(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0042295
(43) 공개일자 2020년04월23일

(51) 국제특허분류(Int. Cl.)
G16B 25/00 (2019.01) *G16B 30/00* (2019.01)
G16B 50/00 (2019.01)
(52) CPC특허분류
G16B 25/00 (2019.02)
G16B 30/00 (2019.02)
(21) 출원번호 10-2018-0122701
(22) 출원일자 2018년10월15일
심사청구일자 2018년10월15일

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
이인석
서울특별시 서대문구 연세로 50, 연세대학교 (신촌동)
한헌종
서울특별시 서대문구 연세로 50, 연세대학교 (신촌동)
(74) 대리인
특허법인 하나

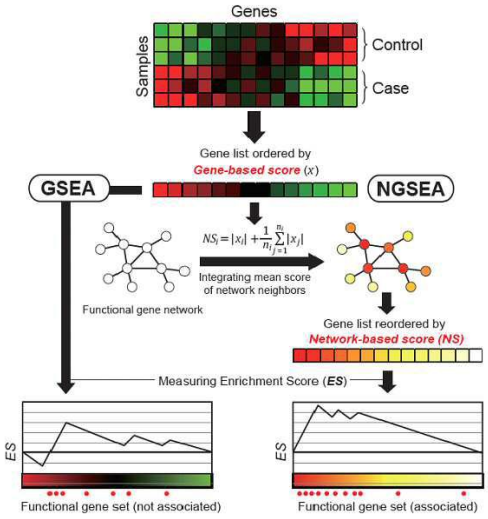
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **네트워크 기반의 유전자 세트 증강 분석 방법을 이용한 약물 재창출 방법**

(57) 요 약

본 발명은 네트워크 기반의 유전자 세트 증강 분석 방법에 관한 것으로, 단독 유전자뿐만 아니라 이웃하는 유전자의 기능적 네트워크를 분석하여 유전자 발현 표현형과 일치하는 경로 유전자 세트를 효과적으로 확인할 수 있는 방법을 제공한다. 또한, 본 발명의 방법을 활용한 알려진 약물의 재창출 방법이 제공된다.

대 표 도 - 도1



(52) CPC특허분류
G16B 50/00 (2019.02)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711076001
부처명 과학기술정보통신부
연구관리전문기관 한국연구재단
연구사업명 포스트게놈신산업육성을위한다부처유전체사업(과기정통부)
연구과제명 유전체 빅데이터 활용을 위한 네트워크증강분석 웹서비스 개발(1/2, 1단계)
기 여 율 1/2
주관기관 연세대학교
연구기간 2018.07.01 ~ 2018.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 2018025079
부처명 과학기술정보통신부
연구관리전문기관 한국연구재단
연구사업명 이공분야기초연구사업
연구과제명 만성난치질환 시스템의학 연구센터
기 여 율 1/2
주관기관 연세대학교
연구기간 2018.06.01 ~ 2019.02.28

명 세 서

청구범위

청구항 1

(a) 유전자 발현 데이터를 포함하는 유전자 세트 정보를 상용화된 데이터베이스로부터 수집하는 단계;

(b) 상기 수집된 유전자 세트와 상호작용하는 유전자 세트를 선별하는 단계; 및

(c) 네트워크 기반 점수(Natwork-based score; NS) 측정법에 기반하여 상기 상호작용하는 유전자 세트 사이의 기능적 연관성을 통합하는 단계;를 포함하는 네트워크 기반의 유전자 세트 증강 분석(Network-based gene set enrichment analysis; NGSEA)을 수행하는 방법.

청구항 2

제1항에 있어서,

상기 네트워크 기반 점수는 하기 수식 1에 의해 산출되는, 방법.

[수식 1]

$$NS_i = |x_i| + \frac{1}{n_i}\sum_{j=1}^{n_i}|x_j|$$

상기 n_i는 i 번째 유전자의 네트워크 이웃의 수이고, 상기 x_i 및 x_j는 각각 I 및 j 번째 유전자의 발현점수이다.

청구항 3

제1항에 있어서,

상기 유전자 세트에 대한 정보는 KEGG PATHWAY Database(https://www.genome.jp/kegg/pathway.html), Drug SIGnatures DataBase(http://tanlab.ucdenver.edu/DSigDB/DSigDBv1.0/), Gene Ontology Consortium(http://www.geneontology.org), DisGeNET(http://www.DisGeNET.org) 및 Diseases(https://diseases.jensenlab.org)를 포함하는 데이터베이스 군으로부터 선택되는 어느 하나 이상의 데이터베이스로부터 획득하는, 방법.

청구항 4

제1항에 있어서,

상기 유전자 세트 사이의 기능적 연관성은 게놈 규모의 기능 유전자 네트워크로 구현하는, 방법.

청구항 5

제4항에 있어서,

상기 게놈 규모의 기능 유전자 네트워크는 HumanNet(www.inetbio.org/humannet) 또는 MouseNet(www.inetbio.org/mousenet)의 데이터 베이스에서 획득하는, 방법.

청구항 6

제1항에 있어서,

상기 네트워크 기반의 유전자 세트 증강 분석은 합산 점수 접근법(aggregate score approach)을 통해 수행하는, 방법.

청구항 7

(a) 약물에 대한 질병 유전자 발현 데이터 세트의 정보를 상용화된 데이터베이스로부터 수집하는 단계; 및

(b) 상기 질병 유전자 발현 데이터 세트를 네트워크 기반 점수에 따라 우선순위를 나열하여 질병과의 연관성을 평가하는 단계;를 포함하는 약물 재창출 방법.

청구항 8

제7항에 있어서,

상기 약물에 대한 정보는 Comparative Toxicogenomics Database(http://ctdbase.org/) 또는 PubChem 데이터베이스(https://ftp.ncbi.nlm.nih.gov/pubchem/Compound/Extras/)를 포함하는 데이터베이스에서 획득하는, 약물 재창출 방법.

청구항 9

제7항에 있어서,

상기 약물에 대한 질병 유전자 발현 데이터 세트는 Bioconductor(https://bioconductor.org/packages/release/data/experiment/html/KEGGdzPathwaysGEO.html), 를 포함하는 데이터베이스 군으로부터 선택되는 어느 하나 이상의 데이터베이스로부터 획득하는, 약물 재창출 방법.

청구항 10

제7항에 있어서,

(c) 상기 약물을 질병 세포에 처리하여 치료 효과를 평가하는 단계;를 더 포함하는, 약물 재창출 예측 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 네트워크 기반의 유전자 세트 증강 분석 방법 및 이를 이용한 약물 재창출 방법에 관한 것이다.

배 경 기 술

[0002] 임상 표본의 분자 표현형은 질병 진단, 환자 계층화, 약물 발견 등에 유용하게 활용되고 있다. 유전자 발현 프로파일링은 임상 표본의 분자 표현형 분석을 위한 가장 접근하기 쉬운 전략이다.

[0003] DNA 칩 기술 및 RNA sequencing은 환자 유래 일차 세포 및 세포주의 분자 프로파일링에 사용된다.

[0004] 임상 표본의 수많은 유전자 발현 프로파일은 Gene Expression Omnibus(GEO) 및 NCI Genomic Data Commons(GDC)와 같은 공공 데이터베이스에서 자유롭게 이용할 수 있다. 게놈 전반의 표현형에 대한 기능 분석은 일반적으로 개개의 유전자가 아닌 주석이 달린 유전자 세트로 해석이 가능하다.

[0005] 따라서 최근 유전자 세트 분석(Gene Set Analysis)을 위한 다양한 알고리듬들이 개발되었다. 이들 중 많은 방법들이 임상표본에서 특이적으로 발현하는 유전자들과 유전자 세트 사이 중복의 통계적인 유의미성을 측정하는 방법으로 이들을 over-representation 접근법으로 분류한다. 상기 분석 방법들은 상당히 합리적이지만 임상시료 특이적 발현 정도에 의한 중요성이 낮은 유전자를 무의미한 유전자로 취급하는 문제점이 있다. 이를 보완하기 위해 개발된 방법이 유전자 세트 증강 분석(Gene Set Enrichment Analaysis; GSEA)이다.

[0006] 그러나 상기 방법도 실제 질환에 원인이 되는 유전자들 보다는 원인유전자의 조절을 받아 발현에 영향을 크게 보이는 유전자들을 중심으로 분석이 되는 단점을 가지고 있다.

[0007] 본 발명자들은 질병과 관련된 유전자들은 실제로 기능적으로 연관된 다른 유전자의 발현을 더 크게 변화시킬 가능성이 높으므로 유전자의 발현 데이터를 기반으로 한 유전자 세트 분석은 유전자들의 기능적인 네트워크 상에서 각 유전자의 이웃하는 모든 유전자들의 발현 정보를 고려하여 분석해야 한다고 가정하였다.

[0008] 이에, 본 발명자들은 각 유전자의 이웃하는 유전자들의 발현 정보를 통합하여 유전자 세트 증강 분석을 진행하기 위한 네트워크 기반 점수(Natwork-based score; NS)를 개발하였고, 상기 점수를 이용하여 네트워크 기반의 유전자 세트 증강 분석(Network-based Gene Set Enrichment Analysis)을 수행하였다.

[0009] 또한, 본 발명자들은 상기 네트워크 기반의 유전자 세트 증강 분석을 통해 알려진 약물 중에서 새롭게 질병을 치료할 수 있는 약물을 재창출하는 예측 방법을 개발하였다.

발명의 내용

해결하려는 과제

[0010] 본 발명은 전술한 종래기술의 문제점을 해결하기 위한 것으로, 본 발명의 목적은 유전자 네트워크를 통해 유전자 세트 증강 분석을 개선하고, 질병에 대한 약물을 재창출하는 시스템을 제공하는 것이다.

과제의 해결 수단

[0011] 본 발명의 일 측면에 따르면, (a) 유전자 발현 데이터를 포함하는 유전자 세트 정보를 상용화된 데이터베이스로부터 수집하는 단계; (b) 상기 수집된 유전자 세트와 상호작용하는 유전자 세트를 선별하는 단계; 및 (c) 네트워크 기반 점수(Natwork-based score; NS) 측정법에 기반하여 상기 상호작용하는 유전자 세트 사이의 기능적 연관성을 통합하는 단계;를 포함하는 네트워크 기반의 유전자 세트 증강 분석(Network-based gene set enrichment analysis; NGSEA)을 수행하는 방법이 제공된다.

[0012] 일 실시예에 있어서, 상기 네트워크 기반 점수는 하기 수식 1에 의해 산출될 수 있다.

[0013] [수식 1]

[0014] $$NS_i = |x_i| + \frac{1}{n_i}\sum_{j=1}^{n_i}|x_j|$$

[0015] 상기 n_i는 i 번째 유전자의 네트워크 이웃의 수이고, 상기 x_i 및 x_j는 각각 i 및 j 번째 유전자의 발현 점수이다.

[0016] 일 실시예에 있어서, 상기 유전자 세트에 대한 정보는 KEGG PATHWAY Database(https://www.genome.jp/kegg/pathway.html), Drug SIGnatures DataBase(http://tanlab.ucdenver.edu/DSigDB/DSigDBv1.0/), Gene Ontology Consortium(http://www.geneontology.org), DisGeNET(http://www.DisGeNET.org) 및 Diseases(https://diseases.jensenlab.org)를 포함하는 데이터베이스 군으로부터 선택되는 어느 하나 이상의 데이터베이스로부터 획득할 수 있다.

[0017] 일 실시예에 있어서, 상기 유전자 세트 사이의 기능적 연관성은 게놈 규모의 기능 유전자 네트워크로 구현될 수 있다.

[0018] 일 실시예에 있어서, 상기 네트워크 기반의 유전자 세트 증강 분석은 합산 점수 접근법(aggregate score approach)을 통해 수행될 수 있다.

[0019] 본 발명의 다른 측면에 따르면, (a) 약물에 대한 질병 유전자 발현 데이터 세트의 정보를 상용화된 데이터베이스로부터 수집하는 단계; 및 (b) 상기 질병 유전자 발현 데이터 세트를 네트워크 기반 점수에 따라 우선순위를 나열하여 질병과의 연관성을 평가하는 단계;를 포함하는 약물 재창출 예측시스템이 제공된다.

[0020] 일 실시예에 있어서, 상기 약물에 대한 정보는 Comparative Toxicogenomics Database(http://ctdbase.org/) 또는 PubChem 데이터베이스(https://ftp.ncbi.nlm.nih.gov/pubchem/Compound/Extras/)를 포함하는 데이터베이스에서 획득할 수 있다.

[0021] 일 실시예에 있어서, 상기 약물에 대한 질병 유전자 발현 데이터 세트는 Bioconductor(https://bioconductor.org/packages/release/data/experiment/html/KEGGdzPathwaysGEO.html), 를 포함하는 데이터베이스 군으로부터 선택되는 어느 하나 이상의 데이터베이스로부터 획득할 수 있다.

[0022] 일 실시예에 있어서, 상기 약물 재창출 방법은 (c) 상기 약물을 질병 세포에 처리하여 치료 효과를 평가하는 단계;를 더 포함할 수 있다.

발명의 효과

[0023] 본 발명의 일 측면에 따른 유전자 세트 증강 분석 방법은 이웃하는 유전자 세트간의 연관성을 분석하고 이를 정량화할 수 있으므로, 질병 연관 유전자 발굴 및 약물 재창출 방법을 위한 유의적인 데이터를 효과적으로 제공할 수 있다.

[0024] 본 발명의 효과는 상기한 효과로 한정되는 것은 아니며, 본 발명의 상세한 설명 또는 특허청구범위에 기재된 발명의 구성으로부터 추론 가능한 모든 효과를 포함하는 것으로 이해되어야 한다.

도면의 간단한 설명

[0025] 도 1은 본 발명에 따른 네트워크 기반의 유전자 세트 증강 분석(NGSEA) 방법의 모식도를 나타낸 것이다.

도 2는 GSEA, AE 및 NGSEA에 의해 일치된 질병 발현 데이터 세트에 대한 관련 KEGG pathway의 예측력을 나타낸 것으로, (A) 는 GSEA, AE 및 NGSEA에서 KEGG pathway 용어와 일치하는 순위 분포를 나타낸 것이고, (B) 는 GSEA 및 NGSEA에서 KEGG pathway 용어와 일치하는 순위 구성을 나타낸 것이며, (C) 는 동일 또는 다른 질병에서 Pearson's correlation coefficient(PCC)의 normalized enrichment scores(NES) 분포를 나타낸 것이며, (D) 는 Alzheimer's disease(HSA05010) 및 Staphylococcus aureus infection(HSA05150)에서 KEGG pathway 용어의 서브네트워크를 나타낸 것이고, (E) 는 acute myeloid leukemia(HSA05221) 및 taste transduction(HSA04742) 에서 KEGG pathway 용어의 서브네트워크를 나타낸 것이다.

도 3은 CMap 및 NGSEA에 의해 일치된 질병 발현 데이터에 대한 알려진 약물 검색 결과를 나타낸 것으로, (A)는 NGSEA에 따른 약물 검색 방법의 모식도이고, (B)는 CMap 및 NGSEA의 검색 능력을 AUROC로 비교한 결과를 나타낸 것이다.

도 4(A)는 CMap 및 NGSEA에서 대장암(GSE9348) 치료를 위해 알려진 약물 검색 결과를 ROC 곡선으로 나타낸 것이고, 도 4(B)는 NGSEA로 대장암 치료를 위해 예측된 상위 30 가지 화학 물질을 나타낸 것이고, 도 4(C)는 다양한 농도(0 내지 250 μM)의 budesonide를 HCT-116 세포주에 처리 후 세포 생존력을 나타낸 것이며, 도 4(D)는 다양한 농도(0 내지 250 μM)의 budesonide를 HT-29 세포주에 처리 후 세포 생존력을 나타낸 것이다.

발명을 실시하기 위한 구체적인 내용

[0026] 이하에서는 첨부한 도면을 참조하여 본 발명을 설명하기로 한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 따라서 여기에서 설명하는 실시 예로 한정되는 것은 아니다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0027] 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 구비할 수 있다는 것을 의미한다.

[0028] 달리 정의되지 않는 한, 분자 생물학, 미생물학, 단백질 정제, 단백질 공학, 및 DNA 서열 분석 및 당업자의 능력 범위 안에서 재조합 DNA 분야에서 흔히 사용되는 통상적인 기술에 의해 수행될 수 있다. 상기 기술들은 당업자에게 알려져 있고, 많은 표준화된 교재 및 참고저서에 기술되어 있다.

[0029] 본 명세서에 달리 정의되어 있지 않으면, 사용된 모든 기술 및 과학 용어는 당업계에 통상의 기술자가 통상적으로 이해하는 바와 같은 의미를 가진다.

[0030] 본 명세서에 포함되는 용어를 포함하는 다양한 과학적 사전이 잘 알려져 있고, 당업계에서 이용 가능하다. 본 명세서에 설명된 것과 유사 또는 등가인 임의의 방법 및 물질이 본원의 실행 또는 시험에 사용되는 것으로 발견되나, 몇몇 방법 및 물질이 설명되어 있다. 당업자가 사용하는 맥락에 따라, 다양하게 사용될 수 있기 때문에, 특정 방법론, 프로토콜 및 시약으로 본 발명이 제한되는 것은 아니다.

[0031] 본 명세서에서 사용되는 바와 같이, 단수형은 문맥이 명확하게 달리 지시하지 않으면 복수의 대상을 포함한다.

[0032] 이하 본 발명을 더욱 상세히 설명한다.

[0033] 본 발명의 일 측면에 따르면, (a) 유전자 발현 데이터를 포함하는 유전자 세트 정보를 상용화된 데이터베이스로부터 수집하는 단계; (b) 상기 수집된 유전자 세트와 상호작용하는 유전자 세트를 선별하는 단계; 및 (c) 네트워크 기반 점수(Network-based score; NS) 측정법에 기반하여 상기 상호작용하는 유전자 세트 사이의 기능적 연관성을 통합하는 단계;를 포함하는 네트워크 기반의 유전자 세트 증강 분석(Network-based gene set enrichment

analysis; NGSEA)을 수행하는 방법이 제공된다.

[0034] 도 1은 본 발명에 따른 네트워크 기반의 유전자 세트 증강 분석 방법의 전체적인 모식도를 나타낸 것이다.

[0035] 상기 네트워크 기반 점수는 하기 수식 1에 의해 산출될 수 있다.

[0036] [수식 1]

[0037] $$NS_i = |x_i| + \frac{1}{n_i}\sum_{j=1}^{n_i}|x_j|$$

[0038] 상기 n_i는 i 번째 유전자의 네트워크 이웃의 수이고, 상기 x_i 및 x_j는 각각 i 및 j 번째 유전자의 발현점수이다.

[0039] 상기 수식 1에서 유전자의 점수를 절대값으로 책정함으로써 이웃하는 유전자 세트와의 상호작용을 모두 적용할 수 있다.

[0040] 예를 들어, 유전자 세트 A의 네트워크 이웃인 B 및 C가 존재하고, B 및 C는 각각 A와 상호작용을 하여 A의 발현을 상향(+) 및 하향(-) 조절하는 경우, 유전자 점수에 절대값을 씌우지 않고 더하면, B 및 C에 의한 상호작용 값이 상쇄될 수 있으나, 각 값에 절대값을 씌우면 B 및 C에 의한 상호작용 점수가 네트워크 기반 점수에 온전히 적용될 수 있다.

[0041] 상기 유전자 세트에 대한 정보는 상용화된 데이터베이스에서 수집할 수 있으며, 상기 상용화된 데이터 베이스는 유전자 세트의 정보를 저장하는 데이터베이스로서, KEGG PATHWAY Database(https://www.genome.jp/kegg/pathway.html), Drug SIGnatures DataBase(http://tanlab.ucdenver.edu/DSigDB/DSigDBv1.0/), Gene Ontology Consortium(http://www.geneontology.org), DisGeNET(http://www.DisGeNET.org) 및 Diseases(https://diseases.jensenlab.org)를 포함하는 데이터베이스 군으로부터 선택되는 어느 하나 이상의 데이터베이스일 수 있으나, 이에 제한되지 않는다.

[0042] 상기 유전자 세트 사이의 기능적 연관성은 게놈 규모의 기능 유전자 네트워크로 구현될 수 있다.

[0043] 유전자 세트 발현 분석에 있어서 종래에 사용되었던 over-representation 접근법은 중요성이 떨어지는 유전자 세트와의 상호작용이 무시될 수 있고, 차등적으로 발현된 유전자 세트 사이의 상대적인 순서에 대한 정보도 제공할 수 없는 문제점이 있다.

[0044] 이에, 본 발명에서는 유전자 세트의 발현 분석에 있어서 유전자 세트와 이웃하는 모든 유전자 세트에 특이 점수를 기초하여 각 주석이 달린 유전자 집합의 점수를 할당하는 합산 점수 접근법이 사용될 수 있다.

[0045] 본 발명의 다른 측면에 따르면, (a) 약물에 대한 질병 유전자 발현 데이터 세트의 정보를 상용화된 데이터베이스로부터 수집하는 단계; 및 (b) 상기 질병 유전자 발현 데이터 세트를 네트워크 기반 점수에 따라 우선순위를 나열하여 질병과의 연관성을 평가하는 단계;를 포함하는 약물 재창출 예측시스템이 제공된다.

[0046] 상기 약물에 대한 정보는 Comparative Toxicogenomics Database(http://ctdbase.org/) 또는 PubChem 데이터베이스(https://ftp.ncbi.nlm.nih.gov/pubchem/Compound/Extras/)를 포함하는 데이터베이스에서 획득할 수 있다.

[0047] 상기 약물에 대한 질병 유전자 발현 데이터 세트는 Bioconductor(https://bioconductor.org/packages/release/data/experiment/html/KEGGdzPathwaysGEO.html) 를 포함하는 데이터베이스 군으로부터 선택되는 어느 하나 이상의 데이터베이스로부터 획득할 수 있다.

[0048] 본 발명자들은 NGSEA의 유용성을 높이기 위해 웹 기반 유전자 세트 농축 분석 서버(www.inetbio.org/ngsea)를 개발하였다.

[0049] 사용자는 KEGG PATHWAY, Gene Ontology Consortium, DisGeNET 및 Diseases과 같은 다양한 데이터베이스에 의해 생물학적 및 질병 과정을 나타내는 기능적 유전자 세트의 우선 순위를 지정할 수 있다.

[0050] 사용자는 유전자 발현 표현형을 제출하여 GSEA와 NGSEA를 동시에 수행할 수 있다.

[0051] Expression Matrix(.gct 형식) 데이터와 사전 득점 된 유전자 목록(.rnk 형식) 모두를 분석을 위한 입력 데이터로 제출할 수 있다.

[0052] 마우스 유전자에 대한 증강 분석은 게놈 규모의 마우스 기능 유전자 네트워크로 수행할 수 있다.

[0053] 사용자는 ES, NES 및 FDR에 따라 유전자 세트의 우선 순위를 지정할 수 있으며, 농축 플롯도 사용할 수 있다.

[0054] 상기 약물 재창출 방법은 (c) 상기 약물을 질병 세포에 처리하여 치료 효과를 평가하는 단계;를 더 포함할 수 있다.

[0055] 이하 실시예를 통해, 본 발명을 더욱 상술하나 하기 실시예에 의해 본 발명이 제한되지 아니함은 자명하다.

[0056] **실험예 1 : 유전자 발현 프로파일, 주석된 유전자 세트 및 기능적 인간 유전자 네트워크**

[0057] 유전자 발현 표현형에 대한 유전자 세트 분석 성능 평가를 위해 KEGG pathway 용어가 이미 주석으로 표시된 발현 프로파일로 구성된 표준 표현 데이터 세트를 사용하였다.

[0058] Bioconductor (https://bioconductor.org/packages/release/data/experiment/html/KEGGdzPathwaysGEO.html)에서 얻은 GEO (KEGGdzPathwaysGEO)의 KEGG 질병 데이터 세트를 유전자 세트 농축 분석 방법의 평가를 위한 표준 데이터 세트로 사용하였다.

[0059] 예를 들어, KEGGdzPathwaysGEO의 GSE21354 데이터 세트는 KEGG pathway 용어 '신경교종(glioma; hsa05214)'으로 주석을 달고 종양 조직 및 4 개의 정상 조직으로부터 14 개의 샘플을 포함하는 미세배열(microarray) 기반 유전자 발현 데이터를 포함하였다.

[0060] 인간 KEGG pathway(https://www.genome.jp/kegg/pathway.html, 2016 년 6 월)와 Drug Signature Database(DSigDB)의 약물 표적 유전자 세트(http://tanlab.ucdenver.edu/DSigDB, version 1)로부터 경로 유전자 세트를 얻었다.

[0061] GSEA와 동일한 기준의 기본 매개 변수 설정을 위해 15 개 미만의 유전자를 포함하는 유전자 세트는 분석에서 제외하였다.

[0062] DSigDB의 경우 약물 이름은 PubChem 데이터베이스(http://ftp.ncbi.nlm.nih.gov/pubchem/Compound/Extras/)에서 제공한 화합물 ID (compound ID, CID)로 매핑하였다.

[0063] 마지막으로 분석을 위해 276 개의 KEGG pathway 유전자 세트와 165 개의 DSigDB 유전자 세트를 사용했다.

[0064] 웹 서버 구축을 위해 Gene Ontology biological process(GOBP) 주석(http://www.geneontology.org, 2018.04.04.), DisGeNET(http://www.DisGeNET.org, 2018.06.08.)의 curated annotation, 질병 유전자 별 3 개 이상의 별점을 가진 질병(https://diseases.jensenlab.org)의 추가 유전자 세트를 사용하였다.

[0065] 질병에 대한 약물 검색 능력을 벤치마킹하기 위해 비교 독성 유전체학 데이터베이스(CTD)의 '치료'범주에서의 연관에 대한 직접적인 증거를 근거로 하여 2,109 개의 질병과 1,481 개의 화학 물질 사이에 17,063 개의 링크를 작성하였다(http://ctdbase.org/ 2018.10.04.)

[0066] CID를 사용하여 약물의 정보를 동의어와 결합하였다.

[0067] 네트워크 기반의 유전자 발현 분석은 게놈 규모의 기능 유전자 네트워크인 HumanNet-XC(www.inetbio.org/humannet)로 구현하였다.

[0068] 즉, HumanNet-XC는 단백질-단백질의 상호작용뿐만 아니라 Bayesian statistics를 통한 다양한 유형의 omic data로부터 유추된 유전자들 사이의 기능적 연관성을 통합하였다.

[0069] HumanNet-XC는 17,790 개의 인간 유전자(코딩 게놈의 94.6 %) 사이에 424,501 개의 functional link를 포함한다.

[0070] 웹 서버에서 마우스 유전자 발현 표현형을 NGSEA에 이용하기 위해 17,714 개의 마우스 유전자(코딩 게놈의 88 %) 사이에 788,080 개의 링크를 포함하는 MouseNet(www.inetbio.org/mousenet)에 대한 기능 유전자 네트워크를 사용하였다.

[0071] **실험예 2 : GSEA, AE 및 NGSEA 결과 비교**

[0072] Broad Institute(http://software.broadinstitute.org/gsea/downloads.jsp)의 javaGSEA v3.0 소프트웨어를 다운받아 분석 및 웹 서버 구현에 사용하였다.

[0073] 상기 javaGSEA는 GSEA 또는 GSEA-preranked 중 하나의 입력 데이터를 분석할 수 있다

[0074] 유전자 발현 매트릭스는 대조군과 실험군을 모두를 포함하였다. 유전자의 순위를 변경하여 GSEA를 향상시키기 위해 기본 매개 변수인 'weighted GSEA-preranked' 함수를 사용하였다.

[0075] 종래의 GSEA는 유전자 발현율, 신호 대 잡음비(SNR) 또는 발현율의 $\log_2$(Ratio)를 기준으로 가장 높게 발현된 유전자를 선정하였다.

[0076] SNR은 실험군과 대조군 간의 평균 발현값 차이를 각 그룹의 표준 편차 합으로 나눈 것이다.

[0077] $\log_2$(Ratio)는 대조군 시료의 평균 발현값에 대한 실험군 시료의 평균 발현값의 비율을 밑이 2인 로그를 취하여 계산하였다.

[0078] NGSEA는 종래의 유전자 기반 점수를 네트워크 이웃 유전자 기반 점수로 수정하였다.

[0079] 구체적으로, 상기 유전자 기반 점수의 절대값을 상기 네트워크 이웃 유전자 기반 점수의 절대값의 평균으로 통합하였으며, 각 유전자에 대한 네트워크 기반 점수(Network-based score, NS)를 하기 수식 1로 나타내었다.

[0080] [수식 1]

[0081]
$$NS_i = |x_i| + \frac{1}{n_i}\sum_{j=1}^{n_i}|x_j|$$

[0082] 상기 n_i는 i 번째 유전자의 네트워크 이웃의 수이고, 상기 x_j는 j 번째 유전자의 발현점수이다. 유전자 발현 데이터가 없는 경우 유전자 기반 점수를 0으로 하였다.

[0083] SNR과 $\log_2$(Ratio)를 모두 실험해본 결과 $\log_2$(Ratio)가 일반적으로 더 나은 결과값을 제공하였으므로, 모든 결과는 $\log_2$(Ratio)를 유전자 기반 점수를 활용하였다.

[0084] absolute enrichment(AE) 분석을 위해 $\log_2$(Ratio)의 절대값을 기반으로 유전자를 나열하였다.

[0085] GSEA, AE, 및 NGSEA는 GSEA 사전 함수를 사용하여 각각 $\log_2$(Ratio) 값, $\log_2$(Ratio)의 절대값 및 NS를 수행하여 유전자 목록을 나열하였고, 상기 GSEA 사전 함수는 enrichment scores(ES), normalized enrichment scores(NES), P-values 및 FDR(false discovery rate) values for each gene set based on modified Kolmogorov Smirnov(K-S) test로 계산하였다.

[0086] 유전자 세트의 회복 성능을 평가하기 위해, 양성 및 음성 모두 높은 점수를 갖는 유전자 세트를 GSEA에서 동등하게 가중시키는 absolute NES로 유전자 세트의 우선 순위를 나열하였다.

[0087] 도 2A를 참조하면, NGSEA의 순위 분포는 GSEA 및 AE와 비교하여 유의하게 높았다($P=2.35e^{-3}$ and $P=4.0e^{-3}$, respectively, by Wilcoxon signed rank test).

[0088] 도 2B를 참조하면, 일치하는 KEGG 경로 조건의 순위는 질병 발현 데이터 세트를 테스트한 24 개 중 18 개(75 %)에서 GSEA와 비교하여 NGSEA에 의해 향상되었다.

[0089] 예를 들어, KEGG 용어 '신경교종(Glioma)'은 GSEA에서 131 번째로 검색되었지만 신경교종 샘플에서 파생된 유전자 발현 데이터 세트(GSE21354)는 NGSEA에서 18 번째로 검색되었다.

[0090] 한편, AE의 성능은 GSEA로부터 유의하게 개선되지 않았다(P=0.11 by Wilcoxon signed rank test).

[0091] 상기 결과는 NGSEA에서 관찰된 개선의 주요 요인이 유전자 발현 데이터의 네트워크 기반 분석에 의한 것임을 시사한다.

[0092] 동일한 질병에 대하여 서로 다른 발현 프로파일 간 KEGG 경로의 할당 점수를 비교하여 세 가지 농축 분석 방법의 견고성(robustness)을 확인하였다.

[0093] 도 2C를 참조하면, 세 가지 농축 분석 결과 모두 동일한 질병 사이의 경로 점수가 다른 질병 사이의 경로 점수보다 유의미한 상관 관계를 보였다.

[0094] 특히, NGSEA는 GSEA와 비교하여 동일 질병군 및 다른 질병군 사이에서 상관 차이의 유의성을 개선시켰다(각각

$P=2.72e^{-6}$ 및 $P=3.44e^{-5}$, Wilcoxon rank sum test).

[0095] 상기 결과는 NGSEA의 농축 분석이 동일한 질병 과정에 대한 발현 프로파일 간의 다양성에 영향을 덜 미침을 시사한다.

[0096] 예를 들어, 도 2D를 참조하면, 알츠하이머 병에 대한 유전자 발현 데이터(GSE5281_VCX)의 경우, 네트워크 기반 점수 측정 방법이 KEGG 용어 '알츠하이머 병'을 17 번째에서 5 번째로 올려 놓았고, 대다수의 경로 유전자는 NGSEA(붉은 색)에서 높게 평가되었으나, KEGG 용어 'Staphylococcus aureus 감염'의 경우 6 번째에서 267 번째로 내려갔으며, 대다수의 경로 유전자는 GSEA에서 높게 평가되었다.

[0097] 도 2E를 참조하면, 급성 골수성 백혈병의 경우, 관련 및 비관련 경로 사이에서 순위가 유사하게 변화하는 경향을 확인하였다.

[0098] 상기 결과는 네트워크 기반 스코어링이 농축 분석을 위해 정렬된 유전자 목록의 기본 생물학적 과정에서 진정한 관련 유전자 집합에 할당된 점수를 더 증가시켜 진정한 기능 유전자의 순위를 높여주었음을 시사한다.

[0099] **실험예 3 : Connectivity Map(CMap)을 이용한 약물 재창출(Drug repositioning)**

[0100] CMap 웹 서버(https://portals.broadinstitute.org/cmap)에서 검색한 FDA 승인 의약품에 대한 24 개의 KEGG 질병 유전자 발현 데이터 세트를 우선순위를 나열하였다.

[0101] CMap은 입력 데이터로 상향 및 하향 태그(Affymetrix HG-U133a probe ID) 목록이 필요하므로 24 개의 각 질병 발현 데이터 세트에서 50 개의 상향 및 하향 조절된 프로브 ID를 선택하였다.

[0102] 입력 유전자가 Affymetrix HG-U133a 프로브 ID를 기반으로 하지 않으면 CMap 분석을 실행하기 위해 AffyMetrix HG-U133a 프로브 ID로 변환하였다.

[0103] 도 3A를 참조하면, 각 FDA 승인 약물에 대한 표적 유전자를 네트워크 기반 점수에 따라 유전자 목록으로 나열하여 질병과의 연관성을 시험하기 위한 기능적 유전자 세트로 사용하였다.

[0104] DSigDB의 능동적인 생물 검정에 근거하여 약물-표적 링크로부터 약물에 대한 표적 유전자 세트를 수집하였다.

[0105] KEGGdzPathwaysGEO의 12 가지 질병에 대한 24 개의 유전자 발현 데이터 세트와 15 개 이상의 표적을 가진 DSigDB의 165 개의 FDA 승인 의약품에 대한 표적 유전자 세트로 NGSEA에 의한 약물의 우선순위를 결정하였다.

[0106] CMap 및 NGSEA의 24 가지 질병 관련 유전자 발현 데이터 세트 각각에 대해 알려진 약물을 검색할 수 있는 능력을 비교하였다.

[0107] 벤치마킹을 위해 Comparative Toxicogenomics Database(CTD)의 '치료' 범주에서 2,109 가지 질병과 1,481 가지 화학 물질 간에 17,063 개의 연관성을 확인하였다.

[0108] 알려진 약물 회수의 성능은 area under the receiver operating characteristic curve(AUROC)으로 벤치마킹 하였다.

[0109] 시험 약물 차이에 의한 편향된 평가를 방지하기 위해 CMap 및 NGSEA 모두에서 포함된 약물로 AUROC 분석을 수행하였다.

[0110] 도 3B를 참조하면, NGSEA의 약물 치료에 대한 AUROC는 CMap과 비교하여 유의하게 개선되었다($P=9.62e^{-4}$, Wilcoxon signed rank test).

[0111] 구체적으로, NGSEA에서 일치된 질병 유전자 발현 데이터 세트에 대한 알려진 약물의 회복은 CMap과 비교하여 16 건에서 24 건으로 향상되었다.

[0112] NGSEA는 특히 항암제 검색에 효과적이었다.

[0113] NGSEA에 의한 16 건의 암 관련 발현 데이터 중 14건(87.5 %)에서 향상된 성능이 관찰되었는데, 상기 결과는 의약 표적 정보가 있는 NGSEA는 항암제 재조정에 있어서 효과적인 접근법이 될 수 있음을 시사한다.

[0114] **실험예 4 : 약물 치료에 의한 항암 효과 분석**

[0115] MTS(3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium) 분석을 통해 약물 치료 후 세포 생존력을 측정하였다.

[0116] 실험예 3에서 선택한 약물 후보를 24, 48 및 72 시간 동안 50 내지 250 μM 농도로 대장암 세포주인 HCT116 또는 HT-29에 처리하였고, MTS 시약을 추가하였다. ELISA 마이크로 플레이트 판독기(Molecular Devices, USA)상에서 490 nm 흡광도를 측정하여 세포 생존율을 계산하였다. 모든 실험을 6 번 반복하였다.

[0117] 도 4A를 참조하면, 알려진 항암제의 회수 성능은 NGSEA에 의한 대장암(GSE9348)에서 가장 크게 개선되었다. AUROC 값은 CMAP 및 NGSEA에서 각각 0.488 및 0.775로 측정되었다.

[0118] 도 4B를 참조하면, NGSEA에 의한 대장암 치료에 대한 30 가지 예측 중 6 가지 화학 물질이 현재 대장암에 사용되는 약물이었고, 이 중 3 가지 화학 물질은 대장암 치료(https://clinicaltrials.gov/)의 임상 시험을 거쳤다.

[0119] 후속 실험 검증을 위해 대장암에 대한 항암 효과의 증거가 없는 것으로 알려진 나머지 후보 중dobutamin(5 위) 및 budesonide(17 위)의 대장암에 대한 항암 효과를 확인하였다. 상기 dobutamine과 budesonide는 Sigma에서 구입하였다.

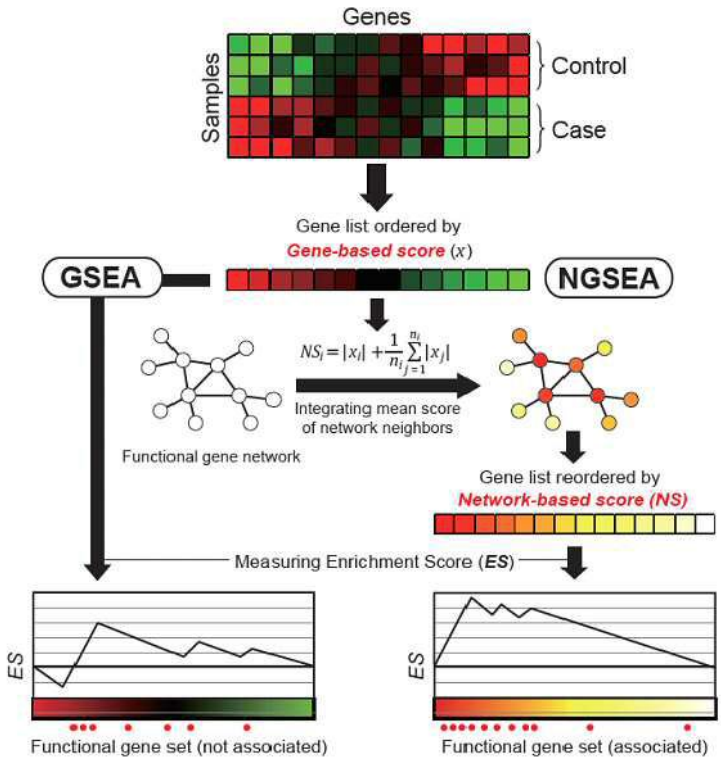
[0120] 도 4C 및 도 4D를 참조하면, 대장암 세포주 HCT116 및 HT-29를 사용한 세포 생존능 분석에서 budesonide를 처리한 경우 암세포 성장을 유의하게 억제하였다.

[0121] 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 실시 예들은 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.

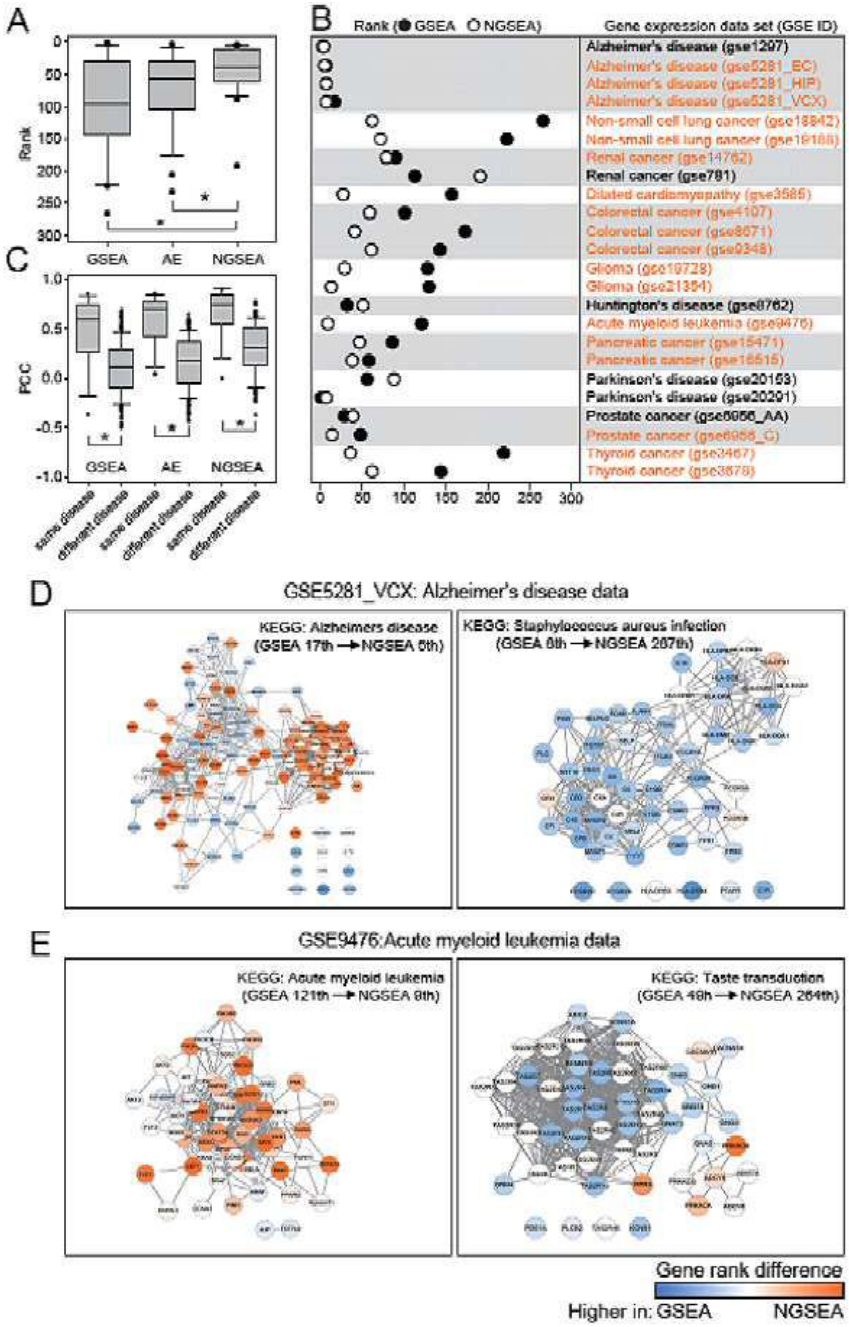
[0122] 본 발명의 범위는 후술하는 특허청구범위에 의해 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

도면

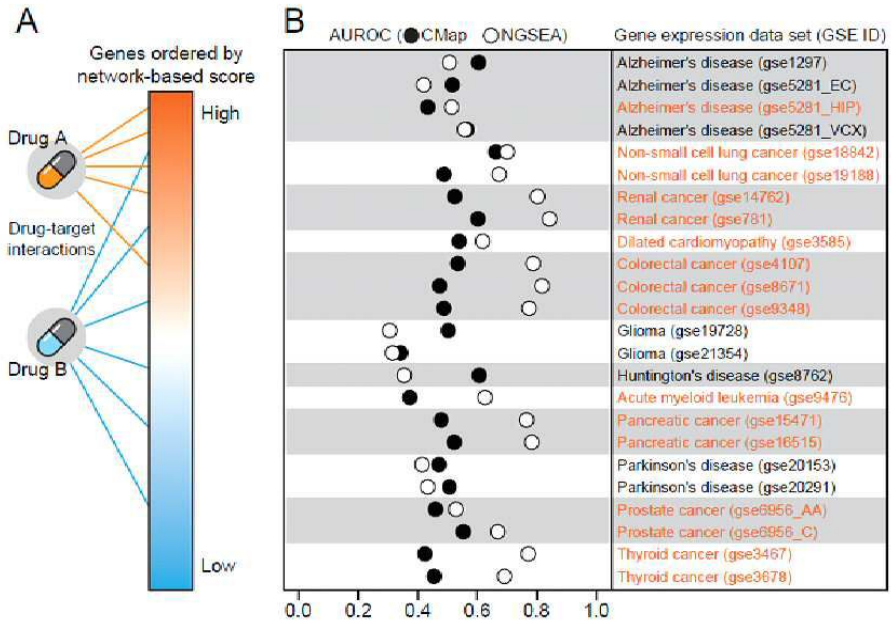
도면1



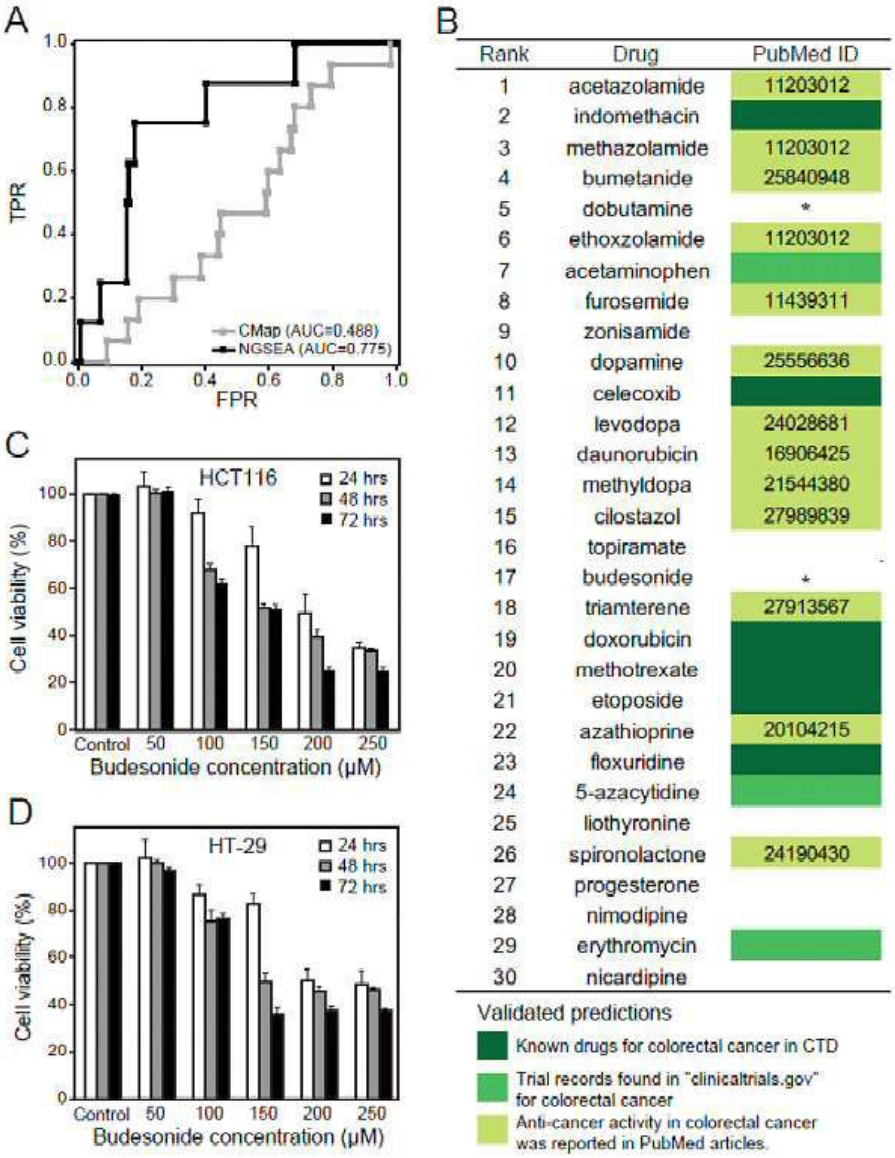
도면2


A
Rank
GSEA
AE
NGSEA
B
Rank (GSEA NGSEA)
Gene expression data set (GSE ID)
Alzheimer's disease (gse1297)
Alzheimer's disease (gse5281_EC)
Alzheimer's disease (gse5281_HIP)
Alzheimer's disease (gse5281_VCX)
Non-small cell lung cancer (gse18842)
Non-small cell lung cancer (gse19188)
Renal cancer (gse14762)
Renal cancer (gse781)
Dilated cardiomyopathy (gse3585)
Colorectal cancer (gse4107)
Colorectal cancer (gse8671)
Colorectal cancer (gse9348)
Glioma (gse19728)
Glioma (gse21354)
Huntington's disease (gse8762)
Acute myeloid leukemia (gse9476)
Pancreatic cancer (gse15471)
Pancreatic cancer (gse16515)
Parkinson's disease (gse20163)
Parkinson's disease (gse20291)
Prostate cancer (gse6956_AA)
Prostate cancer (gse6956_C)
Thyroid cancer (gse2467)
Thyroid cancer (gse3678)
C
PCC
GSEA
AE
NGSEA
same disease
different disease
D
GSE5281_VCX: Alzheimer's disease data
KEGG: Alzheimers disease
(GSEA 17th → NGSEA 5th)
KEGG: Staphylococcus aureus infection
(GSEA 0th → NGSEA 267th)
E
GSE9476:Acute myeloid leukemia data
KEGG: Acute myeloid leukemia
(GSEA 121th → NGSEA 8th)
KEGG: Taste transduction
(GSEA 48th → NGSEA 264th)
Gene rank difference
Higher in: GSEA
NGSEA

도면3


A
Genes ordered by network-based score
High
Low
Drug A
Drug-target interactions
Drug B
B
AUROC (CMap NGSEA)
Gene expression data set (GSE ID)
Alzheimer's disease (gse1297)
Alzheimer's disease (gse5281_EC)
Alzheimer's disease (gse5281_HIP)
Alzheimer's disease (gse5281_VCX)
Non-small cell lung cancer (gse18842)
Non-small cell lung cancer (gse19188)
Renal cancer (gse14762)
Renal cancer (gse781)
Dilated cardiomyopathy (gse3585)
Colorectal cancer (gse4107)
Colorectal cancer (gse8671)
Colorectal cancer (gse9348)
Glioma (gse19728)
Glioma (gse21354)
Huntington's disease (gse8762)
Acute myeloid leukemia (gse9476)
Pancreatic cancer (gse15471)
Pancreatic cancer (gse16515)
Parkinson's disease (gse20153)
Parkinson's disease (gse20291)
Prostate cancer (gse6956_AA)
Prostate cancer (gse6956_C)
Thyroid cancer (gse3467)
Thyroid cancer (gse3678)
0.0
0.2
0.4
0.6
0.8
1.0

도면4


A
TPR
FPR
CMap (AUC=0.488)
NGSEA (AUC=0.775)
B
Rank
Drug
PubMed ID
1 acetazolamide 11203012
2 indomethacin
3 methazolamide 11203012
4 bumetanide 25840948
5 dobutamine *
6 ethoxzolamide 11203012
7 acetaminophen
8 furosemide 11439311
9 zonisamide
10 dopamine 25556636
11 celecoxib
12 levodopa 24028681
13 daunorubicin 16906425
14 methyldopa 21544380
15 cilostazol 27989839
16 topiramate
17 budesonide *
18 triamterene 27913567
19 doxorubicin
20 methotrexate
21 etoposide
22 azathioprine 20104215
23 floxuridine
24 5-azacytidine
25 liothyronine
26 spironolactone 24190430
27 progesterone
28 nimodipine
29 erythromycin
30 nicardipine
Validated predictions
Known drugs for colorectal cancer in CTD
Trial records found in "clinicaltrials.gov" for colorectal cancer
Anti-cancer activity in colorectal cancer was reported in PubMed articles.
C
HCT116
24 hrs
48 hrs
72 hrs
Cell viability (%)
Control 50 100 150 200 250
Budesonide concentration (μM)
D
HT-29
24 hrs
48 hrs
72 hrs
Cell viability (%)
Control 50 100 150 200 250
Budesonide concentration (μM)