KSBi-BIML 2023

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

생물정보학&머신러닝 워크샵(온라인)

Integrative analysis of multi-omics data

정인욱 _ 경북대학교

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

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
Phone Number	053-950-5552

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, Joungmin 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

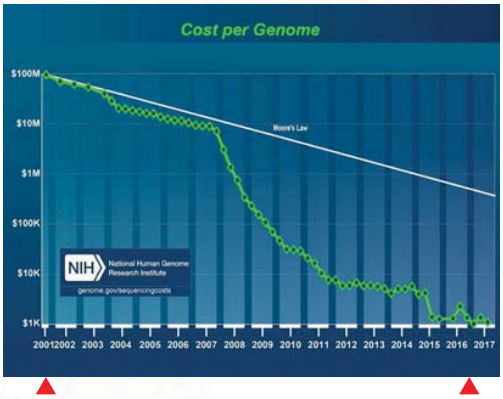
Contents

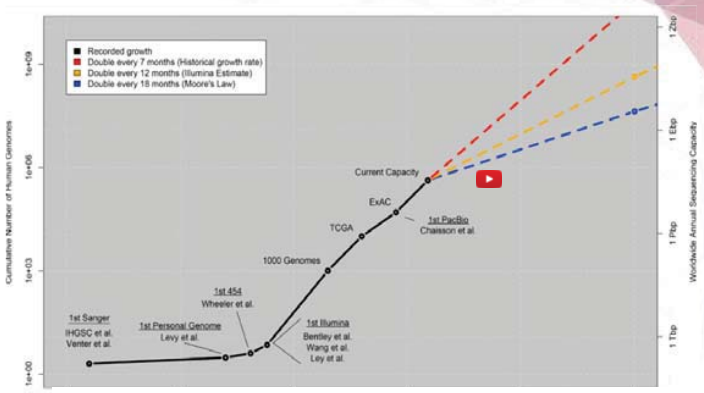
1. Multi-omics overview

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

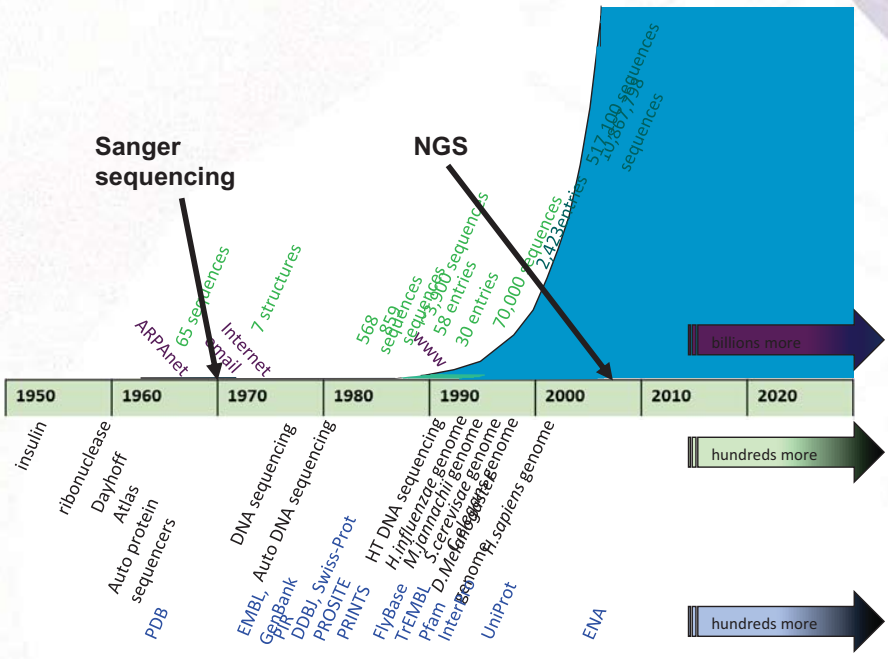
Bio Data| Trend of Cost & Volume





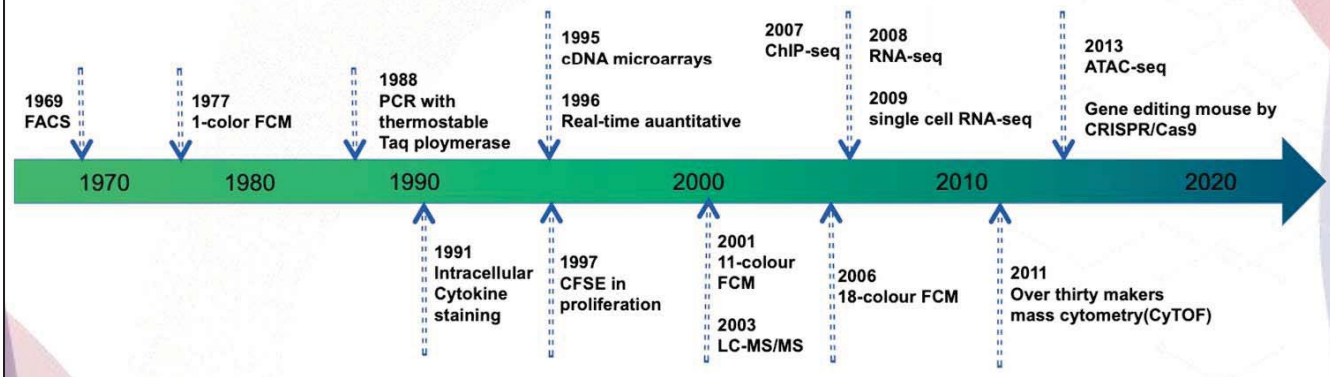
$100M/Sample **$1000/Sample** **2000** **2010** **2018** **2025**

Timeline of Biological data increase



Source from Teresa K.Attwood, University of Manchester

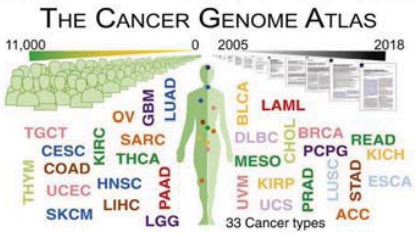
Timeline of multi-omics technologies

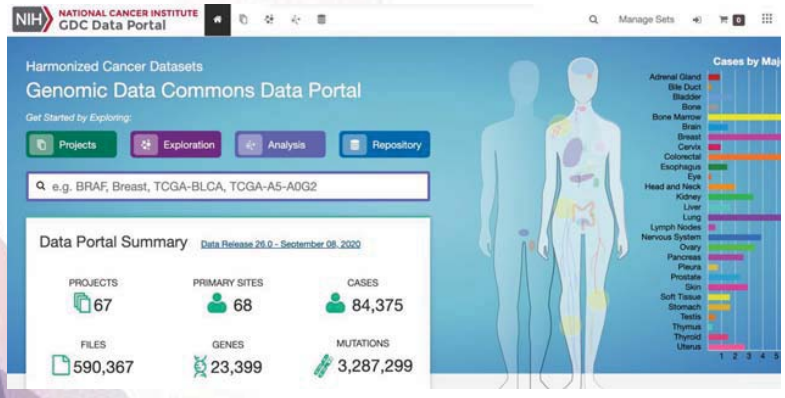


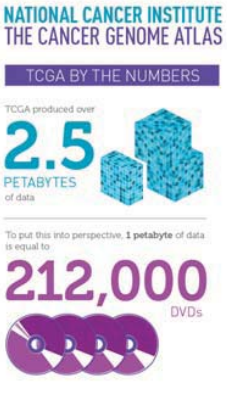
Kang Ning & Yuxue Li , Introduction to Multi-Omics, Translational Bioinformatics book series 2023

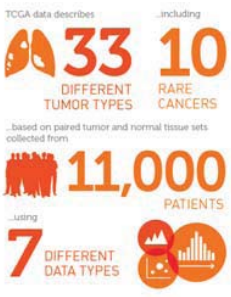
A Big Data Example

- **The Cancer Genome Atlas (TCGA)**









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

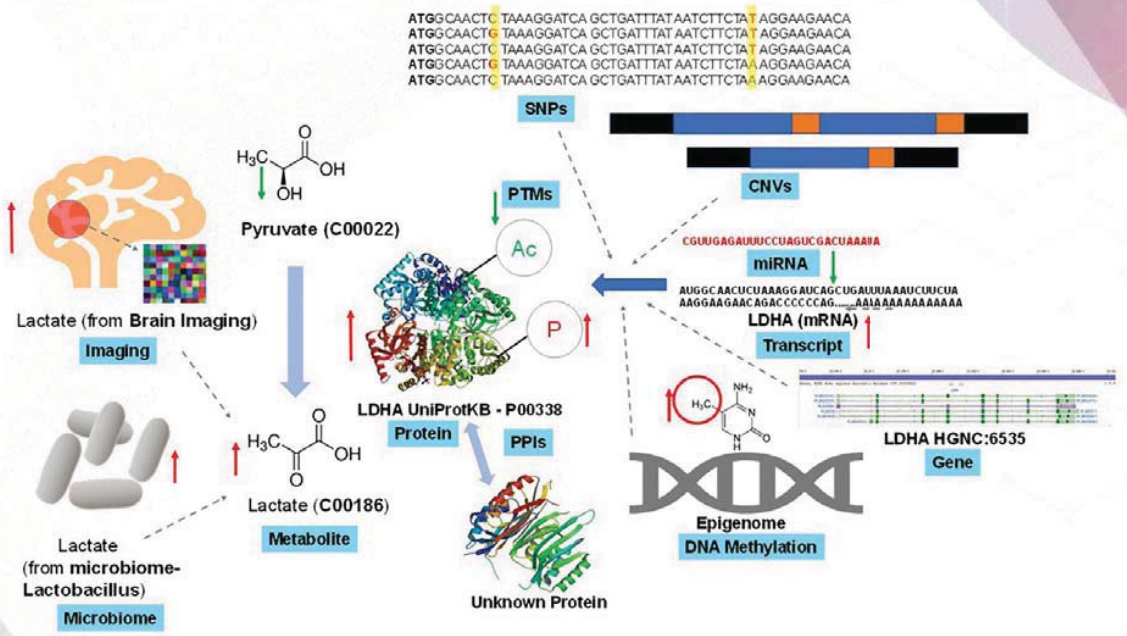
1) Search for patterns

2) Create hypothesis

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

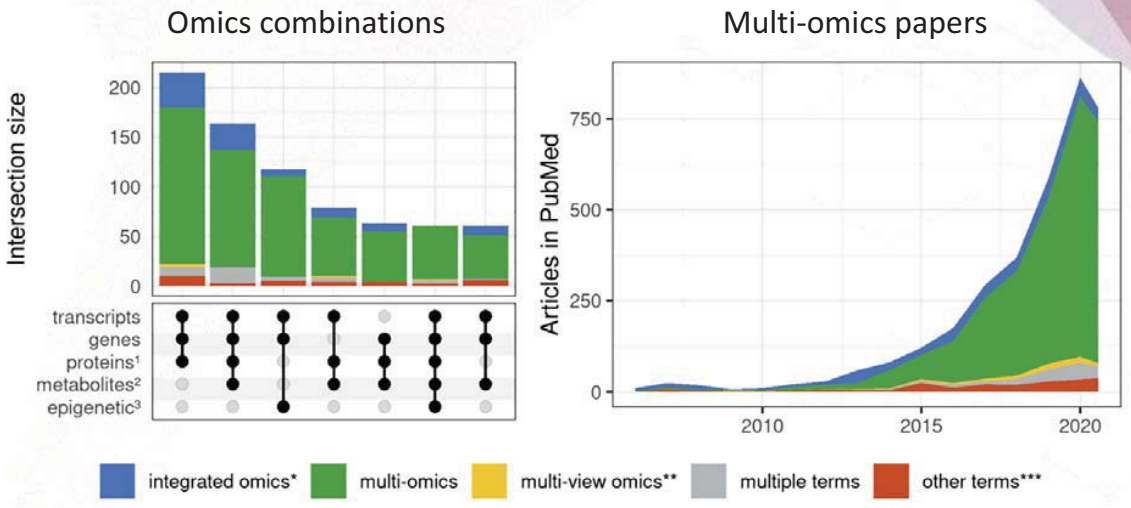
1. Multi-omics BIG DATA

Multi-omics interconnections



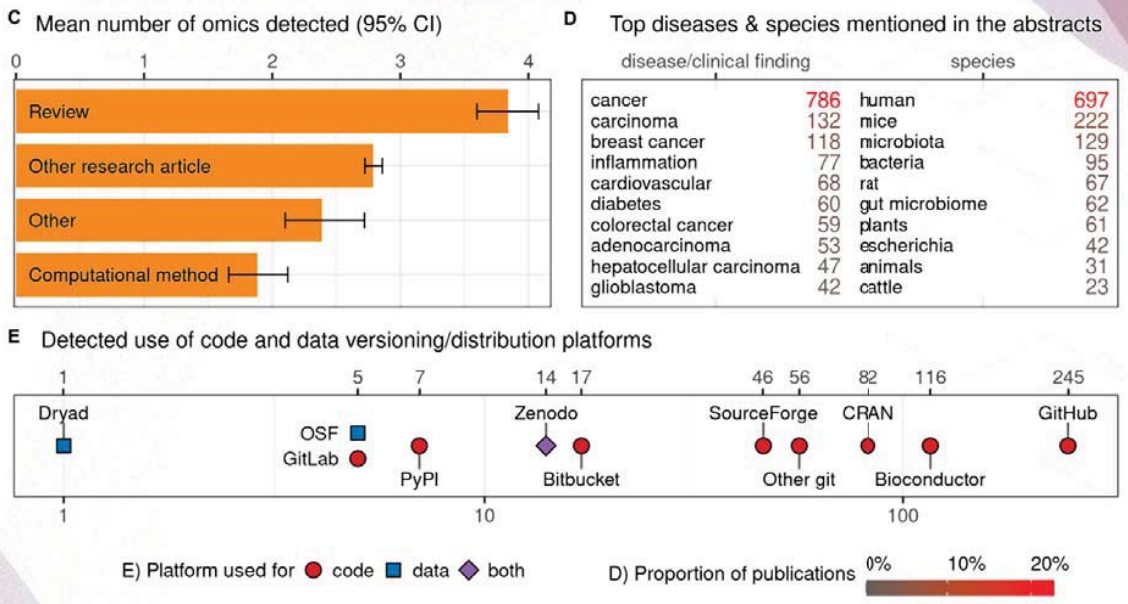
Krassowski, Michal, et al. "State of the field in multi-omics research: From computational needs to data mining and sharing." *Frontiers in Genetics* 11 (2020)

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

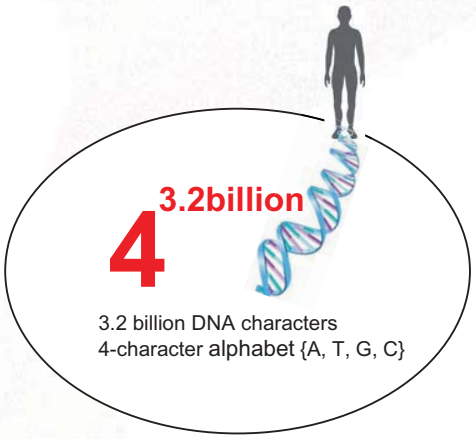


Krassowski, Michal, et al. "State of the field in multi-omics research: From computational needs to data mining and sharing." Frontiers in Genetics 11 (2020)

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



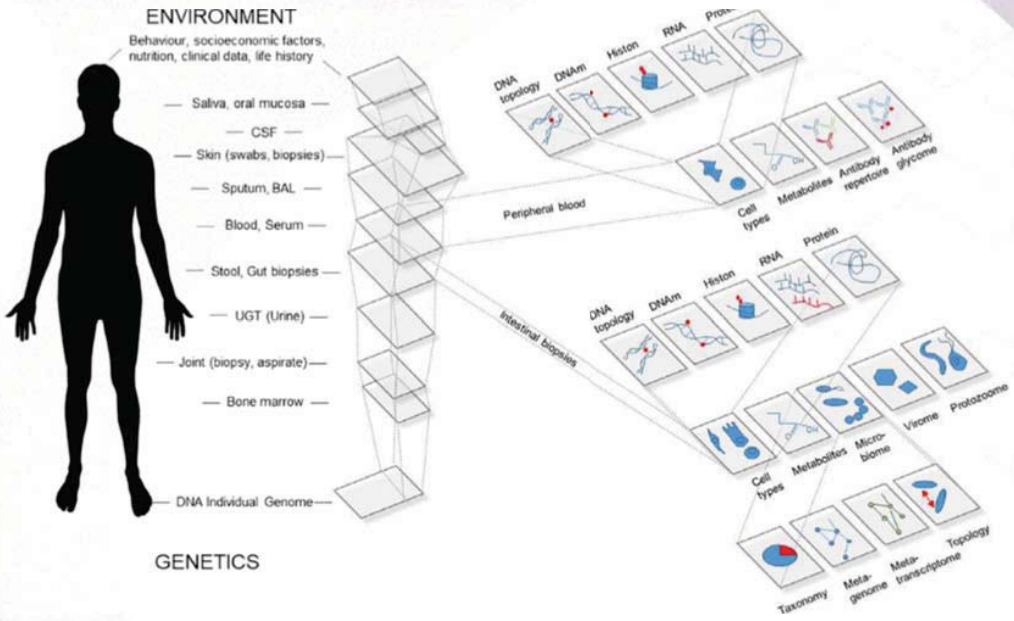
Search space

Low sample,
High dimension problem

1 sample

Genes
B2M
HNRPLL
HPRT
PSMB2
PSMB4
PPIA
PRPS1
PRPS1L1
PRPS1L3
PRPS2
…
35K genes

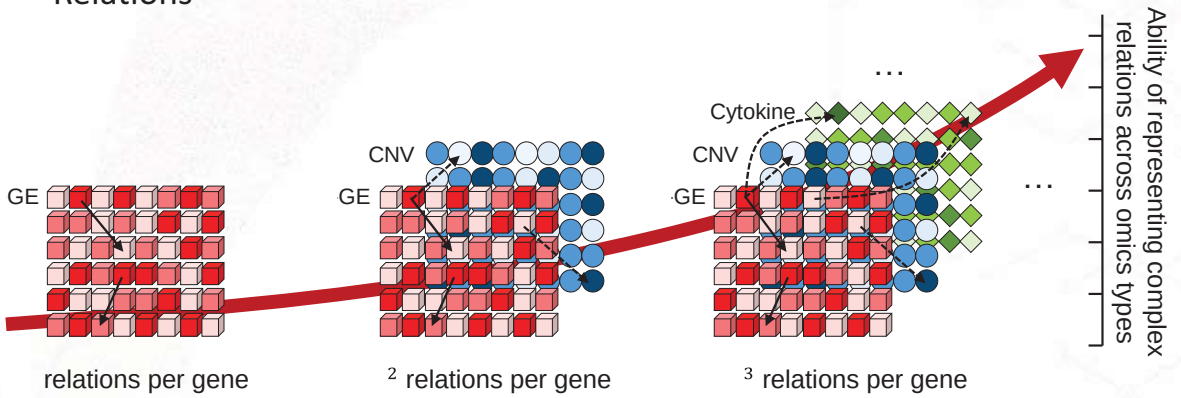
Challenges| Multi-modal data adds extra dimensions



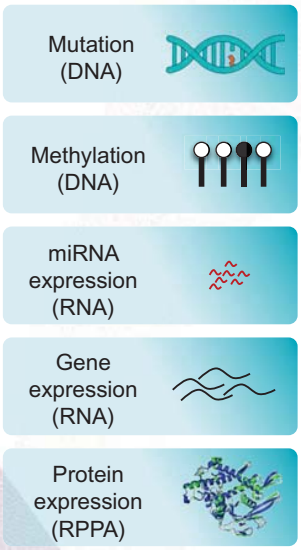
https://www.mynewgut.eu/sites/default/files/3-Hadrich.pdf

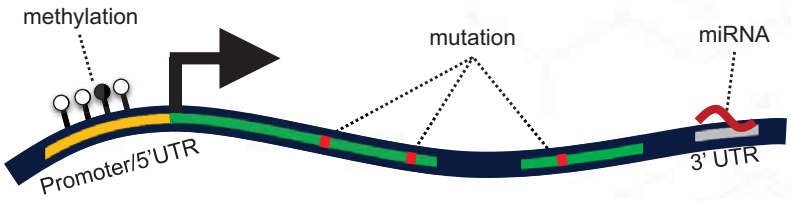
Expectations from multi-omics data

- Interpretability
- Relations



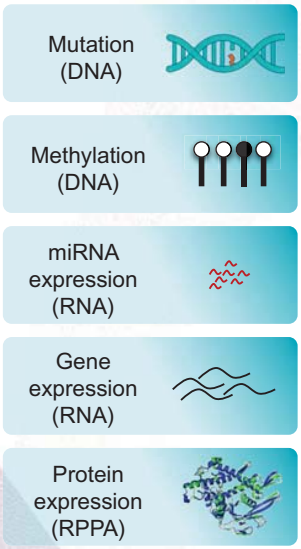
Challenges| Multi-modal data adds extra dimensions

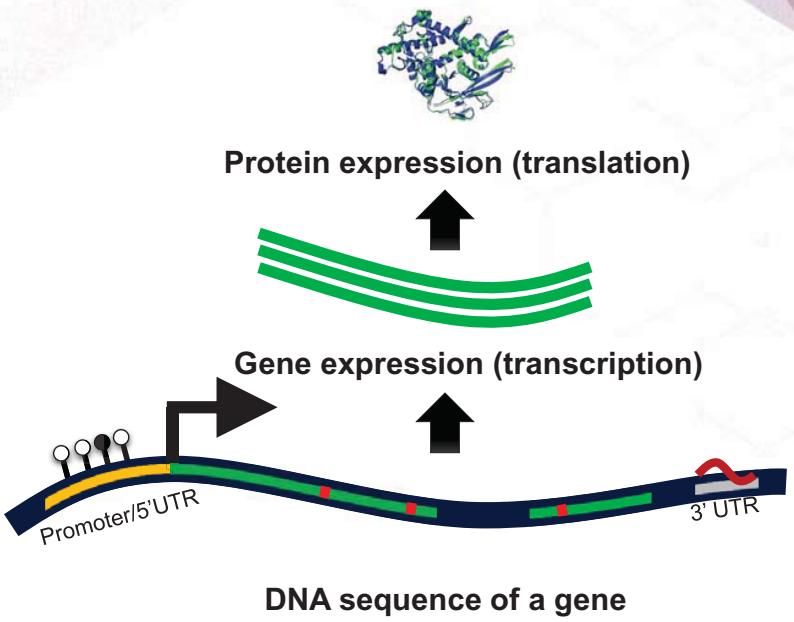




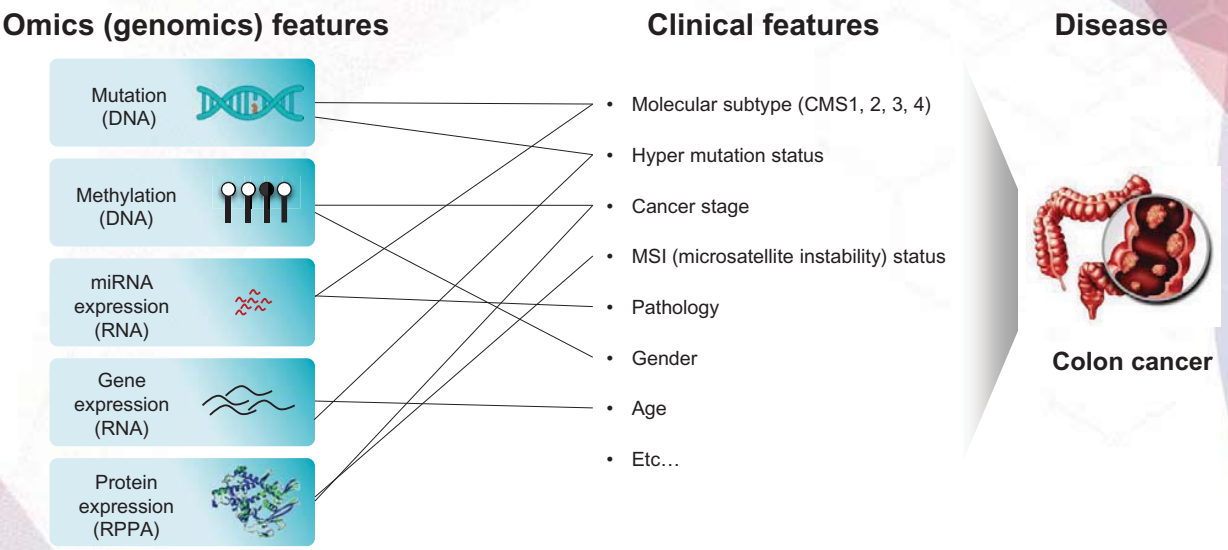
DNA sequence of a gene

Challenges| Multi-modal data adds extra dimensions





Challenges| Searching for explainable omics causality



Challenges| Searching for explainable omics causality

Omics (genomics) features

associations?

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

associations?

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

Omics features

associations?

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

Radiomics 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

Some questions to ponder on

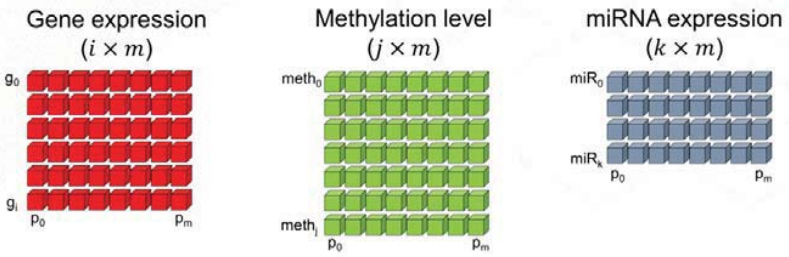
- Is multi-omics better than single-omics?
  - More data = higher quality of result?
- How much (at least) data do we need?
- What type of omics associate well together?
- What types of clinical features are explainable by MO?

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

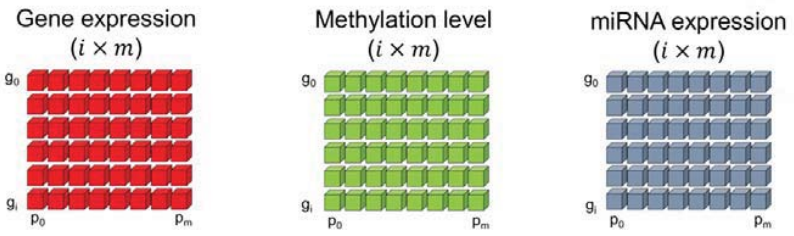
Background| Two ways of integrating multi-omics data

Multi-dimensional integration

(most methods)



Multi-staged integration

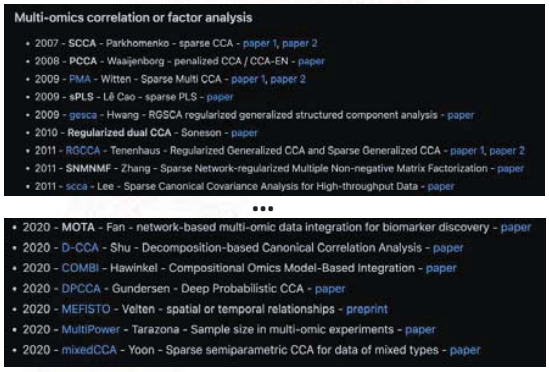


Ritchie, Marylyn D., et al. "Methods of integrating data to uncover genotype–phenotype interactions." Nature Reviews Genetics 16.2 (2015)

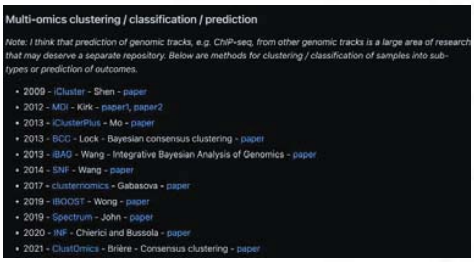
MO Analysis Methods |

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

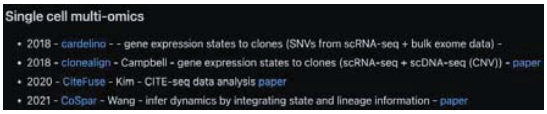
Factor analysis



Clustering



Single-cell



Network analysis

reviews and many more…

MO Factor analysis

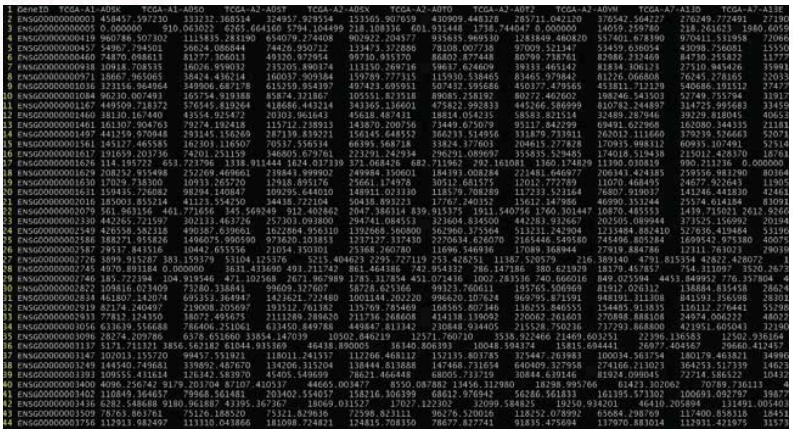
- **Factor analysis** focuses on reducing dimensions for representative learning on a more simple or lower space than the original data
- It's advantage lies in finding strong signals and alleviate interpretation of the result
- In addition to factor analysis, multi-omics is often analyzed using multiview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

- But why reduce dimensions?

MO data characteristics

- MO data is highly complex, large and and heterogenous

Single-omics : a dataset of only bulk RNA-seq samples

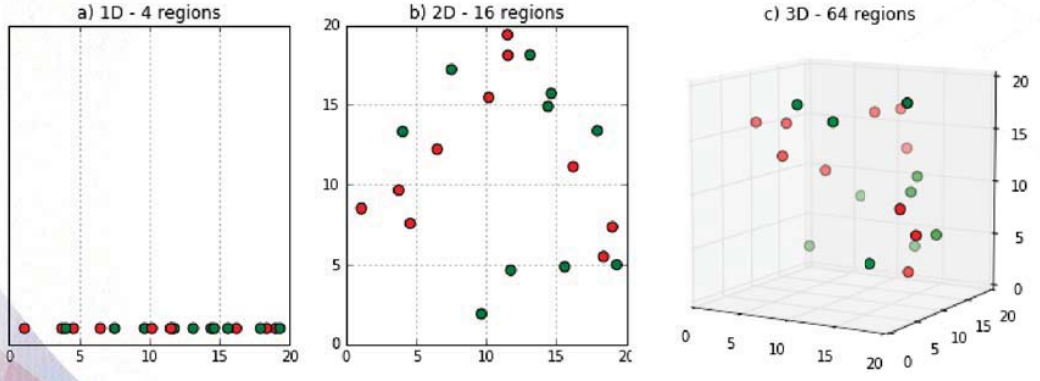


> 15K genes

> 600 patients

The Curse of Dimensionality

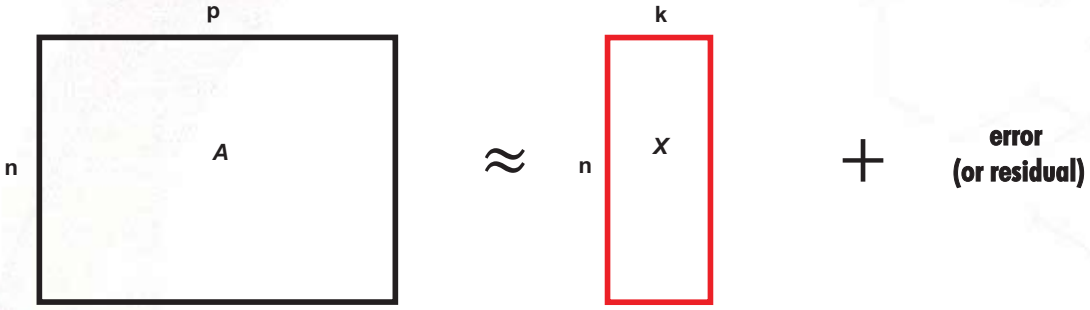
- a) shows 20 data points on a 1 dimensional plane
- Adding a dimension causes the amount of data required to represent it to not double but **square!**
- So, with 2 dimensions, we will need a space of at least 20^2 to represent the 20 data points
- With 3 dimensions, $20^3 = 8{,}000$ is at least needed to represent 20 data points!



A brief overview of factor analysis

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

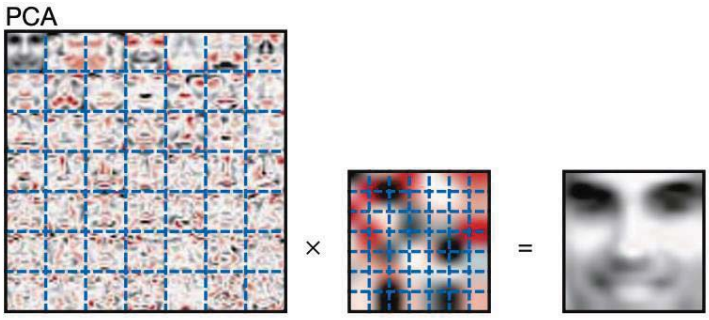
- PCA concept



Original data

Approximated data (compressed)

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

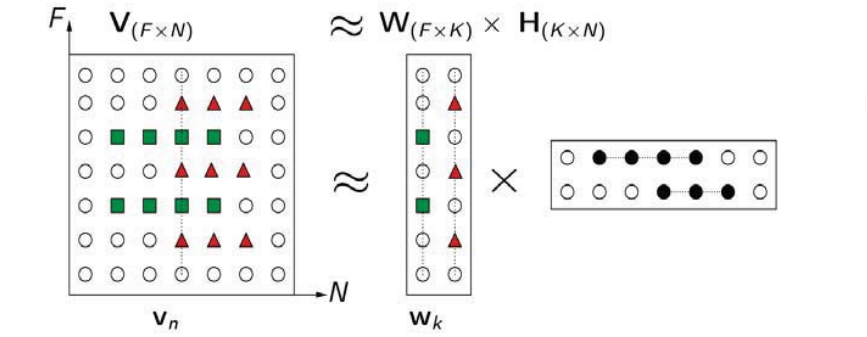
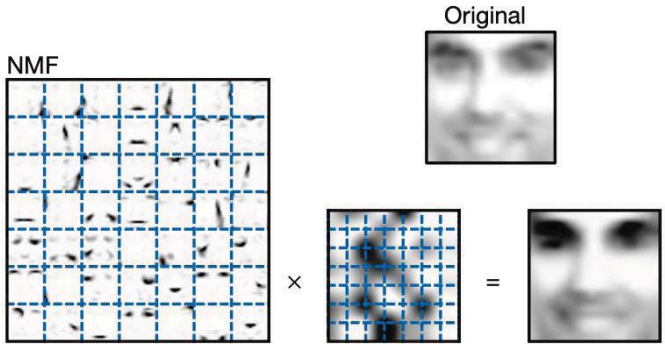


Illustration by C. Févotte

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)

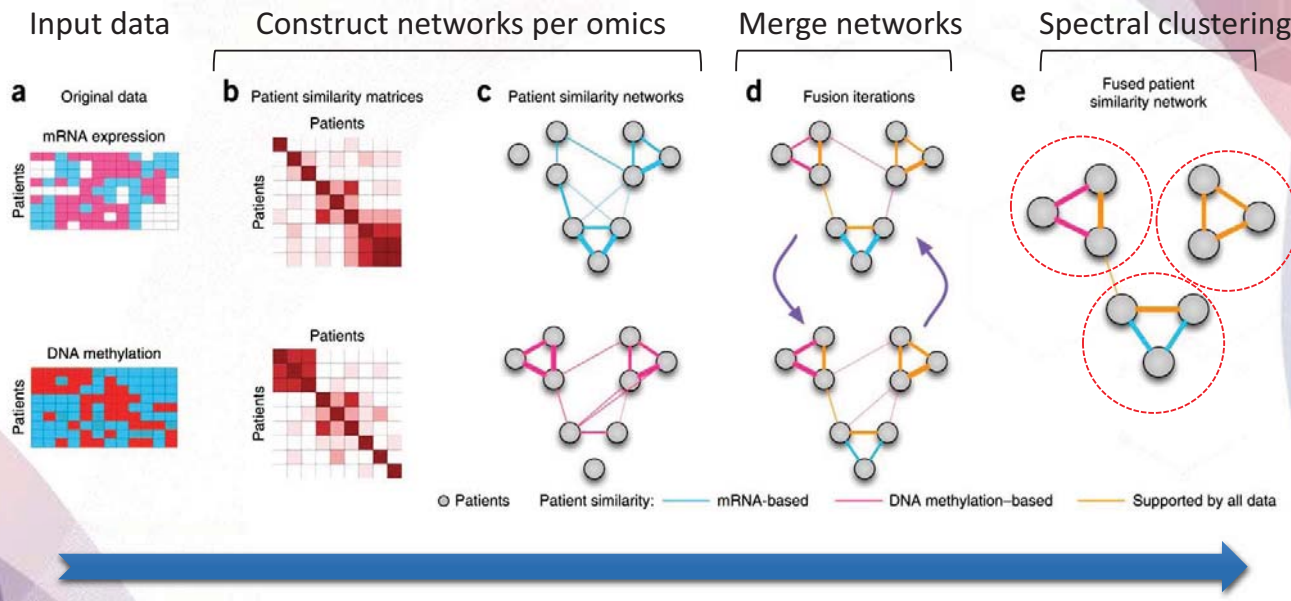
Methods | SNF – Similarity Network Fusion

- 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

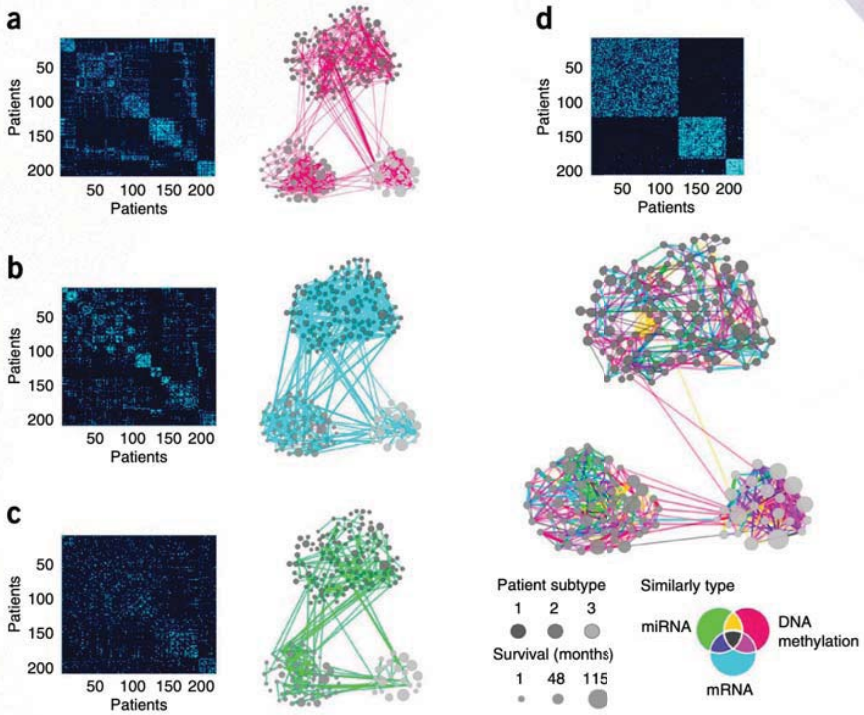


Methods | SNF – Results

Methylation

mRNA

miRNA



Methods | JointNMF

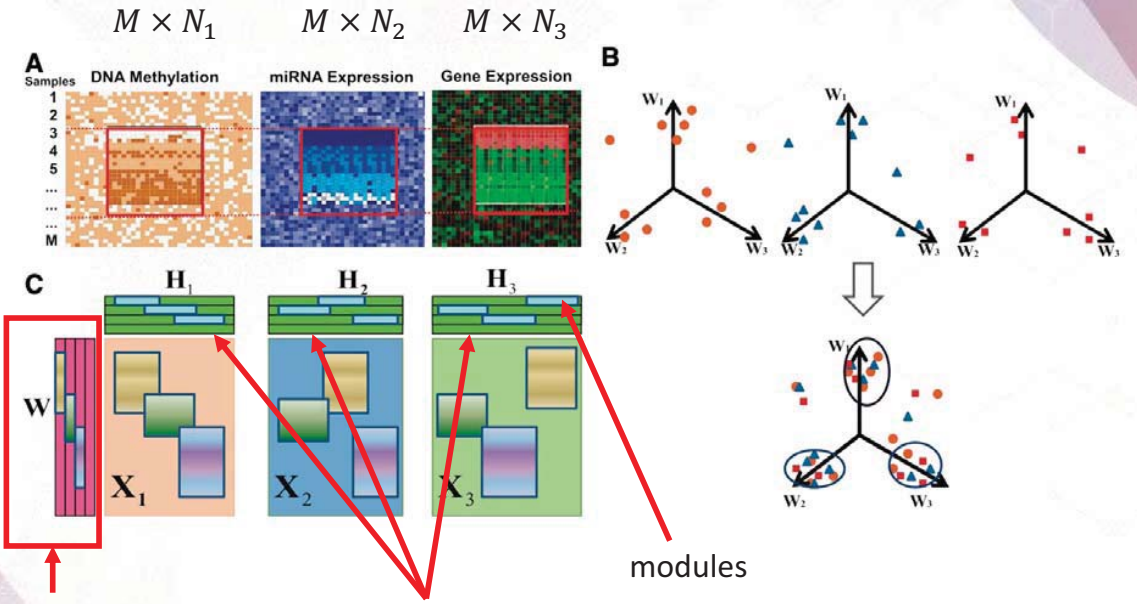
- Based on the Non-negative Matrix Factorization (NMF) method, JointNMF discovers modules (or ranks) that show association between different omics using ovarian cancer samples
- Omics: mRNA, methylation, miRNA
- Samples: 385 ovarian cancer patients (TCGA)

- Advantage:
  - Reports a set of important omics-features
- Disadvantage:
  - Number of ranks is difficult to determine
  - Can become slow and require large memory with large samples and many features
  - May not work well with small no. of samples (constrained by the ranks)

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 – concept

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



Common basis matrix

$W = M \times K\ (rank)$

Omics specific coefficient matrices

$X_1 \sim W \times H_1 \quad H_1 = K \times N_1$

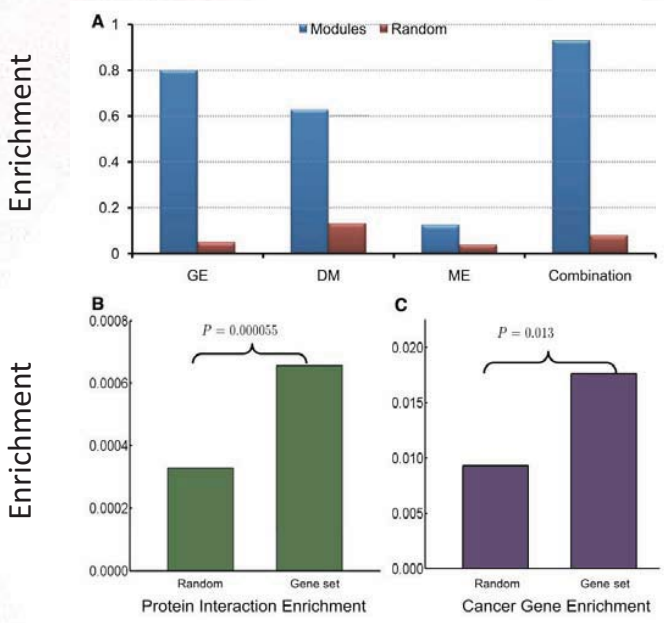
$X_2 \sim W \times H_2 \quad \ldots$

$X_3 \sim W \times H_3$

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)



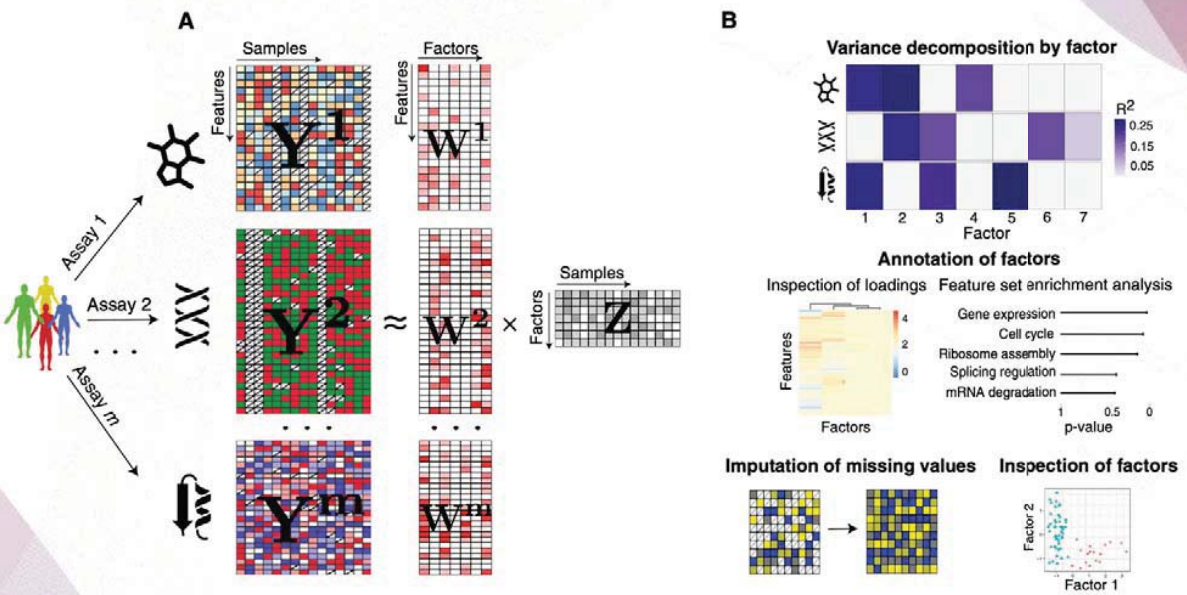
Methods | MOFA – Multi-Omics Factor Analysis

- 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)

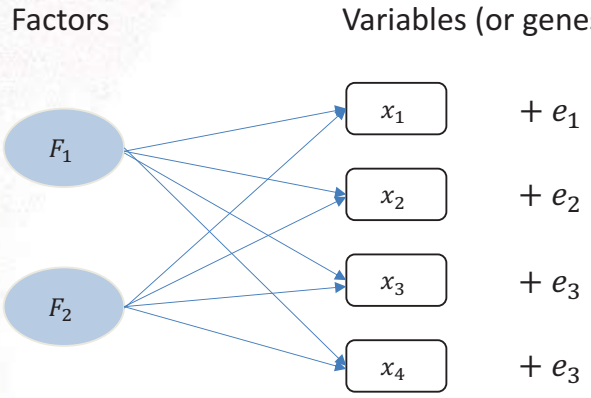
Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." *Molecular systems biology* 14.6 (2018): e8124.

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

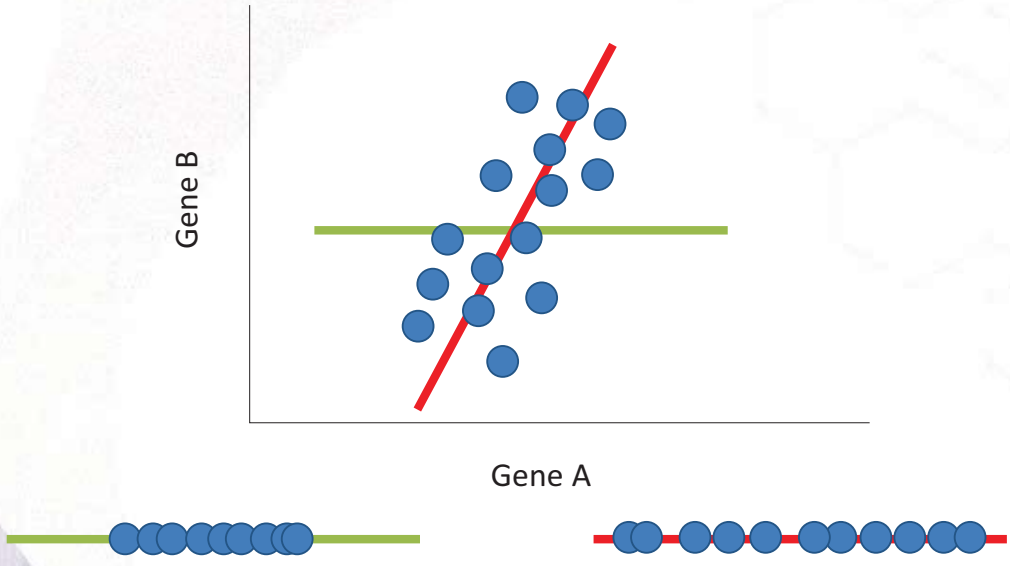


$$x_1 = b_{11} \times F_1 + b_{12} \times F_2 + e_1$$

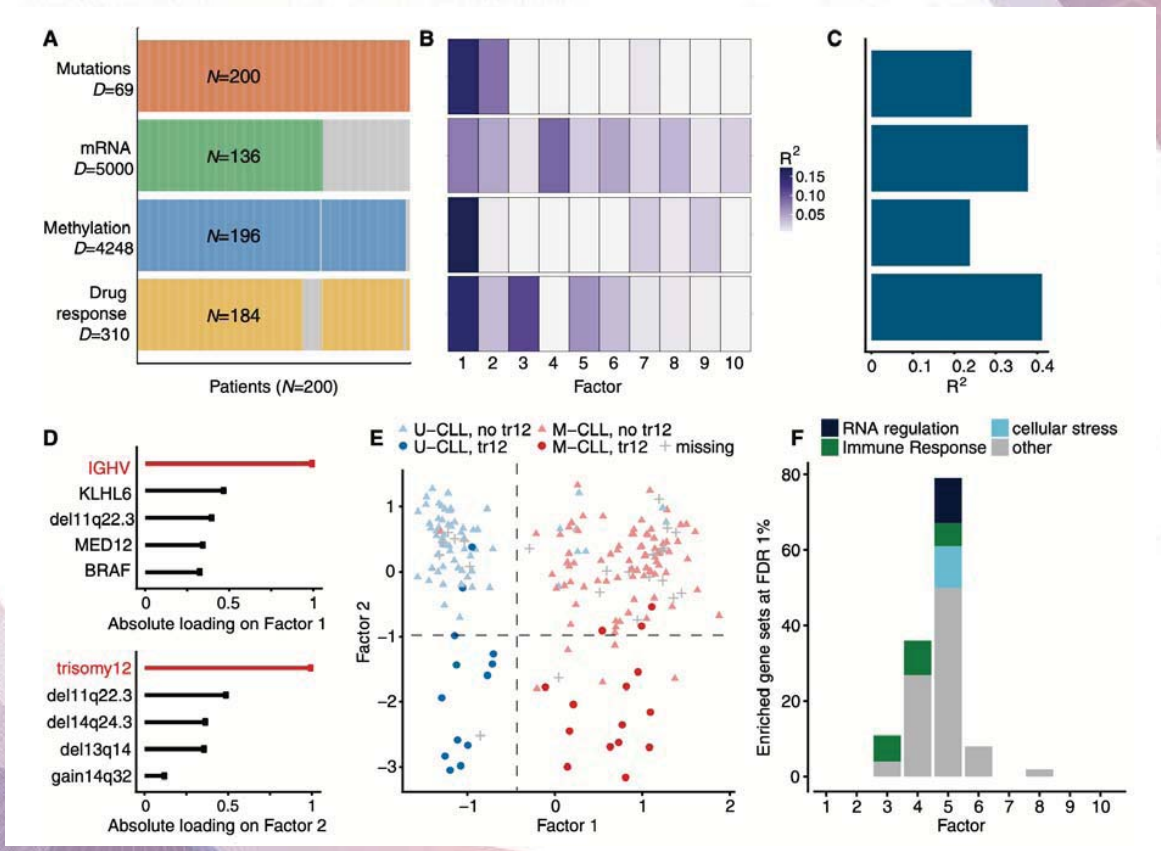
$$F_1 = w_{11} \times x_1 + w_{12} \times x_2 + w_{13} \times x_3 + w_{14} \times x_4$$

Brief overview of PCA

- Finds the line that maximizes variance in the data

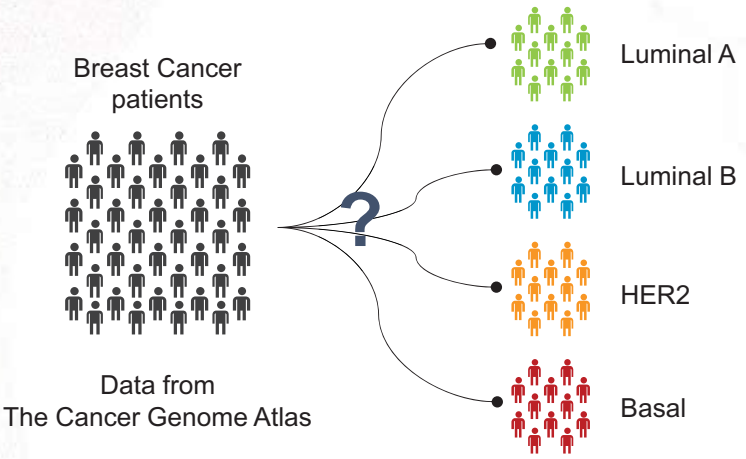


Methods | MOFA – Results



Multi-omics research

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

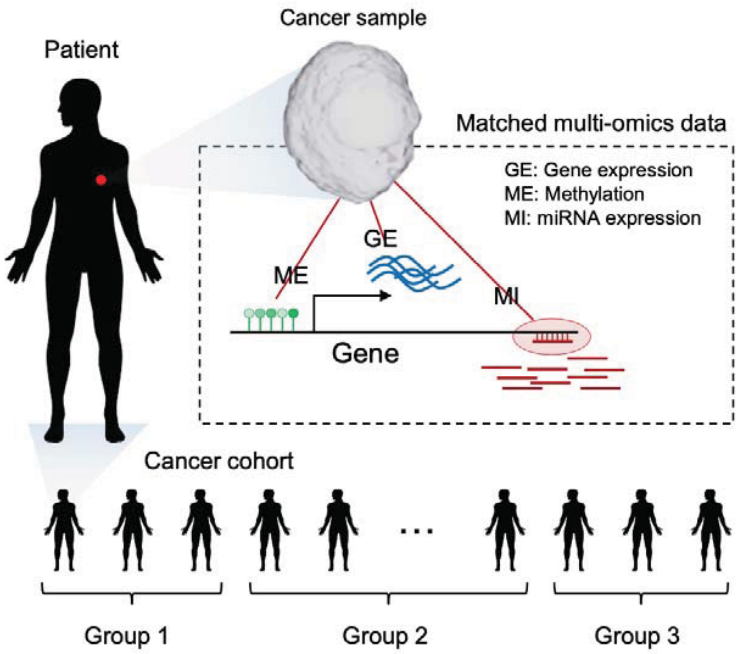


Some recent research topics on MO

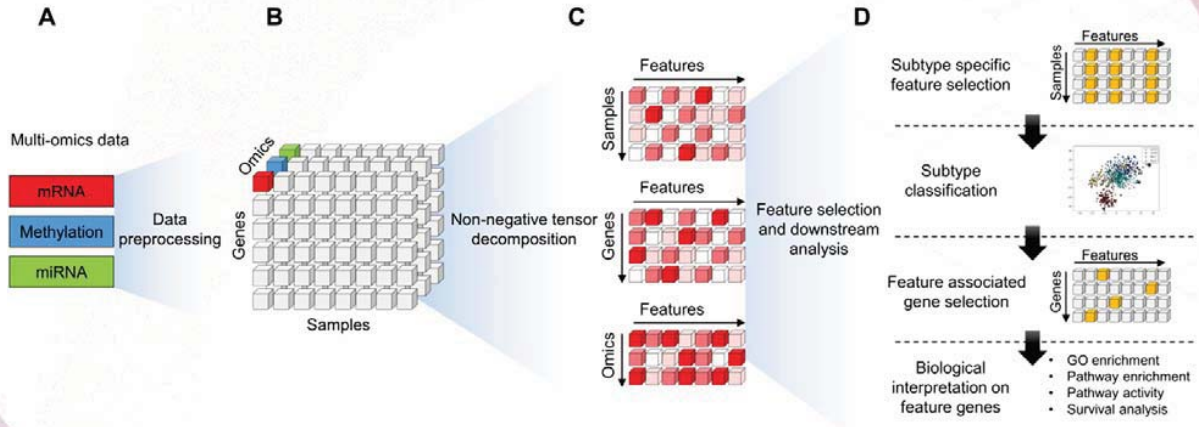
- MONTI: A multi-omics non-negative tensor decomposition framework for the integrated analysis of cancer subtypes (Frontiers in Genetics, 2021)
- MOPA: An Integrative Multi-Omics Pathway Analysis Method for Measuring Omics Activity (in preparation)
- Parametric analysis on large-scale multi-omics data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

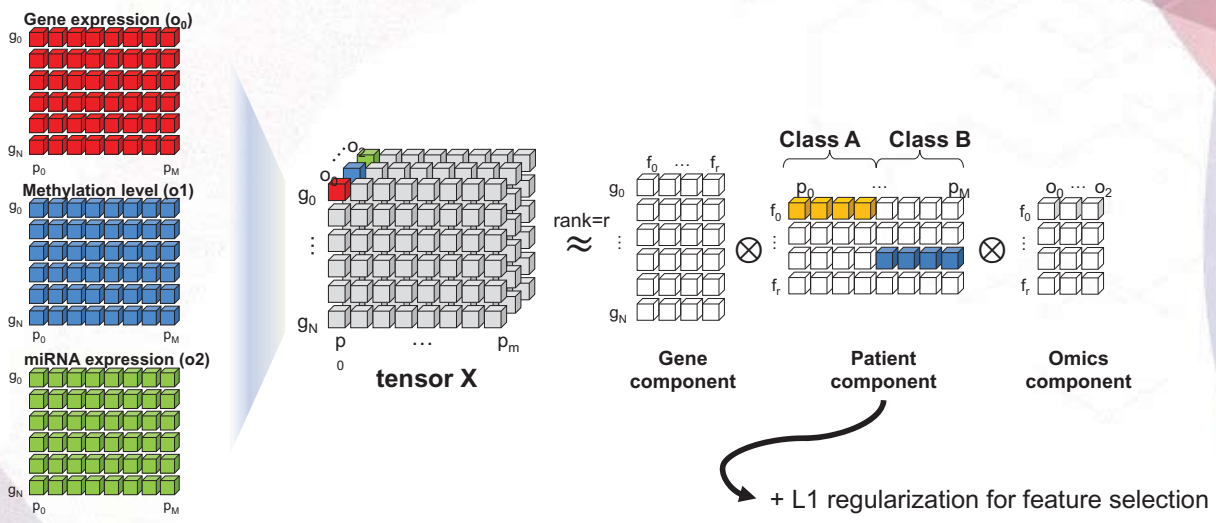


Methods| MONTI Workflow



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

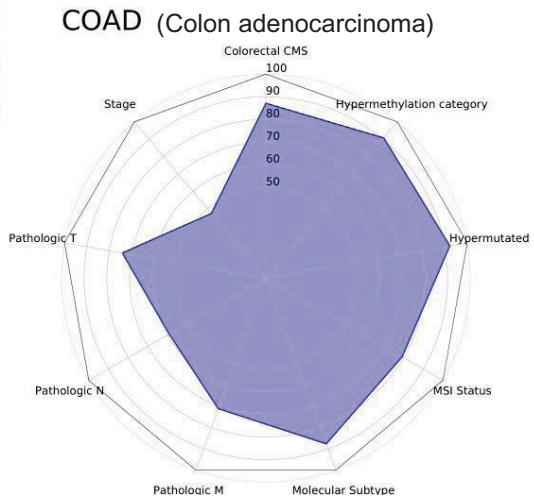


Data| Multi-omics Data of 10 Cancer Types and clinical features

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

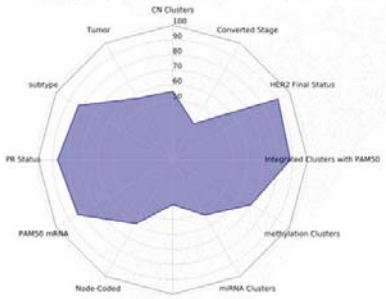
CANCER TYPE	Clinical_Type	Patient_number	Sample_Type	Gene_number
COAD	Colorectal_CMS	206	['CMS1', 'CMS2', 'CMS3', 'CMS4']	14454
STAD	Molecular_Subtype	305	['CIN', 'EBV', 'GS', 'MSI']	
BRCA	subtype	595	['Basal', 'Her2', 'LumA', 'LumB']	
HNSC	gender	298	['FEMALE', 'MALE']	
OV	TUMORSTAGE	320	['IIIC', 'IV']	
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

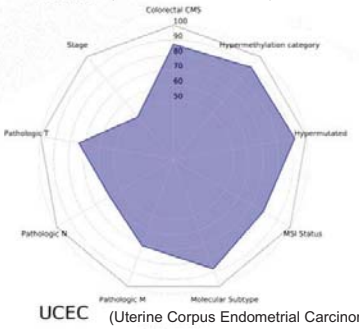


Results| Clinical Feature Classification Accuracy on 10 Cancer types

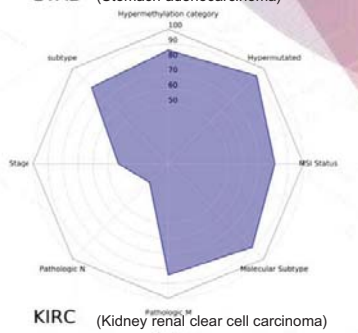
BRCA (Breast invasive carcinoma)



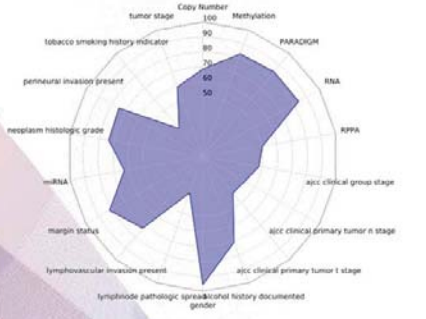
COAD (Colon adenocarcinoma)



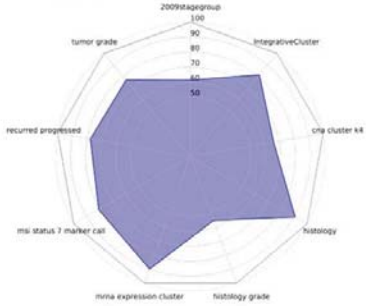
STAD (Stomach adenocarcinoma)



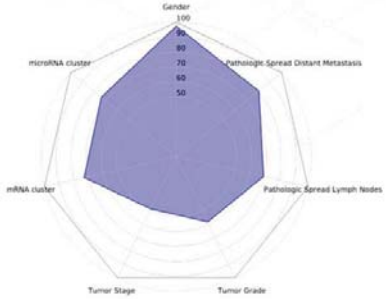
HNSC (Head and Neck squamous cell carcinoma)



UCEC (Uterine Corpus Endometrial Carcinoma)

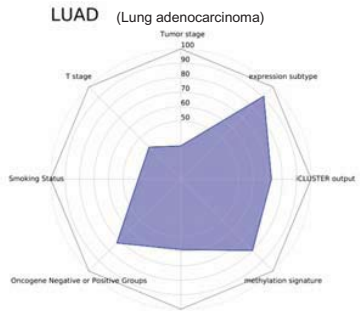


KIRC (Kidney renal clear cell carcinoma)

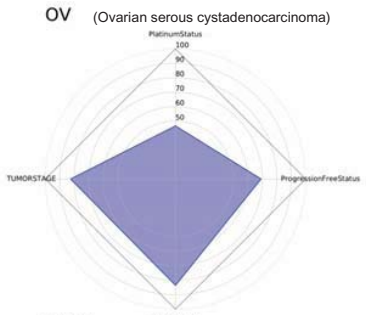


Results| Clinical Feature Classification Accuracy on 10 Cancer types

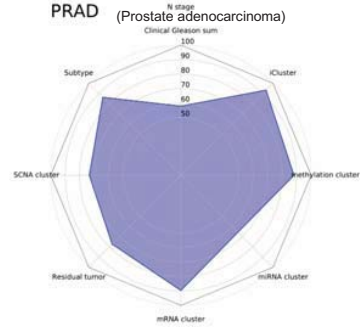
LUAD (Lung adenocarcinoma)



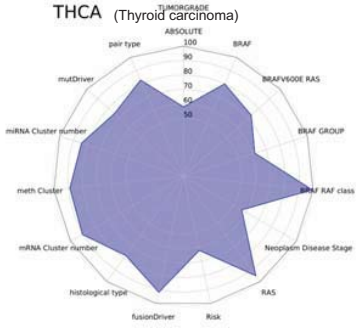
OV (Ovarian serous cystadenocarcinoma)



PRAD (Prostate adenocarcinoma)

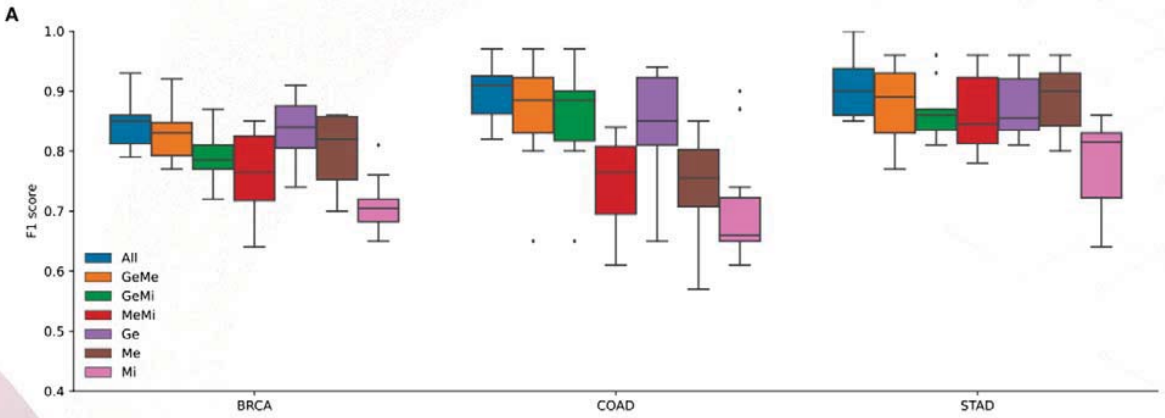


THCA (Thyroid carcinoma)



Results| Omics combinations

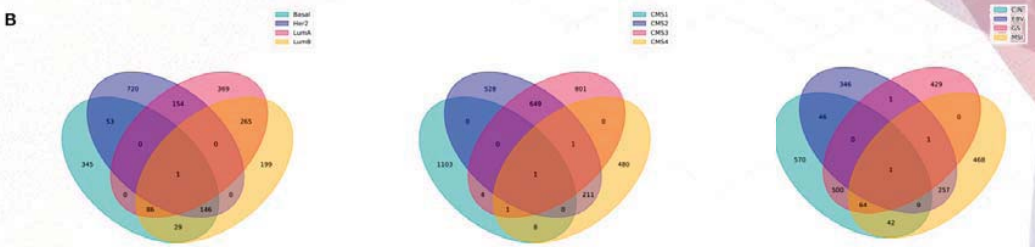
- Different combinations show different results



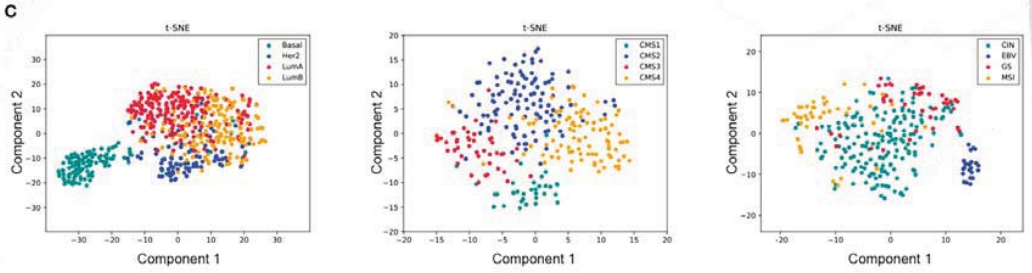
Results| Subtype features

BRCA COAD STAD

- Some features are shared across the subtypes

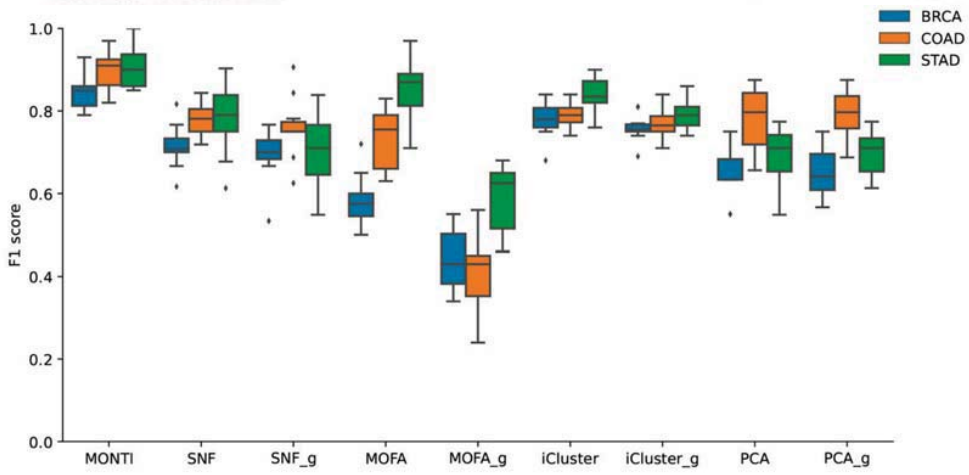


- The subtypes of BRCA, COAD and STAD are well separated



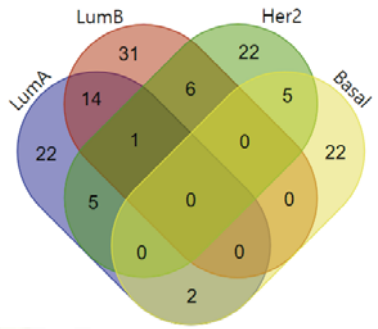
Results| Is gene-level integration helpful? (1)

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

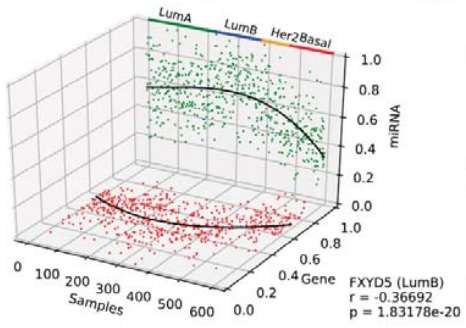


Results| Is gene-level integration helpful? (2)

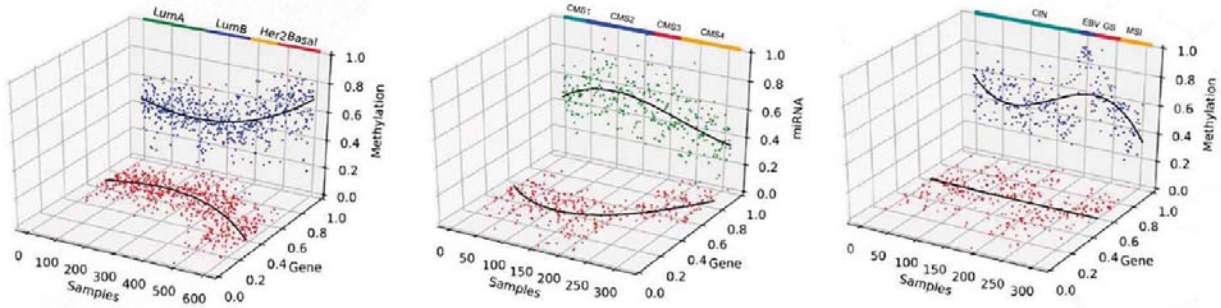
Total of 130 features selected



FXYD5 gene and miRNA expression show significant relationship



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



EXOC6 (Basal) OLFML2B (CMS4) MAPK15 (EBV)

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 |

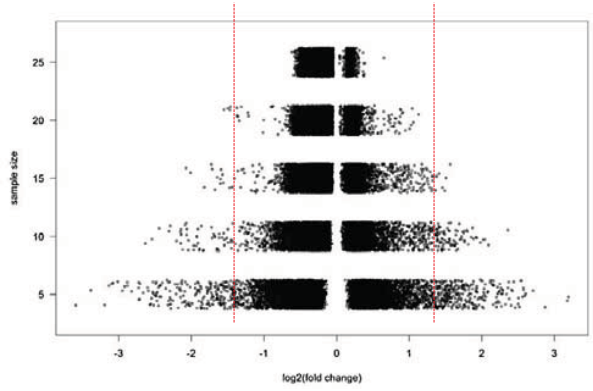
- What clinical features can be explained by MO?
- How many samples are at least required for robust results?
- How many genes or MO features are needed?
- What happens if sample size is not balanced between groups?

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
Original dataset (no differences between patients and controls)					
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
Perturbed dataset (100 miRNAs set to differentially expressed between patients and controls)					
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

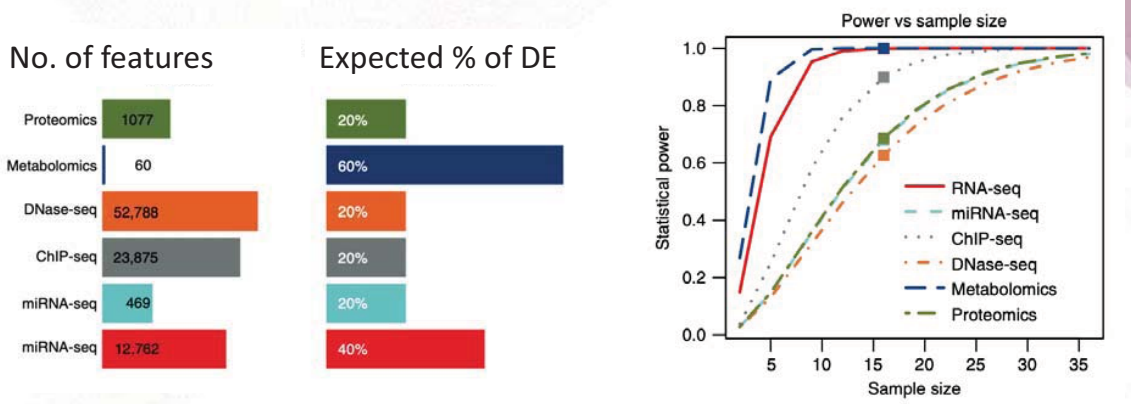


Table 1 MultiPower parameters and results from the STATegra pilot data.

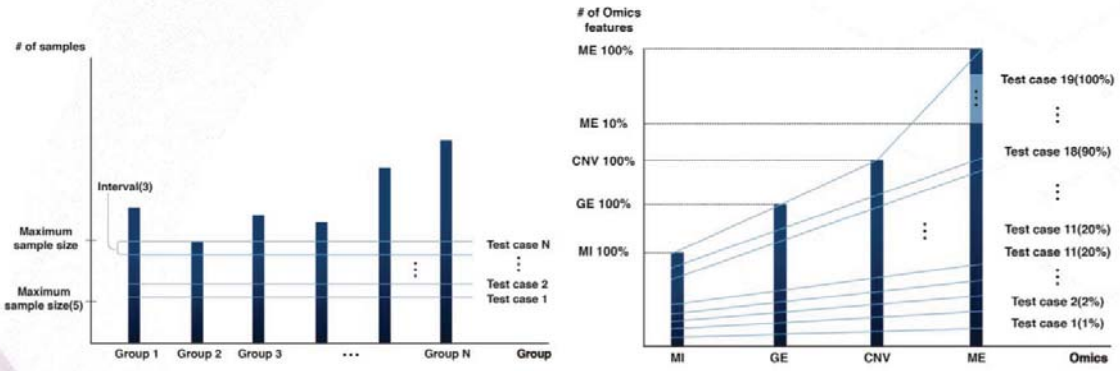
Omic	numFeat[b]	DEperc[b]	Delta[a]	Dispersion[a]	minSampleSize	optSampleSize	Power
RNA-seq	12,762	0.4	0.61	0.32	5	16	0.999
miRNA-seq	469	0.2	0.50	0.46	14	16	0.680
ChIP-seq	23,875	0.2	1.35	0.96	10	16	0.898
DNase-seq	52,788	0.2	0.51	0.49	16	16	0.627
Metabolomics	60	0.6	1.20	0.52	4	16	1.000
Proteomics	1077	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.
numFeat number of omic features, *DEperc* expected proportion of DE features, *delta* difference of means to be detected, *dispersion* pooled standard deviation, *minSampleSize* sample size to achieve the minimum power per omic, *optSampleSize* optimal sample size for the experiment, *power* power reached with the optimal sample size.
[a]Parameter estimated by MultiPower.
[b]Parameter 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

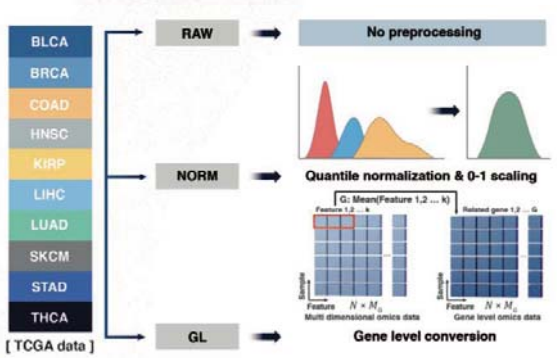


(a) Sample size test

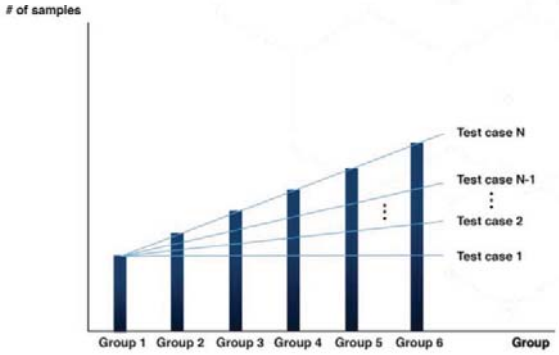
(b) Feature selection test

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

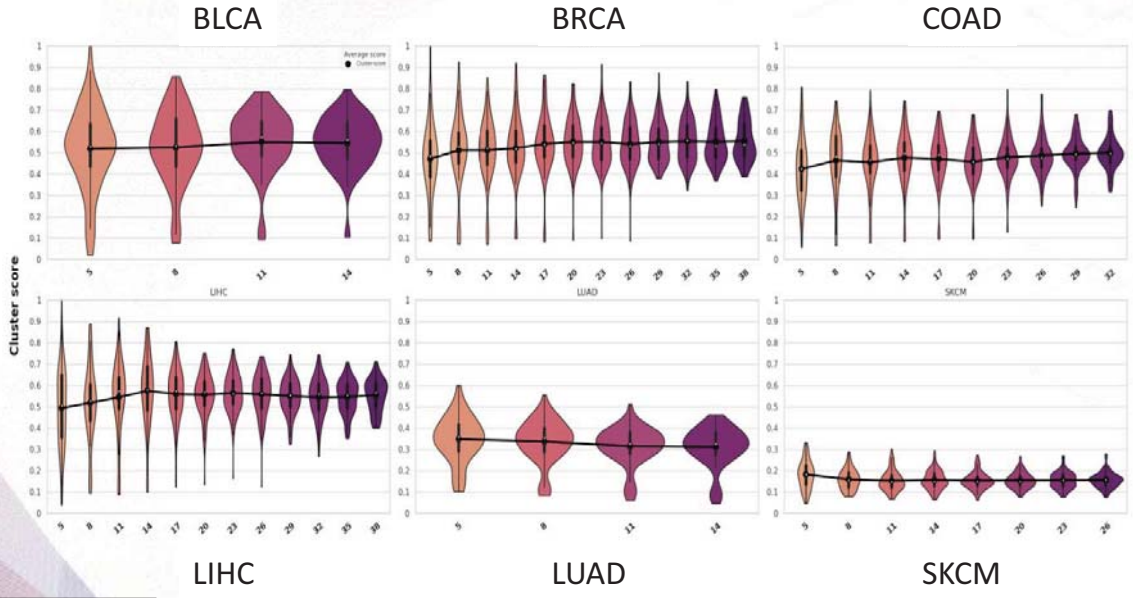


(d) Balance test

(paper in preparation…)

Benchmark tests | Preliminary results – Sample size

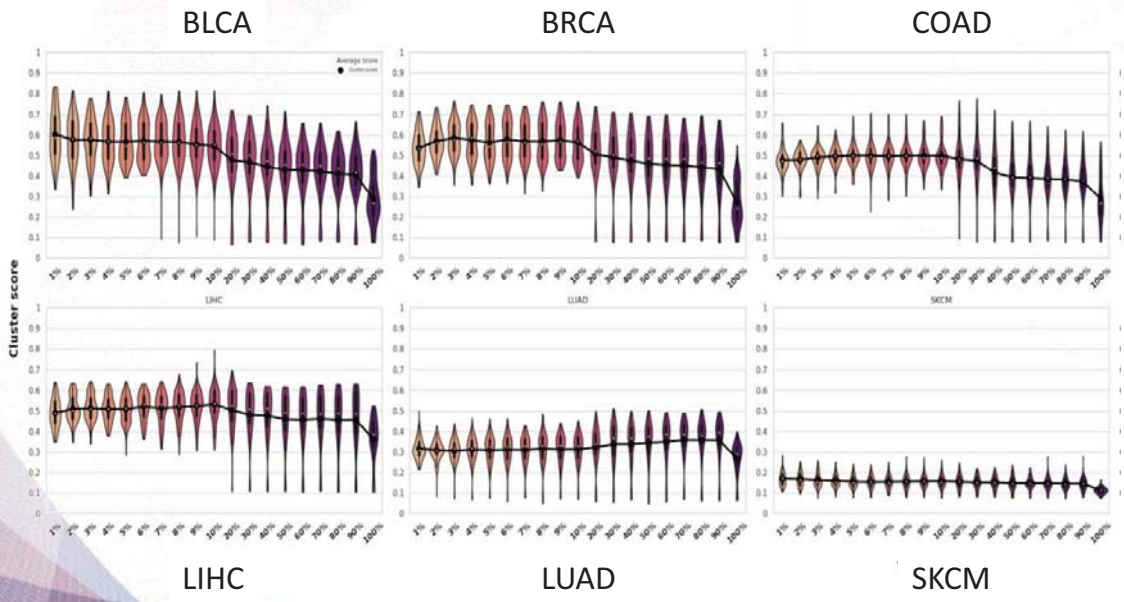
- Using 10 cancer types, the performance started to converge with sample n > 60



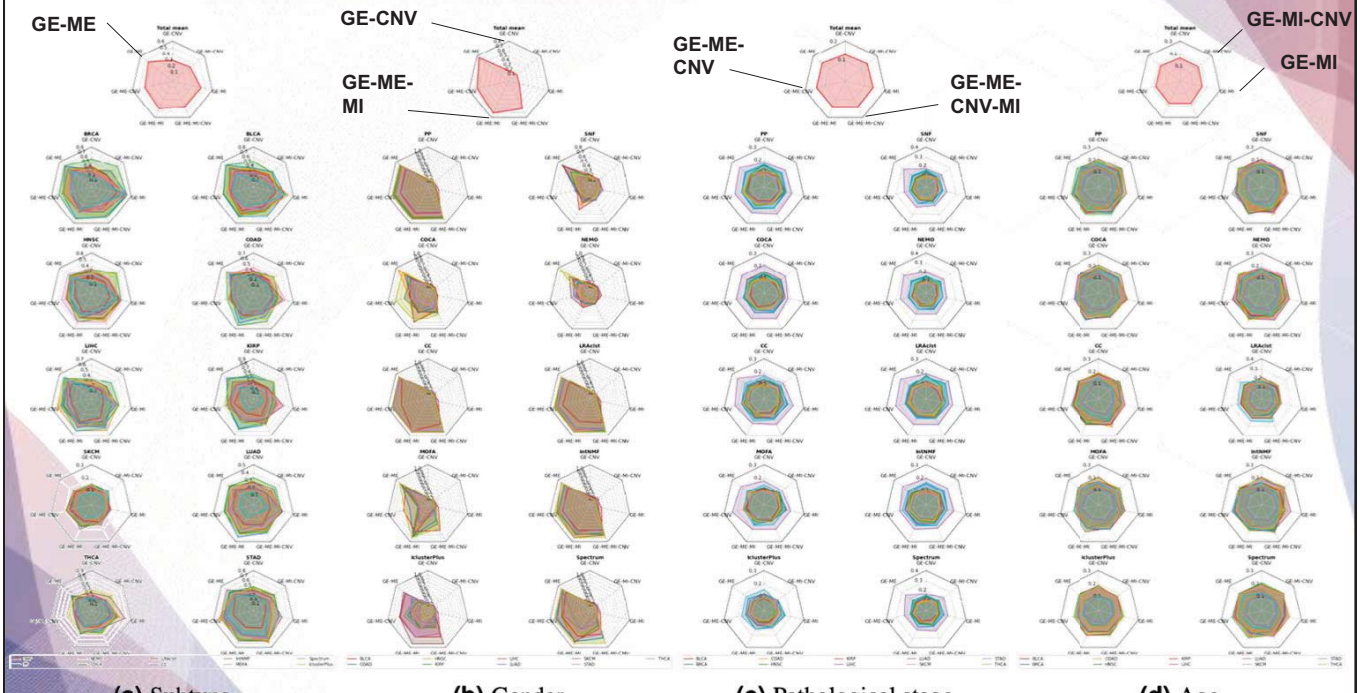
BLCA - Bladder Urothelial Carcinoma
BRCA - Breast invasive carcinoma
COAD - Colon adenocarcinoma
LIHC - Liver hepatocellular carcinoma
LUAD - Lung adenocarcinoma
SKCM - Skin Cutaneous Melanoma

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



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

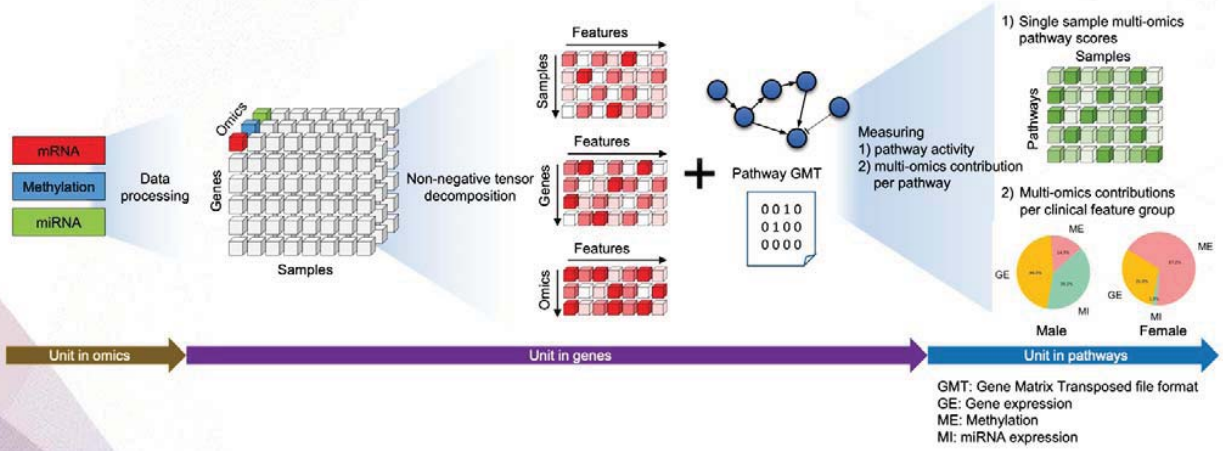
  → biological interpretation needs further work on the given results

- 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

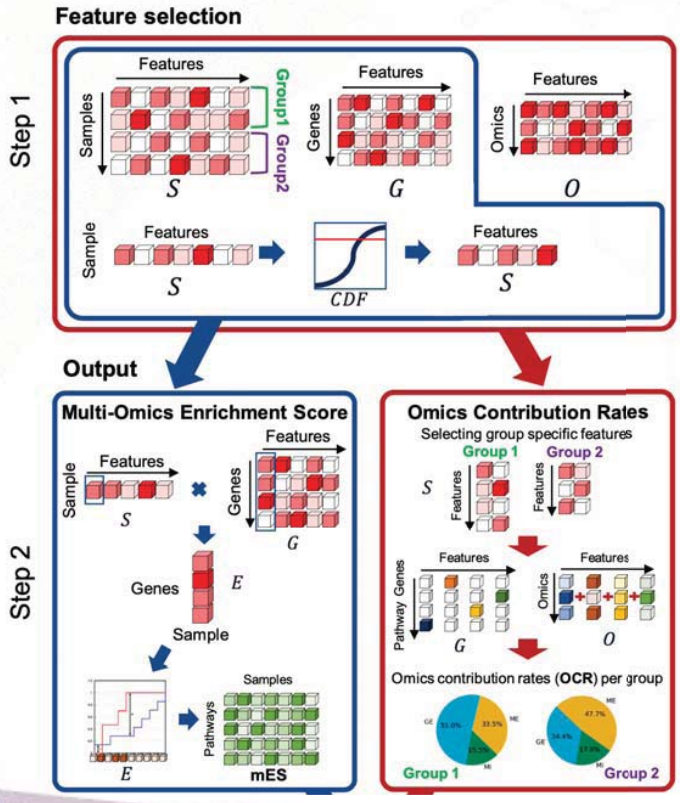
- MOPA is a tool that scores pathway activity based on MO data for each sample and each pathway
- The framework is very similar to GSVA or SSGEA but extended to consider MO data



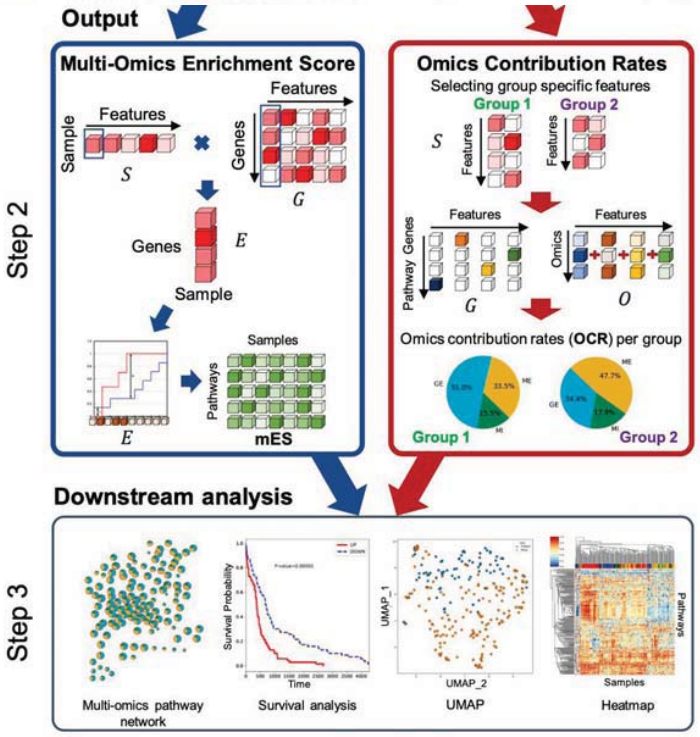
Objective| mES and OCR metrics

- 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)

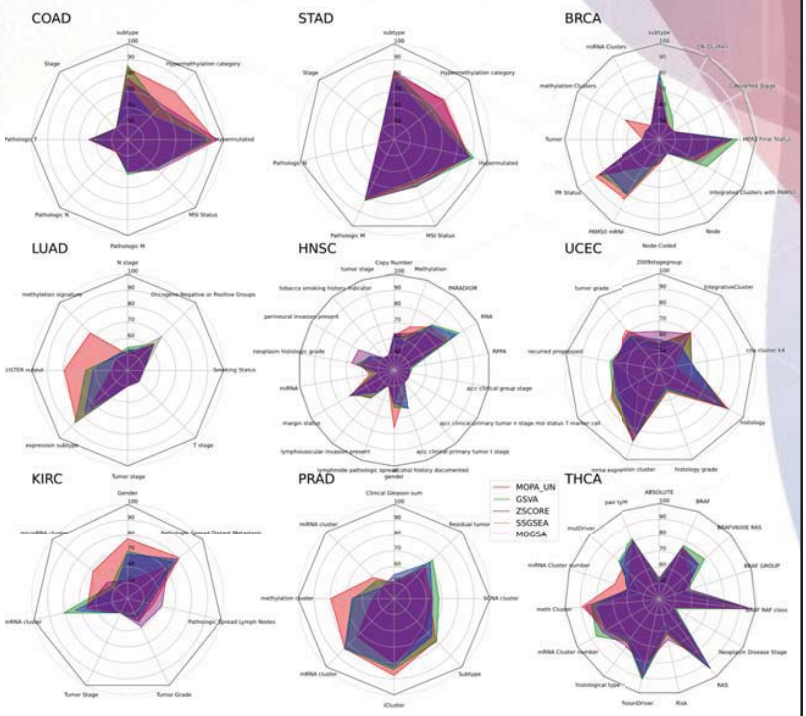


Methods| MOPA Workflow (Step 2~3)

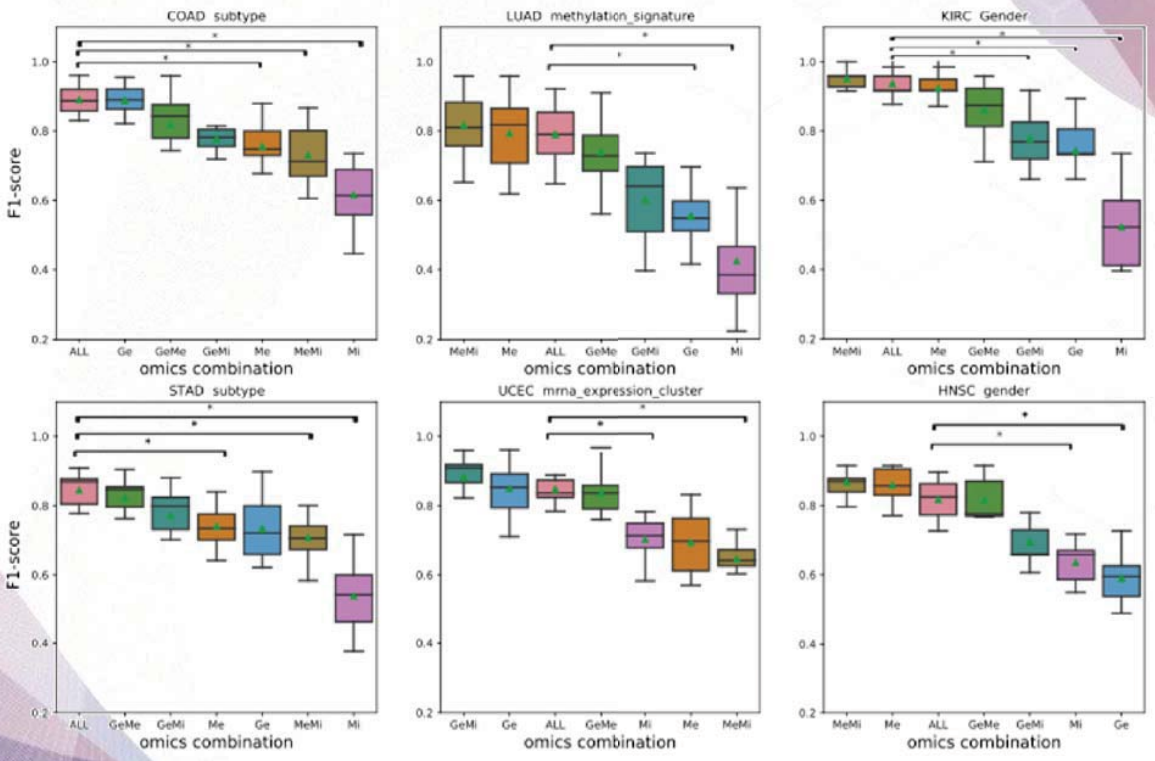


Results | MOPA on Pan-cancer data

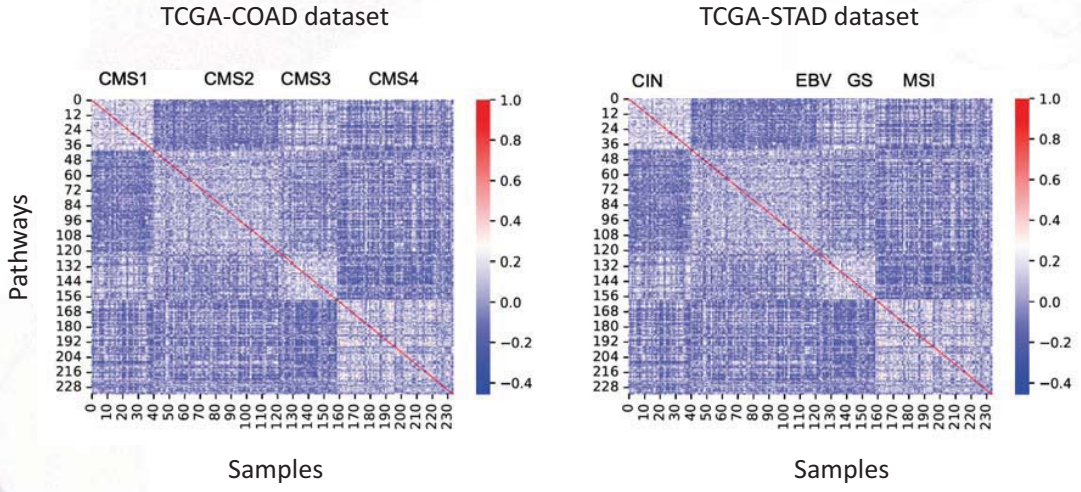
- MOPA was used to analyze 9 cancer types including 95 clinical features (e.g, subtype, cancer stage, gender)
- Some clinical features are well explained while some showed poor classification performance
- In the majority features, MOPA showed higher or equal performance to competing methods



Results | Multi-omics combinations

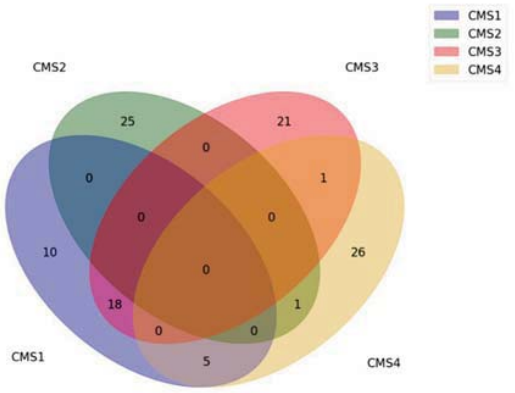


Results | Features detected by MOPA

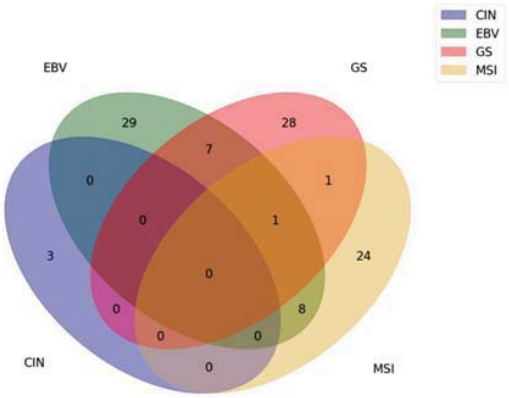


Results | Features detected by MOPA

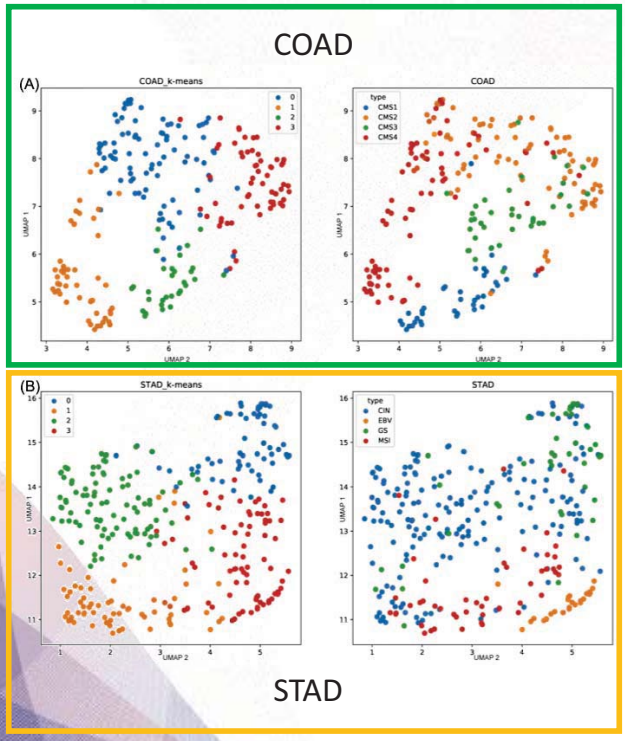
TCGA-COAD

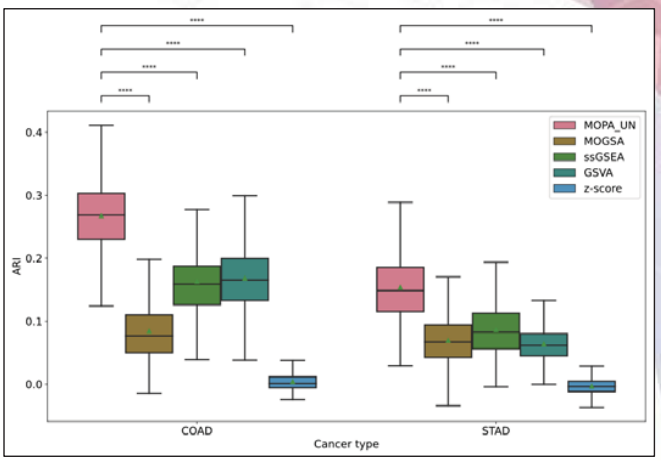


TCGA-STAD

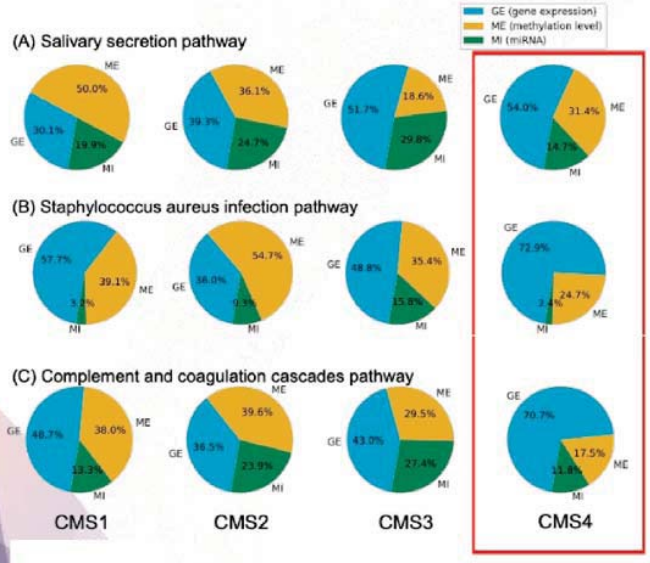


Results | Subtype clustering using *mES*

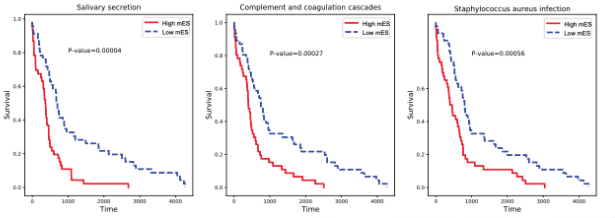




Results | Use Case Study – COAD



Survival plot



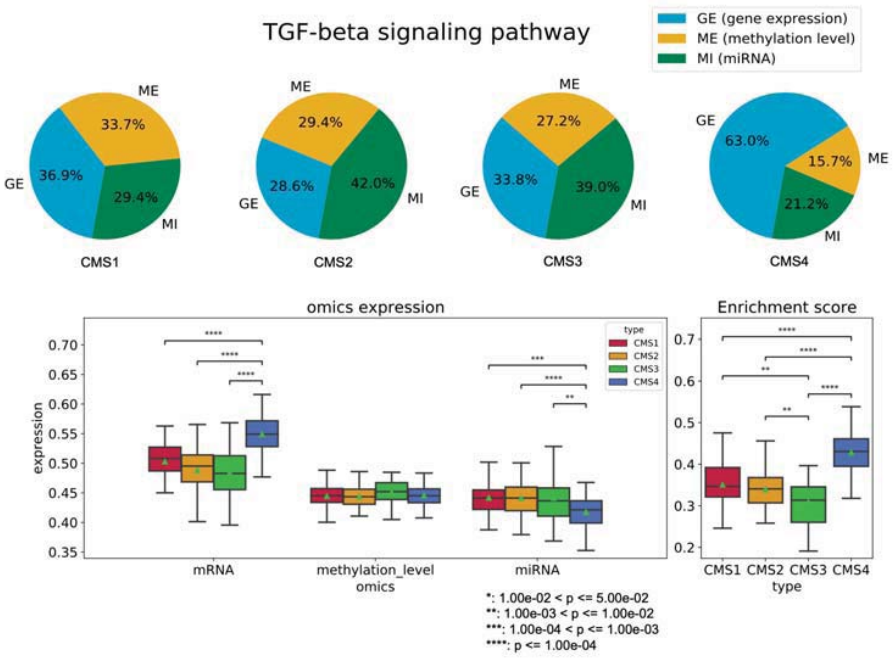
Salivary secretion pathway

Staphylococcus aureus infection pathway

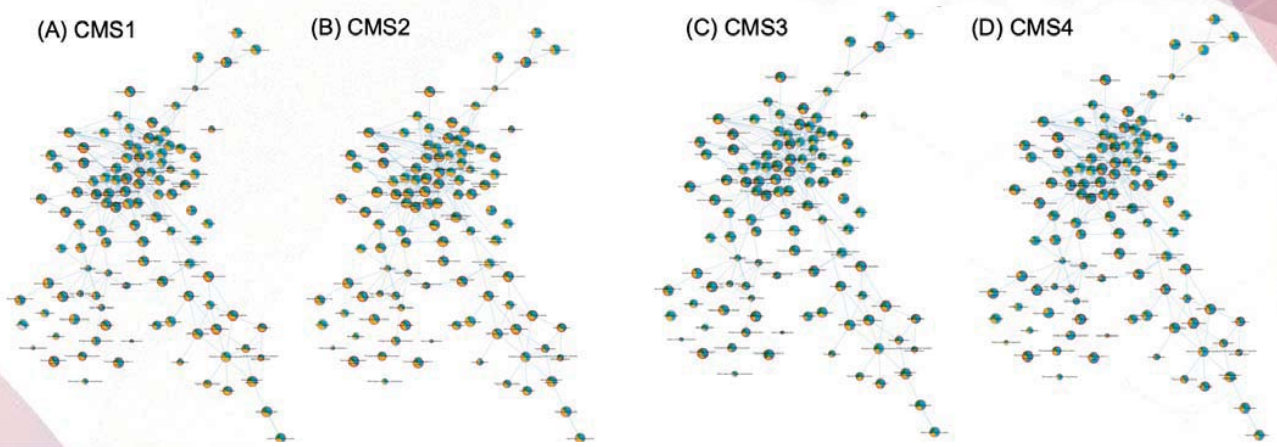
Complement and coagulation cascades pathway

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

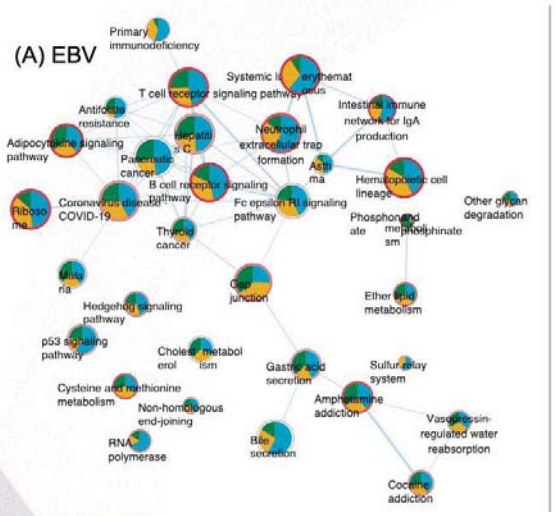
Results | Use Case Study – COAD

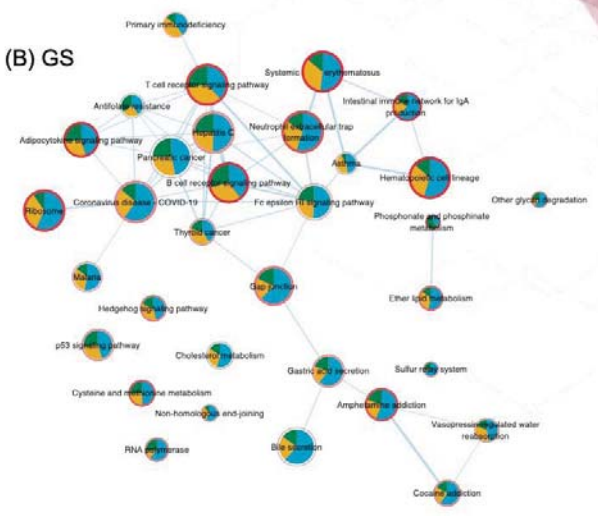


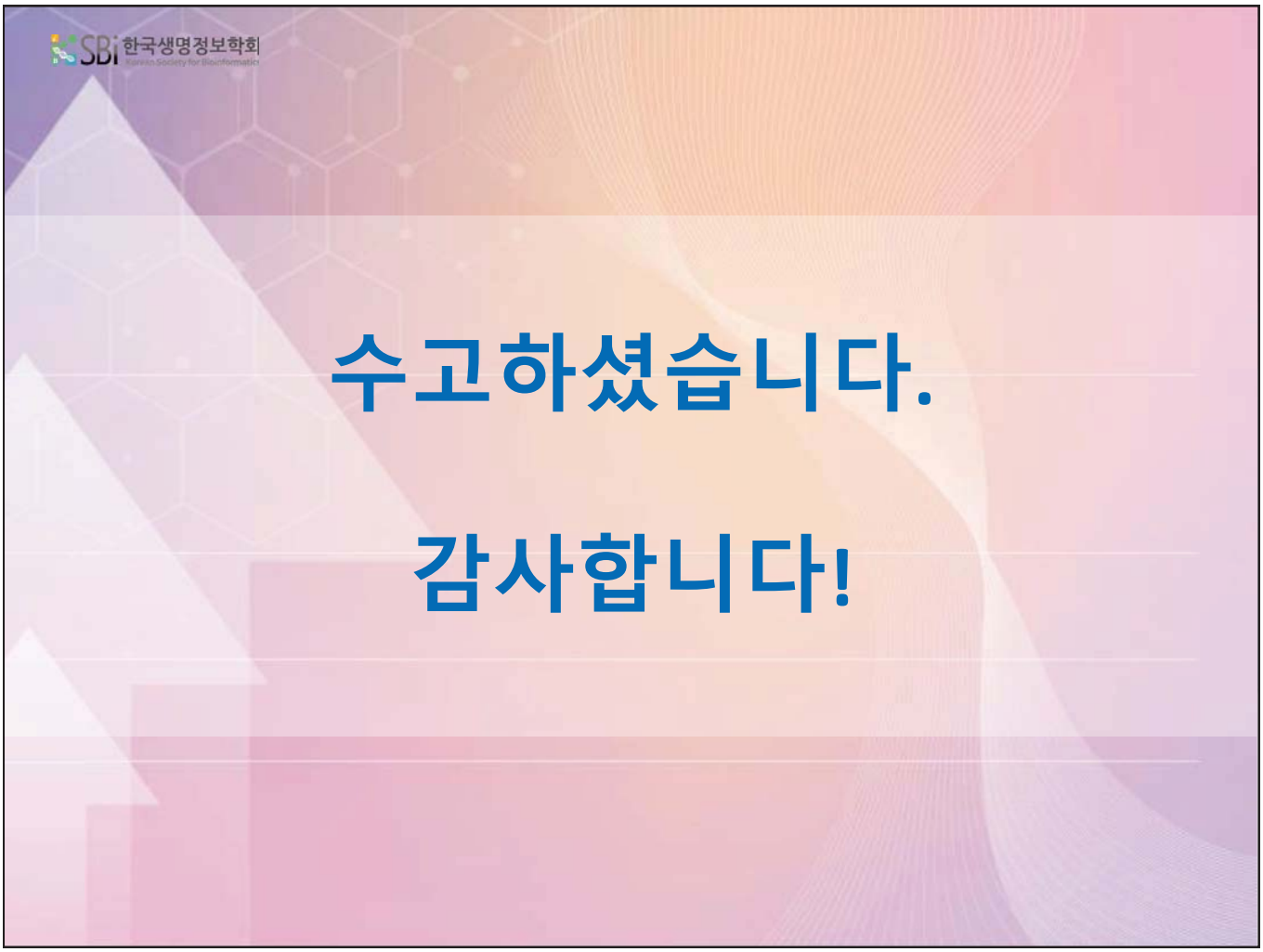
Results | Use Case Study – COAD (MO pathway network)



Results | Use Case Study – STAD






SBi 한국생명정보학회
수고하셨습니다.
감사합니다!