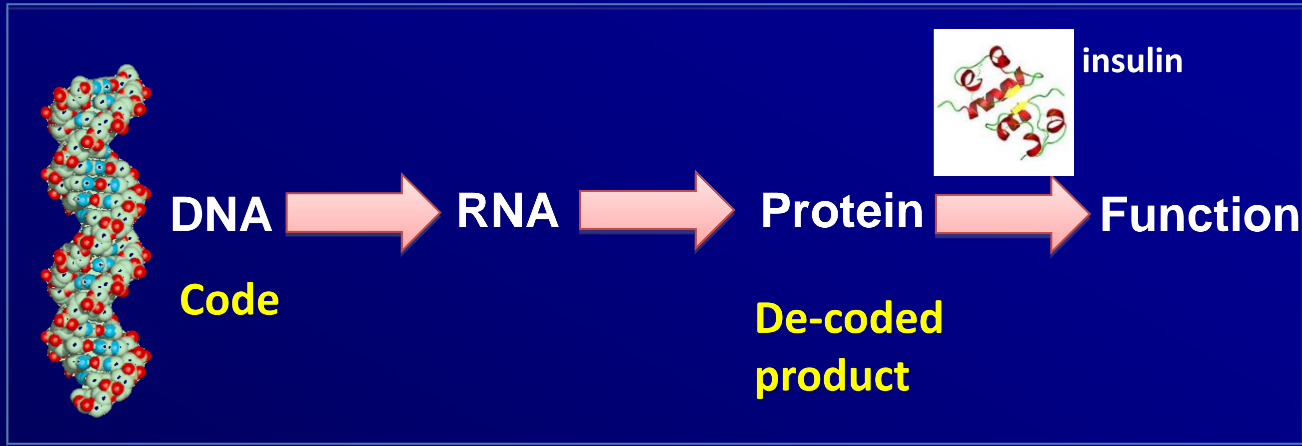
DNA Sequencing

李 余 动
lyd@zjsu.edu.cn

目　　录

- DNA测序基础
- Sanger法测序原理
- Sanger法测序数据分析

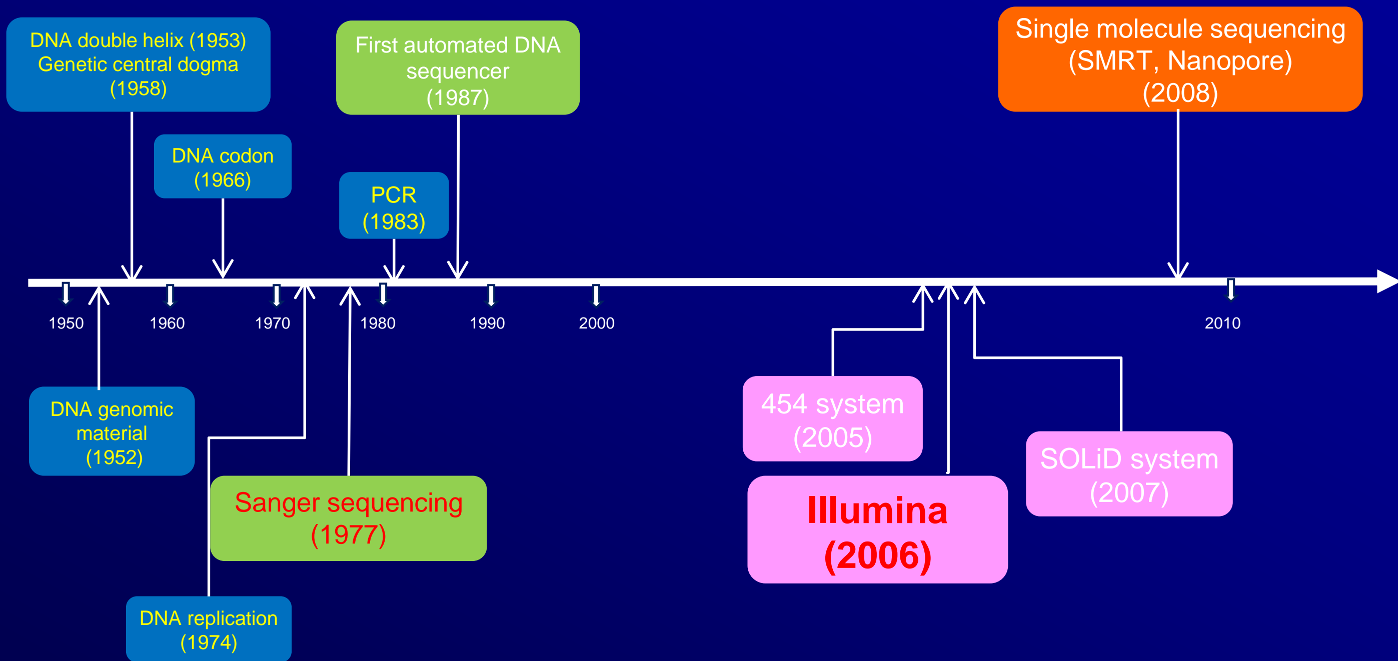
DNA to Protein



- Segments of DNA in the genome called “GENES” code for proteins （遗传信息编码者）
- Proteins are the functional end products in the cell (功能执行者)

第一部分：　DNA测序技术

DNA sequencing is the process of determining the sequence of nucleotides within a DNA molecule.



Next Generation Sequencing (NGS): 下一代测序

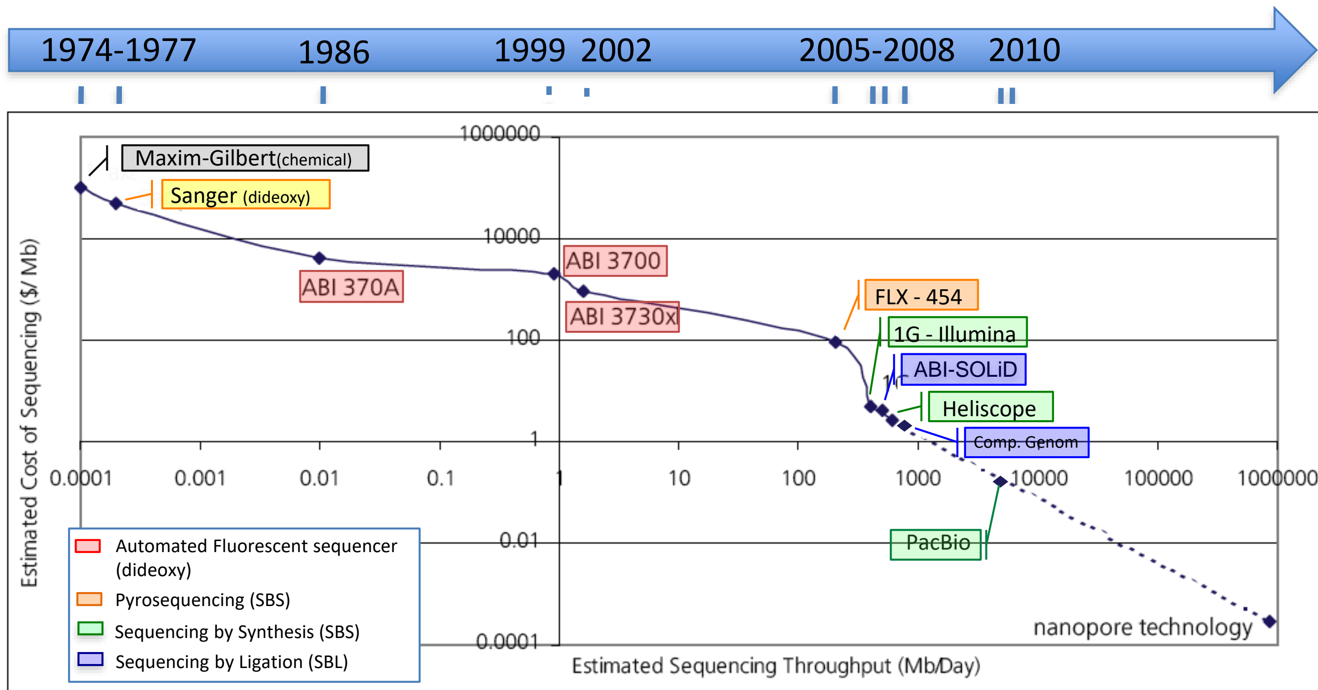
Sequencing technologies

- First generation sequencing technology
  - –Sanger Sequencing
- Second generation sequencing technology
  - –Illumina: HiSeq, MiSeq
- Third generation sequencing technology
  - –PacBio SMRT(单分子实时测序)
  - –Oxford Nanopore

Oxford Nanopore Technology

- Size of USB drive
- May drive the next revolution in genomics
- Whole genome sequencing in 15 minutes for less than $1,000
- because they rely on the detection of electronic, rather than optical, signals.

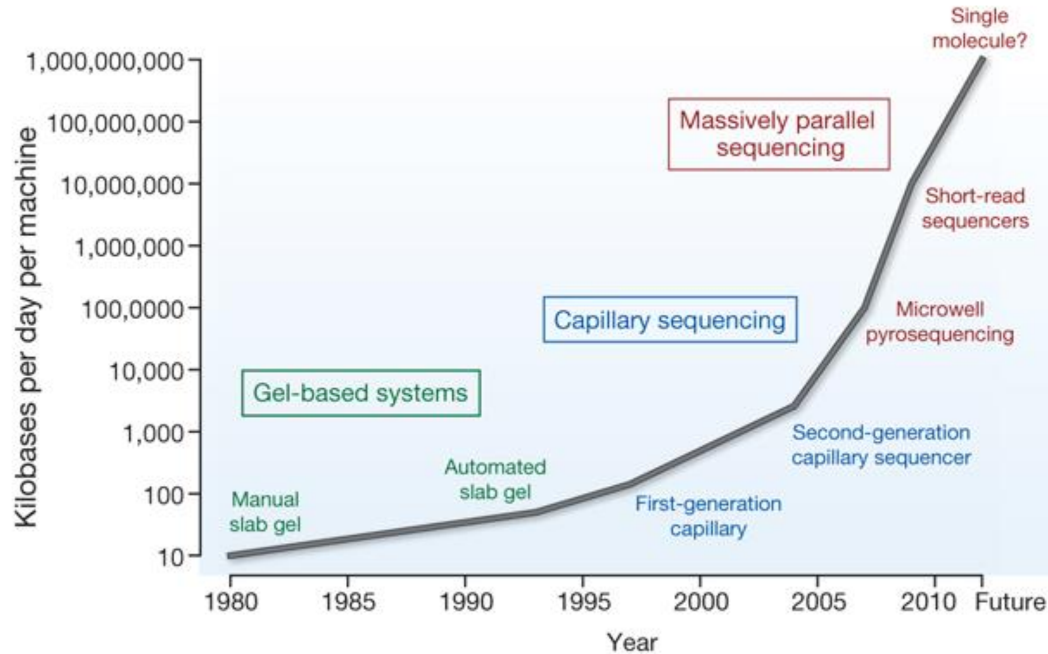
DNA Sequencing Cost vs Throughput Timeline



- 100 human genomes in <=10 days
- 99.99999% accurate with 98% coverage
- $10K/genome

Modified from UBS 2007

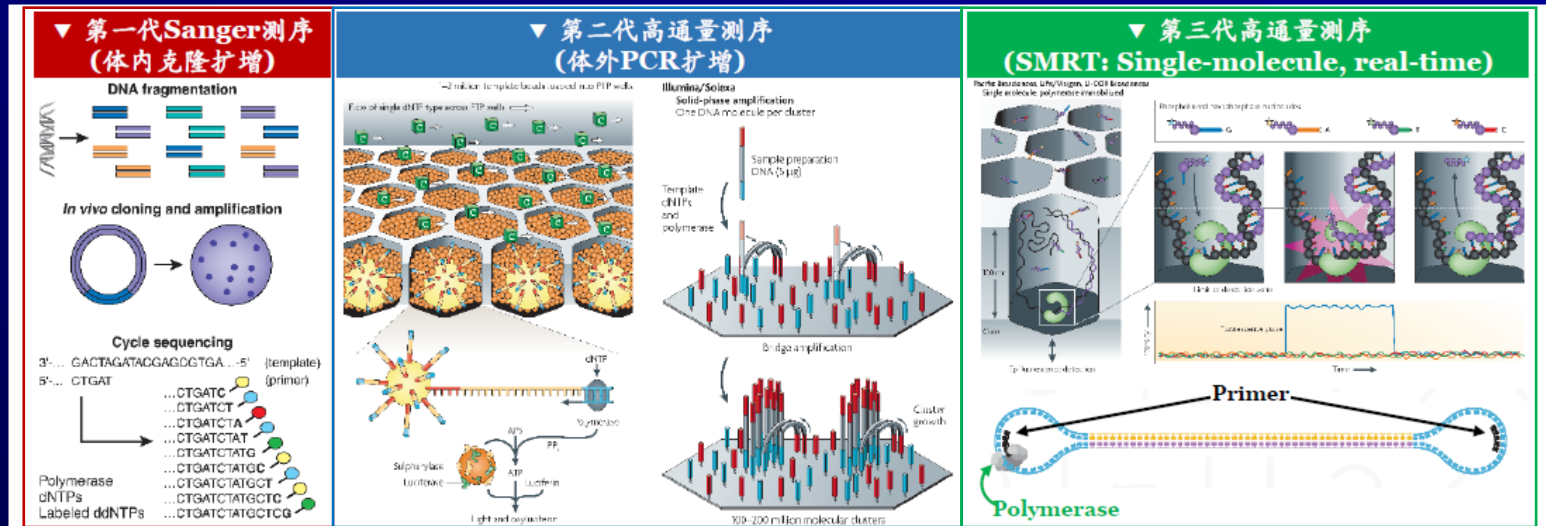
Improvements in the rate of DNA sequencing



MR Stratton *et al. Nature* **458**, 719-724 (2009) doi:10.1038/nature07943

Next Generation Sequencing (NGS) = High-throughput sequencing (HTS)

Next generation sequencer determines the bases of every DNA molecule.



Method	Generation	Read length (bp)	Single pass error rate	No. of reads per run	Time per run	Cost per million bases
Sanger ABI 3730xl	1st	600–1000	0.001%	96	0.5–3 h	$500
454 (Roche) GS FLX+	2nd	700	1%	1×10^6	23 h	$8.57
Illumina HiSeq 2500 (High Output)	2nd	2 × 125	0.1%	8×10^9 (paired)	7–60 h	$0.03
Illumina HiSeq 2500 (Rapid Run)	2nd	2 × 250	0.1%	1.2×10^9 (paired)	1–6 days	$0.04
Ion Torrent	2nd	200	1%	8.2×10^7	2–4 h	$0.1
SOLiD 5500xl	2nd	2 × 60	5%	8×10^8	6 days	$0.11
PacBio RS II: P6-C4	3rd	Avg. 10–15 k	13%	$3.5–7.5 \times 10^4$	0.5–4 h	$0.40–0.80
Oxford Nanopore MinION	3rd	Avg. 2–5 k	38%	$1.1–4.7 \times 10^4$	50 h	$6.44–17.90

(Shendure & Ji, 2008, Nature Biotechnology; Metzker, 2010, Nature Reviews Genetics; Rhoads & Au, 2015, Genomics Proteomics Bioinformatics)

DNA测序方法与应用

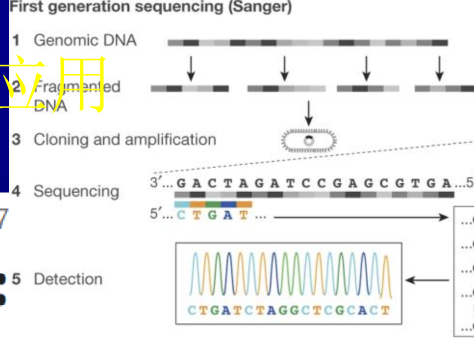
Review Article | Published: 11 October 2017

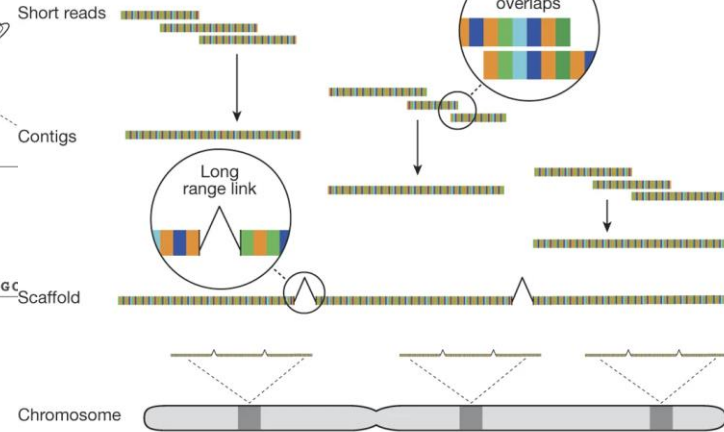
DNA sequencing at 40:

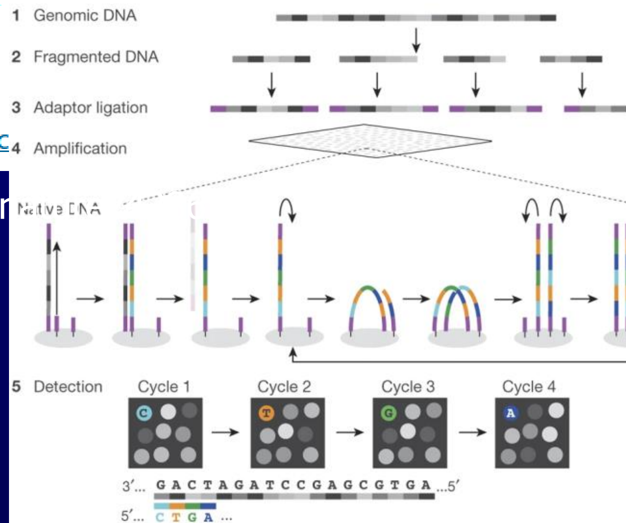
Jay Shendure, Shankar Balasubramanian,
Schloss & Robert H. Waterston

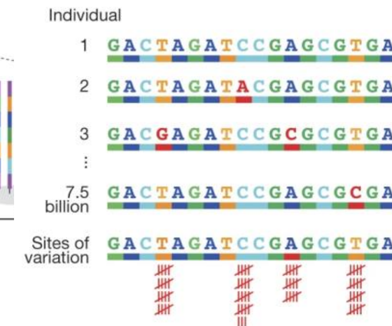
Nature 550, 345–353 (2017) | Cite this artic

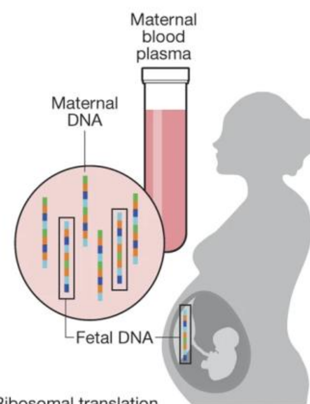
https://www.nature.com/articles/n

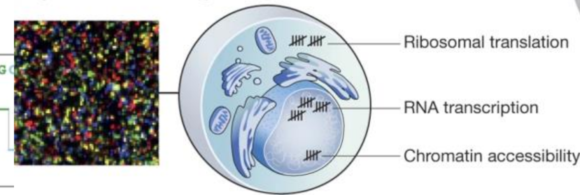


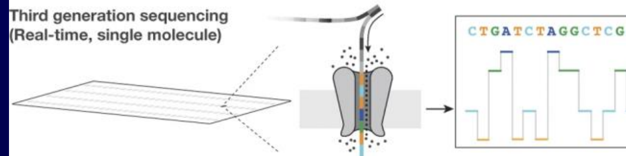








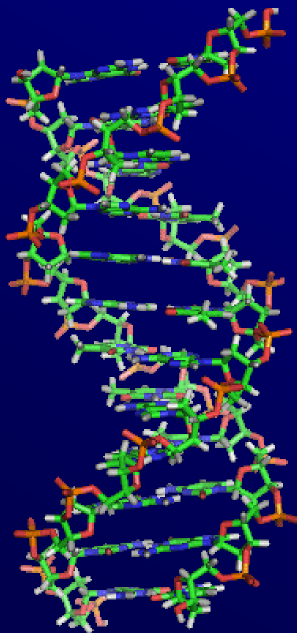


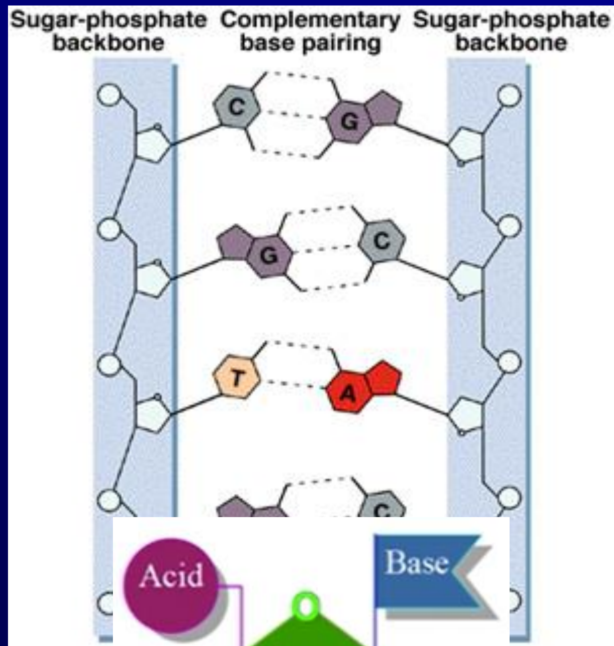


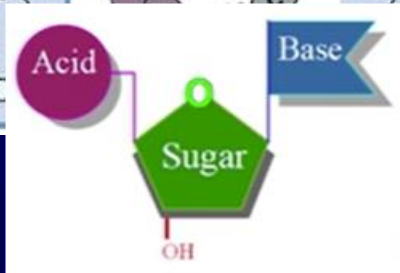
第二部分：　Sanger法测序原理

双脱氧链终止法（Sanger法）

DNA的双螺旋结构



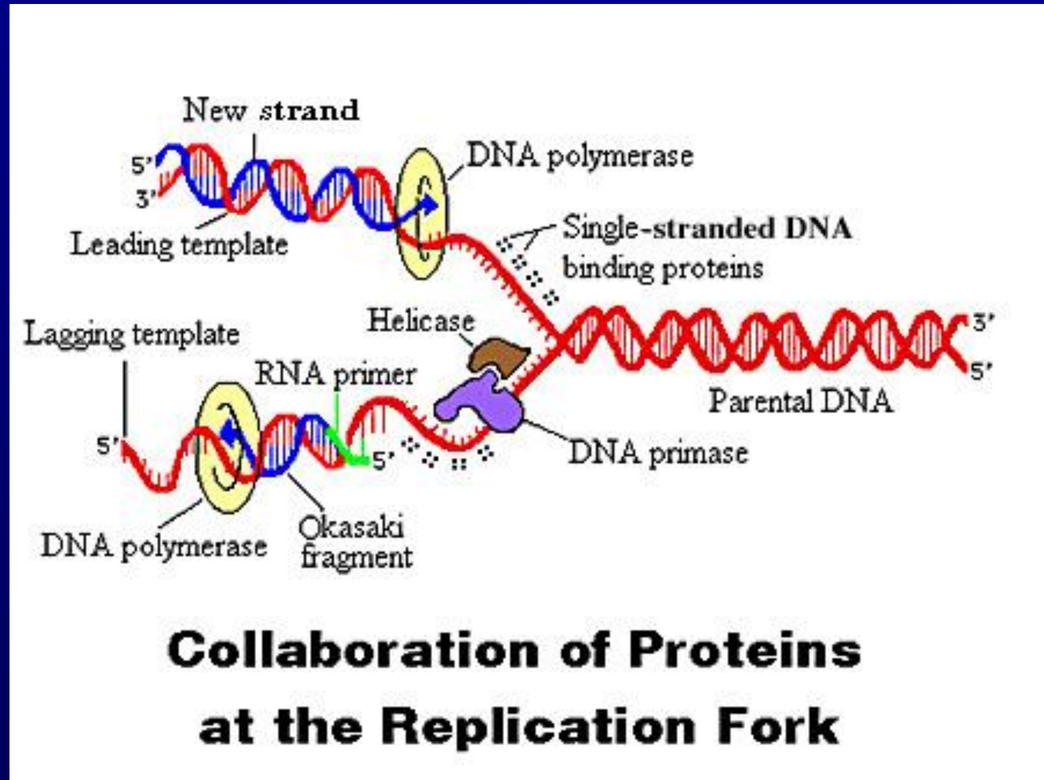




- 通过碱基互补配对原则，单链DNA形成双链结构。
- 碱基互补：
  - A:T
  - C:G
- 核苷酸分子包括：
  - 碱基
  - 核糖(3'-OH)
  - 磷酸

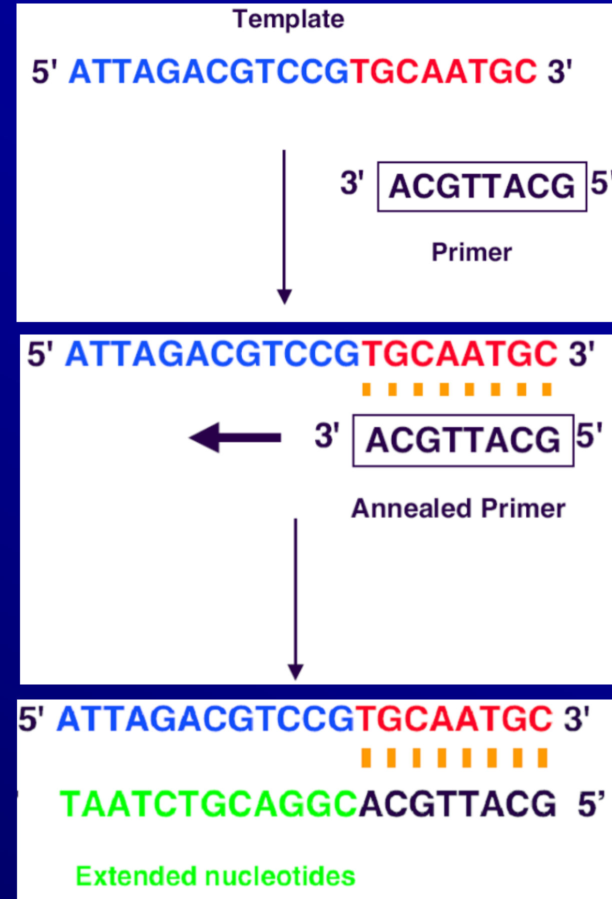
DNA半保留复制

DNA 在进行复制的时候链间氢键断裂，双链解旋分开，每条链作为模板在其上合成互补链，经过一系列酶（DNA聚合酶、解旋酶、连接酶等）的作用生成两个新的DNA分子。



DNA测序反应：引物延伸(primer-extension)

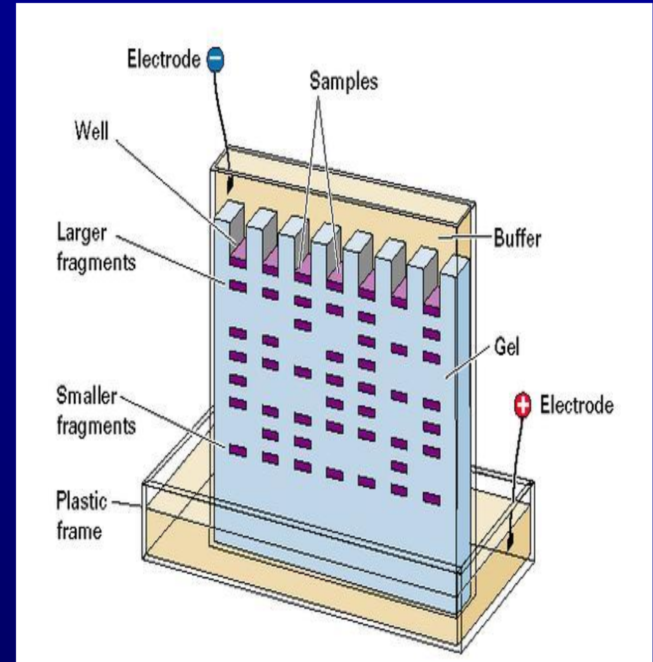
R. Wu (吴瑞) introduced the first method for DNA sequence analysis by introducing the primer-extension approach*



*Wu Ray. Nucleotide sequence analysis of DNA. I. Partial sequence of the cohesive ends of bacteriophage lambda and 186 DNA. J Mol Biol. 1970, 51(3):501-21.

电泳 (Electrophoresis)

- 变性聚丙烯酰胺凝胶电泳(SDS-PAGE)
- 在凝胶一端小槽中放入荧光或同位素标记的DNA片段，两端加电压，短DNA片段移动快，而长DNA片段移动慢。
- 测序时需要区分长度只差一个碱基的DNA片段。

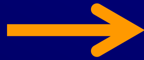


双脱氧链终止法（Sanger法）

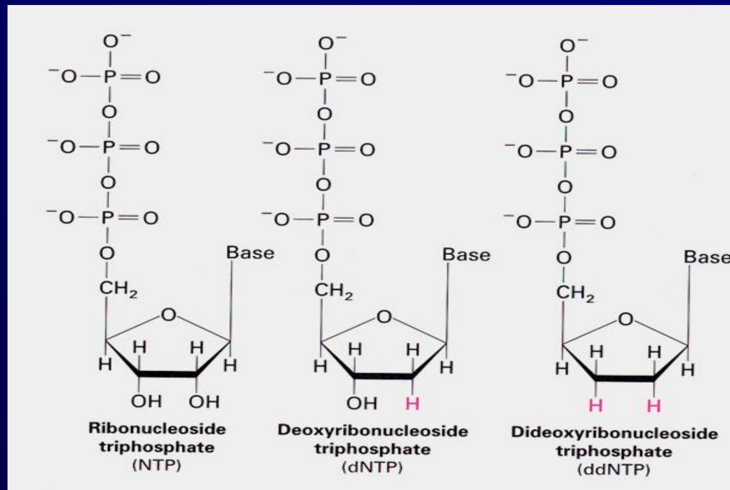
- 1953年，Frederick Sanger 发明了双脱氧核苷酸末端终止测序法。
  - 此方法获1974年的Nobel奖
- 这是第一代DNA测序技术，该方法被用于人类基因组计划的测序。

Sanger法测序的基本原理

Sequencing Primer



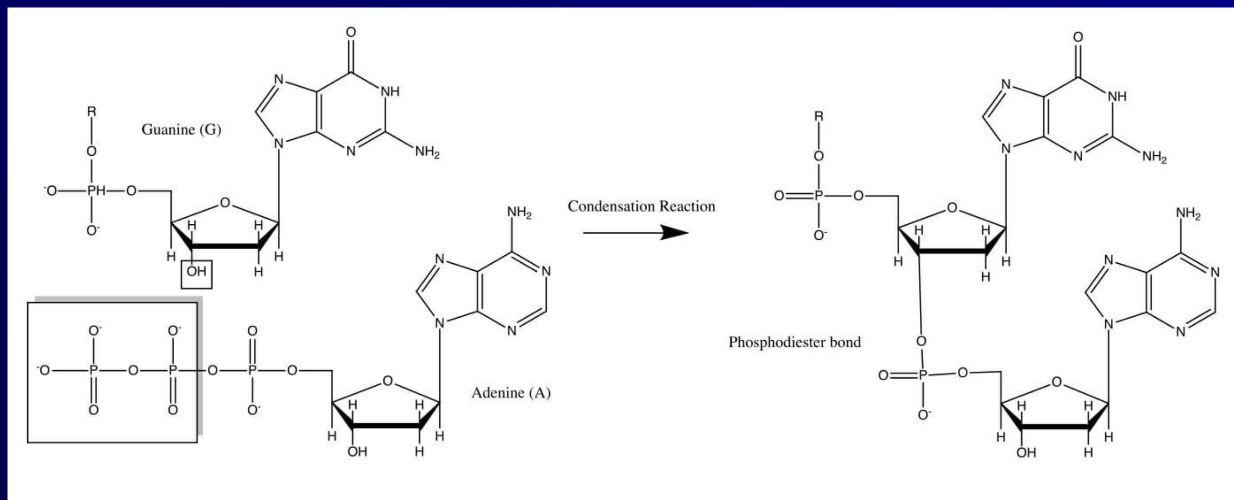
One sequencing reaction is about ~ 1000bp

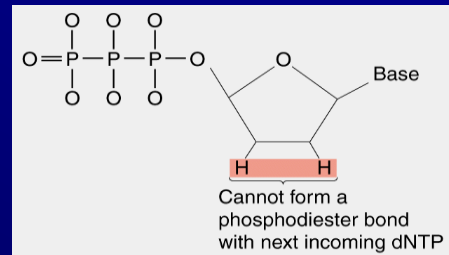


利用脱氧核苷三磷酸（dNTP）的类似物双脱氧核苷三磷酸（ddNTP）取代正常的底物。

Sanger法测序的基本原理

- 利用DNA聚合酶不能够区分dNTP和ddNTP的特性，使ddNTP参入到寡核苷酸链的3’-末端。
- 因为ddNTP 3’不是-OH，不能与下一个核苷酸聚合延伸，从而终止DNA链的增长。

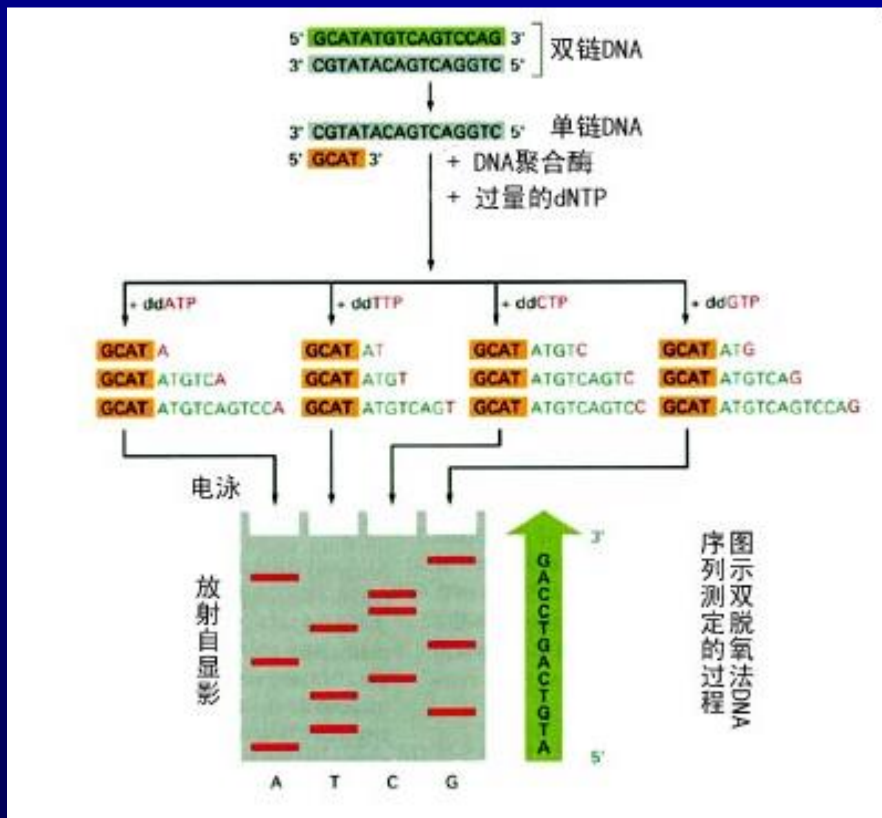




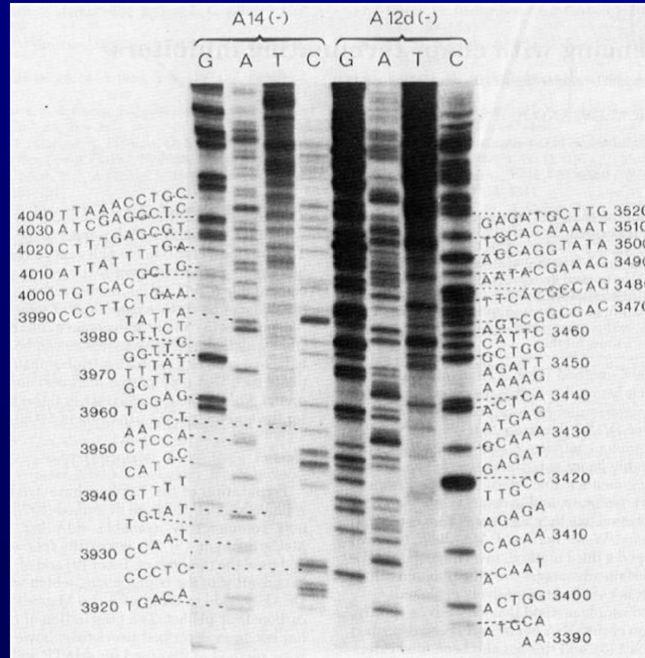
Sanger法测序的基本原理

Sanger法(链终止法)测序:

通过合成与单链DNA互补的核苷酸链，由于合成的互补链可在不同位置随机终止反应，产生只差一个核苷酸的DNA分子，从而来读取待测DNA分子的顺序。



碱基判读(base calling): SDS-PAGE电泳胶图

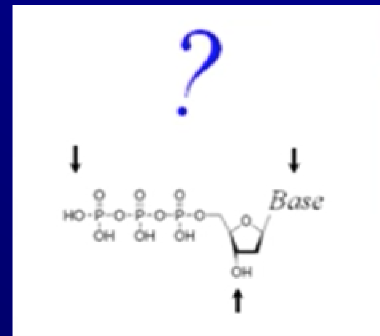


聚丙烯酰胺凝胶电泳(SDS-PAGE)可以区分长度只差一个核苷酸的DNA分子。

- 制得的四组混合物全部平行地点加在变性聚丙烯酰胺凝胶电泳板上进行电泳，
- 每组制品中的各个组分将按其链长的不同得到分离，
- 从而制得相应的放射性自显影图谱。
- 从所得图谱可直接读得DNA的碱基序列。

测序自动化

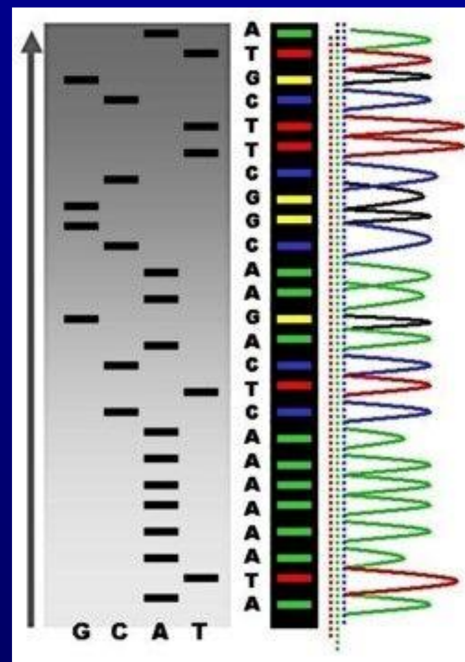
- 与链终止法测序原理相同，只是用不同的荧光色彩标记ddNTP, 如ddATP标记红色荧光, ddCTP标记蓝色荧光, ddGTP标记黄色荧光, ddTTP标记绿色荧光。
- 由于每种ddNTP带有各自特定的荧光颜色，而简化为由1个泳道同时判读4种碱基。



核苷酸分子上修饰的位置？

碱基判读(base calling): 测序峰图

- Base calling是把峰图数据转换成碱基序列：
  - 红色峰代表T，
  - 绿色峰代表A，
  - 蓝色峰代表C，
  - 黄色峰代表G。



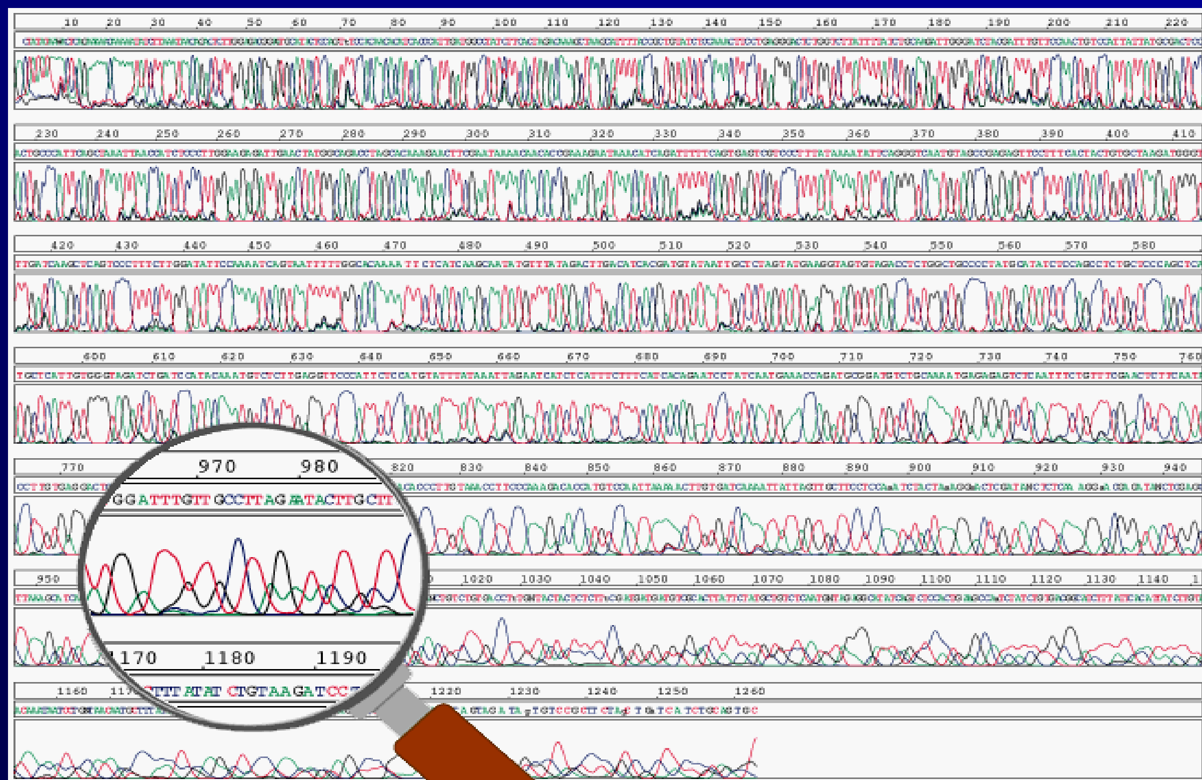
Sanger测序的胶图与峰图比较

全自动的测序仪器：ABI3730x

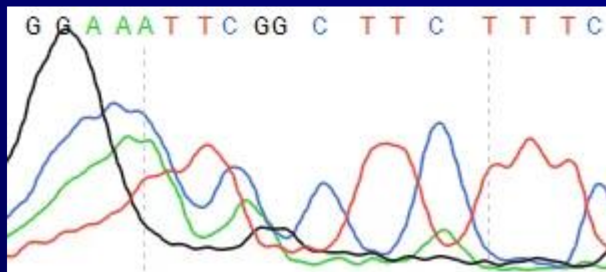
DNA自动测序仪的应用实现了凝胶电泳、初始数据获取、碱基阅读等步骤自动化。

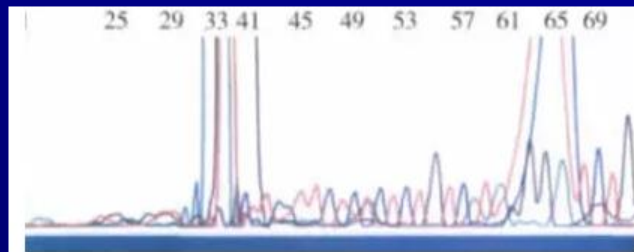
- 速度快：200bp/h
- 并行：同时测64~96样品
- 费用：20RMB/500bp
- 测序长度：500~1000bp

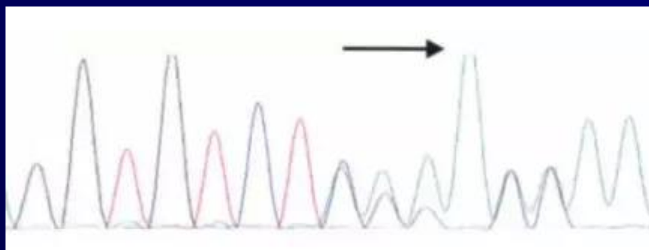
测序峰图(Chromatogram)



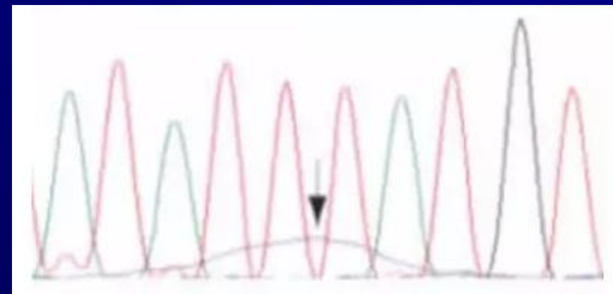
查看峰图，注意杂峰





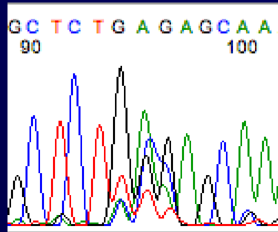


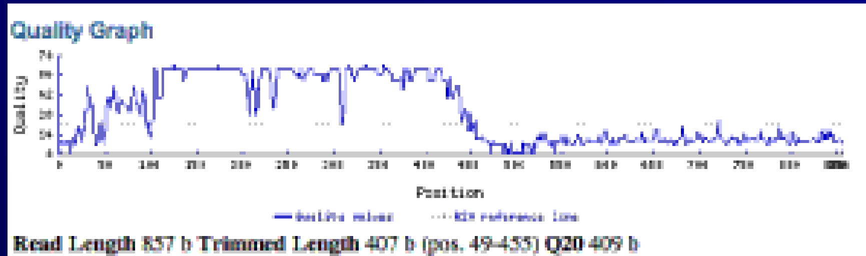
杂合峰



Phred Quality Scores

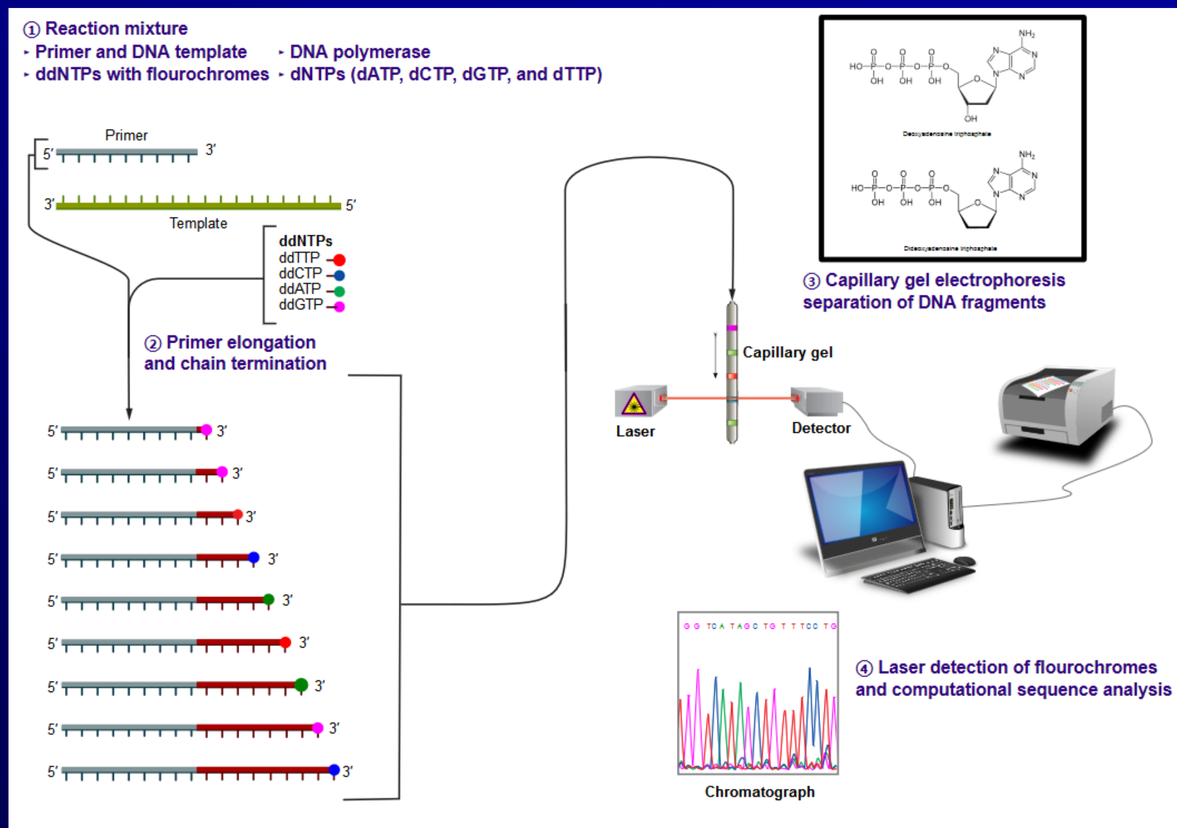
- Phred is a program that assigns a quality score to each base in a sequence. These scores can then be used to trim bad data from the ends, and to determine how good an overlap actually is.
  - there are much improved algorithms now, but the phred quality score is still widely used.
- Phred scores (Q) are logarithmically related to the probability of an error (Pe):
  - $Q = -10 \log_{10} P$, where Q is the phred score and P is the probability that the base was called incorrectly.
  - Example: Q=10 means a 10% error probability; 20 means a 1% chance, 30 means a 0.1% chance, etc.
  - A Q score of 20 is generally considered the minimum acceptable score.





Sanger法荧光自动测序

- 准备测序模板与引物
- 在PCR时加入荧光标记的复制终止剂，比如ddA, ddT, ddC, ddG（相应于4种碱基）
- ddNTP的两个作用：
  - 可以当作正常碱基参与复制
  - 一旦链入DNA中，其后就不能再继续连接
- 电泳
- 谁终止，碱基就是谁



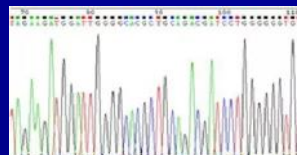
第三部分：Sanger测序分析流程

使用BioEdit或SnapGene软件

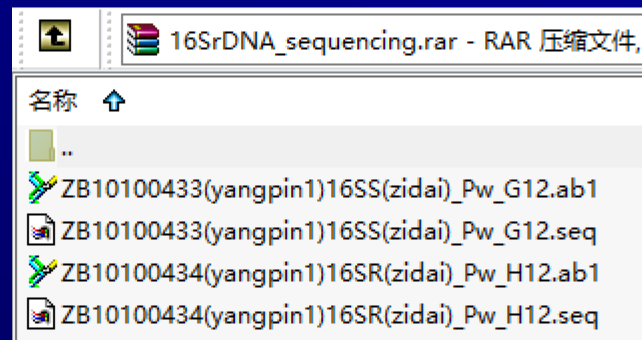
测序结果文件：

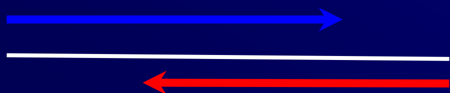
1. *.seq – 序列文件

2. *.ab1 — 峰图文件



前者可由记事本打开，而后者可由BioEdit或Chromas打开察看

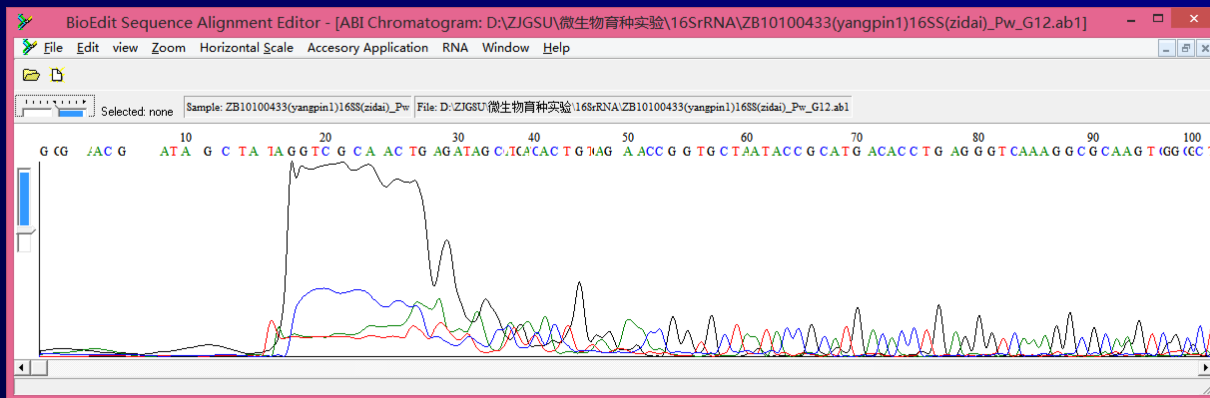




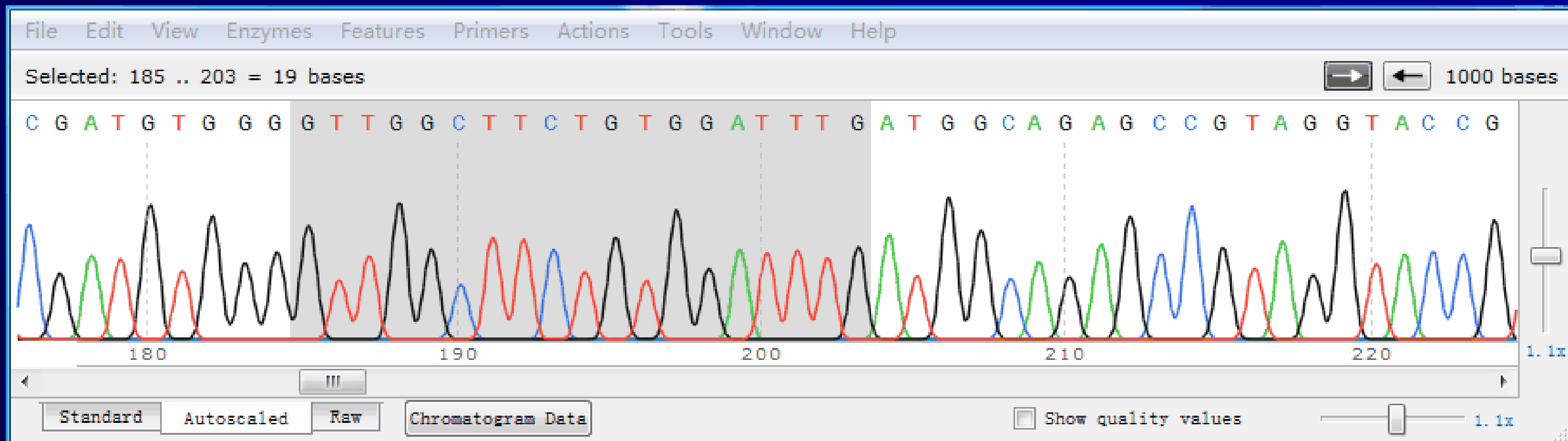
通常细菌16S rDNA的PCR产物要求双向测序，全长测通，使最终DNA序列比较准确。

测序结果分析

- 由于Sanger测序技术限制，每个测序反应仅有800bp左右比较准确。
- 测序的两端大约50个碱基的测序图部分通常杂质的干扰较大，无法判读，这是正常现象。
  - 可能是测序引物干扰、DNA聚合酶活性降低与杂质干扰较大等原因



- SnapGene软件不仅能够通过左右、上下滑块调节图形，并且能够实现反向互补的神奇操作。



切除两端低质量碱基

- 一般Sanger法测序reads的前端与末端碱基的质量会不好，此两部分测序峰图通常无法判读，需要把此两部分碱基切除，留下read中间碱基质量相对较好（峰图规则）的序列用于后续分析：
  - BioEdit 的Sequence菜单->select positions，在弹出窗口中输入56与950，点OK按钮后，就以背景黑的显示已选择的序列。
  - 再选edit菜单->Copy(或直接按Ctrl-C键)，复制序列到一个新的文本文件，保存为16S_rDNA.fasta。增加序列的注释行“>16S-F”(加F代表正向测序序列)。
  - 同上步骤，根据峰图信息，再复制另一个反向测序结果的高质量序列到文本文件16S_rDNA.fasta，并标记序列为“>16S-R”(R代表反向测序序列)。

切除两端低质量碱基后的序列

16S_rDNA.fasta [D:\ZJGSU\微生物育种实验\16SrDNA] - Notepad2-mod

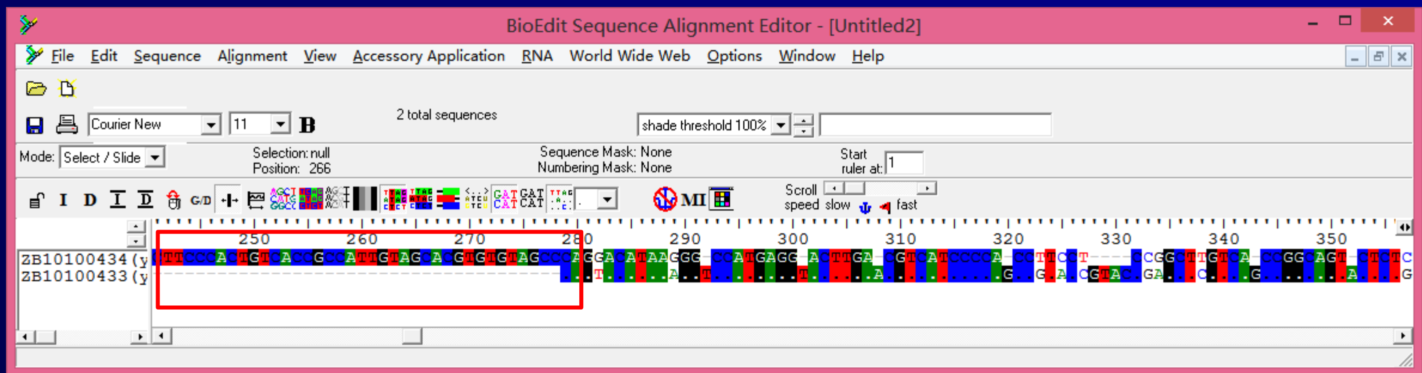
File Edit View Settings ?

```
>16S-F
ACCGGTGCTAATACCGCATGACACCTGAGGGTCAAAGGCGCAAGTCGGCGCTGTGGAGGAGCGCTGCGTTTGATTAGCTAGTTGGTGGGGTA
AGGGCCTACCAAGGCGATGATCAATAGCTGGTTTGAGAGGATGATCAGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAG
CAGTGGGGAATATTGGACAATGGGGGCAACCCTGATCCAGCAATGCCGCGTGTGTGAAGAAGGTCTTCGGATTGTAAAGCACTTTCGACGGG
GACGATGATGACGGTACCCGTAGAAGAAGCCCCGGCTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGGGGCTAGCGTTGCTCGGAATGA
CTGGGCGTAAAGGGCGTGTAGGCGGTTTGTACAGTCAGATGTGAAATCCCCGGGCTTAACCTGGGAGCTGCATTTGATACGTGCAGACTAGA
GTGTGAGAGAGGGTTGTGGAATTCGCAGTGTAGAGGTGAAATTCGTAGATATTGGGAAGAACACCGGTGGCGAAGGCGGCAACCTGGCTCAT
TACTGACGCTGAGGCGCGAAAGCGTGGGGAGCATACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGATGTGTGCTAGATGTTGGGTG
ACTTAGTCATTCAGTGTCGCAGTTAACGCGTGAAGCACACCGCCTGGGGAGTACGGCCGCAAGGTTGAAACTCAAAGGAATTGACGGGGGCC
CGCACAAGCGGTGGAGCATGTGGTATAATTCGAAGCAACGCGCAGAATCTTTACCAGGTCTTGAATGTAGAGGCTGCAAGCAGAGATGTGTG
TTTCCCGCAAGGGACCTCTAACACAGGTGCTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTATGTCCCAGCAACGAGCGCAAC
CTCCTATCTTTAGTTGCCATCAGGGTTGGGCTGGGGCACTCTACAGAGTACTGCCGGCTGACGAGCTCGG
>16S-R
TTCAGCTCCACCGGCTTAAGGTCAAACCAACTCCCATGGTGTGACGGGCGGTGTGTACAAGGCCCGGGAACGTATTCACCGCGGCATGCTGA
TCCGCGATTACTAGCGATTCCACCTTCATGCACTCGAGTTGCAGAGTGCAATCCGAACTGAGACGGCTTTTAGAGATCAGCATGGTGTCACC
ACCTAGCTTCCCACTGTCACCGCCATTGTAGCACGTGTGTAGCCCAGGACATAAGGGCCATGAGGACTTGACGTCATCCCCACCTTCCTCCG
GCTTGTCACCGGCAGTCTCTCTAGAGTGCCCAGCCCAACCTGATGGCAACTAAAGATAGGGGTTGCGCTCGTTGCGGGACTTAACCCAACAT
CTCACGACACGAGCTGACGACAGCCATGCAGCACCTGTGTTAGAGGTCCCTTGCGGGAAACAAACATCTCTGCTTGCAGCCTCTACATTCAA
GCCCTGGTAAGGTTCTGCGCGTTGCTTCGAATTAAACCACATGCTCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTCAACCTTG
CGGCCGTACTCCCCAGGCGGTGTGCTTAACGCGTTAACTGCGACACTGAATGACTAAGTCACCCAACATCTAGCACACATCGTTTACAGCGT
GGACTACCAGGGTATCTAATCCTGTTTGCTCCCCACGCTTTCGCGCCTCAGCGTCAGTAATGAGCCAGGTTGCCGCCTTCGCCACCGGTGTT
CTTCCCAATATCTACGAATTTCACCTCTACACTGGGAATTCCACAACCCTCTCTCACACTCTAGTCTGCACGTATCAAATGCAGCTCCCAGG
TTAAGCCCGGGGATTTCACATCTGACTGTACAAACCGCCTACACGCCCTTTACGCTCAGTCATTCCGAGCACGCTAGTCCCTCTTCGTATTA
CCGCGCTGCTGCACGAAGTTAGCCGGGGCTTCTTCTACGGTACGTCATCATCGTTCCTCGTCGAAGTGCT
```

Ln 4 / 4 Col 991 Sel 0 Sel Ln 0 | 1.95 KB | UTF-8 | CR+LF INS | Default Text

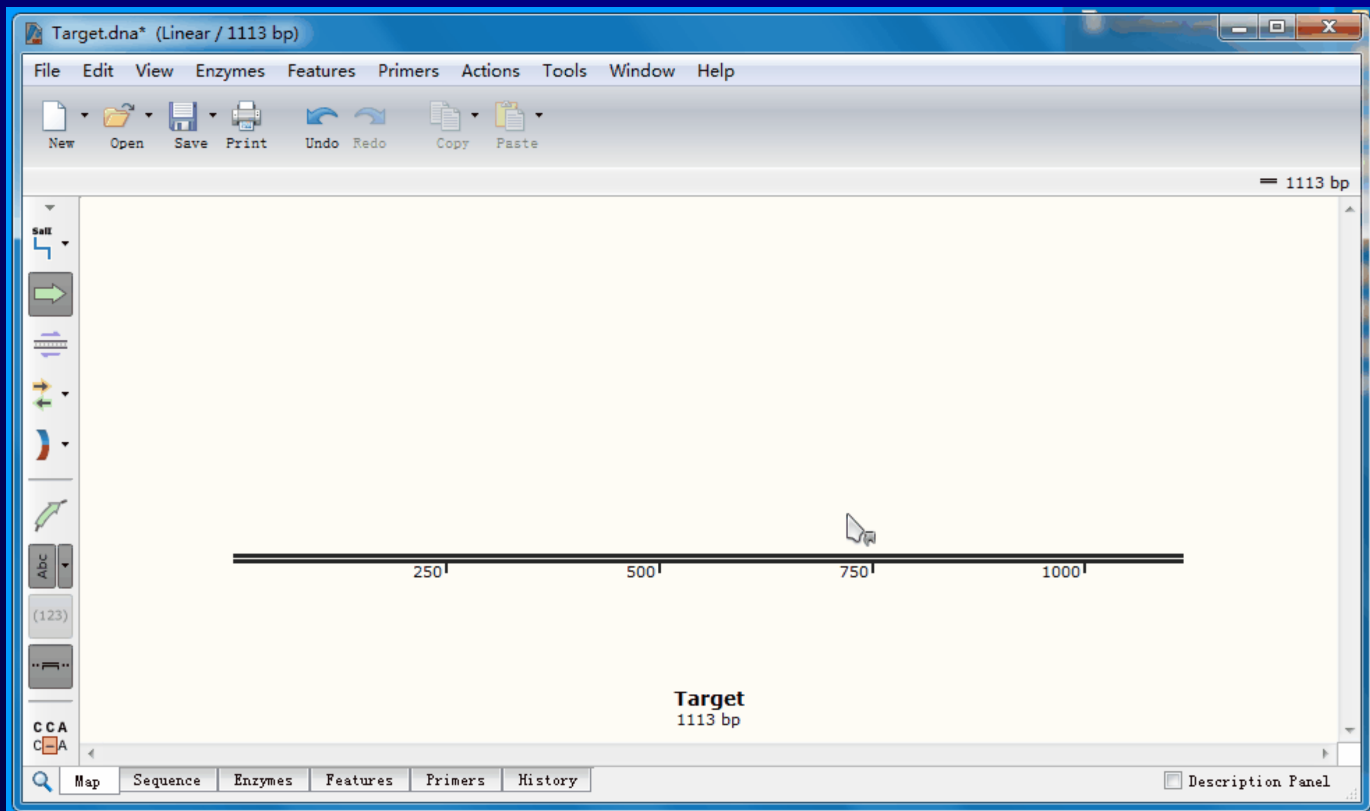
两端测序序列的合并

- 利用BioEdit的alignment功能找重叠区域:
  - Accessory application->ClustalW multiple alignment
- 比对结果不好，再试一下反向互补序列:
  - Sequence->Nuleic Acid->Reverse Complement



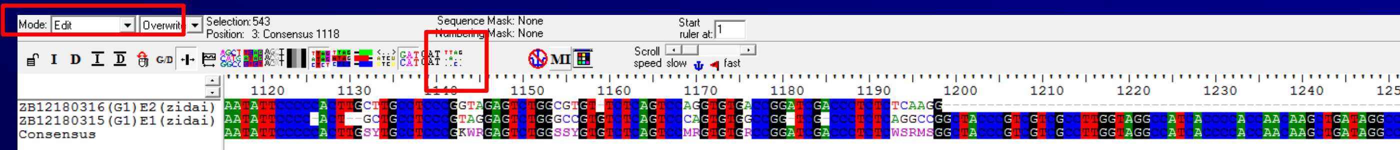
注意：如果重叠区不是大部分碱基相同，且前后有大段空位(-)，可能公司给的测序序列不是反向互补的序列，再试一下将第二条序列反向互补再比对。

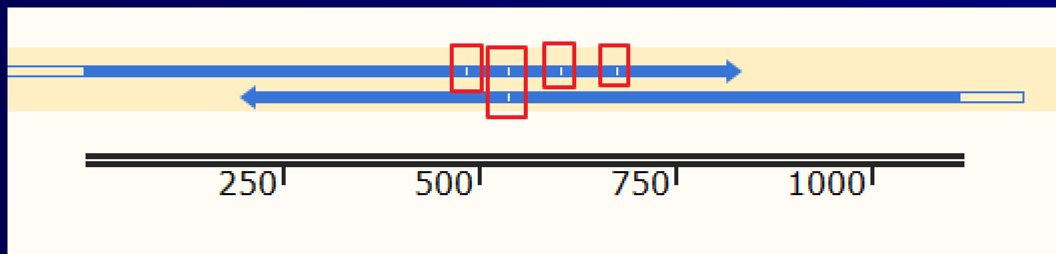
- SnapGene:点击Tools →align multiple sequences，选择正、反向测序的ab1文件，这样就得到了理论序列与正反向测序峰图的比对图：



两端测序序列的合并

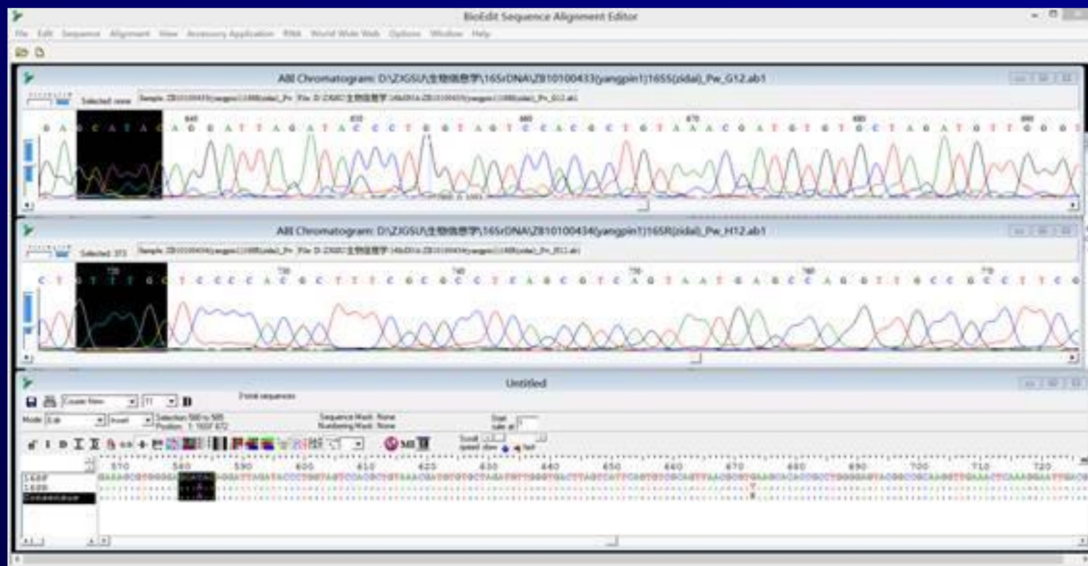
- 利用BioEdit得到一致序列
  - Alignment->Create Consensus Sequence
- 修改碱基前，需要先把bioedit的Mode设置为"Edit"与"Insert"，并选中按钮"view conservation plotting identities [...] with a dot"，以点显示相同的碱基，便于观察差异位点。



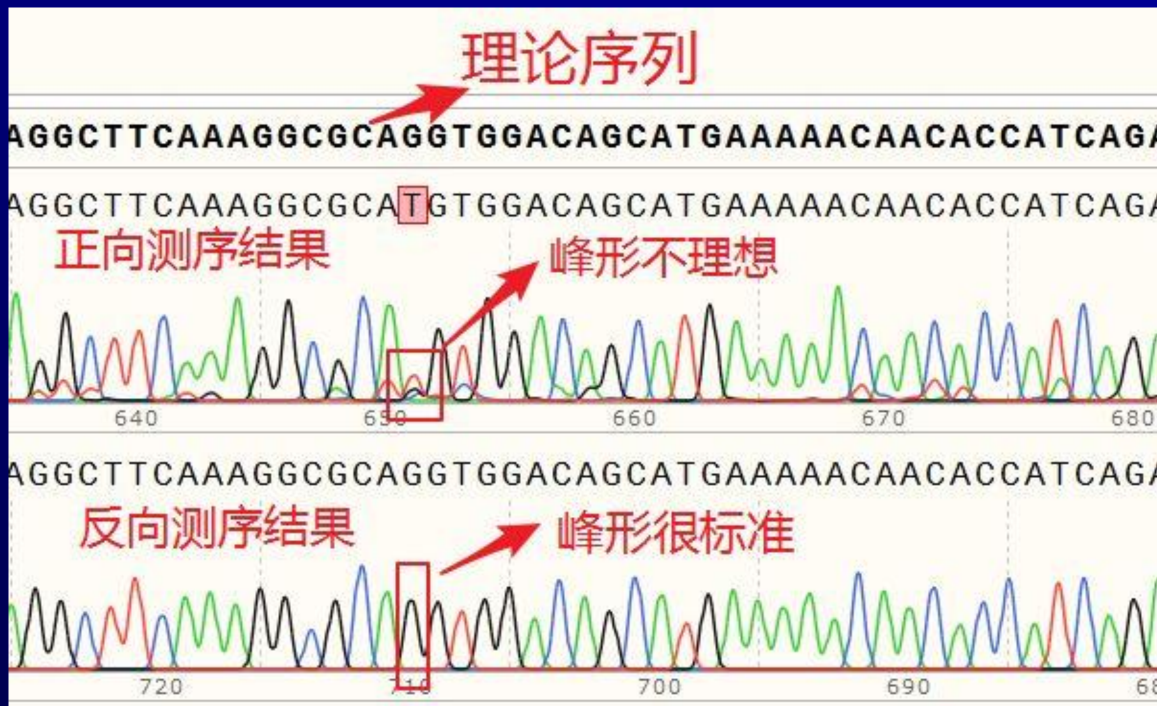


碱基的校准

- 根据对应的峰图文件（*.ab1）的质量，修改碱基
- 在BioEdit中分别打开正向和反向测序结果的AB1文件，并和上面比对结果放同一窗口中(如下图)。
- 定位到差异碱基的位置（下图黑色部分），根据对应的峰图文件ab1的质量情况，修改Consesus序列的对应碱基。

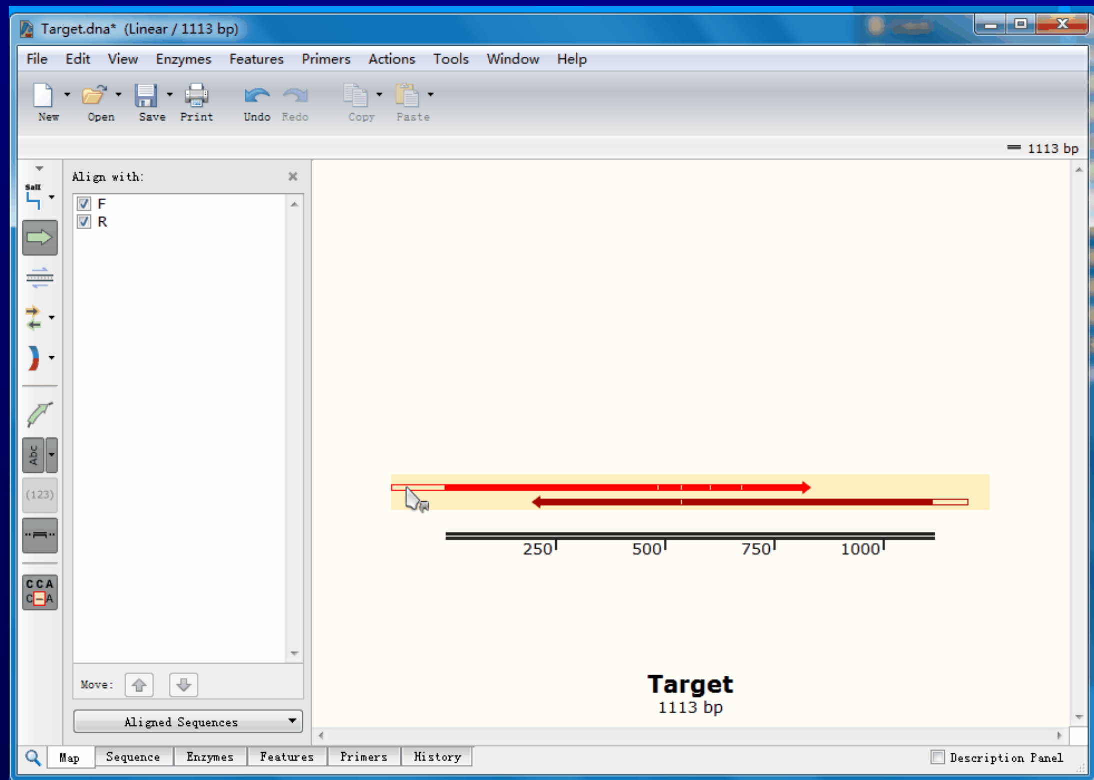


正反向测序校准碱基



正向测序的峰高非常矮，并且出现了套峰，从而产生误读(T)；相比之下，反向测序的峰形十分标准，因此，综合判断，该处反向测序的结果(G)更为可信。

- 点击Sequence标签，切换到序列页面，找到不匹配的位点，可以看到正向测序的峰非常矮，并且出现了套峰，从而产生误读；相比之下，反向测序的峰形十分标准，因此，综合判断，该处反向测序的结果更为可信。



保存校正后的consensus序列

- BioEdit中选择concensus序列->鼠标右键:copy sequences,
- BioEdit中新建Alignment文件->并粘贴序列:Edit->paste sequence，
- 点击concensus的标题改为"16S_rDNA",并保存为文件16S_rDNA.fasta

```
>16S_rDNA_2
GGGAGTGGGGGCATGCTTACCATGCAAGTCGCACGAAGGTTTCGGCCTTAGTGGCGGACGGGTGAGTAACGCGTAGGTATCTATCCATGGGTGGGGG
ATAACACTGGGAAACCGGTGCTAATACCGCATGACACCTGAGGGTCAAAGGCGCAAGTCGCCTGTGGAGGAGCCTGCGTTTGATTAGCTAGTGGTGG
GGGGTAAAGACTATATGCGATGATCATAGGCTGTTGAGAGATGATCAGGCACACTGGACTGAGACACGTCCAGACTCTACGGGAGCAGCAGTGGGGG
AATATGACAATGGGGGCACCTGATCCAGCAATGCCGCGTGTGTGAAGAAGGTCTTCGGATTGTAAAGCACTTTCGACGGGGACGATGATGACGGTAC
CCGTAGAAGAAGCCCCGGCTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGGGGCTAGCGTTGCTCGGAATGACTGGGCGTAAAGGGCGTGTAG
GCGGTTTGTACAGTCAGATGTGAAATCCCCGGGCTTAACATGGGAGCTGCATGTGATACGTGCAGACTAGAGTGTGAGAGAGGGTTGTGGAATTCCCA
GTGTAGAGGTGAAATTCGTAGATATTGGGAAGAACACCGGTGGCGAAGGCGGCAACCTGGCTCATTACTGACGCTGAGGCGCGAAAGCGTGGGGAG
CAAACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGATGTGTGCTAGATGTTGGGTGACTTAGTCATTCAGTGTCGCAGTTAACGCGTTAAGCAC
ACCGCCTGGGGAGTACGGCCGCAAGGTTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCG
CAGAACCTTACCAGGGCTTGAATGTAGAGGCTGCAAGCAGAGATGTTTGTTTCCCGCAAGGGACCTCTAACACAGGTGCTGCATGGCTGTCGTCAGCT
CGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTATCTTTAGTTGCCATCAGGTTGGGCTGGGCACTCTAGAGAGACTGCCGGTGA
CAAGCCGGAGGAAGGTGGGGATGACGTCAAGTCCTCATGGCCCTTATGTCCTGGGCTACACACGTGCTACAATGGCGGTGACAGTGGGAAGCTAGGT
GGTGACACCATGCTGATCTCTAAAAGCCGTCTCAGTTCGGATTGCACTCTGCAACTCGAGTGCATGAAGGTGGAATCGCTAGTAATCGCGGATCAGCA
TGCCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTTGGTTTGACCTTAAGCCGGTGAGCGAACCGCAAGGACGCA
GCCGACCACGTCCGTAACGGTGGT
```

BLAST搜索16S rDNA的同源序列

- NCBI的在线BLAST：http://blast.ncbi.nlm.nih.gov

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

以核苷酸库的Blast为例

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

BLAST参数选择

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a nucleotide query.

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

```
>16S_rDNA_2
GGGAGTGGGGGCATGCTTACCATGCAAGTCGCACGAAGGTTTCGGCCTTAGTGGCGGACGGG
TGAGTAACGCGTAGGTATCTATCCATGGGTGGGGGATAACACTGGGAAACCGGTGCTAATAC
CGCATGACACCTGAGGGTCAAAGGCGCAAGTCGCCTGTGGAGGAGCCTGCGTTTGATTAGCT
AGTGGTGGGGGGTAAAGACTATATGCGATGATCATAGGCTGTTGAGAGATGATCAGGCACAC
TGGACTGAGACACGTCCAGACTCTACGGGAGCAGCAGTGGGGGAATATGACAATGGGGGCAC
```

Or, upload file — Browse... No file selected.

Job Title: 16S_rDNA_2

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: ○Human genomic + transcript ○Mouse genomic + transcript ◉Others (nr etc.):

16S ribosomal RNA sequences (Bacteria and Archaea)

Organism (Optional): ☐ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

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 16S ribosomal RNA sequences (Bacteria and Archaea) using Blastn (Optimize for somewhat similar sequences)

☑ Show results in a new window

BLAST结果说明此菌为:

Acetobacter pasteurianus

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Acetobacter pasteurianus IFO 3283-01 strain IFO 3283 16S ribosomal RNA, complete sequence	2302	2302	98%	0.0	98%	NR_102925.1
☐	Acetobacter pasteurianus subsp. paradoxus strain LMD 53.6 16S ribosomal RNA gene, partial sequence	2296	2296	98%	0.0	98%	NR_104959.1
☐	Acetobacter pomorum strain LMG 18848 16S ribosomal RNA gene, complete sequence	2287	2287	98%	0.0	97%	NR_042112.1
☐	Acetobacter pomorum strain LTH2458 16S ribosomal RNA gene, complete sequence	2287	2287	98%	0.0	97%	NR_114684.1
☐	Acetobacter pasteurianus strain LMG 1262 16S ribosomal RNA gene, partial sequence	2284	2284	97%	0.0	98%	NR_117258.1
☐	Acetobacter pasteurianus strain LMD 22.1 16S ribosomal RNA gene, partial sequence	2271	2271	98%	0.0	97%	NR_026107.1
☐	Acetobacter pasteurianus subsp. ascendens strain LMG 1590 16S ribosomal RNA gene, partial sequence	2269	2269	98%	0.0	97%	NR_117456.1
☐	Acetobacter pasteurianus subsp. paradoxus strain LMG 1591 16S ribosomal RNA gene, partial sequence	2264	2264	98%	0.0	97%	NR_117457.1
☐	Acetobacter pasteurianus strain LMG 1262 16S ribosomal RNA gene, partial sequence	2237	2237	95%	0.0	98%	NR_118169.1
☐	Acetobacter pasteurianus strain DSM 3509 16S ribosomal RNA gene, partial sequence	2232	2232	95%	0.0	98%	NR_117257.1
☐	Acetobacter pomorum strain LMG 18848 16S ribosomal RNA gene, partial sequence	2224	2224	95%	0.0	98%	NR_118171.1
☐	Acetobacter okinawensis strain 1-35 16S ribosomal RNA gene, partial sequence	2170	2170	98%	0.0	96%	NR_113546.1
☐	Acetobacter syzygii strain NBRC 16604 16S ribosomal RNA gene, partial sequence	2170	2170	98%	0.0	96%	NR_113850.1
☐	Acetobacter syzygii strain 9H-2 16S ribosomal RNA gene, complete sequence	2170	2170	98%	0.0	96%	NR_040868.1

相似度98%

作业

- 利用BioEdit或SnapGene对一株细菌的16S rDNA序列Sanger测序结果的碱基质量进行检查，并根据峰图校正低质量碱基，尤其是双向测序的序列重叠部分的碱基要求一致。