生信技能树
Biotrainee.com

转录组分析入门（1h）+实战(2h)

沈梦圆

2017/11/19

目录

入门

实战

RNA-seq数据分析学习资源推荐

Conesa *et al. Genome Biology* (2016) 17:13
DOI 10.1186/s13059-016-0881-8

Genome Biology

REVIEW Open Access

A survey of best practices for RNA-seq data analysis

Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał Wojciech Szcześniak[12], Daniel J. Gaffney[3], Laura L. Elo[13], Xuegong Zhang[14,15]

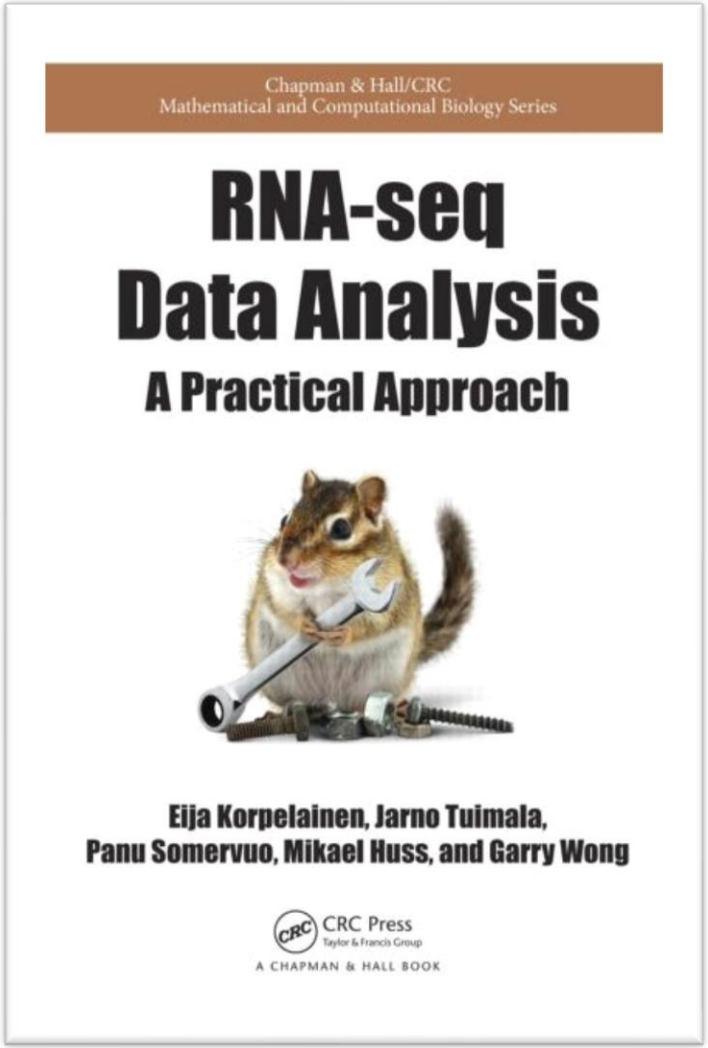
nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-00050-4 OPEN

Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis

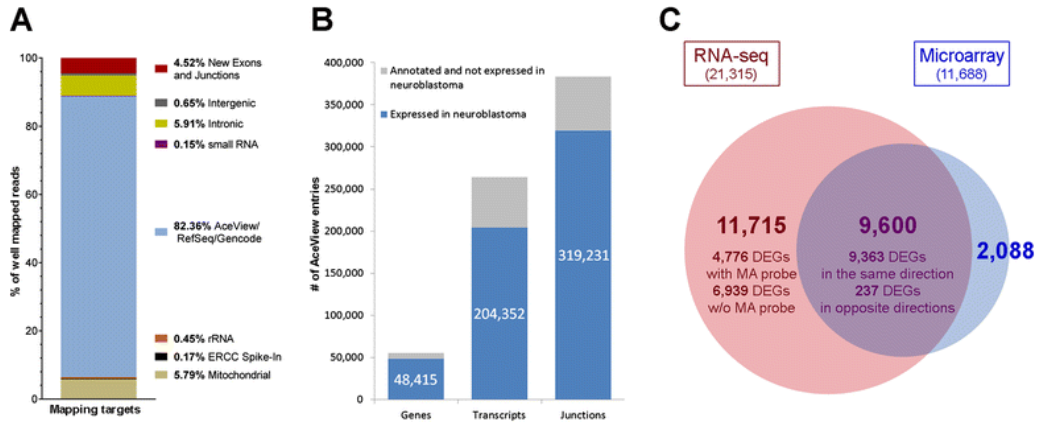
Sayed Mohammad Ebrahim Sahraeian[1], Marghoob Mohiyuddin[1], Robert Sebra[2], Hagen Tilgner[3], Pegah T. Afshar[4], Kin Fai Au[5], Narges Bani Asadi[1], Mark B. Gerstein[6], Wing Hung Wong[7], Michael P. Snyder[3], Eric Schadt[2] & Hugo Y.K. Lam[1]



http://mp.weixin.qq.com/s/wjWu0DhV3GrEySy4hIJPOQ

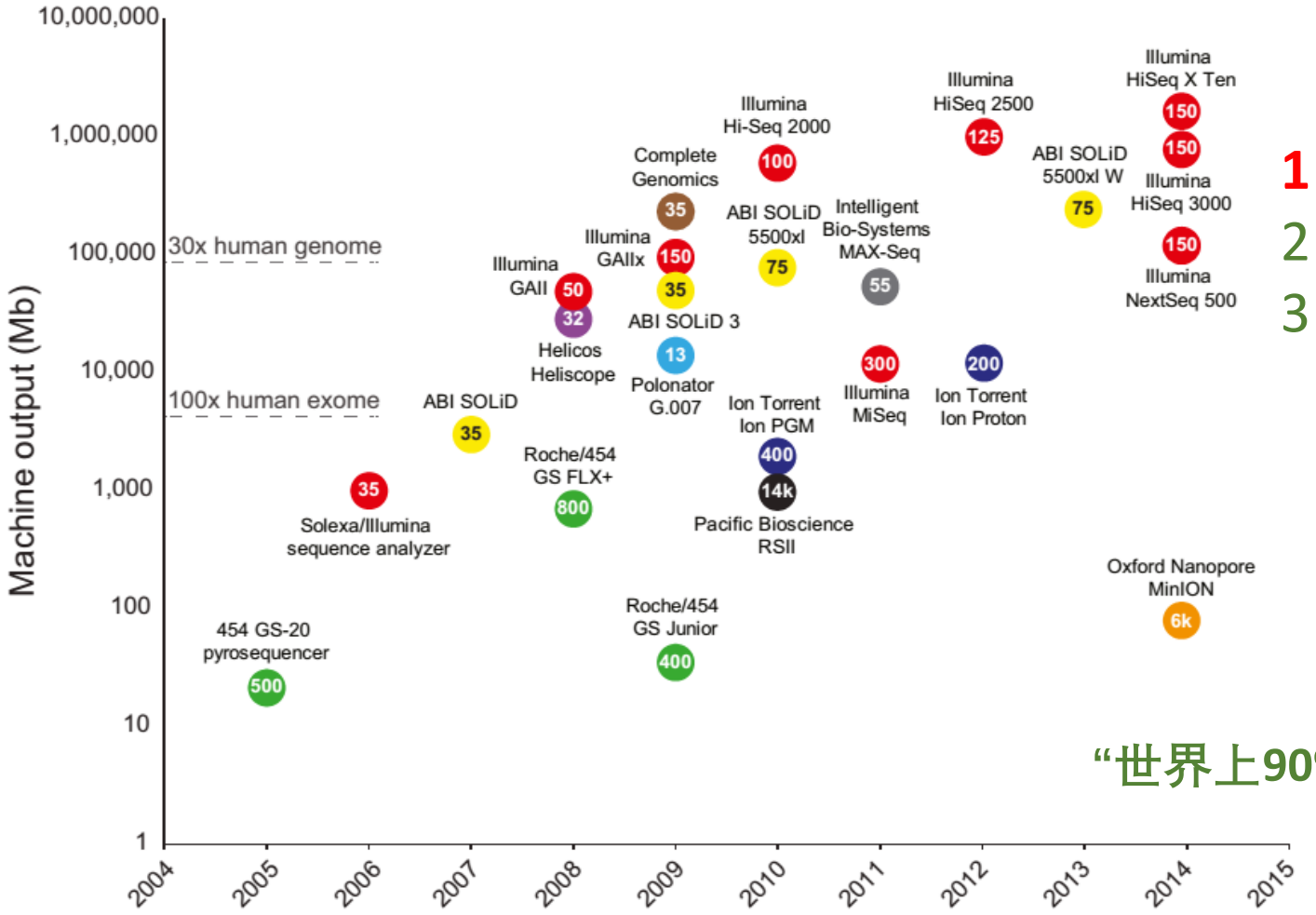
1 背景- 基因表达- RNA-Seq &芯片 & qPCR

- 检测基因数量：RNA-Seq是芯片的2倍以上；
- 检测差异表达基因数量：RNA-Seq是芯片的2倍左右；
- 转录本变异体差异表达检测数量，RNA-Seq是芯片的8倍
- 低丰度基因检测准确性，RNA-Seq的qPCR验证率是芯片的5倍；
- 差异表达倍数准确性，RNA-Seq与qPCR相关性比芯片高14%‘；
- 芯片与RNA测序数据是否可以通用？可以。



- 总的来说，RNA-seq优势更大，更广的检测范围、低表达基因有更高的敏感性，在可变剪切和基因融合检测更具优势；
- qPCR作为验证手段，金标准；

1 背景-高通量测序平台



1 Illumina

2 Life Technologies/ThermoFisher/Iorrent

3 Pacfic Biosciences

“世界上90%以上的测序数据都由Illumina仪器产生”

1 背景-高通量测序平台

	 MiniSeq System	 MiSeq Series	 NextSeq Series	 HiSeq Series	 HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million†	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Run Time	4–24 hours	4–55 hours	12–30 hours	<1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	<3 days
Benchtop Sequencer	Yes	Yes	Yes	No	No
System Versions	• MiniSeq System for low-throughput targeted DNA and RNA sequencing	• MiSeq System for targeted and small genome sequencing • MiSeq FGx System for forensic genomics • MiSeqDx System for molecular diagnostics	• NextSeq 500 System for everyday genomics • NextSeq 550 System for both sequencing and cytogenomic arrays	• HiSeq 3000/HiSeq 4000 Systems for production-scale genomics • HiSeq 2500 Systems for large-scale genomics	• HiSeq X Five System for production-scale whole-genome sequencing • HiSeq X Ten System for population-scale whole-genome sequencing

HiSeq X Ten 最便宜

以后

NovaSeq 更便宜

1 背景-RNA-seq到底测的是什么？

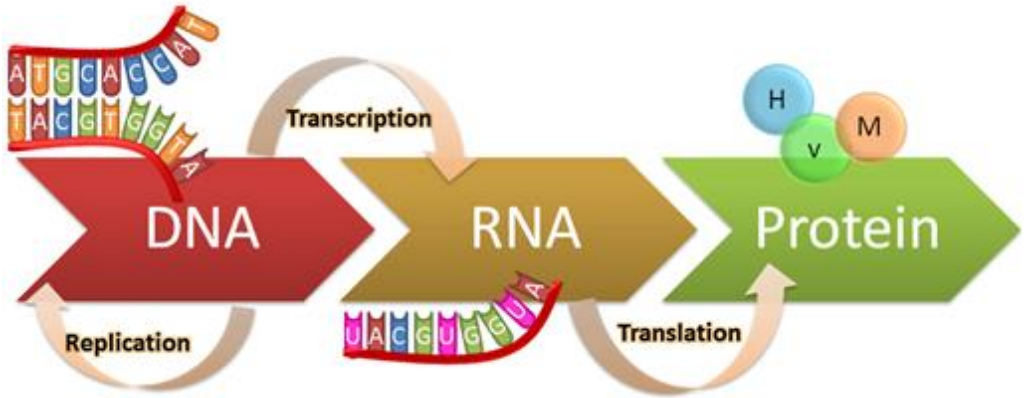
RNA class	Length (nt)	Number of species	Abundance (% RNA mass)
rRNA	160–5,025	6	80–85
tRNA	70–90	~120	10–13
mRNA	2,000–10,000	>20,000	3–5
snoRNA	~90	~200	<2
snRNA	100–300	10	
miRNA	~22	~1,000	
lncRNA	200–17,000	>1,000	
Y RNA	80–110	5	
7SLRNA	~300	1	
telRNA	450	1	
vtRNA	~80–120	5	
scaRNA	~200–300	~20	
piRNA	27	>1,000,000	

Nature Reviews | Molecular Cell Biology

mRNA在生物个体内RNA的组分中只占很小的一部分，rRNA占绝大多数。

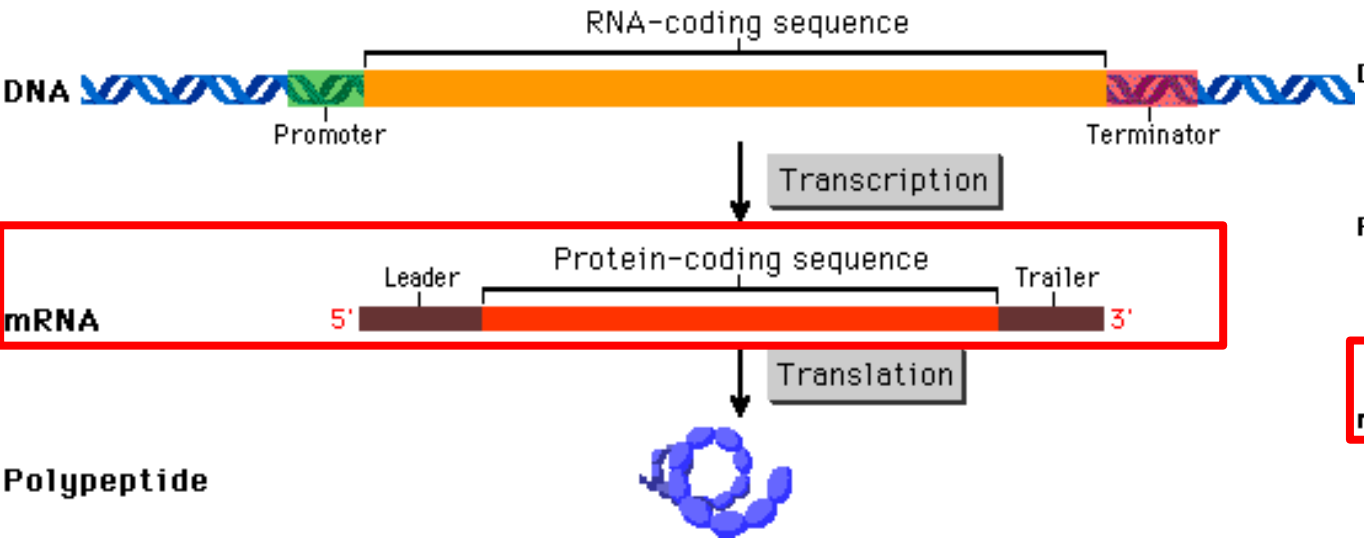
一般我们说RNA-seq指的都是mRNA-seq（编码RNA），后面的流程也都是主要针对mRNA-seq数据分析的。（常规）

在科学家们的努力下，也可以把那些非编码RNA提取出来建库，进行测序。(比较小众，但高级)

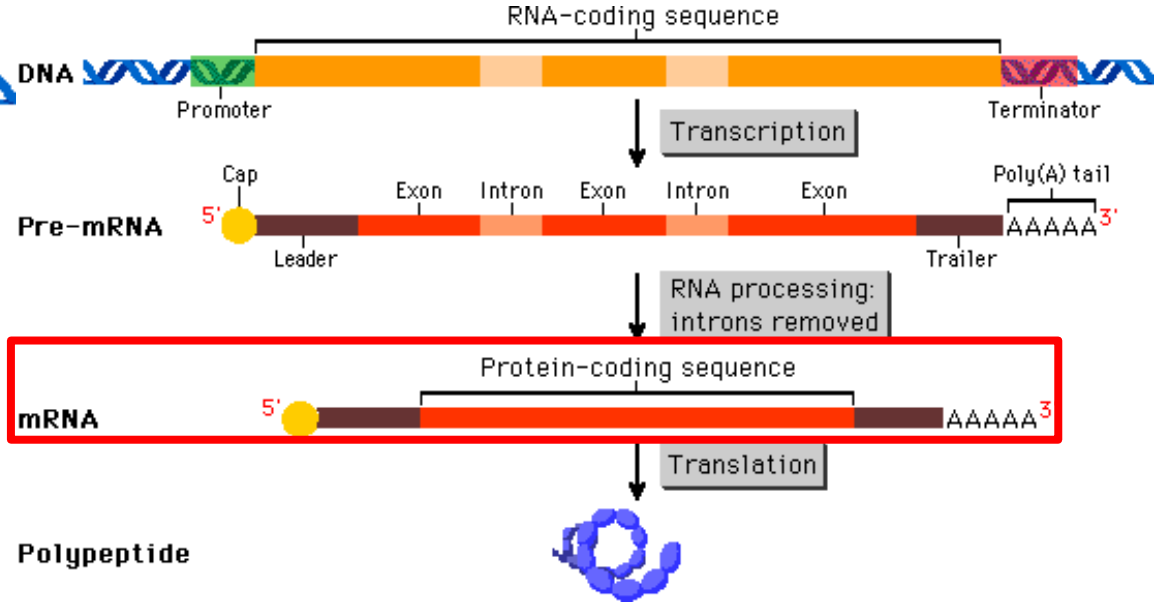


1 背景-RNA-seq到底测的是什么？

mRNA in Prokaryotes



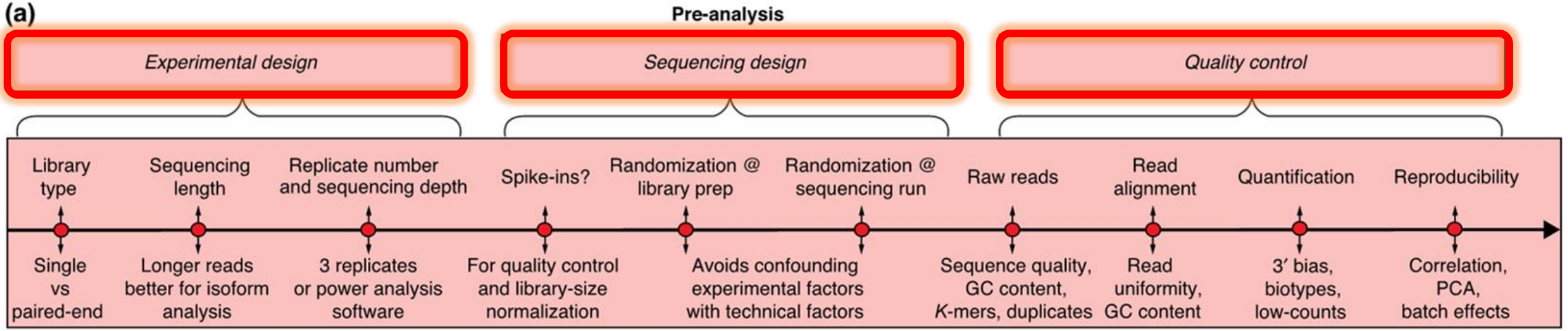
mRNA in Eukaryotes



原核与真核生物mRNA的结构特点也不同:
真核生物mRNA由5’端帽子结构、5’端不翻译区、翻译区、3’端不翻译区和3’端聚腺苷酸尾巴组成,
原核生物mRNA无5’端帽子结构和3’端聚腺苷酸尾巴。

2 数据分析前（Per-analysis）

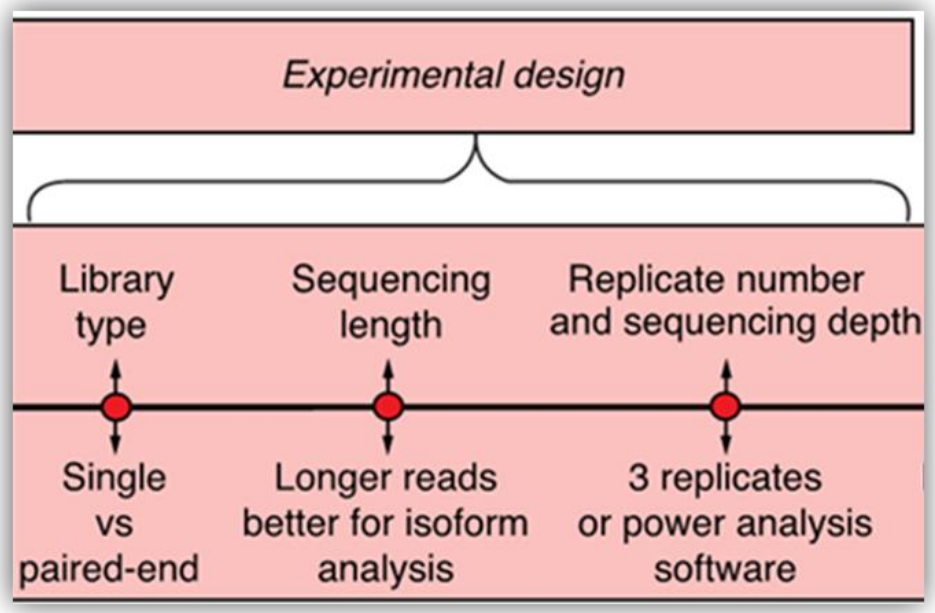
实验设计　　测序设计　　质量设计



2.1 实验设计

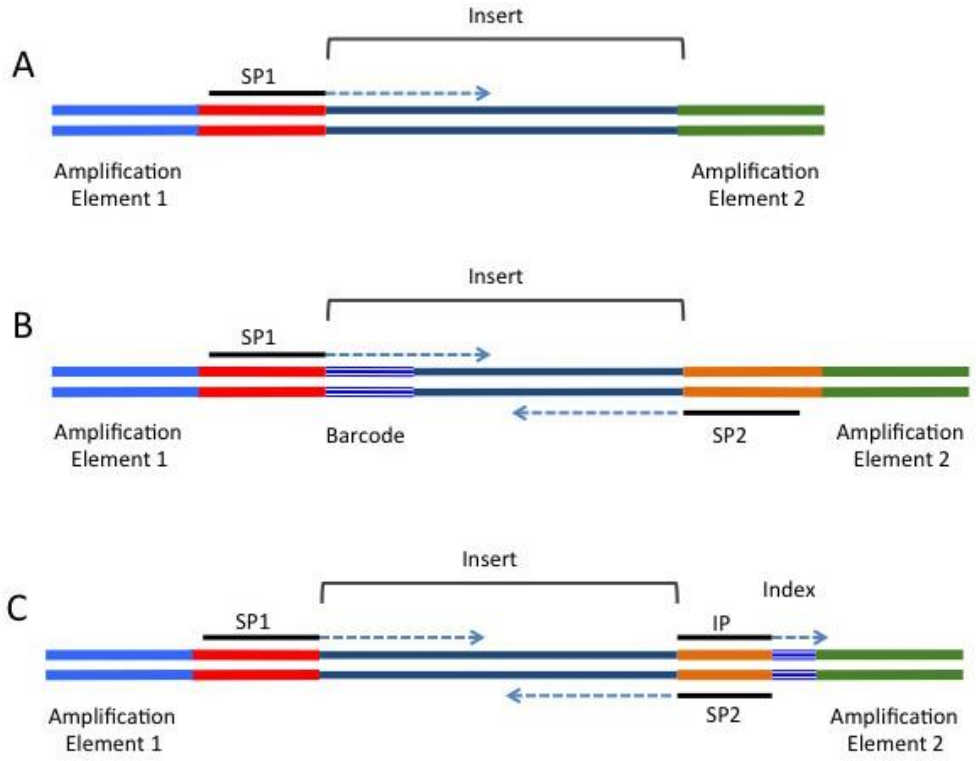
Q：

- 单端测序还是双端测序？
- 测序读长多少？
- 是否建链特异性文库？
- 需要多需要测多少数据量？
  （测序深度多少？）
- 设置多少个生物学重复比较合适？
- 如果我们做了生物重复还要做技术重复么？



2.1 实验设计-文库构建

Q：
- 单端测序还是双端测序？
- 测序读长多少？



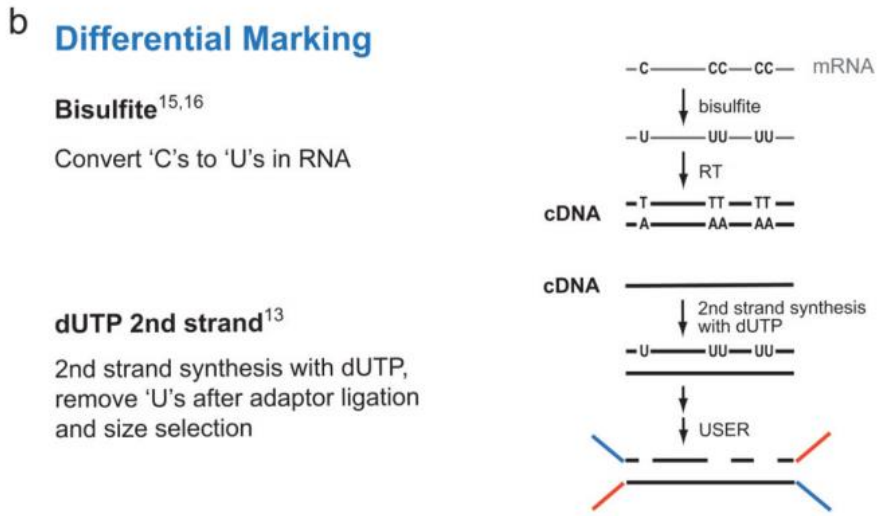
1. 一般生物体中的的RNA中，rRNA占绝大多数，含量超过90%，而mRNA的含量在1-2%左右。
对于真核生物，一般使用加poly(A)选择性富集mRNA或者而原核生物则是通过去除rRNA。
2. 对于Illumina，测序插入片段一般小于500bp。确定合适长度的插入片段是后续测序和分析的关键。
3. 单端还是**双端测序**
如果你研究的某个物种的基因表达水平，并且它的转录组已经被注释很好了，单端测序产生的数据量一般是足够的了。（单端50bp）
双端测序呢，它的读长更长，更适合于那些没有被注释的转录组物种的研究，便于其转录本的从头拼接。(双端150bp)
选双端，因为便宜！

简简单单讲insertion size: https://guangchuangyu.github.io/cn/2014/08/insertion-size/

2.1 实验设计-文库构建

Q：是否建链特异性文库？

	普通	链特异性
表达准确性	高	更高
转录本方向	无	有
检测反义转录本	无	有
分析成本	低	较高
操纵子检测	无	有
非编码RNA	可以，效果差	可以，效果好



【陈巍学基因】视频15：RiboZero和方向性RNA文库：http://mp.weixin.qq.com/s/4KIgbE5PbkKzJTg3lDVxzg

掺U法
ScriptSeq法

链特异建库那点事

几个常用软件的设置

hisat2 --rna-strandness RF
tophat --library-type option fr-firststrand
htseq-count：–s reverse
rsem：--forward-prob0
sXpress --rf-stranded / --fr-stranded
trinity --SSlibtype RF

参数错了又怎样？

链特异性当作普通建库数据：
具体某一个基因而言，影响不会太大，因为绝大多数反义链本身表达量就非常低。

普通建库当作链特异性数据处理：
会损失大量的数据，没剩下几个read。计算出来的结果自然也会有非常大的差异，是不准确的。

来自生信技能树-熊: https://mp.weixin.qq.com/s/7bWu4oAP3Nr2trqYTJWzSA

2.1 实验设计-测序深度和重复数

Q：需要测多少数据量？
（测序深度多少？）

Transcriptome Sequencing	Differential expression profiling	10-25M	Liu Y. et al., 2014; ENCODE 2011 RNA-Seq
	Alternative splicing	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	Allele specific expression	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	De novo assembly	>100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq

http://www.biotrainee.com:8080/thread-1373-1-1.html

Jimmy说:
人一般测20M-50M 条reads
检测低表达基因要至少100M reads == 1亿条

human genome : 3000 M nt
1/30 protein-coding genes 100 M nt
1 M reads gives 150 M nt

2.1 实验设计-测序深度和重复数

Q：
- 设置多少个生物学重复比较合适？
- 如果我们做了生物重复还要做技术重复么？

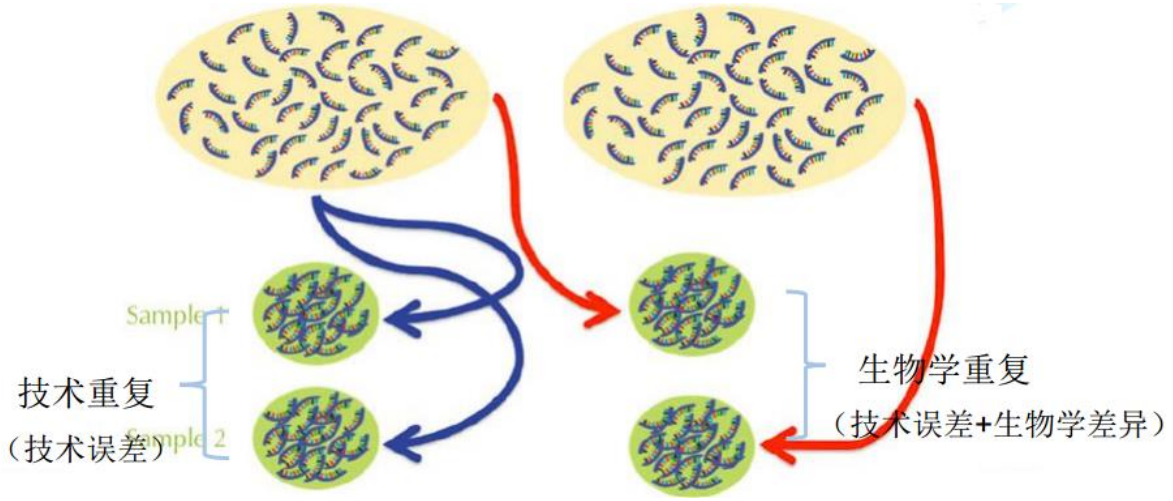


Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

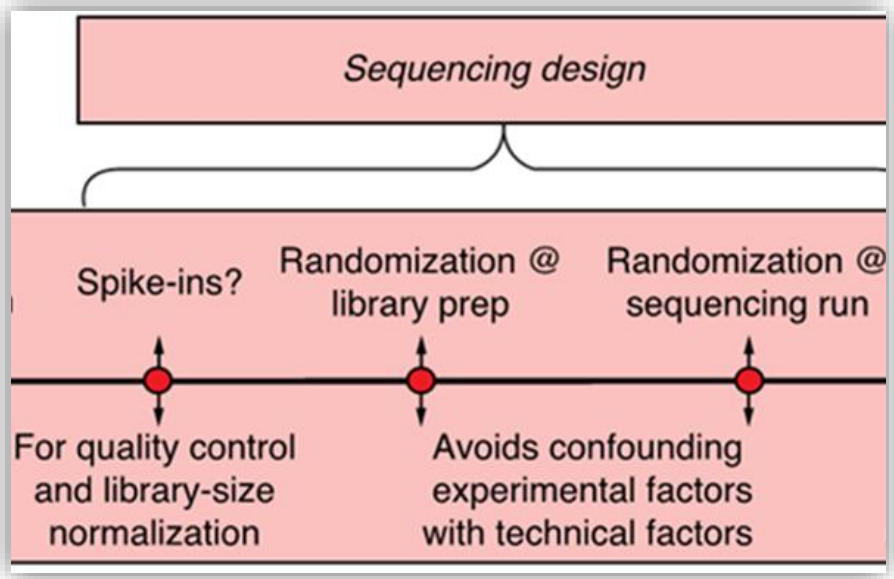
一般不做技术重复；
至少3个生物学重复，最佳6个；

The problem with technical replicates

Stat

2.2 测序设计

Q：
如何减少测序造成的误差？

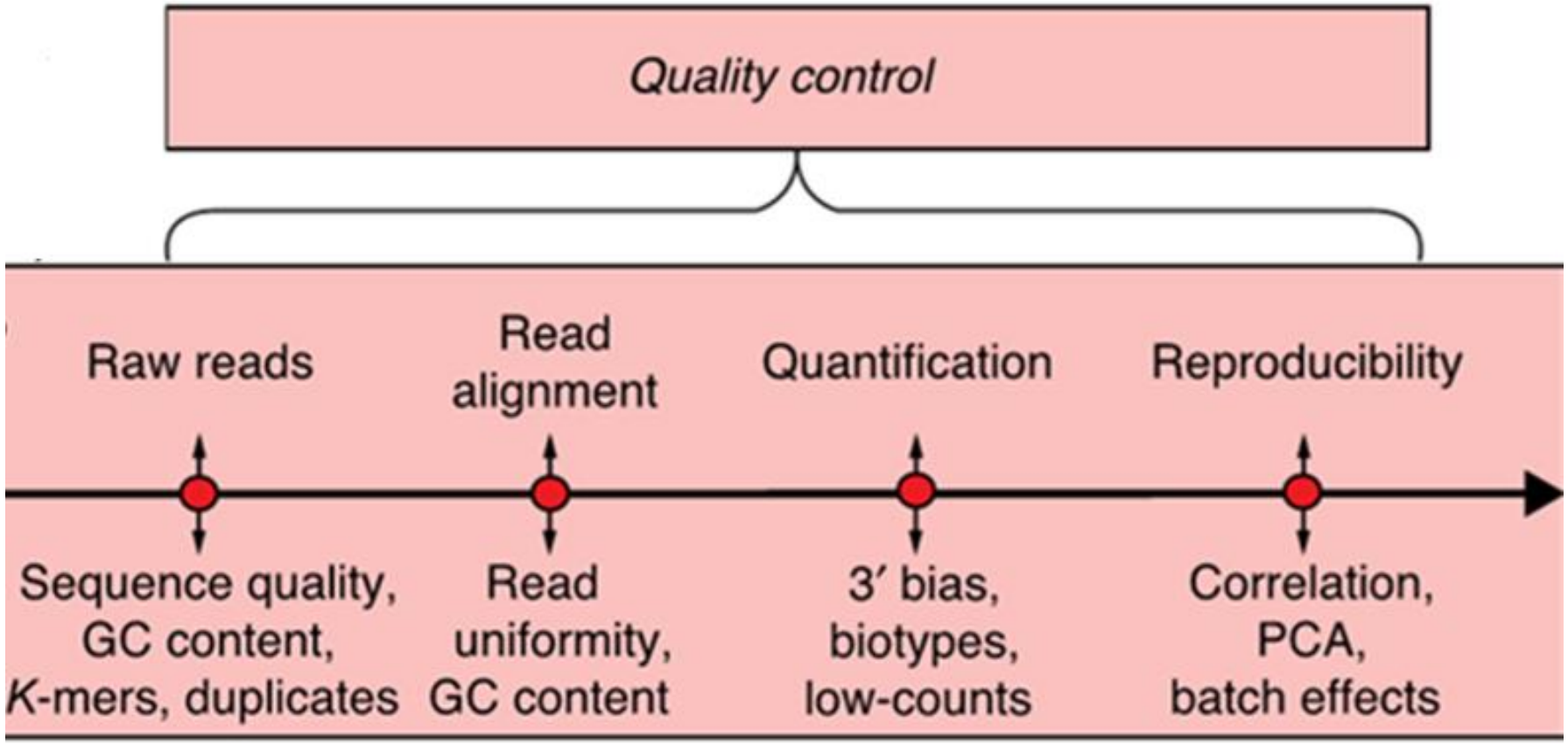


RNA-seq文库的制备和测序过程：RNA碎裂，cDNA合成，接头连接，PCR扩增，加标签（多样品混合测序），上泳池测序
如何减少误差：
1. 使用末端带随机核酸的接头或者使用化学碎裂法代替RnaseIII碎裂法；
2. 不同批次实验或者不同runs
a. 如果样品太多在一个批次或者一个run跑不完，为了避免技术误差造成太大的实验误差，要把样品随机分配到每个批次或runs中;
b.如果你的样品是多样品混合测序，每个样品要单独加上标签，每个lanes要保证足够的测序深度，为了保证所有的样品在每个lane中都有。

如果送给公司去做的话，我们要选择建库水平好些的，并且要求他们这么去做，应该会更好。

2.3 质量控制

Q：
如何进行
原始数据/
比对结果/
定量结果/
重复结果
的质量控制？



2.3 质量控制-Raw reads

Feature\Tools	NGS QC Toolkit v2.2	FastQC v0.10.0	PRINSEQ-lite v0.17[1]	TagDust	FASTX-Toolkit v0.0.13	SolexaQA v1.10	TagCleaner v0.12[1]	CANGS v1.1
Supported NGS platforms	Illumina, 454	FASTQ[2]	Illumina, 454	Illumina, 454	Illumina	Illumina	Illumina, 454	454
Parallelization	Yes	Yes	No	No	No	No	No	No
Detection of FASTQ variants	Yes	Yes	Yes	No	No	Yes	No	No
Primer/Adaptor removal	Yes	No[3]	No	Yes	Yes	No	Yes[4]	Yes
Homopolymer trimming (Roche 454 data)	Yes	No	No	No	No	No	No	Yes
Paired-end data integrity	Yes	No	No	No	No	No	No	No
QC of 454 paired-end reads	Yes	No	No	No	No	No	No	No
Sequence duplication filtering	No	No[5]	Yes	No	Yes	No	No	Yes
Low complexity filtering	No	No	Yes	No	Yes	No	No	No
N/X content filtering	No	No[6]	Yes	No	Yes	No	No	Yes
Compatability with compressed input data file	Yes	Yes	No	No	No	No	No	No
GC content calculation	Yes	Yes	Yes	No	No	No	No	No
File format conversion	Yes	No	No	No	No	No	No	No
Export HQ and/or filtered reads	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Graphical output of QC statistics	Yes	Yes	No[7]	No	Yes	Yes	No[7]	No
Dependencies	Perl modules: Parallel::ForkManager, String::Approx, GD::Graph (optional)	-	-	-	Perl module: GD::Graph	R, matrix2png	-	BLAST, NCBI nr database

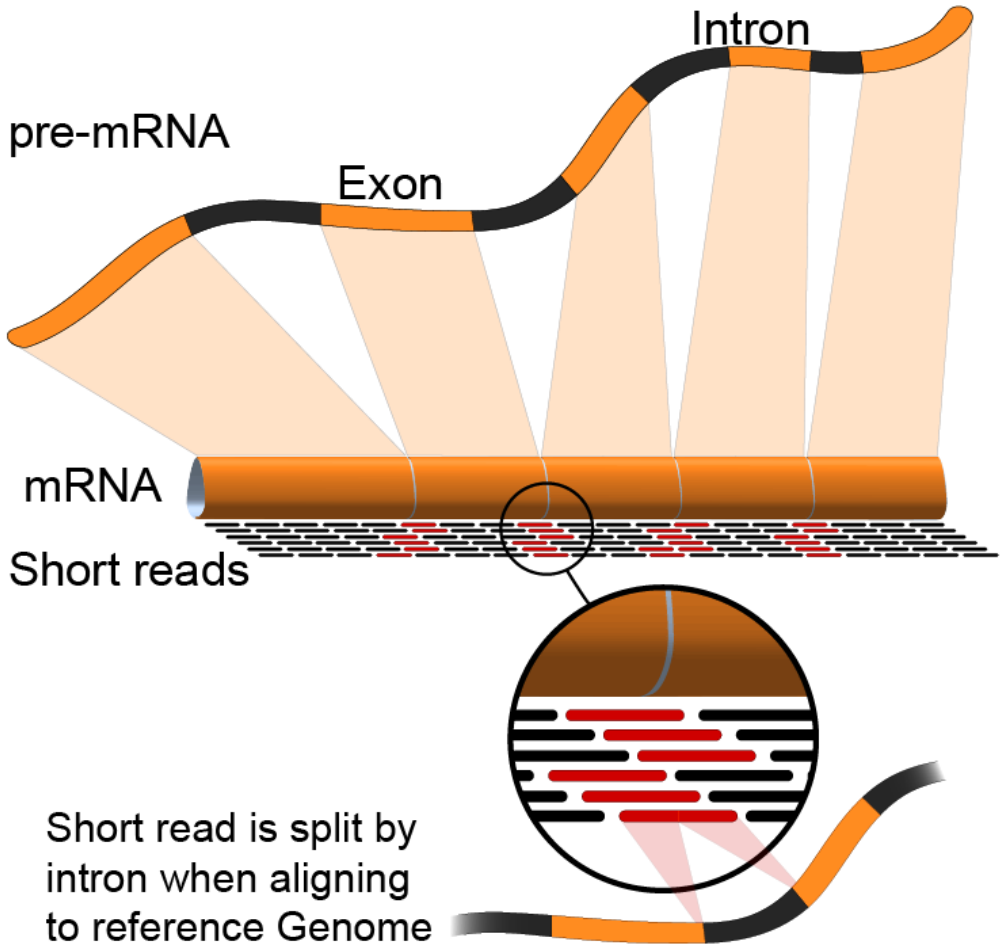
确定数据处理方式：
Trim(Long reads /Short insert)
Filter(Short reads/Large amount)

常用的软件有
FASTX-Toolkit/Trimmomatic

比较详细的介绍：
NGS测序数据的质量控制 (Quality Control，QC)
https://mp.weixin.qq.com/s/aqWz6GWjCA6UfGLwtRV37g

Patel RK, Jain M (2012) NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. PLoS ONE 7(2): e30619. doi:10.1371/journal.pone.0030619
http://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0030619

2.3 质量控制-Read alignment/Quantification/Reproducibility



比对：

1.比对上的reads占总reads的百分比（70 ~ 90 % - 基因组；更低 - 转录本）；

2. Reads比对到外显子和参考链上的覆盖度是否一致；

定量：

GC含量和基因长度偏差；

再现性：

技术重复（Spearman R^2>0.9）

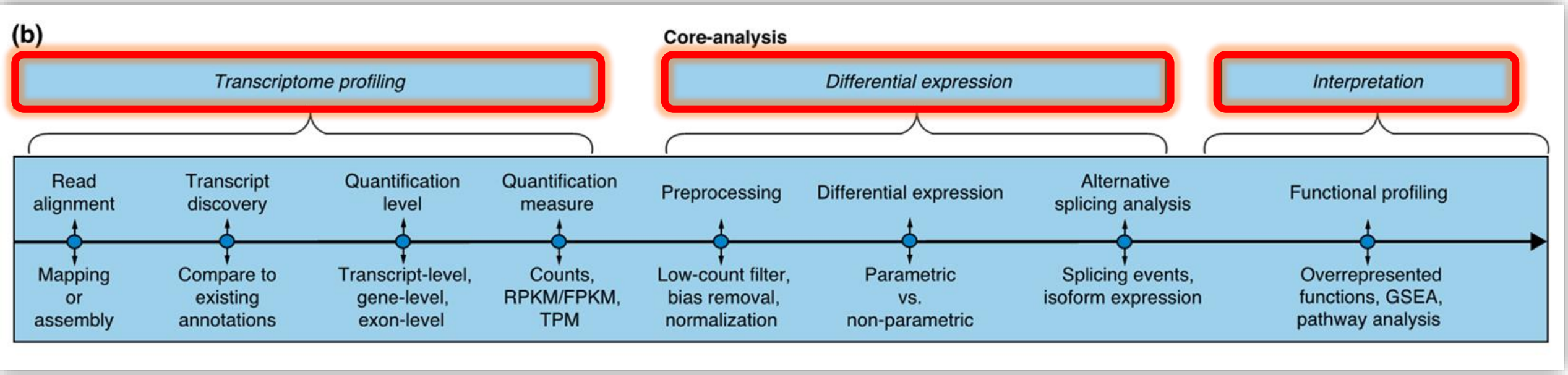
生物重复（principal component analysis :PCA）

常用软件有

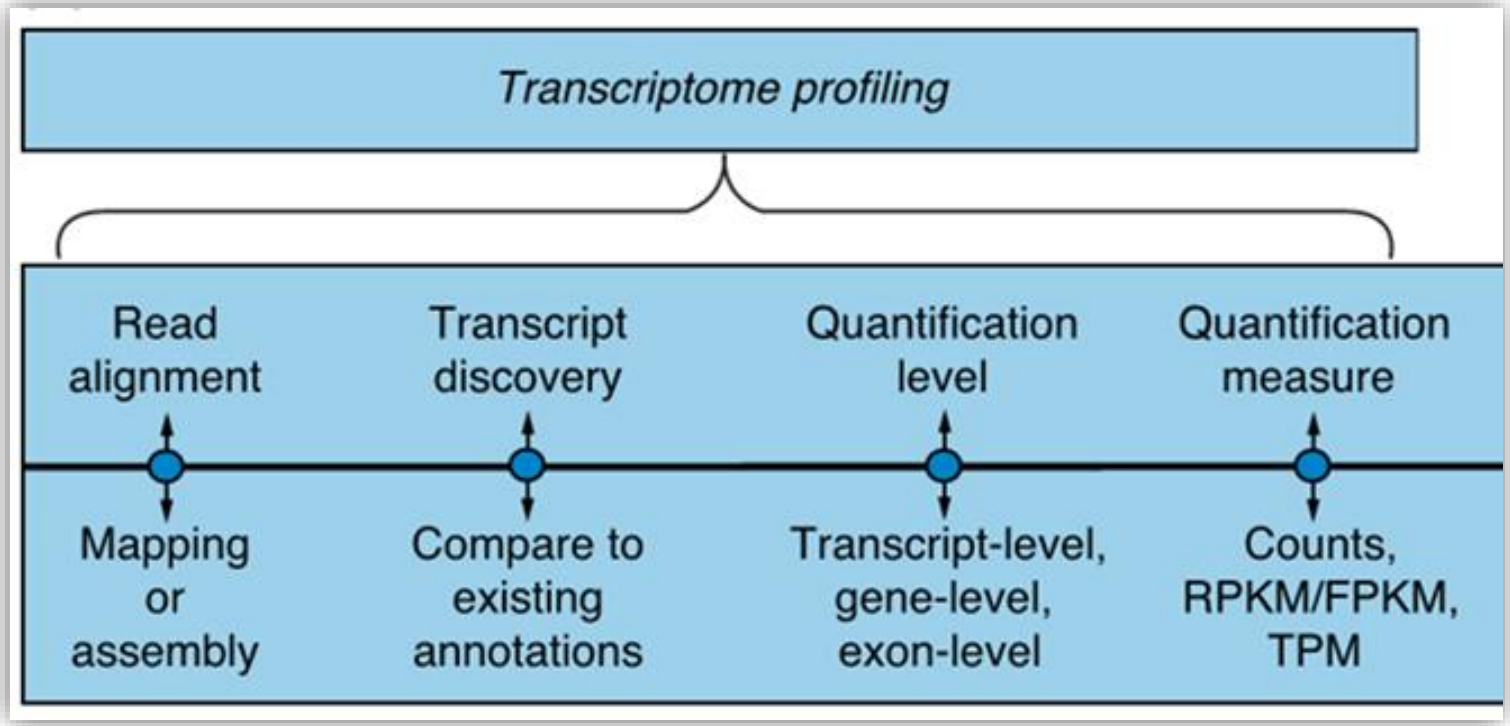
比对：Picard , RSeQC and Qualimap；

定量：NOISeq or EDASeq （R包）

3 RNA-seq数据分析（Core-analysis）



3.1 转录本分析



Q：

- 比对到基因组序列、比对到转录组序列？
- 一条reads比对到多个地方？
- 用什么来表示表达量的高低？

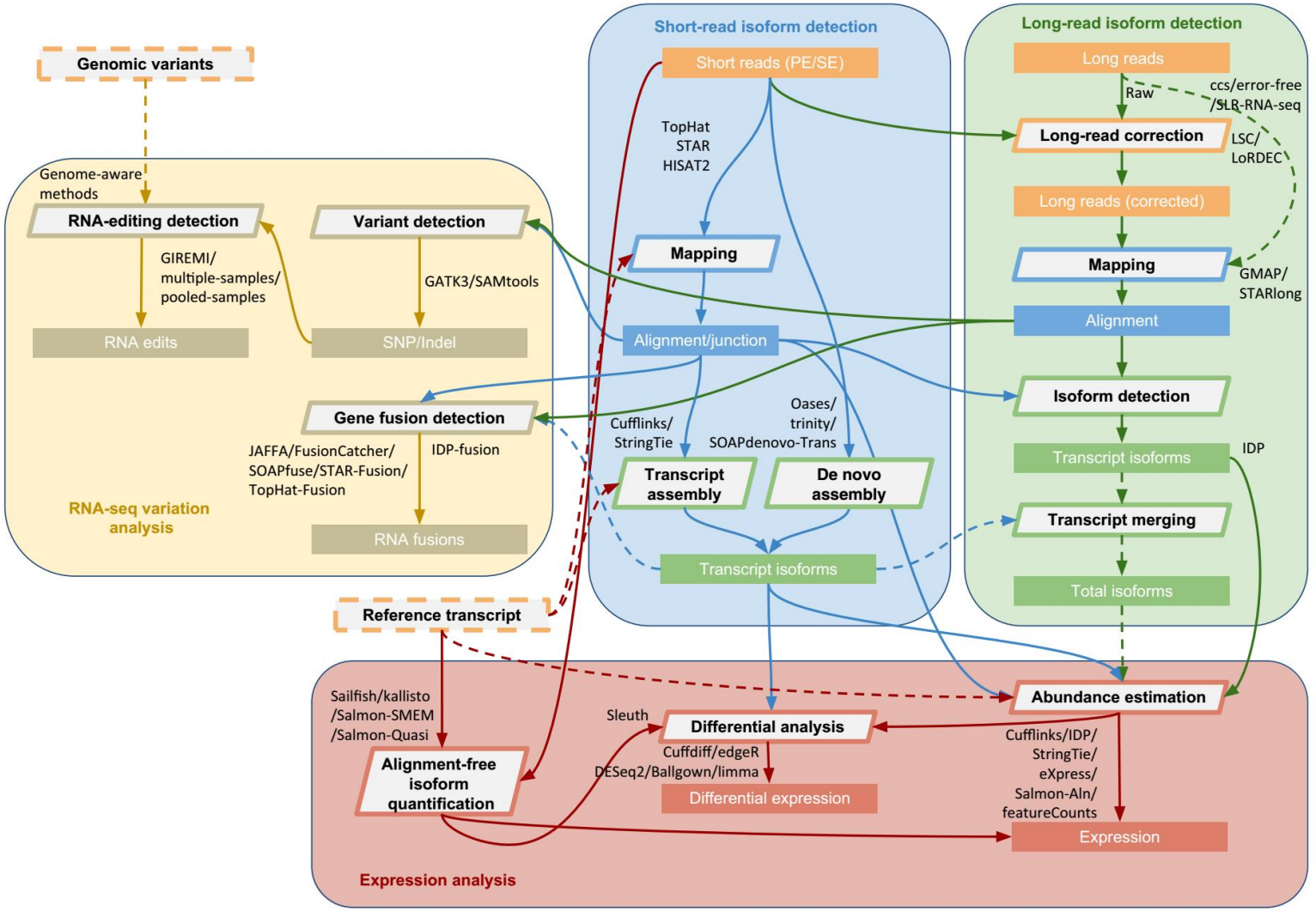
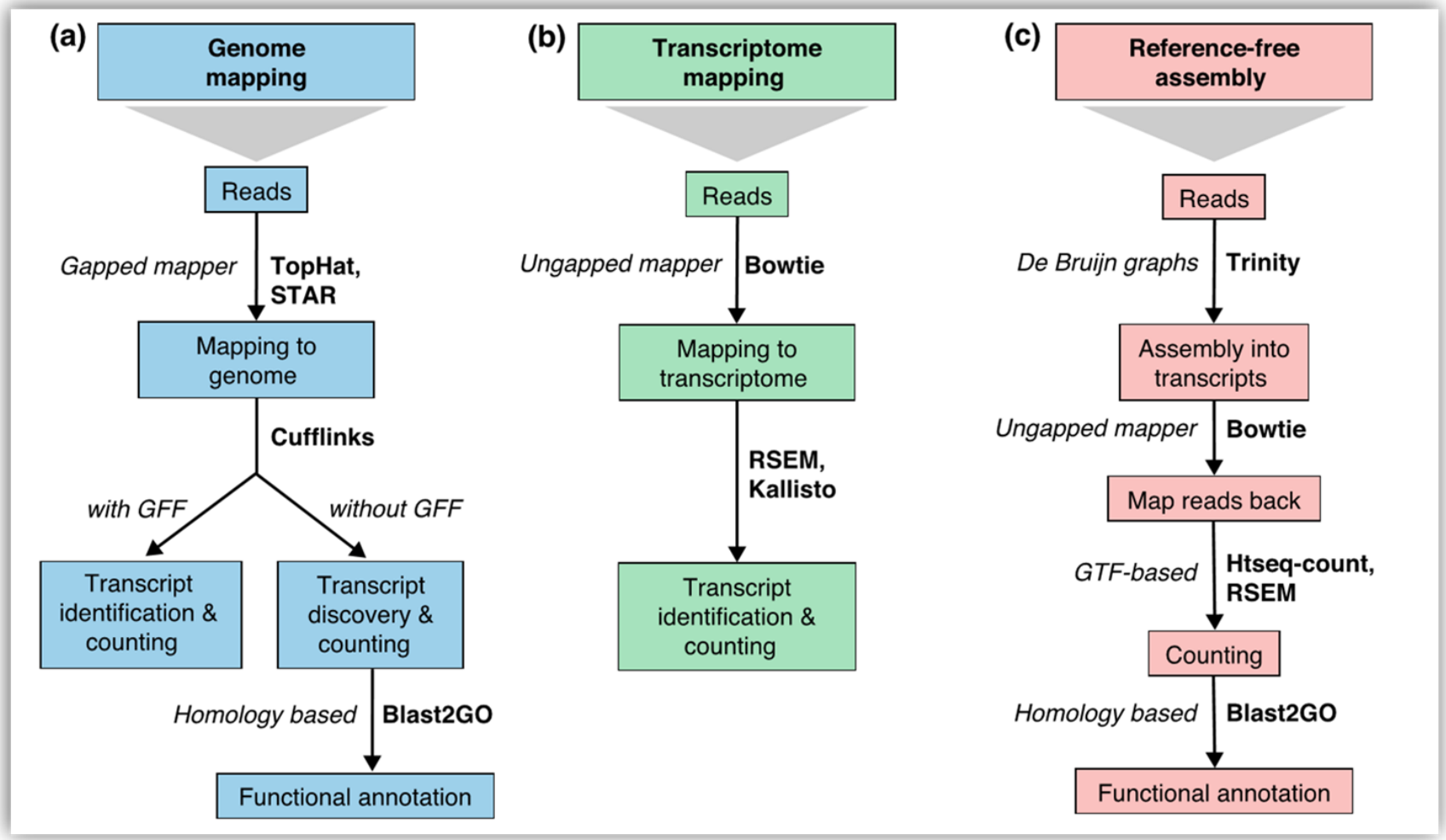


Fig. 1 The RNACocktail analysis protocol. RNACocktail is a comprehensive protocol of RNA-seq data analysis. The figure summarizes the widely used approaches for the key steps over the broad spectrum of RNA-seq analysis and also succinctly captures the possible workflows one can use to analyse RNA-seq data

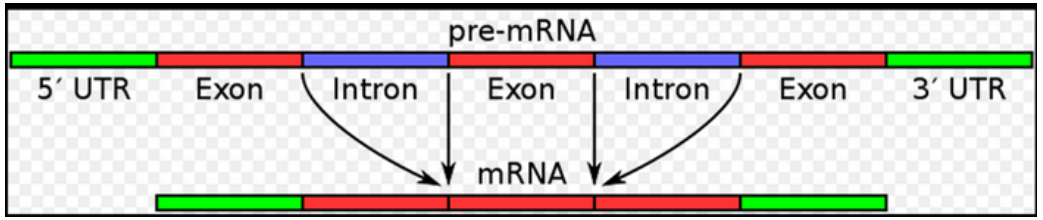
转录组分析工具大比拼（完整翻译版）：http://mp.weixin.qq.com/s/xAaj-d5LRRj0SSMFJ7Yo9Q

3.1 转录本分析-比对

Q:
比对到基因组序列、比对到转录组序列？

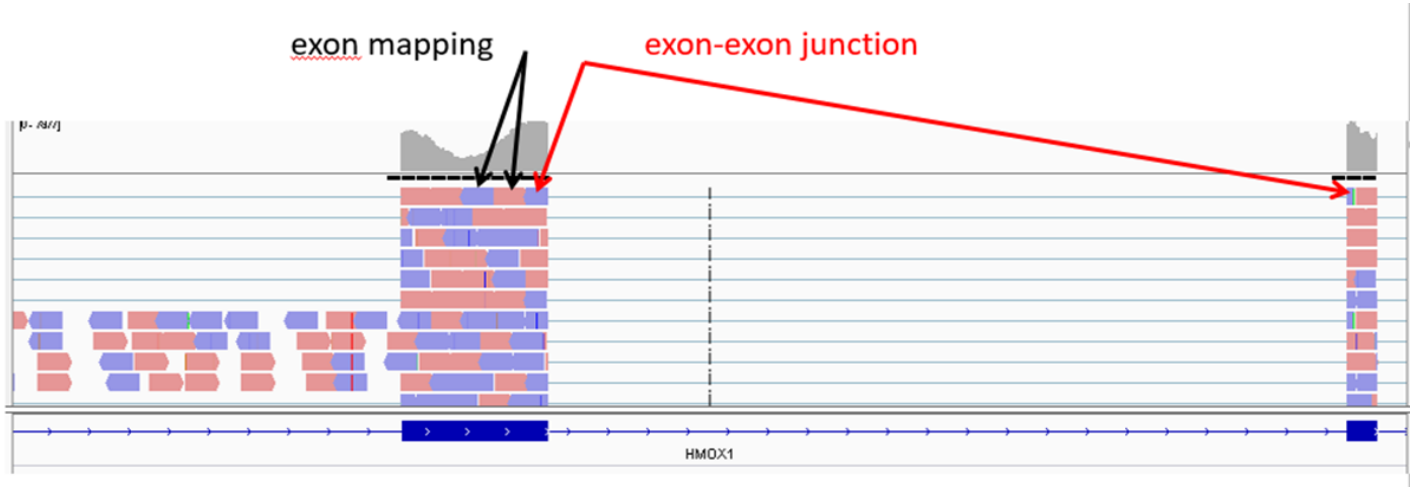


3.1 转录本分析-比对软件



- **gapped or spliced mappers:**
  TopHat(最流行但以升级HISAT2),
- GSNAP, PALMapper , MapSplice
  (可以识别SNPs或indels)
- STAR , MapSplice (检测非标准剪切位点)
- GEM (achieve ultra-fast mapping)
- STAR (map long-reads)

- **no gapped alignment and unspliced:**
- Bowtie



3.1 转录本分析-转录本识别

转录本的注释软件：
GRIT (基于5’ ends from CAGE or RAMPAGE)
Cufflinks, iReckon , SLIDE and StringTie (基于已有的注释信息)
Montebello（还可以定量）
Augustus(Gene-finding tools)(适用于编码RNA)

转录本重构

- 完全使用已知的转录本
- 基于已知转录本进一步优化
- 完全从头拼接

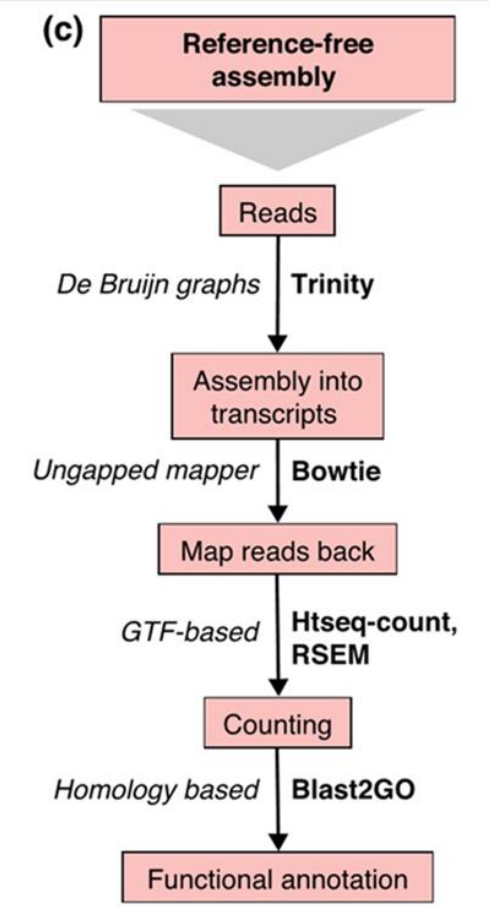
3.1 转录本分析-无参

参考基因组没有或者不完整的情况下：

完全从头拼接转录本软件：
SOAPdenovoTrans , Oases , Trans-ABySS or Trinity

把所有样本的reads混合用于转录本的拼接。
二代测序的转录组reads用于拼接还是存在一些问题的，最终拼接结果不太理想。一个转录本的拼接结果会是10~100contigs。

三代测序的读长直接可以把一个转录本读完了，完全不需要拼接。



3.1 转录本分析-表达定量

- 最简单的定量方法是累计比对到原始reads数->HTSeq-count or featureCounts
- 基因水平的定量>使用 GTF注释文件
- 基因表达量归一算法：
  - **RPKM (reads per kilobase of exon model per million reads)**
  - **FPKM (fragments per kilobase of exon model per million mapped reads)**
  - **TPM (transcripts per million)**
  - Total Count(TC):总reads数矫正
  - Upper Quartile(UQ):上四分之一分位数 总reads矫正
  - Median（Med):中位数 总reads数矫正
  - Quantile(Q):基因芯片软件limma中的矫正算法
  - TMM（edgeR软件中的算法）
  - 几何平均数（Deseq软件中的算法）
- Cufflinks (PE reads GTF)
- 算法 (RSEM ,eXpress , Sailfish and kallisto.)
  - 算法Sailfish计算reads的k-mer值不需要比对，其他定量算法都需要依赖于比对到每个转录本上的reads数；
- NURD（SE 低内存，计算量小）

3.1 转录本分析-表达定量

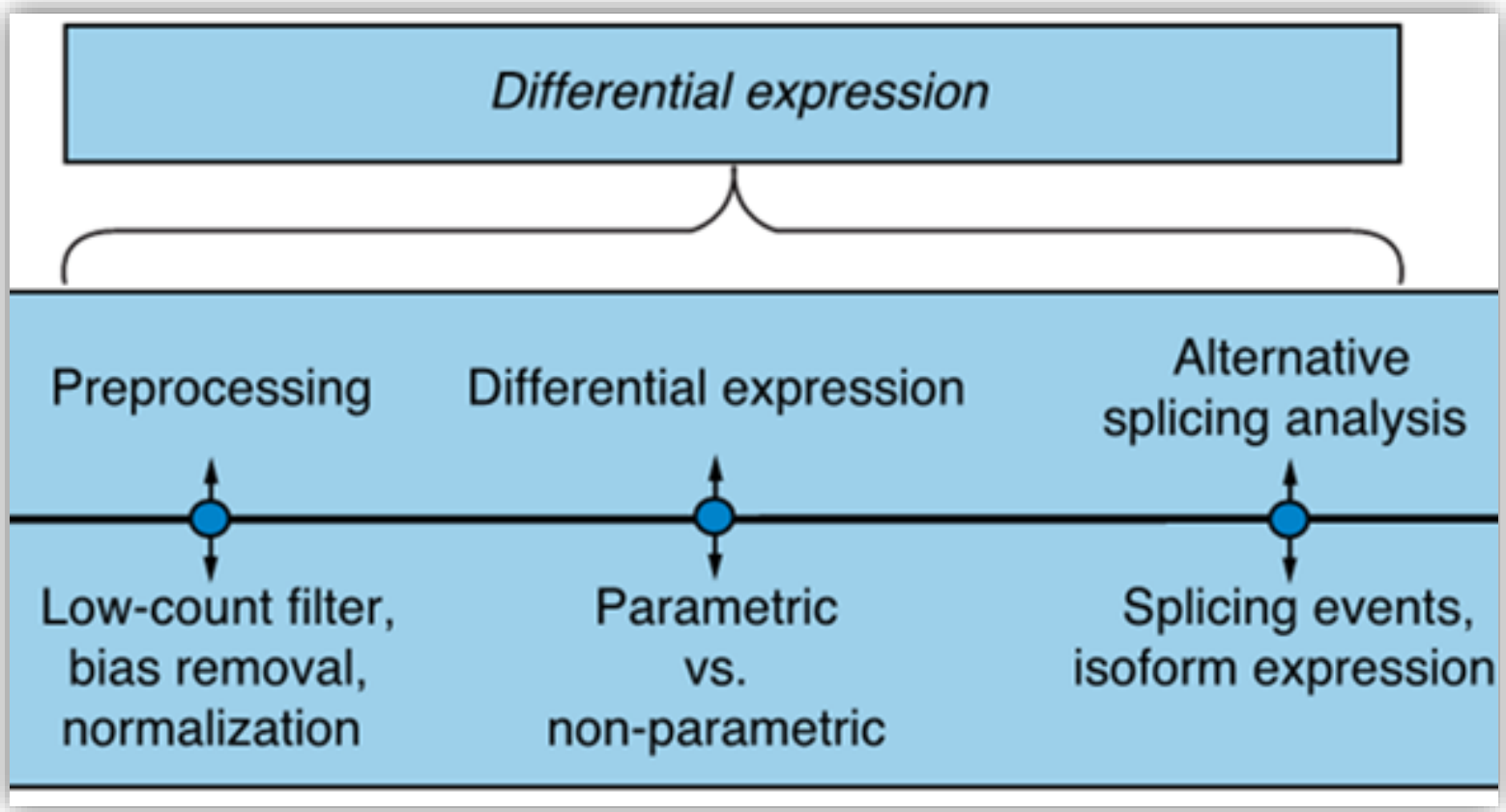
- 先说结论：
- 学术界已经不再推荐RPKM、FPKM；
- 比较基因的表达丰度，例如哪个基因在哪个组织里高表达，用TPM做均一化处理；
- 不同组间比较，找差异基因，先得到read counts，然后用DESeq2或edgeR，做均一化和差异基因筛选；如果对比某个基因的KO组和对照，推荐DESeq2。

RPKM, FPKM and TPM

Stat

3.2 差异表达分析

通过RNA-seq得到最重要的数据是基因/转录本在各个样品中的表达量。



Q：

- 表达量低的基因测不到？
- 基因的表达量怎么算？
- 同一个基因在不同样品间比较
- 不同基因在同一样品比较
- 怎样才算有差异？

3.2 差异表达分析- RNA-seq数据概率分布

混合分布

- 技术误差：泊松分布
- 生物差异：伽马分布
- 那么生物学重复样本间的误差分布则符合：

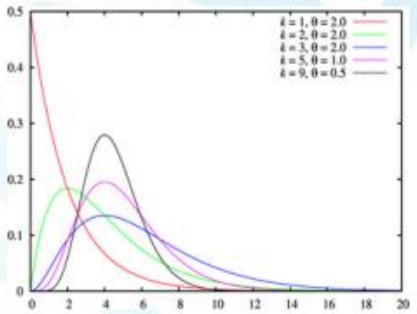
泊松分布 + 伽马分布 = 负二项分布

$$v = \mu + \alpha\mu^2$$

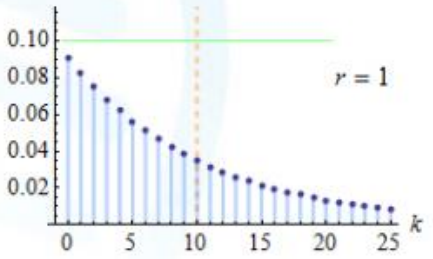
注： v，方差； μ，均值；α，散度因子

例如差异分析软件edgeR、Deseq都使用负二项分布，但对散度因子的估算方式并不相同。

伽马分布



负二项分布



这种算法考虑了低表达量和重复数少的情况；

差异基因分析

- 参数方法:
  - **edgeR** , take as input raw read counts and introduce possible bias sources into the statistical model to perform an integrated normalization as well as a differential expression analysis
  - **DESeq2**, like edgeR, uses the negative binomial as the reference distribution and provides its own normalization approach
  - **baySeq and EBSeq** are Bayesian approaches, also based on the negative binomial model
- 非参数方法: NOISeq or SAMseq
- 两个样本比较或无重复：DEGseq(泊松分布)；NOISeq(经验分布)
- 建议：要通读软件的说明文件，并且多用几个软件来分析，比较一下。

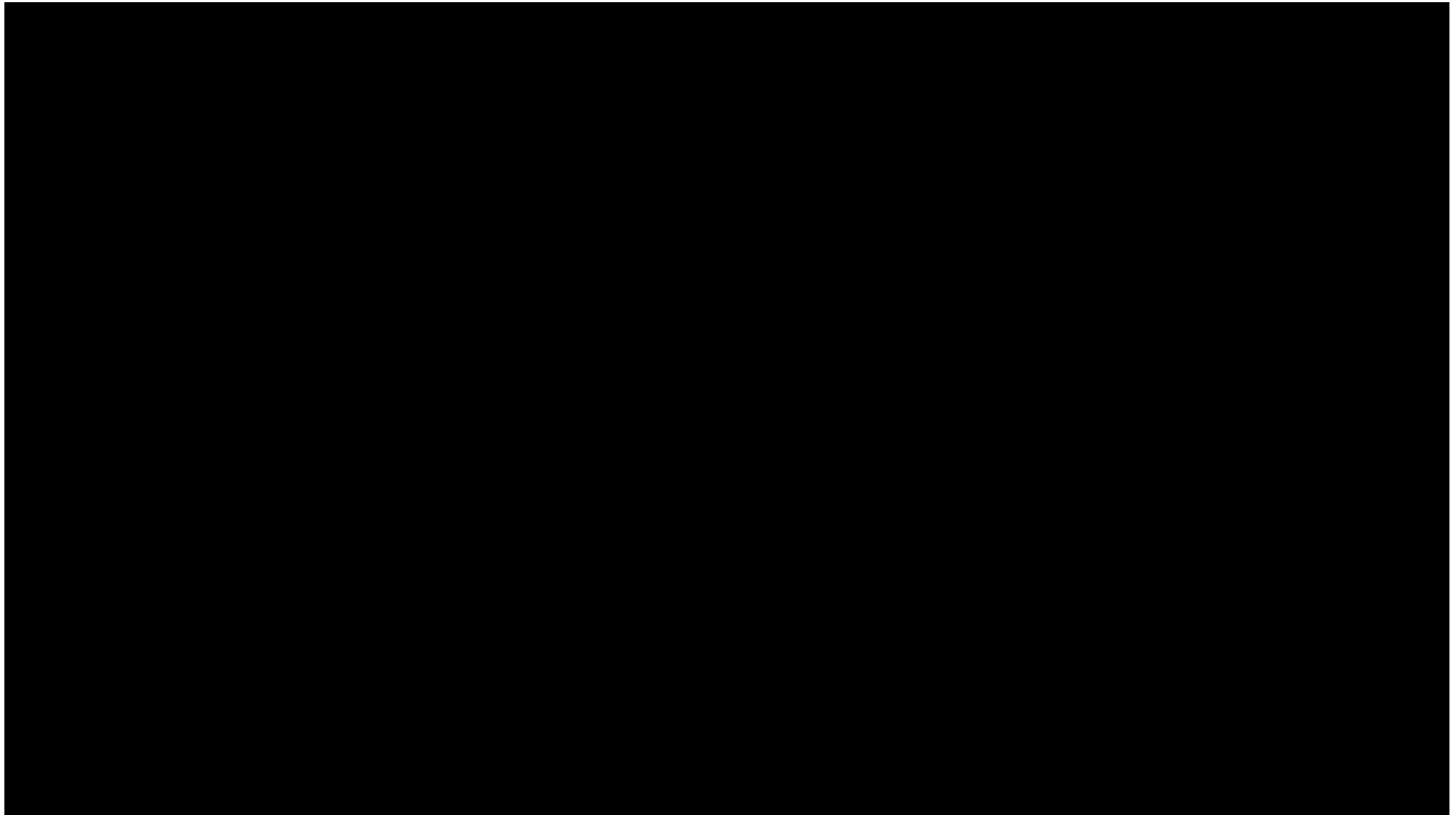
StatQuest DESeq2, part 1, Library Normalization

StatQuest!!!!!

StatQuest edgeR, part 1, Library Normalization

StatQuest!!!!!!!!!!!!!!!!!!!!!!!!

StatQuest edgeR and DESeq2, part 2 - Independent Filtering

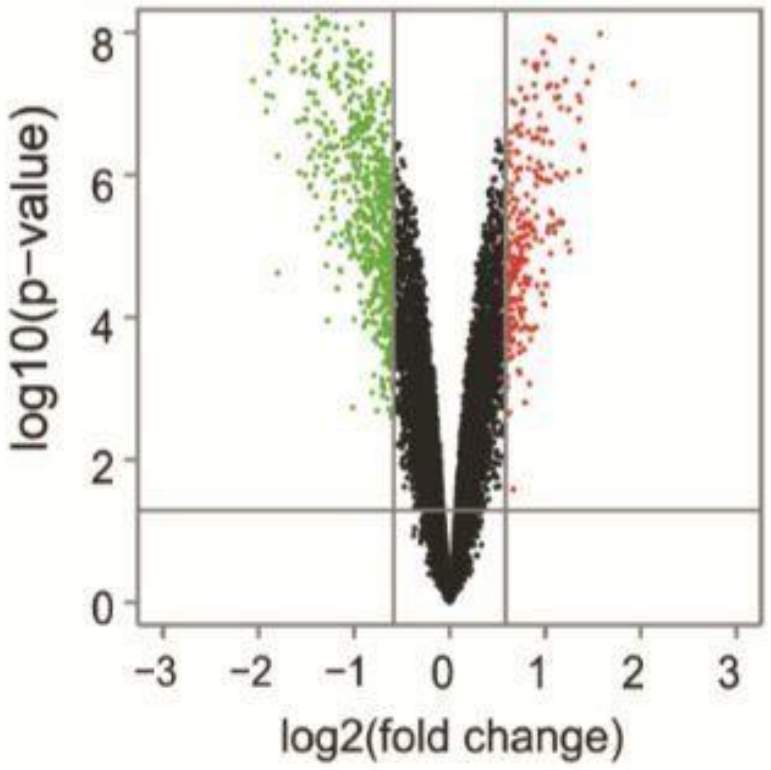


StatQuest FDR and the Benjamini-Hochberg Method clearly explained

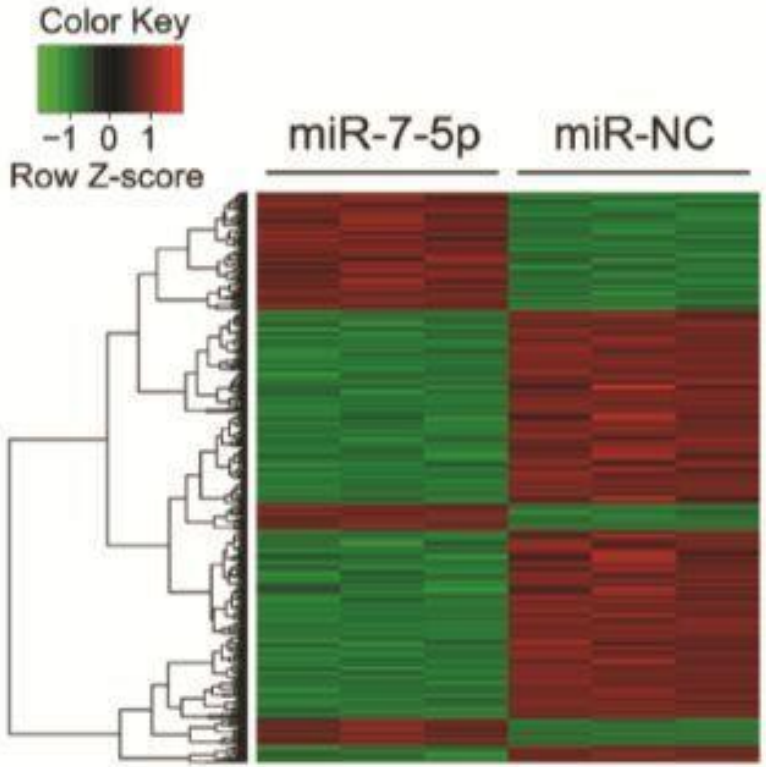
Holy Freaking Smokes!!!!

StatQuest Logs (logarithms), clearly explained

A



B



C

Gene	FC	p-value
CTSK	-6.247	3.57×10^{-11}
IRS2	-1.509	2.69×10^{-5}
PAK1	-1.631	1.74×10^{-6}
POLE4	-6.861	1.77×10^{-9}
RAF1	-2.594	1.77×10^{-7}
RELA	-1.588	3.91×10^{-6}
SMO	-1.533	1.57×10^{-6}
SP1	-1.566	3.30×10^{-4}
STMN3	-1.928	7.02×10^{-6}
TGFA	-1.553	2.42×10^{-5}

火山图

热图（聚类分析）

StatQuest_ Heatmaps - considerations for drawing and interpreting them

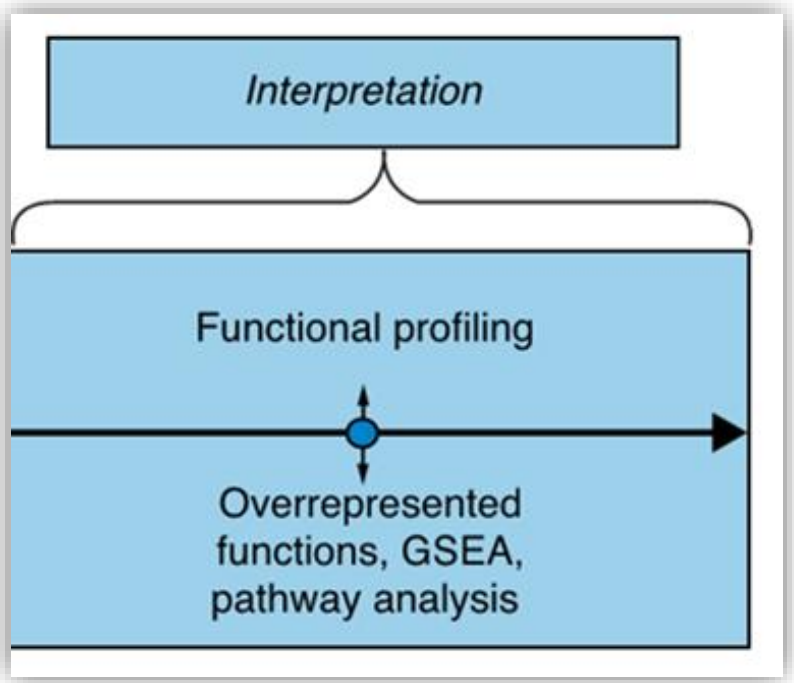
3.2 差异表达分析-可变剪接与差异可变剪接

- exon or junction methods
  - BASIS：使用多水平贝叶斯模型直接推断不同亚型转录本的差异表达；
  - CuffDiff2：先估计不同亚型转录本的表达量，再比较差异；
  - rSeqDiff：可以同时检测有可变剪接和无可变剪接的两种情况的基因差异表达；
- ‘exon-based’ approach
  - DEXseq and DSGSeq
  - rMATS
  - rDiff
  - DiffSplice

    exon or junction methods is their greater accuracy in identifying individual alternative splicing events

    Exon based methods are appropriate if the focus of the study is not on whole isoforms but on the inclusion and exclusion of specific exons and the functional protein domains (or regulatory features, in case of untranslated region exons) that they contain.

3.3 功能分析



新转录本注释：
SwissProt
Pfam
InterPto
NR
Rfam

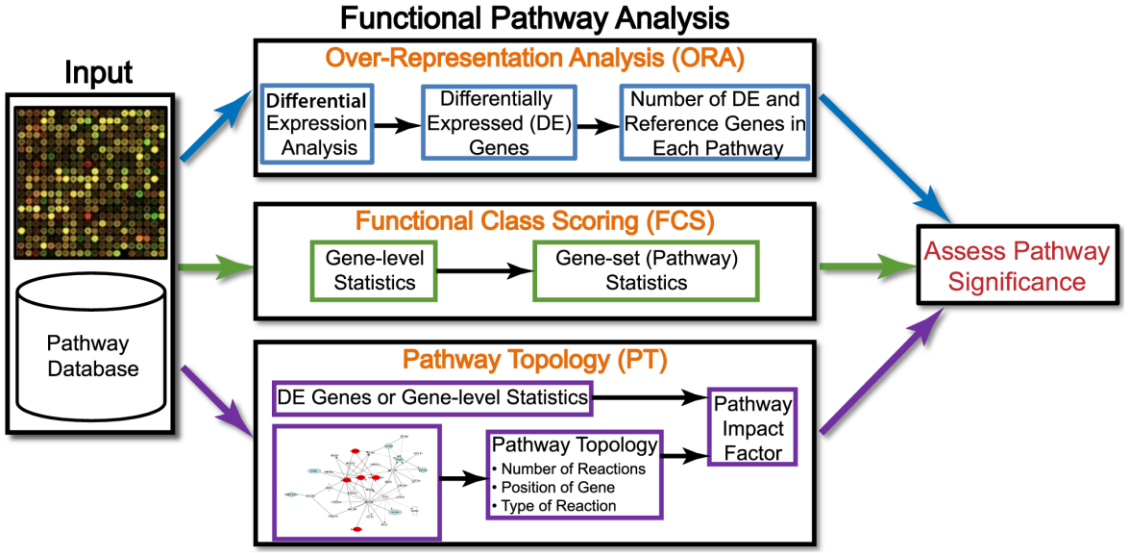
非编码RNA的注释：
miRbase
Miranda

三种功能分析的方法：

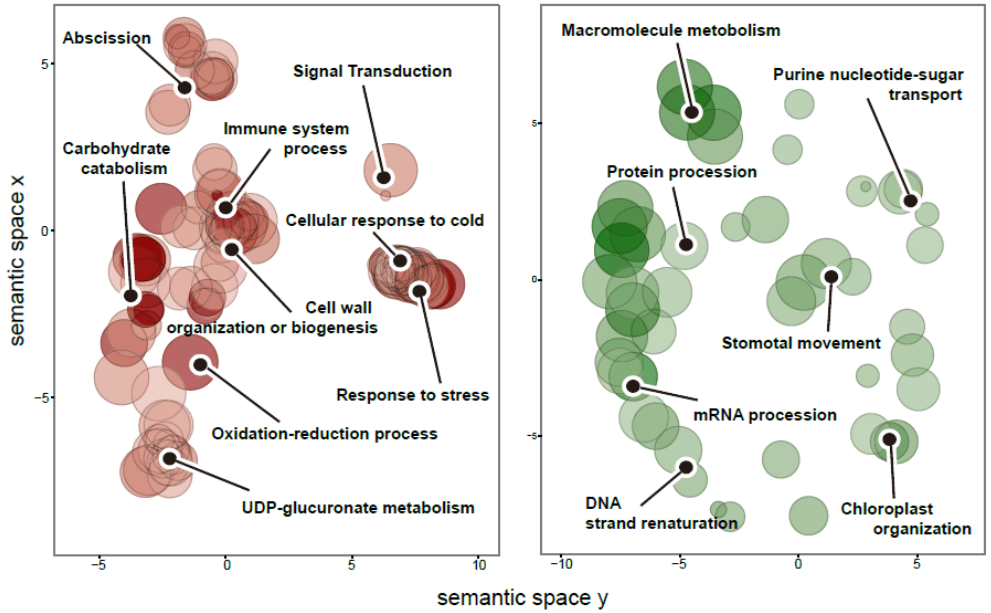
(a) **comparing a list of DEGs against the rest of the genome for overrepresented functions（常见的差异基因GO/KEGG功能富集分析）**

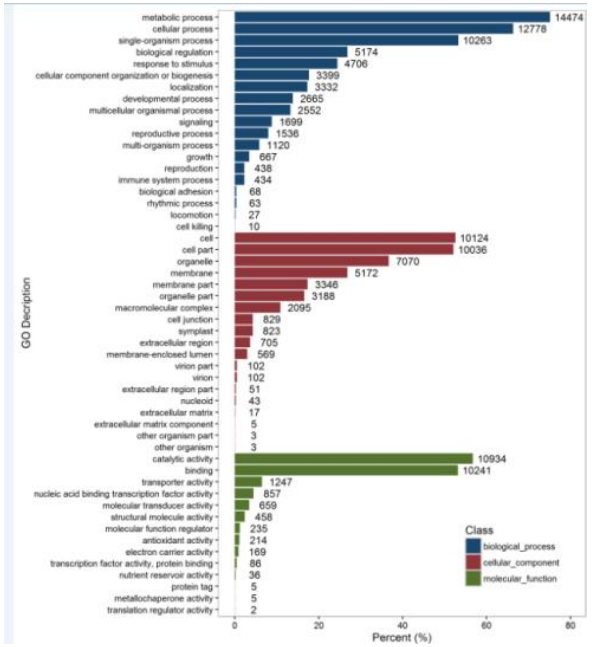
(b) gene set enrichment analysis (GSEA)

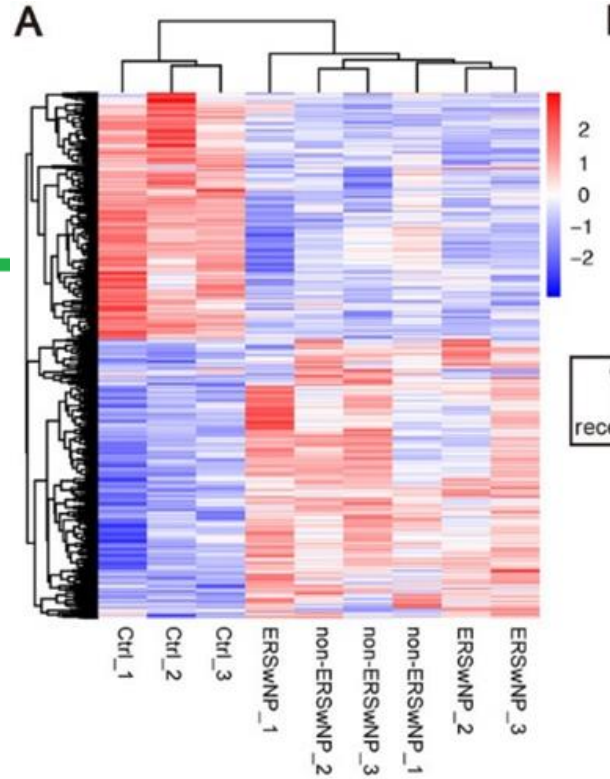
(c) Pathway analysis

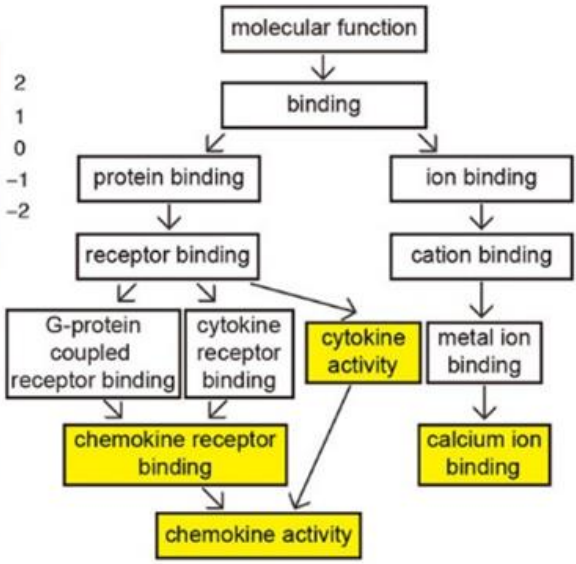


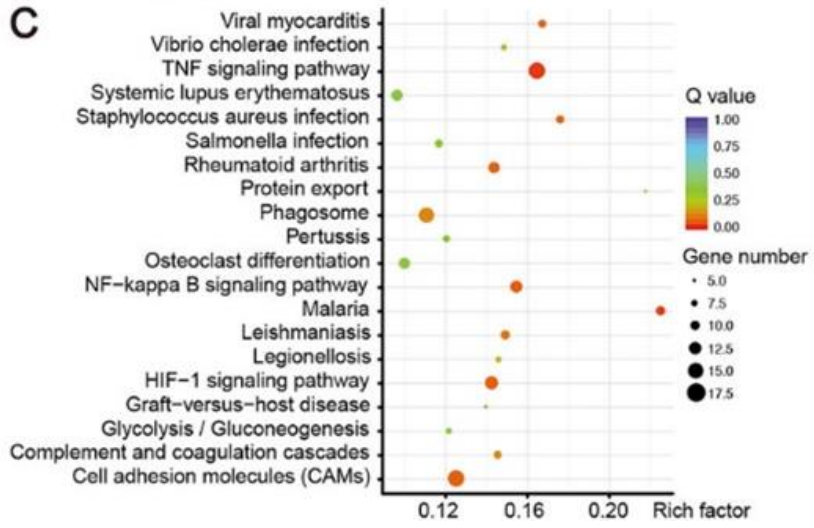
Zhang Y, Lameijer E W, Pa T H, et al. PASSion: a pattern growth algorithm-based pipeline for splice junction detection in paired-end RNA-Seq data.[J]. Bioinformatics, 2012, 28(4):479-486.





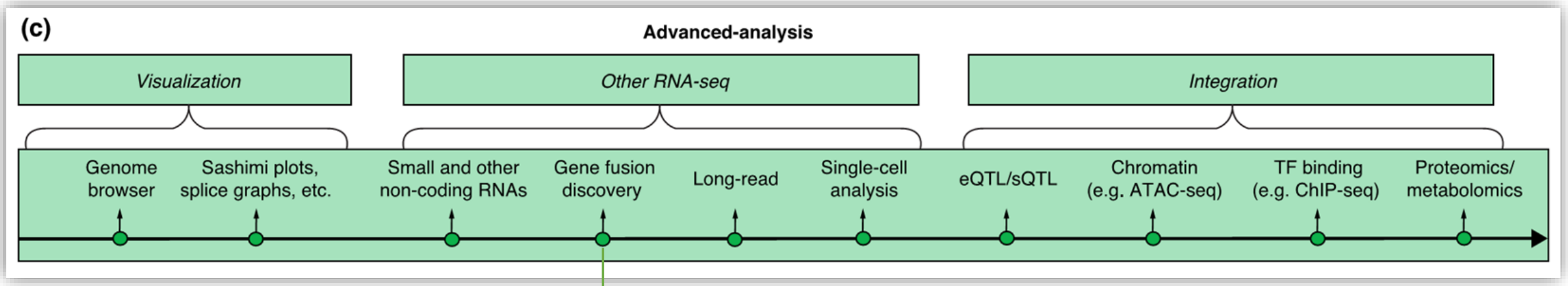






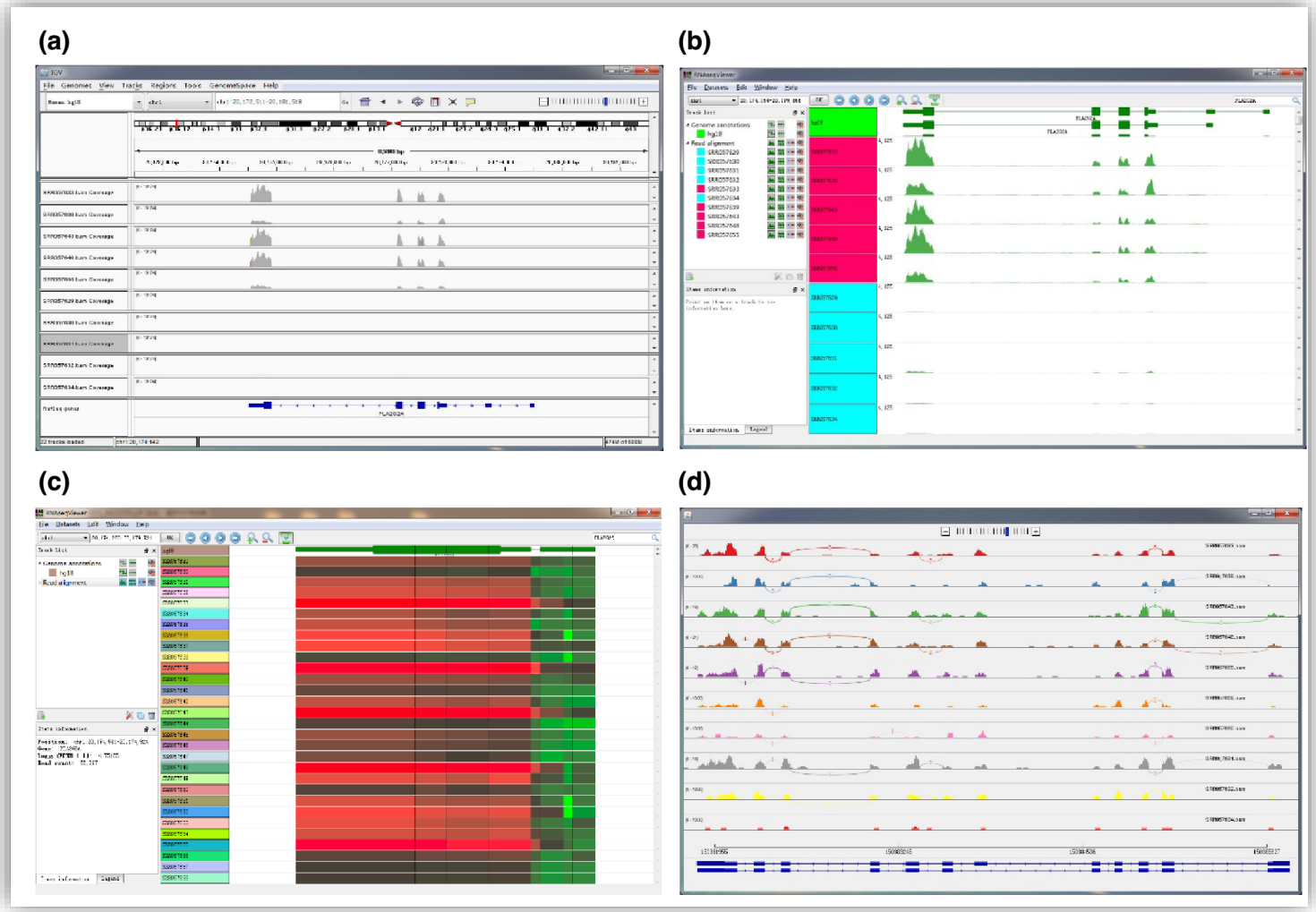
GO有向无环图
KEGG代谢通路富集分析
共表达分析 ……

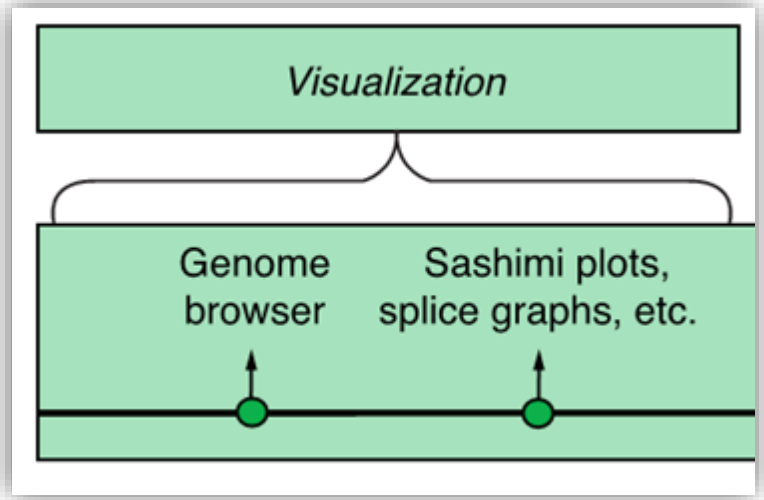
4 其他分析(Advanced-analysis)



我是如何学习Gene Fusion分析的
http://mp.weixin.qq.com/s/4U9EPD9Jw2vNH0S9gqedoA

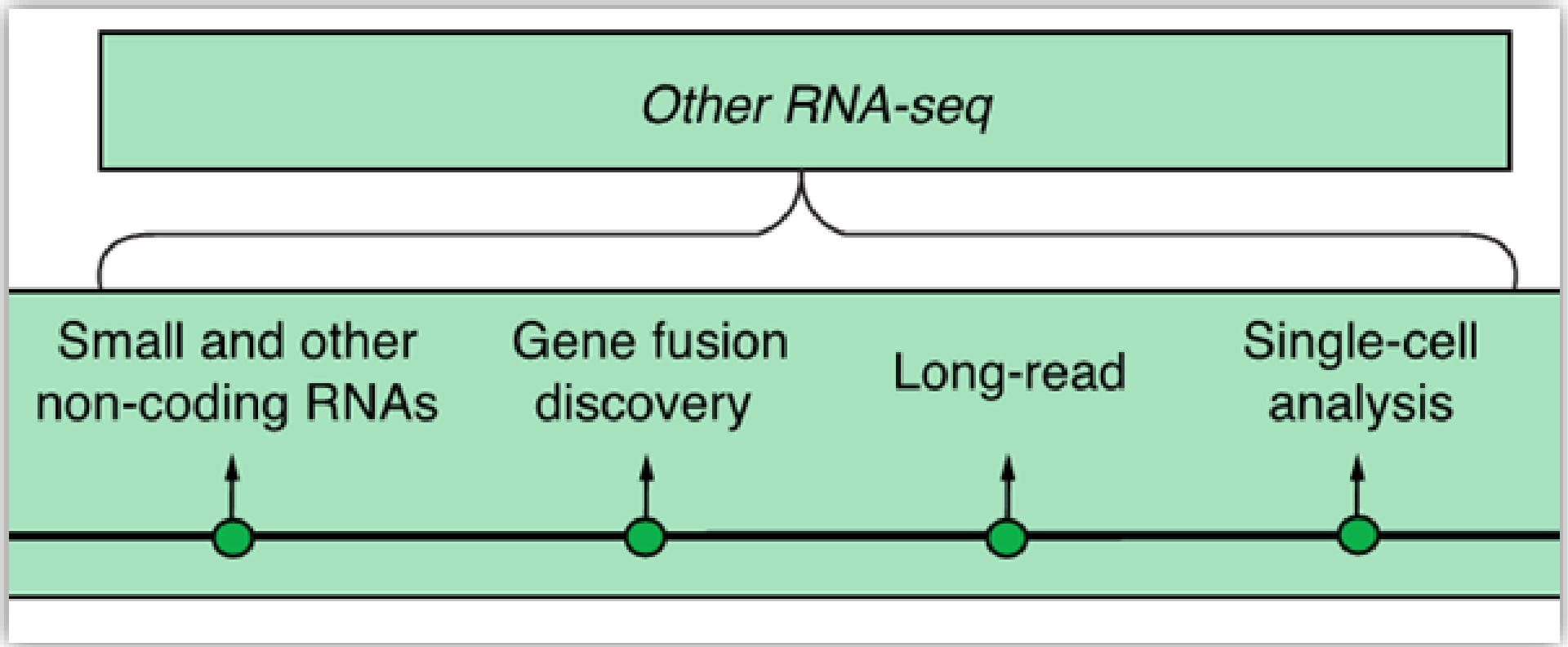
4.1 可视化



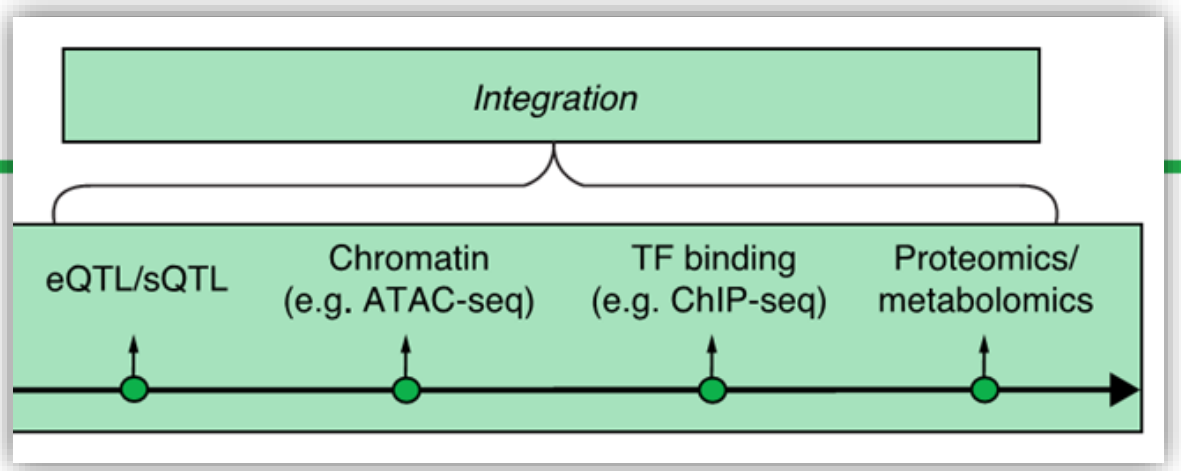


ReadXplorer,UCSC browser,
Integrative Genomics Viewer (IGV) ,
Genome Maps or Savant
RNAseqViewer
DESeq2 or DEXseq
CummeRbund(for CuffDiff) or Sashimi plots
SplicePlot
SpliceSeq, SplicingViewer
TraV

4.2 其他RNA-seq应用



4.3 多种数据整合分析



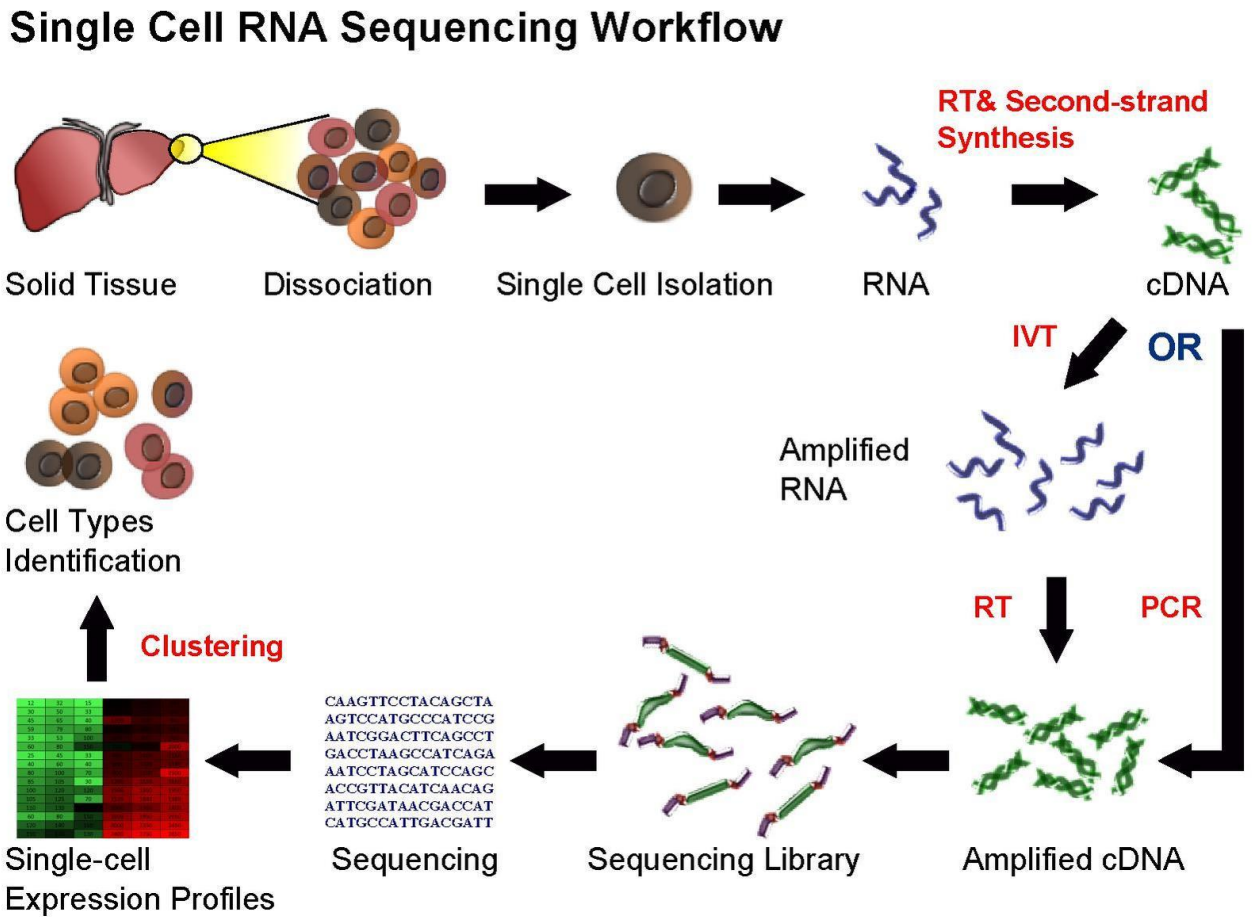
- DNA sequencing
  - 发现SNP位点
  - 研究RNA编辑
  - 定位表达数量性状基因座
- DNA methylation
  - DEGs and methylation pattern
  - The statistically significant correlations that were observed, however, accounted for relatively small effects.
- Chromatin features
  - RNA-seq and transcription factor (TF) chromatin immunoprecipitation sequencing (ChIPseq) data
- MicroRNAs
  - 具有挑战性
- Proteomics and metabolomics
  - Proteomics：low correlation (~0.40)
  - Metabolomics: identify pathways that are regulated at both the gene expression and the metabolite level

4.3 多种数据整合分析

- Integration and visualization of multiple data types
  - 分析技术不成熟
  - 软件：
    - SNMNMF and PIMiM：mRNA and miRNA 加 protein–protein, DNA–protein, and miRNA–mRNA（识别miRNA-gene调控模型）
    - MONA：mRNA,miRNA, DNA methylation, and proteomics data
    - Paintomics：任何组学数据的代谢通路分析
    - 3Omics：transcriptomics, metabolomics and proteomics data
    - Anduril , Galaxy and Chipster

5 展望

- Single-cell RNA-seq



- Long-read sequencing

Pacbio 技术可以直接测得接近全长的转录本，可以有效解决二代测序技术拼接较为零碎以及潜在嵌合拼接的问题；

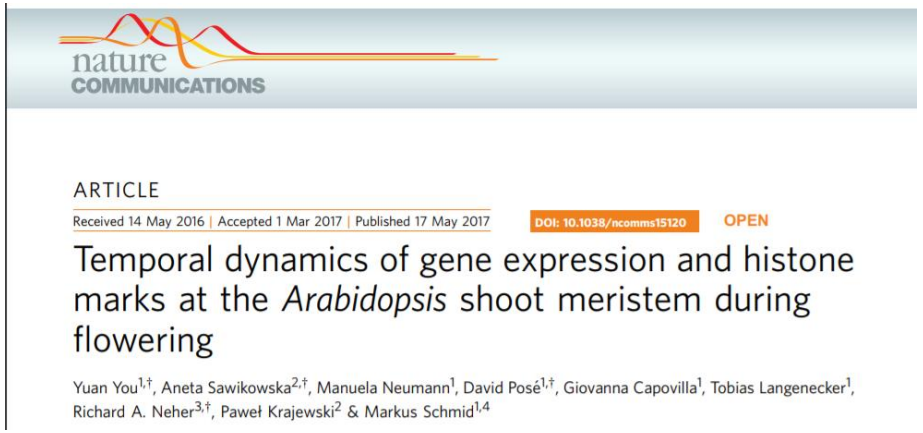
目前的瓶颈：价格（建库价格和测序价格）

（1）需要多种长度的文库；

（2）测序通量有限；

实战

- 项目一：https://github.com/twbattaglia/RNAseq-workflow
- 项目二：http://www.bio-info-trainee.com/2809.html



- 项目三：http://www.biotrainee.com/thread-1750-1-1.html

本节结束

生信技能树

Biotrainee.com