和趣求真 乘实生信

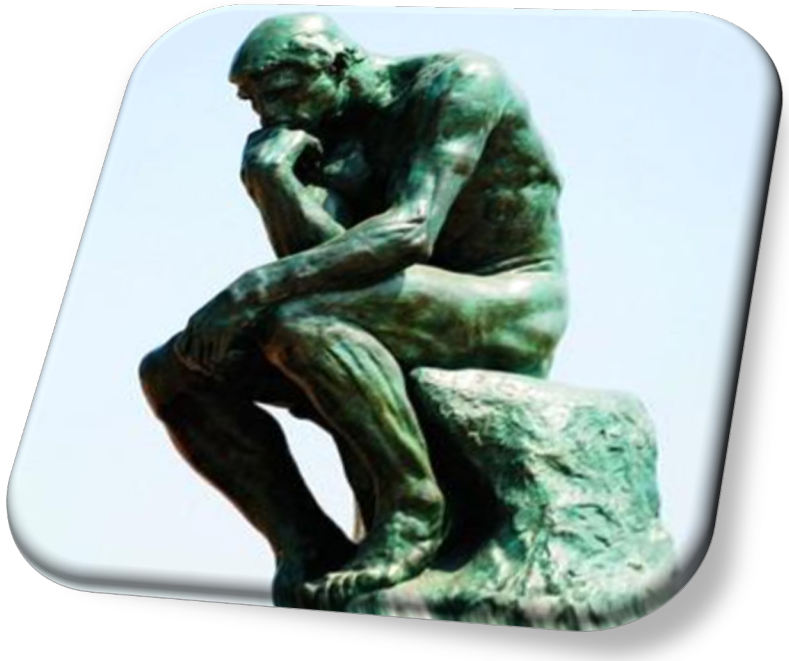
RNA-Seq 分析流程

ZhangLab-小课堂 第一期（试运行）

谢文召

2022.03.04

小课堂因何而来？



时常感到自己一个人的力量太有限了？

时常感到思考会片面，打不开思路？

想了解学习某个板块，却奈没有经验停滞不前？

亦或许，拥有一颗期待发光的心，缺乏舞台展示？

小课堂因何而来？

与人交谈一次，往往比多年闭门劳作更能启发心智。思想必定是在与人交往中产生，而在孤独中进行加工和表达。

——列夫·托尔斯泰

让知识在幸福的海洋中徜徉

（今天是值得纪念的一天）

小课堂Team

Leader | 组长

Postdoctoral fellows | 博士后

Dr. LI/Huan (李环)
Genomics

2019级博士生

ZHOU/Run (周润)

2020级博士生

HUANG/Yicheng (黄臆丞)

XIE/Wenzhao (谢文召)

2021级博士生

YU/Zhichao (于志超)

2019级硕士生

ZHANG/Wenhui (张文慧)

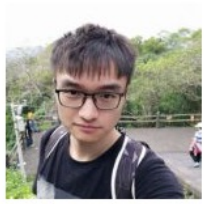
2020级硕士生

GAO/Tingting (高婷婷)

HAN/Wanxin (韩婉欣)

LI/Mengyuan (李梦圆)

LIU/Sishi (刘思诗)



ZHAO/Zhiyuan (赵志远)

LI/Yan (李岩)

2021级硕士生

ZHANG/Yulu (张雨露)

LI/Shanying (李姗莹)

CHENG/Rundong (程润东)

?

Undergraduate Students | 本科生

HU/Xinyue (胡馨月), 2019级

LIU/Jian (刘健), 2019级

小课堂运行模式

和趣求真 Z 秉实生信

活动安排：

1）前期以ZhangLab为主（年级由高往低顺延），轮流打怪；

2）后期按比例邀请主讲人（比如1/4青椒、 1/4博后、1/4博士、1/4硕士等）；

3）暂定每月的第一个周五下午3：00-3：40；

活动主题：

1）走专刊形式，例1月份基因组专题，2月份转录组专题，3月份python教学专题等；

2）附兴趣板块，不断收集听众感兴趣、想了解的模块，收纳总结，做成专场；

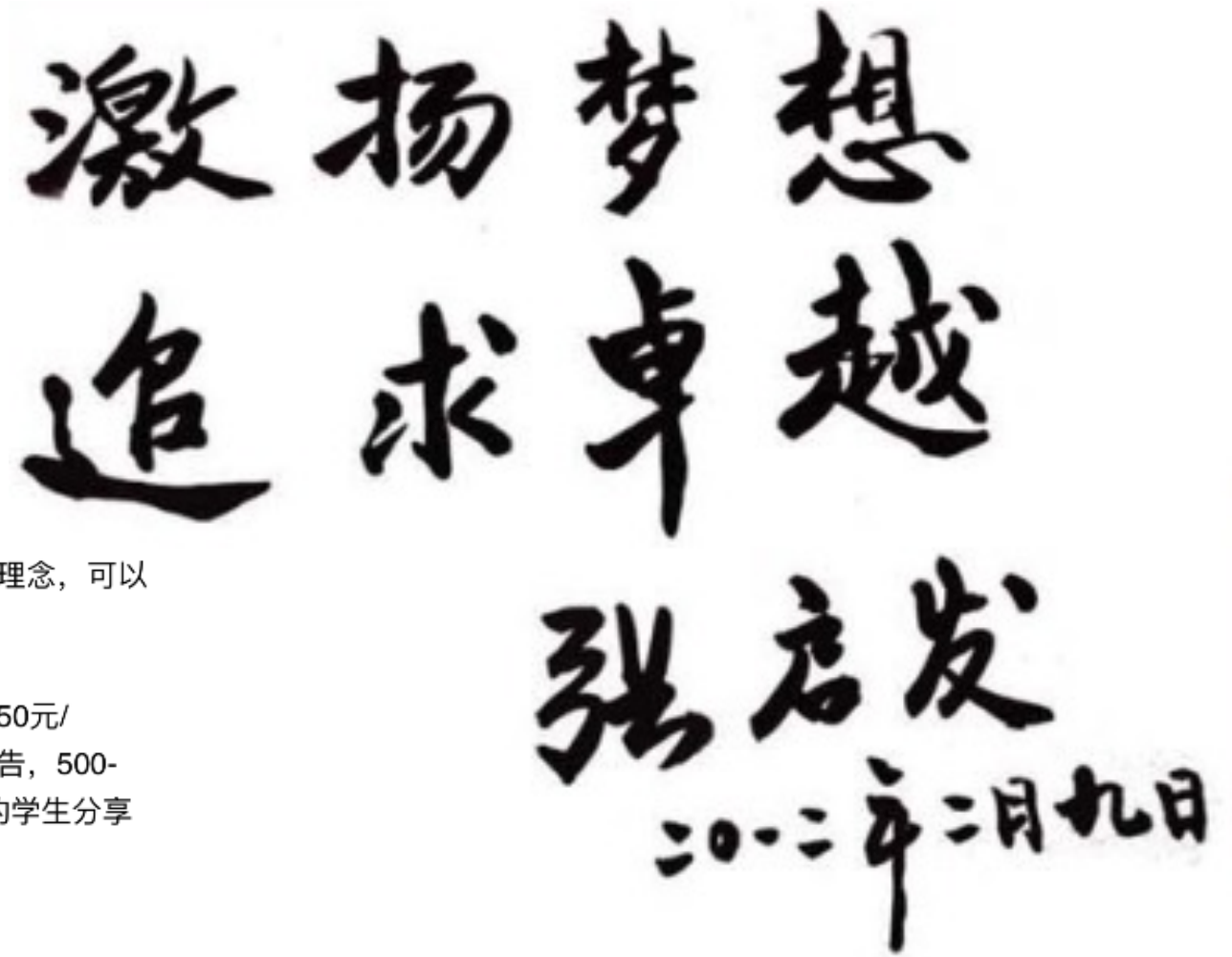
3）随走随做，灵感来了怎能让默默其流失？

活动宣传：

1）海报设计？ 2）Logo设计？ 3）公众号/网站定期推送？

以Zhang Lab为点，连接二综/华农为线，最终铺成华农/华中/全国农林生信/的面。

您将收获什么？



建议结合我们课题组“和趣求真秉实生信”的理念，可以分三类或层次：

“和趣生信”（课题组员分享生信小技巧等，50元/次），“求真生信”（邀请生信大咖做前沿报告，500-1000元/次），“秉实生信”（由发表过论文的学生分享具体课题实战经验，200元/次）。

诸如此类的细节请大家一起谈论。

大主题的名称还需考虑，发挥你们的想象力和创新力吧！

RNA-Seq 分析流程

谢文召

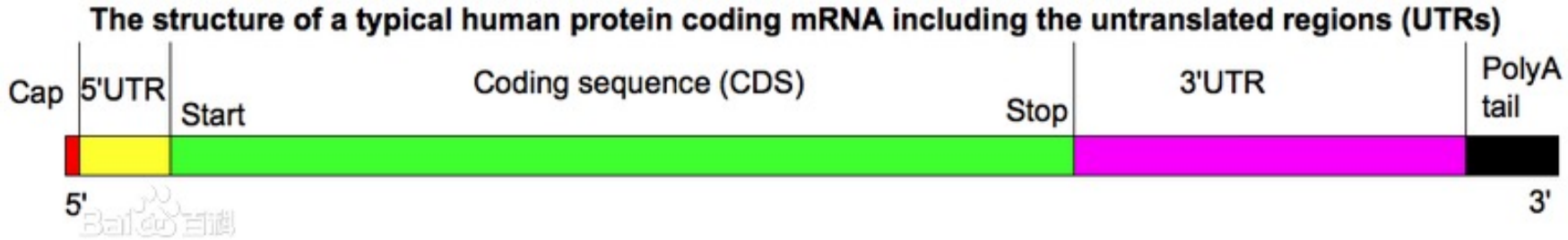
2022.03.04

0) RNA Classification.

真核生物中，最初转录生成hnRNA, 其多属mRNA前体，25%左右可最终加工成mRNA。

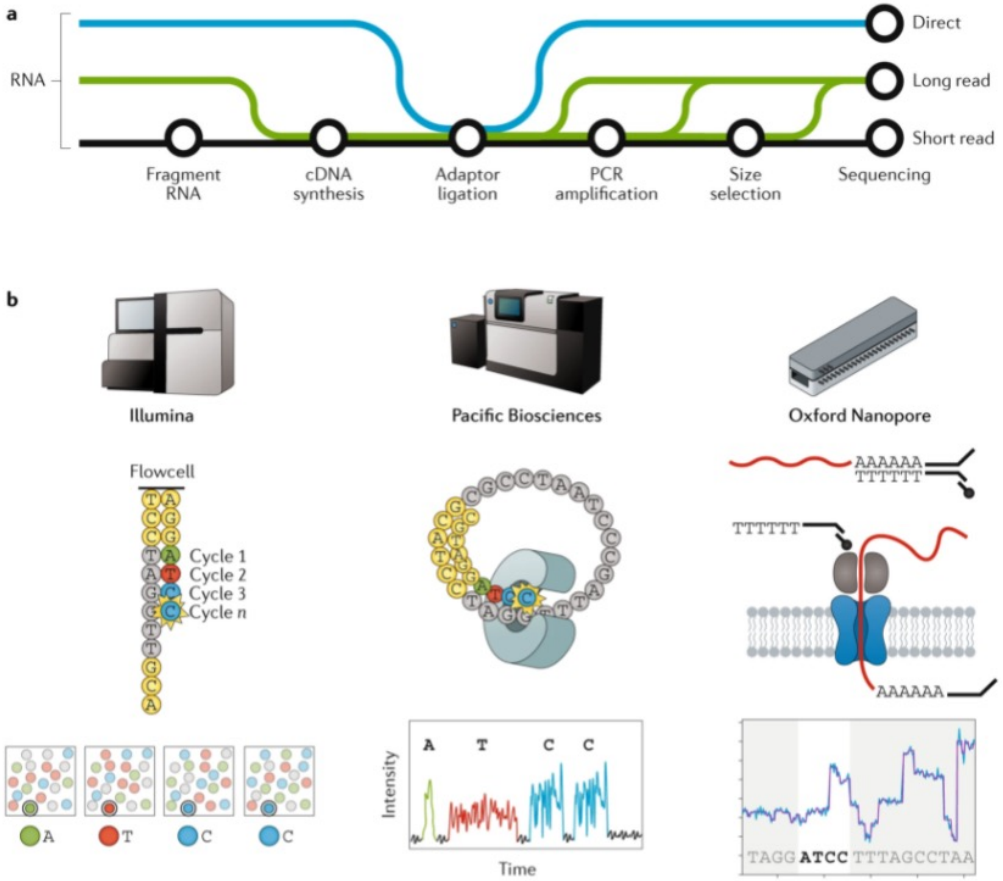
RNA

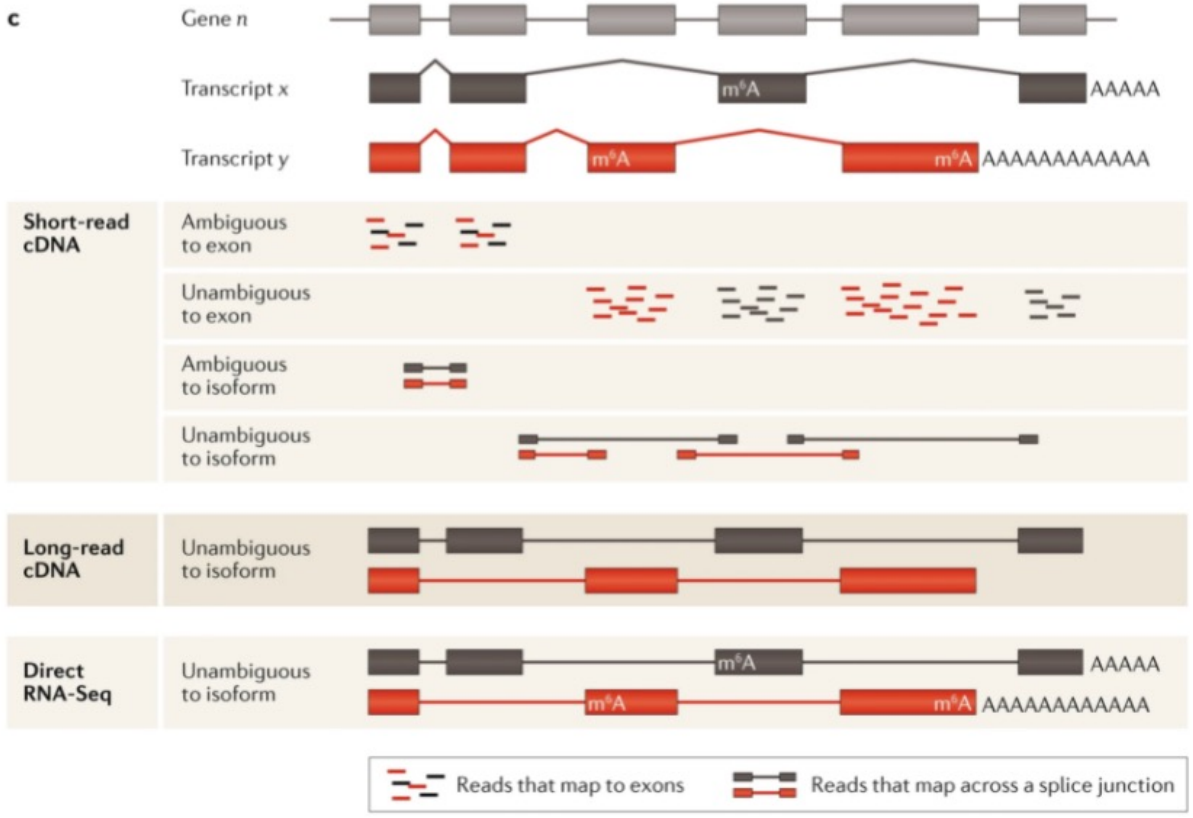
- hnRNA+**mRNA**



- tRNA：转运RNA。 负责把氨基酸搬运到核糖体上
- rRNA：核糖体RNA（ribosomalRNA）， 是组成核糖体的主要成分
- miRNA
- 小分子RNA：在真核生物中发现的一类具有调控功能的非编码RNA, 其大小长约20~25个核苷酸。
- 端粒酶RNA：存在于真核生物细胞核和细胞质中， 长度为100到300个碱基
- 反义RNA：TERC是端粒酶的一部分， 作为端粒继续延伸的模板， 由端粒酶催化实现端粒的延长 。
- lncRNA：可与mRNA互补配对的单链RNA分子。可与mRNA发生互补配对， 抑制mRNA的翻译
- ncRNA：长链非编码RNA(Long non-coding RNA, lncRNA)， 长度大于 200 个核苷酸的非编码 RNA。

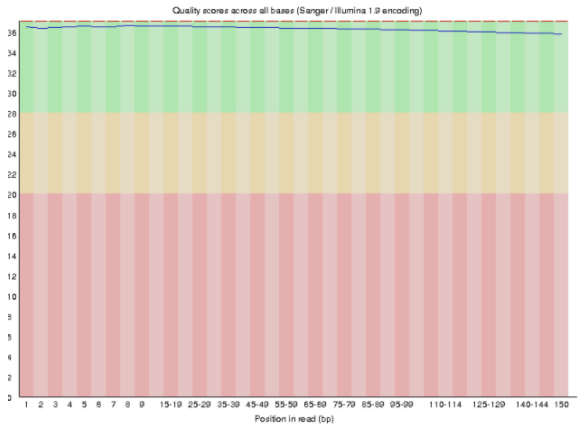
1) RNA Sequence.



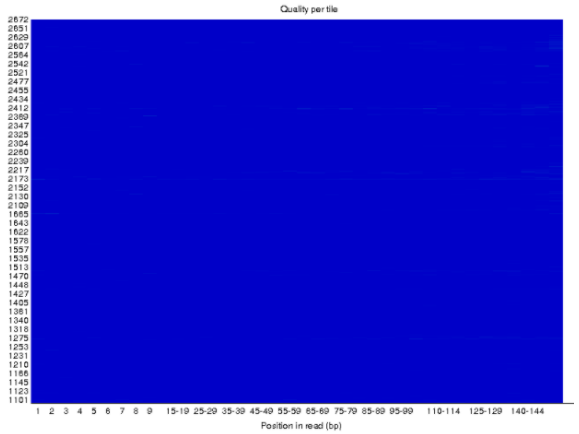


short-read cDNA测序用于差异基因分析

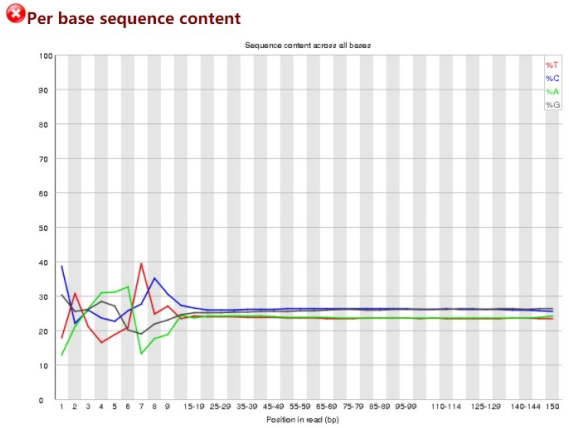
2) **RNA FQ** fastp、fastx-Toolkit、trimmomatic



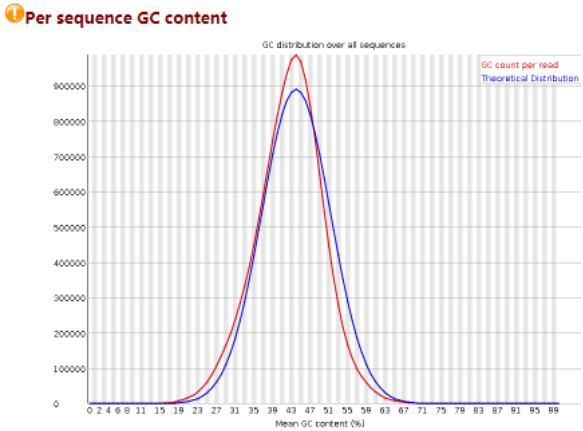
150bp每个碱基位置的质量值统计



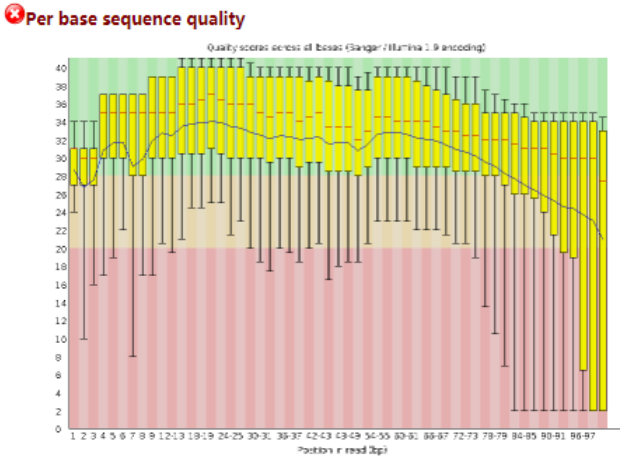
整体质量

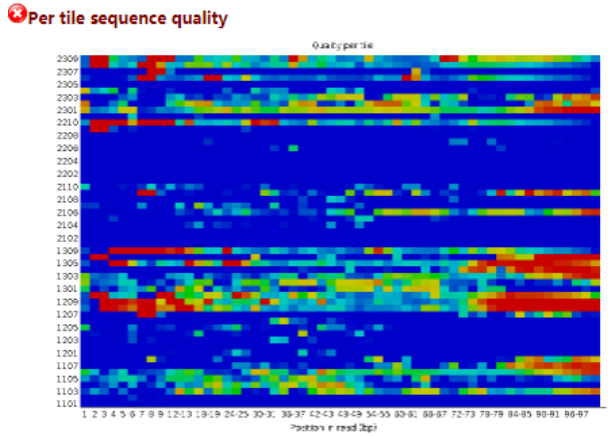


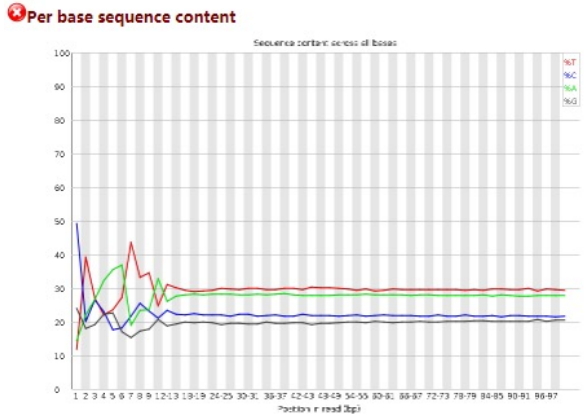
从第一个到最后一个碱基位置上ATGC含量，需要切除前端不稳定的序列

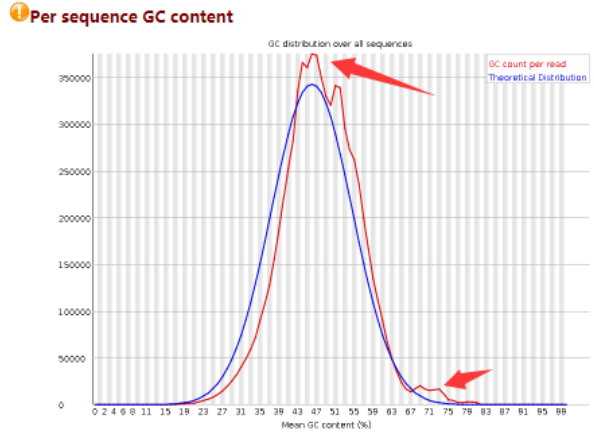


GC含量分布图，GC曲线平滑说明数据干净









3) RNA Correlation

在我们正式进行转录组数据分析之前，需要先对组内生物学重复（一般设置3个生物学重复）进行样本关系分析，判断组内重复性效果的好坏，是否有离群样本。应广大研究者之需，本期针对大家比较关心的**样本重复性**问题进行探讨，力争为各位老师在科研之路上带来帮助。

在进行问题讨论之前，首先我们对可能会困扰大家的**关于什么是生物学重复和技术学重复**的问题进行区分。

①生物学重复：

指同一处理下不同的生物学样品。由于遗传和环境等因素的影响会引起生物体的个体差异，因此需要采用生物重复的实验设计方法来降低该差异。一般的实验设计中，都会包括实验组和对照组。如下图A实验组包含3只小鼠，那么这3只小鼠，经过相同的实验处理，分别测组织的RNA-seq，即为一组生物学重复。

②技术重复：

简单来说就是对同一生物体样品进行重复地检测。如下图B、C，都属于技术重复。对于第一种技术重复，重点是检测RNA-seq方法的准确度。比如当发现了一个新的检测基因表达量的方法，就需要用这种重复来验证（图1 B）；第二种技术重复重点是这个小鼠本身的基因表达水平（图1 C）。

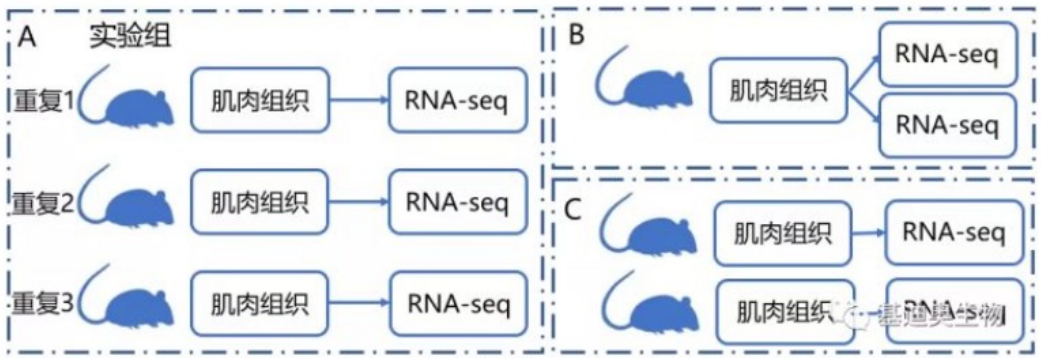
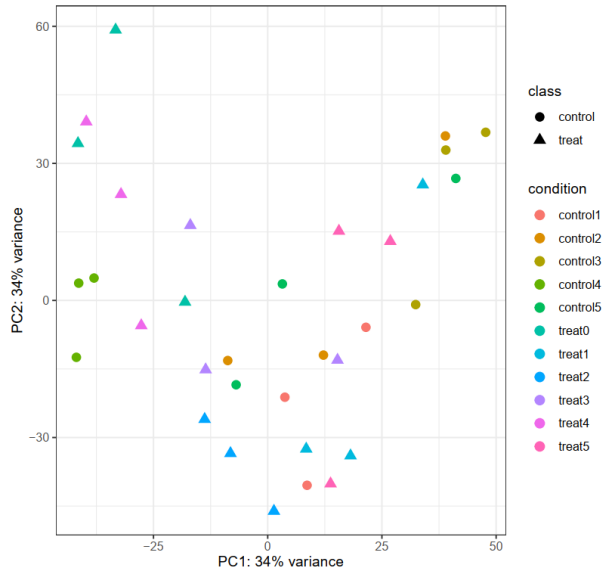
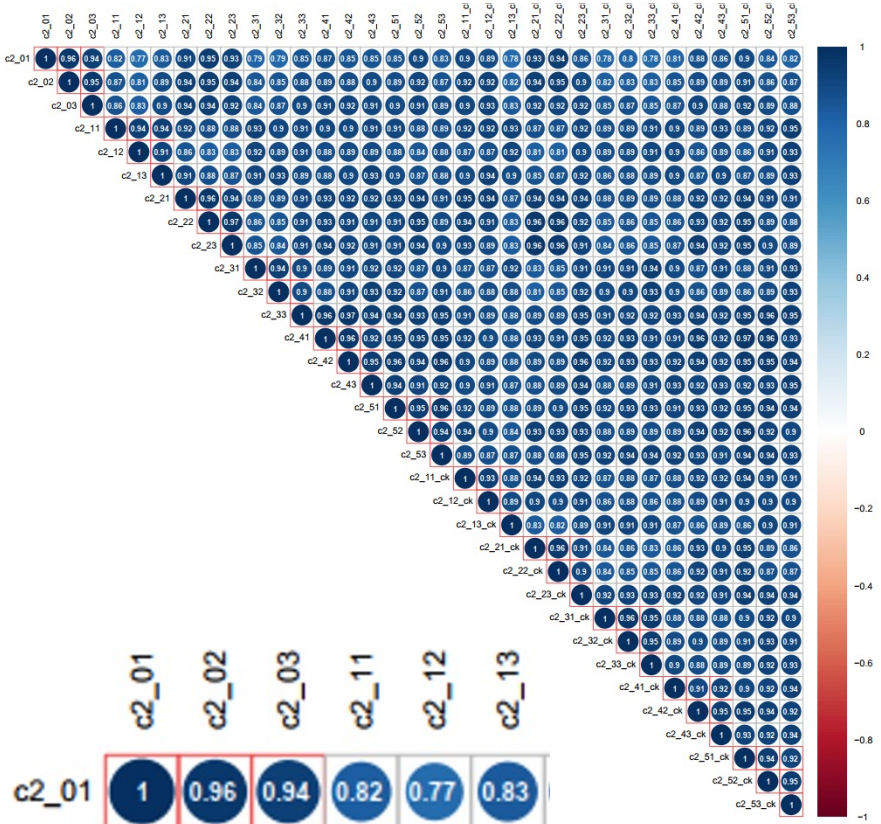


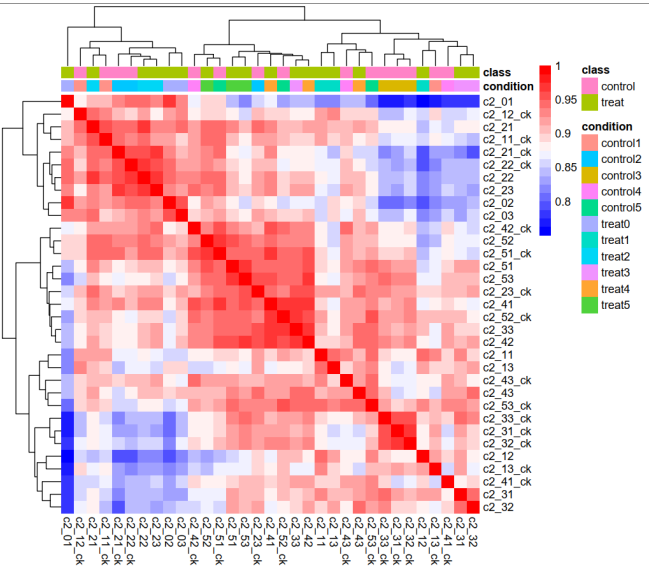
图1 生物学重复和技术重复

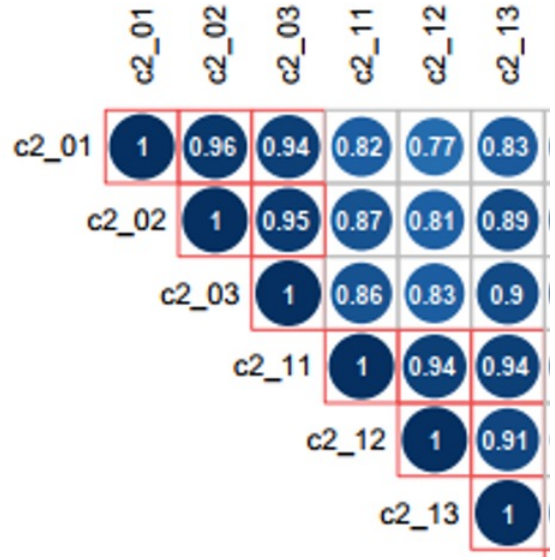
>=3

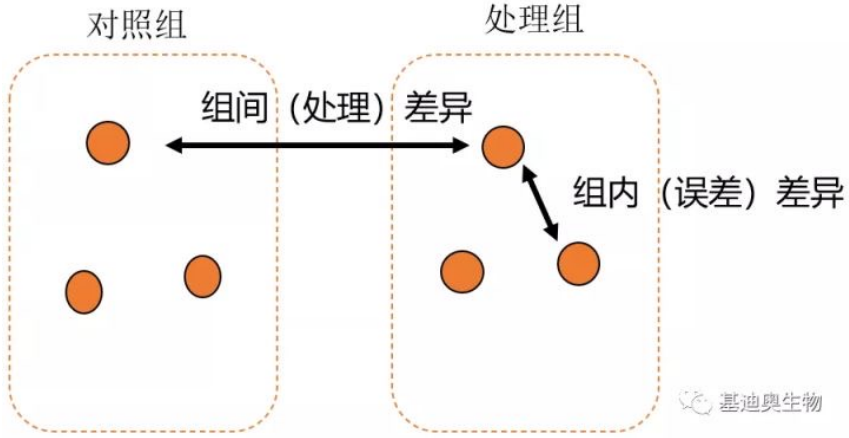
3) RNA Correlation











生物学重复间的R2建议在0.7以上；
技术重复之间的R2建议在0.85以上；

4) RNA Map

比对之前，需要了解比对的目的是什么？RNA-Seq数据比对和DNA-Seq数据比对有什么差异？

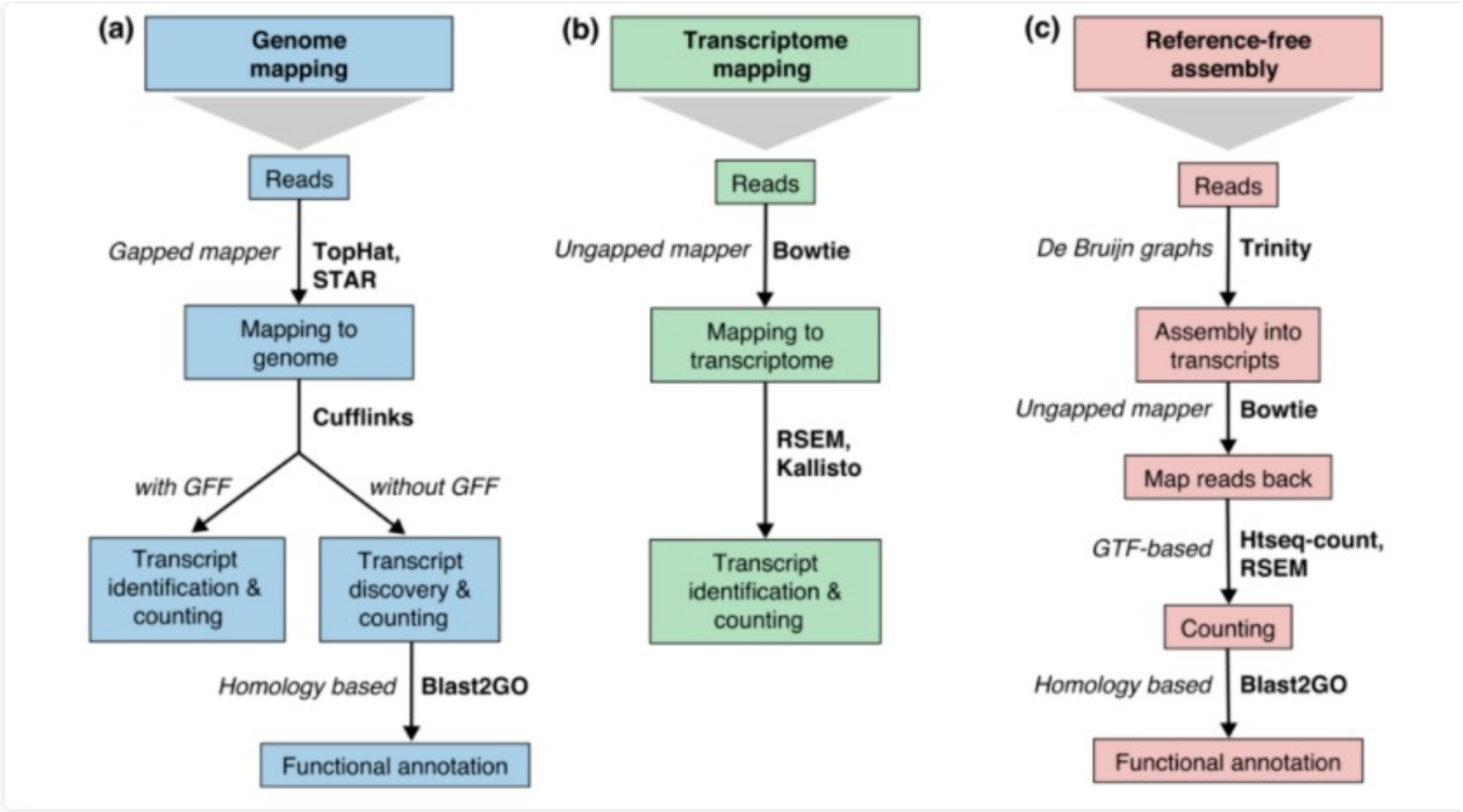
1）如果是找差异表达基因，只需要确定不同的read计数就行的话，可以用bowtie, bwa这类比对工具，或者是salmon这类align-free工具，并且后者的速度更快;

2）如果是找新的isoform，或RNA的可变剪切，就需要TopHat, HISAT2或者是STAR这类工具用于找到剪切位点。

3）比对到基因组：使用间隔比对算法，如 TopHat、STAR等，然后根据是否提供了注释文件（GFF格式文件，包含转录本位置信息），又可以分为转录本识别和转录本发现并进行定量分析。

4）比对到转录组：使用非间隔比对算法，如Bowtie等，然后使用RSEM或Kallisto方法识别转录本并计算定量信息。

4) RNA Map


(a)
Genome mapping
Reads
Gapped mapper
TopHat, STAR
Mapping to genome
Cufflinks
with GFF
without GFF
Transcript identification & counting
Transcript discovery & counting
Homology based
Blast2GO
Functional annotation
(b)
Transcriptome mapping
Reads
Ungapped mapper
Bowtie
Mapping to transcriptome
RSEM, Kallisto
Transcript identification & counting
(c)
Reference-free assembly
Reads
De Bruijn graphs
Trinity
Assembly into transcripts
Ungapped mapper
Bowtie
Map reads back
GTF-based
Htseq-count, RSEM
Counting
Homology based
Blast2GO
Functional annotation

5) Reads quantification

`RNA-seq` 最广泛的应用就是用来评估基因和转录本的表达，这一应用主要是基于比对到转录组区间内的 `reads` 的数量

最简单的方法是，使用 `HTSeq-count` 或 `featureCounts` 计算区间内的 `reads` 数来量化基因的表达。这种基因水平的（不是转录本水平）的量化方法使用的是 `GTF` 文件，这种文件包含外显子和基因在基因组上的坐标。

但一般不能直接使用 `read count` 来比较基因的表达水平，因为该值会受到转录本长度、 `reads` 总数以及测序偏差等因素的影响。所以需要先进行标准化，标准化方法有

1. `RPKM/FPKM`：每百万 `reads` 每一千碱基对中包含的 `reads` 数

该方法先计算测序深度系数，即总 `reads` 数除以 一百万，然后计算基因或转录本的长度（单位为 `kb`），标准化顺序为先消除测序深度的影响，再消除长度的影响：

$$RPKM(x) = \frac{Reads\ per\ transcript}{\frac{total\ reads}{10^6} \times \frac{transcript\ length}{1000}} = \frac{Reads\ per\ transcript}{million\ reads \times transcript\ length(kb)} = \frac{10^9 \times C_x}{R \times L_x}$$

其中

- `x` 表示一个基因或转录本，或基因组上一段特定的区域
- Missing superscript or subscript argument 表示比对到 `x` 外显子区域的 `reads` 数；
- `R` 表示当前样本中包含的全部 `reads` 数
- Missing superscript or subscript argument 表示 `x` 外显子区域包含的碱基数（长度， `bp`）

`FPKM` 与 `RPKM` 的计算公式一样，只是 `RPKM` 用于单端测序， `FPKM` 用于双端测序

1. `TPM`：其与 `RPKM` 最大的区别是，标准化顺序为先消除基因长度的影响，再消除测序深度的影响

首先，将 `reads count` 除以基因或转录本的长度（ `kb` ）得到 `RPK`（`reads per kilobase`），然后将样本中所有的 `RPK` 加起来除以 10^6，得到标准化系数，最后使用 `RPK` 除以标注化系数

$$TPM(x) = \frac{C_x/L_x \times 10^6}{\sum_{i=1}^{N} C_i/L_i}$$

其中

- `x` 表示一个基因或转录本，或基因组上一段特定的区域
- Missing superscript or subscript argument 表示比对到 `x` 外显子区域的 `reads` 数
- Missing superscript or subscript argument 表示 `x` 外显子区域包含的碱基数（ `kp` ）
- `N` 表示基因或转录本总数

这样，每个样本的 `TPM` 总和是一样的，便于比较样本间的差异

6) Diff expression

识别在两个条件下有显著性表达差异的基因，简称差异表达基因；那么怎样才能称得上显著性表达差异？

倍数分析：计算每一个基因在两个条件下的比值，若大于给定阈值，则为差异表达基因；

经典统计模型（如t检验）方法：计算表达差异的置信度，选取一定P值以下的作为差异表达基因；

机器学习：进行特征（基因）选择，包括贝叶斯模型、支持向量机或者随机森林等；

有2点注意：

1）FPKM 和 TPM 标准化方法消除了测序深度和基因或转录本的长度因素的影响，但依赖于总的或有效的 reads 数，当样本的具有异质性转录本分布或当高表达或差异表达的特征扭曲了 count 分布时，表现欠佳；

2）除了这些样本内特异的标准化方法，还需要解决数据集之间的批次效应（不同实验条件下产生的数据之间存在的差异（保证样本测序的一致性）；

计算差异表达的方法：

1）edgeR 将原始的 read counts 作为输入，并在统计模型中加入了标准化

2）DESeq2 使用的是负二项分布作为参考分布，并提供了自己的标准化方法；

```
dds <- DESeq(dds)
vsd <- vst(dds, blind=FALSE)
```

7) GO_KEGG

五：GO　Analysis
要点：
1，差异基因个数
2，GO分析原理：Fisher's　Exact　Test
3，P-Value　or FDR，与Enrichment相比，优选P-Value，P-Value越小代表在这个GO team上富集越显著。
4，Enrichment=(a/a+b)/(a+b/a+b+c+d)
5，GO2Gene with Blast?
6，GO如何解读?
7，CC(cell component)，MF(Molecule Function)，BP(biological process)，这三者可以依次理解为，where（在哪里），what（干了什么），influence（最终影响了什么），而往往BP又和表型（phenotype）结合在一起，所以GO能给我们提供一些研究重点。
8，Down　Up　or All

六：Pathways Analysis
要点：
1，差异基因个数
2，Pathway分析原理：Fisher's　Exact　Test
3，P-Value　or FDR，与Enrichment相比，优选P-Value，P-Value越小代表在这个Pathway上与你的研究方向越相关。
4，Enrichment=(a/a+b)/(a+b/a+b+c+d)
5，Disease Pathway（发现一些pathway和你的研究方向不符，可以选择性的删掉，有些编辑并不能理解这个内容，因为我们不怎么参考FDR的值而是参考P-value值，其只和某一行数值有关，而P-Value并不会受到条目数量的影响，而FDR会随着P-Value的值变化发生变化）
6，KEGG数据库
7，Other Pathway

GO: agriGO2
KEGG: clusterProfiler

Thanks