KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical 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 2024 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

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

KSBi-BIML 2024

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

안녕하십니까?

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

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

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

2024년 2월

한국생명정보학회장 이 인 석

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

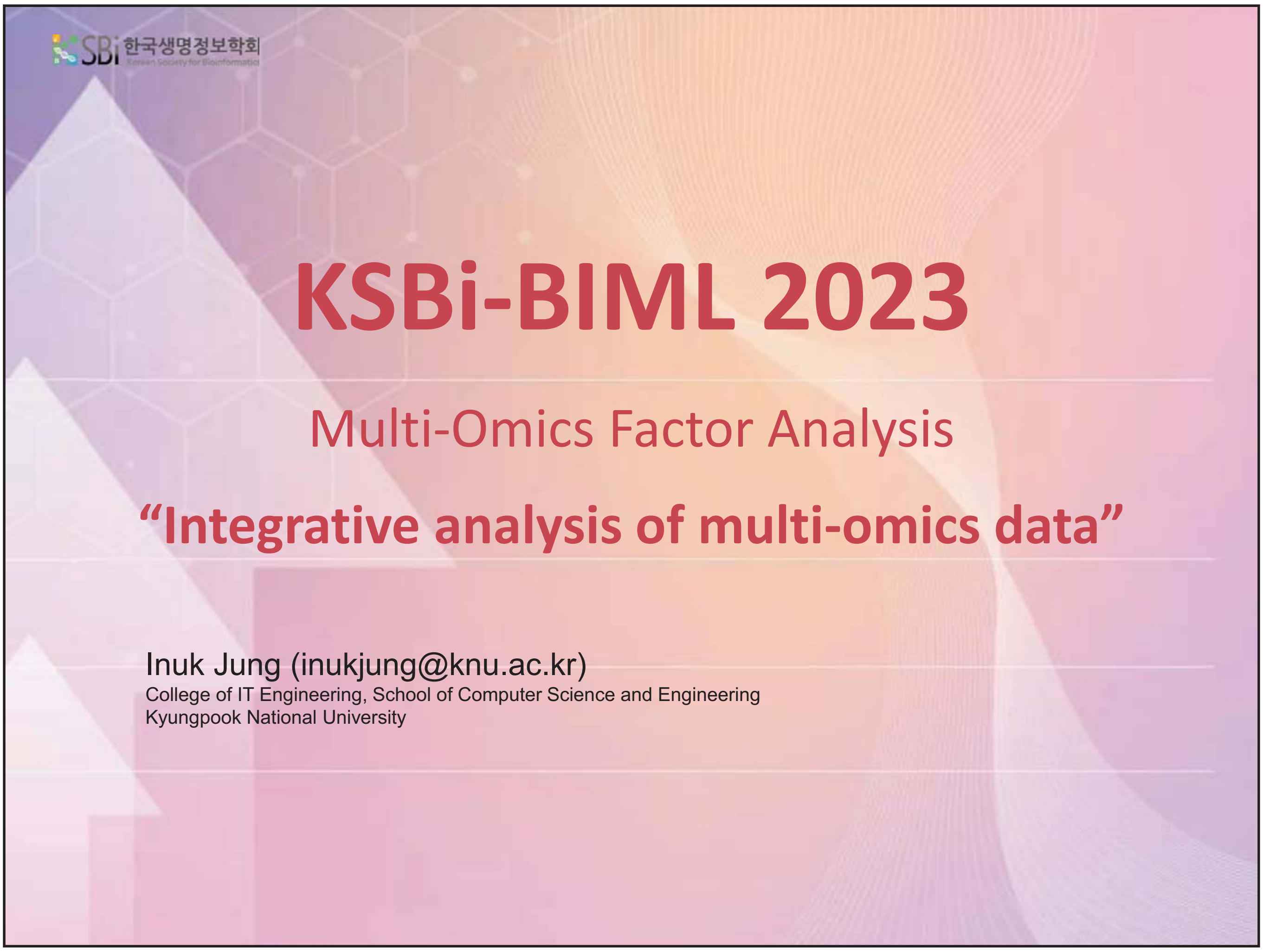
2004	B.S. in Computer Science, Canterbury University, New Zealand
2007	M.S. in Computer Science, Yonsei University, Korea
2017	Ph.D. in Interdisciplinary Program in Bioinformatics, Seoul National University

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

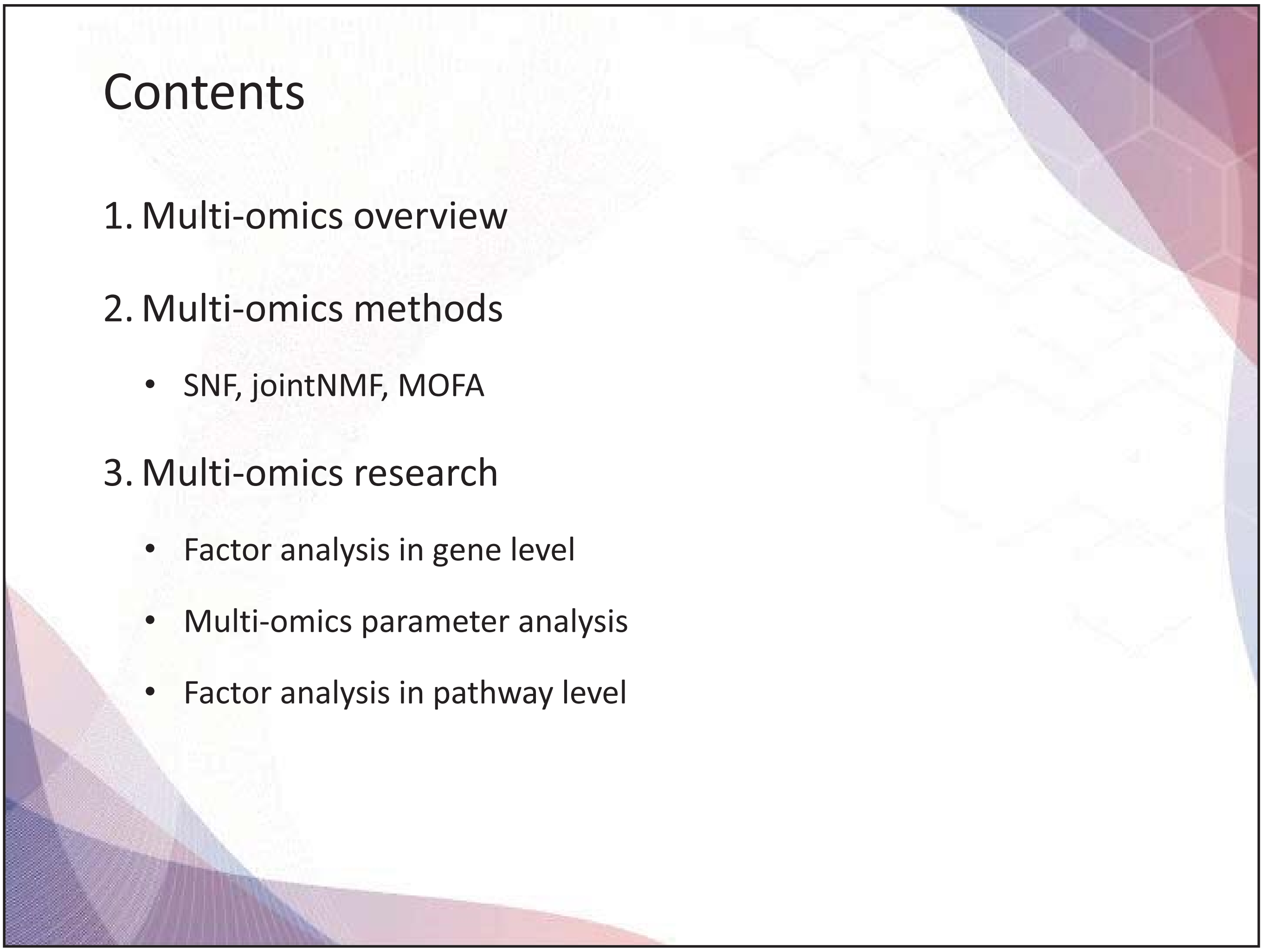


KSBi-BIML 2023

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



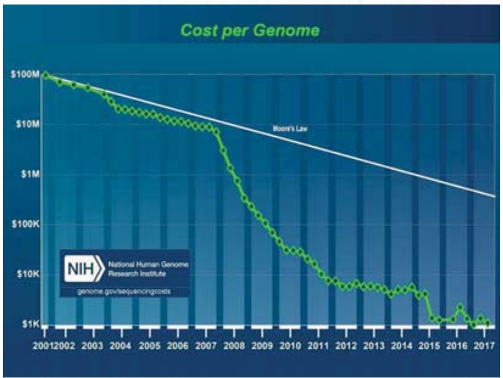
Contents

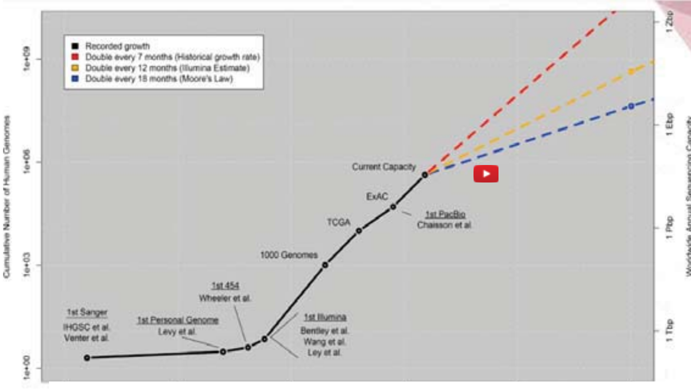
1. Multi-omics overview

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

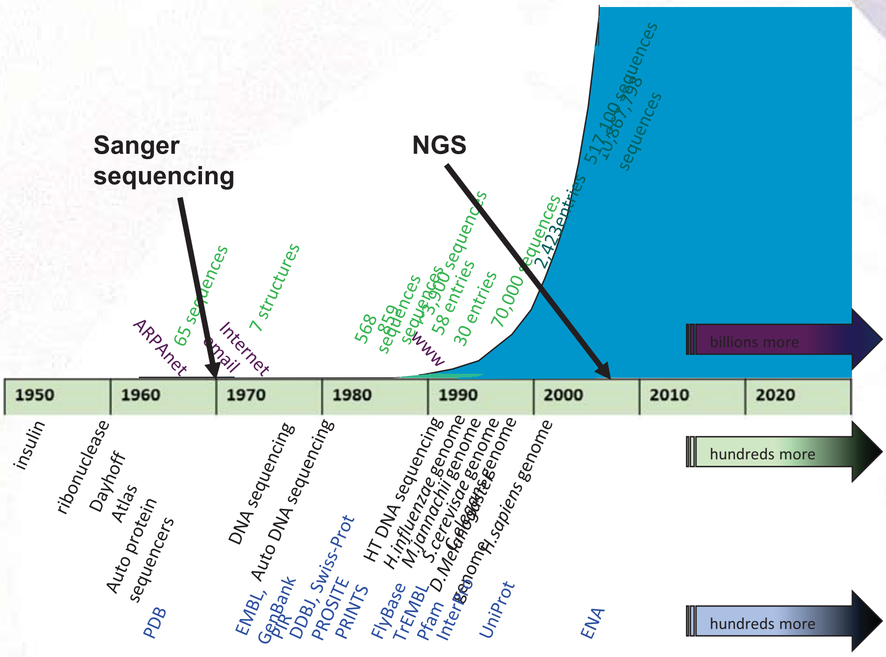
2000

2010

2018

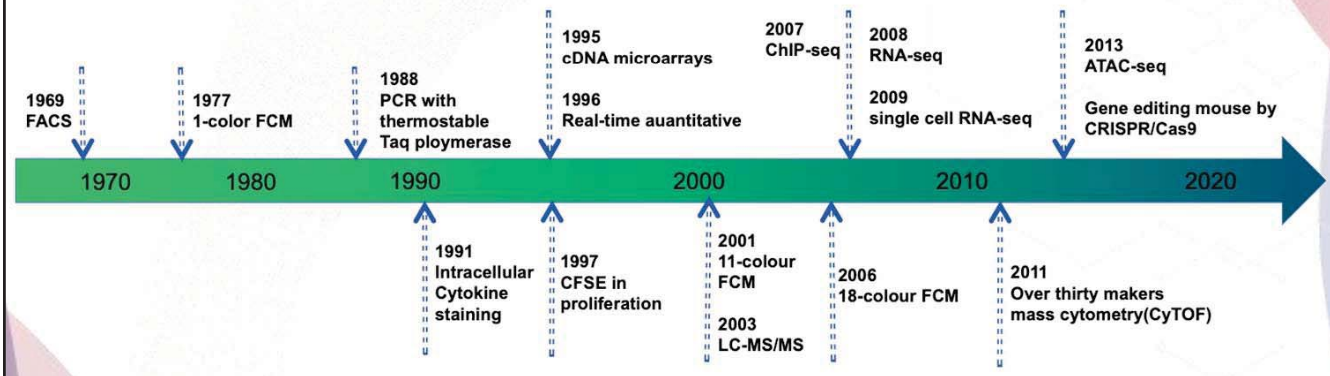
2025

Timeline of Biological data increase



Source from Teresa K.Attwood, University of Manchester

Timeline of multi-omics technologies



1970 1980 1990 2000 2010 2020

1969
FACS

1977
1-color FCM

1988
PCR with
thermostable
Taq ploymerase

1995
cDNA microarrays

1996
Real-time auantitative

2007
ChIP-seq

2008
RNA-seq

2009
single cell RNA-seq

2013
ATAC-seq

Gene editing mouse by
CRISPR/Cas9

1991
Intracellular
Cytokine
staining

1997
CFSE in
proliferation

2001
11-colour
FCM

2003
LC-MS/MS

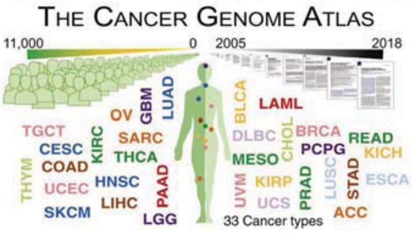
2006
18-colour FCM

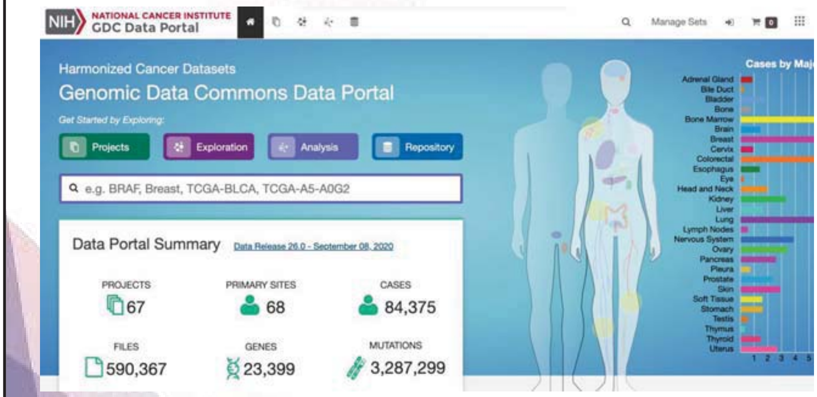
2011
Over thirty makers
mass cytometry(CyTOF)

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

A Big Data Example

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





Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects | Exploration | Analysis | Repository

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary Data Release 26.0 - September 08, 2020

PROJECTS 67

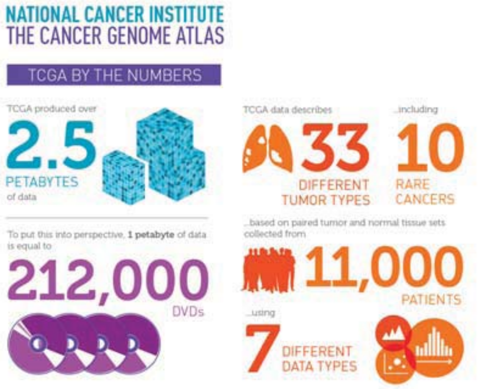
PRIMARY SITES 68

CASES 84,375

FILES 590,367

GENES 23,399

MUTATIONS 3,287,299



NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over 2.5 PETABYTES of data

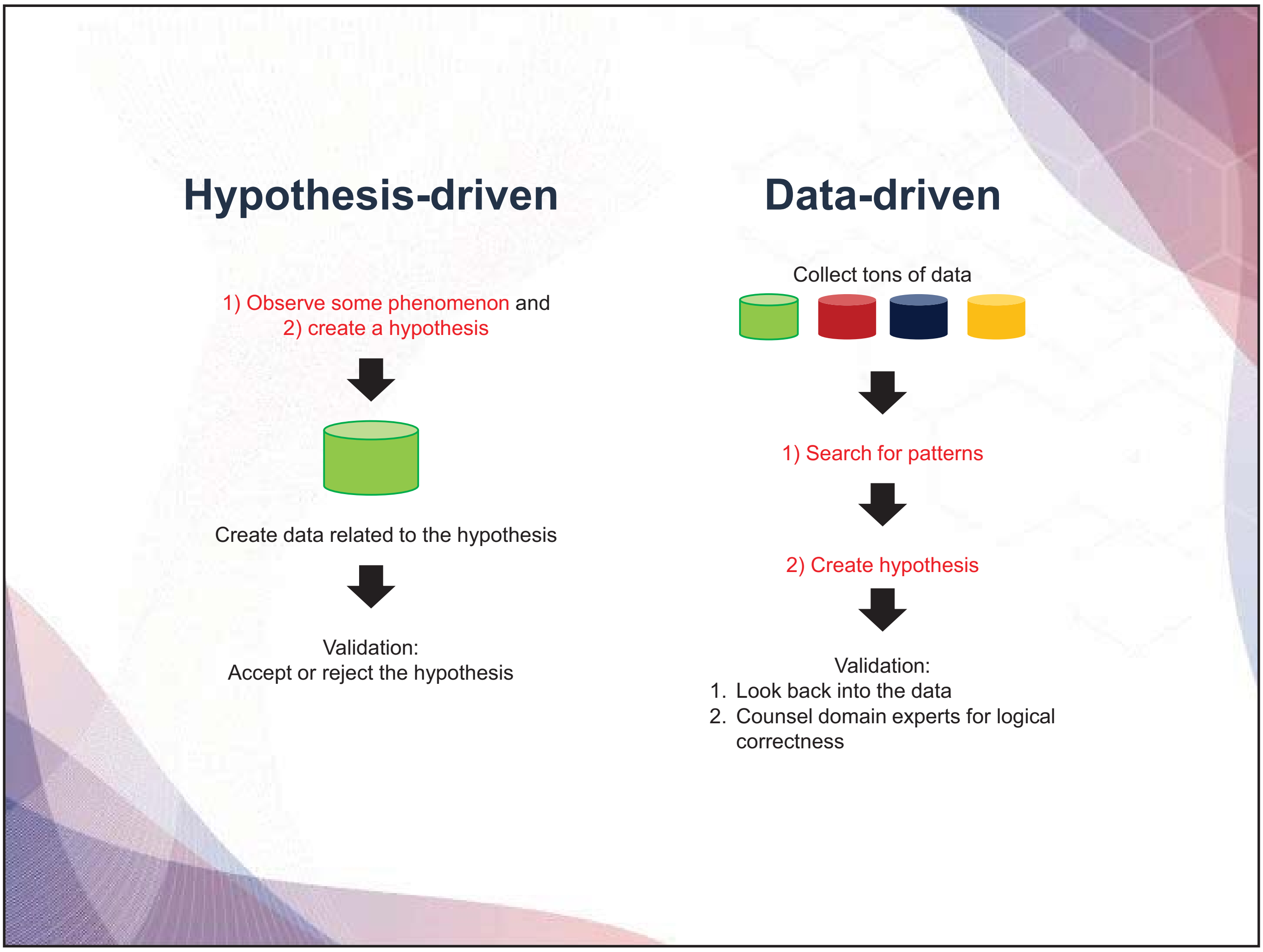
To put this into perspective, **1 petabyte** of data is equal to 212,000 DVDs

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

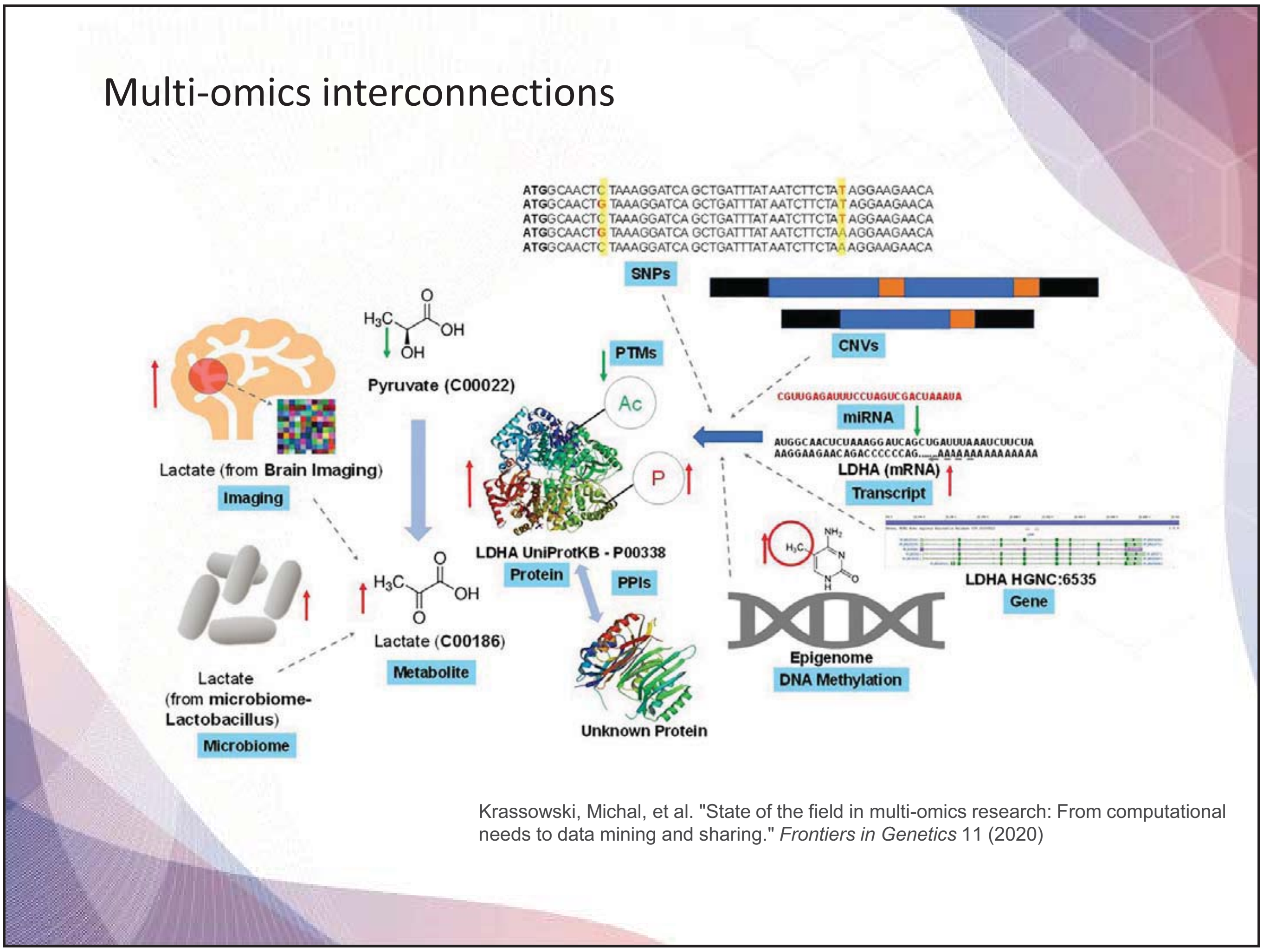
1) Search for patterns

2) Create hypothesis

Validation:

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

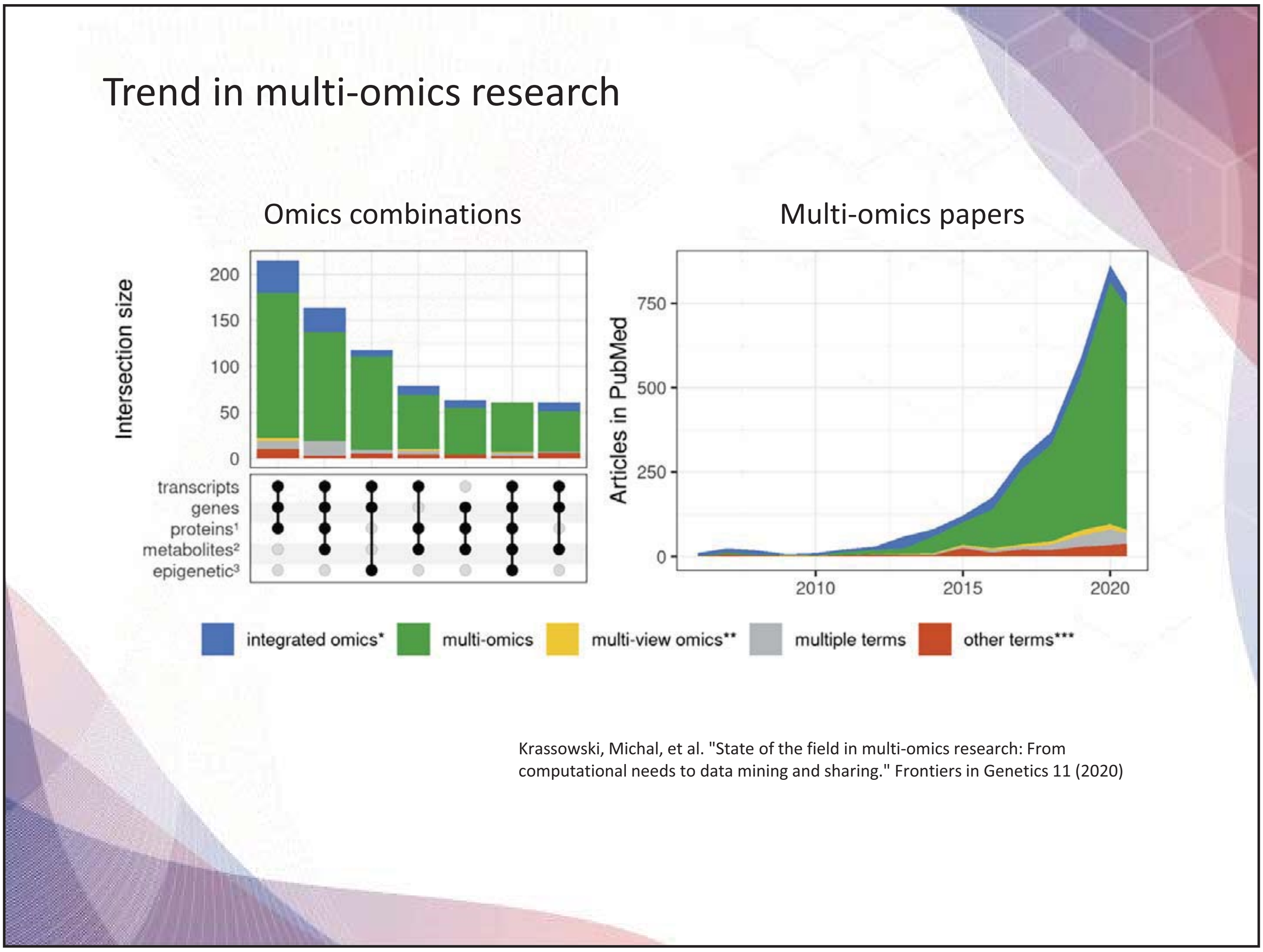
1. Multi-omics BIG DATA



Multi-omics interconnections

ATGGCAACTC TAAAGGATCA GCTGATTTAT AATCTTCTAT AGGAAGAACA
ATGGCAACTG TAAAGGATCA GCTGATTTAT AATCTTCTAT AGGAAGAACA
ATGGCAACTC TAAAGGATCA GCTGATTTAT AATCTTCTAT AGGAAGAACA
ATGGCAACTG TAAAGGATCA GCTGATTTAT AATCTTCTAA AGGAAGAACA
ATGGCAACTC TAAAGGATCA GCTGATTTAT AATCTTCTAA AGGAAGAACA
SNPs
CNVs
H3C
O
OH
OH
Pyruvate (C00022)
PTMs
Ac
P
CGUUGAGAUUUCCUAGUCGACUAAAUA
miRNA
AUGGCAACUCUAAAGGAUCAGCUGAUUUAAAUCUUCUA
AAGGAAGAACAGACCCCCCAG......AAAAAAAAAAAAAAA
LDHA (mRNA)
Transcript
Lactate (from Brain Imaging)
Imaging
LDHA UniProtKB - P00338
Protein
PPIs
H3C
O
OH
O
Lactate (C00186)
Metabolite
NH2
H3C
N
N
H
O
Epigenome
DNA Methylation
LDHA HGNC:6535
Gene
Lactate
(from microbiome-
Lactobacillus)
Microbiome
Unknown Protein

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

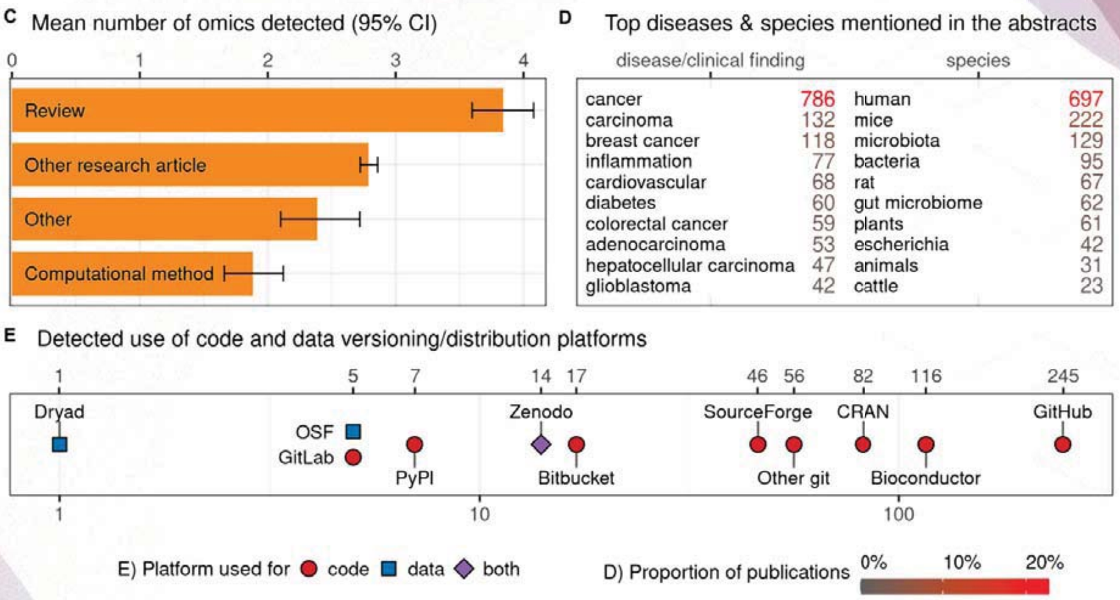
Omics combinations

Multi-omics papers

🟦 integrated omics* 🟩 multi-omics 🟨 multi-view omics** ⬜ multiple terms 🟥 other terms***

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



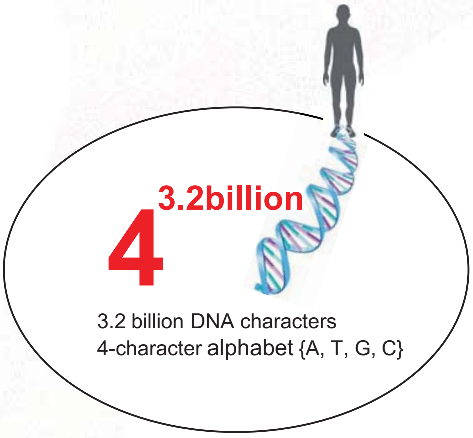
disease/clinical finding		species	
cancer	786	human	697
carcinoma	132	mice	222
breast cancer	118	microbiota	129
inflammation	77	bacteria	95
cardiovascular	68	rat	67
diabetes	60	gut microbiome	62
colorectal cancer	59	plants	61
adenocarcinoma	53	escherichia	42
hepatocellular carcinoma	47	animals	31
glioblastoma	42	cattle	23

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

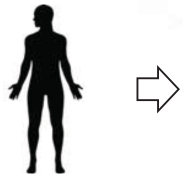


$4^{3.2\text{billion}}$

3.2 billion DNA characters
4-character alphabet {A, T, G, C}

Search space

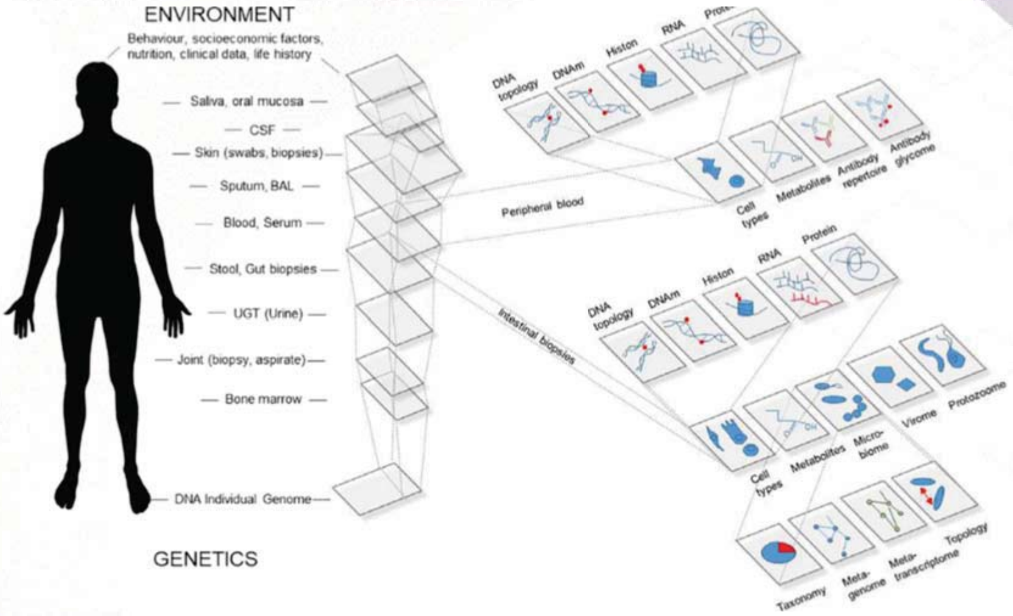
Low sample,
High dimension problem



1 sample

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

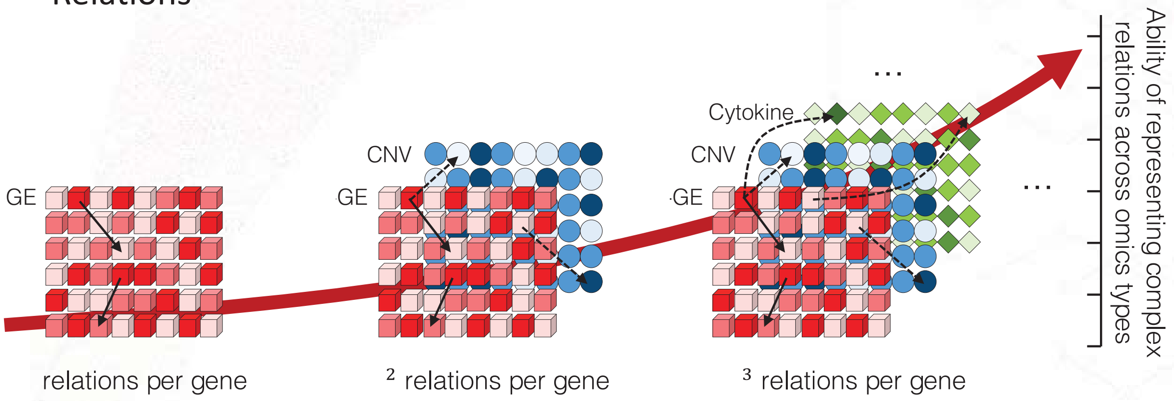
Challenges| Multi-modal data adds extra dimensions



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

Expectations from multi-omics data

- Interpretability
- Relations



GE
CNV
GE
Cytokine
CNV
GE
…
…
relations per gene
^{2} relations per gene
^{3} relations per gene
Ability of representing complex relations across omics types

Challenges| Multi-modal data adds extra dimensions

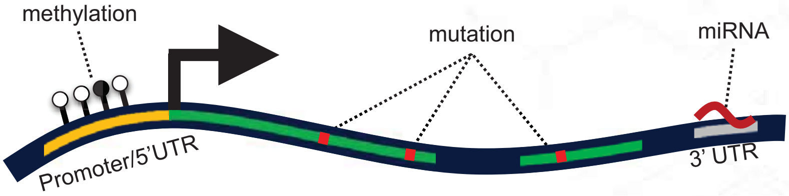
Mutation (DNA)

Methylation (DNA)

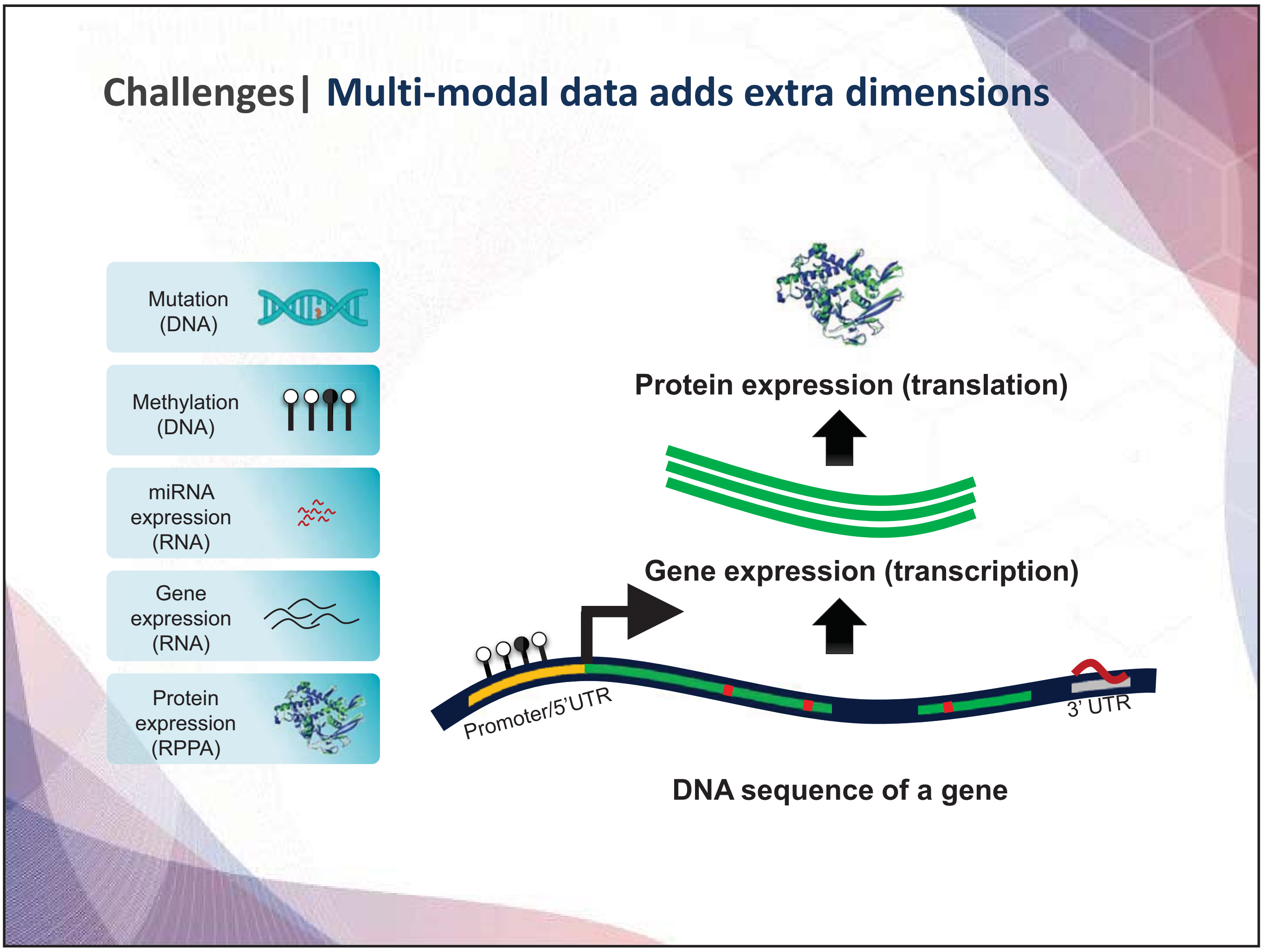
miRNA expression (RNA)

Gene expression (RNA)

Protein expression (RPPA)



DNA sequence of a gene



Challenges| Multi-modal data adds extra dimensions

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

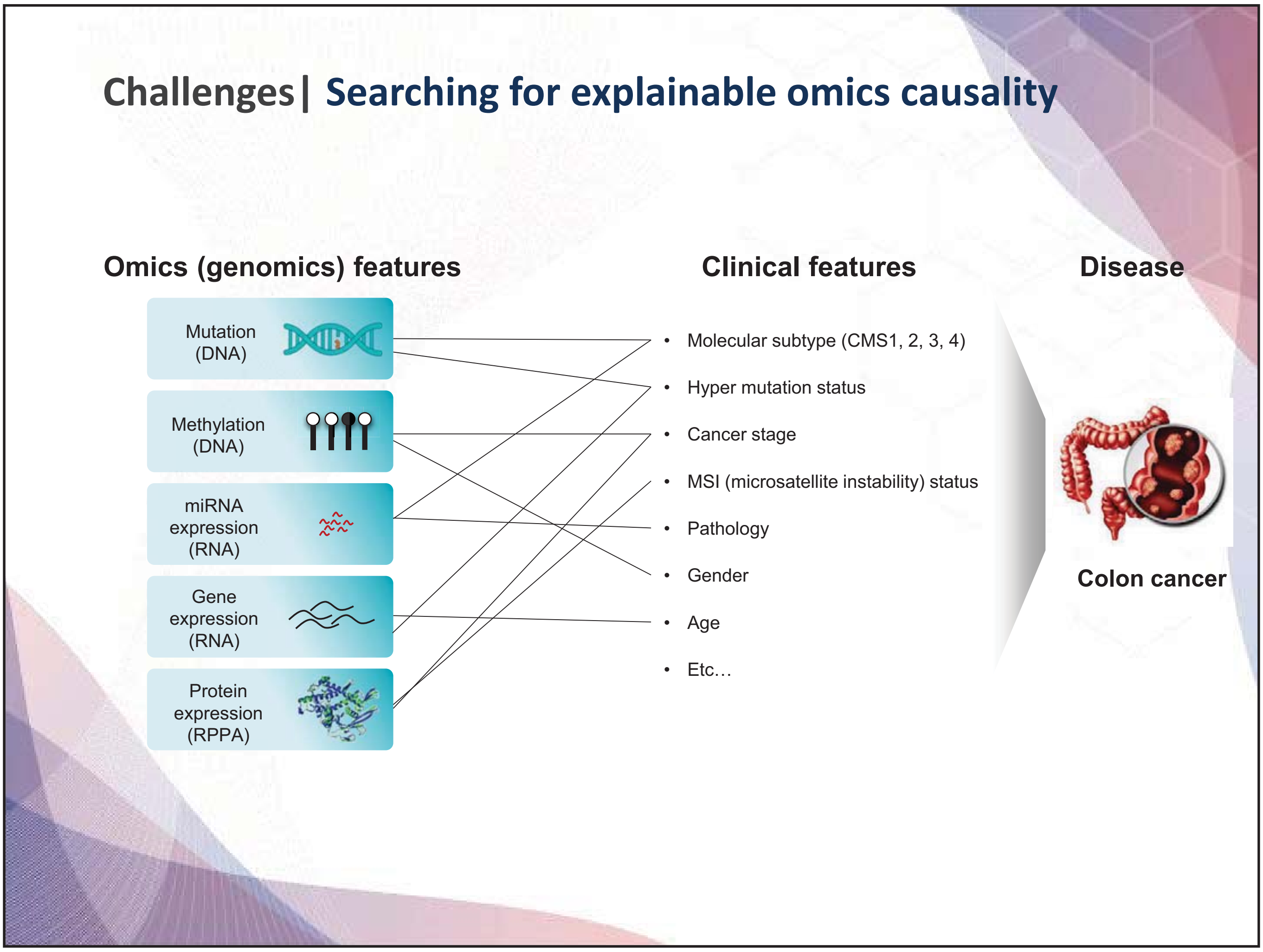
Protein expression (translation)

Gene expression (transcription)

Promoter/5'UTR

3' UTR

DNA sequence of a gene



Challenges| Searching for explainable omics causality

Omics (genomics) features

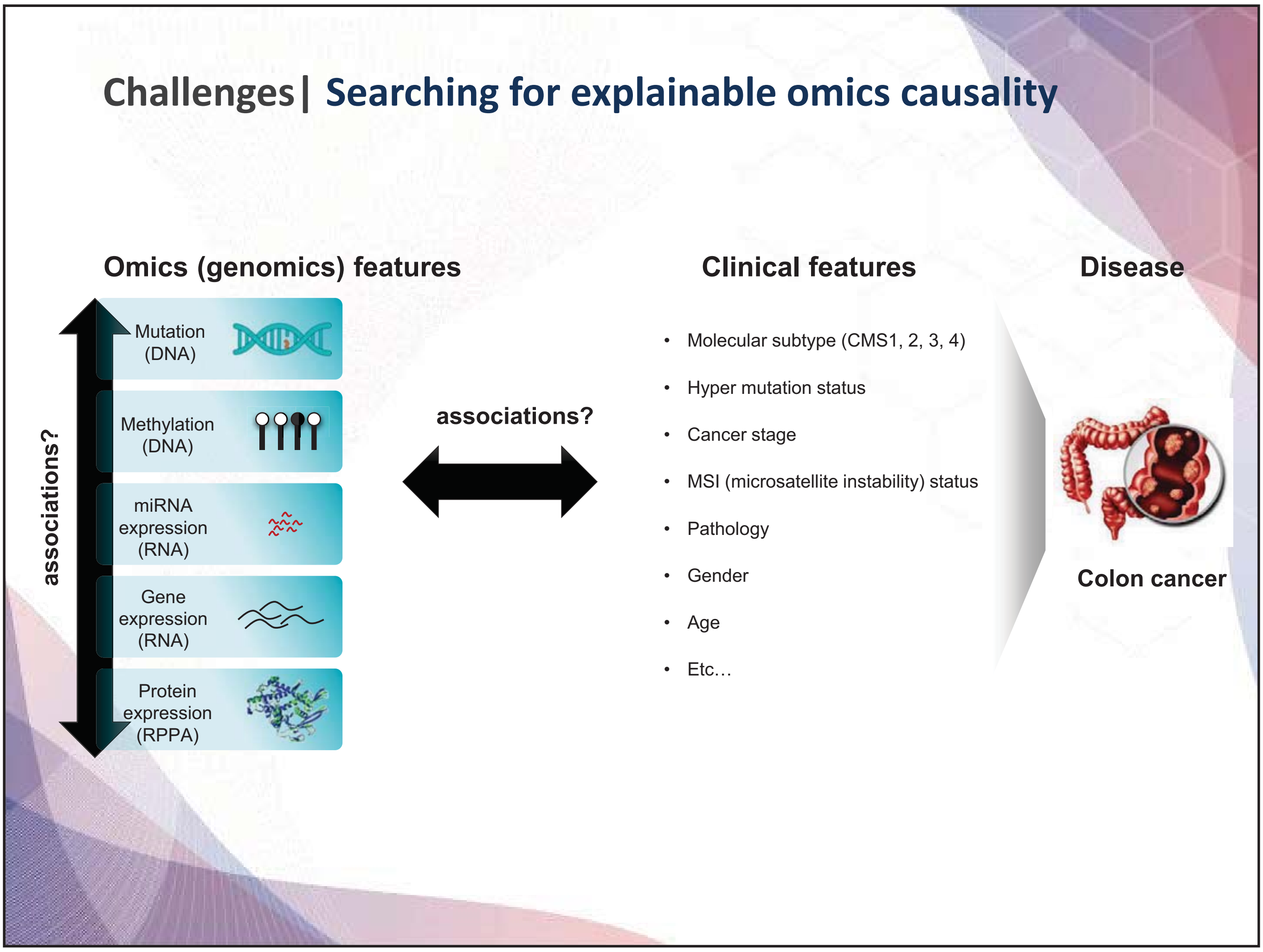
- Mutation (DNA)
- Methylation (DNA)
- miRNA expression (RNA)
- Gene expression (RNA)
- Protein expression (RPPA)

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 (genomics) features

associations?

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

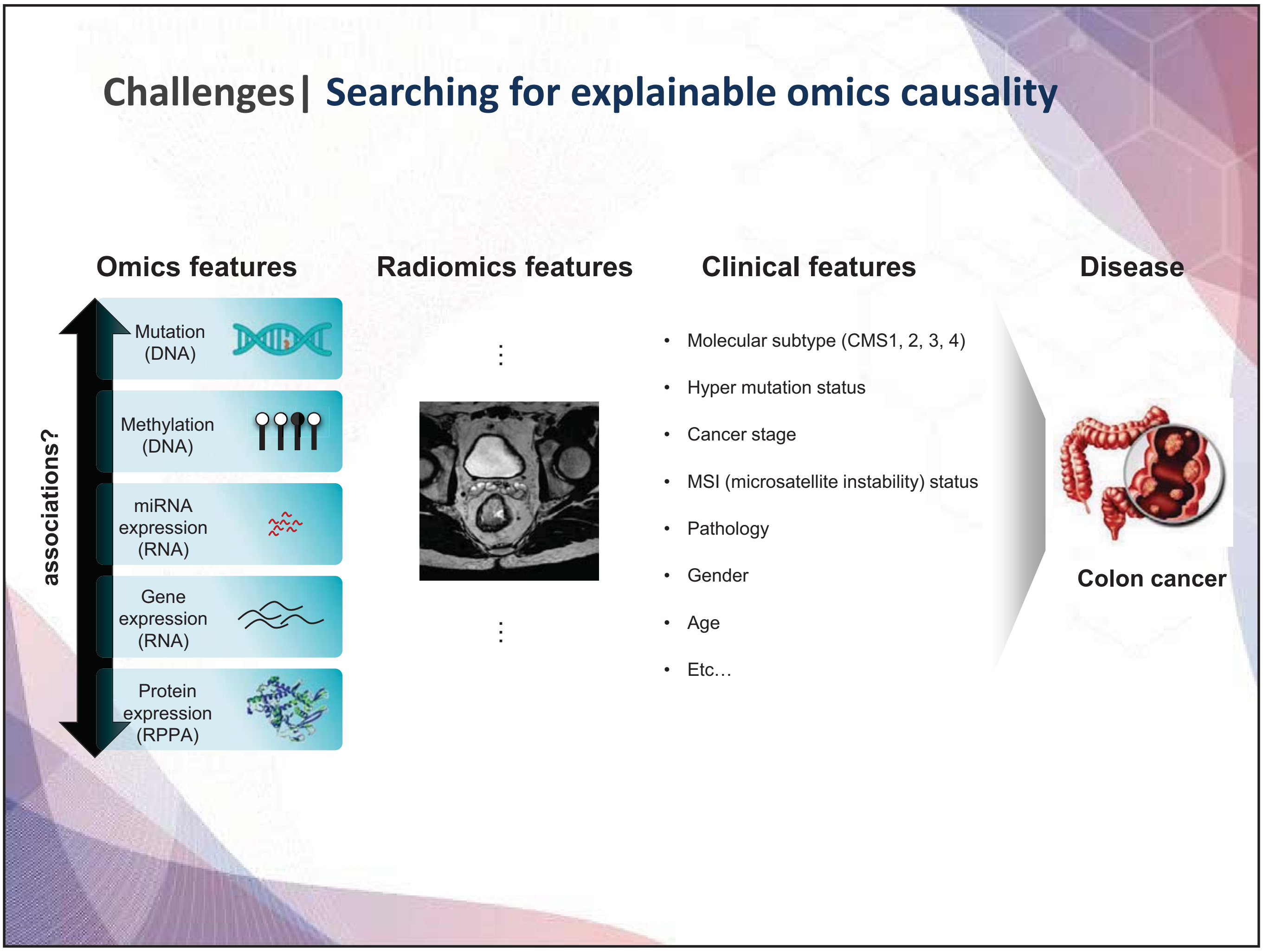
associations?

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

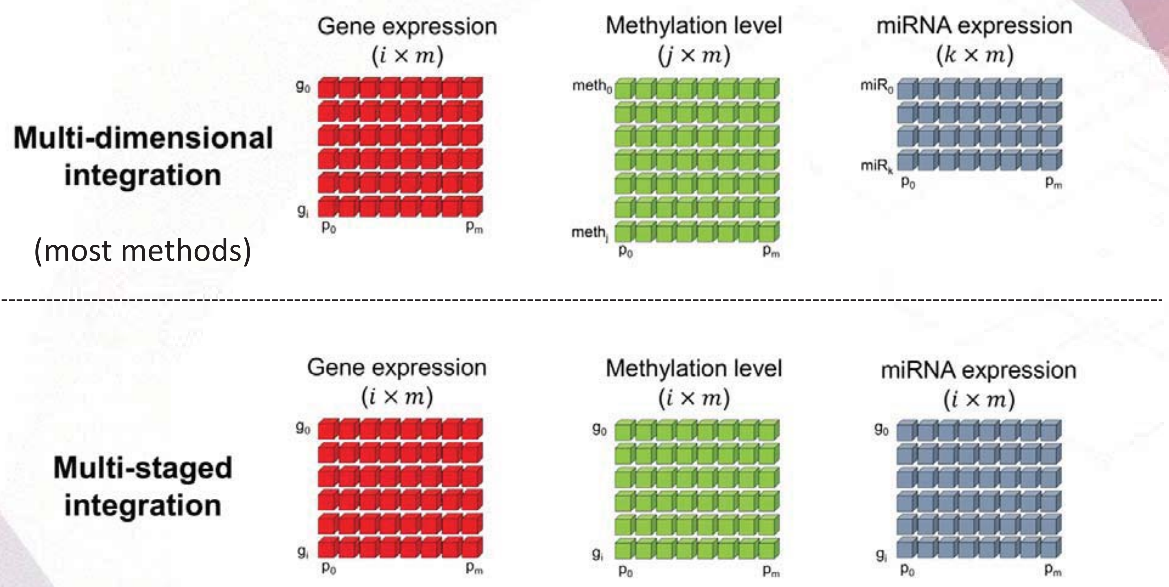
Colon cancer

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



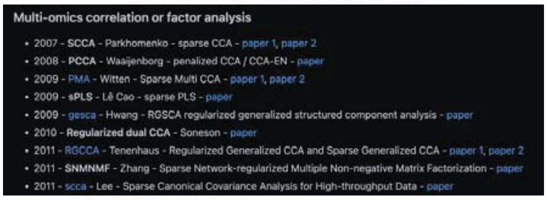
(most methods)

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

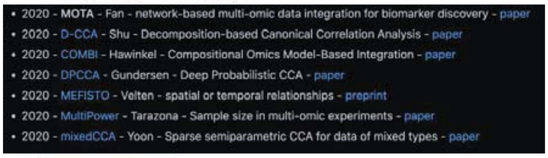
Factor analysis



Multi-omics correlation or factor analysis

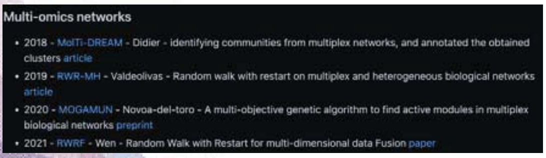
- 2007 - **SCCA** - Parkhomenko - sparse CCA - paper 1, paper 2
- 2008 - **PCCA** - Waaijenborg - penalized CCA / CCA-EN - paper
- 2009 - PMA - Witten - Sparse Multi CCA - paper 1, paper 2
- 2009 - **sPLS** - Lê Cao - sparse PLS - paper
- 2009 - gesca - Hwang - RGSCA regularized generalized structured 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

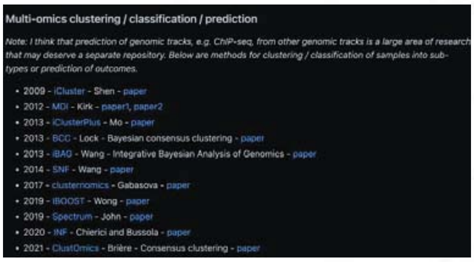
Network analysis



Multi-omics networks

- 2018 - MolTi-DREAM - Didier - identifying communities from multiplex networks, and annotated the obtained clusters article
- 2019 - RWR-MH - Valdeolivas - 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

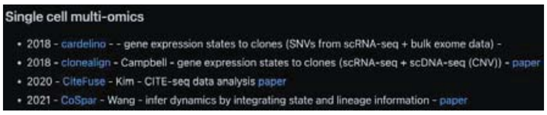


Multi-omics clustering / classification / prediction

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 - clusternomics - Gabasova - paper
- 2019 - iBOOST - Wong - paper
- 2019 - Spectrum - John - paper
- 2020 - INF - Chierici and Bussola - paper
- 2021 - ClustOmics - Brière - Consensus clustering - paper

Single-cell



Single cell multi-omics

- 2018 - cardelino - - 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…

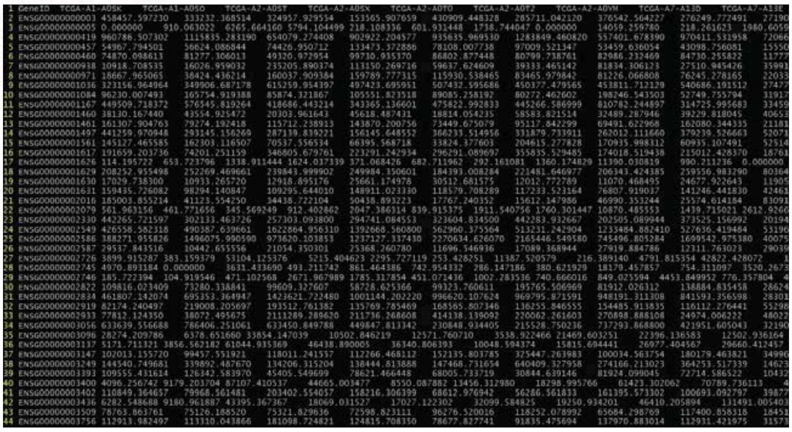
MO Factor analysis

- **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?

MO data characteristics

- MO data is highly complex, large and and heterogenous

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



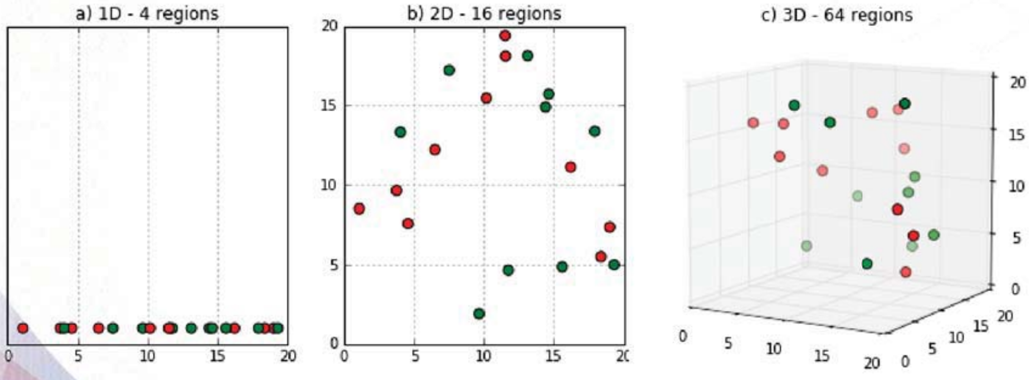
```
GeneID  TCGA-A1-A0SK    TCGA-A1-A0SO    TCGA-A2-A0ST    TCGA-A2-A0SX    TCGA-A2-A0T0    TCGA-A2-A0T2    TCGA-A2-A0YM    TCGA-A7-A13D    TCGA-A7-A13E
ENSG00000000003 458457.597230   333232.368514   324957.929554   153565.907659   430909.448328   285711.042120   376542.564227   276249.772491   27190
ENSG00000000005 0.000000    910.063022  6265.664160 5794.104499 218.108336  601.931448  1738.744047 0.000000    14059.259780    218.261623  1980.6059
ENSG00000000419 960786.507302   1115835.283190  654079.274408   902922.204577   935635.969530   1283849.460820  557401.678390   970411.531958   72066
ENSG00000000457 54967.794501    56624.086844    74426.950712    133473.372886   78108.007738    97009.521347    53459.636054    43098.756081    15550
ENSG00000000460 74870.098613    81277.306013    49320.972954    99710.935370    86802.877448    80799.738761    82986.232469    84730.255822    11777
ENSG00000000938 10918.708535    16026.959032    235205.890374   113150.269716   59637.624609    39333.465142    81834.306123    27510.945426    35991
ENSG00000000971 18667.965065    38424.436214    160037.909384   159789.777315   115930.538465   83465.979842    81226.066808    76245.278165    22033
ENSG00000001036 323156.964964   349906.687178   615259.954397   497423.695951   507432.595686   450377.479565   453811.712129   540686.191512   27477
ENSG00000001084 96230.007493    165754.919388   85874.321867    105551.823518   89085.258192    80272.462602    198246.543503   52749.755794    31917
ENSG00000001167 449509.718372   576545.819264   418686.443214   343365.136601   475822.992833   445266.586999   810782.244897   314725.995683   33459
ENSG00000001460 38130.167440    43554.925472    20303.961643    45618.487431    18814.054235    58583.821514    32489.287946    39229.818045    40653
ENSG00000001461 161307.904763   79274.192418    115712.238913   143870.200756   73449.675079    95117.842299    69491.622968    162080.344335   21181
ENSG00000001497 441259.970948   293145.156269   287139.839221   156145.648552   366233.514956   331879.733911   262012.111660   379239.526663   52071
ENSG00000001561 145127.465585   162303.116507   70537.556534    66395.568718    33824.377603    204615.277828   170935.998312   60935.107491    52514
ENSG00000001617 191659.203736   74201.251159    346805.679761   223291.242934   296291.089697   355835.529485   174018.519438   215012.428370   18761
ENSG00000001626 114.195722  653.723796  1338.911444 1624.037339 371.068426  682.711962  292.161081  1360.174829 11390.030819    990.211236  0.000000
ENSG00000001629 208252.955498   252269.469661   239843.999902   249984.350601   184393.008284   221481.646977   206343.424185   259556.983290   80364
ENSG00000001630 17029.738300    10933.265720    12918.895176    25661.174978    30512.681575    12012.772789    11070.468495    24677.922643    11905
ENSG00000001631 159435.726082   98294.140847    109295.644010   148911.023330   118579.708289   117233.523164   76807.919037    141246.441830   42461
ENSG00000002016 185003.855214   41123.554250    34438.722104    50438.893223    17767.240352    15612.147986    46990.353244    25574.614184    83091
ENSG00000002079 561.963156  461.771656  345.569249  912.402862  2047.386314 839.915375  1911.540756 1760.301447 10870.485553    1439.715021 2612.9266
ENSG00000002330 442265.721597   302133.463726   257303.093800   294741.084553   323604.834500   442283.932667   202505.089944   373525.156992   20194
ENSG00000002549 426558.582318   490387.639661   1622864.956310  1392668.560800  562960.375564   513231.242904   1233484.882410  527636.419484   53196
ENSG00000002586 188271.955826   1496075.990590  973620.103853   1237127.337430  2270634.626070  2165446.549580  745496.805284   1699542.975380  40075
ENSG00000002587 29537.843516    10442.655556    21054.350301    25368.260780    11696.546936    17089.368944    27919.884786    12311.763023    29039
ENSG00000002726 3899.915287 383.159379  53104.125376    5215.404623 2295.727119 253.428251  11387.520579    216.389140  4791.815354 42822.428072    1
ENSG00000002745 4970.893184 0.000000    3631.433690 493.211742  861.464386  742.954332  286.147186  380.621929  18179.457857    754.311097  3520.2673
ENSG00000002746 185.722394  104.919546  471.102568  2671.967989 1785.317854 451.071436  1002.283536 740.666016  849.025594  4453.849952 776.357804  4
ENSG00000002822 109816.023409   73280.338841    99609.327607    58728.625366    99323.760611    195765.506969   81912.026312    138884.835458   28624
ENSG00000002834 461807.142074   695353.364947   1423621.722480  1001144.202220  996620.107624   969795.871591   948191.311308   841593.356598   28301
ENSG00000002919 82174.240497    219008.205697   193512.761382   135769.785469   168565.807346   136255.846555   154485.913835   116112.276441   55298
ENSG00000002933 77812.124350    38072.495675    2111289.289620  211736.268608   414138.139092   220062.261603   270898.888108   24974.006222    48022
ENSG00000003056 633639.556688   786406.251061   633450.849788   449847.813342   230848.934405   215528.750236   737293.868800   421951.605043   32190
ENSG00000003096 28274.209786    6378.651660 33854.147039    10502.846219    12571.760710    3538.922466 21469.603251    22396.136583    12502.936164
ENSG00000003137 5171.711321 3856.562182 61044.935369    46438.890005    36340.806393    10048.594374    15815.694441    26977.404567    29660.412457
ENSG00000003147 102013.155720   99457.551921    118011.241557   112266.468112   152135.803785   325447.263983   100034.563754   180179.463821   34996
ENSG00000003249 144540.749681   339892.487670   134206.315204   138444.813888   147468.731654   640409.327958   274166.213023   364253.517339   14623
ENSG00000003393 109555.431614   126342.583970   45405.549699    78621.466448    68005.733719    30844.639146    81924.099045    72714.586522    10432
ENSG00000003400 4096.256742 9179.203704 87107.410537    44665.003477    8550.087882 13456.312980    18298.995766    61423.302062    70789.736113    4
ENSG00000003402 110849.364657   79968.561481    203402.554057   158216.306399   68612.976942    56286.561833    161395.573302   100693.092797   39877
ENSG00000003436 6282.548688 9180.961887 43395.367367    18069.031527    17027.122302    32099.584825    19250.934201    46410.205894    131491.005403
ENSG00000003509 78763.863761    75126.188520    75321.829636    72598.823111    96276.520016    118252.078992   65684.298769    117400.858318   18451
ENSG00000003756 112913.982497   113310.043866   181098.724821   124815.708150   78677.827741    91835.475694    137970.883014   112931.421975   31573
```

> 15K genes

> 600 patients

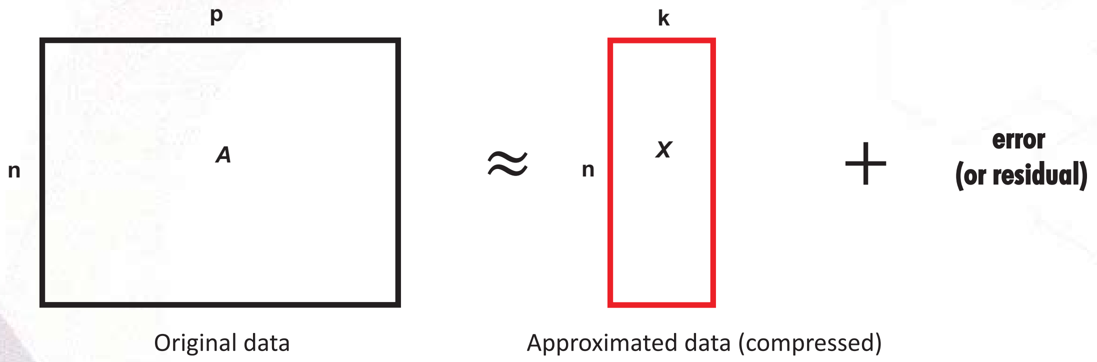
The Curse of Dimensionality

- 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!

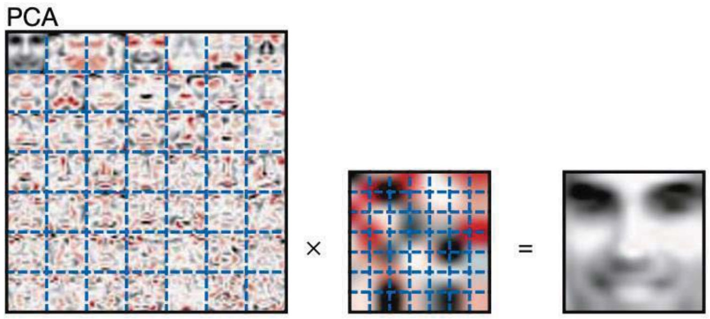


A brief overview of factor analysis

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



PCA example

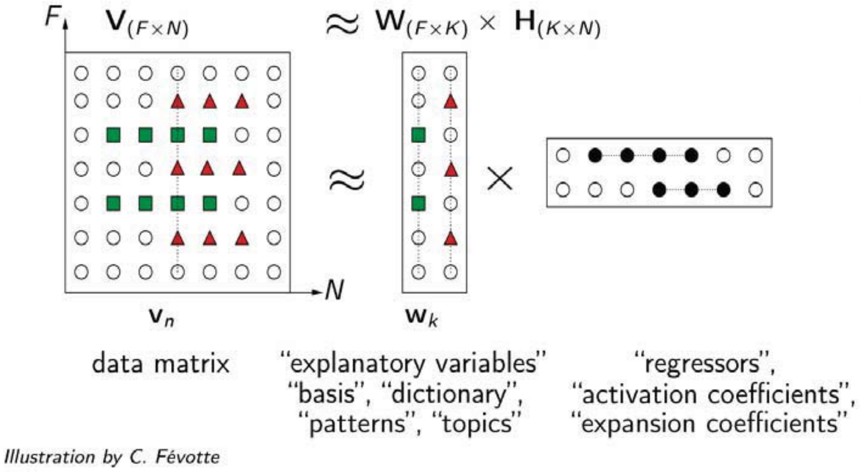


Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401.6755 (1999): 788-791.

A brief overview of factor analysis

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

- NMF concept



F
$\mathbf{V}_{(F\times N)} \approx \mathbf{W}_{(F\times K)} \times \mathbf{H}_{(K\times N)}$
$\approx$
$\times$
N
$\mathbf{v}_n$
$\mathbf{w}_k$

data matrix

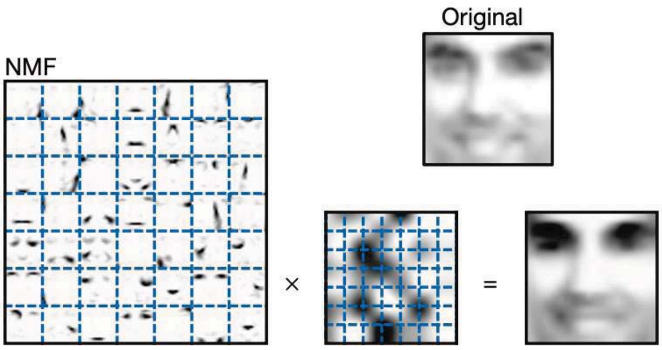
"explanatory variables" "basis", "dictionary", "patterns", "topics"

"regressors", "activation coefficients", "expansion coefficients"

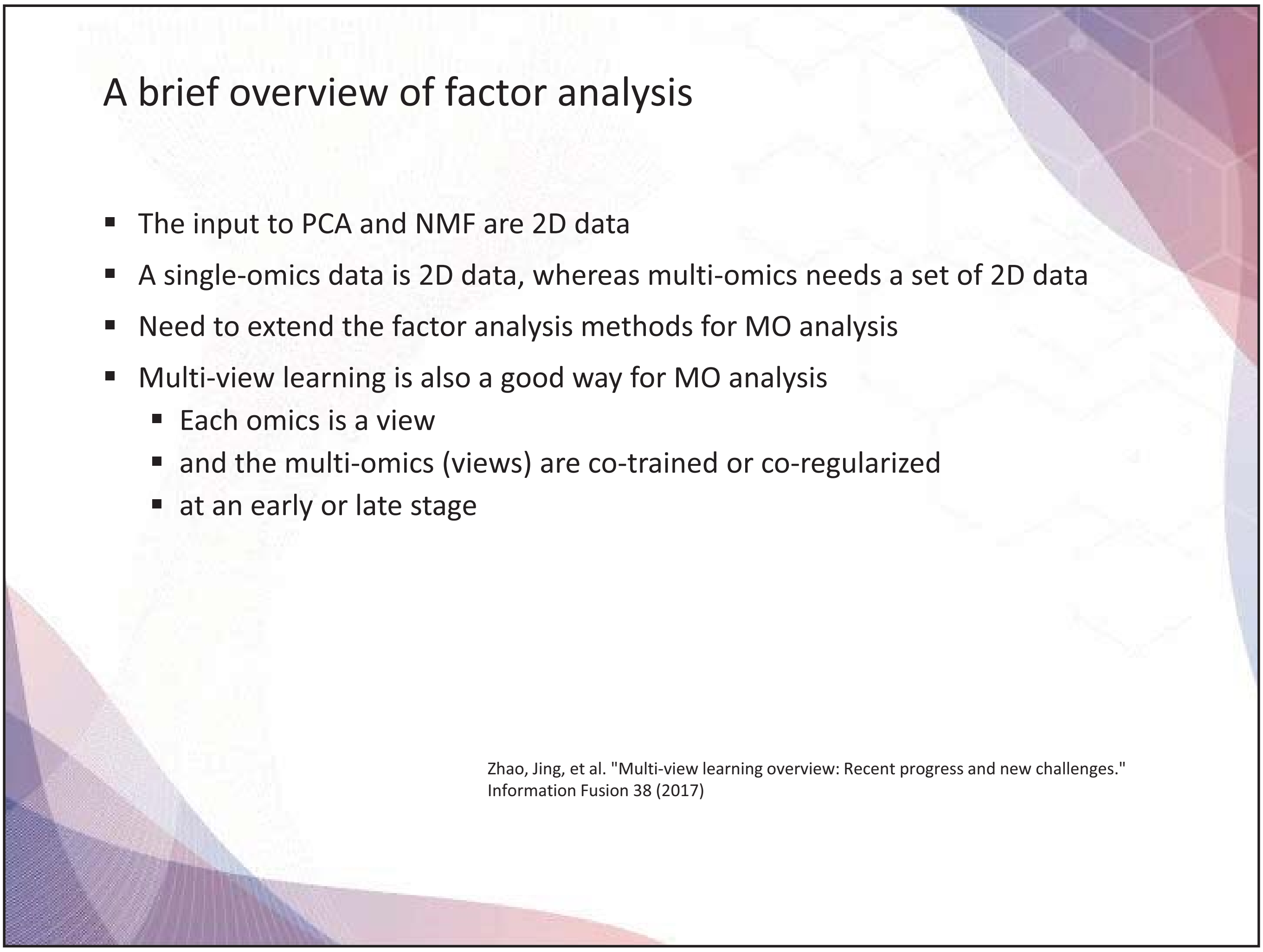
Illustration by C. Févotte

NMF example

- NMF



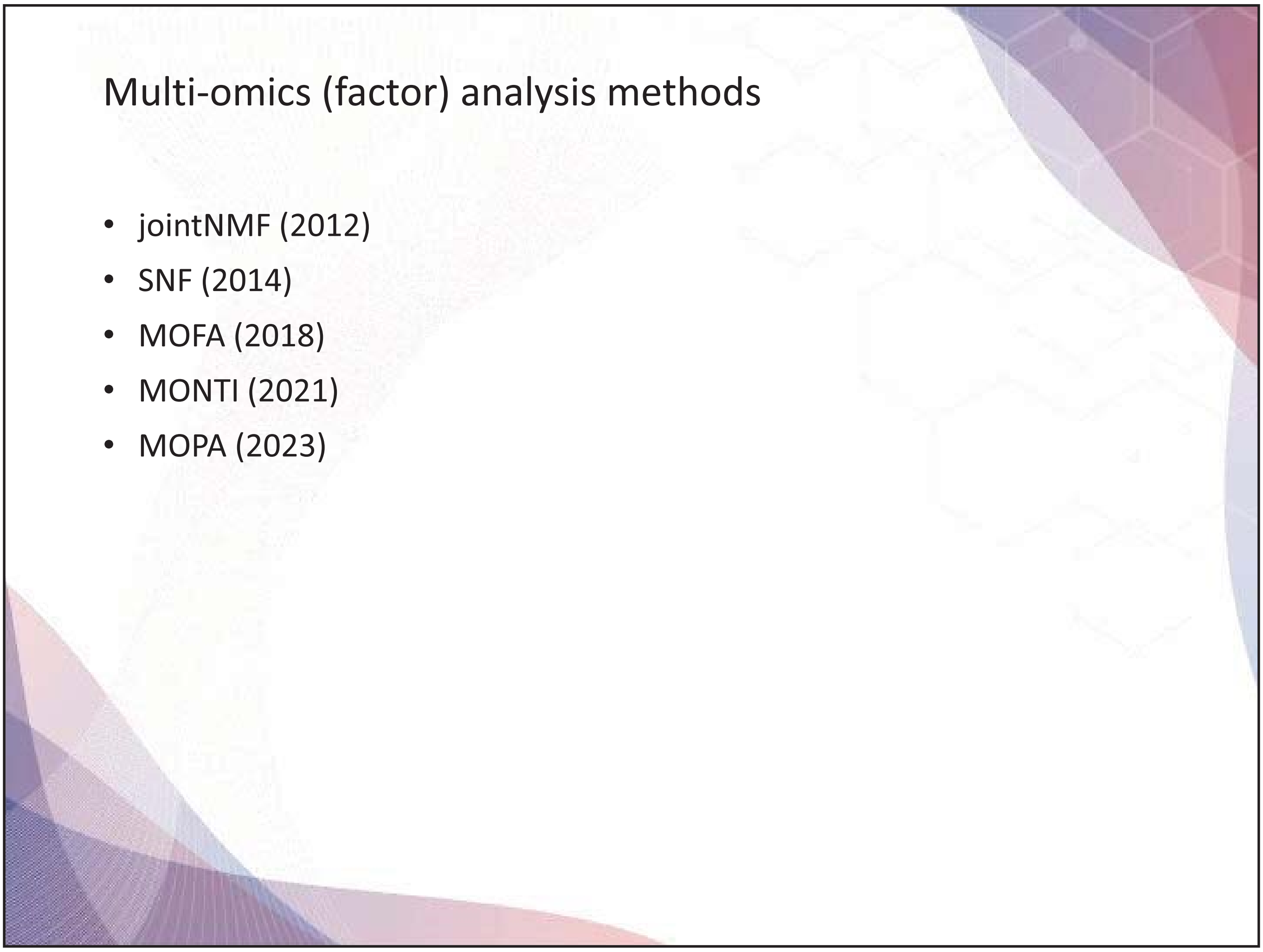
Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401.6755 (1999): 788-791.



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)

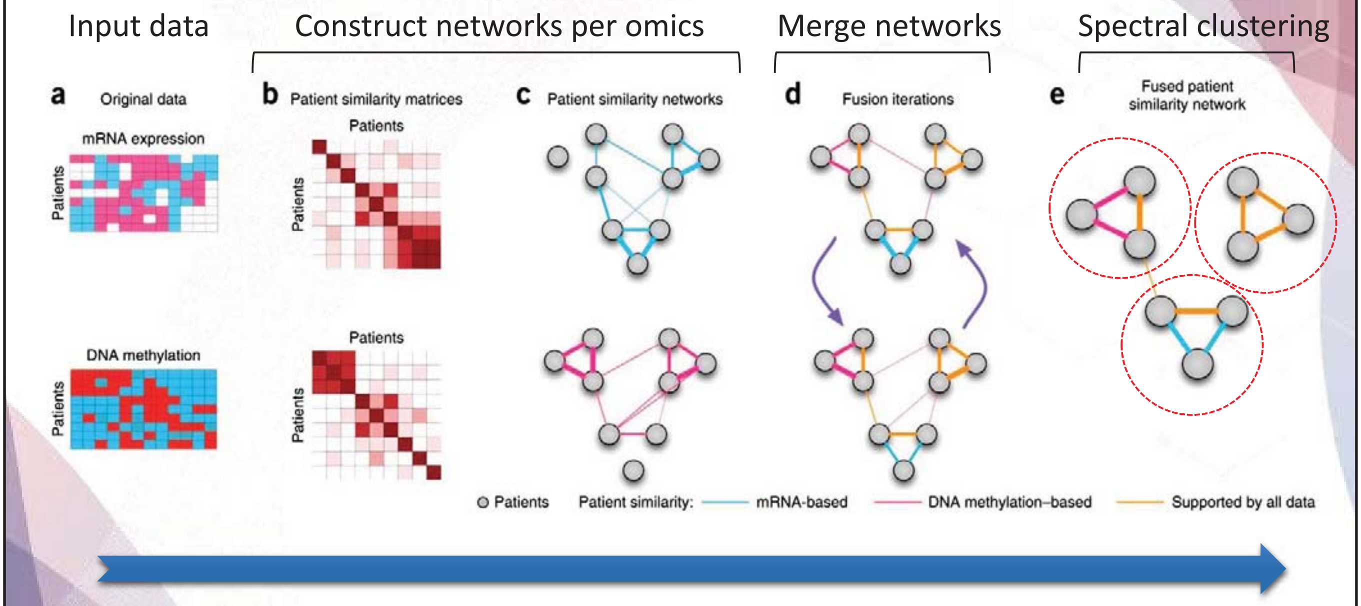
Methods | SNF – Similarity Network Fusion

- 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



Input data

Construct networks per omics

Merge networks

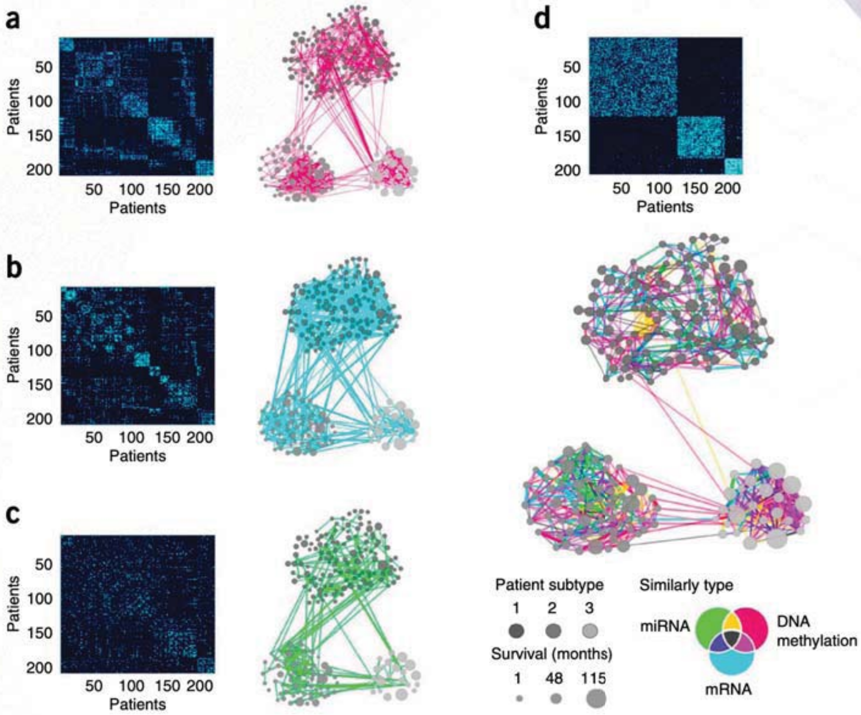
Spectral clustering

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)

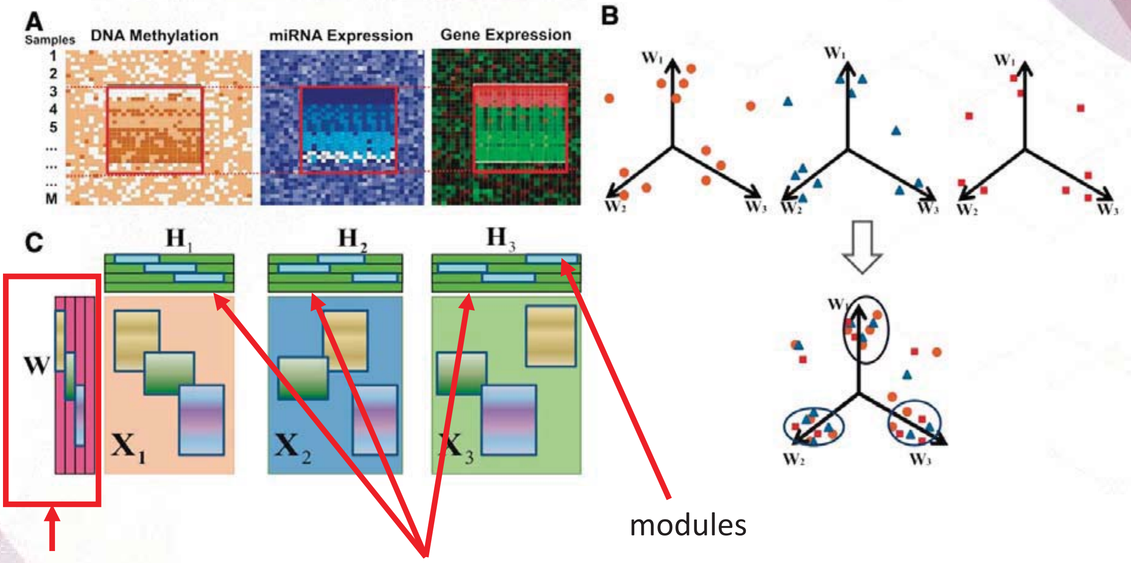
- Advantage:
  - Reports a set of important omics-features
- Disadvantage:
  - Number of ranks is difficult to determine
  - Can become slow and require large memory with large samples and many features
  - May not work well with small no. of samples (constrained by the ranks)

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)

$M \times N_1$ $M \times N_2$ $M \times N_3$



modules

Common basis matrix

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

Omics specific coefficient matrices

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

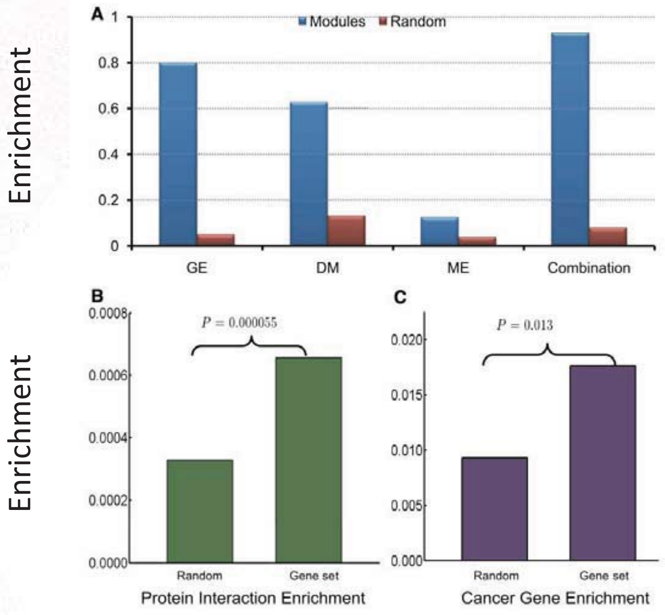
$X_2 \sim W \times H_2 \quad \ldots$

$X_3 \sim W \times H_3$

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)



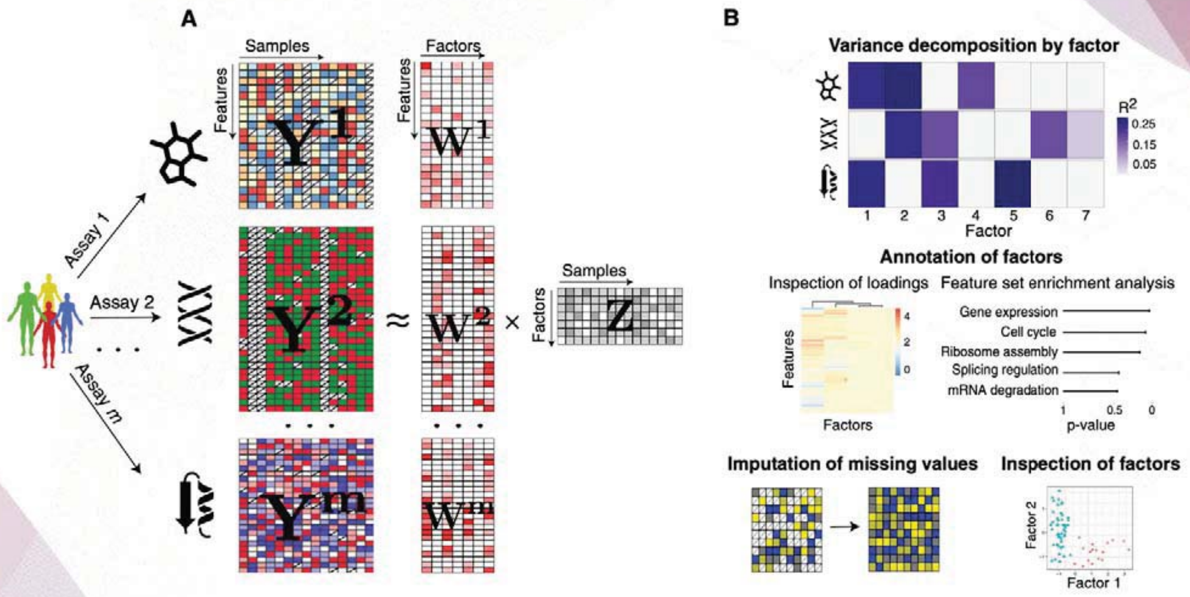
Methods | MOFA – Multi-Omics Factor Analysis

- 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)

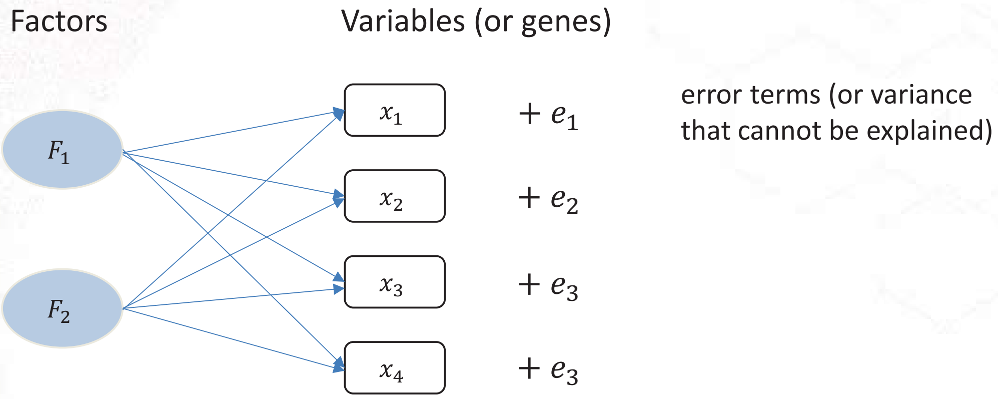
Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." *Molecular systems biology* 14.6 (2018): e8124.

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

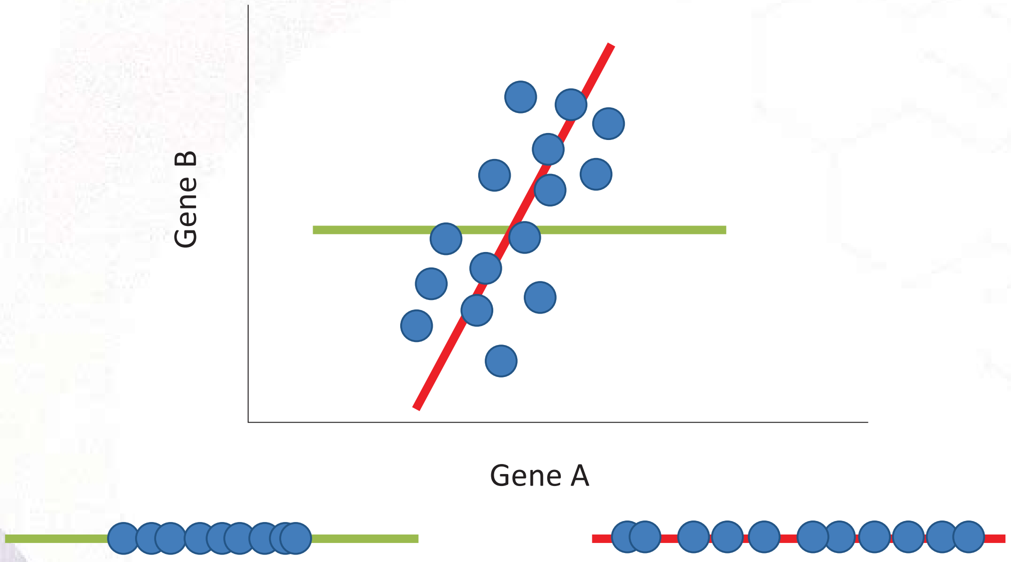


$$x_1 = b_{11} \times F_1 + b_{12} \times F_2 + e_1$$

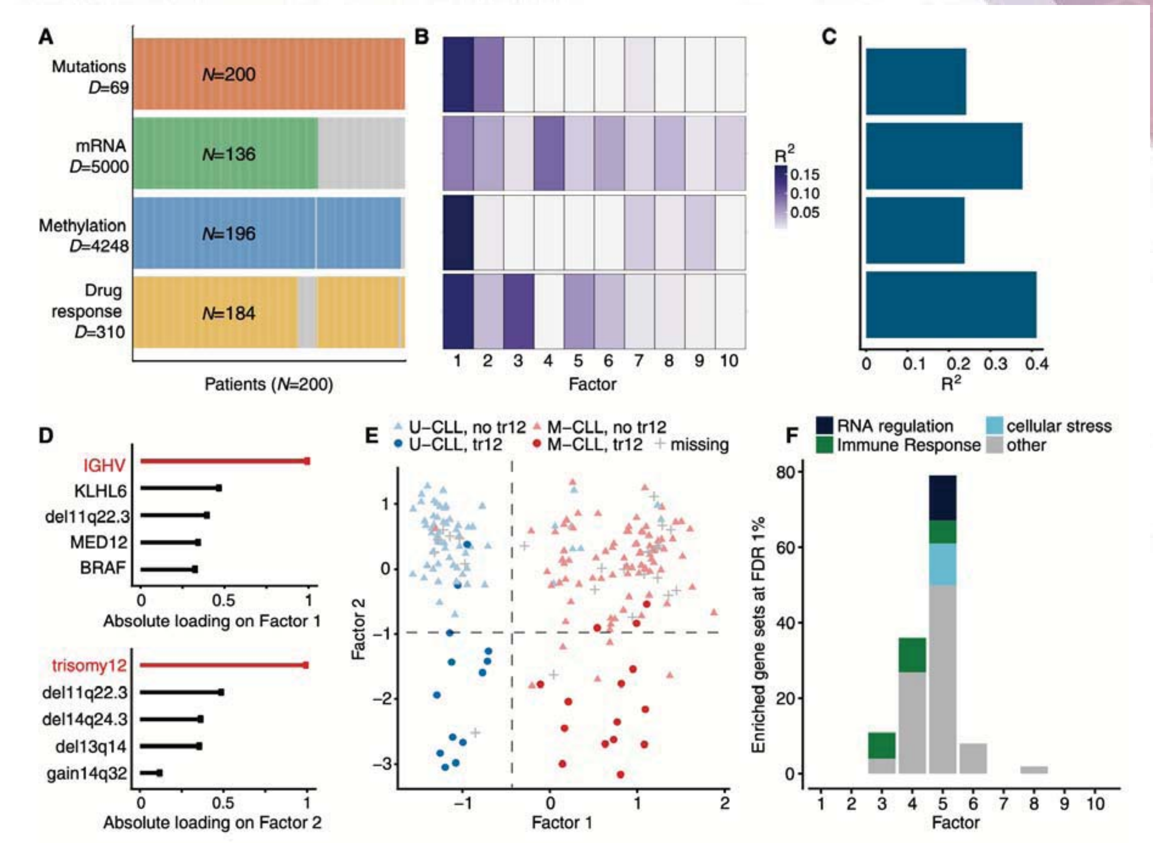
$$F_1 = w_{11} \times x_1 + w_{12} \times x_2 + w_{13} \times x_3 + w_{14} \times x_4$$

Brief overview of PCA

- Finds the line that maximizes variance in the data

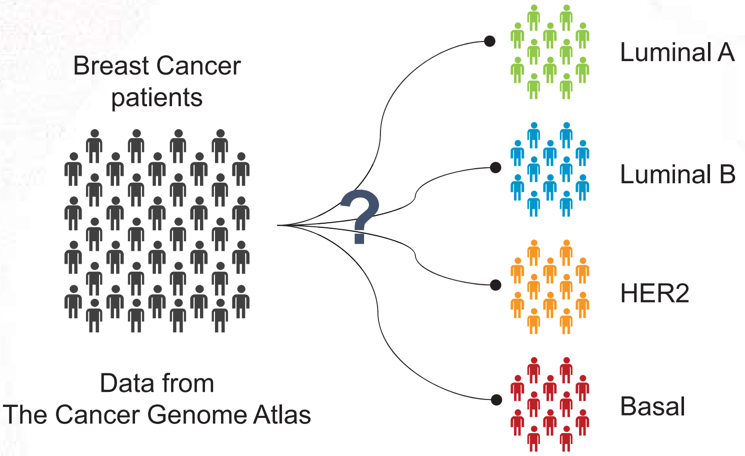


Methods | MOFA – Results



Multi-omics research

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



Breast Cancer patients

Data from
The Cancer Genome Atlas

Luminal A

Luminal B

HER2

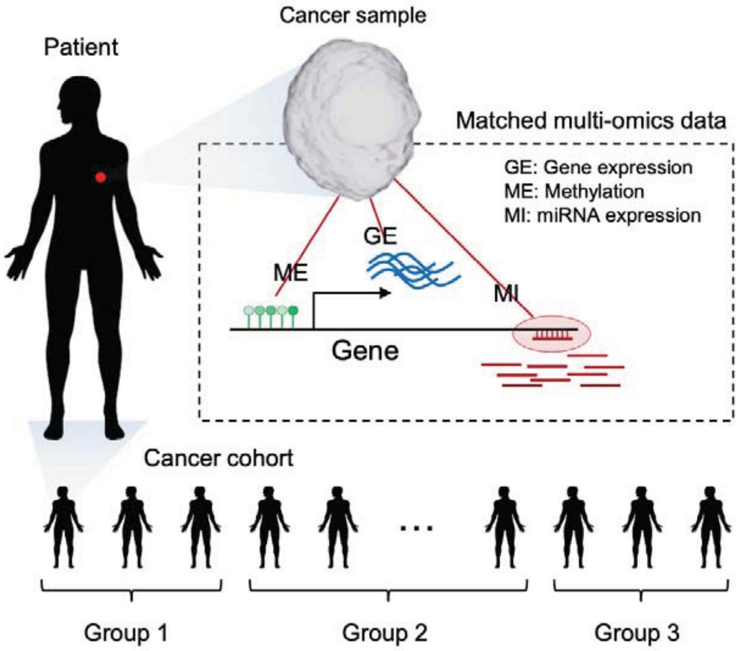
Basal

Some recent research topics on MO

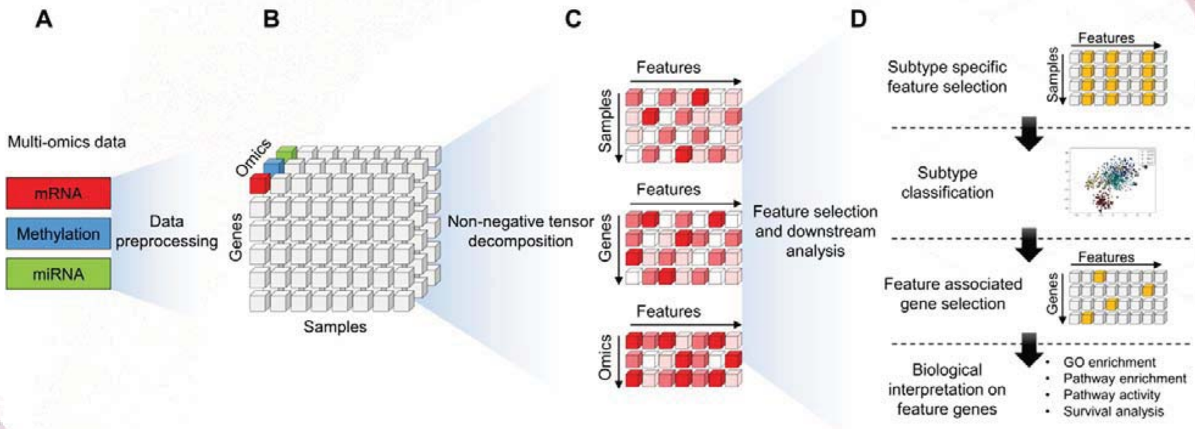
- MONTI: A multi-omics non-negative tensor decomposition framework for the integrated analysis of cancer subtypes (Frontiers in Genetics, 2021)
- MOPA: An Integrative Multi-Omics Pathway Analysis Method for Measuring Omics Activity (in preparation)
- Parametric analysis on large-scale multi-omics data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

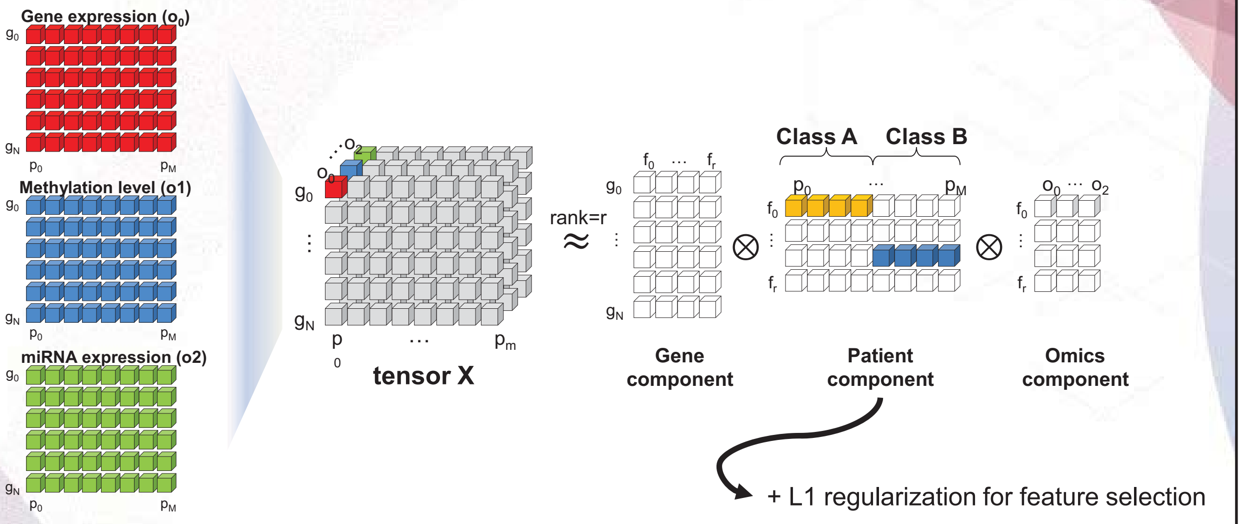


Methods| MONTI Workflow

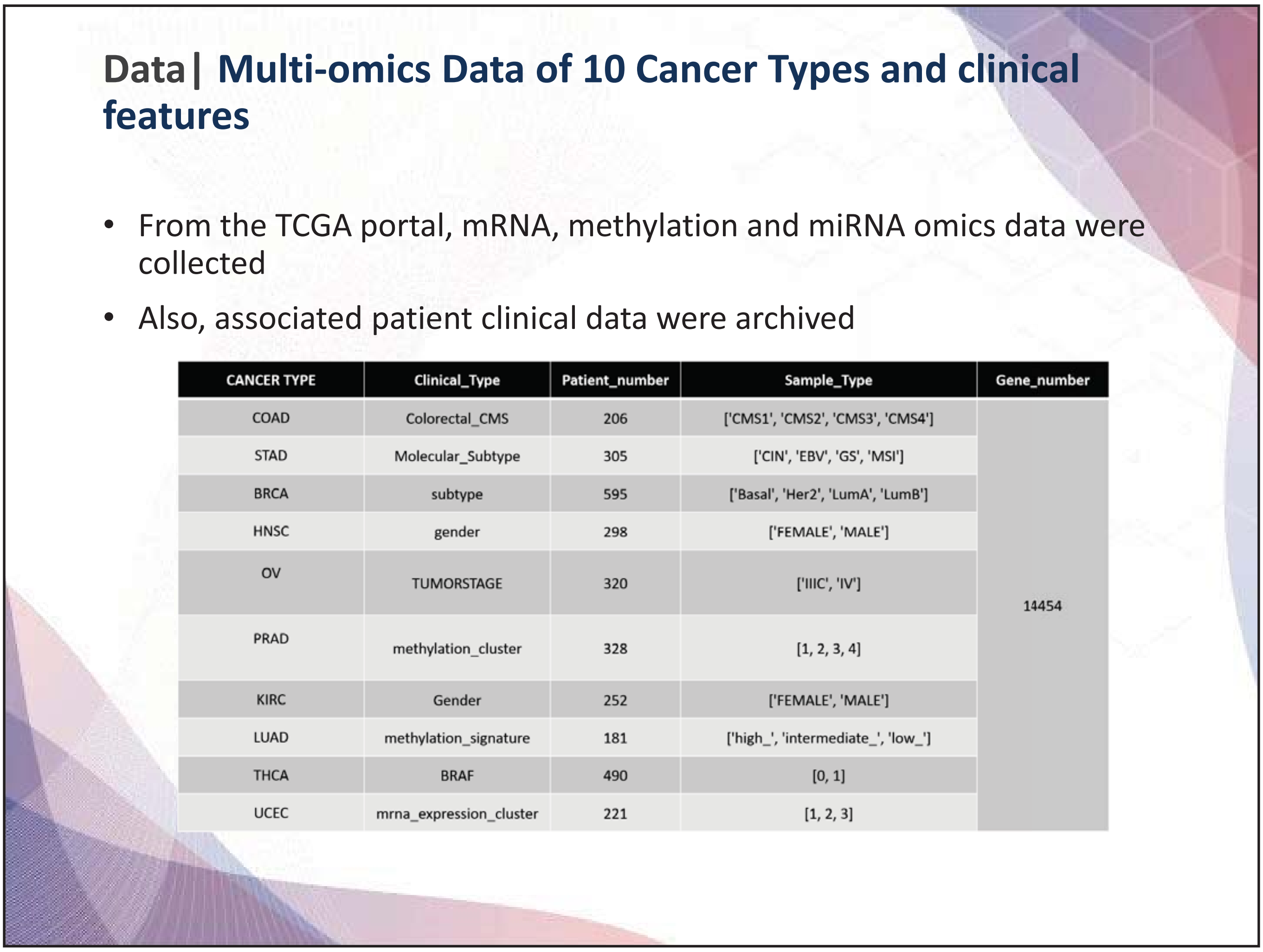


Jung, Inuk, et al. "MONTI: A Multi-Omics Non-negative Tensor Decomposition Framework for Gene-Level Integrative Analysis." Frontiers in Genetics 12 (2021).

Methods| Non-negative Tensor decomposition



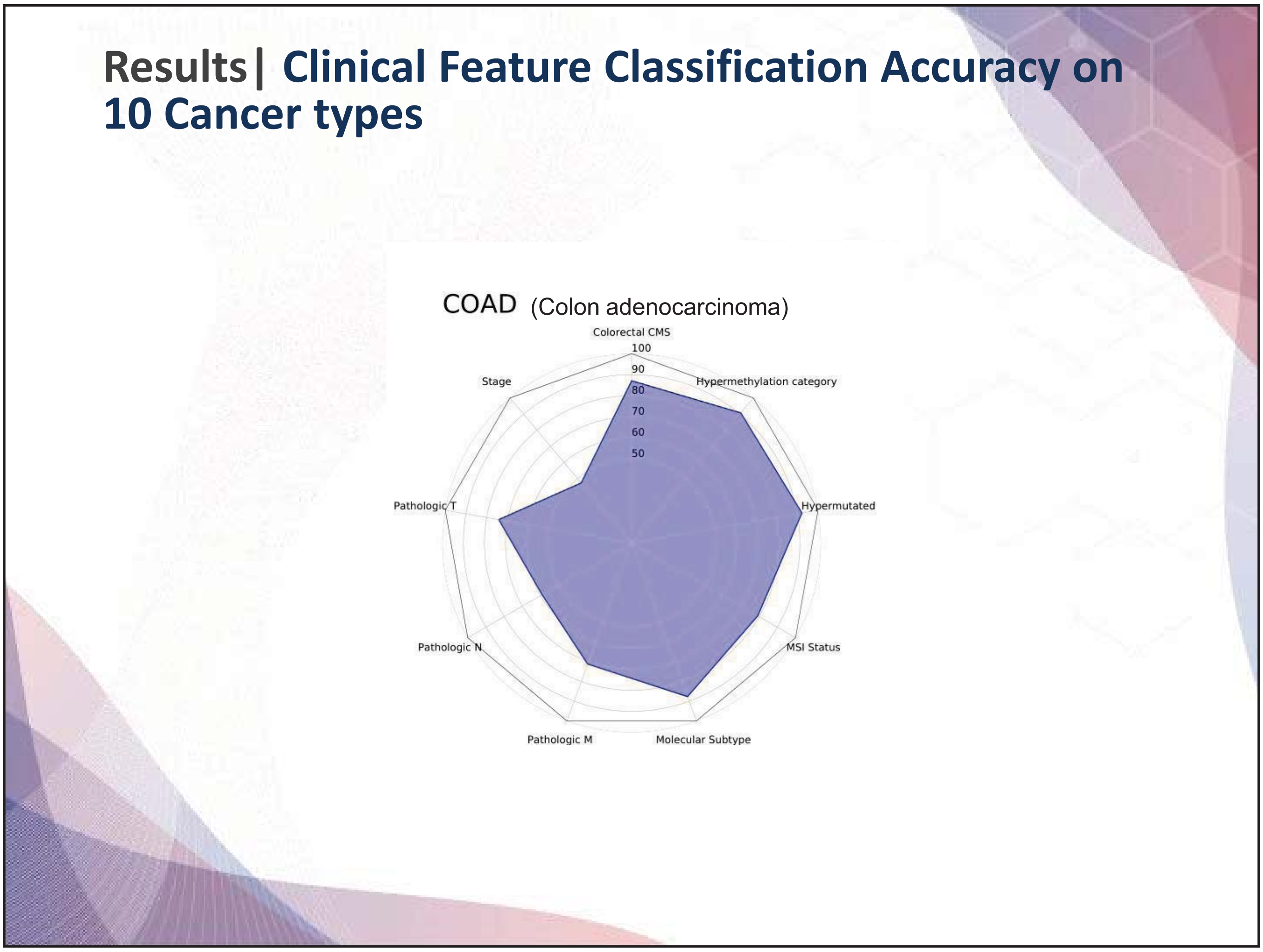
+ L1 regularization for feature selection



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]	

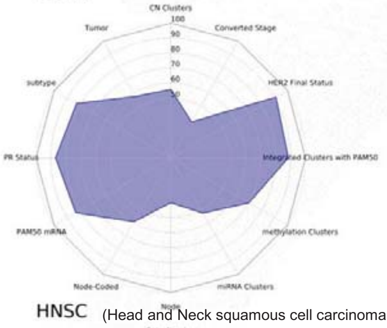


Results| Clinical Feature Classification Accuracy on 10 Cancer types

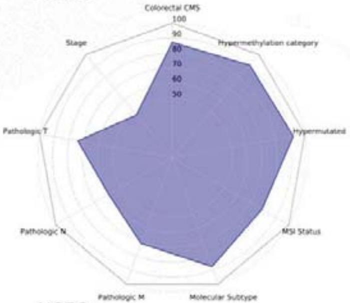
COAD (Colon adenocarcinoma)

Results| Clinical Feature Classification Accuracy on 10 Cancer types

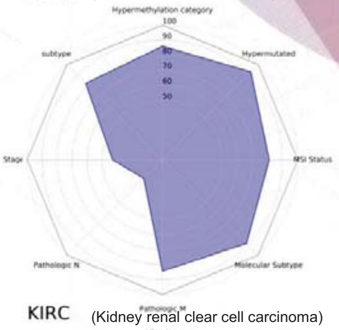
BRCA (Breast invasive carcinoma)



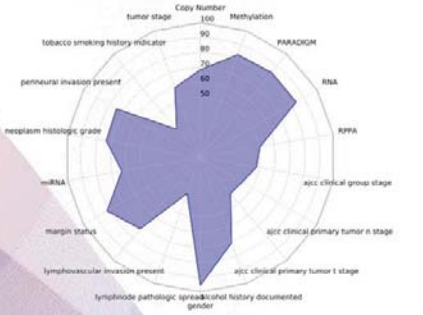
COAD (Colon adenocarcinoma)



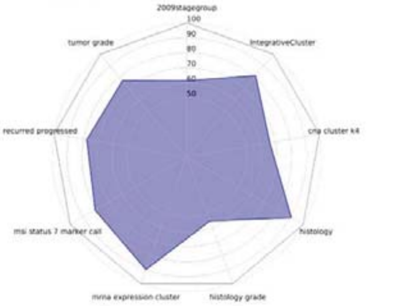
STAD (Stomach adenocarcinoma)



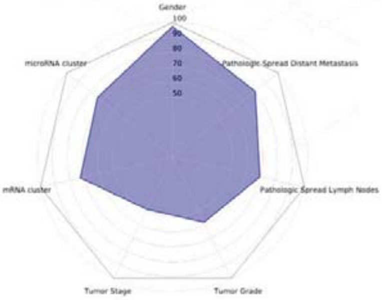
HNSC (Head and Neck squamous cell carcinoma)



UCEC (Uterine Corpus Endometrial Carcinoma)

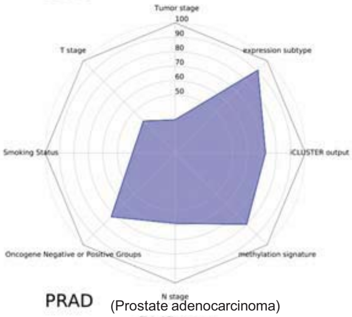


KIRC (Kidney renal clear cell carcinoma)

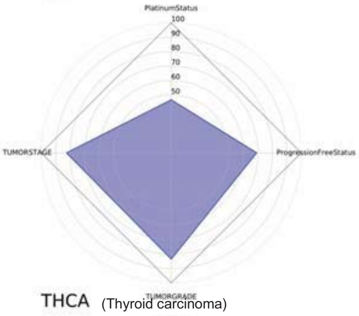


Results| Clinical Feature Classification Accuracy on 10 Cancer types

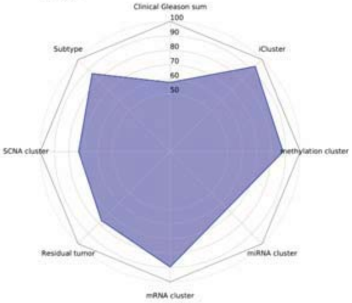
LUAD (Lung adenocarcinoma)



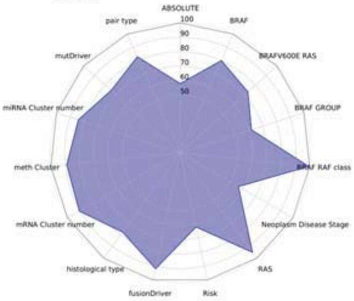
OV (Ovarian serous cystadenocarcinoma)



PRAD (Prostate adenocarcinoma)

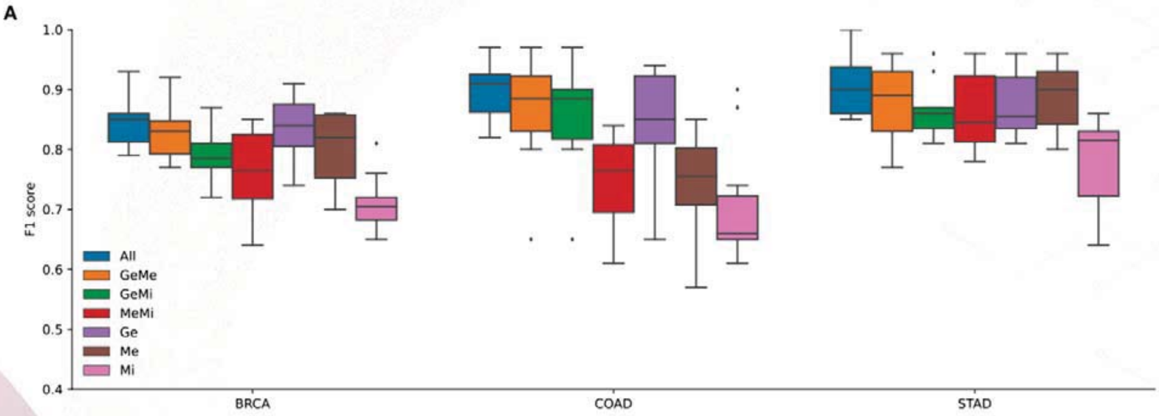


THCA (Thyroid carcinoma)



Results| Omics combinations

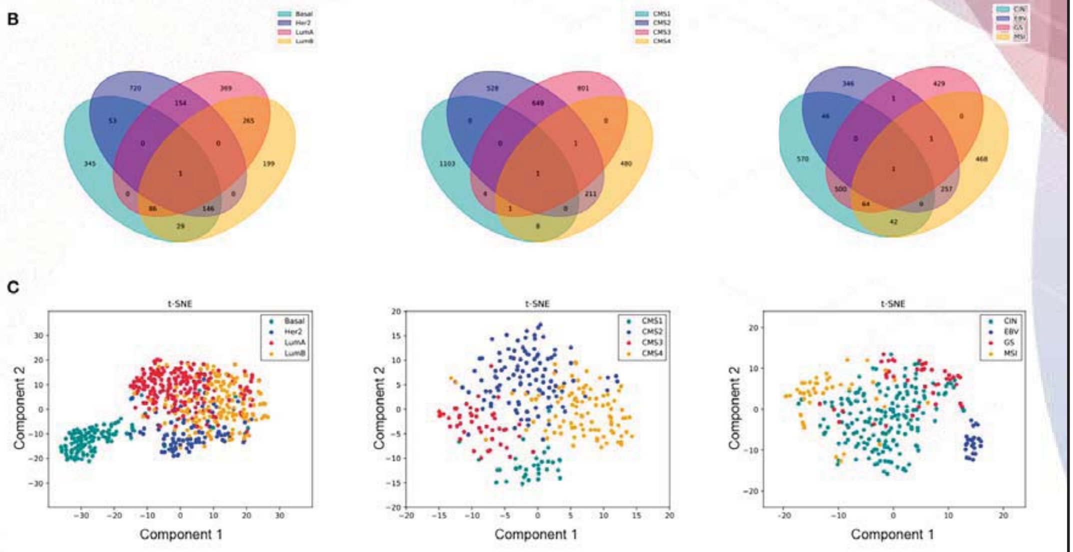
- Different combinations show different results

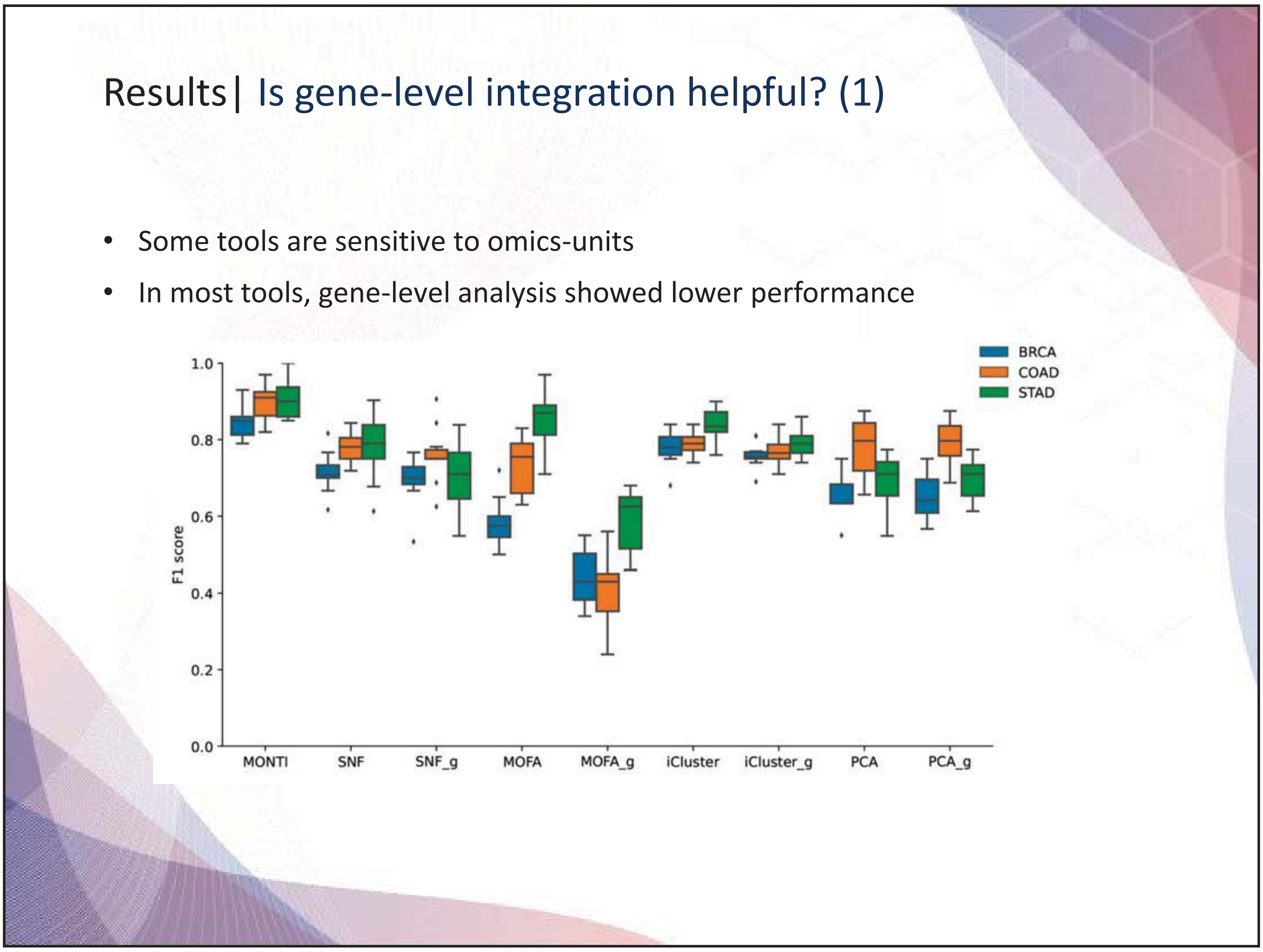


Results| Subtype features

- Some features are shared across the subtypes
- The subtypes of BRCA, COAD and STAD are well separated

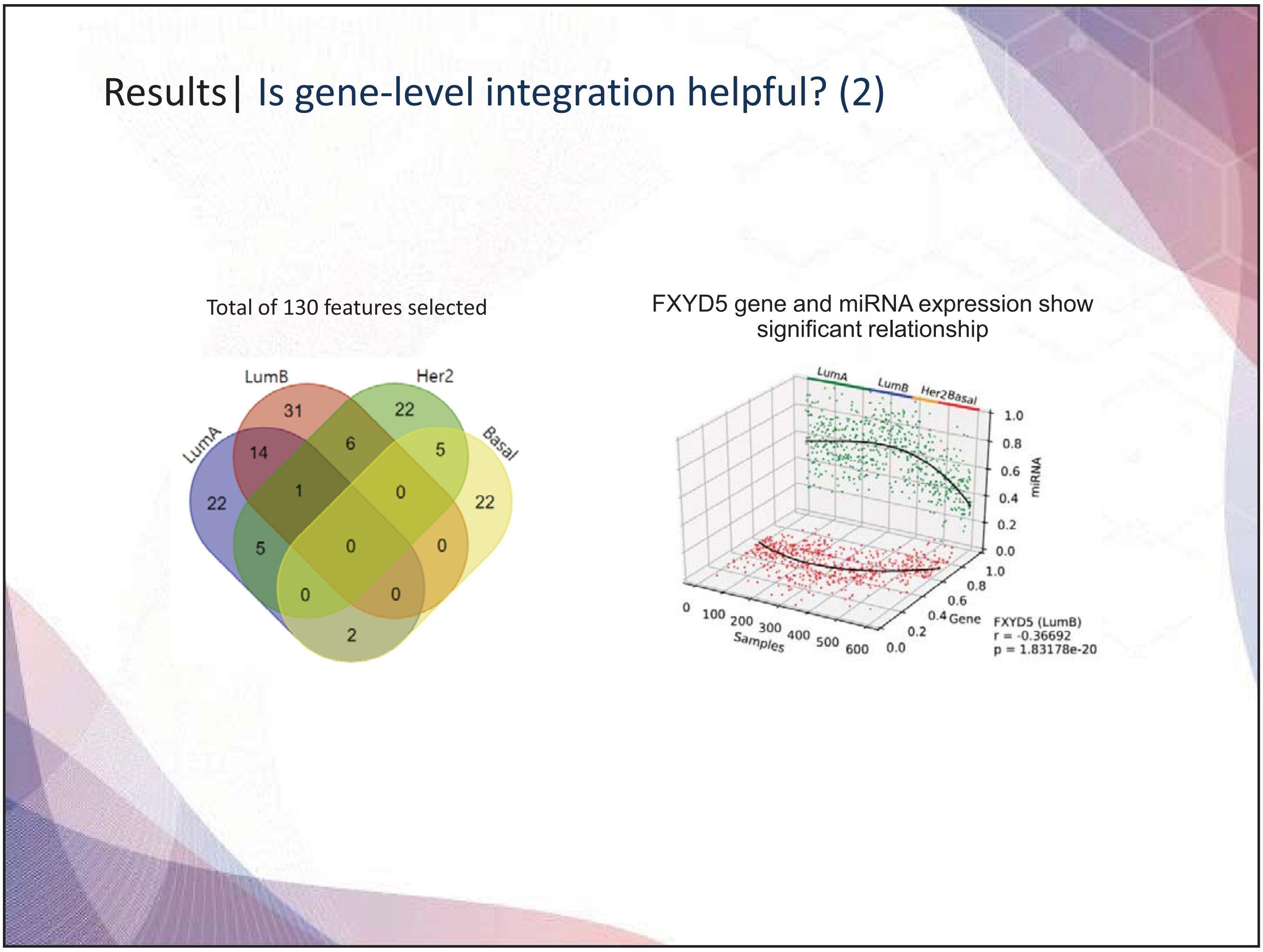
BRCA COAD STAD





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

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

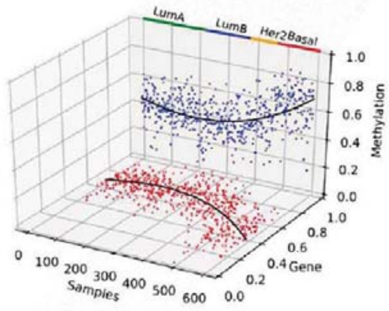


Results| Is gene-level integration helpful? (2)

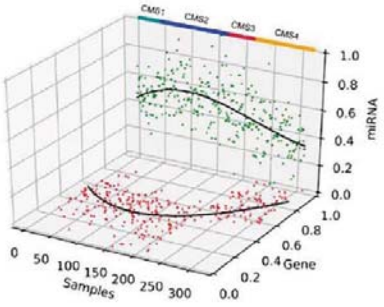
Total of 130 features selected

FXYD5 gene and miRNA expression show significant relationship

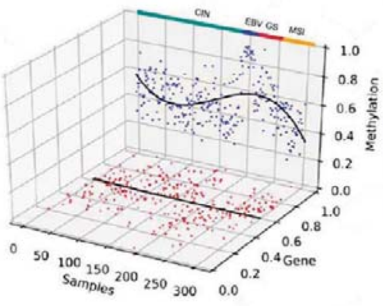
Results| Is gene-level integration helpful? (3)



EXOC6 (Basal)

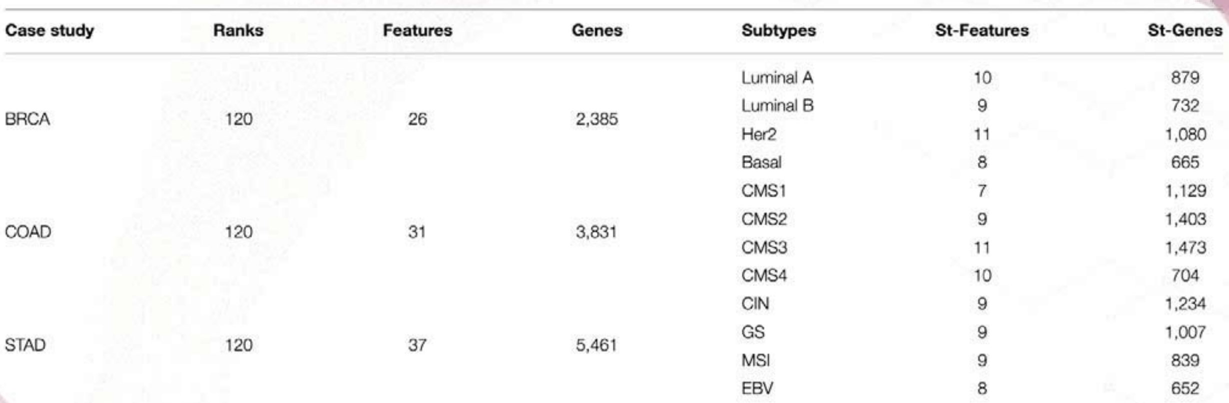


OLFML2B (CMS4)



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

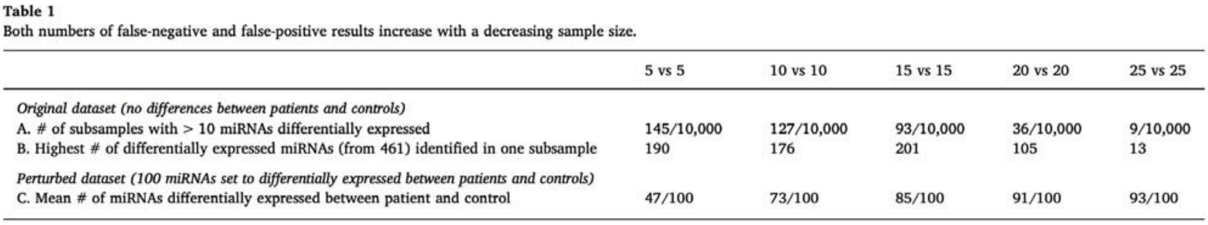
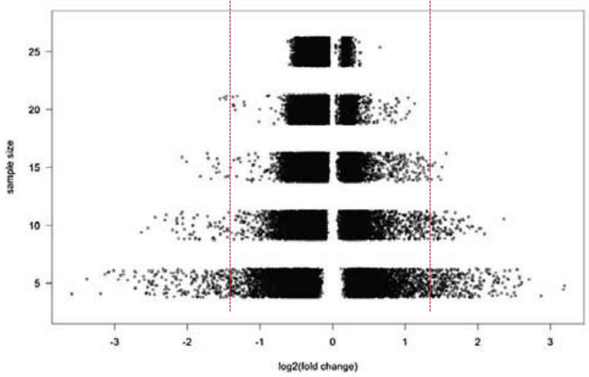


Table 1
Both numbers of false-negative and false-positive results increase with a decreasing sample size.

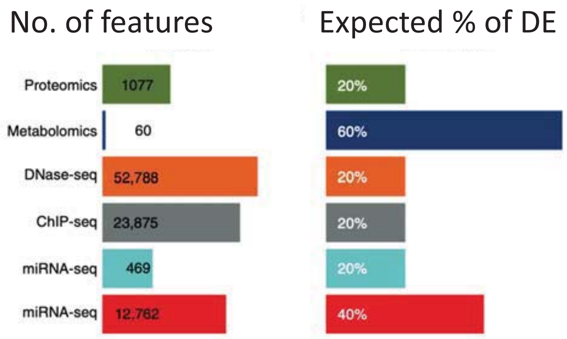
	5 vs 5	10 vs 10	15 vs 15	20 vs 20	25 vs 25
Original dataset (no differences between patients and controls)					
A. # of subsamples with > 10 miRNAs differentially expressed	145/10,000	127/10,000	93/10,000	36/10,000	9/10,000
B. Highest # of differentially expressed miRNAs (from 461) identified in one 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

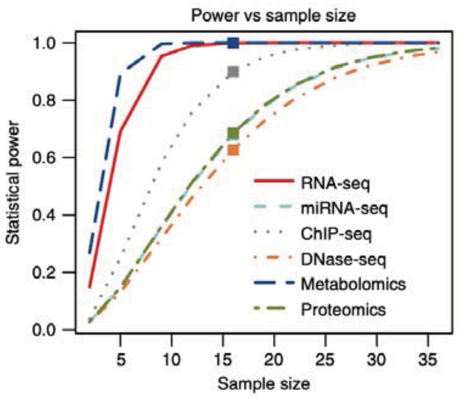


Kok, M. G. M., et al. "Small sample sizes in high-throughput miRNA screens: a common pitfall for the identification of miRNA biomarkers." *Biomolecular detection and quantification* 15 (2018)

Sample size estimation – multi-omics

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





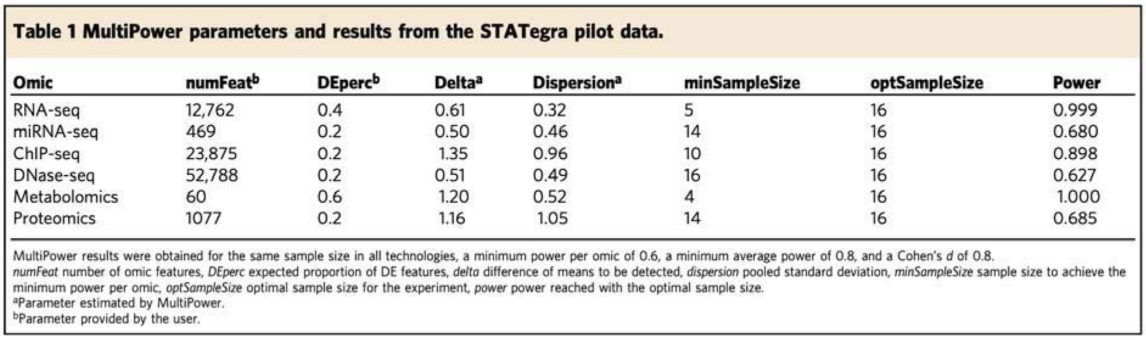


Table 1 MultiPower parameters and results from the STATegra pilot data.

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.

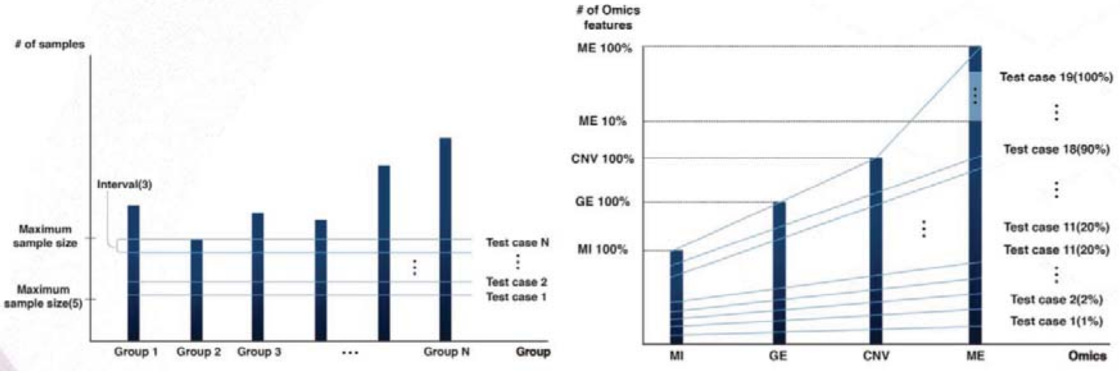
[a] Parameter estimated by MultiPower.

[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

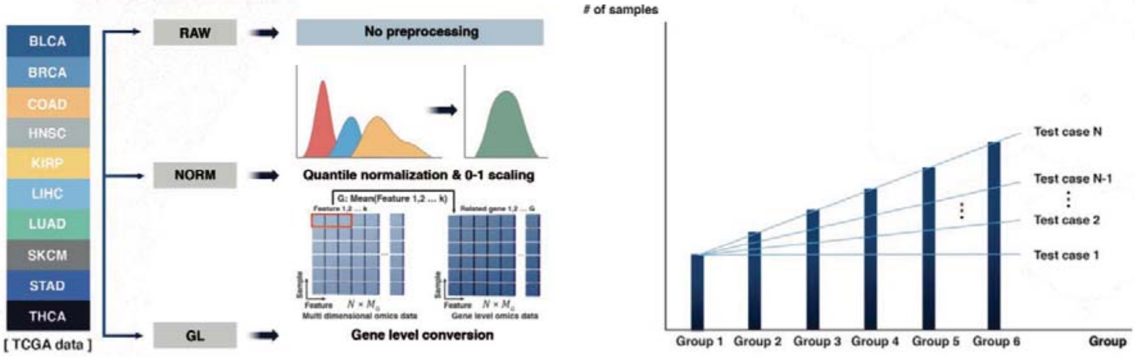


(a) Sample size test

(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



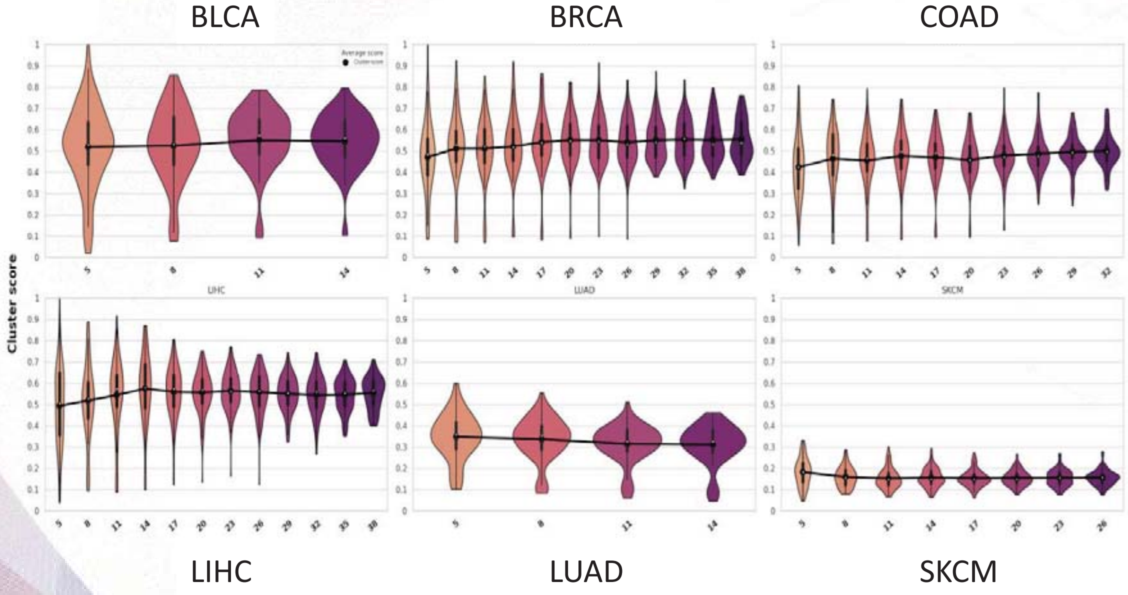
(c) Preprocessing test

(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

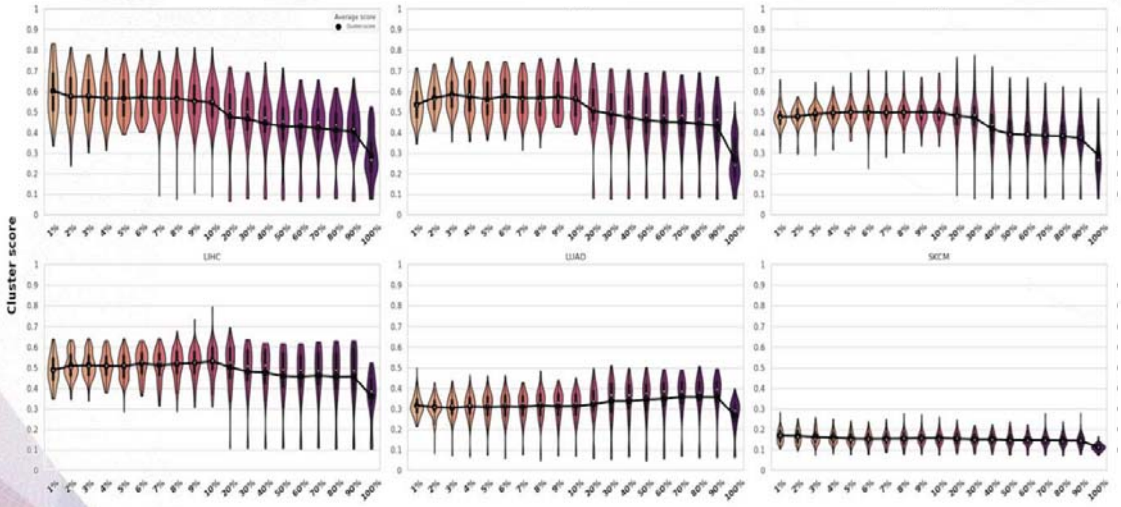


BLCA - Bladder Urothelial Carcinoma
BRCA - Breast invasive carcinoma
COAD - Colon adenocarcinoma
LIHC - Liver hepatocellular carcinoma
LUAD - Lung adenocarcinoma
SKCM - Skin Cutaneous Melanoma

Benchmark tests | Preliminary results – Feature numbers

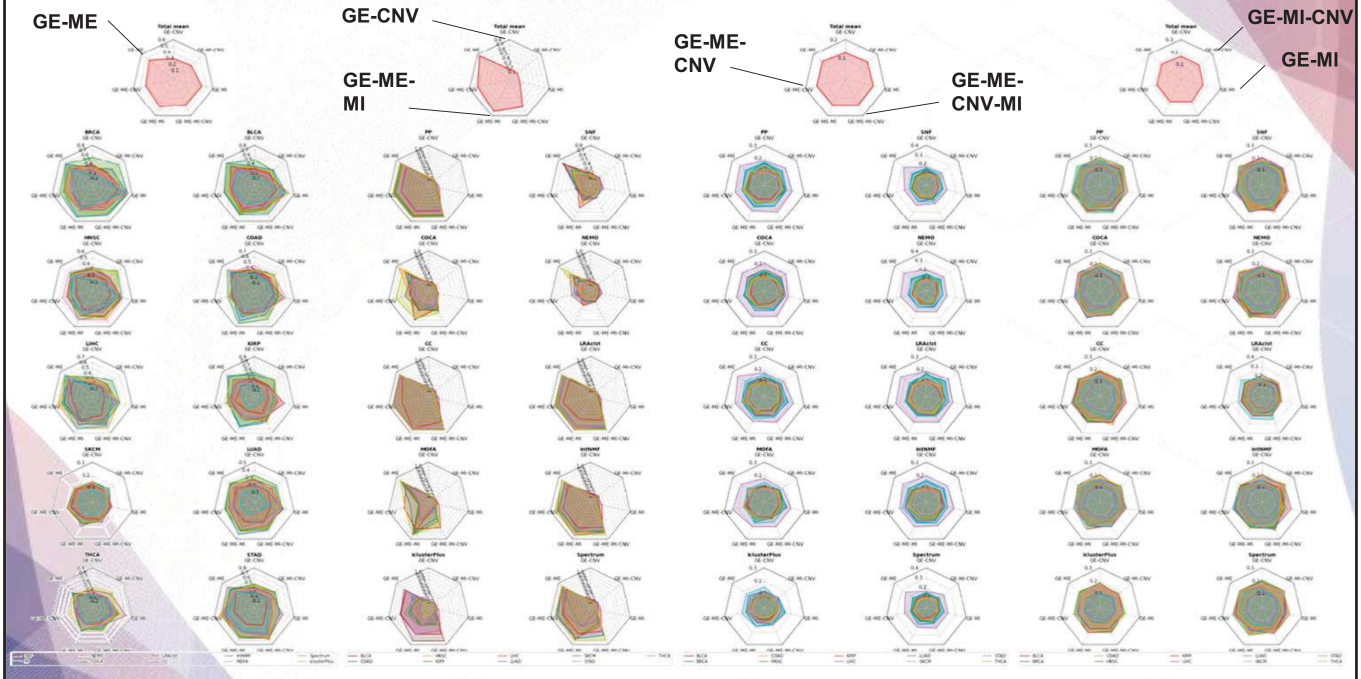
- The performance started to decline with # of features > 10~20%
  - Feature selection is important

BLCA BRCA COAD



LIHC LUAD SKCM

Benchmark tests | Preliminary results – Omics combinations

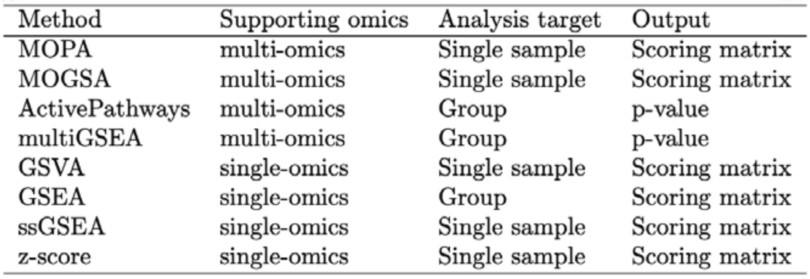


(a) Subtype (b) Gender (c) Pathological stage (d) Age

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

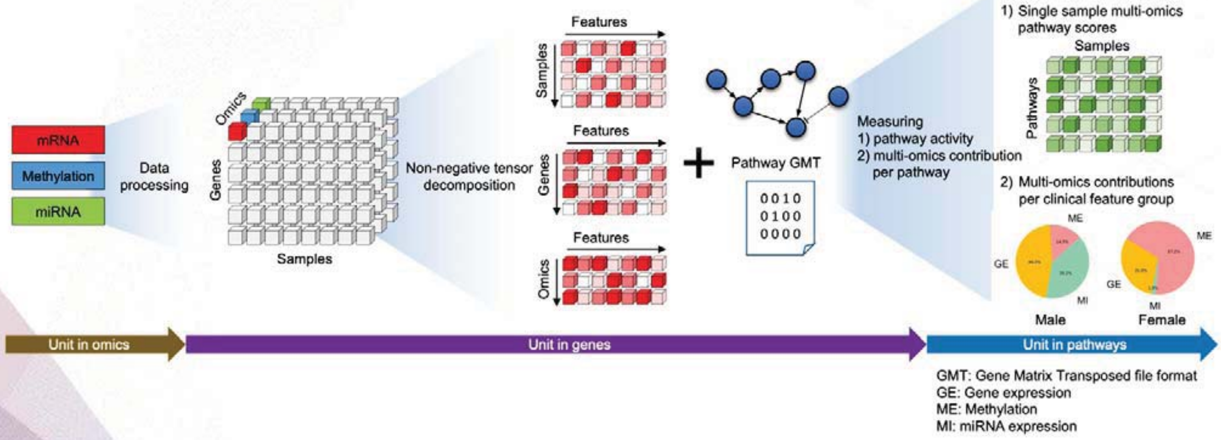
  → biological interpretation needs further work on the given results
- A simple list of genes may not be enough for such purpose
  - Especially, since the genes are selected from MO data (i.e., if genes selected from GE, we can perform GSVA or SSSGEA)
- **A list of multi-omics pathway analysis methods**



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

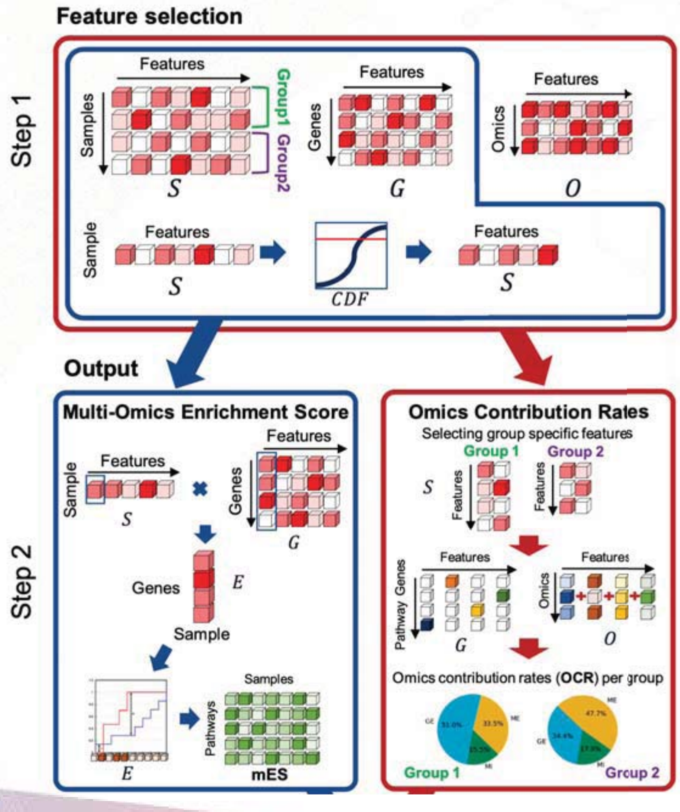
- MOPA is a tool that scores pathway activity based on MO data for each sample and each pathway
- The framework is very similar to GSVA or SSGEA but extended to consider MO data



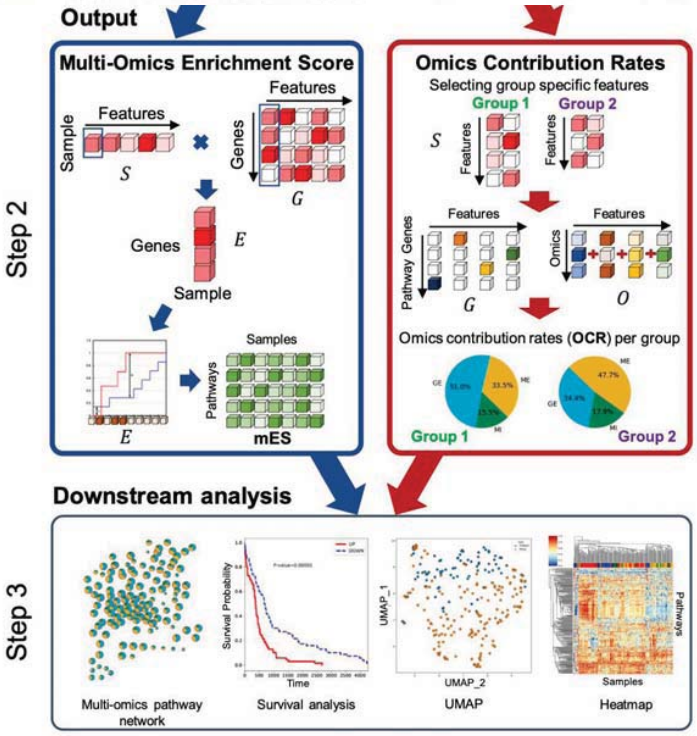
Objective| mES and OCR metrics

- 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)

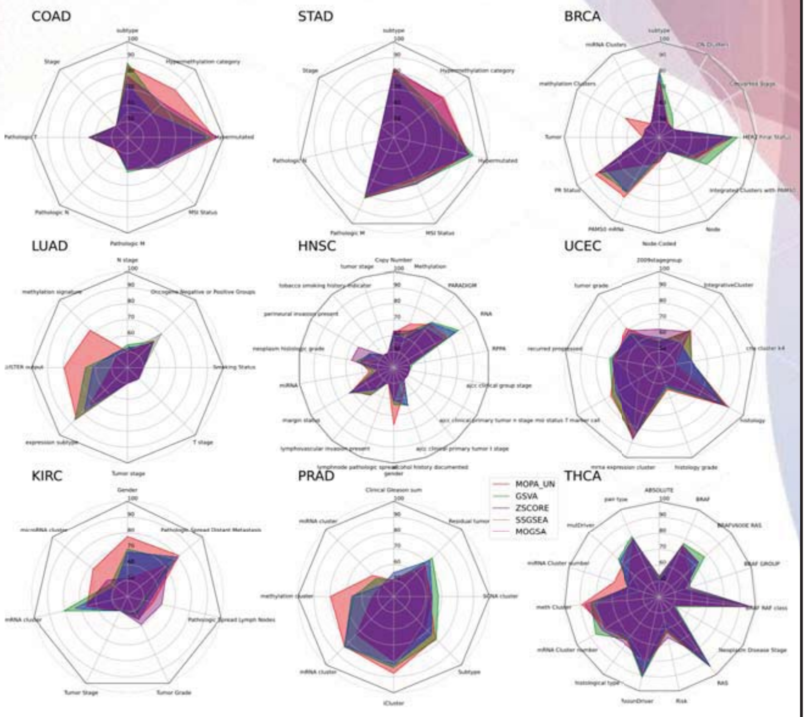


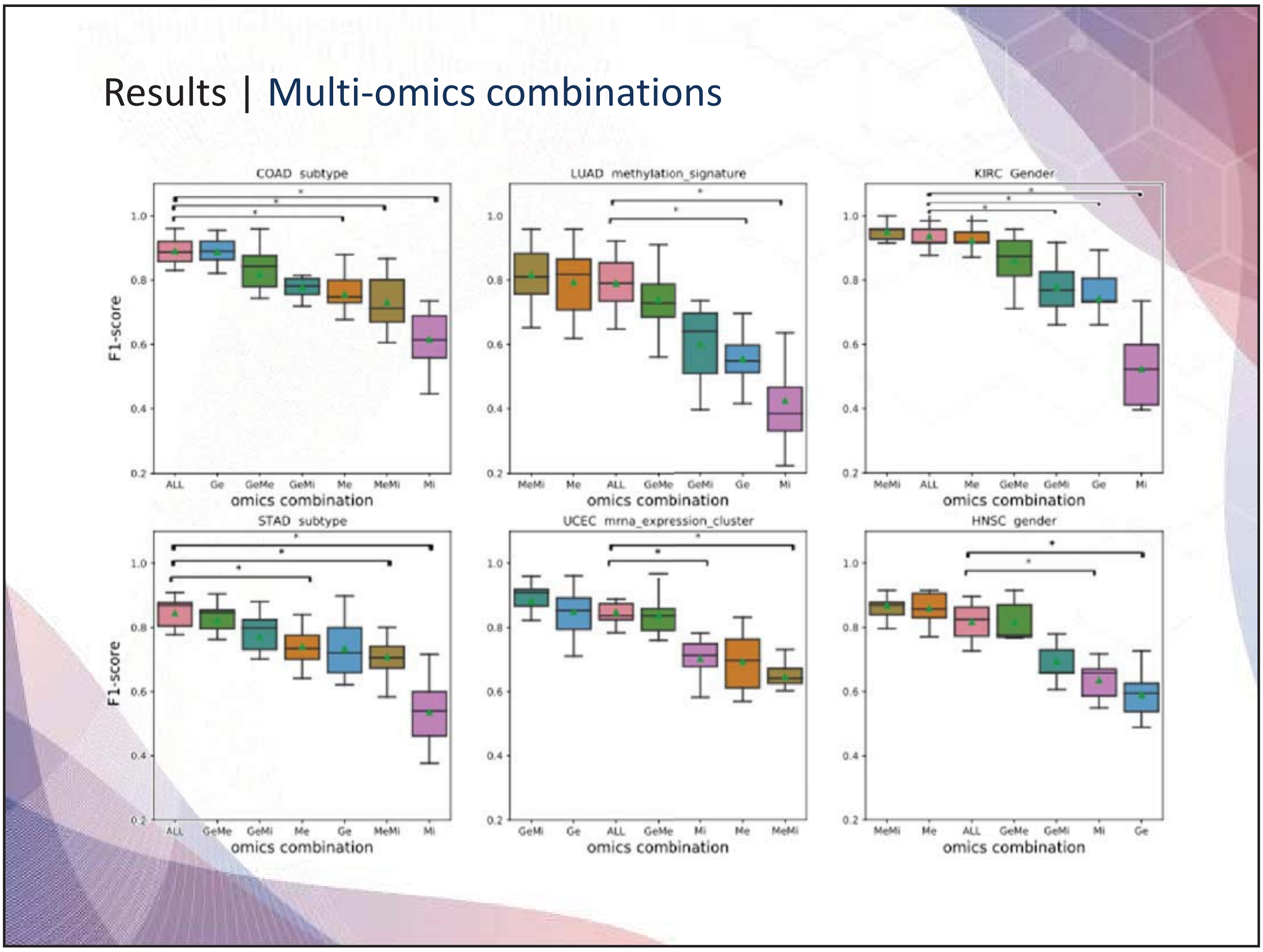
Methods| MOPA Workflow (Step 2~3)



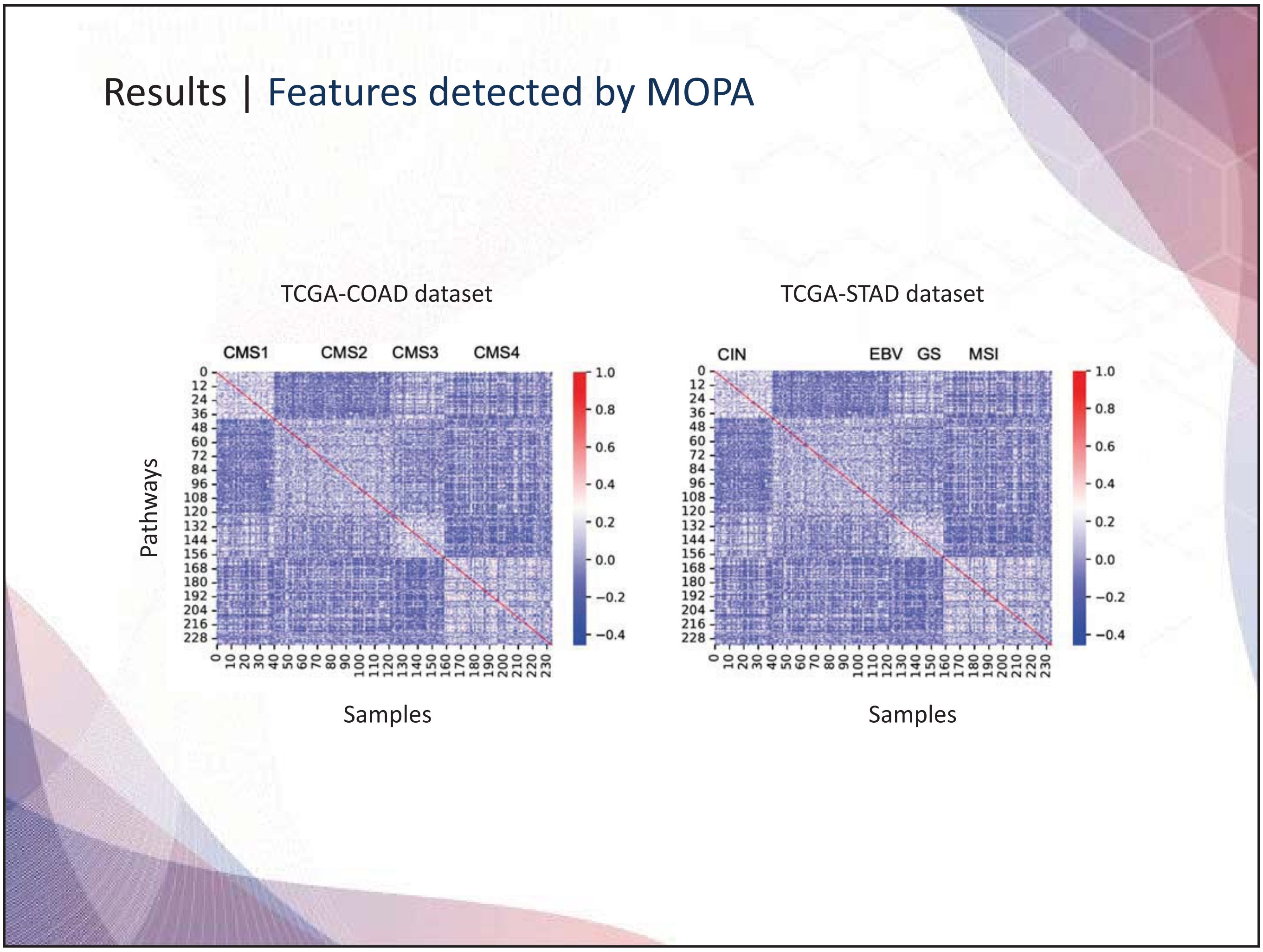
Results | MOPA on Pan-cancer data

- 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





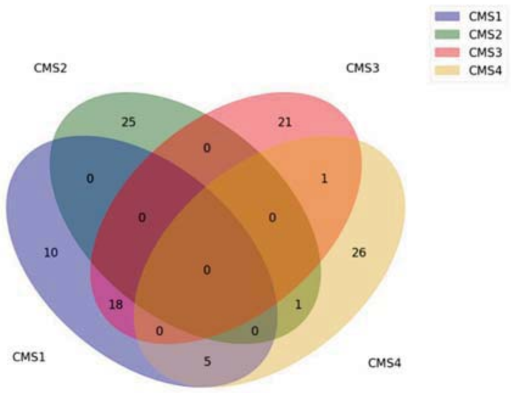
Results | Multi-omics combinations



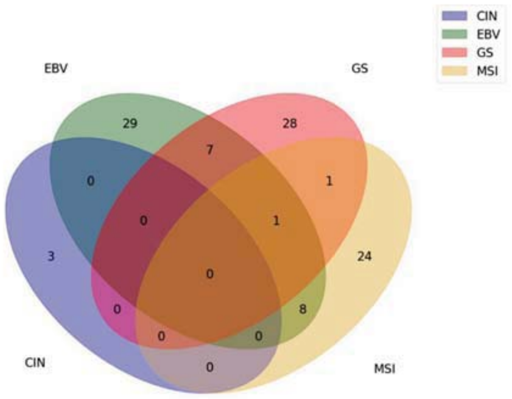
Results | Features detected by MOPA

Results | Features detected by MOPA

TCGA-COAD

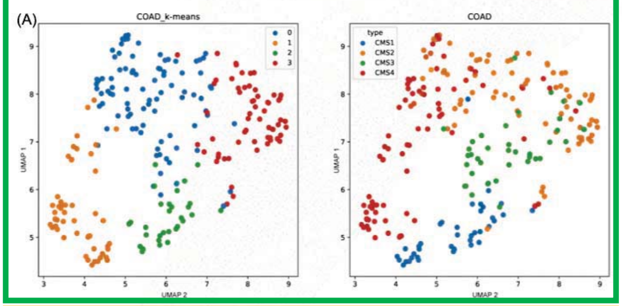


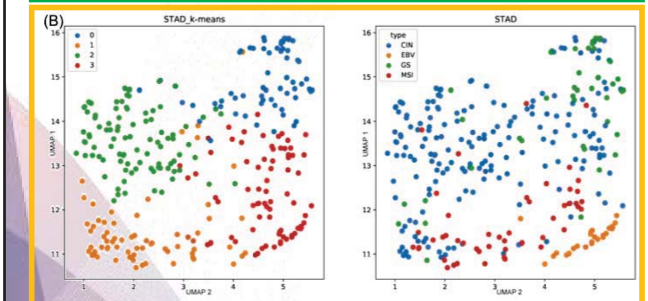
TCGA-STAD



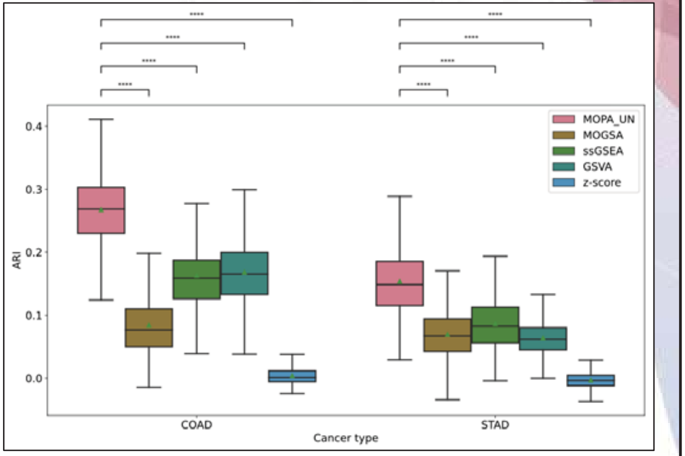
Results | Subtype clustering using *mES*

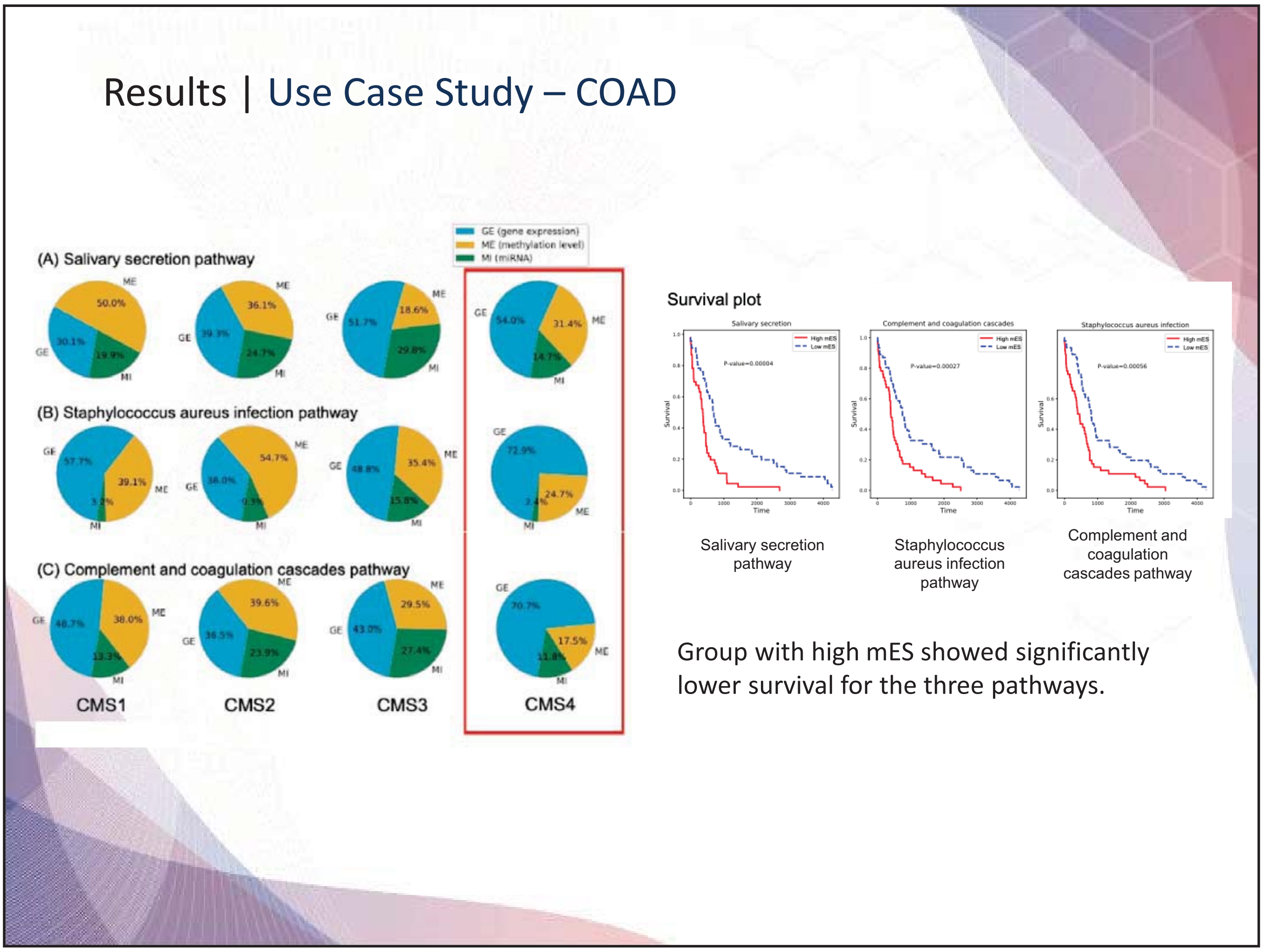
COAD





STAD





Results | Use Case Study – COAD

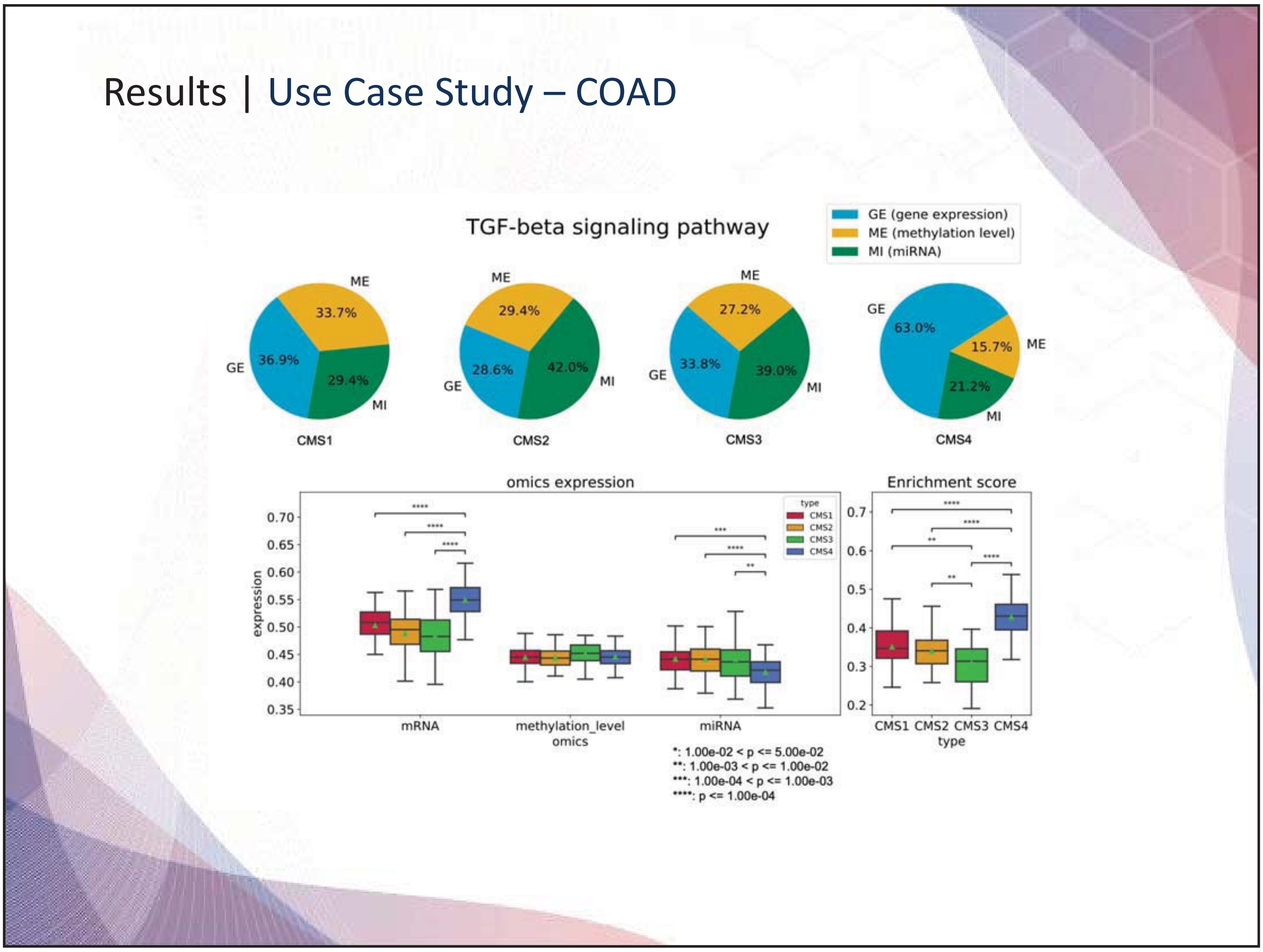
Survival plot

Salivary secretion pathway

Staphylococcus aureus infection pathway

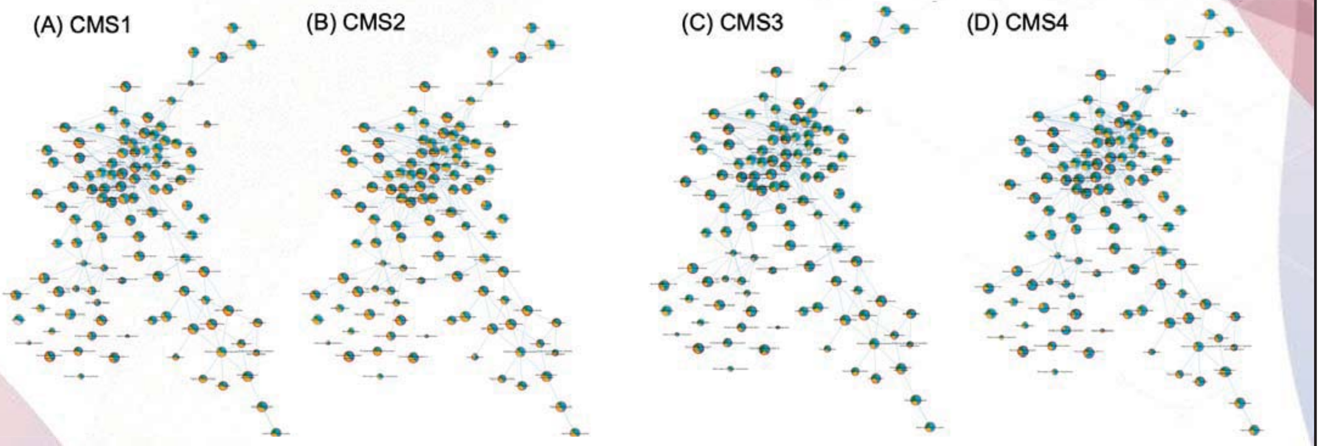
Complement and coagulation cascades pathway

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

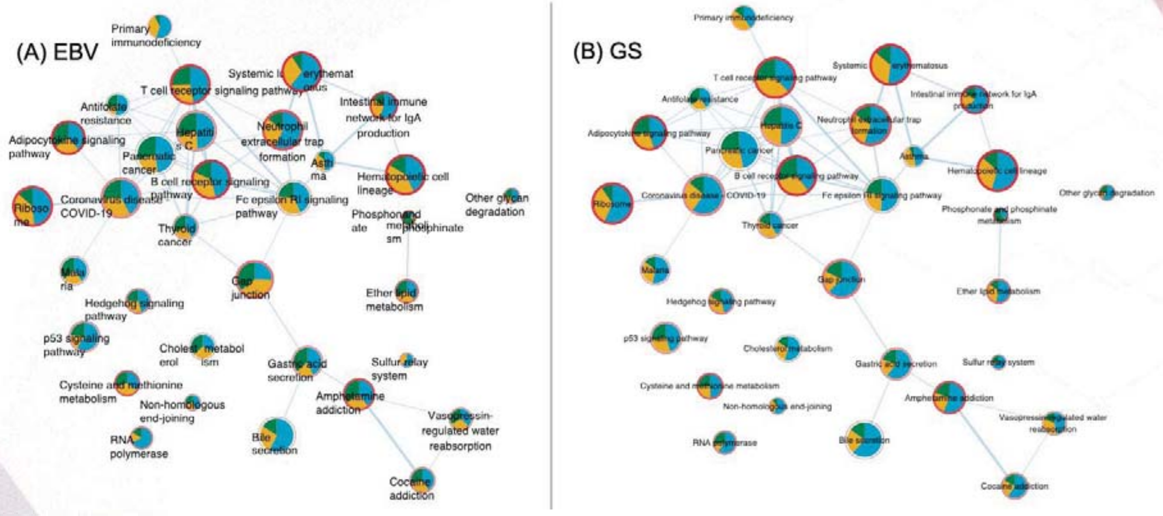


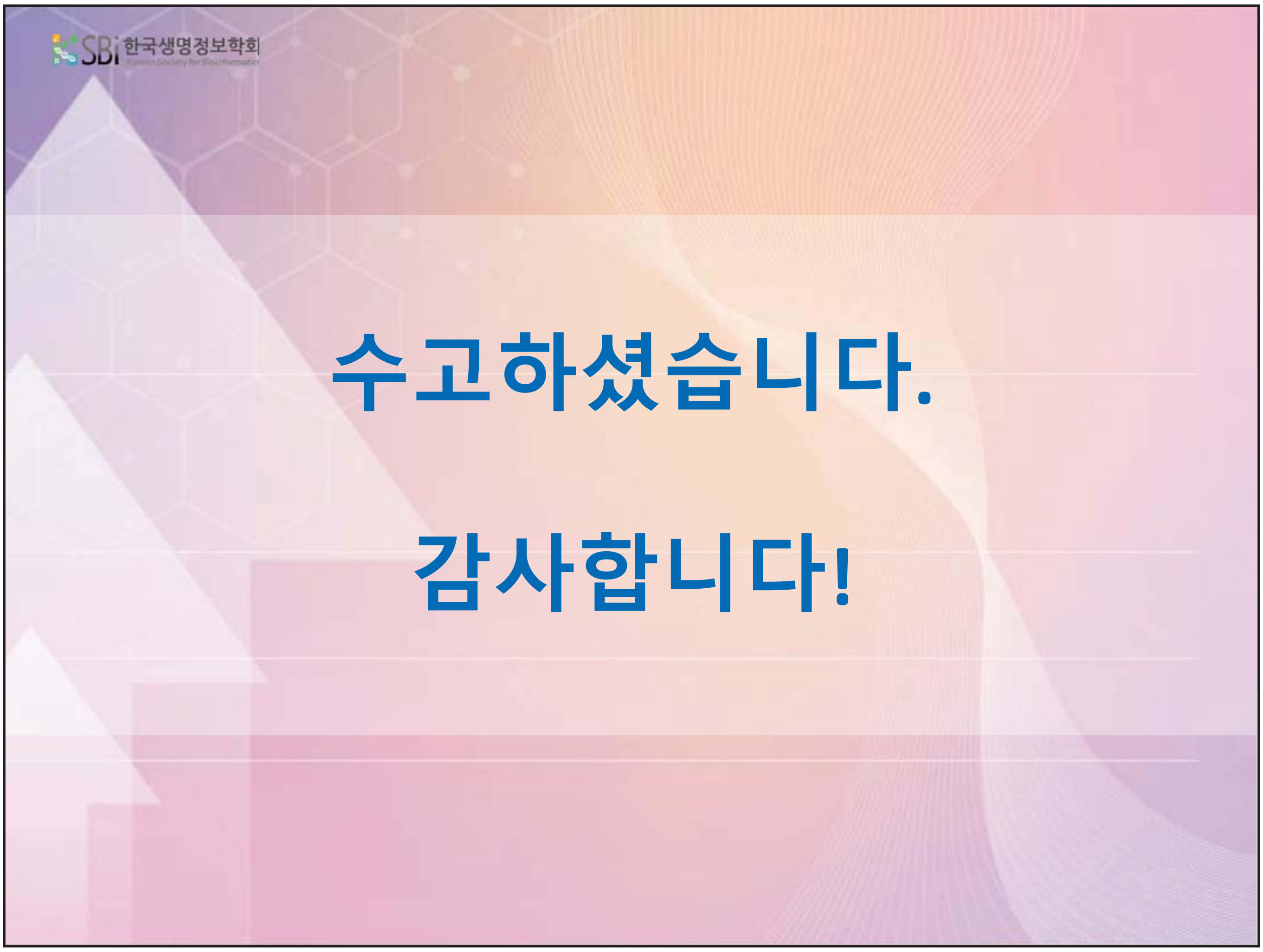
Results | Use Case Study – COAD

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



Results | Use Case Study – STAD


(A) EBV
Primary immunodeficiency
Systemic lupus erythematosus
T cell receptor signaling pathway
Antifolate resistance
Intestinal immune network for IgA production
Adipocytokine signaling pathway
Hepatitis C
Neutrophil extracellular trap formation
Pancreatic cancer
Asthma
B cell receptor signaling pathway
Hematopoietic cell lineage
Ribosome
Coronavirus disease COVID-19
Fc epsilon RI signaling pathway
Other glycan degradation
Phosphonate and phosphinate metabolism
Thyroid cancer
Malaria
Gap junction
Ether lipid metabolism
Hedgehog signaling pathway
p53 signaling pathway
Cholesterol metabolism
Gastric acid secretion
Sulfur relay system
Cysteine and methionine metabolism
Amphetamine addiction
Non-homologous end-joining
Vasopressin-regulated water reabsorption
RNA polymerase
Bile secretion
Cocaine addiction
(B) GS
Primary immunodeficiency
Systemic lupus erythematosus
T cell receptor signaling pathway
Antifolate resistance
Intestinal immune network for IgA production
Adipocytokine signaling pathway
Hepatitis C
Neutrophil extracellular trap formation
Pancreatic cancer
Asthma
Hematopoietic cell lineage
B cell receptor signaling pathway
Ribosome
Coronavirus disease - COVID-19
Fc epsilon RI signaling pathway
Other glycan degradation
Phosphonate and phosphinate metabolism
Thyroid cancer
Malaria
Gap junction
Ether lipid metabolism
Hedgehog signaling pathway
p53 signaling pathway
Cholesterol metabolism
Gastric acid secretion
Sulfur relay system
Cysteine and methionine metabolism
Amphetamine addiction
Non-homologous end-joining
Vasopressin-regulated water reabsorption
RNA polymerase
Bile secretion
Cocaine addiction



KSBi 한국생명정보학회 Korean Society for Bioinformatics

수고하셨습니다.

감사합니다!