

# 花椒茎尖节点转录组测序及基因注释

蒋弘刚<sup>1</sup>, 魏安智<sup>1\*</sup>, 杨途熙<sup>1</sup>, 代惠萍<sup>2</sup>

(1 西北农林科技大学 林学院, 陕西 杨陵 712100; 2 陕西理工学院, 陕西 汉中 723000)

**摘要:**花椒茎枝有尖锐皮刺,在实际生产中增加采摘成本,并降低采摘效率。为培育品质优良的无刺品种,筛选皮刺分化的关键基因具有重要意义。使用二代高通量测序技术对花椒皮刺起始分化的茎尖节点组织转录组进行测序,使用生物学软件对原始数据进行拼装,最终获得 45 057 条转录本序列。使用比对工具将得到的转录本与 Swiss-Prot, NR, Pfam, KEGG, KOG 数据库比对后, 80.34% 的转录本序列至少被一个数据库注释。获取的大量的基因序列及注释信息为进一步探究花椒皮刺分化分子机理提供了良好的基础,为定位皮刺分化关键基因提供依据。

**关键词:**花椒; 皮刺分化; 转录组; 高通量测序

**中图分类号:**S722.39      **文献标志码:**A      **文章编号:**1001-7461(2014)06-0094-06

## Transcriptome Sequencing of Node Tissue of Stem Tip of *Zanthoxylum bungeanum* and Gene Annotation

JIANG Hong-gang<sup>1</sup>, WEI An-zhi<sup>1\*</sup>, YANG Tu-xi<sup>1</sup>, DAI Hui-ping<sup>2</sup>

(1. College of Forestry, Northwest A&F University, Yangling, Shaanxi 712100, China;

2. Shaanxi University of Technology, Hanzhong, Shaanxi 723000, China)

**Abstract:** The fruit of *Zanthoxylum bungeanum* is one of the most popular spices in China. However, there are numerous short prickles on both the stems and the branches, which result high cost and low efficiency of picking the fruit. Therefore, breeding of spine free cultivar is needed. During this process, localization of essential genes in prickle differentiation is above all. We sequenced the node of stem where prickle differentiation initiation by Next-generation sequencing technology. After *de novo* assembly, we obtained 45 057 transcripts of the node of the stem. In addition, we annotated the transcriptome against five public database (Swiss-Prot, NR, Pfam, KEGG and KOG). The massive sequence and annotation information would build a solid base for non-prickle hua-jiao breeding.

**Key words:** *Zanthoxylum bungeanum*; prickle differentiation; transcriptome; high-throughput sequencing

花椒 (*Zanthoxylum bungeanum*) 为芸香科花椒属植物,在全球约有 250 种,我国有 39 个变种,以枝条上皮刺的数量多少可分为多刺、少刺及无刺 3 类<sup>[1]</sup>。目前生产上栽培广泛、品质优良的地方品种常具较多皮刺。从分子水平上解释皮刺分化的机理将有助于培育无刺的优良品质的花椒。目前未见花椒基因组和转录组的相关报道,遗传信息匮乏。为此,利用新一代测序平台进行高通量测序,可以在有

效获取大量花椒转录本信息,利用生物信息分析从海量的转录本数据中筛选候选基因,为后续研究提供良好的基础。

花椒的皮刺结构相对简单,无维管束,其组织细胞形态与茎表皮相近。虽然目前尚无直接阐述皮刺分化分子机理的报道。现已证明植物细胞极性的建立及形态发生依赖于生长素浓度梯度的建立,而浓度梯度的建立需要是由生长素的合成和运输机制协

收稿日期:2014-05-23    修回日期:2014-07-10

基金项目:国家林业局林业公益性行业科研专项(201304706)。

作者简介:蒋弘刚,男,在读硕士,研究方向:林木遗传改良。E-mail: freshair@vip.qq.com

\* 通信作者:魏安智,男,教授,研究方向:林木遗传改良。E-mail: weianzhi@126.com

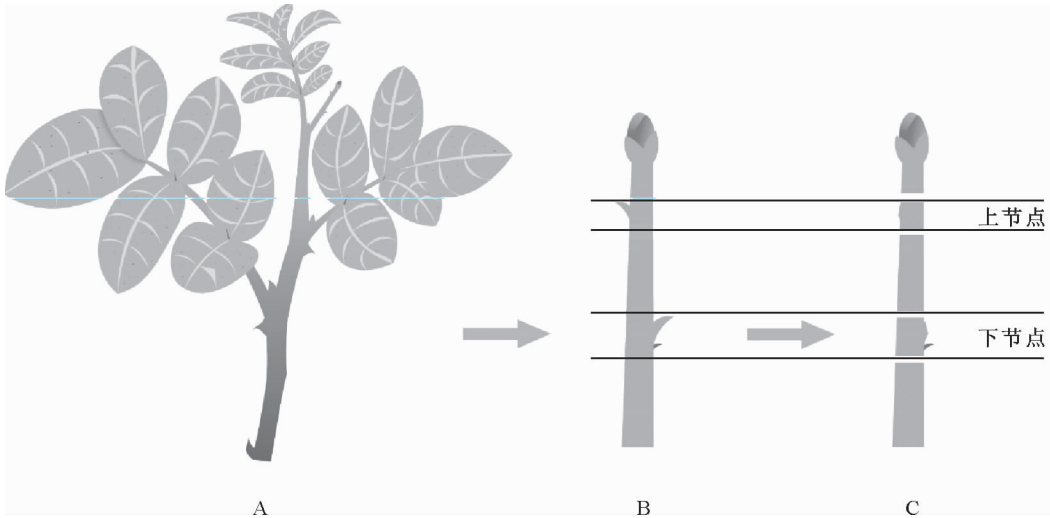
调工作形成<sup>[2]</sup>。植物的生长素在组织间甚至单细胞中常常是不对称分布的,这些浓度梯度的形成是通过膜结合转运蛋白极性分布实现<sup>[3]</sup>。理解生长素调控植物形态发生,首先需理解生长素如何被极性运输。在生长素极性运输的调控网络中有 *PINs*, *Aux/IAAs*, *ARFs* 等多个基因家族参与,目前已有多种假设模型<sup>[4-5]</sup>。这些模型可以预测数个植物发育过程中 PIN 蛋白的行为,PIN 蛋白的极性分布直接导致生长素的不对称分布。简而言之,细胞极性分化是花椒皮刺的分化的第一步,找到建立细胞极性的基因是培育无刺花椒的关键。本研究拟获取花椒茎尖节点的转录本序列,通过公共数据库对基因

序列的注释,获取大量可识别的序列信息,为后续寻找花椒皮刺分化的关键基因提供丰富、可靠的基础信息。

## 1 材料与方法

### 1.1 材料

以长势良好的 7 年生‘大红袍’花椒为材料。在生长季节选嫩枝尖端皮刺即将分化的部位(图 1A 尖端),取皮刺未分化的上节点及皮刺已分化的下节点各 20 个(去幼叶后取其节点,图 1C)。将采集的样品迅速置于小型液氮罐中速冻,运至实验室于-80℃保存,用于提取组织总 RNA。



A. 大红袍嫩茎示意图,B. 嫩茎尖端上节点有嫩叶长出但无皮刺分化,下节点的嫩叶基部分化出幼小皮刺,C. 将上节点及下节点的嫩叶去除。

图 1 花椒茎尖节点样品采集示意

Fig. 1 Sampling schematic of node of *Z. bungeanum* setm tip

### 1.2 方法

1.2.1 RNA 的提取及分离 使用联川生物公司提供的总 RNA 提取试剂盒(TRK-1001)提取样品总 RNA。此提取试剂盒采用离心柱层析法,使用专利的分离介质。提取过程中,RNA 可优先从其他细胞成分中分离纯化后得到总 RNA。利用带有 Oligo-dT 的介质与 mRNA 的 Poly(A)尾结合的原理,从总 RNA 中分离 mRNA。

1.2.2 cDNA 测序文库的构建 通过二价金属离子将 mRNA 样品片段化,回收 200~300 bp 的片段进行反转录反应,用随机引物进行反转录,反转录后消化 RNA,并进行第二链 cDNA 的合成,得到双链

cDNA;后进行末端修复,利用 T4 DNA 聚合酶和 Klenow DNA 聚合酶活性将 DNA 的粘性末端转变成平末端。用相应的酶在 3'末端加一个腺嘌呤后,将与测序仪 flow cell 中锚定的序列和测序引物相匹配的接头连接到 DNA 片段的两端,进行进一步纯化和扩增。

1.2.3 文库质检与上机测序 构建好的转录组测序文库用 Agilent 2100 Bioanalyzer 质检合格后,使用 Illumina HiSeqTM 2500 高通量测序仪进行双末端 2 × 100 bp 测序。

1.2.4 信息分析

1.2.4.1 信息分析流程 信息分析流程见图 2。

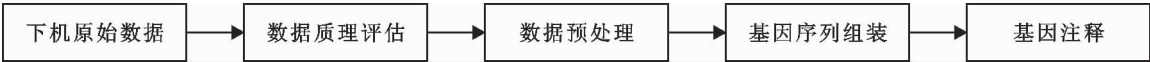


图 2 信息分析流程

Fig. 2 Process of information analysis

1.2.4.2 下机原始数据 测序得到的原始荧光图像数据经计算机识别转化为碱基序列信息,结果以 FASTQ 文件格式存储,存储的原始数据称之为 raw reads。在 FASTQ 格式文件中每个 read 由 4 行描述:

- 1)序列标识以及相关的描述信息,以‘@’开头;
- 2)第 2 行是序列;
- 3)第 3 行以‘+’开头,后面是序列标示符、描述信息,或者空白;
- 4)第 4 行是质量信息,与第 2 行的序列相对应,每一个序列都有一个质量评分,根据评分体系的不同,每个字符的含义表示的数字也不相同

例如:

```
@ILLUMINA-08A740:1:FC650CDAAXX:5:1:1238:1018 1:N:0:GCCAAT
ACATTCGGTCTGAATGCTACAT-
GAAGAACATAAGCC
+
#122166666C@@@C@@@CCC@@@
@@@@# # # # #
```

1.2.4.3 数据质量评估 由于测序过程中化学试剂的消耗,测序错误率随着测序序列的读长的增加而升高,为 Illumina 高通量测序平台都共同的特征<sup>[6]</sup>。鉴于高通量测序错误率对结果的影响,需对原始数据进行质量评估。

1.2.4.4 数据预处理 数据预处理的步骤:

- 1)去除带接头的 reads;
- 2)去除 N 的比例>5%的 reads(N 表示无法确定的碱基);
- 3)去除低质量 reads(质量值  $Q \leq 10$  的碱基数占整个 read 的 20 以上);
- 4)获取纯净的可用数据用于后续分析

1.2.4.5 基因序列组装 使用短 reads 组装软件 Trinity<sup>[7]</sup>拼装过滤后的纯净数据。

Trinity 由 3 个模块组成,组装步骤如下:

1)Inchworm 构建 k-mer 库,  $K=25$ 。过滤低频 k-mer 选择最高频度的 k-mer 作为种子,用来 Contig 组装。以 k-mer 间 overlap 长度等于 k-1 对种子进行延伸,直到不能再延伸,形成线性 Contig。

2)Chrysalis 把可能存在可变剪切及其他平行基因的 Contigs 聚类。每个 Contig 集定义成一个 Component,对每个 Component 构建 de Bruijn graphs。拿 reads 验证,看每个 Component 的 reads 支持情况。

3)Butterfly 合并 de Bruijn 图中有连续节点的线性路径,以形成更长的序列。剔除可能由于测序错误的分叉,使边均匀。用动态规划算法打分,鉴定被 reads 和 read pairs 支持的路径,剔除 reads 支持少的路径。

Trinity 拼接结果为 Transcript,以 FASTQ 形式储存。用 TGICL<sup>[8]</sup>和 Phrap 软件对这些转录本进行同源聚类 and 拼接后得到长度大于 200 bp 非冗余的 unigene。

1.2.4.6 Unigene 功能注释 使用 BLAST 比对工具将样本 unigene 与公共数据库进行比较,通过基因的相似性进行功能注释。将样本基因序列,分别与 Swiss-Prot、NR、KEGG、KOG 和 Pfam 数据库进行比对,以  $e\text{-value} < 10^{-5}$  为阈值获取注释,合并基因得到的所有注释详细信息。

2 结果与分析

2.1 RNA 质量评估

将提取的总 RNA 上机进行质量检测。根据质检报告,共得到 74.94  $\mu\text{g}$  总 RNA。取 40.00  $\mu\text{g}$  总 RNA,使用带有 Oligo-dT 的介质分离得到 mRNA 2.03  $\mu\text{g}$ 。剩余的总 RNA 置于  $-80^{\circ}\text{C}$  冰箱保存,用于后期实时荧光定量 PCR。

2.2 数据产量统计

原始数据的统计(表 1)表明,数据留存率非常高,数据质量较好,最终获得样本有效数据量 6.12G,满足后续数据分析要求。

表 1 数据产量统计

| 原始数据       |               |            |               | 有效数据    |
|------------|---------------|------------|---------------|---------|
| Read       | Base          | Read       | Base          |         |
| 56 242 196 | 5 624 219 600 | 56 242 196 | 5 624 219 600 | 100.00% |

注:表 1 数据量均是 Read1+Read2 的结果。

2.3 数据质量评估

使用 fastQC 软件对原始数据进行质量评估(图 2),横坐标表示测序序列的碱基位置,纵坐标表示碱基测序错误率。 $Q_{20}$ 反映了数据的质量,说明测序结果中由于测序仪器造成的错误概率<1%的碱基所占的比例。 $Q_{30}$ 则表示错误概率<0.1%的碱基所占的比例。其中  $Q_{30}$  比率>99%,表示本次测序结果错误率非常低。

2.4 组装结果分析

2.4.1 De novo 拼接结果统计

2.4.2 Unigene 分布统计

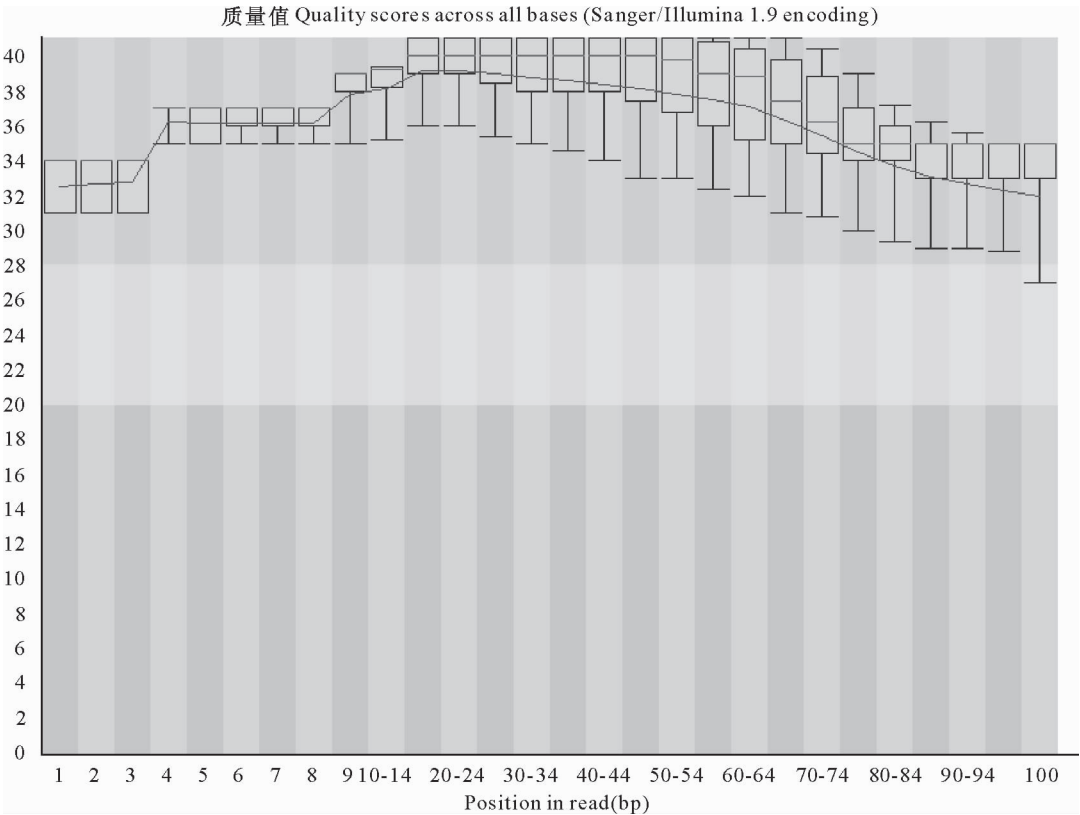


图 3 单碱基序列质量分布

Fig. 3 Per base sequence quality

表 2 *De novo* 拼接结果统计

Table 2 Statistics of *de novo* assembly results

|            | 总条数    | 最短长度 | 中值  | 平均长度 | N50 | 最大长度  | 总长         |
|------------|--------|------|-----|------|-----|-------|------------|
| Transcript | 65 337 | 201  | 419 | 627  | 874 | 6 750 | 41 006 030 |
| Unigene    | 45 057 | 201  | 407 | 610  | 846 | 6 750 | 27 517 689 |

注:设所有序列的长度加和为  $x$ ,将所有序列按长度从小到大排序,并从第 1 条序列开始累加计算总长,当长度达到  $x/2$  时对应的那条序列的长度即为该序列的 N50。

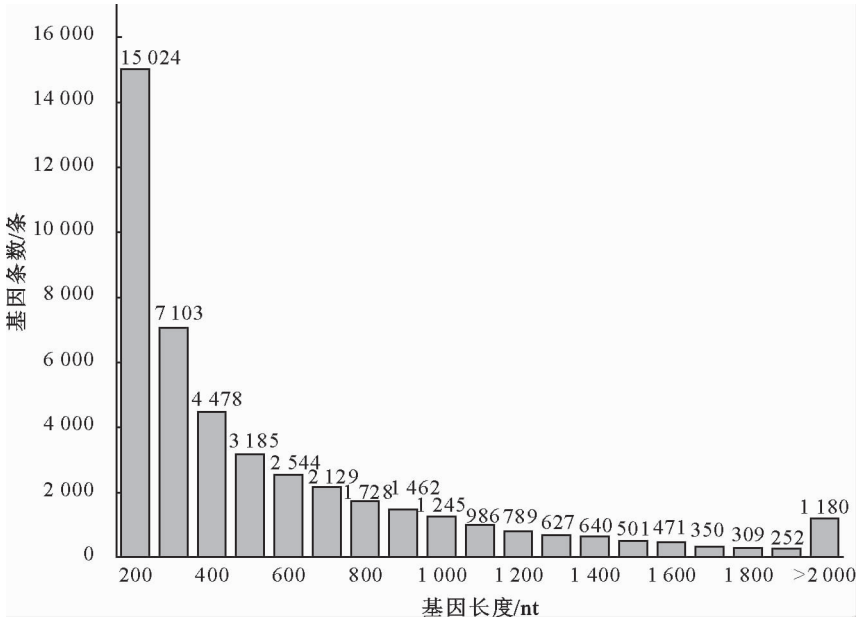


图 4 Unigene 长度分布统计

Fig. 4 Length distribution of unigene

## 2.5 Unigene 的功能注释

2.5.1 注释结果汇总 使用 BLAST 工具将拼装得到的 45 057 条 unigene 与 Swiss-Prot,NR,Pfam,KEGG,KOG 公共数据比对(阈值为  $e\text{-value} < 10^{-5}$ )。其中 80.34% 的 unigene 获得了 NR 数据库

注释,49.82% 获得了 Swiss-Prot 数据库的注释,40.23%,78.83% 和 70.26% 的序列分别得到了 Pfam,KEGG,KOG 数据库的注释(表 3)。

表 3 注释结果统计

Table 3 Summary of annotation results

| Unigene | Swiss-Prot | NR     | Pfam   | KEGG   | KOG    |
|---------|------------|--------|--------|--------|--------|
| 45 057  | 22 448     | 36 198 | 18 128 | 35 517 | 31 657 |
|         | 49.82%     | 80.34% | 40.23% | 78.83% | 70.26% |

2.5.2 物种比例分布 收集有 NR 注释的 unigene,并统计其物种信息(图 5)。可以看出,有 18.5% 的注释信息来源于葡萄(*Vitis vinifera*),18.1% 的信息来源于蓖麻(*Ricinus communis*),17.7% 来源于杨毛果(*Populus trichocarpa*),毛桃(*Prunus persica*)、野草莓(*Fragaria vesca*)、大豆(*Glycine max*)及其他分别占 12.1%、4.8%、4.6% 和 24.2%。

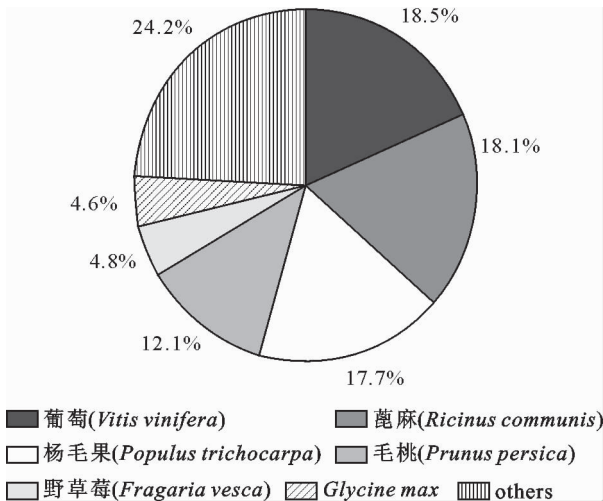


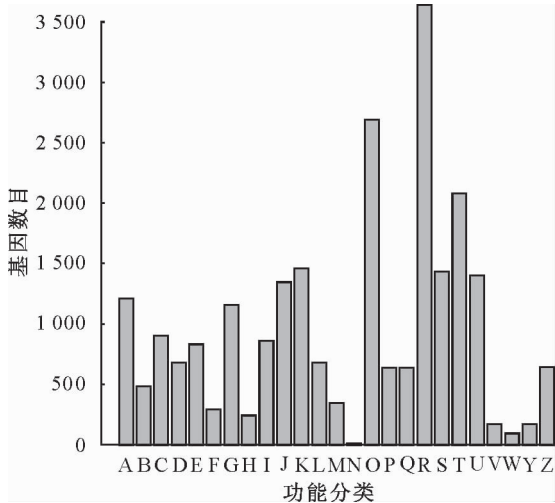
图 5 物种比例分布统计

Fig. 5 Species distribution

2.5.3 KOG 功能分类 (图 6)

## 3 结论与讨论

使用最新的二代高通量测序平台首次对大红袍花椒茎尖节点进行测序,经生物学软件组装,获得了 45 057 条转录本。使用 BLAST 比对公共数据库,80% 以上的序列被注释。通过检索注释结果,可以找到参与植物形态发生的基因,如响应生长素浓度的 ARF 基因家族和 Aux/IAA 基因家族以及参与极轴分化的 YABBY 基因家族等。其中,ARF 家族基因有 16 个,包括生长素响应激活因子 ARF5、ARF6、ARF7、ARF8、ARF19,生长素响应抑制因子 ARF1,ARF2,ARF3,ARF1,ARF2,ARF3,ARF4,ARF9,ARF10,ARF12,ARF17,ARF18,ARF22,



注:A. RNA 加工与修饰 B. 染色质结构与动力 C. 能量生产与转化 D. 细胞周期控制,细胞分裂,染色体分配 E. 氨基酸运输与代谢 F. 核苷酸运输与代谢 G. 糖运输与代谢 H. 辅酶运输与代谢 I. 脂运输与代谢 J. 翻译,核糖体结构和生物发生 K. 转录 L. 复制,重组与修复 M. 细胞壁,细胞膜与囊膜生物发生 N. 细胞运动 O. 翻译后修饰,蛋白质转换与分子伴侣 P. 无机离子的运输与代谢 Q. 次级代谢产物生物合成,运输与代谢 R. 一般功能(预测) S. 未知功能 T. 信号转导机理 U. 胞内运输,分泌与小泡运输 V. 保护机制 W. 胞外结构 X. 细胞核结构 Y. 细胞骨架。

图 6 花椒茎节点 unigene 的 KOG 功能分类

Fig. 6 KOG function classification of *Z. bungeanum* node of stem unigenes

ARF23<sup>[9]</sup>。这些响应因子在植物体中冗余表达,当生长素浓度较低时,Aux/IAAs 蛋白通过与生长素响应激活因子形成二聚结构,并且招聚辅阻遏蛋白 TPL,从而抑制 ARF 介导的转录活性,抑制下游基因的表达。同时,Aux/IAAs 蛋白也可与生长素响应抑制因子形成二聚体,影响其活性。当生细胞内的生太素浓度增高时,受体 TIR1/AFB 与 Aux/IAA 的引力增加,促进了 Aux/IAA 蛋白的降解。这样被释放的生长素响应激活因子就可以激活下游基因的转录<sup>[10]</sup>。

不同的 ARF 蛋白与不同的 Aux/IAA 蛋白互动,其对生长素浓度的敏感度也不同,开启或抑制的下游基因也不尽相同<sup>[10]</sup>。‘大红袍’花椒的皮刺在

嫩茎尖端分布相对较多,在复叶叶柄基部两端长有一对皮刺,复叶的两单叶叶柄间也常长有刺。嫩茎顶端产生较多生长素,幼嫩的叶片也可以产生生长素,产生的生长素向植株生物学下端运输。从外观观察,这些生长素易积累的位置与皮刺生长规律吻合。可以推测生长素响应激活因子可能参与激活了早期皮刺分化的关键基因。

另外,在转录组数据中找到了 3 个 *YABBY* 基因,分别为 *YAB1*,*YAB2* 和 *YAB4*。*YABBY* 基因家族是一类含有 C<sub>2</sub>C<sub>2</sub> 锌指结构域和 *YABBY* 结构域的转录因子<sup>[11]</sup>。*YABBY* 转录因子参与多个生物过程,其中包括叶与花的形态发生,器官生长和顶尖分生组织的活性控制<sup>[12]</sup>。虽目前仍不清楚这些基因的内在联系,但根据现有报道可以推断 *YABBY* 基因与器官极轴分化有密切的关系<sup>[12-13]</sup>。

随着人们对植物形态发生认知的深入,花椒转录组数据将得到更好的解读。后续可使用原位杂交技术研究 *ARF* 和 *YABBY* 基因家族在皮刺分化过程中的表达情况,进一步筛选参与分化的关键基因。目前花椒再生体系已经建立<sup>[14-15]</sup>,可对重点基因进行干扰或敲除,观察植株生长状况,为培育良种无刺花椒提供依据。

参考文献:

[1] 中国科学院中国植物志编辑委员会. 中国植物志(43 卷 2 分册)[M]. 北京: 科学出版社, 1997.

[2] KARIN L, ANDERS Ö, LAETITIA L, *et al.* Developmental regulation of indole-3-acetic acid turnover in Scots pine seedlings[J]. *Plant Physiology*, 2001, 125(1): 464-475.

[3] KALIKA P, PANKAJ D. Polar auxin transport: cell polarity to patterning[M]// CHEN R J, BALU K F. Polar auxin transport. Springer Berlin Heidelberg, 2013:25-44.

[4] BENKOVA E, MICHNIEWICZ M, SAUER M, *et al.* Local, efflux-dependent auxin gradients as a common module for plant organ formation[J]. *Cell*, 2003, 115(5): 591-602.

[5] BHALERAO R P, BENNETT M J. The case for morphogens

in plants[J]. *Nat Cell Biol.*, 2003, 5(11): 939-43.

[6] ERLICH Y, MITRA P P, MCCOMBIE W R, *et al.* Alta-cy-clic: a self-optimizing base caller for next-generation sequencing[J]. *Nature Methods*, 2008, 5(8): 679-682.

[7] GRABHERR M G., HAAS B J, YASSOUR M, *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome[J]. *Nature Biotechnology*, 2011, 29(7): 644-652.

[8] PERTEA G, HUANG X, LIANG F, *et al.* TIGR gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets[J]. *Bioinformatics*, 2003, 19(5): 651-652.

[9] VERNOUX T, BRUNOUD G, FARCOT E, *et al.* The auxin signalling network translates dynamic input into robust patterning at the shoot apex[J]. *Mol. Syst. Biol.*, 2011, 7: 508.

[10] BIANCO M D, KEPINSKI S. Context, specificity, and self-organization in auxin response[J]. *Cold Spring Harbor Perspectives Biol.*, 2011, 3(1): 1-20.

[11] BOWMAN J L. The *YABBY* gene family and abaxial cell fate [J]. *Curr. Opin. Plant Biol.*, 2000, 3(1): 17-22.

[12] BONACCORSO O, LEE J E, PUAH L, *et al.* Filamentous flower controls lateral organ development by acting as both an activator and a repressor[J]. *BMC Plant Biol.*, 2012, 12: 176.

[13] SAWA S, ITO T, SHIMURA Y, *et al.* Filamentous flower controls the formation and development of arabidopsis inflorescences and floral meristems[J]. *Plant Cell*, 1999, 11(1): 69-86.

[14] 王港, 李周岐, 刘晓敏, 等. 花椒组织培养再生体系的建立 [J]. 西北林学院学报, 2008, 23(3): 117-119.

WANG G, LI Z Q, LIU X M, *et al.* Establishment of tissue culture regeneration system of *Zanthoxylum bungeanum* [J]. *Journal of Northwest Forestry University*, 2008, 23(3): 117-119. (in Chinese)

[15] 薛惠丹, 李周岐. 花椒愈伤组织 EMS 诱变及变异研究 [J]. 西北林学院学报, 2012, 27(2): 98-101.

XUE H D, LI Z Q. Variation and EMS mutagenesis in callus of *Zanthoxylum bungeanum* [J]. *Journal of Northwest Forestry University*, 2012, 27(2): 98-101. (in Chinese)