

如何只用万分之一的费用 获得全基因组序列

美国国家人类基因组研究所（NHGRI）的新一代测序技术正指引着测序技术平台的发展方向。随着测序技术的不断进步，在不久的将来，全基因组测序将被普及，成为常规的检测手段。

早在2001年，当人类基因组图谱即将绘制完成之际，位于美国马里兰州贝塞斯达的美国国立健康研究院（NIH）下属的国家人类基因组研究所（NHGRI），早就已经开始筹备下一阶段的基因组研究了^[1,2]。超过600位科学家和相关人员经过两次大会以及10次讨论不同主题的专题会议后，成立了一个特别委员会。该委员会确立了在未来的10到20年间，基因组学的研究方向及所面临的机遇和挑战。会议的成果是，与会的科学家制定了一项非常“大胆的”计划^[3]。该计划由一系列的高新技术组成。在当时看来，这些高新技术简直就像科幻小说一样虚幻、遥不可及。其中一项技术就是要大幅度降低全基因组测序的费用，至少要降至当时费用的万分之一或者十万分之一。也就是说，检测一个人的全基因组序列不能超过1000美元。第四届委员会资金的最大一笔资助项目就与大幅度降低全基因组测序费用的技术有关。现在，价格低廉的测序技术已带领人们迈进了全基因组测序的普及时代。



图片说明：2007年3月，在圣地亚哥举行的NHGRI高级DNA测序技术发展会议。NHGRI每年都会举办一次这种会议，旨在为研究人员提供一个交流信息、寻找合作伙伴、讨论待解决问题的平台。

测序目标

早在2004年，NHGRI的DNA测序技术发展目标^[4]就已经写入了委员会两个阶段的战略规划。那时，美国国内使用的最好的测序技术是基于毛细管电泳芯片（capillary array electrophoresis, CAE）的Sanger测序法。Sanger CAE测序法能获得高质量的哺乳动物基因组序列图，所以该技术也得以在NHGRI下属的各大测序中心得到广泛应用^[5]。不过，当时测一个图谱的花费要高达1000万美元。委员会给NHGRI制定的远期目标就是在10年内开发出一种新的全基因组测序技术，能够将费用降低至当时费用的万分之一，即测一个基因组序列只需花费1000美元。而中期目标是在5年内开发出一种全新的全基因组测序技术，希望能将费用降低至当时费用的百分之一，即测一个基因组序列花费10万美元。这种发展计划也符合自人类基因组计划实施以来，测序技术降价的规律，即经过10年左右费用会降低100倍，从最初的每碱基对花费10美元降至现在的0.1美元^[6]。

委员会的战略目标也同样强调测序质量的重要性。检测测序质量的金标准是无序列缺口 (gap) 的序列测定准确率要达到99.999%^[7]。根据以往的经验, 要获得如此高质量的测序图谱必须经过人工修正。所以, 尽管研究人员的最终目标为获得准确率达99.999%的高质量测序图谱, 但就目前来说, 如果能在有序列重叠等其它情况存在的条件下, 通过自动测序获得和小鼠基因组序列图谱质量一样的结果就非常不错了^[8]。

现在, 学界对委员会的“两步走”战略表示高度关注, 因为在委员会领导下, 已经有超过160项科研进展投入科研工作的实际运用当中。查看委员会过去5年来的资金资助对象, 可以发现, NHGRI已经为50个研究团队投入了超过一亿美元的资金。他

们资助项目的范围很广, 包括从测试新方法的可行性到开发一套全新的系统。同时, 他们也非常注重将科研与实际应用相结合。一般来说, 对于完成被投项目的近期目标来说, NHGRI投入的资金量还是绰绰有余的。

最近, 眼看着NHGRI的中期目标, 即测一个基因组序列花费10万美元马上就要实现了, 他们立即又将资金资助的对象转向了远期目标, 即测一个基因组序列仅需1000美元。尽管有一些被NHGRI资助的研究团队从一开始就已经将目标锁定于1000美元, 但那些为100万美元这个中期目标而工作的研究团队却非常有可能取得更好的成绩。因为他们也许只需要在他们研究进展的基础上进行一些创新、补充或者优化, 说不定就能为人们带来更便宜的测序技术^[9]。

实现方法

现在, 有几种方法都有可能取得远期目标的最终成功。人们可以从www.genome.gov/10000368#6网站上看到最近几年里NHGRI都对哪些100万美元的研究工作和1000美元的研究工作给予了奖励。

此外, 该网页上还有对这些工作的简要介绍。在100万美元这个范围, 使用得最多的是四色DNA合成测序法 (Sequencing by synthesis, SBS, 参见文后小词典1) 以及寡核苷酸探针结合循环测序法 (Sequencing by ligation, 参见文后小词典2), 而在1000美元这个范围占优势地位的则是纳米通道测序技术 (nanopore, 参见文后小词典3)^[10-12]。

其它受资助的项目包括: 对传统Sanger测序法样品准备阶段的整合及小型化处理、毛细管电泳芯片分离及检测^[13]、实时监测DNA多聚酶介导的荧光标记核苷酸掺入反应 (<http://visigenbio.com/technology.html>)^[14,15]、通过监测转录后蛋白质合成序列来推导出核酸序列、使用荧光标记的或其它标记方式标记的寡核苷酸文库芯片杂交测序^[16]、通过监测DNA在合成过程中分子量的改变来测序、通过DNA分子与寡核苷酸探针结合或解离来测序、或者依靠DNA分子能否通过某个特定周长的化学环来测序、以及通过固定在介质表面的DNA分子的延展性来测序 (因为DNA分子的延展性与其分子长度相关)^[17]。

作为NHGRI和其它几个资金资助的科研项目, 这些工作的前景都是非常光明的。最近, 委员会把资金又投向了几个新方法, 例如物理测序法、化学测序法、生物化学测序法、光谱测序法、显微镜测序法^[18]。最后两种方法都是通过直接“看”DNA分子来测序的新方法。

学界观点

NHGRI的这项伟大工程还有一个重要的组成部分, 那就是每年一度的科研大会。NHGRI希望通过这个大会建立一个供科学家交流思想的平台, 既促进合作, 也促进竞争。大会主要讨论学界的科研进展, 科学家可以在大会上畅所欲言, 分享科研心得体会。大会的目的是希望能帮助大家群策群力, 共同攻克难关。现在已经有好几个研究小组相互达成了合作关系。例如, 有一个研究小组在某个技术领域非常专业, 他们已经研制出某种特殊的化学试剂或解决了某项技术中的关键难题, 他们的工作对其他碰到同类问题的研究小组来说是至关重要的。

今年大会上，与会者得出一致结论，那就是目前，要想大幅度降低全基因组测序费用还存在着几个关键难题。例如，使用四色DNA合成测序法，对表面化学试剂的严格控制是保证测序效率和DNA与探针结合特异性的关键，尤其是在使用荧光标记的测序反应中更是如此。对核苷酸进行荧光标记必须非常注意染料的荧光强度和光稳定性，要注意荧光标记核苷酸的纯度，还要注意DNA聚合酶的兼容性（polymerase compatibility）和其它许多方面的因素。监测荧光时又需注意高分辨率、快速、敏感、低背景噪声、高颜色分辨能力等因素。总体来说，使用该方法测序在处理大量样品时需要保证DNA合成的效率以及荧光信号的强度。

今天，尽管的新一代测序技术已经在不同的方面取得了不同程度上的进步，但只有最基本、最

彻底的革新才能取得突破性进展，才能获得更长的测序长度、更低的错误率以及更快的数据处理速度等。对于纳米通道测序技术来说，关键的挑战包括：要有可靠的且成熟的纳米通道制作工艺、传感器的电子化、控制DNA分子通过传感器时的方向与动作，还有最关键的一点就是要能很好地分辨掺入的碱基。

与会者都非常清楚什么问题对于他们的研究领域来说最为关键。在大会上讨论这些问题能帮助与会者修正他们的研究思路，更快、更好地取得进展。召开大会还有一项重要的意义，那就是还有一些科学家或工程师，虽然他们并没有参与测序技术的研究，但他们的专业知识对于其他正在从事这项研究的科学家来说也是非常重要的，说不定就能帮助他们解决大问题。

未来展望

自2004年NHGRI开展了降低基因组测序费用这项宏大工程以来，已经有好几种新型测序技术平台进入了商业化阶段，并且已经在世界各地的实验室中得到了大规模的应用。

NHGRI资助了位于瑞士巴塞尔的Roche/454公司（<http://www.454.com/enabling-technology/index.asp>）^[19]以及位于美国加利福尼亚州福思特市的Applied Biosystems公司（http://marketing.appliedbiosystems.com/mk/get/SOLID_KNOWLEDGE_LANDING）的研发工作。目前，上述两家公司的测序系统以及来自于美国圣地亚哥的Illumina/Solexa公司的系统（www.illumina.com/pages.ilmn?ID=250）^[20]都得以应用于NHGRI资助的实验室以及世界其它的实验室和测序中心。

同时，来自目标用户的反馈信息则有助于仪器制造商更进一步完善测序平台。此外，NHGRI也资助了最近刚刚兴起的剑桥大学的Helicos系统（www.helicosbio.com）。不过，目前还没有太多的实验室使用过该系统，所以也没有太多的相关资料。

尽管目前新一代测序技术所获得的数据质量及其处理的数据规模比较令人满意，但依然存在许多生物信息学方面的问题。最显而易见的问题就是，如何

收集、存贮、整理和分析如此大量的DNA序列信息。接下来，我们面临的稍微复杂一点的问题就是如何整合这些序列信息。在人类基因组计划完成之际，Sanger CAE测序法一次能测700~900个碱基对，并且已经能很好地使用生物信息学工具来分析这些基因序列或集合基因数据^[21]。不过，使用新兴测序技术得到的数据无论从形式上还是从本质上来说，都和以前使用Sanger CAE测序法获得的数据有很大不同。此外，使用新兴技术能保证准确测序的长度要比使用Sanger CAE测序法短得多^[22]。

被广泛接受的质量度量学（quality metrics）尽管也得到了发展，但目前还不能投入到实际应用当中。同时，由于新技术存在通读长度短的问题，所以需要更多的测序次数，而且更多数目的短序列也需要有更好的工具以整合分析，只有这样才能保证测序的质量。

有一些测序实验室已经发现，将几种新技术结合起来就能在花费相同的前提下得到最高质量的测序结果。不过，人们还需要通过生物信息学的帮助来优化这种多技术结合的测序方法。目前，如何处理新测序技术得来的数据还有待探

索，不过最近出现并且日趋普及的新方法——配对端点测序方法（paired-end read protocols）以及其它更有效的新算法和方法）的出现，将有望解决这一困难^[23,24]。

随着科研人员的不断努力，我们将会看到有通读长度更长、准确率更高的测序方法问世。所有这些进展都必须依赖于核酸化学、聚合酶活性、荧光标记及检测手段、表面化学、图像分析以及软件等各领域的通力合作才能实现。

目前，新一代测序技术已经投入使用了，并且得到了大量的在生物医药方面具有重要意义的数据库^[22]。那么下一步该做些什么呢？更新一代的测序技术正在紧张的研发之中，这些更新一代的技术首先要解决的就是如何延长通读长度。尽管研究人员不断完善现有的测序技术，但更新的技术，例如通过激活的DNA聚合酶实时读取序列^[14,15]以及纳米通道测序技术已经表现出了更大的优势。这些技术一次可以通读数千个碱基，并且可以对同一样品重复测序，所以准确率更高^[25~27]。

尽管美国加利福尼亚州门罗帕克的Pacific Biosciences公司宣布，他们将于2010年发布零模式波导密封系统（zero-mode wave-guide confinement system）^[28]，不过我们还是不能确定这些新技术何时才能真正投入到实际应用当中，或者是否真能投入到实用当中。所以，对于新技术来说，挑战与机遇是并存的。

NHGRI的最终目标当然不仅仅是降低测序的费用这么简单，他们是想开发出更好、更快捷且价廉的测序方式以造福人类健康。之所以启动这项计划是因为他们相信，如果能了解一个人的基因组数据，就能改善一个人的健康状况。人类基因组计划已经为我们应用基因组数据打下了坚实的基础，实际上也已经向我们打开了基因药物这扇大门。随着人类基因组数据的公布，以及NHGRI带来的测序技术的进步，目前已有无数科学家发现人类基因序列的改变与疾病有着紧密联系。伴随着人类基因组测序以及基因组功能学、环境微生物学、人体微生物学等学科的进展，人们对基因序列的了解也会更为深入。

接下来，人们要面对的就是个性化医疗的问题了。要提供个性化医疗服务，首先就必须能够对成千上万的患者进行个体基因组测序，并且测序的质量要比现在高才行，然后还必须要从这些个体基因组序列中发现与健康或疾病相关的信息。只有这样，才能够提供个性化医疗服务。个人测序技术的前景是光明的，因为它能帮助我们借助个人基因组信息来进行诊断与治疗。而对于测序技术的研究人员和生物医药研究人员来说，1000美元测序计划或许会是一个终身的挑战。

原文检索：

www.nature.com

 YORK/编译



Jeffery A. Schloss/原作者

Jeffery A. Schloss是NHGRI院外研究处（Division of Extramural Research）技术研发协调中心（Technology Development Coordination）的项目主管。

更多关于Jeffery A. Schloss的介绍，请至：

<http://inc2.inc-conf.net/SpeakersFiles/SpeakerAbstractsBios/SchlossBio.pdf>

参考文献

1. <http://www.genome.gov/10005717>
2. <http://www.genome.gov/12010624>
3. Collins, F.S. *et al. Nature* 422, 835-847 (2003).
4. <http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-04-003.html>
5. <http://www.genome.gov/11508922>
6. Service, R.F. *Science* 311, 1544-1546 (2006).
7. <http://www.genome.gov/10000923>

8. Mouse Genome Sequencing Consortium. *Nature* 420, 520-562(2002).
9. <http://www.genome.gov/Pages/Research/DER/GTP/GTPPubs&Patents.pdf>
10. Shendure, J. et al. *Nat. Rev. Genet.* 5, 335-344(2004).
11. Bayley, H. *Curr. Opin. Chem. Biol.* 10, 628-637(2006).
12. Zwolak, M. & Di Ventra, M. *Rev. Mod. Phys.* 80, 141-165 (2008).
13. Blazej, R. et al. *Anal. Chem.* 79, 4499-4506 (2007).
14. Korlach, J. et al. *Proc. Natl. Acad. Sci. USA* 105, 1176-1181(2008).
15. Bashford, G. et al. *Opt. Express* 16, 3445-3455 (2008).
16. Lizardi, P. *Nat. Biotechnol.* 26, 649-650 (2008).
17. Ashcroft, B.A. et al. *Small* 4, 1468-1475 (2008).
18. <http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-08-008.html>
19. Margulies, M. et al. *Nature* 437, 376-380 (2005).
20. Bentley, D.R. *Curr. Opin. Genet. Dev.* 16, 545-552 (2006).
21. International Human Genome Sequencing Consortium. *Nature* 431, 931-945 (2004).
22. Mardis, E. *Annu. Rev. Genomics Hum. Genet.* 9, 387-402 (2008).
23. Brockman, W. et al. *Genome Res.* 18, 763-770 (2008).
24. Zerbino, D.R. & Birney, E. *Genome Res.* 18, 821-829 (2008).
25. Korlach, J. et al. *Nucleosides Nucleotides Nucleic Acids* 27, 1072-1083(2008).
26. Williams, J.G.K. et al. *Nucleic Acids Res.* published online, 22 August 2008(doi: 10.1093/nar/gkn531)..
27. Gershow, M. & Golovchenko, J.A. *Nat. Nanotechnol.* 2, 775-779(2008).
28. Foquet, M. et al. *J. Appl. Phys.* 103, 034301-1-034301-9(2008).

小词典

1. Sequencing by Synthesis, SBS: 四色DNA合成测序法。

多聚酶反应过程中在一种固体表面上进行的DNA合成测序(SBS)是描绘DNA序列的一种经典方法,研究人员通过将一种可裂解荧光基团连接到碱基上并利用一种小的化学可逆部分诱导3'-OH基团以使它们能够被DNA聚合酶识别为反应底物,从而将四种核苷(A、C、G、T)修饰成一种可逆的终止子。研究人员发现,一种烯丙基能够成功地作为荧光基团和3'-O-烯丙基修饰核苷的连接剂,从而形成化学上可裂解的荧光核苷可逆终止子——3'-O-烯丙基荧光基团以用于SBS。

荧光基团和一个DNA衍生产物上的3'-O-烯丙基(通过在一种聚合酶反应中整合3'-O-烯丙基-dNTPs-烯丙基荧光基团产生)在30秒中能够被缓冲液中钯(Pd)催化的烯丙基化作用同时移除。这种一步完成的双重脱烯丙基反应因此能够控制聚合酶反应,增加SBS的效率。通过使用这种类型的荧光核苷类似物,从而得以在一种DNA芯片和一种四色荧光扫描仪上精确地测序含有均聚物区域的DNA模板。

2. Sequencing by Ligation: 寡核苷酸探针结合循环测序方法。PCR直接测序是指对PCR产物进行的直接序列分析,而不像传统的测序技术先将

DNA待测片段克隆于测序载体上再测序。这不仅大大地简化了操作步骤,节省大量的人力和物力,而且可实现自动化操作,加之新的荧光检测技术的应用,使测序的效率大大提高。采用PCR循环直接测序时,应首先将扩增产物转化为单链测序模板。目前,常用的转化方法为不对称PCR,即在反应体系中引物浓度的差异来形成单链DNA,通常侧引物的浓度为100:1。当某一引物被耗尽后,另一引物扩增的片段即为单链然后即可用于测序。此外,获得单链DNA还有磁珠俘获法、外切酶消化法及Genomic amplification with transcript sequencing法(简称GAWTS法)。PCR直接测序的最新发展是将双脱氧终止技术与PCR技术相结合进行循环测序,在PCR反应体系中同时将ddNTP加入,并利用同位素或荧光素标记的引物引导扩增后,模板的扩增与测序同时进行,其特异性在于使用的模板量小且不需分离单链。

应用PCR测序有以下优点:

- (1)模板需要量小;
- (2)方法简便,操作易于标准化、自动化;
- (3)测序效率高、准确,在短时间即可完成。

3. Nanopore: 纳米孔道测序法是利用当单链DNA分子在外加电压下通过纳米尺寸的孔道时产生的离子电流阻滞来测序,电流阻滞的调制显示出分子的长度、组成、结构和动态行为。