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

CS2220: Introduction to Computational Biology

Lecture 5: Gene Expression and Proteome Analysis

Limsoon Wong

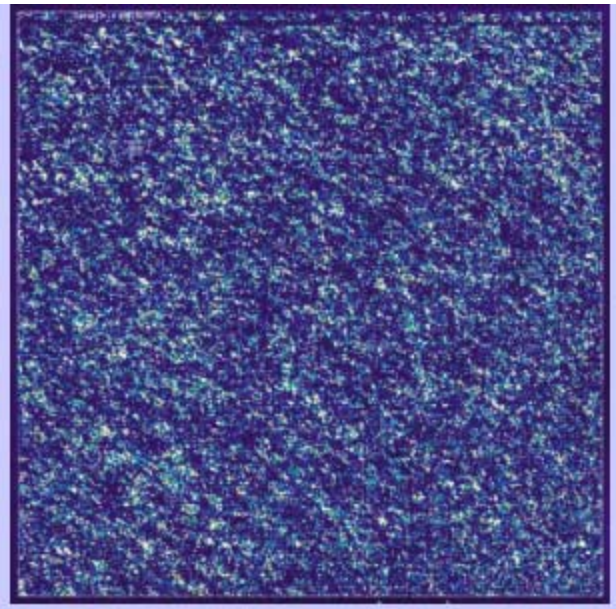
9 February 2007

Background on Microarrays

What's a Microarray?

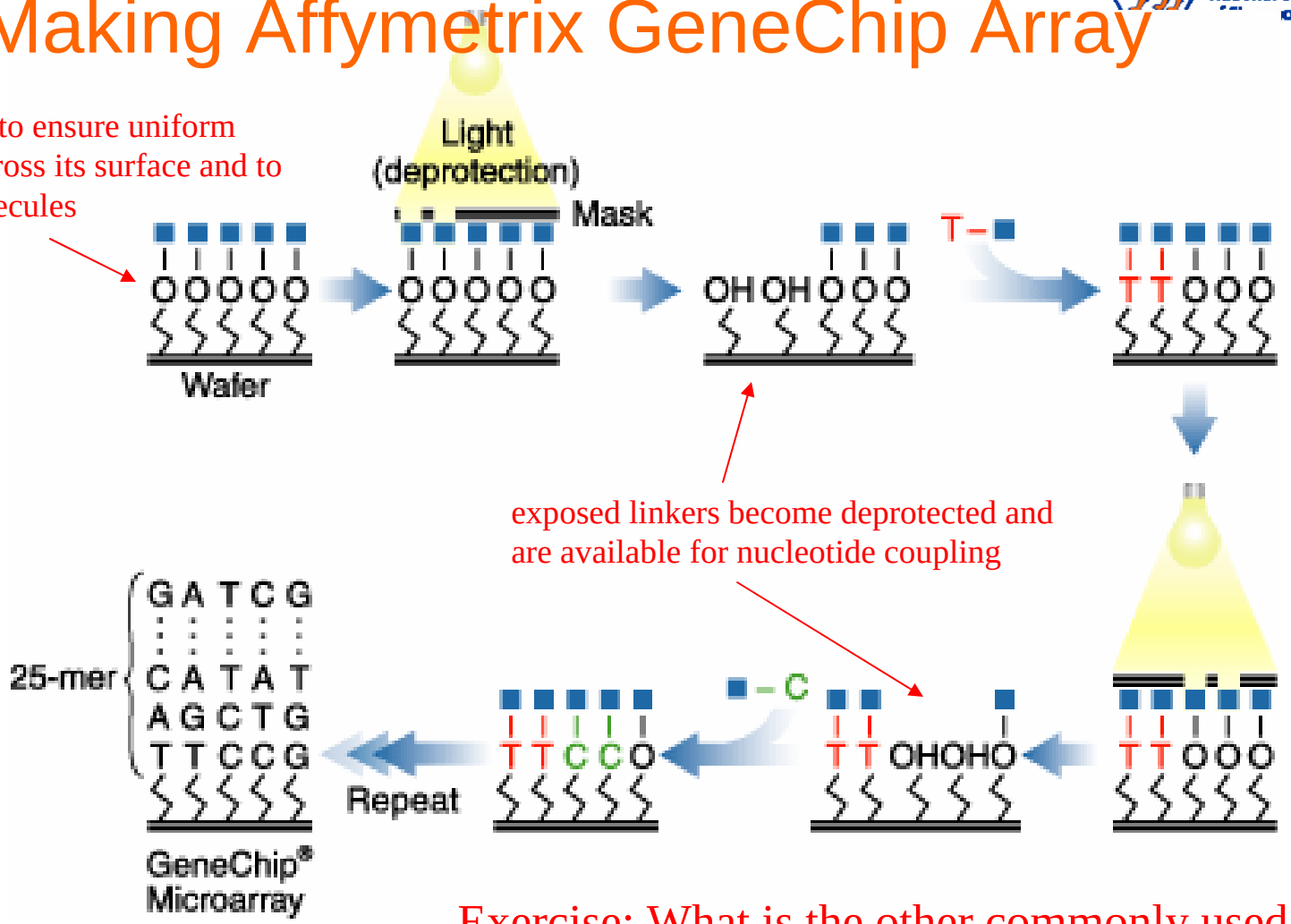
- **Contain large number of DNA molecules spotted on glass slides, nylon membranes, or silicon wafers**
- **Detect what genes are being expressed or found in a cell of a tissue sample**
- **Measure expression of thousands of genes simultaneously**

Affymetrix GeneChip Array



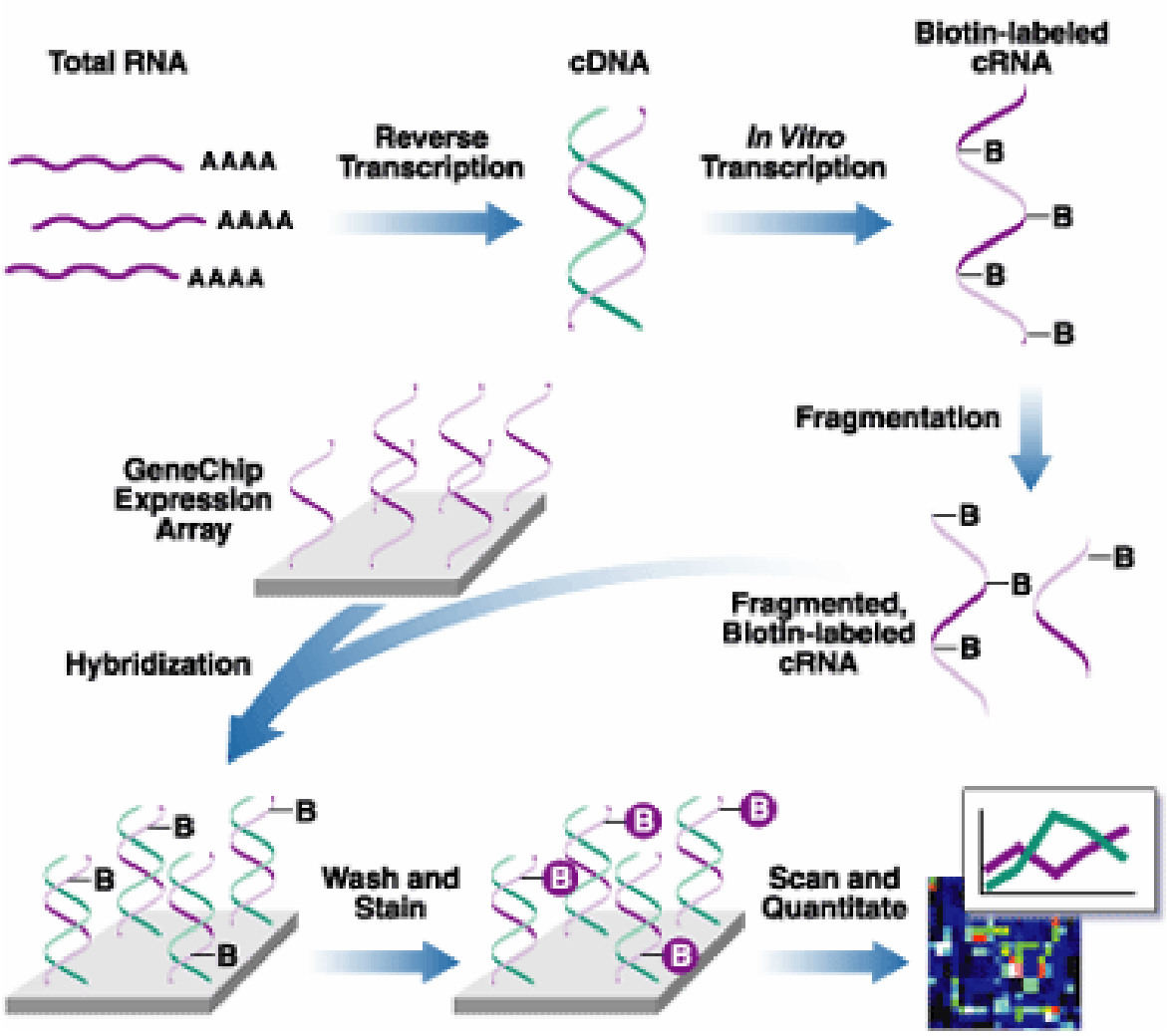
Making Affymetrix GeneChip Array

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?

Gene Expression Measurement by Affymetrix GeneChip Array



A Sample Affymetrix GeneChip Data File (U95A)

	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-BioB	13	1	19	9340.6	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioB	15	0	19	12862.4	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioB	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-BioB	7	5	18	-483	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioB	5	4	18	313.7	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioB	7	6	20	-1016.2	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r

Some Advice on Affymetrix Gene Chip 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**
  - Accuracy of laser scanner
- **Deal with missing values**

Exercise: Suggest 2 ways to deal with missing value

Type of Gene Expression Datasets

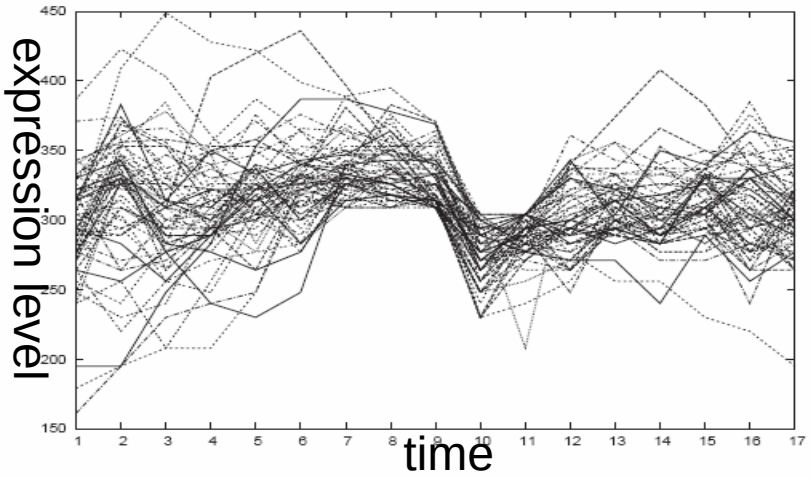
- Gene-Conditions or **Gene-Sample** (**numeric** or discretized)

1000 - 100,000 columns

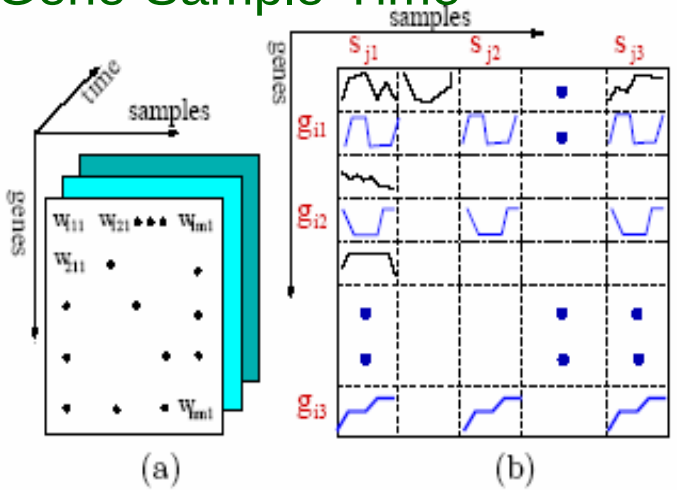
100-500 rows

	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



Type of Gene Expression Datasets

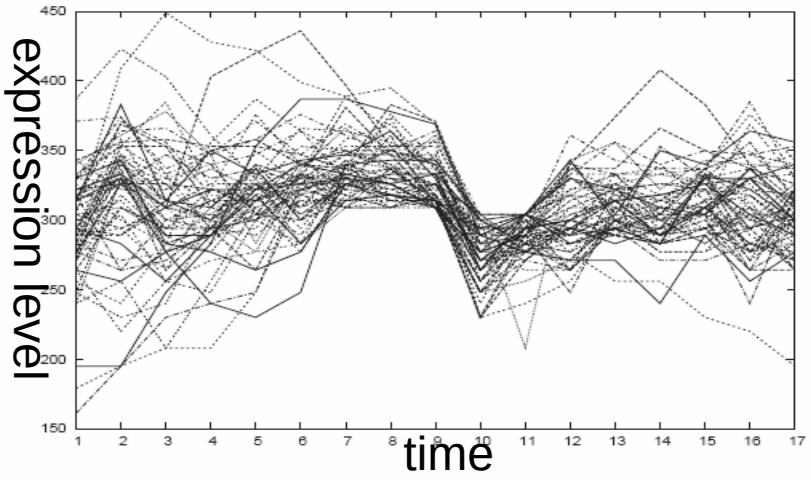
- **Gene-Conditions** or Gene-Sample (**numeric** or discretized)

1000 - 100,000 columns

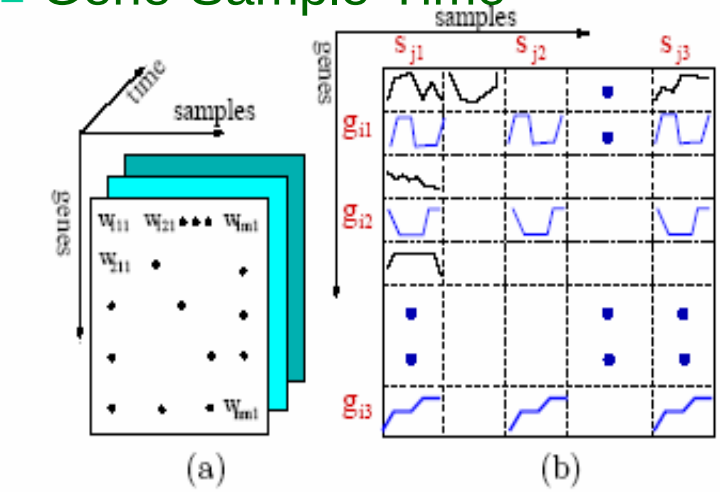
100-500 rows

	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



Type of Gene Expression Datasets

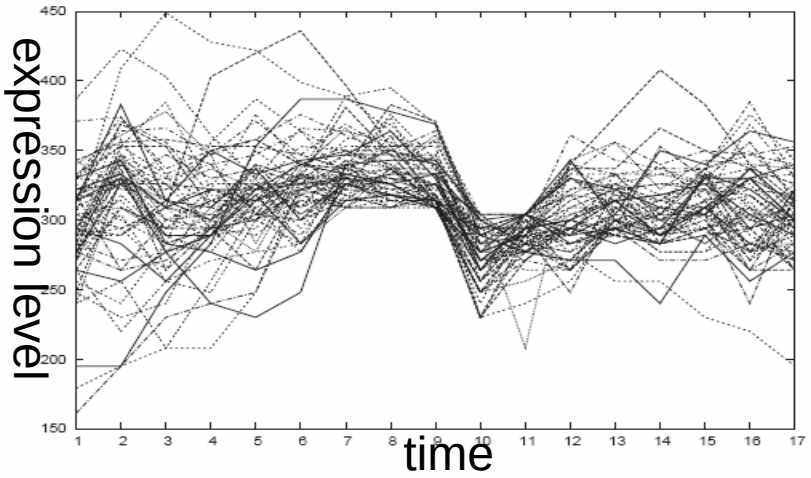
- Gene-Conditions or **Gene-Sample** (numeric or **discretized**)

1000 - 100,000 columns

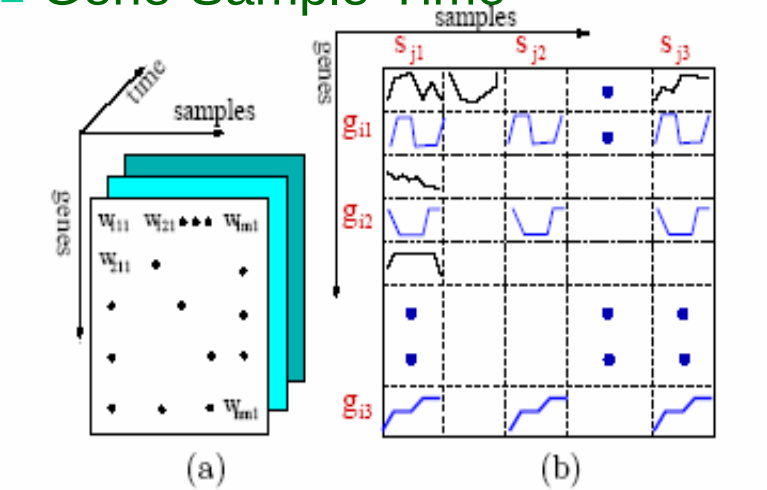
100-500 rows

	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



Gene Expression Profile Classification

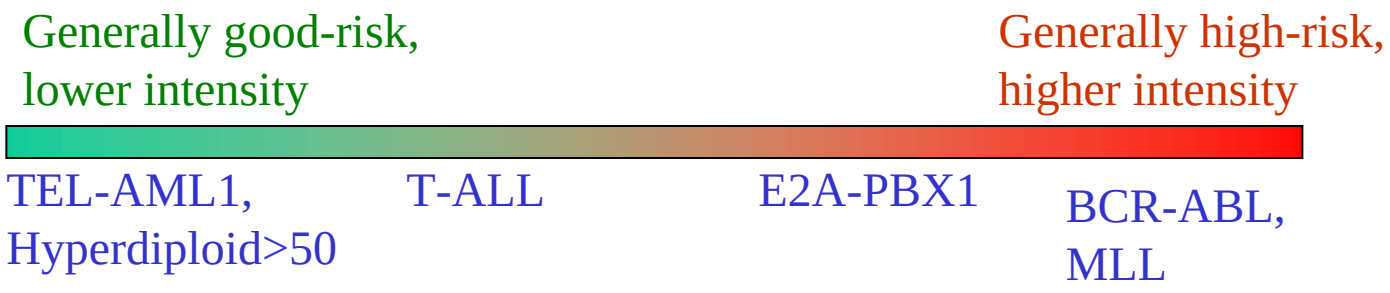
Diagnosis of Childhood Acute Lymphoblastic Leukemia and Optimization of Risk-Benefit Ratio of Therapy

Childhood ALL, A Heterogeneous Disease

- **Major subtypes are**
  - T-ALL
  - E2A-PBX1
  - TEL-AML1
  - MLL genome rearrangements
  - Hyperdiploid>50
  - BCR-ABL

Risk-Stratified Therapy

- **Different subtypes respond differently to the same treatment intensity**



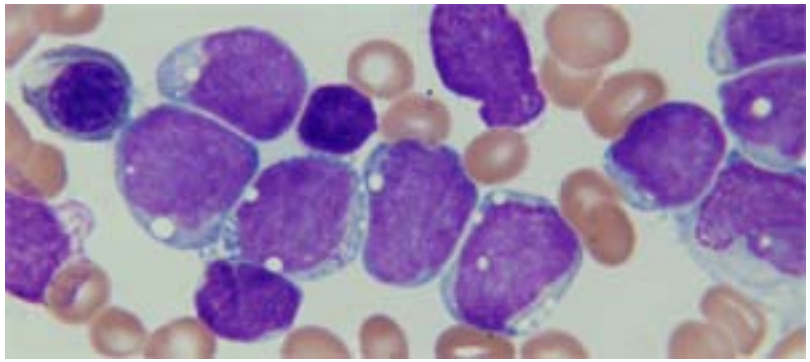
- **Match patient to optimum treatment intensity for his subtype & prognosis**

Treatment Failure

- **Overly intensive treatment leads to**
  - Development of secondary cancers
  - Reduction of IQ
- **Insufficiently intensive treatment leads to**
  - Relapse

Risk Assignment

- **The major subtypes look similar**



- **Conventional diagnosis requires**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics

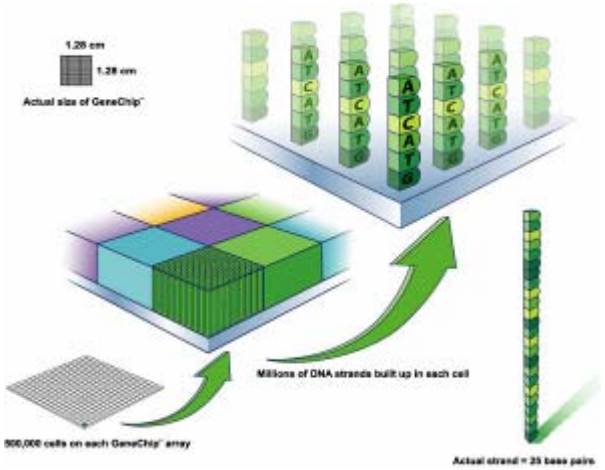
Mission

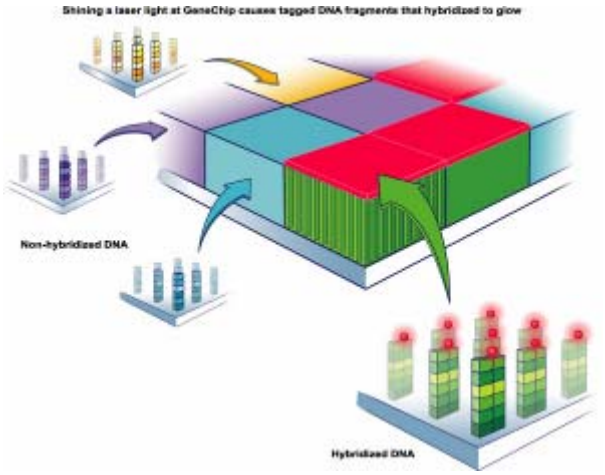
- **Conventional risk assignment procedure requires difficult expensive tests and collective judgement of multiple specialists**

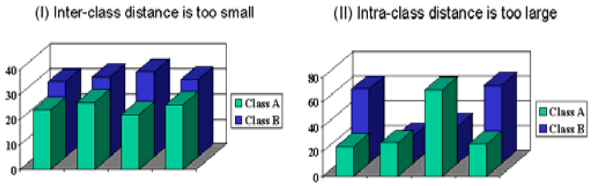
- **Generally available only in major advanced hospitals**

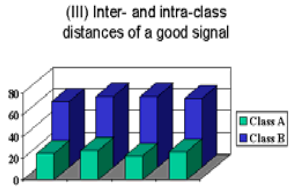
⇒ **Can we have a single-test easy-to-use platform instead?**

Single-Test Platform of Microarray & Machine Learning



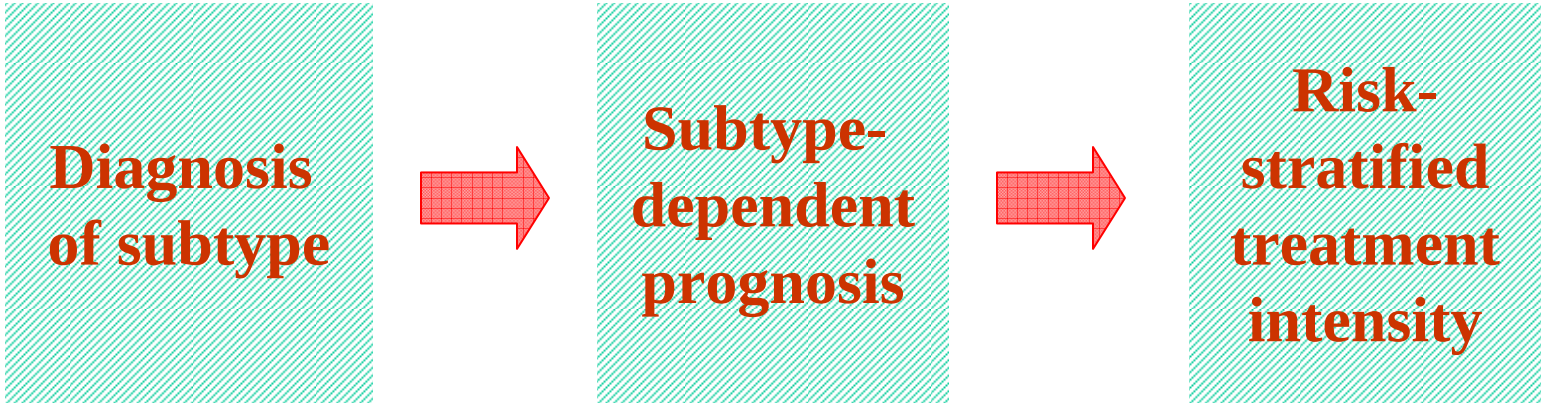






	00-0586-U	00-0586-U	00-0586-U	00-0586-U	00-0586-U	Descriptions
	Positive	Negative	Pairs InAv	Avg Diff	Abs Call	
AFFX-Murl	5	2	19	297.5	A	M16762 Mouse int
AFFX-Murl	3	2	19	554.2	A	M37897 Mouse int
AFFX-Murl	4	2	19	308.6	A	M25892 Mus musc
AFFX-Murl	1	3	19	141	A	M83649 Mus musc
AFFX-BioE	13	1	19	9340.6	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

Overall Strategy



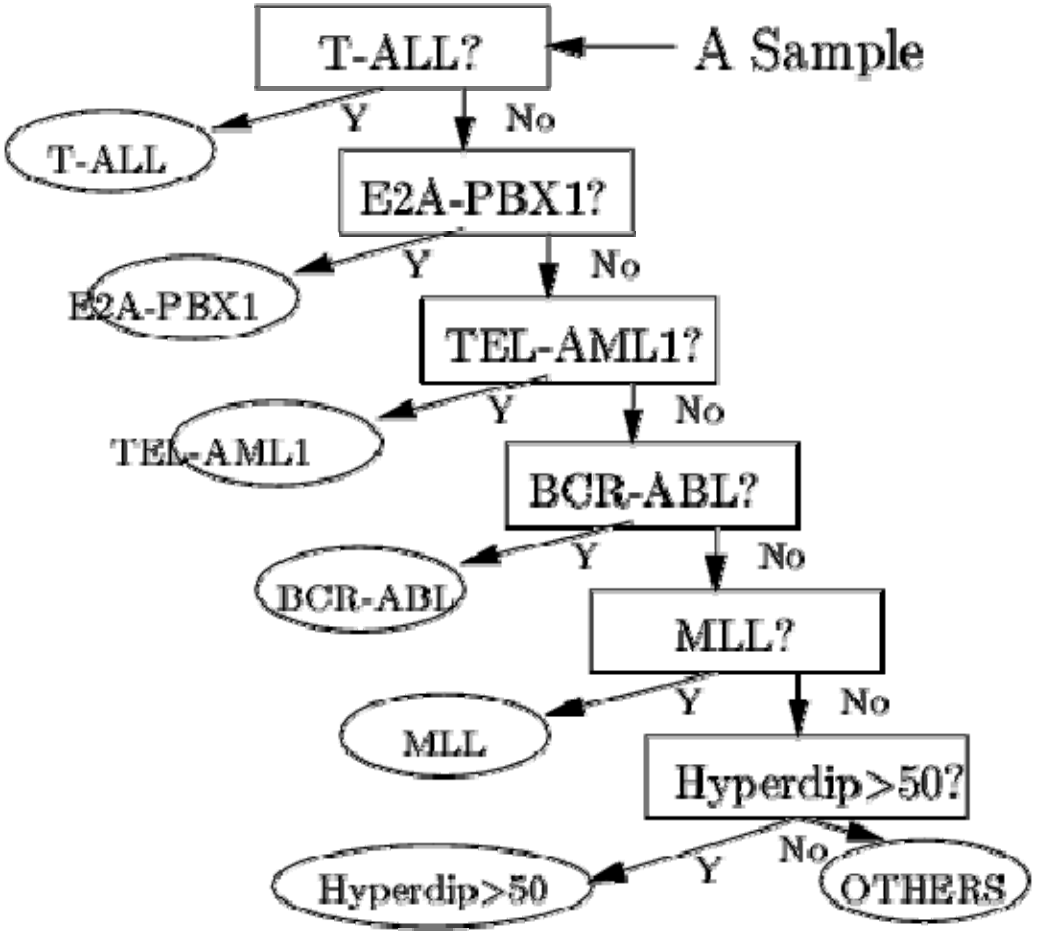
- For each subtype, select genes to develop classification model for diagnosing that subtype
- For each subtype, select genes to develop prediction model for prognosis of that subtype

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**

Childhood ALL Subtype Diagnosis Workflow

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



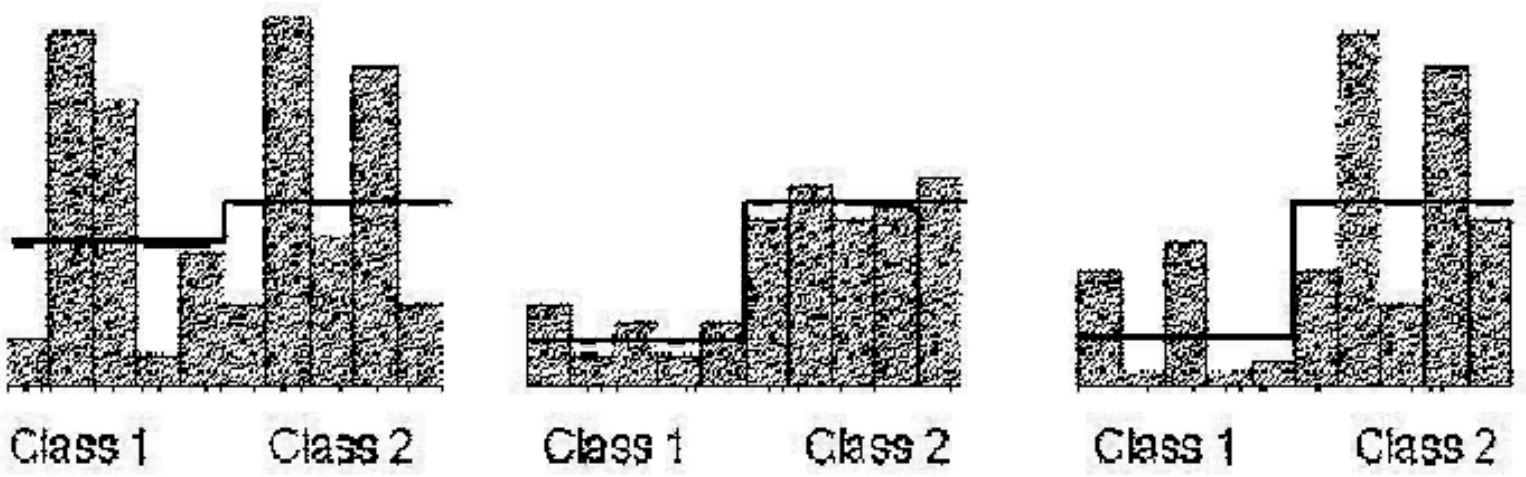
Training and Testing Sets

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

Exercise: Download this data from
http://research.i2r.a-star.edu.sg/rp/Leukemia/Stjude.html
and try your hands on ALL subtype classification using WEKA

Signal Selection Basic Idea

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



Signal Selection by $\chi 2$

The χ^2 value of a signal is defined as:

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

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} ($E_{ij} = R_i * C_j / N$).

Exercise: List the top 10 genes for distinguishing E2A-PBX1 from other ALL subtypes

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**

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

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.

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

PCL Testing

Most freq EP of pos class in the test sample

Exercise: For k =10, what is the ideal $score^P$ and $score^N$?

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

Most freq EP of pos class

Similarly,

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

If $Score^P > Score^N$, then positive class, Otherwise negative class

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

Understandability of PCL

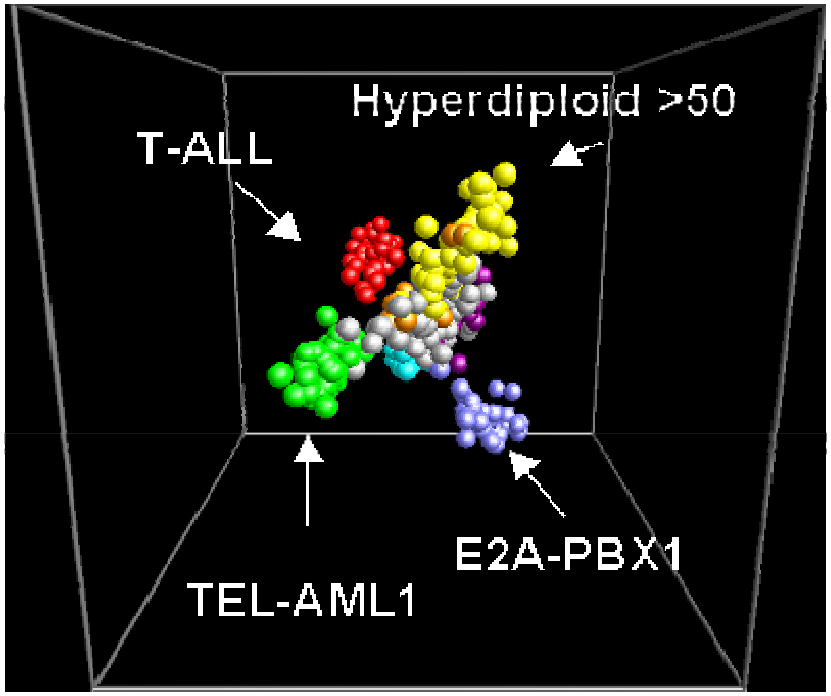
- **E.g., for T-ALL vs. OTHERS, one ideally discriminatory gene 38319_at was found, inducing these 2 EPs**

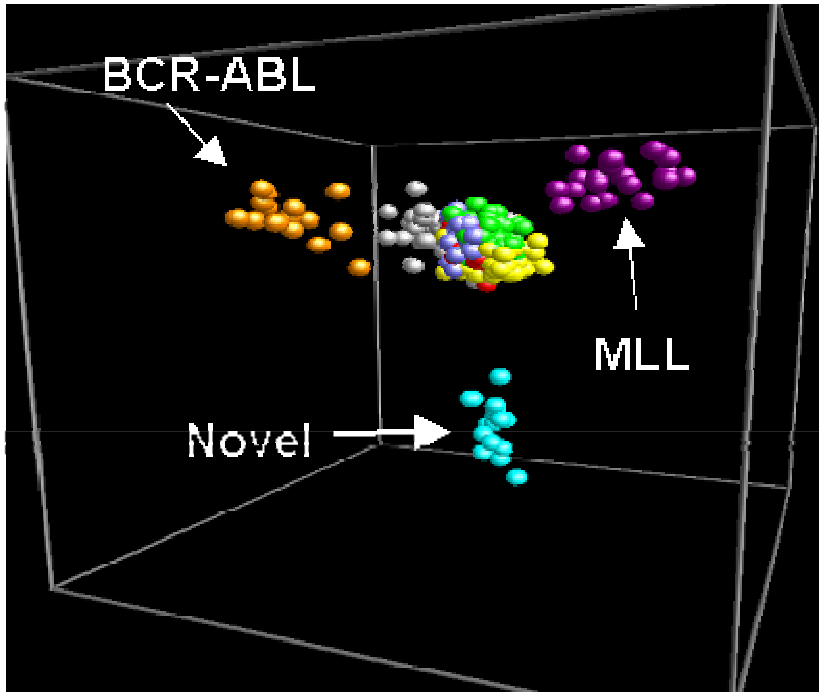
$$\{gene\_(38\,319\_at)\,@\,(-\infty,\ 15\,975.6)\}\ \text{and}$$
$$\{gene\_(38\,319\_at)\,@\,[15\,975.6,\ +\infty)\}.$$

- **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.

Multidimensional Scaling Plot for Subtype Diagnosis



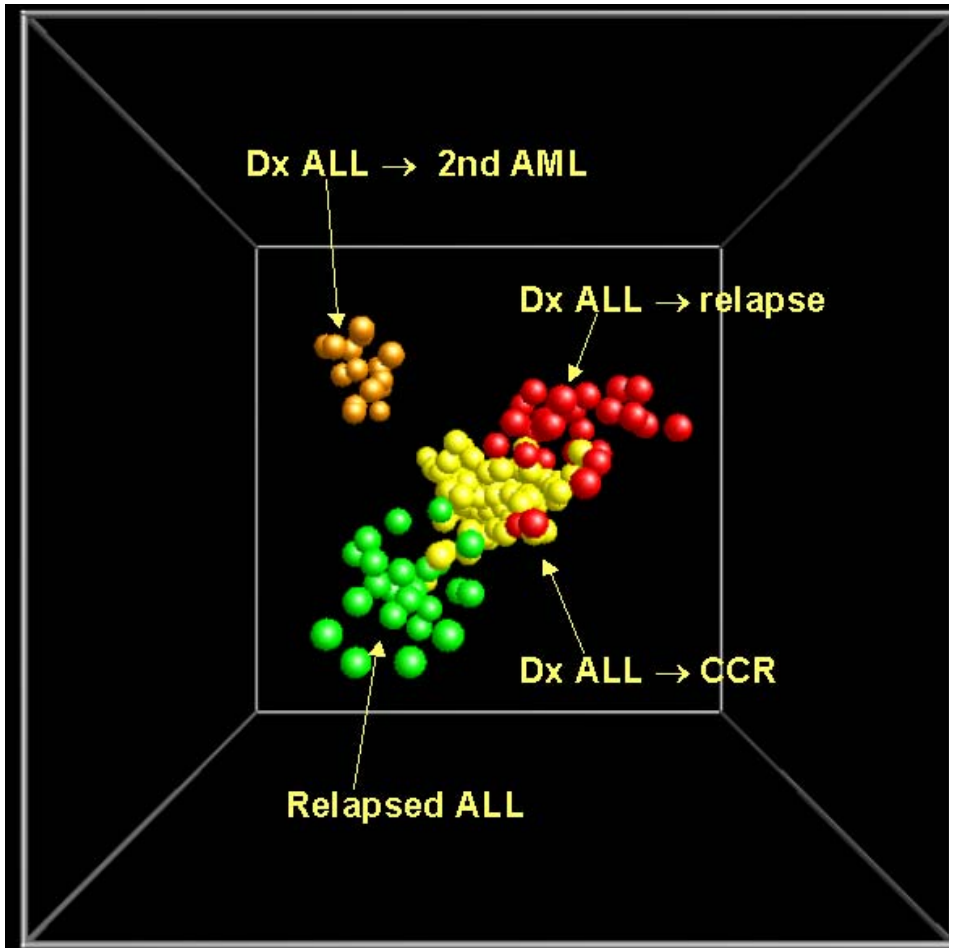


Obtained by performing PCA on the 20 genes chosen for each level

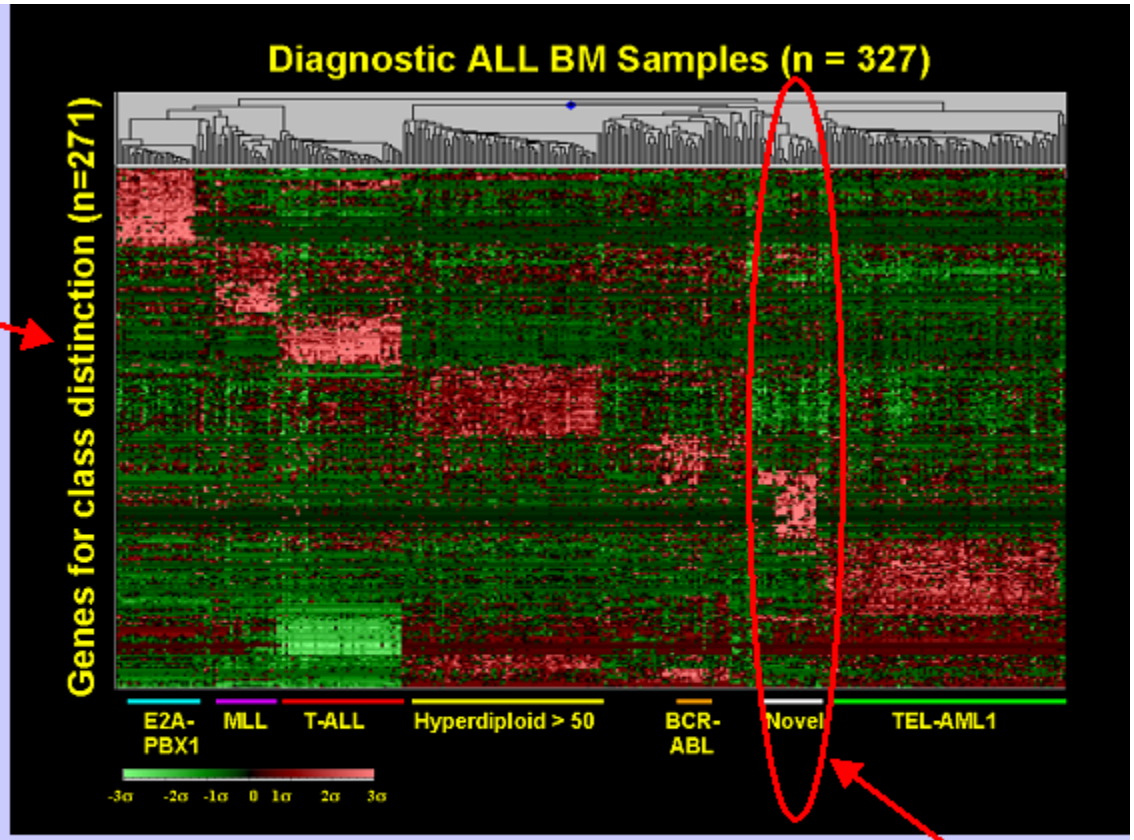
Exercise: What is PCA? Describe the PCA procedure

Multidimensional Scaling Plot Subtype-Dependent Prognosis

- **Similar computational analysis was carried out to predict relapse and/or secondary AML in a subtype-specific manner**
- **>97% accuracy achieved**



Is there a new subtype?



Genes selected by $\chi 2$

New subtype discovered

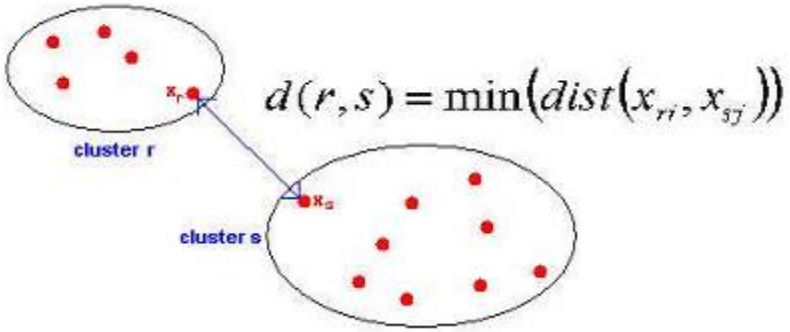
- **Hierarchical clustering of gene expression profiles reveals a novel subtype of childhood ALL**

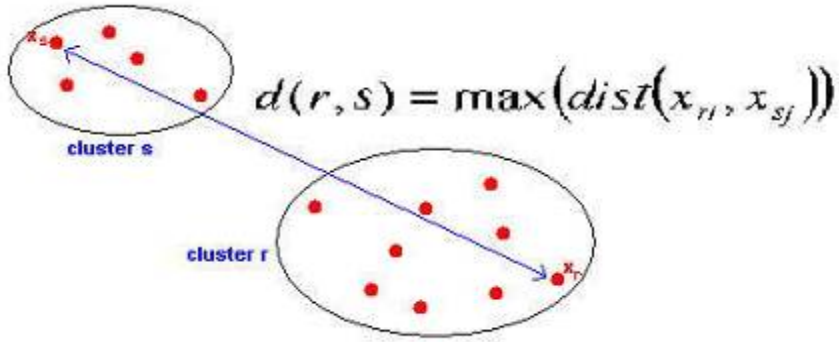
Exercise: Name and describe one bi-clustering method

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**
  - "Similarity" is often defined using
    - **Single linkage**
    - **Complete linkage**
    - **Average linkage**
- **Repeat previous step until all items are clustered into a single cluster of size N**

Single, Complete, & Average Linkage





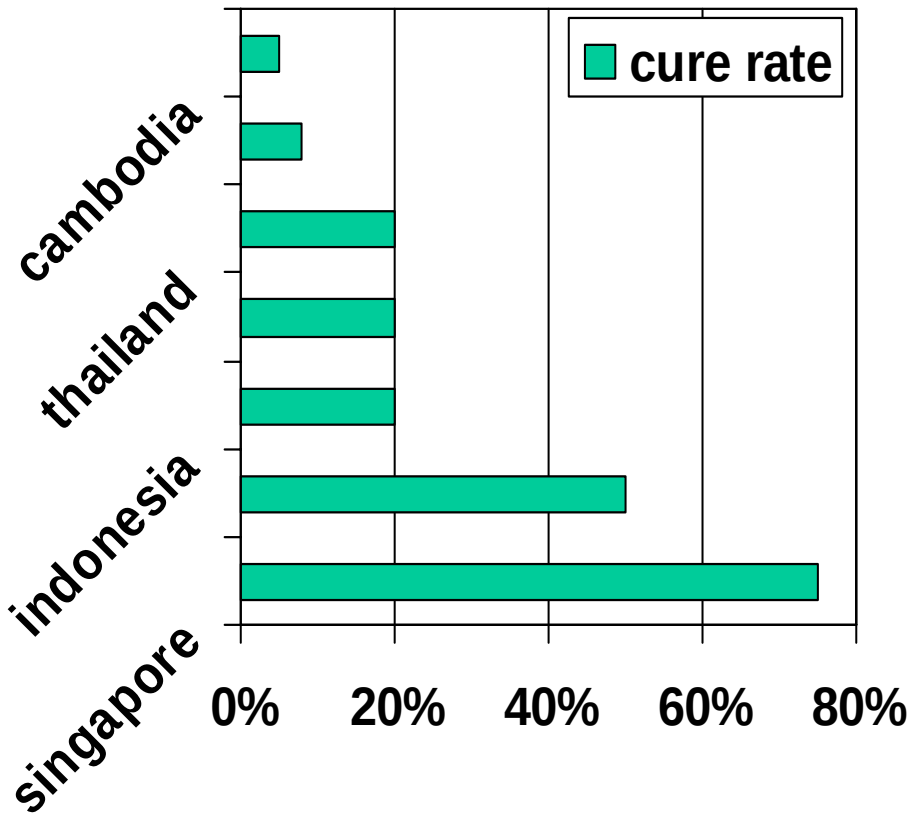
Single linkage defines distance betw two clusters as min distance betw them

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

Exercise: Give definition of "average linkage"

Image source: UCL Microcore Website

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**

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**

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**

⇒ **5-20% cure rates**

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

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

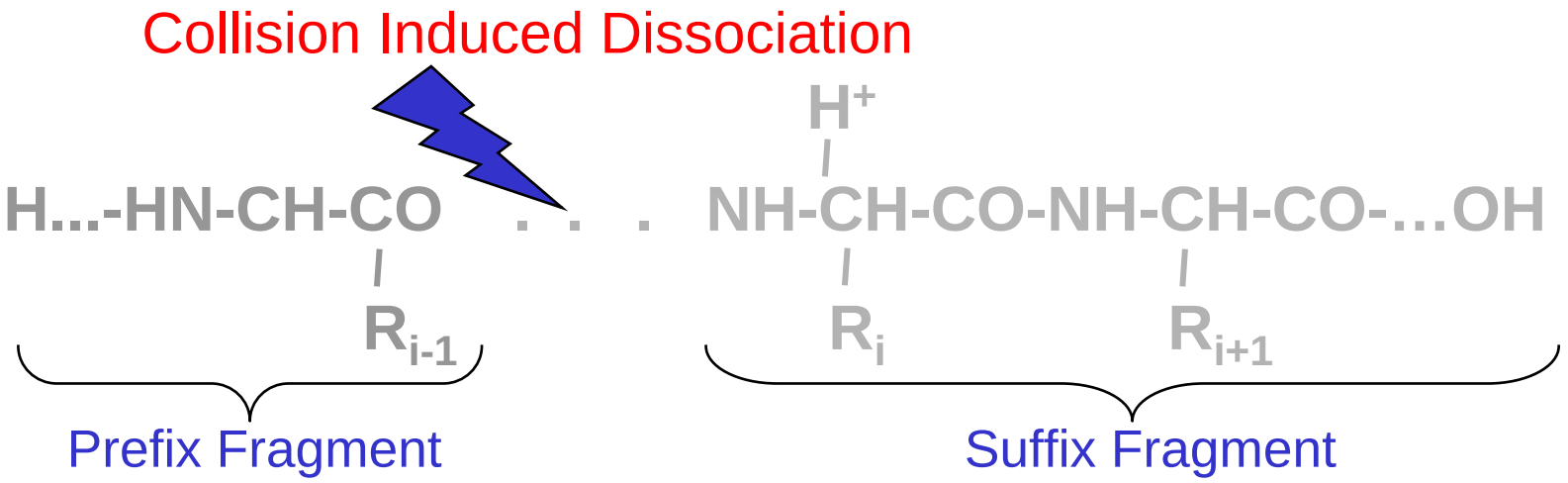
Background on Proteomic Mass-Spec

Motivation for Protein Identification/ Sequencing

- **It is not possible to know the full set of proteins even thought the whole genome is sequenced. Different way of splicing, new undiscovered genes etc.**
- **Important to identify which protein interact in a biological system**
- **Different cells have different expressed protein**

Source: Anthony Tung

Peptide Fragmentation



- **Peptides tend to fragment along the backbone**
- **Fragments can also loose neutral chemical groups like NH_3 and H_2O**

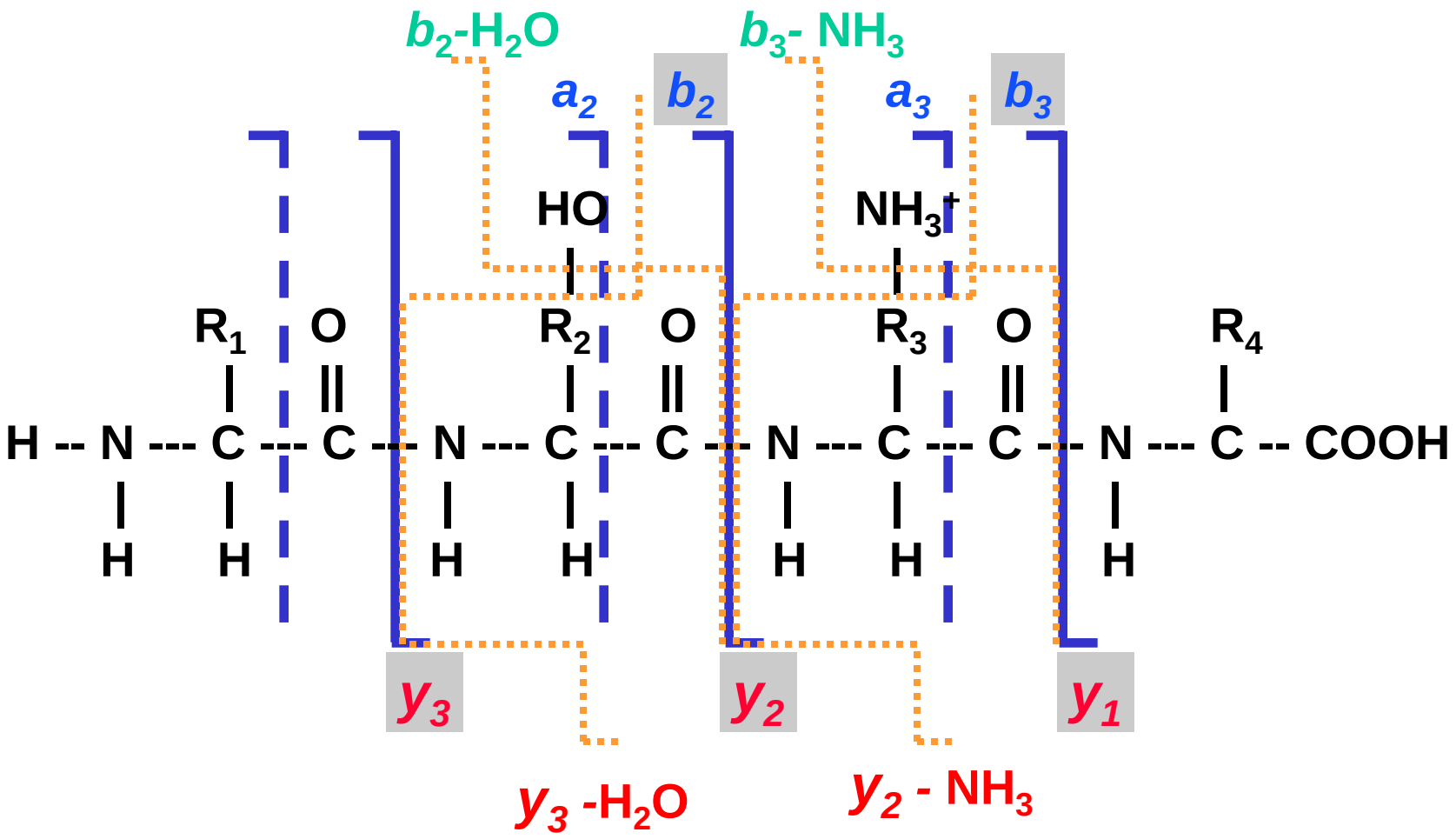
Source: Anthony Tung

Breaking Protein into Peptides and Peptides into Fragment Ions

- **Proteases, e.g. trypsin, break protein into peptides**
- **Tandem Mass Spectrometer further breaks peptides down into fragment ions and measures mass of each piece**
- **Mass Spectrometer accelerates the fragmented ions; heavier ions accelerate slower than lighter ones**
- **Mass Spectrometer measure mass/charge ratio of an ion**

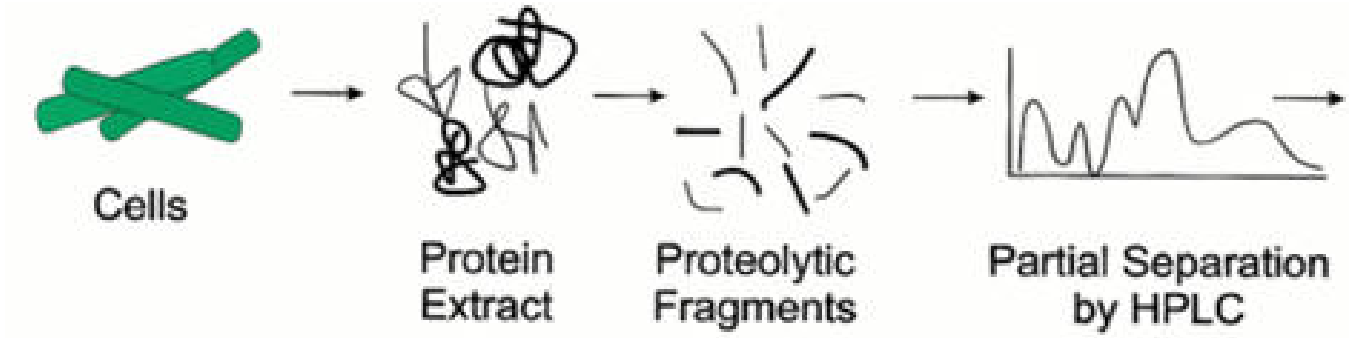
Source: Anthony Tung

Peptide Fragmentation

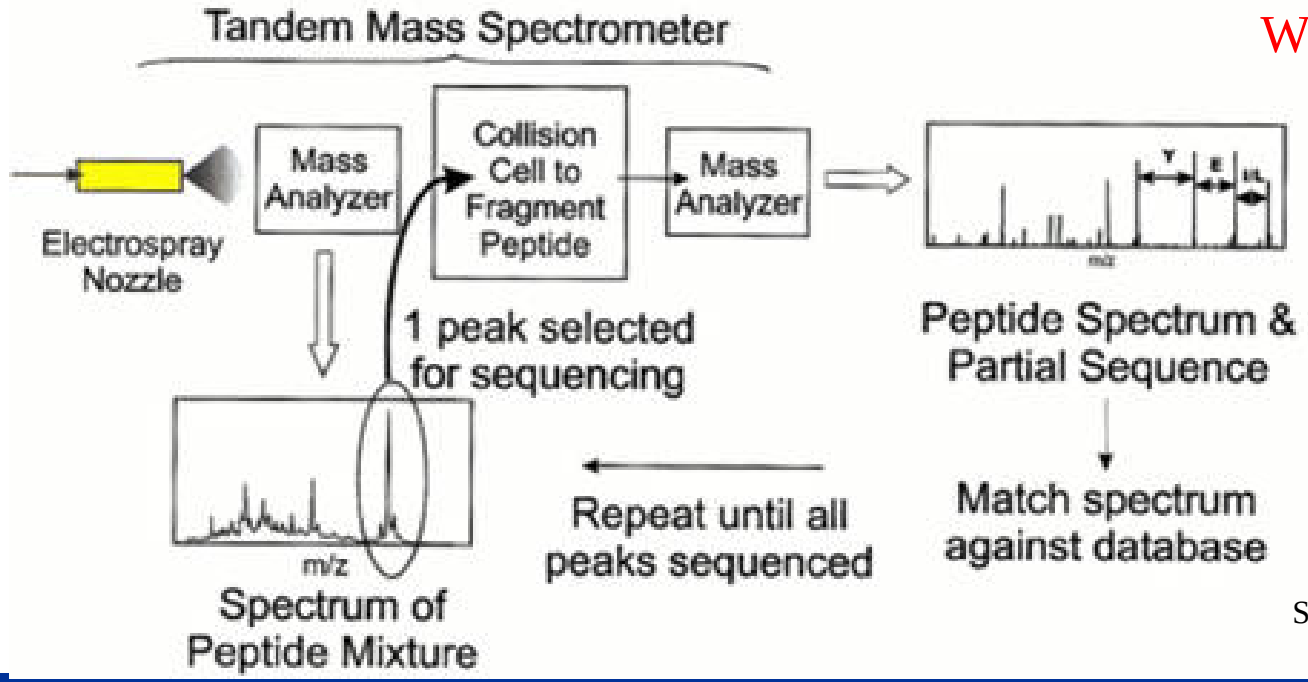


Source: Anthony Tung

Tandem Mass-Spectrometry

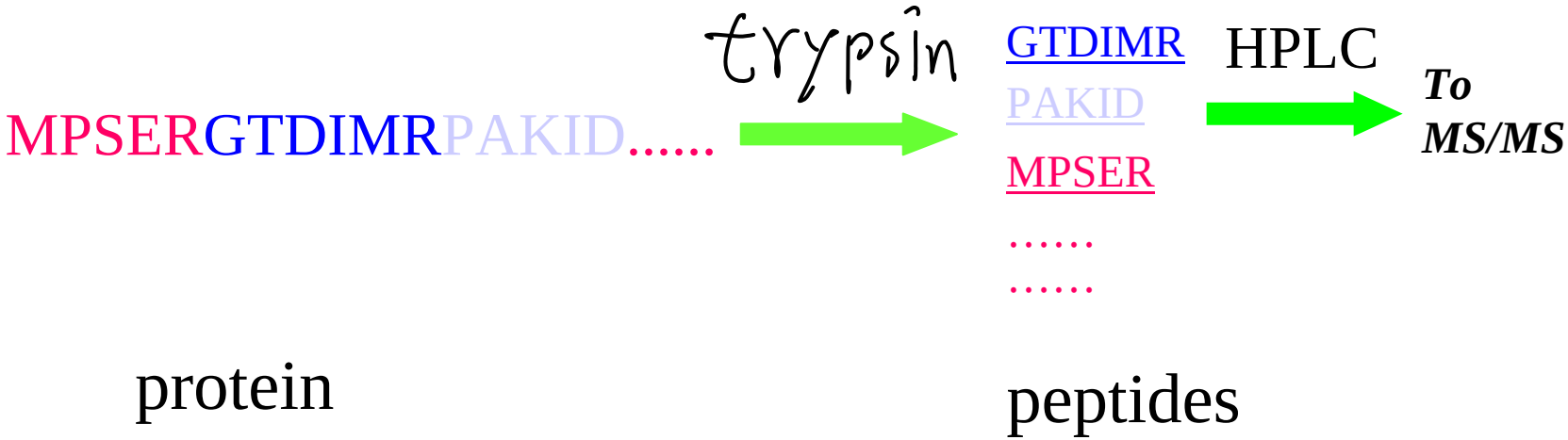


Exercise:
What is HPLC?



Source: Anthony Tung

Breaking Proteins into Peptides



Source: Anthony Tung

Mass Spectrometry

Matrix-Assisted Laser Desorption/Ionization (MALDI)

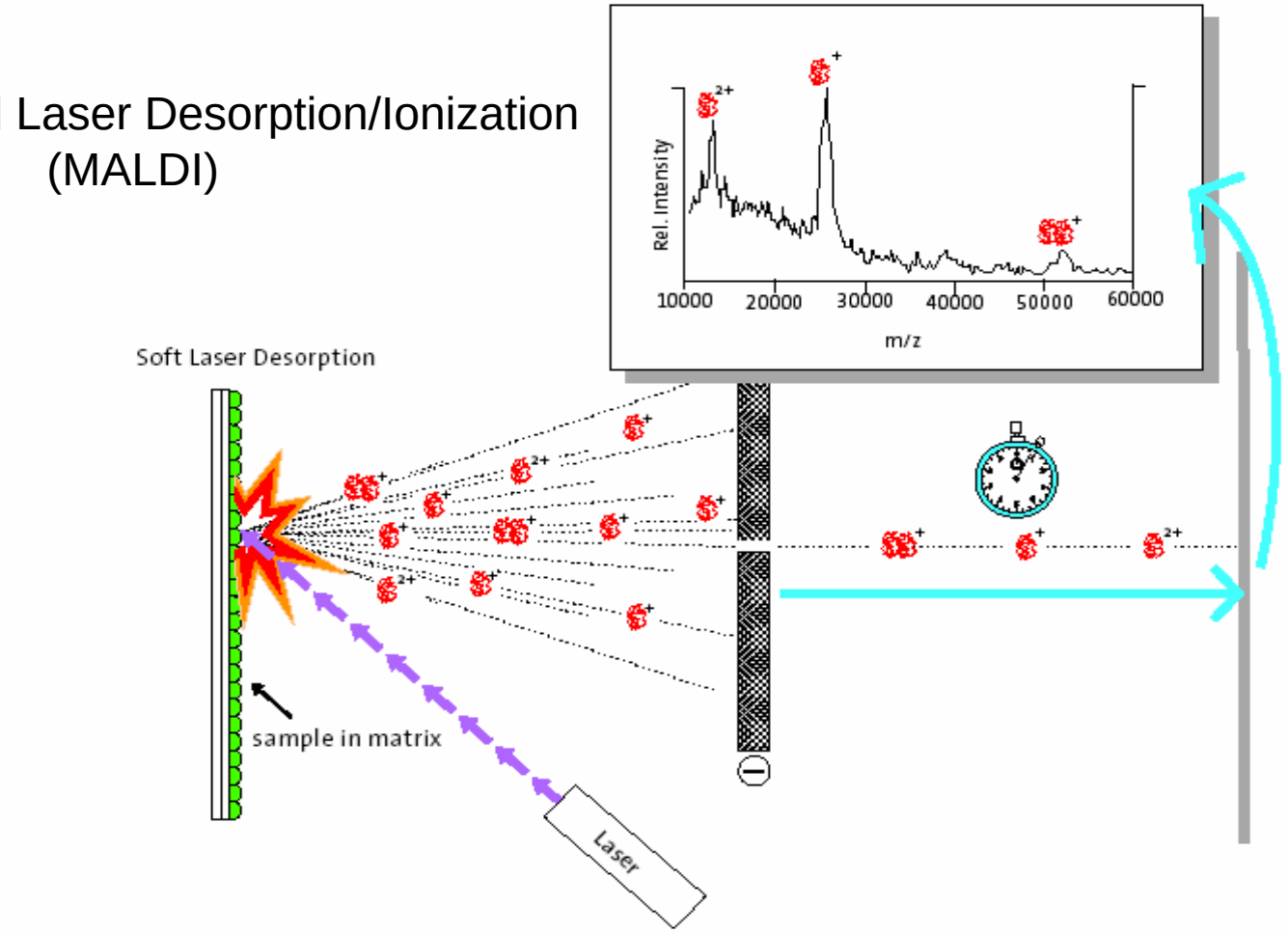
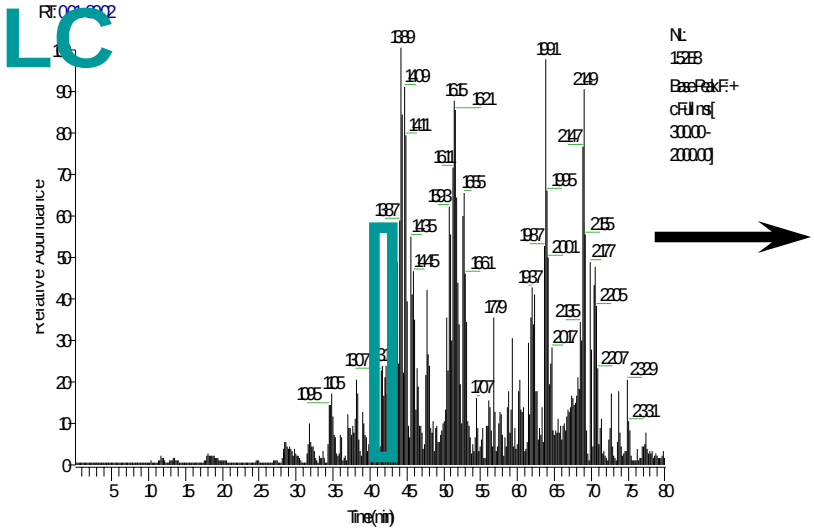
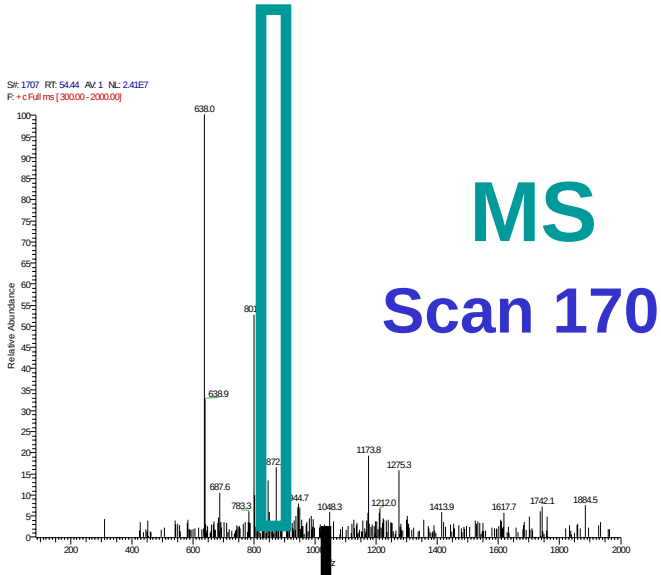


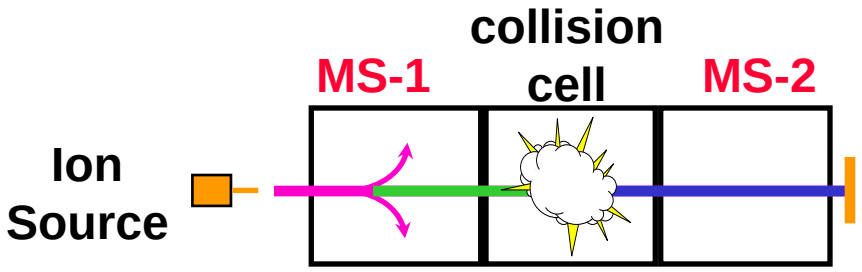
Figure 2. The soft laser desorption process.

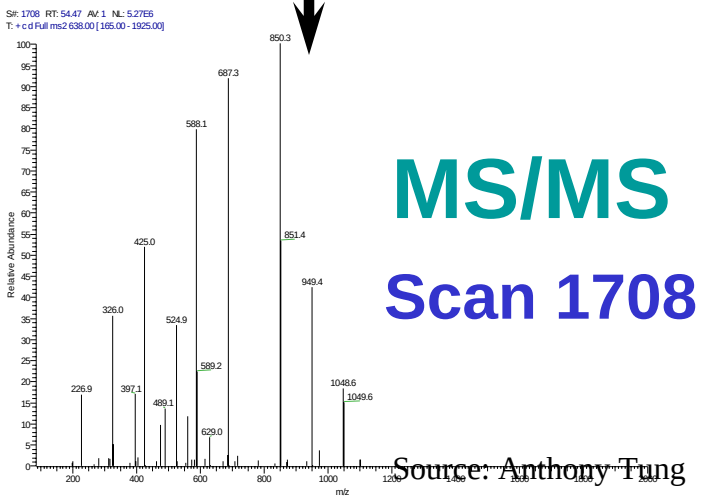
Source: Anthony Tung, & Vineet Bafna

Tandem Mass Spectrometry









Source: Anthony Tung

Proteomic Profile Classification

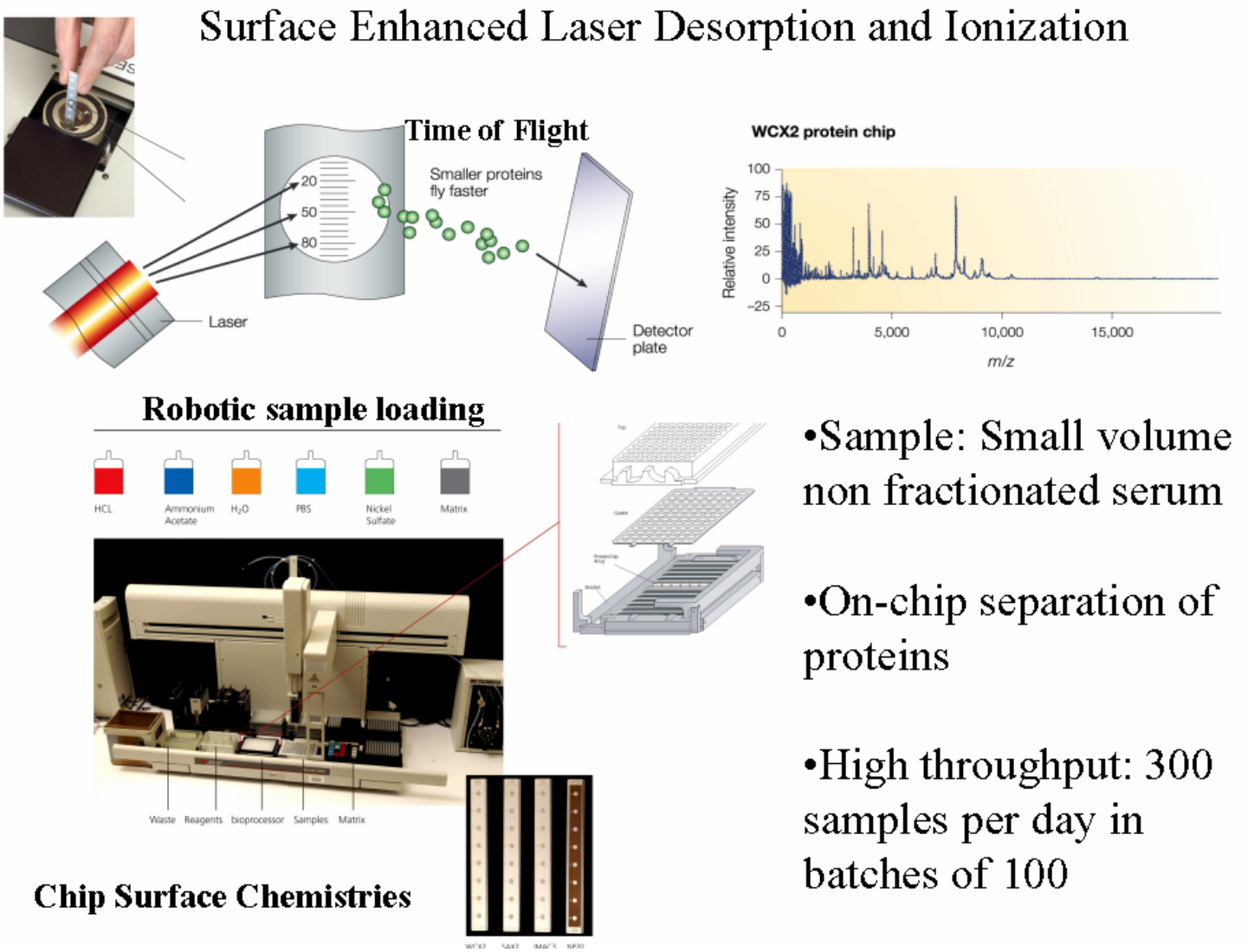
Detection of Ovarian Cancer

Ovarian Cancer Data

Petricoin et al., Lancet 359:572--577, 2002

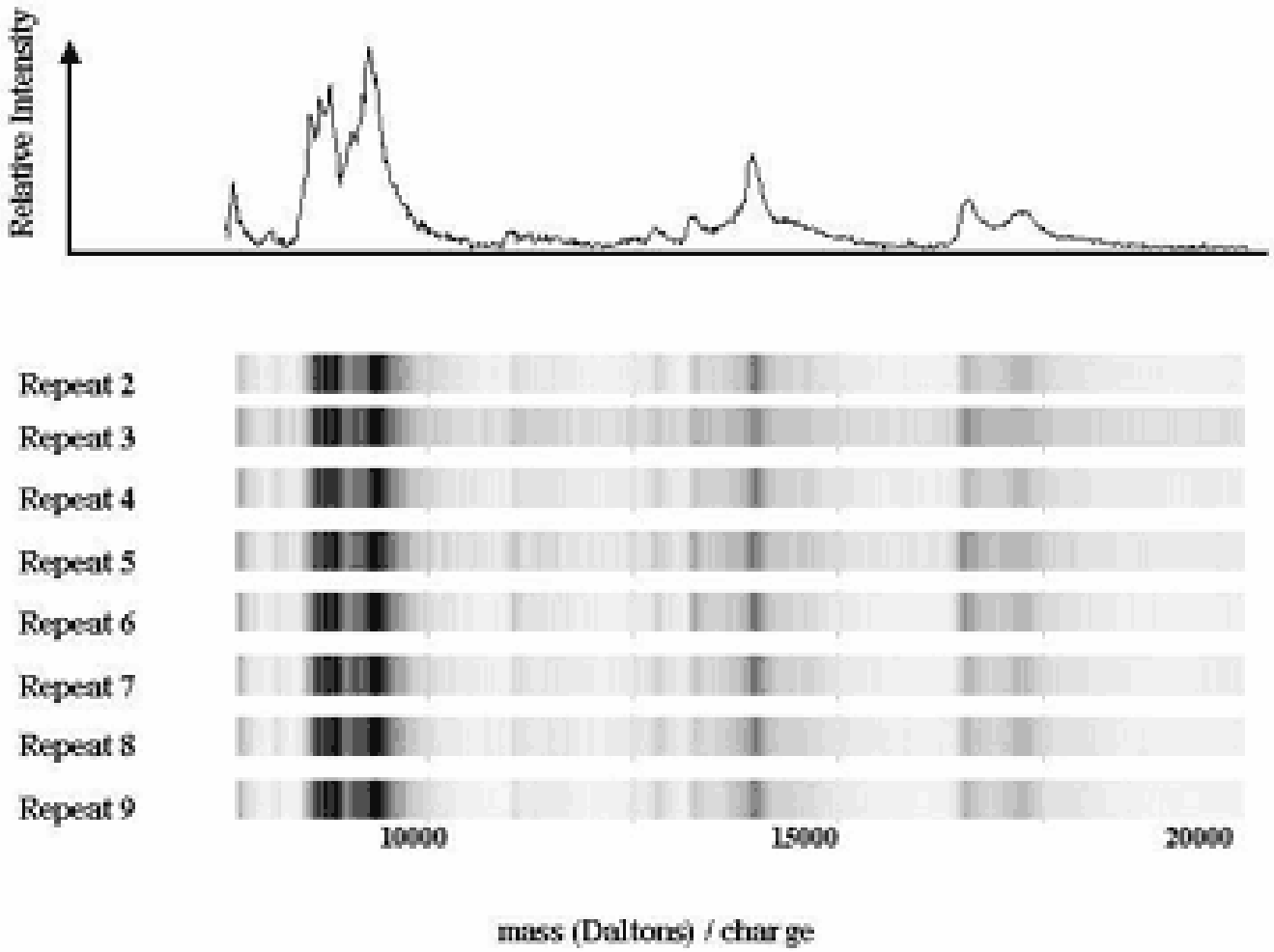
- **Identify proteomic patterns in serum that distinguish ovarian cancer from non-cancer**
- **6-16-02 release**
- **91 non-cancer samples**
- **162 cancer samples**
- **15154 features**
- **Each feature is the amplitude of an ion (aka M/Z identities)**

Proteomic Profiling by Mass Spec



- Sample: Small volume non fractionated serum
- On-chip separation of proteins
- High throughput: 300 samples per day in batches of 100

A Sample Proteomic Profile



Typical Procedure in Analysing Proteomic Profiles for Diagnosis

- **Proteomic data collection**
- **Ion (M/Z) values selection**
- **Classifier training**
- **Classifier tuning (optional for some machine learning methods)**
- **Apply classifier for diagnosis of future cases**

Accuracy

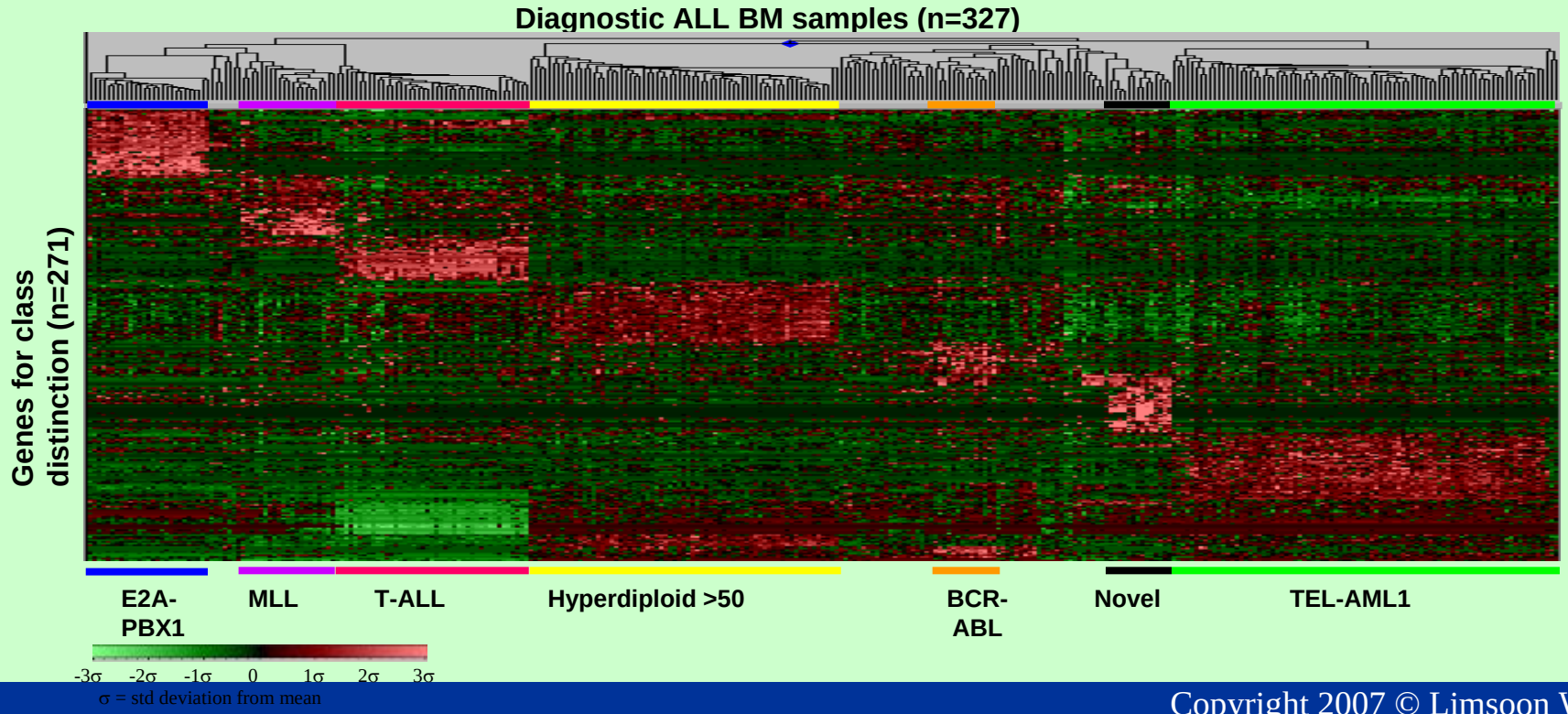
# of features	SVM	NB	*k*-NN	C4.5	PCL
60	0	4	2	5	-
50	0	6	3	6	-
40	1	6	3	4	1
30	4	6	6	5	-
20	5	6	5	10	3
10	8	10	7	9	-

Errors from 10-fold cross validation using the *n* M/Z identities of lowest entropy

Gene Interaction Prediction

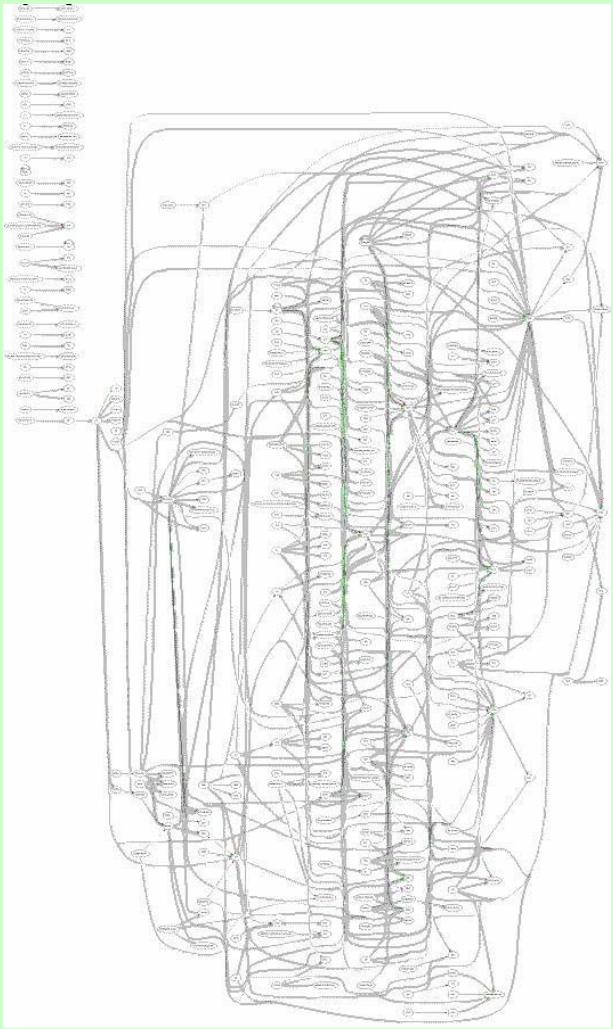
Beyond Classification of Gene Expression Profiles

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



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 data?**



Key Questions

- **For each gene in the network:**
- **Which genes affect it?**
- **How they affect it?**
  - Positively?
  - Negatively?
  - More complicated ways?

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

A Classification-Based Technique

Soinov et al., *Genome Biology* 4:R6.1-9, Jan 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 $x_{ij} \leq a_i$

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**
  - we 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 $\langle s_{ij} \mid i \neq g \rangle$
  - and extract the decision tree or rules used

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**

Deriving Treatment Plan

NUS
National University
of Singapore

Aspire speculation!

Can we do more with EPs?

- **Detect gene groups that are significantly related to a disease**
- **Derive coordinated gene expression patterns from these groups**
- **Derive "treatment plan" based on these patterns**

Colon Tumour Dataset

Alon et al., *PNAS* 96:6745--6750, 1999

- **We use the colon tumour dataset above to illustrate our ideas**
  - 22 normal samples
  - 40 colon tumour samples

Our list	accession number	cutting points
1,2	M26383	59.83
3,4	M63391	1696.22
5,6	R87126	379.38
7,8	M76378	842.30
9,10	H08393	84.87
11,12	X12671	229.99
13,14	R36977	274.96
15,16	J02854	735.80
17,18	M22382	447.04
19,20	J05032	88.90
21,22	M76378	1048.37
23,24	M76378	1136.74
25,26	M16937	390.44
27,28	H40095	400.03
29,30	U30825	288.99
31,32	H43887	334.01
33,34	H51015	84.19
35,36	X57206	417.30
37,38	R10066	494.17
39,40	T96873	75.42
41,42	T57619	2597.85
43,44	R84411	735.57
45,46	U21090	232.74
47,48	U32519	87.58
49,50	T71025	1695.98
51,52	T92451	845.7
53,54	U09564	120.38
55,56	H40560	913.77
57,58	T47377	629.44
59,60	X53586	121.91
61,62	U25138	186.19
63,64	T60155	1798.65
65,66	H55758	1453.15
67,68	Z50753	196.12
69,70	U09587	486.17

Detect Gene Groups

- **Feature Selection**
  - Use entropy method
  - 35 genes have cut points
- **Generate EPs**
  - 9450 EPs in normals
  - 1008 EPs in tumours
- **EPs with largest support are gene groups significantly co-related to disease**

Top 20 EPs

Emerging patterns	Count & Freq. (%) in normal tissues
{25, 33, 37, 41, 43, 57, 59, 69}	17(77.27%)
{25, 33, 37, 41, 43, 47, 57, 69}	17(77.27%)
{29, 33, 35, 37, 41, 43, 57, 69}	17(77.27%)
{29, 33, 37, 41, 43, 47, 57, 69}	17(77.27%)
{29, 33, 37, 41, 43, 57, 59, 69}	17(77.27%)
{25, 33, 35, 37, 41, 43, 57, 69}	17(77.27%)
{33, 35, 37, 41, 43, 57, 65, 69}	17(77.27%)
{33, 37, 41, 43, 47, 57, 65, 69}	17(77.27%)
{33, 37, 41, 43, 57, 59, 65, 69}	17(77.27%)
{33, 35, 37, 41, 43, 45, 57, 69}	17(77.27%)
{33, 37, 41, 43, 45, 47, 57, 69}	17(77.27%)
{33, 37, 41, 43, 45, 57, 59, 69}	17(77.27%)
{13, 33, 35, 37, 43, 57, 69}	17(77.27%)
{13, 33, 37, 43, 47, 57, 69}	17(77.27%)
{13, 33, 37, 43, 57, 59, 69}	17(77.27%)
{13, 32, 37, 57, 69}	17(77.27%)
{33, 35, 37, 57, 68}	17(77.27%)
{33, 37, 47, 57, 68}	17(77.27%)
{33, 37, 57, 59, 68}	17(77.27%)
{32, 37, 41, 57, 69}	17(77.27%)

Emerging patterns	Count & Freq. (%) in cancer tissues
{2, 10}	28 (70.00%)
{10, 61}	27 (67.50%)
{10, 20}	27 (67.50%)
{3, 10}	27 (67.50%)
{10, 21}	27 (67.50%)
{10, 23}	27 (67.50%)
{7, 40, 56}	26 (65.00%)
{2, 56}	26 (65.00%)
{12, 56}	26 (65.00%)
{10, 63}	26 (65.00%)
{3, 58}	26 (65.00%)
{7, 58}	26 (65.00%)
{15, 58}	26 (65.00%)
{23, 58}	26 (65.00%)
{58, 61}	26 (65.00%)
{2, 58}	26 (65.00%)
{20, 56}	26 (65.00%)
{21, 58}	26 (65.00%)
{15, 40, 56}	25 (62.50%)
{21, 40, 56}	25 (62.50%)

Observation

- **Some EPs contain large number of genes and still have high freq**
- **E.g., {25, 33, 37, 41, 43, 57, 59, 69} has freq 72.27% in normal and 0% in cancer samples**
- **I.e., almost every normal cell's gene expression values satisfy all conds. implied by these 8 items**

genes	expression interval
M16937	<390.44
H51015	<84.19
R10066	<494.17
T57619	<2597.85
R84411	<735.57
T47377	<629.44
X53586	<121.91
U09587	<486.17

Treatment Plan Idea

- **Increase/decrease expression level of particular genes in a cancer cell so that**
  - it has the common EPs of normal cells
  - it has no common EPs of cancer cells

Treatment Plan Example

- **From the EP {25,33,37,41,43,57,59,69}**
  - 77% of normal cells express the 8 genes (M16937, H51015, R10066, T57619, R84411, T47377, X53586, U09587) in the corr. Intervals
  - a cancer cell never express all 8 genes in the same way
  - if expression level of improperly expressed genes can be adjusted, the cancer cell can have one common EP of normal cells
  - a cancer cell can then be iteratively converted into a normal one

Choosing Genes to Adjust

- **Consider tumour cell T1**
- **77% of normal cells have this EP**

genes	expression interval
M16937	<390.44
H51015	<84.19
R10066	<494.17
T57619	<2597.85
R84411	<735.57
T47377	<629.44
X53586	<121.91
U09587	<486.17

genes	expression levels in T1
M16937	369.92
H51015	137.39
R10066	354.97
T57619	1926.39
R84411	798.28
T47377	662.06
X53586	136.09
U09587	672.20

If H51015, R84411, T47377, X53586, U09587 in T1 can be down regulated so T1 now contains the EP above, then T1 will have one more common property of normal cells

Doing more adjustments...

Interestingly, the expression change of the 5 genes in **T1** leads to a chain of other changes. These include the change that 9 extra top-ten EPs of normal cells are contained in the adjusted **T1**. So all top-ten EPs of normal cells are contained in **T1** if the 5 genes' expression levels are adjusted. As the average number of top-ten EPs contained in normal cells is 7, the changed **T1** cell will now be considered as a cell that has the most important features of normal cells. Note that we have adjusted only 5 genes' expression level

Next, eliminate common EPs of cancer cells in T1

It is also necessary to eliminate those common properties of cancer cells that are contained in **T1**. By adjusting the expression level of 2 other genes, M26383 and H08393, the top-ten EPs of cancer cells all disappear from **T1**. According to the colon tumor dataset, the average number of top-ten EPs of cancer cells contained in a cancer cell is 6. Therefore, **T1** is converted into a normal cell as it now holds the common properties of normal cells and does not hold the common properties of cancer cells.

"Treatment Plan" Validation

- **"Adjustments" were made to the 40 colon tumour samples based on EPs as described**
- **Classifiers trained on original samples were applied to the adjusted samples**

classifier	no. of misclassifications in original samples	no. of adjusted tumour samples classified as normal
SVM	6	40
HyperPipes	5	39
Voting Feature Intervals	3	39

It works!

A Big But...

- **Effective means for identifying mechanisms and pathways through which to modulate gene expression of selected genes need to be developed**

Any Question?

Acknowledgements

- **Some slides for this lectures are adapted from those given to me by Jinyan Li and Anthony Tung**

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
- E.F. Petricoin et al., “Use of proteomic patterns in serum to identify ovarian cancer”, *Lancet*, 359:572--577, 2002
- U.Alon et al., “Broad patterns of gene expression revealed by clustering analysis of tumor colon tissues probed by oligonucleotide arrays”, *PNAS* 96:6745--6750, 1999
- J.Li, L. Wong, “Geography of differences between two classes of data”, *Proc. 6th European Conf. on Principles of Data Mining and Knowledge Discovery*, pp. 325--337, 2002
- J.Li, L. Wong, “Identifying good diagnostic genes or gene groups from gene expression data by using the concept of emerging patterns”, *Bioinformatics*, 18:725--734, 2002

References

- J.Li et al., “A comparative study on feature selection and classification methods using a large set of gene expression profiles”, *GIW*, 13:51--60, 2002
- M. A. Hall, “Correlation-based feature selection machine learning”, PhD thesis, Dept of Comp. Sci., Univ. of Waikato, New Zealand, 1998
- U. M. Fayyad, K. B. Irani, “Multi-interval discretization of continuous-valued attributes”, *IJCAI* 13:1022-1027, 1993
- H. Liu, R. Sentiono, “Chi2: Feature selection and discretization of numeric attributes”, *IEEE Intl. Conf. Tools with Artificial Intelligence* 7:338--391, 1995
- 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