

# **Lecture 2**

## **Unsupervised Clustering of Multi-Omics Data**

April 13, 2023



**Instructor: Sierra Niemiec**

# Overview

## I. Introduction

I. Purpose/Goal

II. Challenges

## II. Review of Concepts: Unsupervised Machine Learning & Clustering

## III. Multi-Omics Unsupervised Clustering

## IV. Method Highlight: Similarity Network Fusion (SNF)

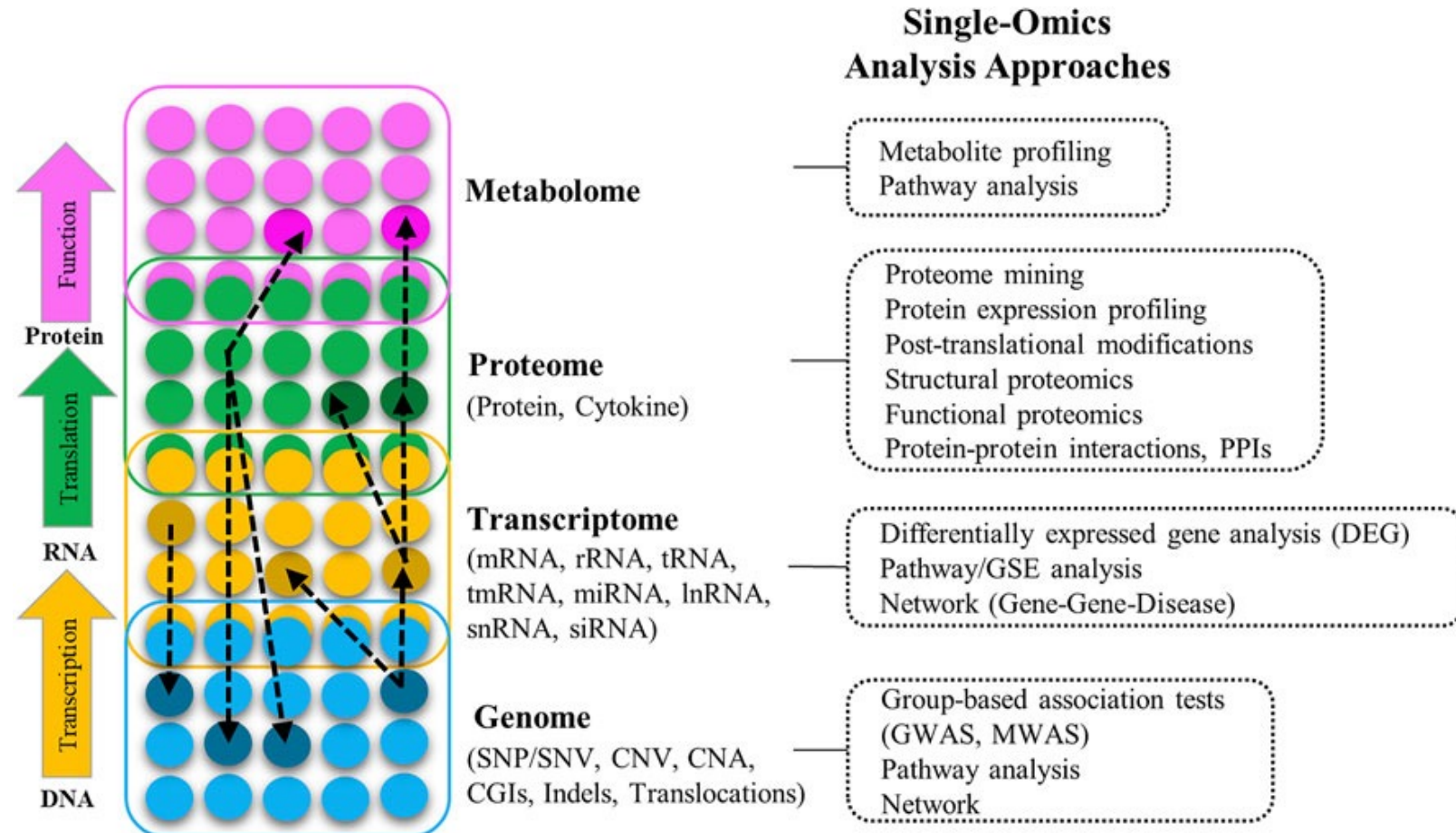
- Introduction
- Overview of Method
- Applicable Data
- Assumptions & Considerations
- Pros/Cons
- Extensions

## V. Conclusion

# I. Introduction

# Purpose/Goal

- Heterogenous diseases states, patient profiles
- Capture unique patterns of sub-populations to guide treatment, understanding of individual's disease
- Precision medicine

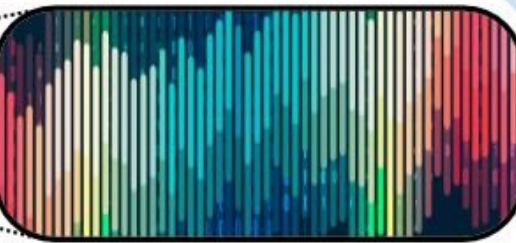
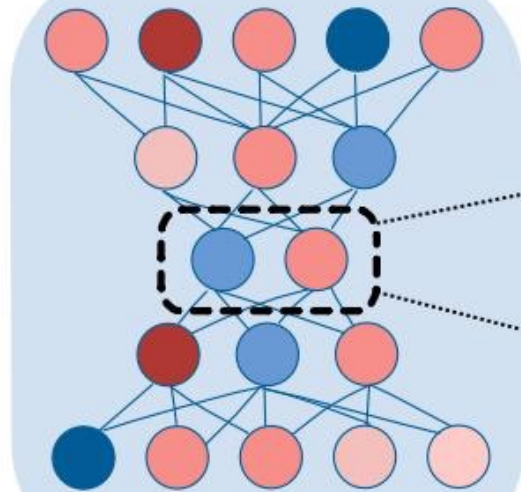


Vahabi & Michailidis, 2022

One or more omics layers are arranged, normalized, and scaled



Dimensionality reduction and multi-omics integration using deep learning

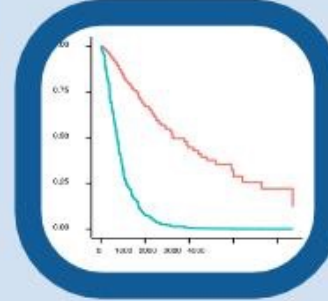


Multi-omic fingerprints detected by MAUI are useful for a wide spectrum of precision oncology applications

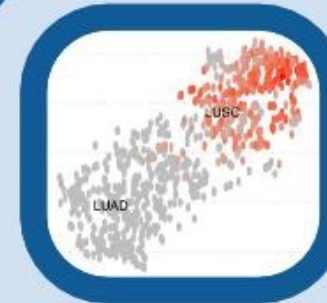
Pan-cancer classification



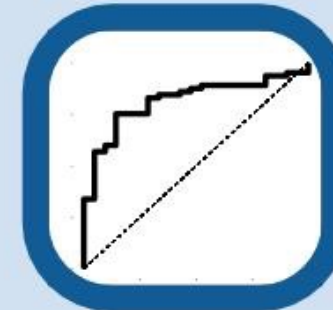
Modeling clinical features



Cancer subtype modeling



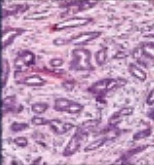

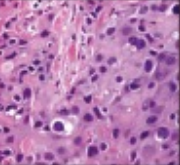
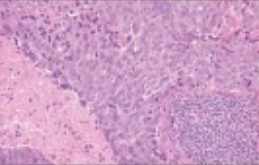
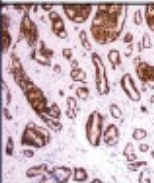
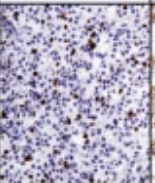
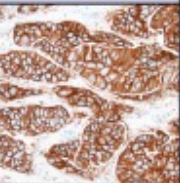
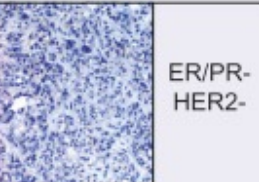
Drug response prediction



Uyar et al., 2021



# Example: Breast Cancer

Molecular subtype	Luminal (A and B)		HER2	Basal
Genetic profile	↑Luminal CKs and ER-related genes (A>B) B↑ in proliferation-related genes		↑HER2-related genes	↑Basal CKs
Histologic correlates				
	A Lower-grade ER+	B Higher-grade ER+	High-grade, ± apocrine features	High-grade, sheet-like, necrosis inflammation
Surrogate markers				
	A Strong ER+, PR±, HER2-, low Ki67	B Weaker ER+, PR±, HER2±, ↑Ki67	HER2+, ± ER/PR	ER/PR-HER2-  CK5/6± EGFR±
Prognosis	Good	Intermediate	Worse	Worse
Response to chemotherapy	Lower	Intermediate	Higher	Higher
Targeted therapies	Hormone therapies		HER2-targeted therapies	Currently investigational

Allison et al., 2017

# Data Integration Challenges for Multi-Omics

- Curse of dimensionality: small number of samples compared to the large number of measurements
- Data heterogeneity: differences in scale, collection bias and noise in each data set
- Complexity of inter-omics variations: complementary nature of the information provided by different types of data.
- Missing data: methods require data for matched subjects
  - Can address with imputation to some extent

# II. Review of Concepts: Unsupervised Machine Learning & Clustering

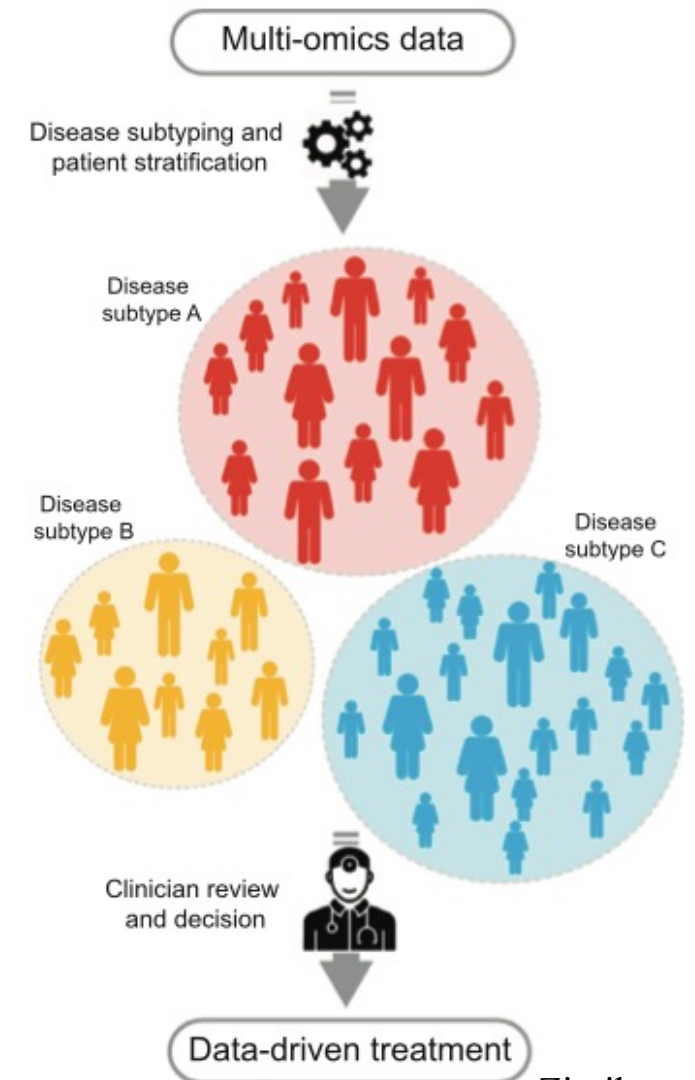


# Review: Unsupervised Machine Learning

- 2 Common forms of UML
- Dimension Reduction
  - Situations where  $p \gg n$  (common in omics)
  - Can a smaller number of features (e.g. 10) adequately represent all  $p$  (e.g. 1000 features)?
  - Principal component analysis (PCA), factor analysis
  - Visual Interpretations
- Clustering Methods
  - Identify groups among set of objects
  - **Cluster samples** (rows of  $X$ ): identify distinct subgroups of disease
  - Cluster features (columns of  $X$ ): identify groups of similar genes
  - Objects within same cluster should be similar while objects in 2 separate clusters should be different

# Unsupervised Learning: Goals for Multi-Omics

- Classify
  - Disease, sample subtype
  - Find meaningful patterns or groups/clusters within (“fingerprint”)
- Discover biomarkers/modules
  - Prioritize genes associated with a disease
  - Find genes that are co-expressed
  - Co-expressed genes may have similar functions or may be co-regulated
  - Clues into gene function
- Define outcomes for subtypes
  - Do these clusters help predict patient survival, treatment, prognosis, more



Zitnik et al., 2019

# Clustering

- Grouping objects based on similarity within the group (internal) and dissimilarity to the objects belonging to other groups (intra-cluster)
- Need:
  - Proximity/Distance metric
  - Criterion to evaluate clustering
  - Algorithm to compute clustering
    - Optimizing criterion function

# Distance/Proximity Metrics

Distance Measure	Equation	Time complexity	Advantages	Disadvantages	Applications
Euclidean Distance	$d_{\text{euc}} = \left[ \sum_{i=1}^n (x_i - y_i)^2 \right]^{\frac{1}{2}}$	O(n)	Very common, easy to compute and works well with datasets with compact or isolated clusters [27,31].	Sensitive to outliers [27,31].	K-means algorithm, Fuzzy c-means algorithm [38].
Average Distance	$d_{\text{ave}} = \left( \frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2 \right)^{\frac{1}{2}}$	O(n)	Better than Euclidean distance [35] at handling outliers.	Variables contribute independently to the measure of distance. Redundant values could dominate the similarity between data points [37].	K-means algorithm
Weighted Euclidean	$d_{\text{we}} = \left( \sum_{i=1}^n w_i (x_i - y_i)^2 \right)^{\frac{1}{2}}$	O(n)	The weight matrix allows to increase the effect of more important data points than less important one [37].	Same as Average Distance.	Fuzzy c-means algorithm [38]
Mahalanobis	$d_{\text{mah}} = \sqrt{(x - y)S^{-1}(x - y)^T}$	<u>O(3n)</u>	Mahalanobis is a data-driven measure that can ease the distance distortion caused by a linear combination of attributes [35].	It can be expensive in terms of computation [33]	Hyperellipsoidal clustering algorithm [30].
Pearson coefficient	$\text{Pearson}(x, y) = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_x)^2} \sqrt{\sum_{i=1}^n (y_i - \mu_y)^2}}$	O(2n)	*Results in accurate outcomes using the hierarchical single-link algorithm for high dimensional datasets.	-	Partitioning and hierarchical clustering algorithms.

\*Points marked by asterisk are compiled based on this article's experimental results.

Shirkhorshidi et al., 2015

# Common Clustering Algorithms

TABLE III COMPARATIVE ANALYSIS OF K-MEANS,  
HIERARCHICAL AND SELF ORGANIZATION MAP ALGORITHM

<b>K-MEANS BASED DATA CLUSTERING</b>	<b>AGGLOMERATIVE HIERARCHICAL ALGORITHM</b>	<b>SELF ORGANIZATION MAP</b>
<b>Partitioning Based Method</b>	<b>Based on hierarchical tree</b>	<b>Based on neural network</b>
<b>Input: k, dataset, randomly chosen k centroids</b>	<b>Input randomly dataset</b>	<b>Random input vector from training dataset</b>
<b>Objective: Minimizing sum of squared distance</b>	<b>Objective: Minimizing sum of squared distance</b>	<b>Multidimensional data is mapped by competitive and unsupervised learning</b>
<b>Final clustering may converge to local optima</b>	<b>Final clustering may converge to local optima.</b>	<b>Final clustering may converge to local optima.</b>
<b>Time complexity: <math>O(n*k*d*i)</math> Where n= no. of data points k= no. of clusters d= dimension of data i= no. of iterations</b>	<b>Time complexity: <math>O(n^2 \lg n)</math></b>	<b>Time complexity: <math>O(m^2 * l)</math> Where m= dimensional input vector l=no of weight vector</b>

Kumar & Asger, 2015

# Axioms for “Good” Clustering

- Scale Invariance:
  - Clustering algorithm should not modify its results when all distances between points are scaled by the factor determined by a constant
- Consistency:
  - Clustering results do not change if the distances within clusters decrease and/or the distances between clusters increase
- Richness:
  - Clustering function must be flexible enough to produce any arbitrary partition/clustering of the input data set
- “Kleinberg proves the following theorem: For every  $n \geq 2$ , there is no clustering function  $f$  that satisfies scale invariance, richness, and consistency. [3]”
- Impossible for any clustering procedure to be able to satisfy all three axioms.
- Practical clustering algorithms must make trade-offs



# Clustering Challenges

- Defining distance metric
- Different data structures
- Identifying number ( $k$ ) of clusters
- Assess clustering when truth is unknown

# Internal Cluster Validation

- No ground truth labels
- Focus on:
  - Cluster cohesion and separation
  - Statistical analysis of the proximity matrix
  - The dendrogram generated by hierarchical clustering algorithms
- Examples:
  - **Silhouette Coefficient**, Calinski-Harabasz Index, Davies-Bouldin Index

# Silhouette Score

$$s(i) = \frac{b_i - a_i}{\max(a_i, b_i)}$$

- Definition:

For object (subject)  $i$

$a_i$ =ave. dist of object  $i$  to other objects in same cluster (mean intra-cluster distance)

$b_i$ =ave. dist of object  $i$  to other objects in closest neighboring cluster (mean nearest-cluster distance)

- Range:  $[-1, 1]$

- Visually inspect the similarities within clusters and differences across clusters
- Assess how well-assigned each individual point is & how close each point in a cluster is to points in the neighboring clusters
  - 1: ‘appropriate’ cluster; well-assigned; further away the cluster’s samples are from the neighboring clusters samples
  - 0: right at the inflection point between two clusters; sample is on or very close to the decision boundary between two neighboring clusters
  - -1: “inappropriate clustering”; assigned to wrong cluster
- Global Silhouette Score: average of Silhouette values
  - Describe the entire population’s performance with a single value
- Disadvantage: can be extremely expensive to compute on all  $n$  points
  - Compute the distance of  $i$  from all other  $n - 1$  points for each  $i$ , complexity of  $O(n^2)$ .

# External Cluster Validation

- Requires ground “truth” labels
- Externally-provided information to evaluate the quality of the clustering results
- External validation metrics are also useful when comparing the results provided by different clustering algorithms
- Examples:
  - Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Jaccard Index, Homogeneity, Completeness and V-measure

# Rand Index

- Similarity measure
- Compares all pairs of samples predicted and true clusterings

- Range: [0, 1]
- 1: perfect match

- Adjusted Rand Index: adjusted for chance

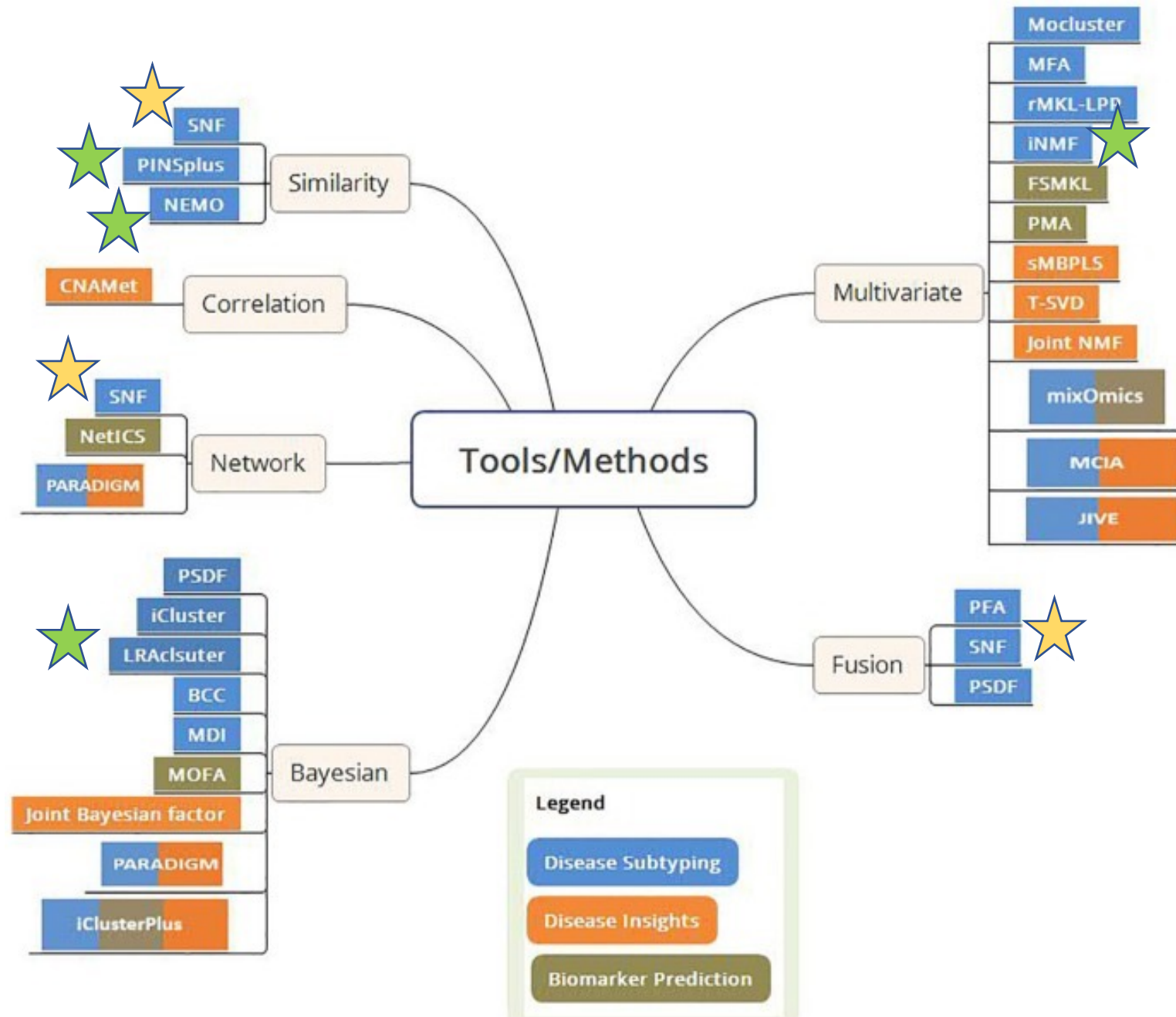
- Range: [0, 1]
- 0: Random labelling
- 1: Clusters identical

$$RI = \frac{\text{Number of Agreeing Pairs}}{\text{Number of Pairs}}$$

$$ARI = \frac{RI - \text{Expected RI}}{\text{Max}(RI) - \text{Expected RI}}$$

# III. Multi-Omics Unsupervised Clustering





Subramanian et al., 2020

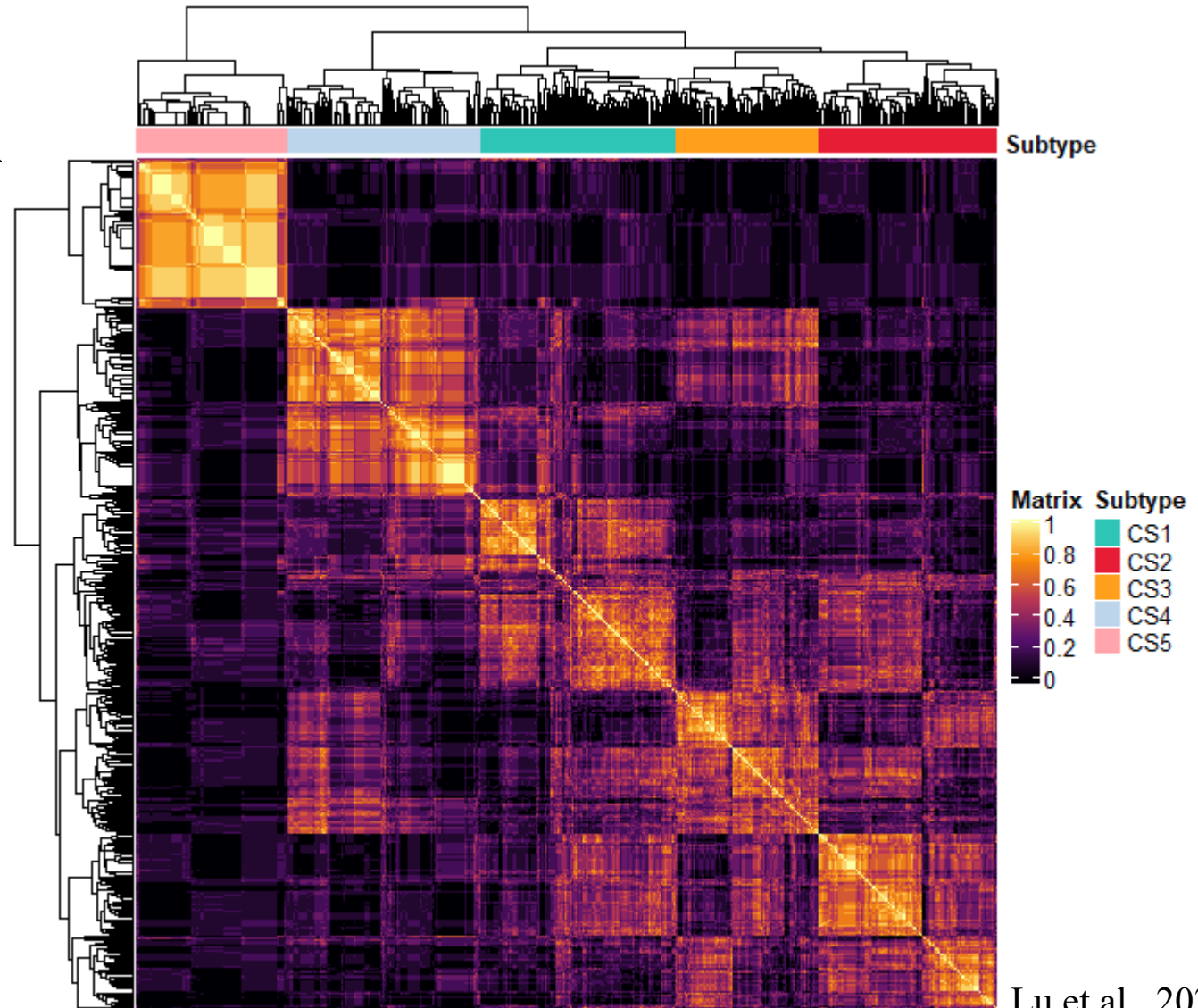
# Recommended Methods

**TABLE 2** Four scenarios with recommended methods

Scenarios	Required characteristics for method	Recommended methods
I. ( <i>Feature selection</i> ): The need to identify clinically relevant disease subtypes and driving molecular signatures which can be targeted for treatment	Performing both sample clustering and feature selection	iCluster; iClusterPlus; iClusterBayes; intNMF; IS-K means; CIMLR; PSDF
II. ( <i>Mixed-type data</i> ): Large scale genomic data of mixed-type in large consortia	Integrating mixed type of data	iClusterPlus; iClusterBayes; moCluster; LRAcluster; MDI; SNF; CIMLR; rMKL-LPP; PINS; PINSPlus
III. ( <i>Computational efficiency</i> ): Concern on the computational resources and consumption of time	Computationally efficient	Spectrum; SNF; ab-SNF; NEMO; CIMLR; rMKL-LPP
IV. ( <i>Knowledge integration</i> ): Leveraging the prior knowledge	Incorporating prior information	IS-K means; PARADIGM

# Consensus Clustering

- Matrix per clustering algorithm  $t$ :
  - $M_{ij}^{(t)} = 1$  : sample  $i$  and  $j$  are clustered in same subtype
  - $M_{ij}^{(t)} = 0$  : otherwise
- Consensus Matrix:  $CM = \sum_{t=1}^{t_{\{max\}}} M^{(t)}$ 
  - Probability matrix represents how many times samples belonging to the same subtype can be clustered together by different multi-omics clustering methods
  - Robust pairwise similarities for samples across different multi-omics integrative clustering algorithms
  - Clustering on  $CM$  (hierarchical)
  - Looking for:
    - Perfect diagonal rectangle
    - Input values are 0 and 1 only because all algorithms derived the same clustering results



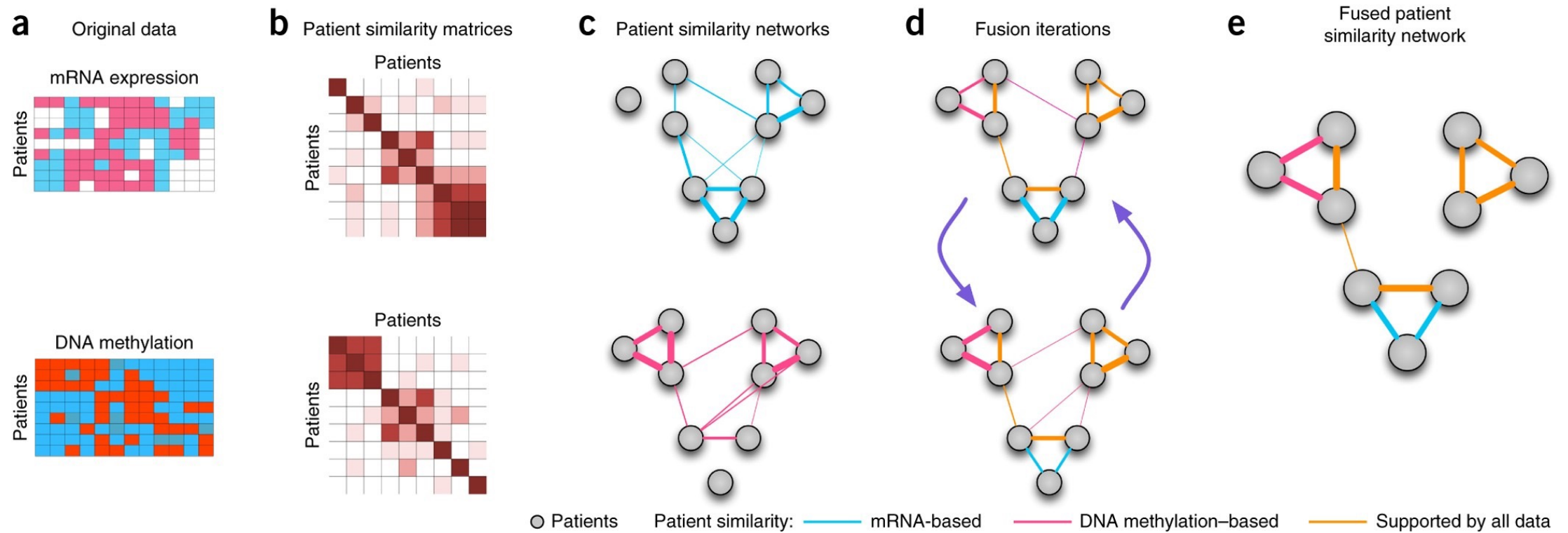
Lu et al., 2021

# IV. Methods Highlight: Similarity Network Fusion (SNF)

# SNF - Introduction

- Uses networks of samples as a basis for integration
- Various omics datasets (both in type and number)
- The fused network captures both shared and complementary information from different data sources
- Insight into how informative each data type is to the observed similarity between samples
- Useful information even from a small number of samples
- Robust to noise and data heterogeneity
- Scales to a large number of genes
- Efficiently identify subtypes among existing samples by clustering and predict labels for new samples based on the constructed network
- **Questions it can address:**
  - Samples' molecular/phenotypic profiles
  - Subtyping and label prediction (sub-populations, drug response)







# SNF - Introduction



Wang et al., 2014



# Step 1: Patient Similarity Matrix, $m_{1,...,M}$

	 $n_1$	 $n_2$	 $n_n$
 $n_1$	0	$w_{\{1,2\}}$	$w_{\{1,n\}}$
 $n_2$	$w_{\{2,1\}}$	0	$w_{\{2,n\}}$
 $n_n$	$w_{\{1,n\}}$	$w_{\{2,n\}}$	0

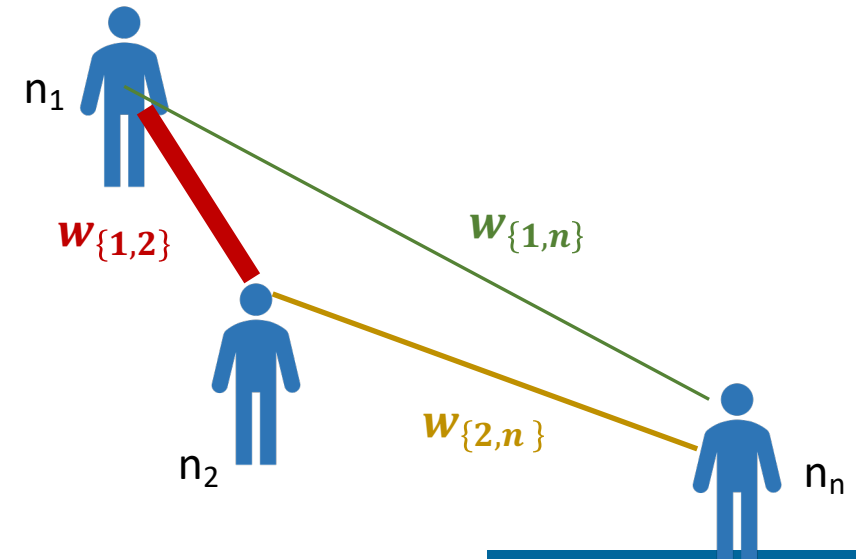
$$W(i,j) = \exp\left(-\frac{\rho^2(x_i, x_j)}{\mu \varepsilon_{\{i,j\}}}\right)$$

Graph:  $G = (V, E)$

$V$  – patients, 1-n

$E$ : edges, weighted by similarity between patients,  $n \times n$  matrix ( $W$ )

Similarity/weights: scaled exponential similarity kernel



# Step 2: Network Fusion

- Full, Normalized Kernel,  $\mathbf{P}$

$$\bullet \mathbf{P}(i, j) = \begin{cases} \frac{W(i, j)}{2 \sum_{k \neq i} W(i, k)}, j \neq i \\ \frac{1}{2}, j = i \end{cases}$$

- Sparse Kernel, Local Affinity,  $\mathbf{S}$

$$\bullet \mathbf{S}(i, j) = \begin{cases} \frac{W(i, j)}{\sum_{k \in N_i} W(i, k)}, j \in N_i \\ \mathbf{0} \text{ otherwise} \end{cases}$$

Iterate (2 data types)

$$\bullet \mathbf{P}_{\{t+1\}}^{(1)} = \mathbf{S}^{(1)} \times \mathbf{P}_t^{(2)} \times (\mathbf{S}^{(1)})^T$$

$$\bullet \mathbf{P}_{\{t+1\}}^{(2)} = \mathbf{S}^{(2)} \times \mathbf{P}_t^{(1)} \times (\mathbf{S}^{(2)})^T$$

Overall Status Matrix

$$\bullet \mathbf{P}^{(c)} = \frac{\mathbf{P}_t^{(1)} + \mathbf{P}_t^{(2)}}{2}$$

Multiple data sets:

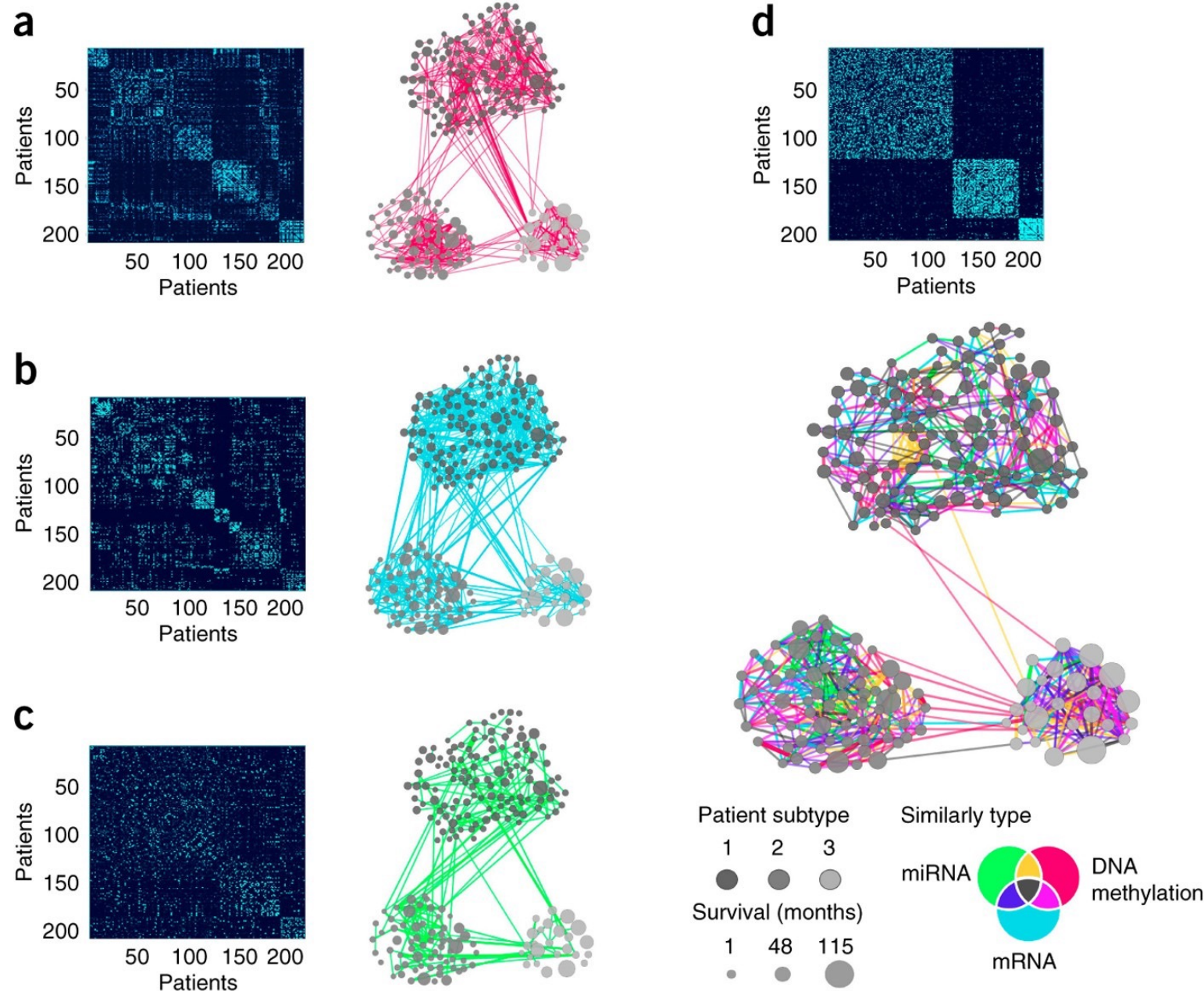
$$\bullet \mathbf{P}_{\{t+1\}}^{(1)} = \sum_{k \in N_i} \sum_{l \in N_j} \mathbf{S}^{(1)}(i, k) \times \mathbf{S}^{(1)}(j, l) \times \mathbf{P}_t^{(2)}(k, l)$$

# SNF – Downstream Analyses

- Network-clustering: disease subtyping
  - Spectral clustering
- Use patient clusters/sub-types to test for associations with outcomes
  - E.g. Cancer types, survival, drug response
- Identify common and complementary signals across data type
- Reduce noise by aggregating across multiple types of data
- Provide insight into the relative importance of each data source for determining patient similarity
  - Refine the understanding of the heterogeneity within each subtype

**Figure 2: Patient similarities for each of the data types independently compared to SNF fused similarity.**

(a–d) Patient-to-patient similarities for 215 patients with GBM represented by similarity matrices and patient networks, where nodes represent patients, edge thickness reflects the strength of the similarity, and node size represents survival. Clusters are coded in grayscale (subtypes 1–3) and arranged according to the subtypes revealed through spectral clustering of the combined patient network. The clustering representation is preserved for all four networks to facilitate visual comparison. DNA methylation (a), mRNA expression (b), miRNA expression (c) and SNF-combined similarity matrix and network (d; see [Supplementary Fig. 11](#) for more information about network edges).



Wang et al., 2014

# SNF – Applicable Data

- Matched data by patients
- N of omics datasets: multiple
- Types: Continuous, binary, and categorical
- Clinical:
  - Such as microbiome and metabolomics data, questionnaires and functional magnetic resonance imaging, together with genomic, clinical and demographic data
  - Just requires: the data can be used to identify similarity between patients

# SNF – Assumptions & Considerations

- Hyperparameters
- Simulations show method not sensitive to different settings of hyperparameters
- $\mu$  : from calculating patient similarity networks (individual data types)
  - Recommend setting  $\mu$  in the range of  $[0.3, 0.8]$
- K: number of clusters
  - The rule of thumb for choosing parameter K:  $K = N/C$
  - N : number of patients
  - C : number of clusters that is believed to be in the data
  - If C is unknown, we usually set  $K \approx N/10$
- KNN implies local similarities are more reliable than remote ones
- Distance/Similarity Metrics
  - Euclidean for continuous data, Chi-squared for binary, categorical
  - Others are possible (correlation, etc.)
- Identifying number of clusters for subtyping with spectral clustering, recommend eigengap



# SNF – Assumptions & Considerations

- Outliers
- Missing data, imputation
  - KNN, same K as in method
  - Removal of patients with more than 20% missing data in a certain data type
- Normalization
  - $\tilde{f} = \frac{f - E(f)}{\sqrt{Var(f)}}$
  - $f$ : any biological feature
  - $\tilde{f}$ : biological feature following normalization
  - $E(f)$ : empirical mean of  $f$
  - $Var(f)$ : empirical variance of  $f$
- Combining data types
  - Normalized mutual information (NMI) to check for concordance of data types

# SNF – Pros

- Flexible: Multiple omics types and multiple number of data sets
- Can integrate various gene-interaction data, such as physical interactions, coexpression and colocalization data
- Provide insight into how informative each data type is to observed similarities of samples
- Small number of samples
- Robust to noise, data heterogeneity
- Computationally efficient
  - Guaranteed to converge
- Number of iterations  $> 20$  always enough to converge
- Scales to large numbers of features

# SNF – Cons

- Requires matched samples, no missingness
- Does not distinguish between data types → false fusion
- Uses Euclidean distance to calculate the similarity matrices between the samples
  - May be incapable of capturing the data's unique structure

# SNF - Extensions

- **DSSF** (Deep Subspace Similarity Fusion) ([Yang et al., 2018](#)): uses an auto-encoder to improve the discriminative similarity between samples
- **AFN** (Affinity Network Fusion) ([Ma and Zhang, 2018](#)) : enables the consideration of patients' pairwise distances
- **NEMO** (NEighborhood based Multi-Omics clustering) ([Rappoport and Shamir, 2019](#)) : enables the computation of global kernel matrix without performing any imputation on the missing observation, handles unmatched samples (different sample sizes in different Omics-types)
- **INF** (Integrative Network Fusion) ([Chierici et al., 2020](#)) : utilizes SNF within a predictive framework including RF ([Breiman, 2001](#)) (Random Forest) and LSVM ([Cortes and Vapnik, 1995](#)) (Linear Support Vector Machine)

# V. Conclusion

# Summary

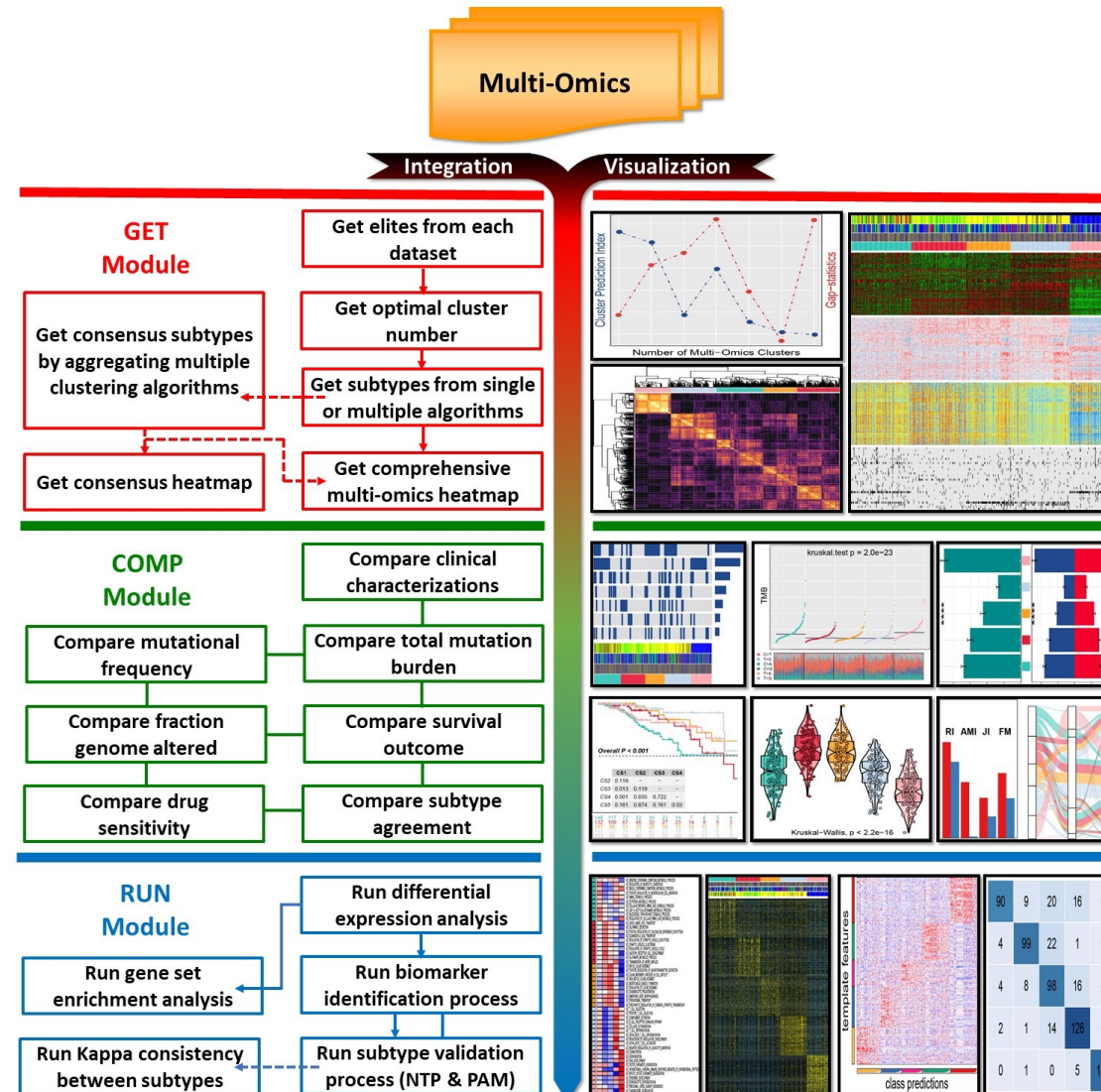
- Identify distinct molecular phenotypes
  - Precision medicine
  - Tailor treatment, prognosis
  - Understand unique biological underpinnings
- Data integration challenges
  - Large  $p$ , small  $n$ ; data heterogeneity; complexity of inter-omics data; missing data
- Review of unsupervised clustering techniques
  - Always trade-offs with each method
  - Use multiple methods to get different views of data
  - Clustering needs to be evaluated with various metrics
- Highlighted SNF as method for multi-omics clustering
- Consensus clustering can provide more stable, robust findings

# Overall Considerations

- Hyperparameter tuning
- Identification of number of clusters
  - Gap statistic, cluster prediction index, grid & random searches
- Definition of clusters, structure
- Feature Selection
- Computational Resources



# Lab Session: MOVICS



Lu et al., 2021

# References

- Allison, K. H. (2017). Chapter 21—Molecular Testing in Breast Cancer. In W. B. Coleman & G. J. Tsongalis (Eds.), *Diagnostic Molecular Pathology* (pp. 257–269). Academic Press. <https://doi.org/10.1016/B978-0-12-800886-7.00021-2>
- Kumar, S., & Asger, D. M. (2015). *Analysis Clustering Techniques in Biological Data with R*. 6.
- Lu, X., Meng, J., Zhou, Y., Jiang, L., & Yan, F. (2021). MOVICS: An R package for multi-omics integration and visualization in cancer subtyping. *Bioinformatics*, 36(22–23), 5539–5541. <https://doi.org/10.1093/bioinformatics/btaa1018>
- Palacio-Niño, J.-O., & Berzal, F. (2019). *Evaluation Metrics for Unsupervised Learning Algorithms* (arXiv:1905.05667). arXiv. <http://arxiv.org/abs/1905.05667>
- Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. *PLOS ONE*, 10(12), e0144059. <https://doi.org/10.1371/journal.pone.0144059>
- Allison, K. H. (2017). Chapter 21—Molecular Testing in Breast Cancer. In W. B. Coleman & G. J. Tsongalis (Eds.), *Diagnostic Molecular Pathology* (pp. 257–269). Academic Press. <https://doi.org/10.1016/B978-0-12-800886-7.00021-2>
- Kumar, S., & Asger, D. M. (2015). *Analysis Clustering Techniques in Biological Data with R*. 6.

# References

- Lu, X., Meng, J., Zhou, Y., Jiang, L., & Yan, F. (2021). *MOVICS: An R package for multi-omics integration and visualization in cancer subtyping*. *Bioinformatics*, 36(22–23), 5539–5541. <https://doi.org/10.1093/bioinformatics/btaa1018>
- Palacio-Niño, J.-O., & Berzal, F. (2019). *Evaluation Metrics for Unsupervised Learning Algorithms* (arXiv:1905.05667). arXiv. <http://arxiv.org/abs/1905.05667>
- Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. *PLOS ONE*, 10(12), e0144059. <https://doi.org/10.1371/journal.pone.0144059>
- Uyar, B., Ronen, J., Franke, V., Gargiulo, G., & Akalin, A. (2021). *Multi-omics and deep learning provide a multifaceted view of cancer* [Preprint]. *Bioinformatics*. <https://doi.org/10.1101/2021.09.29.462364>
- Vahabi, N., & Michailidis, G. (2022). Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review. *Frontiers in Genetics*, 13, 854752. <https://doi.org/10.3389/fgene.2022.854752>
- Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3), 333–337. <https://doi.org/10.1038/nmeth.2810>
- Zhang, X., Zhou, Z., Xu, H., & Liu, C. (2022). Integrative clustering methods for multi-omics data. *WIREs Computational Statistics*, 14(3). <https://doi.org/10.1002/wics.1553>
- Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50, 71–91. <https://doi.org/10.1016/j.inffus.2018.09.012>