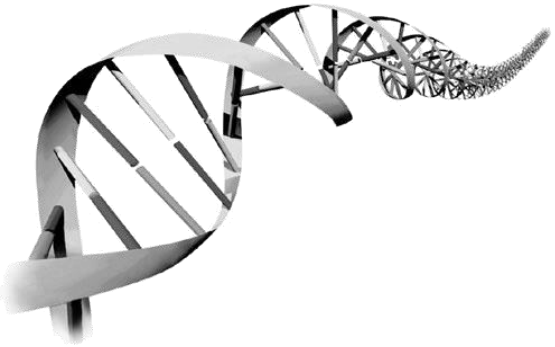


生物信息学

第十一章 转录组分析

转录组与转录调控

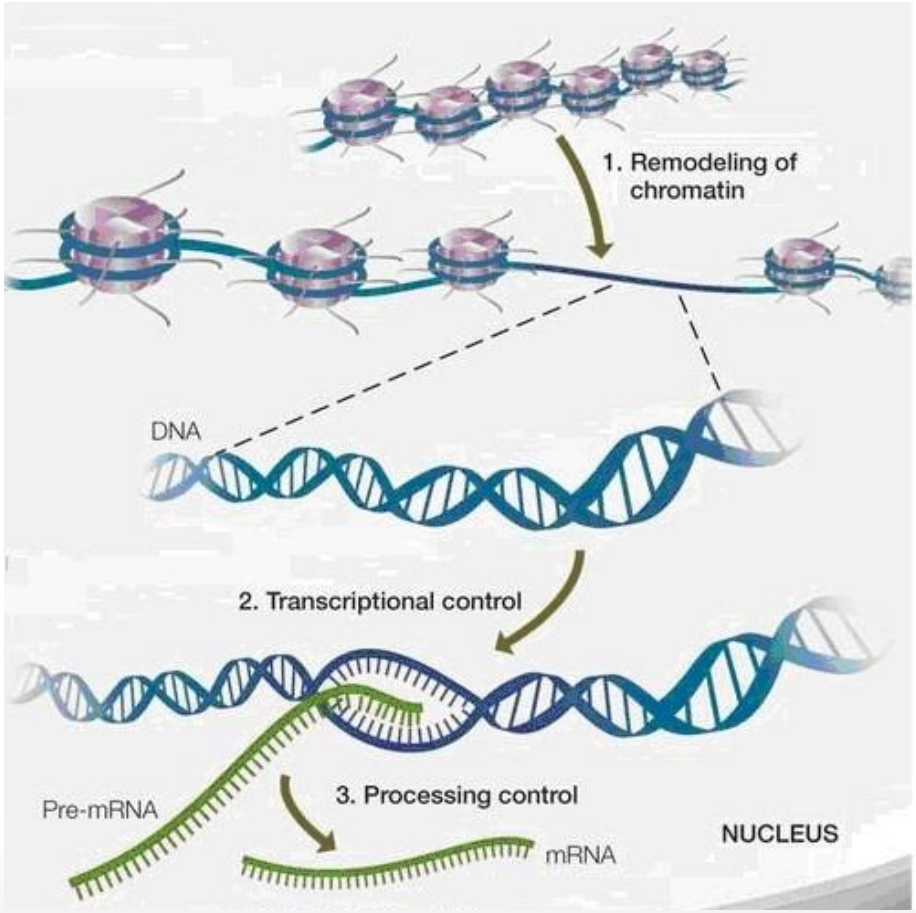
- 基因在多个层次上受到调控
- 转录组 (Transcriptome)
  - 细胞内所有RNA分子，包括mRNA, rRNA, tRNA和其他非编码 (non-coding) RNA
- 转录调控 (Transcriptional regulation)
  - 基因调控 (Gene regulation)

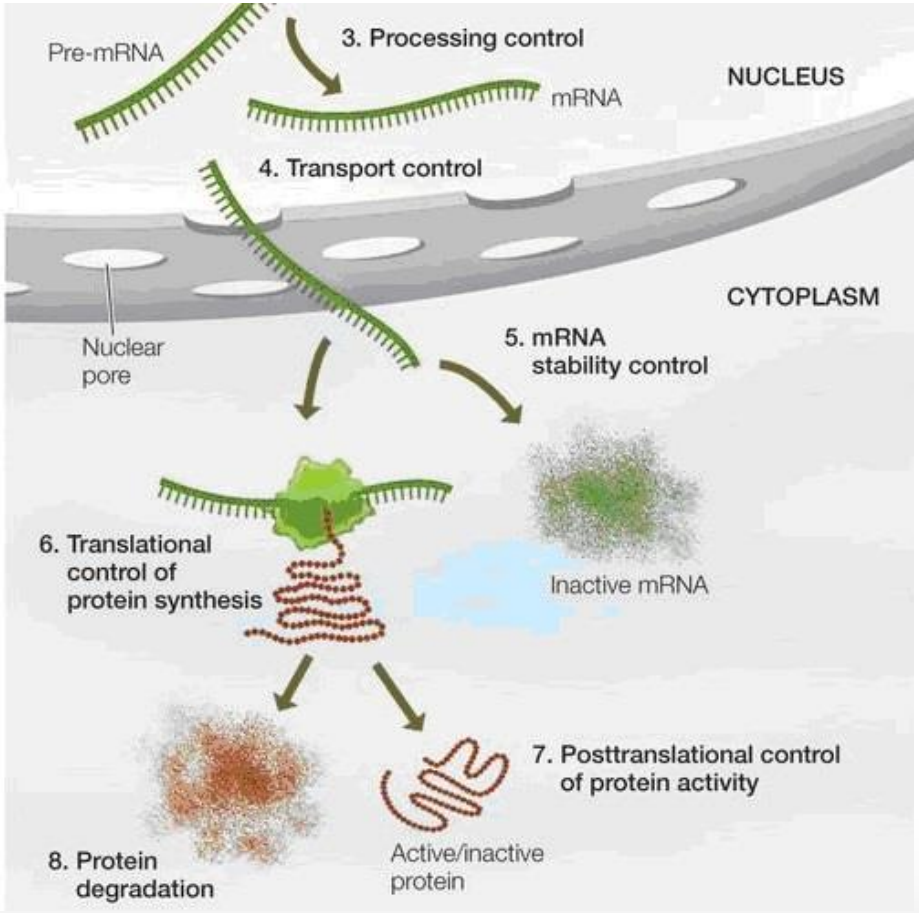
- 转录 (Transcription)
- 转录后 (Post transcription)：RNA稳定性

} “转录组”

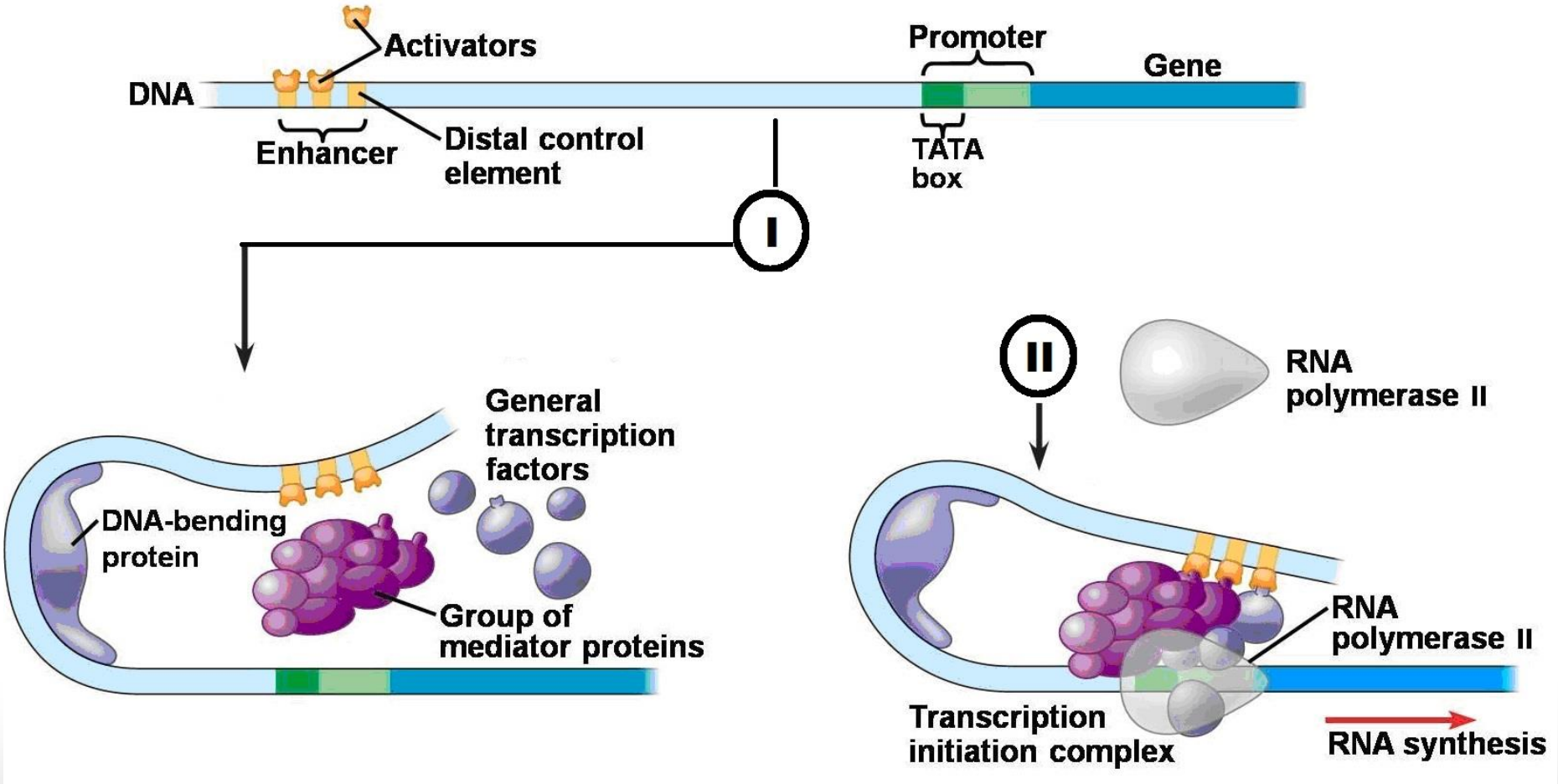
- 翻译 (Translation)
- 翻译后 (Post translation)

真核生物基因表达的基本方式

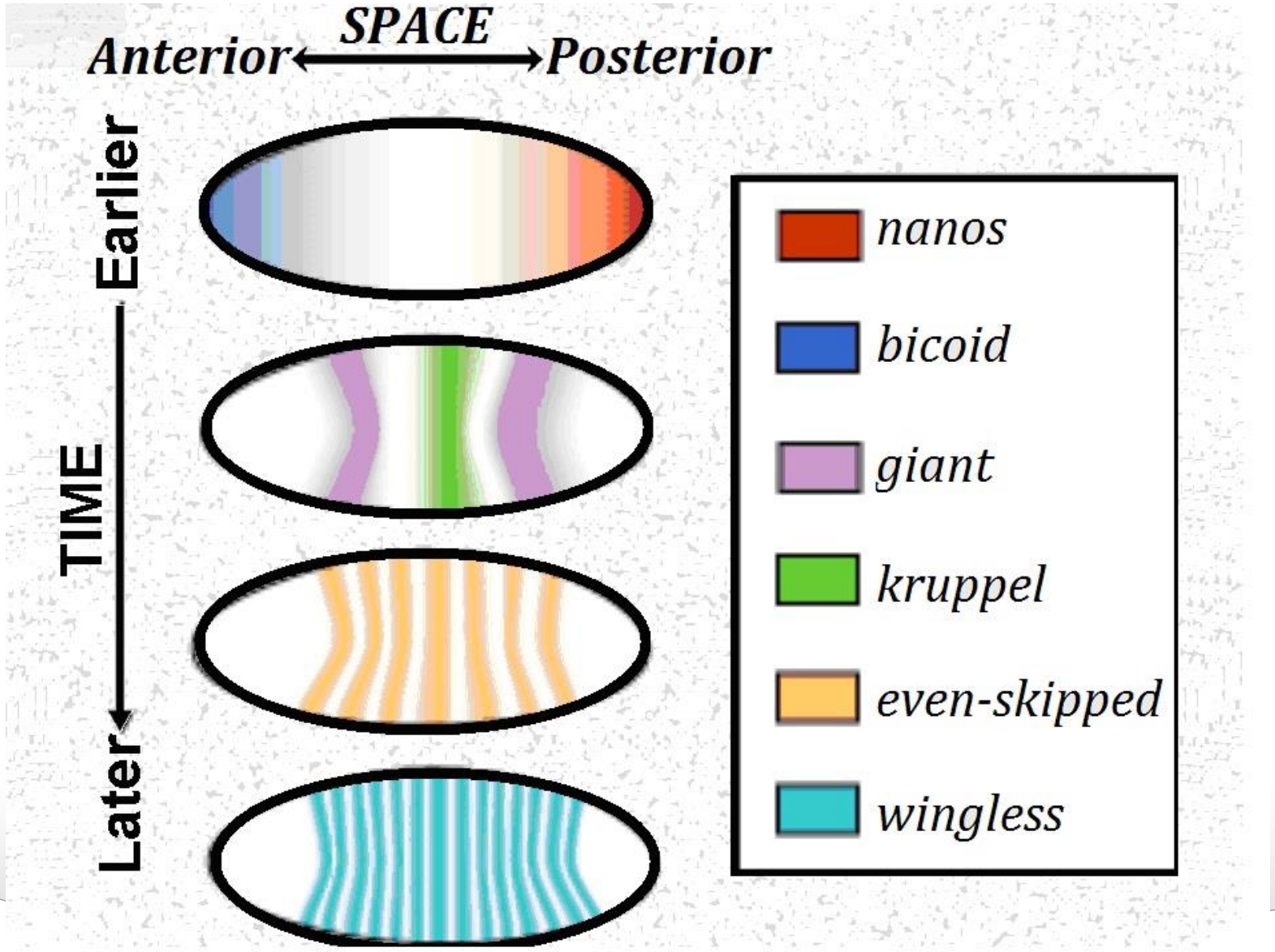




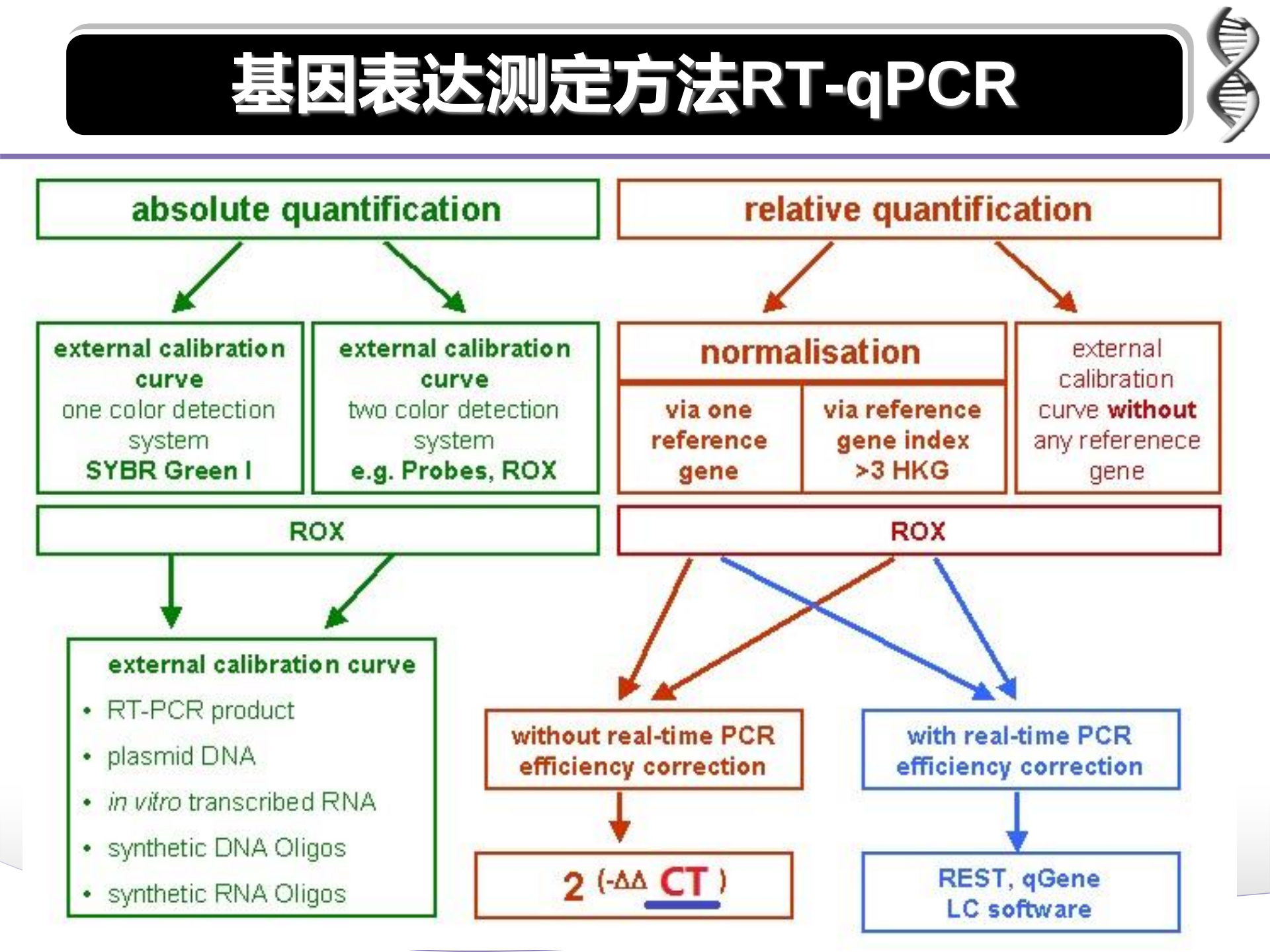
基因表达调控示意图


Activators
DNA
Promoter
Gene
Enhancer
Distal control element
TATA box
I
II
RNA polymerase II
General transcription factors
DNA-bending protein
Group of mediator proteins
RNA polymerase II
Transcription initiation complex
RNA synthesis

基因表达的时空性


SPACE
Anterior
Posterior
TIME
Earlier
Later
nanos
bicoid
giant
kruppel
even-skipped
wingless

基因表达测定方法RT-qPCR



基因表达 (Gene expression) 分析

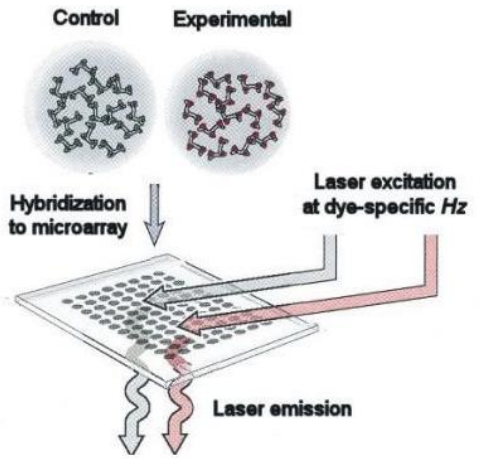
- 快照 (Snapshot)
  - 所有基因的RNA表达水平
  - 提供大量的数据
- 发现在特定生长时期，或者随着环境变化，哪些基因的表达上调或者下调
- 在相同条件下，上调或者下调变化规律相似的基因，可能具有功能上的关联
- 可以从共表达的基因中寻找调控模块
- 基因表达的模式可以用来表征异常的细胞调控
  - 癌症的诊断

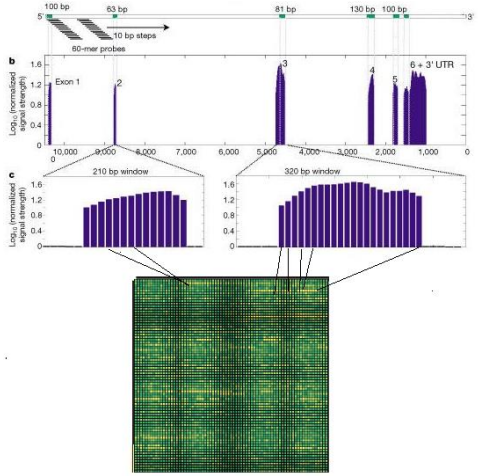
转录组与转录调控的主要研究技术

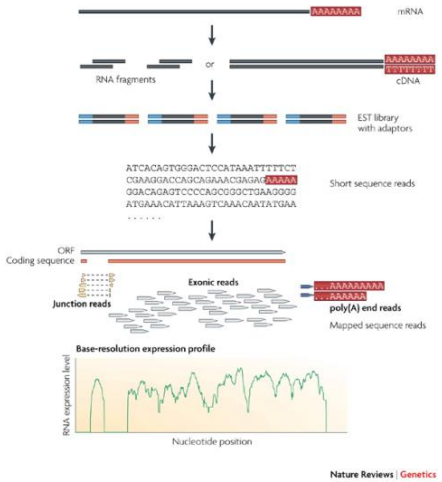
- **转录组**
  - **单个实验中检测整个转录组**
  - **高通量测序：RNA-seq**
  - **基于DNA杂交：微阵列/基因芯片 (Microarray)**
- **转录调控：**
  - **蛋白质-DNA的相互作用关系**
  - **高通量测序：ChIP-seq**
  - **芯片：ChIP-chip**

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 RNA-seq技术的应用：利用下一代测序技术直接测定mRNA

RNA-seq

- 分析流程
  - 分离纯化所有的mRNA
  - 利用反转录酶将mRNA转变成cDNA
  - cDNA测序
  - 将序列回贴到参考序列上
- 序列检测到越多，则表达越高
- RNA-seq数据分析：**挑战！**
  - 读段回贴
  - 定量已知的基因
  - 发现新的转录本
  - 定量可变剪接异构体

RNA-seq数据：序列读段

FASTQ: Illumina测序读段文件

```
Line 1: @EAS042_0001:1:1:1061:20798#0/1
Line 2: TNTCTGTGTCCTGGGGCATCAATGATAGTCACATAGTACTTGCTGGTCTCAAATTTCCACAAGGAGATATCAATGG
Line 3: +EAS042_0001:1:1:1061:20798#0/1
Line 4: aB\^^Y]a^]cde`daaYaaa_bc\\`b^Y\a\aaUQY\]a\`aa\W__]HVZ]VQF^[`UH]\J^F^T^\\I]__
```

Line 1

EAS042_0001	the unique instrument name
1	flowcell lane
1	tile number within the flowcell lane
1061	'x'-coordinate of the cluster within the tile
20798	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

Line 2: 原始序列

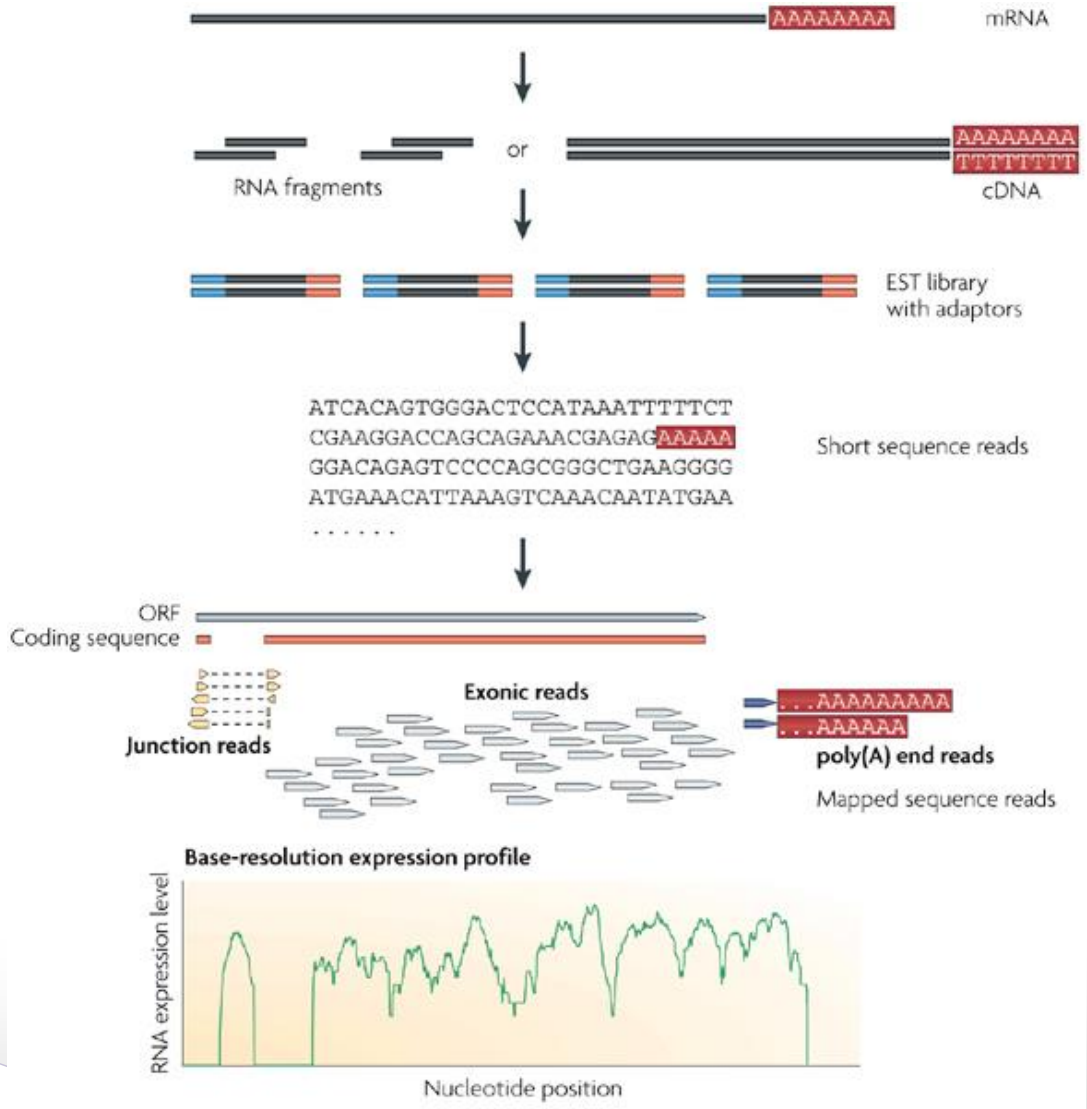
Line 3: + ?

Line 4: 序列的质量分值

-5 ~ 62

使用 ASCII 59 ~ 126

RNA-seq分析流程



Poly(A) end-reads

样品准备

高通量测序

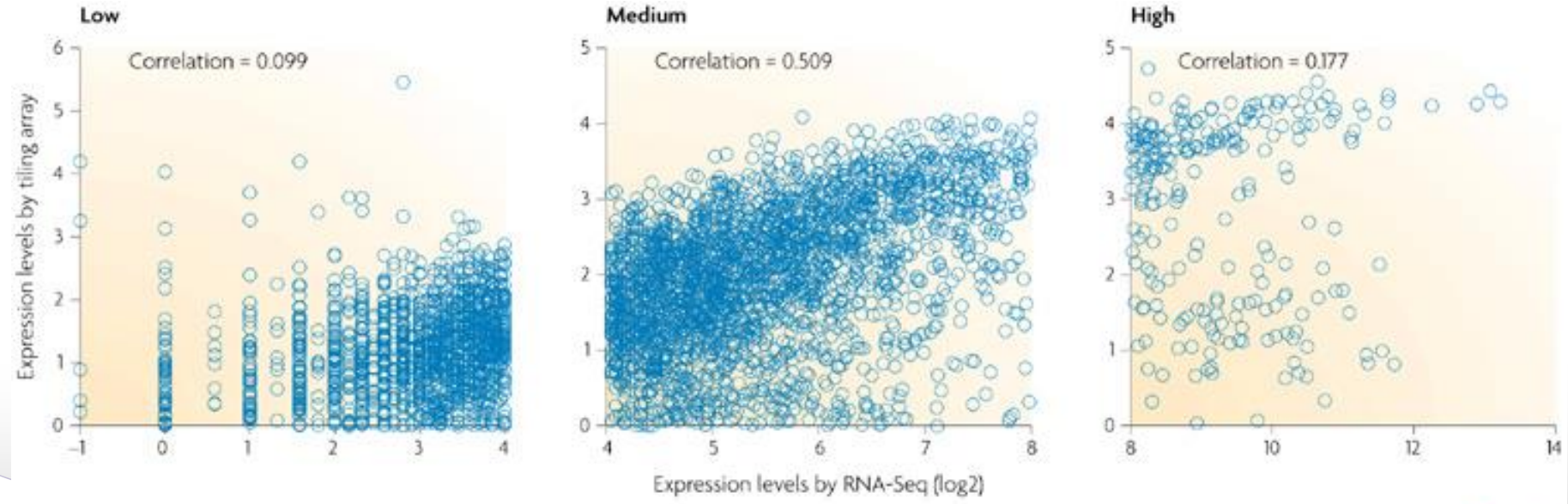
数据分析:
- ✓**读段回贴**
- ✓**可视化**
- ✓**重头组装**
- ✓**定量**

RNA-seq *vs.* DNA芯片

- RNA-seq可以发现新的转录本和可变剪接异构体，DNA芯片只能检测已知转录本的表达
- RNA-seq比全基因组铺瓦芯片的解析度更高：单个碱基
- RNA-seq相同的实验流程可以做不同的分析
  - 单核苷酸多态性 (SNP芯片)
  - 外显子连接点 (连接点芯片)
  - 基因融合 (基因融合芯片)

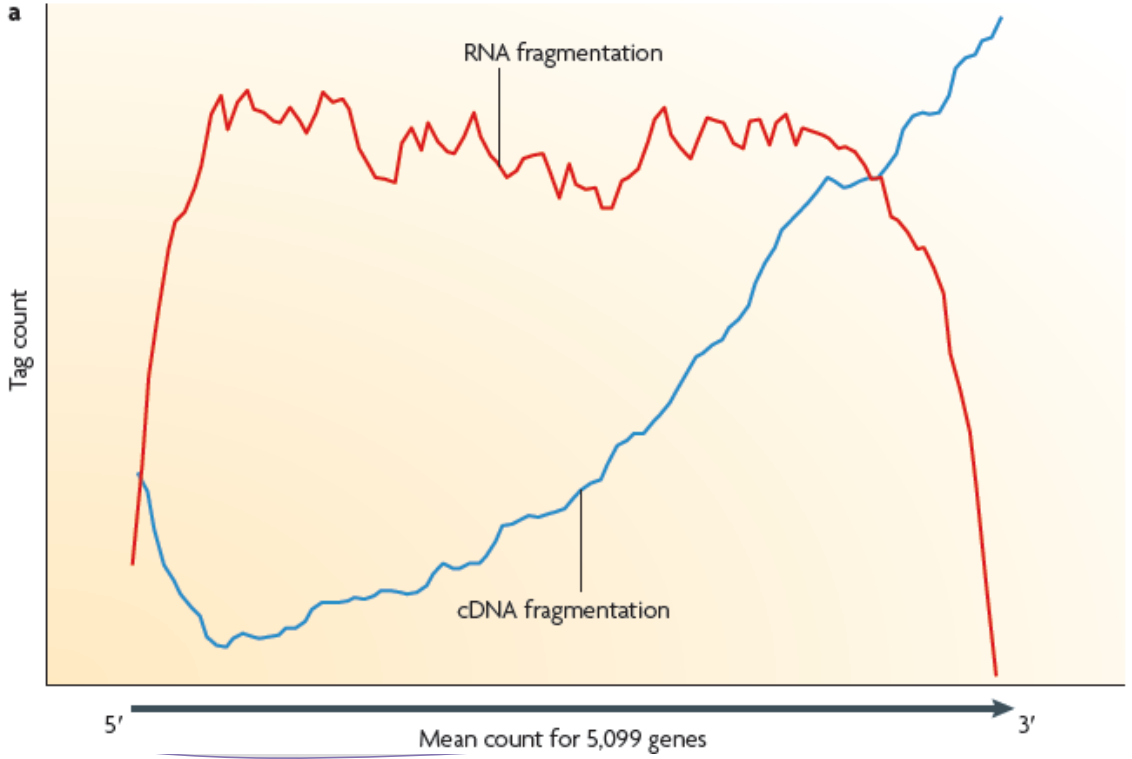
RNA-seq *vs.* DNA芯片

- 营养丰富的培养基里的芽殖酵母
- 仅在中等表达时，两者有较高一致性
- 基因表达低或者高的时候，关联性差
  - DNA芯片的对基因表达低或高的检测不够敏感



RNA-seq的问题：建库

- mRNA与cDNA碎裂的不同
- 利用oligo-dT引物得到的cDNA更倾向于获得转录本的3’端
- RNA碎裂在5’和3’端都较少



RNA-seq数据的比对

❒ 读段回贴的常用软件

软件名称	比对方法	备注
Bowtie	Burrows-Wheeler转换	整合质量得分
BWA	Burrows-Wheeler转换	整合质量得分
Stampy	种子匹配方法	概率模型
SHRiMP	种子匹配方法	Smith-Waterman的扩展
TopHat	Exon-first方法	利用Bowtie比对
MapSplice	Exon-first方法	与多种Unspliced aligners共同运行

转录组的重建

❒ 转录本组装的软件

软件名称	优点	输入	输出
Cufflinks	参考基因组引导装配，可以识别基因的新转录本	比对到参考基因组的reads	转录本结构及表达
Scripture	参考基因组引导装配，可以识别基因的新转录本	比对到参考基因组的reads	转录本结构及表达
TransABySS	不需要参考基因组，可以识别新的基因和新的转录本	测序得到的原始reads	转录本结构及表达
Trinity	不需要参考基因组，可以识别新的基因和新的转录本	测序得到的原始reads	转录本结构及表达

RNA-seq数据的归一化

- **RPM (Reads per million mapped reads)**

$$RPM\ of\ a\ gene = \frac{Number\ of\ reads\ mapped\ to\ a\ gene \times 10^6}{Total\ number\ of\ mapped\ reads\ from\ given\ library}$$

$$RPM\ of\ a\ gene = \frac{5000 \times 10^6}{4 \times 10^6} = 1250$$

- **RPKM (Reads per kilo base per million mapped reads)**

$$RPKM\ of\ a\ gene = \frac{Number\ of\ reads\ mapped\ to\ a\ gene \times 10^3 \times 10^6}{Total\ number\ of\ mapped\ reads\ from\ given\ library \times gene\ length\ in\ bp}$$

$$RPKM\ of\ a\ gene = \frac{5000 \times 10^3 \times 10^6}{4 \times 10^6 \times 2000} = 625$$

- **TPM (Transcript per million)**

$$A \cdot \frac{1}{\sum(A)} \cdot 10^6 Where\ A = \frac{total\ reads\ mapped\ to\ gene \cdot 10^3}{gene\ length\ in\ bp}$$

基因表达的数据分析

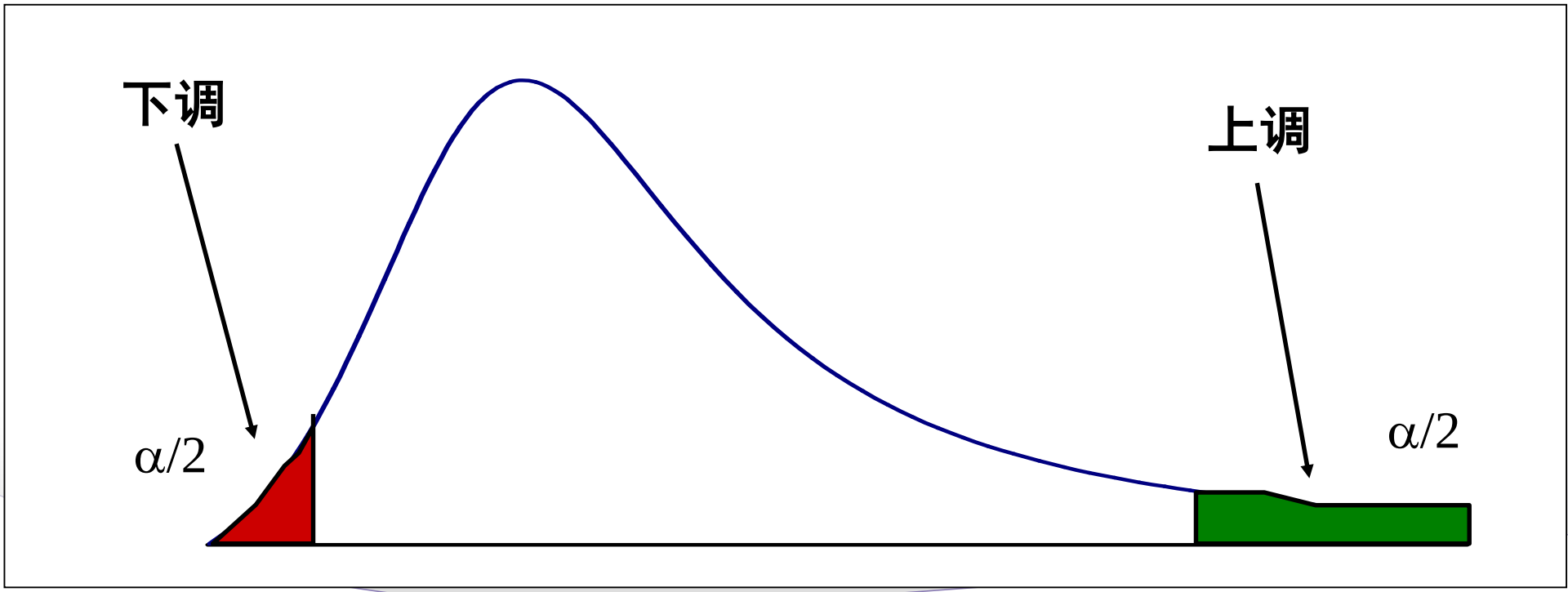
- 差异表达基因分析
- 基因共表达分析
- 基因表达数据的聚类和分类
- 基因集分析
- 生物学通路与网络分析
- RNA可变剪接

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

- 倍数法：通常以2倍差异为阈值，判断基因是否差异表达
- *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RNA-seq实验中，每一个基因都是一个独立试验
- RNA-seq：高通量，>10,000个基因
- 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：10,000个基因RNA-Seq数据，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: $0.01/10{,}000 = 1*10^{-6}$

- 若差异表达基因接受*p*-value < 0.01为显著
- 则10,000个基因的多重检验可设置*p*-value < $1*10^{-6}$ 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

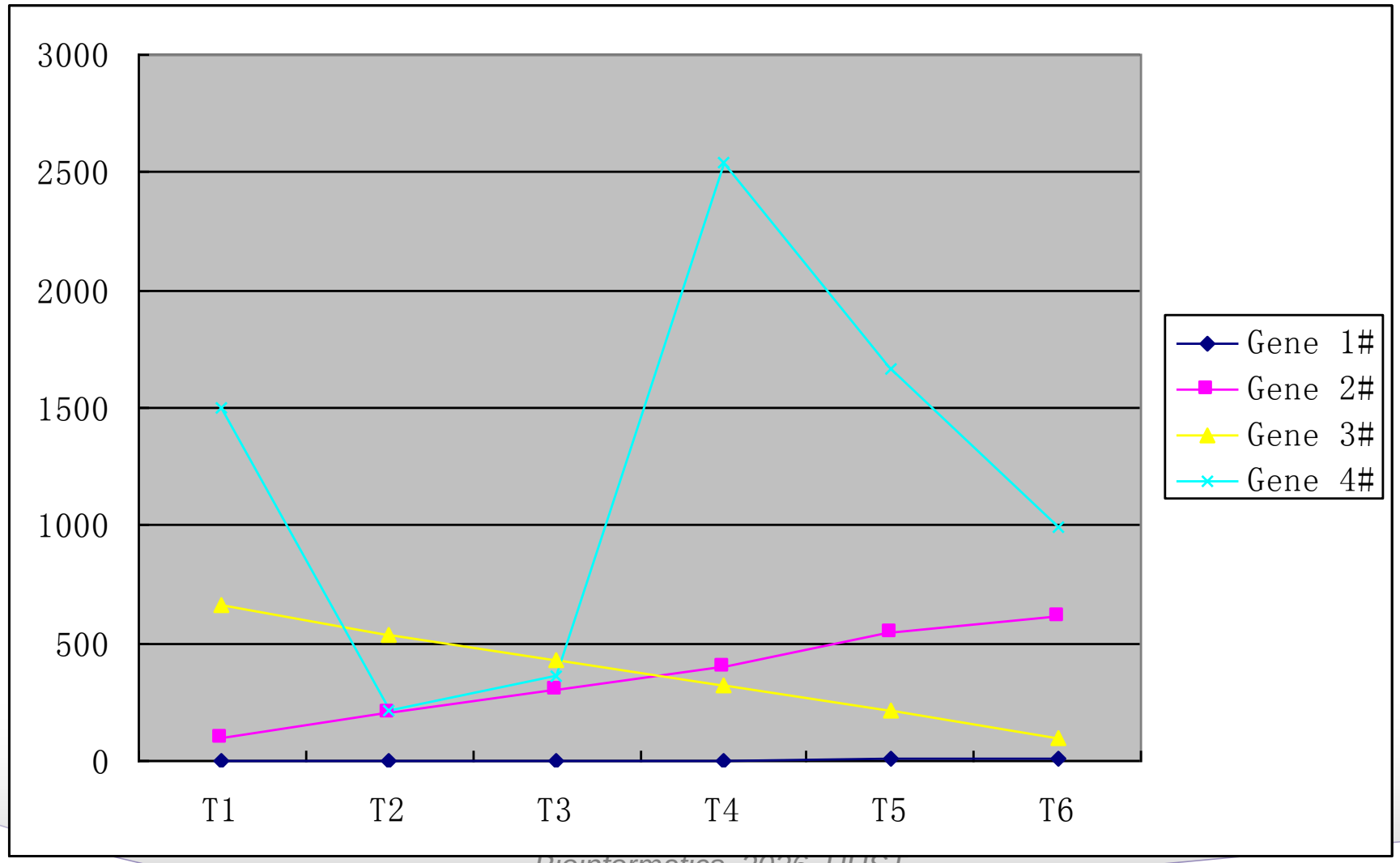
基因共表达分析

- 在N个不同的条件下 (时间序列的表达数据)，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

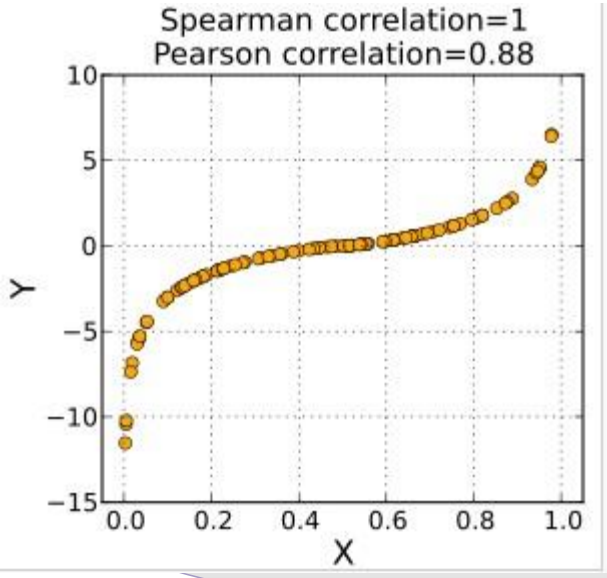
$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$

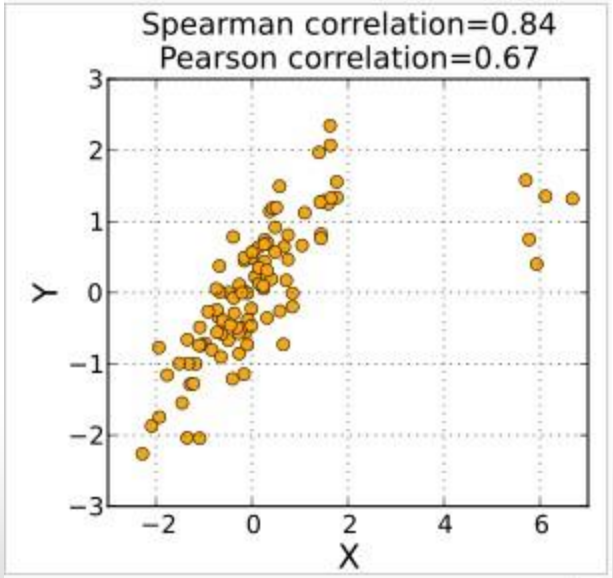
	Gene 1#	Gene 2#	Gene 3#
Gene 1#			
Gene 2#	0.996368		
Gene 3#	-0.99988	-0.99611	
Gene 4#	0.245292	0.254855	-0.2395

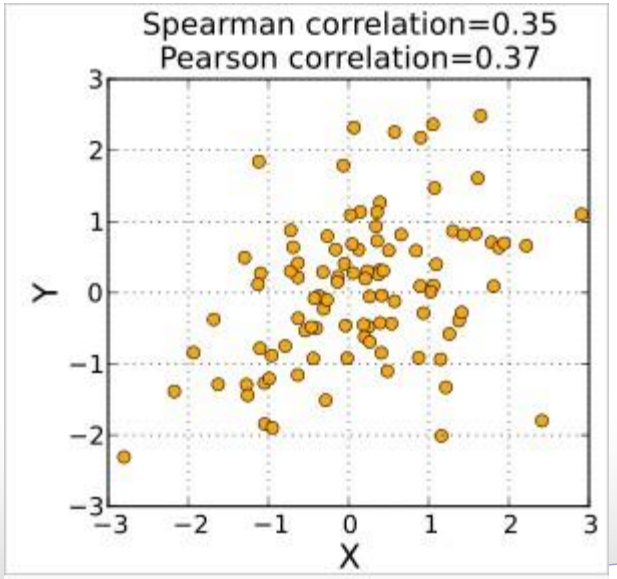
- 结论：**Gene 1#**与Gene 2#表达正相关，与Gene 3#表达负相关，与Gene 4#无关联

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
- 发现新的模式
- 聚类方法：
  - 层次聚类
  - *K*均值聚类
  - 高斯混合聚类
  - 密度聚类

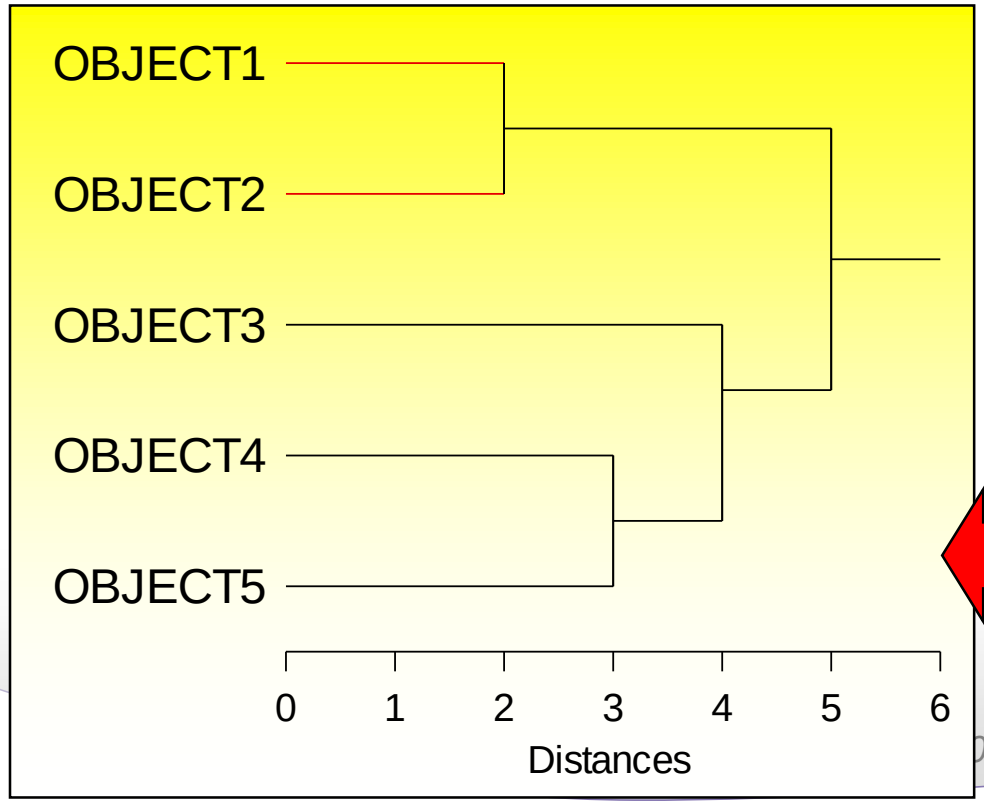
层次聚类

- 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	

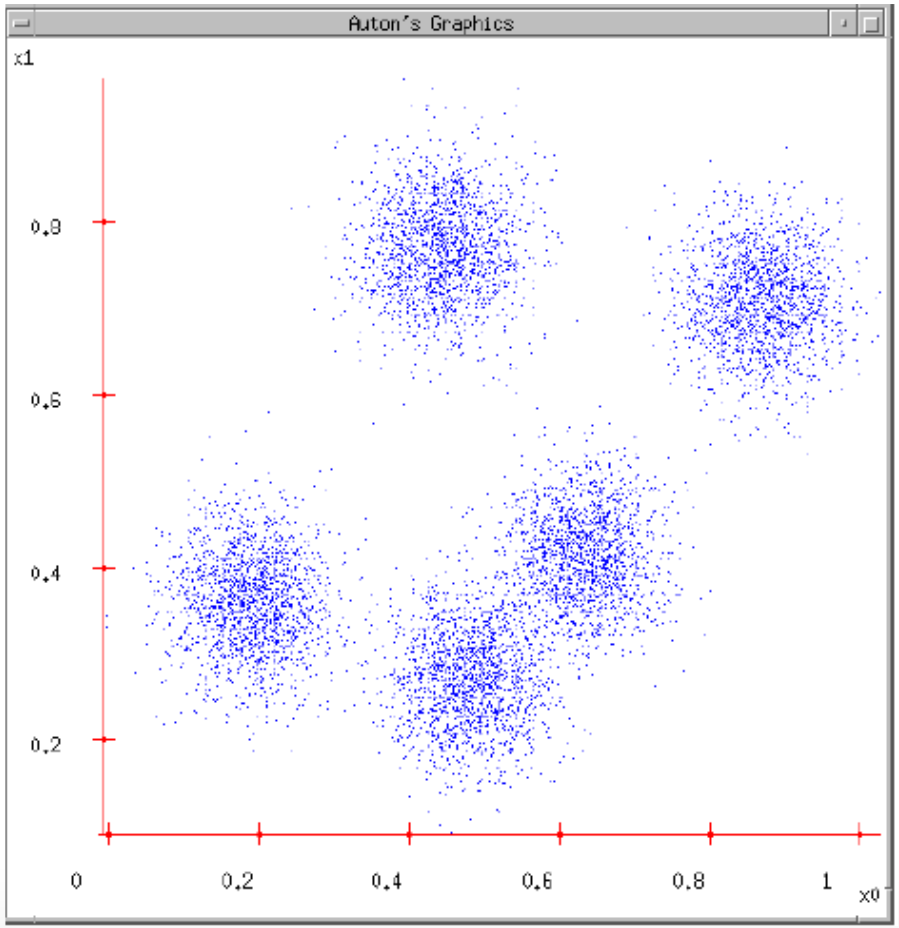
Distance matrix

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)



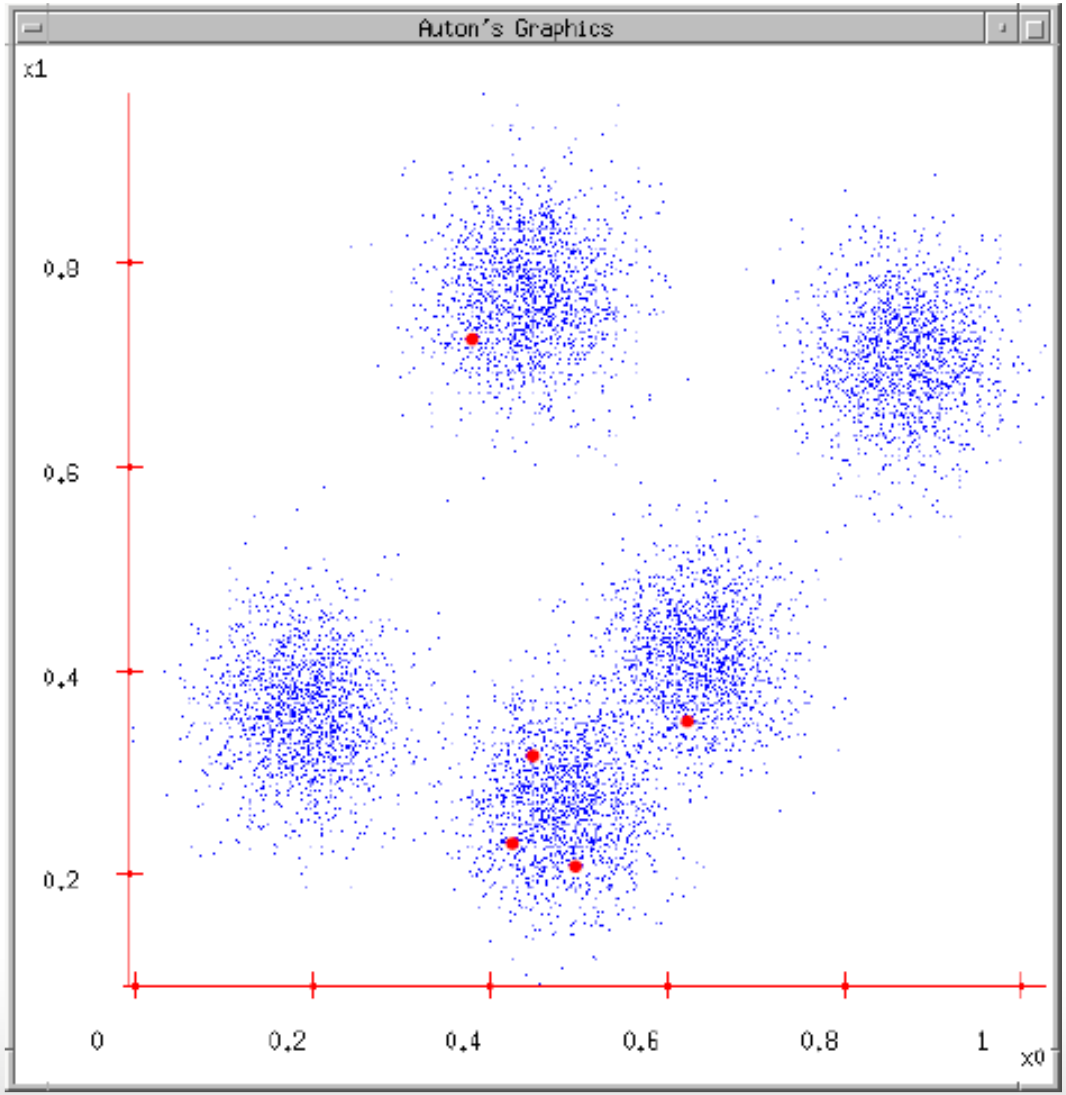
K均值聚类

- 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



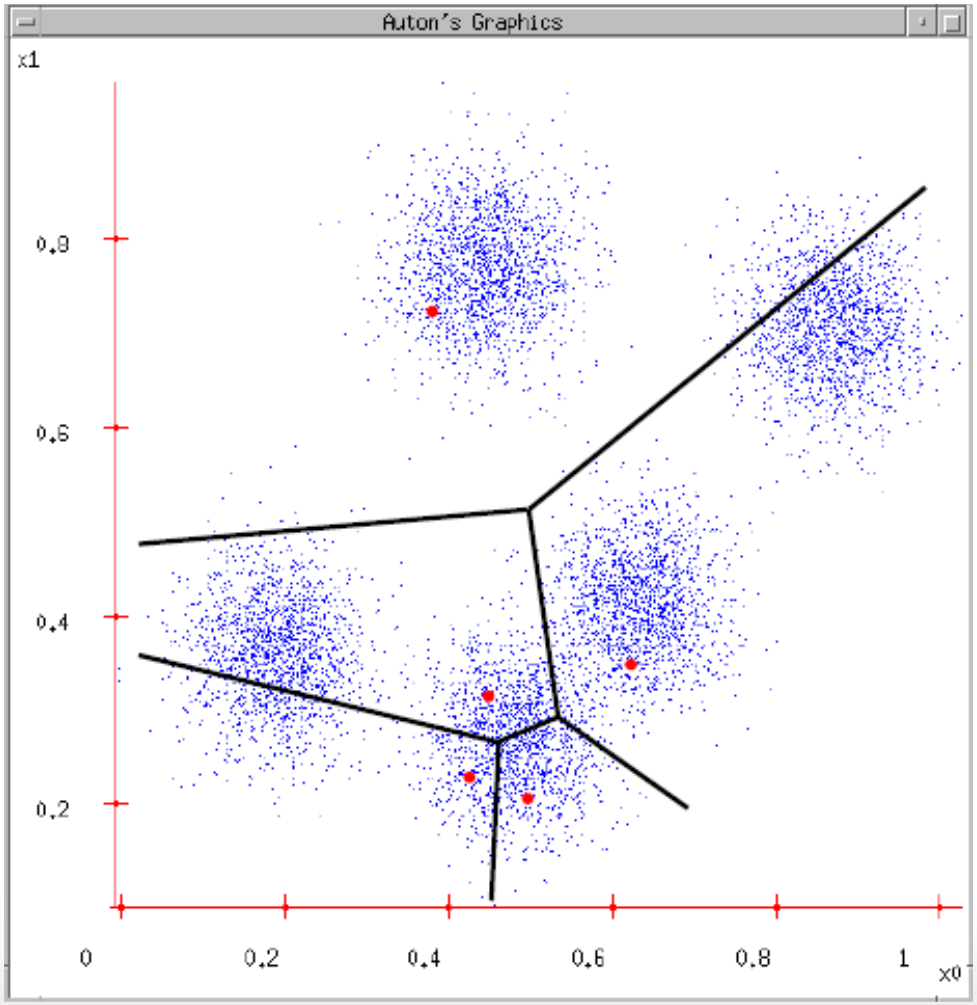
K均值聚类

- 随机选取5个点，作为每一个类的中心点



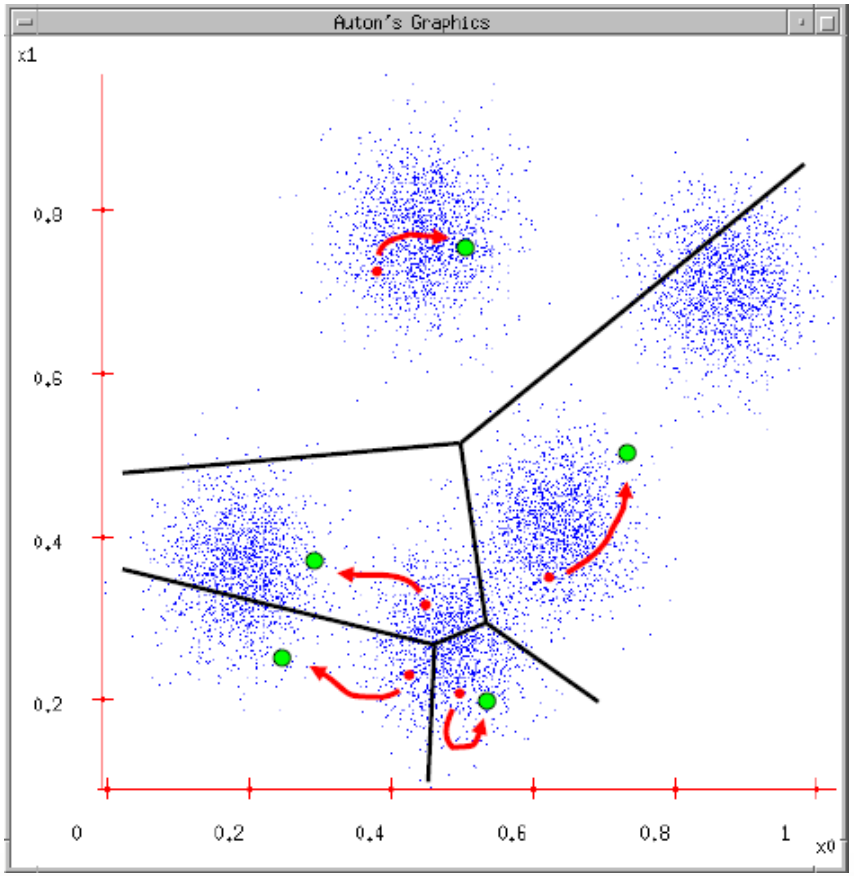
K均值聚类

- 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



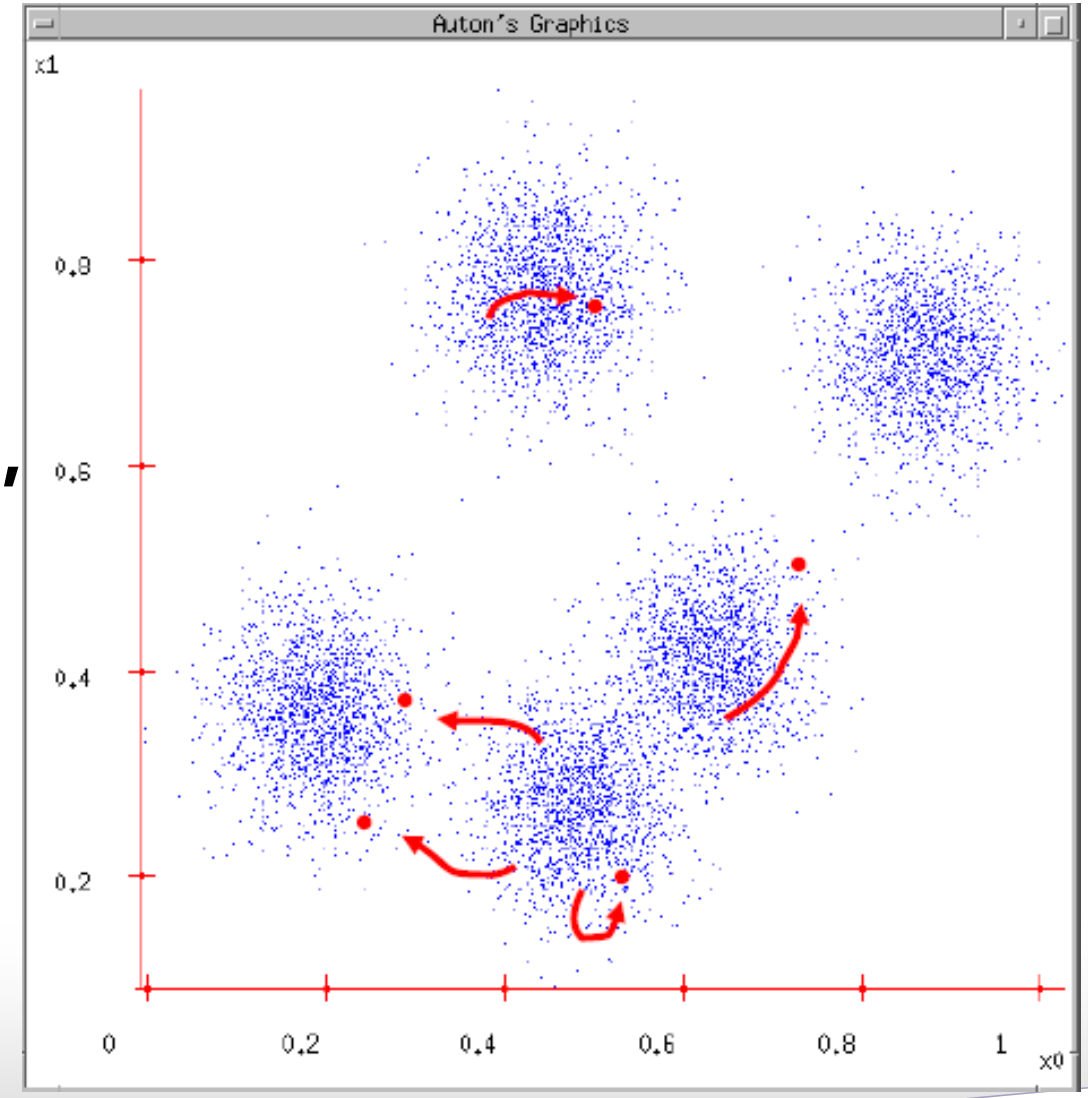
K均值聚类

- 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！



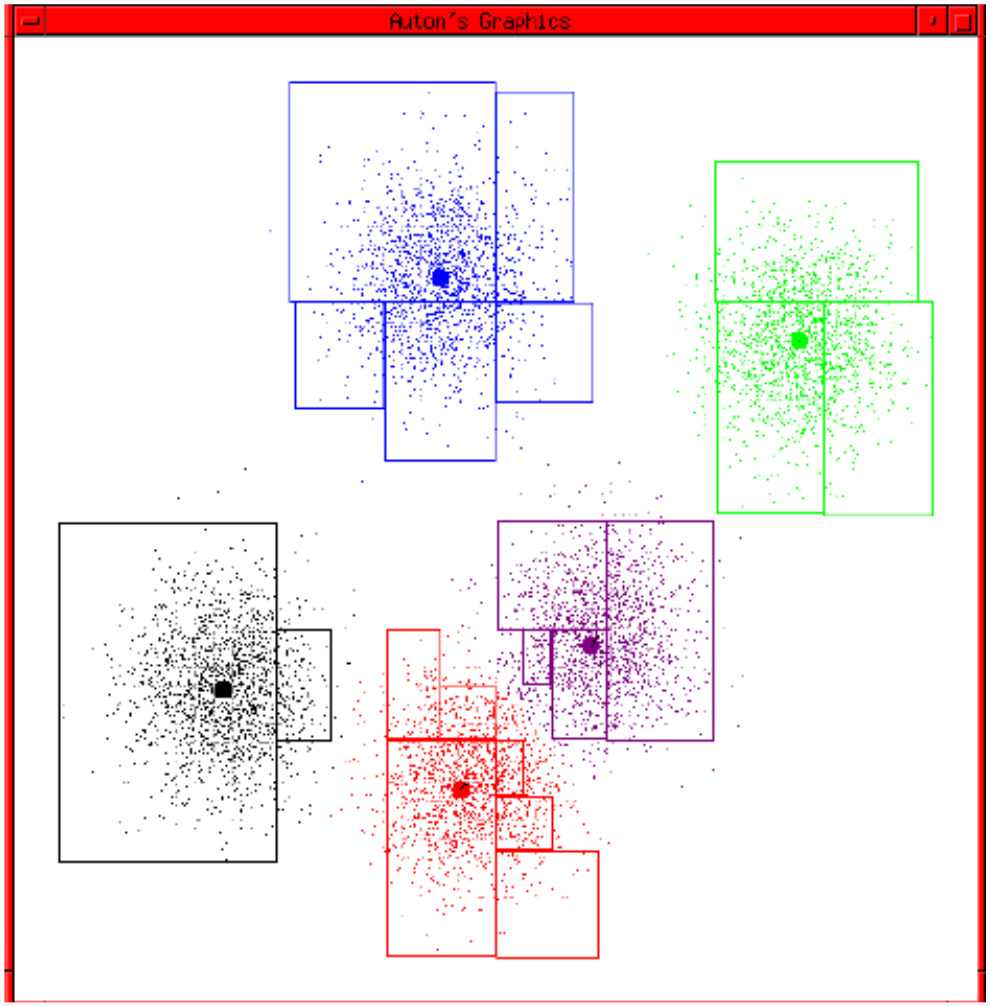
K均值聚类

- 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均值聚类

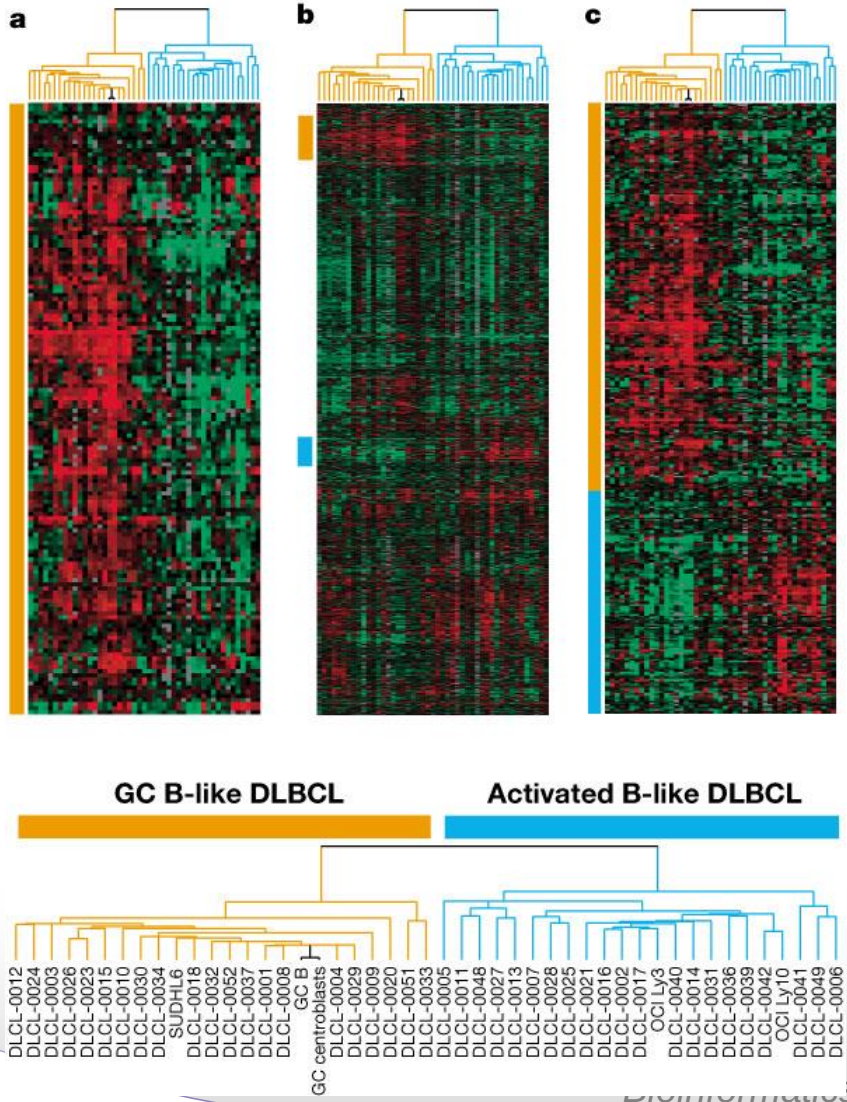
- 最终结果：所有基因芯片数据被聚成5类
- 软件：Cluster 3.0, Michael Eissen, Stanford



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)：根据发现的模式进行预测
- 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

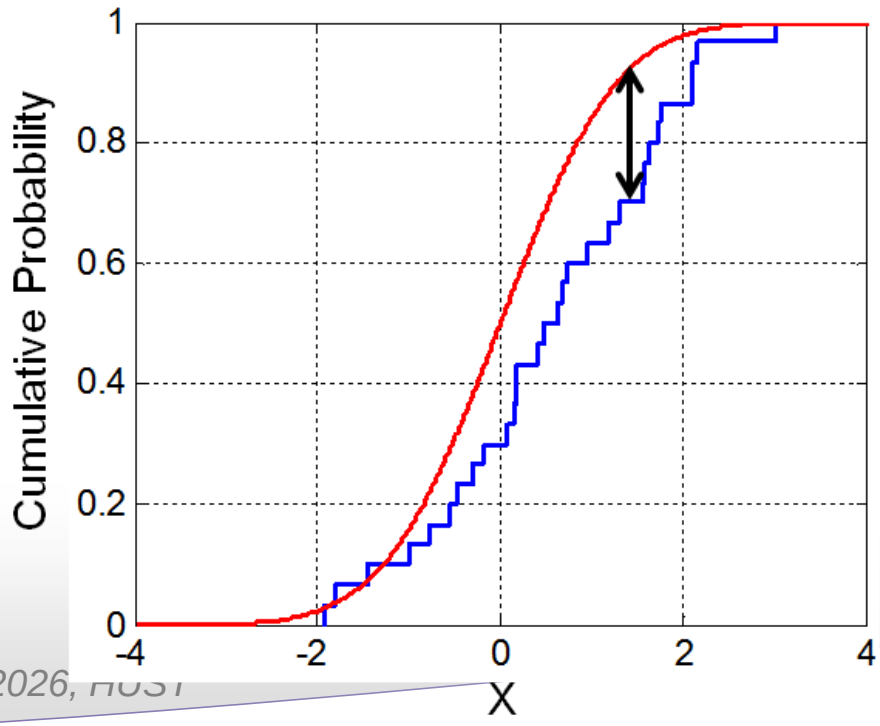
Description of GO term	Number of proteins annotated in group S[a]	Number of proteins annotated in group W[b]	Enrichment ratio	*P*-value
The top 15 most enriched processes in SUMO substrates				
Regulation of transcription, DNA-dependent (GO:0006355)	26.1% (510)	9.0% (2174)	2.89	6.12 E − 121
Transcription 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

基因集富集分析

- Gene Set Enrichment Analysis (GSEA)
- 考虑所有的基因而不仅是差异表达基因
- 根据差异表达水平将所有基因排序
- 发现基因集里的基因是否倾向于分布在所有基因列表的前部 (上调) 或后部 (下调)
- 计算流程:
  - 计算一个富集分值 (Enrichment score, ES)
  - 根据ES分值估计统计显著性
  - 针对多重检验进行校正

Kolmogorov–Smirnov检验

- GSEA, $p = 0$，不考虑p-value的权重
- 非参检验方法 (Nonparametric test)
- 检验分布是否符合正态分布
- 两样本的Kolmogorov–Smirnov检验
  - 差异最大的值算显著性



GSEA的输入

- 基因表达数据*D*，包括*N*个基因和*k*个样本
- 排序产生基因列表*L*
  - 上调-不变-下调
  - 差异表达基因的*p*-value
- 用指数*p*来控制每一步的权重
  - *p*=0，标准Kolmogorov–Smirnov统计
  - *p*=1，GSEA软件
- 独立的基因集*S*，如GO, KEGG
  - 包括N_H个基因

Enrichment Score ES(S)

- 基因集*D*排序为$L = \{g_1, \ldots, g_N\}$，$r(g_j) = r_j$
- 对基因列表*L*，计算每一个位置*i*以上，包括在基因集S中的 (“hits”) 部分，以及不在S中的 (“misses”) 部分

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}.$$

- $ES = P_{hit} - P_{miss}$的最大值

ES计算：实例

基因列表

Gene	E-ratio
1	5
2	4
3	3
4	2
5	1.5
6	0.04
7	0.007

基因集***S*= 1, 3, 4, 5**

$P_{hit}(S, 1) = 5/5=1$

$P_{miss}(S, 1) = 0$

ES(1) = 1

$P_{hit}(S, 2) = 5/5=1$

$P_{miss}(S, 2) =0.33$

ES(2) = 0.67

$P_{hit}(S, 3) = 1 + 3/(5+3) =1.375$

$P_{miss}(S, 3) =0.33$

ES(3) = 1.045

$P_{hit}(S, 4) = 1.375 + 2/(5+3+2)=1.575$

$P_{miss}(S, 4) = 0.33$

ES(4) = 1.245

ES计算：实例

基因列表

Gene	E-ratio
1	5
2	4
3	3
4	2
5	1.5
6	0.04
7	0.007

基因集***S*= 1, 3, 4, 5**

$P_{hit}(S, 5) = 1.575 + 1.5/(5+3+2+1.5)=1.705$

$P_{miss}(S, 5) = 0.33$

ES(5) = 1.375

$P_{hit}(S, 6) = 1.705$

$P_{miss}(S, 6) = 0.66$

ES(6) = 1.045

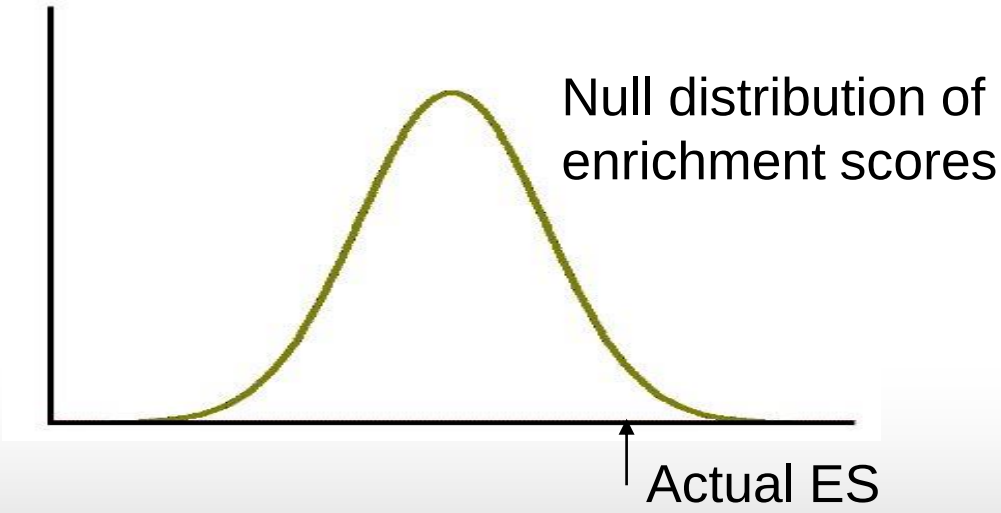
$P_{hit}(S, 7) = 1.705$

$P_{miss}(S, 7) = 1$

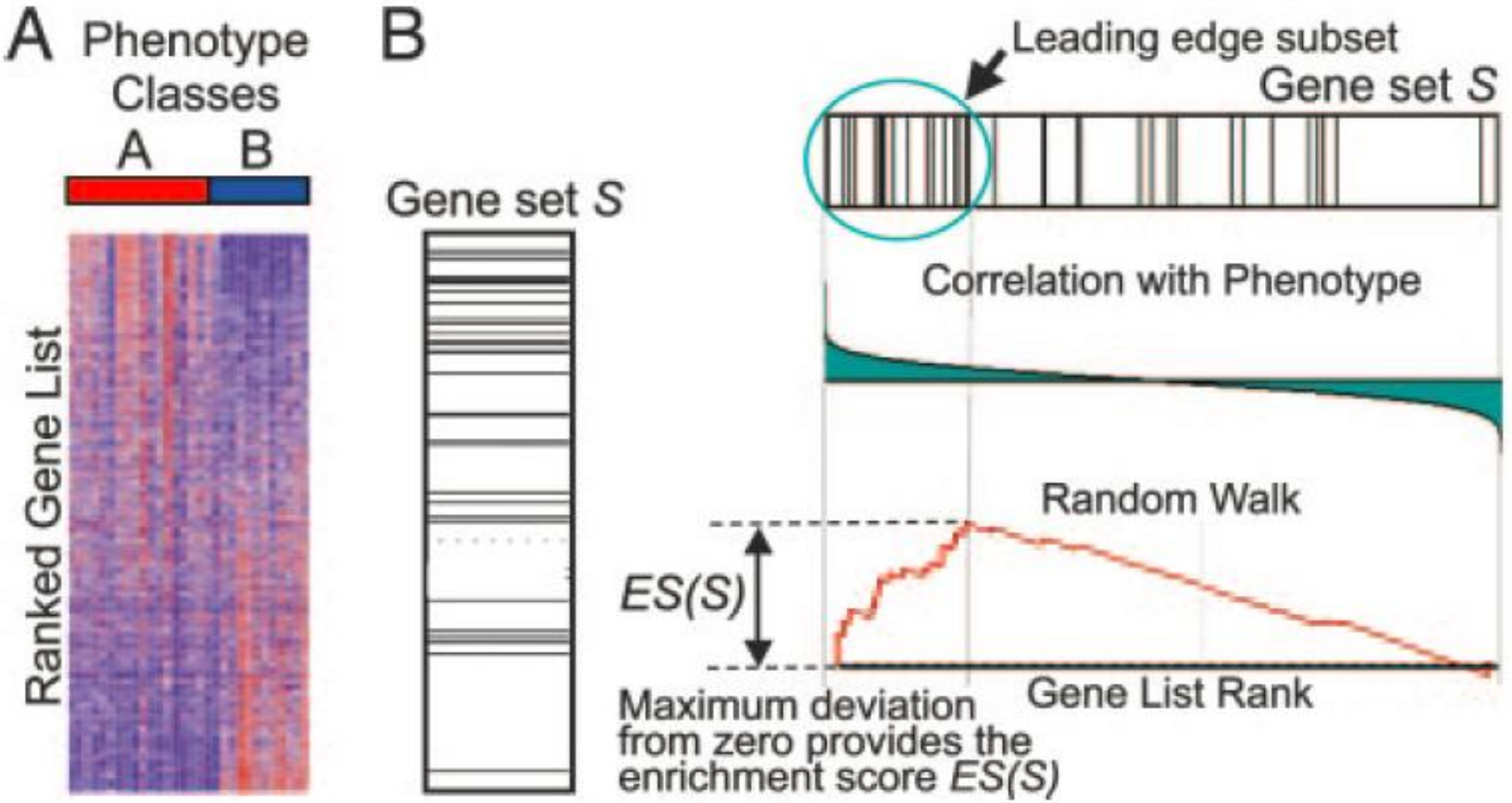
ES(7) = 0.705

显著性估算

- 计算ES的显著性：与ES_{NULL}比较
- 将原数据的*p*-value取出，随机分给每个基因，重新排序
- 重新计算ES值，重复1,000次置换 (Permutation)，建立ES_{NULL}分布的直方图
- 估算*p*-value = 大于或等于ES的个数/总数

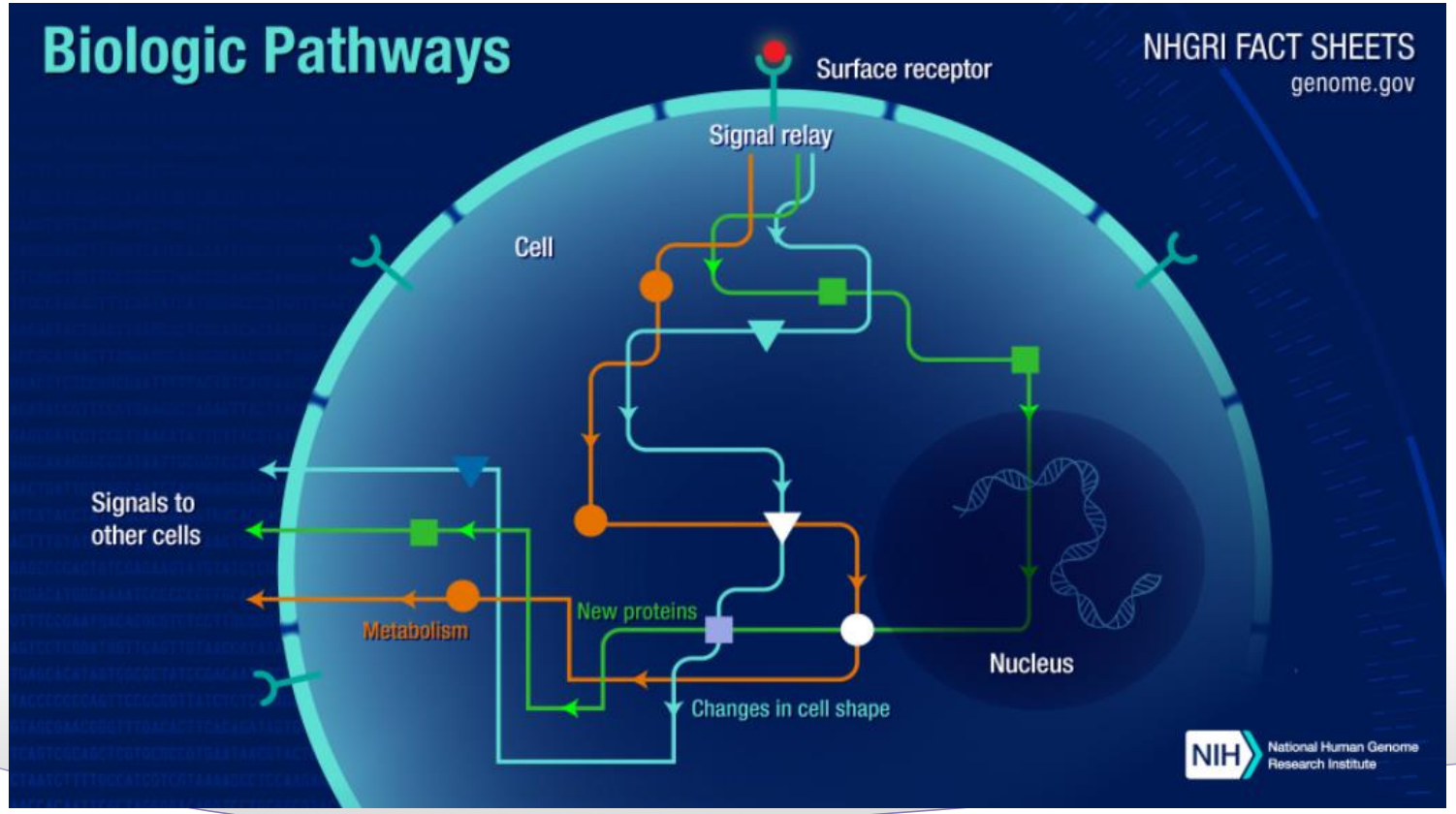


GSEA算法总结

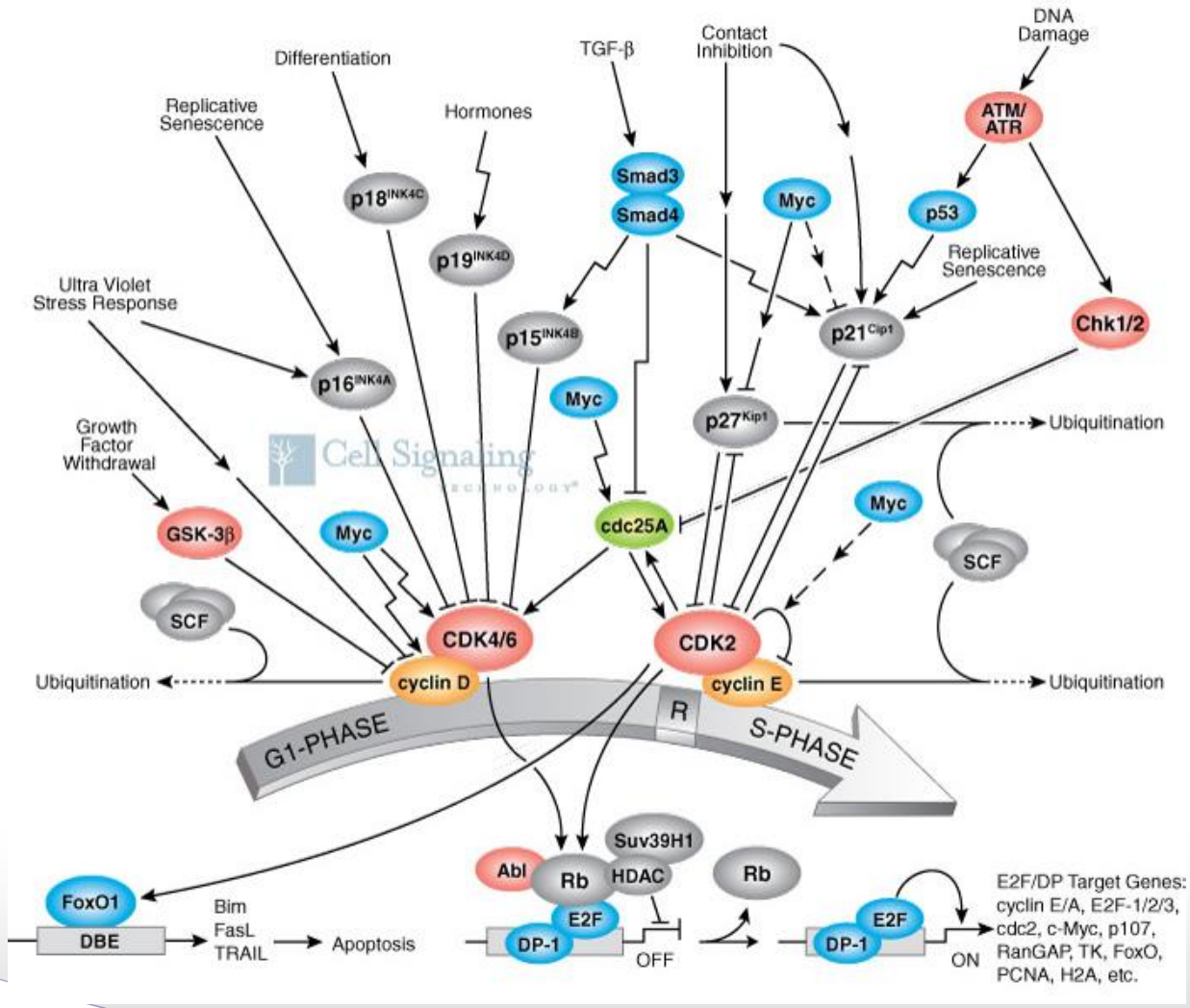

A
Phenotype
Classes
A
B
Ranked Gene List
B
Gene set S
Leading edge subset
Gene set S
Correlation with Phenotype
Random Walk
ES(S)
Gene List Rank
Maximum deviation
from zero provides the
enrichment score ES(S)

生物学通路

- 由生物体内一系列生物化学分子（包括基因，基因产物以及化合物等）通过各种生化级联反应来完成某一具体生物学功能的过程



细胞信号通路



G1/S检验点：有调控方向

生物通路富集分析算法

过表达分析法（ORA）

表达矩阵

单基因分析

基因得分

建立列联表

基因集合得分

显著性检验

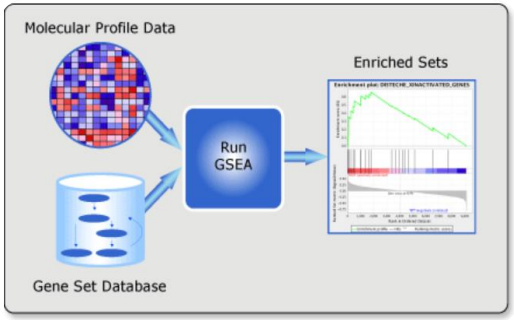
P值

多重检验校正

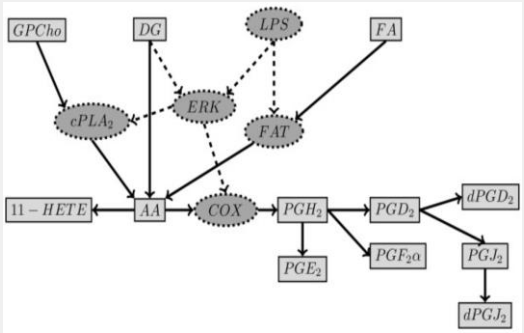
校正后的P值

基因集合

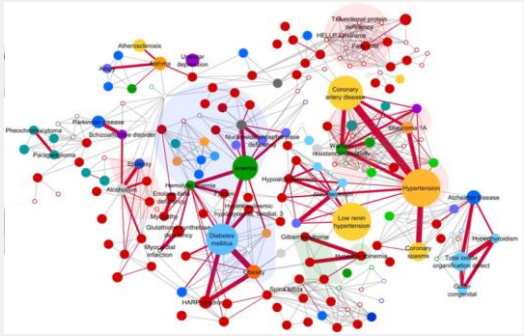
功能集打分法（FCS）



通路拓扑结构法



网络拓扑结构法



D.-S. Lee et al. pnas 105(29)9880-9885(2008)

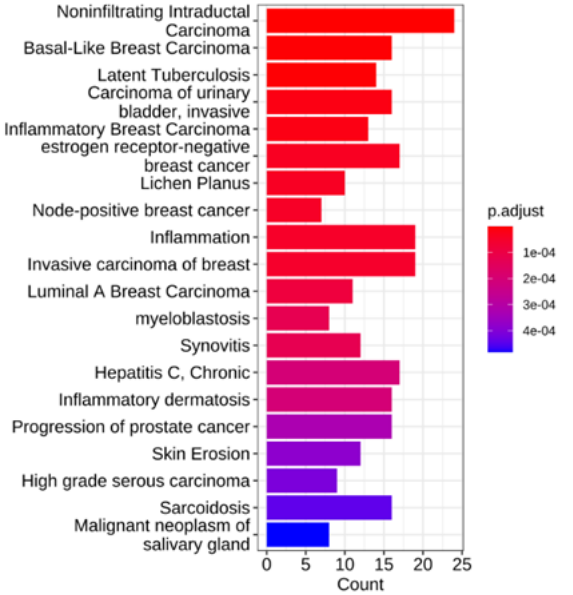
Hasdemir et al. BMC Systems Biology 8(1):61(2014)

生物通路富集分析算法

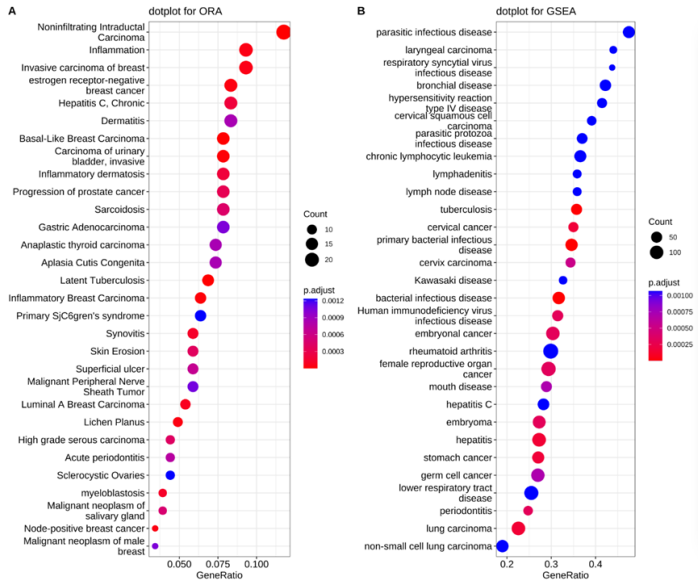
- **过表达分析法：**
  clusterProfiler、DAVID、GeneMAPP等
- **功能集打分法：**
  - **GSEA、Catmap、GlobalTest等**
- **拓扑结构法：**
  - **Pathway-Express、SPIA、NetGSA等**

KEGG通路富集分析

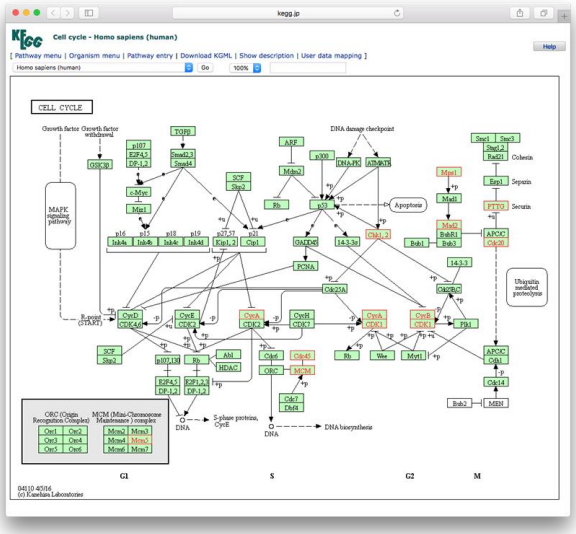
❒ 分析结果可视化



barplot

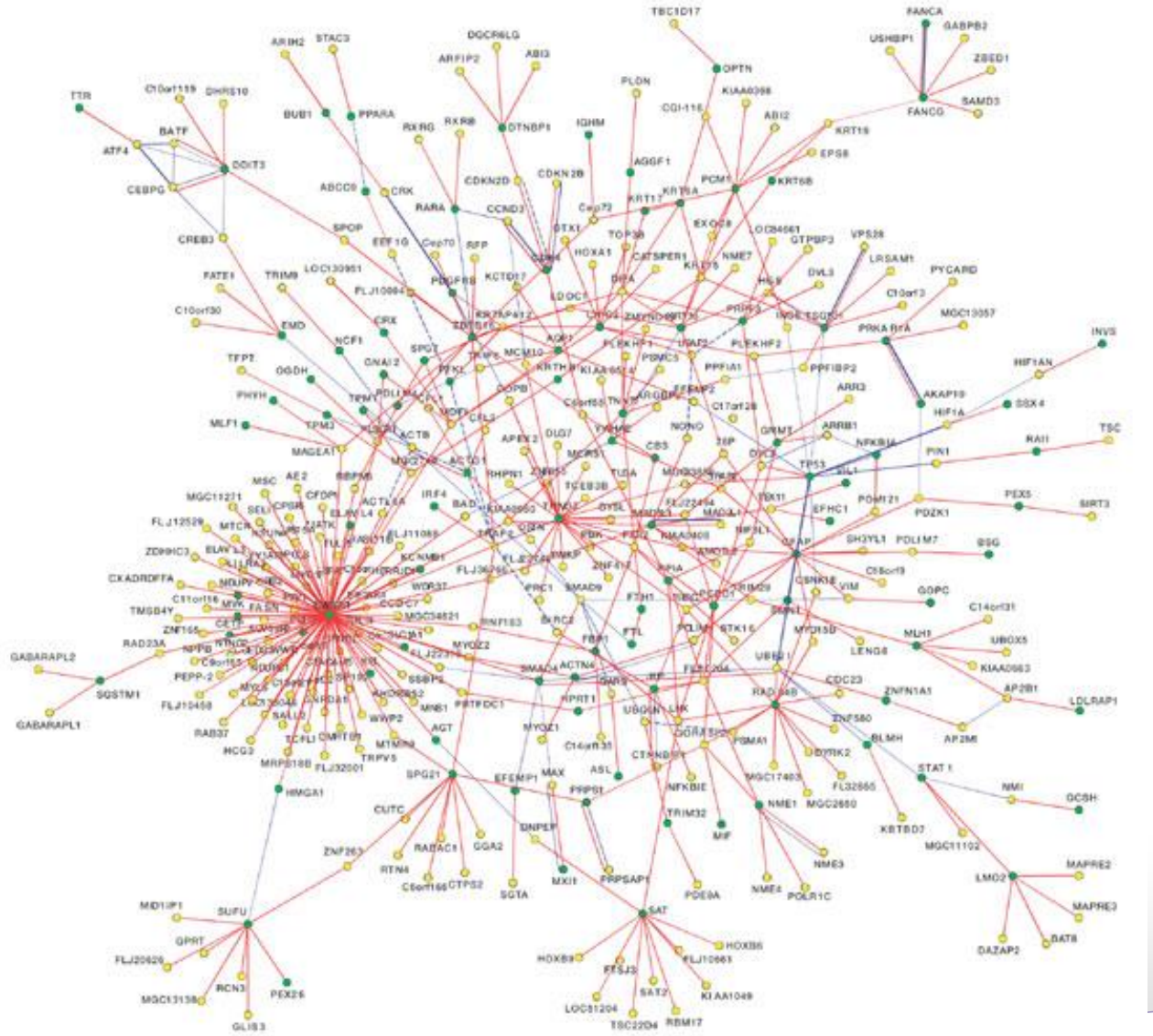


dotplot



KEGG pathway

蛋白质-蛋白质相互作用网络



蛋白质互作数据库BioGRID

蛋白质互作数据库IntAct

EMBL-EBI Services Research Training About us

IntAct

Home | Advanced Search | About | Resources | Download | Feedback

IntAct Molecular Interaction Database

IntAct provides a freely available, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions and are freely available. The IntAct Team also produce the Complex Portal.

Search in IntAct

Enter search term(s)...

Search | Search Tips

Examples

- Gene, Protein, RNA or Chemical name: BRCA2, Staurosporine
- UniProtKB or ChEBI AC: Q06609, CHEBI:15996
- UniProtKB ID: LCK_HUMAN
- RNACentral ID: URS00004C95F4_559292
- PMID: 25416956
- IMEx ID: IM-23318

Dataset of the month: February

Widespread macromolecular interaction perturbations in human genetic disorders..

- Sahni. et al. IntAct PSI-MI 2.5 PSI-MI TAB
- Go to Archive

Sign up for our newsletter

Sign up here

News Follow @intact_project

Tweets by @intact_project

Data Content

- Publications: **14495**
- Interactions: **694486**
- Interactors: **95487**

Submission

Submit your data to IntAct to increase its visibility and usability!

Contributors

Manually curated content is added to IntAct by curators at the EMBL-EBI and the following organisations:

MINT, UniProt, SIB, I2D, innateDB, Molecular Connections, MatrixDB, MBInfo, AgBase, GO, IMEx

IntAct is a member of the IMEx Consortium.

Citing IntAct

The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases.

Orchard S et al
[PMID:24234451]
[Full Text]

Training

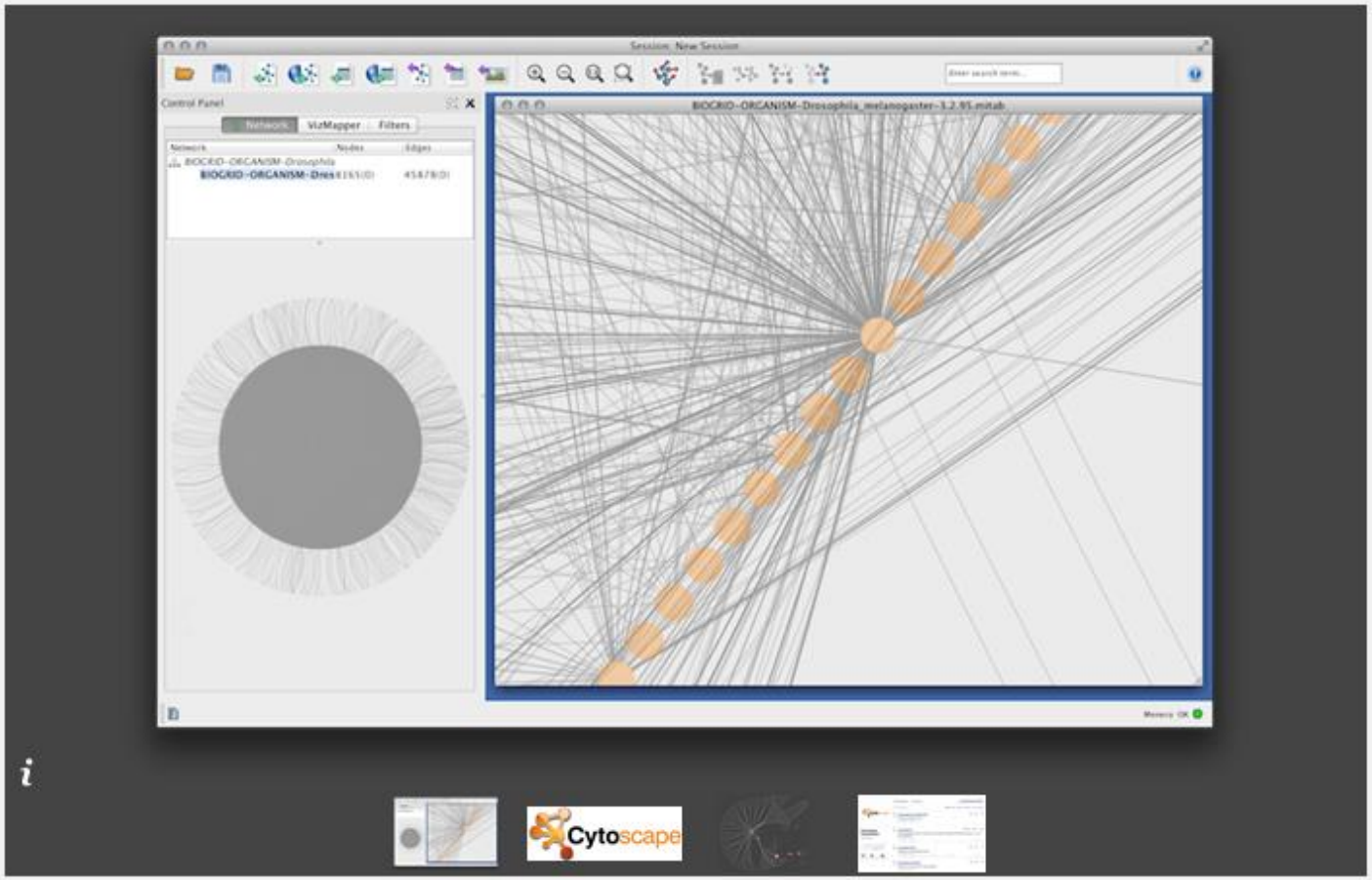
Online & upcoming courses

Cytoscape：网络构建和分析工具

Cytoscape

Search... Go

Home Introduction Download Apps Documentation Community Report a Bug Getting Help



Network Data Integration, Analysis, and Visualization in a Box

Cytoscape is an open source software platform for visualizing complex networks and integrating these with any type of attribute data. A lot of *Apps* are available for various kinds of problem domains, including bioinformatics, social network analysis, and semantic web.

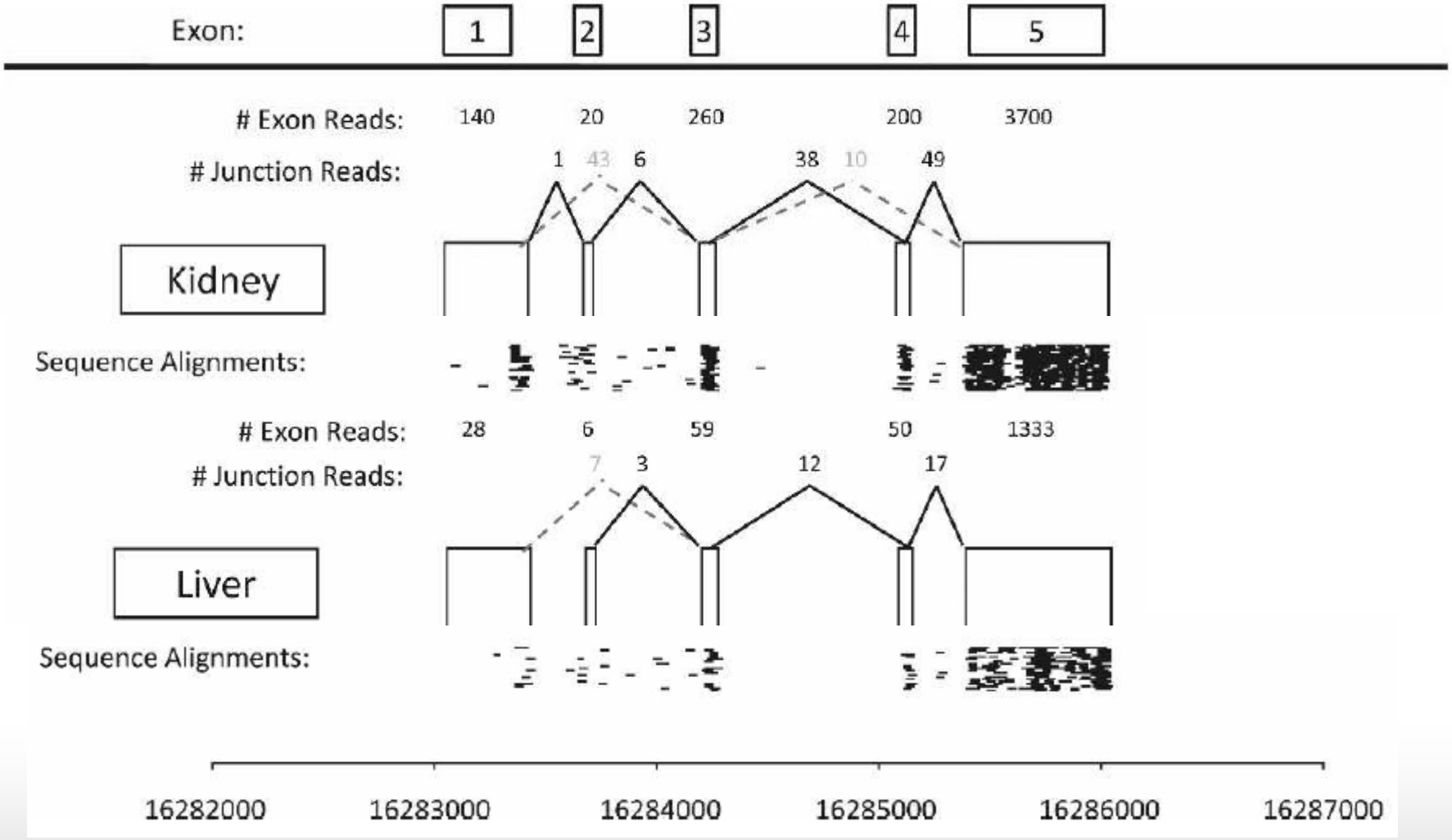
Download Cytoscape

Welcome Letter

Release Notes

Sample Visualizations

基因表达的组织特异性

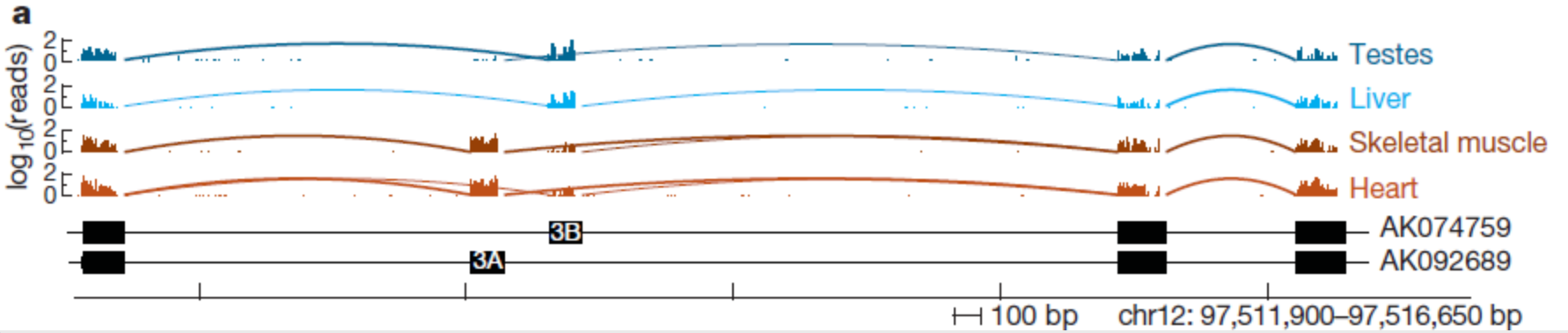


基因C17orf45 (ENSG00000175061)的结构

RNA可变剪接

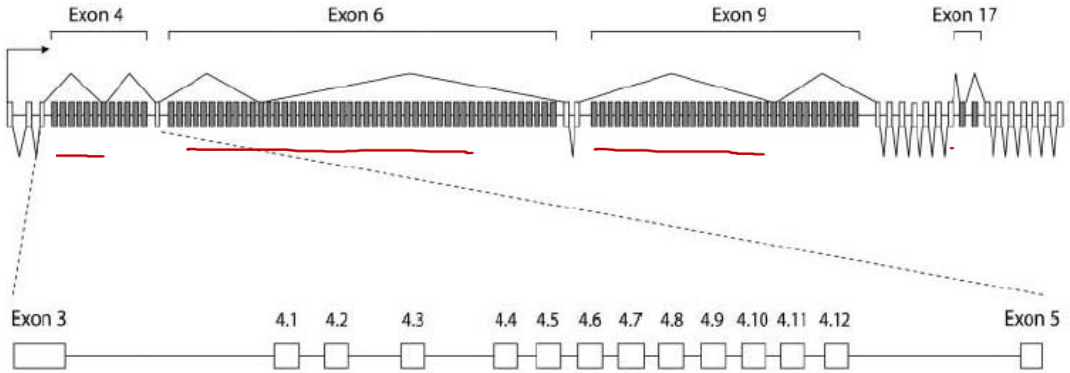
- AS, Alternatively Splicing
  - 92–94%的人类多外显子基因存在可变剪接现象
- Major isoform vs. minor isoform
- 86%包括至少有一个次要异构体 (比例15%)

SLC25A3基因

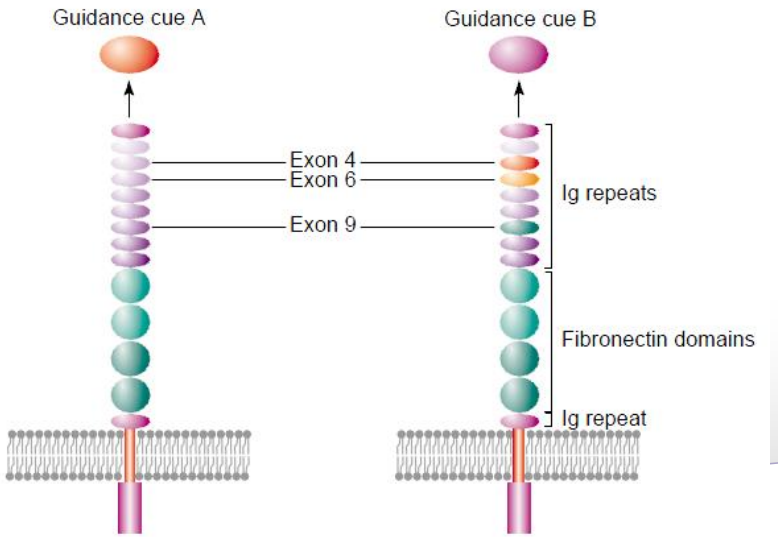


果蝇Dscam基因的可变剪接

- 12*48*33*2=38,016个不同的异构体

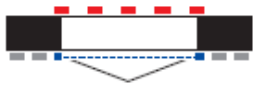
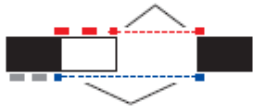
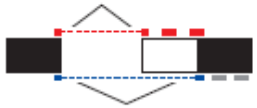
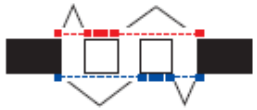
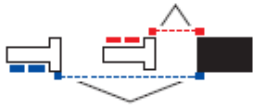
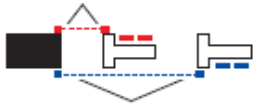


- 果蝇: ~250,000个神经元
- 轴突导向 (Axon guidance)
- 调控神经发育
  - 神经元之间的连接



可变剪接的类型

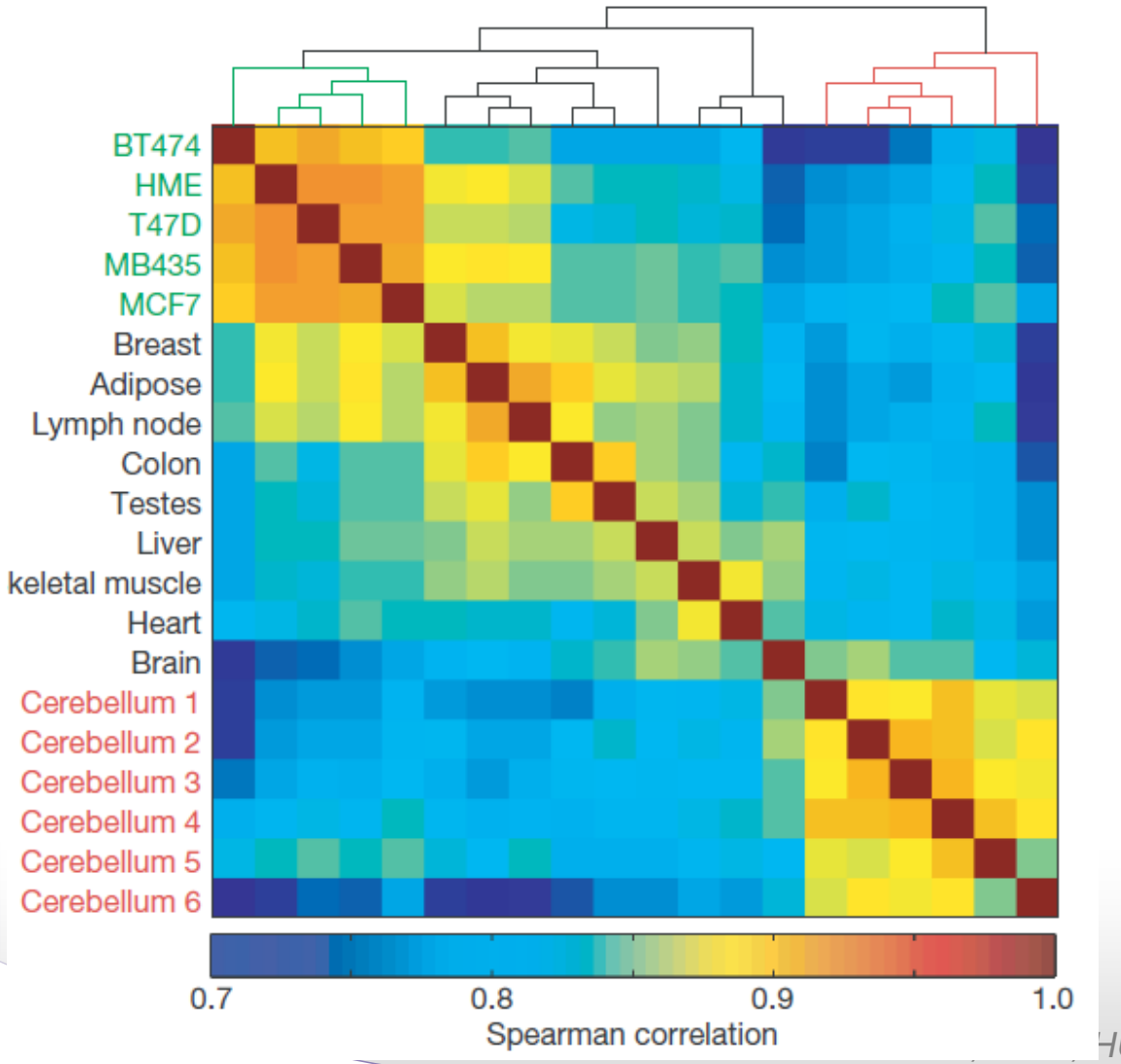
- 8种类型
- 仅统计外显子上的读段数 – 不全面

Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon		37	35	10,436	6,822	65	72
Retained intron		1	1	167	96	57	71
Alternative 5′ splice site (A5SS)		15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)		17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)		4	4	167	95	57	66
Alternative first exon (AFE)		14	13	10,281	5,311	52	63
Alternative last exon (ALE)		9	8	5,246	2,491	47	52
Tandem 3′ UTRs		7	7	5,136	3,801	74	80
Total		105	100	37,782	22,657	60	68

Constitutive exon or region — Body read — Junction read — pA Polyadenylation site

Alternative exon or extension — Inclusive/extended isoform — Exclusive isoform — Both isoforms

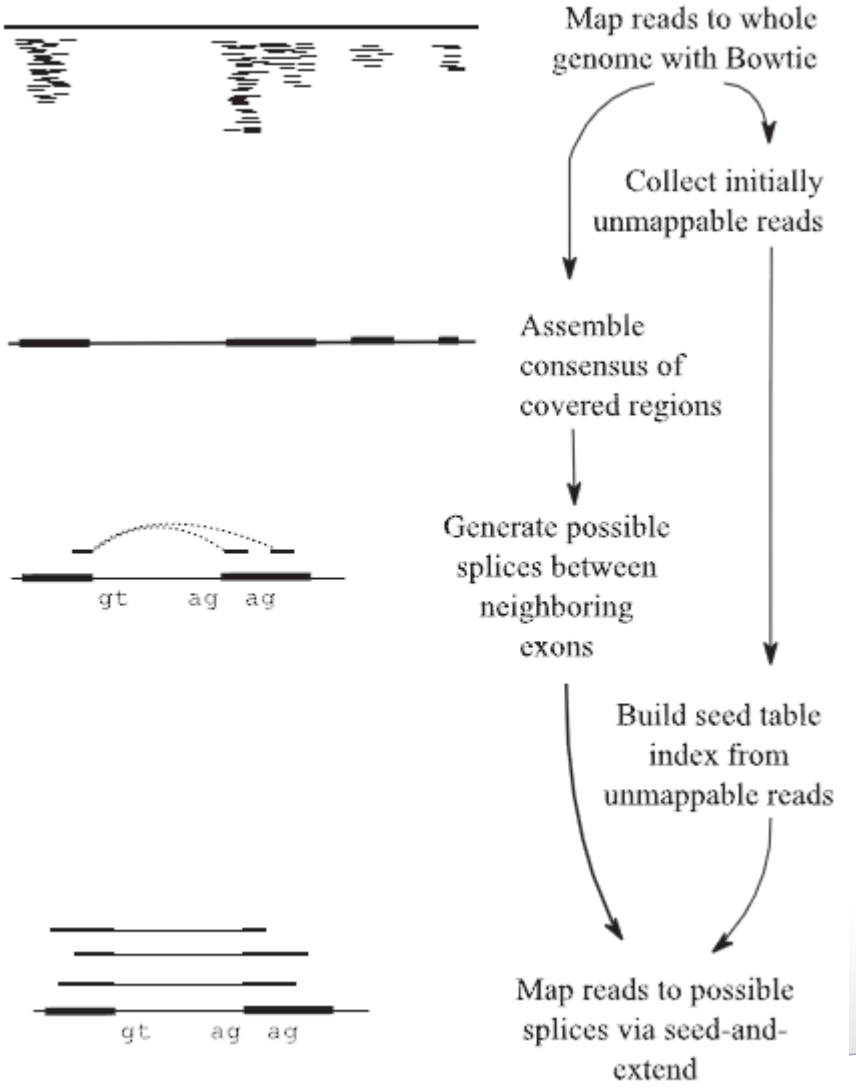
可变剪接的组织特异性



- 脑的可变剪接特异性最高
- 不同个体之间存在差异
- 细胞系有特别的剪接模式

可变剪接鉴定：Cufflinks

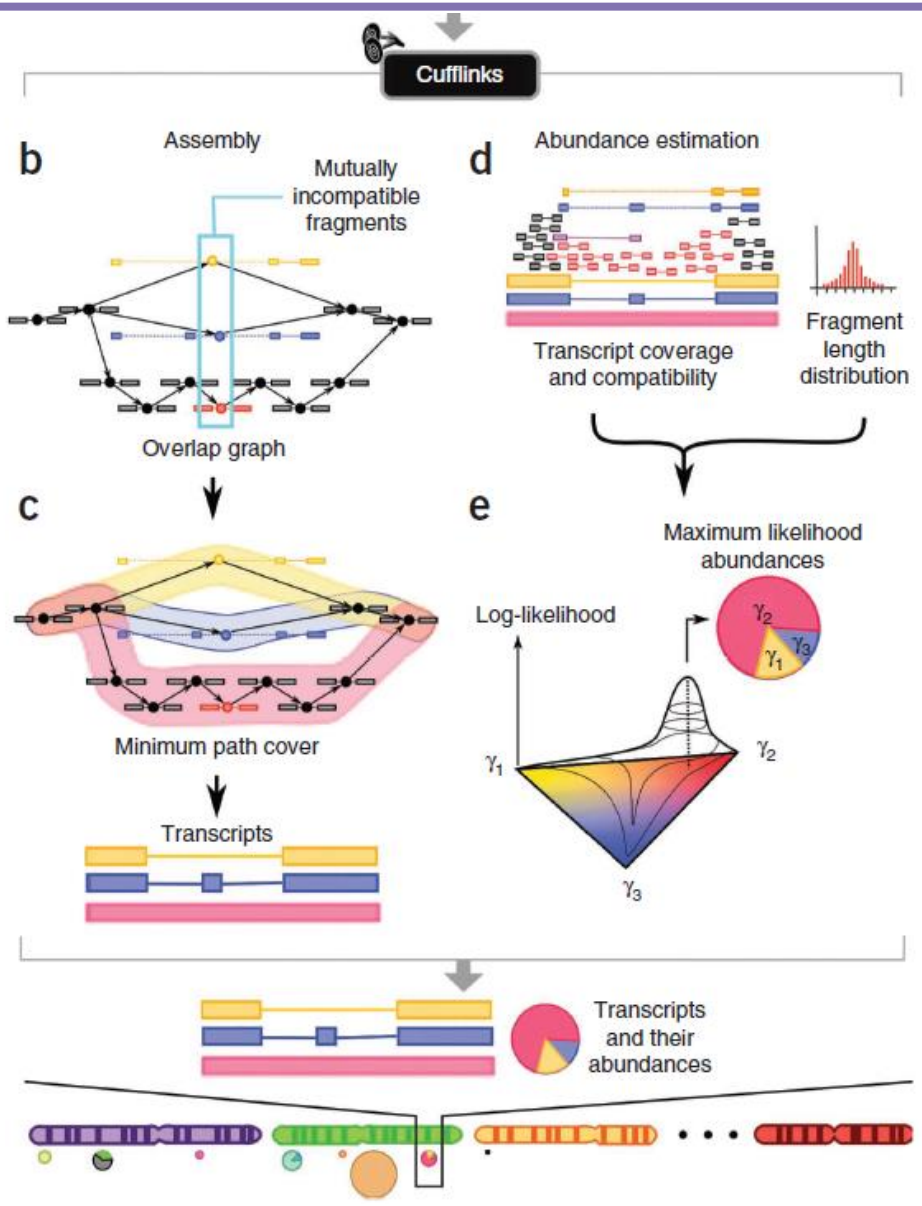
- 用Bowtie (领带) 将所有读段回贴到基因组上
- 不能完美匹配的，用TopHat (礼帽) 发现剪接位点



http://cole-trapnell-lab.github.io/cufflinks/

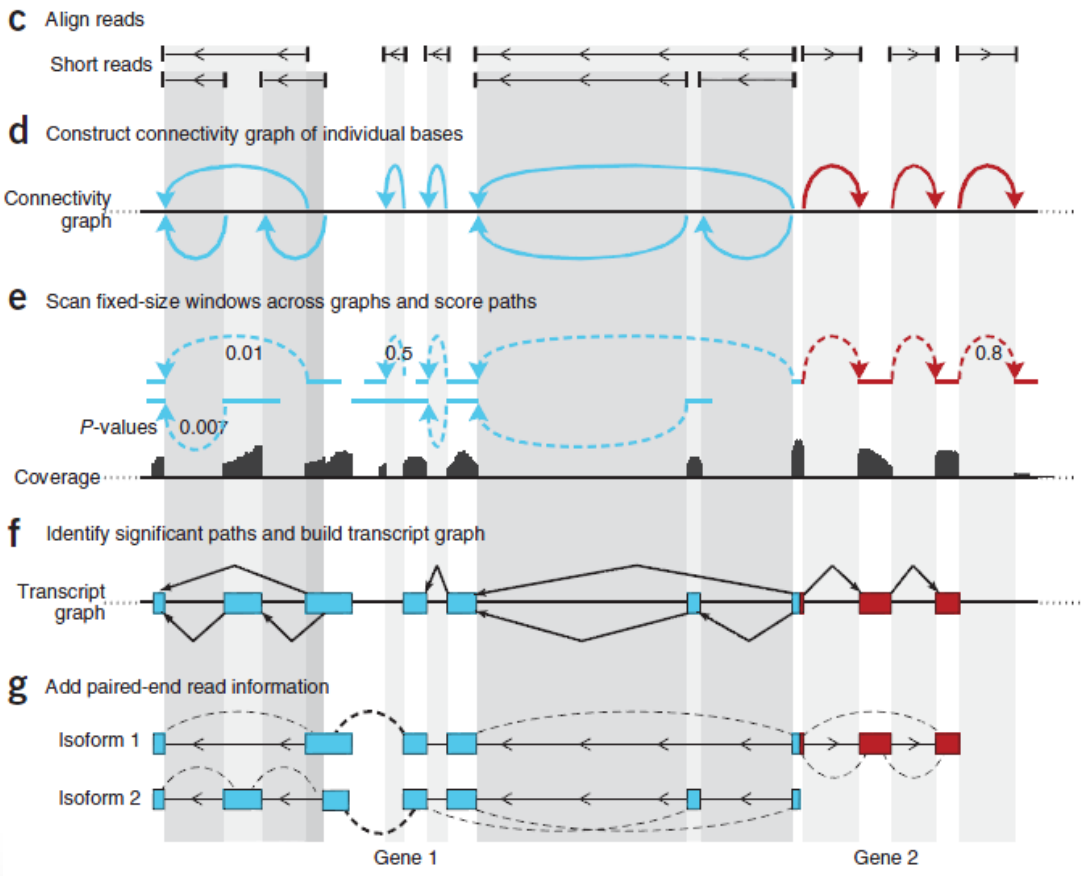
可变剪接鉴定：Cufflinks

- OLC算法 (Overlap graph)
- 读段是图中的结点，
- 若可以匹配则置边
- 用OLC算法计算多条路径
- 用最大似然性方法估算每个异构体的表达水平



可变剪接鉴定：Scripture

- DBG算法 (Connectivity graph)
- 单个碱基是结点，边是两个读段共有的序列
- 用DBG算法算出多个路径
- Cuffdiff：利用双端测序数据校正不同异构体的比例
- RPKM归一化表达



http://www.broadinstitute.org/software/scripture/