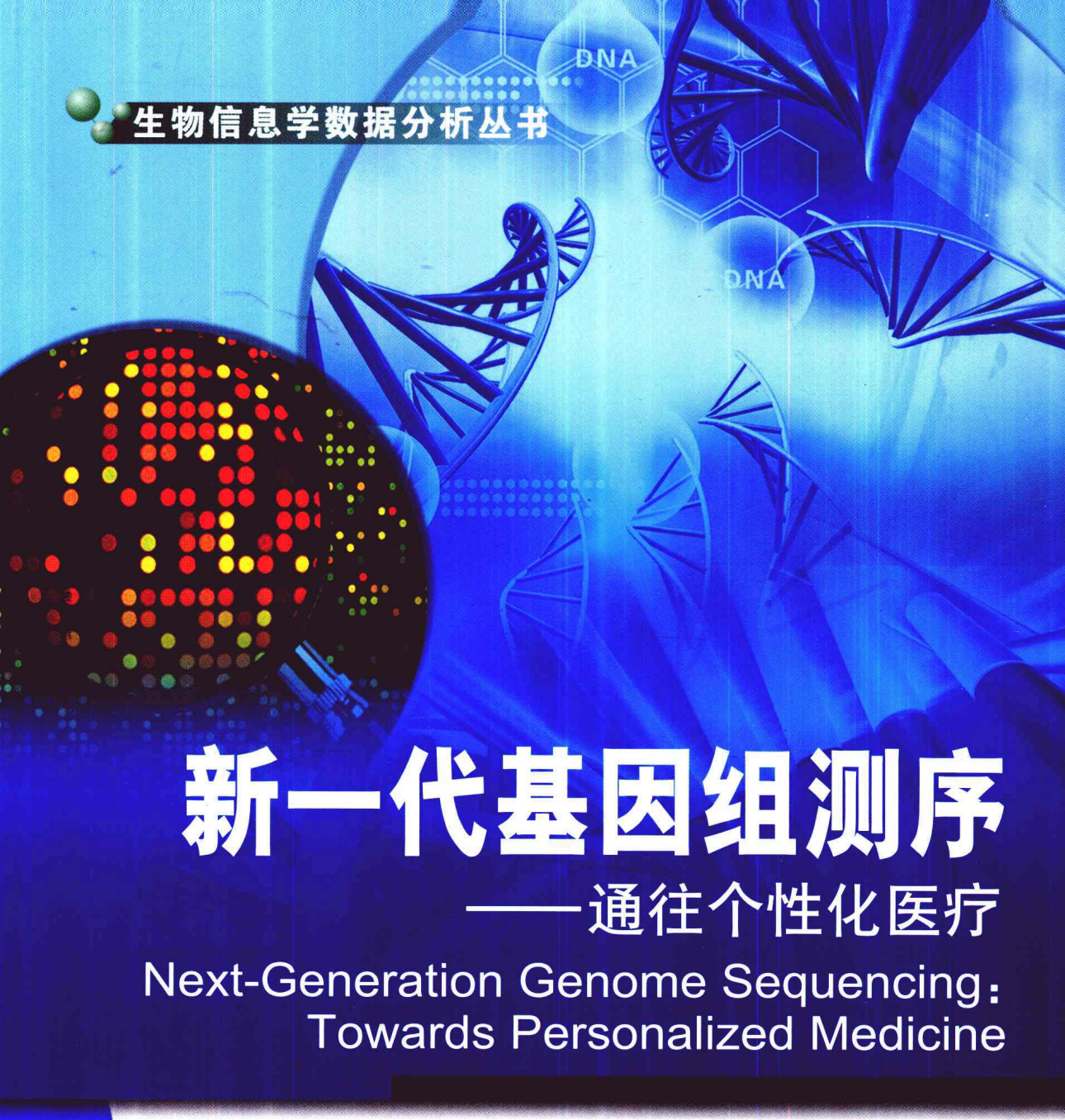


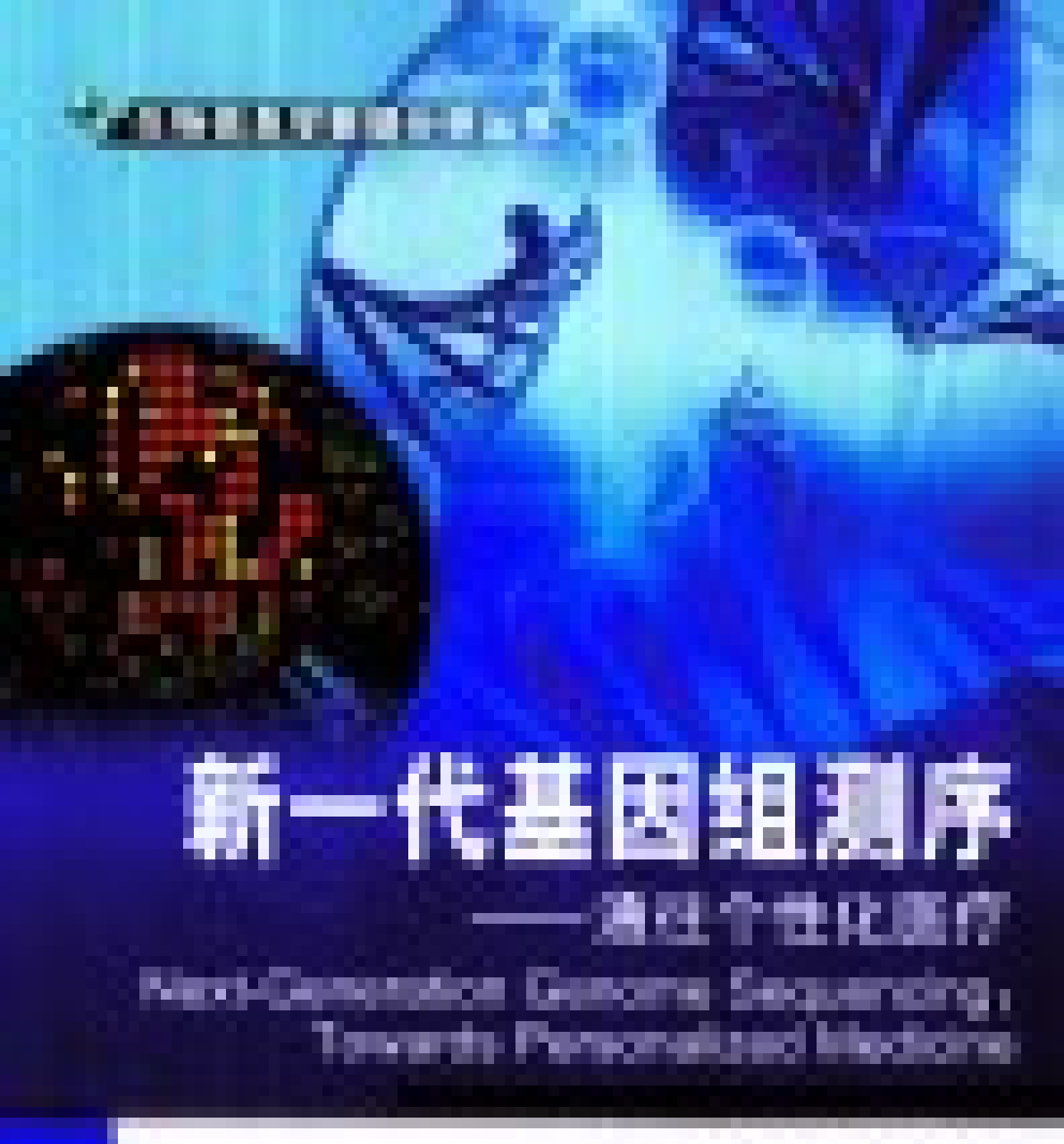
生物信息学数据分析丛书

新一代基因组测序
——通往个性化医疗

Next-Generation Genome Sequencing: Towards Personalized Medicine

〔德〕M. 贾尼特　编著
薛庆中　等　译

科学出版社



新一代基因组测序

——通往个性化医疗

Next-Generation Genome Sequencing: Towards Personalized Medicine

Next-Generation Genome Sequencing:

Towards Personalized Medicine

新一代基因组测序

——通往个性化医疗

〔德〕M. 贾尼特 编著

薛庆中 等 译

科学出版社

北京

图字：01-2011-7446 号

内 容 简 介

与传统测序技术相比，新一代测序（NGS）技术具有超高速、高通量、低成本和高效益的强大优势；这种新兴技术的研发和新一代测序平台的建立对基因组研究、人类健康和社会认知都产生了重大影响，是当今前沿科学发展最为迅猛的领域。全书包括 5 个部分 18 章。第一部分和第二部分分别概述了传统的 Sanger DNA 测序和新一代测序平台的工作原理、方法和特点；第三部分剖析了困扰测序技术瓶颈及其解决方案；第四部分介绍了测序的商业化应用和新兴的测序平台；第五部分探讨了新一代测序技术在基因组学研究中的广泛应用。

本书由参与 NGS 技术开发和应用的研究人员和发明家撰写，是全球第一本介绍新一代 DNA 测序技术的书。可作为高等院校生物学、医学、农学等领域的师生和研究人员学习参考用书，也对希望了解个人基因组信息、个性化医疗、伦理学等问题的读者有启迪和帮助作用。

图书在版编目(CIP)数据

新一代基因组测序：通往个性化医疗 /（德）M. 贾尼特（Janitz, M）编著；薛庆中等译. —北京：科学出版社，2012

（生物信息学数据分析丛书）

书名原文：Next-Generation Genome Sequencing: Towards Personalized Medicine

ISBN 978-7-03-033007-9

Ⅰ. ①新… Ⅱ. ①贾… ②薛… Ⅲ. ①人类基因组计划 Ⅳ. ①Q78

中国版本图书馆 CIP 数据核字（2011）第 258721 号

责任编辑：李 悦 岳漫宇 贺窑青 / 责任校对：张怡君

责任印制：钱玉芬 / 封面设计：耕者设计工作室

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

骏走印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2012 年 1 月第 一 版 开本：B5（720×1000）

2012 年 1 月第一次印刷 印张：15 1/4 插页：6

印数：1—3 500 字数：312 000

定价：58.00 元

（如有印装质量问题，我社负责调换）

译 者 名 单

（按姓氏拼音顺序排列）

丁文超　韩　序　胡杰峰　胡一彦　金呈朦　李　禹　卢　静

苏锟楷　唐　倩　王庭璋　薛庆中　周元飞　周忠静　朱惠琴

译 者 序

人类基因组等模式生物全基因组草图的完成为揭示生命的奥秘奠定了基础，也为研究生物基因组学和蛋白质组学敲开了大门，这一重大科学成就主要应该归咎于 Sanger DNA 测序方法的创立。然而，基于全基因组序列信息得以广泛应用，尤其是探究人类多数疾病和相关基因之间的特定联系，面向个性化医疗的未来，则依赖于近年来新一代测序（next-generation sequencing）技术的研发和新一代测序平台的建立。2010 年 1 月我怀着求知的愿望，参加了在美国加州圣地亚哥举办的植物、动物基因组学研讨会，聆听了著名科学家们的精彩报告，浏览了来自全球科学工作者展示的墙报，观看了跨国公司研发、展示的测序设备和仪表，让我零距离感受到这种新兴技术独特的知识趣味，增强了学习新方法的动力。

会后，我即通过因特网搜寻发现，“新一代 DNA 测序技术”名称最早是在 *Nature Method* 期刊 2008 年 1 月被提出，用“Next-Generation and Sequencing”关键字在 NCBI 数据库检索可找到 385 篇论文，短短两年时间内，几乎每隔一天就有一篇论文问世，其中不少发表在 *Nature* 系列等世界一流期刊上，充分表明这个关键词已成为当今前沿科学上的热点。我也在新书目录中欣喜地发现“新一代基因组测序”（Next-Generation Genome Sequencing）的书名醒目地列在其中。该书作者 Michal Janitz 博士称：这是第一本介绍新一代 DNA 测序技术的书，（他的前言写于 2008 年 7 月）。不久，徐建红博士帮我从美国新泽西州立大学（Rutgers University）图书馆找到此书并借出。我抓紧时间粗读了全文，深深地被这本书的内容所吸引。这本书的主编 Michal Janitz 博士是德国马克斯普朗克分子遗传学研究所脊椎动物基因组学的课题组长（group leader），目前在德国教育与研究部生命科学资助发展计划担任顾问，他是德国国家基因组研究网络的成员。全书分 5 部分 18 章，涵盖了新一代测序在基因组研究中拓展的各个领域，每个章节都是由参与新一代基因组测序技术开发和应用的研究人员和发明家操刀编写，因而具有很高的影响力和广泛的实用性。每章篇幅不长，却写得较为简明。并附有很多彩色图和表，有助于读者对测序原理和过程的理解。为了和国内学者分享新技术的进展，我萌生了翻译此书的想法，并与科学出版社李悦编辑交流，很快得到了她的热心支持和帮助。

“next-generation sequencing（NGS）”本书译为新一代测序，在国内，也有人直译为下一代测序。我们偏爱称之新一代测序是想突显其与传统测序技术的不

同。正如本书作者所述，新一代测序最显著的特性是超高速和高通量，它每次测序能产生数百万个读序列，而每个读序列又只有 30 ~ 400bp 长。这将使用户面临着如何应对高通量数据的存储和处理两大难题。为此，计算机工作者们研发了很多新算法，来适应这些严峻的挑战。更令人惊喜的是，今后采用新一代测序对一个基因组重测序，其目标设为 1000 美元，即是传统测序方法价格的万分之一。充分显示了新一代测序技术的魅力，相信读者阅读此书后会有不少收获和启发。

阅读本书以及相关基因组测序文献时通常会遇到很多新名词，读者不妨查阅书中的中英对照。其中不少名词在国内还没有统一的译名。例如，reads 就有多种译法，我们将其译为读序列，是兼顾了直译和意译两方面，也许比“读、读取、读长、测序片段”等译法更易理解。在翻译中我们力求对同一个单词有统一的译名，但是，个别单词在日常应用中已有惯用的名称，就只能顺其自然，求同存异了，如标记、标签、标志物。

本书翻译工作是本人和浙江大学纳米院、农学系的几位研究生共同完成的，翻译过程中一些较难理解的句子，请教了中科院基因组所于军教授和胡松年教授，得到他们的认真指点。科学出版社李悦等编辑对样稿中表达欠妥的句子做了提醒和修改。本书的出版得到浙江大学纳米院领导的支持，此外，徐建红、陈爱华、刘秋香、周国艳、刘杰为此书的出版、校对做了很多具体工作。在此一并致谢。新一代测序技术反映了当今前沿科学的最新发展，然而，限于我们的知识水平，对于原文理解和翻译不当之处在所难免，望请读者给予批评指正。

薛庆中

2011 年 8 月于浙江大学

前　言

30 年前由 Sanger 及其同事开发了 DNA 快速测序的方法，从而启动了破译基因的进程。随着人类基因组的破译，测序技术已产生了实质性的改进以满足高通量数据快速增长的需求，毛细管电泳的自动化便是一个成功的实例。最近基因组测序已能在自动化程度高，众多人才聚集的大型测序中心完成。但是，DNA 测序规模即便达到了产业化，人类基因组草图的产生仍需耗资 1 千万美元，花费近 10 年的时间。由于测序价格高，基于群体的表型 - 基因型的连锁研究还只能小规模进行，因此很难取得稳健统计的结论。多数疾病和特定基因之间推测的联系还经不起科学验证。2004 年，大规模平行焦磷酸测序技术的商业化已首次实现，它为低成本高效益和快速破译基因组创造了良机。此后不久，进入市场的其他厂商为人类基因组测序所定的目标是只需 1000 美元。

本书的主题是希望给读者全面介绍新一代测序（NGS）技术，突显遗传学对基因组研究、人类健康和社会认知的影响。新一代测序虽然没有明确的定义，但它和基于传统的毛细管测序 NGS 平台的功能有以下区别：新一代测序的显著特性是每次能产生数百万个读序列（reads），而传统测序只有 96 个，完成细菌或果蝇全基因组测序过程分别只需几小时或几天而不是几个月。以毛细管测序为代表的传统测序方式是基于载体的克隆方法，这已经被通过直接粉碎，然后对扩增 DNA 测序的方法所替代。新一代测序的另一个显著特性是测序产物本身的读序列长度较短，介于 30 ~ 400bp。由于这些读序列长度有限，因而对其应用会产生重大影响，如从头测序。新一代测序的明显优势是能把高通量、模板制备容易、短读序列拼接软件、大规模测序数据存储和处理等富有挑战性的新功能加以整合。为满足科学界的广泛需求，本书解读了新一代测序技术及其在未来的基因组研究中的作用。每个章节都是由参与 NGS 技术开发和应用的研究人员和发明家编写的。

本书的第一部分对至今仍然是生命科学中黄金标准的 Sanger DNA 测序进行了精彩的概述。第二部分和第四部分分别介绍了测序的商业化应用和新兴的测序平台。第三部分突显了困扰当前测序技术的两个瓶颈，即数据存储和数据处理。一旦新一代测序技术商业化，它的应用将会取得前所未有的突破。第五部分与读者深入探讨了新一代测序在基因组研究中新拓展的应用，其中部分内容涉及对现有技术的改进，但更多的是显示新一代测序的特点：良好的稳定性、高效益，这

在古基因学研究中的应用尤为突出。

新一代测序技术在全基因组背景中用于基因研究所显示的通用性和稳定性，让许多科学家，包括我自己为之震惊。我们知道，引起大部分疾病的过程都不是由单个遗传性缺陷所产生的。相反，它们并不是数千个基因的单独作用，而是涉及数百个基因的相互作用。以往，遗传学者们都聚焦在那些具有独立、效应显著的主基因上，当这些基因出错时，其效应就容易发现。然而，将那些效应并不明显的微效基因组合起来可能同样是重要的。无论对单细胞或组织，还是对整个生物体而言，将新一代测序技术和系统生物学结合的方法是十分有效的，它可以阐明网络调控复杂的依赖关系。

我们希望这本书将使我们深刻理解基因组的变化，不仅对其研究本身，而且也对我们日常生活的许多方面，包括医疗保健政策、医疗诊断和治疗产生影响。最明显的例子是，心脏病或癌症的患者指望通过基因组学获知其疾病发展的风险，甚至希望知道喝多少咖啡是安全的。这些信息是依据个人的单核苷酸多态性（SNP）模式与带有特定疾病相关 SNP 单体型的相关性。最近公众讨论了有关个人基因组信息的可行性所带来的挑战，揭示了基因组信息及其应用的新认知。渴望了解基因组，并将其视为与己相关的大事，对于科学界以外的人们来说，这还是第一次接近遗传学信息，在获得好处的同时，也会不断遭遇其伦理和法律风险。书的最后部分还向读者介绍了一些日后将会加剧的争议。

最后，我想表达对本书所有作者的衷心感谢，他们为全面而清晰地探索基因组的迷人技术及其应用，做出了非凡的努力。我还要感谢 Hans Lehrach 教授的一贯支持。

Michal Janitz

2008 年 7 月柏林

（薛庆中译）

目　录

第三部分 瓶颈：序列数据分析

第四部分　新兴测序技术

第五部分　新一代测序：真实地整合基因组分析

第一部分　Sanger DNA 测序

1 Sanger DNA 测序

Artem E. Men, Peter Wilson, Kirby Siemering and Susan Forrest

1.1 Sanger 测序的基础

1977 年 Sanger 团队里程碑式地解密了首个噬菌体 ϕX174 的基因组（其 DNA 拼接长度仅 5000 多个碱基）。随后，美国基因组研究所（The Institute of Genome Research，TIGR）于 20 世纪 90 年代早期测出了几个百万碱基大小的细菌基因组序列，欧洲财团于 1996 年测出了第一个真核生物芽殖酵母（*Saccharomyces cerevisiae*）的基因组。一直到后来若干个哺乳动物包括人类自己的 10 亿个碱基大小的基因组几乎完全被测出。在过去 30 年里，Sanger 法无疑已成为一种长期而有效的 DNA 测序方法。现在，测序技术已经使现代生物学的研究现状发生根本变化，为系统地描述生物提供了精确工具。随着表型数据与 DNA 序列计算相结合能力的不断提高，测序领域已经得到了快速的发展，因此即便细微的 DNA 变化［如单核苷酸多态性（SNP）］也能明确地与生物学表型联系起来。这就使从单细胞生物到大多数复杂多细胞生物发展的由核苷酸驱动的基本生命过程的实用监测方法成为可能。

1977 年发表的经典 Sanger DNA 测序方法[1]第一个新颖之处是基于 DNA 合成时 4 个独立碱基特异性链式终止反应分别对应 4 种不同的核苷酸［图 1.1（a）］。在所有 4 种 2′-脱氧三磷酸核苷酸（dNTP）的每个反应液中加入一种特别的 2′，3′-双脱氧三磷酸核苷酸（ddNTP），如反应“A”中加入 ddATP，依次类推。在测序反应中使用 ddNTP 在当时来说是一种很新的方法，与该团队 1975 年开发的“加减”法相比，测序结果更好。每次新合成的 DNA 链延伸的终止都与 ddNTP 相对应。由于加入 ddNTP 的量很少，因此终止反应很少发生，同时由于终止反应的随机性，就产生了延伸产物的混合物，而 N 端碱基每个位点都会产生匹配产物，通过与 3′端 ddNTP 结合终止延长。

该方法的第二个新颖之处是在 DNA 新链合成中引入了放射性磷或硫同位素标记的前体（dNTP 或测序引物），因此使每个产物可以通过放射线显影得以检测。最后，由于每个延伸反应产生一个非常复杂而且含大量放射性的 DNA 混合物，因此，开发一种个别检测这些分子的方法可能就是成功的关键。该发明使用