EBIOTECH

生物通技术周刊
新一代测序专刊

第68期
2009年10月30日
全文下载

[新一代测序]

放眼未来，看新一代测序
新一代测序技术之三国时代(上):Illumina
新一代测序技术之三国时代(中):Roche/454
新一代测序技术之三国时代(下):ABI

[样品制备]

揭开基因组捕获的神秘面纱
基因组捕获之有问有答
Roche/454用户畅谈测序样品制备
Illumina用户分享测序样品制备的经验

[体验新一代测序]

专访BIG测序专家胡松年研究员
国内外牛人评说新一代测序技术

[第三代测序崭露头角]

专访Radoje Drmanac：5000元测序的奥秘
廉价的第三代纳米孔测序
第三代测序技术揭密
第三代单分子测序的开山之作
RNA直接测序指日可待

主办：生物通 www.ebiotrade.com

本期责编:余亮 制作:吴春红
广告联系电话:020-87511980 欢迎访问:www.ebiotrade.com

放眼未来，看新一代测序

在过去几年里，新一代 DNA 测序技术平台在那些大型测序实验室中迅猛发展，各种新技术犹如雨后春笋般涌现。之所以将它们称之为新一代测序技术（next-generation sequencing），是相对于传统 Sanger 测序而言的。Sanger 测序法一直以来因可靠、准确，可以产生长的读长而被广泛应用，但是它的致命缺陷是相当慢。十三年，一个人类基因组，这显然不是理想的速度，我们需要更高通量的测序平台。此时，新一代测序技术应运而生，它们利用大量并行处理的能力读取多个短 DNA 片段，然后拼接成一幅完整的图画。

2006 年底，美国 X 大奖基金会设立了基因组 Archon X 大奖，奖金高达 1000 万美元。这项大奖将颁给第一个能在 10 天之内，用不到 100 万美元的费用，完成 100 个人类基因组测序的民间团队。附加条件是覆盖率不小于 98%，误差不大于 1/10000 bp。重赏之下，必有勇夫。454 生命科学公司自 2005 年推出市场上首个新一代测序平台 Genome Sequencer 20 以来，就成为该奖项的有力竞争者。

之后，454 公司、Solexa 公司和 Agencourt 私人基因组学公司分别被罗氏、Illumina 和 ABI 公司收购，都是瞄准了测序这个潜在的巨大市场。人类基因组测序的成本也在持续下降，也许，我们很快就能看到 Archon X 大奖花落谁家。

新一代测序技术的魅力

新一代测序技术究竟有着什么样的魅力，引各大公司竞折腰？那就让我们先来看看它与 Sanger 测序流程的比较（图 1）。Sanger 测序大家都比较了解，是先将基因组 DNA 片断化，然后克隆到质粒载体上，再转化大肠杆菌。对于每个测序反应，挑出单克隆，并纯化质粒 DNA。每个循环测序反应产生以 ddNTP 终止的，荧光标记的产物梯度，在测序仪的 96 或 384 毛细管中进行高分辨率的电泳分离。当不同分子量的荧光标记片断通过检测器时，四通道发射光谱就构成了测序轨迹。

在新一代测序技术中，片断化的基因组 DNA 两侧连上接头，随后运用不同的步骤来产生几百万个空间固定的 PCR 克隆阵列（polony）。每个克隆由单个文库片段的多个拷贝组成。之后进行引物杂交和酶延伸反应。由于所有的克隆都是系在同一平面上，这些反应就能够大规模平行进行。同样地，每个延伸所掺入的荧光标记的成像检测也能同时进行，来获取测序数据。酶拷问和成像的持续反复构成了相邻的测序阅读片段。

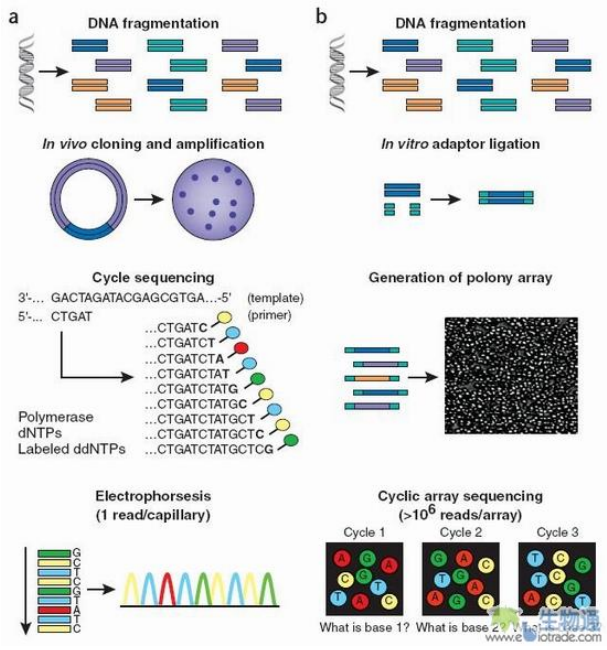


图 1. Sanger 测序与新一代测序的流程比较（图片来自 Nature Biotechnology）。

新一代测序技术备受关注的一个主要原因就是它的通量持续增长，潜力无限。ABI 公司的

SOLiD 3 系统是目前最高通量的系统，单次运行能产生 50GB 的人基因组序列数据，相当于基因组的 17 倍覆盖度。遥想当年 SOLiD 刚刚发布时，通量也只有 2.5GB。短短一年半的时间，随着 SOLiD 系统升级到 SOLiD 3，通量提高了 20 倍，这种可扩展性得益于独特的开放玻片形式和灵活的微珠设计。

Illumina 同样不甘落后，在它的宏伟蓝图中，今年将会实现单次运行获得 95 GB 以上的高质量数据。而依靠近期试剂与软件的升级，Genome Analyzer IIx 能够获得 100 bp 以上的配对末端读长，并在每次运行中产生超过 20 GB 的高质量数据。虽然与 95GB 的目标相距甚远，相信 Illumina 还会在硬件、软件、试剂上有大动作。

新一代测序仪在准确性上也是绝不含糊。SOLiD 系统原始碱基数据的准确度大于 99.94%，而在 15X 覆盖率时的准确度可以达到 99.999%，是目前新一代基因分析技术中准确度最高的。其秘密在于 SOLiD 系统采用专利的双碱基编码技术，在测序过程中对每个碱基分析判读两遍，能够在序列测定中减少原始数据错误，提供内在的校对功能。使用连接酶替代聚合酶方法获得更高的保真度，能够明显减少因碱基误配而出现的错误，可以消除相位不同步的问题。另外，测序过程中定期更换测序引物也能够减少背景噪音和错误率。

虽然还无法与传统 Snager 方法的 1000 bp 读长相抗衡，但新一代测序技术的阅读长度也在稳步提高。Roche 的新一代测序平台 Genome Sequencer FLX 目前的读长为 400 bp，这也是最让它引以为傲的地方。Roche 应用科学市场部经理 Timothy Harkins 曾表示：“我们的平台与其他平台的最大区别就在于测序的读长。其他平台只能产生几十个碱基的读长。”

不仅仅是测序……

新一代测序技术的应用也不再局限于单纯的测序。有了这些测序平台，研究者们能够对未知基因组的生物体样本进行基因表达研究。这意味着他们可以获悉哪些基因被转录，这些基因是否与其他已知的基因同源，抑或它们是全新的。另外，他们还能鉴定表达水平、体细胞突变和剪接变异体。这些都是上一代测序所无法实现的。有关人士认为，新一代的测序方法还会入侵芯片和其他技术领域，给芯片市场带来一定的冲击，但由于价格等问题，要完全取代表达谱芯片还需要一定的时间。

深入的重测序还能让研究人员更多地了解与很多疾病相关的遗传作用。不久前，剑桥大学的研究人员就利用 454 的测序仪，发现一种致病基因的罕见突变有可能降低罹患 I 型糖尿病的风险。这项研究提出了一种从大量候选基因中识别更多 I 型糖尿病特异基因的方法。而高通量测序在耐药性方面的研究也使个性化药物的前景更广阔。

数据分析与储存

虽然测序速度提高了，费用也下降了，但测序产生的海量数据却为后续的分析与储存带来了巨大的挑战。除非你了解如何分析和储存新数据，增加新技术的通量和扩展应用才会变得有用。以千人基因组计划为例，他们不仅面临数据储存的问题，还面对如何比较两个不同个体的基因组的分析障碍。接着就是注释问题——课题组还要对人类基因组进行完整注释。第一个亚洲人基因组图谱的绘制者王俊博士也表示，测序产生的大量数据给后期的生物信息分析带来了巨大的压力，他们面临了现有的生物分析软件无法解决的问题，例如测序数据量较大增长了序列比对的时间、测序序列平均读长较短导致序列很难精确定位，为此，他们独立自主研发出 SOAP、SOAPsnp 软件，“这是我们完成这个项目时最值得骄傲的地方之一”。

新一代的测序平台运行一轮后往往产生 TB 数量级的信息，包括数据和图像。如果你想要存储所有图像，那么在计算机硬件上的花费可能会高于仪器运行所需的费用。Illumina 公司研发副总裁 Tony Smith 认为：“真正大的数据是图像。Illumina 为客户提供储存所有图像的机会，因为有些客户想要这些图像。问题是你每轮获得的图像数据可能是上百 GB 甚至 1TB。而未来数据只会增不会减。客户可以出于质量控制目的存储一组图像，或储存一轮特别重要的图像并备份归档。”

扫清障碍

新一代测序技术尽管优势多多，但价格高也是有目共睹的，一台新一代测序仪的价格大约在 50 万美元，除非实验室测序的工作量非常大，否则是不会考虑购买的。另外，每次开机的费用也不菲。对于 KB 到 MB 范围的小型项目，Sanger 测序无疑还是最佳的选择；但对于全基因组测序、鉴定体细胞突变等大型基因组计划，新一代测序技术则更有魅力。此时，Sanger 测序就不仅仅是试剂的投入了，你还需要购买机器人、处理天文数字的 96 孔板或 384 孔板、维护毛细管测序仪、购置昂贵的生物信息学设备来处理信息流。显然，新一代测序仪就简单得多。

现在的最大挑战就是如何将每个样品的费用降低到更多更小的实验室都可以接受。长江后浪推前浪，所谓 “第三代”或“下下一代”的单分子测序系统将成为新一代测序的有力竞争者。Helicos 公司在去年推出了 HeliScope 测序仪。它的研制是基于 Helicos 公司的单分子测序技术，它可以通过合成互补链技术对数百万个 DNA 片断进行测序而无需对 DNA 链进行扩增。但目前 HeliScope 测序仪正遭受测序错误的困扰，而且其价格惊人，大约是其他测序仪的 2 倍。Pacific Biosciences 打算在明年将产品正式推向市场，它的目标是在 2013 年前实现三分钟读完人类基因组。

新一代测序技术正在以惊人的速度向前发展，而多家公司你追我赶的竞争造就了目前百花齐放的局面。有人预言，五年后个人基因组图谱的价格将是 100 美元。让我们拭目以待吧。（生物通 余亮）

相关阅读：

新一代测序技术之三国时代(上):Illumina

新一代测序技术之三国时代(中):Roche/454

新一代测序技术之三国时代(下):ABI

新一代测序技术之三国时代(上):Illumina

Illumina 公司的新一代测序仪 Genome Analyzer 最早由 Solexa 公司研发，利用其专利核心技术“DNA 簇”和“可逆性末端终结（reversible terminator）”，实现自动化样本制备及基因组数百万个碱基大规模平行测序。Illumina 公司于 2007 年花费 6 亿美金的巨资收购了 Solexa，就是为了促成 Genome Analyzer 的商品化。Genome Analyzer 作为新一代测序技术平台，具有高准确性，高通量，高灵敏度，和低运行成本等突出优势，可以同时完成传统基因组学研究（测序和注释）以及功能基因组学 （基因表达及调控，基因功能，蛋白/核酸相互作用）研究。

Genome Analyzer 自上市以来，已经为千人基因组计划立下了赫赫战功。今年早期，荷兰科学家利用它首次绘出女性的基因组图谱。而就在前两周，《Nature》杂志上一连出现三个人类基因组图谱：炎黄一号-第一个亚洲人图谱；第一个癌症病人图谱；第一个非洲人图谱。它们全是依赖 Genome Analyzer 完成的。哗，一下就来仨！这和第一个人类基因组图谱的 13 年形成了多么鲜明的对照。

根据去年底的数据，Genome Analyzer 已售出约 200 台，估计是市场占有率最广的。前不久，华大基因再添置了 12 台，准备放在香港和深圳的实验室，至此华大基因已经有 29 台 Genome Analyzer。而著名的麻省理工学院和哈佛大学 Broad 研究院拥有 47 台 Illumina 测序仪。众多实验室之所以选择 Illumina，看中的无疑是 Genome Analyzer 的高性价比。

上个月，Illumina 将 Genome Analyzer II 升级到 Genome Analyzer IIx，距年底实现单次运行获得 95 GB 数据的宏伟目标又近了一步。Genome Analyzer IIx 有两个核心特征：其一是更大的试剂冷却器，支持超过 100 个测序循环，进一步提升了系统的易用性和自动化；其二是全新的流动池支架，让每轮运行所得的高质量数据增加 20%。依靠系统软件和试剂的改进，Genome Analyzer IIx 现在能够支持 100 bp 以上的配对末端读长，并在每次运行中产生超过 20 GB 的高质量数据。

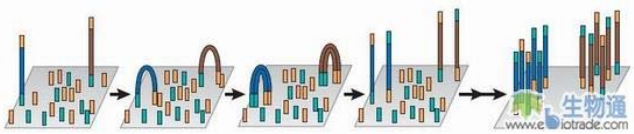
Genome Analyzer 技术的基本原理：

1. 文库制备

将基因组 DNA 打成几百个碱基（或更短）的小片段，在片段的两个末端加上接头(adapter)。

2. 产生 DNA 簇

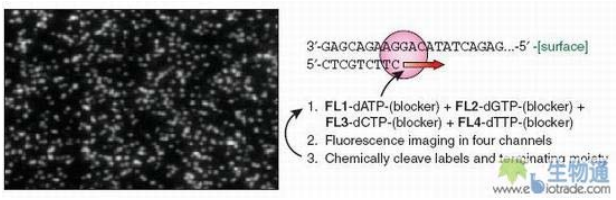
利用专利的芯片，其表面连接有一层单链引物，DNA 片段变成单链后通过与芯片表面的引物碱基互补被一端“固定”在芯片上。另外一端（5’或 3’）随机和附近的另外一个引物互补，也被“固定”住，形成“桥 (bridge) “。反复 30 轮扩增，每个单分子得到了 1000 倍扩增，成为单克隆 DNA 簇。DNA 簇产生之后，扩增子被线性化，测序引物随后杂交在目标区域一侧的通用序列上。



3. 测序

Genome Analyzer 系统应用了边合成边测序（Sequencing By Synthesis）的原理。加入改造

过的 DNA 聚合酶和带有 4 种荧光标记的 dNTP。这些核苷酸是“可逆终止子”，因为 3’羟基末端带有可化学切割的部分，它只容许每个循环掺入单个碱基。此时，用激光扫描反应板表面，读取每条模板序列第一轮反应所聚合上去的核苷酸种类。之后，将这些基团化学切割，恢复 3'端粘性，继续聚合第二个核苷酸。如此继续下去，直到每条模板序列都完全被聚合为双链。这样，统计每轮收集到的荧光信号结果，就可以得知每个模板 DNA 片段的序列。目前的配对末端读长可达到 2×50 bp，更长的读长也能实现，但错误率会增高。读长会受到多个引起信号衰减的因素所影响，如荧光标记的不完全切割。



4. 数据分析

自动读取碱基，数据被转移到自动分析通道进行二次分析。

点击索取Genome Analyzer系统更详细的资料！

Genome Analyzer 系统之所以如此畅销，关键在于其技术上的优势。

1. 可扩展的超高通量

Genome Analyzer 系统目前每次运行后可获得超过 20 GB 的高品质过滤数据。这个技术的可扩展性保证了更高的数据密度和输出，能用更少的经费完成更复杂的项目。到今年底，通量还有望上升到 95 GB，相当于人类基因组的 30 倍覆盖度。

2. 需要样品量少

Genome Analyzer 系统需要的样品量低至 100ng，能应用在很多样品有限的实验（比如免疫沉淀、显微切割等）中。这也是很多研究人员所考虑的因素。

3. 简单、快速、自动化

Genome Analyzer 系统提供了最简单和简洁的工作流程。即使是最小的实验室也能像大型基因组中心一样进行大规模的实验。制备样品文库可以在几小时内完成，一个星期内就能得到高精确度的数据。Cluster Station 可以说是 Genome Analyzer 的核心。由独立软件控制的自动生成 DNA 簇的过程可以在 5 小时之内（30 分钟手工操作）完成。这个自动化的流程不需要进行 Emulsion PCR，减少了手工操作误差和污染可能性，也不需要机器人操作或洁净室。快速的实验流程使 Genome Analyzer 的能力增至最大，而自动化步移降低了项目的时间和费用。

4. 新颖的测序化学技术

Genome Analyzer 通过合成测序来支持大规模并行测序。利用新颖的可逆荧光标记终止子，可以在 DNA 链延伸的过程中检测单个碱基掺入。由于四个可逆终止子 dNTP 在每个测序循环都存在，自然的竞争减少了掺入的误差。

5. 单个或配对末端支持

Genome Analyzer 系统支持单个片段或配对末端文库。文库构建过程简单，减少了样品分离和制备的时间。制备基因组 DNA 的单个片段或配对末端文库需要 6 个小时，只有 3 个小时需要手工操作。2×50 个碱基或更长的读长增加了比对基因组的能力，并拓展了在其他方面的应用。

然而，精明的用户更看重的是性价比，这也是他们选择 Illumina 的重要原因。Illumina 的售价约为 45 万美元，低于 454 GS FLX 的 50 万和 SOLiD 系统的 59 万（以上皆为美国的售价）。此外，运行成本也是一个关键因素。美国凤凰城翻译基因组

学研究院（TGen）的主管 David Duggan 曾表示，当年购买新一代测序仪时，每次运行的费用就成为他下决定的主要因素。他最终选择了 Illumina Genome Analyzer，因为每轮的运行费用为 3000-4000 美元（2007 年的数据），较为合理，而其他测序仪可能更高。当然，他也综合考虑了通量、运行时间和样品量。

弗吉尼亚联邦大学的高原（音译）博士认为：“Genome Analyzer 的操作费用、易用性和可扩展性让我实现了大规模基因组实验。现在，我的小型实验室正在进行过去只能在大型基因组中心才能完成的实验。低样品需求、简单的流程、高质量的数据以及应用灵活性让 Illumina Genome Analyzer 从其他高通量测序技术中脱颖而出。”

高原博士的话已经很好地总结了 Genome Analyzer 系统的特点，那我们就借用来作为本文的结语吧。（生物通 余亮）

相关阅读：

放眼未来，看新一代测序

新一代测序技术之三国时代(中):Roche/454

新一代测序技术之三国时代(下):ABI

新一代测序技术之三国时代(中):Roche/454

454 公司可谓新一代测序技术的奠基人。2005 年底，454 公司推出了革命性的基于焦磷酸测序法的超高通量基因组测序系统——Genome Sequencer 20 System，被《Nature》杂志以里程碑事件报道，开创了边合成边测序（sequencing-by-synthesis）的先河。之后，454 公司被罗氏诊断公司以 1.55 亿美元收购。一年后，他们又推出了性能更优的第二代基因组测序系统—— Genome Sequencer FLX System (GS FLX)。去年 10 月，全新的 GS FLX Titanium 系列试剂和软件的补充，让 GS FLX 的通量一下子提高了 5 倍，准确性、读长也进一步提升。

想当年，GS 20 的出现，揭开了测序历史上崭新的一页。Jonathan Rothberg 博士就是大规模并行测序的发明者，同时也是 454 的创始人。上世纪 90 年代，很多学者也都想到了大规模并行测序，他们试图将 Sanger 测序移到芯片上，但都以失败告终，因为这项技术没有可扩展性。1999 年，Rothberg 的儿子出世，他放了两个星期的陪产假。小家伙出生后被送入婴儿特护病房，Rothberg 非常担心，甚至想获取儿子的基因组信息。这段担惊受怕的经历给了他灵感，他突然意识到焦磷酸测序（pyrosequencing）不仅简单，而且具有可扩展性。两个星期之后，Rothberg 就开始设计芯片和流动室，让测序在更小的反应室中进行，并同时进行几百万个反应。

硬件的设计和制造也只是成功的一半，在样品制备上还有同样漫长的路要走。Rothberg 摒弃了传统的细菌克隆与挑选，将 DNA 打断成随机片段，并寻找一种方法来克隆每个片段。受到其他学者乳液实验的启发，他也想将 DNA 放入油包水的乳液中，这样就省去了反应管。一个好汉三个帮。在 Joel Bader 等人的帮助下，Rothberg 验证了这些想法的可行性，并利用了炸药中的表面活性剂来维持乳液的热稳定性。就这样，乳液 PCR 终于诞生了。

之后，454 生命科学公司用新一代测序仪对 DNA 双螺旋结构的发明者 James Watson 进行了基因组测序。第一份个人基因组图谱的绘制只用了两年时间，花费不到 100 万美元。虽然现在看来这并不算什么，但就当时而言，它相对于人类基因组计划已是质的飞跃。

GS FLX 系统的工作流程

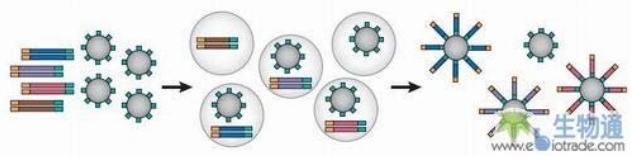
GS FLX 系统的流程概括起来，就是“一个片段 = 一个磁珠 = 一条读长（One fragment = One bead = One read）”。

1）样品输入并片段化：GS FLX 系统支持各种不同来源的样品，包括基因组 DNA、PCR 产物、BAC、cDNA、小分子 RNA 等等。大的样品例如基因组 DNA 或者 BAC 等被打断成 300－800 bp 的片段；对于小分子的非编码 RNA 或者 PCR 扩增产物，这一步则不需要。短的 PCR 产物则可以直接跳到步骤 3)。

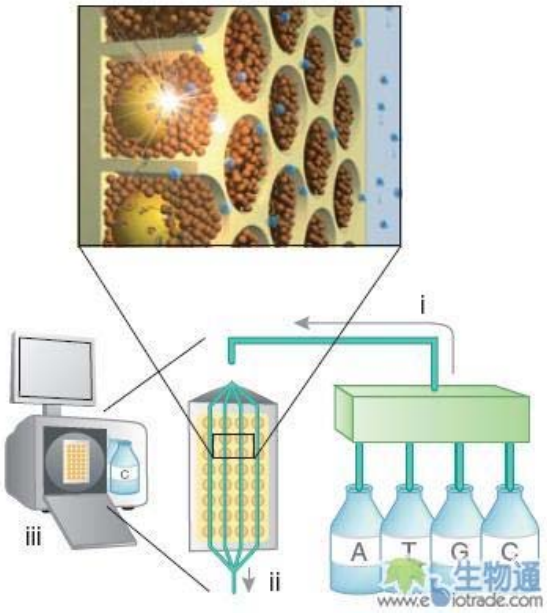
2）文库制备：借助一系列标准的分子生物学技术，将 A 和 B 接头（3’和 5’端具有特异性）连接到 DNA 片段上。接头也将用于后续的纯化，扩增和测序步骤。具有 A、B 接头的单链 DNA 片段组成了样品文库。

3）一个 DNA 片段＝一个磁珠：单链 DNA 文库被固定在特别设计的 DNA 捕获磁珠上。每一个磁珠携带了一个独特的单链 DNA 片段。磁珠结合的文库被扩增试剂乳化，形成油包水的混合物，这样就形成了只包含一个磁珠和一个独特片段的微反应器。

4）乳液 PCR 扩增：每个独特的片段在自己的微反应器里进行独立的扩增，而没有其他的竞争性或者污染性序列的影响。整个片段文库的扩增平行进行。对于每一个片段而言，扩增后产生了几百万个相同的拷贝。随后，乳液混合物被打破，扩增的片段仍然结合在磁珠上。



5）一个磁珠＝一条读长：携带 DNA 的捕获磁珠随后放入 PTP 板中进行后继的测序。PTP 孔的直径（29um）只能容纳一个磁珠（20um）。然后将 PTP 板放置在 GS FLX 中，测序开始。放置在四个单独的试剂瓶里的四种碱基，依照 T、A、C、G 的顺序依次循环进入 PTP 板，每次只进入一个碱基。如果发生碱基配对，就会释放一个焦磷酸。这个焦磷酸在 ATP 硫酸化酶和萤光素酶的作用下，经过一个合成反应和一个化学发光反应，最终将萤光素氧化成氧化萤光素，同时释放出光信号。此反应释放出的光信号实时被仪器配置的高灵敏度 CCD 捕获到。有一个碱基和测序模板进行配对，就会捕获到一分子的光信号；由此一一对应，就可以准确、快速地确定待测模板的碱基序列。这也就是大名鼎鼎的焦磷酸测序。



6）数据分析：GS FLX 系统在 10 小时的运行当中可获得 100 多万个读长，读取超过 4-6 亿个碱基信息。GS FLX 系统提供两种不同的生物信息学工具对测序数据进行分析，适用于不同的应用：达 400 MB 的从头拼接和任何大小基因组的重测序。

GS FLX 系统的准确率在 99%以上。其主要限制来自同聚物，也就是相同碱基的连续掺入，如 AAA 或 GGG。由于没有终止元件来阻止单个循环的连续掺入，同聚物的长度就需要从信号强度中推断出来。这个过程就可能产生误差。因此，454 测序平台的主要错误类型是插入-缺失，而不是替换。

点击索取 GS FLX 系统的更多资料！

新升级让性能全面提升

去年底发布的 Titanium 系列试剂，是对现有 GS FLX 平台的重要升级。升级内容包含耗材、试剂和软件。你无需对仪器的硬件做任何昂贵的升级，只改进试剂和软件，就能立刻实现性能提升。升级之后，每轮测序能产生 100 万个读长片段，高质量（Q20）的读长增加至 400 bp。第 400 个碱基的准确率是 99%，之前的更高。通量也提高了 5 倍，目前每轮运行能获得 4-6 亿个碱基对，所需时间为 10 小时。

- PTP 平板的创新重设计　重新设计之后，PTP 平板上孔的密度更高，利用更小的 DNA 捕获磁珠进行金属覆盖，改善了信号质量，因此读长的数量和长度都明显改善，同时准确性更高。目前孔的直径是 29 um，DNA 捕获磁珠的大小是 20 um。

- 改进的测序试剂　改进的 GS FLX Titanium 试剂显著降低了背景噪音，因此在几乎相同的运行时间内，读长更加长。

- 升级的软件　优化用于超高通量测序的软件，能轻松对更大、更复杂的基因组进行拼接和作图。

- GS FLX 2.0 版 它与以前版本的输出数据也完全兼容，让片段能够共同拼接和作图。

广阔的应用天地

在新一代测序技术中，GS 系统是最多产的。截至 2008 年 9 月，已经发表了 250 多篇高质量的 paper。其中 Nature 20 篇、Science 13 篇、Cell 6 篇、Genome Research 20 篇、PNAS 24 篇。光是这些数据就让人咂舌。这些研究跨越了测序应用的多个方面：82 篇全基因组测序论文包括比较基因组学的从头测序和重测序；54 篇小分子 RNA 研究；37 篇聚焦快速兴起的宏基因组学；27 篇关于转录组图谱分析，包括全转录组拼接和表达图谱；13 篇研究染色体结构和表观遗传学；10 篇有关稀有变异检测的超深度测序这个新领域；11 篇研究古老 RNA。其余的文章关注 454 测序系统的技术和生物信息学。多种多样的应用彰显出 454 测序系统的能力，那些传统意义上无法用测序来解决的问题现在也一并解决了。

454 测序系统除了为多项研究领域开辟了基因组分析之路，同时也加速了探索的步伐。一般来说，研究、分析、撰写并提交论文，经同行评议后发表，需要一年左右的时间。而利用 Genome Sequencer 系统发表论文的速度，显然表明 454 测序结果的数据质量高，且分析简单。超长读长与易用的分析工具相结合，让研究人员能更集中精力于科学研究，而不是研究测序过程中的某个技术细节。这样研究项目能快速完成，接着踏上新的研究道路。

与其他新一代测序平台相比，454 平台的突出优势是读长。目前 GS FLX 系统的序列读长已超过 400 bp。虽然 454 平台的测序成本比其他平台要高很多，不过对于那些需要长读长的应用，如从头拼接和宏基因组学，它仍是最理想的选择。

去年底，美国加利福尼亚大学的研究小组利用全新的 GS FLX Titanium 系列试剂对海洋样品的宏基因组进行测序，发现了一种全新的蓝藻物种，文章发表在 11 月 14 日的《Science》杂志上。这项研究是系统升级后发表的首篇文章。首席研究员 Jonathan Zehr 对于获得数据及分析结果的速度非常震惊。他表示：“多年来我们一直试图培养这种微生物，但都没有成功。有了 GS FLX Titanium，我们在几天之内就通过单次测序运行，从环境样品直接获得了宝贵的基因组信息。这个系统超长的读长对于我们从复杂的微生物群体中鉴定并分析这种独特的细菌基因组来说非常关键。”

最近，在 454 测序平台的协助下，研究人员完成了油棕榈的全基因组测序、拼接和注释。油棕榈是一种重要的经济作物，它的基因组很大，达 17 GB。基因组的测序工作是由 GS FLX Titanium 系统完成的，拼接和分析则是由马来西亚一家生物信息学公司完成的。值得注意的是，这是史上第一次在没有添加传统 Sanger 测序数据的情况下，完成了对大且非常复杂的植物基因组进行从头拼接。这种快速经济的方法为了解多种经济作物的遗传结构打开了大门。

此外，罗氏旗下的另一家公司 NimbleGen 正在全球性地捕获定向重测序市场。NimbleGen 序列捕获芯片与 454 的测序仪结合，能让完整的人外显组测序成为现实，最终将为研究流水线输送技术，并促进个性化医疗的开发。NimbleGen 序列捕获技术将在后文中详细介绍，敬请留意哦。（生物通 余亮）

相关阅读：

放眼未来，看新一代测序

新一代测序技术之三国时代(上):Illumina

新一代测序技术之三国时代(下):ABI

新一代测序技术之三国时代(下):ABI

过去 20 年，美国应用生物系统公司（ABI）在测序方面一直占据着垄断地位。自公司的共同创始人 Leroy Hood 在上世纪 80 年代中期设计了第一台自动荧光测序仪之后，生命科学研究就摆脱了手工测序的繁琐和辛劳，骄傲地迈入自动测序的新时代。直到 2005 年，454 推出了 FLX 焦磷酸测序平台，ABI 的领先地位开始有些动摇。之后，ABI 迅速收购了一家测序公司——Agencourt Personal Genomics，并在 2007 年底推出了 SOLiD 新一代测序平台。从 SOLiD 到如今的 SOLiD 3，短短一年多时间，它已经上演了一出精彩的“一级方程式赛车”。

SOLiD 全称为 supported oligo ligation detetion，它的独特之处在于以四色荧光标记寡核苷酸的连续连接合成为基础，取代了传统的聚合酶连接反应，可对单拷贝 DNA 片段进行大规模扩增和高通量并行测序。就通量而言，SOLiD 3 系统是革命性的，目前 SOLiD 3 单次运行可产生 50GB 的序列数据，相当于 17 倍人类基因组覆盖度。而其无以伦比的准确性、系统可靠性和可扩展性更让它从其他新一代测序平台中脱颖而出。为什么 SOLiD 能轻松实现貌似不可能的任务？让生物通带你从测序原理入手，一探究竟。

SOLiD 工作流程

a. 文库制备

SOLiD 系统能支持两种测序模板：片段文库(fragment library)或配对末端文库(mate-paired library)。使用哪一种文库取决于你的应用及需要的信息。片段文库就是将基因组 DNA 打断，两头加上接头，制成文库。如果你想要做转录组测序、RNA 定量、miRNA 探索、重测序、3’, 5’-RACE、甲基化分析、ChIP 测序等，就可以用它。如果你的应用是全基因组测序、SNP 分析、结构重排/拷贝数，则需要用配对末端文库。配对末端文库是将基因组 DNA 打断后，与中间接头连接，再环化，然后用 EcoP15 酶切，使中间接头两端各有 27bp 的碱基，再加上两端的接头，形成文库。

b. 乳液 PCR/微珠富集

在微反应器中加入测序模板、PCR 反应元件、微珠和引物，进行乳液 PCR（Emulsion PCR）。PCR 完成之后，变性模板，富集带有延伸模板的微珠，去除多余的微珠。微珠上的模板经过 3’修饰，可以与玻片共价结合。看到这里，是不是有一种似曾相识的感觉呢？那就对了，此步骤与 454 的 GS FLX 基本相同。不过 SOLiD 系统的微珠要小得多，只有 1 um。

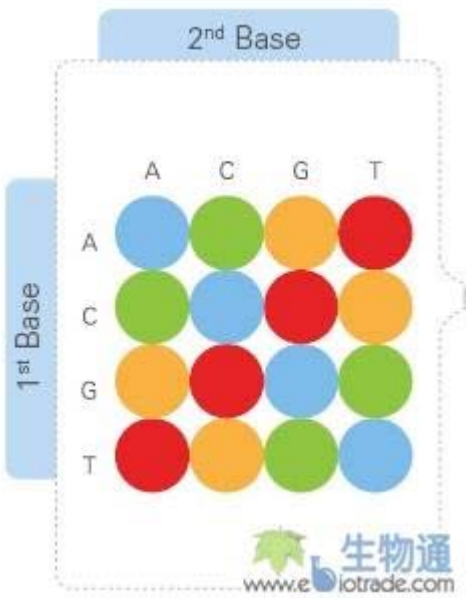
乳液 PCR 最大的特点是可以形成数目庞大的独立反应空间以进行 DNA 扩增。其关键技术是“注水到油”，基本过程是在 PCR 反应前，将包含 PCR 所有反应成分的水溶液注入到高速旋转的矿物油表面，水溶液瞬间形成无数个被矿物油包裹的小水滴。这些小水滴就构成了独立的 PCR 反应空间。理想状态下，每个小水滴只含一个 DNA 模板和一个 P1 磁珠，由于水相中的 P2 引物和磁珠表面的 P1 引物所介导的 PCR 反应，这个 DNA 模板的拷贝数量呈指数级增加，PCR 反应结束后，P1 磁珠表面就固定有拷贝数目巨大的同来源 DNA 模板扩增产物。

c. 微珠沉积

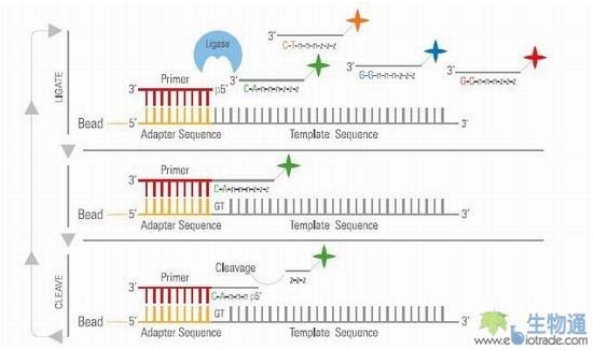
3’修饰的微珠沉积在一块玻片上。在微珠上样的过程中，沉积小室将每张玻片分成 1 个、4 个或 8 个测序区域。SOLiD 系统最大的优点就是每张玻片能容纳更高密度的微珠，在同一系统中轻松实现更高的通量。

d. 连接测序

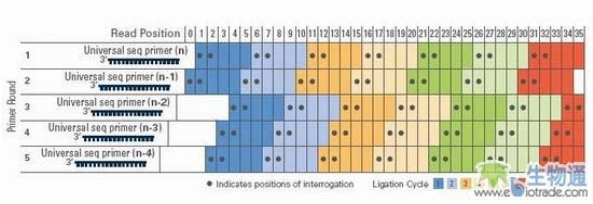
这一步可就是 SOLiD 的独门秘笈了。它的独特之处在于没有采用惯常的聚合酶，而用了连接酶。SOLiD 连接反应的底物是 8 碱基单链荧光探针混合物。连接反应中，这些探针按照碱基互补规则与单链 DNA 模板链配对。探针的 5’末端分别标记了 CY5、Texas Red、CY3、6-FAM 这 4 种颜色的荧光染料。探针 3’端 1～5 位为随机碱基，可以是 ATCG 四种碱基中的任何一种碱基，其中第 1、2 位构成的碱基对是表征探针染料类型的编码区，下图的双碱基编码矩阵规定了该编码区 16 种碱基对和 4 种探针颜色的对应关系，而 3～5 位的“n”表示随机碱基，6～8 位的“z”指的是可以和任何碱基配对的特殊碱基。



单向 SOLiD 测序包括五轮测序反应，每轮测序反应含有多次连接反应。第一轮测序的第一次连接反应由连接引物“n”介导，由于每个磁珠只含有均质单链 DNA 模板，所以这次连接反应掺入一种 8 碱基荧光探针，SOLiD 测序仪记录下探针第 1、2 位编码区颜色信息，随后的化学处理断裂探针 3’端第 5、6 位碱基间的化学键，并除去 6~8 位碱基及 5’末端荧光基团，暴露探针第 5 位碱基 5’磷酸，为下一次连接反应作准备。因为第一次连接反应使合成链多了 5 个碱基，所以第二次连接反应得到模板上第 6、7 位碱基序列的颜色信息，而第三次连接反应得到的是第 11、12 位碱基序列的颜色信息……



几个循环之后，引物重置，开始第二轮的测序。由于第二轮连接引物 n-1 比第一轮错开一位，所以第二轮得到以 0，1 位起始的若干碱基对的颜色信息。五轮测序反应反应后，按照第 0、1 位，第 1、2 位... ...的顺序把对应于模板序列的颜色信息连起来，就得到由“0，1，2，3...”组成的 SOLiD 原始颜色序列。



e. 数据分析

SOLiD 测序完成后，获得了由颜色编码组成的 SOLiD 原始序列。理论上来说，按照“双碱基编码矩阵”，只要知道所测 DNA 序列中任何一个位置的碱基类型，就可以将 SOLiD 原始颜色序列“解码”成碱基序列。但由于双碱基编码规则中双碱基与颜色信息的简并特性（一种颜色对应 4 种碱基对），前面碱基的颜色编码直接影响紧跟其后碱基的解码，所以一个错误颜色编码就会引起“连锁解码错误”，改变错误颜色编码之后的所有碱基。

和其它所有测序仪一样，测序错误在所难免，关键是对测序错误的评价和后续处理。由于 SOLiD 系统采用了双碱基编码技术，在测序过程中对每个碱基判读两遍，从而减少原始数据错误，提供内在的校对功能。这样，双保险确保了 SOLiD 系统原

始碱基数据的准确度大于 99.94%，而在 15X 覆盖率时的准确度可以达到 99.999%，是目前新一代基因分析技术中准确度最高的。

为避免“连锁解码错误”的发生，SOLiD 数据分析软件不直接将 SOLiD 原始颜色序列解码成碱基序列，而是依靠 reference 序列进行后续数据分析。SOLiD 序列分析软件首先根据“双碱基编码矩阵”把 reference 碱基序列转换成颜色编码序列，然后与 SOLiD 原始颜色序列进行比较，来获得 SOLiD 原始颜色序列在 reference 的位置，及两者的匹配性信息。Reference 转换而成的颜色编码序列和 SOLiD 原始序列的不完全匹配主要有两种情况：“单颜色不匹配”和“两连续颜色不匹配”。由于每个碱基都被独立地检测两次，且 SNP 位点将改变连续的两个颜色编码，所以一般情况下 SOLiD 将单颜色不匹配处理成测序错误，这样一来，SOLiD 分析软件就完成了该测序错误的自动校正；而连续两颜色不匹配也可能是连续的两次测序错误，SOLiD 分析软件将综合考虑该位置颜色序列的一致性及质量值来判断该位点是否为 SNP。

在初步了解了 SOLiD 系统的工作原理之后，我们才能明白它的魅力所在。

系统可扩展性

SOLiD 系统采用开放玻片式的结构，使用包被 DNA 样品的微珠来输入基因组信息。微珠密度并不是一成不变的，系统支持更高密度的微珠富集。开放式玻片形式、微珠富集、以及软件算法的结合，能使平台轻松升级到更高的通量，而无需对基础技术和配置做重大改变。这也是 SOLiD 系统平均每季度将通量扩大一倍的原因所在。

无以伦比的通量

目前 SOLiD 3 系统单次运行能产生 50 GB 的人基因组序列数据，相当于基因组的 17 倍覆盖度，这显然是其他任一台新一代测序系统都无法达到的。今年初，ABI 公司和贝勒医学院人类基因组测序中心（HGSC）的科学家总结了他们在千人基因组计划首次数据发布中的贡献。作为商业参与者以及与 HGSC 共同协作，ABI 公司利用 SOLiD 系统产生了超过 460 GB 可作图的序列数据，比这两个机构的预定目标高出了 65%。而通量的升高也有望进一步降低基因组测序的费用，成本只需 1 万美元的人类基因组测序指日可待。

最大的灵活性

SOLiD 3 系统具有两个独立的流动室，让用户能在一台 SOLiD 分析仪中运行两个完全独立的实验——同时提供两套仪器。玻片也能分成 1 个、4 个或 8 个小室。而 20 个条形码序列则提供了额外的灵活性，显著增加了定向重测序、表达和 ChIP 分析的经济性。目前最多能同时运行 320 个样品（2×8×20）。

至此，SOLiD 系统已不再是一台单纯的测序仪，而是成为功能更全面的基因分析仪。除了测序和重测序，还能进行全基因表达图谱分析、SNP、microRNA、ChIP、甲基化等多种分析。

全基因表达图谱分析

芯片大概是目前应用最广泛的从全局角度分析基因表达整体模式的方法。然而，基于杂交技术的微阵列技术只限用于已知序列，无法检测新的 mRNA；而且杂交技术灵敏度有限，难以检测低丰度的目标（需要更多的样品量），难以检测重复序列；也无法捕捉到目的基因表达水平的微小变化------而这恰恰是研究在刺激下或环境变化时的生物反应所必需的。

与芯片技术相比，基于测序的高灵敏 SOLiD 技术可对单个细胞和癌症样品中存在的痕量 RNA 进行整体的全基因组表达图谱分析，每次运行能定位高达 2 亿 4 千万个标签（mRNA 的相对表达水平可通过系统产生的序列标签数目来计算），可检测低至每个细胞中 10-40pg 的总 RNA，即使 mRNA 表达水平很低，SOLiD 系统也能够无偏向性地分析样品中存在的已知和未知 mRNA，从而定量特定 mRNA 的差异表达模式。起始样品比微阵列技术要少得多，尤其适用于来源极为有限的生物样品分析，如癌症干细胞----分析其基因和非编

码 RNA 的表达图谱有助于有助于加速发掘潜在的生物标志物，从而更准确区分不同的疾病类型以及识别疾病易感性，帮助于研究人员更好地了解病变细胞的特性。

更多 RNA 研究

除了单细胞基因表达图谱分析，SOLiD 系统在 RNA 方面的其他应用还包括利用 SOLiD Small RNA Expression Kit 来发现和筛选小分子 RNA，实现在无需预先知道序列信息的情况下高通量发现新的 RNA 分子。这个方案有望显著地提高研究人员鉴别小分子 RNA 的能力，将过去不可能完成的实验变为可能。目前已发现的 microRNAs 还非常有限，SOLiD 可在不知道目标分子 DNA 序列的情况下进行检测和定量小的 RNA 分子，可将样品制备工作从常规方法的四天缩短为仅需一天，是分析在生物样品中表达的已知和未知 miRNA 及其它小分子 RNAs 的有效工具。利用 SOLiD Whole Transcriptome Kit 还可以探索和鉴定全转录本。SOLiD 无可比拟的高通量和测序数据的高精确性使得可以用短序列读长即可测序整个转录组。了解转录组对有助于解开导致复杂疾病的分子通路的秘密。这一系列应用补充使研究人员能在单个超高通量平台上开展综合的 RNA 研究。

SNP 分析

尽管绝大多数的人类遗传信息在所有人中都相同，但是研究人员通常更感兴趣的是研究个体之间微小的遗传差异。这种差异包括单碱基变异，以及被称为结构变异的各种较大片段 DNA 序列变异。结构变异包括 DNA 片段的插入、缺失、倒位和易位，结构变异的 DNA 片段范围可从几个碱基对到数百万个碱基对，可能对基因产生重要影响，并导致人类疾病的发生。SOLiD 流程获得的严密的片段范围，使研究人员可以鉴别出很宽范围内的插入和缺失片段，结构重排也能很容易鉴别出来。这个平台的超高通量使研究人员可轻而易举地获得高度基因组覆盖率的数据，精确鉴定个体基因组中存在的数百万个单碱基多态性 SNP，揭示大量此前未知、具有潜在医学价值的遗传变异，从而促进我们对正常/疾病状态下 DNA 结构变异的了解，以及在更高的分辨率下对结构变异进行深入分析，解释个体之间的易感性差异和对疾病治疗应答的差异，最终实现个性化医疗。

甲基化分析

甲基化是自然发生的 DNA 化学修饰的一种。已知抑癌基因的失活与 DNA 序列特定区域的甲基化有关。而去甲基化则可能导致基因组不稳定和表达模式变化。DNA 甲基化区域可能作为基因在癌症过程中的标记。研究人员一直致力研究从正常到癌变过程中甲基化模式如何变化的，原癌基因异常甲基化模式在癌变过程中扮演怎样的角色。SOLiD 系统运行通量非常惊人，很快就可以做多个样本全基因组甲基化模式检测，使得研究人员可以鉴别基因组中对应元件的甲基化状态，从而帮助研究人员检测甲基化模式是否可以作为癌症的生物标识，以及更好了解甲基化在癌变过程中扮演的角色。

了解SOLiD系统的更多应用！

著名的 Sanger 研究院和 Broad 研究院正利用 SOLiD 系统来探索人类基因组样品中的遗传变异。包括美国华盛顿大学医学院、加利福尼亚大学 Santa Barbara 分校、哥伦比亚大学、澳洲昆士兰大学、日本东京大学、荷兰 Hubrecht 研究院、北京基因组研究院等等研究单位都先后配置了 SOLiD 系统。

SOLiD 系统这个创新的平台将过去种种梦想都变成了现实。未来，它将不仅改变生命科学，甚至可能改变我们的生活。也许，几年后的出生体检报告就是一份个人基因组图谱，告诉你与生俱来了哪些遗传变异，何时以及如何进行及时干预。（生物通 余亮 吴青）

相关阅读：

放眼未来，看新一代测序

新一代测序技术之三国时代(上):Illumina

新一代测序技术之三国时代(中):Roche/454

揭开基因组捕获的神秘面纱

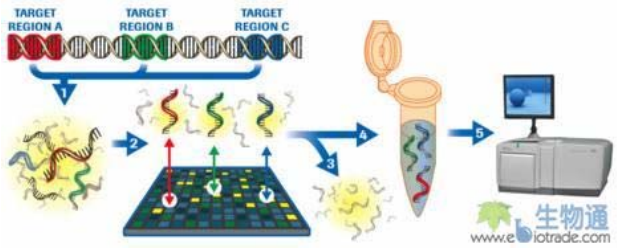
新一代测序技术之所以迷人，是因为它一次运行几天就完成了我们若干人若干年才能完成的任务。在新一代测序的推动下，各种生物的基因组图谱纷纷出炉，一时呈现出百花齐放的繁荣景象。然而，有些研究并不需要对全基因组进行测序，而只需对特定的基因组区域进行研究，比如外显子、SNP 区域或与疾病相关的区域。如何高效地捕获这些区域？在 07 年之前就只能依赖传统的 PCR，不过大量的引物设计、合成、实验及优化，个中艰辛可能只有做过的人才清楚。

不过，科技的魅力就在于其不断创新，生命科学尤其如此。在 2007 年 11 月，《Nature Methods》上出现了三篇文章，都是描述大基因组区域的富集和捕获方法。其中两种是采用基于芯片的杂交捕获方法[1,2]，而另一个则利用分子倒置探针来分离特异的基因组区域[3]。

罗氏公司旗下 NimbleGen 公司立刻注意到了芯片的方法，并开始了最初的验证工作。他们设计了一块芯片，预计能捕获 6726 个基因组区域和一系列大小在 200 KB 到 5 MB 的染色体区域，将片段化和扩增的基因组 DNA 与之杂交。结果显示 65-75%的捕获序列能与目标序列对上。于是，Roche NimbleGen 开始了序列捕获芯片的商业化之路。

经过它的不断完善和扩展，Roche NimbleGen 现在已经拥有多种标准和定制的捕获芯片，包括最新推出的人外显组捕获芯片，它是基于高密度的芯片平台，内含 210 万个寡核苷酸探针，能捕获 18 万个人编码外显子和近 700 个 microRNA 外显子。这些全新的解决方案一步就完成富集步骤，极大地节省了费用、人力和时间。此外，这种芯片可以量身定制，能捕获连续或分散的基因组区域，灵活性非常高。

NimbleGen 序列捕获芯片的原理与一般的芯片类似，不过据罗氏的专家介绍，探针的长度会稍长一些。至于探针的具体信息，那就是 NimbleGen 的专利了，外人不得而知。捕获过程也很简单：基因组 DNA 被打断，然后与定制的序列捕获芯片杂交，没有杂交上的片断被洗掉。富集的目标群体随后被洗脱并扩增，再用 GS FLX 进行高通量测序。



NimbleGen 序列捕获芯片的优势：

- 定向捕获基因组目标区　目前一块芯片最长可捕获 5 MB 的指定基因组区域，特异性和覆盖度都很高。

- 数据可靠　芯片上包含了对照探针，用于验证系统的性能。基于已知、独特的基因座，这些目标区域提供了对照基因座富集水平的定量测定方法。

- 你说，我捕获　每个人感兴趣的区域都不一样，只要你选择出想要捕获的区域，Roche NimbleGen 就会设计并合成一块定制的序列捕获

芯片。捕获区域可以是连续或非连续的基因组长片段、全外显组或其他任何区域。

- 省钱、省时又省事　有了这块芯片，什么引物设计、PCR，都抛到脑后吧，一次就得到了可直接用于测序的所有目标区域。与 PCR 方法相比真是省钱、省时又省事。

今年初，NimbleGen 又突破性地推出了人外显组捕获产品。它利用了优化的设计算法（2008 年 10 月推出的 385K 平台），提供了迄今为止表现最佳，功能最强大的 NimbleGen 序列捕获芯片设计。这种优化设计的目标是以相等的效率捕获所有目标区域，并全面减少所需的测序，从而降低测序费用。人外显组产品是建立在新的高密度 HD2（210 万个长寡核苷酸探针）平台上的 NimbleGen 序列捕获芯片。此产品能在单个芯片上捕获人基因组的几乎所有编码区域（约 18 万个人编码外显子和约 700 个 miRNA 外显子）。

点击索取 NimbleGen 序列捕获芯片的更详细资料

一直以来，人外显组测序被许多研究人员认为是重测序中的“圣杯”，它将带来重大的生物医学突破。外显组测序本身能探索许多功能变化，而这些变化引起了多种常见及稀有的疾病（如癌症和老年痴呆症）。在 NimbleGen 人外显组序列捕获芯片推出之前，外显组的测序无论从技术上还是经济上都是不可行的，因为制备所有人外显子的传统 PCR 方法不但昂贵，而且耗时。Roche NimbleGen 的序列捕获技术和 454 测序系统让完整的人外显组测序成为现实，最终将为研究流水线输送技术并促进个性化医疗的开发。

无独有偶，安捷伦科技公司也瞄准了这个市场，推出了 SureSelect Target Enrichment System。这个试剂盒是源自 Broad 研究院的授权。去年 Broad 研究院的研究人员曾利用生物素化的 RNA 作为诱饵去捕获溶液中的 DNA 靶点。后来，安捷伦获得了此项技术的授权，并加上了质量控制步骤，将它包装成一个试剂盒。安捷伦的产品与 Broad 研究院原始方法的最大区别在于诱饵的长度，现在的 120 聚体取代了原始的 170 聚体。

目标富集，也称为定向重测序、基因组划分或 DNA 捕获，在研究人员只对基因组的某个特定区域感兴趣时相当有用。SureSelect 平台能在测序前捕获一系列外显子或其他基因组目标，并洗掉基因组的其他部分。SureSelect 取代了其他耗时耗力的定向重测序方法，突破了大部分新一代测序流程中的主要瓶颈。

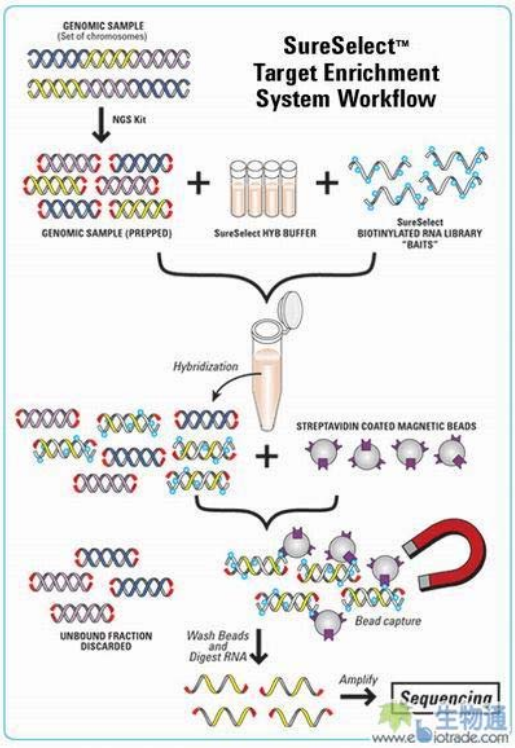
安捷伦的基因组划分产品有别于 NimbleGen 的序列捕获芯片，是一种 In-Solution 的解决方案，包含了客户指定的探针混合物，单管中最多包含 55000 个生物素化的 RNA 探针。这些捕获探针长度为 120 bp，是目前市场上最长的。它们能有效捕获包含未知突变的 DNA，如 SNP、插入或缺失。SureSelect 试剂盒有多种包装，适于几十个到几千个样品，也能与超高通量流程中的自动化相适应。SureSelect 目标富集系统的流程如下：

1. 基因组 DNA 被打断，并组装成测序平台特异的文库形式。在捕获之前对文库大小进行选择，并利用电泳等方法来验证。

2. 精选大小的文库随后与 SureSelect 诱饵共同孵育 24 小时。

3. 加入链酶亲和素标记的磁珠，并通过强磁铁从混合物中钓出 RNA 诱饵-DNA 杂合体。

4. 洗脱磁珠，将 RNA 诱饵降解，剩下的就是目的 DNA。接下来就是常规的扩增和测序了。



用户可利用安捷伦的 eArray 在线设计工具，自行设计 SureSelect 探针混合物。eArray 工具中包含了许多主要的基因组，用户也能上传他们自己的序列。这个直观的在线设计工具是安捷伦定制基因组产品的核心，在芯片研究群体中颇受欢迎，现在又扩展到新的 SureSelect 平台。eArray 让研究人员轻松设计他们所需的工具，不收取任何费用。

Agilent SureSelect 目标富集系统最初是为 Illumina 的 Genome Analyzer 系统而设计的。两天前，安捷伦又和 ABI 公司联合推出适用于 SOLiD 系统的富集系统。

到目前为止，一些早期用户已经试用了安捷伦的 SureSelect 系统，包括 Wellcome Trust Sanger 研究院。其测序技术开发的主管 Daniel Turner 表示："安捷伦 SureSelect 目标富集系统的性能给我们留下了很深的印象。该技术与对手相比，有几个重要的特点：它操作简单，很容易扩展到 96 孔的形式；它比芯片或 PCR 等方法需要更少的基因组 DNA；特异性极佳。实际上，它所需的 DNA 起始量只有其他平台的十分之一，因此对于珍贵样品而言是再适合不过了。"

今年 7 月，安捷伦还与冷泉港实验室合作，推出了以芯片为基础的基因组划分产品。原理与上一代产品相似，SureSelect DNA 捕获芯片也能从完整基因组中提取用户定义的基因组区域，从而大幅度降低测序费用。两者的区别是：溶液型的 SureSelect 目标富集系统适用于样品数量上千的大规模高通量测序研究，包括自动化的高通量流程，而捕获芯片则是它的补充，适合小型化的研究。

在一项合作研究中，安捷伦利用 244K SureSelect DNA 捕获芯片来捕获乳腺癌相关的外显子区域。利用 60 聚体的探针大致捕获到 0.025%的人类基因组，包括 1287 个分离的基因组区域，然后释放并测序。最终，在目标区域获得了 2700 倍的富集。结果验证了 SureSelect DNA 捕获芯片的有效性。测序读数覆盖了目标区域的 99.8%以上，而 98%的目标碱基都至少有一次读取。这些结果证实 DNA 捕获芯片是定向测序的快速有效方案，尤其适合小型研究。

点击索取 SureSelect 捕获产品的详细资料

总的来说，基因组捕获还是一个相当年轻的技术，从出现至今不过两三年的时间。这些新技术都将面临的一个问题是，如果某一天，人类基因组的测序费用真的降到了 1000 美元或以下，基因组捕获该往哪里走。随着测序的通量越来越高，费用越来越低，当有一天测序整个基因组的费用与捕获+测序相同时，捕获还有必要存在吗？有人认为它不会永远存在，也许是一年，也许是五年，没人知道。但无论如何，它让现在的研究人员受益无穷，这是无可非议的。

至于如何选择合适的基因组捕获技术呢，请看生物通后续报道之《基因组捕获之有问有答》。

（生物通 余亮）

参考文献：

1. Albert, T.J. et al. Direct selection of human genomic loci by microarray hybridization. Nat. Methods 4, 903–905 (2007).

2. Okou, D.T. et al. Microarray-based genomic selection for high-throughput resequencing. Nat. Methods 4, 907–909 (2007).

3. Porreca, G.J. et al. Multiplex amplification of large sets of human exons. Nat. Methods 4, 931–936 (2007).

基因组捕获之有问有答

随着新一代测序技术的发展，基因组定向捕获中的学问也就越来越大。PCR 当然也并非不可，过去人们都是用 PCR 来选择性扩增目的片段，但关键是效率不高，引物的设计也相当费神，花费还不小。近两年来，市场上也出现了一些定向捕获基因组的技术。如何选择这些技术？它们有着怎样的优缺点？我们还是来听听专家的意见吧，他们毕竟切身体会过。如果专家们也不能解决你的问题，那希望文末的相关资料能助你一臂之力。

Q1：你选择哪一种捕获方法？为什么？

华盛顿大学基因组测序中心 Jon Armstrong：

我们选择的是液相捕获。溶液中的 DNA-DNA 杂交动力学比固相中更易理解，且液相中的杂交更高效。我们只用 1 ug 左右的 DNA 即可，比目前任一种固相技术更节约样品。而且，液相比固相更容易进行多重和自动化分析。

贝勒医学院 Matthew Bainbridge：

我们用的是罗氏 NimbleGen 序列捕获芯片。目前的 HD2 芯片能在单个芯片上捕获整个 CCDS 外显组（36 MB）。它们提供了目标之间的泊松分布覆盖以及跨越目标长度的均匀覆盖。

范德比特大学医学中心 Shawn Levy：

我们了解过所有的捕获技术，也期望在基因组测序的定向 DNA 富集领域能有持续的发展和进步。到目前为止，我们非常满意芯片的捕获。定制的 NimbleGen 芯片不但高效，使用也很简单，能捕获人和小鼠上中等（250 kb）到大量（6 MB）的基因组 DNA。我们能够使用现有的芯片杂交设备来处理这些芯片，从而降低了资金消耗。样品制备方法的简单性还能对每个芯片设计进行优化，并为上样时严格的 DNA 条件提供灵活性。

埃默里大学医学院 Michael Zwick：

我们只用过固态的方法。基于以下三个理由，我们始终关注这种方法。第一，进行少数实验的花费比溶液方法要低得多。第二，我们已经开发了必需的硬件，能在实验室中开展实验。第三，我们有兴趣捕获所有类型的遗传变异（SNP 和插入缺失），而现在还不知道液相方法是否能有效捕获插入缺失。我们的终极目标是选择最佳的技术，能够解决我们所关注的人类遗传学问题。

Q2：你如何增加捕获的特异性？

华盛顿大学基因组测序中心 Jon Armstrong：

捕获中特异性降低可能是由于：1) 捕获探针的设计不佳；2) 并非最佳的捕获条件；3) 基因组 DNA 中重复序列的封闭不充分；4) 基因组 DNA 与捕获探针的比例不是最优。我们目前正在改进所有条件的参数，以增加特异性。

贝勒医学院 Matthew Bainbridge：

我们做了大量的工作来优化捕获和洗脱的条件，以及封闭 DNA 的使用量。我们还在进行新探针的设计，它能够富集捕获出的 DNA。

范德比特大学医学中心 Shawn Levy：

我们优化了操作步骤中的几个点来改善特异性和整体分析的效率。最终的大小、大小分布和 DNA 的使用量都一一检查过。此外，杂交、洗涤

和洗脱的方法也经过优化，以提高产量和特异性。通过这些努力，我们大约使整体产量增加了 40%。尽管产量有了明显提高，但芯片的特异性似乎变化不大。具体的百分比随芯片设计的不同而不同，但对于同一种芯片设计，这个百分比一般都是非常接近的。

埃默里大学医学院 Michael Zwick：

优化样品的杂交和洗脱有多种方法。其中包括温度、oligo 设计和洗脱液捕获等参数。

相关资料：

Albert T, Molla MN, Muzny DM, Nazareth L, Da¬vid Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock, Gibbs RA.(2007).Direct selection of genomic loci by microarray hybridiza¬tion.Nature Methods.4 (11): 903-905.

Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett.(2005).Direct genomic selection.Nature Methods.2: 63-69.

Bau S, Schracke N, Kränzle M, Wu H, Stähler PF, Hoheisel JD, Beier M, Summerer D.(2009).Targeted next-generation se¬quencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays.Analytical and Bioana¬lytical Chemistry.393: 1, 171-175.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C.(2009).Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.Nature Biotechnol¬ogy.27 (2): 182-189.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C.(2009).Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.Nature Biotechnol¬ogy.27 (2): 182-189.

Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J.(2009).Massively paral¬lel exon capture and library-free resequencing across 16 genomes.Nature Methods.6: 315-316.

Zheng J, Moorhead M, Weng L, Siddiqui F, Carlton VEH, Ireland JS, Lee L, Peterson J, Wilkins J, Lin S, Kan Z, Seshagiri S, Davis RW, Fahama M.(2009).High-throughput, high-accuracy array-based resequencing.PNAS.106 (16): 6712-6717.

Roche/454 用户畅谈测序样品制备

像新一代测序这种复杂实验，光看 Protocol 怎么行？不仅要多看文献，还要自己不断摸索，才能慢慢积累出自己的经验。看看下面几位 Roche/454 用户关于样品制备问题的经验，说不定能让你少走些弯路。

Matthias Meyer 和 Richard Reinhardt 都来自德国马普研究院，不过研究方向不同，Meyer 是进化人类学，而 Reinhardt 是分子遗传学。Bruce Roe 来自俄克拉荷马州立大学。Kenneth Nelson 来自耶鲁大学。Agnes Viale 则是斯隆-凯特琳纪念癌症中心的研究人员。

问题一：当你分离待测序的目的基因组区域时，如何确保准确性和重复性？

Matthias Meyer：除了全基因组鸟枪法测序，我们目前的目标是小基因组区域，它可以通过 PCR 或长距离 PCR 的预扩增而轻松富集。我们认为，长距离 PCR 的成功不仅与 DNA 质量有很大关系，还与 PCR 系统相关，评估几个不同厂家的试剂盒会有很大帮助。

Richard Reinhardt：我们不用 454 进行重测序，主要是用在新物种（de novo）测序和 cDNA/microRNA 文库上。

Bruce Roe：实际上我们很少聚焦特定的基因组区域，一旦需要，我们会利用降落（Touchdown）PCR 以及第二轮的巢式引物和降落 PCR 来扩增目的基因组区域。

Agnes Viale：对于全基因组测序，我们利用几种不同方法（比如 Qiagen 的 Dneasy 和蛋白酶 K/酚-氯仿提取试剂盒）提取的 DNA 来获得高质量的 454 数据。我们不认为纯化方法是一个关键因素，只要获得的 DNA 是高分子量的，且纯度高（260/280>1.7）。对于扩增子重测序，应该在扩增时使用带校正功能的聚合酶。我们一般利用 AMPure Agencourt 试剂盒来纯化 PCR 产物。我们还没有为长片段 PCR 产物的重测序而优化操作步骤。

问题二：如何优化起始 DNA 的量？

Matthias Meyer：如果你使用定量 PCR 来估计测序文库中的拷贝数，那么样品用量是很少的，1 ng 或更少的起始材料就能产生足够的文库，基本上不需要优化起始 DNA 的量。Roche 文库制备手册中建议的定量方法不够灵敏，安捷伦芯片检测则需要微克级的起始样品。因此，在运用 454 平台时，我建议进行定量 PCR。它不仅能够将样品用量显著降低至纳克或皮克，以我们的经验还能得到更为一致的测序产量。我们用这种方法定量了 100 个左右的文库，大部分都给出了最佳的序列数。第一次使用该方法时，最好用一个现有的已滴定过的文库作为起始的参考值。

Kenneth Nelson：滴定是优化起始 DNA 量的最准确、最佳方法。我们还利用 RiboGreen 和 PicoGreen（Invitrogen）分析来定量，以及安捷伦的生物分析仪来筛分。我们发现这些很重要，不能跳过。最近一个德国的小组发表了 qPCR 的方法，但我们还没试过。

Richard Reinhardt：我们一般是用安捷伦的仪器来检查质量，有时候也会用滴定。

Bruce Roe：我们一般用 5 到 10 ug 起始 DNA 来制备文库，在不同阶段，我们会在 Caliper AMS-90 上定量 DNA。在 emPCR 步骤，我们使用的起始 DNA 比 Roche 推荐的量要低一些。

Agnes Viale：此步骤很关键。样品和磁珠的比例不合适会毁了一次运行。如果 DNA 是分离的条带，我们用 PicoGreen 的定量方法来计算样品的摩尔浓度。如果起始材料是弥散的，比如 cDNA，我们还会根据安捷伦 Bioanalyzer DNA 1000 的结果来评估。这个方法是我们的经验，但它真的很好用。

问题三：你采用什么方法来确保样品制备步骤更快速？

Matthias Meyer：我们实验室在测序文库构建前会给多个样品贴上条形码。这一步增加了样品制备所需的时间，于是我们开发出多通道制备步骤，让移液机器人实现半自动化。一旦样品贴上条形码，Roche 的标准测序文库制备步骤只需要几小时。不过，我们发现测序文库会很快降解。将文库分装，并立即冻存，能减少测序失败的风险，并节省很多时间和费用。

Kenneth Nelson：样品制备产生的 DNA 通常足够多次测序，所以一次小心的样品制备是非常高效的。我们还没有发现有任何捷径可走。

Richard Reinhardt：我们通常严格按照 Roche 提供的操作步骤。

Agnes Viale：在这一点上，我们仍是手工处理样品。为了降低试剂费用，我们先用不同的 copy-to-bead 比例对每个样品进行两轮或三轮 emPCR。然后，根据磁珠回收的百分比，我们选择最佳的比例来处理剩余的样品。这个过程绕过了 PTP 的滴定，但是没有减少处理时间。总的来说，我们的样品制备是从星期一到星期四，在星期四晚上开始运行 454。

你是不是也对 454 测序仪有点兴趣，那就赶快索取更多资料吧

（生物通 余亮）

Illumina 用户分享测序样品制备的经验

上一篇说的是 Roche/454 的样品制备，这一次就轮到 Illumina 的用户畅所欲言啦。Ghia Euskirchen 来自耶鲁大学的 Mike Snyder 实验室。高原（音译）来自弗吉尼亚联邦大学。Stephen Kingsmore 是美国国家基因组资源中心的专家。Anoja Perera 来自美国 Stowers 医学研究所。

问题一：当你分离待测序的目的基因组区域时，如何确保准确性和重复性？

Ghia Euskirchen：我们 Solexa（Illumina）仪器的大部分工作是 ChIP 测序。许多为 ChIP-chip 开发的标准也适用于 ChIP-seq，抗体验证对所有 ChIP 实验都很关键。我们用 IP-western 以及质谱来验证抗体。对于重复性，我们会评估三个相同的样品，并关注对照位点。

高原：我们大致通过以下几点来检查准确性和重复性：1. 在目的区域定位读取；2. 利用 Sanger 测序来验证；3. 通过重复实验来查看相关性。

Stephen Kingsmore：国家基因组资源中心目前有两台Illumina的测序仪在运行中，第三台马上也要到了。我们主要有两方面应用：基因组DNA测序和mRNA测序。mRNA步骤是由Illumina的Gary Schroth小组开发的，我们一位同事又稍作修改，而我们的基因组DNA步骤是标准的。对于这些样品类型，我们利用LIMS系统来确保准确性和重复性，它对测序过程中的每个样品进行追踪。另外，我们还在测序仪和Infinium HapMap 550K基因分型芯片上同时运行一套样品，这有助于我们验证SNP检测的准确性。对于核酸变异检测，我们是利用自己开发的Alpheus软件系统（http://alpheus.ncgr.org/）。

Anoja Perera：到目前为止，我们只进行过全基因组范围的实验。以后，如果我们要分离基因组区域，我们也会进行验证实验。验证的类型将取决于分离了什么基因组以及分离的技术。如果我们是用长距离 PCR，那么我们会跑胶验证。我们也可以用 Sanger 测序来验证扩增区域。

问题二：如何优化起始 DNA 的量？

高原：我们曾使用不同的 DNA 起始量来构建文库，并确定哪个浓度的结果更好。我们发现最重要的优化是起始的文库浓度。我们通常使用 3 pM-4 pM 的 DNA 来产生簇。测量文库浓度的办法有很多。我们是用混合法，先用 Nanodrop 来测 DNA 的量，然后和定量 marker 一起跑胶。这两种方法都是高度推荐，因为这个重要参数会决定你的最终测序结果。

Stephen Kingsmore：我们会在两个时候优化起始材料的量。一个是 RNA 文库产生的时候，另一个是生成簇的时候。加入过多或过少的文库都会使序列读取减少。而最佳数量的簇会在每个通道中产生 500 万个读取。我们利用安捷伦的 Bioanalyzer 来确定文库浓度，一般上样 1 pM-3.5 pM。

Anoja Perera：对起始 DNA 来说，数量和质量都很关键。当然，高效的纯化技术是必不可少的。

问题三：你采用什么方法来确保样品制备步骤更快速？

Ghia Euskirchen：我们发现基因组和 ChIP DNA 文库制备相当简单。在文库制备时我们通常会将样品分开，以避免交叉污染。

高原：Solexa（Illumina）的样品制备已经足够简单了。我们几乎是按照它的操作步骤来做的。

Stephen Kingsmore：Illumina 的样品制备过程很快速，大约需要一天，而且能同时制备几个文库。这个过程的瓶颈不在于样品制备，而是簇生成（我们有两台 cluster station 来应对）、序列生成（特别是产生 46 bp 读长时）、碱基检出和基因组比对。

Anoja Perera：提前熟悉一下操作步骤，确保所有的试剂和用品都能用。一定要有备用的。我们就曾试过两次错误扩增，如果没有备用的试剂，我们的实验就会延后。通读操作步骤，划出时间表。基因表达步骤需要三整天，如果准备不充分，你可能还要再多花 8 小时。在等待的时候浏览后续的步骤，了解哪些需要拿出来融化，以便节省时间。

专访 BIG 测序专家胡松年研究员

胡松年，研究员，2003 年至今任中科院北京基因组研究所（BIG）所长助理，基因组生物信息学平台负责人。他于 1996 年毕业于中国农业大学的植物生化系，后前往美国华盛顿大学基因组中心做访问学者。1999 年任中科院遗传所人类基因组中心暨北京华大基因研究中心的总工程师，是国际人类基因组测序计划中国部分（1%项目）具体执行的总负责人。近年来先后承担了“籼稻基因组测序”、“大规模基因组测序技术平台的建立与优化”、“中国－丹麦家猪基因组测序”及“产黄青霉菌基因组测序”、“动植物基因组中可变剪接形式的比较分析”等重大项目。

自 1999 年到 2009 年的十年间，胡松年研究员一直在和测序打交道。谈到新一代测序，我想他是最有发言权的。于是，生物通就新一代测序中大家关心的问题采访了胡老师。当天他正在面试，不过还是抽空接受了我们的访问。

生物通：请您介绍一下中科院北京基因组研究所以及它现有的测序平台？近期主要成果有哪些？

胡老师：北京基因组研究所是在 2003 年正式挂牌成立的。它的前身是华大基因，之前有很多项目是由两个单位共同完成的。到了 2007 年，院里考虑到整个科研的需要，又正好碰上搬家，于是基因组研究所就完全独立出来。研究所已经独立完成了“中国超级杂交水稻基因组计划”、家蚕基因组计划及家鸡基因多态性图谱等一系列的科学项目。目前所里有 9 台 SOLiD、4 台 Solexa（Illumina）以及 3 台 3730 测序仪，还有一台 454 的 GS FLX 将于 8 月份到位。最近主要是应用 SOLiD 来进行一些研究。去年，国家水稻基因组计划的韩斌老师也加入了我们所，我们正在开展一些水稻的测序和研究。

生物通：选择测序新平台的标准是什么？您最看重的是哪些因素？

胡老师：对研究所来说，它更看重的是实验技术和实验方法的提高。这些新一代的测序平台各有各的特点，没有一台机器就好到能包揽所有的项目。最重要的是研究各种仪器的特点，然后针对不同的项目来进行选择，让这些仪器的优势能发挥出来。这些仪器目前的价格都差不多，我看主要还是根据自己的研究目的来选择合适的仪器。

生物通：目前主流的 3 大新一代测序平台，它们各有哪些长处？您的评价如何？它们分别适用于哪些研究？

胡老师：这 3 种测序平台我都用过。454 给大家的感觉是和以前的 3730 更相像。它在升级后，读长达到 400 bp，与其他两种仪器相比，这是它的最大优势。因此更适合于新物种的 de novo 测序。而 SOLiD 和 Solexa 则有些类似，它们的读长都比较短，但通量高，因此更适合于大基因组的重测序和转录组研究，比如真核生物的基因组。SOLiD 是采用双碱基编码的原理，双色球信息对于一些刚开始从事基因组测序的人来说，会有些不适应，觉得不太直观。但是它的通量更高，每次运行能得到 30-50 GB 的数据。这么高的通量对大基因组的测序而言，在成本上就更有优势。Solexa 上市得比较早，所以文献相对多一些，SOLiD 和 Solexa 的读长都比较较短，因此更适合于有 reference 的基因组测序。

其实，在很多时候还是应该将长（读长）和短（读长）的方法结合起来，才能得到一个更为理想的结果。对于 de novo 的基因组测序，一般都是先

用 454 去绘一个草图，再用短读长的仪器来填充，因为它们的通量更高，更为经济。

生物通：比较其它新一代测序仪，您认为 SOLiD 最大的优势是什么？

胡老师：除了通量高，SOLiD 在转录组测序上优势更为明显。它是将 RNA（总 RNA 或 mRNA）直接打断，这样，它就和以往的测序不同，既能测出正义链的表达，也能测出反义链的表达。现在很多人认为 antisense 也像小 RNA 一样，起到重要的调控作用。而且，它还能看到一些非编码重复序列的表达。另外，对于大的基因组测序，从整体上来看，SOLiD 可能更为经济一些。

生物通：您对测序的前景有何展望？对于已经初露端倪的第三代测序您怎么看？

胡老师：第三代测序已经有些苗头了，也可能很快就会面世。目前的主要难点还是样品制备，就是将细胞破碎之后，会有很多影响测序的杂质，在去除杂质的过程中，有可能造成 DNA 的断裂。现在测序技术发展很快，以后读长可能会达到几 K，或者通量更高，但是不能忽视的是样本的处理。即使检测技术再强，如果样本处理时已经丢失了一半 DNA，那么得到的信息也还是不完整的。

生物通：新一代测序中最大的挑战是什么？作为生物信息学的专家，你们怎么解决这个问题？北京基因组研究所主要采用哪些硬件和软件来支持数据分析？

胡老师：数据的存储和分析是新一代测序中最大的两个问题，几天下来的数据量可能就在 TB 级别。为了配合新一代测序仪，研究所正在与浪潮公司合作，配备了 600 TB 的存储空间、十万亿次的刀片服务器以及 128、256 MB 的大内存。在软件方面，测序仪自身也携带软件，但这些软件本身肯定既有优点，又有缺点。SOLiD 和 Solexa 这些仪器都比较开放，它们的软件会开放源代码，而且还有一些用户群，可以互相交流。但是国内与国外的研究方向往往有很大差别，国外主要是研究人和模式动物，而国内研究植物的相当多，动植物的基因组存在较大差别，因此还是要自己根据实验目的来调整。我们通常会以项目为导向，将仪器自身的软件与网上其他一些软件进行整合，开发出自己的流程。

生物通：国内研究所主要经费往往来源是国家拨款，但和国外实验室相比还是远远不够的，利用现有实力提供部分商业化服务是一种有益的补充方案，生物通也留意到华大已经对外开展一些测序的服务。你们有没有考虑这项服务？

胡老师：北京基因组研究所作为一个科研机构，和公司还是有一定区别。我们的主要精力还是会放在科研上，我们与很多科研院所都有合作。在今年秋天，研究所还打算先在北京举办一系列技术应用型的讲座，谈谈什么项目该用什么样的仪器去做。同时，我们一直开办“基因组科学与信息”的培训班。这个培训班已经成功开办了 30 多期，应该是国内最长的吧。学员的反映相当不错。我们的学员遍布全国各地，有一些是研究生、博士生，也有一些老师带着课题来，想与我们交流和协作。我们不讲那些参考文献或者综述上能看到的东西，而是将我们在实际研究中的思路、心得体会传授给大家，而且讲课的老师都是在一线工作的，有着丰富的经验。另外，我们每期都会根据学员的建议来调整，再加入一些大家都关心的问题。

生物通：生命科学领域的学生应该如何转入生物信息学领域？您有何建议？

胡老师：学生物的学生一谈起要编程、搞计算机，都会有些怵，不知道怎么上手。我认为，一开始不必对自己要求太高，也不用看很多算法、概率，其实我们主要还是解决生物问题。首先要学会使用软件，有了一定经验后，再学习如何评估软件，在这个过程中再进一步了解它的大致原理，并了解哪些软件的组合适合分析这一类问题。如此，触类旁通，就能够解决更多问题。但是，必要的计算机知识也是不可少的，我们会要求学生学习 Linux 平台和 Perl 语言。总之，学东西关键是要有目的，去解决问题，慢慢地就熟能生巧了。（生物通 余亮）

国内外牛人评说新一代测序技术

目前，新一代测序技术与 iPS 细胞一样炙手可热。可是，没有调查，就没有发言权。只有使用过，才有资格对新一代测序技术品头论足。究竟这些仪器的性能如何，是不是真的如介绍中那么美好？在新一代测序中又会遇到那些难题？我们还是听一听几位测序达人的评论吧。

王俊，博士，深圳华大基因研究院副院长

论测序，华大基因算是国内至 Top 的研究院，装备精良，人才济济。2008 年底，首个亚洲人基因组出炉。这一研究成果公布在权威期刊《Nature》杂志上，文章的通讯作者和第一作者正是来自深圳华大基因研究院的王俊博士。在这篇文章中，研究人员利用新一代测序仪 Illumina Genome Analyzer 完成了中国人基因组的测序，测序量达到 36 倍覆盖率，并且研究人员还比对了 NCBI 人类相关基因组，短读取序列达到 99.97％覆盖率。

生物通：第一个黄种人基因组图谱的公布是我们的骄傲，相比较于水稻、家蚕、家鸡、家猪等动植物基因组图谱，这个基因组图谱的完成是否更困难一些？还是更容易一些，在这个基因组测定过程中是否遇到了一些技术困难？具体有哪些？

王俊博士：相较于水稻、家蚕、家鸡、家猪等动植物基因组图谱而言，第一个黄种人基因组图谱的总体工作相对更加困难一些。我们在测定第一个黄种人的时候采用了新一代测序仪 Illumina Genome Analyzer，虽然测序价格更便宜，测序速度更快，却给数据的存储、处理、分析、展示带来了巨大的挑战，尤其是面临了现有的生物分析软件无法解决的问题，例如测序数据量较大增长了序列比对的时间，测序序列平均读长较短导致序列很难精确定位，而针对这些困难我们独立自主研发的软件（SOAP、SOAPsnp）是我们完成这个项目时最值得骄傲的地方之一。

生物通：在基因组测定过程中主要采用的技术点有哪些？您认为最关键的一项技术是什么？

王俊博士：在基因组测定过程中主要的技术点是基因组测序和生物信息分析。我认为最关键的技术是生物信息分析，因为随着新一代测序技术的广泛使用，测序的成本大大降低，测序速度有所提高，而测序过程也变得相对简单容易，但是测序产生的大量数据却给后期的生物信息分析带来了巨大的压力，因此我认为生物信息分析是在基因组测定过程中最关键的一项技术。

Harold Swerdlow，博士，Wellcome Trust Sanger 研究院测序技术的主管

世界顶级研究院 Wellcome Trust Sanger 研究院至少拥有 37 台 Illumina 的 Genome Analyzer，5 台 ABI SOLiD 和 2 台 454 GS FLX。但是他们也没有完全抛弃毛细管方法，目前仍有 50 台 ABI 3730，用于斑马鱼和猪的基因组计划。光是看这个数据，就已经让人咂舌了。当然，从下面的访谈中你也可以看出，顶级的基因组中心和一般的实验室还真是不一样，他们资金雄厚，站得高，看得远。因此他们的选购标准不适合普通实验室，仅作了解。

Q：你们选择新平台的标准是什么？

A：我们一定要站在最前沿，无论代价是什么，我们都会做。我们经常测试新仪器和现有仪器的新版本。我们要看到这项技术确实能产生合理量的准确序列，才会进行购买。但同时，我们还有开发资源的任务，因此我们对检验新技术很有兴趣，并且我们还能验证现有技术，它们对于其他实验室或许还不够成熟。因为我们希望站在测序的前沿，我们就要比小实验室跑得更快，它们的经费可能只够买

一台仪器。同时，我们有义务与他人分享我们的经验。

Q：你们有着何种数据储存与分析硬件来支持测序仪？

A：我们的计算机设备特地为支持新一代测序而刚刚更新过。我们有 320 TB 的文件服务器来短期存储图像和序列。整套设备能支持大约 30 台 Illumina 的测序仪。当然，我们还会扩充的。

Q：你们会永久储存测序仪所获得的数据吗？

A：就目前来说，是足够的，因此你不必在每次开始新一轮测序之前删除以前的。我们有足够的容量来储存。但一个月之后的情况呢，我不敢说。

Q：能谈谈 Sanger 研究院的新一代测序平台所参与的计划吗？

A：我们正利用 Illumina 参加 Mike Stratton 的癌症基因组计划以及大猩猩测序计划。Julian Parkhill 正用它进行高通量的病原体研究。例如，在高度可变的细菌群体中，任一群体都有很多突变，你很难知道哪个是真正致病的，但是如果你通览大量的群体，你就能得到其他方法无法获得的大量信息。我们主要利用 454 的仪器进行病原体测序。

Q：在使用这些新平台时，你们遇到的技术及数据处理上的最大挑战是什么？

A：我想对于用惯了 ABI 毛细管测序仪的人们来说，这些新仪器并不是开箱即用的。你不能只是插上插头，然后就等着在电脑上分析数据。它还需要进行许多开发和支持，这是技术上的挑战。就数据方面而言，最近很多人在讨论储存及计算需求。每个人都想储存图像，让问题更加恶化。但是我认为这个问题不难解决，你可以投更多的钱去买更多的硬件。当然这对小型实验室来说比较困难。

最大的挑战是去了解如何以最优的方式提取和分析数据，因为这些数据我们并不熟悉。例如颜色区分、碱基检出、校准、数据的标准化这些问题。如果你能解决这些问题中的一部分，我们就能获得更好的分析技术，也就能从相同的数据中获得更多更高质量的碱基。

另外一个问题是仪器厂商的品质衡量与用户不一致。人们不知道该使用哪个判断阈值（cutoffs threshold）。我们该把判断阈值设在哪里，才能得到最好的数据，但又不抛弃过多的数据？这个问题非常棘手。

David Duggan，博士，凤凰城翻译基因组学研究院（TGen）的主管

David Duggan 负责 TGen 的两个基因分型中心。他们利用 Affymetrix、Illumina、Sequenom 和 ABI 的技术进行着多项基因分型研究。后来，他购买了一台 Illumina 的 Genome Analyzer，将高通量测序融入了实验设计中。Duggan 博士很详细地谈论了当时选购 GA 时的考虑因素，值得国内的实验室借鉴。不过，那已是两年前的事了，目前的选择又多了很多，还需要大家重新评估。

Q：你为何决定购买 Illumina 的测序仪？

A：你别忘了，我们是在（2007 年）3 月做决定的。当时只有 454 和 Illumina 两种选择。我们也和 ABI 联系过。Helicos 也联系了我们，谈到 HeliScope。但我们不想为 SOLiD 再等 9 个月，而 HeliScope 还需要 1 年多的时间。

我们也不是光从便利性考虑。我们很满意 Illumina 系统的一些特征。比如说，能够进行 1 GB 的基因组 DNA 测序；仪器上的运行时间在 3 天。HeliScope 的预计运行时间要长得多。同时，Illumina 的样品量为 0.1-1 mg，与我们的实验设计相符。最后，一个很大的因素就是运行费用。Illumina GA 的运行费用在 3000-4000 美元，比较合理。就这样的费用而言，除了 NIH 的拨款，我们还能从其他地方获得基金。所以，购买 GA 并是出于某种考虑，而是上述种种因素的综合。

Q：你能不能给我们例举一下如何将高通量测序整合到研究中？

A：举个例子，我们正在进行一个合作项目，利用 tag-SNP 方法来筛选 52 个候选基因。理想上

我希望对部分群体中全部 52 个基因进行重测序，不仅鉴定出 SNP 变异体，还有插入和缺失多态性，然后再根据数据设计出实验方法来对 7200 个样品进行基因分型。它比单独的 SNP 研究更全面。

我们设想的另一个实验设计是，目前，我们是分阶段进行基因组范围的研究。在每一个阶段，我们将基因组区域逐渐缩小。一开始，我们研究 4000 个样品的 50 万个 SNP。然后根据预算，鉴定前 1000 个或几百个 SNP，并在一个确认的群体中进行基因分型。之后在第三阶段，我们挑出少数有意义的 SNP，并开始重测序。

而有了新一代测序技术的高通量，我们不再限制在少数候选区域。我们能将管道扩宽一些，对几十个候选基因区域进行重测序。比如之前的一项研究，他们鉴定出人类基因组上 II 型糖尿病的十个致病区域。每次测序一个？不，我想一次全部测序。新一代测序技术也赋予我们这个能力。它比 Sanger 测序更便宜，也更高效。

Mark Skolnick，博士，Myriad Genetics 公司的创始人之一，现任首席科学家

Skolnick 博士是 Myriad Genetics 公司的 CSO、技术奠基人。该公司的发展战略是开发急需的医疗保健产品，主要涉及肿瘤、老年痴呆症和抗病毒等几个领域。他的研究小组克隆了乳腺癌、卵巢癌、前列腺癌、肥胖等疾病的易感基因。另外，他们还利用 Sanger 测序和 454 的 Genome Sequencer 对葡萄藤和苹果的基因组进行了测序。在中国，很多测序工作也是围绕植物展开，那么 Skolnick 博士的经验可能会有一定的借鉴意义。

Q：你为什么选择 454 技术来进行苹果和葡萄藤项目？

A：当时我们受意大利一所研究院的委托，刚完成了葡萄的项目，并开发出一种高度自动化的引物步移平台来填补缺口。那时 454 刚上市，我们就想 454 的 4 倍覆盖度能够很好地填补剩余的缺口。实际上，它完成地非常好，我们也就不需要再进行任何引物步移。对于葡萄和苹果而言，测序都是复杂的项目，因为它们都是非近交的天然生物。复杂度在于你实际上要同时测两个基因组，母本染色体和父本染色体。如果你发现序列差异，你还必须解释到底是错误还是多态性。

Q：对于苹果基因组项目，你使用了与葡萄不同的策略。你能谈一谈这些吗？

A：在葡萄项目中，我们基本完成了拼接，打算开始引物步移时，才决定使用 454。我们利用了 7 倍 Sanger 覆盖度和 4 倍 454 覆盖度。在苹果项目中，我们只利用 BAC 和 fosmid 完成了 4 倍 Sanger 覆盖度，然后，就加入了 10 倍 454 覆盖度，其中大部分是平均 500 个碱基的长读取。现在，我们的总覆盖度是 14 倍，而不是 11 倍，因为有两个染色体，父本和母本，那么每个多态性的平均覆盖度为 7 倍，在确定两个染色体的特定差异上，可靠性是进一步增强了。

Q：谁开发了这些项目的拼接软件？

A：拼接软件是由我们小组的 Andrey Zharkikh 开发的。拼接程序很独特，因为它在拼接两个不同的单倍体。它将显示出序列相似性的重叠群（contig）放在一起，同时，它又试图将它们分成 A、B 两个染色体。因此，当它看到序列差异或缺失时，它必须询问“这是我必须修正的错误吗？还是我要试图去理解的真正序列差异？”

有了这种杂合体的拼接策略，你能得到数百万个遗传标记物，非常棒。接着，你能使用它们中的 1000 或 2000 或 3000 的亚群，来进行互相定位。于是，你得到了海量的生物学信息。

Q：你计划将拼接软件与他人共享吗？

A：当然愿意。不过，我们只能克隆 Andrey 的。它不是一个真正的程序或产品，它是一系列脚本和代码片段。我们所能做的是将所有信息告诉 454，让他们在拼接程序中加入这段。将它变成产品需要巨量的工作。那真的超出了我们的范围。（生物通整理）

专访 Radoje Drmanac：5000 元测序的奥秘

生物通报道：来自 Complete Genomics 公司，华盛顿大学，哈佛医学院等处的研究人员在 11 月 6 日《Science》杂志上发表论文，描述其专利 DNA 测序平台，并公布了对三个完整人类基因组序列分析的结果。

文章的通讯作者是来自 Complete Genomics 公司（CG）Radoje Drmanac 博士和 Dennis G. Ballinger 博士，这家公司位于美国加利福利亚，是世界上首家提供大量人类基因组测序的服务机构。

为了进一步了解这一研究成果，生物通特采访了 Radoje Drmanac 博士，就一些读者感兴趣的问题请教了他。

生物通：第三代测序技术已经逐渐成为了一个热点，这篇 Science 文章给科学家们带来了许多惊喜，您能概况下这篇文章的主要内容吗？

Drmanac 博士：这是第一篇证明 Complete Genomics 公司（以下简称 CG 公司）基因组技术的同行评审文章，这篇文章成功完成了三个人类基因组的序列测定，也说明了人类基因组测序的成本可以低至 1726 美元（45x coverage），并且保持高度的精确性。

生物通：实验研究过程中主要采用的技术有哪些？最重要的技术点在哪里？

Drmanac 博士：CG 公司的专利平台技术采用了复合探针-锚定分子连接（Combinatorial probe anchor ligation，cPAL）化学试剂，以及预制基因组 DNA 纳米芯片，后者能提高成像效率，降低成本。具体而言：

1）预制（准确的框架）基因组 DNA 纳米芯片缩小了所需试剂容量，并且能加快成像——每次成像能捕捉到更多的 DNA 点；

2）新颖的试剂（复合探针-锚定分子连接）能对 70 个 DNA 碱基对进行单独序列阅读——每个碱基对的阅读都是完全独立的，这种序列阅读利用的是低浓度低成本试剂。

生物通：文章中提到的纳米阵列测序技术相对于现在流行的第二代测序，有何优势？

Drmanac 博士：新方法能在消耗更少试剂的前提下获得更多的数据——每台机器能获得 1000Gb 的数据，提高了一个数量级，也降低了整体的成本。

生物通：这项技术的个人基因组测序服务定价多少？预计在短期内，价格是否会显著下降？

Drmanac 博士：CG 公司目前还没有提供任何个人基因组测序的服务，一些小项目的价格是每个基因组 20000 美元（最少 8 个基因组），一些大型项目低至 5000 美元。

生物通：如何处理如此大量的数据？

Drmanac 博士：我们提供给客户的是测序的结果，而不是一堆原始数据，因此我们的客户从 CG 公司拿到的是好用的，可以用于研究的数据，其中包含有每个基因组序列差异的分类和注解列表。这种数据集比直接从测序系统中获得的原始数据更小，也更易于处理。

生物通：这种技术是如何兼顾低成本和速度，效率的?

Drmanac 博士：CG 公司已经证明这项技术在成本低的前提下，确保了高精确性——精确性高达

99.999%，和高通量，而且这一技术的生化基础也为未来测序质量和效率的急速提升奠定了良好的基础。

我们的这项技术能进行大规模，高精确性，低成本的完整人类基因组测序，因此对于成千上万的人类疾病，我们第一次可以进行大量患者的全面遗传学分析。

（生物通：王蕾）

附：

Complete Genomics

Corporate Vision

Many chronic and life-threatening diseases have a genetic basis, but current technology cannot analyze the human genome in a sufficiently complete or cost-effective manner to enable researchers to understand entire disease pathways. This incomplete understanding of the genetic interactions involved in disease limits healthcare outcomes by hindering the development of tailored drugs, diagnostics, and advanced disease prevention techniques.

Origin of an Idea

Complete Genomics was established in March 2006 by Dr. Clifford Reid, Dr. Radoje Drmanac, and Mr. John Curson, who shared a vision to provide high-throughput, affordable, and complete genome sequencing of human populations. Their goal was to enable commercial-scale research of the genetic mechanisms underlying drug responses and complex diseases, ensuring important advances in the diagnostic and therapeutic markets.

Flourishing Company

Complete Genomics sequenced its first genome in early 2009 and that data is publically available in the National Center for Biotechnology Information (NCBI) database. Already, in 2009, Complete Genomics has sequenced and delivered genomes to important collaborators in academic, pharmaceutical and government research institutions. In 2010, the company intends to sequence 10,000 genomes. Complete Genomics' mission is to become the global leader in human genome sequencing. It is currently building the world's largest human genome sequencing center in California. Further expansion is planned by opening sequencing centers worldwide.

Winning Strategy

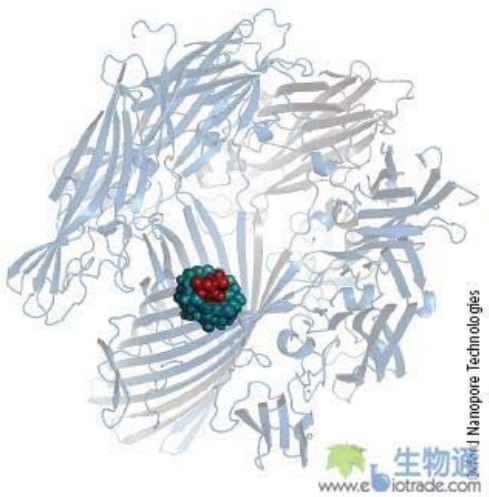
By offering low-cost, high-quality, complete DNA sequencing, Complete Genomics will power large-scale human genome studies that will enable great strides in our understanding of the genetic basis of disease. Pharmaceutical and biotechnology companies that had been previously priced out of the market will finally be able to access population-wide human genomic data for a wide variety of diagnostic and discovery applications. This exploration will provide new avenues for therapeutic and diagnostic discovery to benefit human health.

廉价的第三代纳米孔测序

最早的 Sanger 测序在人类基因组计划中立下赫赫战功，但也给基因组测序贴上了数亿美元的价格标签，让人生畏。这两年发展迅猛的第二代测序仪——Illumina 的 Genome Analyzer、Roche 454 的 GS 系列以及 ABI 的 SOLiD 系统——让人类基因组重测序的费用蹭地降低到 10 万美元以下。现在，能对单个 DNA 分子进行测序的第三代测序仪也加入到这场比赛中，让竞争更加激烈。

目前，第三代测序主要有三种技术平台。两种通过掺入并检测荧光标记的核苷酸，来实现单分子测序。Helicos 的遗传分析系统已上市，而 Pacific Biosciences 准备在明年推出单分子实时（SMRT）技术。第三种 Oxford Nanopore 的纳米孔（nanopore）测序还尚未有推出的时间表，但有可能是这三种当中最便宜的。纳米孔测序的优势在于它不需要对 DNA 进行标记，也就省去了昂贵的荧光试剂和 CCD 照相机。

最近，Oxford Nanopore Technologies 的 Hagan Bayley 及他的研究小组正致力于改善纳米孔。根据他们之前的工作，他们以 a-溶血素来设计纳米孔，并将环式糊精共价结合在孔的内侧（下图）。当核酸外切酶消化单链 DNA 后，单个碱基落入孔中，它们瞬间与环式糊精相互作用，并阻碍了穿过孔中的电流。每个碱基 ATGC 以及甲基胞嘧啶都有自己特有的电流振幅，因此很容易转化成 DNA 序列。每个碱基也有特有的平均停留时间，它的解离速率常数是电压依赖的，+180 mV 的电位能确保碱基从孔的另一侧离开。



a-溶血素纳米孔（剖面图）以及共价结合的环式糊精（浅蓝色）瞬间结合落入孔中的碱基（红色）。

以往对甲基胞嘧啶进行测序，都要先进行重亚硫酸盐转化，而纳米孔技术能直接读出这第五种碱基。这对表观基因组测序的研究人员来说可谓是个好消息。

纳米孔测序预计能满足大部分测序用户的需求：99.8%的准确性相当高，且错误很容易通过计算来纠正。均聚物延伸也没有问题，因为纳米孔记录每一个碱基，而不管其前后的碱基。读长也会很长。Bayley 认为：“它有可能读取数千个碱基，序列质量也不会下降。即使中途有一些小差错，它也可以重新开始。”

但是，Oxford Nanopore 的测序仪仍面临两个重要的技术问题。一是如何将核酸外切酶更好地附着在孔上，让它每次只掉入一个碱基，这是一个大挑战。另一个是并行化。这个问题可能简单一些。他们可以开发出一个芯片，上面有数万个孔，来确保整个测序过程更快速。

在纳米孔测序技术的推动下，实现千元基因组的目标指日可待了。

参考文献：

Clarke, J. et al. Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechnol. advance online publication 22 February, 2009.

第三代测序技术揭密

在第二代测序技术的协助下，个人基因组图谱正在如火如荼地绘制中。但第二代测序技术很快就遇上了强劲的对手——第三代测序技术，也被称为“下、下一代的测序（next-next-generation sequencing）”。第三代测序技术是基于纳米孔（nanopore）的单分子读取技术，有着更快的数据读取速度，应用潜能也势必超越测序。

2 月 5 日，基因组科学家们齐聚美国佛罗里达州的基因组生物和技术进展会议，来了解哪家公司的第三代测序技术能实现人类基因组的 3 分钟测序或以 5000 美元的价格出售。尽管科学家们对公布的数据表示谨慎乐观，但他们对于此类测序仪的优越之处仍心存疑虑。

Complete Genomics

在 2008 年 10 月，美国加利福尼亚州的 Complete Genomics 公司曾宣称他们将在 2009 年以 5000 美元的价格售卖人类基因组，但当时没有公布支持数据。在这次会议上，该公司公布了一个人类基因组，据称是用 9 台仪器在 8 天内完成的。

该公司的 CEO，Clifford Reid 表示，他们将 254GB 的数据拼接成草图，覆盖某个匿名男性基因组的 92%，每个碱基平均读取了 91 次。与目前应用中的高速测序，即第二代测序类似，Complete Genomics 也产生短的 DNA 读长。通过对每个碱基的多次测序，它的目标是排除悄悄混入的可能错误。Reid 认为这项技术非常准确，碱基错误的概率低于 0.33%。这与目前的测序仪相当。

Complete Genomics 并不出售测序仪，但用自己的测序仪来完成所有的内部工作。这让某些科学家质疑，但另一些却深受鼓舞。

速度和费用成为 Complete Genomics 的最大卖点。该公司并没有透露基因组测序的确切费用，但据称每个基因组的原材料费用低至 1000 美元。它的目标是在上个月推出市场，今年对 1000 个基因组进行测序，明年测序数量达到 20000 个。

Pacific Biosciences

在 Complete Genomics 做报告前的一小时，Pacific Biosciences 的首席技术官 Stephen Turner 展示了大肠杆菌的完整基因组，并称每个碱基的平均读取了 38 次，准确率大于 99.9999%。

Pacific Biosciences 利用了单分子技术和 DNA 聚合酶，在反应的同时读取测序产物。尽管目前仪器的读取速度仅为 3 碱基/秒，但它的目标是在 2013 年前实现三分钟读完人类基因组。它还有望实现更长的读长。Tuner 表示大肠杆菌基因组的平均读长是 586 bp，有些能达到 2805 bp。某些科学家期望长读长能排除错误，让他们了解到难以读取的部分。

Pacific Biosciences 打算在明年正式推向市场。同时，目前的测序技术，如 Illumina、Applied Biosystems 和 Roche 也正以惊人的速度制造数据，在单次几天的运行中产生相当于多个人类基因组的数据。速率不断增长的同时，费用也在下降。例如，Illumina 在会议中表示它在今年年底能实现 10000 美元的人类基因组测序。

Helicos Biosciences

并不是每家测序公司都这么幸运。Helicos Biosciences 制造的第三代测序仪就被测序错误所困扰。就在会议前几天，Helicos 透露它的第一名顾客已经将测序仪退还。在这次会议上，该公司表示它已经拼接了线虫的基因组。但是它的历史问题和高昂的仪器费用，即使降低至 99.9999 万美元，与其他测序仪的 50 万美元左右的价格相比，仍然让许多科学家望而却步。

在会议前的研讨会上，来自多伦多安大略癌症研究所的 John McPherson 说出了大家的心声："Helicos 是单分子测序的先锋，但我认为他们还没有达到预定的目标。"但 Helicos 的首席技术官 William Efcavitch 却不认同这种说法，"关于我们不行的传闻太夸张了"。

许多科学家希望他是对的，他们期待着 Helicos 与其他公司继续竞争，以更低的价格获得更多的数据。一位科学家表示："这种竞争是良性的。无论这些公司干得多好，我们都期望更多。"

（生物通 薄荷）

第三代单分子测序的开山之作

斯坦福大学的科学家最近利用 Helicos Biosciences 的 Heliscope 单分子测序仪，对一名白人男子的基因组进行了测序，文章发表在最新一期的《Nature Biotechnology》在线版上。

斯坦福大学的生物工程师 Stephen Quake，同时也是 Helicos 的创始人之一，对他本人的基因组进行了测序。在两名研究人员的协助下，他利用一台 Heliscope 测序仪和 4 次数据收集运行，完成了此次测序。

研究人员报告称，他们产生了数十亿个 Heliscope 序列读取，覆盖了 90%的人参考基因组，覆盖度达 28 倍。序列读长为 24 到 70 个碱基，平均读长为 32 个碱基。到目前为止，他们已经鉴定出 280 万个 SNP 和 752 个拷贝数变异。

这次测序花了 4 个星期的时间，试剂花费为 48000 美元。Quake 认为，这些工作在普通的实验室中就能完成，只需要一台仪器，费用也适中。另一名研究人员 Neff 表示，Heliscope 的主要优势在于高产量以及文库制备简单，不需要 DNA 扩增或连接。“如果有三台 Heliscope，我一个星期就能完成。”

Neff 解释道，因为有小部分核苷酸未标记上，在测序过程中出现了一些“暗色”的核苷酸，好像缺失一样。研究小组利用 Helicos 软件来协助碱基检出，并通过增加基因组覆盖度来校正缺失的错误。与参考基因组比对之后，他们发现读取片段覆盖了 25 亿个碱基，大约 90%。

为了帮助检出 SNP，另一位作者 Pushkarev 开发出一种 UMKA 算法。这个算法预测出基因组中的 2805471 个 SNP。其中大约 76%也在 dbSNP 中找到。他们还将鉴定出的 SNP 与 Illumina Human610-Quad SNP BeadArray 检测到的结果进行比较，发现两种方法的一致性为 99.8%。此外，研究人员还用 Sanger 测序验证了其中 100 个 SNP。他们还发现了基因组中的 752 个 CNV，其中超过半数出现在基因组变异数据库中。

作者们也提到这种测序方法仍有缺陷，包括基因组覆盖不完整，缺乏 SNP 和结构变异的完整数据。

Helicos 的副总裁 Patrice Milos 表示：“显然，这是 Steve Quake 的里程碑式文章，同时这也是 Helicos 的里程碑式文章。我们着实兴奋和激动。”他认为 Helicos 一直致力于开发容易运行的仪器，且前期的样品制备简单，这样小型实验室也能进行基因组测序。

下一步，研究小组还将与斯坦福干细胞生物学和再生医学研究院合作，利用 Heliscope 对癌症基因组进行测序。

原文检索：

Single-molecule sequencing of an individual human genome

Dmitry Pushkarev, Norma F Neff & Stephen R Quake

Nature Biotechnology Published online: 10 August 2009 | doi:10.1038/nbt.1561

摘要：

Recent advances in high-throughput DNA sequencing technologies have enabled order-of-magnitude improvements in both cost and throughput. Here we report the use of single-molecule methods to sequence an individual human genome. We aligned billions of 24- to 70-bp reads (32 bp average) to 90% of the National Center for Biotechnology Information (NCBI) reference genome, with 28 average coverage. Our results were obtained on one sequencing instrument by a single operator with four data collection runs. Single-molecule sequencing enabled analysis of human genomic information without the need for cloning, amplification or ligation. We determined 2.8 million single nucleotide polymorphisms (SNPs) with a false-positive rate of less than 1% as validated by Sanger sequencing and 99.8% concordance with SNP genotyping arrays. We identified 752 regions of copy number variation by analyzing coverage depth alone and validated 27 of these using digital PCR. This milestone should allow widespread application of genome sequencing to many aspects of genetics and human health, including personal genomics.

RNA 直接测序指日可待

Helicos BioSciences（第三代测序仪制造商）的研究人员近日发表了一篇原理验证（proof-of-principle，生物通注）研究，说明利用其单分子测序技术来进行 RNA 直接测序的可行性，文章发表在 9 月 23 日的《Nature》在线版上。

该研究小组利用一台 Helicos 样机，直接对酿酒酵母的 RNA 进行测序，而没有将其转变成 cDNA。在这个过程中，他们还发现了许多酿酒酵母转录本 3'端的异质性，同时有证据表明酵母中至少有一些核仁 RNA 和核糖体 RNA 是聚腺苷酸化的。

随着 RNA 的重要性逐渐被人所认识，研究人员也开发出多种方法来研究它。但是，许多方法仍需将 RNA 反转录成 cDNA，而这一步会引入错误，且效率不高。研究人员写道："人们急需一种方法，而这种方法不会有反转录、扩增、连接以及其他 cDNA 合成步骤的相关困难，而是能利用极少量的总 RNA 来综合、无偏见地浏览转录组。"

为了让"边合成边测序"的反应适应直接的 RNA 测序，Milos（Helicos 的首席科学家）及她的同事优化了一切，从使用的聚合酶到缓冲液再到专利的荧光核苷酸类似物。总的来说，这种方法在 poly(dT)包被的表面捕获聚腺苷酸化的 RNA，并利用大肠杆菌 poly(A)聚合酶 I 来进行边合成边测序的步骤。

该小组首先对一段 40 个碱基的 RNA 序列进行测序，来检验这种方法。大约一半（48.5%）的读长是 20 个核苷酸或更长。最长的无误差读数是 38 个碱基。

随后她们将注意力转向酿酒酵母这种模式生物。因为酿酒酵母的大部分 RNA 本身就存在聚腺苷酸化，所以她们不需要在测序前再加上 poly-A 尾巴。

在这个实验中，研究小组产生了 41261 个读数，每个约为 20 个核苷酸或更长。近一半（48.4%）的读数与酵母基因组配对。90%以上的配对读数是位于已知的酵母开放读码框 3'端的 400 个核苷酸内。

实验过程中，研究人员意外地检测到酵母 RNA 3'端的异质性。她们还发现证据，暗示至少有一些酵母核糖体 RNA 和小核仁 RNA（snoRNA）是聚腺苷酸化的。

研究人员称这种 RNA 直接测序方法目前的错误率约为 4%，大部分是黑暗碱基（dark base）引起的缺失。插入的错误率为 1-2%，而替换的概率只有 0.1-0.3%。

Milos 介绍说，她们还在进行相似的研究，为哺乳动物细胞的直接 RNA 测序开发方法。她补充道，尽管实验是在 Helicos 样机上完成的，但该公司正在开发 Heliscope 的直接 RNA 测序步骤。该步骤有望在明年推出。

上个月，Helicos 的创始人之一 Stephen Quake 在两名研究人员的协助下，利用一台 Heliscope测序仪和 4 次数据收集运行，对他本人的基因组进行了测序。详情请看：第三代单分子测序的开山之作。

原文检索：

Direct RNA sequencing

Nature advance online publication 23 September 2009 | doi:10.1038/nature08390