生物信息学

第十一章 转录组与转录调控分析（1）

转录组与转录调控

- 基因在多个层次上受到调控
- 转录组 (Transcriptome)
  - 细胞内所有RNA分子，包括mRNA, rRNA, tRNA和其他非编码 (non-coding) RNA
- 转录调控 (Transcriptional regulation)
  - 基因调控 (Gene regulation)

- 转录 (Transcription)
- 转录后 (Post transcription)：RNA稳定性

} “转录组”

- 翻译 (Translation)
- 翻译后 (Post translation)

基因表达 (Gene expression) 分析

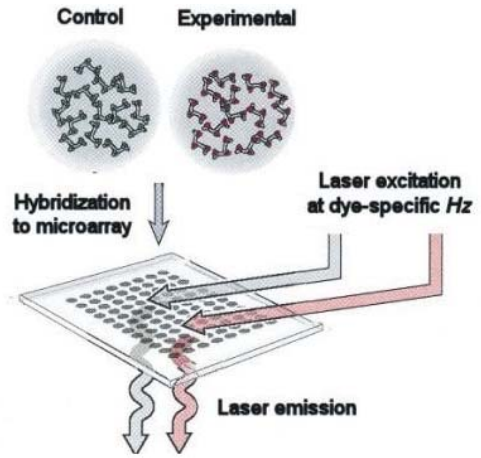
- 快照 (Snapshot)
  - 所有基因的RNA表达水平
  - 提供大量的数据
- 发现在特定生长时期，或者随着环境变化，哪些基因的表达上调或者下调
- 在相同条件下，上调或者下调变化规律相似的基因，可能具有功能上的关联
- 可以从共表达的基因中寻找调控模块
- 基因表达的模式可以用来表征异常的细胞调控
  - 癌症的诊断

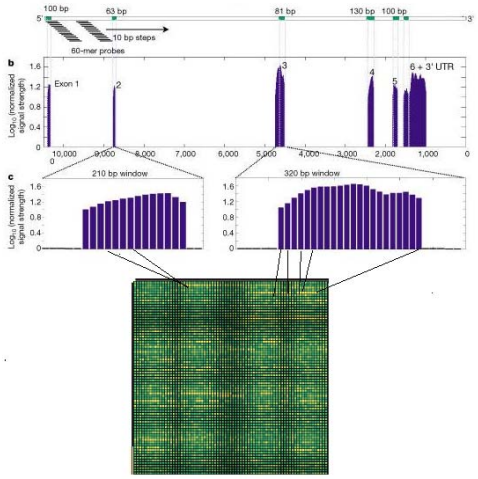
转录组与转录调控的主要研究技术

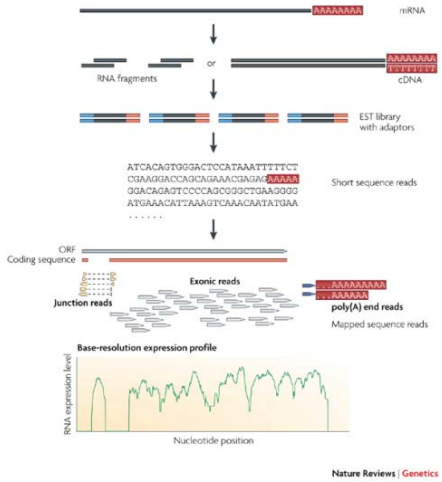
- 转录组
  - 单个实验中检测整个转录组
  - 基于DNA杂交：微阵列/基因芯片 (Microarray)
  - 基于高通量测序：RNA-Seq, Small RNA-Seq等
- 转录调控：
  - 蛋白质-DNA的相互作用关系
  - 芯片：ChIP-chip
  - 基于高通量测序：ChIP-Seq, Methyl-Seq等

转录组学研究技术的发展

基于DNA杂交技术







1995 Patrick O. Brown研究组发明cDNA芯片：检测已知基因的表达水平

2002 Affymetrix公司，发明“铺瓦芯片”(Tiling array)，发现新的基因、异构体并检测其表达

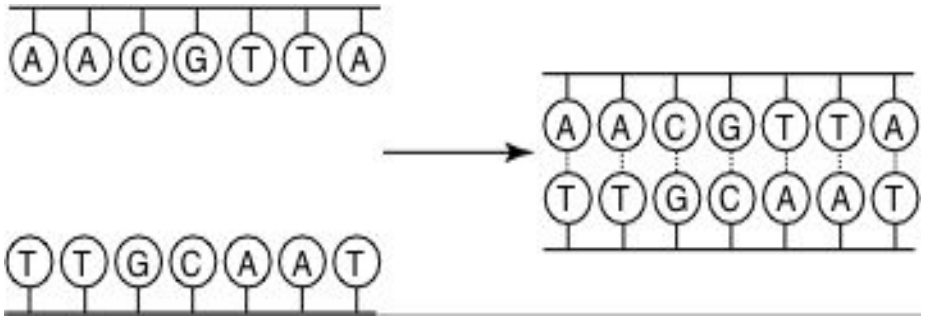
2008 mRNA-Seq技术的应用：利用下一代测序技术直接测定mRNA

RNA-Seq：仍在发展中的技术

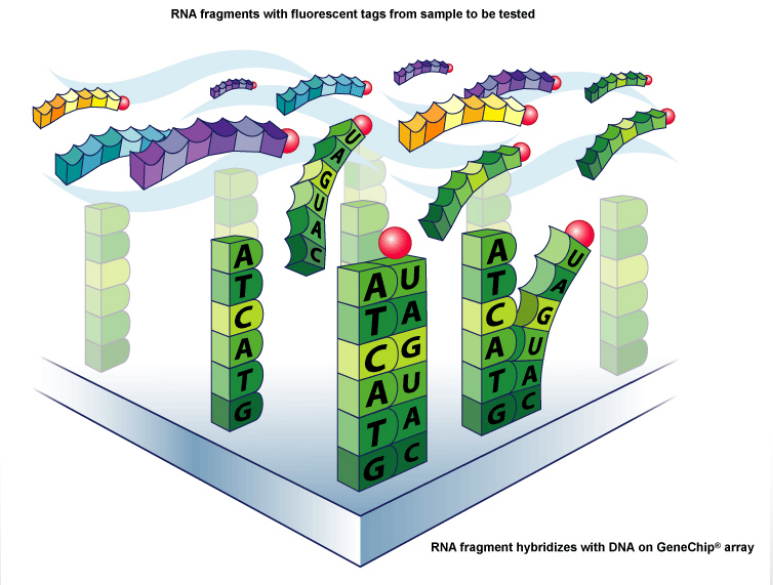
DNA芯片 (DNA microarray)

- 基因芯片 (1987)
  - 根据免疫测定的 (immunoassay)的方法予以改进
- 高通量、点阵以及Northern杂交 (1995)
  - 同时测定细胞内数千个基因的表达情况
  - 将mRNA反转录成cDNA与芯片上的探针杂交
  - 最终将被基于测序的方法取代
- 芯片的体积非常小：微量样品的检测
- 基因表达情况的定量分析
- 其他类型的芯片：
  - 组织芯片
  - 蛋白质芯片

基因芯片的密度：100-1 million DNA 探针/1cm²



碱基互补



将样品中的DNA/RNA标上荧光标记，可以定量检验基因的表达水平

基因芯片技术的类型

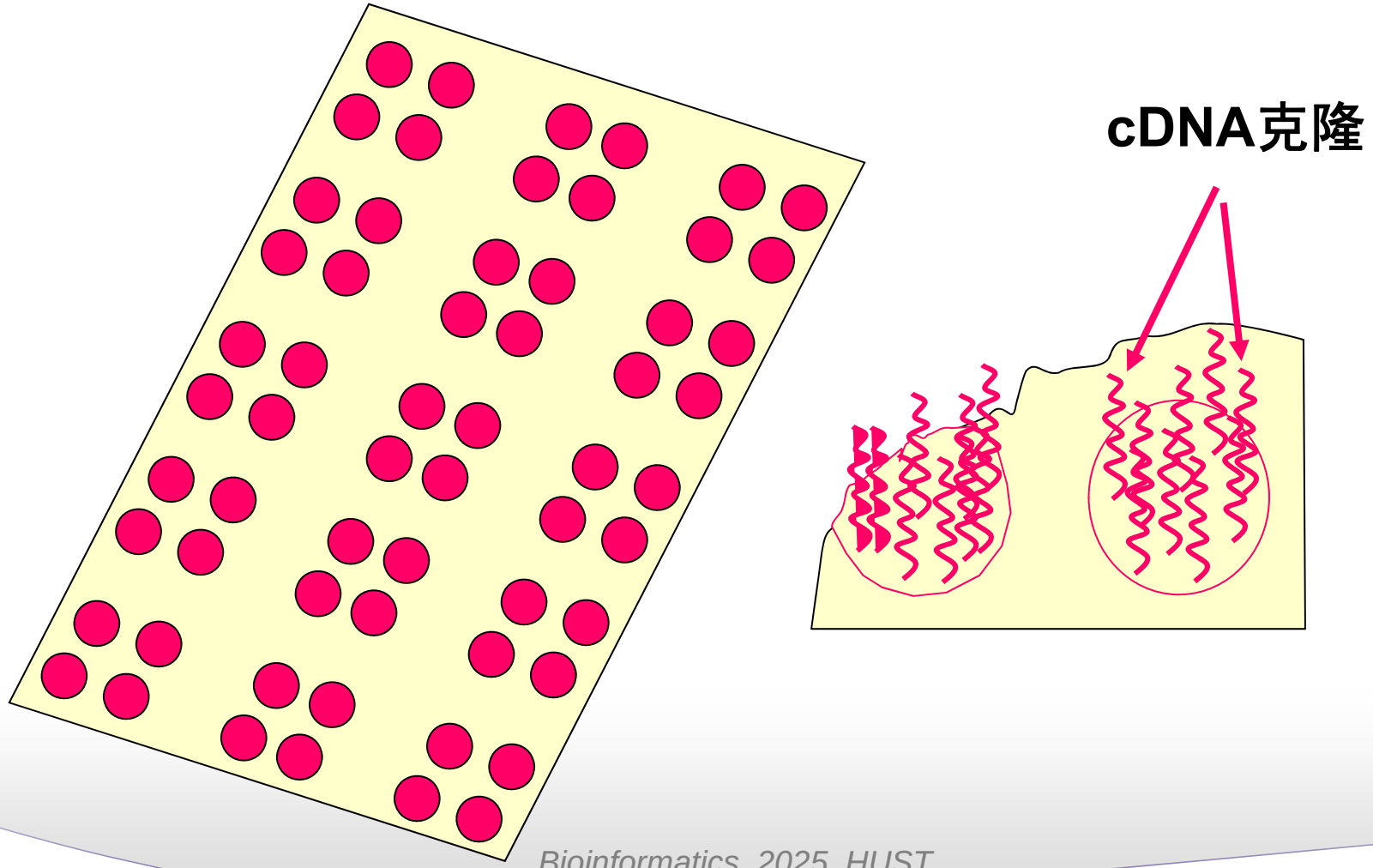
- 按技术手段、探针类型分类
  - Short oligonucleotide arrays (Affymetrix)
  - cDNA arrays (Brown/Botstein)
  - Long oligo arrays (Agilent)
  - Serial analysis of gene expression (SAGE)
- 按实验要求分类
  - 单通道 (Single Channel): 一次检验一种状态
  - 双通道 (Dual Channel): 差异表达基因的筛选

两类主流的DNA芯片

- cDNA microarrays: 将500~5,000bp的cDNA固载到介质上 (例如玻璃)，Stanford开发设计，通常为双通道
- DNA chips: 将寡核苷酸探针 (20~80-mer) 合成到芯片上，Affymetrix开发设计，通常为单通道

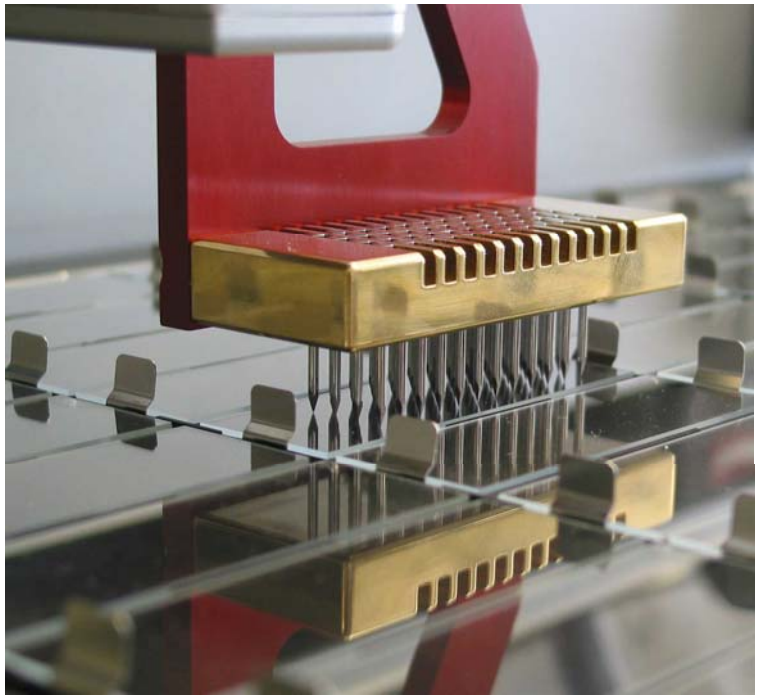
cDNA microarrays

- DNA连接到固态载体：玻璃、塑料或尼龙

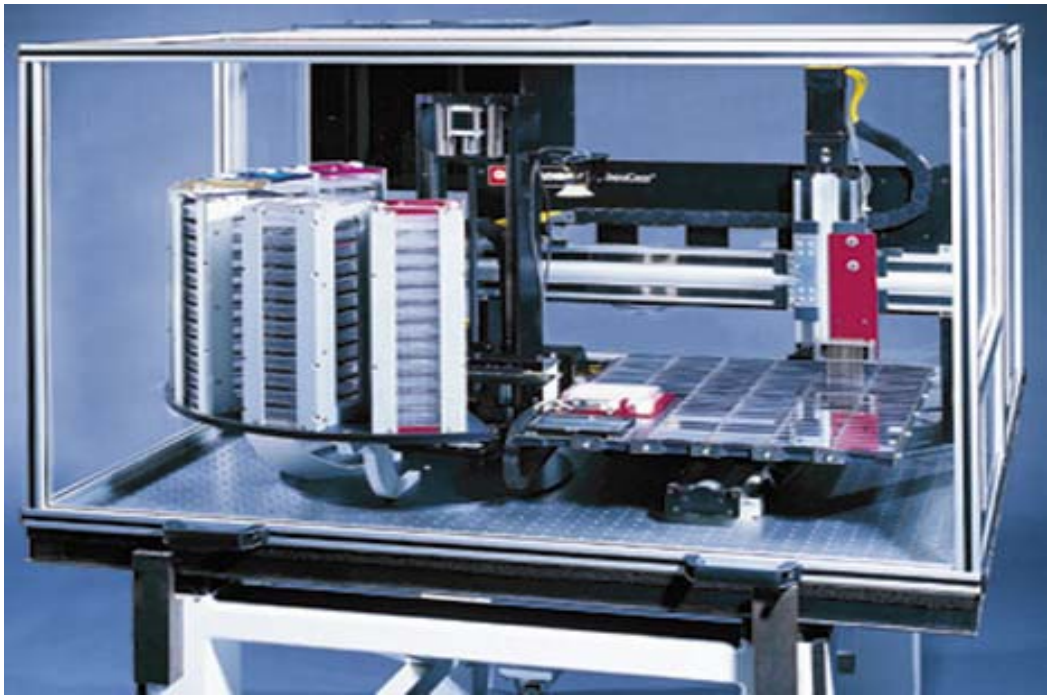


cDNA microarrays的制备

Robot spotter

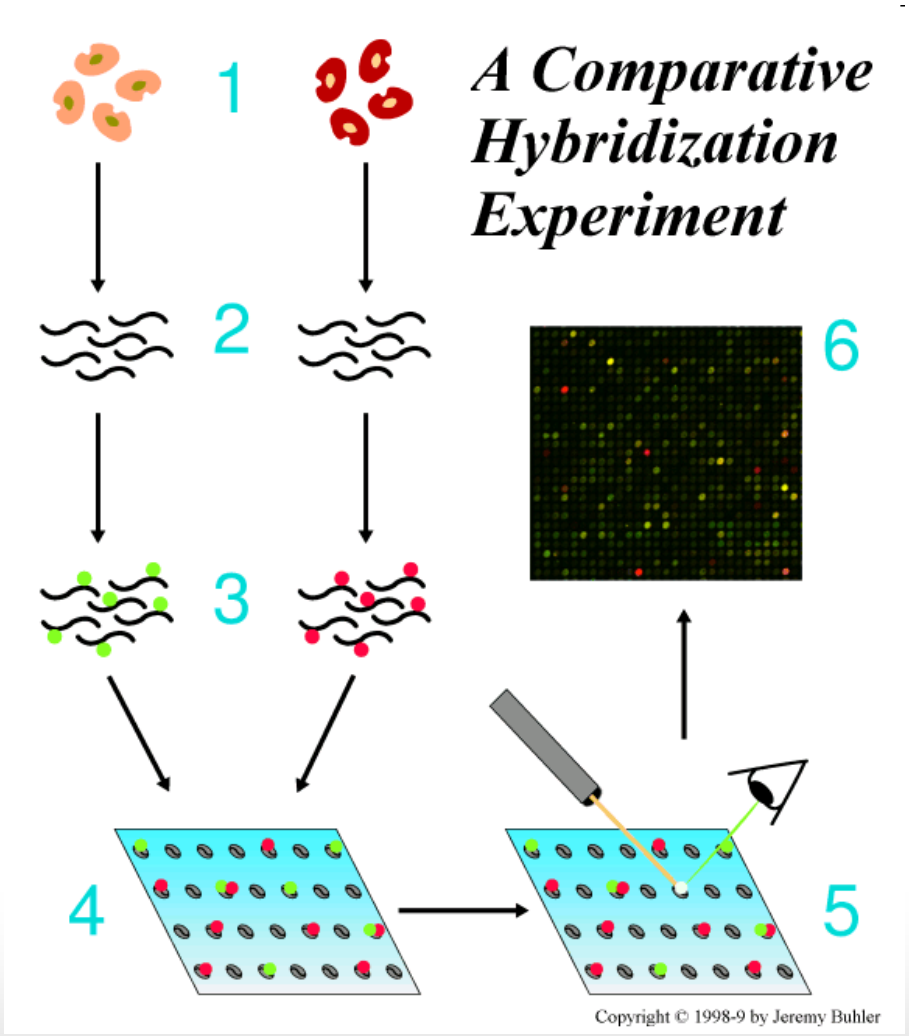


Commercial DNA spotter

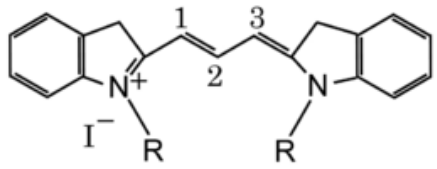


普通盖玻片

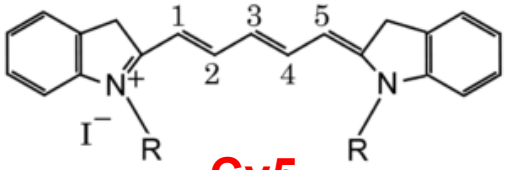
差异表达基因的筛选



荧光染料



Cy3



Cy5

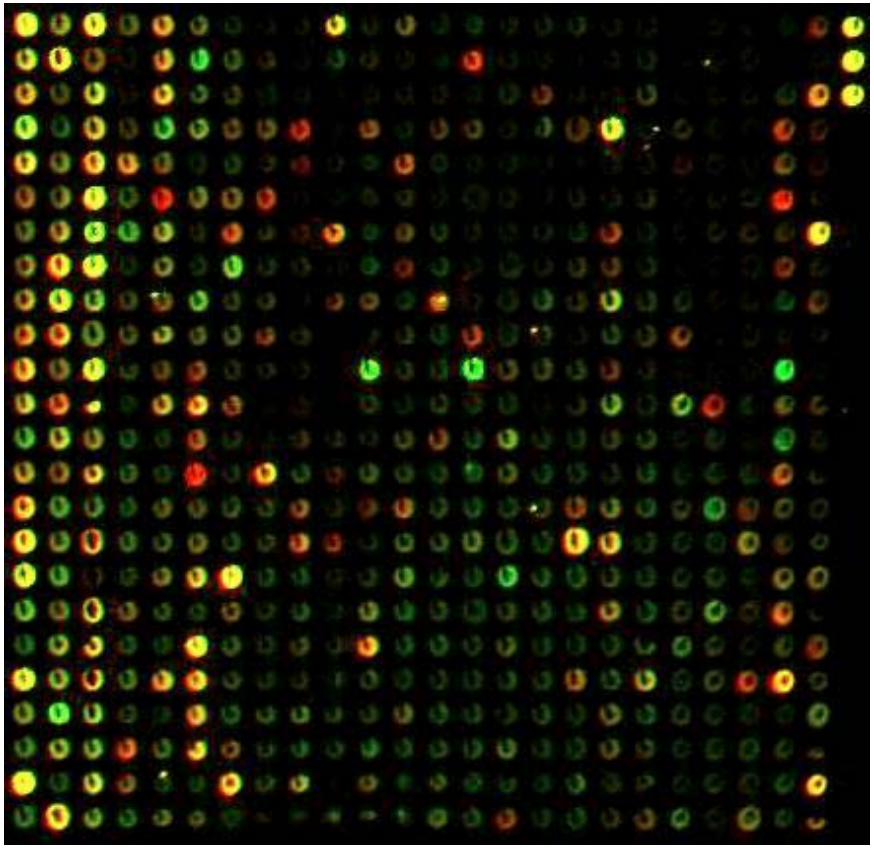
Treatment / control

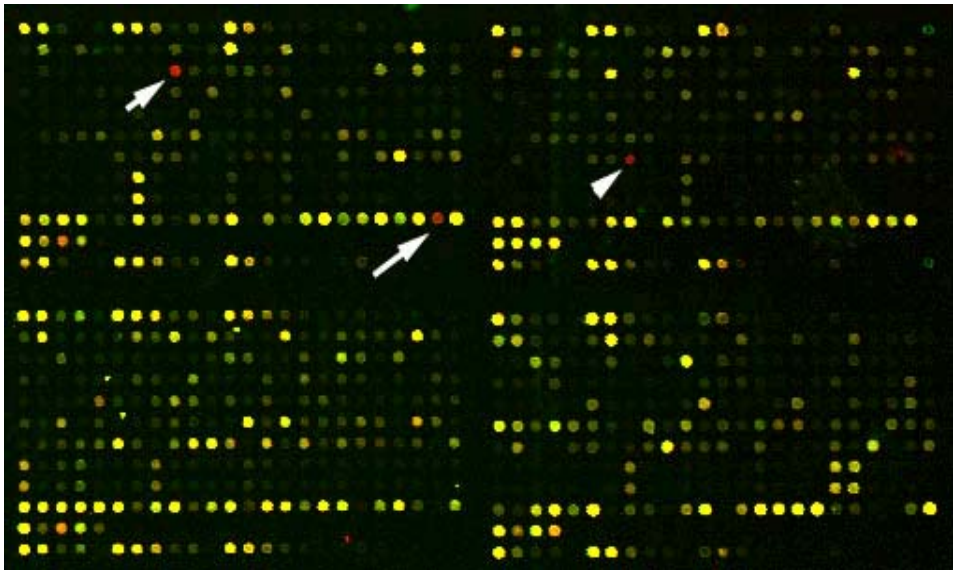
Normal / tumor tissue

Brain / liver

…

点样后的cDNA Microarrays





基因表达的数据

mRNA 样本

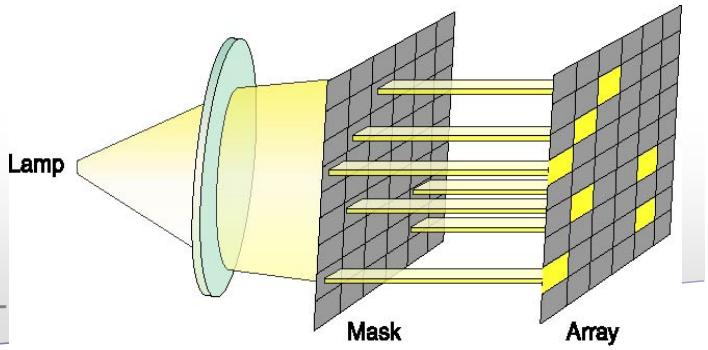
基因列表		sample1	sample2	sample3	sample4	sample5	...
	1	0.46	0.30	0.80	1.51	0.90	...
	2	-0.10	0.49	0.24	0.06	0.46	...
	3	0.15	0.74	0.04	0.10	0.20	...
	4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	5	-0.06	1.06	1.35	1.09	-1.09	...

基因*i*在mRNA样本*j*中的表达水平

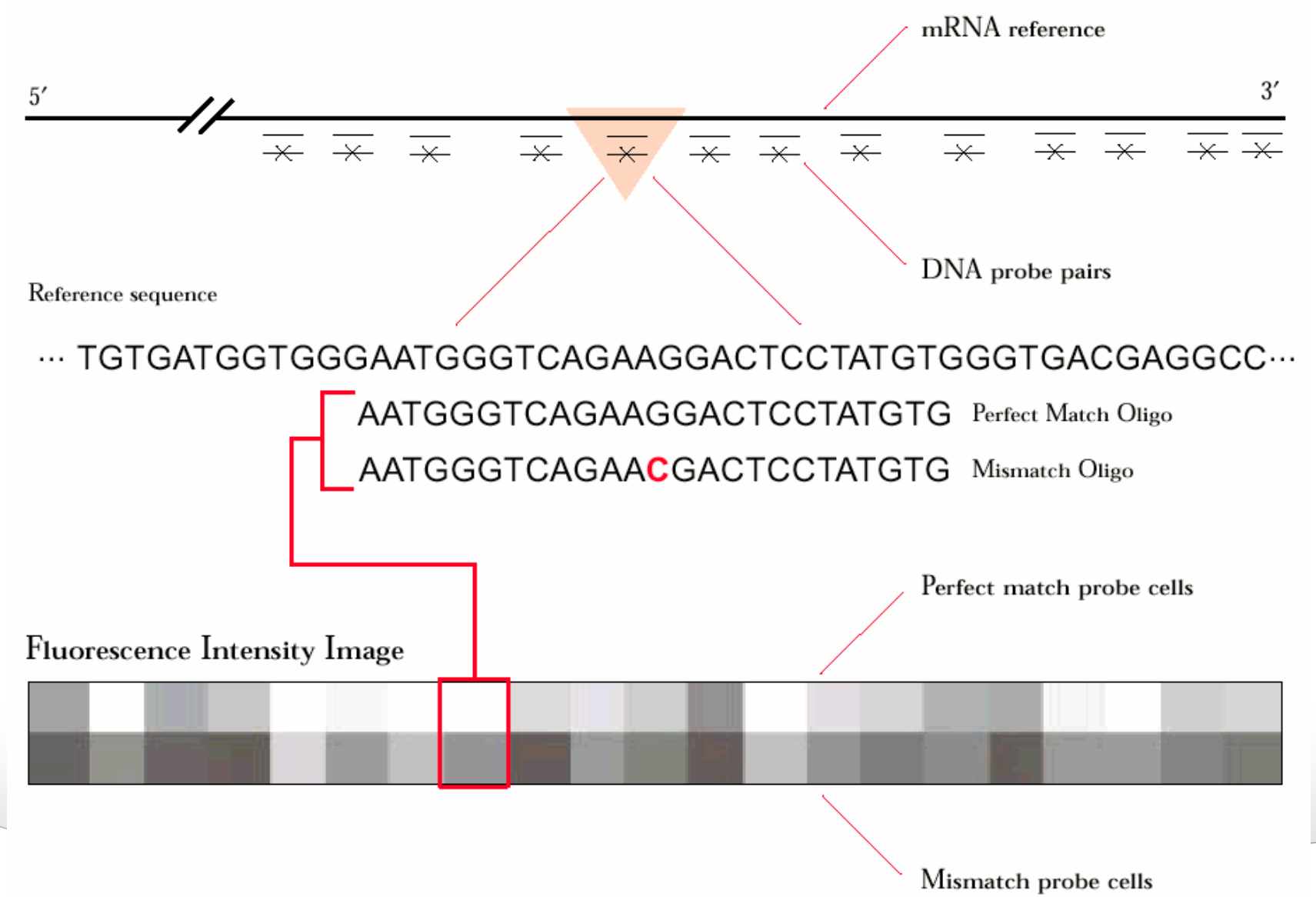
= { Log (Red intensity / Green intensity)
  { Log(Avg. PM - Avg. MM)

Affymetrix GeneChips

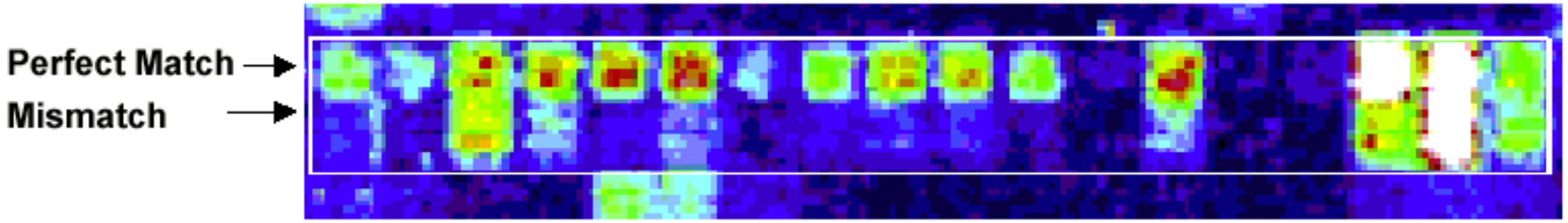
- 寡聚核苷酸链 (Oligonucleotide)
  - 一般长度为20~25bp
  - 每个基因设计10~40个不同的寡聚核苷酸链
- 每个基因寡聚核苷酸链的选择
  - 在基因组上是唯一的
  - 不发生重叠
- GeneChip的制备
  - 硅芯片：光刻 (photolithography)
  - 每次合成一个碱基



Affymetrix GeneChips



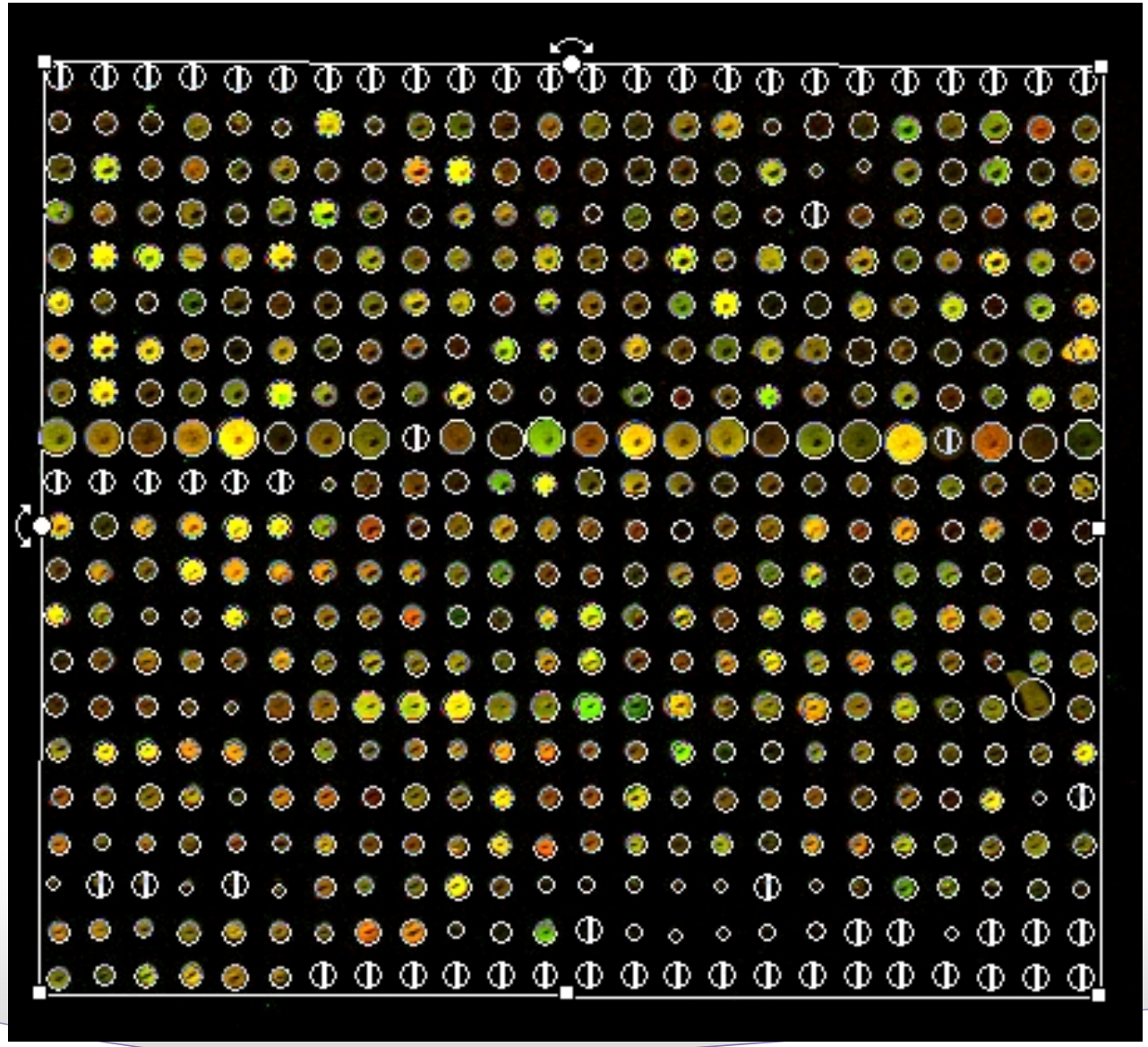
A Probe Set (DNA Chip)



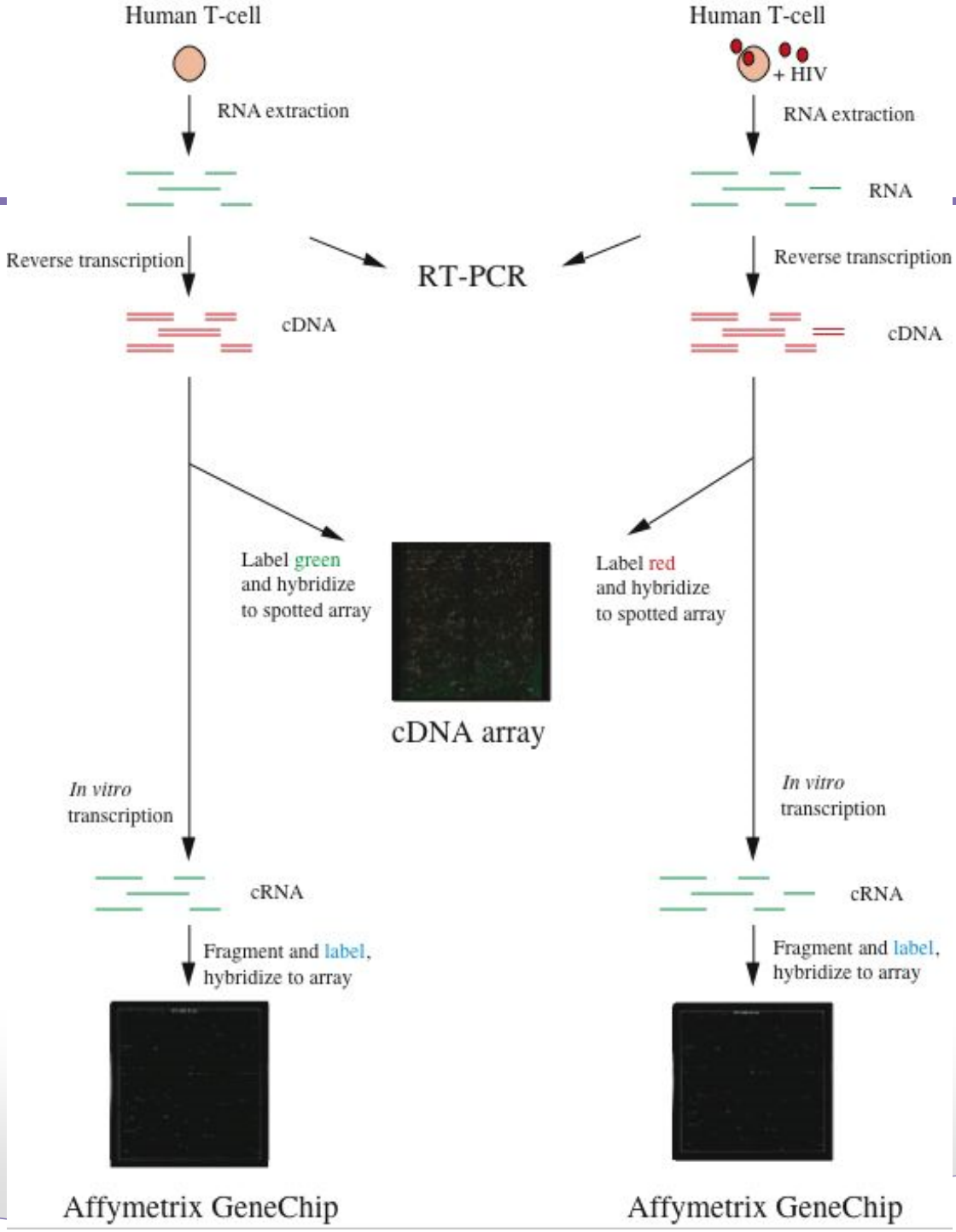
Perfect Match AGGCTATCGCACTCCAGTGG

Mismatch AGGCTATCGTACTCCAGTGG

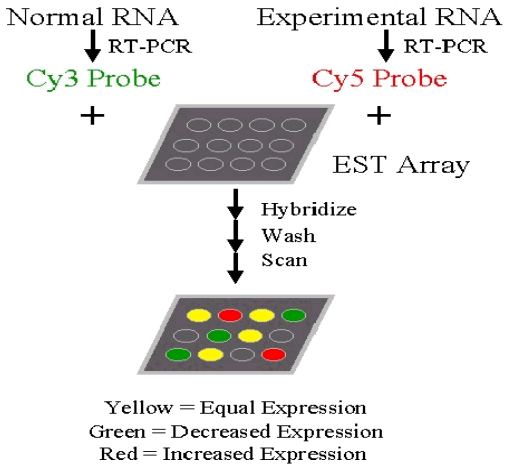
点样后的GeneChip

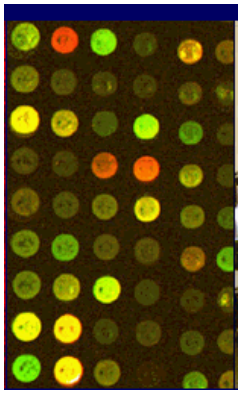


总结


Human T-cell
RNA extraction
Reverse transcription
cDNA
Human T-cell
+ HIV
RNA extraction
RNA
Reverse transcription
cDNA
RT-PCR
Label green
and hybridize
to spotted array
Label red
and hybridize
to spotted array
cDNA array
In vitro
transcription
cRNA
Fragment and label,
hybridize to array
In vitro
transcription
cRNA
Fragment and label,
hybridize to array
Affymetrix GeneChip
Affymetrix GeneChip

基因芯片的实验流程

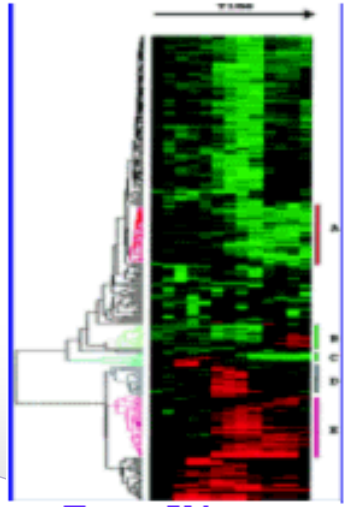





Microarray Scanner

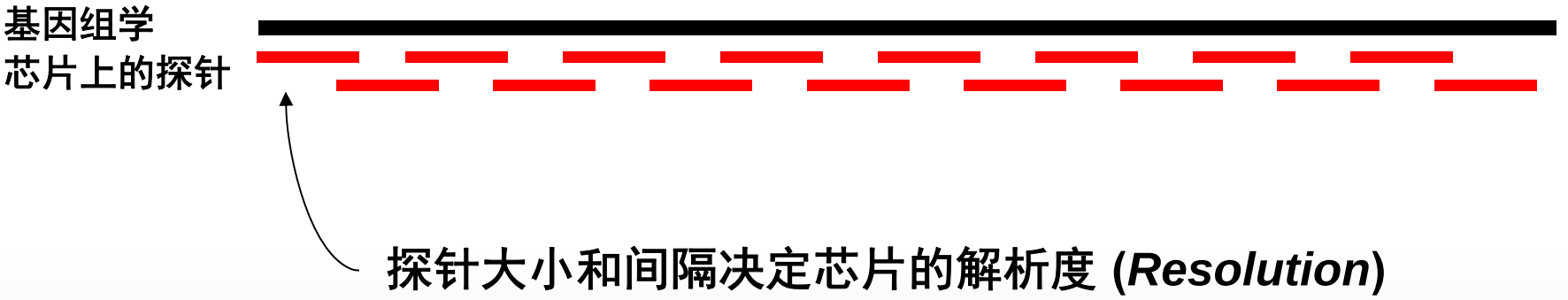
	A	B	C	D	E	F	G
1	YORF	NAME	GWEIGHT	spo0	spo30	spo2	spo5
2	EWEIGHT			1	1	1	1
3	YAL003W	EFB1	1	0.23	-1.79	-1.29	-1.56
4	YAL004W	YAL004W	1	0.41	-0.38	-0.89	-1.06
5	YAL005C	SSA1	1	0.61	-0.07	-1.29	-1.29
6	YAL010C	MDM10	1	0.16	-0.15	-0.76	-1.25


Cluster


TreeView

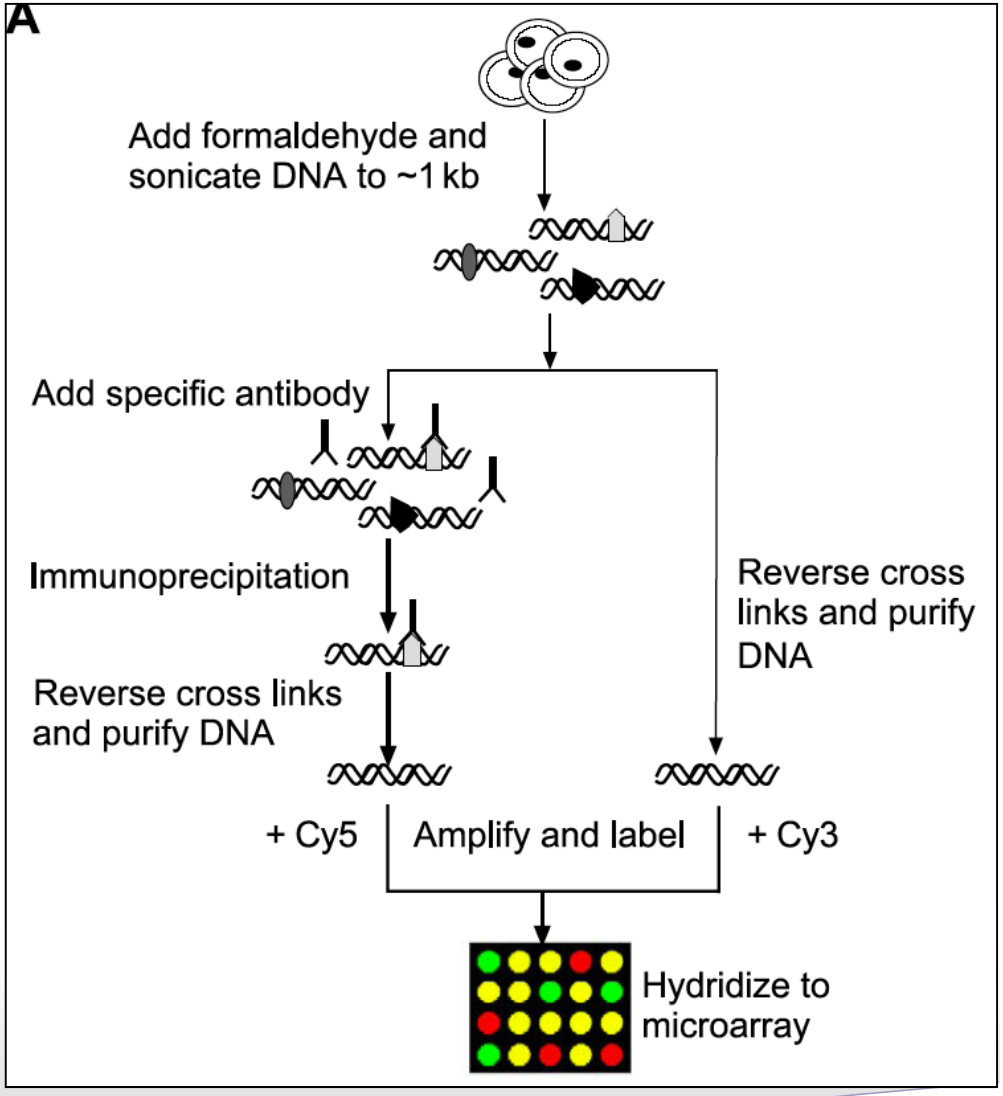
铺瓦芯片 (Tiled microarray)

- 高度覆盖基因组的某个区域，或整个基因组
- 设计的探针能够覆盖序列上的每一个碱基对
- 一般不包括重复序列

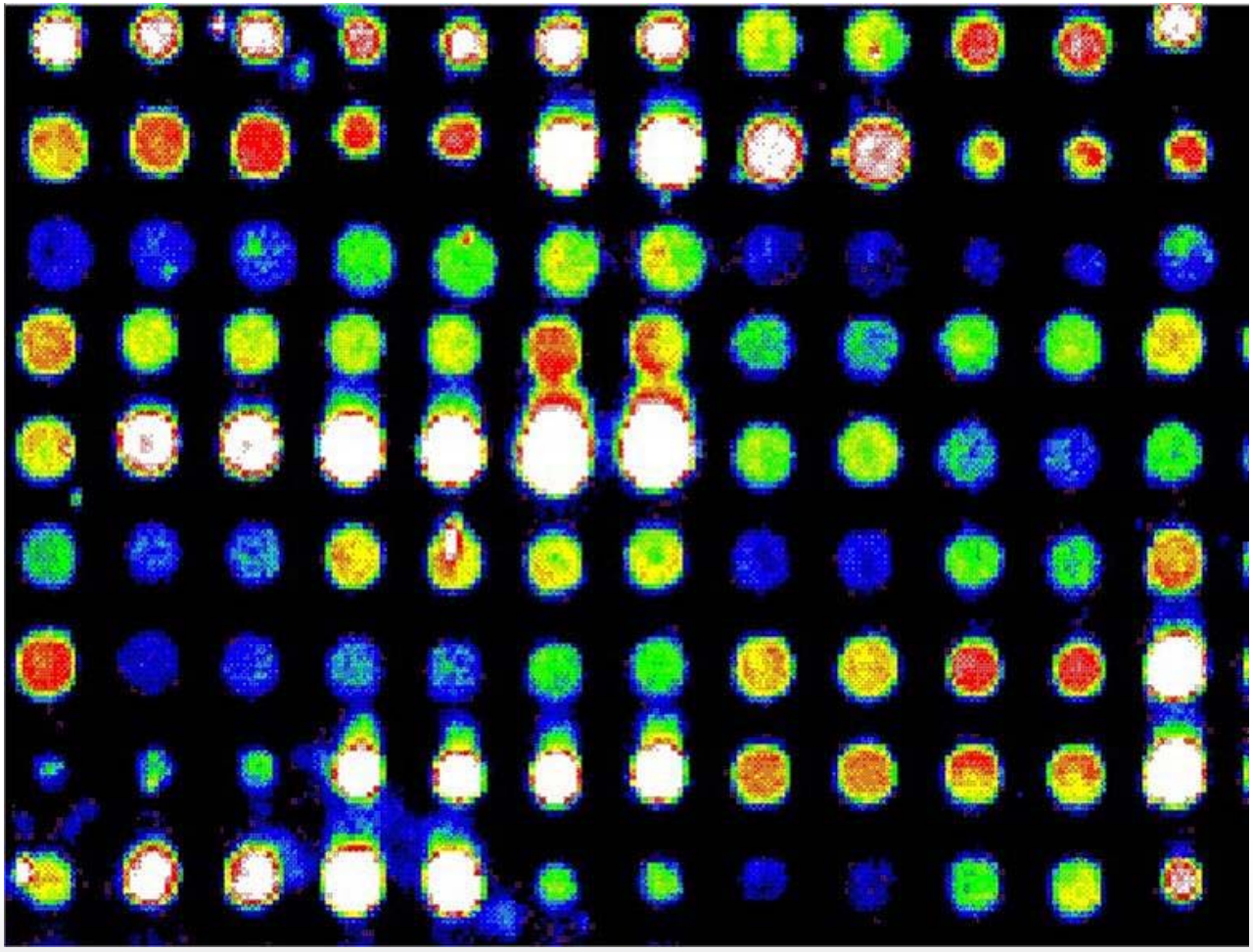


其他芯片：ChIP-chip

- 染色质免疫共沉淀
  - Chromatin immunoprecipitation (ChIP)
- 蛋白质与DNA的结合
  - 转录因子
  - 修饰的组蛋白



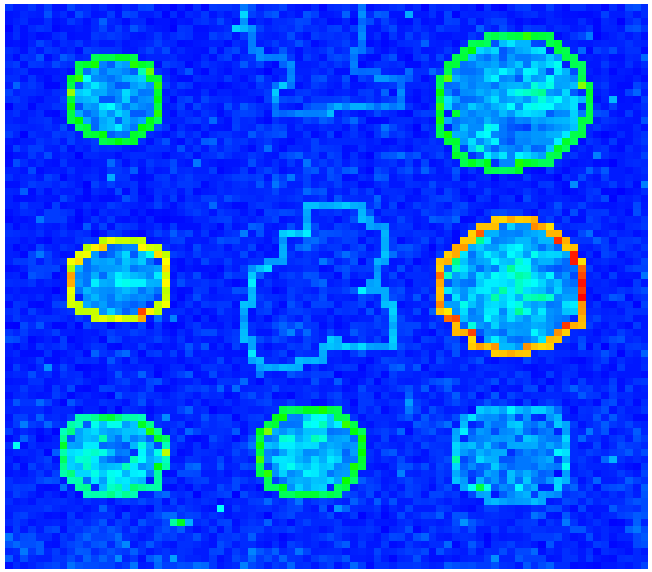
图像处理与数据标准化



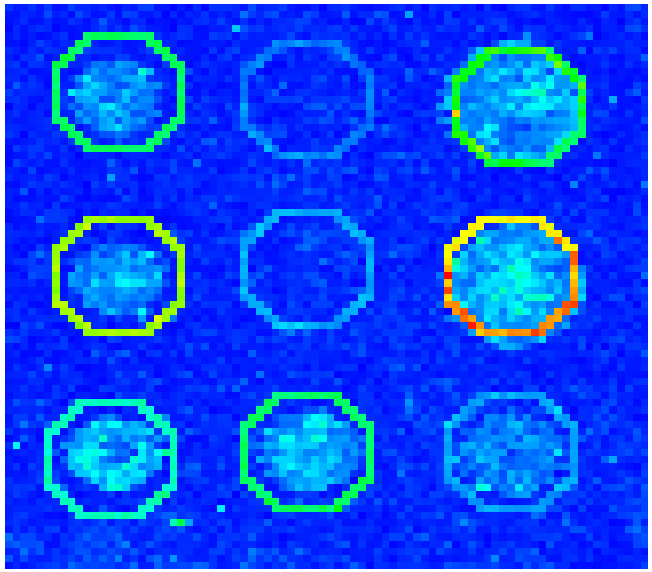
单通道基因芯片
white (very high)
red (high)
Yellow (a little high)
green (medium)
blue (low)
black (no)

图像处理

- 栅格化：确定点的位置
- 图象分割 (Segmentation)：将点从背景中分离出来
- 抽提亮度：各个像素亮度的均值 (mean) 或中位数 (median)
- 背景校正：局部或全局



植根区域生长法(SRG)



Fixed Circle

基因表达的定量

对于每个点，我们可以计算

Red intensity $= R_{fg} - R_{bg}$

fg = foreground, bg = background, and

Green intensity $= G_{fg} - G_{bg}$

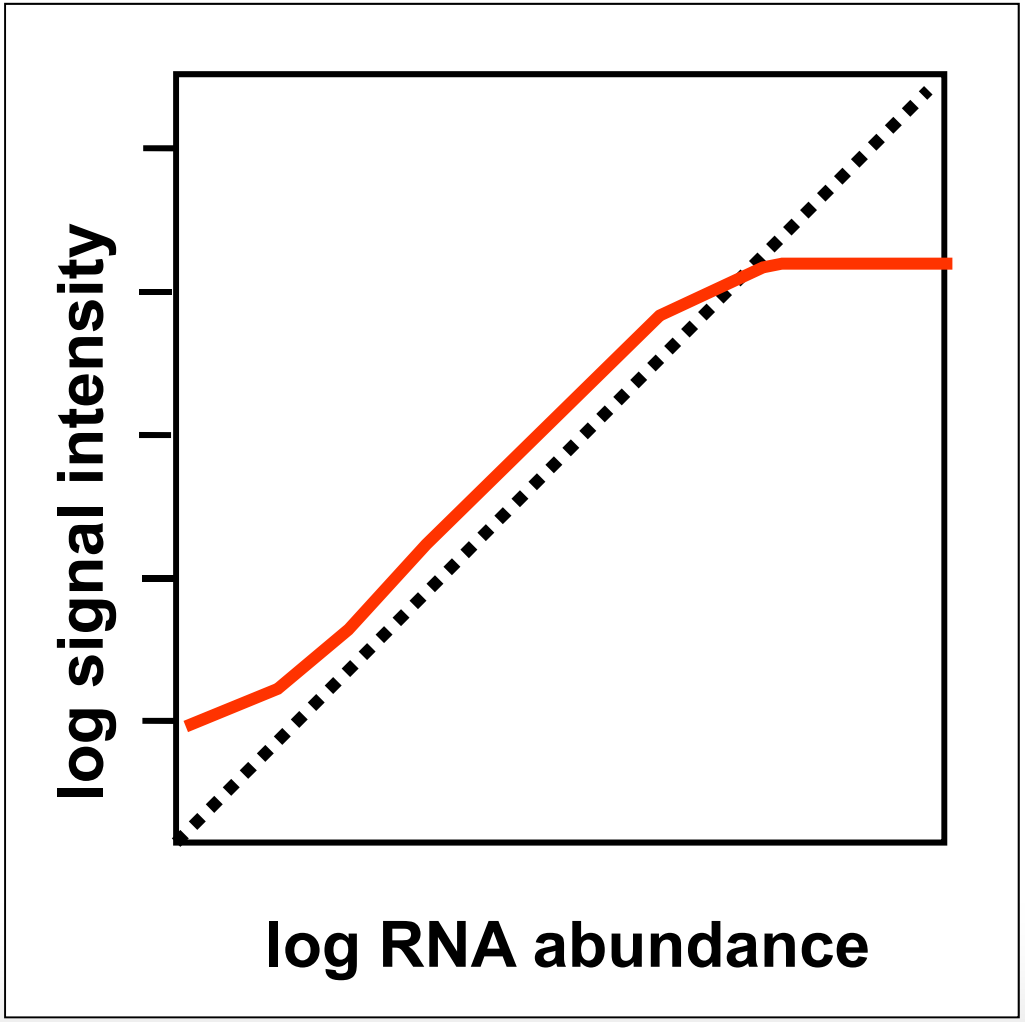
and combine them in the log (base 2) ratio

Log_2(Red intensity / Green intensity)

Green intensity (medium): ~1

Microarray: 误差的来源

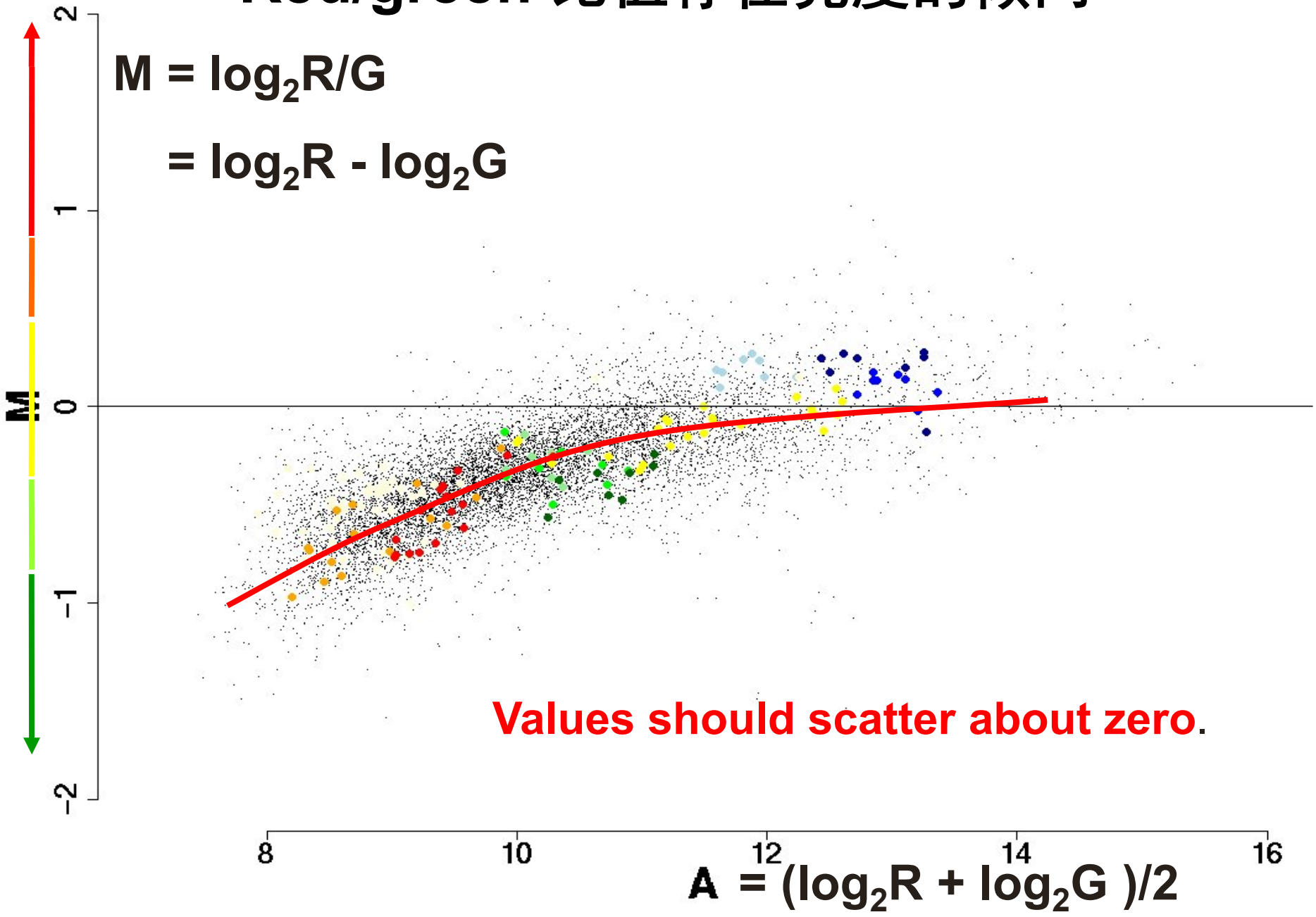
- 图像分析
- 扫描
- DNA杂交过程
  - 温度
  - 时间
  - 混合均匀程度
- 探针的标记
- RNA的抽提
- 加样
- 其他



Red/green 比值存在亮度的倾向

$M = \log_2 R/G$

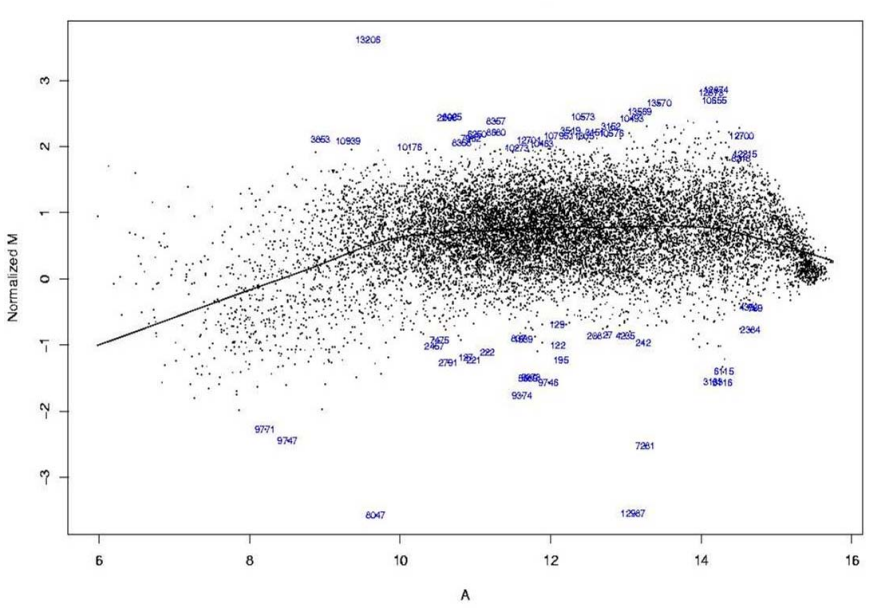
$= \log_2 R - \log_2 G$



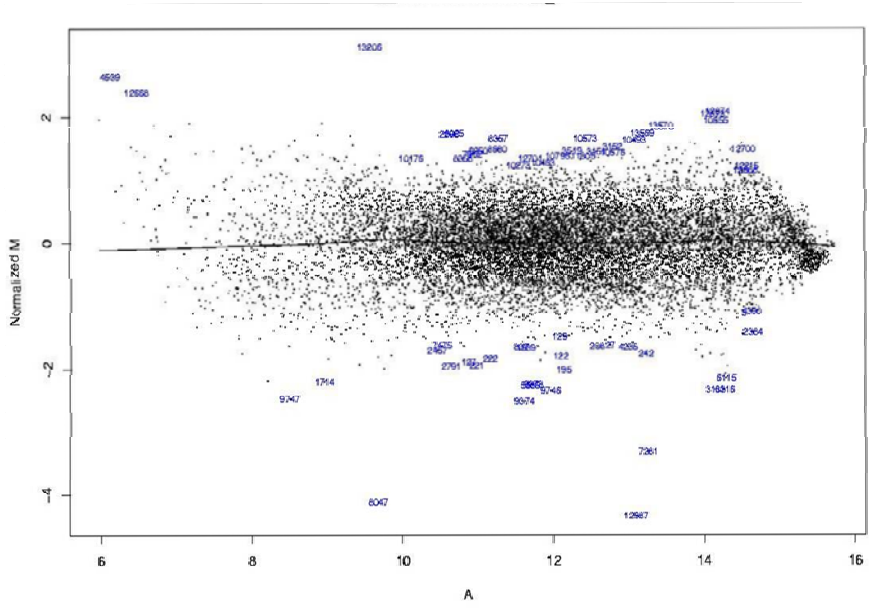
Values should scatter about zero.

数据标准化

标准化前



标准化后



基因芯片的数据分析

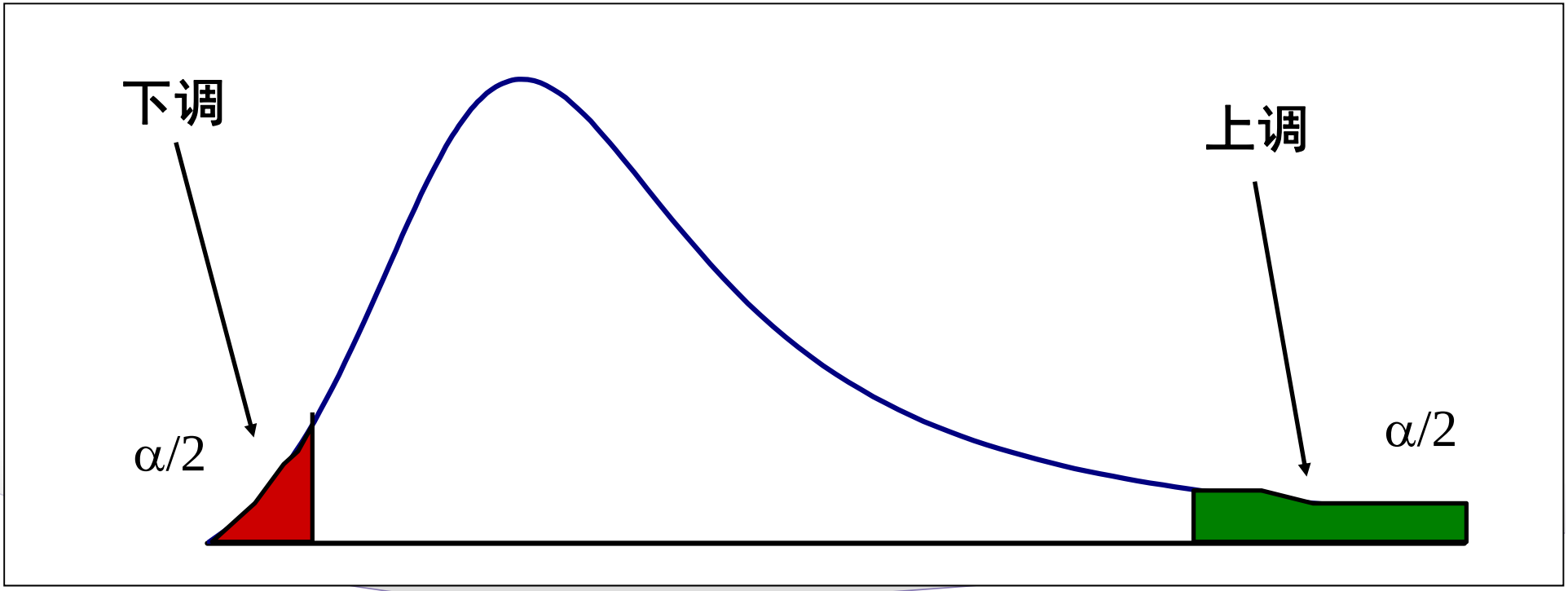
- 差异表达基因分析
- 基因共表达分析
- 基因表达数据的聚类和分类
- 基因集分析
- 基因调控网络

差异表达基因的分析

- 差异表达基因的分析: 寻找处理前后表达上调或者下调的基因
- 处理是否有差异？
- 使用标准的统计学方法检验 (t-test or F-test)，发现统计显著性差异表达的基因，
- 如果处理本身并不显著，则结果无意义

统计学分析

- Fold change, 一般2-fold上调或下调 (平行实验的样本较少)
- *p*-value (平行实验的样本较多)



p-value: 学生分布

- T-test: 学生分布
- Excel函数：TTEST(array1,array2,tails,type)
  - Array1为第一个数据集
  - Array2为第二个数据集
  - Tails指示分布曲线的尾数。如果 tails = 1，函数 TTEST 使用单尾分布。如果 tails = 2，函数 TTEST 使用双尾分布
  - Type为 t 检验的类型
    - 1 成对
    - 2 等方差双样本检验
    - 3 异方差双样本检验

p-value: 学生分布

- 一般选择双尾分布
- 异方差双样本检验
- Excel函数：=TTEST(B2:D2,E2:G2,2,3)
- C：对照组；T：实验组

	C1	C2	C3	T1	T2	T3	TTEST
Gene 1	1.322	1.676	1.457	3.526	4.234	3.879	0.001988

多重比较

- 在基因芯片的实验中，每一个基因/探针，都是一个独立的实验
- 基因芯片：高通量，>1,000个基因/探针
- 因此，无论怎么比较，总会有一些基因会是统计显著性差异表的 —— 可能是随机产生的
- 如何评估表达差异基因预测的有效性？
- 例：1,000个探针的双通道芯片，以*p*-value < 0.01为域值，发现7个上调基因，5个下调基因，分析结果是否具有统计学意义？

Bonferroni correction

- 假阳性预测
  - “Type 1 error" or "False Discovery"
- 总的I型错误率 (Family-wise error rate, FWER)
  - 多重检验的校正

Bonferroni correction

set α to desired α/number of tests
so: $0.01/1{,}000 = 1*10^{-5}$

- 若差异表达基因接受*p*-value < 0.01为显著
- 则1,000个基因的多重检验可设置*p*-value < $1*10^{-5}$为阈值

错误发现率

- False Discovery Rate (FDR)
  - *p*-value：全部样本有多少被预测错误
  - *q*-value：预测的结果中有多少是错的
- 根据*p*-value计算每个差异表达基因的*q*-value
- 将*p*-value按从小到大进行排序，计算rank值
- 例如，可接受*q*-value < 0.05为阈值

Benjamini–Hochberg correction

$$q - value = p - value \times \frac{Count}{Rank}$$

Count：总数

Rank：秩

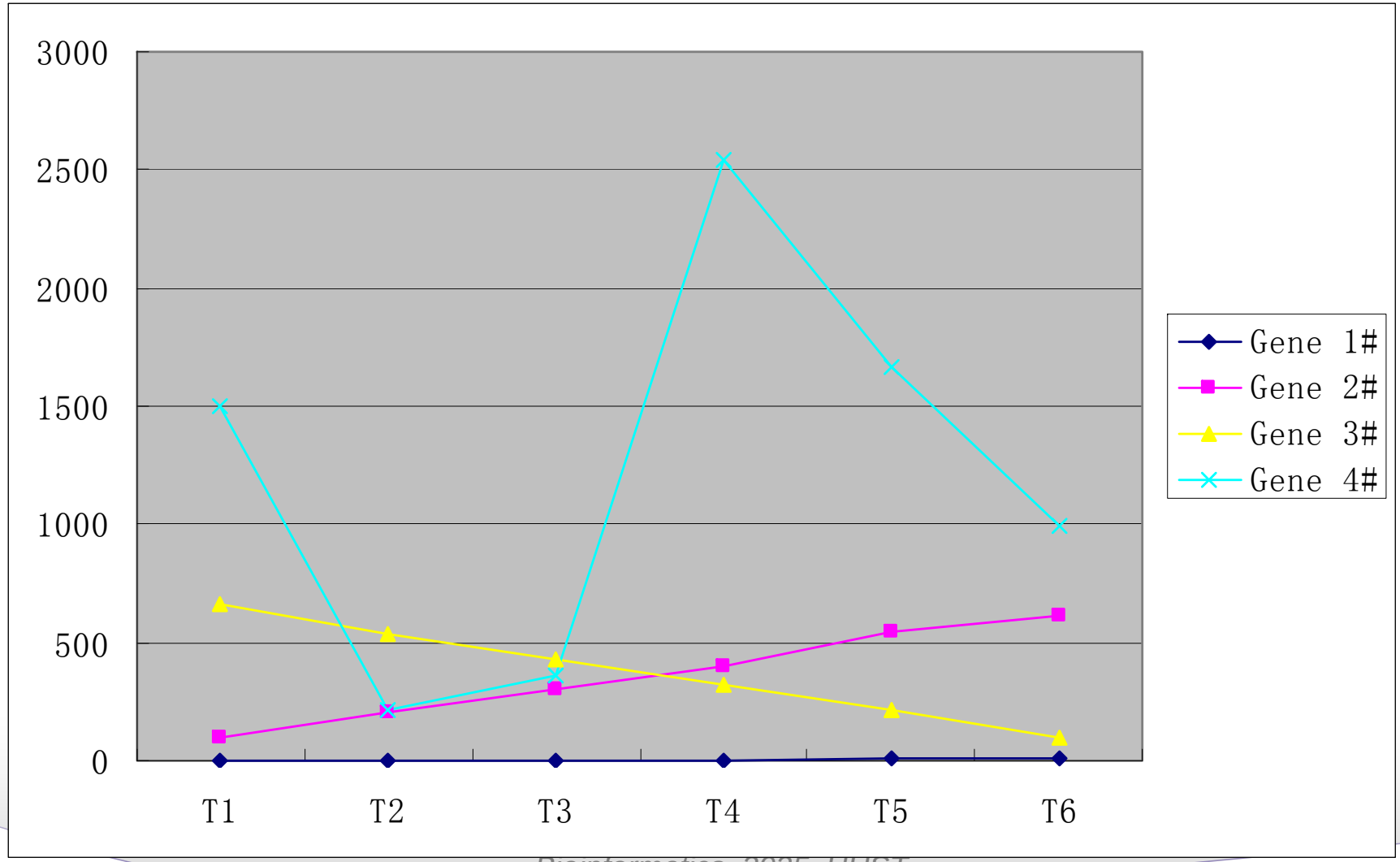
基因共表达分析

- 在N个不同的条件下 (时间序列的芯片数据)，考察基因X和Y的表达是否相似
- Gene 1#是否与Gene 2#、Gene 3#和Gene 4#共表达？
- 共表达：
  - 正相关：相似的表达谱，可能存在正关联
  - 负相关：相反的表达谱，可能存在负调控

Gene Name	T1	T2	T3	T4	T5	T6
Gene 1#	**1**	**2**	**3**	**4**	**5**	**6**
Gene 2#	100	200	300	400	550	610
Gene 3#	660	540	430	320	210	101
Gene 4#	1504	215	357	2545	1670	998

Eisen MB *et al.*, (1998) PNAS 95:14863-14868

没有相关性？


3000
2500
2000
1500
1000
500
0
T1
T2
T3
T4
T5
T6
Gene 1#
Gene 2#
Gene 3#
Gene 4#

基因相关性分析

- Spearman rank correlation
- Kendall's tau
- Euclidean distance
- Pearson correlation coefficient: -1 ~ 1
- Excel函数：=PEARSON(array1,array2)

$$\rho(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i}^{n} y_i$$

Pearson相关系数

- r ~ [-1, 1]
  - r~ 1，正相关
  - r~ -1，负相关

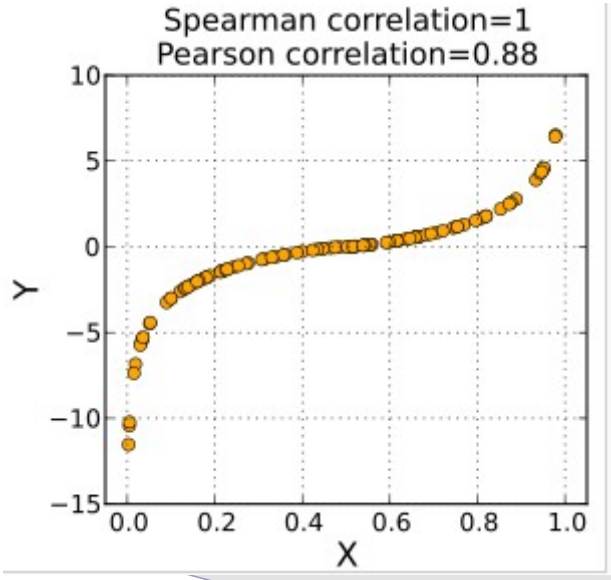
$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

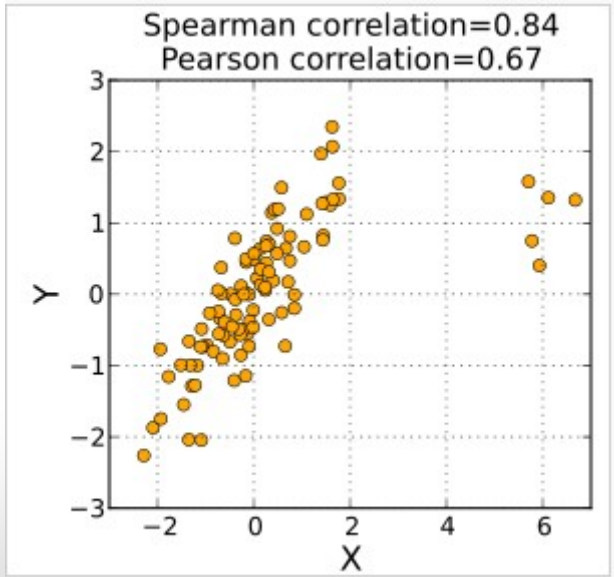
	Gene 1#	Gene 2#	Gene 3#
Gene 1#			
Gene 2#	0.996368		
Gene 3#	-0.99988	-0.99611	
Gene 4#	0.245292	0.254855	-0.2395

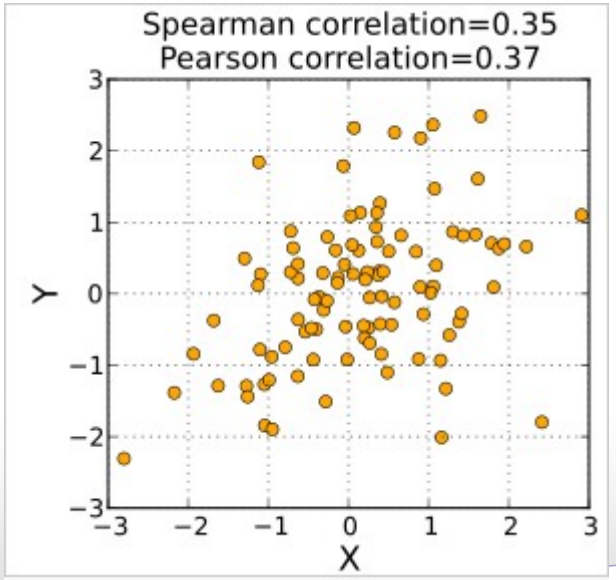
- 结论：**Gene 1#**与Gene 2#表达正相关，与Gene 3#表达负相关，与Gene 4#无关联

Spearman's rank correlation

- 等级相关：两组变量之间是否存在单调的相关性
- 与Pearson相关系数的区别：对数据的敏感性/依赖性较低







公式及计算方法

- 相关系数ρ
- $d_i = x_i - y_i$

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Gene Name	T1	T2	T3	T4	T5	T6
Gene 2# (x_i)	100	200	300	400	550	610
Gene 3# (y_i)	1504	215	357	2545	1670	998
Rank x_i	1	2	3	4	5	6
Rank y_i	4	1	2	6	5	3
d_i	-3	1	1	-2	0	3
d_i^2	9	1	1	4	0	9

- ρ= 1- (6 * 24)/[6(36-1)] = 0.3143

Kendall's tau

- **Kendall tau distance: 统计两个列表中数据的一致性分布情况**

Gene Name	T1	T2	T3	T4	T5	T6
Gene 2#	100	200	300	400	550	610
Gene 3#	1504	215	357	2545	1670	998

Pair	T1, T2	T1, T3	T1, T4	T1, T5	T1, T6	T2, T3	T2, T4	T2, T5
Gene 2#	<	<	<	<	<	<	<	<
Gene 3#	>	>	<	<	>	<	<	<
Count	1	1			1			

Pair	T2, T6	T3, T4	T3, T5	T3, T6	T4, T5	T4, T6	T5, T6
Gene 2#	<	<	<	<	<	<	<
Gene 3#	<	<	<	<	>	>	>
Count					1	1	1

- K= 3*2/ [6(6-1)] =0.2

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}$$

Euclidean distance

- 两组变量在n维空间上的直线距离
- 生物学意义：考察两个基因是否以1：1的关系发生相互关联

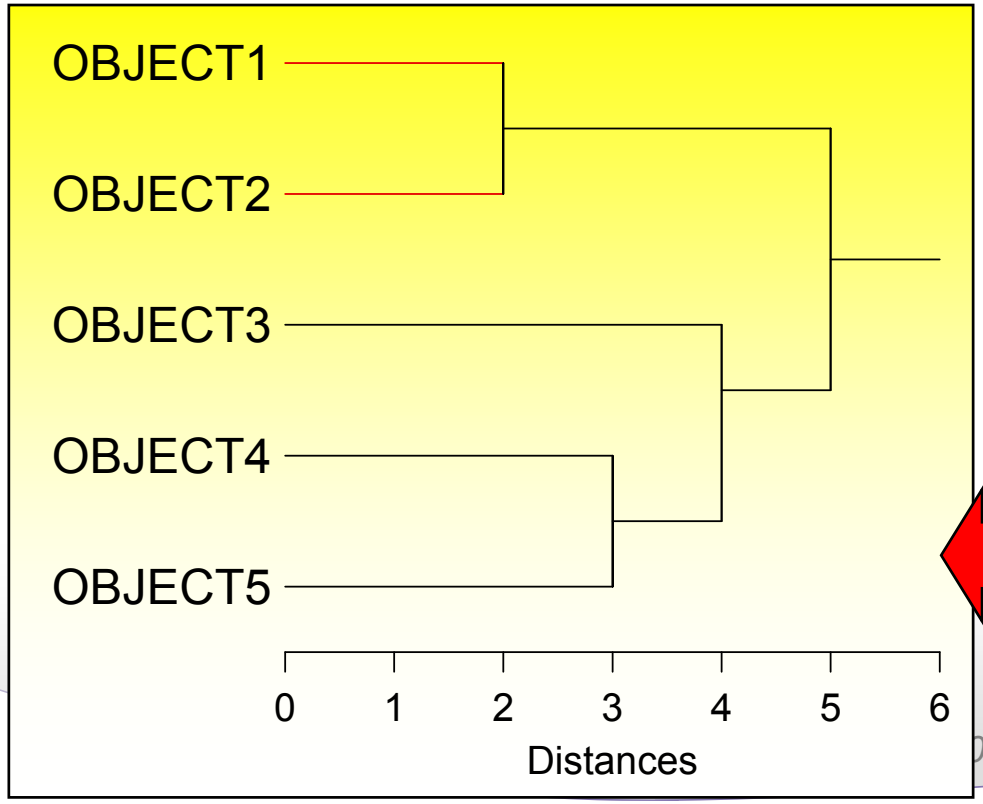
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

基因表达数据的聚类

- 将表达谱相似的基因聚类在一起
- 无督导学习 (Unsupervised learning)
- Pattern finding: 发现新的模式
- 聚类方法：
  - Hierarchical clustering
  - *K*-means clustering

Hierarchical clustering

- 用树状结构来表征基因表达之间的相似性/相关性
- 优点：不需要指定结果有多少类



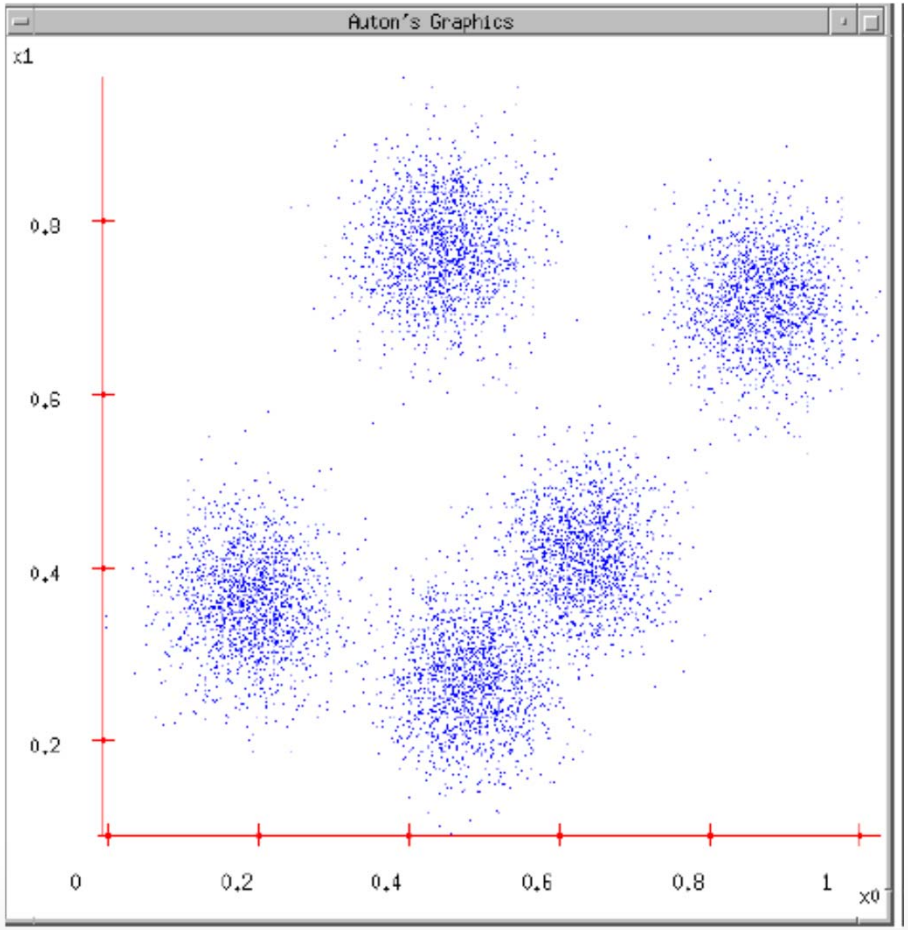
Object	1	2	3	4	5
1					
2	2				
3	6	5			
4	10	9	4		
5	9	8	5	3	

Distance matrix

Distance	Cluster
0	1,2,3,4,5
2	(1, 2), 3, 4, 5
3	(1, 2), 3, (4, 5)
4	(1, 2), (3, 4, 5)
5	(1, 2, 3, 4, 5)

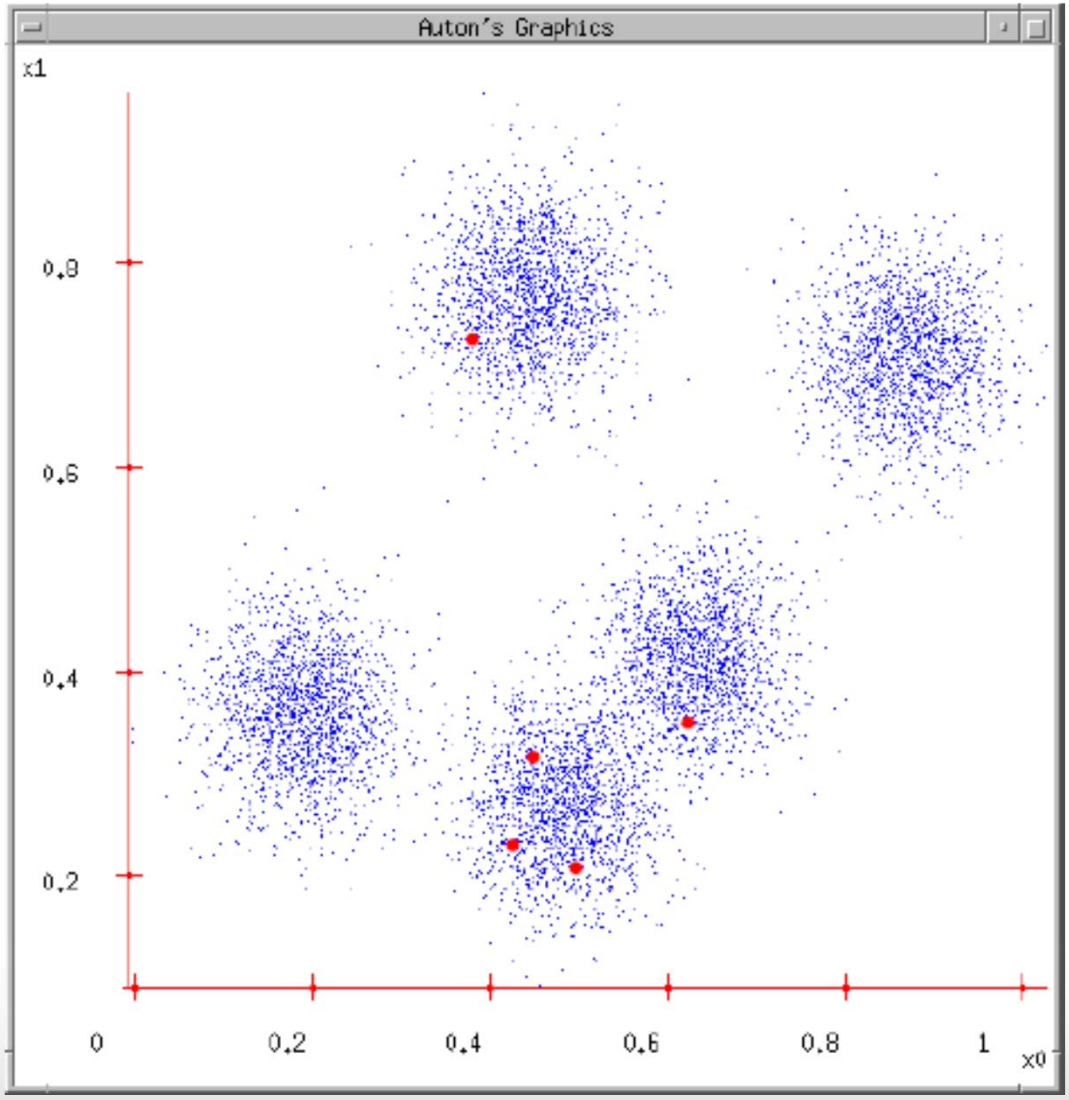
k-means clustering

- 对数据进行聚类
- 必须给定结果分成多少类！
- 假设该例中，指定为聚成5类



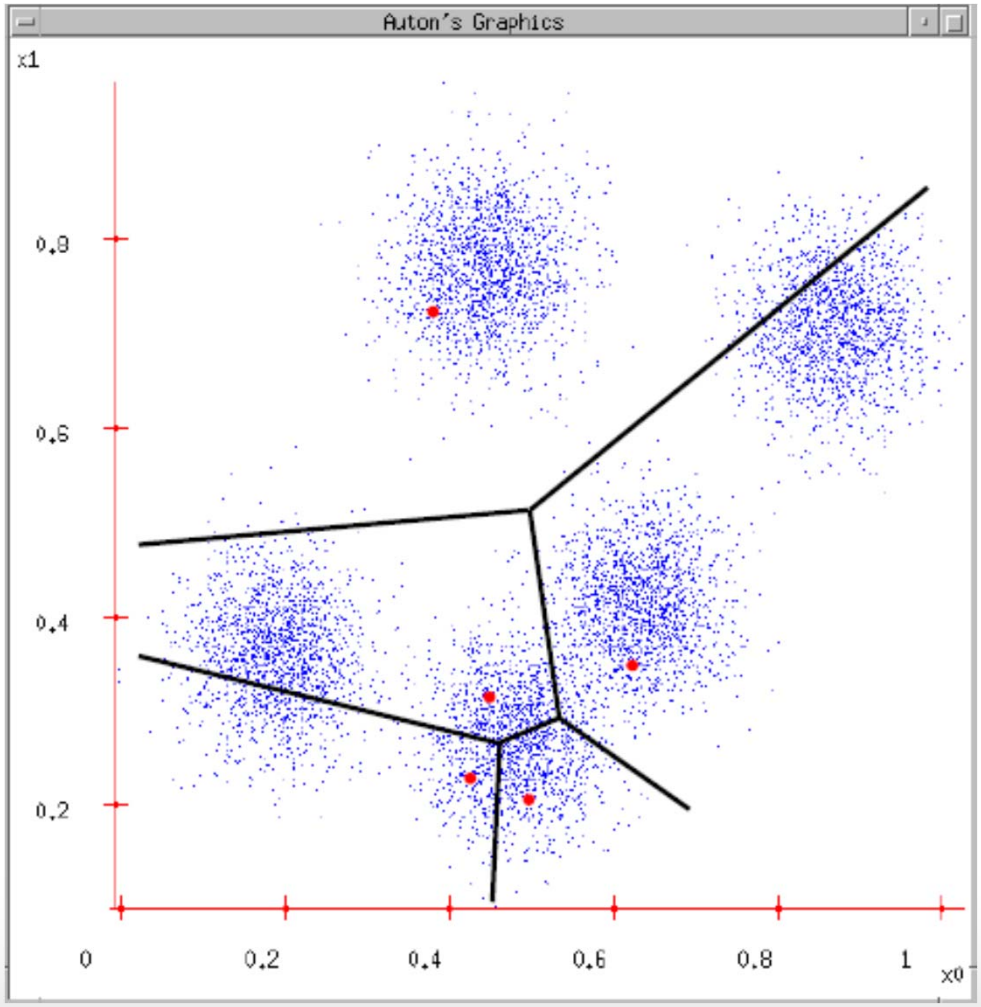
k-means clustering

- 随便选取5个点，作为每一个类的中心点



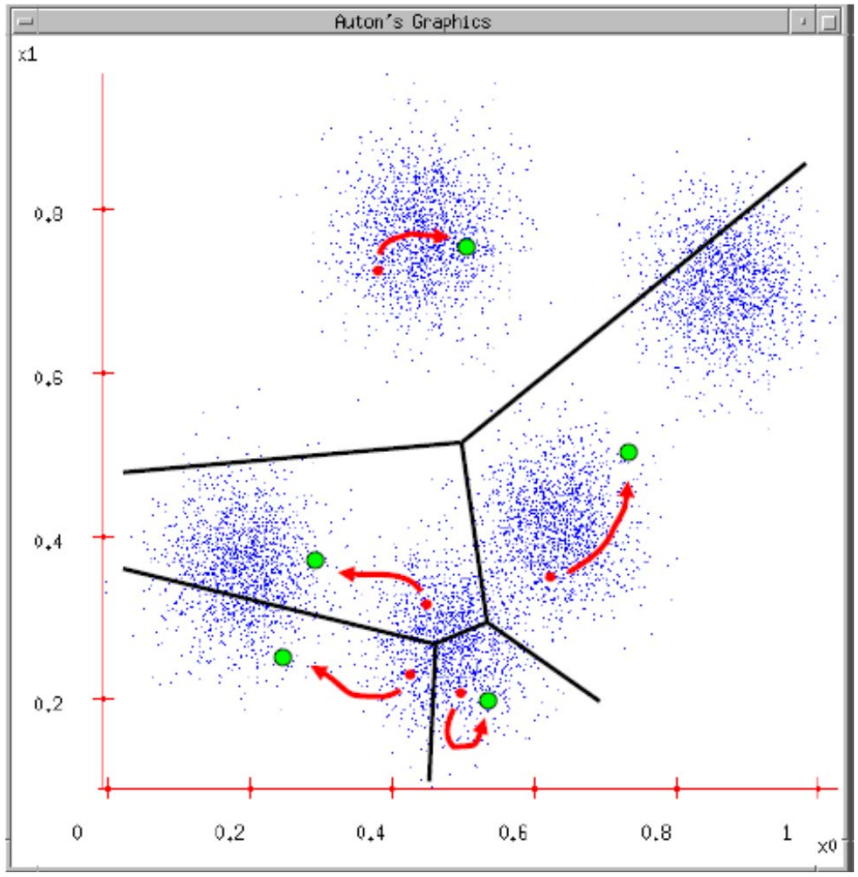
k-means clustering

- 计算其他点与这5个中心点的距离
- 距离：
  - 欧氏距离
  - 马氏距离
  - 皮尔孙相关系数…
- 点的归类：离哪个中心点近，归哪个类



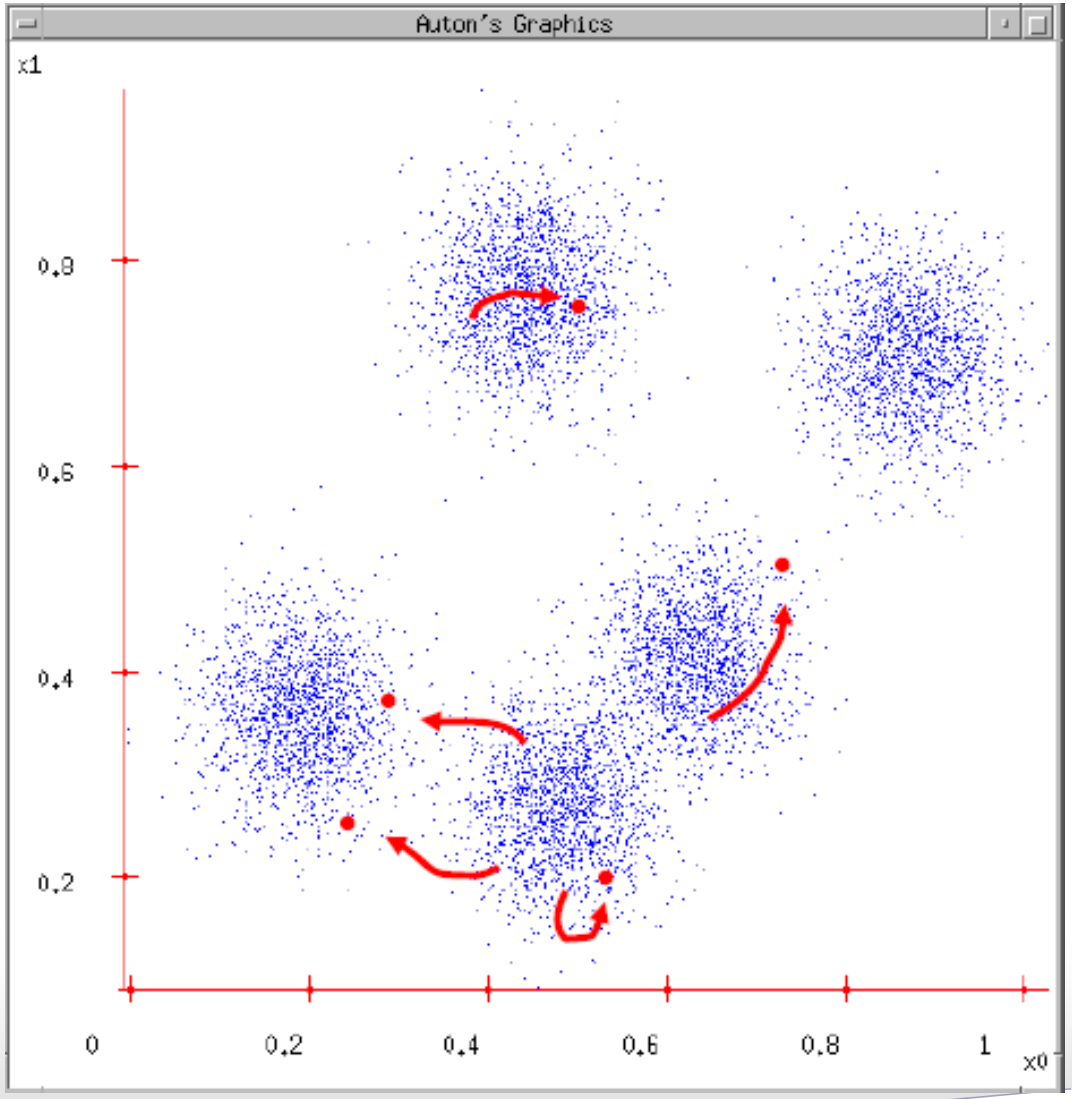
k-means clustering

- 针对每一类中的每一个点，计算其与其他点的距离，加和，除以该类点的数目
- 找到新的中心点，即改点到该类中其他点的平均值最小
- 确定新的5个中心点！



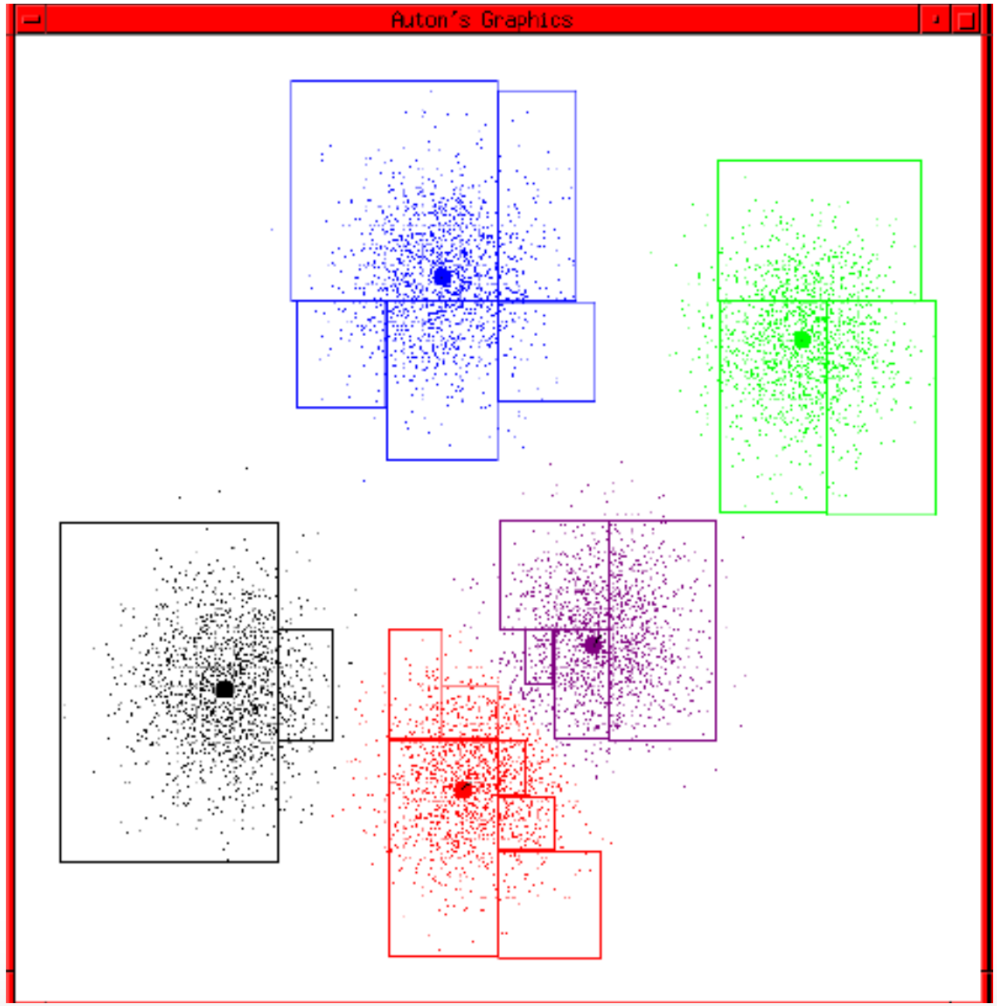
k-means clustering

- 重复2, 3，直到结果收敛
- 实际操作时，因结果完全收敛时间过长，一般指定迭代的次数，如1,000次



k-means clustering

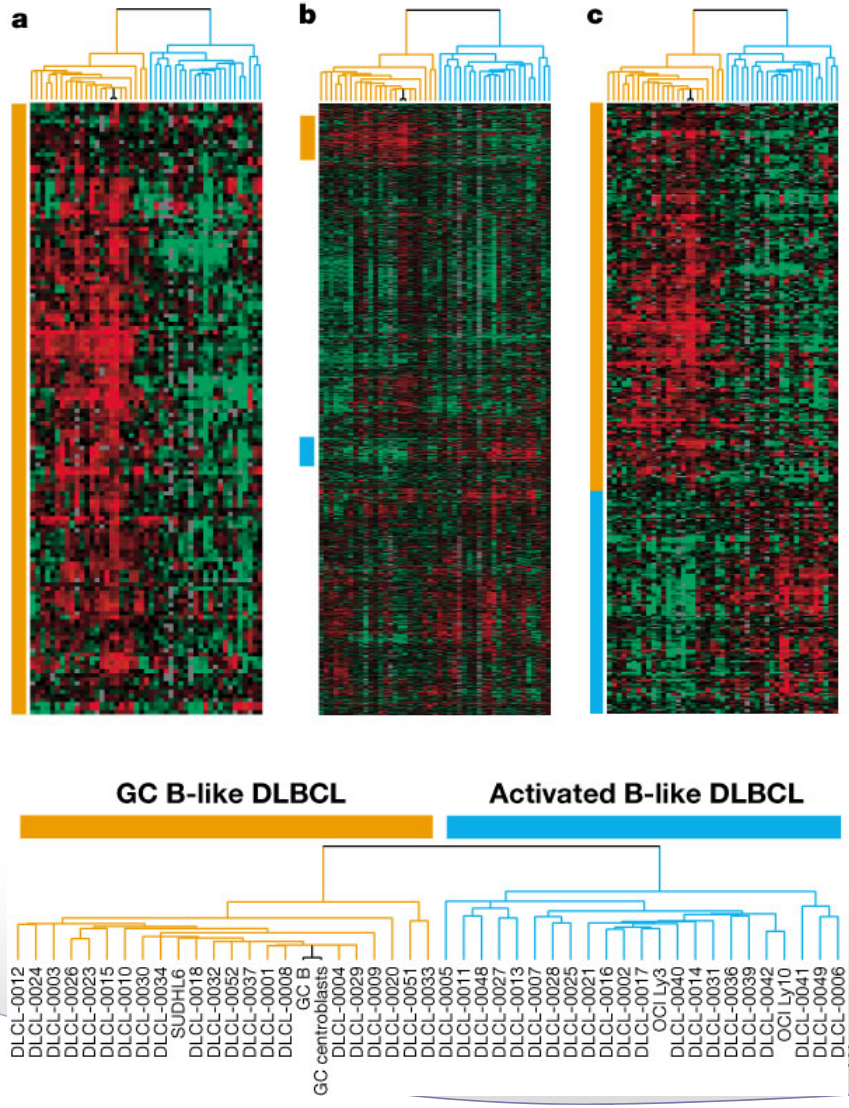
- 最终结果：所有基因芯片数据被聚成5类
- 软件：Cluster 3.0, Michael Eissen, Stanford



基因表达数据的分类

- 根据基因表达的数据将样本分成两类或多类；
- 督导学习 (supervised learning)：根据发现的pattern进行预测
- 应用：
  - 癌症 vs. 正常组织
  - 癌症的亚型、不同阶段 (良性的 vs. 恶性的)
  - 对药物的敏感性 (tamoxifen for breast cancer)

Diffuse large B-cell lymphoma (DLBCL)



1. 通过聚类发现各种亚型之间的关系

2. 根据基因表达模式，能够预测新的基因表达样本

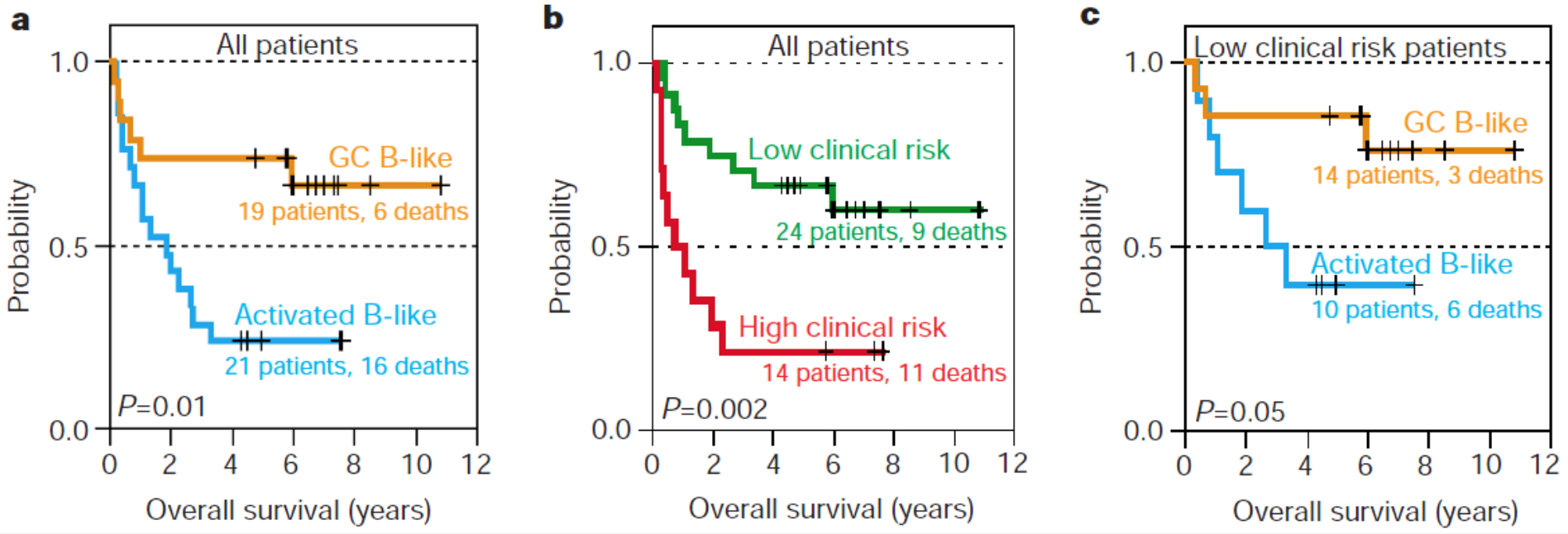
DLBCL: 弥散性大B细胞淋巴瘤

GC B-like:生发中心B细胞样亚型

Activated B-like :外周血活化B细胞样亚型

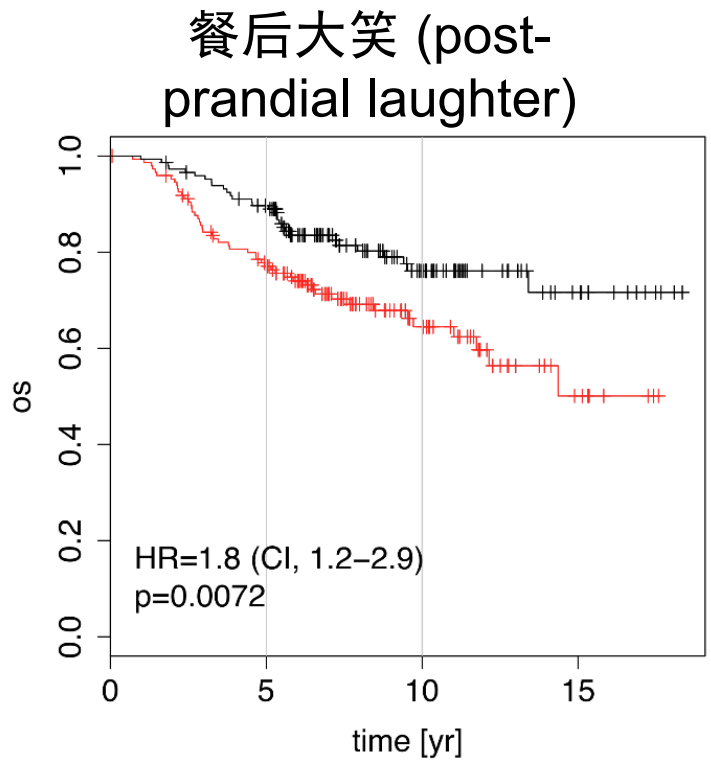
基因表达 *vs.* 诊断

- GC B-like的患者生存率明显比Activated B-like患者高



笑是最好的药？

- 绝大多数随机的基因表达标记与乳腺癌显著相关



Testing whether laughter IS the best medicine

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

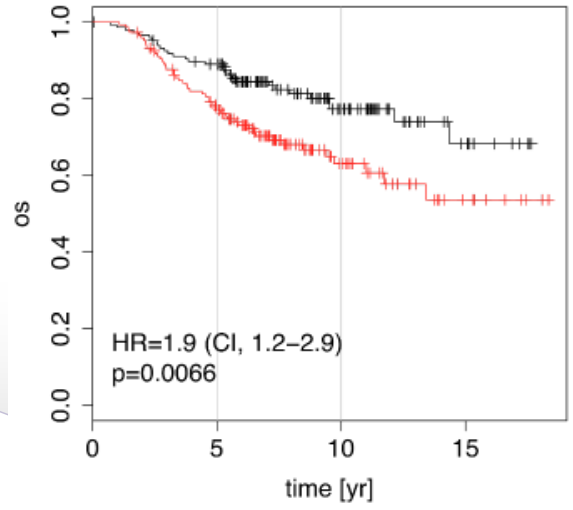
Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

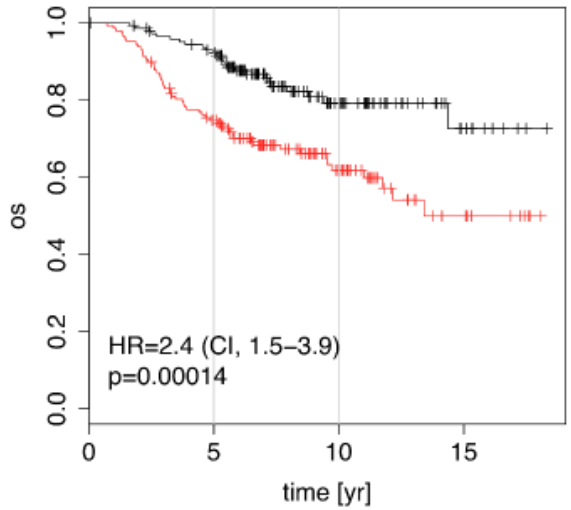
基因表达 *vs.* 诊断

- 随机的基因表达标记比已知的癌症相关标记与乳腺癌更显著相关
- 标记不能只关注准确性而不考虑生物学机制
- 不能只考虑*p*-value，还需要验证
- 动物模型、体外模型等可能有偏差

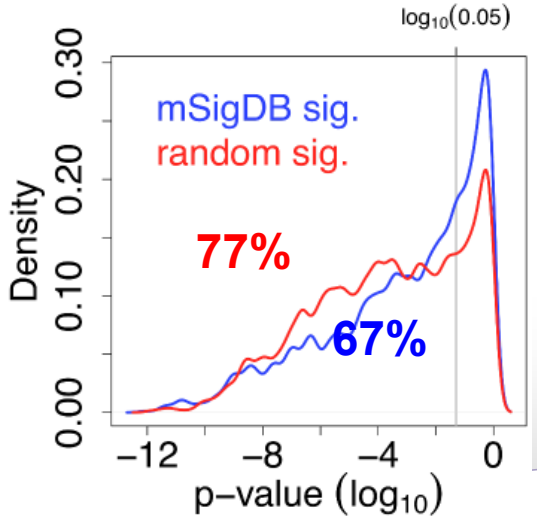
社交挫败



皮肤成纤维细胞定位



已知 vs. 随机



基因集分析

- Gene Set Analysis
- 通过基因芯片，找到了一批“interesting” 的基因
  - 差异表达基因
  - 不做差异基因鉴定，直接做基因集分析
- 生物学功能上是否存在关联？
  - 某种功能是否显著？
- 计算分析方法
  - 基因本体 (Gene Ontology)
  - KEGG (Kyoto Encyclopedia of Genes and Genomes)
  - 超几何分布

基因集的GO分析：超几何分布

- 例如：利用基因芯片技术，在某种条件下检测26,873个人类基因的表达水平，与对照比较之后，发现2,683个基因的表达量显著上调
- 其中所有人类基因中有2255个具有DNA binding (GO:0003677) 的GO注释，上调基因中有530个具有同样注释
- 表达上调基因是否显著具有DNA binding (GO:0003677)的功能？

显著性检验：超几何分布

$$Enrichment\_ratio = \frac{\frac{m}{M}}{\frac{n}{N}}$$

$$p-value = \sum_{m'=m}^{n} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \qquad (Enrichment\_ratio \geq 1)$$

or

$$p-value = \sum_{m'=0}^{m} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \qquad (Enrichment\_ratio < 1)$$

The top 15 most enriched processes and functions in SUMO substrates

Description of GO term	Number of proteins annotated in group S[a]	Number of proteins annotated in group W[b]	Enrichment ratio	*P*-value
The top 15 most enriched processes in SUMO substrates				
Regulation of transcription, DNA-dependent (GO:0006355)	26.1% (510)	9.0% (2174)	2.89	6.12 E − 121
Transcription from Pol II promoter (GO:0006366)	3.5% (69)	0.8% (204)	4.17	1.00 E − 25
Development (GO:0007275)	5.8% (114)	2.6% (631)	2.23	2.96 E − 16
Signal transduction (GO:0007165)	9.1% (178)	5.0% (1207)	1.82	2.06 E − 15
Regulation of transcription from Pol II promoter (GO:0006357)	2.7% (52)	0.8% (192)	3.34	4.13 E − 15
Protein amino acid phosphorylation (GO:0006468)	6.7% (131)	3.5% (850)	1.90	5.45 E − 13
Cell growth and/or maintenance (GO:0008151)	3.4% (67)	1.4% (341)	2.42	9.45 E − 12
Cell cycle (GO:0007049)	2.5% (49)	1.0% (240)	2.51	1.49 E − 09
Intracellular signaling cascade (GO:0007242)	4.6% (90)	2.5% (609)	1.82	2.00 E − 08
Endocytosis (GO:0006897)	1.4% (27)	0.4% (108)	3.08	9.71 E − 08
Mitosis (GO:0007067)	1.3% (26)	0.4% (103)	3.11	1.35 E − 07
Perception of sound (GO:0007605)	1.2% (23)	0.4% (87)	3.26	2.87 E − 07
Morphogenesis (GO:0009653)	1.2% (23)	0.4% (107)	2.65	1.31 E − 05
Frizzled signaling pathway (GO:0007222)	0.5% (10)	0.1% (26)	4.74	1.92 E − 05
Negative regulation of transcription from Pol II promoter (GO:0000122)	0.9% (18)	0.3% (74)	3.00	1.93 E − 05
The top 15 most enriched functions in SUMO substrates				
DNA binding (GO:0003677)	27.1% (530)	9.4% (2255)	2.89	1.00 E − 126
Transcription factor activity (GO:0003700)	15.5% (304)	4.6% (1102)	3.40	3.64 E − 87
Nucleic acid binding (GO:0003676)	14.2% (277)	7.6% (1823)	1.87	7.89 E − 26
Zinc ion binding (GO:0008270)	14.6% (285)	8.2% (1968)	1.78	2.80 E − 23
Protein serine/threonine kinase activity (GO:0004674)	6.1% (119)	2.3% (559)	2.62	7.18 E − 23
Actin binding (GO:0003779)	3.7% (72)	1.1% (259)	3.42	4.25 E − 21
ATP binding (GO:0005524)	13.3% (260)	8.0% (1925)	1.66	3.69 E − 17
Protein kinase activity (GO:0004672)	6.5% (128)	3.2% (776)	2.03	6.38 E − 15
RNA polymerase II transcription factor activity (GO:0003702)	2.1% (41)	0.6% (138)	3.66	1.12 E − 13
Steroid hormone receptor activity (GO:0003707)	1.5% (29)	0.3% (75)	4.76	2.47 E − 13
GTPase activator activity (GO:0005096)	1.8% (35)	0.5% (110)	3.92	7.49 E − 13
Transcription coactivator activity (GO:0003713)	2.2% (43)	0.7% (158)	3.35	8.04 E − 13
Ligand-dependent nuclear receptor activity (GO:0004879)	1.5% (29)	0.3% (79)	4.52	1.17 E − 12
Protein binding (GO:0005515)	11.9% (233)	8.0% (1907)	1.50	7.58 E − 11
Calmodulin binding (GO:0005516)	1.8% (35)	0.5% (132)	3.27	2.31 E − 10

分析工具

Tool	Scope of the analysis	Level of abstraction	User interface	Application type	Platform	Supported input IDs
Onto-Express	All GO categories	Fully flexible; different levels of abstractions in different GO subtrees	Java GUI	Web-based	Any	GenBank, UniGene, Entrez Gene, Affymetrix, Gene symbol
GoMiner	All GO categories	Static global analysis	Java GUI	Stand-alone	Windows only	Organism specific IDs in GO
DAVID	All GO categories	Only lowest level of GO	HTML GUI	Web-based	Any	GenBank, UniGene, Entrez Gene, Affymetrix, RefSeq, UniProt, PIR
EASEonline	All GO categories	User-selected, fixed level	HTML GUI	Both	Any	Affymetrix, GenBank, UniGene, Entrez Gene
GeneMerge	One category	Only lowest level of GO	HTML GUI	Both	Any	Only supports organism specific IDs used in GO
FuncAssociate	All GO categories	Only lowest level of GO	HTML GUI	Web-based	Any	MODB gene products
GOTM	All GO categories	Only lowest level of GO	HTML GUI	Web-based	Any	Affymetrix, UniGene, ENSEMBL, Swiss-Prot, Entrez Gene
FatiGO	One category	User-selected, fixed level and static global analysis	HTML GUI	Web-based	Any	Affymetrix, GenBank
CLENCH	All GO categories	Static global analysis	Command-line input, HTML output	Stand-alone	Windows only	*A.thaliana* MIPS IDs
GOstat	All GO categories	User-selected, fixed level	HTML GUI	Web-based	Any	GenBank, UniGene, Gene symbol, Organism specific IDs in GO
GOToolBox	All GO categories	User-selected, fixed level	HTML GUI	Web-based	Any	Only organism specific IDs in GO
GoSurfer	All GO categories	Only lowest level of GO	C/C++ GUI	Stand-alone	Windows only	Affymetrix, UniGene, Entrez Gene
Ontology Traverser	One category	Only lowest level of GO	HTML GUI	Web-based	Any	Affymetrix
eGOn	One category	Only lowest level of GO	HTML GUI	Web-based	Any	GenBank, UniGene, Clone