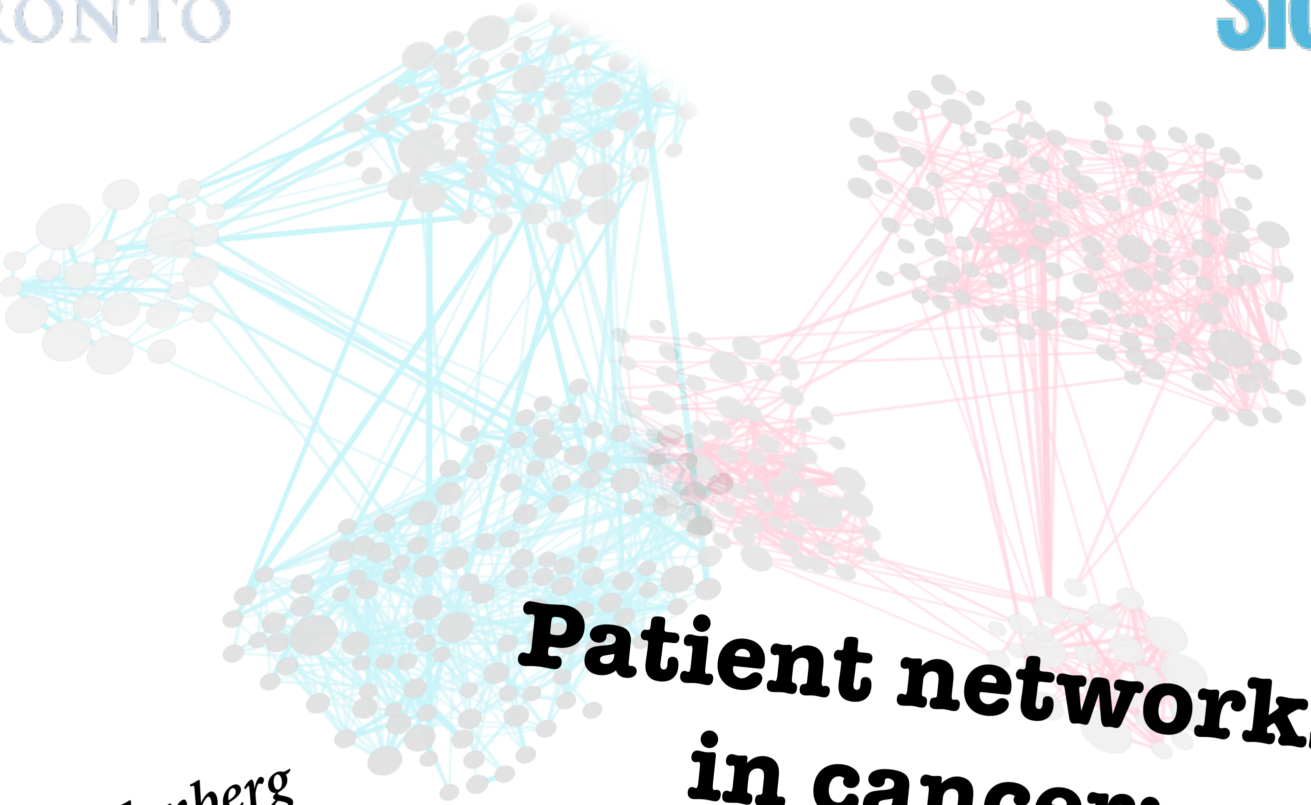
UNIVERSITY OF TORONTO

SickKids



Patient networks in cancer: a platform for data integration

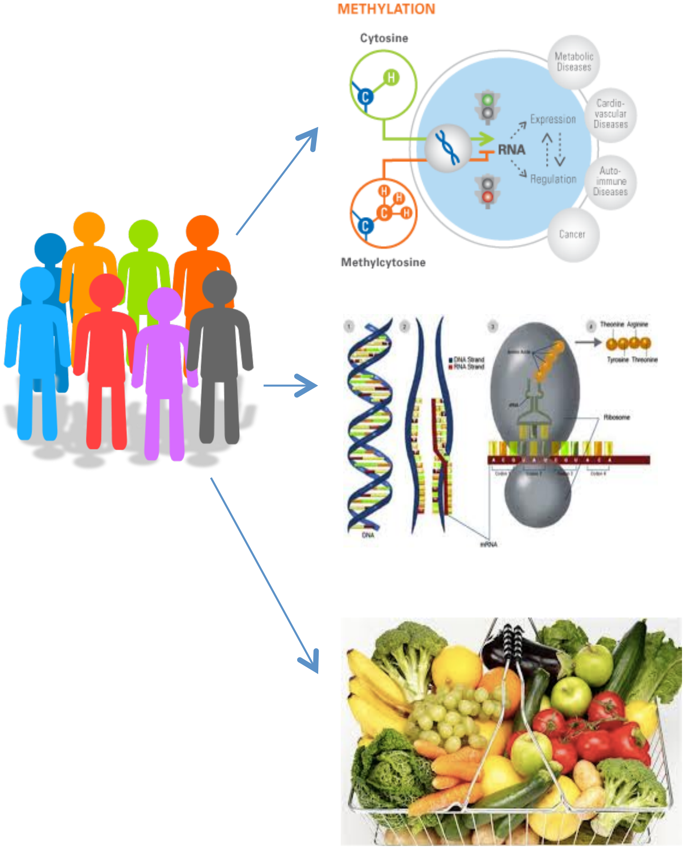
Anna Goldenberg and The Goldenberg Lab

Outline

- Data integration – problem setup
- Patient network representation – why and how
- Similarity Network Fusion – novel integration method
- Network driven analysis:
  - Cancer heterogeneity
  - Differential feature selection
- Missing data
  - Random entries
  - Patients
- Taking networks further:
  - Survival analysis (novel formulation)
  - Personalized medicine

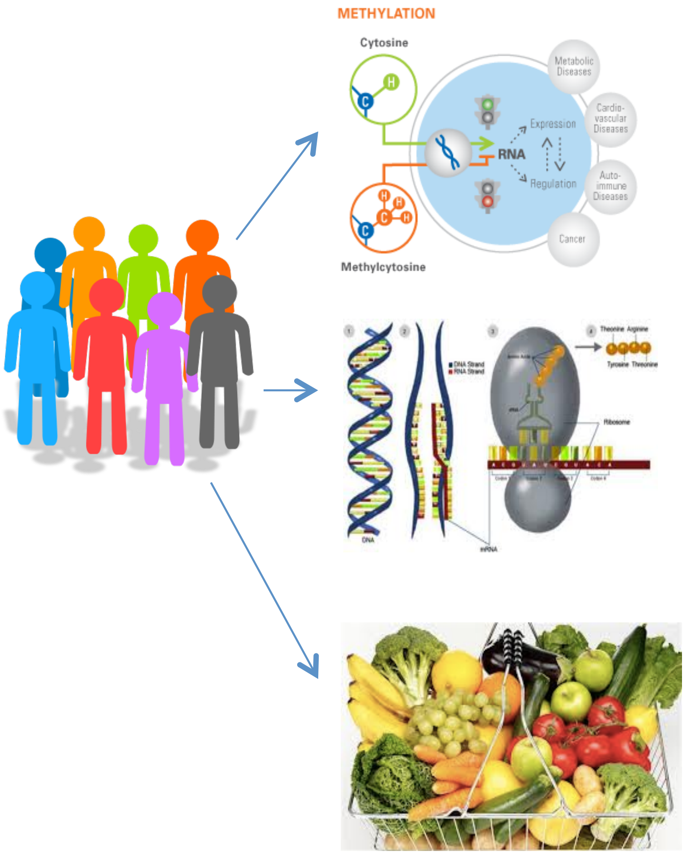
Problem setup

Problem setup



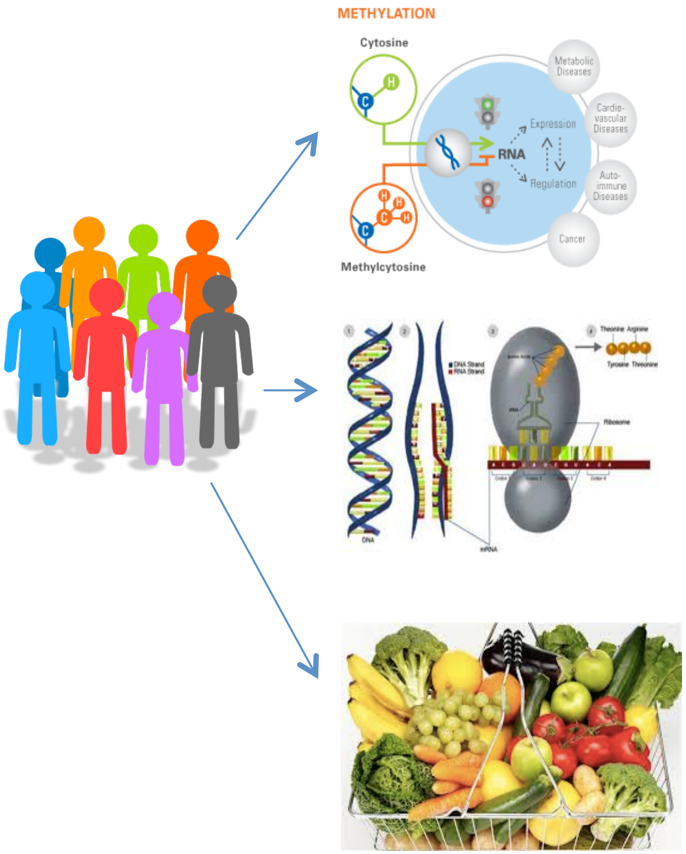
Problem setup

How to combine?



Problem setup

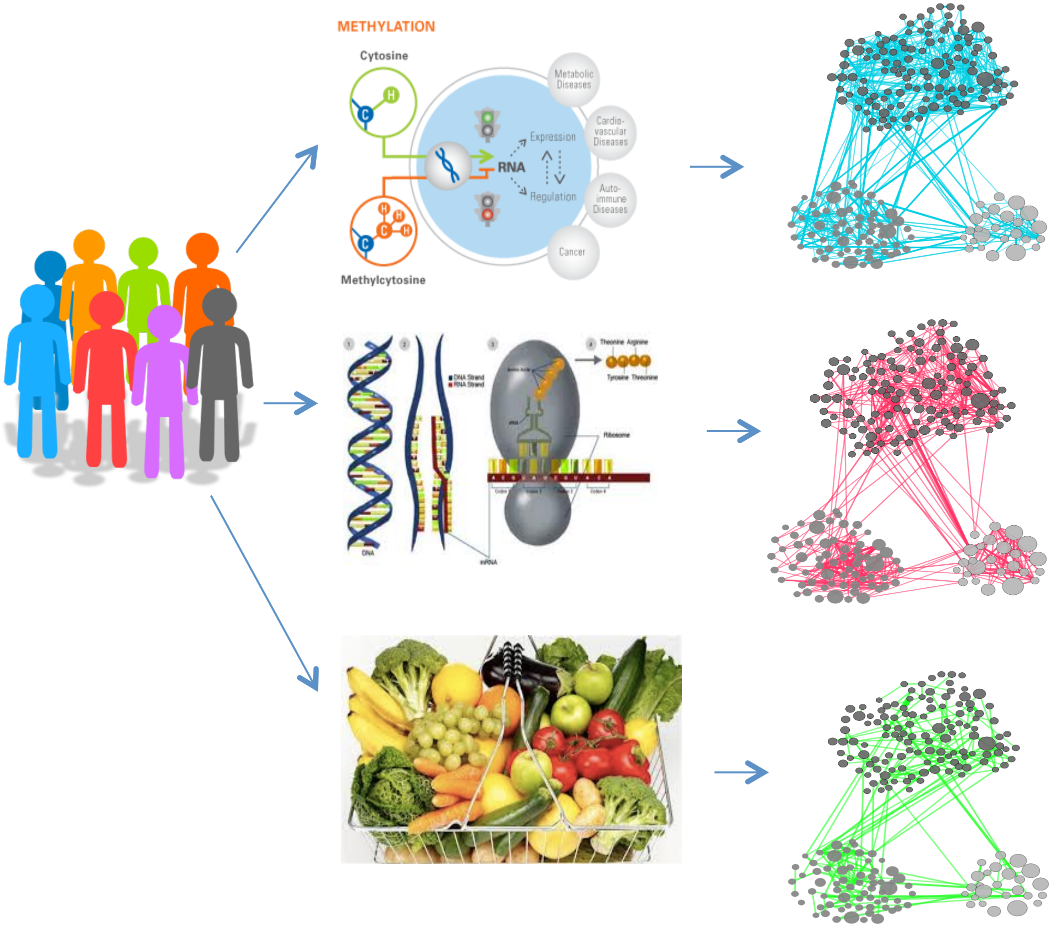
How to combine?



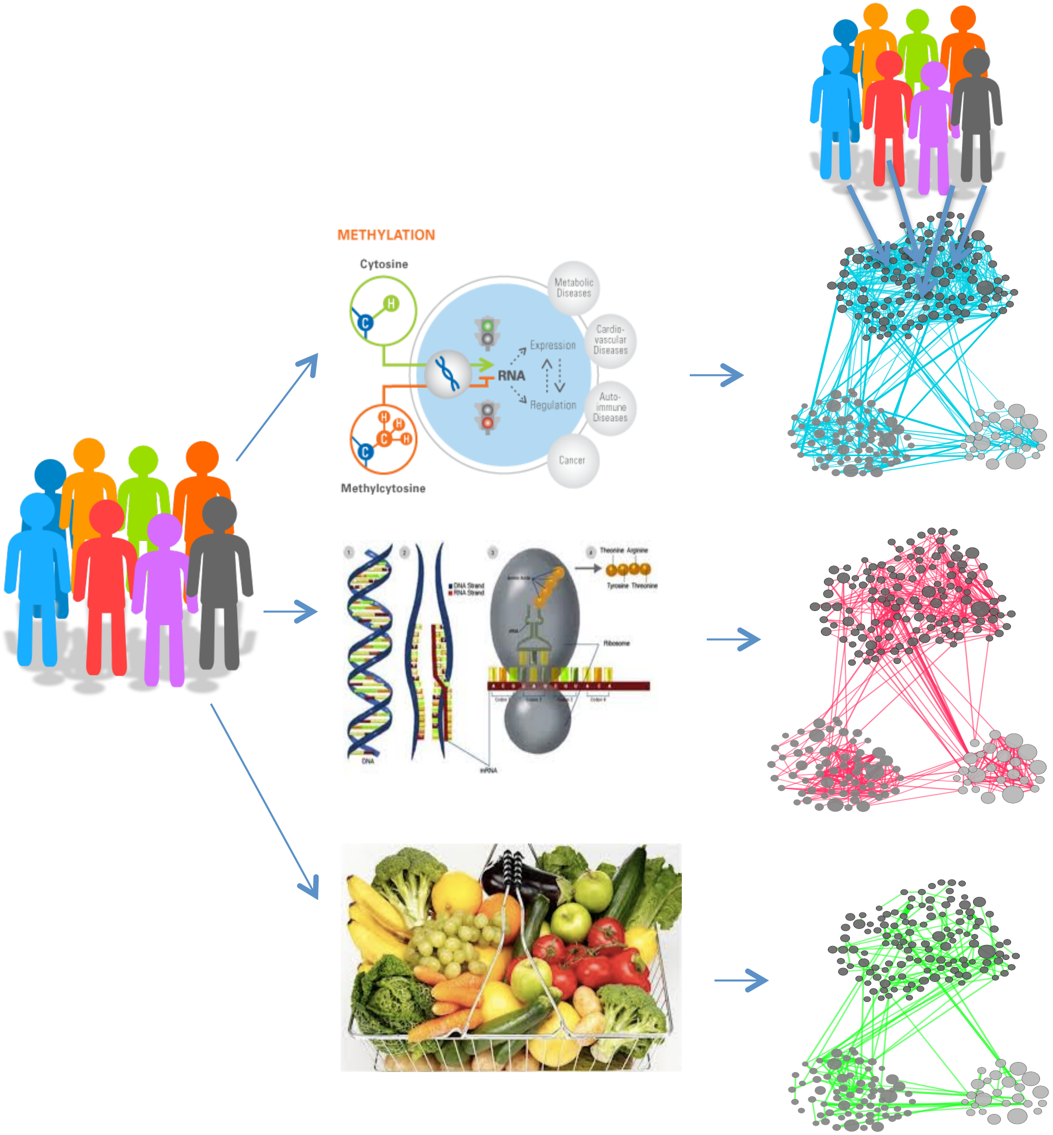
Issues

- Large number of measurements, small sample sizes (p>>n)
- Need to integrate common and complementary information
- Not all measurements can be mapped to the same unit (gene)

Goldenberg Lab: Similarity Network Fusion

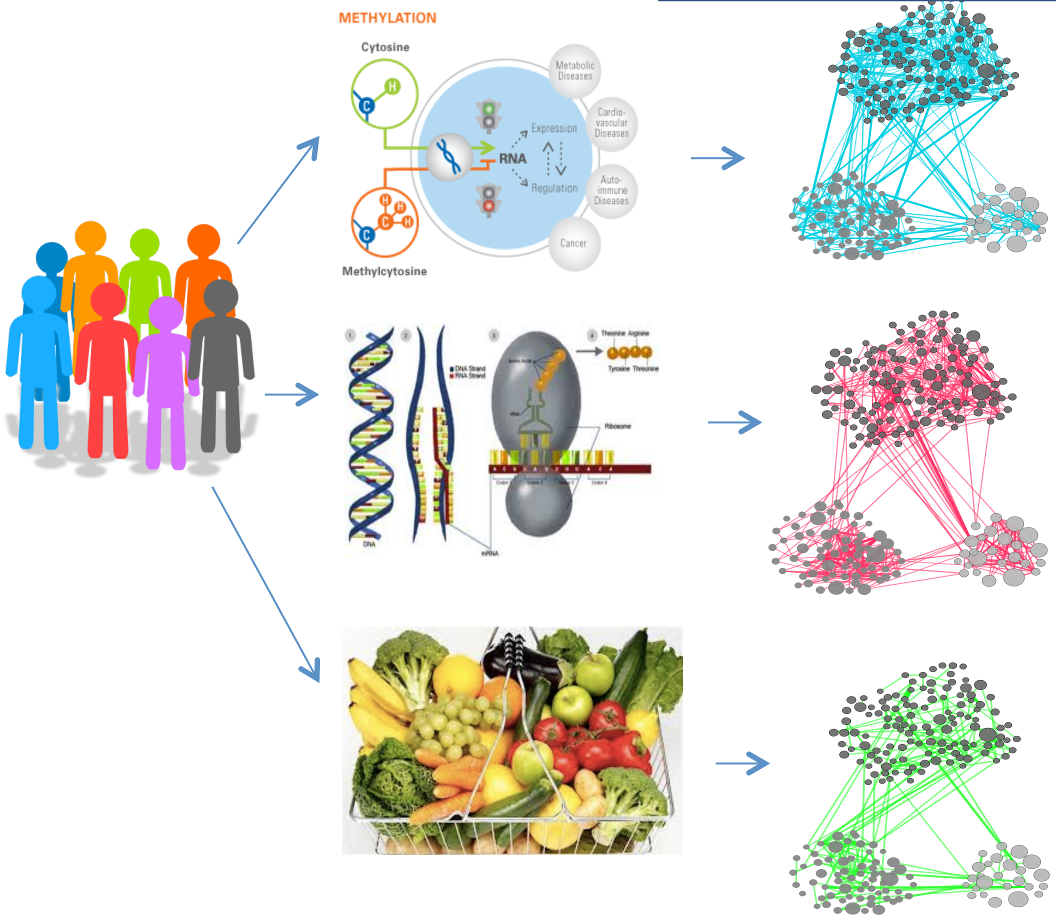


Goldenberg Lab: Similarity Network Fusion



Goldenberg Lab: Similarity Network Fusion

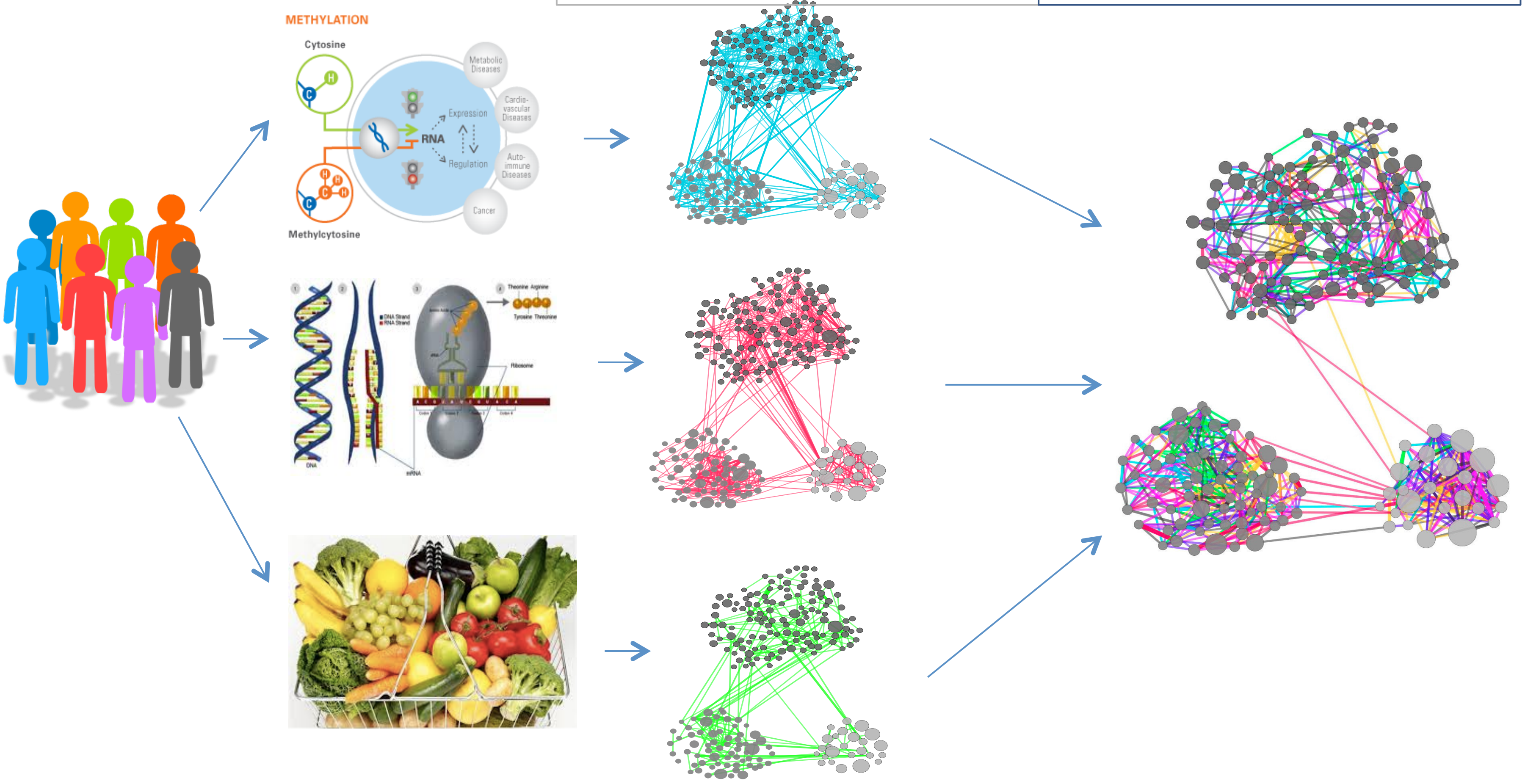
Step 1. Construct a similarity network for each data source



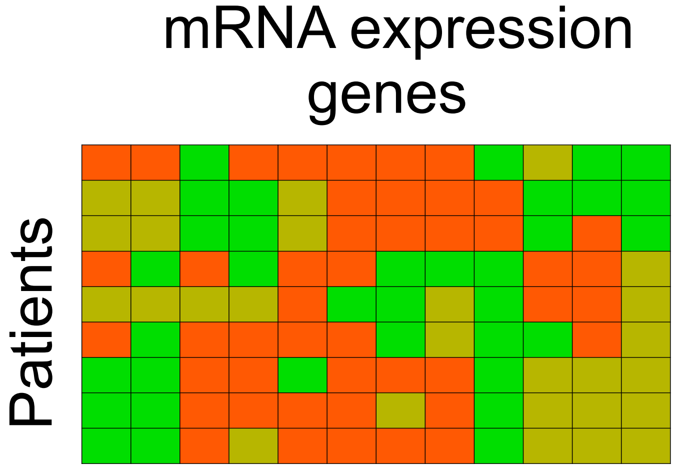
Goldenberg Lab: Similarity Network Fusion

Step 1. Construct a similarity network for each data source

Step 2. Integrate networks using non-linear fusion method



1. Construct similarity networks



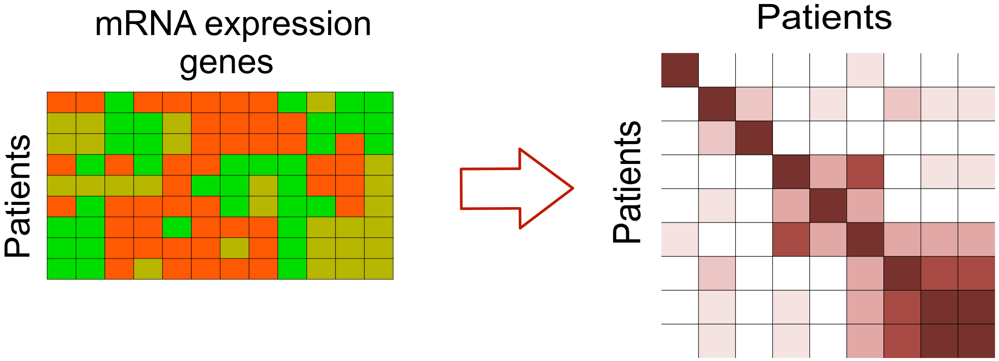
1. Construct similarity networks

Patient similarity:

$$W(i,j) = exp(\frac{\rho(x_i,x_j)^2}{\eta\xi_{ij}^2})$$

Adjacency matrix:

$$P(i,j) = \frac{W(i,j)}{\sum_{k\in V} W(i,k)}$$

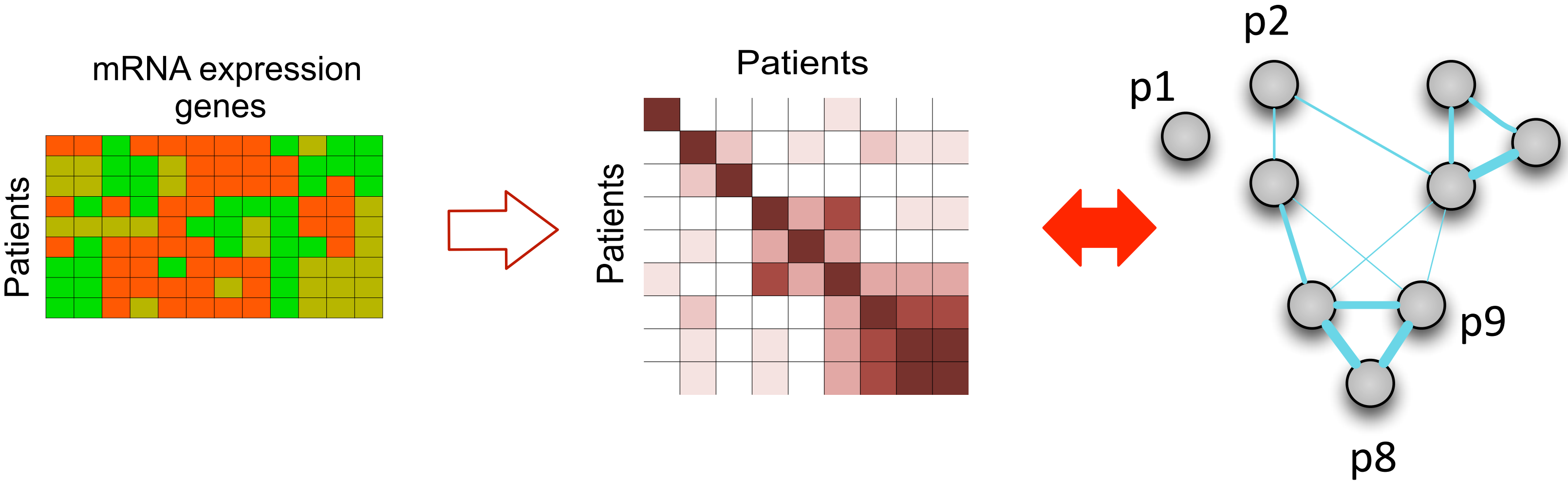


1. Construct similarity networks

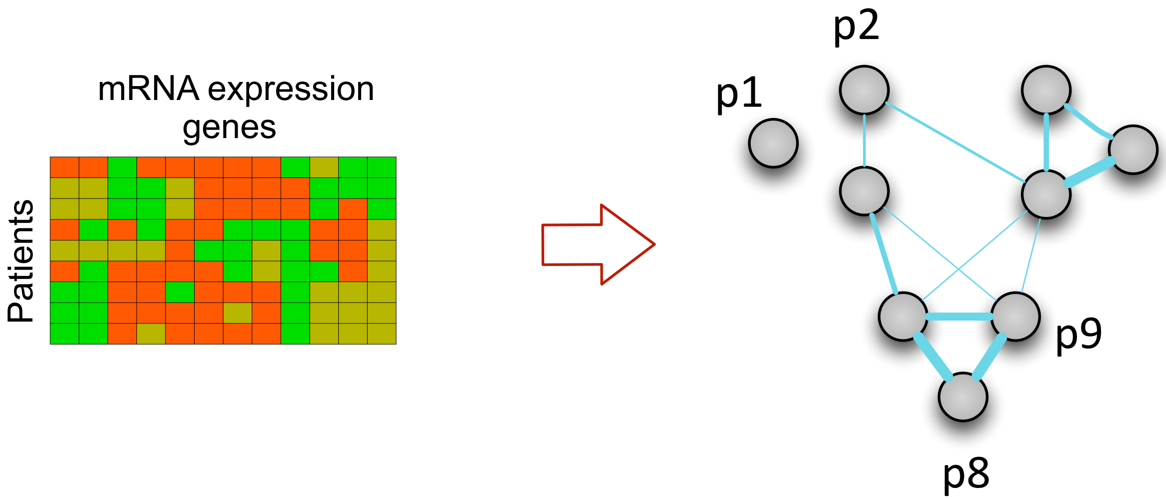
Sparsification

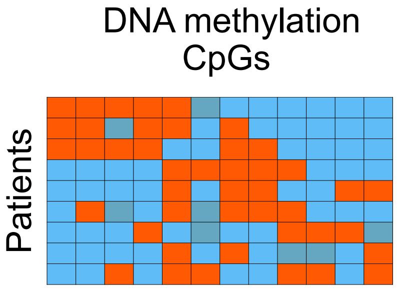
1) $$\mathcal{W}(i,j) = \begin{cases} W(i,j) \text{ if } x_j \in KNN(x_i) \\ 0 \text{ otherwise} \end{cases}$$

2) $$\mathcal{P}(i,j) = \frac{\mathcal{W}(i,j)}{\sum_{x_k \in KNN(x_i)} \mathcal{W}(i,k)}$$

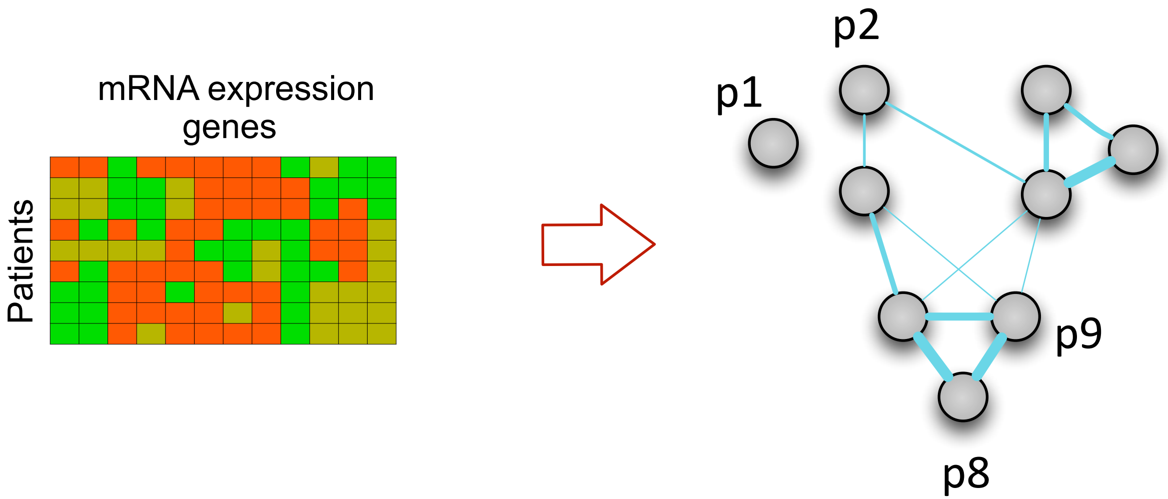


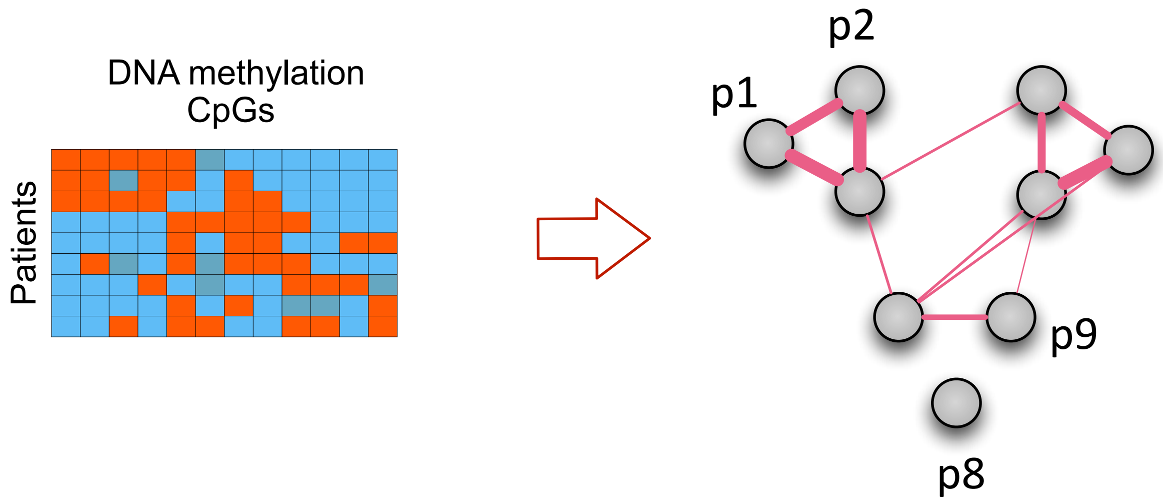
1. Construct similarity networks





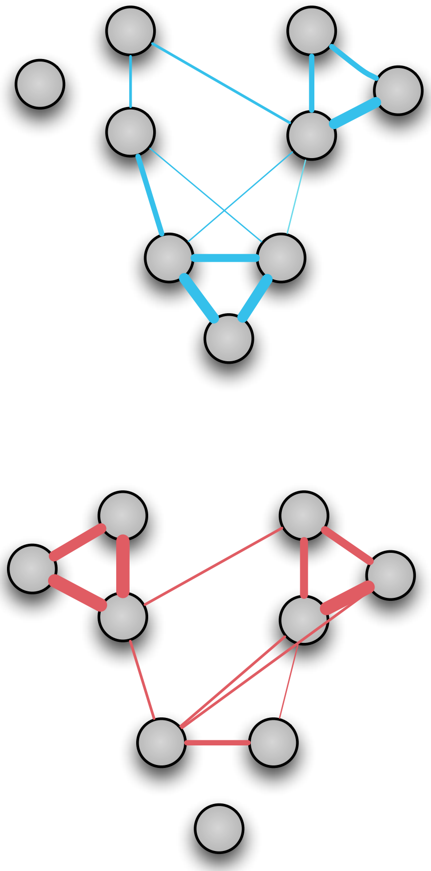
1. Construct similarity networks





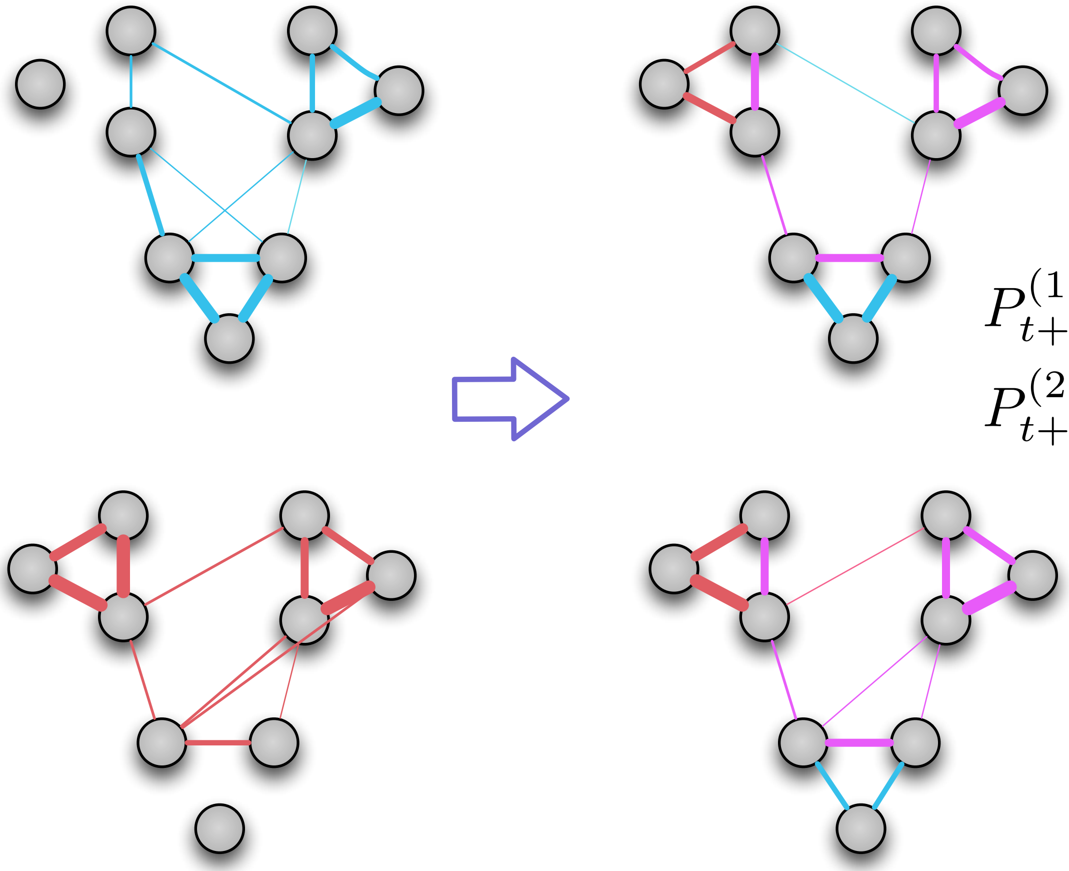
2. Combine networks

Sample Similarity Networks



● Patient   Patient similarity: ▬ mRNA-based ▬ DNA Methylation-based ▬ Supported by all data

2. Combine networks



Sample Similarity Networks

Fusion

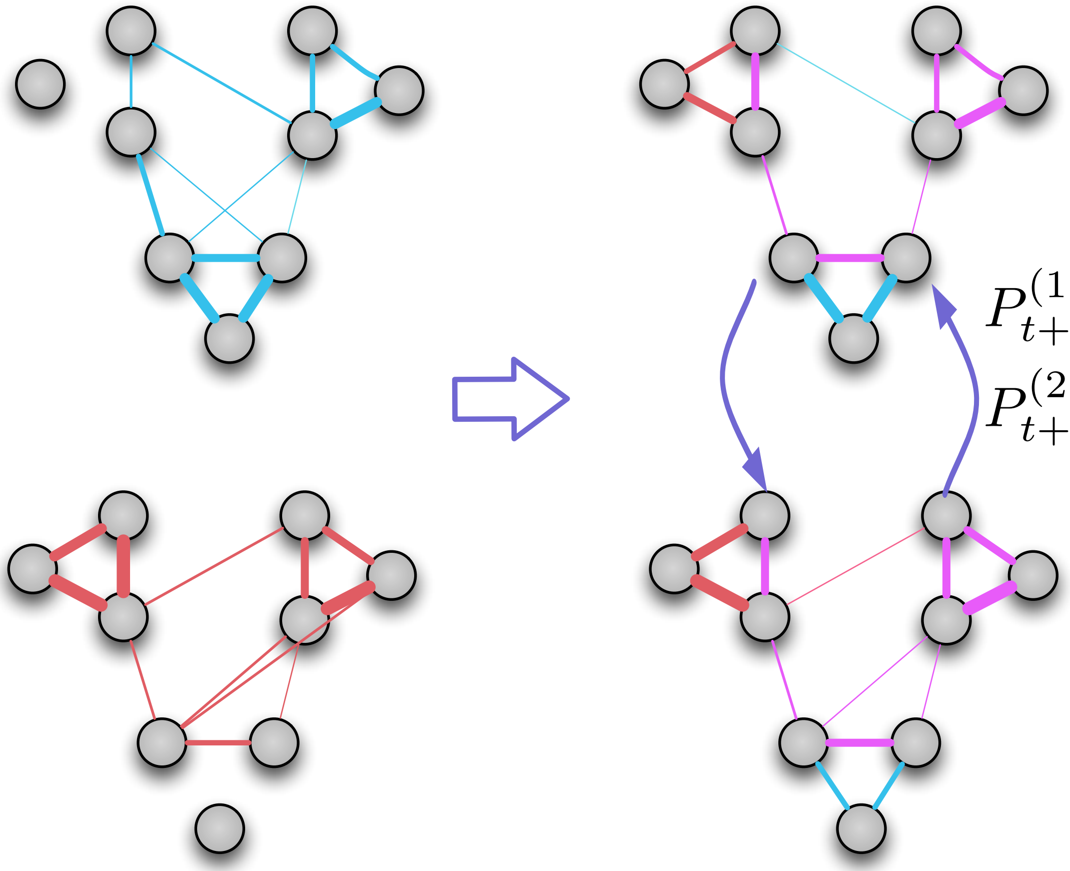
$$P_{t+1}^{(1)} = \mathcal{P}^{(1)} \times (P_t^{(2)}) \times (\mathcal{P}^{(1)})'$$
$$P_{t+1}^{(2)} = \mathcal{P}^{(2)} \times (P_t^{(1)}) \times (\mathcal{P}^{(2)})'$$

● Patient Patient similarity: ▬ mRNA-based ▬ DNA Methylation-based ▬ Supported by all data

2. Combine networks

Sample Similarity Networks

Fusion Iterations



$$P^{(1)}_{t+1} = \mathcal{P}^{(1)} \times (P^{(2)}_t) \times (\mathcal{P}^{(1)})'$$

$$P^{(2)}_{t+1} = \mathcal{P}^{(2)} \times (P^{(1)}_t) \times (\mathcal{P}^{(2)})'$$

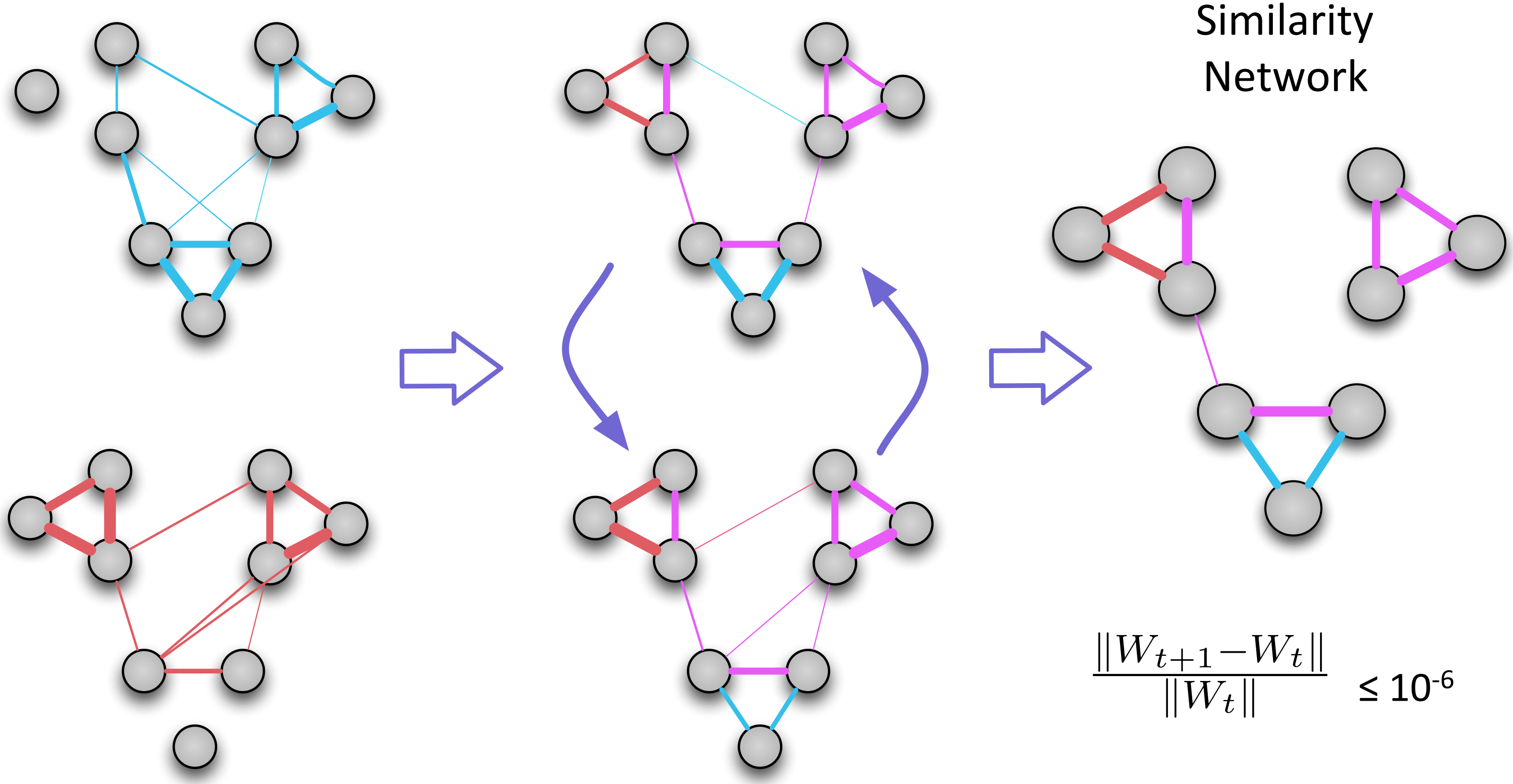
● Patient Patient similarity: 🟦 mRNA-based 🟥 DNA Methylation-based 🟪 Supported by all data

2. Combine networks

Sample Similarity Networks

Fusion Iterations

Fused Similarity Network



$$\frac{||W_{t+1} - W_t||}{||W_t||} \leq 10^{-6}$$

● Patient   Patient similarity: 🟦 mRNA-based   🟥 DNA Methylation-based   🟪 Supported by all data

Network Fusion

Fusing 2 networks:

$$P_{t+1}^{(1)} = \mathcal{P}^{(1)} \times (P_t^{(2)}) \times (\mathcal{P}^{(1)})'$$
$$P_{t+1}^{(2)} = \mathcal{P}^{(2)} \times (P_t^{(1)}) \times (\mathcal{P}^{(2)})'$$

Fusing m networks:

$$P_{t+1}^{(i)} \quad = \quad \mathcal{P}^{(i)} \times (\frac{1}{m-1} \sum_{j \neq i} P_t^{(j)}) \times (\mathcal{P}^{(i)})' + \eta I$$

Experiments

Data:

5 TCGA cancers

METABRIC (Large Breast Cancer db)

Comparative Methods:

Concatenation

iCluster

PDSB

Multiple kernel learning

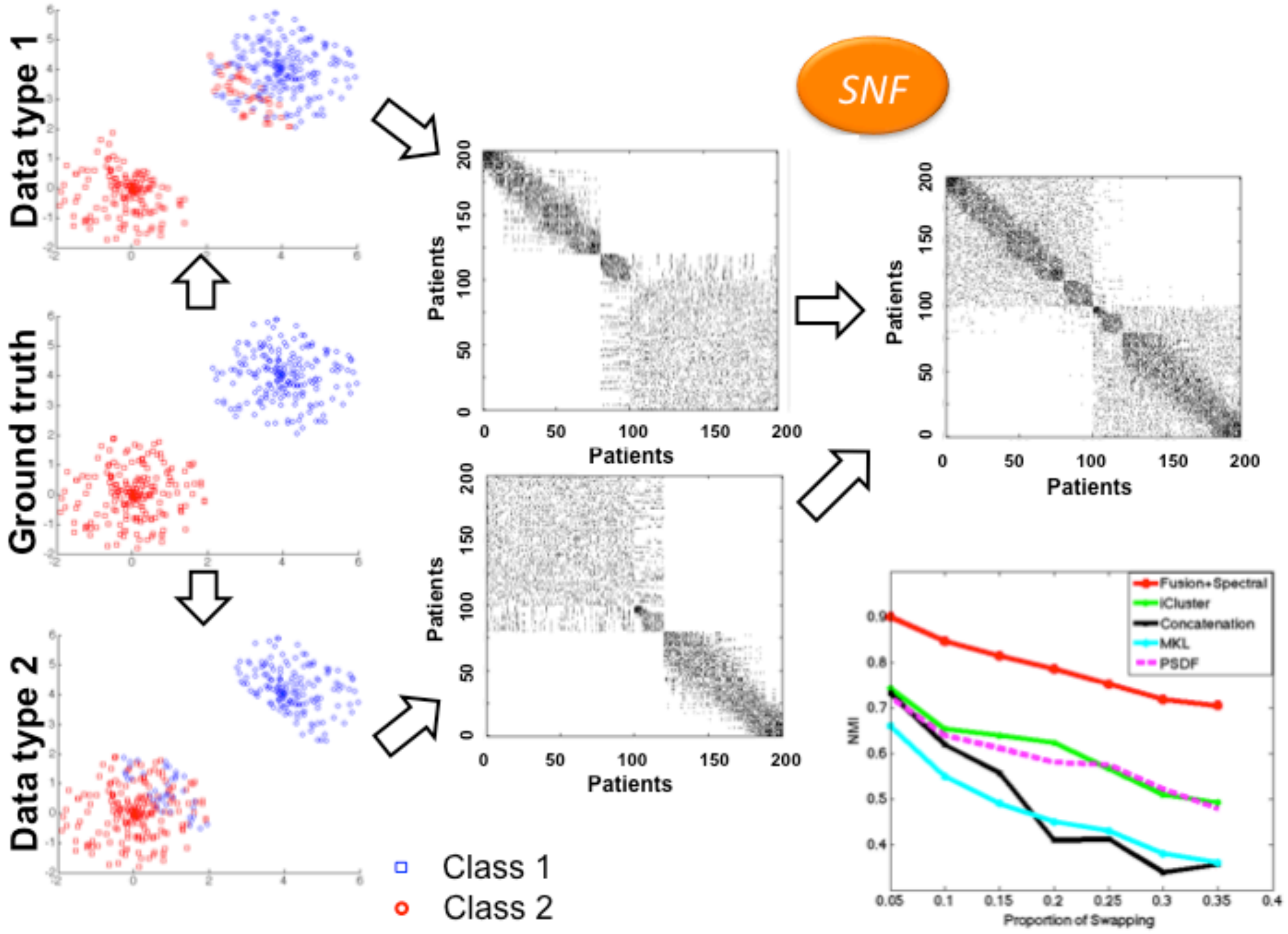
Criteria:

-$\log_{10}$(log rank pvalue)

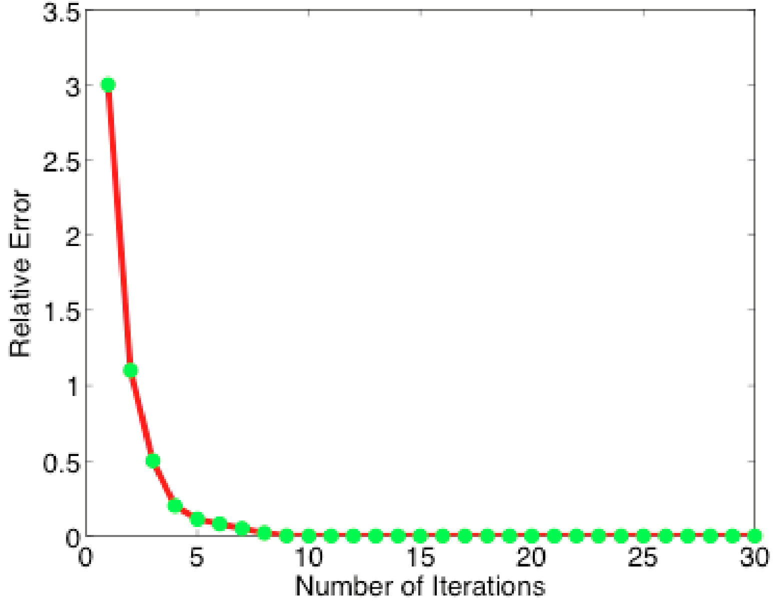
Silhouette score (cluster homogeneity)

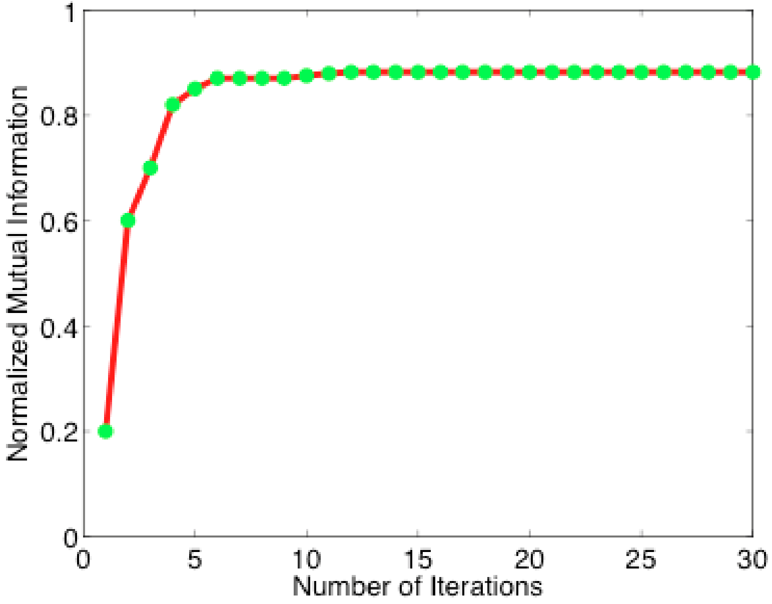
Running time

Simulation 1 – complementarity



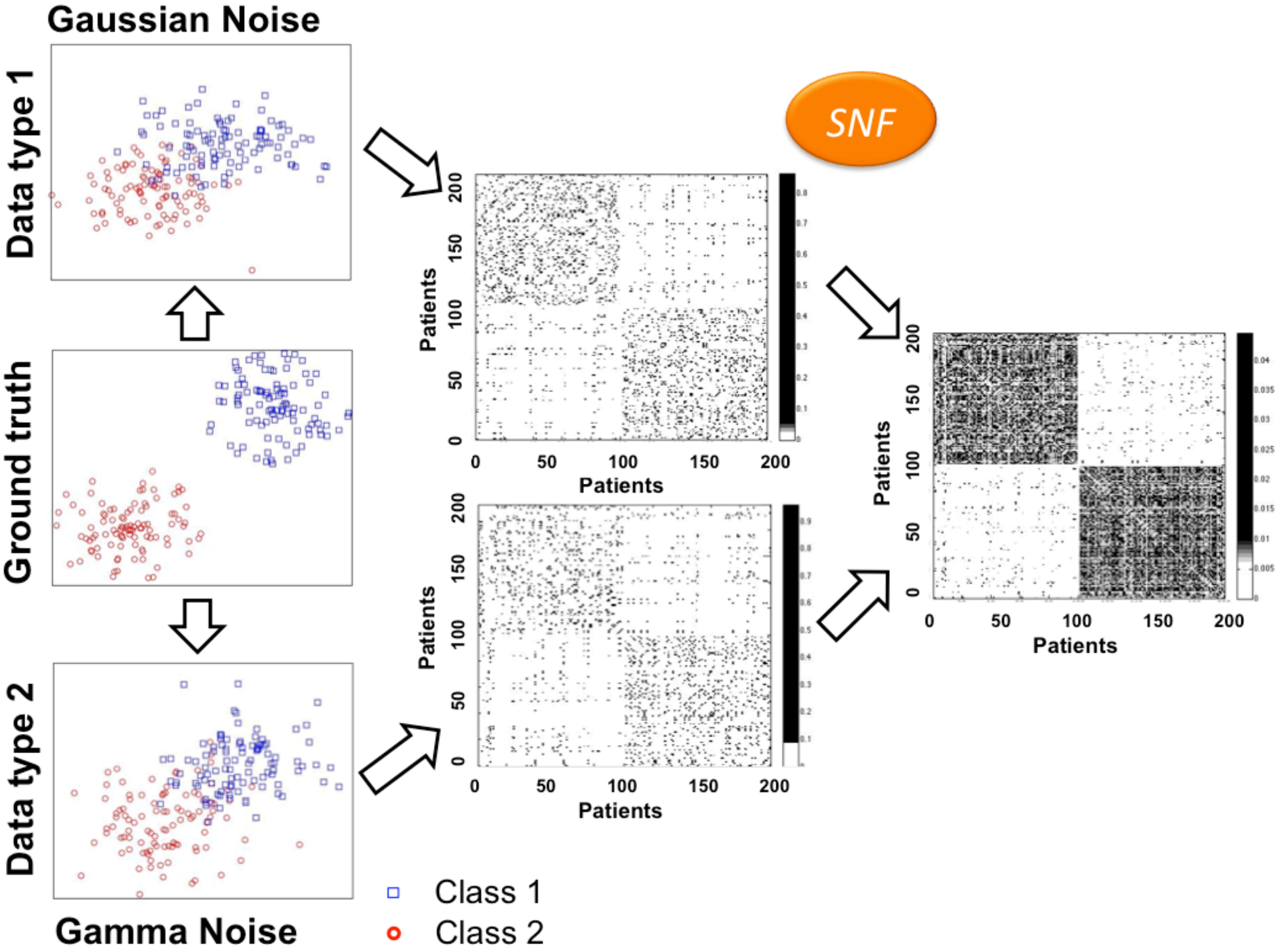
Simulation 1 convergence



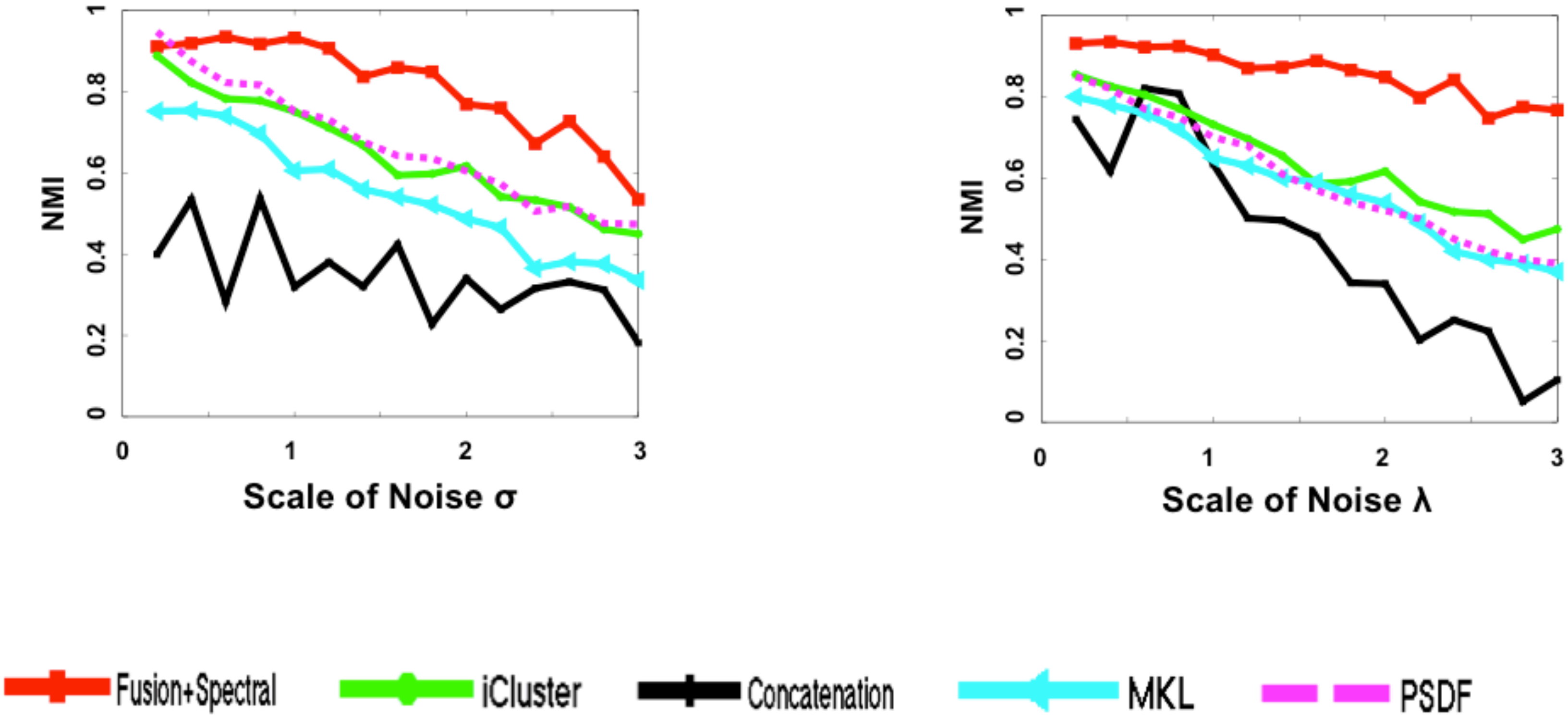


20% of patients are mislabeled

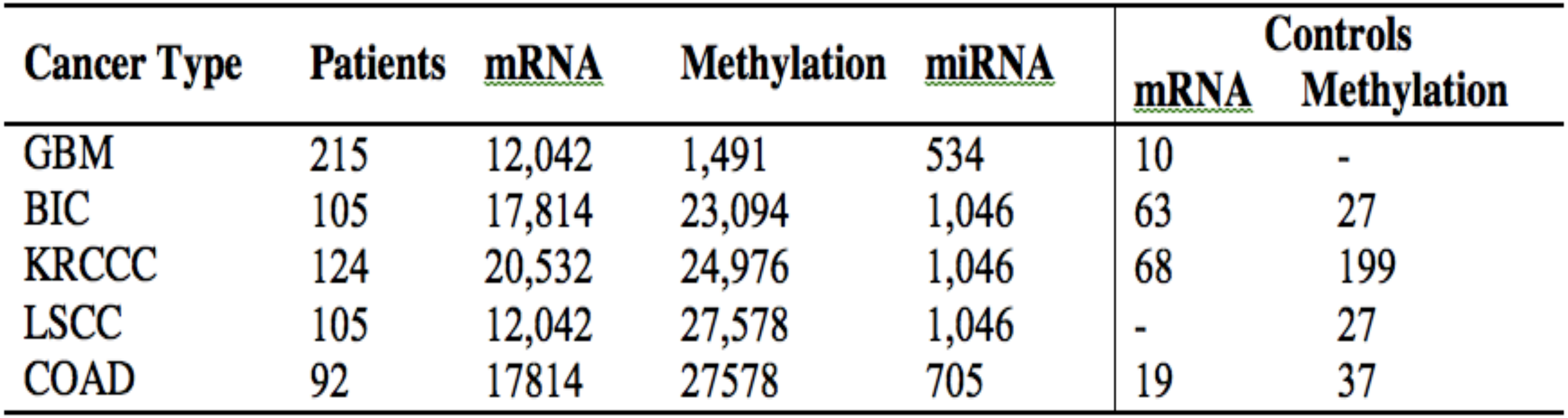
Simulation 2 - removing noise



Simulation 2 - removing noise



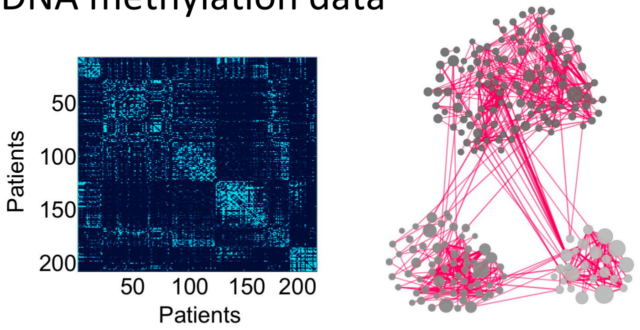
TCGA Data



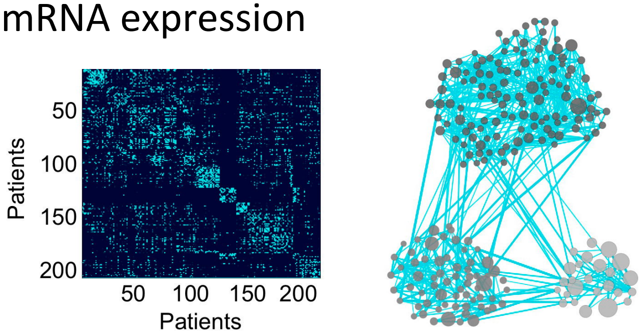
Cancer Type	Patients	mRNA	Methylation	miRNA	Controls	
					mRNA	Methylation
GBM	215	12,042	1,491	534	10	-
BIC	105	17,814	23,094	1,046	63	27
KRCCC	124	20,532	24,976	1,046	68	199
LSCC	105	12,042	27,578	1,046	-	27
COAD	92	17814	27578	705	19	37

Case study: Glioblastoma

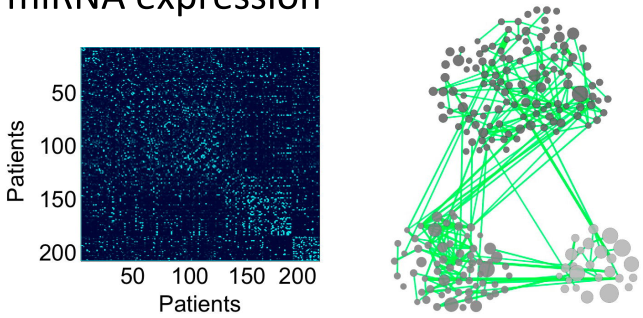
DNA methylation data



mRNA expression



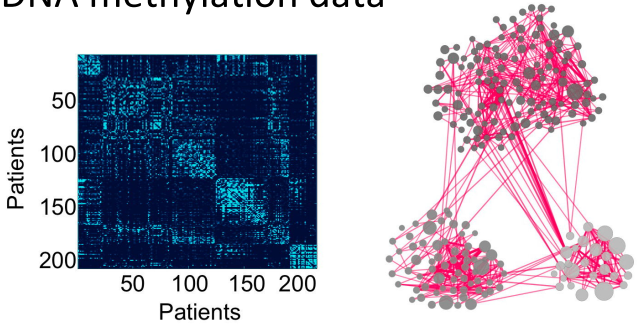
miRNA expression



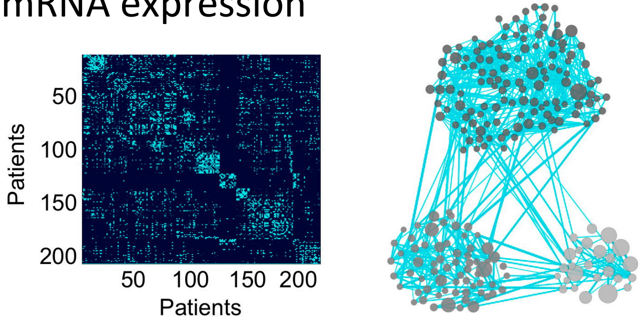
Bo Wang

Case study: Glioblastoma

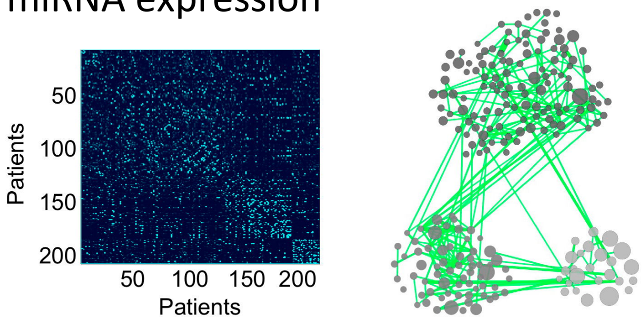
DNA methylation data



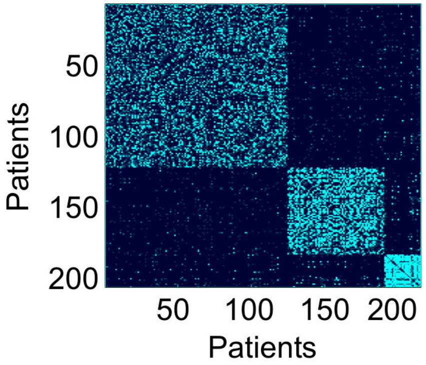
mRNA expression

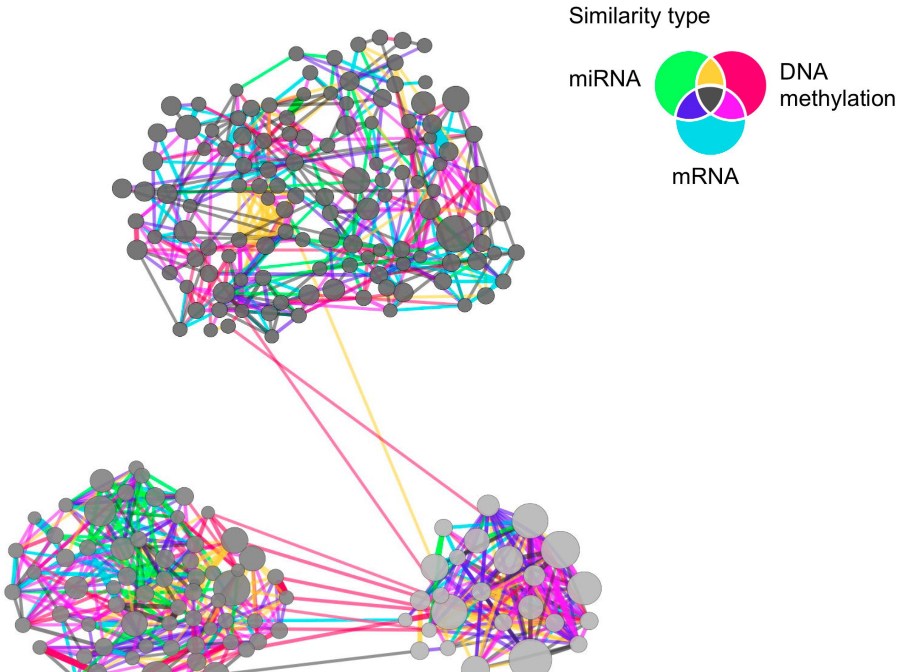


miRNA expression



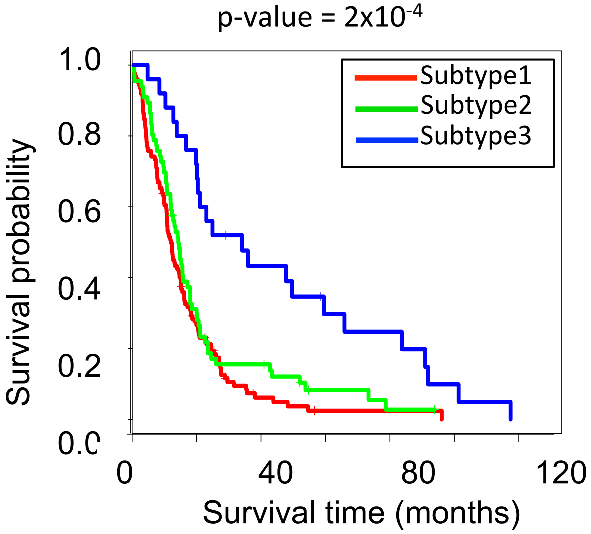
FUSED

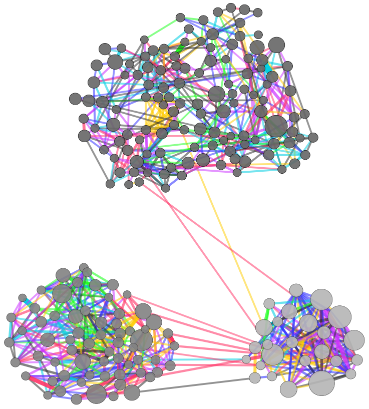




Bo Wang

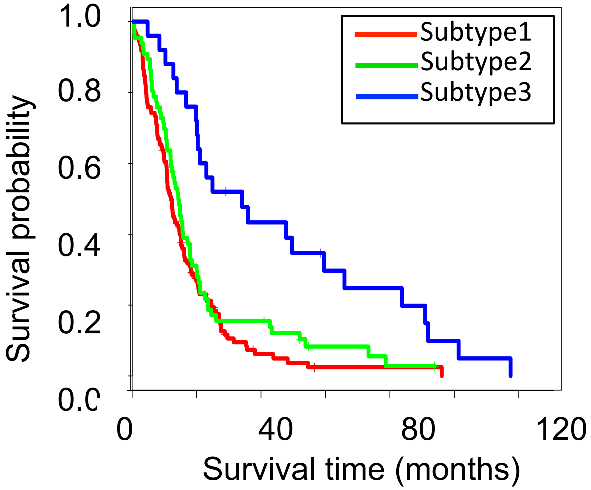
Clinical properties of the subtypes

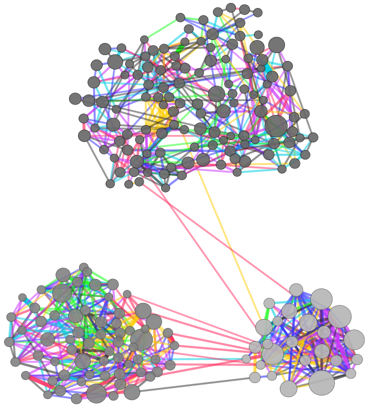


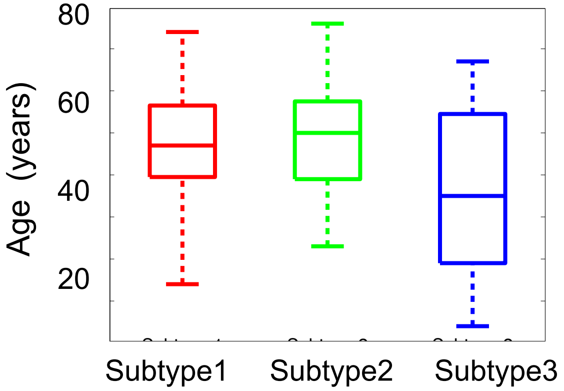


Clinical properties of the subtypes

p-value = $2x10^{-4}$



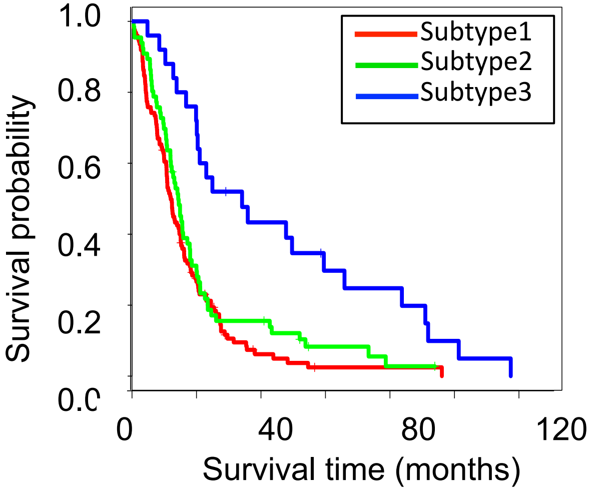


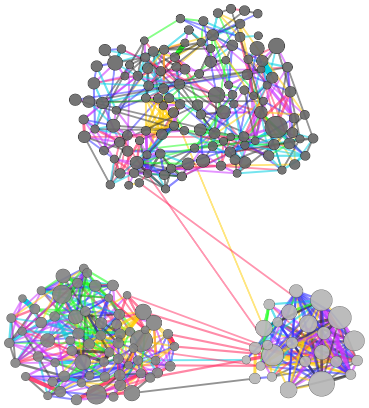


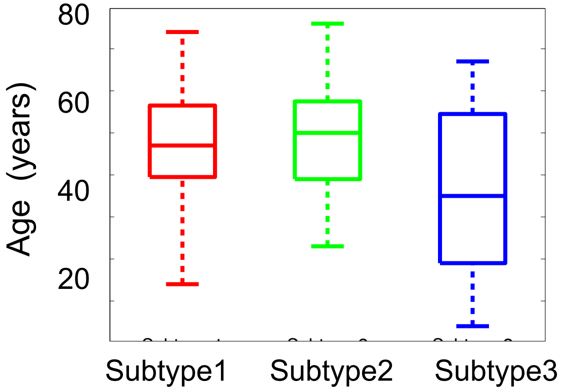
p-value = $3x10^{-5}$

Clinical properties of the subtypes

p-value = $2x10^{-4}$

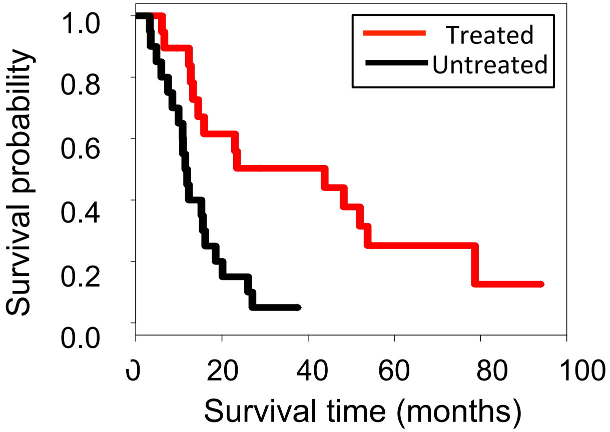






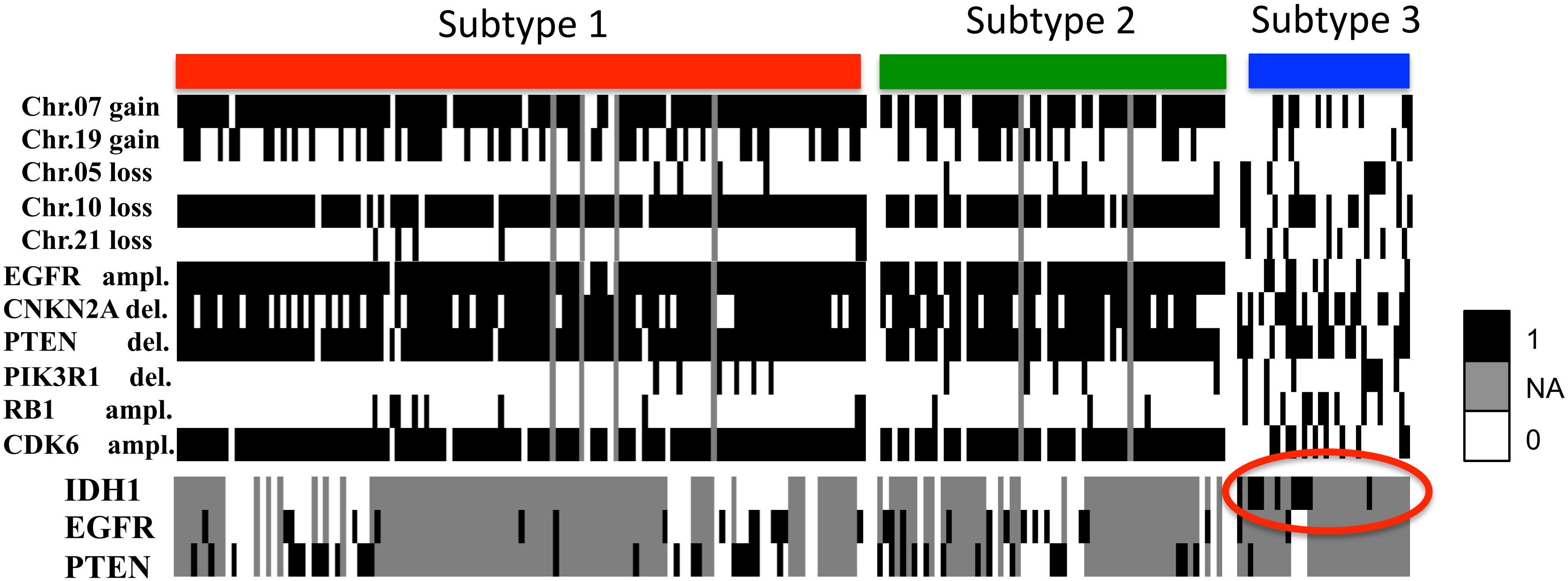
p-value = $3x10^{-5}$

Subtype 1



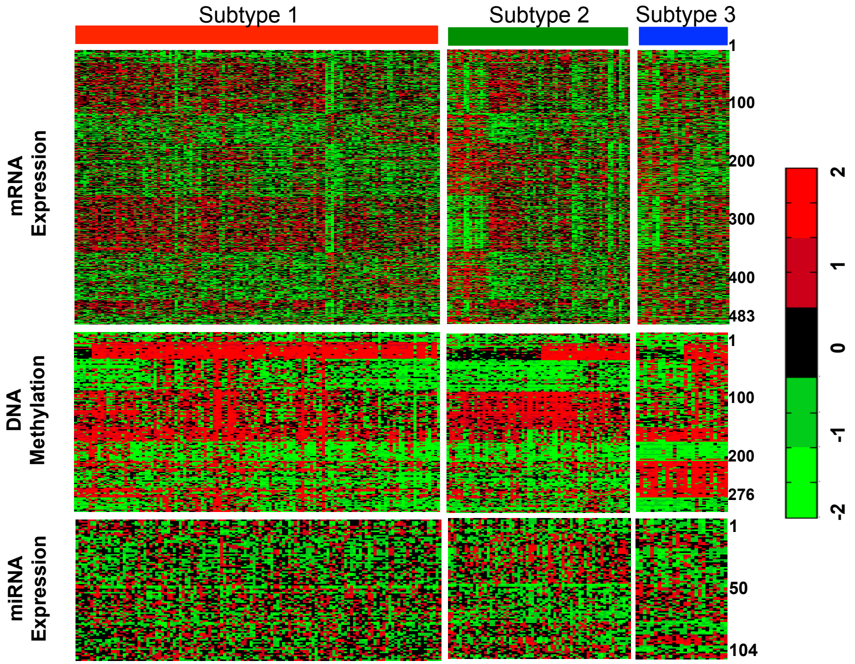
p = 0.001

Biological characterization of the subtypes



Feature Selection

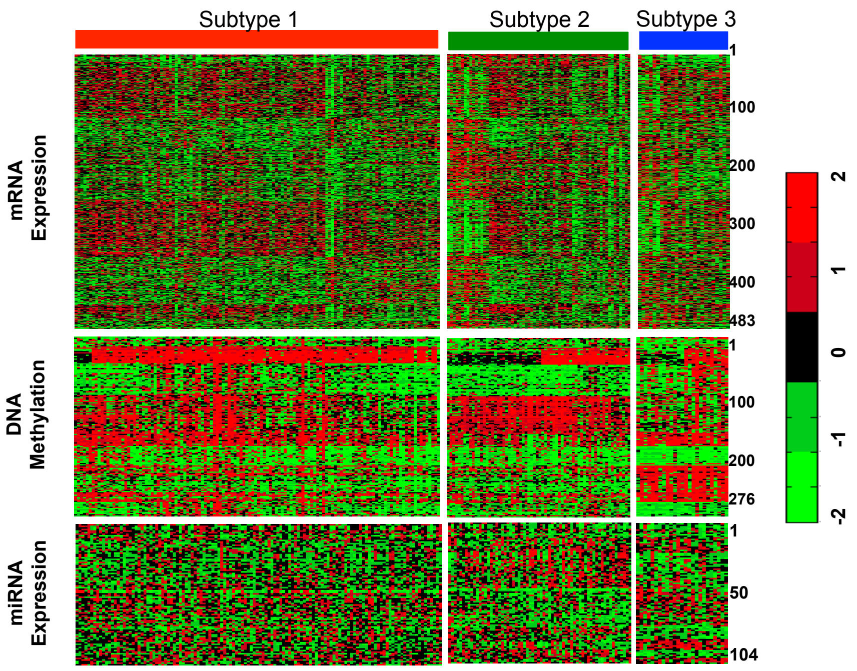
Standard t-test
Differential analysis



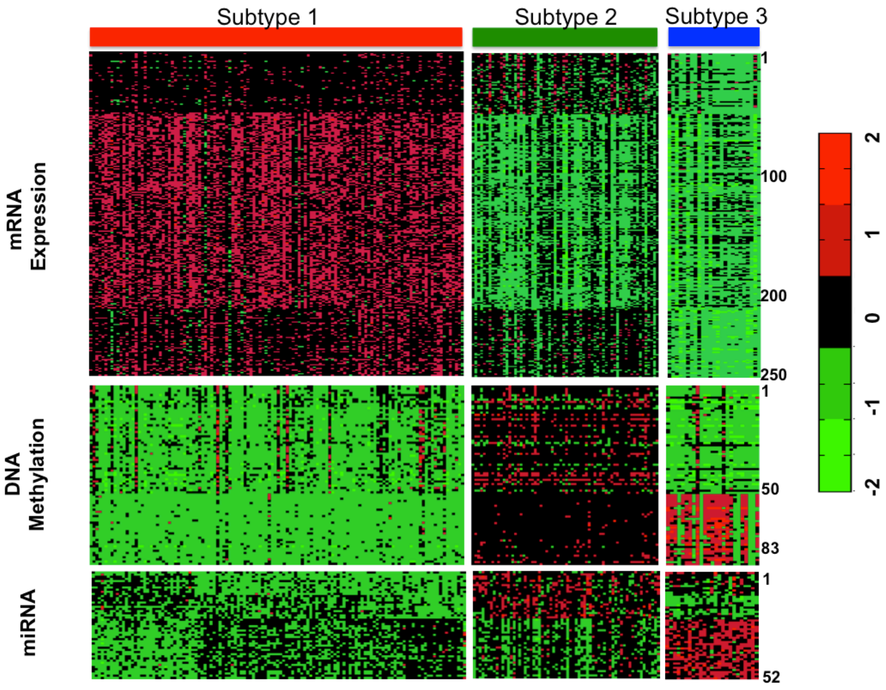
Bo Wang

Feature Selection

Standard t-test Differential analysis

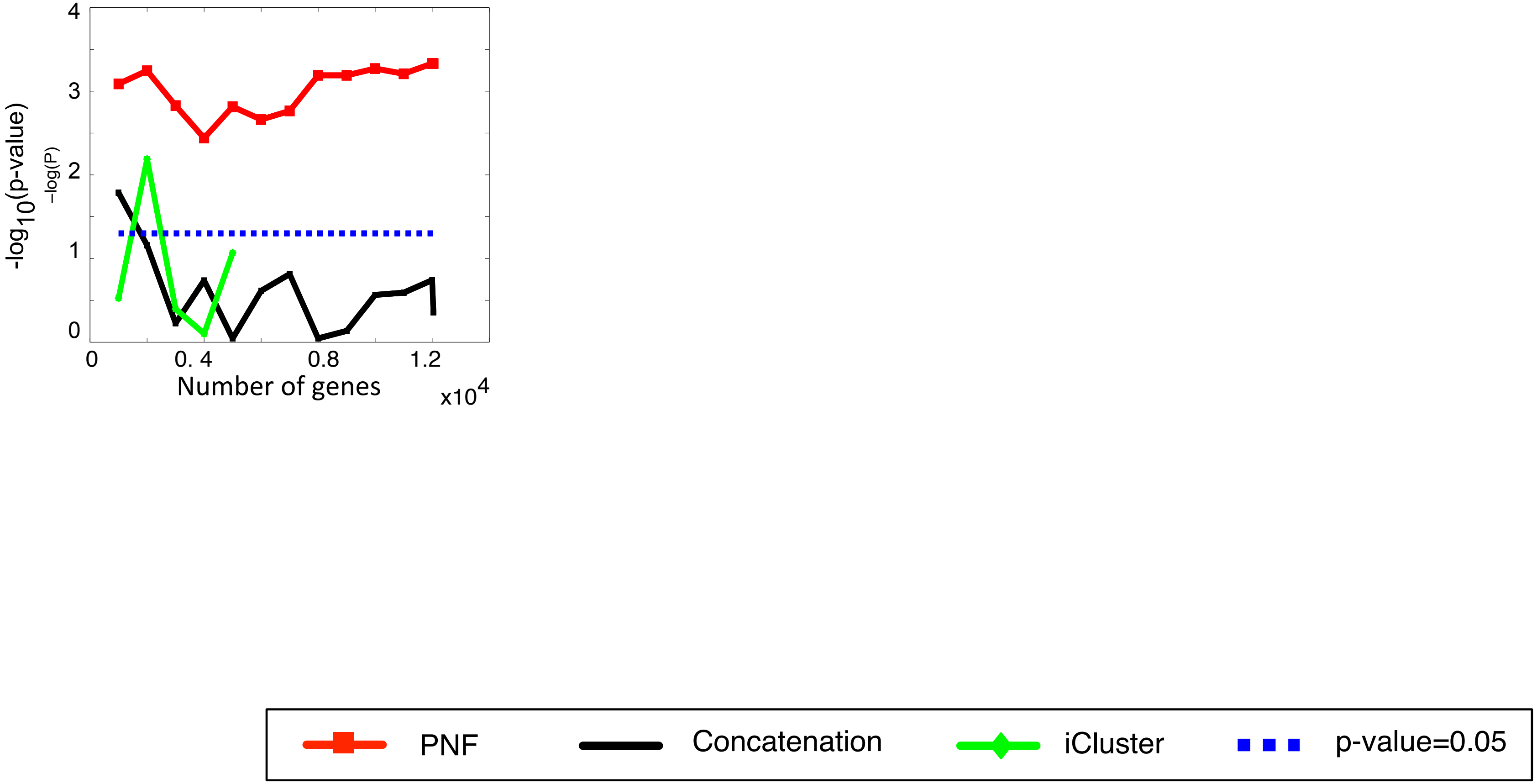


Network-based NMI Differential analysis



Bo Wang

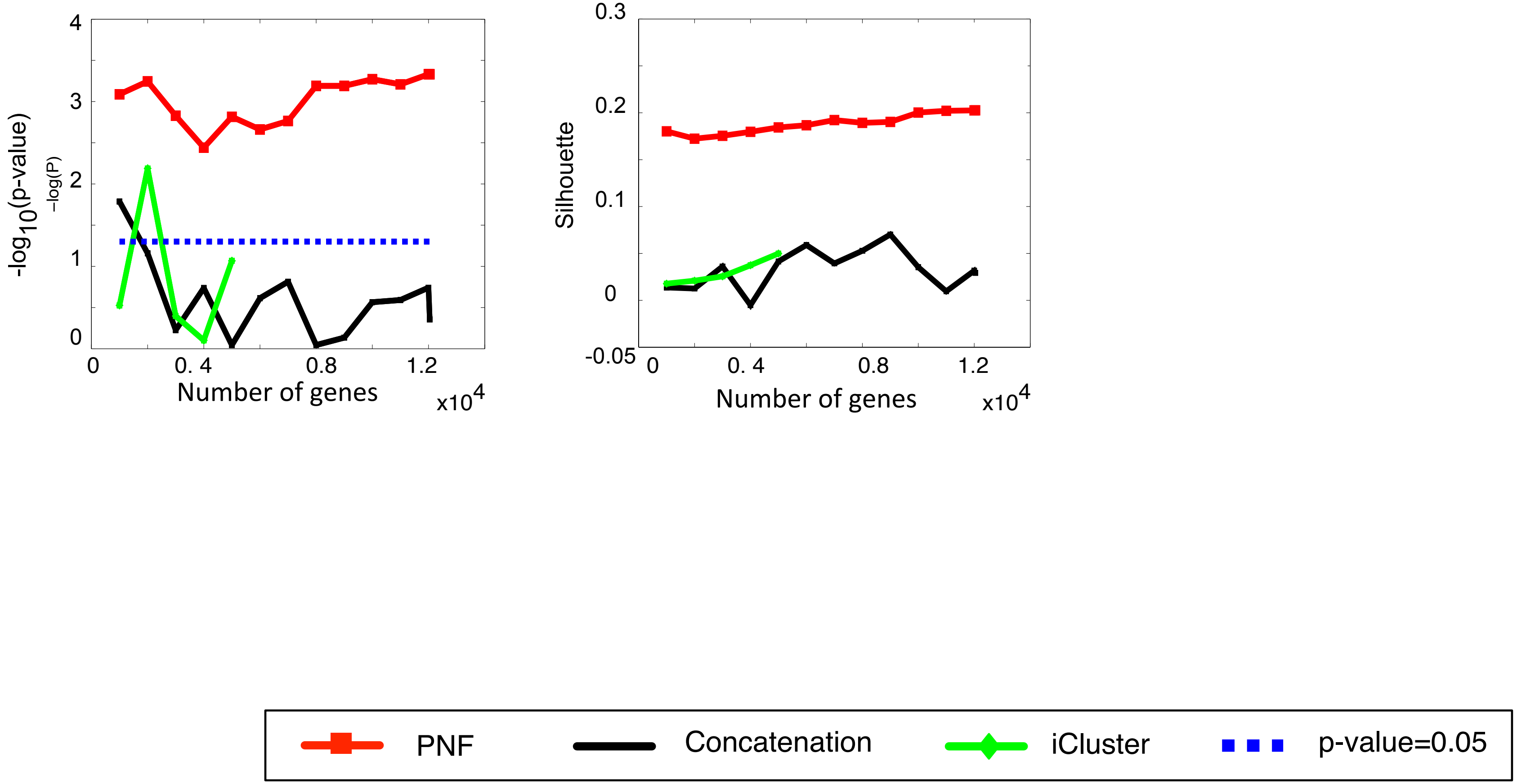
Gene Pre-selection in GBM



Genes are ordered by significance of the differential values between tumor and normal

Bo Wang

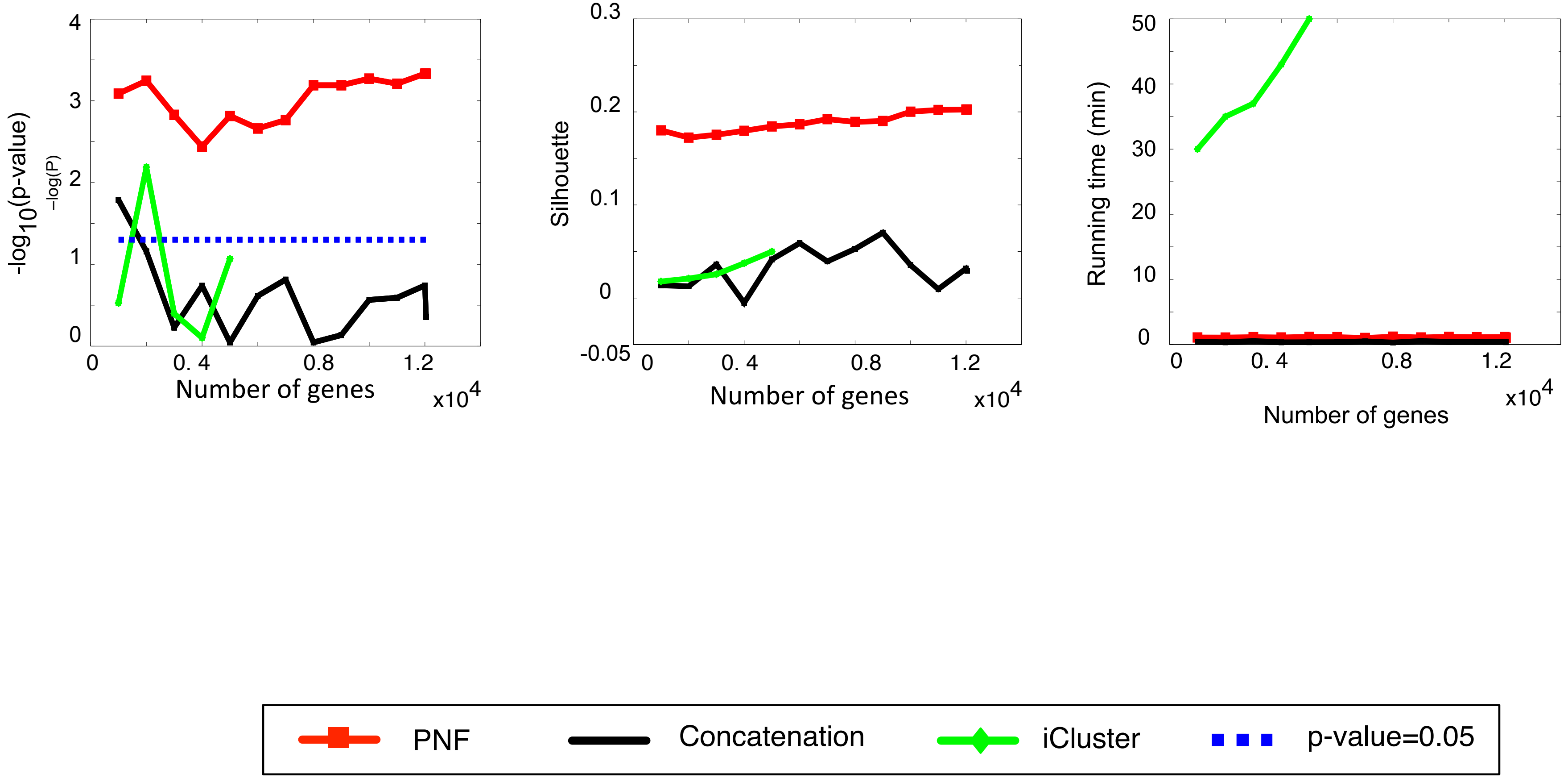
Gene Pre-selection in GBM



Genes are ordered by significance of the differential values between tumor and normal

Bo Wang

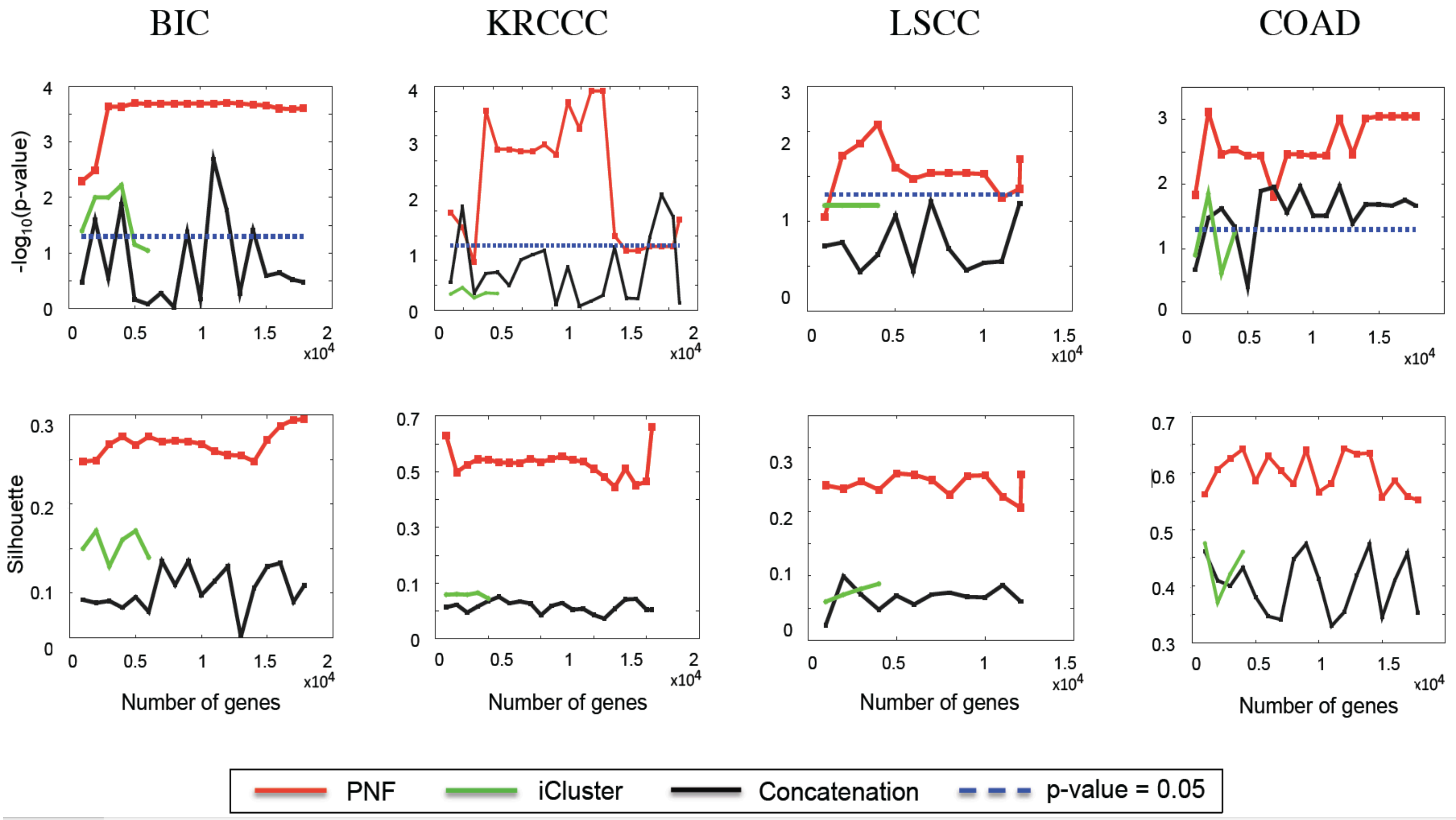
Gene Pre-selection in GBM



Genes are ordered by significance of the differential values between tumor and normal

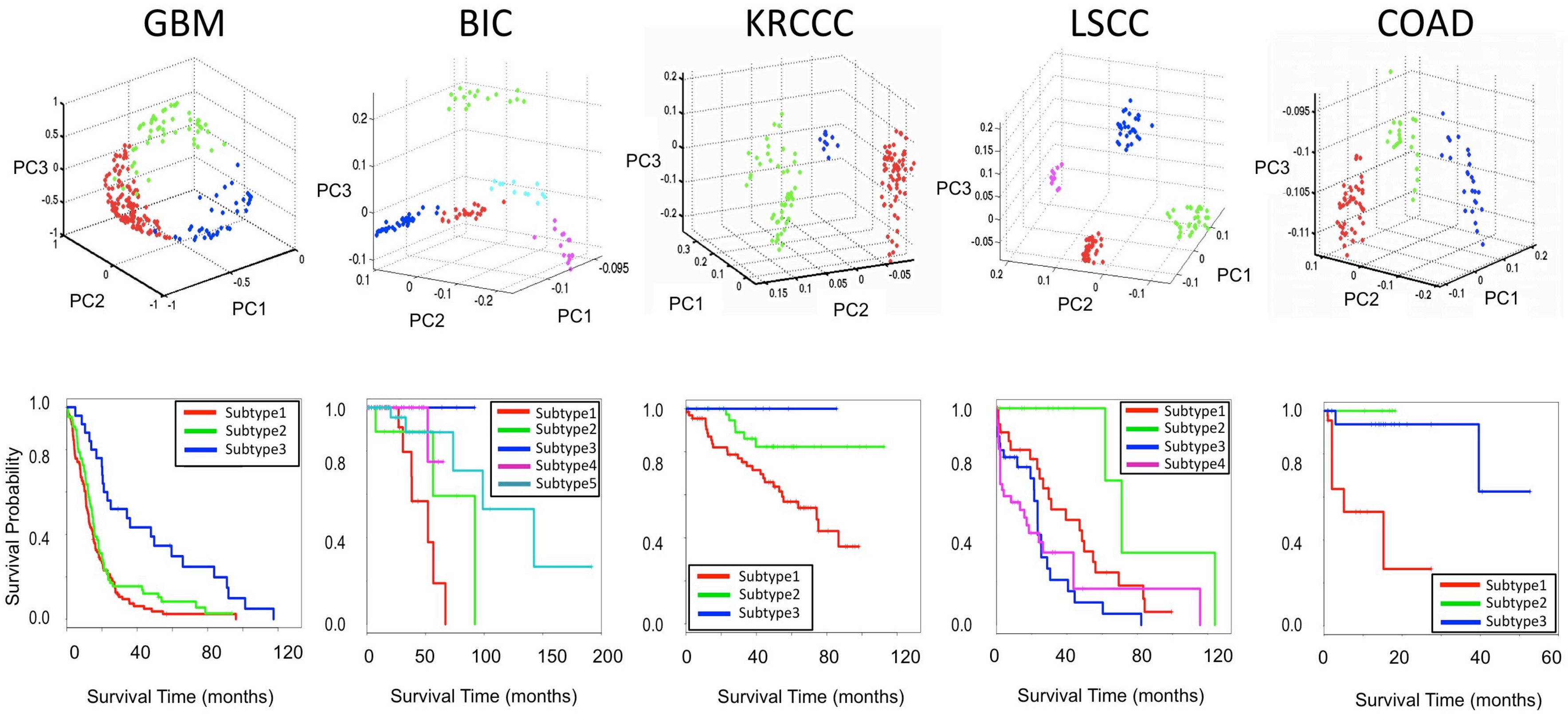
Bo Wang

Gene pre-selection across cancers



Bo Wang

Clustering of the network



Bo Wang

Patient networks framework advantages

✓ Creates a unified view of patients based on multiple heterogeneous sources

✓ Integrates gene and non-gene based data

✓ No need to do gene pre-selection

✓ Robust to different types of noise

✓ Scalable

Patient networks

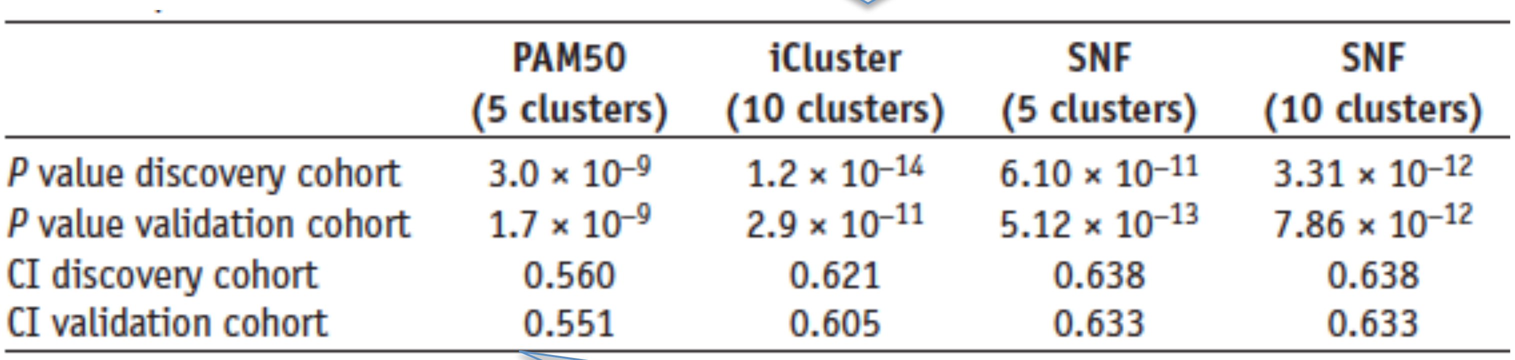
- ✓ Obtain superior results on regular tasks such as subtyping
- ✓ No fea re pre-se ion
- ✓ Imputatio
- ✓ Can e hout imputi

Transformative power of patient networks

Breast Cancer (METABRIC example)

CNV and expression data
Discovery: 997 patients
Validation: 995 patients

Nature, 2012



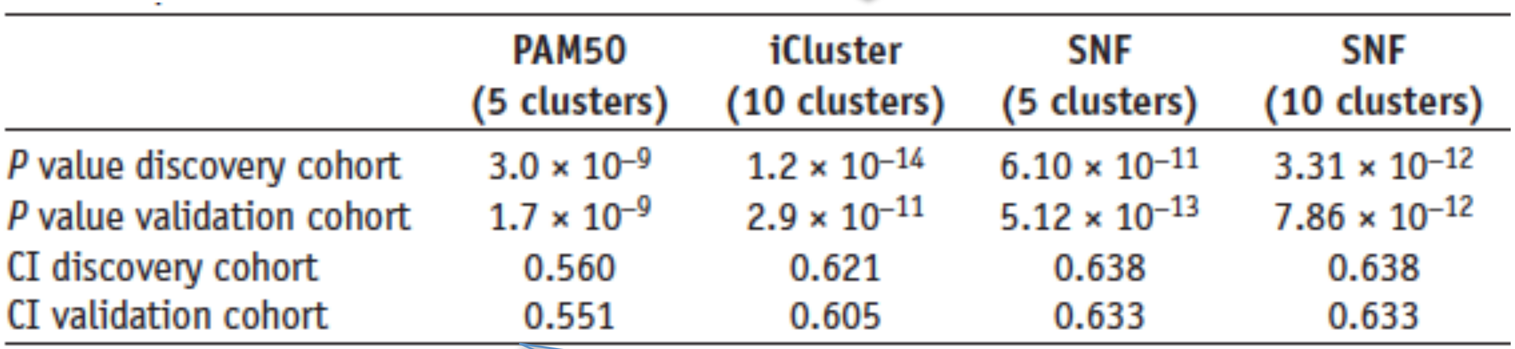
	PAM50 (5 clusters)	iCluster (10 clusters)	SNF (5 clusters)	SNF (10 clusters)
P value discovery cohort	3.0×10^{-9}	1.2×10^{-14}	6.10×10^{-11}	3.31×10^{-12}
P value validation cohort	1.7×10^{-9}	2.9×10^{-11}	5.12×10^{-13}	7.86×10^{-12}
CI discovery cohort	0.560	0.621	0.638	0.638
CI validation cohort	0.551	0.605	0.633	0.633

established

Breast Cancer (METABRIC example)

CNV and expression data
Discovery: 997 patients
Validation: 995 patients

Nature, 2012

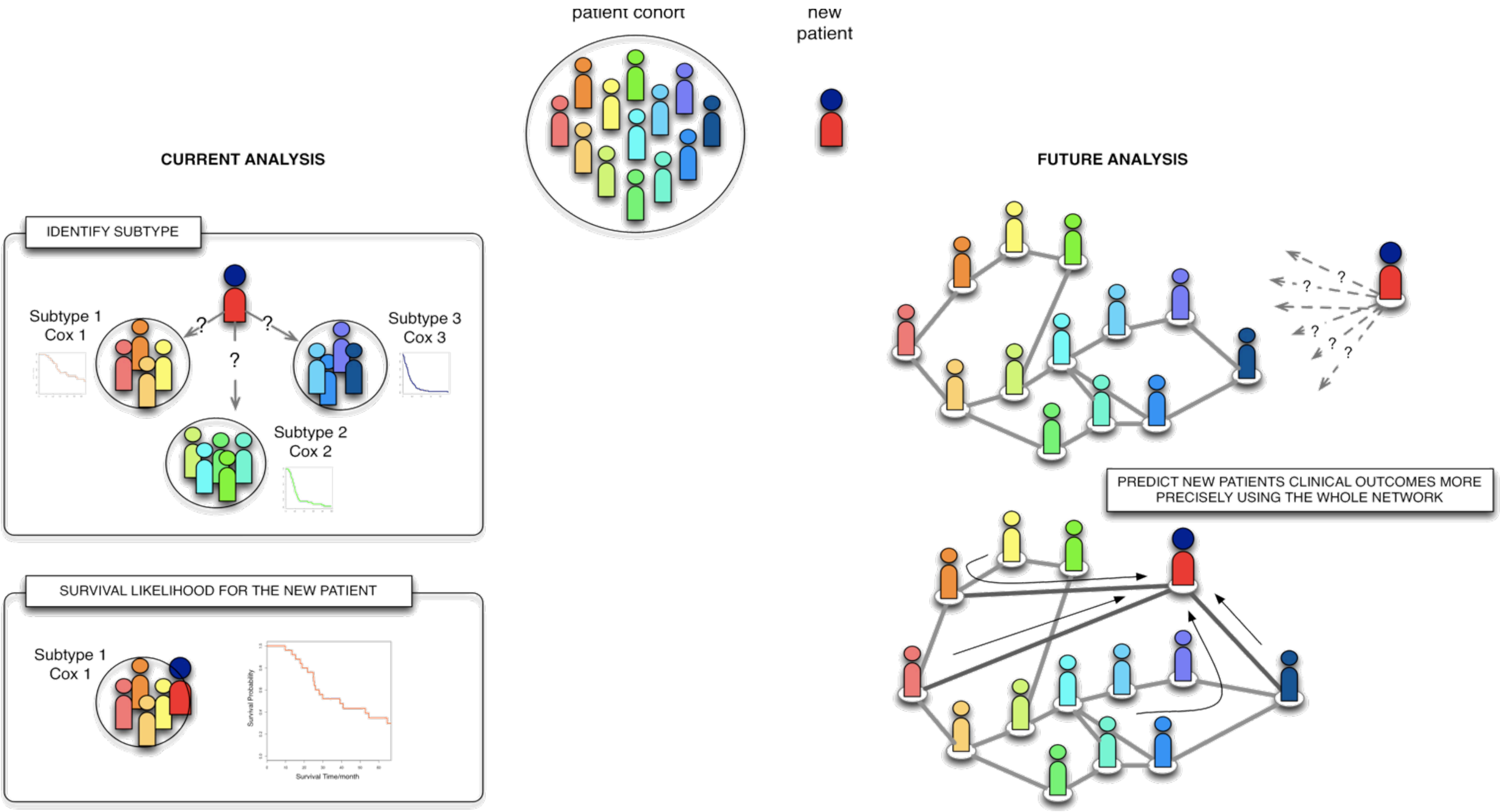


	PAM50 (5 clusters)	iCluster (10 clusters)	SNF (5 clusters)	SNF (10 clusters)
P value discovery cohort	3.0×10^{-9}	1.2×10^{-14}	6.10×10^{-11}	3.31×10^{-12}
P value validation cohort	1.7×10^{-9}	2.9×10^{-11}	5.12×10^{-13}	7.86×10^{-12}
CI discovery cohort	0.560	0.621	0.638	0.638
CI validation cohort	0.551	0.605	0.633	0.633

established

So how many subtypes are there really in breast cancer?

Predicting using the network



patient cohort

new patient

CURRENT ANALYSIS

IDENTIFY SUBTYPE

Subtype 1
Cox 1

Subtype 3
Cox 3

Subtype 2
Cox 2

SURVIVAL LIKELIHOOD FOR THE NEW PATIENT

Subtype 1
Cox 1

FUTURE ANALYSIS

PREDICT NEW PATIENTS CLINICAL OUTCOMES MORE PRECISELY USING THE WHOLE NETWORK

Predicting using the network

Cox objective $$lp(z) = \sum_{i=1}^{n} \delta_i \left(\mathbf{X}_{\boldsymbol{i}}^{\boldsymbol{T}} z - \log\left(\sum_{j \in \boldsymbol{R}(\boldsymbol{t_i})} \exp\left(\mathbf{X}_{\boldsymbol{j}}^{\boldsymbol{T}} z \right) \right) \right)$$

Predicting using the network

Cox objective

$$lp(z) = \sum_{i=1}^{n} \delta_i \left(\mathbf{X}_{\boldsymbol{i}}^{\boldsymbol{T}} z - \log\left(\sum_{j \in \boldsymbol{R}(\boldsymbol{t_i})} \exp\left(\mathbf{X}_{\boldsymbol{j}}^{\boldsymbol{T}} z\right) \right)\right)$$

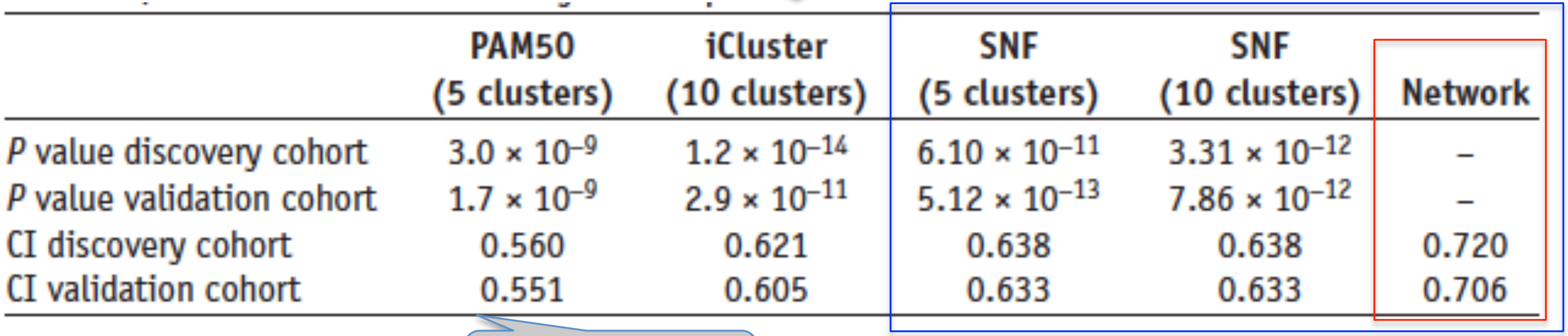
Our network-regularized objective

$$lp(\boldsymbol{z}) = \sum_{i=1}^{n} \delta_i \left(X_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{z} - \log\left(\sum_{j \in \boldsymbol{R}(\boldsymbol{t_i})} \exp(X_{\boldsymbol{j}}^{\boldsymbol{T}} \boldsymbol{z}) \right)\right) - \lambda \sum_{i} \sum_{j} (X_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{z} - X_{\boldsymbol{j}}^{\boldsymbol{T}} \boldsymbol{z})^2 w_{ij}$$

Predicting using the network
Breast Cancer (METABRIC example)

CNV and expression data
Discovery: 997 patients
Validation: 995 patients

Nature, 2012



	PAM50 (5 clusters)	iCluster (10 clusters)	SNF (5 clusters)	SNF (10 clusters)	Network
P value discovery cohort	3.0×10^{-9}	1.2×10^{-14}	6.10×10^{-11}	3.31×10^{-12}	–
P value validation cohort	1.7×10^{-9}	2.9×10^{-11}	5.12×10^{-13}	7.86×10^{-12}	–
CI discovery cohort	0.560	0.621	0.638	0.638	0.720
CI validation cohort	0.551	0.605	0.633	0.633	0.706

established