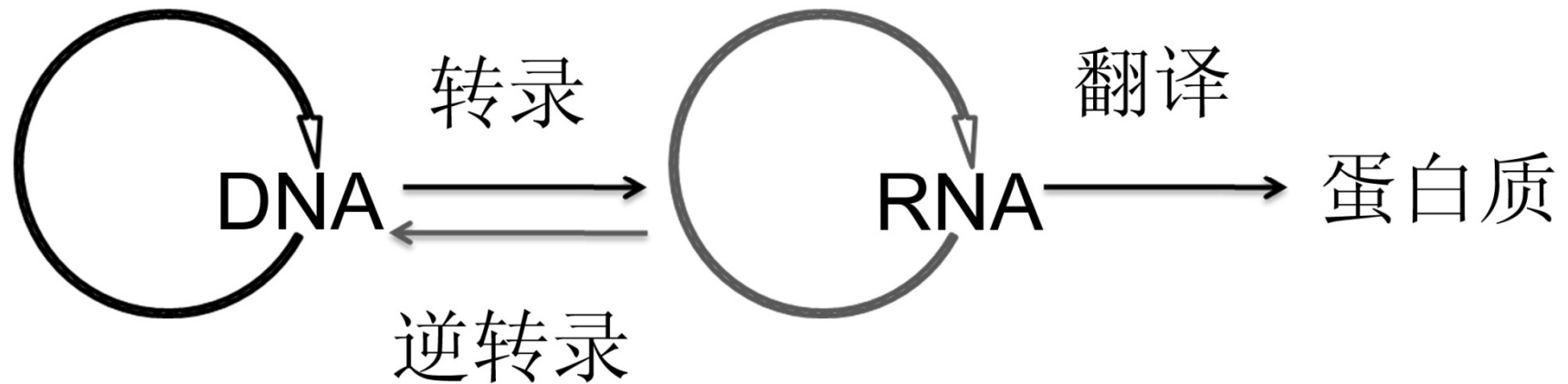




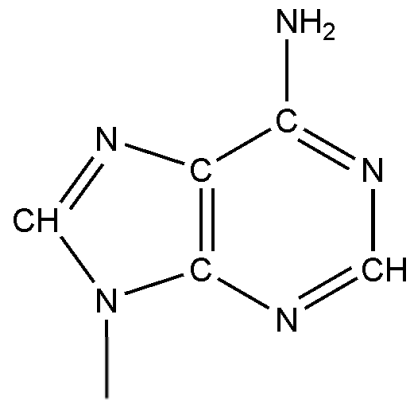
生物信息学

第二章 生物序列获取和存储

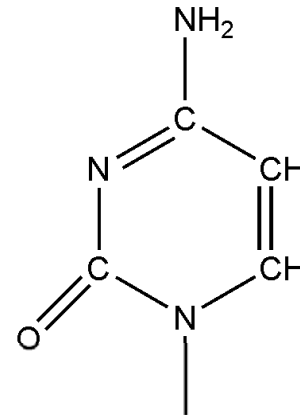
中心法则



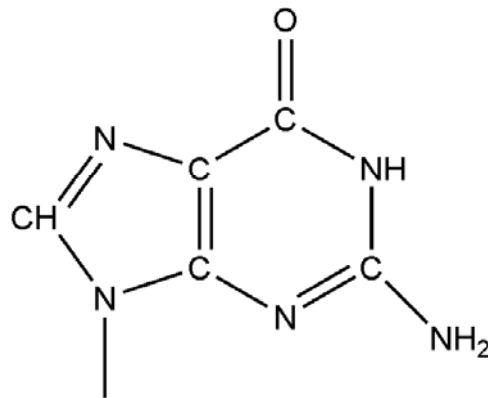
DNA结构：碱基/核苷



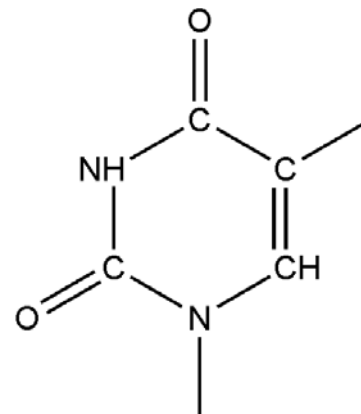
Adenine (A)



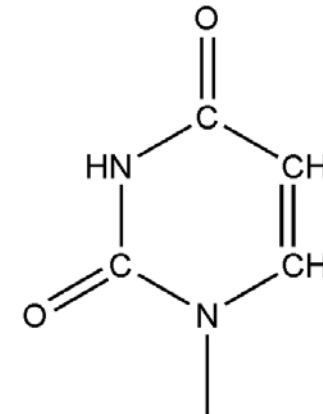
Cytosine (C)



Guanine (G)



Thymine (T)

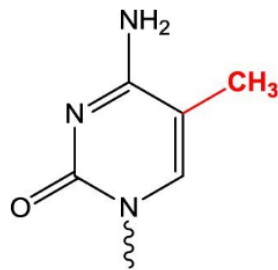


Uracil (U)

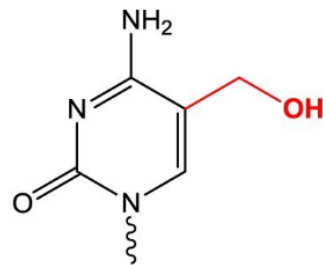
DNA修饰 (~10种)



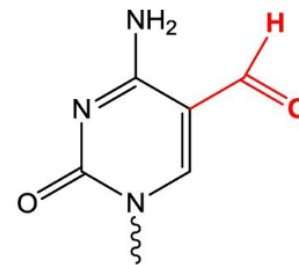
- 5-methylcytosine (5-mC)
- 5-hydroxymethylcytosine (5-hmC)
- 5-formylcytosine (5-fC)
- 5-carboxylcytosine (5-caC)
- 4-methylcytosine(4-mC)
- 6-methyladenine (6-mA)
- 8-oxoguanine (8-oxoG)
- 8-oxoadenine (8-oxoA)



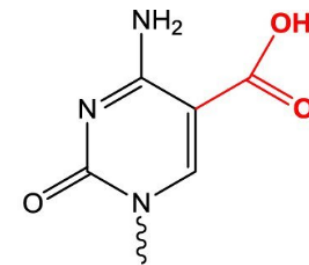
5-mC



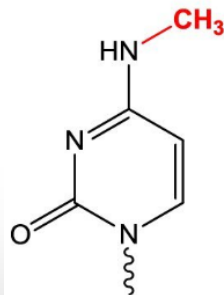
5-hmC



5-fC



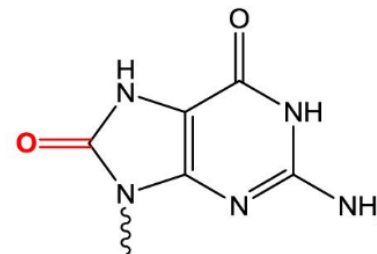
5-caC



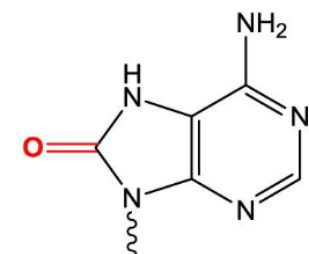
4-mC



6-mA

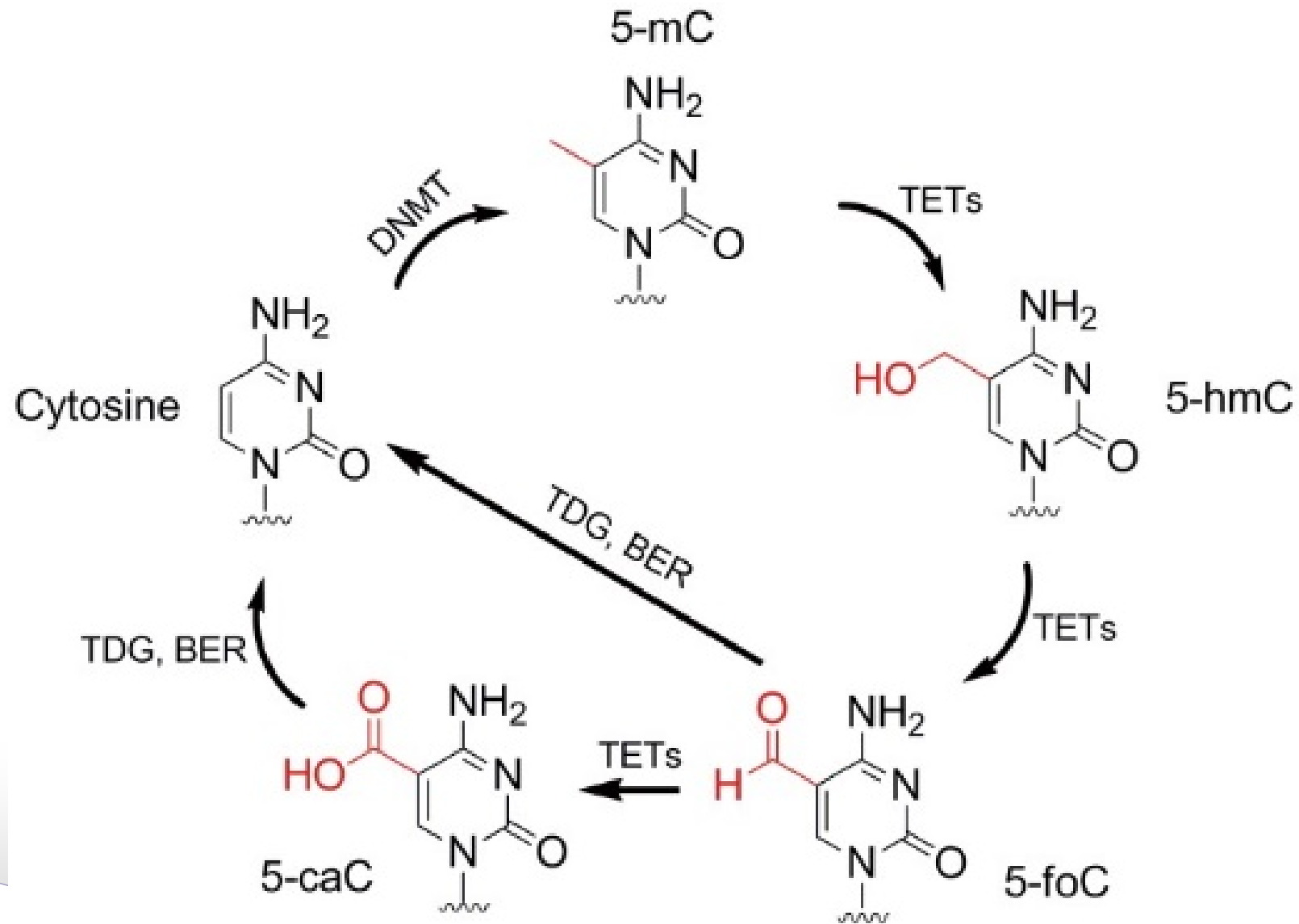


8-oxoG

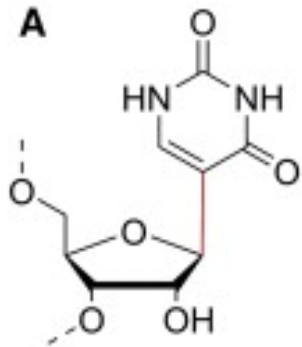


8-oxoA

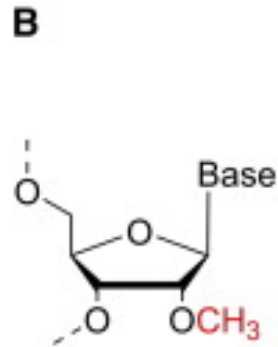
DNA修饰的催化过程



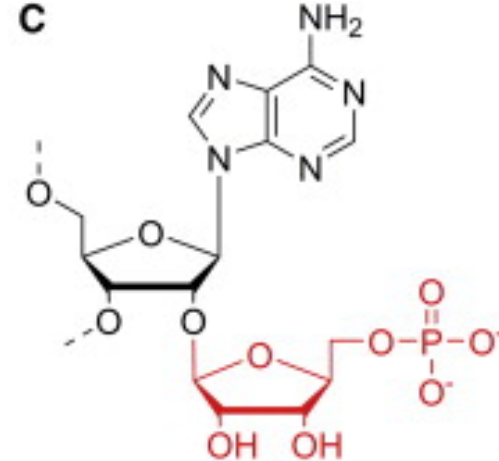
RNA修饰 (~150种)



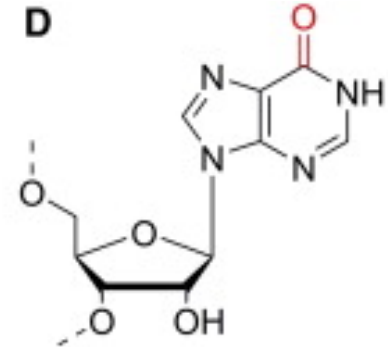
Pseudouridine



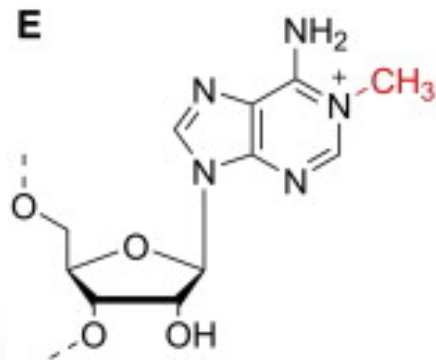
2'-O-methylation



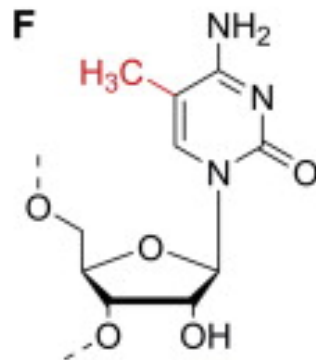
2'-O-ribosylation of adenosine



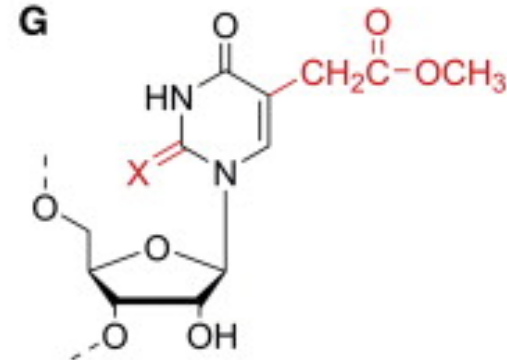
Inosine



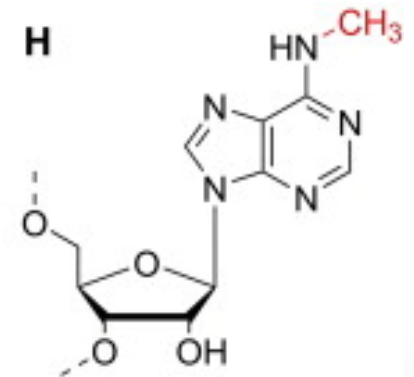
m¹A



m⁵C

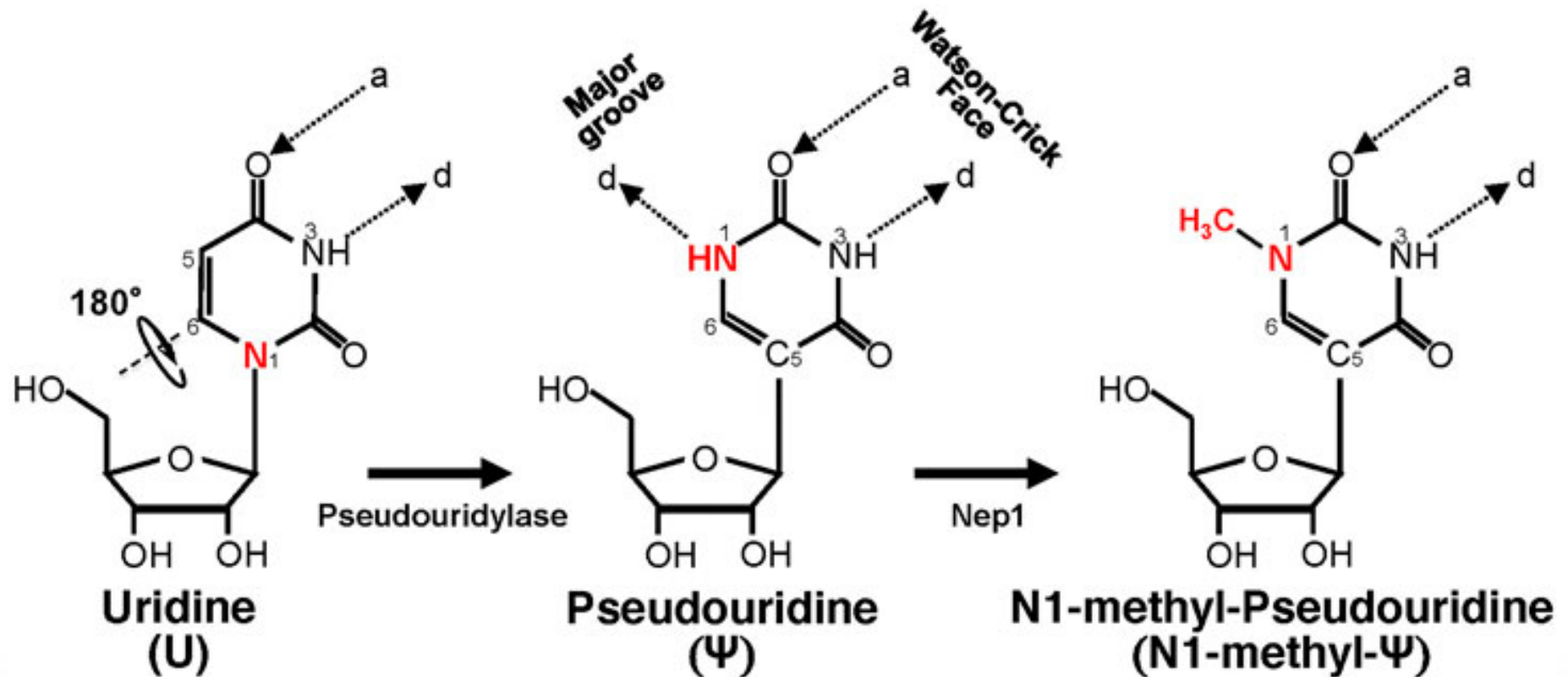


mcm⁵U, X = O
mcm⁵S²U, X = S



m⁶A

mRNA疫苗中的关键修饰

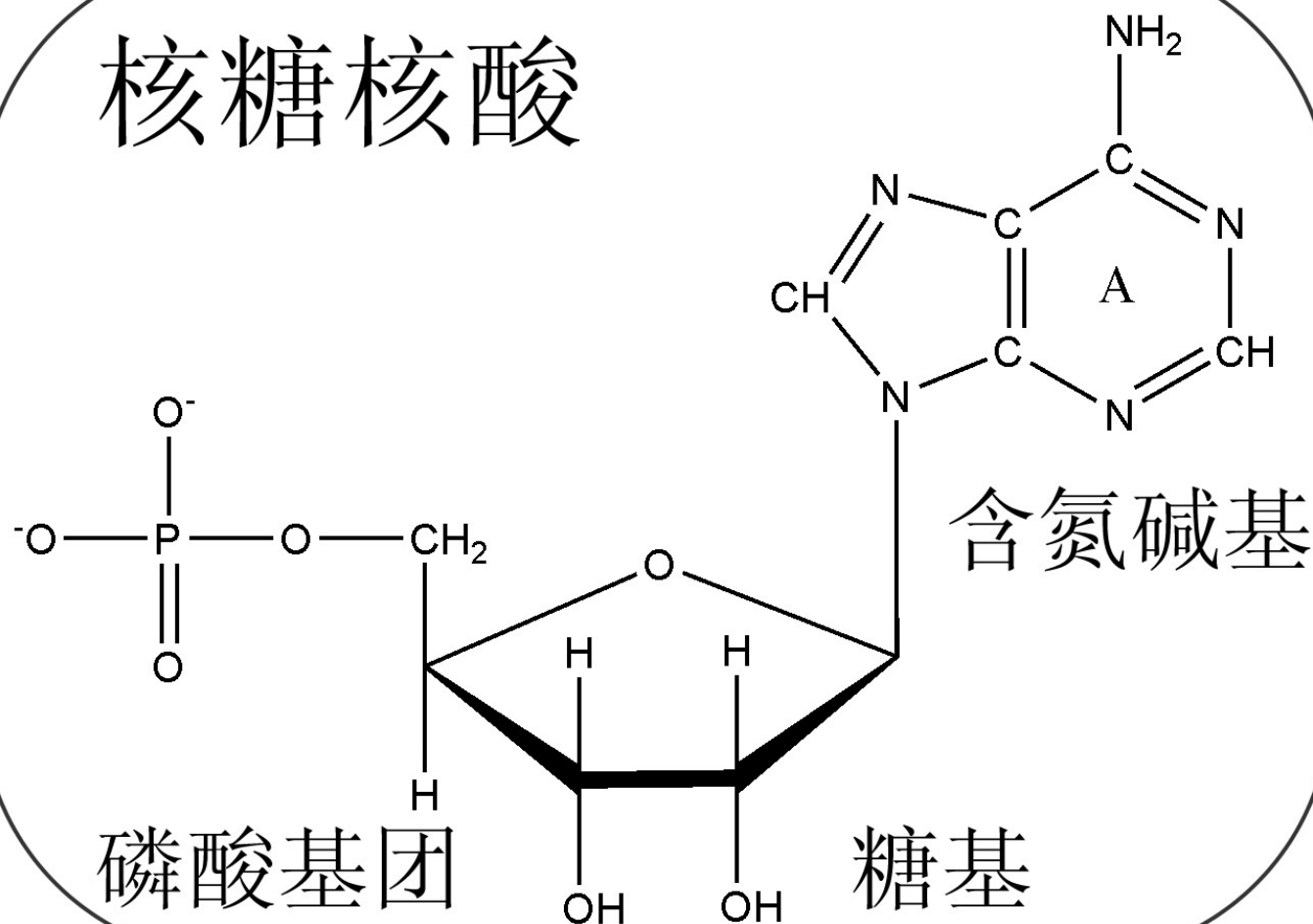


核糖核酸



Ribonucleotide

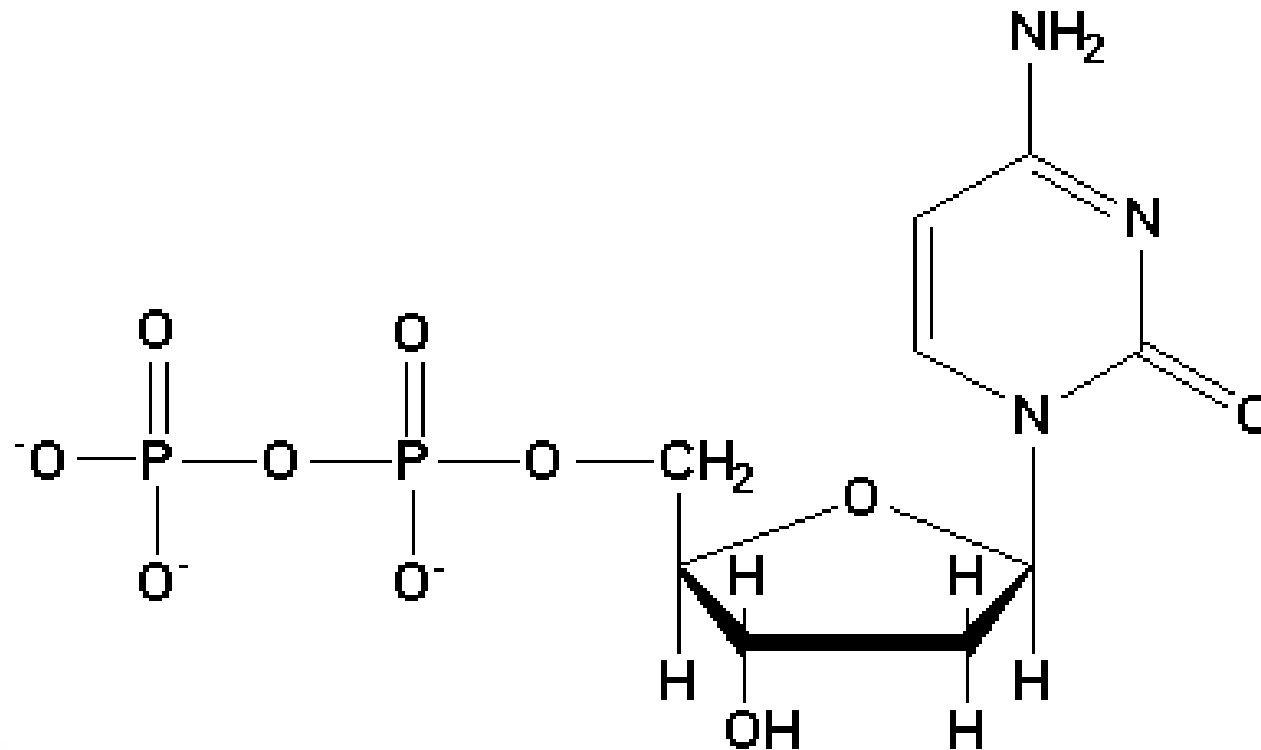
核糖核酸



脱氧核糖核酸



Deoxyribonucleotide



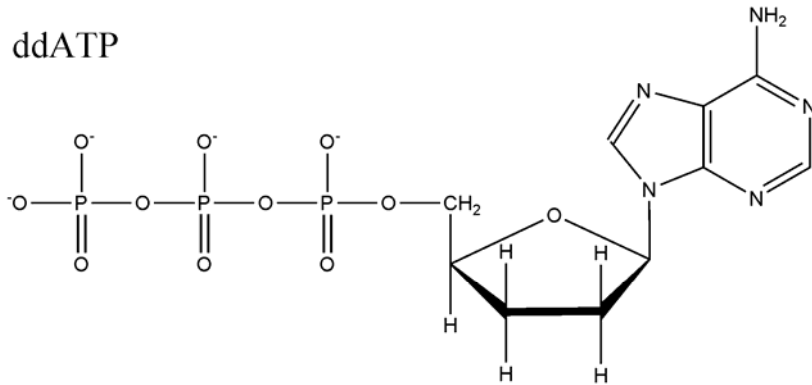
Deoxycytidine diphosphate (dCDP)

双脱氧核糖核苷酸

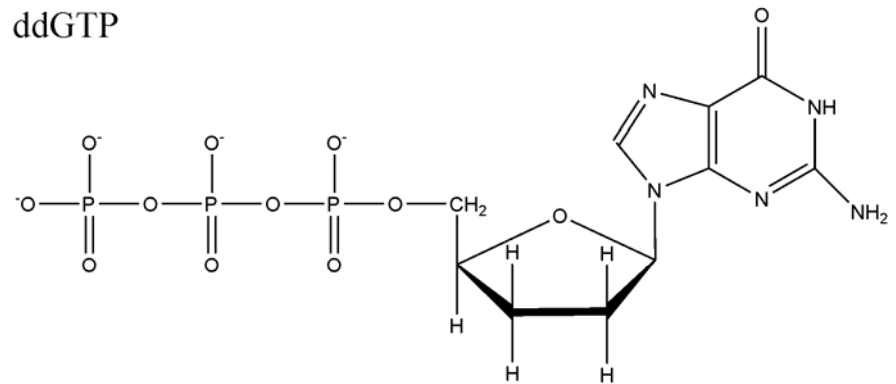


Dideoxynucleotide

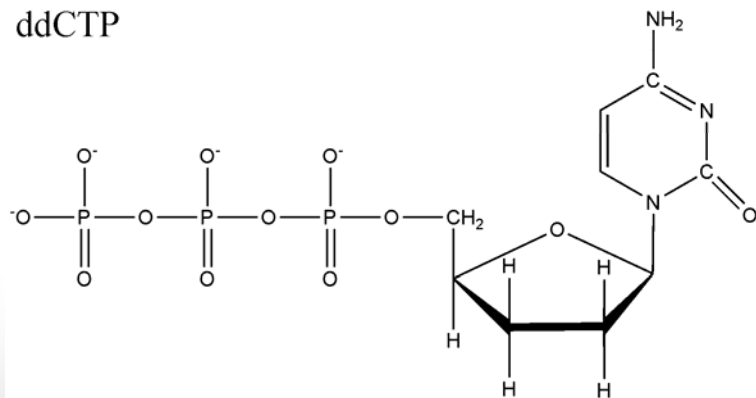
ddATP



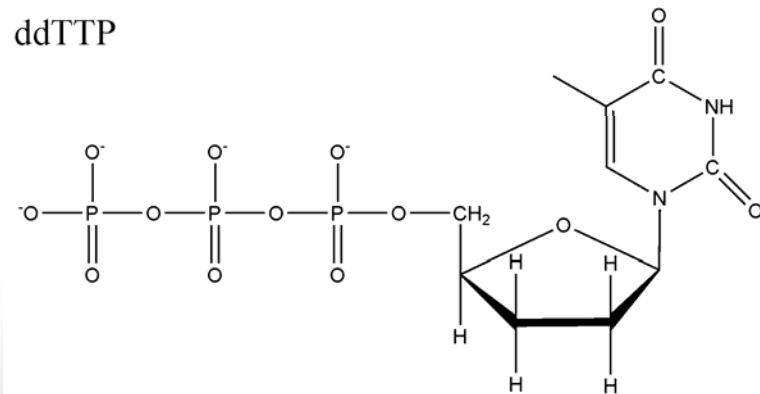
ddGTP



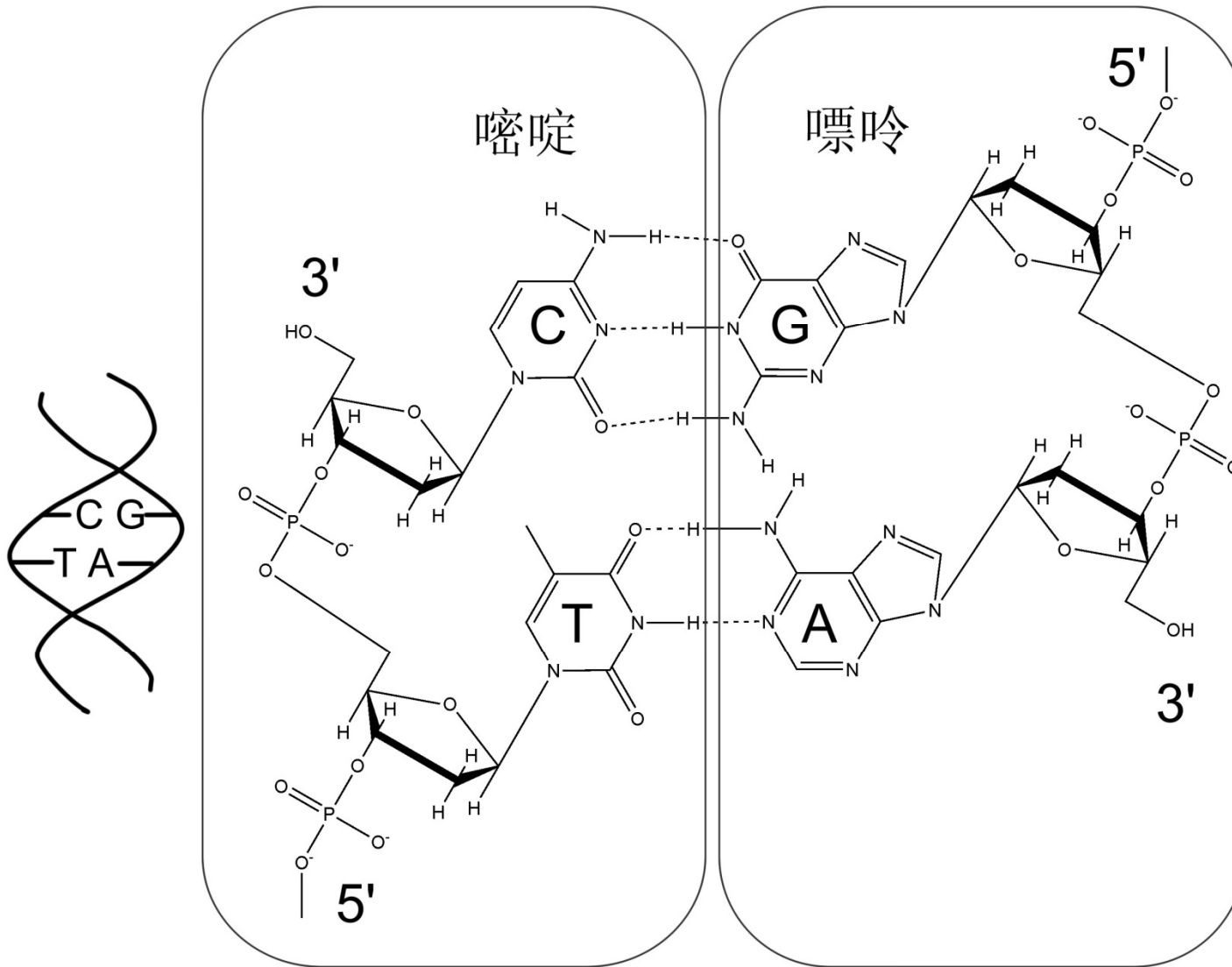
ddCTP



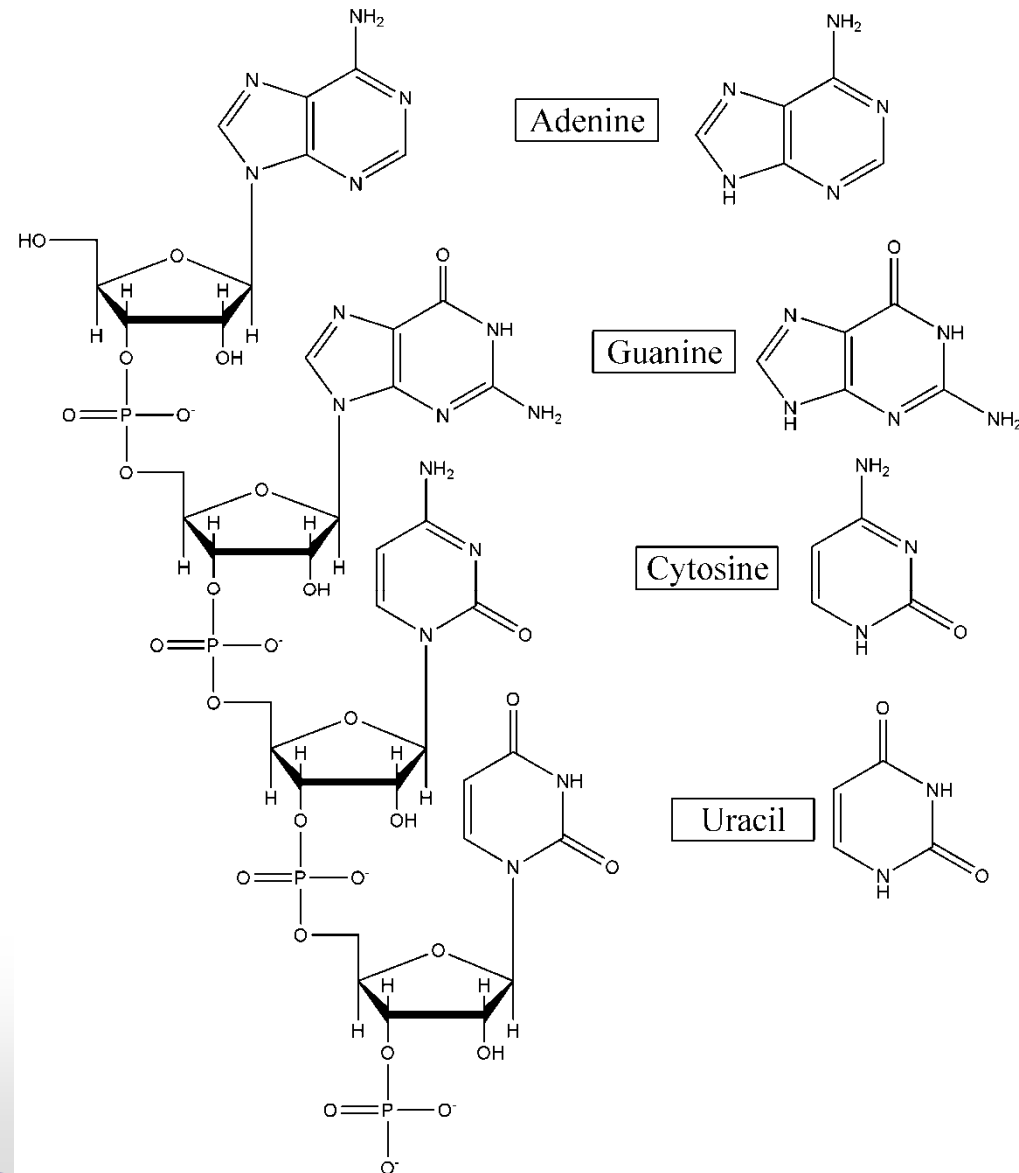
ddTTP



DNA的结构



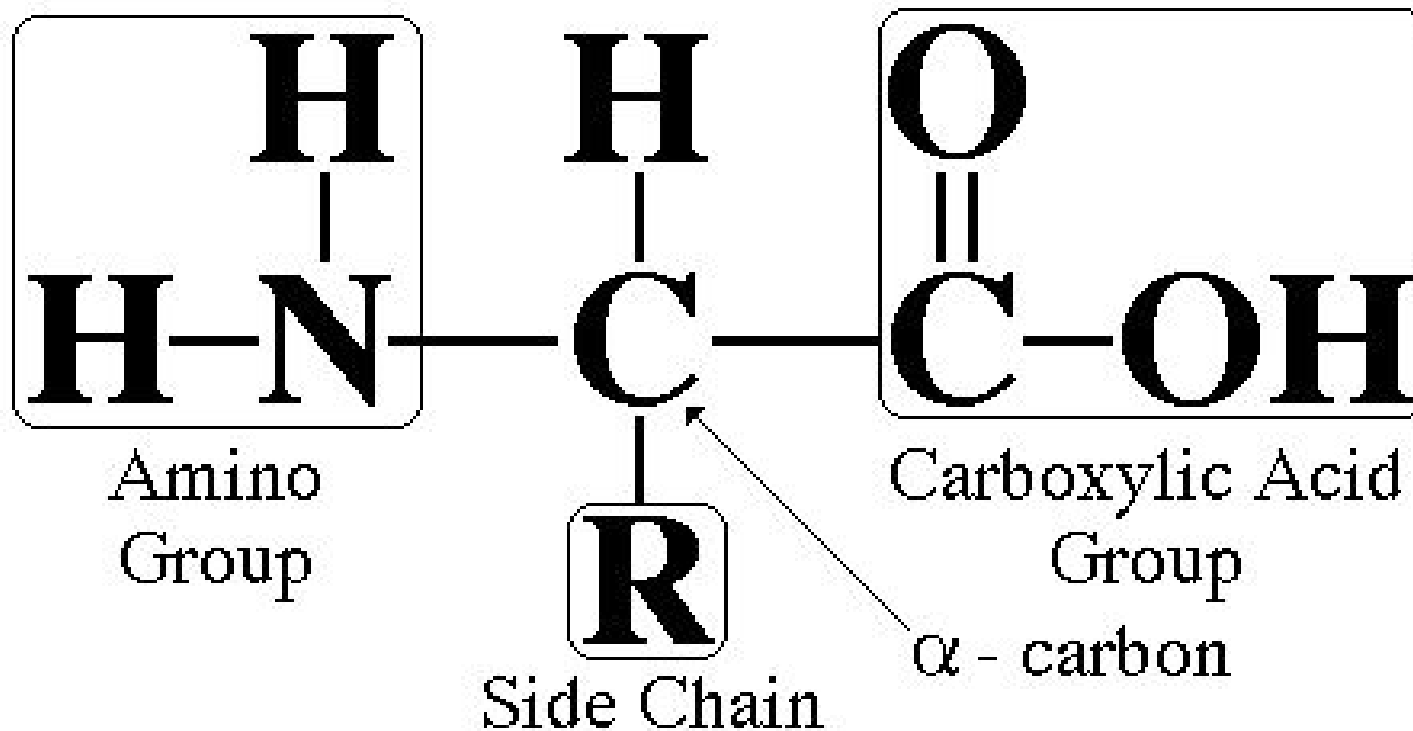
RNA的结构



氨基酸的结构



Amino Acid Structure

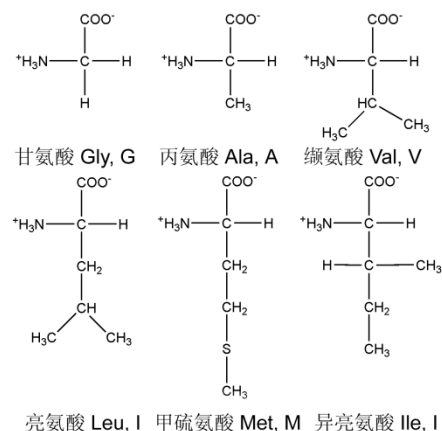


氨基酸的性质及分类

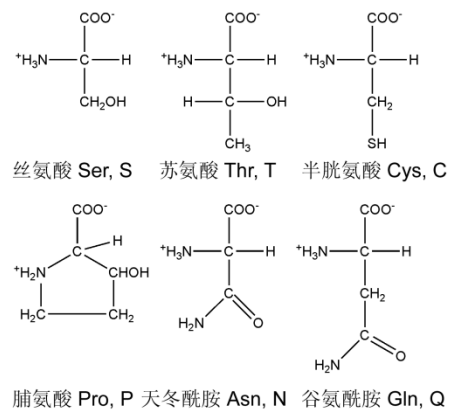


20种常见的蛋白质氨基酸

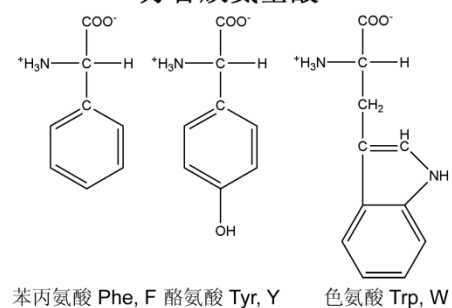
非极性脂肪酸氨基酸



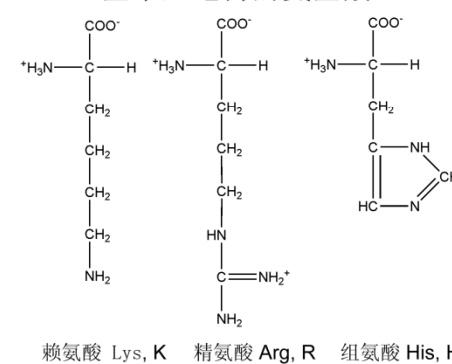
R基不带电荷的氨基酸



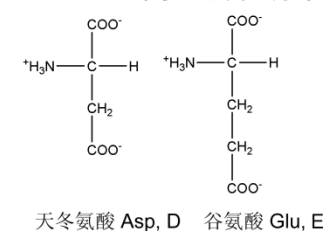
芳香族氨基酸



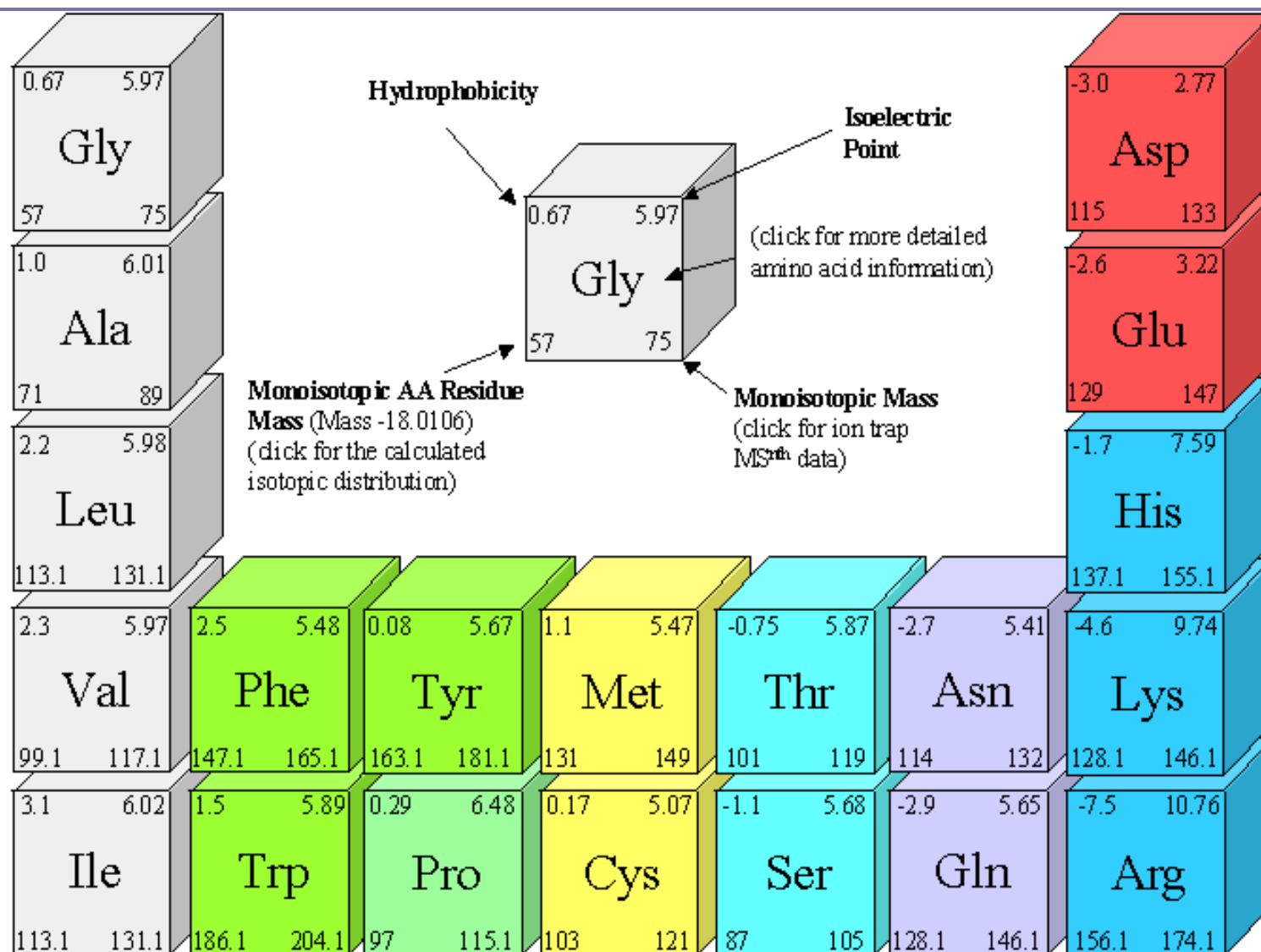
R基带正电荷的氨基酸



R基带负电荷的氨基酸



氨基酸：周期表



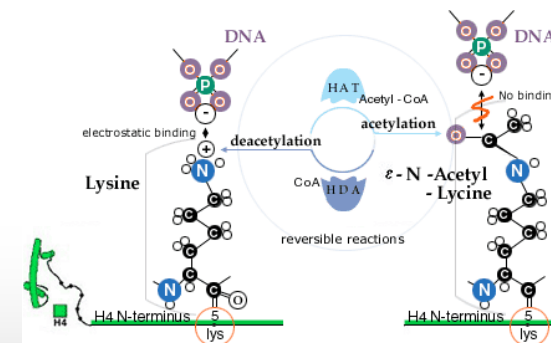
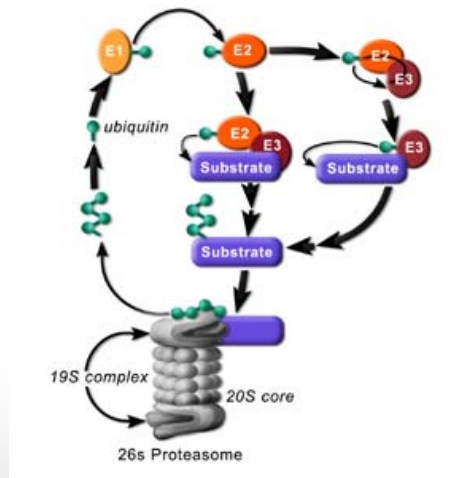
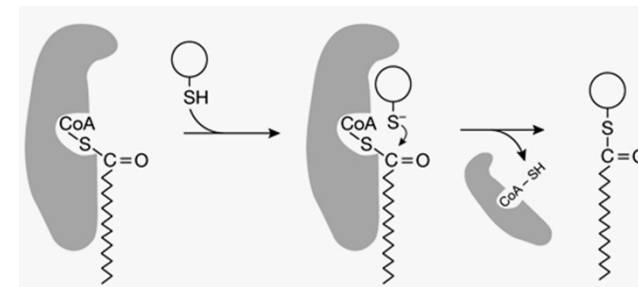
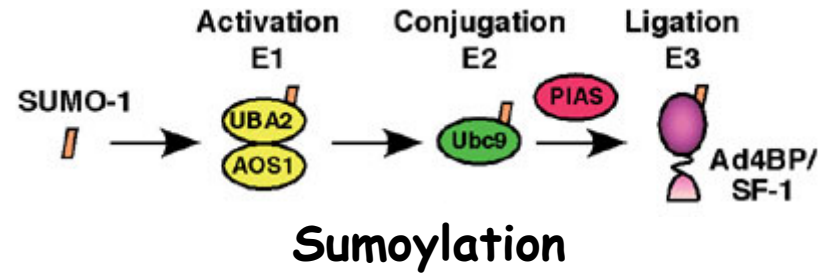
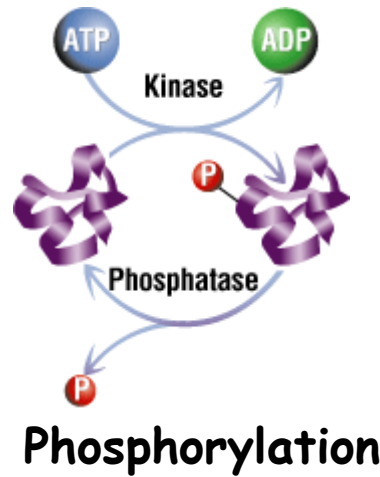
蛋白质翻译后修饰



- ❑ **POST**: DNA转录为RNA, RNA翻译成蛋白质之后发生
- ❑ 生化反应: 产生或破坏共价键
- ❑ 发生在氨基酸的主链或侧链上
- ❑ 通常由特定酶催化
- ❑ 可发生在15种非疏水的氨基酸上 (not F/V/I/L/A)
- ❑ 目前已发现~680种PTMs

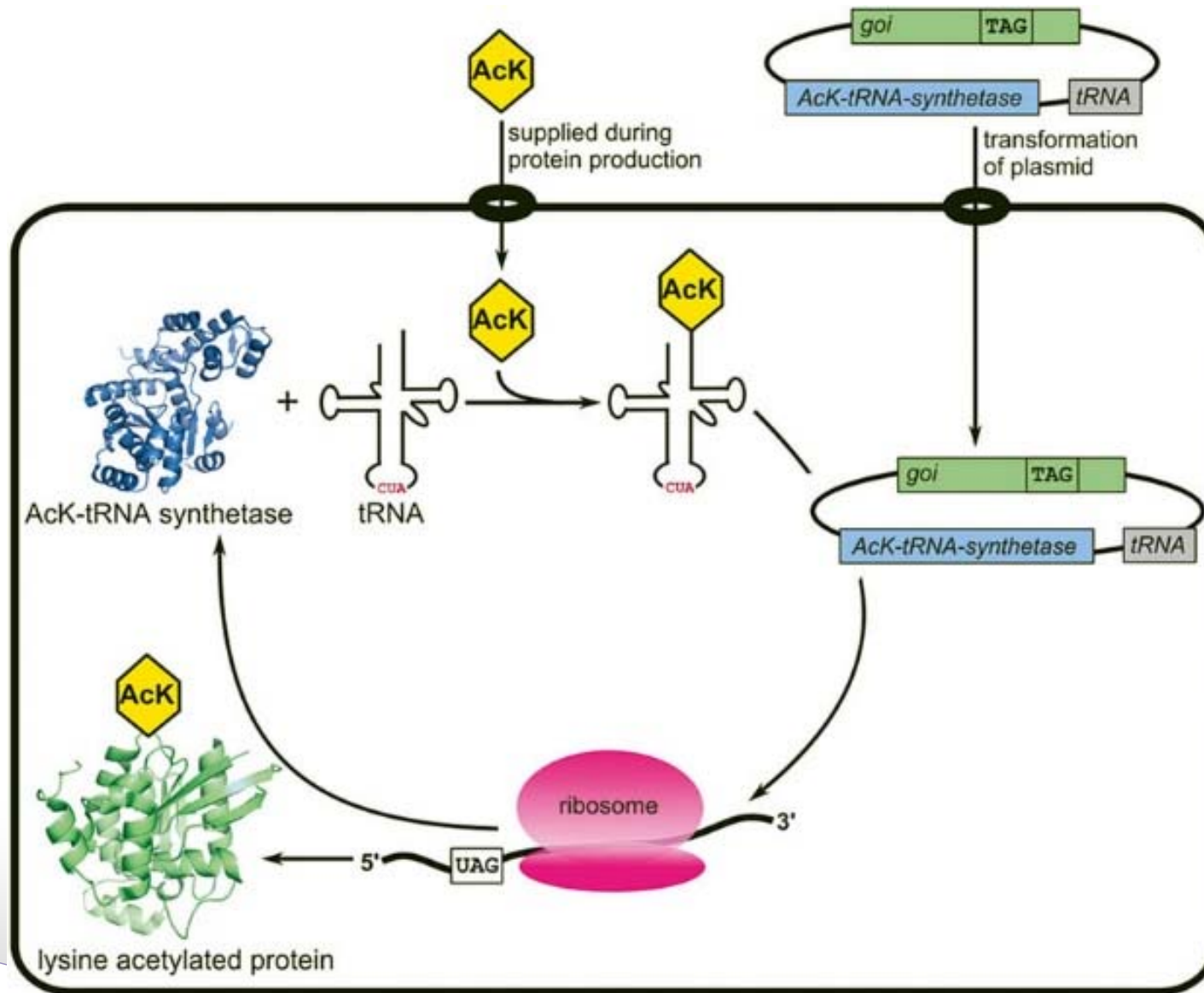
<http://www.uniprot.org/docs/ptmlist>

常见的蛋白质翻译后修饰



	T	C	A	G
T	TTT Phe (F)	TCT Ser (S)	TAT Tyr (Y)	TGT Cys (C)
	TTC ..	TCC ..	TAC ..	TGC ..
	TTA Leu (L)	TCA ..	TAA Ter ochre	TGA Ter opal
	TTG ..	TCG ..	TAG Ter amber	TGG Trp (W)
C	CTT Leu (L)	CCT Phe (P)	CAT His (H)	CGT Arg (R)
	CTC ..	CCC ..	CAC ..	CGC ..
	CTA ..	CCA ..	CAA Gln (Q)	CGA ..
	CTG ..	CCG ..	CAA ..	CGG ..
A	ATT Ile (I)	ACT Thr (T)	AAT Asn (N)	AGT Ser (S)
	ATC ..	ACC ..	AAC ..	AGC ..
	ATA ..	ACA ..	AAA Lys (K)	AGA Arg (R)
	ATG Met (M)	ACG ..	AAA ..	AGG ..
G	GTT Val (V)	GCT Ala (A)	GAT Asp (D)	GGT Gly (G)
	GTC ..	GCC ..	GAC ..	GGC ..
	GTA ..	GCA ..	GAA Glu (E)	GGA ..
	GTG ..	GCG ..	GAA ..	GGG ..

遗传密码子扩展



吴瑞先生：DNA测序



- ❑ Dr. Ray Wu, 1928年出生于北京, 2008年去世
- ❑ 康奈尔大学分子生物学与遗传学教授、植物基因工程创建人之一
- ❑ 《君子爱“生”，得之有道》
- ❑ 1968年设计DNA测序方法



Journal of Molecular Biology

Volume 35, Issue 3, 1968, Pages 523-537



Structure and base sequence in the cohesive ends of bacteriophage lambda DNA

Ray Wu^{1,2}, A.D. Kaiser^{1,2}

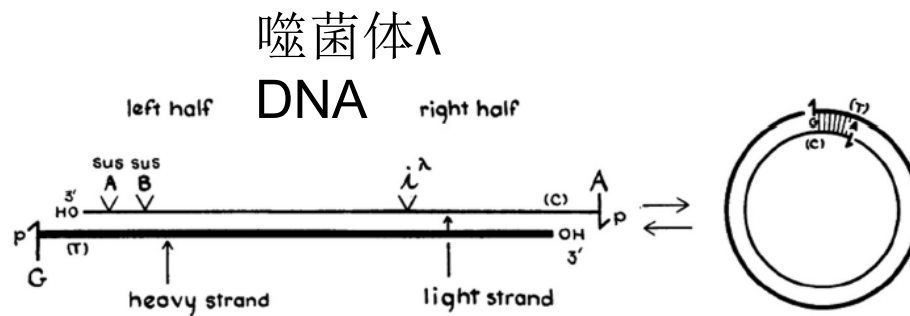
Received 4 March 1968, Revised 6 May 1968, Available online 8 July 2006



吴瑞先生：DNA测序



- ❑ 基本原理：位置特异性的引物延伸（Location specific-primer-extension principle in labeling the DNA）
- ❑ 1973年，Gilbert; 1975年，Sanger



Ray Wu as Fifth Business: Deconstructing collective memory in the history of DNA sequencing

Lisa A. Onaga



1954年



51. 吴瑞先生：第五先生还是DNA测序之父？

第一次知道吴瑞先生（图一）的名字，是看了饶毅老师写的博文《君子爱“生” 得之有道》，这篇博文后来收录在《饶议科学I》里，过年期间又读过一遍。其中有一句写的很有意思：“1971年吴瑞的引物延伸，是测序的一个关键步骤，给奖是可以的”。看到这儿我笑晕了：这都哪儿跟哪儿啊？所有课本上讲的都是Sanger测序法，所以显然是Sanger的贡献最大，况且诺奖都发了，还争这个有意思吗？另外，中国的语言历来有内涵，一般来说，“可以资助”的意思就是“不可以资助”，所以饶老师写博客为华人挣功劳的心意是挺好的，但显然不符合事实，对吧？咱读这篇文章的时候就是这么想的。



吴瑞先生，1954年于宾州大学
Stud Hist Philos Biol Biomed Sci., 2014, 46:1-14



Sci China C Life Sci., 2009, 52(2):99-100

DNA测序



- ❑ 分子生物学研究中最重要工具：确定DNA分子中的碱基组成和顺序
- ❑ 1980, Walter Gilbert & Fred Sanger: 利用DNA聚合酶测定DNA序列

1958: 胰岛素



Frederick Sanger

1980: DNA生化研究 & DNA测序



Paul Berg



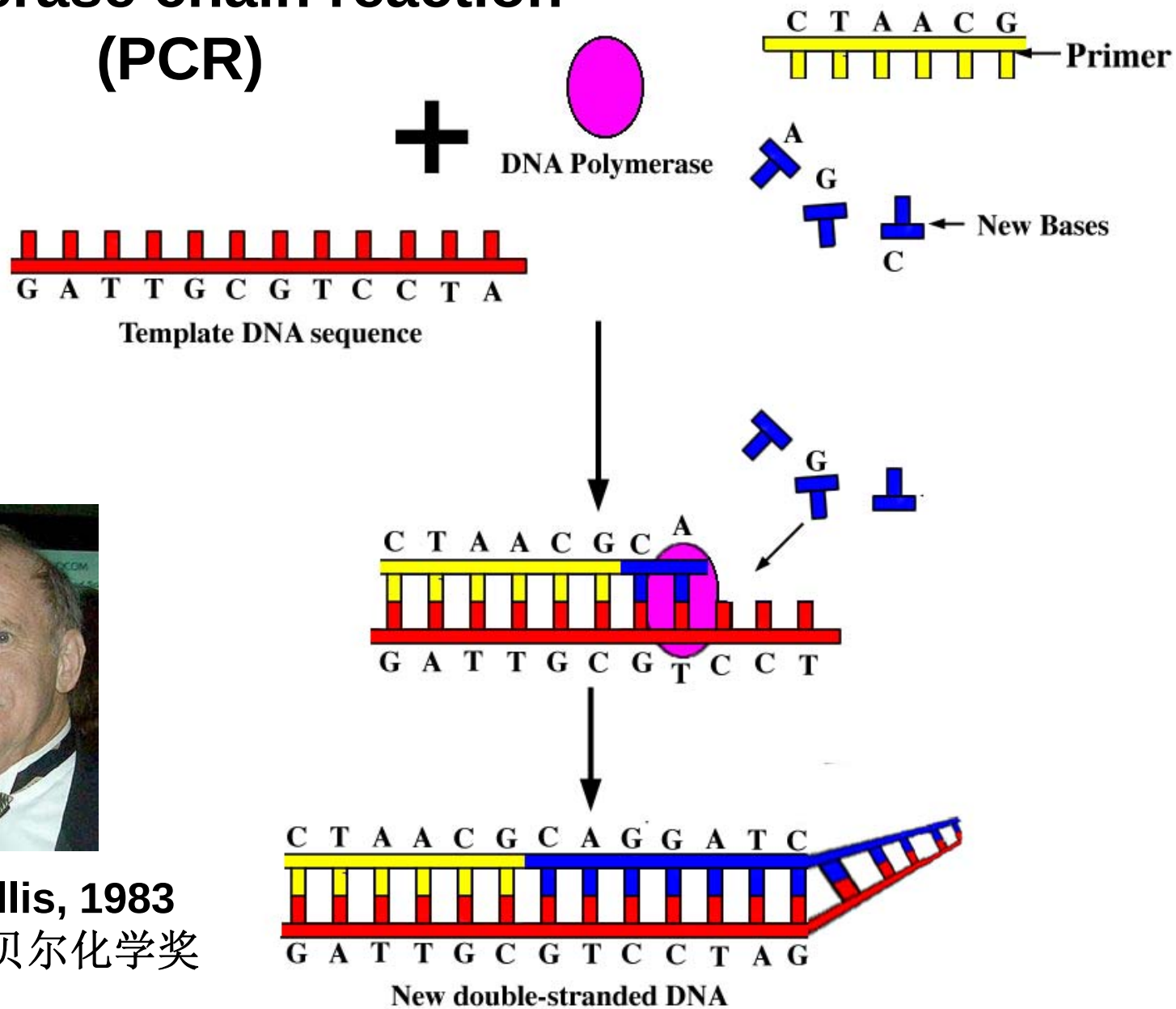
Walter Gilbert



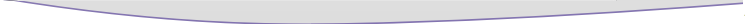
Frederick Sanger



Polymerase chain reaction (PCR)



Kary Mullis, 1983
1993年诺贝尔化学奖



自动测序

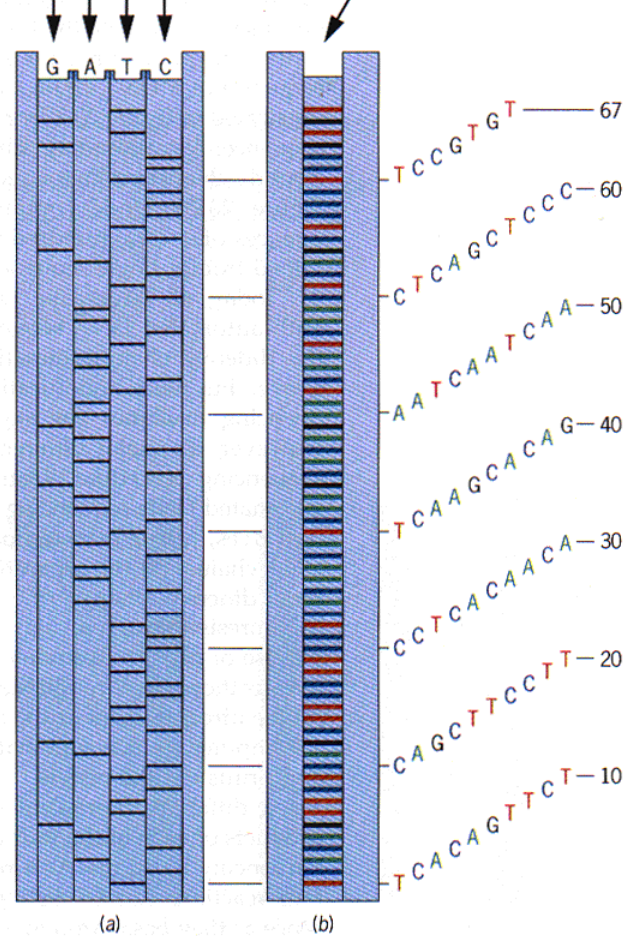


- ❑ “假如你真想改变一门学科，发明一些新的技术，这将使你超越以前人们曾经所看到的”
- ❑ 1980s, Leroy Hood发展了为4种终止核苷酸碱基标定不同荧光颜色的方法，改进了测序技术
- ❑ 4种碱基可在同一个反应中测序，并在一个条带中排列显示
- ❑ Hood: 利用计算机收集整理数据的先驱

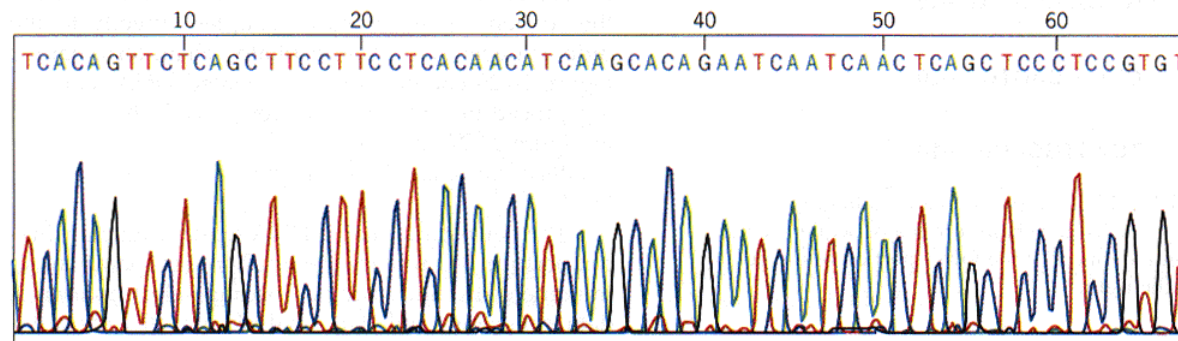


Each dideoxy chain-terminator reaction is loaded into a separate sample well.

All four dideoxy chain-terminator reactions are loaded into the same sample well.



Leroy Hood



第一代测序：自动测序仪



❑ 1986, Model 370A DNA Sequencing System

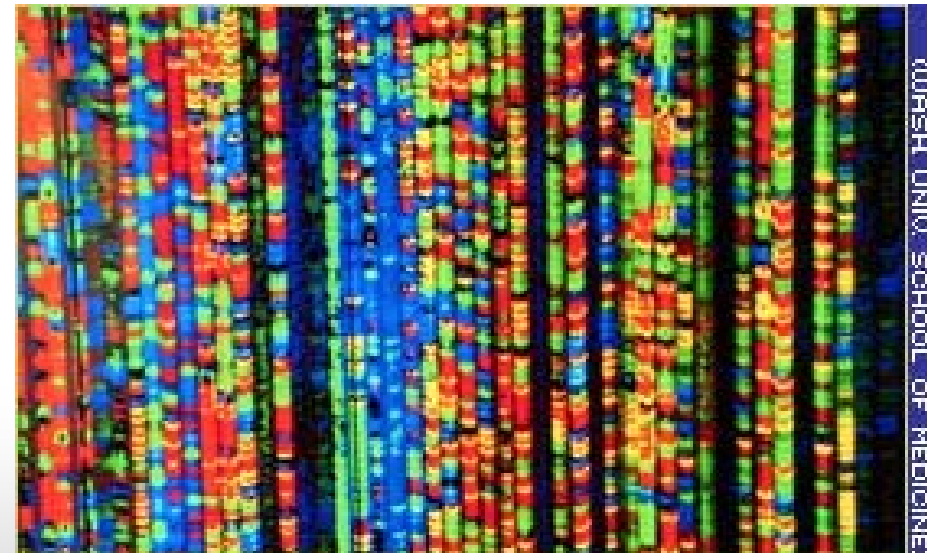
✿ Applied Biosystems, Inc (ABI)

✿ 1987, ABI 370

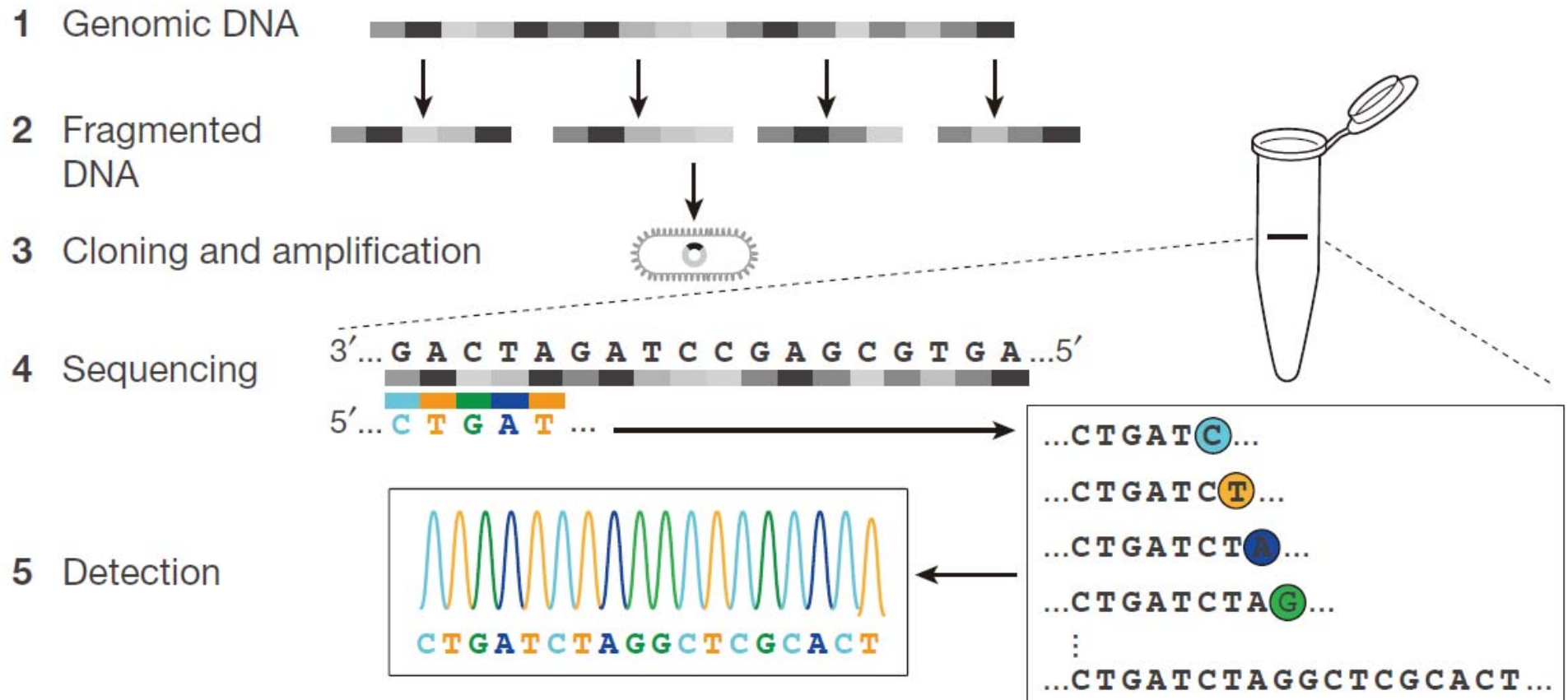
✿ 利用四种颜色，可直接阅读四种碱基

Company History

Year	Company Name
1981	Genetic Systems Company (GeneCo)
1982	Applied Biosystems, Inc. (ABI)
1993	Applied Biosystems, Perkin-Elmer
1996	PE Applied Biosystems
1998	PE Biosystems
2000	Applied Biosystems Group, Applera Corp
2002	Applied Biosystems
2008	Life Technologies
2014	Thermo Fisher Scientific



第一代测序



第二代测序



❑ Next Generation Sequencing, NGS

❑ Read: 读段

454



Illumina/Solexa



ABI-SOLID



纳米技术

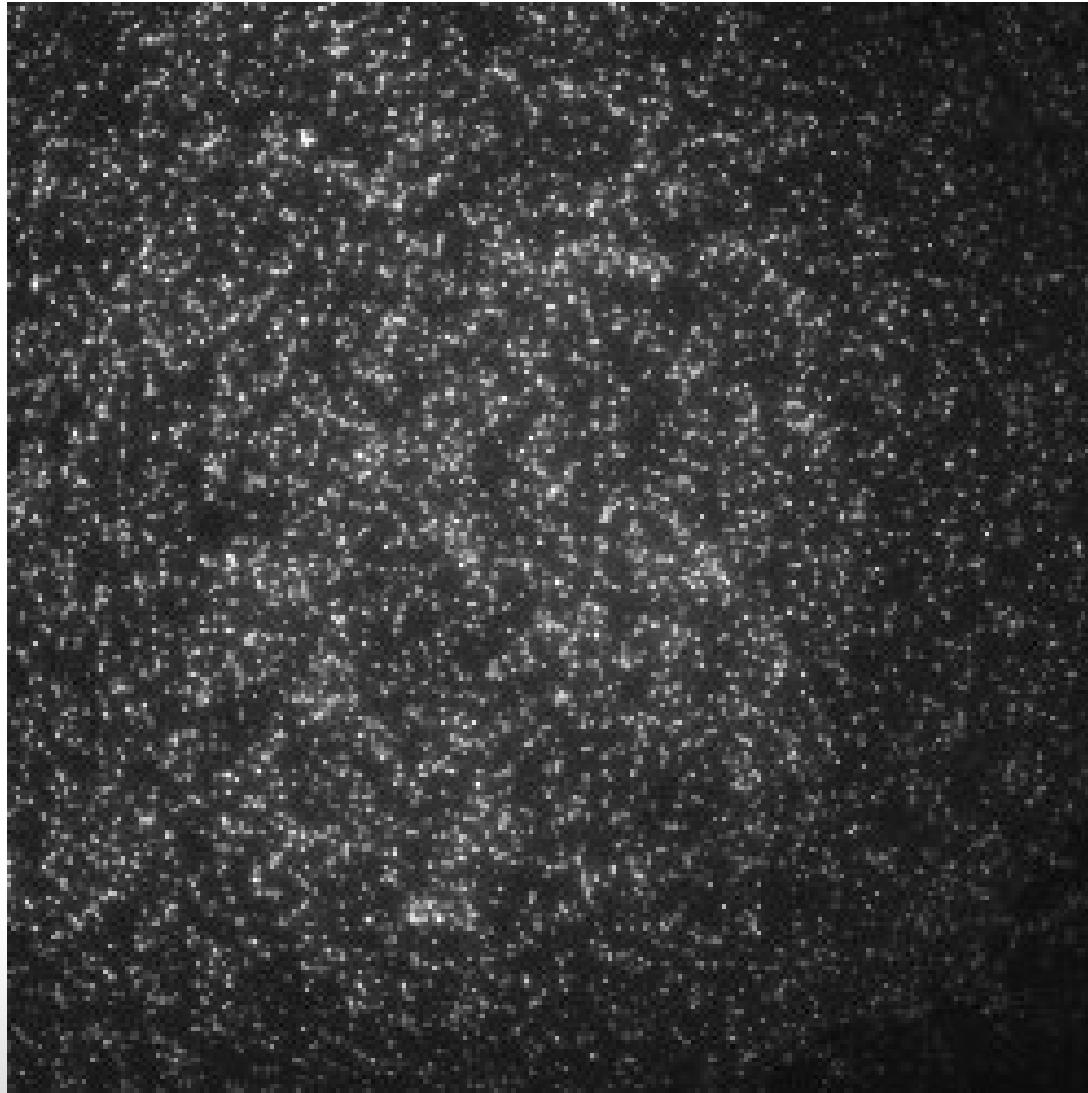


□ 每个系统的测序方式不同，但基本原理相似：

- ✿ 将待测序的DNA切断成小的片段
- ✿ 将每个DNA分子固定到固态材料的表面
- ✿ 同时扩增每一个分子
- ✿ 一次添加一个碱基，然后检测不同的信号：A, C, T, & G
- ✿ 需要高分辨图像处理技术

➡ (Solexa has 800 images @ 4 megapixels each)

Illumina/Solexa测序仪上的小区 (tile)





巨大容量的图像数据

- ❑ 原始图像的数据规模：TB级
 - ✿ Solexa =
 - ✿ ABI-SOLID: >
 - ✿ 454: <
- ❑ 图像数据立即被处理成亮度数据（点的位置以及亮度）
- ❑ 亮度数据需要被处理成碱基（basecall）（**A**, **C**, **T**, or **G**, 以及每一个碱基的概率分值）

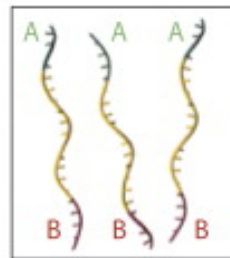
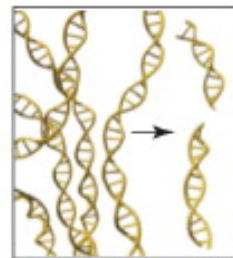
454测序原理

- ❑ 第一种的高通量DNA测序仪
- ❑ 2004年实现商业化
- ❑ 10个小时内可同时测定8个样本
- ❑ 焦磷酸测序（**pyrosequencing**）：
 - ✿ A. 液相反应
 - ✿ B. 一个bead上可扩增并连接百万个单一分子
 - ✿ C. 碱基上分解的焦磷酸引起发光反应，被扫描成图像
 - ✿ D. 每次扩增一个碱基
- ❑ 碱基类型的错误极少，存在插入/缺失错误
- ❑ Polonator: 原理相似，序列较短（26bp）

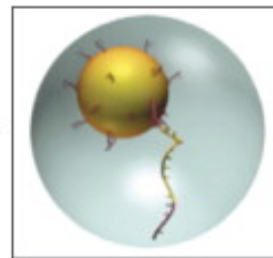


Roche (454) GSFLX Workflow:

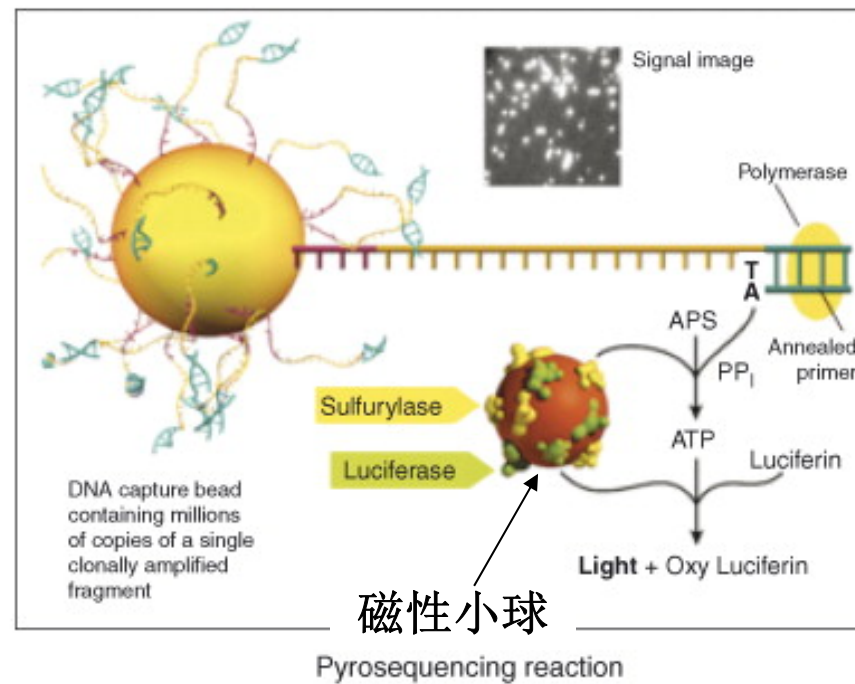
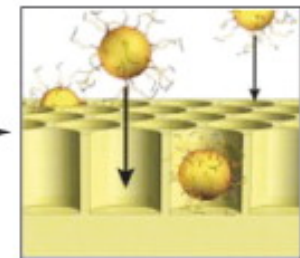
Library construction



Emulsion PCR



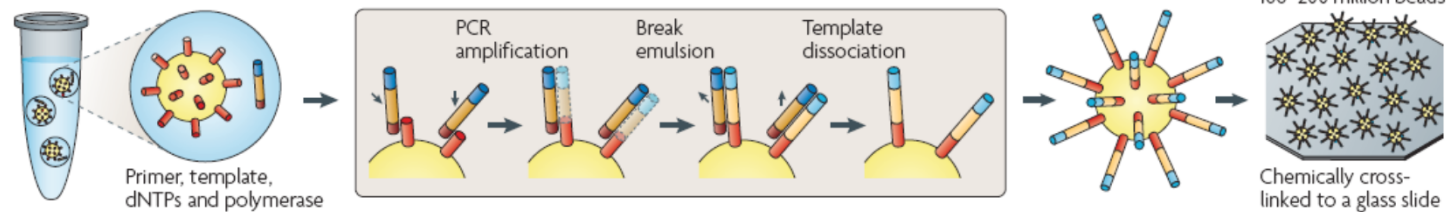
PTP loading



TRENDS in Genetics

a Roche/454, Life/APG, Polonator Emulsion PCR

One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion



GS FLX+ System (454)



状态指示灯

CCD相机

PicoTiterPlate
微孔板

多支吸管

泵、阀、
除泡器

试剂盒

读长：
最大1000bp
平均700bp

覆盖度：
>500bp: 85%碱基
>700bp: 45%碱基

通量：700Mb

每轮读段1M条

准确性：99.997%

每轮时间：23小时

Roche shutting down 454 sequencing business

Category: Medtech Published: 21 October 2013

Roche is shuttering its 454 life sciences sequencing operations and laying off about 100 employees, the company confirmed.



The 454 sequencers will be phased out in mid-2016, and the 454 facility in Branford, Conn., will be closed "accordingly," Roche said in a statement e-mailed to GenomeWeb Daily News.

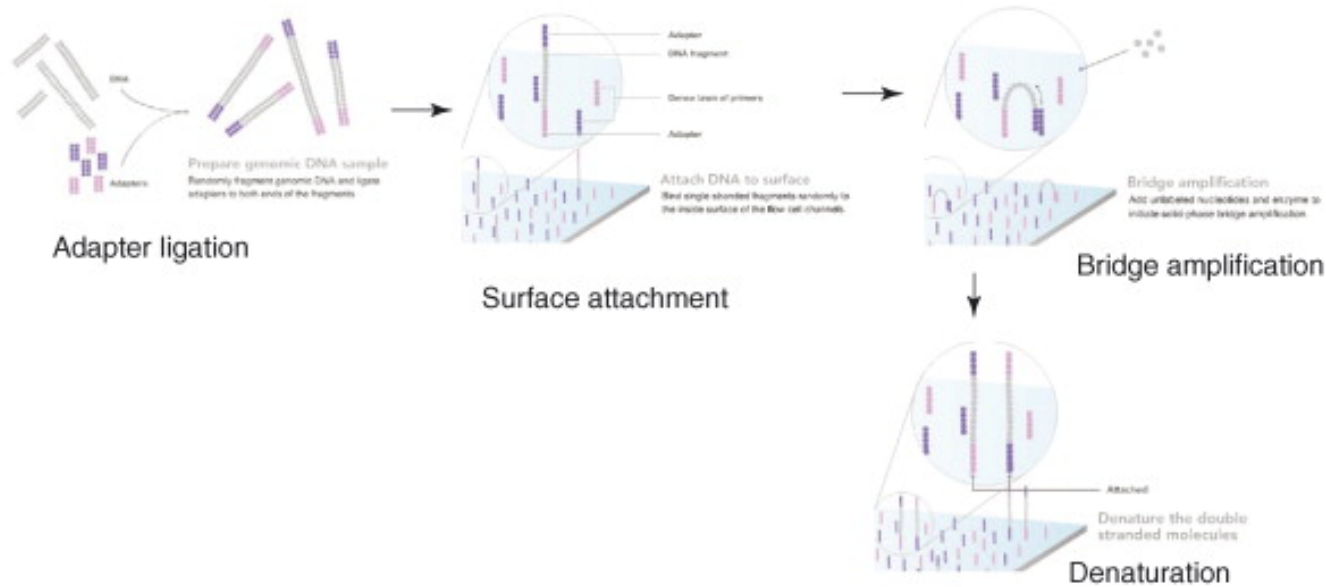
Illumina Genome Analyzer



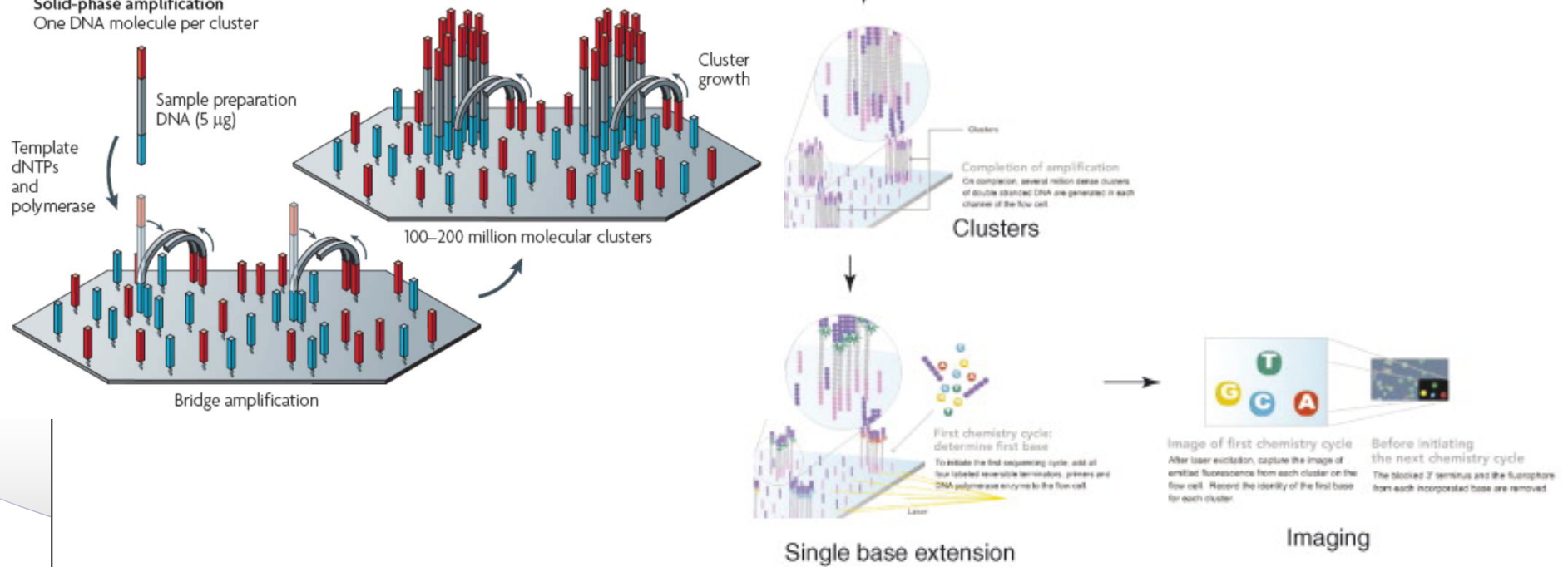
- ❑ 最早由Solexa开发设计，现在是Illumina的子公司
- ❑ 2006年实现商业化
- ❑ 每轮实验：7个样本/3日
- ❑ 低错误率，主要是碱基类型错误，有少量的插入/缺失
- ❑ Sequencing by synthesis



Illumina Genome Analyzer Workflow



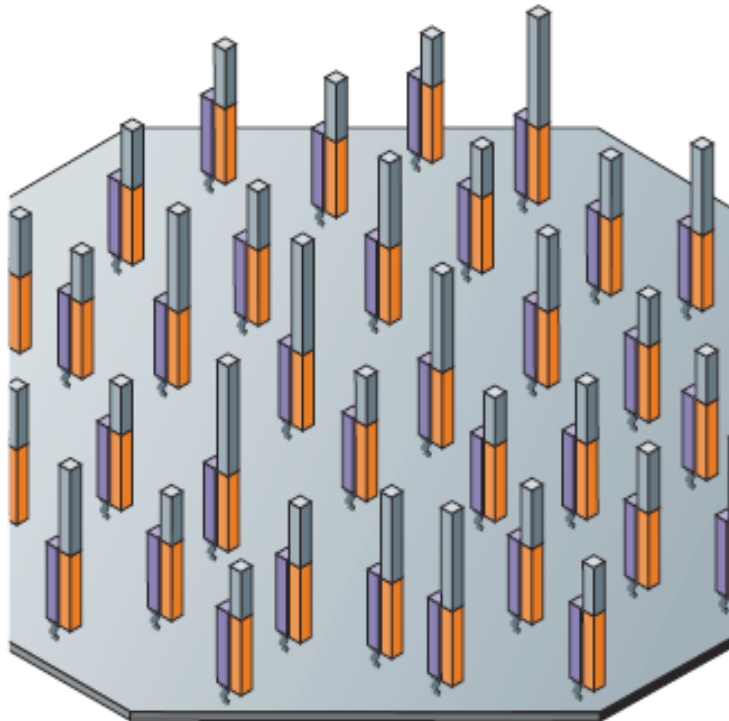
b Illumina/Solexa Solid-phase amplification One DNA molecule per cluster



Helicos BioSciences



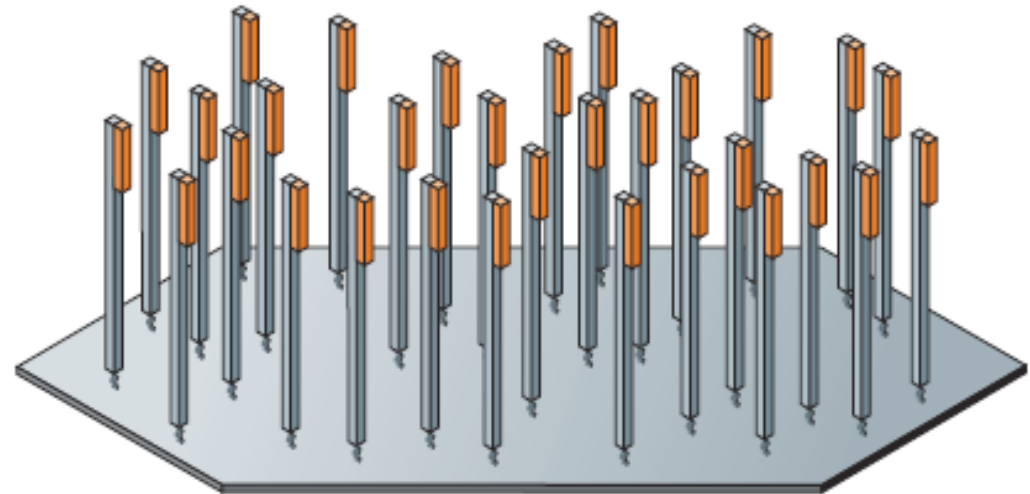
c Helicos BioSciences: one-pass sequencing
Single molecule: primer immobilized



Billions of primed, single-molecule templates

Adaptor固定，然后接模板

d Helicos BioSciences: two-pass sequencing
Single molecule: template immobilized



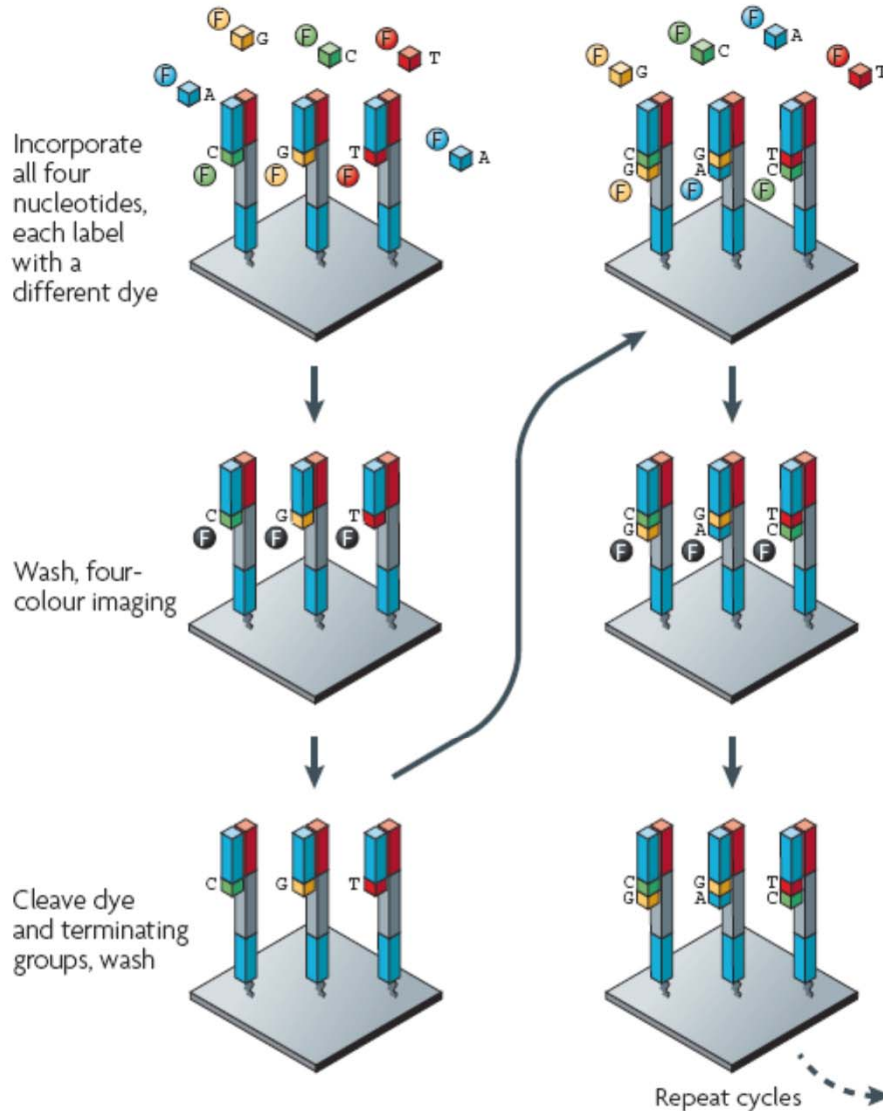
Billions of primed, single-molecule templates

模板先固定，然后接**Adaptor**

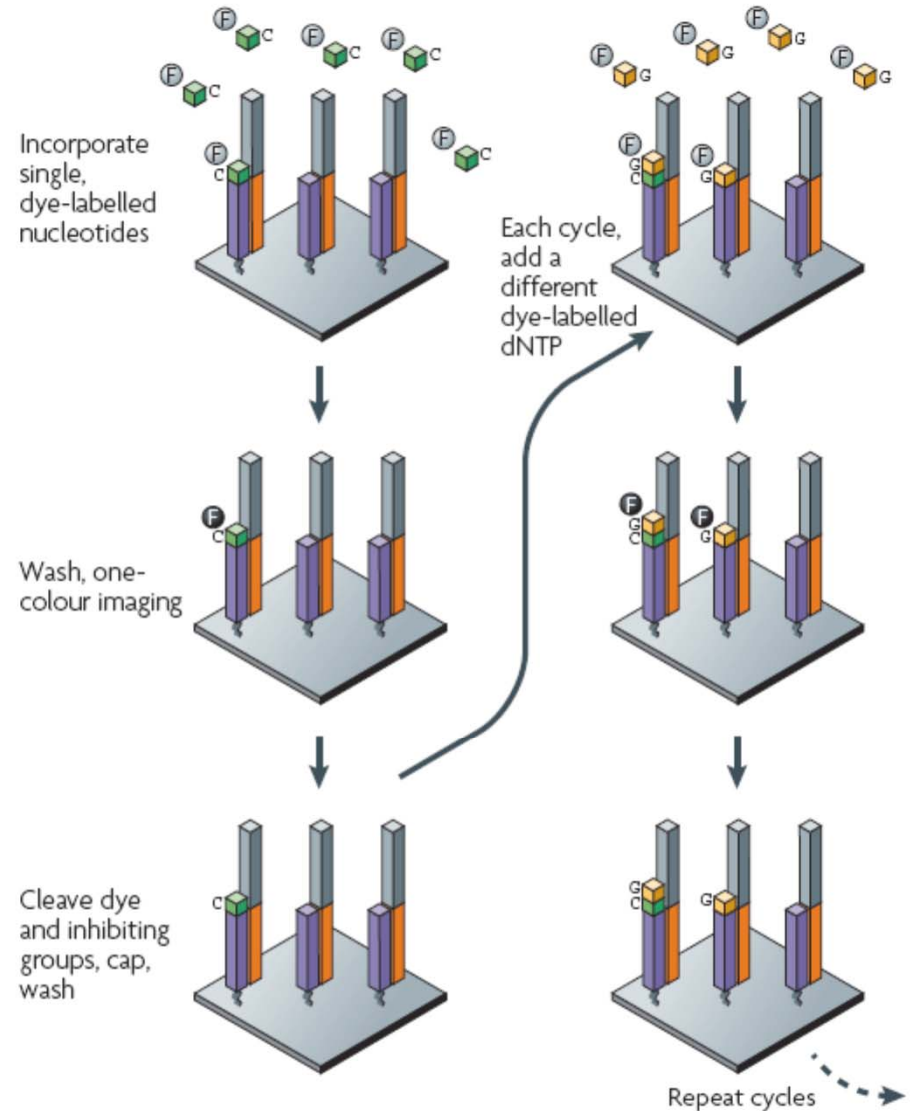
Illumina/Solexa vs. Helicos BioSciences



a Illumina/Solexa — Reversible terminators



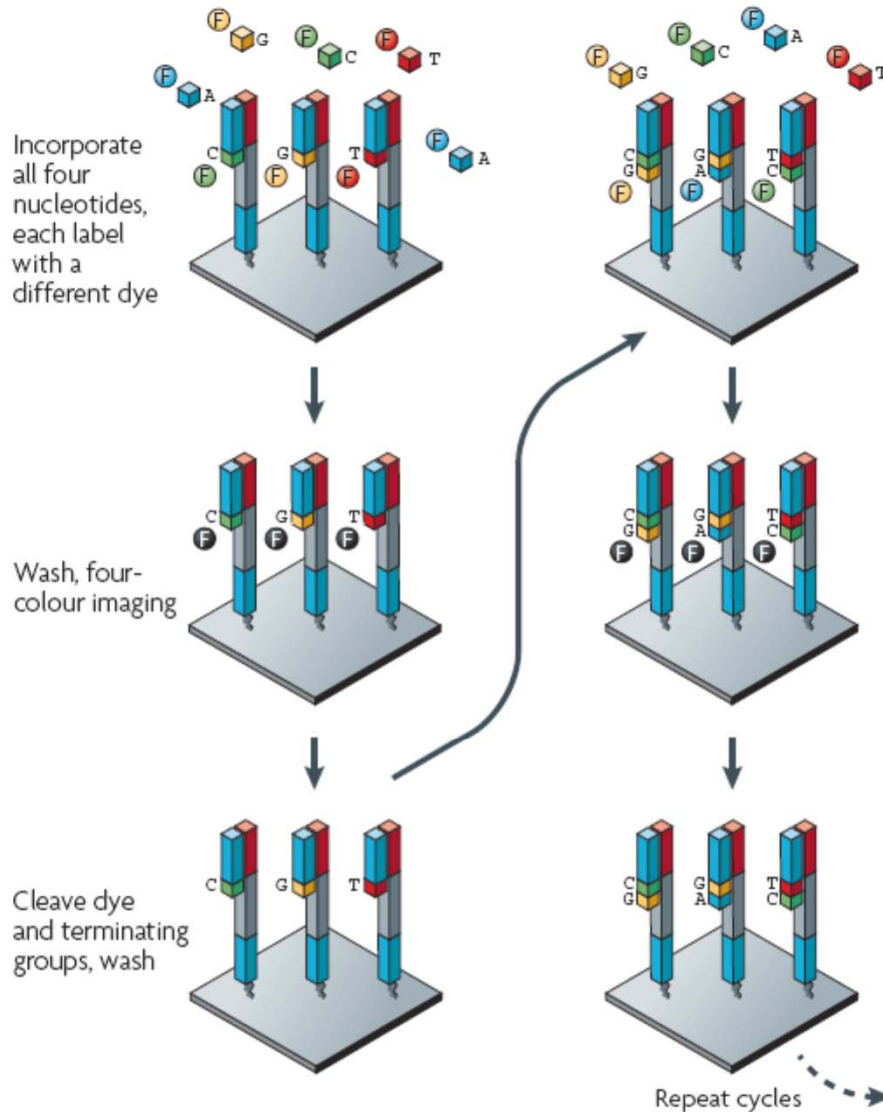
c Helicos BioSciences — Reversible terminators



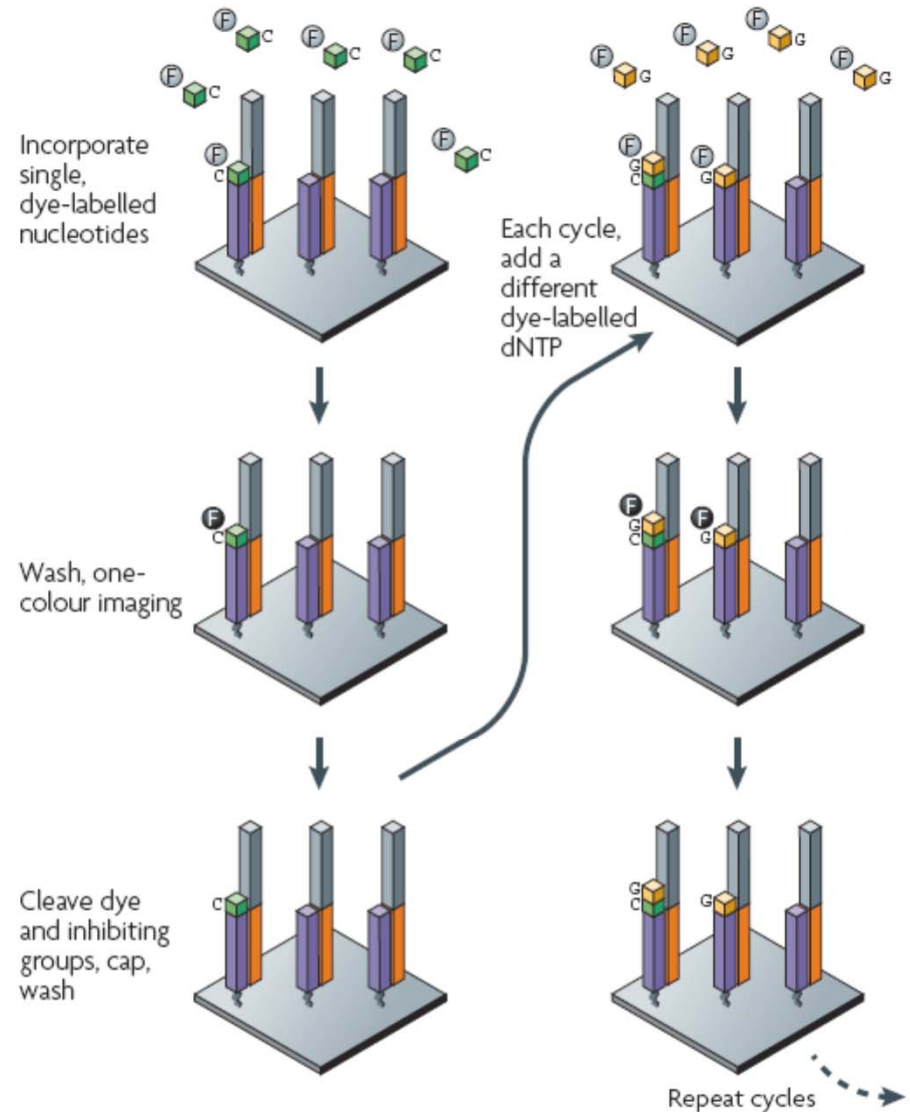
Illumina/Solexa vs. Helicos BioSciences



a Illumina/Solexa — Reversible terminators



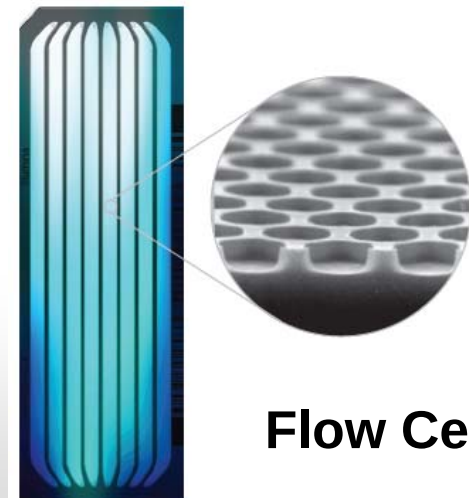
c Helicos BioSciences — Reversible terminators



HiSeq 3000/HiSeq 4000 (illumina)



	HiSeq 3000	HiSeq 4000
Flow Cell数	1	1或2
通量/天	>200Gb	>400Gb
每轮时间	<1-3.5天	<1-3.5天
每轮测人类基因组套数	6	12



Flow Cell

HiSeq X Ten (illumina)



- ❑ 由10台HiSeq X组成
- ❑ 每年测**人类基因组**：>18,000个
- ❑ 单个人类基因组 (30X)：<\$1,000
- ❑ 每轮通量：1.6-1.8Tb
- ❑ 每轮时间：<3天



MiSeq: 第一个FDA批准的NGS设备



- ❑ 2013年11月, FDA: 美国食品和药物管理局 (Food and Drug Administration)
- ❑ “Personalized medicine”: 利用个人的遗传信息为疾病的检测、治疗和预防提供更为精准的方法
- ❑ Illumina MiSeq Dx



Illumina – 桌面式测序仪



桌面式测序仪

生产规模的测序仪

SBS: Sequencing by synthesis



iSeq 100系统



MiniSeq系统



MiSeq系列+



NextSeq系列+

常用应用和方法	关键应用	关键应用	关键应用	关键应用
大型全基因组测序（人类、植物、动物）				
小型全基因组测序（微生物、病毒）				
外显子组测序				
靶向基因测序（扩增子、基因panel）				
全转录组测序				
使用mRNA-Seq进行基因表达谱分析				
靶向基因表达谱分析				
长片段扩增子测序*				
miRNA和Small RNA分析				
DNA-蛋白质相互作用分析				
甲基化测序				
16S宏基因组测序				
运行时间	9–17.5小时	4–24小时	4–55小时	12–30小时
最大数据产出	1.2 Gb	7.5 Gb	15 Gb	120 Gb
每次运行获得的最大read数	4 M	25 M	25 M [†]	400 M
最长读长	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp

Illumina – 生产规模的测序仪



桌面式测序仪

生产规模的测序仪



NextSeq系列+



HiSeq系列+



HiSeq X系列†



NovaSeq系列+

常用应用和方法	关键应用 ■	关键应用 ■	关键应用 ■	关键应用 ■
大型全基因组测序（人类、植物、动物）	●	●	●	●
小型全基因组测序（微生物、病毒）	●	●		●
外显子组测序	●	●		●
靶向基因测序（扩增子、基因panel）	●	●		●
全转录组测序	●	●		●
使用mRNA-Seq进行基因表达谱分析	●	●		●
miRNA和Small RNA分析	●	●		●
DNA-蛋白质相互作用分析	●	●		●
甲基化测序	●	●		●
鸟枪法宏基因组学测序	●	●		●
运行时间	12–30小时	<1–3.5天（HiSeq 3000/HiSeq 4000） 7小时–6天（HiSeq 2500）	<3天	19–40小时§
最大输出	120 Gb	1500 Gb	1800 Gb	6000 Gb ^l
每次运行获得的最大read数	400 M	5 B	6 B	20 B**
最长读长	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp

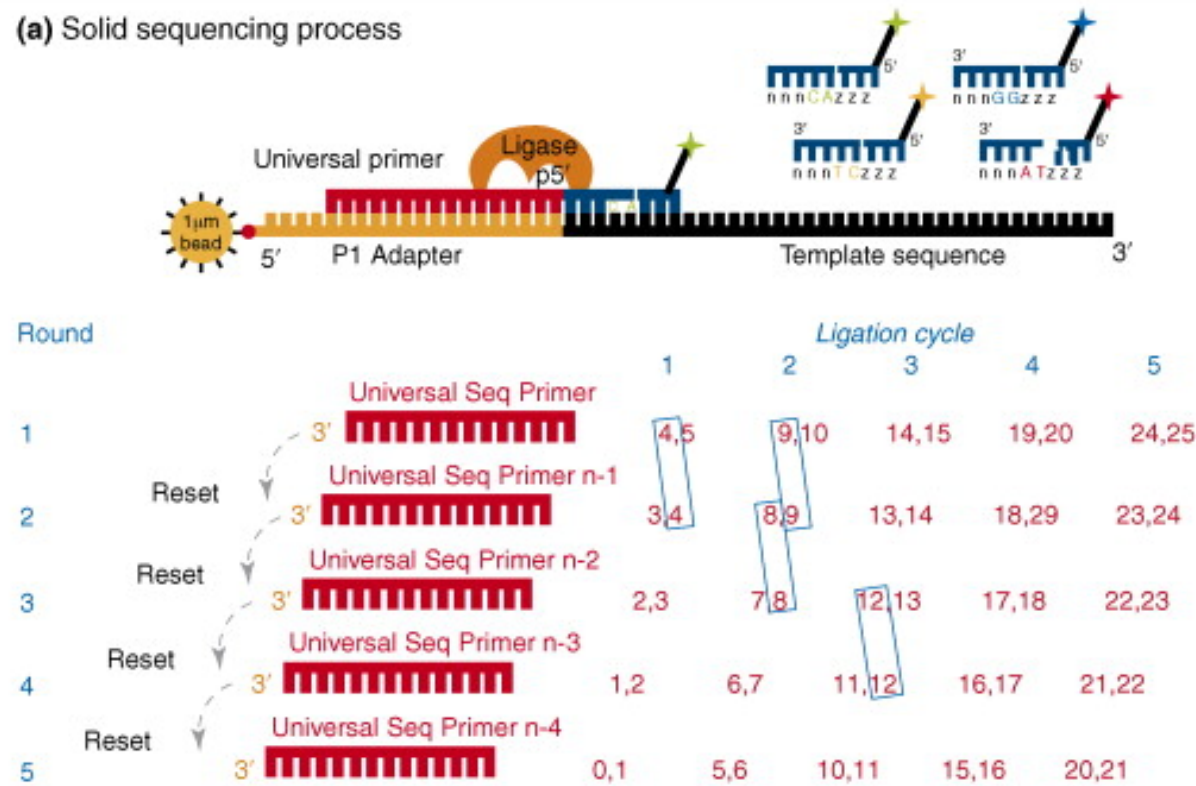
ABI-SOLID



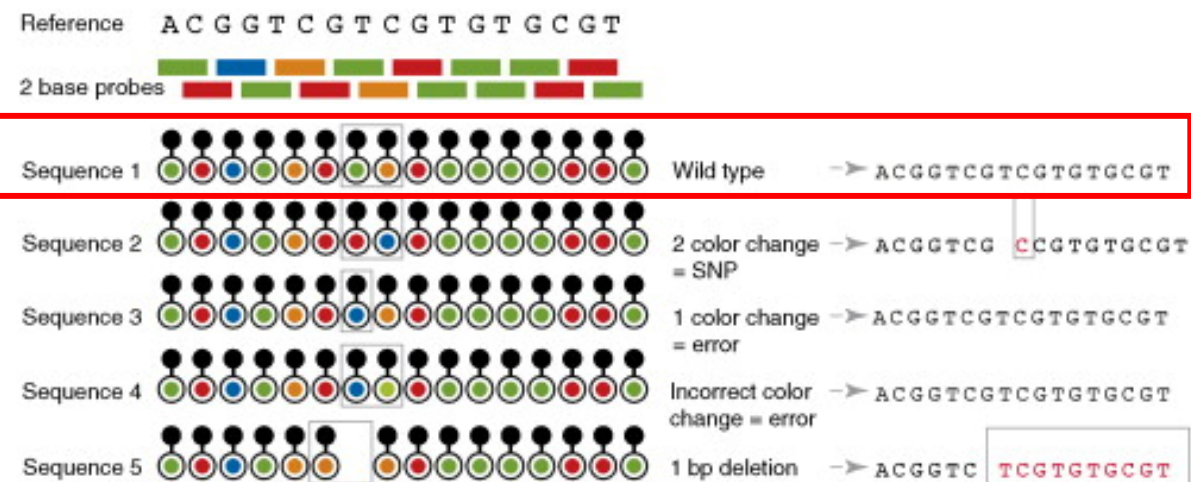
- ❑ 2007年实现商业化
- ❑ 每周产生20GB的数据
- ❑ 每轮: 6GB
- ❑ Sequencing by ligation
- ❑ “2 based encoding”
- ❑ “color-space” data



(a) Solid sequencing process



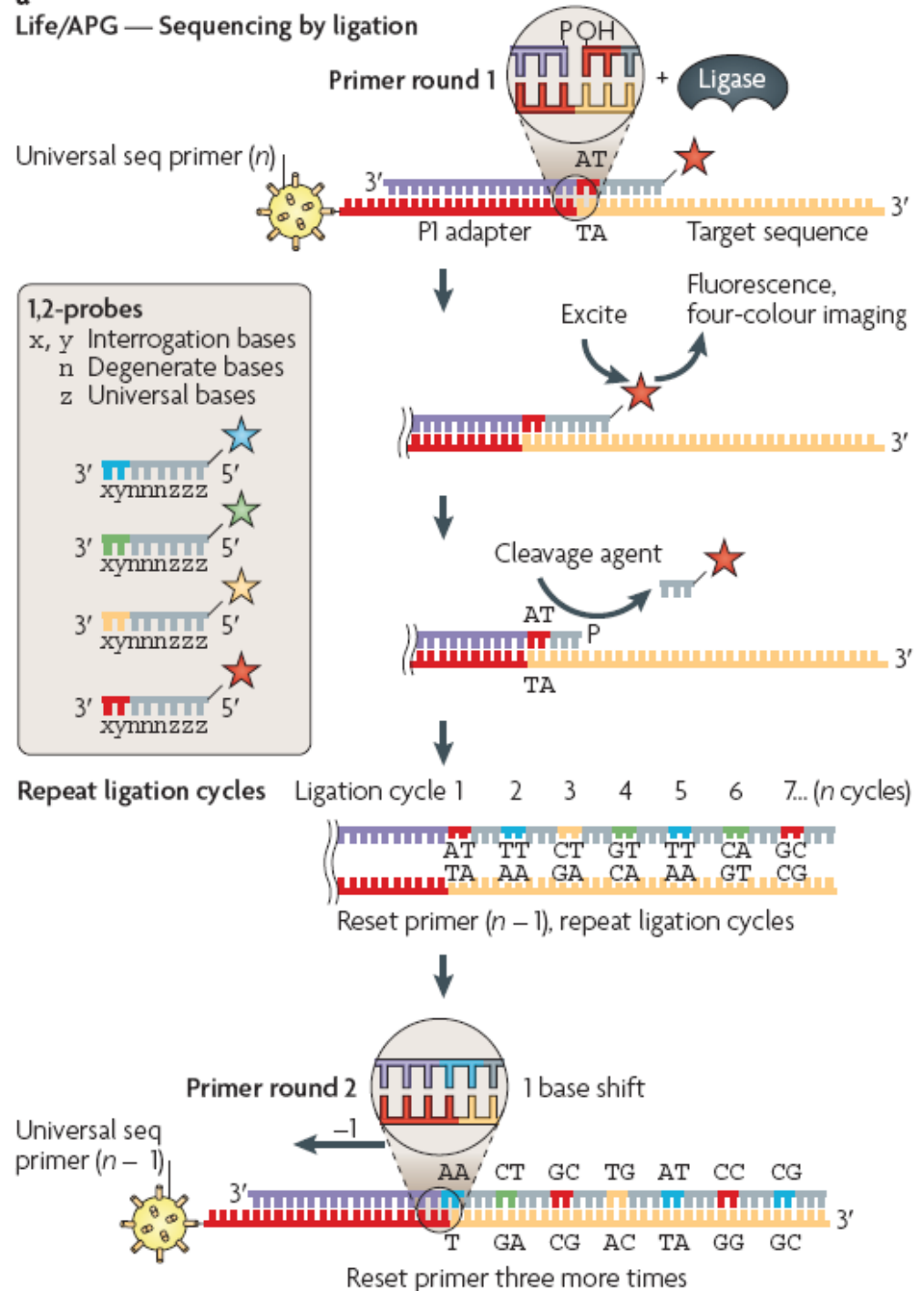
(b) Principles of two base encoding



ABI-SOLID

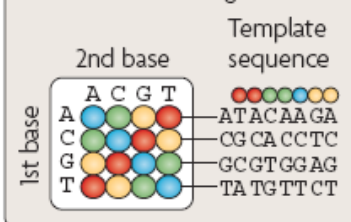
a

Life/APG — Sequencing by ligation

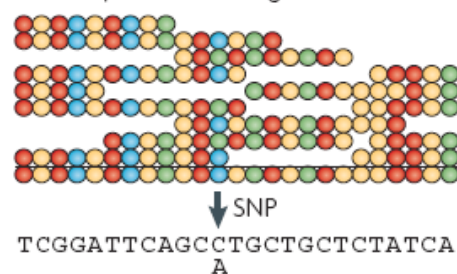


b

Two-base encoding: each target nucleotide is interrogated twice



Alignment of colour-space reads to colour-space reference genome



Bioinform

Revolocity: Complete Genomics

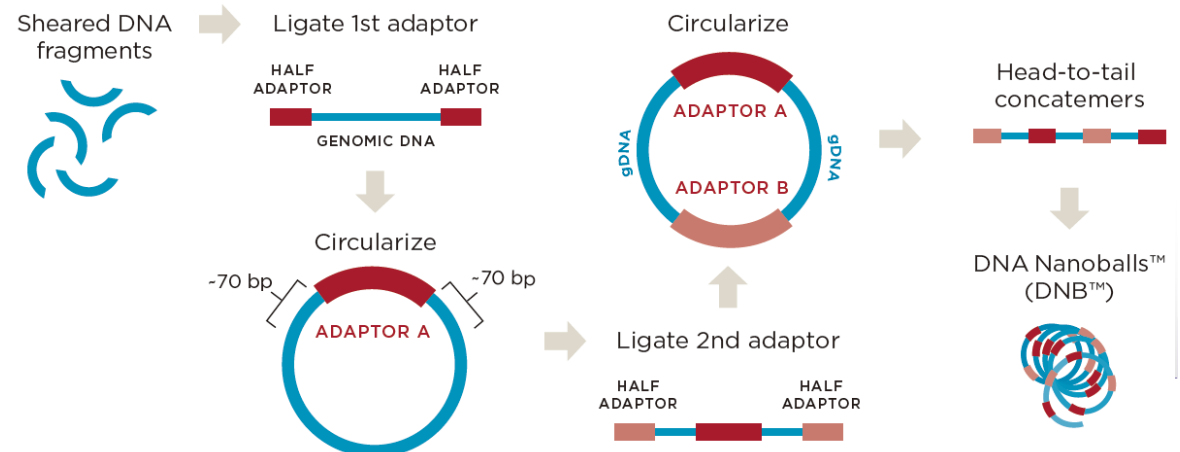


IMPACT OF CORRELATION ANALYSIS ON FALSE-POSITIVE RATE

Outcome	Without correlation analysis	With correlation analysis
SNP Sensitivity	98.5%	98.3%
SNP False-Positive Rate	12%	2.1%
Indel Sensitivity	83.4%	83.1%
Indel False-Positive Rate	26.3%	17.8%

Proprietary library preparation—the first step to quality results

华大基因 (BGI)



BGISEQ: 桌面化测序系统



BGISEQ-500



BGISEQ-50



	BGISeq-500	NextSeq-500
通常	8-200G	25-120G
读长	50SE/50PE/100SE/100PE	75SE/75PE/150PE
质量	Q30 bases > 85%	Q30 bases > 75%/80%
运行时间	~24h	12-30h

第二代测序 (Massively parallel)



1 Genomic DNA

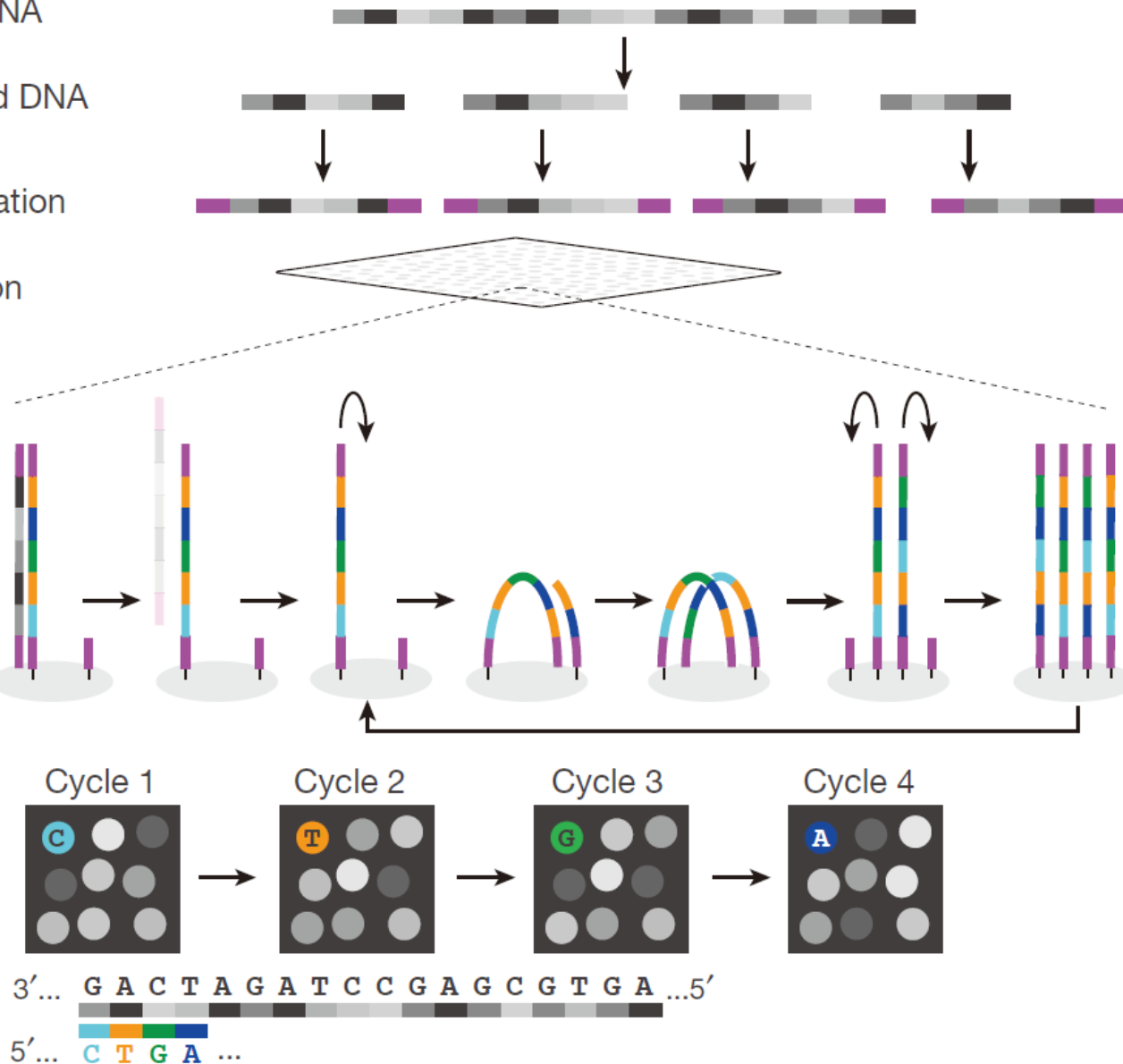
2 Fragmented DNA

3 Adaptor ligation

4 Amplification

Native DNA

5 Detection



每GB数据的花费 (2008)



- ❑ Solexa: ~\$6,000 per GB
- ❑ ABI-SOLID: \$6,000 per GB
- ❑ 454: \$85,000 per GB
- ❑ 读段越长越有价值
- ❑ 人类基因组3GB, 但是1x coverage不能完成拼装
- ❑ 2015: RNA-seq, ~\$1,000/GB
- ❑ 2017: RNA-seq, ~\$10/GB (65~80RMB)

短读段



- 第二代测序仪产生短读段 (Solexa = 36 bp, 2008)

- ✿ 全基因组组装困难

- ✿ 重复区域难以确认和组装

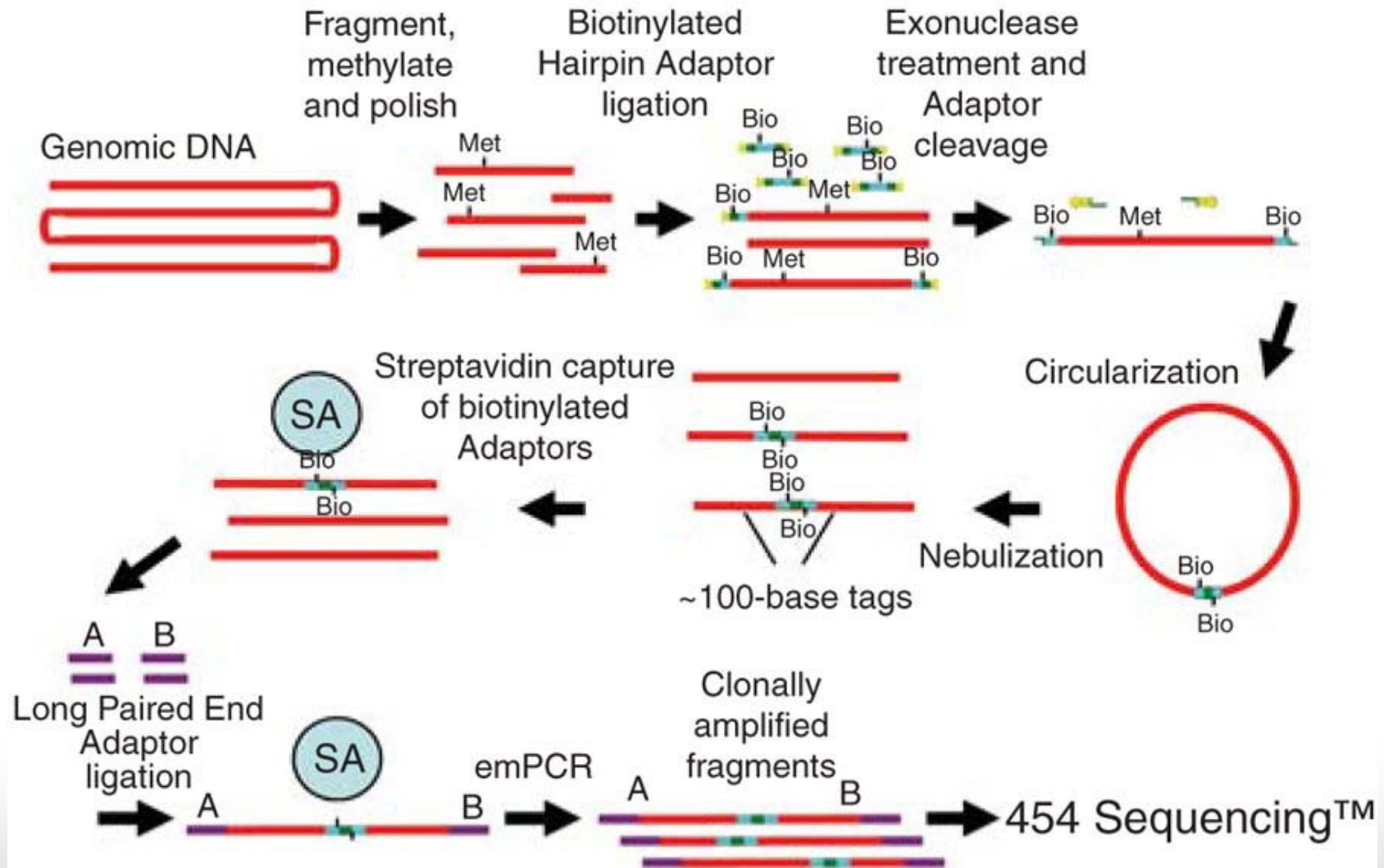
- 需要非常多重的覆盖率

- 解决方案：

- ✿ 新算法

- ✿ 更长的读段

Paired-End Sequencing (3KB)



第三代测序：PacBio Systems



- ❑ The “third-generation sequencing”: long-read sequencing
- ❑ Single Molecule, Real-Time (SMRT) technology
- ❑ DNA聚合酶固定
- ❑ DNA单分子通过纳米尺度的小孔
- ❑ “real-time” sequencing



PacBio RS II



Sequel

	RS II	Sequel
平均读长	10 ~ 15kb	8 ~ 12kb
ZMWs	15万	100万
数据量 / SMRTCell	500Mb~1Gb	5 ~ 10Gb
SMRTCell No./Run	1~16	1~16
Run time / SMRTCell	0.5~6hrs	0.5~6hr
Multiplex Amplicons	384	1536

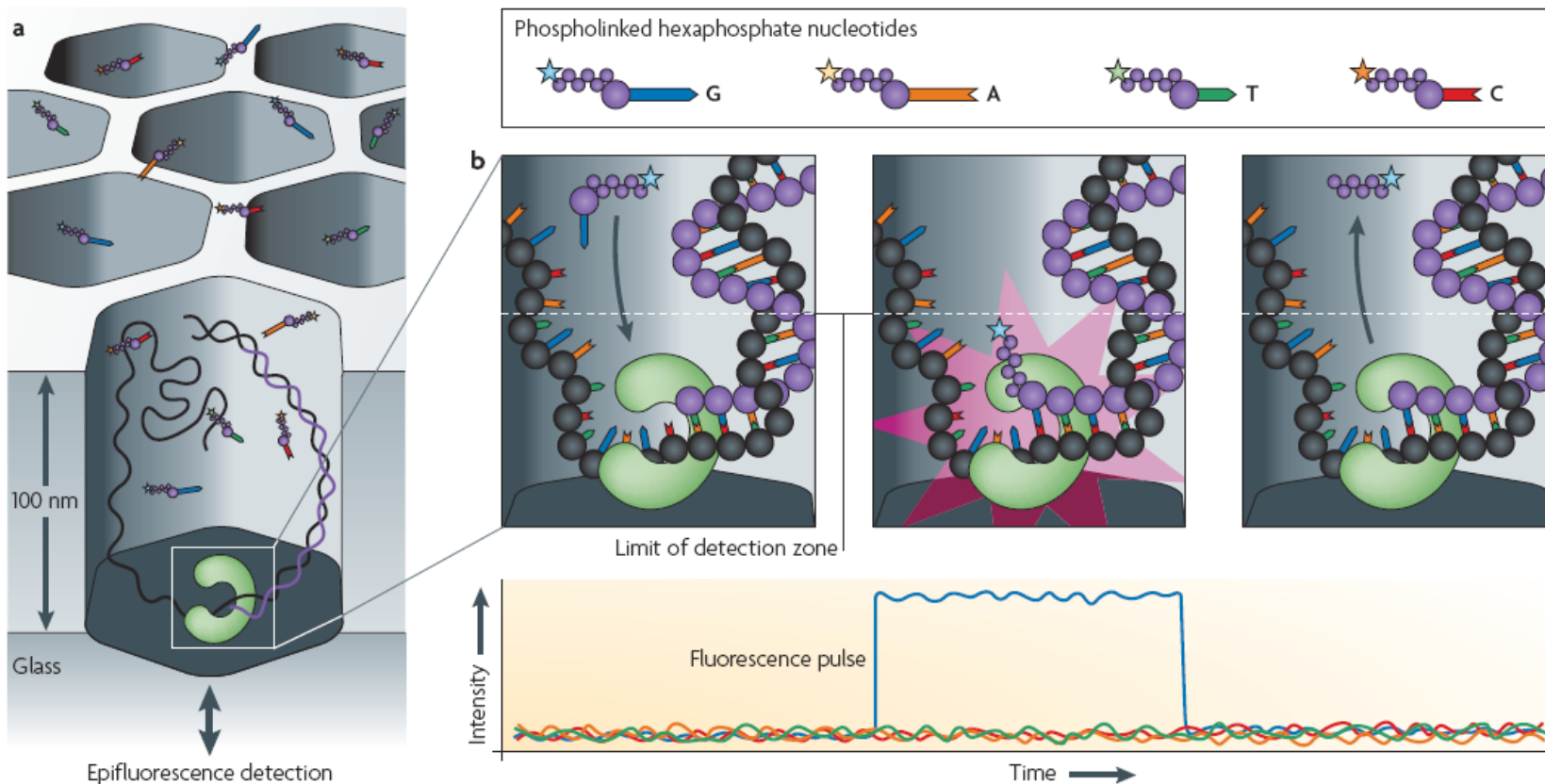
informatics, 2025, HUST

PacBio Systems测序原理



- ❑ ZMW孔（Zero-Mode Waveguides, 零模波导孔）：微孔直径小于光的波长时，光强会急剧衰减
- ❑ 孔底部仅能固定单分子DNA聚合酶和单分子DNA

Pacific Biosciences — Real-time sequencing

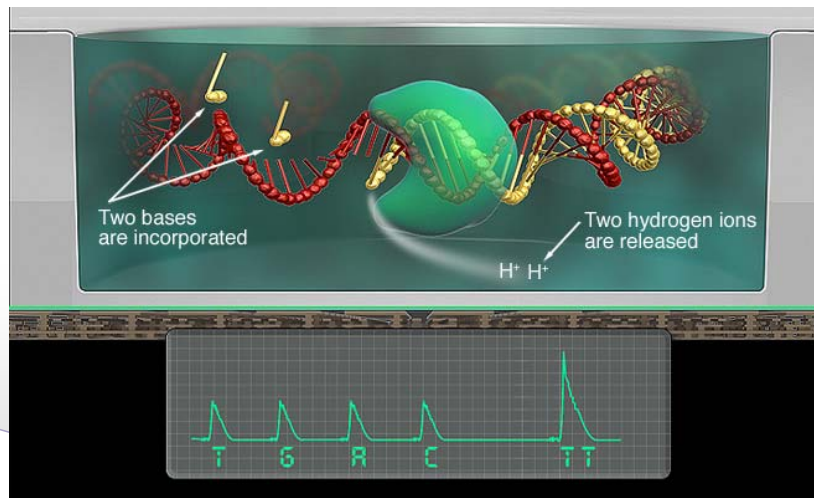


Ion Torrent™ (Life Technologies)



Criteria	Ion Torrent Proton	Illumina HiSeq 2500
System Price	\$243,000	\$740,000
Annual Service cost (yr 2 & 3)	\$19,400	\$59,200
Per Gb cost	\$16.67	\$46.00
Per Run Gb	~ 60 Gb (est.)	120 Gb
Per Run cost	\$1,000	\$5520 (est)
Three year TCO @ 30% utilization	\$731,800	\$2,100,400
Three year TCO w/ HiSeq Upgrade	\$731,800	\$1,410,400
Time from library to data	8 hours	27 hours
Throughput per 40h workweek	600 Gb	480 Gb
Accuracy	99%	99%
Expected readlength at launch	200 base-pairs	150 base-pairs
Potential readlength improvement	400 base-pairs	250 base-pairs
Ease of use	+++	+++

测序原理：向 DNA 聚合物中加入 dNTP，会释放出氢离子。采用半导体测定这些氢离子引起的 pH 变化，通过同时测量数百万起此类变化，可以测定各个片段的序列



Oxford Nanopore Technologies



□ MinION (100 g, USB 3.0)

✿ 512个纳米孔通道，样品准备~10分钟

✿ 读段>200kb，10-20Gb/48小时

□ PromethION

✿ 3000个纳米孔通道，11Tb/48小时

Nanopore devices perform DNA/RNA sequencing directly and in real time.
The technology is scalable from miniature devices to high-throughput installations.

[How it works](#)

[Compare products](#)



SmidgION



Flongle



MinION

GridION X5



GridION



PromethION

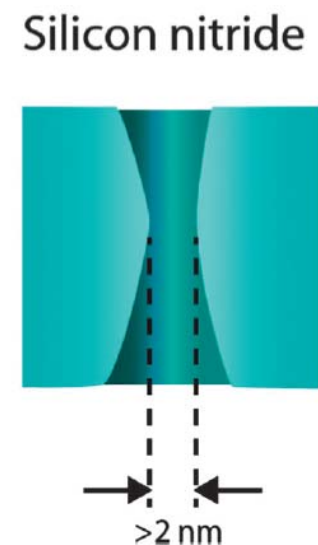
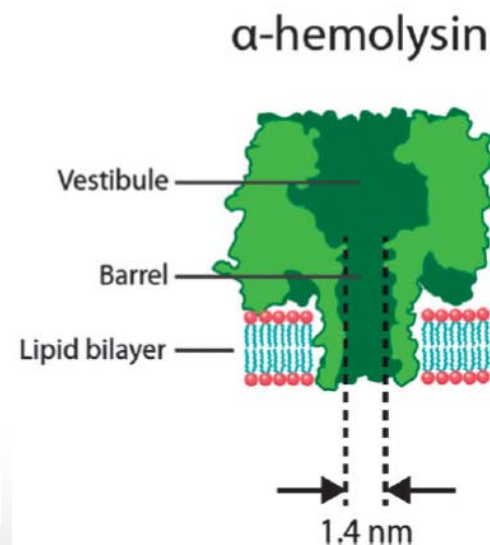
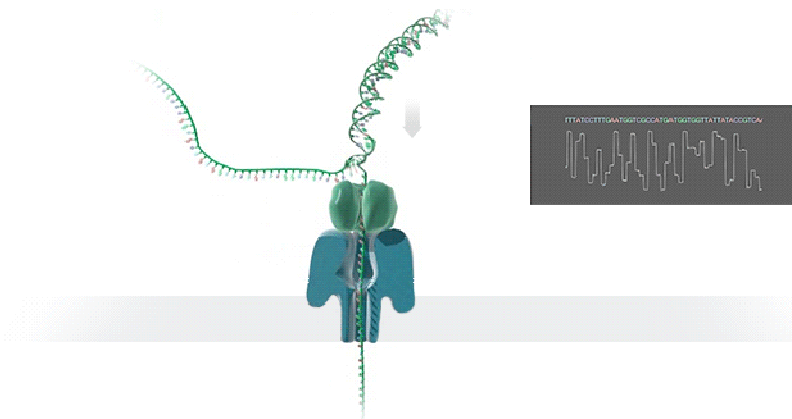
Nanopore测序原理



□ DNA链的直径:

⚙ B-DNA (2.0nm); A-DNA (2.6nm)

□ Nanopore类型: 生物孔和材料孔



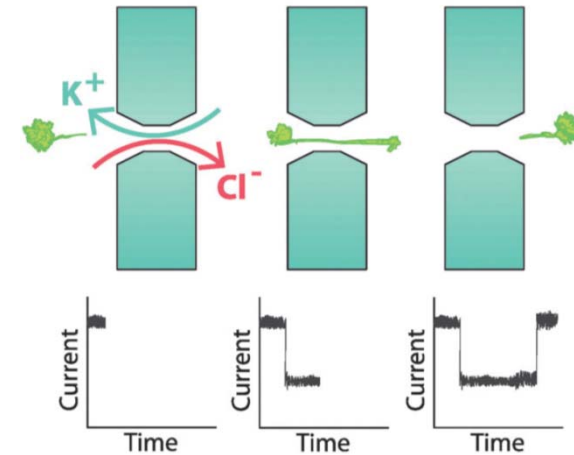
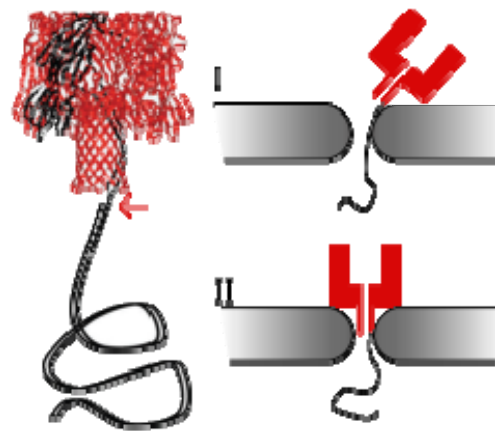
Nanopore测序原理



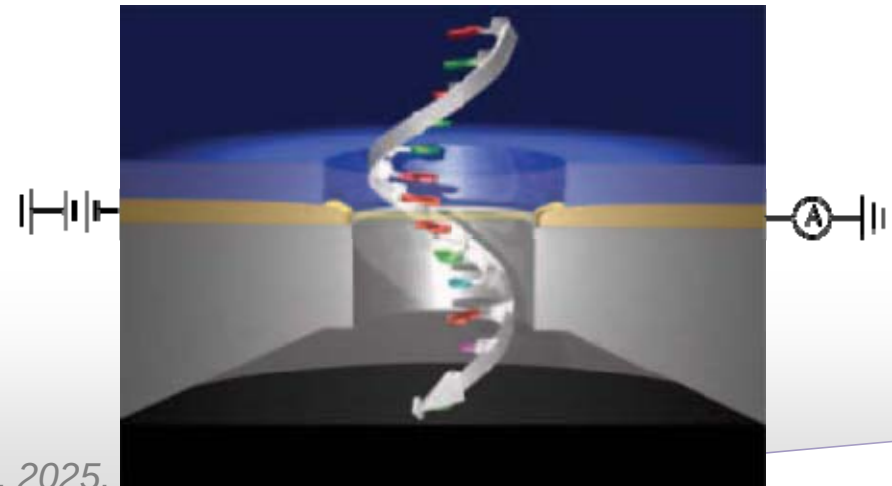
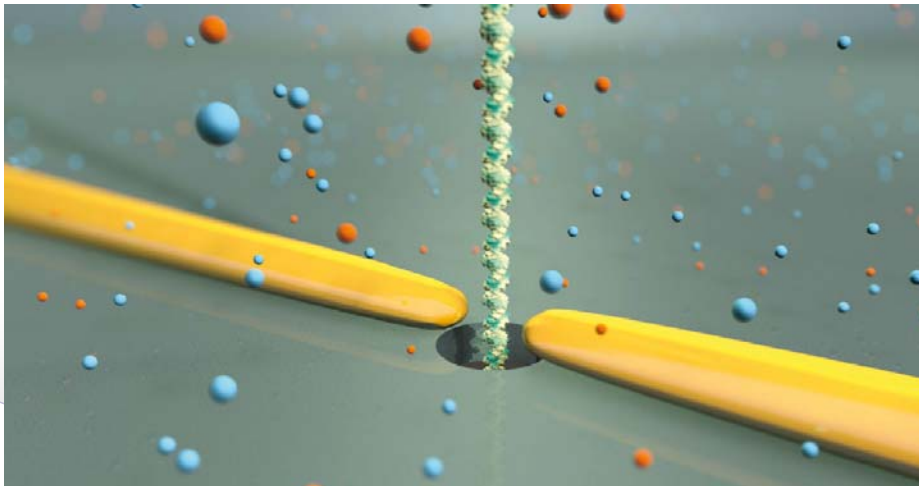
□ 膜两侧加电压

⚙ 开孔电流

⚙ 阻遏电流



□ 纳米孔两侧加电压



高通量测序方法的比较



Comparison of high-throughput sequencing methods^{[63][64]}

Method	Read length	Accuracy (single read not consensus)	Reads per run	Time per run	Cost per 1 million bases (in US\$)	Advantages	Disadvantages
Single-molecule real-time sequencing (Pacific Biosciences)	10,000 bp to 15,000 bp avg (14,000 bp N50); maximum read length >40,000 bases ^{[65][66][67]}	87% single-read accuracy ^[68]	50,000 per SMRT cell, or 500–1000 megabases ^{[69][70]}	30 minutes to 4 hours ^[71]	\$0.13–\$0.60	Fast. Detects 4mC, 5mC, 6mA. ^[72]	Moderate throughput. Equipment can be very expensive.
Ion semiconductor (Ion Torrent sequencing)	up to 600 bp ^[73]	98%	up to 80 million	2 hours	\$1	Less expensive equipment. Fast.	Homopolymer errors.
Pyrosequencing (454)	700 bp	99.9%	1 million	24 hours	\$10	Long read size. Fast.	Runs are expensive. Homopolymer errors.
Sequencing by synthesis (Illumina)	MiniSeq, NextSeq: 75–300 bp; MiSeq: 50–600 bp; HiSeq 2500: 50–500 bp; HiSeq 3/4000: 50–300 bp; HiSeq X: 300 bp	99.9% (Phred30)	MiniSeq/MiSeq: 1–25 Million; NextSeq: 130–00 Million, HiSeq 2500: 300 million – 2 billion, HiSeq 3/4000 2.5 billion, HiSeq X: 3 billion	1 to 11 days, depending upon sequencer and specified read length ^[74]	\$0.05 to \$0.15	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipment can be very expensive. Requires high concentrations of DNA.
Sequencing by ligation (SOLiD sequencing)	50+35 or 50+50 bp	99.9%	1.2 to 1.4 billion	1 to 2 weeks	\$0.13	Low cost per base.	Slower than other methods. Has issues sequencing palindromic sequences. ^[75]
Nanopore Sequencing ^[76]	Dependent on library prep, not the device, so user chooses read length. (up to 500 kb reported)	~92–97% single read (up to 99.96% consensus)	dependent on read length selected by user	data streamed in real time. Choose 1 min to 48 hrs	\$500–999 per Flow Cell, base cost dependent on expt	Longest individual reads. Accessible user community. Portable (Palm sized).	Lower throughput than other machines, Single read accuracy in 90s.
Chain termination (Sanger sequencing)	400 to 900 bp	99.9%	N/A	20 minutes to 3 hours	\$2400	Useful for many applications.	More expensive and impractical for larger sequencing projects. This method also requires the time consuming step of plasmid cloning or PCR.

https://en.wikipedia.org/wiki/DNA_sequencing

序列数据的存储



- ❑ 核酸**四**大数据库：GenBank, EBI-ENA, **NGDC**, DDBJ
- ❑ GSA (Genome Sequence Archive)
 - ✿ Raw sequence read
- ❑ Ensembl数据库：基因组注释
- ❑ Refseq数据库
- ❑ NCBI的Gene信息数据库
- ❑ 蛋白质序列：UniProt数据库

GenBank



❏ <https://www.ncbi.nlm.nih.gov/genbank/>

An official website of the United States government [Here's how you know](#) ✓

National Library of Medicine
National Center for Biotechnology Information

Log in

GenBank Nucleotide Search

GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Documentation Other

GenBank Overview

What is GenBank?

GenBank[®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research, 2013 Jan;41\(D1\):D36-42](#)). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

GenBank Resources

- [GenBank Home](#)
- [Submission Types](#)
- [Submission Tools](#)
- [Search GenBank](#)
- [Update GenBank Records](#)



☐ <https://www.ebi.ac.uk/ena/>

The screenshot shows the top section of the EBI-ENA website. At the top is a dark blue navigation bar with links for EMBL-EBI, Services, Research, Training, and About us, along with a search icon. Below this is a teal banner featuring the ENA logo (European Nucleotide Archive) on the left. On the right side of the banner, there are two search input fields: 'Enter text search terms' with a 'Search' button, and 'Enter accession' with a 'View' button. Examples are provided for both: 'histone, BN000065' for the first and 'Taxon:9606, BN000065, PRJEB402' for the second. Below the banner is a light blue navigation bar with links for Home, Submit, Search, Rulespace, About, and Support, each with a dropdown arrow.

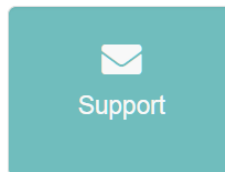
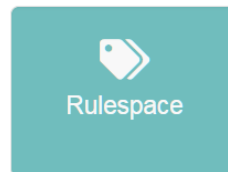
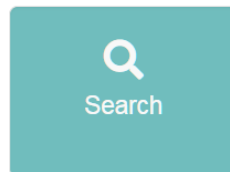
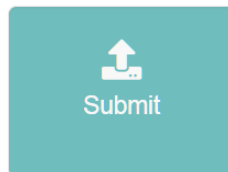
We recommend that you subscribe to the ENA-announce mailing list for updates on services.

For SARS-CoV-2 data submissions, users should contact us in advance of submission at virus-dataflow@ebi.ac.uk for specific advice on options and to access the highest levels of support. We have also launched a [Drag-and-Drop Data Submission Service](#) (currently in Beta) suitable for certain SARS-Cov-2 submissions. We are inviting submitters to try this out. Please contact us at the email above for details.

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. [More about ENA.](#)

Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.



[Tweets by ENA](#)

NGDC



☐ <https://ngdc.cncb.ac.cn/>

DatabasesToolsStandardsPublicationsAbout

国家基因组科学数据中心
National Genomics Data Center

NSTI
国家科技资源共享服务平台
National Science & Technology Information Infrastructure

登录语言 / Language ▾

面向我国人口健康和社会可持续发展的重大战略需求，建立生命与健康大数据汇交存储、安全管理、开放共享与整合挖掘研究体系，研发大数据前沿交叉与转化应用的新方法和新技术，建成支撑我国生命科学发展、国际领先的基因组科学数据中心。

▼ All databases

Find a bioproject, biosample, gene, protein, tool, database...

Q Search

e.g., PRJCA000126; SAMC000385; tp53; EGFR; human; KaKs_Calculator; GenBank

国家基因组科学数据中心发布细胞分类库Cell Taxonomy，欢迎访问使用！

提交

科学项目数据汇交

人类遗传资源信息管理备份

序列搜索比对

新冠病毒信息库

文献库

NGDC/NODE



❏ <https://www.biosino.org/node/>

The screenshot shows the NGDC/BMDC NODE website. The header includes the NGDC/BMDC logo, the NODE logo, and navigation links for Database, Tool & Pipeline, News, and About. Below the header is a search bar and a navigation menu with links for Project, Experiment, Sample, Run, Analysis, Statistics, and Help. The main content area features a featured article titled "Cloning of Macaque Monkeys by Somatic Cell Nuclear Transfer" by Liu, Z., Cai, Y., Wang, Y., Nie, Y., Zhang, C., & Xu, Y., et al. (2018). The article is from the Institute of Neuroscience, CAS Center for Excellence in Brain Science and Intelligence Technology, State Key Laboratory of Neuroscience, CAS Key Laboratory of Primate Neurobiology, Chinese Academy of Sciences, Shanghai, China. The article was published on 8 Feb 2018. The article is accompanied by a cover image of the journal Cell. Below the featured article are three sections: Overview, Submitting to NODE, and Using NODE Data. The Overview section describes NODE as an integrated, compatible, comparable, and scalable multi-omics resource platform. The Submitting to NODE section provides instructions on how to submit data to the platform. The Using NODE Data section provides information on how to use the data for research.

NGDC / BMDC Database Tool & Pipeline News About

NODE Project Experiment Sample Run Analysis Statistics Help Search Login Register

Cloning of Macaque Monkeys by Somatic Cell Nuclear Transfer

Liu, Z., Cai, Y., Wang, Y., Nie, Y., Zhang, C., & Xu, Y., et al. (2018)
Institute of Neuroscience, CAS Center for Excellence in Brain Science and Intelligence Technology, State Key Laboratory of Neuroscience, CAS Key Laboratory of Primate Neurobiology, Chinese Academy of Sciences, Shanghai, China
8 Feb 2018

Overview

NODE (The National Omics Data Encyclopedia) provides an integrated, compatible, comparable, and scalable multi-omics resource platform that supports flexible data management and effective data release. NODE uses a hierarchical data architecture to support storage of multi-omics data including sequencing data, MS based proteomics data, MS or NMR based metabolomics data, and fluorescence imaging data.

Launched in early 2017, NODE has collected and published over 900 terabytes of omics data for researchers from China and all over the world in last three years, 22% of which contains multiple omics data. NODE provides functions around the whole life cycle of omics data, from data archive, data requests/responses to data sharing, data analysis, data review and publish.

Submitting to NODE

Making data available to the research community enhances reproducibility and allows for new discovery by comparing data sets.

- 1. Sign up for NODE and Login
- 2. Data Upload/Transfer
- 3. Release of human genetic resources
- 4. Archives (Submit Metadata)
- FAQ

Contact Us

Whether you're looking for answers, would like to solve a problem, or just want to let us know how we did, please contact us by the following Email Address:

Using NODE Data

Use NODE data to validate experimental results, determine variance and open up new avenues of research.

For Public User:

- NODE Search
- Data Download
- Request for Restricted Data

For Data Owner:

- NODE Architecture
- Data/Metadata Management
- Data/Metadata Security
- Share and Review

DDBJ



❏ <https://www.ddbj.nig.ac.jp/index-e.html>

 **DDBJ** [Services](#) [SuperComputer](#) [Statistics](#) [Activities](#) [About Us](#)

DDBJ Web Sites [Terms](#)

 **DDBJ** Bioinformation and DDBJ Center provides sharing and analysis services for data from life science researches and advances science.



Services

Search, analysis, database services of DDBJ Center



Submission

Navigation for how to submit your data



Super Computer

NIG Supercomputer



Statistics

Statistics of DDBJ Center services



Activities

Training sessions and achievements of DDBJ Center



About us

About Bioinformation and DDBJ Center

NEWS

DDBJ Rel. 129.0, DAD Rel. 99.0 Completed
2023/01/17 [Data Release](#) [DDBJ](#)
[DDBJ Center](#)

[more](#)

Genome Sequence Archive



☐ <https://ngdc.cncb.ac.cn/gsa/>



GSA
Genome Sequence Archive

中文 English

[Home](#) [Submit](#) [Browse](#) [Search](#) [Statistics](#) [Support ▾](#) [Login](#) [Register](#)

新闻动态: 整合INSDC SRA库中全部元数据信息, 提供全局检索及热点数据下载服务。

组学原始数据归档库

组学原始数据归档库 (Genome Sequence Archive) 是组学原始数据汇交、存储、管理与共享系统, 是国内首个被国际期刊认可的组学数据发布平台。目前已整合INSDC组学数据, 提供统一检索、数据下载及数据导向服务。



提交数据到GSA



下载GSA数据



浏览已经公开的GSA信息



查找帮助信息和说明文档

热点数据

- [猴痘病毒序列](#)
- [南非Omicron测序数据](#)
- [新型冠状病毒数据](#)
- [人类遗传资源数据](#)
- [多元数据归档库](#)

帮助和支持

如果您在数据上传过程中遇到问题, 或发现任何系统报错, 请随时联系我们。

Email: gsa@big.ac.cn

QQ群: [548170081](#)

非常感谢您对GSA数据库的支持, 欢迎随时提出宝贵意见或建议, 我们将根据您的需求, 不断完善和改进系统功能。

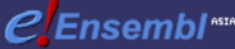
汇交与整合数据

644,547	30,682,422	23,153,989	24,741,900
			
项目	样本	实验	测序反应

Ensembl数据库



☐ <http://asia.ensembl.org/index.html>

 [BLAST/BLAT](#) | [VEP](#) | [Tools](#) | [BioMart](#) | [Downloads](#) | [Help & Docs](#) | [Blog](#)

Login/Register

 Search all species... 

Tools

[All tools](#)

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species ▼ for

Go

e.g. [BRCA2](#) or [rat 5:62797383-63627669](#) or [rs699](#) or [coronary heart disease](#)

All genomes

-- Select a species -- ▼

 **Pig breeds**
Pig reference genome and 12 additional breeds

[View full list of all species](#)

Favourite genomes 

 **Human**
GRCh38.p13
[Still using GRCh37?](#)

 **Mouse**
GRCm39

 **Zebrafish**
GRCz11

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 108 (Oct 2022)

- Changes in the default tracks in the Location view: cDNAs EST cluster (UniGene) CCDS to be removed when MANE Select is available
- RNASeq tracks including data from GeneSWITCH consortium for chicken
- Variation data for crab-eating macaque, pike-perch, prairie vole, Japanese quail and collared flycatcher
- Retirement of postGAP tool

[More release news](#) on our blog

Ensembl Rapid Release

New assemblies with gene and protein annotation every two weeks.

Note: species that already exist on this site will continue to be updated with the full range of annotations.

Go

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

[Rapid Release news](#) on our blog

Ensembl - *Homo sapiens*



http://asia.ensembl.org/Homo_sapiens/Info/Index

BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Login/Register

Search Human...

Human (GRCh38.p13) ▼

Search Human (Homo sapiens)

Search all categories ▼ Search... Go

e.g. [BRCA2](#) or [17:63992802-64038237](#) or [rs699](#) or [osteoarthritis](#)

Genome assembly: GRCh38.p13 (GCA_000001405.28)

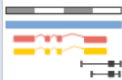
- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh38 coordinates
- Display your data in Ensembl

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart Go



View karyotype



Example region

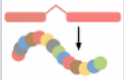
Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download FASTA files for genes, cDNAs, ncRNA, proteins
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
- Update your old Ensembl IDs



Example gene



Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

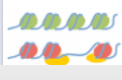
- More about comparative analysis
- Download alignments (EMF)



Example gene tree

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.



Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

- More about variation in Ensembl
- Download all variants (GVF)
- Variant Effect Predictor 



Example variant



Example phenotype

RefSeq数据库



❏ <https://www.ncbi.nlm.nih.gov/refseq/>

 An official website of the United States government [Here's how you know](#) ✓

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

RefSeq

RefSeq: NCBI Reference Sequence Database

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

Using RefSeq	RefSeq Access	RefSeq projects
About RefSeq	Human Genome Resources and Download	Consensus CDS (CCDS)
Human Reference Genome	RefSeq FTP	RefSeq Functional Elements
Prokaryotic RefSeq Genomes	RefSeq genomes FTP	RefSeqGene
FAQ	New RefSeq genomic (last 30 days)	Targeted Loci
NCBI Handbook	New RefSeq transcripts (last 30 days)	Virus Variation
Factsheet	New RefSeq proteins (last 30 days)	RefSeq Select
	Searching for RefSeq records (Queries)	MANE

Refseq数据库



- ❑ 提供高质量无冗余完整的序列信息
- ❑ 包括基因组DNA、RNA转录本以及蛋白质等序列信息
- ❑ 序列文件的标识符：
 - ✿ DNA/RNA序列: NM_004336.5, XM_XXX
 - ✿ 蛋白质序列: NP_004327.1, XP_XXX
 - ✿ 非编码RNA: NR_135914.2
 - ✿ 染色体: NC_XXX

Announcements

January 13, 2023

RefSeq Release 216 is available for FTP

This release includes:

Proteins: 249,868,639

Transcripts: 49,869,497

Organisms: 128,299

Available at: <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>

Documentation: [Release Notes](#)

NCBI Gene



- ❑ 序列从Refseq数据库中得到
- ❑ 详尽的注释信息，包括基因在基因组的定位，基因名称、蛋白质名称，基因结构等

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

Gene Search

Advanced Help

Gene

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

Using Gene

- [Gene Quick Start](#)
- [FAQ](#)
- [Download/FTP](#)
- [RefSeq Mailing List](#)
- [Gene News](#)
- [Factsheet](#)

Gene Tools

- [Submit GeneRIFs](#)
- [Submit Correction](#)
- [Statistics](#)
- [BLAST](#)
- [Genome Workbench](#)

Other Resources

- [OMIM](#)
- [RefSeq](#)
- [RefSeqGene](#)
- [Protein Clusters](#)

<http://www.ncbi.nlm.nih.gov/gene>

UniProt



- ❑ 专家审核的蛋白质序列数据与知识库
- ❑ UniProtKB: UniProt Knowledgebase



The screenshot shows the UniProt website homepage. At the top, the UniProt logo is on the left, and navigation links for BLAST, Align, Peptide search, ID mapping, and SPARQL are in the center. On the right, it says 'Release 2022_05 | Statistics' followed by icons for a printer, a storefront, an envelope, and a help icon. The main heading 'Find your protein' is centered in a large blue box. Below this is a search bar with a dropdown menu set to 'UniProtKB'. To the right of the search bar are links for 'Advanced', 'List', and a 'Search' button. Below the search bar, example search terms are listed: 'Examples: Insulin, APP, Human, P05067, organism_id:9606'. At the bottom of the blue box, a statement reads: 'UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. [Cite UniProt](#)'. On the right side of the blue box, there are vertical buttons for 'Feedback' and 'Help'. At the very bottom, a light blue box contains the URL 'http://www.uniprot.org/'.

UniProtKB ▾ | Advanced | List Search

Examples: Insulin, APP, Human, P05067, organism_id:9606

UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. [Cite UniProt](#)

<http://www.uniprot.org/>

UniProt数据库



□ UniProt两个字库：

- ✿ Swiss-Prot: Reviewed

- ✿ TrEMBL: Unreviewed

□ 蛋白质标识符命名规则（先后顺序）

- ✿ O>P>Q>A>B...

- ✿ 标识符越靠前，注释信息越丰富

