生物信息学

第十一章 转录组与转录调控分析（2）

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

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

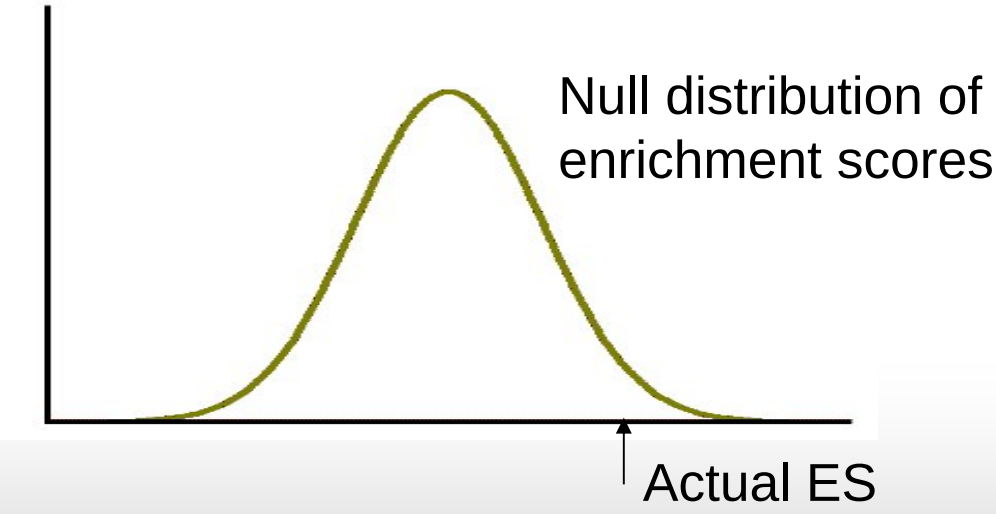
$P_{hit}(S, 7) = 1.705$

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

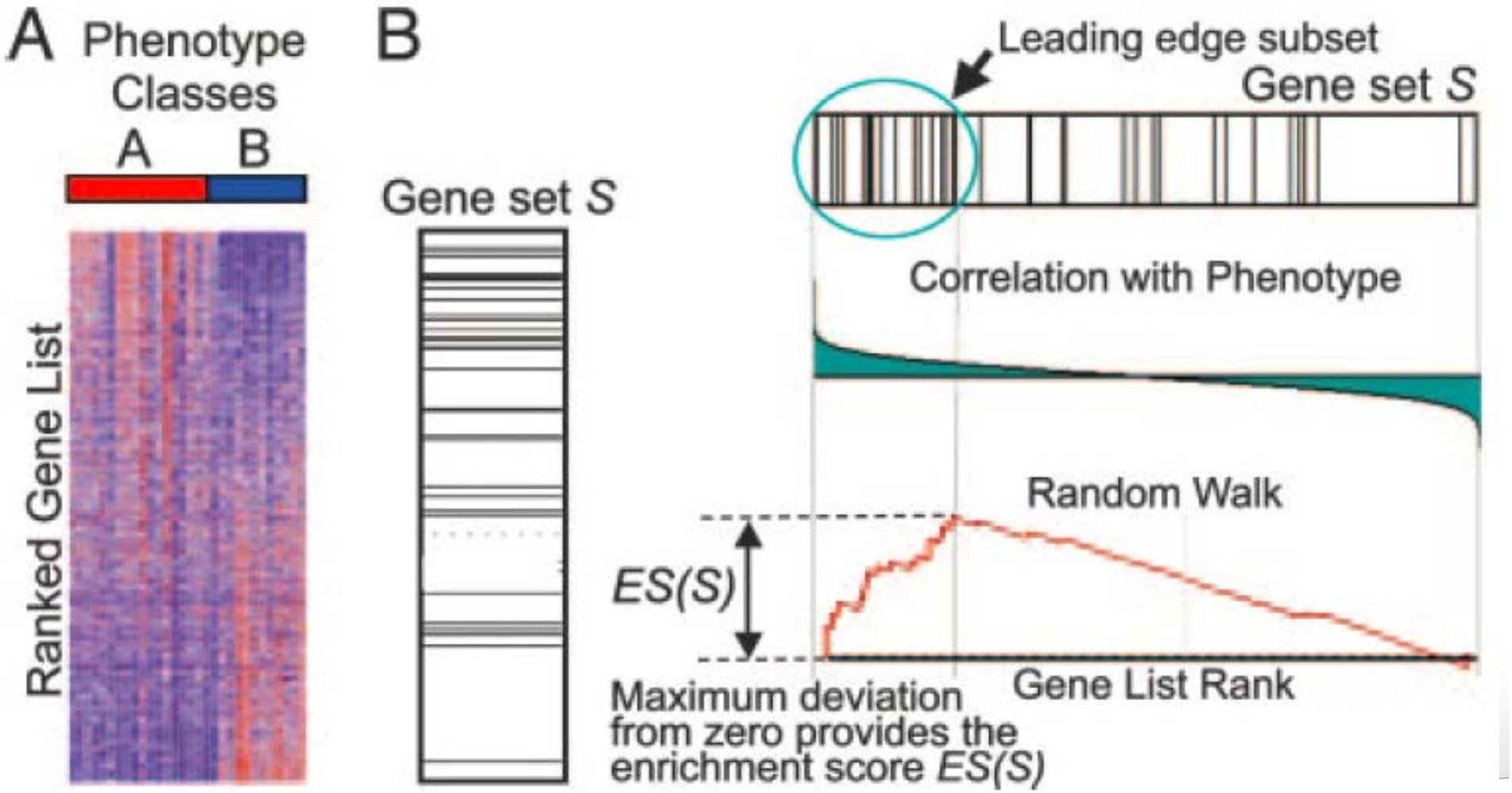
ES(7) = 0.705

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

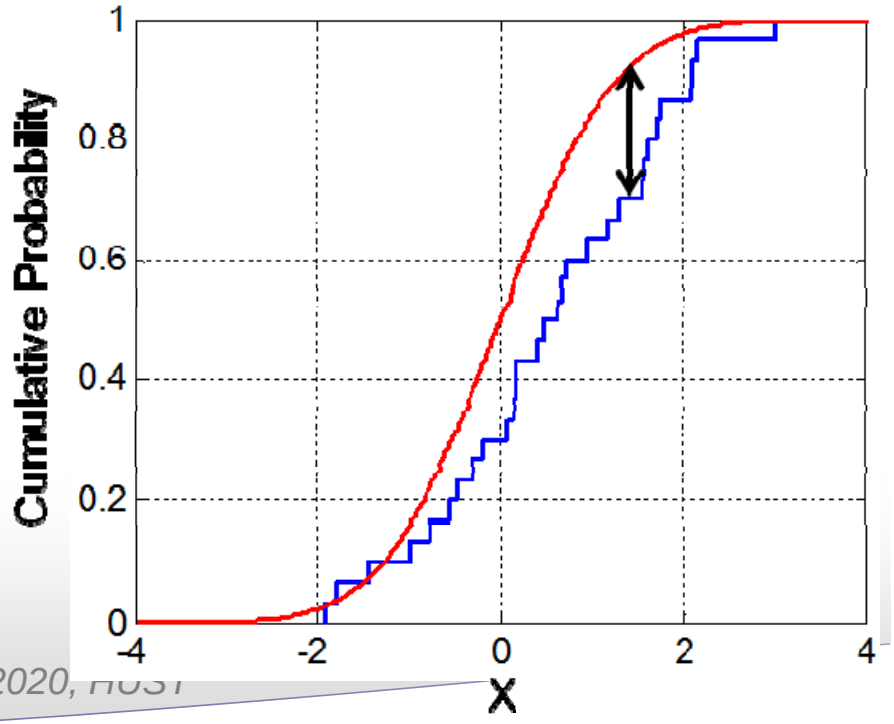


GSEA算法总结


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)

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

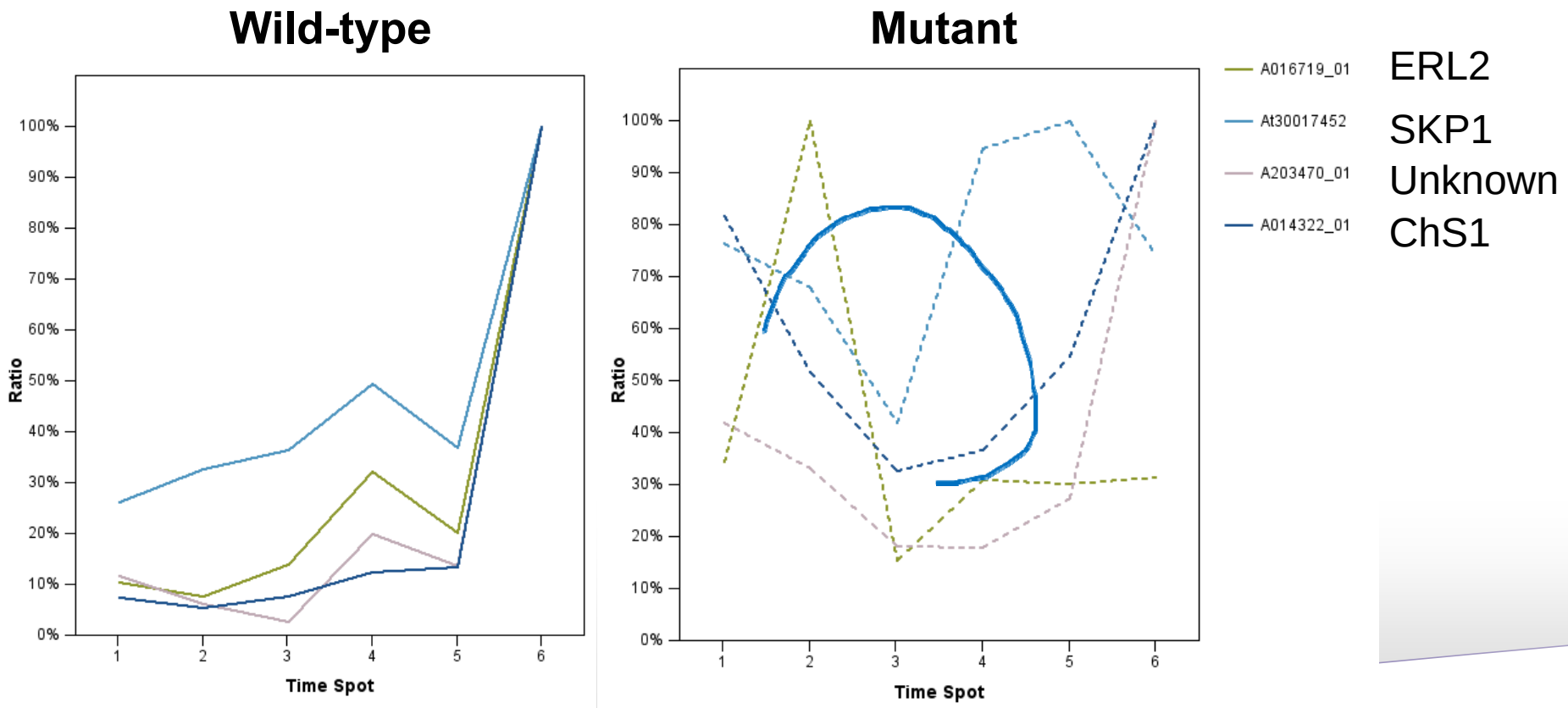


基因调控网络

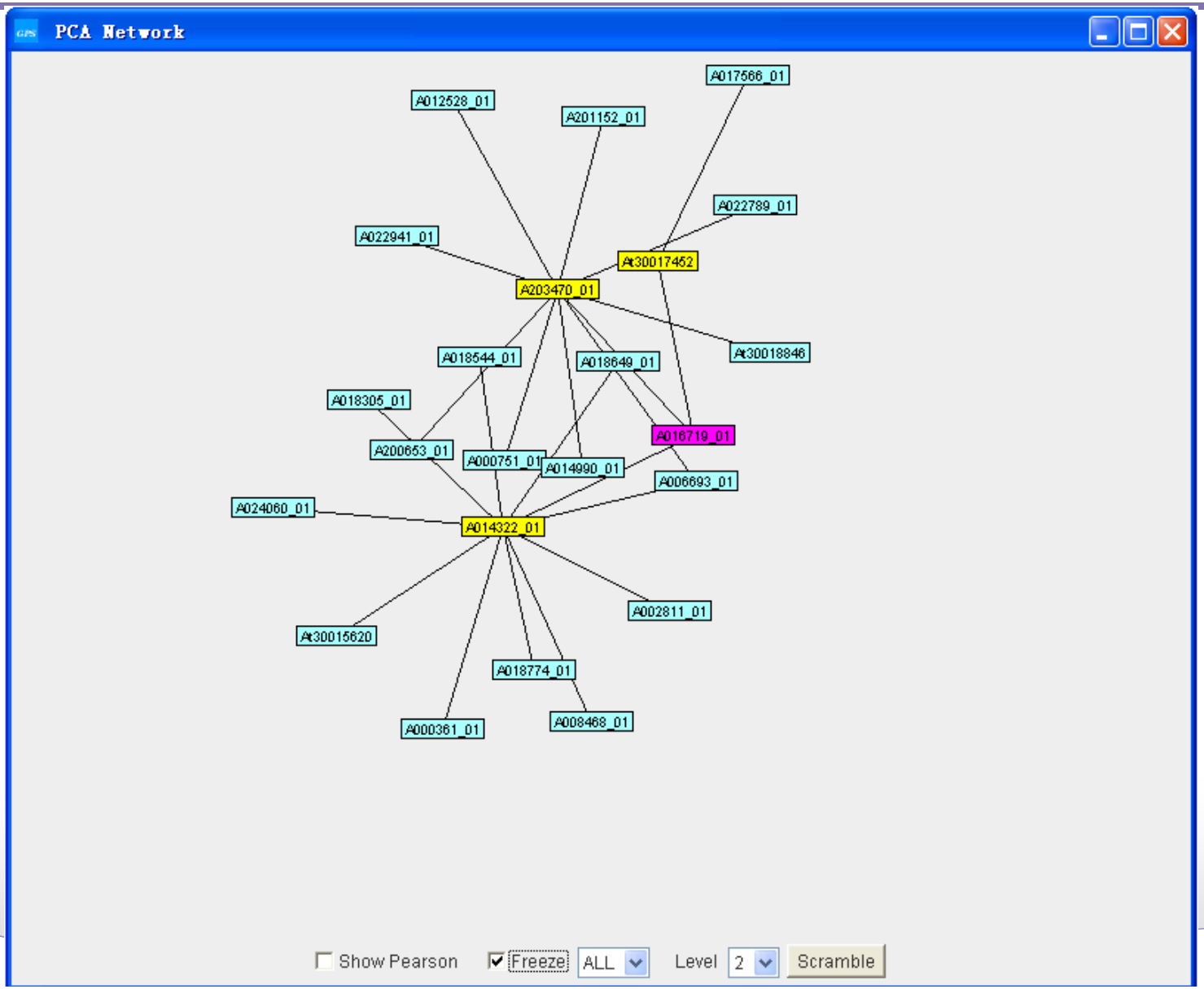
- 早期观点：表达谱相似的基因可能存在功能上的关联，可能相互作用… (直接作用)
- 当前的观点：表达谱相似的基因可能具有共同的调控元件 (基因UTR区域存在共同的启动子), 能够被同一个上游因子所调控

相关系数：基因共表达网络

- 与光合效率和气孔发育相关的基因：**ERL2**
  - 在**Wild-type**中与之显著相关，但在**Mutant**中显著不相关的基因

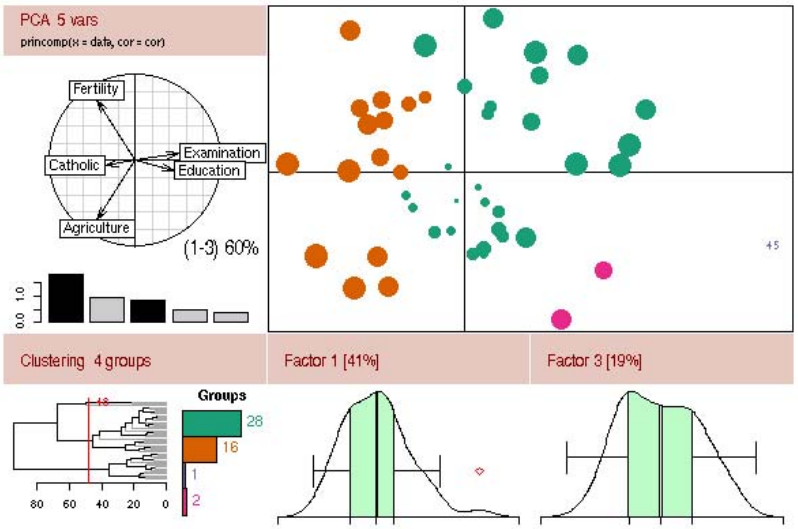


相关系数：基因共表达网络



Microarray: 工具&数据库

The R Project for Statistical Computing



About R
What is R?
Contributors
Screenshots
What's new?

Download
CRAN

R Project
Foundation
Members & Donors
Mailing Lists
Bug Tracking
Developer Page
Conferences
Search

Documentation
Manuals
FAQs
Newsletter
Wiki
Books
Certification
Other

Misc
Bioconductor
Related Projects
Links

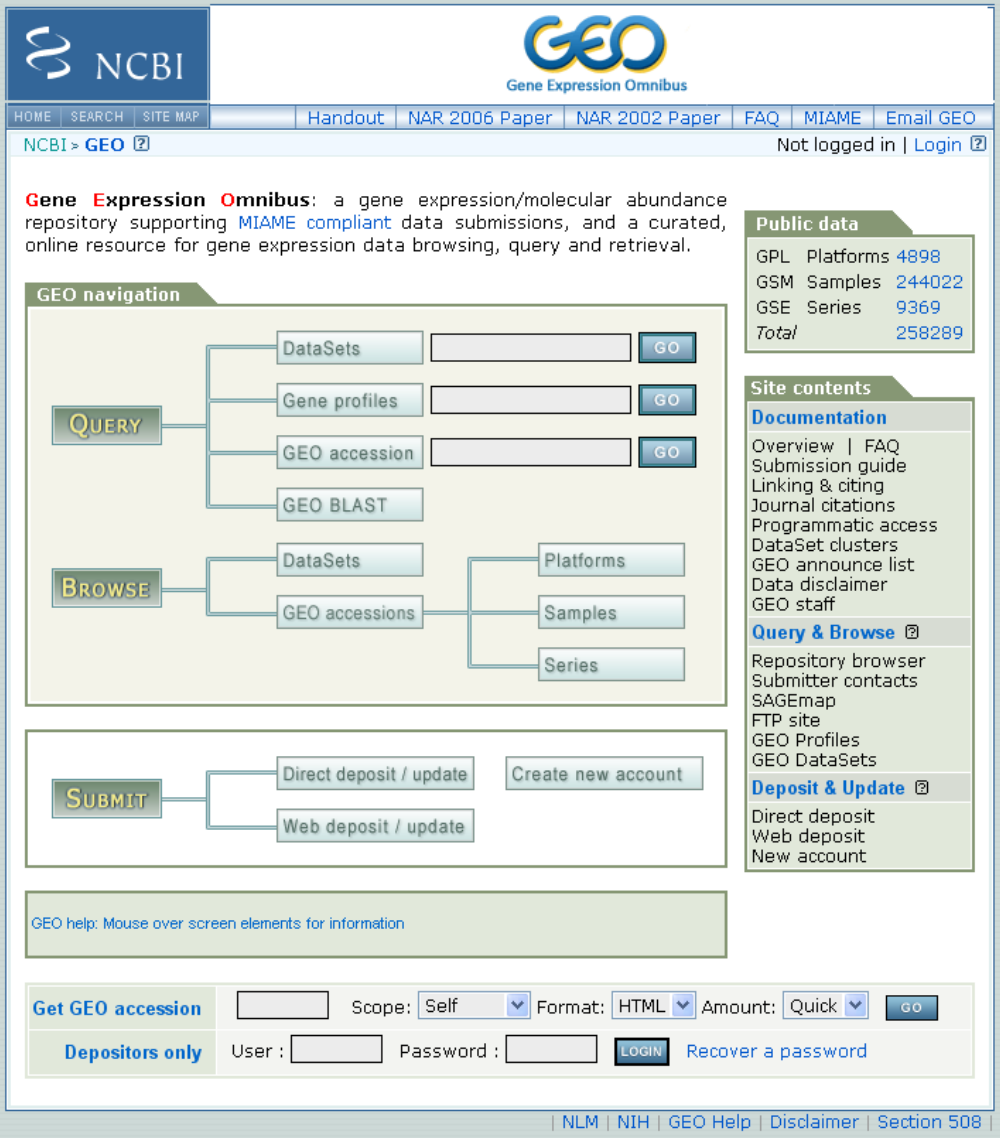
Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

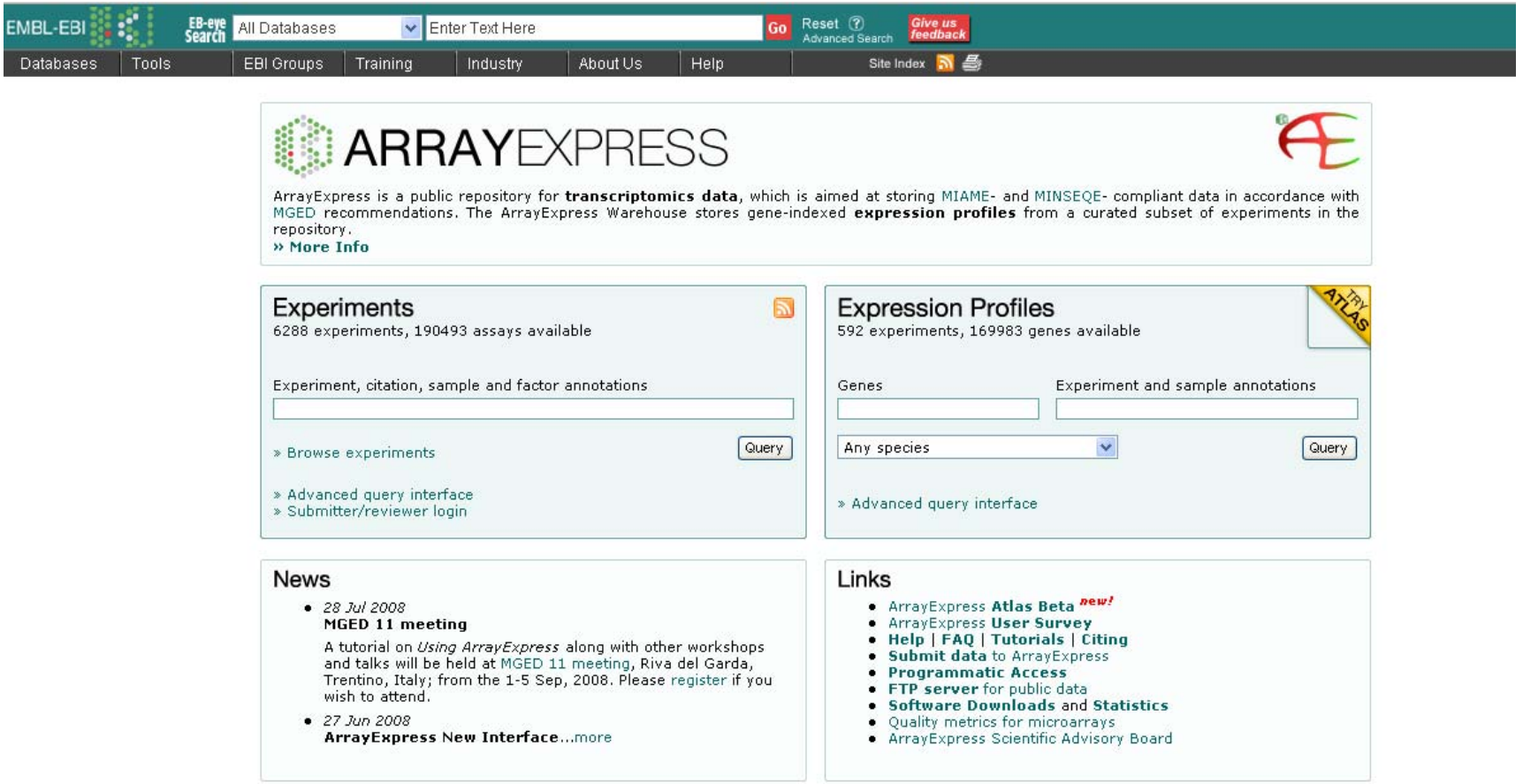
News:

- **R version 2.7.1** has been released on 2008-06-23.
- **R News 8/1** has been published on 2008-05-27.
- The R Foundation as been awarded four slots for R projects in the Google Summer of Code 2008.
- **R 2.7.2 Release Candidates** will appear August 18-24. Final release is scheduled for 2008-08-25.
- **useR! 2008**, the R user conference, will be held at Dortmund University, Germany, August 12-14, 2008.

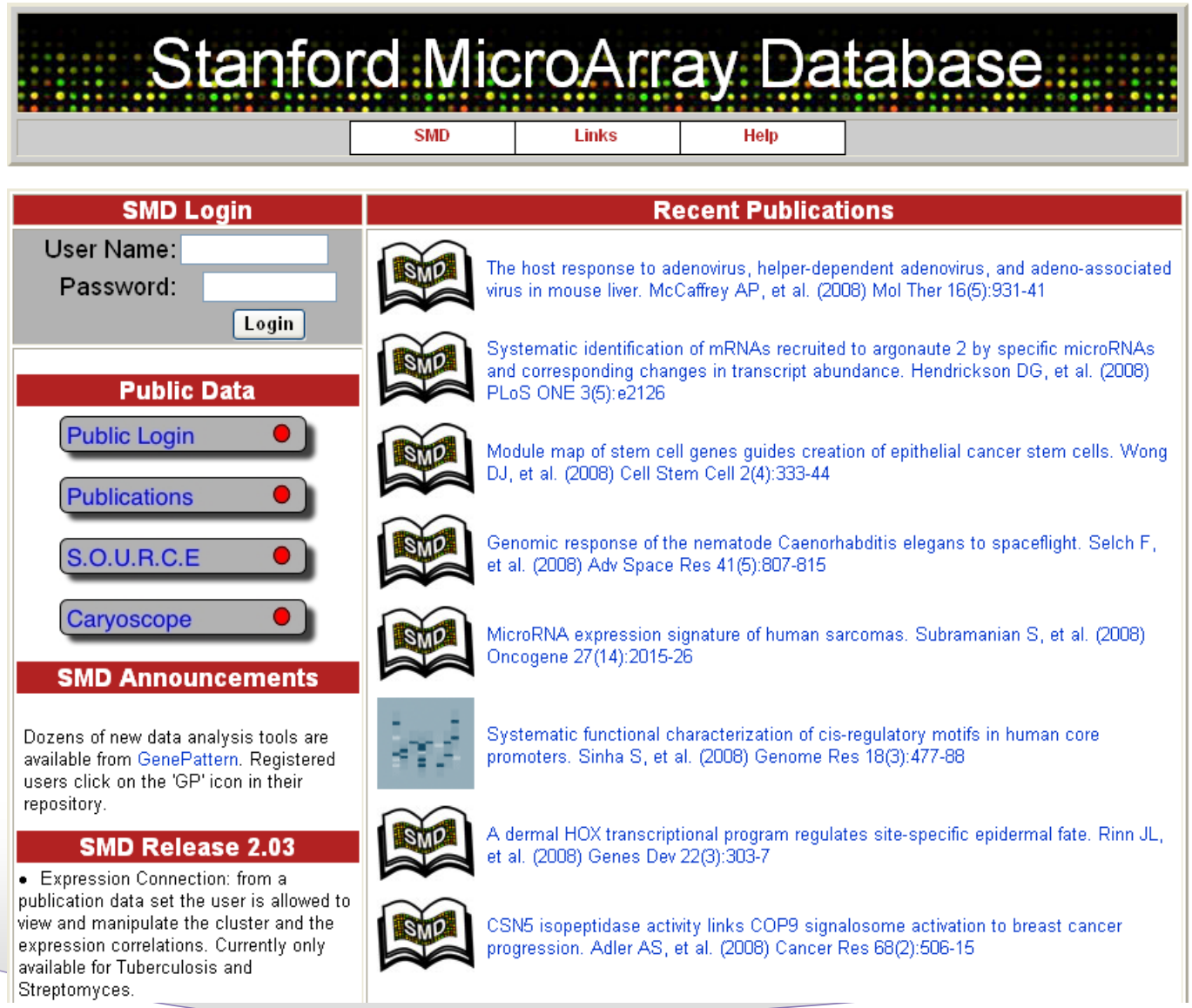
GEO - NCBI


NCBI
Gene Expression Omnibus
HOME | SEARCH | SITE MAP
Handout | NAR 2006 Paper | NAR 2002 Paper | FAQ | MIAME | Email GEO
NCBI » GEO
Not logged in | Login
Gene Expression Omnibus: a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.
GEO navigation
QUERY
DataSets
Gene profiles
GEO accession
GEO BLAST
GO
BROWSE
DataSets
GEO accessions
Platforms
Samples
Series
SUBMIT
Direct deposit / update
Web deposit / update
Create new account
GEO help: Mouse over screen elements for information
Public data
GPL Platforms 4898
GSM Samples 244022
GSE Series 9369
Total 258289
Site contents
Documentation
Overview | FAQ
Submission guide
Linking & citing
Journal citations
Programmatic access
DataSet clusters
GEO announce list
Data disclaimer
GEO staff
Query & Browse
Repository browser
Submitter contacts
SAGEmap
FTP site
GEO Profiles
GEO DataSets
Deposit & Update
Direct deposit
Web deposit
New account
Get GEO accession
Scope: Self
Format: HTML
Amount: Quick
GO
Depositors only
User :
Password :
LOGIN
Recover a password
NLM | NIH | GEO Help | Disclaimer | Section 508

Array Express - EMBL



SMD - Stanford



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
  - 将序列映射到基因组上
- 序列检测到越多，则表达越高
- RNA-Seq数据分析：挑战！
  - 将读段映射到参考基因组上
  - 定量已知的基因
  - 发现新的转录本
  - 定量可变剪切异构体

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

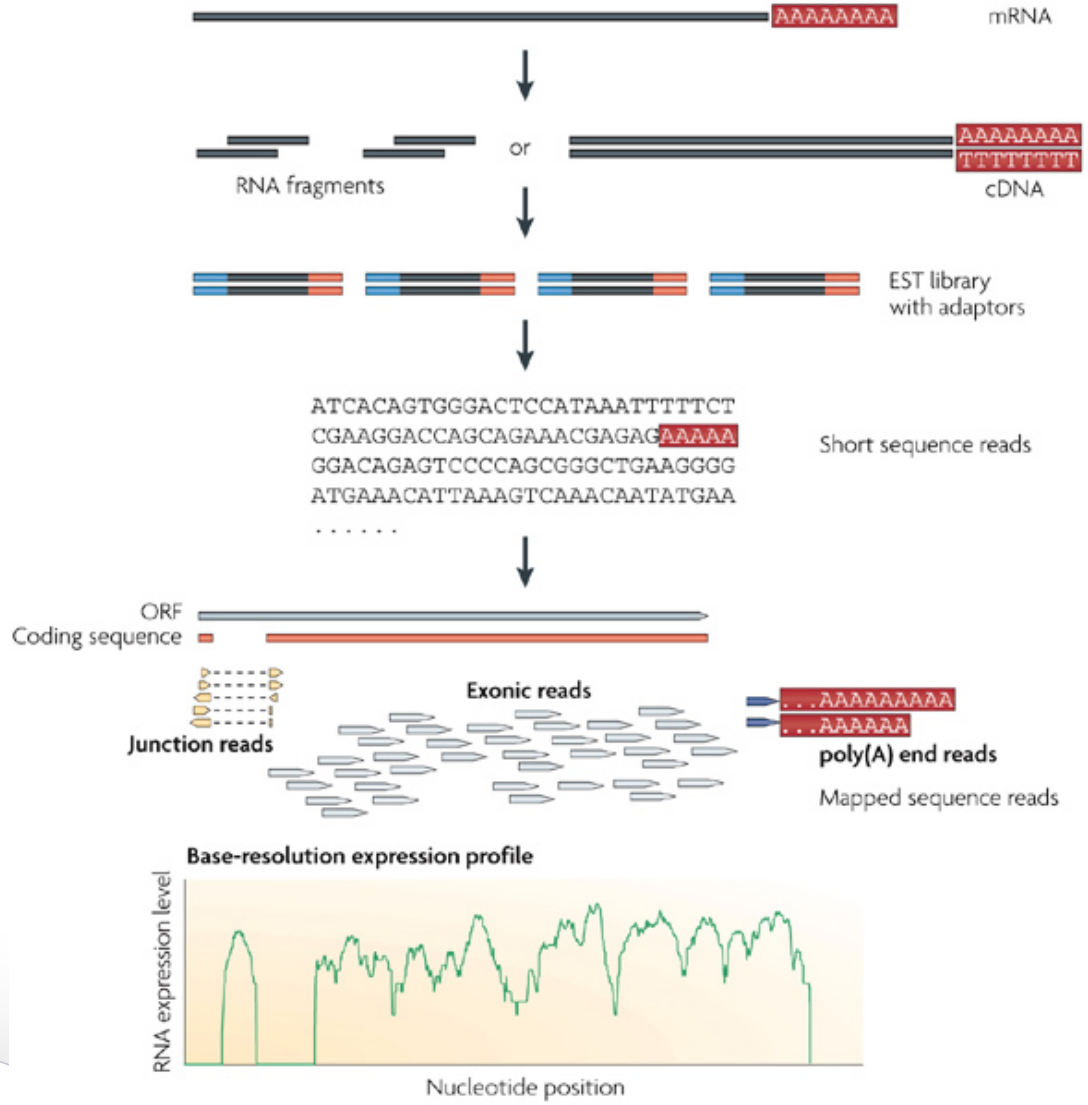
Line 3: + ?

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:
- ✓读段映射
- ✓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切异构体，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的优势

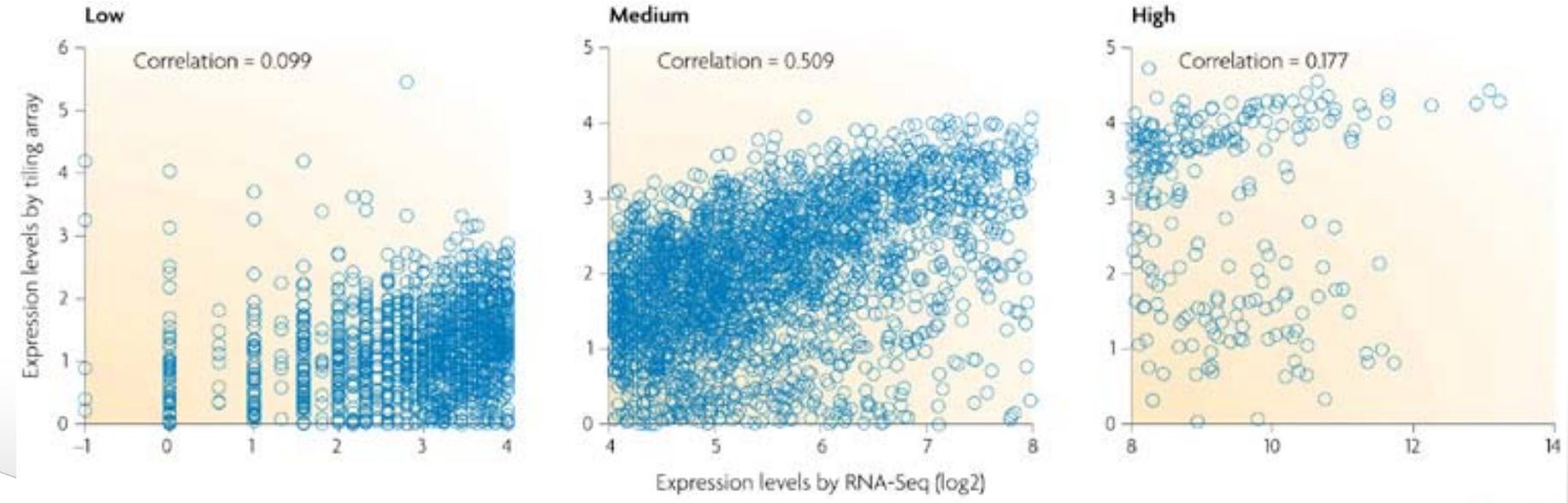
- 转录组分析方法：RNA-Seq比其他方法更有优势

Table 1 | **Advantages of RNA-Seq compared with other transcriptomics methods**

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

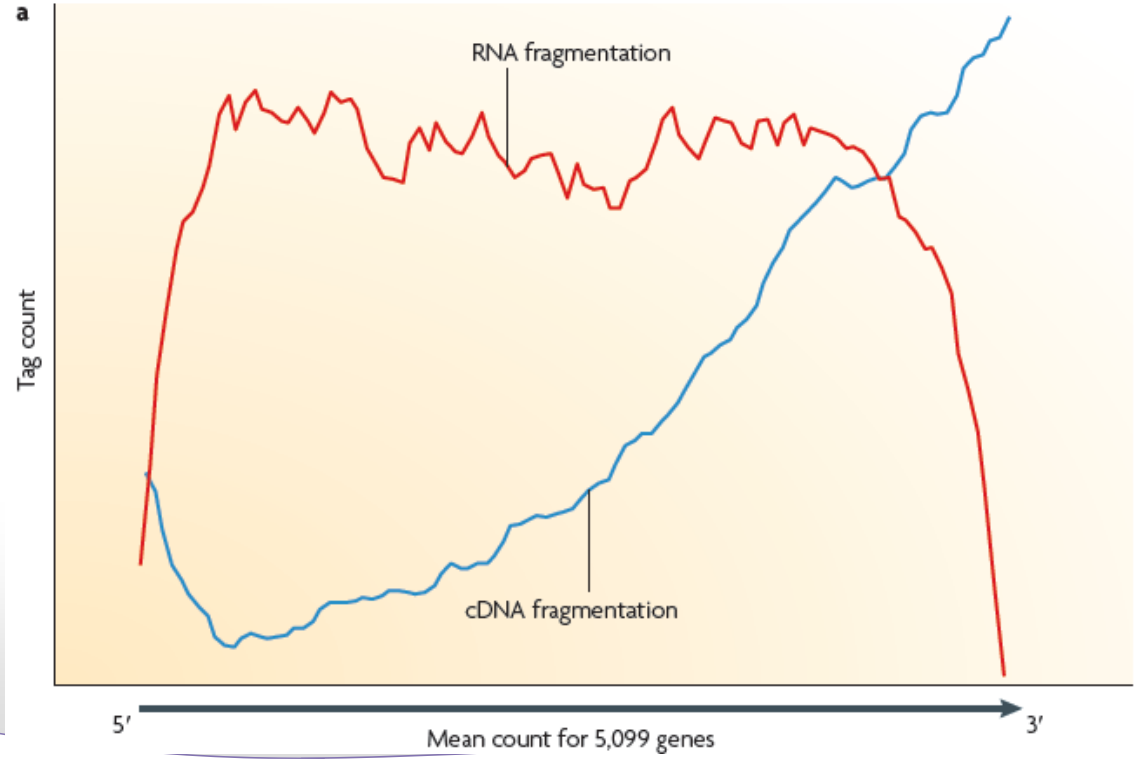
RNA-Seq *vs*. DNA芯片

- 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



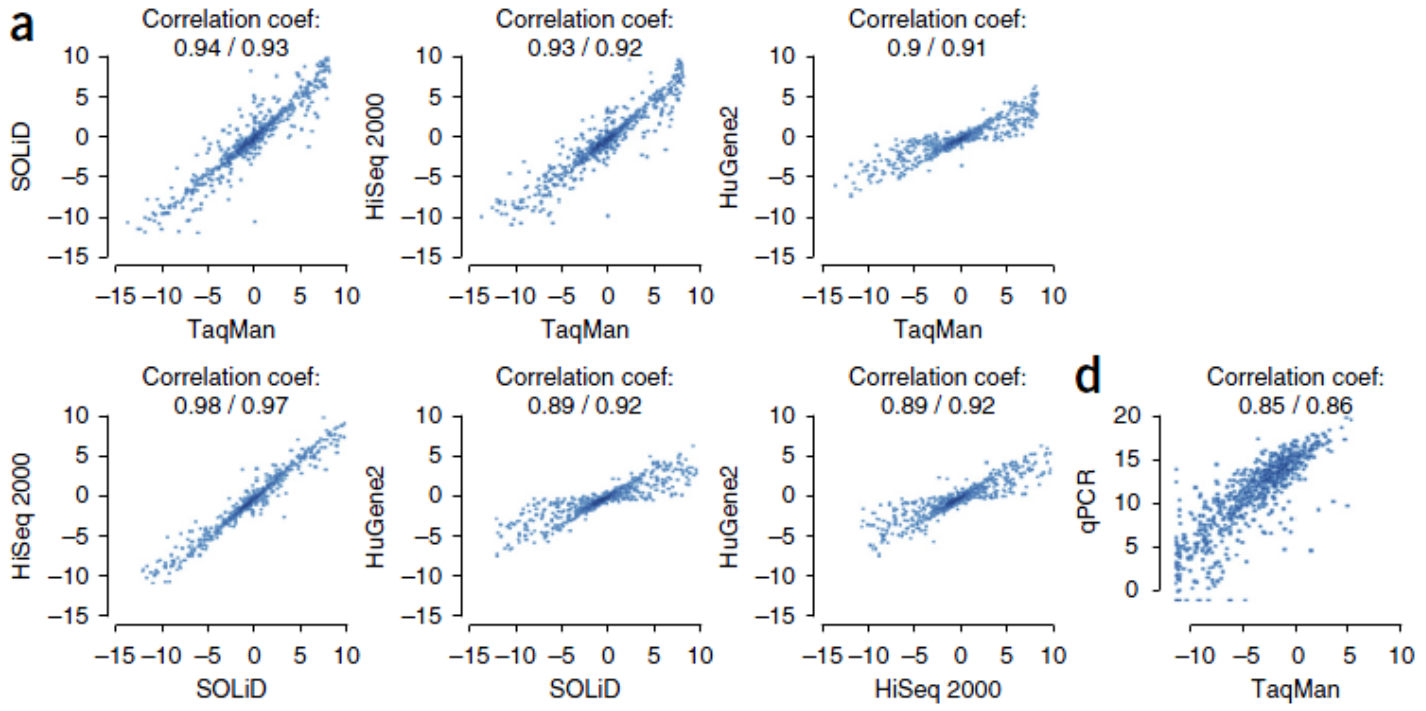
RNA-Seq的问题：建库

- 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

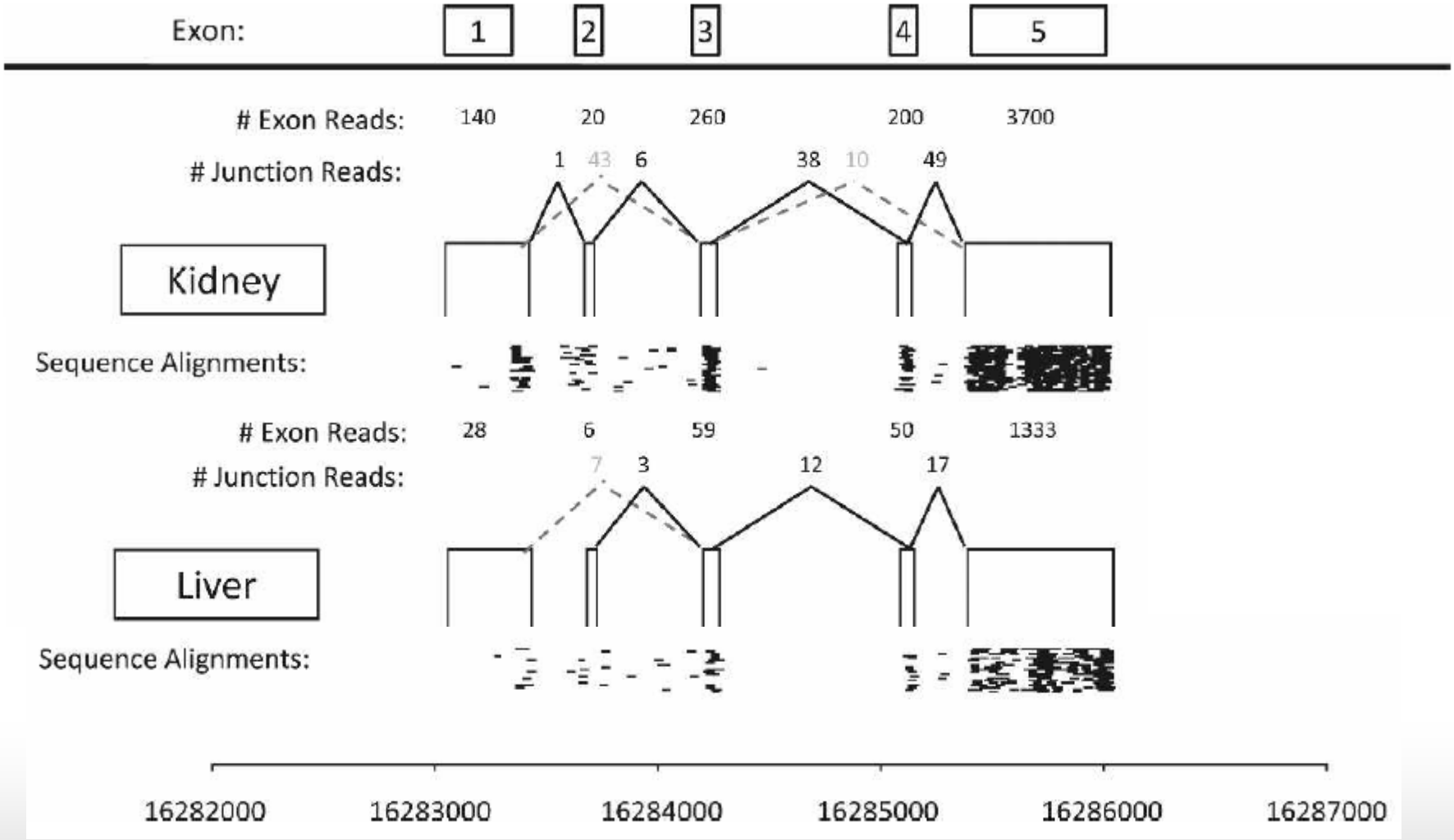


RNA-Seq的准确性

- HiSeq 2000和SOLiD测序分析结果
- HuGene2：Affymetrix芯片数据
- 利用qPCR (Taqman或PrimePCR) 验证表达水平
- 不同方法一致性较高



RNA-Seq数据：基因模型



基因C17orf45 (ENSG00000175061)的结构

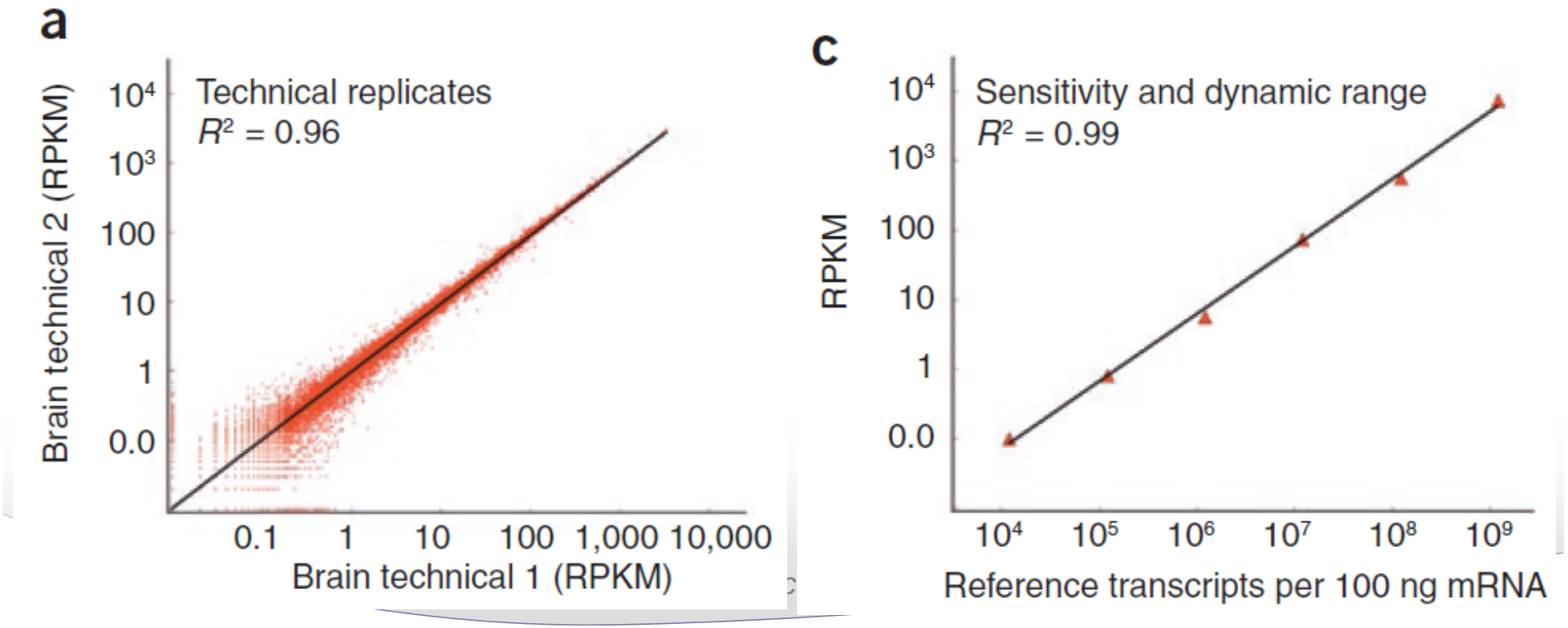
RNA-Seq数据的归一化

- 不同样本之间比较：测序深度不同
- 不同基因之间：表达相同时序列越长读段越多
- RPKM (The number of reads per kilobase of transcript per million mapped reads)
  - 每百万可映射读段每千kb的读段数
  - 例如，测序10^7读段，$8*10^6$可映射 (*N*)
  - 1kb转录本 (*L*)，总共有1000个读段 (*C*)
  - RPKM = 1000/(1 * 8) = 125
- Multiread (多重定位读段)
  - 计数时需要除以定位个数

$$R = \frac{10^9 C}{NL}$$

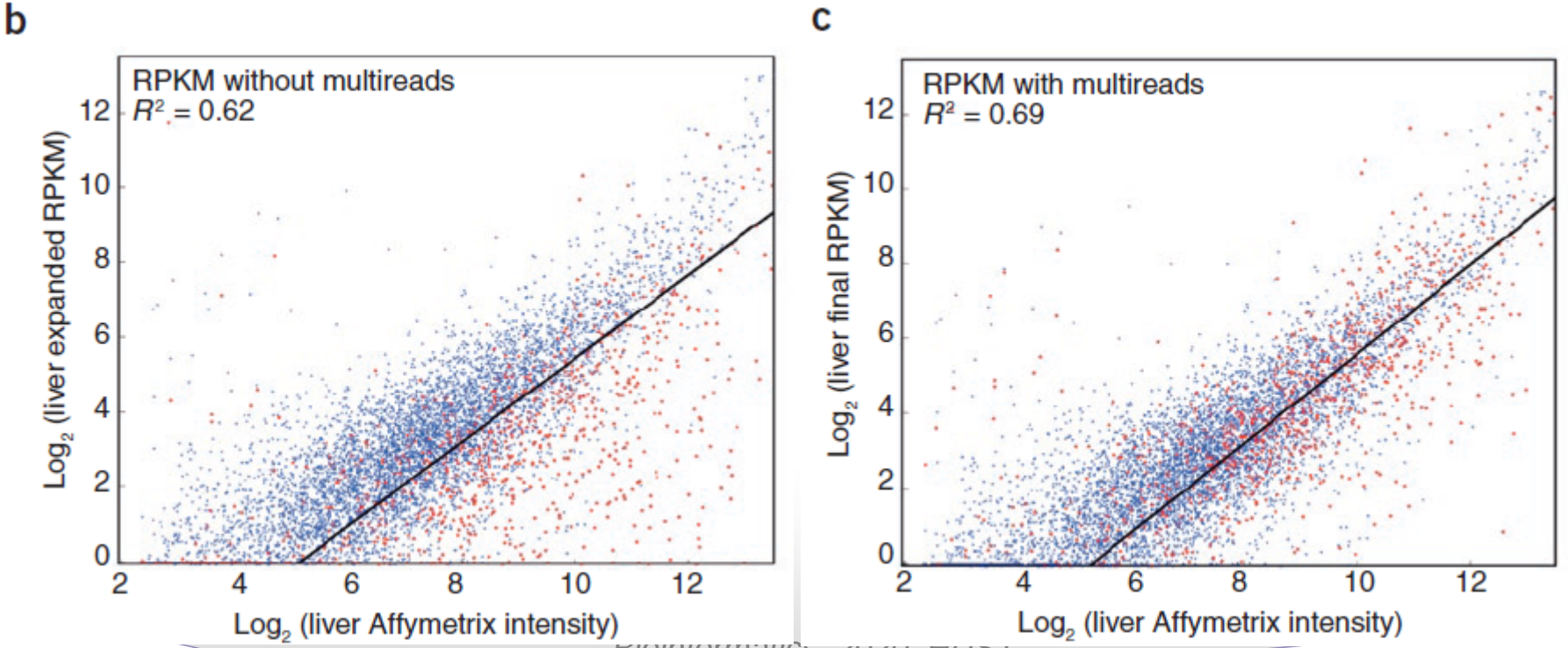
归一化的RNA-Seq数据

- 技术重复的结果一致性较高
- 体外合成的转录本加入样本中检测
  - RPKM与加入量线性相关



多重定位读段

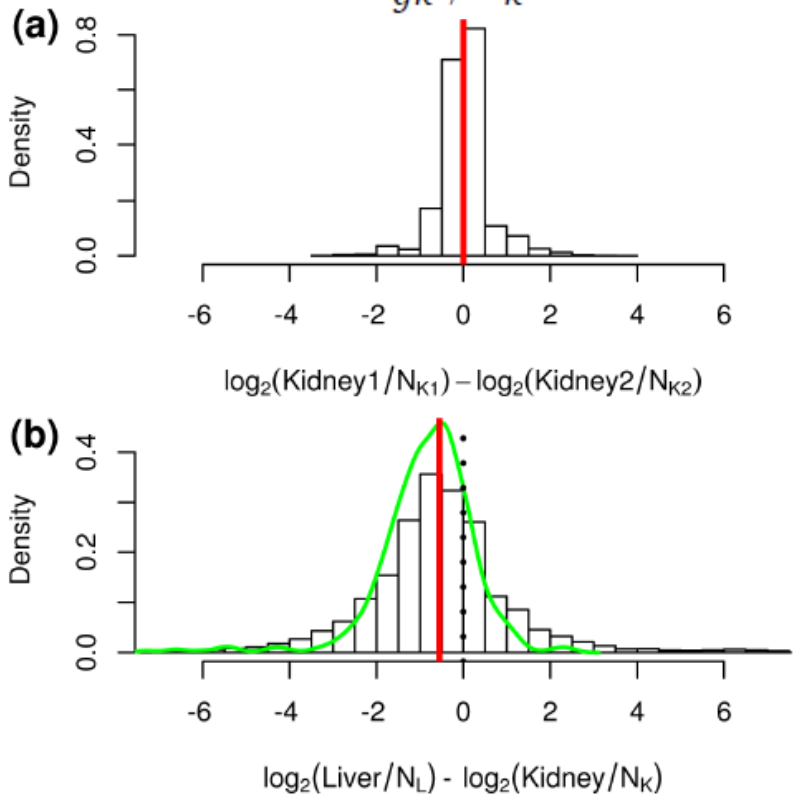
- RNA-Seq *vs*. DNA芯片
  - 不考虑多重定位读段造成部分基因RPKM值偏低
  - 考虑多重定位读段提高与芯片结果的线性关系



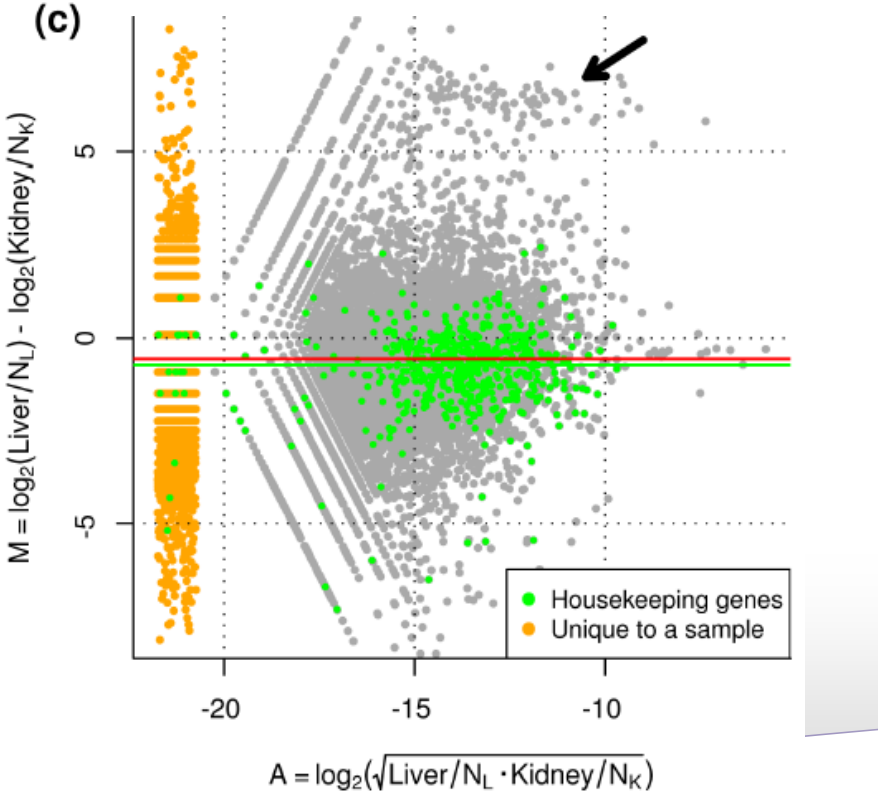
归一化：不同样本之间的比较

- 不同样本相同基因的比较不需要考虑长度因素
- Y_{gk}：基因g的读段数目，N_k：读段总数

$$M_g = \log_2 \frac{Y_{gk}/N_k}{Y_{gk'}/N_{k'}}$$



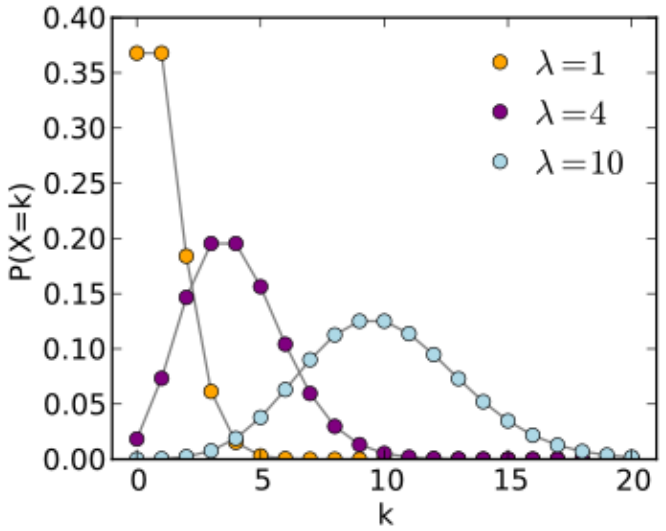
$$A_g = \frac{1}{2}\log_2\left(Y_{gk}/N_k \bullet Y_{gk'}/N_{k'}\right) \text{ for } Y_{g\bullet} \neq 0$$

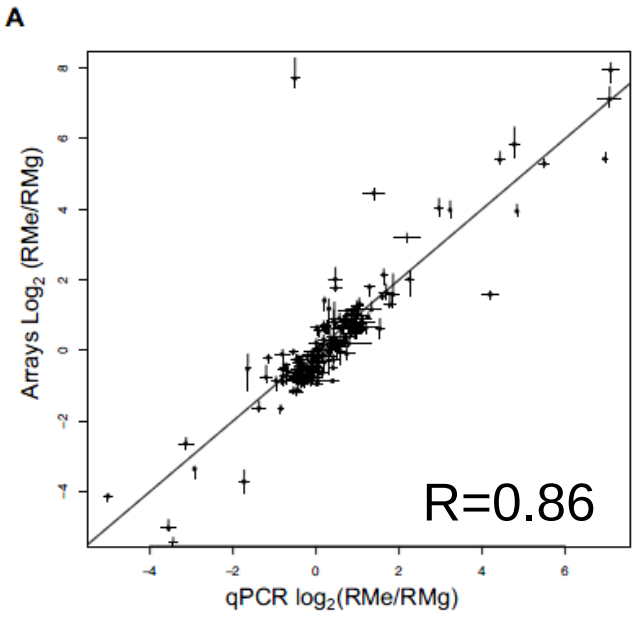


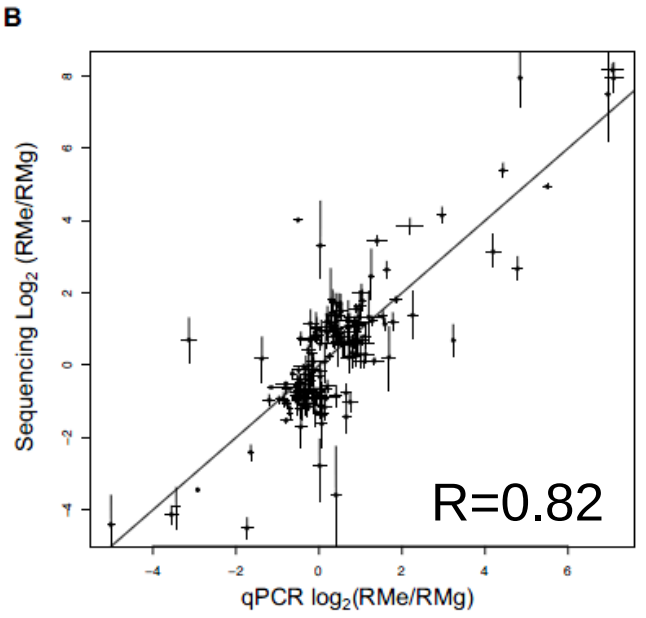
差异表达基因分析

- Lander-Waterman model
  - 转录本中*C*个分子，测序获得*N*个读段
  - 测到某个分子的概率为$\lambda=N/C$
- 截短的 (Truncate) 泊松过程
  - 不知道有多少分子被测0次
  - 读段个数分布在L~R之间
- 第i个分子测到x_i个读段符合泊松分布
  - $\Pr(X_i|\lambda) = \frac{1}{K_{L,R}(\lambda)} \cdot \frac{e^{-\lambda}\lambda^{x_i}}{X_i!}$
  - $K_{L,R}(\lambda) = \sum_{x=L}^{R} Pr(X_i|\lambda) \cong 1 - poisson(0,\lambda)$

- λ < 10：有偏分布
- λ ~ 10：近似为正态分布
- 差异表达基因分析
  - 利用Fisher's exact test
  - qPCR与芯片结果拟合更好！？

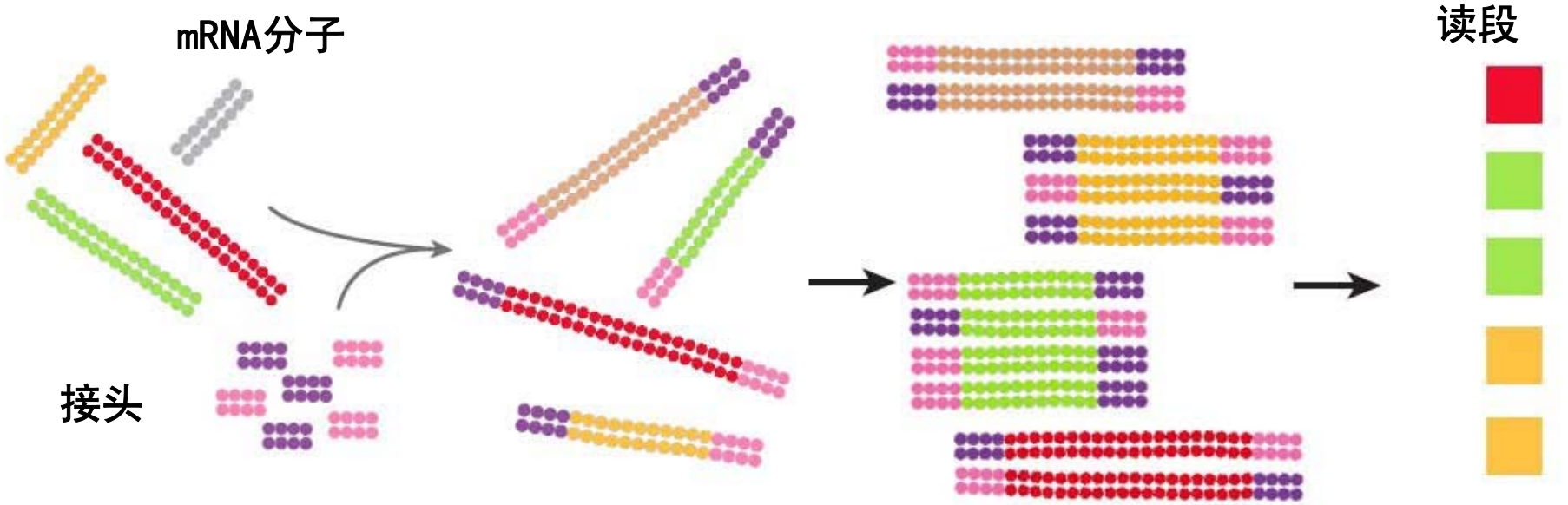






二次泊松分布？

- **TSPM: two-stage Poisson model**
  - 测序可以看成是泊松过程
  - 建库也可以近似看成是泊松过程
  - 两者的λ不同，需要分别估计



负二项分布：DESeq

- NB: Negative Binomial (Pascal) Distribution
  - 每次试验的成功率*p*，失败率1-*p*
  - 经历*r*次失败后观察到*k*次成功事件，则

$$f(k;r,p) \equiv \Pr(X=k) = \binom{k+r-1}{k} p^k (1-p)^r \quad \text{for } k = 0,1,2,\ldots$$

- RNA-Seq数据的归一化
  - 总共*m*个样本
  - 基因*i*在样本*j*中检测到K_{ij}个读段
  - 样本*j*的采样深度s_j

$$s_j = \underset{i}{median} \frac{K_{ij}}{\left(\prod_{v=1}^{m} K_{iv}\right)^{1/m}}$$

负二项分布：DESeq

- RNA-Seq数据的归一化
  - 总共*m*个样本
  - 基因*i*在样本*j*中检测到K_{ij}个读段
  - 样本*j*的采样深度s_j
  - 样本*j*的条件*p(j)*
  - q_{ip} 基因*i*在条件*p*下的平均归一化表达量

$$q_{ip} = \frac{1}{\#\text{ of replicates}} \sum_{\text{j in replicates}} \frac{K_{ij}}{s_j}$$

$$\mu_{ij} = q_{ip(j)} s_j \qquad \sigma_{ij}^2 = \mu_{ij} + s_j^2 v_p\left(q_{ip(j)}\right)$$

用最大似然性估算 (MLE) 获得

$$K_{ij} \sim NB\left(\mu_{ij}, \sigma_{ij}^2\right)$$

DESeq：差异表达基因分析

- 两样本的*p*-value计算
  - $p(a, b) = Pr(K_{iA} = a)\ Pr(K_{iB} = b)$
  - $K_{iS} = K_{iA} + K_{iB}$

$$p_i = \frac{\sum_{\substack{a+b=k_{iS} \\ p(a,b)\le p(k_{iA},k_{iB})}} p(a,b)}{\sum_{a+b=k_{iS}} p(a,b)}.$$

Anders and Huber *Genome Biology* 2010, **11**:R106
http://genomebiology.com/2010/11/10/R106

Genome Biology

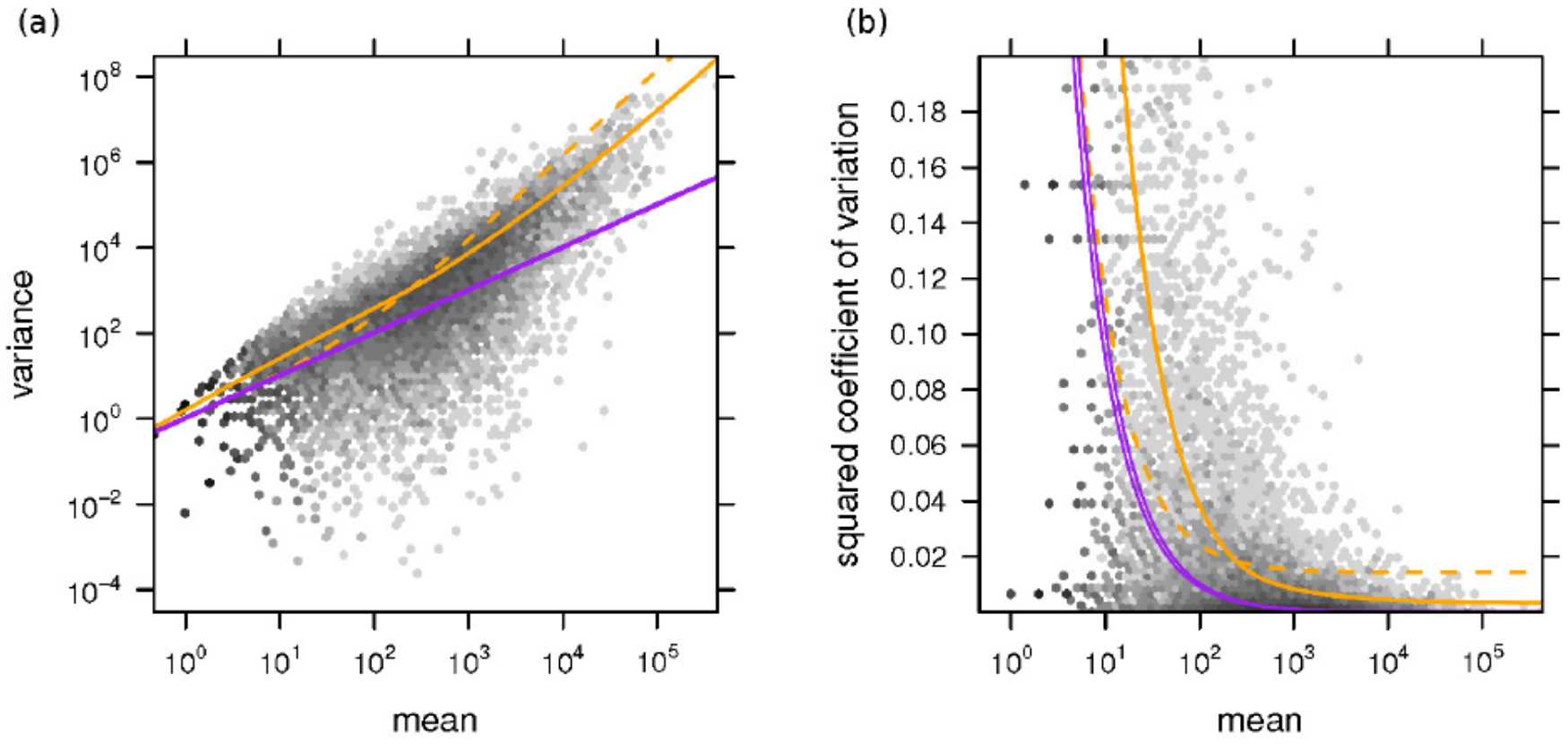
METHOD Open Access

Differential expression analysis for sequence count data

Simon Anders*, Wolfgang Huber

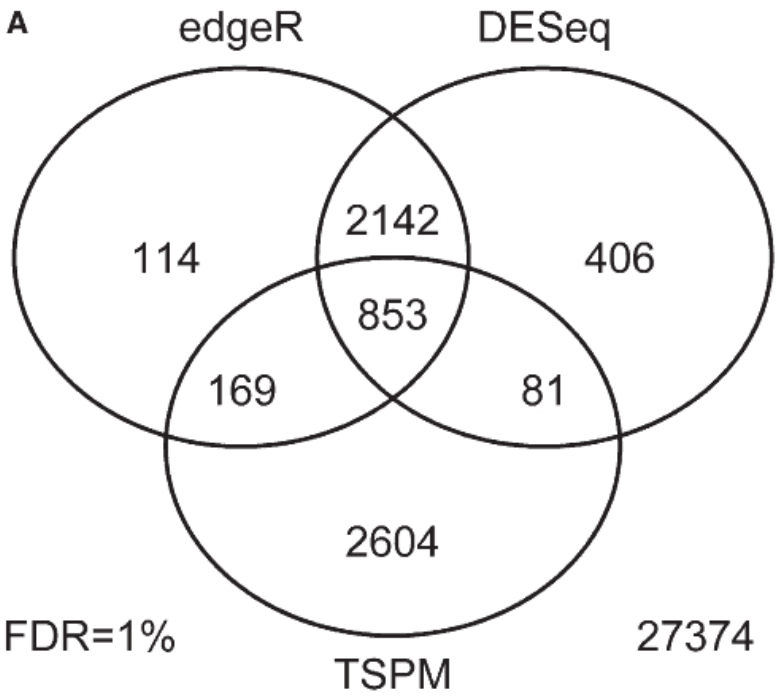
负二项分布 *vs.* 泊松分布

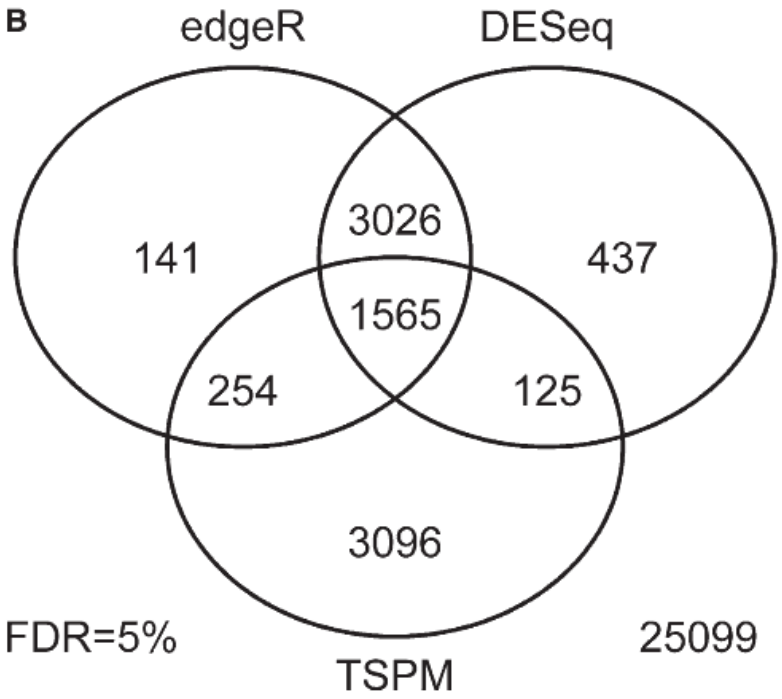
- DESeq (橙)，EdgeR (橙虚线)，泊松 (紫)
  - NB估算的方差更大



负二项分布 vs. 二次泊松分布

- DESeq和EdgeR的重合较高：NB分布
- TSPM (二次泊松) 与其他工具重合较低





natics,

RNA-Seq差异表达基因分析工具

- 目前最好：DESeq和EdgeR
  - DESeq: http://bioconductor.org/packages/release/bioc/html/DESeq.html
  - EdgeR: http://bioconductor.org/packages/release/bioc/html/edgeR.html
- R语言 + Bioconductor实现

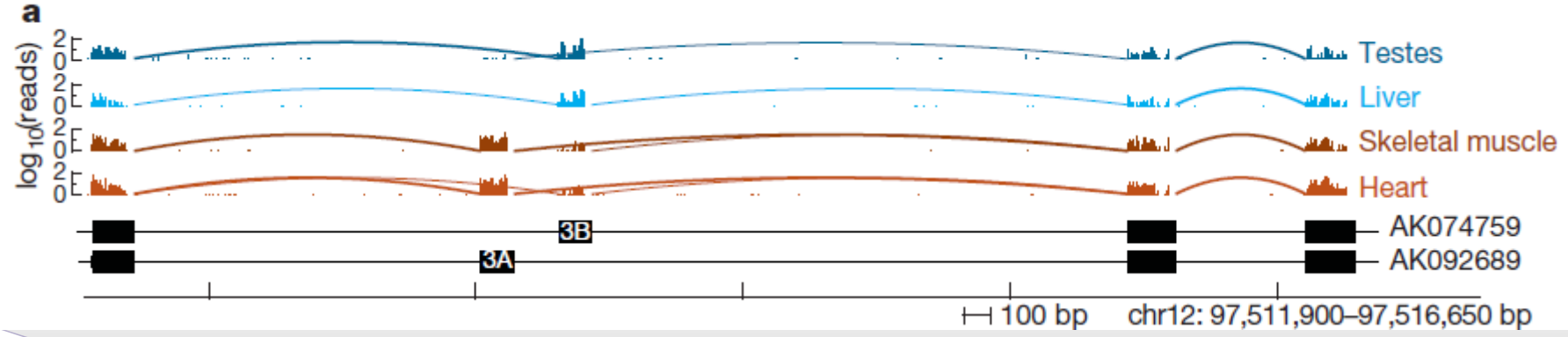
Table 1 A comparison of common statistical methods for RNA-Seq differential gene expression analysis

Method	Underlying distribution	Recommended with biological replicates	Multi-group comparison	R/Bioconductor package	Reference
Fisher's exact Test	Poisson	No	No	No	[27]
Likelihood ratio test	Poisson	No	Yes	No	[21]
edgeR	Negative Binomial	Yes	Yes	Yes	[28]
DESeq	Negative Binomial	Yes	No	Yes	[15]
baySeq	Negative Binomial	Yes	Yes	Yes	[17]
BBSeq	Beta-Binomial	Yes	No	Yes	[29]
Two-stage poisson model	Poisson	Yes	Yes	No	[30]

可变剪切

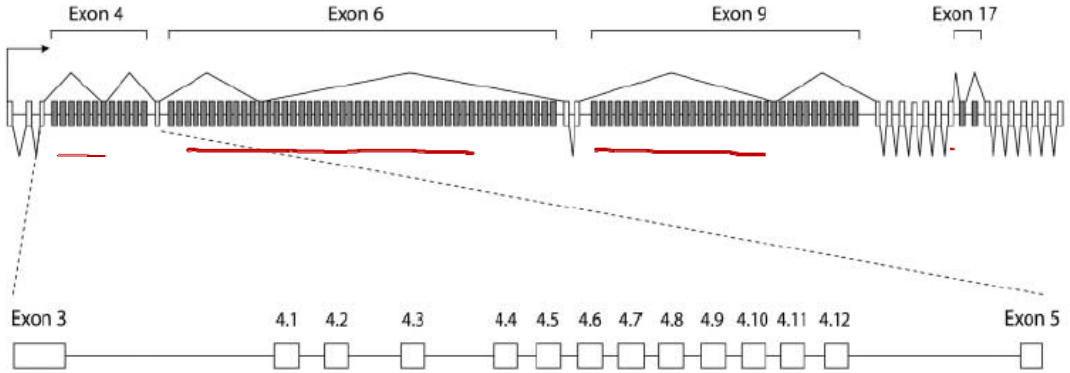
- AS, Alternatively Splicing
  - 92–94%的人类多外显子基因存在可变剪切现象
- Major isoform vs. minor isoform
- 86%包括至少有一个次要异构体 (比例15%)

SLC25A3基因

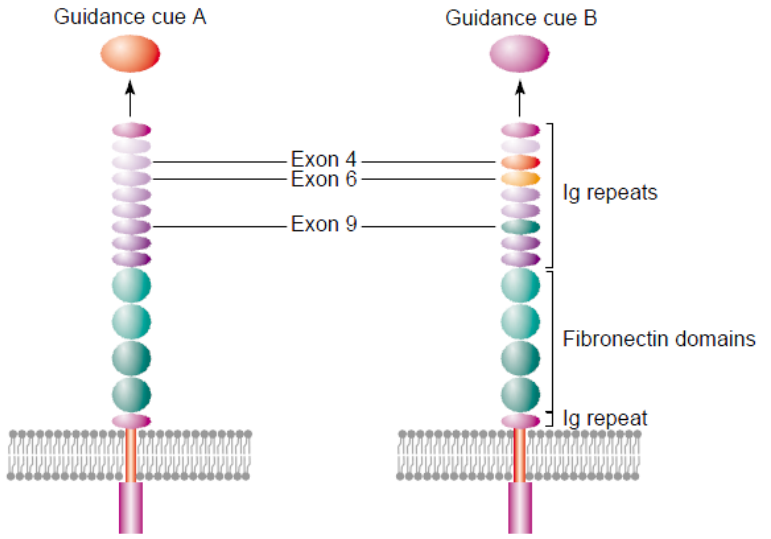


果蝇Dscam基因的可变剪切

- 12*48*33*2=38,016个不同的异构体



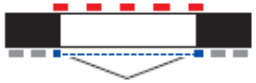
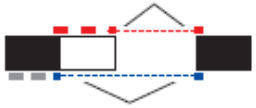
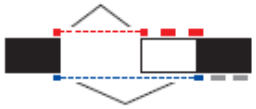
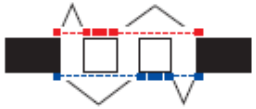
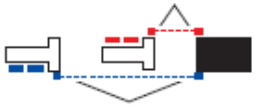
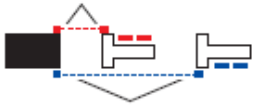
- 果蝇: ~250,000个神经元
- 轴突导向 (Axon guidance)
- 调控神经发育
  - 神经元之间的连接



Bioinformatics, 2020, I

可变剪切的类型

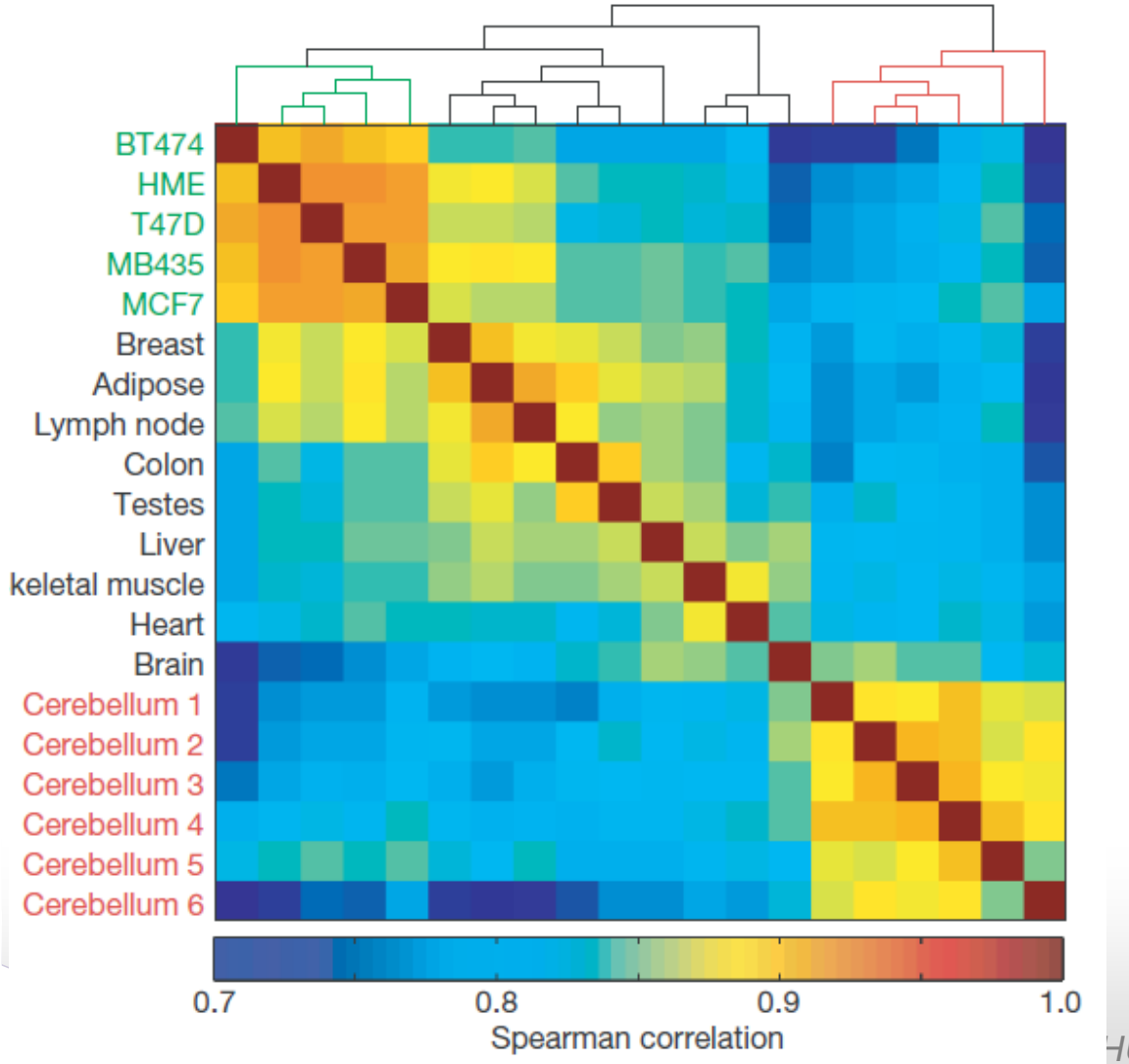
- 8种类型
- 仅统计外显子上的读段数 – 不全面

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

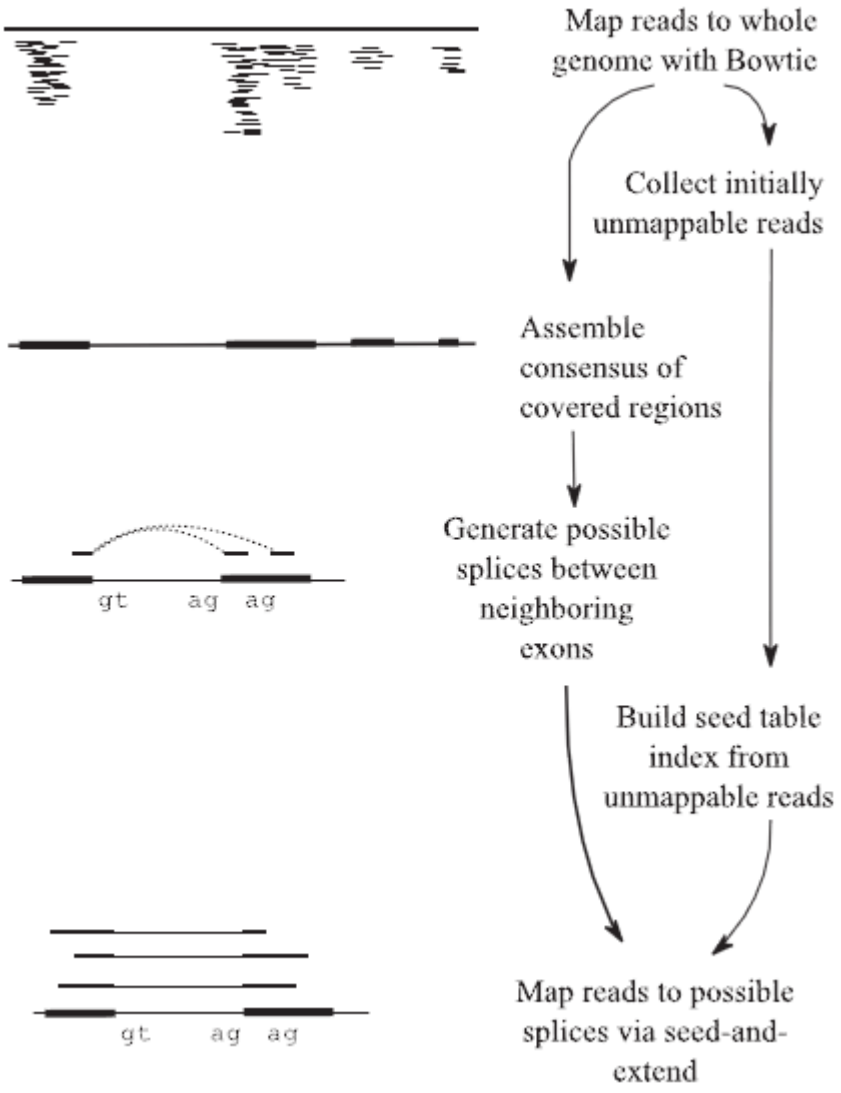
可变剪切的组织特异性



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

可变剪切鉴定：Cufflinks

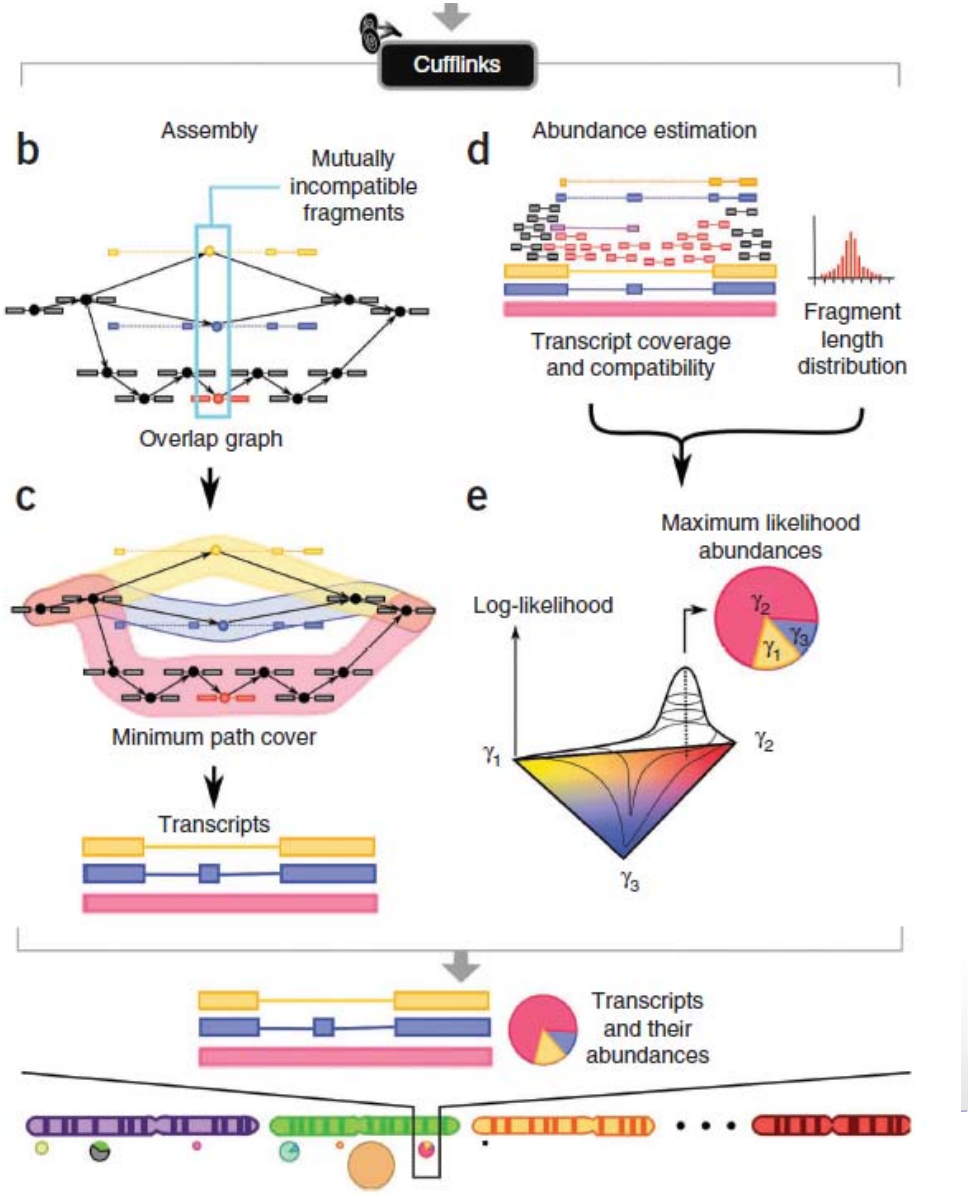
- 用Bowtie (领带) 将所有读段映射到基因组上
- 不能完美匹配的，用TopHat (礼帽) 发现剪切位点



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

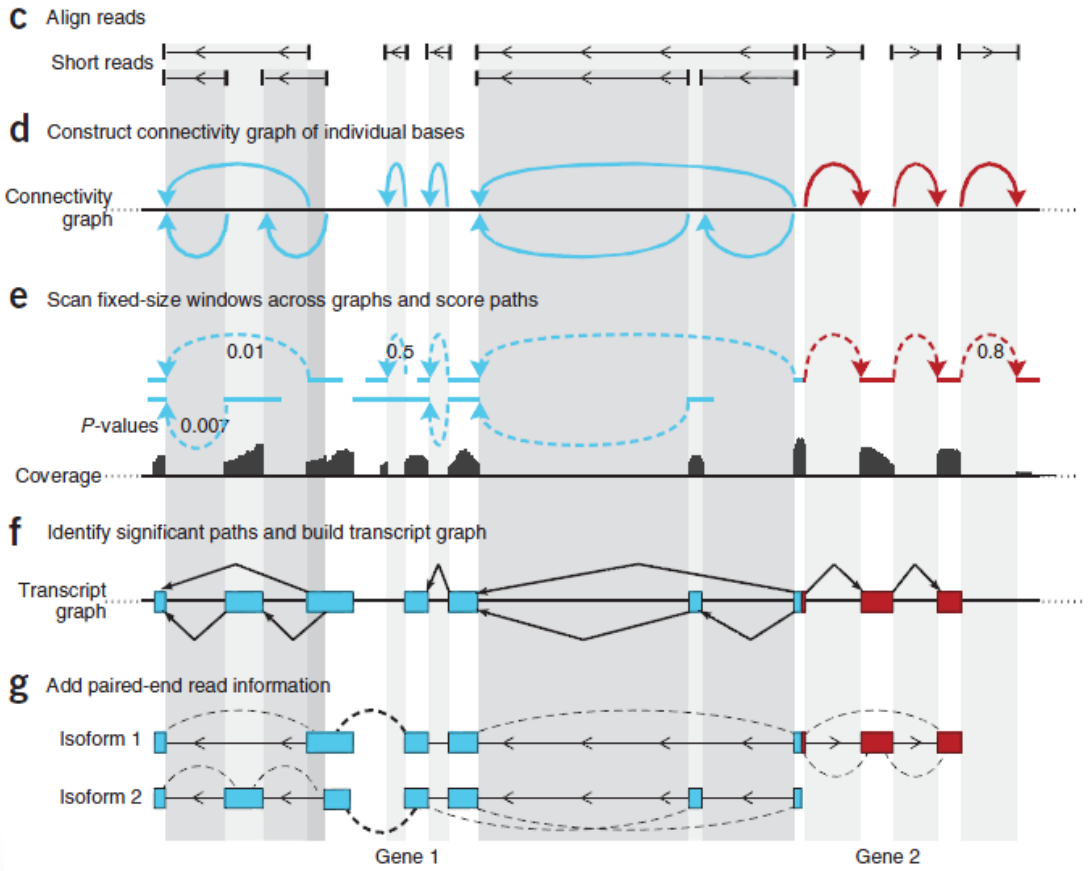
可变剪切鉴定：Cufflinks

- 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



可变剪切鉴定：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/

常用的RNA-Seq分析软件

Table 2 | List of publicly available RNA-seq software packages discussed in this review

	Primary category	Discovery	Need genomic assembly	Associated read mapper	Splice junctions	Quantitation	Reference
ABySS v1.0.11	Short-read assembler	Yes	No	NA	Assembled	Read coverage	33
BASIS V1	Existing transcript quantitation	No	Yes	External	From existing models	Read coverage	41
ERANGE v3.1	Existing and novel gene quantitation	Yes	Yes	Blat Bowtie Eland	From existing models Novel with blat	RPKM from gene annotations and novel transfrags	18
G-Mo.R-Se v1.0	Novel gene model annotation	Yes	Yes	SOAP	Predicted from transfrags	No	37
QPALMA v0.9.9.2	Spliced read mapper	Yes	Yes	Integrated	Predicted from transfrags	No	38
RNA-mate v1.1	Existing and novel gene quantitation	Yes	Yes	Map reads	From existing models	Deprecated in v1.1	36
RSAT v0.0.3	Existing transcript quantitation	No	No; requires transcript sequences	Eland, SeqMap (bundled)	From supplied transcript sequences	RPKM from transcript sequences	40
TopHat v1.0.10	Existing and novel gene quantitation	Yes	Yes	Bowtie	Predicted from transfrags From existing models	RPKM from supplied annotations	32
Velvet v0.7.47	Short read assembler	Yes	No	NA	Assembled	Fold coverage	31

NA, short-read assemblers do not rely on any particular annotated read mapper to assemble the transcripts.

表观遗传学 (Epigenetics)

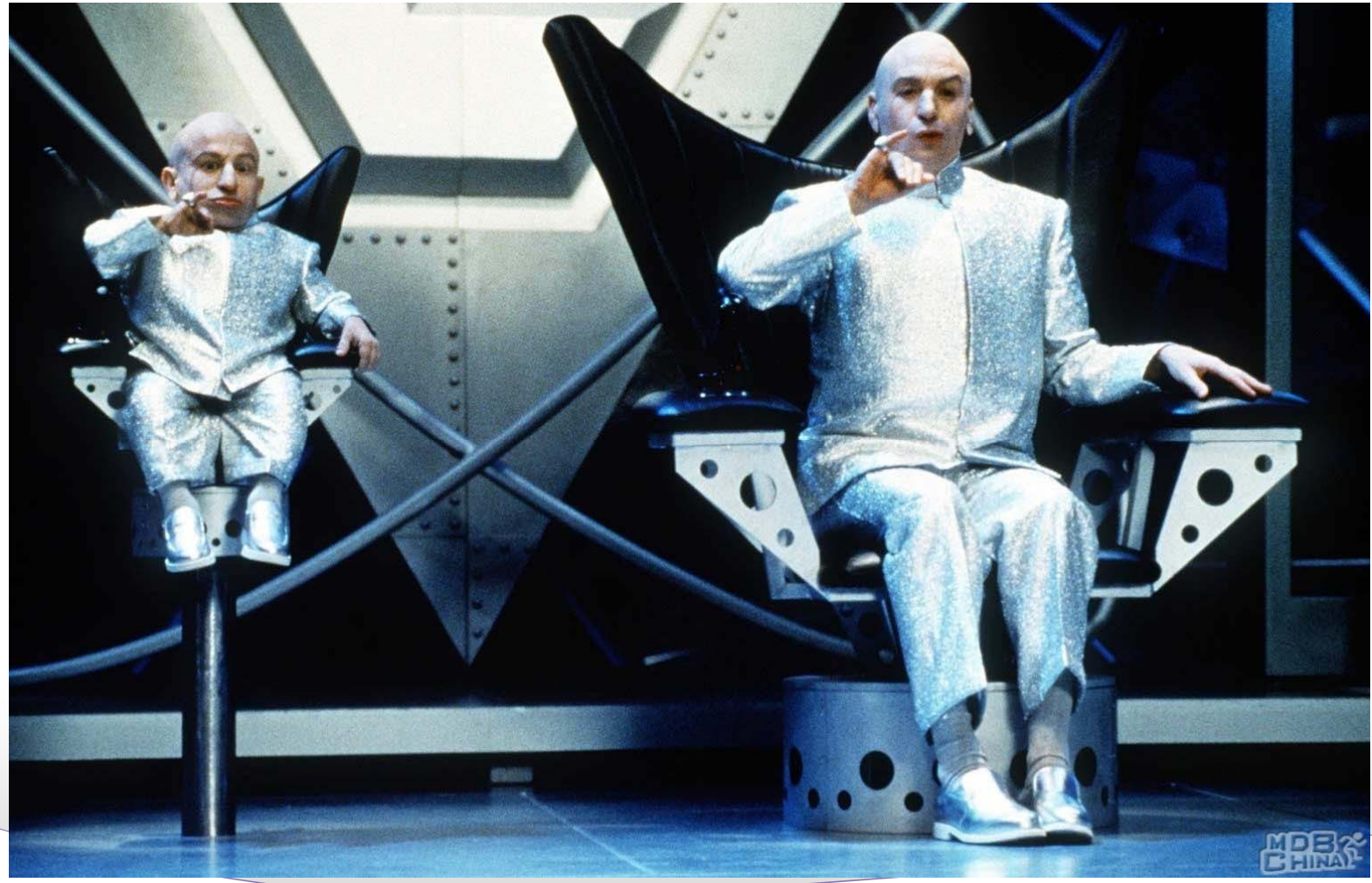
Dolly　　Ian Wilmut

多莉：生于1996年7月5日，死于2003年2月14日

Dr. Evil

Dr. Evil's Clone
"Mini Me"

Dr. Evil



The calico cat

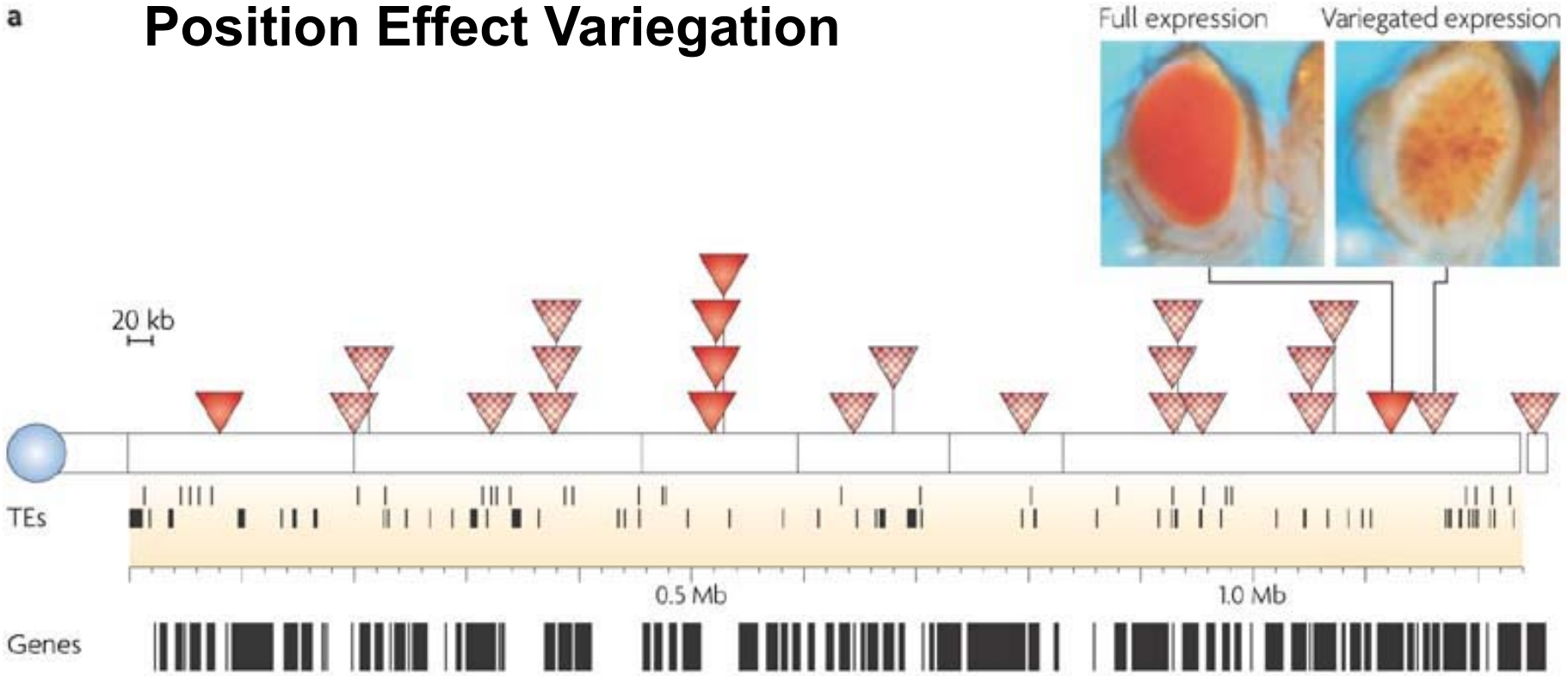
Rainbow

?

Allie

Copy Cat

PEV: 位置效应斑

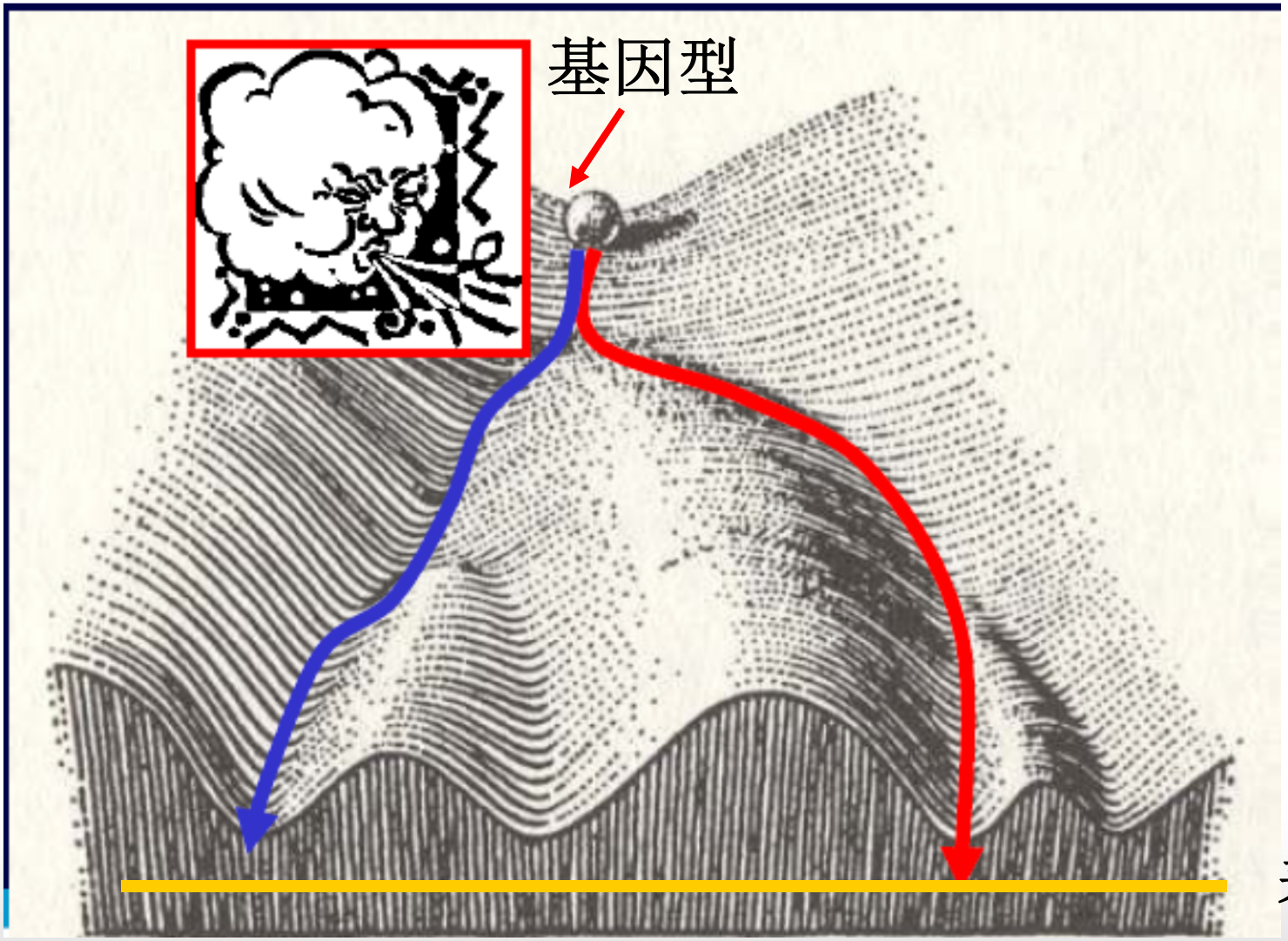


- white 基因 vs. 转座子

表观遗传学

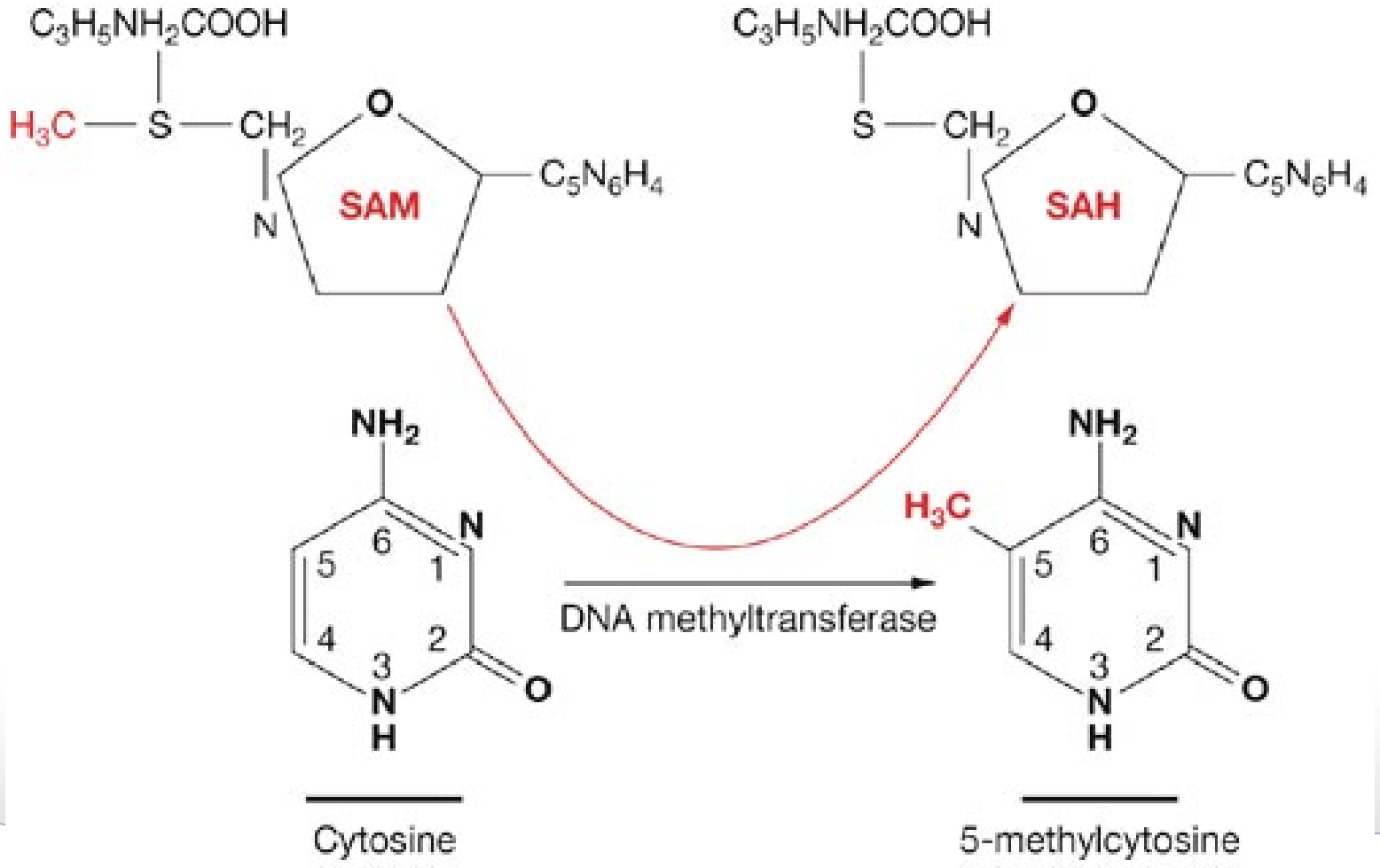
- 基因的DNA序列不发生改变的情况下，基因的表达水平与功能发生改变，并产生可遗传的表型
- 特征: (1)可遗传；(2) 可逆性；(3) DNA不变
- 表观遗传学机制：
  - DNA甲基化
  - 组蛋白修饰
  - MicroRNA
  - 染色质重塑
  - …

Waddington's epigenetics

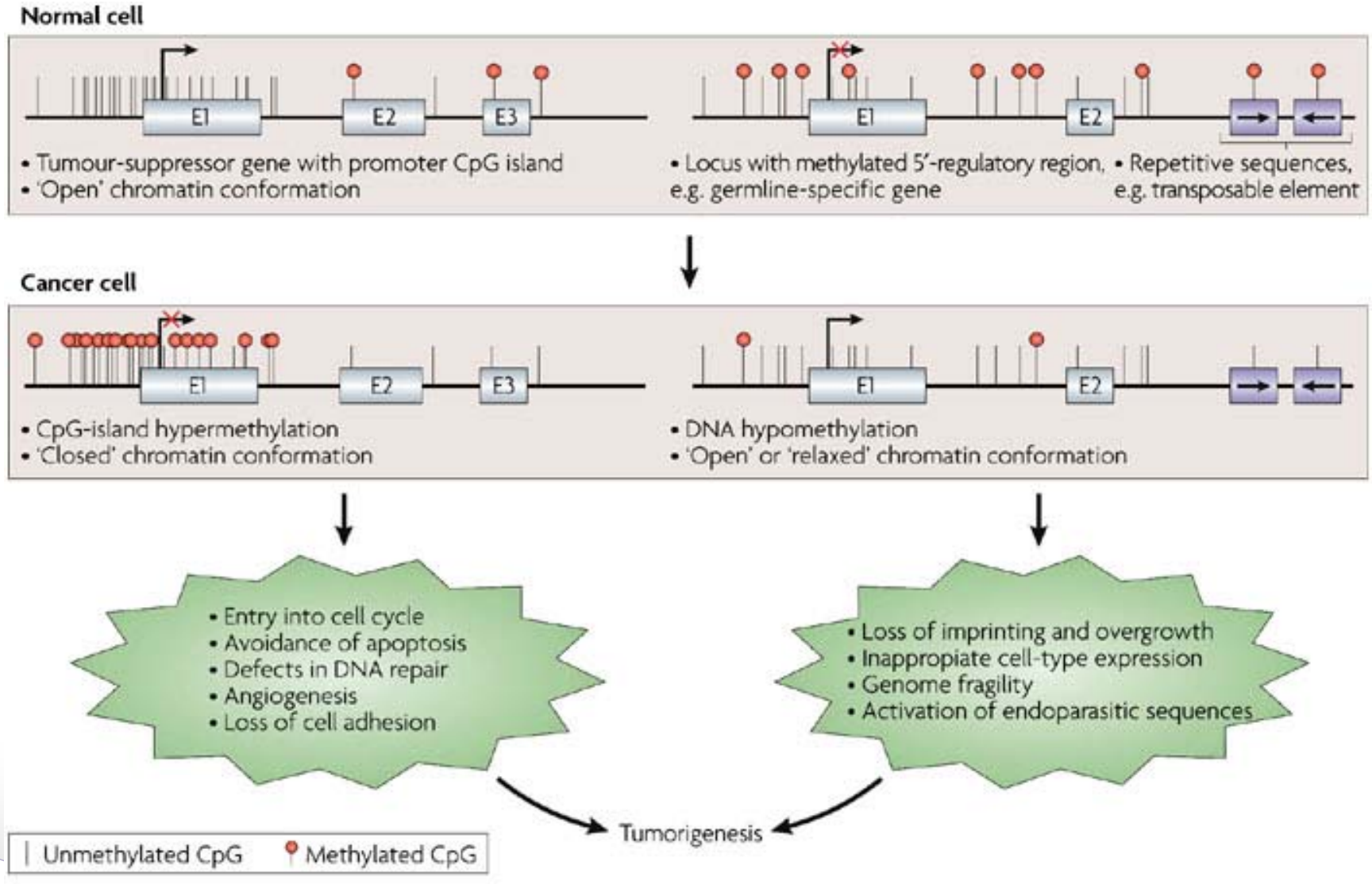


DNA甲基化

SAH, S-adenosylhomocysteine; SAM, S-adenosylmethionine



DNA甲基化



Nature Reviews | Genetics

叶酸: 很重要?

Google 叶酸 Google 搜索 高级搜索 | 使用偏好

◉所有网页 ○中文网页 ○简体中文网页 ○中国的网页

网页 约有**2,460,000**项符合**叶酸**的查询结果，以下是第1-

叶酸_百度百科

叶酸（Folic acid）维生素B复合体之一，相当于蝶酰谷氨酸（pteroylglutamic acid，PGA）。是米切尔（H. K. Mitchell，1941）从菠菜叶中提取纯化的，而命名为叶酸。 ...

baike.baidu.com/view/8114.htm - 41k - 网页快照 - 类似网页

孕期如何补充叶酸

叶酸，是一种B族维生素，也叫做B9。你的身体需要用它来生成红细胞、去甲肾上腺素和色拉托宁（神经系统的化学成分）。叶酸协助合成DNA（身体的遗传物质），维持大脑的 ...

cn.babycenter.com/pregnancy/nutrition/folicacid/ - 38k - 网页快照 - 类似网页

孕妇不要过量服用叶酸

2007年1月21日 ... 据英国报道，新的研究显示，妇女在怀孕期间大剂量服用叶酸，今后患乳腺癌的概率可能增加。 这一发现对妇女怀孕期间广泛服用的维生素的安全性带来疑问。 ...

news.xinhuanet.com/health/2007-01/21/content_5631511.htm - 70k - 网页快照 - 类似网页

巧补叶酸宝宝聪明_新华网

2008年5月22日 ... 近日，一五七医院妇产科郭会平主任接受记者采访时指出——怀孕之后补充叶酸对预防胎儿畸形作用不大，要防胎儿神经管畸形最佳补充叶酸的时间应该是怀孕 ...

news.xinhuanet.com/lady/2008-05/22/content_8211613.htm - 22k - 网页快照 - 类似网页

孕期营养明星之叶酸访谈-搜狐母婴

主持人：各位，今天邀请的嘉宾是孕期营养明星——叶酸！叶酸，你好！ 叶酸：主持人好， 大家好！ 主持人：叶酸，昨天看你还像一汪水，今天怎么又变成片剂了？ ...

baobao.sohu.com/20070412/n249381255.shtml - 107k - 网页快照 - 类似网页

孕妈咪：叶酸虽小别忘补_健康中心_新浪宝贝_新浪网

别看叶酸在人体内来起来似乎不太起眼，可它却是蛋白质和核酸合成的必需因子，血红蛋白、红细胞、白细胞快速增生、氨基酸代谢、大脑中长链脂肪酸如DNA的代谢等都少不了 ...

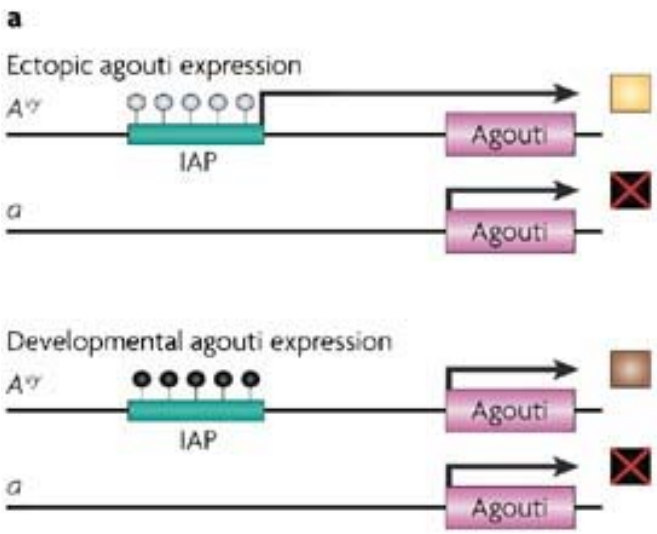
baby.sina.com.cn/health/2004-09-30/47_65377.shtml - 62k - 网页快照 - 类似网页

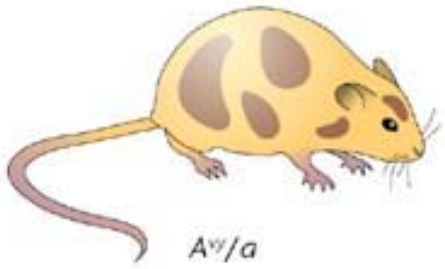
主持人：最近，我看了一篇关于你的报道，说是：成年妇女每天需摄入50—100微克，而妊娠期妇女上升为300—400毫克。如果孕妇体内缺乏叶酸的话，会导致胎儿脊柱裂、脑脊膜膨出等神经管畸形，甚至产下“无脑儿”，也会造成流产。缺乏叶酸，除了会造成先天畸形儿外，还会给孕妇带来危险。孕妇将会出现贫血症状，严重时还会导致贫血性心脏病、妊娠期高血压病、胎盘早剥、早产和产褥感染等。

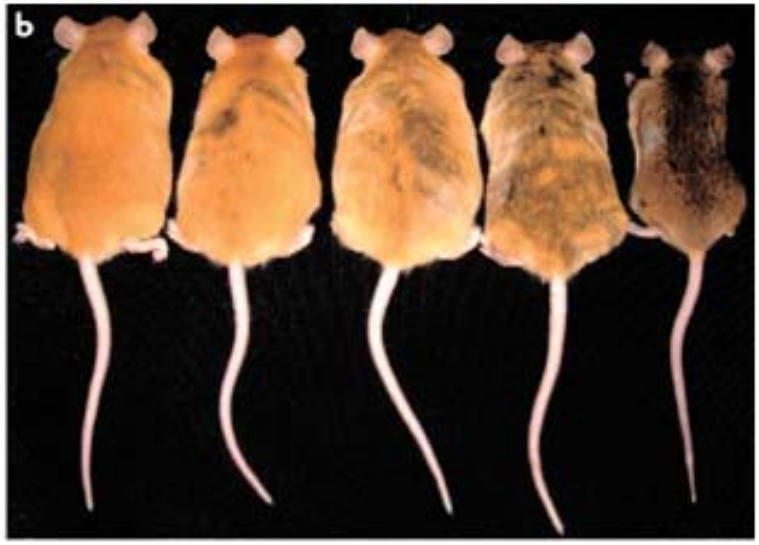
叶酸：是的。卫生部曾经对全国12万多名围产儿进行调查，结果无脑儿、开放性脊柱裂、脑脊膜膨出等神经管畸形，分别居于出生缺陷的第一、三、九位！这些都是因为我叶酸的缺乏而造成的。

DNA甲基化 vs. 转座子

- 叶酸：很重要
- IAP: intracisternal A particle
  - 非甲基化：异常表达Agouti
  - 甲基化：正常表达

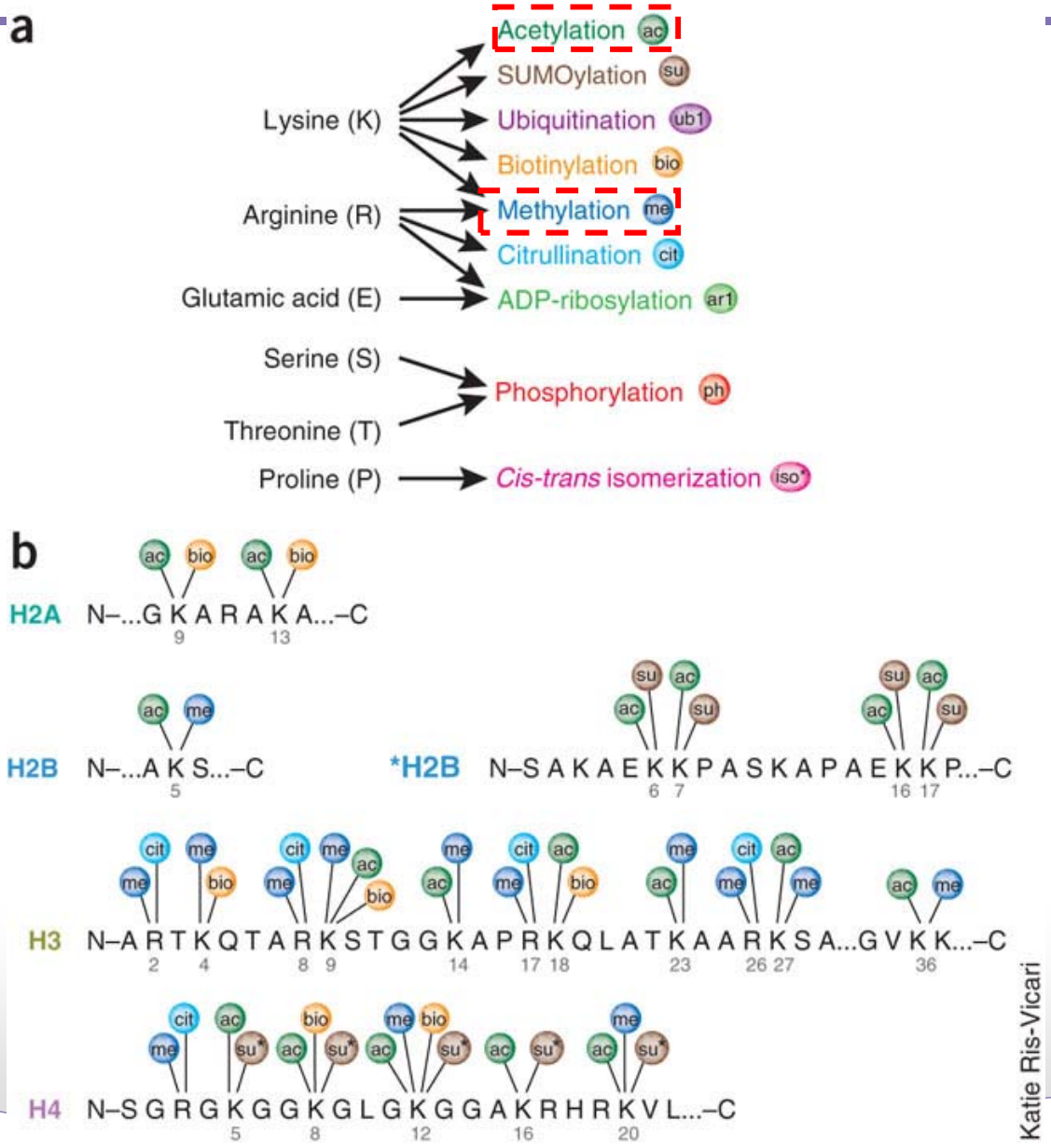






Nature Reviews | Genetics

组蛋白的共价修饰

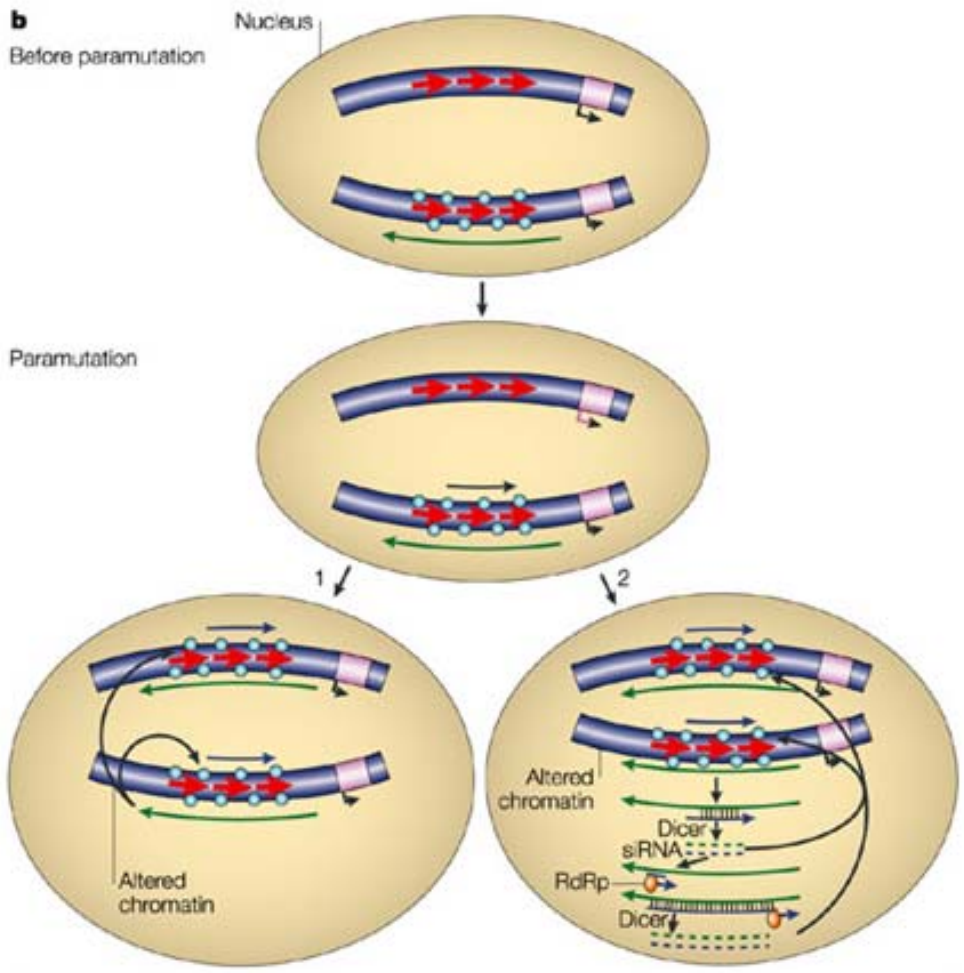

a
Lysine (K)
Arginine (R)
Glutamic acid (E)
Serine (S)
Threonine (T)
Proline (P)
Acetylation ac
SUMOylation su
Ubiquitination ub1
Biotinylation bio
Methylation me
Citrullination cit
ADP-ribosylation ar1
Phosphorylation ph
Cis-trans isomerization iso*
b
H2A N–...G K A R A K A...–C
H2B N–...A K S...–C
*H2B N–S A K A E K K P A S K A P A E K K P...–C
H3 N–A R T K Q T A R K S T G G K A P R K Q L A T K A A R K S A...G V K K...–C
H4 N–S G R G K G G K G L G K G G A K R H R K V L...–C
Katie Ris-Vicari

Paramutation

- 一个沉默的基因常常导致等位基因也发生沉默
- maize P-rr gene：Myb 转录因子

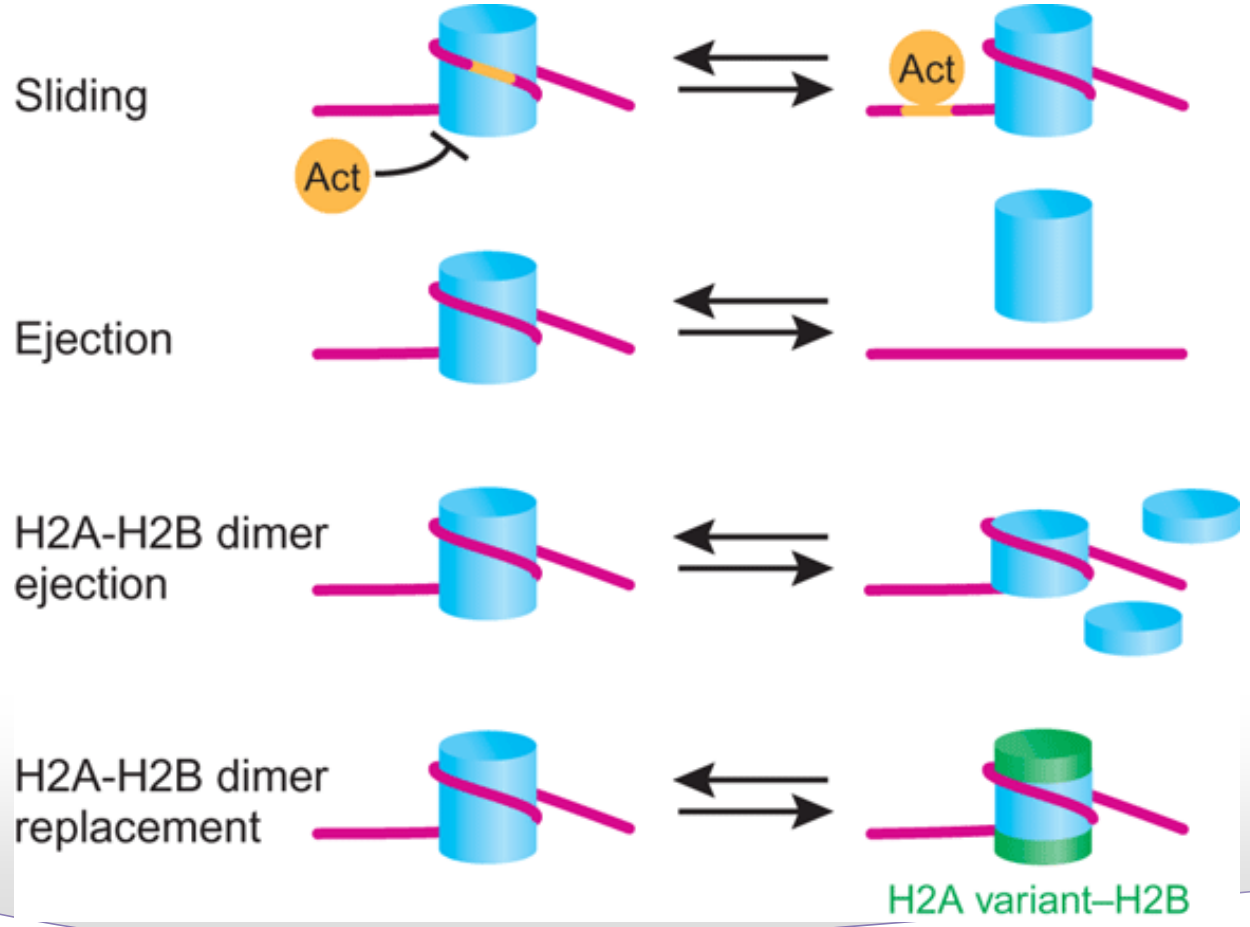
Paramutation

- 分子机理 (RNA-mediated trans-induction of chromatin)：RNA参与调控

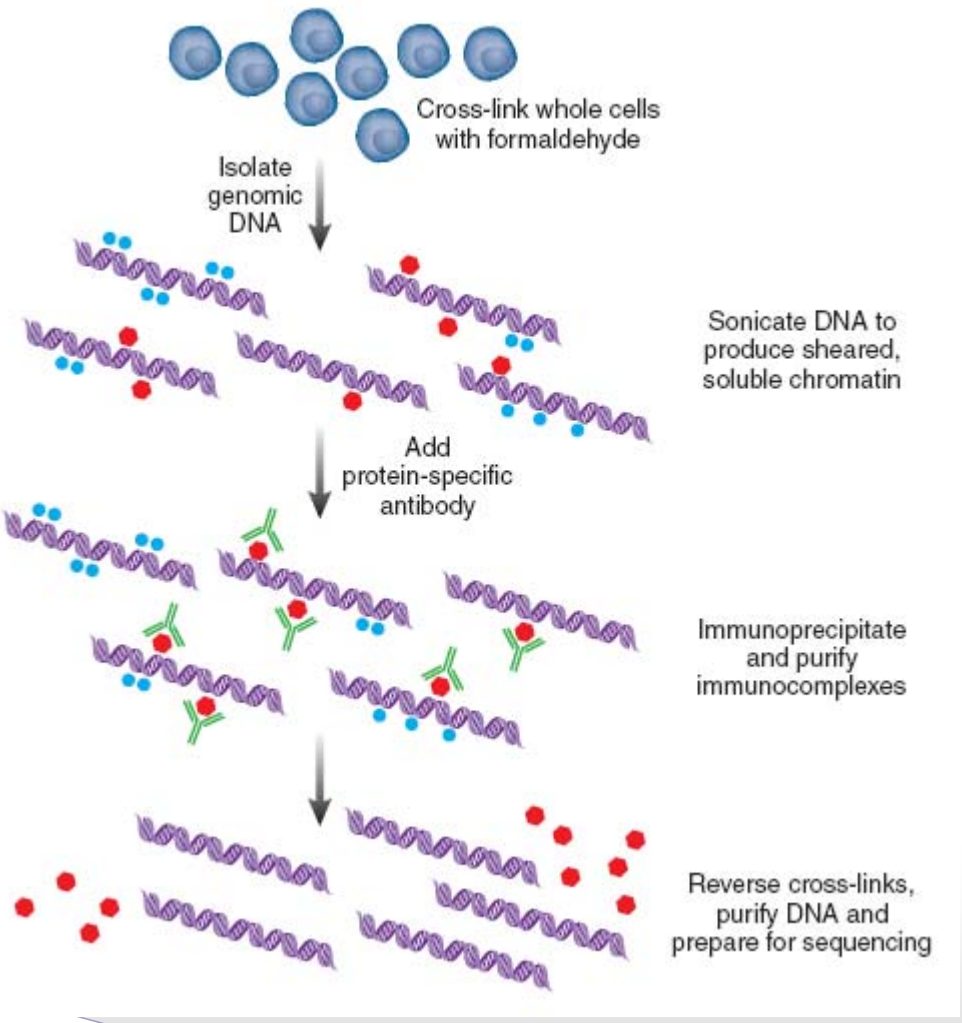


染色质重塑

- 染色质 (Chromatin): 染色体上高度致密的部分，通常不表达基因



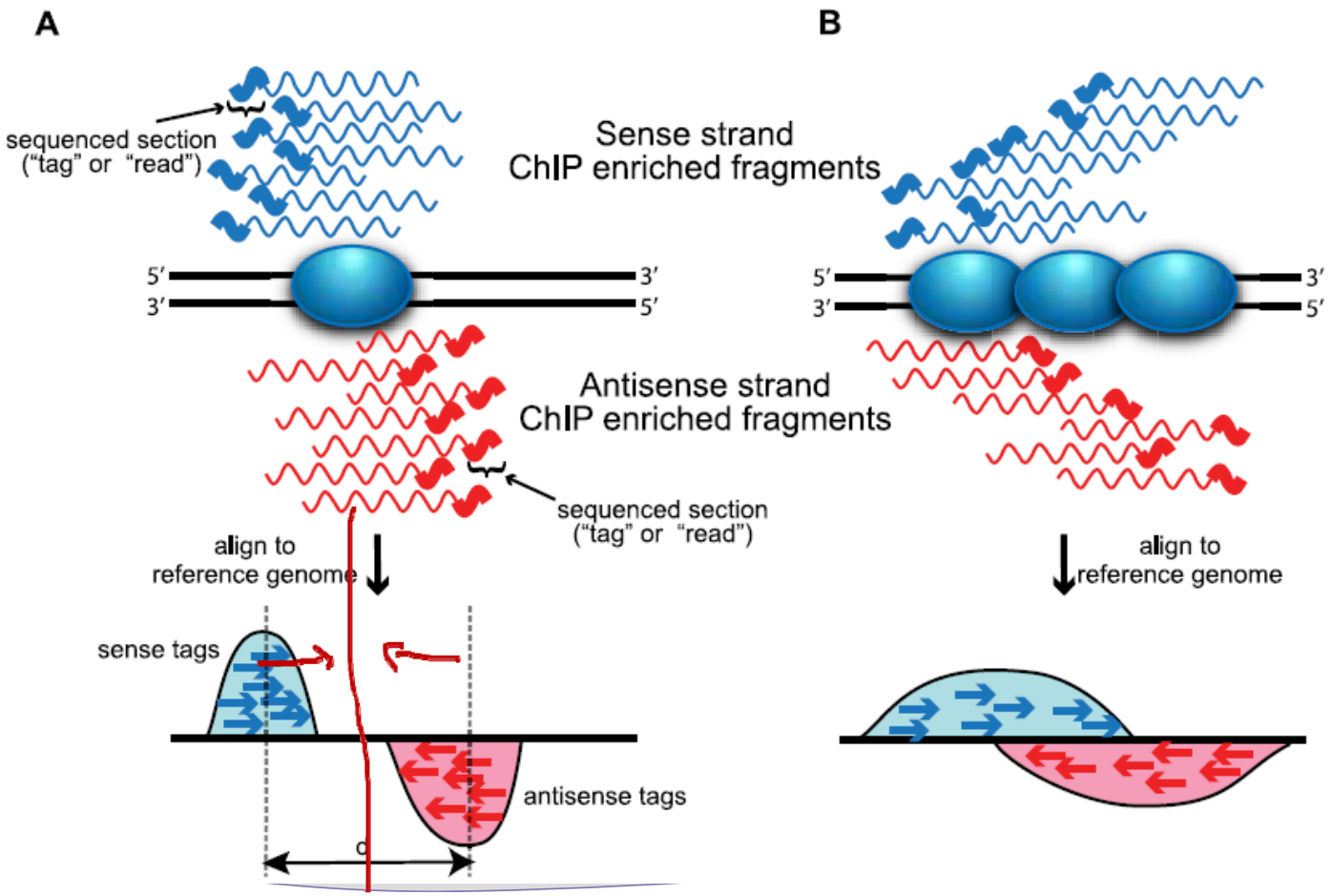
ChIP-Seq



- 转录因子、组蛋白等与DNA结合
- **Chromatin immunoprecipitation (ChIP) 结合高通量测序**

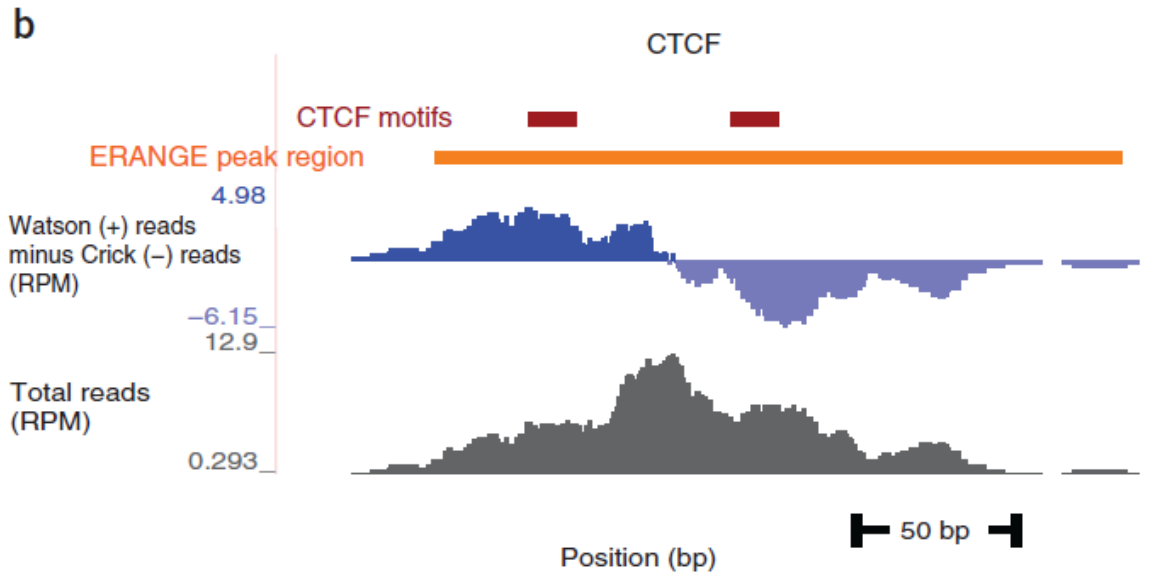
ChIP-Seq的读段分布

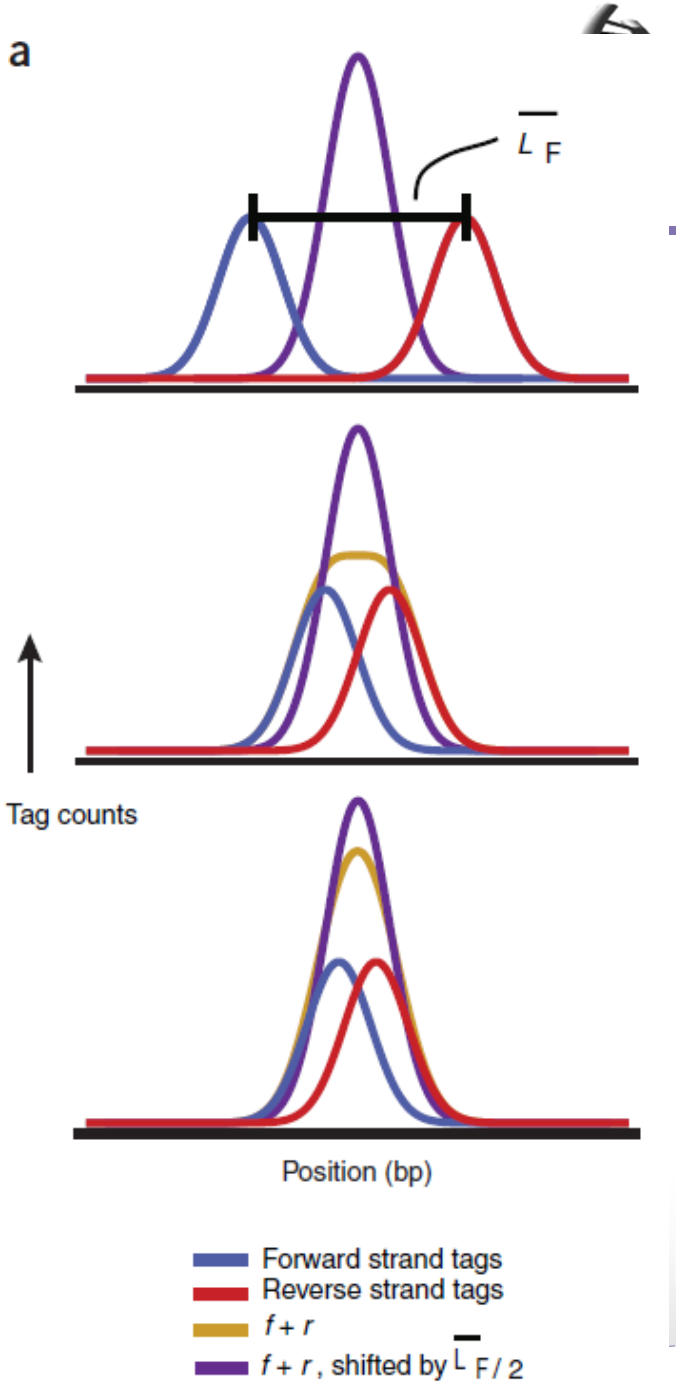
- 不同蛋白质与DNA结合的模式不同
  - A. 转录因子；B. RNA聚合酶和组蛋白质等



ChIP-Seq校正

- 测序读段的双峰校正为单峰
- 例如，CTCF结合双峰的校正





常见的ChIP-Seq数据分析工具

Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

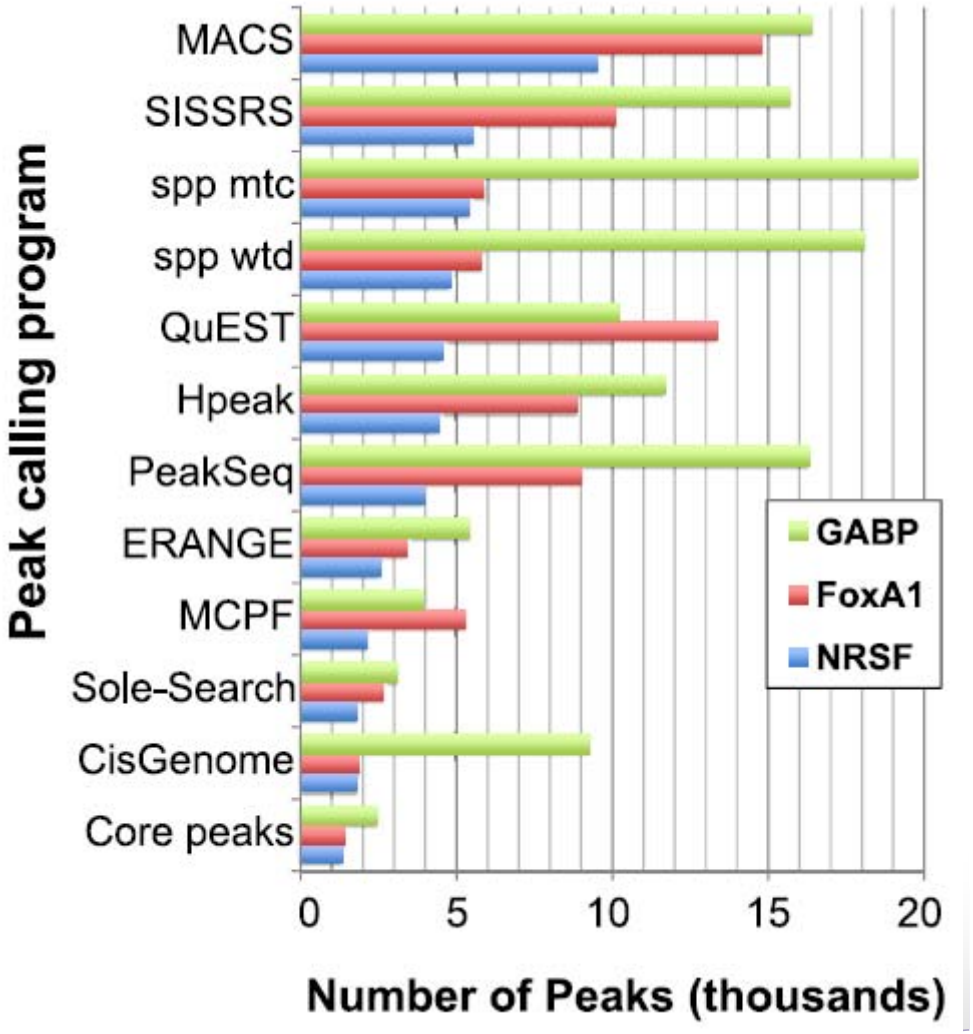
X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method exludes putative duplicated regions, no treatment of deletions

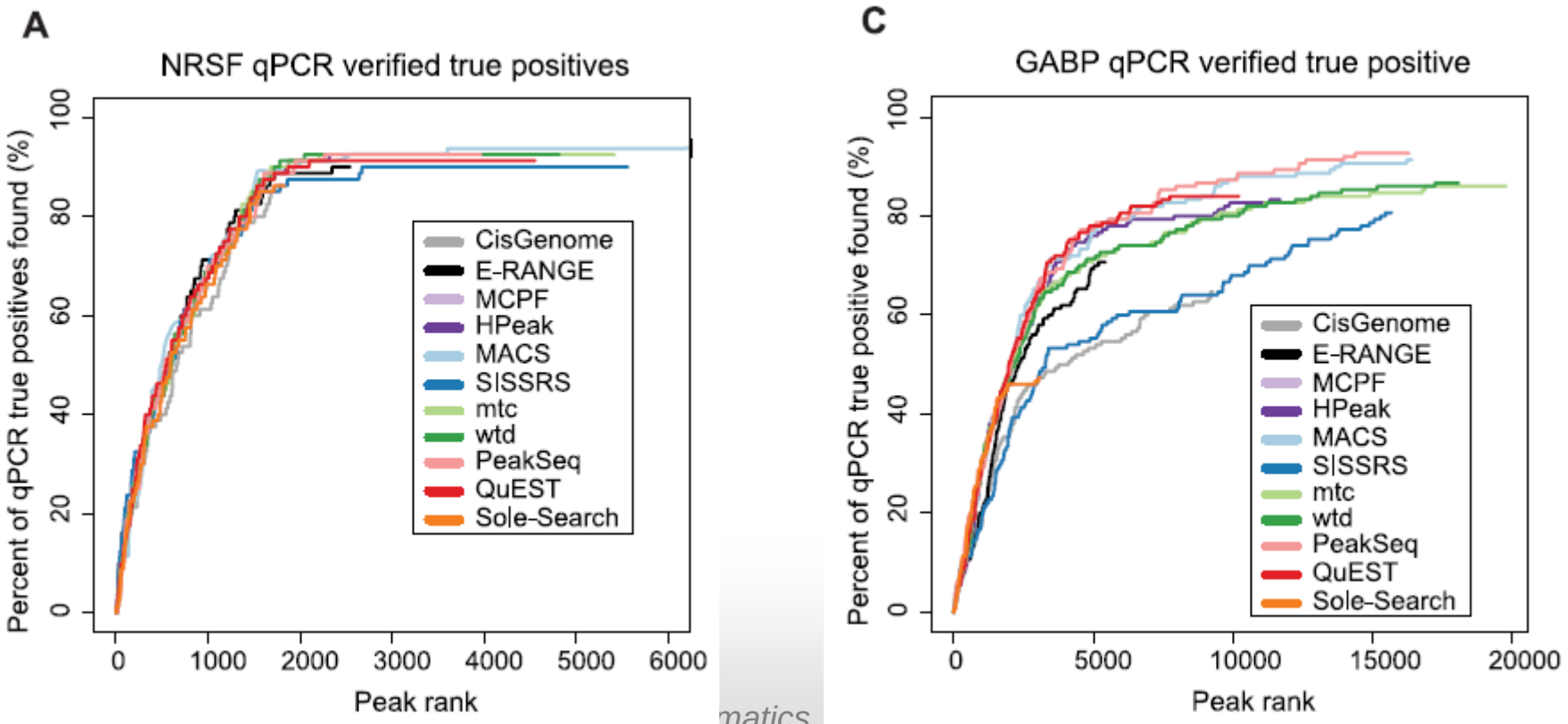
不同软件结果获得的峰不同

- 人类三个转录因子
- NRSF
- (neuron-restrictive silencer factor)
- GABP (growth-associated binding protein)
- FoxA1 (hepatocyte nuclear factor 3α)

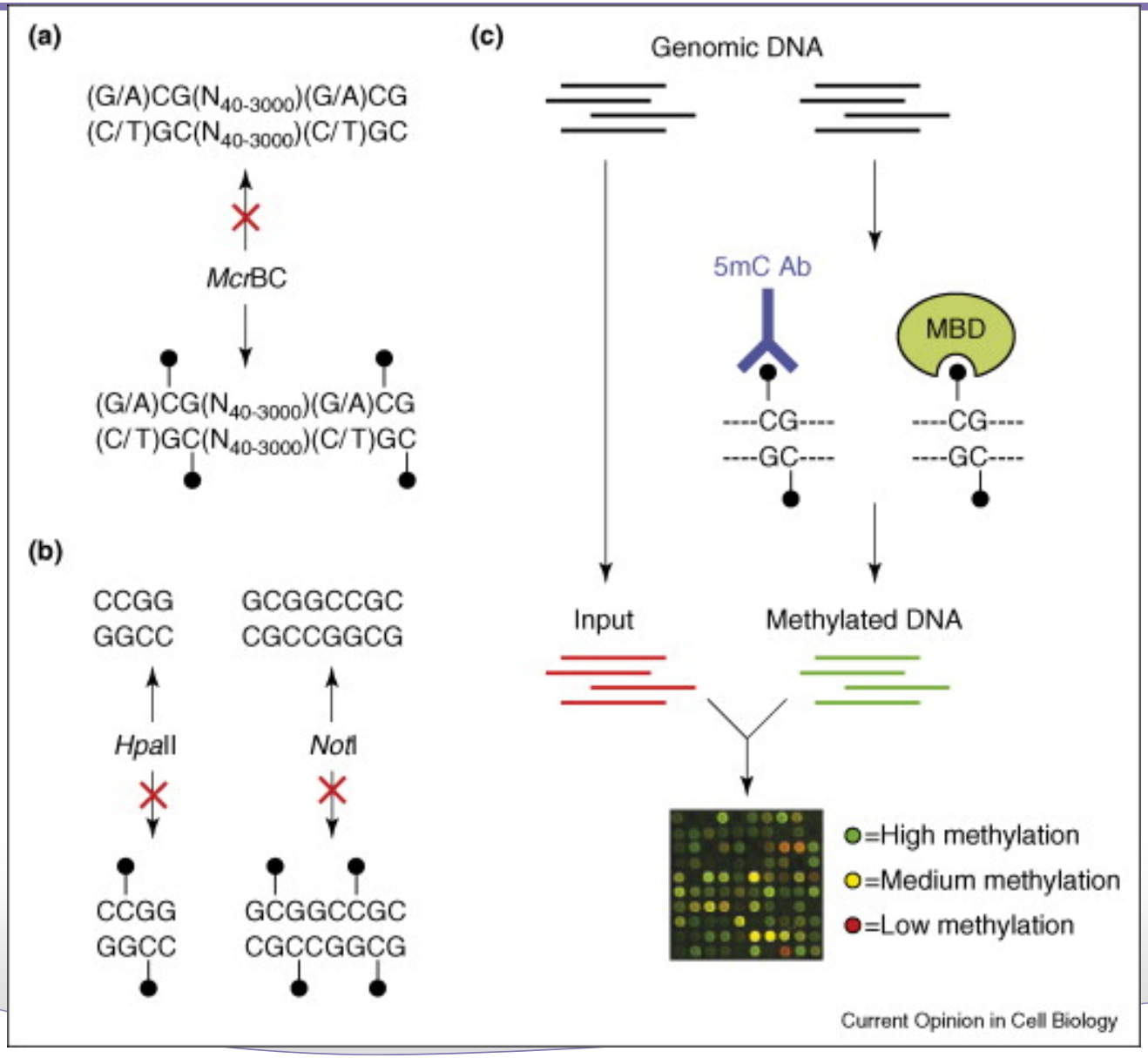


qPCR验证

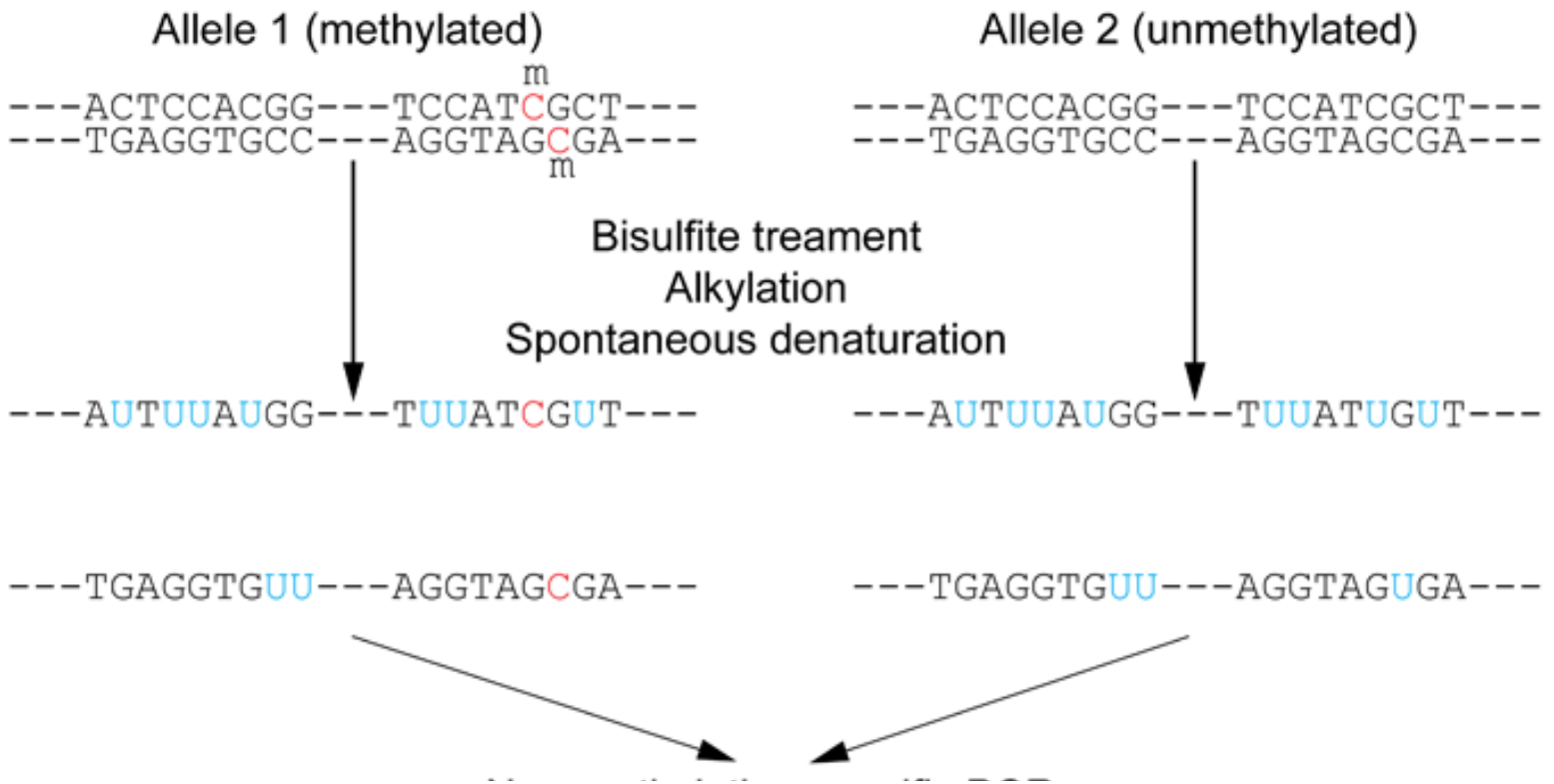
- NRSF: 大致相当
- GABP：SISSRS和CisGenome较差



DNA甲基化的检测：芯片



MethylC-seq: Bisulfite Modification



重亚硫酸盐测序: 非甲基化的C将被测序成T，反链为A

Differentiation of bisulfite-generated polymorphisms

MethyI-Seq技术

- **Bisulfite-based**
  - **MethylC-seq**
  - **Reduced representation bisulfite sequencing (RRBS)**
- **Enrichment-based**
  - **Methylated DNA immunoprecipitation sequencing (MeDIP-seq)**
  - **Methylated DNA binding domain sequencing (MBD-seq)**
- **MethylC-seq全基因组覆盖度高，测CpG岛其他方法有优势**