BerryGenomics
贝瑞基因

二代真核转录组

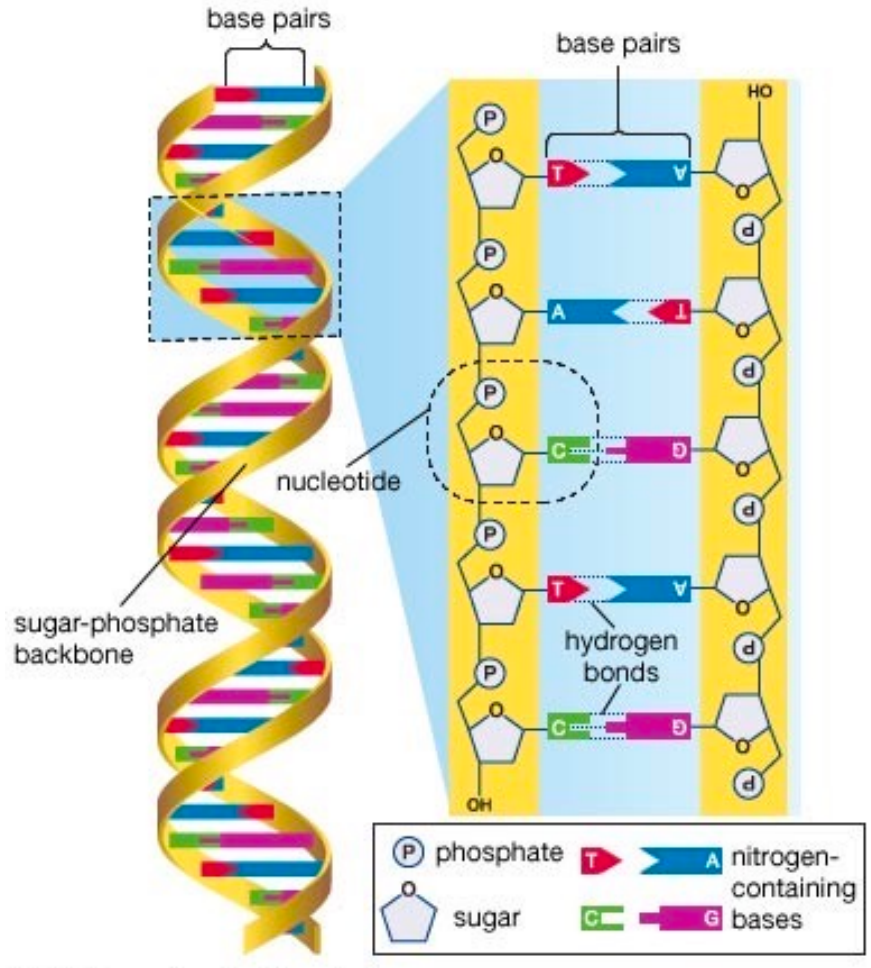
王鹏

科技服务技术支持部

1

Illumina测序原理

什么是测序



DNA测序：不可见的未知序列，转换成可见的ATGC

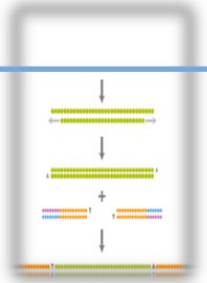
序列分析：可见的ATGC，转换成机器可读的0和1

统计绘图：机器识别的0和1，转换成人可读的图表

Illumina测序步骤

二代测序四步骤

1 Library Preparation



文库构建

2 Cluster Generation

簇的生成

3 Sequencing

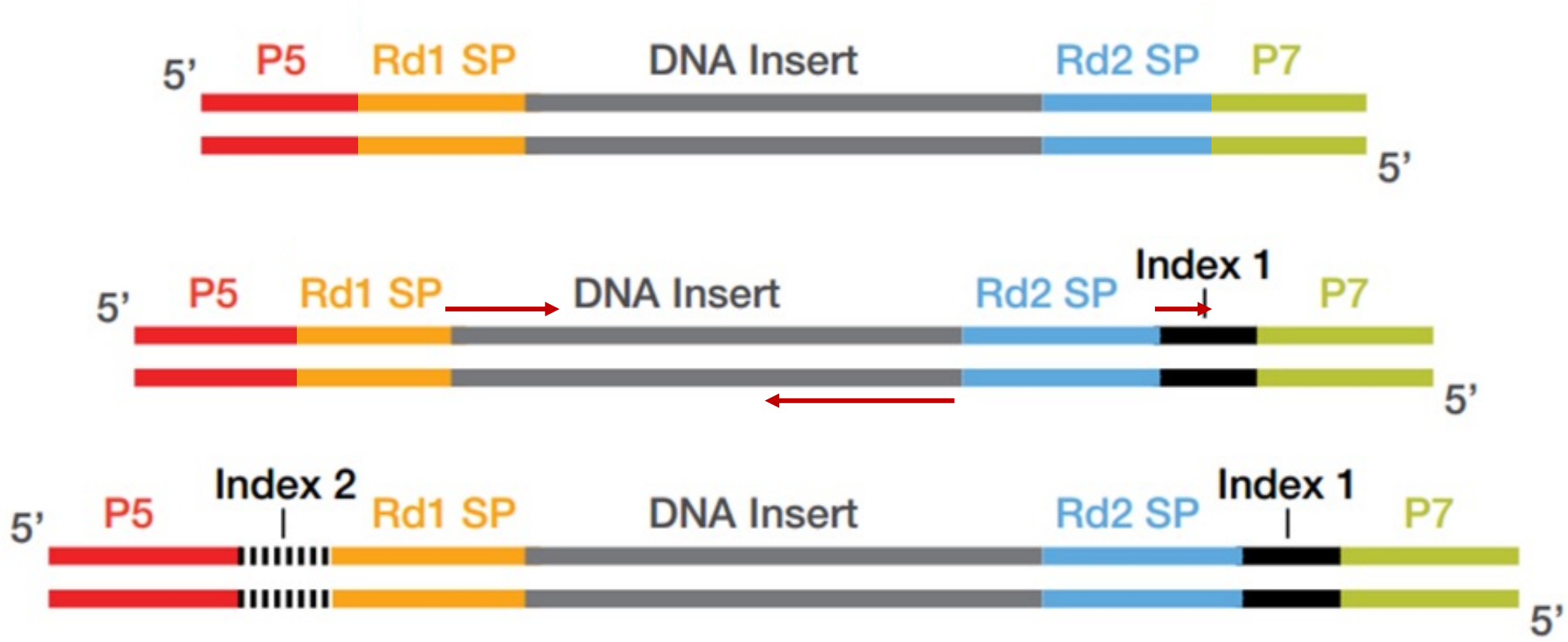
上机测序

4 Data Analysis

数据分析

Illumina文库结构

测序就像走路，首先要确定行程目标，其次要有个引路人



P5/P7接头：行程的起始和终止；Rd1SP：行程的引路人；Index：区分不同行人的标签

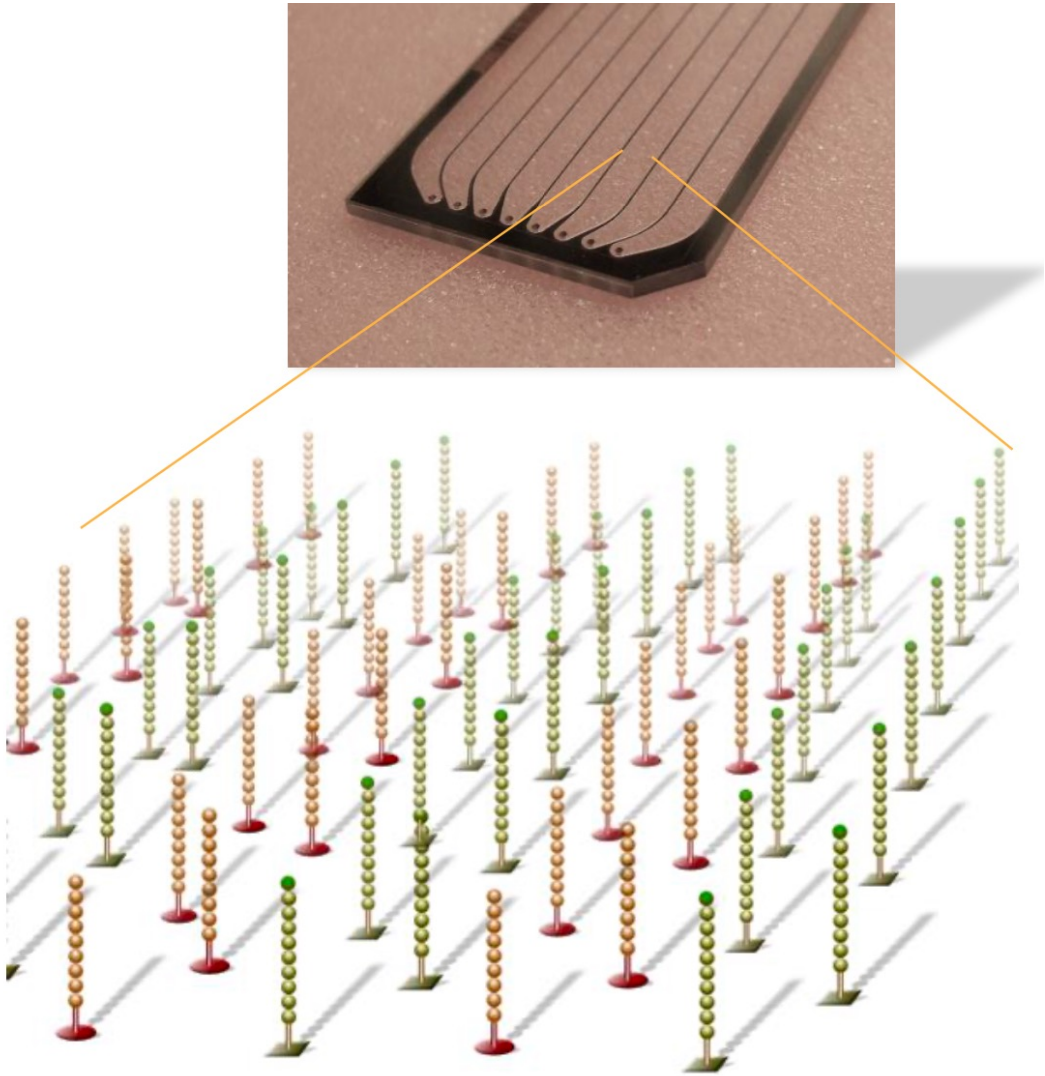
Illumina flowcell

走路要在一定区域行走，不能乱走，flowcell就类似行走的固定区域或载体

Cluster generation occurs on a flow cell

A flow cell is a thick glass slide with channels or lanes

Each lane is randomly coated with a lawn of oligos that are complementary to library adapters

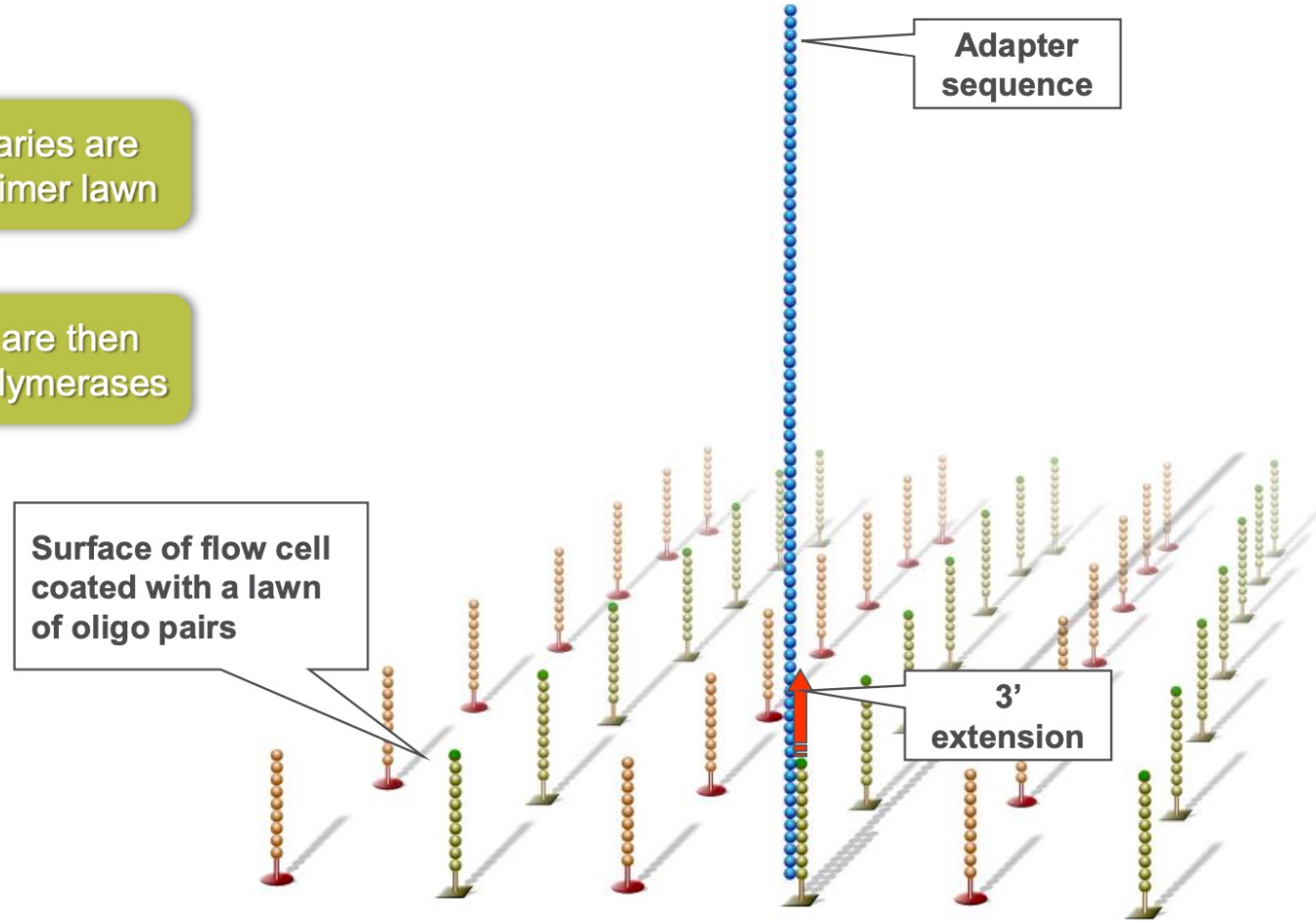


片段杂交&延伸

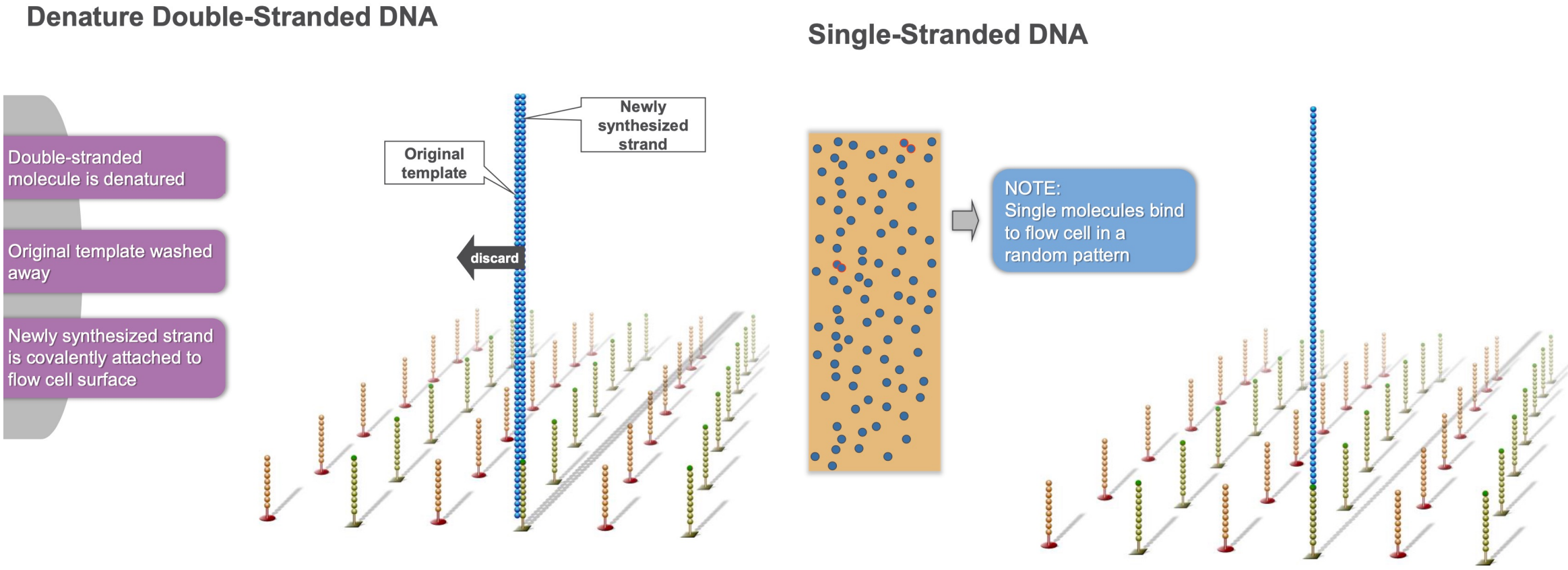
Hybridize Fragment & Extend

Single DNA libraries are hybridized to primer lawn

Bound libraries are then extended by polymerases



双链DNA变性丢掉原始DNA链

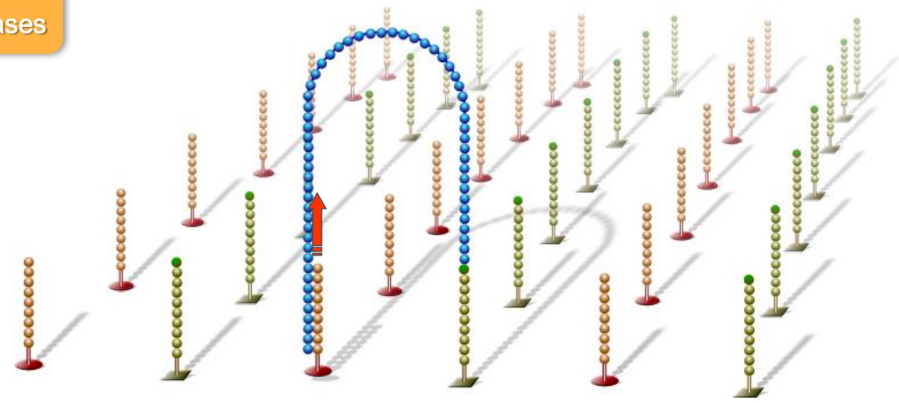


桥式PCR

Bridge Amplification

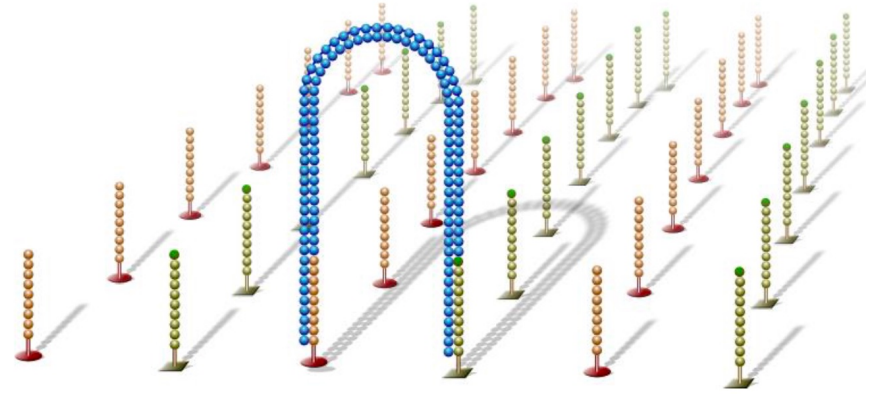
Single-stranded molecule flips over and forms a bridge by hybridizing to adjacent, complementary primer

Hybridized primer is extended by polymerases



Bridge Amplification

Double-stranded bridge is formed

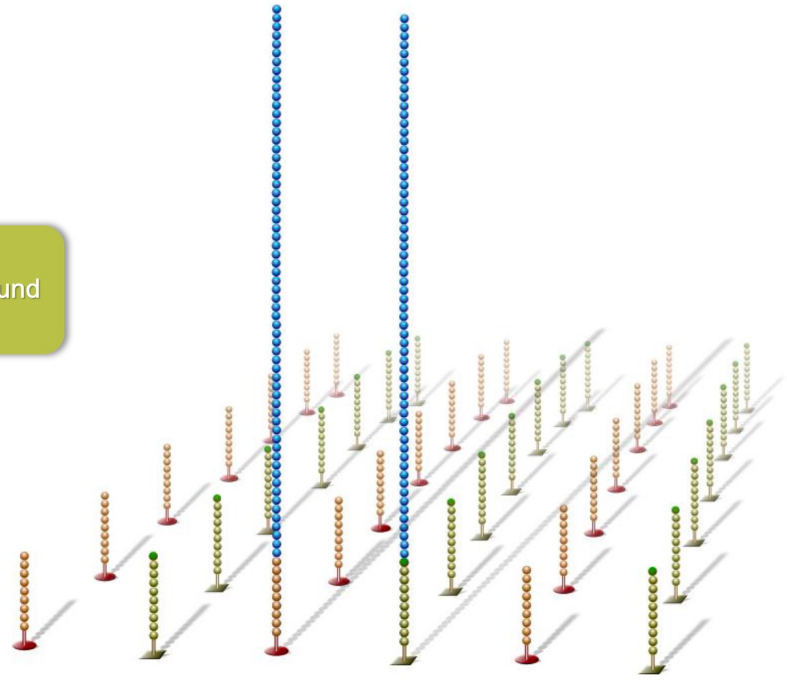


桥式PCR

Denature Double-Stranded Bridge

Double-stranded bridge is denatured

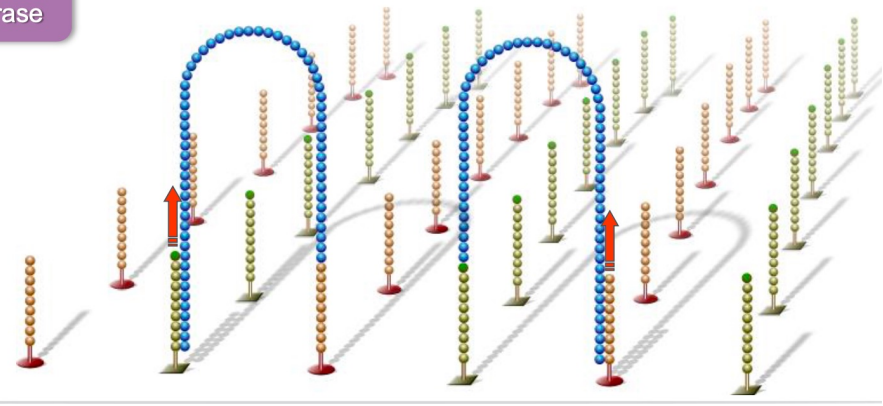
Result:
Two copies of covalently bound single-stranded templates



Bridge Amplification

Single-stranded molecules flip over to hybridize to adjacent primers

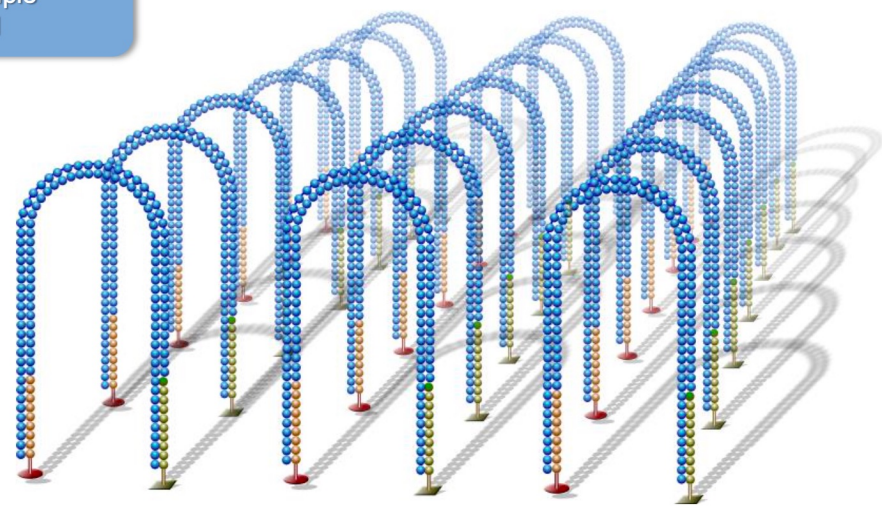
Hybridized primer is extended by polymerase



桥式PCR

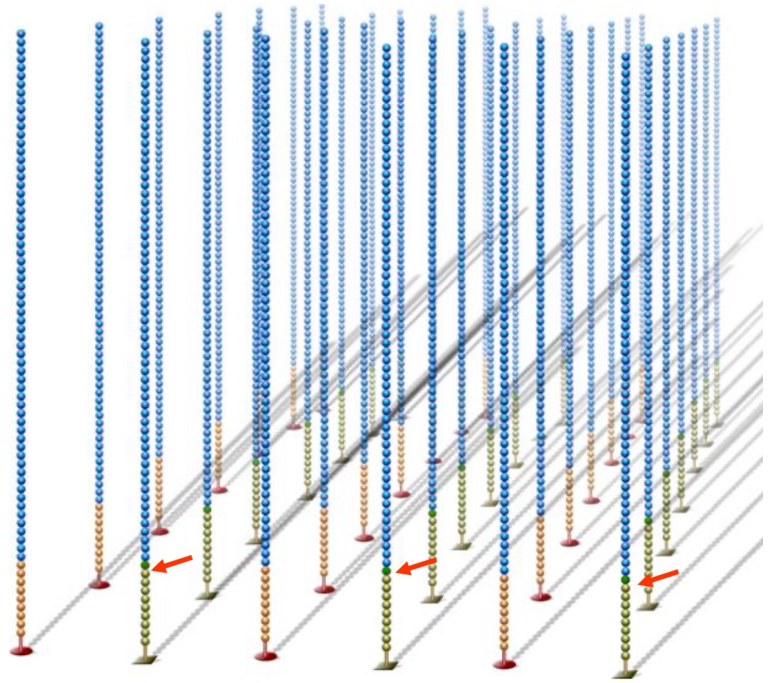
Bridge Amplification

Bridge amplification cycle is repeated until multiple bridges are formed



Linearization

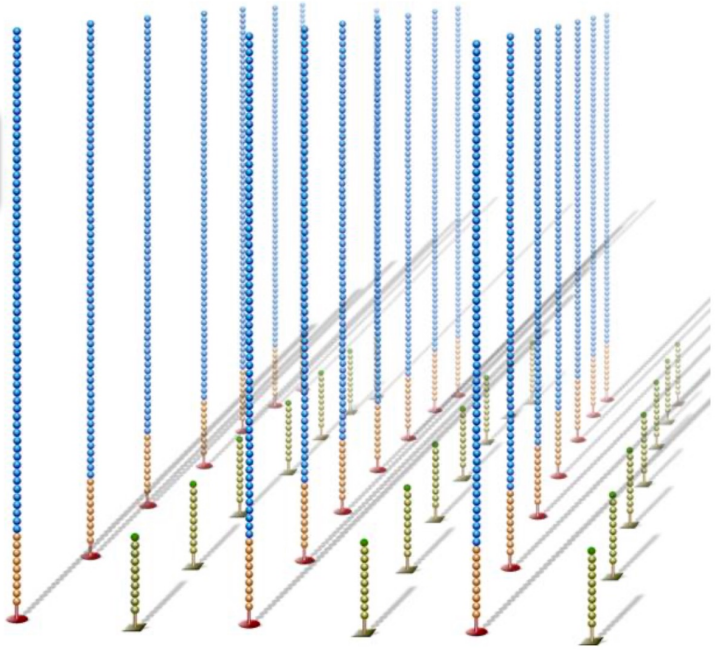
dsDNA bridges are denatured



洗脱第二链，封闭3端

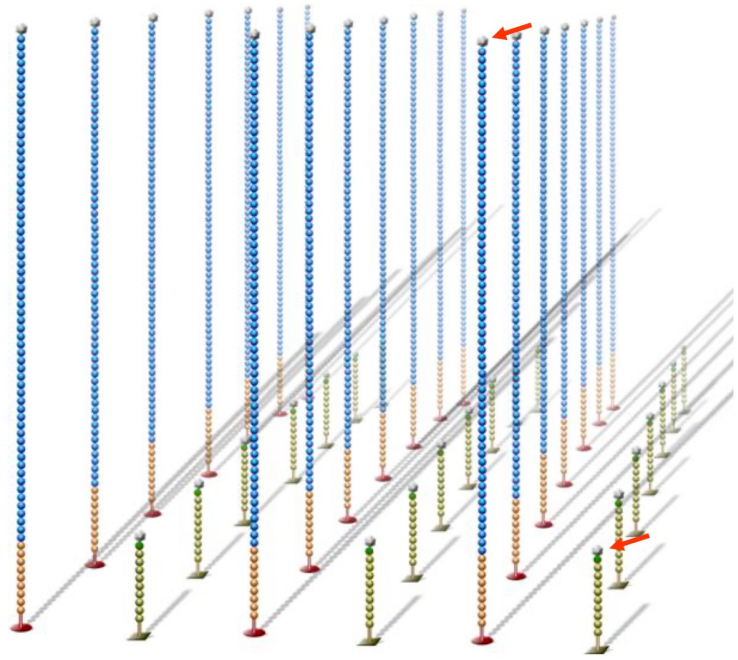
Reverse Strand Cleavage

Reverse strands are cleaved and washed away, leaving a cluster with forward strands only

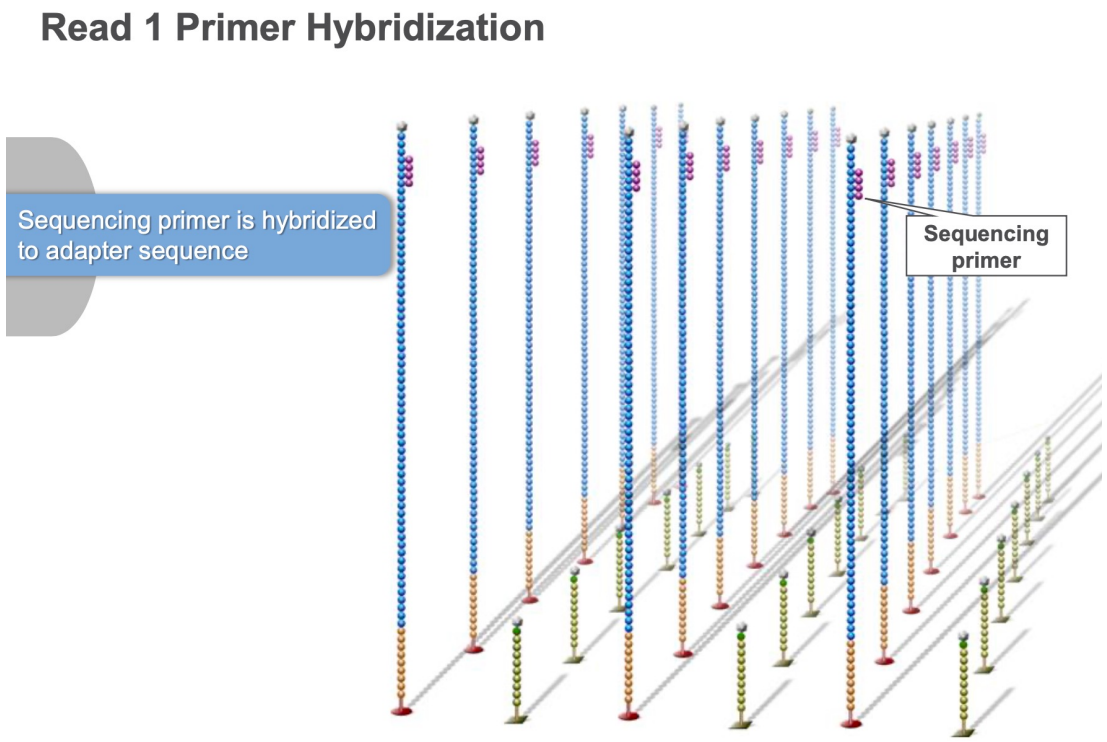


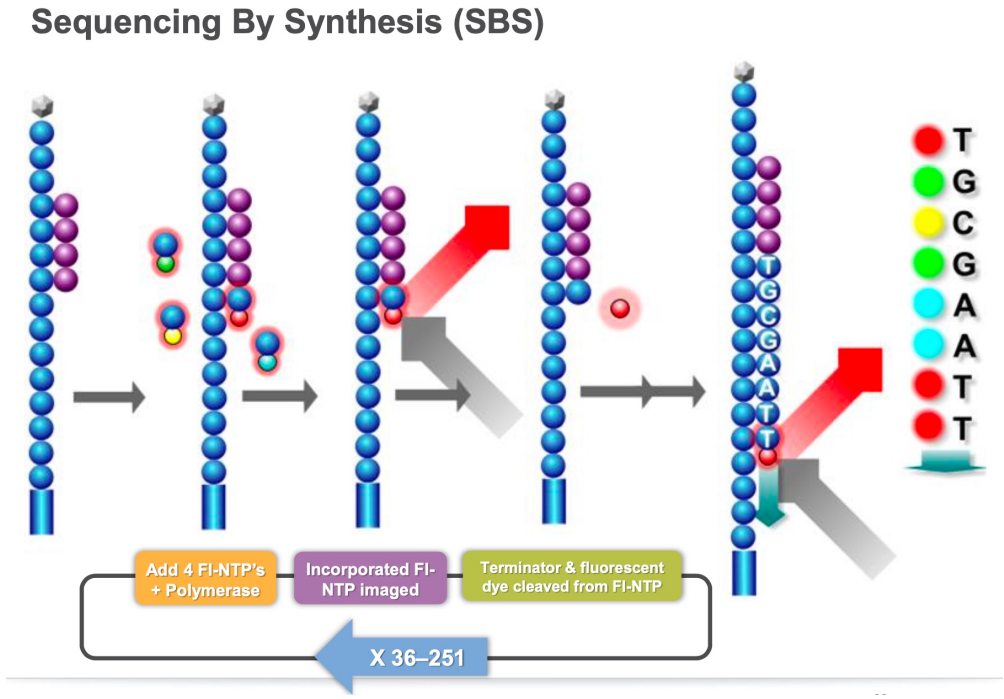
Blocking

Free 3’ ends are blocked to prevent unwanted DNA priming



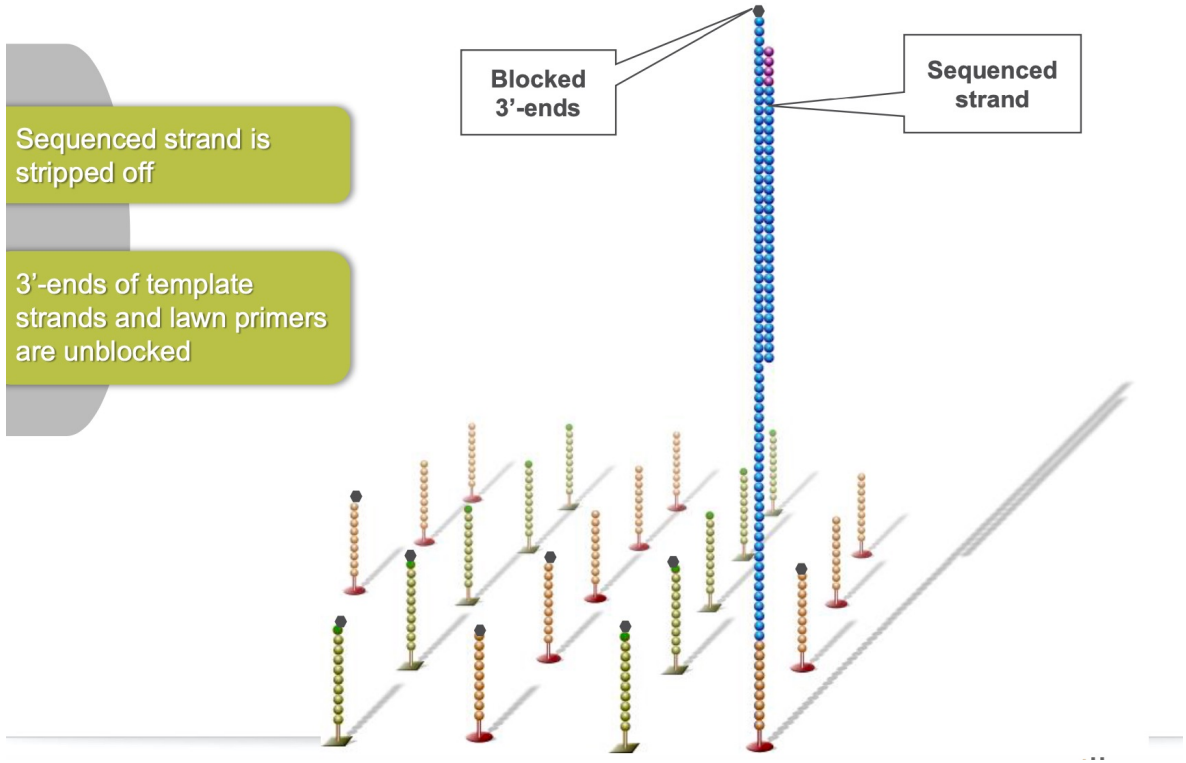
加入引物1，边合成边测序



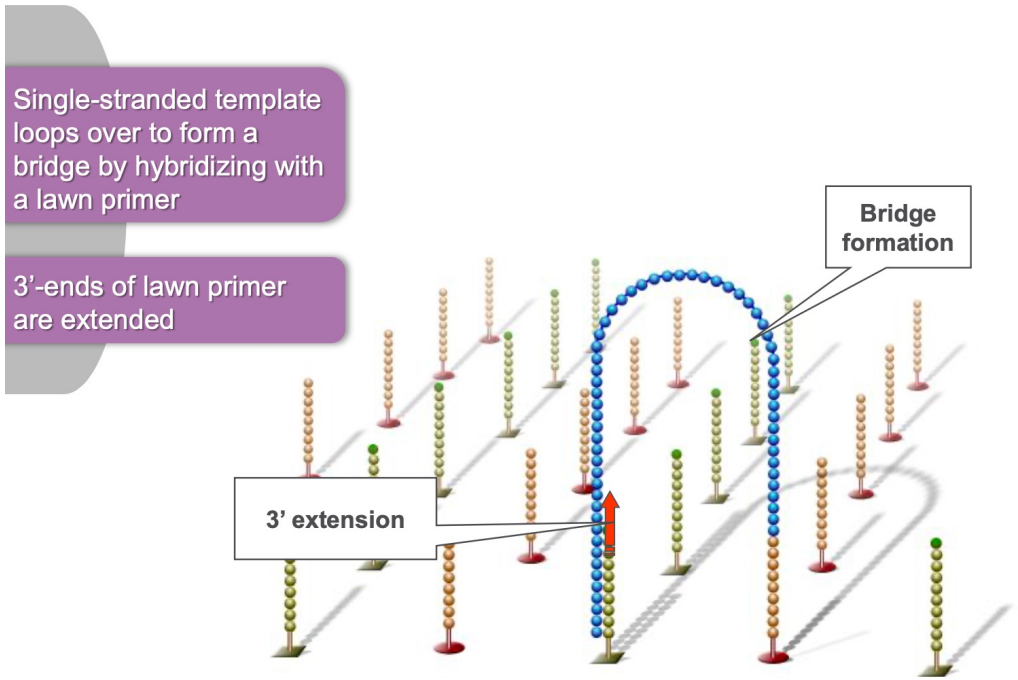


2端测序

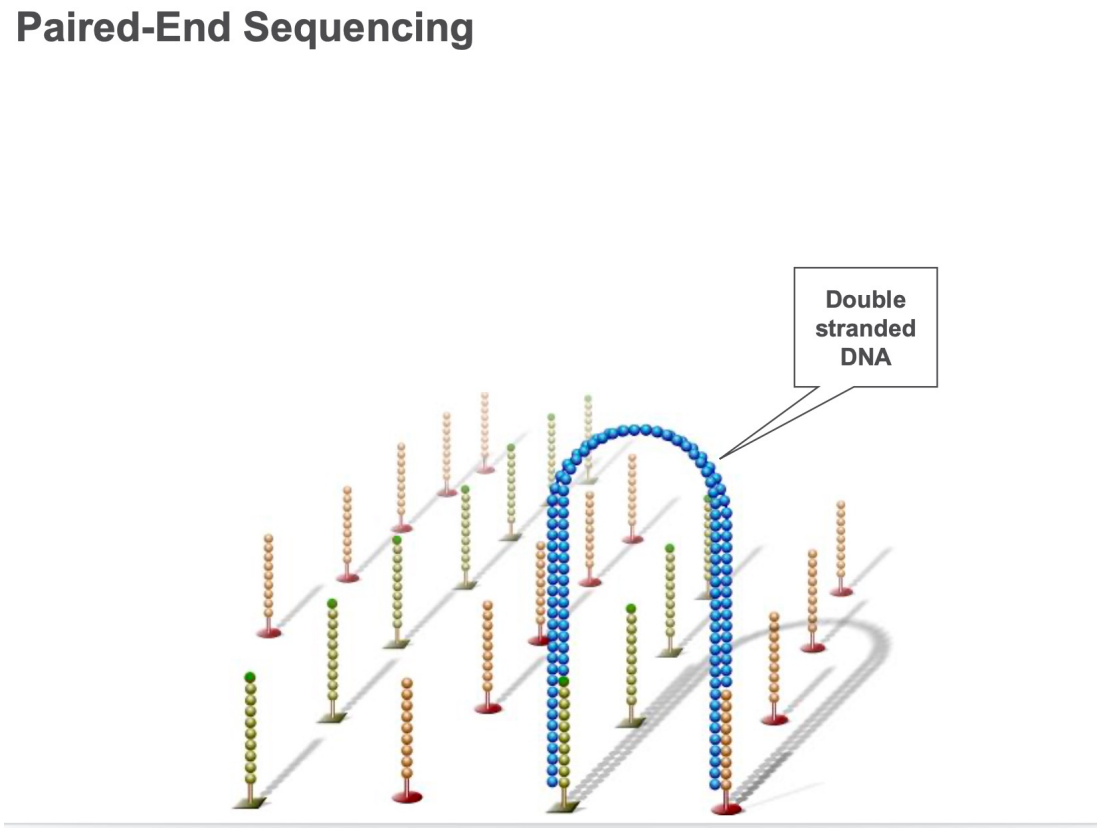
Paired-End Sequencing

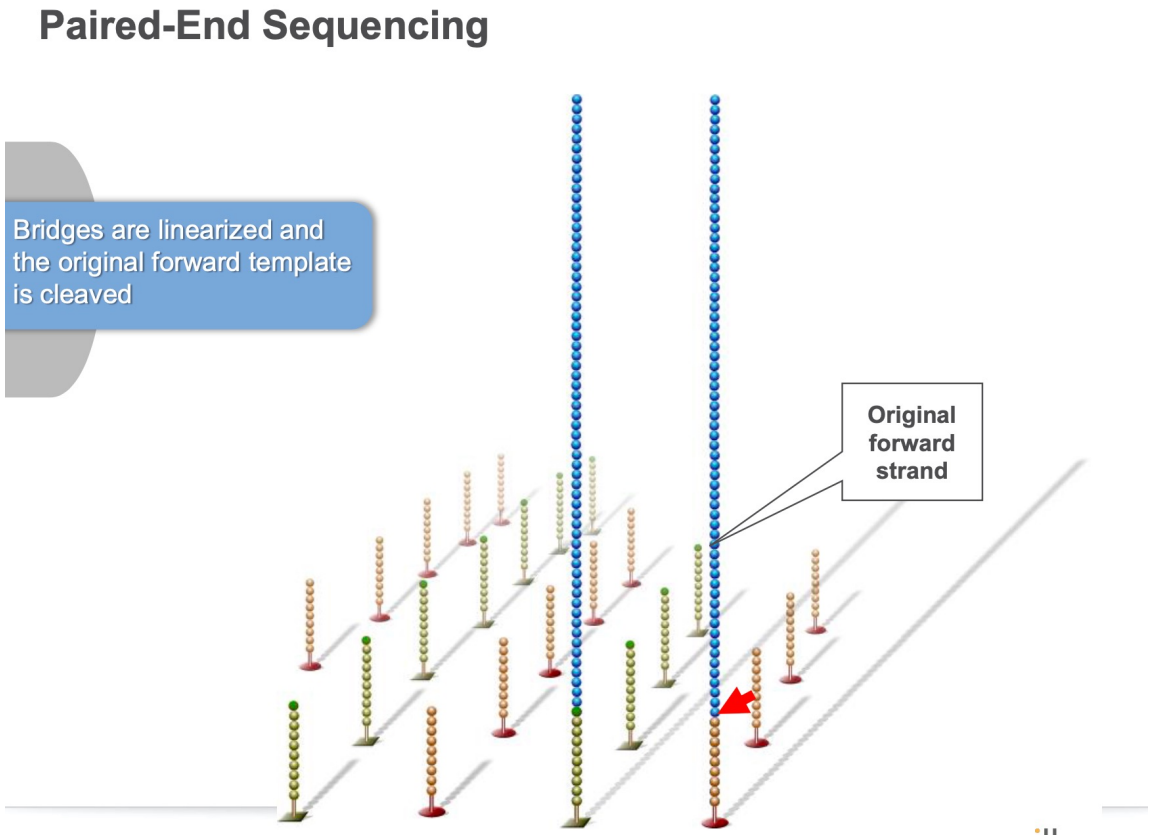


Paired-End Sequencing

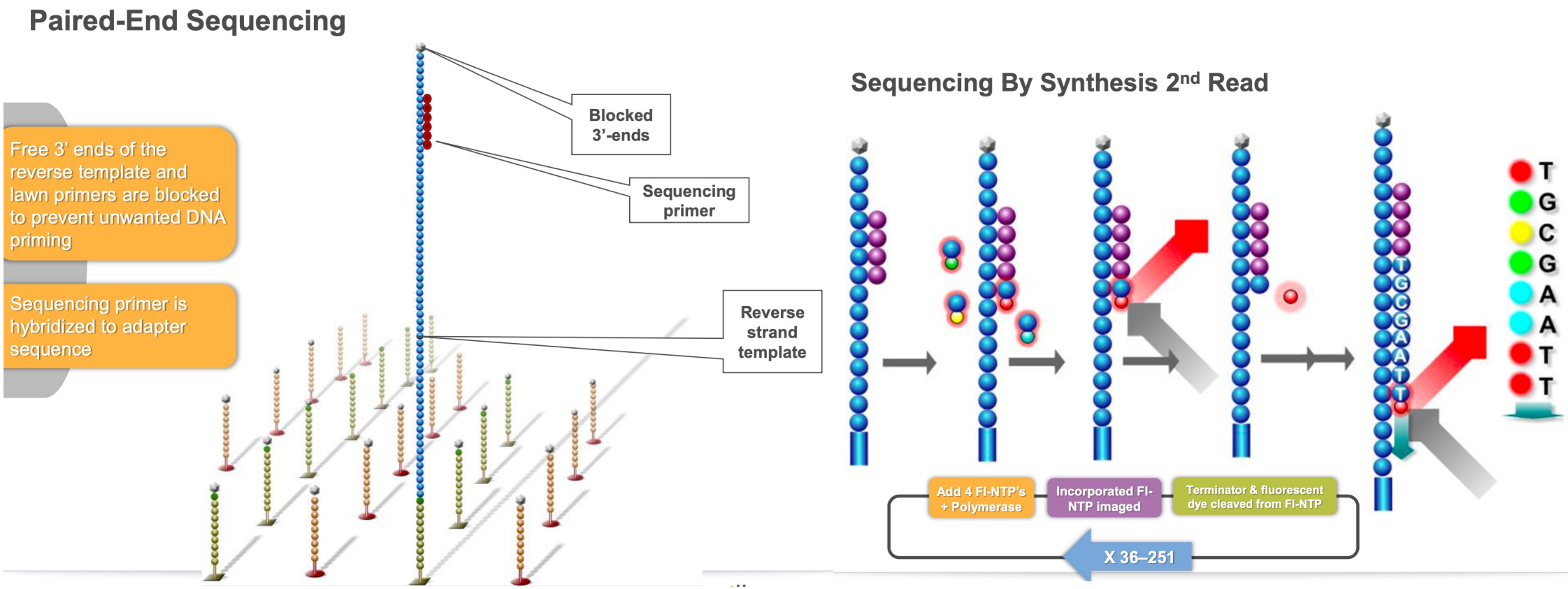


2端测序


Paired-End Sequencing
Double stranded DNA

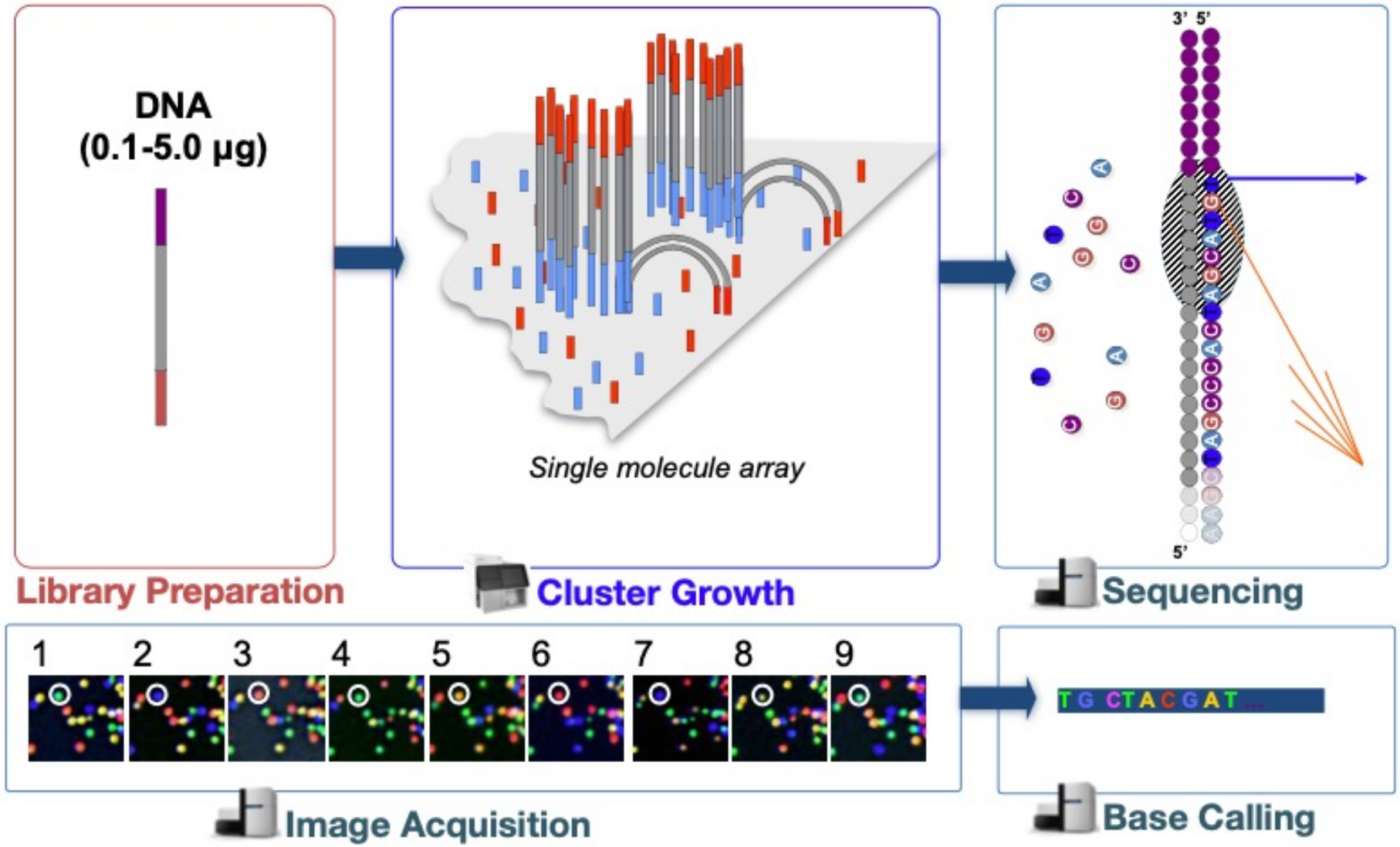

Paired-End Sequencing
Bridges are linearized and the original forward template is cleaved
Original forward strand

2端测序


Paired-End Sequencing
Free 3' ends of the reverse template and lawn primers are blocked to prevent unwanted DNA priming
Sequencing primer is hybridized to adapter sequence
Blocked 3'-ends
Sequencing primer
Reverse strand template
Sequencing By Synthesis 2nd Read
Add 4 FI-NTP's + Polymerase
Incorporated FI-NTP imaged
Terminator & fluorescent dye cleaved from FI-NTP
X 36–251
T
G
C
G
A
A
T
T

Illumina测序动态图

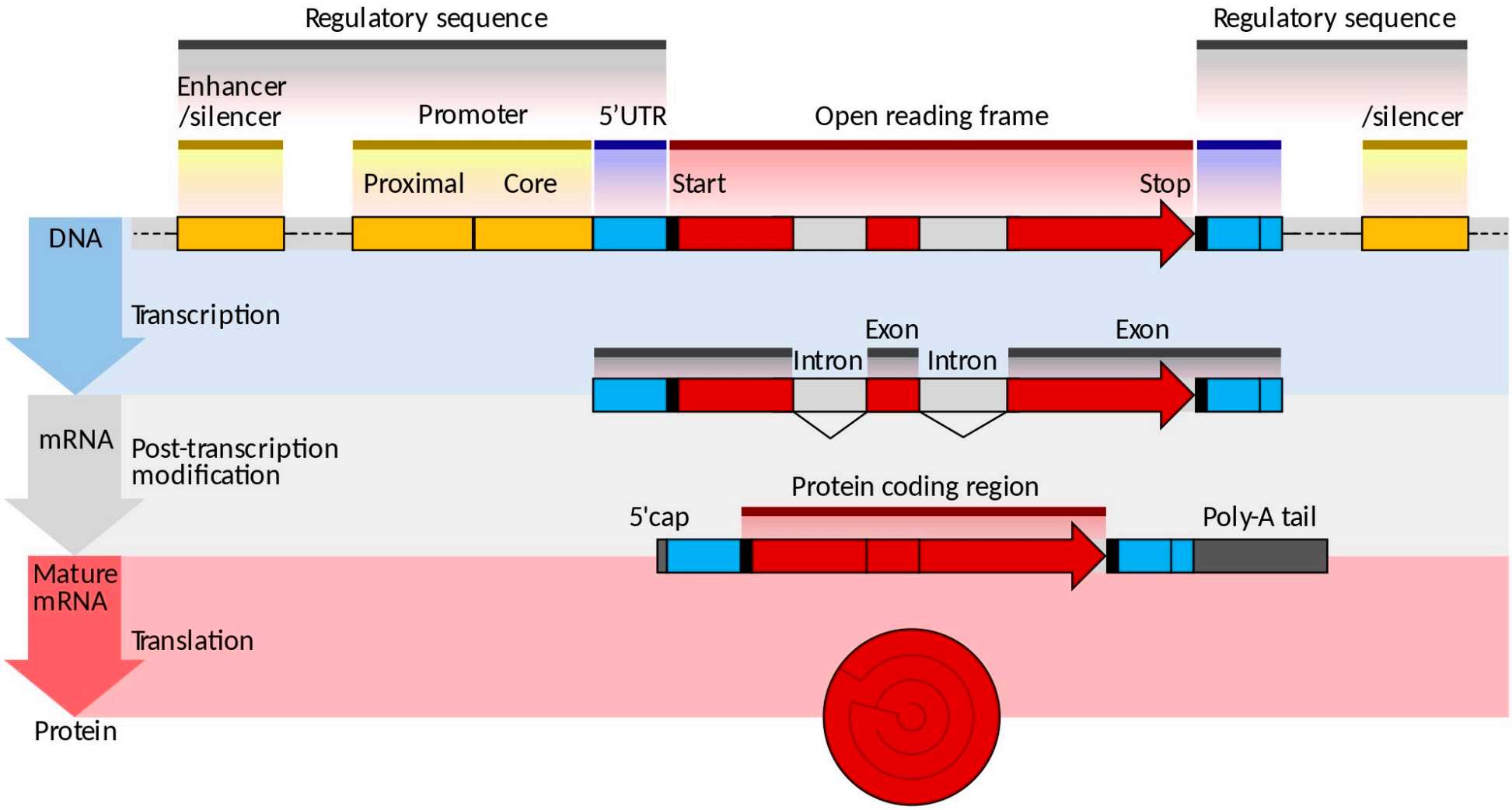
测序示意图


DNA
(0.1-5.0 µg)
Library Preparation
Single molecule array
Cluster Growth
3' 5'
5'
Sequencing
1 2 3 4 5 6 7 8 9
Image Acquisition
TG CTACGAT...
Base Calling

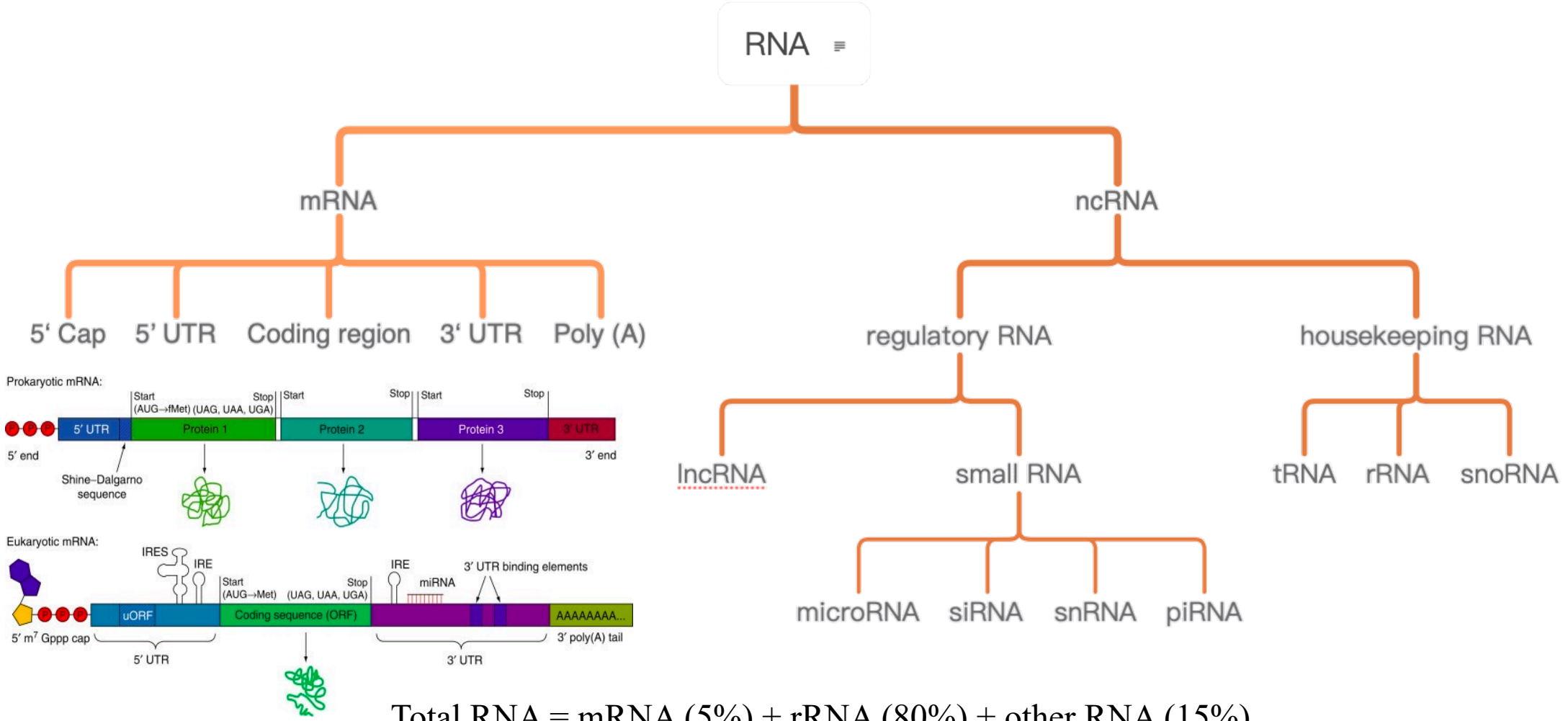
2

RNA常见的文库类型

从DNA到mRNA

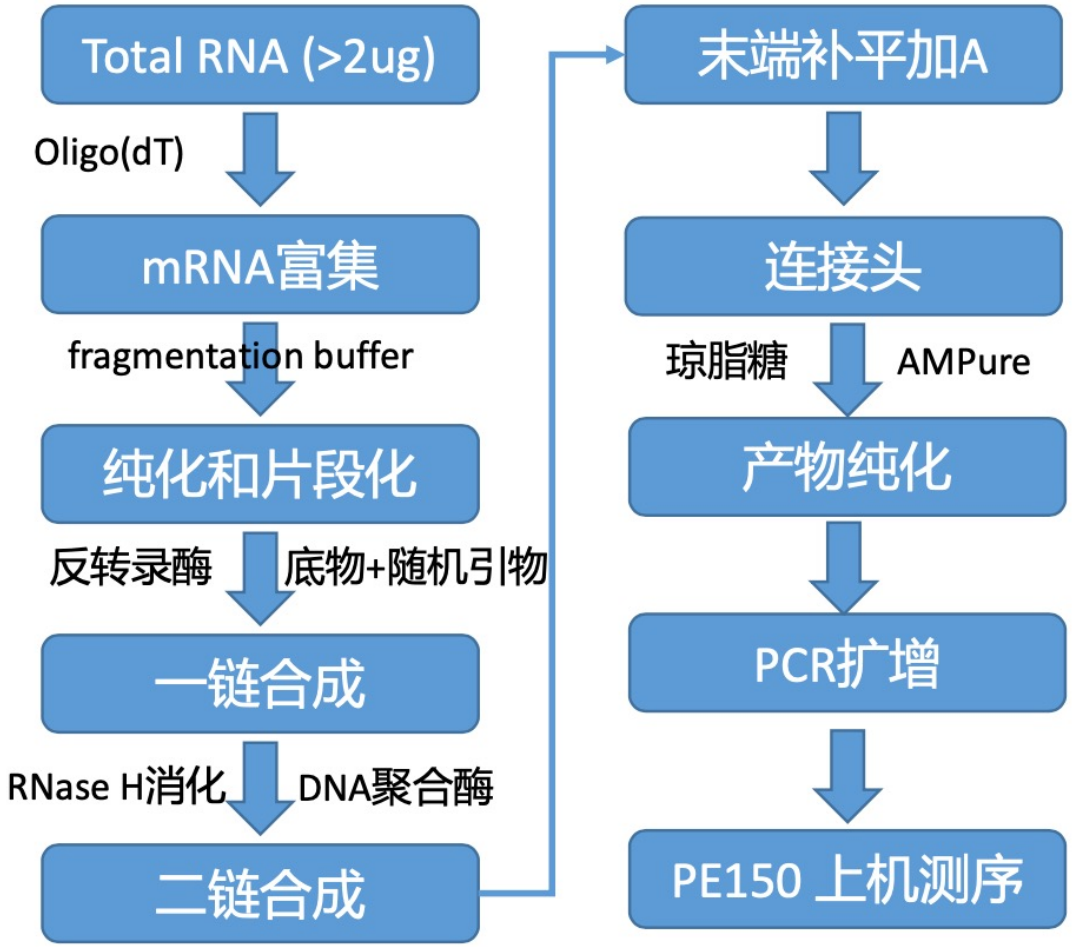

Regulatory sequence
Regulatory sequence
Enhancer
/silencer
Promoter
5'UTR
Open reading frame
/silencer
Proximal
Core
Start
Stop
DNA
Transcription
Exon
Exon
Intron
Intron
mRNA
Post-transcription
modification
Protein coding region
5'cap
Poly-A tail
Mature
mRNA
Translation
Protein

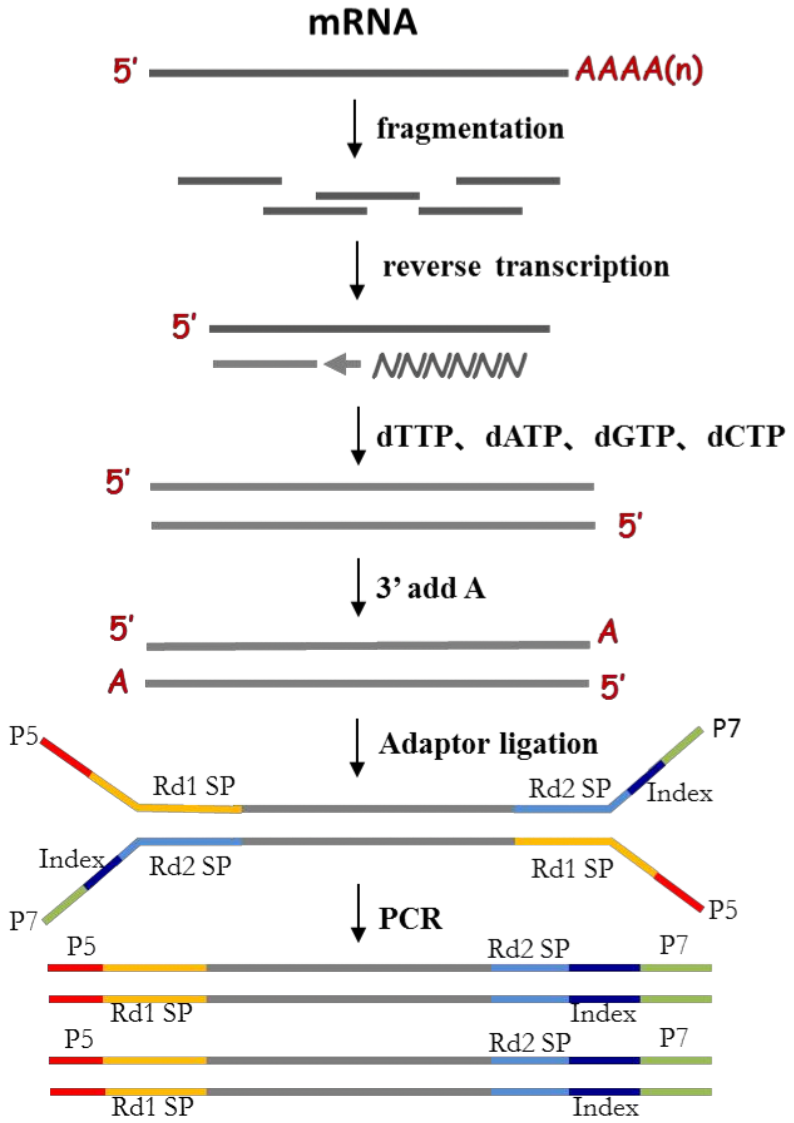
RNA分类



Total RNA = mRNA (5%) + rRNA (80%) + other RNA (15%)

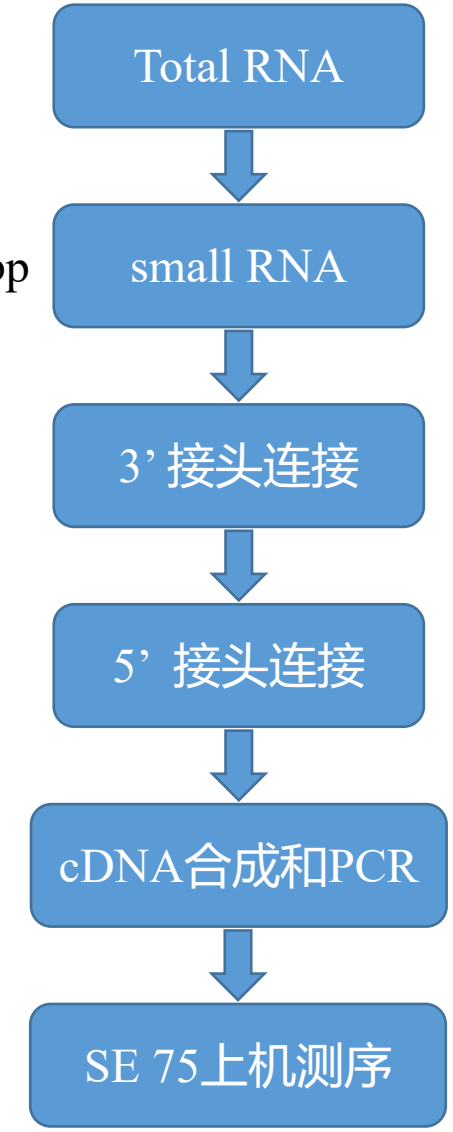
RNA-seq文库构建

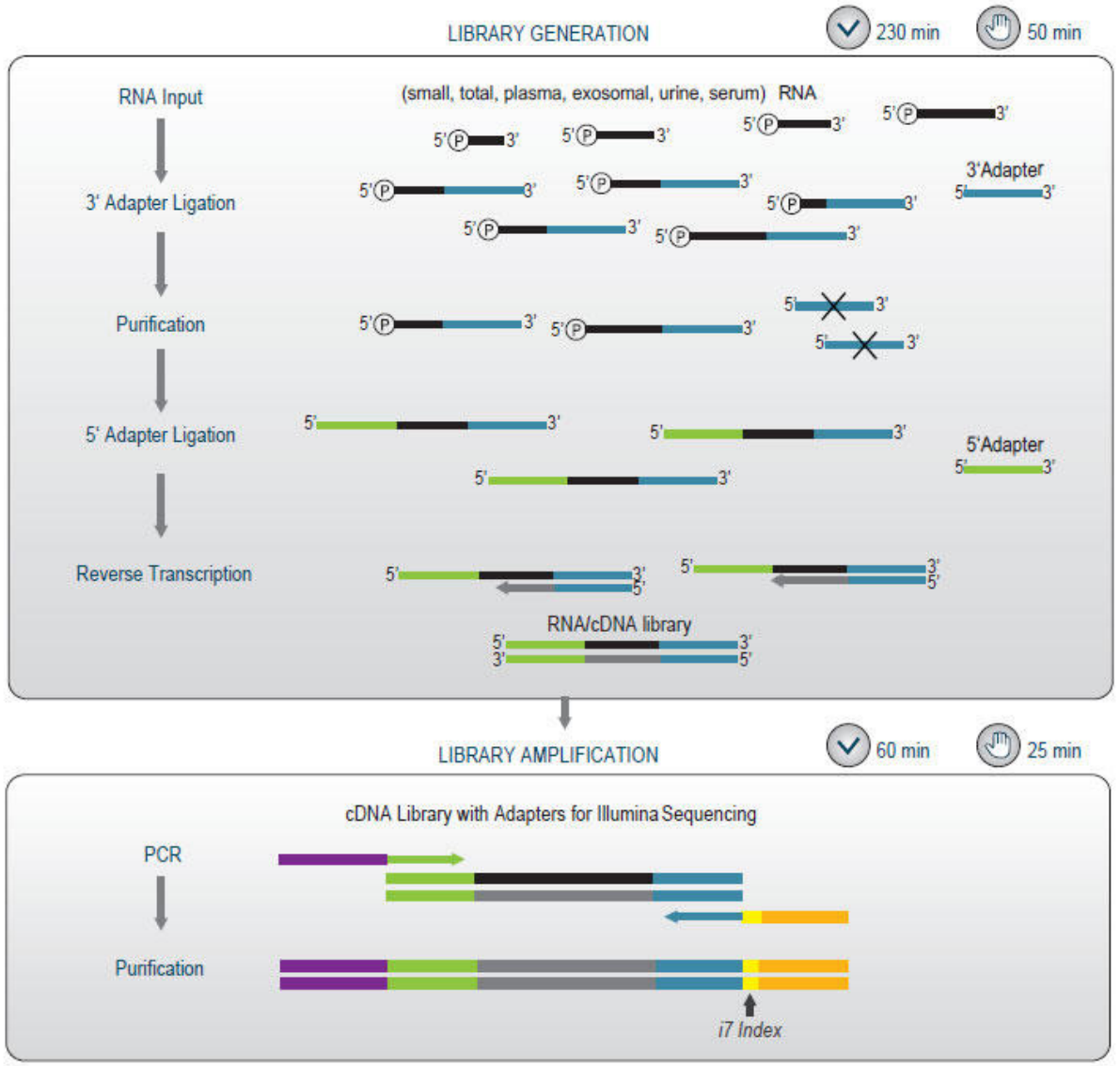

Total RNA (>2ug)
Oligo(dT)
mRNA富集
fragmentation buffer
纯化和片段化
反转录酶
底物+随机引物
一链合成
RNase H消化
DNA聚合酶
二链合成
末端补平加A
连接头
琼脂糖
AMPure
产物纯化
PCR扩增
PE150 上机测序


mRNA
5'
AAAA(n)
fragmentation
reverse transcription
5'
NNNNNN
dTTP、dATP、dGTP、dCTP
5'
5'
3' add A
5'
A
A
5'
Adaptor ligation
P5
P7
Rd1 SP
Rd2 SP
Index
Index
Rd2 SP
Rd1 SP
P7
P5
PCR
P5
Rd2 SP
P7
Rd1 SP
Index
P5
Rd2 SP
P7
Rd1 SP
Index

Small RNA文库构建

18-30bp



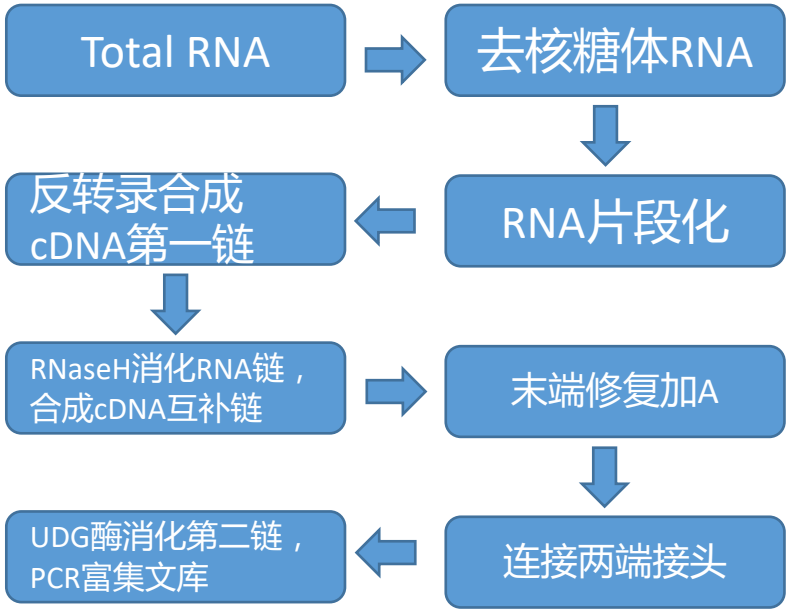


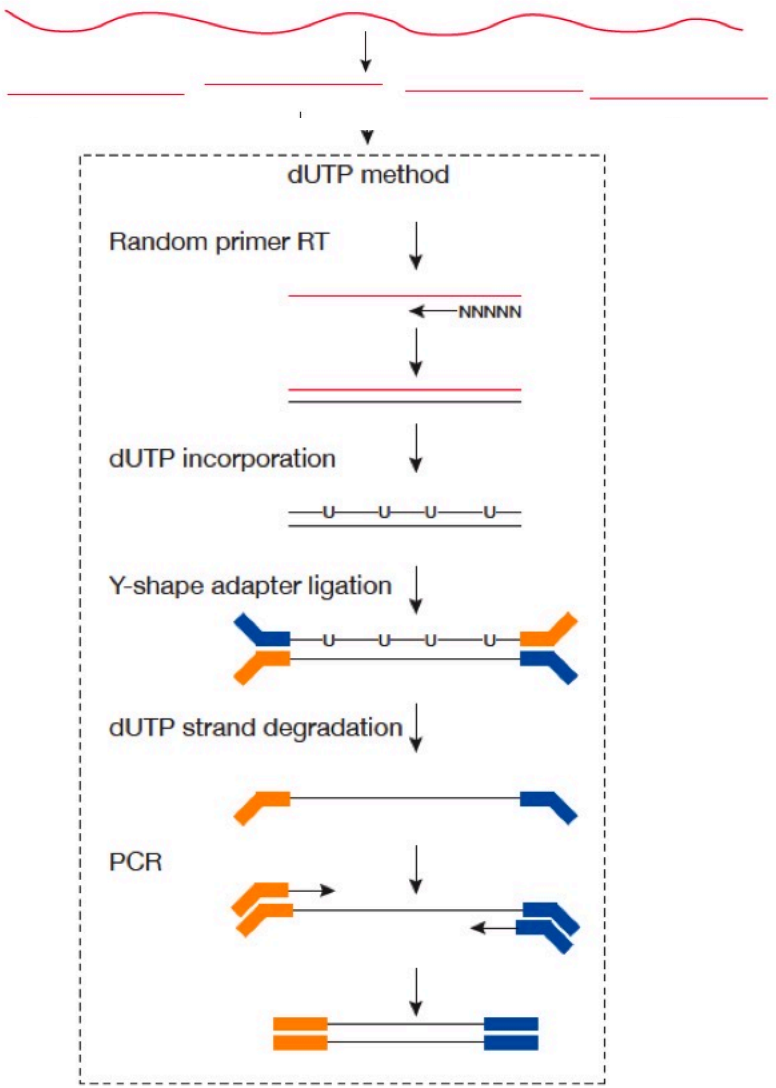
LncRNA文库构建

>200bp

链特建库可区分Reads的正负方向

- 定量更加准确
- 对可变剪切的检测更准确
- 无参转录组更加真实





3

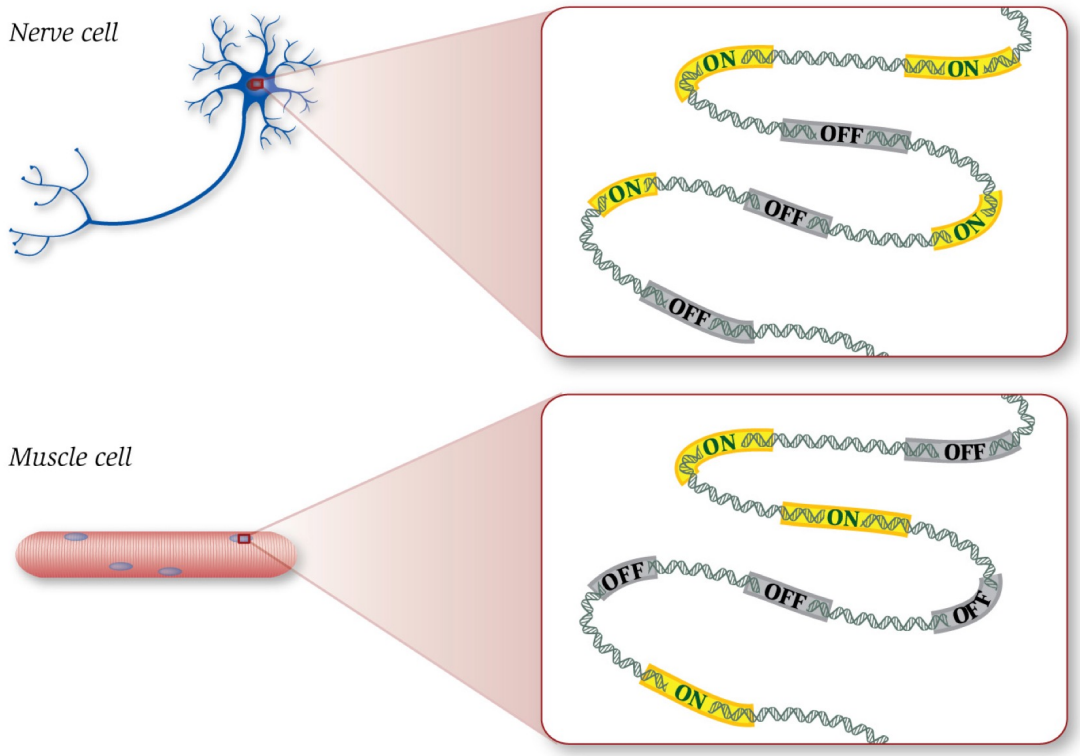
RNA-seq分析内容

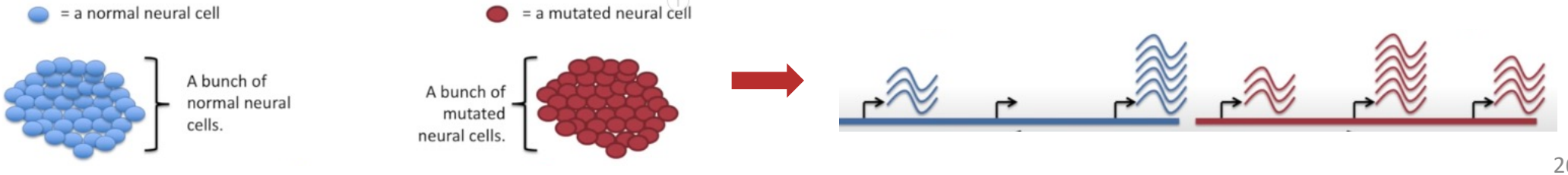
RNA-seq解决的问题

基因的表达具

有时空特异性

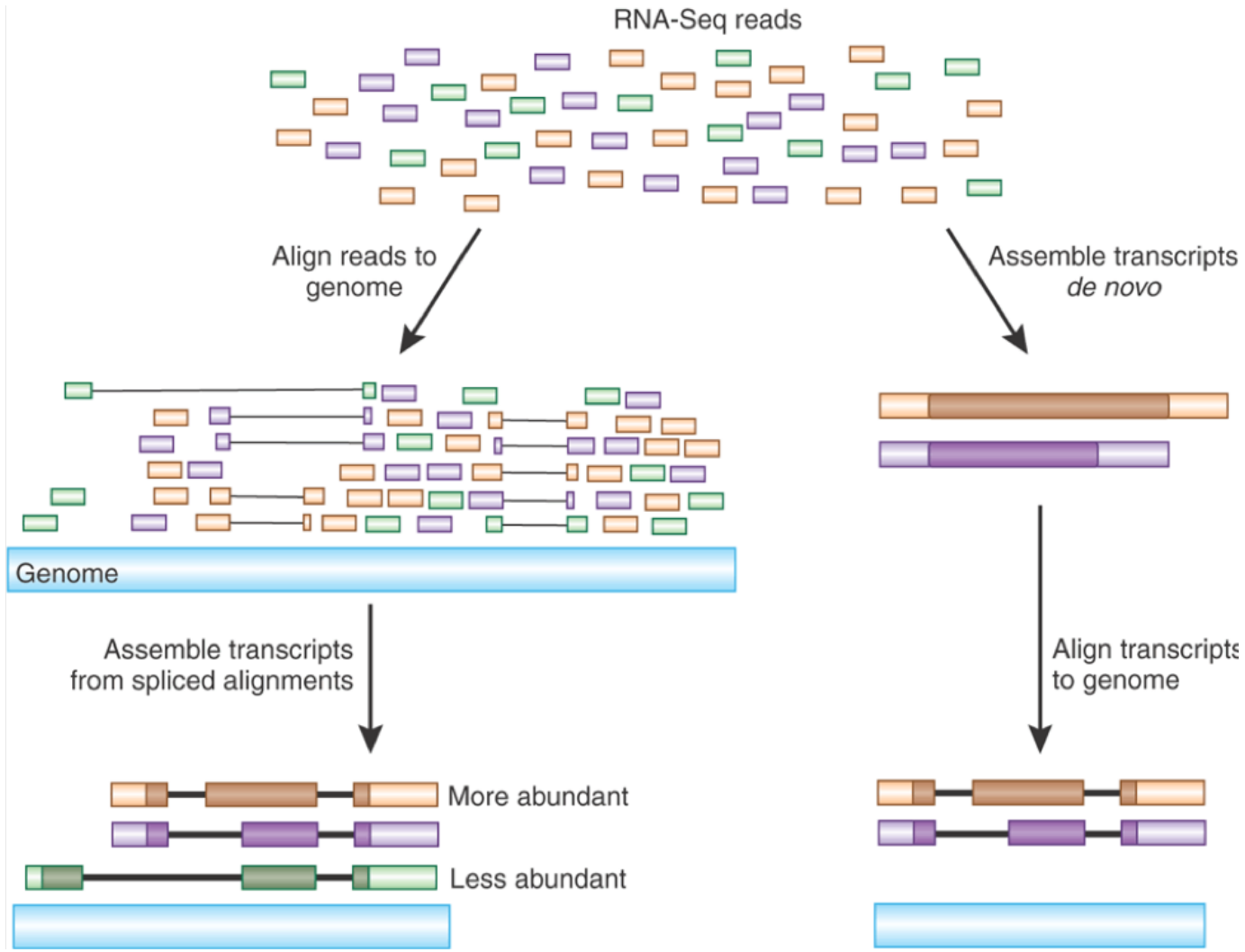
- 在某些样本中，哪些基因是活跃的（转录表达）？
- 不同样本中，差异表达的基因有哪些？





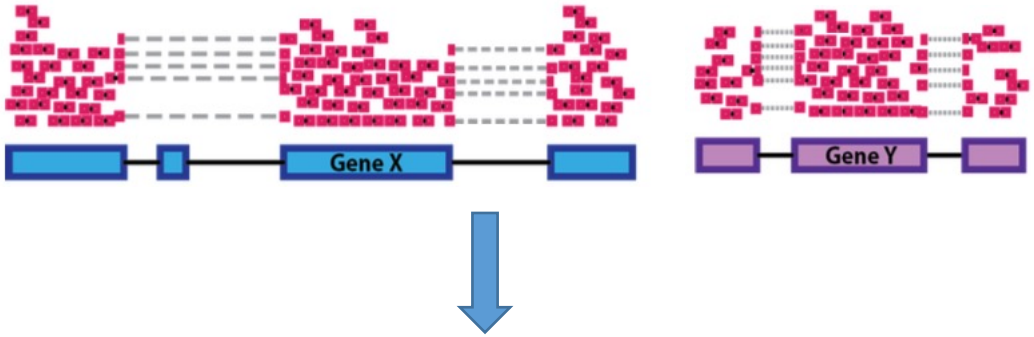
RNA-seq分析流程

有参转录组



无参转录组

基因表达矩阵



	ctrl_1	ctrl_2	exp_1	exp_1
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
…	…	…	…	…
…	…	…	…	…

Count matrix

获得基因表达矩阵后，

可以从哪些维度分析？

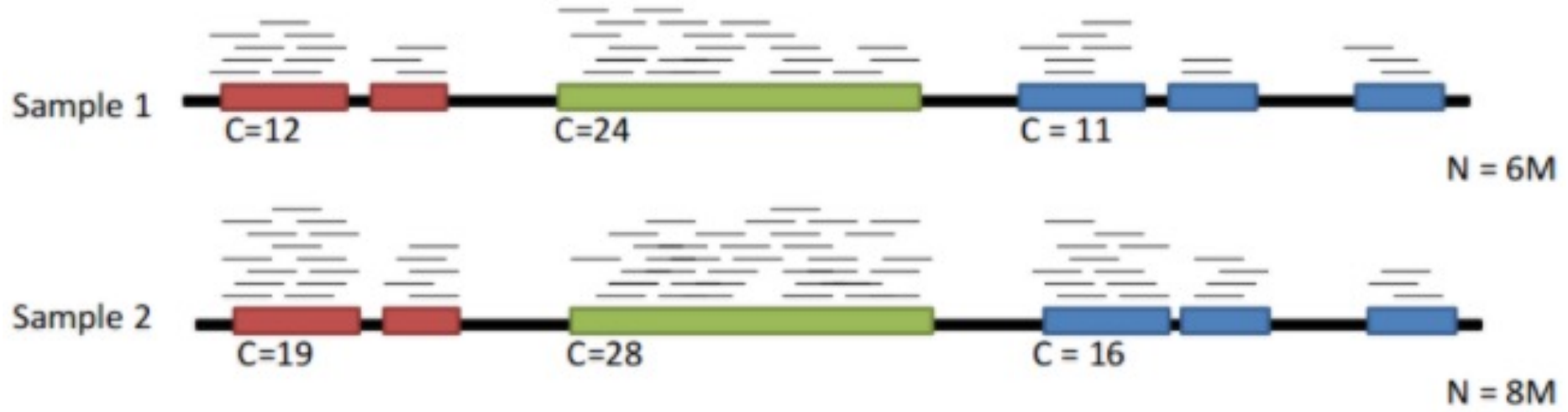
样本维度

不同样本中活跃基因的数目；样本之间的相似性...

基因维度

差异表达基因有多少；基因的具体功能是什么？

数据的标准化



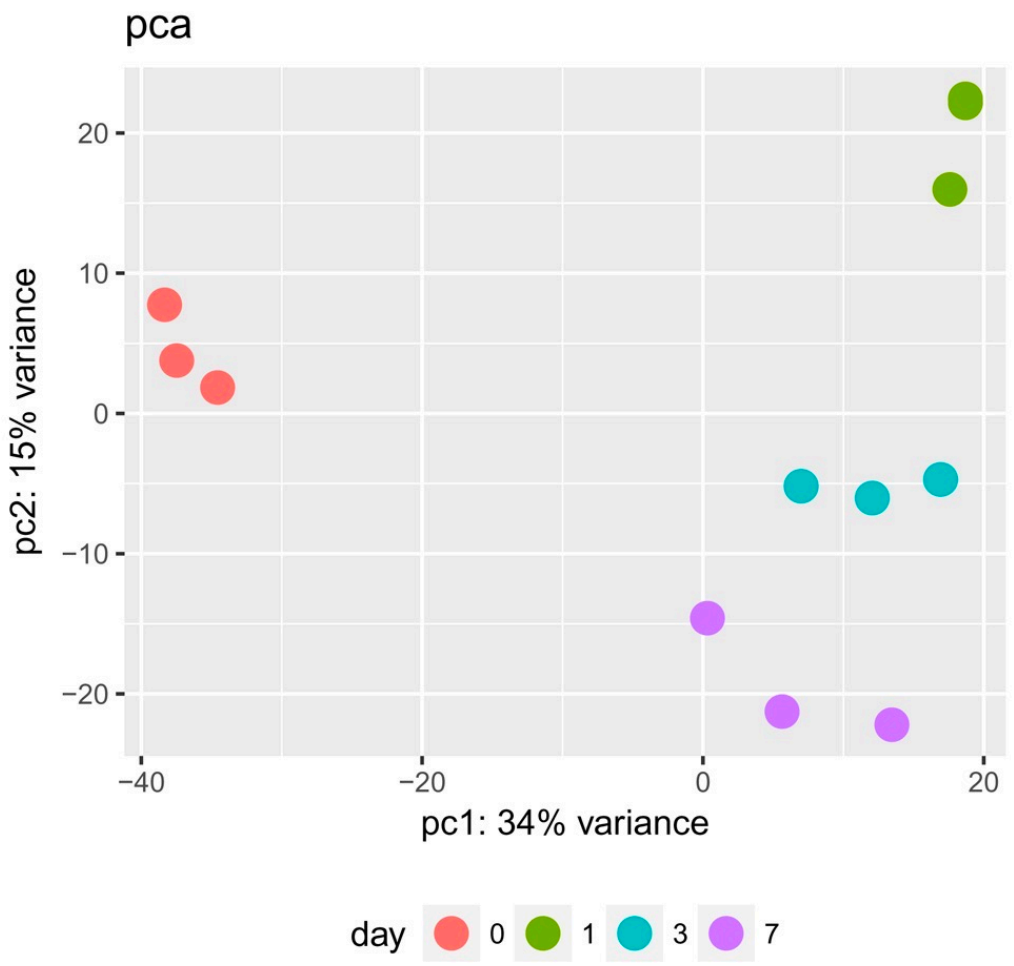
测序深度的影响

基因长度的影响

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

注：n：reads比对上的count数；N：总的测序的reads数；L：基因片段的长度

样本维度-PCA



不同类型的样本，是否有显著的分离？

分离：样本间的基因表达模式差异较大

不分离：样本间的基因表达较为相似

横坐标：主成分1及其方差解释率

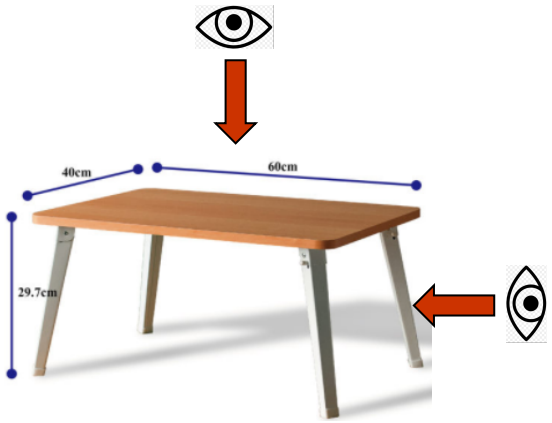
纵坐标：主成分2及其方差解释率

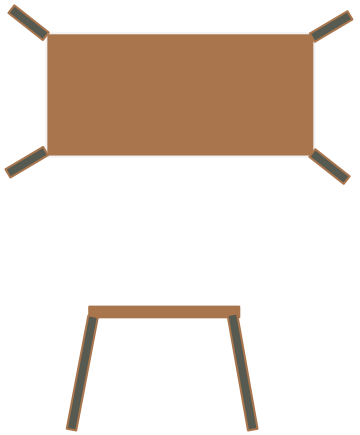
图中的点代表样本

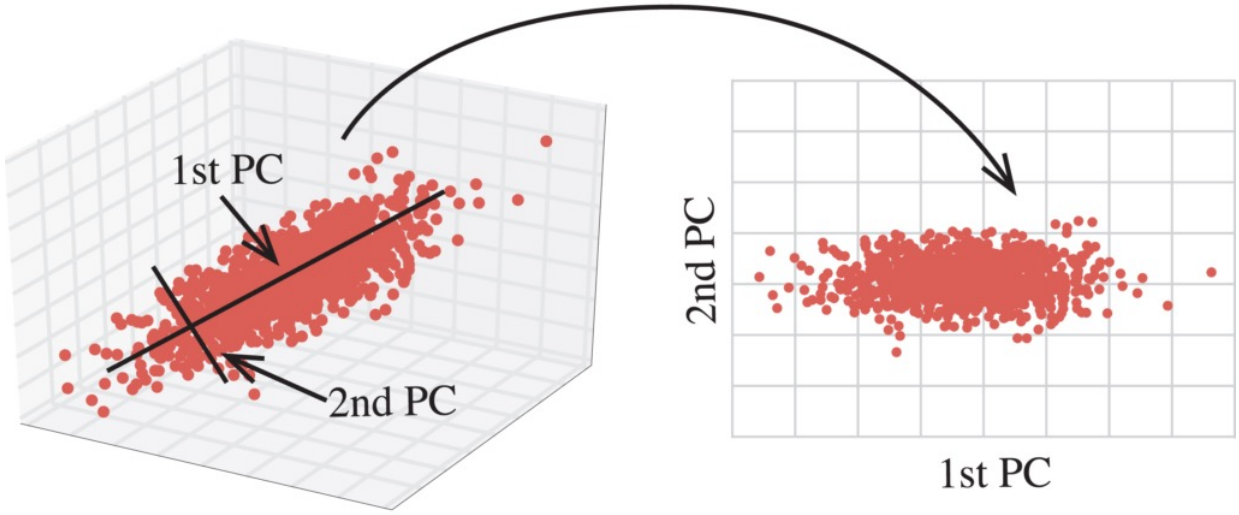
颜色表示不同分组

PCA是什么

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







High dimension

Complexity

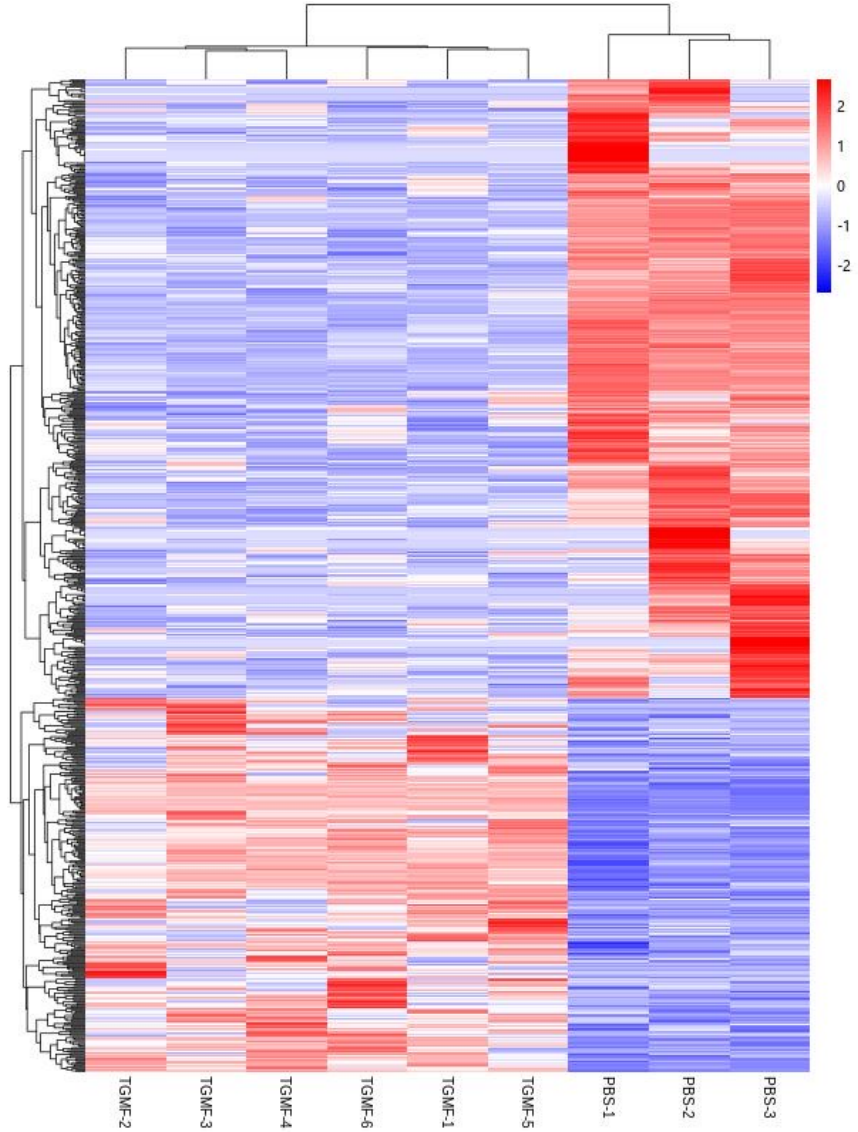
Simplicity

Low dimension

对数据的”降维打击”，以达到视觉可及的程度

也是对数据的简化，方便理解数据的组成

样本维度-表达热图



不同基因和样本的表达量差异

行：基因名；列：样本名

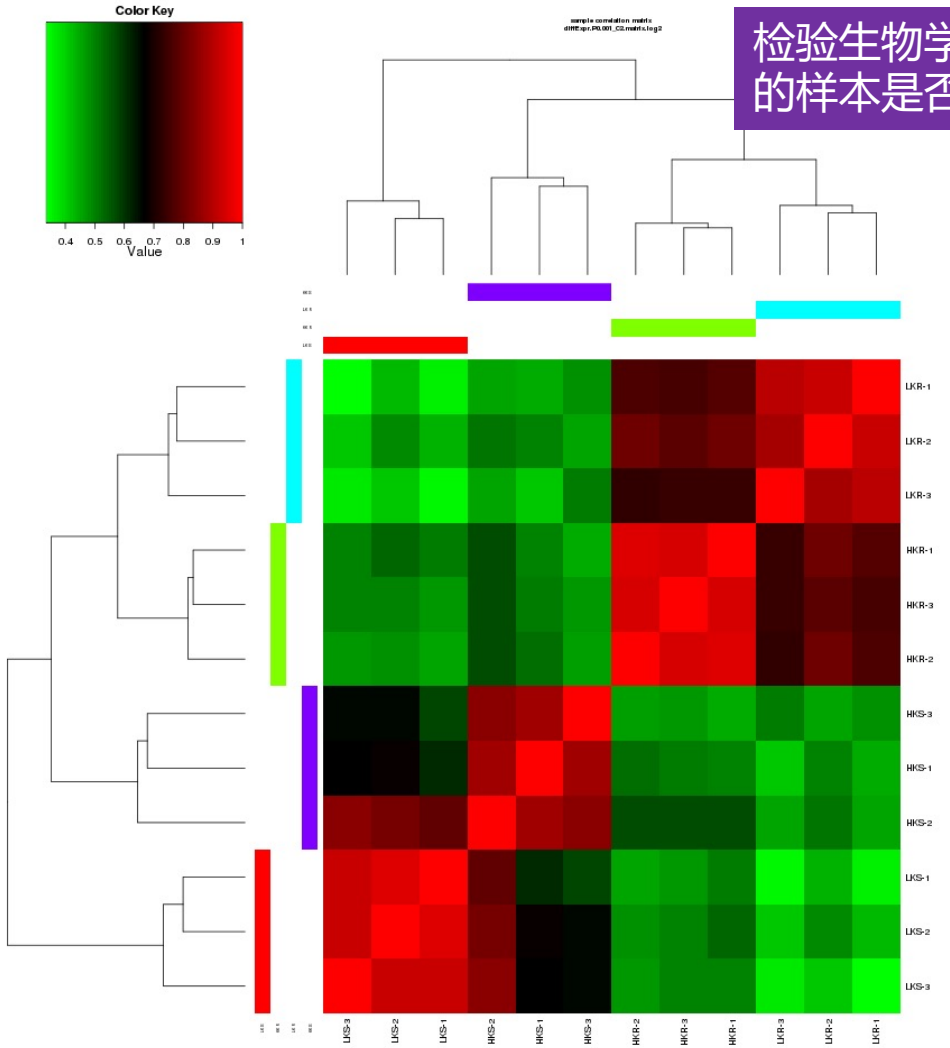
红色：高表达；绿色：低表达

行聚类：基因在不同样本中的表达相似

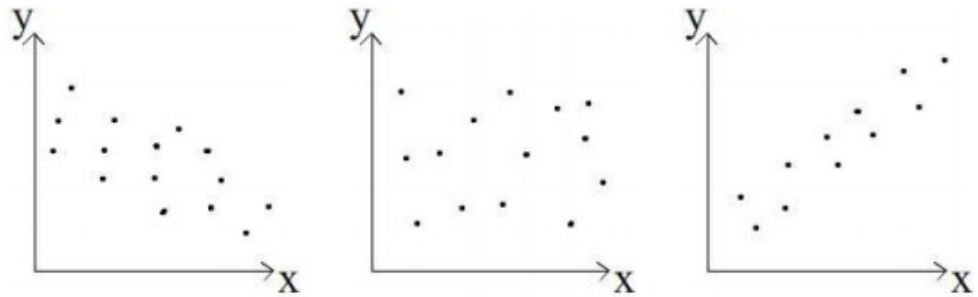
列聚类：不同样本的基因表达比较相似

标准化 Z-score $z = \frac{x - \mu}{\sigma}$

样本维度-样本相关性



检验生物学重复的样本是否合理

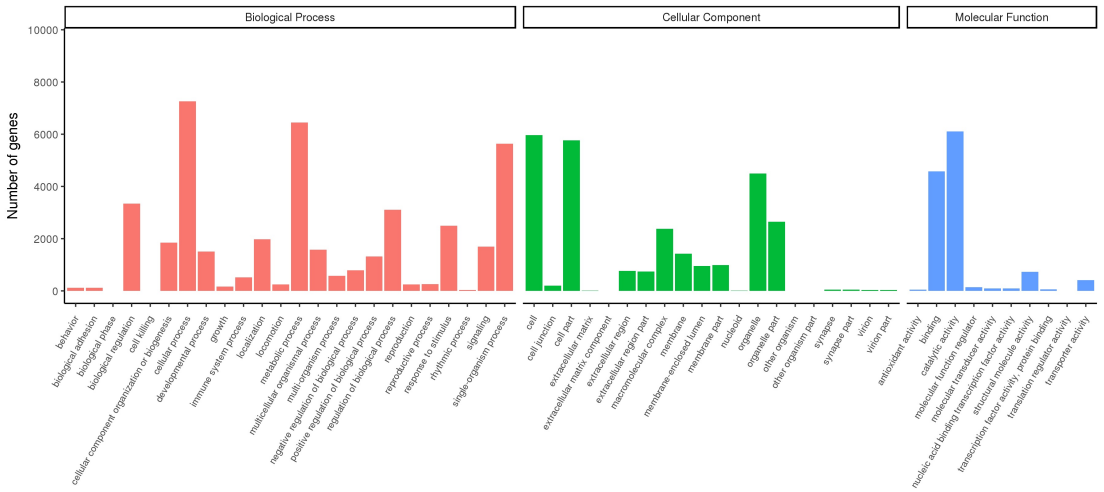


1、只表相关，而不表因果 2、相关亦有强弱之分

基因维度-基因功能注释

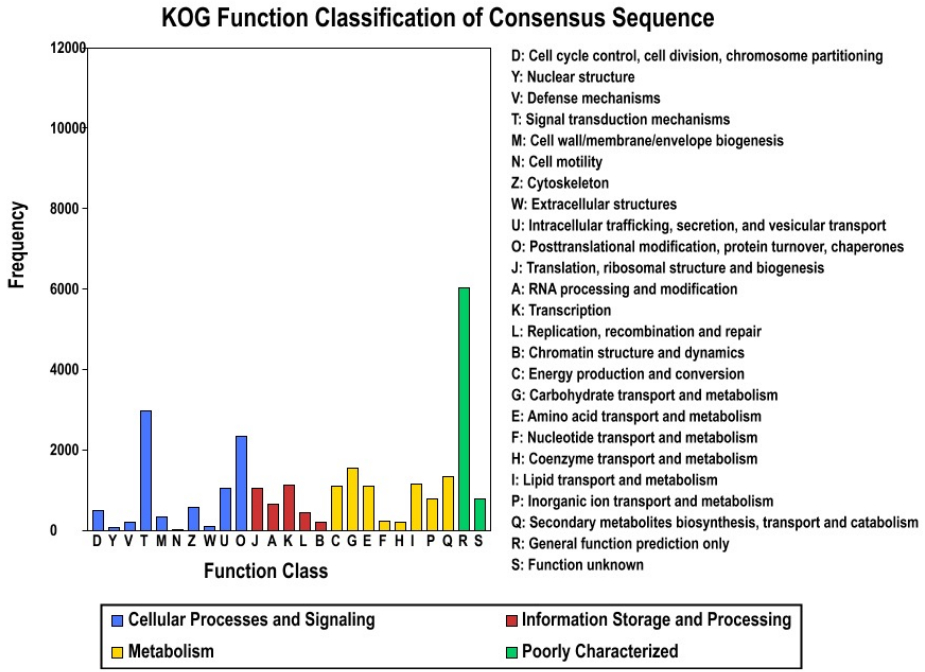
GO Gene Ontology：相似的基因，保守的功能

- 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修复或信号转导】



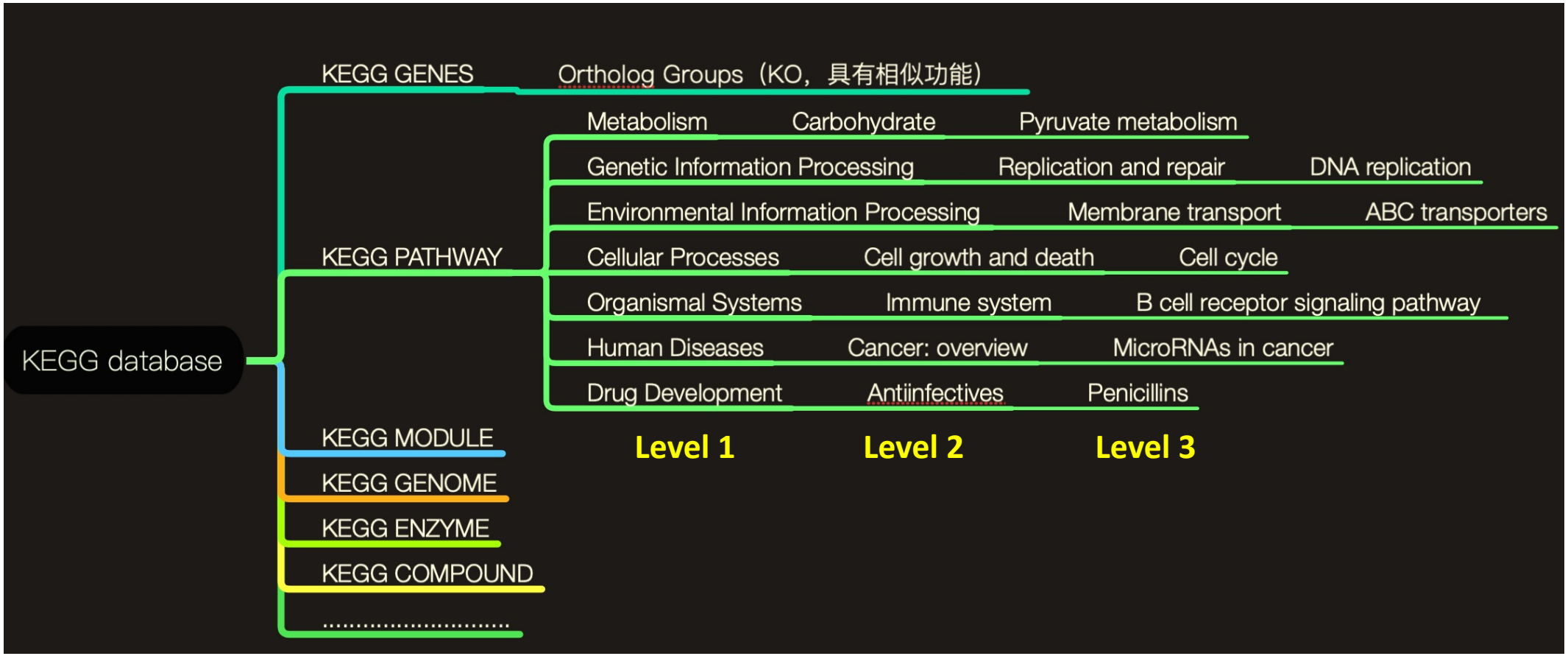
KOG Eukaryotic Orthologous Groups of proteins

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25个功能单元。



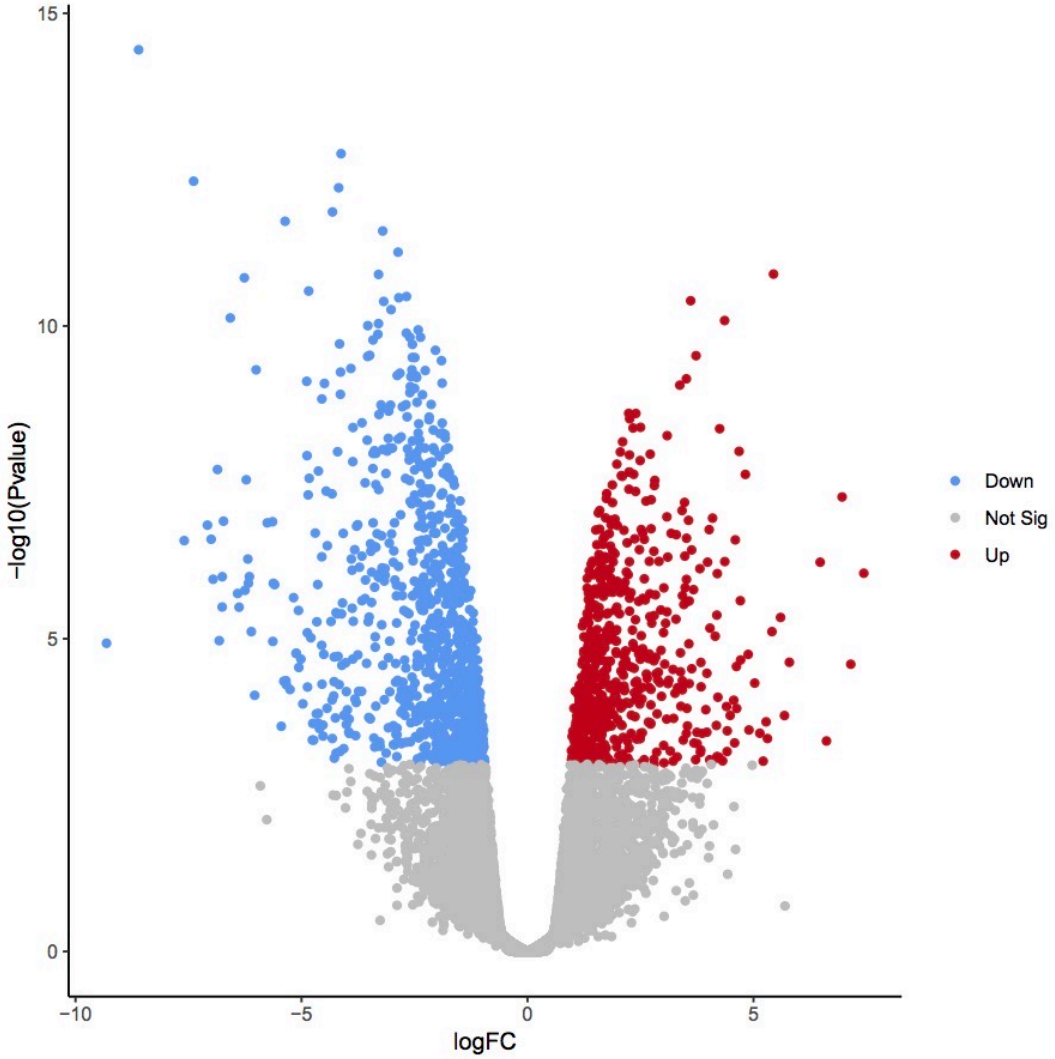
基因维度-基因功能注释

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

基因维度-差异表达基因



对照-处理组的差异基因有多少、有哪些

圆点：某个基因；颜色：蓝色下调；红色：上调；灰色：差异不显著

横坐标：**log2 (fold change)**

纵坐标：纵坐标：-log10 (FDR)

着色：FDR<0.05 & log2|FC|>1

Deseq2 and edgeR

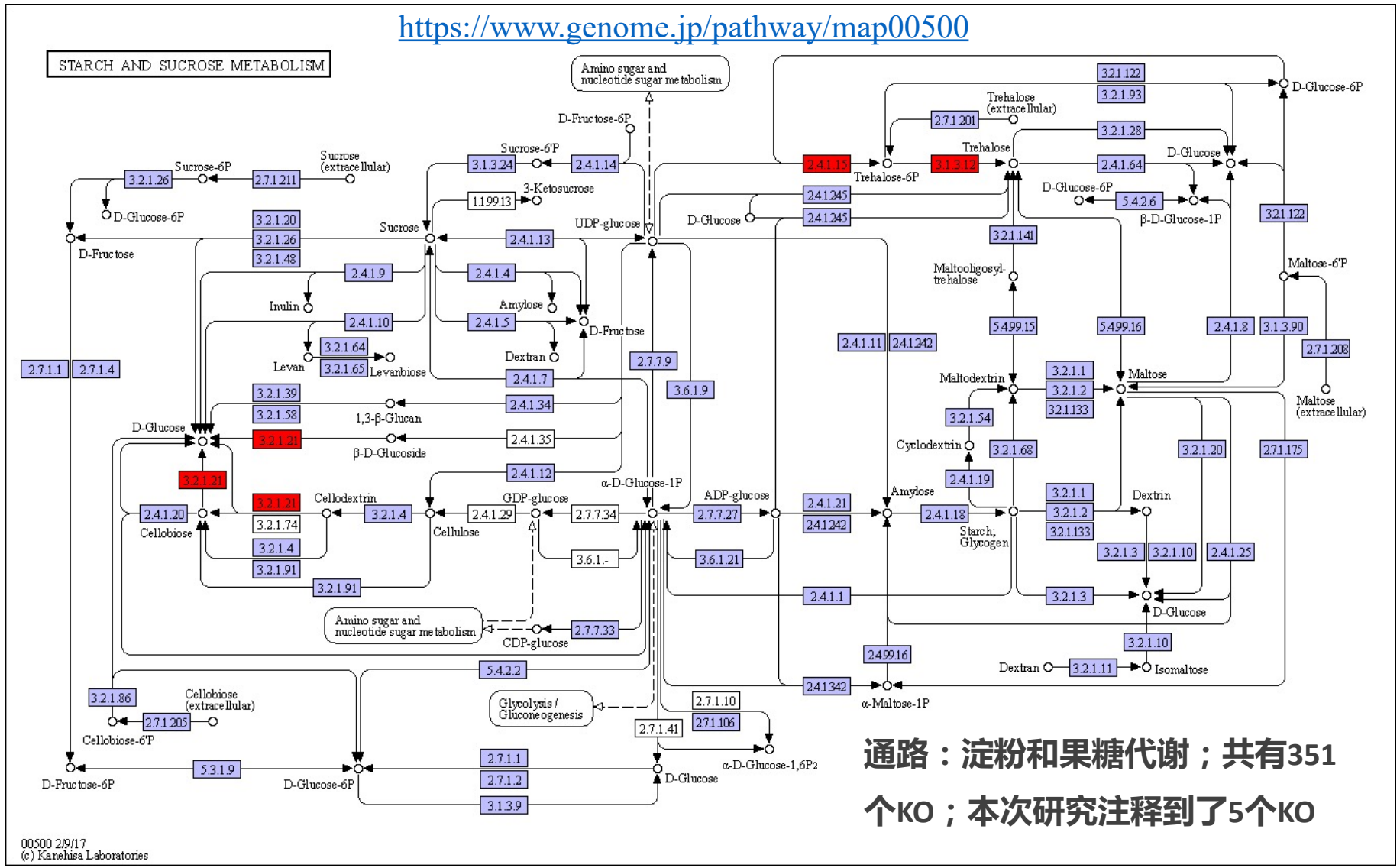
Fold change为什么用常用log转化？

(0, 1)　　(1,+∞)

(-∞, 0)　　(0,+∞)

基因维度-KEGG代谢通路

https://www.genome.jp/pathway/map00500



通路：淀粉和果糖代谢；共有351个KO；本次研究注释到了5个KO

圆圈：化学物质

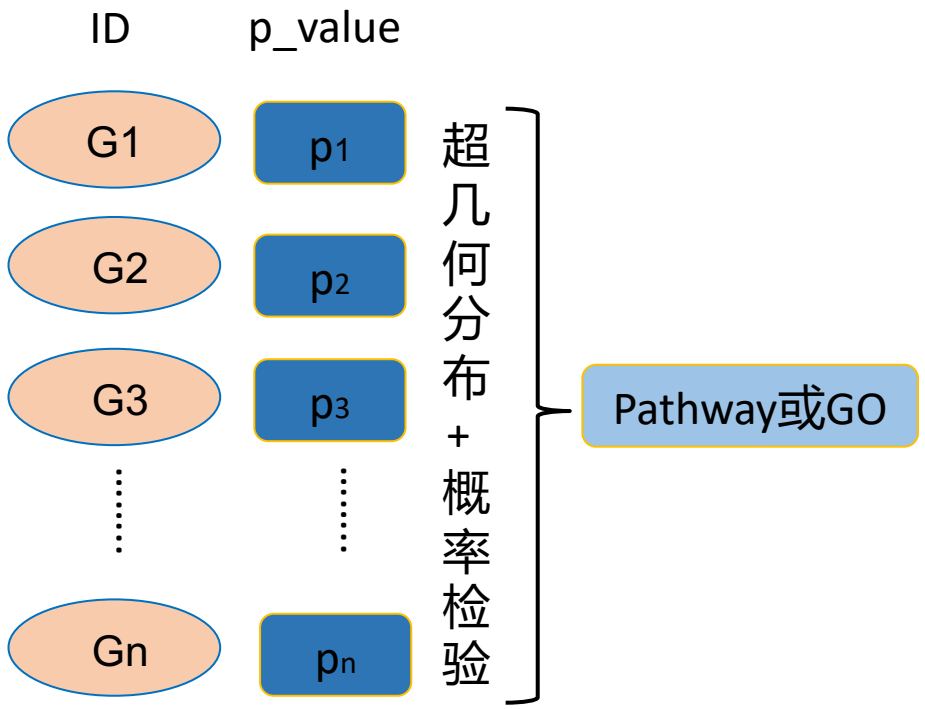
矩形：基因产物（酶或蛋白）

蓝：可超链接；

红：注释到该KO

白：无法链接

基因维度-富集分析



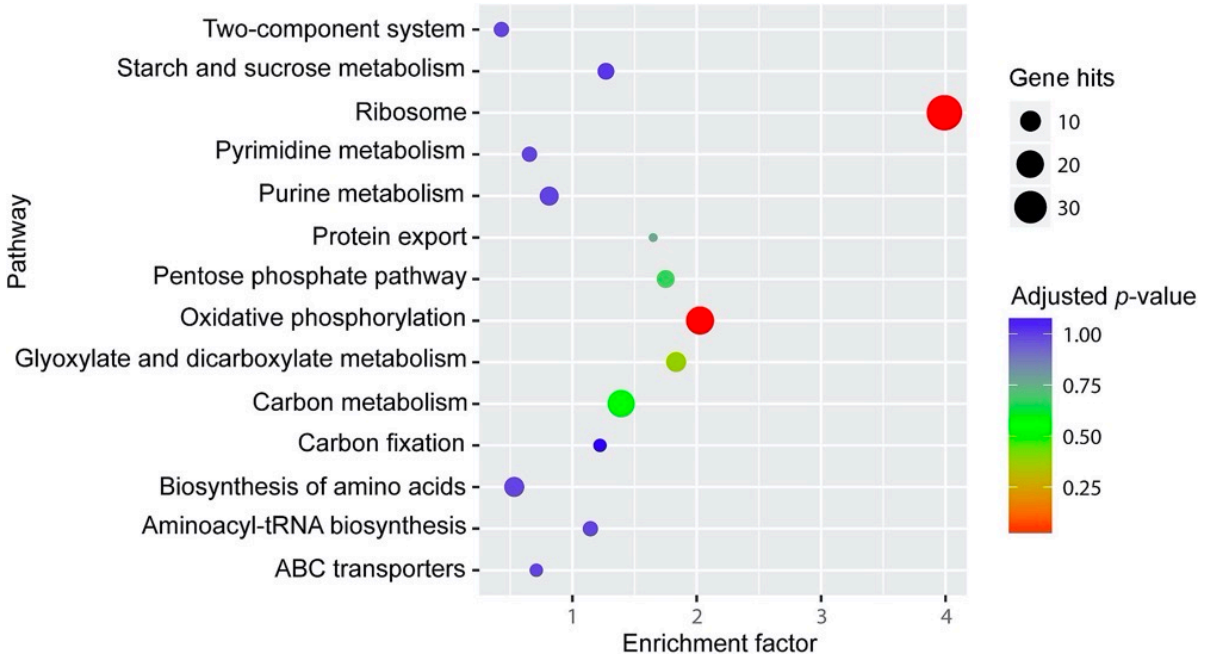
信号：在某个Pathway或GO term中的差异基因比例

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

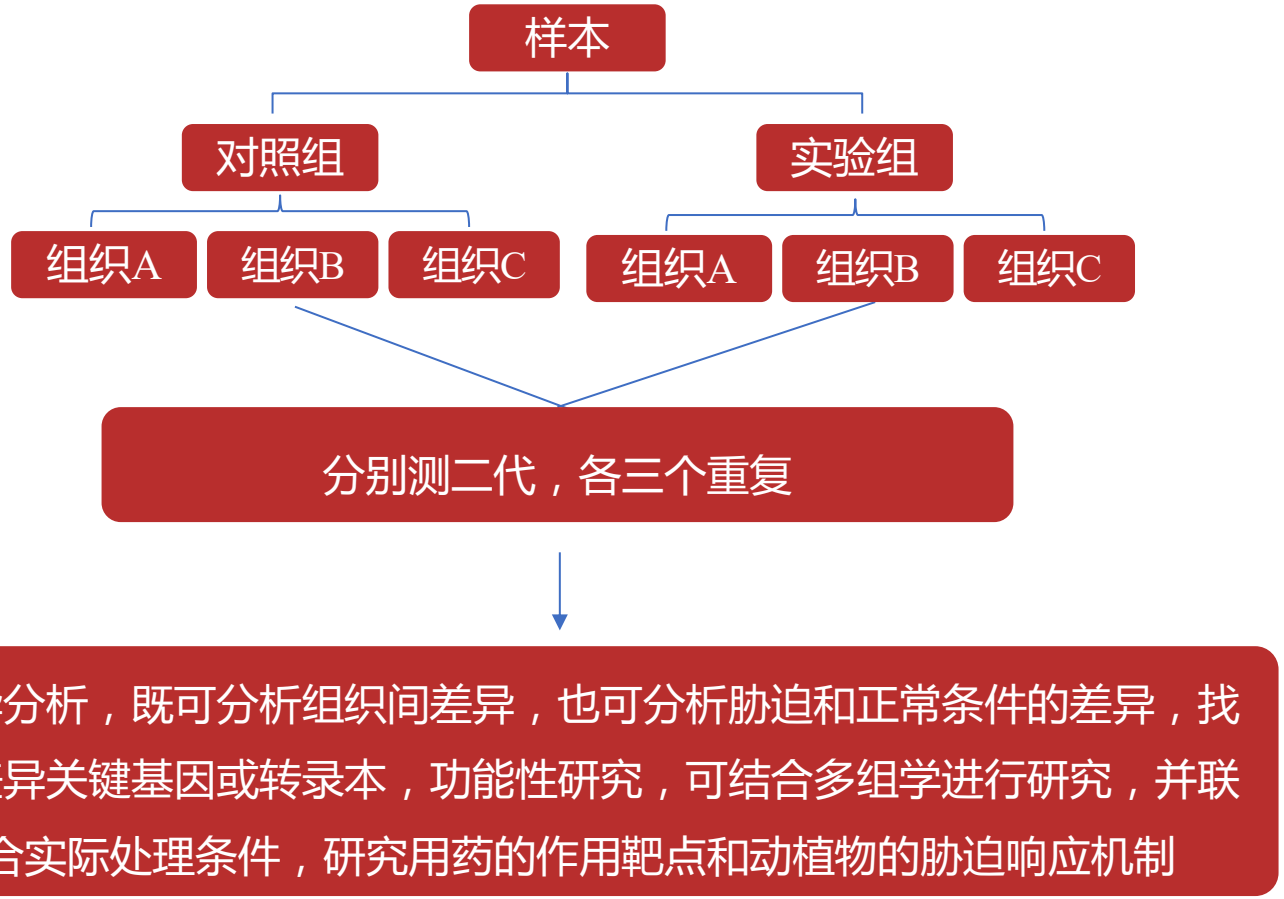
背景：国家富裕人数占总人口的比例



4

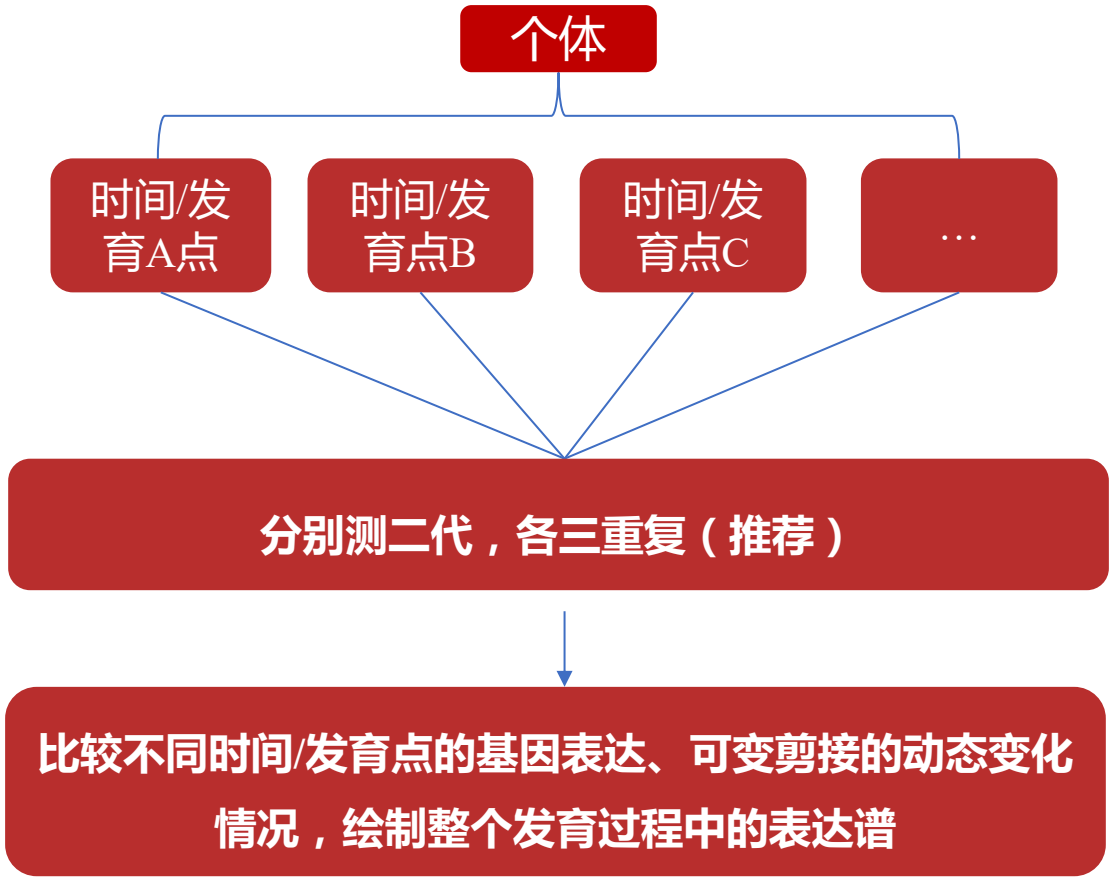
RNA-seq研究思路

差异转录组

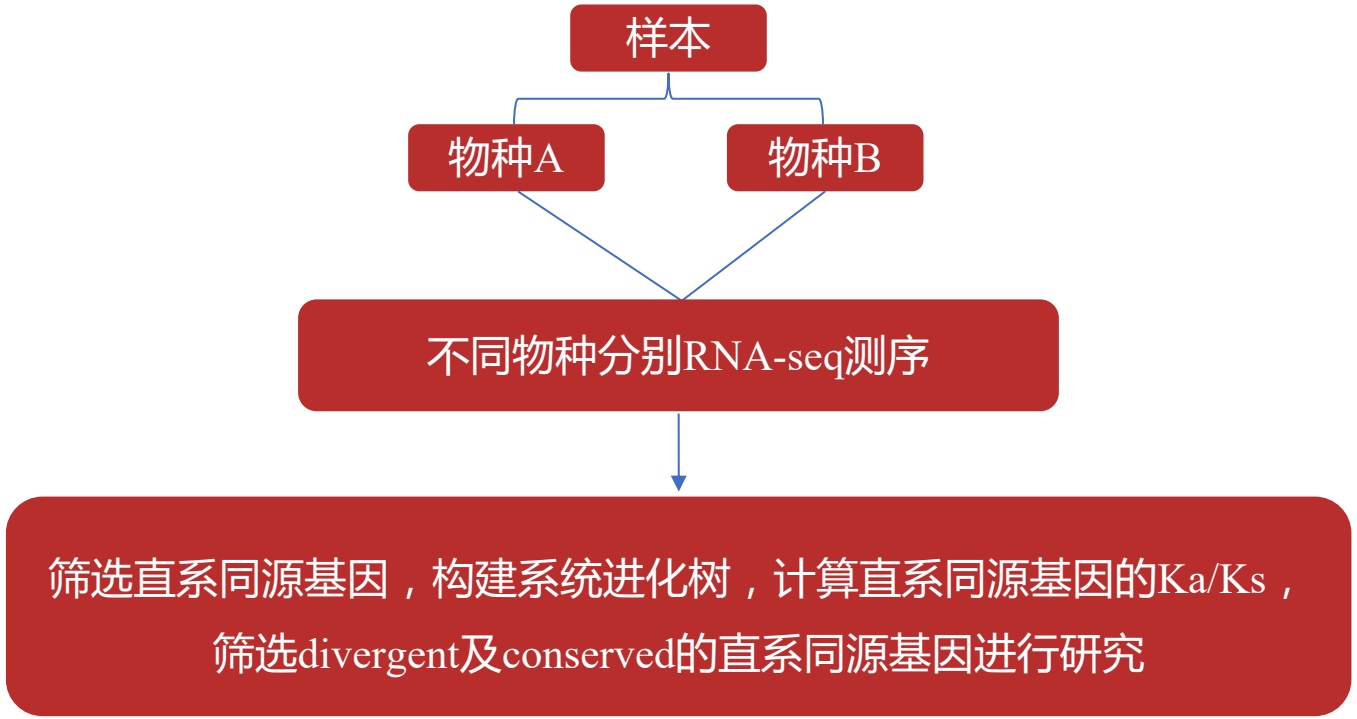


注：样本选择为同一物种的正常组和实验组。

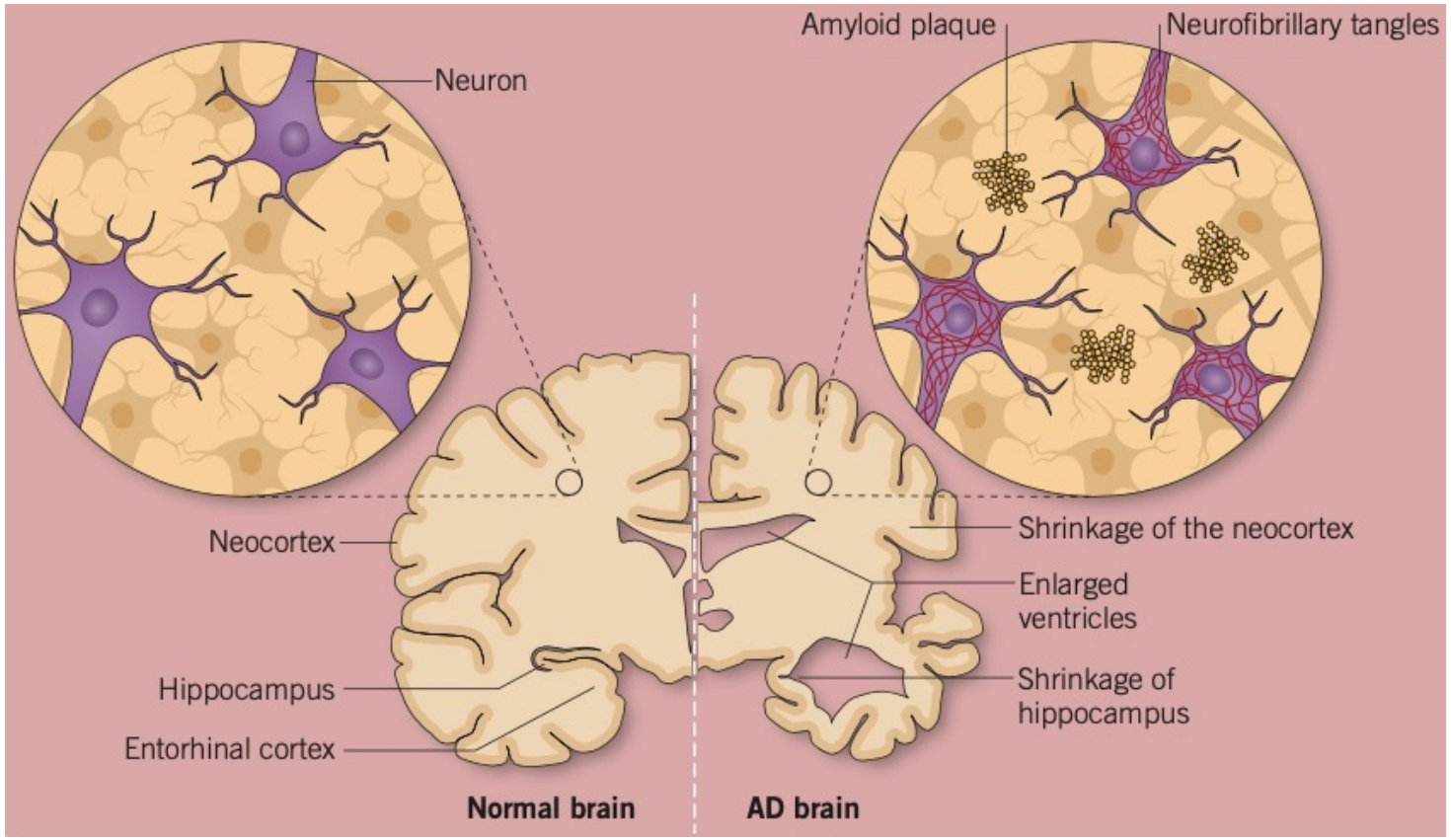
动态转录组


个体
时间/发育A点
时间/发育点B
时间/发育点C
…
分别测二代，各三重复（推荐）
比较不同时间/发育点的基因表达、可变剪接的动态变化情况，绘制整个发育过程中的表达谱

比较转录组


样本
物种A
物种B
不同物种分别RNA-seq测序
筛选直系同源基因，构建系统进化树，计算直系同源基因的Ka/Ks，
筛选divergent及conserved的直系同源基因进行研究

案例解读



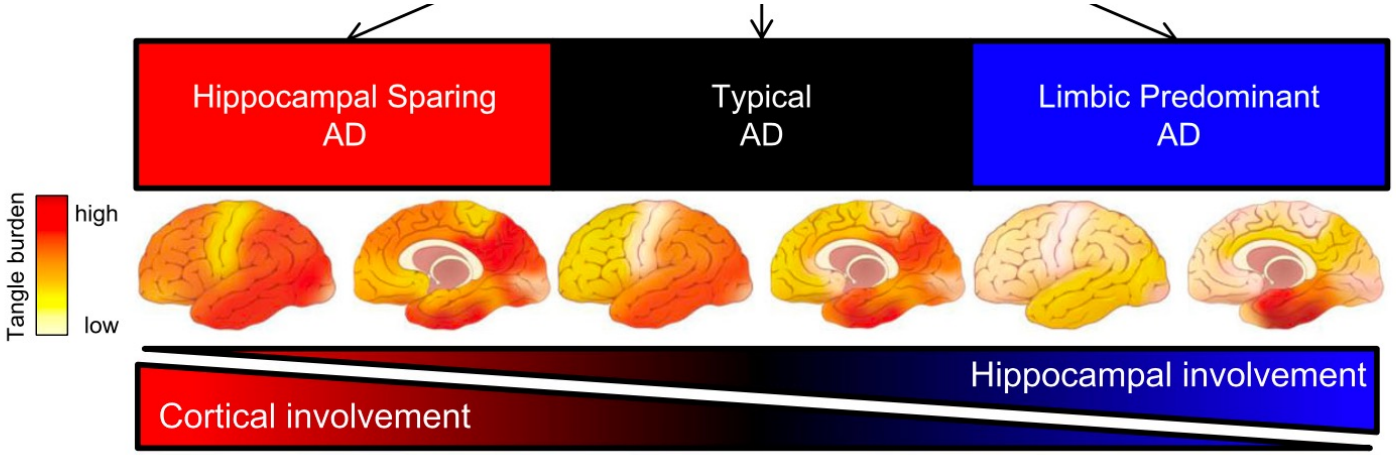
阿尔兹海默症（Alzheimer's disease ，AD）病例假说：β淀粉样蛋白（Aβ）和Tau（微管蛋白）；Aβ形成斑块和Tau过度磷酸化形成神经纤维缠结，进而影响神经元健康，导致神经退行性症状。

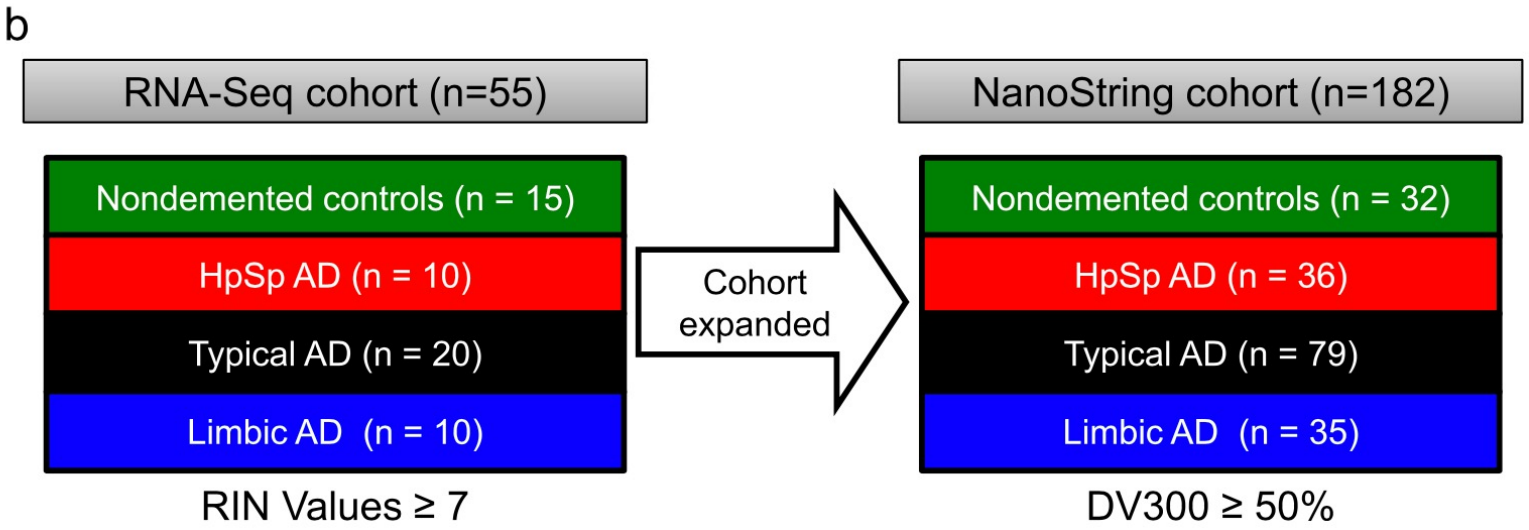
Aβ沉积：皮层到边缘系统

Tau：皮层下核团/内嗅皮层到边缘系统

鉴定不同AD亚型中的基因表达谱

实验设计



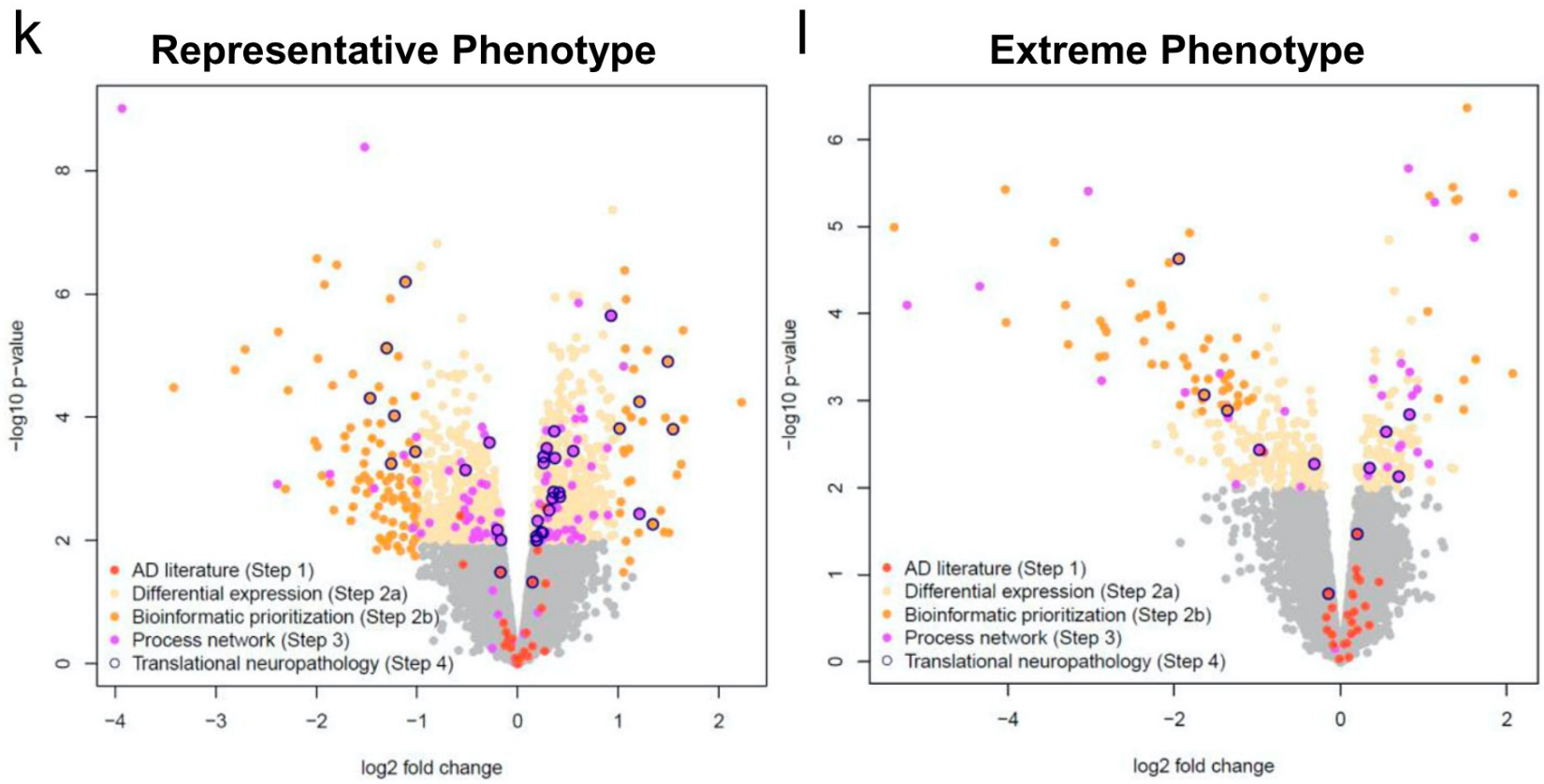


三种AD的亚型：

海马保护型AD（皮层严重，海马较轻）；**典型的AD**（海马和皮层均有病变）；**边缘主导型AD**（海马严重，皮层较轻）

海马区域易感基因

主要结果



AD中海马易感性的基因；发现了SERPINA5, RYBP, SLC38A2, FEM1B和PYDC1可以很好的区分AD和对照组

典型AD（与对照组相比）和边缘主导型AD（与海马保护型AD相比）中，大多数基因都是下调的。

在典型AD中的差异基因多与细胞生长和增殖相关，而在边缘主导型AD中大多数基因与炎症和免疫反应相关

今天讲了什么

Illumina测序原理

mRNA文库构建

RNA-seq解决的问题

RNA-seq的分析流程

RNA-seq的分析内容

RNA-seq的研究思路

一篇AD的案例

BerryGenomics
贝瑞基因

Thank You!

官 方 网 站

官 方 微 信

www.berrygenomics.com