KSBi-BIML 2023

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists, Data Scientists,
and Bioinformatians

생물정보학&머신러닝 워크샵(온라인)

Multi-omics driven systematic approaches to understand cancer complexity

김권일 _ 경희대학교

SBi 한국생명정보학회
Korean Society for Bioinformatics

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 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 2023

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists, Data Scientists, and Bioinformatians

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2023에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 9차를 맞이하게 되었습니다. 지난 2년간은 심각한 코로나 대유행으로 인해 아쉽게도 모든 강의가 온라인으로 진행되어 현장 강의에서만 가능한 강의자와 수강생 사이에 다양한 소통의 기회가 없음에 대한 아쉬움이 있었습니다. 다행히도 최근 사회적 거리두기 완화로 현장 강의가 가능해져 올해는 현장 강의를 재개함으로써 온라인과 현장 강의의 장점을 모두 갖춘 프로그램을 구성할 수 있게 되었습니다.

BIML 워크샵은 전통적으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되었습니다. 올해 AI 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 심층학습(Deep learning) 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체 분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 싱글셀오믹스 분석과 메타유전체분석 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다. 이외에 다양한 생명정보학 분야에 대하여 30개 이상의 온라인 강좌가 개설되어 제공되며 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다. 특히 BIML은 각 분야 국내 최고 전문가들의 강의로 구성되어 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것입니다.

이번 BIML-2023을 준비하기까지 너무나 많은 수고를 해주신 BIML-2023 운영위원회의 남진우, 우현구, 백대현, 정성원, 정인경, 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 연사분께 깊은 감사를 드립니다.

2023년 2월

한국생명정보학회장 이 인 석

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 multi-omics 분야의 태동에서부터 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. 연계된 실습에서는 암 multi-omics 공개 데이터를 대상으로 clustering 및 해석을 통합적으로 수행하고, 생물학 네트워크 기반의 multi-omics data 분석을 배울 것 이다. 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 multi-omics data 분석

* 교육생 준비물 및 필요조건:
  네트워크 사용이 가능한 노트북 또는 데스크탑
  Python 사용 가능자

* 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, Kyungheedae-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 Profeesor, Department of Biology, Kyung Hee University

Selected Publications (5 maximum)

1. 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
2. **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).
3. Jang K*, **Kim K***, Cho A, Lee I, Choi JK. Network perturbation by recurrent regulatory variants in cancer. ***PLoS Computional Biology***. 13. e1005449 (2017). *Co-first
4. **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
5. **Kim K**, Yang W, Lee KS, Bang H, Jang K, Kim SC, Yang JO, Park S, Park K, Choi JK. Global transcription network incorporating distal regulator binding reveals selective cooperation of cancer drivers and risk genes. ***Nucleic Acids Research***. 43, 5716-5729 (2015).

KSBi-BIML

Multi-omics driven systematic approaches to understa
nd cancer complexity

Department of Biology, Kyung Hee University

Kwoneel Kim, PhD

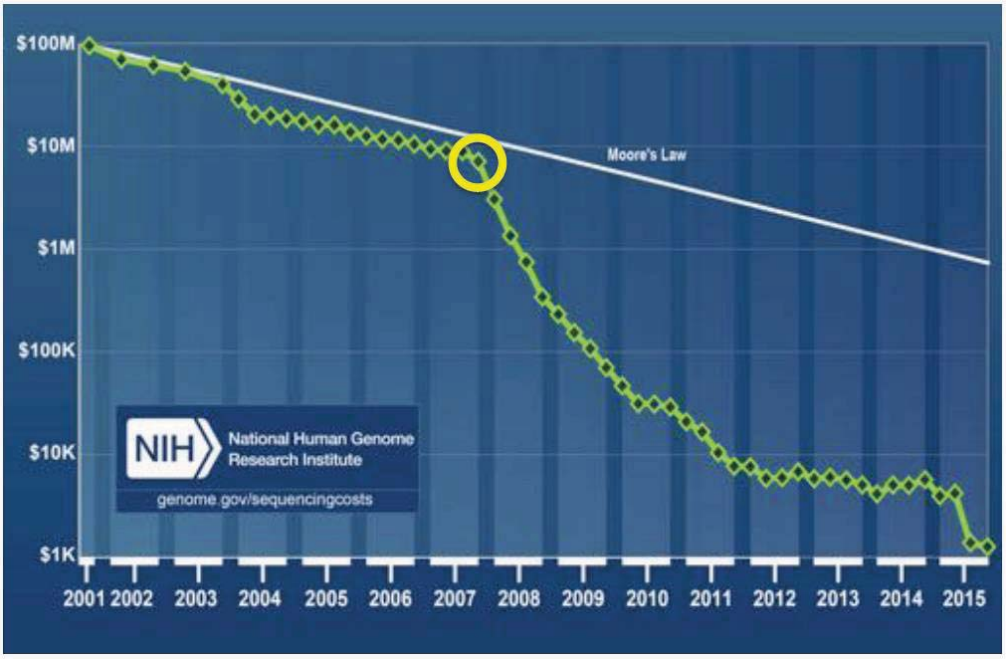
SBi 한국생명정보학회 Korean Society for Bioinformatics

DNA 염기서열분석기술은 인간 게놈 프로젝트 이후 꾸준히 발달해왔다.

#2시간 30 분의 분량으로 준비가 되었기 때문에, 기초 오믹스 배경에 대한 설명은 skip 하거나 빠르게 듣기를 권장 (3~9p 까지)
#본격적인 내용은 슬라이드 제목 "-omics: -ome 을 연구하는 학문 (10p)" 에서부터 시작

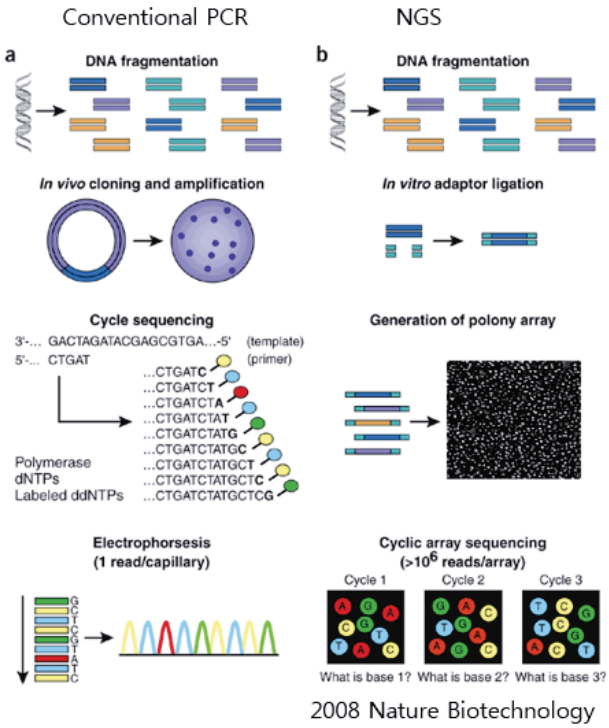
(TIGR/Celera, 1995-2001)

DNA sequencing 기술의 발전



차세대 염기서열 분석법 (Next-generation sequencing, NGS)

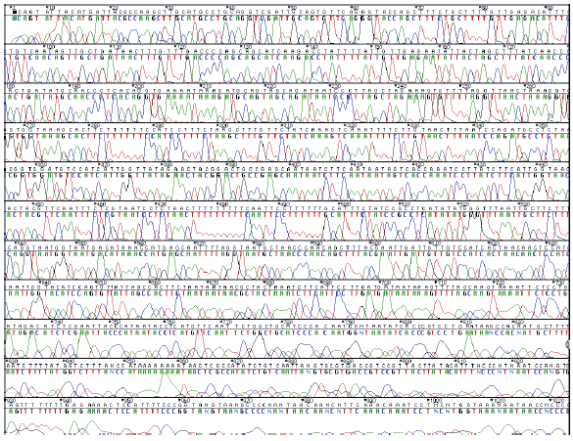
< 기존의 PCR법 VS NGS 법 >

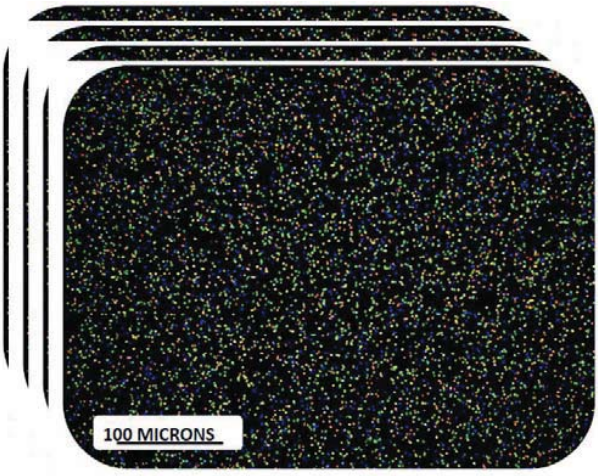


2008 Nature Biotechnology

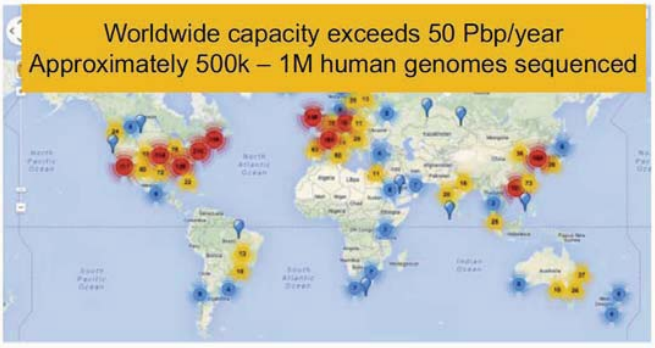
NextSeq 550 Series

차세대 염기서열 분석법: “parallel” and “high throughput”





전세계 염기서열 분석센터 (2014)



Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com

100 GB / Genome
4.7GB / DVD
~20 DVDs / Genome

X

10,000 Genomes

=

1PB Data
200,000 DVDs

787 feet of DVDs
~1/6 of a mile tall

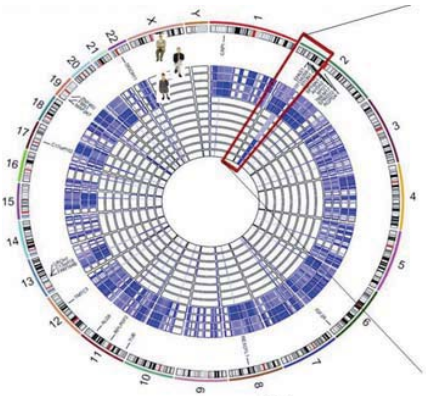
500 2 TB drives
$100k

#엄청 크기가 크다는 것만 이해하면 되겠습니다.

염기서열 분석기술의 발전은 한 명의 하나의 유전자를 분석하는 것에서 **<u>여러 명의 여러 유전자의 서열과 발현</u>** 분석을 가능하게 하였다: 유전체학 (genomics) and 전사체학 (transcriptomics)

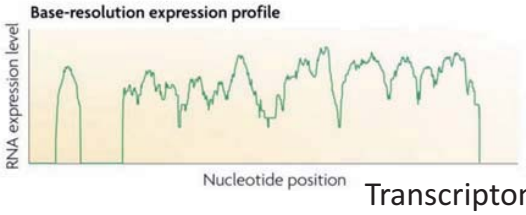
#1-gene sequencing study 가 이미 연구가 너무 많이 되었다고 보면 좋겠습니다

Gene



Genome

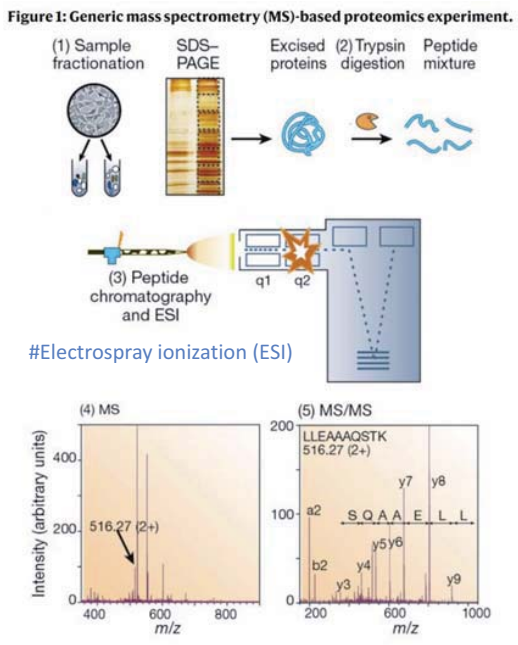
Transcript



Transcriptome

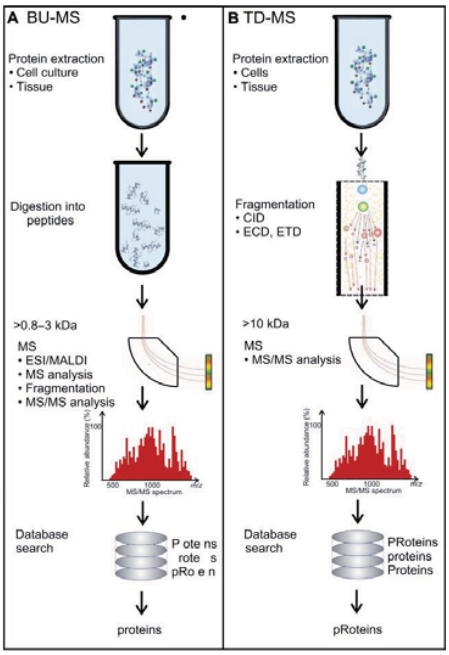
Mass spectrometry (MS) for proteomics

#이 슬라이드에서 말하는 원소, 분자, 물질은 모두 amino acid 를 칭함

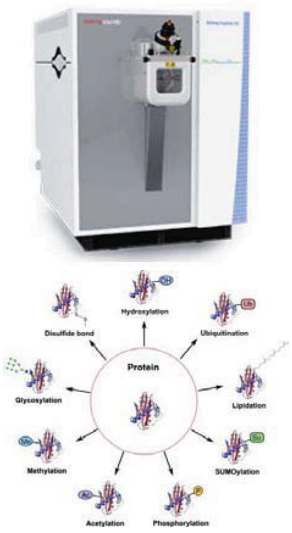


#Electrospray ionization (ESI)

#여러 원소로 구성된 화합물은 각자의 질량에 따라 구별될 수 있음; 구별하기가 쉬운 일은 아님..



#Bottom-up (BU-MS) approach that is used prevalently: 단백질을 짧게 조각낸 peptide를 분석

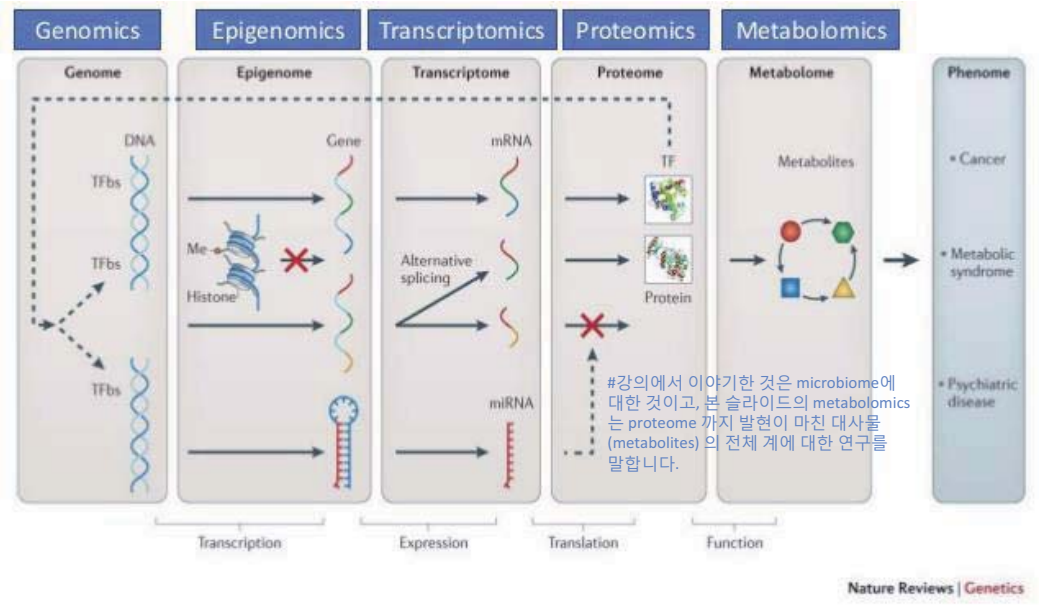


#Top-down (TD-MS) approach that analyzes intact protein <70 kDa: 온전한 형태의 단백질 (intact protein)을 분해없이 직접 분석. **<u>Post-translational modification</u>** (PTM) 을 동정할 수 있음

Nature. 422, 198–207 (2003)
Science Advances. 6, eaax8978 (2020)

-omics: -ome 을 연구하는 학문

"The suffix -ome as used in molecular biology refers to a ~~totality~~ of some sort"
"분자생물학에서 접미사 –ome 은 어떤 종류의 전체를 의미한다"



Large-scale research efforts for omics study

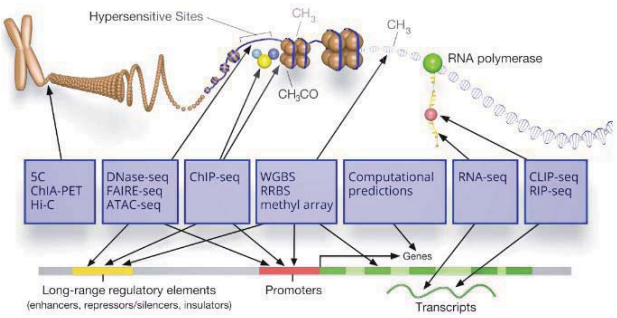
Table 1 | **Data types for integrative omics**

Data type	Large-scale research efforts	Utility and advantages	Major caveats
Genetic variation	Many GWAS consortia, 1000 Genomes, gnomAD and UK Biobank	Unbiased source of genetic basis of disease and direct inference of causality	At least one step removed from the phenotype
Epigenetics	ENCODE and Roadmap Epigenomics Project	Functional impact and typically easy to infer causality	Not applicable for all phenotypes
Gene expression	GTEx and GEUVADIS	Inexpensive assay for an intermediate step towards the phenotype	Not applicable for all phenotypes
Proteomics and metabolomics	CPTAC, EDRN and Common Fund	Likely to be very close to the phenotype	Expensive and difficult to scale (proteomics)
Microbiome	Human Microbiome Project	Likely to be very close to the phenotype and measures a combination of genetic and environmental influences	Combination of genetic and environmental influences makes it difficult to infer the direction of causality

In this table, 'phenotype' refers to an organismal phenotype. CPTAC, Clinical Proteomic Tumour Analysis Consortium; EDRN, Early Detection Research Network; ENCODE, Encyclopedia of DNA Elements; GEUVADIS, Genetic European Variation in Health and Disease; gnomAD, Genome Aggregation Database; GTEx, Genotype–Tissue Expression; GWAS, genome-wide association study.

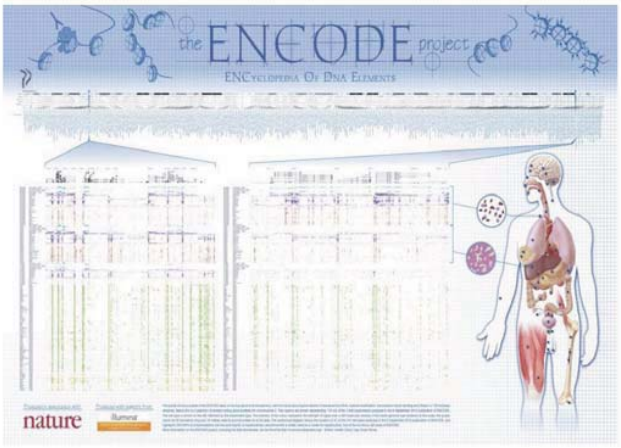
Nature Reviews Genetics. 19, 299–310 (2018)

ENCyclopedia Of DNA Elements (ENCODE) project



#Experiments to analyze the function of DNA elements; coupled with NGS

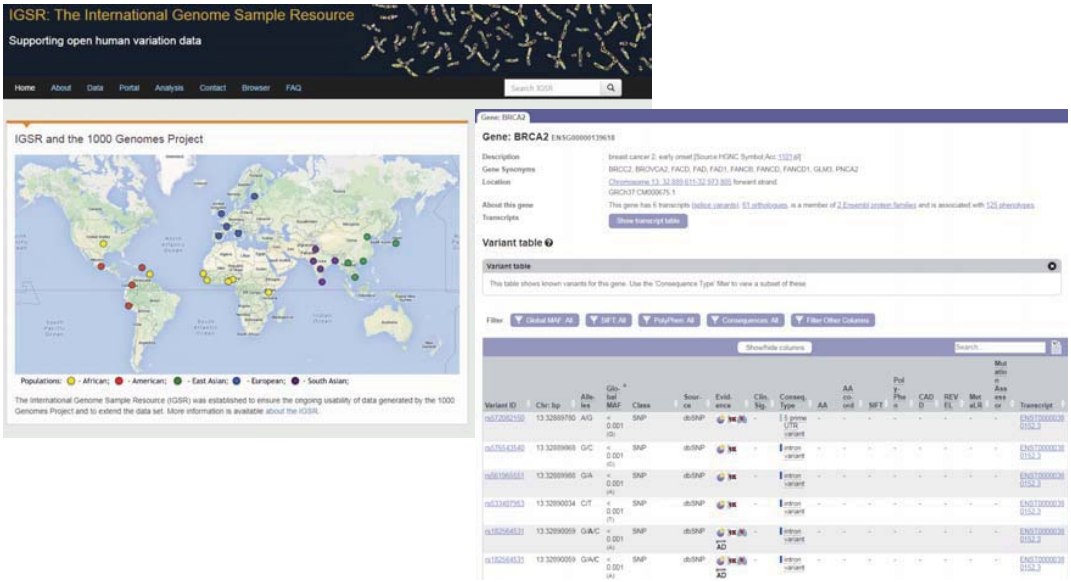
#이후 소개하는 모든 consortium 의 데이터가 오믹스 연구가 가능한 수준의 big data 로 생산이 가능하게 된 것에는 NGS 와 MS technology 의 발전이 있었기 때문



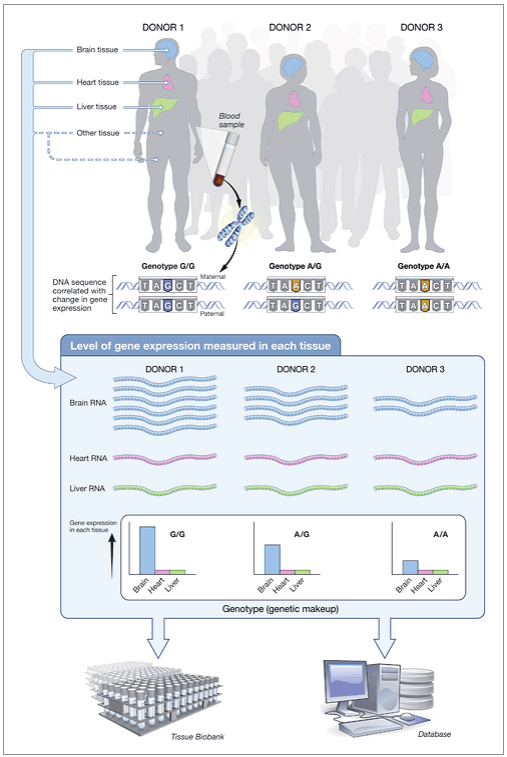
#Epigenome map

~ 9,000 experiments
~ 1,000 Consortium publications
~3,500 community publications using ENCODE data

1000 genome project

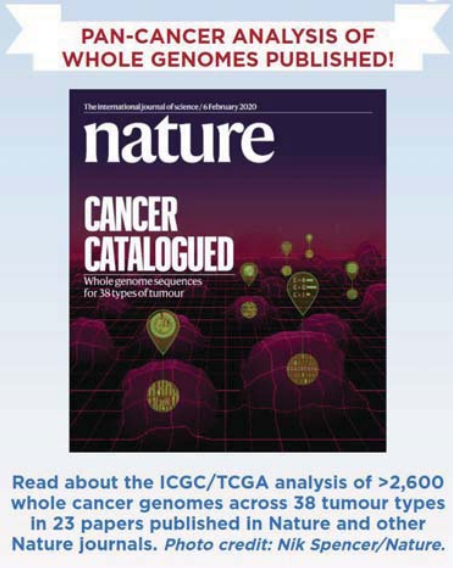


Genotype-tissue expression (GTEX) project



Clinical Proteomic Technology Assessment for Cancer (CPTAC)

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

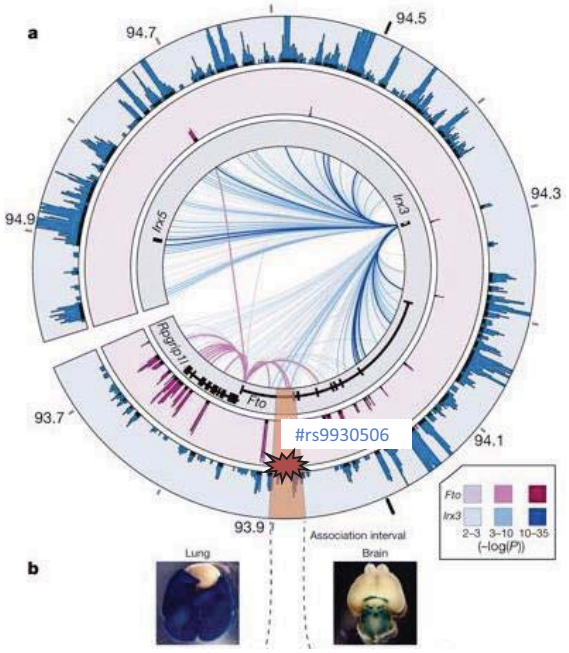


#Sequencing 기술의 발전이 omics 연구를 가능하게 함 -> multiomics

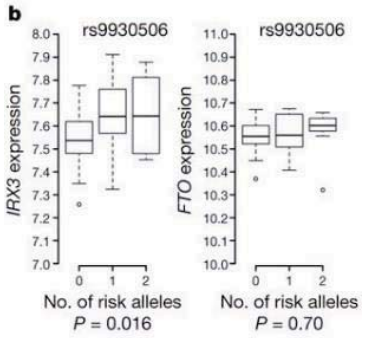
#Global science network 에서 multiomics big data 가 해석가능한 형태로 생산 및 제공되고 있음

Organized by
SBi 한국생명정보학회 Korean Society for Bioinformatics

A case study of integration between epigenomics and genomics



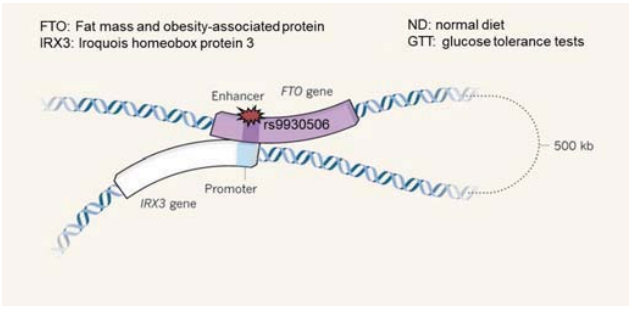
#FTO: Fat mass and obesity-associated protein
#IRX3: Iroquois homeobox protein 3

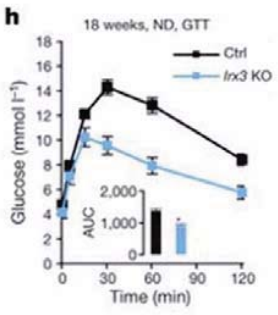


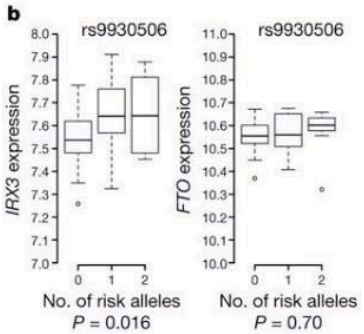
Nature. 507, 371–375 (2014)

Unravel a hidden regulator

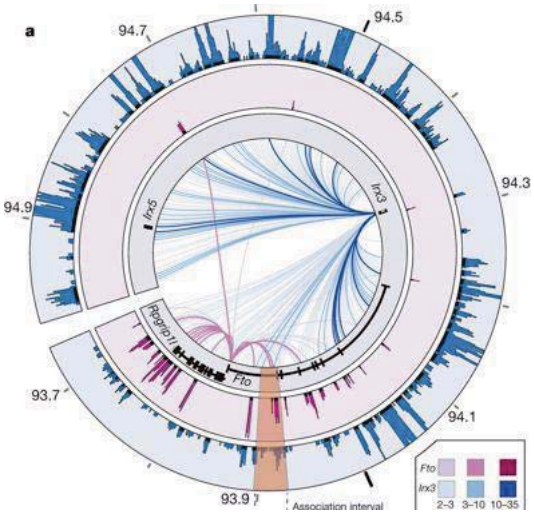
"IRX3 encodes a transcription factor — a type of protein involved in regulating the expression of other genes — and is highly expressed in the brain, consistent with a role in regulating **energy metabolism and eating behaviour**."

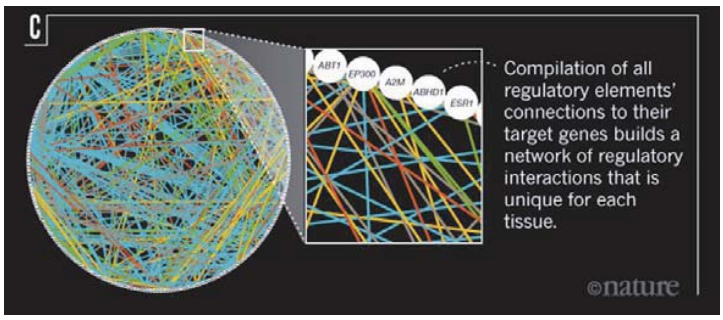






"Systems" biology in transcriptional regulation

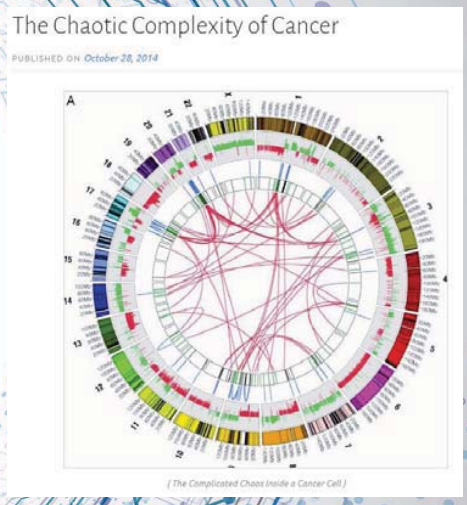




"Dark side of the genome"
Nature. 538, 275–277 (2016)

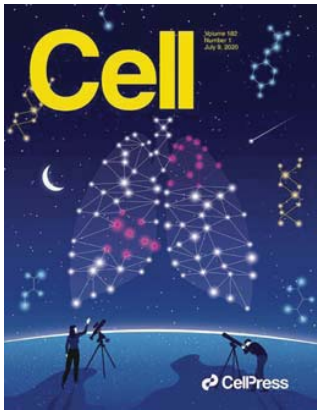
#Genomics 단독으로 해석하지 못하는 생물학적 양상을 epigenomics 와의 통합으로 해석 가능

#Epigenomics 조절은 복잡도가 매우 높아 system 수준의 이해가 필요



Organized by
SBi 한국생명정보학회 Korean Society for Bioinformatics

Milestone paper of multiomics study in lung cancer (2020)

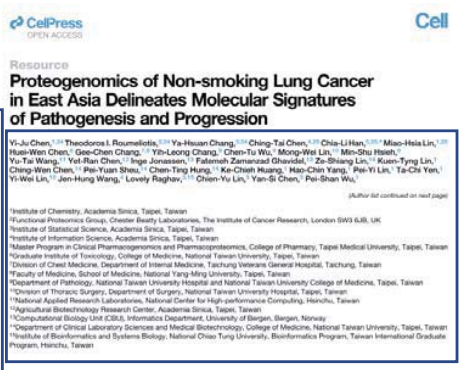



CellPress
OPEN ACCESS

Cell

Resource

Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma


CellPress
OPEN ACCESS

Cell

Resource

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

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

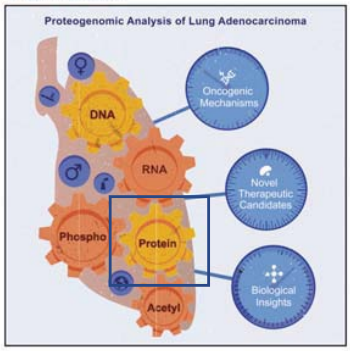
Proteogenomics

Cell

Resource

Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma

Graphical Abstract



Authors

Michael A. Gillette, Shankha Satpathy, Song Cao, ..., D.R. Mani, Steven A. Carr, Clinical Proteomic Tumor Analysis Consortium

Correspondence

gillette@broadinstitute.org (M.A.G.), shankha@broadinstitue.org (S.S.), scarr@broad.mit.edu (S.A.C.)

In Brief

Comprehensive proteogenomic characterization of lung adenocarcinomas and paired normal adjacent tissues from patients of diverse smoking status and country of origin yields insights into cancer taxonomy, oncogenesis, and immune response; offers novel candidate biomarkers and therapeutic targets; and provides a community resource for further discovery.

Highlights

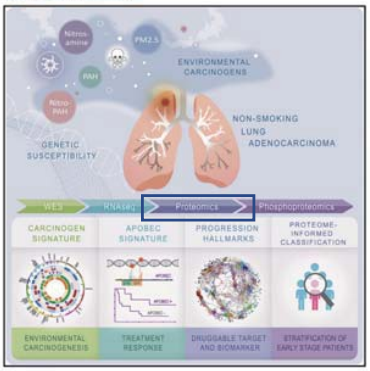
- Comprehensive LUAD proteogenomics exposes multi-omic clusters and immune subtypes
- Phosphoproteomics identifies candidate *ALK*-fusion diagnostic markers and targets
- Candidate drug targets: PTPN11 (*EGFR*), SOS1 (*KRAS*), neutrophil degranulation (*STK11*)
- Phospho and acetyl modifications denote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

Cell

Resource

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

Graphical Abstract



Authors

Yi-Ju Chen, Theodoros I. Roumeliotis, Ya-Hsuan Chang, ..., Hsuan-Yu Chen, Pan-Chyr Yang, Yu-Ju Chen

Correspondence

was@tmu.edu.tw (C.-L.H.), tsung@iis.sinica.edu.tw (T.-Y.S.), chenjs@ntu.edu.tw (J.-S.C.), slyu@ntu.edu.tw (S.-L.Y.), jyoti.choudhary@icr.ac.uk (J.S.C.), hychen@stat.sinica.edu.tw (H.-Y.C.), pcyang@ntu.edu.tw (P.-C.Y.), yujuchen@gate.sinica.edu.tw (Y.-J.C.)

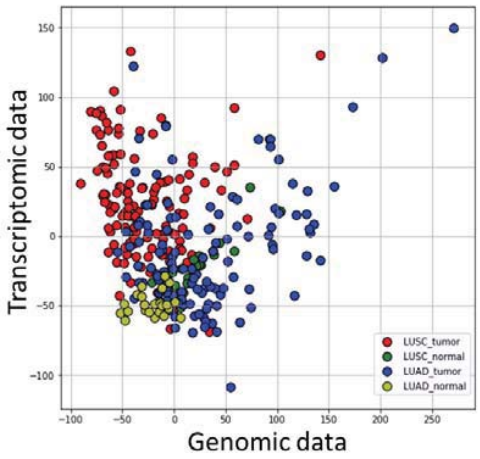
In Brief

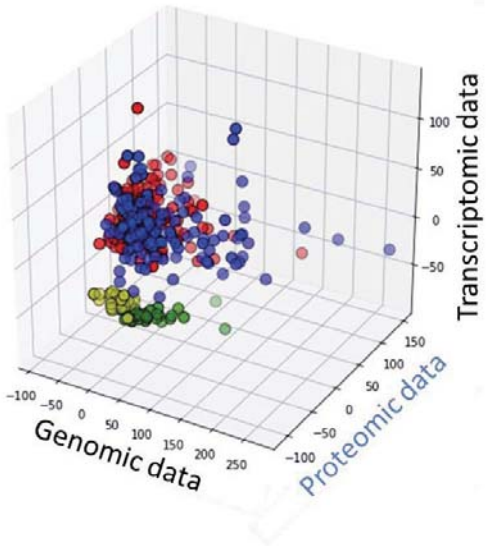
Deep proteogenomic landscape of early stage lung adenocarcinoma in a cohort of mostly non-smokers reveals unique drivers and biomarkers, as well as gender-associated mutagenesis.

Highlights

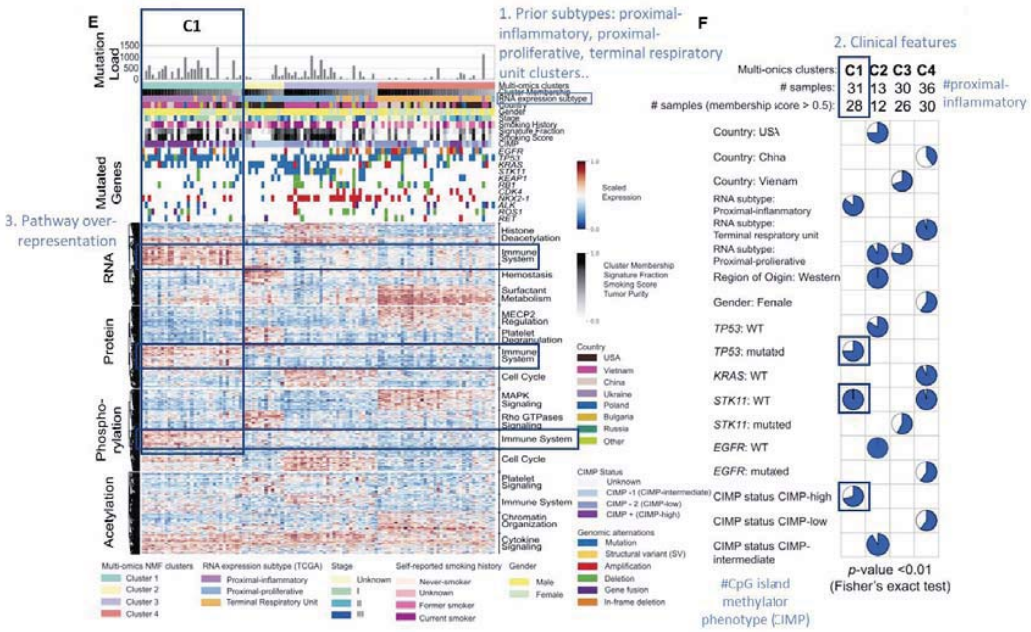
- First deep proteogenomic landscape of non-smoking lung adenocarcinoma in East Asia
- Identified age, sex-related endogenous, and environmental carcinogen mutagenic processes
- Proteome-informed classification distinguished clinical features within early stages
- Protein networks identified tumorigenesis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

Multi-omics clustering





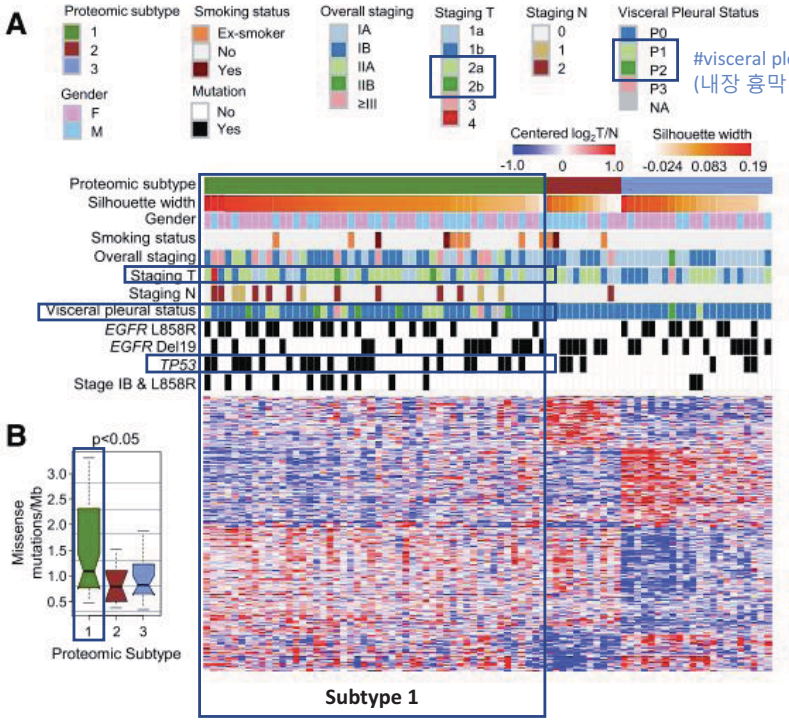
Non-negative matrix factorization (NMF)-based unsupervised "multi-omics clustering": samples of the clusters were significantly associated with distinctive clinical and molecular features



#3. 관련하여 다시한번 정리: multiomics clustering 을 통해 발굴한 sample cluster 가 독특한 clinical and molecular 특징을 나타내면서, 이미 알려진 유전자 구성 (pathway) 특성을 재현성 있게 나타냄

Cell. 182, 200–225 (2020)

Alignment of the proteomic subtypes with clinical features revealed a strong separation by tumor staging, as well as by driver mutations

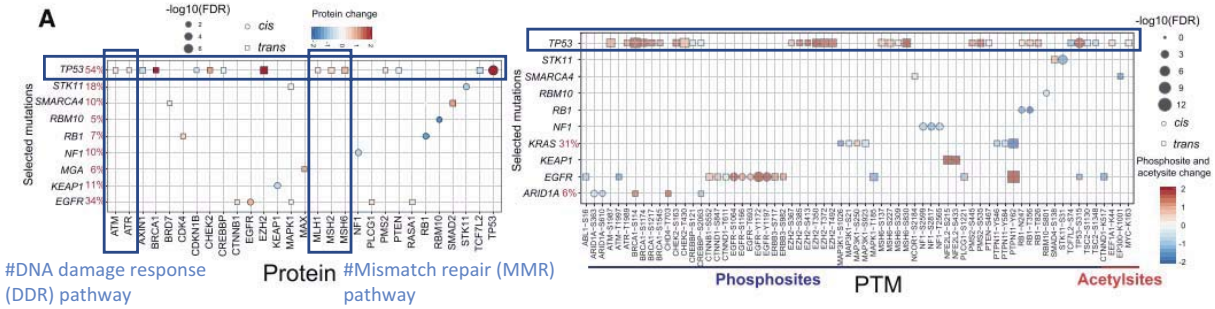


#visceral pleural invasion
(내장 흉막 침범)

Cell. 182, 226–244 (2020)

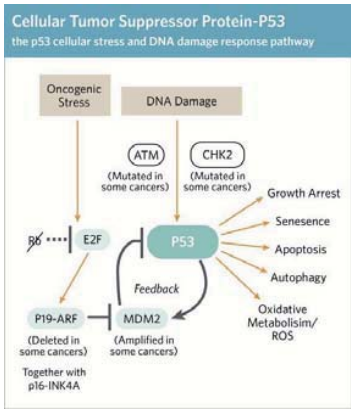
Connecting driver mutations to proteome, phosphoproteome, and pathways

#SKT11 은 중요하지 않다고 하였지만 해당 논문에서는 의미 있게 다룸. 본 워크샵에서는 분량상 생략



#DNA damage response (DDR) pathway

#Mismatch repair (MMR) pathway

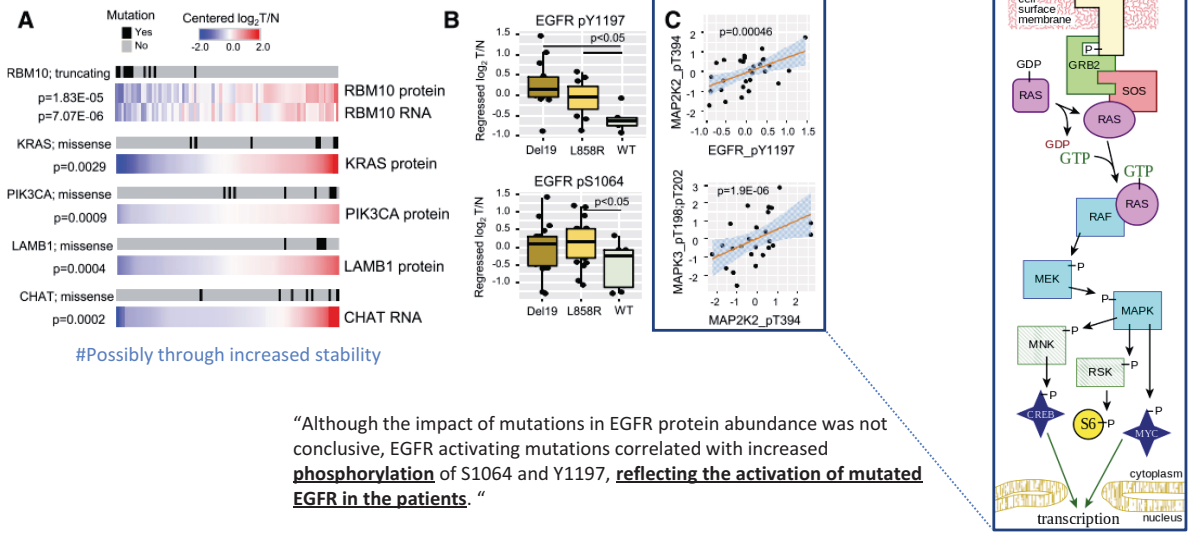


Cell. 182, 200–225 (2020)

Direct and indirect consequences of genomic aberrations at the transcriptome and proteome levels

#EGFR activating mutations

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

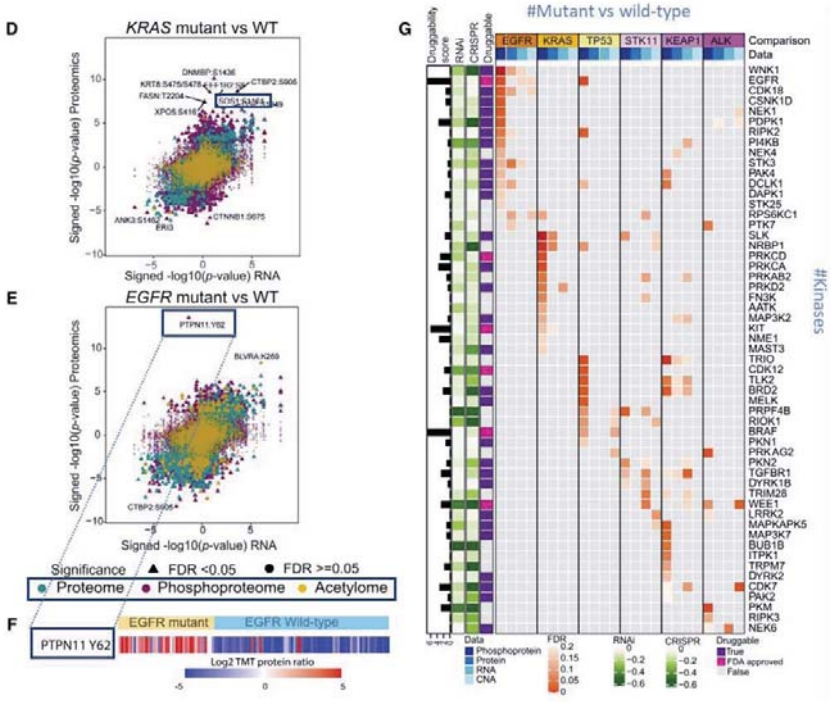


#Possibly through increased stability

"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)

Identification of therapeutic strategies from proteogenomics analyses



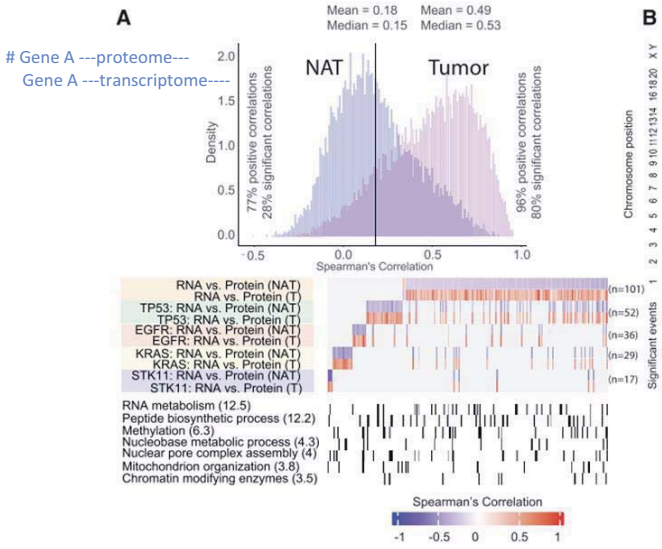
"SOS1 is a guanine exchange factor (GEF) that activates KRAS (Vigil et al., 2010), and inhibition of **SOS1 and KRAS is an emerging therapeutic strategy for KRAS mutant cancers** (Hillig et al., 2019; O'Bryan, 2019)."

"**EGFR mutant** tumors showed highly significant and remarkably consistent **tyrosine phosphorylation of PTPN11/Shp2 at Y62.**

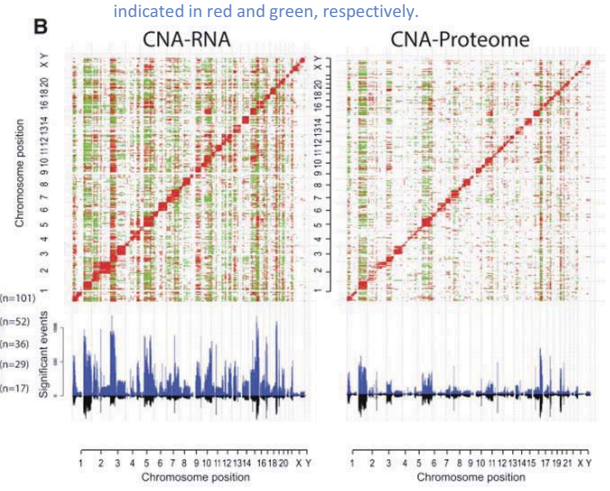
Although prior studies have associated PTPN11/Shp2 phosphorylation with important biological consequences in NSCLC cell lines and xenograft models, this is, to our knowledge, **the first report of such phosphorylation in a large set of primary treatment-naive** LUADs."

Cell. 182, 200–225 (2020)

Gene-wise CNA-mRNA-protein correlations displayed striking differences



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

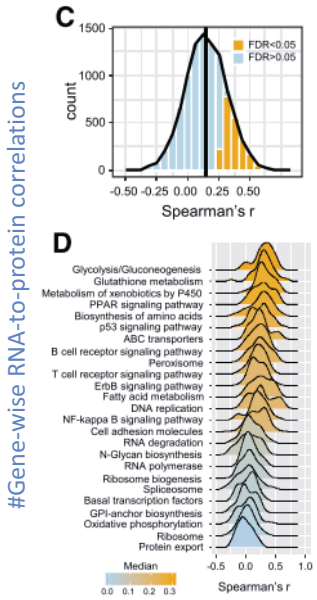


"Although sample-wise mRNA-protein correlations were fairly consistent between tumors and NATs, **gene-wise correlations displayed striking differences**"

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

Cell. 182, 200–225 (2020)

Gene-wise mRNA-protein correlations displayed striking differences

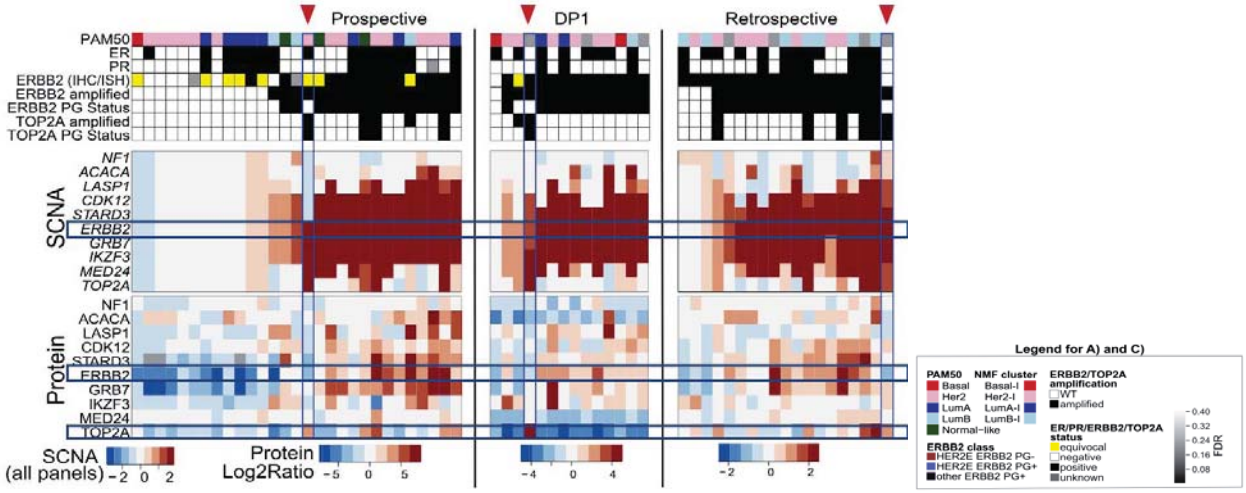


"**Only 22% proteins displayed significant positive correlations with the cognate RNA**. Enrichment analysis showed a pathway-dependent RNA-to-protein correlation, with **basic cellular functions poorly corresponding to RNA**. Taken together, these analyses indicate transcriptionally modulated upregulation of DNA replication, glycolysis, glutathione metabolism, and immune-related pathways, while upregulation of DNA repair, protein processing and transport pathways, and downregulation of cell-adhesion-related pathways were more apparent at protein level (전사체와 단백체가 pathway 별 변화 양상이 다르다)."

Cell. 182, 226–244 (2020)

"Pseudo-ERBB2+'' status in resistance for anti-ERBB2 antibody therapy

#해당 연구를 lung cancer model 로 언급하였는데 breast cancer 에서 진행된 것임을 정정합니다.

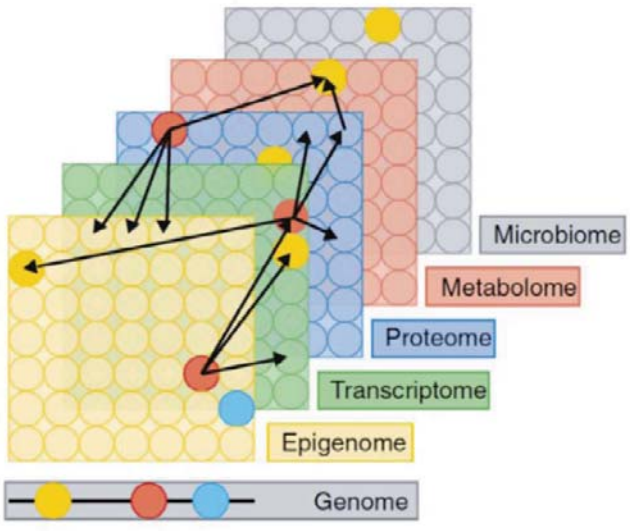


''Because these **pseudo-ERBB2+ samples are examples where anti-ERBB2 treatment may not have been effective** because of lack of drug target expression, proteogenomics approaches were used to assess ERBB2 driver status in the current dataset and our earlier cohort.

"..amplication and overexpression of TOP2A, suggesting an **alternative chromosome 17 amplicon driver in some cases (Harris et al., 2009).**"

Cell. 183, 1436–1456 (2020)

Changes of a layer of one omics do not always transfer into other layer



Genome Biology. 18, 83 (2017)

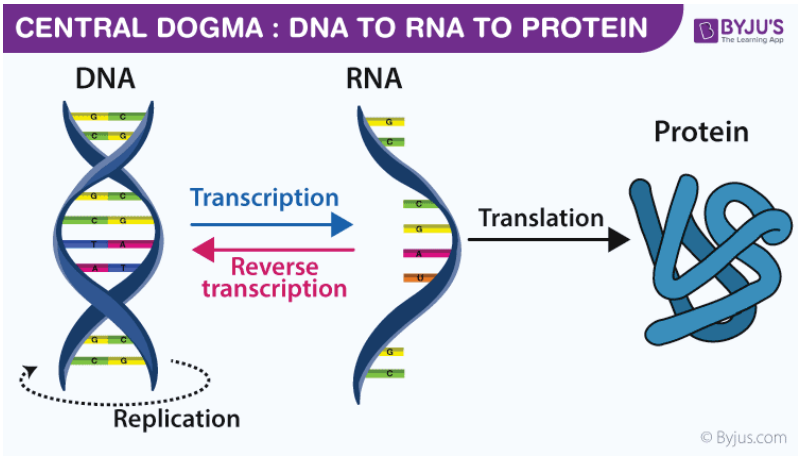
#Genomics 단독으로 해석하지 못하는 생물학적 양상을 proteomics 와의 통합으로 해석 가능

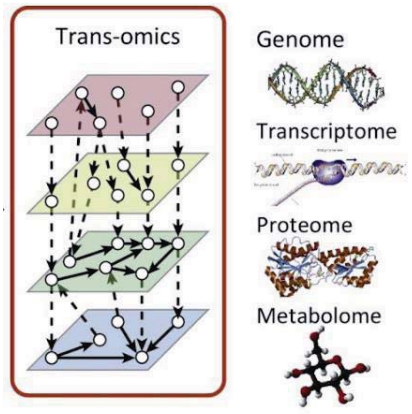
#각 omics 에서의 변화가 그대로 전달되는 것이 아니므로 system 수준의 이해가 필요

Organized by

SBi 한국생명정보학회 Korean Society for Bioinformatics

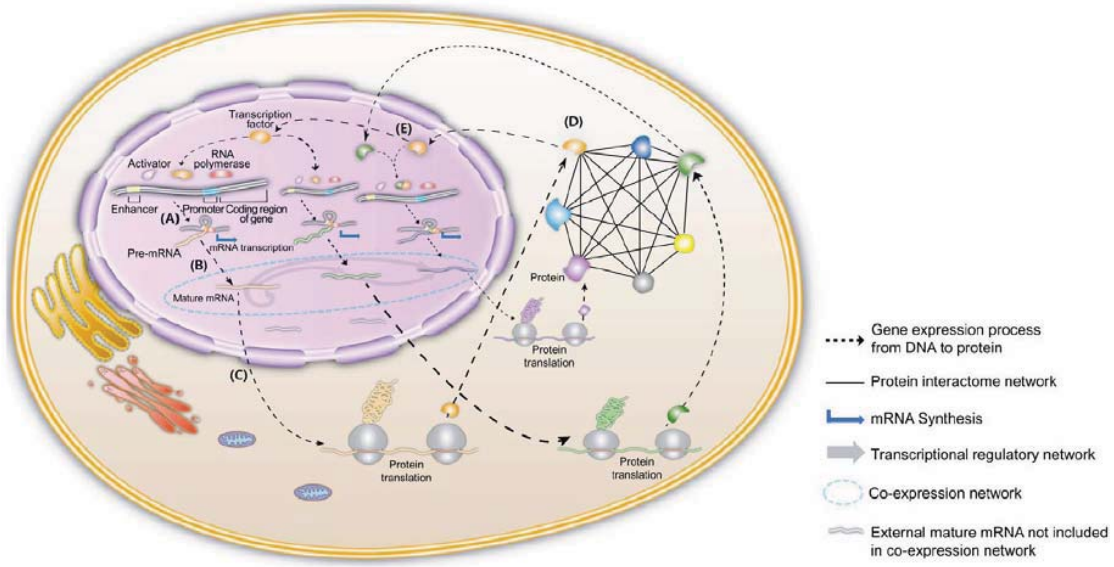
Complexity of controlling quantity in central dogma





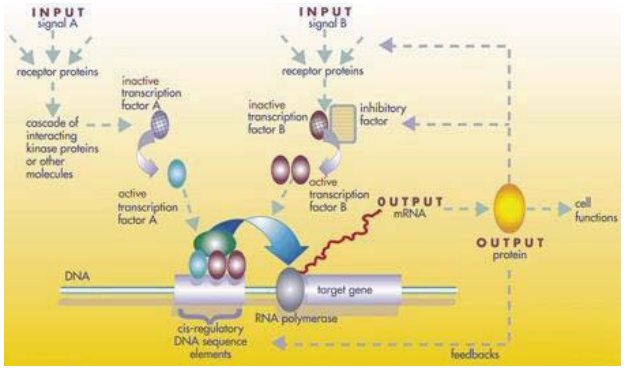
Trends in Biotechnology. 34, 276-290 (2016)

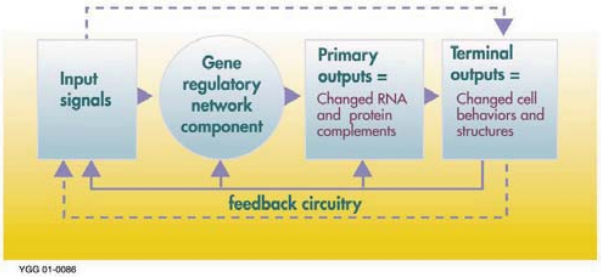
Multi-omics data for systems biology in cell systems

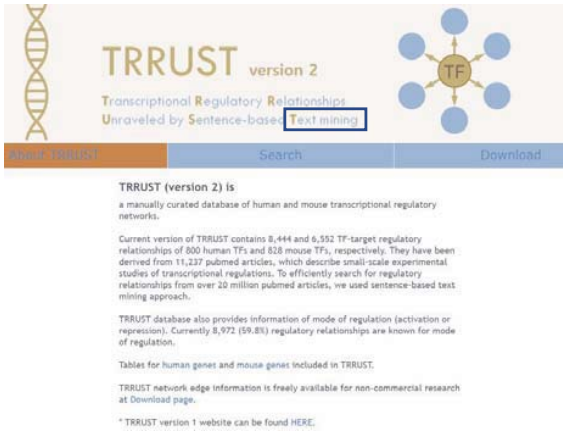


Anim Cells Syst. 30, 1-7 (2020)

Gene regulatory network

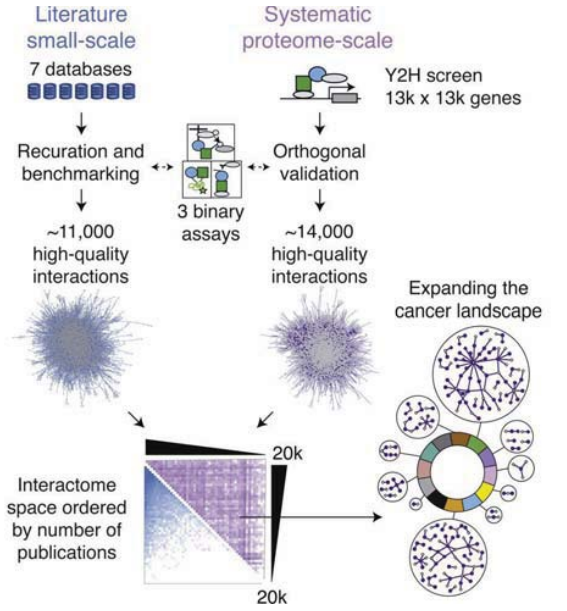






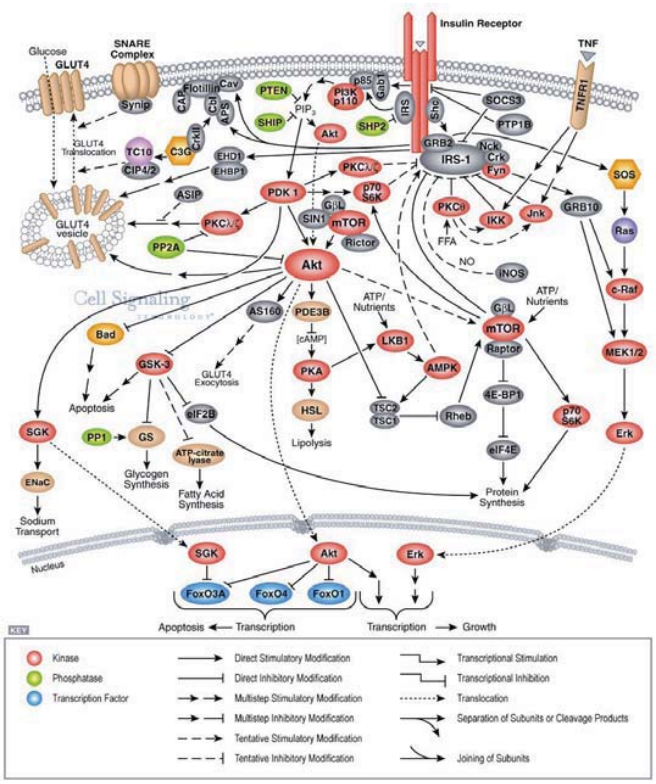
"..the potential regulations inferred based on the **transcription factor binding sites** (TFBSs)."

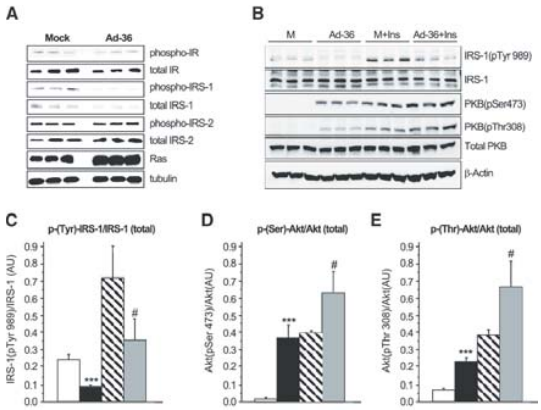
Interactome network



Cell. 159, 1212-1226 (2014)
Nature. 580, 402–408 (2020)

Signaling network





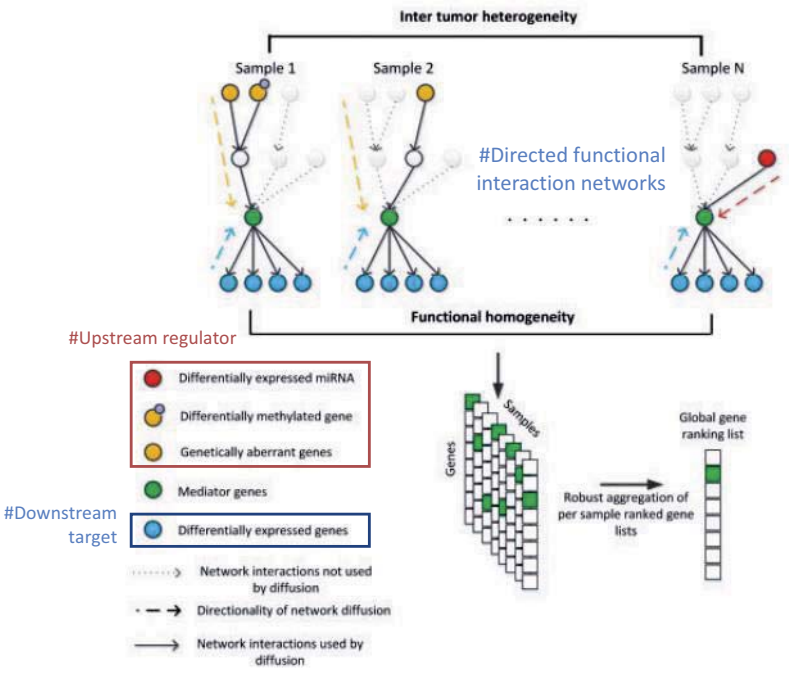
#Multiomics system 에는 다양한 층위 및 network 가 존재

#**이것을 multiomics 해석에 어떻게 적용할지는 case-dependent**

Organized by

SBi 한국생명정보학회 Korean Society for Bioinformatics

Network-based integration of multi-omics data for prioritizing cancer genes



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

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

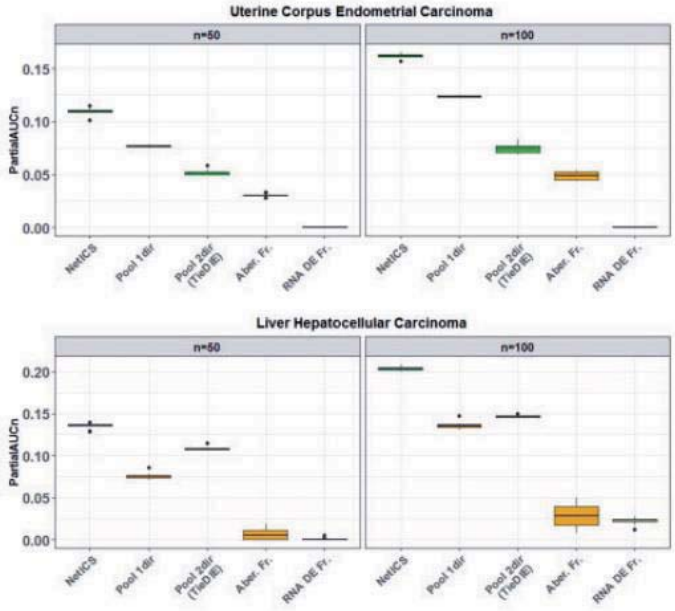
"The vector E determines the **mediator effect** for each gene.

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.**"

#원론, 청사진 등으로 언급한 대상은 생명현상의 정보를 담고 있는 DNA

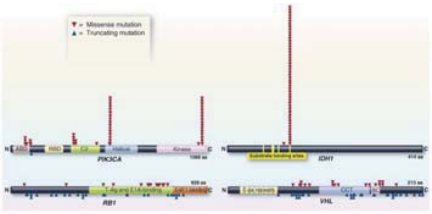
Bioinformatics. 34, 2441–2448 (2018)

Prediction of known cancer genes



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

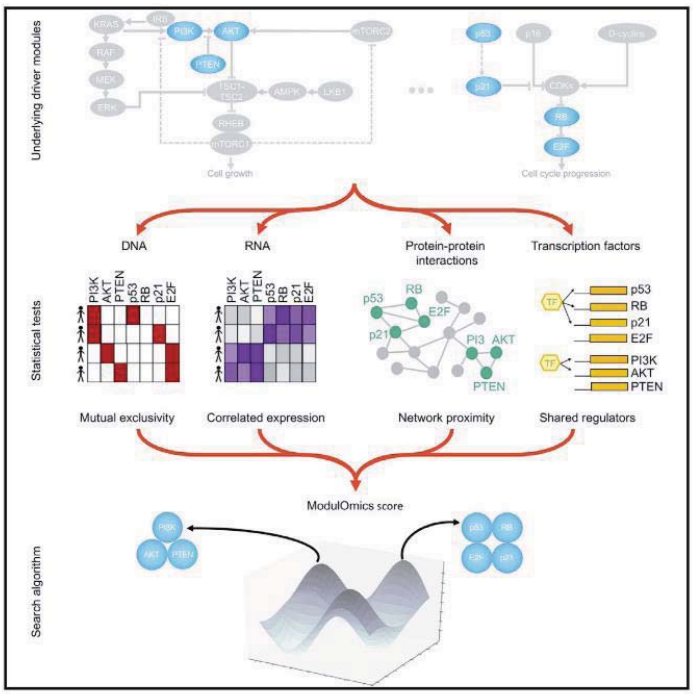
#RNA DE Fr.: ranking by **frequency of differentially expressed genes** across all samples



Science. 339, 1546–1558 (2013)

Simultaneous integration of multi-omics data improves the identification of cancer driver modules

Graphical Abstract



Cell Systems. 8, 456–466 (2019)

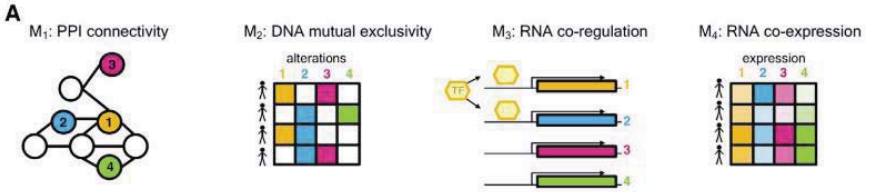
ModulOmics

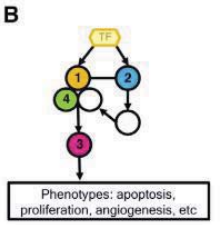
#M1 = frequency of connectivity (proteome)

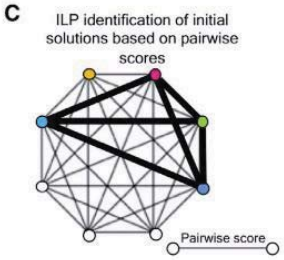
#M2 = DNA mutual exclusivity (genome)

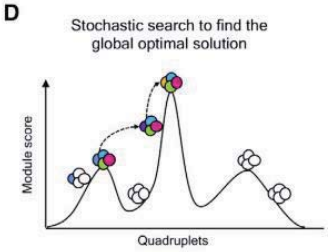
#M3 = co-regulation for common TF (transcriptome)

#M4 = co-expression correlation (transcriptome)









$$S_G = \frac{1}{m}\sum_{k=1}^{m} P(G|M_k)$$

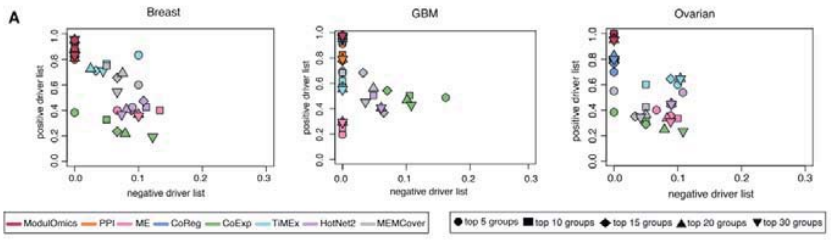
#Highest sum of **pairwise** ModulOmics scores , computed as the average of the four scores corresponding to models M1–M4

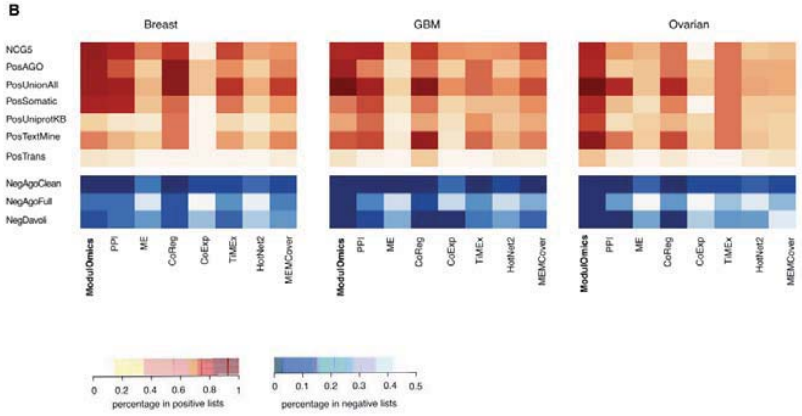
The space of initial solutions is clustered, and genes are exchanged between clusters to identify modules with high **global scores**.

"Given a set G={G1,.,Gn} of genes and a collection M={M1,.Mm} of models for different data types, we are interested in computing S_G reflecting **how likely are the genes in G to be functionally connected**.

S_G is the genes in G are, under different models: computed as the mean of m probabilistic scores $P(G|M_k)$. Each of these m scores represents how strongly functionally connected the genes in G are, under different models."

The driver module inferred by ModulOmics are enriched with cancer driver genes: no single omics data type dominated the ModulOmics score

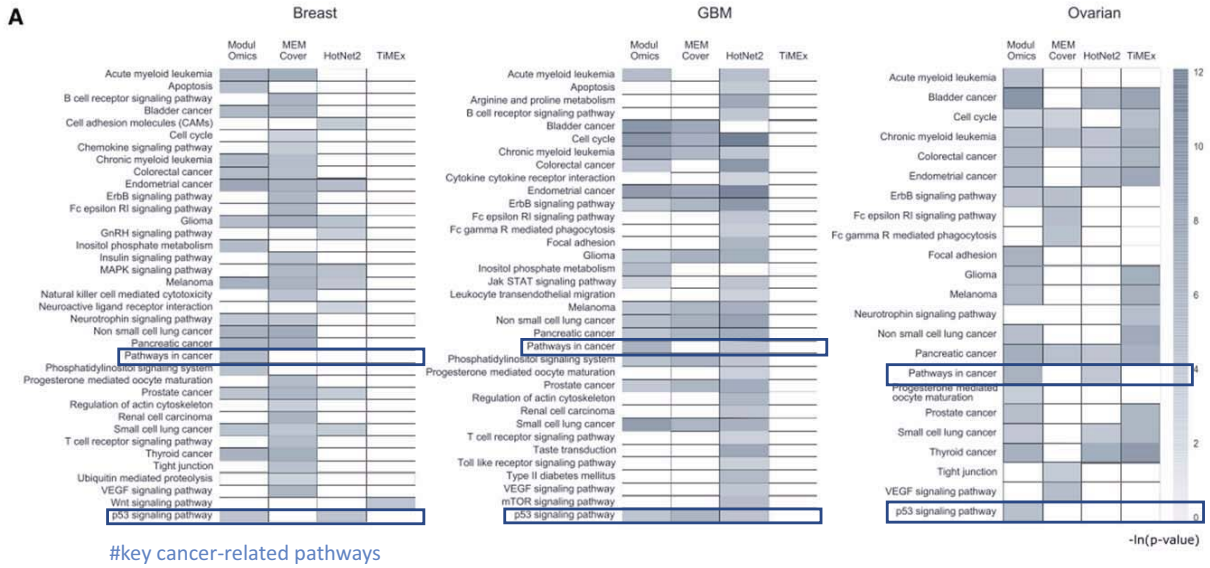




#Positive control lists for cancer driver genes

#Negative control lists for cancer driver genes

The driver modules inferred by ModulOmics are enriched with cancer driver pathways



#key cancer-related pathways

Driver modules in breast cancer subtypes recapitulate known mechanisms

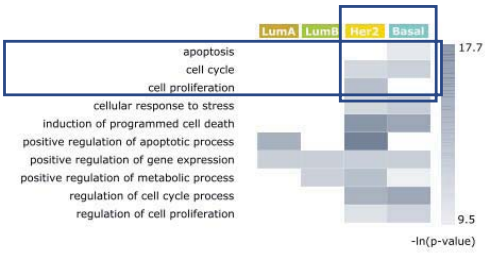
#BRCA1, 2: basal-specific

#CDH1: luminal A and B-specific

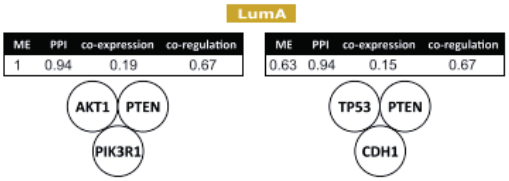
#EGFR: her2-specific

#MAP3K1: liminal B-specific

#Explanation for tumor aggressiveness of Her2 and basal tumors by enriched in tumor progression pathways

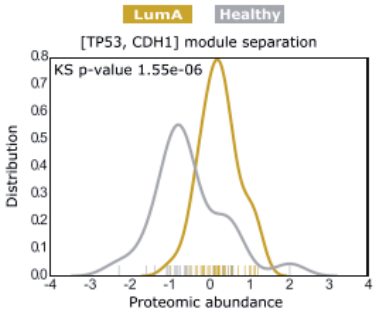


Driver modules in breast cancer subtypes suggest unexplored functionalities



"**The canonical module PTEN, AKT1, PIK3R1** recapitulates the known mutual exclusivity pattern of mutations within the **PI3K pathway**."

"In contrast, the module suggesting the **noncanonical role (PTEN, CDH1, TP53)** supports the hypothesis that PTEN regulates cell proliferation by increasing the binding of CDH1 to APC\C, a complex known for its tumor-suppressive function, and by increasing TP53 acetylation following DNA damage."



#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)

Organized by

SBI 한국생명정보학회 Korean Society for Bioinformatics

#Sequencing 기술의 발전이 omics 연구를 가능하게 함 -> multiomics

#Global science network 에서 multiomics big data 가 해석가능한 형태로 생산 및 제공되고 있음

#Omics 단독으로 해석하지 못하는 생물학적 양상을 multiomics 로 해석 가능

#Multiomics 간 조절은 매우 복잡도가 매우 높고, 각 omics 에서의 변화가 그대로 전달되는 것이 아니므로 system 수준의 이해가 필요

#생물학적 system 에는 다양한 층위가 존재하므로, 이것을 multiomics 해석에 어떻게 적용할지는 case-dependent

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

Organized by

SBi 한국생명정보학회 Korean Society for Bioinformatics

Thank you

실습: NMF clustering of multiomics data and network analysis

Department of Biology, Kyung Hee University

Kwoneel Kim, PhD

Kwoneelkim@gmail.com

SBi 한국생명정보학회 Korean Society for Bioinformatics

KSBi-BIML

NMF clustering of LUAD multi-omics data and network analysis

Kyung Hee University

Jiyeon Kim

SBi 한국생명정보학회 Korean Society for Bioinformatics

실습 목표

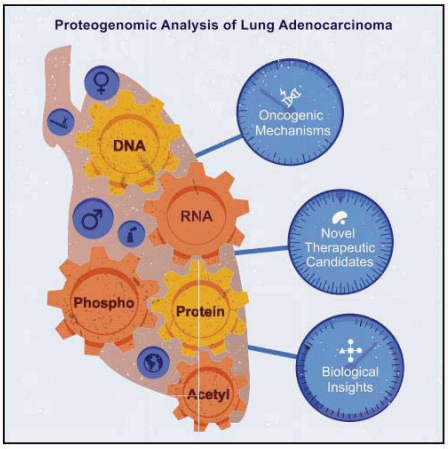
- Multi-omics data를 NMF method를 이용하여 clustering
- Analysis of network-based mechanism

Cell

Resource

Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma

Graphical Abstract



Authors

Michael A. Gillette, Shankha Satpathy, Song Cao, ..., D.R. Mani, Steven A. Carr, Clinical Proteomic Tumor Analysis Consortium

Correspondence

gillette@broadinstitute.org (M.A.G.), shankha@broadinstitue.org (S.S.), scarr@broad.mit.edu (S.A.C.)

In Brief

Comprehensive proteogenomic characterization of lung adenocarcinomas and paired normal adjacent tissues from patients of diverse smoking status and country of origin yields insights into cancer taxonomy, oncogenesis, and immune response; offers novel candidate biomarkers and therapeutic targets; and provides a community resource for further discovery.

SBi 한국생명정보학회 Korean Society for Bioinformatics

Data preparing

Data Download

#Data는 논문에서 제공하는 supplementary data 사용
#제공된 multiomics 데이터를 사용하여 NMF clustering 수행

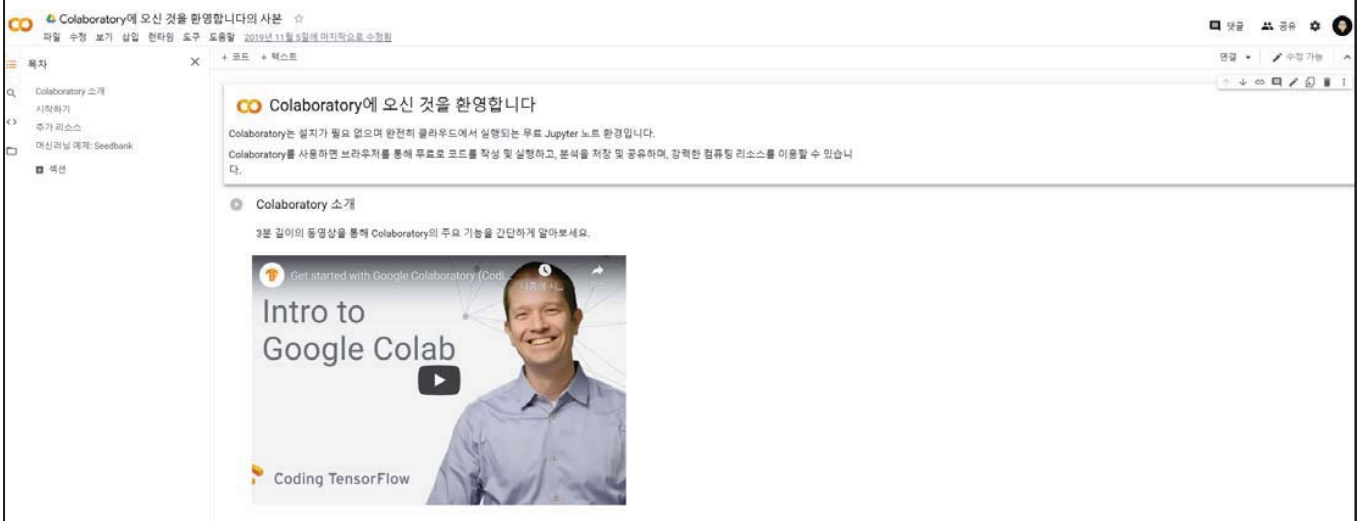
- RNA expression
- Protein expression
- Copy number variation

Data pre-processing

- 전체 샘플의 30% 이상 NA값을 가진 feature 제거
- Feature name filtering
- Sample filtering
- Feature 별 std < 0.5인 feature 제거

RNA	9259
Protein	7107
CNV	19

Google Colaboratory



#Colab 기본 정보 참고 자료

https://drive.google.com/drive/folders/1jxOCihLdevxyBiLt9lKidrb_x74PWSyl?usp=sharing

NMF clustering

- Nimfa library 사용

#Clustering score 에 기반하여 최적의 cluster number 를 찾으려 하는데, score 간 scale 이 상당히 차이가 나기 때문에 다양한 scoring 기법을 시도함

Nimfa

Star 437

Navigation

Models (**models**)
Methods (**methods**)
Utils (**utils**)
Examples (**examples**)
Datasets (**datasets**)

Quick search

Go

Welcome to Nimfa

Nimfa is a Python library for nonnegative matrix factorization. It includes implementations of several factorization methods, initialization approaches, and quality scoring. Both dense and sparse matrix representation are supported.

Nimfa is distributed under the BSD license.

The sample script using Nimfa on medulloblastoma gene expression data is given below. It uses alternating least squares nonnegative matrix factorization with projected gradient method for subproblems [Lin2007] and Random Vcol [Albright2006] initialization algorithm. The returned object is fitted factorization model through which user can access matrix factors and estimate quality measures.

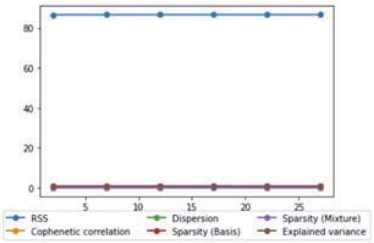
#Rank = sample cluster

```
import nimfa

V = nimfa.examples.medulloblastoma.read(normalize=True)

lsnmf = nimfa.Lsnmf(V, seed='random_vcol', rank=50, max_iter=100)
lsnmf_fit = lsnmf()

print('Rss: %5.4f' % lsnmf_fit.fit.rss())
print('Evar: %5.4f' % lsnmf_fit.fit.evar())
print('K-L divergence: %5.4f' % lsnmf_fit.distance(metric='kl'))
print('Sparseness, W: %5.4f, H: %5.4f' % lsnmf_fit.fit.sparseness())
```

Running this script produces the following output, where slight differences in reported scores across different runs can be attributed to randomness of the Random Vcol initialization method:

```
Rss: 0.2668
Evar: 0.9997
K-L divergence: 38.8744
Sparseness, W: 0.7297, H: 0.8796
```

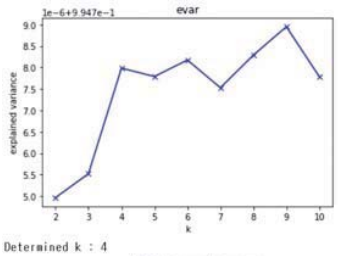
```
range(2, 30, 5)
```



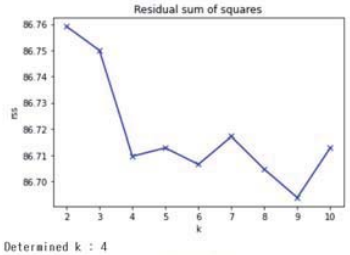
SBi 한국생명정보학회 Korean Society for Bioinformatics

NMF clustering

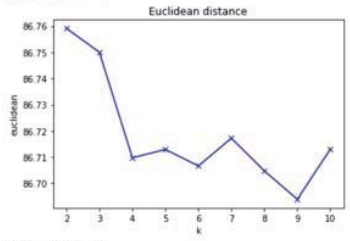
#k = the number of sample clusters



Determined k : 4

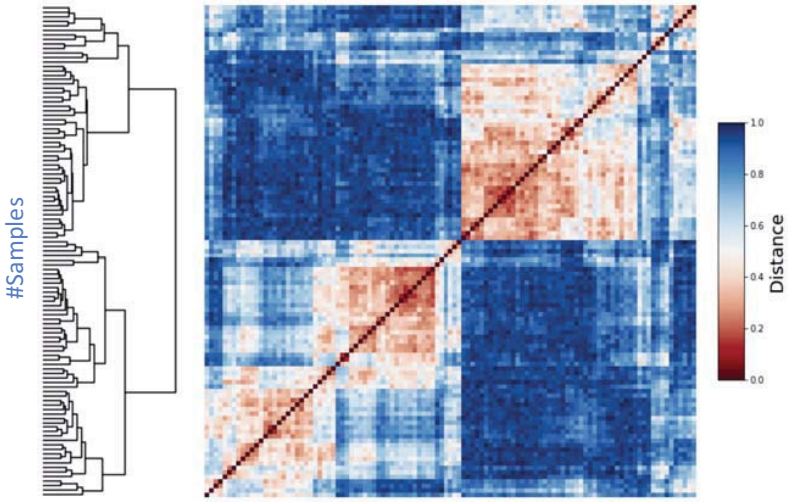


Determined k : 4



Determined k : 4

#Evaluated variance(evar) 가 증가하고 residual sum of squares와 euclidean distance 가 감소할수록 clustering 이 잘 되었다고 판단할 수 있음
#K 초기값의 설정에 따라 scoring이 달라질 수 있음



#유전자별 각 clustering에 대한 기여도를 알고리즘에서 산출해 줌

SBi 한국생명정보학회 Korean Society for Bioinformatics

Protein-protein interaction

https://string-db.org/cgi/download?sessionId=bnEQRR5VwVKB

#본 실습에서는 가정을 매우 단순히 하여 protein interactome 만을 사용하여 clustering 결과에 대한 system 수준의 접근을 시도

	node1	node2	combined_score
0	ARF5	SPTBN2	909
1	ARF5	KIF13B	910
2	ARF5	KIF21A	910
3	ARF5	TMED7	906
4	ARF5	ARFGAP1	971
...	...	...	...
648299	OR6Q1	REEP1	900
648300	OR6Q1	REEP4	900
648301	OR6Q1	GNB1	900
648302	OR6Q1	RTP3	900
648303	OR6Q1	REEP2	900

648304 rows × 3 columns

#combined_score: 두 유전자가 interacting 할 확율 값. 확율x103 값으로 반환됨.

```
Features = [features_0, features_1, features_2, features_3]

def change_ID(features):
  """change proteinID to geneSymbol"""
  result_name = []
  for name in features:
    if (name.split('_')[-1] == 'RNA') or (name.split('_')[-1] == 'CNV'):
      result_name.append(name.split('_')[0])
    elif name.split('_')[-1] == 'Protein':
      result_name.append(proid2gene[name[:-8]])
  return result_name

def get_ppis(features, ppis):
  feature_rst = change_ID(features)
  feature_rst = list(set(feature_rst))
  print(len(features),'->',len(feature_rst))

  ppi_tmp = ppis[(ppis['node1'].isin(feature_rst)) & (ppis['node2'].isin(feature_rst))]
  ppi_list = list(set(ppi_tmp['node1']))
  print(len(ppi_list))
  return ppi_list

for c,features in enumerate(Features):
  setattr(mod,f'ppis_{c}',get_ppis(features, ppis))
```

#feature = gene

```
1000 -> 903
324
1000 -> 889
537
1000 -> 873
456
1000 -> 894
265
```

#각 omics data 간의 gene symbol – gene id mapping 에 중복이 있을 수 있음

Pathway enrichment analysis

https://maayanlab.cloud/Enrichr/#stats

#NMF cluster 에서 protein interaction 을 하는 그룹들이 어떤 biological pathway 에 enrichment 되어 있는지 확인

- Gseapy library 사용

```
len(gp.get_library_name())
```

```
171
```

Enrichr

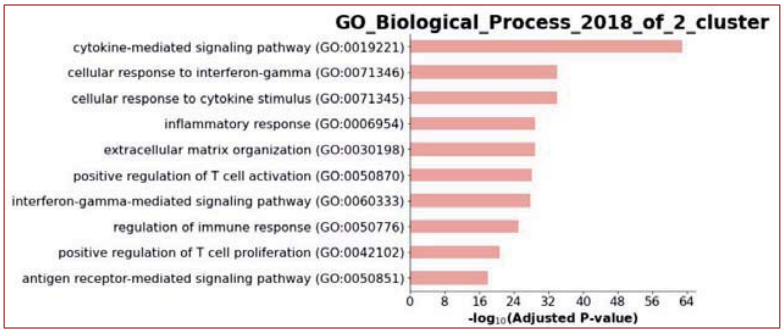
Login | Register

31,333,825 lists analyzed
339,127 terms
171 libraries

Analyze | What's new? | Libraries | Gene search | Term search | About | Help

Gene-set Library	Terms	Gene Coverage	Genes per Term
Genes_Associated_with_NIH_Grants	32876	15886	9
Cancer_Cell_Line_Encyclopedia	967	15797	176
Achilles_fitness_decrease	216	4271	128
Achilles_fitness_increase	216	4320	129
Aging_Perturbations_from_GEO_down	286	16129	292
Aging_Perturbations_from_GEO_up	286	15309	308
Allen_Brain_Atlas_10x_scRNA_2021	766	12361	124
Allen_Brain_Atlas_down	2192	13877	304
Allen_Brain_Atlas_up	2192	13121	305
ARCHS4_Cell-lines	125	23601	2395
ARCHS4_IDG_Coexp	352	20883	299
ARCHS4_Kinases_Coexp	498	19612	299
ARCHS4_TFs_Coexp	1724	25983	299
ARCHS4_Tissues	108	21809	2316
GO_Biological_Process_2018		5103	14433

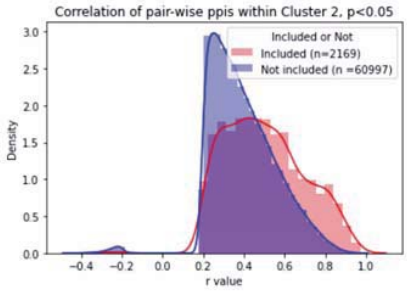
#실습에서는 immune-related pathway 에 enrichment 를 보인 2번째 cluster 에 집중



DEG correlation

#DEG(differentially expressed gene) correlation 분석을 통해 normal-tumor 사이에 유전자 발현 변화 양상이 cluster 2의 protein interacting group 내에서 유전자 간 연관성이 있는 지를 protein interacting 하지 않는 group 과의 비교를 수행; 이를 통해 cluster 2 내부의 protein interacting gene 들은 서로 간 normal-tumor 의 변화의 양상이 연관성을 갖고 있음을 간접적으로 증명

All pair-wise of PPIs



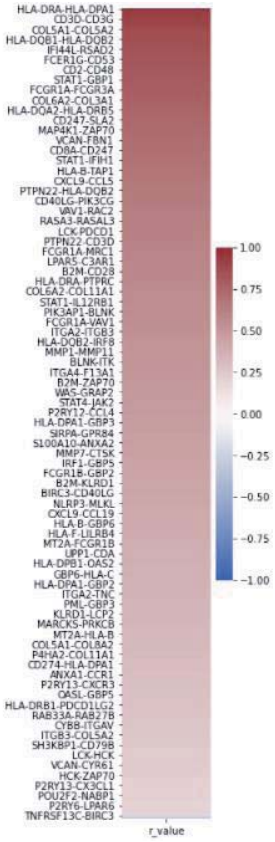

```
ranksums(list(pyr.values()),list(pnr.values()))
```

```
RanksumsResult(statistic=24.842295742953336, pvalue=3.13237299847862e-136)
```

Degree of PPIs

```
B2M        61
HLA-DRB1   60
HLA-DRA    60
LCK        56
PTAFR      55
           ..
CD226       1
AK4         1
CLEC7A      1
HAVCR2      1
NRK         1
Name: node1, Length: 456, dtype: int64
```

#feature = gene



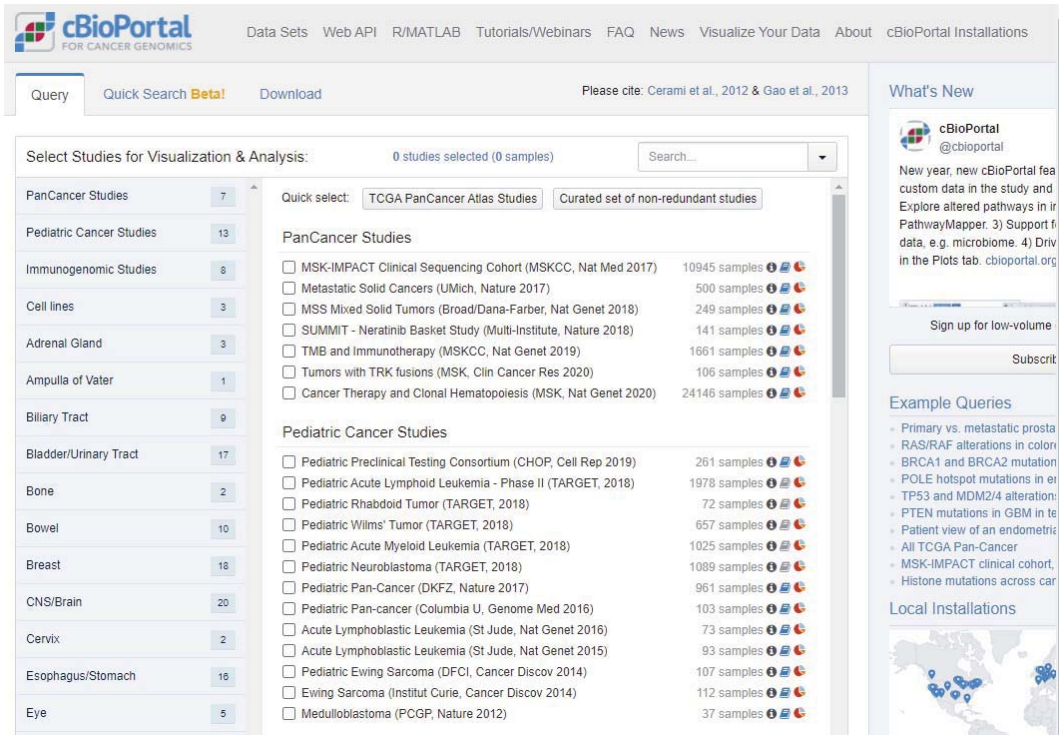
SBi 한국생명정보학회 Korean Society for Bioinformatics

9

cBioPortal

#cBioportal 은 cBioPortal은 다양한 층위의 암 유전체 데이터를 탐색하고 분석하며, 그 결과를 가시적으로 확인할 수 있는 대표적인 온라인 포탈 사이트임
#System 연구에서 중요한 역할을 하는 유전자에 대한 근거 및 단서를 얻을 수 있음 (이후 실습에서 몇 가지 단편적인 예시 다룰 예정)

https://www.cbioportal.org/



SBi 한국생명정보학회 Korean Society for Bioinformatics

10


Thank you
SBi 한국생명정보학회
Korean Society for Bioinformatics