For written notes on this lecture, please read chapter 14 of *The Practical Bioinformatician*.

CS2220: Introduction to Computational Biology
Unit 3: Gene Expression Analysis

Wong Limsoon

Plan

- Microarray background
- Gene expression profile classification
- Gene expression profile clustering
- Normalization
- Extreme sample selection
- Gene regulatory network inference

Copyright 2019 © Wong Limsoon

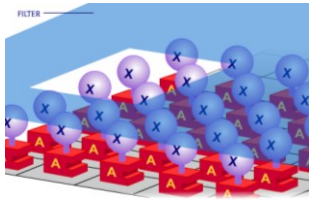
Background on microarrays

NUS
National University
of Singapore

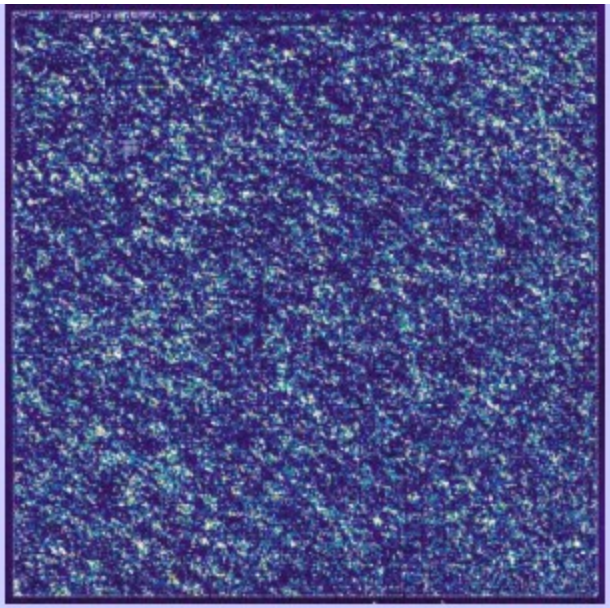
What is a microarray?

- **Contain large numbers of DNA molecules spotted on glass slides, nylon membranes, or silicon wafers**
- **Detect what genes are being expressed or found in a cell of a tissue sample**
- **Measure expression of thousands of genes simultaneously**

Copyright 2019 © Wong Limsoon



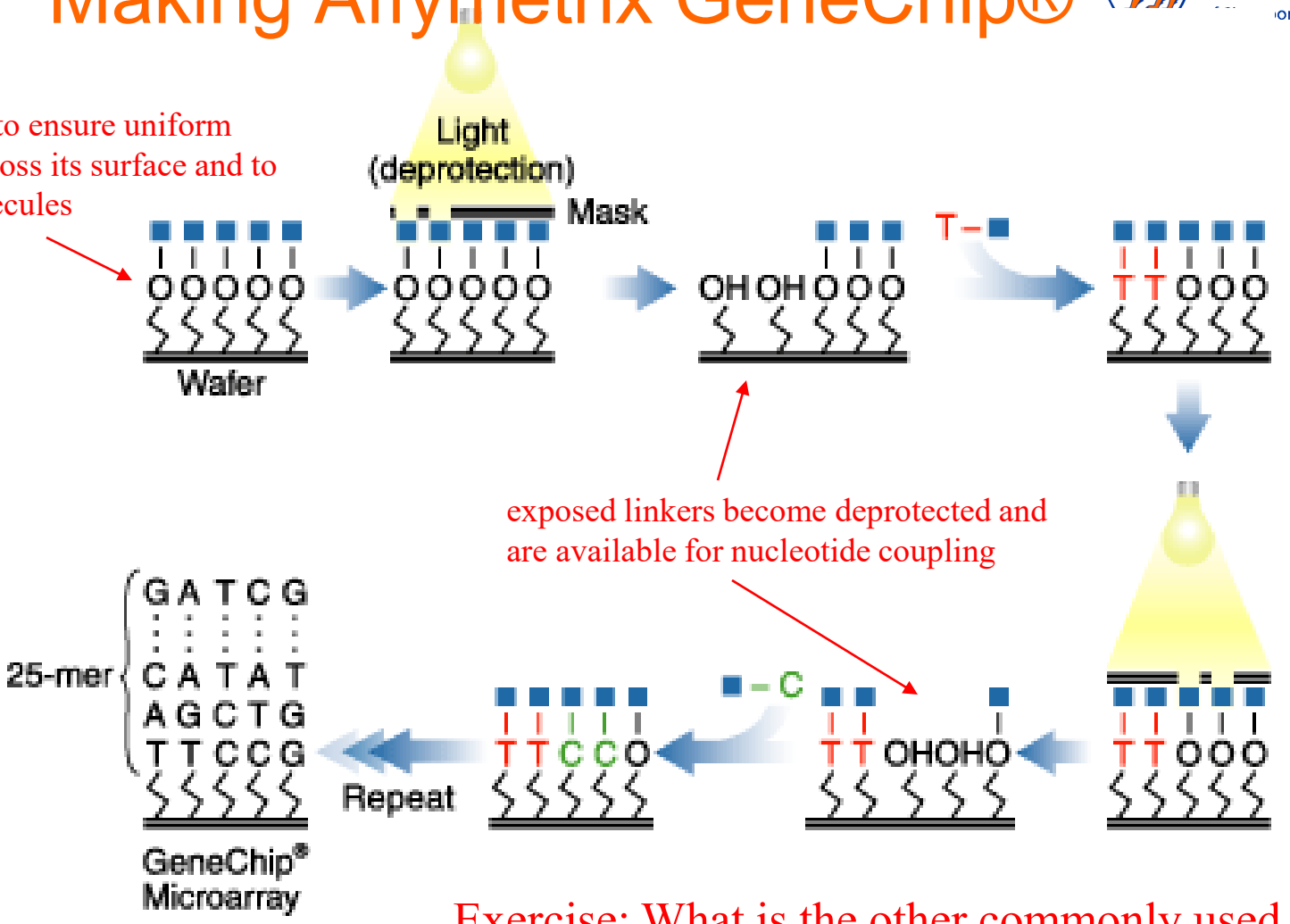
Affymetrix GeneChip®



Copyright 2019 © Wong Limsoon

Making Affymetrix GeneChip®

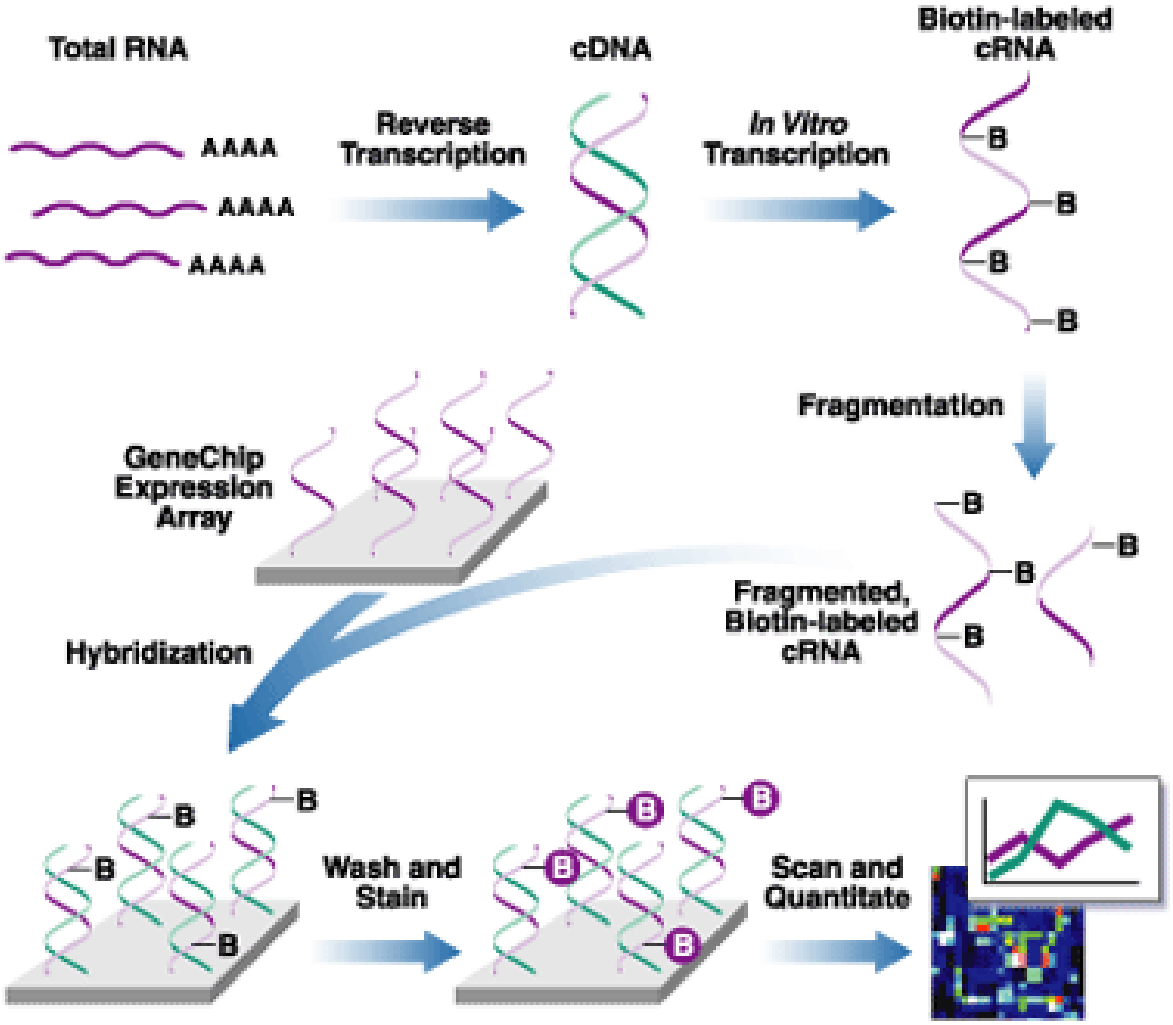
quartz is washed to ensure uniform hydroxylation across its surface and to attach linker molecules



Exercise: What is the other commonly used type of microarray? How is that one different from Affymetrix's?

Copyright 2019 © Wong Limsoon

Gene expression measurement by Affymetrix GeneChip®



Click to watch an interesting movie explaining the working of microarray

Copyright 2019 © Wong Limsoon

Sample Affymetrix GeneChip® data file (U95A)

NUS National University of Singapore

	00-0586-U	00-0586-U	00-0586-U	00-0586-U	00-0586-U	Descriptions
	Positive	Negative	Pairs InAv	Avg Diff	Abs Call	
AFFX-MurI	5	2	19	297.5	A	M16762 Mouse interleukin 2 (IL-2) gene, exon 4
AFFX-MurI	3	2	19	554.2	A	M37897 Mouse interleukin 10 mRNA, complete cds
AFFX-MurI	4	2	19	308.6	A	M25892 Mus musculus interleukin 4 (Il-4) mRNA, comp
AFFX-MurF	1	3	19	141	A	M83649 Mus musculus Fas antigen mRNA, complete
AFFX-BioE	13	1	19	9340.6	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioE	15	0	19	12862.4	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioE	12	0	19	8716.5	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioC	17	0	19	25942.5	P	J04423 E coli bioC protein (-5 and -3 represent transcr
AFFX-BioC	16	0	20	28838.5	P	J04423 E coli bioC protein (-5 and -3 represent transcr
AFFX-BioD	17	0	19	25765.2	P	J04423 E coli bioD gene dethiobiotin synthetase (-5 ar
AFFX-BioD	19	0	20	140113.2	P	J04423 E coli bioD gene dethiobiotin synthetase (-5 ar
AFFX-CreX	20	0	20	280036.6	P	X03453 Bacteriophage P1 cre recombinase protein (-5
AFFX-CreX	20	0	20	401741.8	P	X03453 Bacteriophage P1 cre recombinase protein (-5
AFFX-BioE	7	5	18	-483	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioE	5	4	18	313.7	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioE	7	6	20	-1016.2	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r

Copyright 2019 © Wong Limsoon

Some advice on processing Affymetrix GeneChip® data

- **Ignore AFFX genes**
  - These genes are control genes
- **Ignore genes with "Abs Call" equal to "A" or "M"**
  - Measurement quality is suspect
- **Upperbound 40000, lowerbound 100**
  - Saturation of laser scanner
- **Deal with missing values**

Exercise: Suggest 2 ways to deal with missing value

Copyright 2019 © Wong Limsoon

Type of gene expression datasets

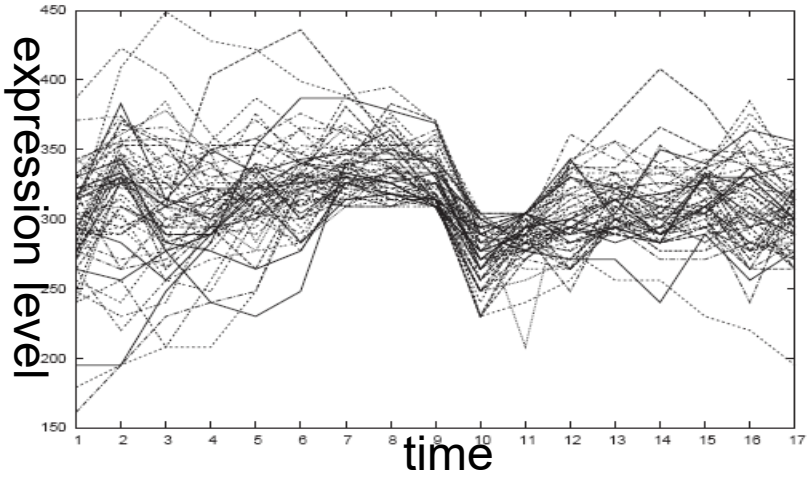
- Gene-Conditions or **Gene-Sample** (**numeric** or discretized)

1000 - 100,000 columns

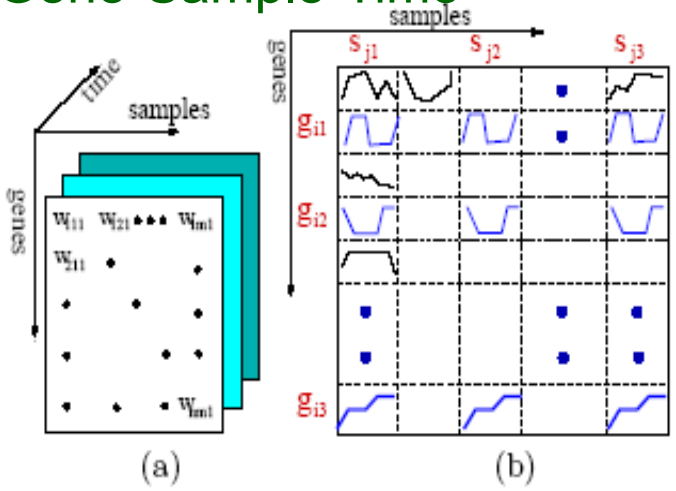
100-500 rows

	Class	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	.....	
Sample1	Cancer	0.12	-1.3	1.7	1.0	-3.2	0.78	-0.12		
Sample2	Cancer							1.3		
.										
	~Cancer									
SampleN	~Cancer									

- Gene-Time



- Gene-Sample-Time



Copyright 2019 © Wong Limsoon

Type of gene expression datasets

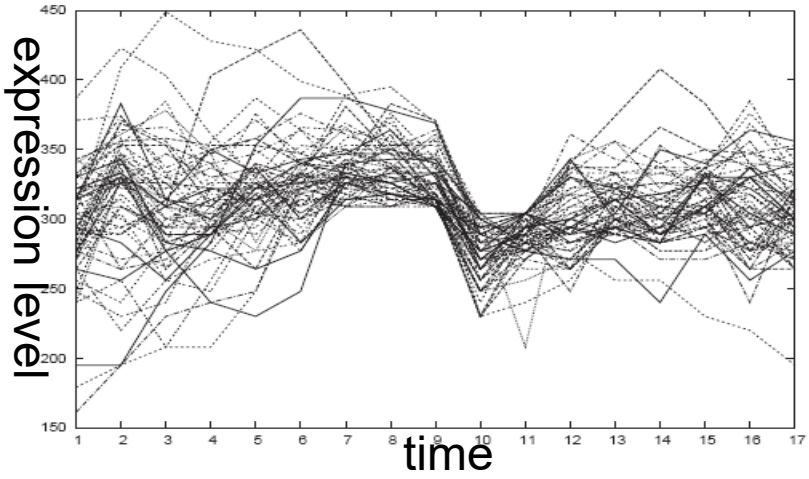
- Gene-Conditions or **Gene-Sample** (numeric or **discretized**)

1000 - 100,000 columns

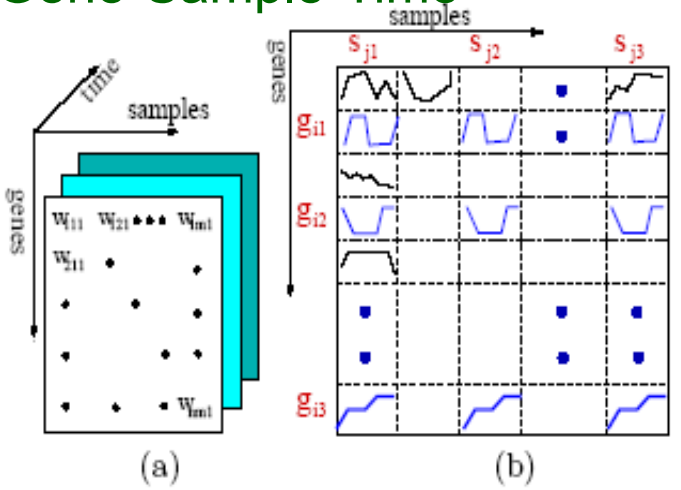
100-500 rows

	Class	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	.....	
Sample1	Cancer	1	0	1	1	1	0	0		
Sample2	Cancer							1		
.										
	~Cancer									
SampleN	~Cancer									

- Gene-Time

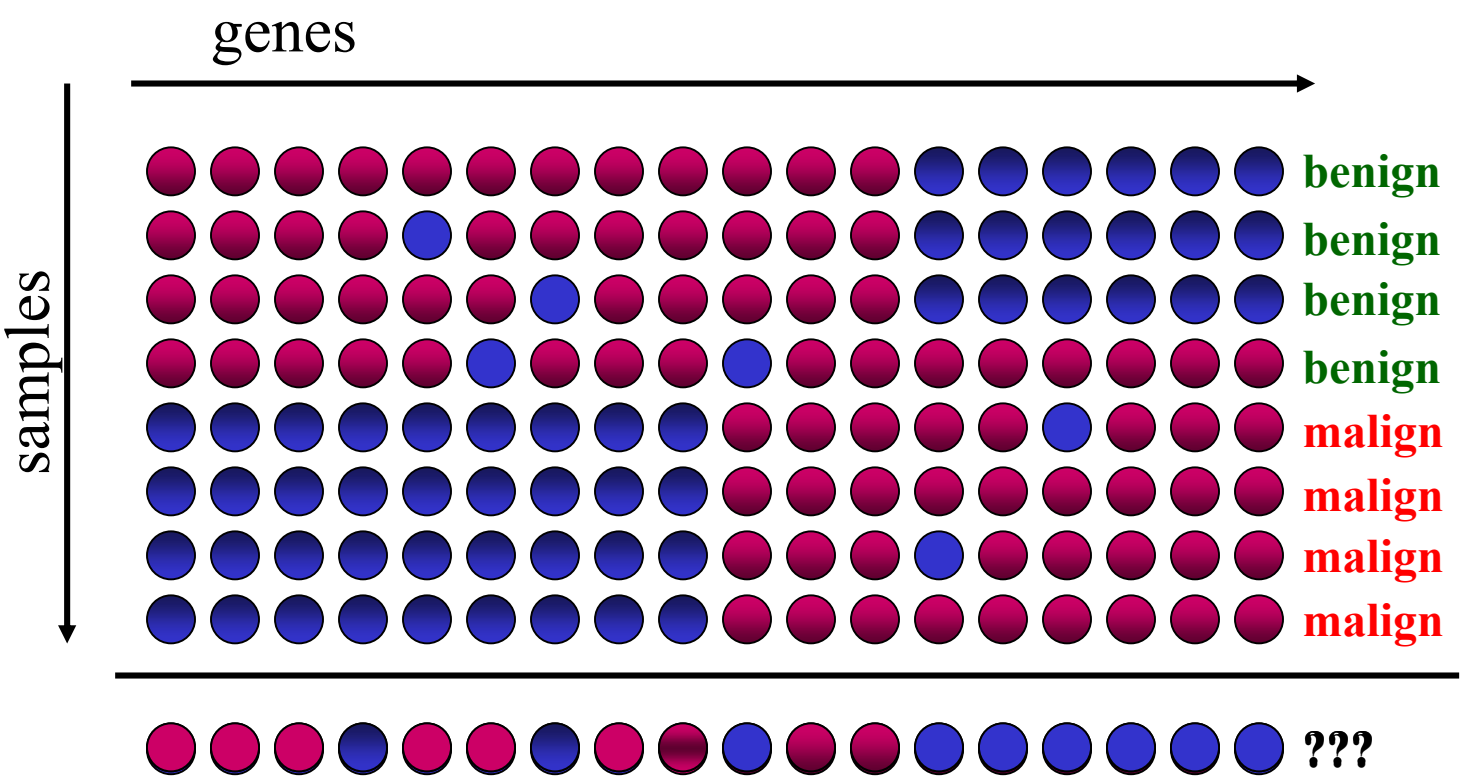


- Gene-Sample-Time



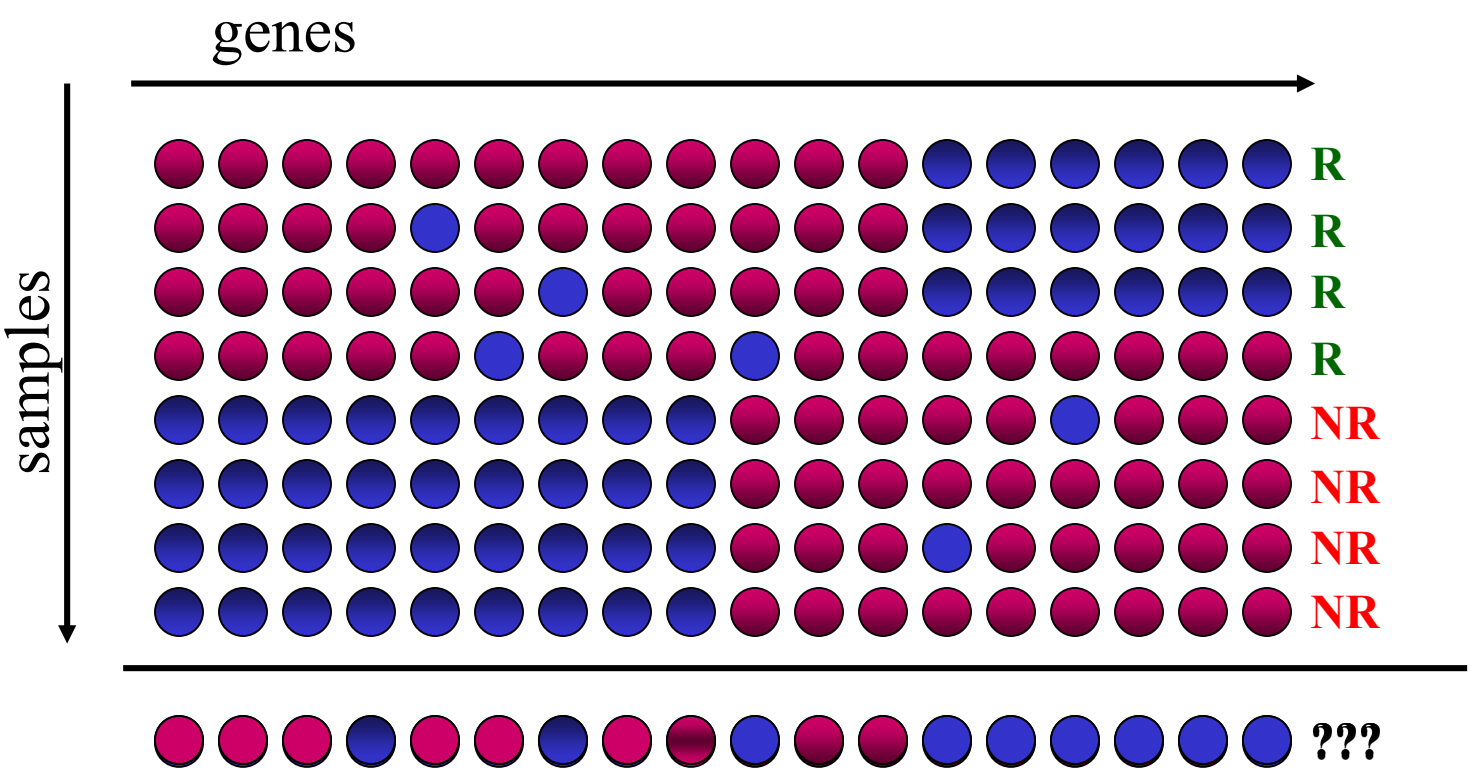
Copyright 2019 © Wong Limsoon

Application: Disease subtype diagnosis



Copyright 2019 © Wong Limsoon

Application: Treatment prognosis



Copyright 2019 © Wong Limsoon

Type of gene expression datasets

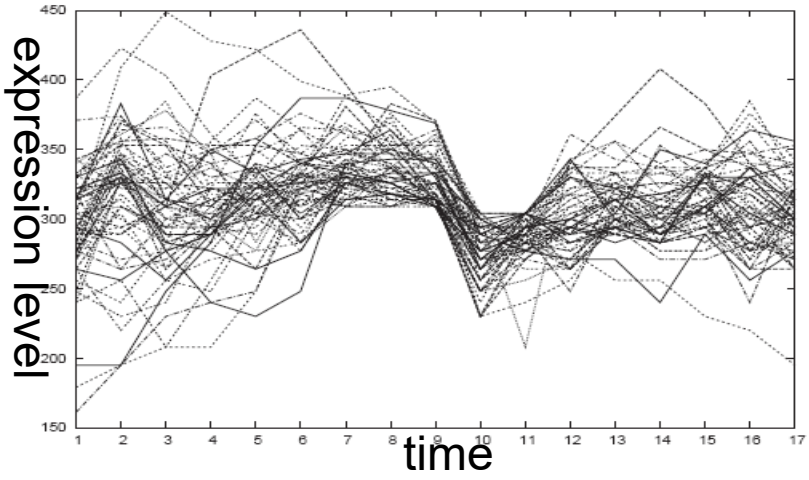
- **Gene-Conditions** or Gene-Sample (**numeric** or discretized)

1000 - 100,000 columns

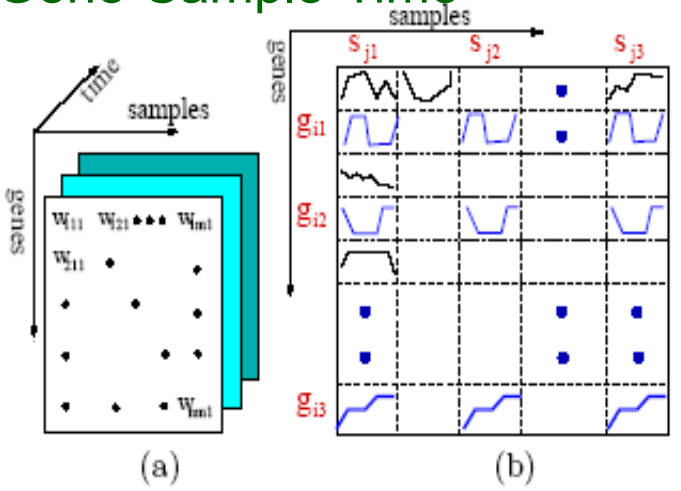
100-500 rows

	Gene1	Gene2	Gene3	Gene 4	Gene5	Gene6	Gene7		
Cond1	0.12	-1.3	1.7	1.0	-3.2	0.78	-0.12		
Cond2							1.3		
CondN									

- Gene-Time

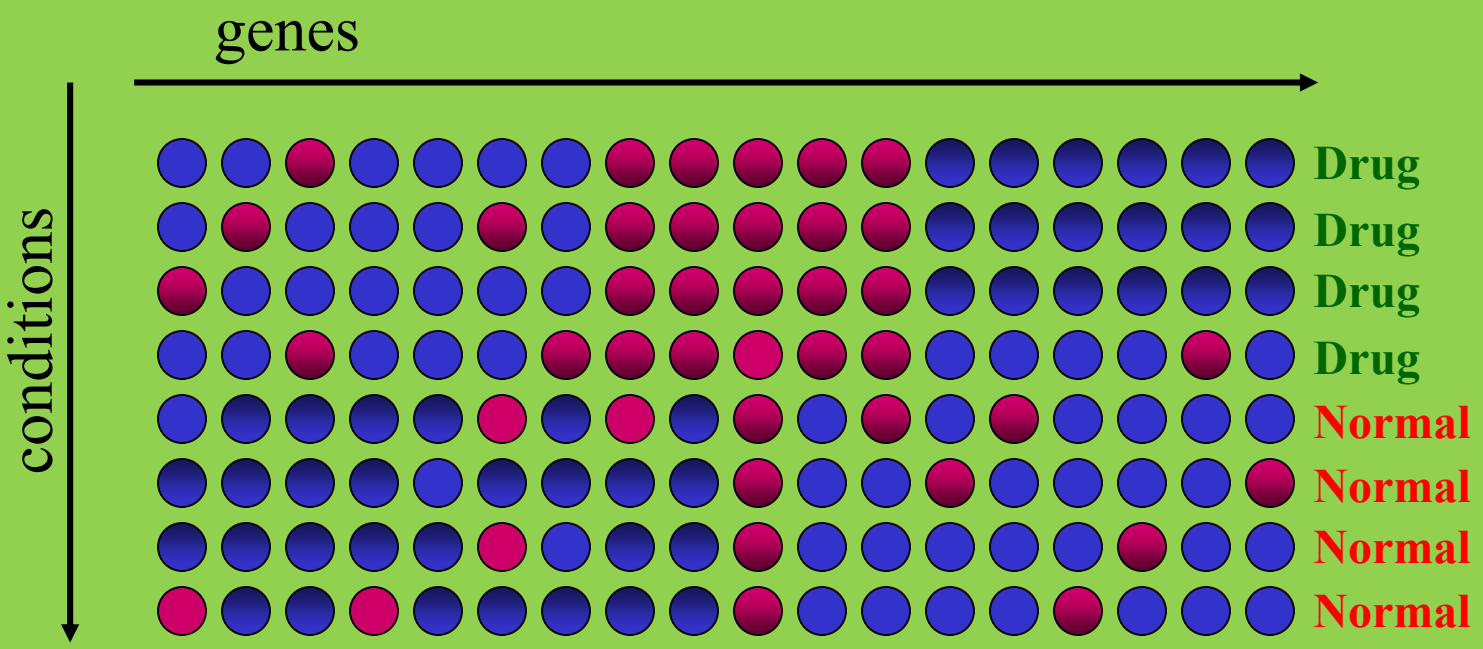


- Gene-Sample-Time



Copyright 2019 © Wong Limsoon

Application: Drug-action detection



- **Which group of genes does the drug affect? Why?**

Exercise #1

Copyright © Wong Limsoon

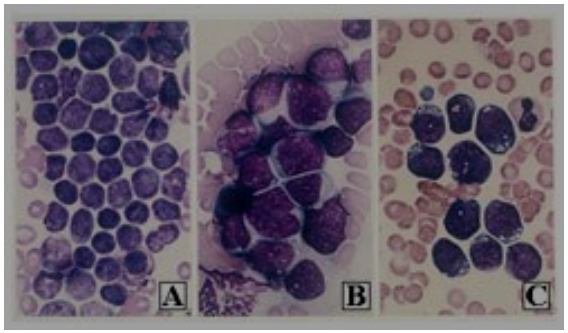
Gene expression profile classification

Childhood acute lymphoblastic leukemia subtype diagnosis

Childhood ALL

- **Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid>50**
- **Diff subtypes respond differently to same Tx**
- **Over-intensive Tx**
  - Development of secondary cancers
  - Reduction of IQ
- **Under-intensiveTx**
  - Relapse

- **The subtypes look similar**



- **Conventional diagnosis**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics
- **Unavailable in most ASEAN countries**

Copyright 2019 © Wong Limsoon

Mission

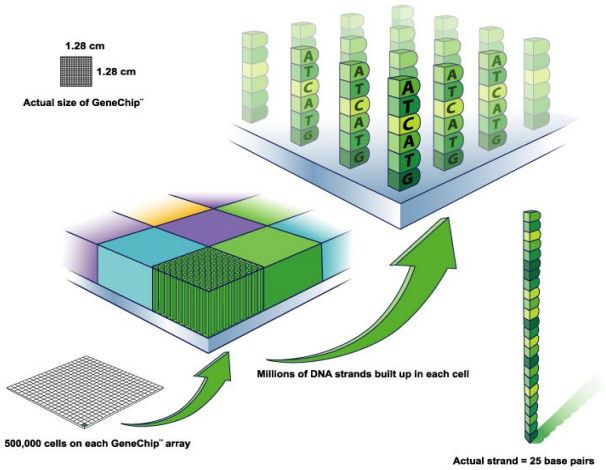
- **Conventional risk assignment procedure requires difficult expensive tests and collective judgement of multiple specialists**

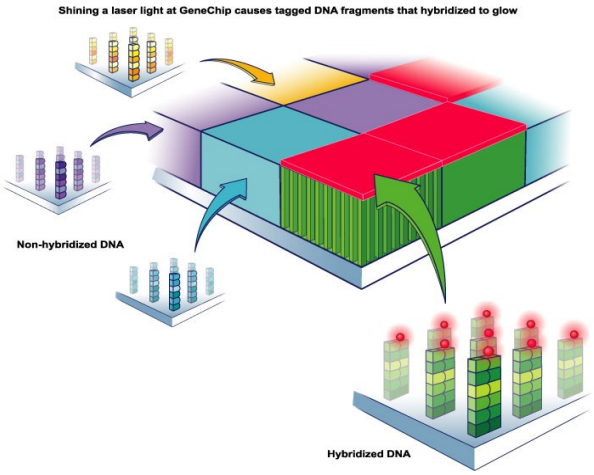
- **Generally available only in major advanced hospitals**

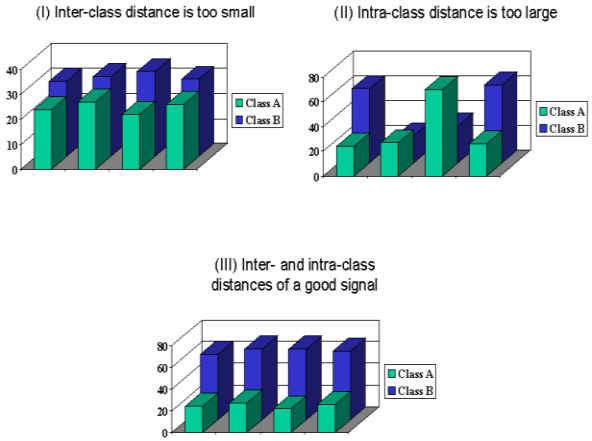
⇒ **Can we have a single-test easy-to-use platform instead?**

Copyright 2019 © Wong Limsoon

Single-test platform of microarray & machine learning



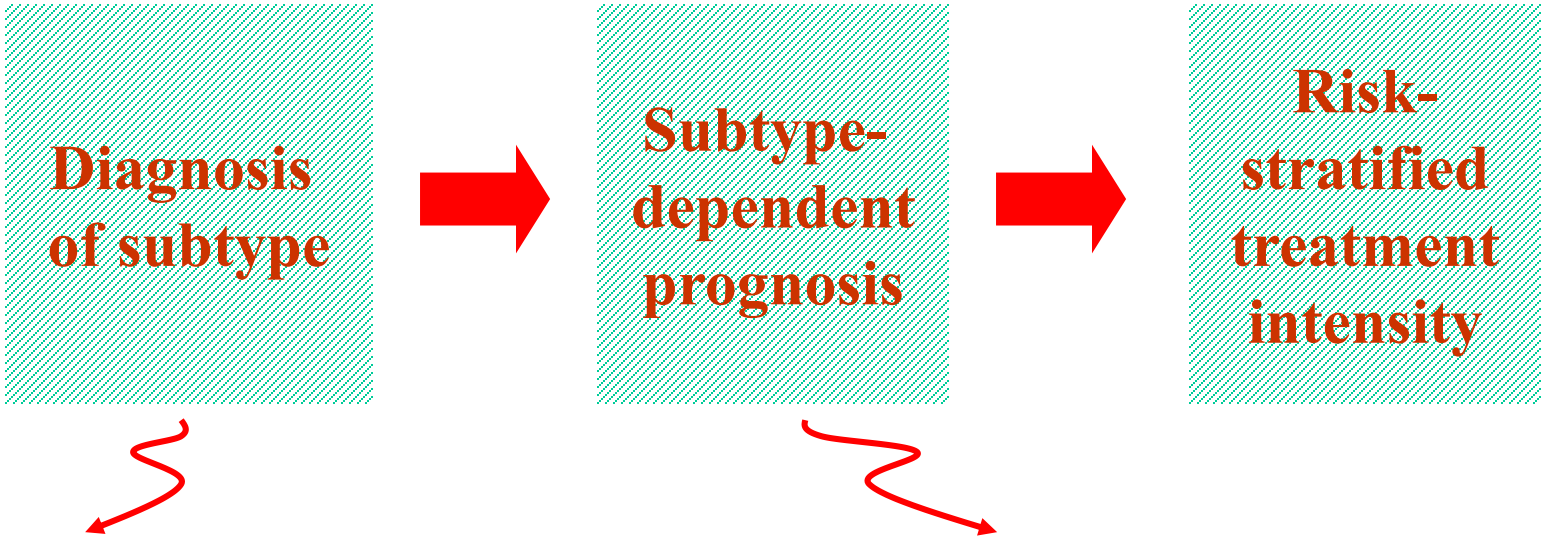




	00-0586-U	00-0586-U	00-0586-U	00-0586-U	00-0586-U	Descriptions
	Positive	Negative	Pairs InAv	Avg Diff	Abs Call	
AFFX-Murl	5	2	19	297.5	A	M16762 Mouse int
AFFX-Murl	3	2	19	554.2	A	M37897 Mouse int
AFFX-Murl	4	2	19	308.6	A	M25892 Mus musc
AFFX-Murf	1	3	19	141	A	M83649 Mus musc
AFFX-BioE	13	1	19	9340.6	P	J04423 E coli bioB
AFFX-BioE	15	0	19	12862.4	P	J04423 E coli bioB
AFFX-BioE	12	0	19	8716.5	P	J04423 E coli bioB
AFFX-BioC	17	0	19	25942.5	P	J04423 E coli bioC
AFFX-BioC	16	0	20	28838.5	P	J04423 E coli bioC
AFFX-BioD	17	0	19	25765.2	P	J04423 E coli bioD
AFFX-BioD	19	0	20	140113.2	P	J04423 E coli bioD
AFFX-CreX	20	0	20	280036.6	P	X03453 Bacterioph
AFFX-CreX	20	0	20	401741.8	P	X03453 Bacterioph
AFFX-BioE	7	5	18	-483	A	J04423 E coli bioB
AFFX-BioE	5	4	18	313.7	A	J04423 E coli bioB
AFFX-BioE	7	6	20	-1016.2	A	J04423 E coli bioB

Copyright 2019 © Wong Limsoon

Overall strategy



- For each subtype, select genes to develop classification model for diagnosing that subtype
- For each subtype, select genes to develop prediction model for prognosis of that subtype

Copyright 2019 © Wong Limsoon

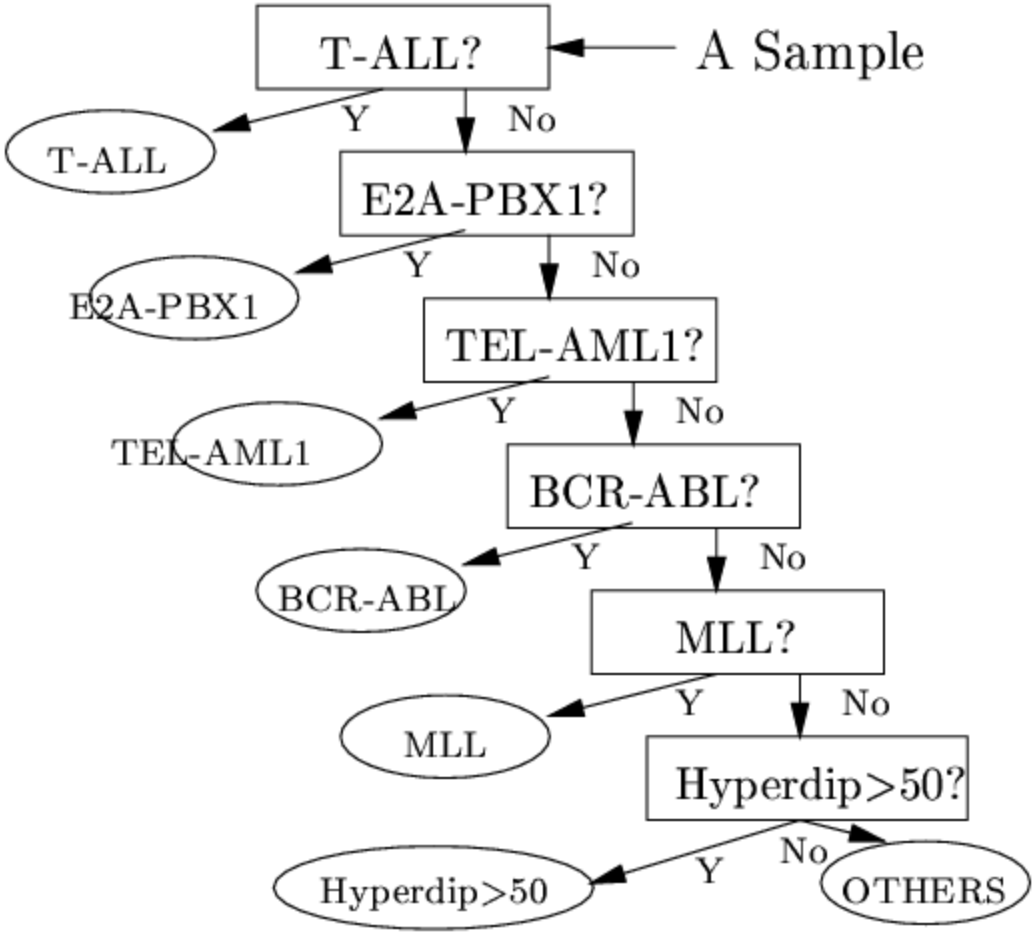
Subtype diagnosis by PCL

- Gene expression data collection
- Gene selection by $\chi 2$
- Classifier training by emerging pattern
- ~~Classifier tuning (optional for some machine learning methods)~~
- Apply classifier for diagnosis of future cases by PCL

Copyright 2019 © Wong Limsoon

Childhood ALL subtype diagnosis workflow

A tree-structured diagnostic workflow was recommended by our doctor collaborator



Copyright 2019 © Wong Limsoon

Training and testing sets

Paired datasets	Ingredients	Training	Testing
T-ALL vs OTHERS1	OTHERS1 ={E2A-PBX1, TEL-AML1, BCR-ABL, Hyperdip>50, MLL, OTHERS}	28 vs 187	15 vs 97
E2A-PBX1 vs OTHERS2	OTHERS2 = {TEL-AML1, BCR-ABL Hyperdip>50, MLL, OTHERS}	18 vs 169	9 vs 88
TEL-AML1 vs OTHERS3	OTHERS3 = {BCR-ABL Hyperdip>50, MLL, OTHERS}	52 vs 117	27 vs 61
BCR-ABL vs OTHERS4	OTHERS4 = {Hyperdip>50, MLL, OTHERS}	9 vs 108	6 vs 55
MLL vs OTHERS5	OTHERS5 = {Hyperdip>50, OTHERS}	14 vs 94	6 vs 49
Hyperdip>50 vs OTHERS	OTHERS = {Hyperdip47-50, Pseudodip, Hypodip, Normo}	42 vs 52	22 vs 27

Copyright 2019 © Wong Limsoon

Signal selection basic idea

- **Choose a signal w/ low intra-class distance**
- **Choose a signal w/ high inter-class distance**

Copyright 2019 © Wong Limsoon

Signal selection by $\chi 2$

The $\mathcal{X}^2$ value of a signal is defined as:

$$\mathcal{X}^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}},$$

where m is the number of intervals, k the number of classes, A_{ij} the number of samples in the ith interval, jth class, R_i the number of samples in the ith interval, C_j the number of samples in the jth class, N the total number of samples, and E_{ij} the expected frequency of A_{ij} ($E_{ij} = R_i * C_j / N$).

Copyright 2019 © Wong Limsoon

Emerging patterns

- **An emerging pattern is a set of conditions**
  - usually involving several features
  - that most members of a class satisfy
  - but none or few of the other class satisfy
- **A jumping emerging pattern is an emerging pattern that**
  - some members of a class satisfy
  - but no members of the other class satisfy
- **We use only jumping emerging patterns**

Copyright 2019 © Wong Limsoon

Examples

Patterns	Frequency (P)	Frequency(N)
{9, 36}	38 instances	0
{9, 23}	38	0
{4, 9}	38	0
{9, 14}	38	0
{6, 9}	38	0
{7, 21}	0	36
{7, 11}	0	35
{7, 43}	0	35
{7, 39}	0	34
{24, 29}	0	34

Easy interpretation

Reference number 9: the expression of gene 37720_at > 215
Reference number 36: the expression of gene 38028_at ≤ 12

Copyright 2019 © Wong Limsoon

PCL: Prediction by Collective Likelihood

- Let $EP_1^P, \ldots, EP_i^P$ be the most general EPs of D^P in descending order of support.
- Suppose the test sample T contains these most general EPs of D^P (in descending order of support):

$$EP_{i_1}^P, EP_{i_2}^P, \cdots, EP_{i_x}^P$$

- Use k top-ranked most general EPs of D^P and D^N. Define the score of T in the D^P class as

$$score(T, D^P) = \sum_{m=1}^{k} \frac{frequency(EP_{i_m}^P)}{frequency(EP_m^P)}$$

- Ditto for $score(T, D^N)$.
- If $score(T, D^P) > score(T, D^N)$, then T is class P. Otherwise it is class N.

Copyright 2019 © Wong Limsoon

PCL learning

Top-Ranked EPs in Positive class	Top-Ranked EPs in Negative class
EP_1^P (90%)	EP_1^N (100%)
EP_2^P (86%)	EP_2^N (95%)
.	.
.	.
EP_n^P (68%)	EP_n^N (80%)

The idea of summarizing multiple top-ranked EPs is intended to avoid some rare tie cases

Copyright 2019 © Wong Limsoon

PCL testing

Most freq EP of pos class in the test sample

$$\text{Score}^{P} = EP_1^{P'} / EP_1^{P} + \ldots + EP_k^{P'} / EP_k^{P}$$

Most freq EP of pos class

Similarly,

$$\text{Score}^{N} = EP_1^{N'} / EP_1^{N} + \ldots + EP_k^{N'} / EP_k^{N}$$

If $\text{Score}^{P} > \text{Score}^{N}$, then positive class, Otherwise negative class

Copyright 2019 © Wong Limsoon

Accuracy of PCL (vs. other classifiers)

Testing Data	Error rate of different models			
	C4.5	SVM	NB	PCL
T-ALL vs OTHERS1	0:1	0:0	0:0	0:0
E2A-PBX1 vs OTHERS2	0:0	0:0	0:0	0:0
TEL-AML1 vs OTHERS3	1:1	0:1	0:1	1:0
BCR-ABL vs OTHERS4	2:0	3:0	1:4	2:0
MLL vs OTHERS5	0:1	0:0	0:0	0:0
Hyperdiploid>50 vs OTHERS	2:6	0:2	0:2	0:1
Total Errors	14	6	8	4

The classifiers are all applied to the 20 genes selected by $\chi 2$ at each level of the tree

Copyright 2019 © Wong Limsoon

Understandability of PCL

- **E.g., for T-ALL vs. OTHERS, one ideally discriminatory gene 38319_at was found, inducing these 2 EPs**

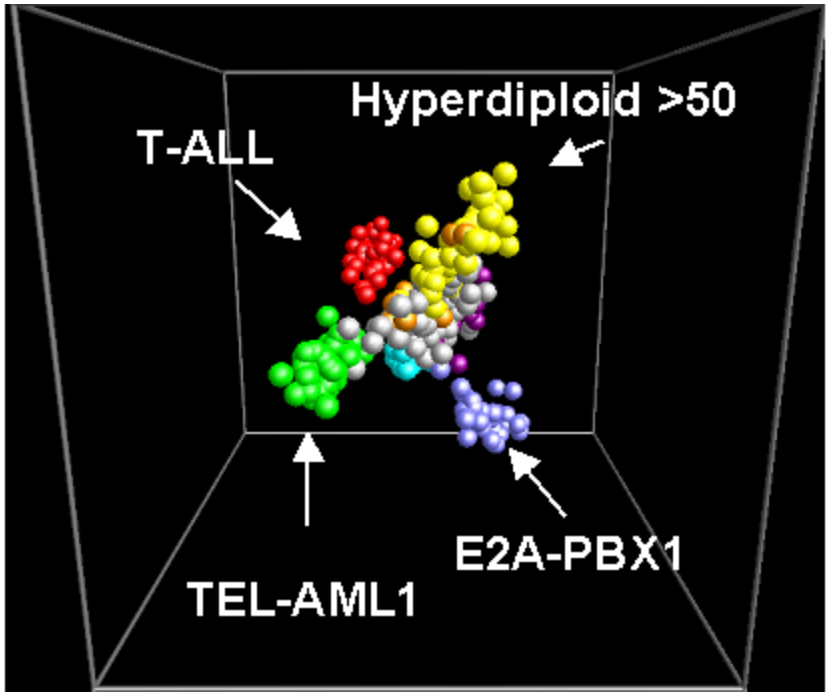
$$\{gene_(38\,319_at)\,@\,(-\infty,\ 15\,975.6)\} \text{ and}$$
$$\{gene_(38\,319_at)\,@\,[15\,975.6,\ +\infty)\}.$$

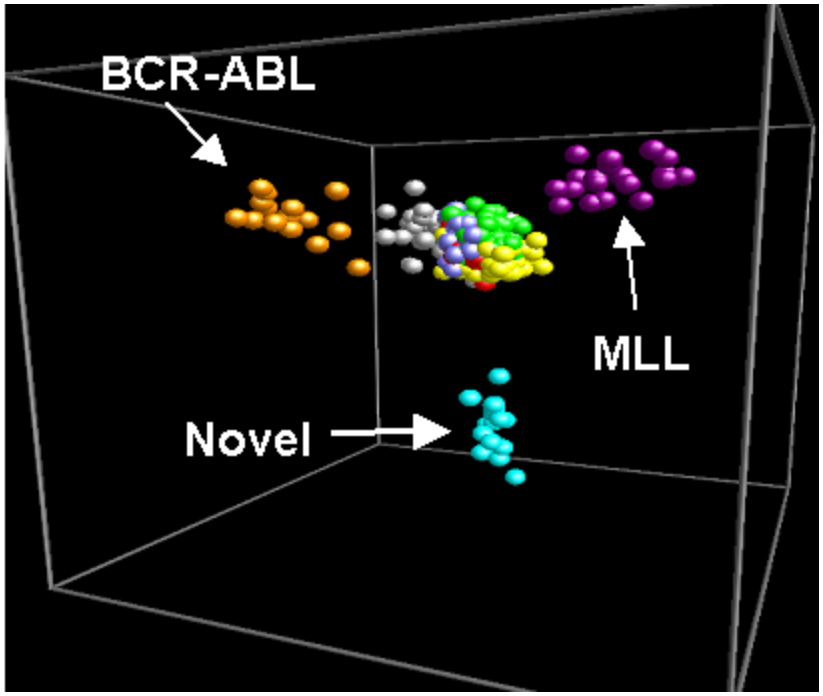
- **These give us the diagnostic rule**

If the expression of 38 319_*at* is less than 15 975.6, then
this ALL sample must be a T-ALL.
Otherwise it must be a subtype in OTHERS1.

Copyright 2019 © Wong Limsoon

Multidimensional scaling plot for subtype diagnosis

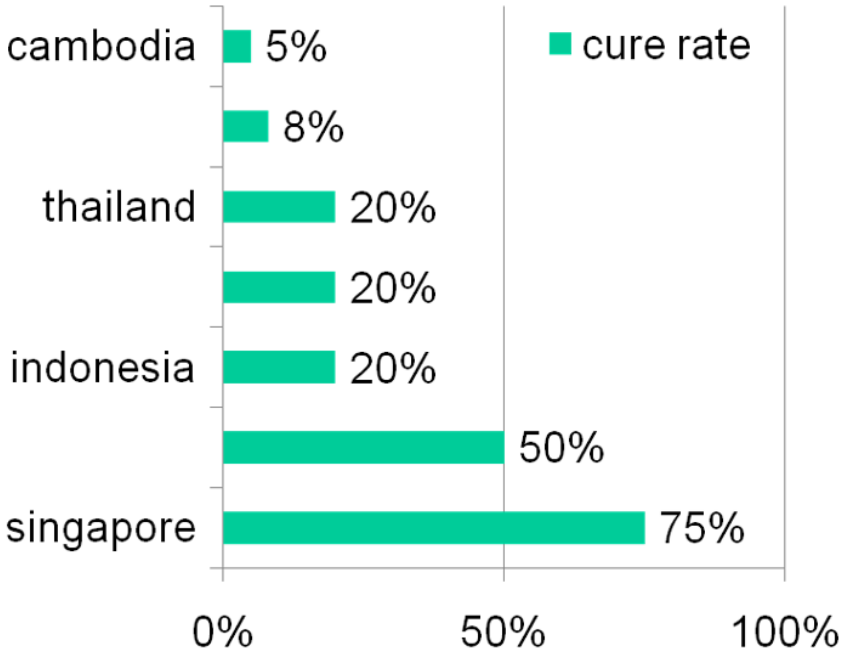




Obtained by performing PCA on the 20 genes chosen for each level

Copyright 2019 © Wong Limsoon

Childhood ALL cure rates



- **Conventional risk assignment procedure requires difficult expensive tests and collective judgement of multiple specialists**

⇒ **Not available in less advanced ASEAN countries**

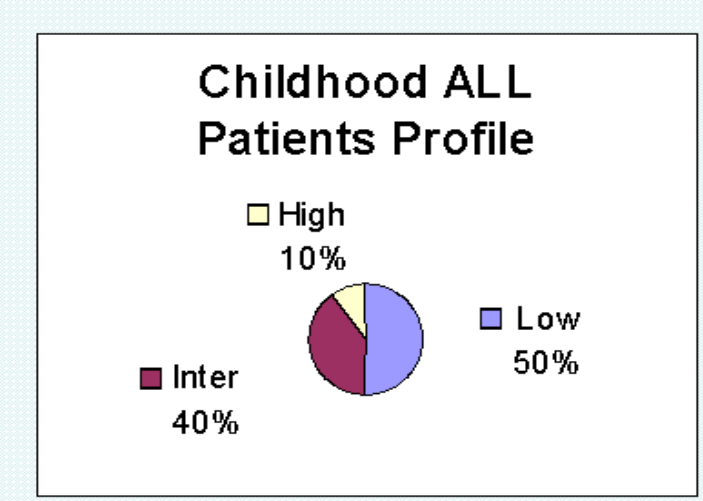
Copyright 2019 © Wong Limsoon

Childhood ALL treatment cost

- **Treatment for childhood ALL over 2 yrs**
  - Intermediate intensity: US$60k
  - Low intensity: US$36k
  - High intensity: US$72k
- **Treatment for relapse: US$150k**
- **Cost for side-effects: Unquantified**

Copyright 2019 © Wong Limsoon

Current situation (2000 new cases / yr in ASEAN)



- Intermediate intensity conventionally applied in less advanced ASEAN countries

- Over intensive for 50% of patients, thus more side effects
- Under intensive for 10% of patients, thus more relapse

- US$120m (US$60k * 2000) for intermediate intensity tx
- US$30m (US$150k * 2000 * 10%) for relapse tx
- Total US$150m/yr plus un-quantified costs for dealing with side effects

Copyright 2019 © Wong Limsoon

Using our platform

- Low intensity applied to 50% of patients
- Intermediate intensity to 40% of patients
- High intensity to 10% of patients

⇒ Reduced side effects

⇒ Reduced relapse

⇒ 75-80% cure rates

- US$36m (US$36k * 2000 * 50%) for low intensity
- US$48m (US$60k * 2000 * 40%) for intermediate intensity
- US$14.4m (US$72k * 2000 * 10%) for high intensity

- Total US$98.4m/yr

⇒ Save US$51.6m/yr

Copyright 2019 © Wong Limsoon

A nice ending…

- **Asian Innovation Gold Award 2003**

The Winners 2003

ALL is a heterog

SOLID GOLD: Allen Yeoh (*front*) with teammates (*from left*) Huiqing Liu, Limsoon Wong and Jinyan Li

GOLD AWARD

Kids' Cancer Heroes

A Singapore team has invented a simple test for childhood leukaemia that promises safer treatments and higher cure rates for kids in the developing world

By Trish Saywell/SINGAPORE

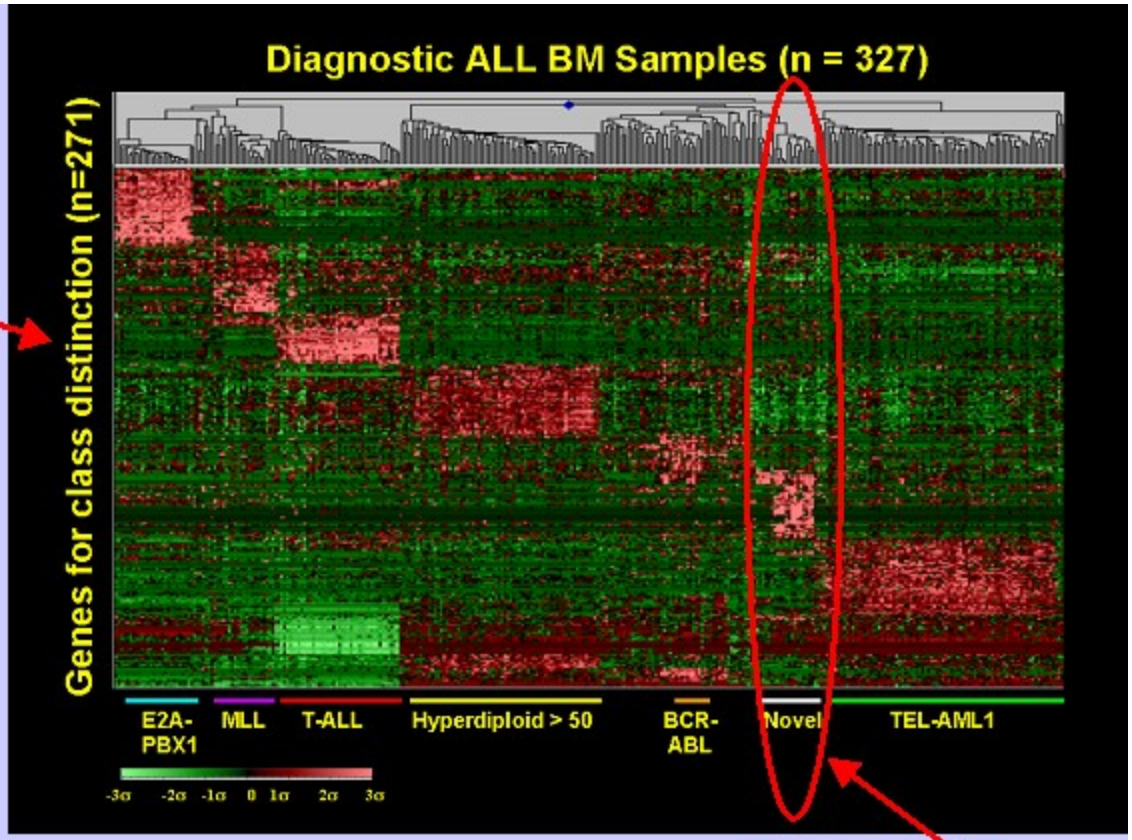
Photos by Munshi Ahmed/REVIEW

Copyright 2019 © Wong Limsoon

Gene expression profile clustering

Novel disease subtype discovery

Is there a new subtype?



Genes selected by $\chi 2$

- **Hierarchical clustering of gene expression profiles reveals a novel subtype of childhood ALL**

New subtype discovered

Exercise: Name and describe one bi-clustering method

Copyright 2019 © Wong Limsoon

Hierarchical clustering

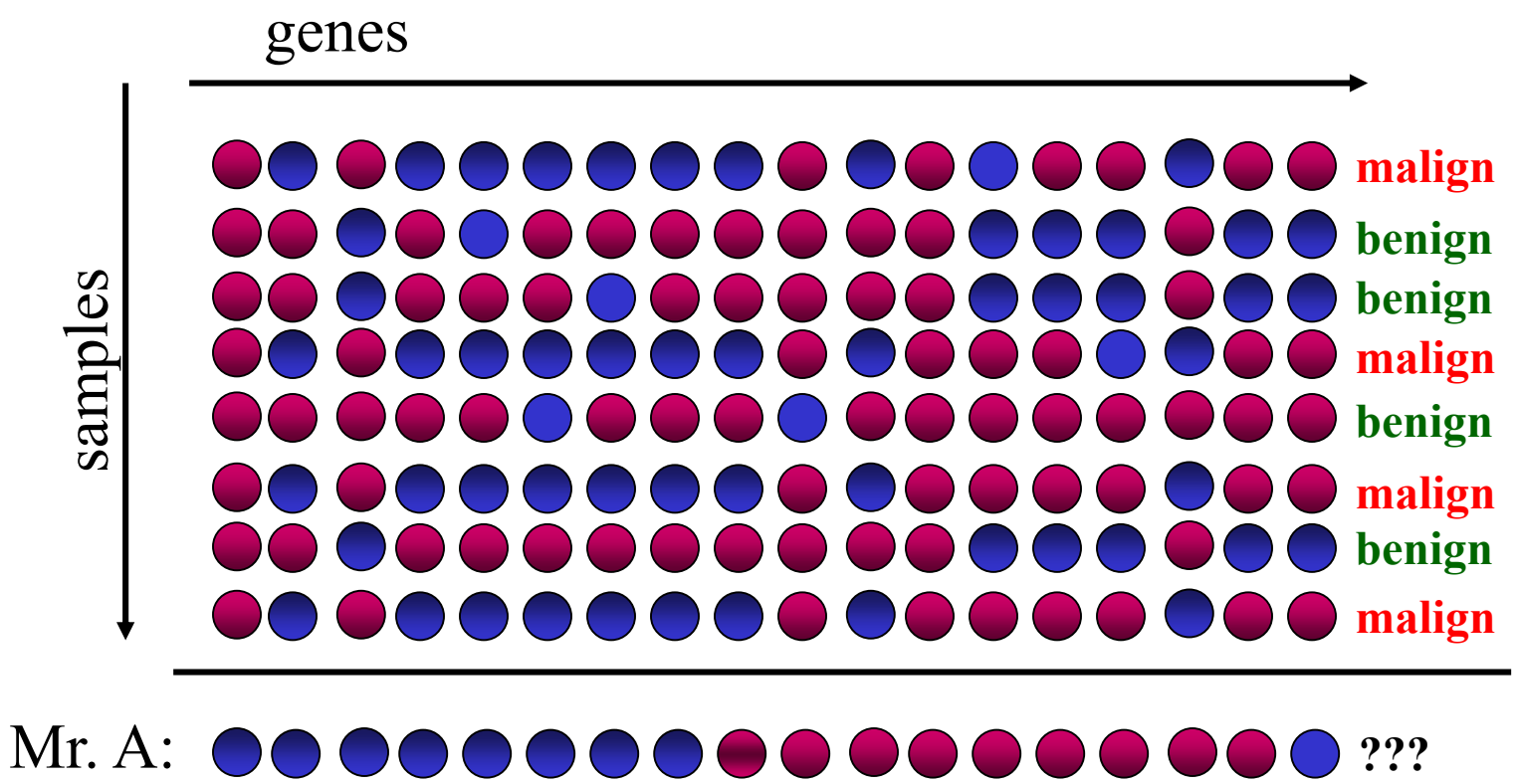
- **Assign each item to its own cluster**
  - If there are N items initially, we get N clusters, each containing just one item
- **Find the "most similar" pair of clusters, merge them into a single cluster, so we now have one less cluster**
- **Repeat previous step until all items are clustered into a single cluster of size N**

Copyright 2019 © Wong Limsoon

Gene expression profile clustering

Diagnosis via guilt-by-association

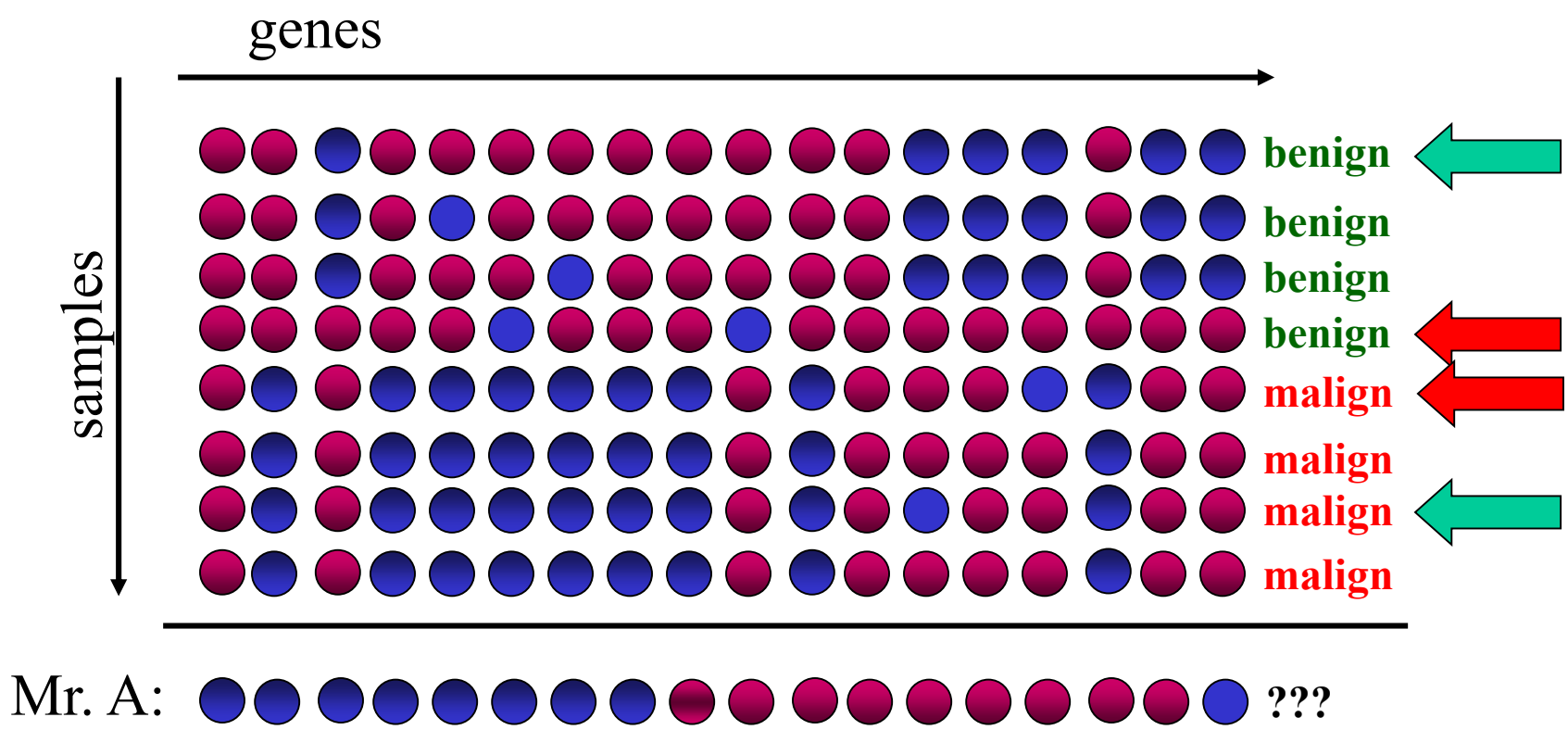
Some patient samples



- **Does Mr. A have cancer?**

Copyright 2019 © Wong Limsoon

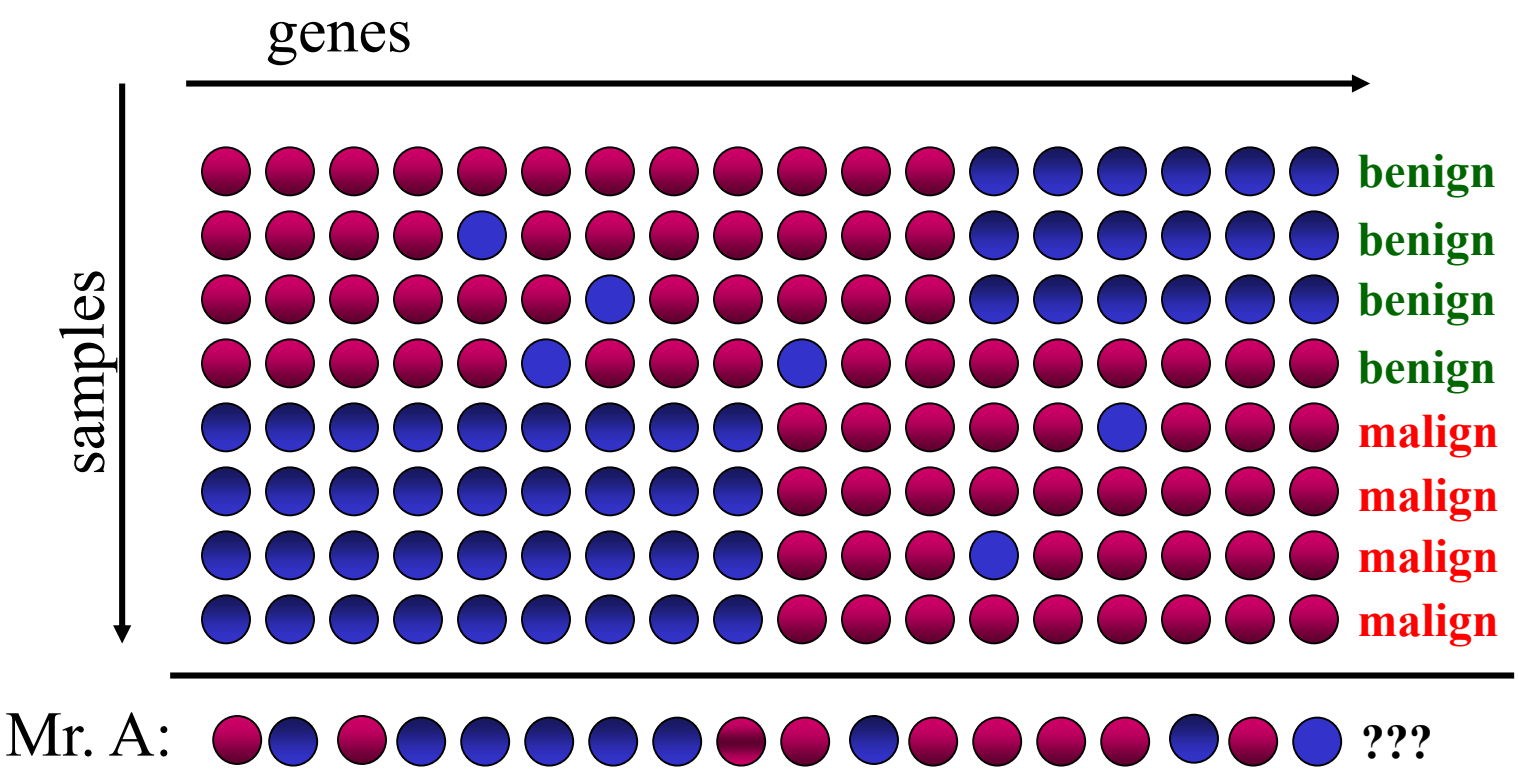
Let's rearrange the rows…



- **Does Mr. A have cancer?**

Copyright 2019 © Wong Limsoon

and the columns too…



- **Does Mr. A have cancer?**

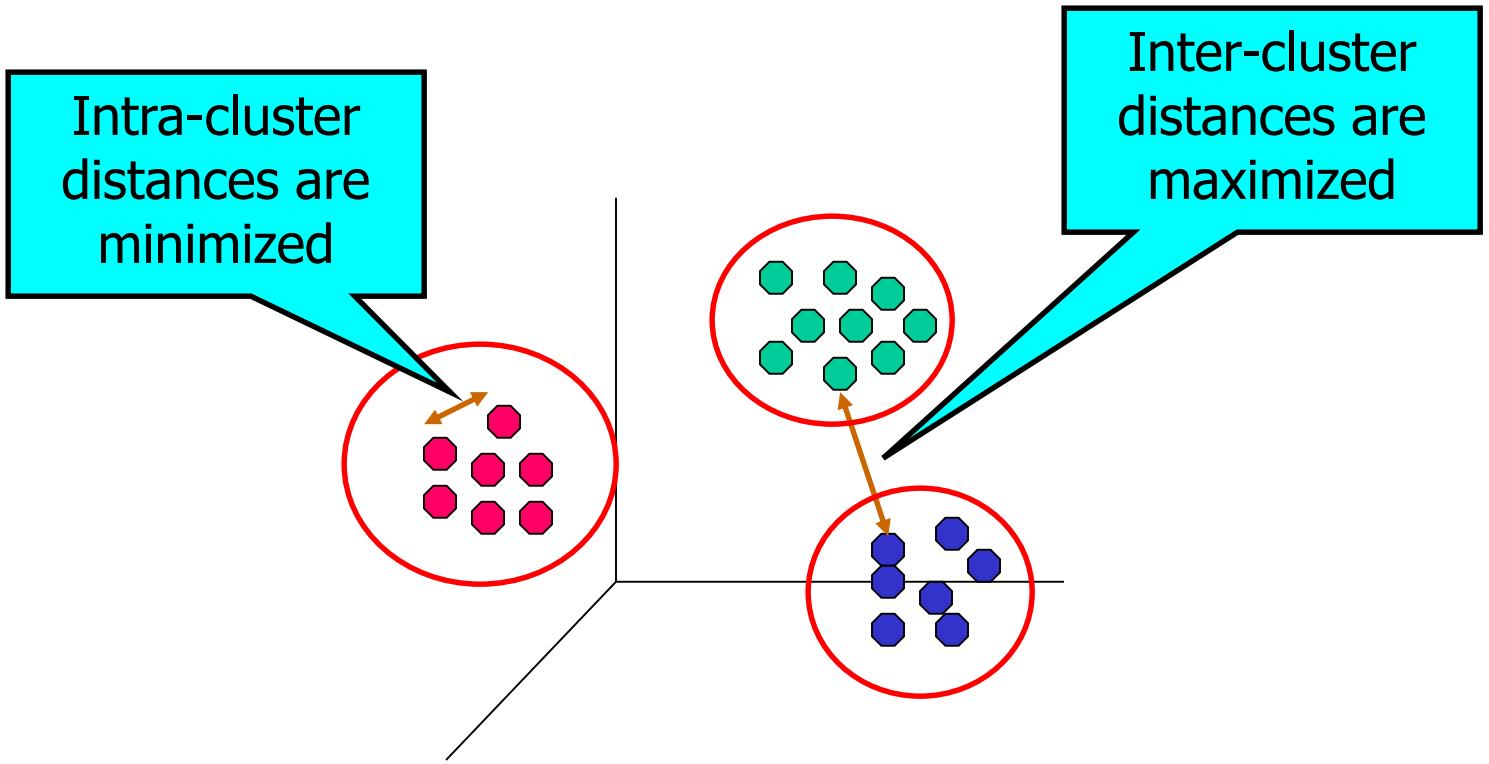
Copyright 2019 © Wong Limsoon

Introduction to simple clustering methods

NUS
National University
of Singapore

What is cluster analysis?

- **Finding groups of objects such that objects in a group are similar to one another and different from objects in other groups**



Copyright 2019 © Wong Limsoon

Notion of a cluster can be ambiguous

How many clusters?

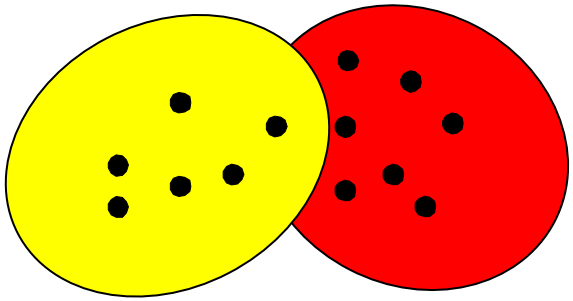
Six Clusters

Two Clusters

Four Clusters

Copyright 2019 © Wong Limsoon

We can also have



Copyright 2019 © Wong Limsoon

K-means clustering

- **Partitional clustering approach**
- **Each cluster is associated with a centroid**
- **Each point is assigned to the cluster with the closest centroid**
- **# of clusters, K, must be specified**

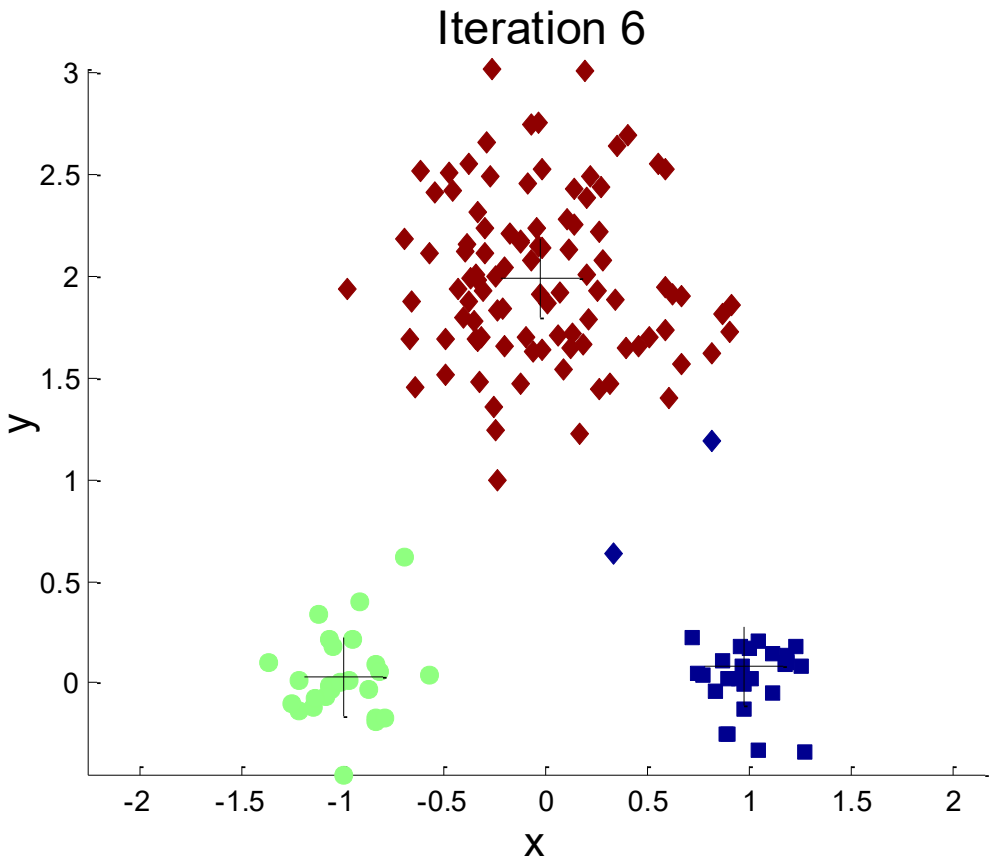
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Assignment

Update

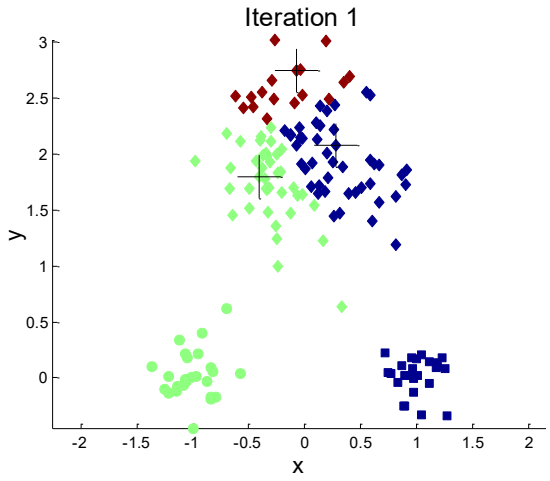
Copyright 2019 © Wong Limsoon

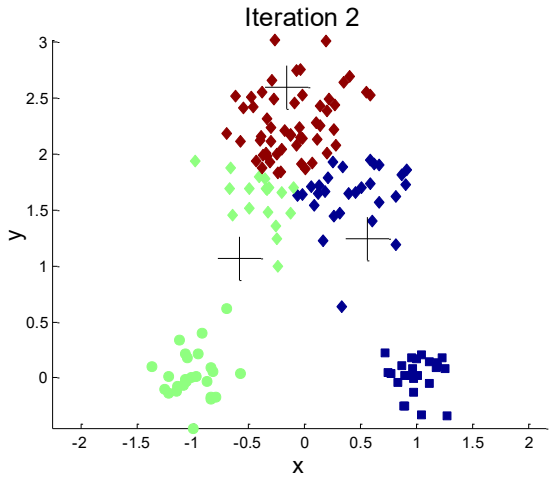
K-means clustering illustration

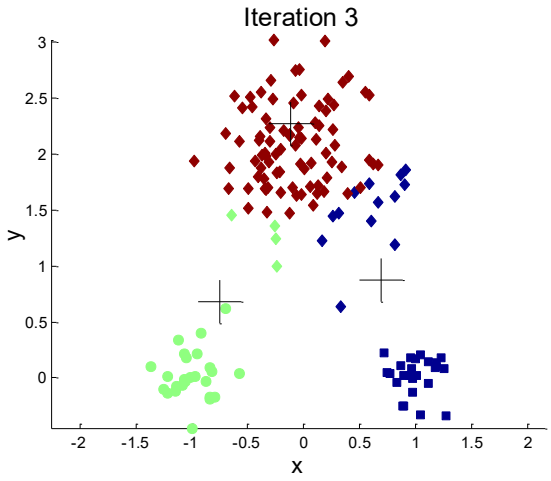


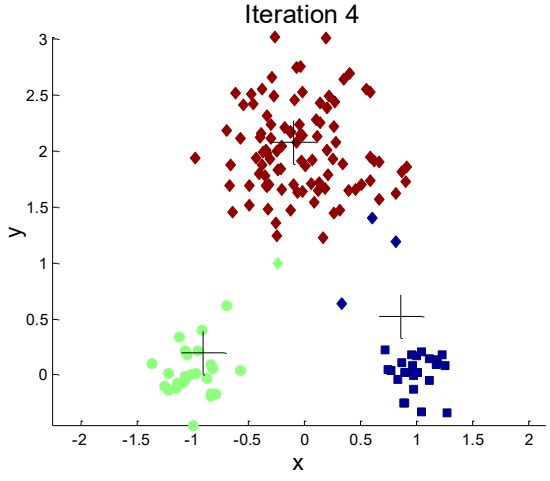
Copyright 2019 © Wong Limsoon

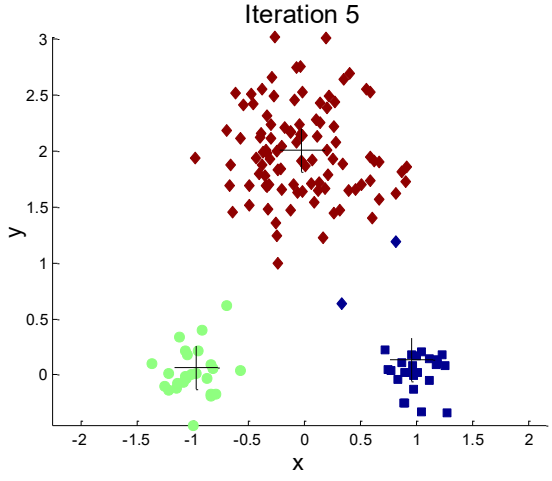
K-means clustering illustration

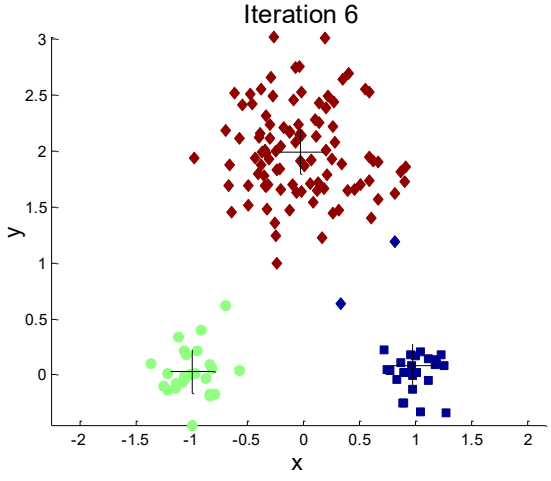








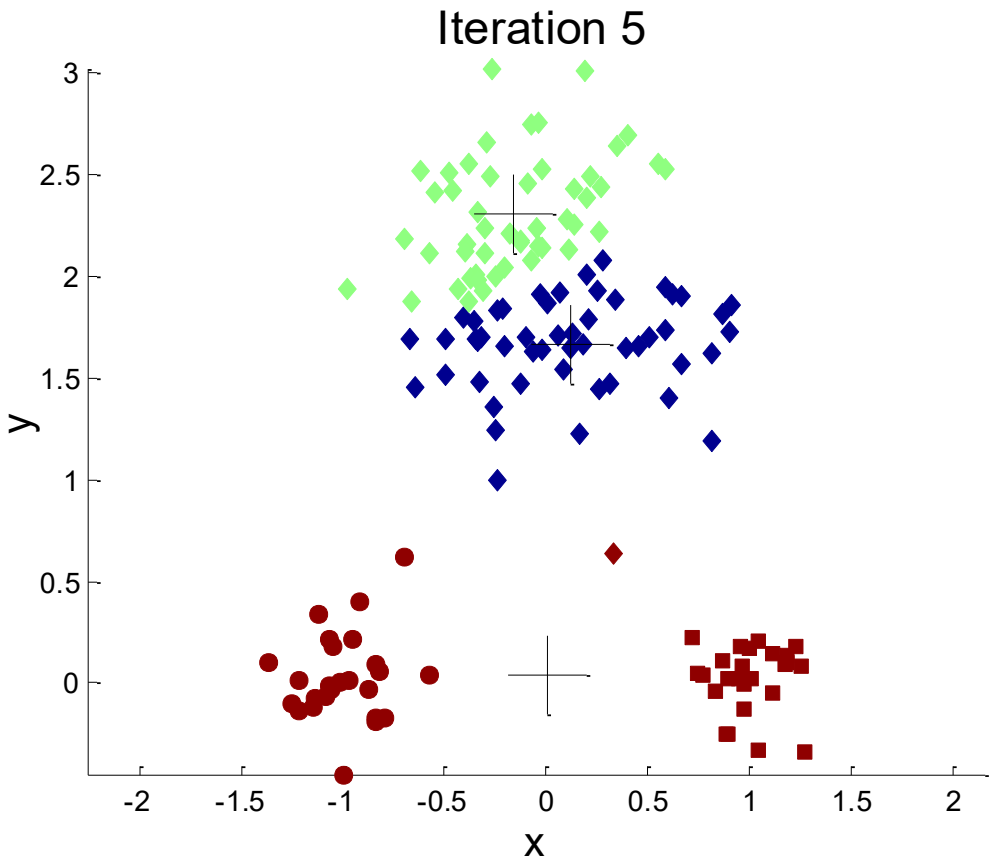




Copyright 2019 © Wong Limsoon

NUS National University of Singapore

Importance of choosing initial centroids



Copyright 2019 © Wong Limsoon

Hierarchical clustering

- **Two main types of hierarchical clustering**
  - Agglomerative:
    - **Start with the points as individual clusters**
    - **At each step, merge the closest pair of clusters until only one cluster (or k clusters) left**
  - Divisive:
    - **Start with one, all-inclusive cluster**
    - **At each step, split a cluster until each cluster contains a point (or there are k clusters)**
- **Traditional hierarchical algorithms use a similarity or distance matrix**
  - Merge or split one cluster at a time

Copyright 2019 © Wong Limsoon

Agglomerative hierarchical clustering

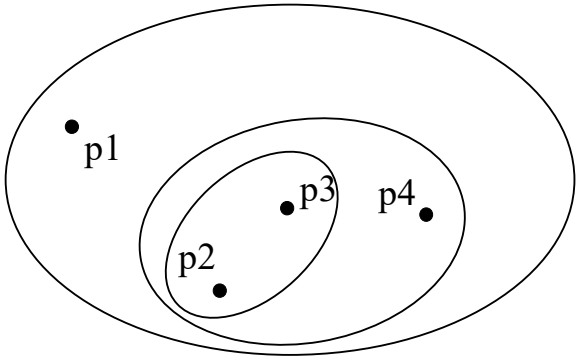
- **More popular hierarchical clustering technique**

- **Basic algorithm**
  - **Compute the proximity matrix**
  - **Let each data point be a cluster**
  - **Repeat**
    - **Merge the two closest clusters** — Merge
    - **Update the proximity matrix** — Update
  - **Until only a single cluster remains**

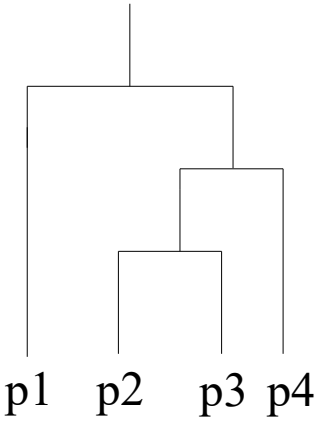
- **Key is computation of proximity of two clusters**
  - Different approaches to defining the distance / similarity between clusters

Copyright 2019 © Wong Limsoon

Visualization of agglomerative hierarchical clustering



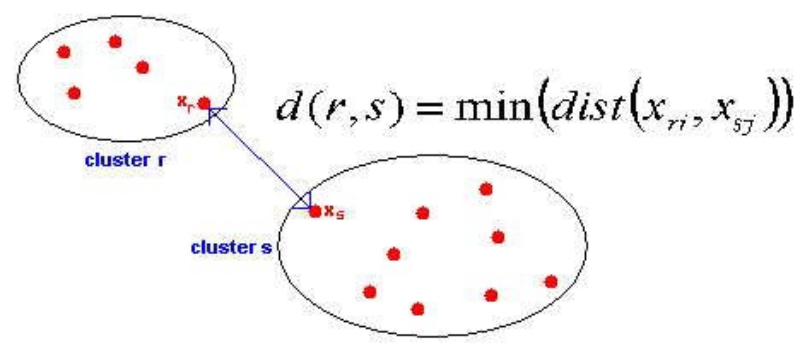
Traditional Hierarchical Clustering

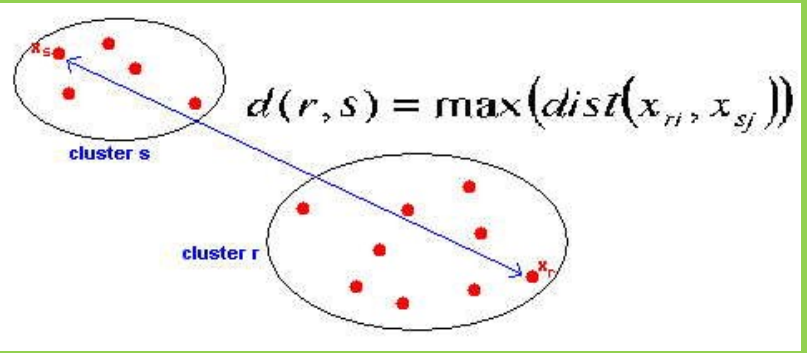


Traditional Dendrogram

Copyright 2019 © Wong Limsoon

Single, complete, & average Linkage





Single linkage defines distance betw two clusters as min distance betw them

Complete linkage defines distance betw two clusters as max distance betw them

Exercise: Give definition of "average linkage"

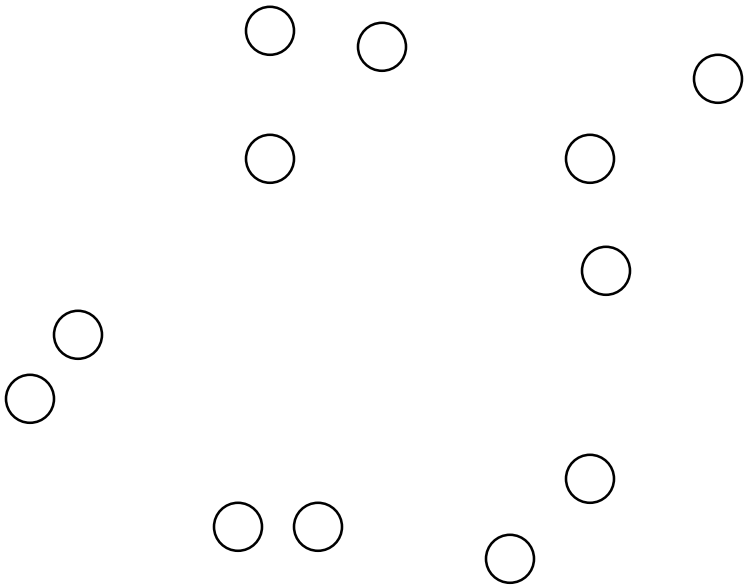
Exercise #2

Image source: UCL Microcore Website

Copyright © Wong Limsoon

Simulation: Starting situation

- **Start with clusters of individual points and a proximity matrix**



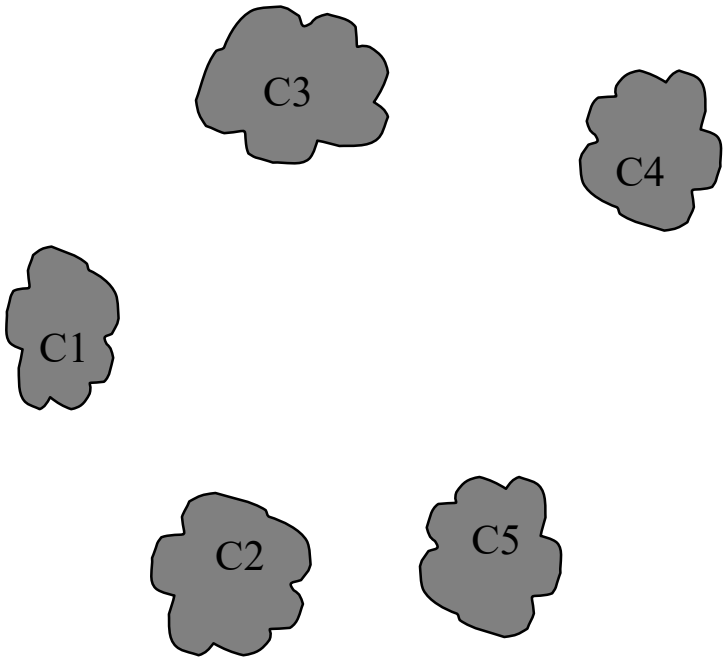
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

Copyright 2019 © Wong Limsoon

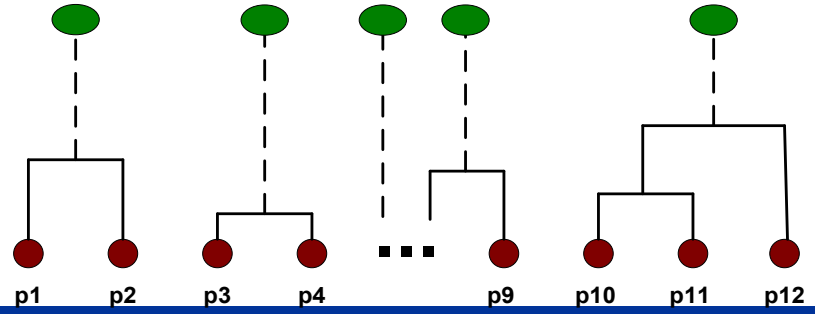
Intermediate situation

- **After some merging steps, we have some clusters**



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

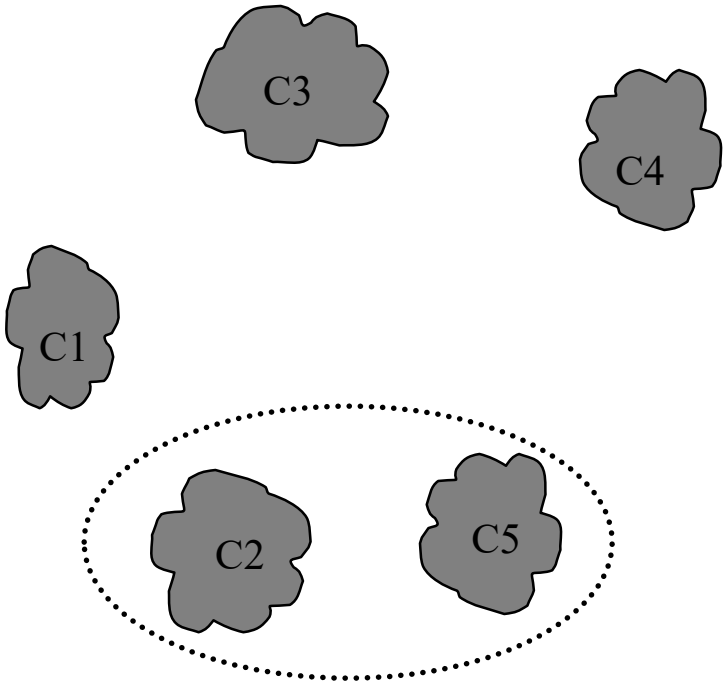
Proximity Matrix



Copyright 2019 © Wong Limsoon

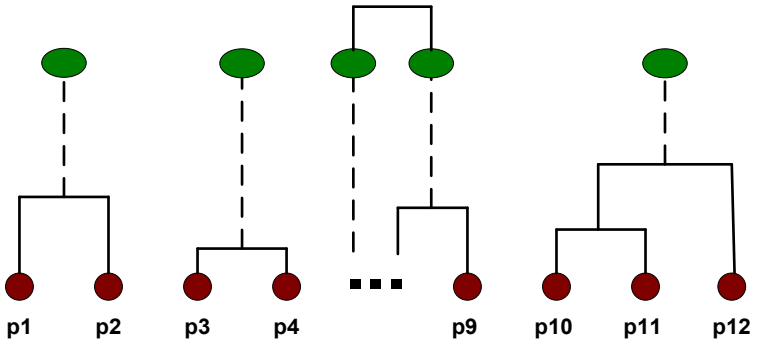
Intermediate situation

- **We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.**



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

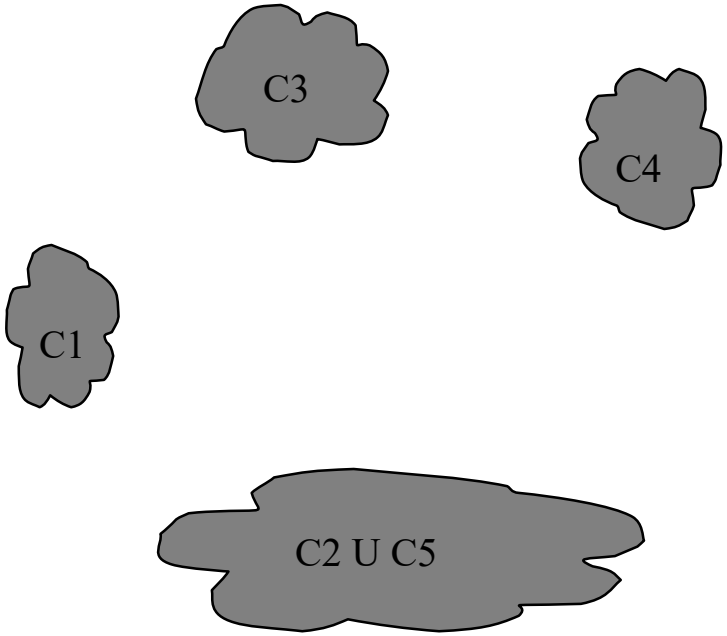
Proximity Matrix



Copyright 2019 © Wong Limsoon

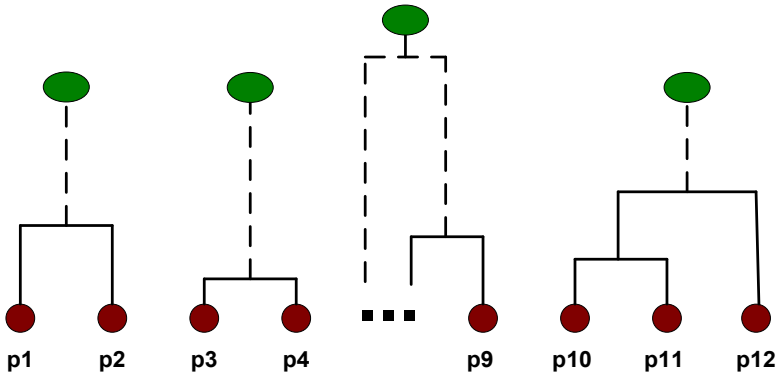
After merging

- **The question is "How do we update the proximity matrix?"**



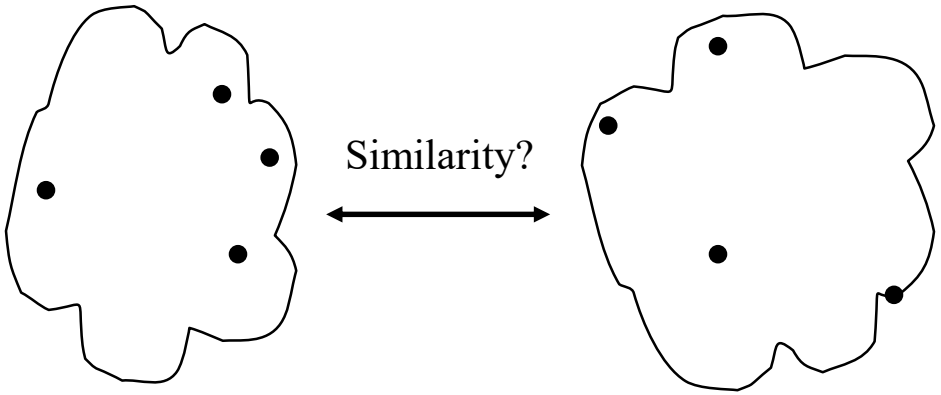
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



Copyright 2019 © Wong Limsoon

How to define inter-cluster similarity



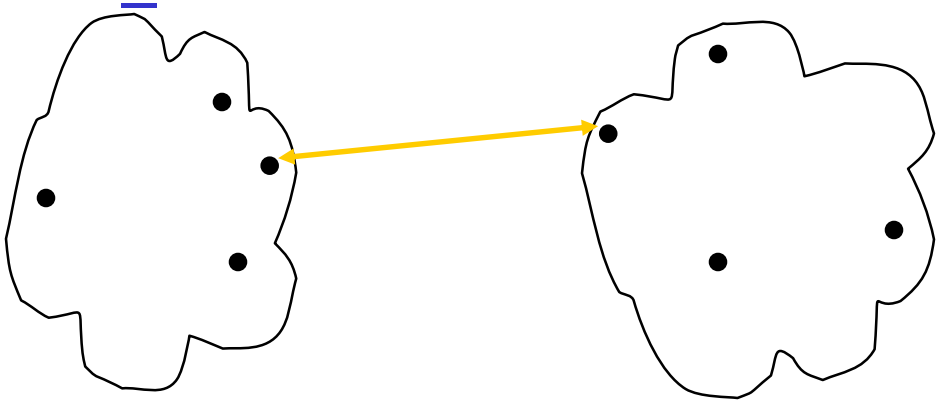
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- Min
- Max
- Group average
- Distance between centroids

Copyright 2019 © Wong Limsoon

How to define inter-cluster similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

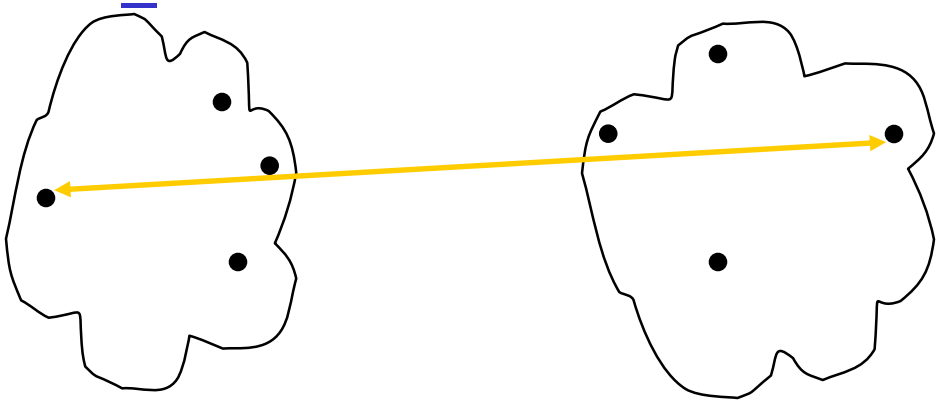
. .

. Proximity Matrix

- Min
- Max
- Group average
- Distance between centroids

Copyright 2019 © Wong Limsoon

How to define inter-cluster similarity



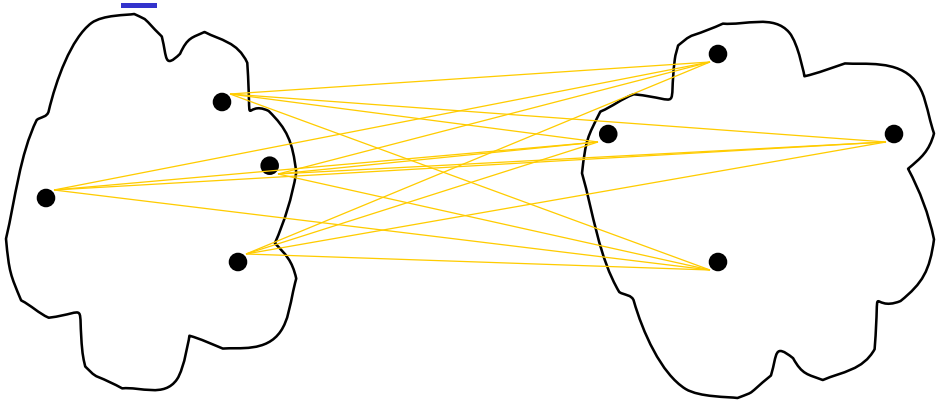
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

Proximity Matrix

- Min
- Max
- Group average
- Distance between centroids

Copyright 2019 © Wong Limsoon

How to define inter-cluster similarity



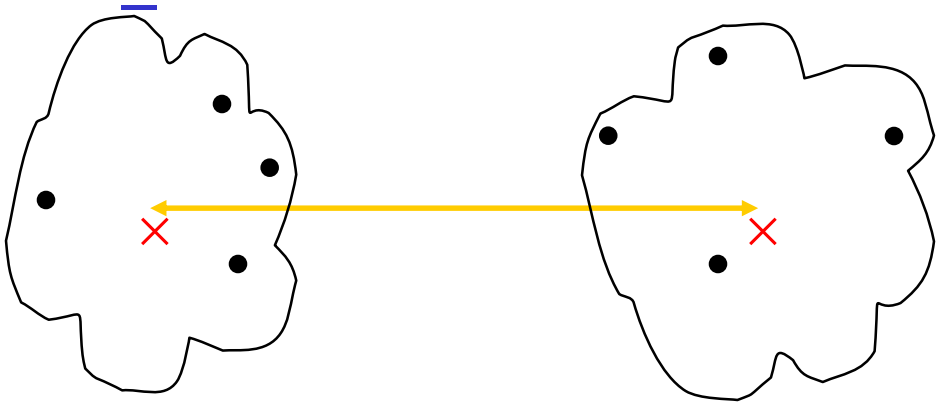
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- Min
- Max
- Group average
- Distance between centroids

Copyright 2019 © Wong Limsoon

How to define inter-cluster similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- Min
- Max
- Group average
- Distance between centroids

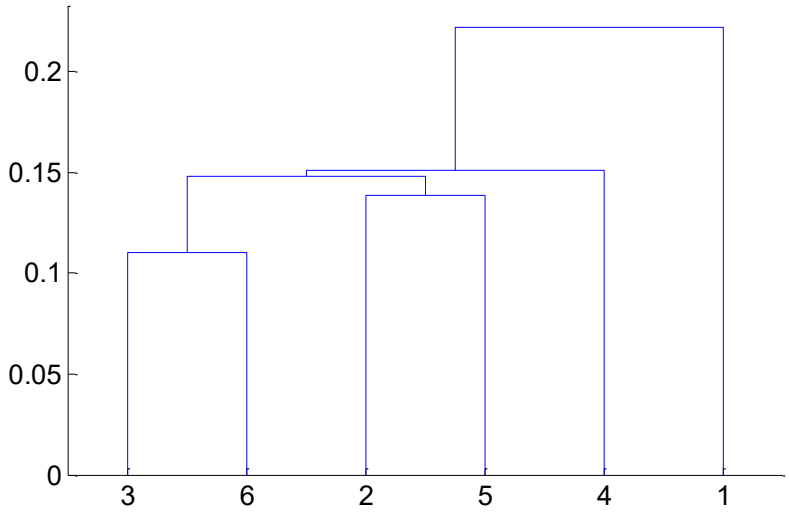
Copyright 2019 © Wong Limsoon

Cluster similarity: Min / single linkage

- **Similarity of two clusters is based on the two most similar (closest) points in the different clusters**
  - Determined by one pair of points, i.e., by one link in the proximity graph

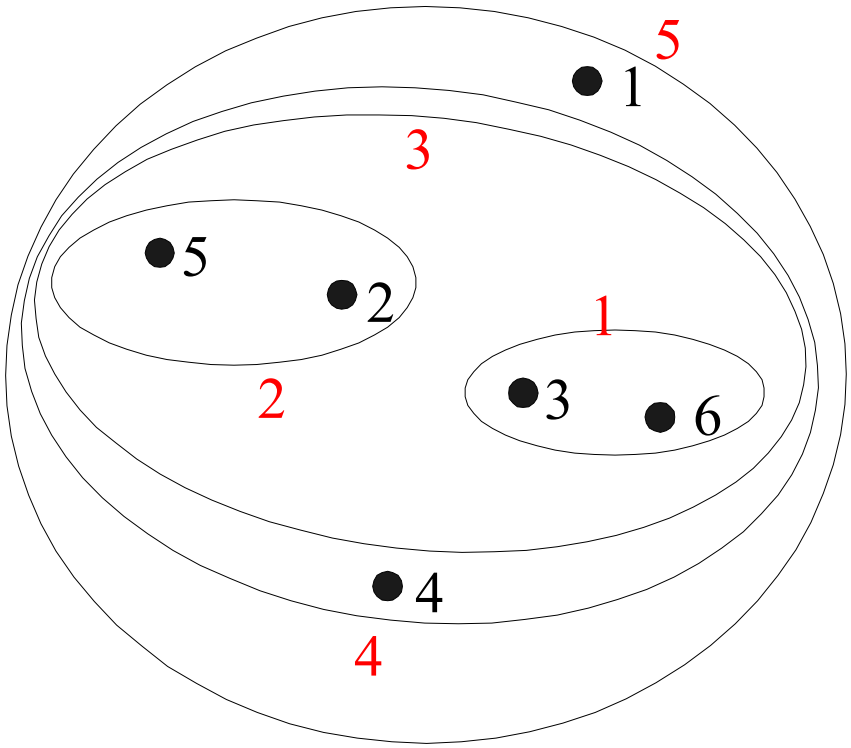
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 8.4. Euclidean distance matrix for 6 points.

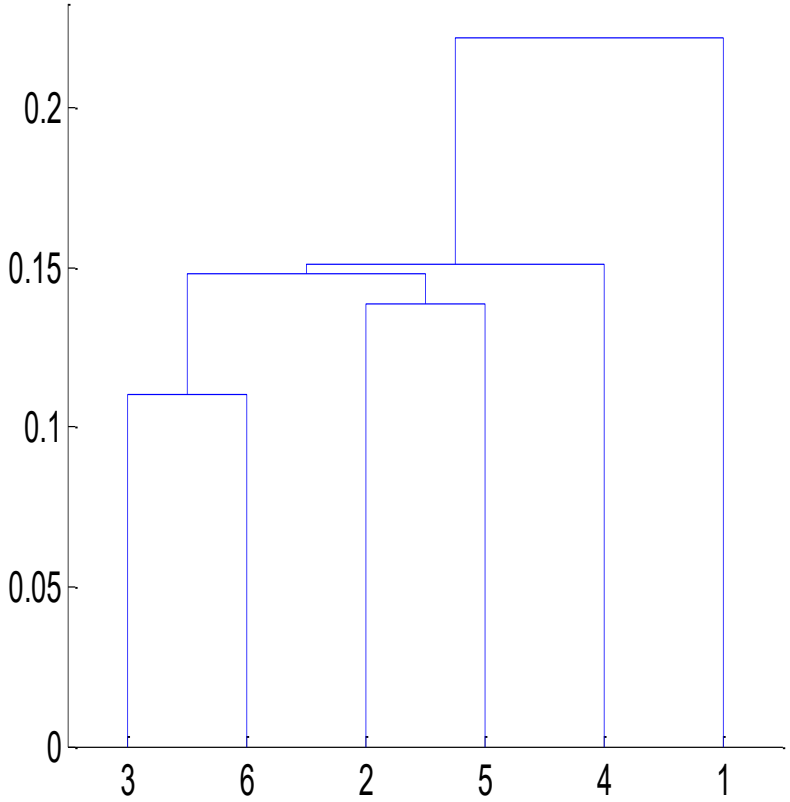


Copyright 2019 © Wong Limsoon

Hierarchical clustering: Min



Single-linkage clustering



Single-linkage dendrogram

Copyright 2019 © Wong Limsoon

Food for thought

- **What are the key strengths of single-linkage clustering?**

- **What are the key weaknesses of single-linkage clustering?**

Exercise #3

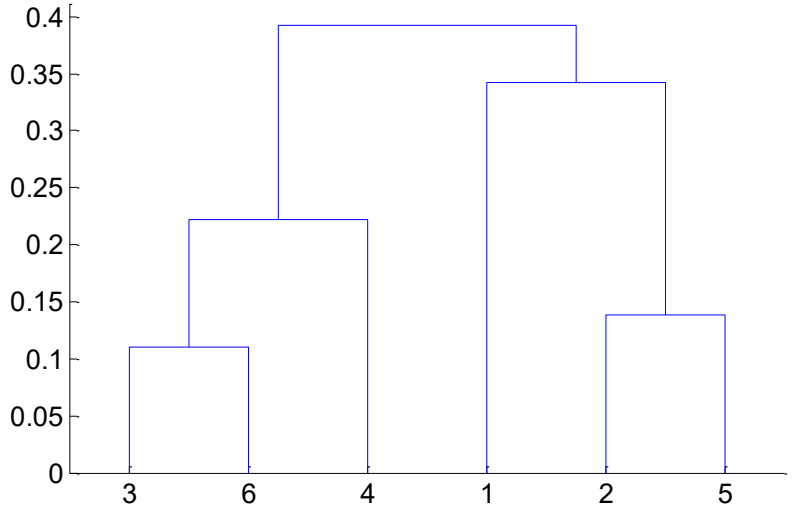
Copyright 2019 © Wong Limsoon

Cluster similarity: Max / complete linkage

- **Similarity of two clusters is based on the two least similar (most distant) points in the different clusters**
  - Determined by all pairs of points in the two clusters

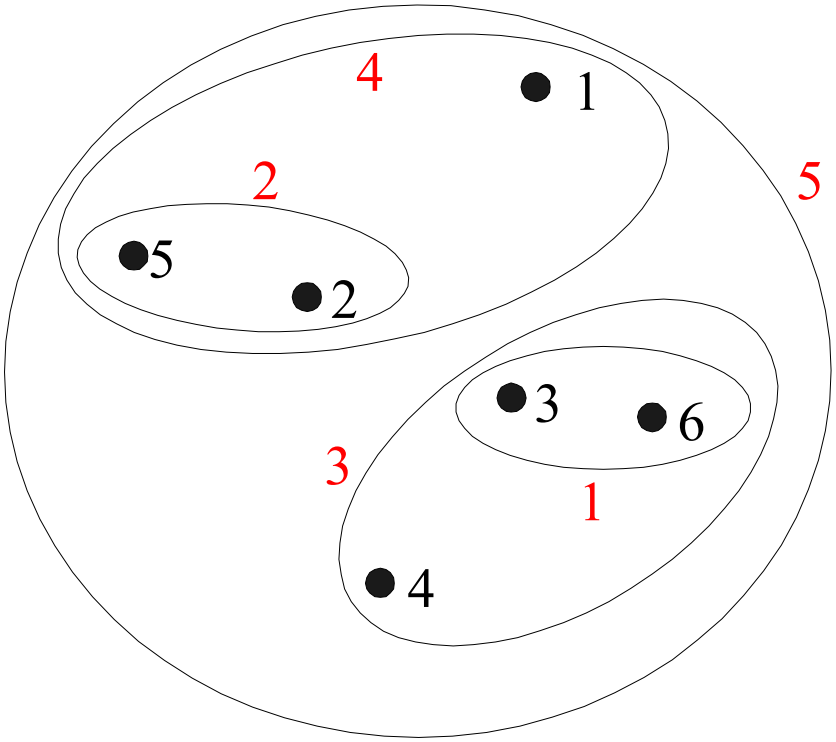
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 8.4. Euclidean distance matrix for 6 points.

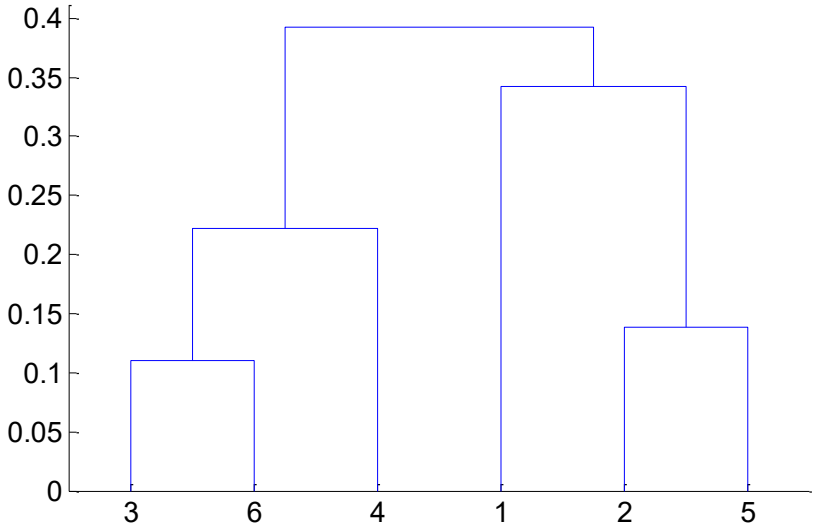


Copyright 2019 © Wong Limsoon

Hierarchical clustering: Max



Nested Clusters



Dendrogram

We still want to merge two most similar clusters each time.
But we define the distance between clusters based on MAX

Copyright 2019 © Wong Limsoon

Food for thought

- **What are the key strengths of complete-linkage clustering?**

- **What are the key weaknesses of complete-linkage clustering?**

Exercise #4

Copyright 2019 © Wong Limsoon

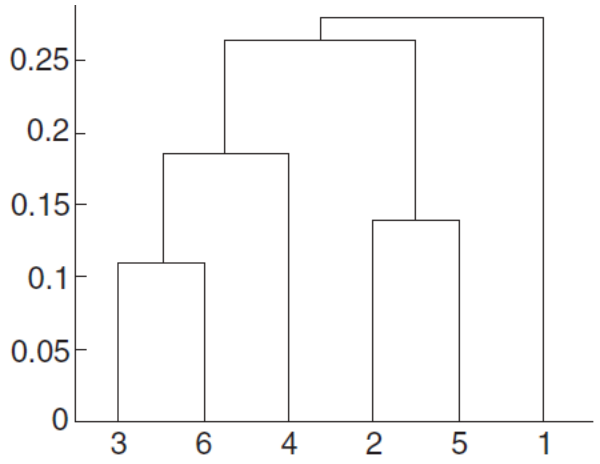
Cluster similarity: Group average

- **Proximity of two clusters is the average of pairwise proximity between points in the two clusters**

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{p_i \in \text{Cluster}_i, p_j \in \text{Cluster}_j} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

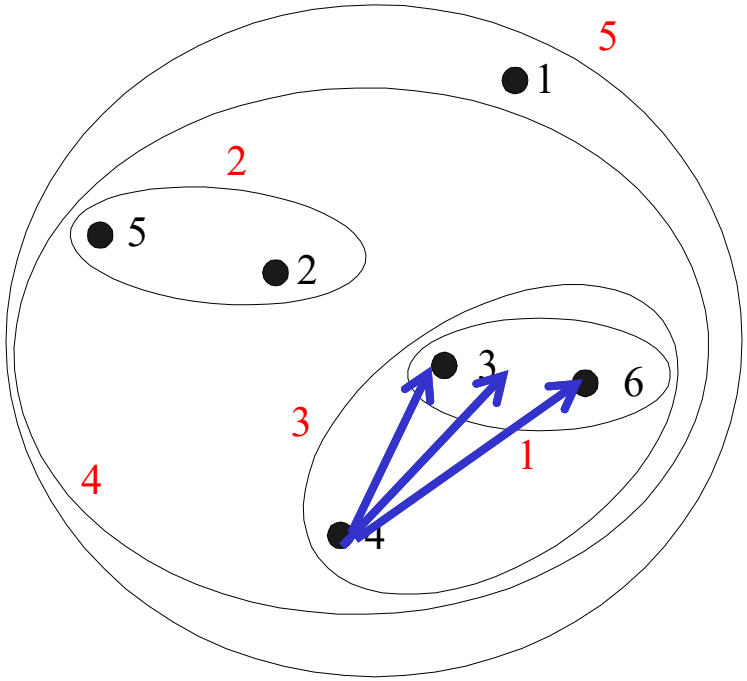
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 8.4. Euclidean distance matrix for 6 points.

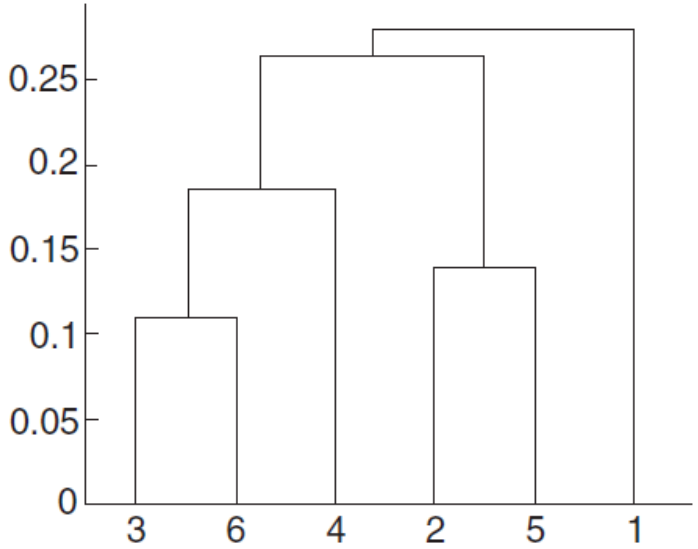


Copyright 2019 © Wong Limsoon

Hierarchical clustering: Group average



Group Average Clustering



Group Average Dendrogram

Copyright 2019 © Wong Limsoon

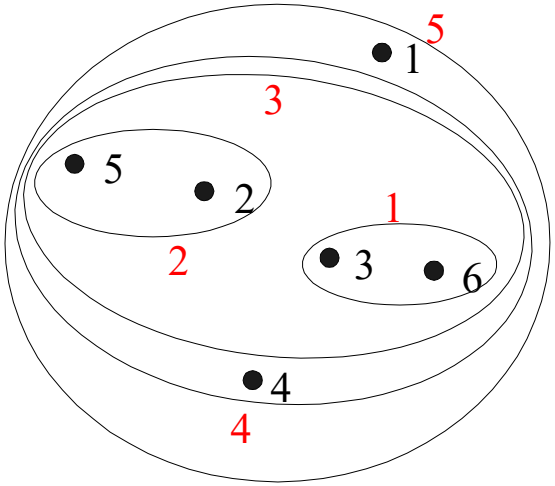
NUS National University of Singapore

Hierarchical clustering: Group average

- **Compromise between single and complete linkage**
- **Strengths**
  - Less susceptible to noise and outliers
- **Limitations**
  - Biased towards globular clusters

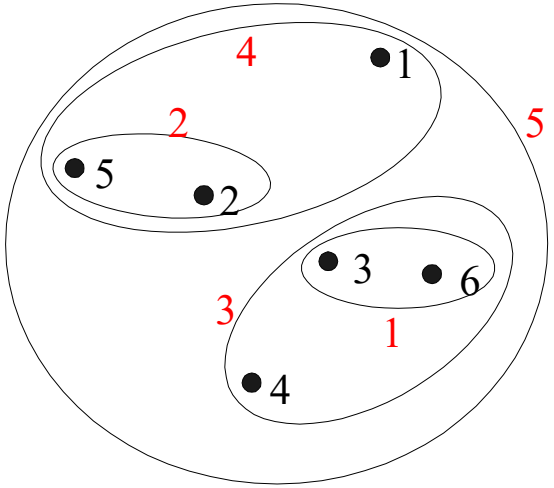
Copyright 2019 © Wong Limsoon

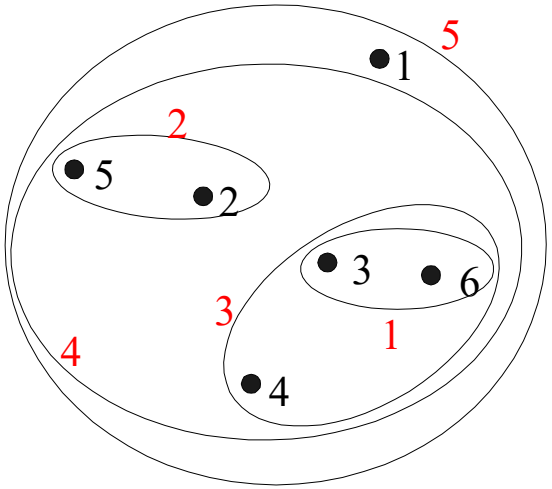
Hierarchical clustering: Comparison



Min

Max





Group average

Copyright 2019 © Wong Limsoon

Food for thought

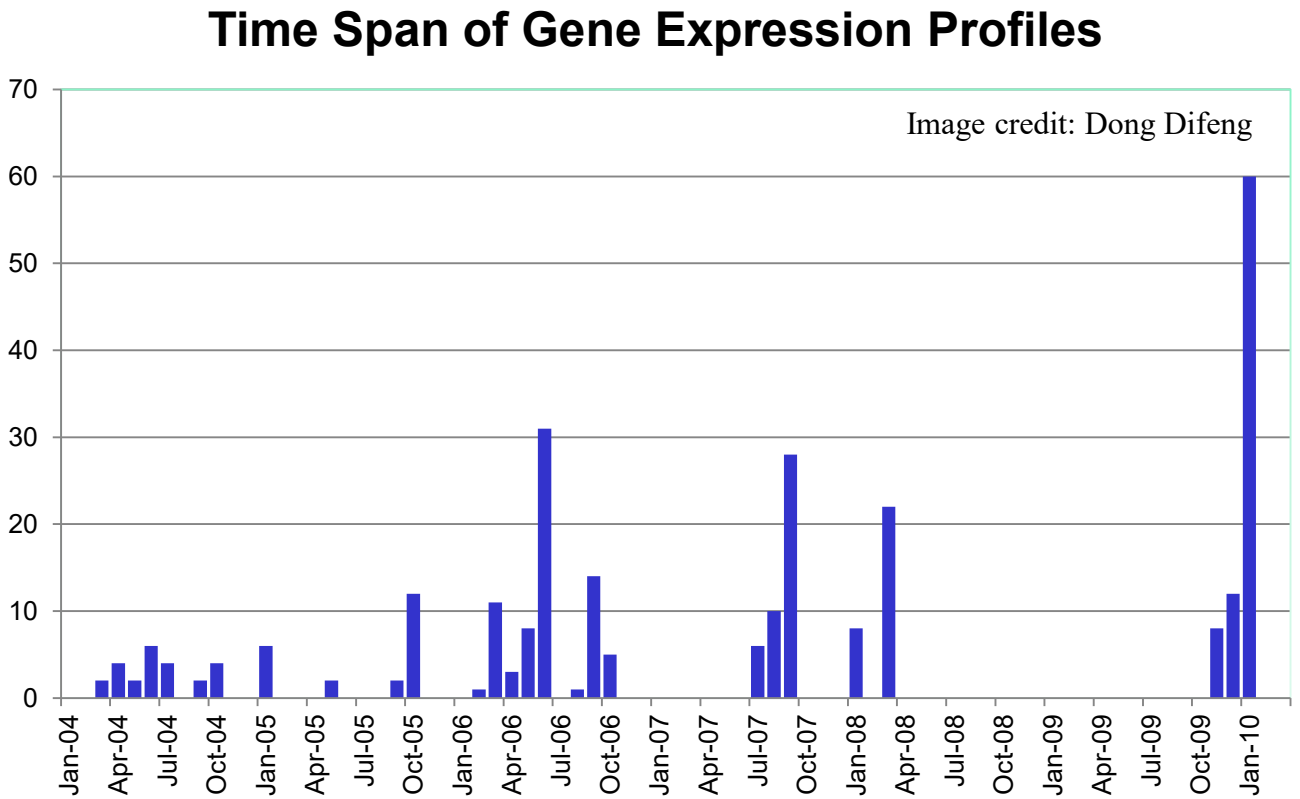
- **What are the space and time complexity of hierarchical clustering?**

Exercise #5

Copyright 2019 © Wong Limsoon

Normalization

Sometimes, a gene expression study may involve batches of data collected over a long period of time…



Copyright 2019 © Wong Limsoon

In such a case, batch effect may be severe… to the extent that you can predict the batch that each sample comes!

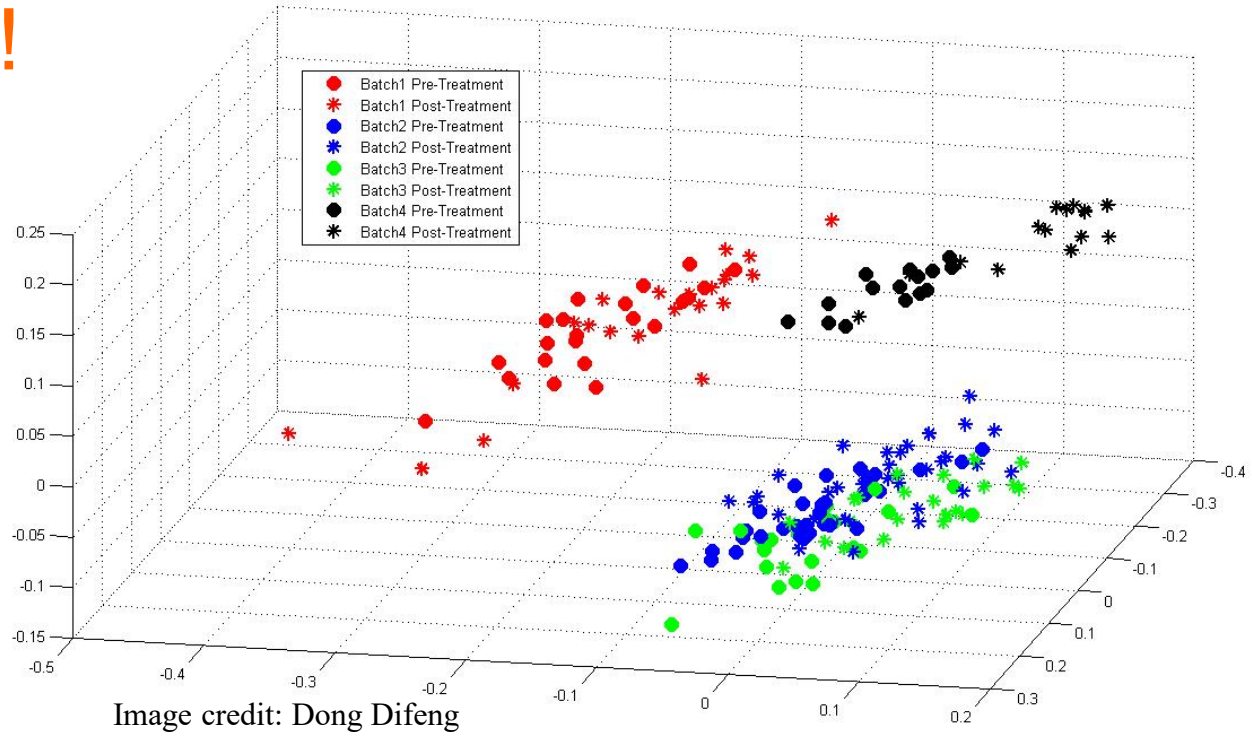


Image credit: Dong Difeng

⇒ Need normalization to correct for batch effect

Copyright 2019 © Wong Limsoon

Normalization approaches

- **Aim of normalization:**
  Reduce variance w/o increasing bias

- **Xform data so that distribution of probe intensities is same on all arrays**
  - E.g., $Z = (x - \mu) / \sigma$

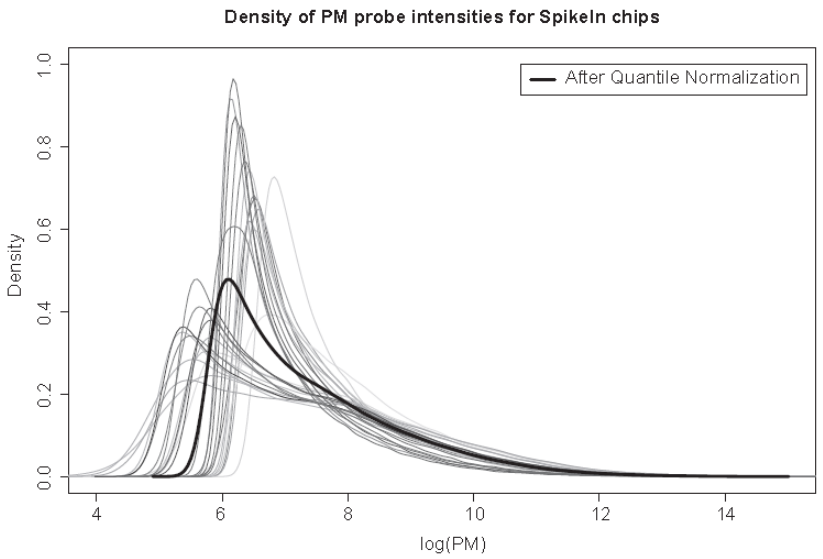
- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's

- **Quantile normalization**

Copyright 2019 © Wong Limsoon

Quantile normalization

- Given *n arrays of length p, form X of size p × n where each array is a column*
- Sort each column of *X to give* X_{sort}
- Take means across rows of X_{sort} *and assign this* mean to each elem in the row to get X'_{sort}
- Get $X_{normalized}$ *by arranging each column of* X'_{sort} to have same ordering as *X*

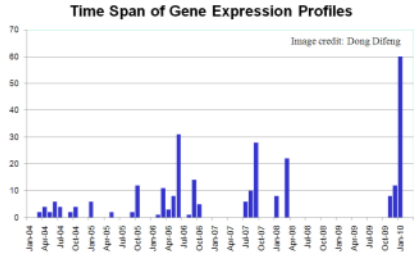


- Implemented in some microarray s/w, e.g., EXPANDER

Copyright 2019 © Wong Limsoon

47

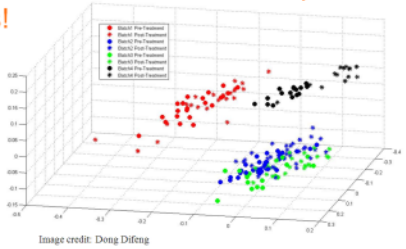
Sometimes, a gene expression study may involve batches of data collected over a long period of time…



Copyright 2011 © Limsoon Wong

48

In such a case, batch effect may be severe… to the extent that you can predict the batch that each sample comes!



⇒ Need normalization to correct for batch effect

Copyright 2011 © Limsoon Wong

After quantile normalization

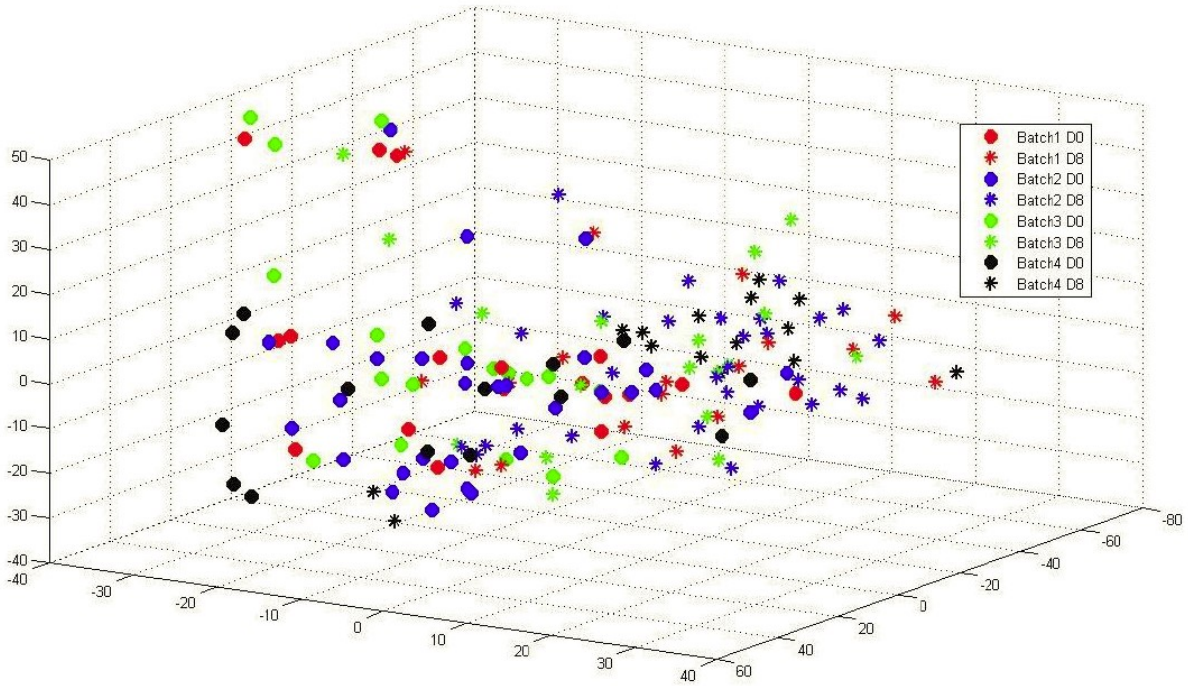
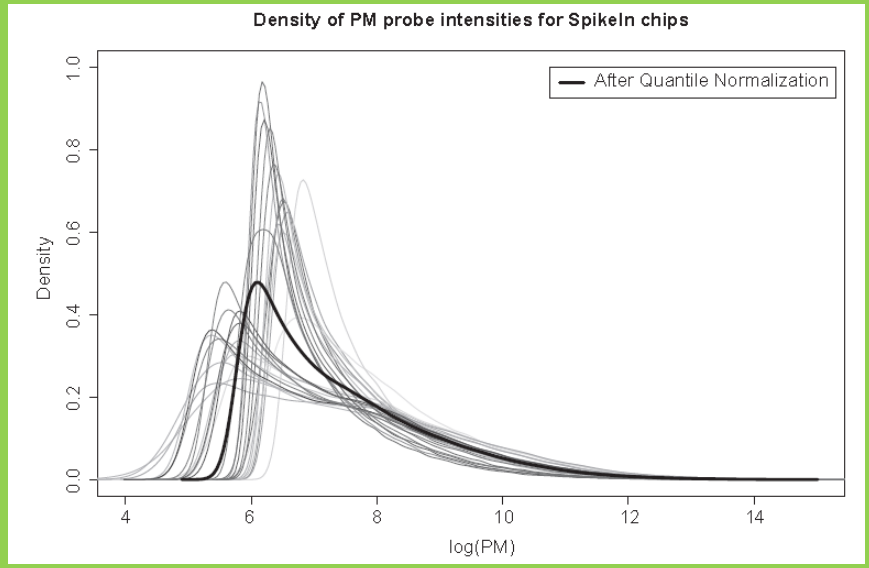


Figure 3.6: GEPs after the batch effects removing.

Copyright 2019 © Wong Limsoon

Food for thought

- Given a cancer vs normal dataset
- Should you apply quantile normalization to the dataset as a whole or should you apply quantile normalization to the cancer and the normal part separately? Why?



Exercise #6

Copyright © Wong Limsoon

Food for thought

- **Given a cancer vs normal dataset**
- **Should you apply Z-normalization to each phenotype separately or to the whole dataset in one go?**
- **Should you apply Z-normalization in a patient-wise or gene-wise manner? Why?**

Exercise #7

Copyright © Wong Limsoon

Selection of patient samples and genes for disease prognosis

Gene expression profile + clinical data ⇒ outcome prediction

- **Univariate & multivariate Cox survival analysis** (Beer et al 2002, Rosenwald et al 2002)
- **Fuzzy neural network** (Ando et al 2002)
- **Partial least squares regression** (Park et al 2002)
- **Weighted voting algorithm** (Shipp et al 2002)
- **Gene index and "reference gene"** (LeBlanc et al 2003)
- ……

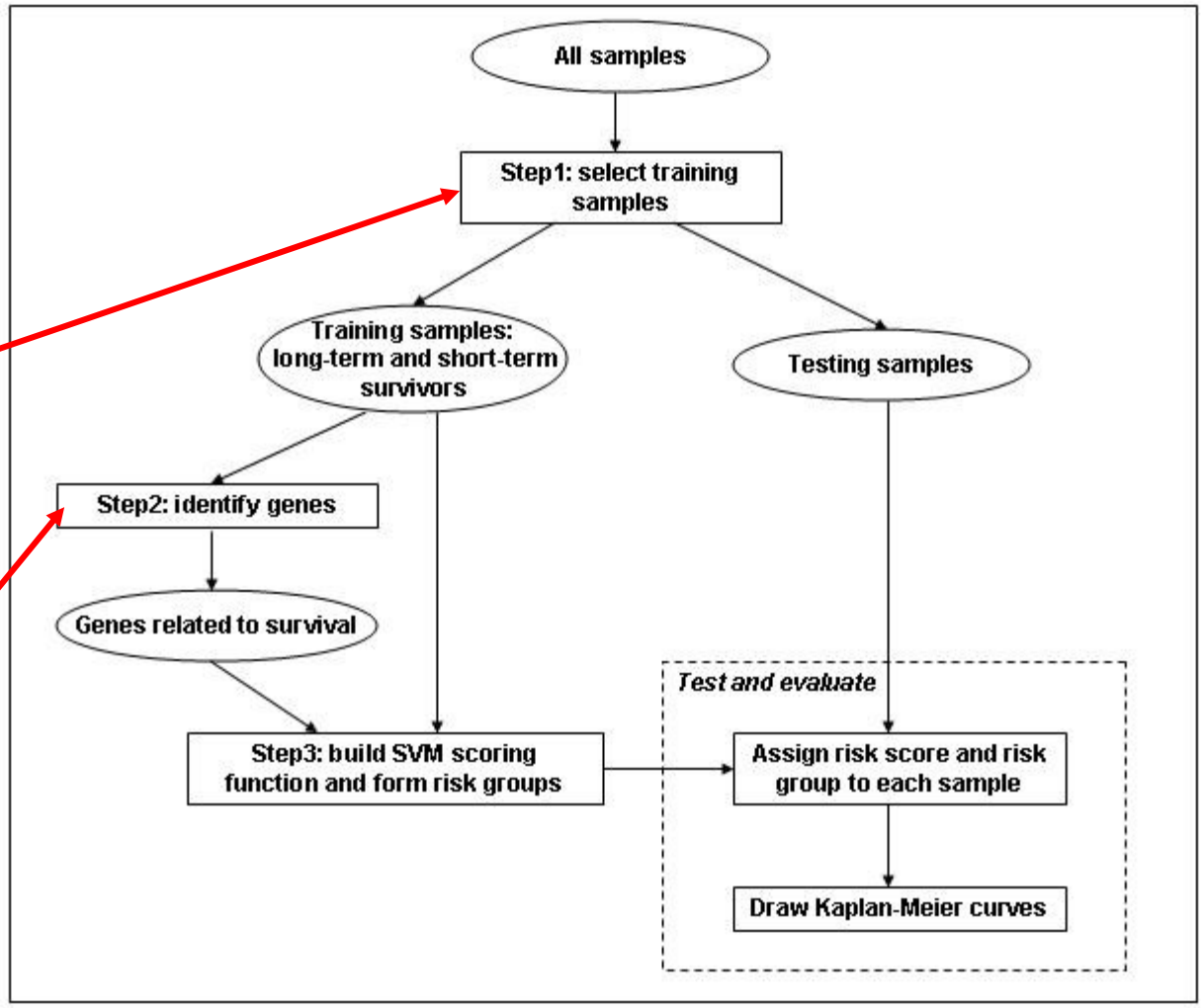
Copyright 2019 © Wong Limsoon

Liu et al. "Use of extreme patient samples for outcome prediction from gene expression data. *Bioinformatics*, 21(16):3377--3384, 2005

Our approach

"extreme" sample selection

ERCOF



Copyright 2019 © Wong Limsoon

Extreme sample selection

Short-term Survivors *v.s.* Long-term Survivors

Short-term survivors
who died within a *short* period

$\Downarrow$

$F(T) < c_1$ and $E(T) = 1$

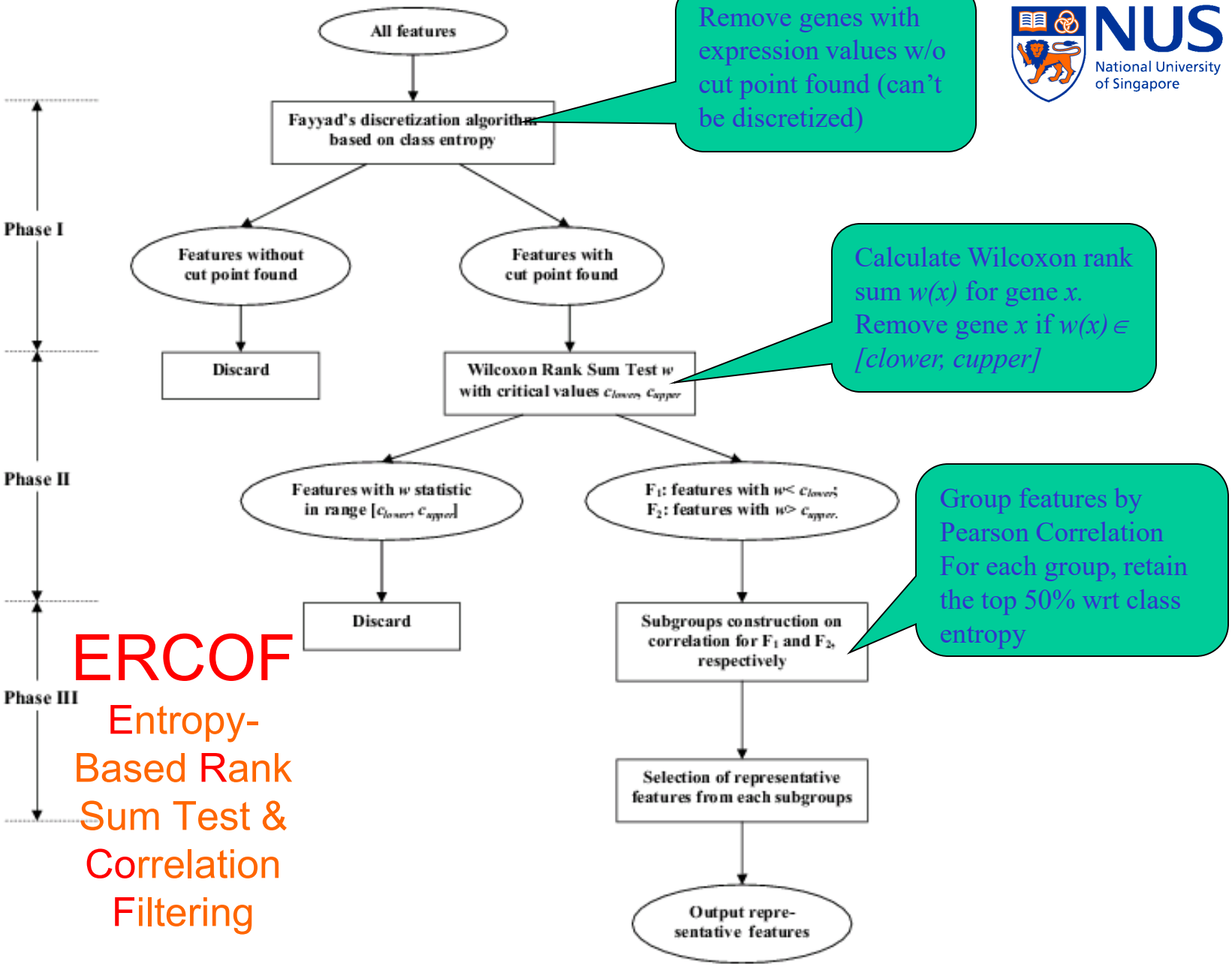
Long-term survivors
who were alive after a *long* follow-up time

$\Downarrow$

$F(T) > c_2$

T: sample
$F(T)$: follow-up time
$E(T)$: status (1:unfavorable; 0: favorable)
c_1 and c_2: thresholds of survival time

Copyright 2019 © Wong Limsoon



Copyright 2019 © Wong Limsoon

Risk score construction

Linear Kernel SVM regression function

$$G(T) = \sum_i a_i y_i K(T, x(i)) + b$$

T: test sample, $x(i)$: support vector,
y_i: class label (1: short-term survivors; -1: long-term survivors)

Transformation function (*posterior probability*)

$$S(T) = \frac{1}{1 + e^{-G(T)}} \qquad (S(T) \in (0,1))$$

$S(T)$: ***risk score*** of sample T

Copyright 2019 © Wong Limsoon

Diffuse large B-cell lymphoma

- **DLBC lymphoma is the most common type of lymphoma in adults**
- **Can be cured by anthracycline-based chemotherapy in 35 to 40 percent of patients**

⇒ **DLBC lymphoma comprises several diseases that differ in responsiveness to chemotherapy**

- **Intl Prognostic Index (IPI)**
  - age, "Eastern Cooperative Oncology Group" Performance status, tumor stage, lactate dehydrogenase level, sites of extranodal disease, ...
- **Not very good for stratifying DLBC lymphoma patients for therapeutic trials**

⇒ **Use gene-expression profiles to predict outcome of chemotherapy?**

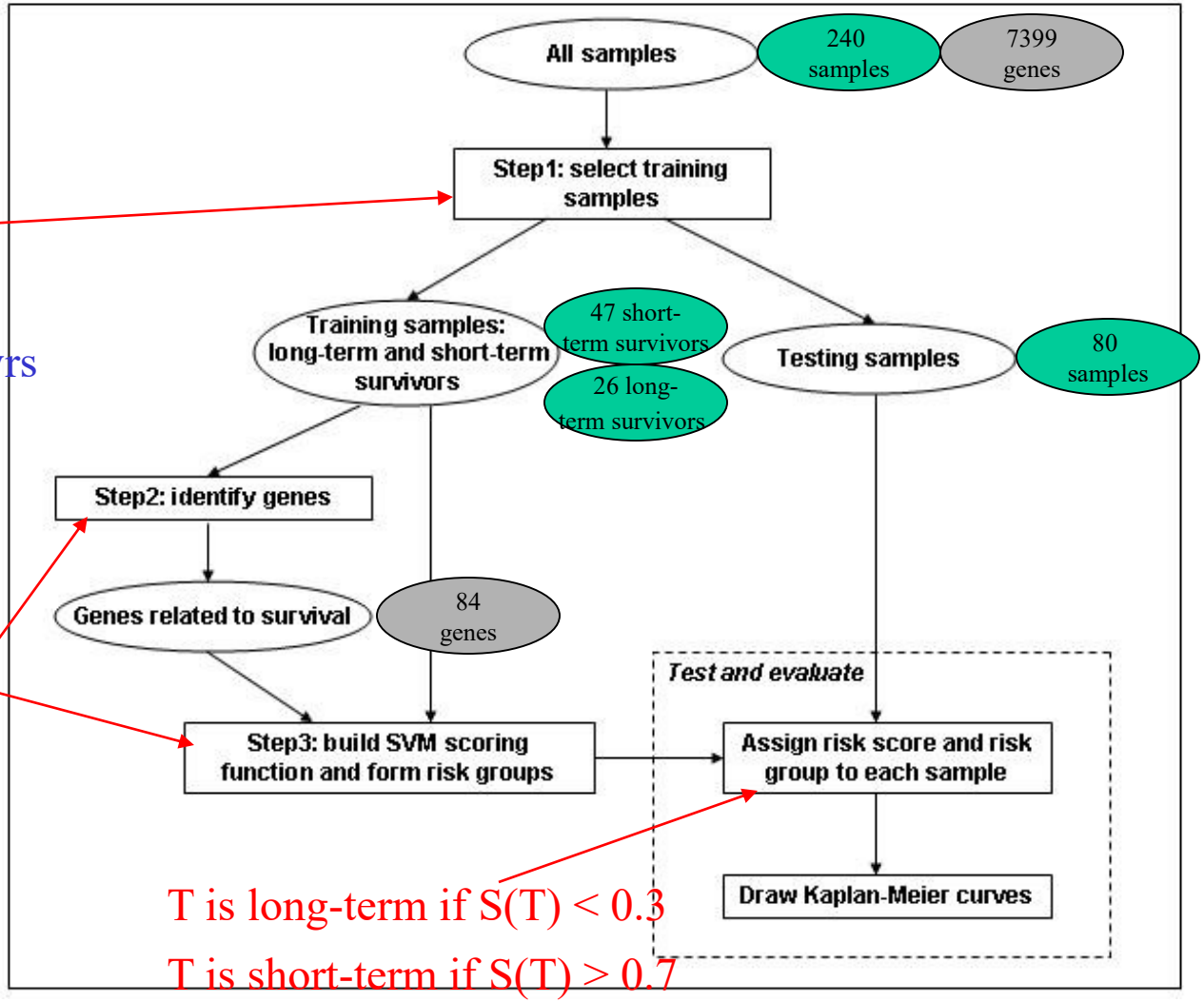
Copyright 2019 © Wong Limsoon

Rosenwald et al., *NEJM* 2002

- **240 data samples**
  - 160 in preliminary group
  - 80 in validation group
  - each sample described by 7399 microarray features
- **Rosenwald et al.'s approach**
  - identify gene: Cox proportional-hazards model
  - cluster identified genes into four gene signatures
  - calculate for each sample an outcome-predictor score
  - divide patients into quartiles according to score

Copyright 2019 © Wong Limsoon

Knowledge discovery from gene expression of "extreme" samples



"extreme" sample selection: < 1 yr vs > 8 yrs

knowledge discovery from gene expression

T is long-term if S(T) < 0.3

T is short-term if S(T) > 0.7

Copyright 2019 © Wong Limsoon

Discussions: Sample selection

Application	Data set	Status		Total
		Dead	Alive	
DLBCL	Original	88	72	160
	Informative	47+1(*)	25	73

Number of samples in original data and selected informative training set.
(*): Number of samples whose corresponding patient was dead at the end of follow-up time, but selected as a long-term survivor.

Copyright 2019 © Wong Limsoon

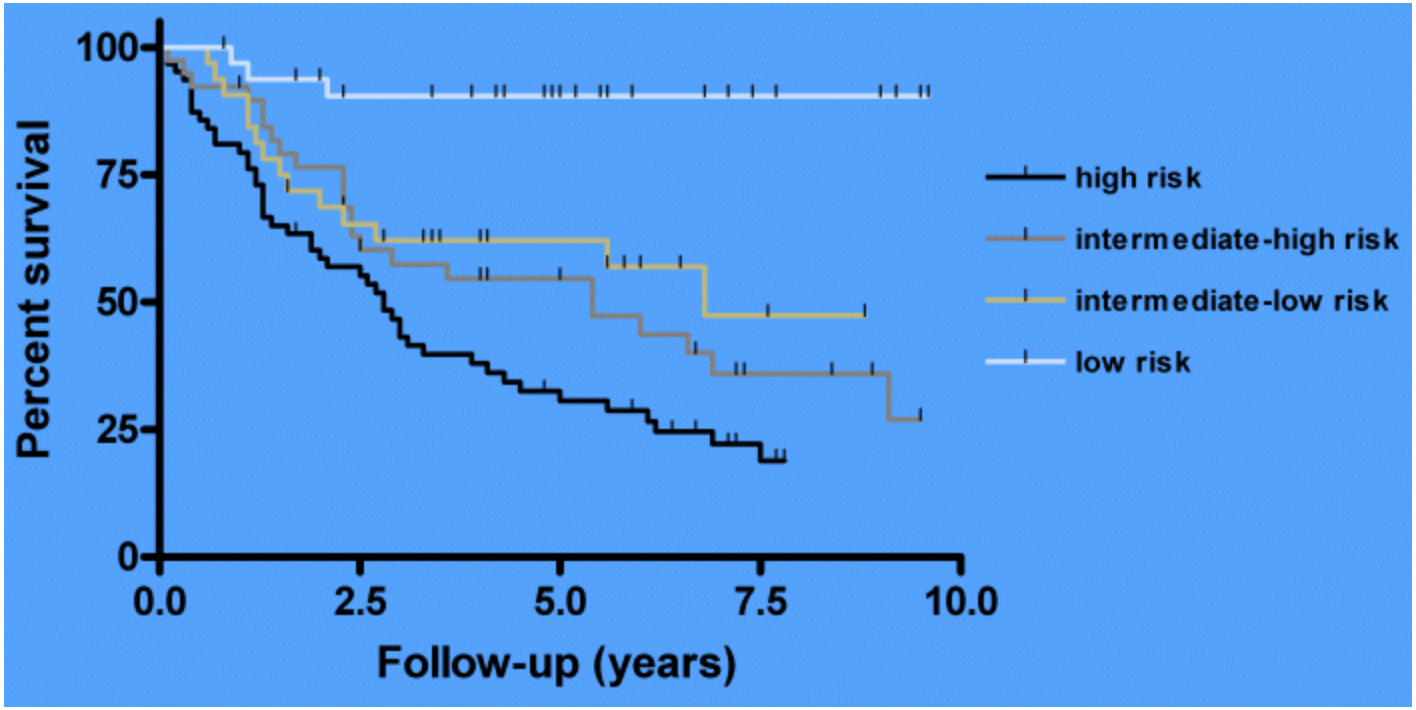
Discussions: Gene identification

Gene selection	DLBCL
Original	4937(*)
Phase I	132(2.7%)
Phase II	84(1.7%)

Number of genes left after feature filtering for each phase.
(*): number of genes after removing those genes who were absent in more than 10% of the experiments.

Copyright 2019 © Wong Limsoon

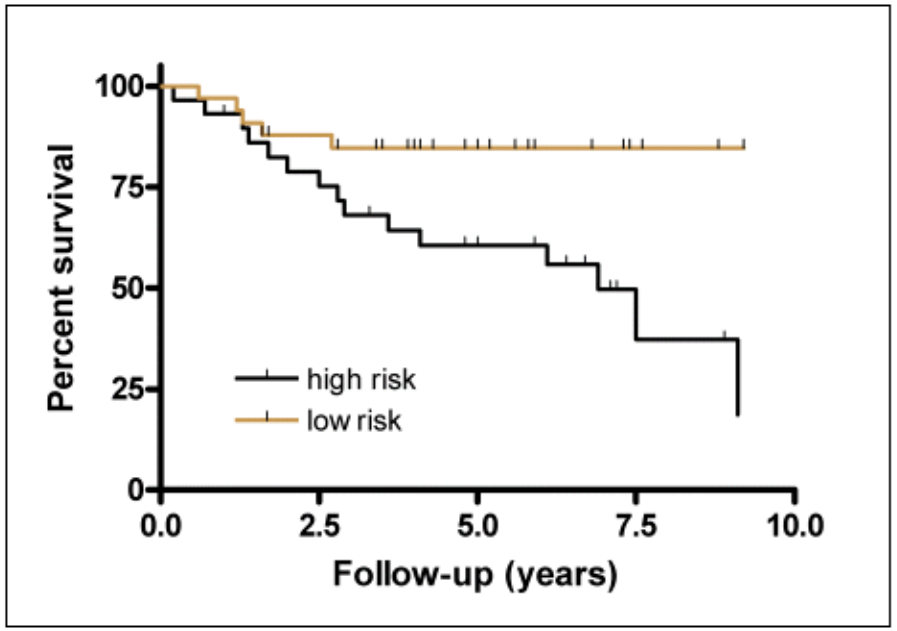
Kaplan-Meier plot for 80 test cases



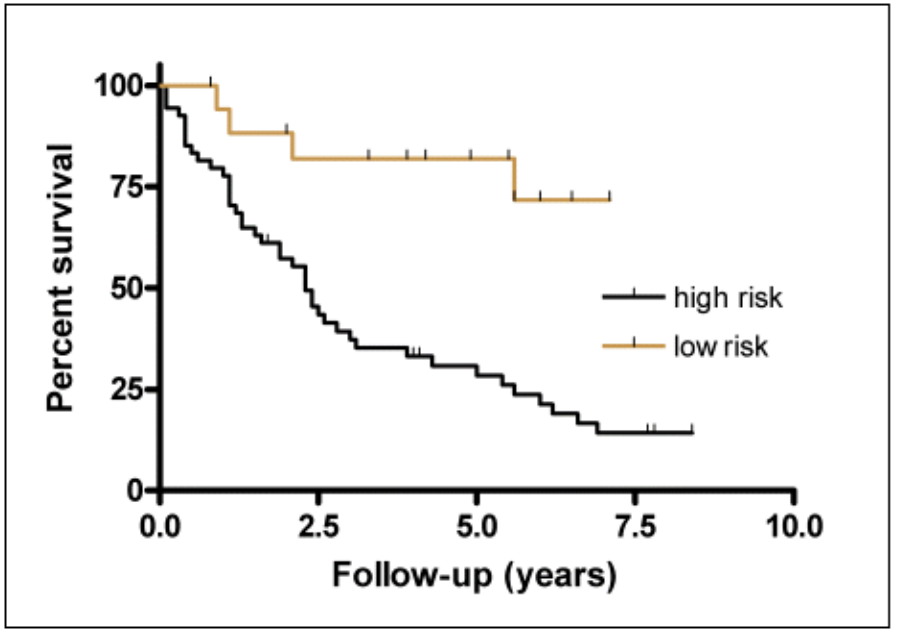
p-value of log-rank test: < 0.0001
Risk score thresholds: 0.7, 0.3

Copyright 2019 © Wong Limsoon

Improvement over IPI



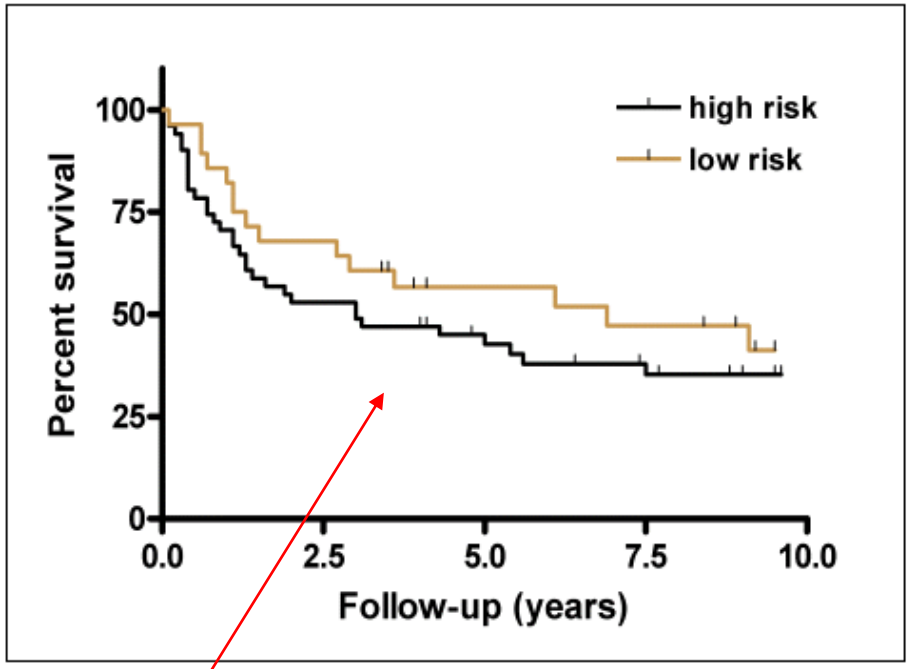
(A) IPI low,
p-value = 0.0063



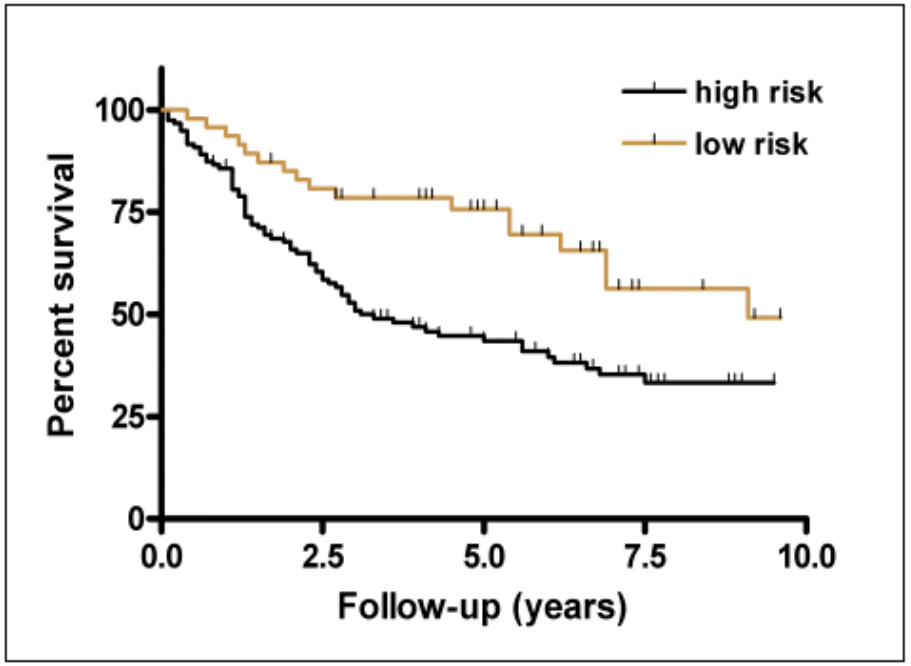
(B) IPI intermediate,
p-value = 0.0003

Copyright 2019 © Wong Limsoon

Merit of "extreme" samples



(A) W/o sample selection (p =0.38)



(B) With sample selection (p=0.009)

No clear difference on the overall survival of the 80 samples in the validation group of DLBCL study, if **no training sample selection conducted**

Copyright 2019 © Wong Limsoon

About the inventor: Huiqing Liu

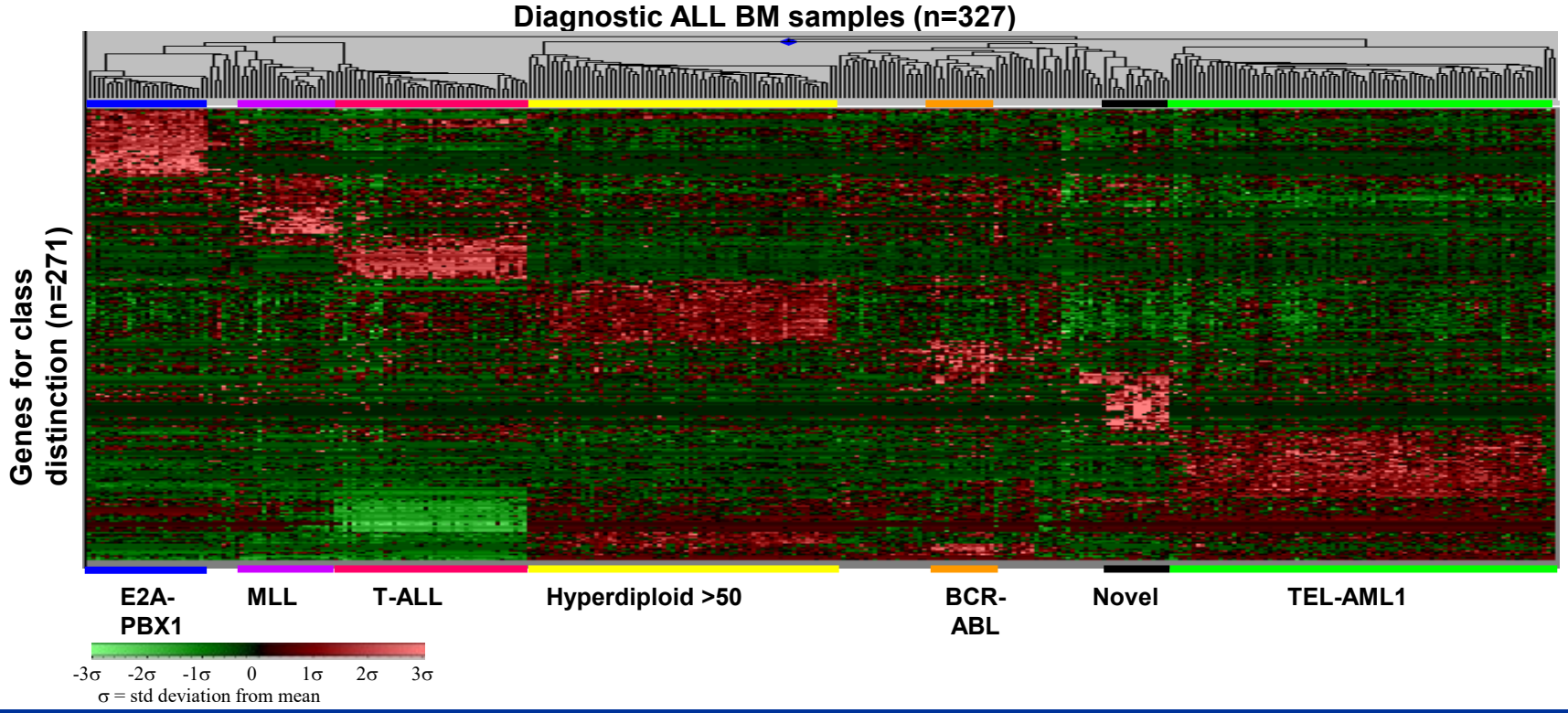
- **Huiqing Liu**
  - PhD, NUS, 2004
  - Currently PI at Incyte
  - Asian Innovation Gold Award 2003
  - New Jersey Cancer Research Award for Scientific Excellence 2008
  - Gallo Prize 2008

Copyright 2019 © Wong Limsoon

Beyond disease diagnosis & prognosis

Beyond classification of gene expression profiles

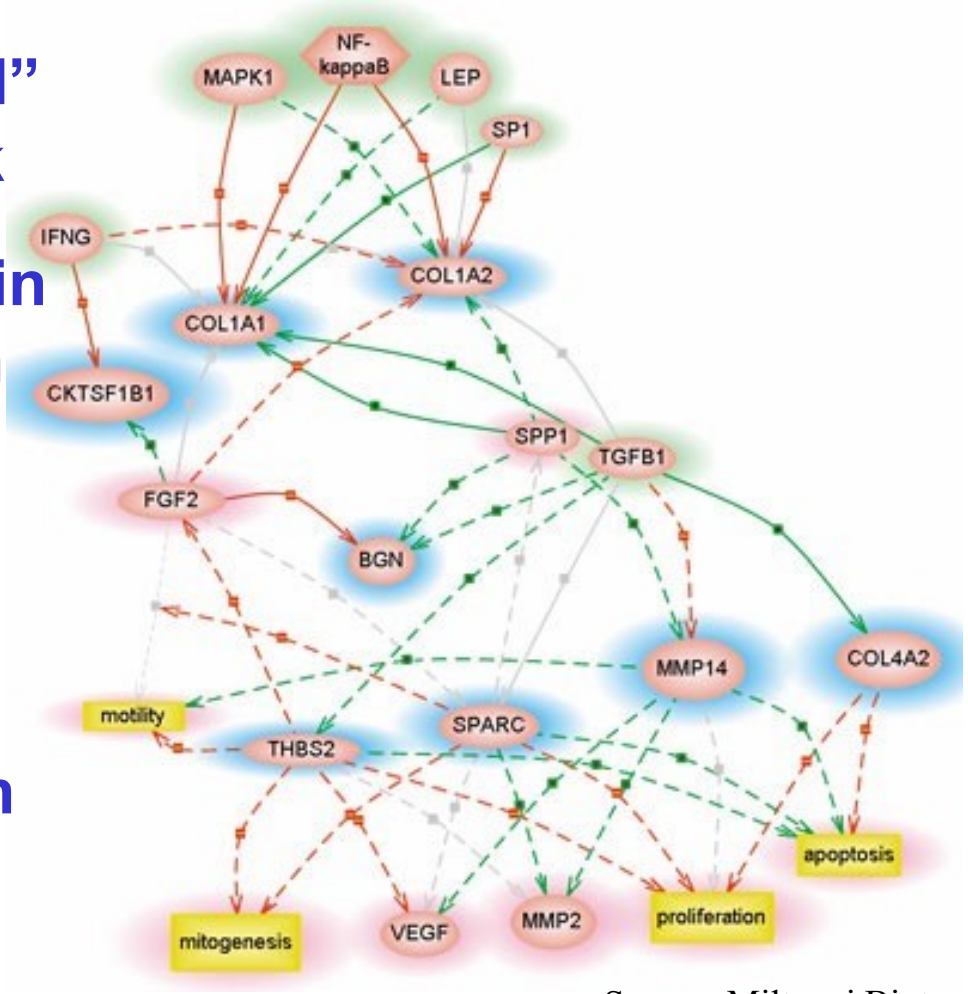
- **After identifying the candidate genes by feature selection, do we know which ones are causal genes, which ones are surrogates, and which are noise?**



Copyright 2019 © Wong Limsoon

Gene regulatory circuits

- Genes are "connected" in "circuit" or network
- Expression of a gene in a network depends on expression of some other genes in the network
- Can we "reconstruct" the gene network from gene expression and other data?



Source: Miltenyi Biotec

Copyright 2019 © Wong Limsoon

Key questions

For each gene in the network:

- **Which genes affect it?**
- **How they affect it?**
  - Positively?
  - Negatively?
  - More complicated ways?

Copyright 2019 © Wong Limsoon

Some techniques

- **Bayesian Networks**
  - Friedman et al., *JCB* 7:601--620, 2000
- **Boolean Networks**
  - Akutsu et al., *PSB* 2000, pages 293--304
- **Differential equations**
  - Chen et al., *PSB* 1999, pages 29--40
- **Classification-based method**
  - Soinov et al., "Towards reconstruction of gene network from expression data by supervised learning", *Genome Biology* 4:R6.1--9, 2003

Copyright 2019 © Wong Limsoon

A classification-based technique

Soinov et al., *Genome Biology* 4:R6.1-9, 2003

- **Given a gene expression matrix X**
  - each row is a gene
  - each column is a sample
  - each element x_{ij} is expression of gene i in sample j
- **Find the average value a_i of each gene i**
- **Denote s_{ij} as state of gene i in sample j,**
  - s_{ij} = up if $x_{ij} > a_i$
  - s_{ij} = down if $x_{ij} \leq a_i$

Copyright 2019 © Wong Limsoon

A classification-based technique

Soinov et al., *Genome Biology* 4:R6.1-9, Jan 2003

- **To see whether the state of gene g is determined by the state of other genes**
  - See whether $\langle s_{ij} \mid i \neq g \rangle$ can predict s_{gj}
  - If can predict with high accuracy, then "yes"
  - Any classifier can be used, such as C4.5, PCL, SVM, etc.
- **To see how the state of gene g is determined by the state of other genes**
  - Apply C4.5 (or PCL or other "rule-based" classifiers) to predict s_{gj} from $\langle s_{ij} \mid i \neq g \rangle$
  - Extract the decision tree or rules used

Copyright 2019 © Wong Limsoon

Advantages of this method

- Can identify genes affecting a target gene
- Don't need discretization thresholds?
- Each data sample is treated as an example
- Explicit rules can be extracted from the classifier (assuming C4.5 or PCL)
- Generalizable to time series

- Discuss the point "Don't need discretization thresholds". Is it true?

Exercise #8

Copyright © Wong Limsoon

Concluding remarks

NUS
National University
of Singapore

Bcr-Abl

- **Targeted drug dev**
  - Know what molecular effect you want to achieve
    - **E.g., inhibit a mutated form of a protein**
  - Engineer a compound that directly binds and causes the desired effect
- **Gleevec (imatinib)**
  - 1st success for real drug
  - Targets Bcr-Abl fusion protein (ie, Philadelphia chromosome, Ph)
  - NCI summary of clinical trial of imatinib for ALL at

http://www.cancer.gov/clinicaltrials/results/ALLimatinib1109/print

Copyright 2019 © Wong Limsoon

What have we learned?

- **Technologies**
  - Microarray
  - PCL, ERCOF
- **Microarray applications**
  - Disease diagnosis by supervised learning
  - Subtype discovery by unsupervised learning
  - Disease diagnosis via guilt-by-association
  - Gene network reconstruction
- **Important tactic**
  - Extreme sample selection

Copyright 2019 © Wong Limsoon

Useful packages

- **EXPANDER (EXPression Analyser & DisplayER)**
  - http://acgt.cs.tau.ac.il/expander
- **BRB-Array Tools**
  - http://linus.nci.nih.gov/BRB-ArrayTools.html
- **NetProt**
  - http://rpubs.com/gohwils/204259
  - https://github.com/gohwils/NetProt/releases/

Copyright 2019 © Wong Limsoon

Any question?

References

- E.-J. Yeoh et al., "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling", *Cancer Cell*, 1:133--143, 2002
- H. Liu, J. Li, L. Wong. Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data. *Bioinformatics*, 21(16):3377--3384, 2005.
- L.D. Miller et al., "Optimal gene expression analysis by microarrays", *Cancer Cell* 2:353--361, 2002
- J. Li, L. Wong, "Techniques for Analysis of Gene Expression", *The Practical Bioinformatician*, Chapter 14, pages 319—346, WSPC, 2004
- B. Bolstad et al. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". *Bioinformatics*, 19:185–193. 2003

Copyright 2019 © Wong Limsoon