19 October 2023

MOPA: An integrative multi-omics pathway analysis method for measuring omics activity

Taewan Goo

Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea

서울대학교 통계학과
생물정보통계연구실
BIBS

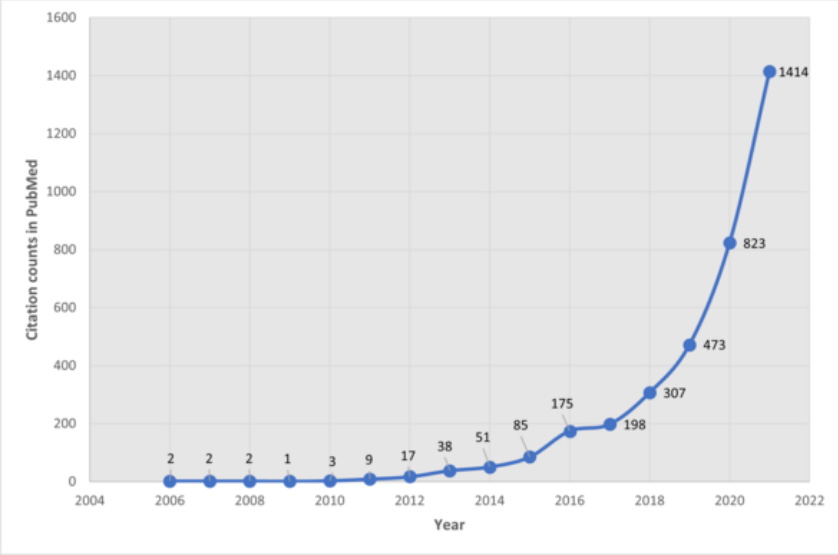
Contents

- Introduction
  - Background
  - Limitations and Challenges
  - MOPA (mES and OCR)
- Materials and methods
- Results
- Discussion

Introduction

Introduction to multi-omics

- The rise of multi-omics
  - Interest in combining different omics data types is growing, leading to a surge in multi-omics data.
  - Multi-omics is a method in biology that integrate data from various ‘omics’ layer like genomics, transcriptomics, and proteome.
  - By looking at many types of data together, scientists get a clearer picture of how cells work.
  - This approach helps in understanding diseases better, finding new markers for them, and seeking treatments.
  - It's very useful for complex diseases, where just looking at one type of data might not give the full story.

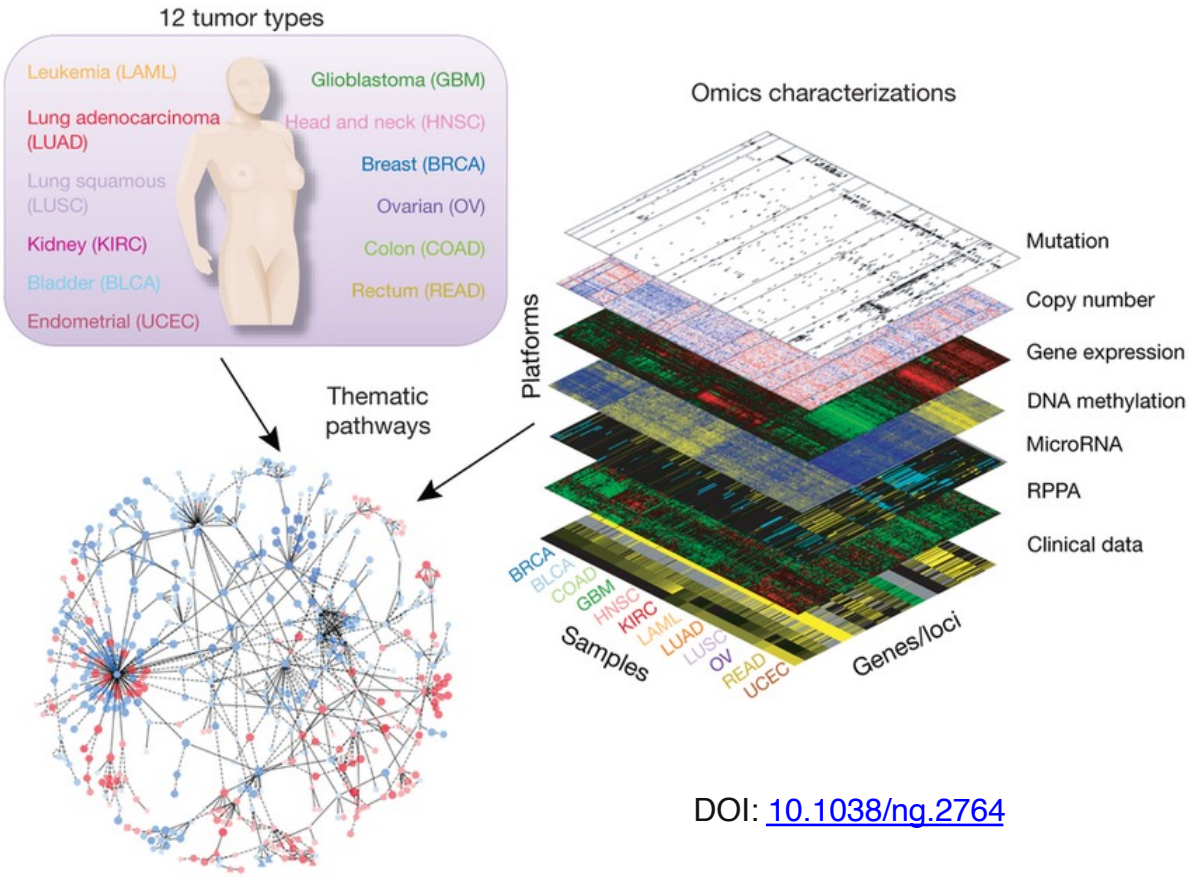


Source: https://en.wikipedia.org/wiki/Multiomics

Introduction

Introduction to multi-omics

- The rise of multi-omics
  - While data collection in multi-omics is advancing quickly, the development of analytical methods is lagging.
  - Interpreting data from multiple sources is complex and can pose challenges.
  - Notable multi-omics databases include TCGA, ENCODE, and GTEx, which provide rich sources of data.



DOI: 10.1038/ng.2764

Introduction

Existing Integrative Methods

- Current tools for multi-omics analysis
  - Recently, a number of methods has been proposed performing multi-omics data analysis:
    - **Similarity Network Fusion (SNF)**
    - **MOFA**
    - **iCluster+**
  - These methods, while helpful, have their limitations: they often tend to produce outputs focused on genes or specific omics features.
  - For deeper biological insights, researcher can further employ enrichment tests such as **GSVA** and **GSEA**

Introduction

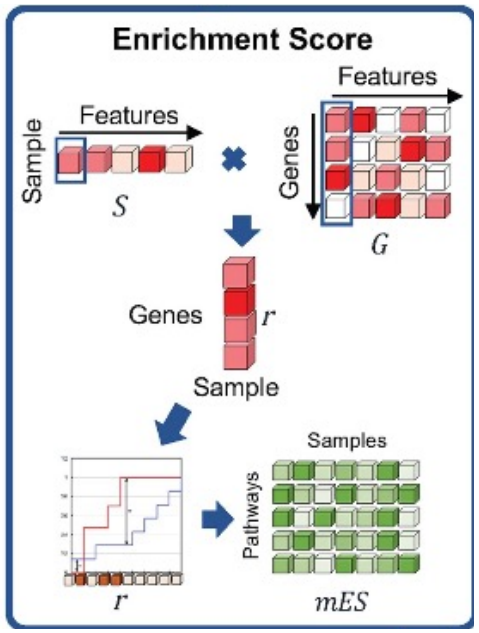
Pathway-based analysis

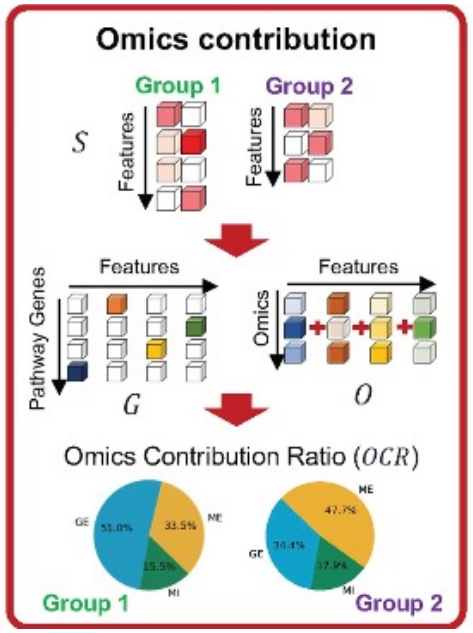
- Pathway as **self-explanatory biological mechanisms**:
  - Pathways provide a comprehensive view of how different genes and proteins interact in a coordinated manner.
  - By looking at pathways, researchers can quickly grasp the **broader biological context**, rather than getting lost in individual genes or proteins.
- Current methods providing pathway outputs:
  - Tools like ActivePathway, multiGSEA, and MOGSA are already paving the way in generating pathway-centric outputs.
- Advantages of **Pathway Enrichment Scores**:
  - Pathway enrichment scores allow for a quantitative understanding of how significantly a certain pathway is affected or altered.
  - This not only aids in identifying crucial pathways but also gives a relative measure of its significance in the biological context.

Introduction

MOPA - The Next Step in Multi-omics Analysis

- Advantages
  - **Pathway Ranking**: MOPA prioritizes pathways based on their relevance and significance in the context of the multi-omics data and associated clinical features.
  - **mES & OCR Metrics:** These innovative metrics introduced by MOPA enable a deeper understanding of pathway involvement and provide a clearer picture of the biological processes at play.
- MOPA's edge over other tools
  - While other tools provide pieces of the multi-omics puzzle, MOPA stands out by **offering a more holistic view**, seamlessly integrating diverse data types and emphasizing pathways that are crucial in clinical contexts.

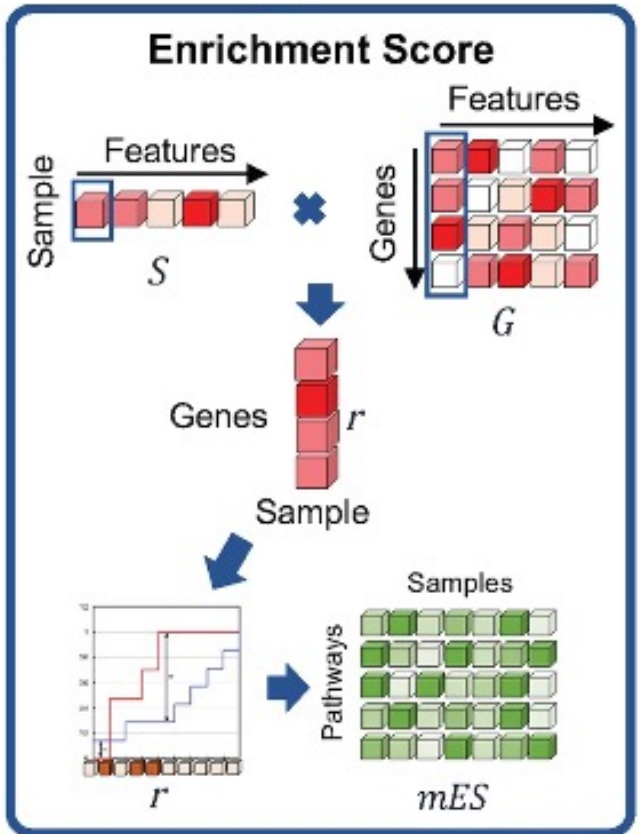




Introduction

Understanding mES

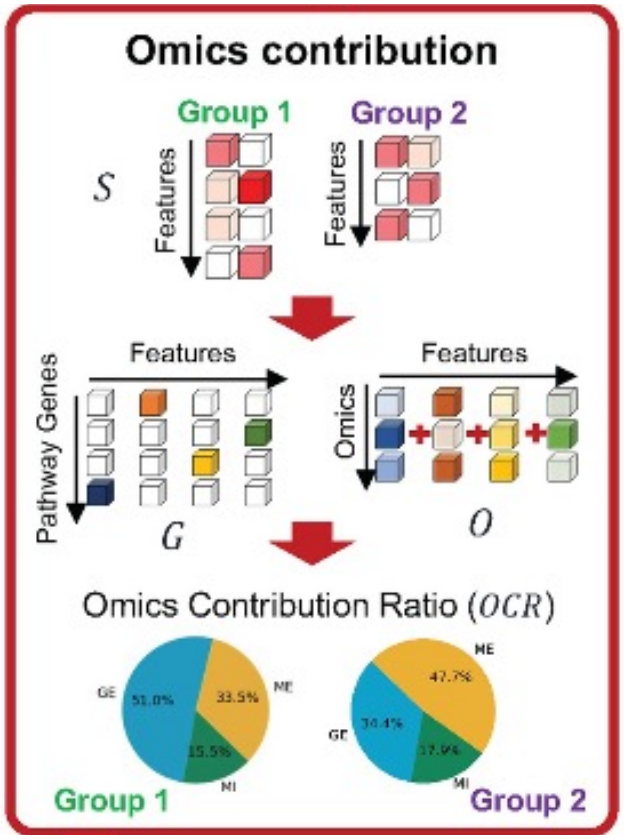
- Definition and Importance of mES
  - **mES** (Multi-Omics Enrichment Score) provides a single value that encapsulates the collective impact of all omics data on a particular pathway.
  - It offers a streamlined and simplified metric that combines the diverse omics layers into one coherent signal.
  - mES simplifies complex multi-omics data, enabling researchers to pinpoint crucial pathways without getting overwhelmed by the intricacies of each individual omics layer.



Introduction

Understanding OCR

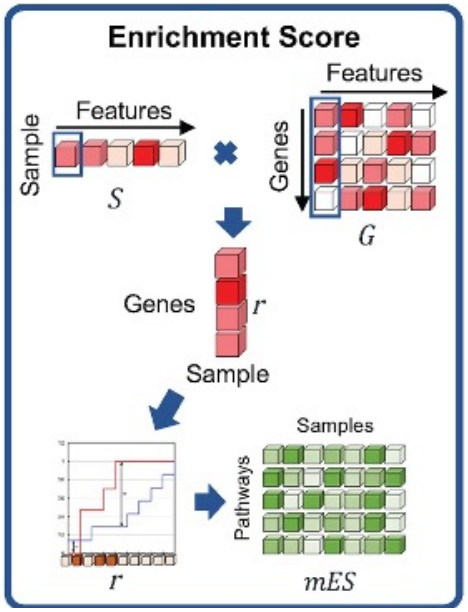
- Definition and Importance of OCR
  - OCR (Omics Contribution Ratio) dissects the mES score, **revealing the contribution of each specific omics layer to the overall score.**
  - It provides a breakdown of how much each omics type influences a pathway's activity.
  - While mES gives an overall view, OCR delves deeper, ensuring researchers understand the underlying dynamics of each omics layer, offering a clear picture of their interplay and respective impacts.

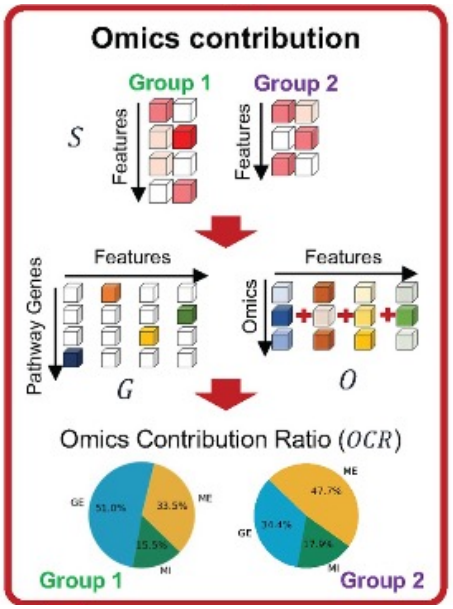


Introduction

MOPA - The Next Step in Multi-omics Analysis

- Definition and Importance of mES
  - **Pathway Ranking**: MOPA prioritizes pathways based on their relevance and significance in the context of the multi-omics data and associated clinical features.
  - **mES & OCR Metrics:** These innovative metrics introduced by MOPA enable a deeper understanding of pathway involvement and provide a clearer picture of the biological processes at play.
- MOPA's edge over other tools
  - While other tools provide pieces of the multi-omics puzzle, MOPA stands out by **offering a more holistic view**, seamlessly integrating diverse data types and emphasizing pathways that are crucial in clinical contexts.





Materials

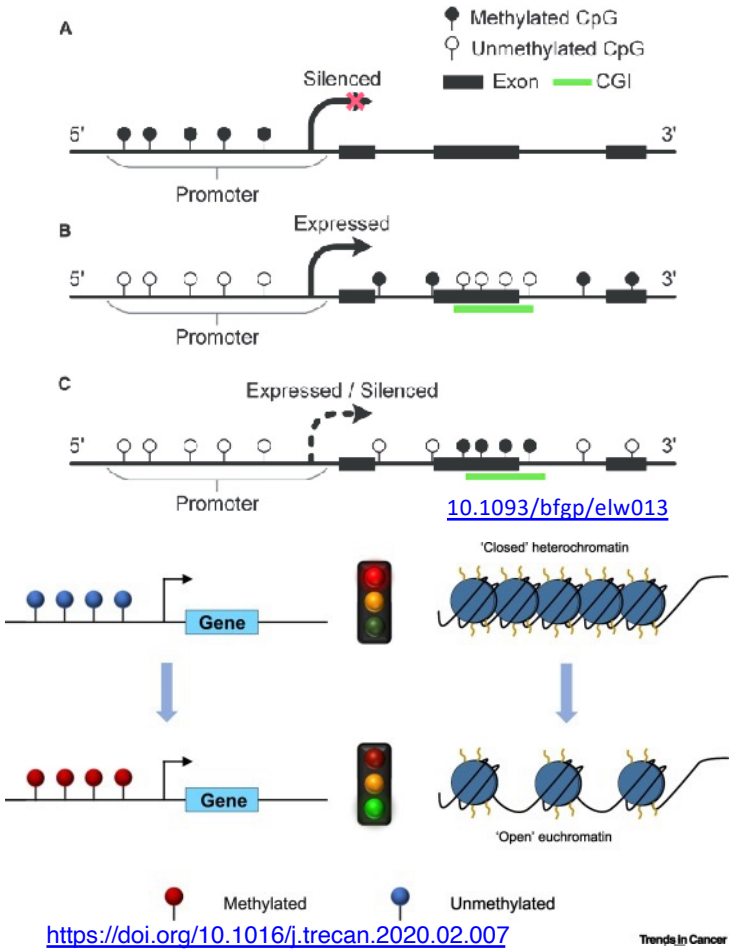
Multi-omics dataset

- **Gene expression (Transcriptomics)**
  - Process by which the information stored in genes is used **to produce functional products**, mainly proteins.
- **Methylation (Epigenomics)**
  - DNA methylation is a chemical modification in which a methyl group is added to the DNA molecule, typically at cytosine residues.
  - DNA methylation is one aspect of epigenomics, which studies **heritable changes in gene expression** without changes in the DNA sequence itself.
- **Micro RNAs (microRNomics)**
  - Micro RNAs(miRNAs) are small non-coding RNAs that play a crucial role in regulating gene expression post-transcriptionally.
  - They bind to the mRNA and either **inhibit its translation** or **lead to its degradation**.

Materials

Multi-omics relationship

- **Gene expression – Methylation – miRNA**
  - DNA methylation influences both transcription of gene and miRNA.
    - Methylation in promoter region of the mRNA and miRNA can lead to its reduced expression.
  - miRNA can regulate gene expression post-transcriptionally.
    - They bind to target mRNAs and either inhibit their translation or lead to their degradation.
  - **Paradoxical mechanism in Cancer**
    - Recent studies have observed that Hypermethylation sometimes correlates with gene activation. The phenomenon may introduce new gene regulation mechanisms, particularly in development, tumor formation, and metastasis[1].
    - Likewise, miRNA biogenesis is influenced by DNA methylation around its coding sequence. Removing DNA methylation from miRNA loci results in their downregulation[2].



10.1093/bfgp/elw013

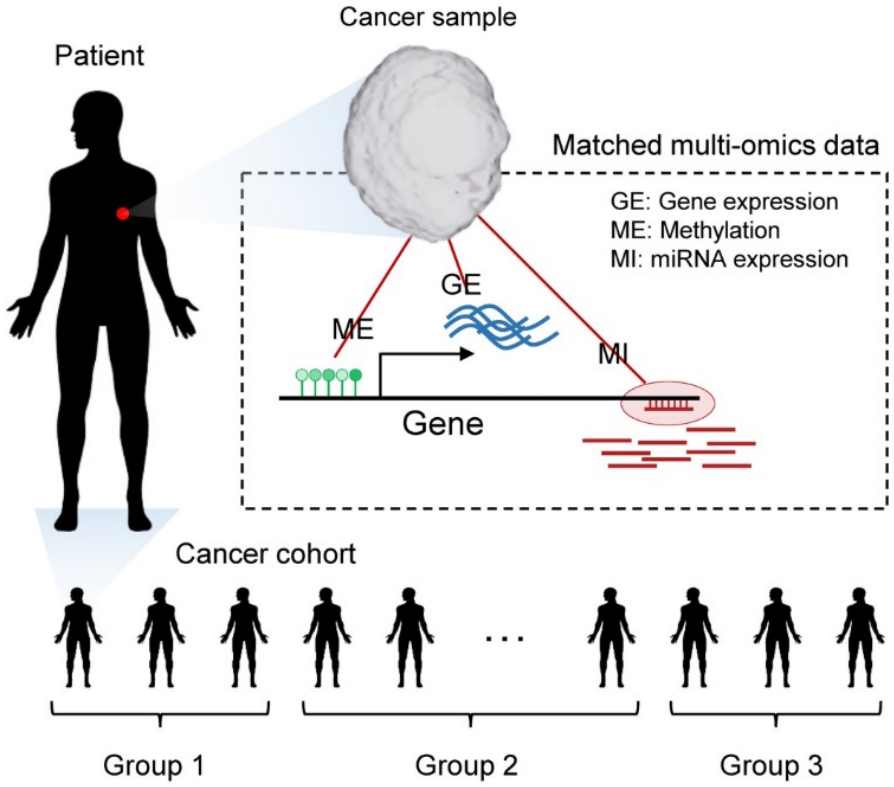
https://doi.org/10.1016/j.trecan.2020.02.007

1. Smith J, Sen S, Weeks RJ, Eccles MR, Chatterjee A. Promoter DNA Hypermethylation and Paradoxical Gene Activation. Trends in Cancer. 2020;6(5):392-406.
2. Yang X, Shao X, Gao L, Zhang S. Comparative DNA methylation analysis to decipher common and cell type-specific patterns among multiple cell types. Briefings in Functional Genomics. 2016;15:elw013.

Materials

Multi-omics dataset

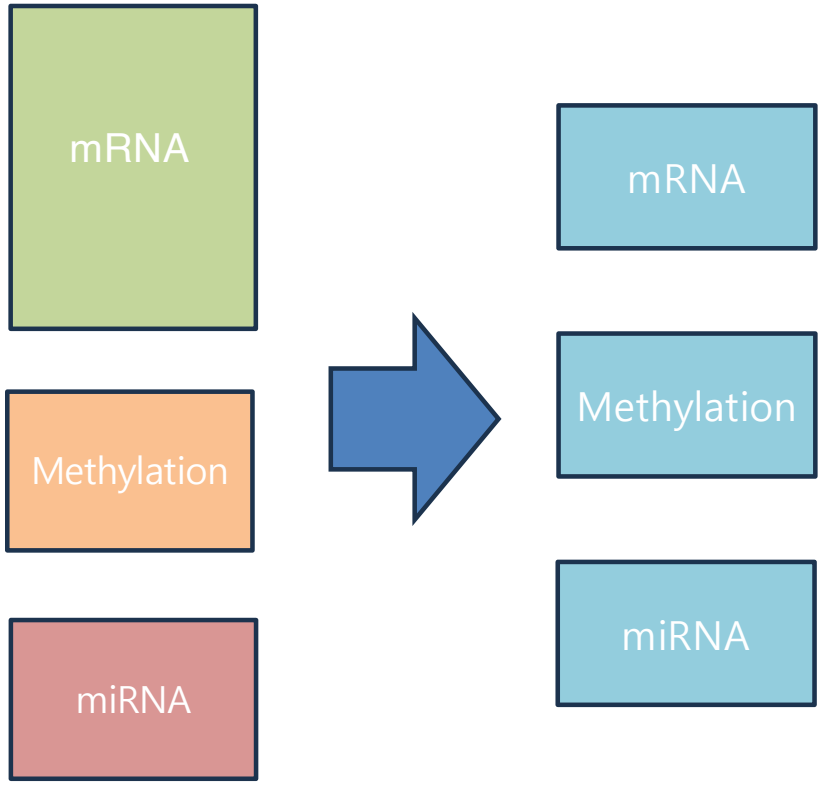
- **Data collection**
  - Multi-omics data was collected in patient-matched manner.
  - The multi-omics data of patients from various cohorts in TCGA data portal was used.
  - Multi-omics data captured gene-regulatory relations between different omics layers that significantly varied between clinical feature groups.



Materials

Multi-omics dataset

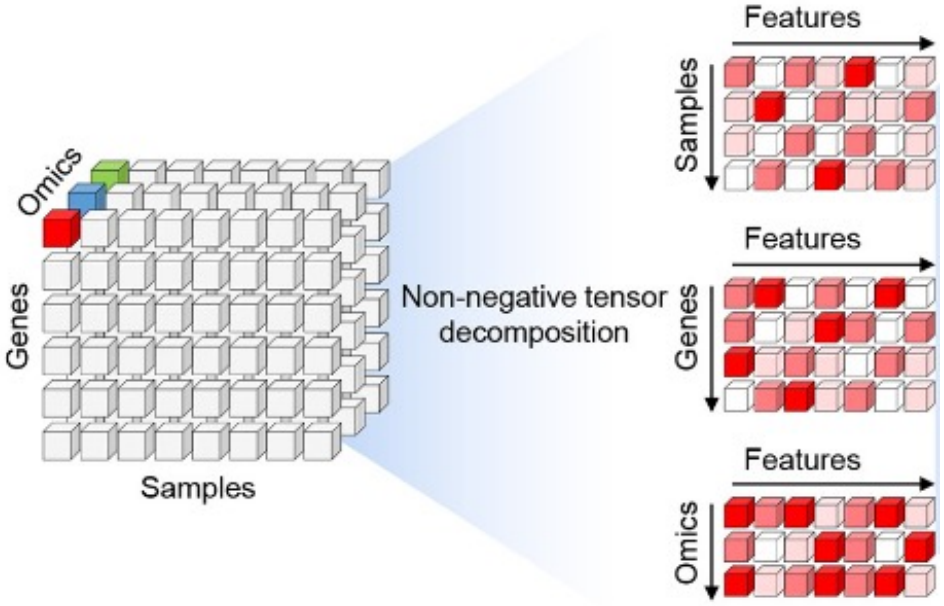
- **Feature scaling and conversion**
  - Each omics layer contains **varying features, scales, and data types.**
  - MOPA detects gene-regulatory **cis-relations** across multi-omics layers.
  - Omics data is transformed into **gene-level** data, streamlining pathway analysis and ensuring each omics layer shares consistent dimensions.



Materials

Multi-omics dataset

- **Tensor decomposition**
  - Combines omics slices into a tensor (or cube) for latent feature identification.
  - By using **MONTI** for **non-negative tensor decomposition,** which selects features related to a specific clinical feature[1].
  - Omics data is transformed into **gene-level** data, streamlining pathway analysis and ensuring each omics layer shares consistent dimensions.

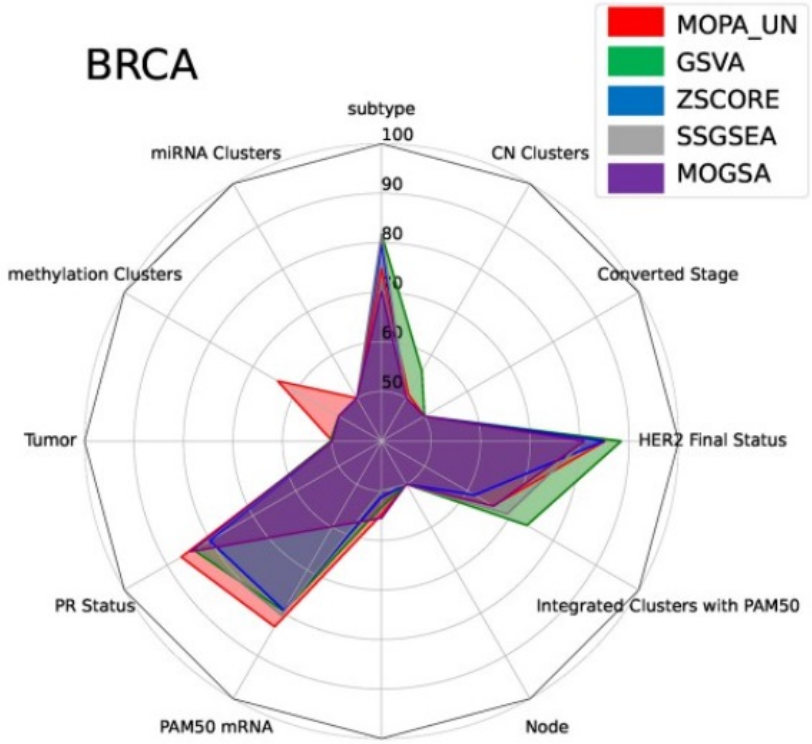


1. Jung I, Kim M, Rhee S, Lim S, Kim S. MONTI: A Multi-Omics Non-negative Tensor Decomposition Framework for Gene-Level Integrative Analysis. Front Genet. 2021;12:682841.

Materials

Multi-omics dataset

- **Application & Evaluation**
  - MOPA was used on nine cancer types, employing mRNA, miRNA and methylation data.
  - Aimed to demonstrate the superiority of MOPA's results using the OCR metric.
  - **Performance of MOPA** was gauged against **four other pathway scoring methods.** This comparison highlighted the efficacy of multi-omics over single-omics analysis.



Materials

Cancer dataset

- Nine cancer types were investigated.
- Samples were compiled based on available clinical feature labels.
- Clinical features for each cancer type and their values presented table in next table.

Cancer type	Clinical feature	No. of groups	Clinical feature groups	No. of samples
COAD	Molecular subtype	4	CMS1, CMS2, CMS3, CMS4	234
STAD	Molecular subtype	4	CIN, EBV, GS, MSI	305
BRCA	Subtype	4	LumA, LumB, HER2, Basal	595
HNSC	Gender	2	Female, male	298
PRAD	Methylation cluster	4	1, 2, 3, 4	328
KIRC	Gender	2	Female, male	252
LUAD	Methylation Signature	3	Low, intermediate, high	181
THCA	BRAF mutation group	2	0, 1	490
UCEC	mRNA expression cluster	3	1, 2, 3	221

Materials

Use Case Study Dataset

- **Objective:**
  - Validate and demonstrate the utility of MOPA
- **Studies performed on:**
  - Molecular subtypes in colon and stomach adenocarcinoma cohorts.
- **Findings:**
  - MOPA reproduced significant biological results specific to each cancer type and its clinical feature groups.
  - Clinical feature groups are attributes from medical records like cancer subtype, age, gender, and stage.

Materials

Pathway and Annotation data

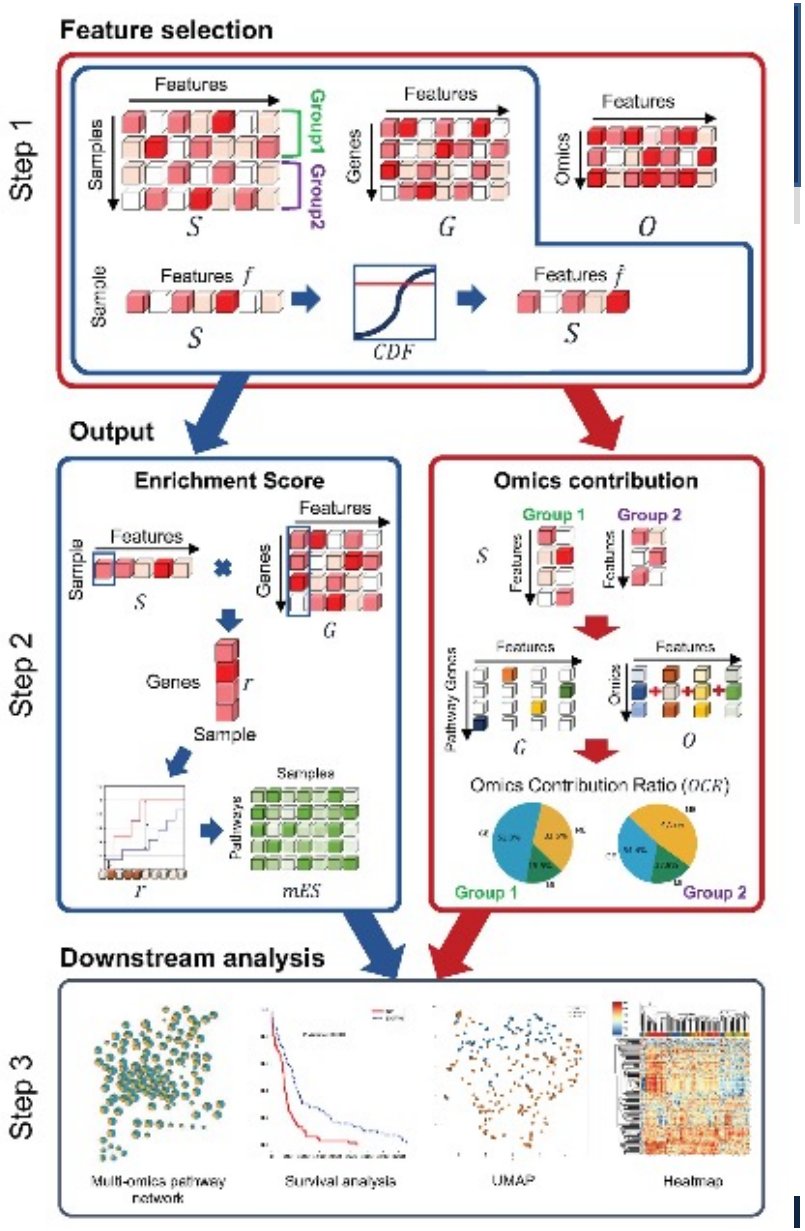
- **Pathway source:**
  - Human KEGG pathway database[1]
- **Analysis method:**
  - miRNA and methylation values were quantified per genes
- **Data processing for pathway analysis**
  - **Average expression of miRNAs** that target a specific gene was assigned to that gene.
  - **miRDB** was used for grouping miRNAs per target gene[2].
  - **Average beta value of probes** located **within 2Kbp upstream** of a gene's transcription start site was assigned.
  - Genes without **associated miRNAs or methylation probes were assigned a value of zero**.
  - **Non-coding genes** were excluded from pathways.

1. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research. 2017;45(D1):D353-D61.
2. Wang X. Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. Bioinformatics. 2016;32(9):1316-22.

Methods

Analysis Workflow Overview

- **Three main steps of MOPA workflow**
  - Preprocessing multi-omics data and detect latent gene-level features.
  - Compute pathway enrichment scores from selected features.
  - Conduct downstream analyses on pathway enrichment scores.



Methods

Step1. Multi-omics Feature Selection

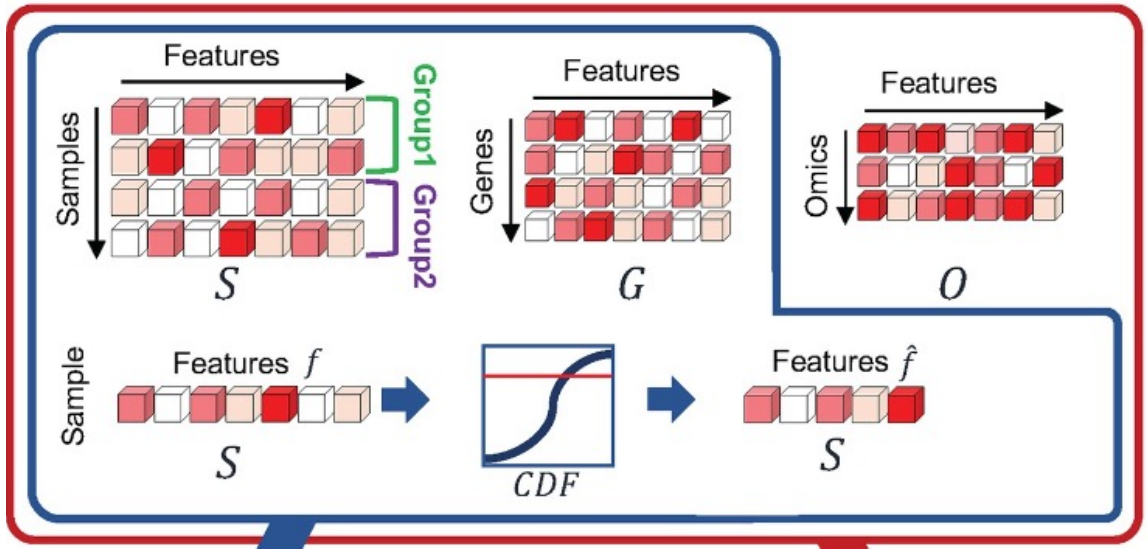
- **Objective:** Gene-level multi-omics feature selection
- **Method**: MONTI – Integrates multi-omics data and outputs latent gene features.
  - With sample labels: **MONTI (supervised)** selects features associated with them.
  - Without labels: MOPA proceeds in an **unsupervised manner**.
  - **Input**: Three dimensional Tensor (X_{ijk})
    - i: number of genes
    - j: number of samples
    - k: number of omics
- **Latent features**: Computed using the PARAFAC tensor factorization method[1].
  - **Tensor decomposition**: Result in three loading matrices: S, G, and O
  - Rank determine latent gene-level omics features; **the R is predetermined.**

1. Bro R. PARAFAC. Tutorial and applications. Chemometrics and Intelligent Laboratory Systems. 1997;38(2):149-71.

Methods

Step1. Multi-omics Feature Selection

- **Tensor decomposition:** method used to break down a tensor into its **constituent parts**, allowing us to represent complex multi-dimensional data in a more **interpretable**
- **S (Sample component)**
  - This matrix represent how each sample associates with the latent features derived from the tensor decomposition.
  - Each row pertains to a sample, and each column pertains to a feature.
  - A high value in a specific cell indicates a strong association of that sample with corresponding feature



Methods

Step 1. Multi-omics Feature Selection

- **Min-max scaling(?) of sample feature association**

$$s_i' = \frac{s_i - min(s_i)}{max(s_i) - min(s_i)}, \forall i \in Samples$$

- **In the code, they used Quantile normalization not min-max scaling**

```
### normalization sample matrix
sample_tensor_selec=qnorm.quantile_normalize(sample_tensor_selec, axis=1, ncpus=8)
sample_tensor_norm=pd.DataFrame(data=sample_tensor_selec)
```

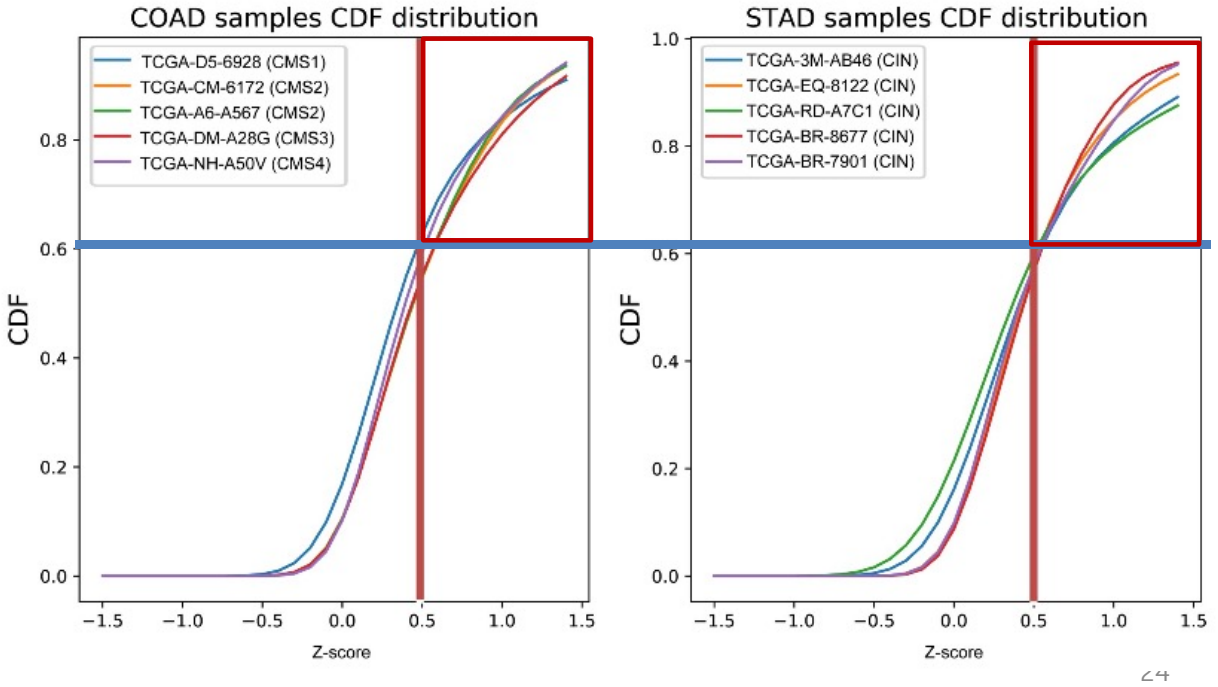
Methods

Step 1. Multi-omics Feature Selection

- **After then, they performed kernel density estimation with S_i' of each sample**
  - Kernel type: Gaussian Kernel

$$CDF_i = \int_{-\infty}^{s_i'} P_{kde}(s_i')dx$$

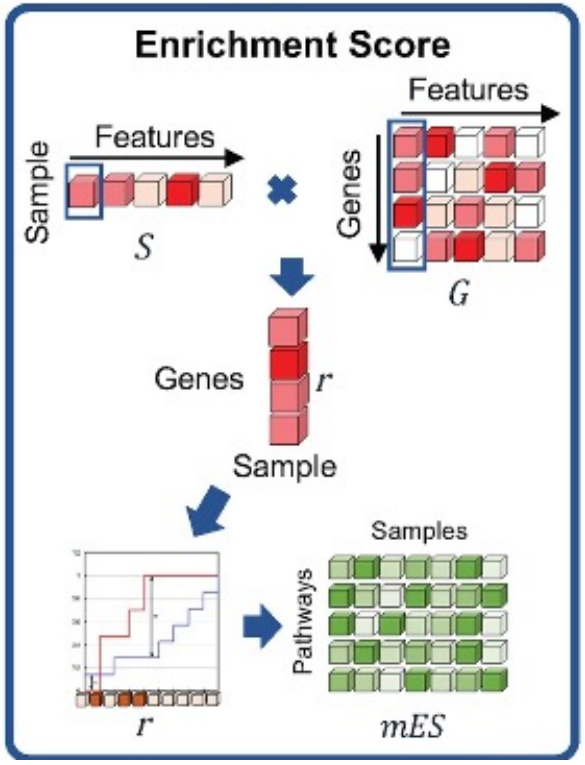
  - By testing a range of CDF thresholds for selecting informative features, 0.6 showed robust results across several different datasets as shown in next Figure.
  - They used 0.6 as threshold for feature selection.



Methods

Step 2. Compute pathway enrichment scores from selected features

- **Calculating mES**
  - The multi-omics Enrichment Score (mES) measures the multi-omics signal strength of a pathway in each sample. A high mES score indicates that a significantly large portion of genes belonging to a specific pathway are highly activated in terms of multi-omics, as compared to those not part of the pathway.
- **Three types of input**
  - Decomposed sample (S) and gene (G) matrix.
  - Gene Matrix Transposed file: indicating the gene memberships to pathways
  - The sample assigned features $\hat{f}$



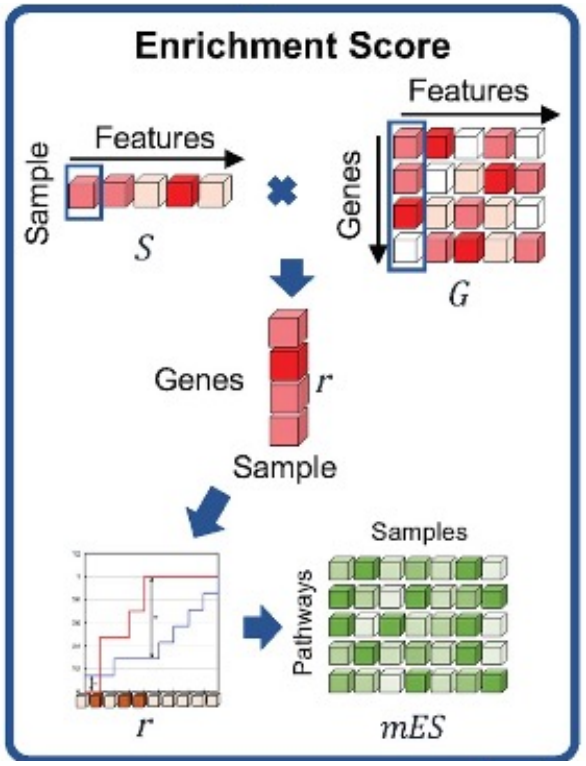
Methods

Step 2. Compute pathway enrichment scores from selected features

- **Methods**
  - For each gene j and sample i, we need to calculate r_{ij}

$$r_{ij} = \sum_{f=1}^{\hat{f}} g'_{jf} \times s'_{if} \qquad \vec{r}_i = (g'_{\hat{f}})^T \cdot s'_{i\hat{f}}$$

    - $s'_{i\hat{f}}$: sample feature values calculated in step 1.
    - $g'_{\hat{f}}$: standardized and positive scaled of $g_{\hat{f}}$.
  - The vector r_i is sorted to order genes.
  - **mES** is computed using the Kolmogorov–Smirnov (KS) random walk statistic, measuring similarity between two distributions: genes with and without membership to a pathway.



Methods

Step 2. Compute pathway enrichment scores from selected features

- **Methods**
  - The result will be the cumulative difference, d_{ijt}, up to the j-th ordered gene between the two distributions of pathway t in sample I as shown in below:

$$d_{ijt} = \frac{\sum_{l=1}^{j} r_{il} I(G_{(l)} \in p_{(t)})}{\sum_{l=1}^{p} r_{il} I(G_{(l)} \in p_{(t)})} - \frac{\sum_{l=1}^{j} I(G_{(l)} \notin p_{(t)})}{q - |p_{(t)}|},$$

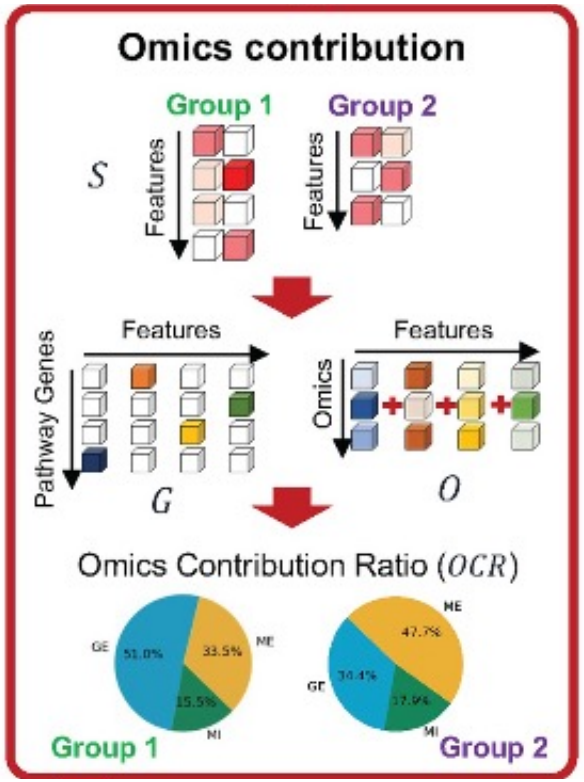
    - $p_{(t)}$: the set of genes in pathway t
    - $I(G_I \in p_{(t)})$: the indicator function that outputs 1 if the l-th gene is a member of pathway t and 0 otherwise.
    - q refers to the number of genes in the dataset.
    - d_{ijt} is computed for each sample, gene, and pathway.
  - The gene with a high r_{il} value starts calculation and d_{ijt} value shows the difference between genes belong to the pathway and genes not.

$$mES_{it} = \max_{j=1,\ldots,n}(0, d_{ijt}) - \left|\min_{j=1,\ldots,n}(0, d_{ijt})\right|$$

Methods

Step 2. Compute pathway enrichment scores from selected features

- **Calculating OCR**
  - The Omics Contribution Rate (OCR) shows the extent to which each type of omics (e.g., genomics, proteomics, etc.) contributes to the mES. It aims to interpret how much a pathway's activity is influenced by each type of omics data.
- **Three types of input**
  - Decomposed sample (S), gene (G) and omics (O) matrix.
  - Gene Matrix Transposed file: indicating the gene memberships to pathways.
  - The sample assigned features $\hat{f}$

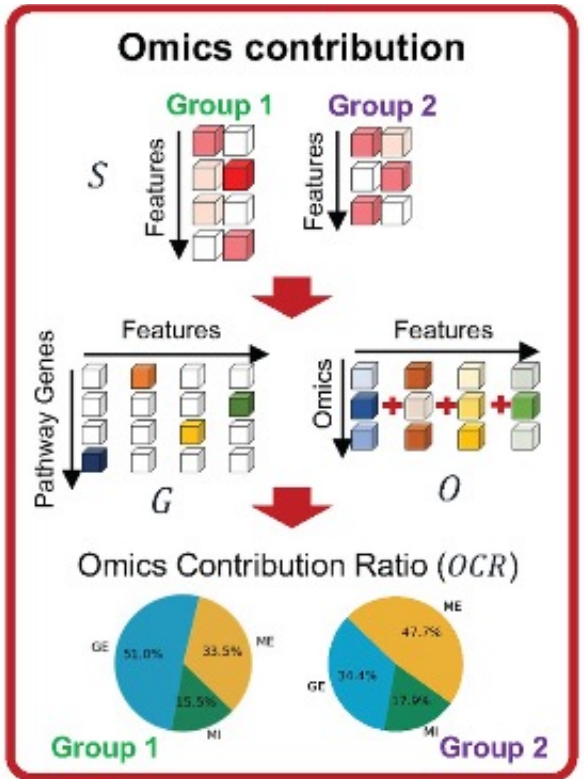


Methods

Step 2. Compute pathway enrichment scores from selected features

- **Methods**
  - Features **commonly assigned to samples within a clinical feature group** are collected.
    - Features shared by 50% of the samples in a group are gathered from S'.
  - The **strongest associated feature of each gene** is selected from g'_j for every gene in $p(t)$
    - The omics profiles of features are then summed to compute below:

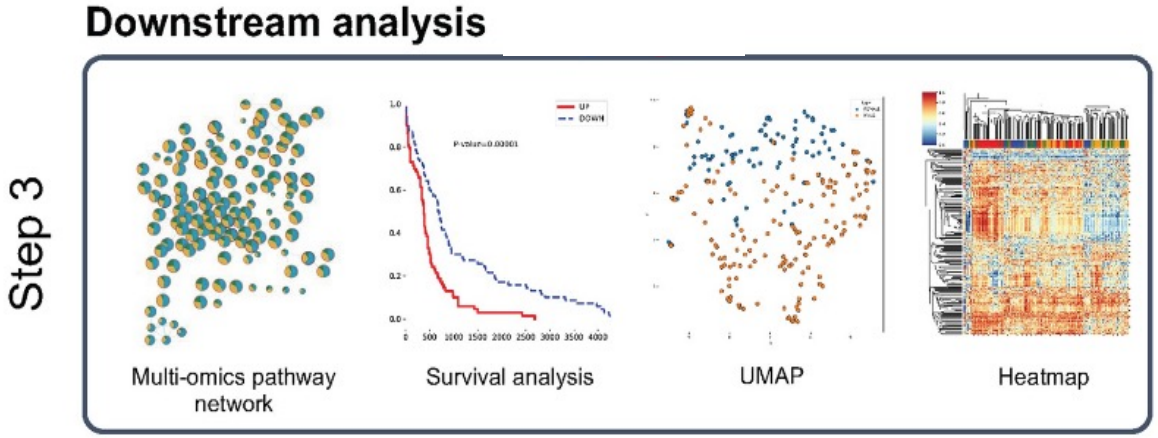
$$\vec{C}_{mt} = \sum_{j=1}^{p_{(t)}} O'_{argmax(g_{\hat{jf}\,m})} \qquad OCR_{kmt} = C_{kmt} / \sum_{k=1}^{L} C_{kmt}$$



Methods

Step 3. Downstream analysis using mES and OCR

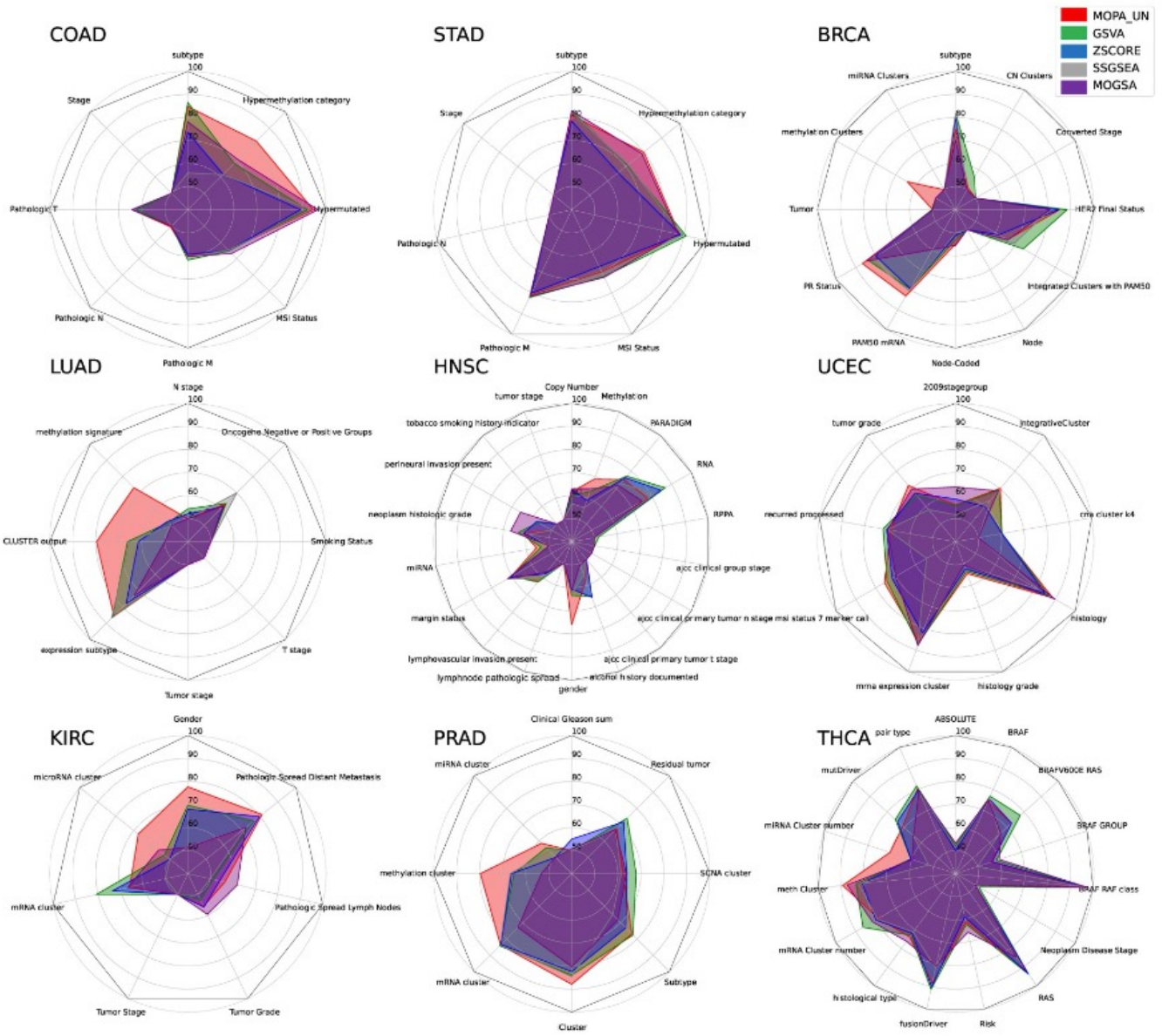
- **Types of downstream analysis**
  - Survival analysis with mES
  - Pathway network visualization (Cytoscape)
  - Multi-omics characteristic visualization (UMAP)
  - Association test among clinical feature group with r_j



Results

The classification performance comparison

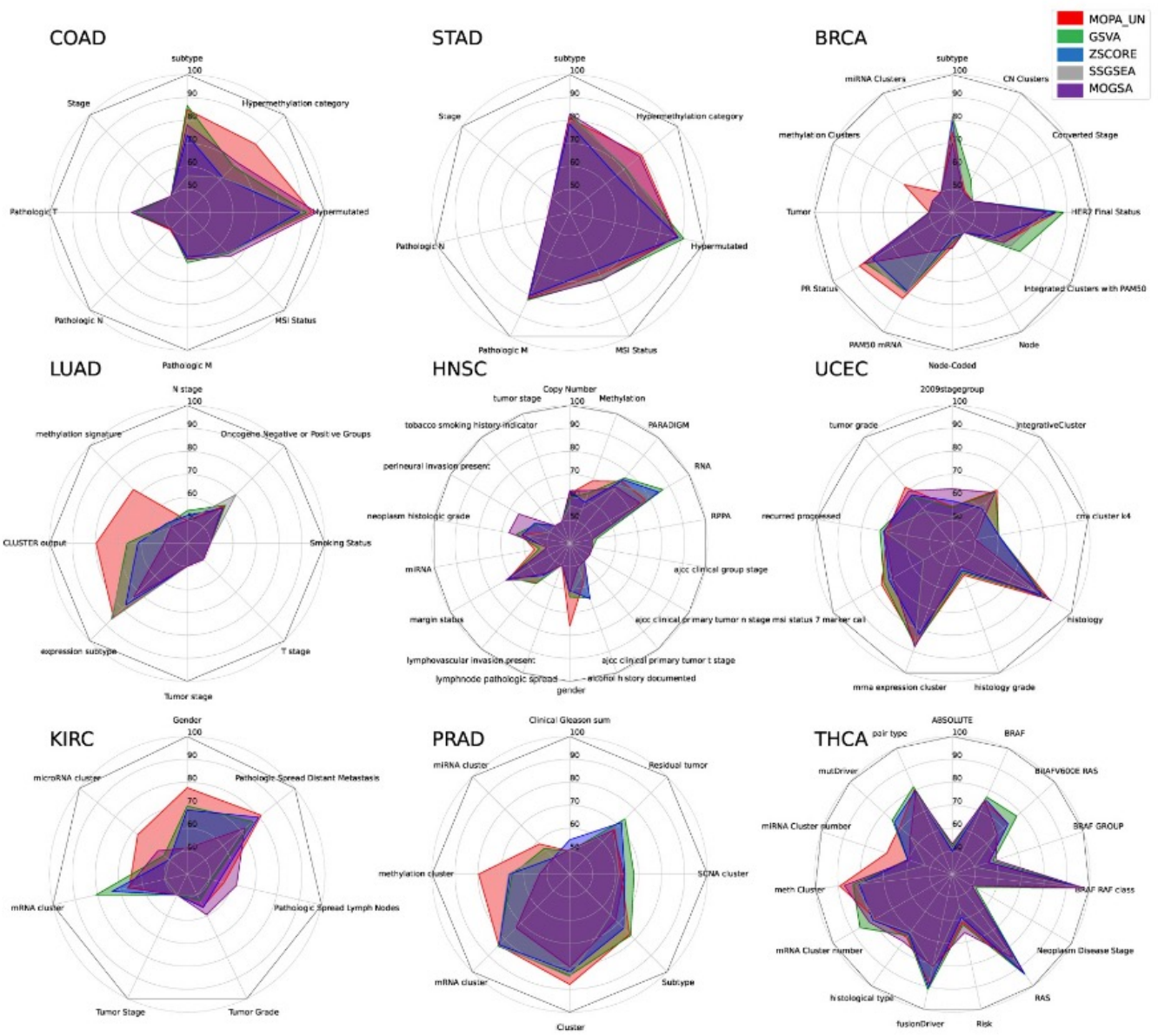
- The results were compared to other pathway enrichment tools
- By training a Multi-Layer Perceptron model on the acquired biological features for predicting target labels, **the 10-cross validated average f1 score of sample classification** results was measured.
- Among the clinical features subtype, mutation clusters showed high F1 scores in all the tools.



Results

The classification performance comparison

- MOPA showed the highest F1 scores in COAD, STAD subtypes, mutations, and hypermethylation clusters.
- For performance evaluation they used four different classification methods as shown next slide.
  - Random Forest
  - Support Vector Machine
  - K-Nearest Neighbor
  - Multi-layer Perceptron



Results

The classification performance comparison

- Overall, MOPA_UN classifiers achieved higher F1-scores in all the tasks.
- The results show that utilizing multi-omics data is advantageous over single-omics data.

Table 3. The F1-score for classifying molecular subtypes in COAD and STAD were measured using four different classification methods.

Cancer	Group	Method	RF	SVM	KNN	MLP
COAD	Subtype	MOPA_UN	0.822	**0.87**	**0.837**	**0.865**
		GSVA	0.75	0.854	0.71	0.86
		ZSCORE	0.663	0.766	0.804	0.712
		ssGSEA	**0.853**	0.584	0.746	0.560
		MOGSA	0.732	0.487	0.654	0.789
STAD	Subtype	MOPA_UN	**0.818**	**0.856**	**0.818**	0.814
		GSVA	0.71	0.805	0.659	**0.83**
		ZSCORE	0.673	0.766	0.65	0.791
		ssGSEA	0.78	0.584	0.776	0.83
		MOGSA	0.793	0.466	0.727	0.826
COAD	Hypermethylation cluster	MOPA_UN	**0.723**	**0.808**	**0.72**	**0.814**
		GSVA	0.581	0.618	0.53	0.688
		ZSCORE	0.513	0.659	0.401	0.614
		ssGSEA	0.673	0.555	0.641	0.48
		MOGSA	0.671	0.273	0.494	0.709
STAD	Hypermethylation cluster	MOPA_UN	0.767	**0.82**	**0.816**	**0.808**
		GSVA	0.56	0.68	0.615	0.693
		ZSCORE	0.57	0.66	0.531	0.684
		ssGSEA	0.612	0.543	0.657	0.708
		MOGSA	**0.783**	0.446	0.722	0.778

Results

Comparison of significant pathways with SOTA

- ActivePathways: A state-of-the-art method that calculates a pathway's p-value for each omics individually. Outputs a list of significant pathways.
- ActivePathways identified 7 significant pathways.
  - 6 out of 7 pathways were significant in both MOPA and ActivePathways.
  - Least Agreement: "Lysosome" pathway
- While MOPA's performance isn't superior, it matches or surpasses compared methods.
- MOPA offers richer context interpretation for multi-omics data.

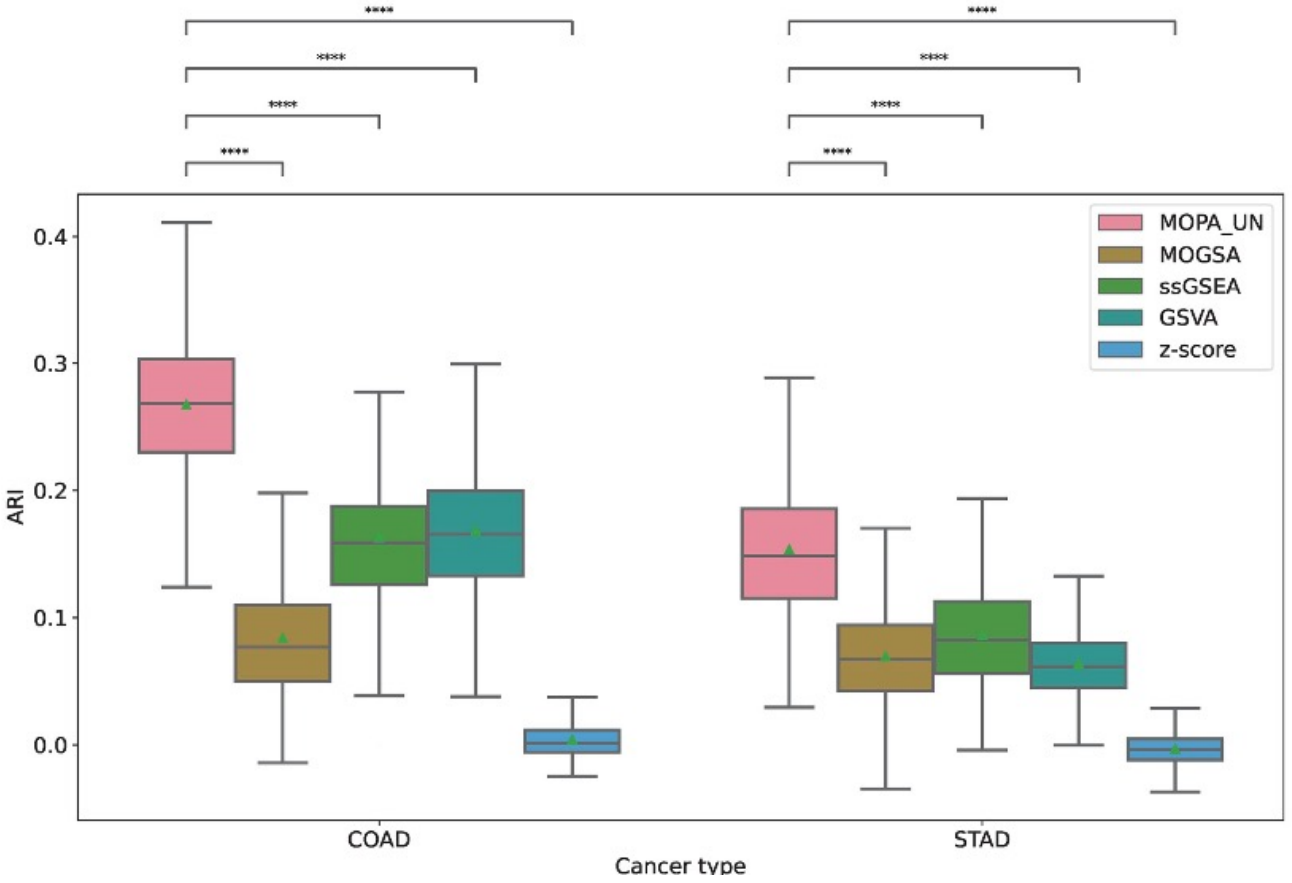
Table 4. Comparison btween ActivePathways and MOPA.

Pathway term	ActivePathways adj. p-value	ActivePathways Supported omics	MOPA adj. p-value
Focal adhesion	3.335E-4	ALL	1.055E-27
ECM-receptor interaction	5.856E-4	Gene	5.341E-25
Axon guidance	8.269E-3	Gene	1.002E-24
Protein digestion and absorption	1.979E-2	Gene	1.522E-30
AGE-RAGE signaling pathway in diabetic complications	2.540E-2	Gene	4.334E-17
Osteoclast differentiation	3.981E-2	Gene	1.716E-38
Lysosome	4.944E-2	Methylation, miRNA	1.348E-1

Results

Clustering Quality Anlalysis

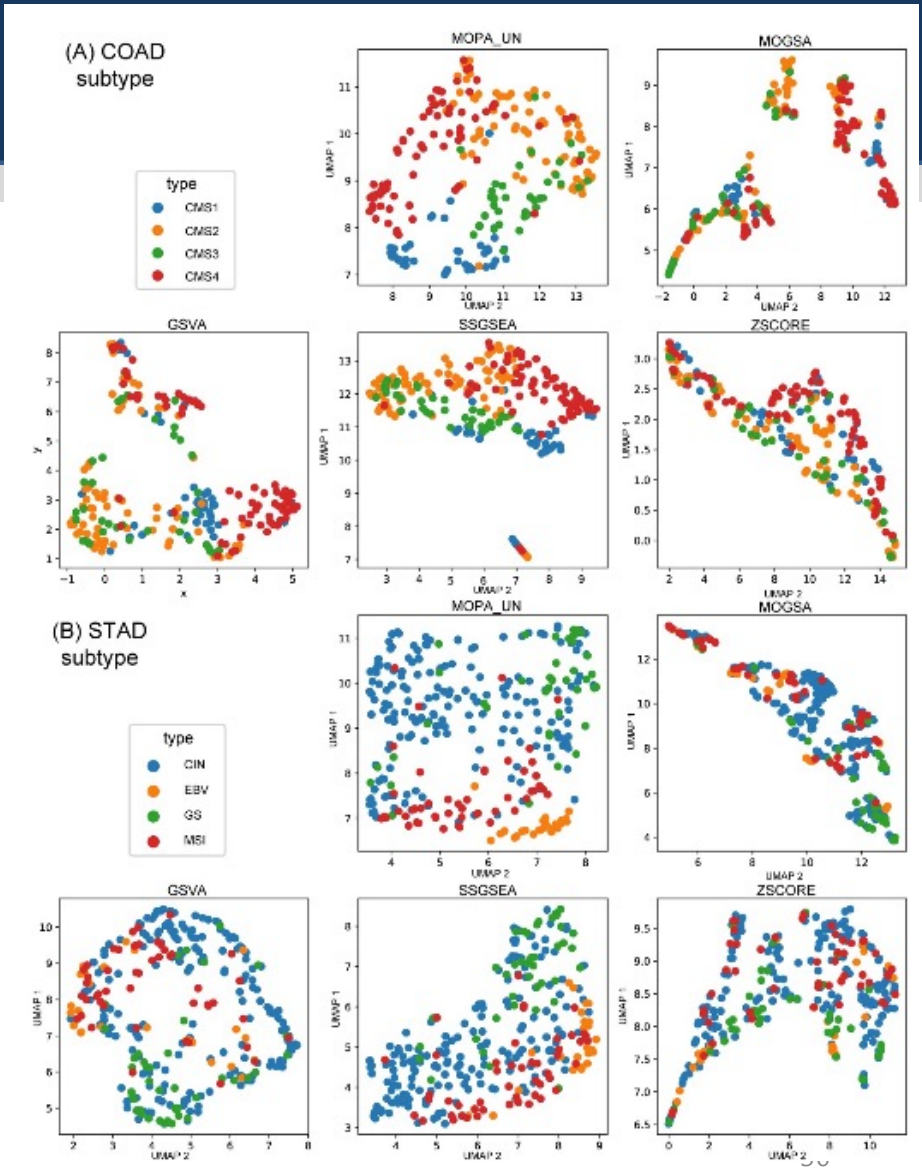
- Adjusted Rand Index (ARI) was employed to measure the quality of clustering
- Molecular subtypes served as the ground truth labels.
- ARI was calculated from bootstrap sampling (30% samples, 1,000 times).
- MOPA showed promising results in detecting sample subgroups in an unsupervised manner.



Results

MOPA's Flexibility in Label-less Situations

- In scenarios without clinical label info:
  - mES can be computed without any label info.
  - OCR requires labels; can use K-means clustering on mES matrix to create sample groups.
- MOPA's ability highlighted: COAD and STAD sample clusters closely matched the actual subtype sample groups.



Results

Use case Study: COAD

- The study aimed to understand pathways associated with COAD molecular subtypes using mES and OCR metrics via MOPA.
- From COAD data, 106 pathways had **significant survival p-values.**
- Samples were divided into high mES and low mES groups. Three most significant pathways were : 'Salivary secretion', 'Complement and coagulation cascades', and 'Staphylococcus aureus infection'.

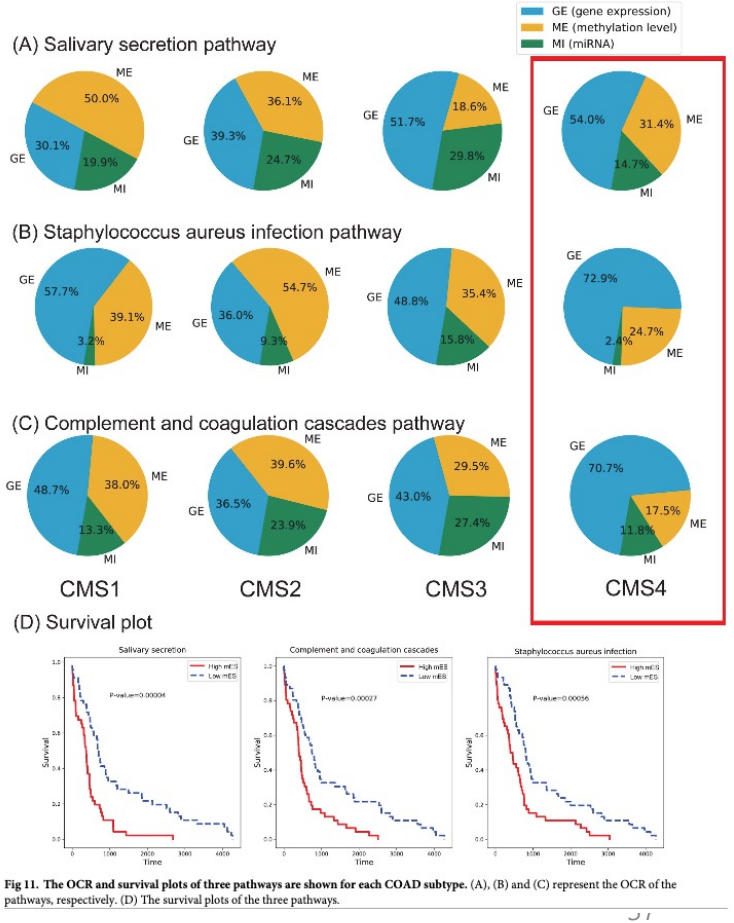


Fig 11. The OCR and survival plots of three pathways are shown for each COAD subtype. (A), (B) and (C) represent the OCR of the pathways, respectively. (D) The survival plots of the three pathways.

Results

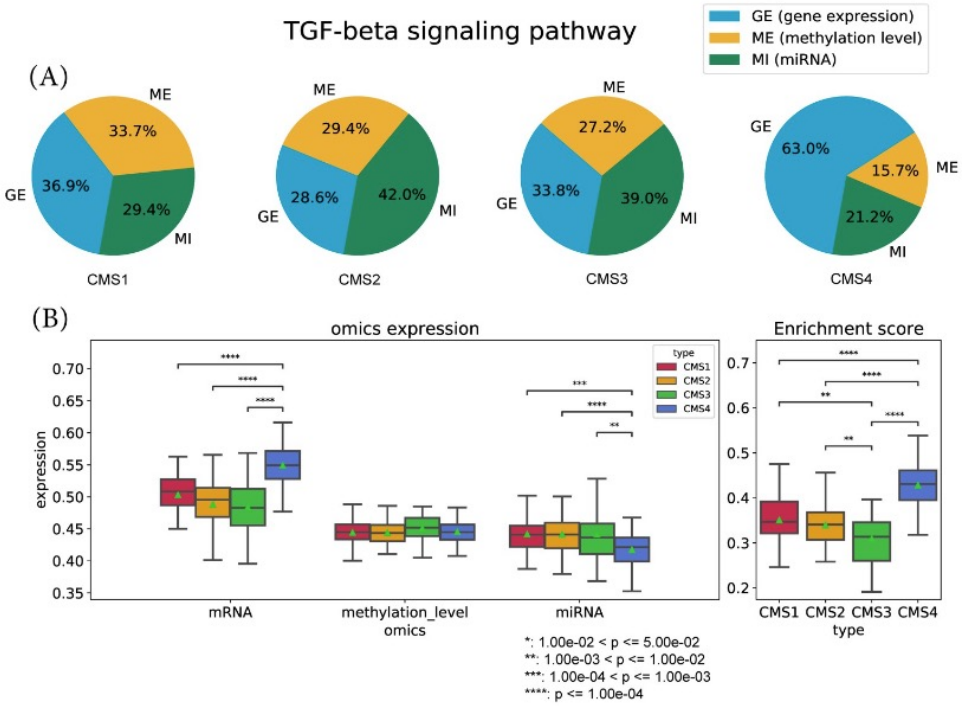
Use case Study: COAD

- Survival analysis
  - From the result, they observed that the survival probability of high mES group was significantly lower than other subtype samples.
  - The three pathways related to “TGF-beta signaling” pathway.
  - However, the p-value of “TGF-beta signaling” pathway was not as significant as the others.

Results

Use case Study: COAD

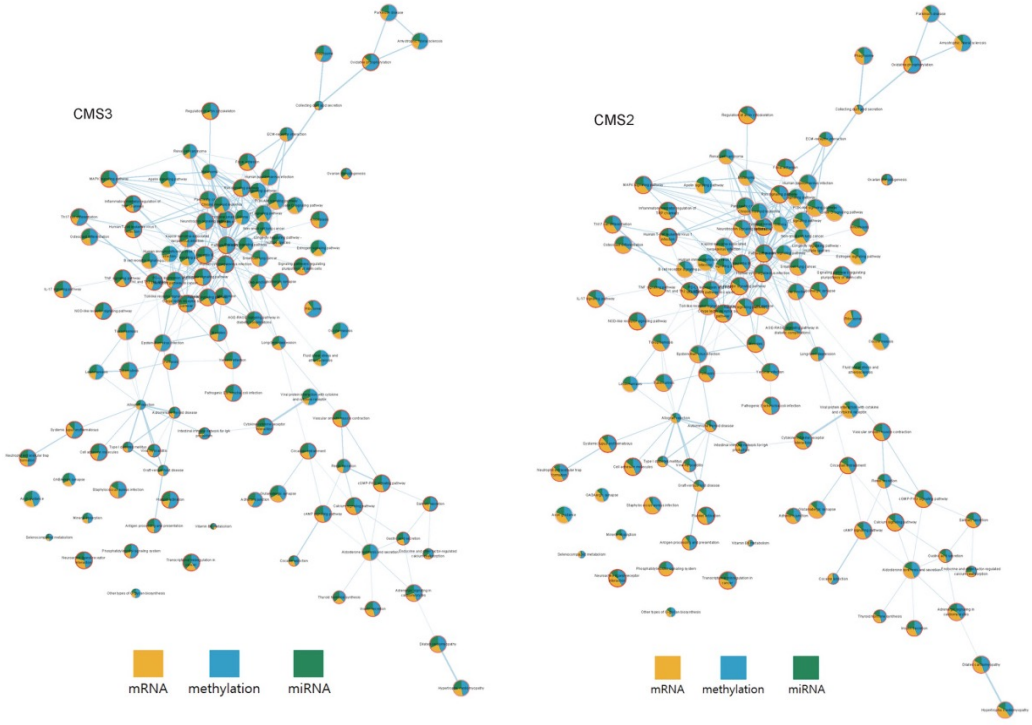
- According to the OCR of the “TGF-beta signaling pathway”, they observed that the CMS4 subtype had a distinctively different ratio of omics activation.
- The gene and miRNA expression significantly differed between CMS4 and the other subtypes.
- This was also observed in three other pathways in previously mentioned.
- Collectively, it implies that the CMS4 subtype yields a very different multi-omics landscape.



Results

Use case Study: COAD

- Network visualization
  - To compare the complete set of pathways with significant p-values, a pathway network specific to each subtype was constructed using Cytoscape.



Discussion

- **Objective**
  - **Interpret pathways** using multi-omics in terms of omics activation in cis-relation.
- **Approach**
  - Comparison of multi-omics activity across clinical feature (sub-group) using mES and OCR metrics.
  - Benefits over traditional methods: Easier interpretation than listing genes, which would require further enrichment analysis.
- **Performance**
  - Tested on nine different cancer multi-omics datasets. MOPA showed equal or superior performance compared to other tools.
- **Flexibility**
  - Not limited to just three omics types and not solely designed for cancer studies.

Discussion

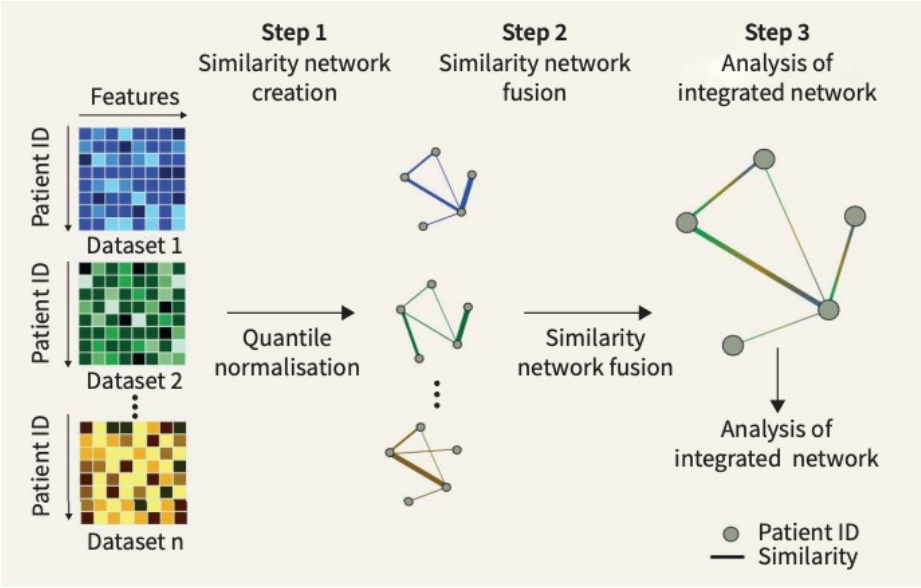
- **Limitations**
  - MOPA may not be suitable for small datasets due to tensor decomposition constraints.
  - Optimal performance observed with a rank of 120 across the three studied omics types.
  - Longer execution time compared to other methods:
- **Future Applications**
  - Extendable to other domains like single-cell COVID studies.
  - Importance: As multi-omics data grows in complexity and dimensionality, MOPA offers an accessible way to understand underlying biology from multiple perspectives.

Thank you for listening

Appendix 1. SNF

Similarity Network Fusion

- Step1. Creation of similarity network
  - Standardization of each omics
  - For each omic dataset, create a network:
    - Nodes: individual patients
    - Edges: Measure of similarity between patients
  - Quantile normalization to address differences in metric ranges
- Step2. Fusion of similarity network
  - To combine multiple patients' similarity networks from various –omics into one integrated network
  - Decompose each dataset's similarity into two
    - Global structure: overall similarity of a patient to all others
    - Local structure: similarity of a patients to its "K"-most similar patients
  - Iteratively fuse decomposed network by diffusing similarity information through common edges



Appendix 2. Performance measure

F1 Score

- **Definition:**
  - The F1 Score is a performance metric for binary classification. It is the harmonic mean of precision and recall**.**
- **Formula:**
  - $\boldsymbol{F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}}$
  - Precision is the ratio of correctly predicted positive observations to the total predicted positives.
    - $Precision = \frac{True\ positives}{True\ positives + False\ positives}$
  - Recall (or Sensitivity) is the ratio of correctly predicted positive observations to the all actual positives.
    - $Recall = \frac{True\ positives}{True\ positives + False\ negatives}$

Appendix 3. Performance measure

F1 Score for multi-class classification

- **Micro-average F1 Score:**
  - First, calculate the aggregate false positives, false negatives, and true positives across all the classes.
  - Then, use these aggregated counts to compute the overall precision and recall, and subsequently the F1 score.
- **Macro-average F1 Score:**
  - Compute the F1 score independently for each class and then take the average (without considering the class distribution).
  - This gives equal weight to each class, irrespective of its frequency.

Appendix 4. Performance measure

Adjusted Rand Index

- **Rand Index**
  - The Rand Index computes the **similarity** between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters.
  - Given two clusterings U (true labels) and V (predicted labels) of a set of n objects
    - a: **the number of pairs of objects** that are in the same set in U of and in the same set in V.
    - b: the number of pairs of objects that are in the different sets in U and V
    - $RI = \frac{a+b}{\binom{n}{2}}$

Appendix 5. Performance measure

Adjusted Rand Index

- **Expected Rand Index**
  - Let's denote by P_{ij} the number of object pairs that are both in cluster U_i of U and cluster V_i of V.
    - $E[a] = \sum_i \sum_j \binom{P_{ij}}{2}$
    - $E[a] = \sum_i \sum_j \binom{n_i}{2} - \binom{P_{ij}}{2}$
    - $E[RI] = \frac{E[a]+E[b]}{\binom{n}{2}}$
- **Adjusted Rand Index**
  - The ARI adjusts the Rand Index by considering the random chance of any two points being clustered together.
  - Mathematically, the Adjusted Rand Index is given by: $ARI = \frac{RI - Expected\ RI}{Max\ RI - Expected\ RI}$

Appendix 6. Performance measure

Example of ARI calculation

- Let's assume we have 5 objects {A,B,C,D,E}
  - True labels
    - Cluster U1: {A,B}
    - Cluster U2: {C,D,E}
  - Predicted labels
    - Cluster V1: {A,C}
    - Cluster V2: {B,D,E}
- Let's compute the values needed:
  - (A,C) is in Cluster V1, but they are in different clusters in U. So they don't contribute to a.
  - (D,E) are in Cluster V2 and also in Cluster U2. This is the only pair that contributes to a.
  - $a = 1$
  - $b = 5$ |{(A,B), (A,D), (A,E), (B,C), and (C,D)}|
  - In case of simplified version of $E[RI]$, we can assume $E[RI]$ = 0.5.
  - Then $ARI = \frac{0.6 - 0.5}{1 - 0.5} = 0.2$

Appendix 7. UMAP

Uniform Manifold Approximation

- **Purpose**
  - The primary purpose of UMAP is to capture both local and global structures of data in lower-dimensional space.
  - This serves Visualization: Making it easier to visualize and interpret complex high-dimensional data by projecting it into 2D or 3D.
- Methods:
  - **Construct a Graph(Fuzzy simplical set construction)**: For each data point in the high-dimensional space, UMAP build a neighborhood graph where nearby points are connected by edges.
  - **Optimize the Embedding**: The algorithm then seeks a low-dim representation where the distance between points in the new space respects their proximity in the original high-dim.
    - This is done by minimizing the ”cross-entropy” between the **distributions of distances** in the two spaces.