# Non-parametric integrative unsupervised clustering with multiple 'omic datasets

Prabhakar Chalise
Brooke L. Fridley

June 18, 2021

KU MEDICAL CENTER
The University of Kansas

# Introduction

PLOS ONE

RESEARCH ARTICLE

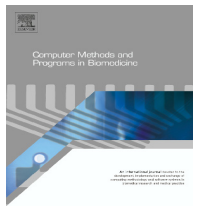
## Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm

**Prabhakar Chalise[1]*, Brooke L. Fridley[2]**

**1** Department of Biostatistics, University of Kansas Medical Center, Kansas City, Kansas, United States of America, **2** Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, Florida, United States of America

ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/cmpb




## *InterSIM*: Simulation tool for multiple integrative 'omic datasets'

CrossMark

*Prabhakar Chalise*[*], *Rama Raghavan*[1], *Brooke L. Fridley*[2]

*Department of Biostatistics, University of Kansas Medical Center, 3901 Rainbow Blvd, Kansas City, KS 66160, United States*

# Introduction

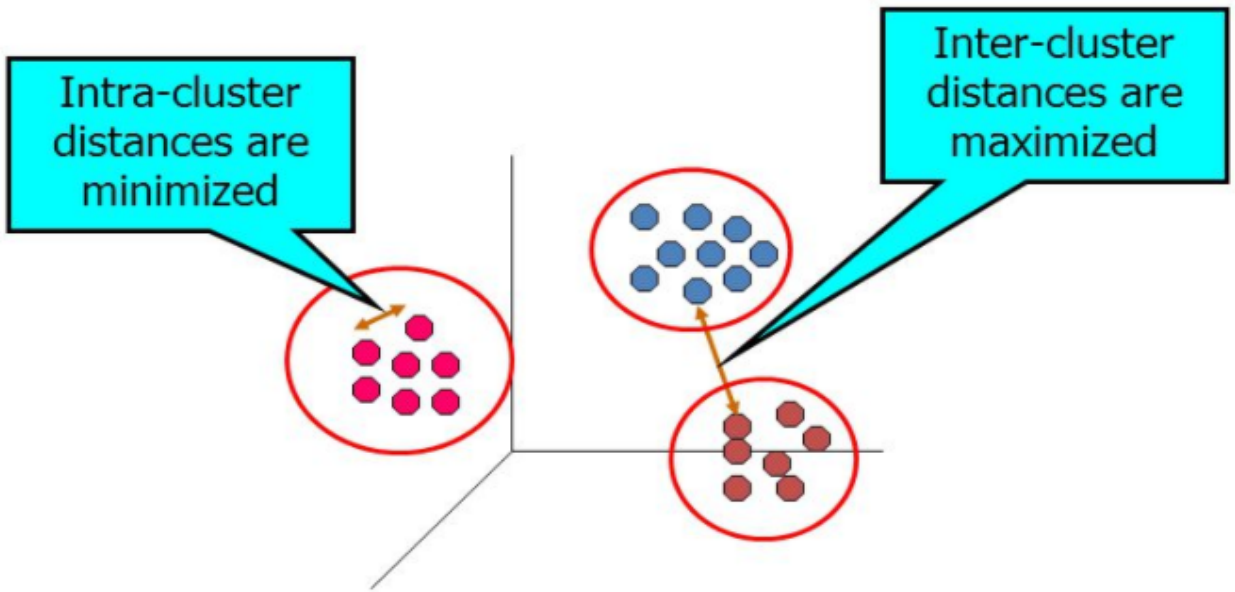
• A fundamental problem in many high dimensional data-analysis tasks is to find a suitable representation of the data.

• A useful representation involves making latent structure in the data more explicit so that further analytic methods can be applied.

• Clustering analysis is a method that extracts such latent structure.

# Clustering

• Clustering is a method of grouping the objects to a discrete set of classes (i.e. clusters) such that the objects within the same class are more similar (or related) to one another compared to objects in different classes (or unrelated)




• Hierarchical Clustering, K-means, Non negative matrix factorization (NMF), model based clustering etc

# Clustering

- Clustering can be done for genomic features or subjects.
  - use on genes, groups with similar expression pattern across the samples $\Rightarrow$ unknown gene-network function or pathway (e.g. WGCNA)
  - use on subjects, group of subjects with similar genomic profiles across all genes $\Rightarrow$ subtypes of disease

# Why do we care?

• A tumor may look like histologically similar but it may be different at molecular level.

• The response trajectory to drug or survival rate may be very different for the subgroup of same cancer.

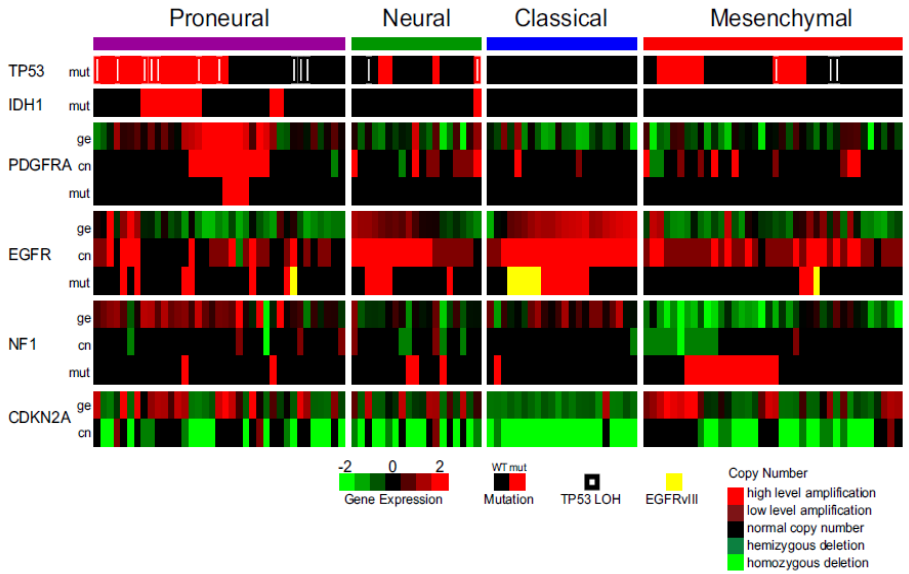
• If clinically meaningful subgroups exist, perhaps we might need different treatment for patients having different subtypes.

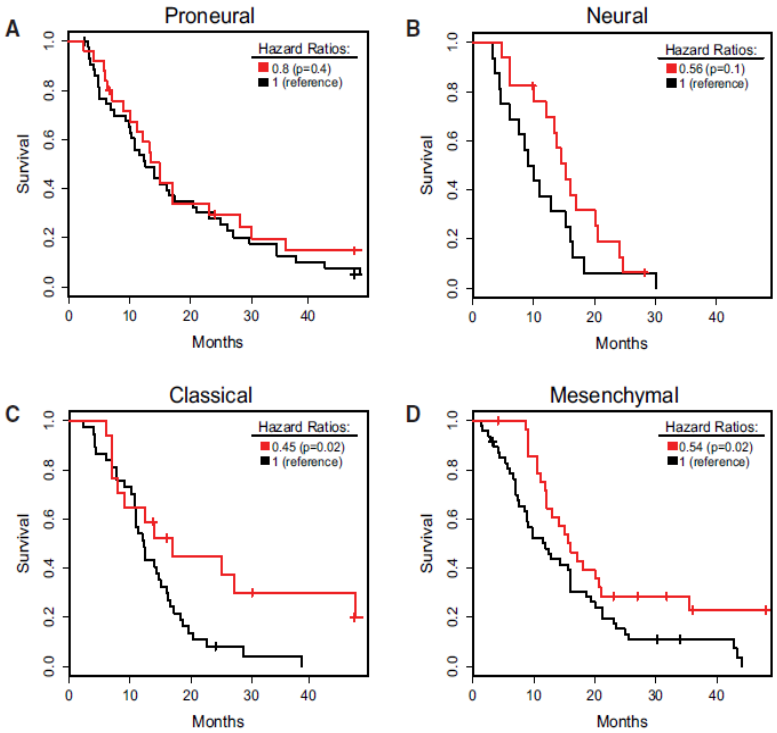
# A few published examples

Cancer Cell
Article

## Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*

Roel G.W. Verhaak,[1,2,17] Katherine A. Hoadley,[3,4,17] Elizabeth Purdom,[7] Victoria Wang,[8] Yuan Qi,[4,5]



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

Archive > Volume 490 > Issue 7418 > Articles > Article

*NATURE* | ARTICLE OPEN

日本語要約

# Comprehensive molecular portraits of human breast tumours

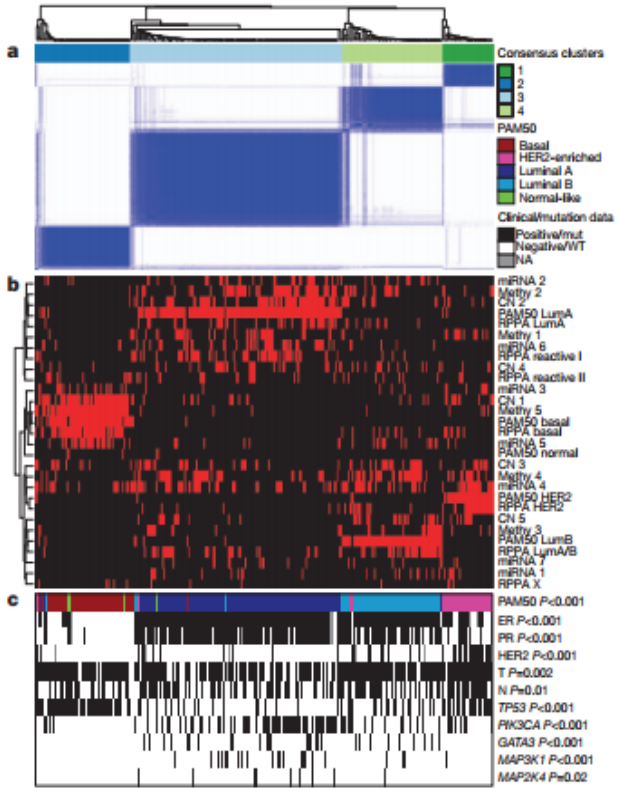
The Cancer Genome Atlas Network

Affiliations | Contributions

*Nature* **490**, 61–70 (04 October 2012) | doi:10.1038/nature11412

Received 22 March 2012 | Accepted 11 July 2012 | Published online 23 September 2012

| Corrected online **03 October 2012**




## Four subtypes:

1. **Luminal A**
2. **Luminal B**
3. **Her2**
4. **Basal**

Proc Natl Acad Sci U S A. 2001 Sep 11; 98(19): 10869–10874. PMCID: PMC58566
doi: 10.1073/pnas.191367098
Medical Sciences

## Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Therese Sørlie,[a,b,c] Charles M. Perou,[a,d] Robert Tibshirani,[e] Turid Aas,[f] Stephanie Geisler,[g] Hilde Johnsen,[b] Trevor Hastie,[e] Michael B. Eisen,[h] Matt van de Rijn,[i] Stefanie S. Jeffrey,[j] Thor Thorsen,[k] Hanne Quist,[l] John C. Matese,[c] Patrick O. Brown,[m] David Botstein,[c] Per Eystein Lønning,[g] and Anne-Lise Børresen-Dale[b,n]

Author information ▶ Article notes ▶ Copyright and License information ▶

Proc Natl Acad Sci U S A. 2004 Jan 20; 101(3): 811–816. PMCID: PMC321763
Published online 2004 Jan 7. doi: 10.1073/pnas.0304146101
Medical Sciences

## Gene expression profiling identifies clinically relevant subtypes of prostate cancer

Jacques Lapointe,[a,b,c] Chunde Li,[d] John P. Higgins,[a] Matt van de Rijn,[a] Eric Bair,[e] Kelli Montgomery,[a] Michelle Ferrari,[c] Lars Egevad,[d] Walter Rayford,[f] Ulf Bergerheim,[g] Peter Ekman,[d] Angelo M. DeMarzo,[h] Robert Tibshirani,[e,i] David Botstein,[j] Patrick O. Brown,[b,k] James D. Brooks,[c,l] and Jonathan R. Pollack[a,m]

Ang et al. BMC Cancer 2010, 10:227
http://www.biomedcentral.com/1471-2407/10/227


RESEARCH ARTICLE Open Access

## Comprehensive profiling of DNA methylation in colorectal cancer reveals subgroups with distinct clinicopathological and molecular features

Pei Woon Ang[1,2], Marie Loh[1,2], Natalia Liem[3], Pei Li Lim[3], Fabienne Grieu[1], Aparna Vaithilingam[2], Cameron Platell[1,4], Wei Peng Yong[3], Barry Iacopetta[1] and Richie Soong*[2]

# Motivation

- High methylation for a CpG island upstream of a gene ⇒ lower mRNA expression "gene silencing";
- Higher mRNA gene expression ⇒ higher downstream protein expression.

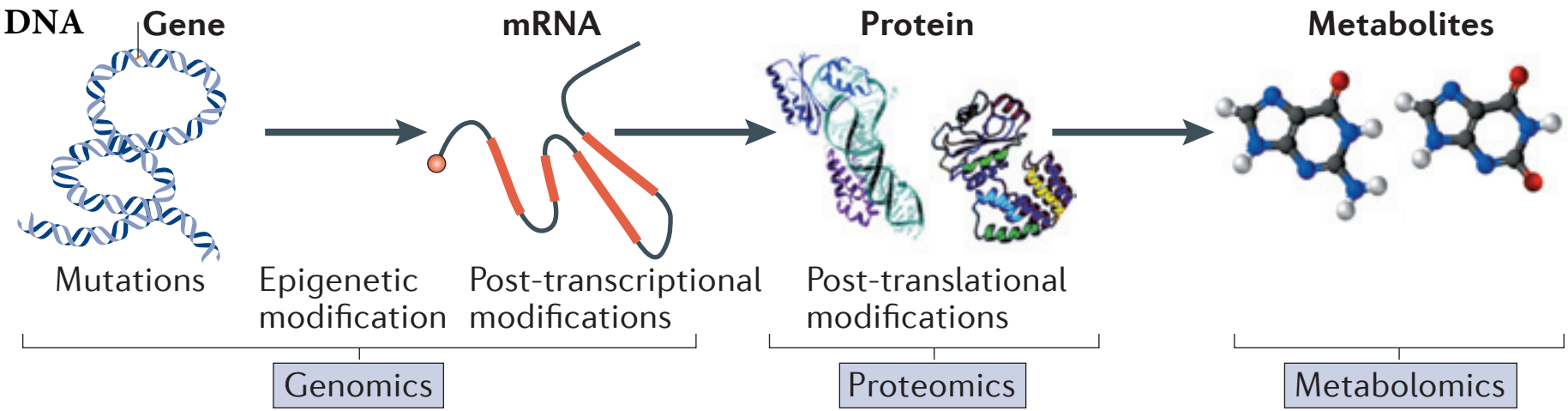



Figure source: http://www.nature.com/nrm/journal/v13/n4/fig_tab/nrm3314_F2.html

# Why should we integrate data?

Focussing only on one platform risks missing obvious signal!

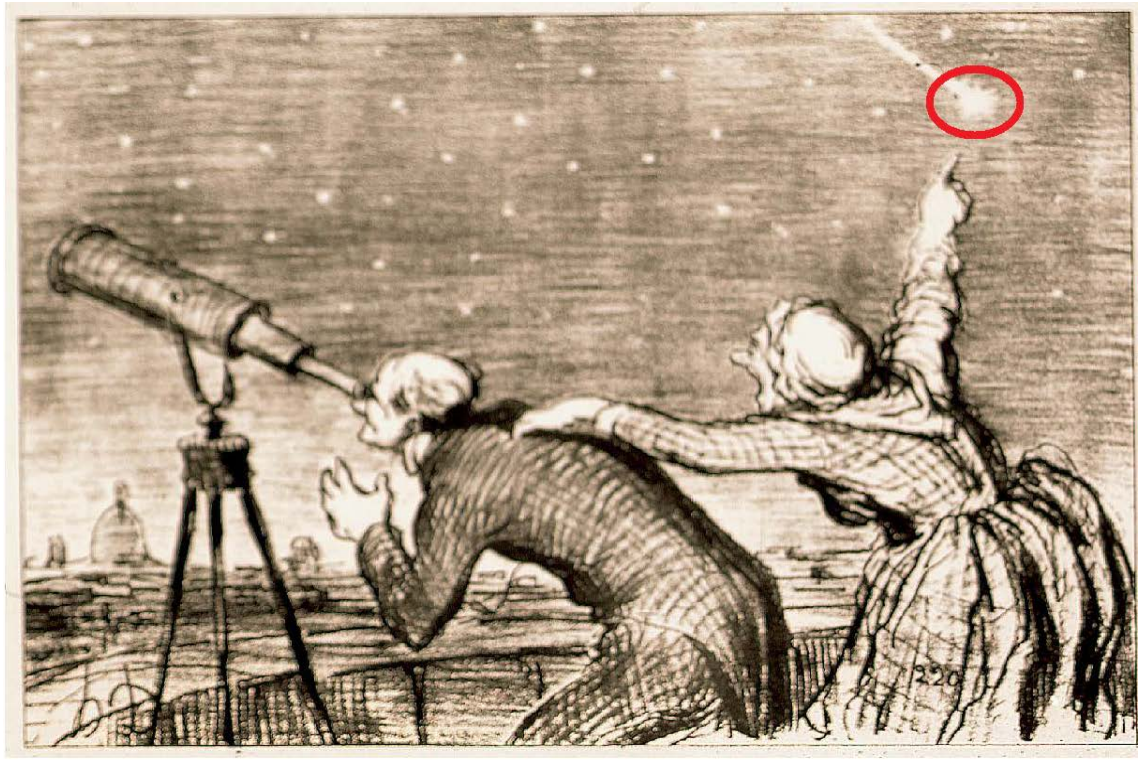


Figure source: https://web.fdi.ucm.es/posgrado/conferencias/AlexSanchez-slides.pdf

# Integrative Clustering

Collective understanding of both inter- and intra- relationships among several types of molecular assays is important.

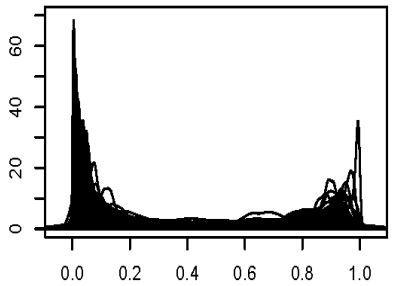
Example of relationships:

- Covariance structure within data
- Correlation between data

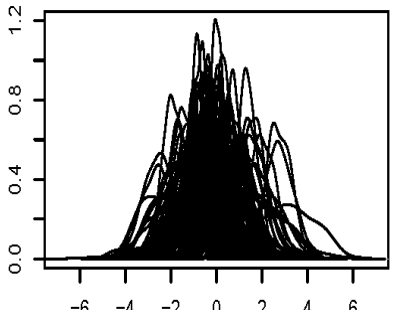
# Challenges of data integration

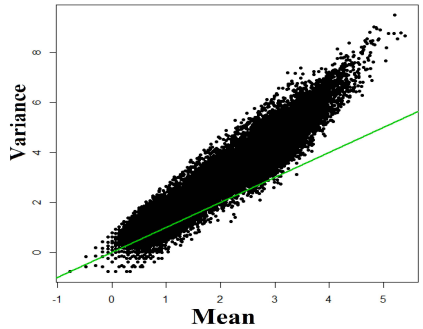
Distribution and scale of data varies:

1. **Mutation** : Present/Absent (1/0)
2. **Methylation**: Ranges (0,1)

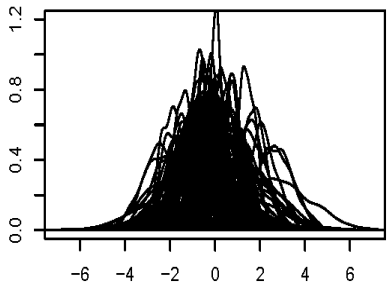



3. **Gene expression**:
   - Microarray: Ranges $(-\infty, \infty)$
   - Sequencing: Ranges $(0, \infty)$




4. **Protein** : Ranges $(-\infty, \infty)$



# Integrative Clustering Approaches

- Traditional:
  - Separate clustering of each data type at a time: Hierarchical clustering, k-means, NMF, model based clustering
  - Manual integration of results.
  - Inconclusive in the assignment of subjects to molecular cancer subtypes. Very subjective, non-reproducible.

# Integrative Clustering Approaches

- Traditional:
  - Separate clustering of each data type at a time: Hierarchical clustering, k-means, NMF, model based clustering
  - Manual integration of results.
  - Inconclusive in the assignment of subjects to molecular cancer subtypes. Very subjective, non-reproducible.
- Comprehensive:
  - iCluster (Shen et al. 2009)
  - Bayesian methods
  - Non-parametric methods: SNF, PINS

# Non-negative Matrix Factorization (NMF)

NMF is a non-parametric method.

Idea of NMF: Either a factor is present with a certain positive effect or it is absent simply having a zero effect.

- Pixels in digital image: Biomedical image processing
- Molecular concentration in bioinformatics (e.g. mRNA, protein, etc)
- Signal intensities in mass spectrometry: Computational Proteomics

# NMF, early works

- Paatero and Tapper (*Environmetrics*, 1994), "Positive Matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values"

  - Positivity-constrained least square algorithm (*Solving Least Squares Problems* by Lawson and Hanson)

# NMF, early works

- Paatero and Tapper (*Environmetrics*, 1994), "Positive Matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values"

  - Positivity-constrained least square algorithm (*Solving Least Squares Problems* by Lawson and Hanson)

- Lee and Seung (*Nature*, 1999) [1], "Algorithms for Non-negative Matrix Factorization"

# NMF, early works

- Paatero and Tapper (*Environmetrics*, 1994), "Positive Matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values"

  - Positivity-constrained least square algorithm (*Solving Least Squares Problems* by Lawson and Hanson)

- Lee and Seung (*Nature*, 1999) [1], "Algorithms for Non-negative Matrix Factorization"
- Brunnet et al. (*PNAS*, 2004) [2], "Metagenes and molecular pattern discovery using matrix factorization"

# Non-negative Matrix Factorization (NMF)

$$X_{n\times p} \approx W_{n\times k} H_{k\times p} \quad \text{s.t.} \quad W \geq 0 \quad H \geq 0$$

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ n \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & \cdots & p \end{array} \\ \begin{pmatrix} . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ . & . & \mathbf{X} & \cdots & . \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ . & . & . & \cdots & . \end{pmatrix} \end{array} \approx \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ n \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 1 & \cdots & k \end{array} \\ \begin{pmatrix} . & \cdots & . \\ . & \cdots & . \\ . & \mathbf{W} & . \\ \vdots & \cdots & \vdots \\ . & \cdots & . \end{pmatrix} \end{array} \times \begin{array}{c} \\ 1 \\ \vdots \\ k \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & \cdots & p \end{array} \\ \begin{pmatrix} . & . & . & \cdots & . \\ \vdots & \vdots & \mathbf{H} & \vdots & \vdots \\ . & . & . & \cdots & . \end{pmatrix} \end{array}$$

Columns of $W$ are the underlying basis vectors, i.e., each of the $n$ columns of $X$ can be built from $k$ columns of $W$.

Columns of $H$ give the weights associated with each basis vector

$$x_{.j} \approx \sum_{k=1}^{K} w_{.k} h_{kj} = W h_{.j}$$

# Application in image processing

Much of the appeal of NMF comes from its empirical success in learning meaningful features from a diverse collection of real-life data sets

**NMF** : $W$ basis and $H$ contain a large fraction of 0, so both the $W$ and $H$ are sparse.

**SVD** : Matrices contain positive and negative values, **Positive**,**Negative**

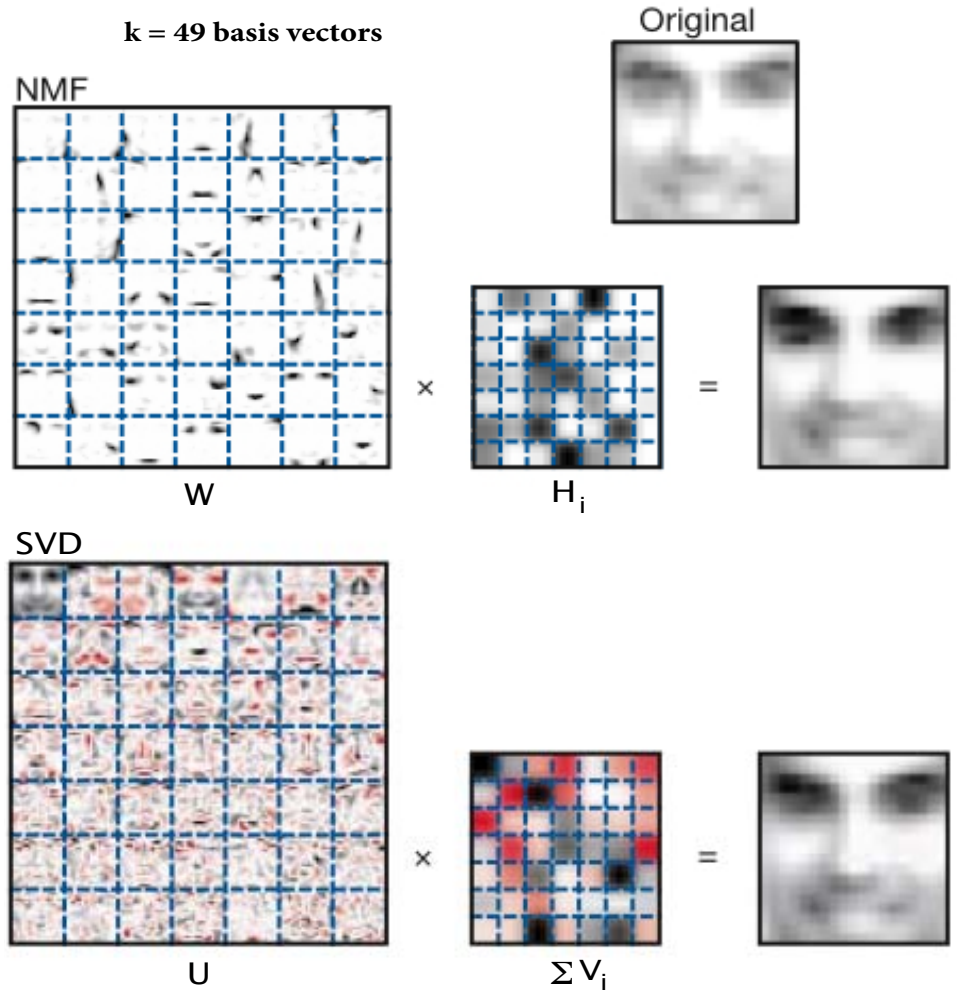



Figure source : Lee and Seung, Nature, Vol. 401, pp. 788-791, 1999 (Citation, 12632)

# Algorithm Provided by Lee and Seung (Multiplicative algorithm)

• Given non negative matrix $X$ and a desired rank $k$, NMF solves the following to estimate $W$ and $H$.

$$F(W, H) \quad = \quad \underset{W,H}{\operatorname{argmin}} \|X - WH\|^2$$

The function is convex in $W$ only (given $H$) or $H$ only (given $W$)

# Algorithm Provided by Lee and Seung (Multiplicative algorithm)

- Given non negative matrix $X$ and a desired rank $k$, NMF solves the following to estimate $W$ and $H$.

$$F(W, H) = \underset{W,H}{\operatorname{argmin}} \|X - WH\|^2$$

The function is convex in $W$ only (given $H$) or $H$ only (given $W$)

- Initialize with random matrices $W$ and $H$ and iterate until convergence

$$H_{\alpha\mu} = H_{\alpha\mu} \odot \frac{(W^T X)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}}$$

$$W_{i\alpha} = W_{i\alpha} \odot \frac{(XH^T)_{i\alpha}}{(WHH^T)_{i\alpha}}$$

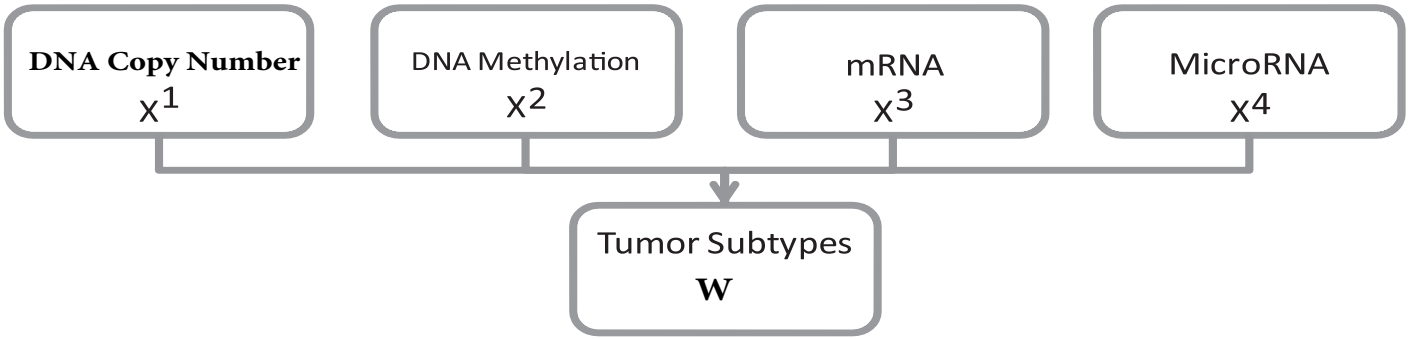
# Limitations:

Limitations: (*Gonzalez and Zhang (2005), Lin (2005) and Berry (2007)*)

- No rescue from 0.
- Non-convex problem. Slower rate of convergence

Moreover, integration of $n$ data sets requires $n + 1$ random initial matrices to begin with

# IntNMF framework




$$
\begin{aligned}
X^1_{n\times p_1} &\approx W_{n\times k} H^1_{k\times p_1} \\
X^2_{n\times p_2} &\approx W_{n\times k} H^2_{k\times p_2} \\
X^3_{n\times p_3} &\approx W_{n\times k} H^3_{k\times p_3} \\
X^4_{n\times p_4} &\approx W_{n\times k} H^4_{k\times p_4}
\end{aligned}
$$

$W$ = Common basis matrix that connects the 4 sets of data, inducing dependencies

$H^1, H^2, H^3$ and $H^4$ are coefficient matrices

# Proposed algorithm: Integrative NMF (intNMF) using ALS

Objective function: $F = \sum_{i=1}^{m} \|X_i - WH_i\|^2$

- Initialize $W$ from $U[0, 1]$ or use Non negative double singular value decomposition (NNDSVD) method proposed by *Boutsidis and Gallopoulos (2008)*

# Proposed algorithm: Integrative NMF (intNMF) using ALS

Objective function: $F = \sum_{i=1}^{m} \|X_i - WH_i\|^2$

- Initialize $W$ from $U[0, 1]$ or use Non negative double singular value decomposition (NNDSVD) method proposed by *Boutsidis and Gallopoulos (2008)*
- Fix $W$ and solve for $H^i, i = 1, 2, \cdots m$ (*Lawson et al. (1974), Bro et al. (1997) and Bentham et al (2004)*)

$$\underset{H^i}{\operatorname{argmin}} \|X^i - WH^i\|^2 \quad s.t. \quad H_i \geq 0$$

# Proposed algorithm: Integrative NMF (intNMF) using ALS

Objective function: $F = \sum_{i=1}^{m} \|X_i - WH_i\|^2$

- Initialize $W$ from $U[0,1]$ or use Non negative double singular value decomposition (NNDSVD) method proposed by *Boutsidis and Gallopoulos (2008)*
- Fix $W$ and solve for $H^i, i = 1, 2, \cdots m$ (*Lawson et al. (1974), Bro et al. (1997) and Bentham et al (2004)*)

$$\underset{H^i}{\operatorname{argmin}} \|X^i - WH^i\|^2 \quad s.t. \quad H_i \geq 0$$

- Fix $H_i$ and solve for common $W$

$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{m} \|X^i - WH^i\|^2 \quad s.t. \quad W \geq 0$$

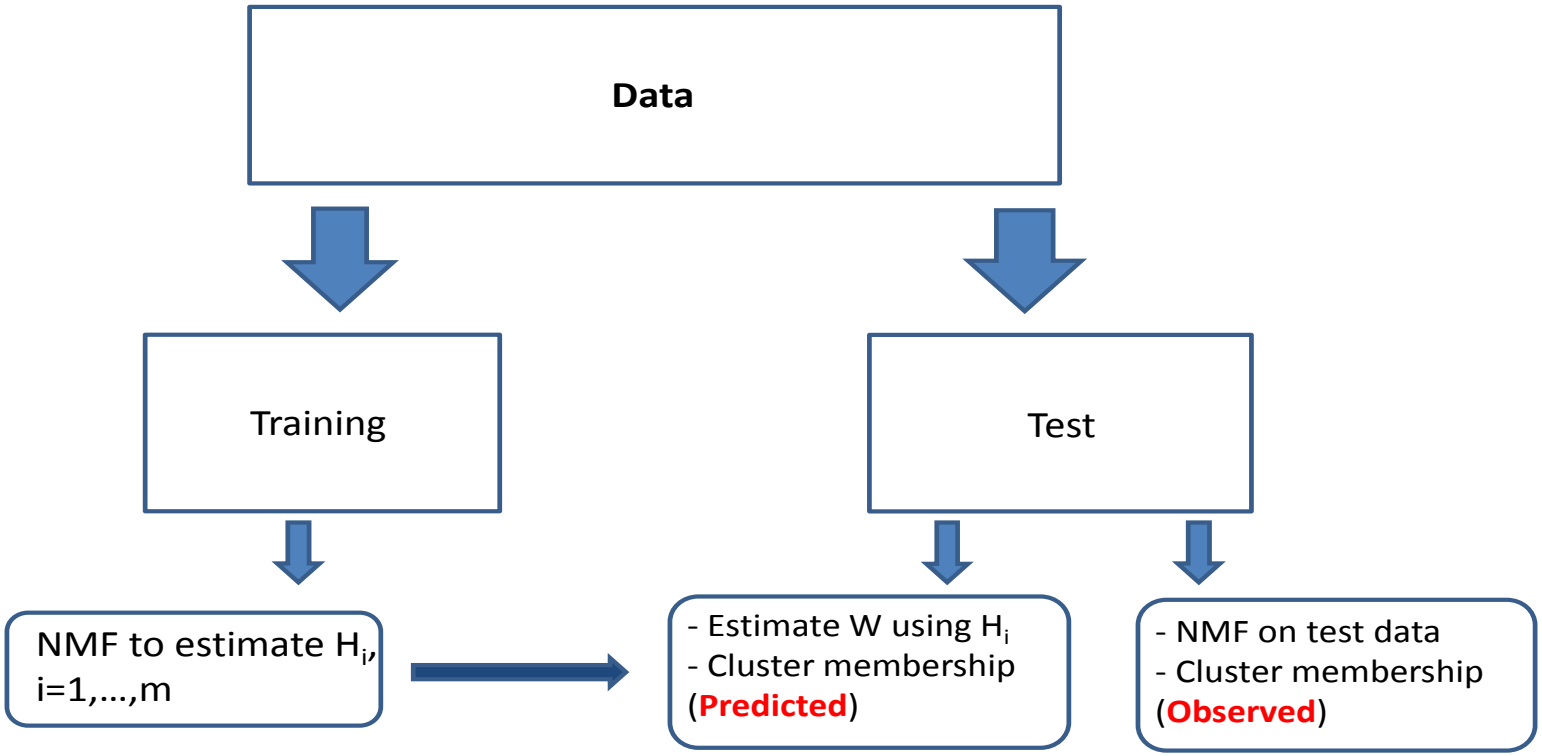
# Two steps

Any clustering method involves

1. Identification of optimum number of clusters $k$
2. Assignment of cluster membership to samples

# Optimum number of clusters (k)

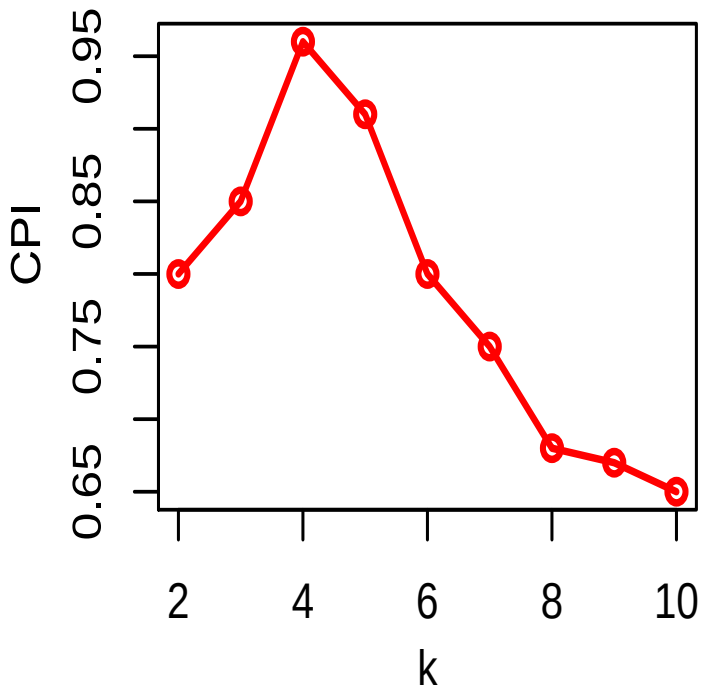
**Cluster Prediction Index (CPI)**: Cross validation technic in estimating optimum number of clusters $k$.




CPI = Adjusted Rand Index

# Optimum number of clusters (k)

**Cluster Prediction Index (CPI) va Clusters**:



# How NMF works in clustering application

Brunet et al (2004) "Metagenes and molecular pattern discovery using matrix factorization"

**Example**: $X_{4\times5} \approx W_{4\times3}H_{3\times5}, \quad n = 4,\ p = 5,\ k = 3$

$$\begin{pmatrix} 5 & 1 & 0 & 2 & 0 \\ 1 & 0 & 3 & 1 & 2 \\ 2 & 1 & 3 & 3 & 1 \\ 3 & 1 & 0 & 4 & 3 \end{pmatrix} \approx \begin{array}{c} \\ \text{Sample1} \\ \text{Sample2} \\ \text{Sample3} \\ \text{Sample4} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Clust1} & \text{Clust2} & \text{Clust3} \end{array} \\ \begin{pmatrix} 4.55 & 0.00 & 0.00 \\ 0.00 & 6.23 & 1.07 \\ 2.37 & 5.60 & 0.56 \\ 3.24 & 0.00 & 4.77 \end{pmatrix} \end{array} \times \begin{pmatrix} 0.93 & 0.30 & 0.00 & 0.53 & 0.00 \\ 0.12 & 0.00 & 0.51 & 0.14 & 0.16 \\ 0.02 & 0.00 & 0.00 & 0.46 & 0.64 \end{pmatrix}$$

Matrix $W$ is used for the cluster membership. Each sample is placed into a cluster corresponding to the largest value in the column of that sample. i.e. sample $j$ is placed in cluster $i$ if $w_{ij}$ is the largest entry in column $j$.

- Sample 1 $\Rightarrow$ Cluster 1
- Sample 2 $\Rightarrow$ Cluster 2
- Sample 3 $\Rightarrow$ Cluster 2
- Sample 4 $\Rightarrow$ Cluster 3

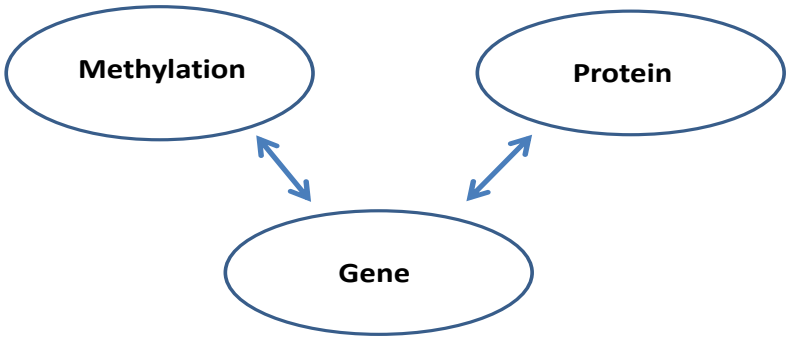
# IntNMF Method Assessment

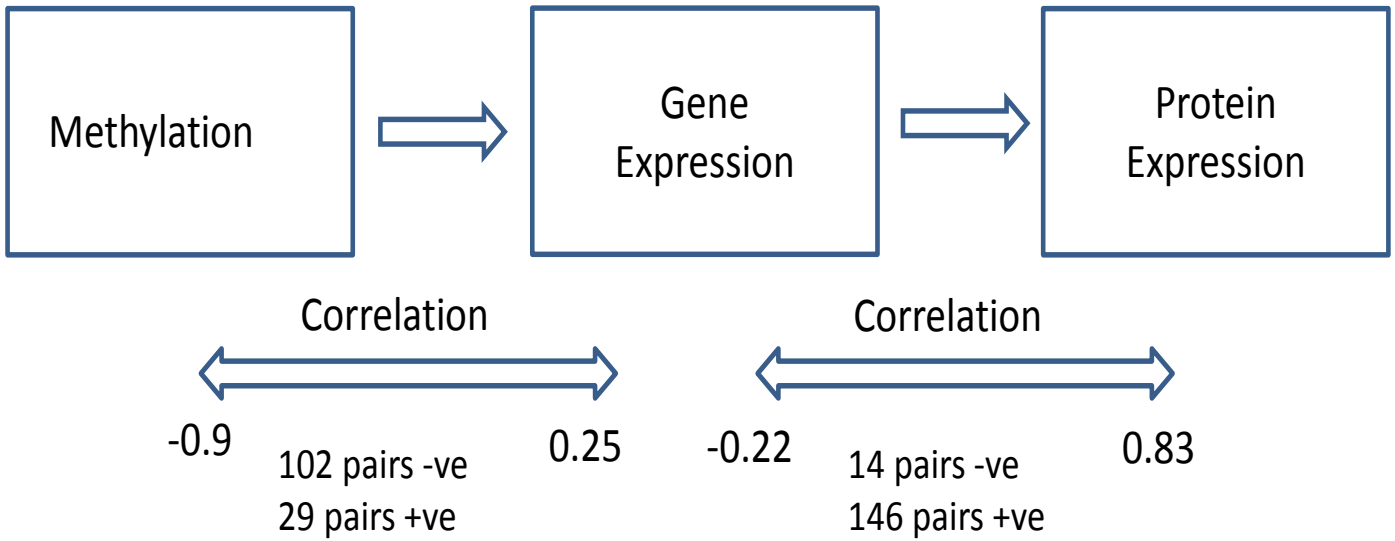
- Simulation Study
- Application to TCGA studies on Glioblastoma data

# Simulation Study

Three data on 384 subjects

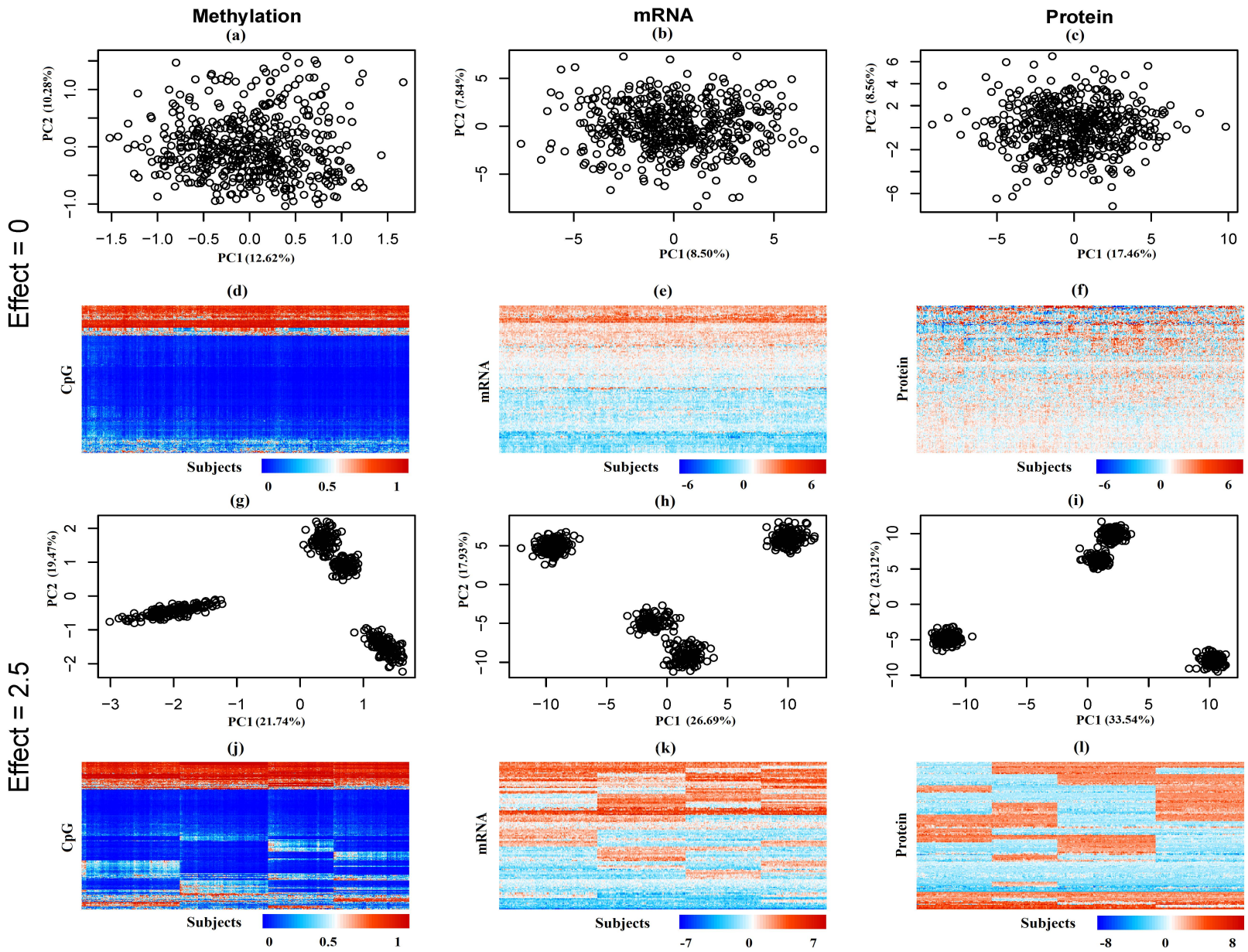
- Methylation: 367 CpGs
- Gene expression: 131 Genes
- Protein expression: 160 Protein




R package "**InterSIM**":
https://cran.rstudio.com/web/packages/InterSIM/index.html

# Simulated data with 0 and 4 groups



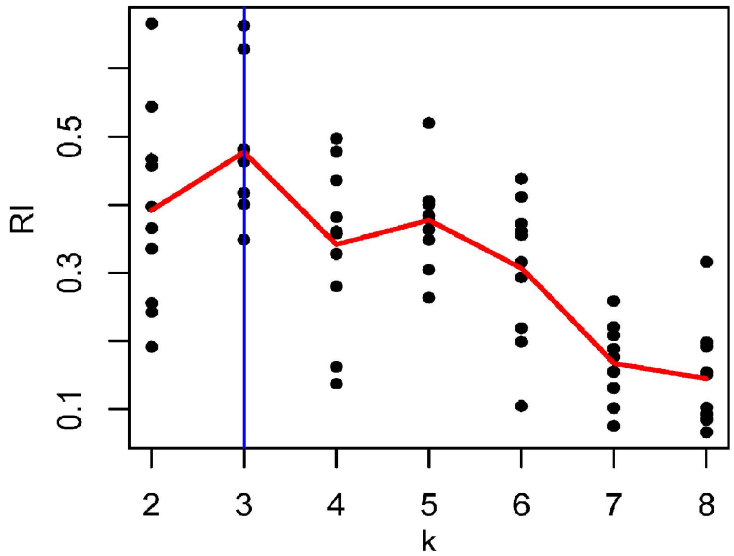
# Application - Example 1:TCGA studies on Glioblastoma

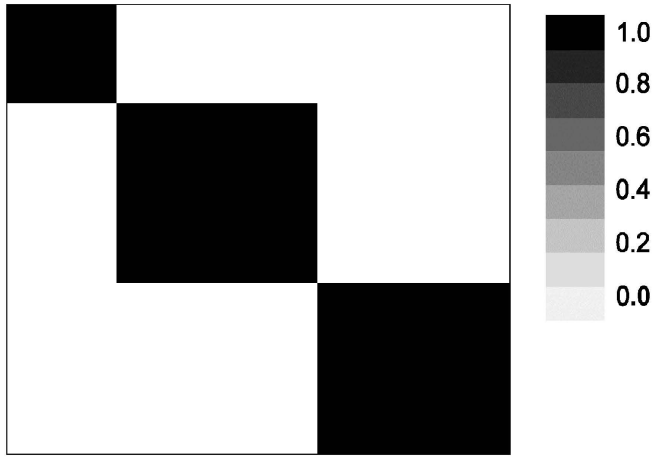
Publicly available data on 55 cancer tissues from The Cancer Genome Atlas (TCGA). https://gdc.cancer.gov/

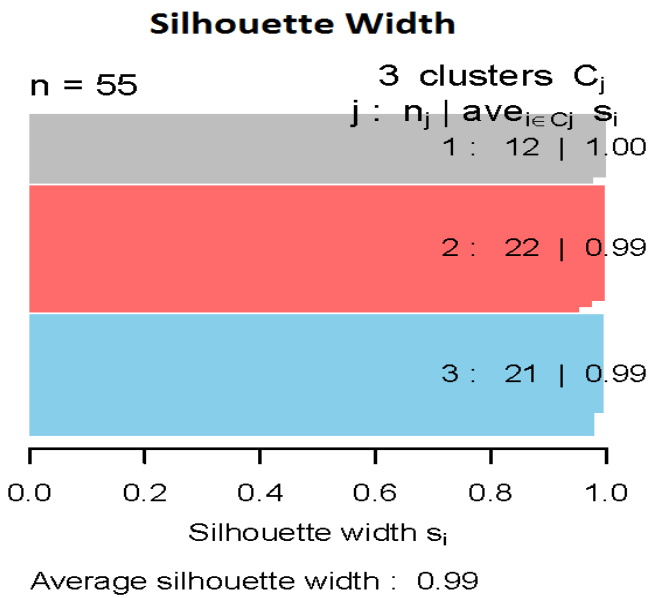
- CNV : 1599 genes
- Methylation : 1515 CpGs
- mRNA : 1740 genes

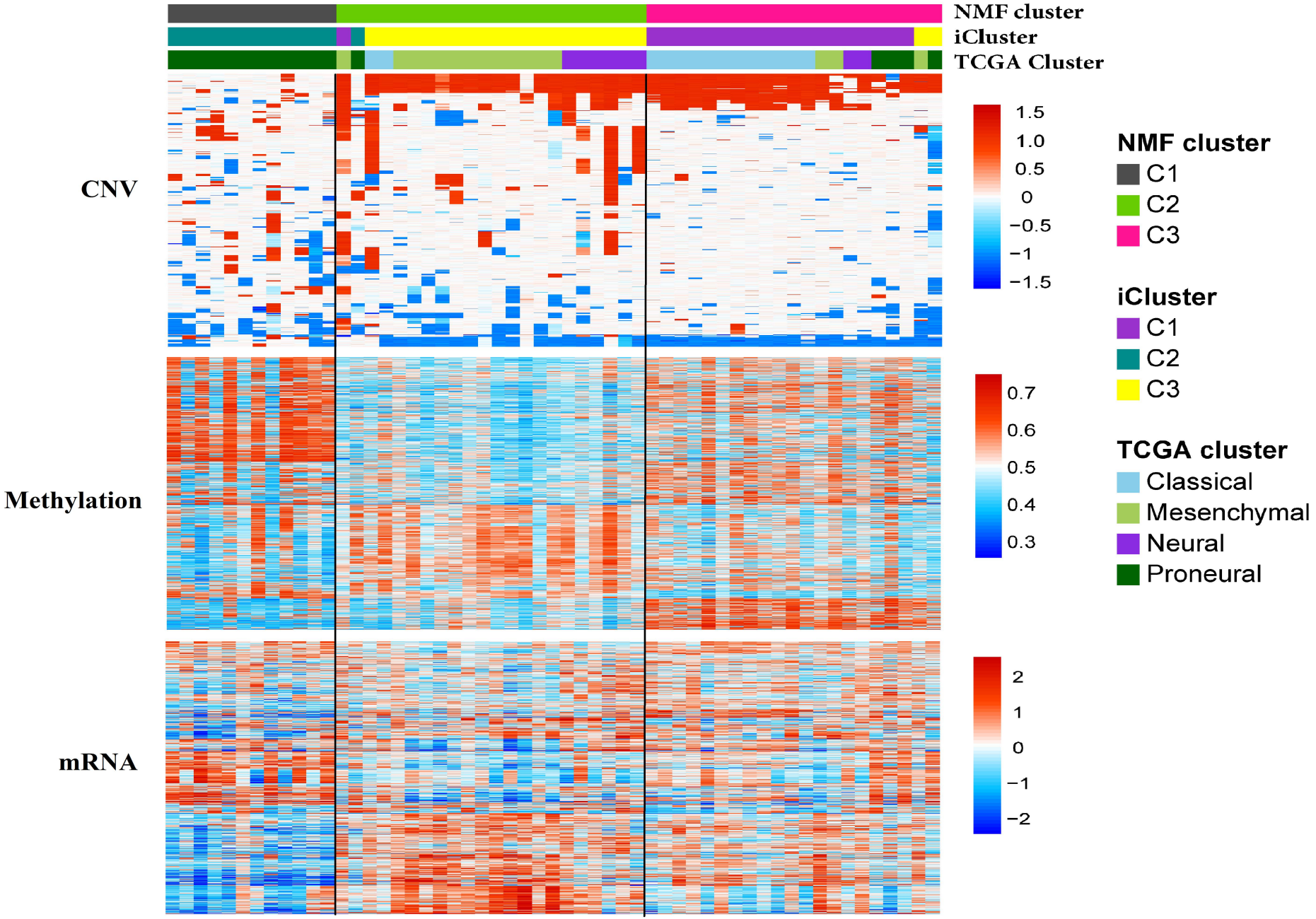
TCGA data portal: https://tcga-data.nci.nih.gov/tcga/

# Optimum $k$


RI
0.1
0.3
0.5
2
3
4
5
6
7
8
k


1.0
0.8
0.6
0.4
0.2
0.0


Silhouette Width
n = 55
3 clusters $C_j$
j : $n_j$ | $ave_{i \in Cj}$ $s_i$
1 : 12 | 1.00
2 : 22 | 0.99
3 : 21 | 0.99
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width $s_i$
Average silhouette width : 0.99


NMF cluster
iCluster
TCGA Cluster
CNV
Methylation
mRNA
1.5
1.0
0.5
0
−0.5
−1
−1.5
0.7
0.6
0.5
0.4
0.3
2
1
0
−1
−2
NMF cluster
C1
C2
C3
iCluster
C1
C2
C3
TCGA cluster
Classical
Mesenchymal
Neural
Proneural

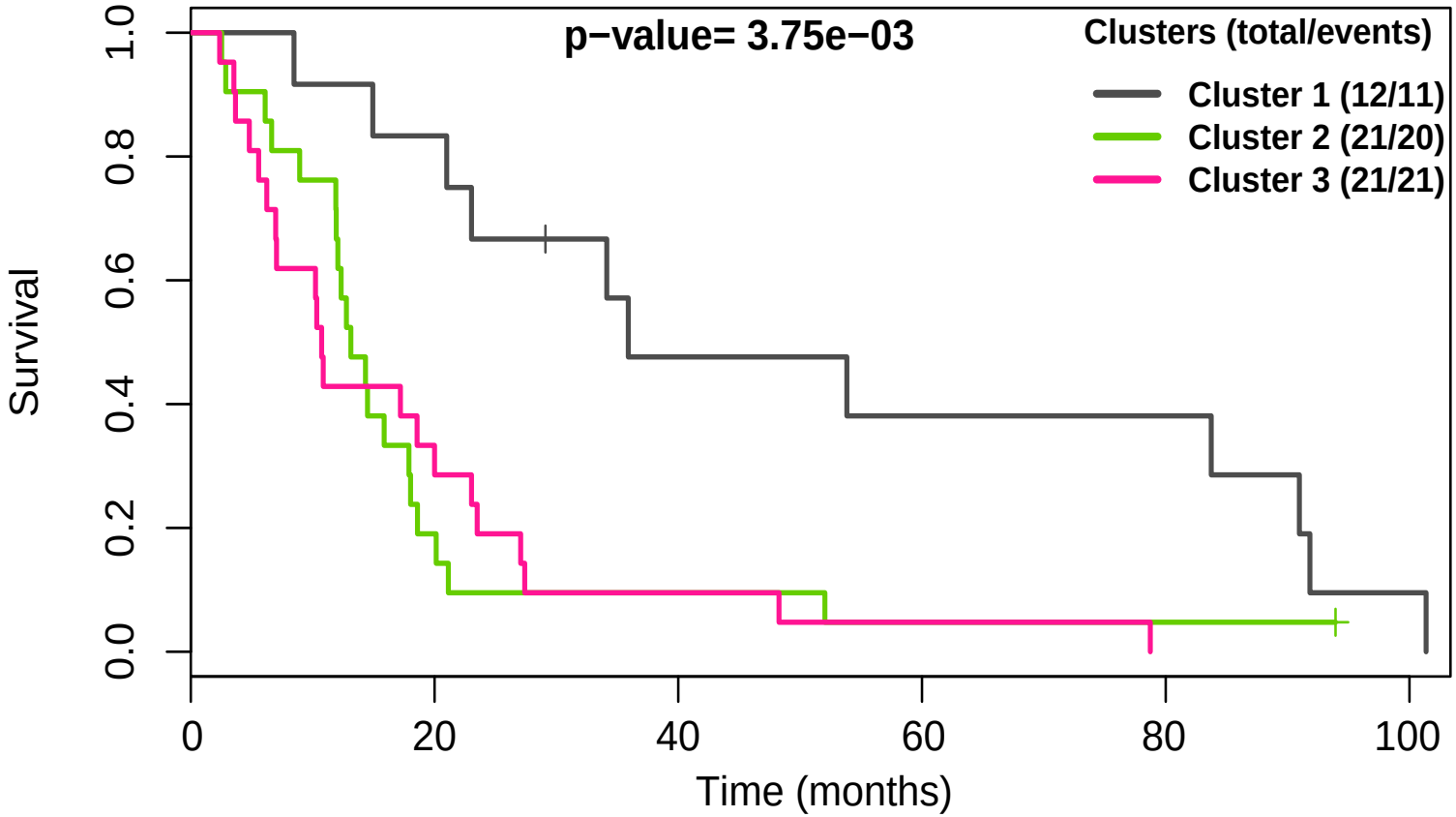
# Clustering results

Cross tabulation of Integrative NMF cluster subtypes

1. Expression subtype (Verhaak 2010)
2. iCluster subtype (Shen et. al)

| | | Integrative NMF | | | |
|---|---|---|---|---|---|
| | | C1 | C2 | C3 | Total |
| Expression Subtype | Classical | 0 | 2 | **12** | 14 |
| | Messenchymal | 0 | **13** | 3 | 16 |
| | Neural | 0 | 6 | 2 | 8 |
| | Proneural | **12** | 1 | 4 | 17 |
| | Total | 12 | 22 | 21 | 55 |
| iCluster | C1 | 0 | 1 | **19** | 20 |
| | C2 | **12** | 1 | 0 | 13 |
| | C3 | 0 | **20** | 2 | 22 |
| | Total | 12 | 22 | 21 | 55 |

# Kaplan Meier Survival Curves



**Thank you!**


Lee D and Seung HS.
Algorithms for Non-negative Matrix Factorization.
*Adv Neural Inform Process*, 13:556–562, 2001.


Brunet JP, Tamayo P, Golub TR, and Mesirov JP.
Metagenes and molecular pattern discovery using matrix factorization.
*PNAS*, 101:4164–4169, 2004.