비생물학적 차이로부터 오는 배치 효과를 확인하는 단계; 상기 보정된 마이크로어레이 데이터로부터 확인된 배치 효과를 제거하고, 대상 질병의 유무로부터 오는 메틸화 수준 차이를 탐지하기 위한 선형모델을 구축하는 단계; 상기 신호 강도 값을 기반으로 복제 수 변이(Copy-number variation, CNV)를 탐지하고, 상기 구축된 선형모델로부터 차등 메틸화 프로브(differentially methylated probe, DMP) 및 차등 메틸화 영역(differentially methylated region, DMR)을 식별하는 단계; 상기 식별한 차등 메틸화 프로브를 이용하여 통계적 신뢰도를 계산하는 단계; 상기 식별한 차등 메틸화 프로브를 이용하여 유전자 세트 농축 분석을 수행하는 단계;를 포함할 수 있다.

[0007] 상기 보정하는 단계는, 상기 통합된 마이크로어레이 데이터 내 복수의 샘플 그룹 중 일정 비율의 하위 샘플 그룹들을 기 정의된 횟수만큼 무작위 복원 반복 추출하는 단계; 상기 하위 샘플 그룹들 내 각 샘플에 대해 측정된 마할라노비스 거리(Mahalanobis distance)를 기반으로 이상치를 탐지하는 단계;를 포함할 수 있다.

[0008] 상기 배치 효과를 확인하는 단계는, 상기 보정된 마이크로어레이 데이터에서 실험군 데이터와 대조군 데이터의 설정 상태를 디스플레이하는 단계;를 포함할 수 있다.

[0009] 상기 배치 효과를 제거하는 단계는, 상기 확인된 배치 효과가 알려진 경우인지 및 동일한 배치 효과를 가지는 샘플이 복수 개 존재하는지 여부를 판단하는 단계; 상기 판단 결과에 따라 상기 배치 효과를 제거할 알고리즘을 선택하는 단계;를 포함할 수 있다.

[0010] 상기 마이크로어레이 데이터를 통합하는 단계 이후에, 상기 통합된 마이크로어레이 데이터에 포함된 프로브들을 신호 강도 값, 비드 개수, p-값, CpG 사이트, 성 염색체 연관 여부 및 탐지 재현률 중 적어도 하나에 기초하여 필터링하는 단계;를 포함할 수 있다.

[0011] 상기 필터링하는 단계는, 상기 프로브들에 대한 분석 여부를 설정하는 사용자 입력을 수신하는 단계를 포함할 수 있다.

[0012] 본 발명의 일 실시예에 따른 DNA 메틸레이션 마이크로어레이 데이터의 분석을 수행하는 전자장치에 있어서, 서로 다른 플랫폼들로부터 각각 마이크로어레이 데이터를 수신하여 표지 염료의 신호 강도 값을 추출하고, 동일한 시토신-구아닌 디뉴클레오타이드(cytosine-guanine dinucleotides; CpG)를 탐지하도록 디자인된 프로브들 중 상기 서로 다른 플랫폼들에 존재하는 프로브들의 프로브 시퀀스, 염색체 위치정보, 프로브 타입 및 프로브 채널을 포함하는 고유정보와 공통된 고유정보를 가지는 제1프로브를 선별하고, 상기 제1프로브를 이용하여 상기 마이크로어레이 데이터를 통합하고, 상기 통합된 마이크로어레이 데이터를 적어도 하나의 통계 기법을 이용하여 보정하고, 상기 보정된 마이크로어레이 데이터를 주성분 분석하여 비생물학적 차이로부터 오는 배치 효과를 확인하고, 상기 보정된 마이크로어레이 데이터로부터 확인된 배치 효과를 제거하고, 대상 질병의 유무로부터 오는 메틸화 수준 차이를 탐지하기 위한 선형모델을 구축하고, 상기 신호 강도 값을 기반으로 복제 수 변이(Copy-number variation, CNV)를 탐지하고, 상기 구축된 선형모델로부터 차등 메틸화 프로브(differentially methylated probe, DMP) 및 차등 메틸화 영역(differentially methylated region, DMR)을 식별하고, 상기 식별한 차등 메틸화 프로브를 이용하여 통계적 신뢰도를 계산하고, 상기 식별한 차등 메틸화 프로브를 이용하여 유전자 세트 농축 분석을 수행하는 프로세서를 포함할 수 있다.

[0013] 상기 프로세서는, 상기 통합된 마이크로어레이 데이터 내 복수의 샘플 그룹 중 일정 비율의 하위 샘플 그룹들을 기 정의된 횟수만큼 무작위 복원 반복 추출하고, 상기 하위 샘플 그룹들 내 각 샘플에 대해 측정된 마할라노비스 거리(Mahalanobis distance)를 기반으로 이상치를 탐지할 수 있다.

[0014] 상기 프로세서는, 상기 보정된 마이크로어레이 데이터에서 실험군 데이터와 대조군 데이터의 설정 상태를 디스플레이 할 수 있다.

[0015] 상기 프로세서는, 상기 확인된 배치 효과가 알려진 경우인지 및 동일한 배치 효과를 가지는 샘플이 복수 개 존

재하는지 여부를 판단하고, 상기 판단 결과에 따라 상기 배치 효과를 제거할 알고리즘을 선택할 수 있다.

[0016] 상기 프로세서는, 상기 통합된 마이크로어레이 데이터에 포함된 프로브들을 신호 강도 값, 비드 개수, p-값, CpG 사이트, 성 염색체 연관 여부 및 탐지 재현률 중 적어도 하나에 기초하여 필터링할 수 있다.

[0017] 상기 프로세서는, 상기 프로브들에 대한 분석 여부를 설정하는 사용자 입력을 수신할 수 있다.

발명의 효과

[0018] 본 발명의 일 실시예에 따르면, 다양한 플랫폼에서 수신하는 마이크로어레이 데이터를 통합하여 보다 정확하고 빠르게 분석할 수 있다.

[0019] 본 발명의 일 실시예에 따르면, 실험군과 대조군 설정을 보여줌으로써 오입력을 방지하고 분석 단계 중간에 실험군과 대조군을 자유롭게 바꾸어 비교할 수 있다.

[0020] 본 발명의 일 실시예에 따르면, 무작위 표본을 반복적으로 추출하는 방법을 통해 이상치 존재 여부에 영향을 크게 받는 마할라노비스 거리 값의 단점을 보완하고 계산량을 줄여 빠르게 탐색 가능하다.

도면의 간단한 설명

[0021] 도 1은 본 발명의 일 실시예에 따른 전자장치의 구성을 도시한 블럭도이다.

도 2는 본 발명의 제1실시예에 따른 DNA 메틸레이션 마이크로어레이 데이터의 분석 방법을 도시한 도면이다.

도 3은 본 발명의 제2실시예에 따른 DNA 메틸레이션 마이크로어레이 데이터의 분석 방법을 도시한 도면이다.

도 4는 본 발명의 제1실시예에 따른 분석 결과를 도시한 도면이다.

도 5는 본 발명의 제2실시예에 따른 분석 결과를 도시한 도면이다.

도 6은 본 발명의 제3실시예에 따른 분석 결과를 도시한 도면이다.

도 7은 본 발명의 제4실시예에 따른 분석 결과를 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0022] 이하, 본 발명에 따른 바람직한 실시 형태를 첨부된 도면을 참조하여 상세하게 설명한다. 첨부된 도면과 함께 이하에 개시될 상세한 설명은 본 발명의 예시적인 실시형태를 설명하고자 하는 것이며, 본 발명이 실시될 수 있는 유일한 실시형태를 나타내고자 하는 것이 아니다. 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략할 수 있고, 명세서 전체를 통하여 동일 또는 유사한 구성 요소에 대해서는 동일한 참조 부호를 사용할 수 있다.

[0023] 본 발명의 일 실시예에 따른 DNA 메틸레이션 마이크로어레이 데이터(이하, 마이크로어레이 데이터라고도 한다.)의 분석 파이프라인의 주된 목적은 분석가가 편리하게 다양한 분석 알고리즘을 사용하도록 하나의 패키지(package)에 여러 알고리즘을 포함시킴에 있다. 또한, 편의성을 고려한 코드를 추가하여 분석에 소요되는 시간을 줄이거나 간단한 상호작용으로 복잡한 데이터 구조 변경이 가능하도록 설계하며, 이와 동시에 각 함수에 다양한 선택지를 제공함으로써 복잡하고 전문적인 분석을 보다 편하게 실행하는 것에 초점을 맞추었다.

[0024] 본 발명의 일 실시예에 따른 마이크로어레이 데이터의 분석은 전자장치(100)에 의해 수행될 수 있다. 전자장치(100)는 컴퓨터, 서버, 스마트 폰, 태블릿 PC, 스마트 패드, 노트북 등으로 구현될 수 있다.

[0025] 이하, 도면들을 참조하여 본 발명의 일 실시예에 따른 전자장치의 구성 및 동작에 대해 구체적으로 설명한다.

[0026] 도 1은 본 발명의 일 실시예에 따른 전자장치의 구성을 도시한 블럭도이다.

[0027] 본 발명의 일 실시예에 따른 전자장치(100)는 입력부(110), 통신부(120), 표시부(130), 저장부(140) 및 프로세서(150)를 포함한다.

[0028] 입력부(110)는 전자장치(100)의 사용자 입력에 대응하여 입력데이터를 발생시킨다. 예를 들어, 사용자 입력은 전자장치(100)의 동작을 시작하게 하는 사용자 입력, 동일한 CpG를 탐지하도록 디자인된 프로브들 중 품질 높은 프로브들을 선별하기 위한 사용자 입력, 하위 샘플 그룹의 비율이나 반복 추출 횟수를 설정하는 사용자 입력, 실험군 데이터와 대조군 데이터를 설정하는 사용자 입력 등일 수 있으며, 이 외에도 마이크로어레이 데이터의

분석을 위해 필요한 사용자 입력인 경우 제한하지 않고 적용 가능하다.

[0029] 입력부(110)는 적어도 하나의 입력수단을 포함한다. 입력부(110)는 키보드(key board), 키패드(key pad), 돔 스위치(dome switch), 터치패널(touch panel), 터치 키(touch key), 마우스(mouse), 메뉴 버튼(menu button) 등을 포함할 수 있다.

[0030] 통신부(120)는 마이크로어레이 데이터, 분석 결과 등을 송수신하기 위해 분석 플랫폼, 서버 등 외부장치와의 통신을 수행한다.

[0031] 이를 위해, 통신부(120)는 5G(5th generation communication), LTE-A(long term evolution-advanced), LTE(long term evolution), Wi-Fi(wireless fidelity), Bluetooth 등의 무선 통신, 혹은 LAN(local area network), WAN(Wide Area Network), 전력선 통신 등의 유선 통신을 수행할 수 있다.

[0032] 표시부(130)는 전자장치(100)의 동작에 따른 표시 데이터를 표시한다. 표시부(130)는 분석 과정이나 분석 결과를 표시하는 화면, 사용자 입력을 수신하는 화면 등을 표시할 수 있다.

[0033] 표시부(130)는 액정 디스플레이(LCD; liquid crystal display), 발광 다이오드(LED; light emitting diode) 디스플레이, 유기 발광 다이오드(OLED; organic LED) 디스플레이, 마이크로 전자기계 시스템(MEMS; micro electro mechanical systems) 디스플레이 및 전자 종이(electronic paper) 디스플레이를 포함한다. 표시부(130)는 입력부(110)와 결합되어 터치 스크린(touch screen)으로 구현될 수 있다.

[0034] 저장부(140)는 전자장치(100)의 동작 프로그램들을 저장한다. 저장부(140)는 전원의 제공 유무와 무관하게 데이터(정보)를 보존할 수 있는 비휘발성 속성의 스토리지(storage)와, 프로세서(150)에 의해 처리되기 위한 데이터가 로딩되며 전원이 제공되지 않으면 데이터를 보존할 수 없는 휘발성 속성의 메모리(memory)를 포함한다. 스토리지에는 플래시메모리(flash-memory), HDD(hard-disc drive), SSD(solid-state drive) ROM(Read Only Memory) 등이 있으며, 메모리에는 버퍼(buffer), 램(RAM; Random Access Memory) 등이 있다.

[0035] 저장부(140)는 마이크로어레이 데이터, 분석 과정에서 필요한 알고리즘, 함수 소스 코드 등을 저장할 수 있다. 저장부(140)는 신호 강도 값 추출, 프로브 선별, 마이크로어레이 데이터 통합, 보정, 배치 효과 확인 및 제거, 선형 모델 구축, 차등 메틸화 프로브(differentially methylated probe, DMP) 및 차등 메틸화 영역(differentially methylated region, DMR) 식별 등을 수행하는 과정에서 필요한 연산 프로그램 등을 저장할 수 있다.

[0036] 프로세서(150)는 프로그램 등 소프트웨어를 실행하여 전자장치(100)의 적어도 하나의 다른 구성요소(예: 하드웨어 또는 소프트웨어 구성요소)를 제어할 수 있고, 다양한 데이터 처리 또는 연산을 수행할 수 있다.

[0037] 본 발명의 일 실시예에 따른 프로세서(150)는 서로 다른 플랫폼들로부터 각각 마이크로어레이 데이터를 수신하여 표지 염료의 신호 강도 값을 추출하고, 동일한 시토신-구아닌 디뉴클레오타이드(cytosine-guanine dinucleotides; CpG)를 탐지하도록 디자인된 프로브들 중 상기 서로 다른 플랫폼들에 존재하는 프로브들의 프로브 시퀀스, 염색체 위치정보, 프로브 타입 및 프로브 채널을 포함하는 고유정보와 공통된 고유정보를 가지는 제1프로브를 선별하고, 상기 제1프로브를 이용하여 상기 마이크로어레이 데이터를 통합하고, 상기 통합된 마이크로어레이 데이터를 적어도 하나의 통계 기법을 이용하여 보정하고, 상기 보정된 마이크로어레이 데이터를 주성분 분석하여 대조군 데이터의 주성분과의 비생물학적 차이로부터 오는 배치 효과를 확인하고, 상기 보정된 마이크로어레이 데이터로부터 확인된 배치 효과를 제거하고, 대상 질병의 유무로부터 오는 메틸화 수준 차이를 탐지하기 위한 선형모델을 구축하고, 상기 신호 강도 값을 기반으로 CNV를 탐지하고, 상기 구축된 선형모델로부터 DMP 및 DMR을 식별하고, 상기 식별한 차등 메틸화 프로브를 이용하여 통계적 신뢰도를 계산하고, 상기 식별한 차등 메틸화 프로브를 이용하여 유전자 세트 농축 분석을 수행할 수 있다.

[0038] 한편, 프로세서(150)는 상기 동작들을 수행하기 위한 데이터 분석, 처리, 및 결과 정보 생성 중 적어도 일부를 규칙 기반 또는 인공지능(Artificial Intelligence) 알고리즘으로서 기계학습, 신경망 네트워크(neural network), 또는 딥러닝 알고리즘 중 적어도 하나를 이용하여 수행할 수 있다. 신경망 네트워크의 예로는, CNN (Convolutional Neural Network), DNN (Deep Neural Network), RNN (Recurrent Neural Network), Vision Transformer과 같은 모델을 포함할 수 있다.

[0039] 도 2는 본 발명의 제1실시예에 따른 DNA 메틸레이션 마이크로어레이 데이터의 분석 방법을 도시한 도면이다.

[0040] 분석 방법은 서로 다른 플랫폼들로부터 각각 마이크로어레이 데이터를 수신하여 표지 염료의 신호 강도 값을 추

출하는 단계(S10)를 포함할 수 있다.

[0041] 서로 다른 플랫폼들은, 예를 들어, 일루미나 사에서 배포하고 있는 플랫폼일 수 있으며, 버전이 상이한 경우도 포함한다. 프로세서(150)는 서로 다른 플랫폼들을 통해 마이크로어레이 데이터를 수신할 수 있다. 프로세서(150)는 마이크로어레이 데이터를 포함하는 파일 크기, 고유 바코드 넘버, 어레이 유형(array type) 등 분석에 불필요한 내용은 저장하지 않고, 가장 중요한 각 프로브에 대한 비드 개수(NBead)와 평균 신호 강도(Mean)를 추출하여 각 파일 별로 저장할 수 있다.

[0042] 프로세서(150)는 마이크로어레이 데이터로부터 표지 염료, 예를 들어 cy3 및 cy5 염료의 신호 강도 값을 추출할 수 있다. 프로세서(150)는 마이크로어레이 데이터를 플랫폼 전용 스캐너(scanner)로 스캔하여 R 언어(language) 환경에 불러들일 수 있는 함수(MeCall.ReadIdat)를 이용하여 신호 강도 값을 추출할 수 있다.

[0043] 분석 방법은 동일한 시토신-구아닌 디뉴클레오타이드(cytosine-guanine dinucleotides; CpG)를 탐지하도록 디자인된 프로브들 중 상기 서로 다른 플랫폼들에 존재하는 프로브들의 프로브 시퀀스, 염색체 위치정보, 프로브 타입 및 프로브 채널을 포함하는 고유정보와 공통된 고유정보를 가지는 제1프로브를 선별하고, 제1프로브를 이용하여 마이크로어레이 데이터를 통합하는 단계(S20)를 포함할 수 있다.

[0044] 앞서 서술한 바와 같이, 마이크로어레이 데이터는 여러 플랫폼들로부터 수신할 수 있고, 서로 다른 플랫폼은 다른 개수의 프로브들을 탐지할 수 있다. 이때, 플랫폼 고유의 프로브들이 존재할 수 있고, 여러 플랫폼들 간에 중복되는 프로브들도 존재한다. 연구자가 이전부터 모아온 방대한 양의 데이터를 한 번에 분석하기 위해 서로 다른 플랫폼의 데이터를 합쳐 분석하는 경우, 통계 분석을 위해 이를 통합할 필요성이 있다.

[0045] 한편, 동일 CpG를 탐지하는 프로브라 하더라도 검출 품질(detection quality)을 위해서 여러 프로브에서 이를 탐지하도록 디자인된 경우가 있다. 따라서, 동일 CpG를 탐지하는 여러 프로브들의 값이 모두 포함된 데이터 테이블을 얻게 되면, 사용자는 해당 프로브들을 찾아서 전부 제거해야 사용이 가능한 번거로움이 발생한다. 만일, 하나의 CpG를 탐지한 여러 프로브들의 신호 강도 값들이 필터링을 거치지 않고 통계분석에 모두 포함되면 분석에 편향을 발생시킬 수 있다.

[0046] 따라서, 본 발명의 일 실시예에서, 프로세서(150)는 함수(MeCall.ShiftArray)를 통해 중복된 프로브들을 자동적으로 제거할 수 있다. 구체적으로, 프로세서(150)는 동일 CpG를 탐지하는 프로브들 중 상기 서로 다른 플랫폼들에 존재하는 프로브들의 프로브 시퀀스, 염색체 위치정보, 프로브 타입 및 프로브 채널을 포함하는 고유정보와 공통된 고유정보를 가지는 제1프로브를 선별할 수 있다. 이 외에도, 프로세서(150)는 특이도 및 민감도가 우수한 것으로 밝혀진 제1프로브를 선별하는 사용자입력을 수신할 수 있다.

[0047] 이 외에도, 프로세서(150)는 탐지 품질이 낮거나, 분석과 연관이 없는 정보들을 삭제하여 데이터의 품질을 높이는 필터링을 수행할 수 있다. 이하, 필터링을 수행하는 기법은 도 3을 참조하여 설명한다.

[0048] 분석 방법은 통합된 마이크로어레이 데이터를 적어도 하나의 통계 기법을 이용하여 보정하는 단계(S30)를 포함할 수 있다.

[0049] CpG 사이트(site)의 메틸화 수준(methylation level)을 측정하기 위해 존재하는 다른 2가지 타입의 프로브로부터 오는 신호 강도의 경향성의 차이나, 마이크로어레이 데이터 자체에 포함된 노이즈(noise)들을 통계적인 방법으로 보정할 수 있다. 그 구체적인 통계 기법은 다음과 같다.

[0050] 1. 프로세서(150)는 제어 프로브(control 프로브)의 분포와 각 두 메틸화 채널과 연관된 한계 분포(marginal distribution)의 변동성을 제거하여 정규화를 진행할 수 있다(Functional normalization).

[0051] 2. 프로세서(150)는 비메틸화 채널(Unmethylated channel)과 메틸화 채널(methylated channel)에서 각각 프로브 내 CpG 함량이 같은 프로브들을 무작위로 뽑아 신호 강도 순으로 하위집합을 생성 및 정렬할 수 있다. 프로세서(150)는 신호 강도의 평균 및 사분위수를 할당하고 선형보간법을 이용하여 2가지 타입의 프로브 간의 차이를 보정할 수 있다(SWAN).

[0052] 3. 프로세서(150)는 메틸레이션 값의 특징인 바이모달(bimodal) 분포를 지수(exponential) 분포와 정규(normal) 분포의 합으로 표현할 수 있다. 프로세서(150)는 관측한 신호 강도 값의 평균 및 분산과 배경 노이즈(background noise)를 측정할 일련의 프로브의 신호 강도 값의 평균 및 분산을 계산 후, 노이즈를 제거한 실제 신호 강도 값을 expectation-maximization (EM) 알고리즘을 활용하여 추정할 수 있다(Enmix).

[0053] 4. 프로세서(150)는 배경 노이즈 조정, 어레이 간 정규화, 염료 편향에 대한 보정 적용여부를 알파벳으로 나타

낼 수 있다. 프로세서(150)는 배경 노이즈 조정은 2가지 타입의 프로브 간의 차이를 표현하는 오프셋(offset) 값을 더하는 것으로 수행하고, 실행한 경우 'd', 그렇지 않은 경우 'n'으로 표현할 수 있다. 어레이 간 정규화는 메틸화 여부를 표현하는 프로브와, 비메틸화 여부를 표현하는 프로브의 샘플간 분위수 정규화(quantile normalization)를 수행하고, 각 타입별 프로브 개별로 적용한 경우 's', 타입에 관계없이 적용한 경우 't', 정규화를 수행하지 않은 경우 'n'으로 표현할 수 있다. 여러 보정 조합 중 가장 보편적으로 사용되는 배경 노이즈 조정, 각 타입별 프로브 개별로 적용된 샘플간 분위수 정규화를 진행을 조합시킨 정규화를 수행할 수 있다(dasen).

[0054] 5. Noob 기법은 관측한 신호 강도 값은 지수 분포를 따르는 실제 신호 강도 값과 정규 분포를 따르는 배경 노이즈 값의 합으로 표현할 수 있음에 기초한다. 배경 노이즈의 값의 분포를 알기 위해 마이크로어레이의 품질 관리를 위해 존재하는 out-of-band(OOB) 프로브들의 신호 강도 값을 이용할 수 있다. 노이즈가 없는 실제 신호 강도 값을 알기 위해 OOB 프로브의 평균 및 표준 편차 값을 적용한 표준 정규분포와 정규 누적 분포를 이용하여 주어진 관측된 신호 강도 값의 조건 하 실제 신호 강도 값을 추정할 수 있다.

[0055] 6. 프로세서(150)는 주어진 베타 값 메트릭스(beta-value matrix)의 모든 프로브들을 프로브 타입 별로 최대우도법을 통해 비메틸화(U), 반-메틸화(hemi-methylated)(H), 그리고 메틸화(M)의 세가지 상태로 분류할 수 있다. 각 프로브 타입(타입 1, 타입 2) 및 메틸레이션 상태(U, H, M)의 (2x3=6개) 프로브 집합마다 EM 알고리즘을 통해서 베타 분포(beta distribution)의 매개변수를 추정할 수 있다. U와 M 그룹에 한해서 프로브 타입 2의 프로브들이 가지는 베타 값을 해당 메틸레이션 상태에 대한 매개변수를 통해 생성한 누적 베타 분포(cumulative beta distribution) 함수에 입력해 존재 확률을 얻고 이 확률을 반대로 같은 메틸레이션 상태를 가졌으나 프로브 타입 1 그룹으로부터 추정된 매개변수를 통해 생성한 누적 베타 분포 함수의 역함수에 입력하여 정규화된 베타 값을 계산할 수 있다. H 그룹에 속하는 프로브들의 베타 값을 정규화 하기 위해 정규화 되기 전의 U 및 M 상태의 베타 값과 정규화 된 U 및 M 상태의 베타 값의 최댓값 및 최솟값을 각각 구하여 이들의 차이를 가지고 H 상태의 프로브의 베타 값들을 정규화할 수 있다(BMIQ).

[0056] 분석 방법은 보정된 마이크로어레이 데이터를 주성분 분석하여 비생물학적 차이에 따른 배치 효과를 확인하는 단계(S40)를 포함할 수 있다.

[0057] 배치 효과(batch effect)는 임상적 차이가 아닌 다른 실험 환경으로부터 오는 차이로써 실제로 데이터에 유의미한 영향을 미치고 있는지 확인할 수 있다.

[0058] 프로세서(150)는 앞서 정규화까지 전처리가 진행된 마이크로어레이 데이터를 주성분 분석(principal component analysis, PCA)하여 얻어진 주성분과 알고 있는 배치 효과들과의 상관관계분석(correlation analysis)을 통해 배치 효과를 확인할 수 있다.

[0059] 한편, 본 단계를 진행하는 동안 프로세서(150)는 보정된 마이크로어레이 데이터에서 실험군 데이터와 대조군 데이터의 설정 상태를 디스플레이할 수 있다. 디스플레이의 형태는, 간단한 메시지로 출력할 수 있으며, 디스플레이 방식은 어느 하나에 한정되지 않는다. 이를 통해, 사용자는 비교할 데이터가 제대로 입력 되었는지 확인할 수 있다.

[0060] 이 과정은 만일 실험군 데이터와 대조군 데이터가 서로 반대로 설정되어 있는 경우, 대조군 대비 실험군의 특성을 파악해야 할 분석이 실험군 대비 대조군의 특성으로 설정되어 최종 결과에서 결과 수치가 정 반대로 나올 수 있기 때문에 중요하다. 메틸레이션은 과메틸화(hyper-methylated)인지 저메틸화(hypo-methylated)인지에 따라 생물학적인 해석이 달라질 수 있기 때문에 이 과정은 중요하다. 추가적으로 단순히 실험군과 대조군이 아닌, 복수 개의 그룹이 동시에 혼재할 경우, 어떤 그룹을 대조군으로 둘지 어떤 그룹을 실험군으로 둘지 간단한 상호작용을 통해 설정할 수 있다. 본 분석 과정의 주된 목적은 두 그룹 간에 차등 메틸화된 영역을 찾는 것 이므로, 복수 개의 그룹이 혼재할 때는 두 그룹으로 축소시키도록 유도할 필요가 있다.

[0061] 분석 방법은 보정된 마이크로어레이 데이터로부터 확인된 배치 효과를 제거하고, 대상 질병의 유무로부터 오는 메틸화 수준 차이를 탐지하기 위한 선형모델을 구축하는 단계(S50)를 포함할 수 있다.

[0062] 프로세서(150)는 앞서 확인된 배치 효과를 다음 3가지 경우에 따라 제거할 수 있다.

[0063] 먼저, 배치 효과를 알고 있고, 가장 효과적으로 제거하는 것으로 알려진 ComBat 알고리즘을 사용할 수 있는 경우에는, ComBat 알고리즘을 적용하여 알려진 배치 효과를 제거할 수 있다.

[0064] 다른 경우는, 배치 효과를 알고 있으나, Combat 알고리즘을 사용할 수 없는 경우이다. ComBat 알고리즘은 배치

효과를 제거하는 데에 있어서 강력한 수단이지만, 하나의 표현형에 최소 2개 이상의 샘플이 존재해야 사용할 수 있다. 예를 들면, 메틸레이션 마이크로 어레이 분석을 할 때 가장 주요한 배치 효과인 Sentrix ID는 반드시 제거해야 하는 것으로 알려져있다. 하지만 같은 Sentrix ID를 가진 샘플이 존재하지 않고 어떤 특정한 Sentrix ID를 가진 샘플이 단 1개라면 ComBat 알고리즘을 이용하여 Sentrix ID로부터 오는 배치 효과는 제거할 수 없다. 이런 경우에는 ComBat 알고리즘의 "mean.only" 인자를 "TRUE"로 놓아 평균에 대해서만 보정이 가능하다. 프로세서(150)는 사용자의 편의성을 높이기 위해 입력된 변수가 어떤 경우에 해당하는지 판단하여, ComBat 알고리즘의 인자를 변경하여 적용시킬 수 있다.

[0065] 마지막 경우는 배치 효과에 관한 데이터를 제공 받을 수 없거나, 배치 효과를 알지 못하는 경우, 배치 효과를 탐색해주는 알고리즘인 SVA(Surrogate Variable Analysis)를 사용할 수 있다. SVA 알고리즘을 적용하여 얻은 대리 변수(surrogate variable) 중 공변량들과 유의미한 상관관계가 없는 변수만 찾아 보정함으로써 공변량들이 데이터에 미치는 영향은 제거하지 않으면서 배치 효과의 영향을 최소화시킬 수 있다.

[0066] 이때, 너무 많은 대리 변수들이 반환되어 데이터가 과하게 보정되는 것을 방지하기 위해, 반환된 대리 변수들과 SVA를 통해 제거되기를 원치 않는 변수들이 가지는 상관관계를 분석하여 해당 변수들과 독립적인 대리 변수들만 선택하여 보정할 수 있다.

[0067] 프로세서(150)는 공변량을 보정하면서 질병 유무로부터 오는 메틸화 수준 차이만을 탐지하기 위해 선형모델을 구축할 수 있다.

[0068] 분석 방법은 신호 강도 값을 기반으로 복제 수 변이(Copy-number variation, CNV)를 탐지하고, 구축된 선형모델로부터 차등 메틸화 프로브(differentially methylated probe, DMP) 및 차등 메틸화 영역(differentially methylated region, DMR)을 식별하는 단계(S60)를 포함할 수 있다.

[0069] 프로세서(150)는 DNA copy, conumee등의 R package에 포함된 함수를 이용하여 cy3 및 cy5 염료 신호 강도 값을 기반으로 복제 수 변이를 탐지할 수 있다.

[0070] 프로세서(150)는 마이크로어레이 데이터의 신호 강도 값을 사용하여 세그먼트(segment) 데이터를 생성하고, 대조군을 참조(reference)로 하여 실험군의 전체 프로브를 염색체상 위치 순으로 나열할 수 있다. 프로세서(150)는 관측값들을 계속해서 더해가면서 영 분포(null distribution)의 a번째 quantile(default = 0.001)이상의 값이 도출되면 변경 점(change point)라고 명명할 수 있다. 전체 유전자에서 이러한 변경 점들을 찾기 위해 circular binary segmentation(CBS) 알고리즘을 이용할 수 있다. 선택적으로 프로세서(150)는 세그먼트 데이터를 입력으로 받아 각 case 샘플 당 복제 수 변이를 나타낸 플롯을 생성할 수 있다.

[0071] 프로세서(150)는 구축한 선형 모델을 기반으로 차등 메틸화 프로브와 차등 메틸화 영역을 탐지할 수 있다.

[0072] 구체적으로, 차등 메틸화 영역을 탐지하기 위해, 프로세서(150)는 다음과 같은 알고리즘들을 이용할 수 있다.

[0073] 1. Bumphunter

[0074] 프로세서(150)는 전체 샘플에서 각 유전자 좌위 별로 메틸레이션 수준 값에 대해 공변량 및 오차를 고려하여 선형 모델의 기울기를 추정하고 이 값들을 평활(smoothing)할 수 있다. 프로세서(150)는 기울기 값이 임계값을 초과하는 부분을 선택하여 잠재적인 차등 메틸화 영역으로 설정할 수 있다. 이후, 부트스트랩핑(Bootstrapping)을 통해 생성한 잠재적인 DMR의 영 분포를 토대로 해당 DMR이 통계적으로 유의미한지 판단할 수 있다.

[0075] 2. Seqlm

[0076] 최소 표현 길이(Minimum Description Length, MDL) 기법을 기반으로 기본 신호를 보존하면서 상관관계가 있는 연속된 사이트를 영역으로 결합하여 더 작은 특징 집합을 찾는 알고리즘이다. 프로세서(150)는 CpG 사이트 간 1000bp를 기준으로 전체 프로브들을 작은 하위 그룹으로 나눌 수 있다. 각 하위 그룹 내에서 가능한 세그먼트들을 조사한 후 MDL점수를 기반으로 가장 긴 세그먼트를 가지면서 메틸레이션 수준 차이를 잘 설명하는 모델을 찾는다. 이후 샘플의 특징적인 메틸레이션 수준 베이스 라인을 고려하여 DMR의 통계적 유의성을 판단한다.

[0077] 3. dmrff

[0078] CpG 사이트 간의 상관관계가 DMR을 탐지하는데 미치는 영향을 고려하기 위해서 역분산 가중 통계(inverse-variance weighted statistics)를 이용하는 DMR 탐지 알고리즘이다. B/S statistics를 사용자가 지정한 하위 영역들에서 계산하여 유의미한 값을 가지는 영역들을 겹치지 않게 greedy 알고리즘으로 연결하여 DMR로 반환할 수 있다.

[0079] 4. DMRcate

[0080] 프로세서(150)는 지정된 폭의 가우시안 커널을 사용하여 인근 CpG 부위의 t-검정(t-statistics)을 결합하여 평활할 수 있다. 그 다음 각 CpG 부위에 대한 통계적 유의성을 평활된 t-검정으로부터 다시 계산하고 게놈 전체 유의성에 대한 새로운 컷 오프(cut off)를 선택하여 컷 오프를 통과하는 CpG 부위의 수가 원래 EWAS에서 다중 테스트 보정에서 살아남은 수와 동일하도록 한다. 재계산된 p-값이 임계값보다 낮은 연속적인 CpG 부위를 DMR로 반환할 수 있다.

[0081] 분석 방법은 식별한 차등 메틸화 프로브를 이용하여 통계적 신뢰도를 계산하는 단계(S70)를 포함할 수 있다.

[0082] 구체적으로, 프로세서(150)는 탐지된 차등 메틸화 프로브가 포함된 주어진 실험 조건을 바탕으로 시뮬레이션을 수행하는 pwrEWAS R package에 포함된 pwrEWAS 함수를 활용하여 각 메틸화 수준의 차이가 갖는 통계적 신뢰도를 계산할 수 있다.

[0083] pwrEWAS는 주어진 실험 조건(샘플 수, 테스트한 프로브의 개수, 잠재적 DMP의 개수, p-값의 임계점 등)에서 시뮬레이션을 통해 각 메틸화 수준의 차이가 갖는 통계적 신뢰도를 계산하고 반환할 수 있다. 본 발명의 분석 방법에 해당 함수를 활용하는 단계를 추가함으로써, 사용자가 더 중요하고 해석할 가치가 높은 차등 메틸화 프로브를 선택하는데 도움을 줄 수 있다.

[0084] 분석 방법은 식별한 차등 메틸화 프로브를 이용하여 유전자 세트 농축 분석을 수행하는 단계(S80)를 포함할 수 있다.

[0085] 구체적으로, 프로세서(150)는 윌레니우스 비중심 초기하 분포(Wallenius non-central hypergeometric distribution)를 사용하여 CpG site와 유전자 간의 복잡한 관계를 고려하는 missMethyl R package에 포함된 gometh 함수를 활용하여 탐지된 차등 메틸화 프로브를 바탕으로 유전자 세트 농축 분석을 수행할 수 있다.

[0086] gometh함수는 윌레니우스 비중심 초기하 분포(Wallenius non-central hypergeometric distribution)를 사용하여 CpG site와 유전자 간의 복잡한 관계를 고려한 유전자 세트 농축 분석을 수행하는 함수이다.

[0087] 유전자 세트 농축 분석은 사용자가 식별한 차등 메틸화 프로브와 연관된 유전자들이 주로 어떤 기능과 연관되어 있는지 확인하여 추가적인 정보를 얻는데 유용하다.

[0088] 본 발명의 일 실시예에 따르면, 비교적 최근에 발표된 알고리즘을 쉽게 이용하도록 함수에 포함시킬 수 있다.

[0089] 본 발명의 일 실시예에 따르면, 최신 매니페스트(manifest) 파일을 이용하여 결과를 도출하고 GRch38/hg38 버전으로 포지션(position)을 제시할 수 있으며, 다른 파이프라인에서는 미지원 하는 플랫폼을 이용한 분석이 가능하다.

[0090] 본 발명의 일 실시예에 따르면, 실험군과 대조군 설정을 보여줌으로써 오입력을 방지하고 분석 단계 중간에 실험군과 대조군을 자유롭게 바꾸어 비교할 수 있다.

[0091] 도 3은 본 발명의 제2실시예에 따른 DNA 메틸레이션 마이크로어레이 데이터의 분석 방법을 도시한 도면이다. 도 3의 분석 방법 중 도 2의 분석 방법과 동일한 단계들은 도 2를 참조하여 서술한 내용을 차용하는 바, 구체적인 설명은 생략한다.

[0092] 분석 방법은 서로 다른 플랫폼들로부터 각각 마이크로어레이 데이터를 수신하여 표지 염료의 신호 강도 값을 추출하는 단계(S401)를 포함할 수 있다.

[0093] 분석 방법은 탐지 품질이 낮거나, 분석과 연관이 없는 정보들을 삭제하여 데이터의 품질을 높이는 필터링을 수행하는 단계(S402)를 포함할 수 있다.

[0094] 필터링을 수행하는 기법은 다음과 같다.

[0095] 1. 저품질 샘플(Bad Quality Sample)

[0096] 프로세서(150)는 minfi R package에 포함되어 있는 fixMethOutliers function과 log2 scale의 신호 강도의 중간값을 활용하여 기 정의된 기준(예: default = 10.5) 보다 낮은 신호 강도 값을 가진 샘플은 품질이 낮은 샘플로 판단하여 분석에서 제외시킬 수 있다.

[0097] 2. 탐지 P-값(Detection P-value)

[0098] 프로세서(150)는 CpG를 탐지하도록 디자인된 프로브의 신호 강도 값이 가지는 p-값을 계산할 수 있다. 이를 위

해 프로세서(150)는 마이크로어레이 데이터의 배경 노이즈를 측정하기 위한 네거티브 제어 프로브(negative control probe)의 신호 강도 값의 확률분포에서 CpG 사이트를 탐지할 수 있다.

[0099] 프로세서(150)는 기 정의된 기준(예: default = 0.01) 보다 높은 값의 detection p-값을 가지는 프로브들은 품질이 낮은 프로브로 판단하여 분석에서 제외시킬 수 있다. 이때, 전자장치(100)는 기준을 설정하는 사용자 입력을 수신할 수 있다.

[0100] 3. 비드 개수

[0101] 특정 CpG 사이트의 정확한 메틸레이션 수준을 측정하기 위해서 측정에 사용된 절대적인 비드(bead)의 개수가 충분할 필요가 있다. 따라서, 프로세서(150)는 모든 샘플 중 특정 프로브를 탐지하기 위한 비드의 개수가 기 정의된 개수(예: 3개) 미만인 경우가 기 정의된 비율 (예: default = 10%)보다 클 경우, 해당 프로브는 품질이 낮은 프로브로 판단하여 분석에서 제외시킬 수 있다. 이때, 전자장치(100)는 기 정의된 값(개수, 비율)을 설정하는 사용자 입력을 수신할 수 있다.

[0102] 4. 비 CG 사이트(Non-CG site)

[0103] 마이크로어레이 데이터에는 CpG 사이트의 메틸레이션 수준을 측정하기 위한 프로브만 존재하는 것이 아닌, CpH 사이트나 어레이의 품질 제어를 위한 네거티브 제어 프로브 등 여러가지 프로브가 존재한다. 메틸레이션 분석에 있어서 가장 관심 있는 것은 CpG 사이트의 메틸레이션 수준이기 때문에 프로세서(150)는 CpG 사이트와 관계없는 프로브는 제거할 수 있다.

[0104] 5. SNP와 연관된 프로브(SNP related probe)

[0105] 다양한 인종 간에는 다양한 단일염기 다형성(Single Nucleotide polymorphism, SNP)이 존재하고 특히 인간에서는 cytosine (C) to thymine (T) polymorphism이 가장 빈번하게 발생한다. 이는 해당 유전자 좌위의 메틸레이션 수준에 큰 영향을 미칠 수 있기 때문에 모든 인종(general)에 연관된 SNP와 연관된 CpG 사이트를 제거하거나 지정된 인종(예: default = general)에 연관된 SNP와 연관된 CpG 사이트를 제거할 수 있다.

[0106] 6. 멀티-히트 프로브(Multi-Hit probe)

[0107] 선행 발표된 논문 등(Nordlund et al., 2013)에 수록된 CpG 사이트 목록에서 해당 CpG들의 포지션(position)은 정렬(alignment)이 불분명한 것으로 나타났기 때문에 분석에서 제외할 수 있다. 이때, 프로세서(150)는 분석 여부에 관한 사용자 입력을 수신하여 제외할 수 있다.

[0108] 7. 성염색체(Sex chromosome) CpGs

[0109] 성별에 따라서 성 염색체 상의 메틸레이션 수준은 X염색체 불활성화(X-chromosome inactivation) 등의 영향으로 인해 통계적 분석에 영향을 미칠 정도로 큰 차이를 보인다. 따라서 성 염색체상의 CpG 사이트의 메틸레이션 수준을 분석하기 위해서는 샘플 모집 단계에서부터 그 영향을 고려해야 한다. 프로세서(150)는 성 염색체 연관 프로브를 분석에서 제거할 수 있으며, 분석 여부에 관한 사용자 입력(default = TRUE)을 수신하여 제외할 수 있다.

[0110] 8. 플래그 프로브(flagged probes)

[0111] 외부 플랫폼에서 수신할 수 있는 어레이 유형 별 매니페스트 파일에는 특정 CpG 사이트를 반복적으로 탐지했을 때, 이상적인 재현률 보다 낮은 재현률을 보이는 일부 프로브들이 명시(FLAG 프로브)되어 있다.

[0112] 프로세서(150)는 플래그 프로브를 분석에서 제거할 수 있으며, 분석 여부에 관한 사용자 입력(default = TRUE)을 수신하여 제외할 수 있다.

[0113] 9. 매핑 실패 프로브(mapping failed probes)

[0114] 외부 플랫폼에서 수신할 수 있는 어레이 유형 별 매니페스트 파일에는 hg19 버전 포지션(position)은 존재하지만 hg38 버전 기준 포지션은 제공되지 않는 CpG들이 존재한다. 이는 hg38 버전으로 참조 유전체(reference genome)가 업데이트 되면서 변경된 위치에 있던 CpG이다. 즉, hg38 참조 유전체 기준 없는 포지션에 위치한 CpG site이거나 멀티 매핑(multi mapping) 등의 문제 때문에 아직 확실한 포지션을 제공할 수 없는 상태에 놓인 CpG들이다.

[0115] 본 분석 방법은 hg38 버전 포지션을 기본적으로 제공하여, 매핑 실패된 프로브들을 분석에서 제거할 수 있으며,

분석 여부에 관한 사용자 입력(default = TRUE)을 수신하여 제외할 수 있다.

[0116] 분석 방법은 통합된 마이크로어레이 데이터를 적어도 하나의 통계 기법을 이용하여 보정(정규화)하는 단계(S403)를 포함할 수 있다.

[0117] 이때, 보정하는 단계는, 샘플 그룹마다 이상치를 가진 샘플을 탐지하는 단계(S404)를 더 포함할 수 있다.

[0118] 앞서 서술한 필터링 과정 중 품질이 낮은 샘플을 제거하는 단계는, 잘못 라벨링되어 전혀 다른 그룹에 속해 버린 샘플이거나, 올바른 그룹에 속해 있지만 너무나도 다른 특성을 가진 샘플인 경우는 제대로 찾아내어 수정하거나 제거할 수 없다는 단점이 있다.

[0119] 또한, 기존의 일반적인 마할라노비스 거리도 이상치 탐지에 사용될 수 있지만, 마할라노비스 거리는 샘플의 분포까지 고려하여 거리를 계산하기 때문에 극단적인 이상치에 취약하다는 단점이 있다.

[0120] 이를 보완하기 위해서 본 분석 방법에서는 전체 샘플에서 기 정의된 비율(default = 90%)만큼 하위 샘플 그룹을 사용자가 지정한 횟수(default = 1000)만큼 무작위 복원 반복 추출하여 하위 샘플 그룹들의 마할라노비스 거리를 각각 측정하고 이상치를 탐지하는 방법을 포함한다.

[0121] 구체적으로, 분석 방법은 통합된 마이크로어레이 데이터 내 복수의 샘플 그룹 중 일정 비율의 하위 샘플 그룹들을 기 정의된 횟수만큼 무작위 복원 반복 추출하는 단계; 및 하위 샘플 그룹들 내 각 샘플에 대해 측정된 마할라노비스 거리(Mahalanobis distance)를 기반으로 이상치를 탐지하는 단계를 포함할 수 있다.

[0122] 결과적으로 각 샘플들이 n번의 반복적 계산에서 몇 번이나 이상치로 탐지된 이력이 있는지 판단하여 20% 이상으로 이상치로 탐지된 적 있다면 명백한 이상치(obvious outlier), 5%이상 20%미만으로 탐지된 적 있다면 잠재적인 이상치(suggestive outlier)로 반환할 수 있다. 이때, 추가적으로 사용자가 원한다면 2개의 주성분으로 플롯(plot)을 제작할 수 있다.

[0123] 무작위 표본을 반복적으로 추출하는 방법을 통해 이상치 존재 여부에 영향을 크게 받는 마할라노비스 거리 값의 단점을 보완하고 계산량을 줄여 빠르게 탐색 가능하다.

[0124] 분석 방법은 통합된 보정된 마이크로어레이 데이터를 주성분 분석하여 배치 효과를 확인하고, 확인된 배치 효과를 제거하는 단계(S405)를 포함할 수 있다.

[0125] 분석 방법은 샘플의 세포 비율을 보정하는 단계(S406)를 포함할 수 있다. 사용된 샘플이 혈액 샘플일 경우, 환자마다 다른 백혈구의 구성 비율이 차등 메틸화 분석에 영향을 미칠 수 있다. 따라서, 각 샘플의 6가지 타입의 백혈구 구성 비율 (CD8-Tcell, CD4-Tcell, NK cell, Bcell, Monocyte, Neutrophils)을 참조 샘플 기반으로 예측하고 보정할 수 있다. FlowSorted.Blood R package에 수록된 estimateCellCounts2 함수를 이용하여 예측할 수 있다.

[0126] 분석 방법은 메틸화 수준 데이터에 주성분 분석 (principal component analysis)을 적용하여 주성분을 추출하는 단계(S407)를 포함할 수 있다.

[0127] 분석 방법은 대상 질병의 유무로부터 오는 메틸화 수준 차이를 탐지하기 위한 선형모델을 구축하는 단계(S408)를 포함할 수 있다.

[0128] 분석 방법은 신호 강도 값을 기반으로 복제 수 변이(Copy-number variation, CNV)를 탐지하고, 구축된 선형모델로부터 차등 메틸화 프로브(differentially methylated probe, DMP) 및 차등 메틸화 영역(differentially methylated region, DMR)을 식별하는 단계(S409, S410, S411)를 포함할 수 있다.

[0129] 분석 방법은 식별된 차등 메틸화 프로브를 바탕으로 통계적 신뢰도를 계산하는 단계(S412)를 포함할 수 있다.

[0130] 분석 방법은 식별된 차등 메틸화 프로브를 바탕으로 유전자 세트 농축 분석을 수행하는 단계(S413)를 포함할 수 있다.

[0131] 도 4는 본 발명의 제1실시예에 따른 분석 결과를 도시한 도면이다.

[0132] 도 4는 확인된 배치 효과를 히트맵으로 나타낸 도면이다. X축은 마이크로어레이 데이터의 주성분을 의미하고 Y축은 잠재적인 배치 효과를 의미한다. 유의미한 상관관계일수록 유색 타일로 표현되었고, 유색 타일의 경우 해당 배치 효과는 마이크로어레이 데이터에 영향을 미치고 있음을 확인할 수 있다.

[0133] 도 5는 본 발명의 제2실시예에 따른 분석 결과를 도시한 도면이다.

[0134] 도 5는 이상치 탐지에 따른 결과값을 나타낸 것이다. 서로 다른 색의 원은 다른 샘플 그룹을 의미한다. 예를 들어, 붉은색으로 표시된 샘플들은 확실한 이상치 샘플(obvious outlier)이고, 파란색으로 표시된 샘플은 잠재적인 이상치 샘플 (suggestive outlier)을 나타낸다.

[0135] 도 6은 본 발명의 제3실시예에 따른 분석 결과를 도시한 도면이다.

[0136] 도 6은 분석 결과를 활용하여 생성한 맨하탄 플롯(Manhattan plot)을 도시한다. 붉은색 점선은 통계적 유의미성을 판단하는 기준선을 -log10 스케일로 표현하였고, 다른 염색체에 존재하는 CpG들은 다른 색의 점으로 표현된다.

[0137] 도 7은 본 발명의 제4실시예에 따른 분석 결과를 도시한 도면이다.

[0138] 도 7은 차등 메틸화 프로브(differentially methylated probe, DMP)를 시각화(visualization)한 볼케이노 플롯(Volcano plot)을 도시한다. 붉은색 점선은 통계적 유의미성을 판단하는 기준선을 -log10 스케일로 표현하였고, 붉은 점은 실험군에서 메틸화 수준이 낮게 나온 CpG를 의미하고 푸른 점은 실험군에서 메틸화 수준이 높게 나온 CpG를 의미한다.

부호의 설명

[0139] 100: 전자장치

110: 입력부

120: 통신부

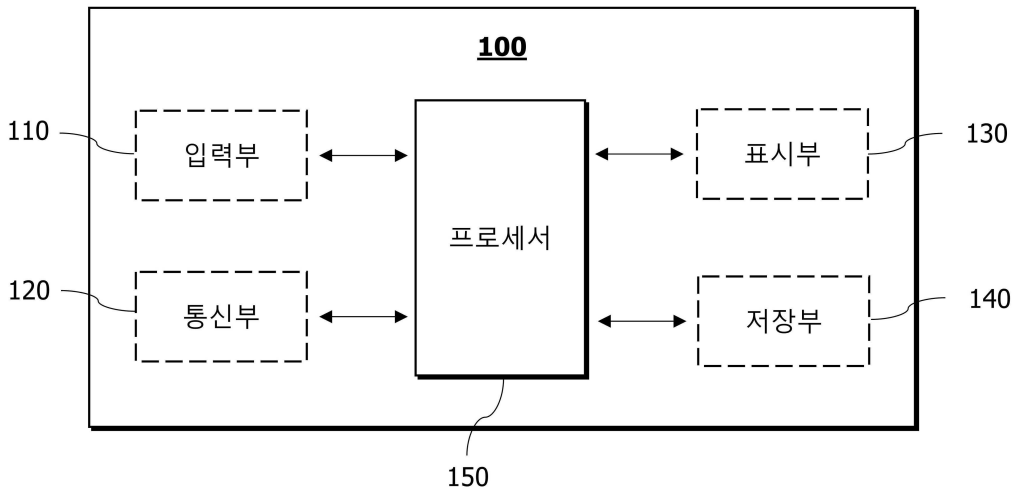
130: 표시부

140: 저장부

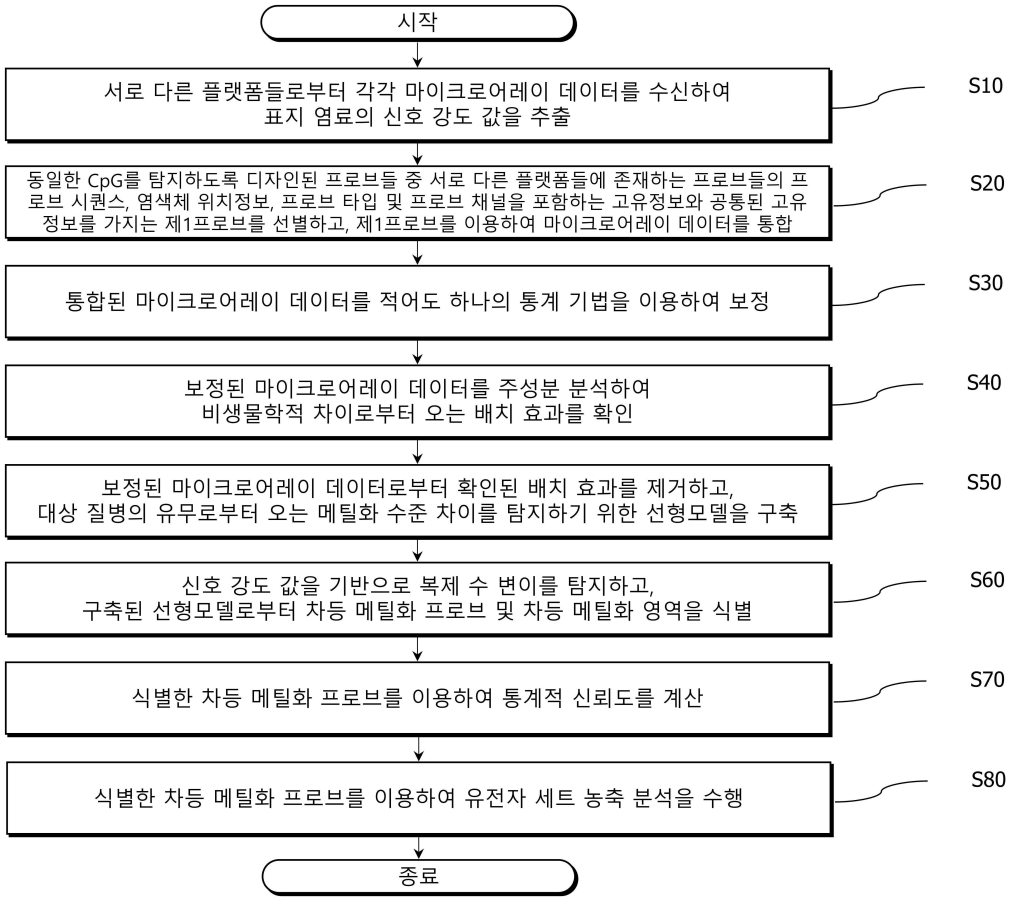
150: 프로세서

도면

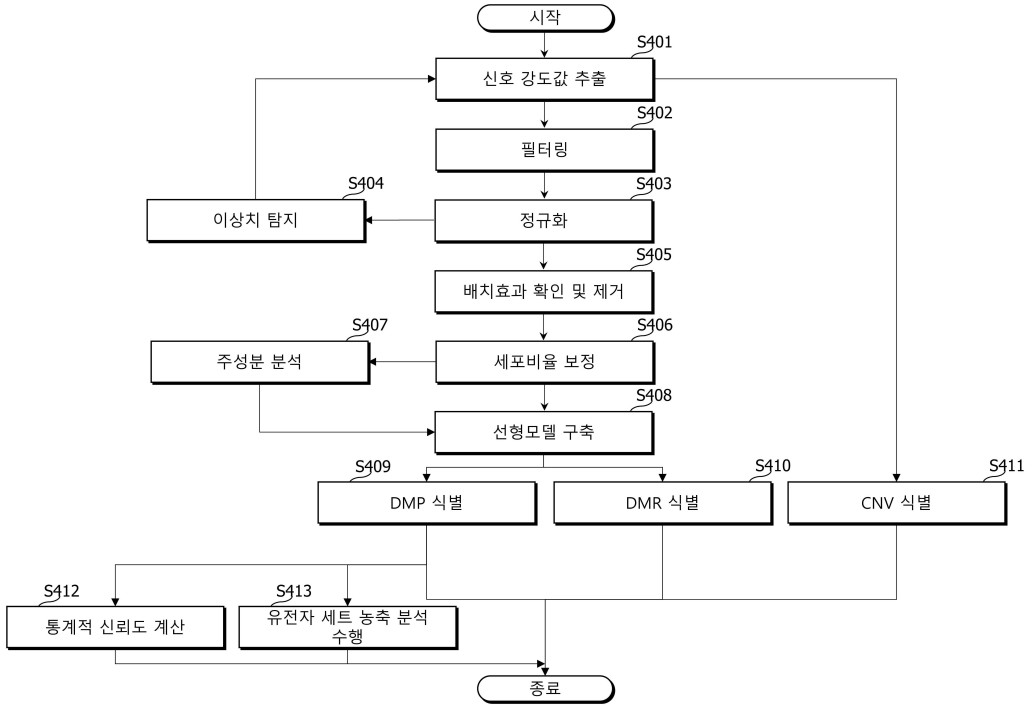
도면1



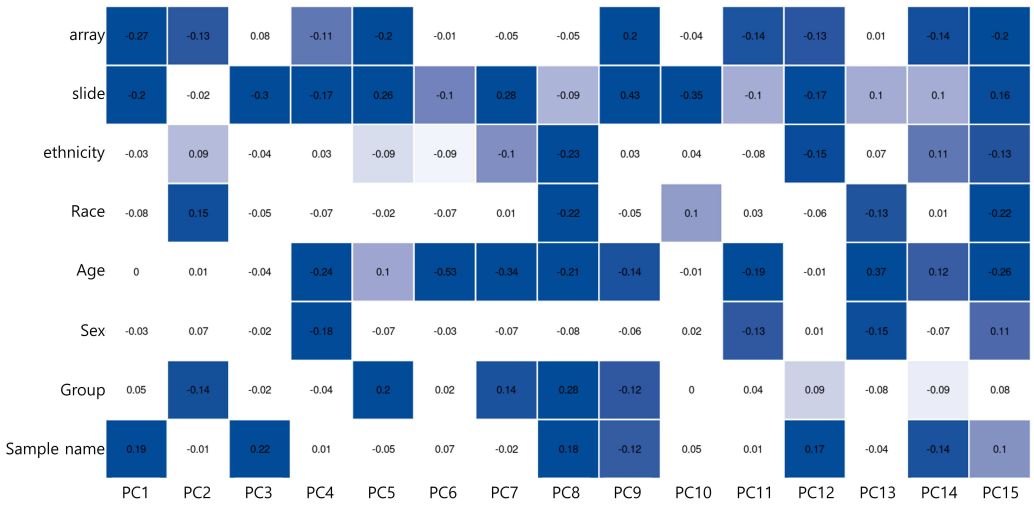
도면2


시작
서로 다른 플랫폼들로부터 각각 마이크로어레이 데이터를 수신하여 표지 염료의 신호 강도 값을 추출
S10
동일한 CpG를 탐지하도록 디자인된 프로브들 중 서로 다른 플랫폼들에 존재하는 프로브들의 프로브 시퀀스, 염색체 위치정보, 프로브 타입 및 프로브 채널을 포함하는 고유정보와 공통된 고유정보를 가지는 제1프로브를 선별하고, 제1프로브를 이용하여 마이크로어레이 데이터를 통합
S20
통합된 마이크로어레이 데이터를 적어도 하나의 통계 기법을 이용하여 보정
S30
보정된 마이크로어레이 데이터를 주성분 분석하여 비생물학적 차이로부터 오는 배치 효과를 확인
S40
보정된 마이크로어레이 데이터로부터 확인된 배치 효과를 제거하고, 대상 질병의 유무로부터 오는 메틸화 수준 차이를 탐지하기 위한 선형모델을 구축
S50
신호 강도 값을 기반으로 복제 수 변이를 탐지하고, 구축된 선형모델로부터 차등 메틸화 프로브 및 차등 메틸화 영역을 식별
S60
식별한 차등 메틸화 프로브를 이용하여 통계적 신뢰도를 계산
S70
식별한 차등 메틸화 프로브를 이용하여 유전자 세트 농축 분석을 수행
S80
종료

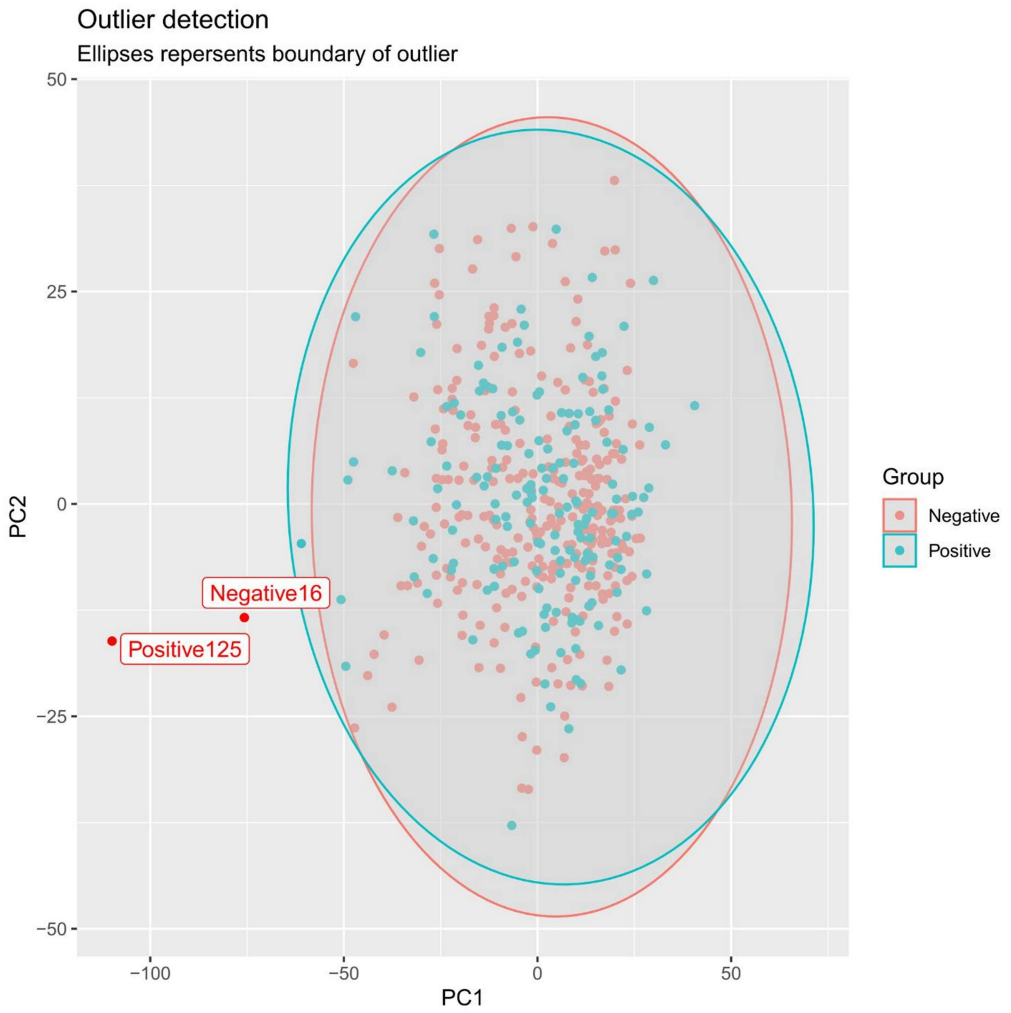
도면3


시작
S401
신호 강도값 추출
S402
필터링
S403
정규화
S404
이상치 탐지
S405
배치효과 확인 및 제거
S406
세포비율 보정
S407
주성분 분석
S408
선형모델 구축
S409
DMP 식별
S410
DMR 식별
S411
CNV 식별
S412
통계적 신뢰도 계산
S413
유전자 세트 농축 분석 수행
종료

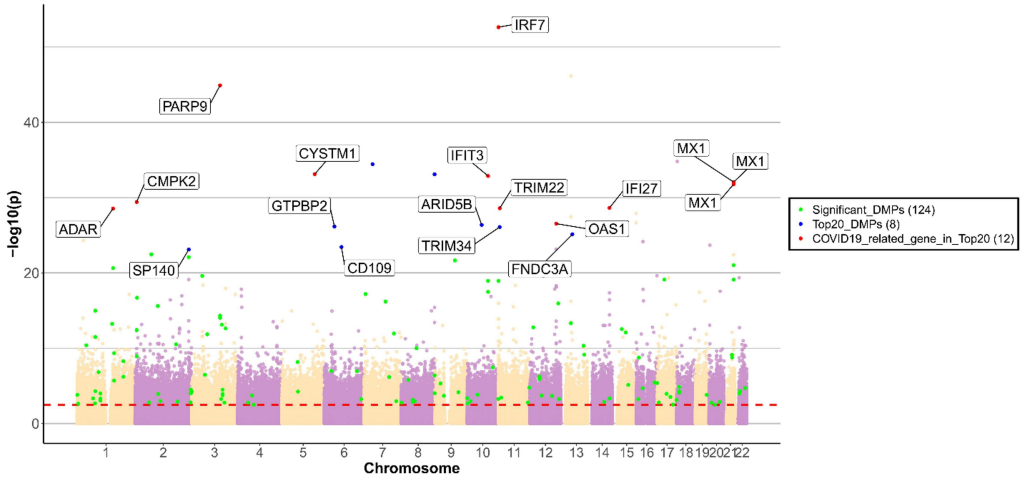
도면4


array -0.27 -0.13 0.08 -0.11 -0.2 -0.01 -0.05 -0.05 0.2 -0.04 -0.14 -0.13 0.01 -0.14 -0.2
slide -0.2 -0.02 -0.3 -0.17 0.26 -0.1 0.28 -0.09 0.43 -0.35 -0.1 -0.17 0.1 0.1 0.16
ethnicity -0.03 0.09 -0.04 0.03 -0.09 -0.09 -0.1 -0.23 0.03 0.04 -0.08 -0.15 0.07 0.11 -0.13
Race -0.08 0.15 -0.05 -0.07 -0.02 -0.07 0.01 -0.22 -0.05 0.1 0.03 -0.06 -0.13 0.01 -0.22
Age 0 0.01 -0.04 -0.24 0.1 -0.53 -0.34 -0.21 -0.14 -0.01 -0.19 -0.01 0.37 0.12 -0.26
Sex -0.03 0.07 -0.02 -0.18 -0.07 -0.03 -0.07 -0.08 -0.06 0.02 -0.13 0.01 -0.15 -0.07 0.11
Group 0.05 -0.14 -0.02 -0.04 0.2 0.02 0.14 0.28 -0.12 0 0.04 0.09 -0.08 -0.09 0.08
Sample name 0.19 -0.01 0.22 0.01 -0.05 0.07 -0.02 0.18 -0.12 0.05 0.01 0.17 -0.04 -0.14 0.1
PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10 PC11 PC12 PC13 PC14 PC15

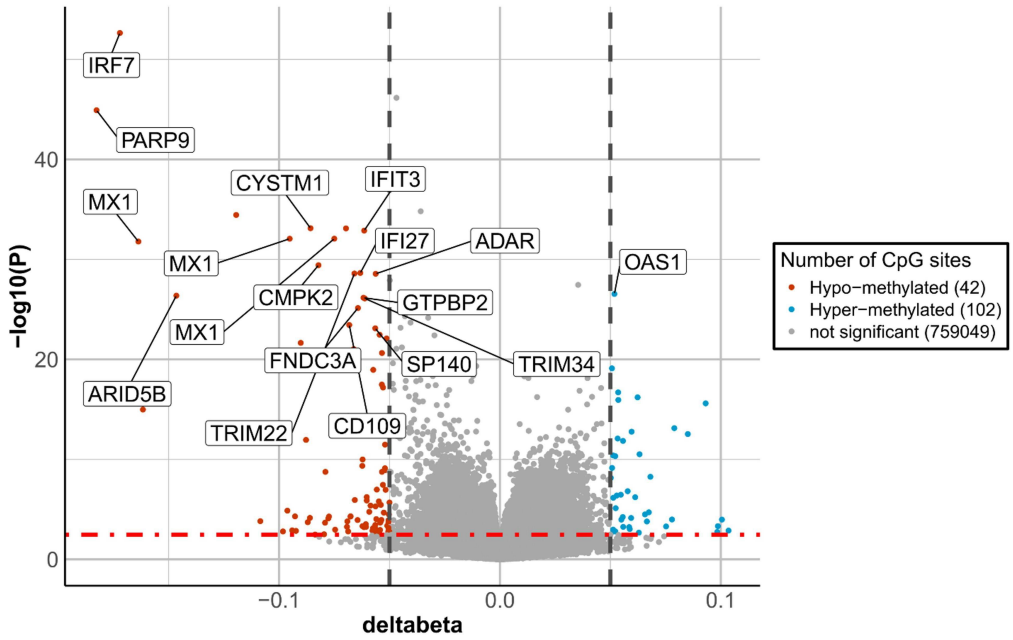
도면5


Outlier detection
Ellipses repersents boundary of outlier
PC2
PC1
Negative16
Positive125
Group
Negative
Positive

도면6


IRF7
PARP9
CYSTM1
IFIT3
MX1
MX1
CMPK2
TRIM22
IFI27
GTPBP2
ARID5B
MX1
ADAR
OAS1
TRIM34
SP140
CD109
FNDC3A
-log10(p)
Chromosome
Significant_DMPs (124)
Top20_DMPs (8)
COVID19_related_gene_in_Top20 (12)

도면7


IRF7
PARP9
MX1
CYSTM1
IFIT3
IFI27
ADAR
OAS1
MX1
CMPK2
GTPBP2
MX1
FNDC3A
SP140
TRIM34
ARID5B
TRIM22
CD109
40
20
0
−log10(P)
−0.1
0.0
0.1
deltabeta
Number of CpG sites
Hypo−methylated (42)
Hyper−methylated (102)
not significant (759049)