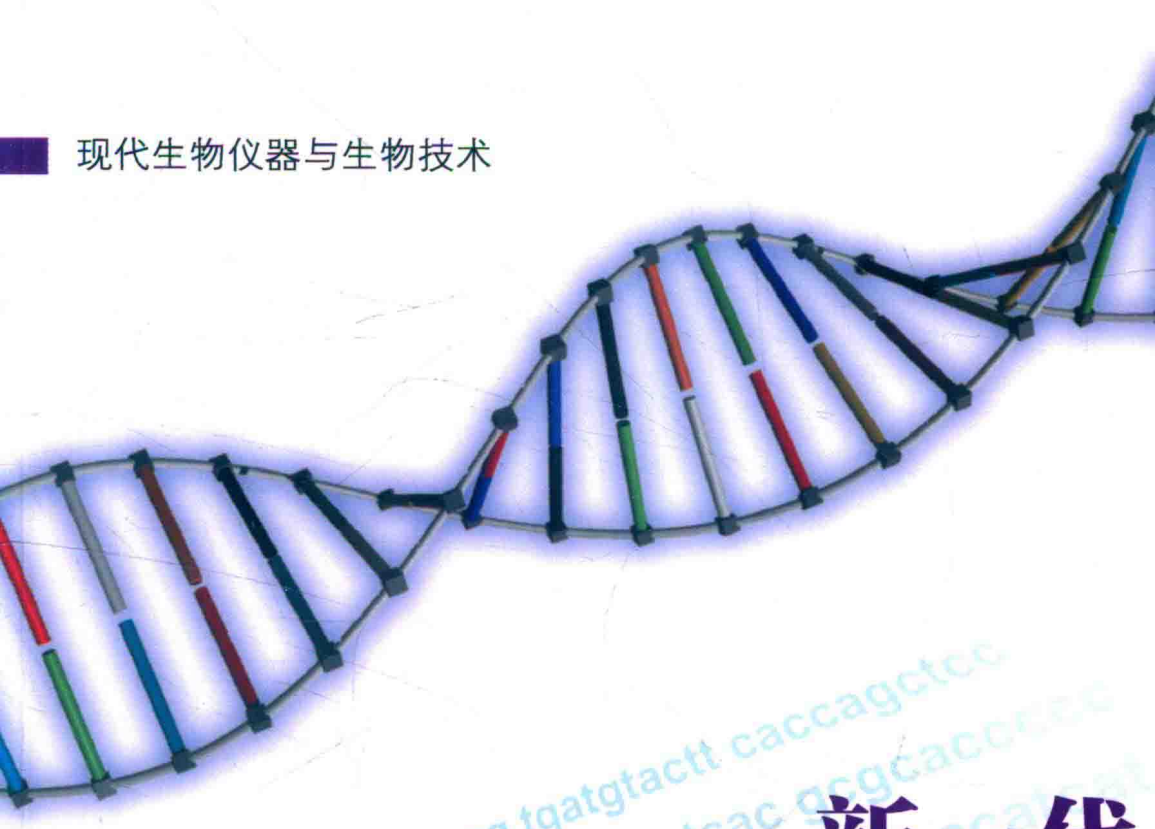
现代生物仪器与生物技术



新一代
基因组测序技术

陈浩峰　主编

科学出版社

新一代基因组测序技术

陈浩峰　主编

科学出版社
北京

内 容 简 介

新一代基因组测序技术是目前生命科学领域中发展非常快的新技术，它的出现极大地推动了生物学、农学和医学诊断等学科各方面的发展，其应用非常广泛。本书以新一代基因组测序技术 Illumina 平台和 PacBio RS 平台为代表，详细阐述了从实验样品处理到数据分析的整个过程，重点介绍了新一代测序技术实践过程中的文库构建与质检、测序仪器操作和测序数据的初步处理及分析。本书是国内第一本详细阐述新一代基因组、转录组等测序技术的中文书，包含了目前最新的技术资料，极具实用性和可操作性。

本书是高等院校师生和科技工作者学习新一代测序技术的最佳参考书。

图书在版编目（CIP）数据

新一代基因组测序技术/陈浩峰主编. —北京：科学出版社, 2016
ISBN 978-7-03-048791-9

I.①新… II.①陈… III. ①基因组–序列–测试–研究
IV.①Q343.1

中国版本图书馆 CIP 数据核字(2016)第 131838 号

责任编辑：罗 静 田明霞 / 责任校对：王 瑞
责任印制：徐晓晨 / 封面设计：北京铭轩堂广告设计有限公司

科 学 出 版 社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京京华虎彩印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2016 年 6 月第 一 版 开本：720×1000 1/16
2016 年 9 月第二次印刷 印张：22 1/2
字数：450 000

定价：128.00 元

(如有印装质量问题, 我社负责调换)

《新一代基因组测序技术》编者名单

主编 陈浩峰

编者(按姓氏拼音排序)

曹英豪　陈　旭　陈浩峰　高　强　韩　瑶

雷　猛　李　妍　李　珍　孟　菲　齐　洺

王　静　王剑峰　杨　鑫　于　莹　张　兵

前　言

自从2005年新一代基因组测序技术诞生以来，短短十余年间，其发展日新月异。今天，新一代测序技术已经被广泛应用到生命科学研究的各个方面，可以说，该技术的迅速发展标志着生命科学组学研究时代的到来。目前，学习、掌握和应用新一代测序技术已经成为广大生命科学工作者的迫切需求，对于国内开展新一代测序研究的实验室来说，手头有一本详细介绍新一代测序各方面技术的中文书籍非常必要。鉴于此，我们编写了本书，供生命科学有关专业的高等院校师生、科研院所的研究人员及其他生命科学从业人员参考。

本书以新一代基因组测序技术Illumina平台和PacBio RS平台为代表，详细阐述了从实验样品处理到数据分析的整个过程，重点介绍了测序过程中的文库构建与质检、测序仪器操作和测序数据的初步处理及分析。全书共分为五章：第一章为绪论，概述了测序技术发展简史，以及各个主要的新一代测序平台的技术特点；第二章到第四章详细介绍了目前新一代测序技术的主流平台——Illumina测序技术的各个方面，第二章介绍了Illumina建库技术，第三章介绍了Illumina测序操作，第四章介绍了Illumina测序数据的初步分析；第五章则是针对单分子测序技术PacBio RS的建库、测序和数据分析过程的介绍。本书的编写人员都是工作在科研第一线，有着多年第一代和新一代测序实践经验和数据分析经验的中青年科研工作者，希望我们从科研工作中得来的经验和体会对广大读者学习新一代测序技术起到借鉴和参考作用。

我们编写本书的目的是，希望具有一定分子生物学与遗传学背景的读者，通过阅读本书了解新一代测序技术的概念及其发展历史，以及各种新一代测序技术平台所独有的特点；并且能够参照书中所述的建库流程指导，在自己的实验室成功完成测序文库的构建和质量检测。此外，我们还希望在有条件进行新一代测序工作的实验室，读者可以参照本书了解Illumina和PacBio RS测序仪的基本工作流程和测序数据的初步分析过程。这样一来，读者在进行此类科研项目时，就可以做到对研究规划心中有数，在实验设计上有的放矢；而不是把测序建库、测序操作和数据分析过程完全交给测序服务公司，把这一部分工作作为“黑箱”来对待。对于其他各种生物学、医学等实验室的研究人员和工作人员来说，即使没有机会上机操作，至少也可以通过阅读本书获得一些新一代测序的知识，了解各种

新一代测序方法的优点与局限性，并且有能力进行测序项目实验的追踪纠错。

感谢中国科学院遗传与发育生物学研究所和北京基因组研究所提供的良好科研工作条件；特别感谢遗传发育所基因组生物学研究中心与植物基因组学国家重点实验室对编写工作的支持；感谢 Illumina 中国公司、New England Biolabs(NEB)中国公司和凯杰(Qiagen)中国公司的大力赞助；感谢本书责任编辑罗静女士的出色工作；最后还要感谢我海内外的亲朋好友与同行，他们在本书的写作过程中提出了宝贵的修改意见。没有上述机构和人员的共同努力，本书不可能很快面世。

限于编者的知识水平，本书所列举的测序研究案例主要集中于植物学、农学与育种学等方面。目前，新一代测序在生命科学的各个领域，尤其是医学研究和疾病的临床诊断中的应用越来越广泛，由于我们缺少这方面的研究和诊断实例，本书对测序诊断方面的应用涉及不多，这是我个人感到遗憾的地方，希望将来有机会增添这部分的内容。

衷心希望广大读者对本书内容提出宝贵意见，以便于我们将来修订和提高。

陈浩峰

2016 年 1 月 5 日于北京

目　录

第一章　测序技术发展概述

随着现代科学技术的发展，生命科学的研究已经进入了组学时代。基因和基因组测序技术已经成为现代生命科学研究，特别是基因组学研究中不可或缺的手段。近年来新一代基因组测序技术突飞猛进的发展带来了基因组学研究的空前繁荣。

自从 1977 年 Fredrick Sanger 等建立了双脱氧链终止法（dideoxy chain-termination method）测序技术以来，基因测序技术经历了几十年的快速发展，在此期间出现了两次技术上的飞跃：第一次飞跃是 Sanger 测序技术实现了大规模测序的自动化，科学家利用该技术完成了“人类基因组计划（Human Genome Project，HGP）”等重大科学研究项目；第二次飞跃是自 2005 年以来，以 Roche 454、Illumina GA/HiSeq、Life SOLiD/Ion Torrent、PacBio RS 为代表的新一代测序技术（next-generation sequencing，NGS）的出现，使得基因组测序通量快速增加，测序成本极大降低。近 10 年来，测序技术的发展速度已经远远超越了半导体信息技术进步的速度（摩尔定律，Moore's Law），它将生命科学研究带入了基因组学时代。新一代测序技术已经在生命科学各个领域及农学、医学、环境保护、法医学等领域中得到了广泛的应用。

那么，新一代测序是怎样在实验室中实现的？测序实验成败的关键是什么？测序数据的产生过程是怎样的？拿到海量的测序数据之后该怎样处理？这些都是正在使用和将要学习使用新一代测序技术的广大科研人员及生命科学相关专业的高等院校师生所关心的问题。目前，对于国内绝大多数生物学和医学实验室来说，新一代测序仪和与测序相关的实验技术仍然是比较昂贵和陌生的，因此，我们编写这本书，向读者介绍一些新一代基因组测序技术的原理、测序实验操作和初步的数据分析方法，就显得十分必要了。

在本书的各个章节中，我们将针对测序技术的发展历史、测序建库、测序仪器操作、数据的初步分析等各个方面逐一为读者详细介绍。

第一节　第一代基因测序方法简介

第一代基因测序方法，即双脱氧链终止法，是由 Fredrick Sanger 等在 1977 年创立的（Sanger et al.，1977），因此，也被称为“Sanger 测序法”。该方法是一种基于 DNA 聚合酶合成反应的测序技术。其测序原理可以简述如下（图 1.1）：在 4 个

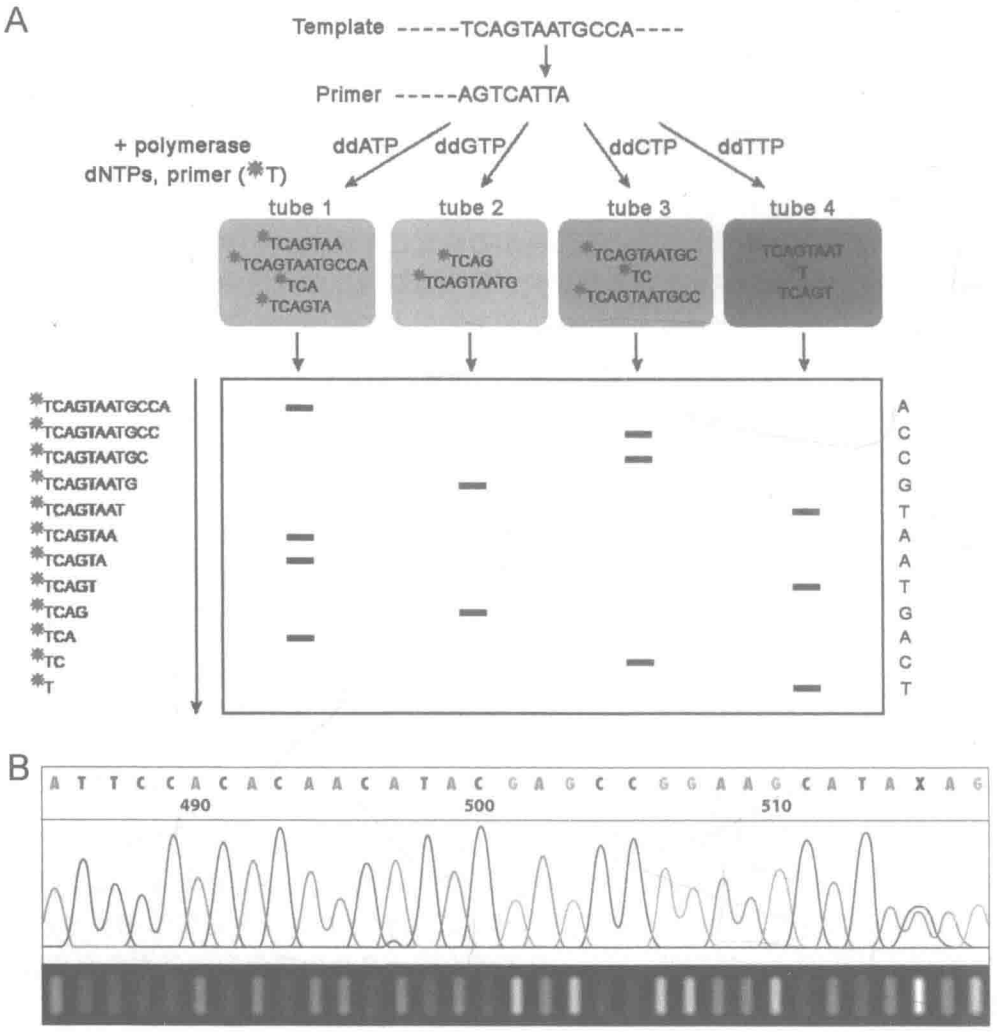


图 1.1　双脱氧链终止（Sanger）测序原理图（选自 Sequencing forensic analysis and genetic analysis 和 Genes and genomics：a short course（3e），略有改动）

A. 454 测序流程图：在 4 个测序反应系统（tube 1、2、3 和 4）中加入待测 DNA 模板、DNA 合成酶、dNTP、反应引物及带有放射性同位素的 ddATP、ddCTP、ddGTP、ddTTP。经过 DNA 合成反应后，就形成了一组长度差为一个核苷酸的 DNA 片段。聚丙烯酰胺凝胶电泳放射自显影后，根据电泳所得到的 DNA 片段大小可反向依次读出被合成的碱基排列顺序，从而得到待测的 DNA 序列。B. 基于“Sanger 测序法”的荧光自动 DNA 测序仪直接将信号转化为 DNA 序列的显示图

测序反应系统中加入待测的 DNA 模板、DNA 合成酶及 DNA 合成反应所需的其他成分，如脱氧核苷三磷酸（dNTP）、反应引物和缓冲液等，并且将少量的 4 种带有放射性同位素的双脱氧核苷三磷酸（ddATP、ddCTP、ddGTP、ddTTP）按一

定比例分别加入相应的反应系统中，然后进行 DNA 合成反应。因为 ddNTP 中包含的是双脱氧核糖，其 3 位碳原子上连接的不是羟基（—OH），而是脱氧后的氢（—H），所以当 ddNTP 被加入到正在合成的 DNA 链中后，系统中后续的 dNTP 就不能再被结合到这条 DNA 链上了，这条 DNA 链的合成就会随机终止在任何碱基处。这样，经过几十个循环的合成反应后，就形成了一组由短到长的 DNA 片段，这些片段之间的长度差为一个核苷酸，并且 3′端碱基以带有放射性同位素标记的 A、C、G 或 T 作为结束。测序合成反应终止后，将合成的产物分为 4 个泳道进行聚丙烯酰胺凝胶电泳，电泳结果经过放射自显影处理后，根据电泳所得到的 DNA 片段大小来排列反应产物带有的末端双脱氧核苷酸类型，即可反向依次读出被合成的碱基排列顺序，从而得到待测的 DNA 序列。

此后，人们在上述最初的“Sanger 测序法”基础上发展出多种 DNA 测序技术，其中最重要的是荧光自动检测技术。该技术基于 Sanger 测序原理，用荧光标记代替同位素标记，并用成像系统自动检测，从而大大地提高了 DNA 测序的速度和准确性。代表性的测序仪器如 ABI 3730XL 测序仪拥有 96 条电泳毛细管，4 种双脱氧核苷酸的碱基分别用不同的荧光进行标记，在通过毛细管末端时由激光激发不同的 DNA 片段上的 4 种荧光基团，从而发出不同颜色的荧光，荧光信号被 CCD（charge coupled device）照相检测系统识别后直接将信号转换成为 DNA 序列。Sanger 测序法在出现之后的大约 30 年间，因其操作简便、测序读长长（为 800 bp～1 kb）、数据准确性高，一直是应用最为广泛的 DNA 测序方法，甚至至今仍是验证新一代测序结果的金标准，常用于验证由新一代测序方法发现的新变异位点。

第一代测序技术的产生，使人们拥有了“阅读”生物基因组秘密的有力工具，在 20 世纪末和 21 世纪初的几年间，科学家利用第一代测序技术完成了一系列物种的全基因组测序，如水稻（Goff et al.，2002）、拟南芥（The *Arabidopsis* Genome Initiative，2000）等模式植物和秀丽线虫（The *C. elegans* Sequencing Consortium，1998）、果蝇（Adams et al.，2000）等模式动物的基因组图谱。该技术最大的成就是保证了“人类基因组计划（Human Genome Project，HGP）”的顺利实施（Lander et al.，2001），这项跨国研究计划开始于 1990 年，2000 年美国国立卫生研究院（National Institutes of Health，NIH）和美国 Selera 公司共同宣布人类基因组草图绘制成功；2003 年由美、日、德、法、英、中六国科学家宣布人类基因组序列图谱绘制成功。人类基因组计划历时 13 年，花费约 30 亿美元，由全世界几千个实验室协力共同完成。美国著名的《时代》（*TIME*）杂志在 2000 年发表文章评论这项成就的意义时这样写道，“……无论怎样评价这项成就都不为过。以遗传密码作武器，科学家现在可以以很轻松的方式（teasing out）在分子水平上获得人类健康和疾病的秘密——至少可以在阿尔茨海默病、心脏病和癌症的诊断和治疗等方面

引发一场革命……历史将记载下这一基因组时代开启的时刻”。

尽管 Sanger 测序法至今仍然被公认为测序的“金标准”，但是它也存在着相当大的局限性。第一是“测序偏好（sequencing bias）”，由于 Sanger 测序法是将待测 DNA 加入到载体（vector）上并在大肠杆菌（*Escherichia coli*）等细菌中进行克隆，因此被克隆的 DNA 不能对细菌有害，并且要与细菌 DNA 的复制机制兼容。测序实验证明，基因组的某些区域，如着丝点和端粒附近的区域很难被克隆，从而导致在基因组测序数据中出现缺失（gap）。第二是 Sanger 测序法处理和分析等位基因频率的能力有限，用这种测序方法在 PCR 扩增产物中发现并区分杂合的单核苷酸多态性（single nucleotide polymorphism，SNP）是很困难的。第三是 Sanger 测序法通量太低，从而导致基因组测序实验成本过高。据估算，用 Sanger 测序法完成一个人基因组（约为 3×10^9 个碱基）的重测序大约需要 1000 万美元，这样就使得一般的实验室无力单独承担大规模的测序实验研究项目。

第二节　新一代测序技术概述

人类基因组计划的顺利实施，是全世界科学家利用第一代测序技术所取得的辉煌成就，同时也标志着生命科学研究进入了后基因组时代，即功能基因组时代。传统的第一代测序技术因其通量低、成本高和时间长的局限性，已经不能满足生物物种的深度测序和重测序等大规模基因组测序的需要，这就促使了新一代基因组测序技术的诞生。依照其在测序市场上出现的时间顺序，新一代测序技术包括 Roche 454 公司的 Genome Sequencer FLX 测序平台，Illumina 公司的 Genome Analyzer、HiSeq 系列、MiSeq、NextSeq500 和 MiniSeq 等测序平台，Life 公司的 SOLiD 测序平台、Ion Torrent Personal Genome Machine（PGM）、Proton 等测序平台，Helicos Biosciences 公司的 Heliscope 测序平台，Pacific BioSciences 公司的 RS 测序平台，以及新近出现在测序市场上的 Oxford Nanopore 公司的 MinION、PromethION、GridION 等测序平台。新一代测序技术最显著的特点是通量高，单碱基测序成本低，一次测序运行可以对几十万至数亿条 DNA 模板进行测序。利用这些特点，人们可以方便地对各种生物物种进行全基因组深度测序、转录组测序、甲基化测序和 ChIP 测序等研究。

在第一代测序技术中，测序合成反应，即通过测序 PCR 产生长度不同（不同片段之间相差一个脱氧核苷酸）的扩增片段，与序列读取（通过电泳方法分离与检测片段长度）及产生序列数据的过程是分离的。与之相比，新一代测序技术通常又被称为大规模平行测序技术（massively parallel sequencing，MPS），它可以同时完成测序模板互补链的合成与序列数据的读取。一般来说，新一代测序包含下

列连续的步骤：①向测序系统加入脱氧核苷酸；②检验和确定被加入的脱氧核苷酸类型；③去除测序反应的各种酶、荧光标记物或脱氧核苷酸的 3′阻断基团等的洗脱反应（Zhang et al.，2011）；这样就实现了“边合成边测序（sequencing by synthesis，SBS）”，如 454、Illumina、Ion Torrent 和 PacBio 等测序技术；或者“边连接边测序（sequencing by ligation，SBL）”，如 SOLiD 技术。

新一代测序技术的发展初期，得到的测序读长相比第一代测序数据（0.8～1 kb）来说都比较短，因此在当时新一代测序方法也被称为“短序列测序方法（short reads sequencing）”。例如，在 2009 年前后，454 测序读长是 400～500 bp，在同时期的新一代测序平台中是最长的。与之相比，Illumina GA 和 SOLiD 当时的读长是 35～50 bp，它们读长的主要受限因素是信噪比（signal-to-noise ratio）低。在这之后的数年间，Illumina 推出了 HiSeq 系列和 MiSeq 测序平台，逐渐把测序读长加长到双端 150 bp（HiSeq）和双端 300 bp（MiSeq），拉近了与第一代测序法读长的距离。最近 1～2 年，随着 PacBio RS 和 Oxford Nanopore 单分子测序平台的推出，新一代测序的读长也迅速提高，达到了几 kb 到数十 kb 的水平，远远超越了第一代测序法的读长。

目前有关新一代测序技术的名称有多种说法，有人把 PacBio RS 和 Oxford Nanopore 的测序平台称为第三代测序技术，以区别于 Illumina、SOLiD 等第二代测序技术，其根据是它们实现了单分子实时测序，省去了第二代测序技术中的模板扩增步骤。但作者认为，这两种测序技术虽然与第二代测序技术有所区别，但仍然存在通量偏小，测序初始数据准确率不高的问题，在测序方法上和第二代测序技术相比并没有可以称得上“代差”水平的改进。所以在本书中我们没有把 PacBio 和 Nanopore 称为第三代测序技术，而是和原有的第二代测序技术一起统称为“新一代测序技术”。

下面按照测序技术出现的时间顺序，简要介绍 6 种主要的新一代测序平台的测序原理和技术特点。

一、Roche 454 焦磷酸测序技术

2005 年，454 公司推出了第一款二代测序仪 Genome Sequencer 20。2007 年，又推出了改进型测序仪 Genome Sequencer FLX 和小型化测序仪 GS Junior。该测序平台利用了焦磷酸测序原理和边合成边测序技术，得到的序列平均长度为 400～500 bp，通量为每次测序运行产出 0.7 Gb 左右数据。相比同期的 Illumina Genome Analyzer 和 SOLiD 测序平台，454 GS 平台得出的序列读长较长（同期的 Illumina 和 SOLiD 读长只有 35～50 bp），有利于基因组数据的拼接，尤其适用于小型基因组如细菌基因组的从头测序（*de novo* sequencing），因而在一段时间内得

到了广泛的应用（Mardis，2008；Rothberg and Leamon，2008）。但在新一代测序平台中，454 的测序通量相对较小，单碱基测序成本高，很快被 Illumina 等后续出现的测序平台超越，最终罗氏公司于 2014 年宣布 454 退出测序技术竞争，遗憾离场。虽然如此，454 平台仍然在新一代测序技术的发展史上留下了浓重的一笔，科学家利用 454 测序技术作出过很多出色的工作，获得过一些重大的发现，如美国贝勒医学院人类基因组中心的 Wheeler 等(2008)利用 454 测序技术完成了 DNA 双螺旋结构发现者之一詹姆斯·沃森（James Watson）的个人基因组测序等。所以读者仍然有必要对 454 测序技术的原理有所了解。

454 测序技术原理（以基因组 DNA 测序为例，图 1.2）(Margulies et al.，2005）如下。

测序文库制备：将符合测序要求的基因组 DNA 用物理剪切的方法（例如，Covaris、nebulizer 或 Bioruptor 等方法）打断为 400～800 bp 的片段，经一系列建库操作，在单链 DNA 的 3′端和 5′端加上 454 建库的接头，形成测序文库。

乳化 PCR 扩增（emulsion PCR，emPCR）：按照一定比例将单链测序文库与 454 测序特有的微球（bead）混合，该微球直径为 28 μm，表面带有和文库一端接头的 DNA 互补的寡核苷酸，文库与微球连接固定后，在理想的状态下，一个微球只与一条单链文库 DNA 结合。然后将带有文库 DNA 的微球置于油相与水相的混合系统中，其中水相部分带有 PCR 扩增所需的所有成分（DNA 酶、dNTP 和扩增引物等），经机械振荡形成乳化混合物（俗称“油包水”混合物），即微小的水滴散落在油相中。由于微球具有亲水性，它们存在于“油包水”混合物的“水相”中。在绝大多数情况下，一个水滴中只含有一个微球。这样每一个水滴就形成了一个独特的 PCR 扩增微反应器，每个文库片段在各自的微反应器中进行 PCR 扩增（50 个循环），最终产生数百万个相同的拷贝。接着利用一系列特定的方法打破“油包水”混合物，选择出带有扩增文库的微球用于测序。

测序：454 系统使用 PTP 板（pico titer plate）作为测序的承载体，PTP 板是经蚀刻技术处理的玻璃板，其中一面约有 4×10^6 个直径为 45 μm 的小孔（well）。将带有扩增文库的微球放入 PTP 测序板上，经离心处理后使微球进入小孔中，每个小孔的直径大小（45 μm）使其只能容纳一个微球（直径 28 μm）。同时，测序反应和荧光发生所需要的各种酶和发光反应底物等也以更加微小的微珠承载，被置入小孔中，围绕在测序微球周围。这样，每个小孔就成为一个微型测序单元。454 测序系统按照 T、A、C 和 G 的固定顺序依次将测序所需的 dNTP 加入到 PTP 板上，每次只加入一种碱基，在 DNA 合成酶的催化作用下发生 DNA 合成反应。如果在某个测序单元内发生与测序模板碱基配对的合成反应，该反应就会释放一个焦磷酸。焦磷酸在 ATP 磷酸化酶的作用下与相应底物合成 ATP，而 ATP 在荧光

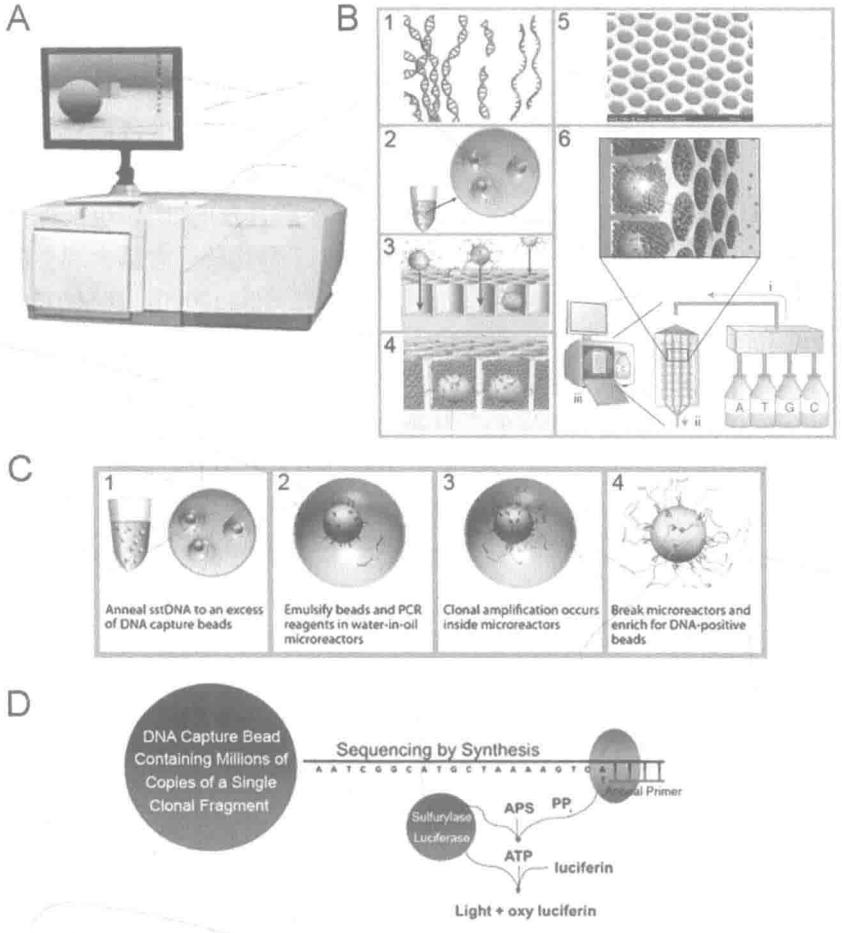


图 1.2　454 GS FLX 测序仪及测序原理图（选自 Rothberg and Leamon，2008 和 Mardis，2008，略有改动）

A. 454 GS FLX 测序仪。B. 454 测序流程图：基因组 DNA 片段化后两端加测序接头（1）；结合一条 DNA 文库单链的微球乳化 PCR 扩增（2）；携带相同 DNA 拷贝的微球进入测序 PTP 板表面微孔（3）；更小的测序反应试剂微球进入微孔（4）；测序载体表面微孔图像（5）；开始进行测序反应和荧光信号采集（6）。C. 454 乳化 PCR 扩增原理图：变性后的 DNA 文库单链与过量微球混合，每个微球结合一条 DNA 单链（1）；微球乳化形成“油包水”反应器结构（2）；反应器内进行 DNA 扩增（3）；打破油包水结构，富集携带扩增产物的微球（4）。D. 454 边合成边测序原理图：微球表面 DNA 克隆变性后退火与测序引物互补结合，当碱基发生配对合成反应时，释放一个焦磷酸，焦磷酸在 ATP 磷酸化酶的作用下与相应底物合成 ATP，然后 ATP 在荧光素酶的作用下释放能量，氧化荧光素放出荧光

素酶的作用下释放能量，氧化荧光素放出荧光。荧光信号被 454 系统配置的高灵敏度 CCD 照相机捕获，这样就得到了该反应循环中被合成的核苷酸信息。如此多次循环之后，测序系统就获得了待测 DNA 模板的序列信息（Rothberg and Leamon，2008）。

二、Illumina 测序技术

Illumina测序平台是继Roche 454测序平台之后第二个出现在高通量测序市场上的测序平台，也是目前应用最为广泛的新一代基因组测序平台，它使用边合成边测序技术实现了大规模平行测序。到目前为止，绝大多数的高通量测序数据是由 Illumina 测序平台产生的。它最早是由 Solexa 公司开发的，所以也被称为 Solexa 测序技术（Metzker，2010）。Illumina 测序平台有多种选择，有超大通量的、适合测序中心和测序公司使用的 HiSeq 系列测序仪，如 HiSeq2000、HiSeq1500/2500、HiSeq3000/4000 和 HiSeq X-Ten 测序仪，也有适合中小型实验室测序和医院医疗诊断测序使用的 MiSeq 和 NextSeq500 测序仪，最近又推出了针对医疗诊断测序实验室的 MiniSeq 测序仪（图 1.2）。从测序读长来看，HiSeq2500 的高通量模式可以读出双端 125 bp，快速模式可以读出双端 250 bp，HiSeq3000/4000 读长可达双端 150 bp，最长的读长是由 MiSeq 产生的，为双端 300 bp。

Illumina 数据的准确率很高，一般在 99.5%以上；其最突出的特点是测序通量高，从而极大地降低了测序成本。Illumina HiSeq X-Ten 是目前唯一可以做到仅花费 1000 美元即可对一个人的全基因组进行重测序的新一代测序仪。当然，如果加上生物信息分析的费用，整体的费用还是远远高于 1000 美元。

Illumina 测序技术是本书重点介绍的内容，本章将只对其测序原理和流程作如下简述。有关具体的建库、测序操作及数据的分析处理过程，请参阅本书第二章至第四章的详细介绍。

Illumina 测序原理（以基因组 DNA 测序为例，如图 1.3、图 1.4 所示）如下。

测序的第一步是建立测序文库（sequencing library preparation），简称建库。以基因组测序为例，利用物理方法或酶切方法，将被测序物种的基因组 DNA 打断成一定长度的片段（200～800 bp），经过片段选择、末端补平及加 A 尾后，用连接酶在 DNA 片段的两端加上 Illumina 测序专用的接头（adapter），连接的产物经过扩增、片段选择和纯化，就形成了可以用来上机测序的文库。在这里必须强调的是，**建库步骤是测序成功与否的关键**！Illumina 测序的建库技术将在本书第二章中作为重点内容加以详细介绍。

测序的第二步是文库的扩增成簇过程（cluster generation），成簇是在 Illumina 特定的仪器 cBot 上实现的（MiSeq 与 HiSeq 的快速模式不需要 cBot，成簇过程与测序过程都在测序仪上完成）。测序文库在有 8 个泳道（lane）的芯片（flow cell）上，与固化在泳道玻片壁上的寡核苷酸特异性互补结合，经桥式扩增（bridge amplification）把带有待测 DNA 片段的文库扩增到 1000 个拷贝左右，每个拷贝都具有相同的 DNA 序列，这样就形成了簇（cluster）。测序文库的成簇过程实际上

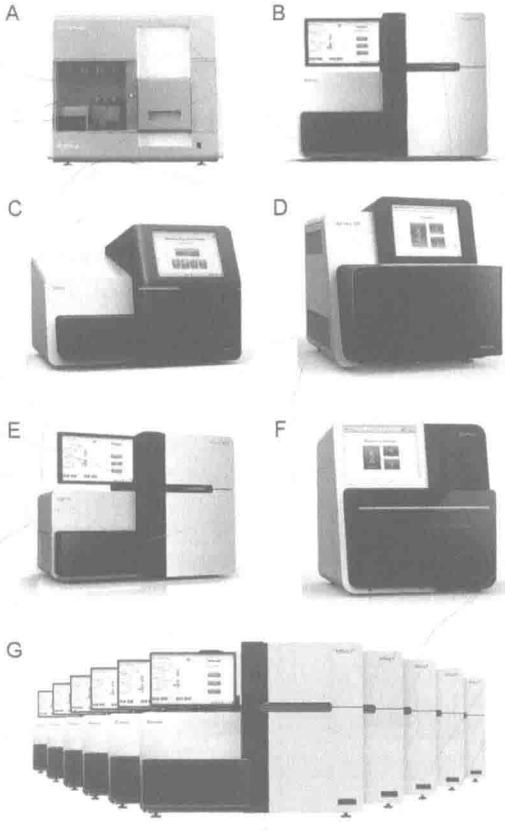


图 1.3　Illumina 测序仪图片（由 Illumina 公司提供）

A. Illumina GAII 测序仪；B. HiSeq2500 测序仪；C. MiSeq 测序仪；D. NextSeq500 测序仪；E. HiSeq4000 测序仪；F. MiniSeq 测序仪；G. HiSeq X-Ten 测序仪

可以看作一个测序荧光信号的放大过程，它可以使测序仪的光学成像系统清楚地捕捉并记录每一步合成测序的荧光激发信号，从而得到高质量的序列数据。

成簇过程完成后，就可以从 cBot 仪器上取出测序芯片，放置到 Illumina 测序仪上进行测序（MiSeq 与 HiSeq 的快速模式不需要挪动测序芯片，可以在仪器上直接测序）。Illumina 测序平台使用 SBS 技术和 3′端可逆屏蔽终结子技术（3′-blocked reversible terminator）进行测序（Bentley et al.，2008）。简单地说，就是 4 种带有不同荧光标记的特殊核苷酸（A、C、G 和 T），与 DNA 合成酶同时加到测序芯片的各个泳道中。在 DNA 合成酶的催化作用下，从测序引物结合部位开始合成与测序模板互补的新 DNA 链。同时，用于测序反应的特殊核苷酸在 3′端的羟基位置被化学基团屏蔽，导致每次 DNA 链合成都只能加入一个核苷酸。一次合