

III 下一代测序技术将改变生物学现状

新一代非基于“桑格技术”的基因测序法以空前的快速测序速度问世了，它为人类带来了重大的科学成果和先进的生物学应用软件。然而，要研发出新一代基因测序法，必须克服30年来以Sanger技术为基础的惯性思维。



应用桑格技术的哺乳动物基因组测序中心拥有工厂式全套设备以及大量工作人员

图片来源: *Maria Nemenuk, Broad Institute*

1977年, Fred Sanger 和 Alan R. Coulson 发表了两篇关于快速测序技术的论文^[1, 2], 该技术能够

破译完整基因片断和整个基因组, 大大促进了生物学的发展。该技术是以 Sanger 和 Coulson 发现的加減法原理 (1975年) 为基础, 大大减少了化学毒性物质以及放射性同位素的处理, 取代同年早期 Maxam 和 Gilbert^[3] 发明的化学降解法成为近30年来唯一的 DNA 测序法^[4]。

完成人类基因组计划的终极目标需要大量基因序列, 这种需求推动了测序技术的发展, 例如毛细管电泳仪的出现。实验自动化和程序并行化导致了一些拥有大规模测序设备与专家小组的测序中心出现。尽管目前已完成了两个不同的人类基因组计划, 但是生物学家渴望知道更多的基因序列, 更重要的是, 他们希望有更经济的测序技术。

2005年, 介绍合成测序 (sequencing-by-synthesis) 方法 (454 生命科学公司研发^[5]) 以及自动测序软件 (George Church 实验室研发^[6]) 的划时代性论文发表是测序市场创新的第一个标志。这两种测序技术应用铬铁尖晶石板或者琼脂糖薄板同时分析序列, 与同期桑格毛细管测序仪 (极限为96个样本) 相比, 具有以更小的反应体积产生更多的序列数据的优势。

454 公司首推的测序仪能轻易的产生与50台以上的全自动化3730XL毛细管测序仪 (Applied Biosystem's 3730XL capillary sequencers) 相当的序列数据, 成本花费却只是后者的1/6左右。尽管如此, 科学界对这项新技术却并不热衷。许多习惯用桑格技术的科学家怀疑新技术的准确度、阅读能力、成本消费、实用性。代理桑格型测序硬件的经销商害怕其投资失败而首先提出了这些怀疑。



最新的新一代测序仪操作
只需一个人，24小时内可
产生与几百台桑格毛细管
测序仪同样多序列数据

批评家为新测序技术的推广设置了种种障碍，然而大多数人却忽略了一个事实，即桑格技术的普及最初也遇到同样的阻碍。桑格技术刚开发出来时，阅读能力很难超过25bp，即使在Fred Sanger双脱氧终止法发明后也只达到80bp，如今却达到了750bp；而新发展的合成测序技术，应用焦磷酸测序方法，其阅读能力最初只有100bp，推向市场16个月后增加至250bp，随着技术的不断完善，目前已达到了400bp，很快就接近桑格技术目前的水平。

除了阅读能力外，能否以有限的成本用一台仪器产生足够数量的序列标记也是另一个需要改善的重要问题。这个问题已经被454公司解决了，应用他们的系统，仅花费阅读35bp或者更小片段的成本就能产生比35bp多10倍的序列标记。当今，抢滩新一代测序市场的仪器主要有三种：罗氏公司的454 GS FLX测序仪，Illumina公司的Solexa 1G测序仪和最新开发的全自动生化SOLiD系统，以及VisiGen公司和Helicos公司预期最近两年内推出的第三代（也叫新二代）单分子测序技术。

尽管Margulies等人^[5]应用原理论证了应用一台或者两台仪器能分析中小型细菌基因组，但大

多数人不相信应用焦磷酸测序方法可以弥补桑格技术的不足。最初，罗氏454 GS20测序仪被用来对细菌基因组再测序或者为正在进行的庞大的“桑格工程”补充数据。这项工程初步阶段需要做环境基因组学研究，不仅需要分析比人类基因组大得多的序列数据，而且要解决文库构建和克隆产生的取舍偏差等问题。这个时候，454技术利用乳滴PCR结合焦磷酸测序的优势迅速的打开了市场。乳滴PCR将单个DNA片段分子的扩增片段分配到一个乳滴中，从而提高了产物的纯度。另一方面，焦磷酸测序利用计算机高通量检测反应中产生的光信号。2006年，早期研究公开证实了新一代测序仪的多功能性能够为深矿的微生物^[7]、深海中的稀有生物圈^[8]或者海洋中的有毒微生物进行基因测序^[9]。

在2005年末，一个进行环境基因组学与古生物DNA研究^[10]的科研小组仅用一台454 GS20测序仪就分析出了28,000年前长达13Mb的长毛象基因组^[10]，从而为完成更有挑战性的计划——解密穴居人基因组铺垫了道路^[11,12]。解密穴居人基因组比解密古象基因组更难，因为从可用的样本中提取的穴居人DNA数量不到从现代人样本中提取量的5%。因此，进行穴居人基因组计划需要比进行

现代人基因组多20多倍的序列测定工作量。

此外，样品中DNA损伤与常温贮藏的状态以及新一代测序误差等因素往往超过了序列变异，因此很难确定样品来源于现代人类还是古穴人。相比较而言，断言长毛象的某一特定顺序是来源于古代标本，而不是现代污染物更容易，因为现代大象不像人类会经常处在实验室环境中。要在基因组范围内得到古代哺乳动物真正的序列，需要对某一区域进行反复的分析或者结合多种方法来确定其来源，这又要求大量削减成本来额外开展该项目。如果能完成这一项目而且能突破为复杂的多来源的DNA混合物测序的难题，将使测序地球上的任何生态系统成为可能，它也为测序至少十万年前的古代动植物打开了一扇窗户，这些都远远超出了我们以前的预想。

新一代测序仪已被用来再测序先前公布的参考细胞株，同时它也首次被允许查找基因组水平上的所有突变。在2005年，Velicer^[14]等人首次从一种进化了1000代的结核分枝杆菌株^[13]的基因组（9Mb）上发现了所有的突变体，同时也研究确定了该菌株上的耐药性等位基因。从这些早期的成果中，人们清楚地看到，新技术虽然能发现新的基因突变^[14]，但是它也必须解决许多测序中出现的错误，如焦磷酸测序时出现的同聚物干扰和阅读短片段时3'端序列迅速降解等问题。

初步的解决方法是联合运用桑格和焦磷酸测序数据^[15]。不管进行何种项目，只运用桑格技术，其人力和财力的消耗将是巨大的。许多实验室现在要么只依靠新一代测序数据，要么结合焦磷酸测序法阅读长片段与illumina的solexa低营运成本的优势，要么应用生物系统的SOLiD平台，独自来考察各种体系的性能。随着更多有效的非桑格测序方法的发明，现在已经能评估新一代测序的准确性和评估绝大多数公布的桑格数据的正确性。

大量生产密切相关的有机体的序列数据推动了再测序的应用。所谓再测序，就是以不同方式处理序列数据而不是重新组装基因组。再测序以一个参考序列做对照，对目标区域进行8-12次排序，排序次数比从头组装基因组（25-70次）少得多。

研究表明用这种方法能够测序10个哺乳动物线粒体基因组^[16]，从而使群体遗传学能够建立在完整的线粒体基因组而不是短序列片段研究的基础上。目前，许多微生物测序项目已在开展，这将不仅有利于扩大可用的基因组数据库，而且还使许多在基因组水平上比较基因型和表现型的研究成为可能。

甚至，研究目前尚未测序的生物体也能应用新一代测序法，直接从序列水平上破译细胞的转录子。在许多方面，用基因序列来表示转录子的特性优于用基因芯片表示。最重要的是，在排序前并不需要一定了解基因序列方面的知识，因为可以借助计算机比照数据库中最接近的参考序列，从而得到转录子序列。因此，转录子序列的获得将会给生命科学带来革命性的进展。例如，参考豆科植物*meticago truncatula*时代的基因组和植物标本*Arabidopsis thaliana*^[17]，科学家能够为*Zea mays*（玉米）^[18]的cDNA排序，而且发现了大量以前未描述的序列标签。

类似转录组学（transcriptomics）方法可以避免巨大基因组所造成的问题。尽

管桑格技术已经成功完成病毒，微生物和大型哺乳动物的测序计划，但还是留下一个问题——无法解析多倍体植物以及其后代的基因组。如小麦有16Gb的六倍体基因组，这些巨大的基因组，往往存在于农作物中，仅利用旧测序法是无法得到其序列的。然而，应用现在已证明的概念——新一代已表达序列标志测序法花费低得多，就至少能在功能水平上对植物基因组进行评估^[18]。

最后，新一代测序法的应用与医学领域密切相关。例如在癌症遗传学方面的应用，对于某些情况下用桑格技术不能检测的癌基因^[19]，现在应用超深测序法（ultra-deep sequencing）就可在组织中检测到。当桑格技术主要用来分析700bp以上的片段时，新一代测序法利用其阅读片段短的特点，在测序领域得到重用。由于癌症遗传学不遵守孟德尔遗传定律，激光捕获显微切割技术（laser-capture microdissection）用于收集相关等位基因而且必须使用PCR产物和/或扩增子序列定向，这样避免了传统的克隆及PCR误差。

虽然新一代测序仪已经应用于多项研究，但还有许多不足有待科学家与工程师们进一步改进。首先是降低成本：若要实现进行个人基因组研究的愿望，减少1-2个订单的规模是必要的，个人的基因组再测序的目标花费为1000美元。此外，降低测序错误率，这不仅是为了所有新一代测序技术的发展，而且还为了在不久的将来桑格测序技术能继续被采用。今后可能会采用特定的DNA聚合酶发射光波的形式直接读取DNA序列，但即使有了这些改进，我们还是不大可能看到DNA序列翻译成机器可读代码。价格下降，数据量有可能会飞涨，这就会造成解析瓶颈。因此，大部分由未来新型测序仪器提供的数据增量，将抵消生物信息学方面增加的人力和财力消耗。

在许多生命科学家认为后基因组学研

究已经到来的时候，在短短不到两年的时间里出版了1000多篇相关科研论文，新一代测序已显示出巨大的潜力。它也使得基因组学返回到由单个科学家或小型科研单位来研究的方式，事实证明，大多数的新一代测序法论文来源于小型研究个体而非基因组中心。在不久的将来回头看，我们肯定会惊讶为什么最初新测序技术在科学界乃至商业界不受欢迎。当第三代的测序仪器被推广的时候，我们可从中窥探到创新能打破桑格技术垄断测序市场30多年的格局。

原文检索：<http://www.nature.com/nmeth/journal/v5/n1/full/nmeth1156.html>

参考文献

- [1] Sanger, F. et al. Nature 24, 687–695 (1977) .
- [2] Sanger, F., Nicklen, S. & Coulson, A.R. Proc. Natl. Acad. Sci. USA 74, 5463–5467 (1977) .
- [3] Maxam, A.M. & Gilbert, W. Proc. Natl. Acad. Sci. USA 74, 560–564 (1977) .
- [4] Sanger, F. & Coulson, A.R. J. Mol. Biol. 94, 441–448 (1975) .
- [5] Margulies, M. et al. Nature 437, 376–380 (2005) .
- [6] Shendure, J. et al. Science 309, 1728–1732 (2005) .
- [7] Edwards, R.A. et al. BMC Genomics 7, 57 (2006) .
- [8] Sogin, M.L. et al. Proc. Natl. Acad. Sci. USA 103, 12115–12120 (2006) .
- [9] Angly, F.E. et al. PLoS Biol. 4, 2121–2131 (2006) .
- [10] Poinar, H.N. et al. Science 311, 392–394 (2006) .
- [11] Green, R.E. et al. Nature 444, 330–336 (2006) .
- [12] Noonan, J.P. et al. Science 314, 1113–1118 (2006) .
- [13] Andries, K. et al. Science 307, 223–227 (2005) .
- [14] Velicer, G.J. et al. Proc. Natl. Acad. Sci. USA 103, 8107–8112 (2006) .
- [15] Goldberg, S.M. et al. Proc. Natl. Acad. Sci. USA 103, 11240–11245 (2006) .
- [16] Gilbert, M.T.P. et al. Science 317, 1927–1930 (2007) .
- [17] Cheung, F. et al. BMC Genomics 7, 272 (2006) .
- [18] Ohtsu, K. et al. Plant J. 52, 391–404 (2007) .
- [19] Thomas, R.K. et al. Nat. Genet. 39, 347–351 (2007) .

 小会 编译



视频教学

www.LifeOmic.com