



高通量测序技术在动植物研究领域中的应用

岳桂东[†], 高强[†], 罗龙海, 王军一, 许姣卉, 尹烨^{*}

深圳华大基因研究院, 深圳 518083

[†] 同等贡献

^{*} 联系人, E-mail: yinye@genomics.cn

收稿日期: 2011-09-09; 接受日期: 2012-01-16

doi: 10.1360/052011-634

摘要 高通量测序是核酸测序研究的一次革命性技术创新, 该技术以极低的单碱基测序成本 and 超高的数据产出量为特征, 为基因组学和后基因组学研究带来了新的科研方法和解决方案. 在动植物研究领域, 高通量测序引领了一次具有里程碑意义的科学研究模式革新, 科研人员可利用该技术在基因组、转录组和表观基因组等领域展开多层次多方面多水平研究. 本文就高通量测序技术应用于动植物基因组学和功能基因组学研究进展进行了系统阐述, 并对当前高通量测序技术的现状和热点及未来的发展趋势作了深入剖析和讨论.

关键词

新一代测序
基因组重测序
转录组测序
表观基因组学

第二代测序技术, 又称新一代测序技术(next-generation sequencing), 相应于以 Sanger 测序法为代表的第三代测序技术而得名. 第二代测序中 3 种主流测序技术分别为依次出现的 Roche/454 焦磷酸测序(2005 年)、Illumina/Solexa 聚合酶合成测序(2006 年)和 ABI/SOLiD 连接酶测序(2007 年)技术. 与 Sanger 测序相比, 3 种新一代测序技术共有的突出特征是, 单次运行(run)产出序列数据量大, 故而又被通称为高通量测序技术^[1].

高通量测序技术一般由模板准备、测序和成像、序列组装和比对等部分组成. 3 种二代测序技术的原理各不相同, 其数据量产出、数据质量和单 Run 运行成本也不一样^[2]. Roche/454 测序序列读长(reads)在 3 种测序技术中最长(600~1000 bp), 但其通量最低(0.5~1 Gb/run); Illumina/Solexa 测序通量大, 其新机型 HiSeq2000 产出量为 600 G/run, 但读长比 454 测序短(通常为 100 bp); ABI/SOLiD 读长最短(50 bp), 但其创新之处在于双碱基编码的应用, 降低了测序的

错误率, 而且由于其双碱基编码和校正系统的原理与重测序相似, 所以 SOLiD 特别适用于具有高质量参考基因组序列物种的重测序.

高通量测序技术的出现, 使得获得核酸序列数据的单碱基测序费用相对于 Sanger 测序急剧下降, 随之也给基因组学研究带来了更多的新方法和新方案. 目前, 高通量测序技术已广泛应用于动植物全基因组测序、基因组重测序、转录组测序、小 RNAs 测序和表观基因组测序等方面. 本文分别对高通量测序技术在多研究水平的具体应用和最新发表的文献进行了综述.

1 全基因组 *de novo* 测序

通过对物种基因组序列进行系统的研究, 可以获得该物种的基因组和重要功能基因的序列信息, 阐述该物种的进化史, 了解该物种生长发育和适应环境的分子机制. 全基因组 *de novo* 测序也称为从头测

序, 它不依赖任何现有的序列资料, 而直接对某个物种的基因组进行测序, 然后利用生物信息学分析手段对序列进行拼接、组装, 从而获得该物种的基因组序列图谱。

1990 年人类基因组计划的启动, 标志着生命科学研究向基因组学领域迈出了重要一步; 多种实验室研究常用的模式生物物种如线虫(*Caenorhabditis elegans*)、拟南芥(*Arabidopsis thaliana*)、小鼠(*Mus musculus*)、水稻(*Oryza sativa*)等全基因组测序计划也得以实施; 伴随着利用 Sanger 测序技术进行全基因组测序的热潮, 人类基因组计划与多种重要模式动植物基因组计划相继顺利完成^[3-7]。在这些重要物种

基因组计划顺利完成后, 科学家们发现对于全面解析众多具有重要科研和经济价值的动植物物种的基因组序列信息来说, 速度慢、通量低和成本高的第一代测序技术远远满足不了需求, 这也使得全新测序技术的开发显得更为紧迫和必要。

第二代测序技术出现后, 科学家们开始选择使用新一代测序技术来进行全基因组 *de novo* 测序, Li 等人^[8]发表在 Nature 上的熊猫(*Ailuropoda melanoleura*)基因组文章, 是第一次完全采用新一代测序技术完成大型物种的全基因组 *de novo* 测序。如表 1 所示, 黄瓜(*Cucumis sativus*)、苹果(*Malus domestica* Borkh)、金小蜂(*Nasonia vitripennis*, *N. giraulti*, *N.*

表 1 已发表的采用高通量测序的动植物基因组

物种名	所属科	基因组大小	所用测序技术
黄瓜 ^[9]	葫芦科	350 M	Sanger+Illumina/Solexa
熊猫 ^[8]	熊科	2.25 G	Illumina/Solexa
金小蜂 ^[10]	金小蜂科	295 M	Illumina/Solexa
苹果 ^[11]	蔷薇科	700 M	Sanger+Roche/454
弓背蚁	蚁科	300 M	Illumina/Solexa
印度跳蚁 ^[12]			
火鸡(<i>Meleagris gallopavo</i>) ^[13]	吐绶鸡科	1.1 G	Roche/454+Illumina/Solexa
野生大豆(<i>Glycine soja</i>) ^[14]	豆科	915.4 M	Illumina/Solexa
森林草莓 ^[15]	蔷薇科	240 M	Roche/454+Illumina/Solexa+ABI/SOLID
可可树 ^[16]	梧桐科	430 M	Sanger, Roche/454+ Illumina/Solexa
阿根廷蚁(<i>Linepithema humile</i>) ^[17]	蚁科	250.8 M	Roche/454+Illumina/Solexa
红色收割蚁(<i>Pogonomyrmex barbatus</i>) ^[18]	蚁科	280 M	Roche/454
火蚁(<i>Solenopsis invicta</i>) ^[19]	蚁科	484.2 M	Roche/454+Illumina/Solexa
麻风树(<i>Jatropha curcas</i>) ^[20]	大戟科	410 M	Sanger+Roche/454
大头切叶蚁(<i>Atta cephalotes</i>) ^[21]	蚁科	300 M	Roche/454
顶切叶蚁(<i>Acromyrmex echinator</i>) ^[22]	蚁科	300 M	Illumina/Solexa
枣椰树(<i>Phoenix dactylifera</i>) ^[23]	棕榈科	685 M	Illumina/Solexa
马铃薯(<i>Solanum tuberosum</i>) ^[24]	茄科	844 M	Illumina/Solexa
中国猕猴(<i>Macaca mulatta</i>) ^[25]	猴科	2.8 G	Illumina/Solexa
珊瑚(<i>Acropora digitifera</i>) ^[26]	鹿角珊瑚	420 M	Roche/454+Illumina/Solexa
巨蟒(<i>Python molurus bivittatus</i>) ^[27]	蟒科	1.4 G	Roche/454
鳕鱼(<i>Atlantic cod</i>) ^[28]	鳕鱼科	830 M	Roche/454
袋鼠(<i>Macropus eugenii</i>) ^[29]	袋鼠科	2.6 G	Sanger+Roche/454+Illumina/ Solexa +ABI/SOLID
白菜(<i>Brassica rapa</i>) ^[30]	十字花科	485 M	Illumina/Solexa
裸鼯鼠(<i>Heterocephalus glaber</i>) ^[31]	鼯鼠科	2.6 G	Illumina/Solexa
食蟹猴(<i>Macaca fascicularis</i>)	猴科	2.8 G	Illumina/Solexa
中国恒河猴(<i>Macaca mulatta lasiota</i>) ^[32]			
大麻(<i>Cannabis sativa</i>) ^[33]	胡麻科	534 M	Illumina/Solexa
木豆(<i>Cajanus cajan</i>) ^[34]	豆科	833 M	Illumina/Solexa
食蟹猴(<i>Macaca fascicularis</i>) ^[35]	猴科	2.8 G	Roche/454
猪蛔虫(<i>Ascaris suum</i>) ^[36]	蛔科	273 M	Illumina/Solexa
帝王蝶(<i>Danaus plexippus</i>) ^[37]	凤蝶科	273 M	Roche/454+Illumina/Solexa
蒺藜苜蓿(<i>Medicago truncatula</i>) ^[38]	豆科	500 M	BAC+optical mapping+ Roche/454

Longicornis)、蚂蚁(弓背蚁 *Camponotus floridanus*, 印度跳蚁 *Harpegnathos saltator*)、森林草莓(*Fragaria vesca*)、可可树(*Theobroma cacao*)等多种物种基因组序列密码解析过程中都采用了新一代测序的序列数据. 如今, 应用第二代测序技术发表的基因组文章也越来越多, 仅 2011 年就有 20 多篇关于不同动植物物种全基因组 *de novo* 测序的文章报道^[17-38]. 随着新一代测序技术的飞速发展, 基因组测序所需的成本较传统技术大大降低, 时间周期也大大缩短, 大规模地物种全基因组 *de novo* 测序渐入佳境, 基因组学研究也迎来新的发展契机和革命性突破.

2 全基因组重测序

全基因组重测序是对基因组序列已知物种的个体进行基因组测序, 并进行差异信息分析的方法. 基于全基因组重测序, 研究人员能够快速的进行资源普查筛选, 寻找到大量的遗传差异, 实现遗传进化分析及重要性状候选基因的预测. 随着测序成本降低和拥有参考基因组序列的物种增多, 基因组重测序也成为育种研究中迅速有效的方法之一, 在全基因组水平扫描并检测出与动植物重要性状相关的变异位点, 具有重大的科研价值和产业价值.

2.1 核心种质资源品系基因组重测序

核心种质是种质资源的核心子集, 它们最大限度地保存了整个资源群体的遗传多样性, 同时代表了整个群体的地理分布. 对核心种质资源进行重测序, 可以深入了解其基因与基因型最大范围的遗传多样性, 同时对于促进种质交流、利用和管理具有重要的学术和实用意义.

Lai 等人^[39]对 6 个玉米(*Zea mays*)骨干自交系(Zheng58, 5003, 478, 178, Chang7-2 和 Mo17)进行了全基因组重测序. 平均每个品系产出 5.4X 的数据, 共发现 1273124 个单核苷酸多态性位点(single nucleotide polymorphisms, SNPs), 得到 30178 个 1~6 bp 的插入缺失位点(insertion and deletion, InDels). 新发现的这些 SNPs 和 InDels 提供了 1 个高密度的全基因组标记信息. 同时鉴定出数百个基因获得与丢失变异(presence/absence variations, PAVs), 其中 296 个 B73 的基因在 6 个品系中的至少 1 个品系中发生丢失,

而 6 个玉米品系中也有数百个基因在 B73 中不存在. 推测这些骨干亲本组合基因组的组合可以弥补另一方功能元件的缺失, 这种 PAVs 的多态变化和其他无义突变的互补作用可能与杂种优势有关.

Zheng 等人^[40]对 3 个高粱(*Sorghum bicolor*)品系(1 个美国甜高粱、1 个中国甜高粱和 1 个中国籽实高粱)品系进行了全基因组重测序, 每株测序深度为 12X, 以已测的美国籽实高粱基因组序列为参考进行信息分析; 发掘出 1057018 个 SNPs, 99948 个 1~10 bp 长的 InDels, 16487 个 PAVs 和 17111 个拷贝数变异(copy number variations, CNVs). 同时, 在甜高粱和籽实高粱序列中鉴定出近 1500 个序列结构差异基因, 这些基因参与糖与淀粉代谢、木质素和香豆素合成、核酸代谢、胁迫应答和 DNA 修复等生物学过程.

2.2 群体基因组重测序

在已知物种基因组的情况下, 对种内群体中的不同个体进行基因组重测序, 可以在全基因组水平上发现群体内个体之间的差异. 通过这种方法, 可以寻找出大量的 SNPs, InDels, 结构变异(structure variations, SVs)等变异信息, 从而获得生物群体的遗传特征. 这对在群体水平上研究物种的进化历史、环境适应性等方面具有重大意义. 利用全基因组重测序有助于快速发现与动植物重要性状相关的遗传变异, 缩短分子育种的实验周期.

Xia 等人^[41]运用新一代测序技术对 29 种家蚕(*Bombyx mori*)和 11 种野蚕(*Bombyx mandarina*)进行了基因组重测序, 构建了一个单碱基分辨率的家蚕遗传变异图谱. 每个个体测序约 3 X, 覆盖基因组序列的 99.88%, 鉴定出 1600 多万个 SNPs, InDels 和 SVs. 分析结果表明, 驯化家蚕确实由野生蚕分化而来, 但它们也保持了很高水平的遗传变异性, 这表明了一个较短驯化事件的发生. 在驯养过程中, 人为选择优良品种, 性状相对单一, 而自然环境选择能够适应环境、生存能力强的个体, 选择性状多样化, 由此导致野蚕 LD(linkage disequilibrium)衰减比家蚕更快. 同时该工作还发现 354 个受到驯化和人工选择压力影响的蛋白编码基因, 它们主要参与调控蚕的丝蛋白合成、能量代谢、生殖特性和飞行能力.

Lam 等人^[42]对 17 株野生大豆(*Glycine soja*)和 14

株栽培大豆(*Glycine max*)进行了全基因组重测序, 鉴定出 630 多万个 SNPs, 建立了大豆高密度分子标记图谱. 通过生物信息软件分别对野生大豆和栽培大豆序列数据进行组装, 鉴定出了 18 万多个 PAVs, 得到了在栽培大豆中获得以及丢失的基因. 与栽培大豆相比, 野生大豆有着更高水平的遗传多样性, 这表明人类的选择导致了栽培大豆狭窄的生物多样性. 此研究还发现了大豆基因组不同于其他作物的 2 个显著特点: 存在较高度度的基因连锁不平衡和较高的单核苷酸非同义替换/同义替换比例. 这表明, 在大豆育种方面, 分子标记育种比基因图位克隆可能更具优势.

通过核心种质资源品系基因组重测序或进行群体基因组重测序分析, 可对所研究物种的遗传变异信息进行全基因组范围地鉴定, 所获得的高密度 SNPs 和 InDels 标记, 可为该物种的遗传学研究以及分子育种提供重要的多态性标记, 也有助于将来在该物种中开展基因与表型关联研究. 同时, 群体遗传学分析所找回的丢失基因对物种基因集也是很好的补充, 可为该物种进一步开展功能基因组学研究提供大量的宝贵数据.

2.3 重测序构建遗传图谱

在过去的 20 多年, 大量的科研工作聚焦于通过数量性状位点(quantitative trait loci, QTLs)分析来探索不同农作物的复杂农艺性状(如产量)的遗传基础. 然而, 几乎所有研究都基于低通量的分子标记(如 RFLPs 和 SSRs)所构建的连锁遗传图谱, 这些图谱大都因分子标记密度较低而不能提供准确和完全的控制性状 QTLs 的数目和座位信息. 新一代测序技术的应用带来基因组序列数据的大幅增长, 为更有效进行基因作图和全基因组范围分析的基因分型策略提供了机遇, 并可在此基础上构建出超高密度的遗传图谱.

Huang 等人^[43]开发了一种通过基因组重测序对重组群体进行基因分型的新方法. 该工作对 9311 与日本晴杂交后自交 11 代得到的 150 个 RILs 进行测序, 每个 RIL 测得 0.02 X 数据, 并将所得数据与 2 个亲本基因组数据相对比, 从而找出 SNPs. 在 150 个 RILs 中, 共鉴定出 1493461 个 SNPs, 平均每个 RIL 的 SNPs 密度是 1 个 SNP/40 kb. 通过滑动窗口方法, 进

行基因分型和重组断裂点的鉴定, 随之构建出含有高密度 SNPs 和高准确性重组断裂位点的重组图谱. 这个水稻重组自交系的遗传图谱, 基因分型准确度为 99.94%, 重组断点分辨率平均为 40 kb. 根据这一基于测序的遗传图谱, 可将一个株高主效 QTL 位点定位在 100 kb 的区域内, 而该区域含有已知的水稻“绿色革命”基因.

Huang 等人^[44]对一个水稻 RILs 群体进行测序, 群体亲本是珍汕 97 和明恢 63. 该研究利用 RIL 个体的基因组数据来反推亲本的基因型, 得到了高质量高密度的遗传图谱. 第一, 所有可能的 SNPs 被鉴定出来, 使用最大简约法在参考数据下, 获得亲本的基因型草图. 第二, 通过使用 SNPs“滑动窗”重采样和贝叶斯推论过滤掉低质量 SNPs 而鉴定出高质量的 SNPs. 第三, 使用高质量的 SNPs 并借助 HMM 算法对群体中株系完成基因分型. 如此构建出一个超高密度的连锁图谱, 使用该图谱, 一个水稻粒宽的 QTL 被定位在一个 200 kb 大小的 bin 中, 该区域含有已知的 *GW5* 基因, 这也反映了该图谱的高质量和高准确性.

Yu 等人^[45]对一个水稻的 RILs 群体个体低覆盖 (0.06 X) 测序得到 SNPs 数据库, 并基于此构建出超高密度的遗传图谱. 用于该研究的水稻重组自交系群体的两个亲本为珍汕 97 和明恢 63, 这 2 个骨干籼稻亚种品系, 通过单粒传的方法构建的群体. 为了评估遗传图谱的质量, 对一些已克隆基因的位置进行了验证, 包括了 *GS3* 和 *GW5/qSW5* 这 2 个分别控制谷粒长度和宽度的主效 QTLs 位点, 以及 *OsC1* 这一控制色泽的质量性状位点. 在所有被选择用于验证的案例中, 目标位点都被精确地定位在基因实际所在的 bins 中, 这表明本研究所得图谱具有很高的质量和精确性. 这个 SNP 图谱结合多年大田调查数据, 对水稻的产量和 3 个产量组成性状(分蘖数、穗粒数、千粒重)进行 QTL 分析, 并与基于 RFLPs/SSRs 的 QTL 定位进行了对比分析. 结果显示, SNP 图谱可以检测出更多的和更精确定位的 QTLs 位点, 这也证实了与 RFLPs/SSRs 图谱相比, SNP 图谱在检测限和分辨率上所具有的优势.

全基因组重测序检测得到 SNPs 信息进而基因分型的方法, 可适用于具有参考基因组序列质量不一的物种的不同类型的作图群体, 并可用于大基因组

和低多态性物种^[43]. 随着测序技术的持续革新, 采用测序构建遗传图谱的技术可能会替代传统的作图方法, 为大范围发掘基因和探索更广泛的生物学问题提供更有力的工具.

2.4 重测序发现突变位点

正向遗传突变、适应进化和表型筛选是创造出带有希望性状的新变异有机体的有力工具和途径, 而这些突变往往不能轻易地用传统遗传学技术鉴定出来. 新一代高通量大规模平行测序技术, 在突变体的亲本株系拥有参考基因组的情况下, 可以快速和准确地得到这一个突变体的基因组信息, 有利于突变位点的初定位和鉴定.

Sarin 等人^[46]使用传统研究手段将一个经 EMS(甲基磺酸乙酯)诱导所得突变体线虫的突变位点定位于 5 号染色体上的 4 Mb 区域内. 随后使用 illumina 深度测序技术得到该线虫突变体的基因组序列, 发现在初定位的 4 Mb 区域中, 突变体基因组和线虫参考基因组存在 80 个变异, 其中有 4 个突变位点处于外显子区, 而唯一的无义突变则存在于 *R07B5.9* 基因中, 该基因在随后的验证实验中得以证实是目标基因.

Cuperus 等人^[47]将 EMS 诱导的 *col-0* 型拟南芥隐性突变体与 *Ler* 型野生拟南芥构建了 F2 群体, 鉴定出 93 个突变体纯合系, 对 93 个样本进行 DNA 提取, 混合后进行高通量测序, 数据经生物信息学分析发现在 2 号染色体上存在一个大小为 1.52 Mb 富含 *col-0* 的 SNPs 的峰(peak). 经进一步研究, 他们在该区域成功找到了目标突变位点.

Ashelford 等人^[48]对一个拟南芥突变体 *ebi-1* 的回交系进行基因组重测序, 随后又通过对突变体的表达数据进行了调查使得候选 SNPs 数目得以有效地缩小. 最终成功鉴定出 1 个在 *AtNFXL-2* 基因中引起 *ebi-1* 突变表型的 SNPs 位点. 该研究证实利用回交系材料可以降低遗传背景噪音, 对其进行测序分析可有效减少候选 SNPs 数目.

2.5 基因组重测序用于 GWAS 研究

近年来出现的全基因关联分析 (genome-wide association study, GWAS) 是一种在医学研究中被广泛使用于鉴定复杂疾病的关键易感基因的新技术,

但该技术未能够广泛应用于作物复杂性状的研究, 其主要原因是植物中基因分型技术的不完善, 以及各作物中缺少类似人类单体型图的高密度单体型图谱.

2010 年, Huang 等人^[44]从近 50000 份的中国起源的水稻品种中, 挑选 517 个具有全面表型调查数据的水稻地方品种进行基因分型工作(基因组重测序). 通过高通量测序总计得到 27 亿条序列读长, 每个品种约测得 1 X 基因组数据. 重测序鉴定出了约 3.6 M 的 SNPs, 通过对 SNPs 的遗传图距计算, 能够把群体分成 3 个的亚群: 粳稻、籼稻以及混合品种. 利用一种新的数据填补方法(k-nearest neighbor algorithm)构建了一个高密度的水稻单体型图谱, 并在 373 个籼稻地方品种中对 14 个重要农艺性状进行 GWAS 分析, 找到了各性状的相关关键位点, 其中有 6 个位点的信号峰与先前用突变体或重组群体研究定位的基因紧密连锁, 而用 GWAS 所鉴定的位点平均来看, 大约可以解释 36% 的表型变异.

最近, Zheng 等人^[40]继续采用这种基于测序的 GWAS 方法, 对一个更大、多样性更高的 950 份世界范围内水稻栽培种(包括粳稻和籼稻亚种), 再次进行 GWAS 分析. 总计鉴定出 32 个与开花期和十粒重相关联的新位点, 这表明更大的样本量可以有效提高 GWAS 检测性状相关变异的能力. 同时, 通过详细注释, 鉴定出 18 个关联位点的候选基因. 这一研究表明, 基于测序的 GWAS 和功能基因组注释的联合应用, 可以在水稻中找出复杂性状与相应遗传多态性因素一一匹配的相应关系.

上述研究表明, GWAS 技术可在水稻中应用以鉴定重要农艺性状的相关基因. 而基于测序的 GWAS 分析方法来鉴定关键基因位点, 是与传统的构建作图群体分析复杂性状方法互为补充的研究策略.

2.6 外显子组和目标区域测序

外显子组(exome)是一个物种基因组中全部外显子区域的总和, 它是基因行使其功能最直接的体现. 通过高通量测序技术进行外显子组测序, 能够直接发现与蛋白质功能变异相关的遗传突变. 目标区域深度测序是指对感兴趣的特定基因组区域进行高通量测序, 因此外显子组测序也属于目标区域测序的一种类型. 可以将感兴趣的基因组区域定制成特异性的探针, 通过这些特异性探针与基因组 DNA 进行

杂交, 富集基因组目标区域, 最后将捕获到的基因组 DNA 进行高通量测序. 相比于全基因组重测序, 外显子组和目标区域测序更加经济、高效. 目前, 在医学基因组学研究领域, 外显子组和目标区域测序技术已经应用到寻找人类各种疾病相关的致病基因和易感基因的研究中, 而在动植物研究中已有报道主要集中在小鼠^[49]中, 在大豆^[50,51]、牛(*Bos taurus*)^[52]、果蝇(*Drosophila melanogaster*)^[53]等物种中也有部分报道.

3 简化基因组测序

随着新一代大规模平行基因分型技术的发展, 对重要和复杂性状基因的 QTLs 精细定位和图位克隆变得更容易获得成功. 在特定作图群体的 QTLs 区域开发高密度的遗传标记是对经济性状重要基因精细定位和图位克隆的基础. SNPs 是最丰富的遗传变异形式, 存在于任何不同基因型之间, 常被用于 QTLs 定位研究. 大规模平行测序技术, 已经被广泛用于鉴

定全基因组范围内的序列变异, 然而对于复杂基因组的物种, 用最经济实惠的方法去鉴定一些 QTLs 区域的 SNP 标记很有必要, 简化基因组测序的研究思路也应运而生.

3.1 简化代表文库测序

简化代表库(reduced representation libraries, RRLs)的使用是在人类研究中首次提出, 可以通过 Sanger 测序有效地发现 SNPs. 通过选择一种内切酶进行酶切, 然后进行文库片段大小的选择, 使用一定大小的酶切片段所对应的序列作为整个基因组序列的部分代表, 来降低基因组的复杂性. 对群体中不同基因型的个体采用相同的内切酶进行酶切, 回收相同大小范围的酶切片段, 建立文库, 然后进行高通量测序. 对于有参考基因组序列的物种进行测序片段的比对, 而对于没有参考序列的物种先进行序列组装, 然后对组装产生的序列进行序列比对. 通过 RRLs 测序和分析, 可以准确地发现 SNPs(图 1).



图 1 简化代表库测序原理

van Tassell 等人^[54]将代表 3 个群体的 66 头牛的基因组 DNA 进行混合, 采用 *Hae* 酶切后建立 RRLs 文库, 并进行了高通量测序. 通过酶切测序获得 62042 个候选的 SNPs, 应用基因分型验证, 在全基因组范围随机挑选 23357 个 SNPs, 验证率为 92%. 该工作建立了一种快速、高效、低成本地大规模开发 SNPs 的新方法.

Kerstens 等人^[55]将来自 2 种品系的 6 个火鸡 (*Meleagris gallopavo*) 个体的基因组 DNA 进行混合, 使用 *Sau3A* 酶消化后构建了一个 RRLs 文库. 该文库经高通量测序及生物信息学分析后, 鉴定出 11000 多个的 SNPs, 在一个代表性的样本中对 340 个 SNPs 进行基因分型发现 SNPs 的转换率为 95%. 该工作是在火鸡基因组数据公布之前所做的, 这表明 RRLs 测序可以高效且低廉地在参考基因组序列未知物种中鉴定出上千个高质量 SNPs.

Wu 等人^[56]对一个栽培大豆作图群体的 2 个亲本基因组 DNA 进行混合, 经 *CviR* 酶切后构建了一个 RRLs 文库并进行了测序分析, 鉴定出 39022 个候选 SNPs. 从得到的 SNPs 标记中选出 164 个 SNPs 用

于定位分析一个已知的 QTL, 可以将该区域的标记密度加密到每 42 Kb 一个标记. 该工作证实使用高通量测序技术和基因组复杂性简化技术相结合的方法, 可以快速地鉴定出很多用于 QTL 区域精细定位的 SNPs 标记.

3.2 酶切位点相关 DNA 测序

酶切位点相关 DNA 测序(restriction-site associated DNA sequencing, RAD-seq)技术是随着高通量测序的出现应运而生的一种新技术, RAD 标签是一种以芯片杂交为基础的分子标记, 而 RAD-seq 则是直接对 RAD 标签进行测序分析, 从而进行 SNPs 的开发和分型, 如图 2 所示, RAD 标签的获得过程, 简单来说是对酶切后 DNA 片段加上可与酶切所得黏性末端序列互补的接头 1, 随机打断后, 再加上非特异性接头 2, 利用识别接头 1 和接头 2 上的引物对进行 PCR 反应, 只有分别带有接头 1 和接头 2 的酶切位点周边的 DNA 片段会得到有意义的扩增, 即为 RAD 标签, 对其进行高通量测序反应, 便被称为 RAD-seq 技术. RAD-seq 技术在模式和非模式生物的遗传分析包括

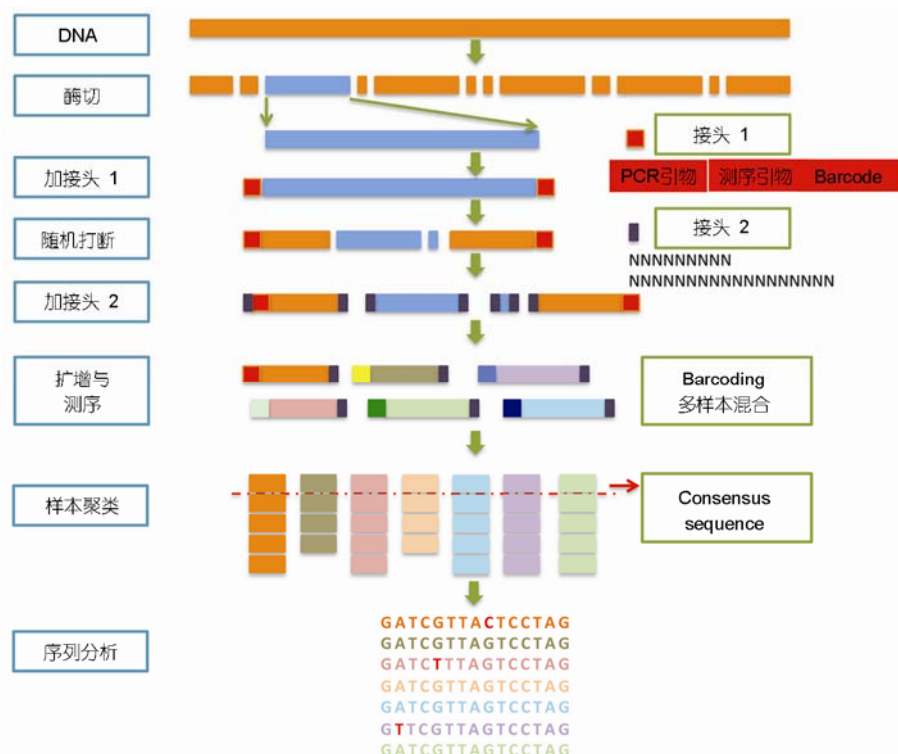


图 2 RAD-seq 技术原理

基因型-表型关联图谱、系统进化、群体遗传等研究领域具有广泛的应用前景。

Baird 等人^[57]第一次报道了 RAD-seq 技术, 他们在三刺鱼(*Gasterosteus aculeatus*)中成功应用该技术鉴定出 13000 多个 SNPs. 同时开发了适用于样本混合的标签系统, 并应用于三刺鱼 F2 群体中个体的重组断点的鉴定, 从而确定了该鱼种侧骨板缺失的遗传基础. 标签系统也被用于另外个性状(弱骨盆结构退化)的遗传定位. 为了进一步证实 RAD-seq 技术可行性, 在链孢霉(*Neurospora crassa*)中利用该方法鉴定了多态性分子标记并成功定位了一个该菌的诱导突变体.

Hohenlohe 等人^[58]利用 RAD-seq 技术对三刺鱼的自然群体间核苷酸的差异进行了基因组范围的扫描. 对来自 2 个海洋和 3 个淡水群体的 100 个个体, 进行 RAD-seq 分析, 每个个体鉴定和分类出 45000 多个 SNPs. 该工作全面地对三刺鱼群体间遗传差异进行了评估并确定了生物地理假说, 表明棘鱼中重复出现的平行类似的表型进化, 可能源自全基因组范围内自始至终广泛并行发生着的遗传进化. 一些此类基因组区域与先前实验室的作图群体鉴定出来的棘鱼表型差异的 QTLs 共定位, 确认了先前鉴定出来的适应性相关的基因组区域.

Emerson 等人^[59]利用 RAD-seq 技术对瓶草草(*Wyeomyia smithii*)进行测序分析, 揭示了瓶草草(起源于 22000 年到 19000 年前劳伦冰盖衰退期间的阿巴拉契亚山脉南部)的遗传结构和进化演变方向. 该工作证实 RAD-seq 技术可用于确定一个物种的近期分化群体的系统发生关系.

Baxter 等人^[60]对小菜蛾(*Plutella xylostella*)回交系进行了 RAD-seq 分析, 成功构建了小菜蛾的遗传连锁图谱, 并完成了小菜蛾和家蚕染色体的比较基因组学分析. Chutimanitsakun 等人^[61]利用 RAD-seq 技术对 93 株大麦(*Hordeum vulgare*)DH(双单倍体)作图群体进行研究, 获得 530 个 SNPs, 应用其中的 445 个 SNPs, 结合之前的标记构建了大麦连锁图谱, 并对生殖适度形状进行了 QTL 定位分析. Pfender 等人^[62]对一个黑麦草(*Lolium perenne*)F1 群体的 193 个体进行 RAD-seq, 并结合 SSRs 和 STSs 分子标记, 利用双假测交原理进行连锁分析, 构建了遗传连锁图谱并检测到 3 个黑麦草抗杆锈病 QTLs.

4 转录组测序

转录组测序, 又称 RNA-seq 或 mRNA-seq, 即从总 RNA 中富集出单链 mRNA 经反转录得到双链 cDNA, 而后对其进行高通量测序分析(图 3). 在已完成基因组测序的物种中, 可将 RNA-seq 结果与基因组 DNA 序列数据进行对比, 从而得到基因表达、可变剪切、基因结构优化、新基因发现等分析结果. 对于没有参考基因组的物种, 可以进行 *de novo* 转录组测序研究, 即对所得测序读长片段进行 *de novo* 组装, 进而得到该物种的单一基因序列集(unigenes)数据.

4.1 有参考基因组物种的转录组测序

Zhang 等人^[63]以水稻 9311 的愈伤组织、根尖、茎尖、叶、稻花/稻穗为材料, 进行转录组测序, 展示了栽培水稻不同器官的转录组图谱. 采用高通量双末端测序, 检测到了 7232 个新转录本, 这些转录本表达丰度低, 且具有组织特异性. 共发现了 23800 个可变剪接, 这些可变剪切事件发生在水稻 33% 的基因中. 同时发现 1356 个融合基因, 鉴定了 234 个候选嵌合转录本, 它们可能是由反式剪切产生地, 这表明转录融合事件比预期地更常见. 这些数据为研究远比预期复杂的水稻转录调节机制提供了广泛的依据.

Li 等人^[64]使用高通量测序分析了玉米叶片的转录组. 作者将玉米 B73 小苗的第 3 叶分为 4 个发育梯度部分(叶基区、过渡区、正在成熟区和成熟区), 并从叶尖中分离出维管束鞘细胞群和叶肉细胞群, 共对 6 个样品进行转录组测序分析. 得到约 120 M 的序列读长去界定基因的结构和可变剪接事件, 并且在不同发育梯度的叶片中及在成熟叶片的维管束鞘和叶肉细胞中进行转录本丰度的定量分析. 检测到大部分基因存在不同的 mRNA 加工事件, 发现在叶片不同发育梯度中有 64% 的基因呈现差异表达, 而 21% 的基因在维管束与叶肉细胞中表达量不同. 对数据聚类分析发现从叶基部的初生细胞壁、基本细胞代谢到过渡到叶尖的次级细胞壁生物合成、C4 植物光合作用的发育的转录组呈现动态变化状态. 这些工作通过对处于不同发育状态的 4 个区域进行研究, 产生了一个高分辨率和高准确性的转录本图谱, 建立了一个可整合生理和代谢数据集的系统框架, 为今

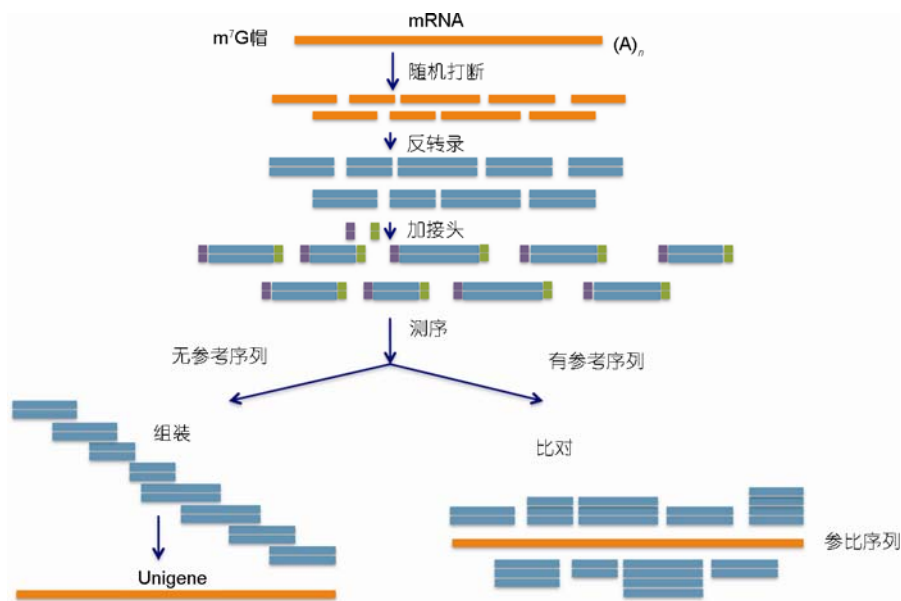


图 3 转录组测序原理

后开展光合作用发育相关研究提供了基础。

4.2 无参考基因组物种的转录组测序

Graham 等人^[65]利用 Roche/454 测序平台对青蒿 (*Artemisia annua*) 的转录组进行了测序分析, 快速找出了大量的 SNPs (1 个 SNP/104 bp), 从中选取 1536 个 SNPs, 结合其他已有 marker 用于遗传连锁图谱的构建。所用的 2 个亲本品种 C4 与 C1 是两个高度杂合且基因型差异很大的青蒿品种, 构建了 F1 群体。Graham 使用了 242 个 F1 个体的数据对 2 个亲本分别构建了遗传图。将表型数据如青蒿素浓度、叶面积、表皮毛状体密度、以及植株鲜重等与基因型数据相关联, 找到了 7 个 QTLs 位点。其中与青蒿素产量相关联的 QTLs 有 3 个, *LG1* 与 *LG9* 上的 QTL 与青蒿素产量正相关, *LG4* 上的 QTL 与青蒿素产量负相关。这一研究结果可以有助于今后对青蒿进行分子标记辅助选育青蒿素高产品种。

Wang 等人^[66]对甘薯 (*Ipomoea batatas*) 的块状根进行高通量转录组测序, 通过 illumina 双末端测序技术产出了超过 59 M 的序列。De novo 组装出 56516 条单一基因序列 (unigenes), 平均长度为 581 bp。基于与已知蛋白的序列相似性搜索, 总计有 35051 个基因被鉴定出来。KEGG 分析有 11056 个是属于代谢途径, 尤其以碳水化合物代谢和次生代谢物质生物合成途

径最突出。此外, 从单一基因序列集中开发出的 4114 个 cDNA SSRs (cSSRs) 候选分子标记可以丰富甘薯的分子标记和有助于将来的分子标记辅助育种。

Chen 等人^[67]报道了飞蝗 (*Brugia malayi*) 的转录组情况, 对测序得到的 21.5 Gb 的测序序列进行了组装, 得到了 72977 条转录本 (N50, 2275 bp), 并鉴定了 11490 个蝗虫蛋白编码基因, 全面发现了具有代表性的核心基因集。与其他 8 种已测序昆虫进行比较基因组学分析, 从而第一次对不完全变态和完全变态的昆虫进行了基因组分歧的鉴定, 且发现了 18 个发育相关基因。

这些研究结果表明基于高通量测序的 *de novo* 转录组分析可在非模式动植物物种, 特别是在基因组大且复杂的物种中, 有效地用于新基因的发现和新分子标记的开发。

5 数字基因表达谱技术

数字基因表达谱 (digital gene expression, DGE) 技术的基本原理 (图 4) 是利用对 mRNAs 经反转录所得 cDNAs 进行双酶切, 使得一条 mRNA 得到一个相对应的标签, 而后进入高通量测序和分析流程, 经过生物信息均一化后比较不同样本间各种标签条数, 找出差异表达标签。早在 1995 年诞生的基因表达系列

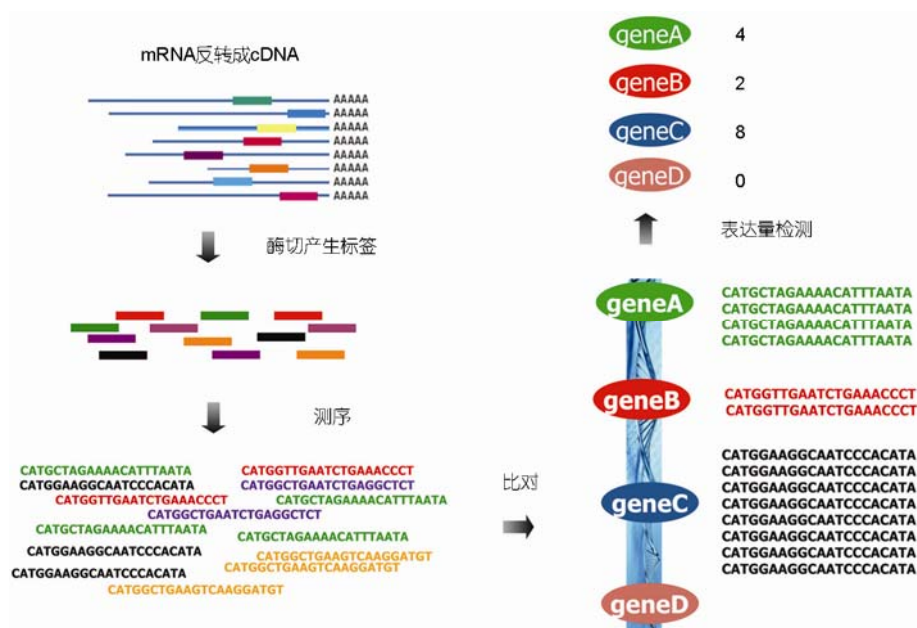


图4 DGE 技术原理

分析(serial analysis of gene expression, SAGE)技术, 是将酶切产生的标签, 用连接酶进行连接后, 进行 Sanger 测序; 而基于高通量测序的 DGE 技术, 是对标签直接进行测序, 故而 DGE 又被称为 DeepSAGE.

DGE 得到的差异表达的标签, 需要与数据库进行对比, 才可以得到基因的差异表达结果, 所以在具有参考基因组的物种和一些没有参考基因组但有单一基因序列集数据库的物种中, 可以直接进行 DGE 的应用, 而在既没有参考基因组又没有单一基因序列集数据库的物种中, 需要先采用 RNA-seq 技术构建一个本物种的单一基因序列集数据库, 用作 DGE 所得标签的对比参考数据库.

5.1 有 unigenes 数据库物种中进行 DGE 分析

Xiao 等人^[68]第一次应用 DGE 技术研究了宿主家猪(*Sus scrofa*)在感染 N-PRRSV(经典北美型 PRRSV)CH1a 株病毒后转录水平做出的响应. 该工作共取 9 头 6 周大的健康小猪, 3 头猪为对照, 在感染实验前一天杀掉取肺; 6 头猪感染病毒, 分别在感染后第 3 天和第 7 天, 各杀 3 头猪取肺. 每 3 头猪混合提取 RNA, 进行 DGE 实验. 研究发现受病毒诱导而上调表达的早期炎症相关细胞因子、趋化因子、黏附分子、发炎相关酶蛋白、发炎细胞、抗体以及补体激活等等很可能导致了 N-PRRSV 感染过程中

发炎响应的发展发生. 而 N-PRRSV 诱导的免疫抑制则可能介导了感染细胞的凋亡, 这导致了免疫细胞的损耗, 也诱导了抗炎细胞因子反应从而不能有效消除初期病毒感染.

Wang 等人^[69]利用 DGE 技术分析了野生型棉花(*Gossypium hirsutum*) Xuzhou142 和它的 fl M(fuzzless/lintless)突变体的胚珠纤维起始阶段和纤维伸长阶段的表达谱. 在开花期当天对花做标记, 收获-2 到 8 DPA(开花期后天数)的棉花胚珠, 剥离的棉花胚珠液氮速冻后-80℃保存. 从-2, -1, 0, 1, 2, 3, 5 和 8 DPA 的野生型和突变体的棉花胚珠中提取总 RNA. -2, -1, 0 和 1 DPA 的总 RNA 混合作为 stage (纤维起始)的样品, 野生型和突变体的分别标记为 WT1 和 M1; 2, 3, 5 和 8 DPA 的总 RNA 混合作为 stage (纤维伸长)的样品, 野生型和突变体分别标记为 WT2 和 M2. 分别对 WT1, M1, WT2 和 M2 进行 DGE 建库测序, 发现野生型和突变体之间表达差异水平最大的 20 个基因是纤维素合成酶基因、磷酸(脂)酶基因和脱氢酶基因, 这些基因都参与了纤维细胞的发育进程, 从而证实了纤维早期发育中基因转录的高度复杂性.

Wu 等人^[70]使用 DGE 研究了葡萄(*Vitis amurensis*)叶片受霜霉病病原菌(*Plasmopara viticola*)感染前后

基因差异表达情况. 该工作分别对 PV 感染葡萄叶片和照叶片进行 DGE 测序. 研究发现 513 个基因和 167 个基因分别在感染叶片中上调和下调表达超过 5 倍. KEGG 富集分析发现, 差异表达基因与核糖体结构、光合作用、氨基酸和糖代谢等代谢通路相关.

5.2 无 unigenes 数据库物种中进行 RNA-seq 和 DGE 联合分析

Wang 等人^[71]从烟粉虱(*Bemisia tabaci*)卵与幼虫混合样本、蛹和成虫中分别进行 RNA 的提取, 将 3 份 RNA 样本混合后, 进行了 3 G 数据量的 RNA-seq; 同时对 3 份 RNA 分别进行 DGE 测序. RNA-seq 所得 43 M 条测序读长, 组装成 168900 个单一基因序列(平均长度为 266 bp), 是此前在 GenBank 数据库中登录注册的烟粉虱的全部序列的 10 倍多. 通过与已知蛋白的相似性搜索, 27290 条序列被鉴定出来. 调查了白粉虱整个发育过程中转录组的变化情况. 而 DGE 实验中, 对每个样本测序得到多于 2.5 M 的标签, 调查了白粉虱整个发育过程中转录组的变化情况, 鉴定出来一大批特异发育时期和抗杀虫剂相关基因.

Mu 等人^[72]用 RNA-seq 和 DGE 分析了大黄鱼(*Pseudosciaena crocea*)受嗜水气单胞菌攻击前后的转录组谱. 对感染后大黄鱼进行转录组测序, 经在 NCBI 数据库中注释分析后, 得到了 8216 条鉴定的单一基因序列. 使用 DGE 技术, 对遭受嗜水气单胞菌感染的大黄鱼和对照大黄鱼对比分析发现, 1996 个基因呈现差异表达, 其中有 727 个上调表达的基因和 489 个下调表达的基因差异在 1.5 倍以上, 而差异变化倍数较大的基因多是炎症反应相关基因.

Hao 等人^[73]对红豆杉(*Taxus mairei*)转录组进行了 *de novo* 测序组装, 基于同已知蛋白的相似性搜索, 23515 个单一基因序列被鉴定; 使用 DGE 调查了红豆杉的根、茎、叶 3 种组织的基因差异表达情况, 鉴定出一批组织特异性功能相关基因和紫杉烷生物合成途径相关基因. 同样, Tang 等人^[74]从开花后 50 和 70 d 的罗汉果(*Siraitia grosvenorii*)中提取 RNA, 等量混合后进行转录组测序. 组装成 43891 个单一基因序列, 平均长度为 668 bp, 从这个文库中鉴定出了所有已知参与罗汉果苷骨架合成的酶的 cDNA 序列. 对开花 3, 50 和 70 d 这 3 个重果的果实发育时期进行 DGE 分析, 选出 7 个 CYP450 和 5 个 UDP 基因作为最可能

参与罗汉果苷合成的候选基因.

新近发展起来的 RNA-seq 和 DGE 技术是高通量的基因发现和表达分析的方法, 是强大的可在转录组水平开展功能基因组学研究的工具. 高通量测序可用于缺少基因组序列信息的物种中转录组的 *de novo* 组装和基因表达分析. 可预计出新一代测序技术在其他非模式生物的转录组研究方面拥有巨大的应用潜力.

6 小 RNAs 测序

MicoRNAs(miRNAs)是一类内源小分子 RNAs, 通常在转录后水平负调节基因表达来发挥作用, 控制了多种生物和代谢途径中众多基因的表达, 在植物生长和发育中扮演着重要角色. 以检测 miRNAs 为主要目的的小 RNAs 测序技术现已经普遍被用于动植物表观遗传学研究.

Wei 等人^[75]对飞蝗进行了小 RNAs 测序. 通过与 miRBase 数据库比对鉴定出 50 个保守的 miRNA 家族, 并在没有飞蝗参考基因组序列的情况下, 通过生物信息分析技术发现了 185 个飞蝗特有的 miRNAs 家族. 对飞蝗 miRNAs 的进化分析表明 miRNAs 的产生主要集中在生物进化过程的 3 个阶段, 无脊椎动物、体腔动物、虫类. 此外还发现了一些转座子来源的内源小 RNAs 和 piRNAs, 也调控了很多基因的表达. 对独居型和群居型 2 个生活型态飞蝗的小 RNAs 的表达情况进行分析, 独居型飞蝗的小 RNAs 表达更为丰富, 找出了在两型飞蝗中差异表达的小 RNAs, 并绘制出 2 个不同生活型态的小 RNAs 表达谱.

Li 等人^[76]采用高通量测序与计算分析相结合的方法, 调查了正常条件下和过氧化氢氧化胁迫处理条件下水稻幼苗的 miRNAs 组. 通过 miRNAomes 和 Northern blot 的分析比较, 确定在氧化胁迫条件下有 7 个 miRNAs 家族呈现显著的差异表达. 这些过氧化氢响应的 miRNAs 的靶基因参与了不同的细胞反应和代谢过程, 包括转录调控、养分运输、生长素的动态平衡、细胞增殖和细胞程序化死亡, 表明多样化的 miRNAs 形成一个复杂的植物氧化应激反应调控网络. 此外, 在水稻幼苗中还发现了 32 个新的 miRNAs, 有趣的是, 首次发现了一个植物外显子小 RNA,

miR3981, 其前体位于基因 *AK106348* 的最后一个外显子中, 这表明植物也可以使用一些外显子作为一个 miRNA 的来源.

Kwak 等人^[77]分别构建了野生型棉花 Xuzhou142 和它的 fl M 突变体的胚珠小 RNA 文库, 并进行高通量测序, 111 个保守的 miRNAs 被鉴定出来, 此外, 还鉴定出 2 个新的细胞特异性的候选 miRNAs. 研究证实了野生型棉花与 fl 突变体中 miRNAs 的表达丰度存在差异, 暗示这些差异表达的 miRNAs 在棉花纤维发育过程中发挥调节转录的作用.

Wang 等人^[78]构建了一个来自浸水 24 h 后玉米种子的小 RNAs 文库并使用 Solexa 技术对其进行了测序分析, 发现了 115 个已知玉米 miRNAs 和 167 个新的 miRNAs. 已知 miRNAs 和新发现 miRNAs 的信息被生物学重复实验所证实. 本研究成功鉴定出的新 miRNAs, 大大丰富了玉米 miRNAs 数据库.

近年来科研工作者也分别将小 RNAs 测序技术用于多种动植物物种, 普查了许多动植物, 如花生 (*Arachis hypogaea*)^[79], 小麦 (*Triticum aestivum*)^[80], 地黄 (*Rehmannia glutinosa*)^[81], 柑橘 (*Citrus trifoliata*)^[82] 和紫菜 (*Porphyra yezoensis*)^[83], 家鸡^[84], 家猪^[85,86], 褐牙鲈 (*Paralichthys olivaceus*)^[87], 松材线虫 (*Bursaphelenchus xylophilus*)^[88], 肝吸虫 (*Clonorchis sinensis*)^[89], 旋毛虫 (*Trichinella spiralis*)^[67], 海胆 (*Strongylocentrotus nudus*)^[90] 等物种的 miRNAs 表达谱, 或者研究了这些物种在不同发育时期、不同胁迫处理等条件下体内 miRNAs 的差异表达情况.

7 降解组测序

鉴定 miRNAs 作用的靶标 mRNAs 是对其功能开展研究的必要内容之一, 传统的鉴定方法主要是依赖于生物信息学方法预测及随后对预测结果进行的实验验证. 随着高通量测序技术的问世, 现在可以直接地对 miRNAs 介导降解 mRNAs 形成的 3 成端带有 ploy(A) 的片段进行测序, 得到这种降解片段数据, 用于新的 miRNA-mRNA 配对关系的发现, 这一新方法被称为降解组测序 (degradome sequencing) (图 5).

在植物体内绝大多数的 miRNAs 是利用剪切作用调控靶基因的表达, 且剪切常发生在 miRNA 与 mRNA 互补区域的第 10 位核苷酸上. 靶基因经剪切产生 5' 剪切片段和 3' 剪切片段. 其中 3' 剪切片段, 含

有自由的 5' 单磷酸和 3' poly(A) 尾巴, 可被 RNA 酶连接, 连接产物可用于下游高通量测序; 而含有 5' 帽子结构的完整基因, 含有帽子结构的 5' 剪切片段或是其他缺少 5' 单磷酸基团的 RNA 是无法被 RNA 酶连接, 因而无法进入下游的测序实验; 对测序数据进行深入地比对分析, 可以直观地发现在 mRNA 序列的某个位点会出现一个波峰 (t-plots), 而该处便是候选的 miRNA 剪切位点.

German^[91]以拟南芥的花组织 (WT 型和 *xrn4* 突变体) 为研究对象, 进行了降解组测序, 得到包括了 800 万个非冗余标签的 2.7 万个转录本信息, 很多以前预测而未被验证的 miRNAs 靶序列在此得到了验证. 与已验证的靶基因相似, 大多数该研究所得 miRNAs 靶基因在 miRNAs 的剪切位点得到单一而丰富的标签, 尤其是在缺少 5'→3' 核外切酶 *AtXRN4* 基因的 *xrn4* 突变体的文库中. 尽管拟南芥的 miRNAs 相关研究已有很多, 该工作仍然给我们一个新的启示, 从反向思路利用降解组测序技术来对新 miRNAs 进行鉴定和验证.

Zhou 等人^[92]报道了在水稻中通过使用降解组测序来鉴定 miRNAs 靶基因的研究. 从粳稻 93-11 的 4~6 cm 长的幼嫩花序中, 鉴定出属于 87 个 miRNAs 靶基因的 177 个转录本. 在拟南芥和水稻中保守的 miRNAs 的靶基因中, 转录因子基因占到 70% (58/82), 这表明, 这些 miRNAs 在水稻的基因调节结点中发挥主要作用. 而非保守的 miRNAs 则调节了影响着更复杂调节网络的众多多样的靶基因.

作为一种技术创新, 降解组测序直接对 miRNAs 形成的 mRNA 的 3' 端降解片段进行高通量测序, 来寻找 miRNAs 作用的靶基因. 该技术可广泛应用于植物突变体、不同组织、胁迫条件、不同基因型及不同的植物物种中, 产生大量有价值的数据, 用于分析 miRNAs 的生物学功能、特异性调控, 甚至是小 RNAs、miRNAs 和靶 mRNAs 的进化. 在动物中, miRNAs 作用靶基因以翻译抑制为主, 但是也有一些 miRNAs 引起靶基因的剪切, 用本方法可以在动物中发现这些稀少但重要的 mRNAs 剪切事件.

8 基于测序的表观基因组学技术

表观遗传学改变是指基于非基因序列改变所致基因表达水平变化, 如 DNA 甲基化和组蛋白修饰等;

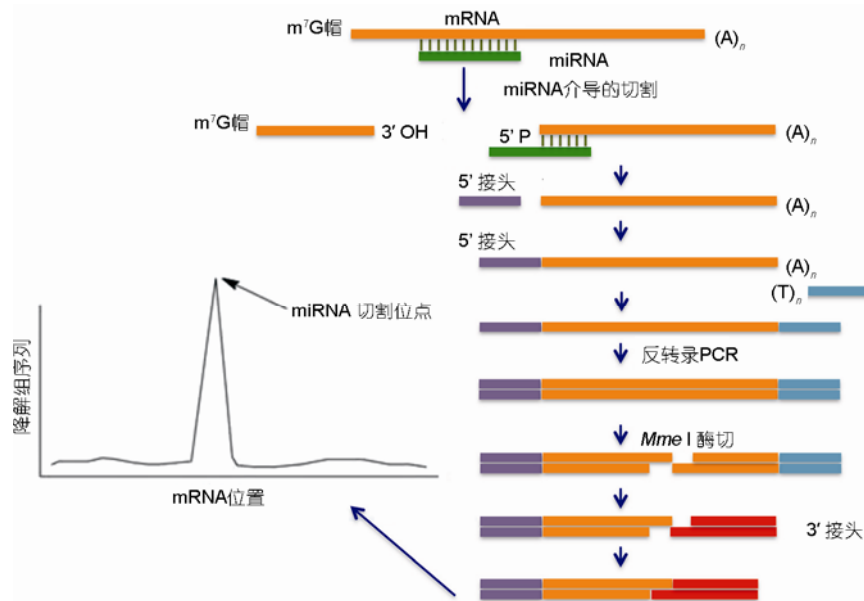


图 5 降解组测序原理

表观基因组学则是在基因组水平上对表观遗传学改变的研究. DNA 甲基化、组蛋白修饰是表观基因组学的重要研究内容.

亚硫酸氢盐修饰结合测序法(bisulfite sequencing)是由 Frommer 等人^[93]提出的研究 DNA 甲基化方法, 原理是亚硫酸氢盐使基因组中未发生甲基化的胞嘧啶脱氨基转变成尿嘧啶, 而甲基化的胞嘧啶保持不变, 然后在所研究的 CpG 位点两侧设计引物进行 PCR 扩增, 对 PCR 产物进行测序并且与未经处理的序列比较, 判断 CpG 位点是否发生甲基化. 目前, 高通量测序技术应用于甲基化测序带来了全基因组甲基化测序(whole genome bisulfite sequencing, WGBS)技术, WGBS 可以在全基因组范围内检测每个 CpG 位点的甲基化状态以及频率, 得到全基因组水平单碱基分辨率的甲基化图谱, 这推动了表观基因组学的发展.

Xiang 等人^[94]对家蚕的 Dazao 品种(基因组序列测序完成的品种)的 2 个个体进行了 WGBS 测序, 测序量分别为 5.9 和 9.9 Gb, 深度为 7.4 和 9.0 X, 覆盖了 92% 的基因组中的胞嘧啶(C). 整合 2 组数据, 得到了家蚕丝腺甲基化图谱, 共鉴定 17 万个 mCs 甲基化位点, 99.2% 位于 CG 位点, BS-PCR 验证表明阳性率高于 85.2%, 大约 0.11% 的基因组胞嘧啶发生甲基化修饰. 家蚕中, 甲基化区域主要富集在基因内部, 与

基因表达水平正相关. 而转座元件, 基因启动区域以及 rDNA 区域甲基化程度很低, 表明在动植物中发挥重要作用的启动子区、核糖体 rDNA 区甲基化调控, 以及转座子区的甲基化抑制等, 这些调控机制可能在昆虫中没有进化出来. 家蚕甲基化谱的完成为理解家蚕及其他昆虫的表观水平的基因调控提供重要参考资料, 也为进一步研究表观遗传系统在家蚕人工驯化过程所起到的贡献奠定了基础.

RRBS(reduced representation bisulfite sequencing)是由 Meissner 等人^[95]提出的一种简化的甲基化测序方法. 结合高通量测序技术的 RRBS, 它类似于构建 RRL 文库, 在进行 DNA 的亚硫酸氢盐修饰之前, 通过选择特定内切酶进行酶切, 选择一定大小的酶切片段所对应的序列作为整个基因组序列的部分代表, 来进行亚硫酸氢盐修饰和后续测序分析. RRBS 的研究方法目前主要应用于小鼠^[96]和人类的研究中, 在动植物方面的相关报道较少, 不过可以预见将来这种性价比很高的方法也会在动植物研究中得到广泛应用.

MeDIP-Seq 技术(methylation DNA immunoprecipitation sequencing)也是研究甲基化的一种有效手段. 其原理是将基因组 DNA 经超声破碎成小片段, 利用抗 5-甲基胞嘧啶的抗体富集高甲基化的 DNA 片段, 该过程称为甲基化 DNA 免疫共沉淀(methylated DNA

immunoprecipitation, MeDIP), 而富集得到的 DNA 片段可进入文库构建和高通量测序过程. 该方法可同时检测整个基因组和特定位点的甲基化状态. Li 等人^[97]利用红色原鸡和爱拔益加肉鸡(美国 AA 肉鸡)的肝脏和肌肉组织, 通过 MeDIP-seq 的方法获得了鸡全基因组甲基化图谱. 该研究发现, 总体上鸡基因组的 DNA 甲基化模式与人类以及拟南芥等物种的 DNA 甲基化模式相似: 基因内部 DNA 甲基化程度较高, 基因的 5'和 3'端甲基化程度相对较低, 在转录起始位点处 DNA 甲基化水平最低. 鸡基因组中大部分 CG 岛是未甲基化的. 基因启动子 DNA 甲基化水平和基因表达水平负相关, 表明其在调控基因转录中起抑制作用. 该研究第一次绘制了鸟类甲基化图谱, 也证明 MeDIP-seq 可以有效的分析鸟类基因组 DNA 甲基化的分布, 有助于对鸟类的表观遗传学的了解.

染色质免疫共沉淀技术(chromatin immunoprecipitation, ChIP)是研究体内蛋白与 DNA 相互作用的一种常用方法, 将其与二代测序技术相结合的 ChIP-seq 技术是, 先通过 ChIP 特异性地富集与目的蛋白相结合的 DNA 片段, 而后对所得 DNA 片段进行高通量测序. 通过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不同修饰的组蛋白、不同转录因子等互作的 DNA 区段信息. Wang 等人^[98]以 15 h 光照 9 h 黑暗条件下培养 14 d 后的 10 株玉米 B73 幼苗为研究对象, 采用 ChIP-seq 技术和甲基化酶切测序技术对玉米基因组中四种组蛋白修饰(H3K4me3, H3K9ac, H3K36me3, H3K27me3)和 DNA 甲基化如何影响基因表达进行了系统的分析. 结果表明, 3 种组蛋白(H3K4me3, H3K9ac, H3K36me3)对基因表达起到正调控的作用; 而另一种组蛋白(H3K27me3)和 DNA 甲基化对基因表达具有负调控

的作用, 这 2 种调控是相互排斥的, 有 H3K27me3 的地方 DNA 甲基化水平低, 反之亦然. 同时发现玉米 DNA 甲基化 peaks 出现在 ATG 周围, 其甲基化模式与水稻非常相似, 而与拟南芥不同.

9 结语

高通量测序技术的诞生可以说是基因组学研究领域一个具有里程碑意义的事件. 该技术使得核酸测序的单碱基成本与第一代测序技术相比急剧下降, 以人类基因组测序为例, 上世纪末进行的人类基因组计划花费 30 亿美元解码了人类生命密码, 而第二代测序使得人类基因组测序已进入万(美)元基因组时代. 如此低廉的单碱基测序成本使得我们可以实施更多物种的基因组计划从而解密更多生物物种的基因组遗传密码. 同时在已完成基因组序列测定的物种中, 对该物种的其他品种进行大规模地全基因组重测序也成为了可能. 现在高通量测序已被广泛应用于以转录组测序等为代表的功能基因组学研究中. 对有参考基因组序列的物种, 转录组测序数据可以大大丰富和验证对基因组数据的注释, 而该技术本身可用于不同样本间基因表达差异、可变剪接等的比较. 在无参考基因组序列的物种中, 转录组测序可以大量且快速地充实该物种的遗传数据库, 这些数据资源可有助于在该物种中进一步开展分子生物学研究. 随高通量测序技术而出现的数字基因表达谱(DGE)测序、小 RNAs 测序、降解组测序、DNA 甲基化测序、染色质免疫共沉淀 DNA 测序等新方法为科学家们进行分子生物学相关研究提供了更多的选择. 总而言之, 高通量测序技术给基因组学研究带来了一个高效的新平台和巨大的发展机遇.

参考文献

- 1 Ansorge W J. Next-generation DNA sequencing techniques. *N Biotechnol*, 2009, 25: 195–203
- 2 Glenn T C. Field guide to next-generation DNA sequencers. *Mol Ecol Resour*, 2011, 11: 759–769
- 3 Hillier L W, Marth G T, Quinlan A R, et al. Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods*, 2008, 5: 183–188
- 4 TAGI. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 2000, 408: 796–815
- 5 Check E. Mouse genome: the real deal. *Nature*, 2002, 420: 457
- 6 Waterston R H, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 2002, 420: 520–562

- 7 Yu J, Hu S, Wang J, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, 2002, 296: 79–92
- 8 Li R, Fan W, Tian G, et al. The sequence and *de novo* assembly of the giant panda genome. *Nature*, 2010, 463: 311–317
- 9 Huang S, Li R, Zhang Z, et al. The genome of the cucumber, *Cucumis sativus* L. *Nat Genet*, 2009, 41: 1275–1281
- 10 Werren J H, Richards S, Desjardins C A, et al. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science*, 2010, 327: 343–348
- 11 Velasco R, Zharkikh A, Affourtit J, et al. The genome of the domesticated apple (*Malus x domestica* Borkh.). *Nat Genet*, 2010, 42: 833–839
- 12 Bonasio R, Zhang G, Ye C, et al. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science*, 2010, 329: 1068–1071
- 13 Dalloul R A, Long J A, Zimin A V, et al. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biol*, 2010, 8: e1000475
- 14 Kim M Y, Lee S, Van K, et al. Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proc Natl Acad Sci USA*, 2010, 107: 22032–22037
- 15 Shulaev V, Sargent D J, Crowhurst R N, et al. The genome of woodland strawberry (*Fragaria vesca*). *Nat Genet*, 2011, 43: 109–116
- 16 Argout X, Salse J, Aury J M, et al. The genome of *Theobroma cacao*. *Nat Genet*, 2011, 43: 101–108
- 17 Smith C D, Zimin A, Holt C, et al. Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). *Proc Natl Acad Sci USA*, 2011, 108: 5673–5678
- 18 Smith C R, Smith C D, Robertson H M, et al. Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *Proc Natl Acad Sci USA*, 2011, 108: 5667–5672
- 19 Wurm Y, Wang J, Riba-Grognuz O, et al. The genome of the fire ant *Solenopsis invicta*. *Proc Natl Acad Sci USA*, 2011, 108: 5679–5684
- 20 Sato S, Hirakawa H, Isobe S, et al. Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. *DNA Res*, 2011, 18: 65–76
- 21 Suen G, Teiling C, Li L, et al. The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. *PLoS Genet*, 2011, 7: e1002007
- 22 Nygaard S, Zhang G, Schiott M, et al. The genome of the leaf-cutting ant *Acromyrmex echinator* suggests key adaptations to advanced social life and fungus farming. *Genome Res*, 2011, 21: 1339–1348
- 23 Al-Dous E K, George B, Al-Mahmoud M E, et al. *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat Biotechnol*, 2011, 29: 521–527
- 24 Xu X, Pan S, Cheng S, et al. Genome sequence and analysis of the tuber crop potato. *Nature*, 2011, 475: 189–195
- 25 Fang X, Zhang Y, Zhang R, et al. Genome sequence and global sequence variation map with 5.5 million SNPs in Chinese rhesus macaque. *Genome Biol*, 2011, 12: R63
- 26 Shinzato C, Shoguchi E, Kawashima T, et al. Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature*, 2011, 476: 320–323
- 27 Castoe T A, de Koning A J, Hall K T, et al. Sequencing the genome of the *Burmese python* (*Python molurus bivittatus*) as a model for studying extreme adaptations in snakes. *Genome Biol*, 2011, 12: 406
- 28 Star B, Nederbragt A J, Jentoft S, et al. The genome sequence of *Atlantic cod* reveals a unique immune system. *Nature*, 2011, 477: 207–210
- 29 Renfree M B, Papenfuss A T, Deakin J E, et al. Genome sequence of an Australian kangaroo, *Macropus eugenii*, provides insight into the evolution of mammalian reproduction and development. *Genome Biol*, 2011, 12: R81
- 30 Wang X, Wang H, Wang J, et al. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet*, 2011, 43: 1035–1039
- 31 Kim E B, Fang X, Fushan A A, et al. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature*, 2011, 479: 223–227
- 32 Yan G, Zhang G, Fang X, et al. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nat Biotechnol*, 2011, 29: 1019–1023
- 33 van Bakel H, Stout J M, Cote A G, et al. The draft genome and transcriptome of *Cannabis sativa*. *Genome Biol*, 2011, 12: R102
- 34 Varshney R K, Chen W, Li Y, et al. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat Biotechnol*, 2012, 30: 83–89
- 35 Ebeling M, Kung E, See A, et al. Genome-based analysis of the nonhuman primate *Macaca fascicularis* as a model for drug safety assessment. *Genome Res*, 2011, 21: 1746–1756
- 36 Jex A R, Liu S, Li B, et al. *Ascaris suum* draft genome. *Nature*, 2011, 479: 529–533
- 37 Zhan S, Merlin C, Boore J L, et al. The monarch butterfly genome yields insights into long-distance migration. *Cell*, 2011, 147: 1171–1185
- 38 Young N D, Debelle F, Oldroyd G E, et al. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature*, 2011,

- 480: 520–524
- 39 Lai J, Li R, Xu X, et al. Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet*, 2010, 42: 1027–1030
- 40 Zheng L Y, Guo X S, He B, et al. Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). *Genome Biol*, 2011, 12: R114
- 41 Xia Q, Guo Y, Zhang Z, et al. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science*, 2009, 326: 433–436
- 42 Lam H M, Xu X, Liu X, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet*, 2010, 42: 1053–1059
- 43 Huang X, Feng Q, Qian Q, et al. High-throughput genotyping by whole-genome resequencing. *Genome Res*, 2009, 19: 1068–1076
- 44 Huang X, Wei X, Sang T, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet*, 2010, 42: 961–967
- 45 Yu H, Xie W, Wang J, et al. Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PLoS One*, 2011, 6: e17595
- 46 Sarin S, Prabhu S, O'Meara M M, et al. *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. *Nat Methods*, 2008, 5: 865–867
- 47 Cuperus J T, Montgomery T A, Fahlgren N, et al. Identification of MIR390a precursor processing-defective mutants in *Arabidopsis* by direct genome sequencing. *Proc Natl Acad Sci USA*, 2010, 107: 466–471
- 48 Ashelford K, Eriksson M E, Allen C M, et al. Full genome re-sequencing reveals a novel circadian clock mutation in *Arabidopsis*. *Genome Biol*, 2011, 12: R28
- 49 Fairfield H, Gilbert G J, Barter M, et al. Mutation discovery in mice by whole exome sequencing. *Genome Biol*, 2011, 12: R86
- 50 Haun W J, Hyten D L, Xu W W, et al. The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol*, 2011, 155: 645–655
- 51 Bolon Y T, Haun W J, Xu W W, et al. Phenotypic and genomic analyses of a fast neutron mutant population resource in soybean. *Plant Physiol*, 2011, 156: 240–253
- 52 Cosart T, Beja-Pereira A, Chen S, et al. Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics*, 2011, 12: 347
- 53 Wang H, Chattopadhyay A, Li Z, et al. Rapid identification of heterozygous mutations in *Drosophila melanogaster* using genomic capture sequencing. *Genome Res*, 2011, 20: 981–988
- 54 Van Tassell C P, Smith T P, Matukumalli L K, et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods*, 2008, 5: 247–252
- 55 Kerstens H H, Crooijmans R P, Veenendaal A, et al. Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics*, 2009, 10: 479
- 56 Wu X, Ren C, Joshi T, et al. SNP discovery by high-throughput sequencing in soybean. *BMC Genomics*, 2010, 11: 469
- 57 Baird N A, Etter P D, Atwood T S, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, 2008, 3: e3376
- 58 Hohenlohe P A, Bassham S, Etter P D, et al. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet*, 2010, 6: e1000862
- 59 Emerson K J, Merz C R, Catchen J M, et al. Resolving postglacial phylogeography using high-throughput sequencing. *Proc Natl Acad Sci USA*, 2010, 107: 16196–16200
- 60 Baxter S W, Davey J W, Johnston J S, et al. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS One*, 2011, 6: e19315
- 61 Chutimanitsakun Y, Nipper R W, Cuesta-Marcos A, et al. Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *BMC Genomics*, 2011, 12: 4
- 62 Pfender W F, Saha M C, Johnson E A, et al. Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *Theor Appl Genet*, 2011, 122: 1467–1480
- 63 Zhang G, Guo G, Hu X, et al. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res*, 2010, 20: 646–654
- 64 Li P, Ponnala L, Gandotra N, et al. The developmental dynamics of the maize leaf transcriptome. *Nat Genet*, 2010, 42: 1060–1067
- 65 Graham I A, Besser K, Blumer S, et al. The genetic map of *Artemisia annua* L. identifies loci affecting yield of the antimalarial drug artemisinin. *Science*, 2010, 327: 328–331

- 66 Wang Z, Fang B, Chen J, et al. *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (*Ipomoea batatas*). BMC Genomics, 2010, 11: 726
- 67 Chen M X, Ai L, Xu M J, et al. Identification and characterization of microRNAs in *Trichinella spiralis* by comparison with *Brugia malayi* and *Caenorhabditis elegans*. Parasitol Res, 2011, 109: 553–558
- 68 Xiao S, Jia J, Mo D, et al. Understanding PRRSV infection in porcine lung based on genome-wide transcriptome response identified by deep sequencing. PLoS One, 2010, 5: e11377
- 69 Wang Q Q, Liu F, Chen X S, et al. Transcriptome profiling of early developing cotton fiber by deep-sequencing reveals significantly differential expression of genes in a fuzzless/lintless mutant. Genomics, 2010, 96: 369–376
- 70 Wu J, Zhang Y, Zhang H, et al. Whole genome wide expression profiles of *Vitis amurens* grape responding to downy mildew by using Solexa sequencing technology. BMC Plant Biol, 2010, 10: 234
- 71 Wang X W, Luan J B, Li J M, et al. *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics, 2010, 11: 400
- 72 Mu Y, Ding F, Cui P, et al. Transcriptome and expression profiling analysis revealed changes of multiple signaling pathways involved in immunity in the large yellow croaker during *Aeromonas hydrophila* infection. BMC Genomics, 2010, 11: 506
- 73 Hao D C, Ge G B, Xiao P G, et al. The first insight into the tissue specific taxus transcriptome via illumina second generation sequencing. PLoS One, 2011, 6: e21220
- 74 Tang Q, Ma X J, Mo C M, et al. An efficient approach to finding *Siraitia grosvenorii* triterpene biosynthetic genes by RNA-seq and digital gene expression analysis. BMC Genomics, 2011, 12: 343
- 75 Wei Y, Chen S, Yang P, et al. Characterization and comparative profiling of the small RNA transcriptomes in two phases of locust. Genome Biol, 2009, 10: R6
- 76 Li T, Li H, Zhang Y X, et al. Identification and analysis of seven HO-responsive miRNAs and 32 new miRNAs in the seedlings of rice (*Oryza sativa* L. ssp. indica). Nucleic Acids Res, 2011, 39: 2821–2833
- 77 Kwak P B, Wang Q Q, Chen X S, et al. Enrichment of a set of microRNAs during the cotton fiber development. BMC Genomics, 2009, 10: 457
- 78 Wang L, Liu H, Li D, et al. Identification and characterization of maize microRNAs involved in the very early stage of seed germination. BMC Genomics, 2011, 12: 154
- 79 Zhao C Z, Xia H, Frazier T P, et al. Deep sequencing identifies novel and conserved microRNAs in peanuts (*Arachis hypogaea* L.). BMC Plant Biol, 2010, 10: 3
- 80 Xin M, Wang Y, Yao Y, et al. Diverse set of microRNAs are responsive to powdery mildew infection and heat stress in wheat (*Triticum aestivum* L.). BMC Plant Biol, 2010, 10: 123
- 81 Yang Y, Chen X, Chen J, et al. Differential miRNA expression in *Rehmannia glutinosa* plants subjected to continuous cropping. BMC Plant Biol, 2011, 11: 53
- 82 Song C, Wang C, Zhang C, et al. Deep sequencing discovery of novel and conserved microRNAs in trifoliate orange (*Citrus trifoliata*). BMC Genomics, 2010, 11: 431
- 83 Liang C, Zhang X, Zou J, et al. Identification of miRNA from *Porphyra yezoensis* by high-throughput sequencing and bioinformatics analysis. PLoS One, 2010, 5: e10698
- 84 Yao J, Wang Y, Wang W, et al. Solexa sequencing analysis of chicken pre-adipocyte microRNAs. Biosci Biotechnol Biochem, 2011, 75: 54–61
- 85 Xie W, Feng Q, Yu H, et al. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc Natl Acad Sci USA, 2010, 107: 10578–10583
- 86 Li G, Li Y, Li X, et al. MicroRNA identity and abundance in developing swine adipose tissue as determined by Solexa sequencing. J Cell Biochem, 2011, 112: 1318–1328
- 87 Fu Y, Shi Z, Wu M, et al. Identification and differential expression of microRNAs during metamorphosis of the Japanese flounder (*Paralichthys olivaceus*). PLoS One, 2011, 6: e22957
- 88 Huang Q X, Cheng X Y, Mao Z C, et al. MicroRNA discovery and analysis of pinewood nematode *Bursaphelenchus xylophilus* by deep sequencing. PLoS One, 2011, 5: e13271
- 89 Xu M J, Liu Q, Nisbet A J, et al. Identification and characterization of microRNAs in *Clonorchis sinensis* of human health significance. BMC Genomics, 2010, 11: 521
- 90 Wei Z, Liu X, Feng T, et al. Novel and conserved micornas in Dalian purple urchin (*Strongylocentrotus nudus*) identified by next

- generation sequencing. *Int J Biol Sci*, 2011, 7: 180–192
- 91 German M A, Pillay M, Jeong D H, et al. Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol*, 2008, 26: 941–946
- 92 Zhou M, Gu L, Li P, et al. Degradome sequencing reveals endogenous small RNA targets in rice (*Oryza sativa* L. ssp. indica). *Front Biol*, 2010, 5: 24
- 93 Frommer M, McDonald L E, Millar D S, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA*, 1992, 89: 1827–1831
- 94 Xiang H, Zhu J, Chen Q, et al. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nat Biotechnol*, 2011, 28: 516–520
- 95 Meissner A, Gnirke A, Bell G W, et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res*, 2005, 33: 5868–5877
- 96 Hirst M, Marra M A, Next generation sequencing based approaches to epigenomics. *Brief Funct Genomics*, 2011, 9: 455–465
- 97 Li Q, Li N, Hu X, et al. Genome-wide mapping of DNA methylation in chicken. *PLoS One*, 2011, 6: e19428
- 98 Wang X, Elling A A, Li X, et al. Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. *Plant Cell*, 2009, 21: 1053–1069

The Application of High-throughput Sequencing Technology in Plant and Animal Research

YUE GuiDong, GAO Qiang, LUO LongHai, WANG JunYi, XU JiaoHui, YIN Ye

Beijing Genomics Institute-Shenzhen, Shenzhen 518083, China

High-throughput sequencing technology is a revolutionary technological innovation of sequencing technology. This technology, with low cost and high throughput data, brought new research methods and solutions for genomics and post genomics research. In plant and animal research, the high-throughput sequencing, leading a landmark innovation in scientific research model, can improve researches in the genome, transcriptome and epigenome, and multi-level system. In this review, we described the research progress of high-throughput sequencing applications in plant and animal genomics and functional genomics, and the current status and hot points of high-throughput sequencing technology. We also discussed the trends in the further application of high-throughput sequencing.

next-generation sequencing, genome resequencing, RNA sequencing, epigenomics