

KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크숍 (온라인)



**Multi-omics driven systematic
approaches to understand
cancer complexity**

김권일 _ 경희대학교



KSBI
KOREAN SOCIETY FOR
BIOINFORMATICS

| 한국생명정보학회



본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBI-BIML 2024

Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크숍인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크숍은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크숍은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의를 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의를 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구 중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의를 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이 인 석

Multi-omics driven systematic approaches to understand cancer complexity

생물정보 흐름의 핵심요소인 DNA, RNA, 그리고 단백질의 서열과 발현양에 대한 data는 관련 기술의 비약적인 발전에 힘입어 big data 수준으로 축적되고 있다. 이에 따라 각 요소의 전체(계) 수준의 양상을 연구하는 omics 분야가 태동하게 되고, 나아가 각 계 사이의 상호작용을 통합 연구하는 multi-omics 분야가 현대 생물학에 자리잡게 되었다. 최근 이러한 상호작용을 강조하는 trans-omics 라는 개념도 등장하였고, multi-omics 연구에는 다양한 접근과 해석이 공존하고 있는 상태이다.

본 강의에서는 최근 발표된 주요한 multi-omics 연구 사례를 대표적인 복잡 질병(complex disease)인 암을 대상으로 하여 소개하고자 한다. 특히, 각 계 내의 복잡계가 중첩된 양상을 나타내는 multi-omics 복합 층계를 이해하기 위해서는 시스템 생물학적 이해가 필수적인데, 이와 관련된 시스템 생물학 기반의 multi-omics 융합 연구 내용을 공유할 것이다. 이는 환자군을 분류하고 각 환자군을 해석하는 데 있어 multi-omics 특화된 클러스터링 등의 방법론들을 포함한다. 마지막으로 multi-omics 요소간 상호작용을 투영시켜 환자 예후를 예측하거나 치료 타겟을 발굴하는 AI 확장 적용 사례들을 다루고자 한다. 이를 통하여 multi-omics data에 내포된 생물학적 상호작용을 해석하고 이해하는 핵심 역량을 갖추는 것을 목표로 한다.

강의는 다음의 내용을 포함한다:

- Multi-omics 개요 및 암 생물학에서의 대표적인 multi-omics 연구
- 시스템 생물학 기반의 multi-omics data 해석
- Multi-omics data clustering 및 해석
- 생물학 network 기반 AI 확장 적용

* 강의 난이도: 중급

* 강의: 김권일 교수 (경희대학교 생물학과)

Curriculum Vitae

Speaker Name: Kwoneel Kim, Ph.D.



► Personal Info

Name Kwoneel Kim
Title Assistant Professor
Affiliation Kyung Hee University

► Contact Information

Address 24, Kyunghedae-ro, Dongdaemun-gu, Seoul
Email kwoneelkim@khu.ac.kr
Phone Number 02-961-9612

Research Interest

Translational bioinformatics, Machine learning and computational genomics, Cancer genomics

Educational Experience

2009 B.S. Dept. Applied Bioscience, Konkuk University, Korea
2011 M.S. Dept. Functional Genomics, UST, Korea
2015 Ph.D. Dept. Bio and Brain Engineering, KAIST, Korea

Professional Experience

2015-2017 Post-Doctoral Researcher, Dept. Bio and Brain Engineering, KAIST
2017-2018 Senior Research Scientist, Asan Institute for Life Sciences, Asan Medical Center
2018- Assistant Professor, Department of Biology, Kyung Hee University

Selected Publications (5 maximum)

1. Bang, H., Park, J. S., Kim, J. Y., Sung, C., An, J., Cho, D. Y., ... & **Kim, K.** Gene essentiality for tumour growth influences neoantigen-directed immunoediting. *Clinical and Translational Medicine*, 12(1). (2022)
2. Kim J-H*, **Kim K***, Yeom J*, Lee E, Kang M-J, Lee S-H, Kim K, Lee S-Y, Hong S-B, Oh DK, Lee K, Choi, S-J, Yang M-J, Kim J, Hong S-J. Integrative multi-omics approach for mechanism of humidifier disinfectant-associated lung injury. *Clinical and Translational Medicine*, 11, e562 (2021) *Co-first
3. **Kim K**, Kim HS, Jeong YK, Jung H, Sun J-M, Ahn JS, Ahn M-J, Park K, Lee S-H, Choi JK. Predicting clinical benefit of immunotherapy by antigenic or functional mutations affecting tumour immunogenicity. *Nature Communications*, 11. 951 (2020).
4. Jang K*, **Kim K***, Cho A, Lee I, Choi JK. Network perturbation by recurrent regulatory variants in cancer. *PLoS Computational Biology*, 13. e1005449 (2017). *Co-first
5. **Kim K***, Jang K*, Yang W*, Choi EY, Park SM, Bae M, Kim YJ, Choi JK. Chromatin structure-based prediction of recurring noncoding mutations in cancer. *Nature Genetics*, 48.1321-1326 (2016). *Co-first

KSBi-BIML 2024

Multi-omics driven systematic approaches to understand cancer complexity

Department of Biology, Kyung Hee University

Kwoneel Kim, PhD

□강의개요.

Multi-omics driven systematic approaches to understand cancer complexity.

생물정보 흐름의 핵심요소인 DNA, RNA, 그리고 단백질의 서열과 발현양에 대한 data는 관련 기술의 비약적인 발전에 힘입어 big data 수준으로 축적되고 있다. 이에 따라 각 요소의 전체(계) 수준의 양상을 연구하는 omics 분야가 태동하게 되고, 나아가 각 계 사이의 상호작용을 통합 연구하는 multi-omics 분야가 현대 생물학에 자리잡게 되었다. 최근 이러한 상호작용을 강조하는 trans-omics 라는 개념도 등장하였고, multi-omics 연구에는 다양한 접근과 해석이 공존하고 있는 상태이다.

본 강의에서는 최근 발표된 주요한 multi-omics 연구 사례를 대표적인 복잡 질병(complex disease)인 암을 대상으로 하여 소개하고자 한다. 특히, 각 계 내의 복잡계가 중첩된 양상을 나타내는 multi-omics 복잡 증계를 이해하기 위해서는 시스템 생물학적 이해가 필수적인데, 이와 관련된 시스템 생물학 기반의 multi-omics 융합 연구 내용을 공유할 것이다. 이는 환자군을 분류하고 각 환자군을 해석하는 데 있어 multi-omics 특화된 클러스터링 등의 방법론들을 포함한다. 마지막으로 multi-omics 요소간 상호작용을 투영시켜 환자 예후를 예측하거나 치료 타겟을 발굴하는 AI 확장 적용 사례들을 다루고자 한다. 이를 통하여 multi-omics data에 내포된 생물학적 상호작용을 해석하고 이해하는 핵심 역량을 갖추는 것을 목표로 한다.

강의는 다음의 내용을 포함한다.

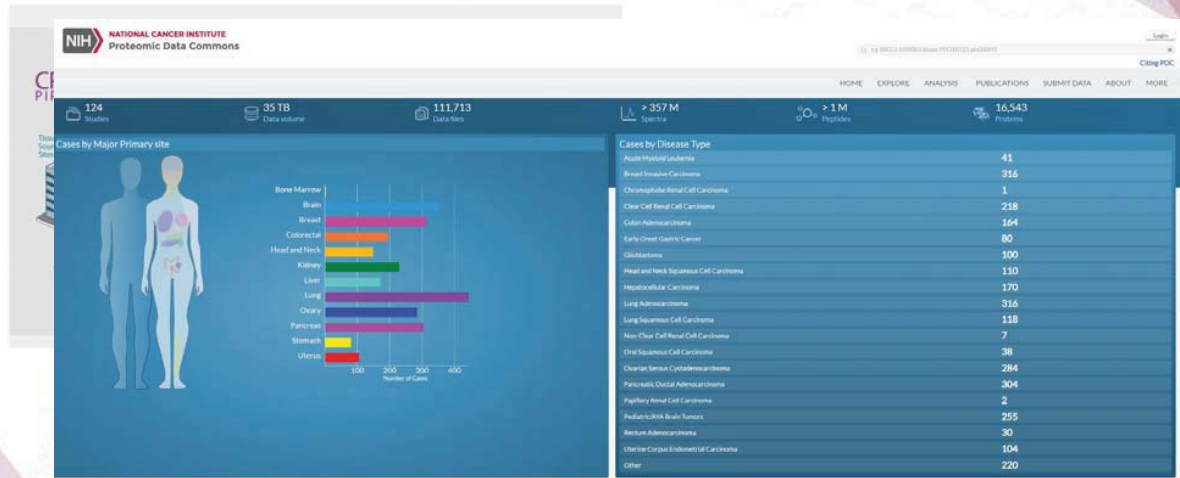
- Multi-omics 개요 및 암 생물학에서의 대표적인 multi-omics 연구.
- 시스템 생물학 기반의 multi-omics data 해석.
- Multi-omics data clustering 및 해석.
- 생물학 network 기반 AI 확장 적용.

* 강의 난이도: 중급

* 강의: 김권일 교수 (경희대학교 생물학과)

Clinical Proteomic Technology Assessment for Cancer (CPTAC)

CPTAC



5

The cancer genome atlas (TCGA) International cancer genome consortium (ICGC) Pan-cancer analysis of whole genomes (PCWAG)

PAN-CANCER ANALYSIS OF WHOLE GENOMES PUBLISHED!

nature
The International Journal of Science (4 February 2020)
CANCER CATALOGUED
Whole genome sequences for 38 types of tumour

Read about the ICGC/TCGA analysis of >2,600 whole cancer genomes across 38 tumour types in 23 papers published in Nature and other Nature journals. Photo credit: Nik Spencer/Nature.

ICGC Data Portal

ADVANCED SEARCH

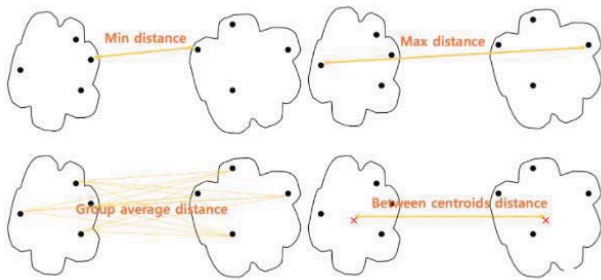
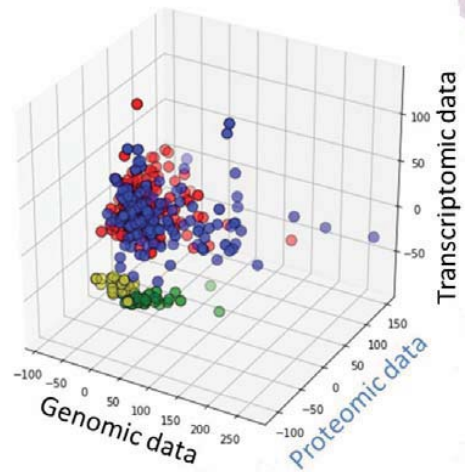
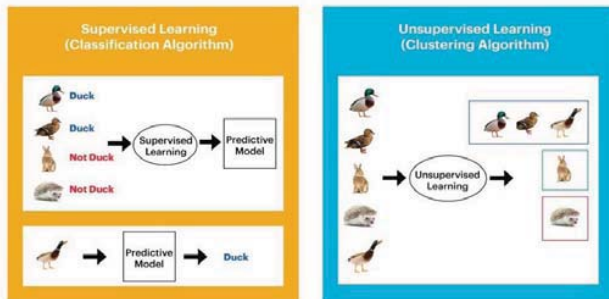
Filters: Donors, Genes, Mutations, Donors, Genes, Mutations, Vital Status, Primary Site, Gender, Tumour Stage, Disease Status, Relapse Type, Age, Data Types, Analysis Types.

Donors
Showing 1 - 10 of 24,289 donors

ID	Project	Site	Gender	Age	Stage	Subtotal (56/57)	seq	copy	copy	copy	meth	meth	meth	meth	meth	meth	meth	meth	# Mutations	# Genes		
D00232761	PBCA-US	Brain	Male	3		2,992	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1,378,562	58,205
D00232224	PBCA-US	Brain	Female	7		547	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1,615,334	53,338
D00232843	MELA-AU	Skin	Male	76	RC	907	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	944,360	31,365
D00229257	MELA-AU	Skin	Male	62	RB	1,110	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	986,186	49,858
D0050970	LICA-IR	Liver	Female	72			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	616,827	48,874
D00229263	MELA-AU	Skin	Male	81	RC	154	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	773,368	42,879
D00230886	MELA-AU	Skin	Male	56	SA/IB	2,736	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	518,365	46,681
D00230877	MELA-AU	Skin	Male	25	IB	427	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	428,219	46,248
D00232875	MELA-AU	Skin	Male	79	IB	1,192	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	819,954	45,993

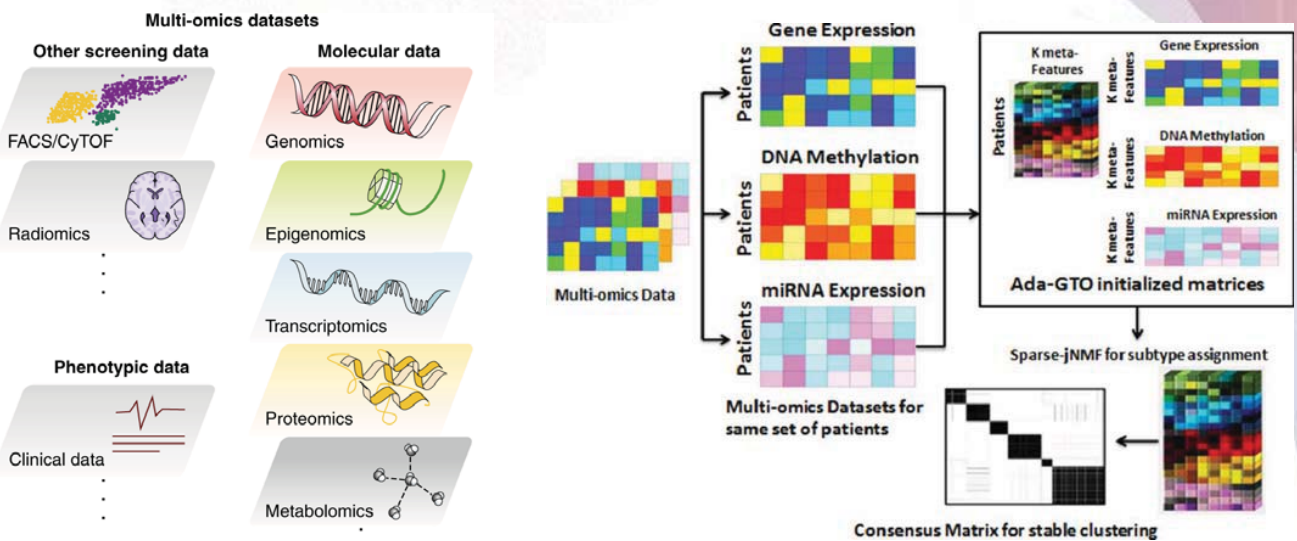
6

멀티오믹스 데이터 클러스터링을 통하여 생물학적 데이터의 숨겨진 분류를 더 잘 찾을 수도 있다



7

멀티오믹스 데이터는 각각의 성질이 매우 다르고 feature 가 매우 많기 때문에 (high dimensionality) 클러스터링이 잘 안될 수 있는데, 차원축소 기반의 접근이 이에 대한 도움을 줄 수 있다



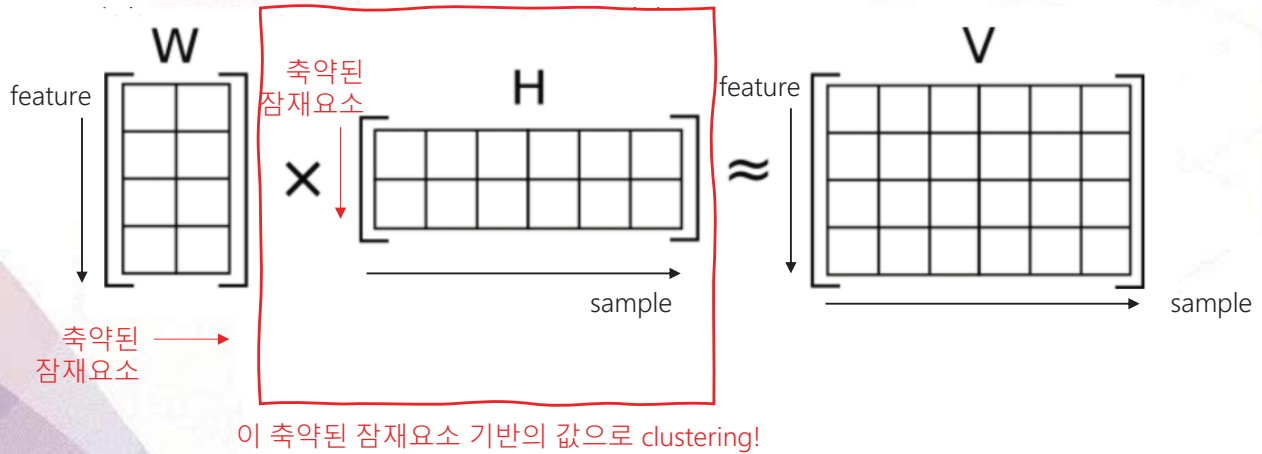
8

Non-Negative Matrix Factorization (NMF) clustering

NMF는 행렬 인수 분해 알고리즘 중 하나:

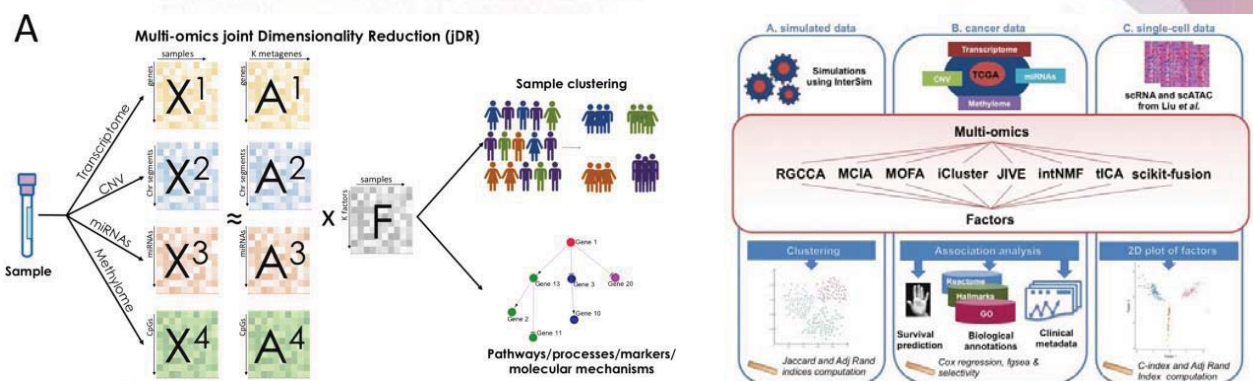
원본 행렬 V 가 있다고 했을 때, 행렬 인수 분해는 이 V 행렬을 두개의 행렬로 분리 하는 것 (이 때, V, H, W 는 모두 음수가 없는 데이터)

분해 행렬 W 는 원본 행에 대해서 이 잠재 요소에 값이 얼마나 되는지에 대응하며, 분해 행렬 H 는 이 잠재 요소가 원본 열 (즉, 원본속성)으로 어떻게 구성됐는지를 나타냄



9

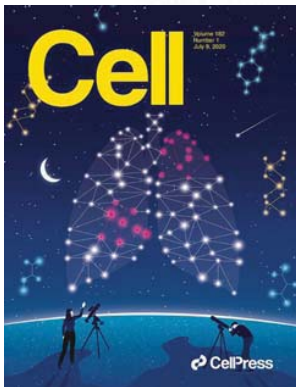
여러 이점 때문에 다양한 차원축소 기반의 클러스터링 알고리즘이 적용되고 있다



1. 축약된 feature selection 이 가능
2. Data 및 noise redundancy 를 줄이는데 탁월
3. computational efficiency

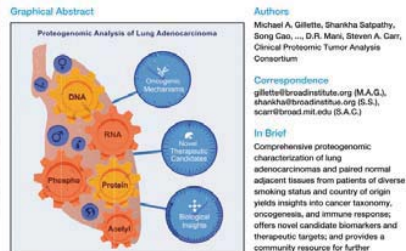
10

Milestone papers of multiomics study in lung cancer (2020): proteogenomics



Cell Resource

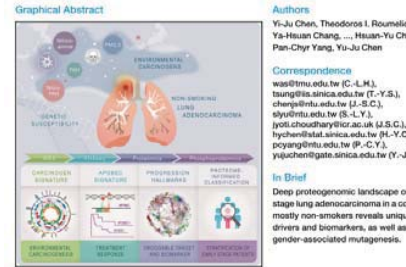
Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma



- Highlights**
- Comprehensive LUAD proteogenomics exposes multi-omic clusters and immune subtypes
 - Phosphoproteomics identifies candidate ALK-fusion diagnostic markers and targets
 - Candidate drug targets: PTPN11 (EGFR), SOS1 (KRAS), neopterin degradation (STK11)
 - Phospho and acetyl modifications denote tumor-specific markers and druggable proteins

Cell Resource

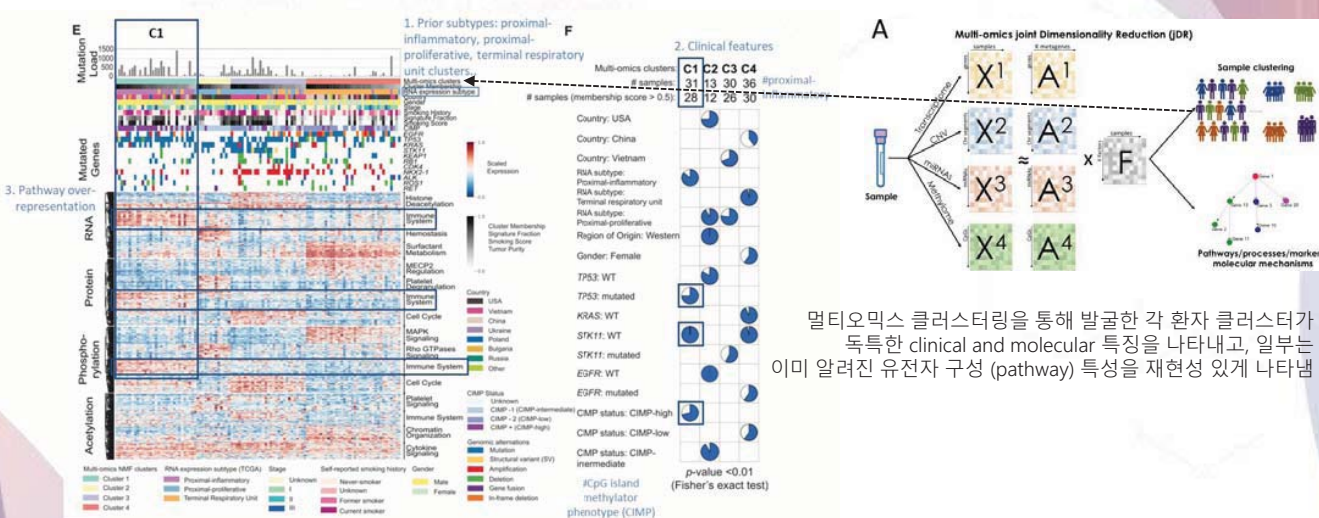
Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression



- Highlights**
- First deep proteogenomic landscape of non-smoking lung adenocarcinoma in East Asia
 - Identified age, sex-related endogenous, and environmental carcinogenic mutagenic processes
 - Proteome-informed classification distinguished clinical features within early stages
 - Protein networks identified tumorigenesis hallmarks, biomarkers, and druggable targets

Cell, 182, 200–225 (2020)
Cell, 182, 226–244 (2020)

NMF-based unsupervised “multiomics clustering”: 각 멀티오믹스 클러스터에 속한 환자들은 고유의 임상적 혹은 분자세포생물학적 특징을 나타낸다

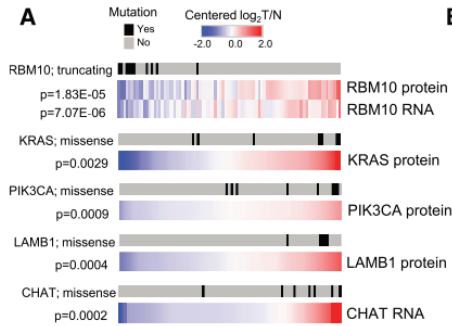


Cell, 182, 200–225 (2020)

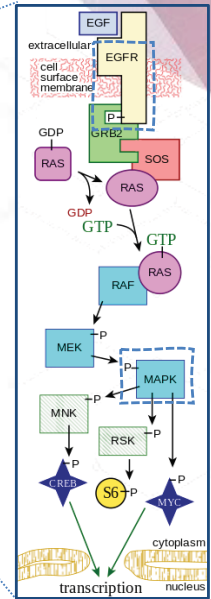
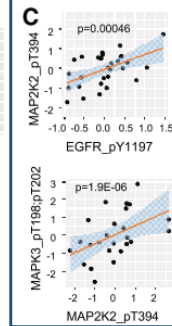
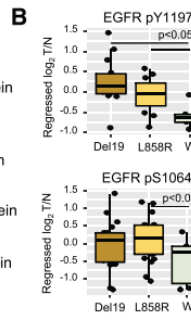
환자의 유전체 변동이 전사체 및 단백질체 변화에 직간접적으로 연관되어 있다

#EGFR activating mutations

#Scatterplots of cophosphorylation within the EGFR-MEK-ERK axis



#Possibly through increased stability



"Although the impact of mutations in EGFR protein abundance was not conclusive, EGFR activating mutations correlated with increased phosphorylation of S1064 and Y1197, reflecting the activation of mutated EGFR in the patients."

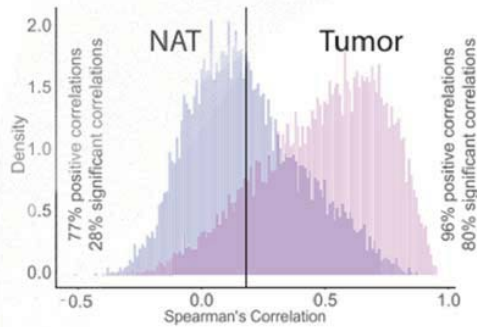
Cell. 182, 226-244 (2020)

13

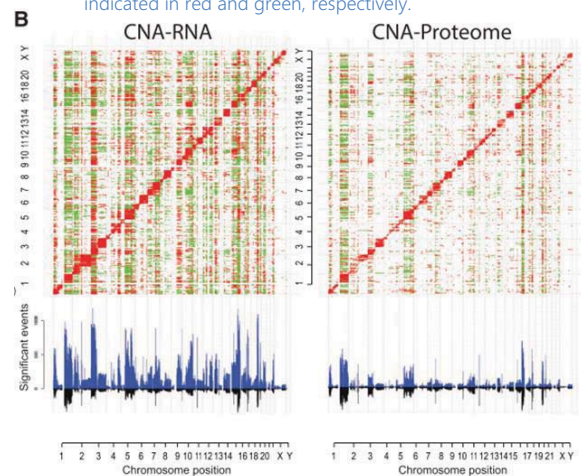
유전자 중심으로 볼 때, CNA-mRNA-protein 간 상관관계가 상당히 다르다

Gene A ---proteome---
Gene A ---transcriptome----

Mean = 0.18 Mean = 0.49
Median = 0.15 Median = 0.53



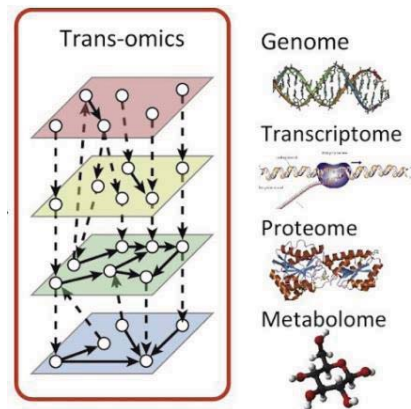
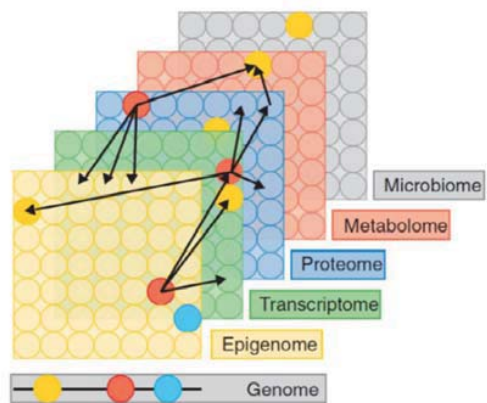
#Significant positive and negative correlations are indicated in red and green, respectively.



"CNA correlations were broadly comparable but considerably dampened at the levels of proteins and PTMs"

14

Changes of a layer of one omics do not always transfer into other layer:
Complexity in regulating expression of RNA and protein



Genome Biology. 18, 83 (2017)
Trends in Biotechnology. 34, 276–290 (2016)

15

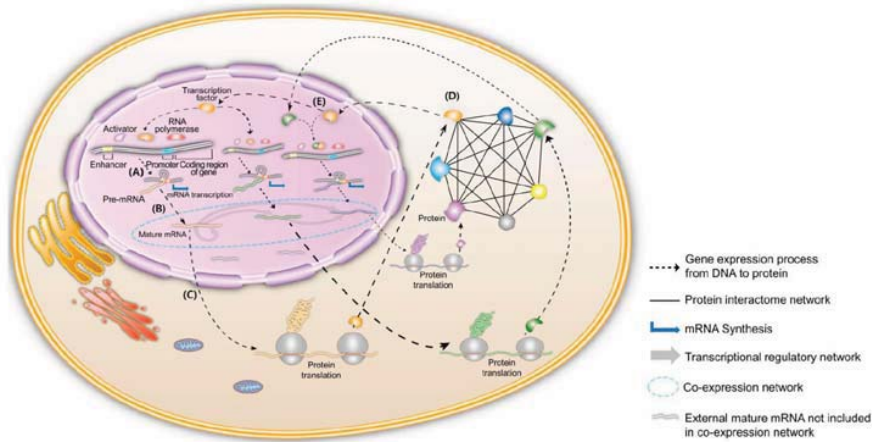
Summary

#단일 오믹스로 밝히지 못한 생물학적 양상을 멀티오믹스 클러스터링을 통해 접근할 수 있다

#클러스터링 분석은 새로운 샘플 군집과 그 군집에 대한 생물학적 해석을 가능하게 한다

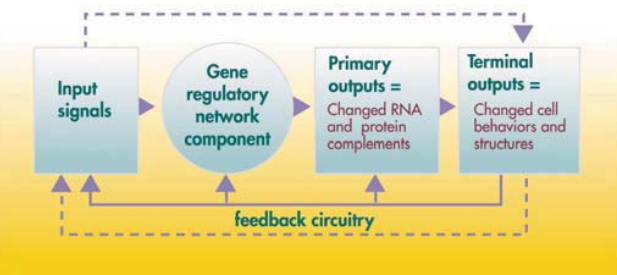
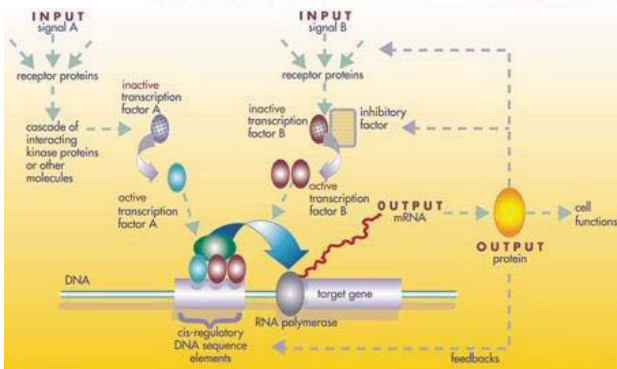
16

Multi-omics data for systems biology in cell systems



Anim Cells Syst. 30, 1-7 (2020)

Gene regulatory network



YGG 01-0086

TRRUST version 2

Transcriptional Regulatory Relationships Uncovered by Sentence-based Text Mining

Home
Search
Download

TRRUST (version 2) is a manually curated database of human and mouse transcriptional regulatory networks.

Current version of TRRUST contains 8,444 and 4,552 TF-target regulatory relationships of 803 human TFs and 428 mouse TFs, respectively. They have been derived from 11,277 published articles, which describe small-scale experimental studies of transcriptional regulations. To efficiently search for regulatory relationships from over 20 million published articles, we used sentence-based text mining approach.

TRRUST database also provides information of mode of regulation (activation or repression). Currently 8,975 (99.8%) regulatory relationships are known for mode of regulation.

Tables for human genes and mouse genes included in TRRUST.

TRRUST network edge information is freely available for non-commercial research at Download page.

* TRRUST version 1 website can be found [HERE](#).

RegNetwork: Regulatory Network Repository

of Transcription Factor and microRNA Mediated Gene Regulation

HOME SEARCH ABOUT SOURCE DOWNLOAD CONTRIBUTION LINKS

*Welcome to RegNetwork! RegNetwork is a free repository of the type transcriptional and posttranscriptional regulatory relationships for human and mouse.

- TF-TF
- TF-gene
- TF-miRNA
- miRNA-TF
- miRNA-gene

RegNetwork integrates the curated regulations in various databases and the potential regulations inferred based on the transcription factor binding sites (TFBS), Transcription factor (TF) and microRNA (miRNA) on central regulations in gene regulation. They function in the transcriptional and posttranscriptional level respectively. Recently, more and more regulatory relationships in databases and literature are available. It is greatly valuable for studying gene regulation system by integrating the prior knowledge of the transcriptional regulation, between TF and target genes, and the posttranscriptional regulation between miRNA and targets. The comprehensive knowledge of transcriptional factor binding sites (TFBS) can also be implemented to compile the potential regulatory relationships between regulators and their targets.

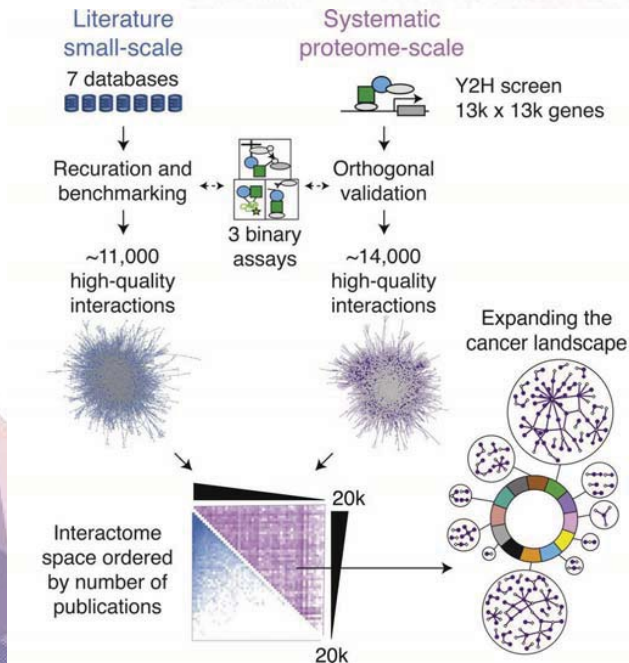
† Update:

Updated with mouse regulatory database. <http://reg.net/> Download Data!

Reference:

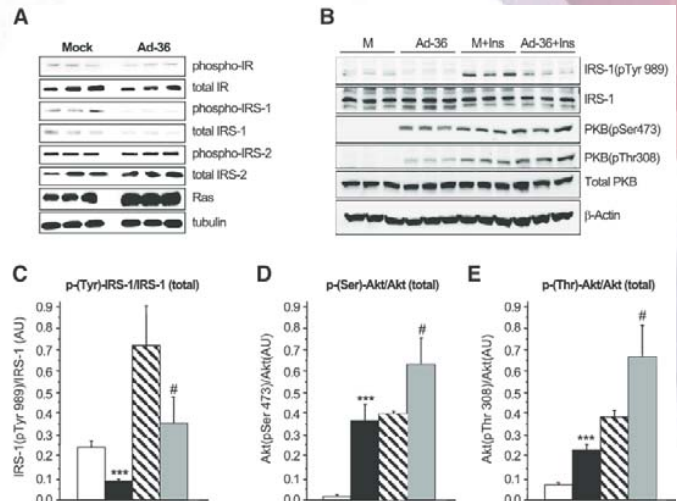
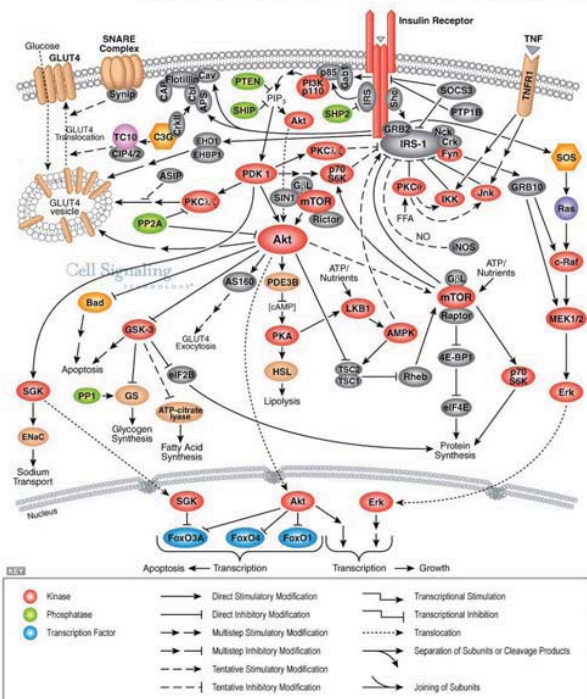
Shi, Peng, Liu, Cheng, Yu, Wang, Xiao and Wu, Yi (2012). RegNetwork: an integrated database of transcriptional and posttranscriptional regulatory networks in human and mouse. Database (Oxford), 2012, pii: bnr012. doi:10.1093/database/bnr012

Interactome network



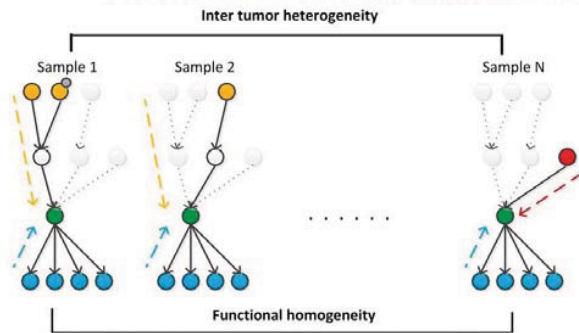
19

Signaling network



20

Network-based Integration of Multi-omics Data (NetICS) for prioritizing cancer genes



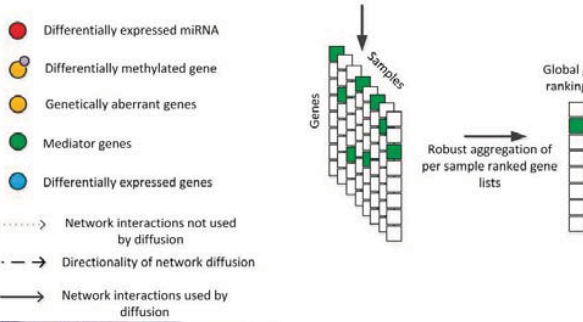
The final scores for all genes are computed as the Hadamard product

$$E = E_M \circ E_D. \quad (7)$$

"The vector E determines the mediator effect for each gene.

A large entry in E_M at position i means that gene i is proximal to many upstream-located aberrant genes or miRNA, and a large entry in E_D at position i means that gene i is proximal to many downstream-located differentially expressed genes."

Bioinformatics. 34, 2441-2448 (2018)



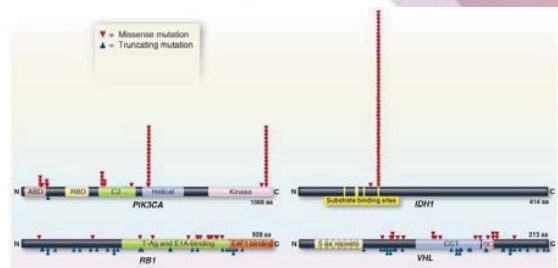
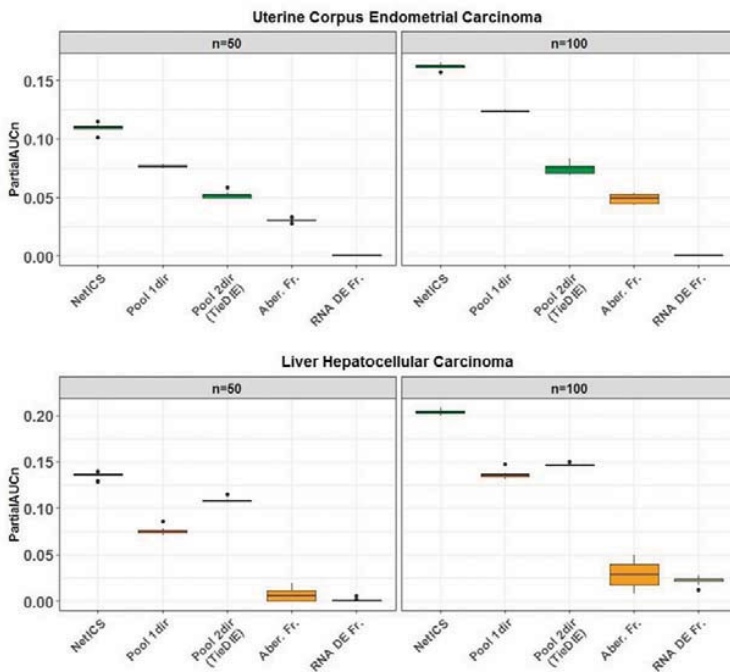
21

Prediction of known cancer genes

#Pool 1dir, 2dir: small directionality effect, no mediator effect

#Aber Fr.: ranking by frequency of aberrant genes across all samples

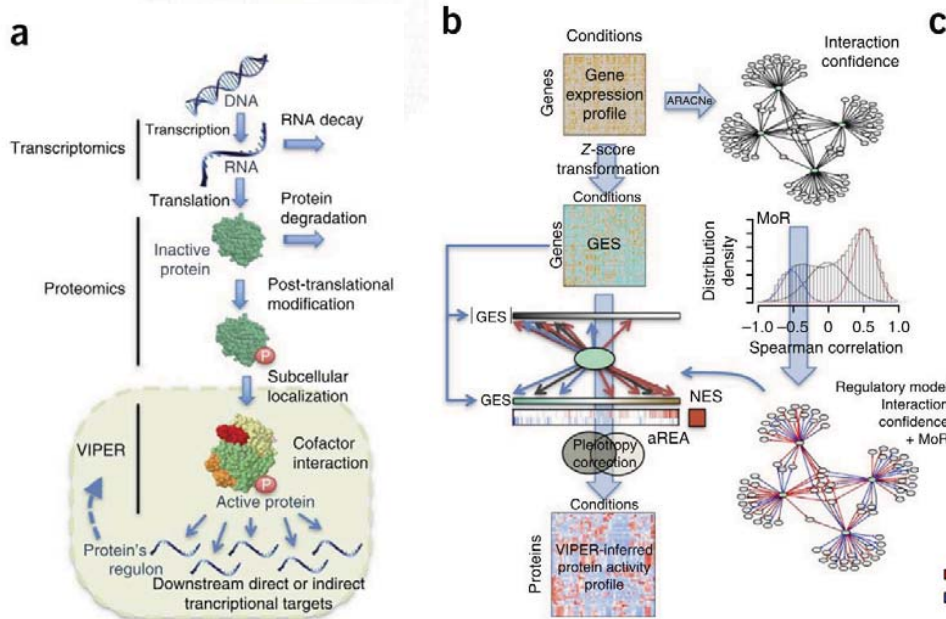
#RNA DE Fr.: ranking by frequency of differentially expressed genes across all samples



22

VIPER: virtual inference of protein activity by enriched regulon analysis

Protein activity 측정은 실험적으로나 생물학적으로나 측정하기 매우 어렵다.



Nature Genetics 48, 838–847 (2016)

23

VIPER inferred loss of protein activity following RNA interference (RNAi)-mediated silencing

RNAi data를 test data로 하여 VIPER의 protein activity inference 성능 검증 수행

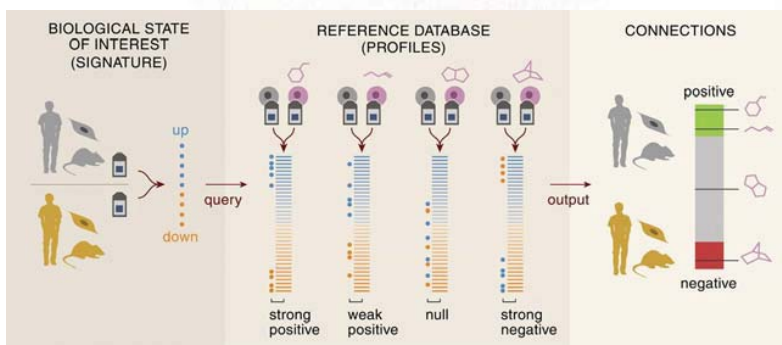
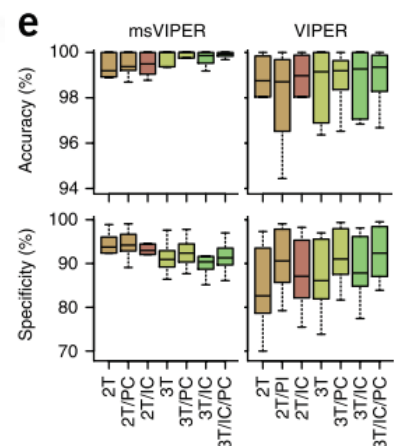


Table 2 Benchmark experiments

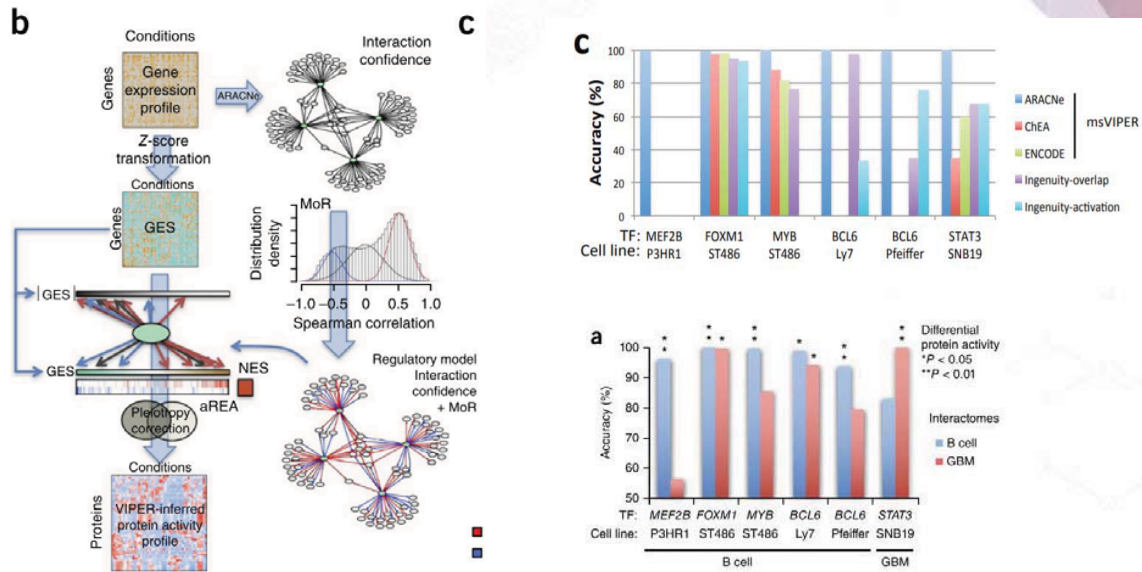
Cell line	Knockdown gene	Technology	Replicates	Profile platform	DEG ^a at $P < 0.01$
P3HR1 (lymphoma)	<i>MEF2B</i>	shRNA ^b	5	HG-U95Av2	960
ST486 (lymphoma)	<i>FOXN1</i>	shRNA ^b	3	HG-U95Av2	276
OCI-Ly7 (lymphoma)	<i>MYB</i>	shRNA ^b	3	HG-U95Av2	469
Pfeiffer (lymphoma)	<i>BCL6</i>	siRNA ^c	3	HG-U133p2	646
Pfeiffer (lymphoma)	<i>BCL6</i>	siRNA ^c	3	HG-U133p2	1,311
SNB19 (glioma)	<i>STAT3</i>	siRNA ^c	6	Illumina HT12v3	501

^aDifferentially expressed genes. ^bShort hairpin RNA. ^cSmall interfering RNA.



24

Context-specific network 사용이 protein activity inference에 중요함



25

Summary

#Multiomics system 에는 다양한 층위 및 network 가 존재

#이것을 multiomics 해석에 어떻게 적용할지는 case-dependent

#Multiomics 에 내재된 생물학적 속성을 system 기반의 분석으로 이해할 수 있음

#알맞게 고안된 생물학적 가설이 중요 (어떤 omics 데이터에 어떤 network 기반의 해석을 할지 등)

26

Predicting drug response



in Bioinformatics

Network Bioinformatics

SECTION ABOUT ARTICLES RESEARCH TOPICS FOR AUTHORS EDITORIAL BOARD

< Articles

ORIGINAL RESEARCH article

Front. Bioinform., 29 April 2021 | <https://doi.org/10.3389/fbinf.2021.639349>



Predicting Anticancer Drug Response With Deep Learning Constrained by Signaling Pathways

Heming Zhang^{1*}, Yixin Chen^{1*} and Fuhai Li^{2,3*}

¹Department of Computer Science, Washington University in St. Louis, St. Louis, MO, United States

²Institute for Informatics, Washington University School of Medicine, St. Louis, MO, United States

³Department of Pediatrics, Washington University School of Medicine, Washington University in St. Louis, St. Louis, MO, United States

27

Model for mechanism of response (MoR) of cancer drugs

#약물의 반응과 multiomics feature 간 존재하는 the mechanism of response; MoR 는 매우 복잡하여 이해하기 어려움

#이 연구에서는 omics feature 의 수를 줄이면서 model 을 해석하기 쉽게 만들기 위하여 46 여가지의 알려져 있는 biological pathway 를 사용하여 anti cancer drug response 를 예측하는 deep learning model 을 구축함

#이 model 에서는 individual gene 의 multiomics data (gene expression, copy number variation, etc.) 가 하나의 feature 로 통합됨

#1,000 여개의 cancer cell line 에 대한 multiomics data 가 있는 Broad Institute cancer cell line encyclopedia (CCLE) database 와 동일한 cancer cell line 에 대한 drug response data 가 있는 the genomics of drug sensitivity in cancer (GDS) database 를 사용

#Biological pathway 를 사용하여 feature integration 을 수행한 모델이 일반적인 deep neural network model (모든 gene feature 를 각각 사용한) 보다 나은 성능을 보임

#또한 model 을 해석하여 anticancer drug response 에 중요한 signaling pathway 의 패턴을 발굴할 수 있음

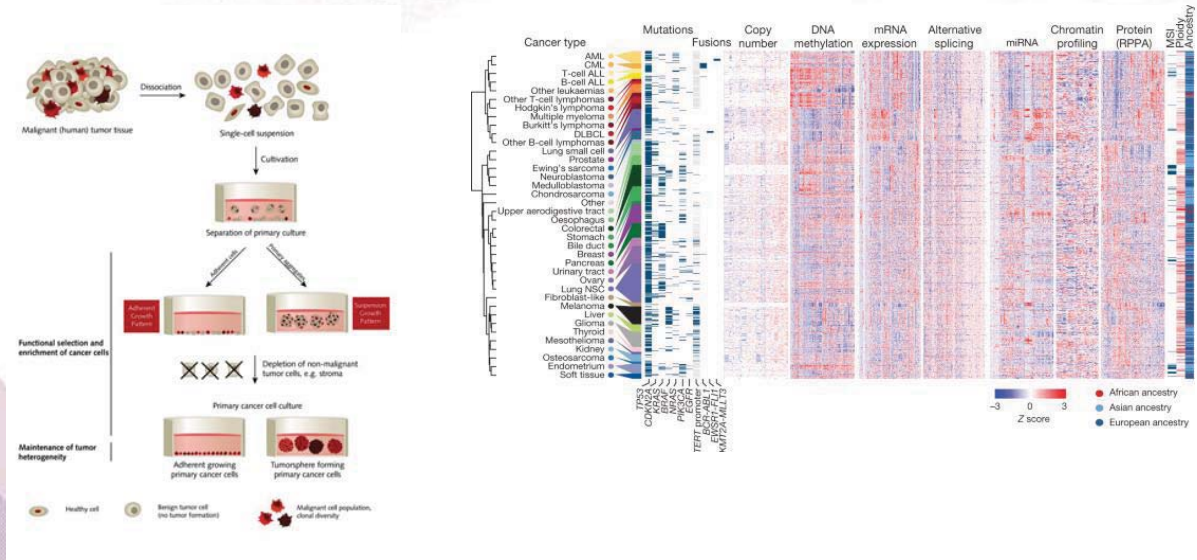


28

Cancer Cell Line Encyclopedia (CCLE)

#Cancer cell line 은 효과적인 약물을 발굴하고 그와 관련된 biomarker 를 평가하는 데 중요한 역할을 하는 experimental model

#CCLE database: 1,000 여개의 cancer cell line 에 대한 multiomics data 생산 및 공유



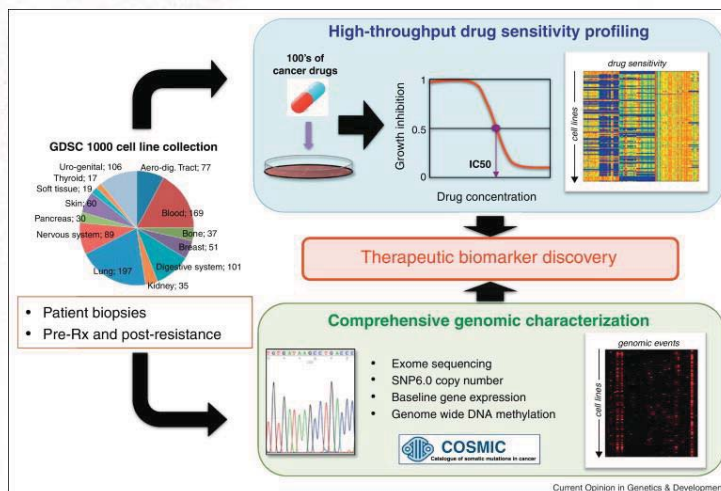
29

The Genomics of Drug Sensitivity in Cancer (GDSC)

#GDSC database: CCLE 의 1,000 여개 cancer cell lines 에 대하여 100 여개 약물 및 화합물 처리에 대한 반응 데이터 생산 및 공유

#이는 약물 반응성과 genetic biomarker 사이의 밝혀지지 않은 연관성을 연구하기 위해 제공

#NCI-ALMANAC drug combination database: 100 여개의 약물의 5,000 여개의 조합을 60 여개의 cancer cell line 에 처리한 반응 데이터 생산 및 공유



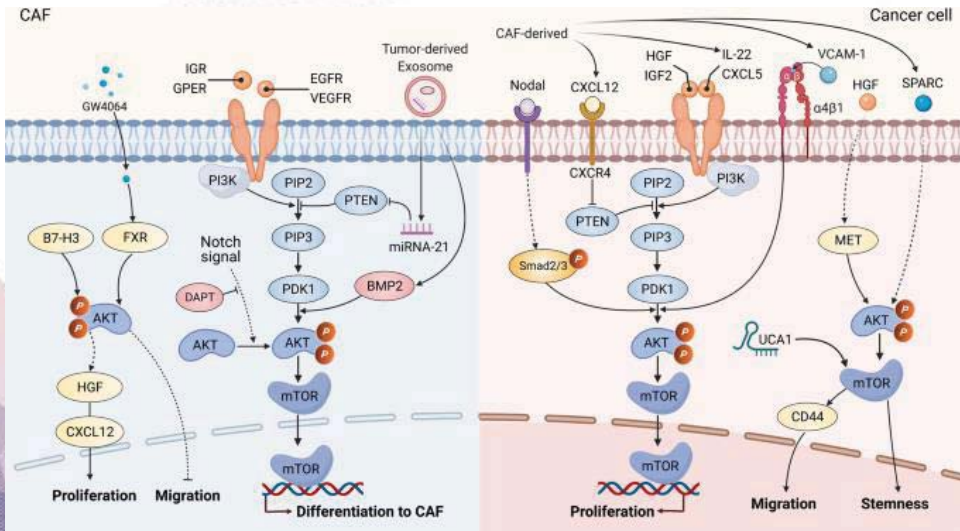
30

consDeepSignaling

#이 연구에서는 개별 gene 의 gene expression 과 copy number variation data 를 사용하여 KEGG signaling pathway database로부터 수집한 46 개의 signaling pathway 의 activity 에 연결시켜 이를 기반으로 약물 반응성을 예측하는 model을 구축

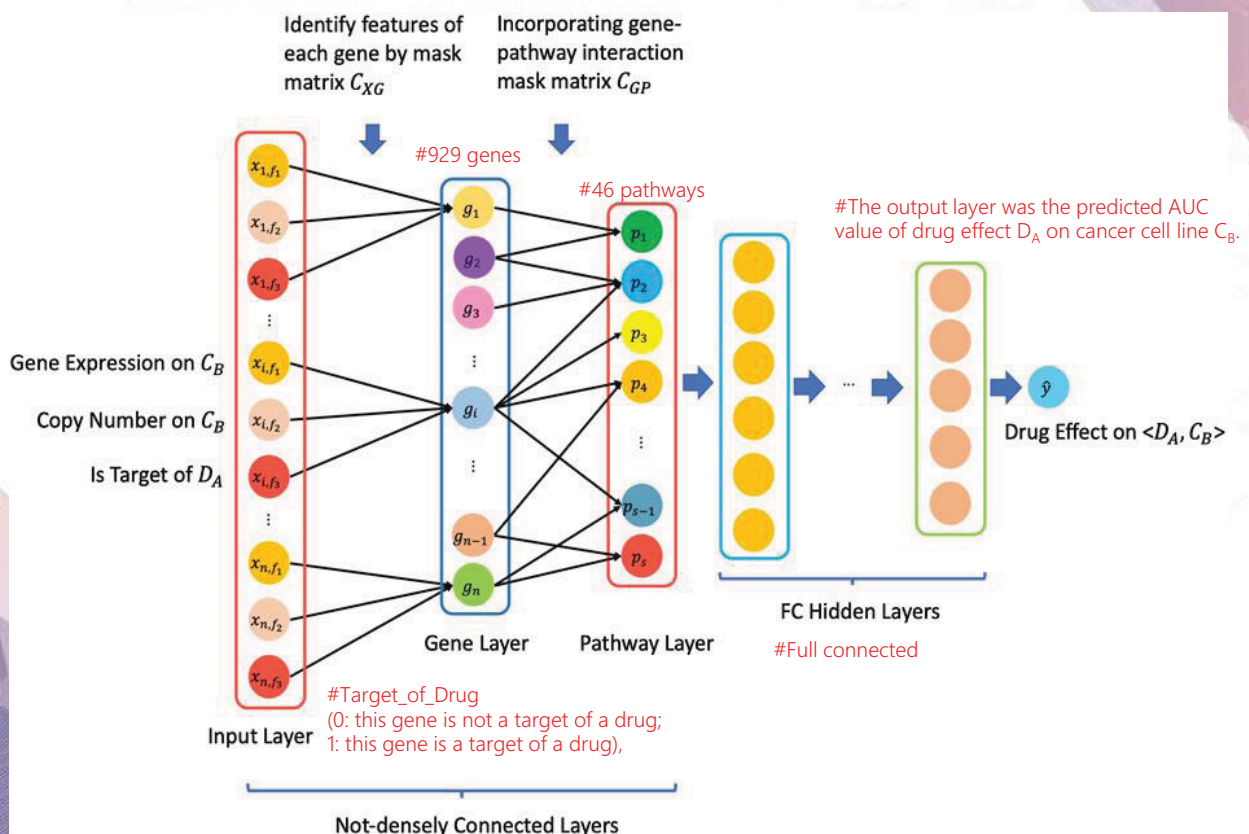
#또한, model 의 예측도에 기여하는 부분을 signaling pathway 수준으로 해석!

#암을 포함한 다양한 질병 연구에서 signaling pathway 는 일련의 gene/protein 사이의 signaling cascade 수준의 이해를 가능하게 하므로 약물 반응을 생물학적으로 이해하기에 적합한 수단임

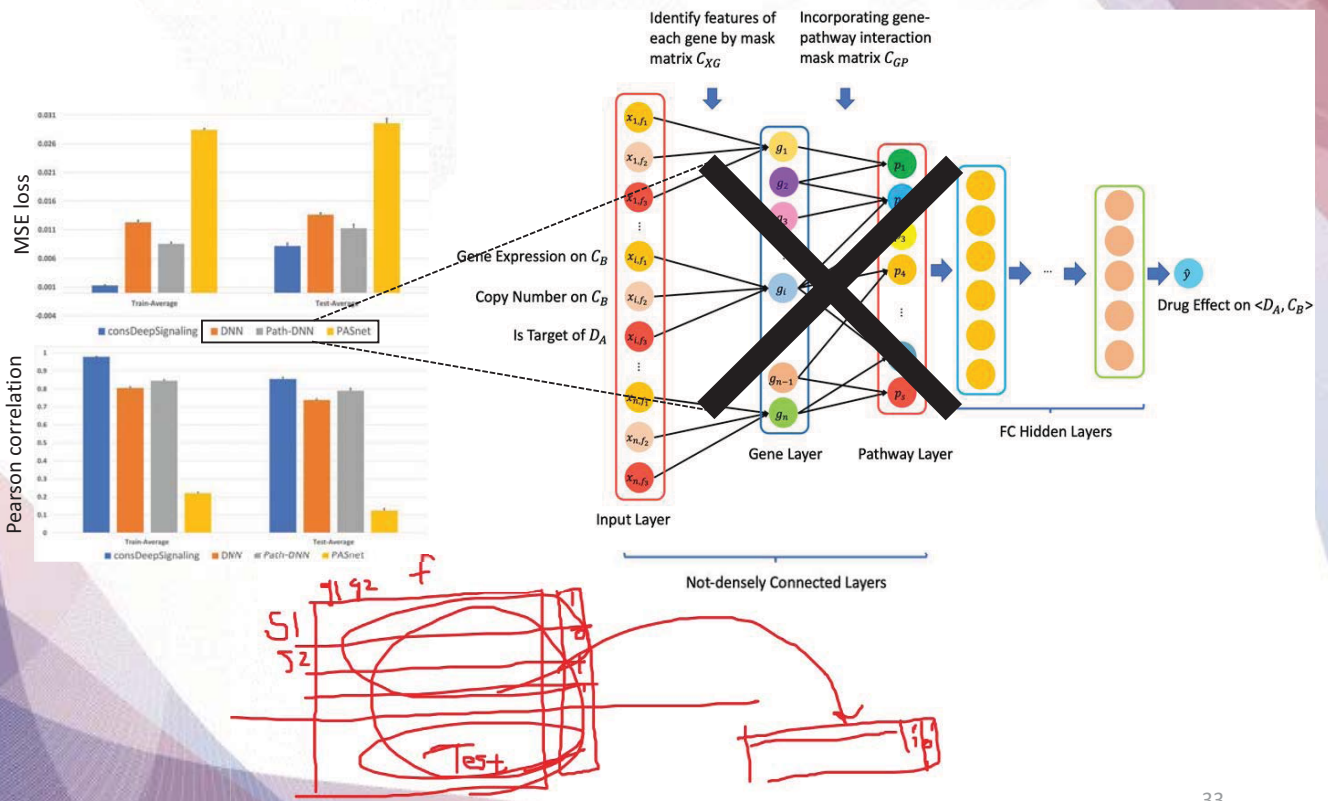


31

ConsDeepSignaling deep learning architecture

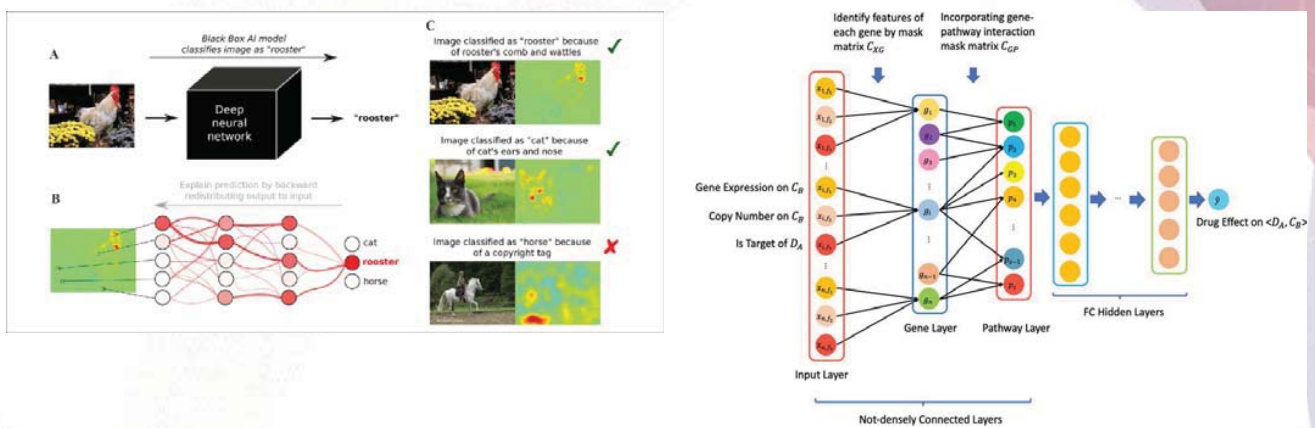


Model evaluation



33

eXplainable AI (XAI)



34

Machine learning of multiomics data

nature

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [articles](#) > [article](#)

Article | [Open Access](#) | [Published: 07 December 2021](#)

Multi-omic machine learning predictor of breast cancer therapy response

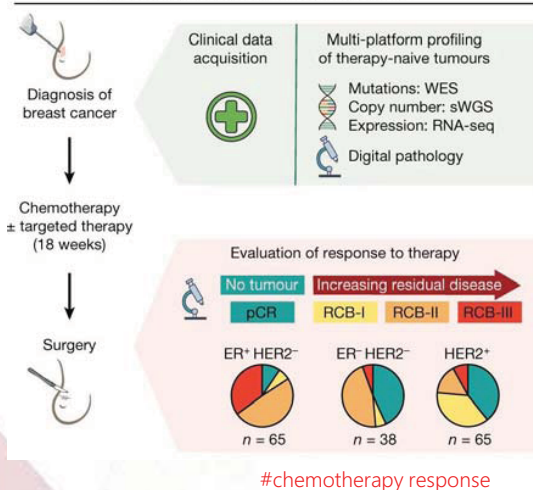
[Stephen-John Sammut](#), [Mireia Crispin-Ortuzar](#), [Suet-Feung Chin](#), [Elena Provenzano](#), [Helen A. Bardwell](#), [Wenxin Ma](#), [Wei Cope](#), [Ali Dariush](#), [Sarah-Jane Dawson](#), [Jean E. Abraham](#), [Janet Dunn](#), [Louise Hiller](#), [Jeremy Thomas](#), [David A. Cameron](#), [John M. S. Bartlett](#), [Larry Hayward](#), [Paul D. Pharoah](#), [Florian Markowitz](#), [Oscar M. Rueda](#), [Helena M. Earl](#) & [Carlos Caldas](#) ✉

[Nature](#) **601**, 623–629 (2022) | [Cite this article](#)

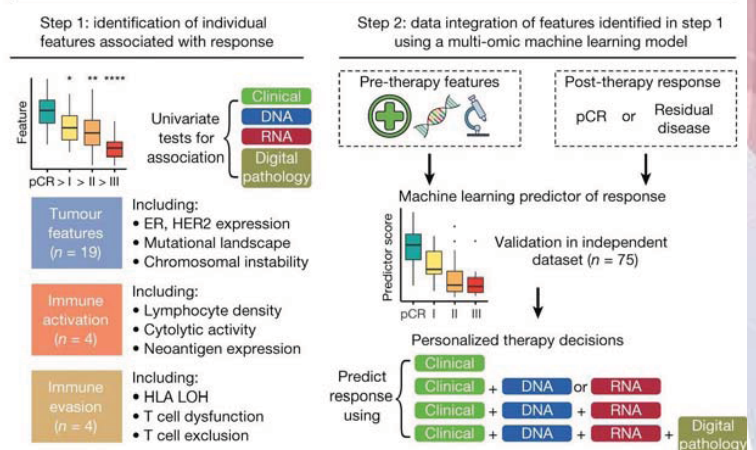
35

Multi-platform profiling of tumor biopsies

Clinical work flow and data acquisition

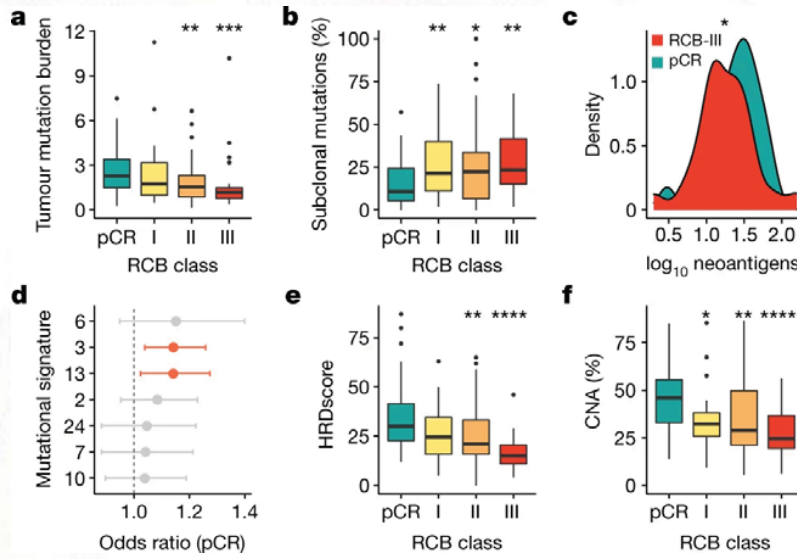


Prediction of response to neoadjuvant therapies



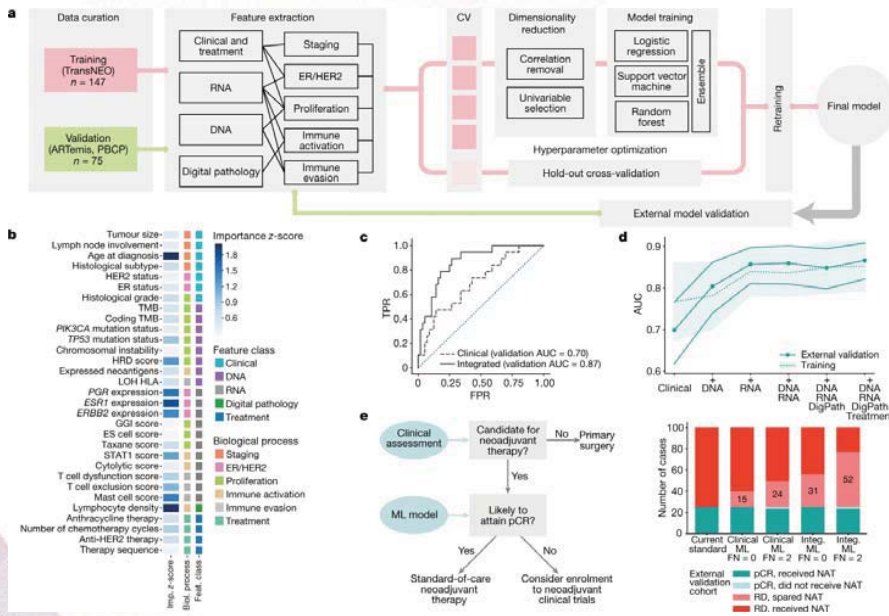
36

Feature selection based on the association with clinical phenotypes



37

Clinical phenotypes are limited predictors



38

Summary

#생물학 네트워크와 멀티오믹스 데이터를 활용하여 다양한 임상적 인공지능 모델이 개발되고 있음

#이는 모든 데이터를 활용하기보다 예측하고자 하는 형질과 연관된 데이터를 잘 추출하는 것이 중요함

#예측 성능이 높은 모델은 만드는 것 만큼이나 모델의 해석도 중요함