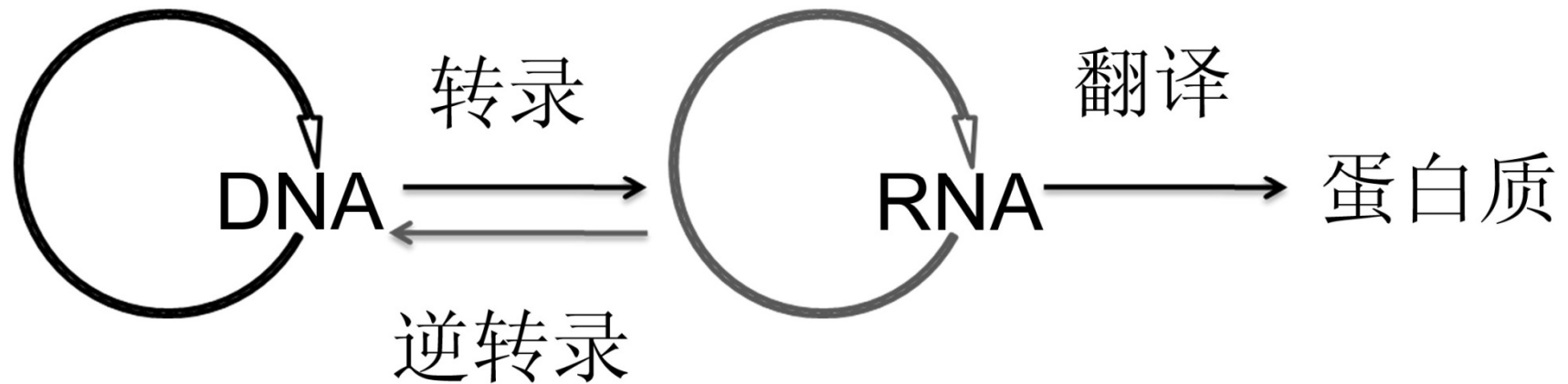




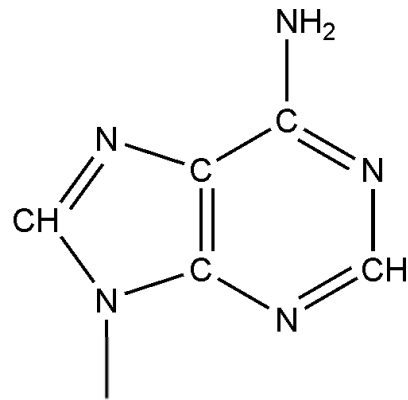
生物信息学

第二章 生物序列数据获取和检索

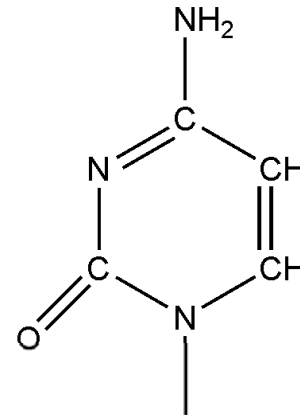
中心法则



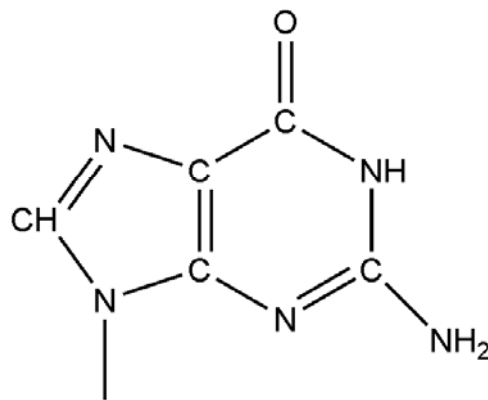
DNA结构：碱基/核苷



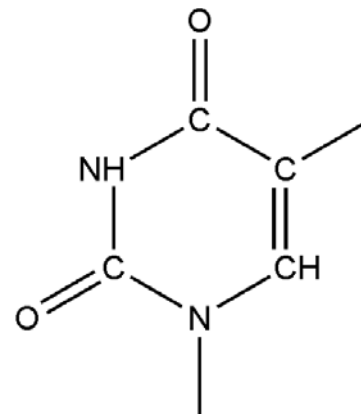
Adenine (A)



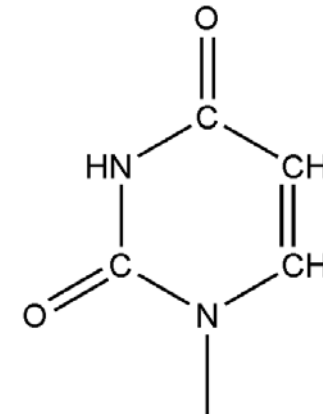
Cytosine (C)



Guanine (G)



Thymine (T)



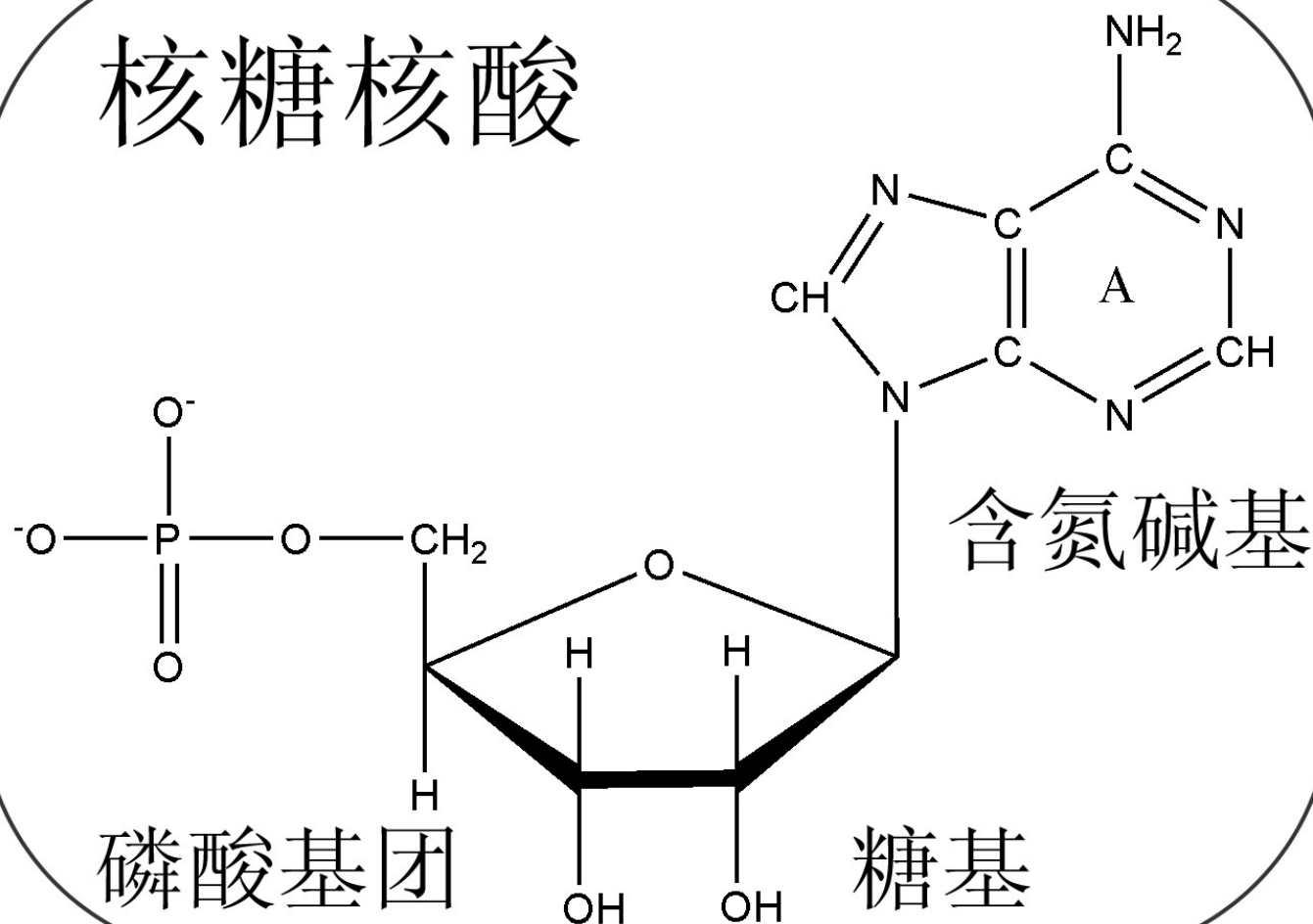
Uracil (U)

核糖核酸



Ribonucleotide

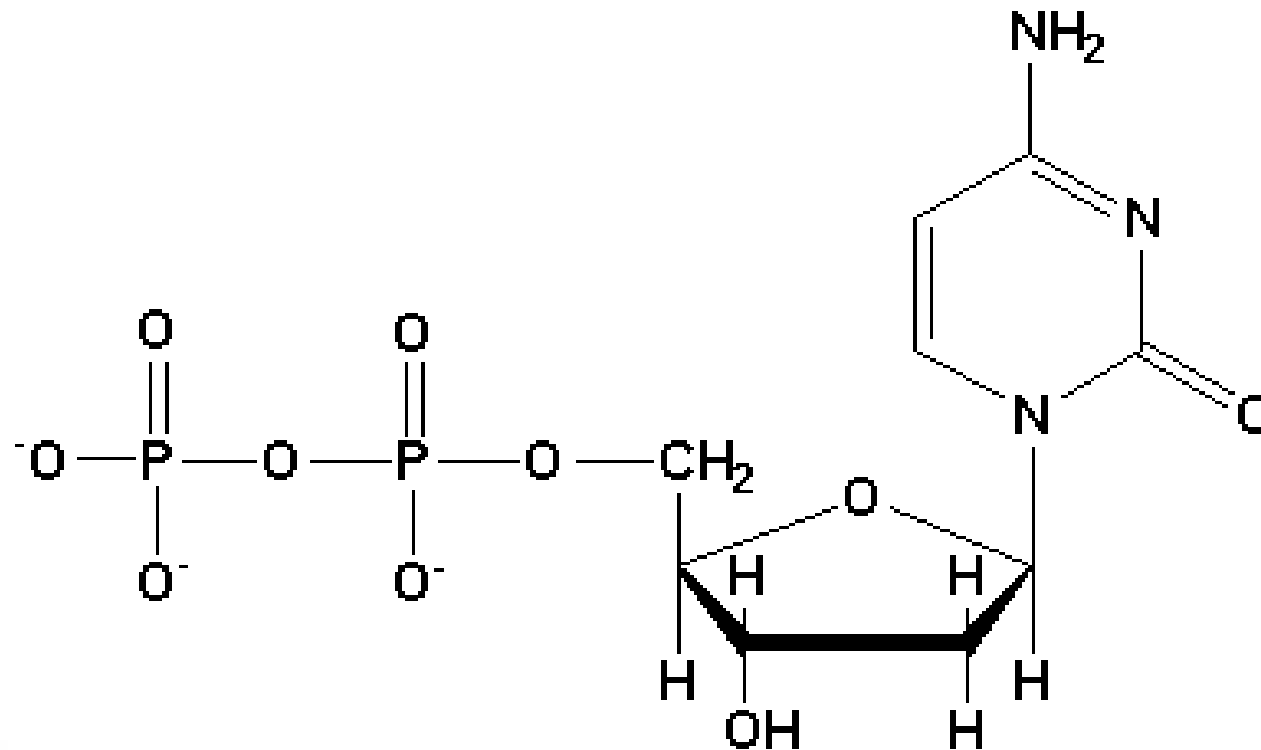
核糖核酸



脱氧核糖核酸



Deoxyribonucleotide



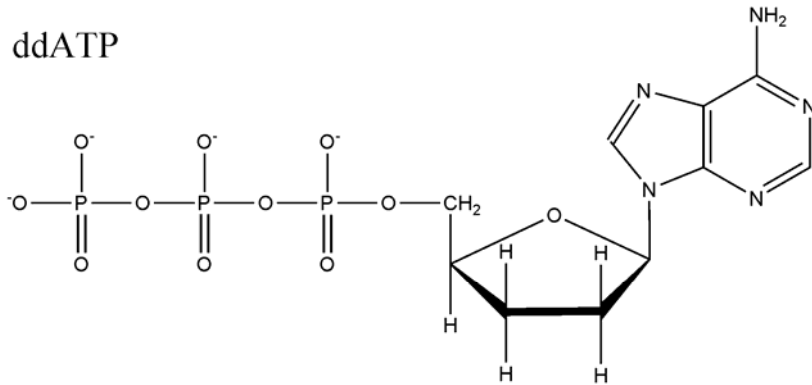
Deoxycytidine diphosphate (dCDP)

双脱氧核糖核苷酸

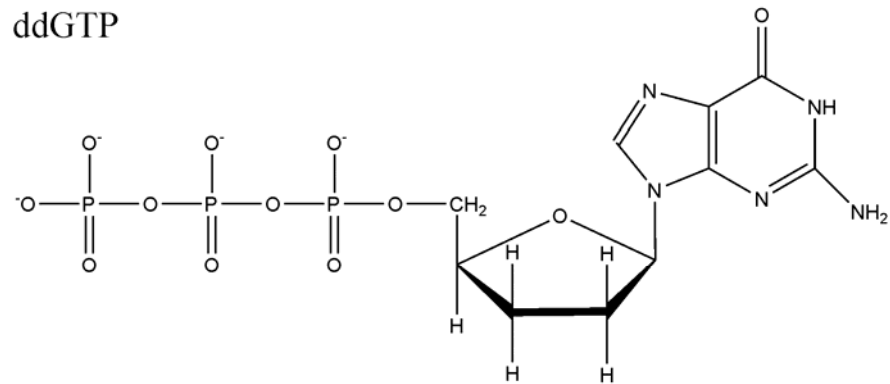


Dideoxynucleotide

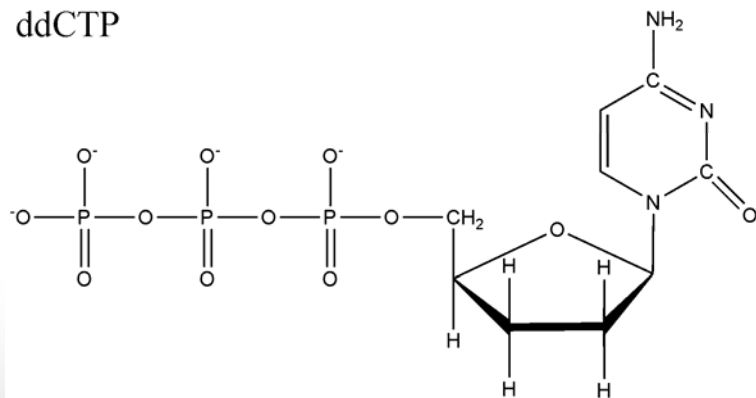
ddATP



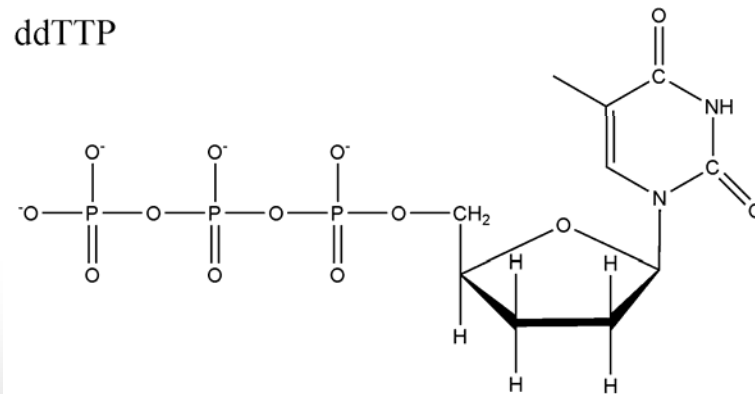
ddGTP



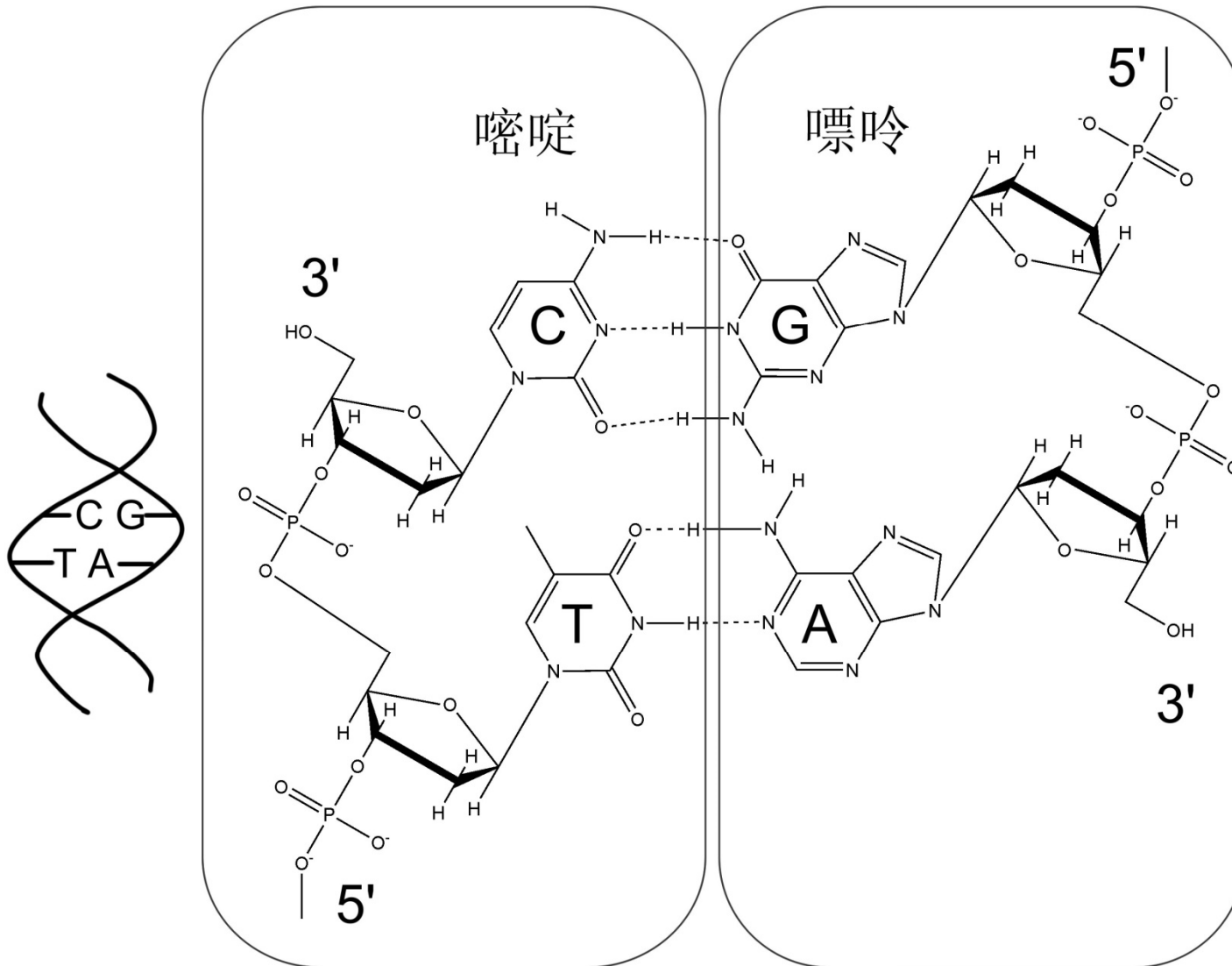
ddCTP



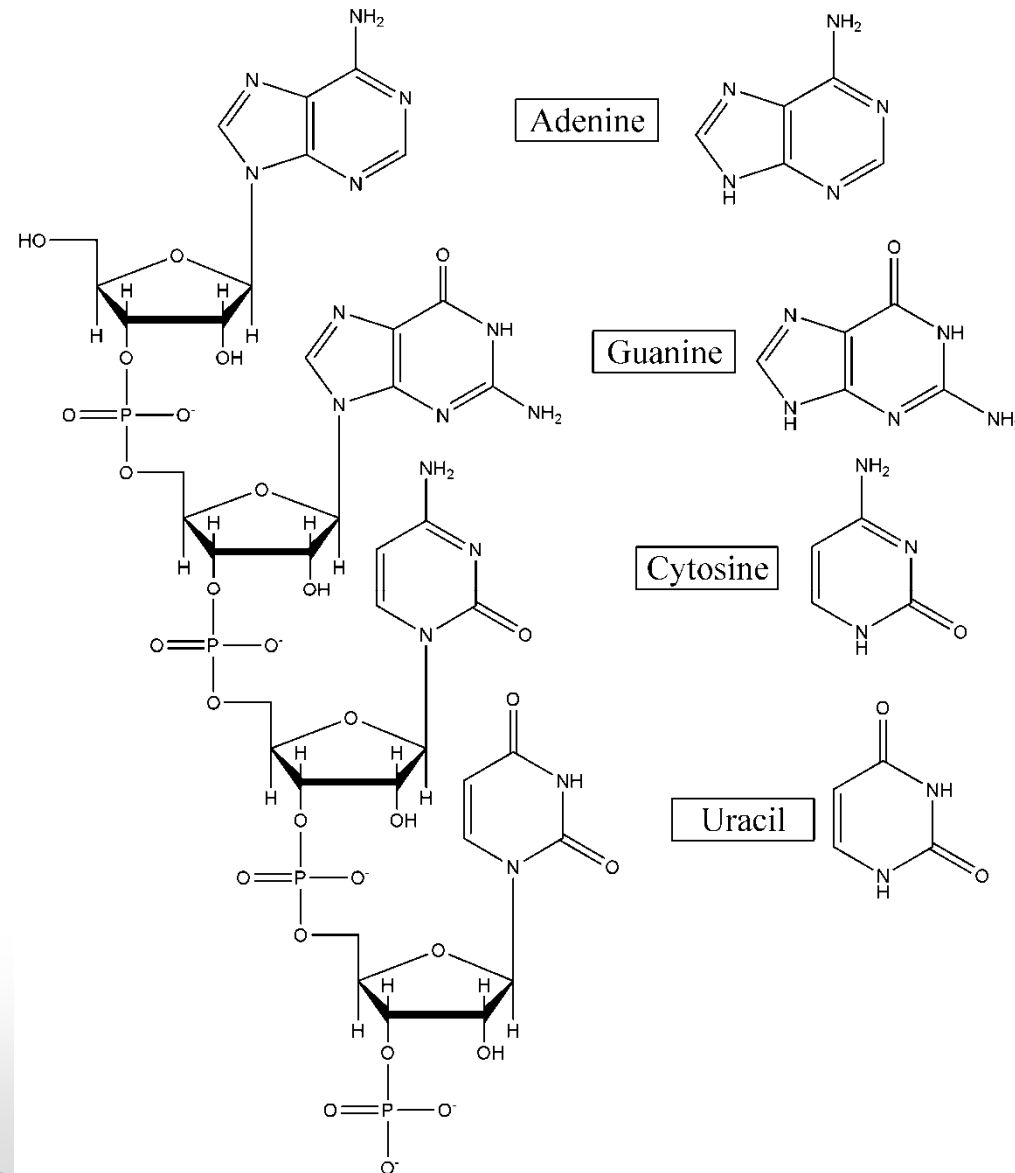
ddTTP



DNA的结构



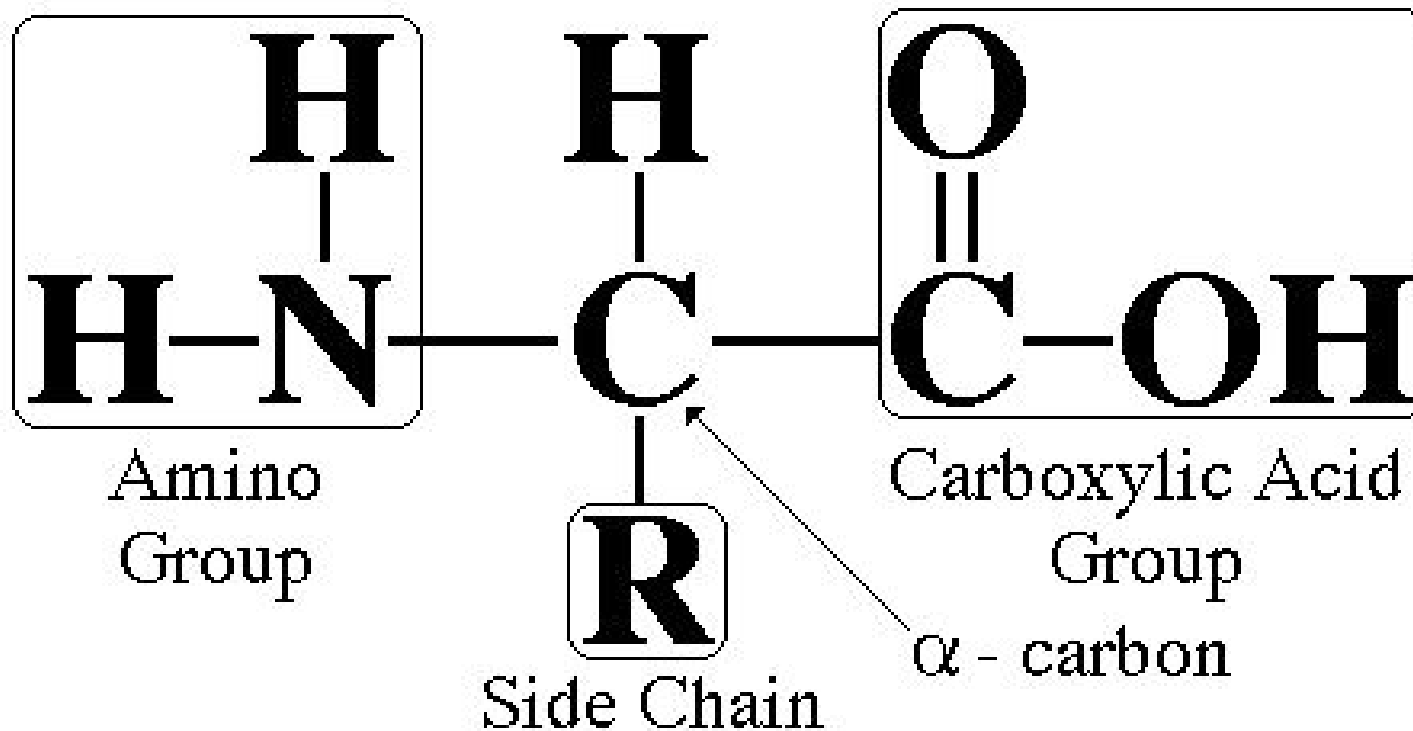
RNA的结构



氨基酸的结构



Amino Acid Structure

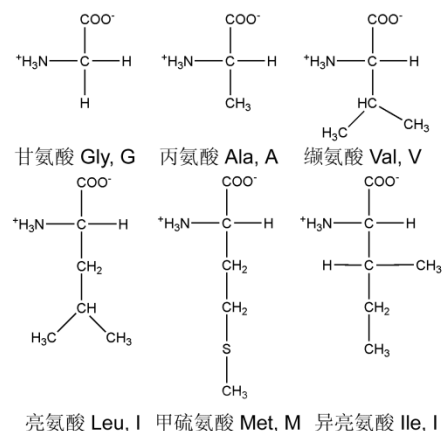


氨基酸的性质及分类

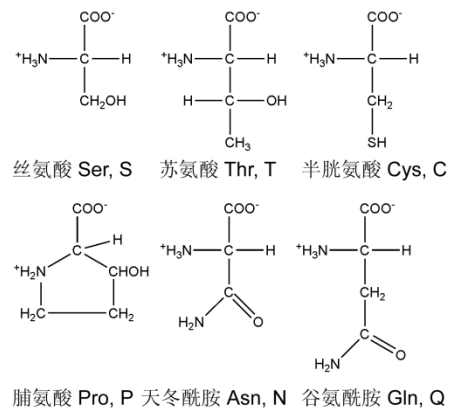


20种常见的蛋白质氨基酸

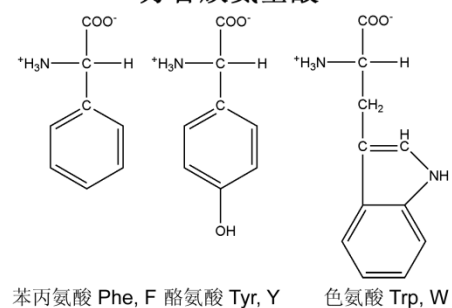
非极性脂肪酸氨基酸



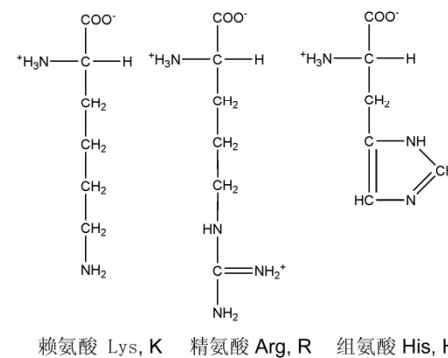
R基不带电荷的氨基酸



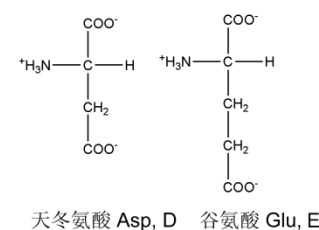
芳香族氨基酸



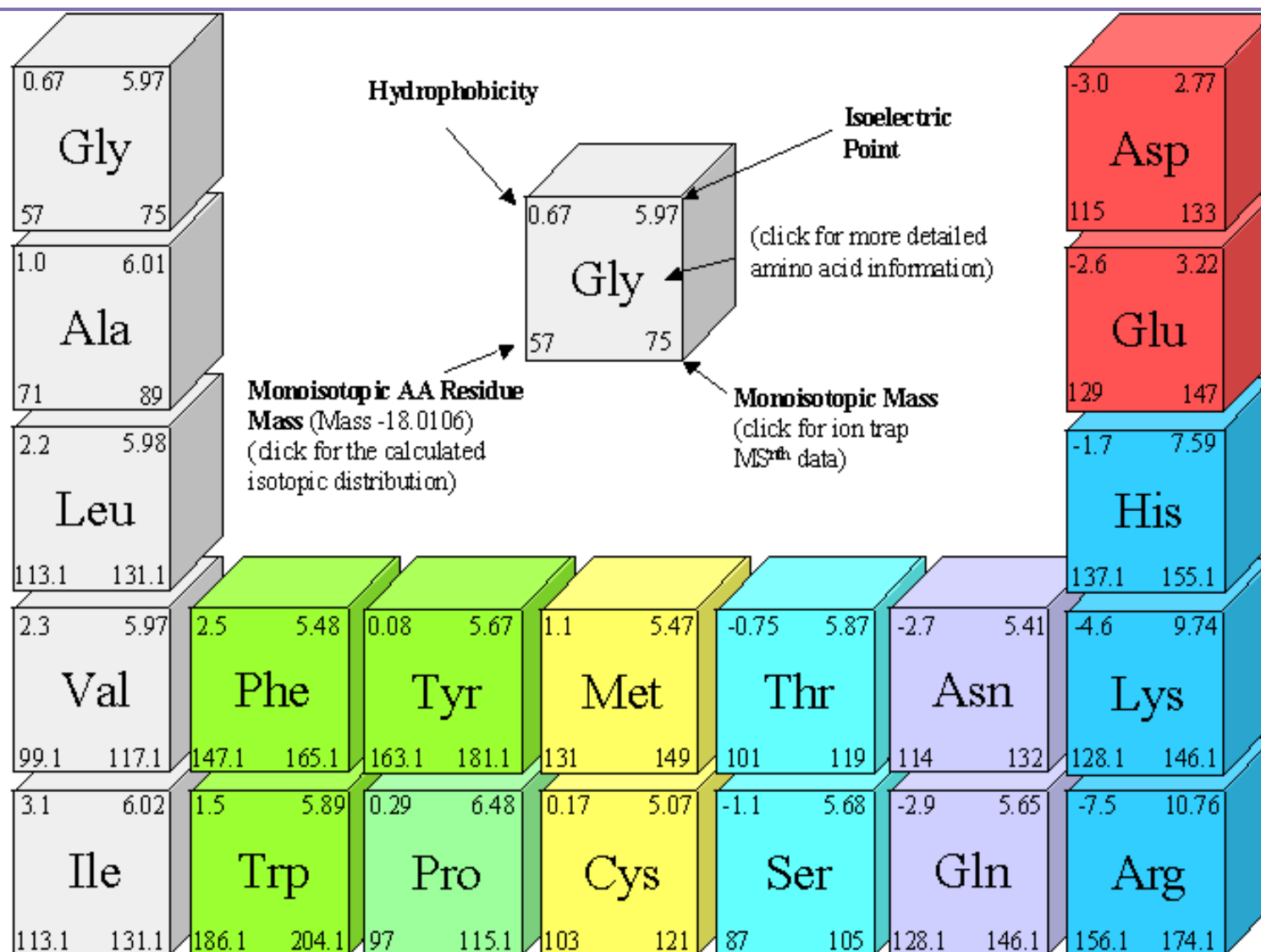
R基带正电荷的氨基酸



R基带负电荷的氨基酸



氨基酸：周期表



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

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



- 噬菌体 λ DNA
- left half right half
- sus A B λ (C) A
- 3' HO P G (T) OH 3' P
- heavy strand light strand
-
- (T) (C)

Lisa A. Onaga



1954年



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



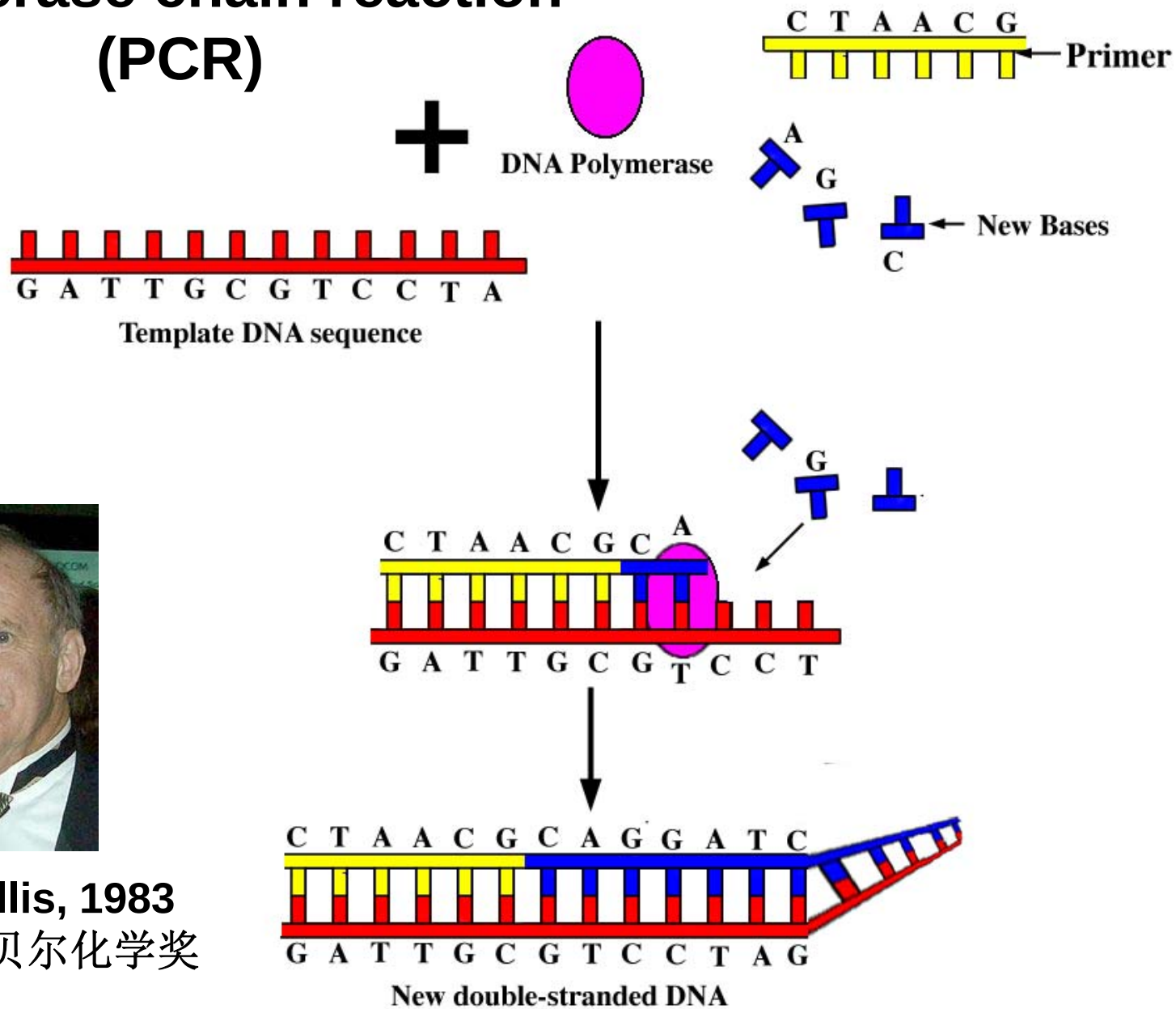
Walter Gilbert



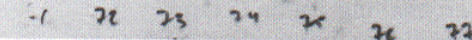
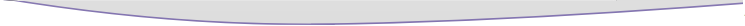
Frederick Sanger



Polymerase chain reaction (PCR)



Kary Mullis, 1983
1993年诺贝尔化学奖



自动测序

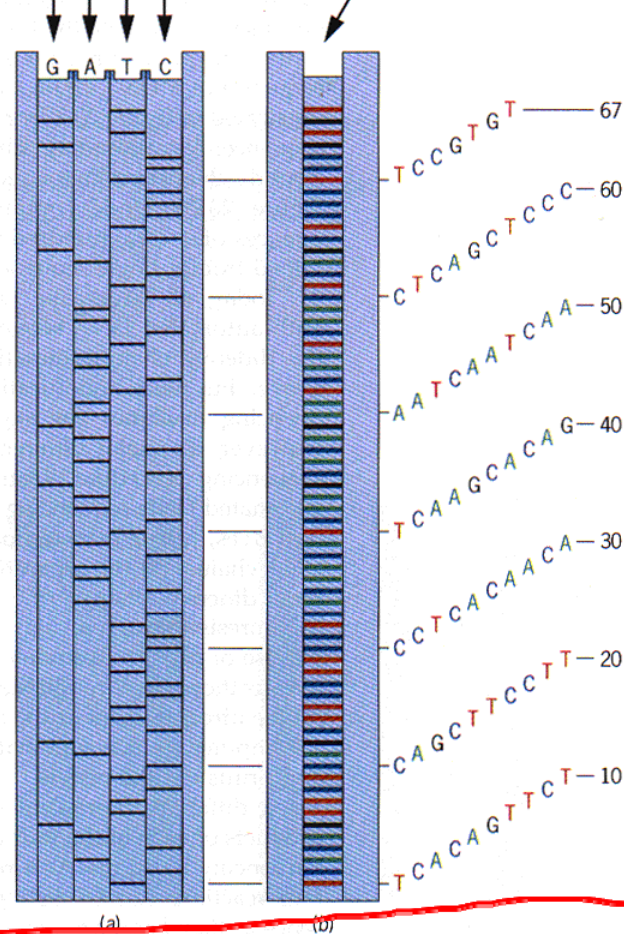


- ❑ “假如你真想改变一门学科，发明一些新的技术，这将使你超越以前人们曾经所看到的”
- ❑ 1980s, Leroy Hood发展了为4种终止核苷酸碱基标定不同荧光颜色的方法，改进了测序技术
- ❑ 4种碱基可在同一个反应中测序，并在一个条带中排列显示
- ❑ Hood: 利用计算机收集整理数据的先驱

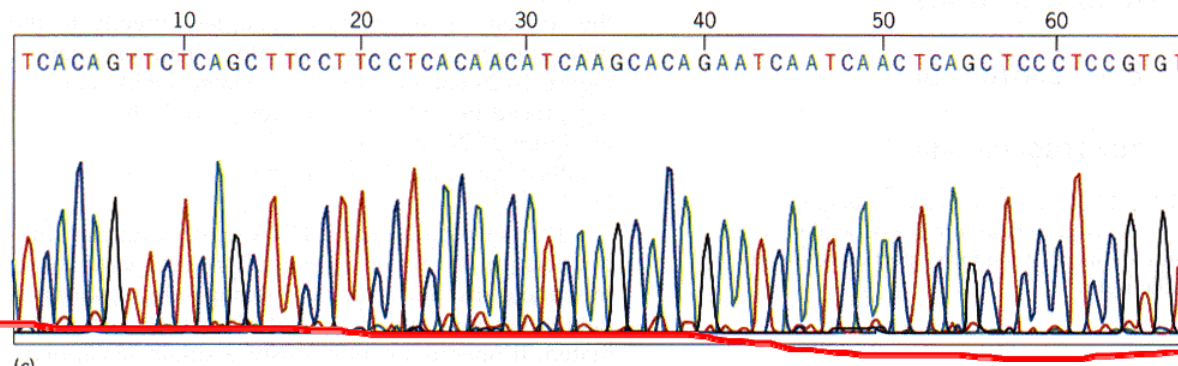


Each dideoxy chain-terminator reaction is loaded into a separate sample well.

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

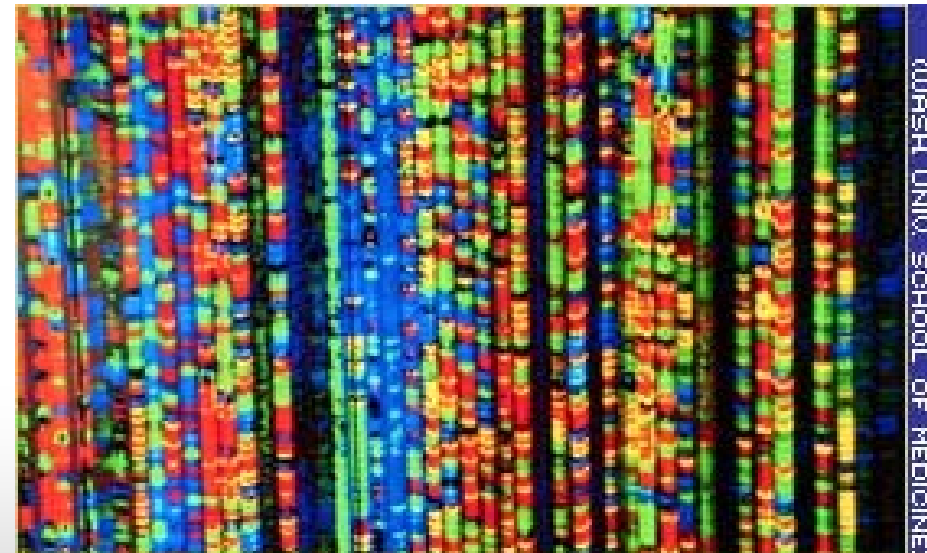
✿ Applied Biosystems, Inc (ABI)

✿ 1987, ABI 370

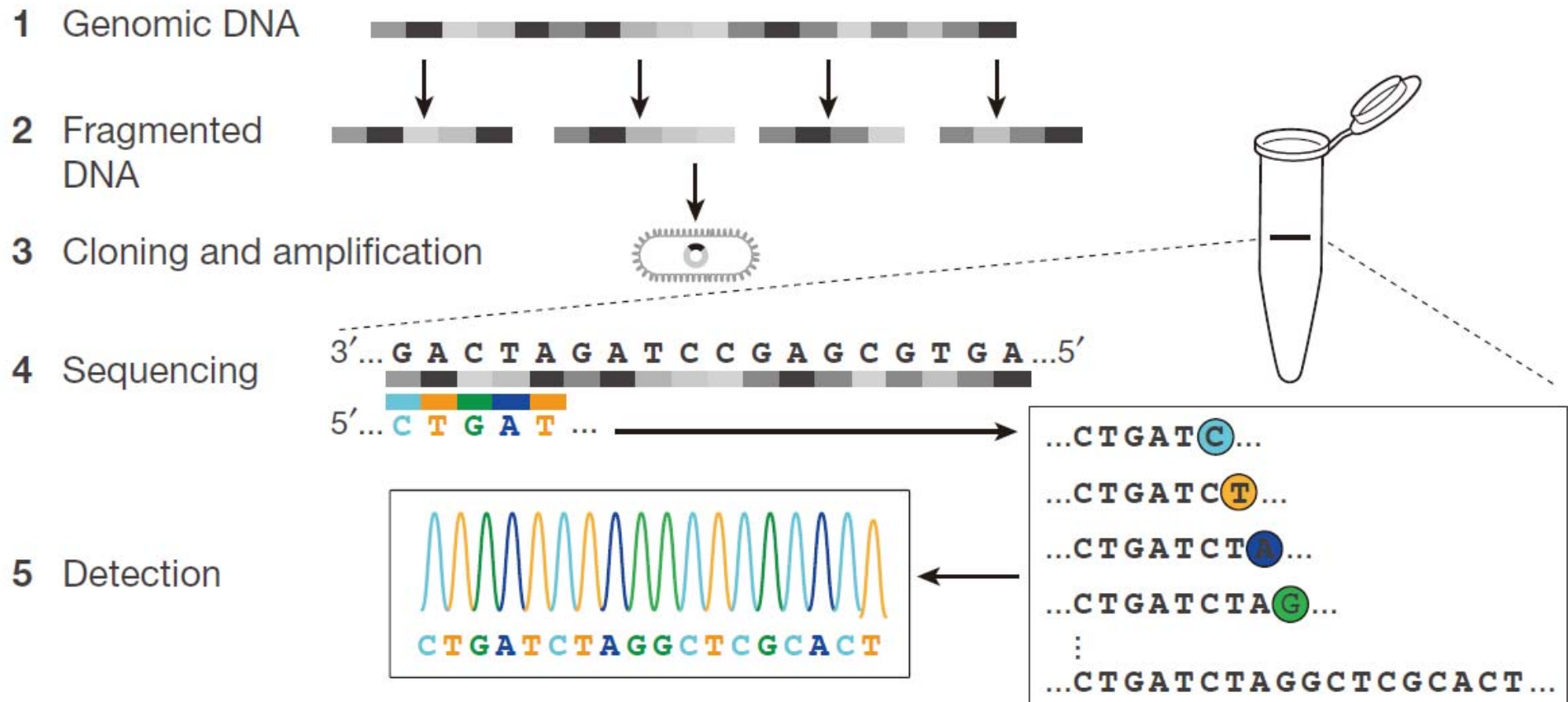
✿ 利用四种颜色，可直接阅读四种碱基

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

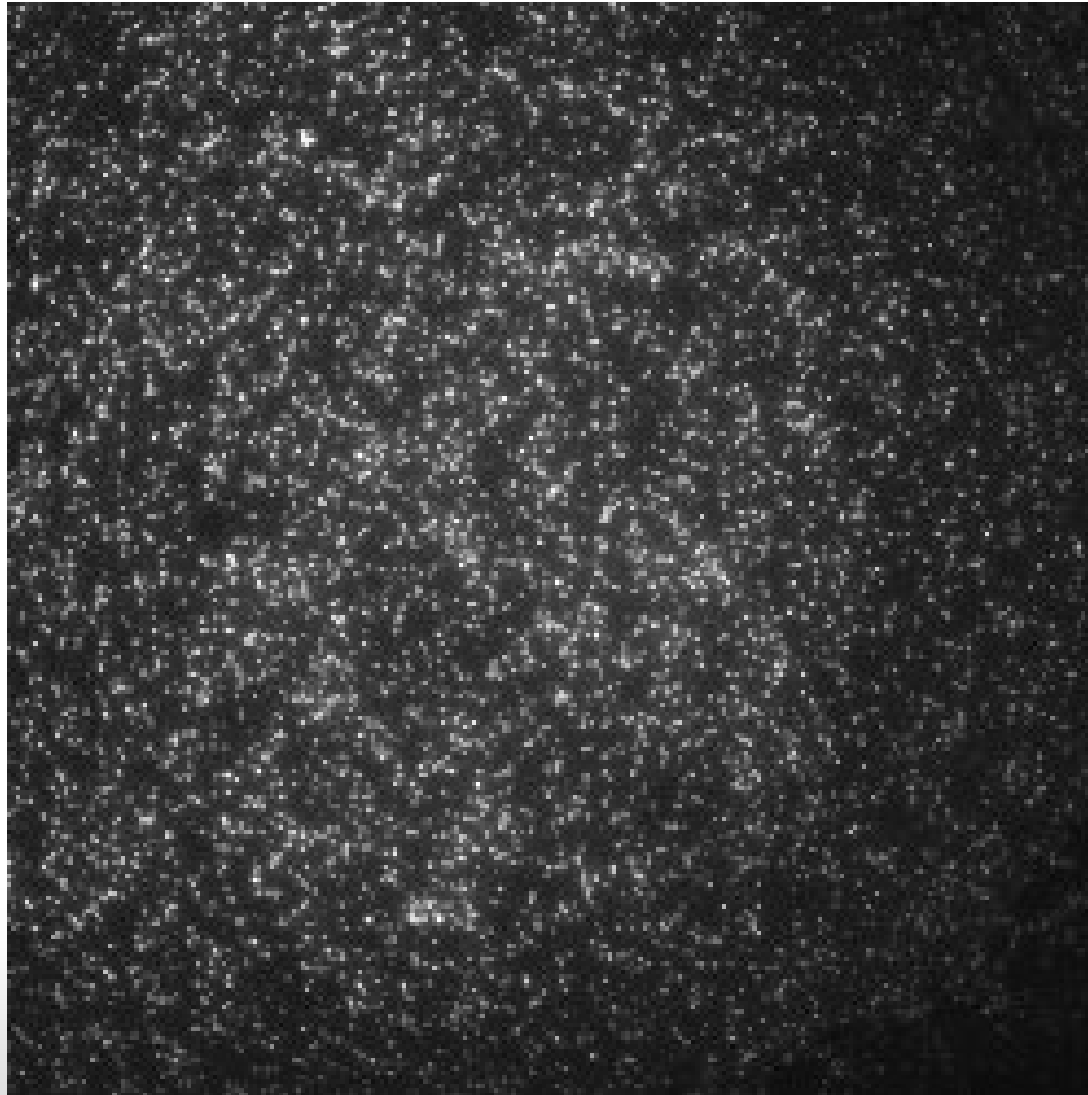


□ 每个系统的测序方式不同，但基本原理相似：

- ✿ 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）

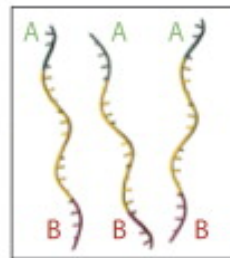
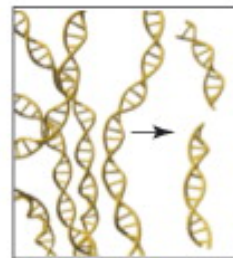
454测序原理

- ❑ 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）

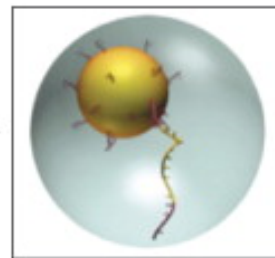


Roche (454) GSFLX Workflow:

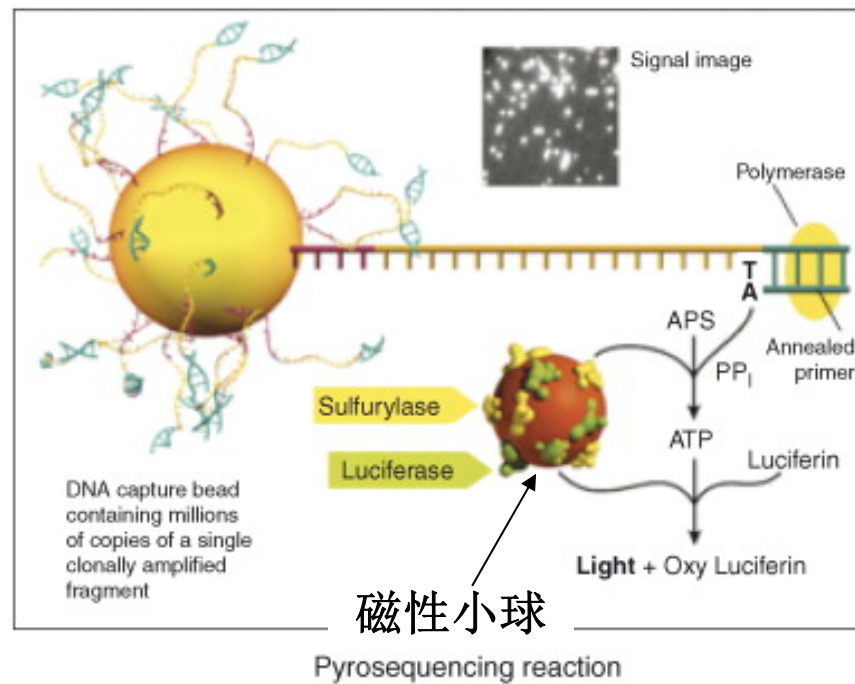
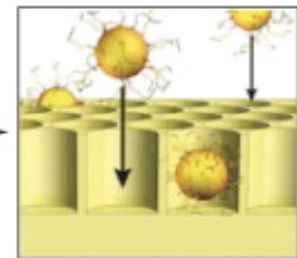
Library construction



Emulsion PCR



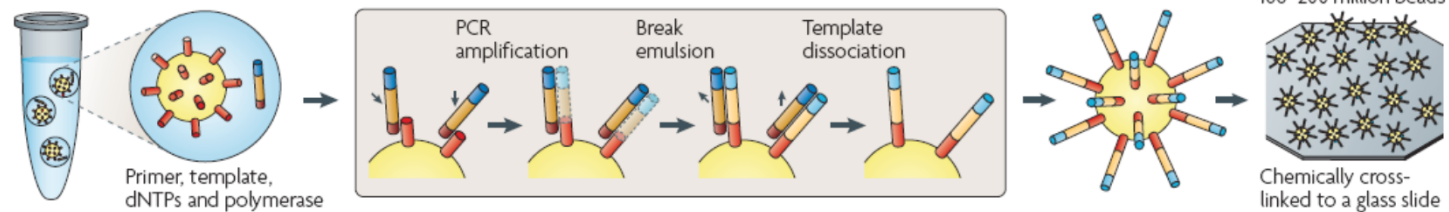
PTP loading



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.

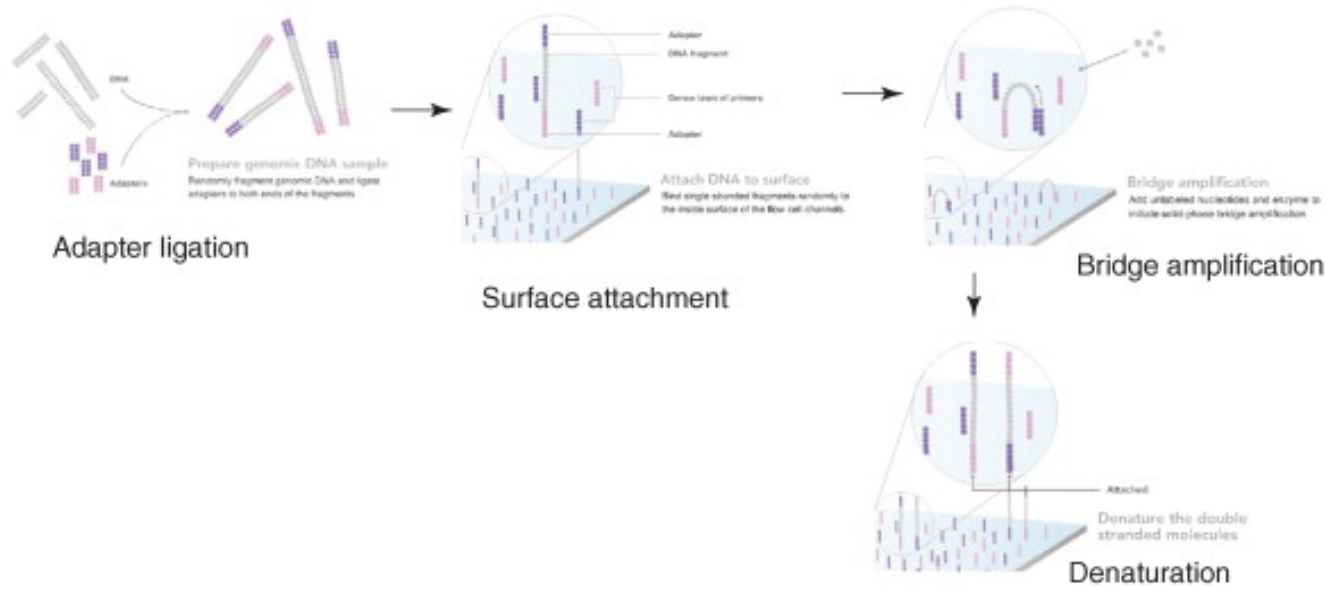
Illumina Genome Analyzer



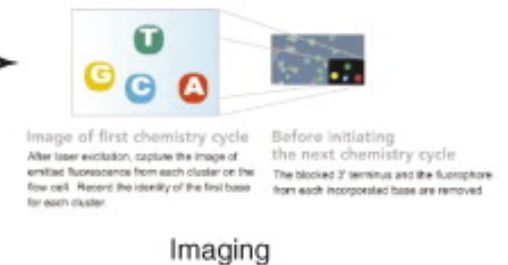
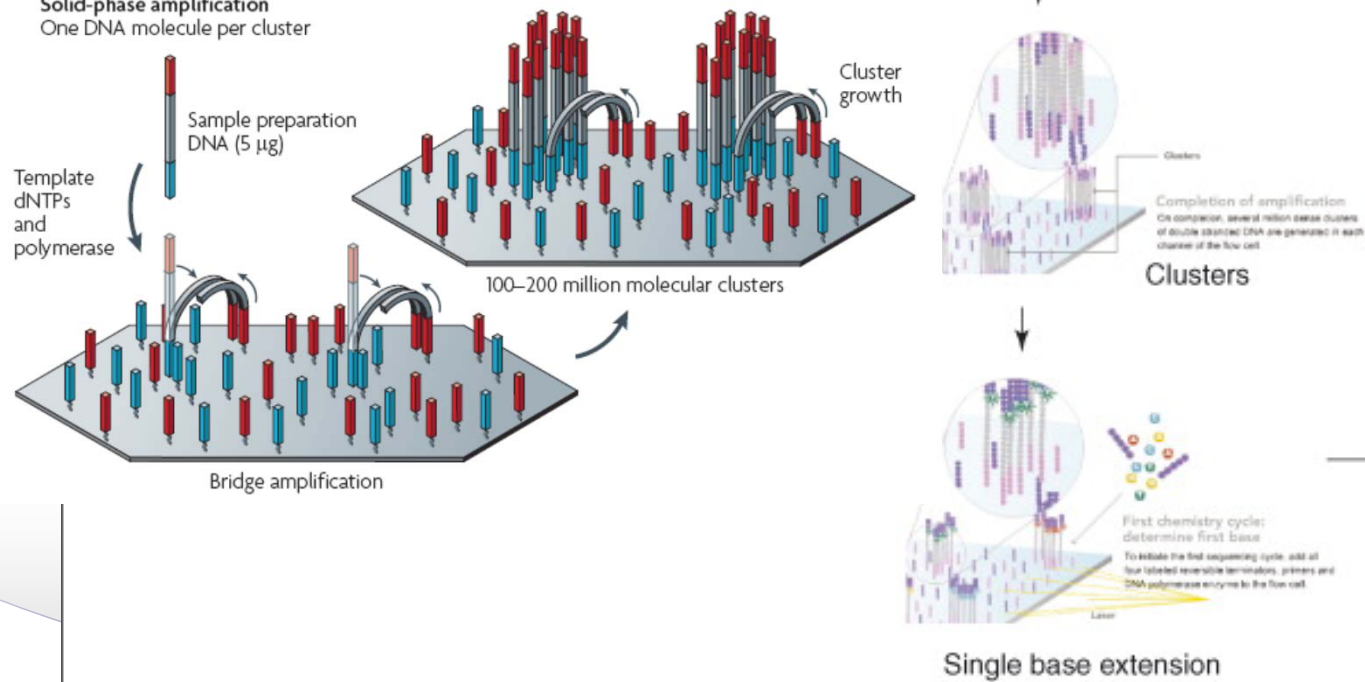
- ❑ 最早由Solexa开发设计，现在是Illumina的子公司
- ❑ 2006年实现商业化
- ❑ 每轮实验：7个样本/3日
- ❑ 低错误率，主要是碱基类型错误，有少量的插入/缺失
- ❑ Sequencing by synthesis



Illumina Genome Analyzer Workflow



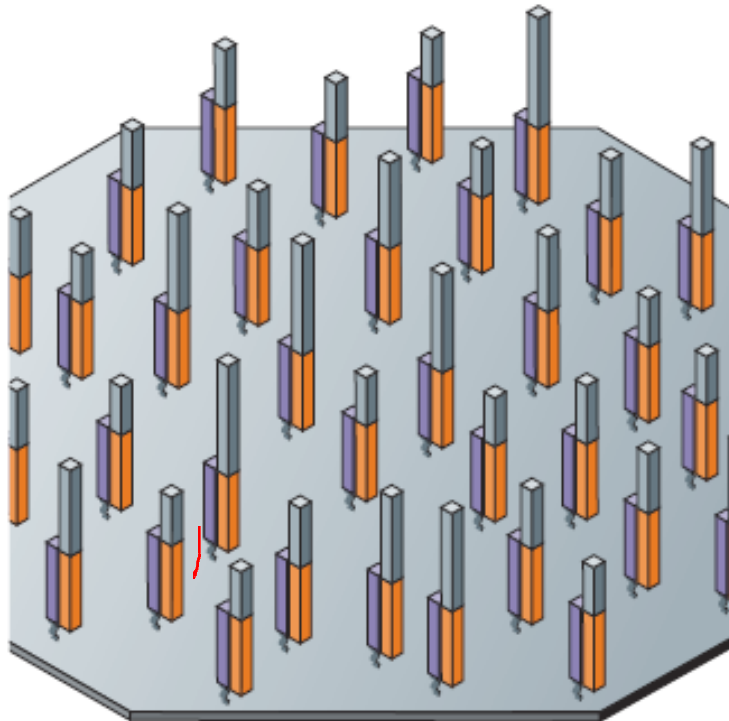
b Illumina/Solexa Solid-phase amplification One DNA molecule per cluster



Helicos BioSciences



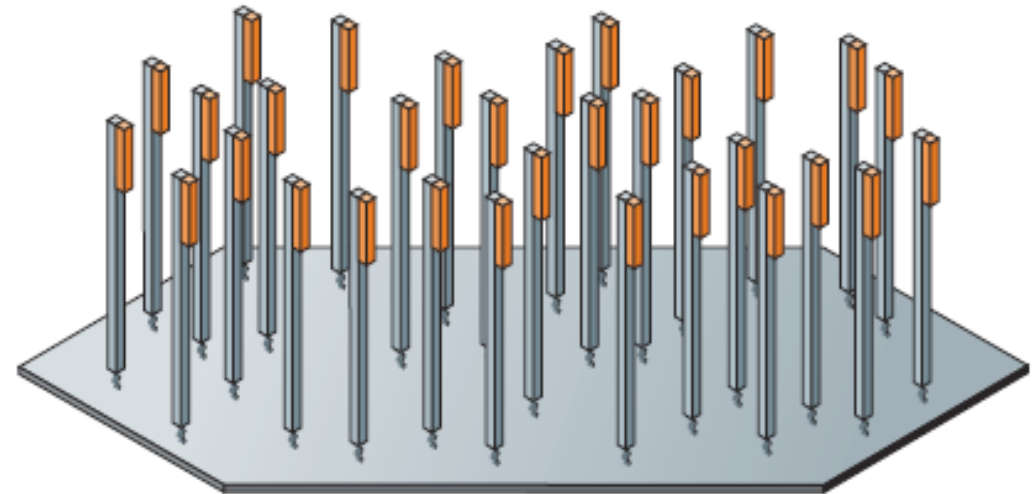
c Helicos BioSciences: one-pass sequencing
Single molecule: primer immobilized



Billions of primed, single-molecule templates

Adaptor固定，然后接模板

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



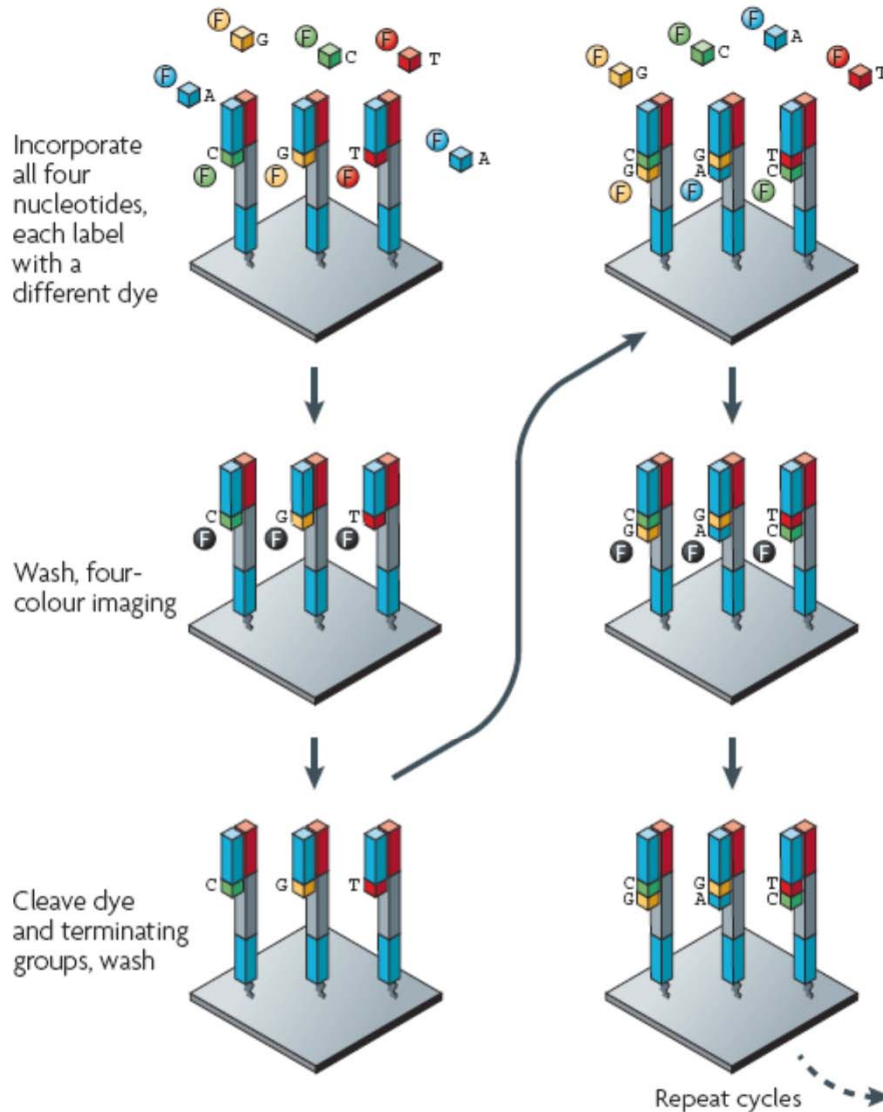
Billions of primed, single-molecule templates

模板先固定，然后接**Adaptor**

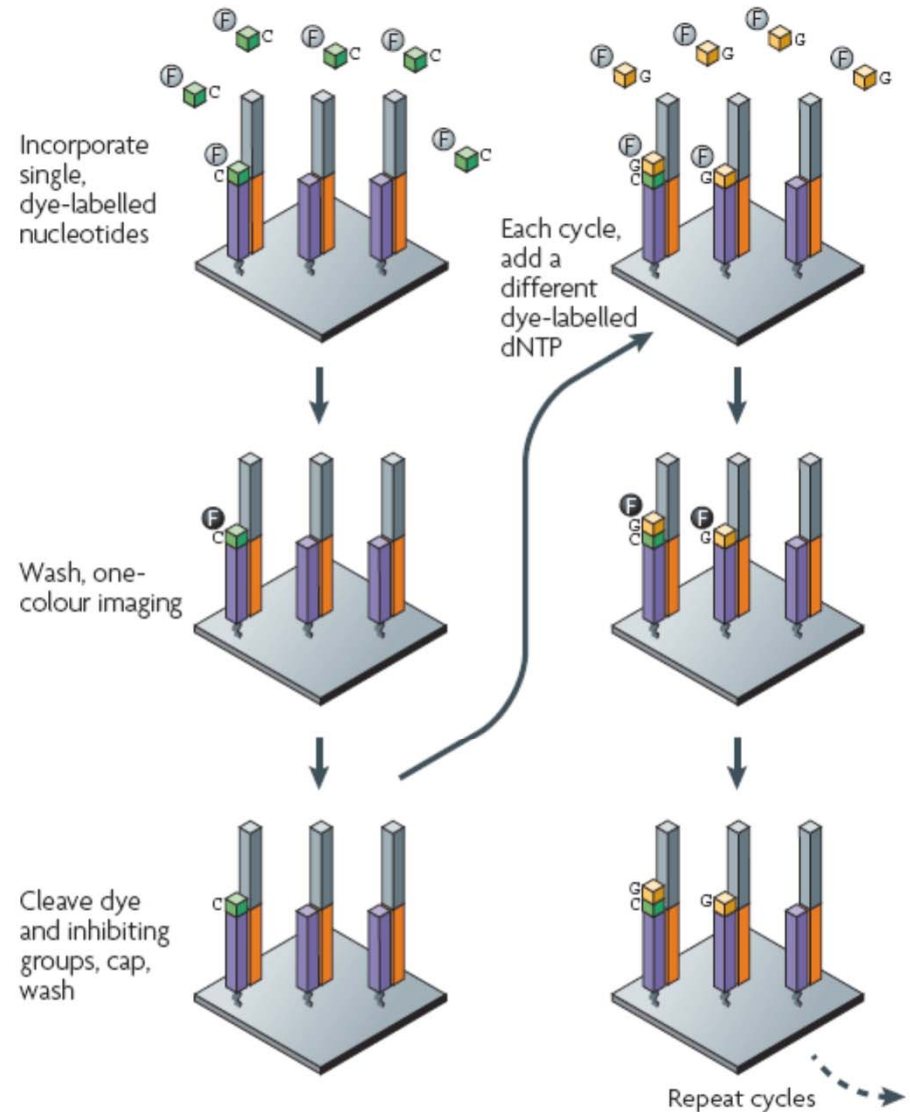
Illumina/Solexa vs. Helicos BioSciences



a Illumina/Solexa — Reversible terminators



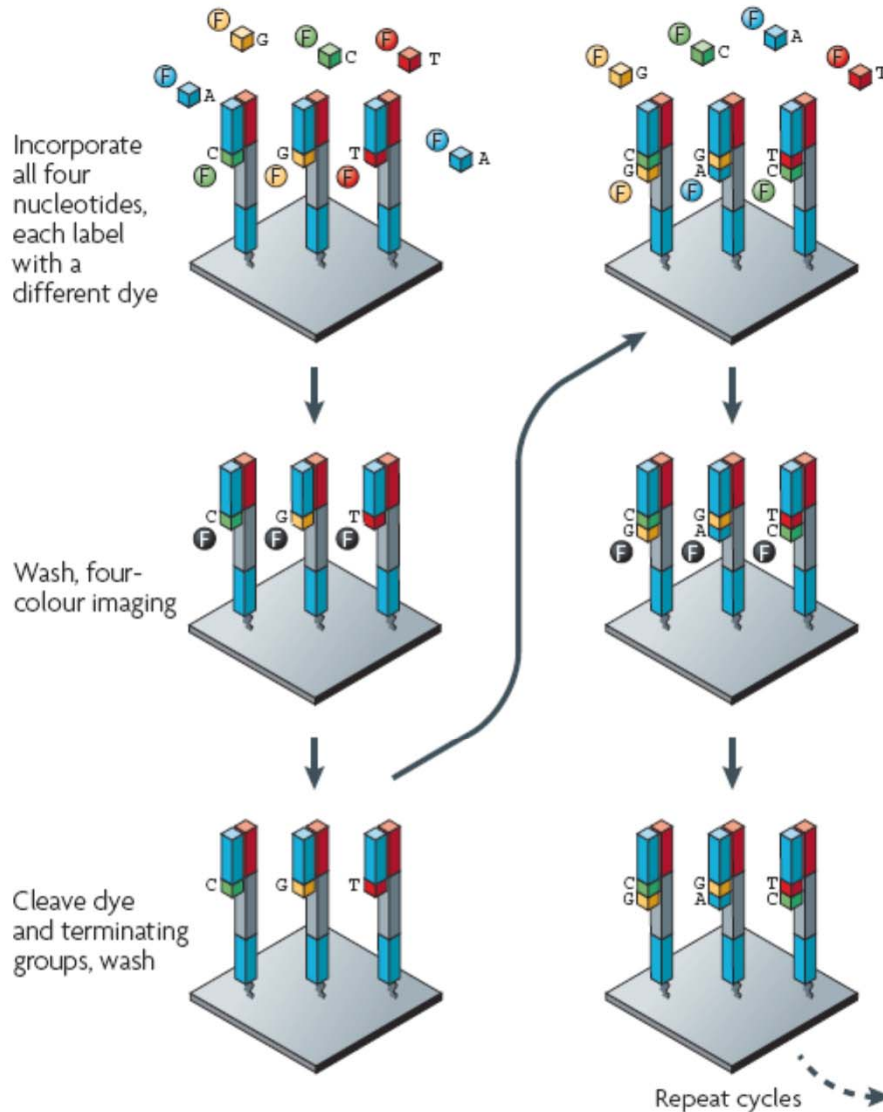
c Helicos BioSciences — Reversible terminators



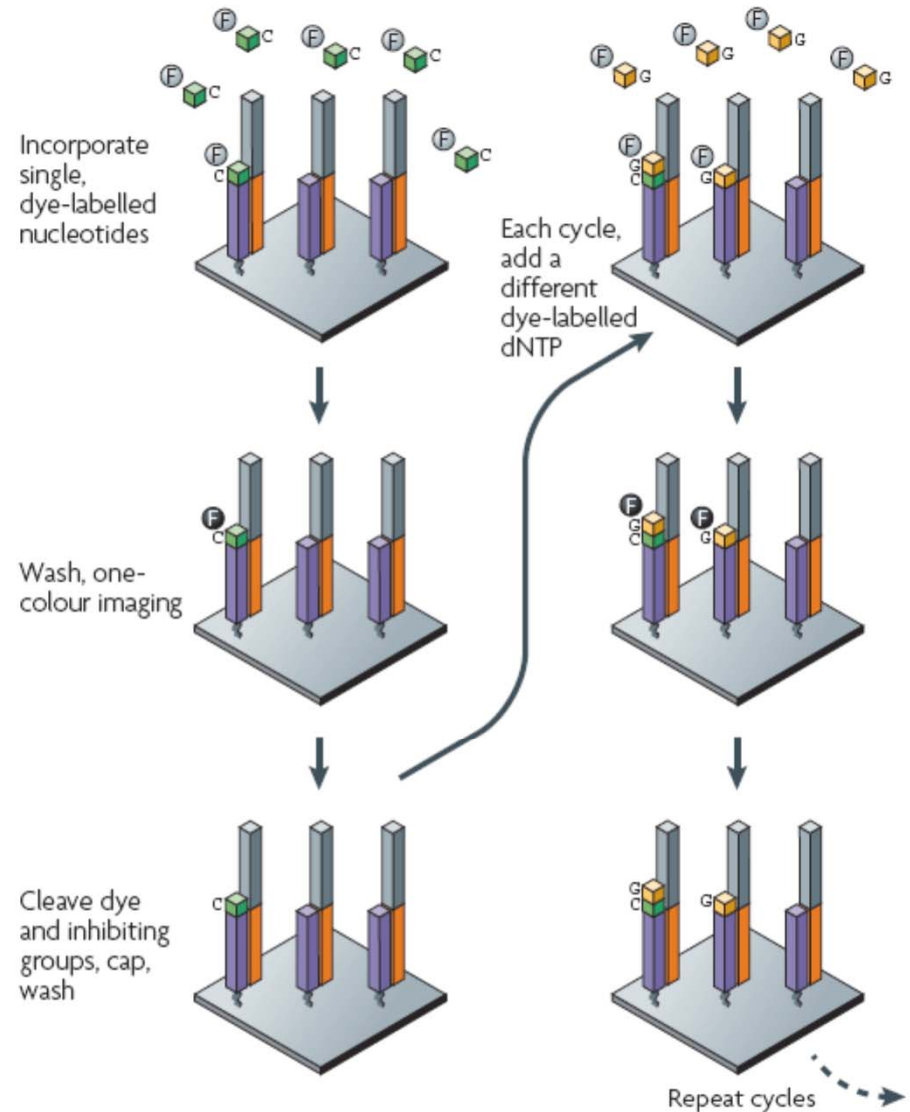
Illumina/Solexa vs. Helicos BioSciences



a Illumina/Solexa — Reversible terminators



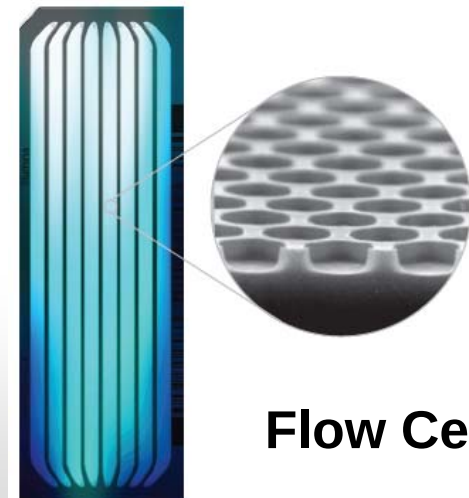
c Helicos BioSciences — Reversible terminators



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	<u>1.2 Gb</u>	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

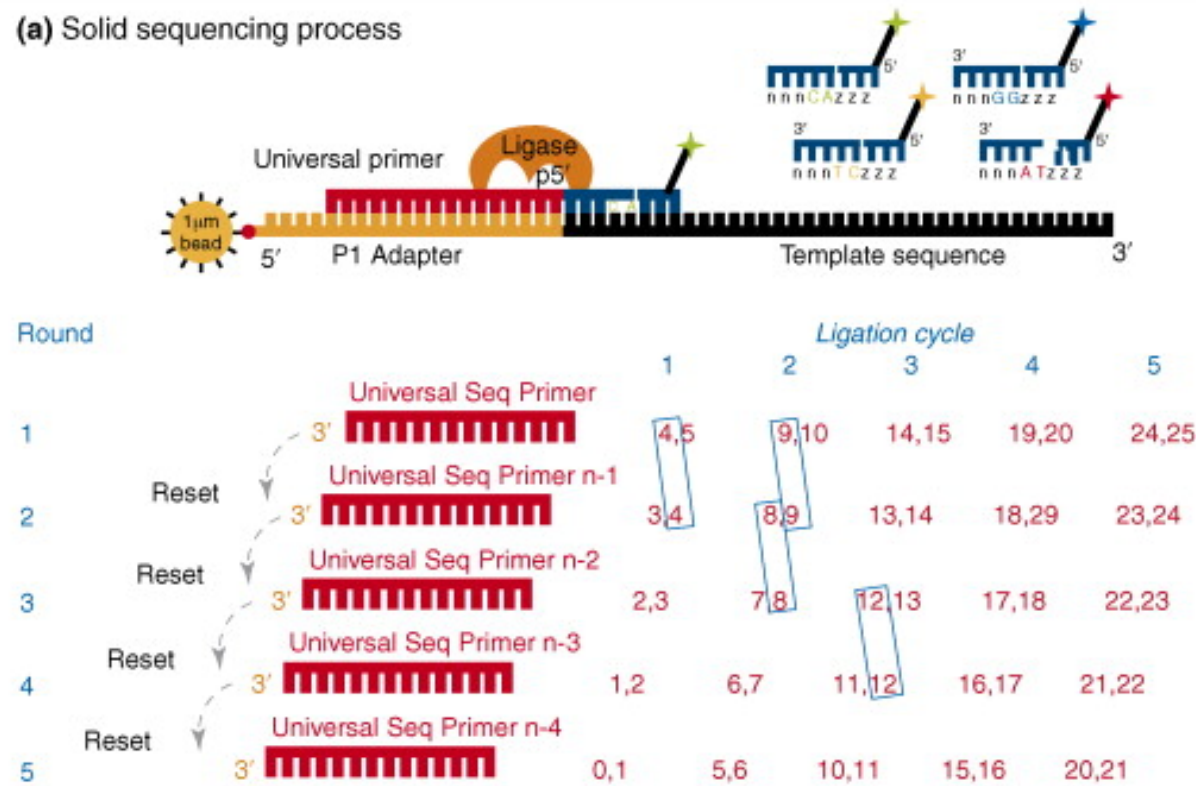
ABI-SOLID



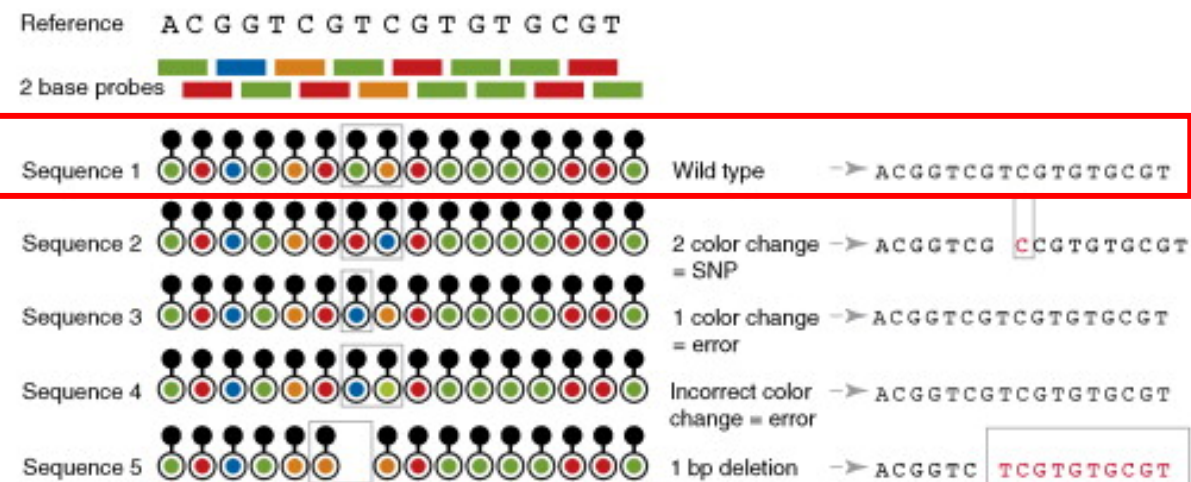
- ❑ 2007年实现商业化
- ❑ 每周产生20GB的数据
- ❑ 每轮: 6GB
- ❑ Sequencing by ligation
- ❑ “2 based encoding”
- ❑ “color-space” data



(a) Solid sequencing process



(b) Principles of two base encoding



ABI-SOLID

Two-base encoding: each target nucleotide is interrogated twice

2nd base

1st base

Template sequence

ATACAAGA
CGCACCTC
GCGTGGAG
TAGTTCCT

TCGGATTCAGCCTGCTGCTCTATCA
Δ

Bioinform

[illegible]

1,2-probes

x, y Interrogation bases
n Degenerate bases
z Universal bases

3' x y n n n n z z z z 5'

3' x y n n n n z z z z 5'

3' x y n n n n z z z z 5'

3' x y n n n n z z z z 5'

Repeat ligation cycles

Ligation cycle 1 2 3 4 5 6 7... (n cycles)

AT TT CT GT TT CA GC
TA AA GA CA AA GT CG

Reset primer ($n - 1$), repeat ligation cycles

Primer round 2

1 base shift

Universal seq primer ($n - 1$)

3' TGA CG AC TA GG GC 3'

3' AACTG AT CC CG 3'

Reset primer three more times

Revolocity: Complete Genomics

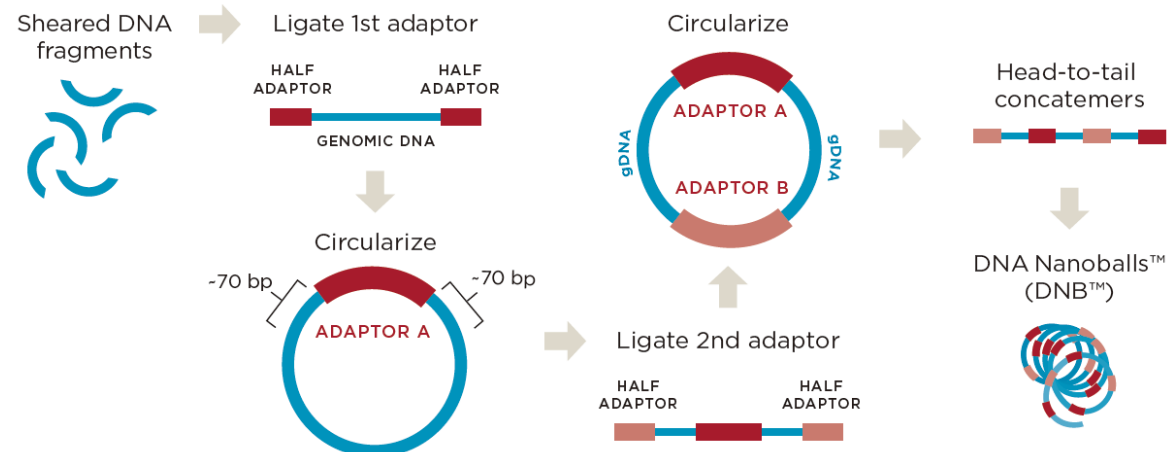


IMPACT OF CORRELATION ANALYSIS ON FALSE-POSITIVE RATE

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

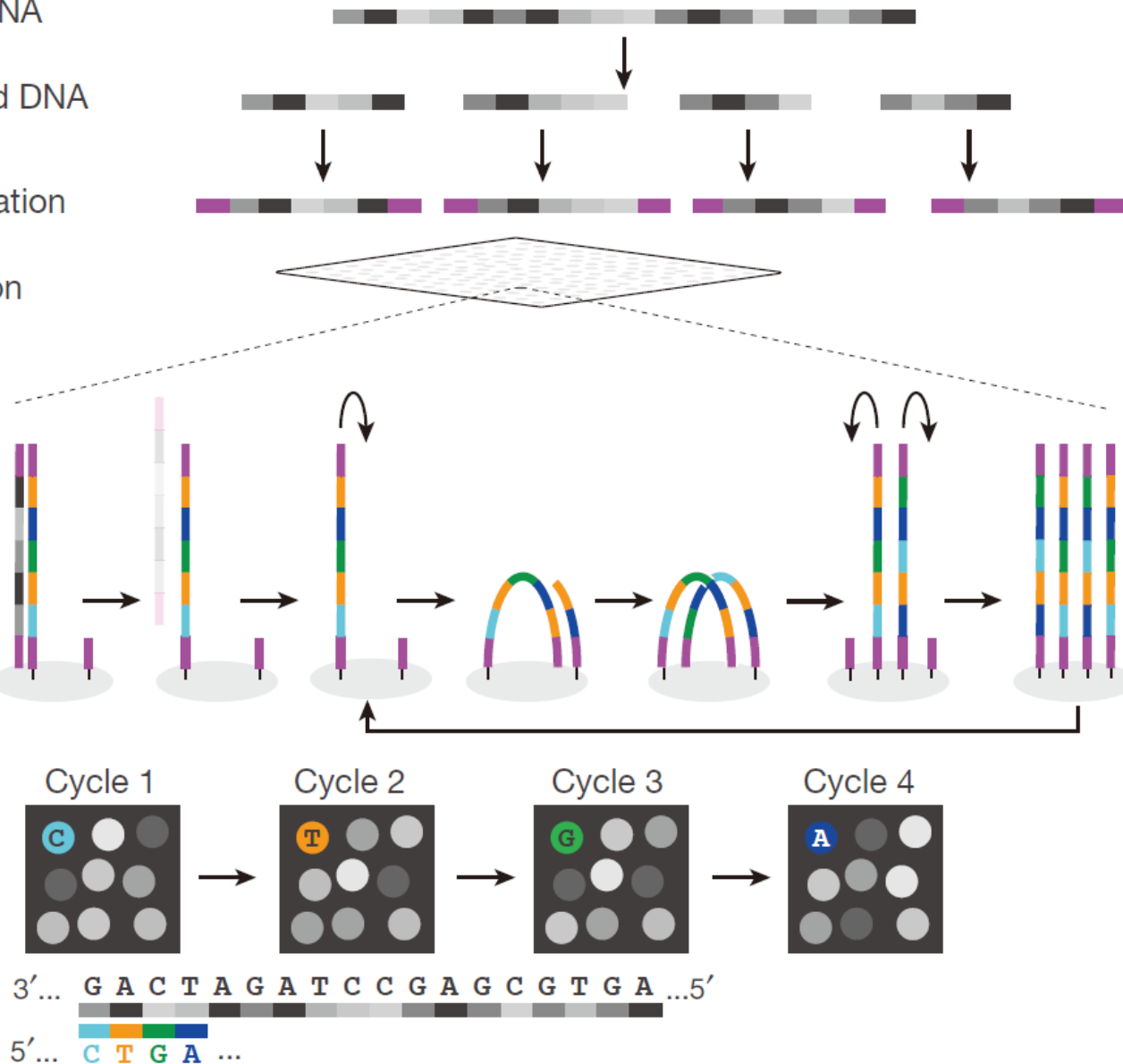
2 Fragmented DNA

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

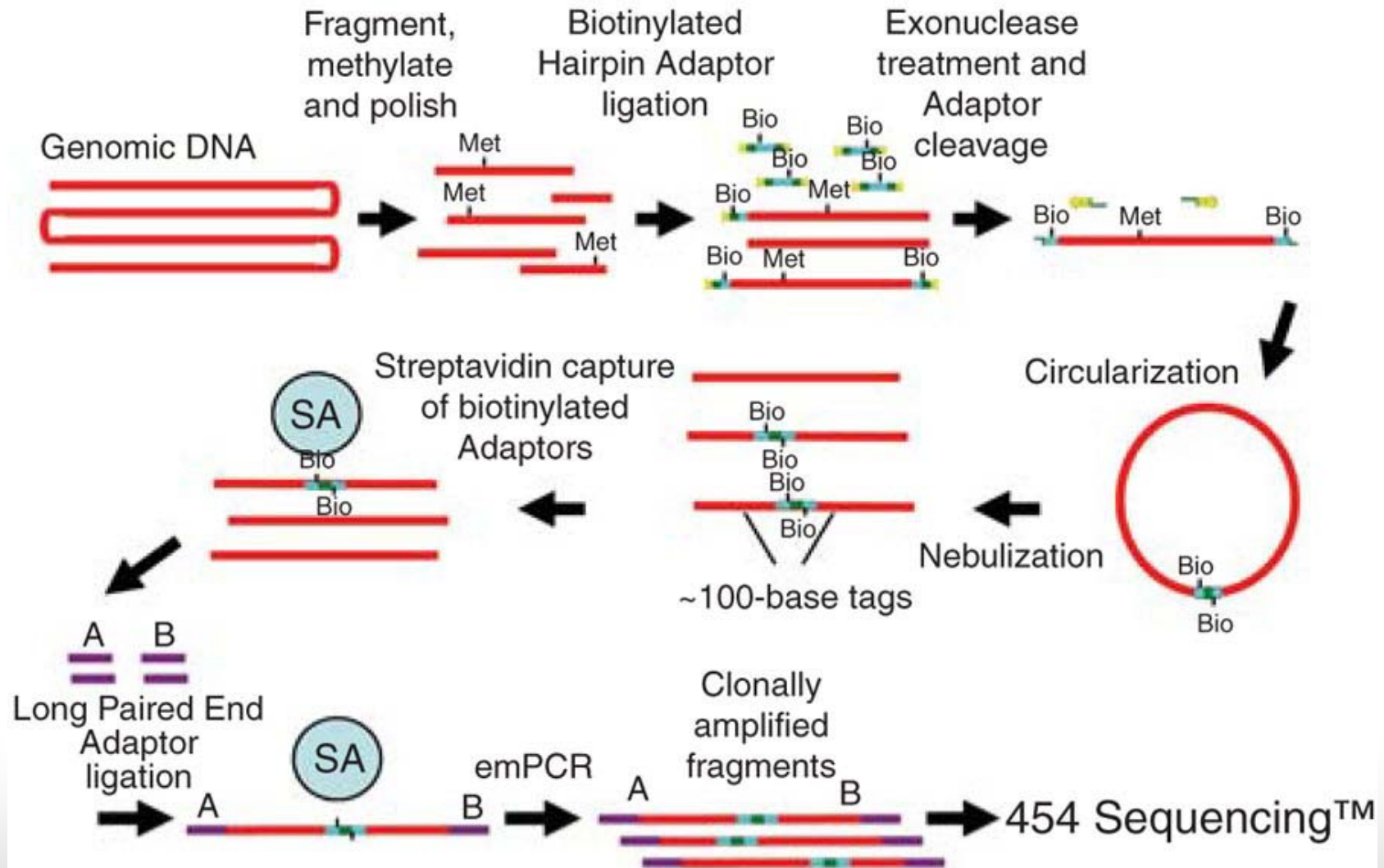
□ 需要非常多重的覆盖率

□ 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

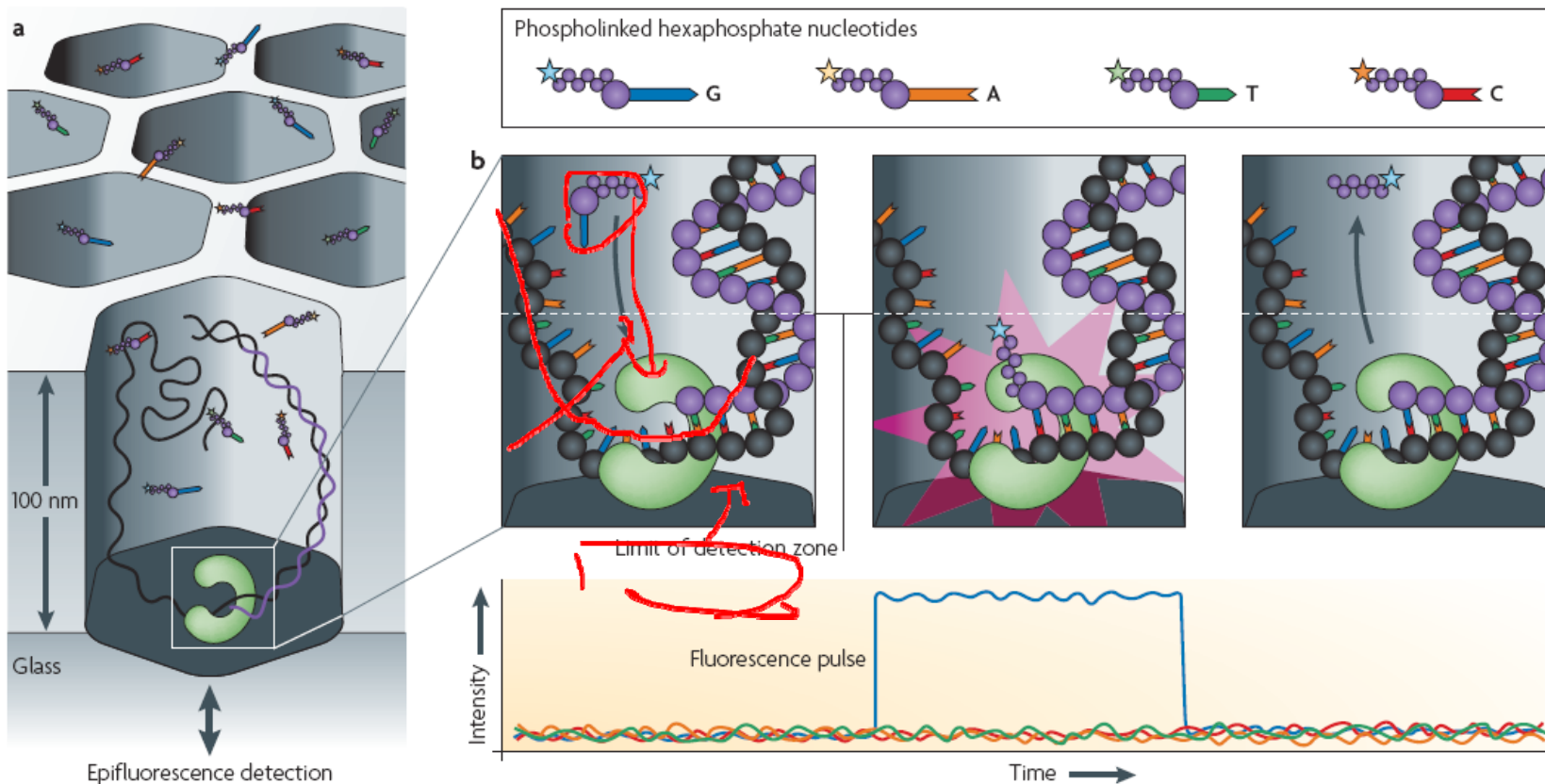
informatics, 2020, HUST

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

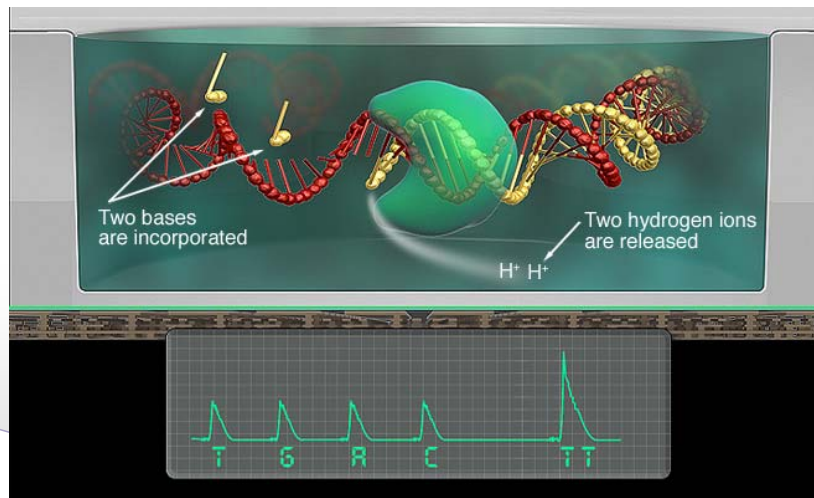


Ion Torrent™ (Life Technologies)



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

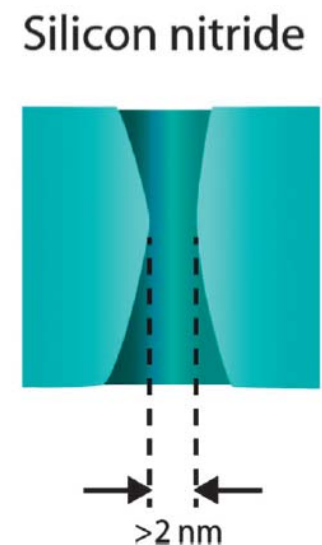
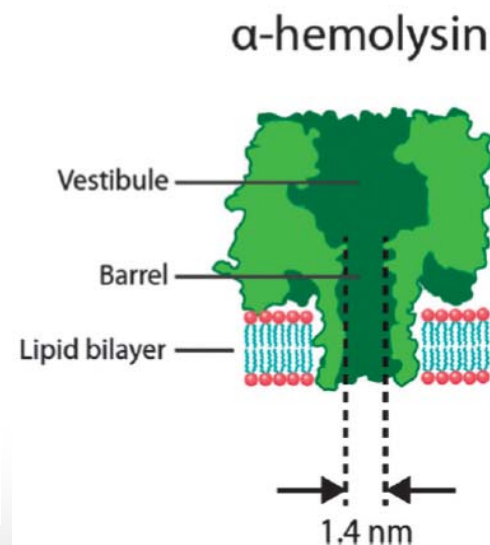
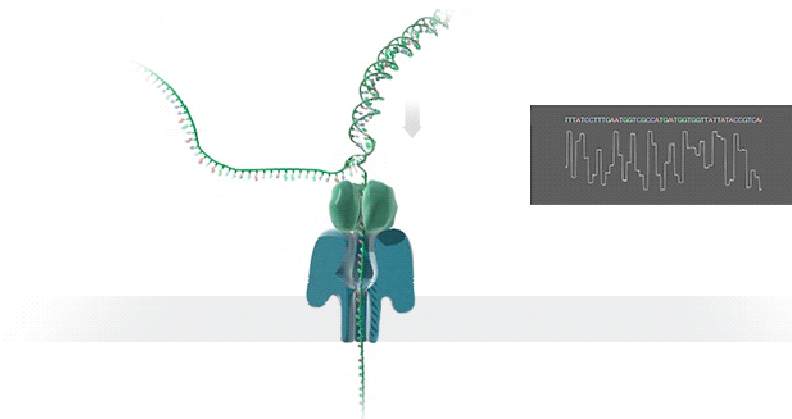
Nanopore测序原理



□ DNA链的直径:

⚙ B-DNA (2.0nm); A-DNA (2.6nm)

□ Nanopore类型: 生物孔和材料孔



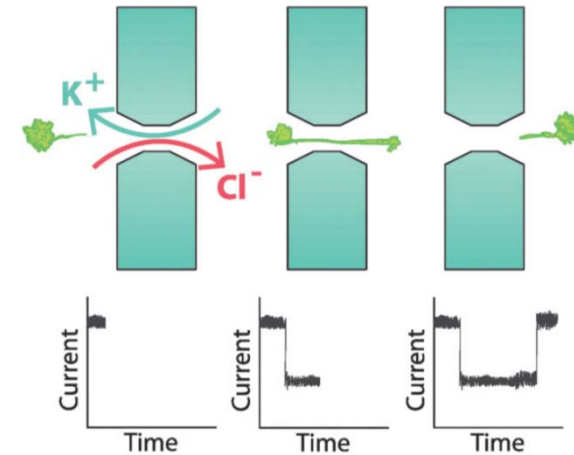
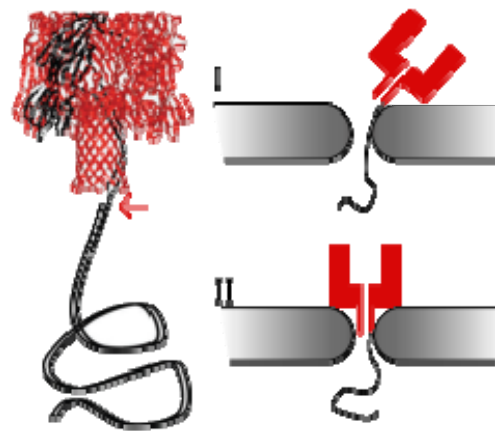
Nanopore测序原理



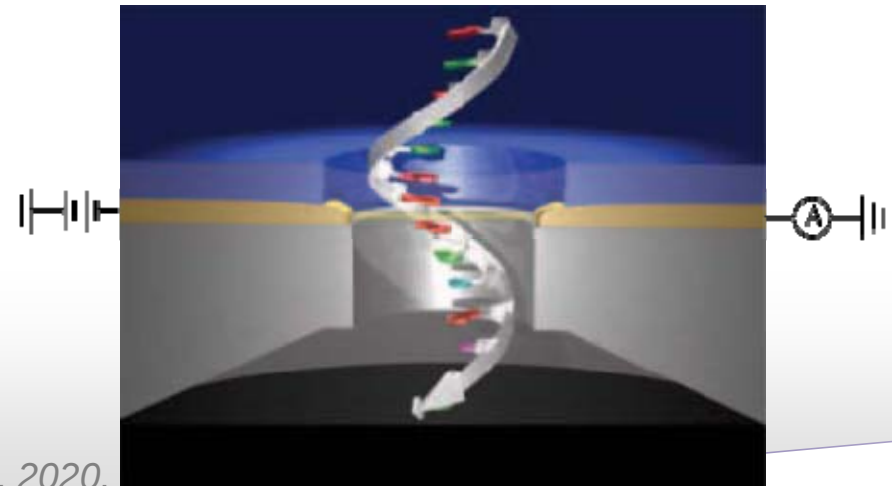
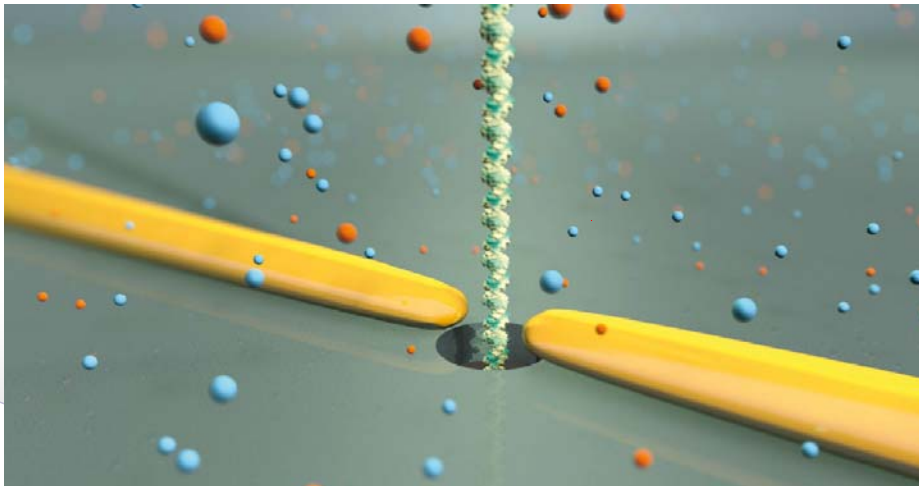
□ 膜两侧加电压

⚙ 开孔电流

⚙ 阻遏电流



□ 纳米孔两侧加电压



cs, 2020,

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.

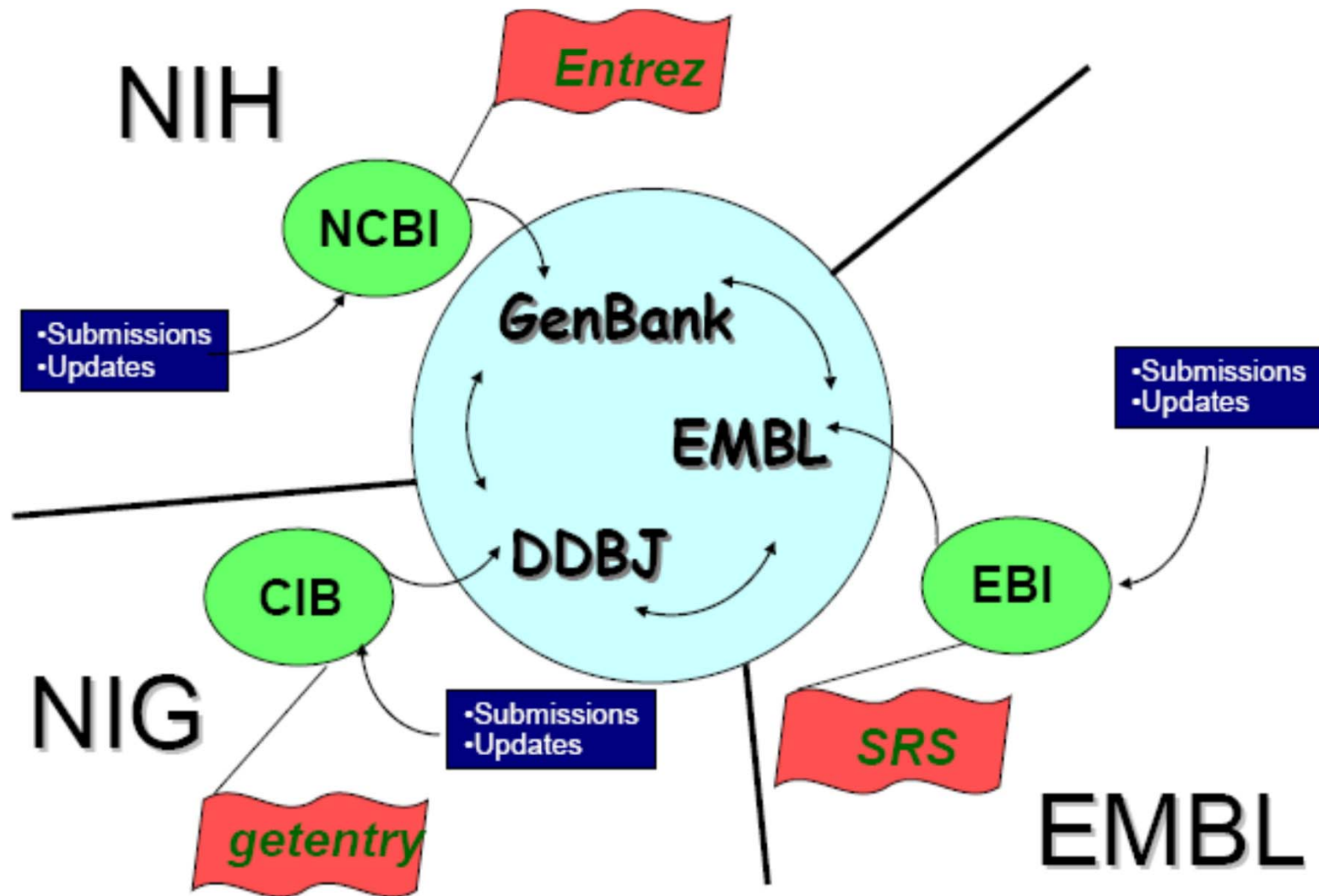
https://en.wikipedia.org/wiki/DNA_sequencing

序列数据的存储



- ❑ 核酸**四**大数据库：GenBank, EBI, **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

三大数据库之间的联系



NGDC



← → × ⓘ 不安全 | bigd.big.ac.cn

NGDC Databases Tools Standards Publications About

English 简体中文

National Genomics Data Center

The National Genomics Data Center advances life & health sciences by providing open access to a suite of resources, with the aim to translate big data into big discoveries and support worldwide activities in both academia and industry.

↑ Data Submission

All databases Find a bioproject, biosample, gene, protein, tool, database... Search

e.g., PRJCA000126; SAMC000385; tp53; EGFR; human; KaKs Calculator; GenBank

New 2019-nCoV Data Resource

<http://bigd.big.ac.cn/>

Popular Resources

- GSA**
Genome Sequence Archive
- BioProject**
Biological Project Library
- GVM**
Genome Variation Map
- BioSample**
Biological Sample Library
- GWH**
- Database Commons**

Featured Resources

- Human Genomics**
EDK EWAS Data Hub PGG.SNV PGG.Han
- Biodiversity Databases**
iDog iSheep IC4R eLMSG PADS Arsenal LSD PED

Genome Sequence Archive



← → × bigd.big.ac.cn/gsa/

NGDC Databases Tools Standards Publications About Sign in

 **GSA**
Genome Sequence Archive

GSA find a GSA accession 
e.g., CRA000112; CRX006656; human

中文 English

Home Submit Browse Search Statistics Documentation Login Register

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。

中国基因组数据共享倡议

提交



提交数据到GSA

下载



下载GSA数据

浏览



浏览已经公开的GSA信息

文档



查找帮助信息和说明文档

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。

GSA支持的期刊

GSA已获得多个国际期刊的认可，并已被国际著名出版商Elsevier收录为指定基因数据归档库。



数据统计

ns 100000 80000 Experiment 1600

<http://bigd.big.ac.cn/gsa/>

Ensembl数据库



← → ↻ ⓘ 不安全 | asia.ensembl.org/index.html

e!Ensembl ASIA | 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 -- ▼

- [View full list of all Ensembl species](#)
- [Edit your favourites](#)

Favourite genomes



Human
GRCh38.p13

[Still using GRCh37?](#)



Mouse
GRCm38.p6

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 99 (January 2020)

- Update to GENCODE 33 for human
- Update to dbSNP153 for human
- Import of updated VISTA enhancers for human and mouse
- New genomes: 10 mammals (including 2 dog breeds), 11 birds, 15 fish and 4 reptiles
- Updated genome assemblies: zebra finch, fugu, Nile tilapia and Asian bonytongue

[More release news](#) on our blog

Other news from our blog

- 30 Jan 2020: [Reaching under-served](#)
- 24 Jan 2020: [Job: Genome Annotator \(Regulation\)](#)
- 24 Jan 2020: [Cool stuff the Ensembl VEP can do: write your own plugin](#)

<http://www.ensembl.org/index.html>

Refseq数据库



- ❑ 提供高质量无冗余完整的序列信息
- ❑ 包括基因组的DNA，转录成的RNA以及蛋白质序列信息
- ❑ 序列文件的标识符：
 - ✿ DNA/RNA序列，NM_XXX, XM_XXX
 - ✿ 蛋白质序列：NP_XXX, XP_XXX
 - ✿ 染色体：NC_XXX

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

Announcements

January 10, 2020

RefSeq Release 98 is available for FTP

This release includes:

Proteins: 161,133,441

Transcripts: 29,134,515

Organisms: 98,406

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

Documentation: [Release Notes](#)

See [previous announcements](#), follow [NCBI on Twitter](#), or subscribe to [NCBI's refseq-announce mail list](#) to receive announcements.

NCBI Gene



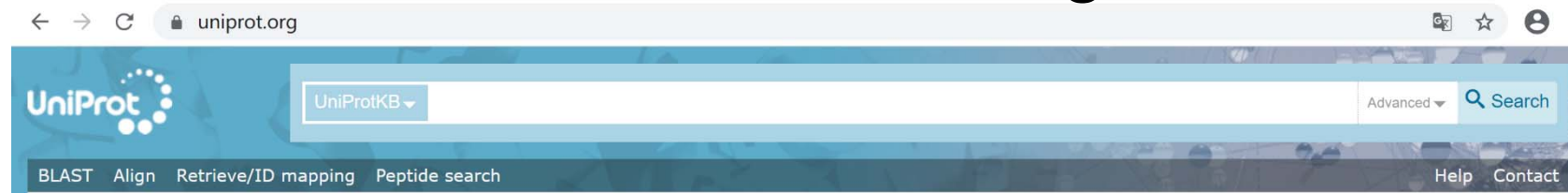
- ❑ 序列从Refseq数据库中得到
- ❑ 详尽的注释信息，包括基因在基因组的定位，基因名称、蛋白质名称，基因结构，等等

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



❑ 专家审核的蛋白质序列数据与知识库

❑ UniProtKB: UniProt Knowledgebase



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

The screenshot displays the UniProt homepage layout. On the left, a vertical sidebar lists database categories: UniProtKB (UniProt Knowledgebase) with Swiss-Prot (561,568) and TrEMBL (179,250,561), UniRef (Sequence clusters), UniParc (Sequence archive), and Proteomes (Proteome sets). A central section titled 'Supporting data' includes links for Literature citations, Taxonomy, and Subcellular locations. On the right, a 'News' section features social media icons (Blog, Twitter, Facebook, RSS) and three news items: 'Pre-release access to 2019 Wuhan Coronavirus protein data', 'Forthcoming changes' (Planned changes for UniProt), and 'UniProt release 2019_11' (Thicker than water | Functional annotation of different gene products | Changes to FT and CC text format | Cross-references to RNaAct | Pr...). A 'News archive' link is located below the news section.

UniProtKB
UniProt Knowledgebase

Swiss-Prot
(561,568)
Manually annotated and reviewed.
Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL
(179,250,561)
Automatically annotated and not reviewed.

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes
Proteome sets

Supporting data

Literature citations Taxonomy Subcellular locations

News

Pre-release access to 2019 Wuhan Coronavirus protein data

Forthcoming changes
Planned changes for UniProt

UniProt release 2019_11
Thicker than water | Functional annotation of different gene products | Changes to FT and CC text format | Cross-references to RNaAct | Pr...

News archive

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

序列数据的文件格式



- ❑ DNA/RNA/氨基酸代码的标识
- ❑ GenBank数据格式
- ❑ UniProt
- ❑ FASTA

DNA代码

Symbol	Meaning
G	G
A	A
T	T
C	C
R	A or G
Y	C or T
M	A or C
K	G or T
S	C or G
W	A or T
H	A, G or T not C
B	C, G or T not A
V	A, C or G not T/U
D	A, G or T not C
N	A, C, G, or T

氨基酸代码



1-letter code	3-letter code	Amin Acid
A	Ala	alanine
C	Cys	cysteine
D	Asp	aspartic acid
E	Glu	glutamic acid
F	Phe	phenylalanine
G	Gly	glycine
H	His	histidine
I	Ile	isoleucine
K	Lys	lysine
L	Leu	leucine
M	Met	methionine
N	Asn	asparagine
P	Pro	proline
Q	Gln	glutamine
R	Arg	arginine
S	Ser	serine
T	Thr	threonine
V	Val	valine
W	Trp	tryptophan
X	Xxx	Any
Y	Tyr	tyrosine
B	Asx	Asp or Asn
Z	Glx	Glu, or Gln

DNA代码

Symbol	Meaning
G	G
A	A
T	T
C	C
R	A or G
Y	C or T
M	A or C
K	G or T
S	C or G
W	A or T
H	A, G or T not C
B	C, G or T not A
V	A, C or G not T/U
D	A, G or T not C
N	A, C, G, or T

氨基酸代码



1-letter code	3-letter code	Amin Acid
A	Ala	alanine
C	Cys	cysteine
D	Asp	aspartic acid
E	Glu	glutamic acid
F	Phe	phenylalanine
G	Gly	glycine
H	His	histidine
I	Ile	isoleucine
K	Lys	lysine
L	Leu	leucine
M	Met	methionine
N	Asn	asparagine
P	Pro	proline
Q	Gln	glutamine
R	Arg	arginine
S	Ser	serine
T	Thr	threonine
V	Val	valine
W	Trp	tryptophan
X	Xxx	Any
Y	Tyr	tyrosine
B	Asx	Asp or Asn
Z	Glx	Glu, or Gln

GenBank的数据类型



1. PRI - primate sequences
2. ROD - rodent sequences
3. MAM - other mammalian sequences
4. VRT - other vertebrate sequences
5. INV - invertebrate sequences
6. PLN - plant, fungal, and algal sequences
7. BCT - bacterial sequences
8. VRL - viral sequences
9. PHG - bacteriophage sequences
10. SYN - synthetic sequences
11. UNA - unannotated sequences
12. EST - EST sequences (expressed sequence tags)
13. PAT - patent sequences
14. STS - STS sequences (sequence tagged sites)
15. GSS - GSS sequences (genome survey sequences)
16. HTG - HTGS sequences (high throughput genomic sequences)
17. HTC - HTC sequences (high throughput cDNA sequences)
18. ENV - Environmental sampling sequences
19. CON - Constructed sequences

UniProt数据格式



UniProtKB

Downloads · Contact · Documentation/Help

Search Blast * Align Retrieve ID Mapping *

Search in Protein Knowledgebase (UniProtKB) Query Search Advanced Search » Clear

P31749 (AKT1_HUMAN) ★ Reviewed, UniProtKB/Swiss-Prot
Last modified July 27, 2011. Version 148. History...

Contribute
Send feedback
Read comments (0) or add your own

Clusters with 100%, 90%, 50% identity | Documents (7) | Third-party data
text xml rdf/xml gff fasta

Names · Attributes · General annotation · Ontologies · Interactions · Sequence annotation · Sequences · References · Cross-refs · Entry info · Documents · Customize order

Names and origin

Protein names	Recommended name: RAC-alpha serine/threonine-protein kinase EC=2.7.11.1 Alternative name(s): Protein kinase B Short name=PKB Proto-oncogene c-Akt RAC-PK-alpha
Gene names	Name: AKT1 Synonyms: PKB, RAC
Organism	Homo sapiens (Human) [Complete proteome]
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo

Protein attributes

Sequence length	480 AA.
Sequence status	Complete.
Protein existence	Evidence at protein level.

Accession number

FASTA格式



Display FASTA Show: 20 Send to File Get Subsequent

1: **FASTA Definition Line**

>gi|460243|gb|U07163.1|SCU07163

gi number

Locus Name

Accession number

Database Identifiers

gb	GenBank
emb	EMBL
dbj	DDBJ
sp	SWISS-PROT
pdb	Protein Databank
pir	PIR
prf	PRF
ref	RefSeq

prote
A
C
CGTT
ATTGC
AACAP
CCATCGAAAAAG
ACAACAAGAC
TACGTCCAGGATCAA
AACCATGGAGAA
TCCCATTTCGC
CGCAG
AATTCAAA
CTTCACCTGTAAGAGA
AATTGAACAC
CAATA
SATATGGA
CAAAATTCCATCACCTA
SGAAAGCC
ATGATACACGAAATAGAA
TCC
ACTTTC
GAGTA
ATATATTTGCGCACT
CA
ATA
TT
ACC
GTACACGGACCCCTC
ATGCATAAAAAGAATATTATTCAT
TTAAGTTAACGGACTTCCGATGGA
AATTGACTACCTTTCTCCAGAAAT
GGCGTCCTGGCGTTTGAACACTG
AAAGGATAGCAGCACTGGATATCA
ACTACTAAAATACGACCCCAAAGA
AACAAAGCCCTTTTGGGAAAATAAG
AATCACTCCCGCACATATCACATA
TATATTGCGATATTTGATTAAATT
AGAGGAAAACAAGCTGAAAATTGC
ATAATAAGAGTAATGAAAGAAAGC
ACATTCAATTGTCTCAA

序列数据的查询



- 例：利用实验学方法（例如，酵母双杂交），发现了与有丝分裂期间某个蛋白可能相互作用的一个基因，测序结果如下（genotype）：

```
CCCCTGCCTGGCAGCCCTTTCTCAAGGACCACCGCATCTCTACATTCAAGA  
ACTGGCCCTTCTTGGAGGGCTGCGCCTGCACCCCGGAGCGGATGGCCGA  
GGCTGGCTTCATCCACTGCCCCACTGAGAACGAGCCAGACTTGGCCCAGT  
GTTTCTTCTGCTTCAAGGAGCTGGAAGGCTGGGAGCCAGATGACGACCCC  
ATAGAGGAACATAAAAAGCATTTCGTCCGGTTGCGCTTTCCTTTCTGTCAAGA  
AGCAGTTTGAAGAATTAACCCTTGGTGAATTTTTGAACTGGACAGAGAAAG  
AGCCAAGAACAAAATTGCAAAGGAAACCAACAATAAGAAGAAAGAATTTGAG  
GAACTGCGGAGAAAGTGCGCCGTGCCATCGAGCAGCTGGCTGCCATGGA  
TTGAGGCCTCTGGC
```

问题：



- ☐ 这是哪个基因？
- ☐ 编码的蛋白质序列是怎样的？
- ☐ 有没有保守的功能结构域（domain）？
- ☐ 它的功能是怎样的？
- ☐ 它在真核生物中保守吗？
- ☐ 有没有三级结构信息？

NCBI: BLAST



← → ↻ ncbi.nlm.nih.gov

NCBI Resources ▾ How To ▾ Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases ▾ Search

NCBI Home

Resource List (A-Z)

- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Develop

Use NCBI APIs and code libraries to build applications

Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST**
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

Important changes to the genomes FTP site in February 07 Feb 2020

We have added the latest NCBI Eukaryotic Genome Annotation Pipeline

Read about NCBI resources in 2020 Nucleic Acids Research database issue 05 Feb 2020

The 2020 Nucleic Acids Research database issue features papers from

NLM announces Curation at Scale Workshop 04 Feb 2020

Data curation plays a critical role in

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

Nucleotide blast



← → ↻ blast.ncbi.nlm.nih.gov/Blast.cgi

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

Search Betacoronavirus Database

We have created a new BLAST database focused on the 2019-nCoV Sequences (Wuhan coronavirus). For further detail please visit [NCBI GenBank](#).

Mon, 03 Feb 2020 10:00:00 EST [More BLAST news...](#)

Web BLAST

Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide

Protein BLAST
protein ► protein

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Megablast: 找基因序列



Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)

[Clear](#) Query subrange [?](#)

From

To

BLAST results will be displayed in a new format by default. You can always switch back to the Traditional Results page. 

Or, upload file 未选择任何文件 [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database ☒ Standard databases (nr etc.): ☐ rRNA/ITS databases ☐ Genomic + transcript databases ☐ Betacoronavirus

Nucleotide collection (nr/nt) [?](#)

Organism [Optional](#) ☐ exclude [+](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude [Optional](#) ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Limit to [Optional](#) ☐ Sequences from type material

Entrez Query [Optional](#) [YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for ☒ Highly similar sequences (megablast)

☐ More dissimilar sequences (discontiguous megablast)

☐ Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

BLAST

Search **database Nucleotide collection (nr/nt)** using **Megablast (Optimize for highly similar sequences)**

☒ Show results in a new window

输入序列

提交序列



http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi

NCBI Blast: lcl|27166 (420 l...)

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

► NCBI/ BLAST/ blastn/ Formatting Results - DJMP5ZKF012 [Formatting options]

Job Title: lcl|27166 (420 letters)

Request ID	DJMP5ZKF012
Status	Searching
Submitted at	Sat Sep 1 07:53:41 2007
Current time	Sat Sep 1 07:53:45 2007
Time since submission	

This page will be automatically updated in 6 seconds

[Copyright](#) | [Disclaimer](#) | [Privacy](#) | [Accessibility](#) | [Contact](#) | [Send feedback on new interface](#)

[NCBI](#) | [NLM](#) | [NIH](#) | [DHHS](#)

NM_001168.3: BIRC5 (Survivin)



[Distance tree of results](#) [MSA viewer](#) [?](#)

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download [v](#)

Manage Columns [v](#)

Show [?](#)

☒ select all 100 sequences selected

[GenBank](#)

[Graphics](#)

[Distance tree of results](#)

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☒	Homo sapiens baculoviral IAP repeat containing 5 (BIRC5), transcript variant 1, mRNA	776	776	100%	0.0	100.00%	NM_001168.3
☒	Homo sapiens baculoviral IAP repeat-containing 5 (BIRC5) mRNA, complete cds	771	771	100%	0.0	99.76%	HM625836.1
☒	Synthetic construct Homo sapiens baculoviral IAP repeat-containing 5 (survivin), mRNA (cDNA clone IMAGE:2964713), **** V	771	771	100%	0.0	99.76%	BC002388.1
☒	Homo sapiens baculoviral IAP repeat-containing 5, mRNA (cDNA clone MGC:75168 IMAGE:5394399), complete cds	771	771	100%	0.0	99.76%	BC065497.1
☒	Pongo abelii baculoviral IAP repeat containing 5 (BIRC5), mRNA	771	771	100%	0.0	99.76%	NM_001132255.1
☒	Homo sapiens inhibitor of apoptosis homolog mRNA, complete cds	771	771	100%	0.0	99.76%	AF077350.1
☒	Homo sapiens baculoviral IAP repeat-containing 5, mRNA (cDNA clone MGC:8592 IMAGE:2961114), complete cds	771	771	100%	0.0	99.76%	BC008718.2
☒	Homo sapiens baculoviral IAP repeat-containing 5, mRNA (cDNA clone MGC:32768 IMAGE:4656567), complete cds	771	771	100%	0.0	99.76%	BC034148.1
☒	PREDICTED: Gorilla gorilla gorilla baculoviral IAP repeat containing 5 (BIRC5), transcript variant X1, mRNA	765	765	100%	0.0	99.52%	XM_019028170.2
☒	PREDICTED: Pan troglodytes baculoviral IAP repeat containing 5 (BIRC5), transcript variant X1, mRNA	765	765	100%	0.0	99.52%	XM_009433243.3
☒	PREDICTED: Pan paniscus baculoviral IAP repeat containing 5 (BIRC5), transcript variant X1, mRNA	765	765	100%	0.0	99.52%	XM_003818274.2

BIRC5 Gene



Descriptions

Graphic Summary

Alignments

Taxonomy

Alignment view

Pairwise



☐ CDS feature ?

Download ▾

100 sequences selected ?

Download ▾

[GenBank](#) [Graphics](#)

▾ Next ▲ Previous ◀ Descriptions

Homo sapiens baculoviral IAP repeat containing 5 (BIRC5), transcript variant 1, mRNA

Sequence ID: [NM_001168.3](#) Length: 2574 Number of Matches: 1

Range 1: 84 to 503 [GenBank](#) [Graphics](#)

▾ Next Match ▲ Previous Match

	Score	Expect	Identities	Gaps	Strand
	776 bits(420)	0.0	420/420(100%)	0/420(0%)	Plus/Plus
Query 1	CCCCCTGCCTGGCAGCCCTTTCTCAAGGACCACCGCATCTCTACATTCAAGAACTGGCCCT				60
Sbjct 84	CCCCCTGCCTGGCAGCCCTTTCTCAAGGACCACCGCATCTCTACATTCAAGAACTGGCCCT				143
Query 61	TCTTGGAGGGCTGCGCCTGCACCCGGAGCGGATGGCCGAGGCTGGCTTCATCCACTGCC				120
Sbjct 144	TCTTGGAGGGCTGCGCCTGCACCCGGAGCGGATGGCCGAGGCTGGCTTCATCCACTGCC				203
Query 121	CCACTGAGAACGAGCCAGACTTGGCCCAGTGTCTTCTGCTTCAAGGAGCTGGAAGGCT				180
Sbjct 204	CCACTGAGAACGAGCCAGACTTGGCCCAGTGTCTTCTGCTTCAAGGAGCTGGAAGGCT				263
Query 181	GGGAGCCAGATGACGACCCCATAGAGGAACATAAAAAGCATTTCGTCCGGTTGCGCTTTCC				240
Sbjct 264	GGGAGCCAGATGACGACCCCATAGAGGAACATAAAAAGCATTTCGTCCGGTTGCGCTTTCC				323
Query 241	TTTCTGTCAAGAAGCAGTTTGAAGAATTAACCTTGGTGAATTTTGAAACTGGACAGAG				300

Related Information

[Gene](#) - associated gene details
[PubChem BioAssay](#) - bioactivity screening
[Genome Data Viewer](#) - aligned genomic context

BIRC5 gene information



Resources ☒ How To ☒ Sign in to NCBI

Gene Create RSS Save search Advanced Help

Full Report

Showing Current items.

BIRC5 baculoviral IAP repeat containing 5 [*Homo sapiens* (human)]
Gene ID: 332, updated on 3-Feb-2020

Official Symbol BIRC5 provided by HGNC

Official Full Name baculoviral IAP repeat containing 5 provided by HGNC

Primary source [HGNC:HGNC:593](#)

See related [Ensembl:ENSG00000089685](#) [MIM:603352](#)

Gene type protein coding

RefSeq status REVIEWED

Organism [Homo sapiens](#)

Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

Also known as API4; EPR-1

Summary This gene is a member of the inhibitor of apoptosis (IAP) gene family, which encode negative regulatory proteins that prevent apoptotic cell death. IAP family members usually contain multiple baculovirus IAP repeat (BIR) domains, but this gene encodes proteins with only a single BIR domain. The encoded proteins also lack a C-terminus RING finger domain. Gene expression is high during fetal development and in most tumors, yet low in adult tissues. Alternatively spliced transcript variants encoding distinct isoforms have been found for this gene. [provided by RefSeq, Jun 2011]

Expression Biased expression in bone marrow (RPKM 13.2), testis (RPKM 10.8) and 12 other tissues [See more](#)

Orthologs [mouse](#) [all](#)

Table of contents

[Summary](#)

[Genomic context](#)

[Genomic regions, transcripts, and products](#)

[Expression](#)

[Bibliography](#)

[Phenotypes](#)

[Variation](#)

[HIV-1 interactions](#)

[Pathways from PubChem](#)

[Interactions](#)

[General gene information](#)
[Markers, Homology, Gene Ontology](#)

[General protein information](#)

[NCBI Reference Sequences \(RefSeq\)](#)

[Related sequences](#)

[Additional links](#)
[Locus-specific Databases](#)

Gene info: 17号染色体



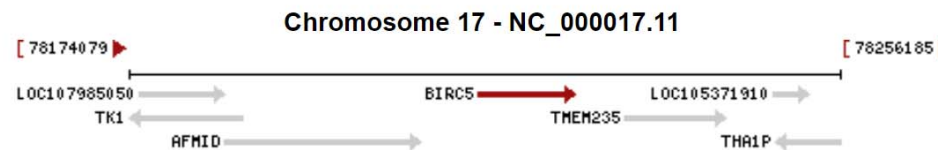
Genomic context

Location: 17q25.3

See BIRC5 in [Genome Data Viewer](#)

Exon count: 6

Annotation release	Status	Assembly	Chr	Location
109.20191205	current	GRCh38.p13 (GCF_000001405.39)	17	NC_000017.11 (78214253..78225635)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	17	NC_000017.10 (76210277..76221716)

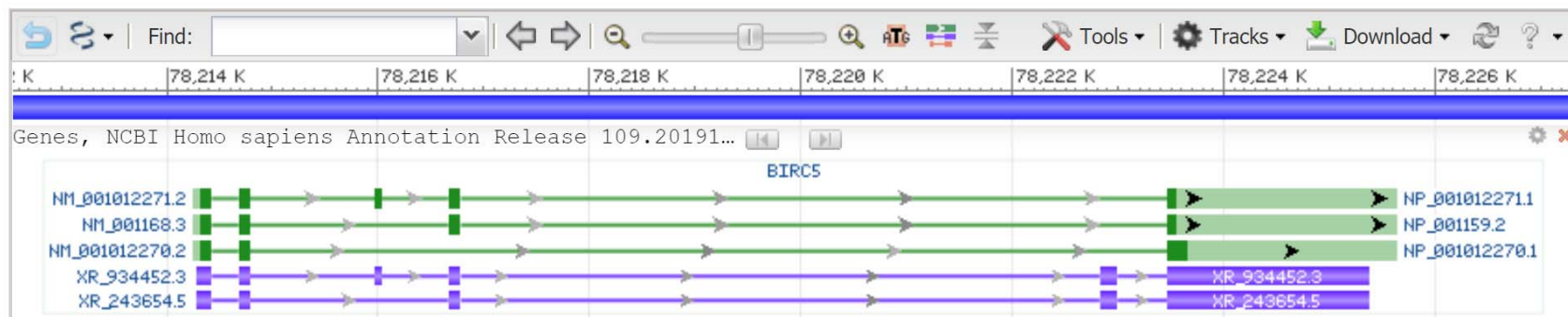


Genomic regions, transcripts, and products

Go to [reference sequence details](#)

Genomic Sequence: NC_000017.11 Chromosome 17 Reference GRCh38.p13 Primary Assembly ▼

Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)



功能注释: Gene Ontology



Process	Evidence Code	Pubs
apoptotic process	IEA	
cell division	IEA	
chromosome segregation	IEA	
cytokine-mediated signaling pathway	TAS	
mitotic spindle assembly	IBA	PubMed
negative regulation of apoptotic process	IDA	PubMed
negative regulation of apoptotic process	IMP	PubMed
negative regulation of cysteine-type endopeptidase activity involved in apoptotic process	IBA	PubMed
negative regulation of transcription, DNA-templated	IMP	PubMed
positive regulation of cell proliferation	TAS	PubMed
protein phosphorylation	IDA	PubMed
protein-containing complex localization	IMP	PubMed
regulation of apoptotic process	TAS	
sensory perception of sound	IEP	PubMed

结论1



- ❑ 核酸序列标识符: NM_001168.3
- ❑ 该基因为人的BIRC5基因
- ❑ 染色体定位: 17q25.3
- ❑ 基因组坐标: 78214253-78225635
- ❑ 初步的功能分析: 凋亡, 细胞分裂, 染色体分离, 等等

NM_001168.3: BIRC5



[Distance tree of results](#) [MSA viewer](#) [?](#)

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download [v](#)

Manage Columns [v](#)

Show

100 [v](#)



☒ select all 100 sequences selected

[GenBank](#)

[Graphics](#)

[Distance tree of results](#)

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☒	Homo sapiens baculoviral IAP repeat containing 5 (BIRC5), transcript variant 1, mRNA	776	776	100%	0.0	100.00%	NM_001168.3
☒	Homo sapiens baculoviral IAP repeat-containing 5 (BIRC5) mRNA, complete cds	771	771	100%	0.0	99.76%	HM625836.1
☒	Synthetic construct Homo sapiens baculoviral IAP repeat-containing 5 (survivin), mRNA (cDNA clone IMAGE:2964713), **** V	771	771	100%	0.0	99.76%	BC002388.1
☒	Homo sapiens baculoviral IAP repeat-containing 5, mRNA (cDNA clone MGC:75168 IMAGE:5394399), complete cds	771	771	100%	0.0	99.76%	BC065497.1
☒	Pongo abelii baculoviral IAP repeat containing 5 (BIRC5), mRNA	771	771	100%	0.0	99.76%	NM_001132255.1
☒	Homo sapiens inhibitor of apoptosis homolog mRNA, complete cds	771	771	100%	0.0	99.76%	AF077350.1
☒	Homo sapiens baculoviral IAP repeat-containing 5, mRNA (cDNA clone MGC:8592 IMAGE:2961114), complete cds	771	771	100%	0.0	99.76%	BC008718.2
☒	Homo sapiens baculoviral IAP repeat-containing 5, mRNA (cDNA clone MGC:32768 IMAGE:4656567), complete cds	771	771	100%	0.0	99.76%	BC034148.1
☒	PREDICTED: Gorilla gorilla gorilla baculoviral IAP repeat containing 5 (BIRC5), transcript variant X1, mRNA	765	765	100%	0.0	99.52%	XM_019028170.2
☒	PREDICTED: Pan troglodytes baculoviral IAP repeat containing 5 (BIRC5), transcript variant X1, mRNA	765	765	100%	0.0	99.52%	XM_009433243.3
☒	PREDICTED: Pan paniscus baculoviral IAP repeat containing 5 (BIRC5), transcript variant X1, mRNA	765	765	100%	0.0	99.52%	XM_003818274.2

Human BIRC5!



Resources ☒

How To ☒

Nucleotide

Nucleotide ▼

Advanced

GenBank ▼

Send to: ▼

Homo sapiens baculoviral IAP repeat containing 5 (BIRC5), transcript variant 1, mRNA

NCBI Reference Sequence: NM_001168.3

[FASTA](#) [Graphics](#)

Go to: ☒

LOCUS NM_001168 2574 bp mRNA linear PRI 13-JAN-2020
DEFINITION Homo sapiens baculoviral IAP repeat containing 5 (BIRC5),
transcript variant 1, mRNA.
ACCESSION NM_001168
VERSION NM_001168.3
KEYWORDS RefSeq; MANE Select.
SOURCE Homo sapiens (human)
ORGANISM [Homo sapiens](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 2574)
AUTHORS Zou J, Liao X, Zhang J and Wang L.
TITLE Dysregulation of miR-195-5p/-218-5p/BIRC5 axis predicts a poor
prognosis in patients with gastric cancer
JOURNAL J. Biol. Regul. Homeost. Agents 33 (5), 1377-1385 (2019)
PUBMED [31663299](#)
REMARK GeneRIF: dysregulation of miR-195-5p/-218-5p/BIRC5 axis predicts a

获取蛋白质的序列信息



CDS

```
65..493
/gene="BIRC5"
/gene_synonym="API4; EPR-1"
/note="isoform 1 is encoded by transcript variant 1;
survivin variant 3 alpha; apoptosis inhibitor 4;
baculoviral IAP repeat-containing protein 5; apoptosis
inhibitor survivin"
/codon_start=1
/product="baculoviral IAP repeat-containing protein 5
isoform 1"
/protein_id="NP_001159.2"
/db_xref="CCDS:CCDS11755.1"
/db_xref="GeneID:332"
/db_xref="HGNC:HGNC:593"
/db_xref="MIM:603352"
/translation="MGAPTLPPAWQPFLKDHRISTFKNWPFLEGCACTERMAEAGFI
HCPTENEPDLAQCFKCFKELEGWEPDDDPIDEEHKKHSSGCAFLSVKKQFEELTLGEFL
KLDREERAKNKIAKETNNKKKEFEETAIEKVRRAIEQLAAMD"
```

BIRC5: 142aa



Resources ☒ How To ☒

Protein

Protein ▼

Advanced

GenPept ▼

Send to: ▼

baculoviral IAP repeat-containing protein 5 isoform 1 [Homo sapiens]

NCBI Reference Sequence: NP_001159.2

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to: ☒

LOCUS NP_001159 142 aa linear PRI 13-JAN-2020

DEFINITION baculoviral IAP repeat-containing protein 5 isoform 1 [Homo sapiens].

ACCESSION NP_001159

VERSION NP_001159.2

DBSOURCE REFSEQ: accession [NM_001168.3](#)

KEYWORDS RefSeq; MANE Select.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (residues 1 to 142)

AUTHORS Zou J, Liao X, Zhang J and Wang L.

TITLE Dysregulation of miR-195-5p/-218-5p/BIRC5 axis predicts a poor prognosis in patients with gastric cancer

JOURNAL J. Biol. Regul. Homeost. Agents 33 (5), 1377-1385 (2019)

结论2



- 人的BIRC5蛋白质包含142个氨基酸，序列标识符为：NP_001159.2

获取FASTA序列



NCBI Resources ☒ How To ☒

Protein

Protein ▼

Advanced

GenPept ▼

Send to: ▼

baculoviral IAP repeat-containing protein 5 isoform 1 [Homo sapiens]

NCBI Reference Sequence: NP_001159.2

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to: ☒

LOCUS NP_001159 142 aa linear PRI 13-JAN-2020

DEFINITION baculoviral IAP repeat-containing protein 5 isoform 1 [Homo sapiens].

ACCESSION NP_001159

VERSION NP_001159.2

DBSOURCE REFSEQ: accession [NM_001168.3](#)

KEYWORDS RefSeq; MANE Select.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (residues 1 to 142)

AUTHORS Zou J, Liao X, Zhang J and Wang L.

TITLE Dysregulation of miR-195-5p/-218-5p/BIRC5 axis predicts a poor prognosis in patients with gastric cancer

JOURNAL J. Biol. Regul. Homeost. Agents 33 (5), 1377-1385 (2019)

FASTA格式的序列



FASTA ▼

baculoviral IAP repeat-containing protein 5 isoform 1 [Homo sapiens]

NCBI Reference Sequence: NP_001159.2

[GenPept](#) [Identical Proteins](#) [Graphics](#)

>NP_001159.2 baculoviral IAP repeat-containing protein 5 isoform 1 [Homo sapiens]

MGAPTLPPAWQPFLKDHRISTFKNWPFLGCACTPERMAEAGFIHCPTENEPDLAQCFKCFKELEGWEPD

DDPIEEHKKHSSGCAFLSVKKQFEELTLGEFLKLDREKAKNKIAKETNNKKKEFEETAEKVRRRAIEQLAA

MD

PHI-BLAST: find domain



blast.ncbi.nlm.nih.gov/Blast.cgi



Specialized searches

SmartBLAST

Find proteins highly similar to your query

Primer-BLAST

Design primers specific to your PCR template

Global Align

Compare two sequences across their entire span (Needleman-Wunsch)

CD-search

Find conserved domains in your sequence

IgBLAST

Search immunoglobulins and T cell receptor sequences

VecScreen

Search sequences for vector contamination

CDART

Find sequences with similar conserved domain architecture

Targeted Loci

Search markers for phylogenetic analysis

Multiple Alignment

Align sequences using domain and protein constraints

MOLE-BLAST

Establish taxonomy for uncultured or environmental sequences

填入蛋白质的FASTA序列



← → ↻ ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

NCBI

Conserved Domains

SH3 SH2

HOME SEARCH GUIDE Structure Home 3D Macromolecular Structures Conserved Domains

Search for Conserved Domains within a protein or coding nucleotide sequence

Enter **protein** or **nucleotide** query as accession, gi, or sequence in [FASTA format](#). For multiple protein queries, use [Batch CD-Search](#). [?](#)

```
>NP_001159.2 baculoviral IAP repeat-containing protein 5 isoform 1 [Homo sapiens]
MGAPTLPPAWQPFLKDHRISTFKNWPFLEGCACTPERMAEAGFIHCPTENEPDLAQCFKFELEGWEPD
DDPIEEHKKHSSGCAFLSVKKQFEELTLGEFLKLDREERAKNKIAKETNNKKKEFEETAEKVRRRAIEQLAA
MD
|
```

OPTIONS

Search against database [?](#): CDD v3.17 - 52910 PSSMs ▾

Expect Value [?](#) threshold: 0.010000

Apply low-complexity filter [?](#) ☐

Composition based statistics adjustment [?](#) ☒

Force live search [?](#) ☐

Rescue borderline hits ☐ Suppress weak overlapping hits ☐

Maximum number of hits [?](#) 500

Result mode ☒ Concise [?](#) ☐ Standard [?](#) ☐ Full [?](#)

Submit

Reset

[Help](#)

Retrieve previous CD-search result

Request ID: [Retrieve](#) [?](#)

BIR domain



NCBI

HOME SEARCH GUIDE NewSearch Structure Home 3D Macromolecular Structures Conserved Domains Pubchem BioSystems

Conserved domains on [lcl|seqsig_MGAPT_9a39546436cfe7208af86fbe7bf45623] View **Concise Results** ?

NP_001159.2 baculoviral IAP repeat-containing protein 5 isoform 1 [Homo sapiens]

Protein Classification ?

BIR domain-containing protein (domain architecture ID 10460345)
BIR domain-containing protein

Graphical summary ☐ Zoom to residue level show extra options » ?

Query seq. MGAPTLPPAQFLKDRISTFNMFFLEGCACRTERMEAGFTHCTENEPCLAQCFQVLEGEPODPIEBHKKSSGAFSLVKWDFEELTLGEPLKLDREPAWKIAKETNNKKKEFEETAEKVRRAIEQLAND

Specific hits

Superfamilies

BIR

BIR superfamily

Search for similar domain architectures ? Refine search ?

List of domain hits ?

	Name	Accession	Description	Interval	E-value
[+]	BIR	pfam00653	Inhibitor of Apoptosis domain; BIR stands for 'Baculovirus Inhibitor of apoptosis protein ...	18-87	7.47e-31

结论3



□ BIRC5具有保守的功能结构域BIR

ExPASy: 生物信息资源



ExPASy
Bioinformatics Resource Portal

Home · About · Contact

Query all databases search help

Visual Guidance

Categories

- proteomics
 - protein sequences and identification
 - proteomics experiment
 - function analysis
 - sequence sites, features and motifs
 - protein modifications
 - protein structure
 - protein interactions
 - similarity search/alignment**
 - genomics
- structure analysis
- systems biology
- evolutionary biology
- population genetics
- transcriptomics
- biophysics
- imaging
- IT infrastructure

Databases

- SIB resources
- External resources - (No support from the ExPASy Team)
- UniProtKB • functional information on proteins • [\[more\]](#)
- MyHits • protein domains database and tools • [\[more\]](#)

Tools

- Alignment tools** • Four tools for multiple alignments • [\[more\]](#)
- BLAST** • sequence similarity search • [\[more\]](#)
- BLAST (UniProt)** • BLAST search on the UniProt web site • [\[more\]](#)
- BLAST - NCBI** • Biological sequence similarity search • [\[more\]](#)
- BLAST - PBIL** • BLAST search on protein sequence databases • [\[more\]](#)
- Blast2Fasta** • Blast to Fasta conversion • [\[more\]](#)
- boxshade** • MSA pretty printer • [\[more\]](#)
- ClustalO (UniProt)** • Align two or more protein sequences • [\[more\]](#)
- Clustal Omega (EBI)** • Multiple sequence alignment program • [\[more\]](#)
- ClustalW** • Multiple sequence alignment • [\[more\]](#)
- ClustalW - PBIL** • Multiple sequence alignment program • [\[more\]](#)
- Decrease redundancy** • Sequence redundancy reduction • [\[more\]](#)
- DIALIGN** • Local multiple sequence alignment • [\[more\]](#)
- Dotlet** • sequence similarity plots • [\[more\]](#)
- FASTA/SSEARCH/GGSEARCH/GLSEARCH** • Sequence similarity searching of protein db • [\[more\]](#)
- GENIO/logo** • RNA/DNA & Amino Acid Sequence Logos • [\[more\]](#)

<https://www.expasy.org/>

BIOINFORMATICS, 2020, HOST

http://web.expasy.org/blast/



Enter a sequence

Examples

```
>NP_001159.2 baculoviral IAP repeat-containing protein  
isoform 1 [Homo sapiens]  
MGAPTLPPAWQPFLKDHISTFKNWPFLEGCACTPERMAEAGFIHCPTENE  
PD  
KELEGWEPD  
DDPIEEHKKHSSGCAFLSVKKQFEELTLGEFLKDRERAKNKIAKETNNKK  
KE  
RRATEQLAA  
MD
```

e.g. [P00750](#), [P05067-5](#), [A4_HUMAN](#) or `accggtggtc`

Run BLAST **Reset**

Primates

Rodentia

Vertebrata

Viridiplantae

Arabidopsis thaliana

Caenorhabditis elegans

Dictyostelium discoideum

Drosophila melanogaster

~~Escherichia coli~~

Homo sapiens

~~Mus musculus~~

Plasmodium falciparum

Rattus norvegicus

Saccharomyces cerevisiae

Schizosaccharomyces pombe

Help on ExPASy: ?

protein sequence database

nucleotide sequence database

nucleotide sequence database

protein sequence database

Choose a database

Protein databases:

[UniProt Knowledgebase \(UniProtKB\)](#) ?

[UniProtKB taxonomic subsets](#) ?

[UniProt prokaryotic reference proteomes](#) ?

[Other databases](#) ?

☒ [UniProtKB/Swiss-Prot only.](#) ?

BIRC5: O15392



Top

Graphical overview of the alignments

List of the matches

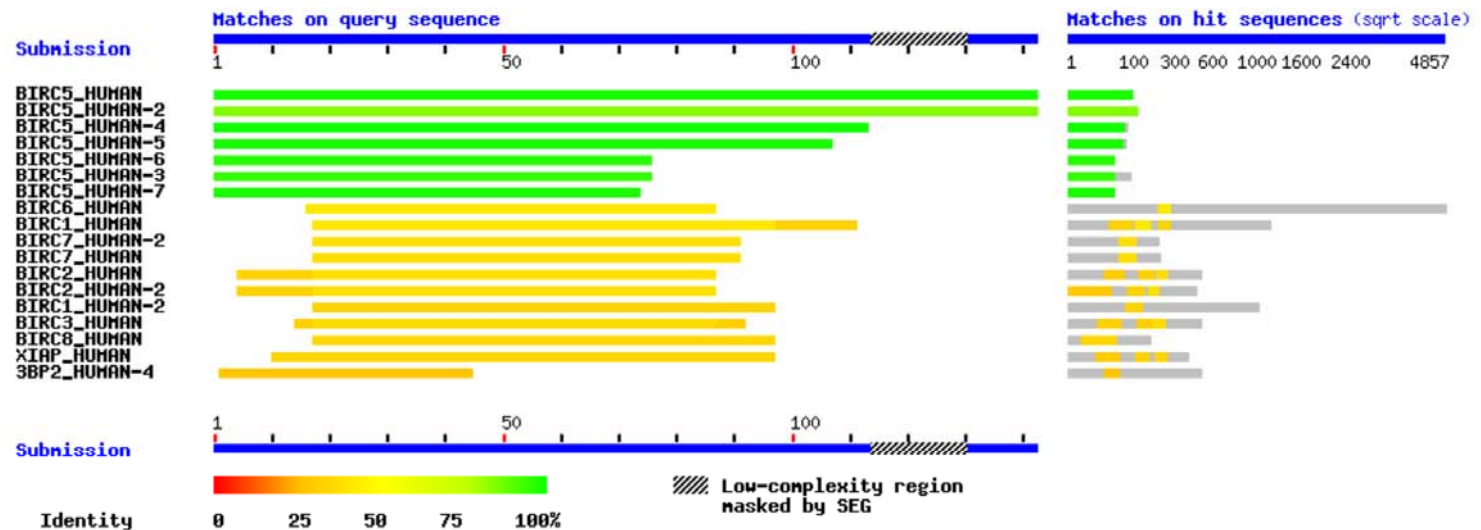
Alignments

BLAST parameters

Text output

Notice, contact and references

Graphical overview of the alignments



List of the matches

Clustal W (multiple alignment)

提交

☐ Select up to...

☐ Include query sequence

	Accession	Length	Db	Description	Score	E-value
☐ 1	O15392 (BIRC5_HUMAN)	142		Baculoviral IAP repeat-containing protein ...	256 bits (654)	1e-87

BIRC5的蛋白质信息



← → × uniprot.org/uniprot/O15392

UniProt

UniProtKB

Advanced Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

UniProtKB - O15392 (BIRC5_HUMAN)

Display

Entry Publications Feature viewer Feature table

Protein **Baculoviral IAP repeat-containing protein 5**

Gene **BIRC5**

Organism *Homo sapiens (Human)*

Status Reviewed - Annotation score: ●●●●●● - Experimental evidence at protein levelⁱ

BLAST Align Format Add to basket History Help video Add a publication Feedback

All None

☒ Function

☒ Names & Taxonomy

☒ Subcellular location

☒ Pathology & Biotech

☒ PTM / Processing

☒ Expression

☒ Interaction

☒ Structure

☒ Family & Domains

Functionⁱ

Multitasking protein that has dual roles in promoting cell proliferation and preventing apoptosis (PubMed:9859993, PubMed:21364656, PubMed:20627126, PubMed:25778398, PubMed:28218735). Component of a chromosome passage protein complex (CPC) which is essential for chromosome alignment and segregation during mitosis and cytokinesis (PubMed:16322459). Acts as an important regulator of the localization of this complex; directs CPC movement to different locations from the inner centromere during prometaphase to midbody during cytokinesis and participates in the organization of the center spindle by associating with polymerized microtubules (PubMed:20826784). Involved in the recruitment of CPC to centromeres during early mitosis via association with histone H3 phosphorylated at 'Thr-3' (H3pT3) during mitosis (PubMed:20929775). The complex with RAN plays a role in mitotic spindle formation by serving as a physical scaffold to help deliver the RAN effector molecule TPX2 to microtubules (PubMed:18591255). May counteract a default induction of apoptosis in G2/M phase (PubMed:9859993). The acetylated form represses STAT3 transactivation of target gene promoters (PubMed:20826784). May play a role in neoplasia (PubMed:10626797). Inhibitor of CASP3 and CASP7 (PubMed:21536684). Essential for the maintenance of mitochondrial integrity and function (PubMed:25778398). Isoform 2 and isoform 3 do not appear to play vital roles in mitosis (PubMed:12773388, PubMed:16291752). Isoform 3 shows a marked reduction in its anti-apoptotic effects when compared with the displayed wild-type isoform (PubMed:10626797). 13 Publications

Sites

翻译后修饰



Amino acid modifications

Feature key	Position(s)	Description	Actions	Graphical view	Length
Modified residue ⁱ	20	Phosphoserine; by AURKC 1 Publication			1
Modified residue ⁱ	23	N6-acetyllysine 1 Publication			1
Modified residue ⁱ	34	Phosphothreonine; by CDK1 and CDK15 Combined sources 2 Publications			1
Modified residue ⁱ	48	Phosphothreonine; by CK2; in vitro 1 Publication			1
Modified residue ⁱ	90	N6-acetyllysine 1 Publication			1
Modified residue ⁱ	110	N6-acetyllysine 1 Publication			1
Modified residue ⁱ	112	N6-acetyllysine 1 Publication			1
Modified residue ⁱ	115	N6-acetyllysine 1 Publication			1
Modified residue ⁱ	117	Phosphothreonine; by AURKB 1 Publication			1
Modified residue ⁱ	121	N6-acetyllysine 1 Publication			1
Modified residue ⁱ	129	N6-acetyllysine 1 Publication			1

Post-translational modificationⁱ

Ubiquitinated by the Cul9-RING ubiquitin-protein ligase complex, leading to its degradation. Ubiquitination is required for centrosomal targeting.

[2 Publications](#)

In vitro phosphorylation at Thr-117 by AURKB prevents interaction with INCENP and localization to mitotic chromosomes (PubMed:[14610074](#)). Phosphorylation at Thr-48 by CK2 is critical for its mitotic and anti-apoptotic activities (PubMed:[21252625](#)). Phosphorylation at Thr-34 by CDK15 is critical for its anti-apoptotic activity (PubMed:[24866247](#)). Phosphorylation at Ser-20 by AURKC is critical for regulation of proper chromosome alignment and segregation, and possibly cytokinesis. [5 Publications](#)

Acetylation at Lys-129 by CBP results in its homodimerization, while deacetylation promotes the formation of monomers which heterodimerize with XPO1/CRM1 which facilitates its nuclear export. The acetylated form represses STAT3 transactivation. The dynamic equilibrium between its acetylation and deacetylation at Lys-129 determines its interaction with XPO1/CRM1, its subsequent subcellular localization, and its ability to inhibit STAT3 transactivation. [1 Publication](#)

KEGG: Pathway



□ <http://www.genome.jp/kegg/kegg2.html>



KEGG - Table of Contents

Menu PATHWAY BRITE MODULE KO GENES LIGAND NETWORK DISEASE DRUG DBGET

Search

KEGG ▼

for

BIRC5

Go

Data-oriented entry points

Category	Entry Point	Content	DBGET Search
Systems information	KEGG PATHWAY	KEGG pathway maps	PATHWAY
	KEGG BRITE	BRITE hierarchies and tables	BRITE
	KEGG MODULE	KEGG modules	MODULE
Genomic information	KO (KEGG Orthology)	Functional orthologs	ORTHOLOGY
	KEGG GENOME	KEGG organisms (complete genomes)	GENOME
	KEGG GENES	Genes and proteins	GENES
	KEGG SSDB	GENES sequence similarity	

Human BIRC5



□ 参与通路：癌症、直肠癌、乙肝

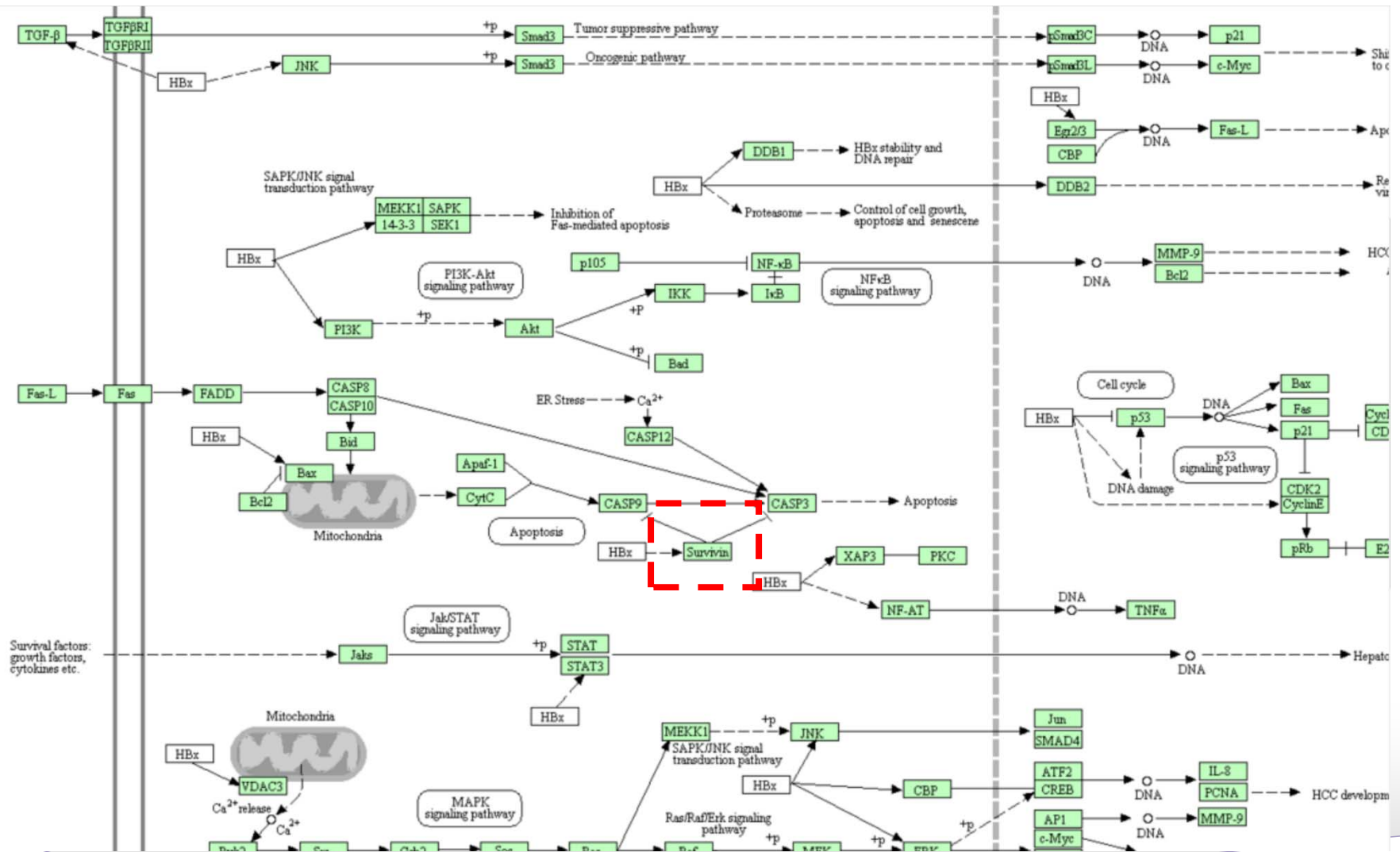


Homo sapiens (human): 332

Help

Entry	332	CDS	T01001
Gene name	BIRC5, API4, EPR-1		
Definition	(RefSeq) baculoviral IAP repeat containing 5		
KO	K08731 baculoviral IAP repeat-containing protein 5		
Organism	hsa Homo sapiens (human)		
Pathway	hsa01524 Platinum drug resistance hsa04210 Apoptosis hsa04215 Apoptosis - multiple species hsa04390 Hippo signaling pathway hsa05161 Hepatitis B hsa05200 Pathways in cancer hsa05210 Colorectal cancer		
Network	nt06131 Apoptosis (viruses and bacteria) nt06162 Hepatitis B virus (HBV) nt06215 WNT signaling nt06260 Colorectal cancer nt06261 Gastric cancer nt06263 Hepatocellular carcinoma nt06271 Endometrial cancer		
Element	N00057 Mutation-inactivated APC to Wnt signaling pathway N00058 Mutation-activated CTNNB1 to Wnt signaling pathway N00533 HBV HBx to Crosstalk between extrinsic and intrinsic apoptotic pathways		

Hepatitis B (hsa05161)



结论4：功能分析



- ❑ 在肿瘤形成过程中可能起一定作用
- ❑ 阻碍G2/M期的细胞编程性凋亡
- ❑ Chromosomal passenger complex (CPC) 的成员之一
- ❑ 细胞亚定位：中心粒、染色体、胞质
- ❑ 参与的生物学通路：癌症、直肠癌、乙肝

人类BIRC5在酵母中有同源序列吗？



UniProtKB - O15392 (BIRC5_HUMAN)

Display

- Entry
- Publications
- Feature viewer
- Feature table

Protein | Baculovirus
Gene | BIRC5
Organism | Homo sapiens
Status | Reviewed

Function

View format

- Text
- FASTA (canonical)
- FASTA (canonical & isoform)
- XML
- RDF/XML
- GFF

获得序列

← → ↻ 🔒 uniprot.org/uniprot/O15392.fasta

```
>sp|O15392|BIRC5_HUMAN Baculoviral IAP repeat-containing protein 5 OS=Homo sapiens OX=9606 GN=BIRC5 PE=1 SV=3
MGAPTLPPAWQPFLKDHRISTFKNWPFLGCACTPERMAEAGFIHCPTENEPDLAQCFFC
FKELEGWEPDDDPIDEEHKKHSSGCAFLSVKKQFEELTLGEFLKLDREKAKNKIAKETNNK
KKEFEETAKKVRRRAIEQLAAMD
```

在酵母中进行序列比对



web.expasy.org/blast/

Enter a sequence

Examples

>sp|Q15392|BIRC5_HUMAN Baculoviral IAP repeat-containing protein 5 OS=Homo sapiens OX=9606 GN=BIRC5 PE=1 SV=3
MGAPTLPPAWQPFLLKDHRISTFKNWPFLGCACTPERMAEAGFIHCPTENEPLI
FKELEGWEPDDPIEEHKKHSSGCAFLSVKKQFEELTLGEFLKIDRFRAKNKI
KKEFEETAKKVRRATEQLAAMD

e.g. P00750, P05067-5, A4_HUMAN or acccgtggtc

Run BLAST Reset

Mammalia

Metazoa

Primates

Rodentia

Vertebrata

Viridiplantae

Arabidopsis thaliana

Caenorhabditis elegans

Dictyostelium discoideum

Drosophila melanogaster

Escherichia coli

Homo sapiens

Mus musculus

Plasmodium falciparum

Rattus norvegicus

Saccharomyces cerevisiae

Schizosaccharomyces pombe

an also use this tool programmatically...

le on ExPASy: ?

protein sequence database

nucleotide sequence database

nucleotide sequence database

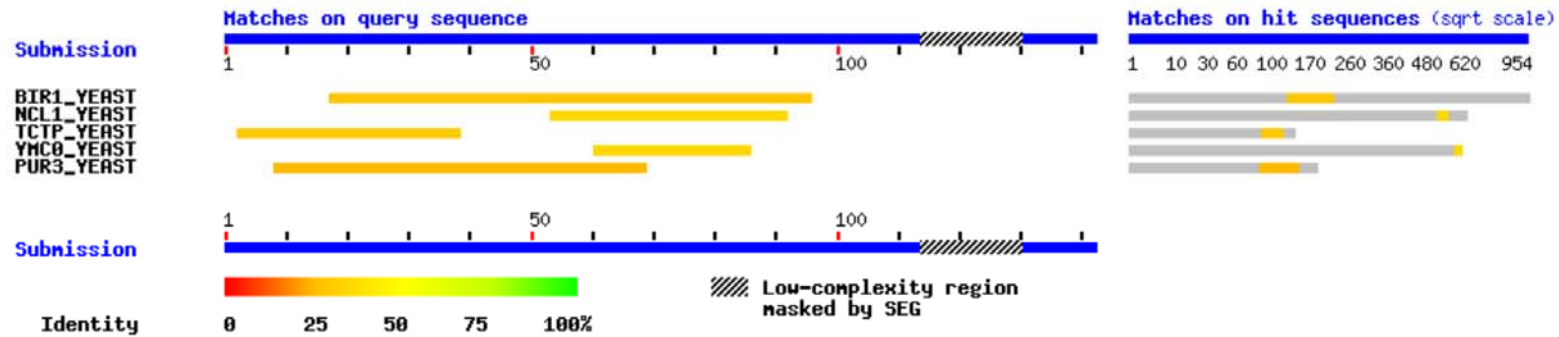
protein sequence database

☐ UniProtKB/Swiss-Prot only. ?

酵母BIR1: P47134



Graphical overview of the alignments



List of the matches

Clustal W (multiple alignment) 提交

☐ Select up to...

☐ Include query sequence

	Accession	Length	Db	Description	Score	E-value
☒	1 P47134 (BIR1_YEAST)	954		Protein BIR1 OS=Saccharomyces cerevisiae (s...	39.7 bits (91)	7e-05
☐	2 P38205 (NCL1_YEAST)	684		Multisite-specific tRNA:(cytosine-C(5))-met...	30.8 bits (68)	0.070
☐	3 P35691 (TCTP_YEAST)	167		Translationally-controlled tumor protein ho...	26.2 bits (56)	1.7
☐	4 Q03722 (YMC0_YEAST)	664		Uncharacterized protein YML020W OS=Saccharo...	24.6 bits (52)	7.7
☐	5 P04161 (PUR3_YEAST)	214		Phosphoribosylglycinamide formyltransferase...	24.3 bits (51)	8.4

酵母BIR1的信息



□ 人类BIRC5在酵母中的同源序列可能是BIR1

← → ↻ uniprot.org/uniprot/P47134

UniProtKB Advanced

BLAST Align Retrieve/ID mapping Peptide search Help Contact

UniProtKB - P47134 (BIR1_YEAST)

Display

Entry **Protein** **Protein BIR1**
Gene **BIR1**
Organism *Saccharomyces cerevisiae* (strain ATCC 204508 / S288c) (Baker's yeast)
Status **Reviewed** - Annotation score: ●●●●● - Experimental evidence at protein levelⁱ

Functionⁱ
Seems to act in the pleiotropic control of cell division. May participate in chromosome segregation events.

Sites

Feature key	Position(s)	Description	Actions	Graphical view	Length
Metal binding ⁱ	208	Zinc			1
Metal binding ⁱ	211	Zinc			1
Metal binding ⁱ	228	Zinc			1
Metal binding ⁱ	237	Zinc			1

None

- ☒ Function
- ☒ Names & Taxonomy
- ☒ Subcellular location
- ☐ Pathology & Biotech
- ☒ PTM / Processing
- ☐ Expression
- ☒ Interaction

PDB: 三级结构数据库



<http://www.uniprot.org/uniprot/O15392>

← → ↻ 🔒 uniprot.org/uniprot/O15392#structure

Display Structureⁱ

Entry

Publications

Feature viewer

Feature table

None

- ☒ Function
- ☒ Names & Taxonomy
- ☒ Subcellular location
- ☒ Pathology & Biotech
- ☒ PTM / Processing
- ☒ Expression
- ☒ Interaction
- ☒ Structure
- ☒ Family & Domains
- ☒ Sequences (7+)
- ☒ Similar proteins



PDB Entry	Method	Resolution	Chain	Positions	Links
1E31	X-ray	2.71 Å	A/B	1-142	PDBe RCSB ... PDBj PDBsum
1F3H	X-ray	2.58 Å	A/B	1-142	PDBe RCSB ... PDBj PDBsum
1XOX	NMR		A/B	1-117	PDBe RCSB ... PDBj PDBsum
2QFA	X-ray	1.40 Å	A	1-142	PDBe RCSB ... PDBj PDBsum
2RAW	X-ray	2.40 Å	A	1-142	PDBe RCSB ... PDBj PDBsum
2RAX	X-ray	3.30 Å	A/F/X	1-120	PDBe

BIRC5的三级结构信息



← → ↻ rcsb.org/structure/1E31

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

PDB-101 WORLDWIDE PDB PROTEIN DATA BANK EMDatabank Unified Data Resource for 3DEM ndb NUCLEIC ACID DATABASE Worldwide Protein Data Bank Foundation

Structure Summary 3D View Annotations Sequence Sequence Similarity Structure Similarity Experiment

Biological Assembly 1 ?

1E31
SURVIVIN DIMER H. SAPIENS
DOI: [10.2210/pdb1E31/pdb](https://doi.org/10.2210/pdb1E31/pdb)
Classification: [APOPTOSIS INHIBITOR](#)
Organism(s): [Homo sapiens](#)
Expression System: [Escherichia coli](#)
Deposited: 2000-06-04 Released: 2001-01-03
Deposition Author(s): [Chantalat, L.](#), [Skoufias, D.A.](#), [Margolis, R.L.](#), [Dideb](#)

Experimental Data Snapshot
Method: X-RAY DIFFRACTION
Resolution: 2.71 Å
R-Value Free: 0.270
R-Value Work: 0.235

3D View: [Structure](#) | [Ligand Interaction](#)

Standalone Viewers
[Protein Workshop](#) | [Ligand Explorer](#)

Display Files Download Files

- FASTA Sequence
- PDB Format
- PDB Format (gz)
- PDBx/mmCIF Format
- PDBx/mmCIF Format (gz)
- PDBML/XML Format (gz)
- Biological Assembly 1

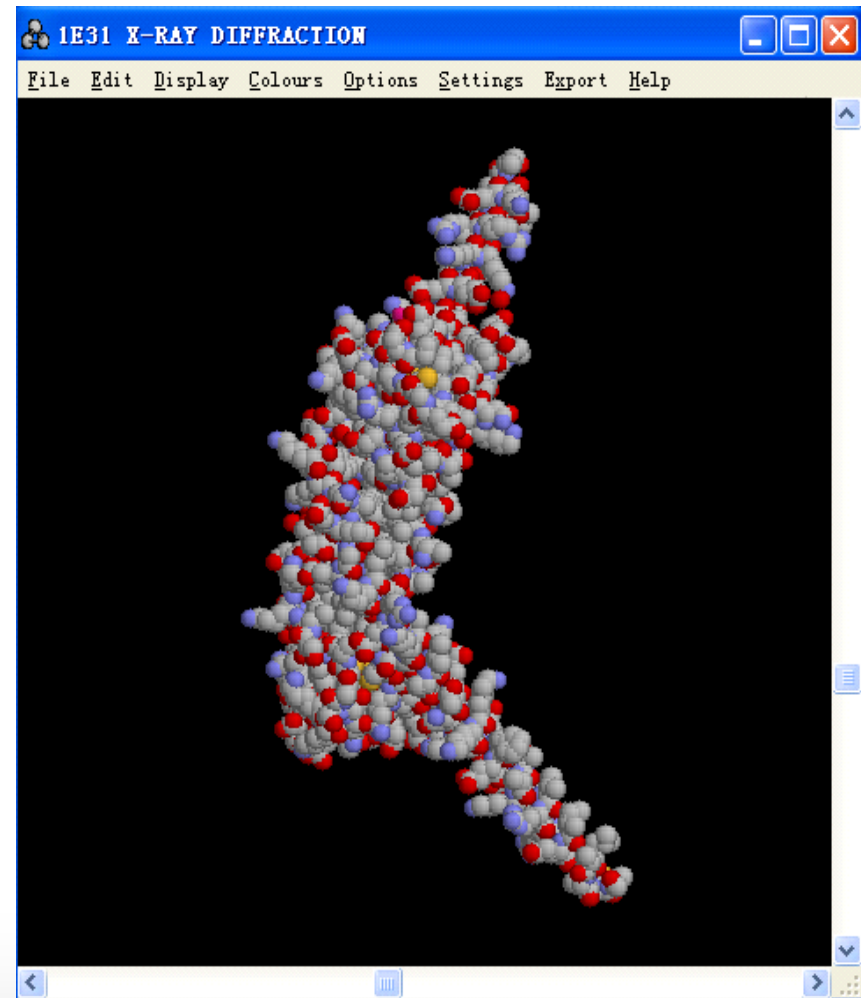
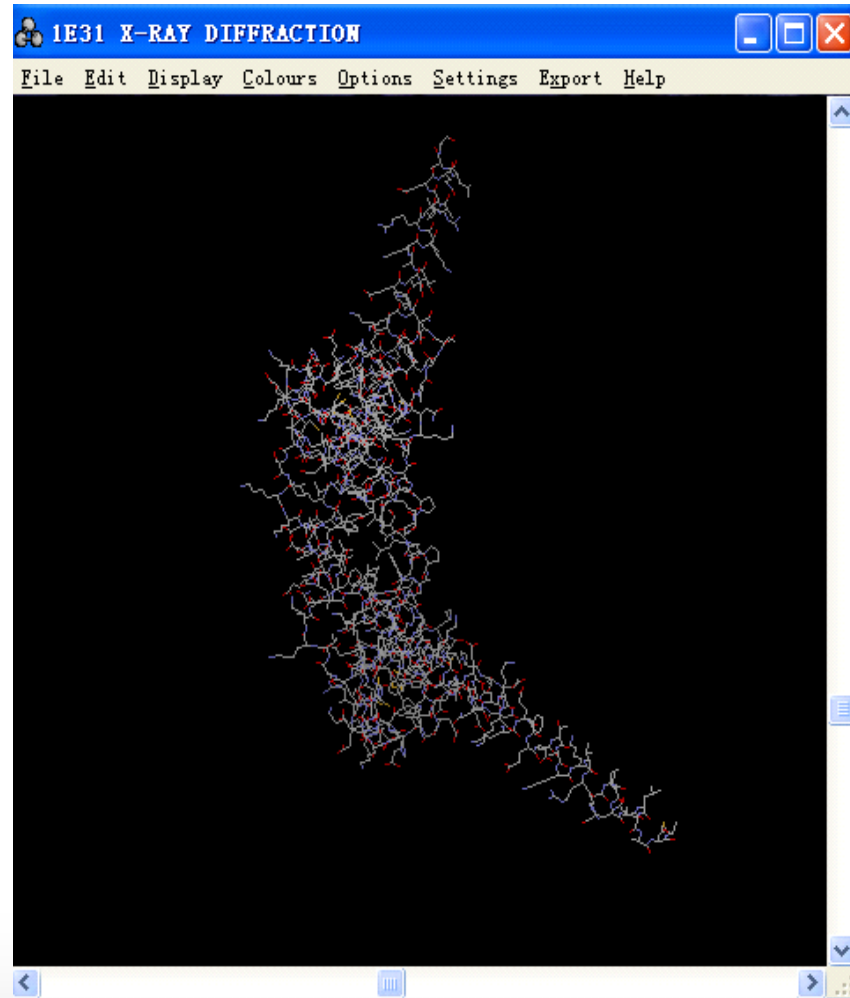
wwwPDB Validation

Metric	Percentile Ranks	Value
Clashscore		13
Ramachandran outliers		3.0%
Sidechain outliers		6.7%

Worse Better
■ Percentile relative to all X-ray structures
□ Percentile relative to X-ray structures of similar resolution

Contact Us

Raswin: 三级结构显示



总结



- ❑ 1. 该基因为人类BIRC5基因，染色体定位：17q25.3，78214253-78225635；基因标识符：NM_001168.3
- ❑ 2. 人的BIRC5蛋白质包含142个氨基酸，序列标识符为：NP_001159.2
- ❑ 3. BIRC5具有保守的功能结构域BIR
- ❑ 4. BIRC5的细胞亚定位：中心粒、染色体等，其功能有：
 - ✿ (1) 在肿瘤形成过程中可能起一定作用
 - ✿ (2) 阻碍G2/M期的细胞编程性凋亡
 - ✿ (3) Chromosomal passenger complex (CPC) 的成员之一
 - ✿ (4) 参与的生物学通路：癌症、直肠癌、乙肝
- ❑ 5. 人的BIRC5在酵母中的同源序列可能是BIR1
- ❑ 6. BIRC5的三级结构已知，在PDB中的标识符为1E31

Homework 1#



- ❑ A mitosis-associated gene *X* was experimentally identified in *Mus musculus*. By DNA sequencing, a partial sequence was obtained as below:

```
GATGAGCTGCTTATCCTACAACGAGAAGTCGGACATCTGGTCCTTG
GGCTGCCTGCTGTATGAGCTGTGTGCACTAATGCCTCCCTTTACAG
CTTTCAACCAAAAAGAGCTAGCTGGGAAAATCAGGGAAGGGAGGT
TCAGGCGCATCCCCTACCGCTACTCTGATGGCTTGAATGACCTCAT
CACTCGGATGCTGAATTTAAAGGACTACCATCGACCTTCAGTGGAA
GAAATTCTGGAGAGCCCTTTGATAGCAGACTTGGTTGCAGAAGAGC
AAAGGAGAAATCTGGAGAGGAGAGGACGGCGCTCAGGCGAGCCT
TCGAAGCTGCCGGACTCCAGCCCTGTGCTGAGCGAGCTCAAGTTG
AAGGAAAGGCAACTGCAGGATCGAGAGCAAGCACTCAGAGCTCGG
GAGGACATCCT
```

Questions:



- ☐ 1. What's the name of gene *X* in *Mus musculus*? Its accession number in GenBank database and coordinates of mouse genome.
- ☐ 2. The homolog of gene *X* in human, and its accession number.
- ☐ 3. In human, the protein product of this gene. Its functions, sub-cellular localizations. Whether it's a enzyme? If so, does it have a conserved functional domain?
- ☐ 4. Whether this gene is conserved in yeast? If so, identify its potential homolog.
- ☐ 5. The 3D structural information of gene *X* in human, but not mouse. It's accession number in PDB.