高通量测序原理及数据处理

王梅　中级产品工程师
上海美吉微生物基因组产品部
电话：021-31050579
E-mail：mdna@majorbio.com

美吉生物
微生物基因组QQ在线大讲堂

批次	日期	主题	主要内容
01 已完成	20170420	微生物全基因组测序数据解读与典型方案设计	（1）微生物全基因组测序背景介绍（研究现状，意义，应用范围） （2）结果解读，包括测序，组装，基因预测，注释，及后续的各种高级分析解读，主要侧重于高级分析部分的结果解读， （3）典型方案设计案例解读 （4）常见问题解答
02 已完成	20170517	线粒体/叶绿体基因组测序新突破	（1）小基因组研究背景（研究现状，意义，应用范围） （2）如何利用高通量技术研究线粒体/叶绿体基因组（思路及流程） （3）研究方向及热点 （4）经典案例分享
03	9月15日	您不可不知的高通量测序知识	（1）高通量测序技术的发展历程 （2）一代测序、二代测序、三代测序技术原理 （3）目前主流的高通量测序平台和基本参数、数据产出 （4）数据下机后的处理与统计
04	9月20日	三代测序技术在微生物组学层面的应用	（1）三代测序技术的原理和发展历程 （2）三代测序技术的优势和短板 （3）三代测序技术在微生物组学领域的广泛应用
05	10月11日左右	细菌基因组分析之泛基因组学研究	（1）细菌基因组学研究的主要方法 （2）泛基因组学研究的概念和意义 （3）泛基因组学研究的方法和策略 （4）泛基因组学研究的经典案例
06	10月20日左右	病原微生物研究之耐药基因分析	（1）病原微生物的生存法则 （2）耐药基因研究的意义 （3）耐药机理剖析 （4）耐药基因的查找和深入分析
07	10月30日之前	线粒体/叶绿体基因组测序策略、组装方法	（1）线粒体、叶绿体基因组的特点 （2）各种样品制备方法的比较 （3）线粒体、叶绿体基因组测序的策略和组装方法
08	11月10日之前	线粒体、叶绿体基因组的个性化分析	（1）线粒体、叶绿体基因组测序的应用价值 （2）针对于不同研究需求的线粒体、叶绿体基因组分析方法 （3）常见的线粒体、叶绿体基因组测序论文写作思路

美吉生物
微生物基因组QQ在线大讲堂

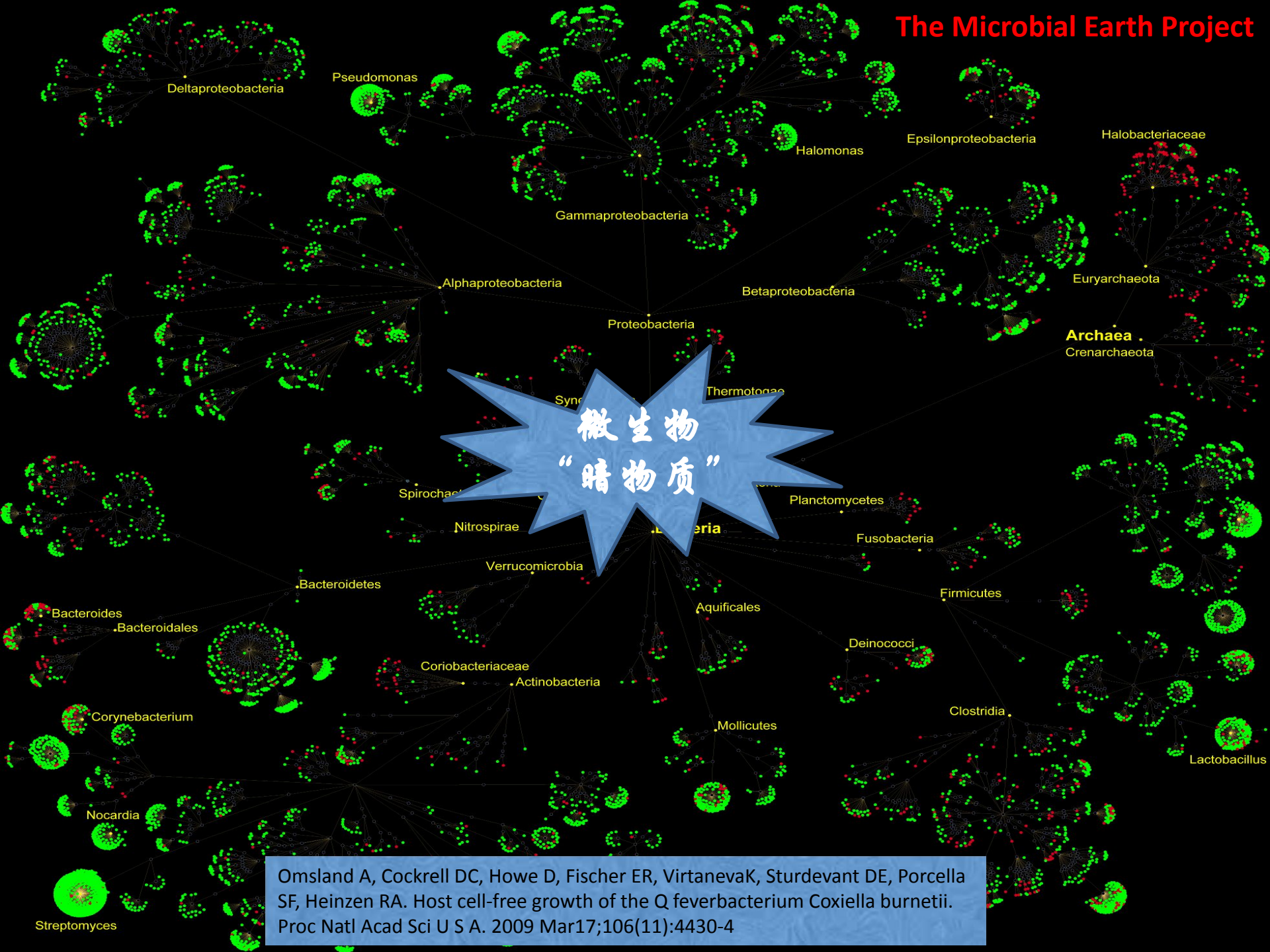
			（4）典型的线粒体、叶绿体基因组测序文章解读
09	11月20日之前	生信课程：如何在 windows 操作系统下玩转微生物基因组学	（1）主要的微生物基因组学分析方法、软件和数据库 （2）基因挖掘层面的分析 （3）基因组大片段层面的研究方法 （4）比较基因组层面的分析方法
10	11月30日之前	病原微生物的基因组学研究热点与分析方法	（1）病原微生物的危害与研究价值 （2）病原微生物的武器装备 （3）病原微生物的研究热点 （4）组学层面上研究病原微生物的法宝
11	12月15日前	环境与工业微生物组学研究热点与分析方法	（1）环境微生物学研究热点与典型代表 （2）工业微生物学研究热点与应用案例 （3）组学层面的分析方法和解决的科学问题
12	12月25日左右	微生物组学层面的基因水平转移研究思路	（1）微生物基因组进化的动力 （2）基因水平转移研究的意义和经典案例 （3）基因水平转移的途径 （4）基因水平转移的研究方法

目录

1. 微生物研究简史
2. 高通量测序技术发展简史
3. 测序技术原理
4. 下机数据处理
5. 美吉与高通量

01 Part One 微生物研究简史

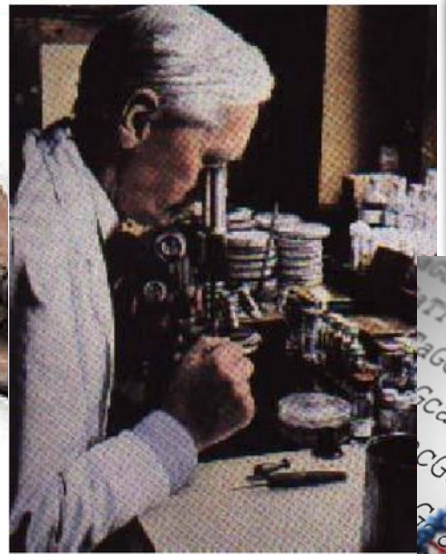

宇宙
“暗物质”



Omsland A, Cockrell DC, Howe D, Fischer ER, VirtanevaK, Sturdevant DE, Porcella SF, Heinzen RA. Host cell-free growth of the Q feverbacterium Coxiella burnetii. Proc Natl Acad Sci U S A. 2009 Mar17;106(11):4430-4

微生物发展史

人类最早发明的酒



经验微生物学

缺乏研究工具

16世纪前

实验生物学

列文虎克发明显微镜

17~19世纪

现代微生物

电子显微镜，同位素示踪原子、PCR

20世纪

分子生物学

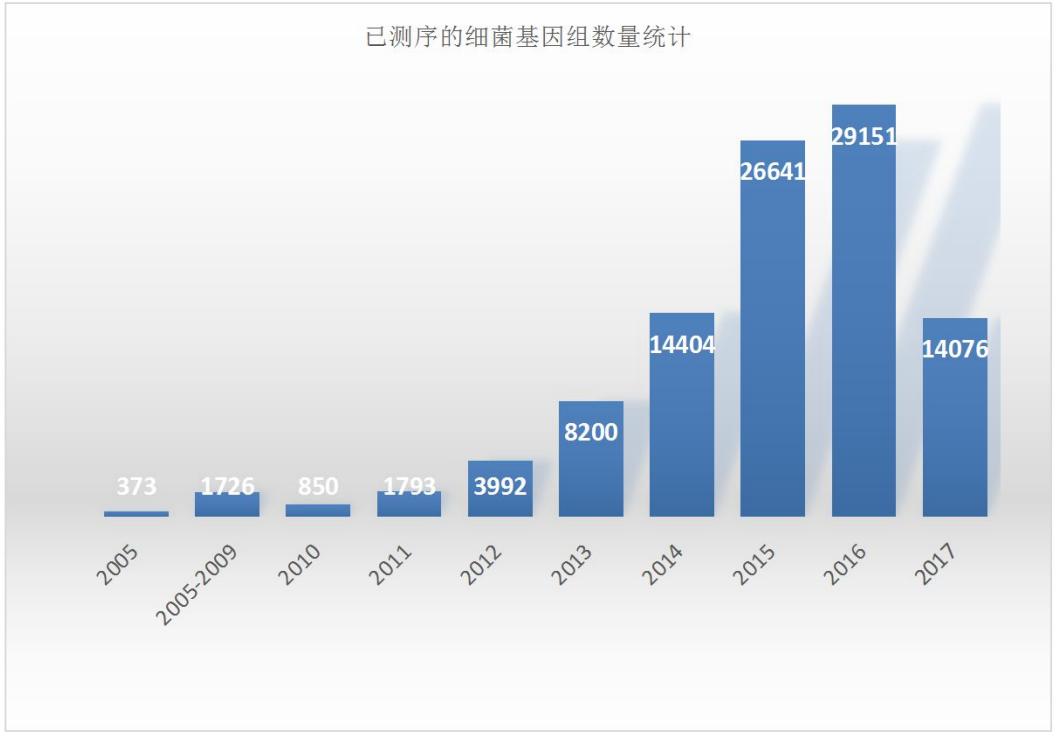
一代测序平台3730

合成基因组学

高通量及合成技术

21世纪

高通量测序技术的发展开创了微生物研究的新纪元



截至2017年6月18日，NCBI收录的微生物基因组已超过6.5万种细菌、1,300种真菌、5,600种病毒总计超过11万株，其中细菌基因组101,206个，真菌基因组2,377个，病毒7,409个。

数据来源 ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/

国际和国家层面上的重大测序项目

1990年：人类基因组计划启动 (human genome project, HGP)

1994年：美国能源部：微生物基因组计划 (mocrobial genome project, MGP)

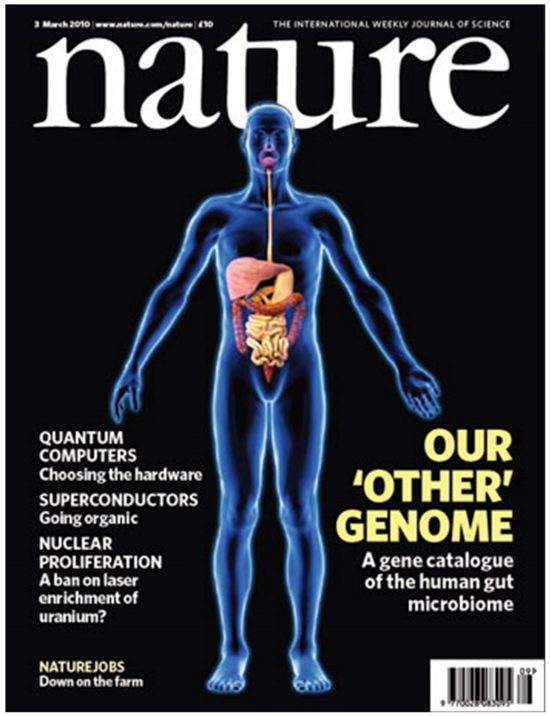
2008年：美国国立卫生研究院：人类微生物组计划（human microbiome project, HMP）；欧盟：人类肠道宏基因组计划（metagenomics of the human intestinal tract, MetaHIT）

2011年：多国：地球微生物计划启动 (earth microbiome project, EMP)

2013年：多国：十万食源性病原微生物基因组计划

2016年：美国国家微生物组计划（national microbiome initiative, NMI）

2017年：中国国家微生物组计划？



超 级 生 物 体

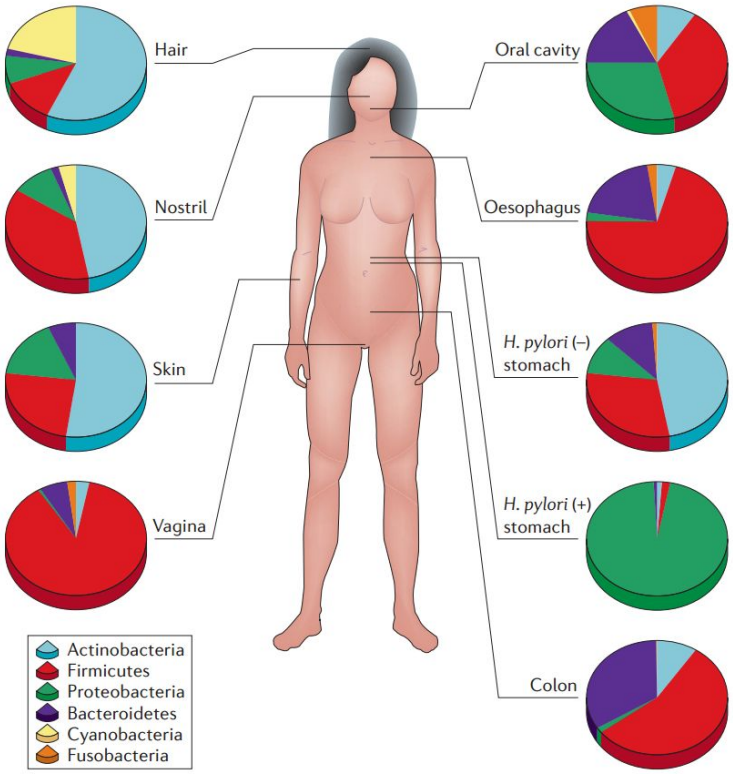
人自身基因组
人体微生物基因组

2007

NIH HUMAN MICROBIOME PROJECT

2008

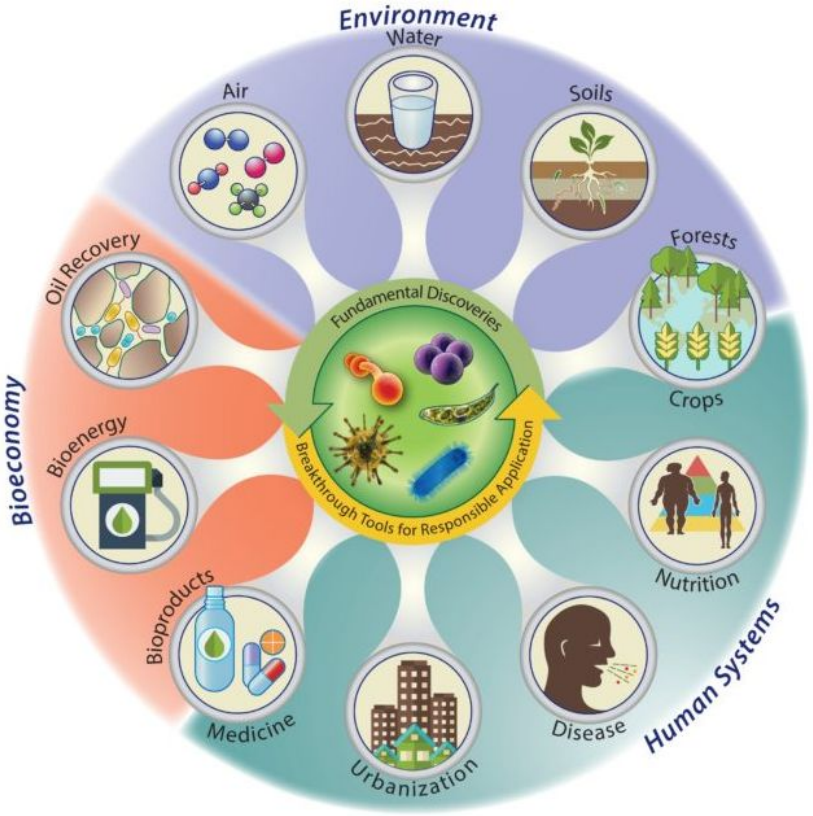
MetaHIT



Peterson J et al., Genome research, 2009; Qin J et al., Nature, 2010

THE NATIONAL MICROBIOME INITIATIVE

2016年5月13日，白宫科学和技术政策办公室（OSTP）宣布启动新的**国家微生物组计划（National Microbiome Initiative，NMI）**



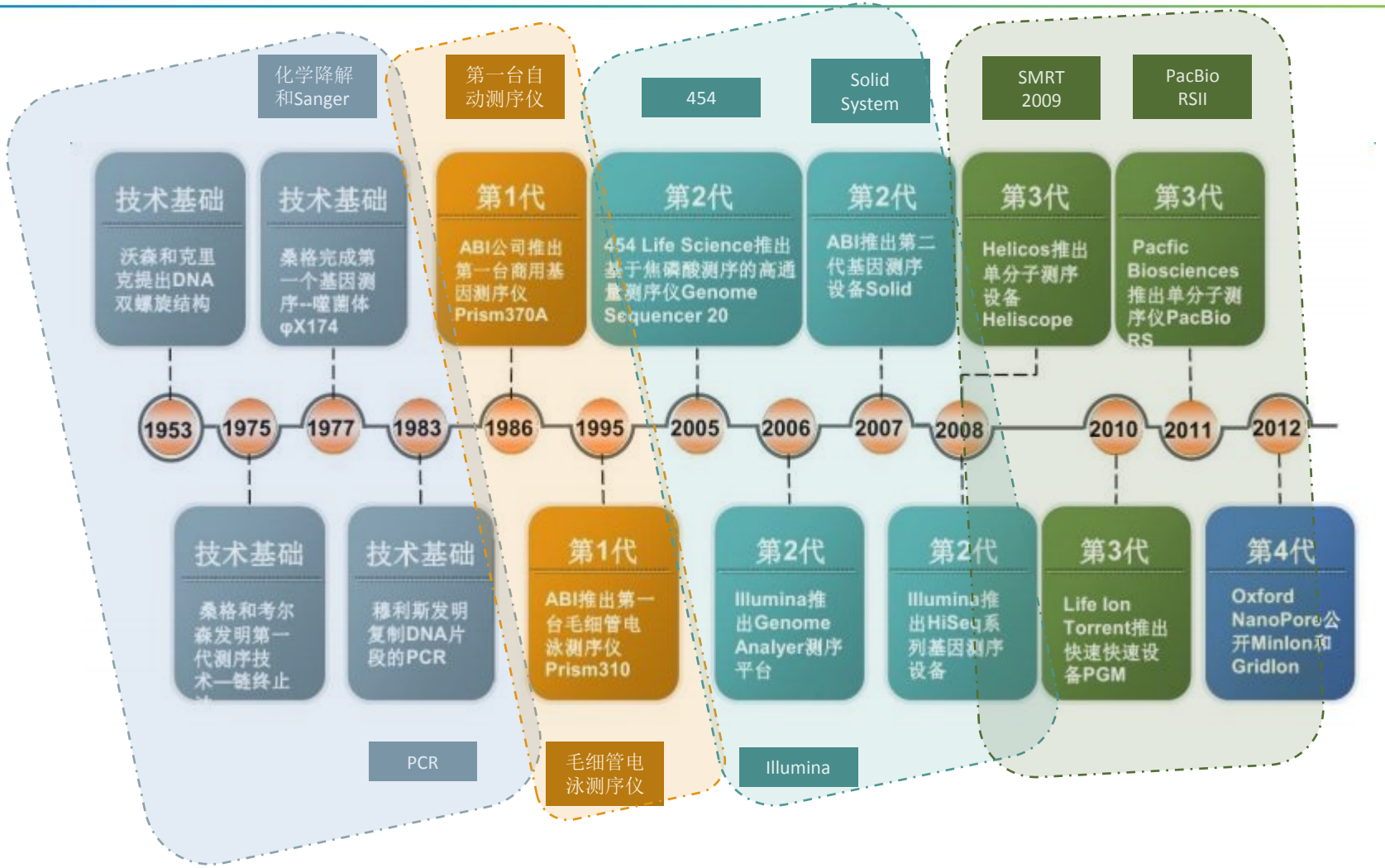
白宫1.21亿美元
民间4亿美元
总计5.21亿美元

➢**支持跨学科研究**：以回答不同生态系统中微生物的基本问题；

➢**开发平台技术**：提升对于不同生态系统中不同微生物组的认识，进行数据累积，并提高微生物基因组数据库访问；

➢**扩大微生物的影响力**：通过全民科普、公众教育和参与的方式进行传播。

02 Part Two 高通量技术发展简史


化学降解和Sanger
第一台自动测序仪
454
Solid System
SMRT 2009
PacBio RSII
技术基础
沃森和克里克提出DNA双螺旋结构
技术基础
桑格完成第一个基因测序--噬菌体φX174
第1代
ABI公司推出第一台商用基因测序仪Prism370A
第2代
454 Life Science推出基于焦磷酸测序的高通量测序仪Genome Sequencer 20
第2代
ABI推出第二代基因测序设备Solid
第3代
Helicos推出单分子测序设备Heliscope
第3代
Pacfic Biosciences推出单分子测序仪PacBio RS
1953
1975
1977
1983
1986
1995
2005
2006
2007
2008
2010
2011
2012
技术基础
桑格和考尔森发明第一代测序技术—链终止
技术基础
穆利斯发明复制DNA片段的PCR
第1代
ABI推出第一台毛细管电泳测序仪Prism310
第2代
Illumina推出Genome Analyer测序平台
第2代
Illumina推出HiSeq系列基因测序设备
第3代
Life Ion Torrent推出快速快速设备PGM
第4代
Oxford NanoPore公开MinIon和GridIon
PCR
毛细管电泳测序仪
Illumina

1975-1977 1975年第一代测序技术诞生（链终止法（Sanger），化学降解法）
1977第一个基因组序列-噬菌体X174测序完成

1980 鸟枪法诞生

1986 荧光ddNTP自动测序仪发明

一代测序

1987 第一台测序仪上市（ABI370A），通量为1kb碱基/天

1990 人类基因组计划启动

1995:1997:1998	ABI推出毛细管电泳测序仪，通量为5kb~15kb/天； MegaBACE1000毛细管电泳测序仪，通量250kb~500kb/天 Prism3700毛细管电泳测序仪，通量500kb~1Mb/天
1996~2000	第一个真核生物-酿酒酵母基因组完成测序 第一个植物基因组-拟南芥完成测序
2001：2004	人类基因组草图序列公布 人类基因组常染色体序列精细图完成
2005~2007 二代测序	454焦磷酸测序仪发明，通量20Mb/天 Solexa发明，完成第一个亚洲人基因组测序 SOLiD发明，每轮运行可产生超过40亿碱基的可定位数据
2008~2009 单分子测序	Helicos Biosciences　Heliscope测序平台,第一款单分子测序仪 Pacific Biosciences　SMRT测序技术发明

2011 Pacific Biosciences推出单分子测序仪Pacbio RSII

2015 Pacific Biosciences推出单分子测序仪Pacbio sequel

03 Part Three

基因组测序原理

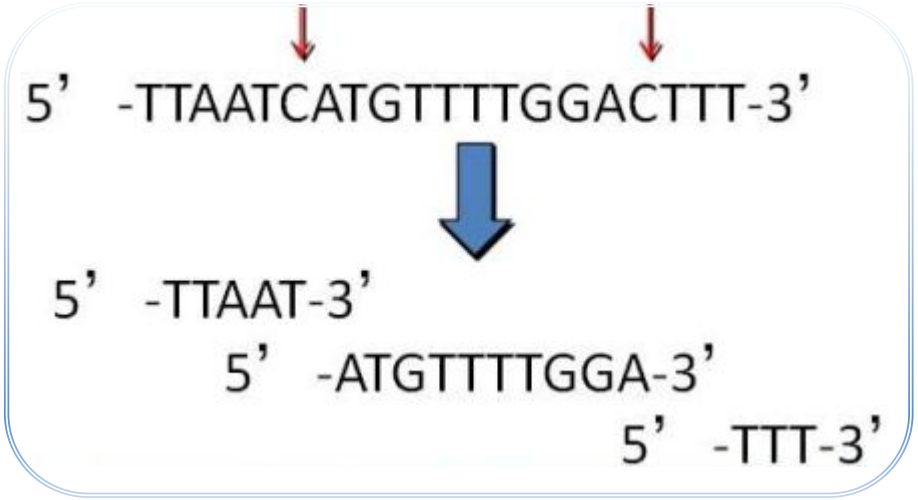
第一代测序技术	第二代测序技术	第三代测序技术
化学法	454（焦磷酸测序）	SMRT
Sanger法	Solexa（SBS）	Nanopore
	SOLiD（SBS）	

Part1

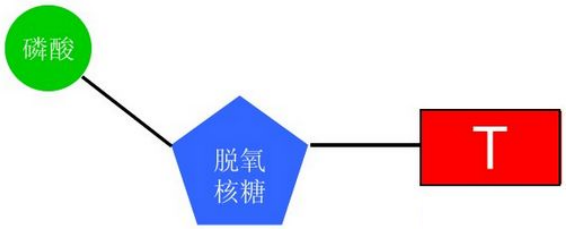
第一代测序

化学法

1976~1977年，由Allan Maxam 和Walter Gilbert提出



原理：通过化学方法将DNA序列在特定位点打断，以特定碱基结尾的片段群通过凝胶电泳分离，再经放射自显影，确定各片段末端碱基，从而得出目的DNA碱基序列

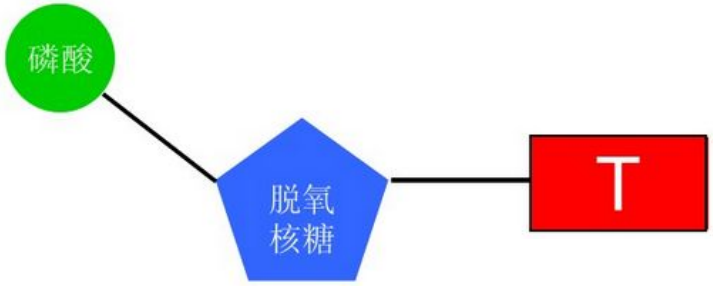


dTTP

Part1

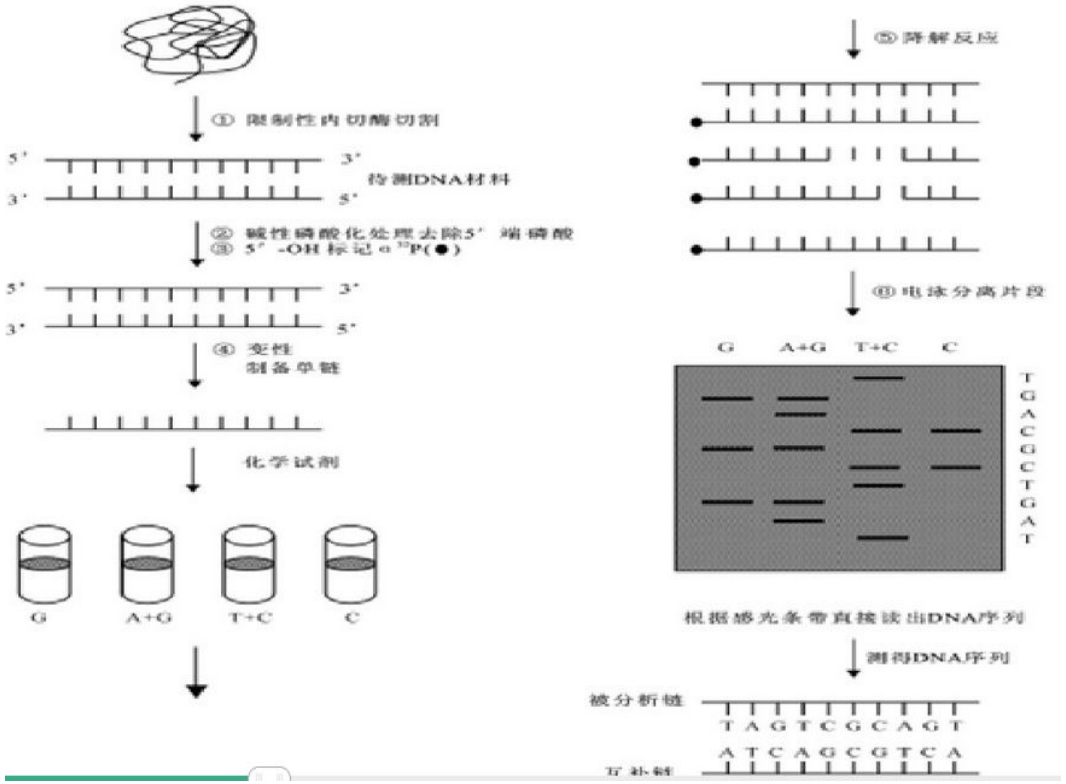
第一代测序

化学法



步骤：

- 纯化的dsDNA变性为ssDNA
- ssDNA的5'端用32-P标记
- 同一样本分为4份，分别用不同的化学试剂切割
- 电泳分离
- 序列读取



化学降解法测序原理

Part1

第一代测序

化学法

目前已不再使用，原因：

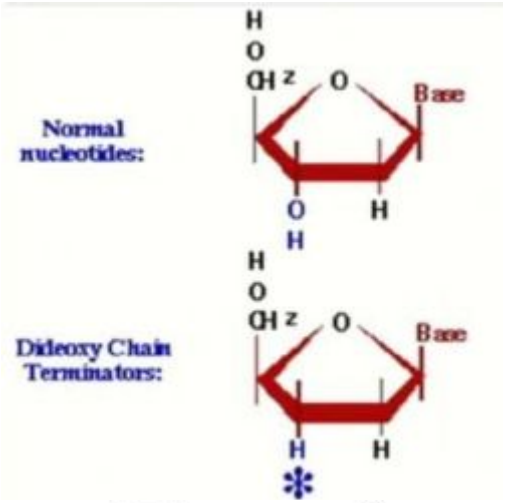
- 链断裂使用的试剂难于制成商业化的kit
- DNA样品需求量大
- 难于实现自动化测序

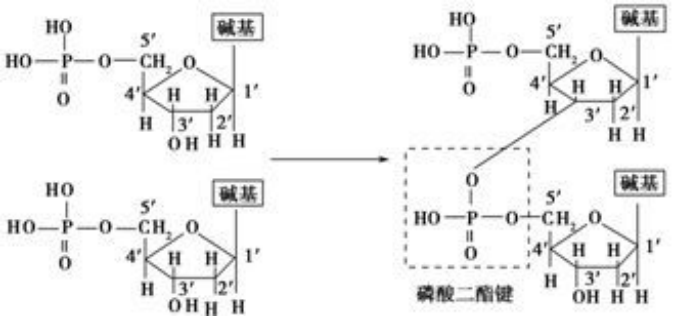
Part1

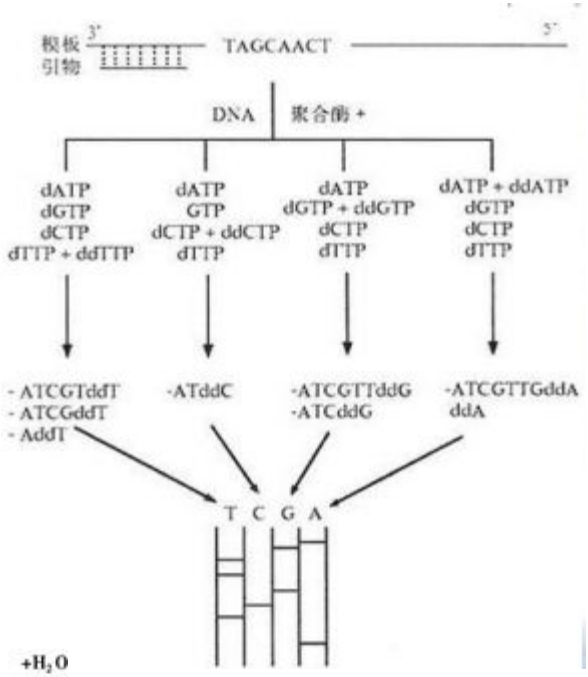
第一代测序

Sanger

1977年，由Fred Sanger发明，又称双脱氧终止法







原理：

- DNA样品分四组，分别在四个体系中合成
- 每个体系中分别加入一种ddNTP
- 得到一系列以ddNTP结尾的片段
- 电泳，读取序列

Part1

第一代测序

Sanger

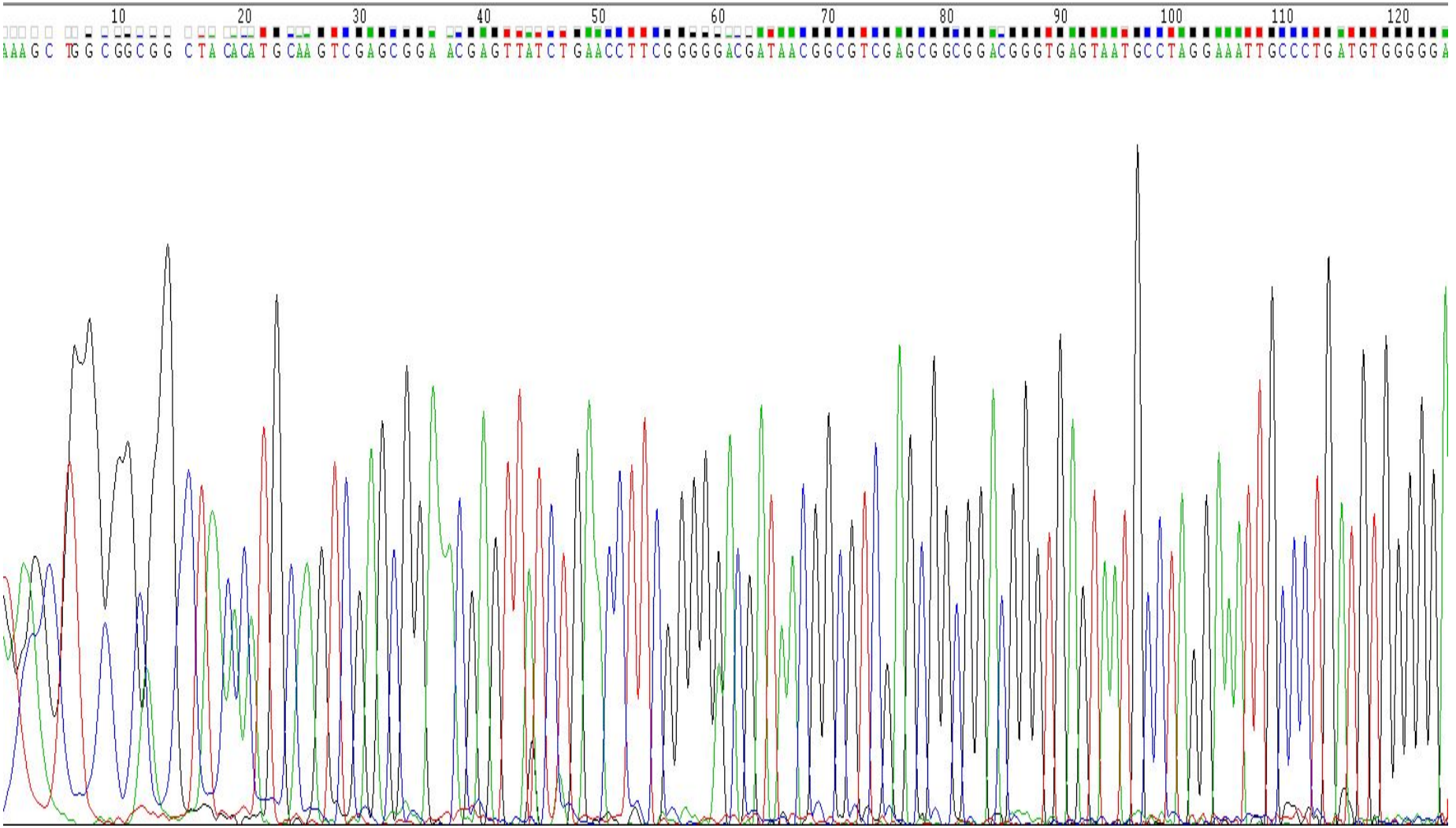
几种Sanger测序仪

3730XL

377

Megabase

一代测序峰图


10 20 30 40 50 60 70 80 90 100 110 120
AAAGC TGGCGGCGG CTACACATGCAAGTCGAGCGGA ACGAGTTATCTGAACCTTCGGGGGACGATAACGGCGTCGAGCGGCGGACGGGTGAGTAATGCCTAGGAAATTGCCCTGATGTGGGGGA

Part1 第一代测序 小结

特征：
- 读长：~1kb
- 数据准确性：99.999%

优势：
- 单序列读长较长
- 可靠性，准确性，可用于基因组测序的gap closing

局限性：
通量低，96reads/run
用于基因组测序的成本高

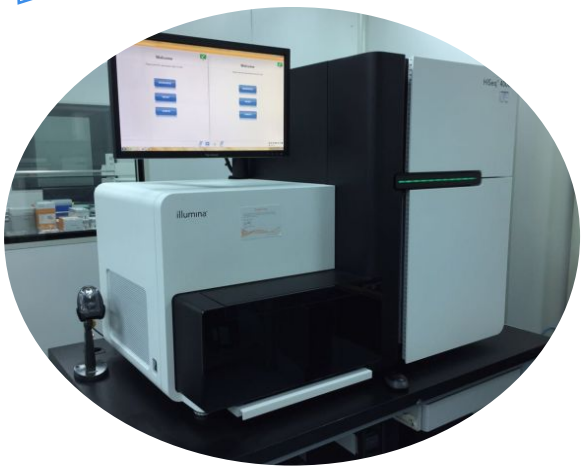
Part2

第二代测序

三国鼎立

illumina

Solexa



Applied Biosystems

SOLiD™

Roche

454 LIFE SCIENCES

Part2

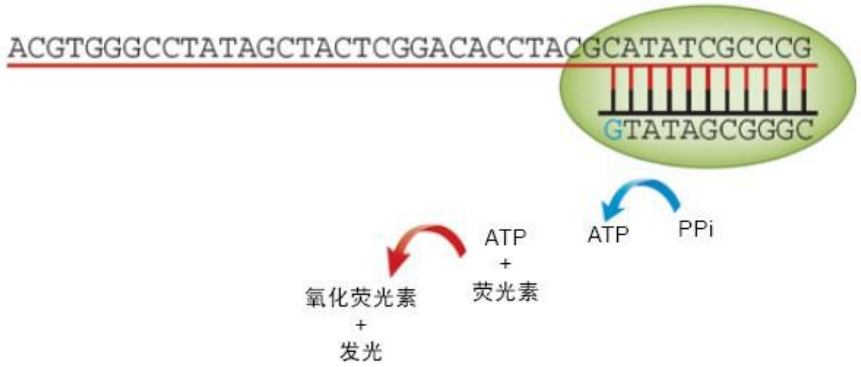
第二代测序

Roche 454

454公司是新一代测序技术的奠基人，2003年首先推出高通量焦磷酸测序技术，2005年3月，罗氏公司以1.55亿美元收购了454公司

Roche 454测序原理：焦磷酸测序（Pyrosequencing）

测序时单链DNA进入PTP（Pico TiterPlate）平板的光线小孔，进行第二链的酶促合成，四种碱基（T\A\C\G）依次循环进入PTP板，每次只进入一个碱基，如果发生碱基配对，就会释放一个焦磷酸，焦磷酸在酶促反应下释放光信号，光信号被实时捕获，每一个碱基和模板配对就会捕获一份子的光信号，一一对应就可读出模板序列。



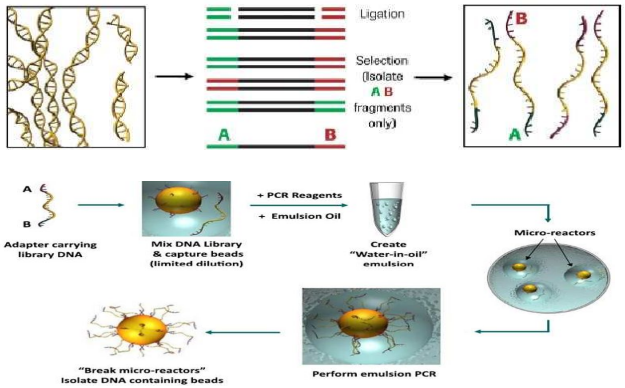
Part2

第二代测序

Roche 454

Roche

Roche 454测序流程：

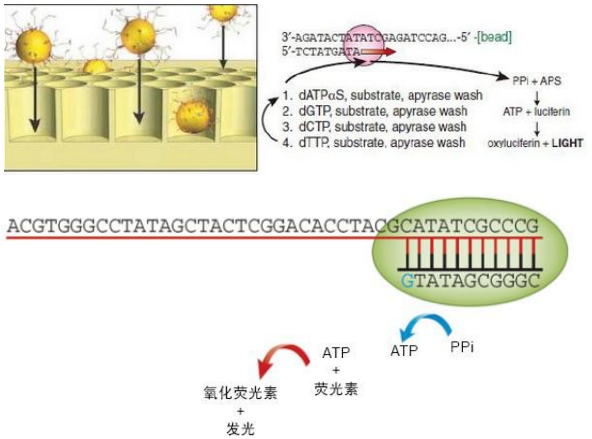


Step1 文库制备（300~800bp）

片段化，末端修复，加接头，回收ssDNA

Step2 乳液PCR

单链DNA与磁珠乳化，每个片段在自己的微反应器中独立扩增



Step3 焦磷酸测序

测序反应以磁珠上大量扩增的ssDNA为模板，每次依次加入一种dNTP进行反应，如果这种dNTP能与待测序列配对，则在与模板结合后释放焦磷酸，焦磷酸与体系中的试剂发生作用，释放特定颜色的荧光，信号被记录并经计算机分析转换为测序结果。

Part2

第二代测序

Roche 454

Roche 454特点：

- 速度快，一个反应10h，读取碱基数超过4~6亿，比Sanger快100倍
- 平均读长450bp
- 通量高，单次反应可获得100万个序列读长
- 准确性高，读长超过400bp时，单一读长的准确性≥99%
- 可以进行双端测序
- 简便高效，实验操作简化，一个人即可完成

Part2

第二代测序

Roche 454

Roche

Roche 454发展：

- 2005年底 Genome Sequencer 20 System
- 2007年 Genome Sequencer FLX System
- 2008年10月 GS FLX Titanium系列试剂和软件
- 2010年，454 GS FLX Titanium读长达到1000bp
- 2013年，由于测序成本高，测序读长被三代排挤，罗氏公司关闭了454 生命科学测序业务
- 2016年，454 测序仪退出市场

Part2

第二代测序

Solexa

illumina

由隶属于剑桥大学的Solexa公司发明，2007年被**Illumina**公司以6亿美元收购；

目前 Illumina 公司的测序仪占据了 70%的市场，是目前性价比最高、应用最广泛的测序技术

核心技术：DNA Cluster和**Reversible terminator**（可逆阻断技术）

测序原理：边合成边测序（Sequencing by Synthesis，**SBS**）

Sequencing-By-Synthesis

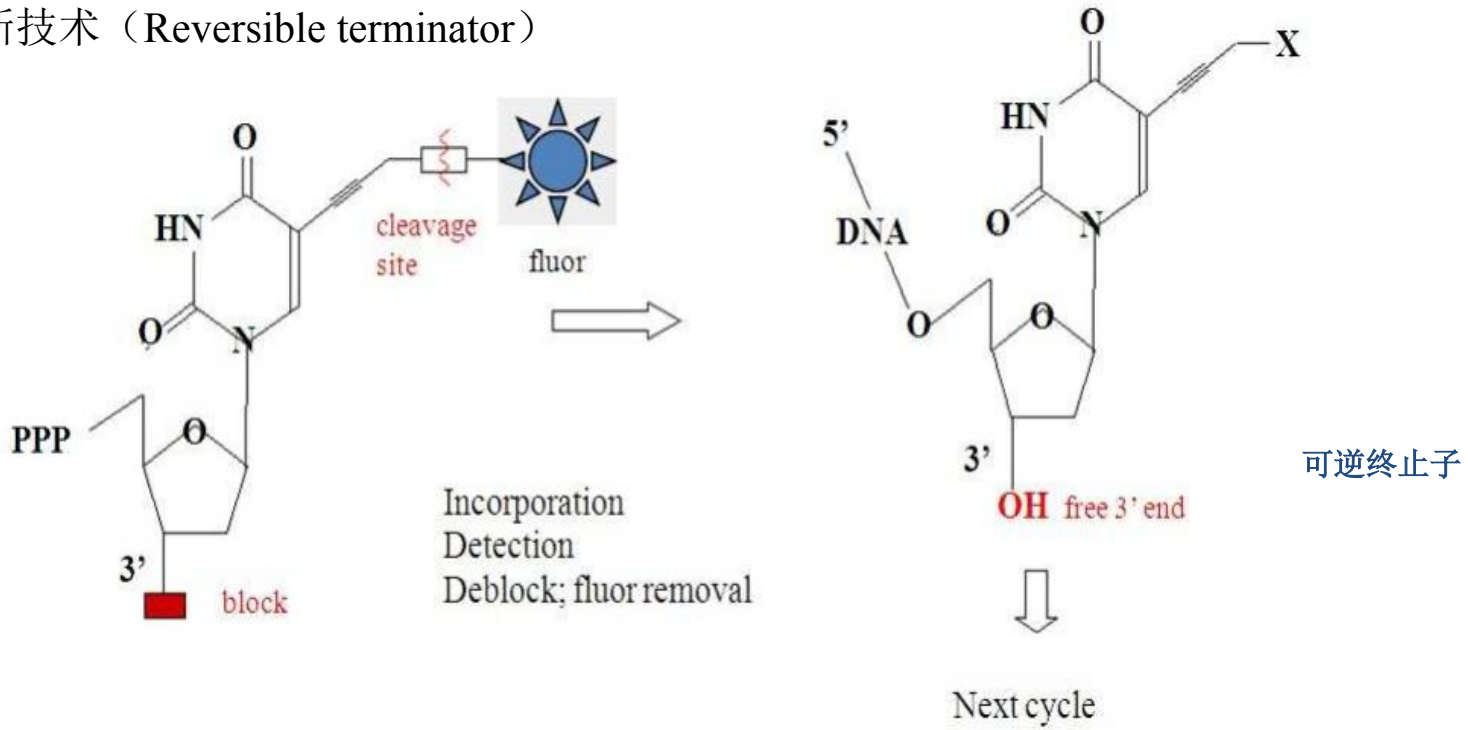
Part2

第二代测序

Solexa

illumina

可逆阻断技术（Reversible terminator）



可逆终止子

利用3'端的阻断基团，确保一次只能合成一个碱基，再利用相应的激光激发荧光基团，捕获激发光，从而读取碱基信息

Part2

第二代测序

Solexa

illumina

边合成边测序技术（SBS）

每个cycle同时投放4种碱基，在阻断基团的作用下每条reads一次只能合成一个碱基，拍照完成后，去掉阻断基团和荧光基团，继续合成下一个碱基

Part2

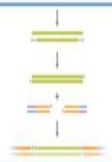
第二代测序

Solexa

illumina

测序流程

1 Library Preparation



2 Cluster Generation

3 Sequencing

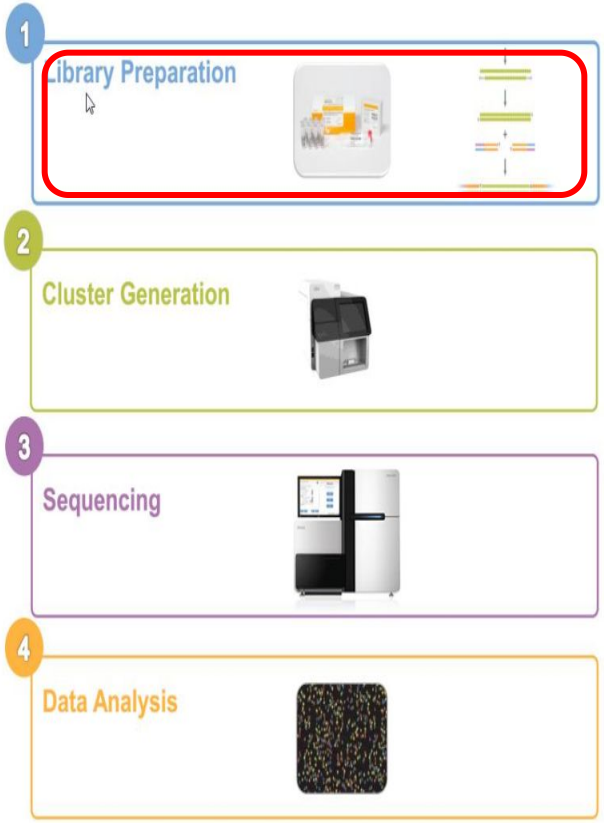
4 Data Analysis

Part2

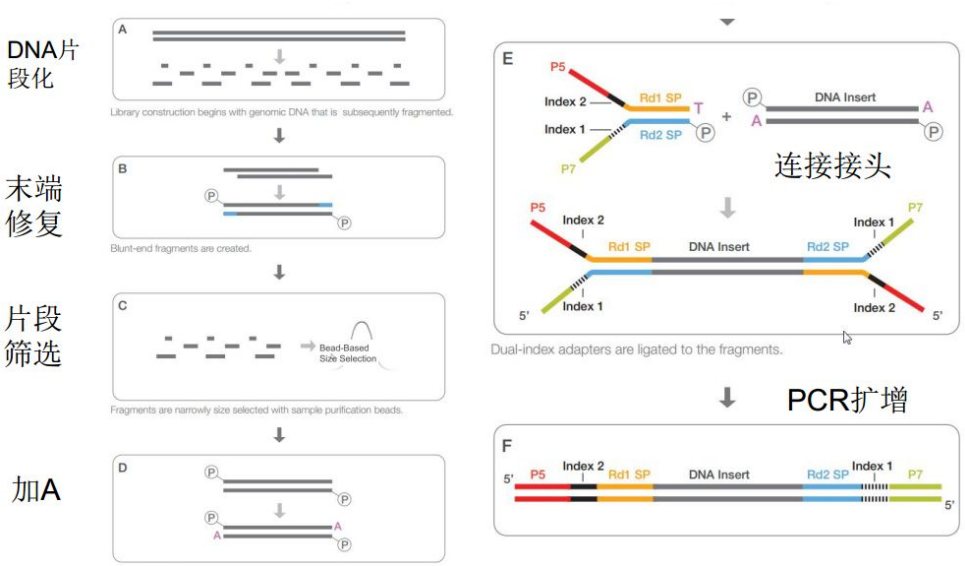
第二代测序 Solexa

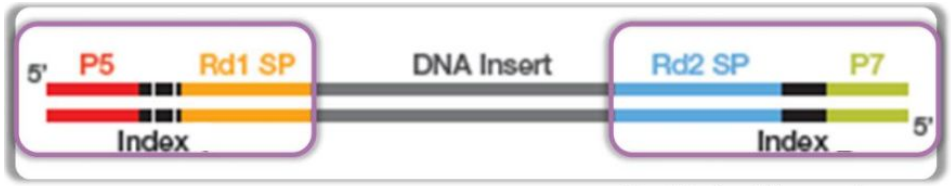
illumina

测序流程



1.文库制备





Dual Index Library shown

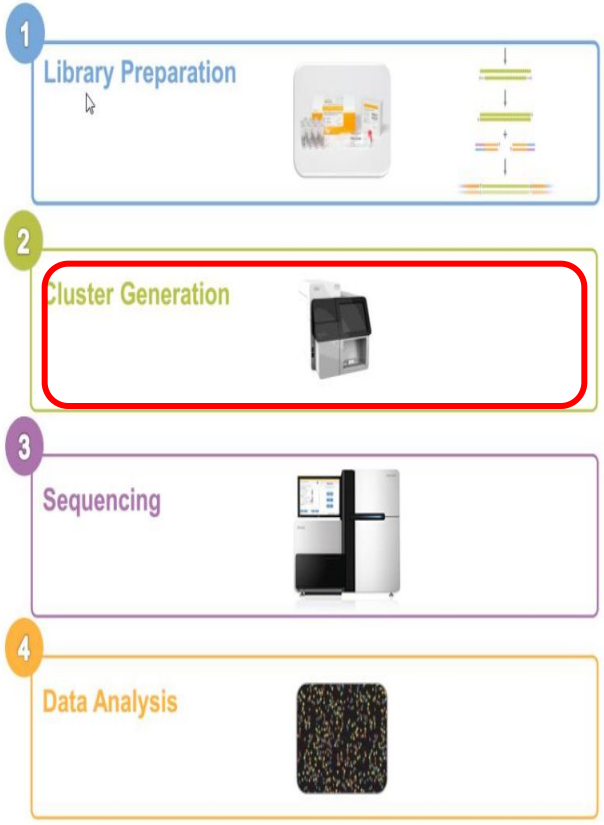
Part2

第二代测序

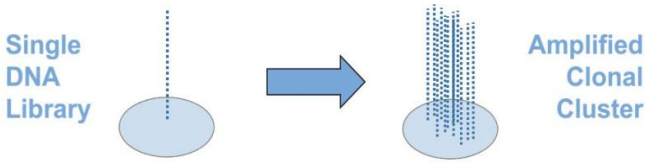
Solexa

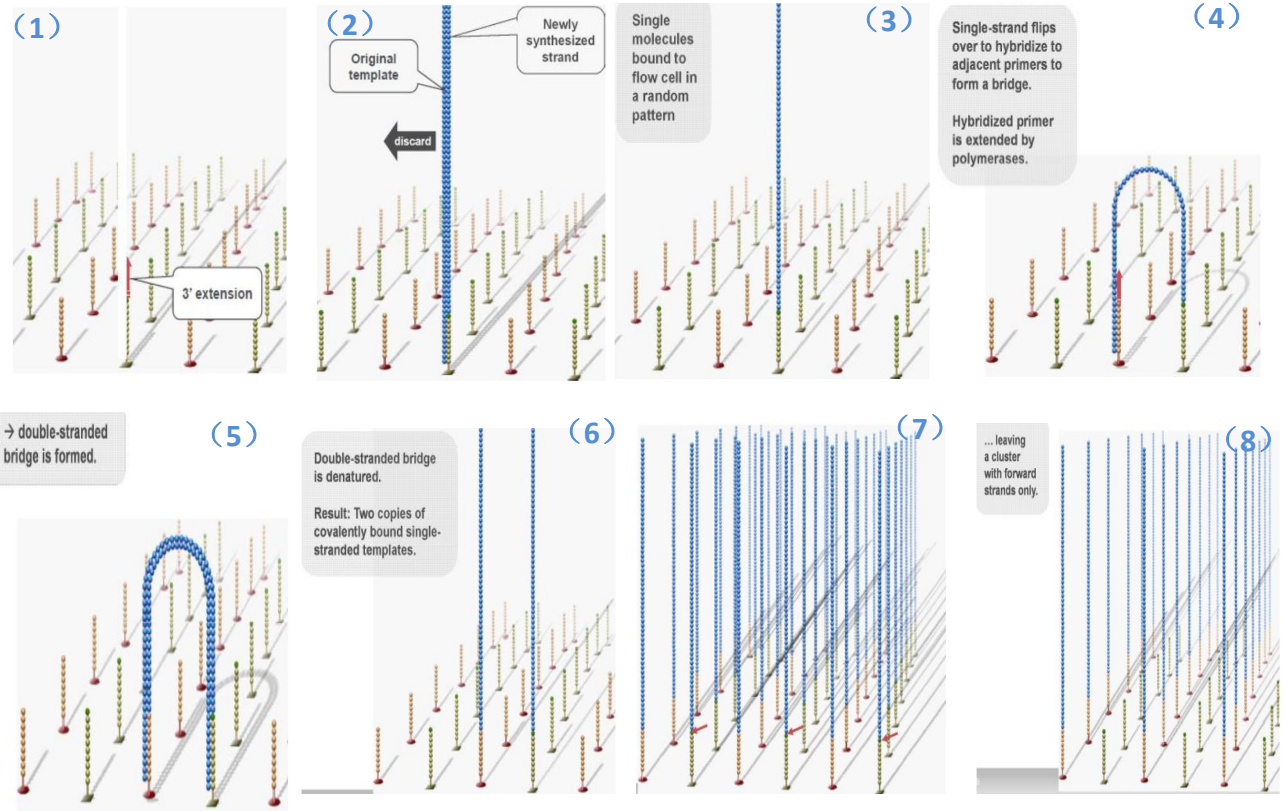
illumina

测序流程



2.桥式PCR，形成cluster





Part2

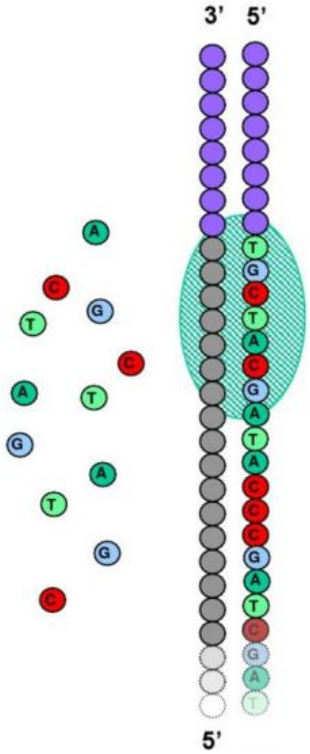
第二代测序

Solexa

illumina

测序流程

3.桥式PCR，形成cluster



边合成，边测序

Part2

第二代测序

Solexa

illumina

测序流程

1. 文库制备：将基因组DNA打成几百个碱基（或更短）的小片段，在片段的两个末端加上接头(adapter)。
2. 产生DNA簇： 利用专利的芯片，其表面连接有一层单链引物， DNA片段变成单链后通过与芯片表面的引物碱基互补被一端“固定”在芯片上。另外一端（ 5’或3’ ）随机和附近的另外一个引物互补，也被“固定”住，形成“桥(bridge) ”。反复30轮扩增，每个单分子得到了1000倍扩增，成为单克隆DNA簇。 DNA簇产生之后，扩增子被线性化，测序引物随后杂交在目标区域一侧的通用序列上
3. 上机测序：利用边合成边测序（ Sequencing By Synthesis）的原理。加入改造过的DNA聚合酶和带有4种荧光标记的dNTP。 这些核苷酸是“可逆终止子”，因为3’羟基末端带有可化学切割的部分，它只容许每个循环掺入单个碱基。此时，用激光扫描反应板表面，读取每条模板序列第一轮反应所聚合上去的核苷酸种类。之后，将这些基团化学切割，恢复3'端粘性，继续聚合第二个核苷酸。如此继续下去，直到每条模板序列都完全被聚合为双链。这样，统计每轮收集到的荧光信号结果，就可以得知每个模板DNA片段的序列
4. 数据分析：自动读取碱基，数据被转移到自动分析通道进行二次分析

Part2

第二代测序

Solexa

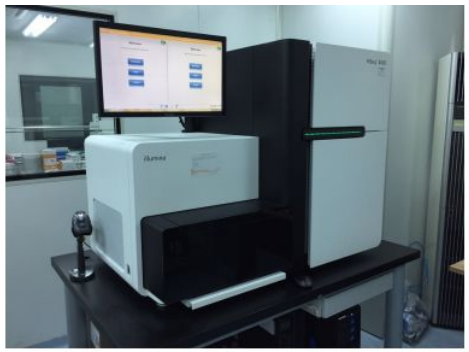
illumina

测序仪器

主要仪器：

Hiseq3000/4000

Hiseq 2500



cBot

NextSeq 500

miseq

Part2

第二代测序

SOLiD

Applied Biosystems

SOLiD（Sequencing by Oligonucleotide Ligation and Detection）

测序原理：**链接法测序（Sequencing by Ligation）**

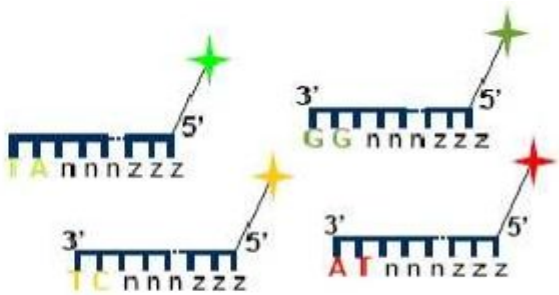
测序引物与文库接头互补配对，引物5’端与体系中游离的八聚体寡核苷酸竞争性连接，该寡聚体被荧光标记。 当其标记颜色被读取后，即在寡核苷酸在第五位和第六位之间切断，以移除标记，进行下一轮反应，以此依次循环。在第一轮反应中，可以得到确定的碱基位点为： 1、 2、6、7、11、 12……位碱基等。重复该反应过程，偏移一位碱基，使用较第一轮少一个碱基的引物进行反应，可以确定的碱基位点包括： 0、 1、5、 6……等，如此往复，直至偏移至引物的第一个碱基（即待测序列0位点碱基）。 由于该位点碱基已知，可通过读取的荧光颜色得知位点1的碱基类型，然后，又以位点1碱基荧光颜色推知位点2的碱基类型，依此类推，直至整个序列读序完成。

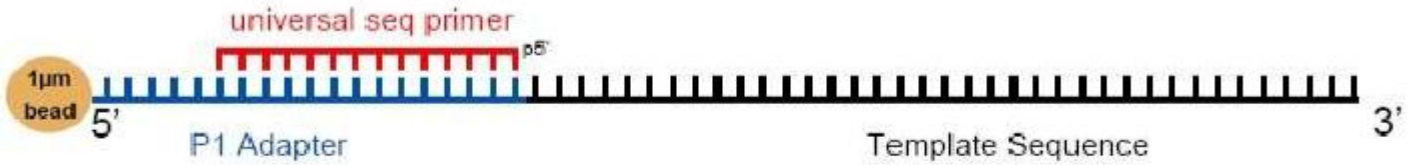
Part2

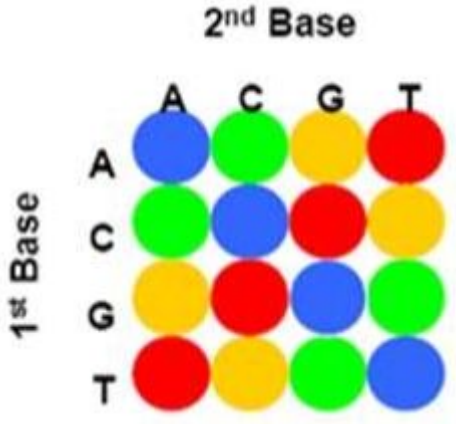
第二代测序

SOLiD

Applied Biosystems







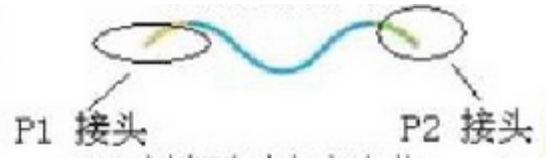
Part2

第二代测序 SOLiD

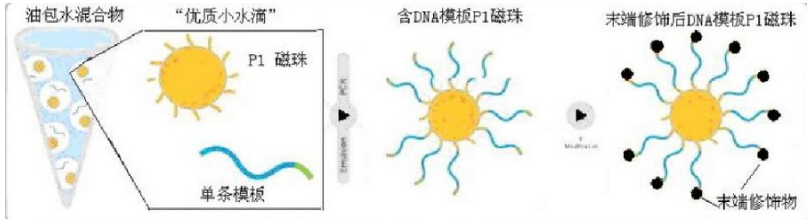
Applied Biosystems

SOLiD流程

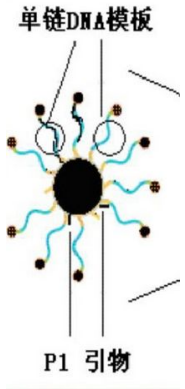
文库构建：基因组打断，连接测序接头



乳液PCR



测序



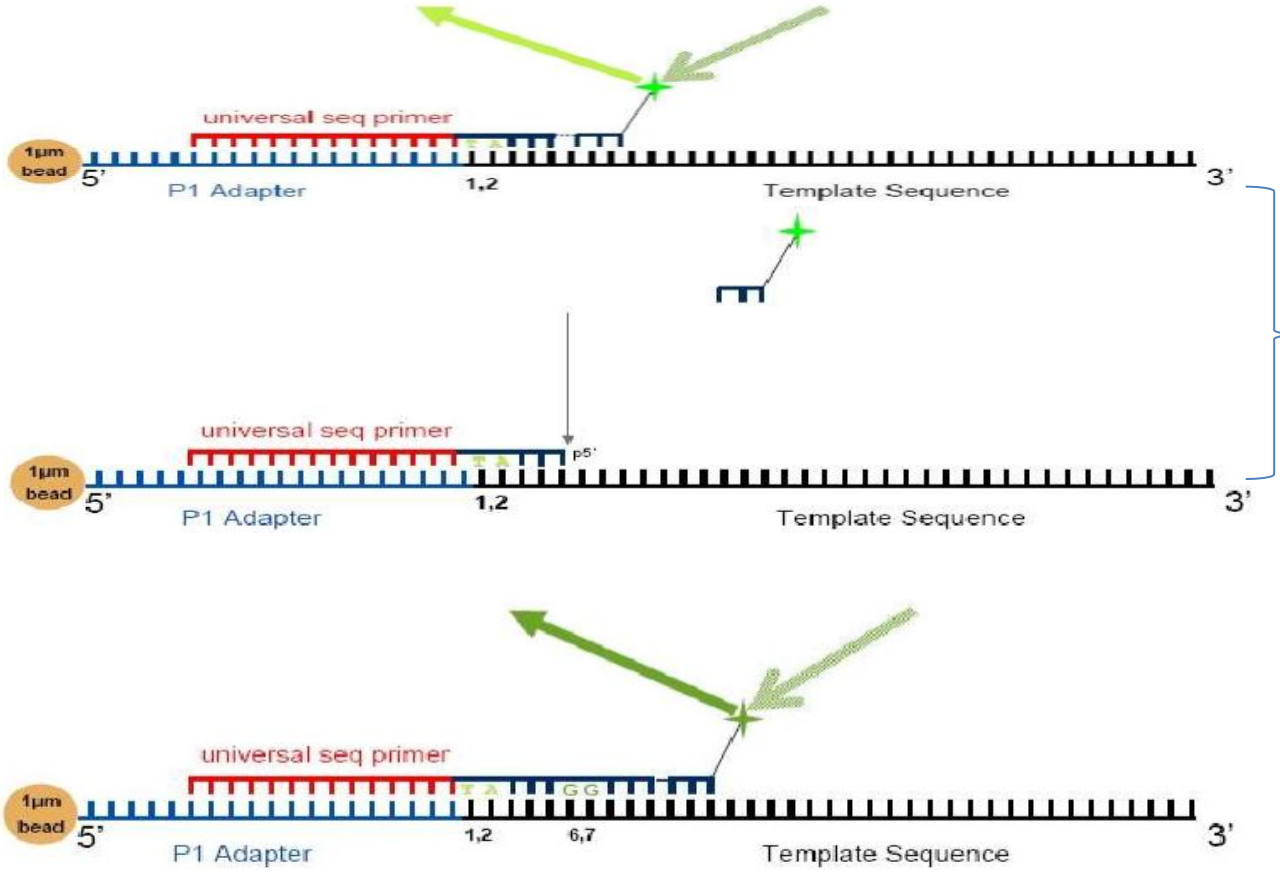
Part2

第二代测序

SOLiD

Applied Biosystems

测序



1st cycle

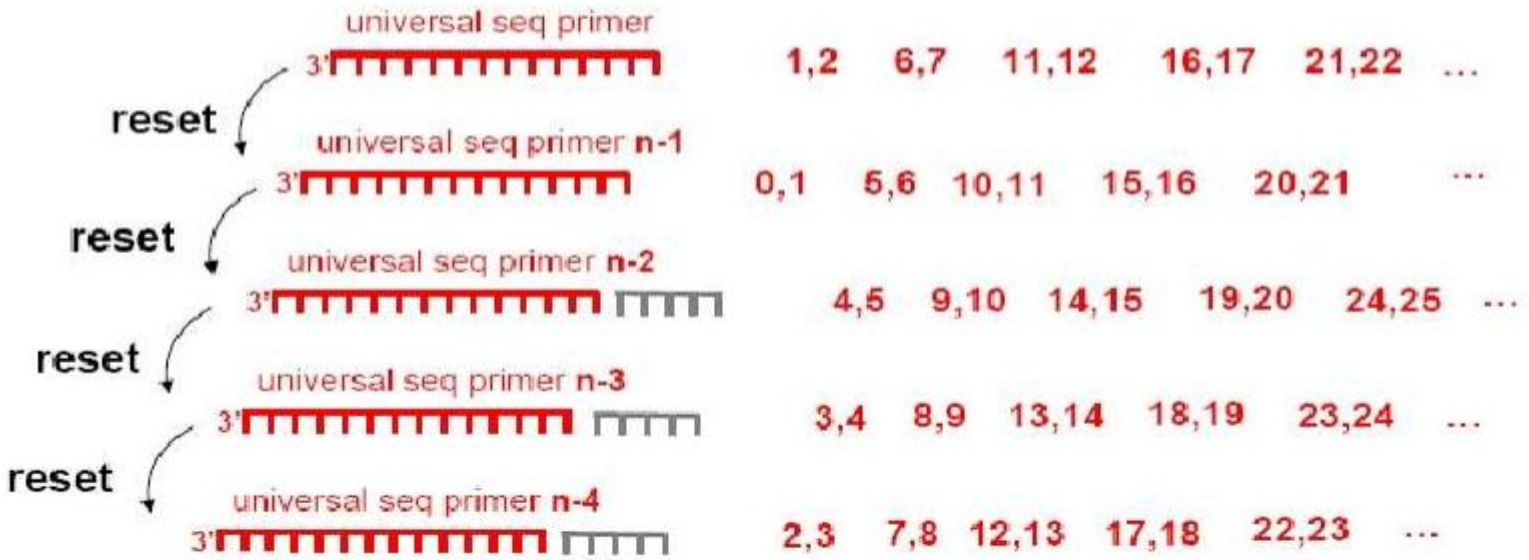
2nd cycle

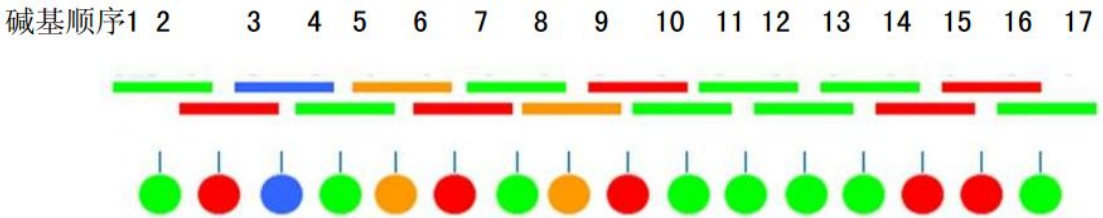
Part2

第二代测序

SOLiD

Applied Biosystems



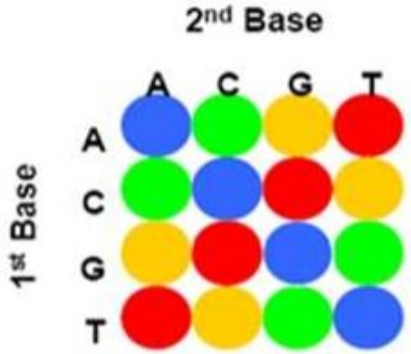


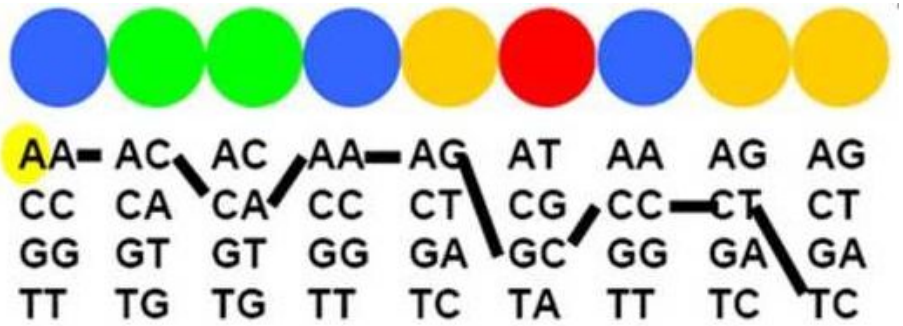
Part2

第二代测序

SOLiD

Applied Biosystems





Part2

第二代测序

SOLiD

Applied Biosystems

SOLiD特点：

- 通量高
- 准确率高，每个碱基检测2次，增加了序列读取的准确性
- 采用双碱基编码技术（two-base encoding），该技术具有误差校正功能，因为它是通过两个碱基来对应一个荧光信号而不是传统的一个碱基对应一个荧光信号，这样每一个位点都会被检测两次，因此错误率明显降低。

Part3

第三代测序

SMRT

PACIFIC BIOSCIENCES

第三代单分子测序技术，划时代的新里程碑，必将为研究开辟出新的领域和思路 不可错过

- **Pacific bioscience: PACBIO RS**
- **Oxford Nanopore**

Part3

第三代测序

SMRT

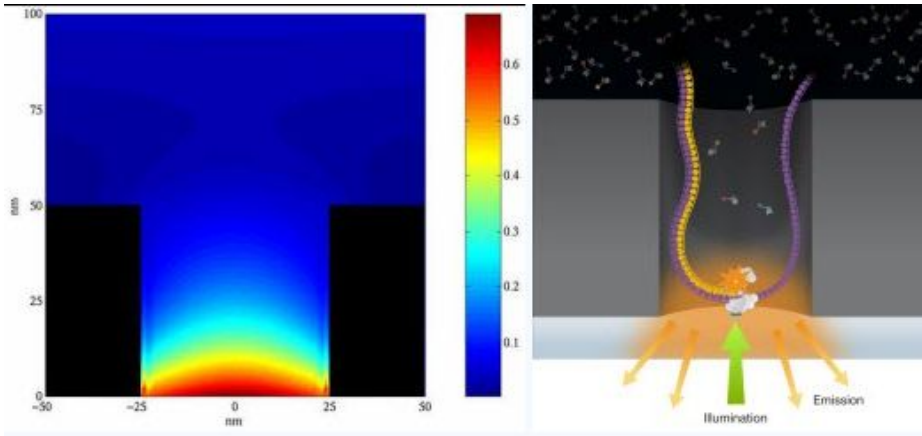
PACIFIC BIOSCIENCES

SMRT：Single Molecular Real Time Sequencing
PacBio RS：RS表示Real time Sequencing

测序原理：

被修饰的dNTP进入测序的纳米孔（ZMW）内，与固定在其中的额DNA聚合酶和DNA模板结合，释放荧光，荧光经CCD检测转为数字信号，进入电脑分析

ZMW（zero-mode waveguides，零模波导孔），在一个反应管（SMRTCell：单分子实时反应孔）中有15万个ZMW，外径100多纳米，比检测激光波长小，有效降低背景值



ZMW

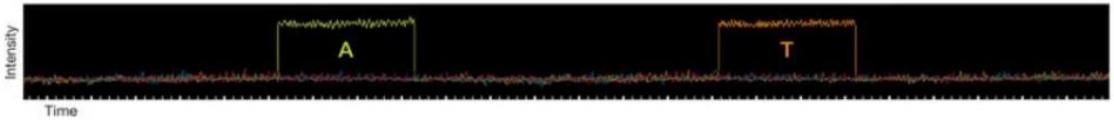
Part3

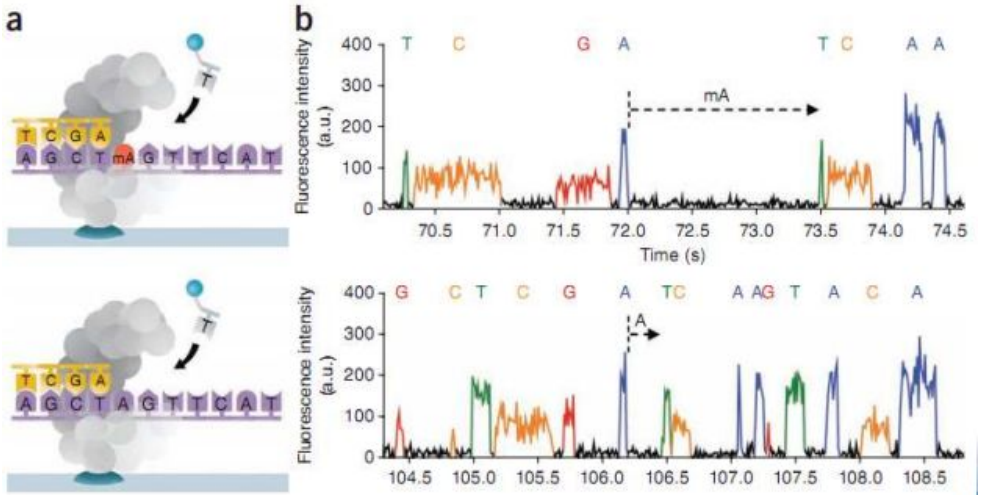
第三代测序

SMRT

PACIFIC BIOSCIENCES

sequencing





Part3

第三代测序

SMRT

PACIFIC BIOSCIENCES

PacBio RSII平台优势：

1. 测序速度快；
2. 读长超长，平均8Kb，最长60kb;
3. 精度高；
4. 可直接进行甲基化测序。

不足：
测序通量较低

Part3

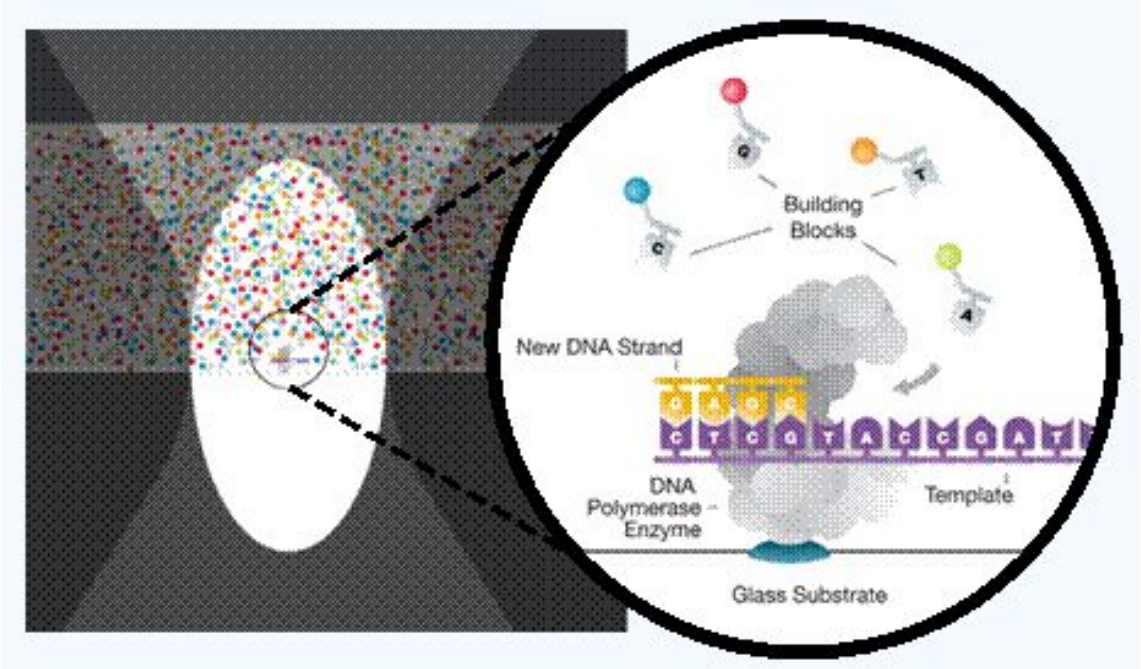
第三代测序

SMRT

PACIFIC BIOSCIENCES®

PacBio测序的技术创新：

- 酶学
- 表面化学
- 检测光学

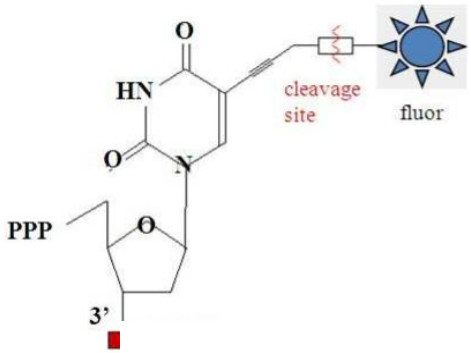


Part3 第三代测序 SMRT

PACIFIC BIOSCIENCES

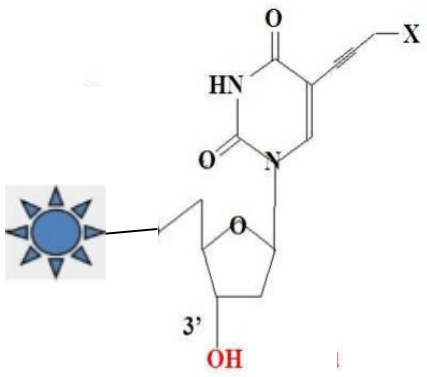
PacBio测序的技术创新：

1. 酶学　DNA聚合酶是实现超长读长的关键；
读长和酶的活性保持有关；



传统的核苷酸标记方法

DNA聚合酶的直径15nm，大分子荧光染料渗入DNA链会干扰DNA聚合酶的活性，造成聚合反应提前终止



PacBio技术创新

在核苷酸的磷酸链上进行荧光标记，一旦核苷酸渗入到新生DNA链中，DNA聚合酶将磷酸基团及其所带荧光标记一并切除，形成天然DNA链。脱落下来的荧光信号迅速衰减至基线以下。

此方法不仅不会影响DNA聚合酶活性，同时游离的荧光信号迅速衰减以降低背景噪音，有利于提高检测过程中的信噪比。

Part3

第三代测序

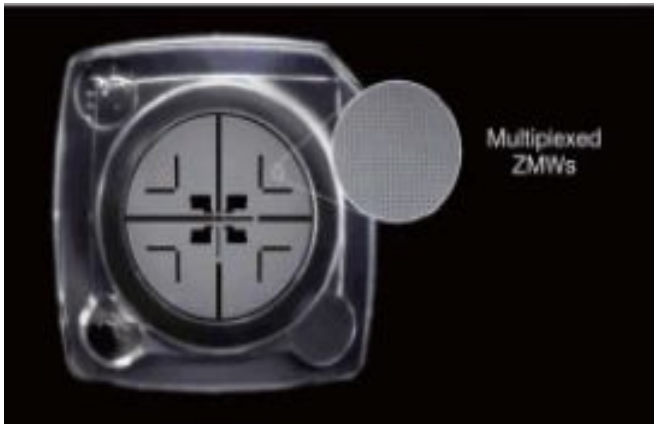
SMRT

PACIFIC BIOSCIENCES

实现**单分子、超长读长**测序的关键：

2. 表面化学问题

实现单个DNA聚合酶分子在基质表面的铆钉，继而完成DNA的合成反应



ZMW
在每个ZMW中，利用专利技术奖单个DNA聚合酶铆钉在底部玻璃的表面，随后dNTP涌入ZMW中，并在阵列表面
扩散

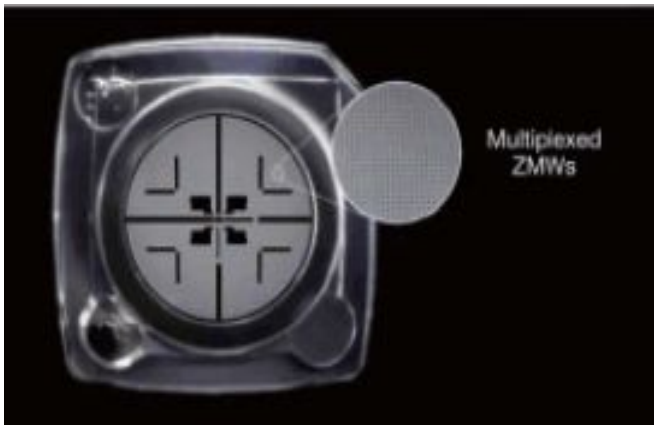
Part3 第三代测序 SMRT

PACIFIC BIOSCIENCES

PacBio测序的技术创新：

3. 检测光学

在DNA合成期间检测**单个**核苷酸的渗入



纳米级直径的ZMW，可阻止波长约为600nm的可见激光完全透过ZMW，造成可见激光在进入ZMW后迅速衰减，保证只有底部的30nm被照亮。

正确的核苷酸渗入新生链需要几毫秒，而单纯的核苷酸扩散只需要几微妙，这种时间差使渗入的核苷酸产生了类似于脉冲信号的高强度信号，该信号随即被转换成相应的碱基类型

Part3

第三代测序

SMRT

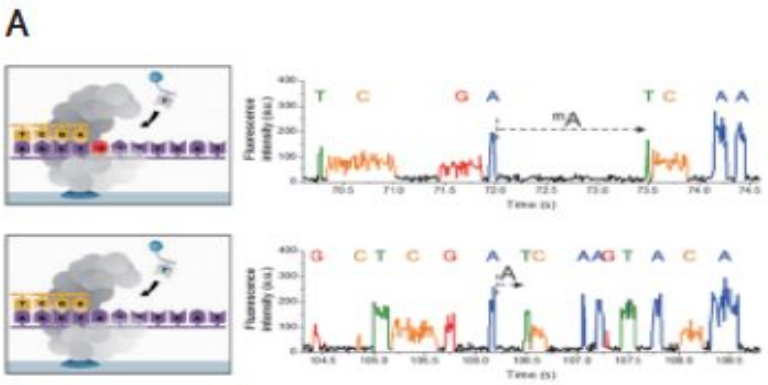
PACIFIC BIOSCIENCES

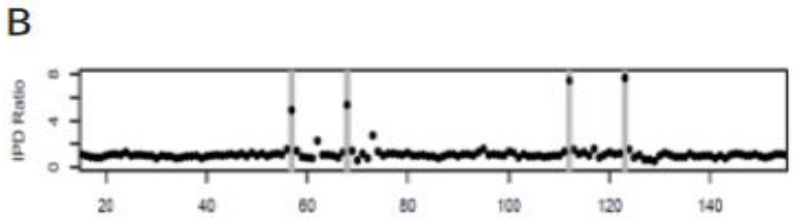
PacBio测序的技术创新：

4. 直接检测碱基修饰

原理：

荧光脉冲的到达时间和持续时间反映了有关聚合酶动力学的信息，修饰碱基相对于未修饰碱基有一个更长的脉冲间隔持续时间（IPD）



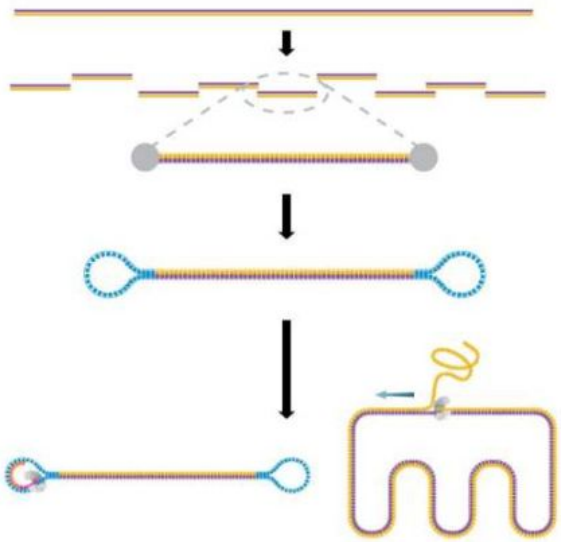


Part3 第三代测序 SMRT

PACIFIC BIOSCIENCES

PacBio测序的技术创新：

4. 长片段文库



通过Covaris的专利g-TUBE，通过离心剪切力，精确地将DNA打断成6~20kb的片段
无需PCR

Part3

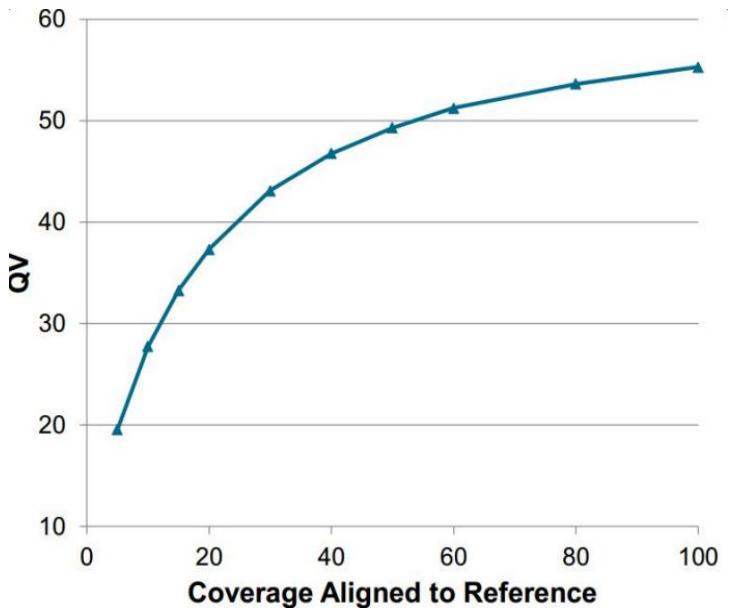
第三代测序

SMRT

PACIFIC BIOSCIENCES

PacBio测序的技术创新：

5. 高精确度



QV10代表90%准确度，20代表99%准确度，30代表99.9%准确度，40代表99.99%准确度，50代表99.999%准确度

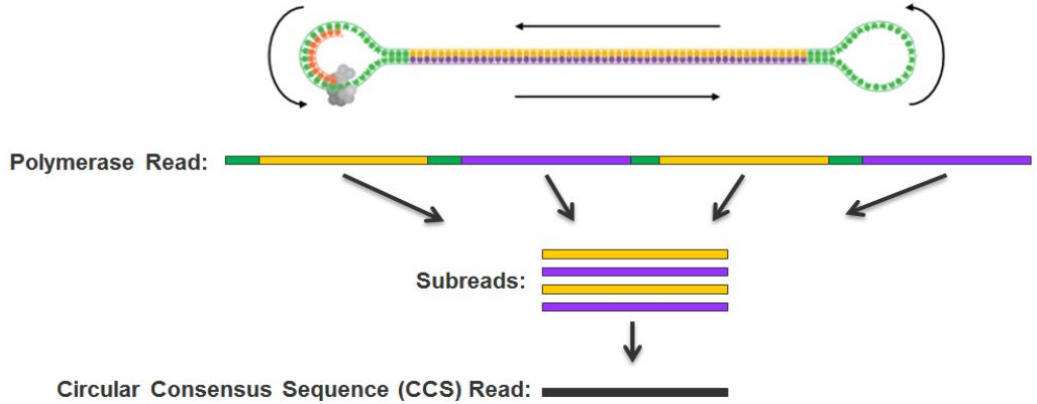
准确率这个概念本身就是指序列一致性，无论一代和二代测序的每一个反应，本来就是N个分子同时叠加反应所得到的平均信号，是一致性序列的结果。 单分子测序1 x覆盖度的精确度为87.5％，这是由于在测序过程中单个分子信号弱，偶尔会出现信号难于分辨的情况。出错几率是随机的，和序列长度、序列组成无关。要提高准确率，只需要提高循环次数，提高单分子覆盖度即可.

Part3

第三代测序

SMRT

PACIFIC BIOSCIENCES



Q:用PCR扩增结果测序是否能通过提高重复拷贝数而提高覆盖度，从而同时达到长片段和高度精确的目的？

A:是，可以通过提高重复拷贝数或对同一单分子环形测序两种方式，或二者结合，达到要求的覆盖度及准确度

如果需要很长的读取，策略是构建3 kb-10 kb的文库，就可以获得长的读长，这就是continuous longread模式。这种模式，很长的读长适合做全基因组序列组装骨架

Part3

第三代测序

Nanopore

Oxford NANOPORE Technologies®

Oxford Nanopore：

纳米单分子测序技术，基于电信号而不是光信号

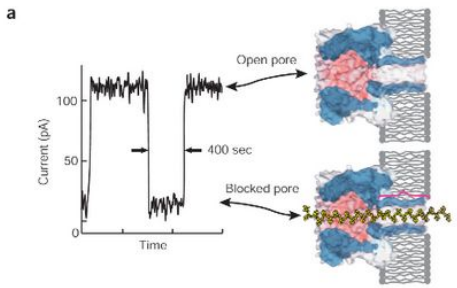
测序原理：

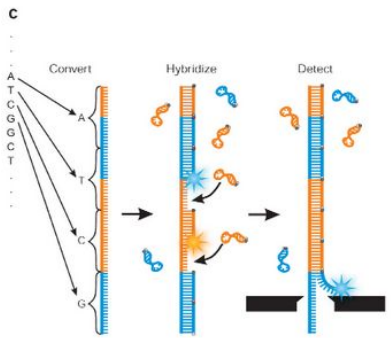
1.核酸外切

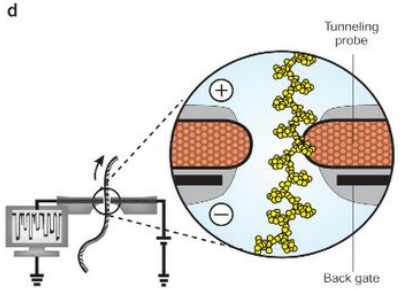
脂质双分子层上含有由 α -溶血素和环化糊精组成的纳米孔，每个孔中含有一个 E. coli 核酸外切酶 I，单链 DNA 通过纳米孔时被依次切成单核苷酸，记录这些单核苷酸引起的电流变化，便可完成 DNA 测序

2.链测序

利用 DNA 解旋酶将双链 DNA 解旋变成单链后通过纳米孔，进行连续测序







特点

读长很长，大约在十几Kb，甚至100Kb，随机错误，能检测甲基化胞嘧啶

总结

	平台	公司	原理	读长
第一代	Sanger	ABI	DNA聚合	1000bp
第二代	454	Roche	焦磷酸测序	~1000bp
	Solexa	Illumina	合成测序	2*150/2*250 /2*300
	Solid	ABI	链接测序	2*50
第三代	SMRT	Pacific bio	合成测序	~8kb
	Nanopore	Oxford	核酸外切	

04 Part Four 下机数据处理

Part1

Illumina数据处理

illumina

下机数据

二代测序得到的原始图像数据经过Base Calling转化为序列数据，结果以FASTQ文件格式存储，FASTQ文件为最原始的数据文件，文件包含测序read的序列信息以及测序质量信息。

```
@HWI-ST531R:144:D11RDACXX:4:1101:1212:1946 1:N:0:ATTCCT
ATNATGACTCAAGCGCTTCCTCAGTTTAATGAAGCTAACTTCAATGCTGAGATCGTTGA
+ HWI-ST531R:144:D11RDACXX:4:1101:1212:1946 1:N:0:ATTCCT
?A#AFFDFFHGFFHJJGIJJJIICHIIIIJJGGHIIJJIIJIIJIHGI@FEHIIJBFFHGJJIIHHHDFFFFDCC
```

Part1

Illumina数据处理

illumina

原始数据质量剪切：

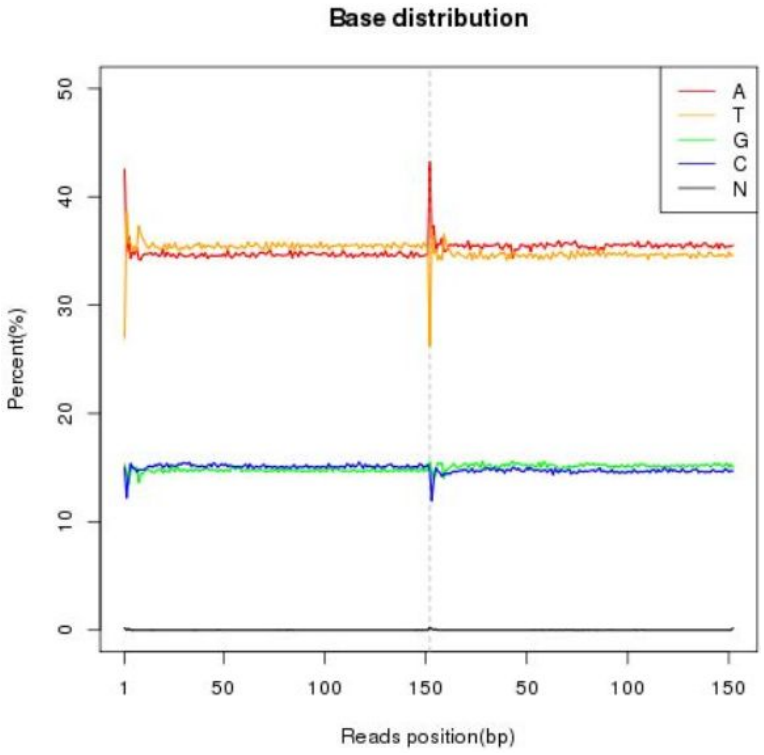
1. 去除 reads 中的 adapter 序列；
2. 剪切前去除 5’ 端含有非 A、G、C、T 的碱基；
3. 修剪测序质量较低的 reads 末端 (测序质量值小于 Q20)；
4. 去除含 N 的比例达到 10%的 reads；
5. 舍弃去 adapter 及质量修剪后长度小于 25bp 的小片段。

Part1

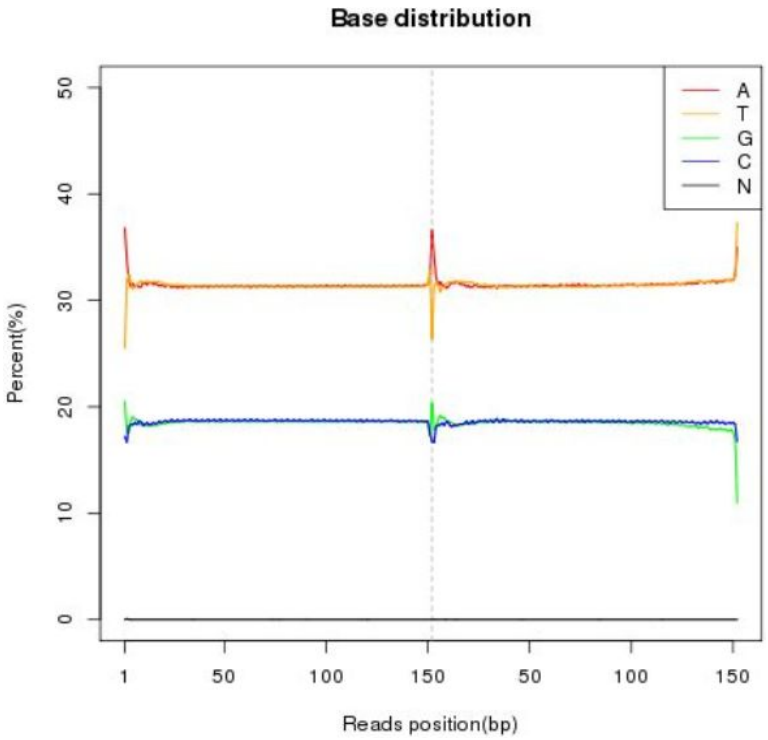
Illumina数据处理

illumina

数据质控——碱基分布统计



raw data

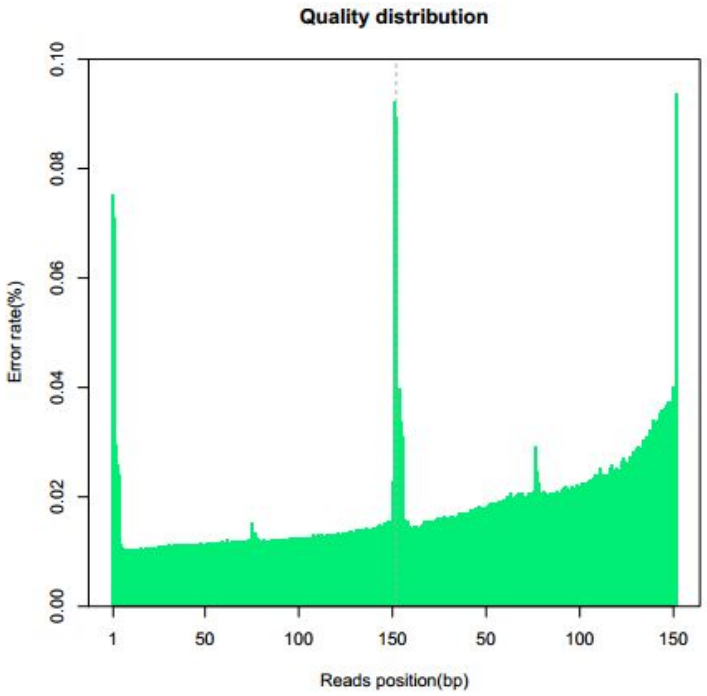


clean data

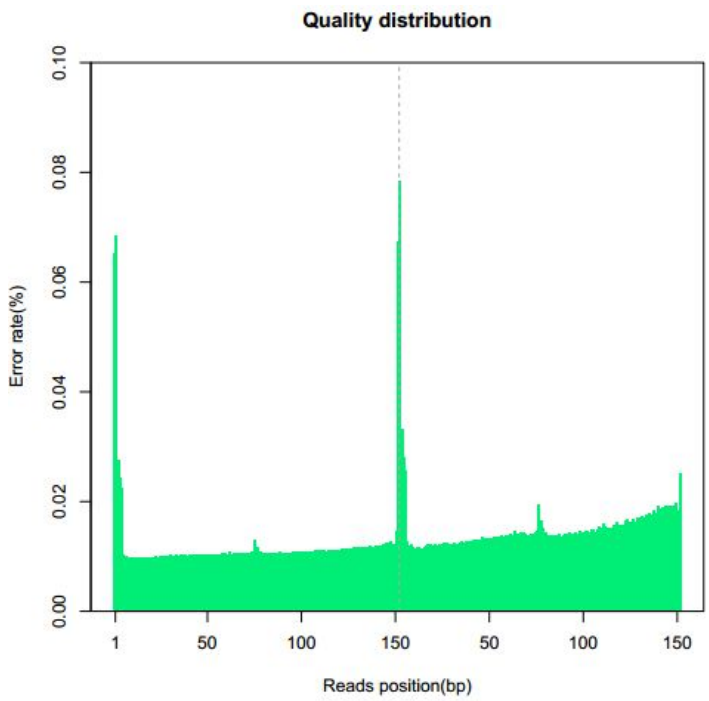
Part1 第二代测序 数据处理

illumina

数据质控——碱基质量统计



raw data

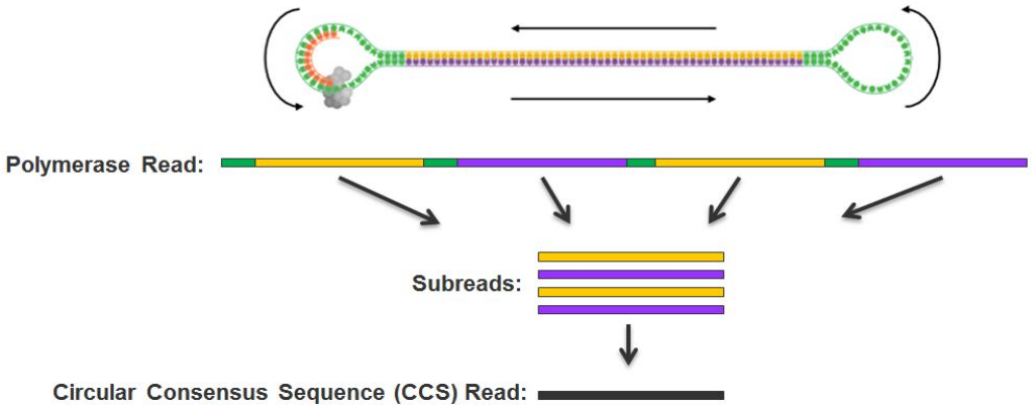


clean data

Part2

PacBio数据处理

PACIFIC BIOSCIENCES



下机数据处理：

PacBio 采用H5文件格式保存原始的下机数据

三代数据自纠 → 使用SMRT的HGAP，通过多重序列比对进行数据的自我校正

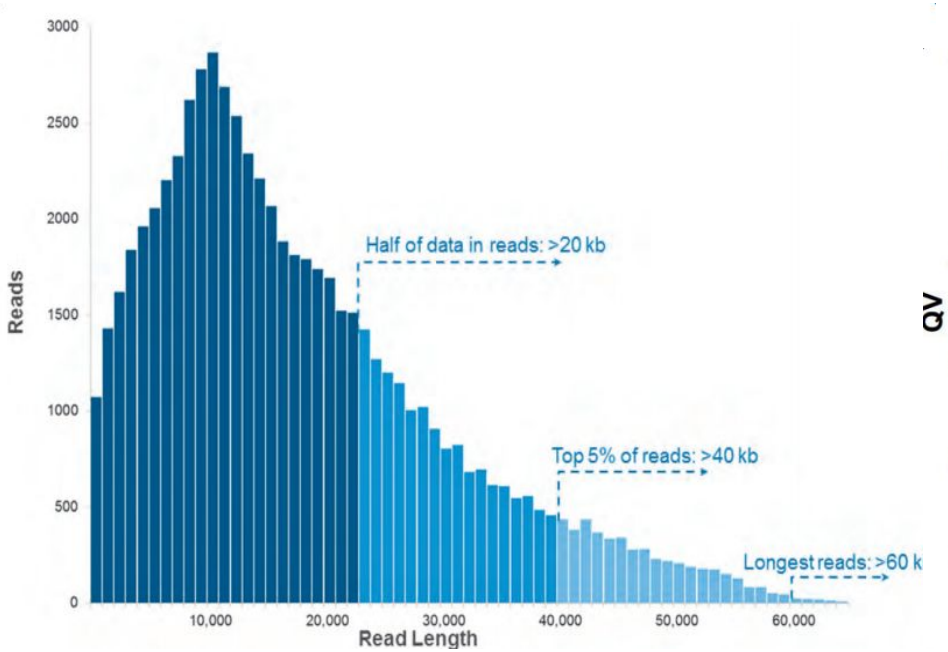
二代纠三代 → 测序同时加测二代数据，用二代数据校正三代数据，得到高质量的三代数据用于后续分析

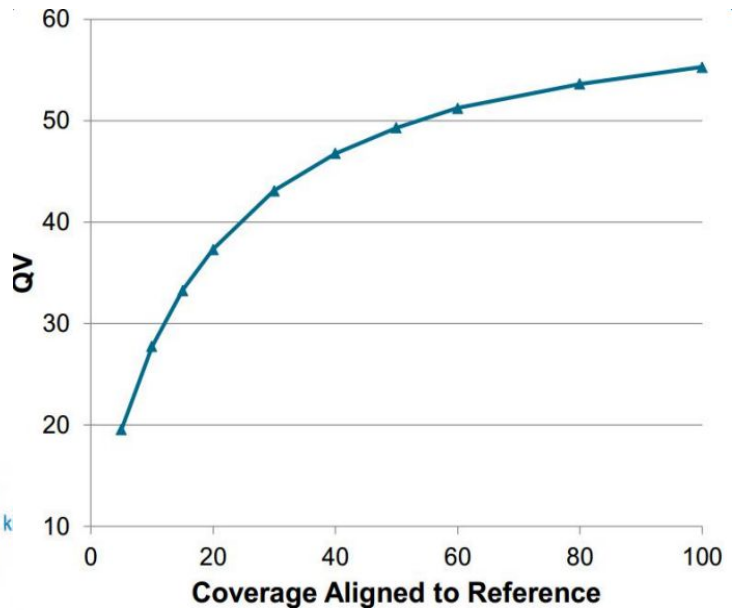
二代三代混合纠错 → 使用二代、三代混合拼接软件，将二代、三代数据数据混合后一起进行组装，在组装层面上进行数据的相互校正

Part2

PacBio数据处理

PACIFIC BIOSCIENCES



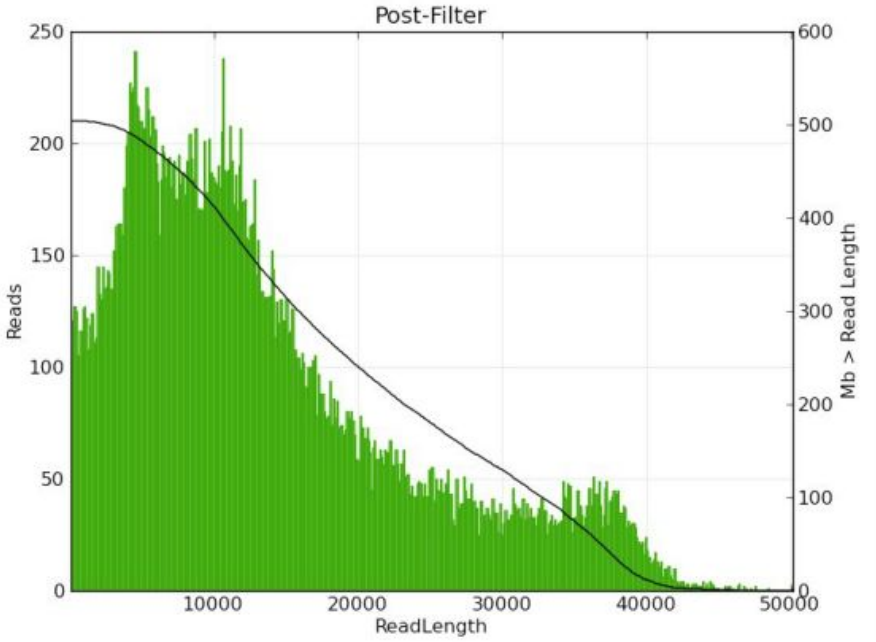


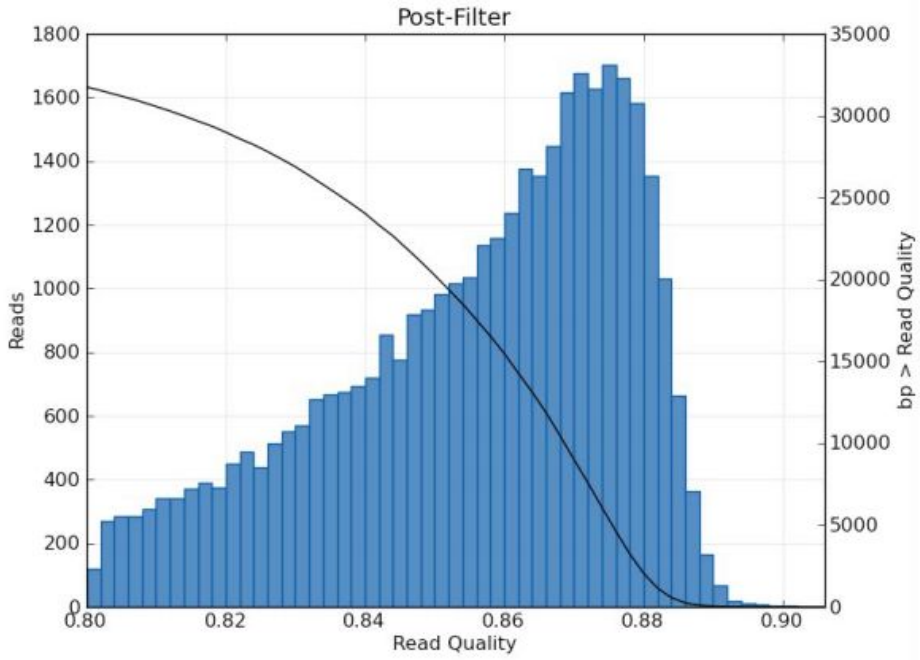
由上图可以看出，当测序深度达到60×的时候，测序的错误率可以达到万分之一以下，美吉为客户提供≥100×的测序，保证基因组的准确组装

Part

PacBio数据处理

PACIFIC BIOSCIENCES





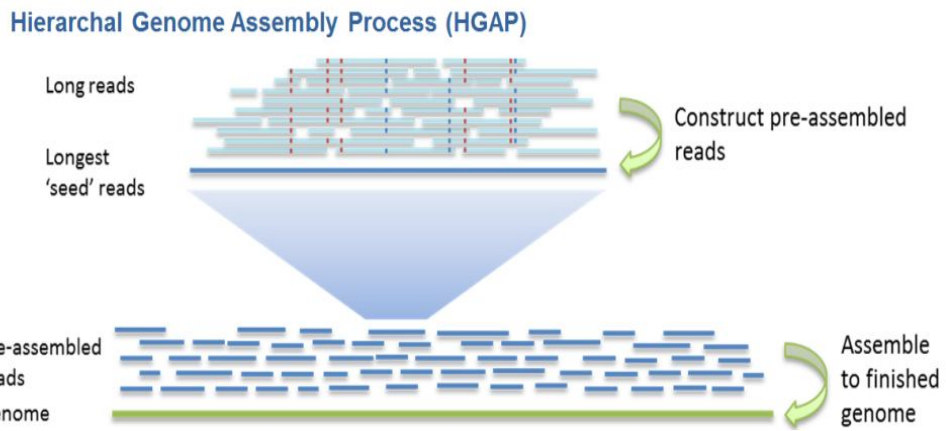
测序reads 的长度是评估三代测序优劣的重要指标

上图为单分子测序 Clean reads 的长度和质量分布统计图。

Part

PacBio数据处理

PACIFIC BIOSCIENCES

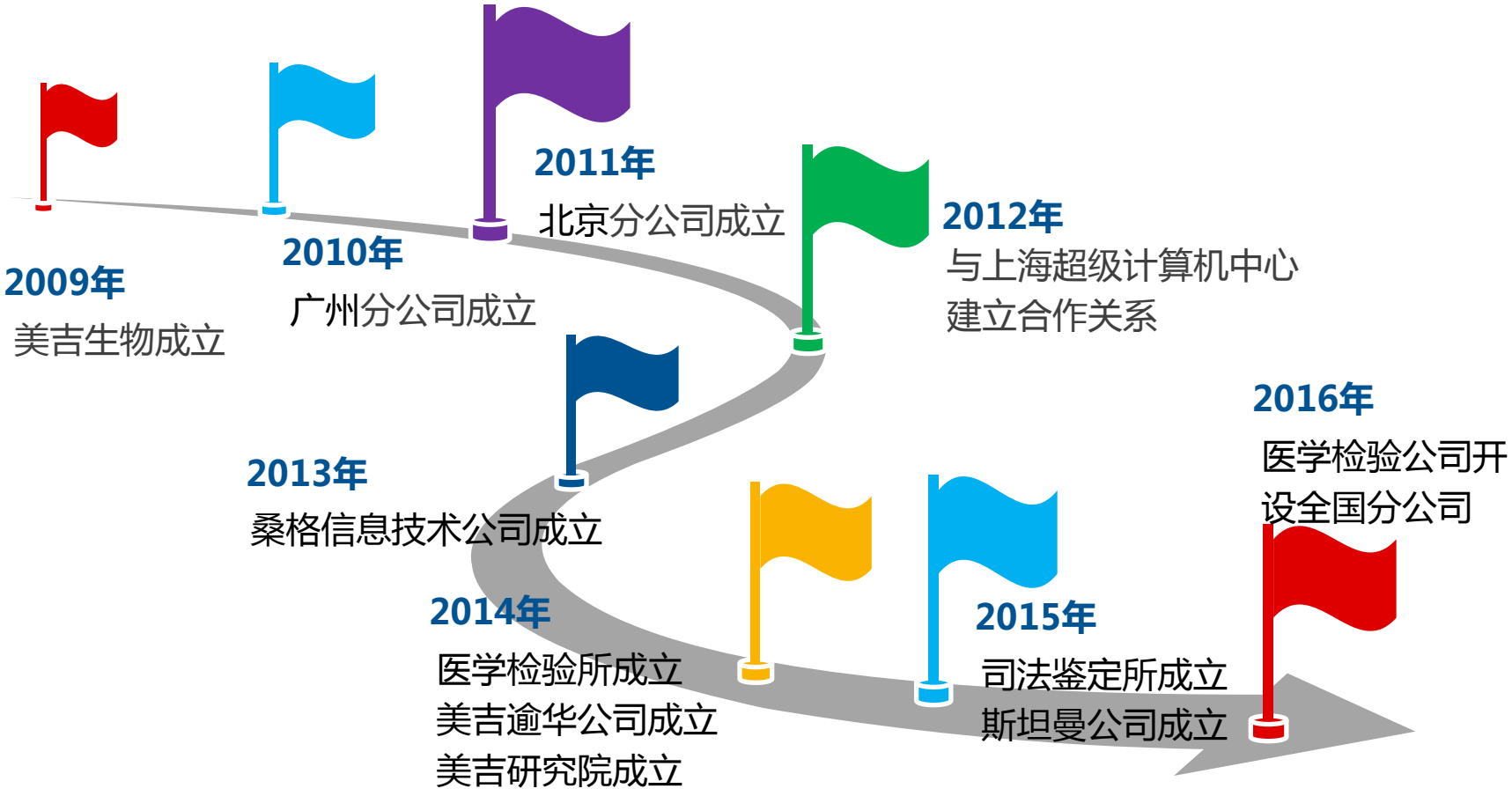



```
>genome
AAGGAGTGACGTCTGTGGACAGCCATACCTCTGAACTATGGCAGCAAATTCTATCCATCATACAAACCAAGCTGAGTAAGCCGAGTT
ACGACACTTGGTTTAAGGCTACCAAGGCAGCGAAACTAAATGACCACTCCATTGTGATTTCTGCACCGACAACTTTTGCCGTGGAAT
GGCTTGAAAGCCGCTATACCAAGTTAGTCGGAGCAACGGTTTATGAGATTTTGGGCAAACATTTAGAGGTCAAGTTCGTTATTGAAG
AGAACAAGCCCGCCGAGGTCGACCTTCAGCAACAACCTCAGCAGCAGCCGGTCGTTCAAGAAGAAGCTGTATCCCATATGCTGAATC
CCAAATATACATTCGATACATTTGTCATTGGATCGGGGAACCGTTTTGCCCACGCGGCATCGCTGGCCGTCGCCGAGGCGCCGGCCA
AAGCTTACAATCCGCTATTTCTATACGGTGGTGTAGGTCTGGGGAAAACGCATCTGATGCATGCTATCGGTCACTATATTTTGGAGC
ATAATCCGACCAGCAAGGTCGTTTATTTATCGTCGGAGAAGTTTACGAACGAGTTCATTAATGCCATCAGGGACAACCGCGGAGAGA
GTTTCCGGAATAAATATCGCAACATTGATATTTTGCTCATTGATGATATTCAATTCATTGCGGGCAAGGAATCGACGCAAGAGGAAT
TTTTCCATACGTTCAACGCGCTTCATGAGGAACGCAAGCAGATTATAATCTCAAGCGATCGACCGCCTAAAGAAATTCCAACGCTGG
AAGAACGGCTGCGCTCTCGCTTCGAGTGGGGACTTATTACGGATATTCAACCGCCGGATTTGGAGACGAGAATTGCTATTCTTCGGA
AAAAGGCGCGGGCAGAAAACCTGGATATTCCTAATGAGGCCATGATGTATATCGCGAATCAAATTGATACAAACATCCGTGAGCTGG
AGGGGGCGCTCATTCGCGTTGTCGCTTATTCTTCCTTAACCAATCAGGATGTATCAAGCCATCTCGCGGCTGAGGCGTTAAAGGATA
```

05 Part Five
美吉与高通量测序

关于美吉

发展历程


2009年
美吉生物成立
2010年
广州分公司成立
2011年
北京分公司成立
2012年
与上海超级计算机中心
建立合作关系
2013年
桑格信息技术公司成立
2014年
医学检验所成立
美吉逾华公司成立
美吉研究院成立
2015年
司法鉴定所成立
斯坦曼公司成立
2016年
医学检验公司开
设全国分公司


美吉生物 Majorbio
科技服务
生物
云计算
医学诊断
司法诊断
美吉
生物集团
质谱
生物医药
生物医疗
研究院
I-Sanger
美吉医学检验 MAJORDX
上海美吉医学检验有限公司
法医物证司法鉴定所
美吉谕华 YH·Health
斯坦曼生物医药 Steinman-Biomed
美吉生物信息研究院
Bio-Informatics Institute

高通量业务 —— 一站式服务


代谢组
蛋白组
转录组
……
基因组
动植物基因组
细菌基因组
真菌基因组
线粒体
叶绿体
病毒
质粒测序
真核生物
原核生物
小RNA测
长链非编码RNA
CirRNA测
链特异性
CeRNA测
全转录组
全长转录
甲基化
转录调控测序
简化基因
遗传图谱
GWAS
BSA
全基因组
群体重测序
16S测序
18S测序
ITS测序
功能基因
宏基因组
宏转录组
宏蛋白组
微生态产品
动植物基因
novo测序
动植物基因测序
转录组测序
基因表达谱
Small RNA测序
简化基因组
RAD测序
全基因组甲测序
蛋白组学测序
外显子
目标区
全基因
人方向
Sanger测序服务
分型检测服务
微生物菌株鉴定服务
TA克隆
基因定点突变服务
全基因合成服务
实时荧光定量PCR技术服务
常规测序及其他

迄今为止，我们的合作单位已达1900余家，合作的课题组已超过6800个……

我们已助力各位客户发表了600+篇SCI论文，总影响因子超过1700……

"M" (majorbio)
"W" (we)
美吉生物® Majorbio 美吉
生命 科技

- 上半部分是M(majorbio)象征美吉，下半部分是W（we）象征我们，寓意我们众志成城、齐心协力撑起一个伟大的美吉。
- 绿色代表生命，蓝色代表科技，寓意我们从事的是基因科技与人类健康事业。

021-31050579
mdna@majorbio.com
2406642459
Majorbio

美吉生物官方微信平台

热诚欢迎您随时垂询！

谢谢您的欣赏

美吉生物
Majorbio

地址/Add:上海市浦东新区国际医学园区康新公路3399号3号楼
电话/Tel：021-51875086　　服务热线：400 660 1216
网址/Web：www.majorbio.com　　传真/Fax：021-51875086-8002