

新一代测序技术及其在植物转录组研究中的应用

李智奕^{1,2}, 宁 维^{1,2}, 陈利平², 李 瑜¹, 韩亚伟^{2*}, 史媛媛¹

(1. 河南农业大学 食品科学技术学院, 河南 郑州 450002; 2. 郑州轻工业学院, 河南 郑州 450002)

摘要: 新一代高通量测序技术发展迅速, 其特点是高通量、低成本, 成为更多研究人员进行转录组分析的重要手段。新一代测序技术主要有 454 测序平台、Solexa 测序平台和 SOLID 测序平台, 目前被广泛应用于生物学研究中。就目前正在发展的 3 种高通量测序技术进行了阐述, 并比较其优缺点, 探讨了新一代测序技术在植物转录组研究中的应用, 并对其发展前景进行了展望。

关键词: 新一代测序技术; 转录组; SNP; EST-SSR

中图分类号: Q52 **文献标志码:** A **文章编号:** 1004-3268(2013)12-0001-05

The Next Generation Sequencing Technology and Its Application in Plant Transcriptome

LI Zhi-yi^{1,2}, NING Wei^{1,2}, CHEN Li-ping², LI Yu¹, HAN Ya-wei^{2*}, SHI Yuan-yuan¹

(1. College of Food Science and Technology, Henan Agricultural University, Zhengzhou 450002, China;

2. Zhengzhou University of Light Industry, Zhengzhou 450002, China)

Abstract: Recently, the next generation sequencing (NGS) technology develops rapidly, which is characterized as high-throughput and low-cost advances and considered by many researchers as an important approach in transcriptome analysis. The main representatives of NGS involve 454 sequencing platform, Solexa sequencing platform and SOLID sequencing platform, which are widely used in biological studies. In the present paper, we reviewed the three high-throughput sequencing technologies in details, compared their advantages and disadvantages, discussed their applications in plant transcriptome, and outlooked their future development.

Key words: next generation sequencing technology; transcriptome; SNP; EST-SSR

20 世纪 70 年代, 以 Sanger 法为代表的第一代测序技术成为现代分子生物学中重要的测序技术^[1]。随着科技进步, 第一代测序技术成本高、速度慢、通量低等缺点凸显, 从而产生了新一代高通量测序技术。新一代测序技术以速度快、测序长、通量高、准确率高为主要优势, 被广泛应用于重复序列较少的基因组测序中(如微生物)。目前, 其在含有大量重复序列的真核生物基因组测序方面也有明显的优势^[2]。在植物转录组研究方面, 被用于新基因的发现、SNP 及分子标记的挖掘、基因家族鉴定及进化分析、转录图谱绘制、代谢途径确定等方面, 为研究基因表达模式、鉴别差异基因、发展分子标记提供

海量数据。文中就新一代高通量测序技术及其在植物转录组研究中的应用进行了综述。

1 新一代测序系统的发展

新一代高通量测序技术一次可对几十万条甚至几百万条 DNA 序列进行测序, 被称为大规模平行测序(massively parallel sequencing); 相比传统测序技术, 无论在通量及效率上都有质的提高, 也被称为下一代测序技术(next generation sequencing); 高通量测序使得对一个物种的转录组和基因组进行全面细致的分析成为可能, 又被称作深度测序(deep sequencing)。如今新一代测序平台已成为主流的

收稿日期: 2013-07-11

基金项目: 国家自然科学基金项目(31112175)

作者简介: 李智奕(1985-), 女, 湖北武汉人, 在读硕士研究生, 研究方向: 烟草转录组。

* 通讯作者: 韩亚伟(1973-), 男, 河南商丘人, 副教授, 博士, 主要从事分子生物学研究。E-mail: ywhan@zzuli.edu.cn

测序技术,有助于研究人员以低廉的价格全面深入地分析基因组、转录组及蛋白质组的各项数据。这将成为一项广泛使用的试验手段,有望给生物学和生物医学研究领域带来革命性的新技术^[3]。

1.1 454 测序平台

新一代测序最早由 454 公司开创,2005 年底 454 公司推出了基于焦磷酸测序(pyrosequencing)的高通量基因组测序系统 Genome Sequencer 20 System。在 454 公司被罗氏诊断公司收购后,推出了性能更优的 Genome Sequencer FLX System(GS FLX)。2008 年又推出了全新的测序试剂 GS FLX Titanium,全面提升了测序的准确性、读长和测序通量。

454 测序仪测序流程如下^[4-5]:(1)将基因组用物理的方法打成较短的 DNA 片段,再将片段两端与接头相连使其容易与磁珠结合;(2)进行乳液 PCR(emulsion PCR),在 PCR 扩增过程中每个片段进行独立扩增;(3)扩增反应完成后,片段和磁珠一起加入到 454 公司发明的 PTP(Pico Titer Plate)板中;(4)在 PTP 板样品孔内有酶类及引物和未经荧光标记的核苷酸进行互补新链的合成;每个核苷酸连接到新链上时,就会释放一分子的焦磷酸盐(PPi),在酶的作用下,经过合成反应和化学发光反应,将荧光素氧化成氧化荧光素,发出可见光,被检测仪器捕获形成峰图。

1.2 Solexa 测序平台

Solexa 公司研发的第二代测序仪,核心专利技术是“DNA 簇(DNA cluster)”和“可逆性末端终结(reversible terminator)”,实现样本自动化制备和大规模并行测序。2007 年 Illumina 公司收购 Solexa 公司。2010 年初,Illumina 将 Genome Analyzer IIx 升级到 HiSeq 2000。

Solexa 测序基本原理如下^[6-8]:(1)将样品 DNA 随机打断成几百个碱基或更短的片段,并在片段的

两头末端加上接头;(2)每个带有接头的 DNA 单链随机结合引物进行 PCR 扩增,通过扩增和变性后得到的单链,其一端连接在芯片上,另一端随机和引物互补连接形成“桥”;(3)形成的单链桥以周围的引物为扩增引物形成双链,双链经变性成单链,再次形成单链桥,进行下一轮扩增;经过 30 轮扩增,每个单分子得到了 1 000 倍扩增,成为单克隆“DNA 簇”;(4)将 DNA 簇进行测序,受激光的激发,可通过标记荧光识别不同核苷酸,读取该次反应颜色后继续下一轮反应,如此反复得到精确的序列片段。

1.3 SOLID 测序平台

美国 ABI 公司在 2007 年底推出了 SOLID 第二代测序平台。2010 年发布最新的 SOLID 5500xl 测序平台。SOLID 以 4 色荧光标记寡核苷酸的连续连接反应为基础,该反应不会出现 DNA 聚合酶合成过程中常有的错配问题。

与其他新一代测序仪不同的是,SOLID 不利用 DNA 聚合酶在合成过程中读取序列,而是利用 DNA 连接酶在连接过程读取序列。SOLID 测序的基本流程为^[9-11]:(1)测序文库的构建;(2)乳液 PCR,(1)、(2)两步骤与 GS FLX 系统相似;(3)连接反应的底物是 8 碱基单链荧光探针,探针的 5'端标记有荧光,3'端 1~2 位碱基与 5'端荧光信号的颜色对应。每次 SOLID 测序共有 5 轮:第 1 轮第 1 次连接反应掺入 1 条探针,测序仪记录探针 3'端 1~2 位碱基的荧光信号,然后除去 6~8 位碱基及 5'端荧光基团,获得 1~2 位颜色信息,第 2 次连接反应得到 6~7 位颜色信息,第 3 次连接反应得到 11~12 位颜色信息;多次连接后,开始第 2 轮测序,使用比第 1 轮少 1 个碱基的引物进行反应,经过 5 轮测序后,得到所有位置的颜色信息从而推断出相应的碱基序列。

1.4 各测序平台优缺点

各测序平台优缺点见表 1。

表 1 各测序平台的比较

项目	454/GS FLX	Solexa/HiSeq 2000	SOLID/SOLID 5500 xl
所属公司	Roche	Illumina	ABI
测序方法	焦磷酸测序法	可逆链终止物和合成测序法	连接测序法
检测方法	光学	荧光/光学	荧光/光学
数据量/(Gb/轮)	0.5	54~60	100
大约读长/bp	300~400	100	30~50
准确率/%	≥99	≥98~99	≥99.9
运行时间	10 h	14 d	8 d
成本/(美元/Mb)	40	2	2
优点	在第二代测序中读长最长;运行速度快	很高测序通量;性价比高	准确率最高,所要拼接出基因组的试剂成本最低
相对局限性	样品制备较难;同源重复序列出错率高;费用高	试剂花费高;用于数据删节和分析的费用很高	测序运行时间长;读长短,数据分析和基因组拼接困难

2 新一代测序平台在植物转录组研究中的应用

转录组测序(RNA sequencing, RNA-Seq)是指利用新一代测序技术对组织或细胞中所有的 mRNA 反转录成的 cDNA 进行测序,该技术能对各物种转录本的结构和表达水平进行分析,同时还能发现未知转录本和稀有转录本,为研究人员提供更为全面和精确的转录组信息;与传统基因芯片相比, RNA-Seq 最大的优势是不需要设计探针,即可对任意物种的整体转录活动进行测序^[12-13]。

2.1 新基因、新转录本的发现

转录组测序能在基因组序列未知的情况下进行,这对研究人员了解许多非模式生物非常有利,如果研究对象尚未完成基因组测序,可采用读段从头拼装(de novo assembly)的方法^[14]。将获得的转录组数据与数据库中已知序列进行比对,研究人员可以发现新基因或新转录本。

Wall 等^[15]通过比较传统毛细管电泳法测序和高通量测序分析拟南芥(*Arabidopsis*)转录组,定位 130 000 个 cDNA 序列读段,经过 de novo assembly 标记 15 000 个基因,包括新剪接变异体和非编码区;其中 1 117 个读段为内含子,3 066 个读段为非编码区,有 12 447 个读段没有出现在 BLASTn 中;试验表明,通过高通量测序得到的标准文库序列标签基因要多于非标准文库,利用高通量测序可以产生相当大数量的基因读段,完成近乎完整的基因组序列测定,这对模式和非模式生物都适用。

Garg 等^[16]利用 Illumina 公司基因组分析平台对鹰嘴豆转录组测序分析,获得 134 954 354 个读段,经过剪切、装配等修饰得到 1 070 000 个高质量读段,再使用短序列拼装工具得到非冗余(non-redundant)转录本 53 409 个;该试验使用短序列基因组装来发掘无参考基因组的大规模基因,通过转录组测序方法获得百万序列,第一次为鹰嘴豆提供了完整的转录组数据,为新基因发现和功能性分子标记的开发提供了资源,也将有助于其他类似的转录组研究。

2.2 SNP 的发现

单核苷酸多态性(single nucleotide polymorphisms, SNP)是在基因组水平上由单个核苷酸的变异而产生的一种 DNA 序列多态性,是人类可遗传变异中最常见的一种,占有已知多态性的 90% 以上。SNP 作为第三代遗传标记,现普遍用于高危群体的发现、相关疾病基因的鉴定、药物设计和测试以

及生物学的基础研究等^[17]。

Novaes 等^[18]比较 Sanger 方法和 454 方法,对巨桉(*E. grandis*)进行分析,在该研究中,通过新一代测序法共获得 1 024 251 个读段,148 Mb 表达序列标签,发现其中一半基因与拟南芥同源,新一代测序方法获得的读段平均长度是一代测序方法的 1/3,而数量则是其 4~5 倍,从而使得 ESTs 数量最大化。经过拼接注释后,获得 23 742 个 SNP,平均每 192 bp 发现 1 个 SNP 位点。其中有 1 个样品中鉴定出 279 个 SNP 位点。

Hansey 等^[19]对 21 个近交品种玉米幼苗进行高通量测序转录组分析,显示 21 个品种中最少的变异数在 57.1%~66.0%。通过 SNP 检查发现,玉米 22 830 个已注释的基因里含有 351 710 个 SNP,其中 329 017 个 SNP 已通过已知基因对其进行注释。6 个国内自交品种中含有 468 966 个 SNP,每 207 bp 中有 1 个 SNP。

2.3 代谢通路的分析

在生物体内,各类酶催化不同生理生化反应。根据转录组测序结果,可知不同生理时期及不同组织内各种酶含量的多少,据此建立相关物质的代谢途径。利用新一代测序可大规模地对组织样本进行分析,有利于建立特定条件下的物质代谢途径。

Feng 等^[20]利用高通量测序平台对杨梅果实在成熟发育期进行分析,得到 41 239 个 unigenes,发现 3 600 个基因差异表达,其中 826 个上调,1 407 个下调,涉及花青素合成的基因全部上调;进行了 GO、KEGG、COG 分析,重点分析了碳水化合物和酸的代谢通路,发现与 75DAF 杨梅果实相比,85DAF 的果实甜度升高、酸度降低,在测序结果中确定有 26 个 unigenes 参与糖代谢,有 5 个编码蔗糖磷酸合成酶的 unigenes 在成熟过程中快速上调,参与转化的 unigenes 中 1 个下调、2 个上调,参与柠檬酸循环的基因无明显变化。

Barrero 等^[21]对狼毒大戟(*Euphorbia fischeriana*)根部进行高通量测序分析,获得 18 180 个转录本。通过 GO 分析,涉及 7 841 个转录本,其中 23.2% 与代谢相关,13.4% 和应激反应相关,其中与蛋白磷酸化过程有关的转录本为 795 个。这些代谢活动都发生在狼毒大戟根部,蛋白质可逆磷酸化与调节生长素信号有关,是植物生长发育所必需的。有 257 个和 370 个转录本分别参与镉响应和防御反应,防御反应中有 191 个转录本与抑菌相关。利用 KEGG 分析得到 3 189 个转录本,涉及 293 个代谢通路,主要涉及碳水化合物代谢途径、能源和脂质代

谢、氨基酸代谢、次级代谢产物合成等。研究还表明,狼毒大戟根部存在活性代谢过程以及各种代谢产物合成,与萜类生物合成途径有关的转录本涉及最多的是次级代谢产物合成。

2.4 EST-SSR 标记的开发

SSR 是目前最常用的微卫星标记之一。微卫星因在染色体上分布均匀、重复性好、变异度高、多态性高、数量丰富、所需 DNA 量少等优点被迅速广泛地应用于生命科学的各领域中。EST(expressed sequence tag)是通过从 cDNA 文库中随机挑选克隆,进行一轮单向测序所获得的序列,通常为几十至 500 bp 左右。由于 EST 的获得快速、简便、廉价,现已成为基因组学研究领域的主要内容^[22]。利用 EST 资源,可开发基于 ESTs 的 SSR 标记,并且由于可以直接获得基因表达的信息,节省了 SSR 引物开发过程中的克隆和测序步骤,降低了引物开发成本。ESTs 数量的迅速增加为开发新的 SSR 标记提供了宝贵资源^[23]。

Wei 等^[24]使用 Illumina paired-end 测序方法分析芝麻开花期的 5 个组织部分:幼根、叶、花、发育中的种子、根尖;对整理过的原始数据组装后的 86 222 个 unigenes 进行了基因功能注释;通过 BLASTx 分析发现,44 750 个 unigenes 与拟南芥高度同源;开发了 7 702 个 EST-SSR 标记,其中双核苷酸 SSR 标记共有 5 166 个,其中 AG/CT 是主要的重复序列;随机选取 50 个 SSR 分析了芝麻的多态性,成功使用 40 对引物扩增片段并且分析了 24 个芝麻品种之间的多态性。高通量测序在非模式生物的基因和分子标记的发现方面具有快速、经济的特点,同时该研究结果为芝麻的序列研究提供了全面的资源。

Kaur 等^[25]使用 454 平台分析了扁豆的组织特异性;样品来自不同的扁豆品种,共产生 1.38×10^6 个 ESTs;将扁豆完整的 unigene 库与模式生物苜蓿和拟南芥对比分析,分别确定 12 639 个和 7 476 个 unigenes;随后又有 25 592 个 unigenes 在 GenBank 里被确认;从已知的 EST-SSR 标记中,设计 2 393 对引物,有 412 个 contig 至少包含 2 个 SSR 位点,而三核苷酸位点占 60.6%;有 192 个 EST-SSR 位点经过筛选验证被确认,其中 51 个显示出 12 个作物品种和 1 个野生品种之间的多态性;有 166 对引物成功扩增,其中 47.5%检测到遗传多态性。

3 前景与展望

转录水平调控是生物体最主要的调控方式,转录组测序研究正逐步取代传统基因芯片技术成为基

因研究的主要方法;对某样品深度测序可以获得低表达的基因,而对不同样品同时测序可以明确样品之间的表达差异;另外,研究人员还可获得转录本表达丰度、转录发生位点、可变剪切、SNP 等重要信息^[29]。

第一代测序平台远没有二代测序平台通量高,但在读取长序列以及原始数据准确方面具有优势;更长的拷贝意味着可在一次测序中测得更长的重复片段,定位更多的外显子,目前,在新一代测序平台中最长的拷贝长度仅为 400 bp。随着新一代测序技术的发展更多问题都有望得到解决,相信改进后的测序系统可提供更高质量的数据和低成本的服务。测序技术研究虽然取得不俗的成绩,但海量数据的产出提出了更大的挑战,例如,怎样充分挖掘隐藏在原始数据中的生物学意义,如何对数据进行高效分类、存档等。随着科学技术的发展,各种新型计算方法的出现将为数据的存储以及交流提供有利条件^[26]。

在植物学研究中,高通量测序技术应用于全基因组测序中能够极大推动更多的非模式植物、农作物和经济作物的全基因组测序工作;现在越来越多的物种基因信息被陆续公布,利于深度发掘新基因、SNP 以及各种分子标记等基因资源,使研究人员逐渐了解非模式生物中的遗传信息^[27]。因此,新一代测序系统在植物转录组分析方面有着广阔的前景。

参考文献:

- [1] Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors[J]. Proceedings of the National Acad Sciences of the United States of America, 1977, 74(12): 5463-5467.
- [2] Wicker T, Schlagenhauf E, Graner A, et al. 454 sequencing put to the test using the complex genome of barley[J]. BMC Genomics, 2006, 7: 275-285.
- [3] Shendure J, Ji H. Next-generation DNA sequencing[J]. Nature Biotechnology, 2008, 26(10): 1135-1145.
- [4] Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate[J]. Science, 1998, 281: 363-365.
- [5] Margulies M, Egholm M, Altman W, et al. Genome sequencing in microfabricated high-density picolitre reactors[J]. Nature, 2005, 437: 376-380.
- [6] Fedurco M, Romieu A, Williams S, et al. BTA, a novel reagent for DNA attachment on glass and efficient generation of SOLID-phase amplified DNA colonies[J]. Nucleic Acids Research, 2006, 34(3): e22.
- [7] Turcatti G, Romieu A, Fedurco M, et al. A new class of

- cleavable fluorescent nucleotides; Synthesis and optimization as reversible terminators for DNA sequencing by synthesis[J]. *Nucleic Acids Research*, 2008, 36(4): e25.
- [8] Mardis E. The impact of next-generation sequencing technology on genetics [J]. *Trends Genet*, 2008, 24(3): 133-141.
- [9] Shendure J, Porreca G, Reppas N, *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome[J]. *Science*, 2005, 309: 1728-1732.
- [10] Kevin M, Alan B, Lev K, *et al.* Reagents, methods, and libraries for Bead-Based sequencing; US, 13-410919[P]. 2006-12-06.
- [11] Smith D, Quinlan A, Peckham H, *et al.* Rapid whole-genome mutational profiling using next-generation sequencing technologies [J]. *Genome Research*, 2008, 18: 1638-1642.
- [12] Costa V, Angelini C, De Feis I, *et al.* Uncovering the complexity of transcriptomes with RNA-Seq [J]. *Journal of Biomedicine and Biotechnology*, 2010: 853-916.
- [13] Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics [J]. *Nature Reviews Genetics*, 2009, 10(1): 57-63.
- [14] Birzele F, Schaub J, Rust W, *et al.* Into the unknown: Expression profiling without genome sequence information in CHO by next generation sequencing [J]. *Nucleic Acids Research*, 2010, 38(12): 3999-4010.
- [15] Wall P, Mack J, Chanderbali A, *et al.* Comparison of next generation sequencing technologies for transcriptome characterization [J]. *BMC Genomics*, 2009, 10: 347-365.
- [16] Garg R, Patel R, Tyagi A, *et al.* De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification [J]. *DNA Research*, 2011, 18(1): 53-63.
- [17] 李延恩, 周艳红. SNP 功能分析的生物信息学方法及其资源[J]. *计算机仿真*, 2007, 24(4): 297-300.
- [18] Novaes E, Drost D, Farmerie W, *et al.* High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome [J]. *BMC Genomics*, 2008, 9: 312-325.
- [19] Hansey C, Vaillancourt B, Sekhon R, *et al.* Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing [J]. *PLoS One*, 2012, 7(3): e33071.
- [20] Feng C, Chen M, Xu C, *et al.* Transcriptomic analysis of Chinese bayberry (*Myrica rubra*) fruit development and ripening using RNA-Seq [J]. *BMC Genomics*, 2012, 13: 19-33.
- [21] Barrero R, Chapman B, Yang Y, *et al.* De novo assembly of *Euphorbia fischeriana* root transcriptome identifies prostratin pathway related genes [J]. *BMC Genomics*, 2011, 12: 600-613.
- [22] 李霞. 生物信息学(八年制)[M]. 2 版. 北京: 人民卫生出版社, 2010: 134.
- [23] 胡重怡, 蔡刘体, 陈兴江. 烟草 ESTs 资源的 SSR 信息分析 [J]. *生物技术通报*, 2009(7): 82-85.
- [24] Wei W, Qi X, Wang L, *et al.* Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers [J]. *BMC Genomics*, 2011, 12: 451-463.
- [25] Kaur S, Cogan N, Pembleton L, *et al.* Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery [J]. *BMC Genomics*, 2011, 12: 265-275.
- [26] 杨晓玲, 施苏华, 唐恬. 新一代测序技术的发展及应用前景 [J]. *生物技术通报*, 2010(10): 76-81.
- [27] 梁烨, 陈双燕, 刘公社. 新一代测序技术在植物转录组研究中的应用 [J]. *遗传*, 2011, 33(12): 1317-1326.