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。

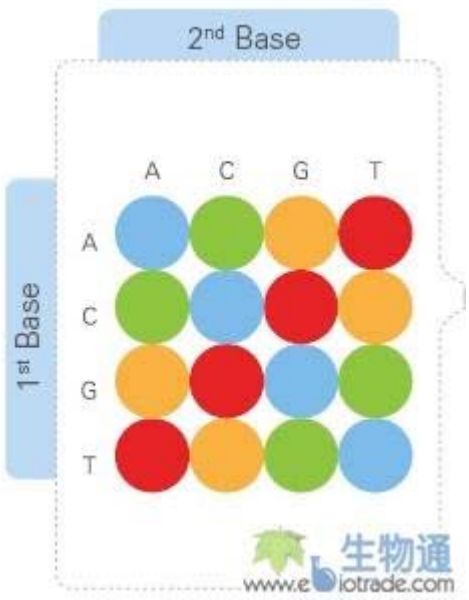
乳液 PCR 最大的特点是可以形成数目庞大的独立反应空间以进行 DNA 扩增。其关键技术是“注水到油”，基本过程是在 PCR 反应前，将包含 PCR 所有反应成分的水溶液注入到高速旋转的矿物油表面，水溶液瞬间形成无数个被矿物油包裹的小水滴。这些小水滴就构成了独立的 PCR 反应空间。理想状态下，每个小水滴只含一个 DNA 模板和一个 P1 磁珠，由于水相中的 P2 引物和磁珠表面的 P1 引物所介导的 PCR 反应，这个 DNA 模板的拷贝数量呈指数级增加，PCR 反应结束后，P1 磁珠表面就固定有拷贝数目巨大的同来源 DNA 模板扩增产物。

c. 微珠沉积

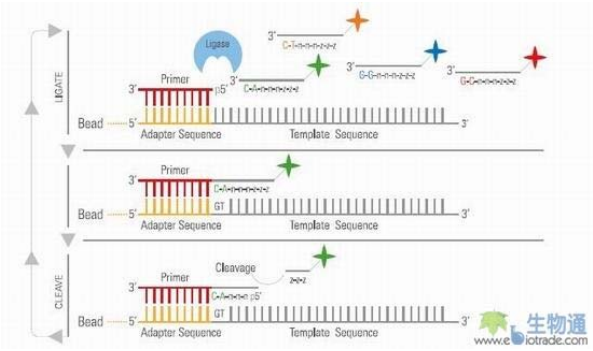
3’修饰的微珠沉积在一块玻片上。在微珠上样的过程中，沉积小室将每张玻片分成 1 个、4 个或 8 个测序区域。SOLiD 系统最大的优点就是每张玻片能容纳更高密度的微珠，在同一系统中轻松实现更高的通量。

d. 连接测序

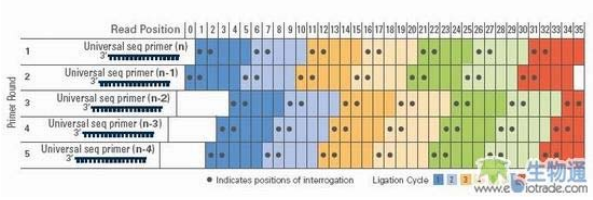
这一步可就是 SOLiD 的独门秘笈了。它的独特之处在于没有采用惯常的聚合酶，而用了连接酶。SOLiD 连接反应的底物是 8 碱基单链荧光探针混合物。连接反应中，这些探针按照碱基互补规则与单链 DNA 模板链配对。探针的 5’末端分别标记了 CY5、Texas Red、CY3、6-FAM 这 4 种颜色的荧光染料。探针 3’端 1～5 位为随机碱基，可以是 ATCG 四种碱基中的任何一种碱基，其中第 1、2 位构成的碱基对是表征探针染料类型的编码区，下图的双碱基编码矩阵规定了该编码区 16 种碱基对和 4 种探针颜色的对应关系，而 3～5 位的“n”表示随机碱基，6～8 位的“z”指的是可以和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