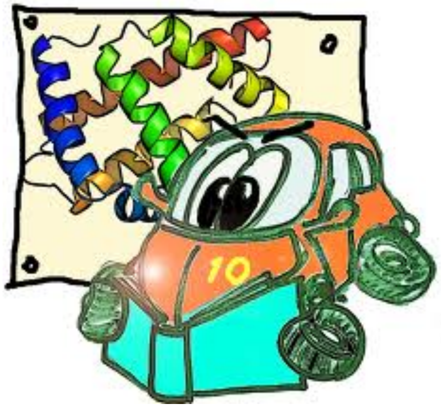
Machine Learning in Computational Biology CSC 2431



Lecture 9: Combining biological datasets

Instructor: Anna Goldenberg

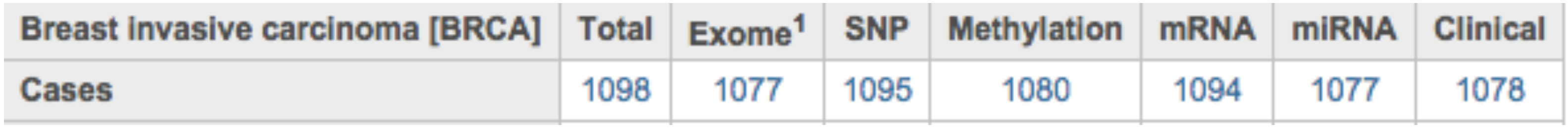
What kind of data integration is there?

What kind of data integration is there?

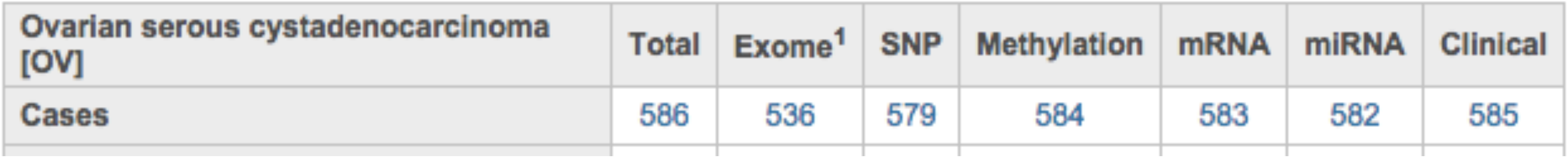
- SNPs and gene expression
- Networks and gene expression (and mutations)
- ENCODE data. Combining different epigenetic signals and binding info
- Ontologies and genome annotations

- Now: integrating patient data

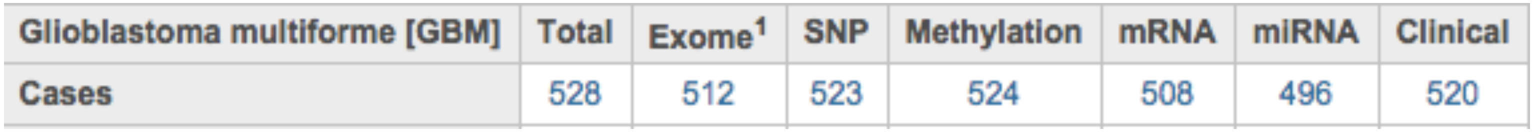
Data is available
E.g. The Cancer Genome Atlas (TCGA)



Breast invasive carcinoma [BRCA]	Total	Exome[1]	SNP	Methylation	mRNA	miRNA	Clinical
Cases	1098	1077	1095	1080	1094	1077	1078



Ovarian serous cystadenocarcinoma [OV]	Total	Exome[1]	SNP	Methylation	mRNA	miRNA	Clinical
Cases	586	536	579	584	583	582	585



Glioblastoma multiforme [GBM]	Total	Exome[1]	SNP	Methylation	mRNA	miRNA	Clinical
Cases	528	512	523	524	508	496	520

Total of 33 cancers.
9 cancers have over 500+ samples
All publicly available!

Why integrate patient data

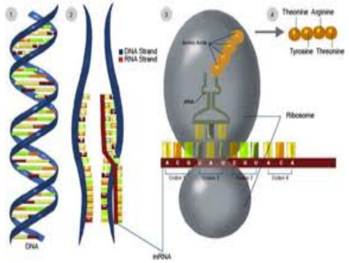
Why integrate patient data

- To identify more homogeneous subsets of patients (that might respond similarly to a given drug)

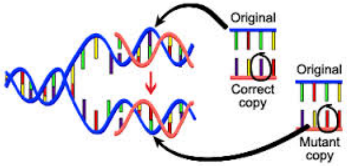
- To help better predict response to drugs

Single data type driven integration

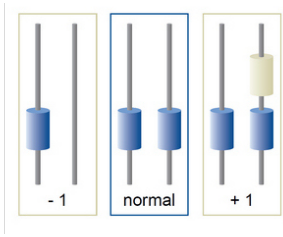
mRNA



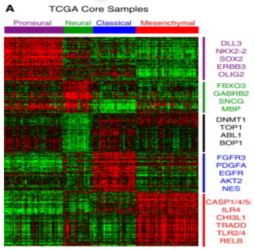
mutations



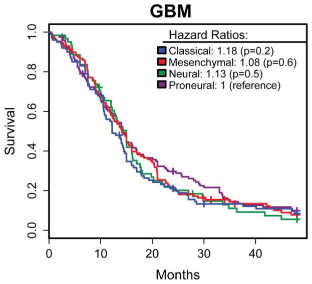
CNV



clinical



+ more genes + more genes →

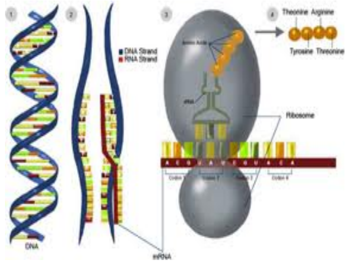


p-value = {0.2,0.6,0.5}

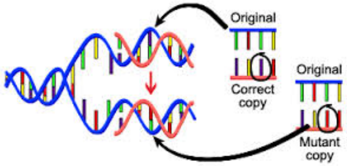
(Verhaak et al, Cancer Cell, 2010)

Single data type driven integration

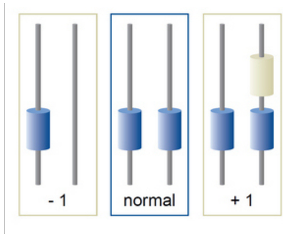
mRNA



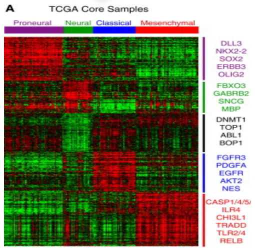
mutations



CNV

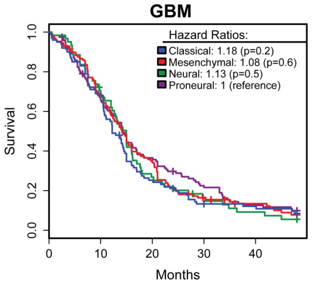


clinical



more genes

more genes

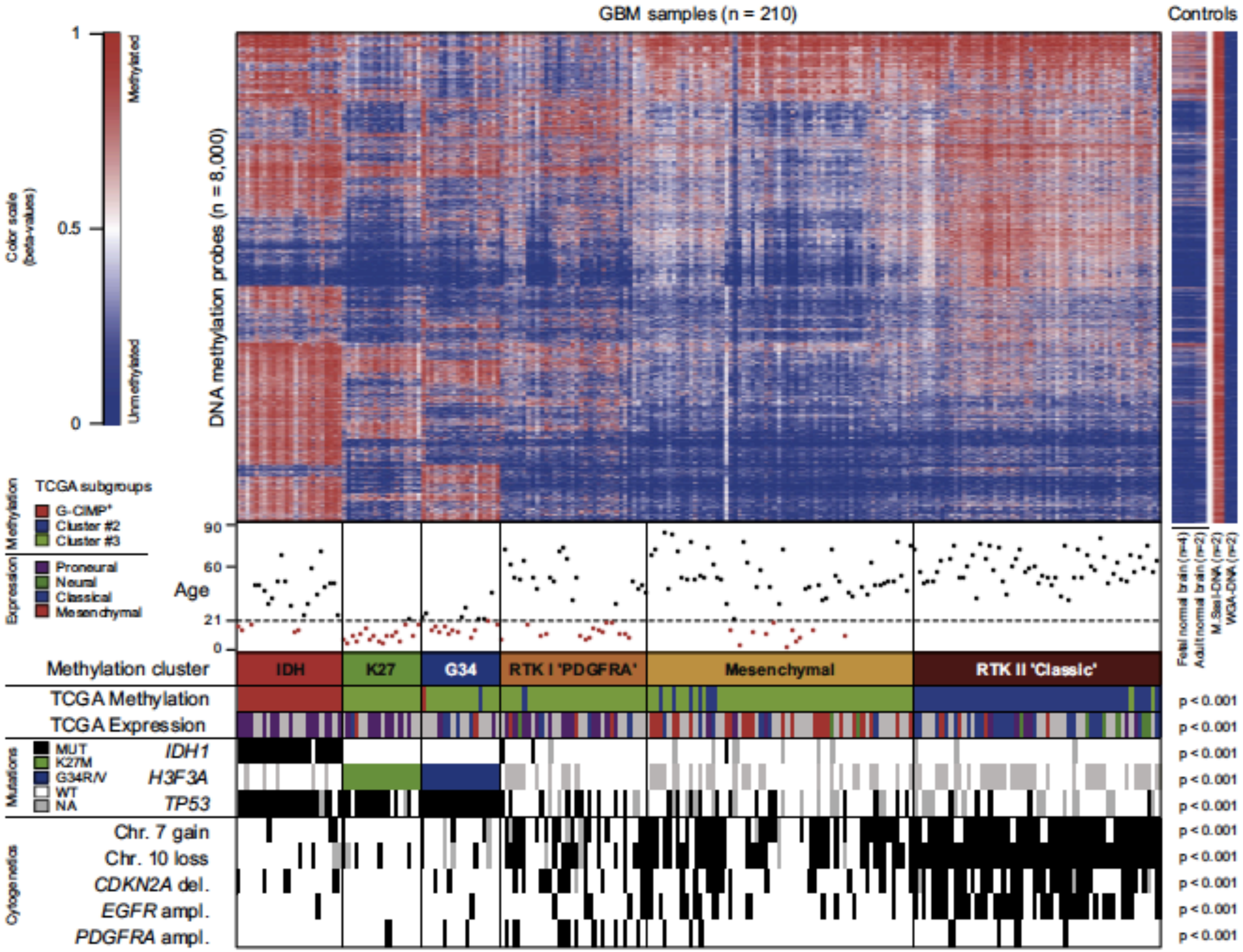


p-value = {0.2,0.6,0.5}

What about methylation data?

(Verhaak et al, Cancer Cell, 2010)

More recent GBM study (Sturm et al, 2012)



Methods used in Verhaak 2010

- Factor analysis – a dimensionality reduction method – used to integrate mRNA data from 3 platforms
- Consensus clustering (consensus average linkage clustering) (Monti et al, 2003)
- SigClust – cluster significance (Liu et al, 2008)
- Silhouette to identify core of clusters (Rousseeuw, 1987)
- ClaNC – nearest centroid-based classifier to identify gene signatures (Dabney, 2006)

More recent GBM study (Sturm, 2012)

- Missing values – imputed using k-NN (Troyanskaya, 2001)
- Unsupevised consensus clustering (R: clusterCons) (Monti, 2003, Wilkerson and Hayes, 2010)
- Consensus matrix was calculated using the k-means algorithm
- Number of clusters is decided by visual assessment

Breast Cancer Analysis (TCGA,2012)

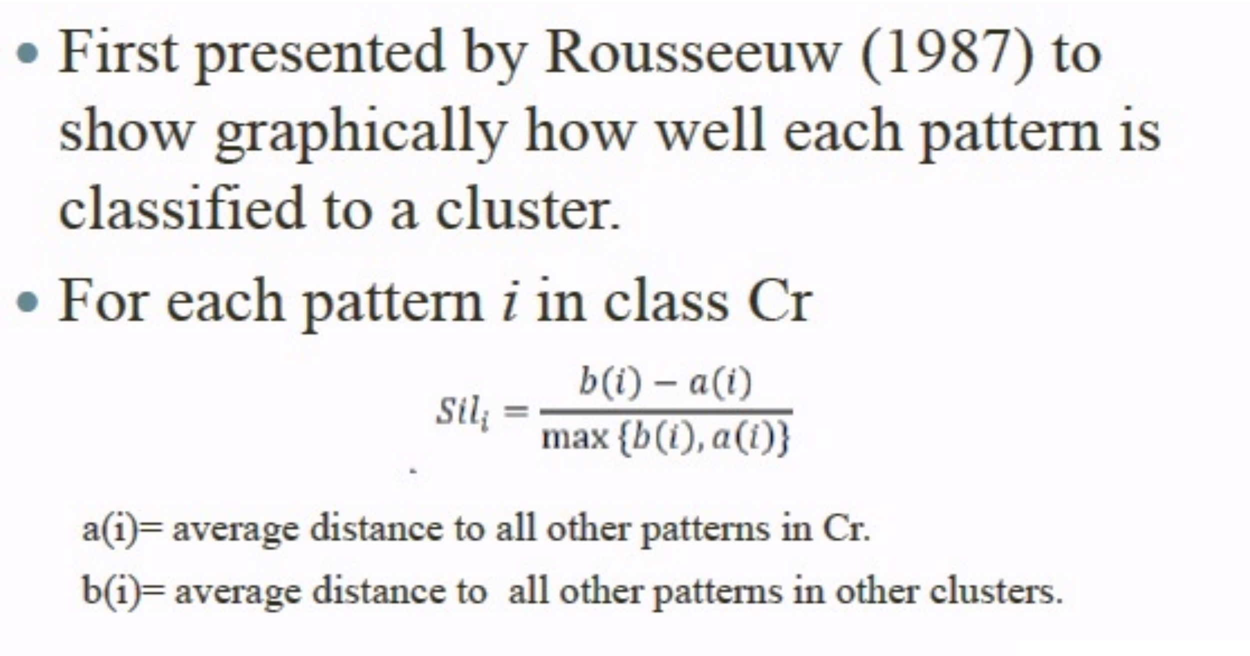
- Integrated pathway analysis using PARADIGM
- Significantly mutated genes were identified using MuSiC package
- NMF for unsupervised clustering of somatic and CNV data, protein expression
- RPMM – recursively partitioned mixture model (RPMM Bioconductor package)
- ConsensusClusterPlus (R-package) to combine clustering based on single data type
- MEMo (Mutual Exclusivity Modules) – identifies mutually exclusive alterations targeting frequently altered genes that are likely to belong to the same pathway

PARADIGM

- *Infers Integrated Pathway Levels (IPLs) for genes, complexes, and processes using pathway interactions and genomic and functional genomic data from a single patient sample.*
- Data:
  - mRNA relative to normal samples
  - CNVs mapped to genes
  - Networks: Biocarta (Biocarta, NCIPID, Reactome) – Superimposed into SuperPathway
- Approach: belief propagation to maximize likelihood (hear more next class!)

Vaske, C. J. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. (2010) Bioinformatics 26

Silhouette statistic



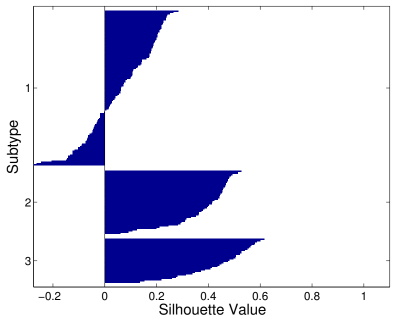
- First presented by Rousseeuw (1987) to show graphically how well each pattern is classified to a cluster.
- For each pattern *i* in class Cr

$$Sil_i = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

a(i)= average distance to all other patterns in Cr.

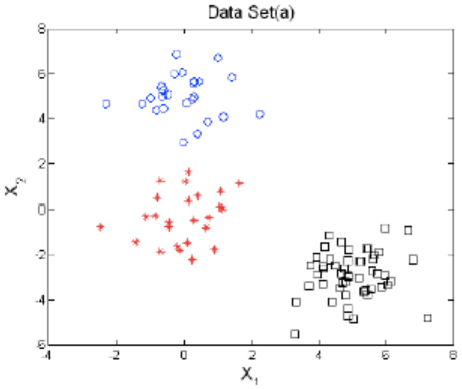
b(i)= average distance to all other patterns in other clusters.

- . $-1 \leq Sil_i \leq 1$
- Sil=1 : good assignment
- Sil=-1: wrong (bad) assignment
- Sil=0 : don't know ; pattern could be belong to either its current cluster or its nearest cluster.

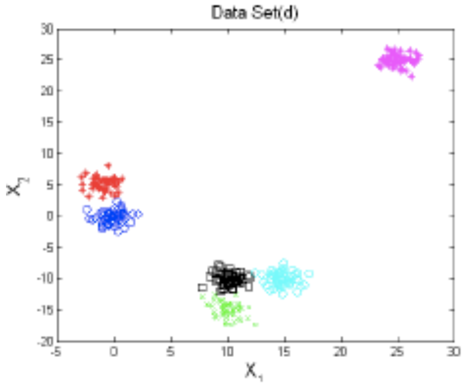


Silhouette statistic

a. Three clusters in 2 dimensions
b. Three clusters in 10 dimensions, each cluster has 50 observations
c. 4 clusters in 10 dimensions with randomly chosen centers
d. Six clusters in 2 dimensions



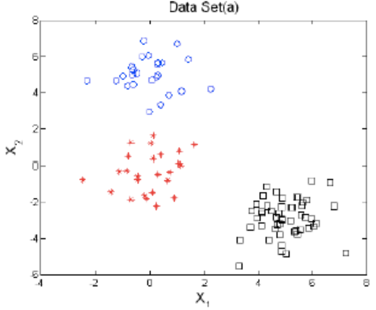
(a)

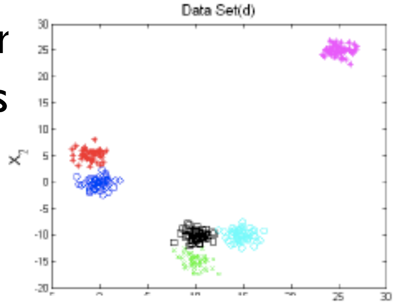


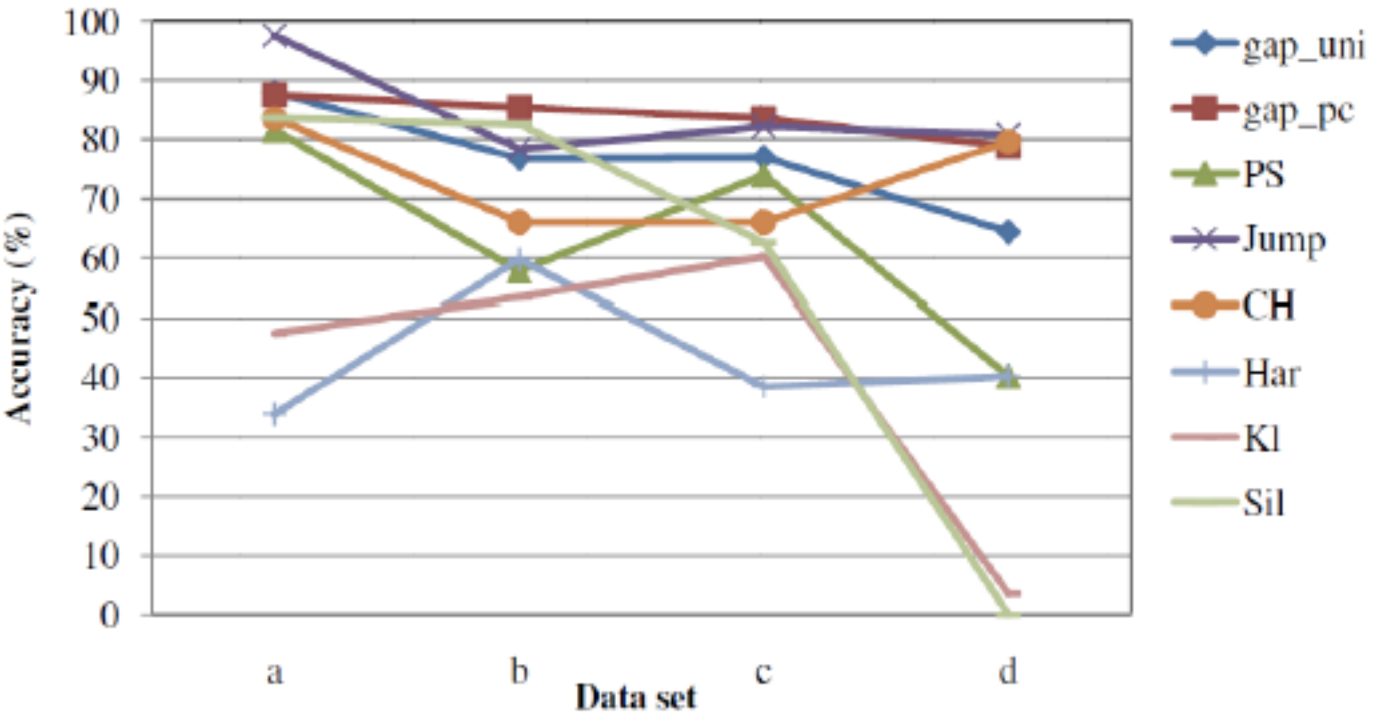
(d)

Silhouette statistic

a. Three clusters in 2 dimensions
b. Three clusters in 10 dimensions, each cluster has 50 obser
c. 4 clusters in 10 dimensions with randomly chosen centers
d. Six clusters in 2 dimensions







Hossein Parsaei. Finding a number of clusters

NMF – non-negative matrix factorization

- Matrix factorization: NMF(V) = WxH
- W and H are *non-negative*
- Current methods (many – gradient descent, alternating non-negative least squares, etc)
- Arora et al (2012) – exact NMF method runs in polynomial time under separability condition of W

Consensus Clustering

- *Resampling based method for class discovery and visualization of gene expression microarray data*
- Goal: assessing stability
- Method:
  - For a 1000 iterations
    1. Resample data
    2. Cluster with fav. clust. method (hier, k-means)
  - Compute consensus matrix $\mathcal{M}(i,j) = \frac{\sum_h M^{(h)}(i,j)}{\sum_h I^{(h)}(i,j)}$
  - Partition D based on Consensus Matrix

Monti, S., Tamayo, P., Mesirov, J., Golub, T. (2003) Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning, 52, 91-118.

SigClust

- Goal: assess statistical signficance of clustering
- H_0: data comes from a single Gaussian
- H_1: not from a single Gaussian
- Statistic: Cluster Index (CI) - sum of within-class sums of squares about the mean of the cluster divided by the total sum of squares about the overall mean (mean-shift and scale invariant)

Liu, Yufeng, Hayes, David Neil, Nobel, Andrew and Marron, J. S, 2008, Statistical Significance of Clustering for High-Dimension, Low-Sample Size Data, Journal of the American Statistical Association 103(483) 1281–1293

Patient Specific Data Fusion (Yuan et al, 2011)

- Nonparametric Bayesian model (gene expression and CNV)
  - Feature selection (each feature is drawn from a multinomial distribution with unknown class proabilities
  - MCMC inference

Multiple Kernel Learning

- Mostly used in supervised cases, but exists in unsupervised scenario (Chuang, CVPR, 2012)
- Linear combination of kernels

$$K_{combine} = \sum_{v=1}^{m} \alpha_v K_v$$

iCluster (Shen et al, 2009)

- Gaussian latent variable model
- Sparsity regularization (Lasso-type)
- Latent variables (embedding is shared)

$$\mathbf{x}_{ik} = \mathbf{W}_k \mathbf{z}_i + \epsilon_{ik}, i = 1, \ldots, n, k = 1, \ldots, m$$

Drawbacks of existing methods

- A lot of manual processing
- Many steps in the pipeline
- Integration mostly done in the feature space – if there is signal in a combination of features, it'll be lost
- Focusing on consensus – what if there is complementary information?

Similarity Network Fusion (Wang et al, 2014)

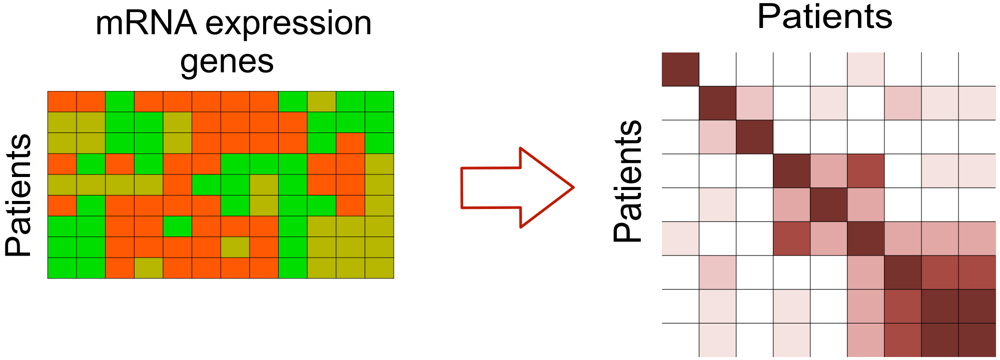
- Integrate data in the patient space

1. Construct patient similarity matrix
2. Fuse multiple matrices

1. Construct similarity networks

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

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

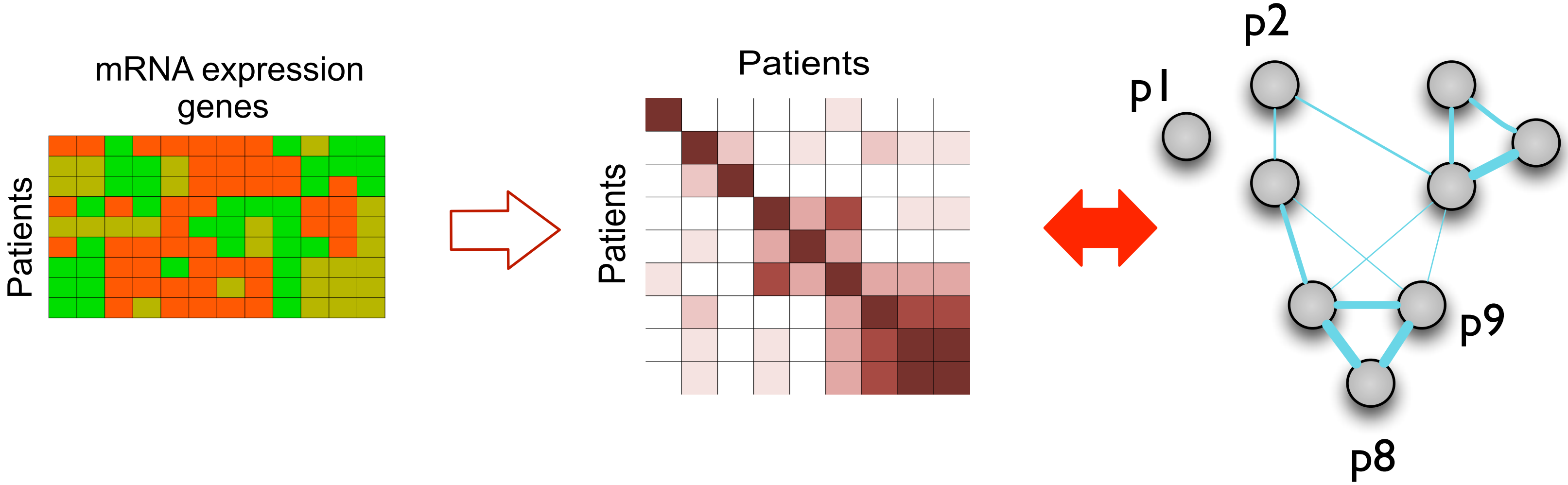


1. Construct similarity networks

Sparsification

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

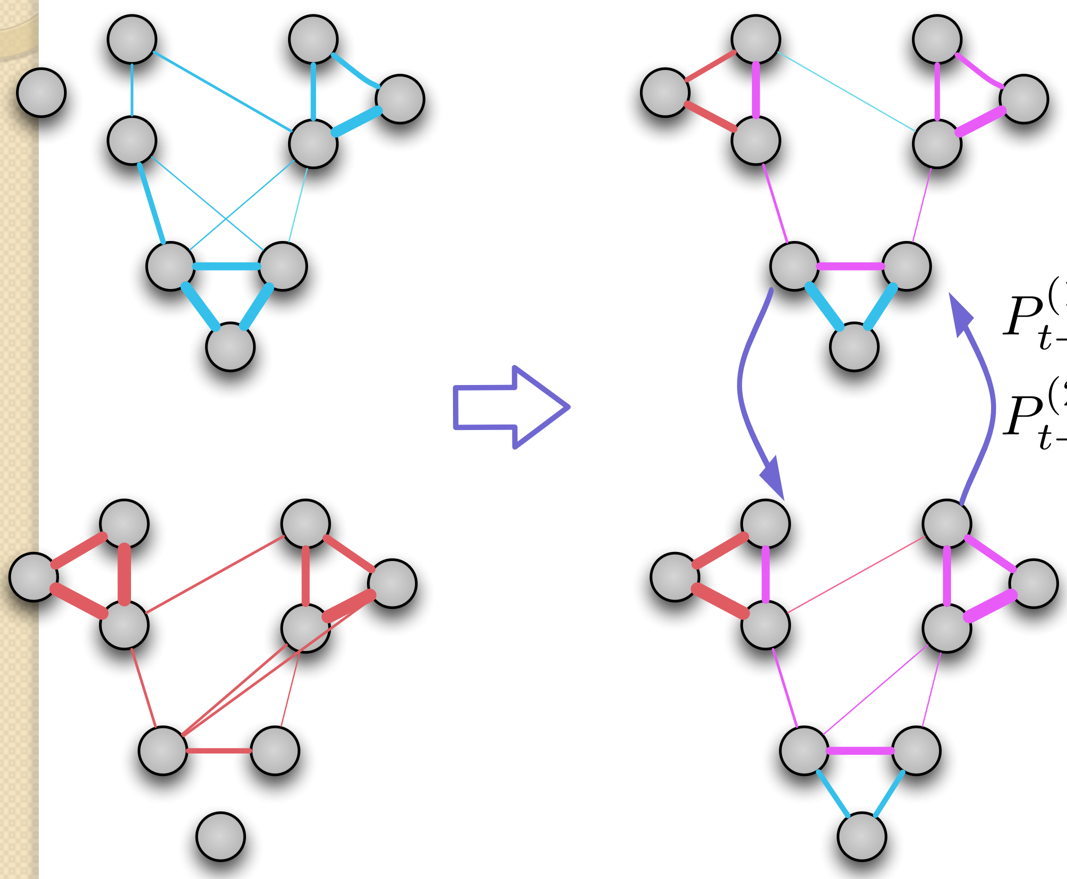
2) $$\mathcal{P}(i,j) = \frac{\mathcal{W}(i,j)}{\sum_{x_k \in KNN(x_i)} \mathcal{W}(i,k)}$$



2. Combine networks

Similarity Networks

Fusion Iterations

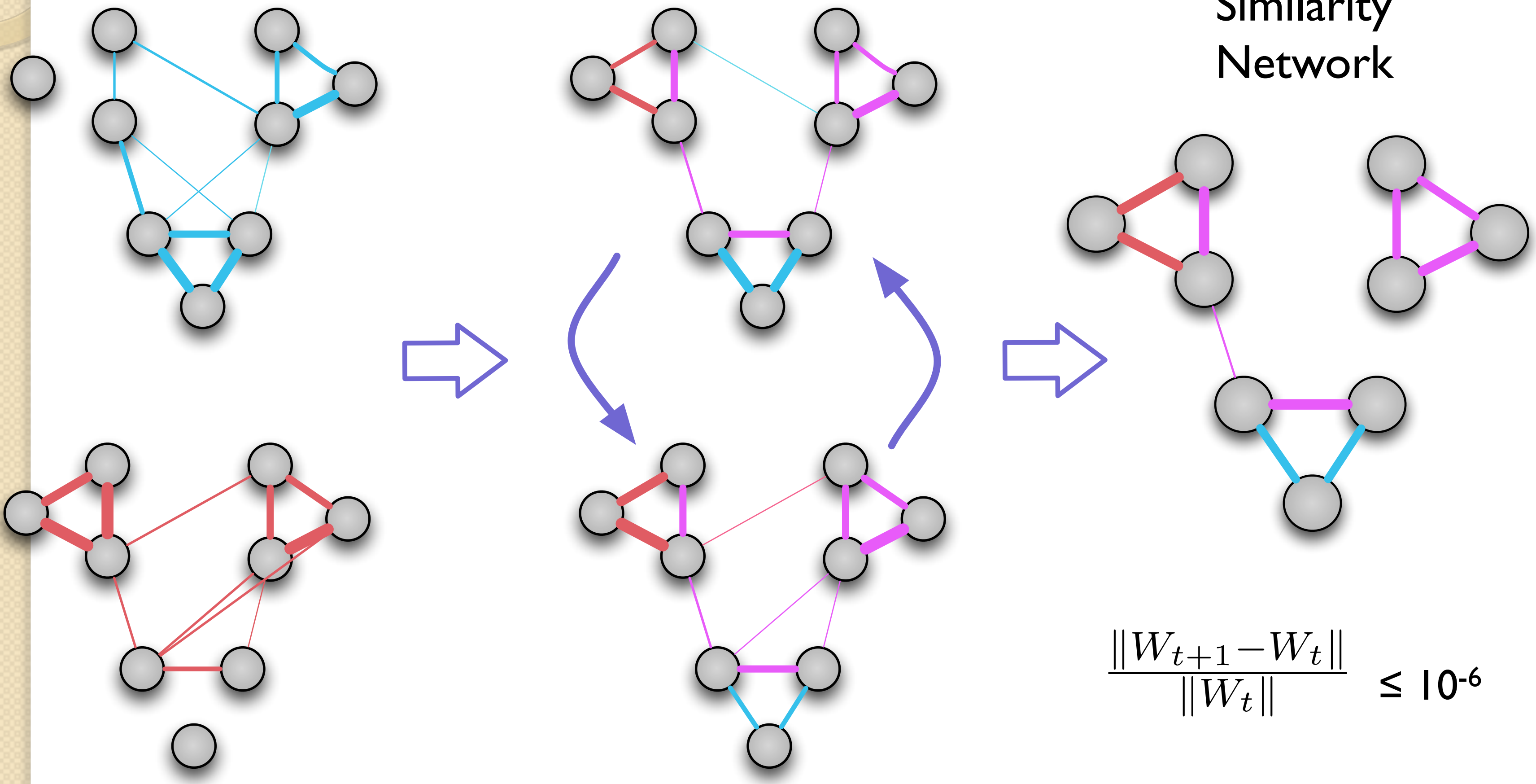


$$P^{(1)}_{t+1} = \mathcal{P}^{(1)} \times (P^{(2)}_t) \times (\mathcal{P}^{(1)})'$$

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

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

2. Combine networks



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:

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

Experiments

Data:

2 simulations

5 TCGA cancers

METABRIC (Large Breast Cancer db)

Comparative Methods:

Concatenation

iCluster

PDSB

Multiple kernel learning

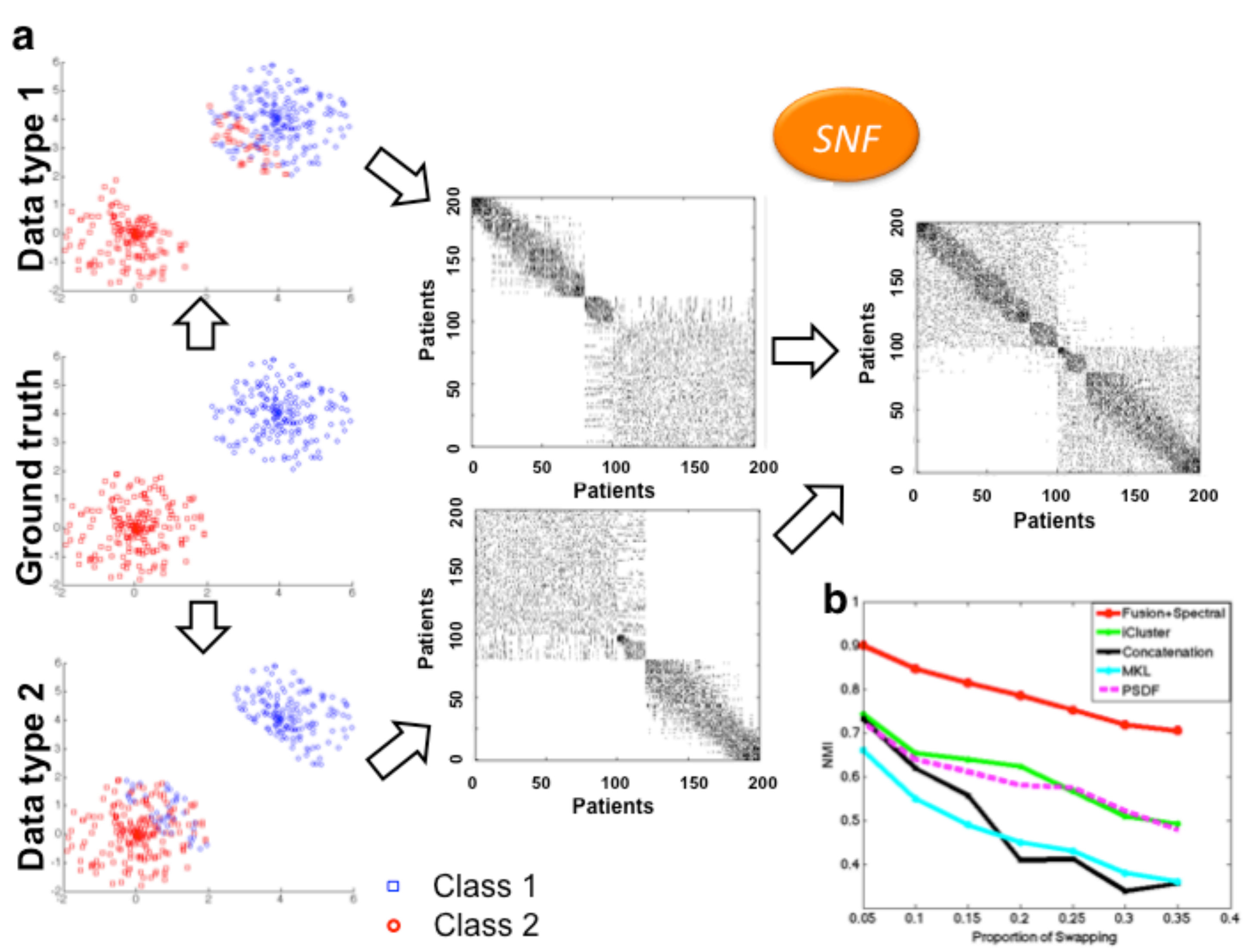
Criteria:

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

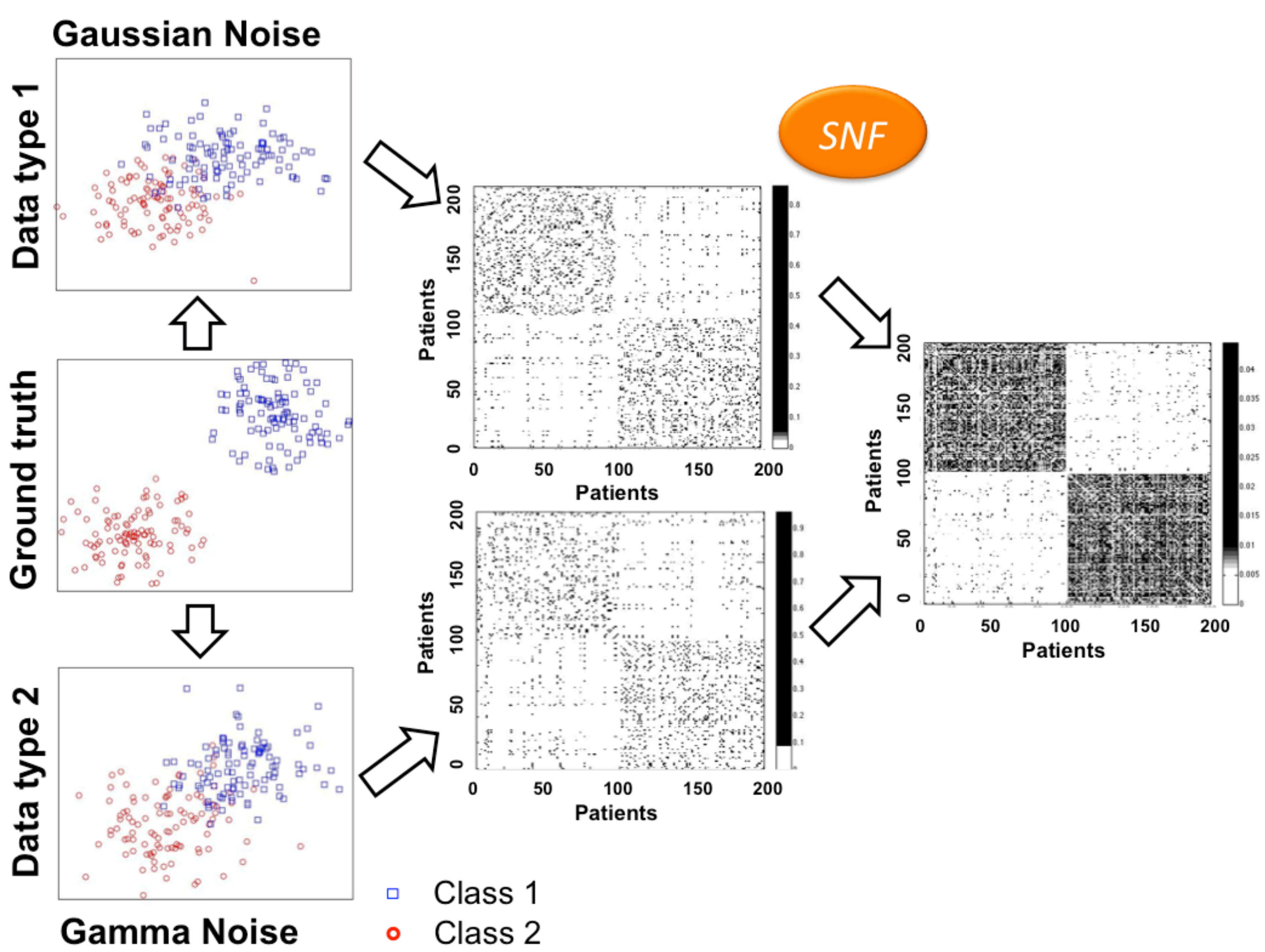
Silhouette score (cluster homogeneity)

Running time

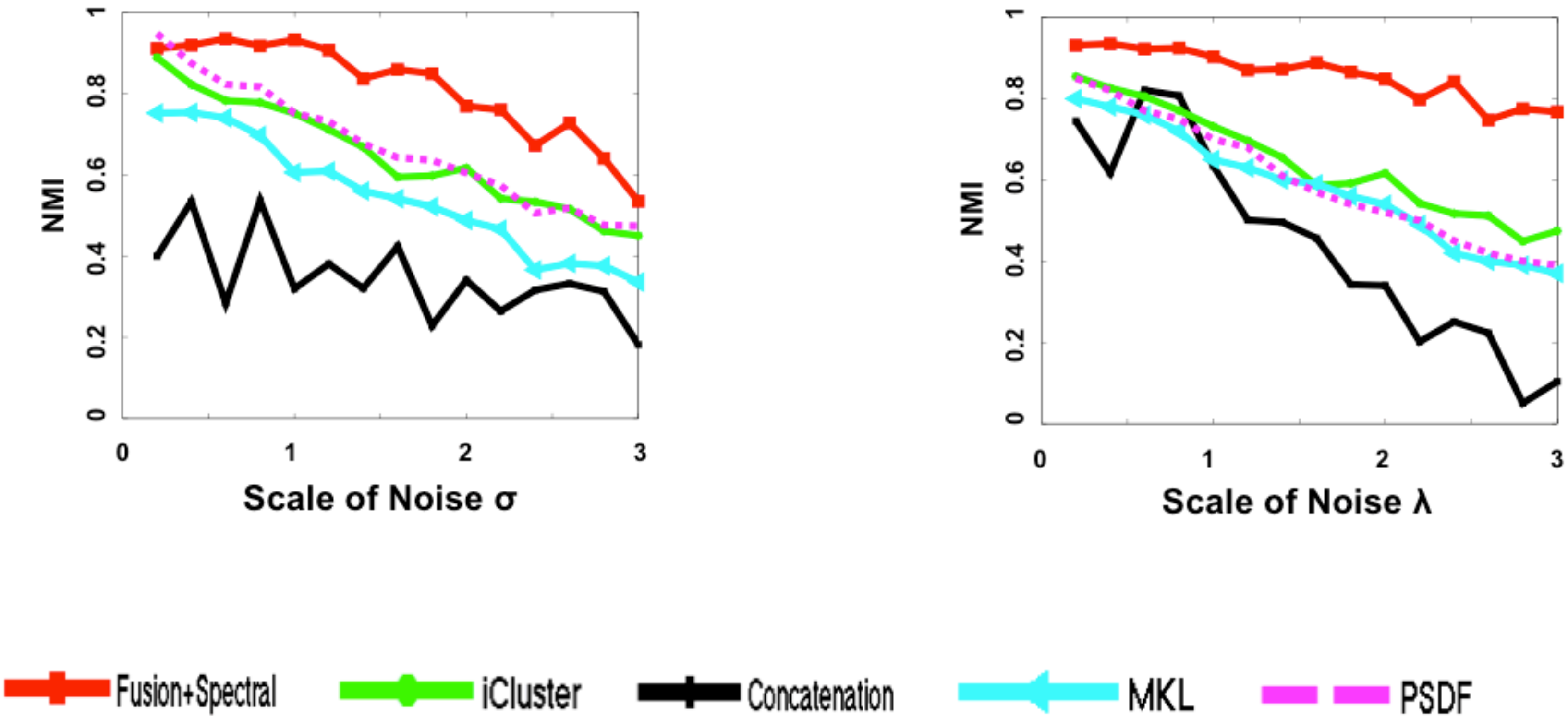
Simulation 1 – complementarity



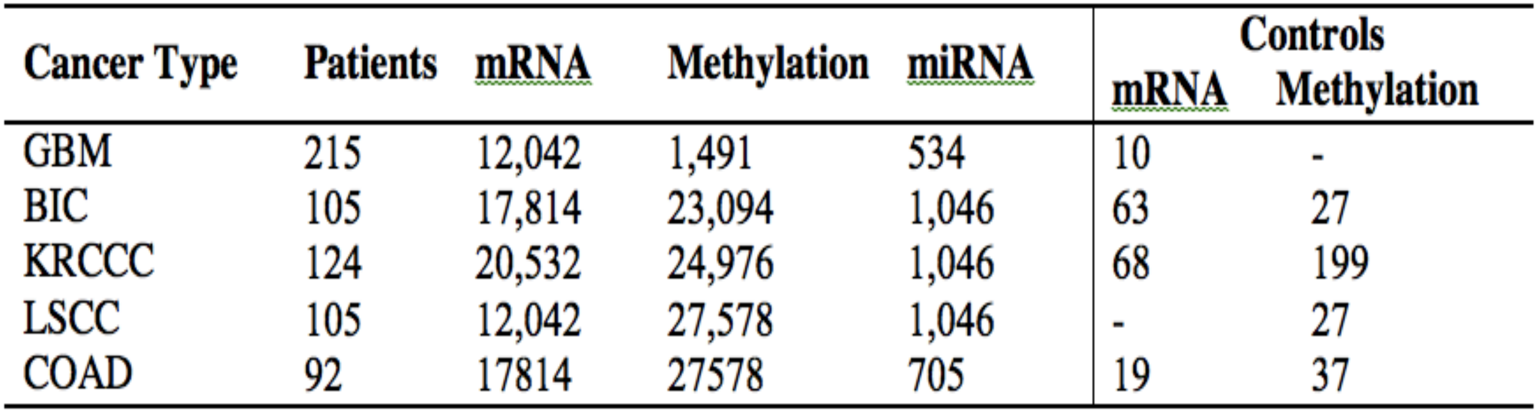
Simulation 2 - removing noise



Simulation 2 - removing noise

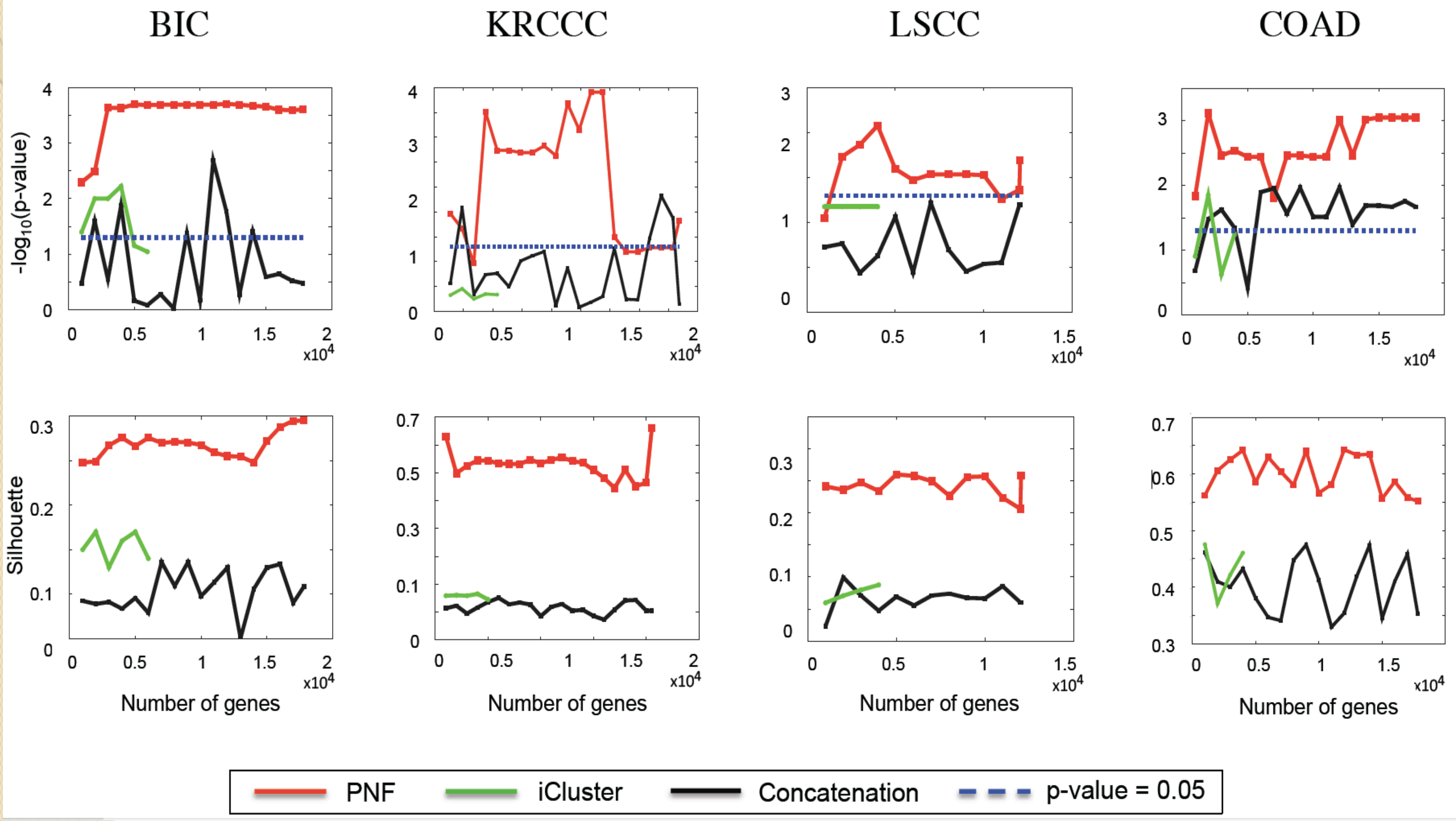


TCGA Data



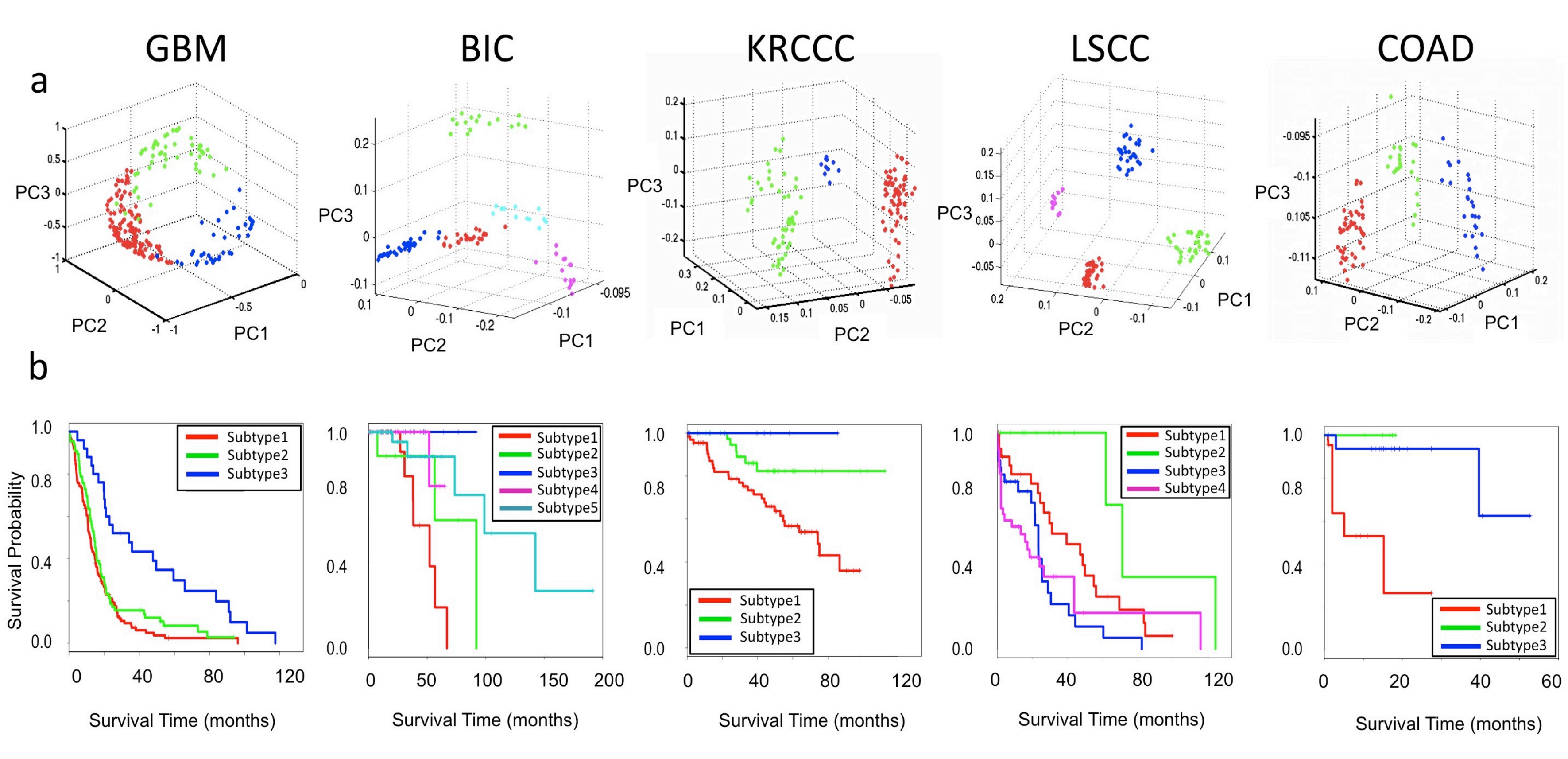
Cancer Type	Patients	mRNA	Methylation	miRNA	Controls	
					mRNA	Methylation
GBM	215	12,042	1,491	534	10	-
BIC	105	17,814	23,094	1,046	63	27
KRCCC	124	20,532	24,976	1,046	68	199
LSCC	105	12,042	27,578	1,046	-	27
COAD	92	17814	27578	705	19	37

Gene pre-selection across cancers



Bo Wang

Clustering of the network



Bo Wang

Patient networks: advantages and disadvantages

- Integrative feature selection
- Growing the network requires extra work
- Unsupervised – hard to turn into a supervised problem

- ✓ 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

Package on CRAN: **SNFtool**

Data integration - future

Data integration - future

- Simultaneous feature selection and data integration
- Supervised vs unsupervised approaches – do we really need unsupervised methods?
- Priors on contributions of different types of data
- Automate feature pre-selection if necessary

Next class

- iCluster – joint latent variable model (Shen et al, 2009) - Ladislav
- PARADIGM – Andrew
- Next topic: pharmacogenomics (guest lecture by Dr Benjamin Haibe-Kains)