

# 转录组测序(RNA-seq)策略及其数据在分子标记开发上的应用

李小白<sup>1</sup> 向林<sup>1</sup> 罗洁<sup>1</sup> 胡标林<sup>2</sup> 田胜平<sup>1</sup> 谢鸣<sup>1\*</sup> 孙崇波<sup>1\*</sup>

(<sup>1</sup>浙江省农业科学院, 杭州 310021; <sup>2</sup>江西省农业科学院, 南昌 330200)

**摘要** 新一代测序技术已经广泛应用于转录组的研究(即RNA-seq), 其数据几乎涵盖了特定组织和特定时期的所有转录本。由于其高通量和低成本的优势, RNA-seq越来越受到研究者的青睐, 特别是在那些缺乏基因组信息的物种上。在此综述中, 讨论了RNA-seq的测序策略, 包括物种遗传背景的考量, 测序平台和数据拼接软件的选择。RNA-seq数据对于SSR和SNP等分子标记的开发是非常有用的, 由于这些标记在相关物种中有很高的通用性, 因此在比较作图中有其不可替代的优势。

**关键词** 转录组; 二代测序; 分子标记; 比较作图

## The Strategy of RNA-seq, Application and Development of Molecular Marker Derived from RNA-seq

Li Xiaobai<sup>1</sup>, Xiang Lin<sup>1</sup>, Luo Jie<sup>1</sup>, Hu Biaolin<sup>2</sup>, Tian Shengping<sup>1</sup>, Xie Ming<sup>1\*</sup>, Sun Chongbo<sup>1\*</sup>

(<sup>1</sup>Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China; <sup>2</sup>Jiangxi Academy of Agricultural Sciences, Nanchang 330200, China)

**Abstract** The next-generation sequencing (NGS) has been extensively applied in transcriptomics, called RNA-seq, which characterizes transcriptional events occurring in a specific tissue and a given period. RNA-seq has the advantages of high-throughput and low cost, and becomes the first choice for researchers to carry out biological researches. It has been proven very useful in non-model species, which often lack the resources of sequenced genome. In this review, we discussed about some strategies of RNA-seq, including genetic background consideration, platform selection for sequencing and bioinformatics tools selection for reads assembly. RNA-seq data are also valuable resource for markers development, including microsatellite (SSR), single nucleotide polymorphism (SNP) and so on. This kind of marker has a high transferability between related species and is a powerful tool in comparative mapping.

**Key words** transcriptome; next-generation sequencing (NGS); molecular marker; comparative mapping

## 1 引言

测序技术在过去几十年间得到了飞速发展, 通过全基因组测序已经得到许多植物的整体基因组信

息, 如拟南芥(*Arabidopsis thaliana*)、大豆(*Glycine max*)、葡萄(*Vitis vinifera*)、玉米(*Zea mays*)和水稻(*Oryza sativa*)等。随着生物技术和计算机技术的

收稿日期: 2012-12-01 接受日期: 2013-01-21

国家自然科学基金(批准号: 31201648)、中国博士后科学基金(批准号: 2012M521203)、浙江省博士后科研项目择优资助(批准号: Bsh1201032)、浙江省花卉新品种选育协作组项目(批准号: 2012C12909-10)和浙江省创新团队(批准号: 2009R50034)资助的课题

\*通讯作者。Tel: 0571-86404336, E-mail: xieming1957@yahoo.com.cn; E-mail: hufanfan1982815@gmail.com

Received: December 1, 2012 Accepted: January 21, 2013

This work was supported by the National Nature Science Foundation of China (Grant No.31201648), Postdoctoral Science Foundation of China (Grant No.2012M521203), Foundation for Selected Postdoctoral Project of Zhejiang Province (Grant No.Bsh1201032), New Variety of Flowers Breeding Group Project of Zhejiang Province (Grant No.2012C12909-10) and Scientific and Technical Innovation Group Project of Flower Industry of Zhejiang Province (Grant No.2009R50034)

\*Corresponding author. Tel: +86-571-86404336, E-mail: xieming1957@yahoo.com.cn; E-mail: hufanfan1982815@gmail.com

网络出版时间: 2013-04-28 11:22

URL: http://www.cnki.net/kcms/detail/31.2035.Q.20130428.1122.001.html

进步, 大规模平行测序平台(massively parallel DNA sequencing platform)发展迅速, 过去需要耗费几年完成的数据现在几天甚至几小时就能够完成, 无需繁琐亚克隆的优势是过去毛细管测序无法相比的。大规模平行测序费用的降低进一步促进了大型测序的推广应用。特别是在转录组的研究上, 包括专门的序列数据库的建立、数据分析方法的开发以及新测序实验策略的应用等已经取得了很大成功。

现如今, 二代测序在转录组的研究上越来越普遍, 大有替代先前的基因芯片(microarrays)和基因表达系列分析技术(serial analysis of gene expression, SAGE)之趋势。由于测序深度的优势, RNA-seq更能全面地揭示生物个体在特定时刻和特定组织的全局基因表达情况<sup>[1-5]</sup>, 如发现新的转录本<sup>[6]</sup>、了解基因表达量、挖掘单核苷酸的多态(single-nucleotide polymorphisms, SNPs)<sup>[1,4]</sup>、选择性剪接(alternative splicing)<sup>[7]</sup>和结构性变异(structural variation)<sup>[8]</sup>。对于序列信息有限的非模式生物, RNA-seq更偏重编码区域。由于相比于基因组, 重复元件和高GC区比较少, 使得拼接相对容易, 所以转录组研究在许多非模式植物中得到了广泛应用<sup>[1-5,9-12]</sup>。本文将讨论植物RNA-seq选择策略及RNA-seq数据在标记开发上的应用。

## 2 植物RNA-seq的研究策略

### 2.1 测序平台选择及成本控制

目前, 诸多生物技术公司开发了新一代测序技术, 如454生命科学公司推出的454测序技术、Illumina公司和ABI公司相继推出的Solexa和SOLiD测序技术等, 其中最早实现商业化的是454测序系统

(表1)。454测序技术发明的末端配对测序法(paired-end sequencing)非常适用于发现基因组中的结构变异。Solexa可采用各种方法制备文库, 文库中片段可长达数百碱基对, 并可通过桥式PCR来扩增模板片段。Solexa测序仪在测序长度方面有所缺陷, 这是因为存在诸如核苷酸上的荧光标记基团或终止或切除不完全等原因导致的测序信号衰减(decay)或移相(dephasing)的问题。但对于检测同聚物而言, Solexa测序仪要比454测序仪更具优势。SOLiD测序仪也可以对任何方法制成的文库进行测序, 其特点是能够将富集模板片段的微珠在芯片上进行高度可控的任意排列和使用双碱基编码技术(two-base encoding)。双碱基编码是通过两个碱基来对应一个荧光信号而不是传统的一对一的荧光信号读取, 这样每一个位点都检测了两次, 因此出错率明显降低。但SOLiD技术还存在测序片段比前二者都短的缺点。前二者作为主流的两大测序系统皆可进行两端测序, 即双向测序平台, 只不过这个双向测序在两个系统中是有差异的。454平台对应的是如何将序列定位到基因组上的“头-头”两个相同方向的测序; 而Solexa平台是“头-尾”两个不同方向的测序。

454平台已经在许多非模式植物的转录组中得到了很好的应用<sup>[1-5,10-11,13-15]</sup>。虽然Solexa测得的序列短于454, 但其测序深度远远大于454, 一定程度上提高了其覆盖率, 消减了其序列短的劣势, 近些年来也在许多非模式植物的转录组研究中得到了广泛应用<sup>[16-21]</sup>。值得注意的是, 通过对花菱草(*Eschscholzia californica*)、酪梨(*Persea americana*)和拟南芥(*Arabidopsis thaliana*)的数据模拟研究发现, 454和Solexa两者的结合往往能够得到很好地从头拼接结

表1 四种测序仪器的比较<sup>[72]</sup>  
Table 1 The comparison among four sequencers<sup>[72]</sup>

| 测序仪<br>Sequencer     | 454 GS FLX                                                      | HiSeq 2000              | SOLiDv4                       | Sanger 3730xl                  |
|----------------------|-----------------------------------------------------------------|-------------------------|-------------------------------|--------------------------------|
| Sequencing mechanism | Pyrosequencing                                                  | Sequencing by synthesis | Ligation and two-base coding  | Dideoxy chain termination      |
| Read length          | 700 bp                                                          | 50SE, 50PE, 101PE       | 50+35 bp or 50+50 bp          | 400~900 bp                     |
| Accuracy             | 99.9%                                                           | 98%(100PE)              | 99.94% raw data               | 99.999%                        |
| Output data/run      | 0.7 Gb                                                          | 600 Gb                  | 120 Gb                        | 1.9~84 Kb                      |
| Time/run             | 24 hours                                                        | 3~10 days               | 7 days for SE; 14 days for PE | 20 mins~3 hours                |
| Advantage            | Read length, fast                                               | High throughput         | Accuracy                      | High quality, long read length |
| Disadvantage         | Error rate with polybase more than 6, high cost, low throughput | Short read assembly     | Short read assembly           | High cost, low throughput      |
| Cost/million bases   | \$10                                                            | \$0.07                  | \$0.13                        | \$2 400                        |
| De novo              | Yes                                                             | Yes                     |                               | Yes                            |

果<sup>[15]</sup>, 抑或借助于更长的测序片段, 如Sanger测序片段, 能够更好地从头拼接<sup>[13-14]</sup>。

尽可能减少成本对转录组测序项目的顺利完成至关重要。首先, 确定所需数据量, 因转录组的大小受基因数目和丰度双重影响, 所以测序量随物种转录组大小的不同而有所差异。针对有参考基因组的物种, 可通过分析基因组、统计编码基因及其碱基数来估算转录组的大小; 针对无参考基因组的物种, 只能参考相近物种的转录组大小。其次, 混合不同样品一起测序也是不错的选择。然后, 选择合适的测序平台, 对于模式植物而言, Solexa和SOLiD可能是比较理想的选择, 因为其测序成本远远低于454。而对于非模式植物而言, 首先需要寻找相近物种的基因组信息或者比较完整的转录组数据, 如果没有相关信息, 在经费允许的情况下454可作为首选。但对于经费有限的实验室Solexa也可考虑, 虽然所读序列比较短, 但在可获大数据量测序的前提下会增加测得跨区域片段的概率, 而且相应的强大软件也在陆续开发, 在一定程度上也减少了短序列的劣势。值得注意的是, 当研究侧重于揭示基因表达差异时, 转录组测序必须设置一定的重复。因为生物学差异是基因自身表达的特性而与检测技术的选择以及数据处理的方式无关, 否则, 实验结果存在很高的假阳性<sup>[22]</sup>。

如果数据分析需要在本地完成, 必须有足够运算能力的计算机CPU和一定当量的内存作为保证。由于拼接软件和其他数据分析软件都是基于Linux下运行的, 这就要求数据处理人员有相应系统的操作经验<sup>[23]</sup>。当然, 与有这方面经验的实验室或者生物信息实验员合作是一个比较明智的选择, 这样能够很大程度地减少处理转录本数据的麻烦。此外, 也可考虑公共数据平台来处理NGS数据, 例如Galaxy(<http://main.g2.bx.psu.edu/>)和iPlant(<http://www.iplantcollaborative.org/>)。Galaxy提供一个界面友好的网络框架, 其整合了处理NGS序列的大量工具, 例如BWA、Bowtie、TopHat和Cufflinks<sup>[24]</sup>。iPlant是一个提供超大物理硬盘的全球硬件平台, 具有聚类计算和存储大数据的能力, 还有各种开方式的软件和跨学科的团队。

## 2.2 测序物种遗传背景的考量

植物本身的遗传背景对转录组测序有很大的影响。由于转录组是通过拼接产生的, 所以很难区

分测序错误、杂合或者重复基因的变异等情况<sup>[23]</sup>。对于模式植物而言, 繁殖方式通常是自交, 因而相对容易得到一些纯合系, 例如拟南芥和水稻等。相反, 对于非模式植物而言, 其繁殖方式通常是异交, 从而导致了其高度的杂合。即使在拼接过程进行了序列质量控制和频率检测, 这种杂合仍给SNP的挖掘产生了不小麻烦。另外, 大部分被子植物在进化上经历了染色体倍增事件<sup>[25]</sup>, 例如, 双子叶植物都经历过三倍体化<sup>[26]</sup>, 另外一些植物如水稻、玉米<sup>[27]</sup>以及大豆<sup>[28]</sup>还经历了额外的基因组倍增。这种普遍的多倍体化和平行基因的重复现象给转录组的拼接带来了一定的麻烦<sup>[29-30]</sup>。

有些研究正试图从基因的非编码区的多态性上区分等位和高度同源的基因<sup>[23]</sup>。例如在玉米中, 研究者通过利用mRNA的3'UTR, 在一些纤维素合成和组蛋白H1基因家族等中鉴定了差异的转录本<sup>[31]</sup>。其次, 还应该考虑到研究物种和模式物种之间的亲缘关系。如果亲缘关系比较近, 模式植物的基因组可辅助研究物种的序列拼接, 如把测序片段直接定位到参考基因组, 以此为依据进行拼接。这一策略需调试序列之间不同的容差和匹配长度参数来最优化这些序列的特异性配对<sup>[32]</sup>。事实上, 参考基因组不需要很完整, 例如在甘蓝型油菜(*Brassica napus*)和六倍体小麦(*Triticum aestivum* L.)中, 研究者成功地把Solexa的测序片段定位到了不同来源的重叠群(contigs)上<sup>[12,33]</sup>。但是, 这个方法的前提是两者序列必须先验性地存在用以设定拼接的参数, 若没有可参考的同源序列, 那么从头拼接可能是其唯一的选择了。

## 2.3 转录组的拼接

转录组的拼接对后续的研究是至关重要的。尤其是NGS的数据量往往比传统毛细管测序更大, 但片段短且有不同程度的错误率, 这就使得针对性的拼接就更为重要了<sup>[23]</sup>。根据是否有参考基因组, 拼接策略可分为从头拼接和参考拼接。在许多没有合适的参考基因组的非模式生物中, 从头拼接就成为了唯一的选择, 其测序片段先拼接成一个重叠群(contig)或一个组(group), 这个群或组包含了一个转录本的所有信息。更为重要的是, 这种策略可以发现由于选择性剪接造成的新转录本。

基于不同的算法, 研究者编译出各种软件来处理NGS数据。有一类是基于重叠布局一致性算法



(overlap layout consensus)的软件, 其能够处理454的较长的序列, 一般用于模式生物的拼接。例如, gsAssembler(454 Life Sciences)就是其中一款使用最为广泛的拼接软件, 它可以把不同的等位基因拼接出一致的转录本, 当然也会产生一些可能来源于选择性剪接的孤立序列(isotigs)。Mira是另一款软件, 虽然运行速度比gsAssembler慢, 但对于并入不同来源的EST序列到拼接群中是非常有用的, 是真正意义上的混合拼接软件<sup>[34]</sup>。CAP3也是此类软件, 虽然在处理大数据量时会遇到一定的困难, 但在454片段的拼接上还是比较有效的<sup>[35]</sup>。另一类基于不同算法的软件原本是为全基因组拼接而开发的, 它按照de Bruijn图数据结构首先定义短测序片段为一个短语(short words or k-mers)<sup>[36]</sup>, 它也能有效地处理相对短的片段。如SOAPdenovo<sup>[37]</sup>、Velvet<sup>[32]</sup>等。另外, Trinity是一款专门处理短片段的软件, 在拼接非模式植物的Solexa数据上展现了惊人的能力<sup>[38]</sup>。相比SOAPdenovo、Velvet和MIRA, 研究者发现有一款基于重叠拼接的TGICL和一款商业化的CLC(CLCBio)软件能得到更好的拼接结果<sup>[39]</sup>。当然, 设置拼接参数也是十分重要的, 例如利用不同的k-mers值或结合重叠群翻译序列比较于参考蛋白组, 证明能够实质性地改善短序列的从头拼接<sup>[40]</sup>。

这里值得注意的是, 一个基因可能产生许多重叠拼接群。这一现象可能是由于等位基因的变异、测序的错误或者是选择性剪接造成的。所以, 一些重叠拼接序列有可能只是同一基因的不同区域, 仅仅是因为缺失跨越这些区域的片段所造成的。利用如CAP3这类软件合并不同来源的重叠群在构建全长转录本是非常有用的, 在蕨类植物(*Pteridium aquilinum*)的研究中已经证实了这点<sup>[9]</sup>。

### 3 来源于转录组的分子标记

转录组数据是开发标记的理想资源, 这些基于转录组建立的标记是根据基因本身的差异而建立的标记, 也称为基因内部标记(genic-markers)。根据其造成多态性的原理以及检验方法的不同, 大致可分为三类: 第一类是genic-Indel和genic-SSR。这类标记多态主要是以片段长度多态为基础的, 是目前研究和应用最多的一类, 检测方法主要是PCR, 操作简便、经济。第二类是genic-SNP(单核苷酸多态性)和genic-RFLP。它们的多态主要是由于序列编码的差

异造成的, 前者指的是单个核苷酸差异, 后者可能是一个或者多个核苷酸差异。它们可通过分子杂交、限制性内切酶来检验。第三类是genic-AFLP。造成其多态性的原因更加复杂, 例如序列编码或者长度的差异。它是以限制性内切酶技术和PCR相结合为基础的标记。

基于转录组的标记除具有一般分子标记的特点外, 还有其特殊优势<sup>[41]</sup>。第一, 信息量大。如果发现一个genic标记与某一性状连锁, 那么该转录本就可能与控制此性状的基因相关<sup>[42]</sup>。第二, 通用性好。由于转录本保守性较高, 故具有较好的通用性, 在亲缘物种(closely related species)之间校正连锁图谱和比较作图方面有很高的利用价值<sup>[43-46]</sup>。另外, SNP覆盖率高和检测方法多样, 特别是在NGS价格一直下降的情况下, 通过基因组或者转录组重测序来挖掘和鉴定这些SNP位点也不失为一种策略。研究者往往习惯把一定量SNP作为一个“集合(bin)”进行研究, 这样多态性就得到了改善, 而且其覆盖率高足以用于区分两个生物体, 或构建遗传图谱, 目前, 在水稻、番茄中已经进行了这方面的研究<sup>[47-48]</sup>。对于长度多态性的标记而言, 由于转录本的高度保守性, 已有的研究证实以PCR为基础的这种基因内部标记在种内的多态性较低<sup>[49-50]</sup>。因此, 可以选择尽量靠近3'或5'端非编码区段的标记, 因为这些区域变异性相对丰富<sup>[51]</sup>。另外, 在选择检测手段上, 采用分辨率较高的聚丙烯酰胺凝胶或变性梯度胶来进行分离。还有毛细管电泳可以检测1-2 bp的长度, 也是一种比较理想的检测工具。

#### 3.1 基于转录组的SSR和SNP标记

SSR和SNP是常用的分子标记, 也是利用转录组数据开发最多的两类标记。特别是前者由于其检测方法简便, 一直为研究者所关注。Luro等<sup>[52]</sup>在柑橘(*Citrus clementina*)的转录组中开发和验证了41个genic-SSR标记, 这些标记大多来源于5'端的UTR区域, 证明了其在柑橘属内种间有很好的通用性, 并且利用这些标记研究了柑橘和其亲缘物种的分类关系。Zeng等<sup>[53]</sup>在箭叶淫羊藿(*Epimedium sagittatum*)的转录本中开发了32个genic-SSR标记, 并且18个标记在淫羊藿属内的其他种也通用, 其中16个具有很高的多态性。Kaur等<sup>[54]</sup>比较了红豌豆(*Pisum sativum* L.)和蚕豆(*Vicia faba* L.)之间的转录组数据, 选取96个标记对两种豆类进行检验。在红

豌豆中的86个标记得到了扩增, 其中59%的标记表现了多态性; 在蚕豆中81个标记得到了扩增, 48%呈现了多态性。Zhang等<sup>[55]</sup>利用芝麻转录组(*Sesamum indicum* L.)开发了32个多态性的genic-SSR标记, 其中14个已经整合到了9个遗传图谱中; Wei等<sup>[17]</sup>利用芝麻的不同组织的转录本数据开发了40个多态性genic-SSR。Li等<sup>[19]</sup>在基于橡胶树(*Hevea brasiliensis*)的转录本中选择110个genic-SSR标记, 它们在13个橡胶树品种中成功扩增率达到了96.36%, 多态性达到了55.45%。Liang等<sup>[56]</sup>和Zhang等<sup>[21]</sup>在利用花生(*Arachis hypogaea* L.)的转录本中分别验证和开发到了251个genic-SSR标记和160个多态性的genic-SSR标记。Sakaguchi等<sup>[57]</sup>和Ueno等<sup>[58]</sup>在针叶树(*Callitris columellaris*)中分别开发了52和44个能够成功扩增并显示多态性的genic-SSR。Kaur等<sup>[54]</sup>、Dutta等<sup>[16]</sup>和Schafleitner等<sup>[59]</sup>等在兵豆(*Lens culinaris*)、木豆(*Cajanus cajan*)和番薯(*Ipomoea batatas*)的转录本中也分别开发了相应的genic-SSR标记。

SNP因其覆盖率极高足以用于区分两个生物样本和定位相应的基因, 是非常热门的一类标记。Blanca等<sup>[60]</sup>在67个黄瓜(*Cucumis melo* L.)品种的转录本中发现了303 883个SNP位点分布于15 064个注释基因上, 通过验证发现其多态性比例高达93%, 利用这些SNP区分了不同的地理来源的黄瓜品种; Guo等<sup>[13]</sup>在对黄瓜孤雌和雌雄共体的花蕾转录组的研究中也得到可信度高的SNP以及一组SSR标记。Barbazuk等<sup>[61]</sup>在对两个玉米品系的茎尖顶端分生组织转录组的研究中都检测到大量的SNP存在, 检测超过36 000个SNP, 发现超过85%在Sanger测序中得到了验证。Shu等<sup>[62]</sup>比较了不同大豆(*Glycine max*)品系间的转录组, 最后鉴定出3 899个SNP, 发现这些SNP广泛分布于大豆基因组中, 涉及各种生理生化过程, 影响众多调控重要农艺性状的基因。例如位于基因Glyma07g03490中3'UTR的分子标记CAPS282与大豆的百粒重相关。Novaes等<sup>[4]</sup>、Cogan等<sup>[63]</sup>和Parchman等<sup>[64]</sup>对巨桉(*Eucalyptus grandis*)、黑麦草(*Lolium perenne* L.)和海滩松(*Pinus Contorta*)的转录组研究中也发现了大量的SNP。

这些研究证明了转录组作为一种有效的资源, 非常适宜于SSR和SNP标记的开发。

### 3.2 转录组的分子标记在比较作图中的应用

正如前面讨论的来源于转录区域的分子标记

有比较好通用性, 可以在相近的物种中穿梭。比较作图正是利用共同的遗传标记对相关物种进行物理或遗传作图<sup>[41,65-67]</sup>, 比较这些标记在不同物种的遗传图谱中的分布, 特别是对非模式植物而言通过参考模式植物的遗传图谱更容易定位相关基因, 同时也为相近物种的基因组结构及基因组演化研究提供了有效工具。例如, 由于小麦多倍体性造成了其全基因组测序的很大困难, Yu等<sup>[68]</sup>利用genic-SSR对小麦和水稻同源性位点进行扩增, 发现其中44%的位点是两者共有的。Varshney等<sup>[46]</sup>利用大麦遗传图谱中的含有SSR的转录本与黑麦、小麦和水稻的EST数据比对, 把在其相应的同源序列中的9个大麦的genic-SSR用于黑麦图谱的绘制, 发现所有的这些标记的位置都在预期的功能区, 且与大麦有很高的共线性。吴磊等<sup>[69]</sup>比对分析拟斯卑尔脱山羊草EST(expressed sequence tag)与小麦UniGene序列, 开发了山羊草InDel标记, 这些genic-indel涉及亚细胞定位、蛋白质结合与催化以及代谢等过程的基因。Lan等<sup>[45]</sup>比较利用genic标记建立的卷心菜(*Brassica oleracea*)和拟南芥的遗传图谱, 发现有57%的位点有很好的对应关系, 这为卷心菜基因定位奠定了基础。Brown等<sup>[43]</sup>利用90个火炬松基因内部标记对松属的单维管束亚属、双维管束亚属和黄杉属的花旗松(*Pseudotsuga menziesii*)进行了比较作图, 发现火炬松和花旗松图谱中有35个标记处于相同位点, 这揭示了这些物种基因组上的共线性关系。Nagamura等<sup>[70]</sup>用水稻cDNA标记延伸因子eFF2同源序列对大豆、高粱和拟南芥等16种植物进行杂交, 结果发现均存在同源性序列。Peng等<sup>[71]</sup>用水稻的转录序列数据与小麦的基因组DNA杂交发现了与定位在同源群4B和4D染色体上的矮化基因共分离的RFLP标记。由此看出, 这些基因内部标记在物种之间进行遗传图谱构建和补充, 及基因定位方面有很大的应用前景。

## 4 小结

虽然相比先前的测序技术, NGS的转录组测序有明显的优势并取得了显著成绩, 但是在拼接上还是有許多改进的余地。拼接的准确性和完整性对利用数据开发分子标记至关重要。例如, 过短的拼接结果往往会增加扩增引物设计的困难, 拼接错误更是造成标记扩增失败的致命因素。因此, 一方面, 许

多研究致力于拼接方法的改进; 另一方面, 研究者通过改进测序工艺从而增加测序片段的长度, 从根本上改善后续的拼接。现在, 测序片段更长和出错率更低的第三代测序技术正呼之欲出, 这为后续拼接的准确性和完整性提供了保证。相信在不久的将来, 转录组在更先进技术的护航下会在各个研究领域发挥更重要的作用, 其数据也会为分子标记的开发更好地利用, 特别是对于非模式植物而言, 更是如此。

### 参考文献 (References)

- Alagna F, D'Agostino N, Torchia L, Servili M, Rao R, Pietrella M, *et al.* Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics* 2009; 10: 399.
- Barakat A, DiLoreto DS, Zhang Y, Smith C, Baier K, Powell WA, *et al.* Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. *BMC Plant Biol* 2009; 9: 51.
- Dassanayake M, Haas JS, Bohnert HJ, Cheeseman JM. Shedding light on an extremophile lifestyle through transcriptomics. *New Phytol* 2009; 183(3): 764-75.
- Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, *et al.* High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 2008; 9: 312.
- Wang W, Wang Y, Zhang Q, Qi Y, Guo D. Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics* 2009; 10(1): 465.
- Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome Biol* 2008; 9(12): R175.
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008; 456(7221): 470-6.
- Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* 2009; 458(7234): 97-101.
- Der JP, Barker MS, Wickett NJ, W dePamphilis C, Wolf PG. *de novo* characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. *BMC Genomics* 2011; 12(1): 99.
- Franssen SU, Shrestha RP, Brautigam A, Bornberg-Bauer E, Weber AP. Comprehensive transcriptome analysis of the highly complex *Pisum sativum* genome using next generation sequencing. *BMC Genomics* 2011; 12: 227.
- Riggins CW, Peng Y, Stewart CN, Tranel PJ. Characterization of *De novo* transcriptome for waterhemp (*Amaranthus tuberculatus*) using GS-FLX 454 pyrosequencing and its application for studies of herbicide target-site genes. *Pest Manag Sci* 2010; 66(10): 1042-52.
- Trick M, Long Y, Meng J, Bancroft I. Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnol J* 2009; 7(4): 334-46.
- Guo S, Zheng Y, Joung JG, Liu S, Zhang Z, Crasta OR, *et al.* Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. *BMC Genomics* 2010; 11: 384.
- Swarbreck SM, Lindquist EA, Ackerly DD, Andersen GL. Analysis of leaf and root transcriptomes of soil-grown *Avena barbata* plants. *Plant Cell Physiol* 2011; 52(2): 317-32.
- Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, Liang H, *et al.* Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* 2009; 10: 347.
- Dutta S, Kumawat G, Singh BP, Gupta DK, Singh S, Dogra V, *et al.* Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* (L.) Millspaugh]. *BMC Plant Biol* 2011; 11: 17.
- Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D, *et al.* Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genomics* 2011; 12(1): 451.
- Zhang XM, Zhao L, Larson-Rabin Z, Li DZ, Guo ZH. *De novo* sequencing and characterization of the floral transcriptome of *Dendrocalamus latiflorus* (Poaceae: Bambusoideae). *PLoS One* 2012; 7(8): e42082.
- Li D, Deng Z, Qin B, Liu X, Men Z. *De novo* assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.). *BMC Genomics* 2012; 13(1): 192.
- Li X, Sun H, Pei J, Dong Y, Wang F, Chen H, *et al.* *De novo* sequencing and comparative analysis of the blueberry transcriptome to discover putative genes related to antioxidants. *Gene* 2012; 511(1): 54-61.
- Zhang J, Liang S, Duan J, Wang J, Chen S, Cheng Z, *et al.* *De novo* assembly and characterisation of the transcriptome during seed development, and generation of genic-SSR markers in peanut (*Arachis hypogaea* L.). *BMC Genomics* 2012; 13: 90.
- Hansen KD, Wu Z, Irizarry RA, Leek JT. Sequencing technology does not eliminate biological variability. *Nat Biotechnol* 2011; 29(7): 572-73.
- Strickler SR, Bombarely A, Mueller LA. Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *Am J Bot* 2012; 99(2): 257-66.
- Goecks J, Nekrutenko A, Taylor J, Galaxy T. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010; 11(8): R86.
- Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, *et al.* Polyploidy and angiosperm diversification. *Am J Bot* 2009; 96 (1): 336-48.
- Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007; 449(7161): 463-7.
- Tang H, Bowers JE, Wang X, Paterson AH. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc Natl Acad Sci USA* 2010; 107(1): 472-7.



- 28 Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* 2010; 463(7278): 178-83.
- 29 Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. *Trends Genet* 2008; 24(3): 142-9.
- 30 Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics* 2010; 95: 315-27.
- 31 Eveland AL, McCarty DR, Koch KE. Transcript profiling by 3'-untranslated region sequencing resolves expression of gene families. *Plant Physiol* 2008; 146(1): 32-44.
- 32 Collins LJ, Biggs PJ, Voelckel C, Joly S. An approach to transcriptome analysis of non-model organisms using short-read sequences. *Genome Inform* 2008; 21: 3-14.
- 33 Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P. Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: A comparative study. *BMC Bioinformatics* 2011; 12 Suppl 14: S2.
- 34 Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, *et al.* Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 2004; 14(6): 1147-59.
- 35 Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res* 1999; 9(9): 868-77.
- 36 Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA* 2001; 98(17): 9748-53.
- 37 Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, *et al.* *de novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res* 2010; 20(2): 265-72.
- 38 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011; 29(7): 644-52.
- 39 Bräutigam A, Mullick T, Schliesky S, Weber APM. Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C3 and C4 species. *J Exp Bot* 2011; 62(9): 3093-102.
- 40 Surget-Groba Y, Montoya-Burgos JI. Optimization of *De novo* transcriptome assembly from next-generation sequencing data. *Genome Res* 2010; 20(10): 1432-40.
- 41 李小白, 崔海瑞, 张明龙. EST分子标记开发及在比较基因组学中的应用. 生物多样性(Li Xiaobai, Cui Hairui, Zhang Minglong. Molecular markers derived from EST: Their development and applications in comparative genomics. *Biodiversity Science*) 2006; 14(6): 541-7.
- 42 Bozhko M, Riegel R, Schubert R, Muller-Starck G. A cyclophilin gene marker confirming geographical differentiation of Norway spruce populations and indicating viability response on excess soil-born salinity. *Mol Ecol* 2003; 12(11): 3147-55.
- 43 Brown GR, Kadel EE 3rd, Bassoni DL, Kiehne KL, Temesgen B, van Buijtenen JP, *et al.* Anchored reference loci in loblolly pine (*Pinus taeda* L.) for integrating pine genomics. *Genetics* 2001; 159(2): 799-809.
- 44 Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. *Plant Sci* 2001; 160(6): 1115-23.
- 45 Lan TH, DelMonte TA, Reischmann KP, Hyman J, Kowalski SP, McFerson J, *et al.* An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana*. *Genome Res* 2000; 10(6): 776-88.
- 46 Varshney RK, Sigmund R, Borner A, Korzun V, Stein N, Sorrells ME, *et al.* Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Sci* 2005; 168(1): 195-202.
- 47 Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, *et al.* The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 2012; 485: 635-41.
- 48 Clark RM. Genome-wide association studies coming of age in rice. *Nature Genet* 2010; 42(11): 926-7.
- 49 Wen M, Wang H, Xia Z, Zou M, Lu C, Wang W. Development of EST-SSR and genomic-SSR markers to assess genetic diversity in *Jatropha Curcas* L. *BMC Res Notes* 2010; 3(1): 42.
- 50 Zhou Y, Gao F, Liu R, Feng J, Li H. *De novo* sequencing and analysis of root transcriptome using 454 pyrosequencing to discover putative genes associated with drought tolerance in *Ammopiptanthus mongolicus*. *BMC Genomics* 2012; 13(1): 266.
- 51 Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS, *et al.* Analysis of SSRs derived from grape ESTs. *Theor Appl Genet* 2000; 100(5): 723-6.
- 52 Luro FL, Costantino G, Terol J, Argout X, Allario T, Wincker P, *et al.* Transferability of the EST-SSRs developed on Nules clementine (*Citrus clementina Hort ex Tan*) to other Citrus species and their effectiveness for genetic mapping. *BMC Genomics* 2008; 9: 287.
- 53 Zeng S, Xiao G, Guo J, Fei Z, Xu Y, Roe BA, *et al.* Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genomics* 2010; 11: 94.
- 54 Kaur S, Pembleton LW, Cogan NO, Savin KW, Leonforte T, Paull J, *et al.* Transcriptome sequencing of field pea and faba bean for discovery and validation of SSR genetic markers. *BMC Genomics* 2012; 13: 104.
- 55 Zhang H, Wei L, Miao H, Zhang T, Wang C. Development and validation of genic-SSR markers in sesame by RNA-seq. *BMC Genomics* 2012; 13: 316.
- 56 Liang X, Chen X, Hong Y, Liu H, Zhou G, Li S, *et al.* Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and Arachis wild species. *BMC Plant Biol* 2009; 9: 35.
- 57 Sakaguchi S, Uchiyama K, Ueno S, Ujino-Ihara T, Tsumura Y, Prior LD, *et al.* Isolation and characterization of 52 polymorphic EST-SSR markers for *Callitris columellaris* (Cupressaceae). *Am J Bot* 2011; 98(12): e363-8.
- 58 Ueno S, Moriguchi Y, Uchiyama K, Ujino-Ihara T, Futamura N, Sakurai T, *et al.* A second generation framework for the analysis of microsatellites in expressed sequence tags and the development of EST-SSR markers for a conifer, *Cryptomeria japonica*. *BMC Genomics* 2012; 13: 136.
- 59 Schafleitner R, Tincopa LR, Palomino O, Rossel G, Robles RF, Alagon R, *et al.* A sweetpotato gene index established by *de novo* assembly of pyrosequencing and Sanger sequences and mining for gene-based microsatellite markers. *BMC Genomics* 2010; 11(1): 604.