Data Integration and Cancer Subtyping

Tin Nguyen

Dept. of Computer Science & Engineering
College of Engineering
University of Nevada, Reno

Background

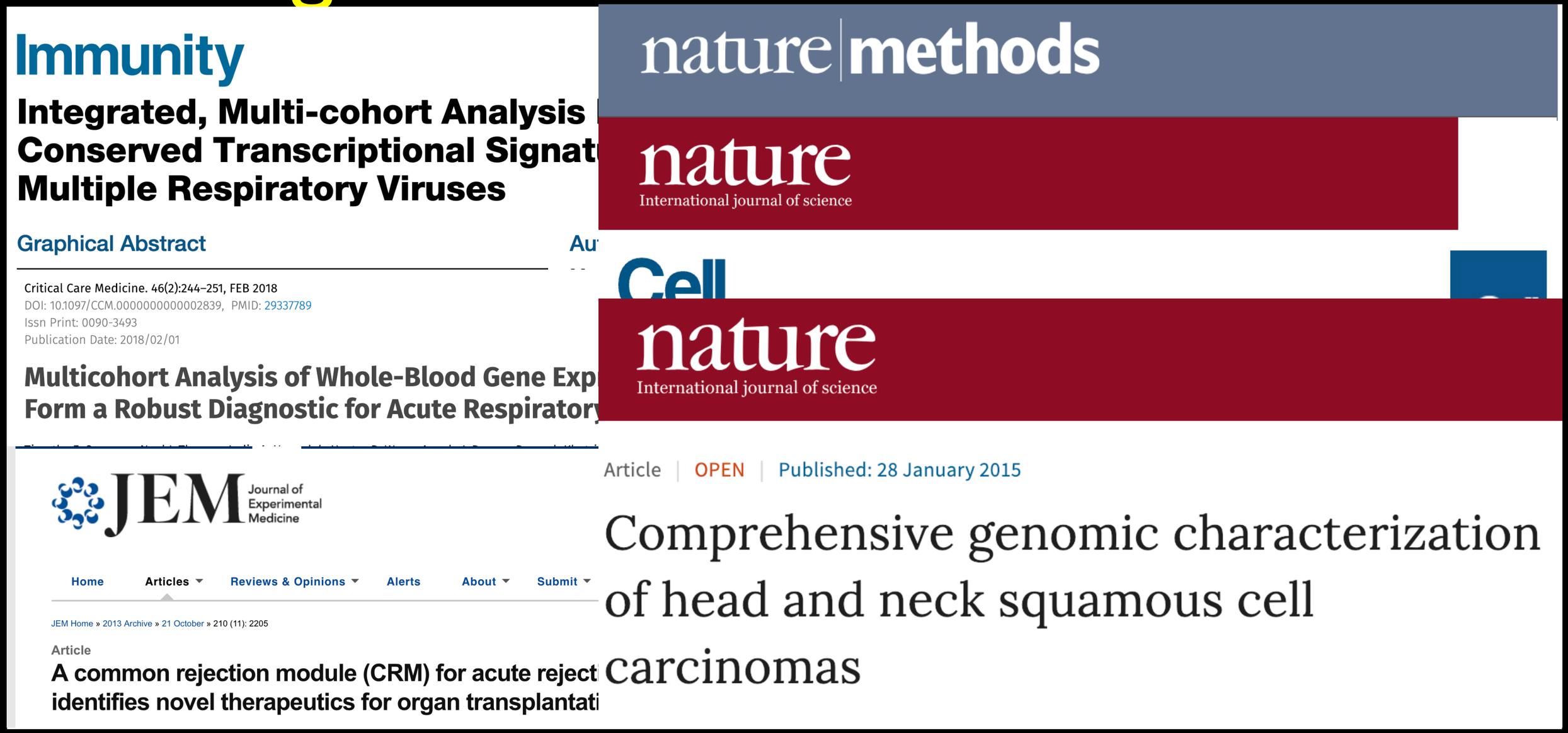
Experience

- MS & BS, 2008, Eotvos Lorand University, Budapest, Hungary
  - Software Developer, Budapest, Bolzano, Oulu, …
- Ph.D., May 2017, Wayne State University, Detroit, Michigan
  - Advisor: Sorin Draghici
- Assistant Professor, July 2017 – present, UNR, Computer Science and Engineering

Research Interest: bioinformatics

- Cancer subtyping, pathway analysis, single-cell RNA sequencing

Integration of bio-molecular data



Comprehensive literature review and statistical considerations for microarray meta-analysis, Tseng et. al., NAR, 2012

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Effect of Three Decades of Screening Mammography on Breast-Cancer Incidence

Archie Bleyer, M.D., and H. Gilbert Welch, M.D., M.P.H.

ABSTRACT

BACKGROUND

To reduce mortality, screening must detect life-threatening disease at an earlier, more curable stage. Effective cancer-screening programs therefore both increase the incidence of cancer detected at an early stage and decrease the incidence of cancer presenting at a late stage.

METHODS

We used Surveillance, Epidemiology, and End Results data to examine trends from 1976 through 2008 in the incidence of early-stage breast cancer (ductal carcinoma in situ and localized disease) and late-stage breast cancer (regional and distant disease) among women 40 years of age or older.

RESULTS

The introduction of screening mammography in the United States has been associated with a doubling in the number of cases of early-stage breast cancer that are detected each year, from 112 to 234 cases per 100,000 women — an absolute increase of 122 cases per 100,000 women. Concomitantly, the rate at which women present with late-stage cancer has decreased by 8%, from 102 to 94 cases per 100,000 women — an absolute decrease of 8 cases per 100,000 women. With the assumption of a constant underlying disease burden, only 8 of the 122 additional early-stage cancers diagnosed were expected to progress to advanced disease. After excluding the transient excess incidence associated with hormone-replacement therapy and adjusting for trends in the incidence of breast cancer among women younger than 40 years of age, we estimated that breast cancer was overdiagnosed (i.e., tumors were detected on screening that would never have led to clinical symptoms) in 1.3 million U.S. women in the past 30 years. We estimated that in 2008, breast cancer was overdiagnosed in more than 70,000 women; this accounted for 31% of all breast cancers diagnosed.

The Opinion Pages | OP-ED CONTRIBUTOR

Cancer Survivor or Victim of Overdiagnosis?

By H. GILBERT WELCH NOV. 21, 2012

Hanover, N.H.

FOR decades women have been told that one of the most important things they can do to protect their health is to have regular mammograms. But over the past few years, it's become increasingly clear that these screenings are not all they're cracked up to be. The latest piece of evidence appears in a study in Wednesday's New England Journal of Medicine, conducted by the oncologist Archie Bleyer and me.

The study looks at the big picture, the effect of three decades of mammography screening in the United States. After correcting for underlying trends and the use of hormone replacement therapy, we found that the introduction of screening has been associated with about 1.5 million additional women receiving a diagnosis of early stage breast cancer.

That would be a good thing if it meant that 1.5 million *fewer* women had gotten a diagnosis of late-stage breast cancer. Then we could say that screening had advanced the time of diagnosis and provided the opportunity of reduced mortality for 1.5 million women.

But instead, we found that there were only around 0.1 million fewer women with a diagnosis of late-stage breast cancer. This discrepancy means there was a lot of overdiagnosis: more than a million women who were told they had early stage cancer — most of whom underwent surgery, chemotherapy or radiation — for a "cancer" that was never going to make them sick. Although it's impossible to know which women these are, that's some pretty serious harm.

Some facts

- During the past 30 years, there were 1.5 million women diagnosed with breast cancer (early diagnosis)
- During the same time the late stage diagnoses of breast cancer were reduced by only 100,000
- More than 1 million women went through surgery, chemotherapy or radiation for a disease that was never going to make them sick
- This cost $23,000/patient for a societal cost of $32.2 billion!!!
- The bottom line: many people receive unnecessary treatment

More facts

- The standard of care for stage I non-small cell lung cancer is surgical resection
- Adjuvant therapy (chemotherapy or radiation) is NOT usually recommended (because clinical trials have shown little statistical improvement in survival)
- However, about 42% of the patients will see a disease recurrence and will eventually die
- The bottom line: many people die because they do not receive necessary treatment

The problem summary

- We are currently unable to distinguish between subgroups of patients (respondent vs. non-respondents) and/or subtypes of disease (aggressive vs. non-aggressive).
- Many attempts to achieve this based solely on gene expression signatures have been undertaken but yielded only modest success so far (very few gene expression tests are FDA-approved as of yet).
- Our hypothesis is that a given disease subtype can be triggered by a number of different events, that may happen at different levels (mRNA, miRNA, epigenetics, etc.).
- Hence, integrating multiple types of data in a single analysis is expected to better distinguish between subtypes of various diseases

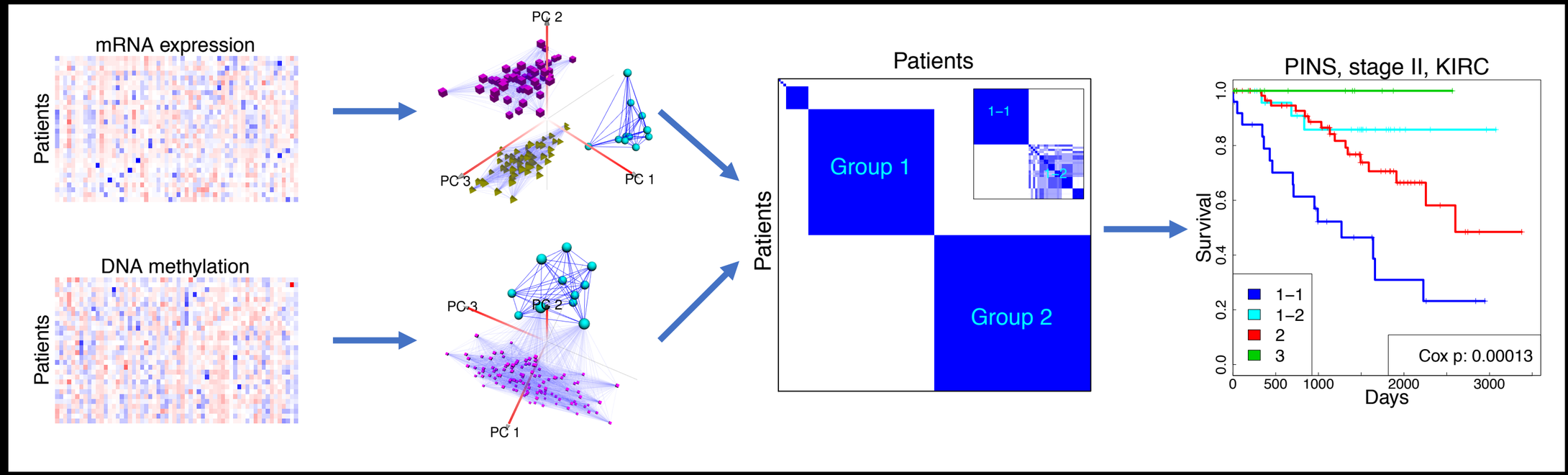
Some state-of-the-art methods

- CC: Consensus Clustering (Monti et al., 2004, Machine Learning)
  - From the Broad Institute
  - 1605 + 782 + … citations on Google Scholar
  - Used in most recent papers in Nature, Science, Cell, and Cancer Cell

- iClusterPlus (Mo et al., 2013, PNAS)
  - 276 Google Scholar citations

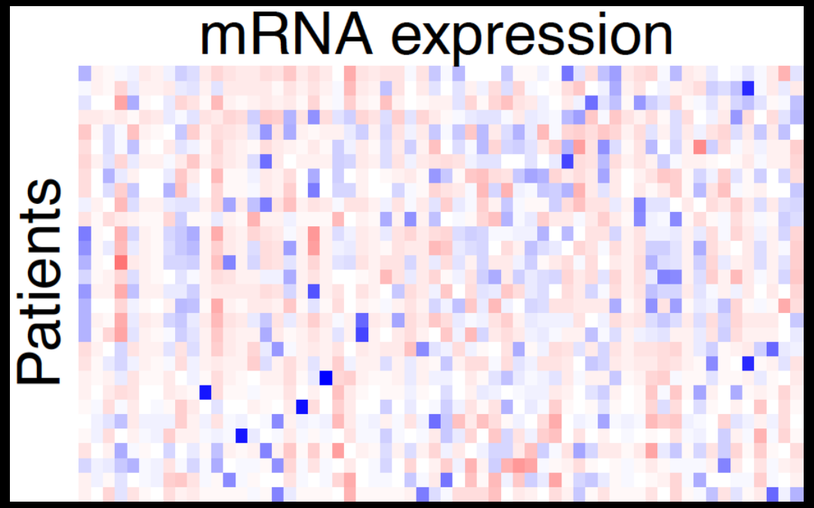
- SNF: Similarity Network Fusion (Wang et al., 2014, Nature Methods)
  - 820 citations on Google Scholar

Disease subtyping using omics data



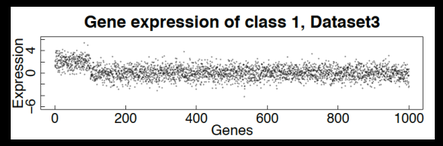
- Data: $\mathbb{D} = \{D_1, \ldots, D_T\}, D_i \in R^{\{N \times M_i\}}$ (N patients and T data types)
- Goal: grouping the patients
- Algorithm I: clustering each data type D_i ($i \in [1, T]$)
- Algorithm II: data integration (T data types)

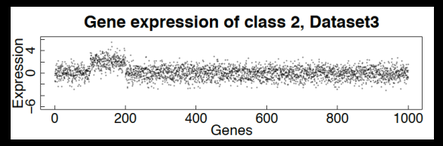
Algorithm I: perturbation clustering

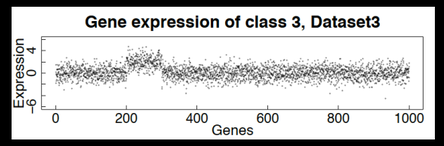


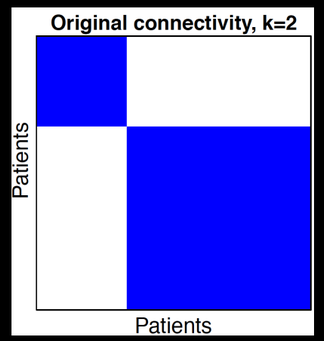
- Small changes in any kind of quantitative assay will be inherently present between individuals, even in a truly homogeneous population
- We are not interested in clusters that form or disappear due to small changes in the data
- Well-defined subtypes have to be stable with respect to data perturbation

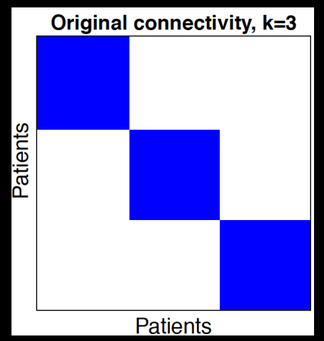
Algorithm I: perturbation clustering

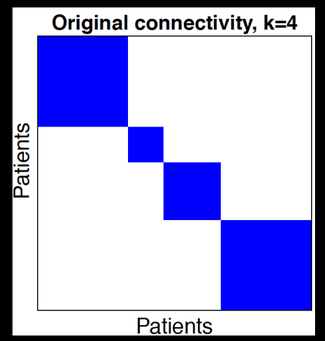


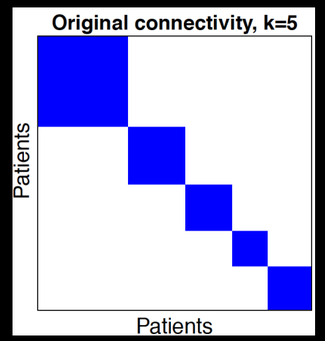


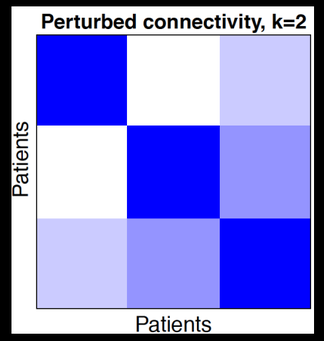


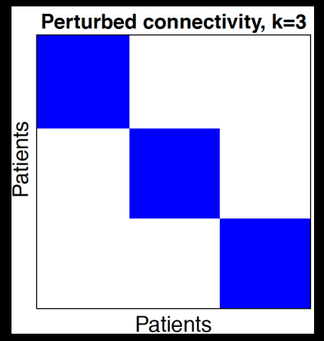


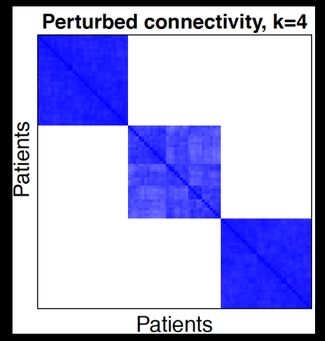


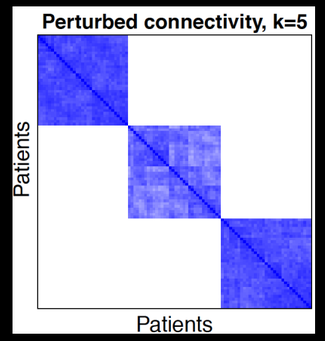


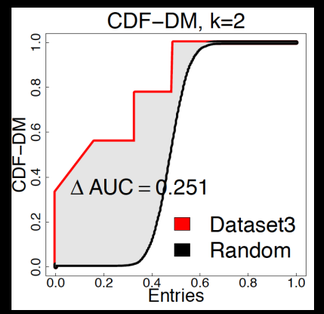


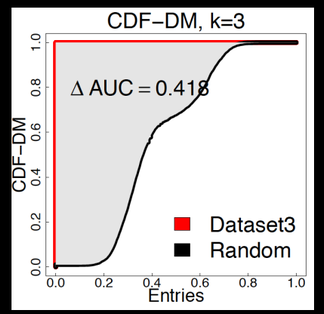


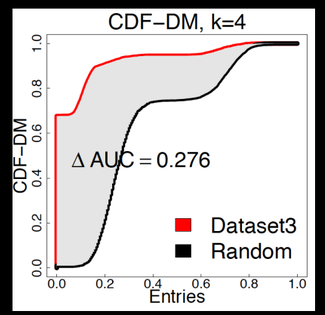


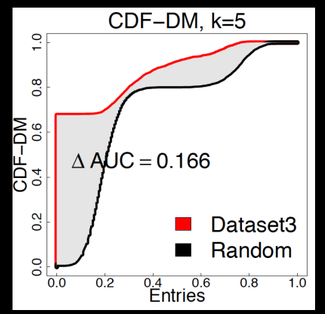


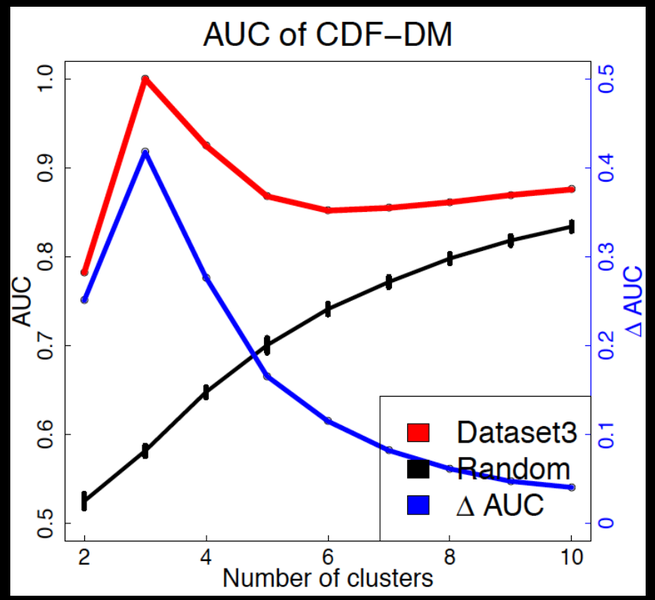






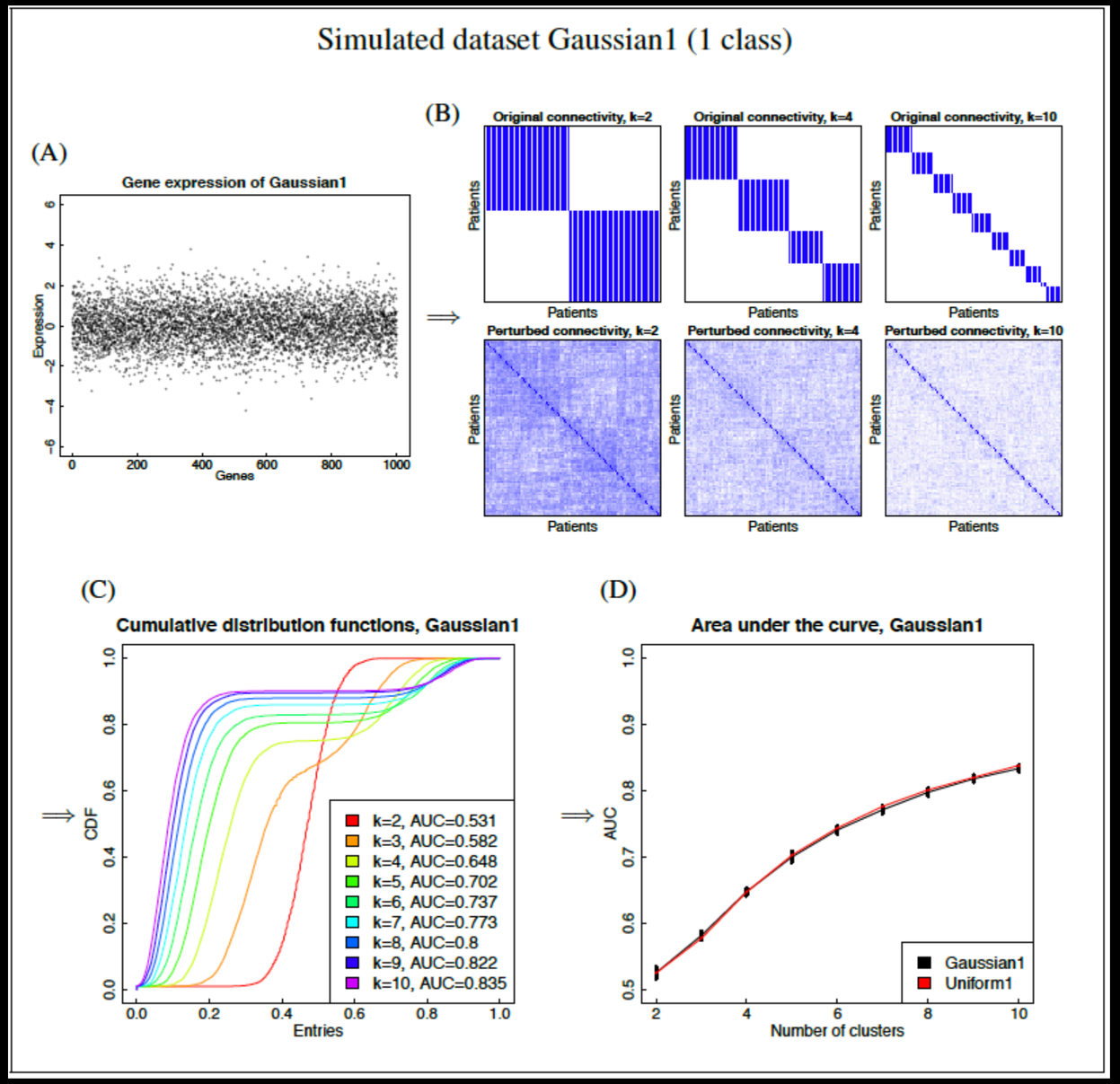


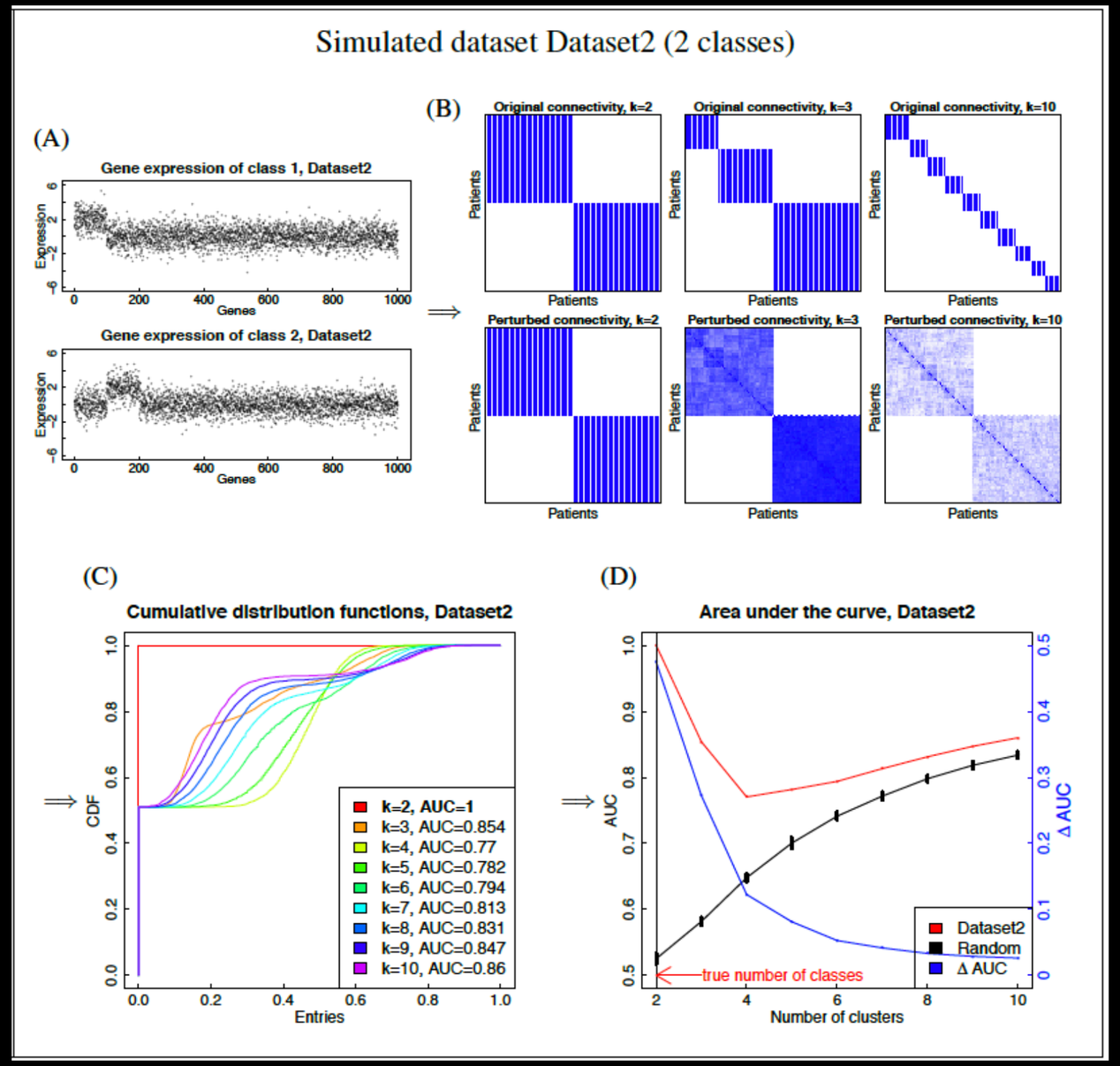




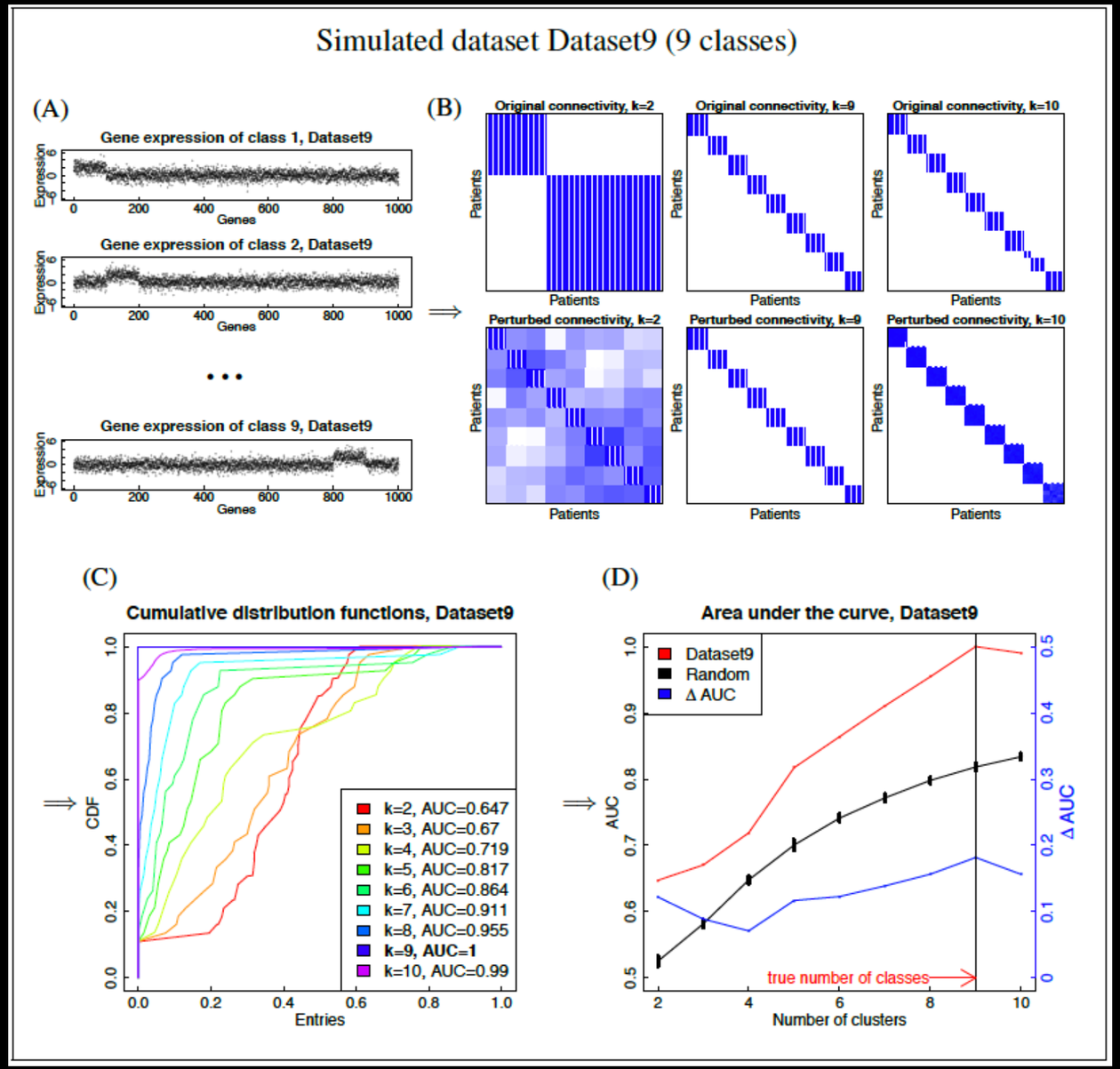
- True structure recovers when the data is perturbed
- AUC, △AUC, connectivity → true subtypes
- Results are consistent regardless of the clustering method (k-means, hc, pam)

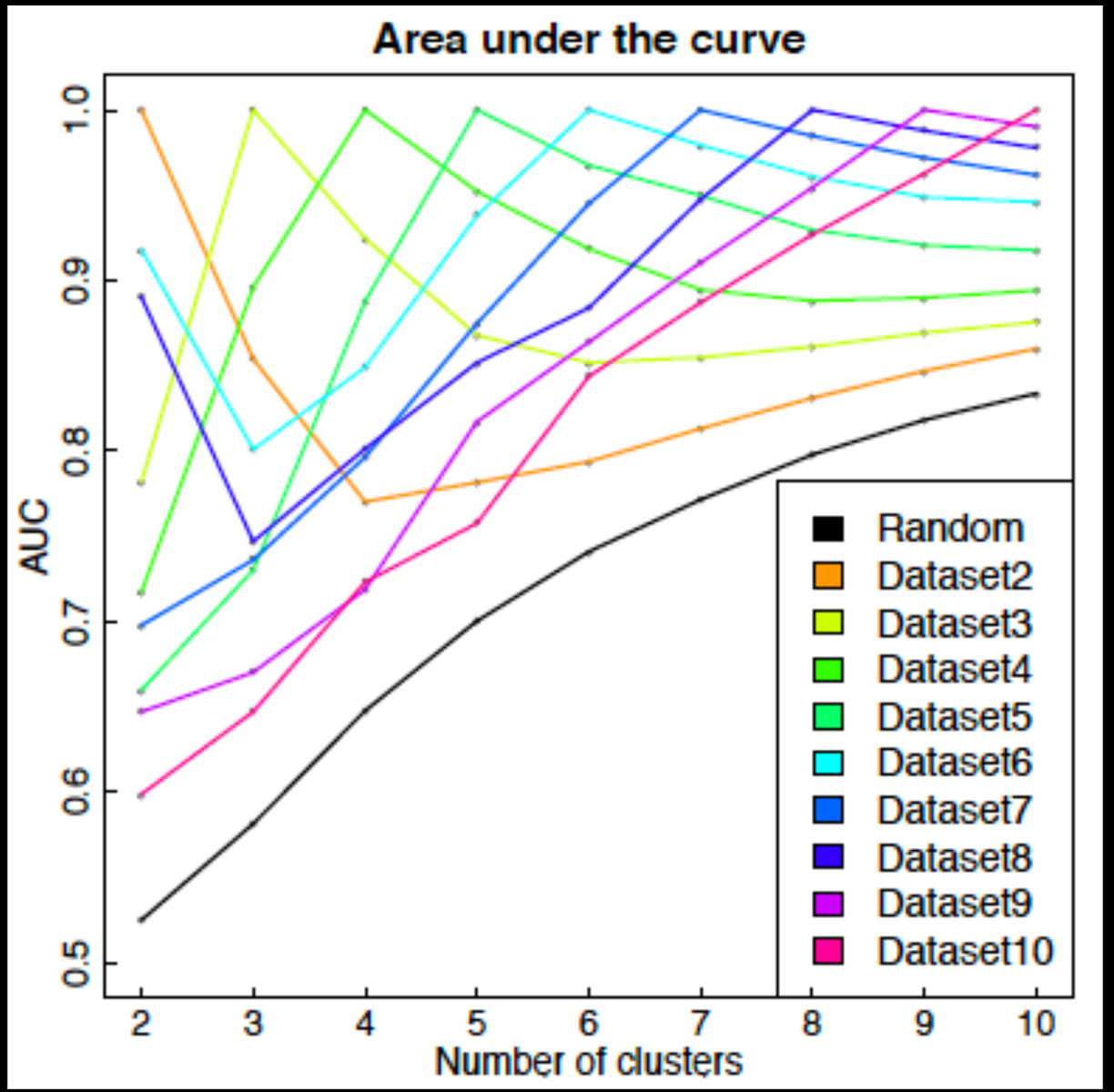
Simulation


Simulated dataset Gaussian1 (1 class)
(A)
Gene expression of Gaussian1
Expression
Genes
(B)
Original connectivity, k=2
Original connectivity, k=4
Original connectivity, k=10
Perturbed connectivity, k=2
Perturbed connectivity, k=4
Perturbed connectivity, k=10
Patients
(C)
Cumulative distribution functions, Gaussian1
CDF
Entries
k=2, AUC=0.531
k=3, AUC=0.582
k=4, AUC=0.648
k=5, AUC=0.702
k=6, AUC=0.737
k=7, AUC=0.773
k=8, AUC=0.8
k=9, AUC=0.822
k=10, AUC=0.835
(D)
Area under the curve, Gaussian1
AUC
Number of clusters
Gaussian1
Uniform1

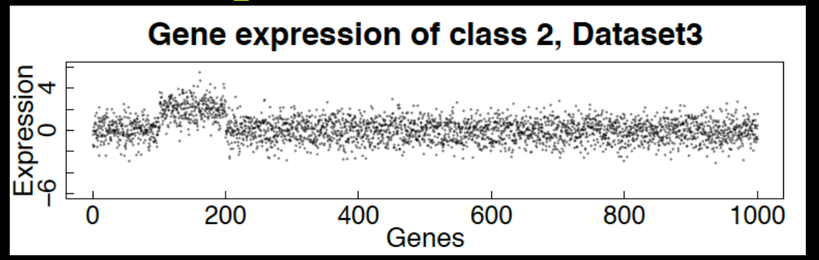

Simulated dataset Dataset2 (2 classes)
(A)
Gene expression of class 1, Dataset2
Gene expression of class 2, Dataset2
Expression
Genes
(B)
Original connectivity, k=2
Original connectivity, k=3
Original connectivity, k=10
Perturbed connectivity, k=2
Perturbed connectivity, k=3
Perturbed connectivity, k=10
Patients
(C)
Cumulative distribution functions, Dataset2
CDF
Entries
k=2, AUC=1
k=3, AUC=0.854
k=4, AUC=0.77
k=5, AUC=0.782
k=6, AUC=0.794
k=7, AUC=0.813
k=8, AUC=0.831
k=9, AUC=0.847
k=10, AUC=0.86
(D)
Area under the curve, Dataset2
AUC
Δ AUC
Number of clusters
Dataset2
Random
Δ AUC
true number of classes

Simulation (cont.)

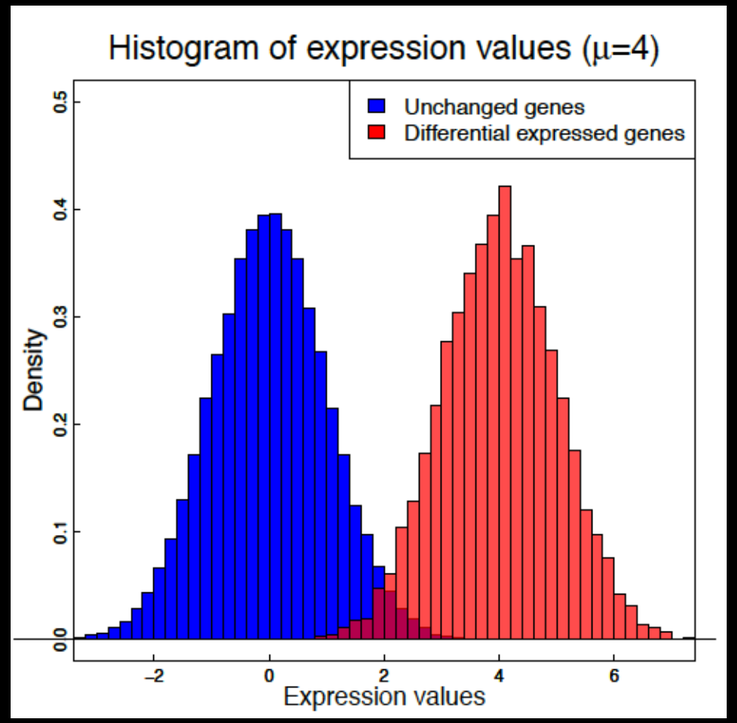


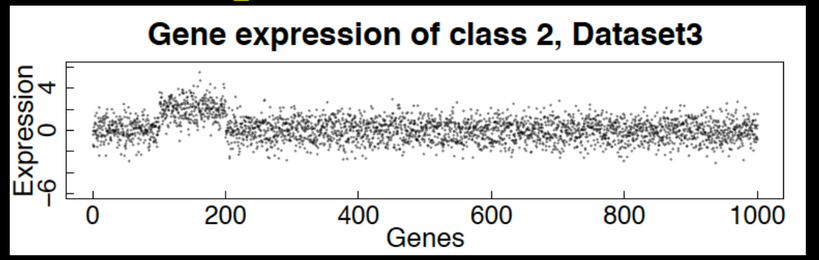


How about stability?

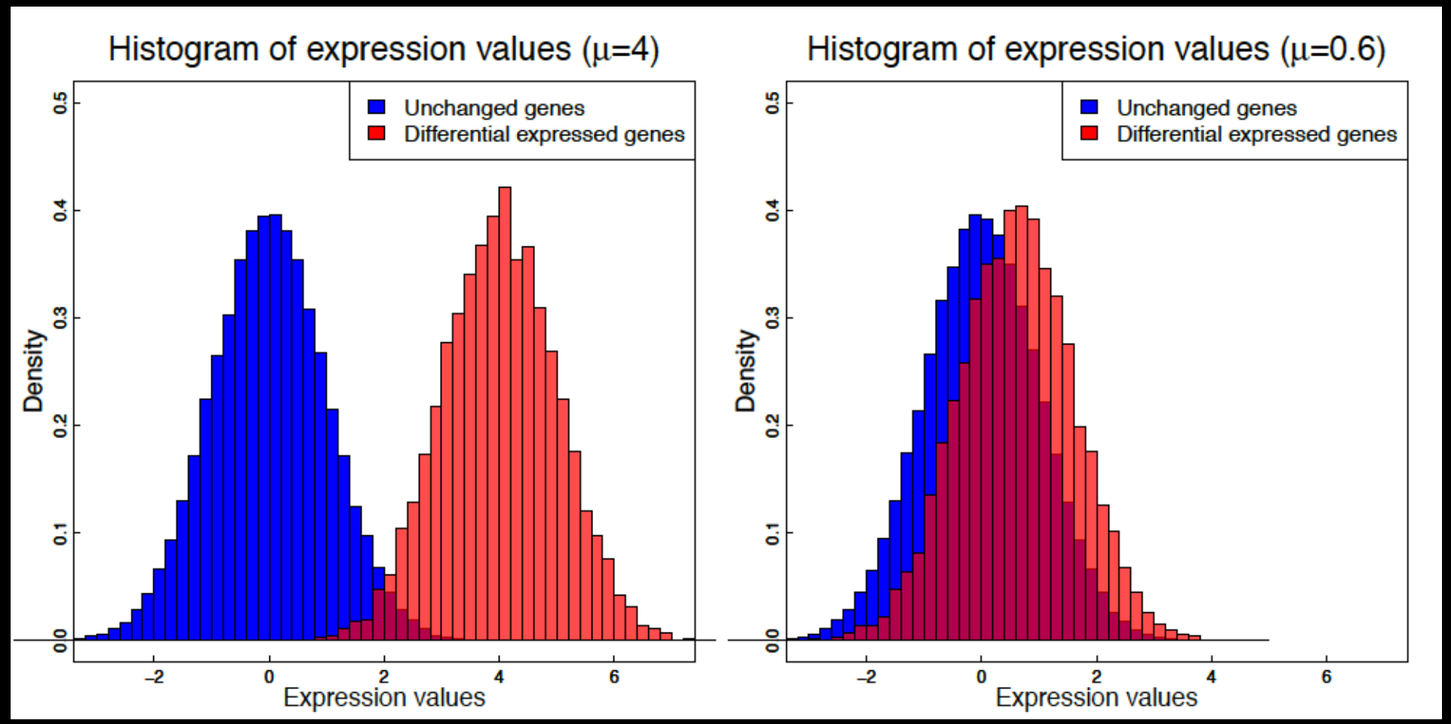


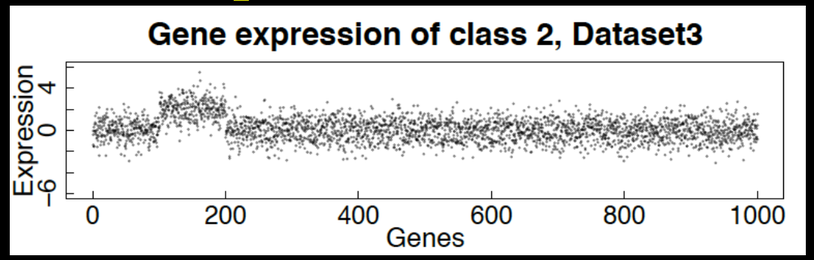
How about stability?



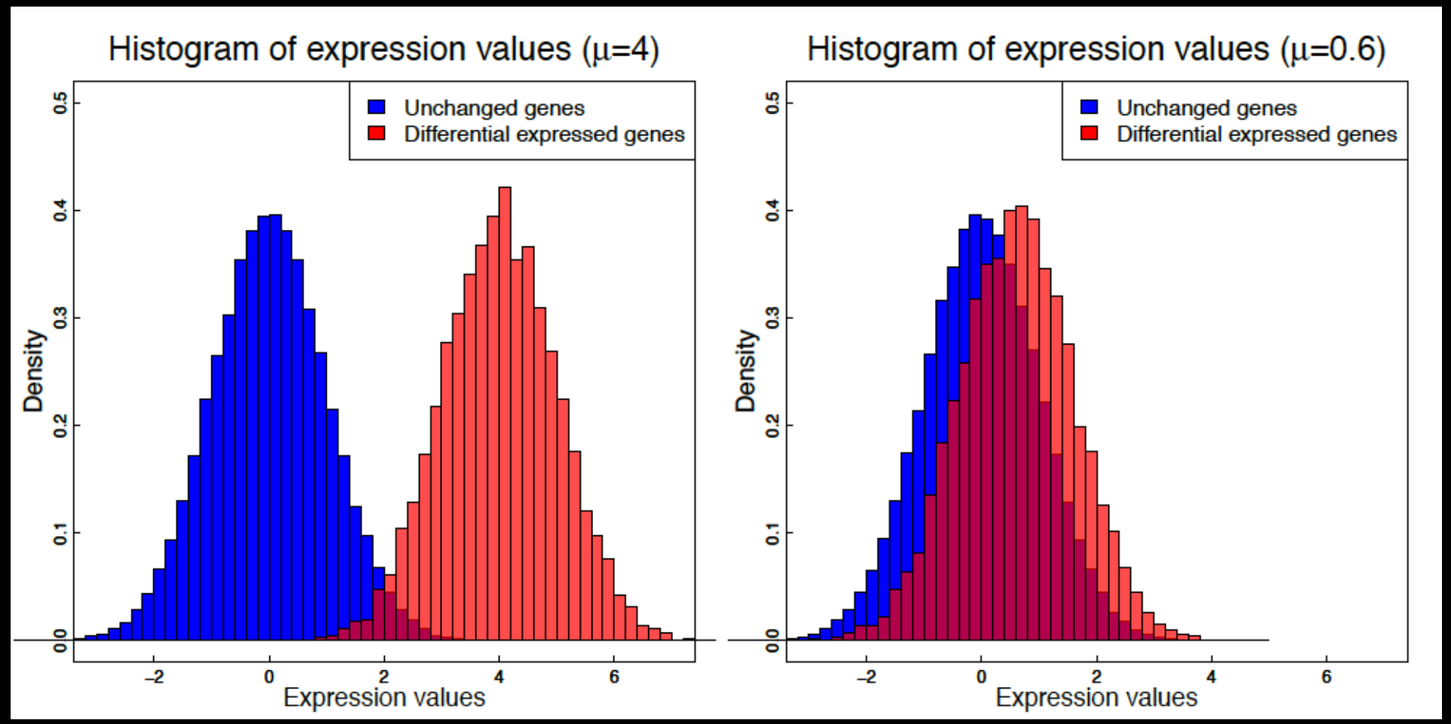


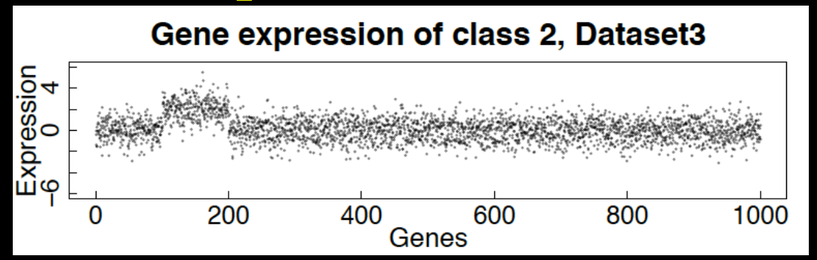
How about stability?





How about stability?





Distance (μ)
4
3
2
1
0.9
0.8
0.7
0.6

How do we assess performance?

Rand Index

Given a set of n elements $S = \{o_1, \ldots, o_n\}$ and two partitions of S to compare, $X = \{X_1, \ldots, X_r\}$, a partition of S into r subsets, and $Y = \{Y_1, \ldots, Y_s\}$, a partition of S into s subsets, define the following:

- a, the number of pairs of elements in S that are in the same set in X and in the same set in Y
- b, the number of pairs of elements in S that are in different sets in X and in different sets in Y
- c, the number of pairs of elements in S that are in the same set in X and in different sets in Y
- d, the number of pairs of elements in S that are in different sets in X and in the same set in Y

The Rand index, R, is:[1][2]

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Intuitively, $a + b$ can be considered as the number of agreements between X and Y and $c + d$ as the number of disagreements between X and Y.

How do we assess performance?

Adjusted Rand Index

Given a set S of n elements, and two groupings (*e.g.* clusterings) of these points, namely $X = \{X_1, X_2, \ldots, X_r\}$ and $Y = \{Y_1, Y_2, \ldots, Y_s\}$, the overlap between X and Y can be summarized in a contingency table $[n_{ij}]$ where each entry n_{ij} denotes the number of objects in common between X_i and Y_j : $n_{ij} = |X_i \cap Y_j|$.

X\Y	Y_1	Y_2	$\ldots$	Y_s	Sums
X_1	n_{11}	n_{12}	$\ldots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\ldots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	a_r
Sums	b_1	b_2	$\ldots$	b_s	

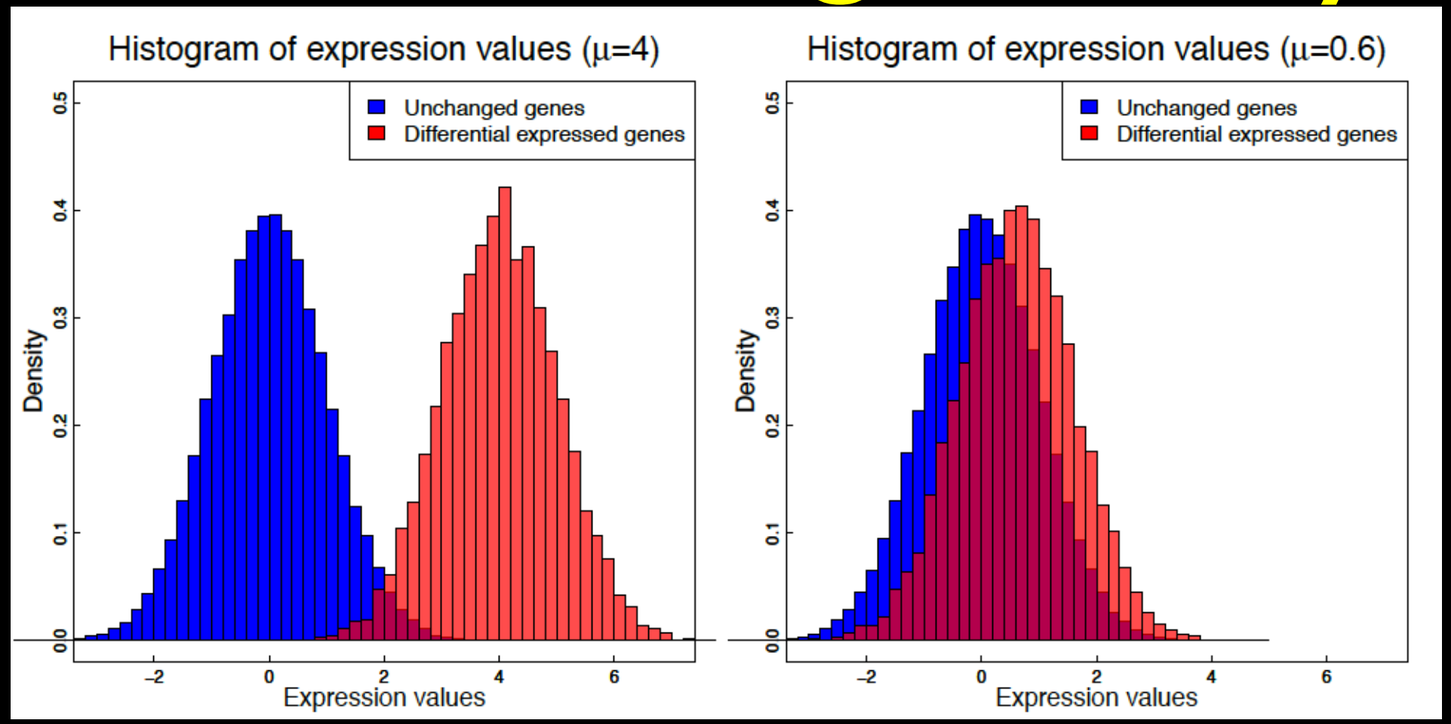
Definition [edit]

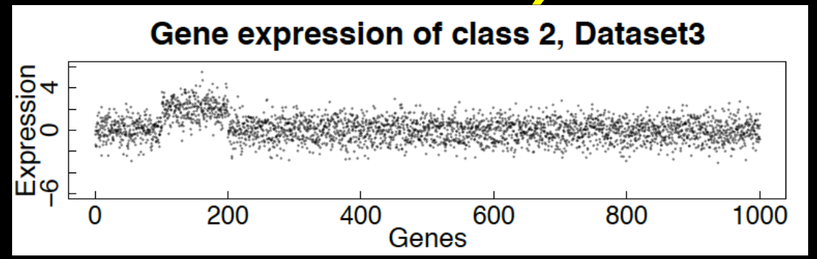
The adjusted form of the Rand Index, the Adjusted Rand Index, is $AdjustedIndex = \dfrac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex}$, more specifically

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}$$

where n_{ij}, a_i, b_j are values from the contingency table.

Clustering stability (simulation)

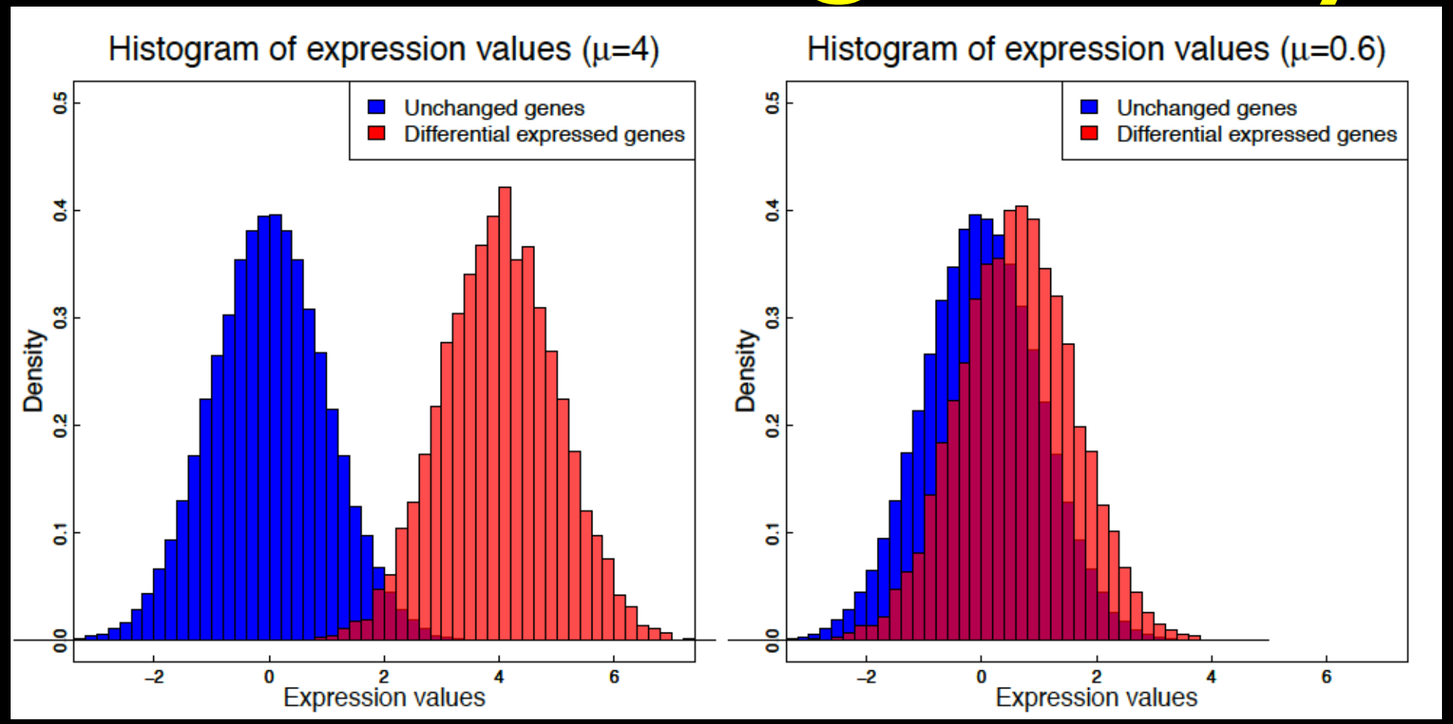


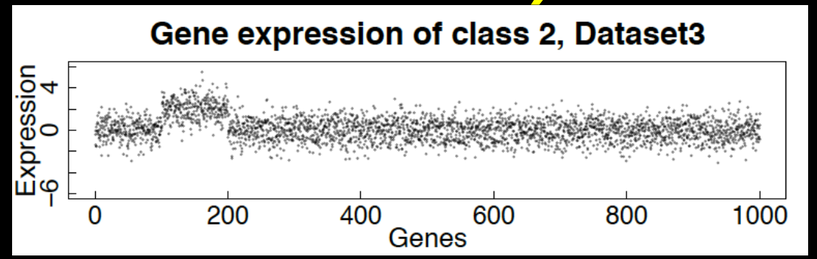


Distance (μ)
4
3
2
1
0.9
0.8
0.7
0.6

- ARI (adjusted Rand index) is the metric used to measure clustering performance

Clustering stability (simulation)

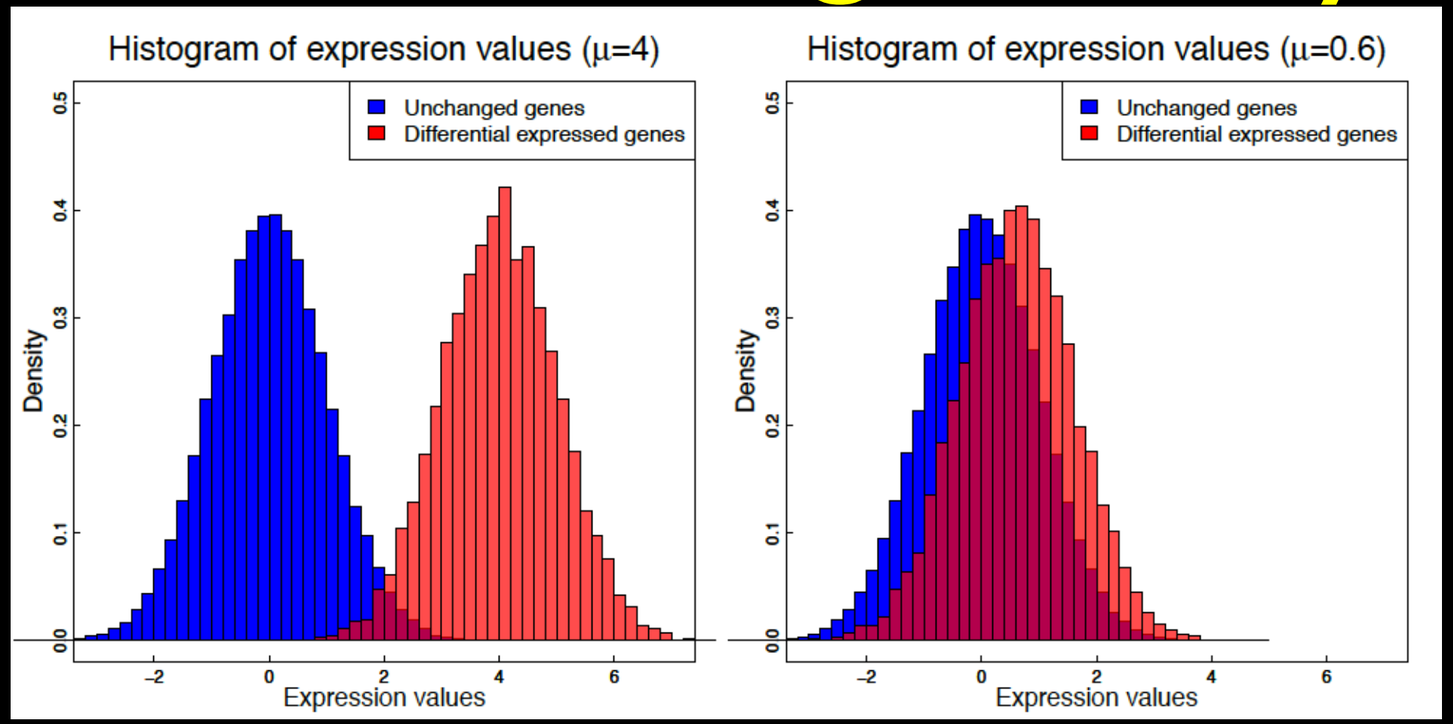


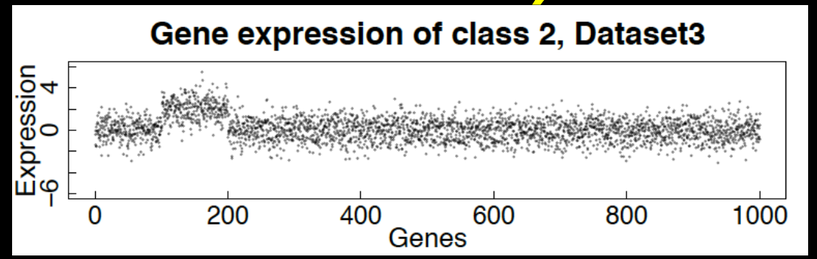


- ARI (adjusted Rand index) is the metric used to measure clustering performance

Distance (μ)	hod iClusterPlus
4	0.968
3	0.966
2	0.964
1	-0.002
0.9	-0.002
0.8	-0.011
0.7	-0.0005
0.6	-0.0005

Clustering stability (simulation)

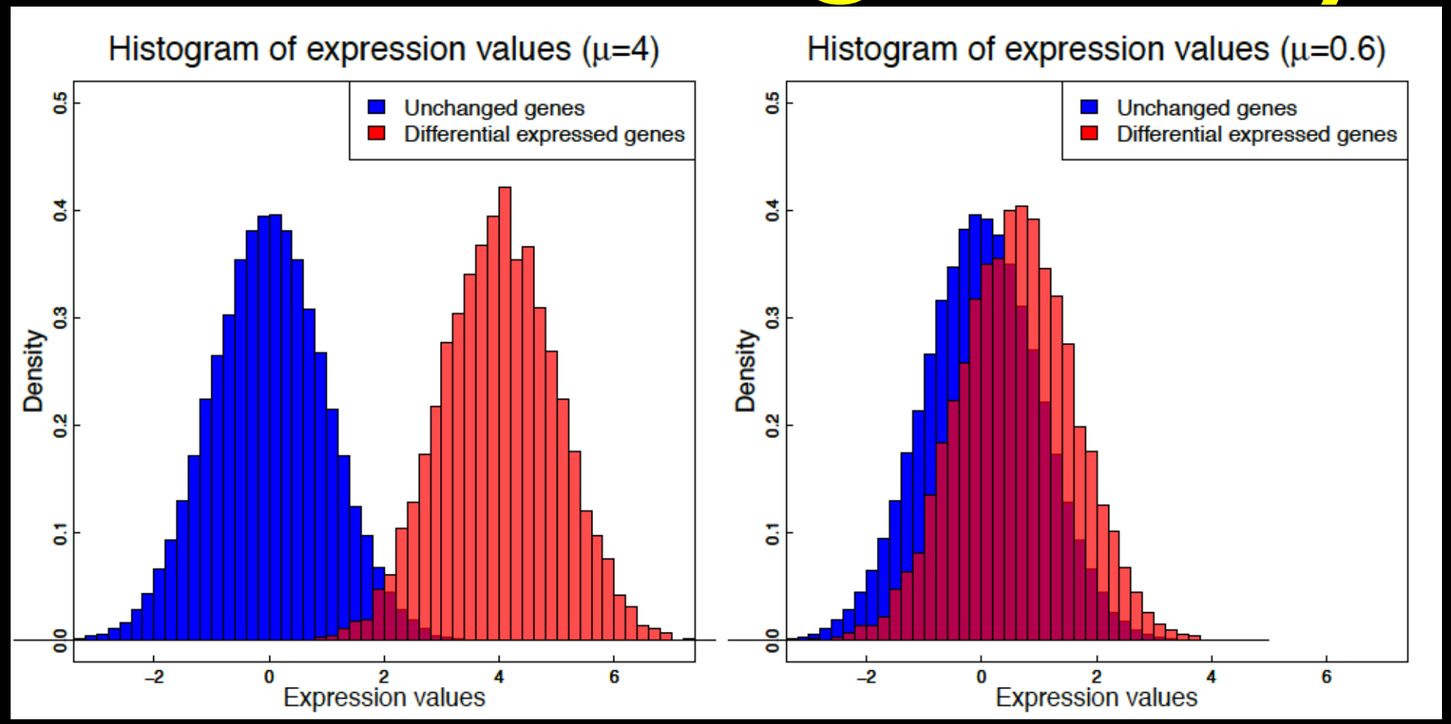


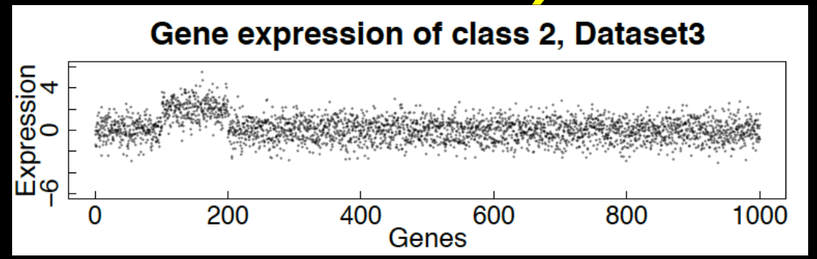


- ARI (adjusted Rand index) is the metric used to measure clustering performance

Distance (μ)	Clust CC	hod iClusterPlus
4	1	0.968
3	1	0.966
2	1	0.964
1	1	-0.002
0.9	1	-0.002
0.8	0.908	-0.011
0.7	0.799	-0.0005
0.6	0.345	-0.0005

Clustering stability (simulation)



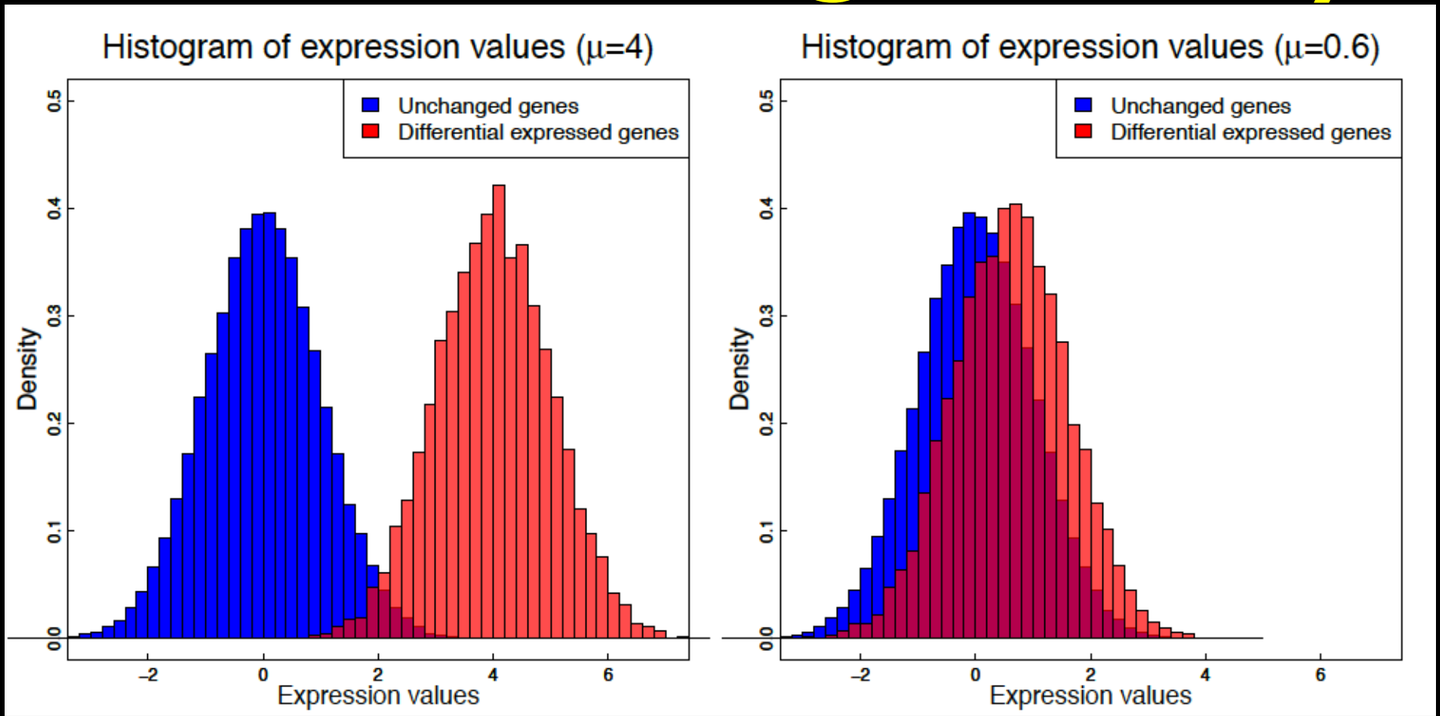


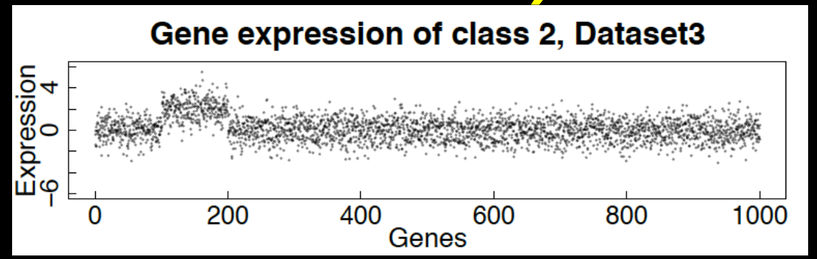
- ARI (adjusted Rand index) is the metric used to measure clustering performance

Distance (μ)
4
3
2
1
0.9
0.8
0.7
0.6

Clustering method		
CC	SNF	iClusterPlus
1	0.06	0.968
1	0.88	0.966
1	0.262	0.964
1	1	-0.002
1	1	-0.002
0.908	1	-0.011
0.799	0.897	-0.0005
0.345	0.221	-0.0005

Clustering stability (simulation)





- ARI (adjusted Rand index) is the metric used to measure clustering performance

Distance (μ)	Noise variance	Clustering method			
		PINS	CC	SNF	iClusterPlus
4	2.577	1	1	0.06	0.968
3	1.877	1	1	0.88	0.966
2	1.384	1	1	0.262	0.964
1	1.08	1	1	1	-0.002
0.9	1.06	1	1	1	-0.002
0.8	1.05	1	0.908	1	-0.011
0.7	1.03	0.94	0.799	0.897	-0.0005
0.6	1.02	0.647	0.345	0.221	-0.0005

Algorithm I validation: mRNA data

- 8 gene expression datasets with known subtypes
  - GSE10245, GSE19188, GSE43580, GSE14924, GSE15061: Gene Expression Omnibus
  - AML2004 (http://www.broadinstitute.org/cancer/pub/nmf)
  - Brain2002 (http://www.broadinstitute.org/MPR/CNS/)
  - Lung2001 (http://www.broadinstitute.org/mpr/lung/)

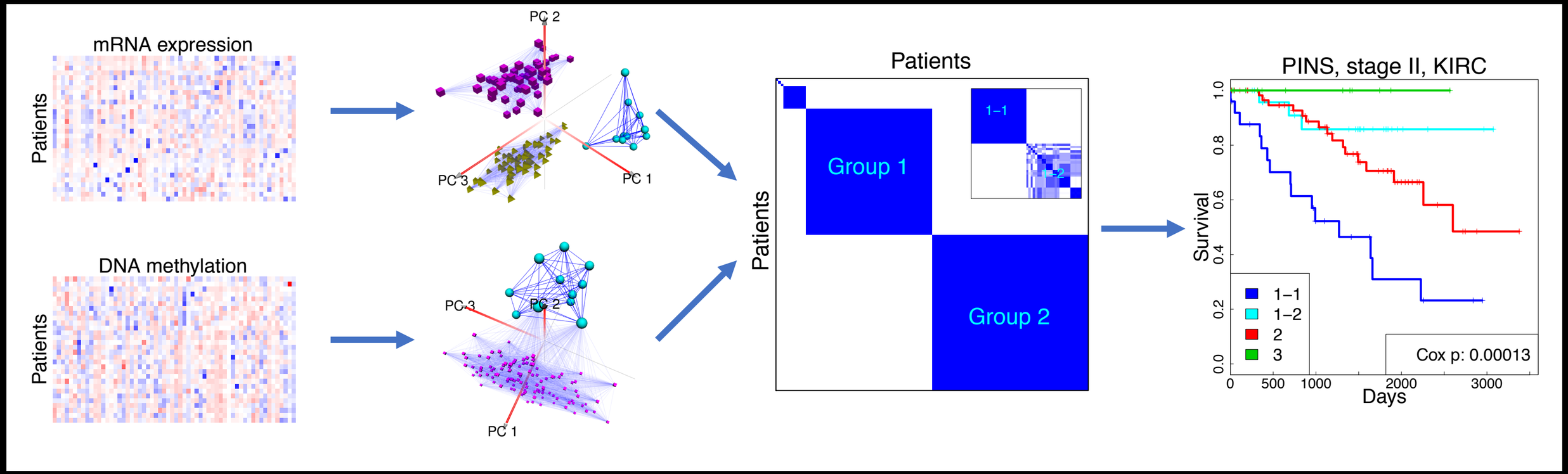
Datasets	#Classes	#Samples	#Components	Platform	Description
GSE10245	2	58	19851	hgu133plus2	40 adenocarcinomas and 18 squamous cell carcinomas
GSE19188	3	91	19851	hgu133plus2	45 adenocarcinomas, 19 large cell carcinomas, and 27 squamous cell carcinomas
GSE43580	2	150	19851	hgu133plus2	77 adenocarcinomas and 73 squamous cell carcinomas
GSE14924	2	20	19851	hgu133plus2	10 acute myeloid leukemia CD4 T cell and 10 CD8 T cell
GSE15061	2	366	19851	hgu133plus2	202 acute myeloid leukemia samples and 164 myelodyplastic syndrome samples
Lung2001	4	237	8641	hgu95a	190 adenocarcinomas, 21 squamous cell carcinomas, 20 carcinoid, and 6 small-cell lung carcinomas
AML2004	3	38	5000	hgu6800	11 acute myeloid leukemia, 19 acute lymphoblastic leukemia B cell, and 8 T cell
Brain2002	5	42	5299	hgu6800	10 meduloblastomas, 10 malignant gliolas, 10 atypical teratoid/rhaboid tumors, 4 normal cerebellums, and 8 primitive neuroectodermal tumors

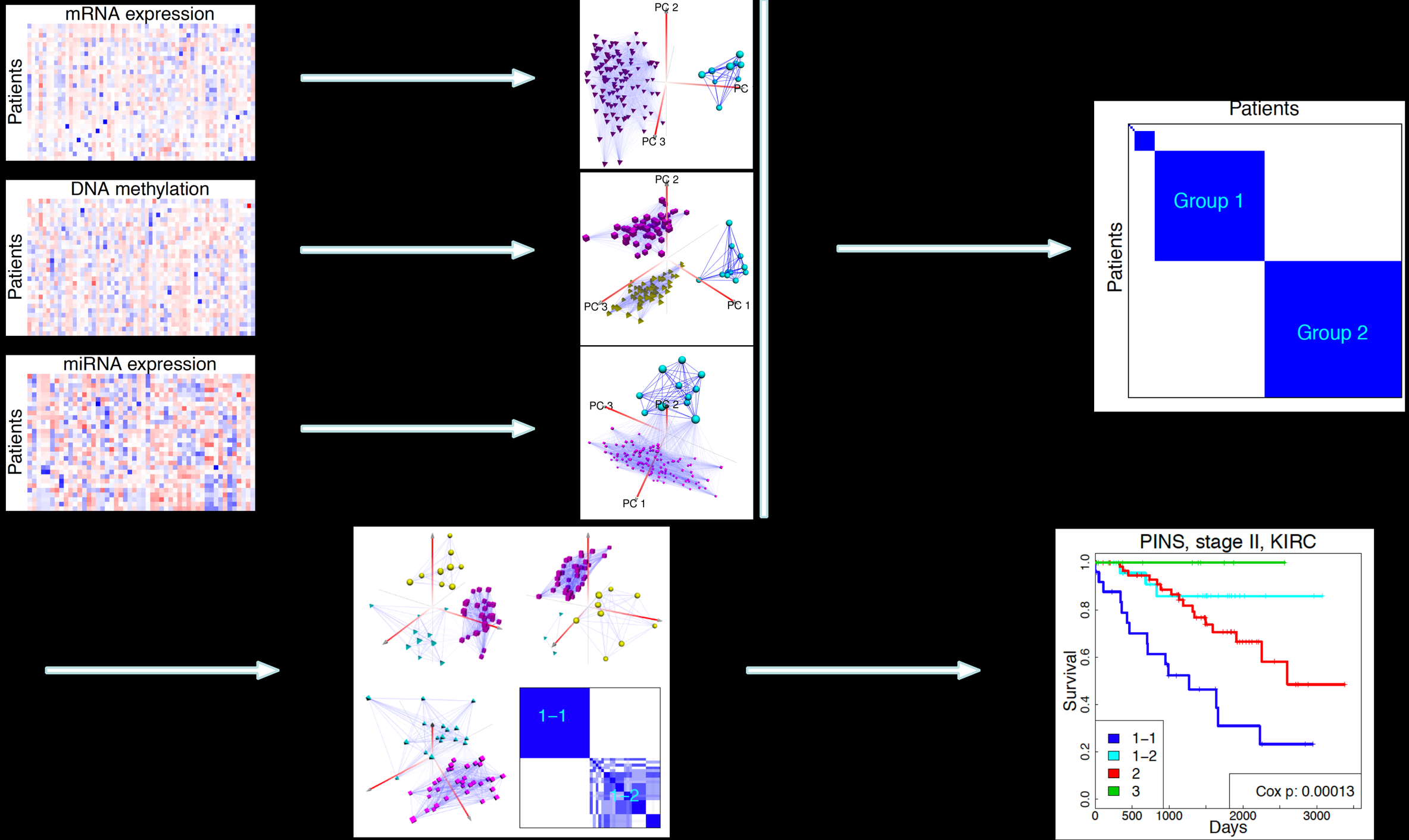
Algorithm I validation: mRNA data

- PINS: Perturbation clustering
- CC: Consensus Clustering (Monti et al., 2004, Machine Learning)
- iClusterPlus (Mo et al., 2013, PNAS)
- SNF: Similarity Network Fusion (Wang et al., 2014, Nature Methods)
- RI (Rand index) and ARI (adjusted Rand index) are metrics to measure clustering performance

Dataset			PINS			CC			SNF			iClusterPlus		
Name	Samples	Subtypes	k	RI	ARI	k	RI	ARI	k	RI	ARI	k	RI	ARI
GSE10245	58	2												
GSE19188	91	3												
GSE43580	150	2												
GSE15061	366	2												
GSE14924	20	2												
Lung2001	237	4												
AML2004	38	3												
Brain2002	42	5												

Subtyping using multi-omics data




mRNA expression
Patients
DNA methylation
Patients
miRNA expression
Patients
PC 2
PC
PC 3
PC 2
PC 3
PC 1
PC 3
PC 2
PC 1
Patients
Patients
Group 1
Group 2
1–1
1–2
PINS, stage II, KIRC
Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
2000
3000
Days
1–1
1–2
2
3
Cox p: 0.00013

Motivation for splitting

- Purposes:
  - Check if the data has hierarchical structure
  - To overcome predominant signals (gender, race, etc.)
  - To avoid imbalanced partitioning
- Conditions:
  - Gap statistic*: do we have enough evidence to reject the null the hypothesis that the data has no obvious clustering?
  - All data types points into similar partitioning (new agreement metric)
  - Normalized entropy to check if stage I partitioning was imbalanced: $p_i = \frac{n_i}{N}, H = \sum_{i=1}^{k} p_i \ln p_i, \widehat{H} = \frac{H}{\ln k}$

** Estimating the number of clusters in a data set via the gap statistic, Tibshirani et al., Journal of the Royal Statistical Society, 2001*

Algorithm II validation: TCGA data

- Data types: mRNA, methylation, and miRNA
- Diseases:
  - Glioblastoma multiforme (GBM): 273 patients
  - Lung squamous cell carcinoma (LUSC): 110 patients
  - Breast invasive carcinoma (BRCA): 172 patients
  - Acute myeloid leukemia (LAML): 164 patients
  - Kidney Renal Clear Cell Carcinoma (KIRC): 131 patients
  - Colon adenocarcinoma (COAD): 146 patients

Algorithm II validation: TCGA data

Dataset	#Sample	Data type	#Components	Platform	Data level
KIRC	124	mRNA	17974	Illumina HiSeq RNASeq	3
		Methylation	23165	HumanMethylation27	3
		miRNA	590	Illumina GASeq miRNASeq	3
GBM	273	mRNA	12042	HT HG-U133A	3
		Methylation	22833	HumanMethylation27	3
		miRNA	534	Illumina HiSeq miRNASeq	3
LAML	164	mRNA	16818	Illumina GASeq RNASeq	3
		Methylation	22833	HumanMethylation27	3
		miRNA	552	Illumina GASeq miRNASeq	3
LUSC	110	mRNA	12042	HT HG-U133A	3
		Methylation	23348	HumanMethylation27	3
		miRNA	706	Illumina GASeq miRNASeq	3
BRCA	172	mRNA	20100	Illumina HiSeq RNASeqV2	3
		Methylation	22533	HumanMethylation27	3
		miRNA	718	Illumina GASeq miRNASeq	3
COAD	146	mRNA	17062	Illumina GASeq RNASeq	3
		Methylation	24454	HumanMethylation27	3
		miRNA	710	Illumina GASeq miRNASeq	3

Algorithm II validation: TCGA data

TCGA dataset			PINS		CC		SNF		iClusterPlus	
Name	Patients	Data type	k	Cox p-value	k	Cox p-value	k	Cox p-value	k	Cox p-value
KIRC	124	mRNA Methylation miRNA								
		Integration								
GBM	273	mRNA Methylation miRNA								
		Integration								
LAML	164	mRNA Methylation miRNA								
		Integration								
LUSC	110	mRNA Methylation miRNA								
		Integration								
BRCA	172	mRNA Methylation miRNA								
		Integration								
COAD	146	mRNA Methylation miRNA								
		Integration								

Algorithm II validation: TCGA data

TCGA dataset			PINS		CC		SNF		iClusterPlus	
Name	Patients	Data type	k	Cox p-value	k	Cox p-value	k	Cox p-value	k	Cox p-value
KIRC	124	mRNA	2	0.176	7	0.073	2	0.219	9	0.072
		Methylation	3	0.111	6	0.128	3	0.577	10	0.14
		miRNA	2	0.138	5	0.509	2	0.138	NA	NA
		Integration	**4**	**1.3×10^{-4}**	**6**	**0.104**	**2**	**0.138**	**6**	**0.077**
GBM	273	mRNA								
		Methylation								
		miRNA								
		Integration								
LAML	164	mRNA								
		Methylation								
		miRNA								
		Integration								
LUSC	110	mRNA								
		Methylation								
		miRNA								
		Integration								
BRCA	172	mRNA								
		Methylation								
		miRNA								
		Integration								
COAD	146	mRNA								
		Methylation								
		miRNA								
		Integration								

Algorithm II validation: TCGA data

TCGA dataset			PINS		CC		SNF		iClusterPlus	
Name	Patients	Data type	k	Cox p-value	k	Cox p-value	k	Cox p-value	k	Cox p-value
KIRC	124	mRNA	2	0.176	7	0.073	2	0.219	9	0.072
		Methylation	3	0.111	6	0.128	3	0.577	10	0.14
		miRNA	2	0.138	5	0.509	2	0.138	NA	NA
		Integration	**4**	$\mathbf{1.3 \times 10^{-4}}$	**6**	**0.104**	**2**	**0.138**	**6**	**0.077**
GBM	273	mRNA	2	0.408	5	0.281	2	0.992	10	0.056
		Methylation	2	10^{-4}	6	0.001	2	0.017	10	0.003
		miRNA	4	0.086	6	0.526	2	0.401	10	0.09
		Integration	**3**	$\mathbf{8.7 \times 10^{-5}}$	**7**	**0.039**	**4**	**0.062**	**5**	**0.076**
LAML	164	mRNA								
		Methylation								
		miRNA								
		Integration								
LUSC	110	mRNA								
		Methylation								
		miRNA								
		Integration								
BRCA	172	mRNA								
		Methylation								
		miRNA								
		Integration								
COAD	146	mRNA								
		Methylation								
		miRNA								
		Integration								

Algorithm II validation: TCGA data

TCGA dataset			PINS		CC		SNF		iClusterPlus	
Name	Patients	Data type	k	Cox p-value	k	Cox p-value	k	Cox p-value	k	Cox p-value
KIRC	124	mRNA	2	0.176	7	0.073	2	0.219	9	0.072
		Methylation	3	0.111	6	0.128	3	0.577	10	0.14
		miRNA	2	0.138	5	0.509	2	0.138	NA	NA
		Integration	**4**	$\mathbf{1.3 \times 10^{-4}}$	**6**	**0.104**	**2**	**0.138**	**6**	**0.077**
GBM	273	mRNA	2	0.408	5	0.281	2	0.992	10	0.056
		Methylation	2	10^{-4}	6	0.001	2	0.017	10	0.003
		miRNA	4	0.086	6	0.526	2	0.401	10	0.09
		Integration	**3**	$\mathbf{8.7 \times 10^{-5}}$	**7**	**0.039**	**4**	**0.062**	**5**	**0.076**
LAML	164	mRNA	5	0.003	6	8×10^{-4}	2	0.327	6	0.01
		Methylation	6	0.239	7	0.049	2	0.993	10	0.002
		miRNA	2	0.072	6	0.017	3	0.183	NA	NA
		Integration	**4**	$\mathbf{2.4 \times 10^{-3}}$	**8**	**0.035**	**3**	**0.037**	**5**	**0.017**
LUSC	110	mRNA								
		Methylation								
		miRNA								
		Integration								
BRCA	172	mRNA								
		Methylation								
		miRNA								
		Integration								
COAD	146	mRNA								
		Methylation								
		miRNA								
		Integration								

Algorithm II validation: TCGA data

TCGA dataset			PINS		CC		SNF		iClusterPlus	
Name	Patients	Data type	k	Cox p-value	k	Cox p-value	k	Cox p-value	k	Cox p-value
KIRC	124	mRNA	2	0.176	7	0.073	2	0.219	9	0.072
		Methylation	3	0.111	6	0.128	3	0.577	10	0.14
		miRNA	2	0.138	5	0.509	2	0.138	NA	NA
		Integration	**4**	$\mathbf{1.3 \times 10^{-4}}$	**6**	**0.104**	**2**	**0.138**	**6**	**0.077**
GBM	273	mRNA	2	0.408	5	0.281	2	0.992	10	0.056
		Methylation	2	10^{-4}	6	0.001	2	0.017	10	0.003
		miRNA	4	0.086	6	0.526	2	0.401	10	0.09
		Integration	**3**	$\mathbf{8.7 \times 10^{-5}}$	**7**	**0.039**	**4**	**0.062**	**5**	**0.076**
LAML	164	mRNA	5	0.003	6	8×10^{-4}	2	0.327	6	0.01
		Methylation	6	0.239	7	0.049	2	0.993	10	0.002
		miRNA	2	0.072	6	0.017	3	0.183	NA	NA
		Integration	**4**	$\mathbf{2.4 \times 10^{-3}}$	**8**	**0.035**	**3**	**0.037**	**5**	**0.017**
LUSC	110	mRNA	3	0.125	5	0.782	3	0.095	7	0.588
		Methylation	8	0.019	9	0.129	2	0.376	10	0.606
		miRNA	2	0.117	6	0.938	2	0.001	NA	NA
		Integration	**5**	$\mathbf{9.7 \times 10^{-3}}$	**6**	**0.794**	**3**	**0.428**	**4**	**0.36**
BRCA	172	mRNA								
		Methylation								
		miRNA								
		Integration								
COAD	146	mRNA								
		Methylation								
		miRNA								
		Integration								

Algorithm II validation: TCGA data

TCGA dataset			PINS		CC		SNF		iClusterPlus	
Name	Patients	Data type	k	Cox p-value	k	Cox p-value	k	Cox p-value	k	Cox p-value
KIRC	124	mRNA	2	0.176	7	0.073	2	0.219	9	0.072
		Methylation	3	0.111	6	0.128	3	0.577	10	0.14
		miRNA	2	0.138	5	0.509	2	0.138	NA	NA
		Integration	**4**	$\mathbf{1.3 \times 10^{-4}}$	**6**	**0.104**	**2**	**0.138**	**6**	**0.077**
GBM	273	mRNA	2	0.408	5	0.281	2	0.992	10	0.056
		Methylation	2	10^{-4}	6	0.001	2	0.017	10	0.003
		miRNA	4	0.086	6	0.526	2	0.401	10	0.09
		Integration	**3**	$\mathbf{8.7 \times 10^{-5}}$	**7**	**0.039**	**4**	**0.062**	**5**	**0.076**
LAML	164	mRNA	5	0.003	6	8×10^{-4}	2	0.327	6	0.01
		Methylation	6	0.239	7	0.049	2	0.993	10	0.002
		miRNA	2	0.072	6	0.017	3	0.183	NA	NA
		Integration	**4**	$\mathbf{2.4 \times 10^{-3}}$	**8**	**0.035**	**3**	**0.037**	**5**	**0.017**
LUSC	110	mRNA	3	0.125	5	0.782	3	0.095	7	0.588
		Methylation	8	0.019	9	0.129	2	0.376	10	0.606
		miRNA	2	0.117	6	0.938	2	0.001	NA	NA
		Integration	**5**	$\mathbf{9.7 \times 10^{-3}}$	**6**	**0.794**	**3**	**0.428**	**4**	**0.36**
BRCA	172	mRNA	2	0.902	8	0.114	2	0.969	9	0.101
		Methylation	4	0.048	8	0.578	5	0.878	10	0.083
		miRNA	3	0.218	5	0.142	2	0.105	NA	NA
		Integration	**7**	$\mathbf{3.4 \times 10^{-2}}$	**7**	**0.667**	**2**	**0.398**	**10**	**0.416**
COAD	146	mRNA								
		Methylation								
		miRNA								
		Integration								

Algorithm II validation: TCGA data

TCGA dataset			PINS		CC		SNF		iClusterPlus	
Name	Patients	Data type	k	Cox p-value	k	Cox p-value	k	Cox p-value	k	Cox p-value
KIRC	124	mRNA	2	0.176	7	0.073	2	0.219	9	0.072
		Methylation	3	0.111	6	0.128	3	0.577	10	0.14
		miRNA	2	0.138	5	0.509	2	0.138	NA	NA
		Integration	**4**	$\mathbf{1.3 \times 10^{-4}}$	**6**	**0.104**	**2**	**0.138**	**6**	**0.077**
GBM	273	mRNA	2	0.408	5	0.281	2	0.992	10	0.056
		Methylation	2	10^{-4}	6	0.001	2	0.017	10	0.003
		miRNA	4	0.086	6	0.526	2	0.401	10	0.09
		Integration	**3**	$\mathbf{8.7 \times 10^{-5}}$	**7**	**0.039**	**4**	**0.062**	**5**	**0.076**
LAML	164	mRNA	5	0.003	6	8×10^{-4}	2	0.327	6	0.01
		Methylation	6	0.239	7	0.049	2	0.993	10	0.002
		miRNA	2	0.072	6	0.017	3	0.183	NA	NA
		Integration	**4**	$\mathbf{2.4 \times 10^{-3}}$	**8**	**0.035**	**3**	**0.037**	**5**	**0.017**
LUSC	110	mRNA	3	0.125	5	0.782	3	0.095	7	0.588
		Methylation	8	0.019	9	0.129	2	0.376	10	0.606
		miRNA	2	0.117	6	0.938	2	0.001	NA	NA
		Integration	**5**	$\mathbf{9.7 \times 10^{-3}}$	**6**	**0.794**	**3**	**0.428**	**4**	**0.36**
BRCA	172	mRNA	2	0.902	8	0.114	2	0.969	9	0.101
		Methylation	4	0.048	8	0.578	5	0.878	10	0.083
		miRNA	3	0.218	5	0.142	2	0.105	NA	NA
		Integration	**7**	$\mathbf{3.4 \times 10^{-2}}$	**7**	**0.667**	**2**	**0.398**	**10**	**0.416**
COAD	146	mRNA	2	0.113	8	0.048	2	0.148	6	0.29
		Methylation	2	0.741	8	0.034	2	0.389	10	0.194
		miRNA	4	0.452	7	0.318	3	0.131	NA	NA
		Integration	**5**	**0.201**	**5**	**0.225**	**2**	**0.296**	**10**	**0.445**

Algorithm II validation: METABRIC

- Molecular Taxonomy of Breast Cancer International Consortium
- Cohorts: discovery (997) and validation (983)
- Data source: European Genome-Phenome Archive (http://www.ebi.ac.uk/ega/)
  - Gene expression (Illumina HT 12 v3)
  - Copy number variation (Affymetrix SNP 6)
- High quality clinical data: cBioPortal (http://www.cbioportal.org)
  - Disease free survival (DFS) and overall survival

Data	#Patients	Metric	Survival	PINS	CC	SNF	iClusterPlus
Discovery	997	P-value	DFS				
			Overall				
		CI	DFS				
			Overall				
Validation	983	P-value	DFS				
			Overall				
		CI	DFS				
			Overall				

GENOME RESEARCH

A novel approach for data integration and disease subtyping

Tin Nguyen, Rebecca Tagett, Diana Diaz, et al.

Genome Res. 2017 27: 2025-2039 originally published online October 24, 2017
Access the most recent version at doi:10.1101/gr.215129.116

More results

Consortium	Dataset	#Patients	PINSPlus2	CC	SNF	iClusterBayes	CIMLR
METABRIC	Discovery	997	3.00E-10	0.022	2.30E-05	0.130	8.79E-13
	Validation	983	2.66E-05	0.096	0.010	0.543	5.74E-5
TCGA	KIRC	124	5.98E-05	0.118	0.691	0.832	0.091
	GBM	273	8.75E-05	0.014	0.021	0.115	0.081
	LAML	164	8.72E-04	0.292	0.002	0.898	1.43E-04
	LUSC	110	0.008	0.688	0.087	0.264	0.039
	BLCA	404	0.019	0.089	0.109	0.511	0.470
	HNSC	228	0.046	0.428	0.366	0.372	0.404
	STAD	362	1.86E-04	0.428	0.041	0.658	0.269
	THYM	119	0.013	0.139	0.097	0.009	0.115
	GBMLGG	510	7.48E-17	5.2E-04	4.8E-14	0.080	6.36E-10
	LGG	510	4.26E-15	2.0E-06	1.6E-14	0.108	8.27E-15
	PAAD	178	2.73E-04	0.013	7.4E-04	0.002	0.002
	COADREAD	294	0.003	0.946	0.660	0.210	0.135
	UCEC	234	0.005	0.105	0.018	0.059	0.046
	CESC	304	0.030	0.376	0.510	0.020	0.190
	COAD	220	0.001	0.419	0.128	0.220	0.561
	BRCA	622	0.002	0.008	0.119	0.027	0.005
	STES	545	0.015	0.301	0.157	0.004	0.034
	KIRP	271	1.15E-09	0.367	0.005	0.003	0.019
	KICH	65	0.028	0.955	0.701	0.692	0.463
	UVM	80	0.006	0.005	1.7E-04	0.066	0.001
	ACC	79	0.013	0.014	4.3E-05	0.001	0.338
	SARC	257	0.030	0.148	0.044	0.043	0.056
	MESO	86	7.34E-04	0.272	4.2E-04	0.037	0.011
	READ	74	0.024	0.737	0.762	0.897	0.335
	LUAD	428	0.066	0.926	0.501	0.022	0.373
	SKCM	439	0.105	0.604	0.478	0.008	7.36E-05
	LIHC	366	0.704	0.622	0.334	0.093	0.186
	UCS	56	0.426	0.207	0.859	0.959	0.359
	OV	286	0.681	0.859	0.445	0.457	0.536
	ESCA	183	0.330	0.791	0.392	0.793	0.558
	PCPG	179	0.866	0.938	0.322	0.667	0.457
	PRAD	493	0.349	0.638	0.475	0.373	0.310
	THCA	499	0.088	0.640	0.620	0.784	0.009
	TGCT	134	0.842	0.758	0.838	0.711	0.839

Time complexity

Consortium	Dataset	#Patients	PINSPlus2	PINSPlus	PINS	CC	SNF	iClusterBayes	CIMLR
METABRIC	Discovery	997	2m	10m	1153m	15m	2m	14m	11m
	Validation	983	2m	9m	581m	14m	2m	14m	10m
TCGA	KIRC	124	<1m	<1m	6m	<1m	<1m	8m	<1m
	GBM	273	<1m	2m	53m	<1m	<1m	16m	2m
	LAML	164	<1m	<1m	10m	<1m	<1m	10m	1m
	LUSC	110	<1m	<1m	5m	<1m	<1m	7m	<1m
	BLCA	404	2m	7m	112m	3m	3m	24m	4m
	HNSC	228	1m	4m	32m	3m	2m	14m	2m
	STAD	362	1m	5m	97m	4m	3m	24m	5m
	THYM	119	1m	1m	6m	2m	1m	8m	1m
	GBMLGG	510	2m	7m	192m	7m	4m	39m	6m
	LGG	510	2m	12m	188m	8m	6m	31m	8m
	PAAD	178	1m	3m	20m	2m	1m	11m	2m
	COADREAD	294	1m	5m	61m	4m	3m	22m	3m
	UCEC	234	1m	4m	34m	4m	2m	19m	2m
	CESC	304	1m	6m	60m	5m	2m	22m	4m
	COAD	220	<1m	3m	30m	3m	2m	16m	2m
	BRCA	622	2m	13m	236m	10m	5m	47m	11m
	STES	545	2m	12m	171m	14m	5m	36m	10m
	KIRP	271	1m	4m	33m	3m	1m	16m	3m
	KICH	65	<1m	1m	4m	1m	<1m	5m	<1m
	UVM	80	<1m	<1m	3m	1m	1m	7m	<1m
	ACC	79	<1m	<1m	3m	1m	<1m	6m	<1m
	SARC	257	2m	6m	43m	3m	1m	17m	3m
	MESO	86	<1m	<1m	4m	2m	<1m	6m	<1m
	READ	74	<1m	1m	3m	2m	<1m	5m	<1m
	LUAD	428	2m	7m	128m	5m	3m	32m	5m
	SKCM	439	2m	7m	144m	3m	3m	26m	5m
	LIHC	366	1m	5m	96m	4m	3m	20m	4m
	UCS	56	<1m	1m	2m	1m	<1m	6m	<1m
	OV	286	<1m	3m	52m	2m	1m	17m	2m
	ESCA	183	2m	5m	23m	5m	2m	11m	2m
	PCPG	179	<1m	2m	16m	3m	1m	14m	2m
	PRAD	493	1m	11m	205m	10m	5m	33m	8m
	THCA	499	1m	9m	213m	5m	3m	33m	6m
	TGCT	134	<1m	2m	9m	2m	1m	9m	1m

PINSPlus: CRAN R package

CSH PRESS GENOME RESEARCH

A novel approach for data integration and disease subtyping

Tin Nguyen, Rebecca Tagett, Diana Diaz, et al.

Genome Res. 2017 27: 2025-2039 originally published online October 24, 2017
Access the most recent version at doi:10.1101/gr.215129.116

Bioinformatics, 2019, 1–4
doi: 10.1093/bioinformatics/bty1049
Advance Access Publication Date: 24 December 2018
Applications Note

OXFORD

Genome analysis

PINSPlus: a tool for tumor subtype discovery in integrated genomic data

Hung Nguyen[1], Sangam Shrestha[1], Sorin Draghici[2] and Tin Nguyen [1,*]

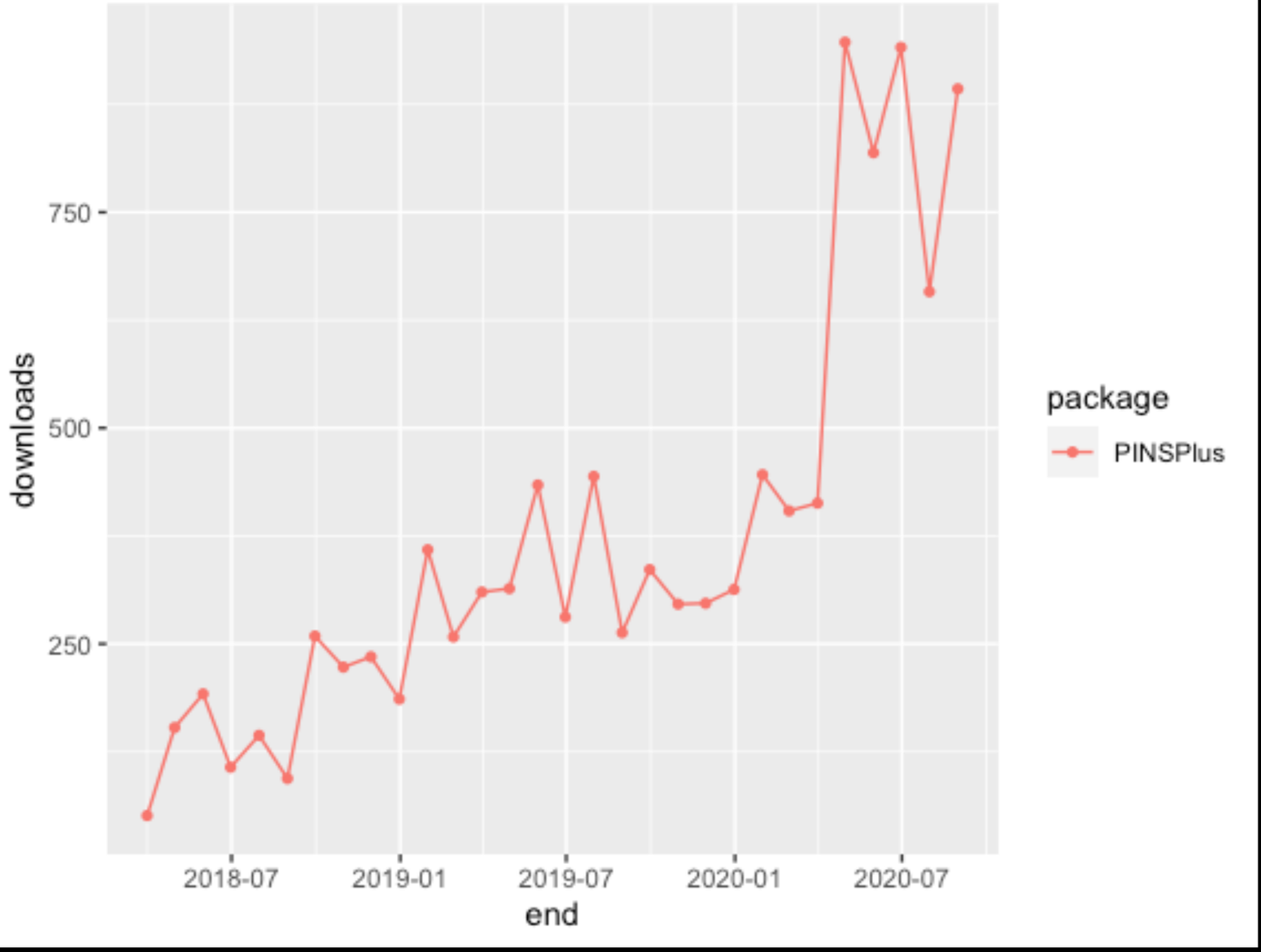
frontiers in Oncology

METHODS
published: 24 June 2020
doi: 10.3389/fonc.2020.01052

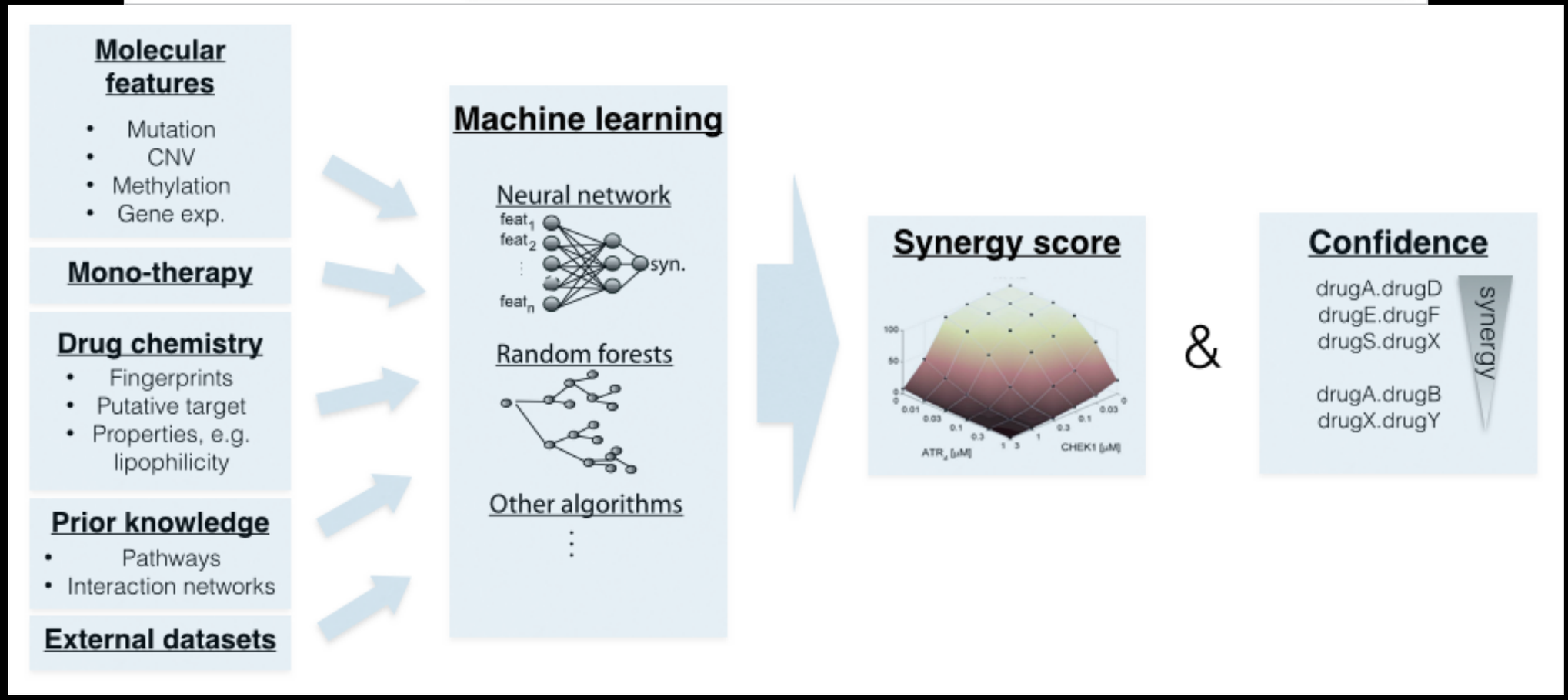
Check for updates

A Novel Method for Cancer Subtyping and Risk Prediction Using Consensus Factor Analysis

Duc Tran[1], Hung Nguyen[1], Uyen Le[2], George Bebis[1], Hung N. Luu[3,4] and Tin Nguyen[1]*



Application: predict drug synergy



Some recent pubs. (2017-2020)

Cancer subtyping

CSH PRESS GENOME RESEARCH

A novel approach for data integration and disease subtyping

Tin Nguyen, Rebecca Tagett, Diana Diaz, et al.

Genome Res. 2017 27: 2025-2039 originally published online October 24, 2017
Access the most recent version at doi:10.1101/gr.215129.116

Bioinformatics, 2019, 1–4
doi: 10.1093/bioinformatics/bty1049
Advance Access Publication Date: 24 December 2018
Applications Note

OXFORD

Genome analysis

PINSPlus: a tool for tumor subtype discovery in integrated genomic data

Hung Nguyen[1], Sangam Shrestha[1], Sorin Draghici[2] and Tin Nguyen[1,*]

frontiers in Oncology

A Novel Method for Cancer Subtyping and Risk Prediction Using Consensus Factor Analysis

Duc Tran[1], Hung Nguyen[1], Uyen Le[2], George Bebis[1], Hung N. Luu[3,4] and Tin Nguyen[1*]

Pathway analysis

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09799-2 OPEN

Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen

Nguyen *et al. Genome Biology* (2019) 20:203
https://doi.org/10.1186/s13059-019-1790-4

Genome Biology

RESEARCH Open Access

Identifying significantly impacted pathways: a comprehensive review and assessment

Check for updates

NBIA: a network-based integrative analysis framework – applied to pathway analysis

Tin Nguyen[1*], Adib Shafi[3], Tuan-Minh Nguyen[3], A. Grant Schissle

nature SCIENTIFIC REPORTS

Single-cell RNA seq.

nature COMMUNICATIONS

Fast and precise single-cell data analysis using hierarchical autoencoder

Duc Tran, Hung Nguyen, Bang Tran, Carlo La Vecchia, Hung N. Luu, Tin Nguyen

OXFORD Briefings in Bioinformatics

A comprehensive survey of regulatory network inference methods using single-cell RNA sequencing data

Hung Nguyen, Duc Tran, Bang Tran, Bahadir Pehlivan and Tin Nguyen

Published Online: 24 September, 2020 | Supp Info: http://doi.org/10.26508/lsa.202000867

Research Article

TRANSPARENT PROCESS OPEN ACCESS Life Science Alliance

Check for updates

Gene selection for optimal prediction of cell position in tissues from single-cell transcriptomics data

Acknowledgement

Intelligent Systems and Bioinformatics Laboratory, Wayne State


Sorin Draghici


Rebecca Taggett


Diana Diaz

GeneLab Open Science for Life in Space

WAYNE STATE UNIVERSITY

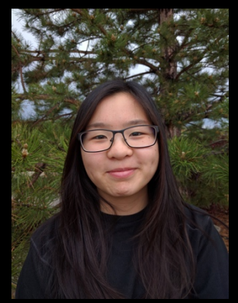
Bioinformatics Laboratory, UNR


Hung Nguyen


Sangam Shrestha


Bang Tran


Duc Tran


Alena Lee

https://bioinformatics.cse.unr.edu

sas