高通量测序技术及其应用*

王兴春[1,2**] 杨致荣[3] 王 敏[1] 李 玮[1] 李生才[2]
(1 山西农业大学生命科学学院 太谷 030801 2 山西农业大学农学院 太谷 030801)
(3 山西农业大学文理学院 太谷 030801)

摘要 高通量测序技术是DNA测序发展历程的一个里程碑,它为现代生命科学研究提供了前所未有的机遇。详细介绍了以454、Solexa和SOLiD为代表的第二代高通量测序技术,以HeliScope TIRM和Pacific Biosciences SMRT为代表的单分子测序技术,以及最近Life Science公司推出的Ion Personal Genome Machine(PGM)测序技术等高通量测序技术的最新进展。在此基础上,阐述了高通量测序技术在基因组测序、转录组测序、基因表达调控、转录因子结合位点的检测以及甲基化等研究领域的应用。最后,讨论了高通量测序技术在成本和后续数据分析等方面存在的问题及其未来的发展前景。

关键词 高通量测序 深度测序 下一代测序 基因组测序 转录组测序
中图分类号 Q3

作为最重要的分子生物学分析方法之一,DNA测序不仅为遗传信息的揭示和基因表达调控等基础生物学研究提供重要数据,而且在基因诊断和基因治疗等应用研究中也发挥着重要的作用。随着科学的发展,传统的Sanger测序技术的局限性日益突出。一方面,该技术依赖于毛细管电泳,不但费时,而且测序反应数也受到限制(目前常用的ABI 3730测序仪每次只能分析96个样品);另一方面,该技术基于酶法测序,成本较高。自2005年以来,以Roche公司的454技术、Illumina公司的Solexa技术和ABI公司的SOLiD技术为标志的高通量测序技术相继诞生。虽然高通量测序技术建立的时间不长,但发展非常快。已经应用于基因组,包括测序和表观基因组学以及功能基因组学研究的许多方面[1]。高通量测序技术已经将人们带到了真正的高通量测序时代,这些技术必将在生命科学领域得到广泛的应用,并产生极大的推动作用。然而,许多科研人员仅知道高通量测序技术的大概原理,而对于该技术在生命科学领域的应用却知之甚少。为此,本文详细阐述了高通量测序技术的最新进展及其在生命科学研究中的应用,以期使科研工作者尽早地了解该技术,并将其更好地应用到科研工作中。

收稿日期:2011-04-11 修回日期:2011-06-02
*山西省青年科技研究基金(2010021030-1)、国家自然科学基金(31100235)资助项目
**通讯作者,电子信箱:wxingchun@163.com; sxaulsc@126.com

1 高通量测序技术及其发展现状

高通量测序技术堪称测序技术发展历程的一个里程碑,该技术可以对数百万个DNA分子进行同时测序。这使得对一个物种的转录组和基因组进行细致全貌的分析成为可能,因此也称其为深度测序(deep sequencing)[2]或下一代测序技术(next generation sequencing, NGS)[3]。目前,所说的高通量测序技术主要是指454 Life Sciences公司、ABI公司和Illumina公司推出的第二代测序技术以及Helicos Heliscope™和Pacific Biosciences推出的单分子测序技术。

1.1 第二代测序技术

2005年,454 Life Sciences公司(现已被Roche公司收购)首先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,开创了第二代测序技术的先河。该技术的原理是酶级联化学发光反应:首先将PCR扩增的单链DNA与引物杂交,并与DNA聚合酶、ATP硫酸化酶、荧光素酶、三磷酸腺苷双磷酸酶、底物荧光素酶和5′-磷酸硫酸腺苷共同孵育。在每一轮测序反应中只加入一种dNTP,若该dNTP与模板配对,聚合酶就可

以将其掺入到引物链中并释放出等摩尔数的焦磷酸。焦磷酸盐被硫酸化酶转化为 ATP,ATP 就会促使氧合荧光素的合成并释放可见光。CCD 检测后通过软件转化为一个峰值,峰值与反应中掺入的核苷酸数目成正比[4]。此后,Illumina 公司和 ABI 公司相继推出了 Solexa[5] 和 SOLiD (supported oligo ligation detetion)[6] 测序技术。它们与焦磷酸测序法的原理类似,核心思想都是边合成边测序(sequencing by synthesis),即生成新 DNA 互补链时,要么加入的 dNTP 通过酶促级联反应催化底物激发出荧光,要么直接加入被荧光标记的 dNTP 或半简并引物,在合成或连接生成互补链时释放出荧光信号。通过捕获光信号并转化为一个测序峰值,获得互补链序列信息。由此可见第二代测序技术无需进行电泳,操作极为简便。这不但大大节省了成本和时间,而且可以在芯片上进行高通量分析。该测序技术单次反应可以对数以百万计的样品进行分析,这样庞大的测序能力是传统测序仪所不能比拟的。假如利用这种方法进行人类基因组的测序,那么在数天内就可以完成,而且成本也将大大降低。

Illumina 公司目前拥有三种测序平台,分别为 HiSeq 2000、HiSeq 1000、Genome Analyzer IIx (http://www.illumina.com/); ABI 公司则主要是 SOLiD 3 和 SOLiD 4 两个测序平台。HiSeq 2000 测序平台单次反应可以读取 200G 的数据,而 SOLiD 4 仅为 100G 左右。从通量这个最直观的数字看来,HiSeq 2000 领先于 SOLiD 4。更高的通量就意味着更低的成本,利用 HiSeq 2000 以 30 倍的覆盖度对两个人类基因组进行测序,每个基因组的费用不到 1 万美元(http://www.illumina.com/systems/hiseq_2000.ilmn)。就测序读长来说,454 测序平台读长最长,目前已经达到 400nt。因此,454 平台比较适合对未知基因组从头测序,但是在判断连续单碱基重复区时准确度不高。Solexa 测序读长较 454 短,仅为 100nt 左右,但测序通量高、价位低,适合基因组重测序等。SOLiD 读长也较短,但测序精度较高,特别适合 SNP 检测等。

1.2 单分子测序技术

在第二代测序平台不断完善和广泛应用的同时,以对单分子 DNA 进行非 PCR 测序为主要特征的更新的测序技术也初显端倪。2008 年 4 月,Helico BioScience 公司的 Harris 等[7] 在 *Science* 上报道了他们开发的基于全内反射显微镜(total Internal reflection microscopy, TIRM)的测序技术—单分子测序技术。该技术完全摒弃了上述测序平台所基于的 PCR 扩增的信号放大过程,真正达到了读取单分子荧光的能力。具体原理为:首先,将待测 DNA 样品随机打断成小片段,在每个小片段的末端加上 poly-dA;然后,将小片段 DNA 模板与固定在检测芯片上的 poly-dT 引物进行杂交并精确定位,并逐一加入荧光标记的末端终止子。在掺入了单个荧光标记的核苷酸后,洗涤、成像,之后切开荧光染料和抑制基团,洗涤、加帽,允许下一个核苷酸的掺入。这样通过掺入、检测和切除的反复循环,即可实时读取大量序列。随后,他们利用该技术对 M13 基因组进行重测序。M13 噬菌体是一种单链 DNA 病毒,基因组全长 6 407 个核苷酸。他们共获得了 28 万条序列,开创了单分子高通量测序的先河。这之后不久,Pacific Biosciences 公司又开发出了另一种单分子测序技术——单分子实时技术(single molecule real time, SMRT)[8]。该技术利用单分子技术和 DNA 聚合酶,在聚合反应的同时就可以读取测序产物。SMRT 测序技术在测序速度、读长和成本方面有着巨大的优势和潜力。虽然目前的读取速度为 3bp/s,但他们声称将在 2013 年前实现三 min 测完人类基因组的目标。从而又向 1 000 美元测定一个人类基因组的目标迈出了一大步。

1.3 Ion PGM 测序技术

与 Sanger 双脱氧链终止法测序技术相比,上述测序技术在测序速度和成本方面都有了革命性的变革。但是,测序仪的价格非常昂贵。例如,目前 Helico BioScience 公司的 HeliScope 测序仪售价近百万美元,这是一般实验室和科研单位所不能承受的。2010 年年底, Life Techologies 公司推出的 Ion Personal Genome Machine (PGM™) 测序仪价格仅为普通测序仪的 1/10,很好的解决了这一问题。Ion PGM 测序仪的设计是基于半导体芯片技术,在半导体芯片的微孔中固定 DNA 链,随后依次掺入 ACGT。随着每个碱基的掺入,释放出氢离子,在它们穿过每个孔底部时能被检测到。与其它新一代测序仪相比,它不需要激光、照相机或标记,价格当然要便宜很多。这种独特的流体体系、微体系机械设计和半导体技术的组合,使得研究人员能够在 2h 内获取从 10Mb 到 1Gb 以上高精确度序列(http://www.iontorrent.com/products-ion-pgm/)。

1.4 高通量测序技术的优点

高通量测序技术有三大优点是传统 Sanger 测序法所不具备的。第一,它利用芯片进行测序,可以在数百

万个点上同时阅读测序,把平行处理的思想用到极致,因此也称之为大规模平行测序(massively parallel signature sequencing, MPSS),这一点是大家所熟知的。第二,高通量测序技术有完美的定量功能,这是因为样品中某种 DNA 被测序的次数反映了样品中这种 DNA 的丰度。这一点有望取代以前的基因表达芯片技术用于基因表达的研究。第三,成本低廉。利用传统 Sanger 测序法完成的人类基因组计划总计耗资 27 亿美元,虽然比预计的 30 亿美元节省了不少,但仍是一个耗资巨大的工程。而现在利用高通量测序技术进行人类基因组测序,耗资不到传统测序法的 1%[9]。

2　高通量测序技术的应用

高通量测序技术的迅猛发展,将基因组学水平的研究带入了一个新的时期,也使经典分子生物科学家对基因组学的认识和思考上升到一个新的水平。高通量测序技术不仅可以进行大规模基因组测序,还可用于基因表达分析、非编码小分子 RNA 的鉴定、转录因子靶基因的筛选和 DNA 甲基化等相关研究。

2.1　高通量测序技术在全基因组测序中的应用

全基因组测序对全面了解一个物种的分子进化、基因组成和基因调控等有着非常重要的意义。但是,高通量测序技术在发展初期由于读长较短,使其在对未知基因组从头测序(de novo Sequencing)的应用受到限制,只能用于基因组重测序。基因组重测序是指对已知基因组序列的物种进行不同个体的基因组测序,并在此基础上对个体或群体进行差异性分析。2008 年 4 月 17 日的 *Nature* 杂志上,美国的科学家发表了首个利用新一代高通量测序技术得到的人类全基因组,这个基因组正是“DNA 之父”James D. Watson 的[9]。整个测序过程在 2 个月内就完成了,花费不到 100 万美元。

随着高通量测序技术的不断完善,独立应用该技术进行全基因组从头测序成为可能。2010 年 1 月 21 日,由深圳华大基因研究院发起,中国科学院昆明动物研究所、中国科学院动物研究所、成都大熊猫繁育研究基地和中国保护大熊猫研究中心参与的大熊猫基因组测序成果以封面文章的形式发表在国际权威杂志 *Nature* 上[10]。这是全球第一个完全使用高通量测序技术完成的基因组序列图。

2.2　高通量 RNA 测序及其在转录组和基因表达调控研究中的应用

转录组研究是基因功能及结构研究的基础和出发点,是全基因组测序完成后首先要面对的问题。最近科学家们将高通量测序技术应用于转录组分析开发出了 RNA 测序技术(RNA-Seq),该技术能够在全基因组范围内检测基因表达情况,进行差异基因筛选分析。由于 RNA-Seq 技术具有通量高、可重复性高、检测范围宽、定量准等特点,已经广泛应用于细菌、拟南芥、水稻和人类等生物转录组的研究。我国在应用 RNA-Seq 进行水稻转录谱分析方面走在了世界的前列。王俊和韩斌领导的研究小组利用 RNA-Seq 技术分别对水稻的转录组进行高分辨率的分析,发现水稻具有可变剪切模式基因数目远远高于预期[11-12]。

转录组研究是从整体水平研究基因功能以及基因结构。但是,多数研究人员感兴趣的是某一特定的生物过程、发育阶段或处理后的基因表达情况。基因芯片技术曾在该领域的研究中发挥了重要的作用,但它只能检测已知序列的特征,对于未知的序列无能为力[13]。而建立在高通量测序基础上的数字化基因表达谱(digital gene expression profiling)分析无需预先针对已知序列设计探针,即可对任何生物整体转录活动进行检测,因此应用范围更加广泛。最近,Eveland 等[14]成功地将数字基因表达谱应用于玉米雌穗发育的研究,并发现了一批调控花序结构的基因。

高通量 RNA 测序技术另一个广泛应用的领域是小 RNA 的研究。小 RNA 在植物的生长、发育和外界胁迫应答等方面具有重要功能。但由于小 RNA 序列短、同源性高,因此利用基因芯片检测小 RNA 非常困难。高通量测序技术不但能够克服这一难题,而且能够发现新的小 RNA。目前,高通量测序技术已经成功的应用于拟南芥[15]、水稻[16]和小麦[17]等生物小 RNA 的研究。

2.3　ChIP-Seq 技术及其在 DNA 和蛋白质相互作用研究中的应用

染色质免疫共沉淀(chromatin immunoprecipitation, ChIP)技术是研究体内蛋白质与 DNA 之间相互作用的强有力工具,在转录因子结合位点或组蛋白特异性修饰位点的研究中被广泛应用[18]。最近诞生的染色质免疫共沉淀测序(chromatin immunoprecipitation sequencing, ChIP-Seq)技术充分结合了 ChIP 和高通量测序技术的优势,能够在全基因组范围内高效地研究

目的蛋白的结合位点。该技术的原理是:首先通过ChIP技术利用抗体特异性地富集交联的目的蛋白-DNA复合体;然后再将得到的DNA片段进行高通量测序;最后将获得的数百万条序列标签精确定位到基因组上,从而获得全基因组范围内与组蛋白或转录因子等互作的DNA区段信息。该方法克服了以往染色质免疫共沉淀-芯片(chromatin immunoprecipitation chip, ChIP-chip)技术依赖于实验者选择探针的不足,可以为任何生物测序[19]。2007年,应用ChIP-Seq技术获得研究成果分别发表在*Science*[20]、*Nature*[21]和*Cell*[22]三大顶级刊物上。我们在体细胞胚胎发生的研究中发现了一个PGA37基因,该基因编码一个R2R3-MYB转录因子,在植物体细胞胚胎发生和脂肪酸生物合成过程中起着重要的调控作用[23]。分离并鉴定PGA37的靶基因是阐明该基因生物学功能的关键。为此,我们正与有关测序公司洽谈,利用ChIP-Seq技术来筛选PGA37的靶基因。

2.4 高通量测序技术在基因组DNA甲基化分析中的应用

DNA甲基化在维持细胞正常功能、遗传印记和胚胎发育过程中起着极其重要的作用。新一代高通量测序技术使得基因组整体水平高精度的甲基化检测成为现实。目前,已经建立了至少三种依赖于高通量测序的DNA甲基化分析技术:甲基化DNA免疫共沉淀测序(methylated DNA immunoprecipitation sequencing, MeDIP-Seq)[24]、甲基结合蛋白测序(methyl-binding protein sequencing, MBD-Seq)和亚硫酸氢盐测序(bisulfite-sequencing, BS-Seq)[25]。它们之间的区别在于前两种方法首先通过特异性结合甲基化DNA的MBD2b或5′-甲基胞嘧啶抗体富集高甲基化的DNA,然后再对富集到的片段测序,而第三种方法不经过富集直接测序。虽然MeDIP-Seq和MBD-Seq都是基于富集的原理,但二者是相辅相成的策略。MeDIP-Seq对高度甲基化和高密度的CpG更加敏感,而MDB-Seq对高度甲基化和中等CpG密度更加敏感[26]。目前,高通量测序技术已经广泛应用于拟南芥[25]、水稻[27]、蚕[28]和人[26]等生物DNA甲基化的研究,并取得了丰硕的成果。

3 高通量测序技术存在的问题及其发展趋势

尽管高通量测序技术有诸多的优势,但其局限性也不容忽视。第一,测序速度提高了,但后续的海量测序数据的分析却成为一大难题[29]。第二,高通量测序技术不适合小规模测序。虽然高通量测序技术的价格不断降低,但一次反应仍需几千至数万元的花费。这对于PCR产物和质粒等数十到数千个碱基的测序,一般客户将是很难接受的,这时传统Sanger测序法无疑还是最佳的选择。最后,新一代测序仪价格昂贵,动辄几百万元,一般的小型实验室难以承受。因此,传统Sanger将与高通量测序技术长期并存,在短期内还不会被淘汰[30]。另外,高通量测序技术只是研究的开端,现在我们所能解释的生物学现象和机制还很有限,即使获得了基因组信息,如何去解释和应用它,仍是一个长远的问题。

测序技术的发展日新月异,本文所述的焦磷酸测序等测序平台都依赖生物化学反应。这不但加大了测序成本,而且浪费了时间,不利于测序速度的提升。因此,非光学显微镜成像、纳米孔和生物孔等直接测序技术是未来的发展方向[30]。相信随着高通量测序成本的进一步降低和对海量数据处理能力的不断提高,高通量测序将成为一项常规的实验手段,并为生物学和生物医学研究领域带来革命性的变革。

参考文献

[1] Mardis E R. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet, 2008, 9:387-402.

[2] Sultan M, Schulz M H, Richard H, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 2008, 321(5891):956-960.

[3] Schuster S C. Next-generation sequencing transforms today's biology. Nat Methods, 2008, 5(1):16-18.

[4] Margulies M, Egholm M, Altman W E, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature, 2005, 437(7057):376-380.

[5] Porreca G J, Zhang K, Li J B, et al. Multiplex amplification of large sets of human exons. Nat Methods, 2007, 4(11):931-936.

[6] Ondov B D, Varadarajan A, Passalacqua K D, et al. Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. Bioinformatics, 2008, 24(23):2776-2777.

[7] Harris T D, Buzby PR, Babcock H, et al. Single-molecule DNA sequencing of a viral genome. Science, 2008, 320(5872):106-109.

[8] Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. Science, 2009, 323(5910):133-

138.

[9] Wheeler D A, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature, 2008, 452(7189):872-876.

[10] Li R, Fan W, Tian G, et al. The sequence and de novo assembly of the giant panda genome. Nature, 2010, 463(7279):311-317.

[11] Lu T T, Lu G J, Fan D L, et al. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res, 2010, 20(9):1238-1249.

[12] Zhang G J, Guo G W, Hu X D, et al. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res, 2010, 20(5):646-654.

[13] 滕晓坤,肖华胜. 基因芯片与高通量 DNA 测序技术前景分析. 中国科学 C 辑:生命科学, 2008, 38(10):891-899.
Teng X K, Xiao H S. Perspectives of DNA microarry and next-generation DNA sequencing technologies. Science in China Series C-Life Science, 2008, 38(10):891-899.

[14] Eveland A L, Satoh-Nagasawa N, Goldshmidt A, et al. Digital gene expression signatures for maize development. Plant Physiol, 2010, 154(3):1024-1039.

[15] Lu C, Tej S S, Luo S, et al. Elucidation of the small RNA component of the transcriptome. Science, 2005, 309(5740):1567-1569.

[16] Sunkar R, Zhou X, Zheng Y, et al. Identification of novel and candidate miRNAs in rice by high throughput sequencing. BMC Plant Biol, 2008, 8:25-41.

[17] Yao Y, Guo G, Ni Z, et al. Cloning and characterization of microRNAs from wheat (Triticum aestivum L.). Genome Biol, 2007, 8(6):96-108.

[18] Solomon M J, Larsen PL, Varshavsky A. Mapping protein-DNA interactions in vivo with formaldehyde:evidence that histone H4 is retained on a highly transcribed gene. Cell, 1988, 53(6):937-947.

[19] Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet, 2009, 10(10):669-680.

[20] Johnson D S, Mortazavi A, Myers R M, et al. Genome-wide mapping of in vivo protein-DNA interactions. Science, 2007, 316(5830):1497-1502.

[21] Mikkelsen T S, Ku M, Jaffe D B, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature, 2007, 448(7153):553-560.

[22] Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. Cell, 2007, 129(4):823-837.

[23] Wang X, Niu Q W, Teng C, et al. Overexpression of PGA37/MYB118 and MYB115 promotes vegetative-to-embryonic transition in Arabidopsis. Cell Res, 2009, 19(2):224-235.

[24] Down T A, Rakyan V K, Turner D J, et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol, 2008, 26(7):779-785.

[25] Cokus S J, Feng S, Zhang X, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature, 2008, 452(7184):215-219.

[26] Li N, Ye M, Li Y, et al. Whole genome DNA methylation analysis based on high throughput sequencing technology. Methods, 2010, 52(3):203-212.

[27] Yan H H, Kikuchi S, Neumann P, et al. Genome-wide mapping of cytosine methylation revealed dynamic DNA methylation patterns associated with genes and centromeres in rice. Plant J, 2010, 63(3):353-365.

[28] Xiang H, Zhu J, Chen Q, et al. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat Biotechnol, 2010, 28(5):516-520.

[29] Gilad Y, Pritchard J K, Thornton K. Characterizing natural variation using next-generation sequencing technologies. Trends in Genetics, 2009, 25(10):463-471.

[30] Zhou X, Ren L, Li Y et al. The next-generation sequencing technology:a technology review and future perspective. Sci China Life Sci, 2010, 53(1):44-57

High-throughput Sequencing Technology and Its Application

WANG Xing-chun[1,2] YANG Zhi-rong[3] WANG Min[1] LI Wei[1] LI Sheng-cai[2]

(1 College of Life Sciences, Shanxi Agricultural University, Taigu 030801, China)

(2 College of Agriculture, Shanxi Agricultural University, Taigu 030801, China)

(3 College of Arts and Sciences, Shanxi Agricultural University, Taigu 030801, China)

Abstract As a milestone in the development of DNA sequencing, high-throughput sequencing technology

provides an unprecedented opportunity for the modern life sciences. The recent progress on this technology, including the second generation sequencing technology (represented by 454, Solexa and SOLiD), the third generation sequencing technology (represented by HeliScope TIRM and Pacific Biosciences SMART) and the Ion Personal Genome Machine sequencing technology are summarized. Then, the application of the high -throughput sequencing technology in genome sequencing, transcriptome sequencing, gene expression regulation, detection of binding locations for transcription factors and methylation analysis are summarized. Finally, the disadvantages and the prospects of this technology were discussed.

Key words High-throughput sequencing Deep sequencing Next generation sequencing Genome Sequence Transcriptome sequence