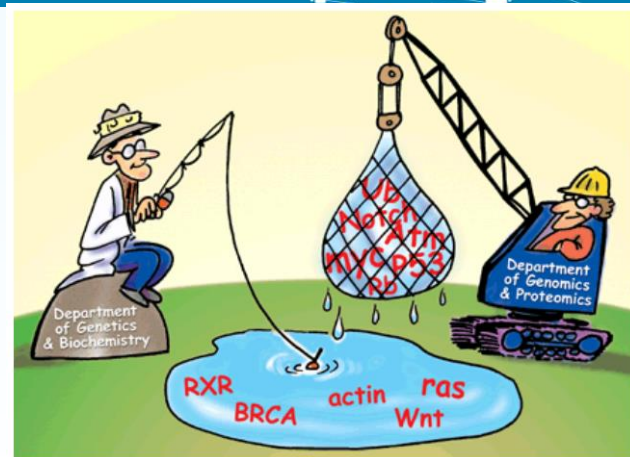


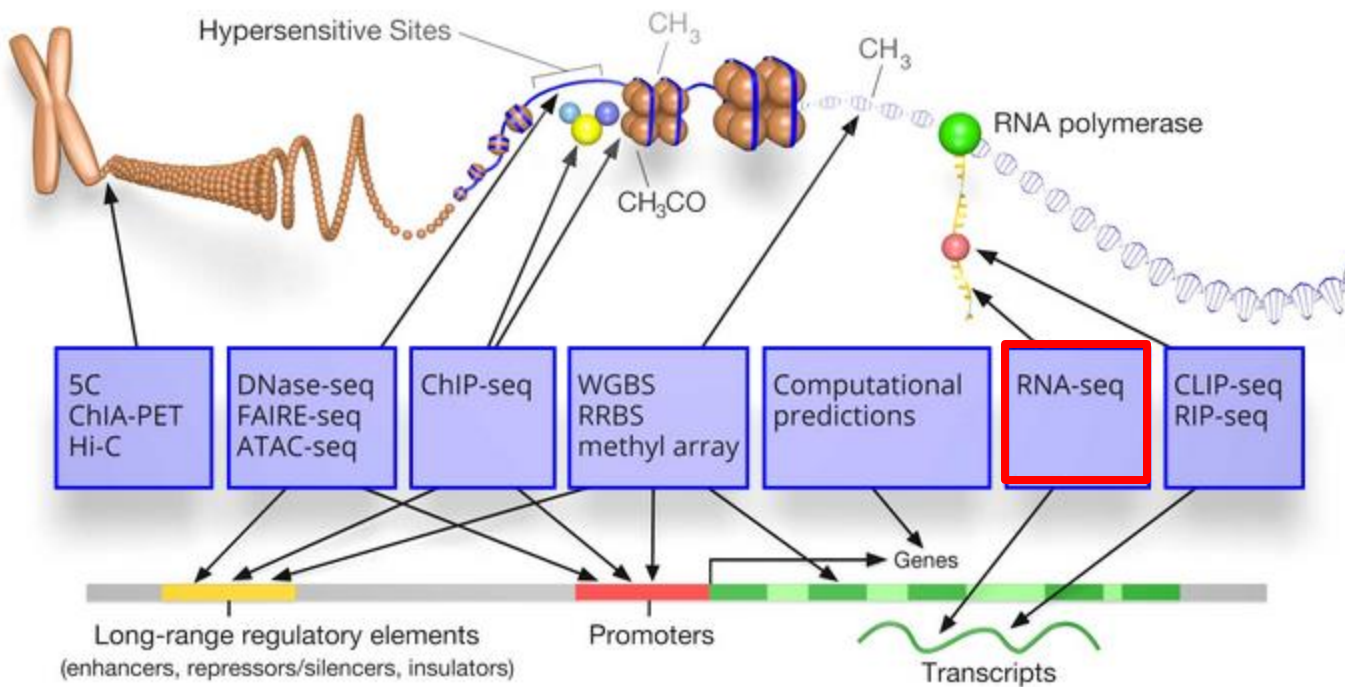


## 11 转录组概述



图片来源: Science

# 测序技术的应用



宏基因组



Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

易生信，毕生缘，培训版权所有。

# 转录组 (transcriptome)的概念

- The transcriptome is the set of all RNA molecules in one cell or a population of cells. It is sometimes used to refer to all RNAs, or just mRNA, depending on the particular experiment.
- Transcriptome这个概念最早在 20 世纪 90 年代提出；从简单的PCR、表达序列标签 (EST: expressed sequence tag)、基因表达芯片，到454测序仪推出后，2006年发表第一篇基于高通量测序的转录组文章，2008年首次出现RNA-seq这个概念。

Cell, 1997 Jan 24;88(2):243-51.

## Characterization of the yeast transcriptome.

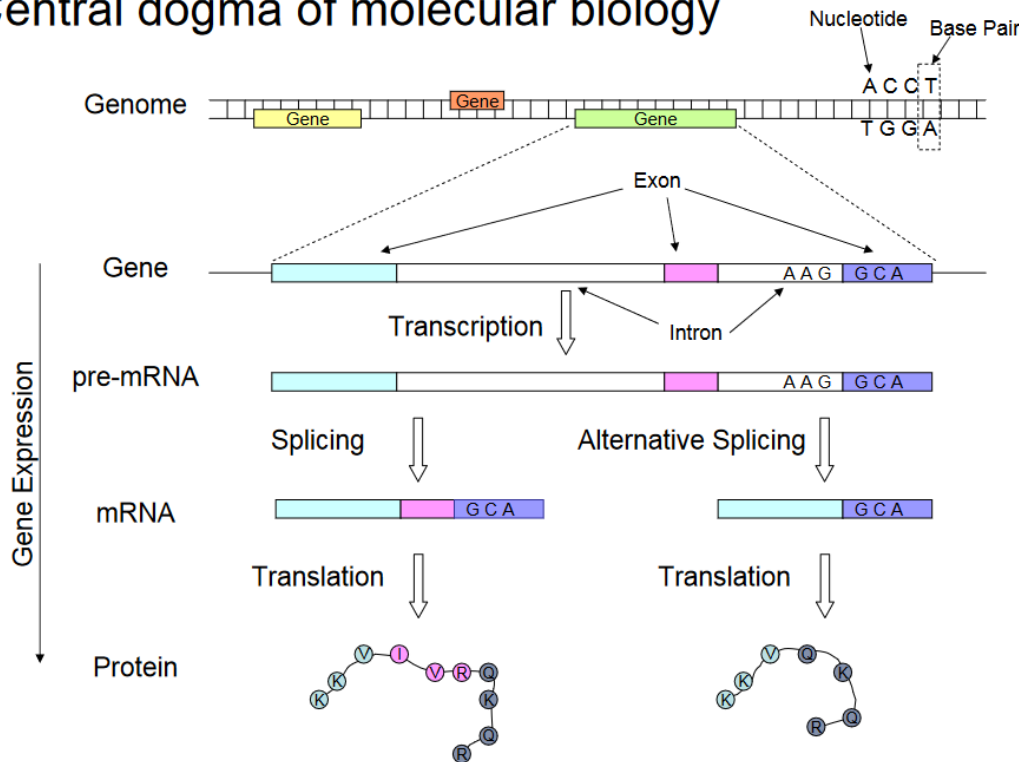
Genome Res. 1999 Feb;9(2):195-209.

Velculescu VE<sup>1</sup>, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW.

## The Genexpress IMAGE knowledge base of the human brain transcriptome: a prototype integrated resource for functional and computational genomics.

Piétu G<sup>1</sup>, Mariage-Samson R, Fayein NA, Matingou C, Eveno E, Houlgatte R, Decraene C, Vandenbrouck Y, Tahi F, Devignes MD, Wirkner U, Ansorge W, Cox D, Nagase T, Nomura N, Auffray C.

## Central dogma of molecular biology

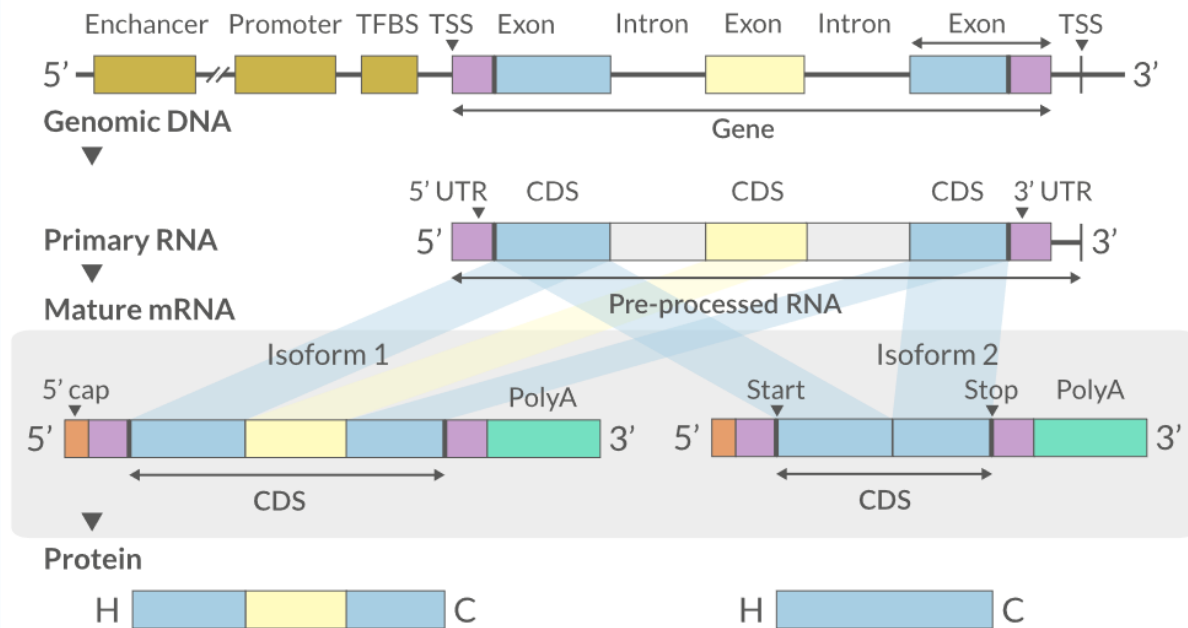


92-94% of human genes undergo alternative splicing,  
86% with a minor isoform frequency of 15% or more  
*E.T. Wang, et al, Nature 456, 470-476 (2008)*

宏基因组

信使RNA

# 为什么测序RNA?



- The transcriptome is spatially and temporally dynamic
- Data comes from functional units (coding regions)
- Only a tiny fraction of the genome

# 人体内细胞中的DNA都一样吗？

- 淋巴前体细胞存在 *VDJ* 重组现象。*VDJ* 是 *variable* (V), *diversity* (D) 和 *joining* (J) 基因区域的缩写，目的是编码产生不同的抗体。
- 生殖细胞中 等位基因的交换。
- 皮肤上皮细胞：25% 的上皮细胞中，携带着与癌症相关的基因突变。
- 神经元：单细胞测序表明，人体 13%-41% 的神经元，存在大片段的基因拷贝数变化，有的增多，有的减少。
- 肝细胞：在正常的肝细胞中，存在由异常有丝分裂以及细胞融合形成的多倍染色体和非整倍染色体的现象。最高可占全部肝细胞的 50%。





# 人类泛基因组整合 47 个个体基因组反应遗传多样性

- 为当下的人类参考基因组(GRCh38)添加了1.19亿碱基对(常染色体多态性序列,其中 9 千万来自于结构变异)和1115个重复基因。

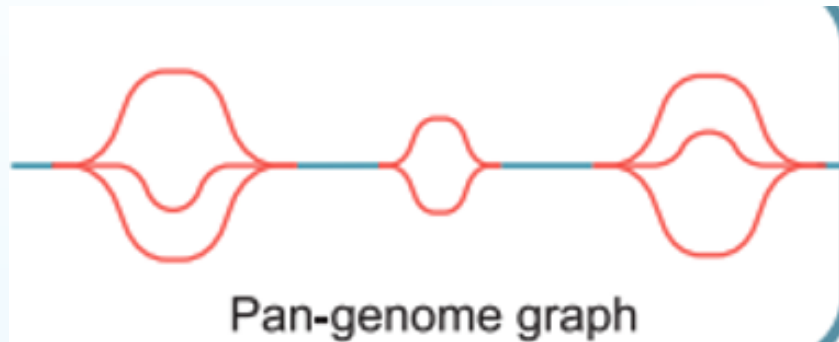
<https://mp.weixin.qq.com/s/KcBU07XQgHM2ozBvs3lIA> Nature — 人类首个“泛基因组”旨在编目人类遗传多样性

<https://www.nature.com/collections/aebdjihcda>



# 不同个体的基因组一致吗？ Pangenome

- 在水稻3K项目中应用“`map-to-pan`”策略产生了在植物泛基因组中解析的首批序列之一 (约630 Mb)，揭示了粳稻品种缺少的约268 Mb的新序列。最近的两项研究使用来自111种和251种不同水稻材料的基因组组合将这一估计值分别提升到约1250 Mb和约1520 Mb。



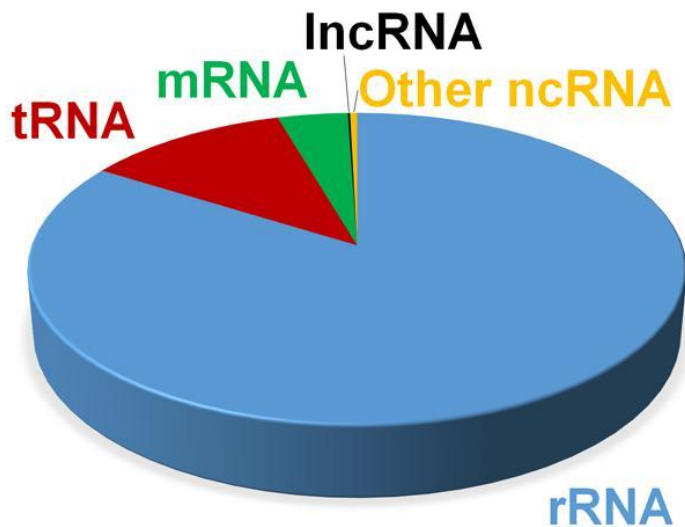
- 一般来说，随着更多基因组的加入，所有个体中存在的“核心”基因的百分比都会下降，最终到达一个约为35%的基准水平，就像在二穗短柄草 (*Brachypodium distachyon*) 中一样。这些“可有可无”的基因通常更值得注意，在水稻中的两项独立研究报告了类似数量的在粳稻品种中缺失的基因 (约1万个)；这些基因通常富集在免疫和防御反应途径中，具有很大的抗药性育种潜力。玉米和小麦的泛基因数量更高，超过100,000 (表1)，而玉米代表性参考基因组B73仅注释了约40000个基因。





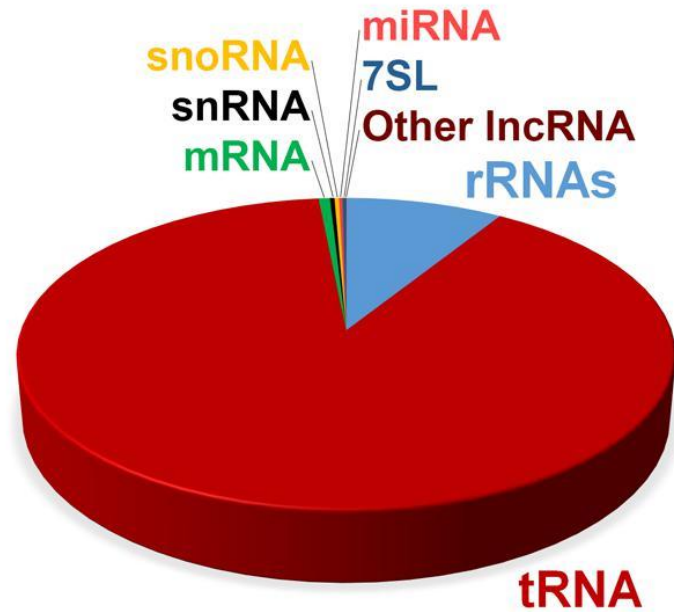
# 不同RNA种类总量不同，分子数不同

A



RNA by mass

B

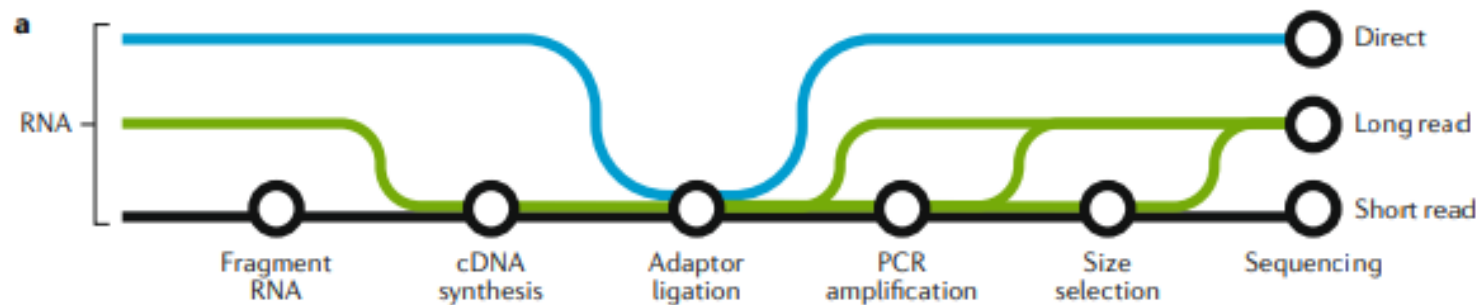


RNA by number of molecules

宏基因组

易生信

# Short read, long read, direct RNA sequencing



基因组

生信信
















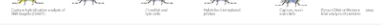



易生信

# 不同转录组测序技术

- 常规转录组
- 新生转录本 GRO-seq
- 正在翻译的转录本 Ribo-seq
- RNA与蛋白的结合 \*-clip/RIP
- RNA-DNA的结合ChIRP
- 转录本两端鉴定 PET-seq
- sRNA-seq/circRNA/lncRNA

## RNA Transcription

select a method below

|                                       |   |
|---------------------------------------|---|
| Ribo-Seq/ART-Seq/GTI-Seq              |    |
| PAR-CLIP                              |    |
| GRO-Seq/BRIC-Seq/Bru-Seq/BruChase-Seq |    |
| HITS-CLIP/CLIP-Seq/PTB-Seq            |    |
| RIP-Seq                               |    |
| ChIRP-Seq                             |    |
| NET-Seq                               |    |
| CAGE-Seq                              |    |
| PARE-Seq                              |    |
| AGO-CLIP                              |    |
| TIF-Seq                               |    |
| FRT-Seq                               |    |
| Repli-Seq                             |    |
| RAP                                   |    |
| 5'-GRO-Seq                            |    |
| BruDRB-Seq                            |    |
| CHART                                 |   |
| RAP-RNA                               |  |
| 4sUDRB-Seq                            |  |

宏基因组



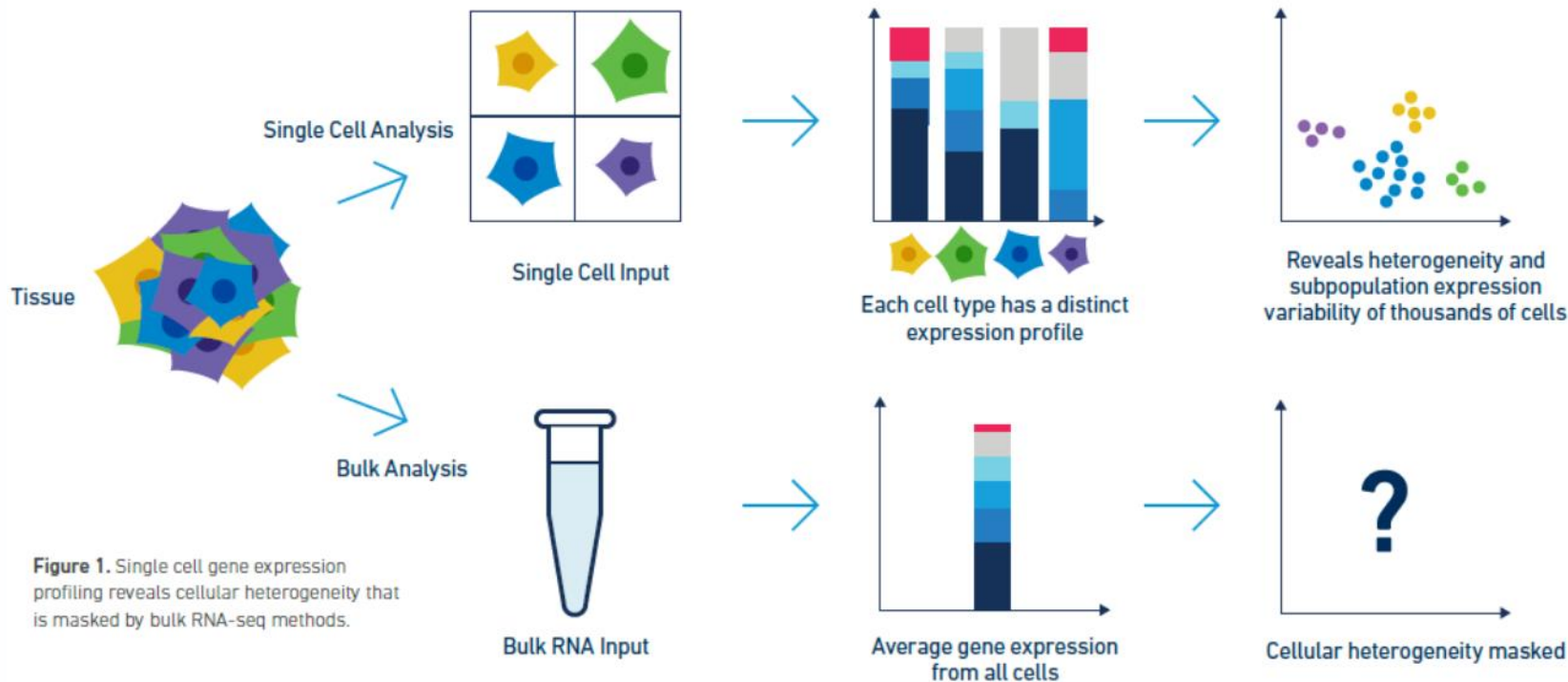
- sRNA: miRNA, piRNA (切胶筛选或磁珠分选, 本身是链特异性库)
- 常规转录组: poly-A捕获
- circRNA: rRNA移除, RNase消化
- 较“全”转录组测序: rRNA移除, 测序lncRNA, mRNA, circRNA

宏基因组  
生信宝典

易生信



# 普通转录组测序结果是混合样品的结果



**Figure 1.** Single cell gene expression profiling reveals cellular heterogeneity that is masked by bulk RNA-seq methods.

基因组

易生信

# 细胞中RNA量的估计

- 单细胞里面有1-50 pg RNA (单个典型哺乳动物细胞有10-30 pg RNA)
- T细胞中有1-2 pg RNA;  $10^5$ - $10^6$ 个T细胞可以提取约200 ng RNA
- 一个哺乳动物细胞中大约有360,000个mRNA分子, 12,000不同种类的转录本
- 低丰度mRNA可能在单个细胞中只有1-15个拷贝。

QIAGEN

*RNA Sequencing and Analysis*



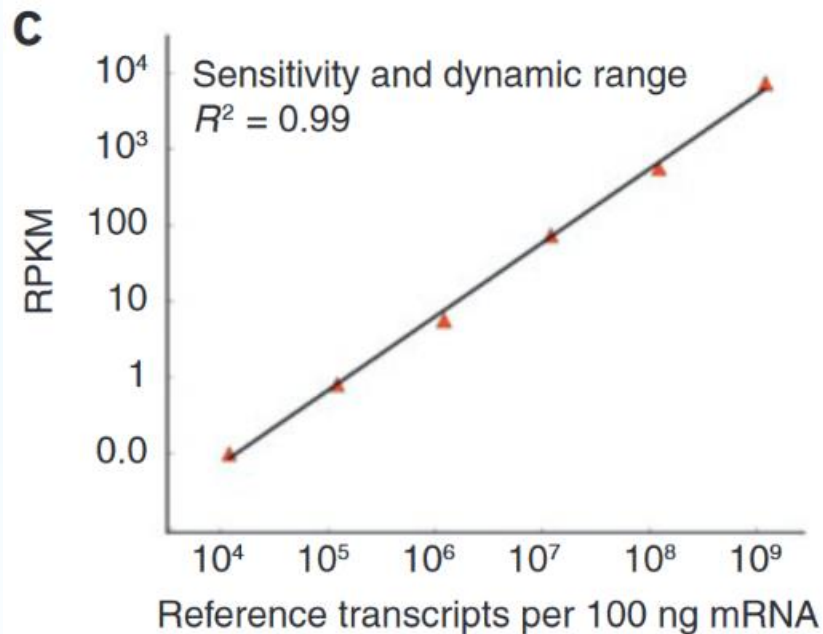
易汉博基因科技(北京)有限公司  
EHBIO Gene Technology (Beijing) co., LTD

Evaluation of ultra-low input RNA sequencing for the study of human T cell transcriptome



# 转录组检测技术的敏感性

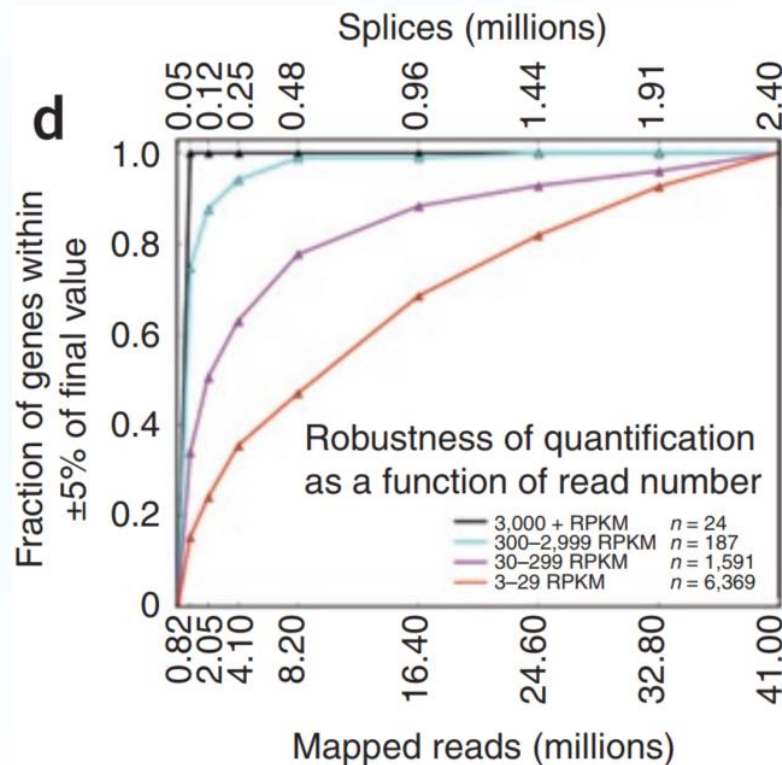
- 把6个体外合成的长度为0.3-10 kb 长度的转录本加到肝脏样品中，检测其检出率
- 加入 $1.2 \times 10^4$ 条转录本时，几乎检测不到，RPKM值为0
- 增加更多的转录本时，检测出的RPKM值与加入的转录本量正相关
- 加入 $1.2 \times 10^9$ 条转录本后仍未达到检测上限



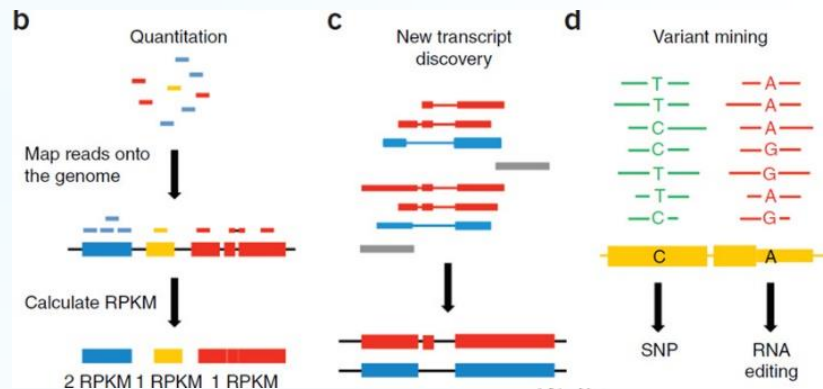
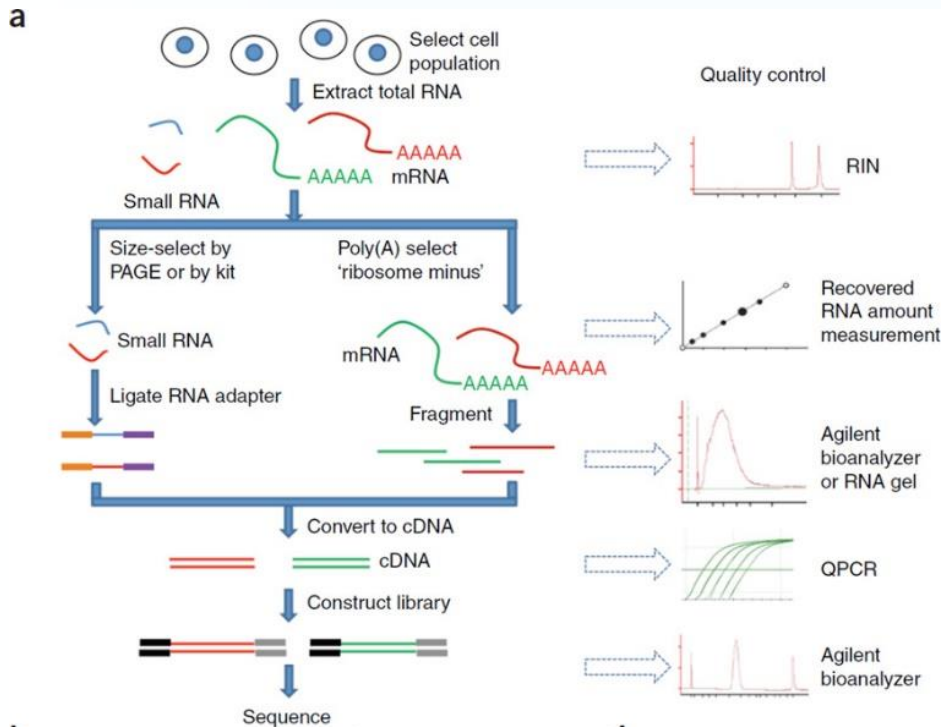
易生信

# 转录组RPKM量化的稳定性

- 211 个最高表达的基因（黑色、浅蓝色）在8 million有效reads下即可检测准确
- 丰度低的基因（紫色、红色）一直未达到检测饱和
- RPKM为3大体等同于一个肝脏细胞中 存 在 一 个 转 录 本



# 转录组流程及应用



[Technical considerations for functional sequencing assays](#)

# RPKM的计算

Gene A 600 bases

Gene B 1100 bases

Gene C 1400 bases

$$\text{RPKM} = 12 / (0.6 * 6) = 3.33$$

$$\text{RPKM} = 24 / (1.1 * 6) = 3.64$$

$$\text{RPKM} = 11 / (1.4 * 6) = 1.31$$



$$\text{RPKM} = 19 / (0.6 * 8) = 3.96$$

$$\text{RPKM} = 28 / (1.1 * 8) = 1.94$$

$$\text{RPKM} = 16 / (1.4 * 8) = 1.43$$

易生信

请语音提问!



**It's QUESTION TIME !!**

# 转录组测序实验设计原则和注意事项-1

- 明确要解决的生物学问题
  - 实验驱动（课题需要转录组数据的支持和验证）
  - 问题驱动（从转录组数据出发开展课题研究）
- 尽可能一次性完成测序数据收集，避免批次效应
- 合理的对照样本
  - 同一组织不同时间或发育阶段的比较
  - 化合物处理 vs. 非处理
  - 癌组织 vs. 癌旁或正常组织

宏基因组

生信宝典

易生信





# 转录组测序实验设计原则和注意事项-2

## ○ 生物学重复

- In all cases, experiments should be performed with **two or more biological replicates**. In general, detecting and quantifying low prevalence RNAs is inherently more variable than high abundance RNAs. As part of the ENCODE pipeline, annotated transcript and genes are quantified using RSEM and the values are made available for downstream correlation analysis. Replicate concordance: the gene level quantification should have a **Spearman correlation of  $> 0.9$**  between isogenic replicates and  **$> 0.8$**  between anisogenic replicates.
- 建议至少**3个**生物学重复

易生信 毕生缘 培训版权所有

易生信



# 为什么需要生物学重复？

- Technical variability in the RNA-seq procedures
- The biological variability of the system
- For a proper statistical power analysis

宏基因组

生信宝典

易生信

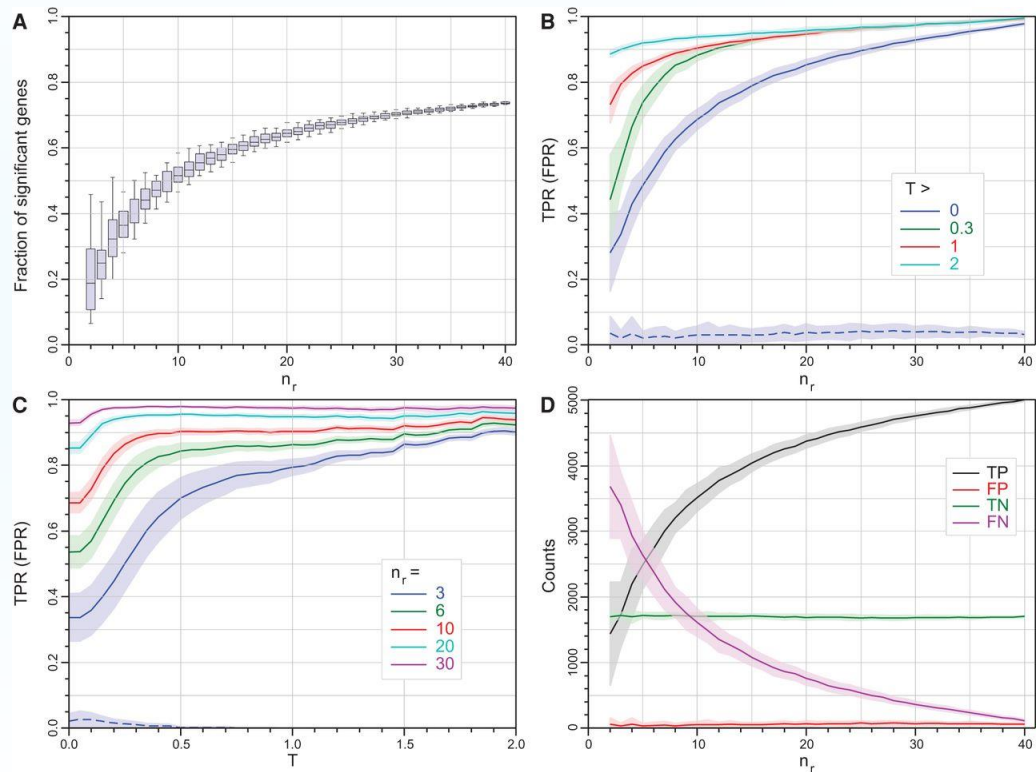


# 但实际3个重复远不够，重复数取决于项目预算

在控制相同的假阳性率条件下测序  
3个生物学重复鉴定的差异基因平均数目只占测序48个生物重复鉴定的差异基因数目的35%左右；

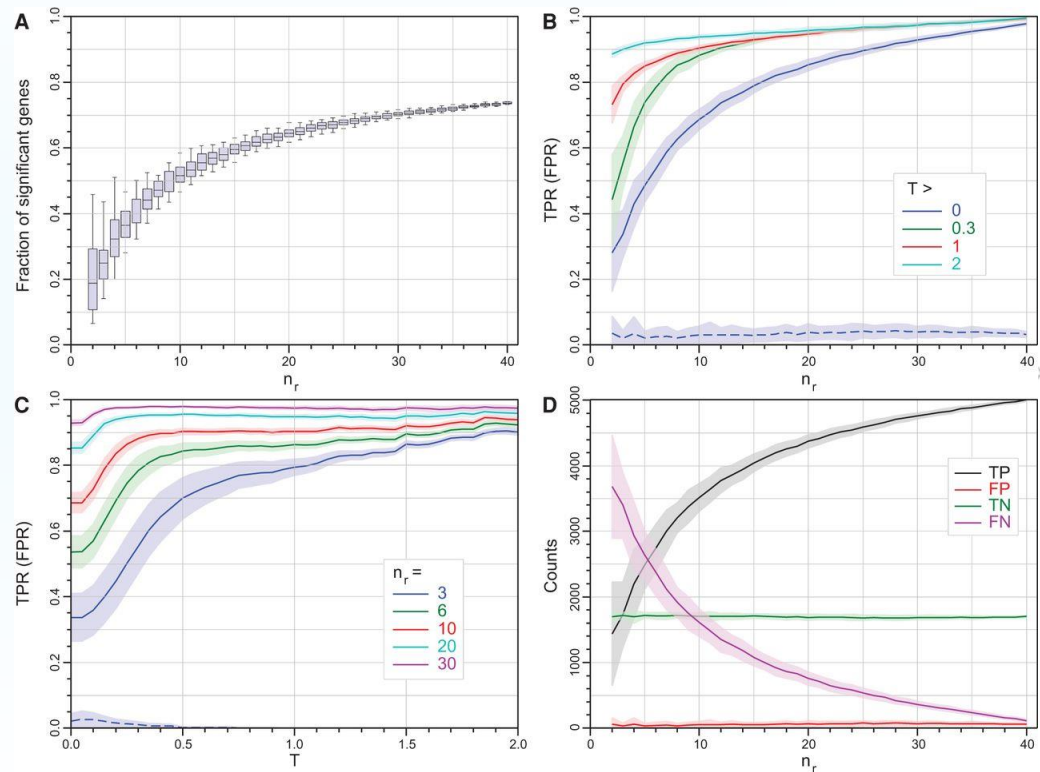
从48个生物重复中多次随机抽取不同组合的3个生物学重复样品鉴定出的差异基因最多时是最少时的3倍，所以测序3个生物学重复获得的差异基因分析结果存在很大的随机性。

而测序20个生物学重复鉴定的差异基因占测序48个生物学重复鉴定的差异基因的92%以上，且多次抽样结果稳定。

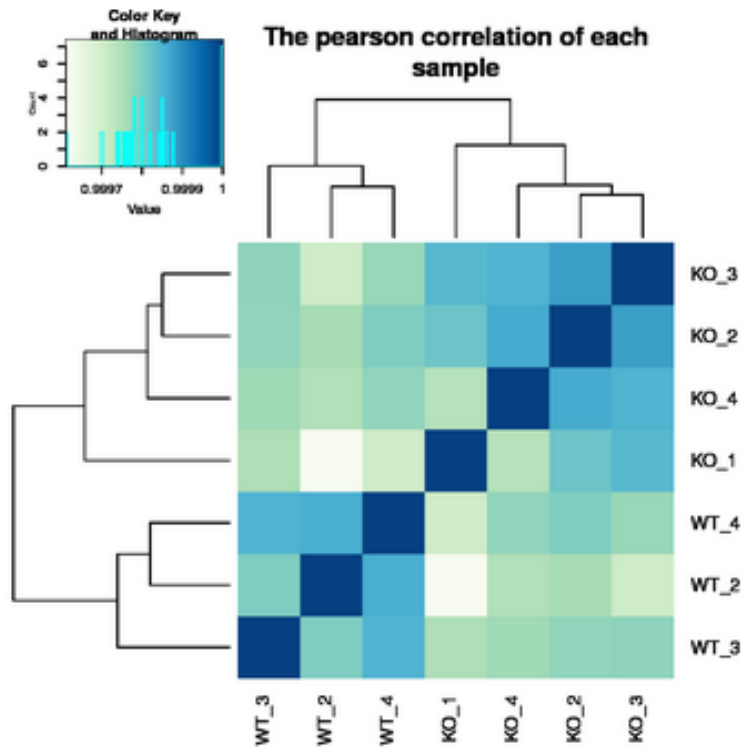
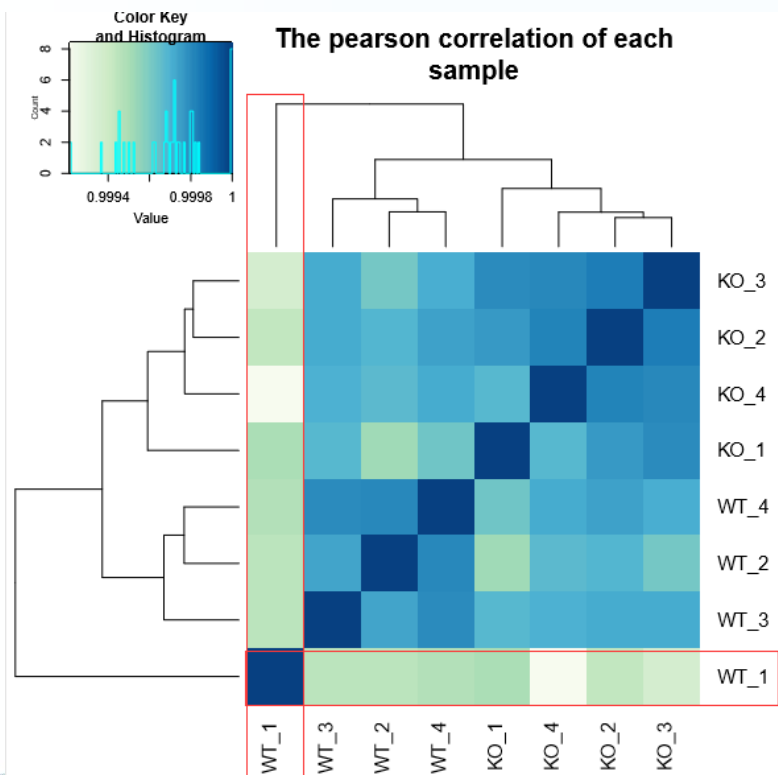


# 但实际3个重复远不够，重复数取决于项目预算

A 48-replicate yeast study showed that many of the tools available for DGE analysis detected only **20–40%** of differentially expressed genes when only 3 replicates were included in the analysis.

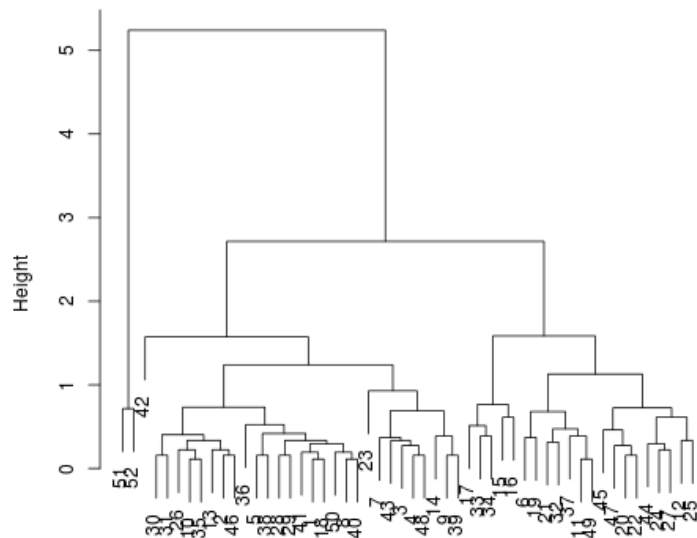


# 什么样的生物重复是合适的？ 组间差异和组内差异？



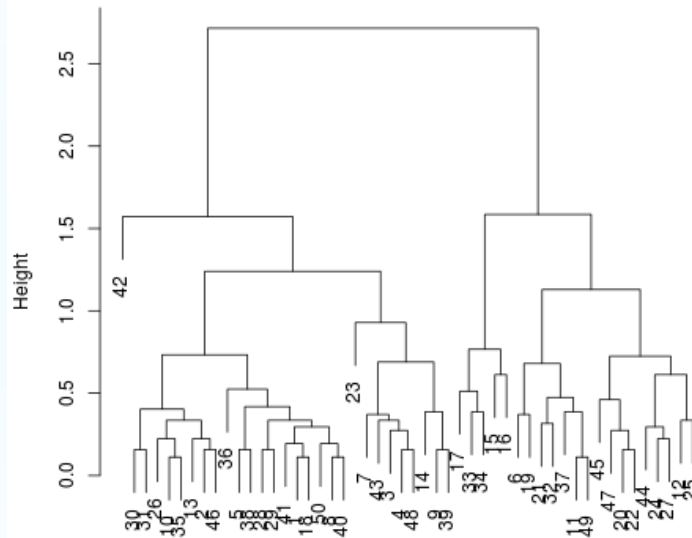
7/12

Iris data. Obs 1-50 from one species, 51-52 from another.



D  
hclust (\*, "complete")

Iris data. Obs 1-50 from same species.



D  
hclust (\*, "complete")

<https://stats.stackexchange.com/questions/101254/outlier-detection-using-clustering-and-dissimilarity-matrix-in-r>



# 不同距离计算方法和聚类算法会得到不同的聚类结果

## Cluster method ⓘ

☒ Complete ☐ Ward.D ☐ Ward.D2 ☐ Single ☐ Average ☐ Mcquitty ☐ Median ☐ Centroid

## Row distance matrix method ⓘ

☒ Pearson ☐ Euclidean ☐ Manhattan ☐ Maximum ☐ Canberra ☐ Binary ☐ Minkowski ☐ Spearman ☐ Bray

☐ Kulczynski ☐ Jaccard ☐ Gower ☐ AltGower ☐ Morisita ☐ Horn ☐ Mountford ☐ Raup ☐ Binomial

☐ Chao ☐ Cao ☐ Mahalanobis

## Column distance matrix method ⓘ

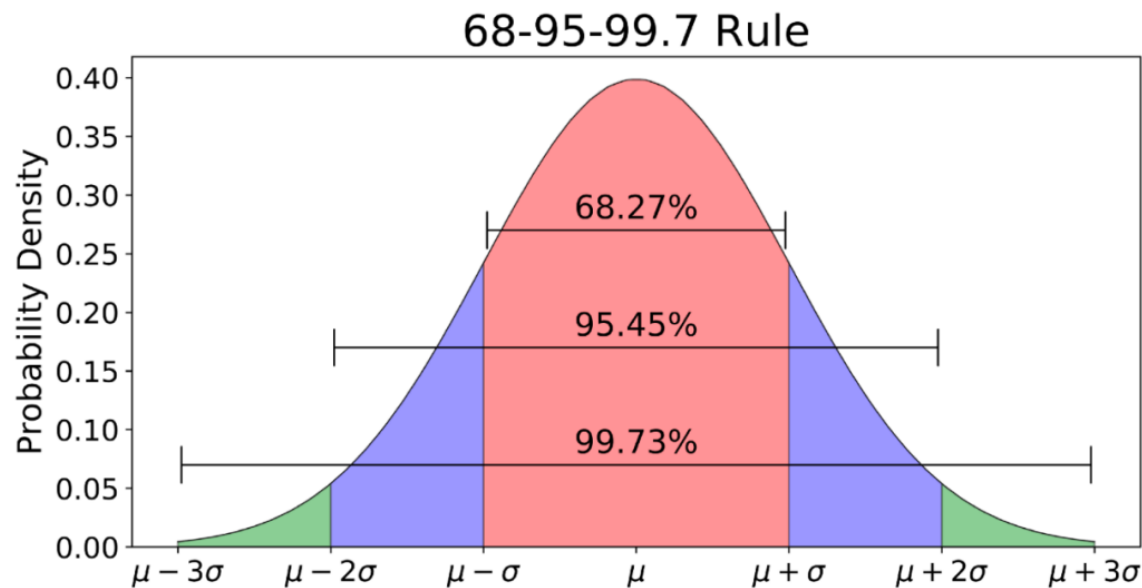
☒ Pearson ☐ Euclidean ☐ Manhattan ☐ Maximum ☐ Canberra ☐ Binary ☐ Minkowski ☐ Spearman ☐ Bray

☐ Kulczynski ☐ Jaccard ☐ Gower ☐ AltGower ☐ Morisita ☐ Horn ☐ Mountford ☐ Raup ☐ Binomial

☐ Chao ☐ Cao ☐ Mahalanobis



# 如何鉴定一组数中的异常值?



$$Z = \frac{X - \mu}{\sigma}$$

|          |   |                              |
|----------|---|------------------------------|
| Z        | → | Standard (Normal) or Z score |
| X        | → | member element of group      |
| $\mu$    | → | mean of expectation          |
| $\sigma$ | → | standard deviation           |

如何鉴定异常值

# 不同样品表达谱相关性矩阵

| id            | untrt_N080611 | trt_N080611 | untrt_N61311 | untrt_N052611 | untrt_N061011 | trt_N61311 | trt_N052611 | trt_N061011 |               |
|---------------|---------------|-------------|--------------|---------------|---------------|------------|-------------|-------------|---------------|
| untrt_N080611 | 1             | 0.9771      | 0.9779       | 0.9796        | 0.9757        | 0.9619     | 0.9564      | 0.9549      |               |
| trt_N080611   | 0.9771        | 1           | 0.9555       | 0.9636        | 0.9569        | 0.9722     | 0.9746      | 0.9727      |               |
| untrt_N61311  | 0.9779        | 0.9555      | 1            | 0.9838        | 0.9826        | 0.9764     | 0.9581      | 0.9574      |               |
| untrt_N052611 | 0.9796        | 0.9636      | 0.9838       | 1             | 0.9852        | 0.9683     | 0.9745      | 0.9654      |               |
| untrt_N061011 | 0.9757        | 0.9569      | 0.9826       | 0.9852        | 1             | 0.964      | 0.9631      | 0.9727      |               |
| trt_N61311    | 0.9619        | 0.9722      | 0.9764       | 0.9683        | 0.964         | 1          | 0.9803      | 0.9798      |               |
| trt_N052611   | 0.9564        | 0.9746      | 0.9581       | 0.9745        | 0.9631        | 0.9803     | 1           | 0.9867      |               |
| trt_N061011   | 0.9549        | 0.9727      | 0.9574       | 0.9654        | 0.9727        | 0.9798     | 0.9867      | 1           |               |
| Sum           | 6.7835        | 6.7726      | 6.7917       | 6.8204        | 6.8002        | 6.8029     | 6.7937      | 6.7896      | sum(B2:B9)-1  |
| mean          | 6.794325      |             |              |               |               |            |             |             | mean(B10:I10) |
| sd            | 0.013260444   |             |              |               |               |            |             |             | std(B10:I10)  |
| K             | -0.816337673  | -1.6383313  | -0.19795717  | 1.96637458    | 0.443047005   | 0.6466601  | -0.0471327  | -0.3563229  | (B10-B11)/B12 |
| Threshold     | -2            |             |              |               |               |            |             |             |               |
| Outlier       | FALSE         | FALSE       | FALSE        | FALSE         | FALSE         | FALSE      | FALSE       | FALSE       | B13<B14       |

易生信

# 如何鉴定异常样本

- 计算样本两两之间的整体表达谱相似性值，可用**Pearson**相关性或**Spearman**相关性
- 每个样本与其它所有样本的相关性之和记为该样本与整体数据的最终相似度值，写为 $S_i$
- 计算所有 $S_i$ 的均值和标准差，获得每个样本的 $Z_i$ 值
- 基于样本中大部分基因表达不变、整体表达谱相似的原则，定义 $Z_i < -3$ （表达偏离整体的样本）的样本为异常样本

宏基因组  
生信宝典

易生信



# 转录组测序实验设计原则和注意事项-3

## ○ 建库类型

- 随机引物和oligo (dT)反转录引物
- 链特异性文库或非链特异性文库

## ○ 测序深度（指测序得到的碱基数与待测转录组大小的比值）

- 普通差异基因分析只需中等测序深度，一般测序都可以满足需求
- 检测新转录本需深度测序，可变剪接需要50-100 million可用reads
- Long RNA-Seq library: 30 million aligned reads/mate-pairs.
- RAMPAGE library: 20 million aligned reads/mate-pairs.
- small RNA-Seq library: 30 million aligned reads/mate-pairs.



# 测序环节的几个概念

## ○ 什么是Raw data和Clean data?

- Raw data是指测序仪下机产生的数据，可能包括接头和低质量数据
- Clean data是指去掉接头和低质量碱基后的可用数据

## ○ 5 G测序量指什么?

- 5 G是指有 $5 \times 10^9$ 个碱基
- 如果单端100 nt, reads数就是50 Million
- 如果双端100 nt, reads数就是左右两端各25 Million

宏基因组

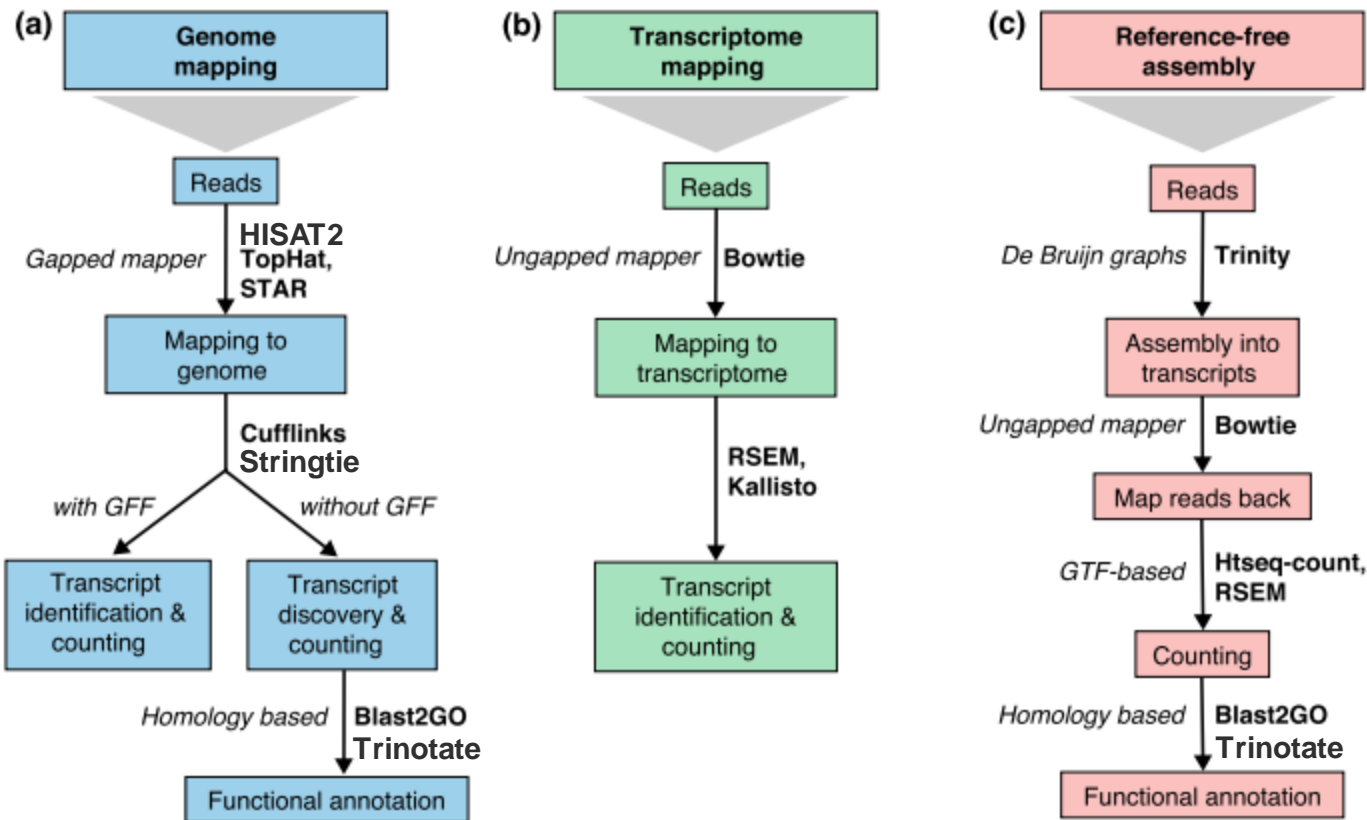
生信宝典

易生信

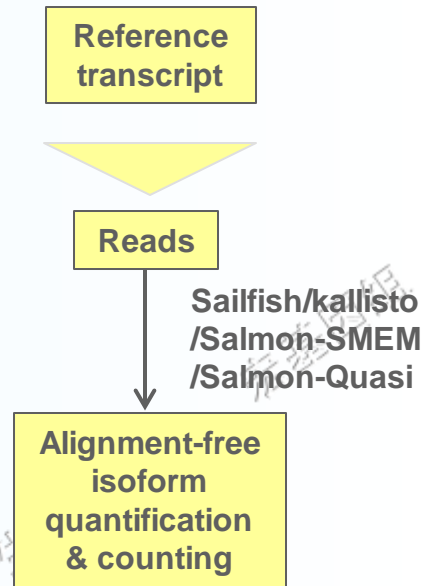




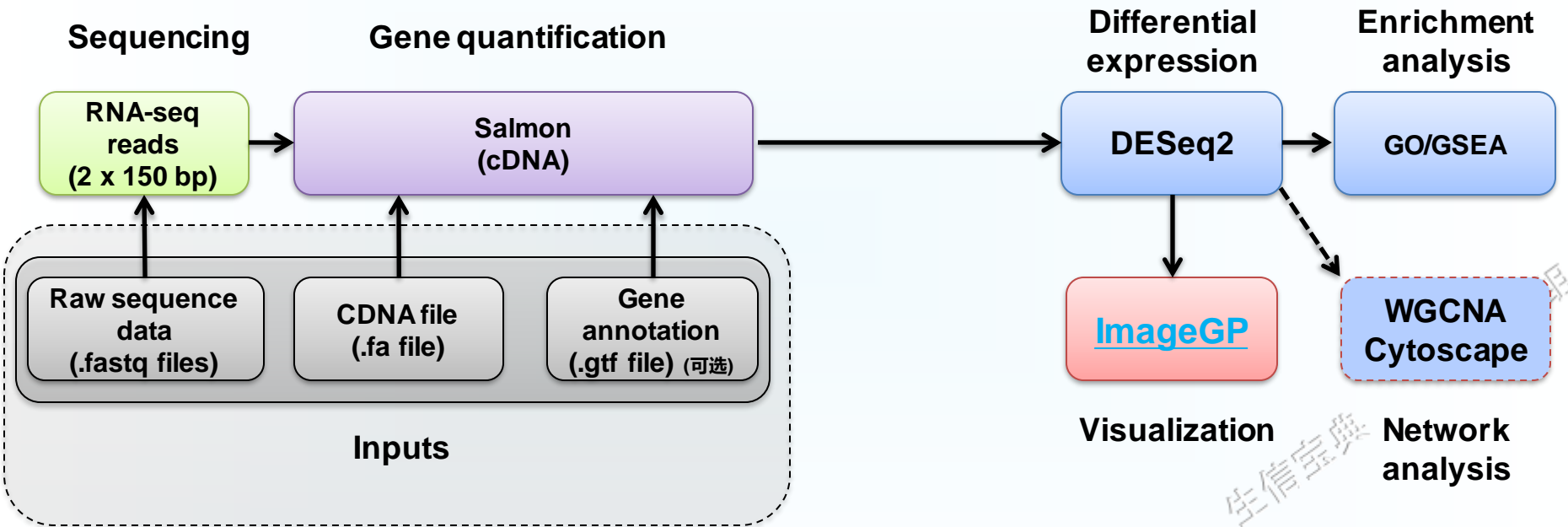
# 四种常规的转录组分析策略



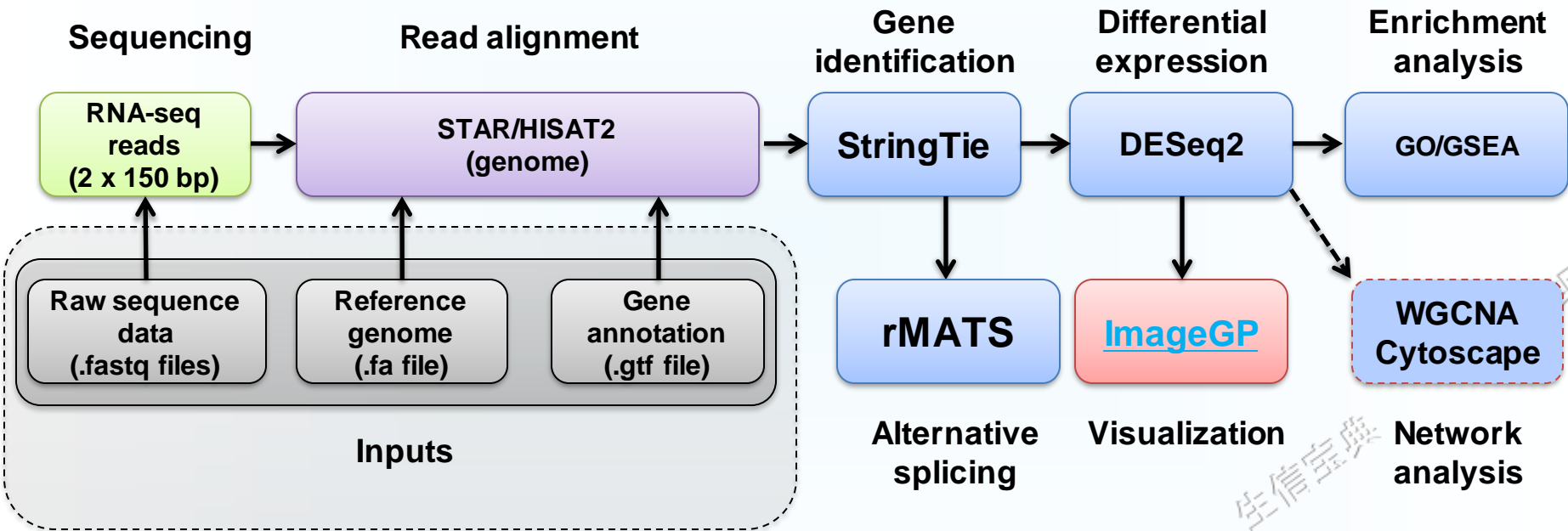
(d) 不基于比对,  
直接定量流程



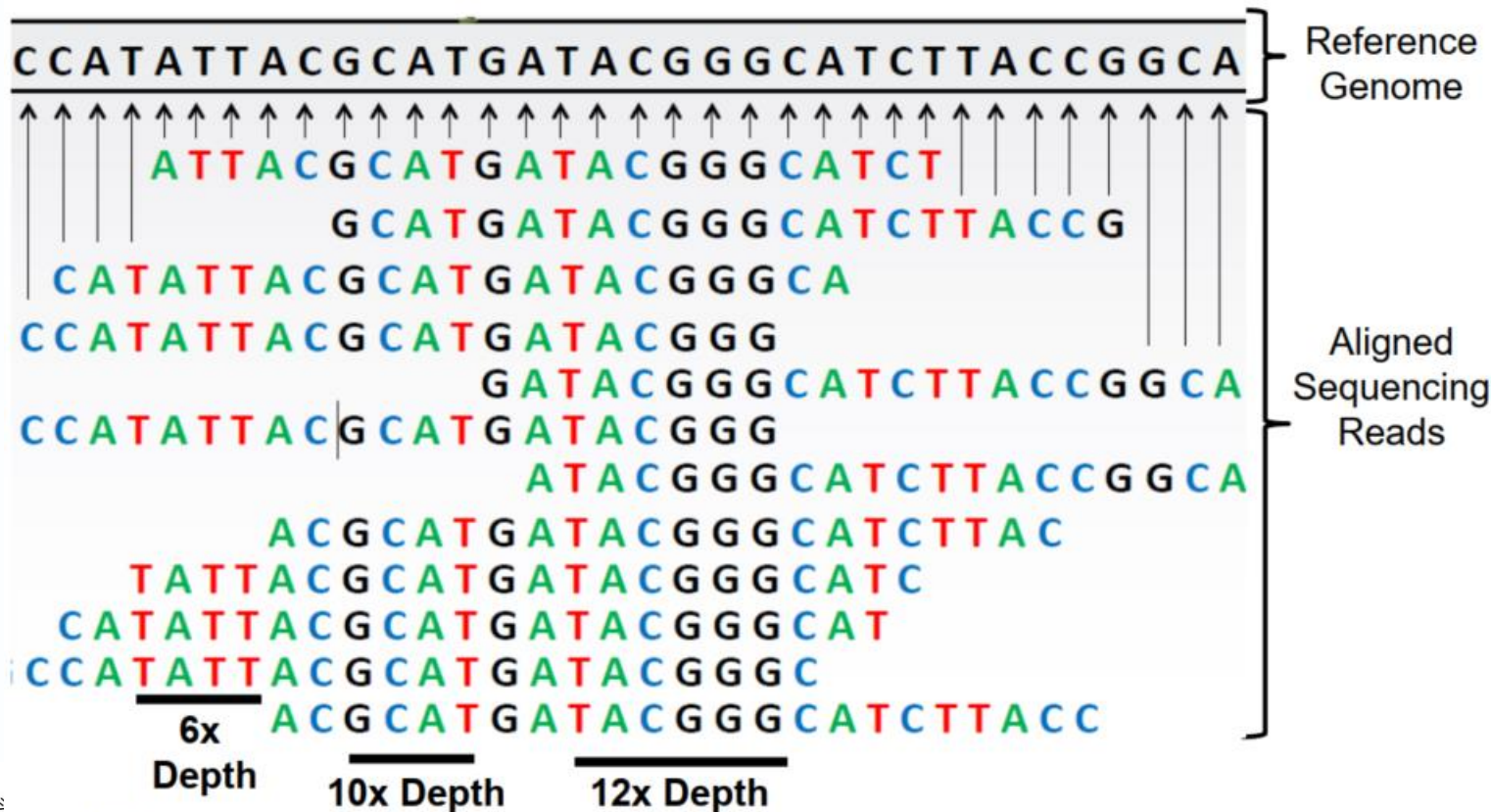
# 有参不比对转录组分析流程



# 有参比对转录组分析流程

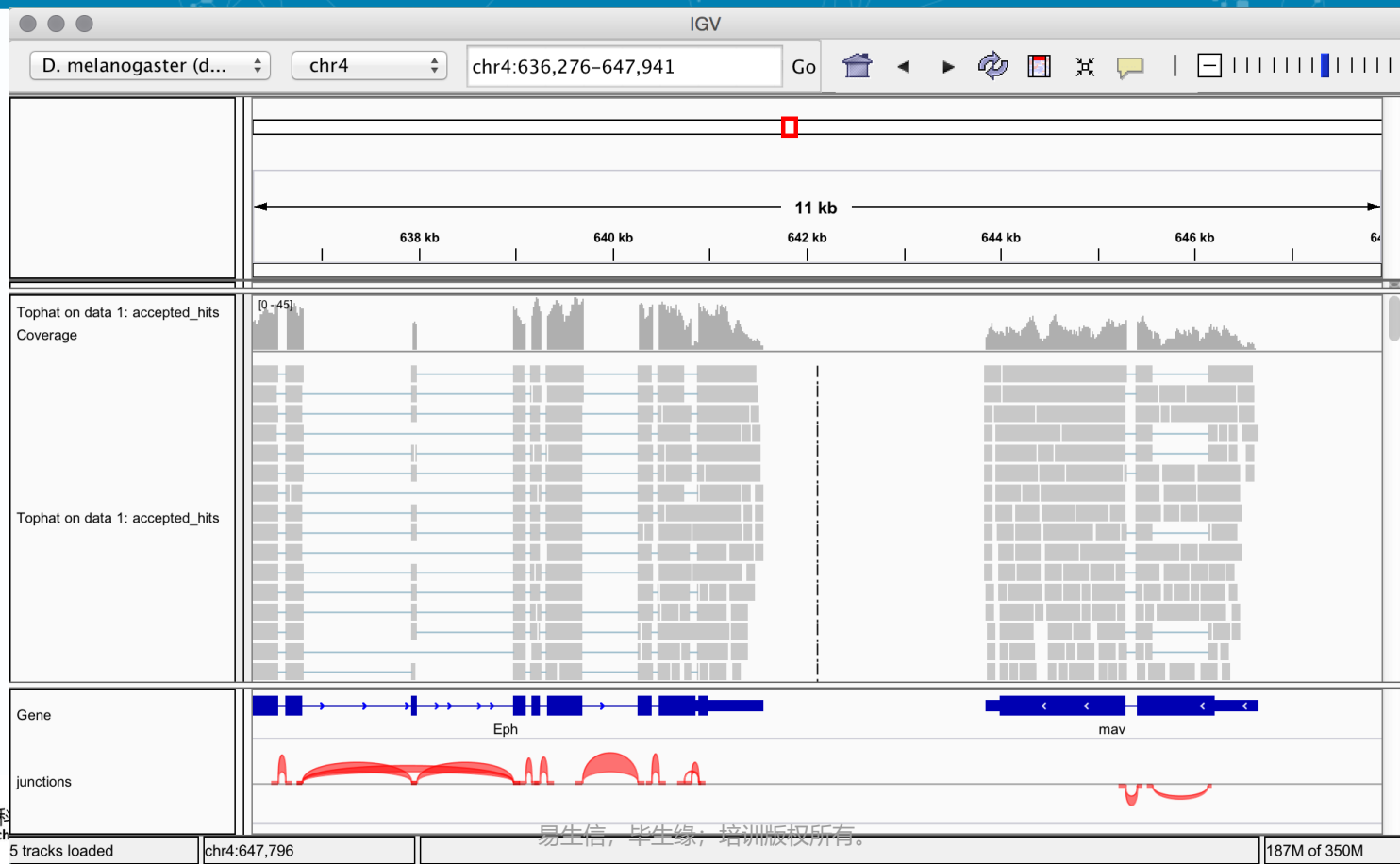


# 测序reads的长度、测序深度、reads分布



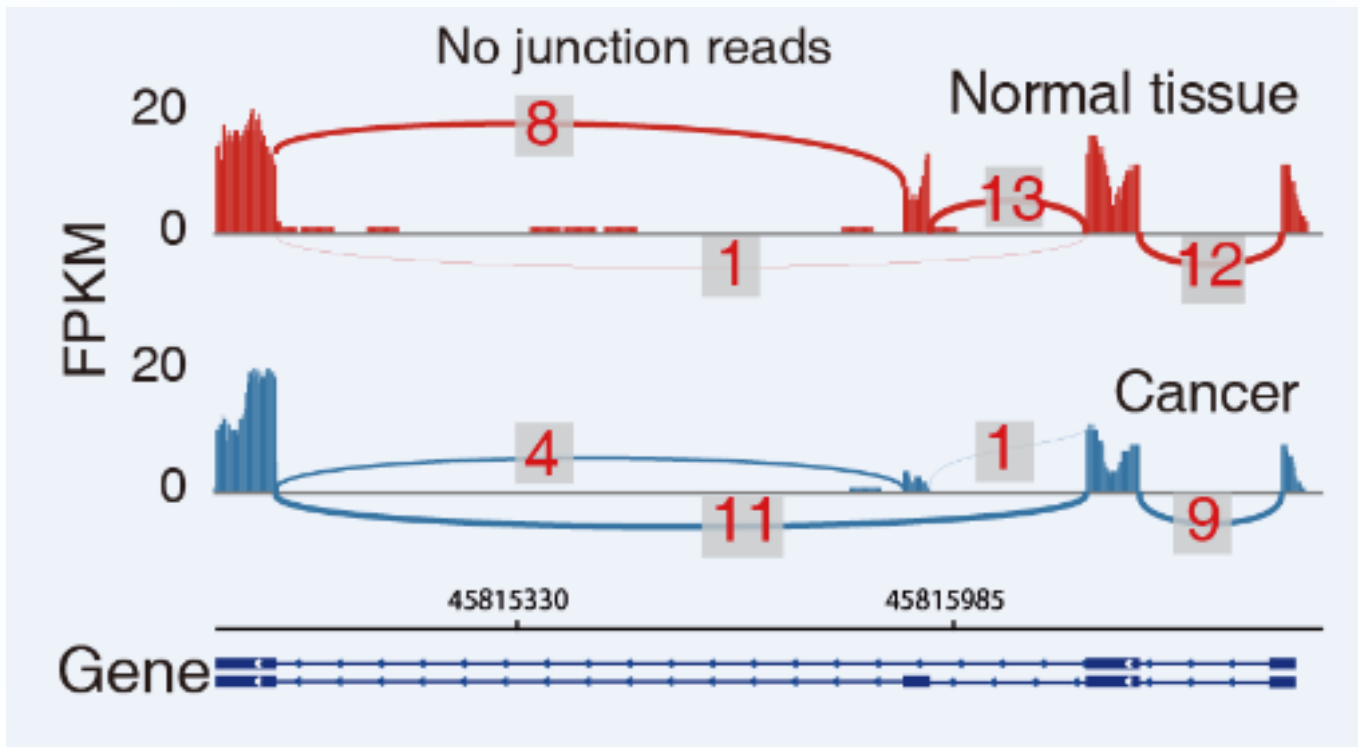
基因组

# 基因组浏览器可视化Reads分布



宏基因组

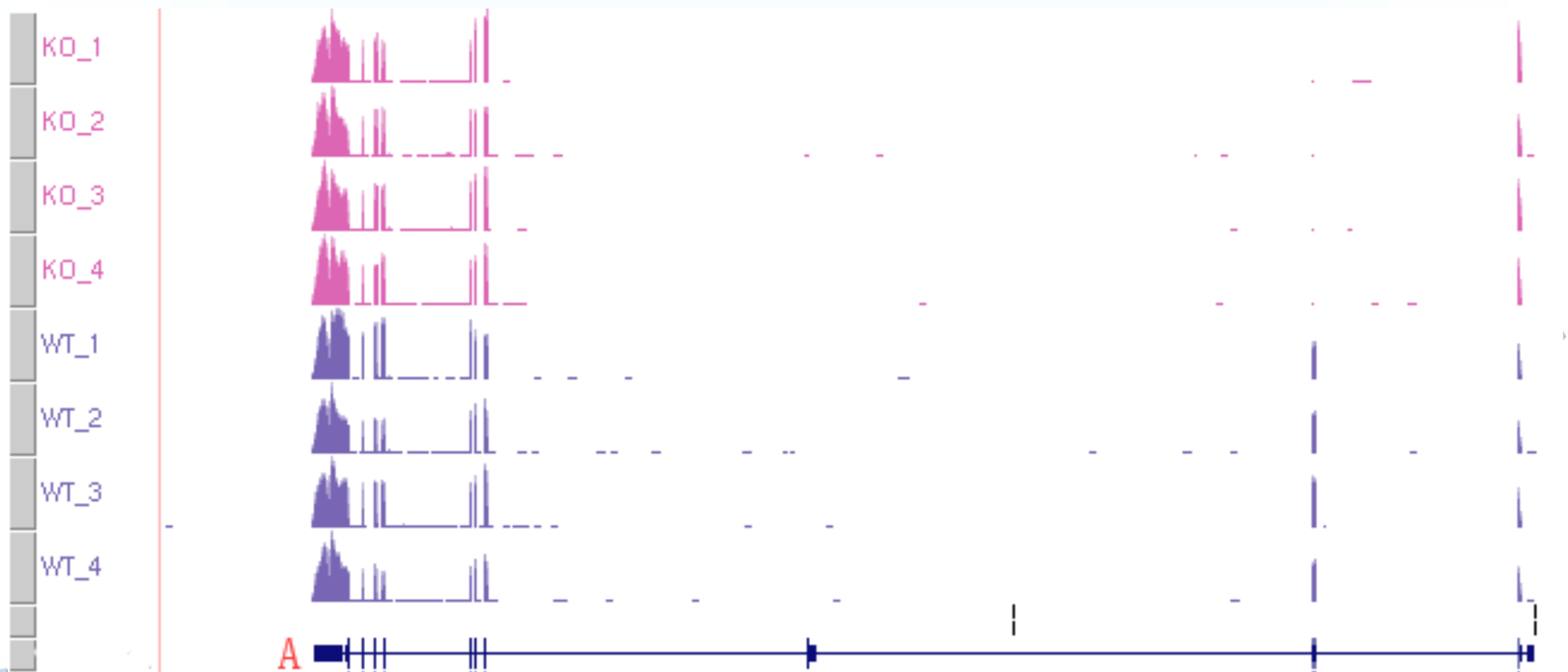
# 基因组浏览器可视化Reads分布 – sashimi plot



宏基因组

11/12

# 转录组峰图的意义 - 基因敲除是否成功?





# Nature重磅综述 | 关于RNA-seq, 你想知道的都在这



nature reviews genetics

Review Article | Published: 24 July 2019

## RNA sequencing: the teenage years

Rory Stark, Marta Grzelak & James Hadfield

*Nature Reviews Genetics* **20**, 631–656(2019) | Cite this article

**47k** Accesses | **61** Citations | **373** Altmetric | Metrics

宏基因组

算

易生信





扫码关注生信宝典，学习更多生信知识



扫码关注宏基因组，获取专业学习资料

# 易生信，没有难学的生信知识

宏基因组