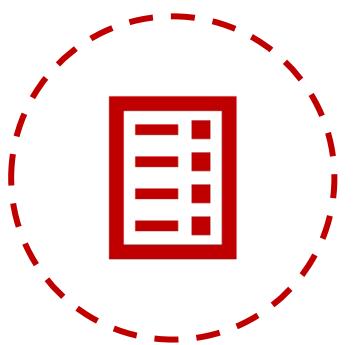


PacBio测序和三代全长转录组

王鹏

科技服务技术支持部



01

PacBio测序介绍

02

全长转录组产品

03

研究领域应用

04

送样和建库策略

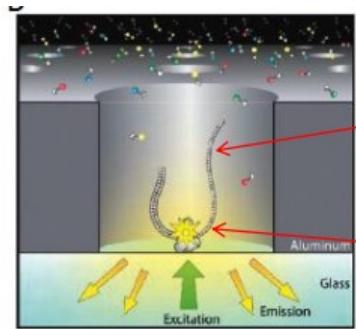
1



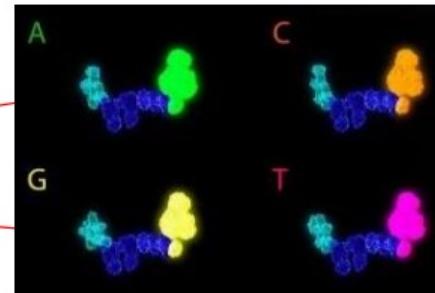
PacBio测序介绍

PacBio测序原理

- ◆ SMRT 测序使用的Cell 是一张厚度为100 nm的金属片，一面带有几十上百万个直径为几十纳米的小孔，称为零模波导孔（ zero-mode waveguide , ZMW ）。
- ◆ 在纳米孔底部，锚定着测序模板（ DNA单链 ）和DNA聚合酶，同时包含着四种被不同荧光基团修饰的dNTP。由于每次添加的dNTP所携带的荧光颜色是不同的，在激光的激发下可以发出不同的荧光，根据散射出的荧光信号可以判断添加的碱基类型。
- ◆ 激光从ZMW的下方进入，由于ZMW的直径小于激光的波长，检测激光会被限制在纳米孔内部，不会进入小孔上方的溶液区，干扰临近ZMW的测序；被激发的荧光也只会从ZMW下方的玻璃散发，被检测器检测。



零模波导孔（ZMW）



磷酸化核苷酸



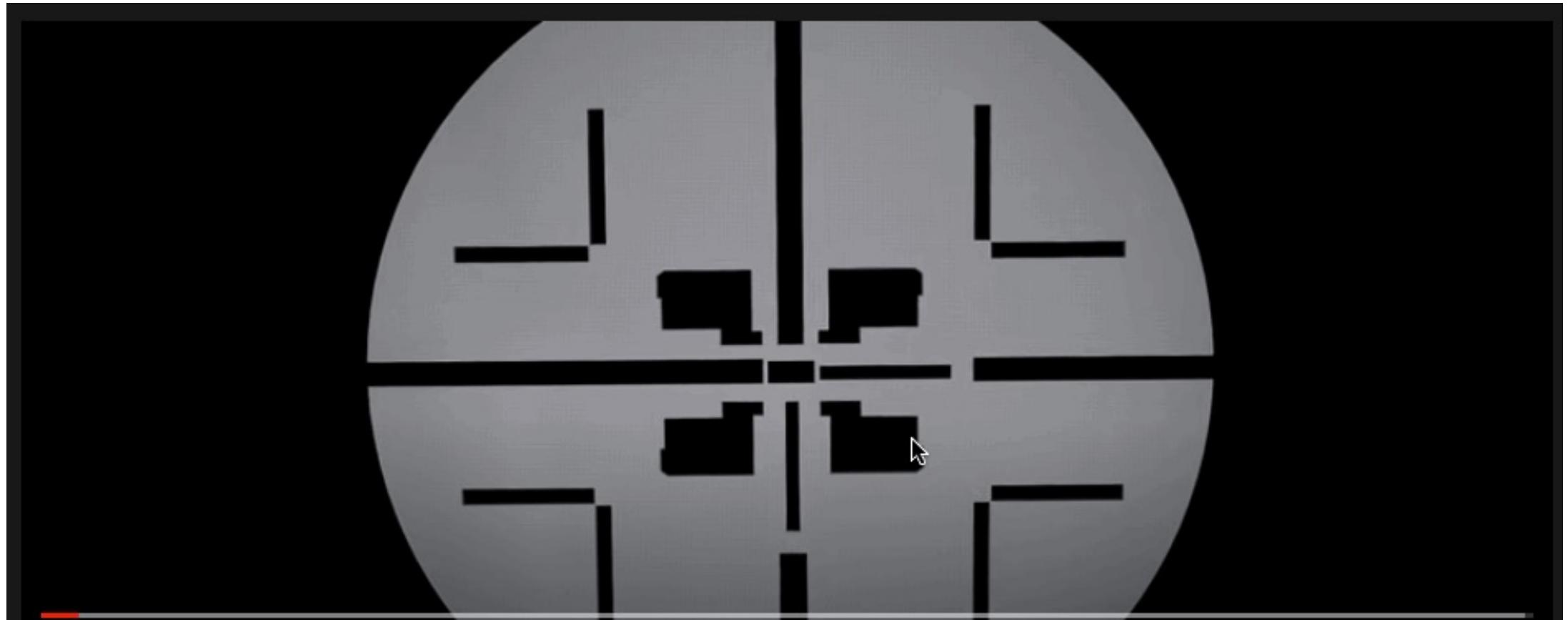
SMRT Cell芯片 8M ZMW



PacBio Sequel II

PacBio测序原理

SMRT® Technology Overview



PacBio两种测序模式

1、Continuous Long Read Sequencing (CLR)

特点：插入片段长（20/40/60kb），测序时间短（默认15h）



2、Circular Consensus Sequencing (CCS)

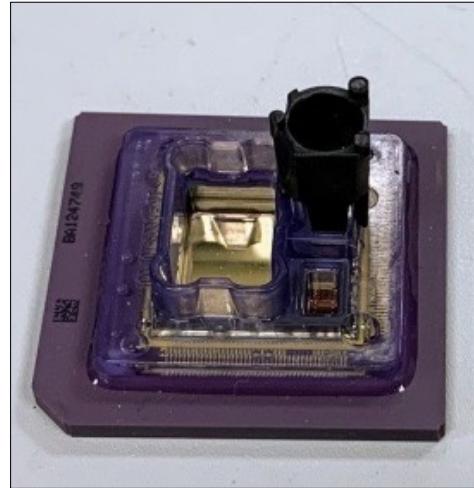
特点：插入片段短（10-15kb），测序时间长（默认30h）



平台升级Sequel→Sequel II



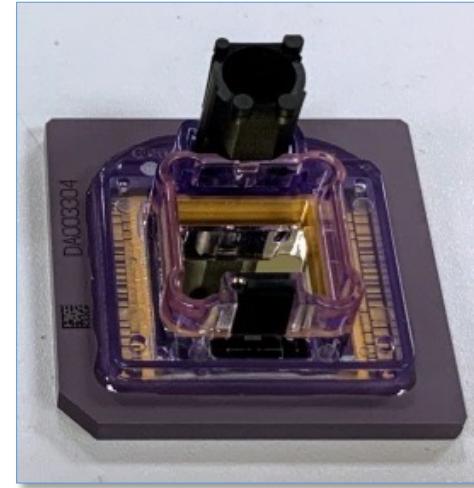
Sequel System



1 million ZMWs
SMRT Cell 1M

- 8× increase in yield
- Reduced project time
- Lower cost
- Equivalent performance

Sequel II System

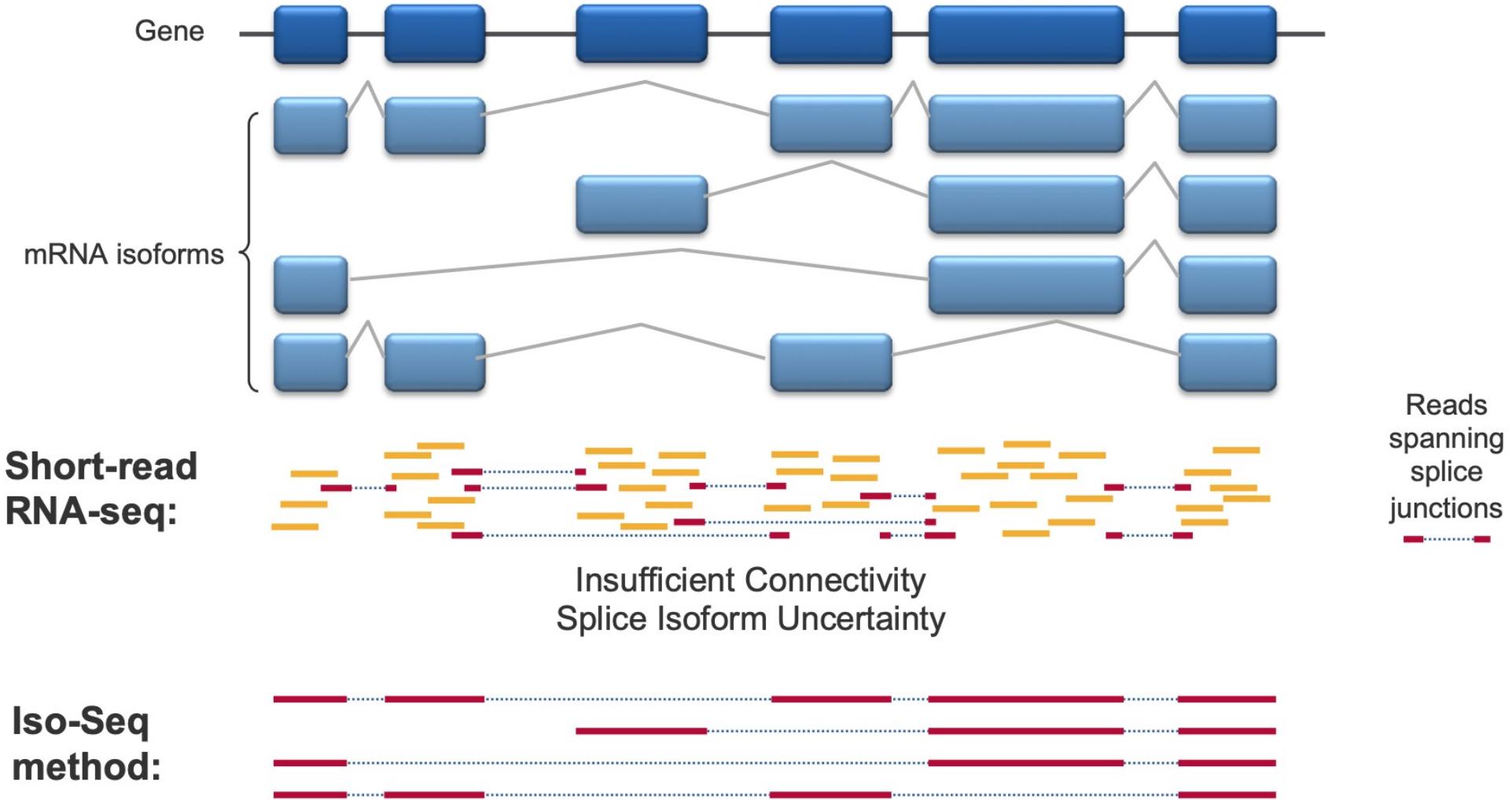


**8 million ZMWs
SMRT Cell 8M**

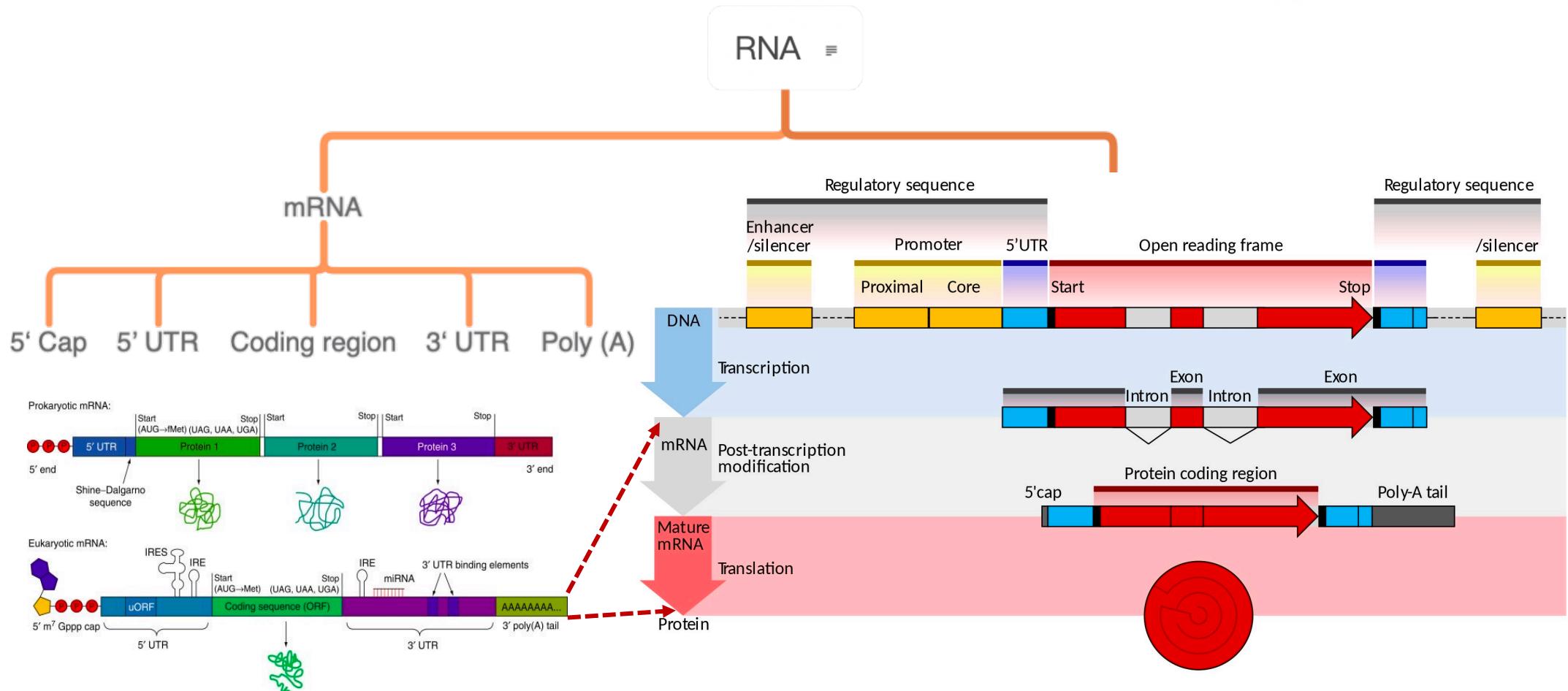
2 |||

全长转录组

为什么全长转录组

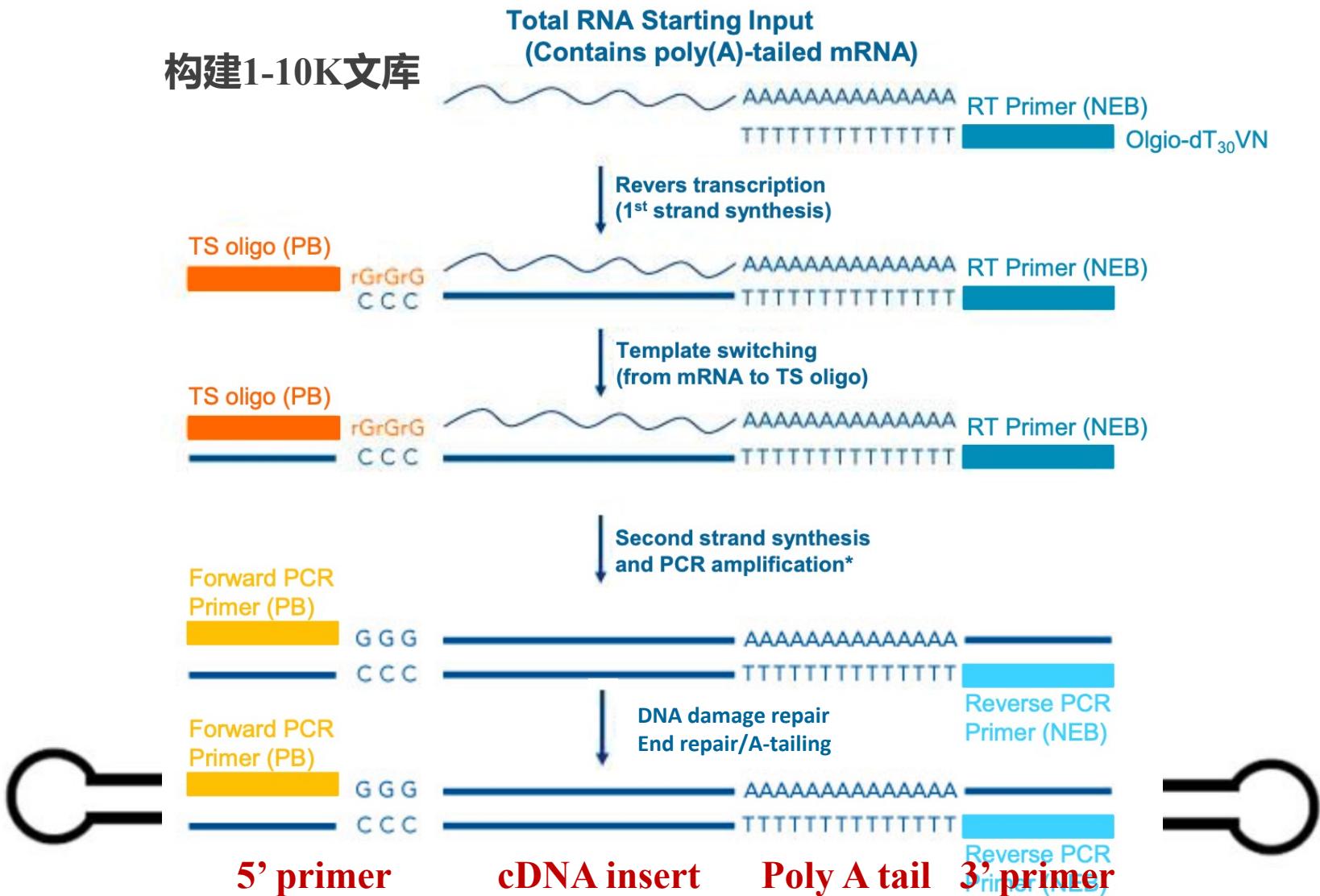


RNA分类

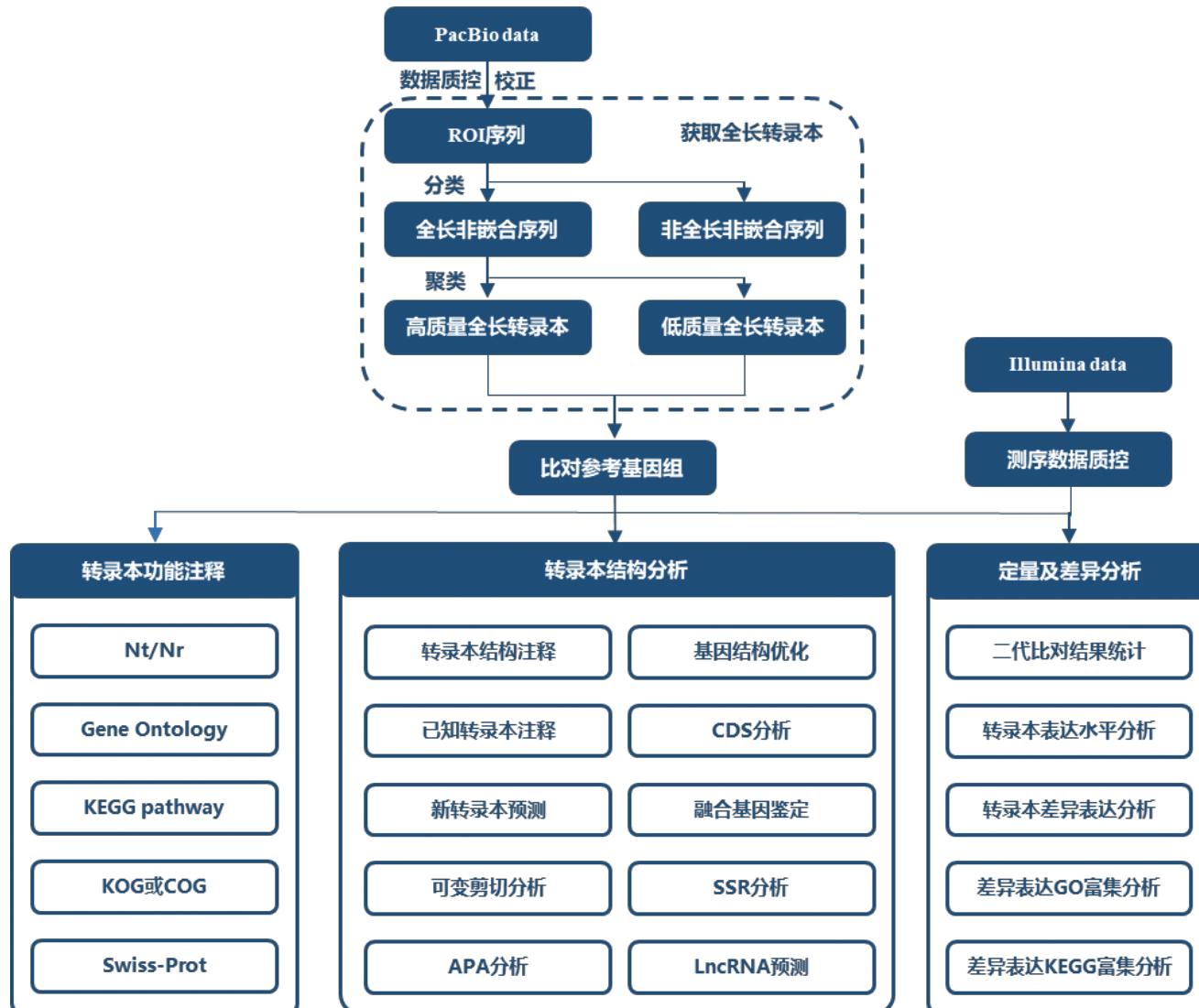


$$\text{Total RNA} = \text{mRNA (5\%)} + \text{rRNA (80\%)} + \text{other RNA (15\%)}$$

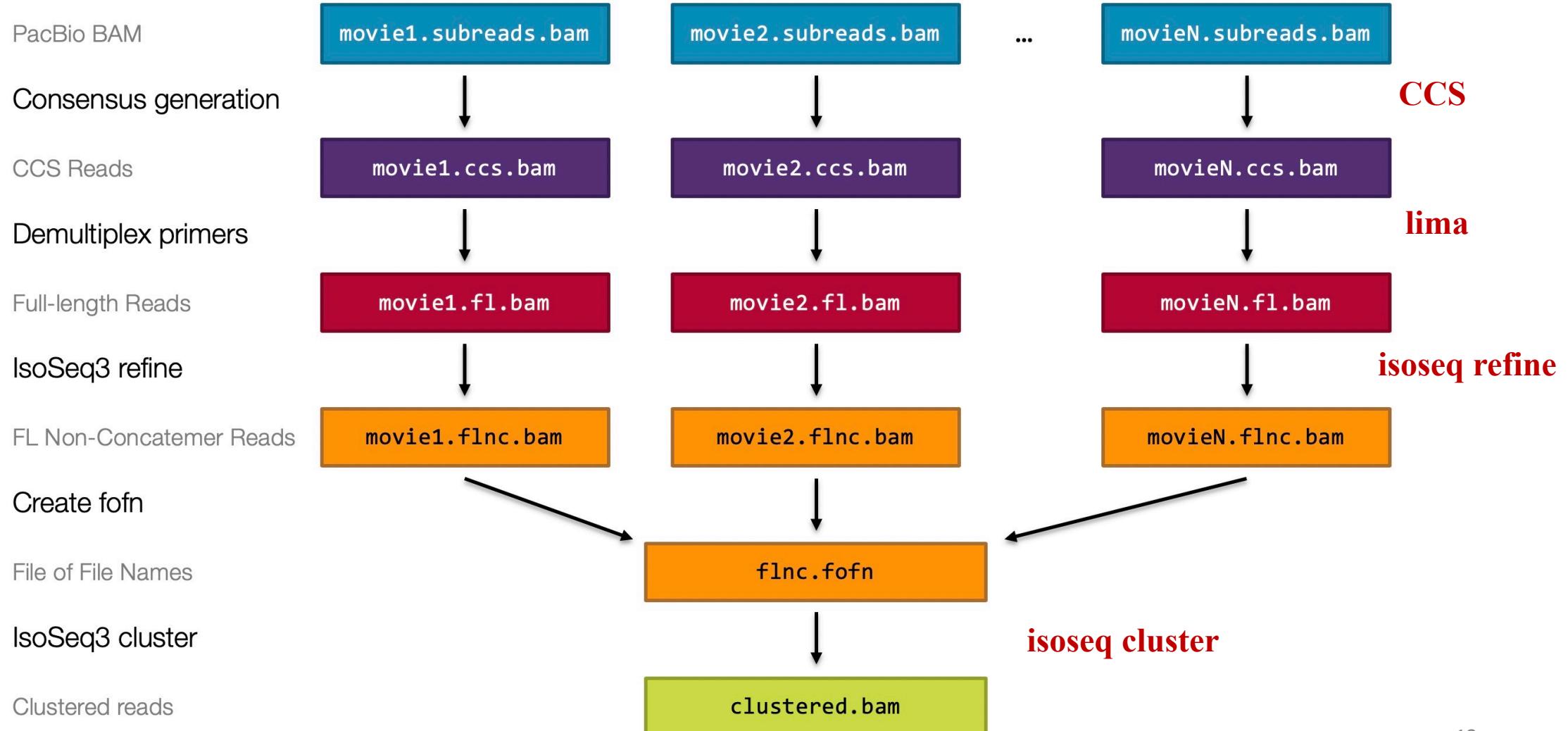
全长转录组-建库测序



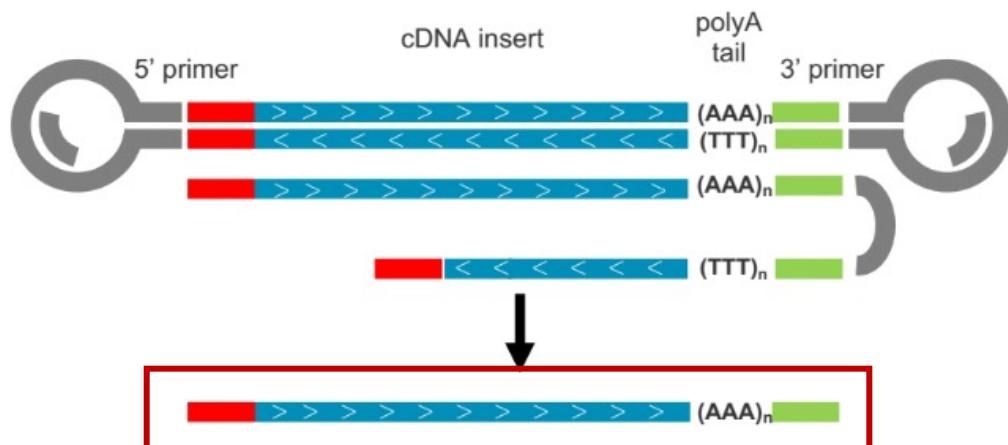
生信分析流程



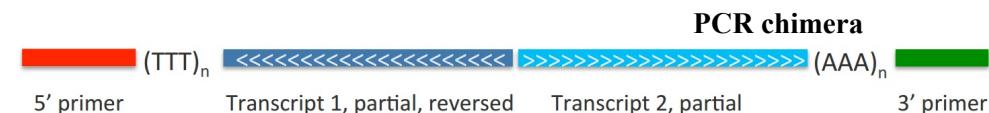
转录本序列处理流程



转录本序列处理



为了提高Reads利用率，针对iso-seq提出了reads of insert



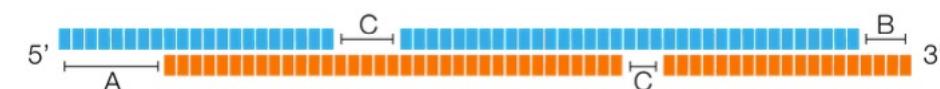
Non-full length reads (NFL) + Full length reads (FL)

| Cause | Outcome | Detection |
|-----------------------------|--------------------------------------|------------------------------------------|
| SMRT® adaptor concentration | Primer-ligated cDNA form concatemers | cDNA primer in the middle |
| PCR amplification | Random fusion of ligated transcripts | Single read maps to different loci/genes |



Two Full-Length reads are considered the same isoform if they are:

- A. <100 bp difference in 5' start
- B. <30 bp difference in 3' end
- C. <10 bp in internal gap (exon), no limit on the number of gaps

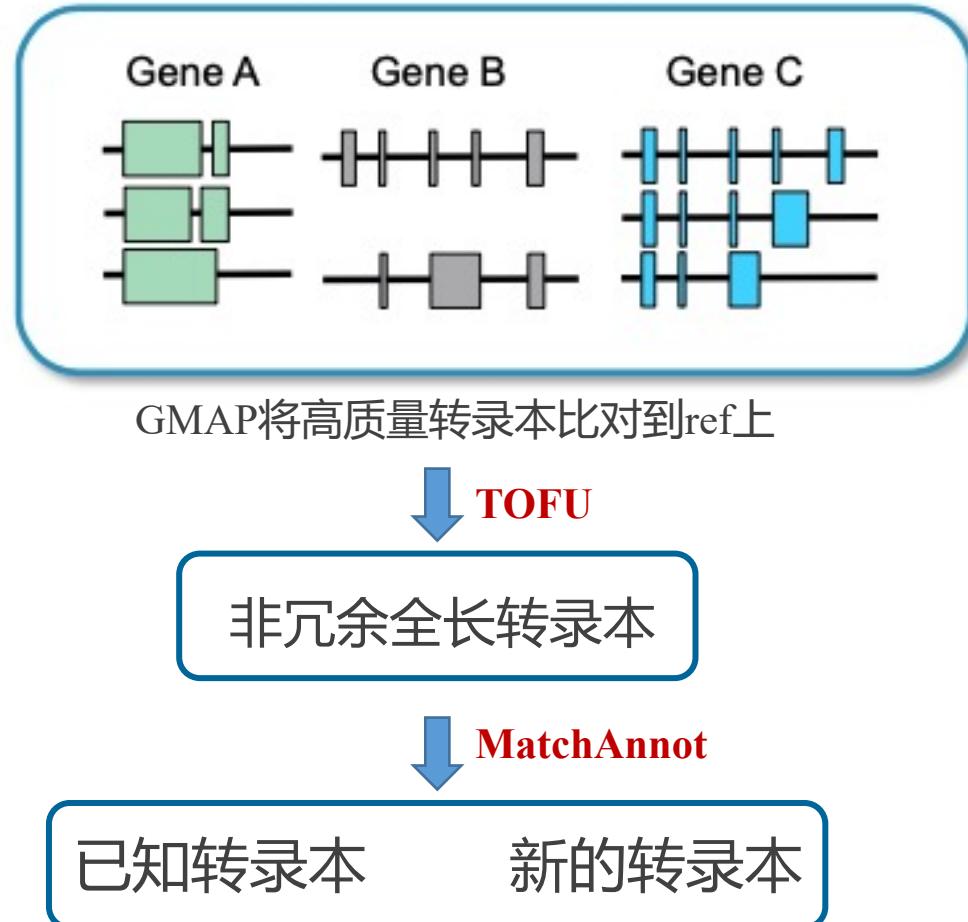


聚类获得一致性全长转录本序列

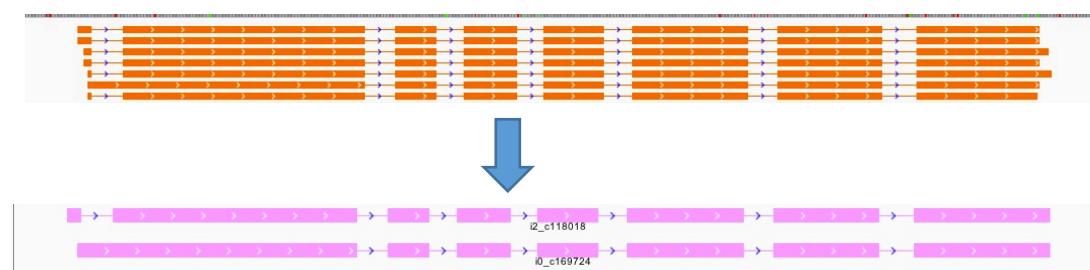
转录本序列处理

| | ccs | lima | refine | cluster |
|----------|-------------------------------------|-------------------------------------|---------------------------------------|-----------------------------------------------------|
| Input | subreads | ROI reads Barcoded primer | FL Reads | FL Reads |
| Output | ROI reads (ccs) | FL reads/sample (No primer) | FL Reads (No polyA and concatemer) | polished HQ consensus polished LQ consensus |
| Count | - | - | ≥1 read for each isoform | ~1 sequence for each isoform |
| Accuracy | Intra-molecule consensus >90% | Intra-molecule consensus >90% | Intra-molecule consensus >90% | Inter-molecule consensus* HQ: ≥ 99% LQ: < 99% |

转录本注释

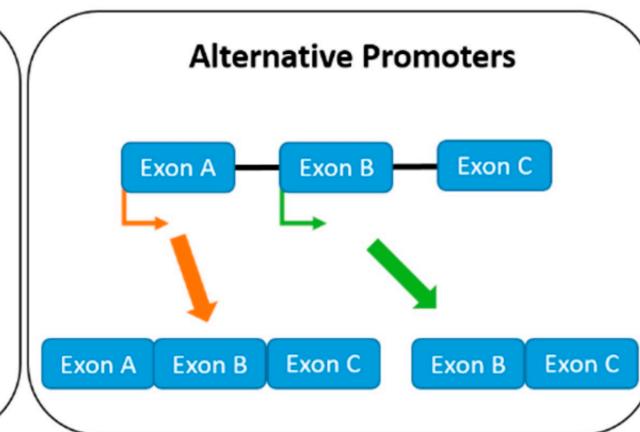
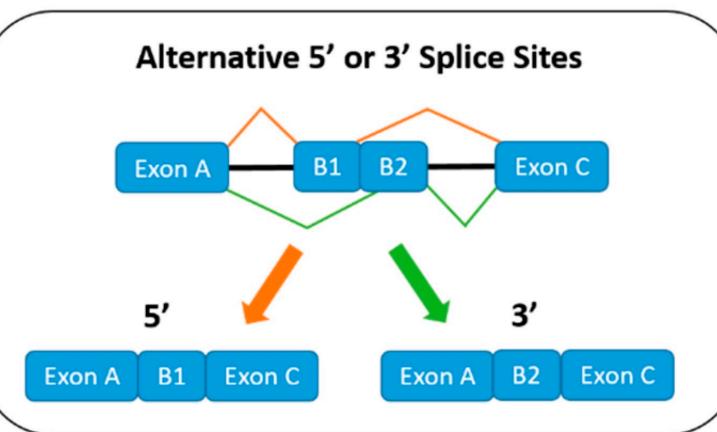
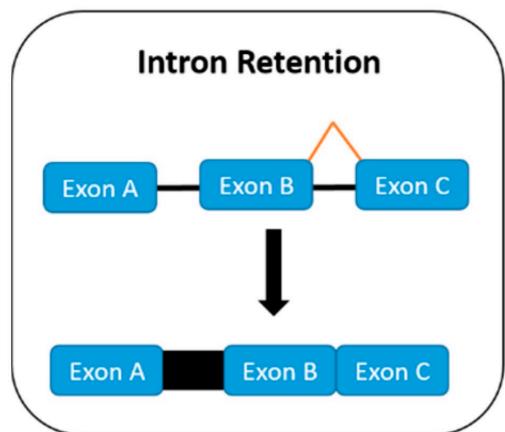
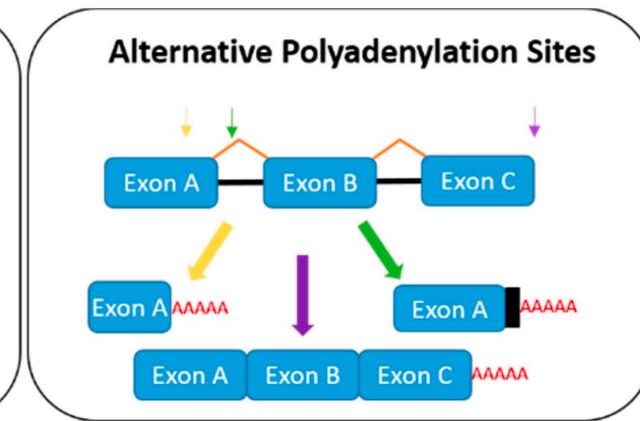
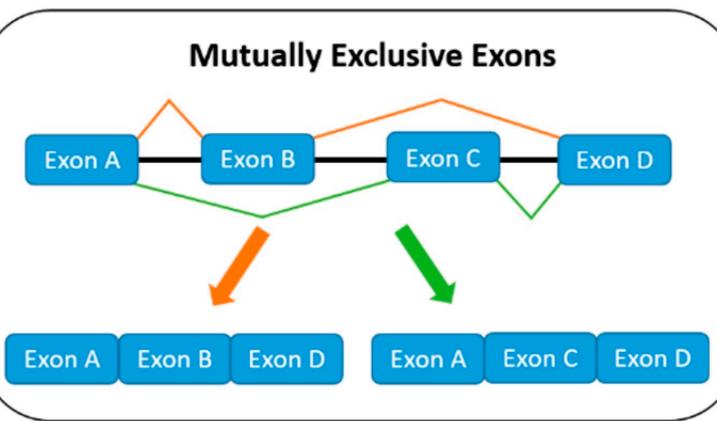
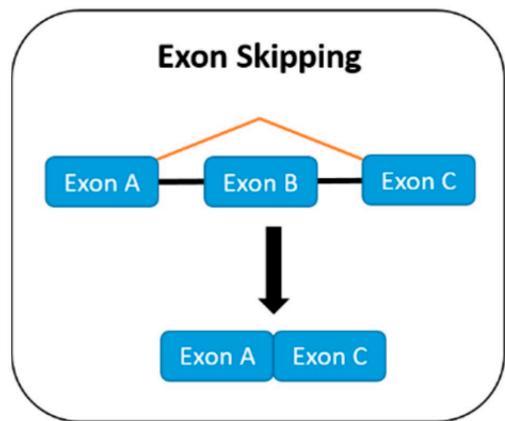


- 同一转录本的多拷贝序列可能分到不同 Cluster，会产生一些冗余的序列
- 由于 3'端存在 Poly-A 结构，可以确定 3'端比较完整，而 5'端序列容易降解，导致同一转录本的不同拷贝分到不同的 Cluster



基因结构分析

- 可变剪切 (Astalavista)



基因结构分析

● 融合基因 (TOFU)

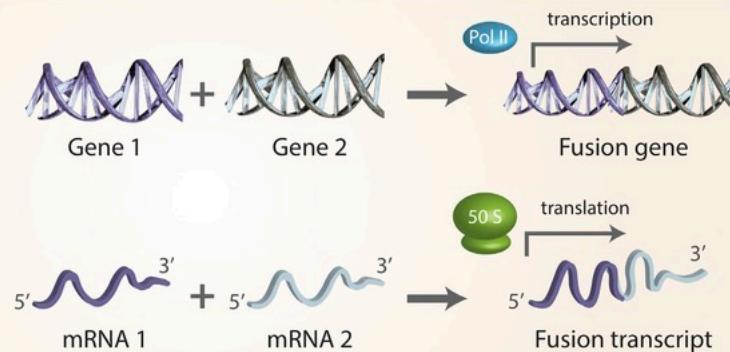
Gene fusion formation

A Fusion by structural rearrangements

Translocations, inversions, deletions and insertions

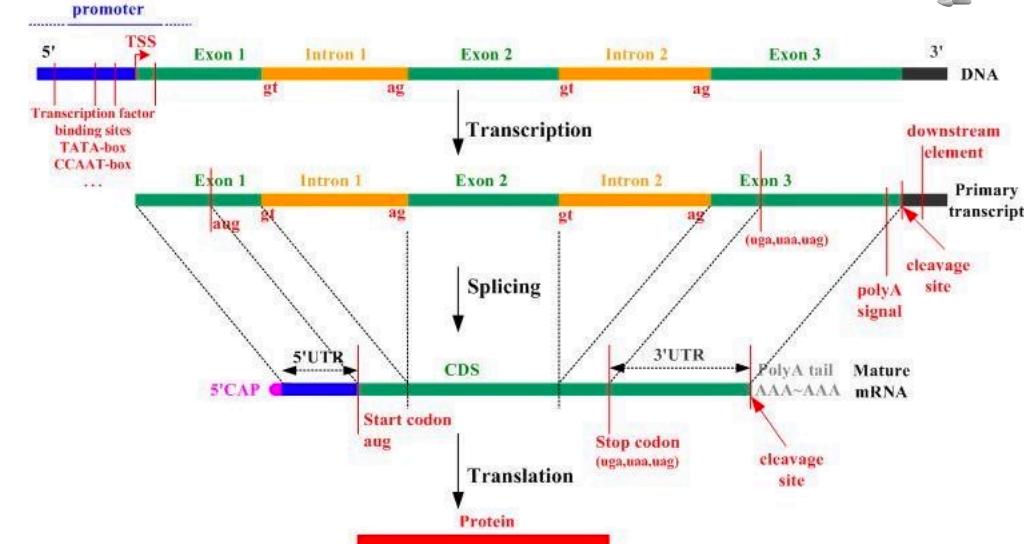
B Fusion by transcription or splicing

Transcription read-through, mRNA trans-splicing or cis-splicing



● CDS (TransDecoder)

ORF vs CDS



- 一条转录本比对到参考上的多个基因位点
- 每个基因位点至少占整条转录本的5%
- 所有位点比对的总长度必须占总长度的99%
- 两个基因位点的距离要在10Kb以上

Coding sequence : 编码序列，包含起始密码子和

终止密码子且能够编码蛋白的序列

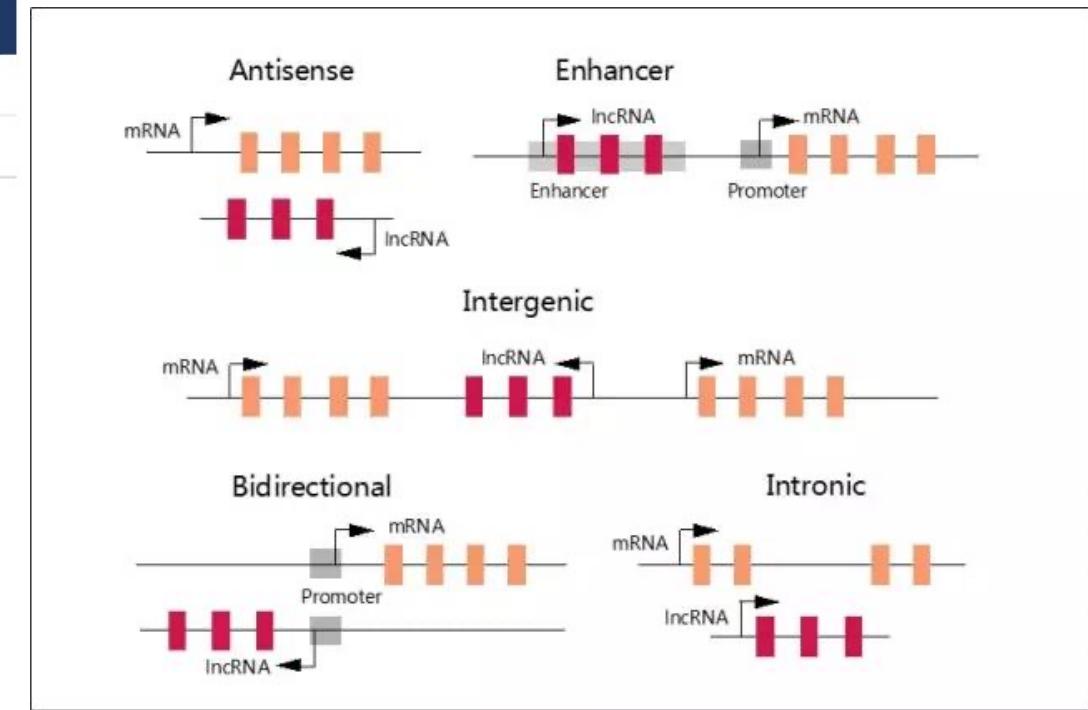
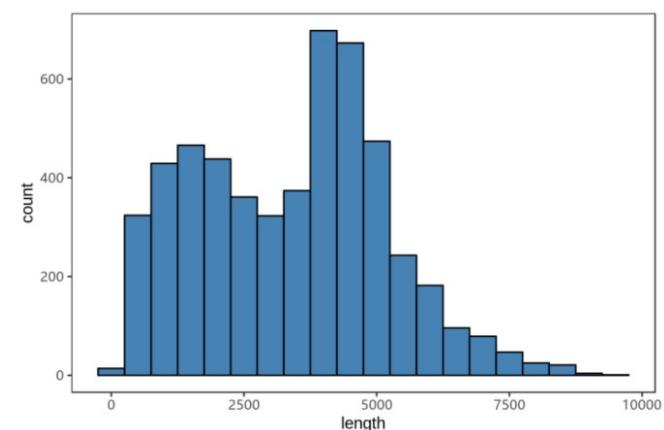
- 6种翻译框搜索编码>100氨基酸的CDS
- 每种翻译框有相应的打分（蛋白库比对）

基因结构分析

● 简单重复序列-SSR (MISA)

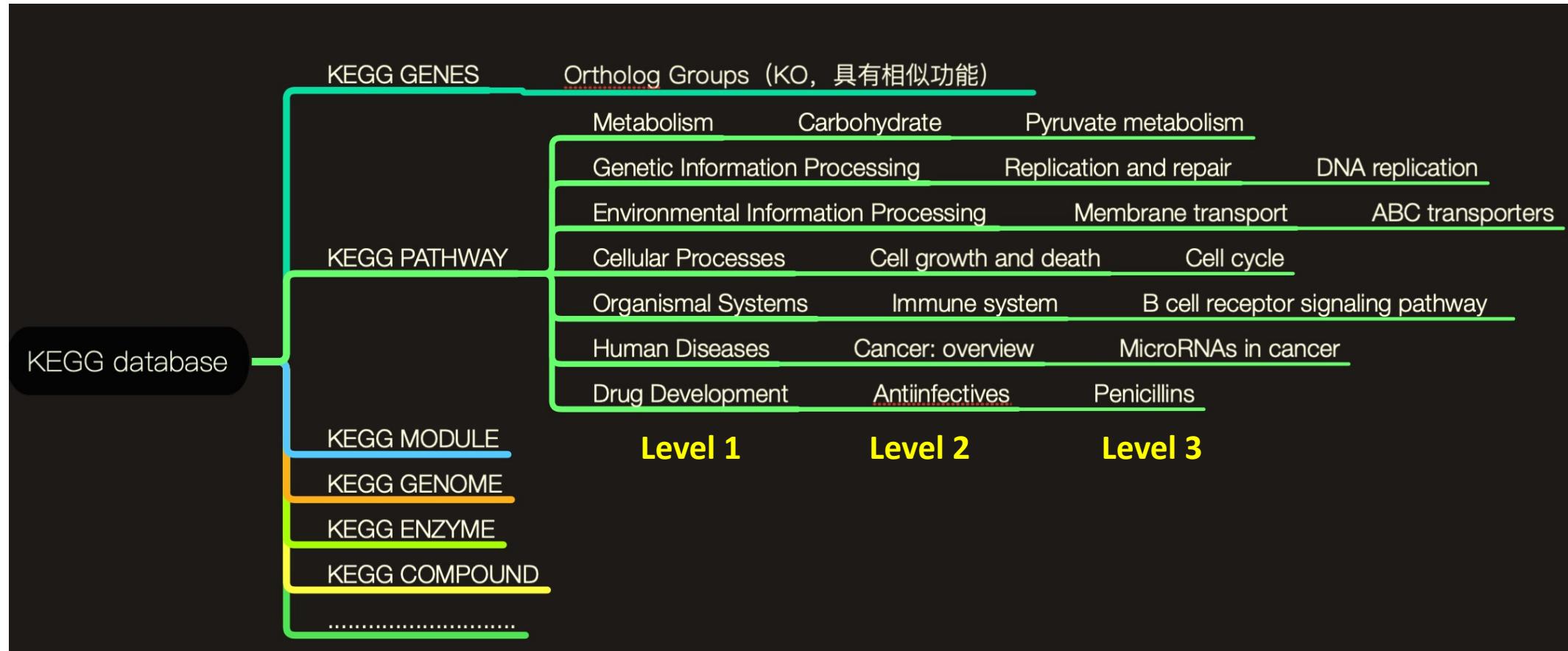
| ID | SSR nr. | SSR type | SSR | Size | Start | End |
|---------|---------|----------|-------|------|-------|-------|
| SSR_304 | 1 | p1 | (A)12 | 12 | 1,382 | 1,393 |
| SSR_337 | 1 | p1 | (T)20 | 20 | 995 | 1,014 |

● 长链非编码RNA-lncRNA (CPAT)



转录本功能注释

Kyoto Encyclopedia of Genes and Genomes



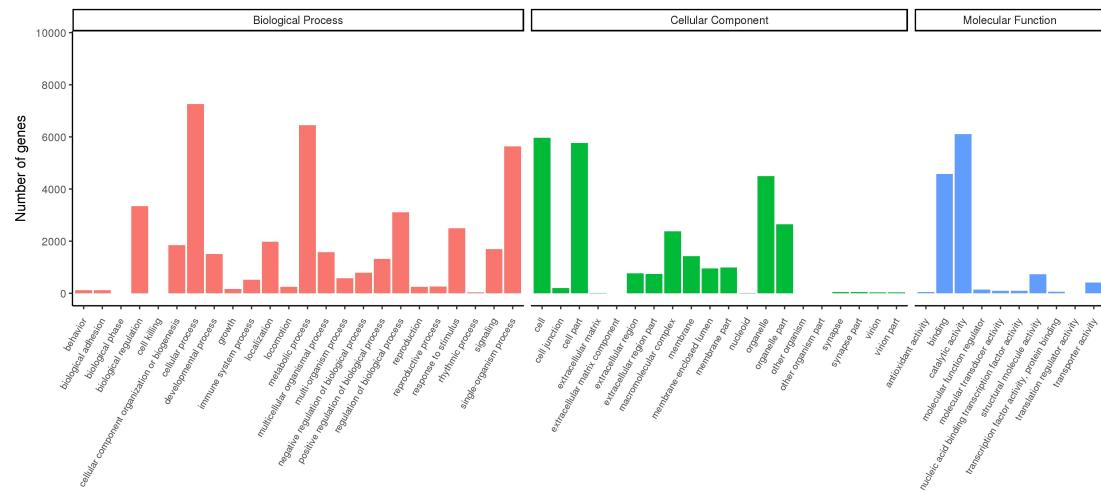
提供全面的功能注释信息，包括基因、酶、化学产物、所属物种等，并系统分类了代谢通路

转录本功能注释

GO

Gene Ontology : 相似的基因，保守的功能

- 分子功能 (Molecular Function) 【催化或转运活性】
- 细胞组分 (Cellular Component) 【线粒体、核糖体】
- 生物学过程 (Biological Process) 【DNA修复或信号转



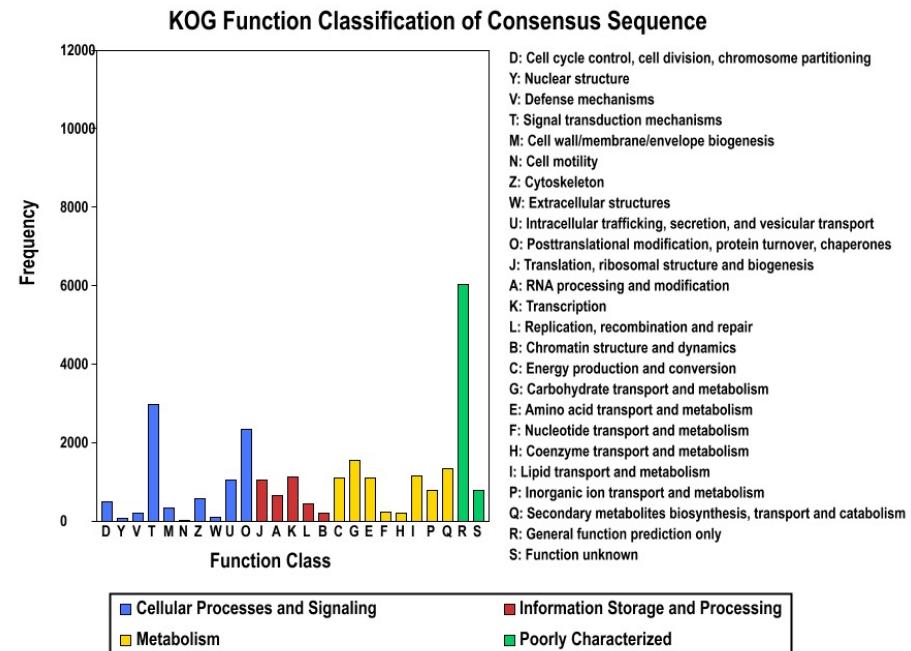
GO term
<http://geneontology.org/docs/GO-term-elements>

id: GO:0000016
name: lactase activity
ontology: molecular_function
def: "Catalysis of the reaction: lactose + H₂O=D-glucose + D-galactose." [EC:3.2.1.108]
synonym: "lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108]
synonym: "lactose galactohydrolase activity" EXACT [EC:3.2.1.108]
xref: EC:3.2.1.108
xref: MetaCyc:LACTASE-RXN
xref: Reactome:20536
is_a: GO:0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds

KOG

Eukaryotic Orthologous Groups of proteins

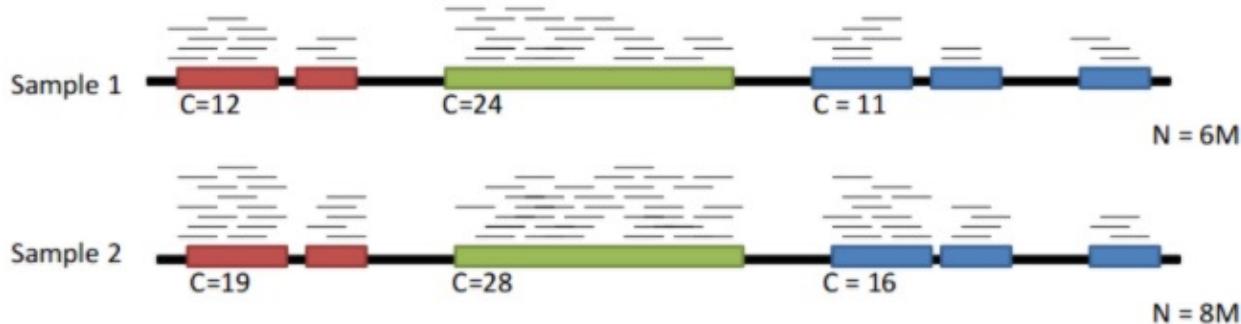
该数据库来源七个真核生物的全基因组，酿酒酵母、裂殖酵母、脑原虫、拟南芥、线虫、果蝇和人的完整基因组。共有4852个KOGs（来自于共同祖先的同源基因），包含60579个蛋白，分成26个功能单元。



转录本功能注释

| Anno_database | Annotated_number | 0<=Length<1000 | 1000<=Length<2000 | 2000<=Length<3000 | 3000<=Length<6000 | Length>=6000 |
|----------------------|------------------|----------------|-------------------|-------------------|-------------------|--------------|
| GO_annotation | 12,255 | 937 | 2,284 | 2,150 | 6,187 | 697 |
| KEGG_annotation | 4,851 | 513 | 967 | 895 | 2,191 | 285 |
| KOG_annotation | 8,947 | 709 | 1,753 | 1,634 | 4,378 | 473 |
| NR_annotation | 12,605 | 990 | 2,336 | 2,203 | 6,357 | 719 |
| NT_annotation | 12,850 | 1,060 | 2,381 | 2,222 | 6,459 | 728 |
| Swissprot_annotation | 12,250 | 927 | 2,295 | 2,162 | 6,168 | 698 |
| All_annotated | 12,850 | 1,060 | 2,381 | 2,222 | 6,459 | 728 |

转录本定量



RPKM : 适用于单端的转录组数据

FPKM : 适用于双端的转录组数据

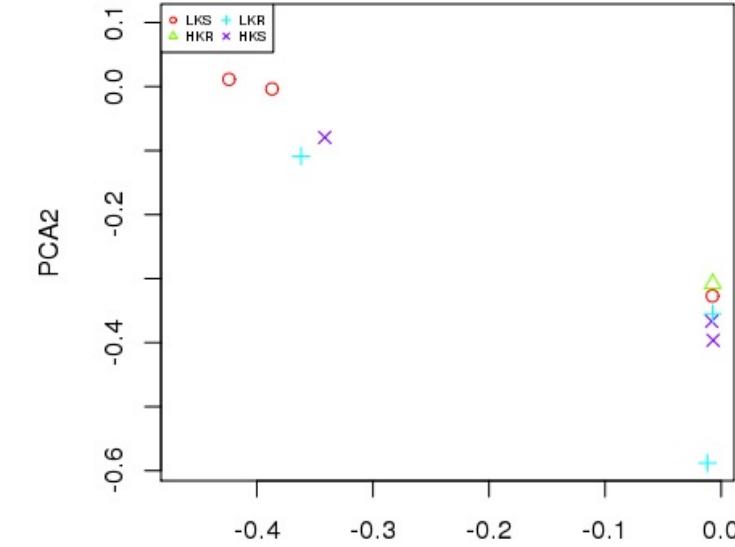
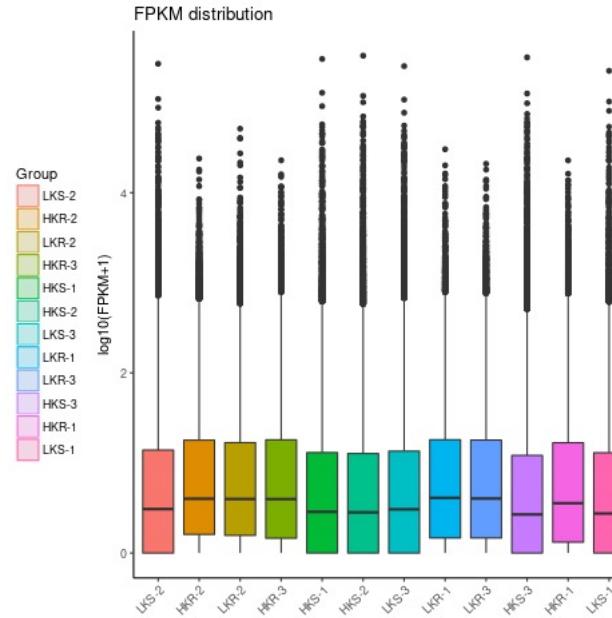
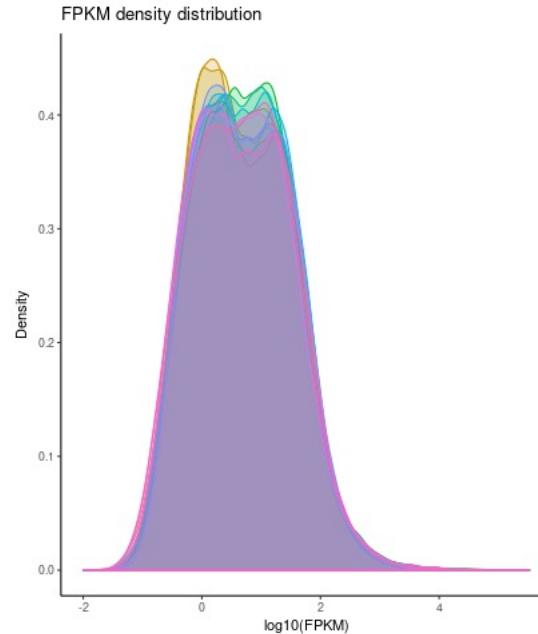
TPM : 适用于RNA和miRNA的校正

转录本的表达量是多少

- 1) RPKM : Reads Per Kilobase of exon model per Million mapped reads
- 2) FPKM : Fragments Per Kilobase of exon model per Million mapped fragments
- 3) TPM : Transcripts Per Kilobase of exon model per Million mapped reads

$$RPKM = \frac{10^6 * n_r}{L * N}$$
$$FPKM = \frac{10^6 * n_f}{L * N}$$
$$TPM = \frac{\frac{n_i}{L_i} * 10^6}{\sum_{i=1}^N \frac{n_i}{L_i}}$$

表达水平分析



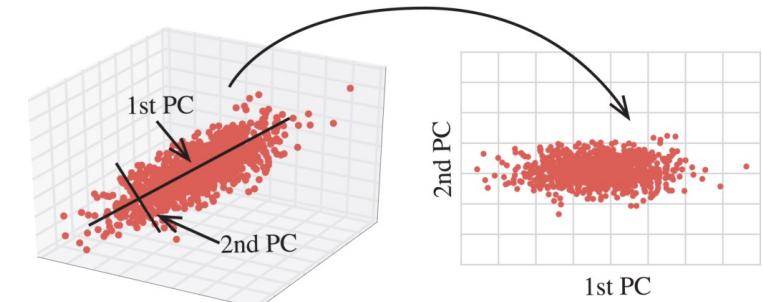
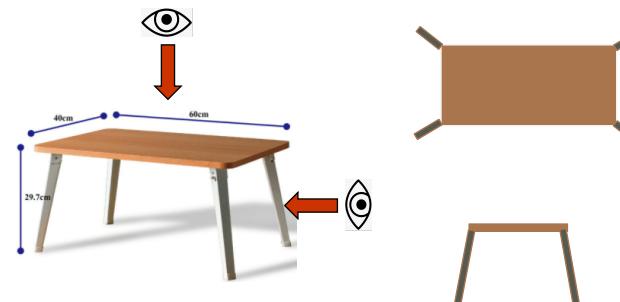
? High dimension

Complexity

Simplicity

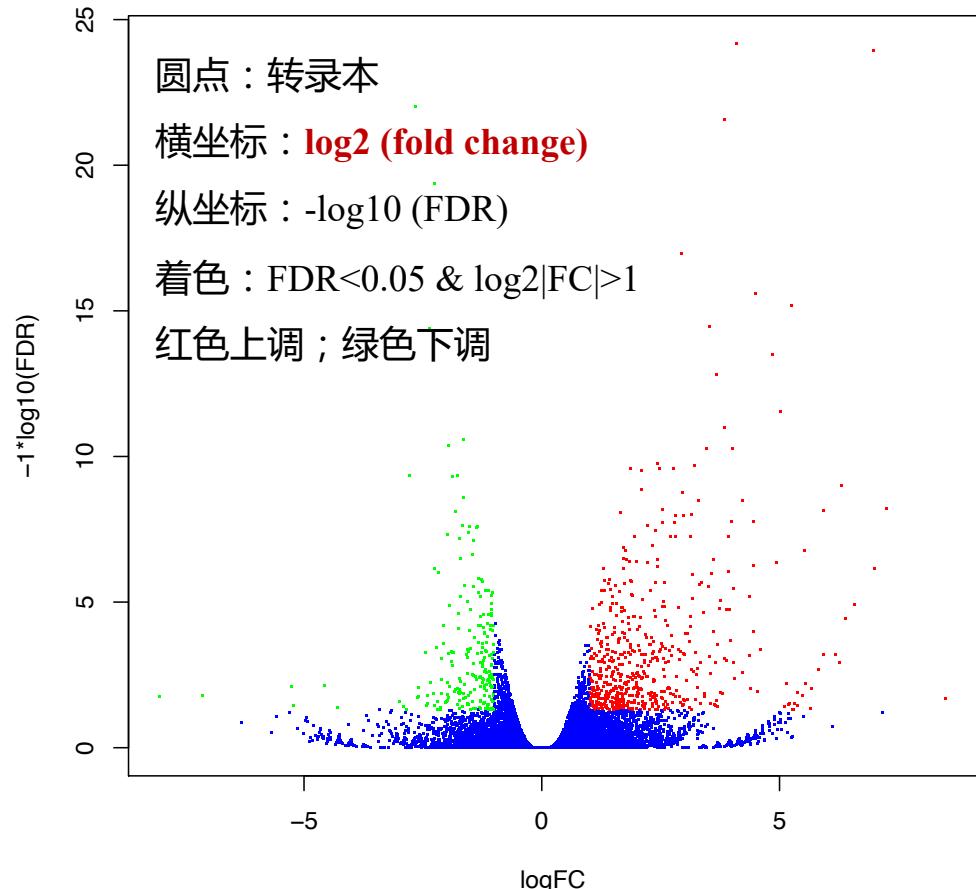


? Low dimension



Principal component analysis : 就是从最佳的角度看数据

差异表达分析

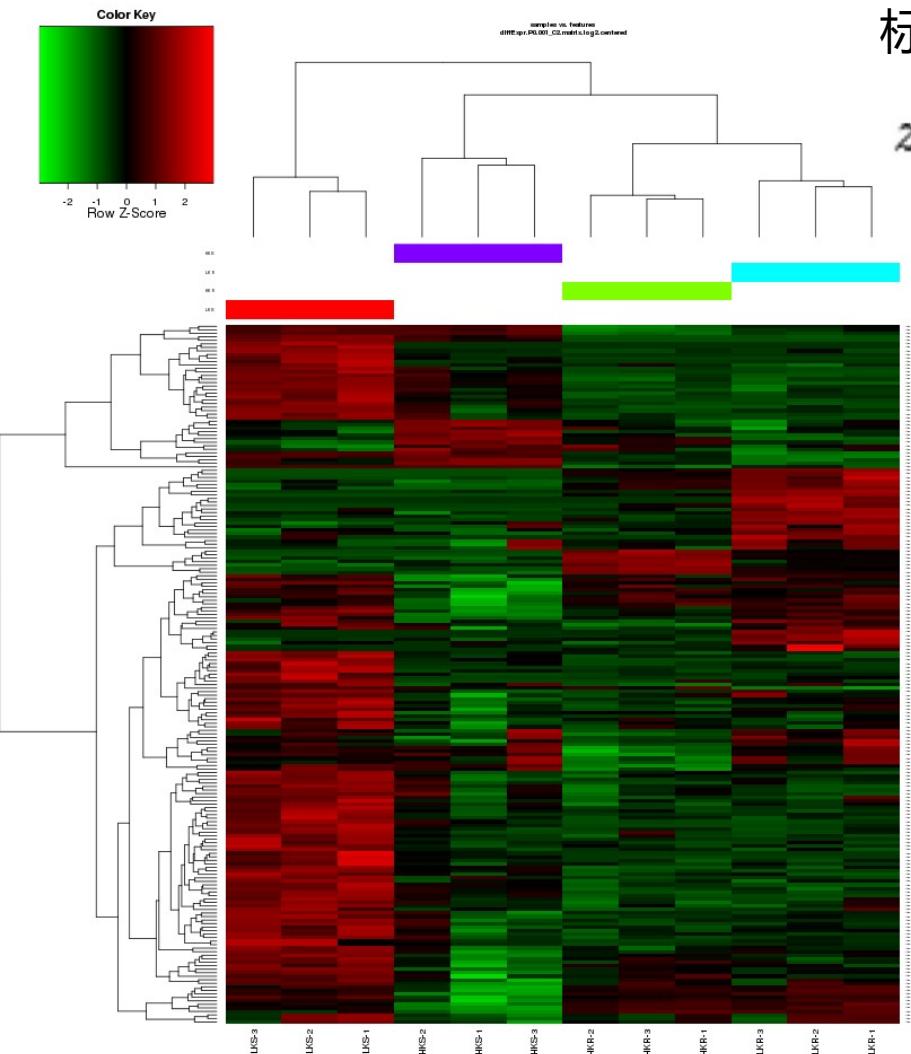


Fold change为什么用常用log转化？



$$\begin{array}{c} (0, 1) \\ (-\infty, 0) \end{array}$$

$$(1, +\infty)$$

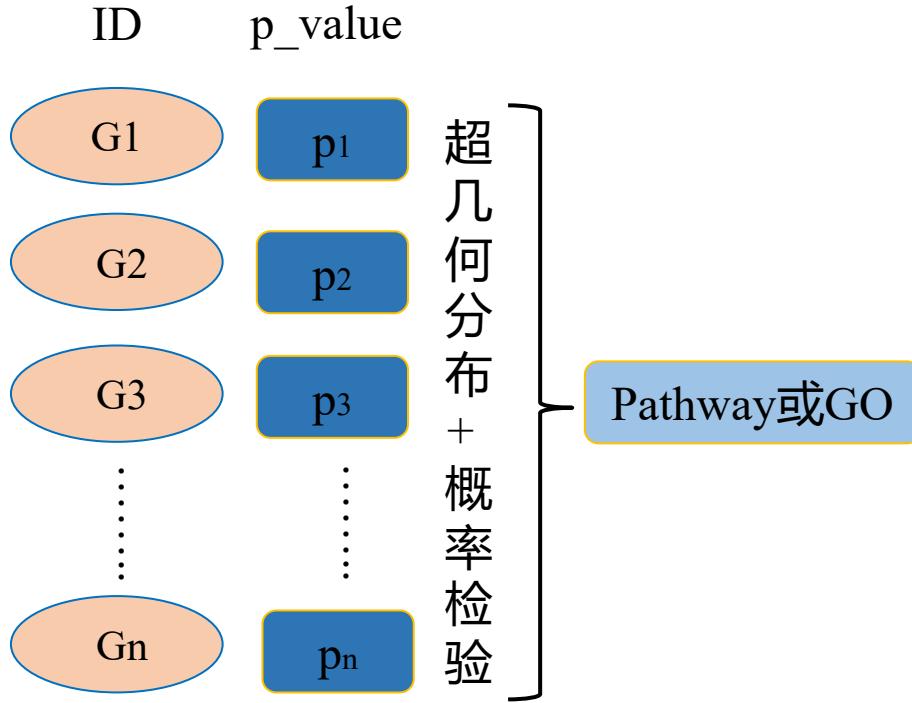


$$z = \frac{x - \mu}{\sigma}$$



红绿色
行聚类
列聚类
数据条

富集分析



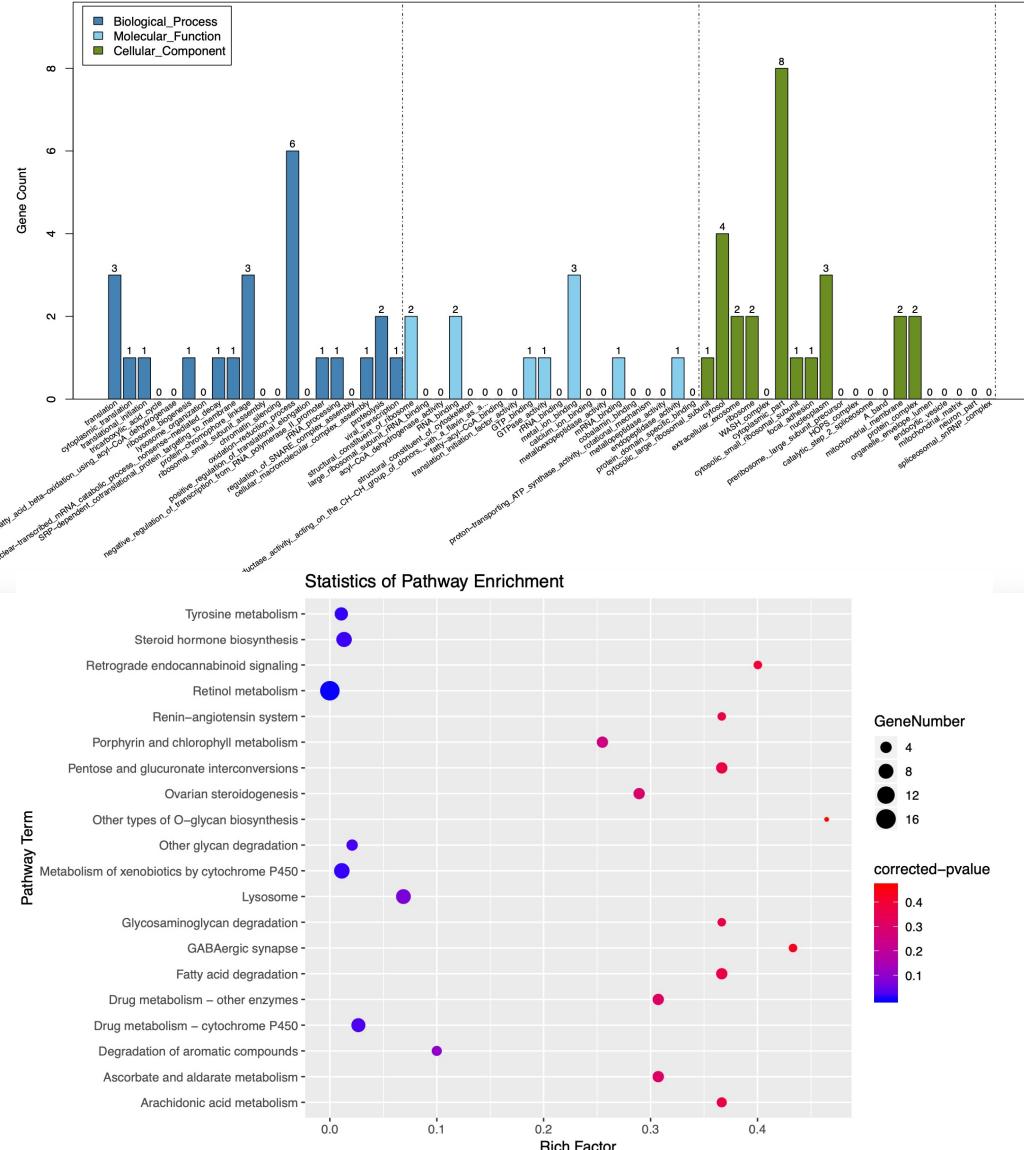
信号：在某个Pathway或GO term中的差异基因比例

背景：所有差异基因数占整个注释基因数目的比例

通俗举例，一个省区是否显著富裕？

信号：富裕人数占该省总人口的比例

背景：国家富裕人数占总人口的比例



3



研究领域应用

应用领域

◆ 1. 完善转录组（有参）

侧重于解决结构和功能分析，如AS、APA、lncRNA、预测新基因、新转录本异构体等。

◆ 2. 功能性研究（有/无参）

将差异表达基因富集到具体代谢通路，对其功能进行研究。

◆ 3. 差异转录组（有/无参）

研究胁迫处理同一物种不同个体或选择表型差异个体进行转录本差异研究，以期了解动植物的适应和驯化的生理机制。

◆ 4. 动态转录组（有/无参）

研究转录组随时间，发育的动态变化，追寻遗传信息表达的变化机理。

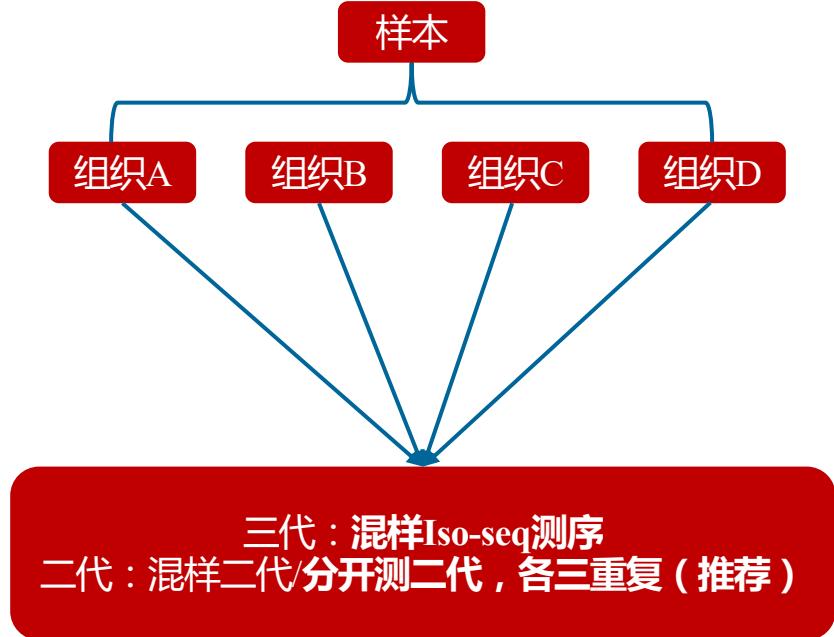
◆ 5. 物种比较转录组（有/无参）

通过Iso-seq技术手段研究少数近缘物种间的亲缘关系以及不同物种或亚种间mRNA序列差异。

◆ 6. 完善基因组注释（有参）

现有的基因预测软件很难准确地预测基因，基于三代全长转录组测序可以促进复杂基因结构的预测，提高其准确性。

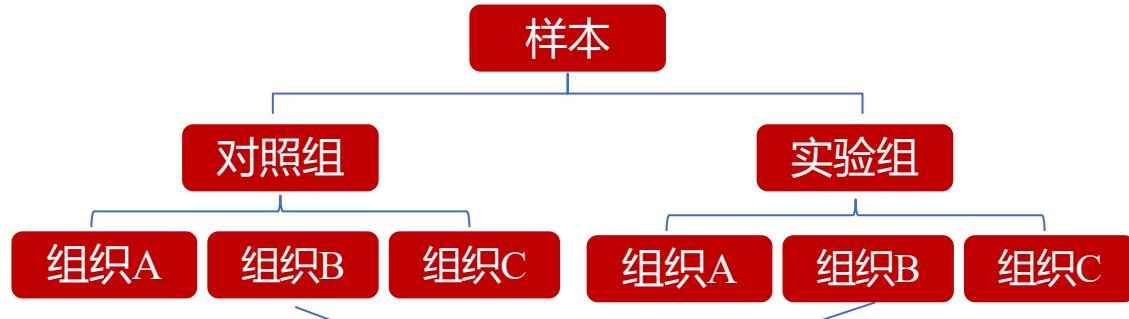
完善注释信息



- ◆可多组织混样（如不同发育阶段的组织或者相同发育阶段的不同组织，同一组织的不同发育时期，甚至不同个体进行混样测序），侧重于解决结构和功能分析。混样建议不超过4个，混样增多需要加大测序数据量，且有可能掩盖低丰度转录本信息。
- ◆混合好的样本混样进行三代全长测序，二代分别进行转录组测序，三代进行结构解析，二代校正并进行定量分析，完善注释的同时还可分析不同组织间的差异表达。

- Li Y, Dai C, Hu C, Liu Z, Kang C. Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. Plant J. 2017 Apr;90(1):164-176. doi: 10.1111/tpj.13462. Epub 2017 Feb 11. PubMed PMID: 27997733.
- Lagarde J , Uszczynska-Ratajczak B , Carbonell S , et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing[J]. Nature Genetics, 2017.

差异转录组



三代：实验组和对照组分别Iso-seq测序

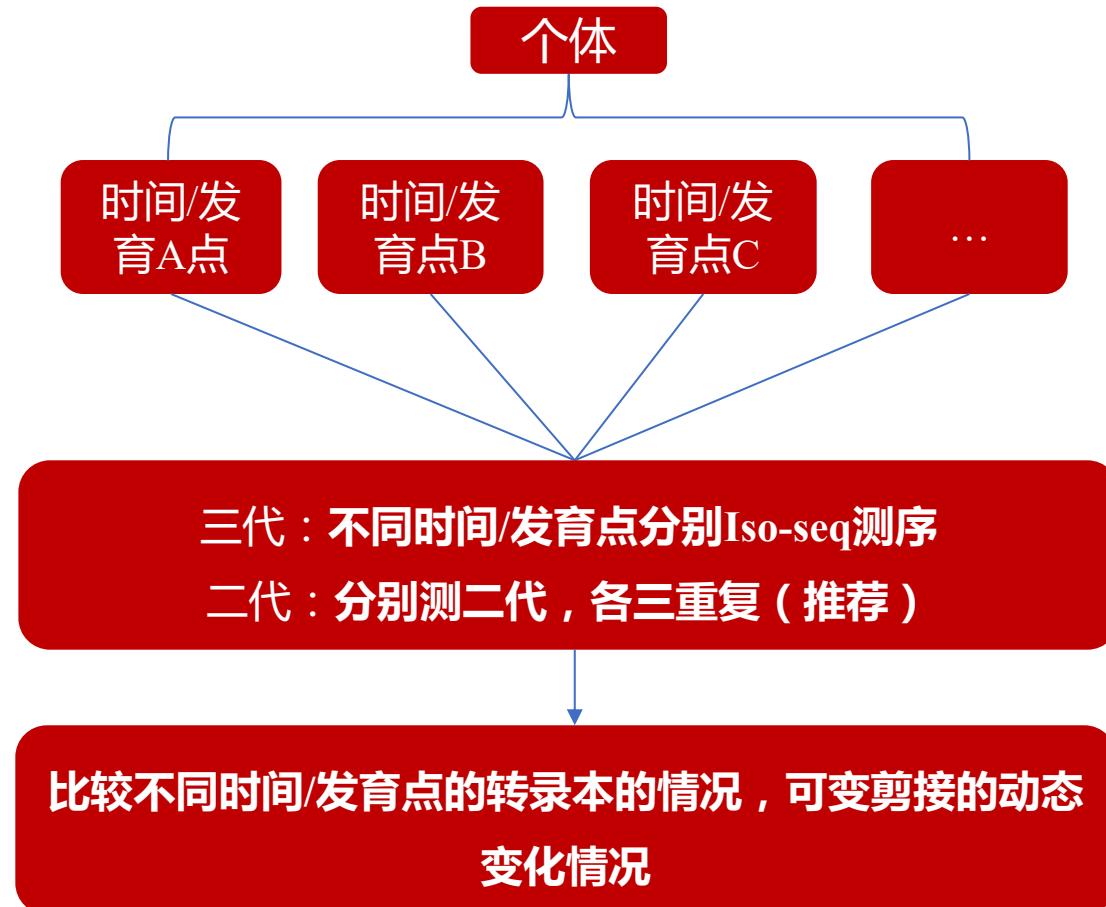
二代：分别测二代，各三个重复

差异分析，既可分析组织间差异，也可分析胁迫和正常条件的差异，找出差异关键基因或转录本，功能性研究，可结合多组学进行研究，并联合实际处理条件，研究动植物的胁迫、适应驯化机制

注：样本选择为同一物种的正常组和实验组。

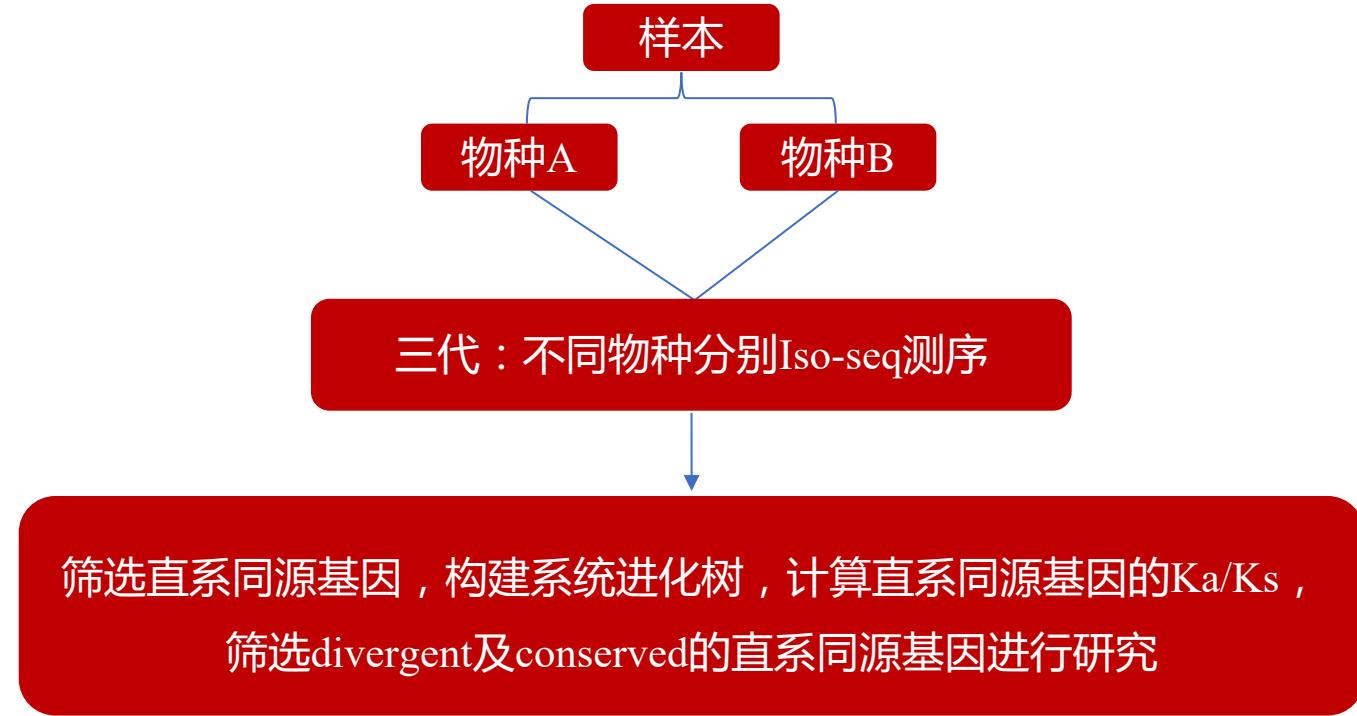
Zhu F Y , Chen M X , Ye N H , et al. Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in *Arabidopsis* seedlings[J]. The Plant Journal, 2017, 91(3):518-533.

动态转录组



Kuang Z , Boeke J D , Canzar S . The dynamic landscape of fission yeast meiosis alternative-splice isoforms[J]. Genome Research, 2017, 27(1):145-156.

比较转录组



通过Iso-seq技术手段研究少数近缘物种间的亲缘关系以及不同物种或亚种间mRNA序列差异。

Wang B , Regulski M , Tseng E , et al. A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing[J]. Genome Research, 2018:genome;gr.227462.117v1.

案例解析1

Hybrid sequencing reveals insight into heat sensing and signaling of bread wheat

发表时间：2019.4.23 IF：6.141

小麦旗叶和籽粒
37°C热胁迫下分别于0m、5m、10m、30m、1h、4h取样

PaBios RS II平台：全长转录组测序，分片段建库（0.5–1, 1–2, 2–3, 3–6 kb）；二代：Illumina HiSeq X

1. 鉴定新基因和新转录本；
2. 差异表达基因和可变剪接基因；
3. 差异基因的富集；
4. 转录因子鉴定；
5. LncRNA的鉴定。

1. 新发现4947个基因和70285个转录本，生成了热胁迫应答全长转录本的全面和动态列表。
2. 补充完善了最近发布的小麦参考基因组。
3. 热休克因子依赖和热信号途径以及早期热应答的代谢出现变化。
4. 器官和亚基因组之间的差异反应和功能分区，并发现了HS反应中转录调控和选择性剪接调控的差异模式。

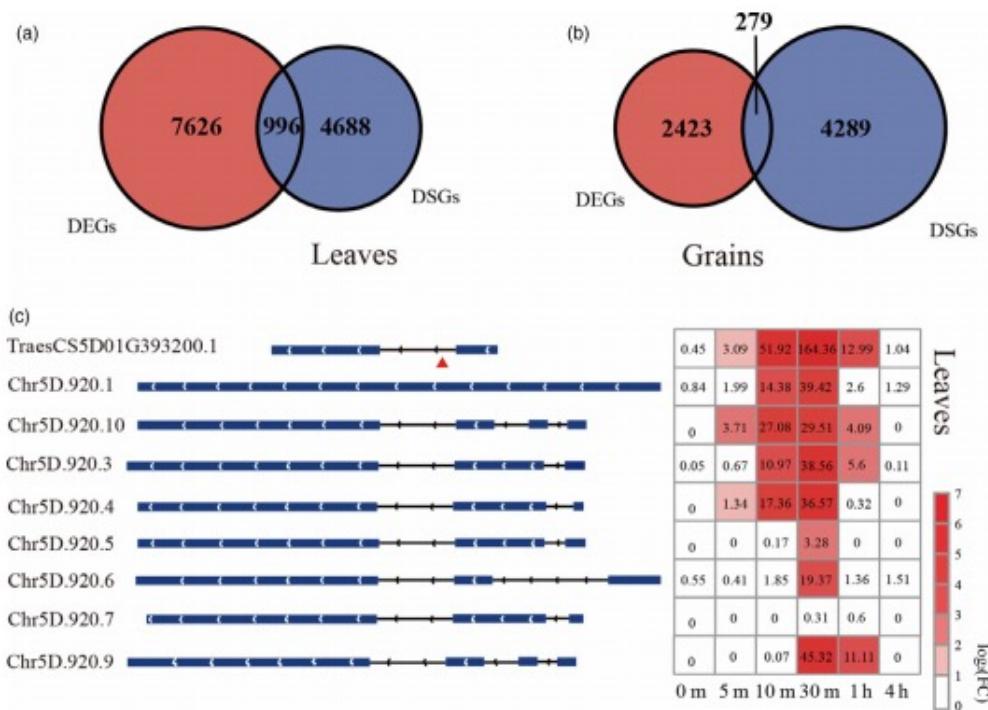


图 差异表达基因(DEGs)和差异剪接基因(DSGs)的韦恩图，以及热休克因子编码基因对热胁迫应答的转录调控和选择性剪接调控

案例解析2

Parallels between natural selection in the cold-adapted crop-wild relative *Tripsacum dactyloides* and artificial selection in temperate adapted maize

发表时间：2019.03 IF：6.141

野生磨擦草种子发芽生长的单一植株
(根，叶和茎)

Pacbio RSII：1-2kb、2-3kb和3-6kb文库，
2cell/文库

- 1.与玉米参考基因组比对
2. lncRNA分析
3. 构建摩擦草进化树
4. 脂肪酸代谢影响代谢情况

1. 磷脂代谢基因加速的蛋白质序列进化提供了可行的预测机制，以提高摩擦草相对于玉米的耐冷和耐冷性。
2. 估计摩擦草属和玉蜀黍属的分化发生大约770万年，显著早于此前报道的使用来自球蛋白-1的序列数据的450-480万年。

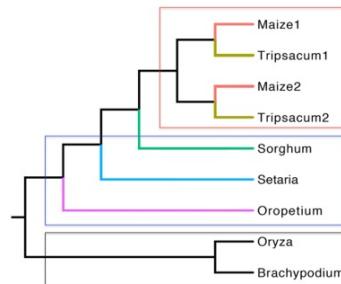


图 系统进化树

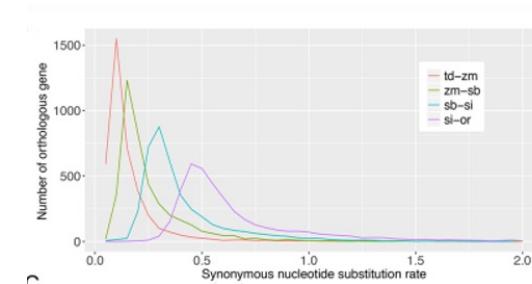


图 直系同源基因同义替换速率 (Ks) 分布

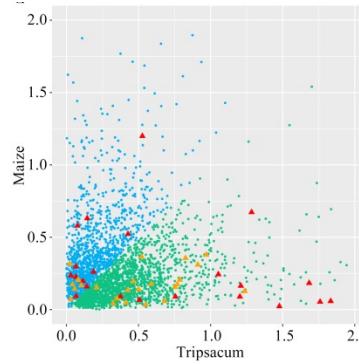


图 玉米与摩擦草Ka / Ks比值分布散点图

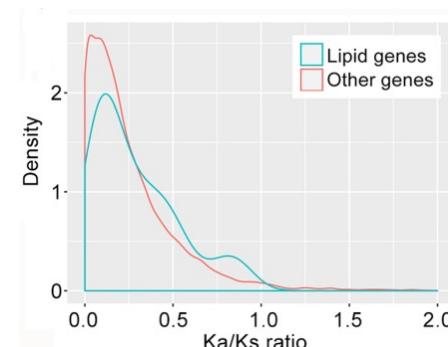
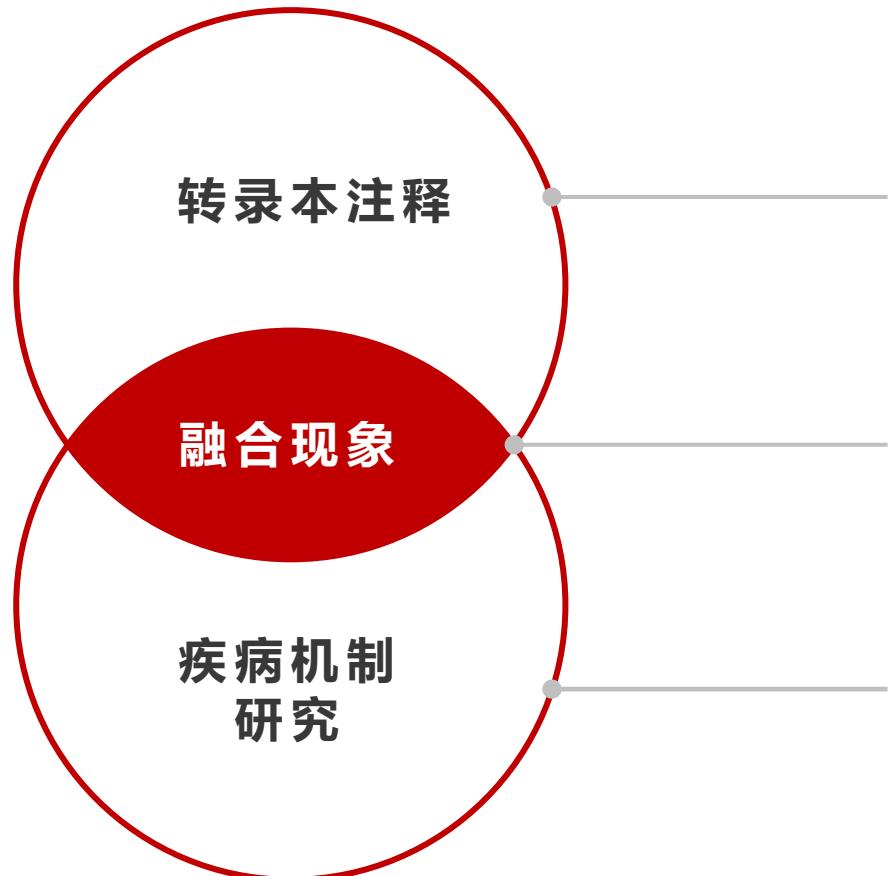


图 脂质代谢基因与其他功能基因之间Ka / Ks分布

医学中的应用



侧重于解决结构分析，如可变剪切，选择性多聚腺苷酸化，lncRNA，预测新基因，新转录本异构体。

对于部分可能发生融合基因的癌症如前列腺癌，白血病等，找到癌症致病机理。

寻找患者个体的不同时间转录本随用药，手术等因素变化，了解病情发展情况及治疗手段的有效性。

4



送样和建库策略

测序策略

◆ 数据量推荐：

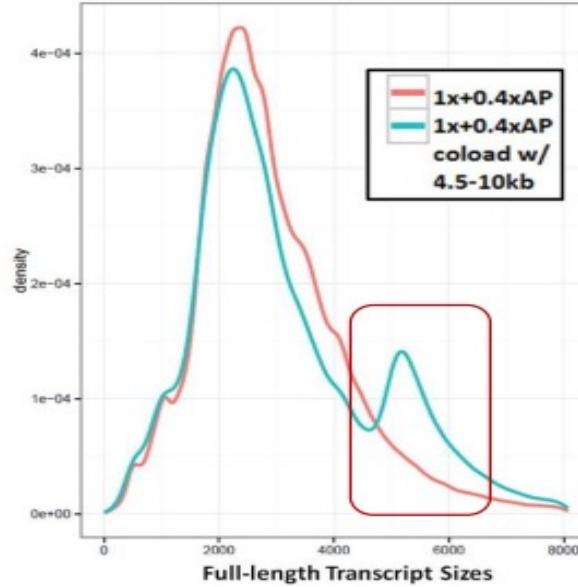
- 通常全长转录组数据量推荐30-60G，如果要想达到转录本的饱和需要更高的数据量；如果只关注高频转录本可以推荐20-30G。
- 可支持同项目同物种多样品pooling测序，数据拆分效率80-85%。

◆ 二+三测序

- 两个及以上样本：二代转录组+三代全长转录组数据，三代测30G，二代至少三个重复，每个至少6G。
- 三代测序样本取样条件必须和二代一致！同一批样本，同一时间段，同一处理。

◆ 建库方案推荐：

- 建一个1-10K的文库，流程简单，成本更低，2.0试剂升级后对长片段的捕获效率也提升了



送样要求

RNA 收样标准

| 文库名称 | RIN值要求 | 定量方法 | 浓度要求 | 总量要求 | 质量要求 |
|------------|--------|-----------------|-----------|------|-----------|
| PB-Iso-seq | ≥7.5 | Agilent 2100 | ≥300ng/μl | 2ug | 无DNA或杂质污染 |

样本送样需求

| 提取核酸 | 物种分类 | 部位 | 建议送样量 | 目标提取量 |
|-----------------|------|-------|---------------------------------------------------------|-------|
| 二代或三代 Total RNA | 植物 | 根 | 常规材料≥500mg, 特殊材料≥1g | ≥2ug |
| | | 茎 | | ≥2ug |
| | | 叶 | | ≥2ug |
| | | 花 | | ≥2ug |
| | | 果实/种子 | | ≥2ug |
| | 人/动物 | 幼苗 | 常规材料≥200mg, 特殊材料≥500mg ≥2*10 ⁶ 细胞 ≥4mL | ≥2ug |
| | | 组织 | | ≥2ug |
| | | 细胞 | | ≥2ug |
| | | 全血 | | ≥2ug |

仪器平台

作为国内最早引入PacBio Sequel II的公司，贝瑞基因目前自主拥有19台最新的PacBio Sequel II测序仪，相比于其他公司，在仪器数量上有明显的优势，贝瑞是目前国内规模最大的PacBio三代测序服务商。能够为科研人员稳定、高效的提供读长更长，通量更高、数据质量更优的三代测序服务。



国内最大的三代测序服务商





TCGA ANALYSIS Thank You!



官方网站



官方微 信

www.berrygenomics.com