新一代DNA测序技术的发展现状

本期《生命奥秘》以“新一代DNA测序技术的发展现状”为专题，重点介绍目前最先进的大规模并行测序技术的发展现状，并报道它们在生物学、生物医学领域以及测序行业产生的巨大影响。

一、我们将如何应付海量的基因信息

新一代测序技术带给人们大量遗传信息的同时，却成为限制其广泛应用的一个障碍。

1980年，英国生物化学家Frederick Sanger与美国生物化学家Walter Gilbert建立了DNA测序技术并获得诺贝尔化学奖，至今已有近三十年了。在这三十年，DNA测序技术取得了令人瞩目的进展。目前已进入市场的循环阵列测序平台采用的是与Sanger生物化学测序方法完全不同的原理。在过去几年，应用极为广泛的毛细管电泳测序法采用的则是多线并行阵列格式，它运用尖端的荧光成像技术进行碱基识别。上述各类新技术为生物学研究领域开辟了新的视角，也使实验研究达到一个新的水平。学界对开发这类新技术的兴趣持续高涨，与此同时，人们却发现这些技术存在一定的不足——大量信息数据的产生限制了技术更加广泛的应用，并降低了其市场价值。

过去，研究人员使用Applied Biosystems（ABI）公司的3730XL毛细管电泳测序仪进行基因分析，每年至多能完成六千万碱基的测序量。随着测序技术日新月异的发展，这种情况已经成为历史。在2005年刚刚开始进行新一代测序技术开发时，Roche公司和454公司联合开发的焦磷酸测序仪的分析速度就已经达到了上述提及的ABI仪器速度的50倍之上。也就是从那时起，因基因数据过多而产生的问题凸显了出来，而且这个问题随着其他制造商开

发出更多更快的测序仪而愈加严重。举个例子，ABI的新一代测序平台SOLiD（supported oligonucleotide ligation and detection）单次运行，便可以分析6Gb的碱基序列；而Roche/454测序仪单次运行可以将上述结果转换成12-15个千兆字节（gigabytes）的数据信息；Illumina Genome Analyzer（GAII）测序系统仅在两个小时的运行时间里，就得到10兆兆字节（terabytes）的信息。尽管对于像Applied Biosystems这样的制造商而言，可以为用户提供高达11.25TB的存储量，但对于多数实验室所具有的信息管理系统来说，规模如此庞大的数据信息，就好像是迎面而来的洪水，让人感到难以控制。

过量信息所带来的一个副作用在于，用户无法将初始图像数据进行分类存档，而必须交给相关公司，利用软件对数据进行读取，然后才能对数据进行保存。对于大多数研究人员来说，像这样在每次实验后对原始数据进行处理的方式既繁琐又不经济。与花费上万美元对每一段序列进行备份分析相比，对每一次测序结果进行重新测定显然是一个更简单、更便宜的选择。测序仪制造商称，对原始数据再次进行分析并不能得到更多新的信息。但是，对于454测序仪而言，用户至少可以通过更新的软件从原始数据得到质量更高的序列，从而提高碱基识别分辨率，减少误差。

除数据处理问题之外，研究人员还需要拥有一个足够强大的计算机平台，以便将来自多个测序技术的短小基因片段进行组合，形成基因组外显子。目前问题在于，测序仪生产商仅仅提供用于某些特定基因信息分析的软件，如靶标重测序、基因表达分析、染色质免疫沉淀反应或基因组从头测序等，而并未提供任何其它类型的下游生物学信息分析软件。研究界越来越熟悉这些测序平台对循证生物学的巨大潜力，这也就产生了新的研究问题以及全新类型的试验方法，而这单凭依赖目前的生物学信息是无法满足的。

从这个角度看，SOLiD软件研发公司（http://solidsoftwaretools.com/gf/）于今年七月刚刚兼并了两个新的软件公司，这一举动无疑朝正确的方向迈进了一步。该公司在开放源码许可证下开发软件分析工具，目的就是为了给生物信息学领域提供支持，并为其开发新的算法。

图片来源：Applied Biosystems.

对用户而言，如果能够将数据格式与不同测序平台获得的结果进行比较所得的统计数字进行标准化，无疑具有重大的意义。特别是由于目前以测序平台为核心的市场竞争激烈，因此每个生产商都努力提供最好的数据结果。

在这样的大环境下，对数据及不同产品的比较结果进行标准化，便显得尤为重要。有一个方法可以更好地对不同的新一代测序技术进行比较，那就是建立一个微阵列定性分析小组（Microarray Quality Control consortium），不仅可以对不同的技术结果进行比较，而且还可以将新技术结果与DNA微阵列或定量PCR进行比较。

综合以上各类因素，可以预见的是，新一代测序平台在近几年内，仍然会局限于少数实验室及研究者，而大多数缺少能够对基因信息进行进一步分析的实验室则无法从新测序技术中获益。对大多数实验室而言，即使新一代的测序平台能够提供更多的信息，DNA微阵列分析仍然是一个相对便宜的选择。例如，在转录分

析中，虽然新一代测序结果不仅能给出具有很大动态范围的基因丰度信息，同时还可提供剪切变异信息以及SNP数据，但是这些数据结果都需要进行不同的DNA微阵列分析才能获得。

那么，有没有什么方法可以解决这些问题呢？首先，相关的资金授予机构应该对生物信息学的发展予以与测序技术同等的关注；此外，由于生物信息学发展中的瓶颈已经限制了测序机器的销售，测序仪生产商也应该联合起来解决这一难题。同时，制造商应该致力于制定以研究领域为基础而不是以不同公司为基础的生物信息学解决方案。

因此，如果新一代测序平台真的能够带动基因组测序“普及化”——让基因组测序从大型测序中心走入每个研究人员的实验室或者小型研究小组，那么还需要研究人员付出更多努力，开发出经济实惠的分析软件以及数据管理系统。目前的状况是，与新一代测序技术相关的生物信息学分析工作仅仅掌握在少数人手里，但是这一具有重要价值的技术毫无疑问应该由大多数人掌握。如果数据处理问题不能得到有效解决，那么ABI公司的SOLiD系统、454公司的超高通量基因组测序系统——GS FLX、Illumina公司的GAII系统等新一代测序仪就永远无法真正出现在能够展现其价值的舞台上。

原文检索：Editorial. (2008) Prepare for the deluge. *Nature Biotechnology*, 26(10): 1099.

YORK/编译

二、传统的DNA测序技术——Sanger测序法

自上世纪90年代初，所有的DNA测序操作几乎无一例外地全部采用半自动化毛细管电泳Sanger测序法（图1a）。而后来出现的高通量测序方法则首先采用以下两种方法中的一种对DNA进行预处理（表1）。

表1 对DNA进行预处理的两种方法

此法也可见于从头测序的鸟枪法（shotgun *de novo* sequencing），即先将大量随机切割产生的DNA小分子片段克隆到高拷贝数质粒（即松弛型质粒）中，随后再将这些质粒用于大肠杆菌转化；
此法是在对特定序列进行再次测序时使用的方法，即采用该特定序列的特异引物进行PCR扩增，从而获得该序列。

无论采用以上哪种方法处理后，我们均可以得到大量的待测序模板片段——质粒或PCR产物。随后，测序仪会进行“循环测序”反应。在每一轮测序反应的引物延伸步骤中，会随机引入已被四种不同颜色荧光分别标记的ddNTP（ddATP、ddTTP、ddGTP、ddCTP）以终止延伸反应。这样就形成了大量末端被荧光标记的、长短不一（终止位点不同）的延伸产物。接着，再用高分辨率的毛细管凝胶电泳分离这些延伸产物，通过对延伸产物末端四种不同荧光颜色的区分，计算机软件会自动“读出”DNA序列。不过，该方法在“读取”每一个碱基信息时都有可能出错。后续操作中，比如基因组组装或者找出变异位点等就是具体情况具体解决了。一般，这种高通量测序仪一次最多只能同时进行96个或384个样品测序。

Sanger DNA测序技术经过了30年的不断发展与完善，现在已经可以对长达1,000bp的DNA片段进行测序了，而且对每一个碱基的读取准确率高达99.999%。在高通量基因组鸟枪法测序操作当中，使用Sanger测序法的费用大约为0.5美元/1,000个碱基。

原文检索：Jay Shendure & Hanlee Ji. (2008) Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135-1145.

YORK/编译