

# Microarray (I)

Ueng-Cheng Yang

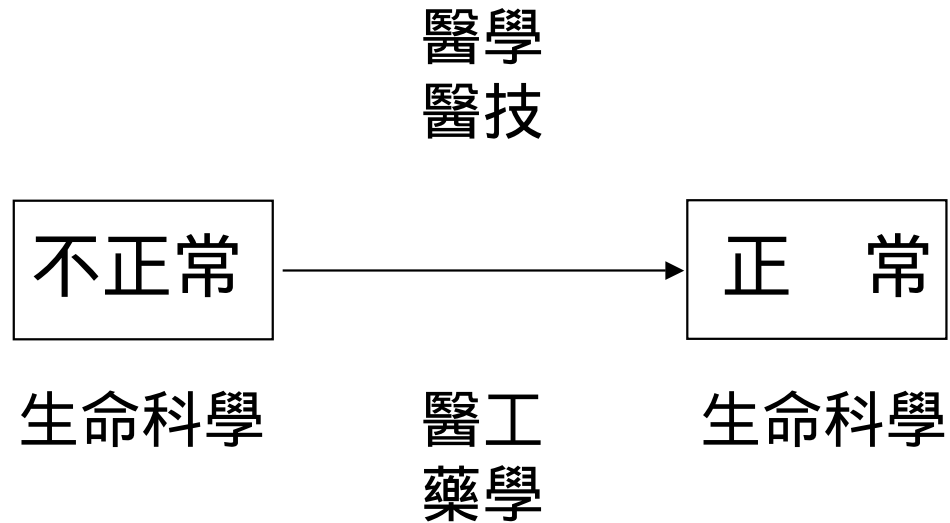
Bioinformatics Research Center

National Yang-Ming University

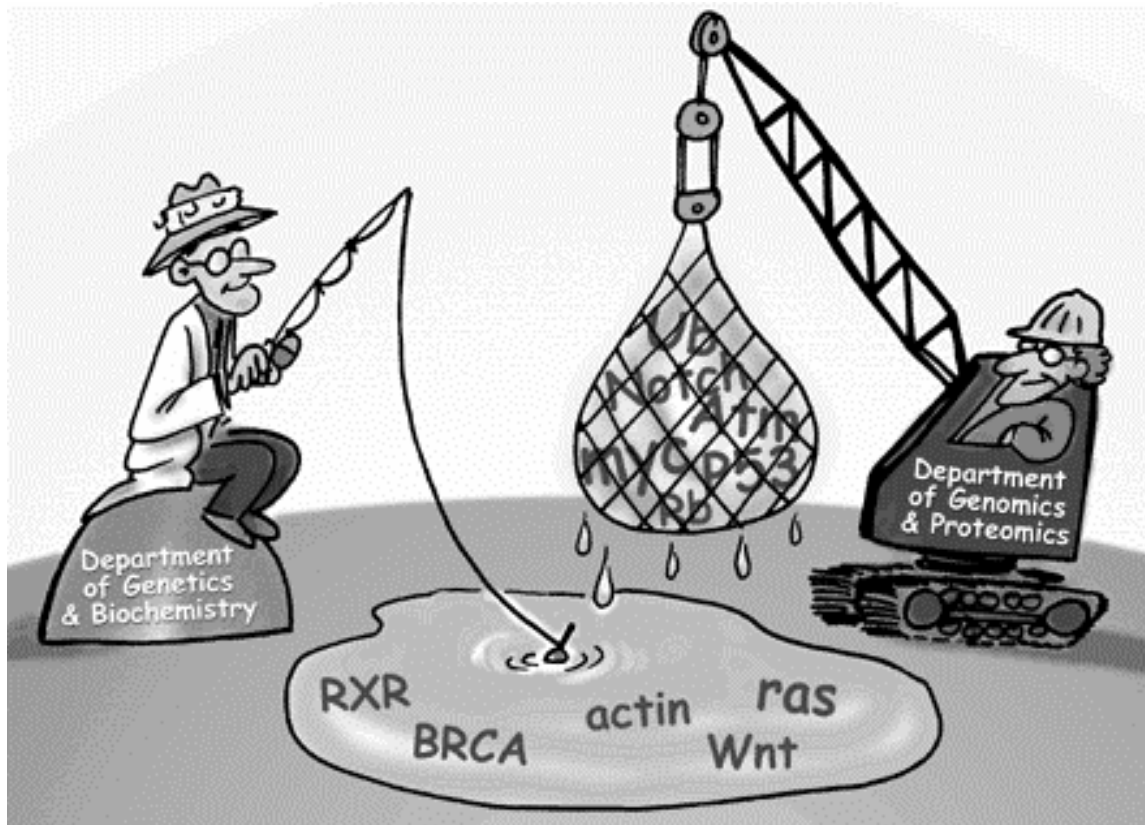
# What you will learn in this course?

- How to use the bioinformatics tools to analyze the data generated by high-throughput techniques?
- How to perform all kinds of statistical analysis by using S-plus?
- How to use the YM tools to setup a user-centric bioinformatics environment by using GeneSpring<sup>TM</sup> as a viewer?

# 醫學與生命科學的關係



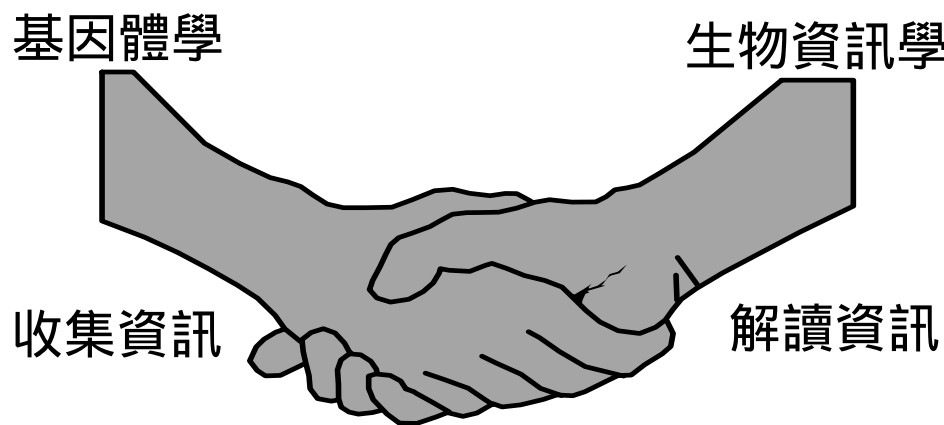
# 傳統分析與巨量分析的對比



<http://www.sciencemag.org/cgi/content/full/291/5507/1221/F1>

YM-Biochem

# 生物資訊學是跨領域的新學門，需整合生物、 數學、理化、統計、與資訊技術



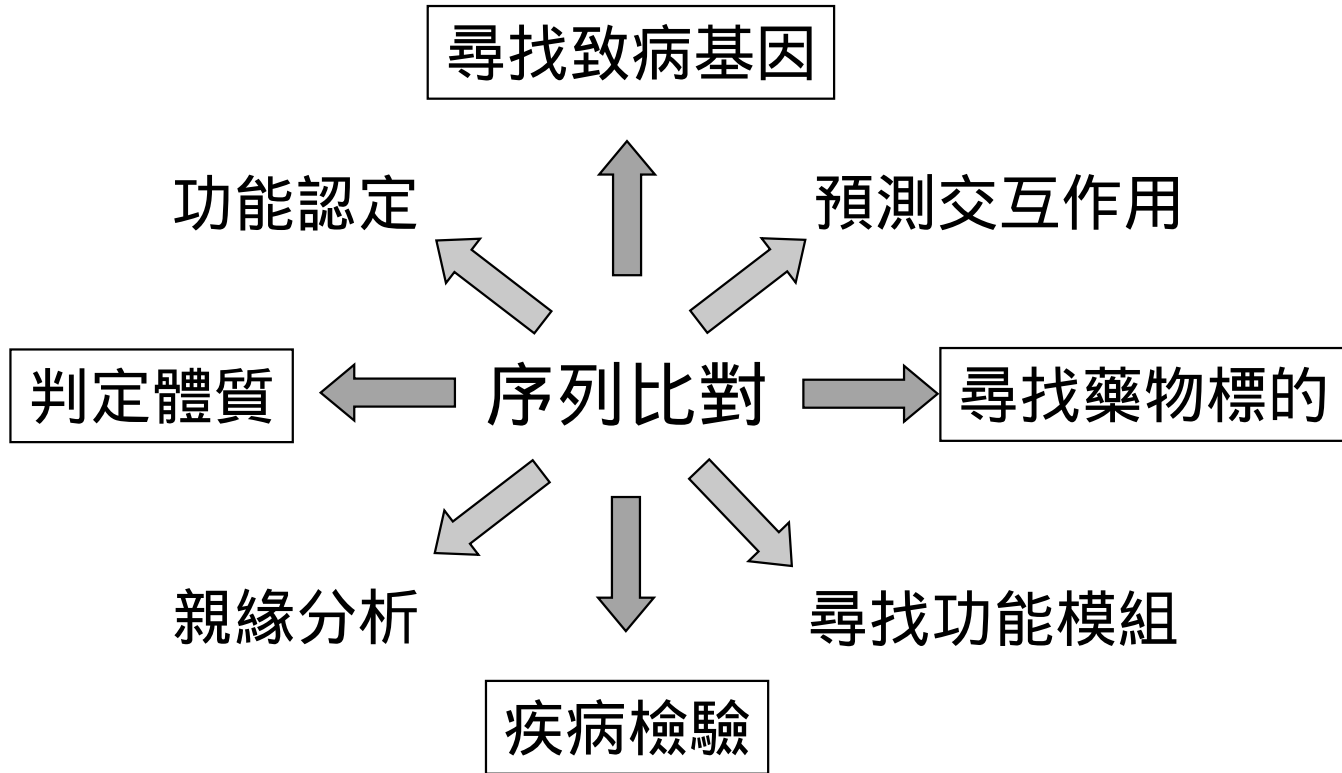
↓ 生物資訊學是發展生物  
科技的催化劑

數據 => 資訊 => 知識 => 技術 => 經濟

# 生物資訊的核心策略



需要生物醫學專業知識



# 生醫應用是發展生物 資訊學的動力

- 尋找致病基因
- 發現新基因
- 尋找藥物作用標的
- 疾病診斷
  - 基因晶片設計與數據分析
  - 個人化的醫療(SNP的生醫應用)

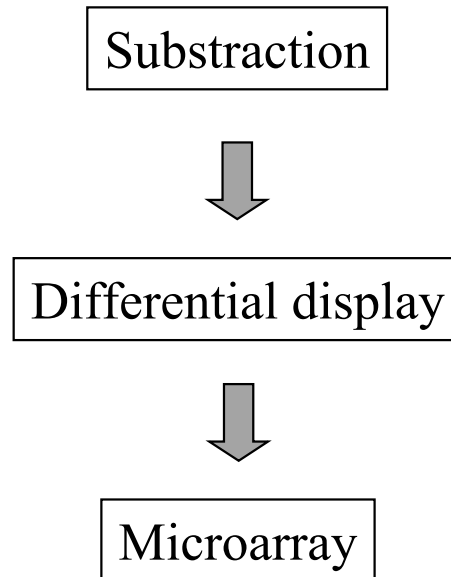
# 研究策略

尋找差異

判定因果關係

找到主控因子/步驟

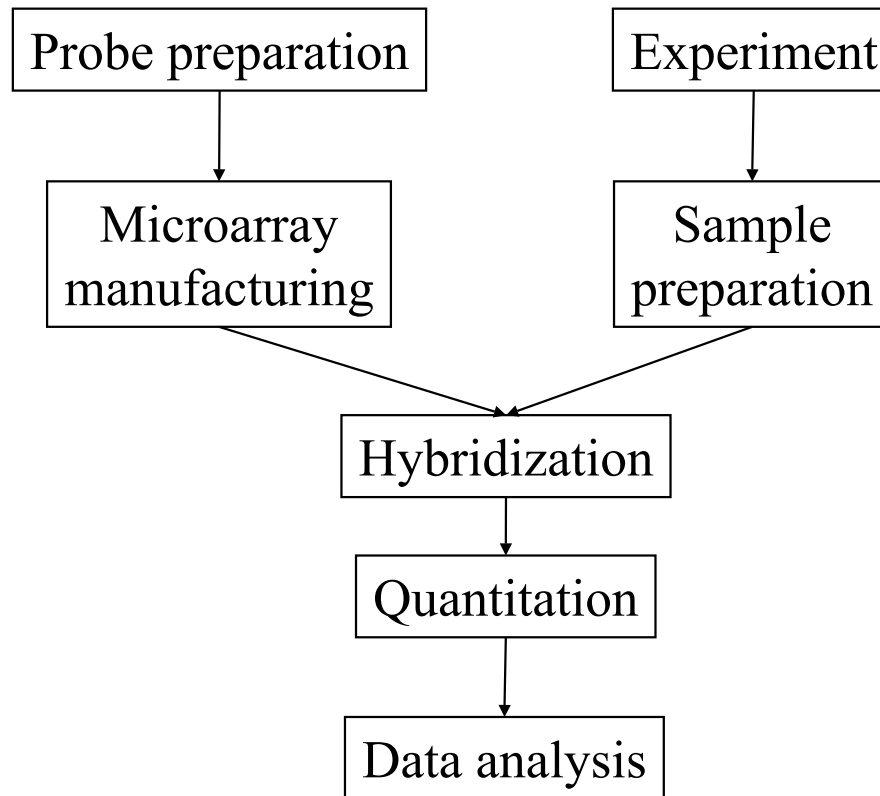
# Evolution of Concepts



# Microarrays

- Oligonucleotide array
  - Synthesized on a chip: *e.g.* Affymetrix
  - Spot on a solid matrix: *e.g.* Compugen
- cDNA array: *e.g.* Incyte

# Major steps in microarray analysis

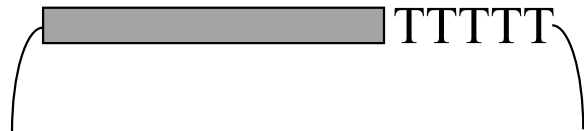


# Jargon

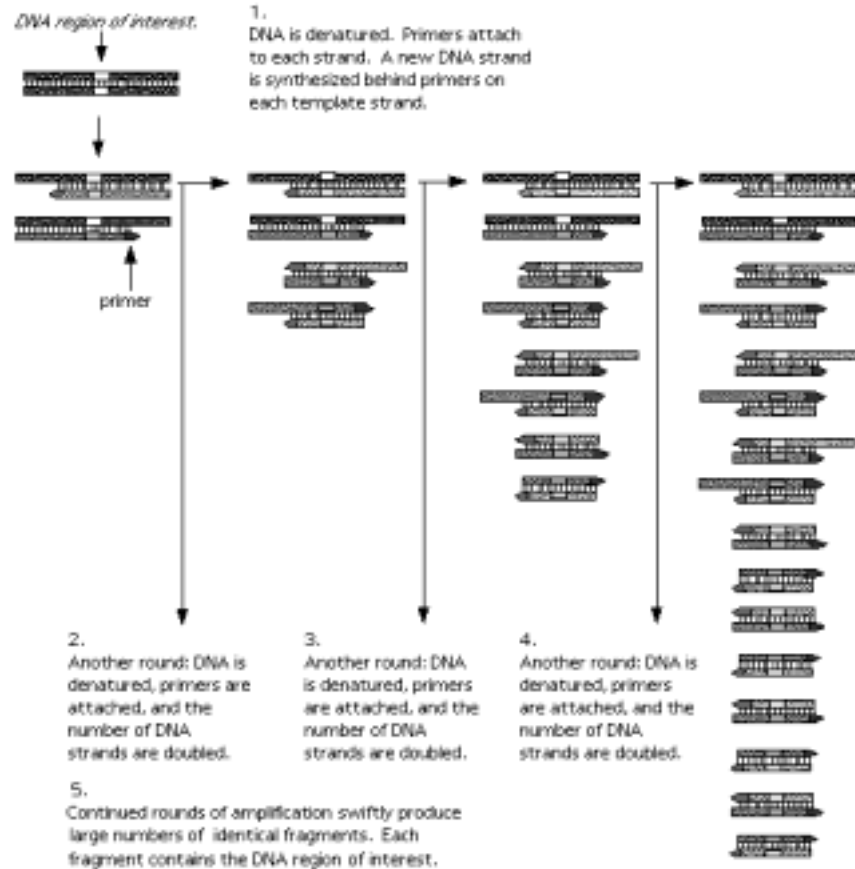
Target: derived from mRNA

Probe: DNA on the solid support

# cDNA Probe preparation



## POLYMERASE CHAIN REACTION



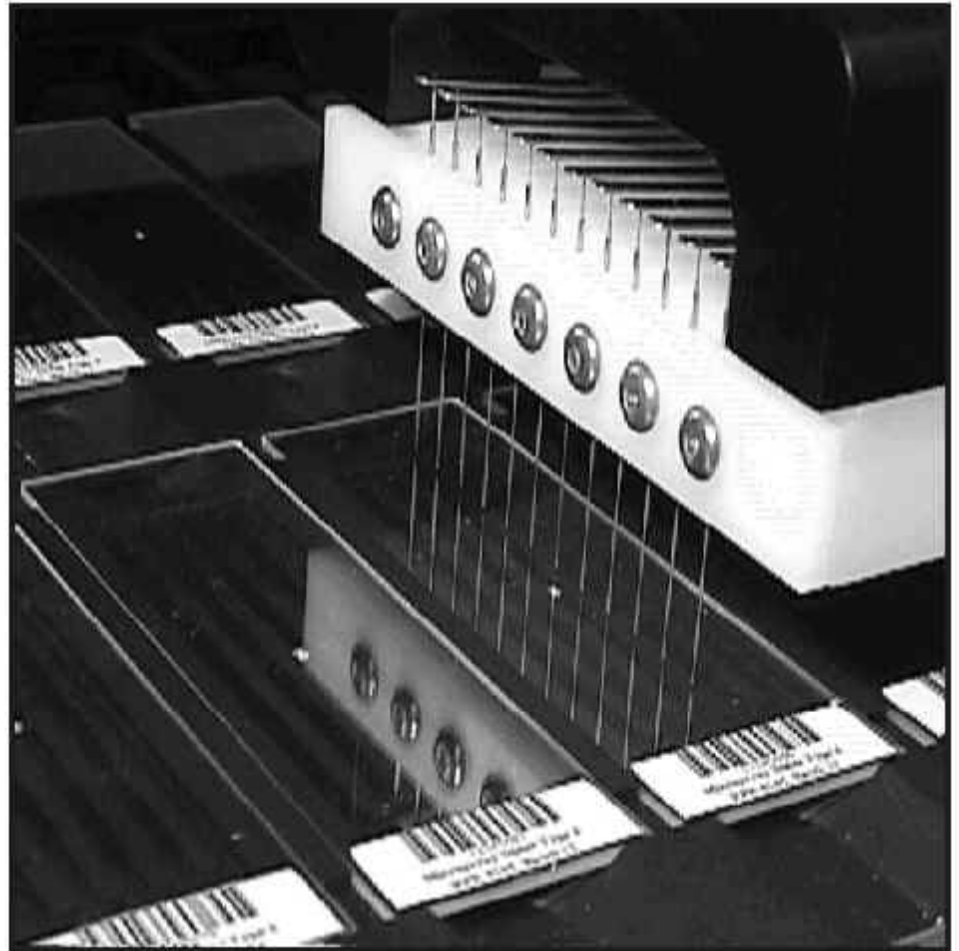
<http://www.accessexcellence.org/AB/GG/polymerase.html>

# Microarray manufacturing: Array spotter



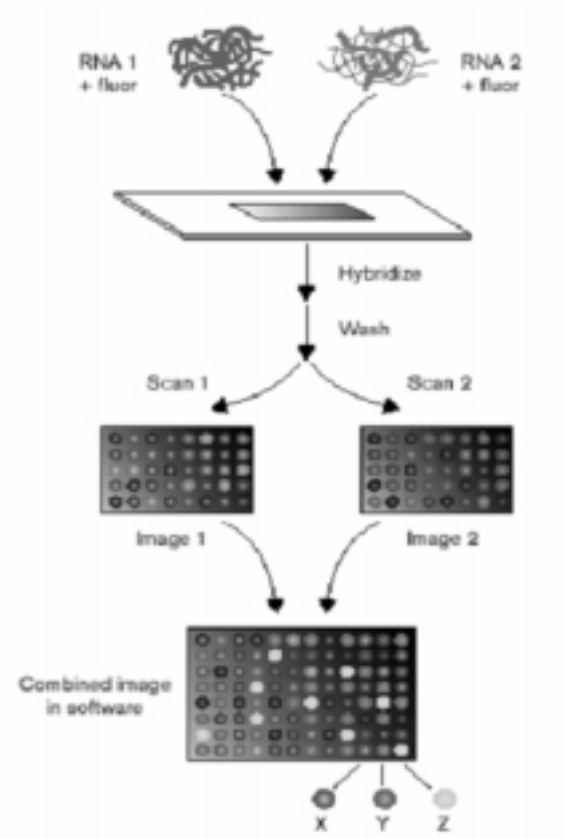
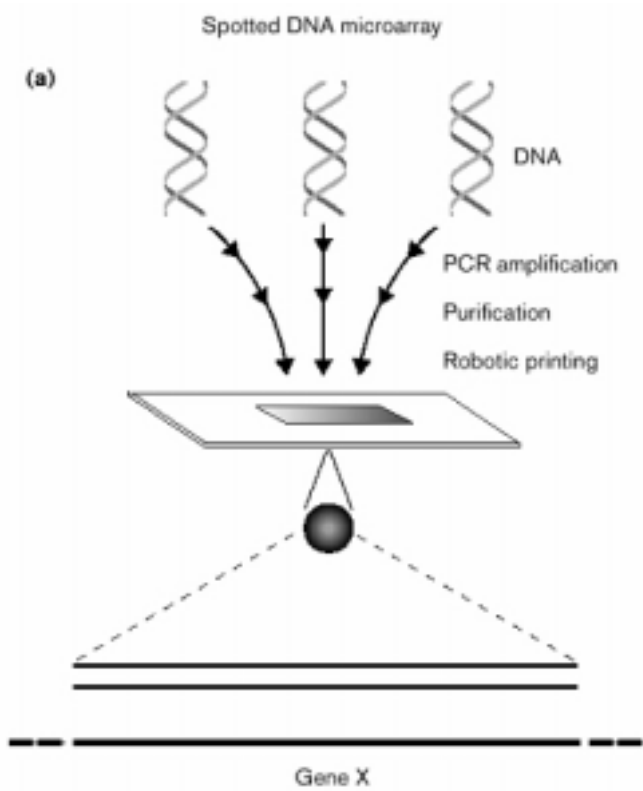
YM-Biochem

# Array spotter (zoom in)



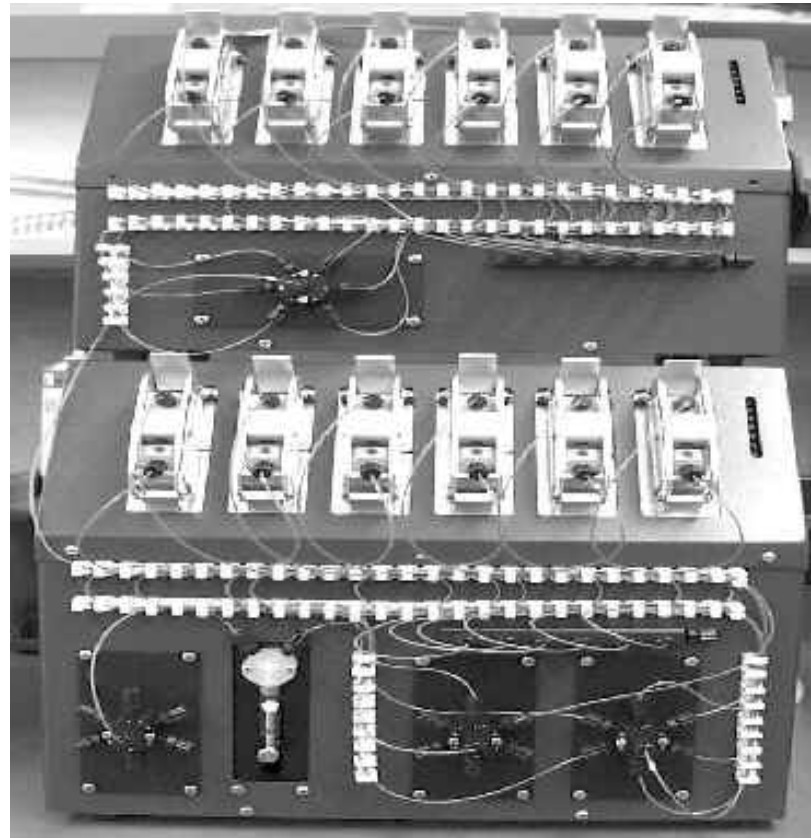
# Steps for a typical microarray experiment

- Experimental design
- Sample preparation
- RNA extraction
- Prepare cDNA
- Labeling cDNA with dye
- Hybridization
- Quantitation



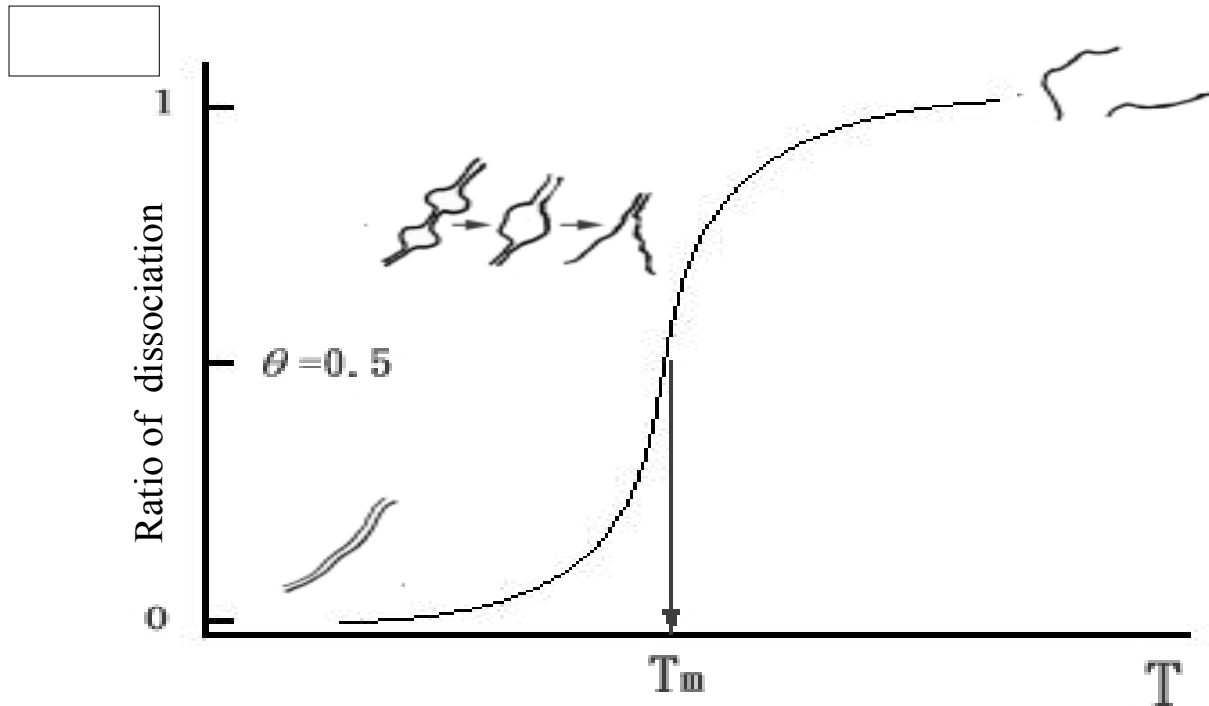
(Harrington *et al.* 2000)

# Automatic hybridization processor



YM-Biochem

# Denaturation and definition of $T_m$

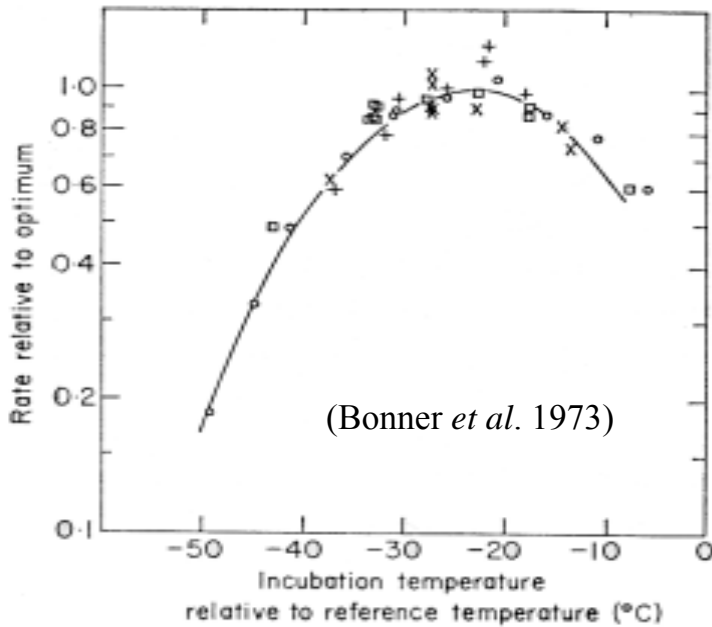


# Effect of other factors on Hybridization

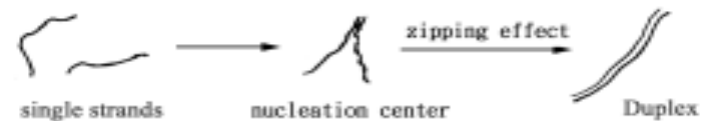
$$T_m = 81.5^\circ\text{C} + 16.6 \log M + 0.41(\%G + C) - \frac{500}{n} - 0.61(\% \text{ formamide})$$

(Bolton *et al.* 1962)

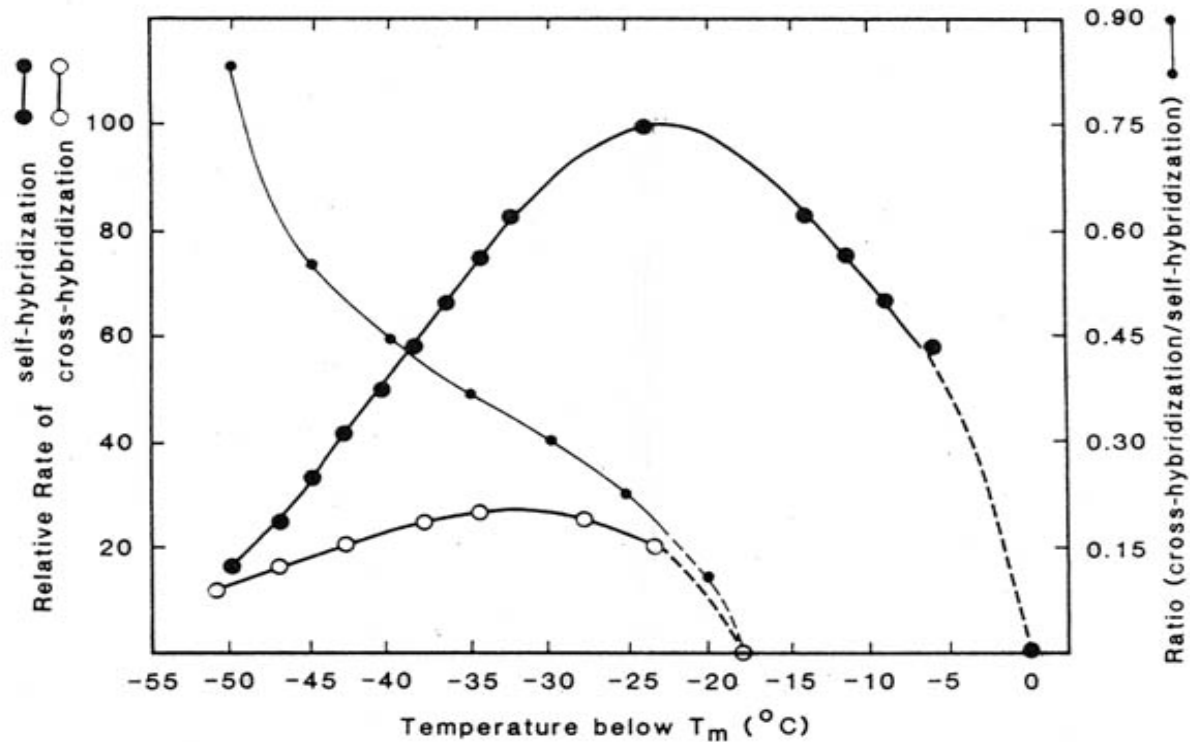
# Hybridization kinetics



1. High temperature favors collision (more chance to form nucleation center)
2. Low temperature stabilize the nucleation center



# Cross-hybridization is slower than perfect hybridization

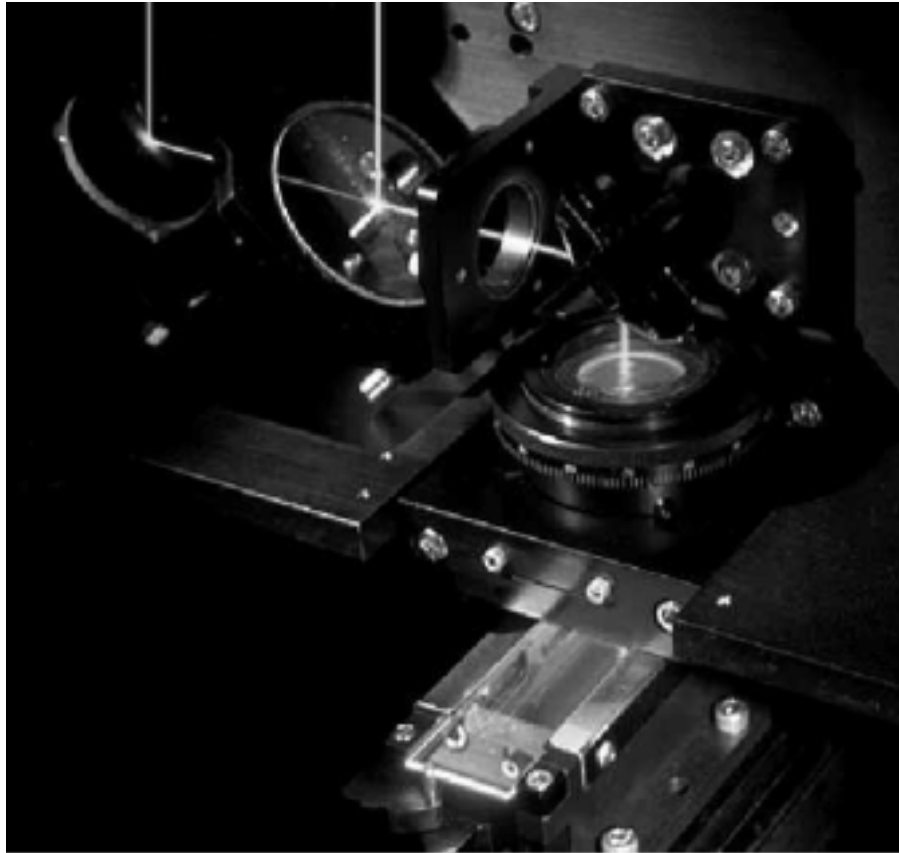


(Beltz *et al.* 1983)  
YM-Biochem

# Decision of Hybridizing Temperature

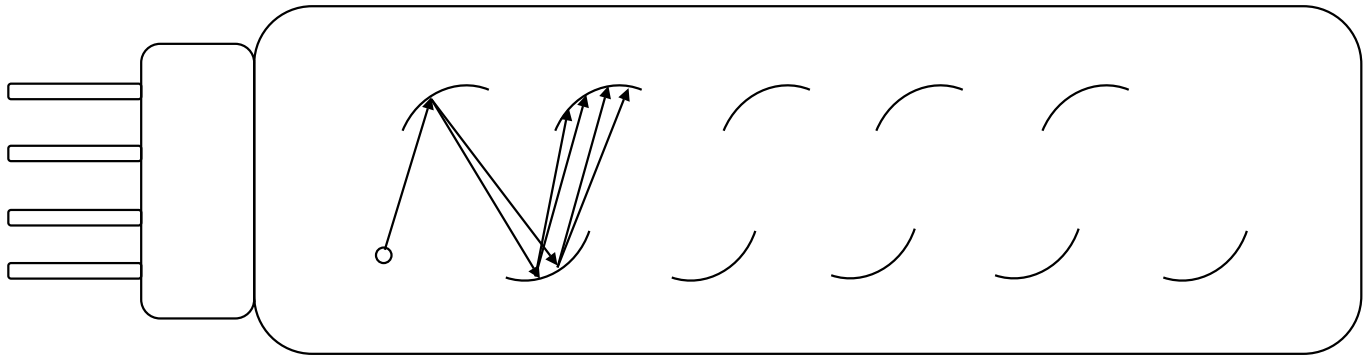
- Probes is variable.
  - Length : 100bp~thousands
  - GC content : 30~70%
- 3X SSC
  - $T_{m(100bp,30\%)}=84.9^{\circ}\text{C} \Rightarrow T_{hyb}=59.9^{\circ}\text{C}$
  - $T_{m(100bp,50\%)}=93.1^{\circ}\text{C} \Rightarrow T_{hyb}=68.1^{\circ}\text{C}$
- Microarray protocol
  - $T_{hyb}=60^{\circ}\text{C} \sim 68^{\circ}\text{C}$

# Laser scanner



YM-Biochem

# Photo-multiplier tube (PMT)

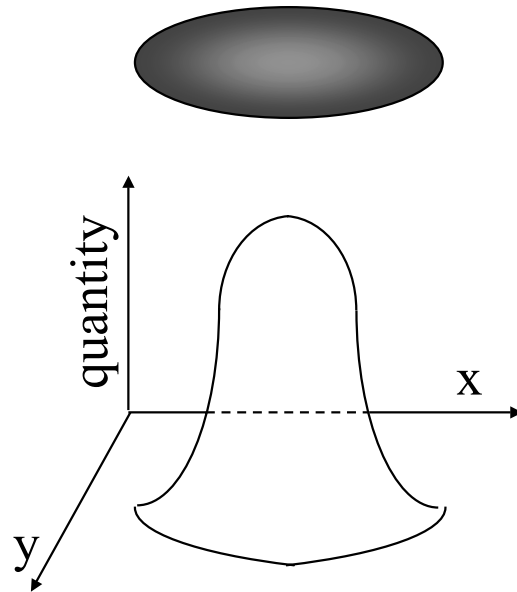


# Quantitation

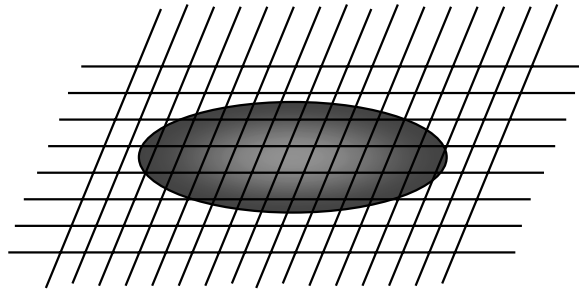
Volume integration

Pixel statistics

# Volume integration



# Pixel statistics



0.1, 0.12, 0.13, 0.15, ....., (5.3), ....., 9.86, 9.88, 9.80, 9.9, 13  
↑  
median

# Data Analysis - Reveal the Difference

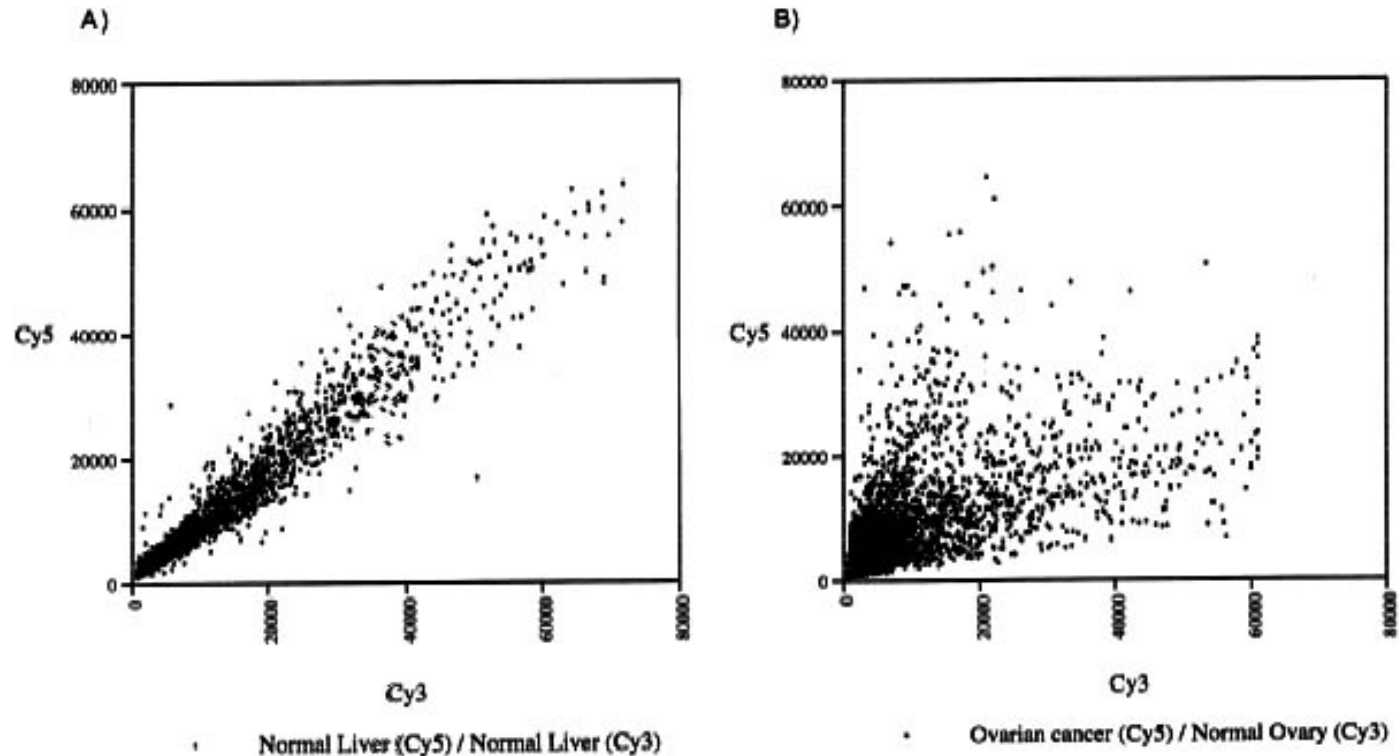
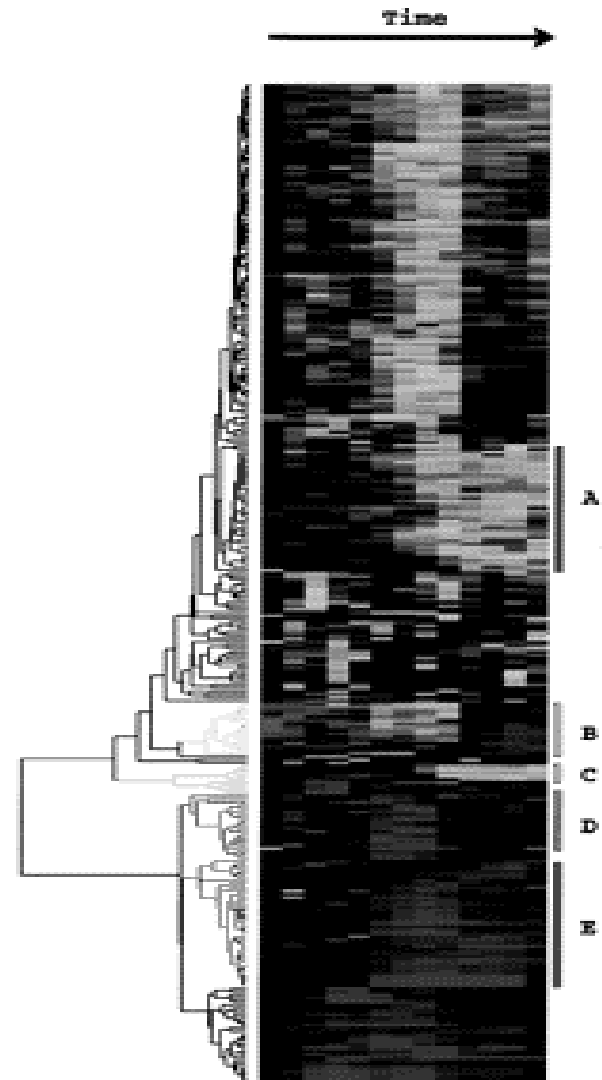


Fig. 1. Scatter plots for a Cy5-liver/Cy3-liver control hybridization (A) and a Cy5-ovarian tumor/Cy3-normal ovary hybridization (B). The value of Cy3 and Cy5 hybridization signals from each clone were plotted directly onto the plot.

# Medical application of $\mu$ -array data

- Fingerprinting
  - Distinguish states (disease *vs* normal; classification of species, strains, cell lines, drugs, --- *etc.*)
- Mechanism studies
  - Pathways and regulatory circuits  
=> Look for drug lead
  - Dynamic changes among developmental stages
  - --- *etc.*

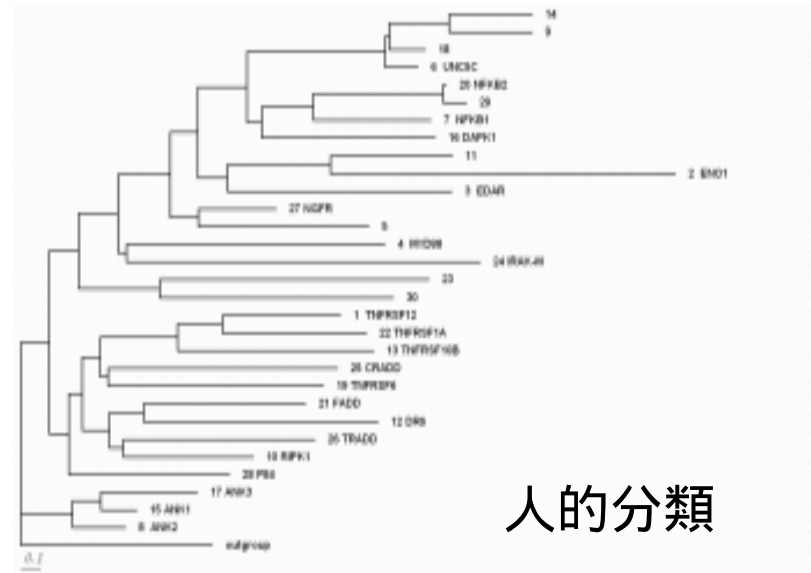
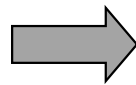
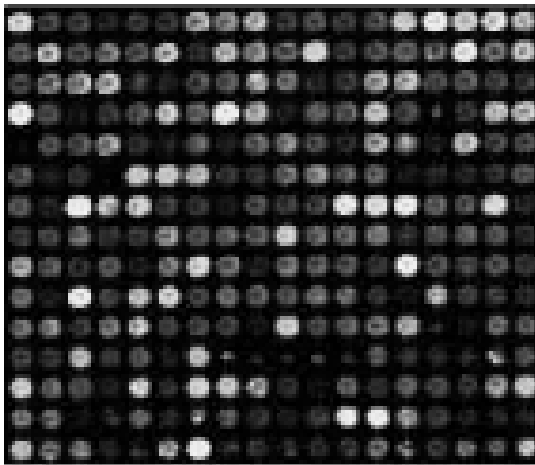
# Discrimination, classification, and clustering



*PNAS, Vol. 95, 14863-14868*

# 疾病檢驗與個人化之疾病治療

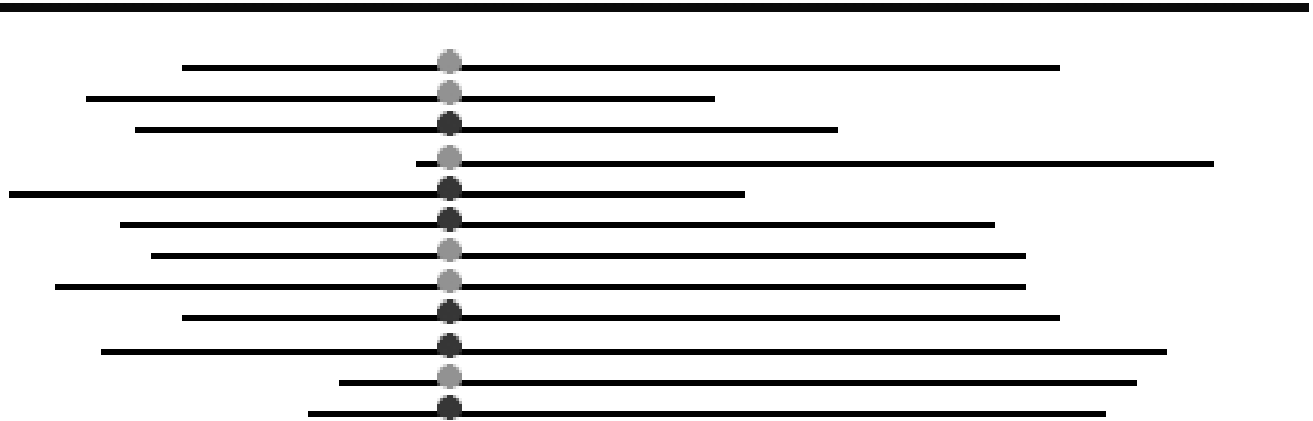
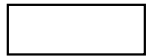
基因晶片



人的分類

# 由序列比對尋找SNP

基因體DNA



修改自張猷忠博士之圖片

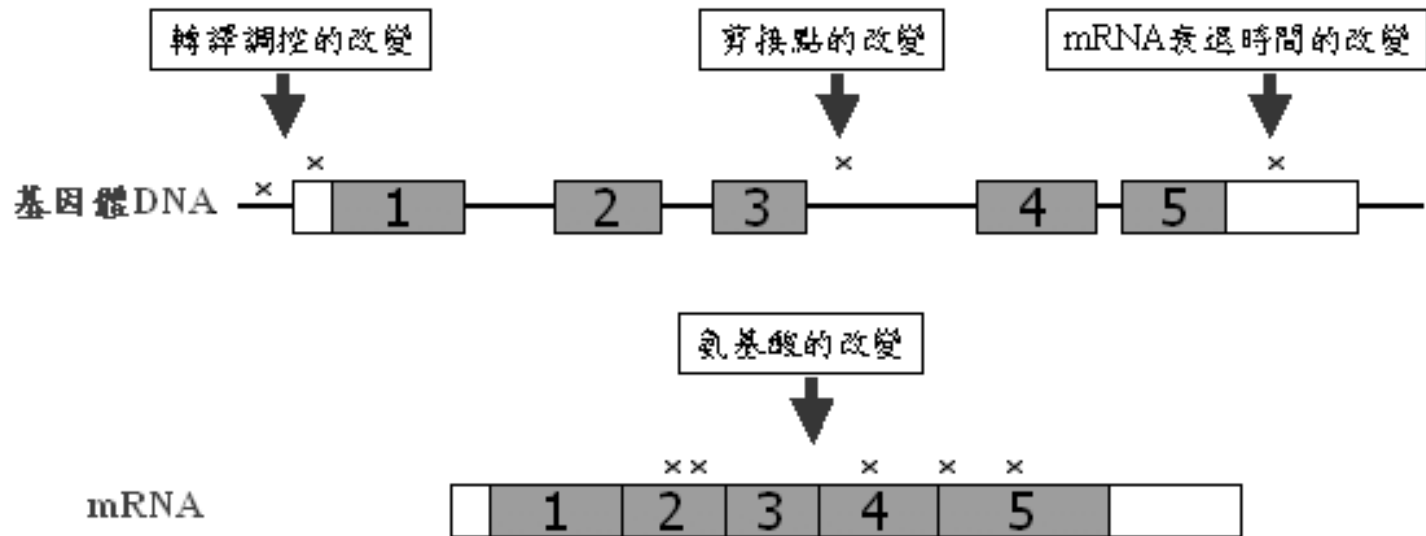
# 由序列比對 尋找SNP

The image displays a bioinformatics software interface for sequence alignment and SNP detection. The main window, titled "AF006386.fasta.screen.2.cluster\_1 ace", shows a sequence alignment view. The top menu includes "Tools", "View", and "Reports". Below the menu, there are buttons for "Contigl", "SNPs: Best", "Next", "Prev", "1/2: 1.0", "Overview", "Find", "Libraries", and "Exit". The main display area shows a consensus sequence and several reads. The consensus sequence is: `TTATTGGCTGCCAGAGGAACACAGTAAATAACTC*CCAAGTGTCTTGTGGAAATTAATCATGTGGAAATTATCA`. The reads are: `qg67h09.xl`, `qr11h10.xl`, `qs75h10.xl`, `qx69a06.xl`, `qy47f04.xl`, `qy91e03.xl`, `yp64h08.sl`, `yu77h08.sl`, and `za67a05.sl`. A secondary window, titled "Assembly overview", shows a graph of SNP quality. The graph has a y-axis labeled "SNP Quality:" and a peak at position 1. The x-axis is labeled "992". The graph shows a peak at position 1, indicating a high quality SNP. The main window also shows a warning message: "警告: Applet 視窗".

The secondary window, titled "AF006386.fasta.screen.2.cluster\_1 ace", shows a sequence alignment view. The top menu includes "Tools", "View", and "Reports". Below the menu, there are buttons for "Contigl", "SNPs: Best", "Next", "Prev", "2/2: 0.99", "Overview", "Find", "Libraries", and "Exit". The main display area shows a consensus sequence and several reads. The consensus sequence is: `ACTOCCCTGGGGGTAGTATGGCATTCAAGATTTCTTCTGCTGCTTGTAGGATCTGGGACACAGGGAGTTGAGG`. The reads are: `AA379733`, `AA383086`, `AF006386`, `yu77h08.rl`, `za77d12.rl`, `zF48g07.rl`, `zk31g05.rl`, `zw81e06.rl`, and `zw82a06.rl`. A secondary window, titled "Assembly overview", shows a graph of SNP quality. The graph has a y-axis labeled "SNP Quality:" and a peak at position 1. The x-axis is labeled "992". The graph shows a peak at position 1, indicating a high quality SNP. The main window also shows a warning message: "警告: Applet 視窗".

本圖片由張猷忠  
博士提供

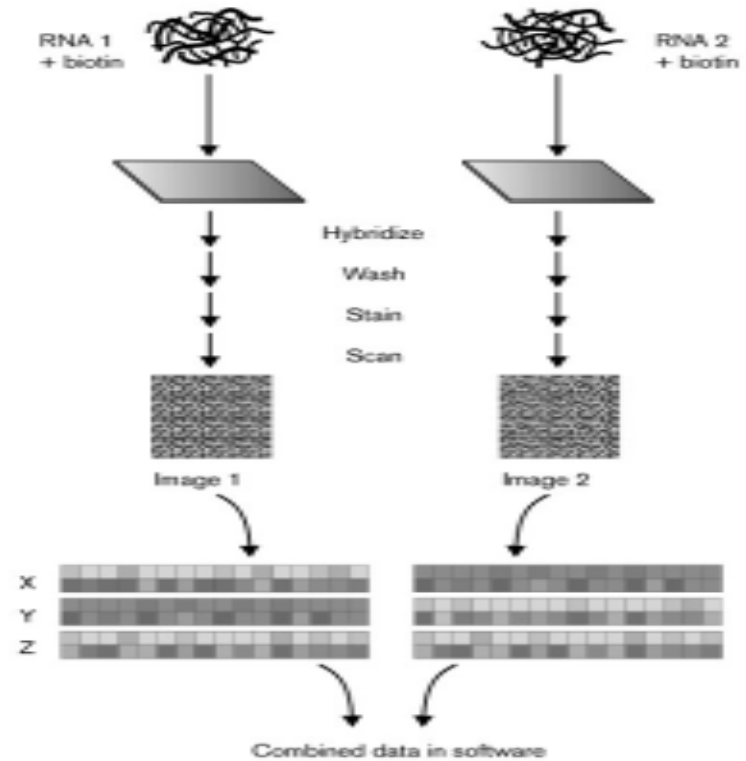
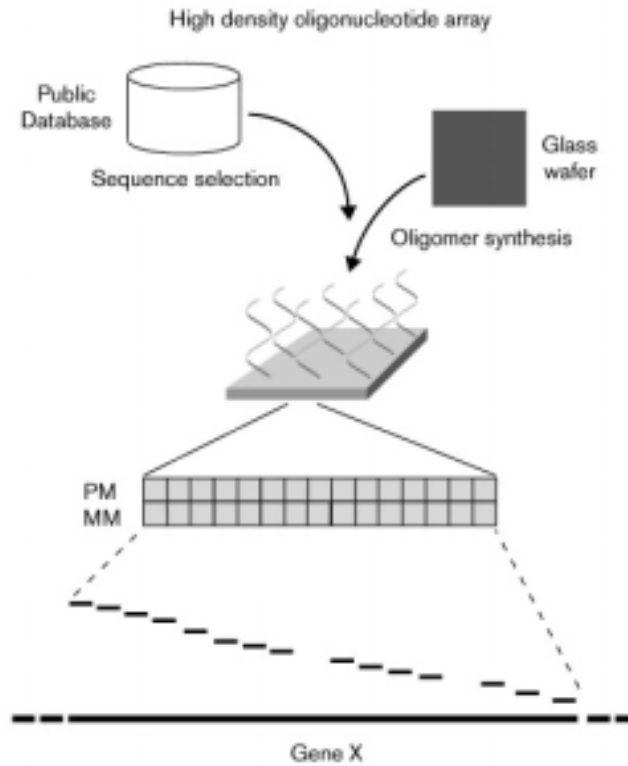
# 發生在基因不同部位的SNP造成的影響



X 代表單核苷酸多型性發生點

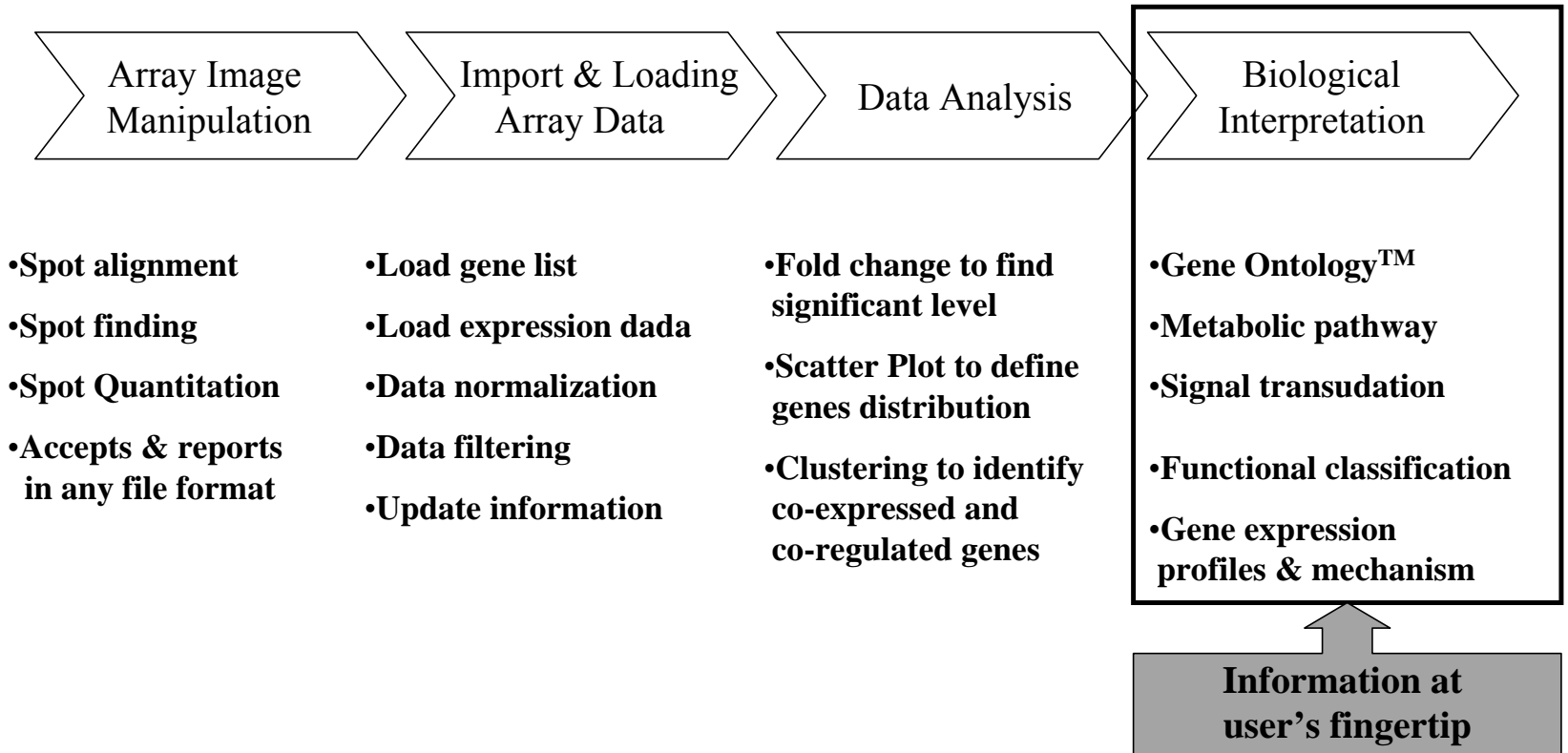
本圖片由張猷忠博士提供

# Oligonucleotide array



Current Opinion in Microbiology

# Data => Information => Knowledge => Technology



# Microarray (II)

Ueng-Cheng Yang  
Bioinformatics Research Center  
National Yang-Ming University

# Outline

Problems assoc. with microarray

Data interpretation

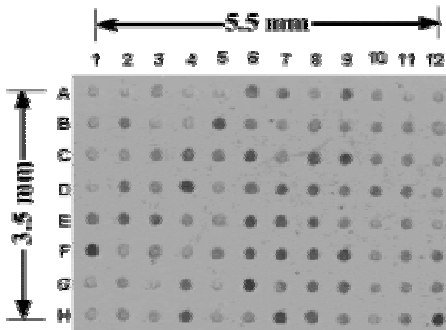
Gene Chip Analysis Tools (GCAT)

# Problems assoc. with microarray

# Errors that can be improved by repeating the experiment

- Experimental condition, such as
  - incubation temperature of your cell line
  - hybridization temperature of each exp.
- Experimental material: such as the healthiness of your cell line.
- Quality of the gene chips
- Data acquisition process

# The problems of cDNA array



1. Competition among targets
2. Alternative splicing
3. Do you have all the genes?  
(*e.g.* transiently expressed genes)

Picture taken from

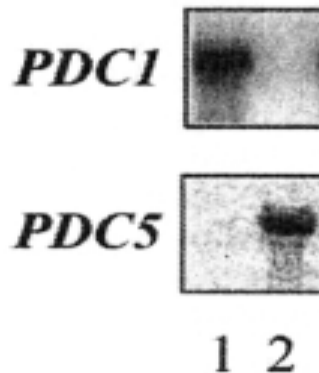
[http://www.ibms.sinica.edu.tw/~peck/chinese/marray\\_4.htm](http://www.ibms.sinica.edu.tw/~peck/chinese/marray_4.htm)

# Effect of Cross-Hybridization (cont.)

Is that caused by regulatory factor?

PDC5 is only expressed in *pdc1* deletion mutants. No mRNA transcribed from PDC5 could be detected in wild-type cells.

~Hohmann et al., 1990, *Eur. J. Biochem.*, 188, 615-621



Northern blot analysis

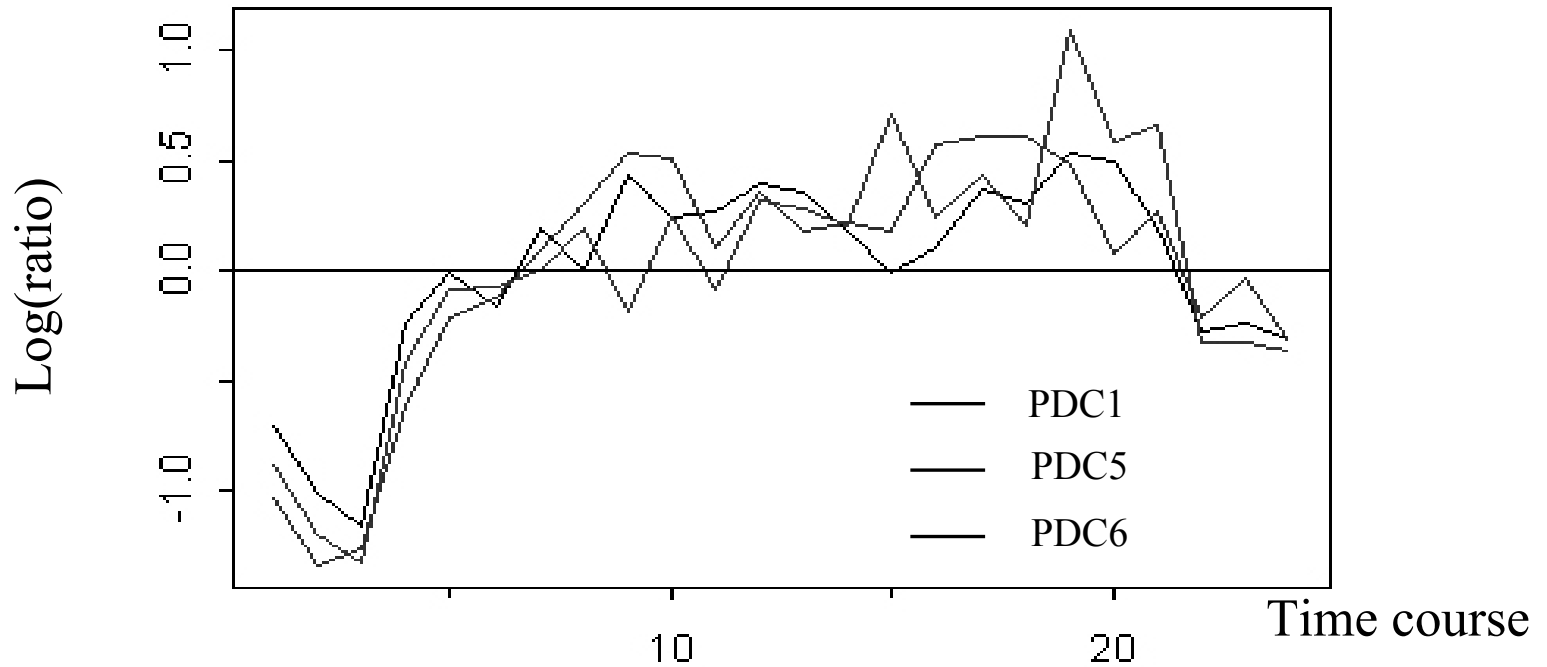
(1) wild-type

(2) a *pdc1*Δ mutant

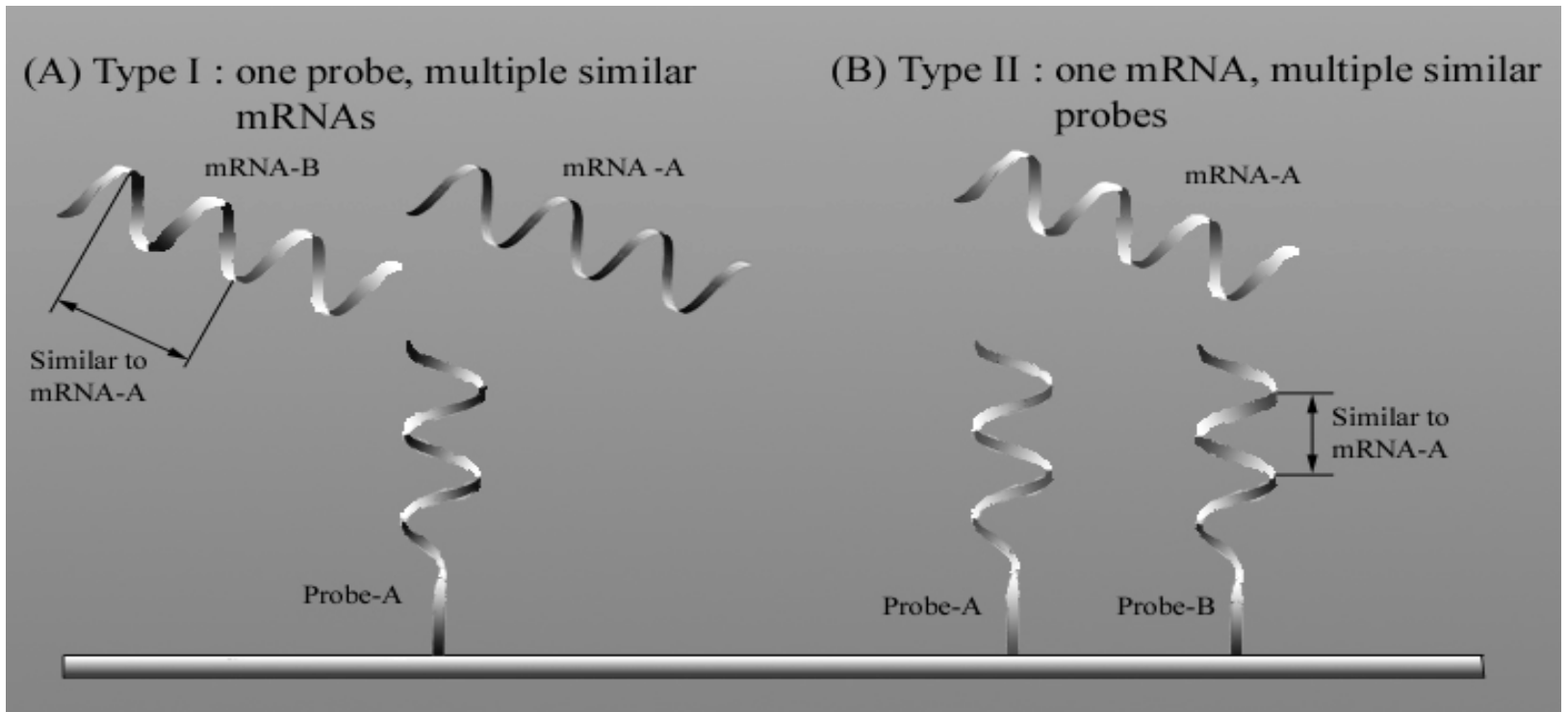
~Eberhardt et al., 1999, *Eur. J. Biochem.*, 262, 191-201

# Effect of Cross-Hybridization (cont.)

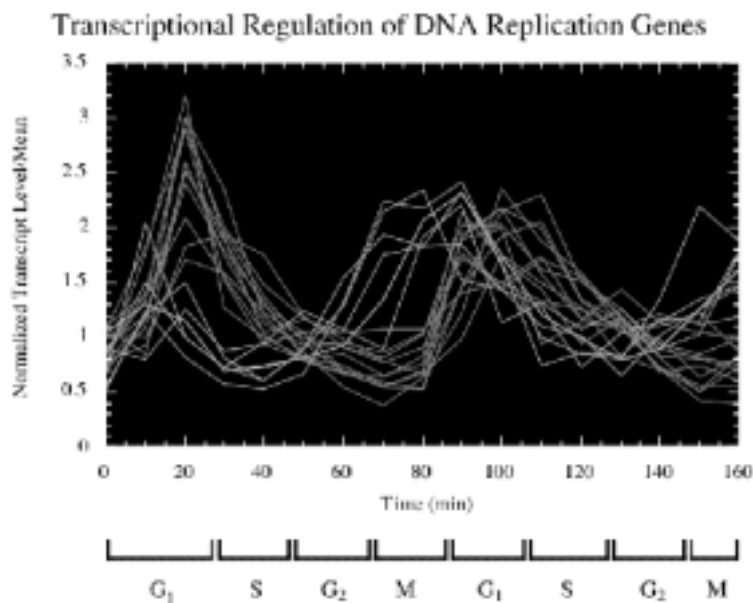
A similar expression profile in PDC6



# Two Types of Cross-Hybridization on cDNA Microarray



# Does clustered time course really mean co-expression?



orange=pre-replication complex genes: mcm2,  
mcm3, cdc46, cdc47, cdc54, cdc6  
blue=replication genes involved in DNA synthesis

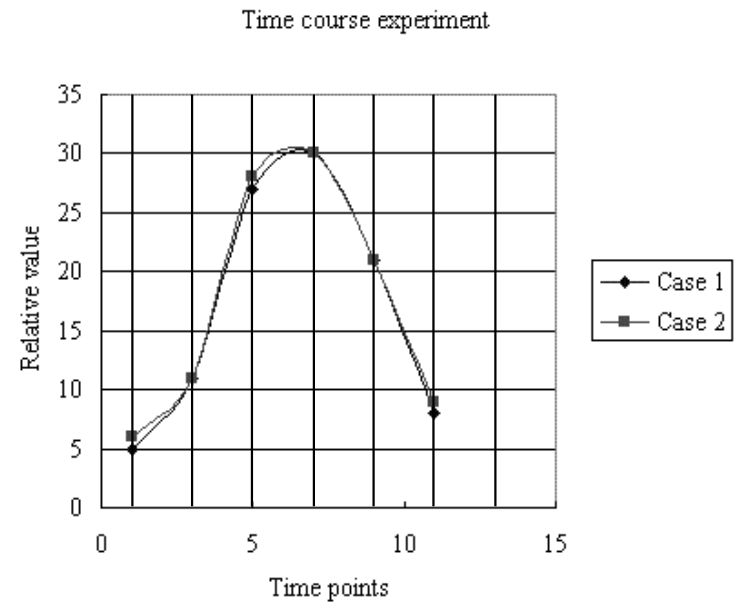
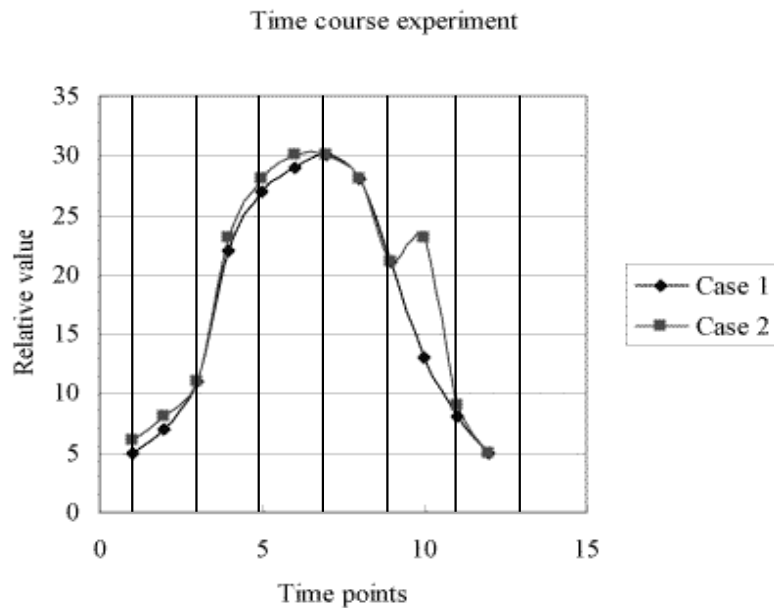
Yes, you can study known system (such as cell cycle) this way; but, how about the unknown systems?

Picture taken from

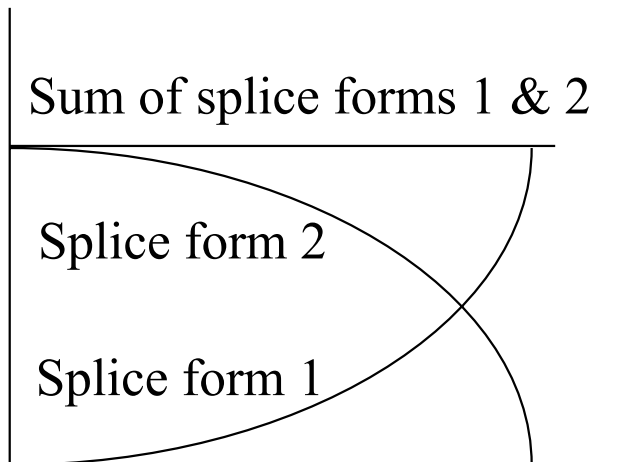
[http://genomics.stanford.edu/yeast/additional\\_figures\\_link.html](http://genomics.stanford.edu/yeast/additional_figures_link.html)

# Meaning of a time course experiment

Misinterpretation of time course data when there are not sufficient points



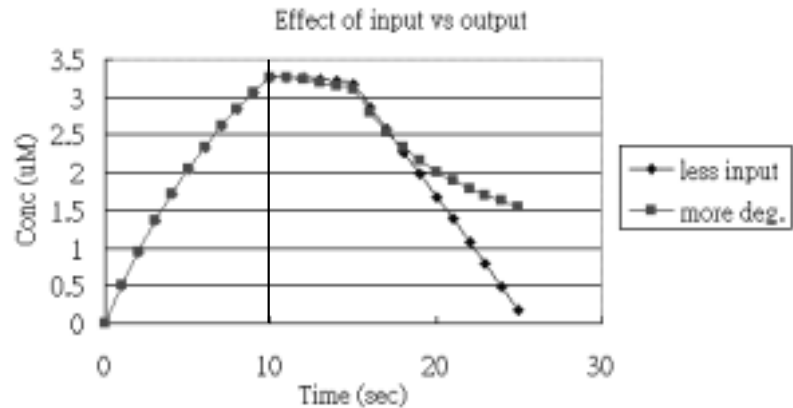
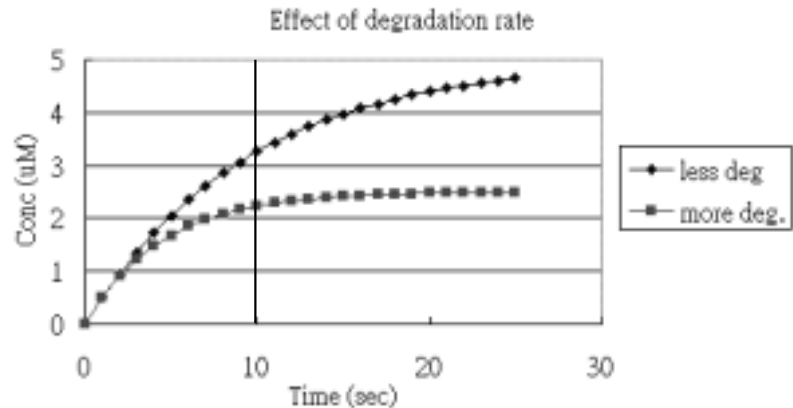
# cDNA array cannot distinguish alternatively spliced forms



Failure to distinguish  
different splicing  
forms

# Meaning of the relative intensity

Misinterpretation of time course data when different transcripts have different degradation rates



If the degradation rate of different genes are not identical, what's the meaning of clustering?

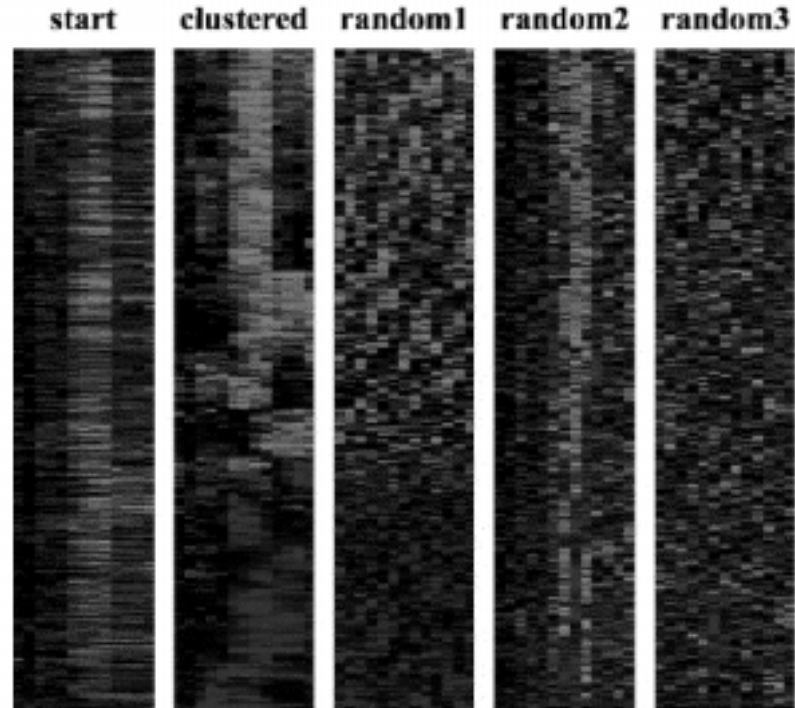
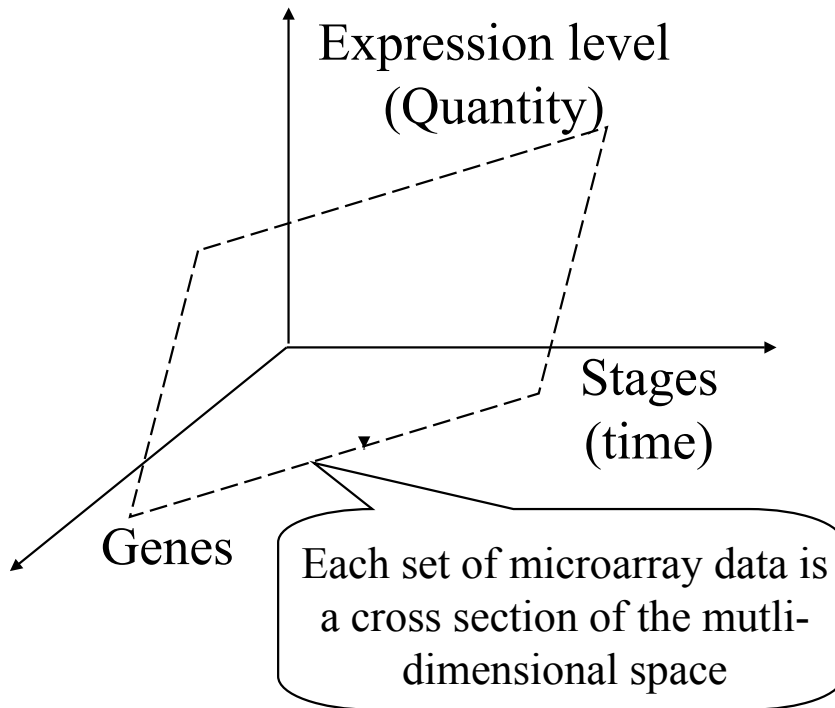


FIG. 3. To demonstrate the biological origins of patterns seen in Figs. 1 and 2, data from Fig. 1 were clustered by using methods described here before and after random permutation within rows (random 1), within columns (random 2), and both (random 3).

# Data interpretation

# Life phenomena can be represented in a multi-dimensional space



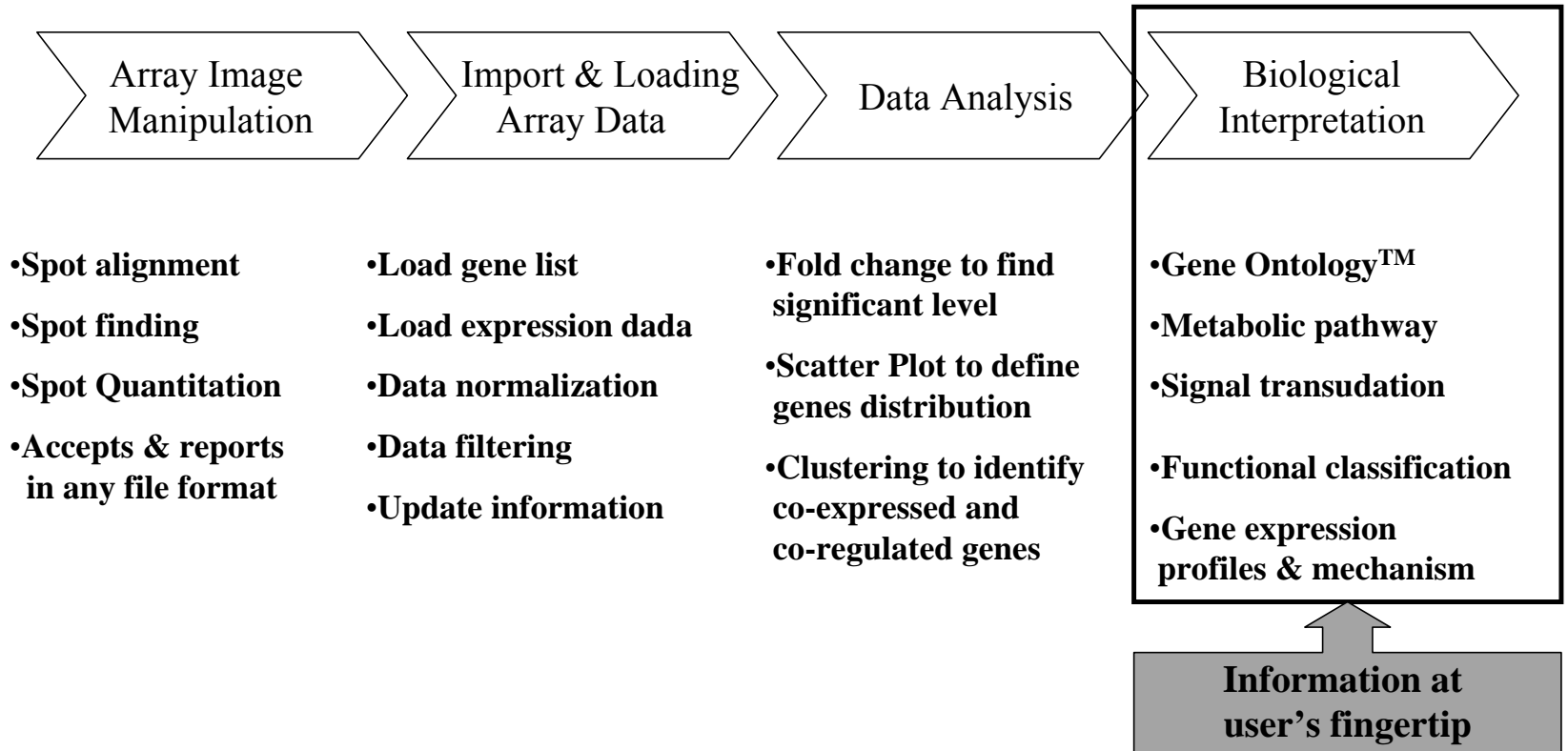
Stages: developmental, transformation, time after treatment, *etc.*

Tissue distribution  
(position, 3D)

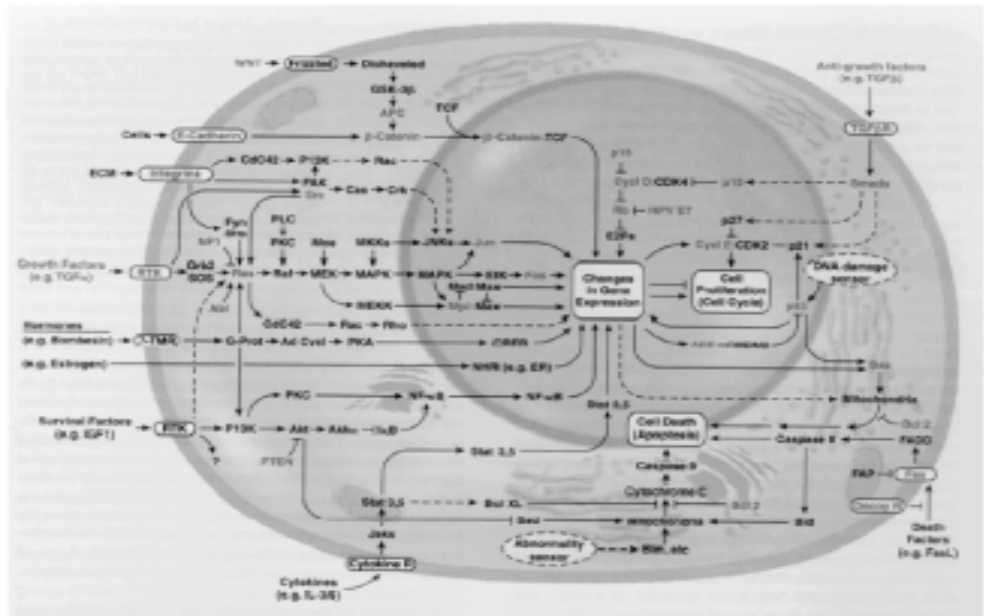
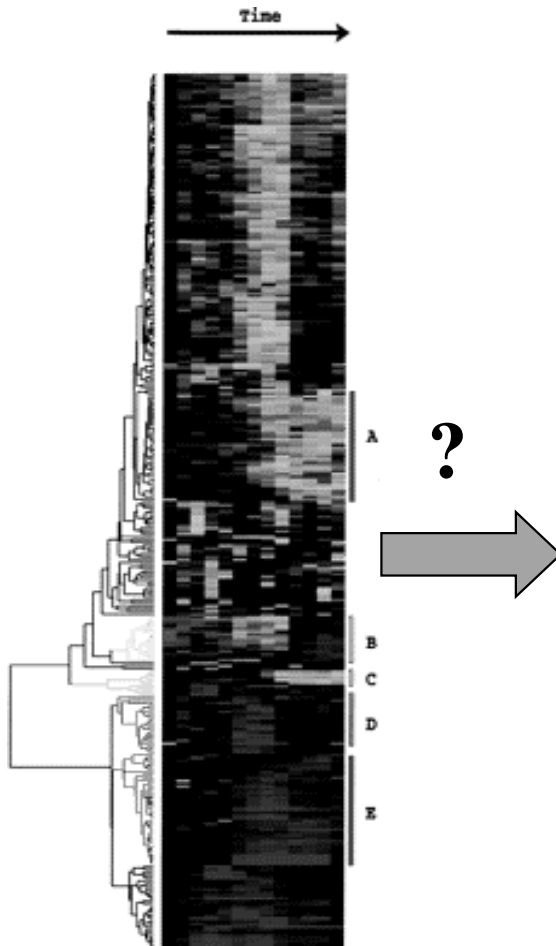
Treatments (stimuli, drugs, nutrients, *etc.*)

Physiological states  
(stressed, fasting, *etc.*)

# Data => Information => Knowledge => Technology

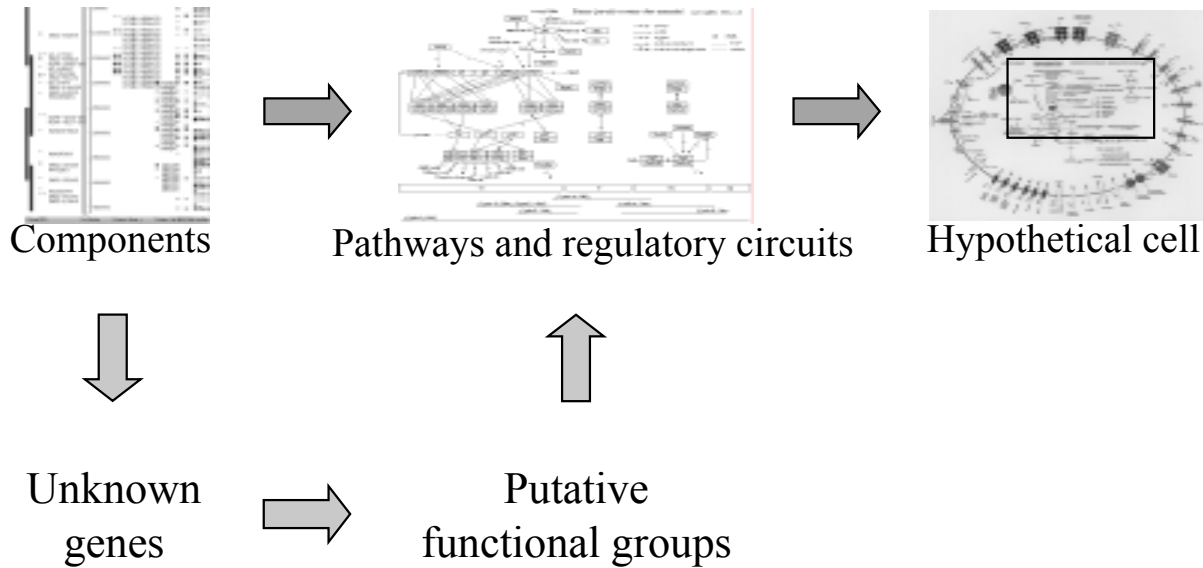


# Goals for system biology



*PNAS, Vol. 95, 14863-14868*

# Why do you want to do clustering?



# Mechanism Studies

Factors involved  $\Rightarrow$  Components

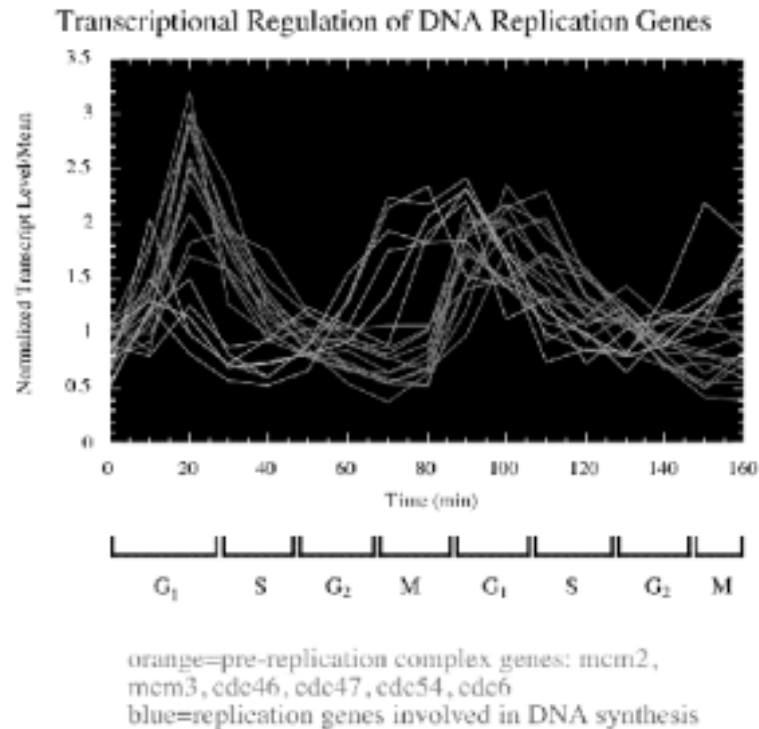
Order of events  $\Rightarrow$  Pathways

Interactions  $\Rightarrow$  Circuit

# Bioinformatics for microarray/ gene chip analysis?

- What should be put on the microarray?
- How many time points do you need?
- Data analysis (assume the quantitation is correct)

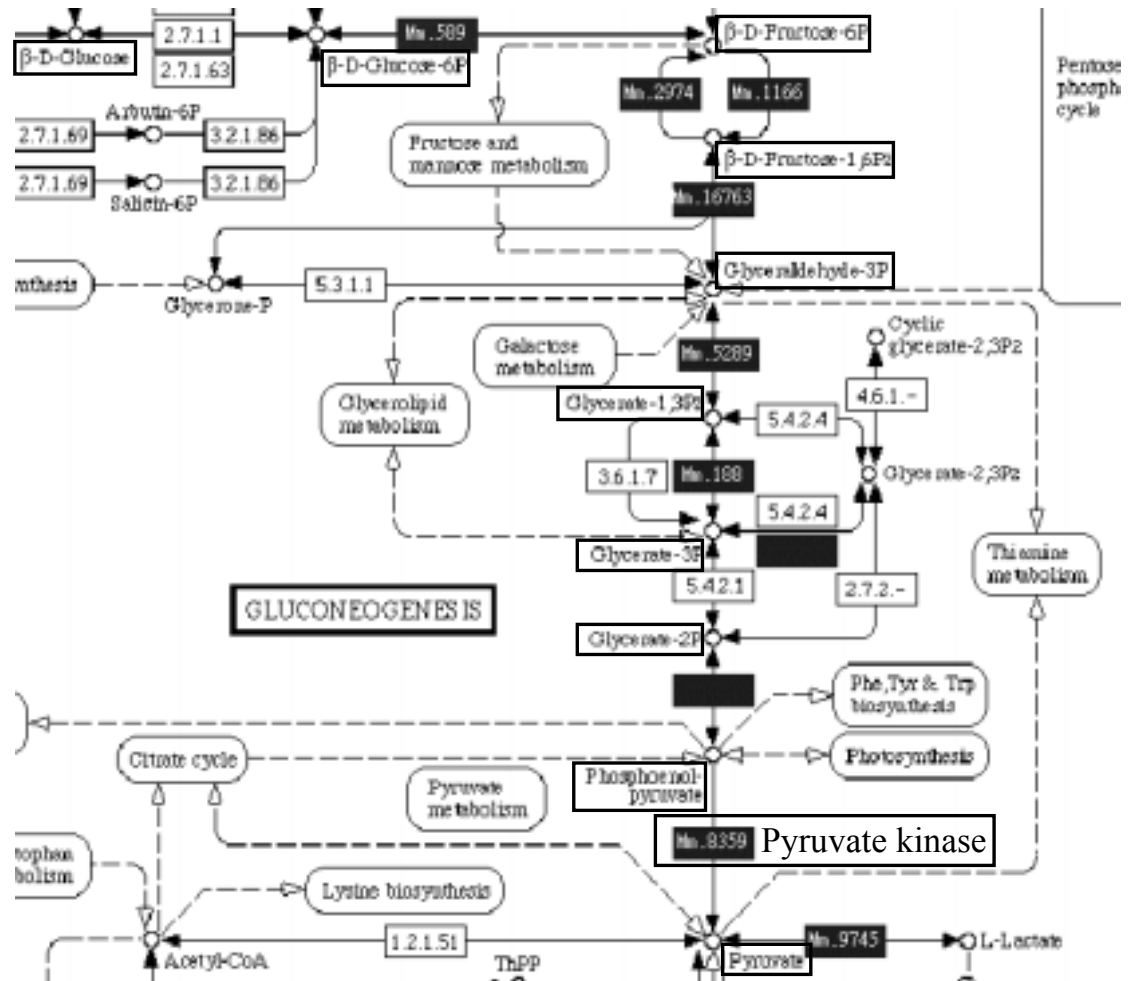
# Inducible gene sets are co-regulated.



Picture taken from  
[http://genomics.stanford.edu/yeast/additional\\_figures\\_link.html](http://genomics.stanford.edu/yeast/additional_figures_link.html)

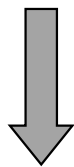
**Most constitutively expressed genes are not regulated**

Rate-limiting step is usually the target for regulation



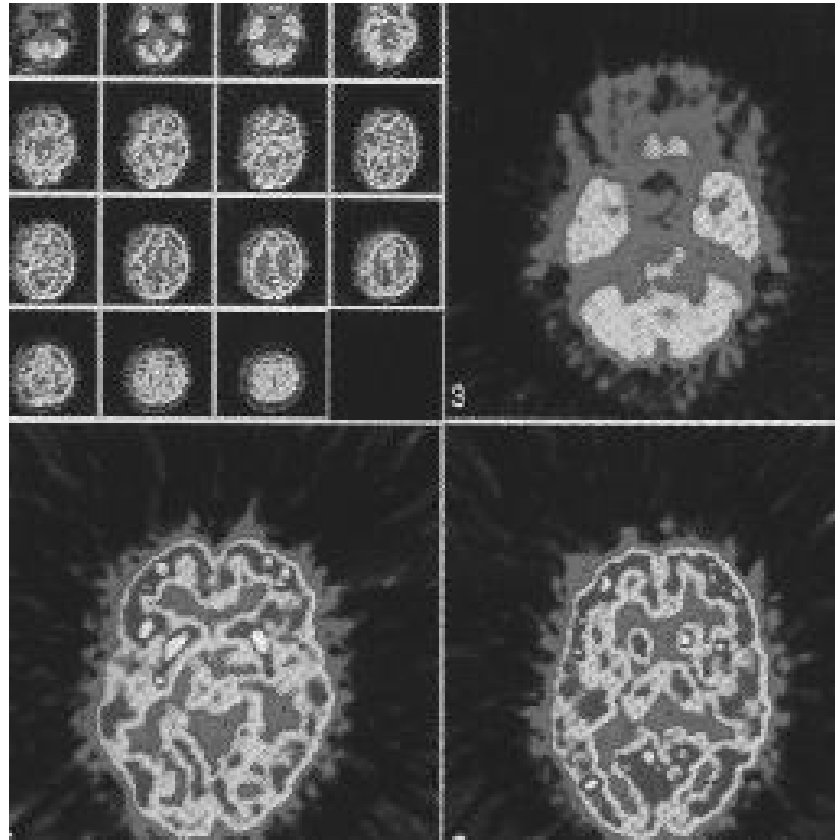
Microarray exp. is  
the nature's way to  
classify genes

Collect sections from  
different angles



Tomography  
(斷層掃描)

Image reconstruction



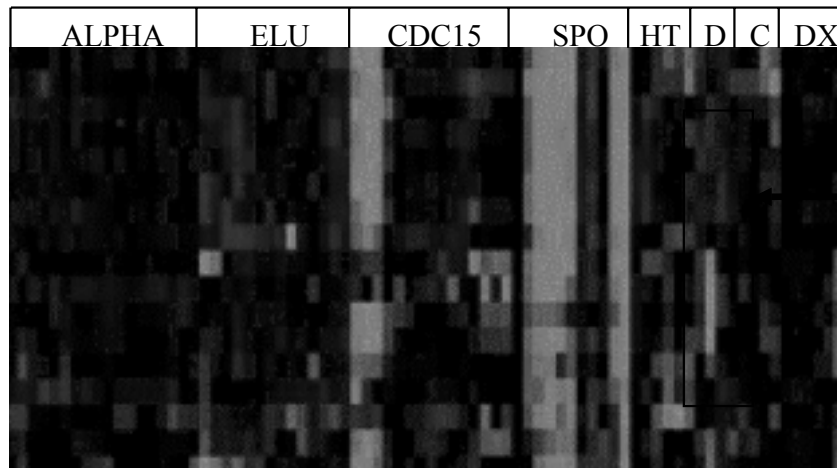
<http://www.npcc.gov.tw/npcc/chn/imaging/imaging.htm>

# How to classify constitutively expressed genes?

- Multiple data sets should be used.
- When multiple conditions are compared together, genes that are only regulated under a given condition will not be clustered. Because they will have different values under different environment.
- When the whole gene sets are activated or repressed under an extreme condition, it will be the driving force for clustering.

# In extreme environment, the whole pathway can be affected

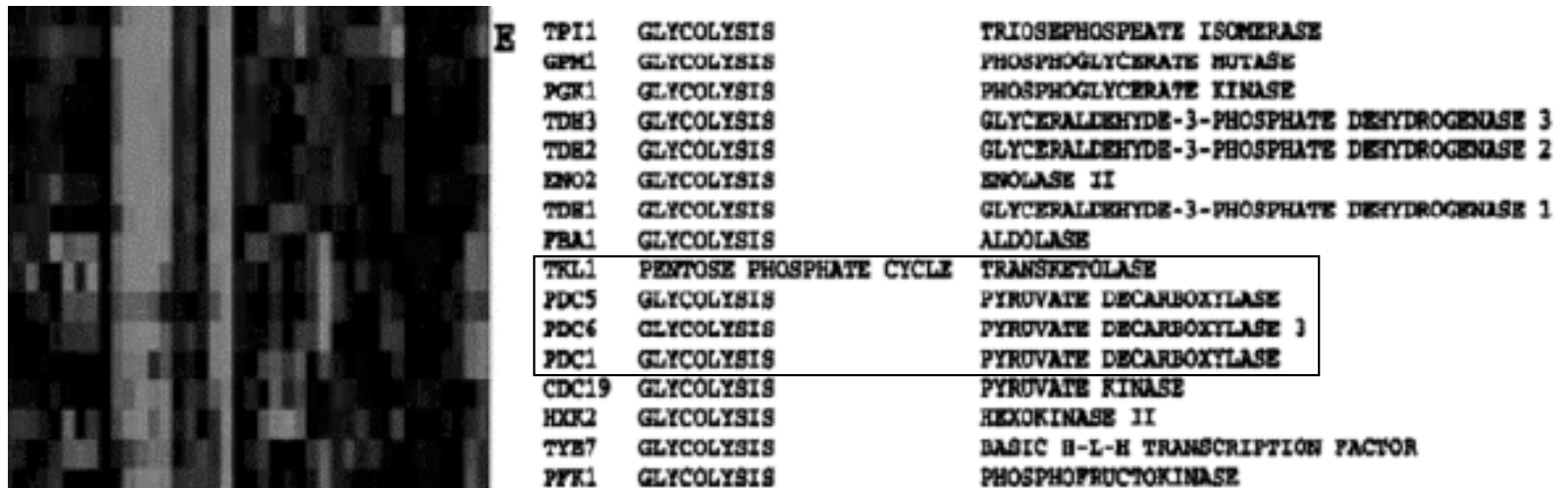
ALPHA = alpha factor arrest 18; ELU = centrifugal elutriation 14;  
CDC15 = cdc15 ts 15; SPO = sporulation 7; HT = shock by high temp 6;  
D = reducing agent 4; C = low temp 4; DX = diauxic shift 7



Clustering is driven by these features

YM-Biochem

# Unrelated sequences of similar function cluster together



Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression pattern. Proc. Natl. Acad. Sci. USA 95, 14863-14868.

# How good is the classification?

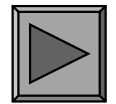
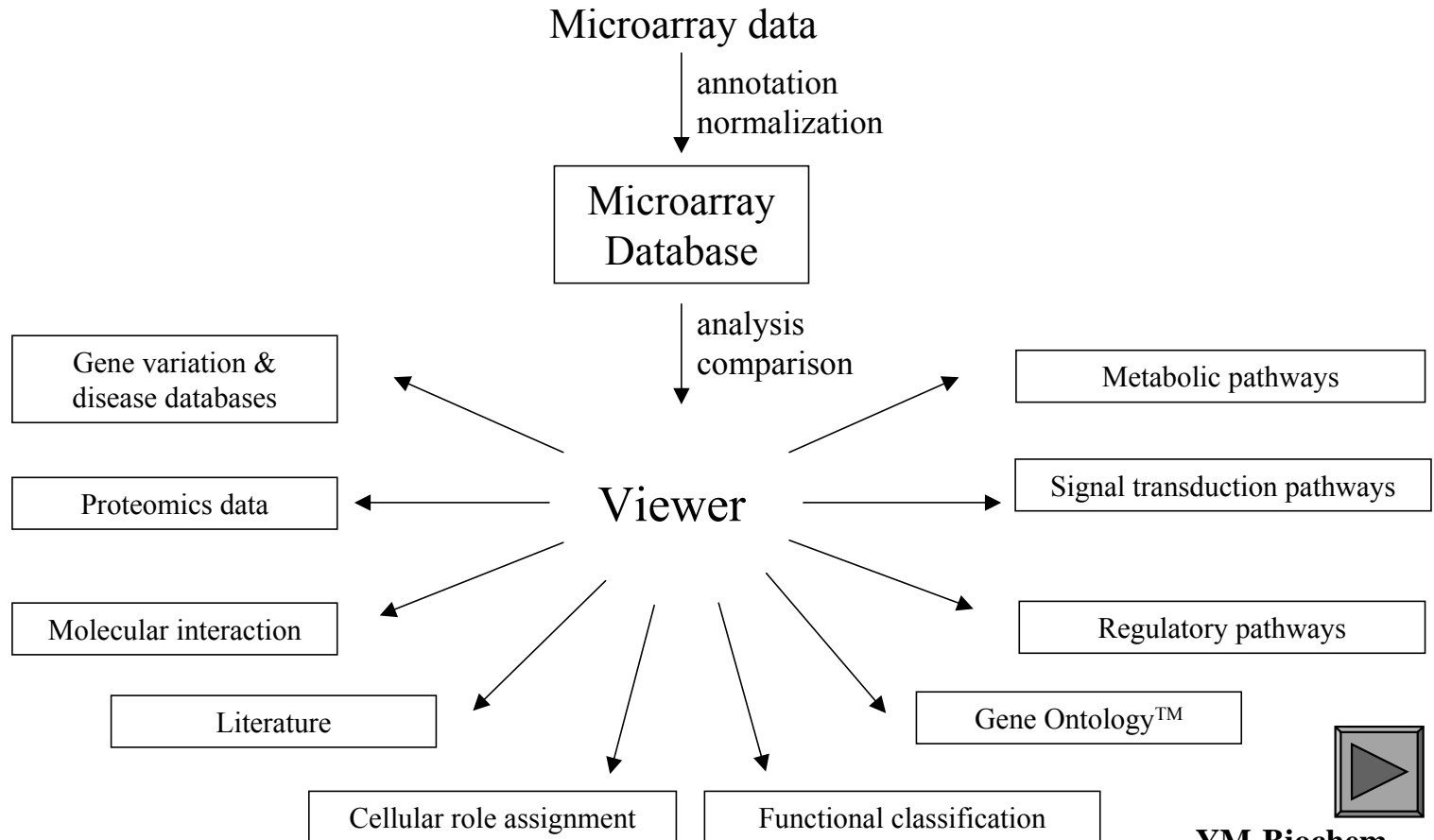
- In microarray clustering
  - hexokinase II
  - phosphofructokinase
  - aldolase
  - triose phosphate isomerase
  - GAPDH 1, 2, 3
  - phosphoglycerate kinase
  - phosphoglycerate mutase
  - Enolase II
  - pyruvate kinase
- In glycolysis, in total there are 10 enzymes involved
- Microarray experiment only missed phosphoglucoisomerase
- Pyruvate (de)carboxylase and transaldolase are misplaced

Pretty good

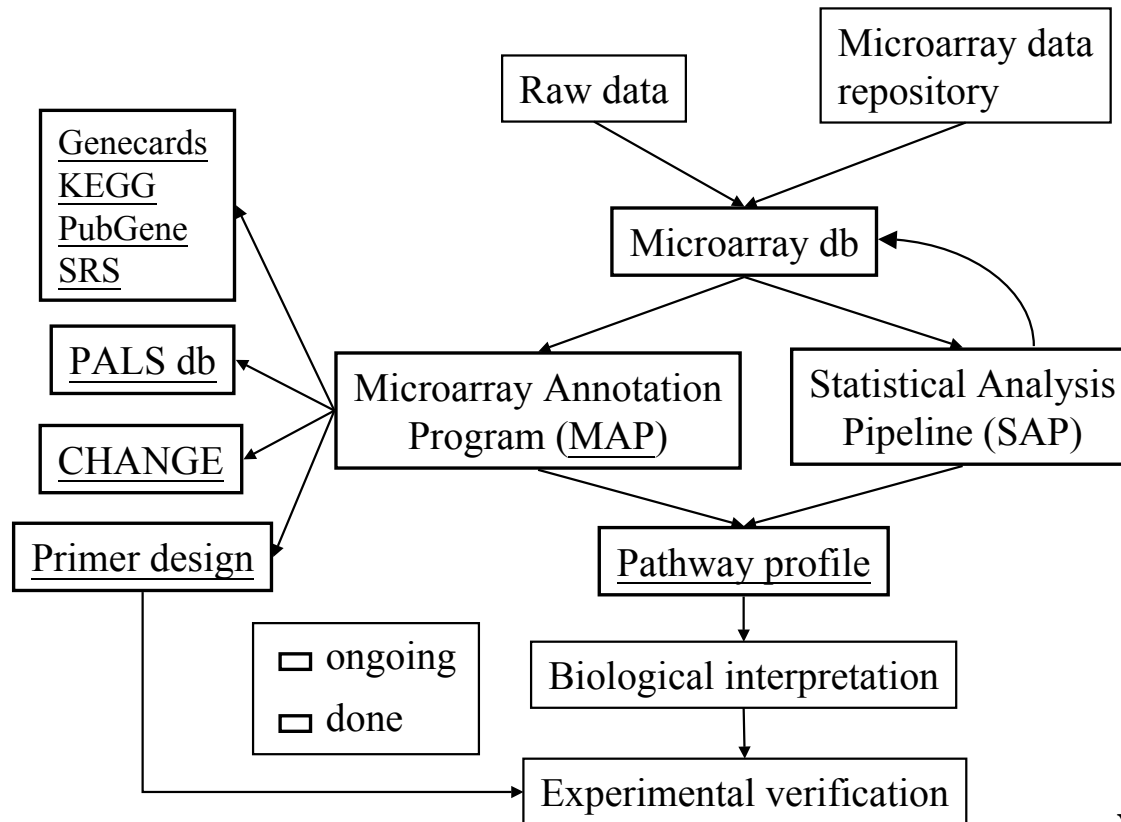


# Gene Chip Analysis Tools (GCAT)

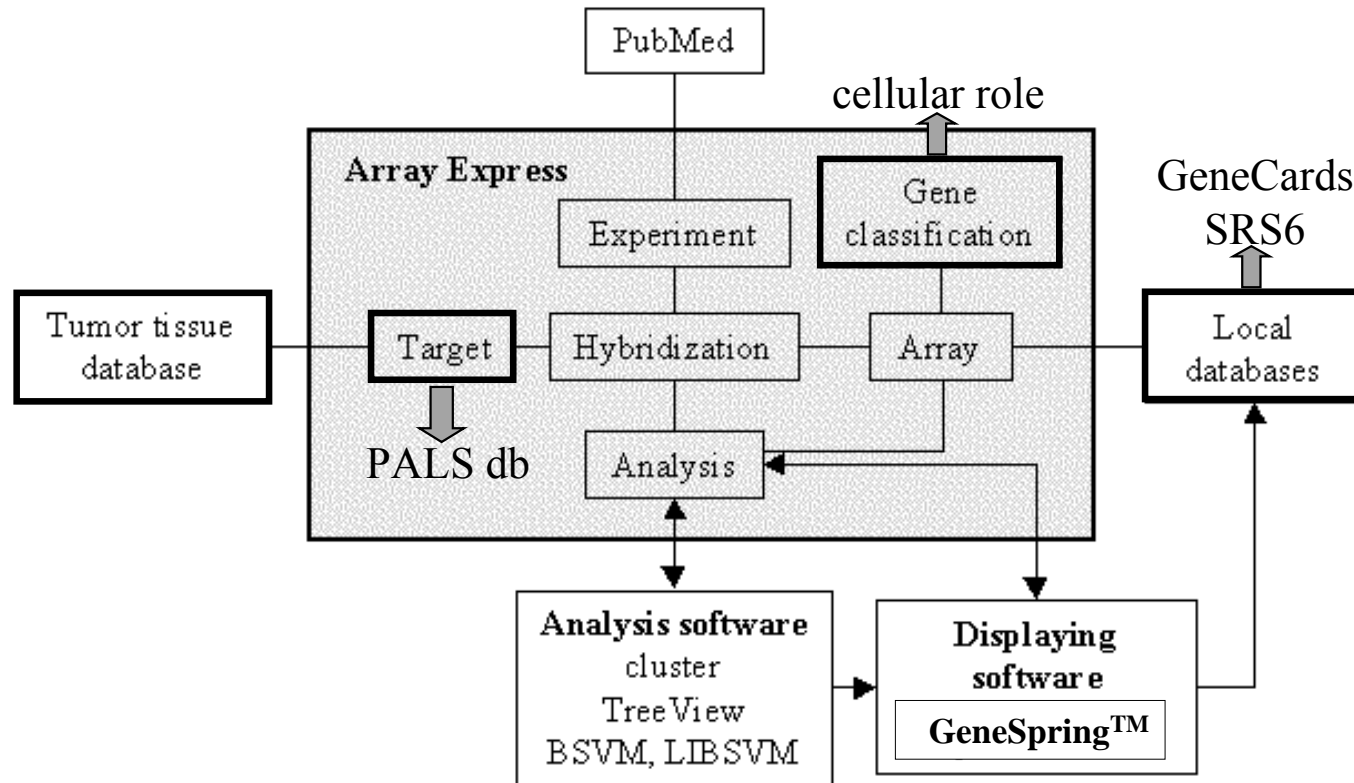
# Goal: Establish a user-centric bioinformatics environment for microarray data analysis

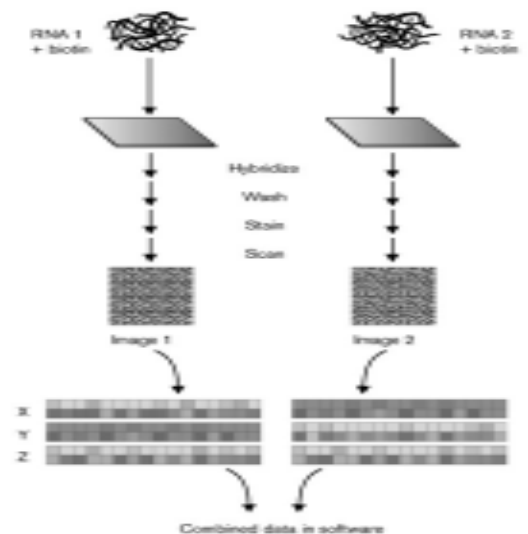
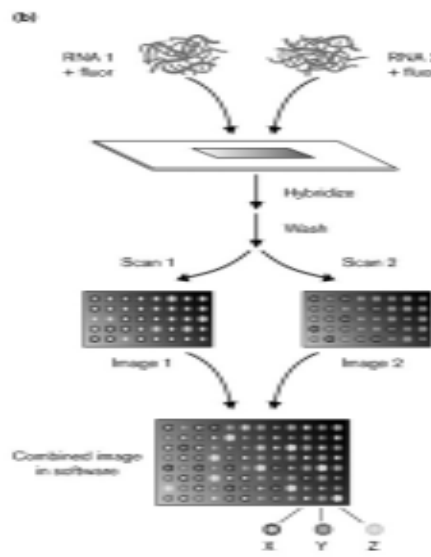
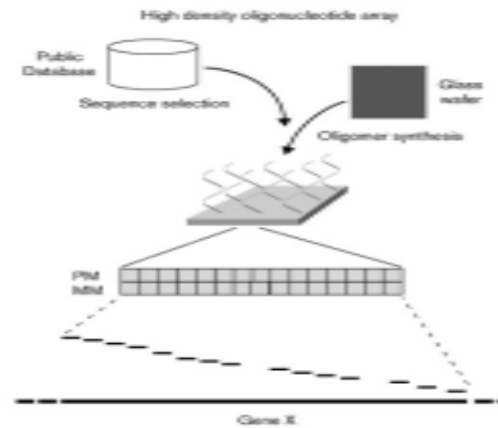
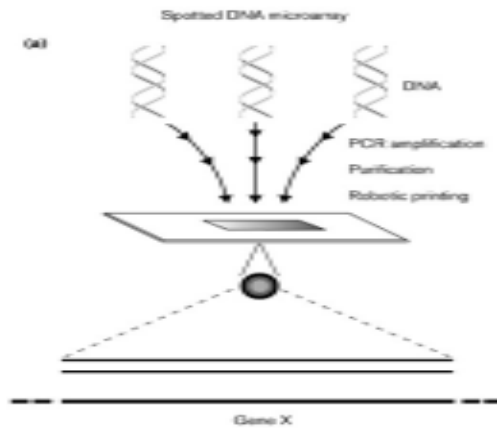


# System architecture for “Gene Chip Analysis Tools (GCAT)”



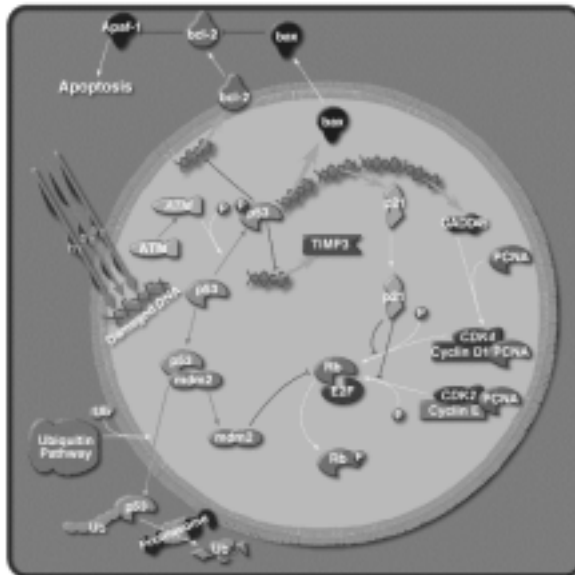
# Database and cross-experiment comparison





Current Opinion Microbiology

# Microarray Annotation Program (MAP)



1. This gene is: Hs.95577 (GeneCard)

[1] Transcriptional regulation:

=>PALS(Alternative Alternative Splicing Database)

=>TRANSFAC(Underconstruct!!)

[2] Biological pathways and disease phenotype:

BioCarta (Signaling pathways)

Pathway_name	Gene_symbol
<u>rbPathway</u>	CDK4
<u>p53Pathway</u>	CDK4
<u>g1Pathway</u>	CDK4
<u>cellcyclePathway</u>	CDK4

KEGG (Metabolic pathways & regulatory pathways)

Pathway_name	EC_number
<u>Inositol phosphate metabolism</u>	2.7.1.-
<u>Sphingoglycolipid metabolism</u>	2.7.1.-
<u>Cell cycle</u>	2.7.1.-
<u>Nicotinate and nicotinamide meta</u>	2.7.1.-
<u>Starch and sucrose metabolism</u>	2.7.1.-

OMIM (Online Mendelian Inheritance in Man)

... etc.



YM-Biochem

Data mining can discover hidden relations among different categories

Gene Ontology™ (GO) classify genes

The first level of GO:

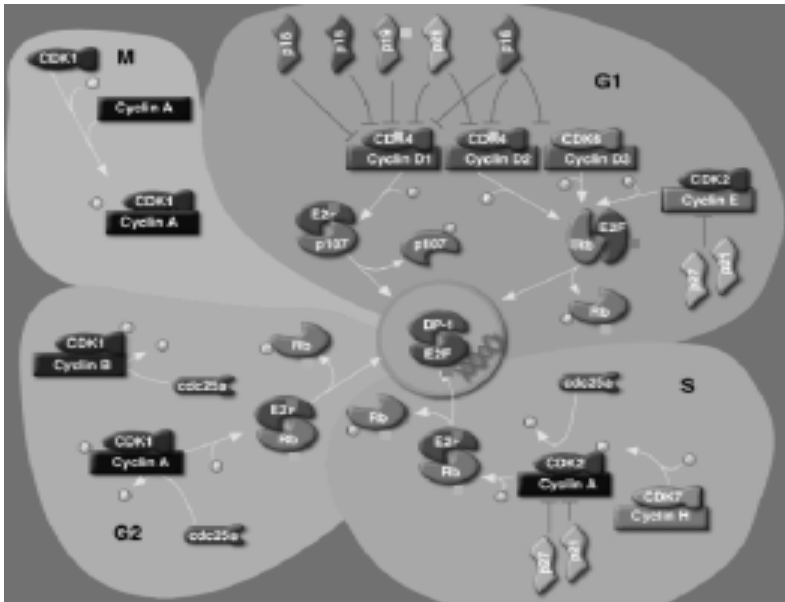
1. Cellular components
2. Molecular function
3. Biological process

Navigator in GeneSpring™

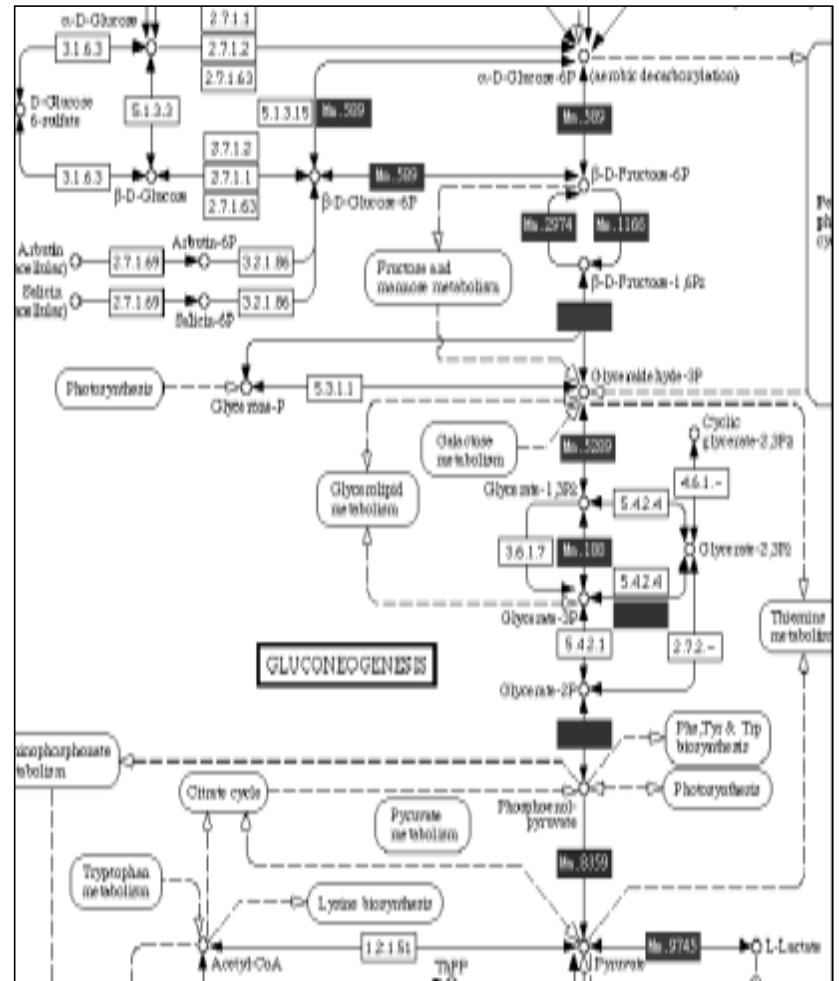


YM-Biochem

Pathway presents the relation of genes

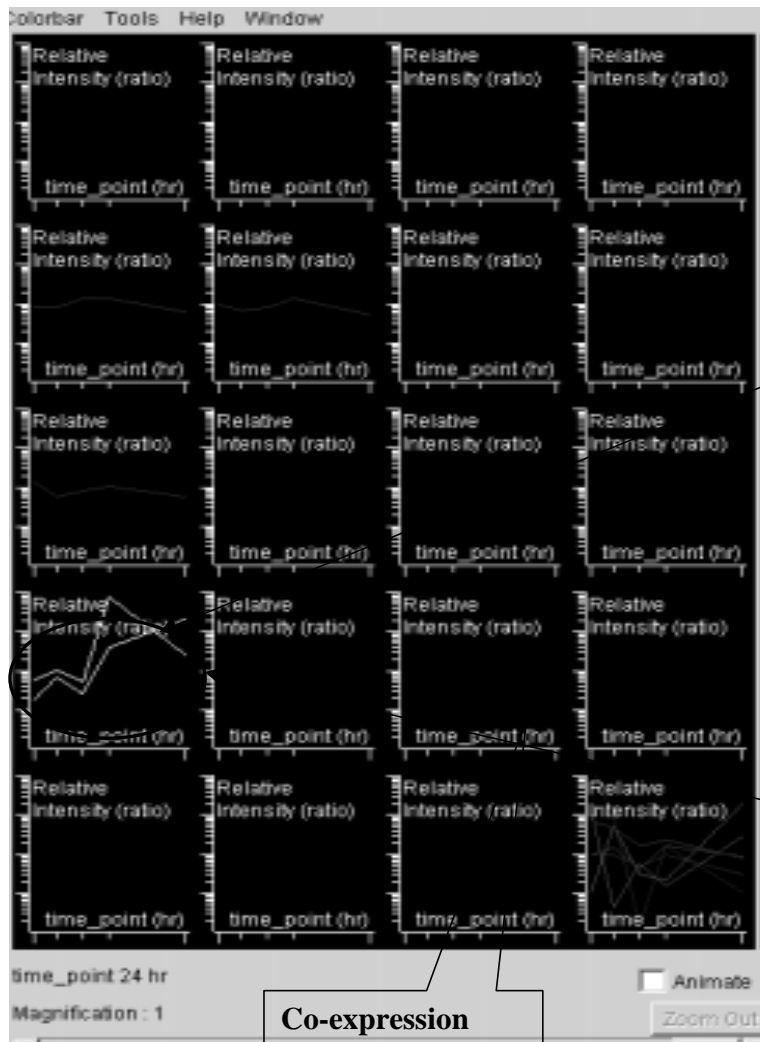


Genome browser in GeneSpring™



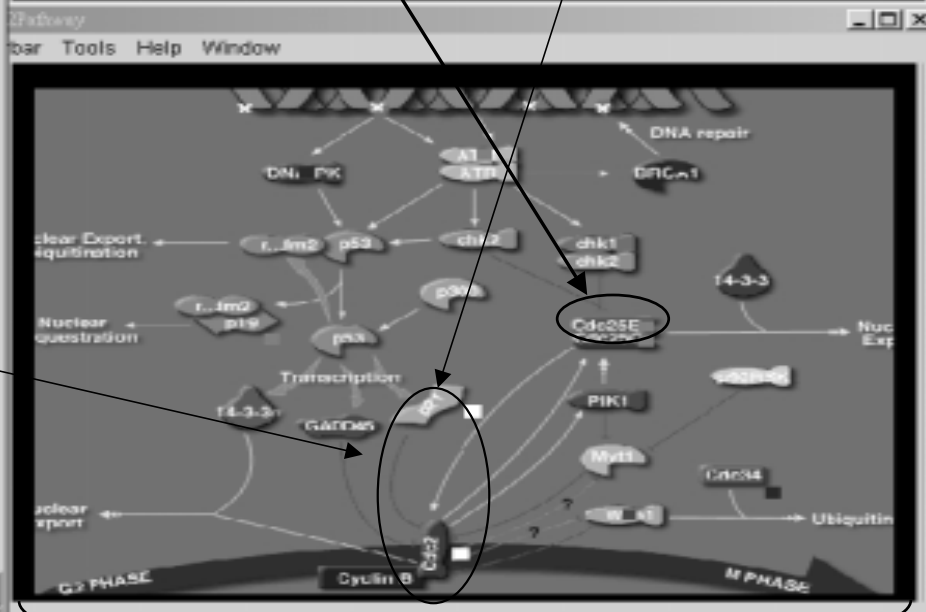
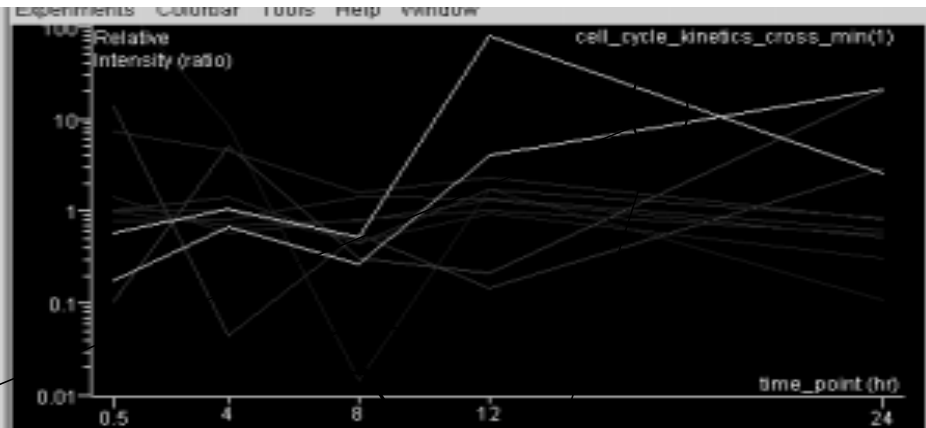
YM-Biochem

### Clustering



Co-expression

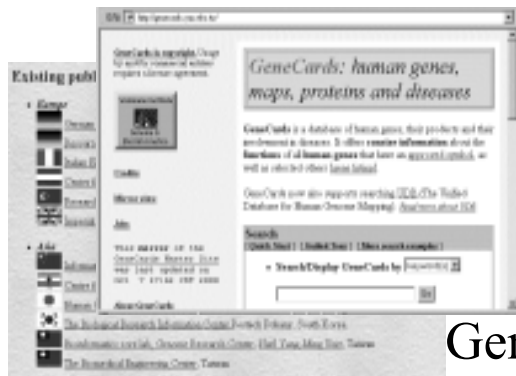
### Expression pattern



Pathways

YM-Biochem

# Mirror sites that support the user-centric environment



GeneCards



SRS6



KEGG



Gene ontology browser



PubGene



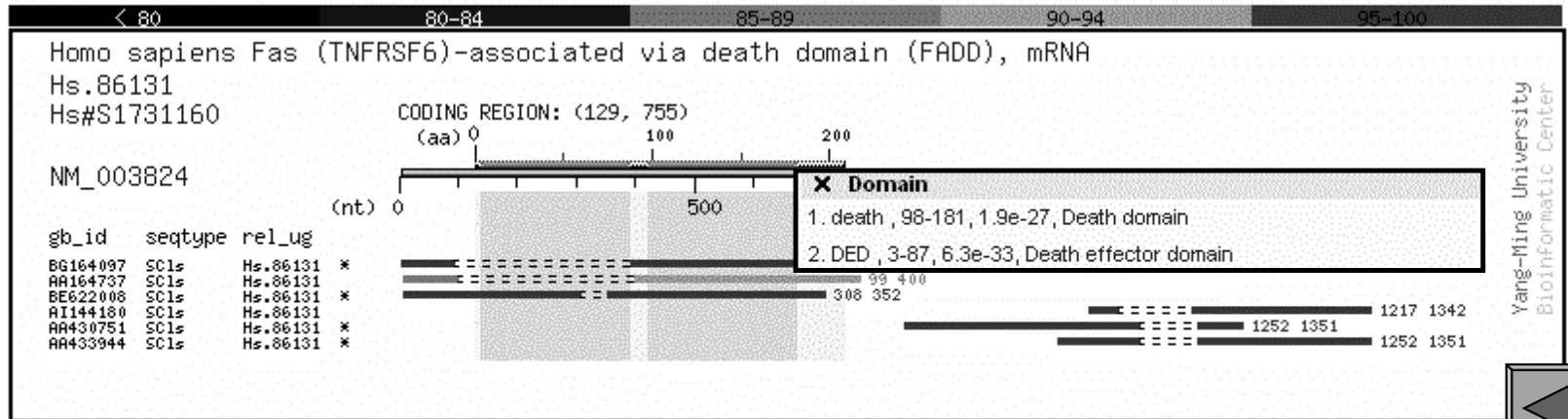
YM-Biochem

# PALS db: Putative ALternative Splicing database

Putative Alternative Splicing Database (PALS db)

Release 3, Web Interface Ver. 0.9.5.1

<a href="#">Summary</a>	<a href="#">Text</a>	<a href="#">OMIM</a>	<a href="#">HUGO</a>	<a href="#">dbSNP</a>	<a href="#">SAGE</a>	<a href="#">GeneCards</a>	<a href="#">Ensembl</a>	<a href="#">Search again!</a>	<a href="#">Help</a>
<a href="#">iProClass</a>	<a href="#">InterPro</a>	<a href="#">Dart</a>	<a href="#">CDD</a>	<a href="#">PSORT</a>	<a href="#">TMHMM</a>	<a href="#">NetOGlyc</a>	<a href="#">ma_mRNA</a>	<a href="#">Homologs</a>	<a href="#">Similarities</a>

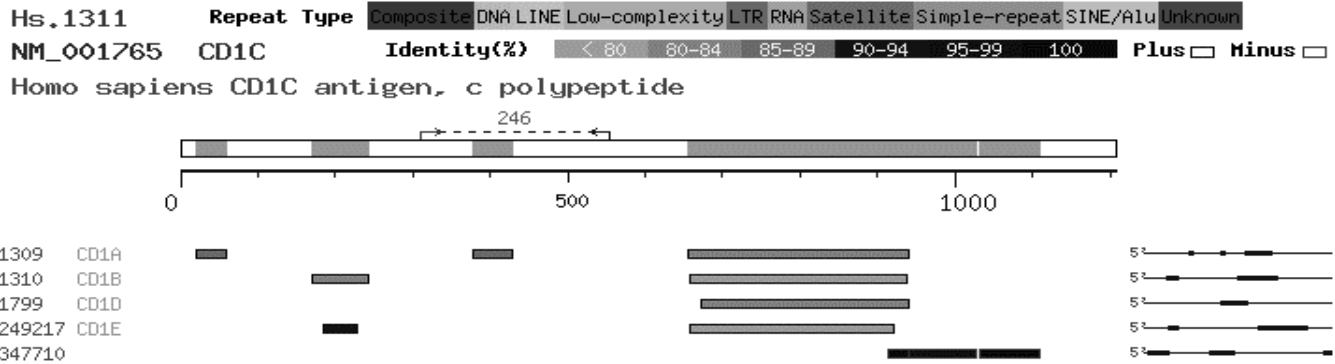


Yang-Ming University  
Bioinformatic Center

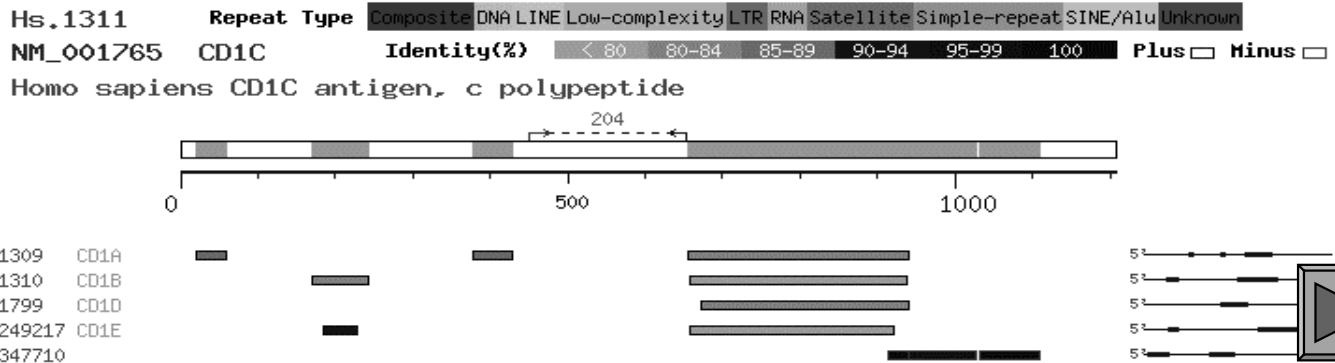


# Primers for qRT-PCR and Hybridization Probe

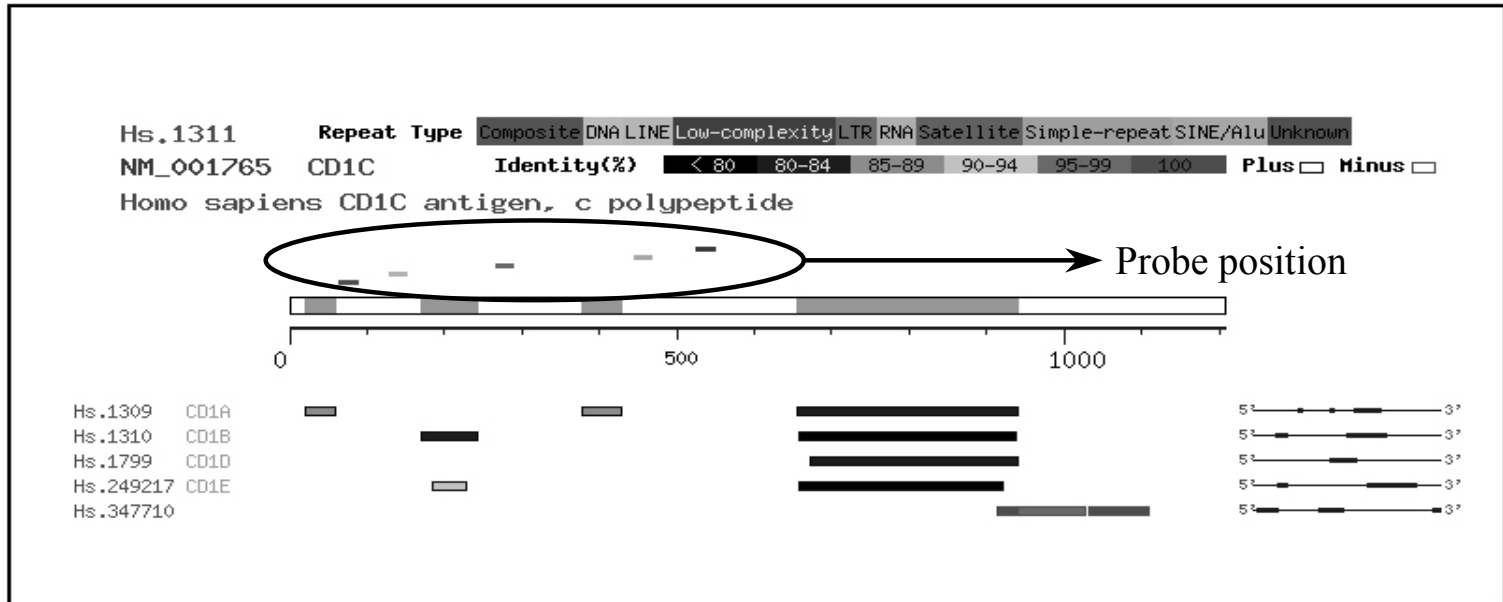
## qRT-PCR primer



## PCR probe for Northern hybridization



# Oligonucleotide Array Probe Design

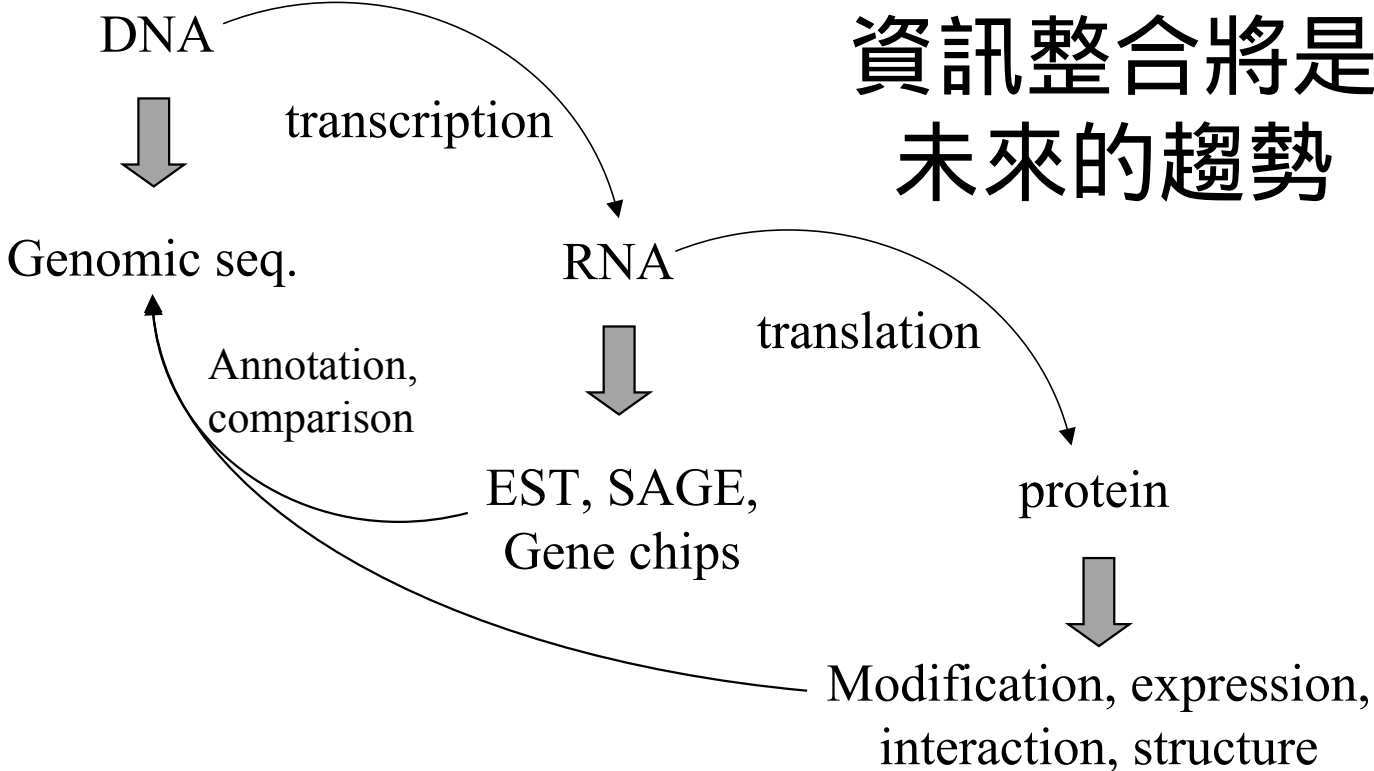


Serial Number	Probe Sequence	Probe Start Position	Probe End Position	Probe Melting Temperature
1	GCCCAAAGTGTCTGTCATCTACT	526	548	63.5236
2	TTTCAGGTAGCTTTC AACGGATT	445	467	61.8043
3	CCAAGGGCAACTTCAGCAATGAA	266	288	65.3689
4	CTCCTTCCATGTCATCCAGATCT	129	151	62.81
5	AGTTTCTGCTGCTAGCTCTTCTT	65	87	63.28

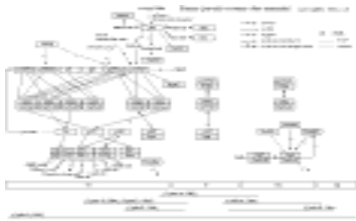
# Summary

- Established a pipeline to annotate human and mouse genes on microarray
- Establish the relations among genes, Gene Ontology<sup>TM</sup>, and the pathways of KEGG and BioCarta.
- Create an user-centric bioinformatics environment by using GeneSpring<sup>TM</sup> as a viewer.

# 資訊整合將是 未來的趨勢



Components



Pathways and regulatory circuits



Hypothetical cell

數據 => 資訊 => 知識 => 技術

