

KSBI-BIML 2026

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists

생명정보학 & 머신러닝 워크샵 (온라인)



Integrative analysis of multi-omics data

정인욱 _ 경북대학교



KSBI
KOREAN SOCIETY FOR
BIOINFORMATICS

한국생명정보학회



본 강의 자료는 한국생명정보학회가 주관하는 BIML 2026 워크샵을 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 행위자 본인에게 있음**을 알립니다.

KSBI-BIML 2026

Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists

한국생명정보학회가 주최하는 BIML-2026 동계 Bioinformatics & Machine Learning 교육 워크숍에 여러분을 초대합니다.

BIML 워크숍은 생명정보학 연구자들이 최신 AI바이오 분야의 인공지능 기반 분석 기술과 바이오 데이터 분석 기법을 이론과 실습을 통해 체계적으로 배울 수 있는 전문 교육 프로그램입니다. 2015년에 시작된 BIML 워크숍은 올해로 12년 차를 맞이하며, 국내 생명정보학 분야의 최초이자 최고 수준의 교육 프로그램으로 자리 잡았습니다. 이번 워크숍은 크게 인공지능바이오(AI바이오) 분야와 디지털바이오 분야, 두 분야로 구성됩니다.

AI바이오 분야에서는 생명정보 분석에 폭넓게 응용되고 있는 다양한 인공지능 기반 자료 모델링 기법을 다룰 예정입니다. 특히, 인공지능 심층학습을 활용한 단백질 구조 예측, 유전체 분석, 신약 개발에 대한 이론 및 실습 강의를 진행됩니다.

또한 디지털바이오 분야에서는 단일세포오믹스, 공간오믹스, 멀티오믹스, 메타오믹스에 대한 강의도 마련되어 있어, 연구자들의 분석 역량 강화에 실질적인 도움을 줄 것으로 기대됩니다.

또한 2024년부터 추가된 의료정보 자료 분석을 다루는 강의를 올해도 지속해서 운영하고자 합니다. 이는 최근 의료정보 자료 분석에 관한 연구 수요 증가를 반영한 것으로, 관련 연구를 수행하는 의과학자 및 의료정보 연구자들에게 유용한 지침을 제공할 것입니다.

또한, 올해도 생명정보학 기술의 다양화에 발맞춰 온라인 강좌를 대폭 확대했습니다. 올해는 무료 강좌 10개를 포함한 총 40개 이상의 강좌가 개설되며, 연구 주제에 맞는 강좌 추천과 강연료 할인 혜택도 제공합니다.

BIML-2026는 국내 주요 연구 중심 대학의 전임 교수 및 각 분야 최고 전문가들의 강의로 구성되어 있으며, 기초 이론부터 최신 연구 동향까지 아우르는 심도 있는 교육의 장이 될 것으로 확신합니다.

여러분의 많은 관심과 참여를 기대합니다!

2026년 2월

한국생명정보학회장 류 성 호

Integrative analysis of multi-omics data

이질적이고 빅데이터인 다중오믹스 데이터는 다양한 생물학 현상을 측정하는데 활용된다. 그러나 다중오믹스 데이터들의 수치와 유전체 적인 요소의 의미가 다르므로 생물학적으로 의미가 있도록 통합 및 분석돼야 한다. 현재 다중오믹스 데이터를 분석한 연구들이 활발히 수행되고 있으며 단일 세포 영역까지 분석분야를 넓히고 있다.

관련 전처리, 통합 및 분석 방법들을 살펴보고 최근에 수행한 다중오믹스 유전자 조절 방법 및 패스웨이 분석 방법을 소개하고자 한다. TCGA의 다양한 암에 대한 다중오믹스 데이터를 활용하여 암의 하위유형을 잘 구분할 수 있는 오믹스 요소 및 패스웨이 발굴을 예시로 강의를 구성하였다.

* 강의 난이도: 초급

* 강의: 정인욱교수 (경북대학교 컴퓨터공학부)

Curriculum Vitae

Speaker Name: Inuk Jung, Ph.D.



► Personal Info

Name Inuk Jung
Title Assistant Professor
Affiliation Department of Computer Science, College of IT,
Kyungpook National University

► Contact Information

Address 80 Daehak-ro, Buk-gu, Daegu 41566
Email inukjung@knu.ac.kr

Research Interest

Machine learning and computational genomics

Educational Experience

2004 B.S. in Computer Science, Canterbury University, New Zealand
2007 M.S. in Computer Science, Yonsei University, Korea
2017 Ph.D. in Interdisciplinary Program in Bioinformatics, Seoul National University

Professional Experience

2007-2011 Research Engineer at LG Electronics, Anyang, Korea
2017-2019 Research Fellow, Bioinformatics Institute, Seoul National University, Korea
2019- Assistant Professor at Department of Computer Science, College of IT,
Kyungpook National University

Selected Publications (5 maximum)

1. Jaemin Jeon, Eon Yong Han and Inuk Jung, "MOPA: An Integrative Multi-Omics Pathway Analysis Method for Measuring Omics Activity", PLOS ONE 2022 (in publication)
2. Inuk Jung, Minsu Kim, Sungmin Rhee, Sangsoo Lim and Sun Kim, MONTI: A Multi-Omics Non-negative Tensor Decomposition Framework for Gene-Level Integrative Analysis, *Frontiers in Genetics*, 10 September 2021
3. Minsik Oh, Sungjoon Park, Sangseon Lee, Dohoon Lee, Sangsoo Lim, Dabin Jeong, Kyuri Jo, Inuk Jung and Sun Kim, "DRIM: A Web-Based System for Investigating Drug Response at the Molecular Level by Condition-Specific Multi-Omics Data Integration", *Frontiers in Genetics*, 12 November 2020
4. Inuk Jung, Jounghmin Choi, and Heejoon Chae, "A non-negative matrix factorization based framework for the analysis of multi-class time-series single-cell RNA-seq data." *IEEE Access* 2020
5. Sangsoo Lim, Sangseon Lee, Inuk Jung, Sungmin Rhee, Sun Kim, "Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data", *Briefings in Bioinformatics* 2018

KSBi-BIML 2023

Multi-Omics Factor Analysis

“Integrative analysis of multi-omics data”

Inuk Jung (inukjung@knu.ac.kr)

College of IT Engineering, School of Computer Science and Engineering
Kyungpook National University

Contents

1. Multi-omics overview

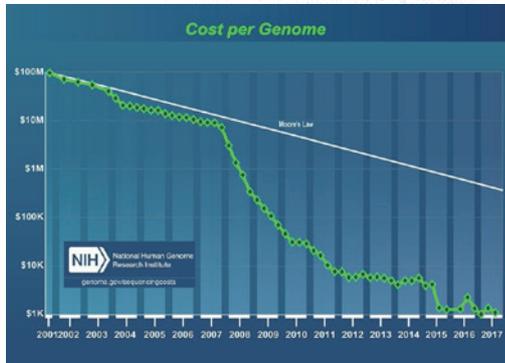
2. Multi-omics methods

- SNF, jointNMF, MOFA

3. Multi-omics research

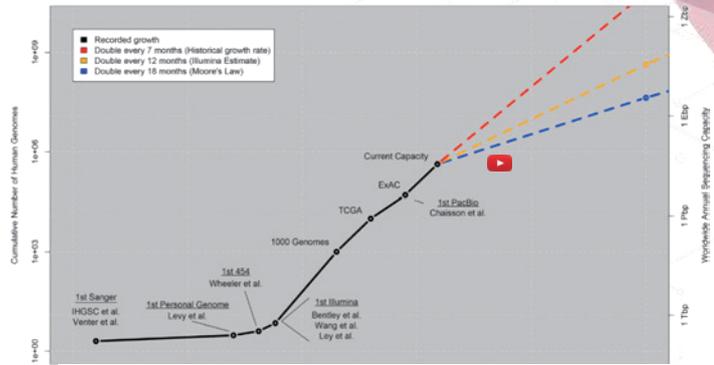
- Factor analysis in gene level
- Multi-omics parameter analysis
- Factor analysis in pathway level

Bio Data | Trend of Cost & Volume



\$100M/Sample

\$1000/Sample



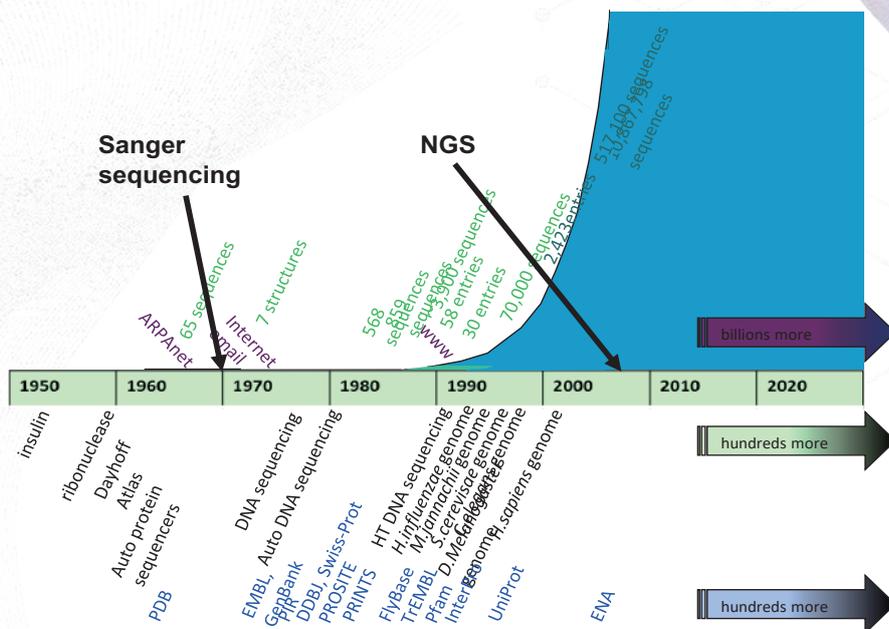
2000

2010

2018

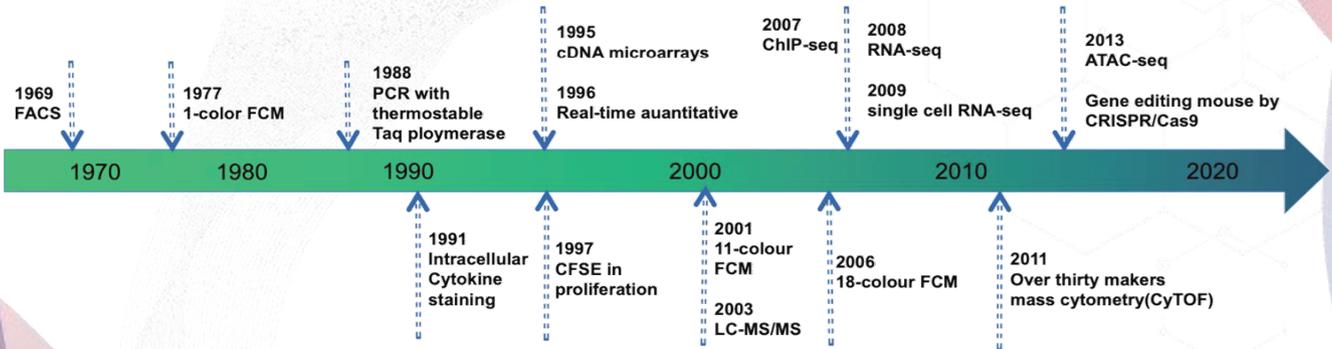
2025

Timeline of Biological data increase



Source from Teresa K. Attwood, University of Manchester

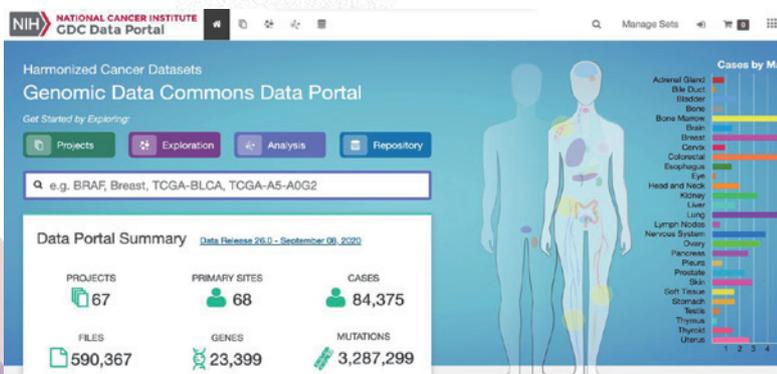
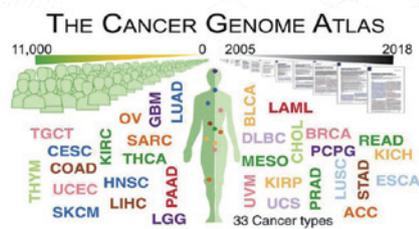
Timeline of multi-omics technologies



Kang Ning & Yuxue Li, Introduction to Multi-Omics, Translational Bioinformatics book series 2023

A Big Data Example

- The Cancer Genome Atlas (TCGA)



NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over **2.5** PETABYTES of data

To put this into perspective, 1 petabyte of data is equal to **212,000** DVDs

TCGA data describes **33** DIFFERENT TUMOR TYPES

...including **10** RARE CANCERS

...based on paired tumor and normal tissue sets collected from **11,000** PATIENTS

...using **7** DIFFERENT DATA TYPES

Hypothesis-driven

- 1) Observe some phenomenon and
- 2) create a hypothesis



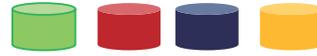
Create data related to the hypothesis



Validation:
Accept or reject the hypothesis

Data-driven

Collect tons of data



- 1) Search for patterns



- 2) Create hypothesis

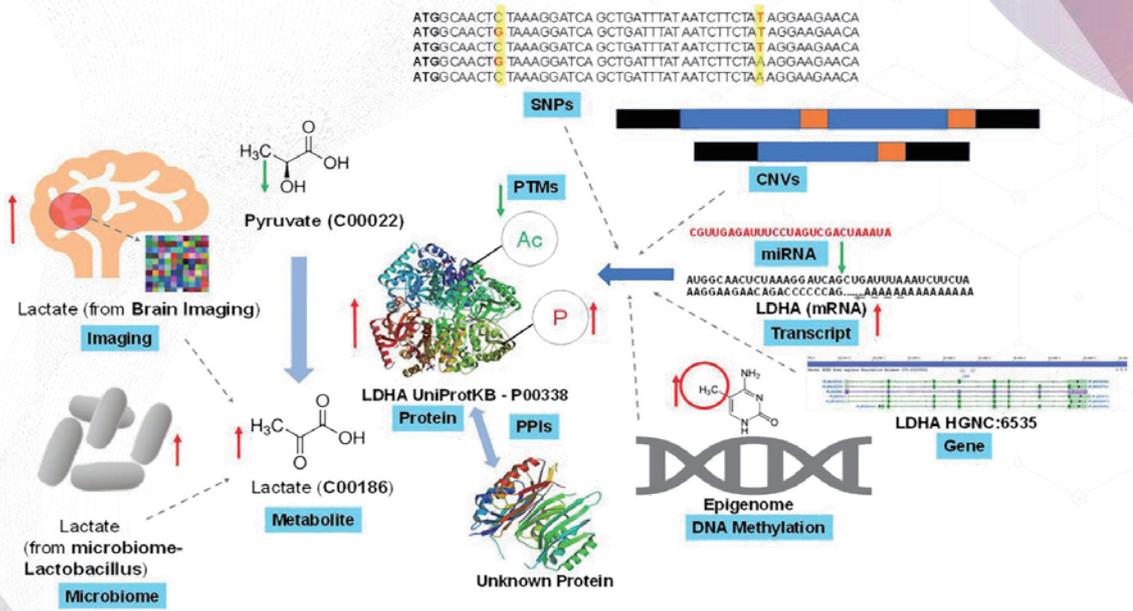


Validation:
1. Look back into the data
2. Counsel domain experts for logical correctness

1. Multi-omics **BIG DATA**

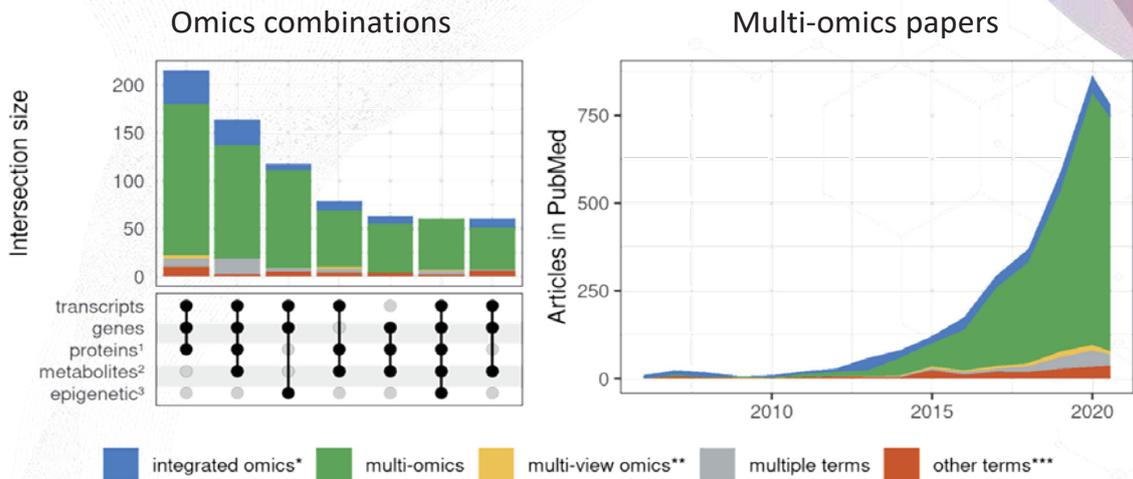


Multi-omics interconnections



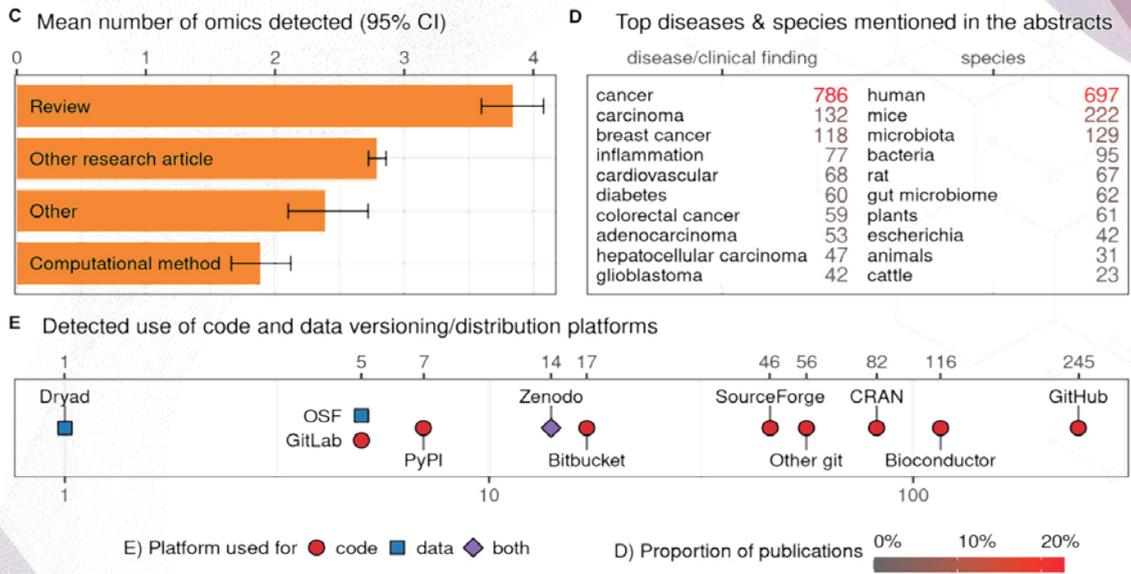
Krassowski, Michal, et al. "State of the field in multi-omics research: From computational needs to data mining and sharing." *Frontiers in Genetics* 11 (2020)

Trend in multi-omics research



Krassowski, Michal, et al. "State of the field in multi-omics research: From computational needs to data mining and sharing." *Frontiers in Genetics* 11 (2020)

Trend in multi-omics research

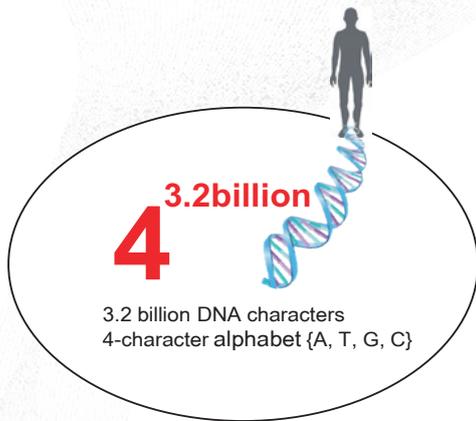


Krassowski, Michal, et al. "State of the field in multi-omics research: From computational needs to data mining and sharing." *Frontiers in Genetics* 11 (2020)

Some challenges in MO analysis

1. Each MO data are big and needs MO-specific preprocessing
2. Heterogeneous and high dimensional data handling
3. Integration is not easy and each method focuses on a different issue
(need to decide what to look for)
4. Selection of appropriate ML method

Challenges | Large Search Space and High Dimensional



Search space

Low sample,
High dimension problem



1 sample

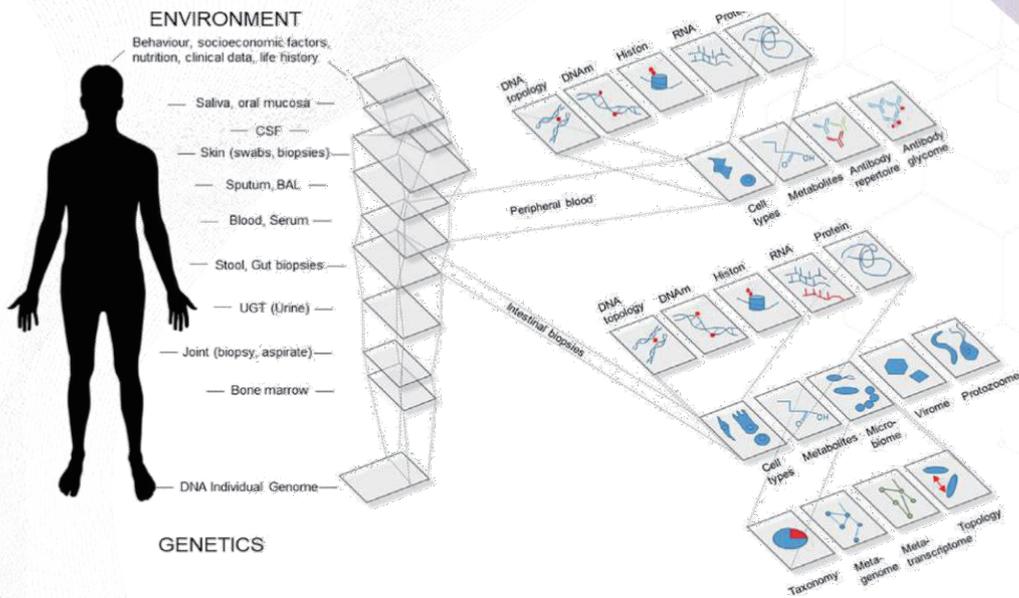


Genes

- B2M
- HNRPLL
- HPRT
- PSMB2
- PSMB4
- PPIA
- PRPS1
- PRPS1L1
- PRPS1L3
- PRPS2
- ...

35K genes

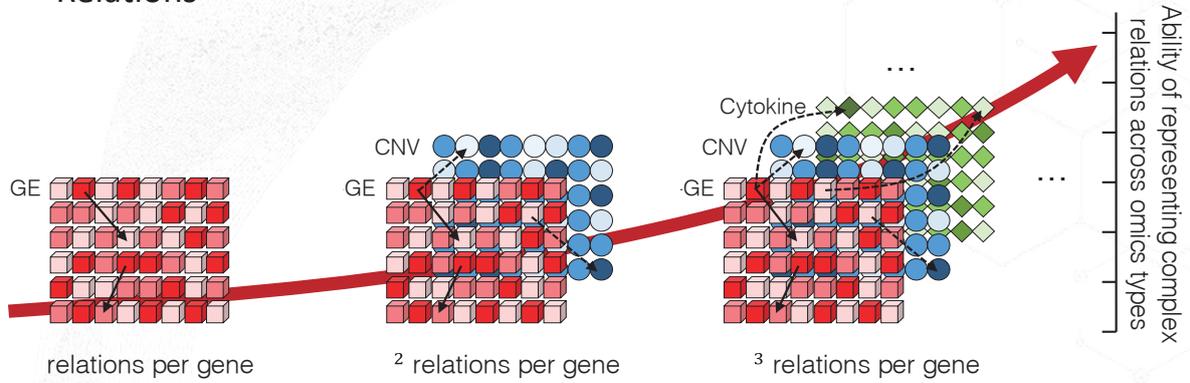
Challenges | Multi-modal data adds extra dimensions



<https://www.mynewgut.eu/sites/default/files/3-Hadrich.pdf>

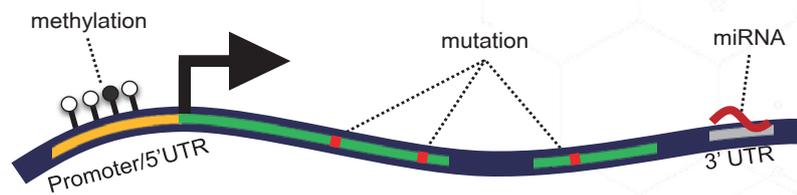
Expectations from multi-omics data

- Interpretability
- Relations



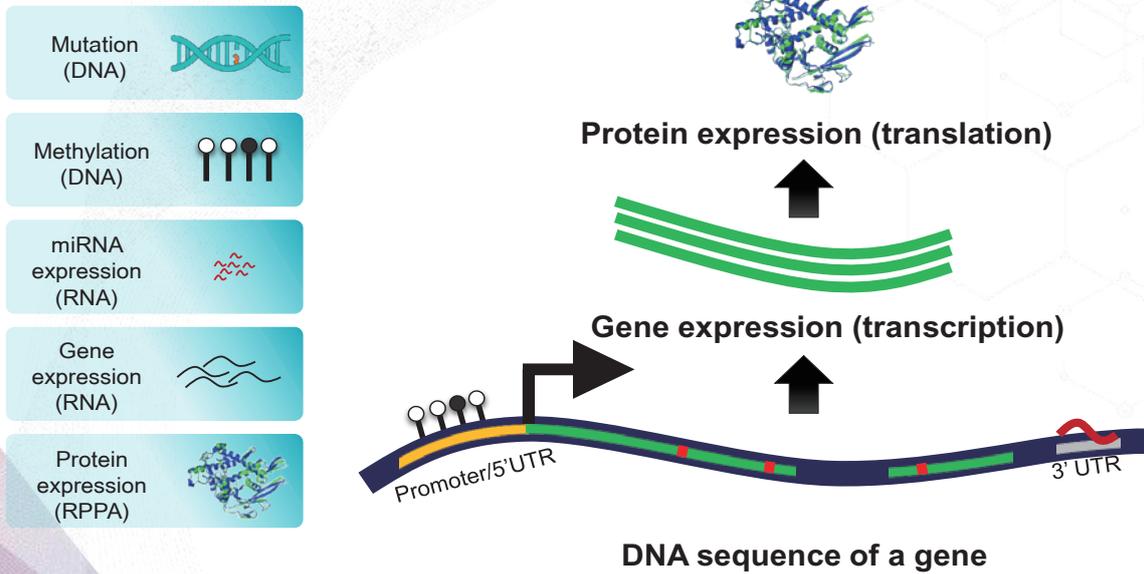
Challenges | Multi-modal data adds extra dimensions

- Mutation (DNA) 
- Methylation (DNA) 
- miRNA expression (RNA) 
- Gene expression (RNA) 
- Protein expression (RPPA) 



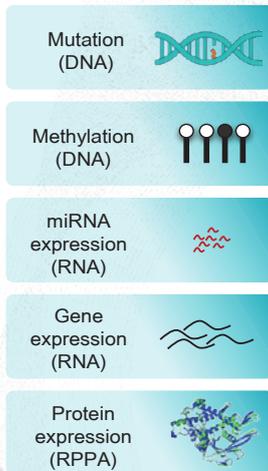
DNA sequence of a gene

Challenges | Multi-modal data adds extra dimensions



Challenges | Searching for explainable omics causality

Omics (genomics) features



Clinical features

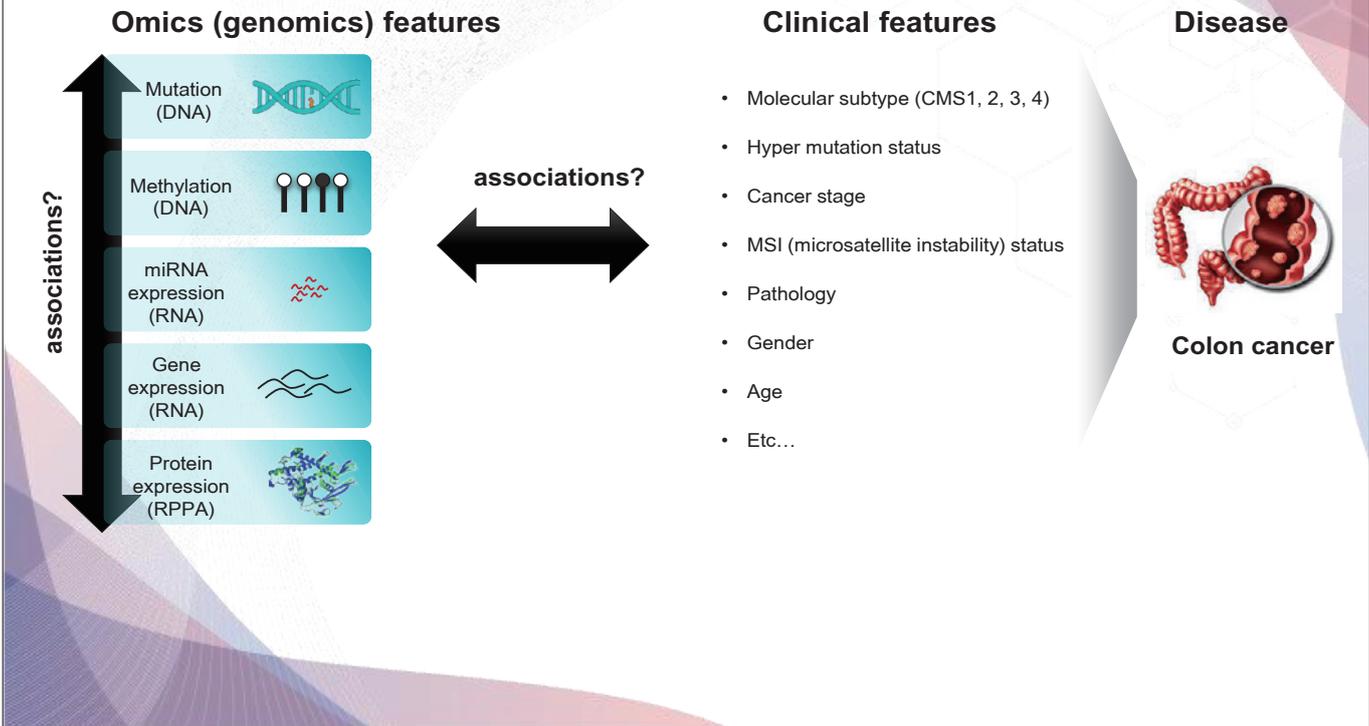
- Molecular subtype (CMS1, 2, 3, 4)
- Hyper mutation status
- Cancer stage
- MSI (microsatellite instability) status
- Pathology
- Gender
- Age
- Etc...

Disease

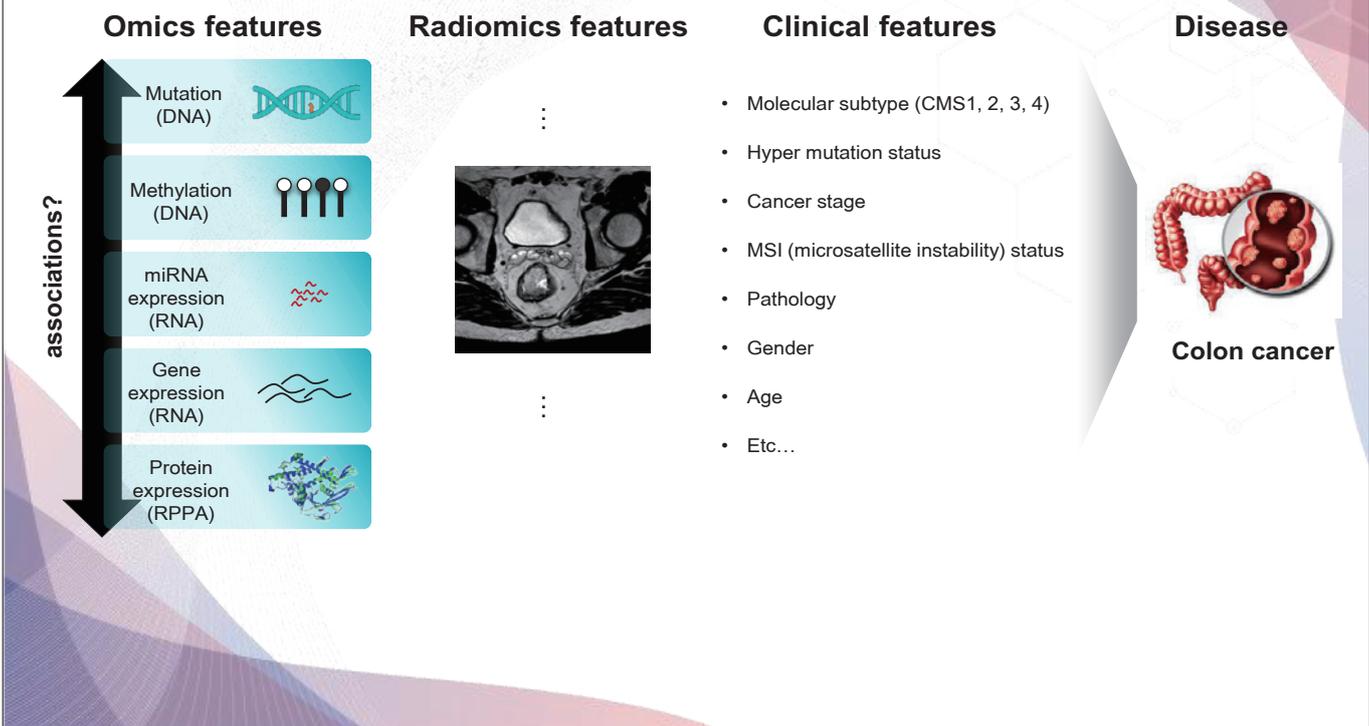


Colon cancer

Challenges | Searching for explainable omics causality



Challenges | Searching for explainable omics causality



Some questions to ponder on

- Is multi-omics better than single-omics?
 - More data = higher quality of result?
- How much (at least) data do we need?
- What type of omics associate well together?
- What types of clinical features are explainable by MO?

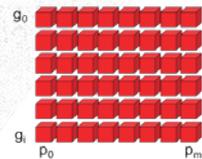
- How do we associate the multiple-omics concepts?
- How do we analyze the integrated data?
- and how do we interpret the results?

Background | Two ways of integrating multi-omics data

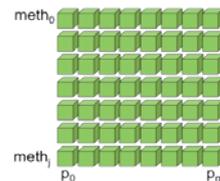
Multi-dimensional integration

(most methods)

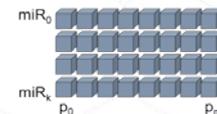
Gene expression
($i \times m$)



Methylation level
($j \times m$)

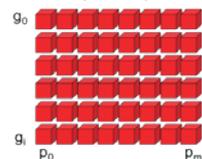


miRNA expression
($k \times m$)

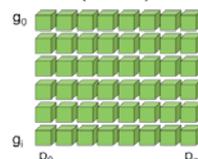


Multi-staged integration

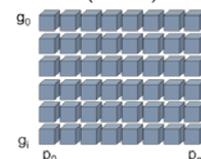
Gene expression
($i \times m$)



Methylation level
($i \times m$)



miRNA expression
($i \times m$)



MO Analysis Methods |

- <https://github.com/mikelove/awesome-multi-omics>

Factor analysis

Multi-omics correlation or factor analysis

- 2007 - SCCA - Parkhomenko - sparse CCA - paper 1, paper 2
- 2008 - PCCA - Waaijenborg - penalized CCA / CCA-EN - paper
- 2009 - PMA - Witten - Sparse Multi CCA - paper 1, paper 2
- 2009 - sPLS - Lê Cao - sparse PLS - paper
- 2009 - pesca - Hwang - RGSCA regularized generalized structured component analysis - paper
- 2010 - Regularized dual CCA - Soneson - paper
- 2011 - RGCCA - Tenenhaus - Regularized Generalized CCA and Sparse Generalized CCA - paper 1, paper 2
- 2011 - SNMNMf - Zhang - Sparse Network-regularized Multiple Non-negative Matrix Factorization - paper
- 2011 - scca - Lee - Sparse Canonical Covariance Analysis for High-throughput Data - paper

...

- 2020 - MOTA - Fan - network-based multi-omic data integration for biomarker discovery - paper
- 2020 - D-CCA - Shu - Decomposition-based Canonical Correlation Analysis - paper
- 2020 - COMBI - Hawinkel - Compositional Omics Model-Based Integration - paper
- 2020 - DPCCA - Gundersen - Deep Probabilistic CCA - paper
- 2020 - MEFISTO - Velten - spatial or temporal relationships - preprint
- 2020 - MultiPower - Tarazona - Sample size in multi-omic experiments - paper
- 2020 - mixedCCA - Yoon - Sparse semiparametric CCA for data of mixed types - paper

Network analysis

Multi-omics networks

- 2018 - MoTI-DREAM - Didier - identifying communities from multiplex networks, and annotated the obtained clusters article
- 2019 - RWR-MH - Valdeolivos - Random walk with restart on multiplex and heterogeneous biological networks article
- 2020 - MOGAMUN - Novoa-del-toro - A multi-objective genetic algorithm to find active modules in multiplex biological networks preprint
- 2021 - RWRf - Wen - Random Walk with Restart for multi-dimensional data Fusion paper

Clustering

Multi-omics clustering / classification / prediction

Note: I think that prediction of genomic tracks, e.g. ChIP-seq, from other genomic tracks is a large area of research that may deserve a separate repository. Below are methods for clustering / classification of samples into sub-types or prediction of outcomes.

- 2009 - iCluster - Shen - paper
- 2012 - MDI - Kirk - paper1, paper2
- 2013 - iClusterPlus - Mo - paper
- 2013 - BCC - Lock - Bayesian consensus clustering - paper
- 2013 - IBAG - Wang - Integrative Bayesian Analysis of Genomics - paper
- 2014 - SNF - Wang - paper
- 2017 - clusteromics - Gabasova - paper
- 2019 - iBOOST - Wong - paper
- 2019 - Spectrum - John - paper
- 2020 - NF - Chiarici and Busato - paper
- 2021 - CustOmics - Birke - Consensus clustering - paper

Single-cell

Single cell multi-omics

- 2018 - cardellino - gene expression states to clones (SNVs from scRNA-seq + bulk exome data) -
- 2018 - clonealign - Campbell - gene expression states to clones (scRNA-seq + scDNA-seq (CNV)) - paper
- 2020 - CiteFuse - Kim - CITE-seq data analysis paper
- 2021 - CoSpar - Wang - infer dynamics by integrating state and lineage information - paper

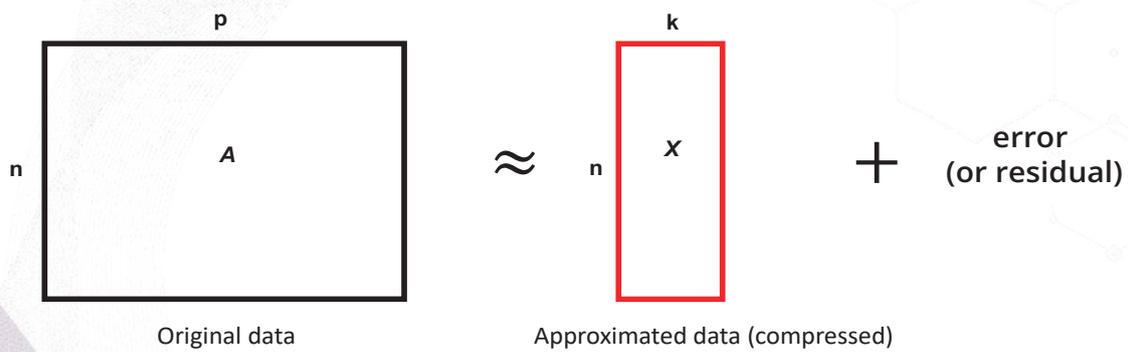
reviews and many more...

MO Factor analysis

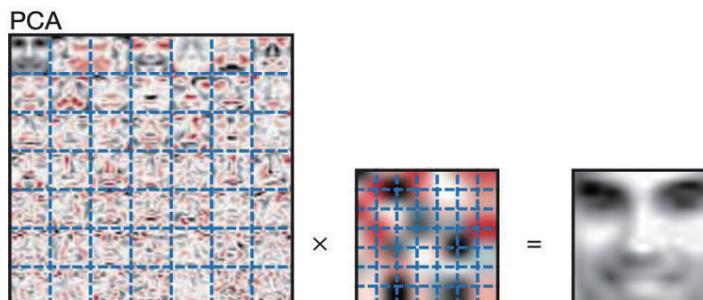
- **Factor analysis** focuses on reducing dimensions for representative learning on a more simple or lower space than the original data
- It's advantage lies in finding strong signals and alleviate interpretation of the result
- In addition to factor analysis, multi-omics is often analyzed using multiview learning
- But why reduce dimensions?

A brief overview of factor analysis

- Principal Component Analysis (PCA) and Non-negative matrix factorization (NMF) are well known dimension reduction methods
- PCA concept



PCA example



Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401.6755 (1999): 788-791.

A brief overview of factor analysis

- Principal Component Analysis (PCA) and Non-negative matrix factorization (NMF) are well known dimension reduction methods
- NMF concept

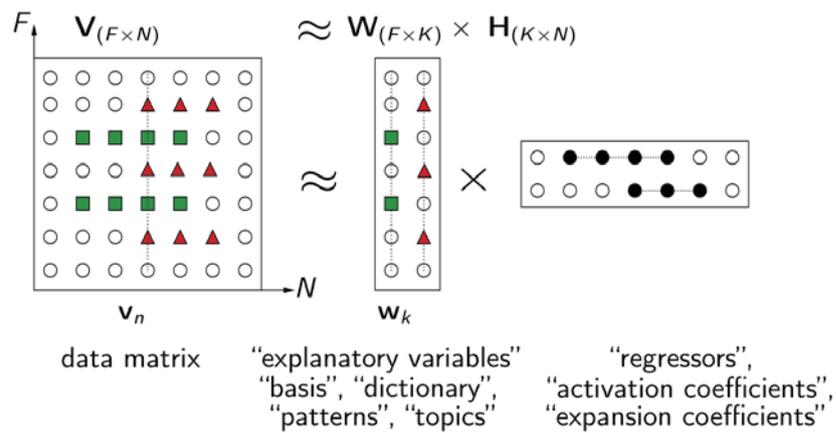
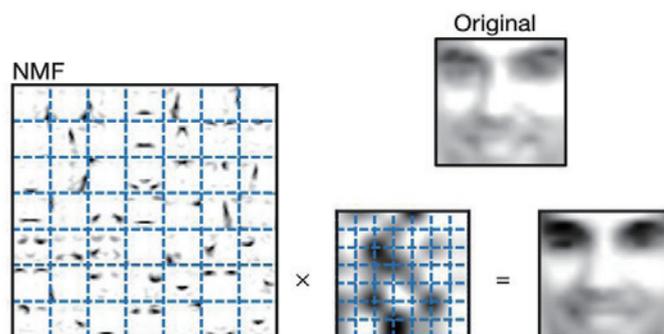


Illustration by C. Févotte

NMF example

- NMF



Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401.6755 (1999): 788-791.

A brief overview of factor analysis

- The input to PCA and NMF are 2D data
- A single-omics data is 2D data, whereas multi-omics needs a set of 2D data
- Need to extend the factor analysis methods for MO analysis
- Multi-view learning is also a good way for MO analysis
 - Each omics is a view
 - and the multi-omics (views) are co-trained or co-regularized
 - at an early or late stage

Zhao, Jing, et al. "Multi-view learning overview: Recent progress and new challenges."
Information Fusion 38 (2017)

Multi-omics (factor) analysis methods

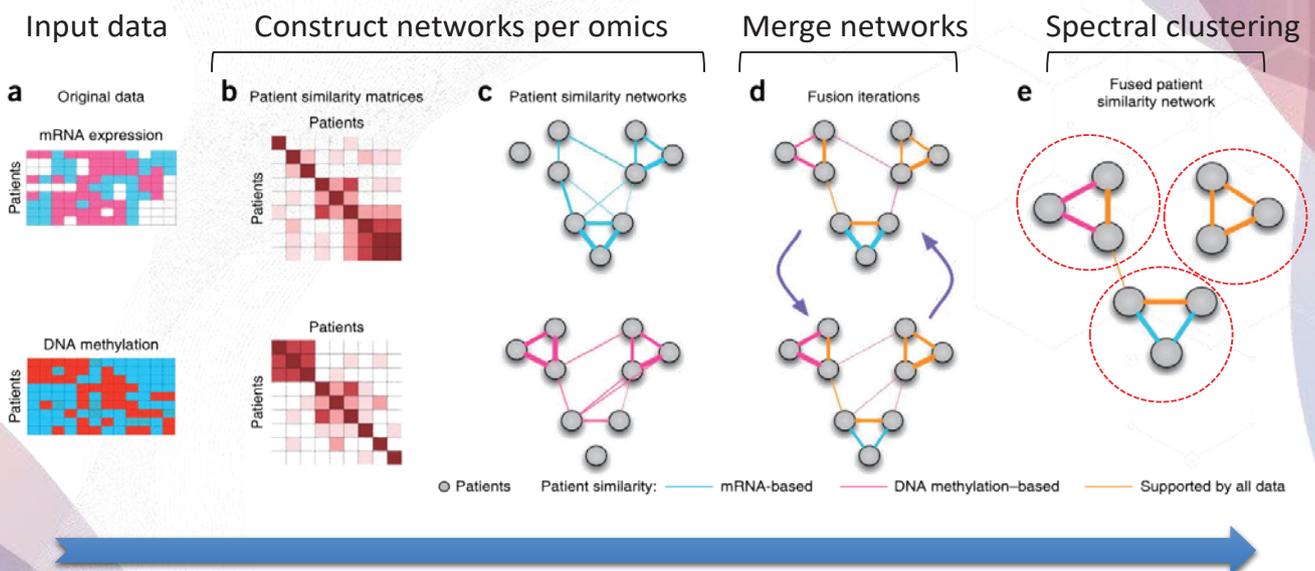
- jointNMF (2012)
- SNF (2014)
- MOFA (2018)
- MONTI (2021)
- MOPA (2023)

Methods | SNF – Similarity Network Fusion

- A network based multi-omics integration method for detecting patient groups with high omics-network similarity
- Omics: mRNA, methylation, miRNA
- Samples: 215 glioblastoma multiforme patients (and BIC, KRCCC, LSCC and COAD cancer types)
- Advantage:
 - Not limited to certain omics type
 - Very fast
 - Works with small no. of samples
 - Provides patient clustering function for patient module detection
- Disadvantage:
 - Merges patients, thus does not pinpoint specific omics features (post-processing required)

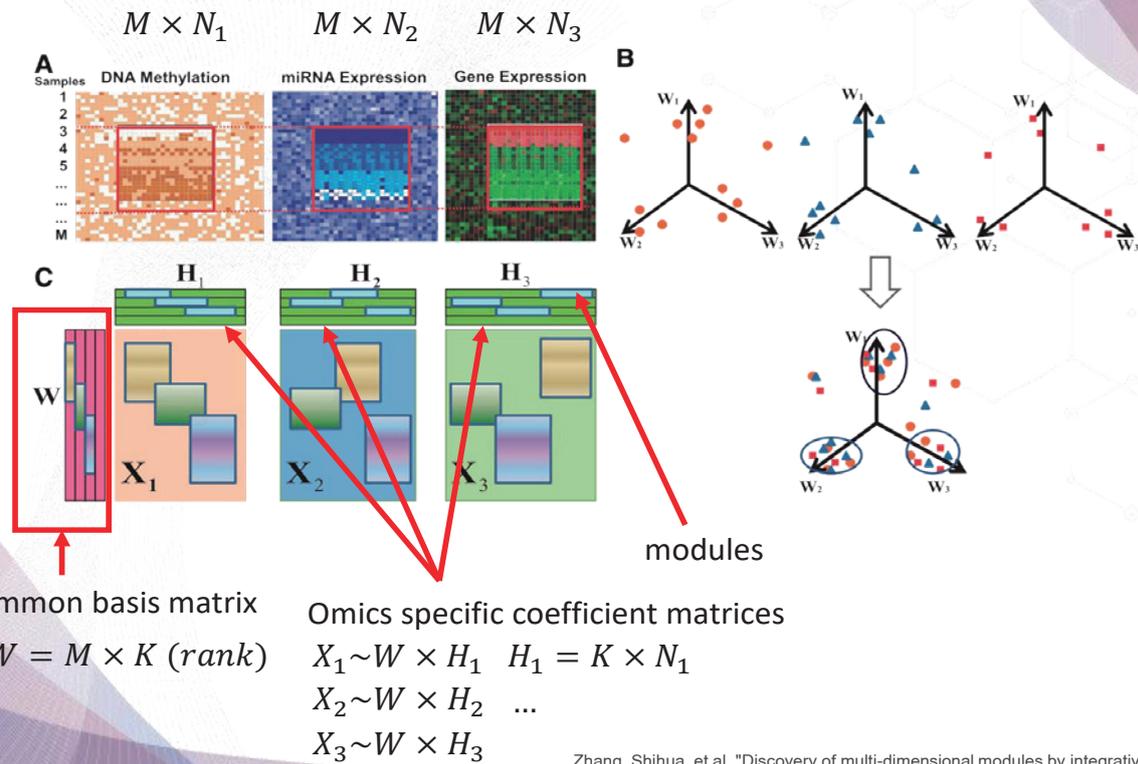
Wang, Bo, et al. "Similarity network fusion for aggregating data types on a genomic scale." *Nature methods* 11.3 (2014): 333-337.

Methods | SNF – concept



Methods | JointNMF – concept

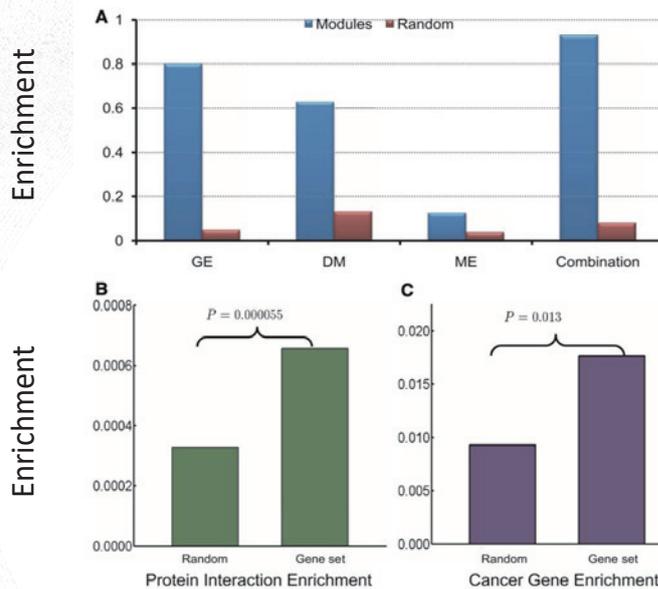
M = Samples
 N_l = Omics features (gene, methylation, miRNA)



Zhang, Shihua, et al. "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data." *Nucleic acids research* 40.19 (2012): 9379-9391.

Methods | JointNMF – Results

- Module associated omics features had high enrichment scores when using the ovarian cancer dataset (K=200 modules, ~80% modules were biologically relevant)

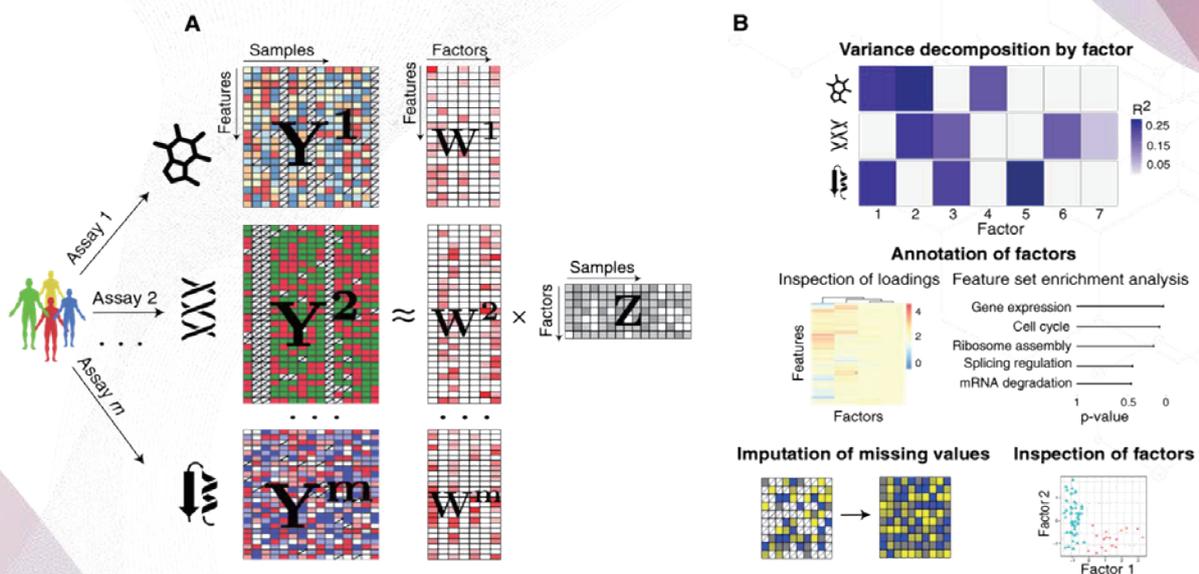


Methods | MOFA – Multi-Omics Factor Analysis

- A factor analysis method for integrating MO data and detecting important factors (or components) related to a specific group
- Samples: 200 chronic lymphocytic leukaemia (CLL)
- Omics: mRNA, methylation, mutation, ex vivo drug response screens
- Advantage:
 - Not limited to certain omics type
 - Able to impute missing values
 - Outputs important omics features with association to some interest
- Disadvantage:
 - Slow with large number of samples and features
 - Constrained number of max. factors
 - Omics features are selected without correlation (post-processing required)

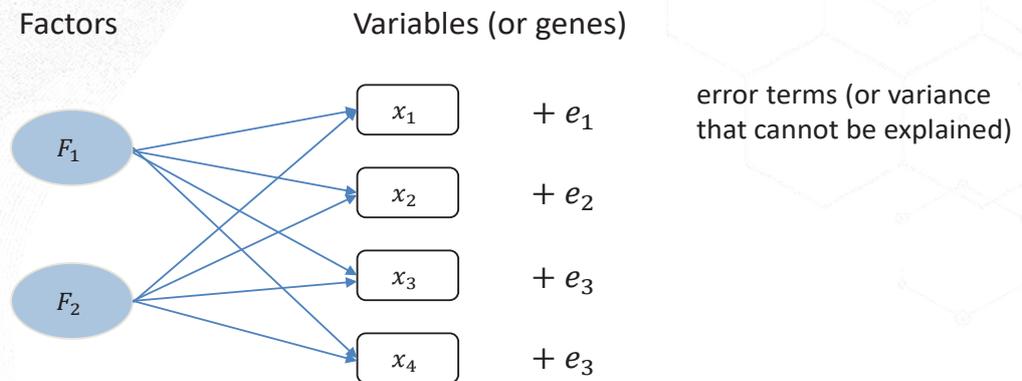
Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." *Molecular systems biology* 14.6 (2018): e8124.

Methods | MOFA – concept



Methods | MOFA – concept

- Factor analysis (or PCA) is different from matrix factorization
- FA is based on variance and learns weights (eigen values) accordingly

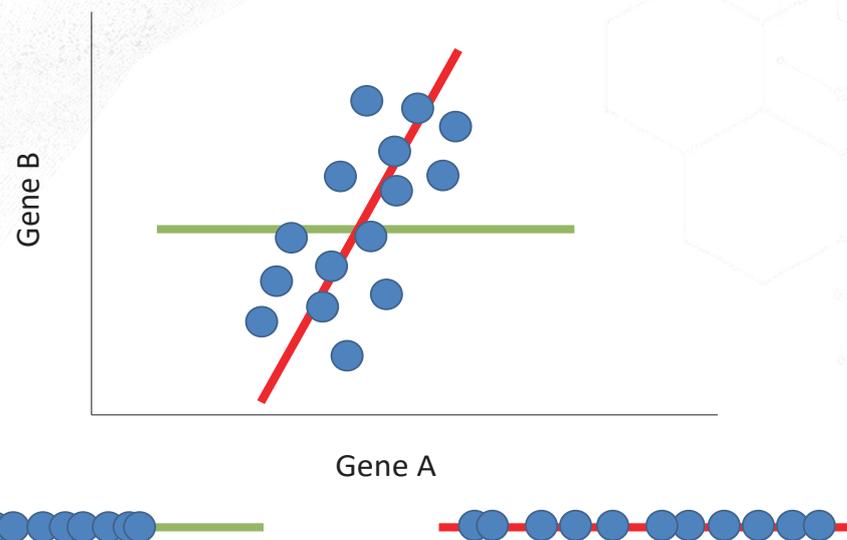


$$x_1 = b_{11} \times F_1 + b_{12} \times F_2 + e_1$$

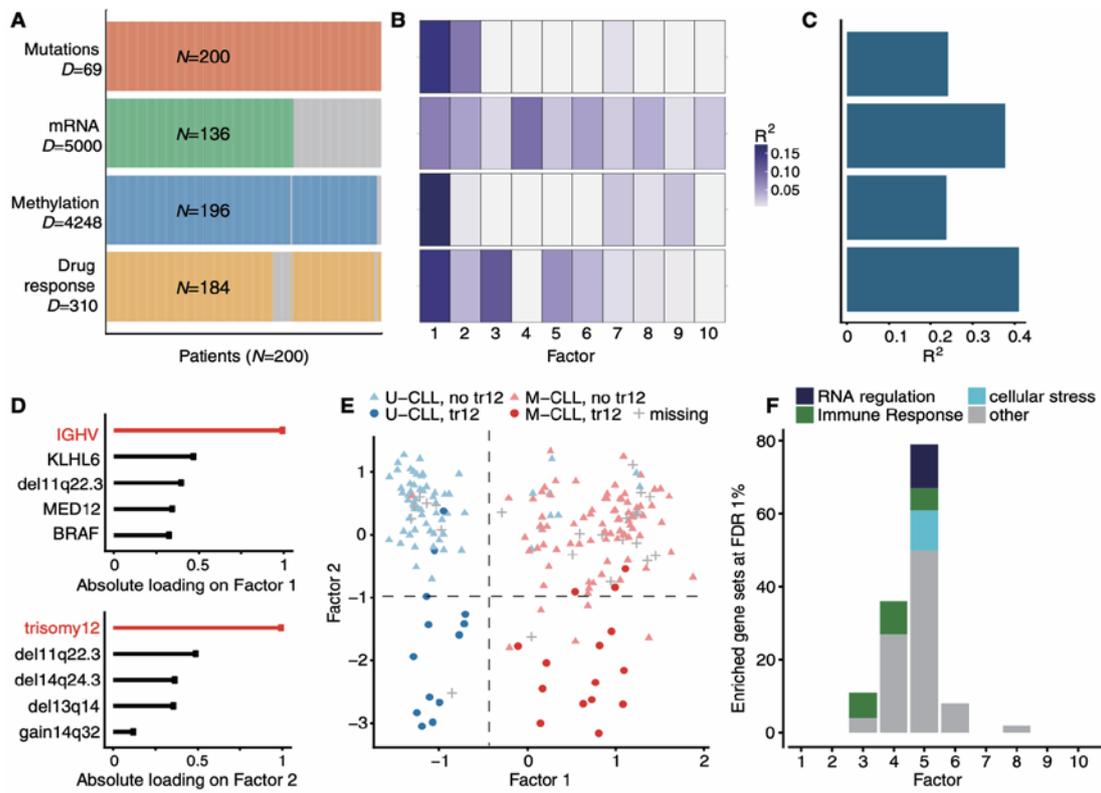
$$F_1 = w_{11} \times x_1 + w_{12} \times x_2 + w_{13} \times x_3 + w_{14} \times x_4$$

Brief overview of PCA

- Finds the line that maximizes variance in the data

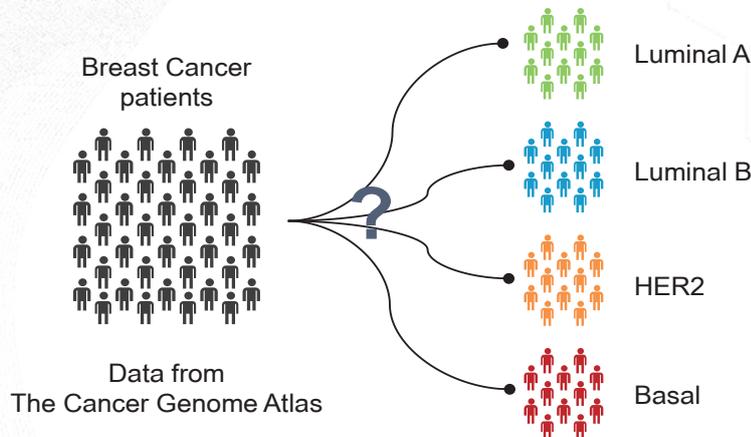


Methods | MOFA – Results



Multi-omics research

“Clinically observable plasticity and heterogeneity occurs within, and not across, the major biological **subtypes** of breast cancer”
TCGA, Nature 2012

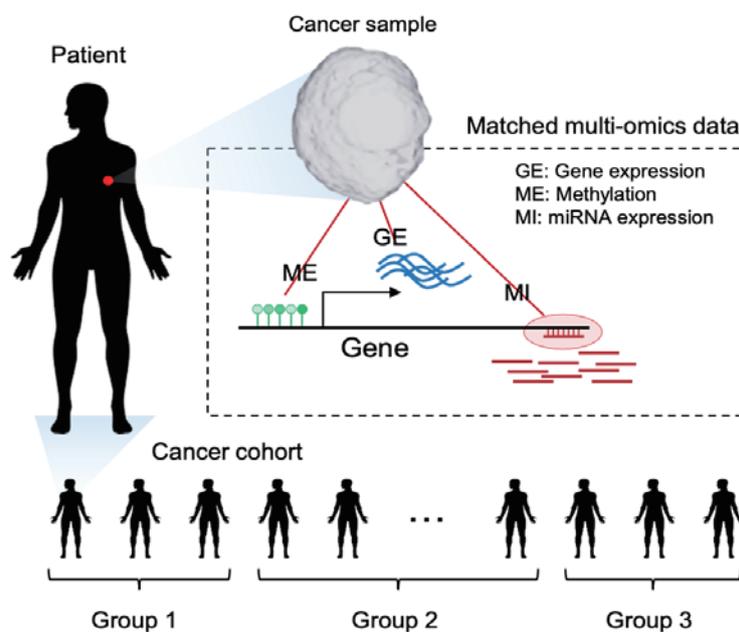


Some recent research topics on MO

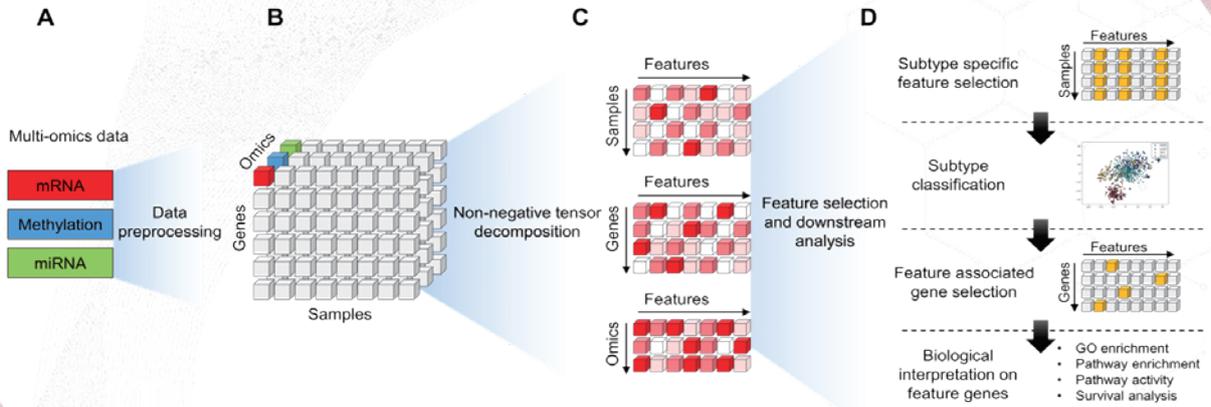
- MONTI: A multi-omics non-negative tensor decomposition framework for the integrated analysis of cancer subtypes (Frontiers in Genetics, 2021)
- MOPA: An Integrative Multi-Omics Pathway Analysis Method for Measuring Omics Activity (in preparation)
- Parametric analysis on large-scale multi-omics data
- Graph based autoencoder for omics relation discovery (under development)

Data mining omics relationships that are specific to some patient group = **interpretation of result**

Multi-Omics Data |

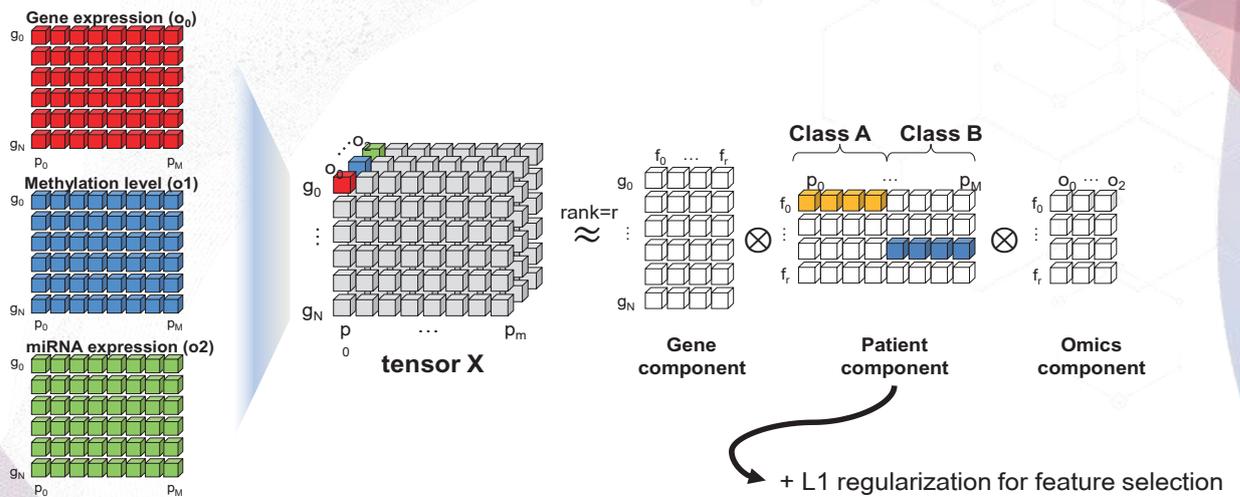


Methods | MONTI Workflow



Jung, Inuk, et al. "MONTI: A Multi-Omics Non-negative Tensor Decomposition Framework for Gene-Level Integrative Analysis." *Frontiers in Genetics* 12 (2021).

Methods | Non-negative Tensor decomposition

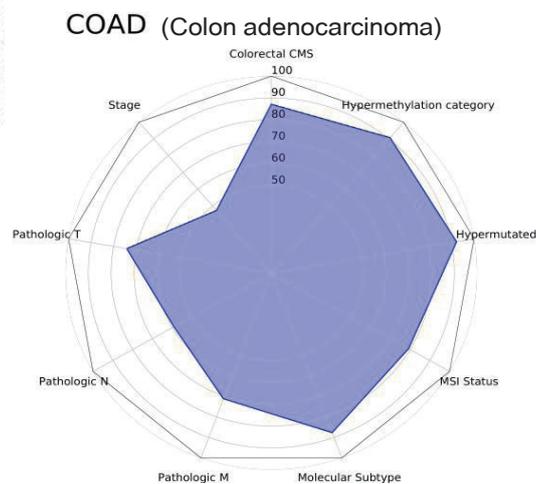


Data | Multi-omics Data of 10 Cancer Types and clinical features

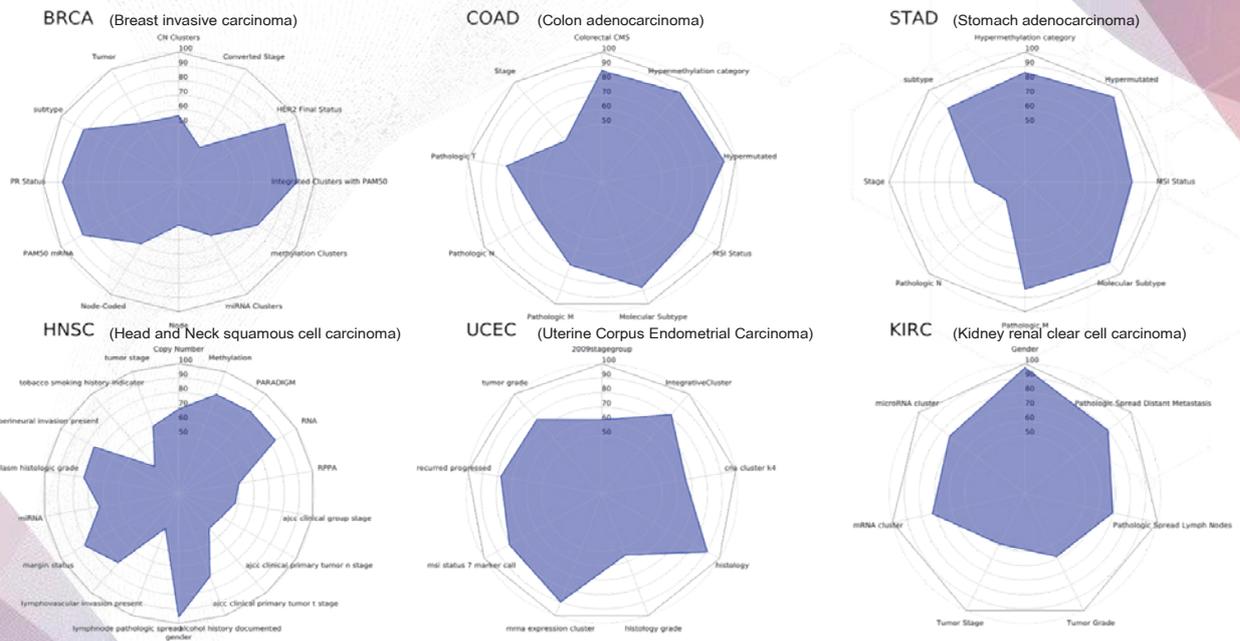
- From the TCGA portal, mRNA, methylation and miRNA omics data were collected
- Also, associated patient clinical data were archived

CANCER TYPE	Clinical_Type	Patient_number	Sample_Type	Gene_number
COAD	Colorectal_CMS	206	['CMS1', 'CMS2', 'CMS3', 'CMS4']	14454
STAD	Molecular_Subtype	305	['CIN', 'EBV', 'GS', 'MSI']	
BRCA	subtype	595	['Basal', 'Her2', 'LumA', 'LumB']	
HNSC	gender	298	['FEMALE', 'MALE']	
OV	TUMORSTAGE	320	['IIIC', 'IV']	
PRAD	methylation_cluster	328	[1, 2, 3, 4]	
KIRC	Gender	252	['FEMALE', 'MALE']	
LUAD	methylation_signature	181	['high_', 'intermediate_', 'low_']	
THCA	BRAF	490	[0, 1]	
UCEC	mrna_expression_cluster	221	[1, 2, 3]	

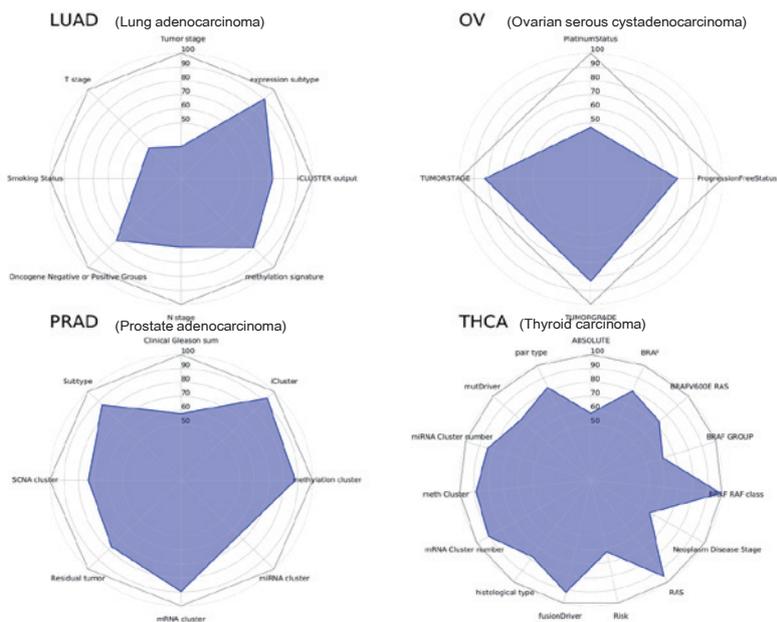
Results | Clinical Feature Classification Accuracy on 10 Cancer types



Results | Clinical Feature Classification Accuracy on 10 Cancer types

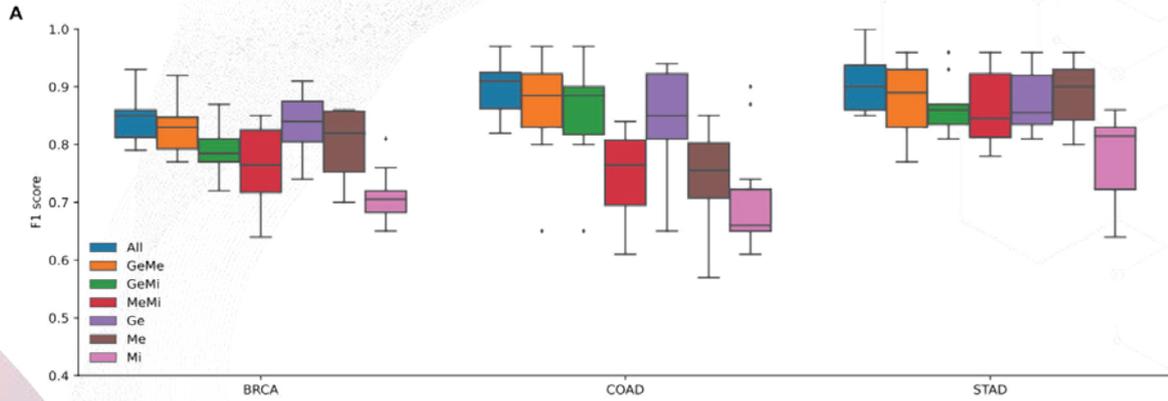


Results | Clinical Feature Classification Accuracy on 10 Cancer types



Results | Omics combinations

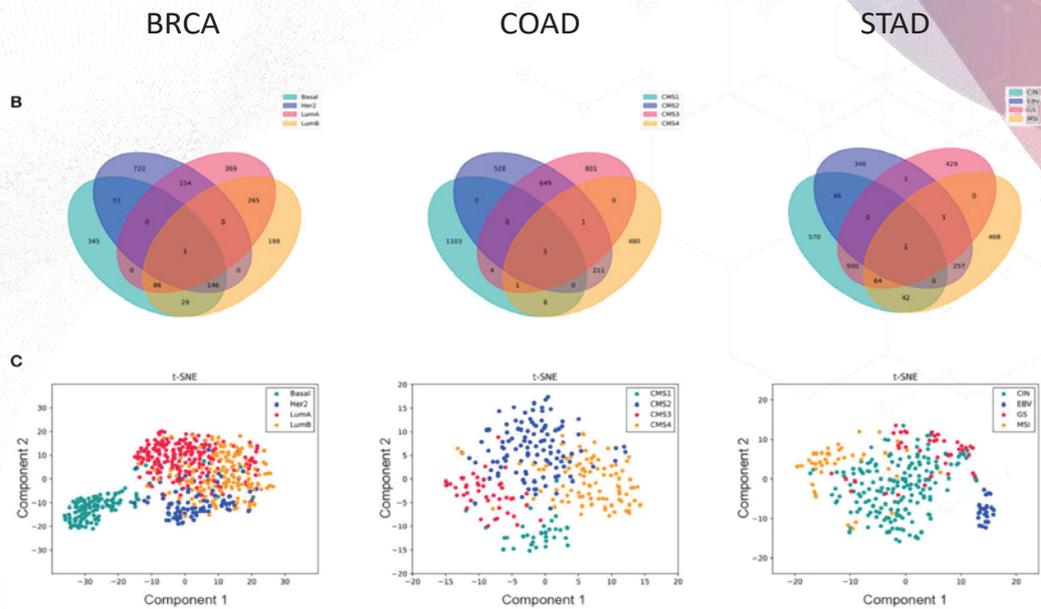
- Different combinations show different results



Results | Subtype features

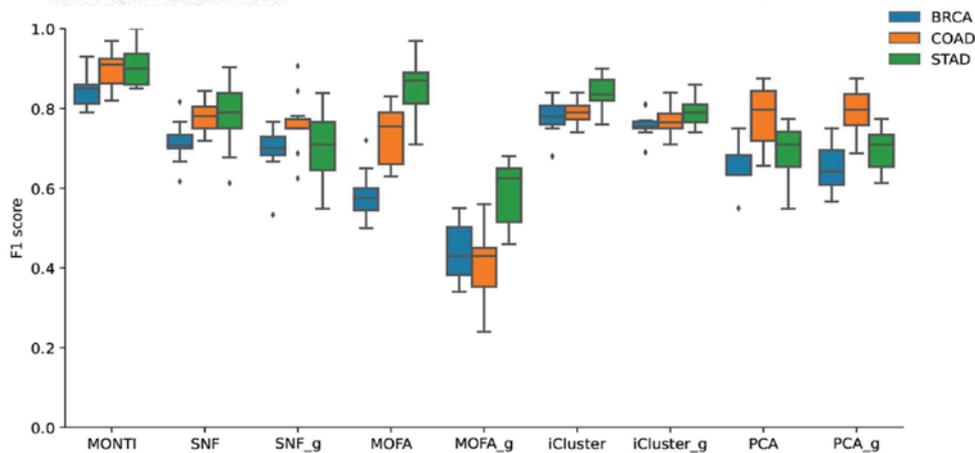
- Some features are shared across the subtypes

- The subtypes of BRCA, COAD and STAD are well separated



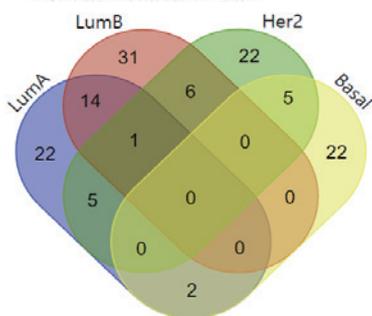
Results | Is gene-level integration helpful? (1)

- Some tools are sensitive to omics-units
- In most tools, gene-level analysis showed lower performance

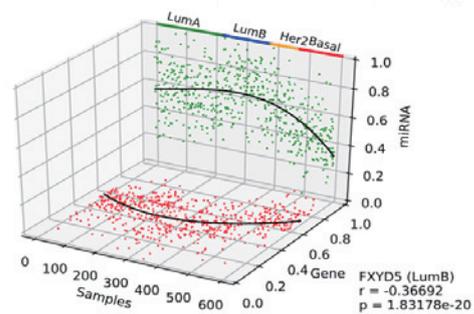


Results | Is gene-level integration helpful? (2)

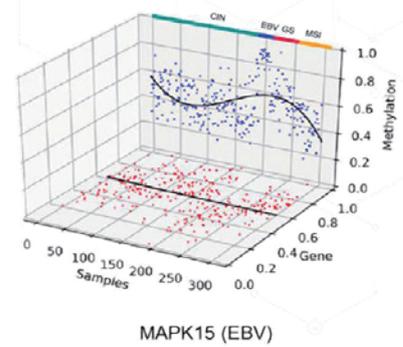
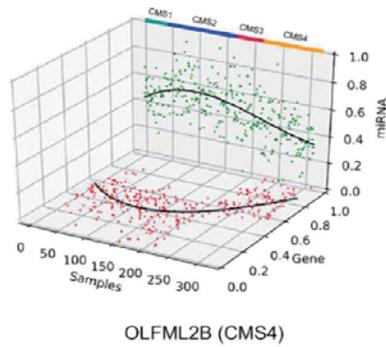
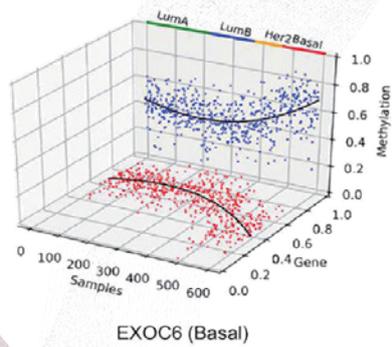
Total of 130 features selected



FXD5 gene and miRNA expression show significant relationship



Results | Is gene-level integration helpful? (3)



Results | Cancer-subtype associated genes

Case study	Ranks	Features	Genes	Subtypes	St-Features	St-Genes
BRCA	120	26	2,385	Luminal A	10	879
				Luminal B	9	732
				Her2	11	1,080
				Basal	8	665
COAD	120	31	3,831	CMS1	7	1,129
				CMS2	9	1,403
				CMS3	11	1,473
				CMS4	10	704
STAD	120	37	5,461	CIN	9	1,234
				GS	9	1,007
				MSI	9	839
				EBV	8	652

Some questions |

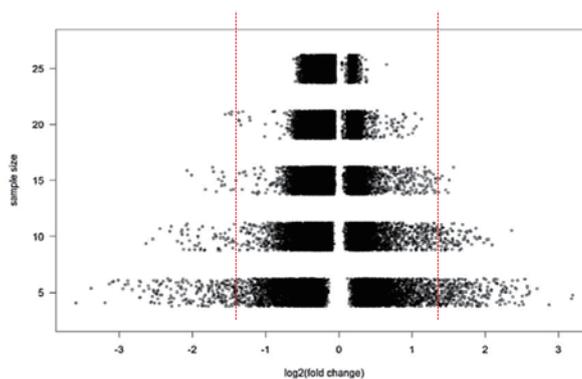
- What clinical features can be explained by MO?
- How many samples are at least required for robust results?
- How many genes or MO features are needed?
- What happens if sample size is not balanced between groups?

Sample size estimation – miRNA

- For miRNAs, at least 19 samples per experimental group is needed to achieve a power of 0.8 at a fold change of 1.5 with FDR < 0.1

Table 1
Both numbers of false-negative and false-positive results increase with a decreasing sample size.

	5 vs 5	10 vs 10	15 vs 15	20 vs 20	25 vs 25
<i>Original dataset (no differences between patients and controls)</i>					
A. # of subsamples with > 10 miRNAs differentially expressed	145/10,000	127/10,000	93/10,000	36/10,000	9/10,000
B. Highest # of differentially expressed miRNAs (from 461) identified in one subsample	190	176	201	105	13
<i>Perturbed dataset (100 miRNAs set to differentially expressed between patients and controls)</i>					
C. Mean # of miRNAs differentially expressed between patient and control	47/100	73/100	85/100	91/100	93/100

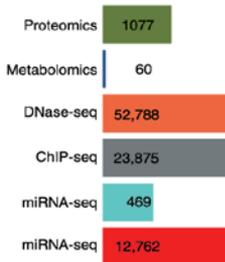


Kok, M. G. M., et al. "Small sample sizes in high-throughput miRNA screens: a common pitfall for the identification of miRNA biomarkers." *Biomolecular detection and quantification* 15 (2018)

Sample size estimation – multi-omics

Data: STATegra – a comprehensive multi-omics dataset of B-cell differentiation in

No. of features



Expected % of DE

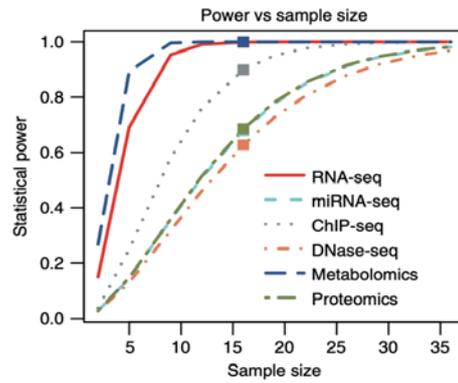
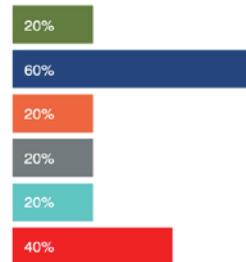


Table 1 MultiPower parameters and results from the STATegra pilot data.

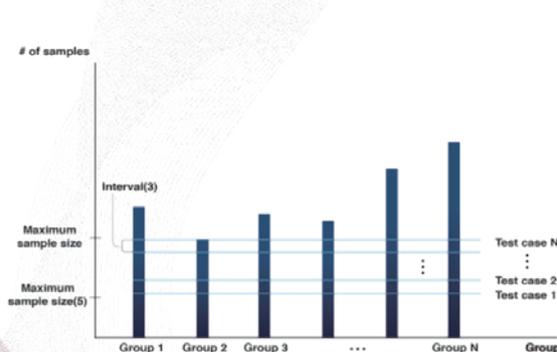
Omic	numFeat ^b	DEperc ^b	Delta ^a	Dispersion ^a	minSampleSize	optSampleSize	Power
RNA-seq	12,762	0.4	0.61	0.32	5	16	0.999
miRNA-seq	469	0.2	0.50	0.46	14	16	0.680
ChIP-seq	23,875	0.2	1.35	0.96	10	16	0.898
DNase-seq	52,788	0.2	0.51	0.49	16	16	0.627
Metabolomics	60	0.6	1.20	0.52	4	16	1.000
Proteomics	1077	0.2	1.16	1.05	14	16	0.685

MultiPower results were obtained for the same sample size in all technologies, a minimum power per omic of 0.6, a minimum average power of 0.8, and a Cohen's *d* of 0.8.
^anumFeat number of omic features, ^bDEperc expected proportion of DE features, ^adelta difference of means to be detected, ^adispersion pooled standard deviation, ^aminSampleSize sample size to achieve the minimum power per omic, ^aoptSampleSize optimal sample size for the experiment, ^apower power reached with the optimal sample size.
^aParameter estimated by MultiPower.
^bParameter provided by the user.

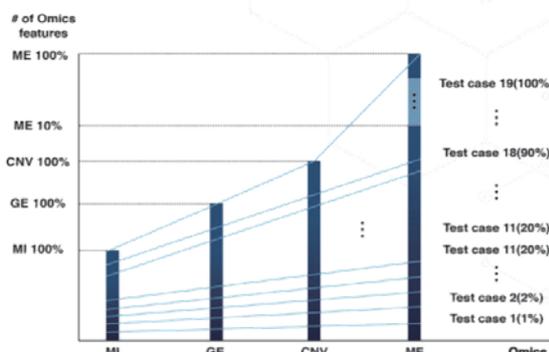
Tarazona, Sonia, et al. "Harmonization of quality metrics and power calculation in multi-omic studies." *Nature communications* 11.1 (2020)

Benchmark tests for the questions

- 7 benchmark tests were designed and analyzed
 - Sample size, feature numbers, preprocessing type, sample balance, noise ratio, biological groups and omics combinations



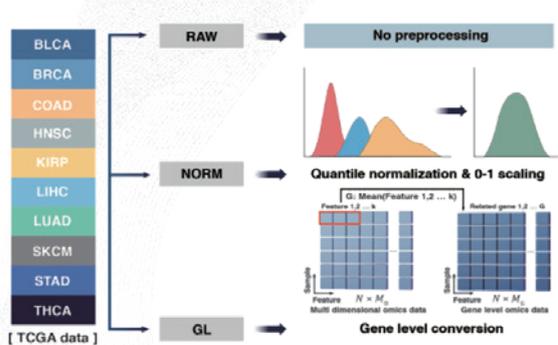
(a) Sample size test



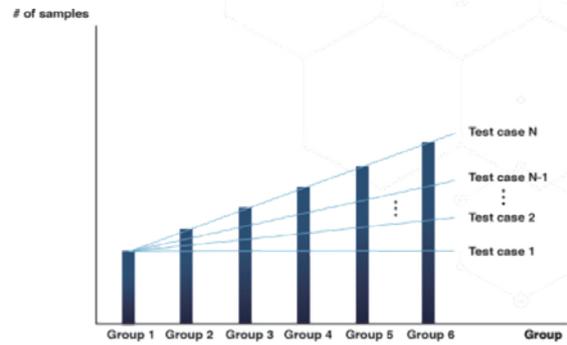
(b) Feature selection test

Benchmark tests for the questions

- 7 benchmark tests were designed and analyzed
 - Sample size, feature numbers, preprocessing type, sample balance, noise ratio, biological groups and omics combinations



(c) Preprocessing test

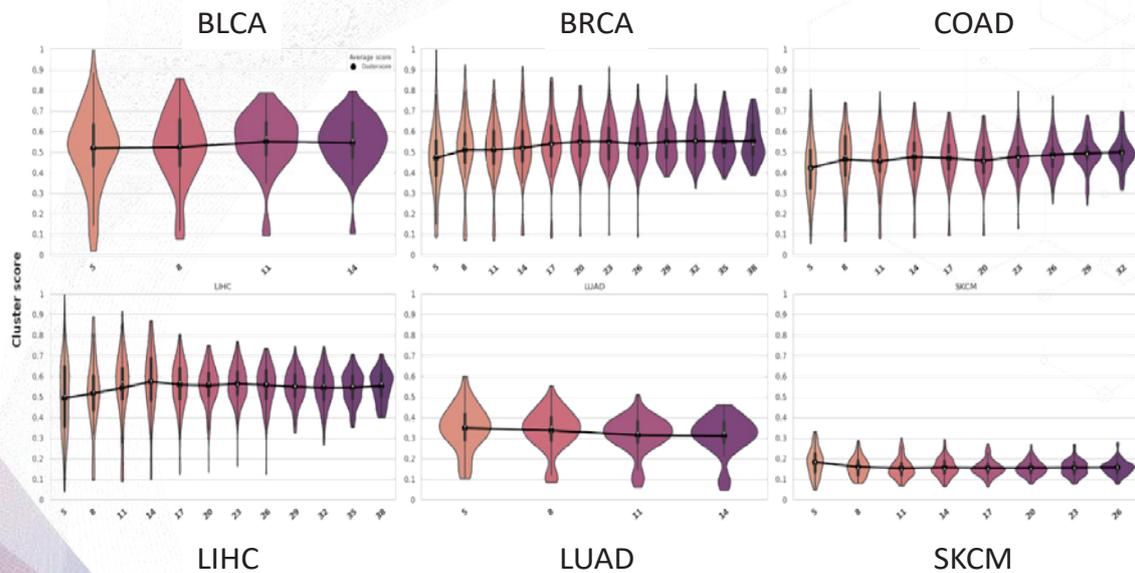


(d) Balance test

(paper in preparation...)

Benchmark tests | Preliminary results – Sample size

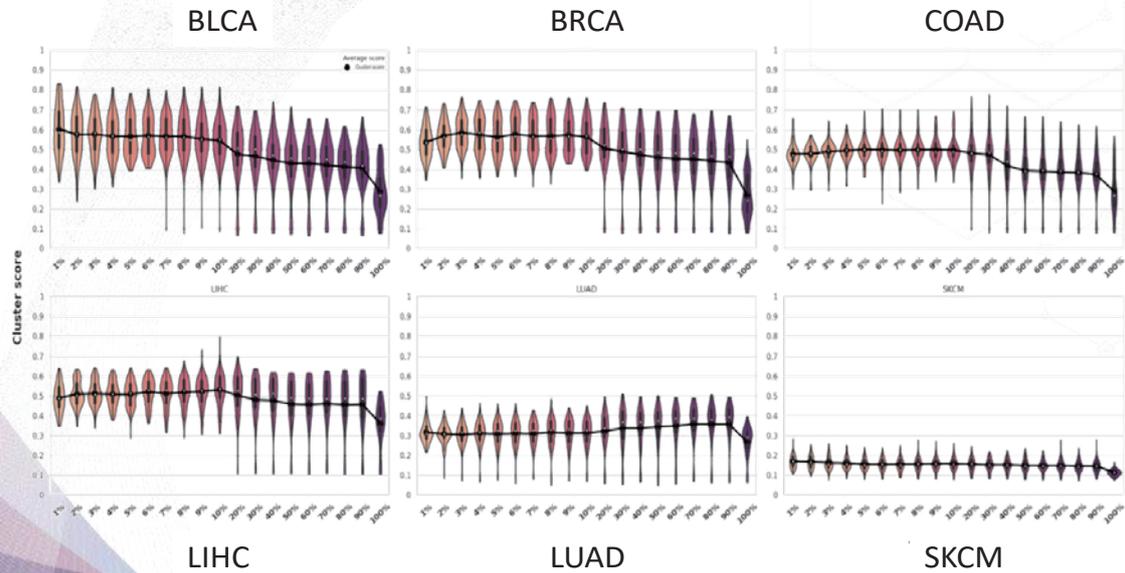
- Using 10 cancer types, the performance started to converge with sample $n > 60$



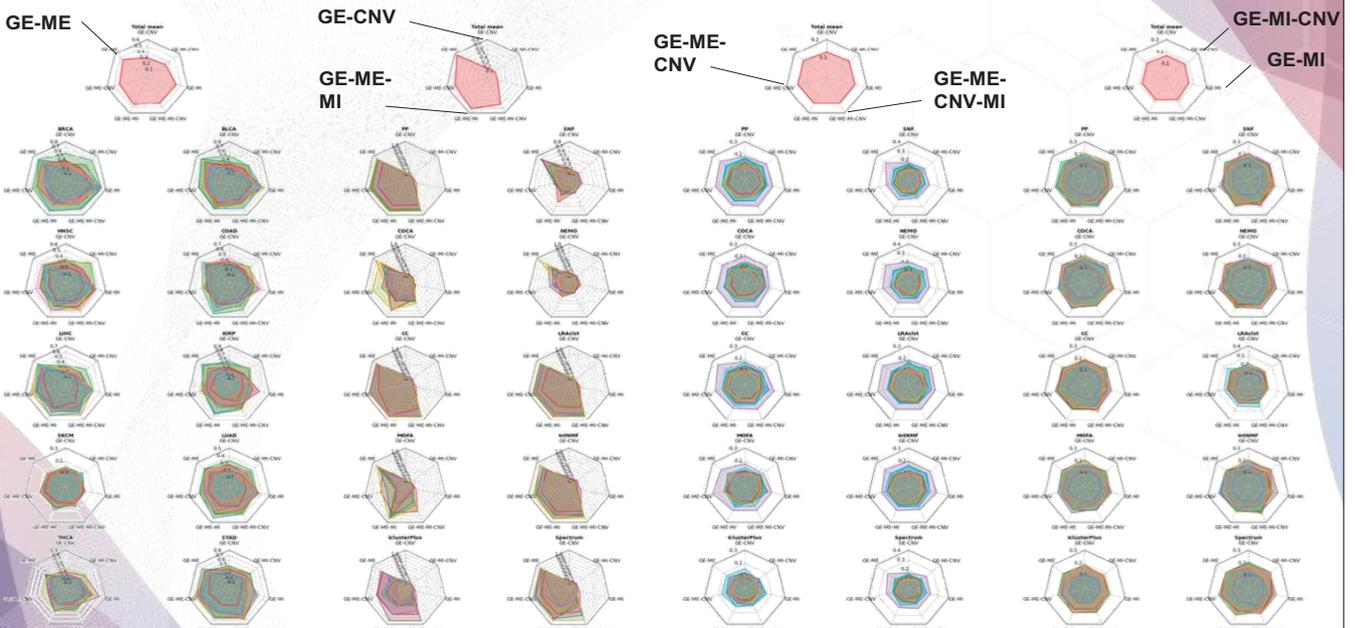
BLCA - Bladder Urothelial Carcinoma
 BRCA - Breast invasive carcinoma
 COAD - Colon adenocarcinoma
 LIHC - Liver hepatocellular carcinoma
 LUAD - Lung adenocarcinoma
 SKCM - Skin Cutaneous Melanoma

Benchmark tests | Preliminary results – Feature numbers

- The performance started to decline with # of features > 10~20%
 - Feature selection is important



Benchmark tests | Preliminary results – Omics combinations



(a) Subtype

(b) Gender

(c) Pathological stage

(d) Age

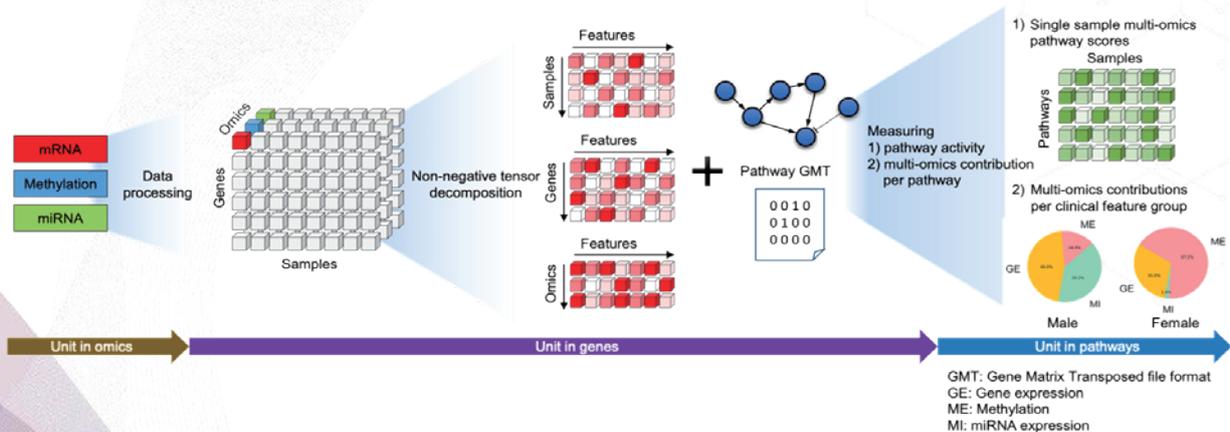
Background | Pathway Analysis

- While there are a number of MO analysis tools, most output a list of genes or accuracy score from clustering or classification results
 - biological interpretation needs further work on the given results
- A simple list of genes may not be enough for such purpose
 - Especially, since the genes are selected from MO data (i.e., if genes selected from GE, we can perform GSVA or SSSGEA)
- **A list of multi-omics pathway analysis methods**

Method	Supporting omics	Analysis target	Output
MOPA	multi-omics	Single sample	Scoring matrix
MOGSA	multi-omics	Single sample	Scoring matrix
ActivePathways	multi-omics	Group	p-value
multiGSEA	multi-omics	Group	p-value
GSVA	single-omics	Single sample	Scoring matrix
GSEA	single-omics	Group	Scoring matrix
ssGSEA	single-omics	Single sample	Scoring matrix
z-score	single-omics	Single sample	Scoring matrix

Methods | MOPA

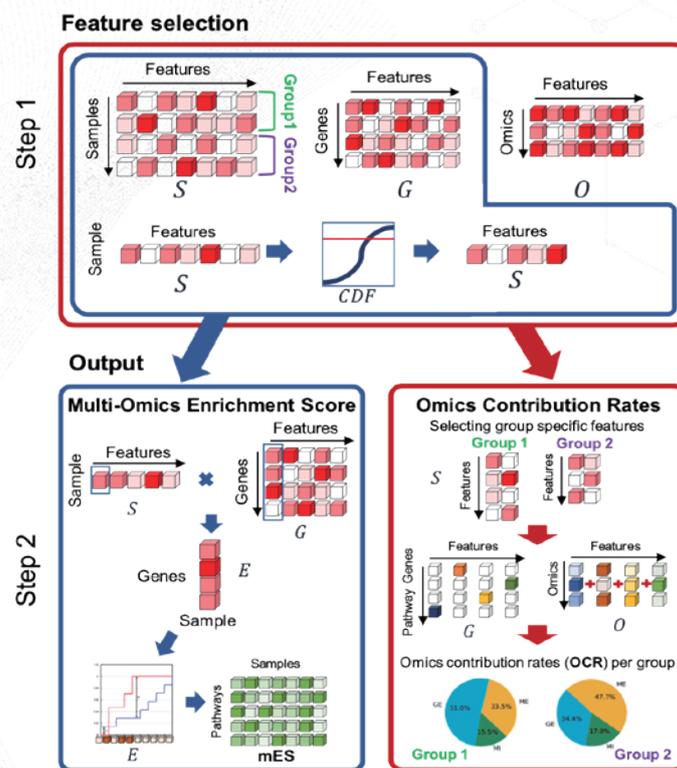
- MOPA is a tool that scores pathway activity based on MO data for each sample and each pathway
- The framework is very similar to GSVA or SSGEA but extended to consider MO data



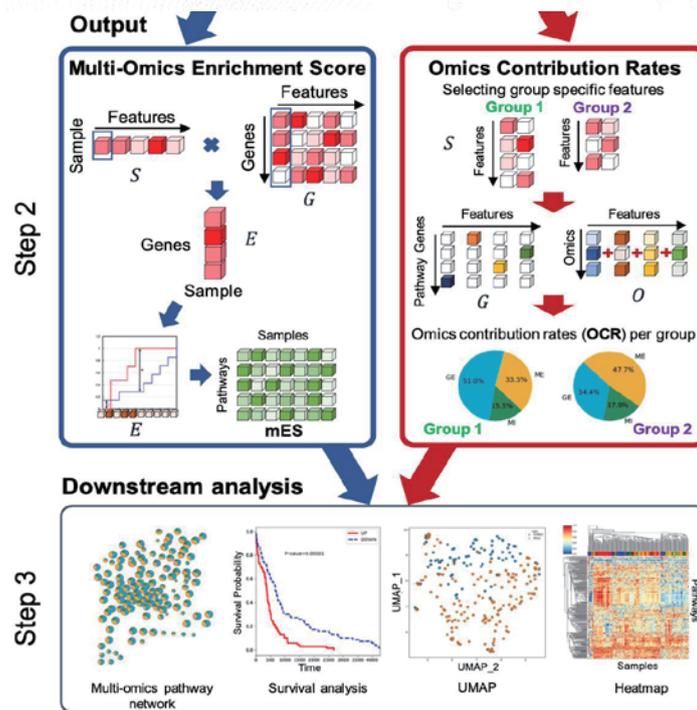
Objective | mES and OCR metrics

- The major objective was to provide metrics that can interpret the pathway results in context of multi-omics data
- For such matter, the **multi-omics Enrichment Score (mES)** and **OCR (Omics Contribution Rate)** were developed

Methods | MOPA Workflow (Step 1~2)

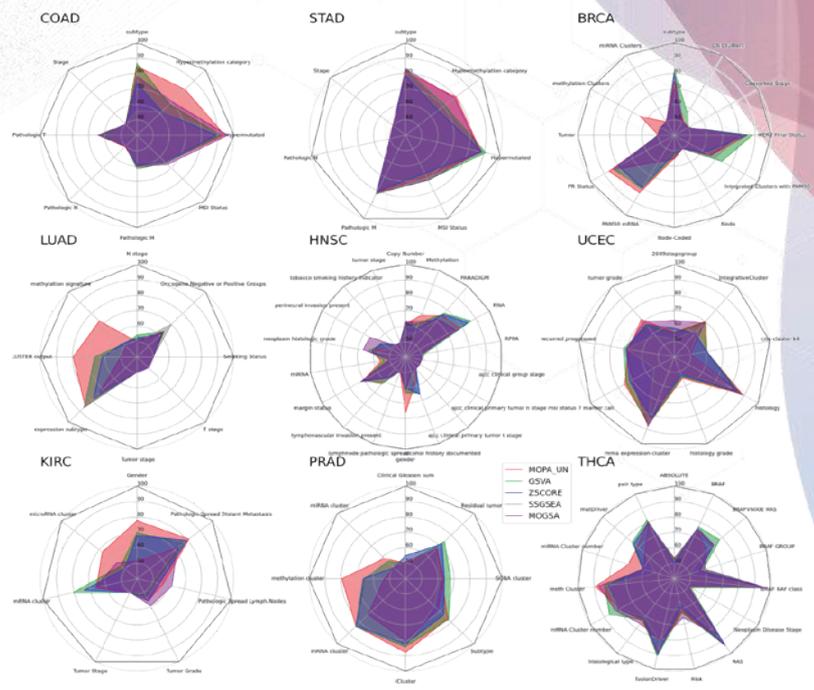


Methods | MOPA Workflow (Step 2~3)

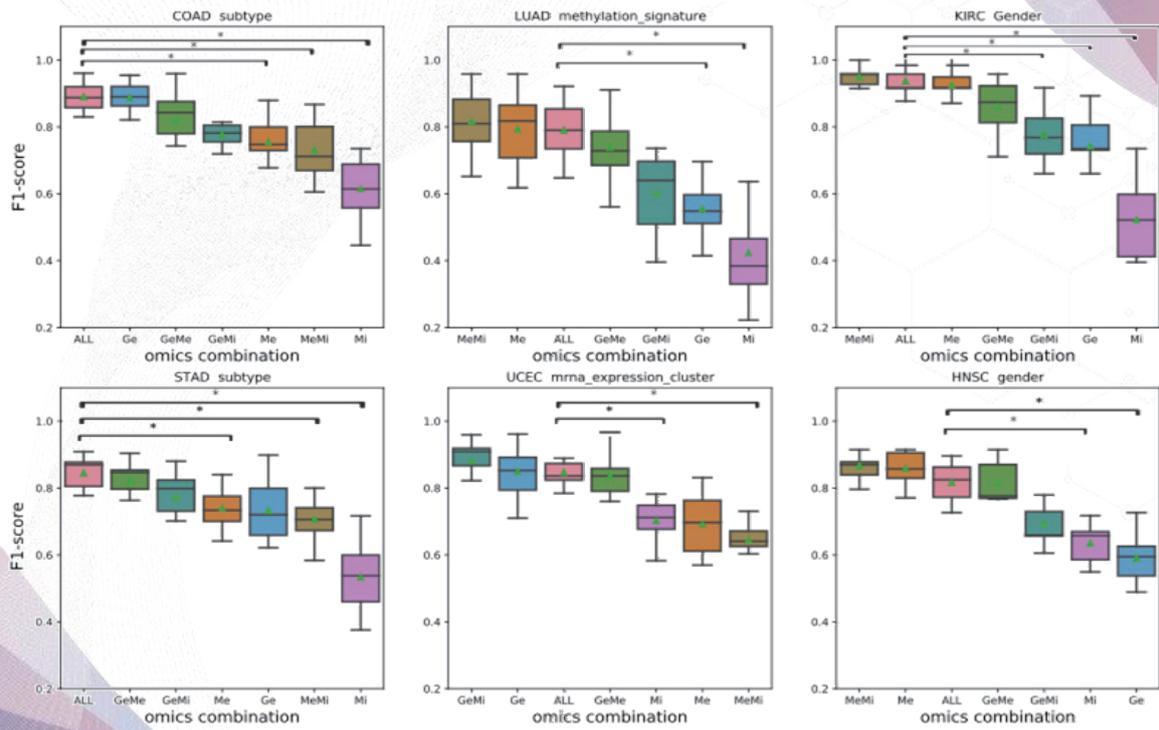


Results | MOPA on Pan-cancer data

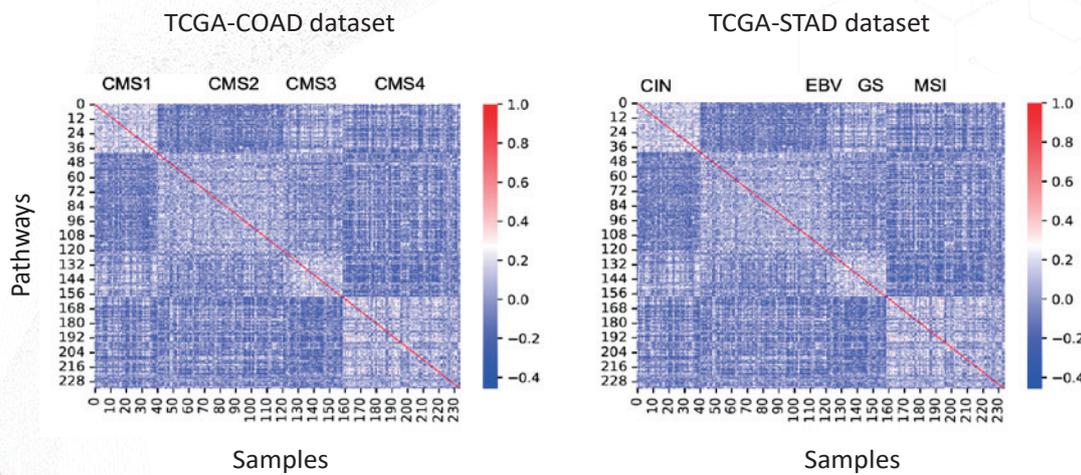
- MOPA was used to analyze 9 cancer types including 95 clinical features (e.g, subtype, cancer stage, gender)
- Some clinical features are well explained while some showed poor classification performance
- In the majority features, MOPA showed higher or equal performance to competing methods



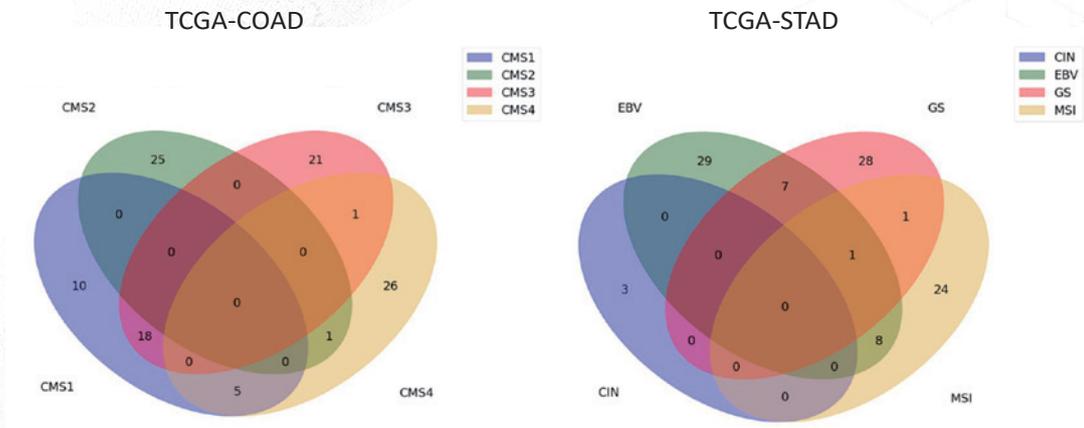
Results | Multi-omics combinations



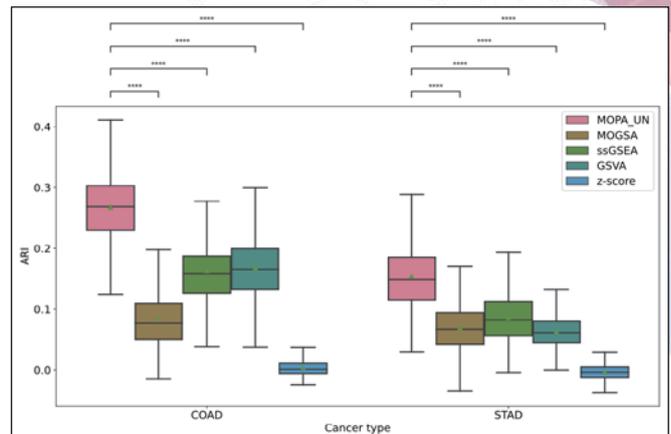
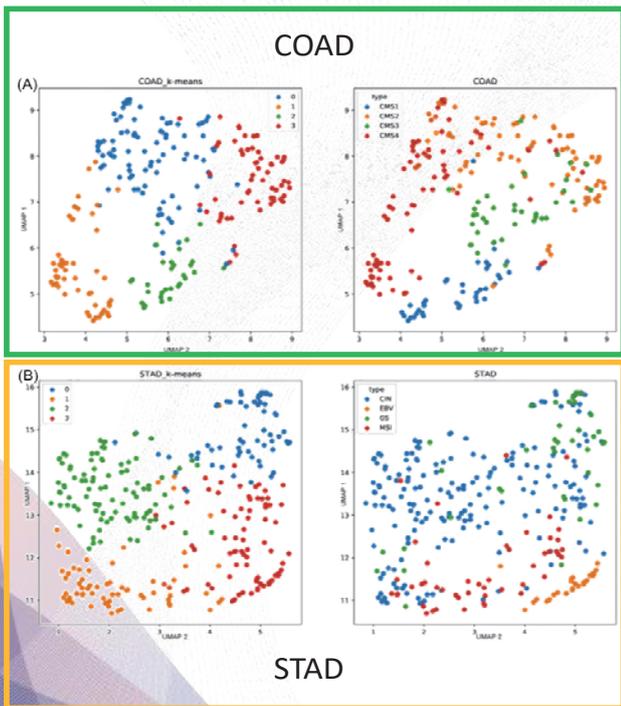
Results | Features detected by MOPA



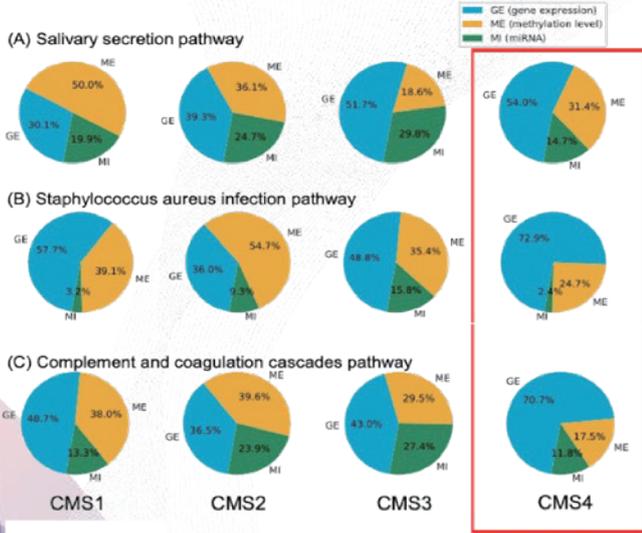
Results | Features detected by MOPA



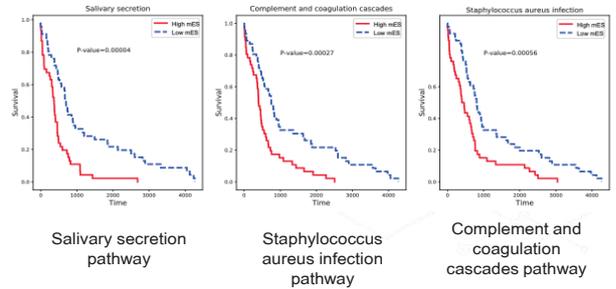
Results | Subtype clustering using *mES*



Results | Use Case Study – COAD

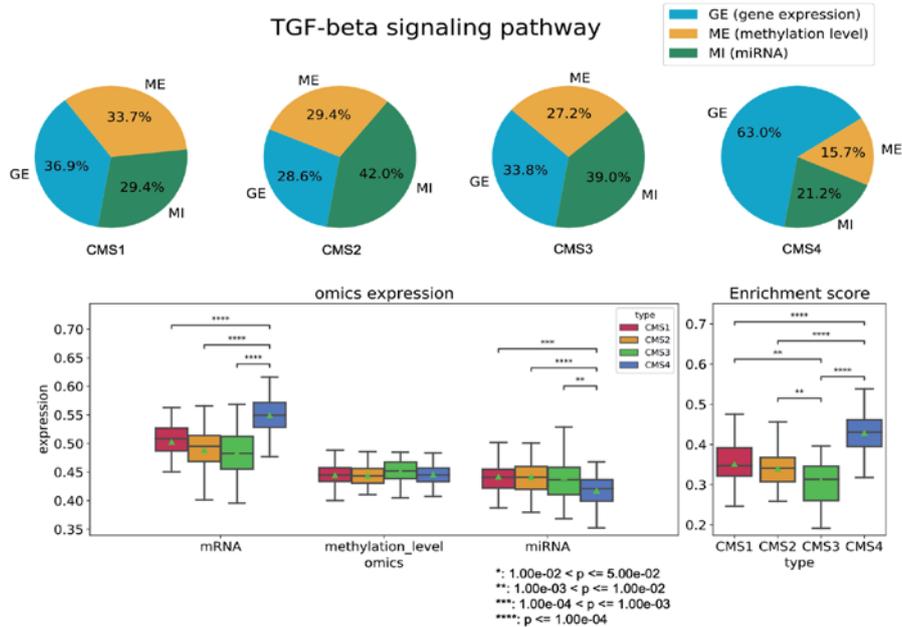


Survival plot



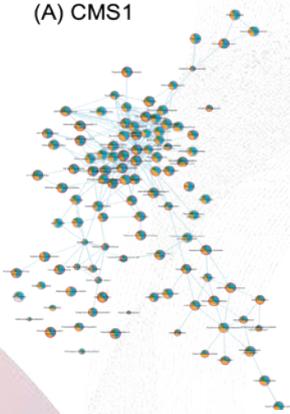
Group with high mES showed significantly lower survival for the three pathways.

Results | Use Case Study – COAD

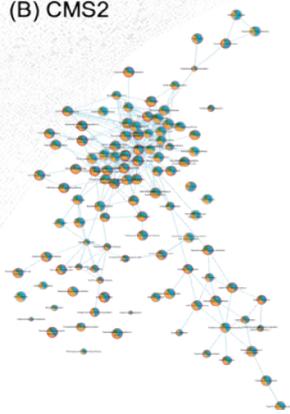


Results | Use Case Study – COAD (MO pathway network)

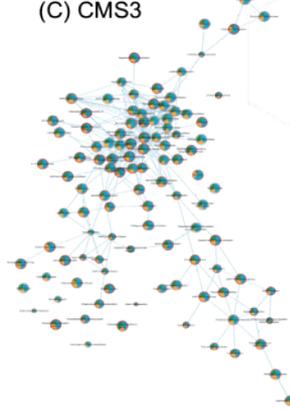
(A) CMS1



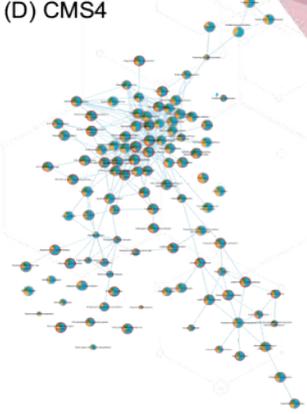
(B) CMS2



(C) CMS3

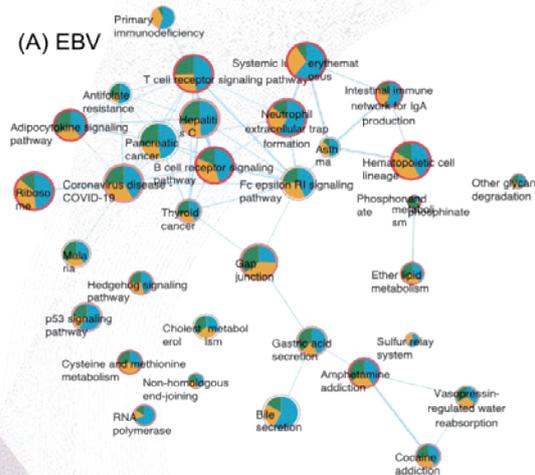


(D) CMS4



Results | Use Case Study – STAD

(A) EBV



(B) GS



수고하셨습니다.

감사합니다!