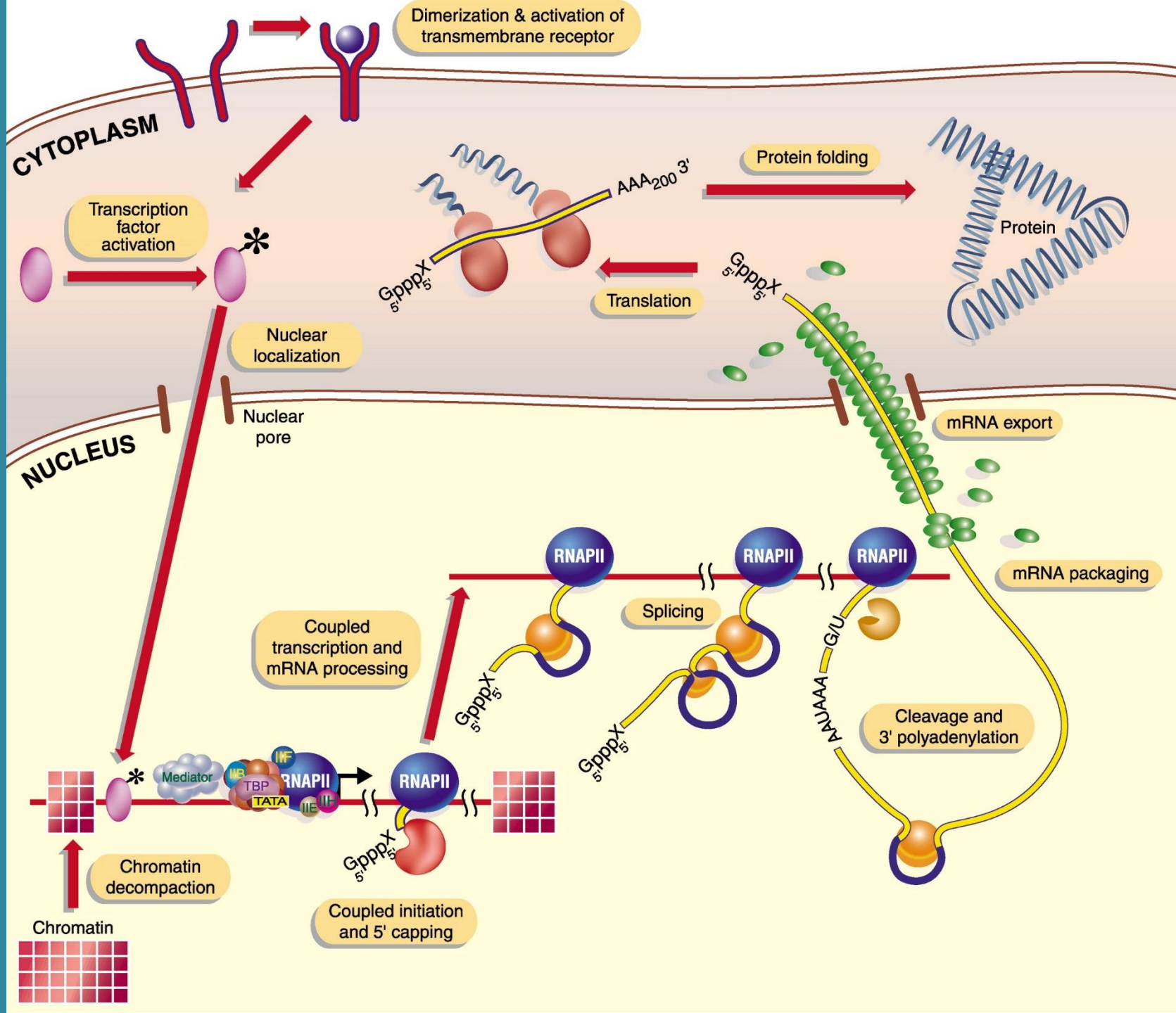
Chapter-06. 基因表达数据分析

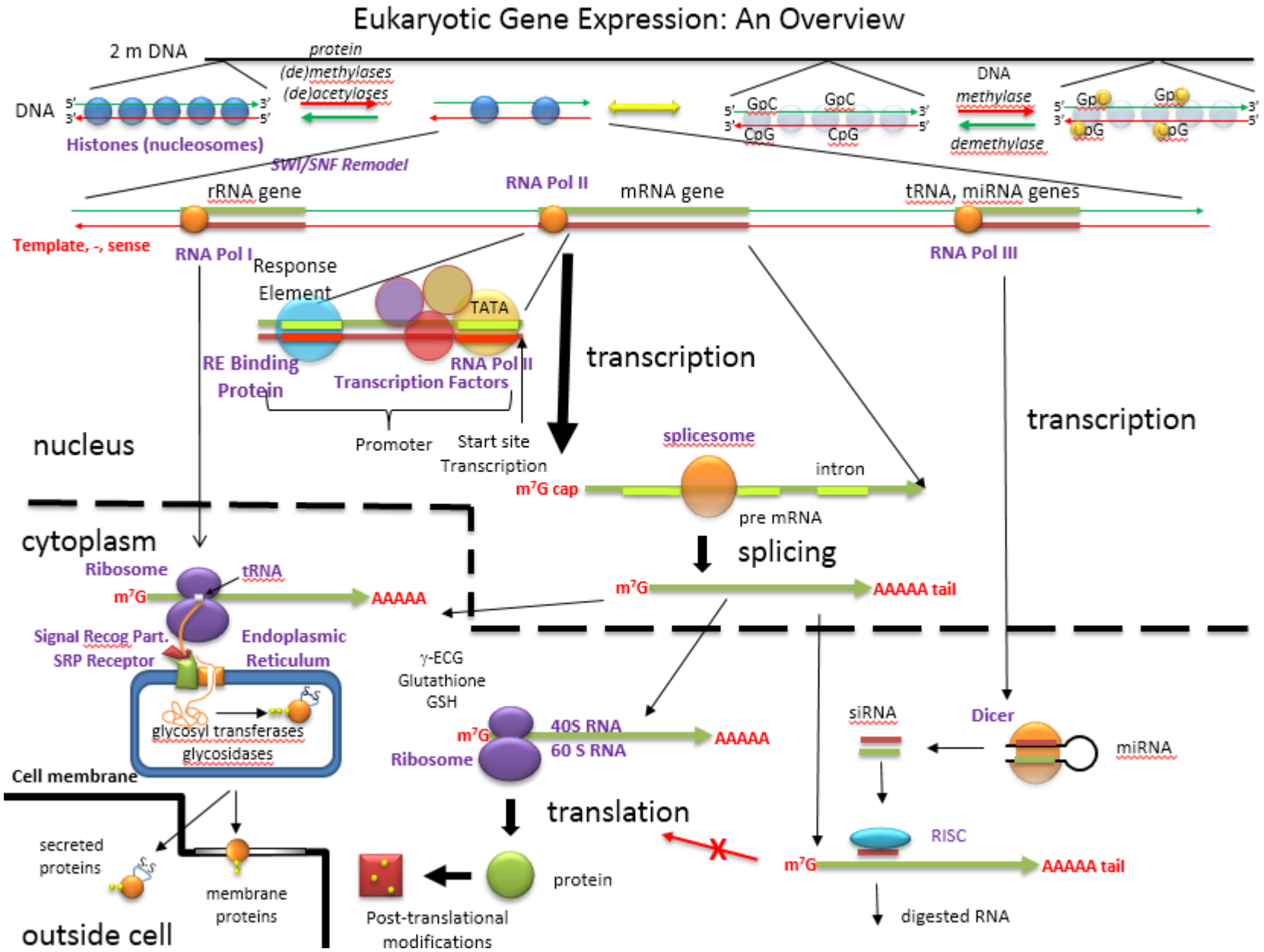


本章内容提要

- 6.1 引言
- 6.2 什基因表达测定平台及数据库
- 6.3 基因芯片数据的基本处理
- 6.4 基因芯片数据的综合分析
- 6.5 基因表达数据分析案例

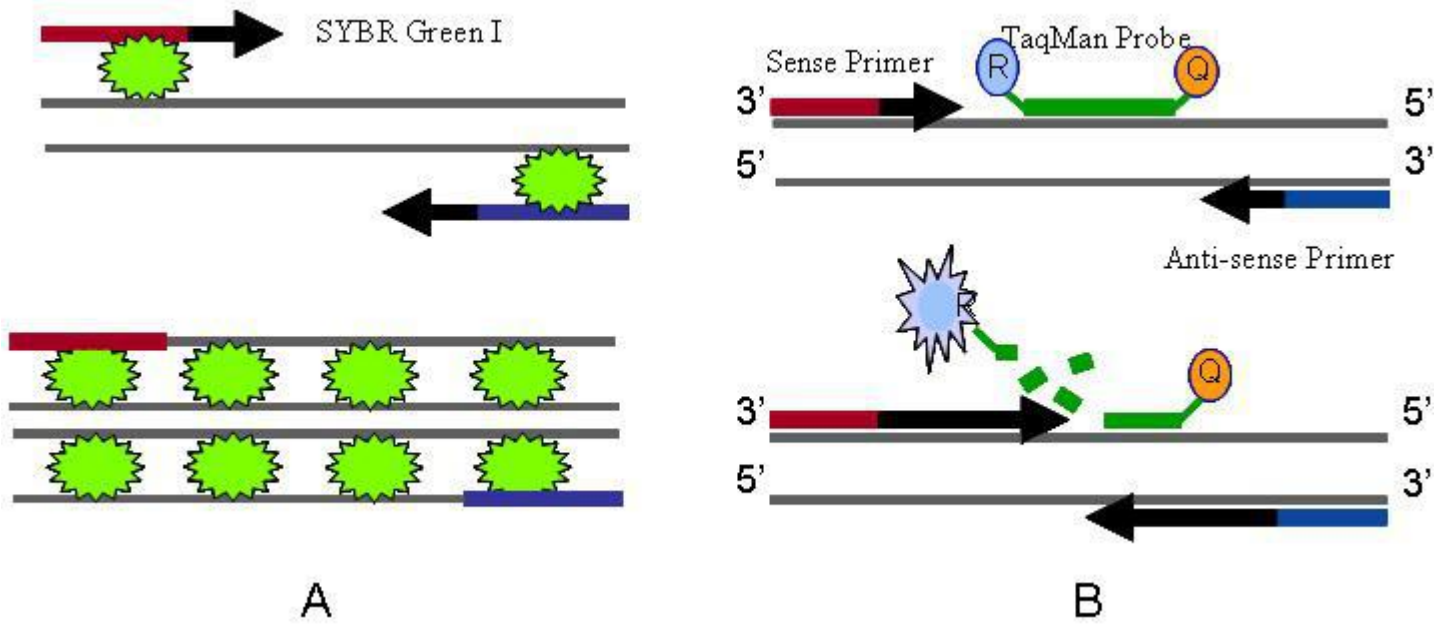
第1节：引言

- 基因表达
- 基因表达测定的原理
- 基因表达测定的应用



1.1 基因表达的基本过程及其可能的调控环节

（1）低通量方法，如RT-qPCR等



上图中，A和B中使用的荧光探针有所不同，导致原理略有差异

（2）高通量方法，如DNA芯片和二代测序（NGS）技术等

Platforms

Microarray Platforms

Affymetrix
GeneChip
Platform

Agilent
SurePrint
Platform

Illumina
BeadArray
Platform

HT Sequencing Platform

Illumina
HiSeq 2000
MiSeq

HT RT-PCR System

Applied Biosystems
7900HT Fast
Real-Time PCR System

Applications

DNA

Genomic Variation	Epigenomic profiling	ChIP applications
SNP	DNA methylation	ChIP-seq ChIP-on-chip

RNA

Transcriptomics			
IVT and WT expression	WT Exon expression	RNA-seq applications	miRNA analysis

Data analysis

Basic analysis

Data Pre-processing	Quality Control	Statistical comparisons

Bioinformatics

Biological networks	Data visualizations	Customized support

Additional Instrumentation

Beckman Coulter
Biomek FXP robot
PCR automation

QIAGEN
QIAcube robot
Automatic RNA/DNA
purifications

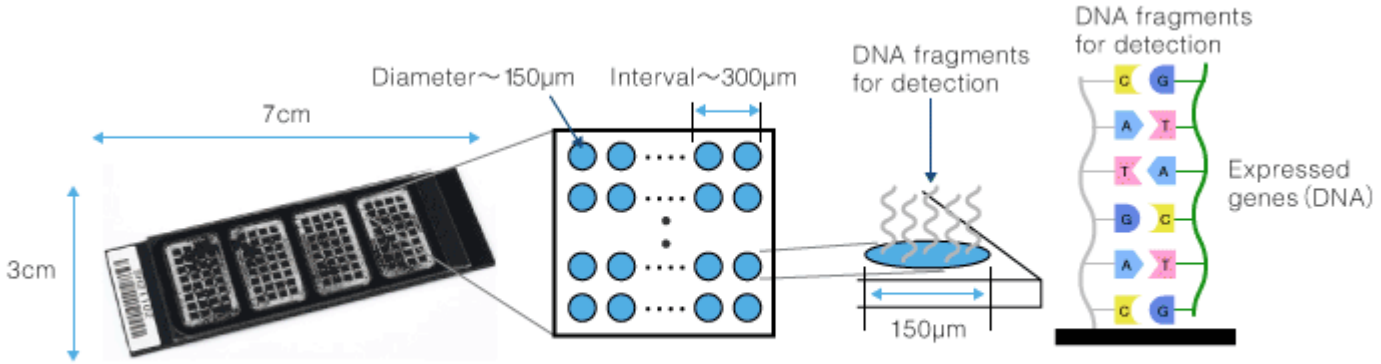
Agilent
2100 Bioanalyzer
2200 Tapestation
RNA/DNA QC

Ref: http://www.bea.ki.se/resources_platforms.html

illumina CSPro CERTIFIED SERVICE PROVIDER

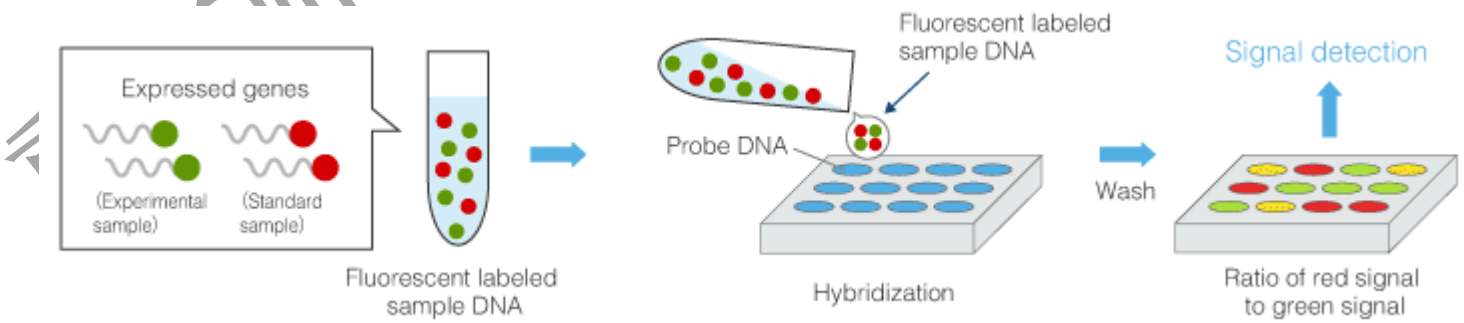
Agilent Technologies

affymetrix Biology for a better world

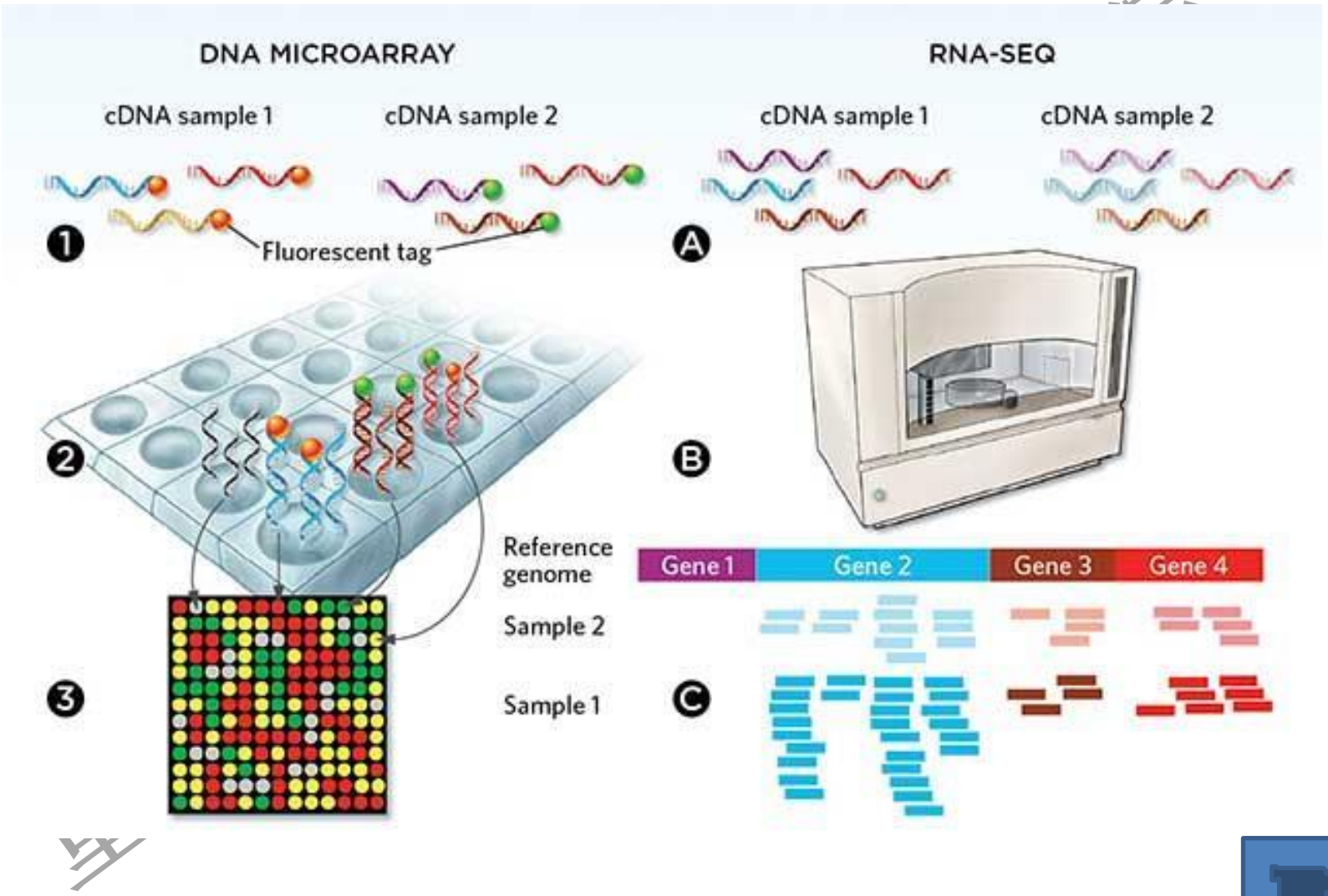


(Several tens to several tens of thousands of types of DNA fragments for detection are immobilized)

Analysis is conducted utilizing hybridization between DNA fragments for detection and DNA expressed



基因芯片与二代测序的原理比较



两者各有千秋。目前，RNA-seq可能更加流行。

DNA测序技术的发展历程

- **第一代测序技术：Sanger法等**

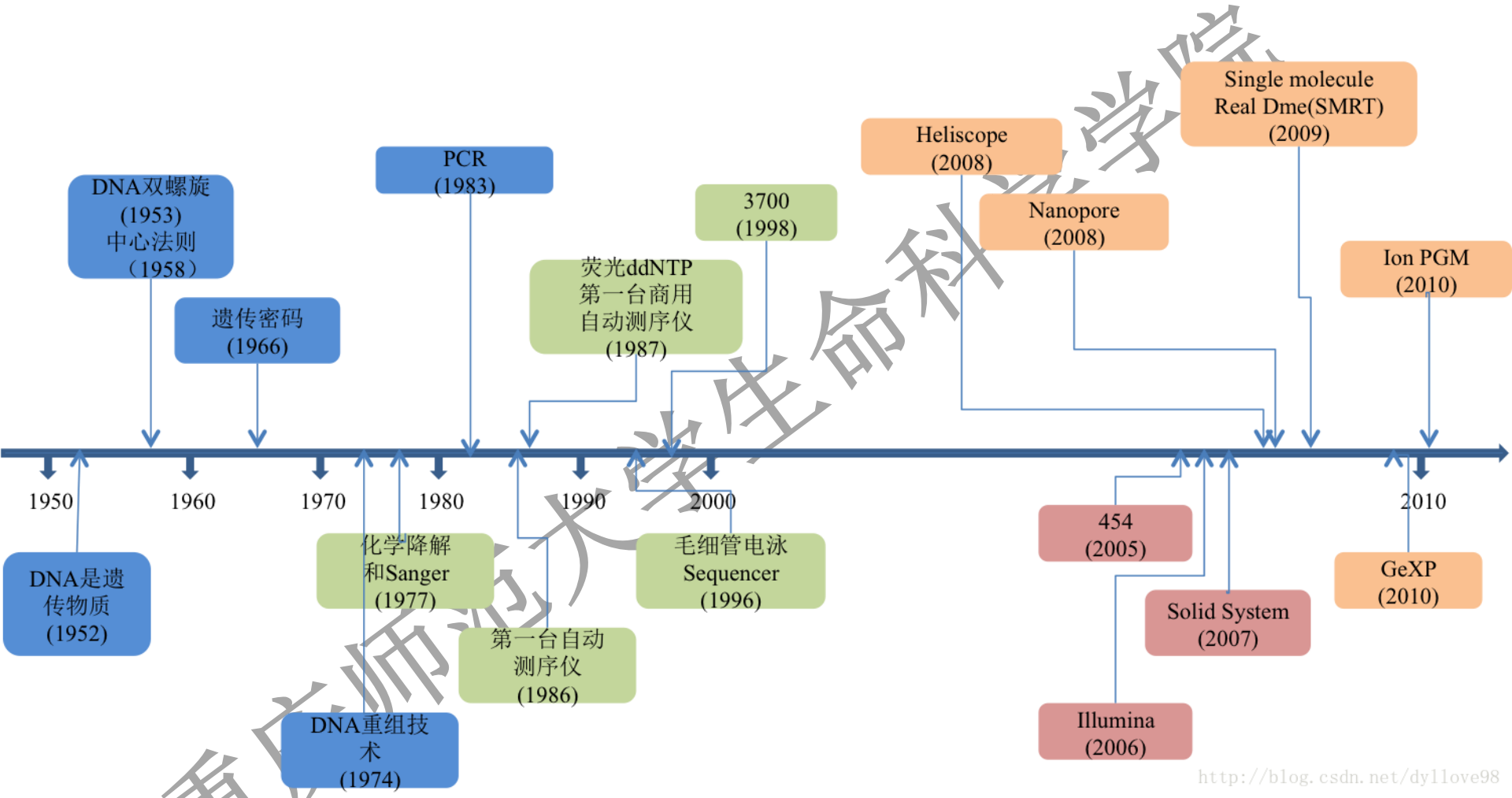
  速度快，但是一次只能测一条单一的序列，且最长也就能测1000-1500bp，被广泛应用在单序列测序上。

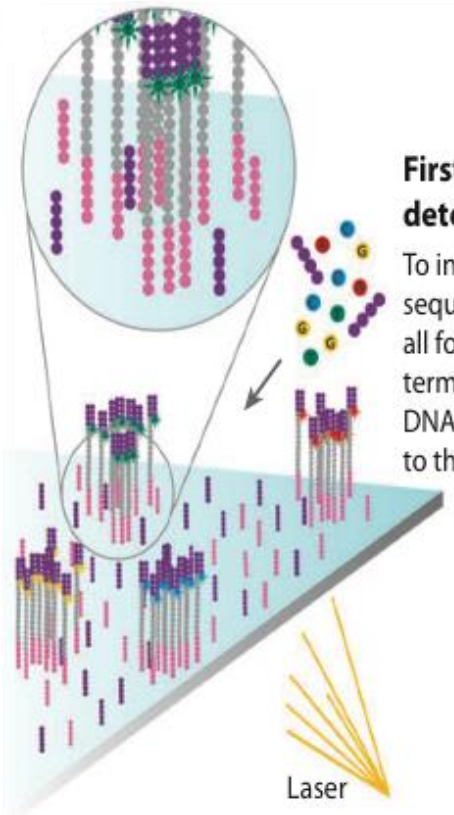
- **第二代测序技术：illumina-Solexa/Hiseq, Roche-454以及ABI Solid等**

  一次能够测大量的序列，但是片段被限制在了250-300bp，由于是通过序列的重叠区域进行拼接，所以有些序列可能被测了好多次。由于建库中利用了PCR富集序列，因此有一些量少的序列可能无法被大量扩增，造成一些信息的丢失，且PCR中有概率会引入错配碱基。

- **第三代测序技术：如Nanopore测序技术。**

  一次能测好多序列，但是测序的长度达到了10kb左右，并且不需要PCR富集序列，直接测序，这就解决了信息的丢失，以及碱基错配的问题。但目前来说三代测序依然有一定的缺陷：三代测序技术依赖DNA聚合酶的活性，且成本很高，目前的错误率在15%-40%，极大地高于二代测序技术的错误率不过好在三代的错误是完全随机发生的，可以靠覆盖度来纠错。





First chemistry cycle: determine first base

To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.

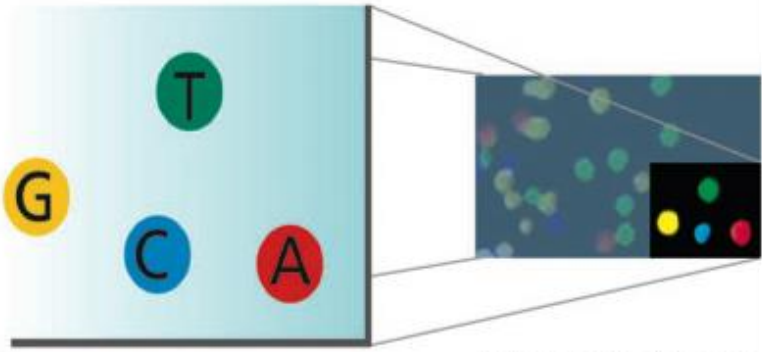
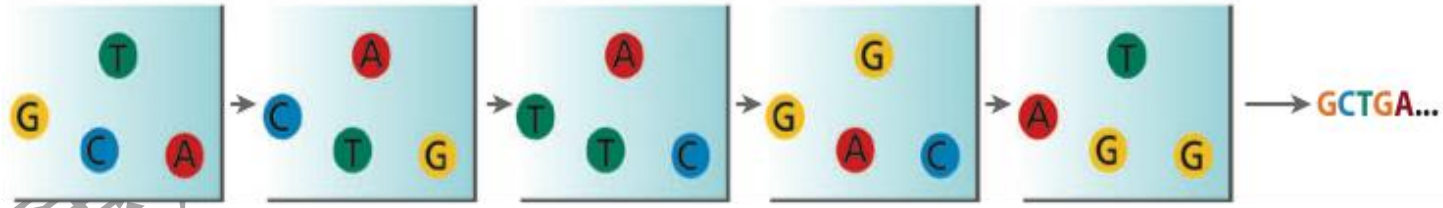


Image of first chemistry cycle

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Before initiating the next chemistry cycle

The blocked 3' terminus and the fluorophore from each incorporated base are removed.



Sequence read over multiple chemistry cycles

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

- ✓ 测定组织特异性表达基因
- ✓ 基因功能的分类
- ✓ 癌症的分类和预测
- ✓ 临床治疗效果预测
- ✓ 基因与小分子药物、疾病之间的关联
- ✓ 干细胞的全能型、自我更新和细胞命运决定研究
- ✓ 动植物的发育研究
- ✓ 环节对细胞基因表达的作用
- ✓ 环境监测
- ✓ 物种的繁育

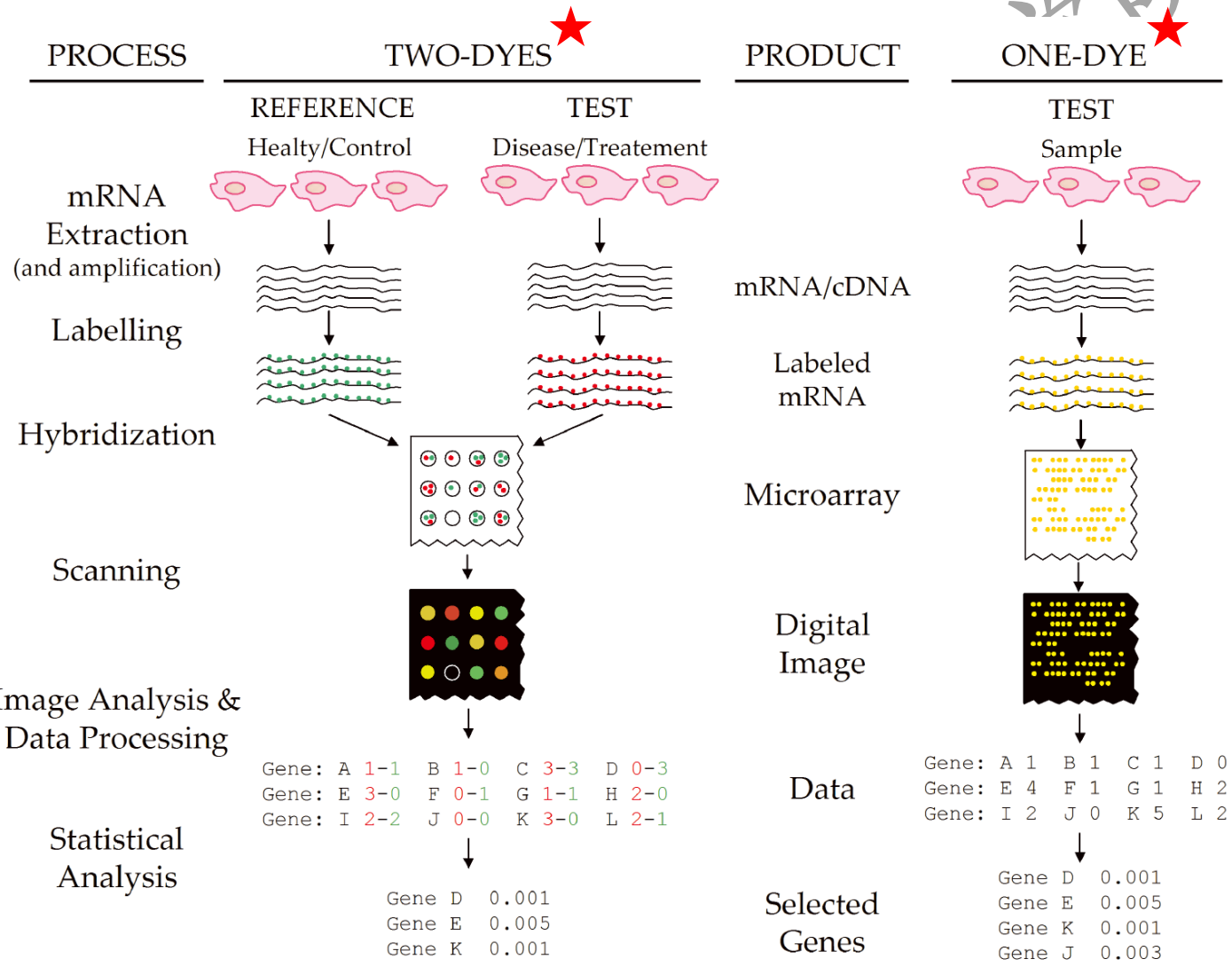
Table 1 | **Status of microarray-based processes**

Process	Status*
Transcriptional profiling	Mature, but still to be improved
Genotyping	Mature, but still to be improved
Splice-variant analysis	In progress
Identification of unknown exons	Early stages
DNA-structure analysis	Pilot phase
ChIP-on-chip	In progress
Protein binding	Under development
Protein–RNA interaction	Idea
Chip-based CGH	In progress
Epigenetic studies	Under development
DNA mapping	Mature
Resequencing	In progress
Large-scale sequencing	Under development
Gene/genome synthesis	Early stages
RNA/RNAi synthesis	Pilot phase
Protein–DNA interaction	Under development
On-chip translation	Under development
Universal microarray	Under development

Hoheisel, J.D. Microarray technology: beyond transcript profiling and genotype analysis. *Nature reviews genetics* **7**, 200 (2006).

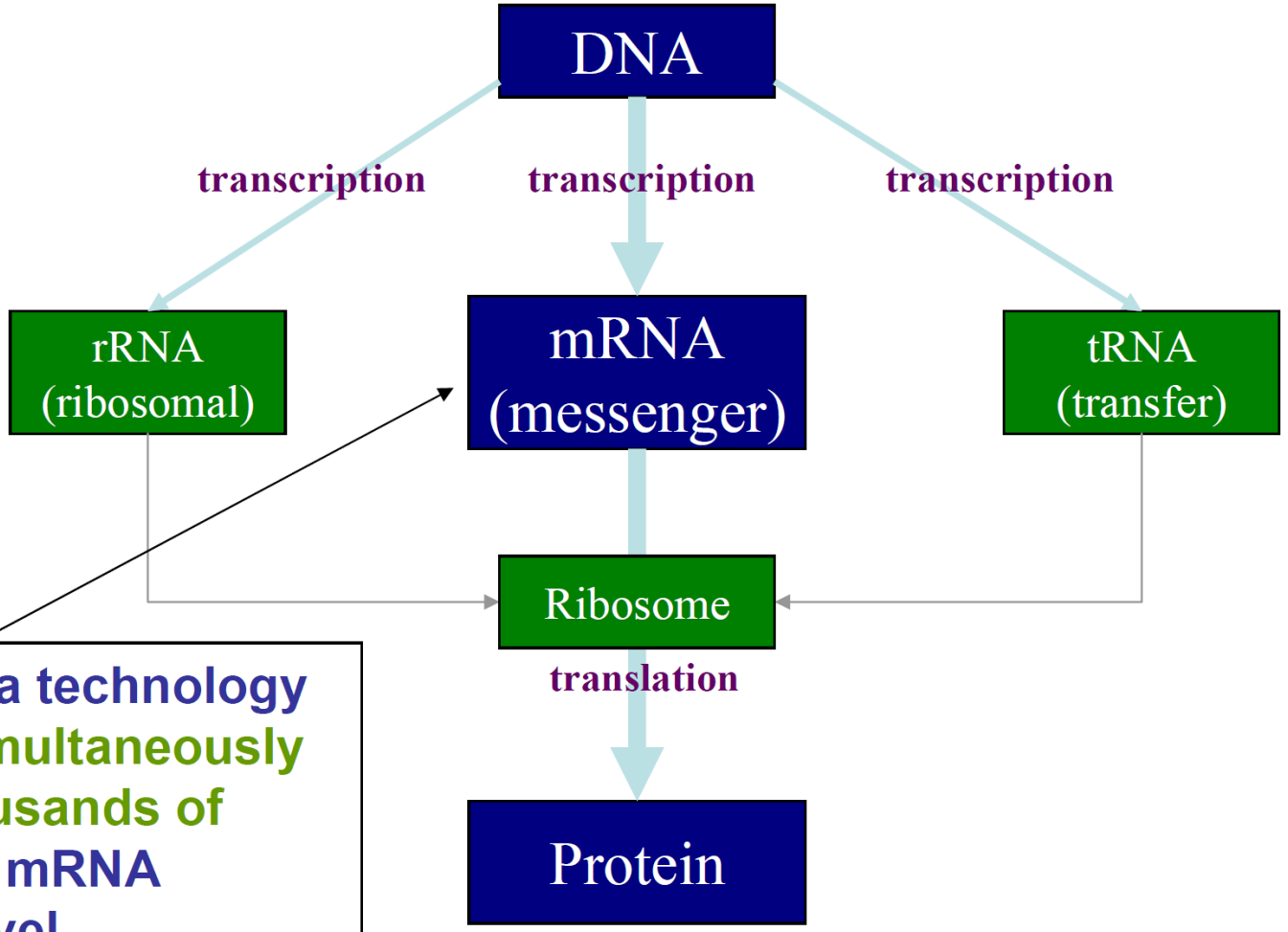
第2节：基因表达测定平台及数据库

一、DNA微阵列（DNA芯片）

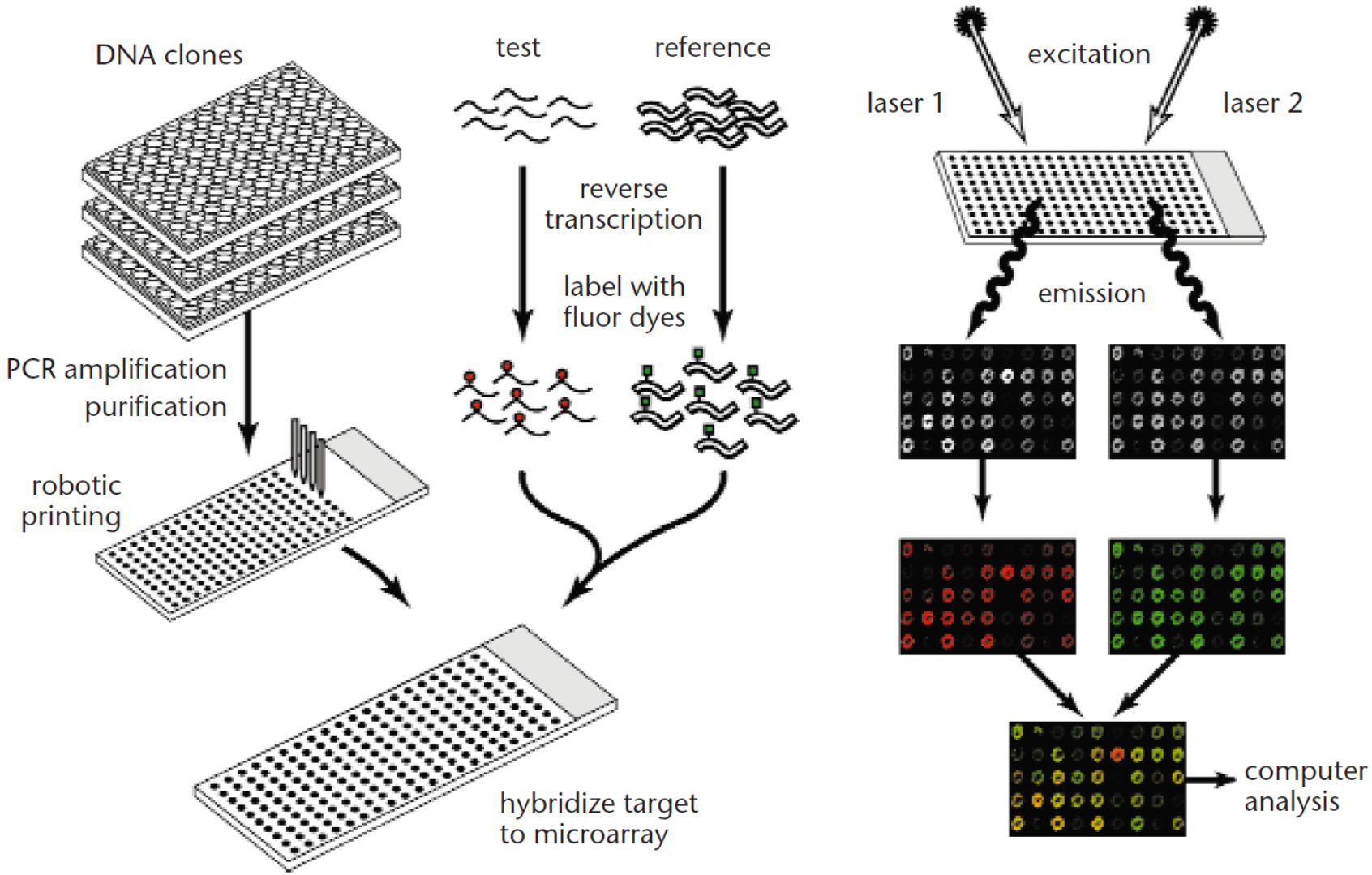


The central dogma of molecular biology:

DNA —*transcription*→ **RNA** —*translation*→ **Protein**



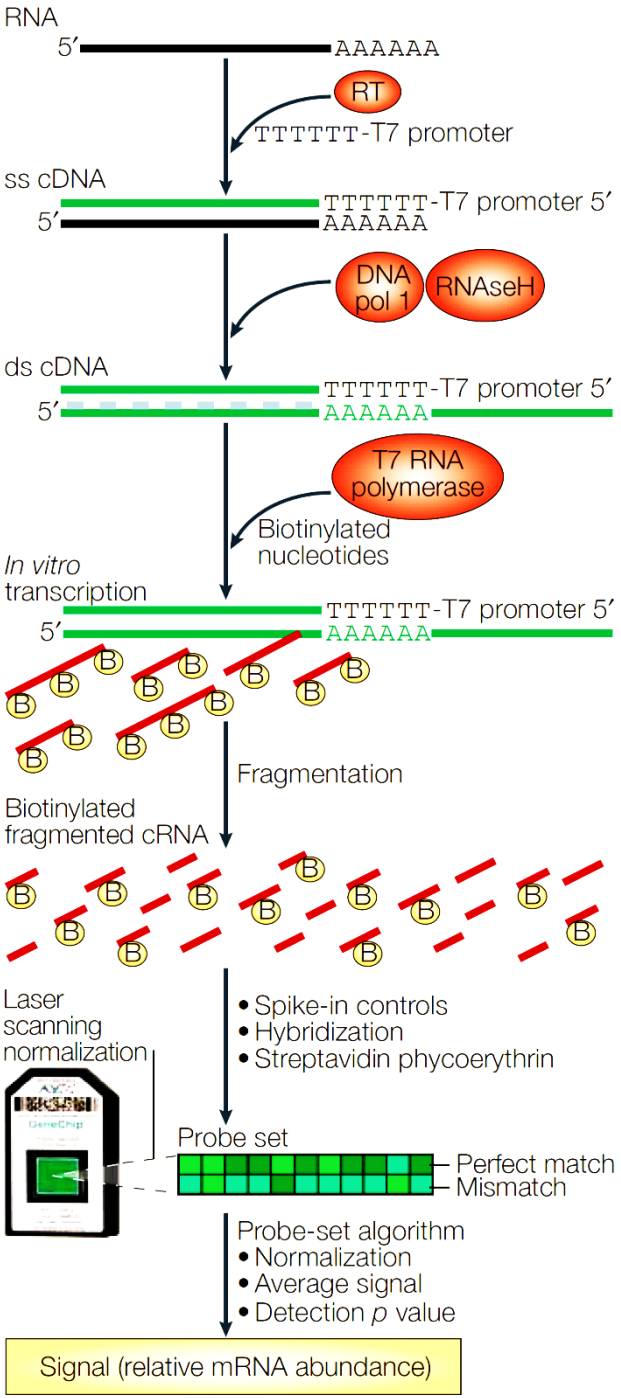
Microarray is a technology to globaly (simultaneously detecting thousands of genes) detect mRNA expression level.



1. 样品准备；2. 杂交；3. 图像扫描；4. 数据分析

二、寡核苷酸芯片

- ✓ 寡核苷酸芯片类探针的设计上优于cDNA芯片；
- ✓ 它的探针预先设计并合成的代表每个基因特异片段的约50mer左右长度的序列；
- ✓ 将其点样到特定的基质上制备成芯片；
- ✓ 克服了探针序列太长导致的非特异性交叉杂交和由于探针杂交条件变化巨大导致的数据结果的不可靠



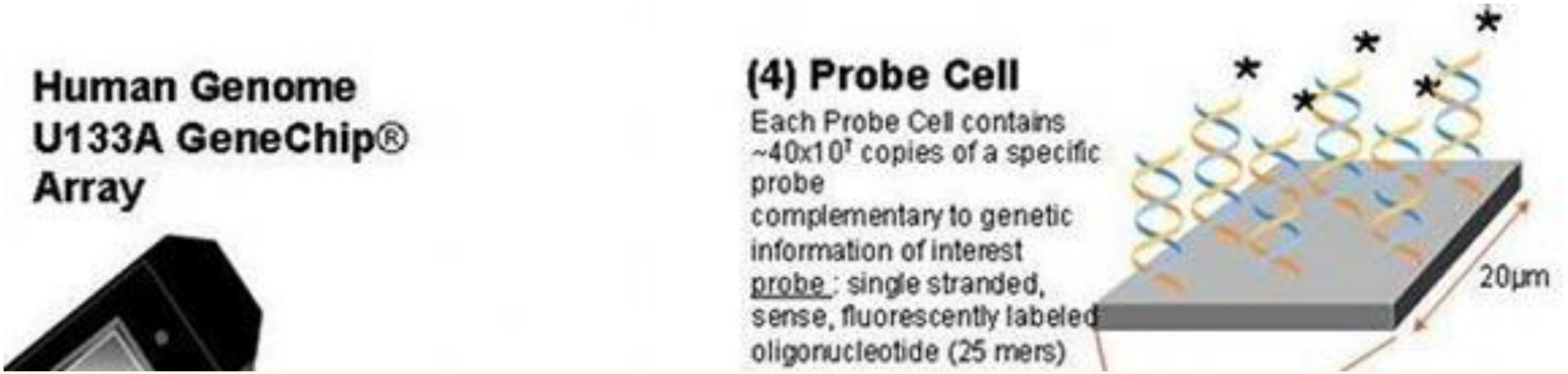
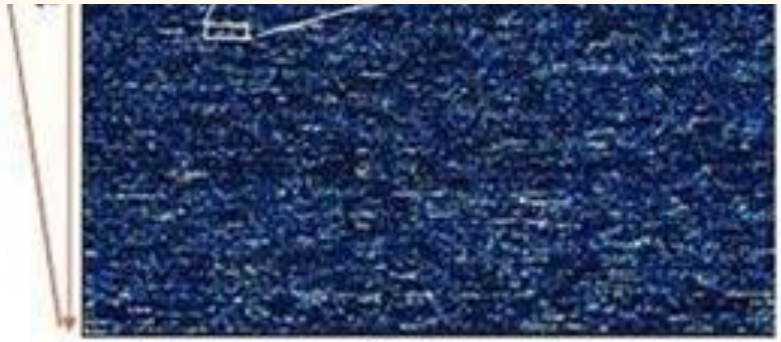


Table 1 | **Comparisons of probe-set analysis algorithms**

Algorithm	Penalty for mismatch signal	Normalization method	Outlier detection and correction	Sensitivity*	Specificity‡
Affymetrix MAS 5.0	High	Individual chips	Little	Good	Excellent
dCHIP difference model	High	Cross-project	Moderate	Good	Excellent
dCHIP	None	Cross-project	Moderate	Excellent	Good
RMA	None	Cross-project	Moderate	Excellent	Good
ProbeProfiler	Moderate	Extensive	Extensive	Good	Good

*Sensitivity is based primarily on ROC (receiver operating characteristic) curves of spike-in mRNA data based on published reports (see http://www.bioconductor.org)[21,23].
‡Specificity measurements are based both on expectations from mismatch weights and published observations in experimental data sets[17,18].

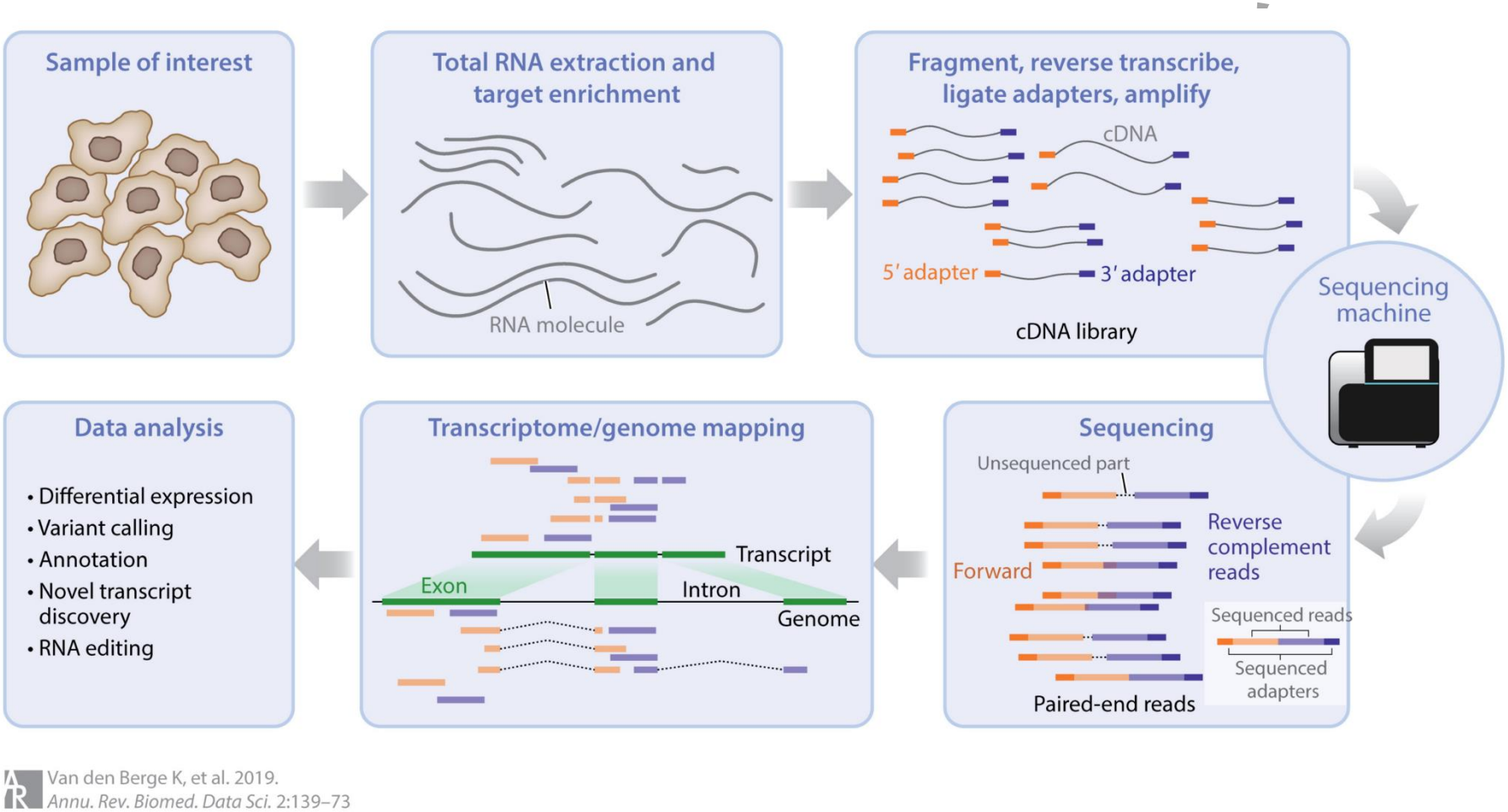


associated by pairs

The Human Genome U133 A GeneChip® array represents more than 22,000 full-length genes and EST clusters.

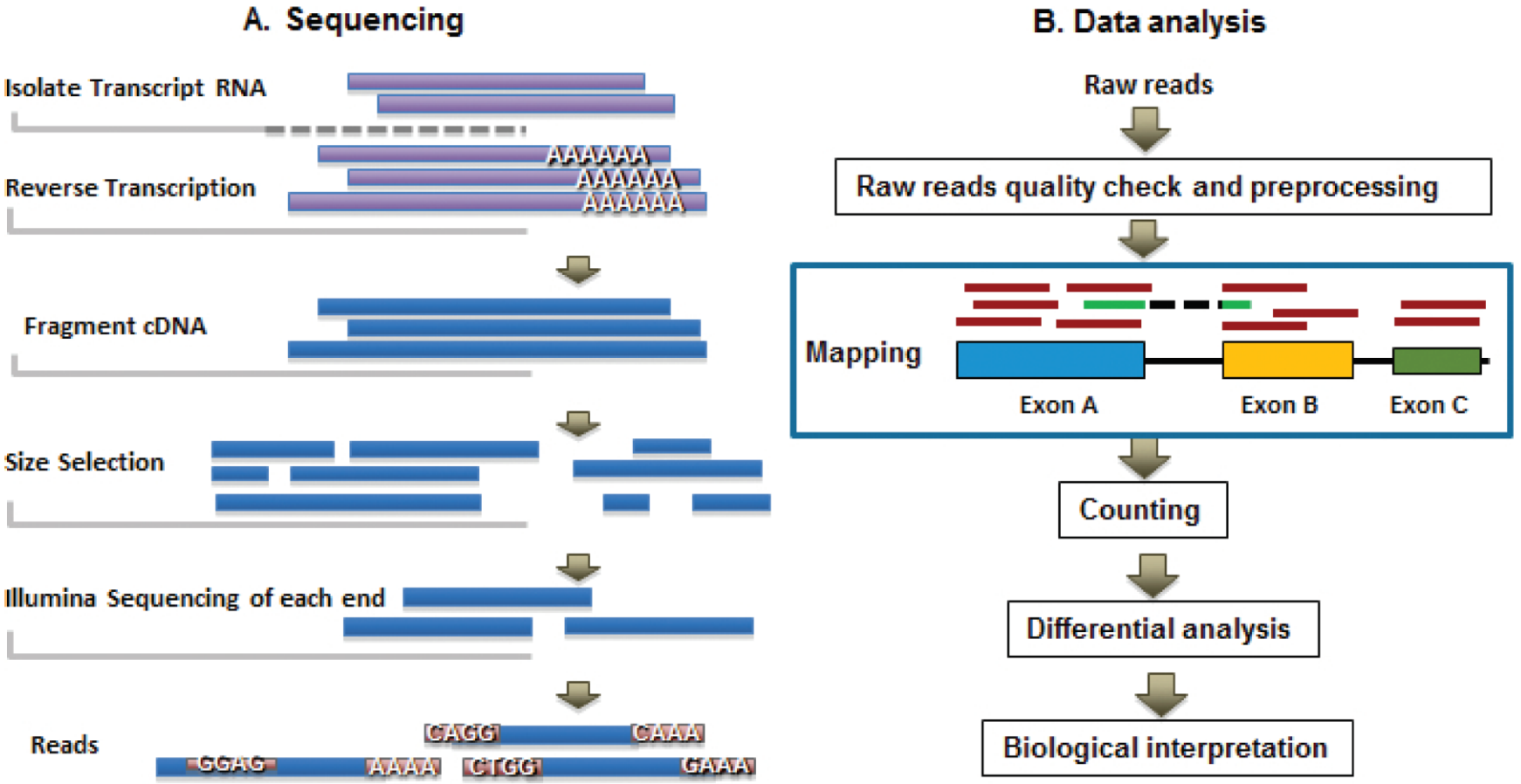
Courtesy: Weizmann Institute of Science
http://www.weizmann.ac.il

三、RNA-seq技术



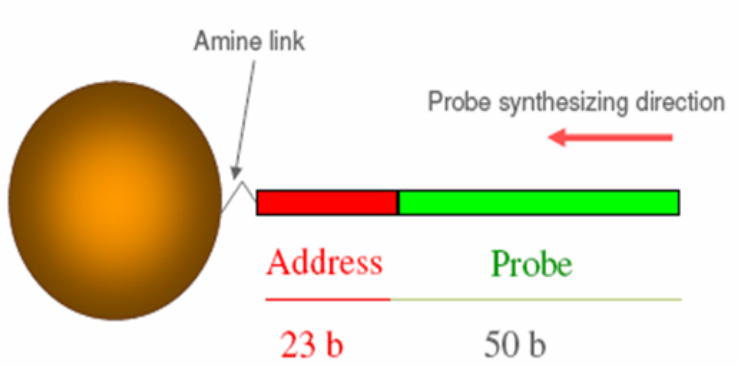
Van den Berge K, et al. 2019.
Annu. Rev. Biomed. Data Sci. 2:139–73

Fig. Summary of RNA-seq data generation. Isolation of RNA from samples of interest, preparation of sequencing libraries, use of a high-throughput sequencer to produce hundreds of millions of short reads, alignment of reads against a reference genome or transcriptome.



RNA-sequencing + *Data analysis*

四、光纤微珠芯片(Bead Array)



五、基因表达仓库

Gene Expression Omnibus, GEO

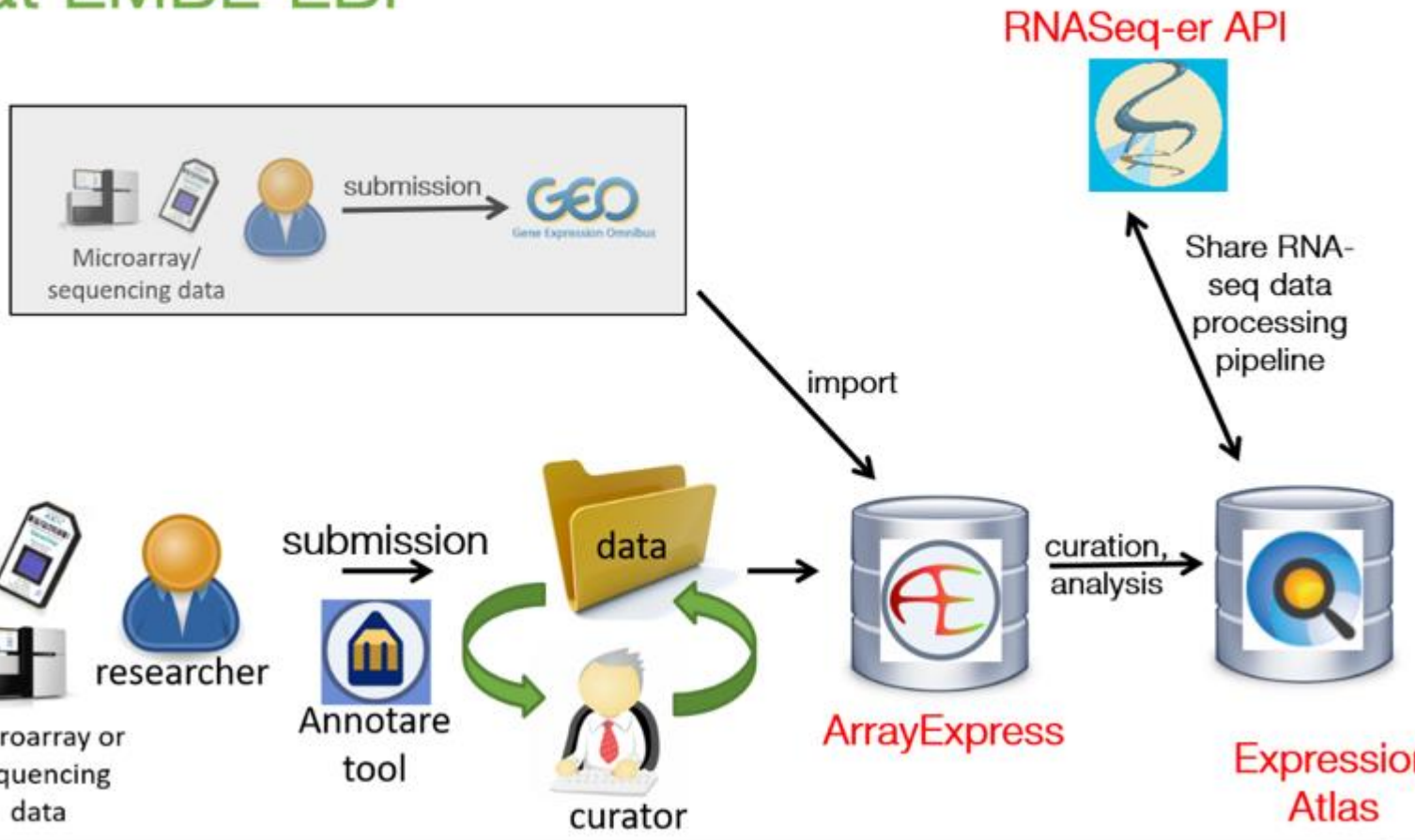
六、EMBL基因表达谱数据库

ArrayExpress

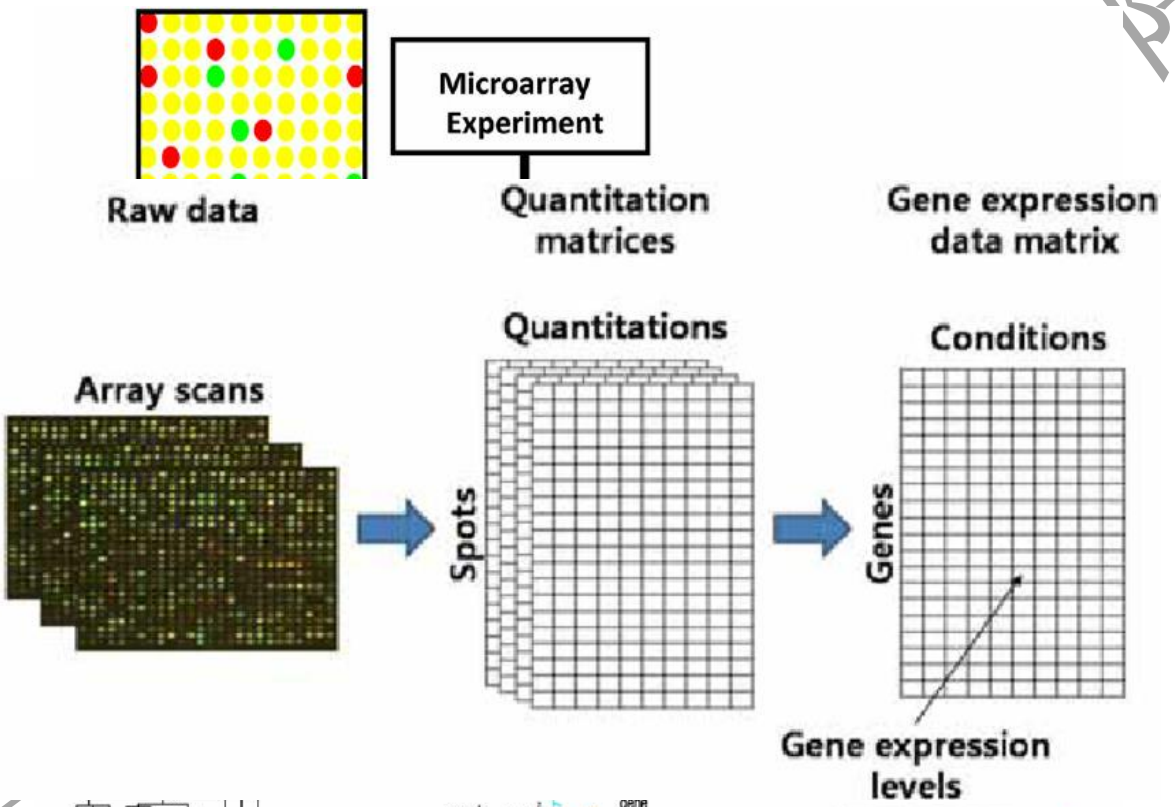
七、其他常用基因表达数据库

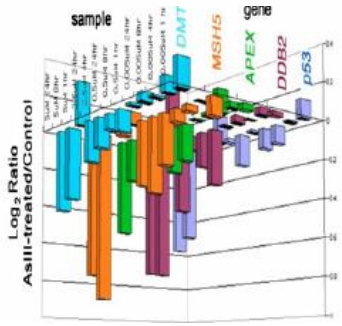
CGED

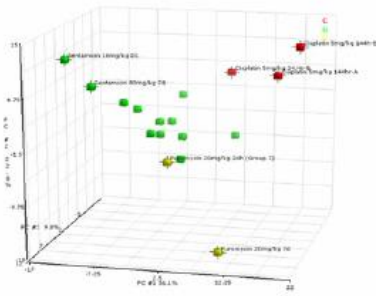
Summary of functional genomics resources at EMBL-EBI

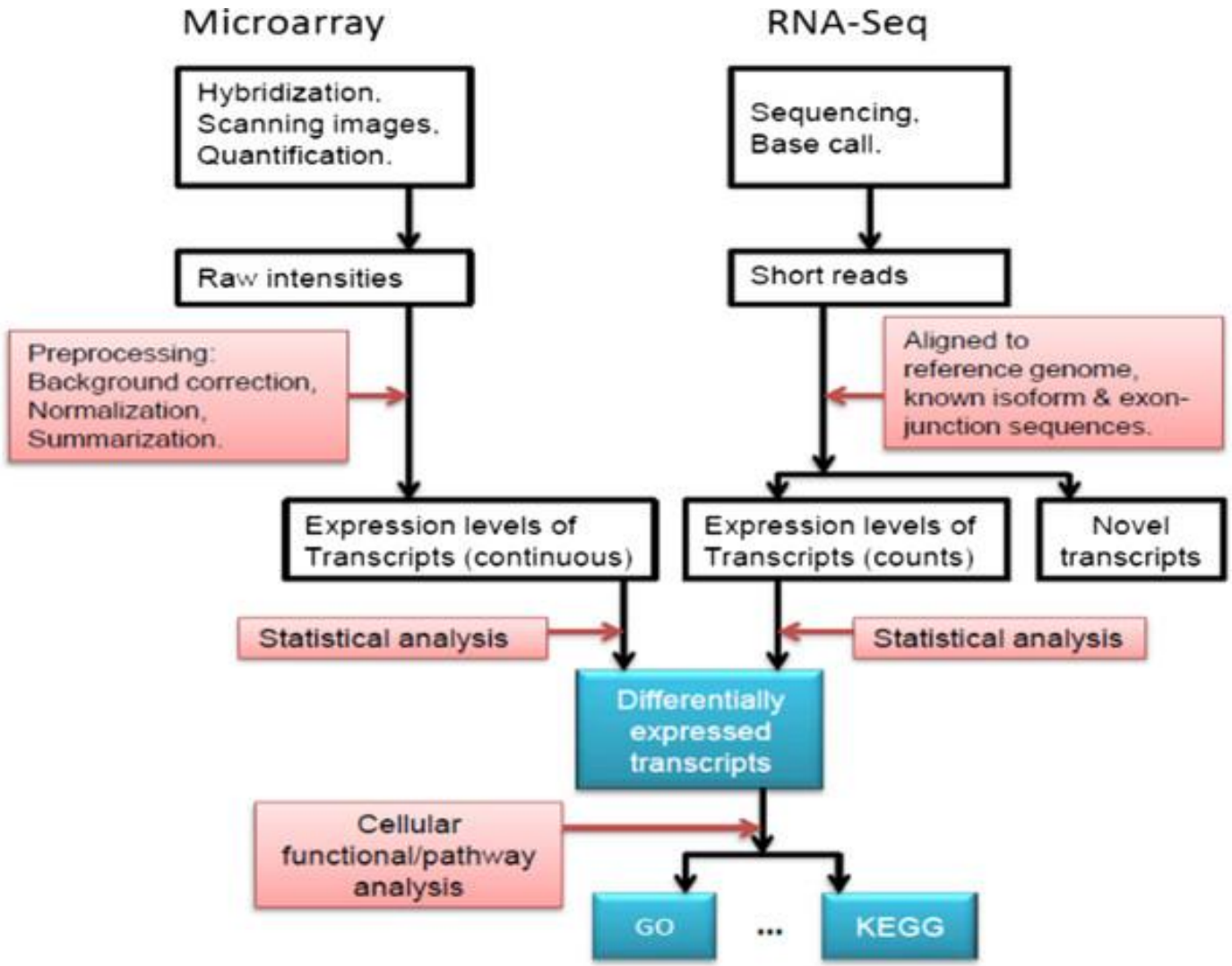


第3节：基因芯片数据分析的基本原理



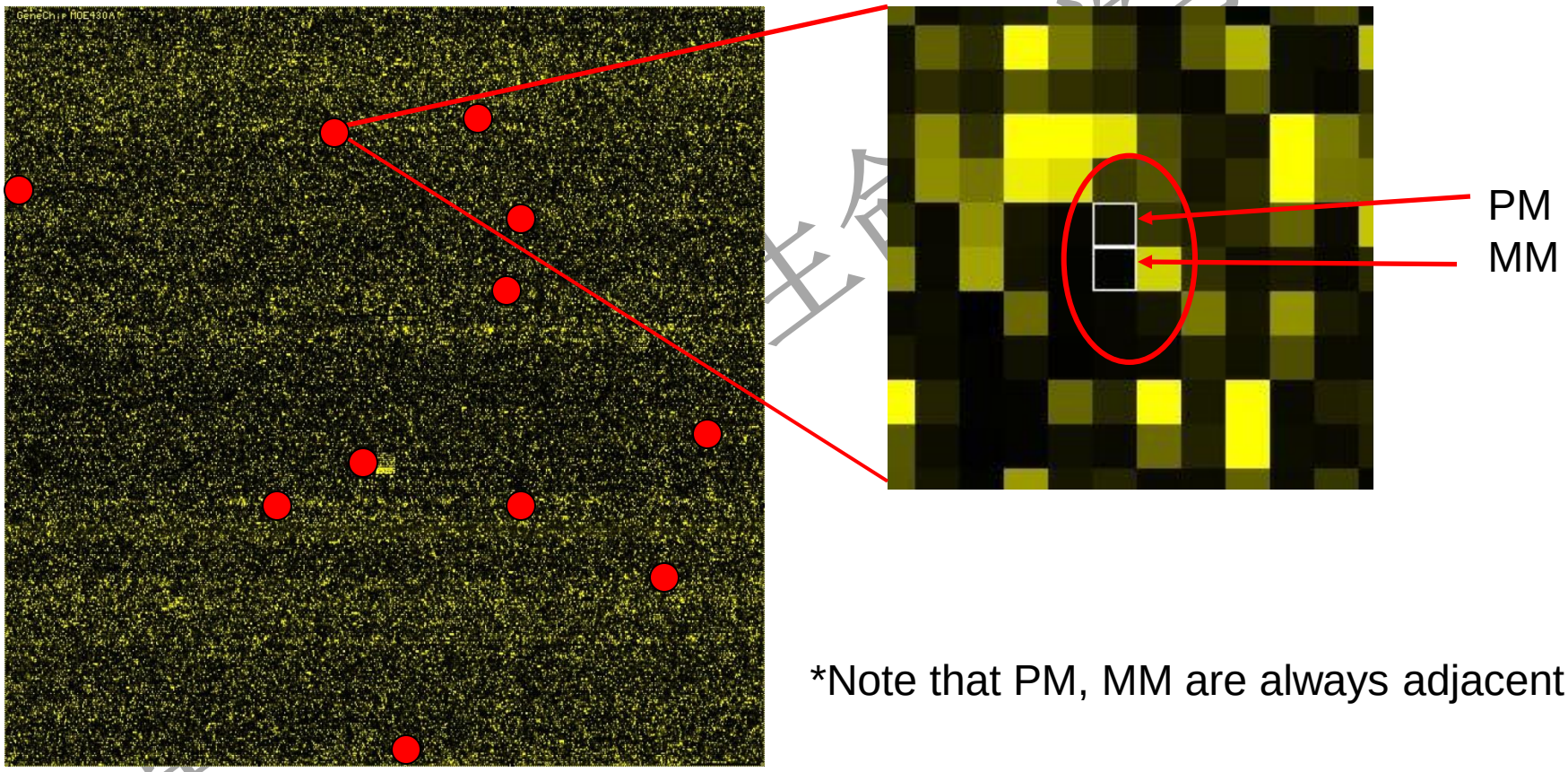






DNA microarray *vs* RNA-seq

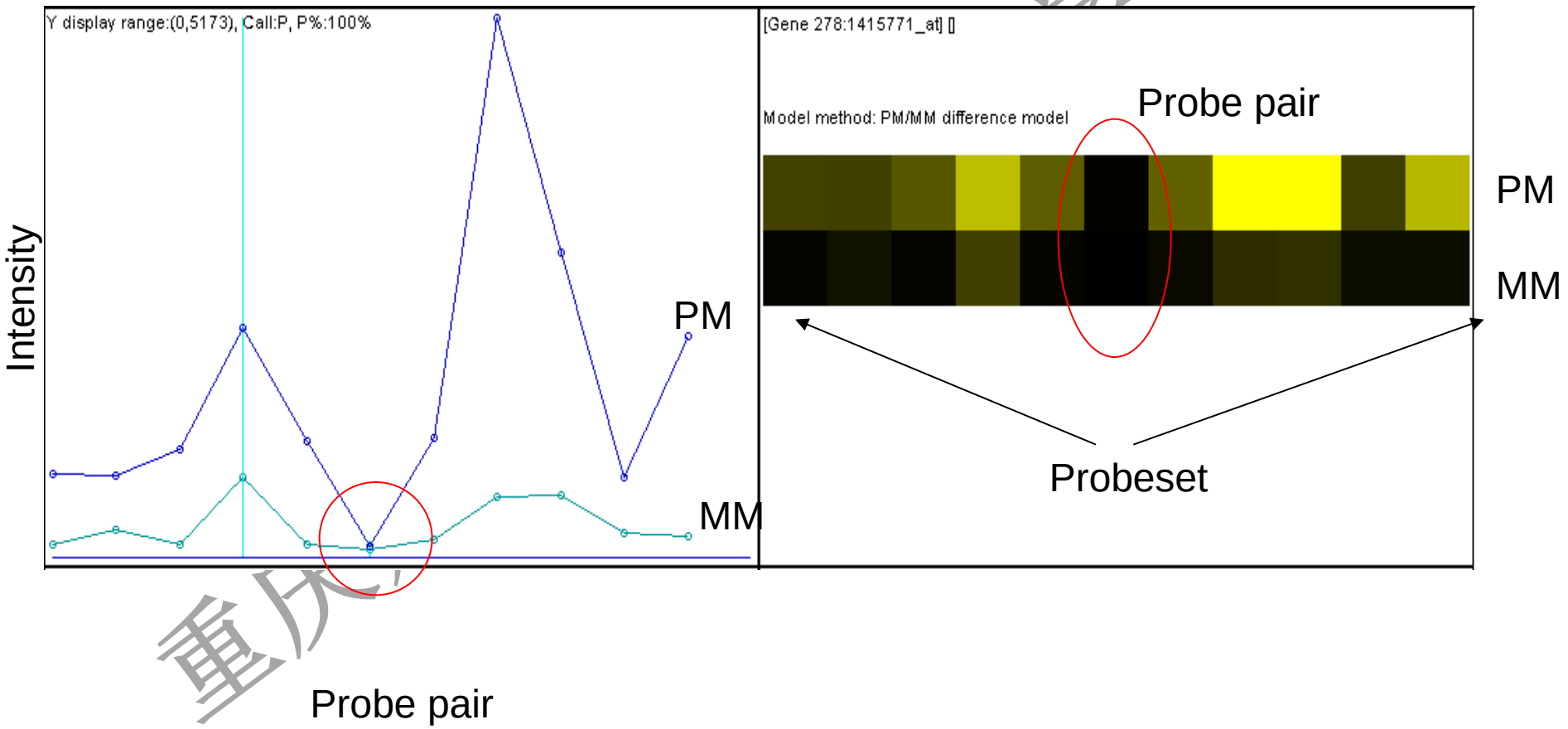
1415771_at on MOE430A



*Note that PM, MM are always adjacent

*image created using dChip software

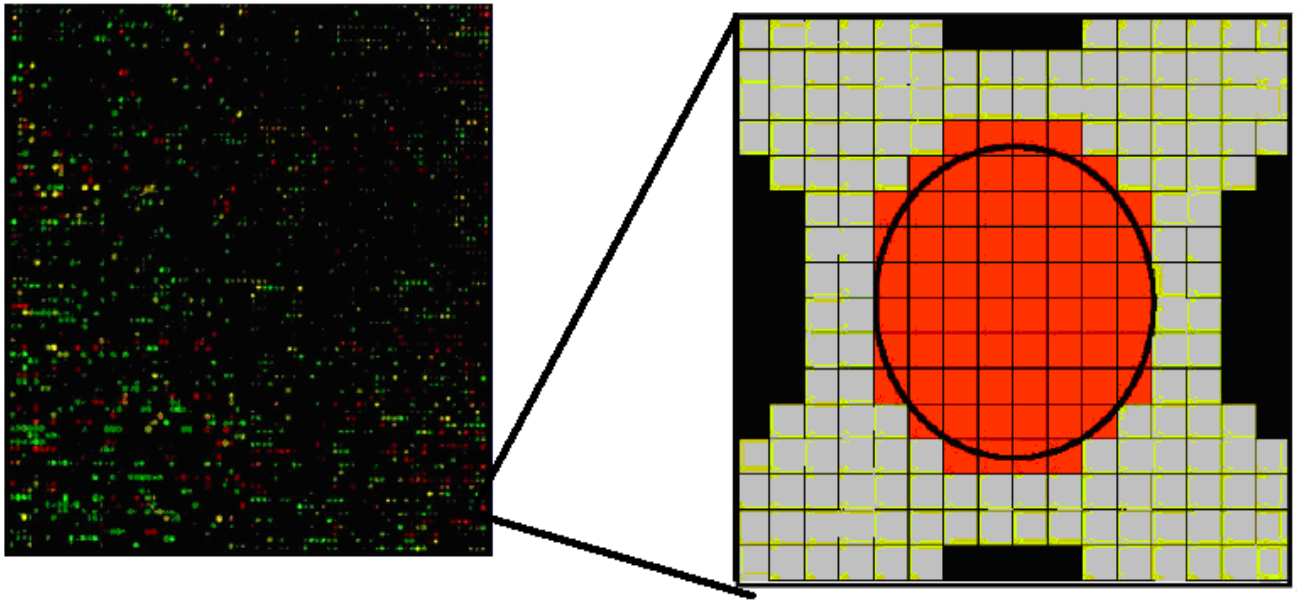
1415771_at on MOE430A



*images created using dChip software

一、基因芯片数据提取

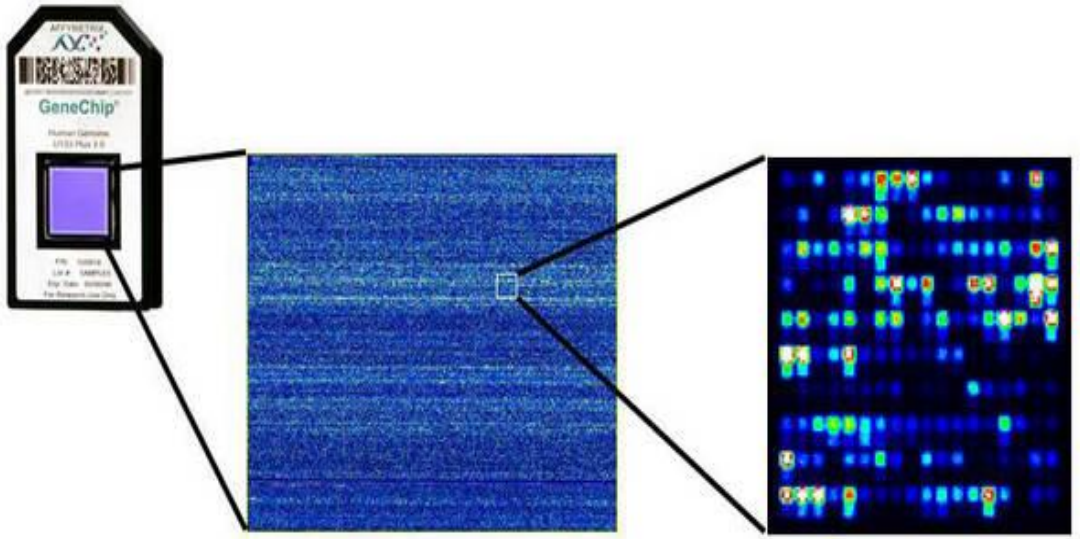
以cDNA微阵列芯片为例



$$Ratio = (CH1I - CH1B)/(CH2I - CH2B)$$

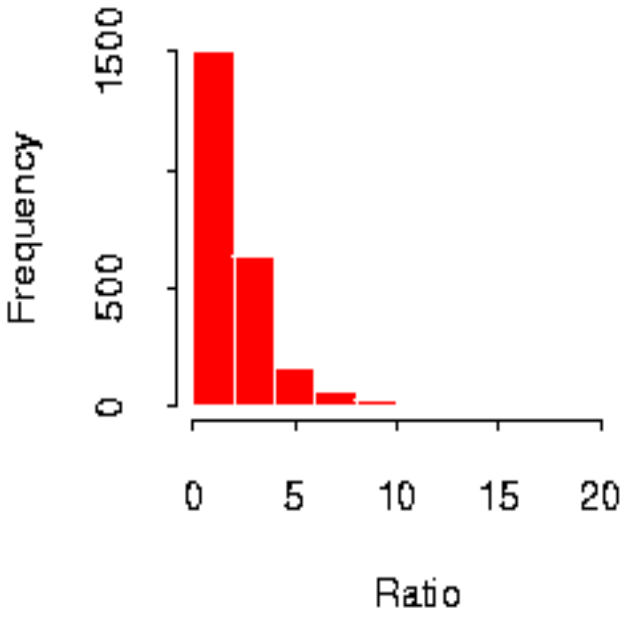
定性信息提取： P/A/M(Present/Absent/Marginal）

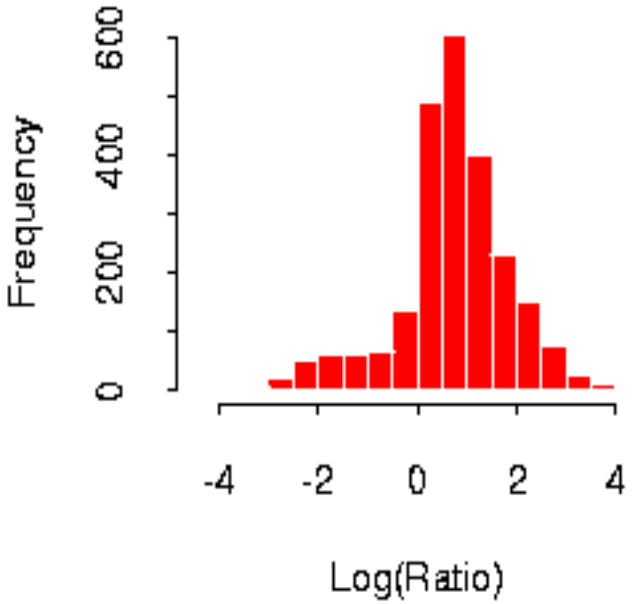
定量信息提取： 基于探针集汇总后的基因水平荧光信号强度值



Probe	Symbol	Sample 1	Sample 2	Sample 3	Sample 4
Probe A	Gene A	5614	6446	5756	5498
Probe B	Gene B	592	401	459	619
Probe C	Gene C	246	238	261	207
Probe D	Gene D	1233	813	647	663

二、对数转换





对芯片数据做对数化转换后，数据可近似正态分布

三、数据过滤

- 数据过滤：去除表达水平是负值或很小的数据或者明显的噪声数据。

1. 过闪耀现象
2. 物理因素导致的信号污染
3. 杂交效能低
4. 点样问题
5. 其他

四、补缺失值

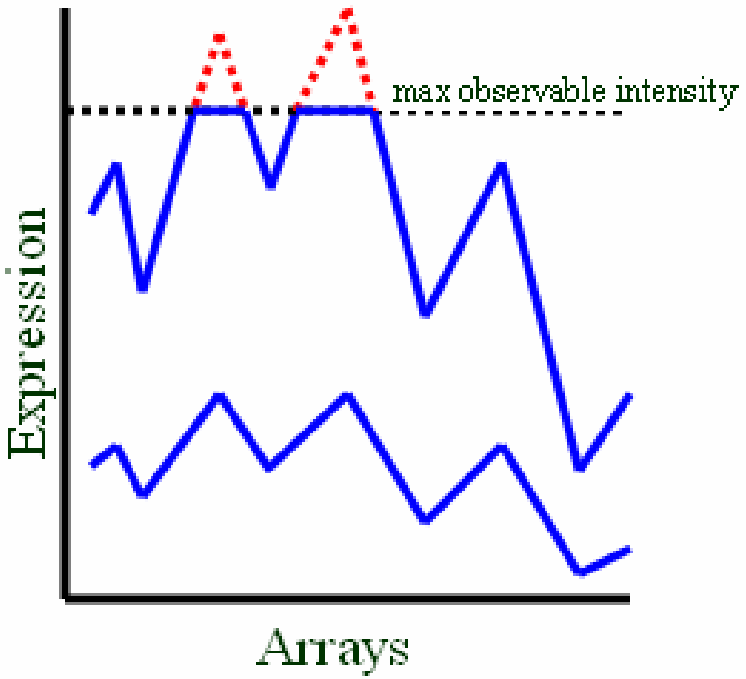
(一)数据缺失类型

- 非随机缺失

  基因表达丰度过高或过低

- 随机缺失

  与基因表达丰度无关，数据补缺主要针对随机缺失情况



(二)数据补缺方法

1. 简单补缺法

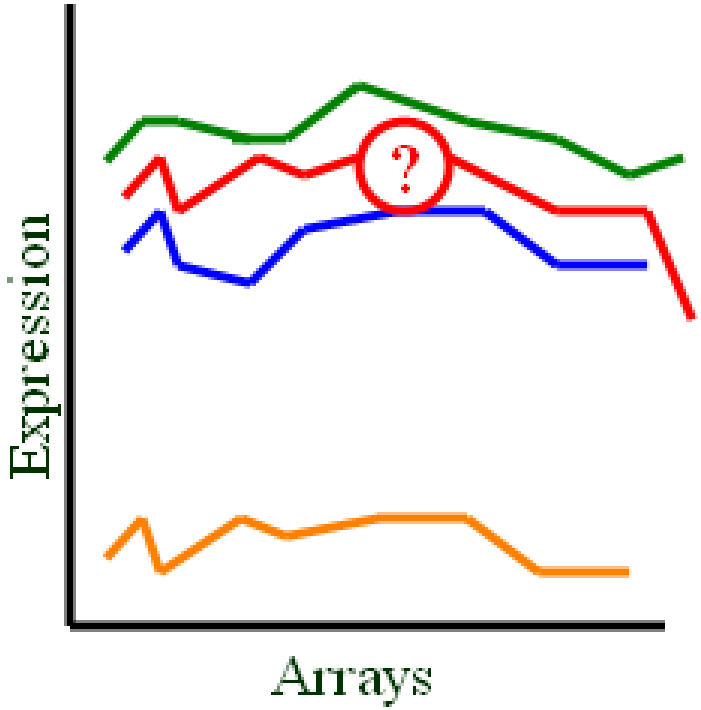
- **missing values = 0 expression**
- **missing values = 1 expression (arbitrary signal)**
- **missing values = row (gene) average**
- **missing values = column (array) average**

2. *k*近邻法

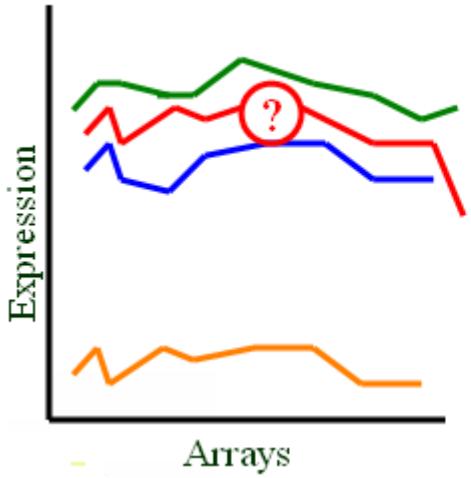
- 选择与具有缺失值基因的*k*个邻居基因
- 用邻居基因的加权平均估计缺失值

参数：

- 邻居个数
- 距离函数



3. 回归法



$$y_1 = a_1 + b_1 x_1$$

$$y_2 = a_2 + b_2 x_2$$

$$y = w_1 y_1 + w_2 y_2$$

4. 其他方法

五、数据标准化

(一)为什么要进行数据标准化

存在不同来源的系统误差

1. 染料物理特性差异(热光敏感性，半衰期等)
2. 染料的结合效率
3. 点样针差异
4. 数据收集过程中的扫描设施
5. 不同芯片间的差异
6. 实验条件差异

(二)运用哪些基因进行标准化处理

- 芯片上大部分基因(假设芯片上大部分基因在不同条件下表达量相同)
- 不同条件间稳定表达的基因(如持家基因)
- 控制序列(spiked control)

  在不同条件下表达水平相同的合成DNA序列或外源的DNA序列。

Intensity to Expression

- Now we have thousands of intensity values associated with probes, grouped into probesets.
- How do you transform intensity to expression values?
  - Algorithms
    - ***MAS5***
      - Affymetrix proprietary method
    - ***RMA/GCRMA***
      - Irizarry, Bolstad
    - ..many others
- Often called "**normalization**"

(三) cDNA芯片数据标准化处理

1. 片内标化(within-slide normalization)

(1) 全局标化(global normalization)

- 假设： $R=k*G$
- 方法：

$$\log_2 R/G \rightarrow \log_2 R/G - c = \log_2 R/(kG).$$

- $c=\log_2 k$：中值或均值

(2) 荧光强度依赖的标化(intensity dependent normalization)

- 为什么
- 方法：scatter-plot smoother lowess拟合

$$\log_2 R/G \rightarrow \log_2 R/G - c(A) = \log_2 R/(k(A)G),$$

- c(A) 为M 对A 的拟合函数
- 标化后的数据

(3) 点样针依赖的标化(within-print-tip-group normalization)

- 为什么

  一张芯片的不同区域运用不同的点样针点样，从而引入点样针带来的系统误差。

- method

$$\log_2 R/G \rightarrow \log_2 R/G - c_i(A) = \log_2 R/(k_i(A)G),$$

(4) 尺度调整(scale adjustment)

- 为什么

  调整不同栅格(grids)间的数据离散度

- 方法：计算不同栅格的尺度因子

$$\hat{a}_i = \frac{MAD_i}{\sqrt[I]{\prod_{i=1}^{I} MAD_i}},$$

$$MAD_i = \text{median}_j\{|M_{ij} - \text{median}_j(M_{ij})|\}.$$

2. 片间标化(multiple-slide normalization)

- 线性标化法(linear scaling methods)

  与芯片内标化的尺度调整(scale adjustment)方法类似
- 非线性标化法(non-linear methods)
- 分位数标化法(quantile normalization)

  两张芯片的表达数据的分位数标化至相同，即分布于对角线上

3. 染色互换实验(dye-swap experiment)的标化

	实验组	对照组
芯片1	cy5(R)	cy3(G′)
芯片2	cy3(G)	cy5(R′)

- 前提假设：$c \approx c'$
- 方法：

$$\log_2 R/G - c \approx -(\log_2 R'/G' - c').$$

$$c \approx \frac{1}{2}\left[\log_2 R/G + \log_2 R'/G'\right] = \frac{1}{2}(M + M').$$

(四) 芯片数据标准化

1. 提取定性信号

(1)对每个探针对计算R

$$R = (PM - MM) / (PM + MM)$$

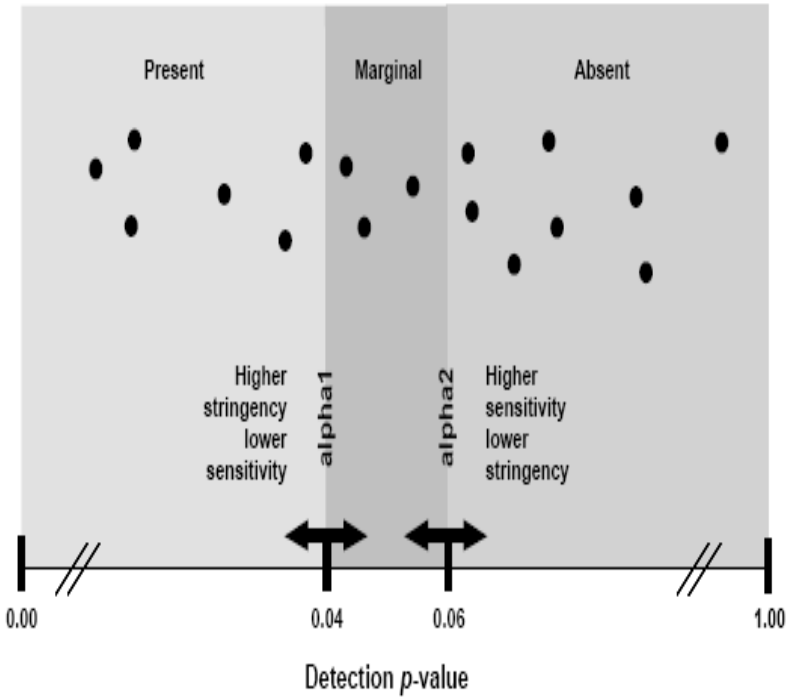
(2)比较R与定义的阈值Tau(小的正值，默认值为0.015).

(3) 单侧的Wilcoxon's Signed Rank test产生*p*值，根据*p*值定义定量信号值

Present call

Marginal call

Absent call



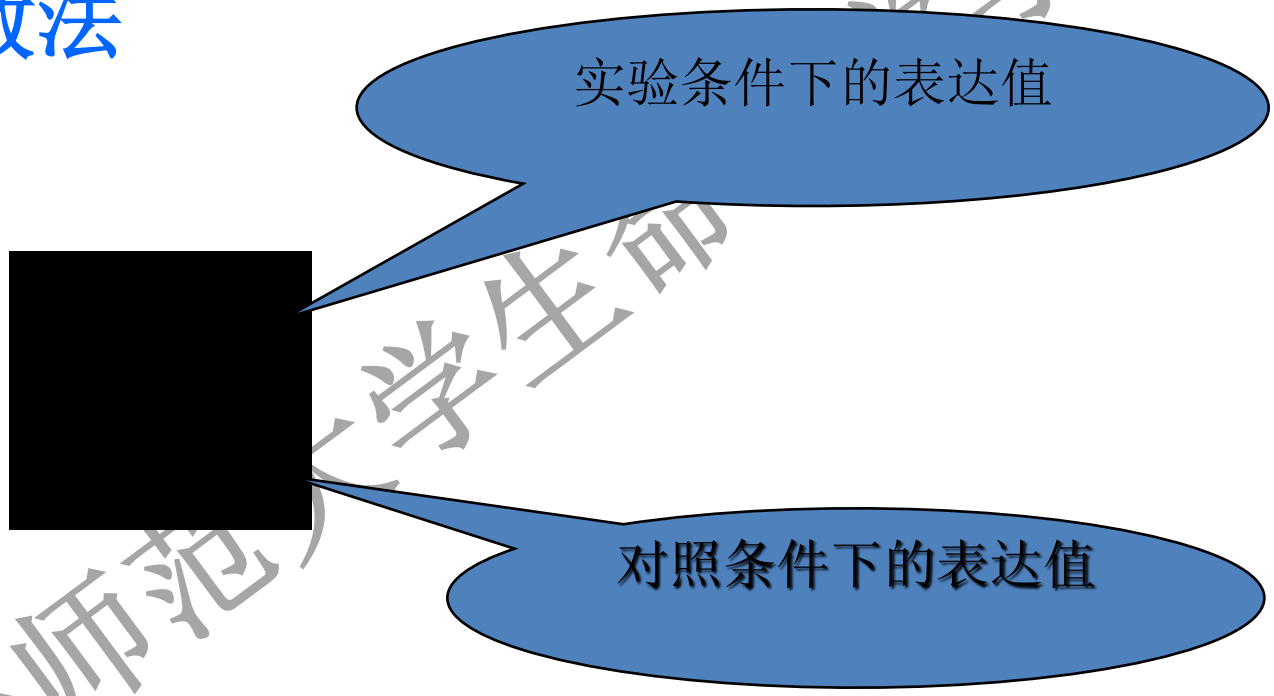
2. 提取定量信号

(1)分析步骤

- 获取探针水平数据
- 背景值效正
- 标准化处理
- 探针特异背景值效正
- 探针集信号的汇总

第四节：基因差异表达分析

一、倍数法



通常以2倍差异为阈值，判断基因是否差异表达

二、t检验法

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{{s_1}^2 / n_1 + {s_2}^2 / n_2}}$$

运用 *t*检验法可以判断基因在两不同条件下的表达差异是否具有显著性

三、方差分析

方差分析可用于基因在两种或多种条件间的表达量的比较，它将基因在样本之间的总变异分解为组间变异和组内变异两部分。通过方差分析的假设检验判断组间变异是否存在，如果存在则表明基因在不同条件下的表达有差异。

四、SAM (significance analysis of microarrays)

(一) 多重假设检验问题

- Ⅰ型错误（假阳性）即在假设检验作推断结论时，拒绝了实际上正确的检验假设，即将无差异表达的基因判断为差异表达。
- Ⅱ型错误（假阴性）即不拒绝实际上不正确的，即将有差异表达的基因判断为无差异表达。
- 在进行差异基因挑选时，整个差异基因筛选过程需要做成千上万次假设检验，导致假阳性率的累积增大。对于这种多重假设检验带来的放大的假阳性率，需要进行纠正。常用的纠正策略有Bonferroni效正，控制FDR（false discovery rate）值等。

(二) 分析步骤

- 计算统计量

$$d = \frac{\overline{x_1} - \overline{x_2}}{s + s_0}$$

- 扰动实验条件，计算扰动后的基因表达的相对差异统计量

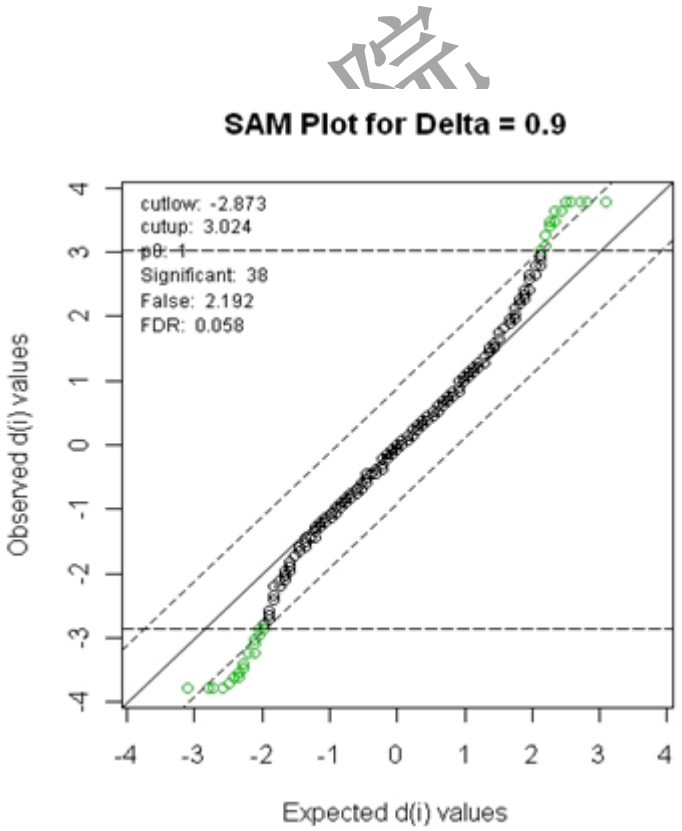
$$d_p$$

- 计算扰动后的平均相对差异统计量

$$d_E = \frac{1}{|P|}\sum d_p$$

- 确定差异表达基因阈值：以最小的正值和最大的负值作为统计阈值，运用该阈值，统计在值中超过该阈值的假阳性基因个数，估计假阳性发现率FDR值。

- 通过调整FDR值的大小得到差异表达基因。



五、信息熵

运用信息熵进行差异基因挑选时，不需要用到样本的类别信息，所以运用信息熵找到的差异基因是指在所有条件下表达波动比较大的基因。

$$H = -\sum_{i=1}^{m} p_i \log p_i$$

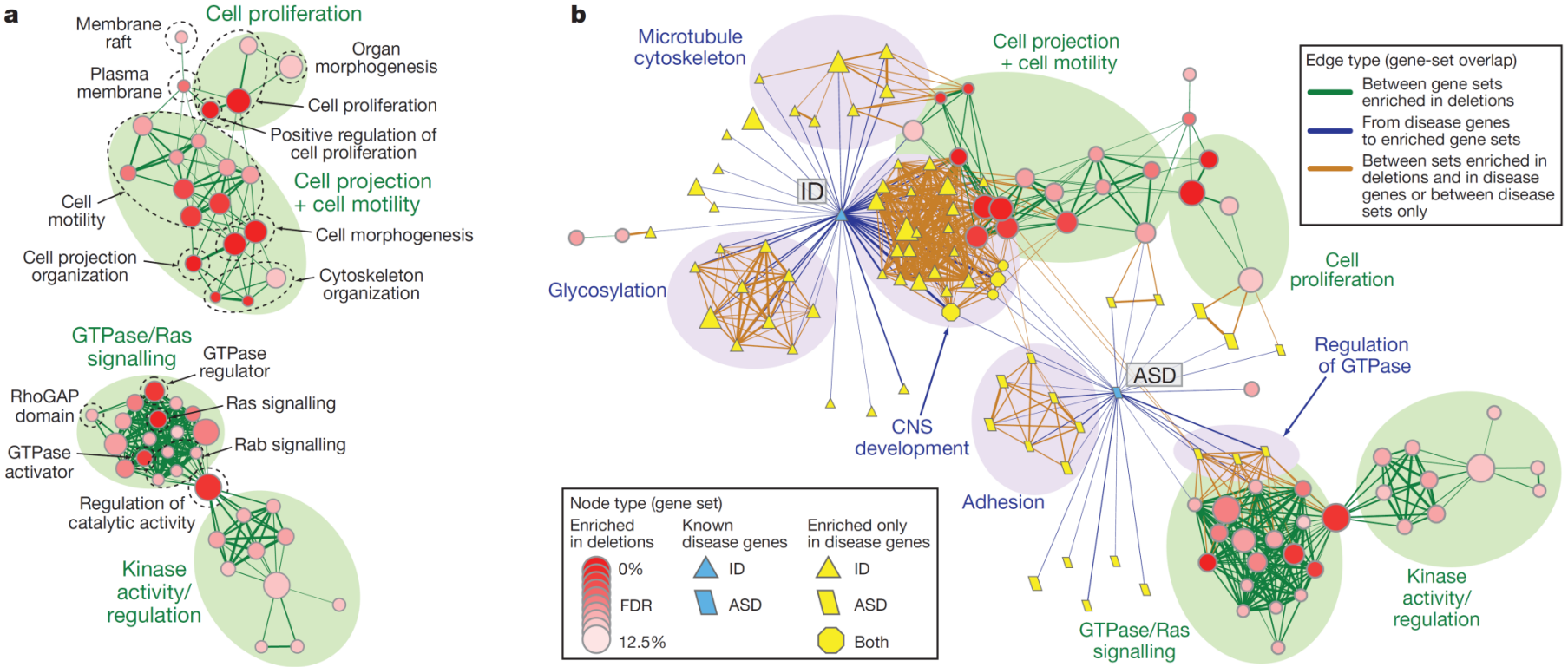


Figure 3 | A functional map of ASD. Enrichment results were mapped as a network of gene sets (nodes) related by mutual overlap (edges), where the colour (red, blue or yellow) indicates the class of gene set. Node size is proportional to the total number of genes in each set and edge thickness represents the number of overlapping genes between sets. **a**, Gene sets enriched for deletions are shown (red) with enrichment significance (FDR *q*-value) represented as a node colour gradient. Groups of functionally related gene sets are circled and labelled (groups, filled green circles; subgroups, dashed line). **b**, An expanded enrichment map shows the relationship between gene sets enriched in deletions (**a**) and sets of known ASD/intellectual disability genes. Node colour hue represents the class of gene set (that is, enriched in deletions, red; known disease genes (ASD and/or intellectual disability (ID) genes), blue; enriched only in disease genes, yellow). Edge colour represents the overlap between gene sets enriched in deletions (green), from disease genes to enriched sets (blue), and between sets enriched in deletions and in disease genes or between disease gene-sets only (orange). The major functional groups are highlighted by filled circles (enriched in deletions, green; enriched in ASD/intellectual disability, blue).

Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368 (2010).

How many replicates?

3 or more Biological Replicates is a minimum!

Biological Replicates

- Recreate the experiment several times. This gives a sense of biological variability.

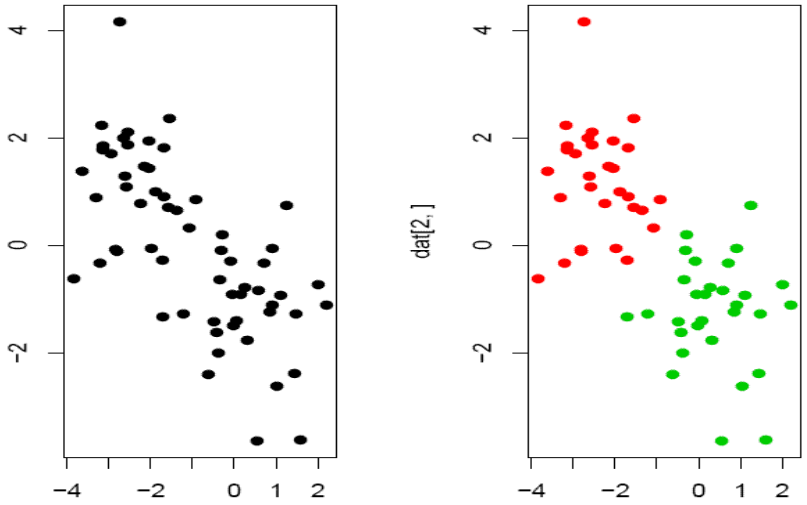
Technical Replicates

- Don't bother unless you're doing a technical study of microarray variability.

第5节：基因表达数据的聚类分析

一、聚类目的

基于物体的相似性将物体分成不同的组



二、基因表达谱数据的聚类

- 对基因进行聚类

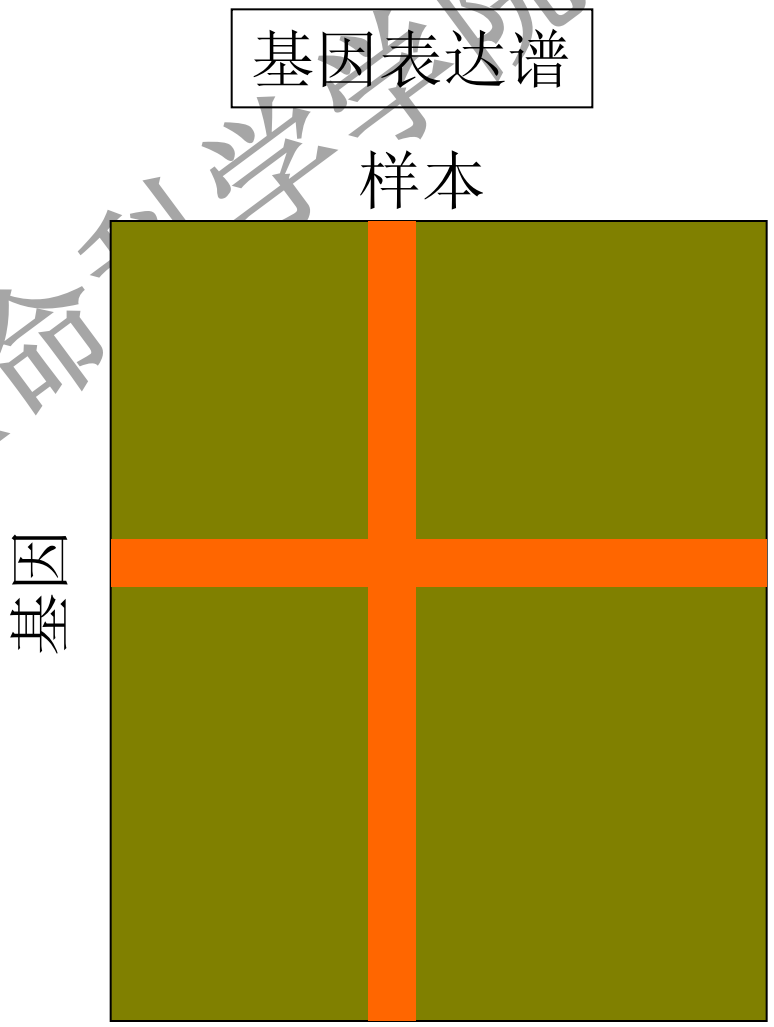
  识别功能相关的基因

  识别基因共表达模式

- 对样本进行聚类

  质量控制

  检查样本是否按已知类别分组发现亚型



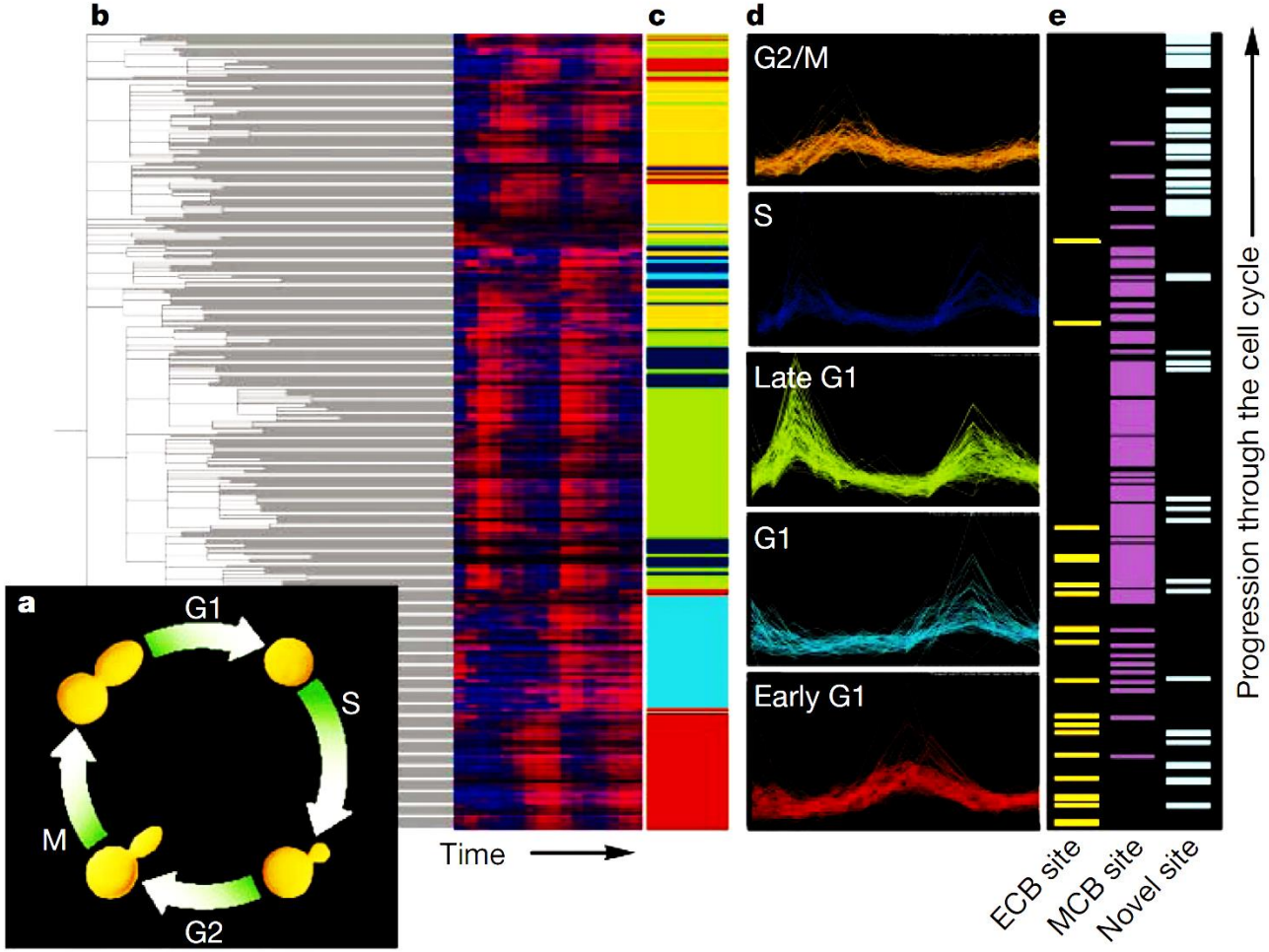


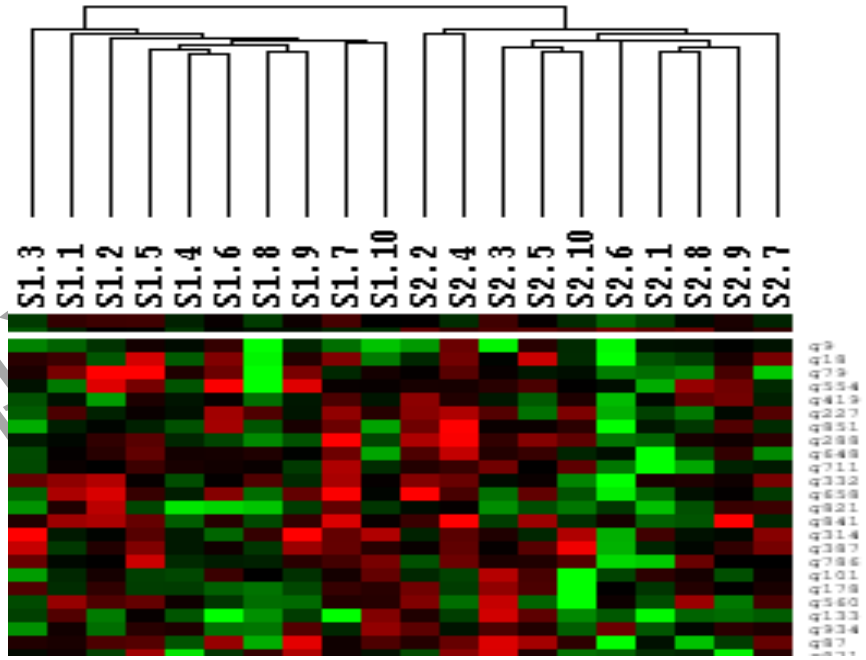
Figure 3 Methods for analysing gene expression data shown for measurements of expression in the cell cycle of *S. cerevisiae*. **a**, Yeast cells were synchronized and cells were collected every ten minutes throughout two complete synchronous cycles (18 time points in total are shown). Expression data were collected by hybridizing labelled cDNA samples to high-density oligonucleotide arrays. Transcript levels were determined for almost every gene in the genome for every time point[24]. A sample of 409 genes (from a total of 6,000) that showed both a significant (more than twofold) fluctuation in transcript levels during the time course and cell cycle-dependent periodicity were selected for further analysis. **b**, Dendrogram indicating similarity of expression profiles, calculated using the Pearson correlation function in the GeneSpring software package (Silicon Genetics, San Carlos, CA). For display purposes, the relative expression levels were plotted in red (high) and blue (low). **c**, The genes were divided into five different temporal expression classes (red, early G1; light blue, G1; green, late G1; dark blue, S; orange, G2/M) using K-tuple means clustering (also using GeneSpring software) and the clusters were named according to their time of peak expression within the cell cycle. **d**, Line graphs for all genes in the clusters defined in **b**. **e**, Location of cell cycle-regulated genes within the dendrogram in **a** that have *cis*-regulatory sequence elements in the 500 bp upstream of their promoter. Column 1, MCB sites (ACGCGT); column 2, ECB sites (TTWCCCNNNNAGGAA); column 3, a new sequence (GTAAACAA or TTGTTTAC) was identified that was statistically associated ($p = 1.77 \times 10^{-7}$ for the forward direction, $p = 0.003$ for the reverse) with the promoter regions of genes whose expression peaked in G2/M phase.

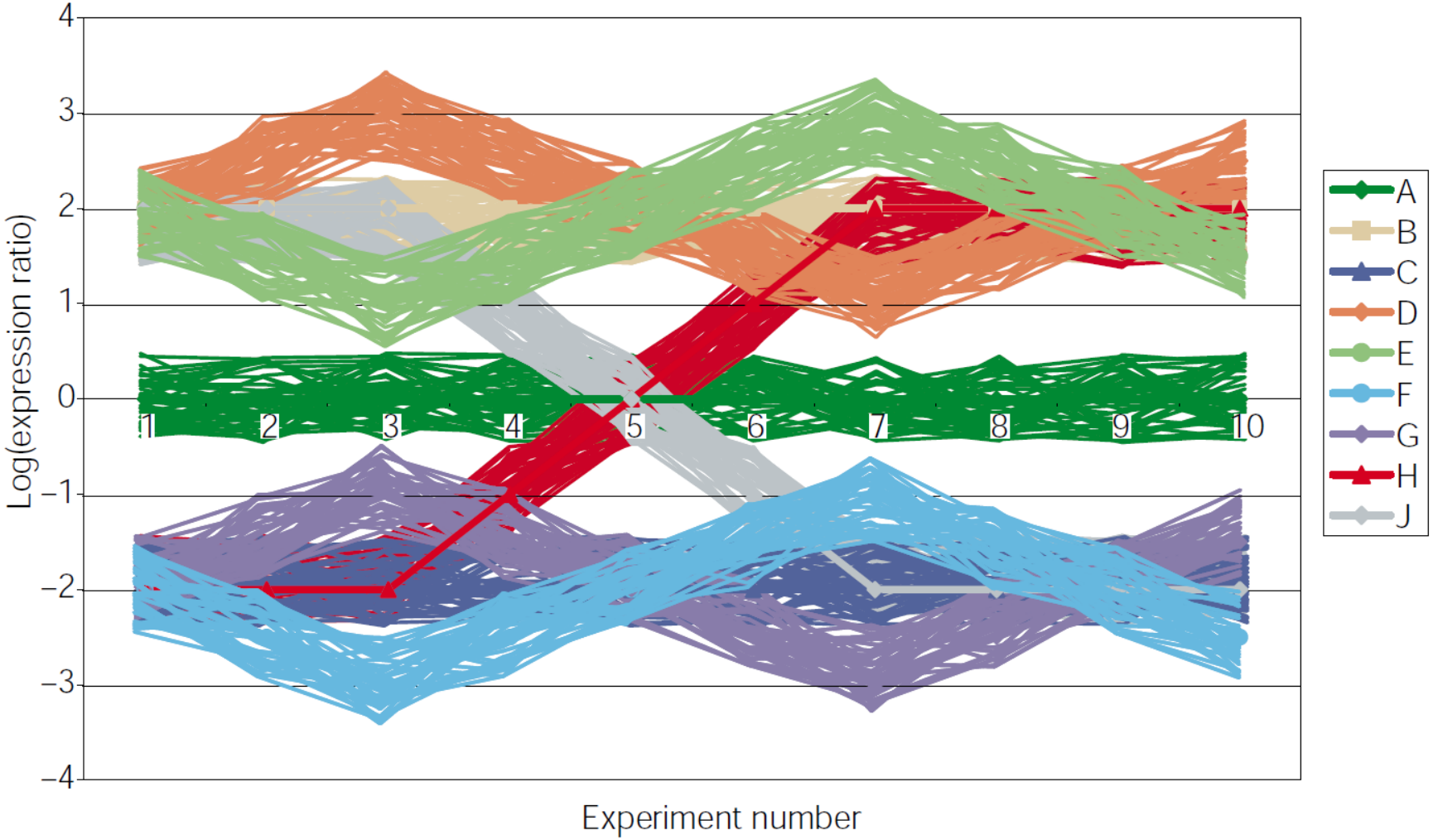
三、距离尺度函数

- 几何距离
- 线性相关系数
- 非线性相关系数
- 互信息
- 其他

四、聚类算法
(一)层次聚类

- 层次聚类算法将研究对象按照它们的相似性关系用树形图进行呈现，进行层次聚类时不需要预先设定类别个数，树状的聚类结构可以展示嵌套式的类别关系。

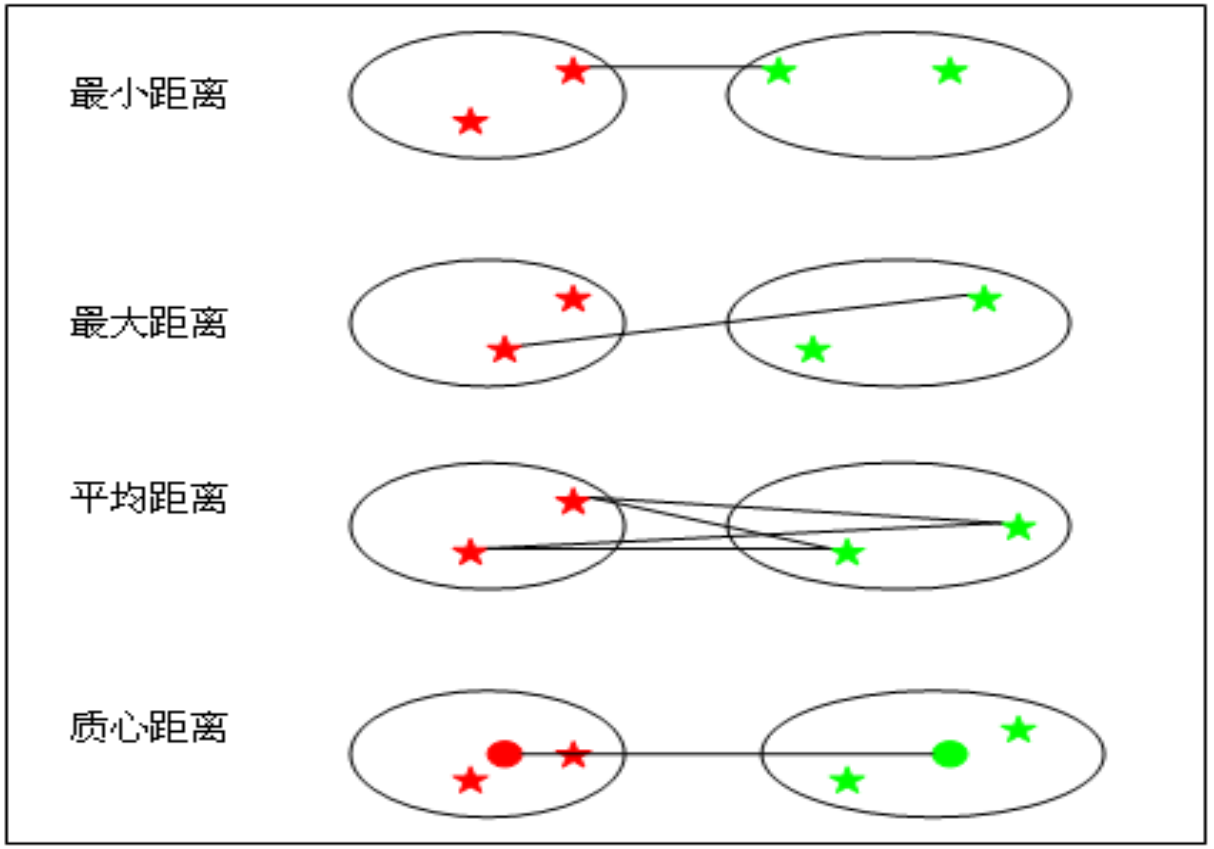




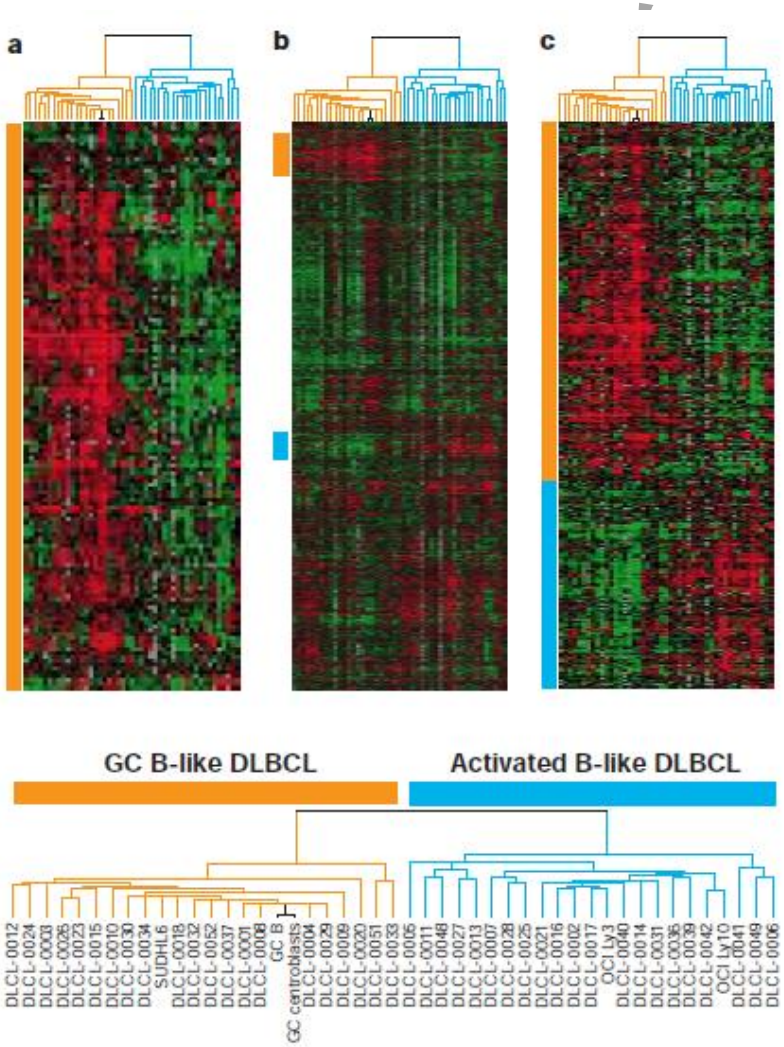
基因共表达模式鉴定示意图

Quackenbush, J. Computational analysis of microarray data. *Nat Rev Genet* **2**, 418-27 (2001).

- 在对含非单独对象的类进行合并或分裂时，常用的类间度量方法

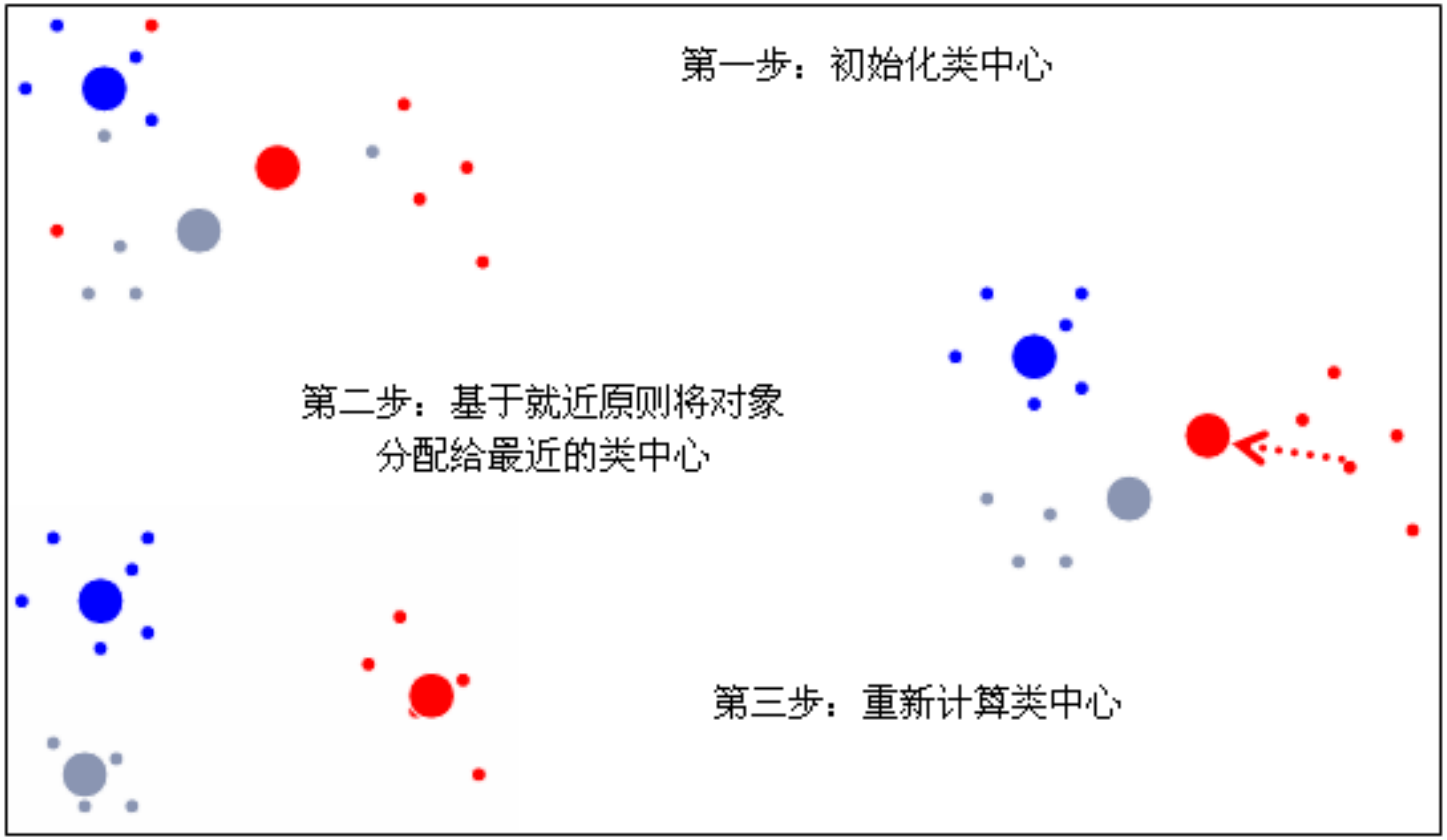


- 2000年Alizadeh等运用基因芯片数据，基于层次聚类算法证实了DLBCL肿瘤病人在mRNA层面确实存在两种亚型



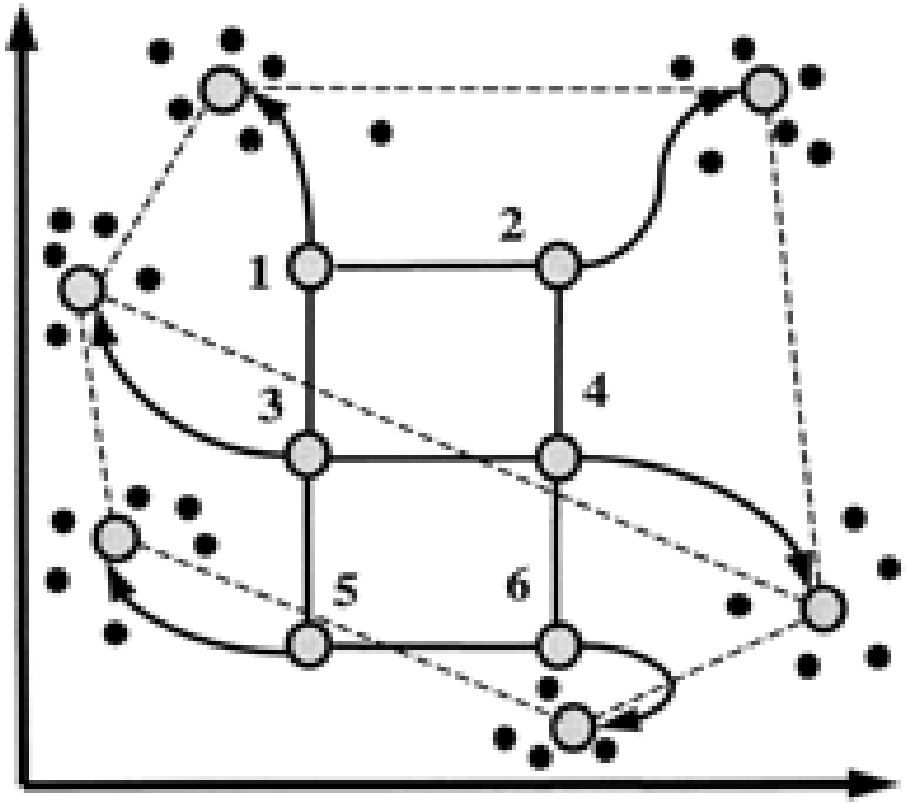
(二)k均值聚类

基本思想



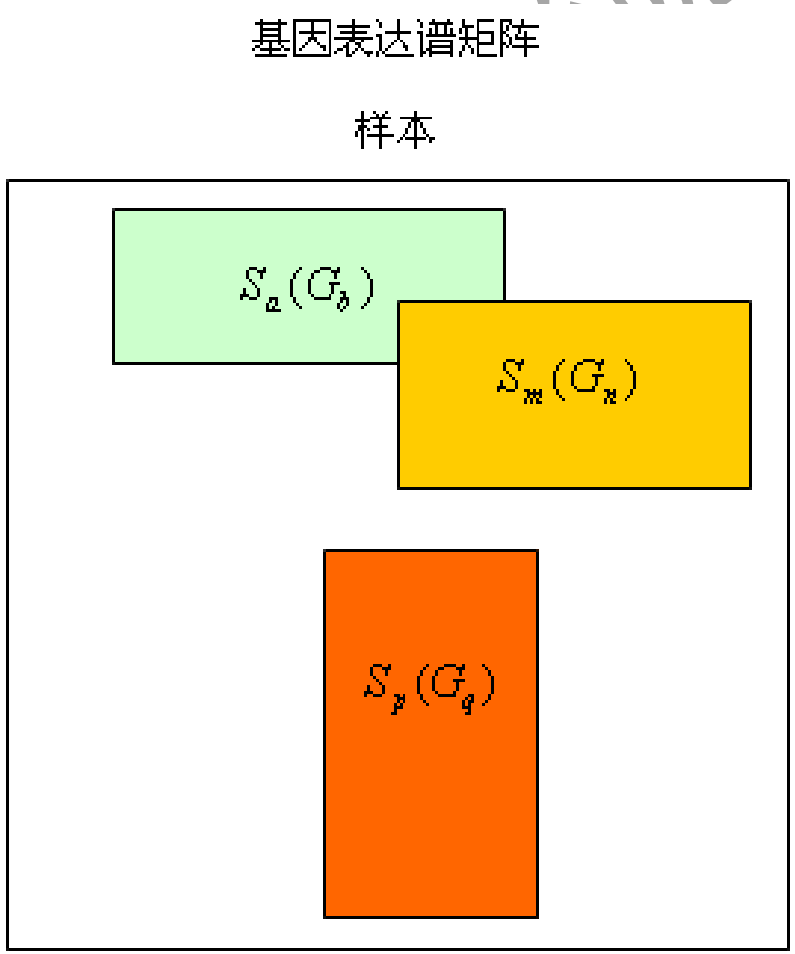
(三)自组织映射聚类

基本思想：在不断的学习过程中，输出层的神经元根据输入样本的特点进行权重调整，最后拓朴结构发生了改变



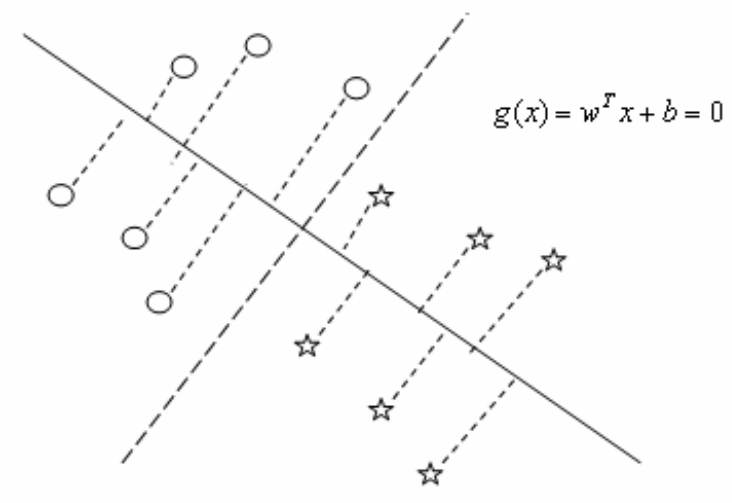
(四)双向聚类

双向聚类就是识别基因表达谱矩阵中同质的子矩阵，运用特定的基因子类识别样本子类。



第6节：基因表达数据的分类分析

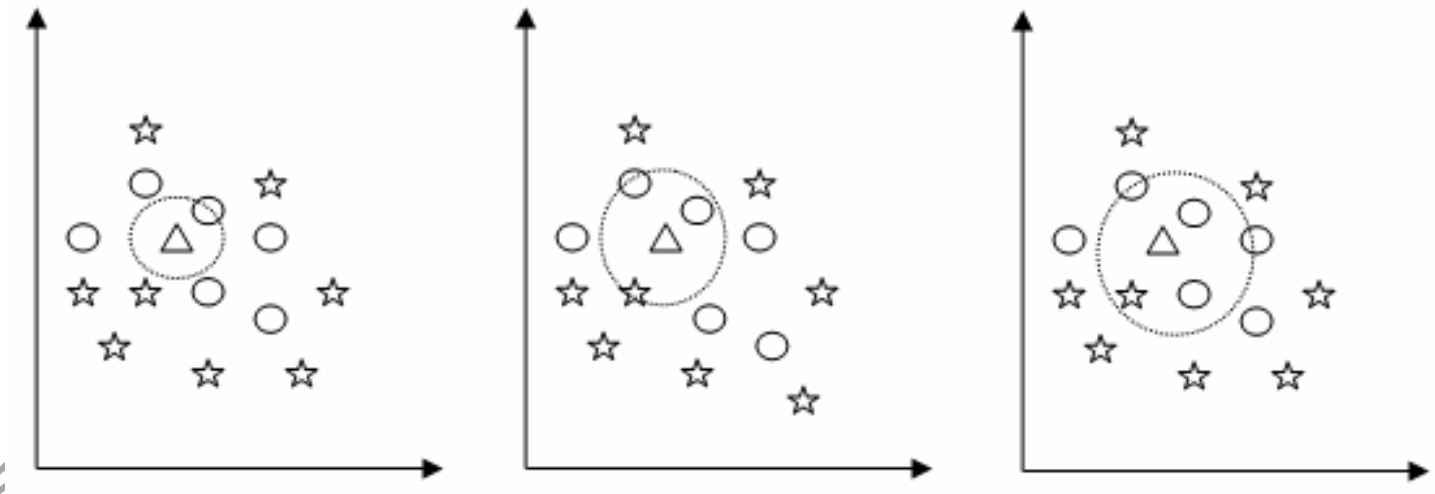
一、线性判别分类器



$$g(x)=w^T x+b\begin{cases}>0, L_1\\<0, L_2\end{cases}$$

二、k 近邻分类法

基本思想

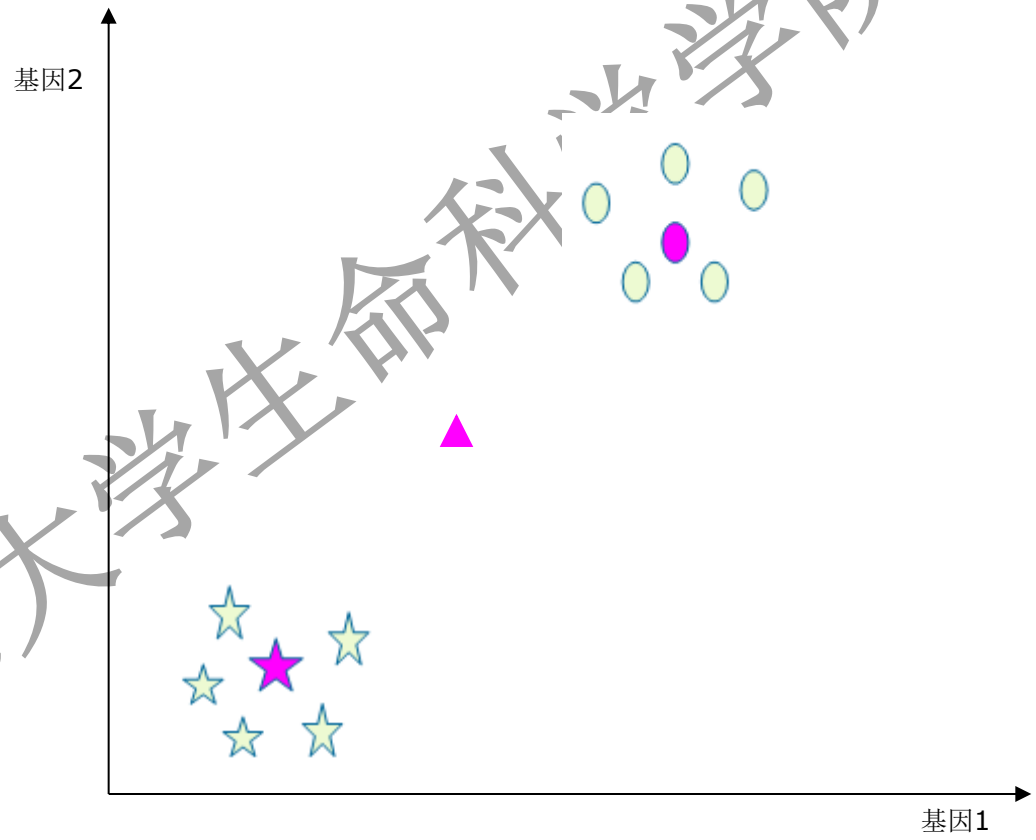


三、PAM分类法

Prediction Analysis for Microarray

基本思想

每类样本的质心向所有样本的质心进行收缩，即收缩每个基因的类均值，收缩的数量由值决定。当收缩过程发生时，某些基因在不同类中将会有相同的类均值，这些基因就不具有类间的区别效能。



分析步骤

- 计算统计量

$$d_{ik}=\frac{\bar{x}_{ik}-\bar{x}_i}{m_k\cdot(s_i+s_0)}$$

- 对公式经过变换得到

$$\bar{x}_{ik}=\bar{x}_i+m_k(s_i+s_0)d_{ik}$$

- 收缩各类的均值

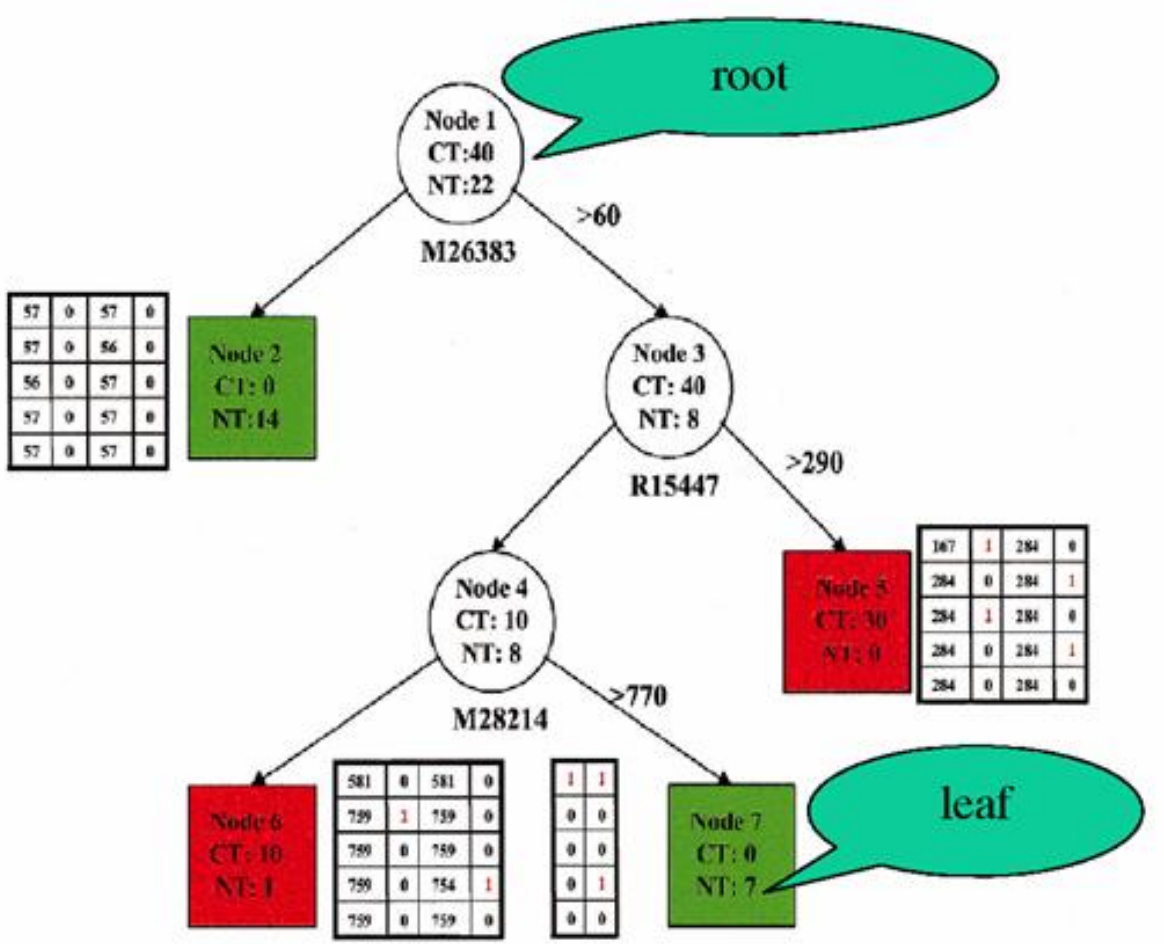
$$\bar{x}'_{ik}=\bar{x}_i+m_k(s_i+s_0)d'_{ik}$$

$$d'_{ik}=sign(d_{ik})(|d_{ik}|-\Delta)_+$$

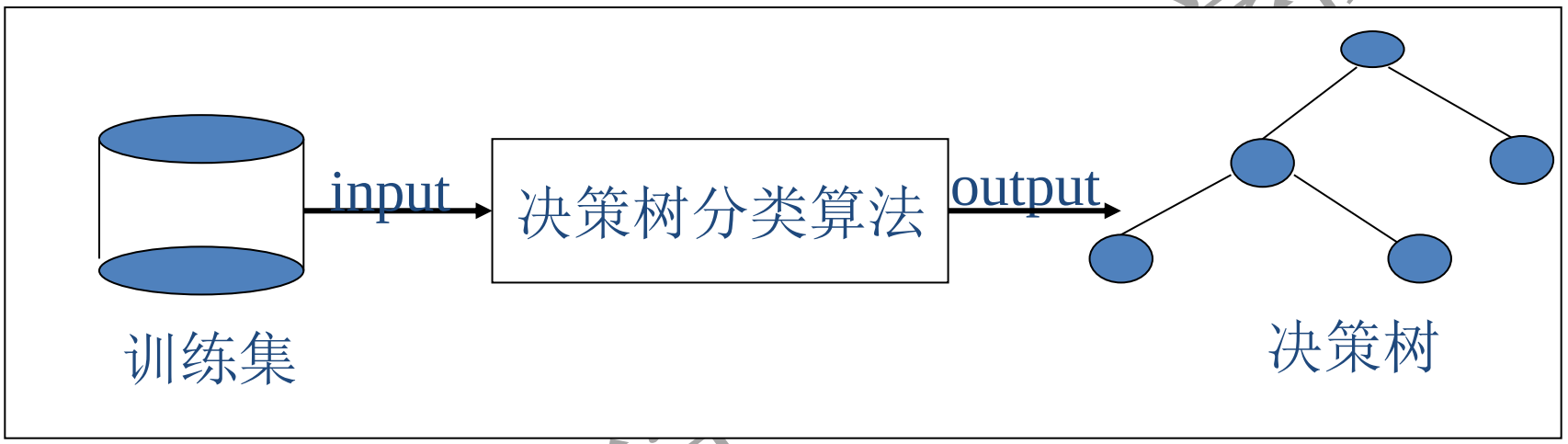
四、决策树

(一)基本思想

- 决策树又称为多级分类器，利用决策树分类可以把一个复杂的多类别分类问题转化为若干个简单的分类问题来解决
- 决策树的结构：一个树性的结构，内部节点上选用一个属性进行分割，每个分叉都是分割的一个部分，叶子节点表示一个分布



(二)分析步骤：提取分类规则，进行分类预测



- 在构造决策树的过程中最重要的一点是在每一个分割节点确定用哪个属性来分类(或分裂）
- 这就涉及到关于使用什么准则来衡量使用A属性比使用B属性更合理

(三)衡量准则

- 信息增益——information gain
- 基尼指数——Gini index

(四)决策树的修剪

- 消除决策树的过适应问题
- 消除训练集中的异常和噪声
- 所涉及的方法很多，比如先剪枝算法（print）与后剪枝（sprint 算法）等等

思考题

1. 如何通过表达谱数据去鉴定house-keeping gene?

2. 看家基因筛选过程中可能会遇到哪些困难？

重庆师范大学
CHONG QING NORMAL UNIVERSIYT

Thanks for your attention!

Acknowledgement

College of Life Sciences, Chongqing Normal University
2022, Chongqing of P. R. C