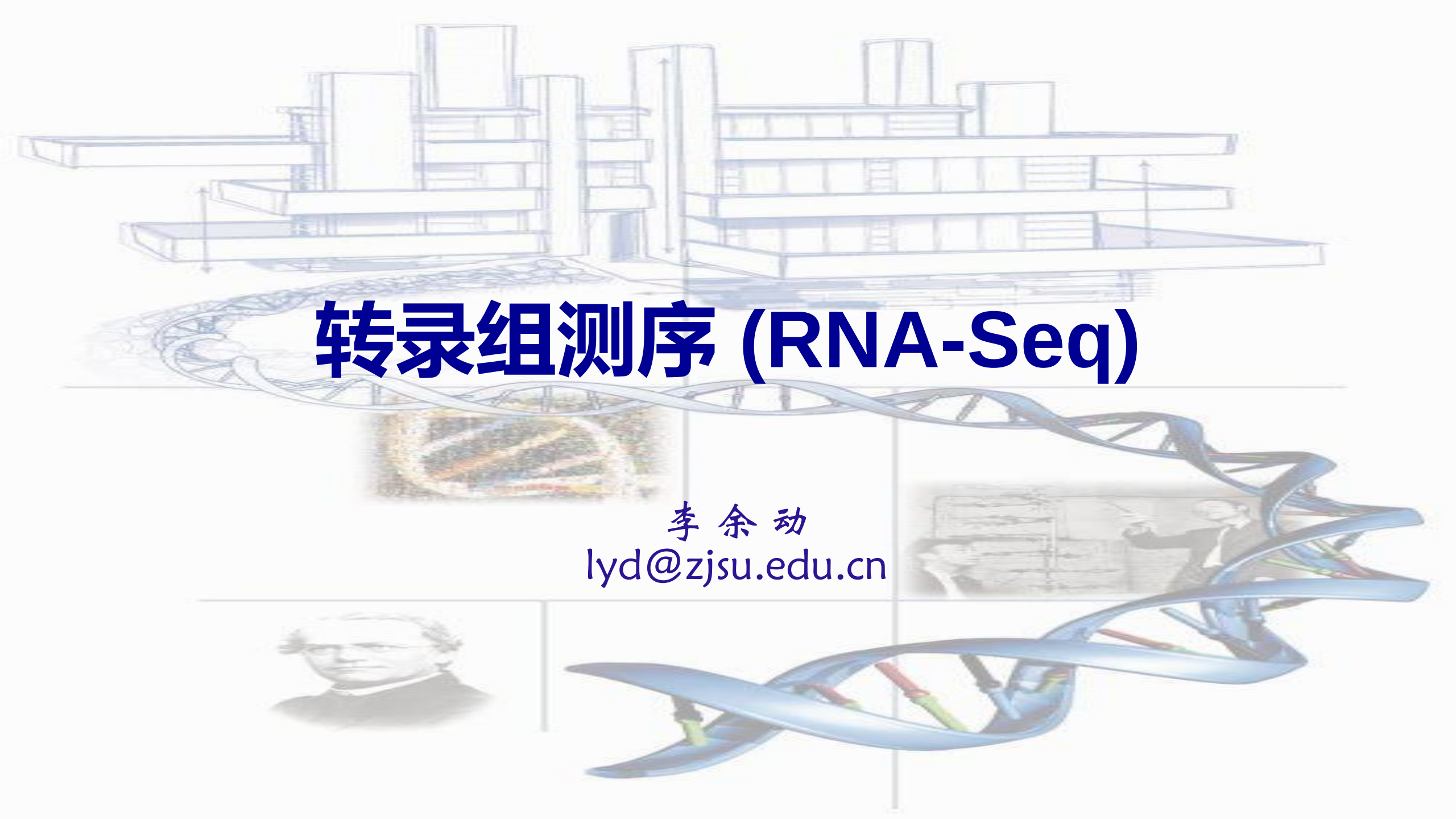
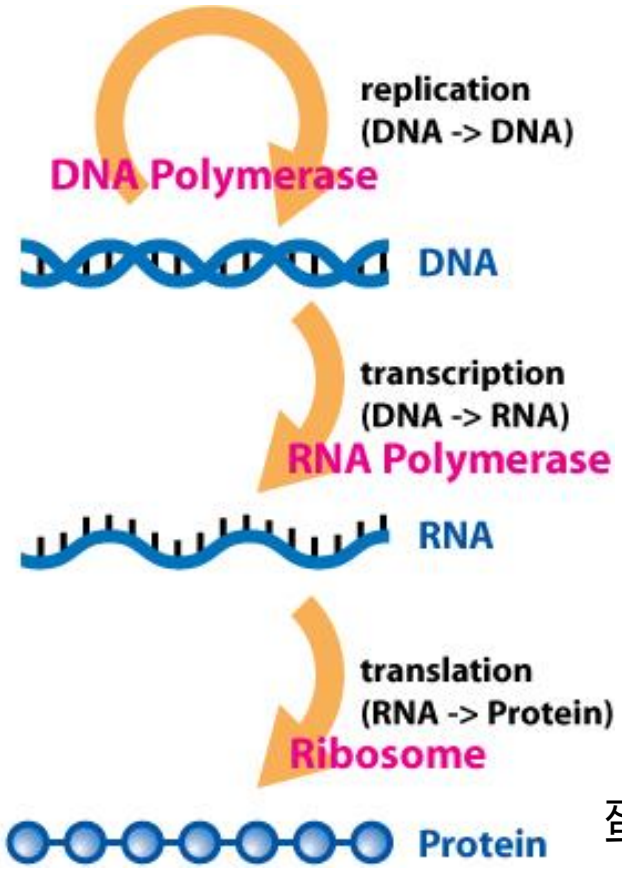


转录组测序 (RNA-Seq)

李 余 动
lyd@zjsu.edu.cn

Topics

简介



基因组

转录组

蛋白质组

中心法则：遗传信息传递

转录组transcriptome

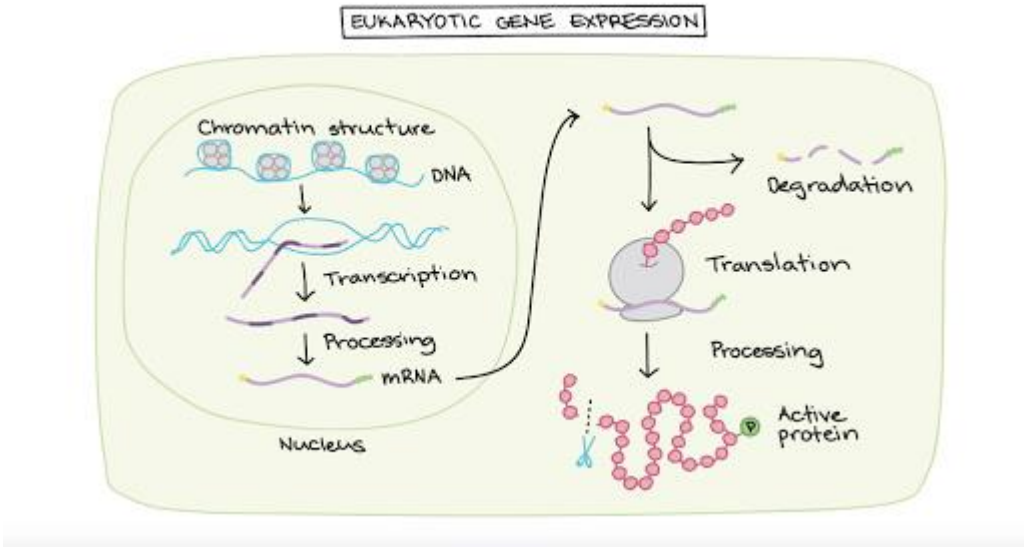
- 某一时期生物体内单个细胞或者特定组织细胞中所有转录的RNA的总和，包括信使RNA(mRNA)、核糖体RNA(rRNA)、转运RNA(tRNA)及其它非编码RNA(ncRNA)等。
- 狭义转录组指转录出的所有mRNA。

“A transcriptome is a collection of all the transcripts (转录本) present in a given cell.” (NHGRI factsheet, NIH, US)

转录组学

- 转录组学研究在单个细胞，或特定类型细胞、组织、器官或发育阶段的细胞群内所产生的各类RNA分子的类型和数量。
- 转录组学也被称为“基因表达谱(gene expression profiles)”，检测在一个特定条件下的RNA表达水平。

- **Gene expression** is the process by which information from a gene is used in the synthesis of a functional **gene product (RNA或蛋白质).**
- Regulation of a gene expression refers to the control of the **amount and timing** of appearance of the functional product of a gene.
- Gene expression is the most fundamental level at which the genotype gives rise to the phenotype, i.e., observable trait.



What is RNA-seq?

- RNA-seq利用高通量测序技术来检测细胞或组织在特定状态中所有的转录产物。

RNA-Seq: a revolutionary tool for transcriptomics

Zhong Wang, Mark Gerstein & Michael Snyder

Nature Reviews Genetics **10**, 57–63 (2009) | Cite this article

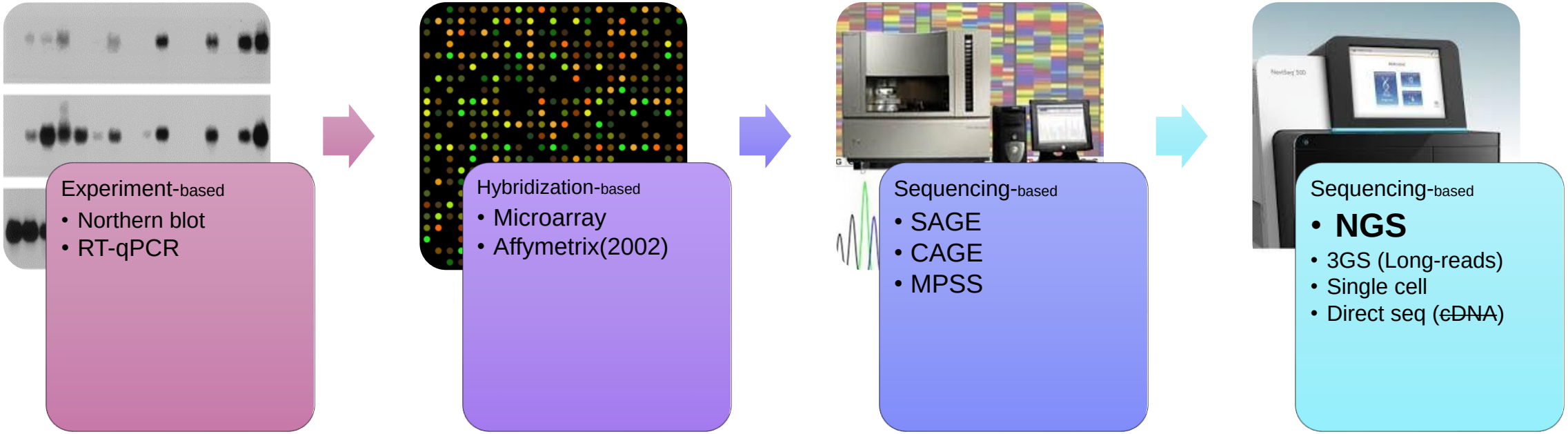
Abstract

RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. Studies using this method have already altered our view of the extent and complexity of eukaryotic transcriptomes. RNA-Seq also provides a far more precise measurement of levels of transcripts and their isoforms than other methods. This article describes the RNA-Seq approach, the challenges associated with its application, and the advances made so far in characterizing several eukaryote transcriptomes.

https://www.nature.com/articles/nrg2484

- Also called “Whole Transcriptome Shotgun Sequencing (WTSS)
- Refers to the use of high-throughput sequencing technologies
- Sequencing **cDNA** to get information about a sample’s RNA content
- A powerful tool to detect the whole transcriptome in cell and tissue.

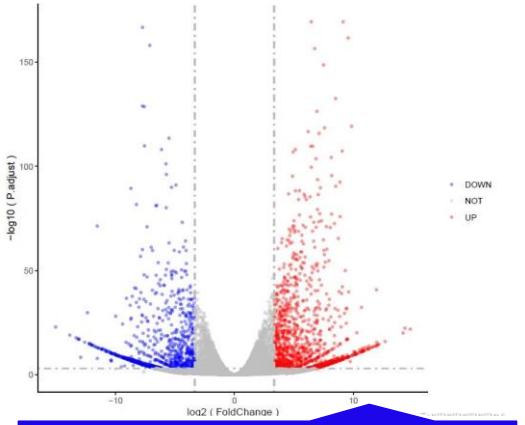
转录研究技术发展



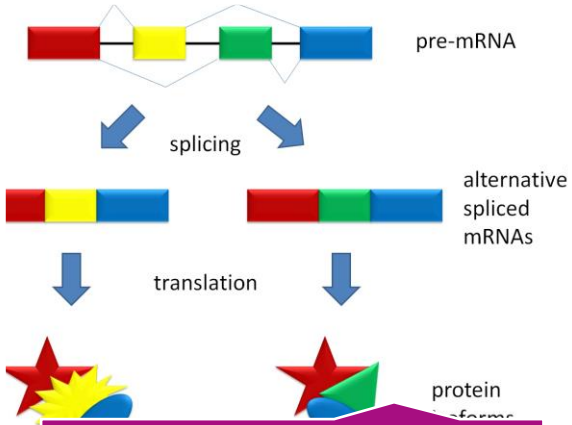
- RNA-Seq可以发现新的转录本和可变剪切异构体，而DNA芯片只能检测已知转录本的表达
- RNA-Seq比芯片的解析精度更高：单碱基水平

RNA-Seq测序应用

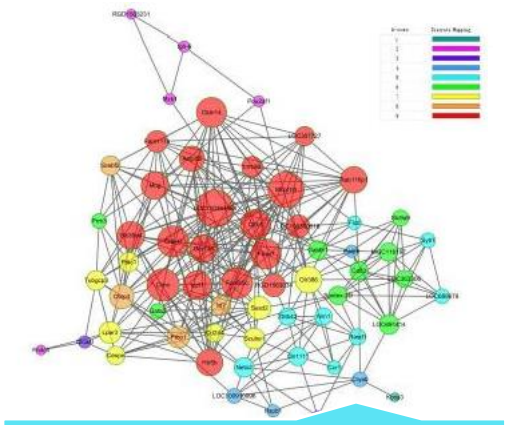
- RNA-Seq技术是发现未知的新转录本、选择性剪接和基因融合等转录现象的有力工具
- RNA-Seq可以解析疾病条件下转录组的差异，识别转录组中新的转录本，为理解人类疾病的基因调控机制提供新的视角



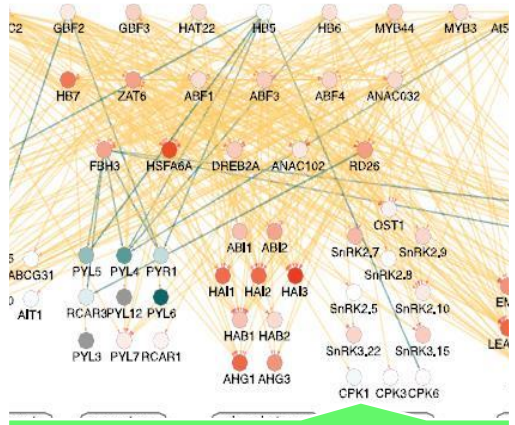
差异表达



可变剪切



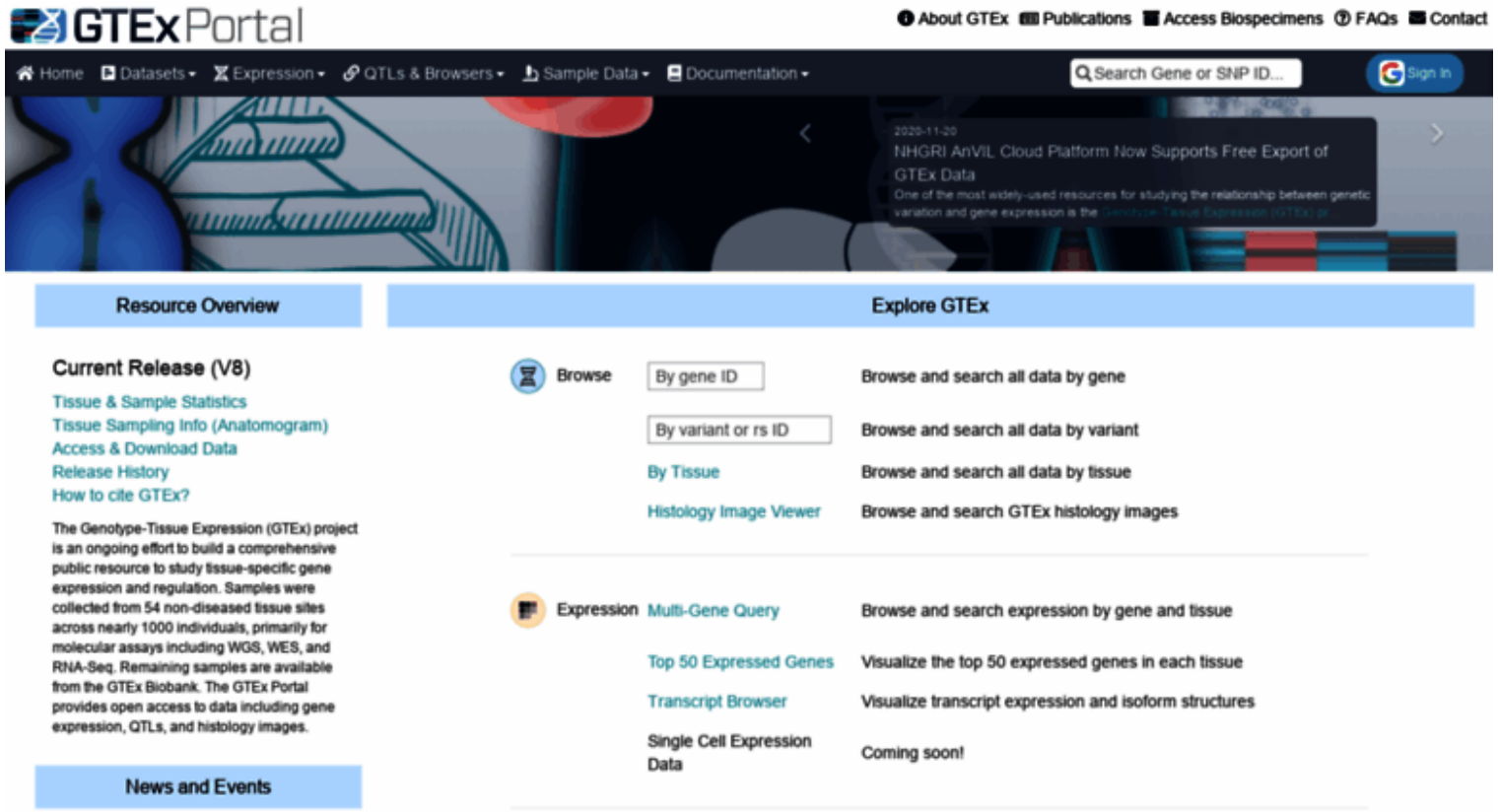
共表达网络



转录调控

GTEx：Genotype-Tissue Expression

- GTEx (Genotype-Tissue Expression)，基因型-组织表达研究联盟，研究人体不同组织样本的基因表达与个体差异的关联。
  - https://gtexportal.org



RNA-Seq流程

1.试验设计

2.测序流程

3.数据分析

4.验证实验

1.试验设计

实例：双因素实验设计

	正常细胞	肿瘤细胞
药物处理	8	2
安慰剂处理	34	18

任一基因的变化都不能归功于单一因素，
而是它们之间的多重相互作用！！

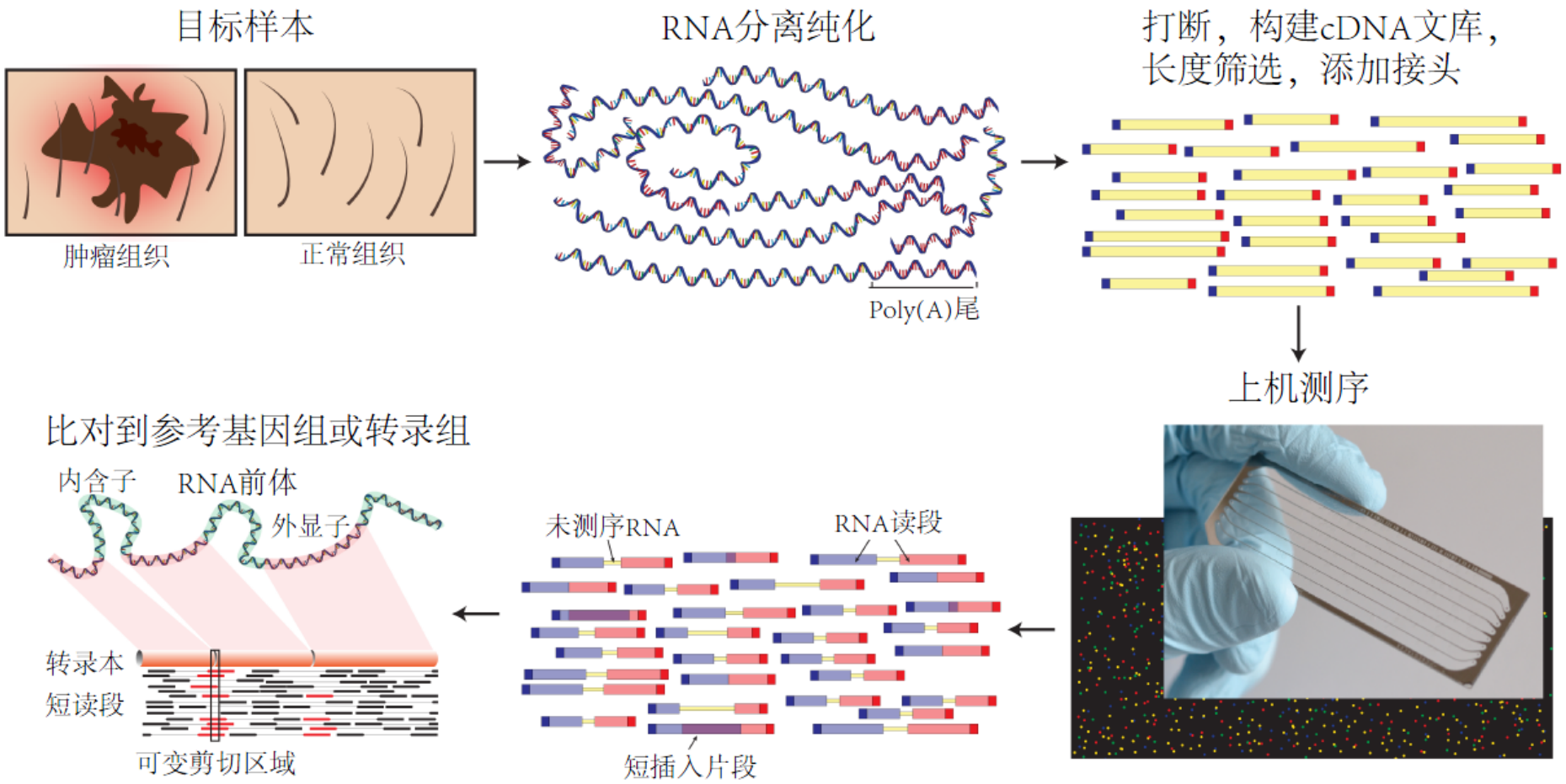
随机化Randomization

指定实验对象到每一个组，从而避免在实验收集过程导入非期望的偏差

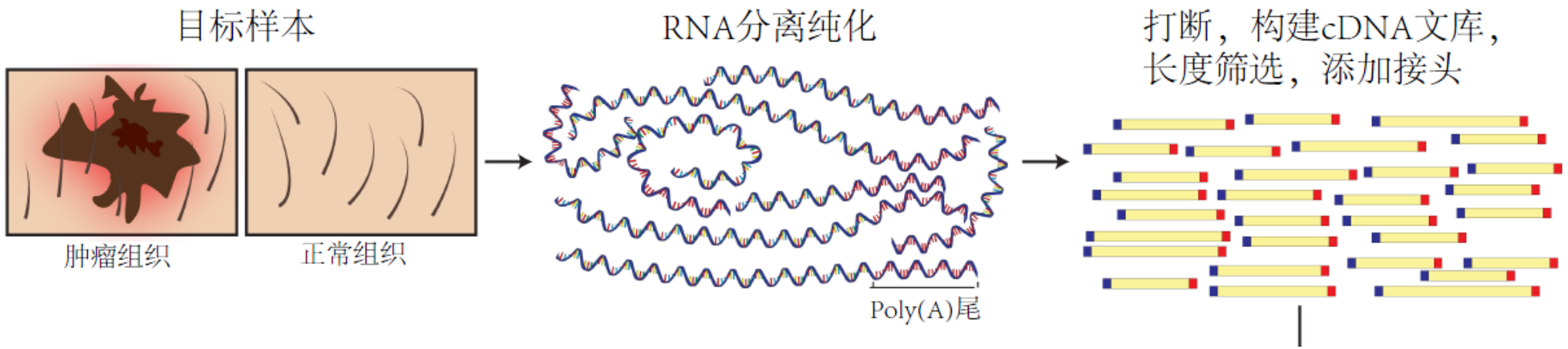
重复Repetition

为了把观察样品的分析结果外推到相应的总体，必须有重复实验来估计组内的基因表达变异。
一般要求至少有三次重复实验。

2.测序流程


目标样本
肿瘤组织
正常组织
RNA分离纯化
Poly(A)尾
打断，构建cDNA文库，
长度筛选，添加接头
上机测序
未测序RNA
RNA读段
短插入片段
比对到参考基因组或转录组
内含子
RNA前体
外显子
转录本
短读段
可变剪切区域

2.1样品准备



样品收集

尽量减少无关因素的影响；如果不可避免，也要在各组内做到均匀分布。

总RNA提取

提取1~2 μg RNA可满足RNA-seq测序需求

RNA分离纯化

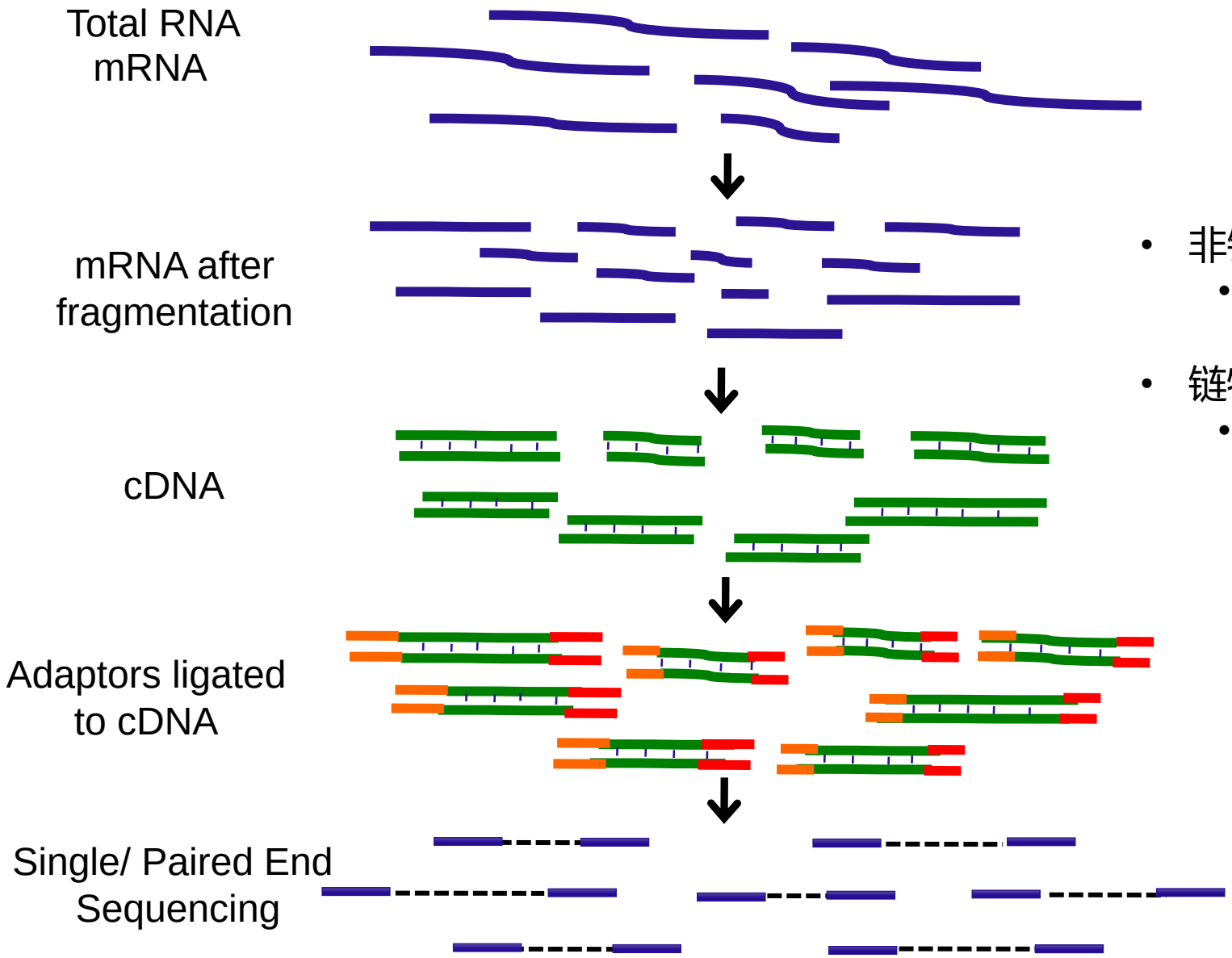
方法1：由于真核生物的mRNA一般有poly(A)尾，通过poly(T)探针富集；

方法2：采用特殊的RNA探针结合rRNAs后，通过核酸酶降解rRNA。

cDNA文库构建

采用poly(T)寡核苷酸结合真核mRNA的poly(A)尾，易造成3’端偏差，也排除了无poly(A)尾结构的mRNA与ncRNA。可以用随机引物进行反转录来研究后者。

2.2测序文库构建流程

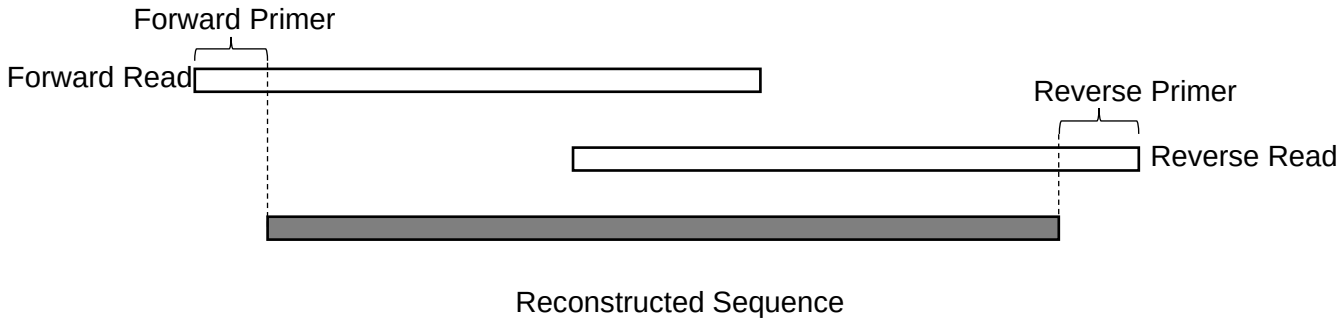


- 非链特异性文库
  - 无法区分测序片段转录自DNA正链还是负链
- 链特异性文库（strand-specific）
  - 建库时通过化学标记保留转录本的方向信息

2.3测序策略

PE vs. SE

采用双端(Pair-End, PE)能利用重叠区段获得比单端(Single-End, SE)测序更可靠的比对结果。



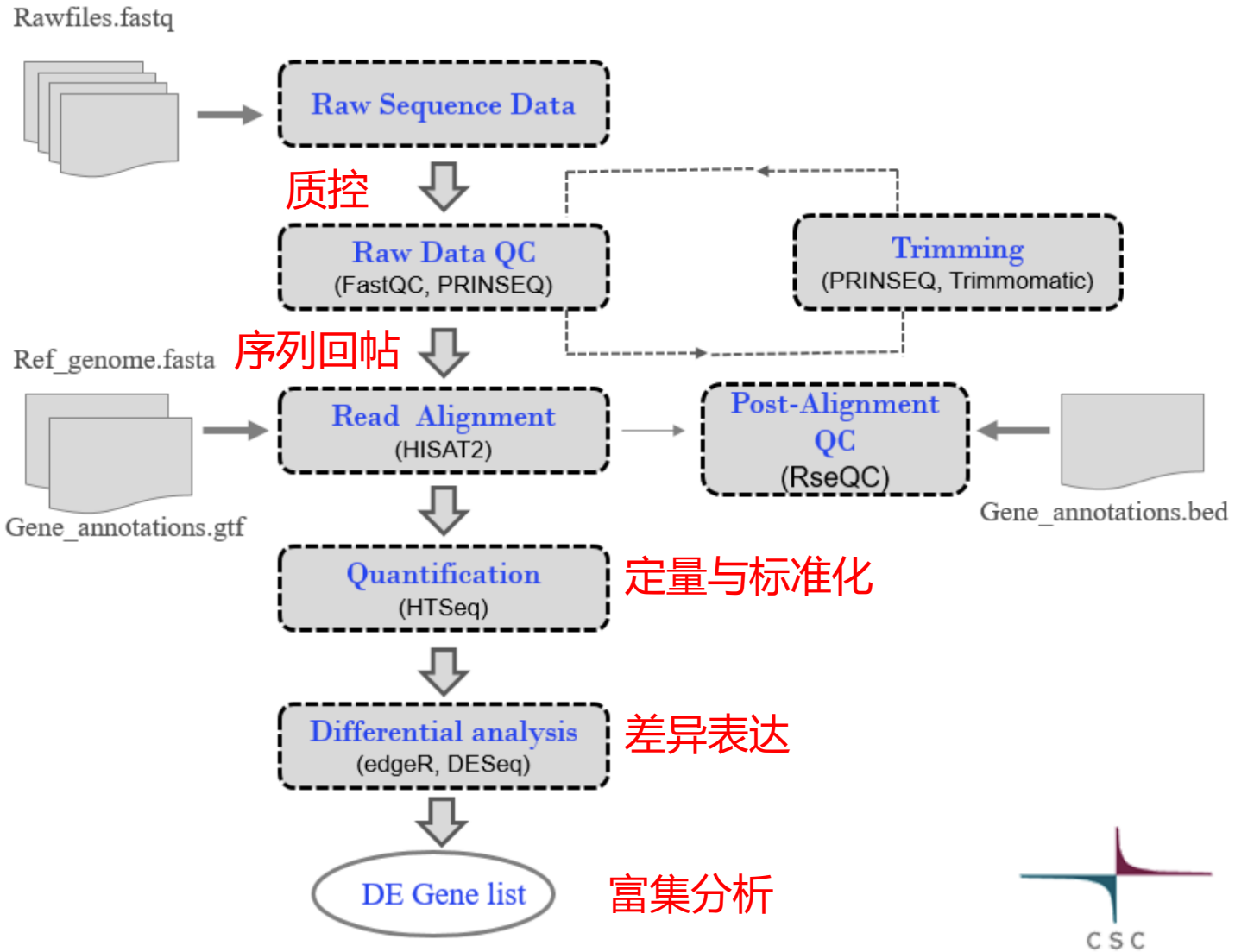
测序深度

- 测序深度要根据实验的具体要求而定，例如，与相对差异表达分析，鉴定新转录本需要更多的测序深度。
- 更大的基因组需要更高的测序深度：
  - 对酵母而言，30M条35bp的reads能够观察到大多数基因(>90%)；
  - 对人类而言，100M条reads才能检测到～80%表达基因，300M条reads才能检测到～80%的差异表达基因。

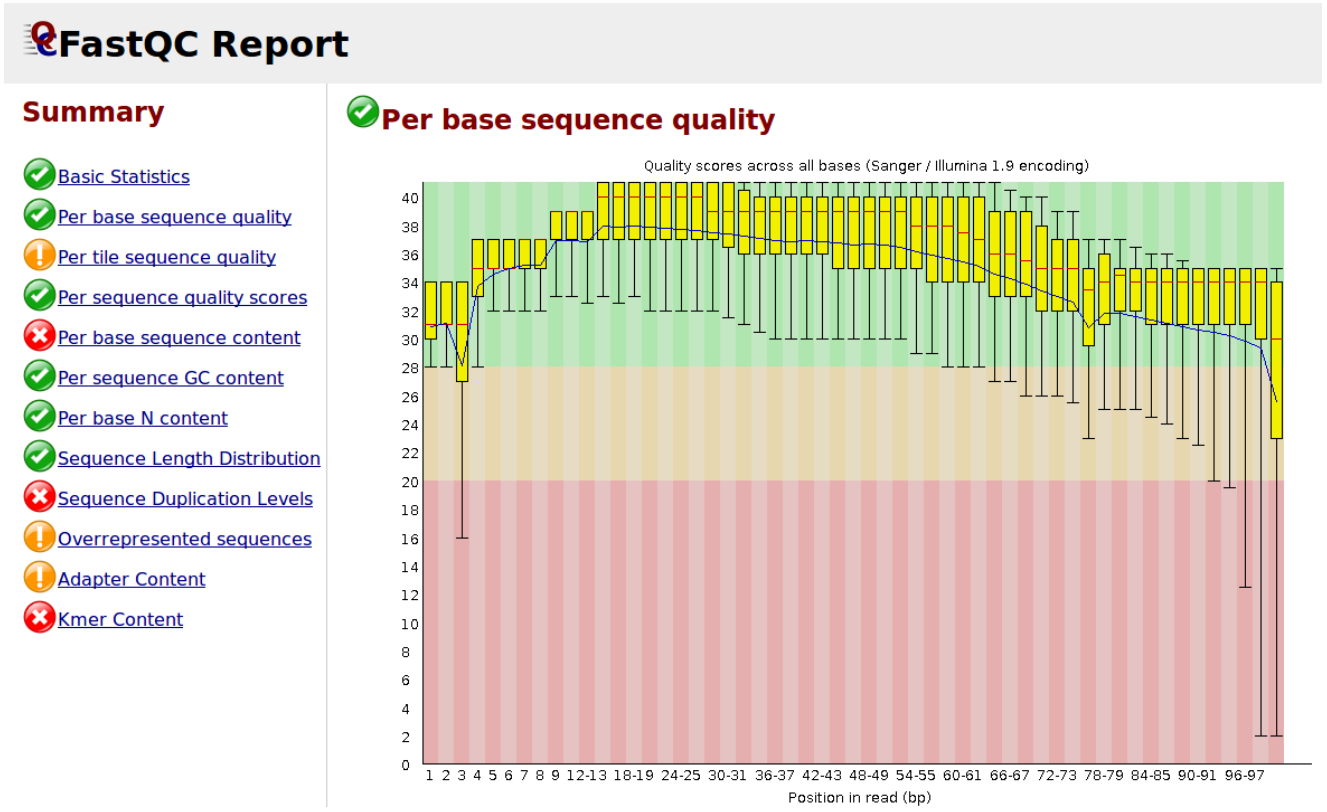
增加样本重复数比增加测序深度更有助于提高检测效果。

3.数据分析

- RNA-seq数据分析一般流程



3.1数据的质量控制



测序数据的质量控制

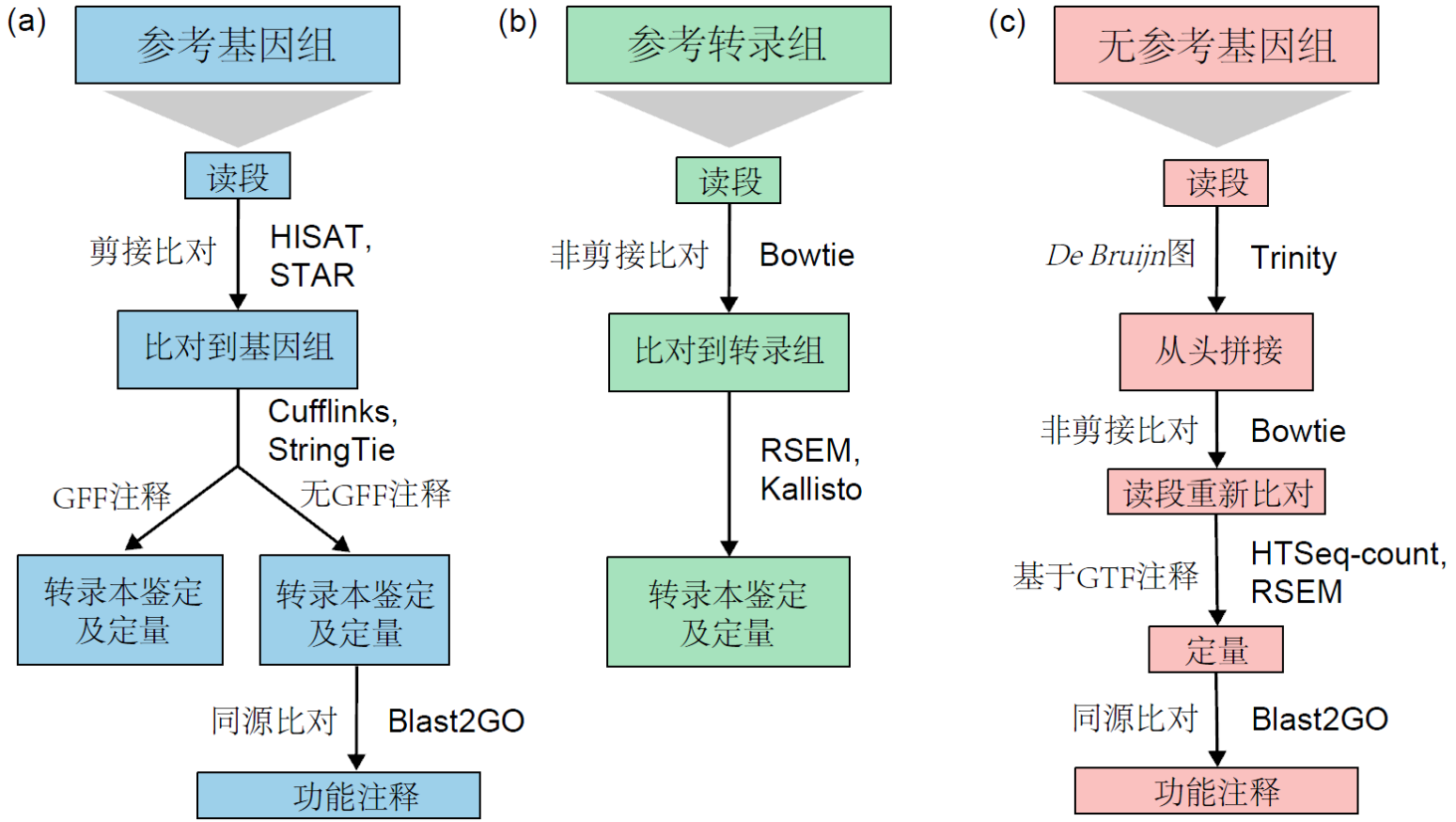
- FastQC—测序质量评估
- FASTP，Trimmomatic—质量控制
- RNA-SeQC

数据分析过程中的质量控制

- Aligned rate
- rRNA rate
- Duplicated rate
- Coverage
- …

3.2读段回帖(mapping)

RNA-seq数据分析的三种策略



Conesa, A. (2016) *Genome Biology*

3.2读段回帖(mapping)

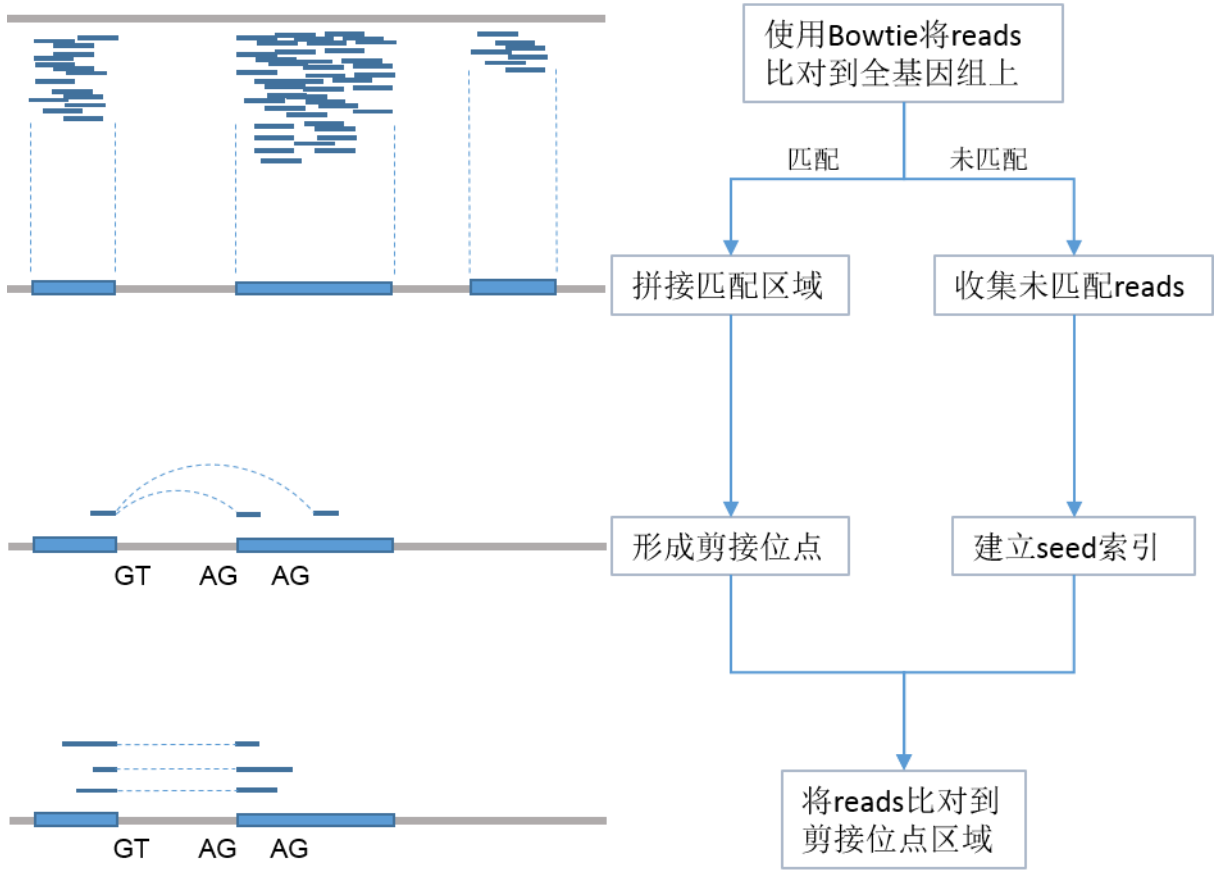
比对策略（有参考基因组）

方法一：剪切比对(spliced aligning)

- 先用mapped reads建立潜在的剪切体结构(Exon-first)，后在unmapped reads中预测splice junction(Tophat/HISAT)；
- Seed-and-extend

  利用reads的部分序列(k-mers)开始比对过程，随后对可能的比对位点(hits)的延伸来定位剪切位点。

剪接比对—**TopHat**，**STAR**，**HISAT2**，**MapSplice** (SNP)



TopHat工作原理 Trapnell, C. (2009) *Bioinformatics*

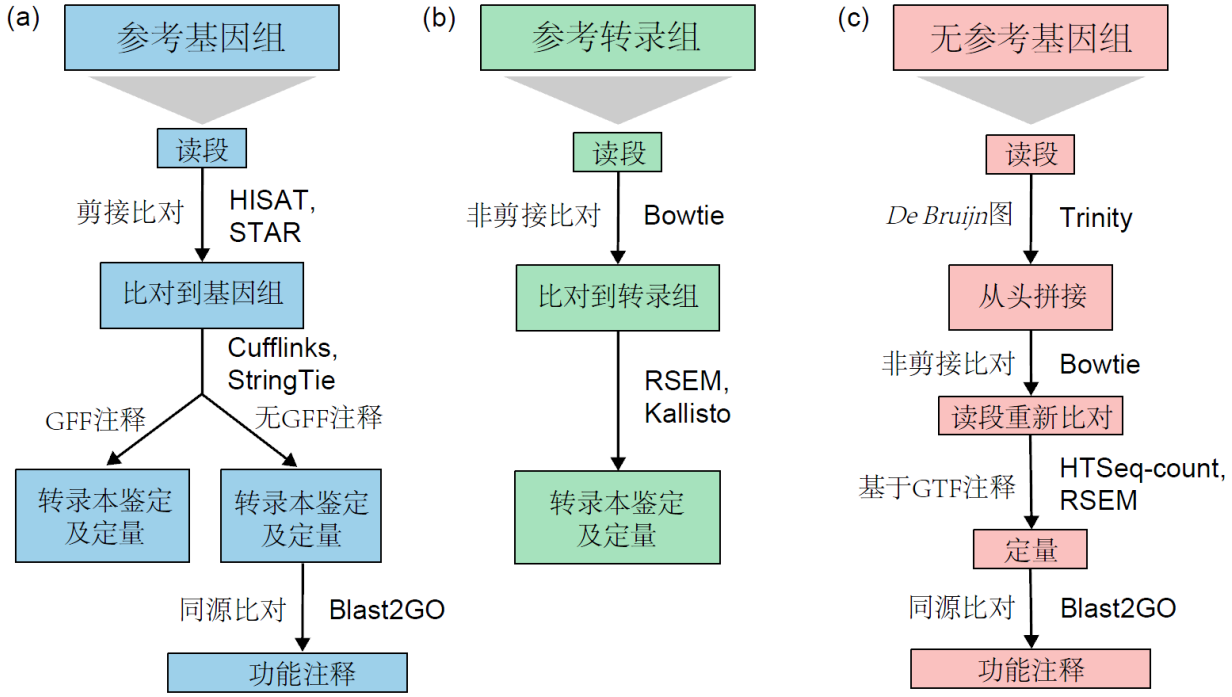
3.2读段回帖(mapping)

比对策略（有参考转录组）

方法二：非剪切比对(unspliced aligning)

- 把所有exon注释放到一个所有转录本异构体数据库，再将reads比对这个数据库。
- 适用人、小鼠等注释信息相对完善的模式生物

非剪接比对—**Bowtie，BWA**（不考虑可变剪切）



Conesa, A. (2016) *Genome Biology*

3.2读段回帖(mapping)

比对策略（无参考基因组）

方法一：

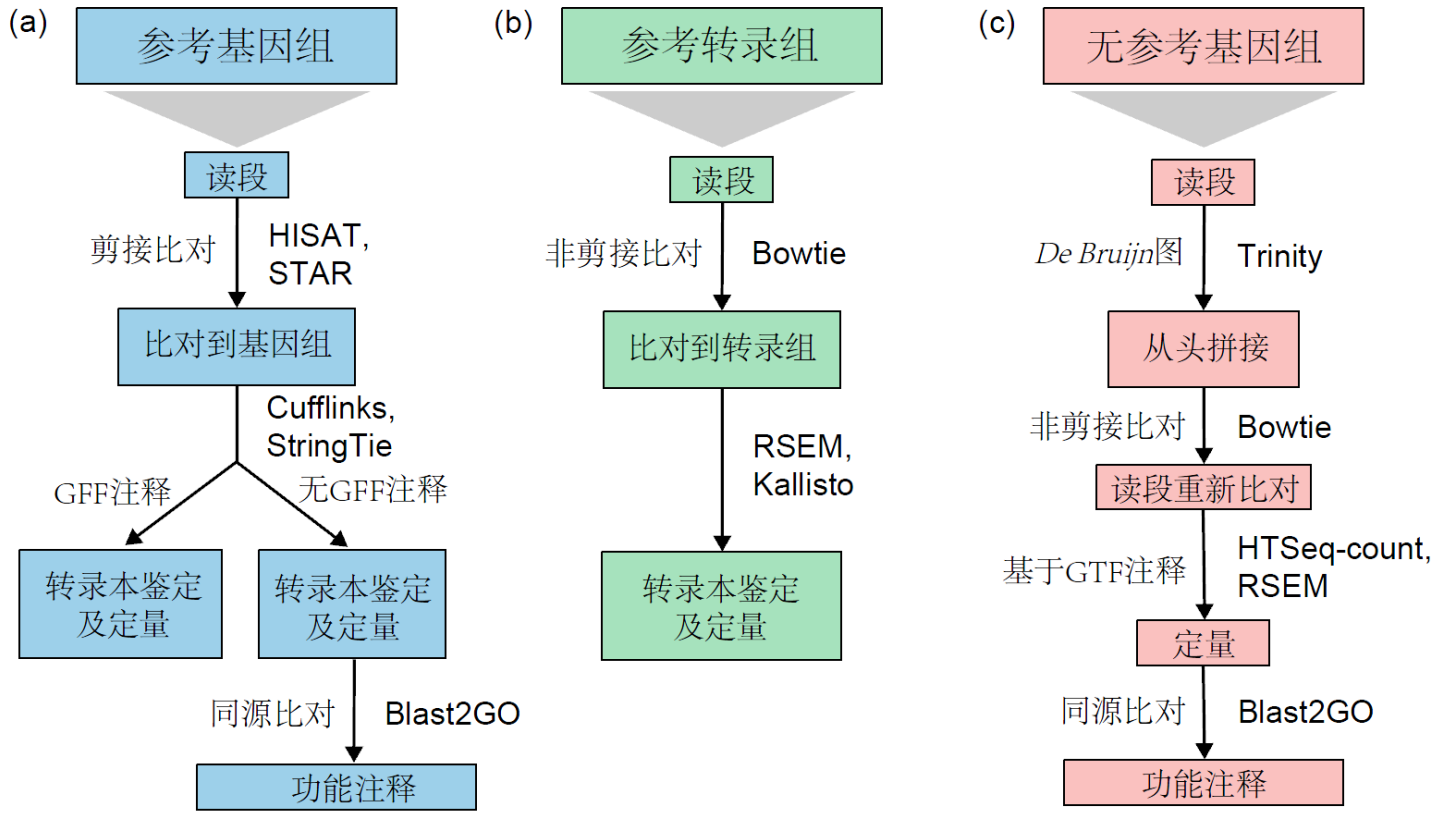
利用亲缘关机相近的物种的基因组。

方法二：

从头(*de novo*)组装（拼接）转录组

Trinity

RNA-seq数据分析的三种策略



Conesa, A. (2016) *Genome Biology*

3.3 转录本定量

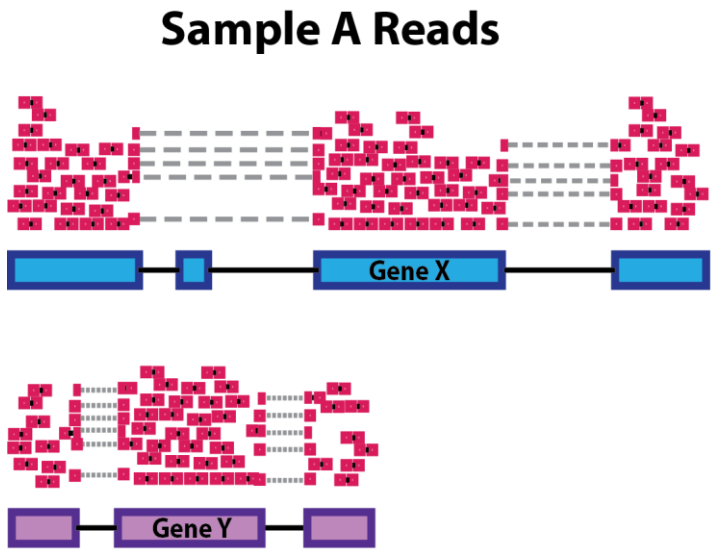
读段数(reads counts)

- 只保留唯一匹配reads

  HTSeq-count，featureCounts

- 保留多重匹配reads，利用统计模型将多重比对的reads定位到对应的转录本异构体上

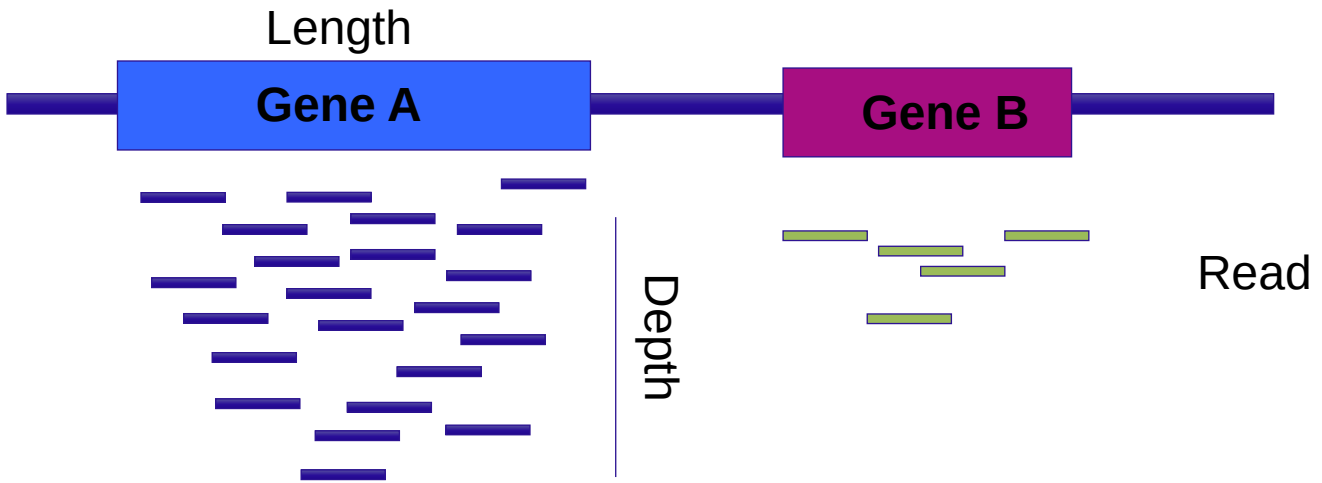
  Cufflinks, StringTie, RSEM



3.3 转录本定量

定量标准化 (Normalization)

- 校正测序**深度**、基因**长度**



- **RPKM**: Reads Per Kilobase of exon model per Million mapped reads (每百万reads回帖到每千个碱基外显子的reads数目)

RPKM计算公式如下：

$$RPKM = \frac{total\ exon\ reads}{mapped\ reads(millions) * exon\ length(KB)}$$

其中，total exon reads 是定位到某个基因外显子(exon)的读段数，mapped reads指回帖到基因组的总读段数，两者单位都是百万(millions)。

RNA-Seq的基因表达定量计算

- RPKM标准化方法：
  - Step 1:对每个样本的测序深度进行标准化。例如GeneA在样本rep1的RPM: 10/3.5=2.86
  - Step 2:对每个基因的长度进行标准化。例如GeneA在样本rep1的RPKM: 2.86/2=1.43

RPKM – step 1: normalize for read depth

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
Gene A (2kb)	10	12	30
Gene B (4kb)	20	25	60
Gene C (1kb)	5	8	15
Gene D (10kb)	0	0	1
Total reads	35	45	106
Tens of reads	3.5	4.5	10.6

注：由于以四个基因为例，我们用total reads除10，在RPKM(per million reads)中应除1,000,000

RPM-scaled using the "per million" factors

Gene Name	Rep1 RPM	Rep2 RPM	Rep3 RPM
Gene A (2kb)	2.86	2.67	2.83
Gene B (4kb)	5.71	5.56	5.66
Gene C (1kb)	1.43	1.78	1.43
Gene D (10kb)	0	0	0.09

RPKM – step 2: normalize for read length

Gene Name	Rep1 RPM	Rep2 RPM	Rep3 RPM
Gene A (2kb)	2.86	2.67	2.83
Gene B (4kb)	5.71	5.56	5.66
Gene C (1kb)	1.43	1.78	1.43
Gene D (10kb)	0	0	0.09

Now we need to scale per kb.

Reads are scaled for depth (M) and gene length (K).

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
Gene A (2kb)	1.43	1.33	1.42
Gene B (4kb)	1.43	1.39	1.42
Gene C (1kb)	1.43	1.78	1.43
Gene D (10kb)	0	0	0.009

3.3 转录本定量

标准化策略

- RPKM, FPKM, TPM
- 常用工具：Cufflinks, StringTie

- **FPKM**: Fragments Per Kilobase of exon model per Million mapped fragments(每百万reads回帖到每千个碱基外显子的fragments数目)
  - Fragments指双末端测序的两端配对片段，FPKM将一个fragment的两个reads计算一次，其它与RPKM相同
- **TPM**: Transcripts Per Million mapped reads (每百万reads回帖的Transcripts)
  - TPM与FPKM的处理顺序不同，先基因长度，后测序深度。

RPKM vs TPM

RPKM

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009
Total:	4.29	4.5	4.25

... the sums of each column are very different.

TPM

Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02
Total:	10	10	10

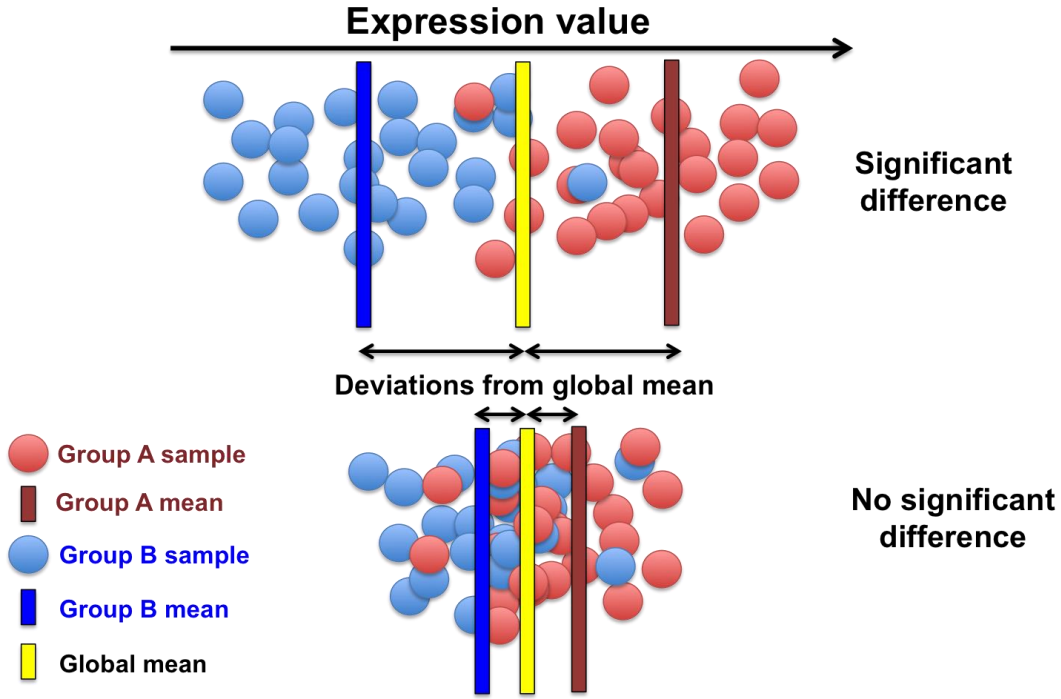
$$TPM = \frac{X_i}{L_i} * \frac{1}{\sum_j \frac{X_i}{L_j}} = \frac{\frac{X_i}{L_i}}{\sum_j \frac{X_i}{L_j}}$$

$$FPKM = \frac{X_i}{L_i} * \frac{1}{\sum_j X_j} = \frac{\frac{X_i}{L_i}}{\sum_j X_j}$$

X_i代表比对到基因上的reads数，单位为Millions;
L_i代表基因外显子的总长度，单位为Kilobase。
两个公式的分子是相同的，$\frac{X_i}{L_i}$代表外显子长度为1Kb的基因的reads数目，即转录本丰度。

3.4差异表达分析

- **差异表达分析(Differential expression analysis)** is是指基于一些统计学模型，对不同样本处理条件下的基因表达差异进行分析，区分这种差异是源于处理效应还是随机误差。
- 差异表达分析的结果一般用差异倍数(fold change)和统计检验显著性值(p-value)来描述。



The average expression of a gene in group A with the average expression of this gene in group B.

3.4鉴定差异表达基因

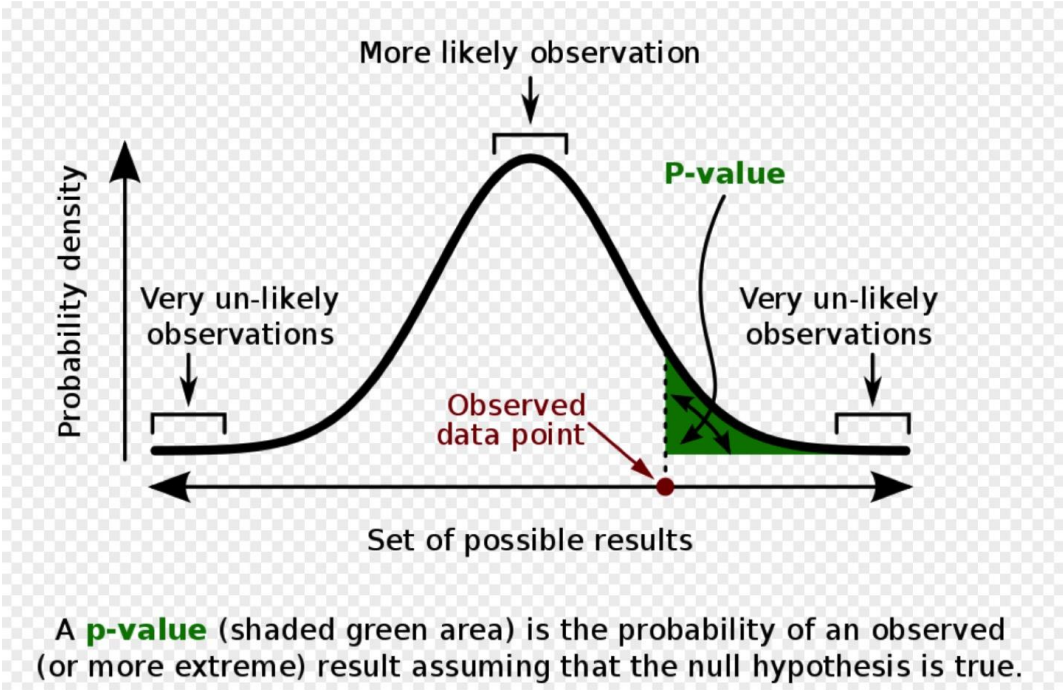
- **基因表达水平变化倍数(Fold Change): log_2(sample1/sample2)**
- Gene X in samples 表达水平是800/200(stress/control)
  - Log2(800/200) = log2(4) = 2，变化倍数为4，正调控(Up-regulated)，用**红色(red)**表示。
- Gene Y in samples 表达水平是200/800(stress/control)
  - Log2(200/800) = log2(1/4) = -2，变化倍数为-2，负调控(Down-regulated)，用**绿色(green)**表示。
- Gene Z in samples 表达水平是400/400(stress/control)
  - Log2(400/400) = log2(1) = 0，变化倍数为0，表达无变化，用**黑色(black)**表示。

ID	S1/Control	S2/Control	S3/Control
Gene1	-2.03	-1.5	0
Gene2	0.01	5.3	0.002
Gene3	4.6	999.34	1.2

3.4鉴定差异表达基因

- ***P*-value**: 在统计检验中，p值是指当原假设为真时，得到样本统计量或更极端情况的概率。
- **Adjusted p-value**: **corrected for multiple genes testing**. In the result, we can say that all genes with adjusted p-value < 0.05 are significantly differentially expressed in these two samples.

- Statistical hypothesis testing (统计假设检验)
- Each test has a known distribution： 正态分布、泊松分布……



https://biocorecrg.github.io/CRG_Bioinformatics_for_Biologists/differential_gene_expression.html

3.4鉴定差异表达基因

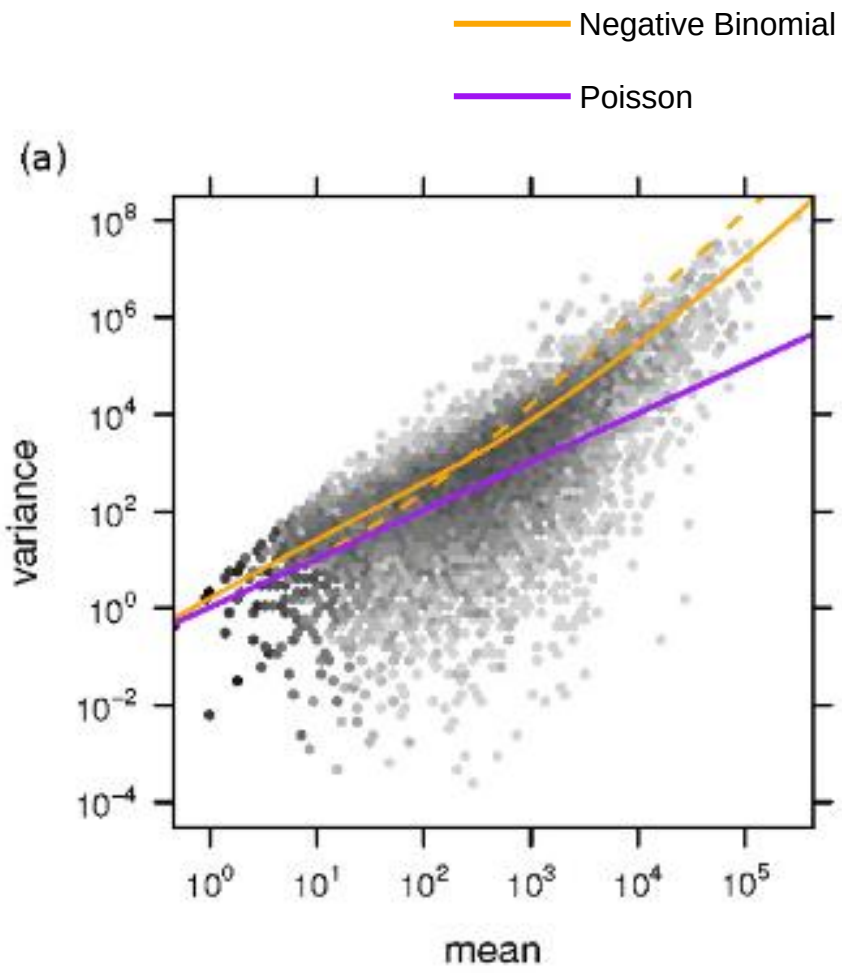
数据差异分析策略：

- 对原始计数(counts)数据应用负二项（Negative Binomial）分布统计模型

  DESeq2, edgeR

- 对标准化数据(FPKM/TPM)进行统计

  Cuffdiff2（T检验）, Ballgown



RNA-Seq数据的过度离散问题

3.4鉴定差异表达基因

RNA-Seq基因差异表达分析步骤：

① 统计基因或转录本对应的reads计数；

② 对reads计数进行标准化，使样本间和样本内的表达水平能够进行精确比较；

③ 对标准化后reads分布进行统计学模型拟合，利用统计学检验评估基因的差异表达，得到对应的*P*值(*p*-value)和差异倍数(fold change)，并完成多重检验校正；

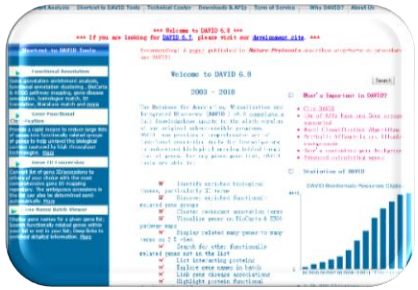
④ 根据差异倍数和校正后的*P*值的特定阈值确定显著差异表达基因。

Differential Expressed Genes (DEGs):

- Log2(Fold Change) > 1
- Adjusted *P*-value < 0.05

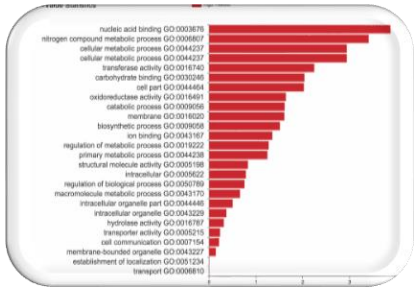
3.5差异表达基因的功能分析

- 常用的基因功能注释信息数据库：
  - Gene Ontology (GO)：基因本体数据库，旨在建立基因及其产物知识的标准体系，涵盖了细胞组分、分子功能和生物学过程三个方面。
  - Kyoto Encyclopedia of Genes and Genomes(KEGG)：KEGG代谢通路数据库，包含新陈代谢、遗传信息加工、环境信息加工、细胞过程、生物体系统、人类疾病和药物开发等多种分子相互作用和反应网络。



DAVID

- 物种覆盖较全；数据更新慢

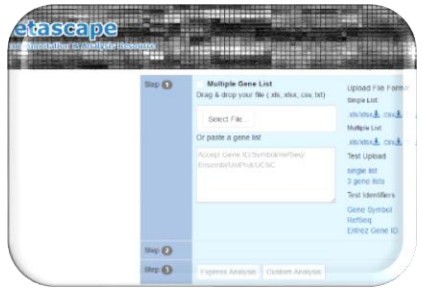


WEGO

- 富集结果可视化

clusterProfiler

- 实时抓取；富集方法全面；R语言



MetaScape

- 操作简单；可视化效果好；物种较少

3.5差异表达基因的功能分析

功能富集(Enrichment)分析：

- 利用已知的基因功能注释信息作为先验知识，对目标基因集进行功能富集。

$$p = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{N-M}{k-i}}{\binom{N}{k}}$$

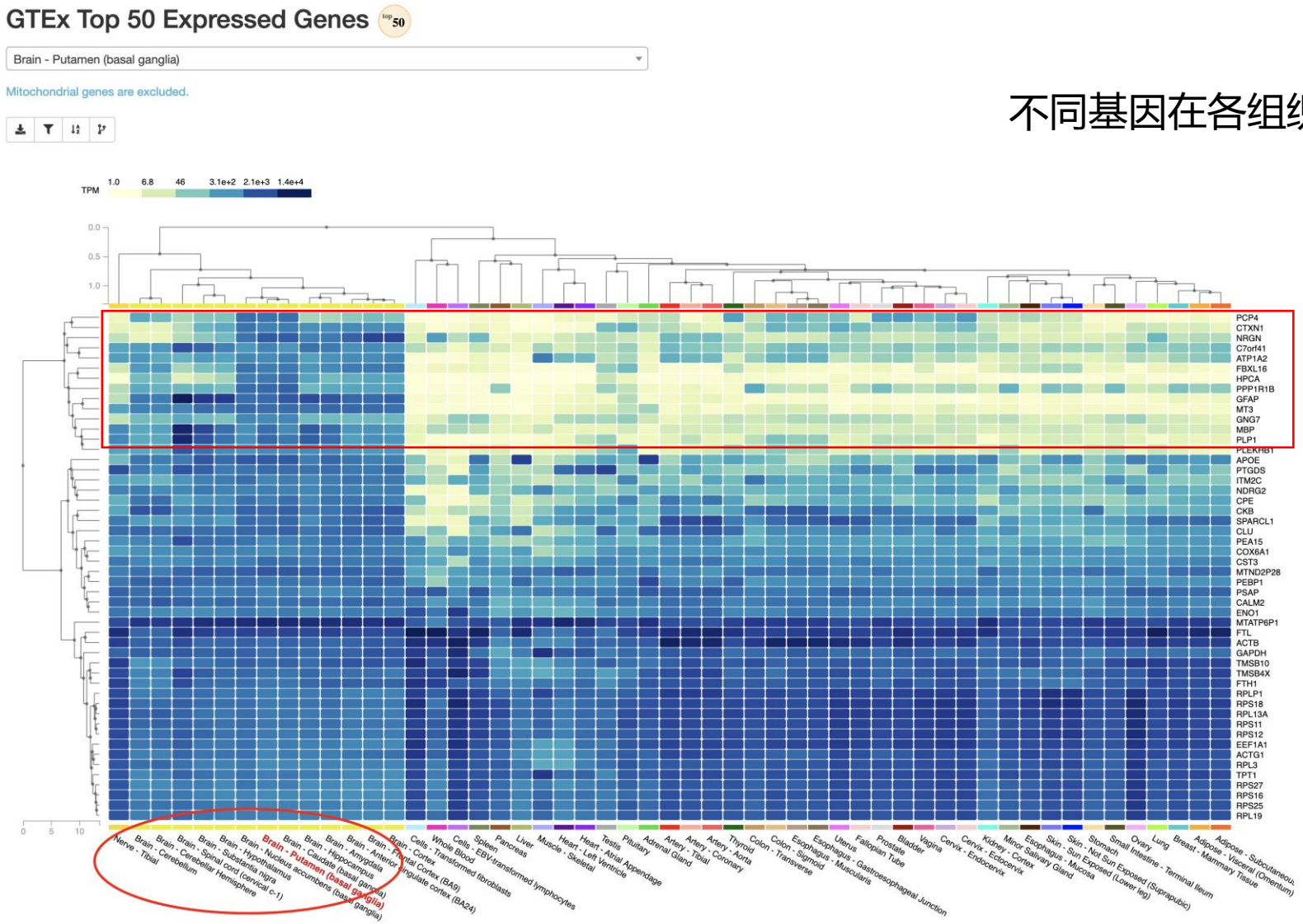
超几何分布检验公式

Fisher’s精确检验

	特定功能集	其他功能集	总基因数
目标基因	x	k-x	k
背景基因	M	N-M	N

含义：从N个基因中随机抽取k个基因，其中有x个属于特定功能基因集M的概率，如检验P<0.05说明为小概率事件，即筛选得到的差异表达基因显著富集于该特定功能集。

转录调控影响基因表达模式

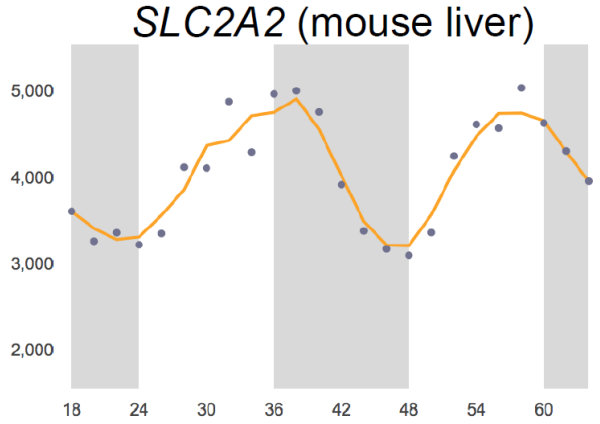


不同基因在各组织中受到差异化转录调控

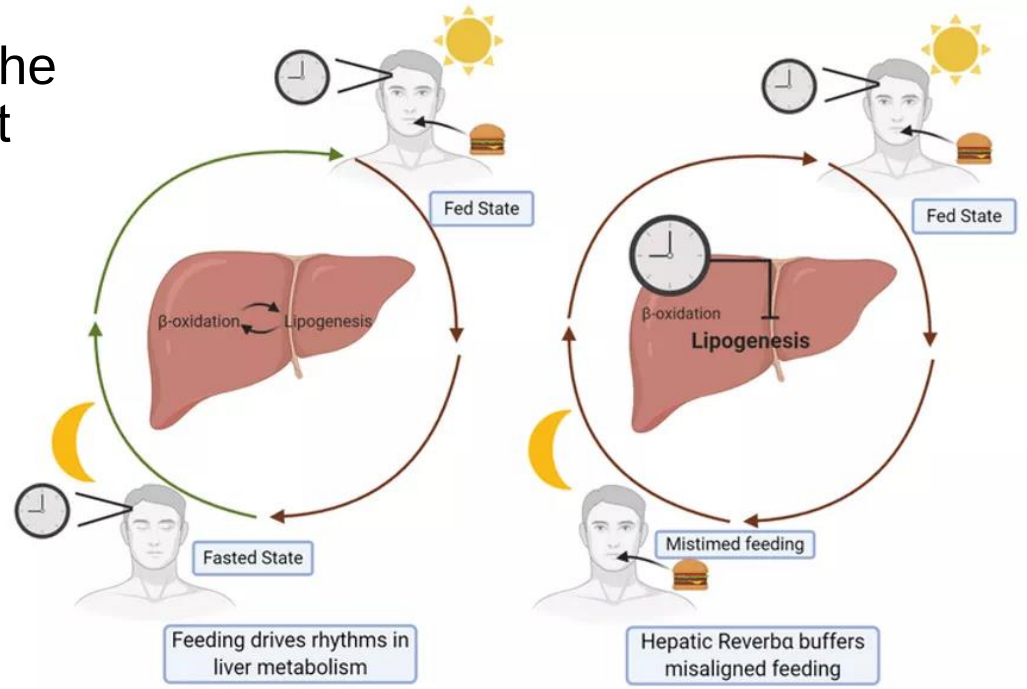
- 管家基因(House-keeping genes, HKG)
- 组织特异基因 (Tissue-specific genes, TSG)

Circadian gene（节律基因）

- Circadian genes are defined as genes whose protein products are necessary components for the generation and regulation of circadian rhythms.



- REVERBα gene is linked to obesity in humans.
- disordered feeding caused a major change in the expression of REVERBα genes that control fat metabolism



RNA-Seq数据分析实战

数据介绍

- 数据来源：为酿酒酵母在不同氮源培养基中培养后进行mRNA测序的数据。
- 对照组：在尿素为唯一氮源的培养条件下的表达数据（尿素为对照氮源，不会引起NCR反应）；
- 实验组：在精氨酸为唯一氮源的培养条件下的表达数据（精氨酸是酵母菌的偏好性氮源，会引起NCR反应）；
- 测序平台：Illumina HiSeq2500，双端测序(2×100 nt)；
- 测序方案：非链特异性(unstranded)，即不确定reads要比对到双链DNA的哪一条链。

RNA-Seq分析的数据集

数据	文件	说明信息
参考基因组序列及基因信息[a]	genome.fa/genes.gtf	酿酒酵母的基因组序列及基因注释文件(GTF)
测序数据[b]	scer_arg_R1.fq/scer_arg_R2.fq	精氨酸为氮源的样品
	scer_urea_R1.fq/scer_urea_R2.fq	尿素为氮源的样品
接头序列[c]	adapter.fa	Illumina测序接头序列

[a]参考基因组序列及GTF文件可从ENSEMBL数据库获得：http://ensembl.org/info/data/ftp/index.html
[b]一个好的RNA-Seq实验设计至少需要3个生物学重复。此练习每个样品只有一个数据，以加快程序运行速度。
[c]更多关于接头序列信息，可查阅https://support.illumina.com/sequencing/documentation.html

1.准备测序数据

将本练习所需要数据文件都复制到一个目录ch16RNAseq

```
$ cd ch16RNAseq
$ ll -h #查看目录下的文件

$ cat scer_arg_R1.fq |grep "^@" |wc –l
1441645 #scer_arg_R1.fq中一共有1441645条reads
```

查看文件里有多少reads

cat ：显示文件的内容
wc -l ：显示文件行数
grep “^@”：FASTQ中每个read的序列信息都以@开头

2. 读段比对

RNA-Seq分析要把原始Reads序列与基因组或转录组参考序列做比对。与DNA测序数据不同，将转录组测序数据比对到参考基因组上时，需要考虑基因中的内含子(intron)区域对比对过程的影响。比对软件要能够处理大片段的缺失(gap)，用于跨过参考基因组中的内含子区域。

```
#安装Tophat
$ sudo apt install tophat

#构建基因组索引文件
$ bowtie2-build genome.fa genome

#运行Tophat
$ tophat2 \
   -p 4 \ #最多占用四个进程
   --library-type fr-unstranded \ #测序文库的类型
   -G genes.gtf \ #参考基因注释
   -o tophat_arg \ #输出的目录
   genome \ #酿酒酵母的基因组索引
   scer_arg_R1.fq.gz scer_arg_R2.fq.gz

#Tophat运行完成后，检查输出目录
$ ll tophat_arg
accepted_hits.bam #bam格式文件保存序列比对结果
insertion.bed #bed格式文件保存插入突变结果
deletions.bed #bed格式文件保存缺失突变结果
junction.bed #bed格式文件保存潜在exon-exon结果
```

3.比对结果分析

Inspect with samtools

```
#通过samtools将二进制bam文件转换成人类可读的sam格式
$ samtools view tophat_arg/accepted_hits.bam |more

#输出比对的序列数
$ samtools view tophat_arg/accepted_hits.bam |wc

#输出至少有一个read比对到基因组某位置的read pairs的数量
$ samtools view tophat_arg/accepted_hits.bam |cut -f1 |sort -u|wc

#采用samtools自己的命令输出这些比对统计数据
$ samtools flagstat  tophat_arg/accepted_hits.bam

690294 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
690294 + 0 mapped (100.00%:-nan%)
690294 + 0 paired <pan class="hljs-keyword">in<span> sequencing
345086 + 0 read1
345208 + 0 read2
659560 + 0 properly paired (95.55%:-nan%)
662574 + 0 with itself and mate mapped
27720 + 0 singletons (4.02%:-nan%)
2702 + 0 with mate mapped to a different chr
1104 + 0 with mate mapped to a different chr (
```

```
#索引accepted_hits.bam，以加速存取，得到accepted_hits.bam.bai
$ samtools index tophat_arg/accepted_hits.bam
```

3.比对结果分析

View alignment in IGV

- 启动IGV，双击igv.bat就可运行IGV。
- 确保IGV左上角选中酵母基因组“SacCer3”。
- IGV菜单“View->Preferences",在标签"Alignments"中设置"visibility range threshold (kb)"为100kb。此设置IGV中基因组多长时显示reads与比对覆盖度(coverage)。
- 现在导入.bam文件(File->Load from File...)。
- 试试IGV的操作，先选择I号染色体显示测序reads，再把鼠标移动到alignment track，并点右键，在跳出菜单中选"View as pairs"。通过右上角的ruler放大或缩小，观察到基因的部分。鼠标移到单个reads可以看到更多有关比对的信息，其中"Pair orientation"指示测序reads是“stranded”或“unstranded”。在基因注释track的右键菜单选择“Expanded”可以看到更多的基因剪切注释。

问题: 基因*CAR2*/*GAT1*是否有足够的reads覆盖？它有多少种已知的transcript isoforms表达了?

4.转录本组装与定量

```
#把cufflinks可执行程序的路径加入系统环境变量
$ export PATH=$PATH:/mnt/c/test/cufflinks-2.2.1.Linux_x86_64/

#运行cufflinks
$ cufflinks -p 4 -G genes.gtf -o cufflinks_arg tophat_arg/accepted_hits.bam

#查看输出文件
$ ll cufflinks_arg

#查看表达量数据
$ less -S cufflinks_arg/genes.fpkm_tracking

#寻找样本中最高表达量的基因
$ sort -n -k 10 cufflinks_arg/genes.fpkm_tracking

#查看GAT1的基因表达量
$ cat cufflinks_arg/isoforms.fpkm_tracking | grep -w "GAT1"
```

*GAT1*基因表达量

```
YFL021W -       -       YFL021W GAT1    TSS6346 VI:95965-97498  1533    4.75909 76.9045 50.5672 103.242 OK
```

问题:哪个转录本是基因的最高表达的？此结果与前面对比对reads结果的观察结果是否一致？
(酿酒酵母可能很少有不同转录体？)

5.鉴定差异表达基因

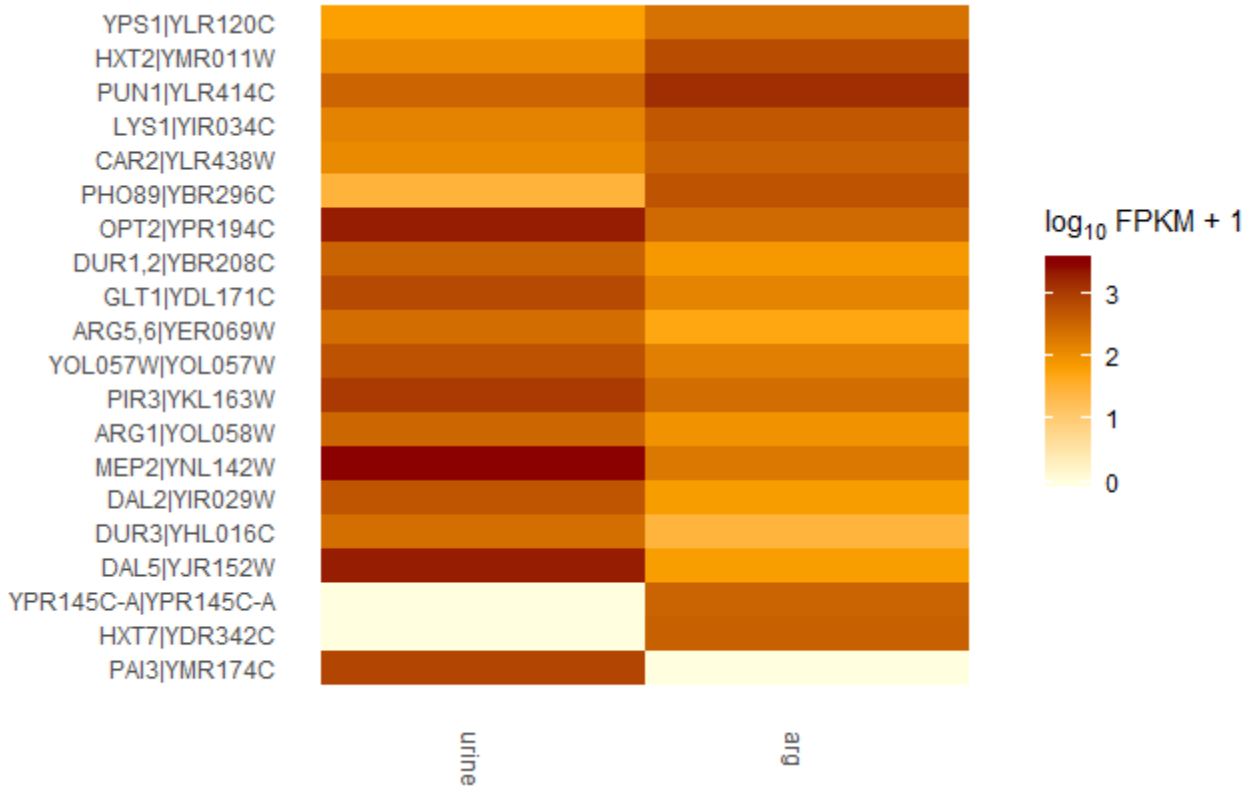
```
$ cuffdiff -p 4 \
   --library-type fr-unstranded \
   -o cuffdiff_out \
   -L urine,arg  \ #-L后为不同样品的名称
   genes.gtf \
   tophat_uri/accepted_hits.bam \#对照组的比对文件
   tophat_arg/accepted_hits.bam  #实验组的比对文件
```

```
#查看输出结果
$ ll cuffdiff_out

$ less -S cuffdiff_out/gene_exp.diff

#显示显著差异表达的基因并按名称降序输出，保存到文件(DE_genes.txt)
$ grep “yes” cuffdiff_out/gene_exp.diff |cut -f1,10,12|sort -n -k 1

$ awk ‘{if($14 == “yes”) print $0}’ cuffdiff_out/gene_exp.diff > DE_genes.txt
```

问题: 这两个样品的差异表达基因列表里有没有你前面看的基因？或你感兴趣的基因？

6.差异基因可视化

```
#安装cummeRbund包
>source("https://www.bioconductor.org/biocLite.R")
>options(BioC_mirror="http://mirrors.ustc.edu.cn/bioc/")
>biocLite("cummeRbund")

#设置工作路径
>setwd("C:/test/ch16RNAseq")

#导入包
>library(cummeRbund)

#读入cuffdiff结果到cuff_data变量中
>cuff_data <- readCufflinks("cuffdiff_out")

#查看数据概要
>cuff_data

>gene_diff <- diffData(genes(cuff_data))

#取前20个差异最显著的基因
>gene_diff_top <- gene_diff[order(gene_diff$p_vaule),][1:20,]

#得到基因的ID
>myGeneIds <- gene_diff_top$gene_id

#根据基因ID得到基因名
>myGenes <- getGenes(cuff_data, myGeneIds)

#绘制热图
>csHeatmap(myGenes, cluster = "both")
```



用cummeRbund绘制的基因差异表达热图

RNA-Seq is still evolving

- Single cell
- Longer reads
- Direct sequencing

Keep updated

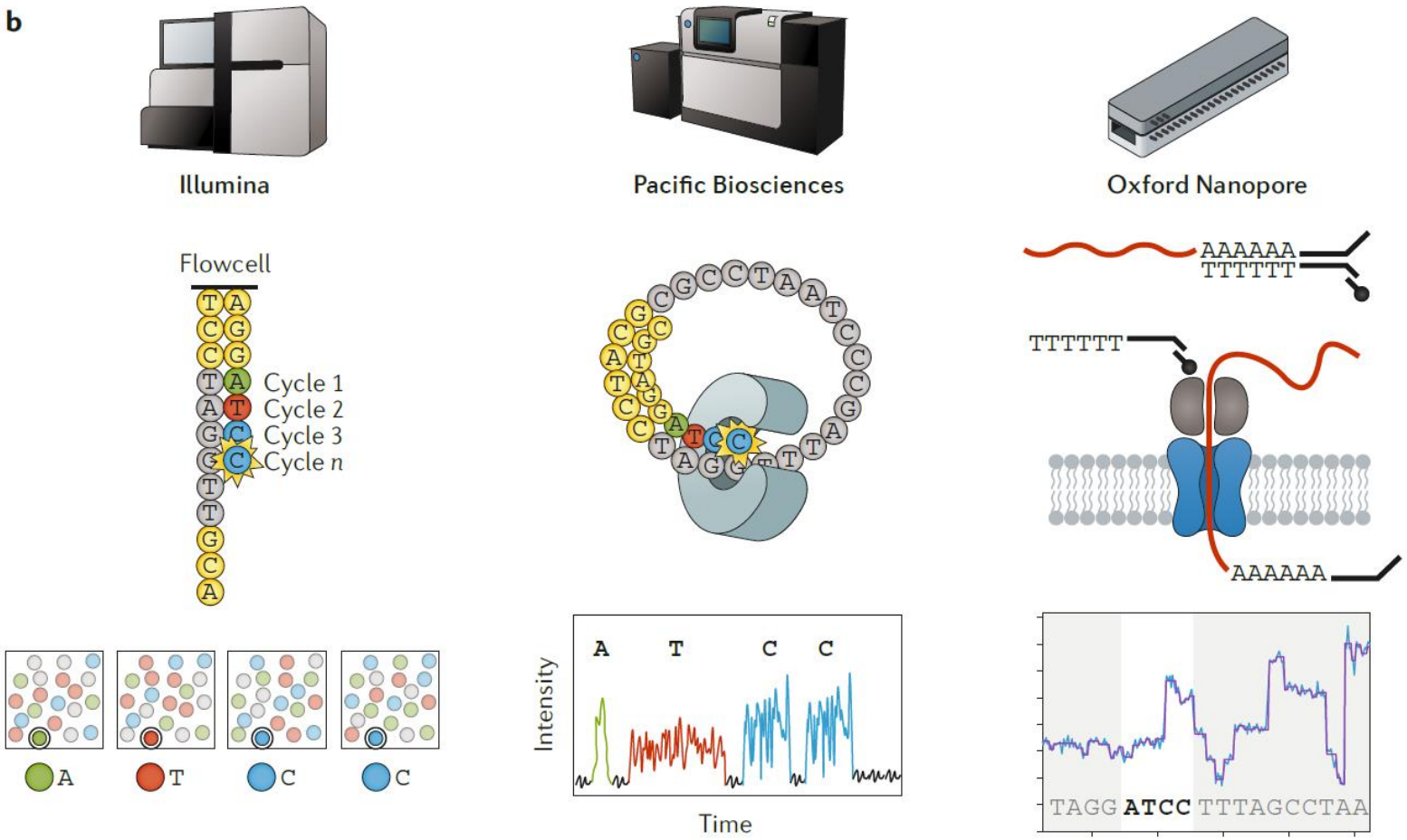
"RNA-Seq is not a mature technology. It is undergoing rapid evolution of biochemistry of sample preparation; of sequencing platforms; of computational pipelines; and of subsequent analysis methods that include statistical treatments and transcript model building." —from ENCODE RNA-Seq analysis guidelines

RNA-seq测序技术进展

二代测序
(Short Reads)

三代测序
(Long Reads)

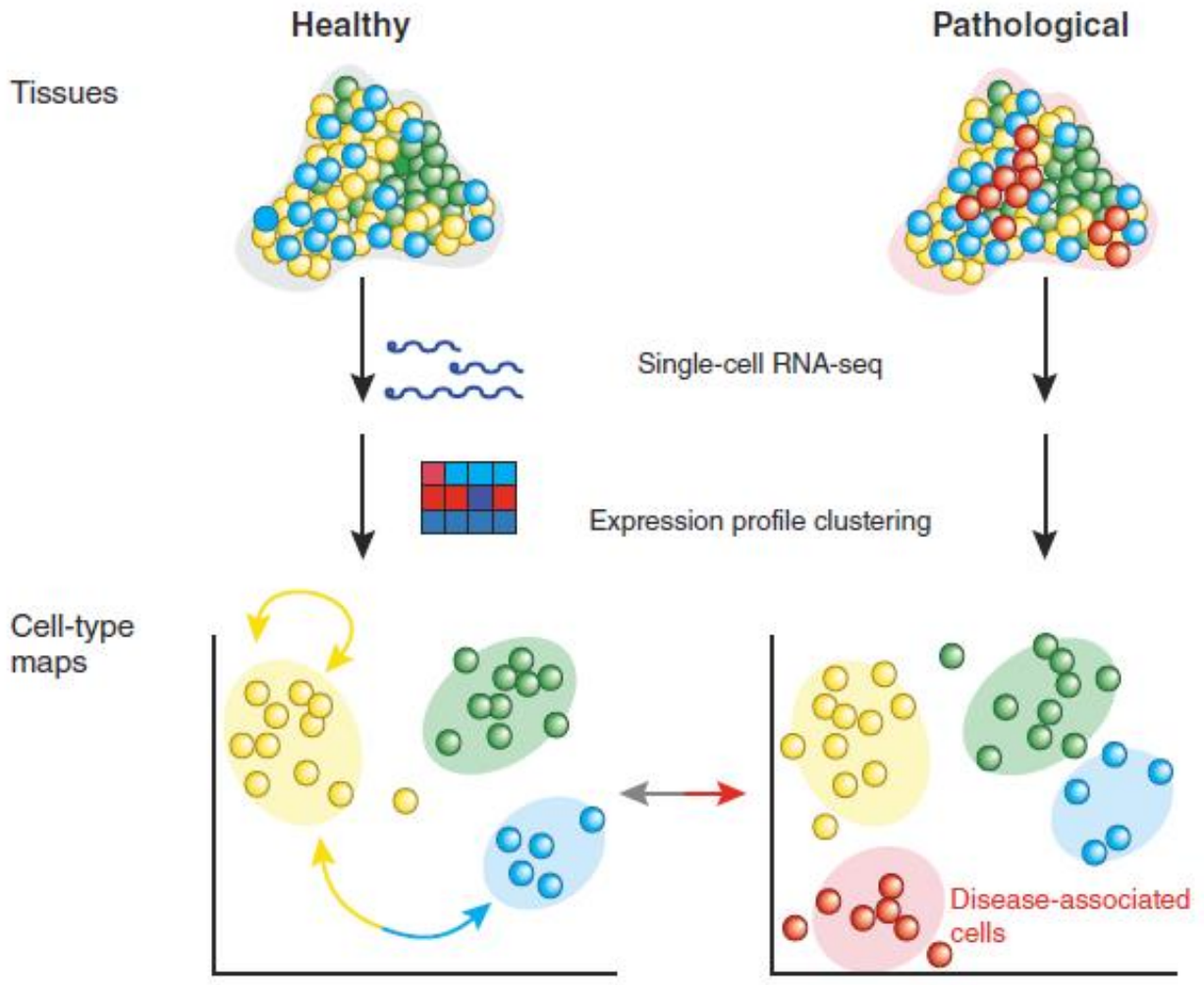
RNA直测
(Direct Sequencing)



Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. Nat Rev Genet 20, 631–656 (2019).

Single-cell transcriptomics (scRNA-Seq)

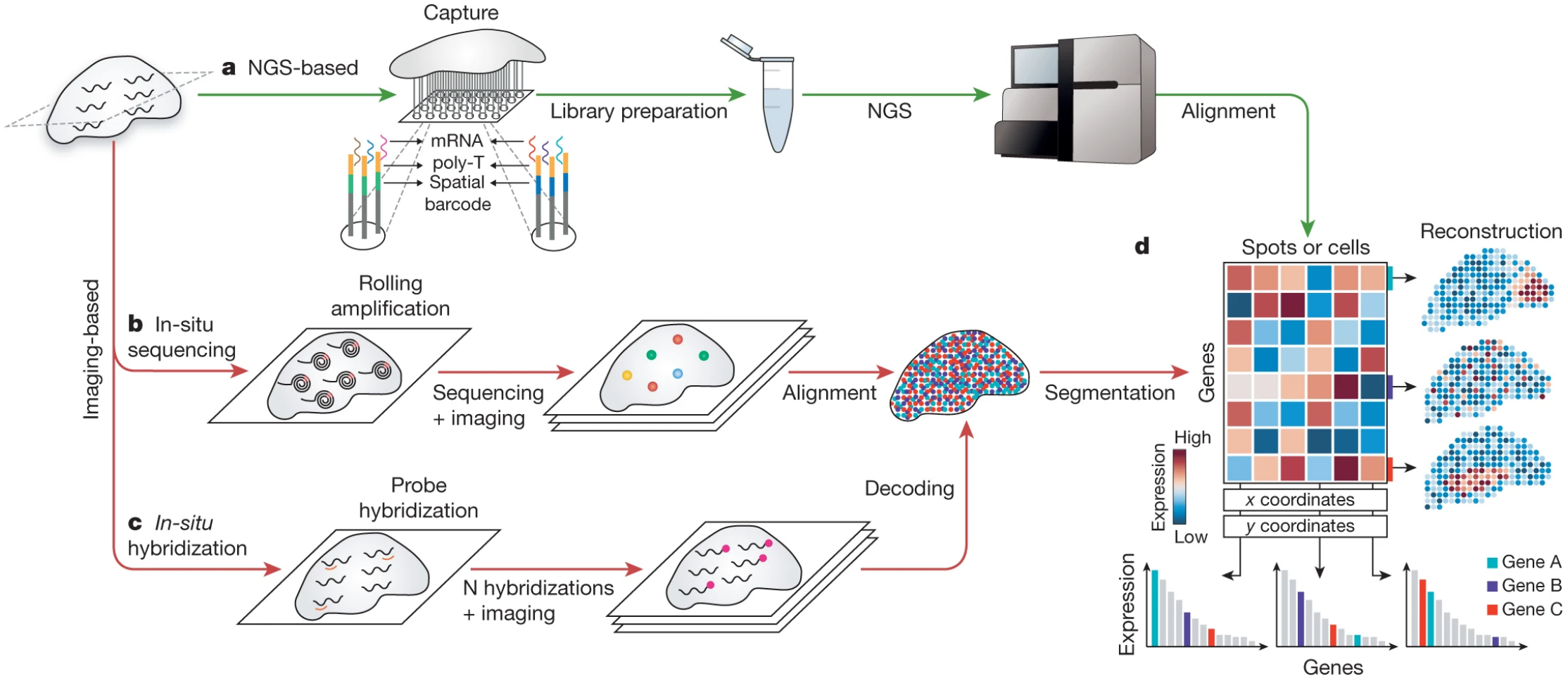
- Single-cell transcriptomics is a powerful way for high-throughput, high-resolution transcriptomic analysis of cell states and dynamics.
- Single-cell RNA-Seq (scRNA-Seq) captures transcriptional profiles in each cell type to describe the genetic basis of their identity and function.



Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods* 2014

Spatial Transcriptomics (STM)

Spatial transcriptomics is a method for assigning cell types to their locations in the histological sections and can also be used to determine subcellular localization of mRNA molecules.



Exploring tissue architecture using spatial transcriptomics: https://www.nature.com/articles/s41586-021-03634-9

RNA-Seq拓展

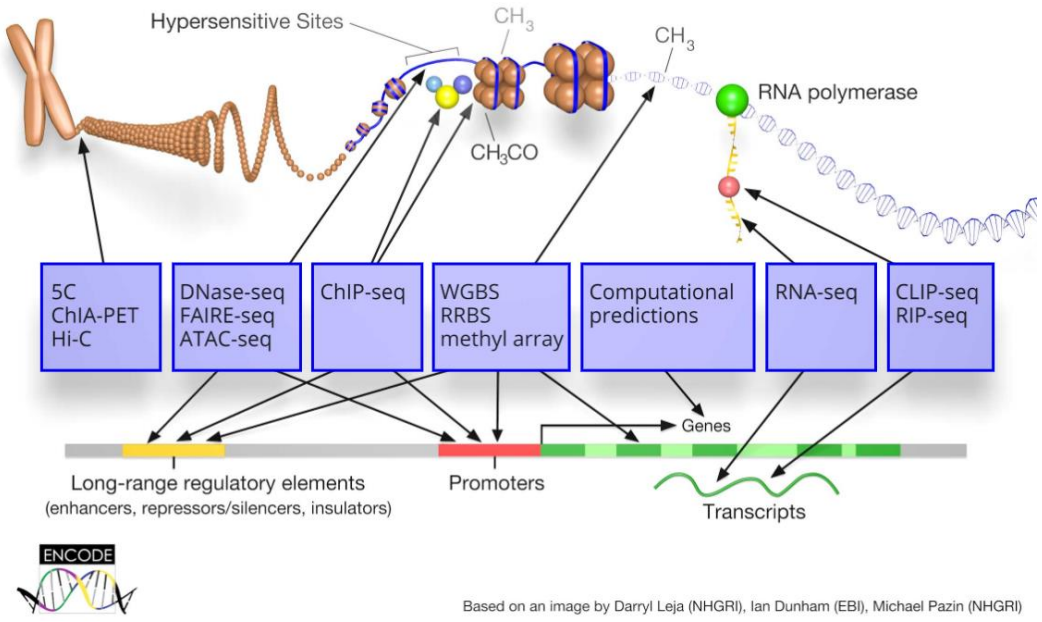
PACIFIC BIOSCIENCES®

10X GENOMICS®

全长转录本（三代测序）

单细胞测序（single cell）

技术革新



多组学整合：基因组、转录组、蛋白组、代谢组、表型组……

非编码RNA：lncRNA，circRNA

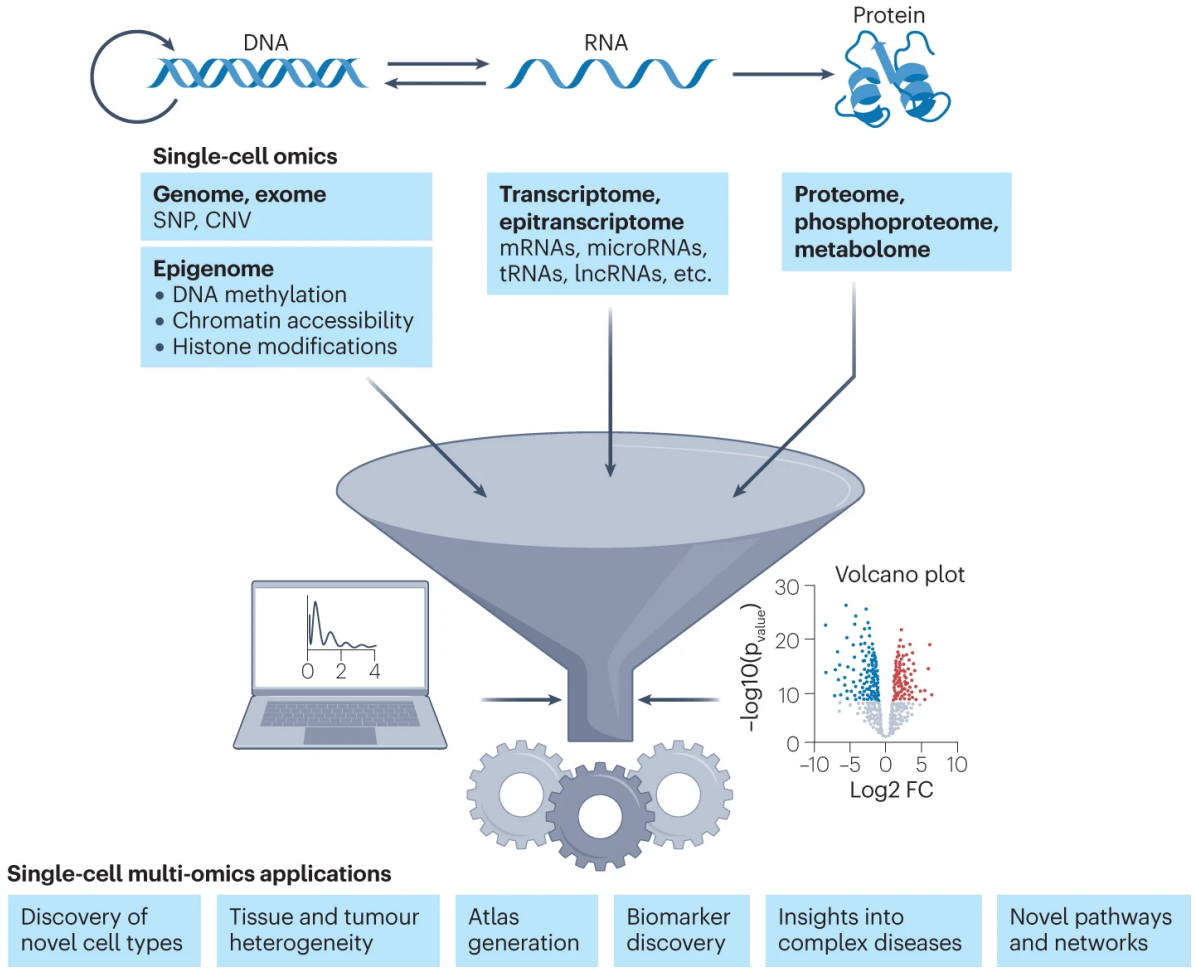
表观转录组：m^6A修饰

空间转录组：时空表达

整合应用

Single-cell multi-omics applications

- 科学家正利用空间单细胞多组学与人工智能技术，以前所未有的细节研究肿瘤，并推动个性化医疗的实现。



Baysoy, A., Bai, Z., Satija, R. *et al.*, Nat Rev Mol Cell Biol 24, 695–713 (2023).

习题

1．简要说明RPKM、FPKM与TPM的概念与计算方法。

2．下载并安装Tophat2、Bowtie2、CUFFLINKS等软件，并用附件的数据进行RNA-Seq分析，掌握软件的常用参数，及结果文件的格式。

3．尝试在R语言中安装CummRbund包，并绘制差异表达基因的热图。