基因组分析-新一代测序技术应用

主要测序技术

- 第一代测序技术
  - Sanger sequencing (1980's)
- 第二代测序技术(next generation sequencing, NGS)
    - Roche/454 (2005)
    - Illumina/Solexa (2006)
    - Life/APG's SOLiD (2007)
    - Life/APG's Ion torrent (2010)
- 第三代测序技术
  - 单分子测序

高通量测序简介

一次性对几百万到十亿多**DNA分子**进行并行测序，又称下一代测序技术，

其使得可对一个物种的转录组和基因组进行深入，细致和全貌的分析，

所以又称为深度测序。

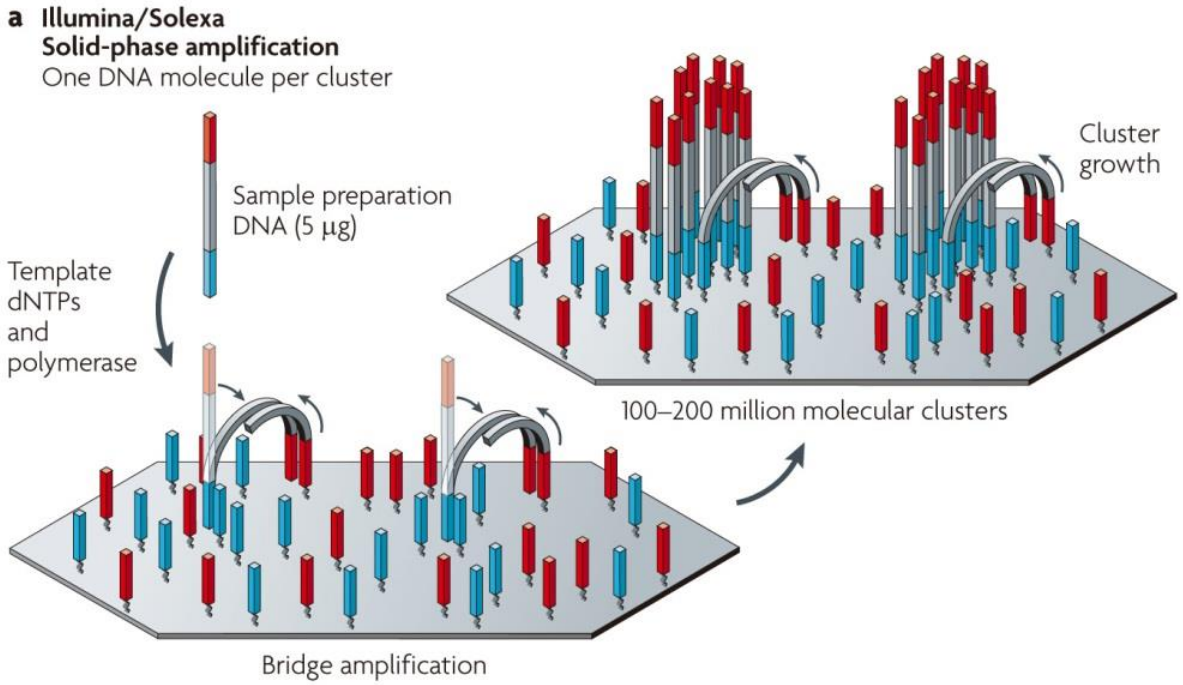
High-throughput Sequencing

Next Generation Sequencing

Deep Sequencing

测序仪品牌	技术原理	开发商
Roche 454	焦磷酸测序	Roche
Illumina Solexa	边合成边测序	Illumina
ABI SOLiD	基于磁珠的大规模并行连接测序	ABI
Helicos	单分子荧光测序	Helicos
Ion Torrent	半导体测序	ABI
SMRT	单分子实时测序	Pacific Bio

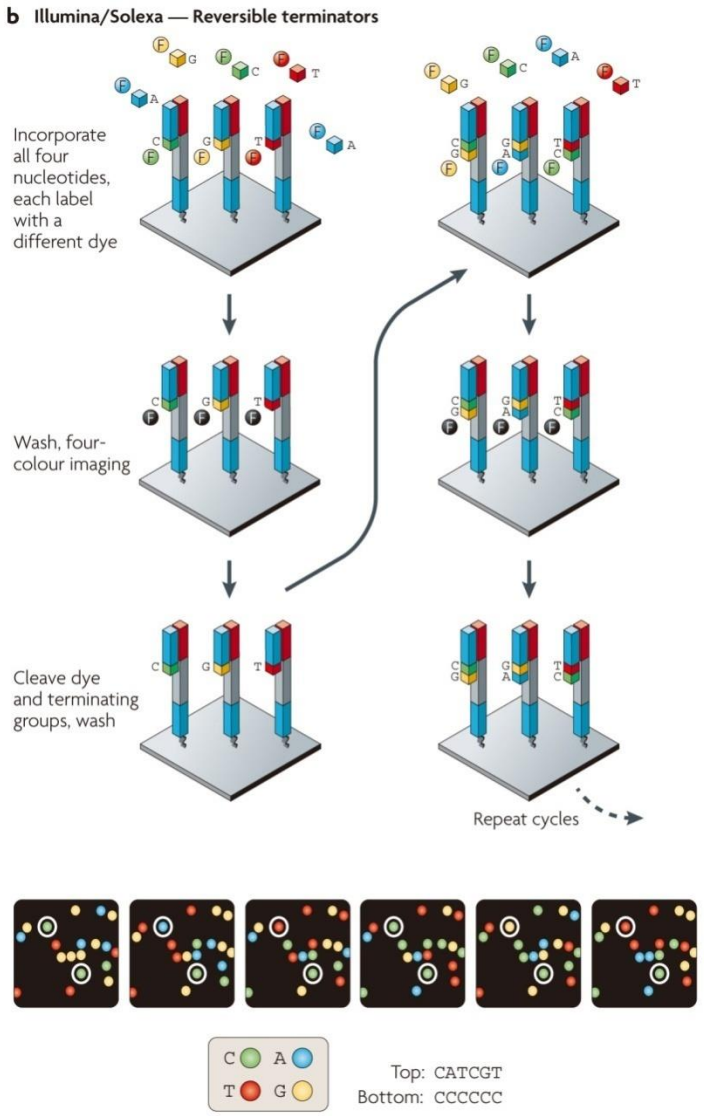
Illumina/Solexa



- 单链DNA两端加上非对称的通用接头(包括测序引物)，接头与事先固定在固相芯片表面的序列互补
- 单链DNA结合到芯片表面形成桥式结构。然后使用接头引物进行PCR扩增
- 变性后在一个芯片上可以形成上亿个不相关的单链DNA分子簇，其一端固定在芯片表面，另一端是自由的

Illumina/Solexa

- 使用测序引物从自由的通用接头一侧开始测序反应。
- 测序使用的dNTP每种碱基被不同的荧光基团标记，同时脱氧核糖的3’-OH被封闭，这样每轮测序循环只能延伸一个核苷酸。读取碱基荧光信号，就能知道这一轮每个簇结合上的是什么核苷酸
- 然后切除荧光基团，打开被封闭的3’-OH，继续进行下一轮反应



Illumina/Solexa

- 主要错误来源：同一个簇内不同DNA链延伸情况不同（相位差），导致读取错误
- 优势：通量最高 (max 600Gb, HiSeq 2500)
- 劣势：读长较短 (max 250bp, HiSeq 2500)，运行时间长(1-14 days，HiSeq 2500大幅提升了运行速度)，数据存储和分析难度大。

HiSeq 2000

Genome Analyzer II

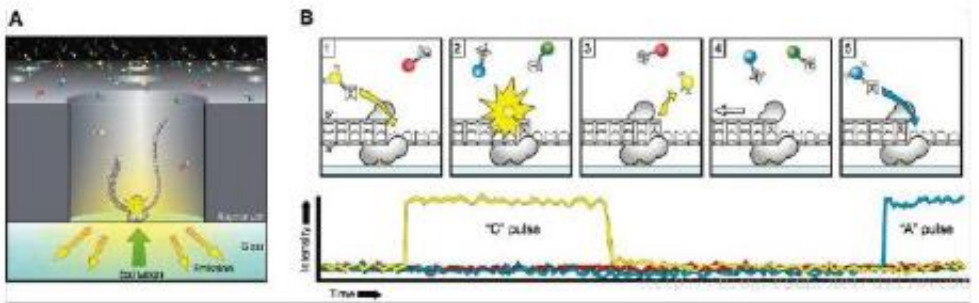
MiSeq

第三代测序技术原理

- 第三代测序技术原理主要分为两大技术阵营：第一大阵营是单分子荧光测序，代表性的技术为美国螺旋生物(Helicos)的TSMS技术和美国太平洋生物(Pacific Bioscience)的SMRT技术。
- 第二大阵营为纳米孔测序，代表性的公司为英国牛津纳米孔公司。新型纳米孔测序法（nanopore sequencing）是采用电泳技术，借助电泳驱动单个分子逐一通过纳米孔 来实现测序的。

三代测序—SMRT

- PacBio SMRT技术应用了**边合成边测序**的思想，以SMRT芯片为测序载体。
- 4色荧光标记 4 种碱基，根据光的波长与峰值可判断进入的碱基类型。
- 通过检测相邻两个碱基之间的测序时间，来检测一些碱基修饰情况，可以通过这个来之间检测甲基化等信息。
- SMRT技术的测序速度很快，每秒约10个dNTP。
- DNA 聚合酶是实现超长读长的关键之一。
- 测序错误率比较高，通过多次测序纠错。

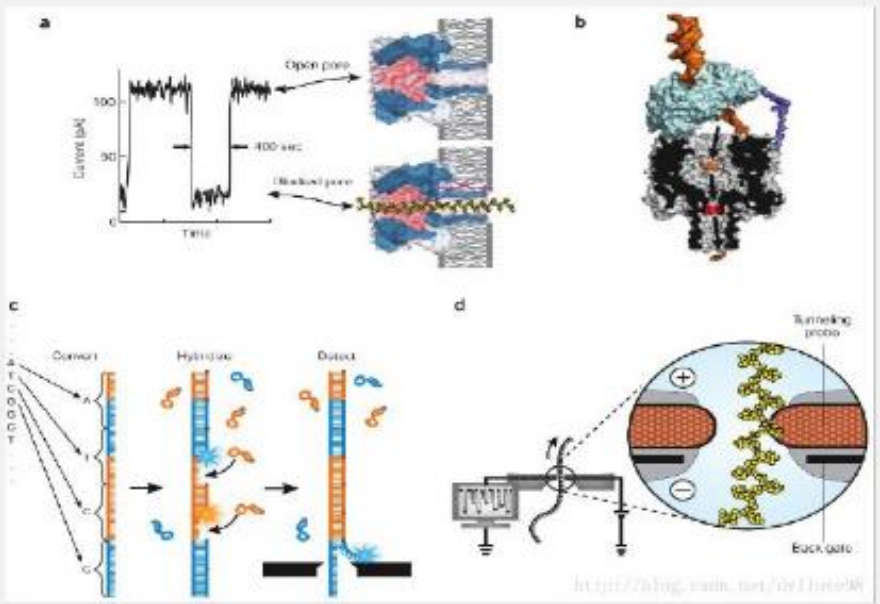


三代测序—Nanopore

Oxford Nanopore Technologies公司所开发的纳米单分子测序技术，是基于电信号的测序技术。

基本原理

当DNA碱基通过纳米孔时，使电荷发生变化，从而短暂地影响流过纳米孔的电流强度（每种碱基所影响的电流变化幅度是不同的），灵敏的电子设备检测到这些变化从而鉴定所通过的碱基。



Nanopore 主要特点

- 读长长，大约在几十kb，甚至100 kb
- 错误率目前介于1%至4%，且是随机错误，而不是聚集在reads的两端;
- 数据可实时读取；
- 通量很高(30x人类基因组有望在一天内完成)；
- 样品制备简单又便宜，不破坏DNA;
- 直接读取甲基化修饰位点，准确率可达99.8%

第二/三代测序技术平台

平台	原理	扩增方法	测序长度(bp)	通量(Gbp)	费用($)	时间
454 FLX	焦磷酸测序	乳滴PCR	250-500	0.1-0.3	8,500	7.5hr
Illumina HiSeq2000	合成测序法	桥式PCR	50-150	200	20,000	7 d
SOLiD	连接介导测序法	乳滴PCR	50	300	20,000	7 d
Ion Torren/Proton	质子测序	乳滴PCR	综合性能与Illumina Hiseq系列相当			
HeliScope	单分子测序	无扩增	30	7.5	18,000	14 d
PacBio SMRT	实时单分子测序	无扩增	读长最长，但错误率高（10%），通量有限			
Oxford Nanopore	纳米孔	无扩增	成本低，仪器体积小，但稳定性有待时间验证			

454 FLX

Illumina HiSeq2000

SOLiD

HeliScope

PacBio SMRT

Oxford Nanopore

Cite Holt A. et al., Genome Res. 2008 and others

高通量测序技术的应用

- DNA测序
  - 全基因组从头 (*de novo*) 测序
  - 基因组重测序 (Resequencing)
  - 宏基因组(Metagenome)测序
  - 外显子测序
  - ChIP-Seq
  - 甲基化测序
  - 大规模筛选遗传标记
- RNA测序
  - 转录组测序
  - 电子表达谱分析
  - 小RNA测序
  - 降解组测序
  - CLIP-Seq
  - PET-Seq

DNA测序

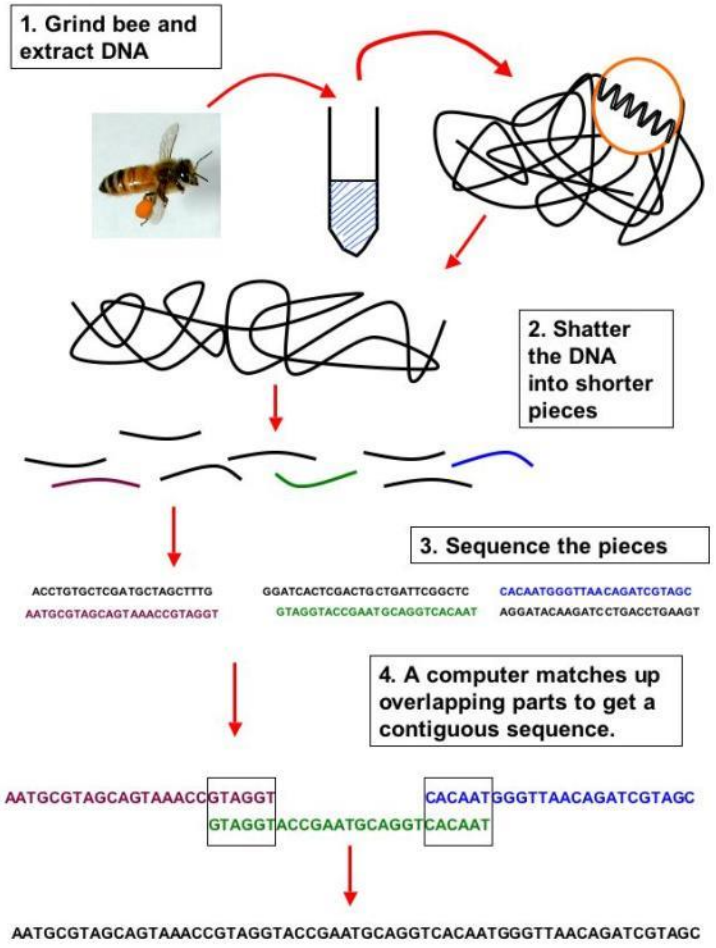
基因组*de novo*测序

- 对未知基因组序列的物种

  取样:

  - 动物：血液、肌肉
  - 植物：叶片（黄化叶，组培植株）
- 估算基因组复杂度（大小、重复序列比例、杂合度）
- 测序技术:
  - Illunima paired-end为主
  - Sanger、454、SOLiD为辅,
  - PacBio目前也开始用于基因组测序补洞
- 文库构建
  - 尽量随机打断

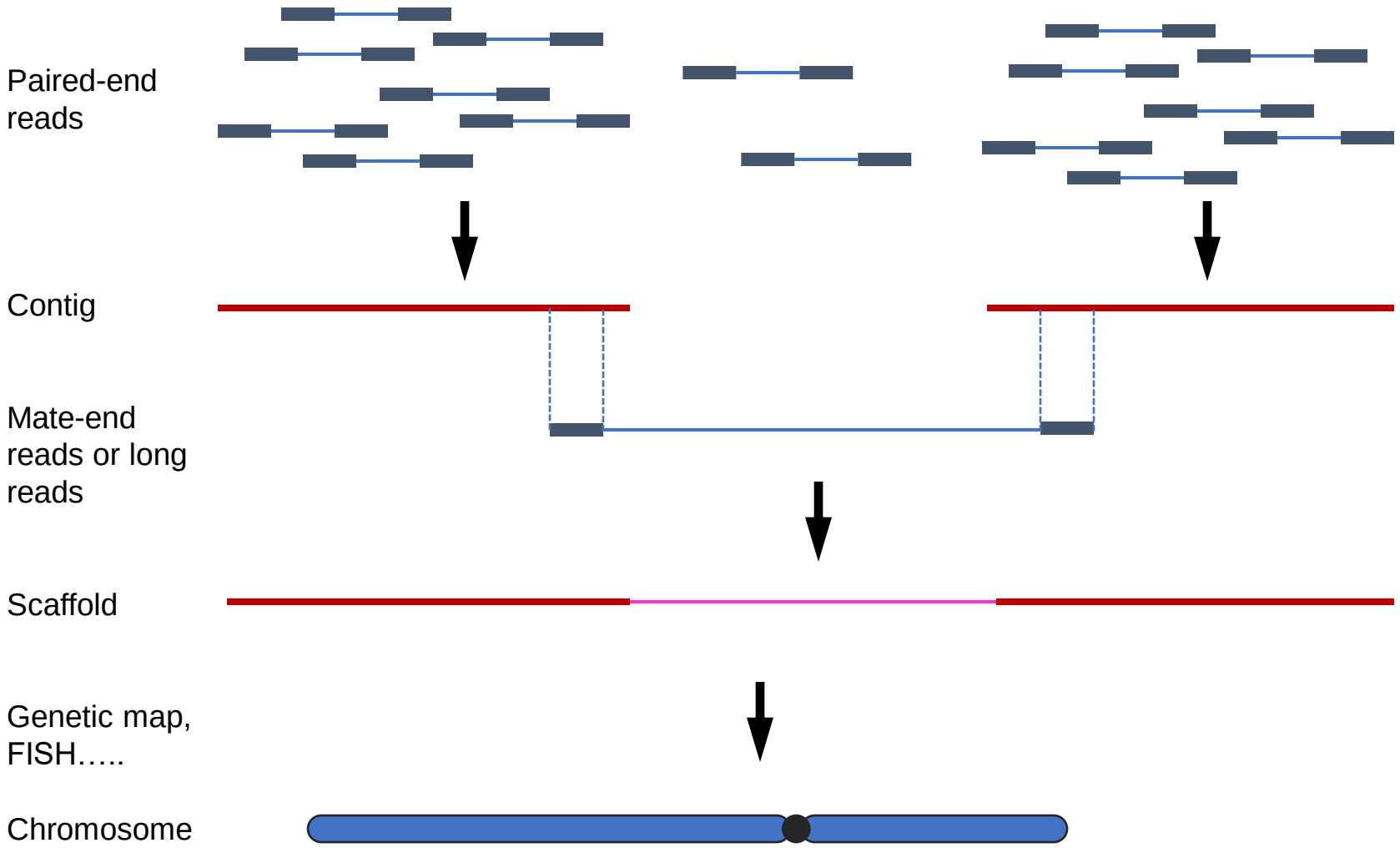
WGS (whole genome shortgun)



- Coverage depth（覆盖深度or测序深度）：
  每个碱基被测序的平均次数，是用来衡量测序数据量的首要参数。
  测序总数据量/基因组大小
- Coverage ratio（覆盖率）：
  被测序到的碱基占全基因组大小的比率。覆盖比率随覆盖深度升高而提高，亦受测序bias的影响，如illumina测序会受到GC bias的影响，而导致测序不均匀。

- 理论上（完全随机打断）测序深度达到20x即可覆盖整个基因组。实际工作中一般需要50x以上（100 bp读长）。
- Reads长度越长越好。

De novo assembly



全基因组成功测序案例1

- 大熊猫基因组（华大基因等，2009）

- 完全使用NGS测序组装

- 确定大熊猫属于食肉目熊科
- 发现大熊猫*T1R1*基因与人和狗相比，两个外显子发生移码突变，成为假基因。该基因与感受鲜味有关，推测是大熊猫不吃肉的主要原因之一。
- 大熊猫个体杂合度较高，表明遗传多样性较高，没有严重的近亲繁殖现象，有利于种群延续。

重测序 (Resequencing)

- 对已有参考基因组物种的不同基因型或不同个体的全基因组或部分区段进行测序，以获得个体之间的基因组和功能差异。

- 用途：
  - 了解物种的起源和演化历程
  - 理解疾病的成因
  - 理解动植物性状的分子机制

} 全基因组关联分析 (Genome-wild association study, GWAS)

重测序分析一般步骤

- 取样
  - 个体数目：足够的个体数目，尤其对于GWAS
  - 测序深度：
    - 研究个体差异，>4-5x覆盖/个体
    - 研究群体差异，>0.5x覆盖/个体，>100x/群体
- 测序：
  - Illunima、SOLiD为主
- 回帖(mapping)
  - 常用软件：BWA, SOAP, Bowtie, MAQ, CLC Genommics Workbench…
- 寻找SNP和Indel (SNP and indel calling)
  - 常用软件：samtools, SOAPsnp, SOAPindel, GATK, CLC Genomics Workbench..

宏基因组 (Metagenome)

- 自然界中群聚的微生物往往很难分离
- 宏基因组:

  利用现代基因组技术直接研究自然状环境中的微生物有机体群落，而不需要经过分离、培养单一种类的微生物。

- 研究来自同一自然环境或同一宿主的整个微生物群落

- 分析其物种分类及功能，以及与环境胁迫或疾病的关联。

外显子组测序

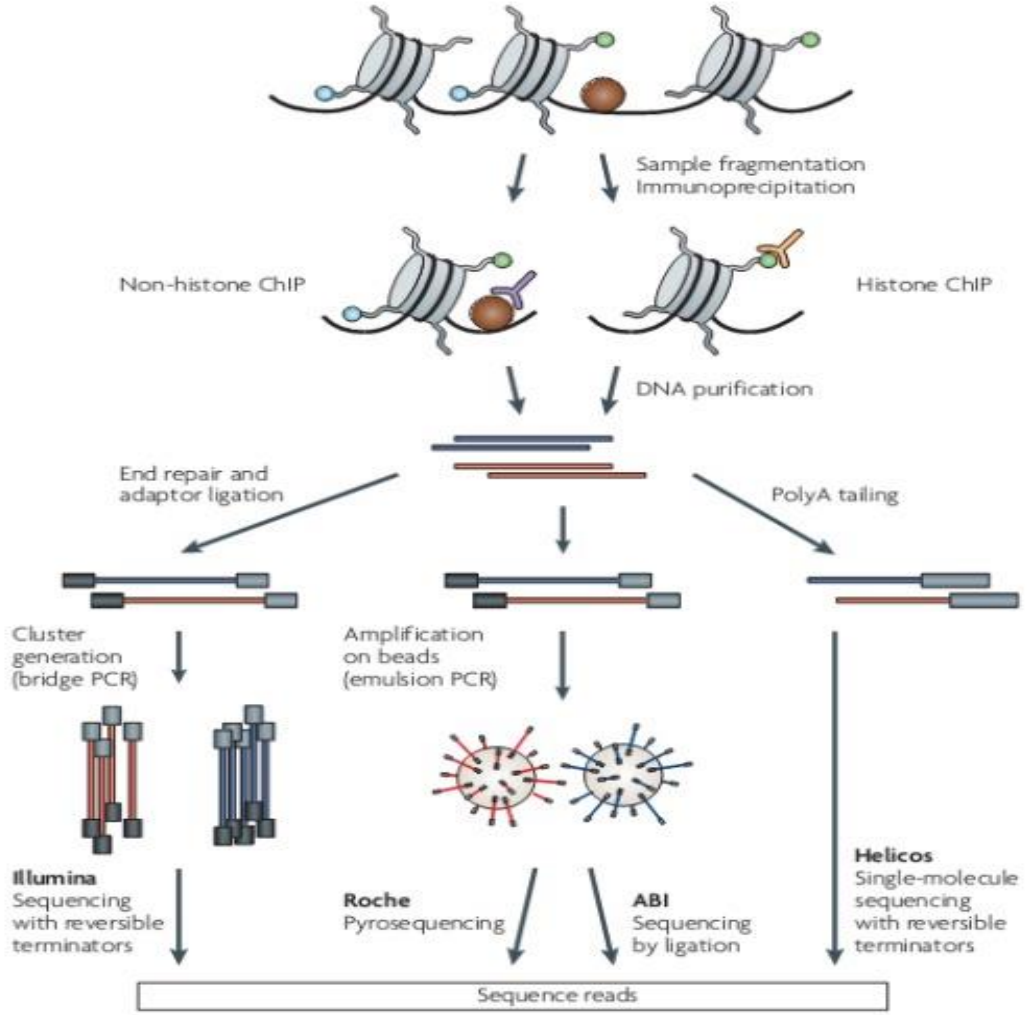
- 外显子组是指全部外显子区域的集合，该区域包含合成蛋白质所需要的重要信息，涵盖了与个体表型相关的大部分功能性变异。

- 与全基因组重测序相比，外显子组测序只需针对外显子区域的DNA，覆盖度更深、数据准确性更高，更加简便、经济、高效。

- 多用于人类癌症分析

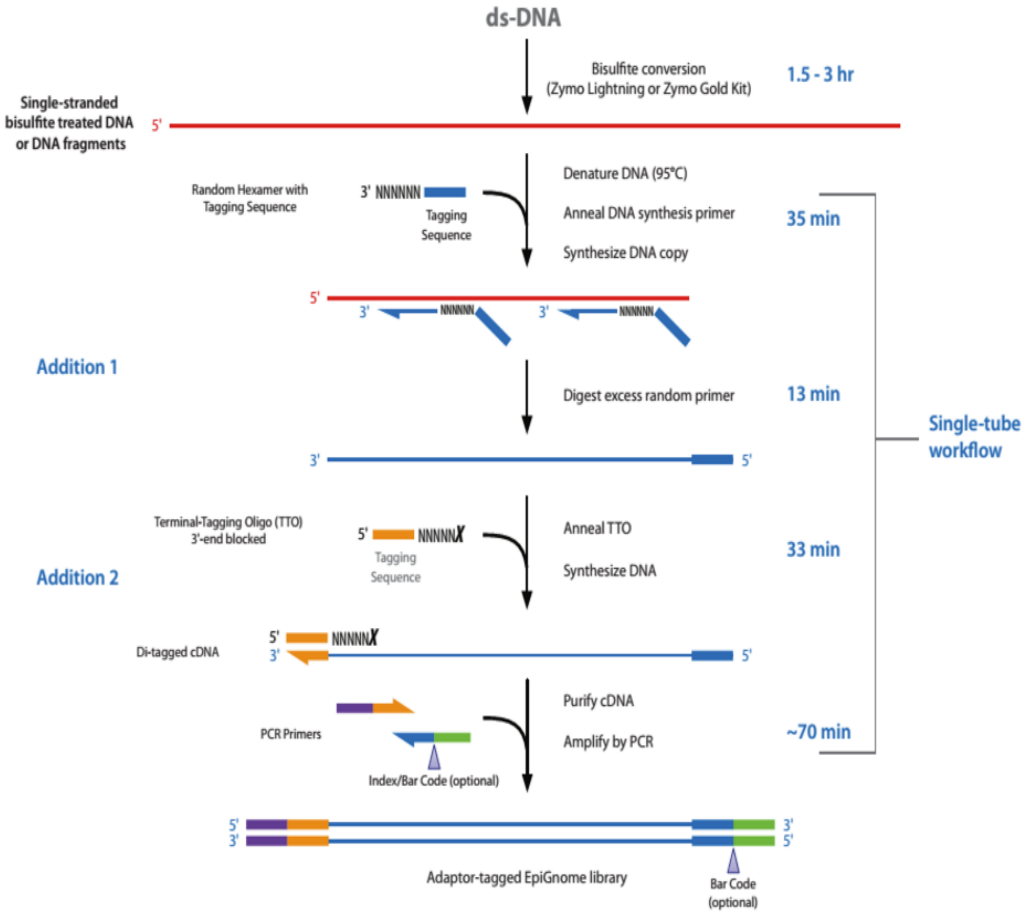
ChIP-Seq

- ChIP (Chromatin Immunoprecipitation) 染色质免疫共沉淀，是指通过蛋白免疫相互作用，用抗体把和染色质相互作用的蛋白，如组蛋白、转录因子等，沉淀下来，从而所获取与其相结合的DNA序列。
- ChIP-Seq就是通过高通量测序对ChIP所得到的序列进行测序，从而进行蛋白和DNA相互作用相关研究。
- 取代以往的ChIP-chip



- 注意：
  - 抗体质量（特异性和灵敏度）
  - 阴性对照

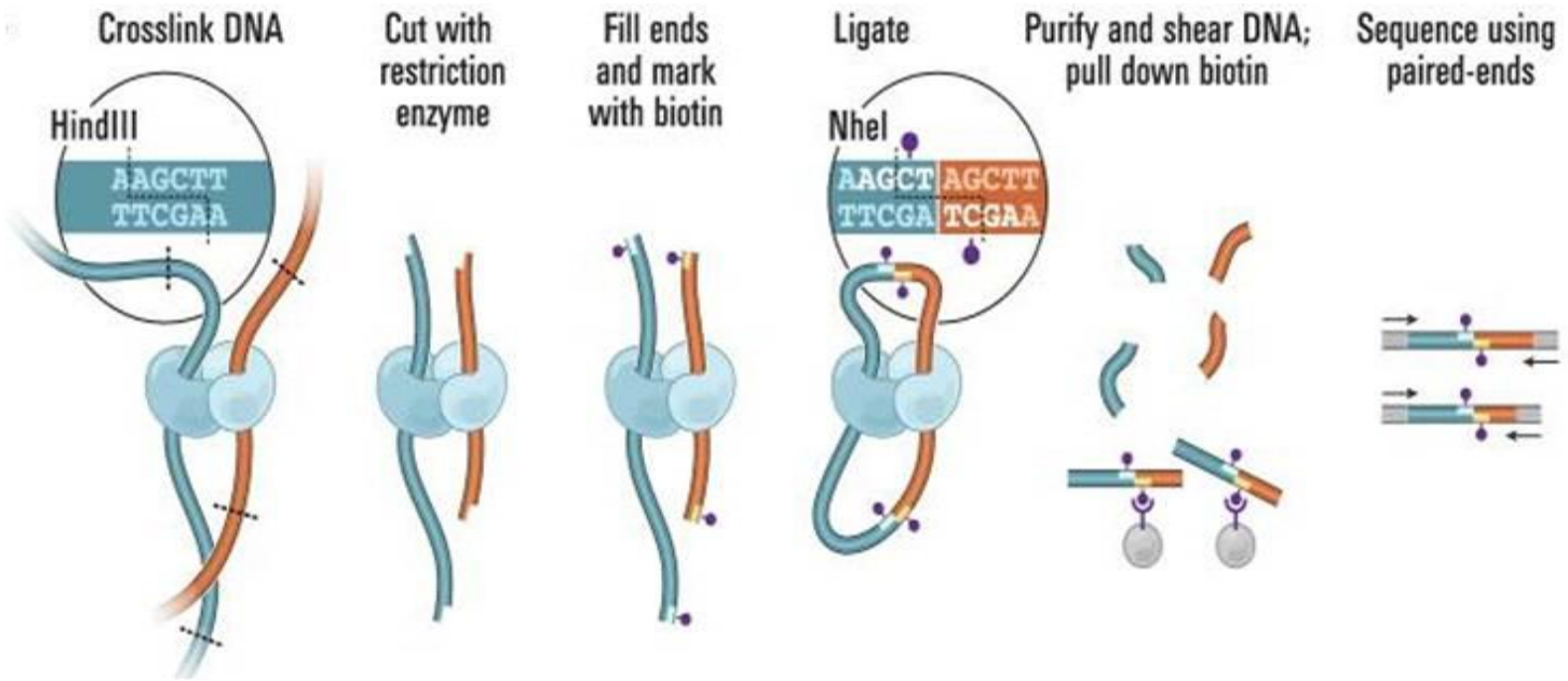
亚硫酸盐测序(BS-seq)



亚硫酸盐测序(Bisulfite sequencing，简称为BS-seq)是一种利用亚硫酸盐处理DNA来测定其甲基化模式的方法(Chatterjee et al., 2012)。DNA甲基化是最初被发现的表观遗传标记，直到现在仍是研究最多的表观遗传标记。在动物中，DNA甲基化主要包括CpG岛的胞嘧啶(C)的C-5位置的甲基化，然后导致转录活性的抑制。

其原理在于，亚硫酸盐可使DNA上的胞嘧啶(C)转变成为尿嘧啶(U)，同时已受甲基化的5-甲基胞嘧啶则不受影响。如此一来，通过亚硫酸盐处理可以使实验 者得知DNA序列上的特定改变从而确定其甲基化的情形。

Hi-C测序（High-throughput chromosome conformation capture）



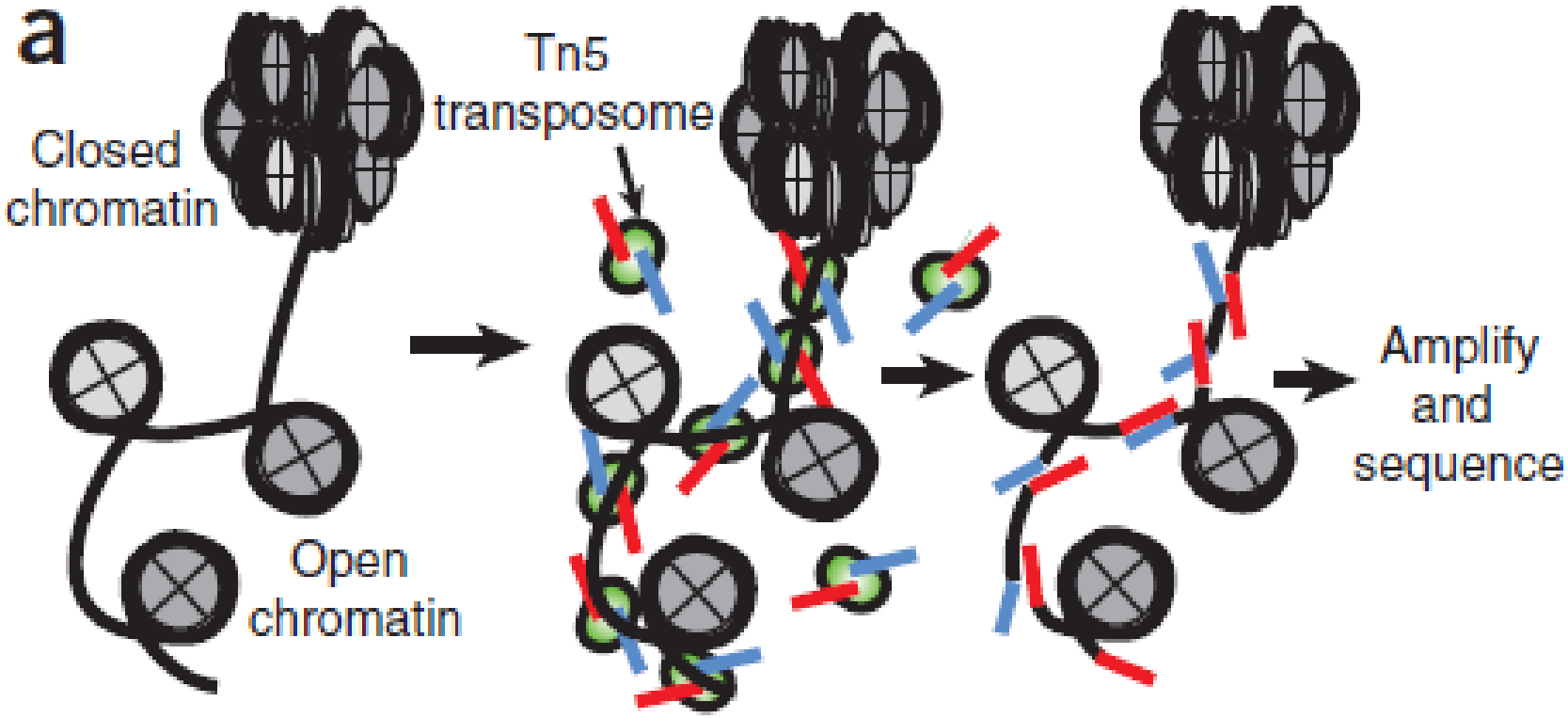
主要应用领域：

1、绘制细胞核内互作染色质的空间构象；

2、解析全基因组范围内基因间整体相互模式；

3、辅助基因组组装拼接

ATAC-seq（Assay for Transposase-Accessible Chromatin with high throughput sequencing）是使用高通量测序对Tn5转座酶可接近性核染色质区域进行测序分析的一种表观遗传学研究技术



ATAC-seq技术的主要应用

（1）获得在特定时空下基因组中所有处于开放状态的序列，分析调控元件

（2）分析染色质开放区域的motif，获得潜在的与其结合的转录因子等调控蛋白

RNA测序

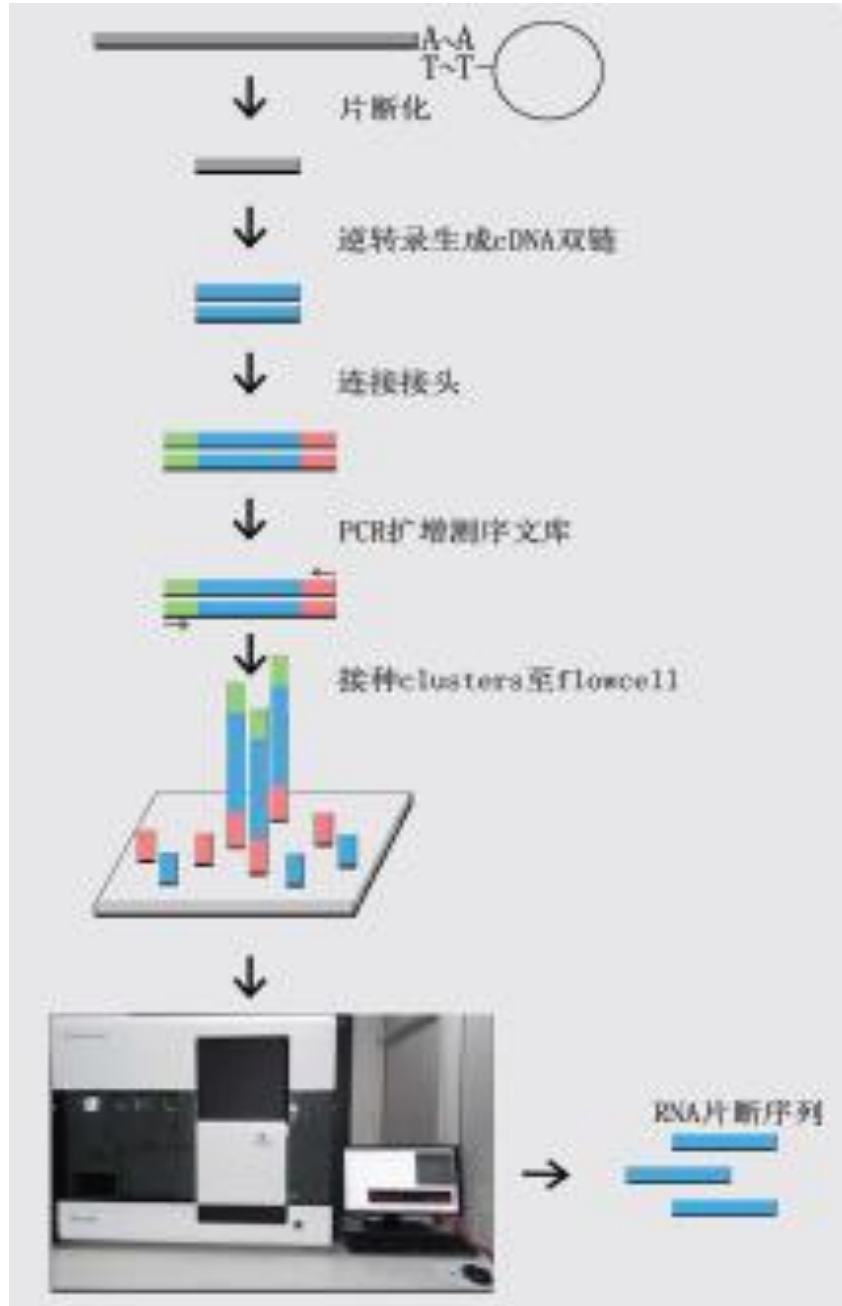
人类基因组草图带给科学家们的困惑

1. 含有30亿对碱基的人类基因组仅含有2－3万个蛋白质基因，是果蝇的两倍，啤酒酵母的4倍。显而易见，生物的复杂性不由编码蛋白质的数目决定。
2. 人类基因组的蛋白质编码区的总和占总基因组长度为1－2%，那么其他98%的基因组有什么功能呢？(1) 24%的基因组是插入编码序列的内含子序列；人类基因平均每个基因有7个内含子。但这么冗长的内含子序列有什么生物学功能呢？ (2) 其他74%的基因组的功能是什么？【注：90%以上的基因组都是转录的！】

转录组测序(RNA-seq)

广义转录组（Transcriptome）系指从一种细胞或者组织的基因组所转录出来的RNA的总和，包括编码蛋白质的mRNA和各种非编码RNA（rRNA, tRNA, snoRNA, snRNA,microRNA 和其他非编码RNA等）。

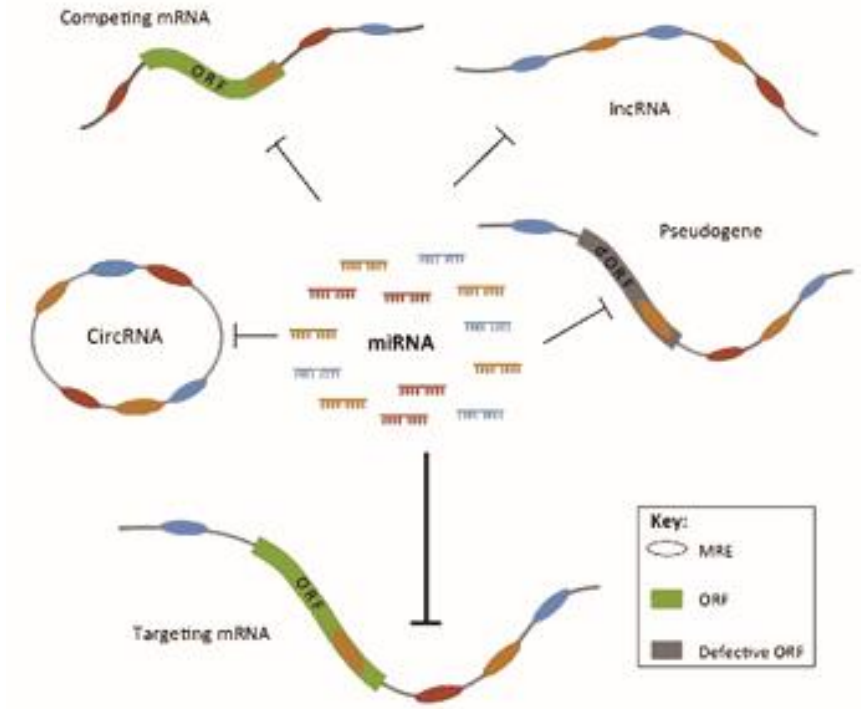
狭义转录组系指所有参与翻译蛋白质的mRNA 总和。

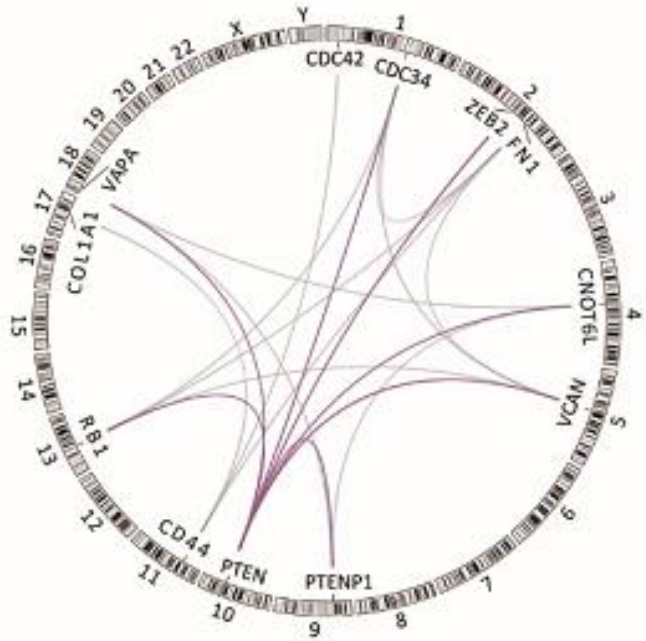

A~A
T~T
片断化
逆转录生成cDNA双链
连接接头
PCR扩增测序文库
装种clusters至flowcell
RNA片断序列

全转录组测序：

是指特定细胞在特定状态下所能转录出来的所有RNA的总和，包括mRNA和非编码RNA（non-coding RNA）。针对非编码RNA的研究主要集中在具有调控作用的miRNA，lncRNA和circRNA。基于二代测序技术的全转录组测序研究，同时分析同一样本中的mRNA，lncRNA，circRNA，miRNA，并且通过两两关联分析、三元关联分析、多元关联分析，使研究内容更加系统化，致力于深入挖掘生命现象背后的转录调控问题。

长链非编码RNA（lncRNA）是一类长度超过200 nt的RNA分子，位于细胞核或胞质内，lncRNA一直以来都是科研领域的热点，吸引着众多研究者的目光。长链非编码RNA按照其来源可分为Antisense lncRNA (反义长非编码RNA)，Intronic transcript (内含子非编码RNA)，Long intergenc noncoding RNA(lincRNA)，Promoter-associated lncRNA(启动子相关lncRNA)，UTR associated lncRNA(非翻译区lncRNA)五种类型。



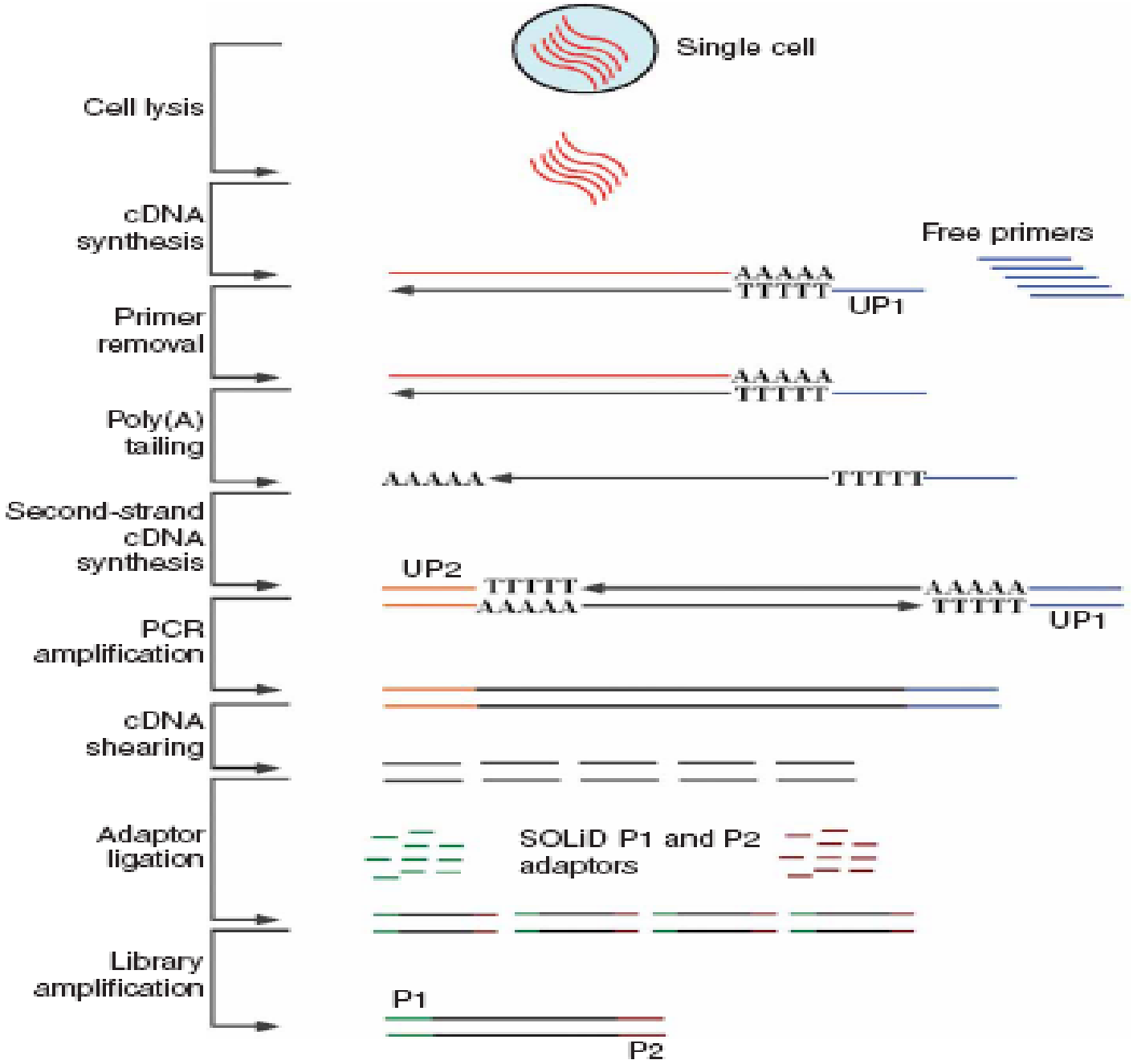


单细胞转录组测序

随着高通量测序技术的不断成熟，转录组高通量测序技术在功能基因组学的研究中发挥了非常重要的作用。

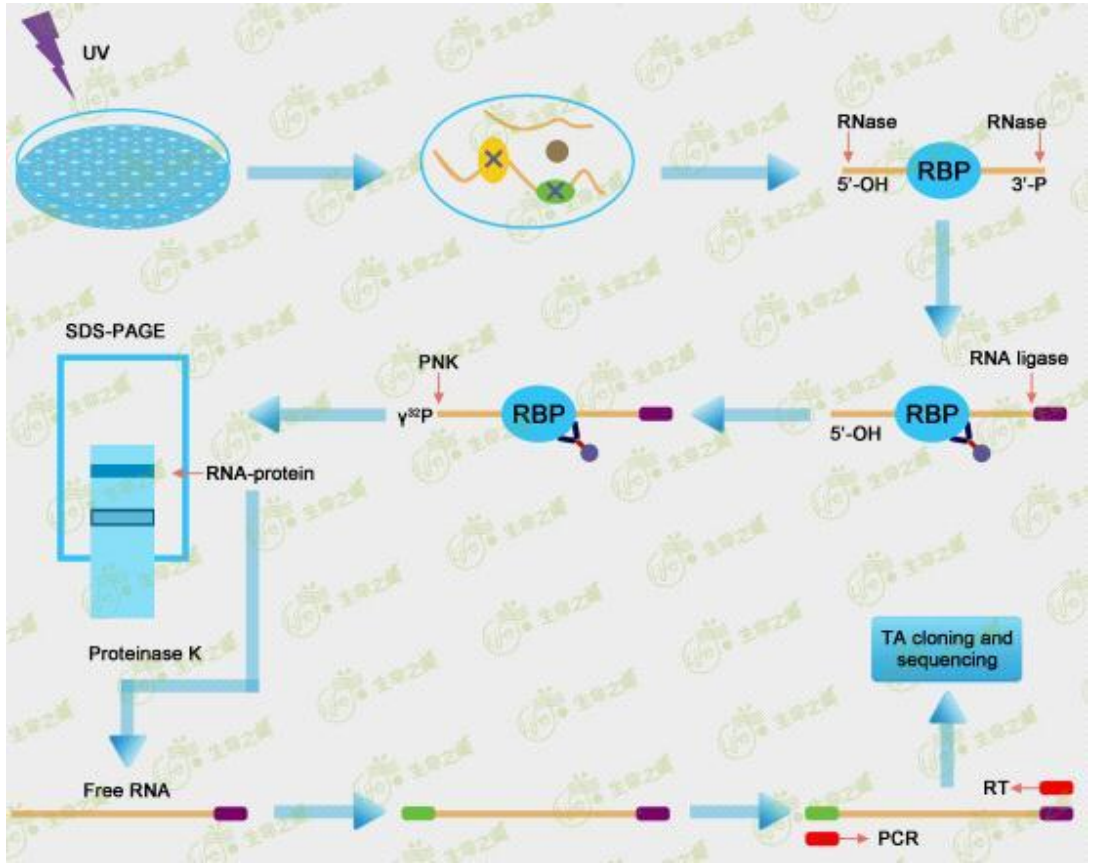
很多研究表明相同的细胞类型中的基因表达水平也常常显示异质性，对于单个细胞来说，转录组的随机变异性可能决定了细胞的最终命运，

因此不同的细胞具有不同的转录组表型，所以从理论上讲，转录组分析应该以单细胞为研究模型，科学家应用单细胞RNA-seq技术即可获得单细胞的转录组信息。

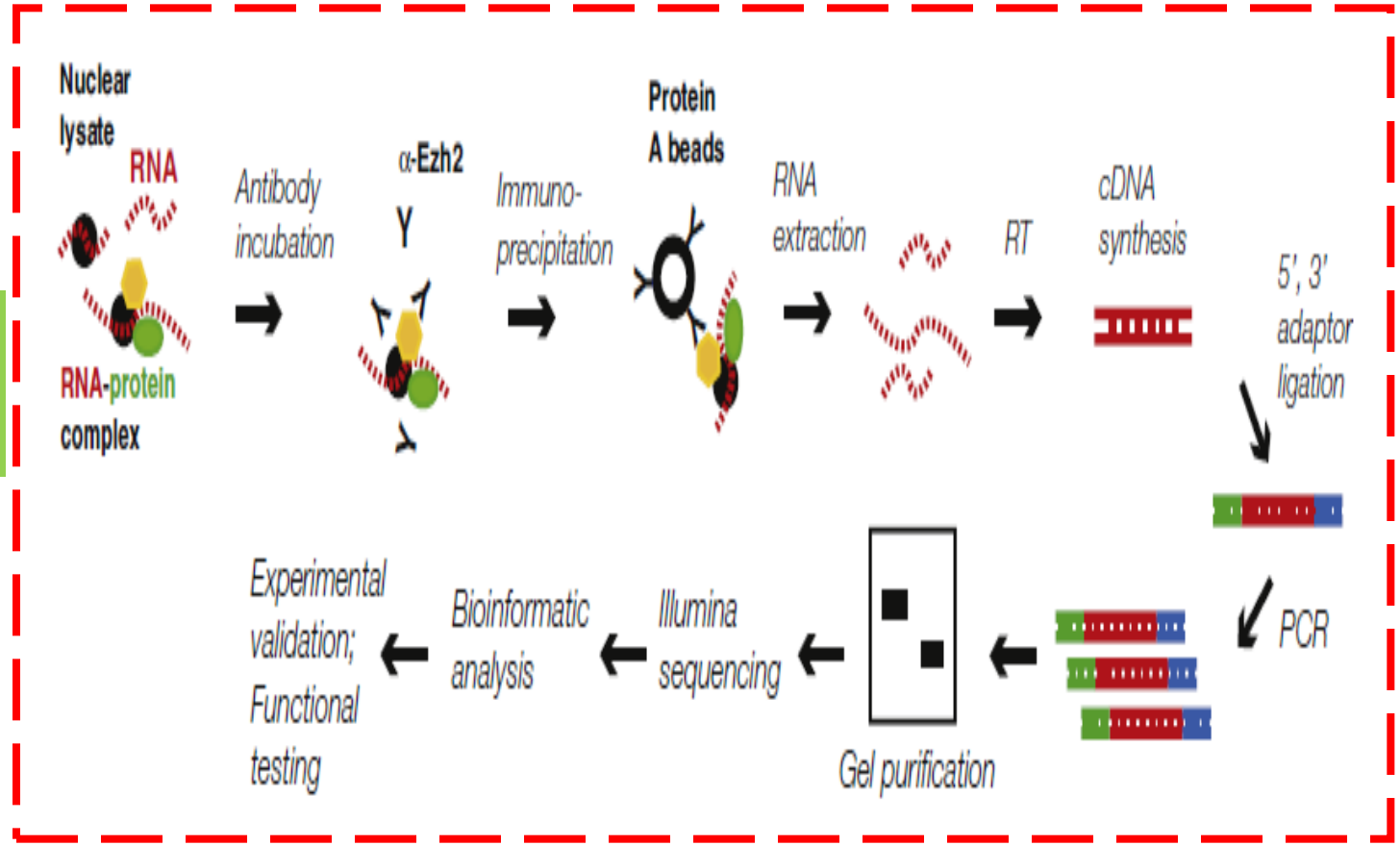

Single cell
Cell lysis
cDNA synthesis
Free primers
AAAAA
TTTTT
UP1
Primer removal
Poly(A) tailing
Second-strand cDNA synthesis
UP2
PCR amplification
cDNA shearing
Adaptor ligation
SOLiD P1 and P2 adaptors
Library amplification
P1
P2

CLIP-Seq

- CLIP-seq (crosslinking-immunprecipitation and high-throughput sequencing)即紫外交联免疫沉淀结合高通量测序，可以高通量研究RNA结合蛋白在体内与众多RNA靶标的结合模式



RIP-seq


Nuclear lysate
RNA
RNA-protein complex
Antibody incubation
α-Ezh2
Immuno-precipitation
Protein A beads
RNA extraction
RT
cDNA synthesis
5', 3' adaptor ligation
PCR
Gel purification
Illumina sequencing
Bioinformatic analysis
Experimental validation; Functional testing

小RNA测序

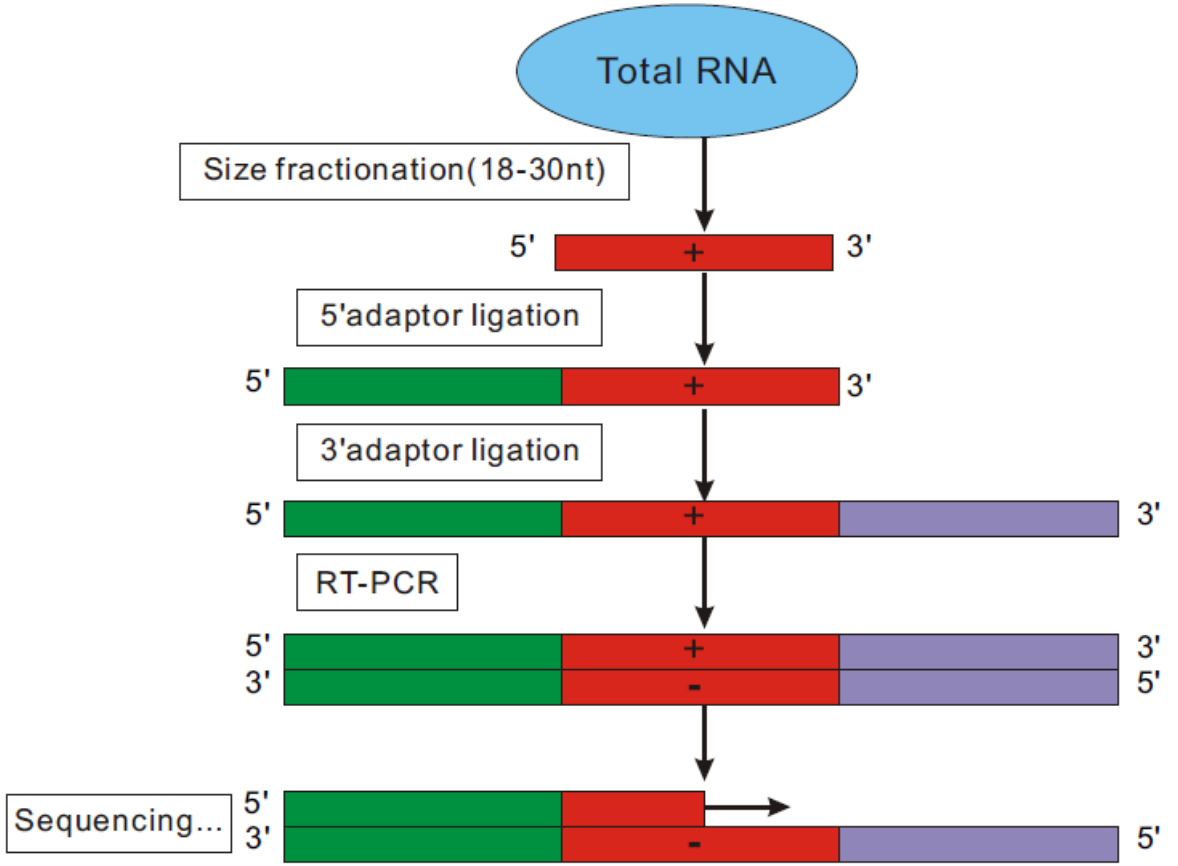
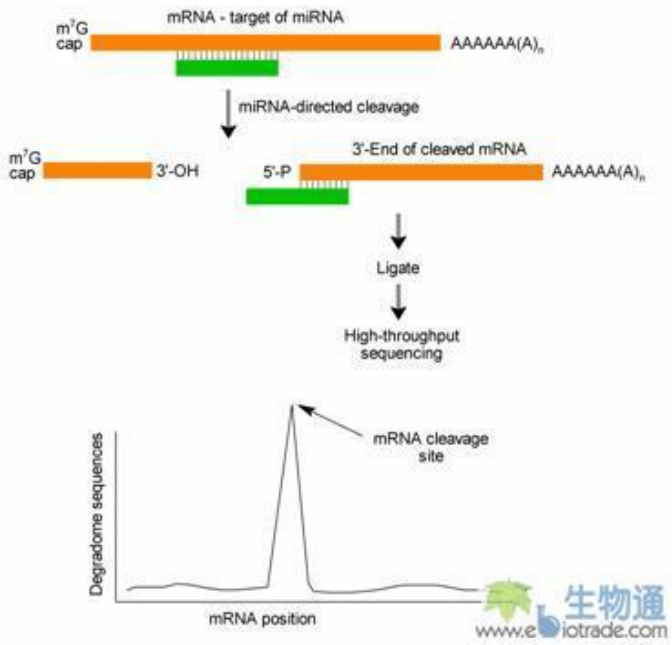


Figure 5-1 Experimental pipeline ofsmall RNA analysis

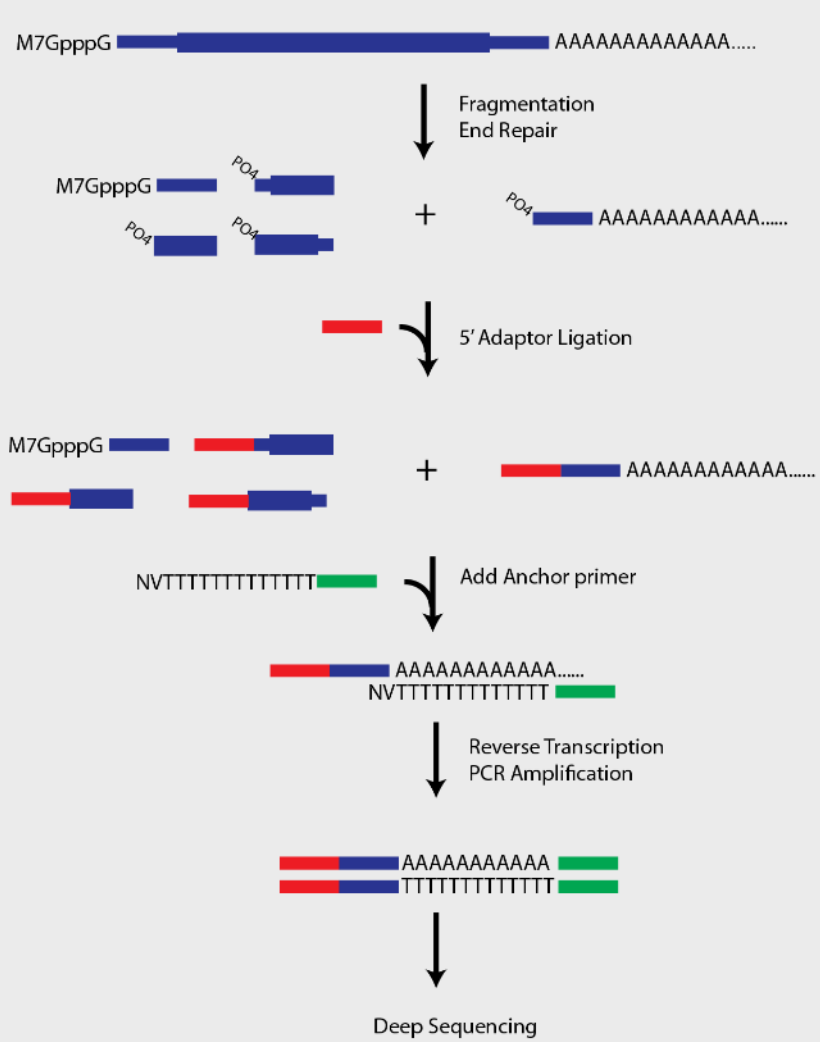
植物降解组 (Degradome) 测序

- 高通量研究植物miRNA靶基因

- 在植物体内绝大多数的miRNA剪切常发生在miRNA与mRNA互补区域的第十位核苷酸上。靶基因经剪切产生二个片段，5’ 剪切片段和3’ 剪切片段。
- 其中3’ 剪切片段，包含有自由的5’ 单磷酸和3’ polyA尾巴，可被RNA连接酶连接，连接产物可用于下游高通量测序;
- 而含有5’ 帽子结构的完整基因，含有帽子结构的5’ 剪切片段或是其他缺少5’ 单磷酸基团的RNA是无法被RNA酶连接，因而无法进入下游的测序实验

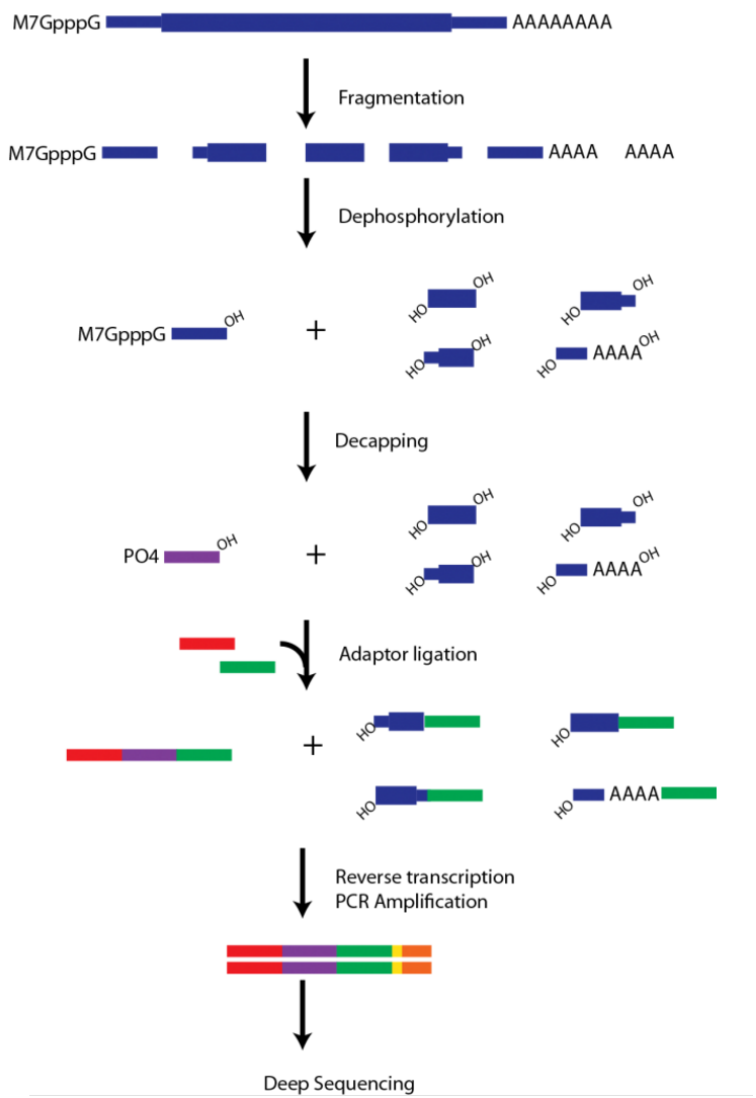


多聚腺苷化位点测序(PAS-seq)



绝大多数真核生物mRNA和lncRNA的3’末端都具有一串连续的腺苷，称为poly(A)尾巴。PAS的选择直接影响到mRNA的3’UTR的长度， 由于3’UTR是mRNA最重要的调控区，是绝大多数RNA结合蛋白和microRNA调控的靶标区域，因此PAS的选择，会直接影响到mRNA的稳定性和翻译效率。

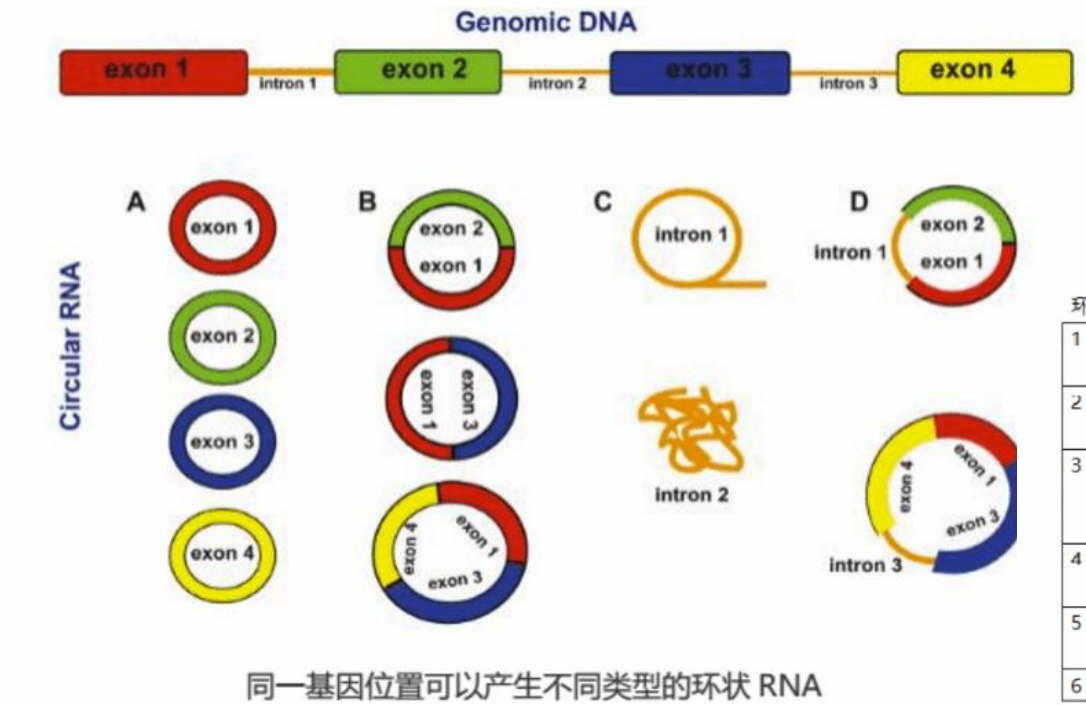
加帽端测序(CAGE-seq)



CAGE-seq (Cap Analysis of Gene Expression AND deep Sequencing)是结合mRNA加帽位点鉴定和高通量测序的手段，可以高通量的鉴定整个细胞内mRNA的转录起始位点，从而获得准确的启动子信息和5’UTR信息。

CircRNA测序：

circRNA（circular RNA，环状RNA）是一类具有闭合环状结构的非编码RNA分子，没有5′帽子结构和3′poly（A）结构（如图1），主要位于细胞质或储存于外泌体中，不受RNA外切酶影响，表达更稳定且不易降解，已被证明广泛存在于多种真核生物体内。大多数circRNA是由外显子环化而成，也有部分circRNA是由内含子环化而成的套索结构（lariat）。同时由于circRNA含有大量的miRNA应答原件（MREs），能与AGO蛋白形成RNA诱导沉默复合体（RISC）的催化核心，最终导致circRNA降解。根据来源，circRNA可大致分为四类：全外显子型的circRNA，内含子和外显子组合的EIcircRNA，内含子组成的套索型ciRNA，由病毒RNA基因组、tRNA、rRNA、snRNA等环化产生的circRNA。

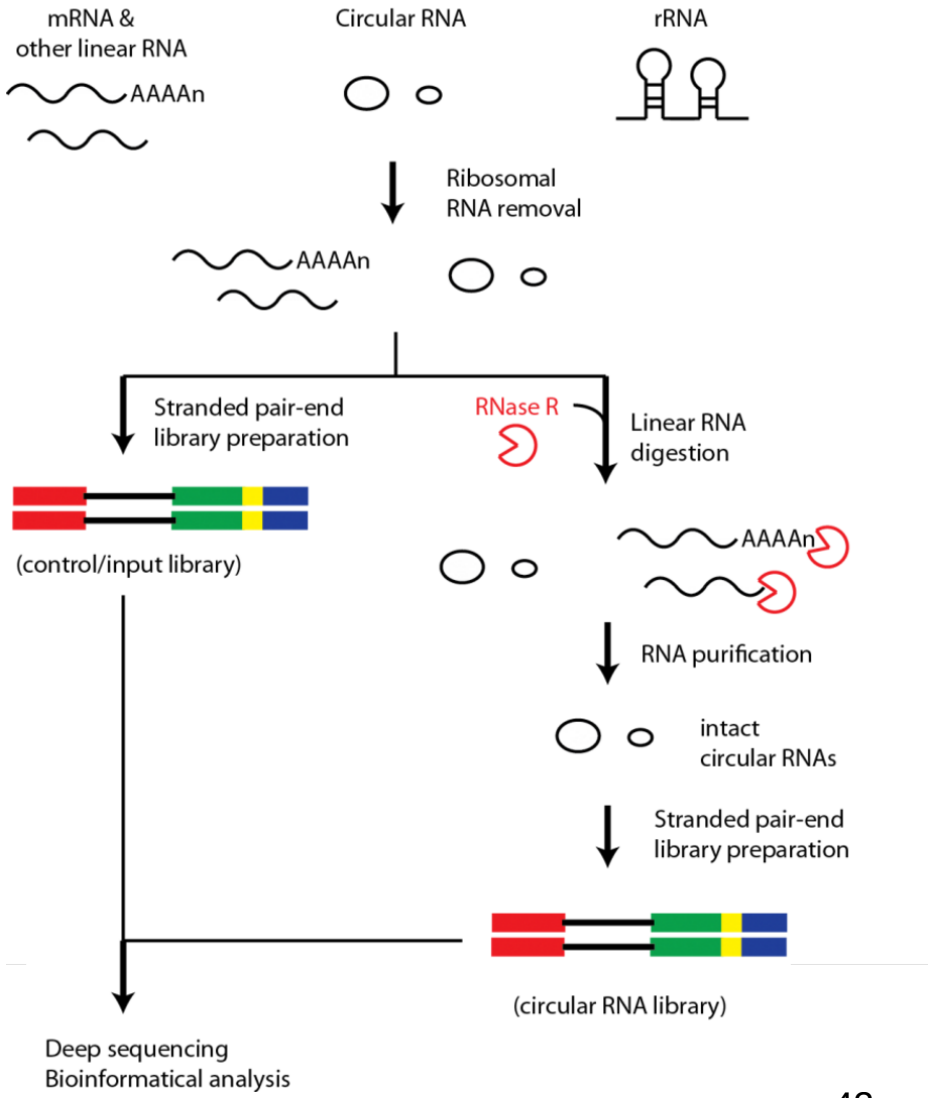


同一基因位置可以产生不同类型的环状 RNA

环状 RNA 的主要类型

1	annot_exon ：环状 RNA 的 breakpoint 位于已知基因的一个外显子的起始和另一个外显子的终止，其序列由 breakpoint 之间的外显子的碱基构成
2	one_exon ：环状 RNA 位于已知基因的某一个外显子内，其序列由 breakpoint 之间的所有碱基构成
3	exon_intron ：环状 RNA 的 breakpoint 之间有已知基因的一个或多个完整外显子但 breakpoint 不位于已知外显了的起始或终止，其序列由上 breakpoint 起始位点到第一个外显子起始位点、breakpoint 之间的外显子和最后一个外显子到 breakpoint 终止位点的碱基构成
4	intronic ：环状 RNA 位于已知基因的某一个内含子内，其序列由 breakpoint 之间的所有碱基构成
5	antisense ：环状 RNA 位于已知基因的反义链，其序列由反义链上 breakpoint 之间的所有碱基构成
6	intergenic ：环状 RNA 位于已知基因的之间，其序列由 breakpoint 之间的所有碱基构成

环状RNA测序(CircRNA-seq)



环形RNA是mRNA在剪接的过程中，上游exon的5’端与下游exon的3’端剪接到一起，从而形成的首尾相接的环状RNA分子。最近的研究发现circular RNA可以作为“microRNA海绵”，竞争结合microRNA从而解除这些microRNA对其他靶标的调控；同时保守型分析发现环状RNA上潜在具有多种RNA结合蛋白的结合位点，暗示circular RNA可以调控RNA结合蛋白的功能。