科学通报 2017 年 第 62 卷 第 19 期: 2042 ~ 2046
《中国科学》杂志社
SCIENCE CHINA PRESS
Science125个科学前沿问题系列解读(XXX)

How much will new technologies lower the cost of sequencing?

新技术能使 DNA 测序的成本降低多少?

石铁流

华东师范大学生命科学学院, 上海市调控生物学重点实验室, 生物信息学及计算生物学中心, 上海 200241
E-mail: tlshi@bio.ecnu.edu.cn

2017-01-06 收稿, 2017-03-01 修回, 2017-03-02 接受, 2017-05-22 网络版发表

摘要 1977年, Sanger及其同事发明出最早的DNA测序技术, 1990年开始的人类基因组计划催生了DNA测序技术的自动化. 近10多年来, 新一代测序技术得到了飞速的发展, 测序速度及测序通量的巨大提升使得个人基因组的测序成本急剧下降, 达到了1000美元一个全基因组的水平, 从而使得DNA测序技术在生命科学研究领域以及临床医学上有着更广阔的应用前景. 近来出现的第三代测序技术具有测序过程中无需PCR扩增而且产生非常长的测序片段的优势. 尽管其测序的准确性只达到85%左右且成本高, 但仍然显示出了较好的前景. "精准医学"时代的来临将进一步促进DNA测序技术的革新, 并使个人基因组测序成本进一步下降, 有望使个人基因组测序成本进入百美元的时代. 随着海量的测序数据的积累, 如何有效地分析和解读这些测序数据面临着巨大的挑战, 生物信息学在此过程中扮演着关键的角色.

关键词 DNA 测序技术, 个人基因组, 测序成本, 新一代测序技术, 第三代测序技术

在2015年1月30日的国情咨文讲话中, 美国总统Obama提出了"精准医学"(precision medicine)的初期计划. 核心就是通过分析上百万名不同年龄层和不同身体状况的男女志愿者, 研究遗传变异对人体健康和疾病的影响, 以便更好地了解疾病形成的机理, 进而为开发相关的药物及实现"精准医学"铺平道路. 精准医学计划的提出得益于这些年来高通量测序技术的飞速发展, 大规模的高通量测序技术在人类基因组学和转录组学上的广泛应用极大地加深了人们对疾病机理的认识.

最早的DNA测序技术是由英国科学家Sanger及其同事[1]于1977年发明的, 他们巧妙地将DNA聚合酶功能应用于DNA序列检测. Sanger法的核心原理是: 由于ddNTP缺少3′-OH, 无法在两个核苷酸之间形成磷酸二酯键, 从而中断DNA链的合成. 因此Sanger及其同事分别在4个DNA合成体系中加入一定量的带有同位素标记的ddNTP (ddATP, ddTTP, ddCTP和ddTTP), 通过凝胶电泳和放射自显影技术, 从而测出DNA序列. 他们开创了一个全新的领域, 为此, Sanger因为在DNA测序技术的成就于1980年第二次获得了诺贝尔化学奖. 在随后的10多年中, 基于电泳的DNA测序方法逐步在生命科学领域得到推广、应用, 极大地推进、丰富了人们对基因组结构及基因结构的认识. 在20世纪80年代后期, Hood研究组[2]进一步发展了DNA测序技术, 他创造性地利用4种不同的荧光物质标记4个不同的脱氧核糖核酸, 并将计算机应用于测序数据的收集, 大大降低了人工进行数据收集的复杂性. 通过将放射性标记物替换为荧光物质, 所需的DNA合成体系从4个降为1个, 且增加了核苷酸的稳定性, 降低了实验人员的健康风险. 随后, 美国政府于1990年开始实施的人类基因组计划把DNA测序技术应用推到了新的高度, 在人类基因组计划实施的过程中, Applied Biosystems公司(美国)推出了第一款自动化的测序仪, 将Hood改

引用格式: 石铁流. 新技术能使 DNA 测序的成本降低多少? 科学通报, 2017, 62: 2042–2046
Shi T L. How much will new technologies lower the cost of DNA sequencing? (in Chinese). Chin Sci Bull, 2017, 62: 2042–2046, doi: 10.1360/N972016- 01107

进的测序技术自动化，从而极大地加快了DNA测序的速度. 之后这种第一代测序技术得到了进一步改进，并行化技术的应用使得测序数据成百倍地提高，为人类基因组测序计划的加速完成打下了坚实的基础. 最终在2001年2月，人类基因组草图得以绘制完成并发布[3,4]，成为了人类生命科学研究史上的一个里程碑.

人类基因组计划一共花费了大约30亿美元，完成了第一个人类参考基因组. 之后的几年间，在花费了一亿美元后，民间人类基因组测序公司Celera Genomics (美国)的总裁，科学奇人J. Craig Venter带领的团队[5]于2007年发布了他的基因组测序. 但这种巨大的花费使得这种技术在人类基因组测序研究领域无法具有推广性，并促使人们寻求测序速度更快、通量更高、价格更便宜的测序技术. 21世纪初，新一代的测序技术(next generation sequencing)陆续地出现，454 Life Sciences公司(瑞士)于2004年推出了454FLX，Solexa公司(美国)于2006年推出了Solexa基因组分析仪，2007年Applied Biosystems公司推出了SOLiD^{TM}系统. 这些新的第二代测序系统跟第一代测序系统相比尽管测序的基本原理上没有太多的改进，但由于第二代测序系统同时对定量序列片段进行大规模并行测序，从而在测序通量上得到了显著的提升，且极大地降低了测序成本. 第二代测序系统与第一代测序系统的成本结构发生了改变. 第二代测序系统推出更有效的和复杂的设备，从而增加固定成本，但降低测序样品成本，从而降低可变成本. 综合来看，总的平均成本是降低的. 454技术当时测序的读段长度(read length)可以达到250 bp[6]，远远超过最初的Solexa的36 bp测序读段长度[7]和SOLiD的36 bp的测序读段长度[8]. 2008年4月，利用454测序仪完成的诺贝尔获得者James Watson的个人基因组的测序总共包含有32亿个碱基，测序深度为8×左右，一共花费了4个半月的时间，少于150万美元的成本[9]. 这与第一代测序技术所花费的时间及成本相比是一个巨大的进步(表1).

近年来，第三代测序技术应运而生. PacBio公司的单分子实时测序技术(single-molecules & real-time sequencing, SMRT)的原理是单分子的DNA通过纳米孔而产生的电流变化或光信号的变化通过仪器捕获，从而可以解析DNA的序列组成. 具体过程是DNA聚合酶将互补的并有荧光标记的核苷酸配对到DNA链上，每当有一个核苷酸被添加，这个核苷酸的荧光颜色就会被记录下来，之后荧光标记就会被消除，然后下一个核苷酸就会被添加和记录. Oxford Nanopore Technologies公司(英国)所开发的纳米单分子测序技术并不是通过合成DNA序列进行测序的，它是通过降解DNA序列进行测序的. 在该系统中，核酸外切酶和一种特殊的纳米孔接合在一起. 当一个核苷酸被核酸外切酶剪切下来后，在通过纳米孔时，不同的核苷酸会产生不同的电流扰动，从而完成测序. 目前还有一种技术是Ion Torrent使用了一种高密度的半导体芯片，当核苷酸在DNA聚合酶的作用下结合到DNA链上时，作为核苷酸延伸产物的氢离子会被释放出来，并被记录下来从而获得碱基的序列信息[10]. 第三代测序技术最大的优势在于样本制备的速度快，

表 1　测序技术不同发展时期的测序成本比较

Table 1　The comparison of sequencing costs among different stages

基因组发表时间	人类基因组计划(2003)	Venter (2007)	Watson (2008)	个人基因组(2016)
花费的时间(开始到结束)	13年	4年	4.5个月	4天
测序技术	跑胶及一代测序	一代测序技术	二代测序技术	二代测序技术
参与研究的科学家	>2800	31	27	3~5
测序成本(开始到结束)	~30亿美元	1亿美元	150万美元	~1000美元
覆盖度	8~10×	7.5×	7.4×	~30×
参与的研究所	16	5	2	1
参与的国家	6	3	1	1
文献来源	[3,4]	[5]	[6]	

测序过程中利用的是天然、未被修饰的DNA分子，并能产生非常长的测序片段，测序过程无需进行PCR扩增，从而能够避免PCR扩增而产生的偏好性. 测序的读段长度越长，所需测序的深度就可以越浅，就越容易拼接出完整的全片段DNA序列. 如第三代的PacBio系统，其测序段长可以达到10 kb以上. 2015年1月发布的新的人类基因组参考序列就是利用这种新的技术填补了160个大的间隔中的55%，同时，解读了26079常染色体结构变异的完整序列，包括倒位、复杂插入和长大片段的串联重复序列[11]. 但目前这种测序技术的准确性只能达到85%左右，而且应用于个人基因组测序的价格还比较贵.

近些年来，随着测序技术的不断改进，测序通量的进一步提升，使得人类全基因组测序成本更显著地下降. 目前，Illumina公司(美国)制造的测序仪占据了全世界测序市场70%以上的份额[12]，2014年初该公司推出的HiSeq X-Ten测序系统包括10台高通量测序仪，以人均1000美元左右的成本每年可完成18000人的全基因组测序. 但要注意的是测序的成本与测序的深度紧密相关，1000美元的花费目前可以产生90 G的数据，对人类基因组来说大约是30×的覆盖度(表1). 2017年1月10日，Illumina公司推出了新的NovaSeq测序仪，使得100美元测序的目标变得更近，进而每个人都可能在精准医疗中获益.

第二代测序技术在通量及价格上有着巨大的优势，但其产生的读段长度与第一代测序技术相比较要短得多(35~200 bp vs. 500~1000 bp)，而且测序的准确性要远低于第一代测序技术，第一代测序技术测序的错误率可以低至10^{-5}，而第二代测序技术的错误率会升至10^{-3}~10^{-2}[7]. 因此，第二代测序技术检测出来的重要的基因突变位点和单核苷酸多态性(single nucleotide polymorphism, SNP)位点需要利用第一代测序技术进一步确认，特别是在临床应用中与致病性和用药相关的基因突变位点和SNP位点更需如此. 同时，因第二代测序技术产生的读段长度短，这给基因组的拼接带来了新的挑战. 一些新的算法被开发出来以满足短读长的拼接需求. 此外，提高新一代测序技术的准确性也将是今后相当长时间内需攻克的难题.

现在精准医学等临床应用的巨大市场正在进一步促进测序技术的革新和发展. 可以预见，新一代的测序技术将不断发展进步，测序技术的创新，测序通量显著的提升将会使测序成本进一步下降，不远的将来，100美元成本的个人基因组测序的梦想将会实现(图1). 新一代的测序技术将会更广泛地应用在生命科学的基础研究及临床应用领域，由此将产生海

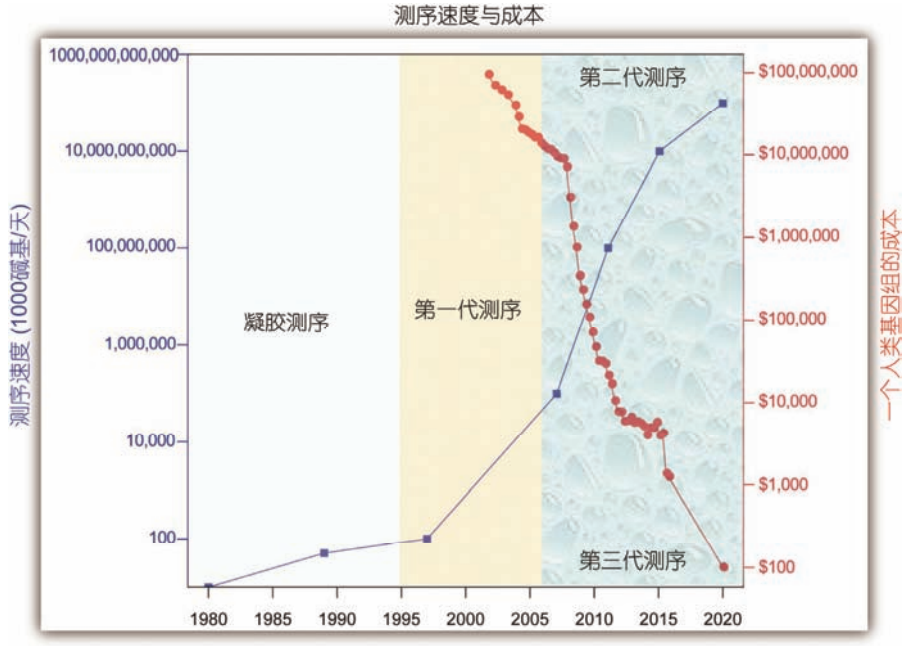


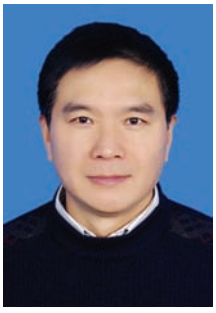
图 1 测序通量的提升与测序成本下降的趋势. 左边*Y*轴为测序通量，右边*Y*轴为测序成本

Figure 1 The trend of the increasing sequencing throughput and decreasing sequencing cost

量的测序数据. 因此, 今后测序领域的瓶颈不是测序数据的获取, 而是测序数据的有效解读. 现有测序技术产生的数据已远超过人们解读数据的能力. 因此, 测序成本大幅度降低并不意味着测序技术应用的成本同样降低, 因为这些成本中都没有考虑数据处理和生物信息学分析的成本. 如Mardis[13]所言: “The $1000 genome, the $100,000 analysis”, 意即花费1000美元进行基因组测序, 需要投入10万美元的成本才能得到应有的分析结果. 为了全方位地解读全基因组的测序数据, 生物信息学扮演着越来越重要的角色. 目前海量的测序数据急需大量的掌握了数学、统计学和计算机技能而又具有生物学知识的生物信息学家来分析、解读, 生物信息学人才的缺乏将是测序技术应用领域, 如精准医学, 发展的主要瓶颈之一.

参考文献

1 Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA, 1977, 74: 5463–5467
2 Smith L M, Sanders J Z, Kaiser R J, et al. Fluorescence detection in automated DNA sequence analysis. Nature, 1986, 321: 674–679
3 The International Human Genome Mapping Consortium. Initial sequencing and analysis of the human genome. Nature, 2001, 409: 934–941
4 Venter J C, Adams M D, Myers E W, et al. The sequence of the human genome. Science, 2001, 291: 1304–1351
5 Levy S, Sutton G, Ng P C, et al. The diploid genome sequence of an individual human. PLoS Biol, 2007, 5: e254–e286
6 Meyer M, Stenzel U, Hofreiter M. Parallel tagged sequencing on the 454 platform. Nat Protoc, 2008, 3: 267–278
7 Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol, 2008, 26: 1135–1145
8 Mardis E R. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet, 2008, 9: 387–402
9 Wheeler D A, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature, 2008, 452: 872–876
10 Munroe D J, Harris T J. Third-generation sequencing fireworks at MarcoIsland. Nat Biotechnol, 2010, 28: 426–428
11 Chaisson M J, Huddleston J, Dennis M Y, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature, 2015, 517: 608–611
12 Zimmerman E, Vogelstein F, Regalado A, et al. The 50 smartest companies. Technol Rev, 2014, 117: 26
13 Mardis E R. The $1000 genome, the $100000 analysis? Genome Med, 2010, 2: 84



石铁流

博士, 教授、博士生导师. 1999年在美国Louisville大学获得计算机硕士学位, 2000年在Louisville大学获得分子生物学博士学位. 2002年加入中国科学院上海生命科学研究院生物信息学中心从事生物信息学的研究工作. 2008年底加入华东师范大学生命医学研究所. 其主要的研究兴趣在以下4个方面: (ⅰ) 疾病临床数据标准化及挖掘和疾病分子检测标记的发现; (ⅱ) 生命大数据整合分析, 包括各种组学数据; (ⅲ) 生物信息学分析方法及高通量分析平台开发; 基因/蛋白质功能和蛋白质相互作用研究; (ⅳ) 药物(包括中药)靶点及作用机理研究.

Summary for “新技术能使 DNA 测序的成本降低多少?”

How much will new technologies lower the cost of DNA sequencing?

SHI TieLiu

Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai 200241, China
E-mail: tlshi@bio.ecnu.edu.cn

Sanger and his colleagues invented the DNA sequencing technology by cleverly applying the function of DNA polymerase to DNA sequence detection in 1977. Human Genome Project (HGP) launched in the early 1990s greatly facilitated the progress of the DNA sequencing technology and prompted the automation of DNA sequencing which conversely accelerate the human genome draft accomplished in 2001. For the past decade, great advancements in the speed and sequencing throughput in the next generation sequencing (NGS) technologies have dramatically reduced the cost of the whole genome sequencing to $1000.

Although there is no great improvement in technique for NGS compared with the first generation sequencing in the basic principle of sequencing, the second generation sequencing system processes the quantified sequencing fragments in massive parallel way, which makes sequencing throughput increased significantly and at the same time greatly reduce the cost.

Recently, the third generation sequencing technologies emerge with the advantages of very long sequencing fragment and no need for PCR amplification during the sequencing process, which significantly decrease the complex of assembling process and avoid the potential bias and false variations caused by PCR technology. Although the accuracy of the third generation sequencing technologies can only reach about 85% and the price is relatively high, the technologies still show their promising future with the gradual improvement in experiment processes and computational methods.

When biomedical researches and clinical applications will benefit from the technology innovation, the enormous applications of NGS in precision medicine and biomedical research fields will further accelerate the progress and innovations in NGS. At the same time, with the vast accumulation of massive sequencing data, how to efficiently analyze and interpret those data becomes a big challenge, bioinformatics plays the key role in the process. Well trained bioinformaticians with good background of mathematics, statistics, computer skills and biology are in high demand. It can be anticipated that the continuous innovations in NGS technology will further significantly decrease the cost for the human genome sequencing and promise the advent of the era with the sequencing cost close to $100, which will push personal genome sequencing widely used in clinical applications and make the precision medicine come true in daily life.

DNA sequencing technology, personal genome, sequencing cost, next generation sequencing, third generation sequencing

doi: 10.1360/N972016-01107