PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2019,46(11):1118~1129
www.pibb.ac.cn

拟南芥不同组织基因表达及可变剪接差异分析*

邢永强[1,2)**] 何泽学[1)] 刘国庆[1,2)] 蔡 禄[1,2)**]
([1)] 内蒙古科技大学生命科学与技术学院，包头 014010；[2)] 内蒙古自治区功能基因组生物信息学重点实验室，包头 014010）

摘要 可变剪接是转录后重要的基因表达调控方式，也是转录组和蛋白质组多样性的重要来源. 近年来随着拟南芥、水稻、玉米等植物转录组测序的完成，研究人员发现植物pre-mRNA可变剪接的发生与组织分化、发育等生物学过程密切相关. 本工作基于GEO数据库的RNA-seq数据，使用高通量测序数据分析常用的Trimmomatic、Salmon、DESeq2、SUPPA2等工具，识别了拟南芥的种子、根、叶、花、花梗、节间、长角果共7种组织的表达基因和可变剪接事件，以及7种组织间的差异表达基因和差异可变剪接事件，并以叶和花为例展示了相应的生物学功能分析. 该工作系统地研究了拟南芥基因表达和可变剪接发生的组织特异性，有助于进一步阐明植物基因组的基因表达调控机制.

关键词 拟南芥，差异表达基因，差异可变剪接，组织特异性
中图分类号 Q6，Q94，Q3 **DOI**：10.16476/j.pibb.2019.0139

可变剪接（alternative splicing，AS），也叫选择性剪接，是指从同一个mRNA前体中通过选择不同的剪接位点组合产生多个不同成熟mRNA的过程[1]. Pre-mRNA的可变剪接是转录后重要的基因表达调控方式，指数式的丰富了转录组和蛋白质组的多样性. 早期由于实验技术限制，人们只能对少数基因的可变剪接事件进行分析. 近年来，随着ChIP-seq和RNA-seq等测序技术的发展，研究人员发现在人类等动物基因组中，约95%的多外显子基因会发生可变剪接；在拟南芥、水稻、玉米等植物基因组中，约60%的多外显子基因会发生可变剪接[2]. 植物pre-mRNA可变剪接的发生与组织分化、发育等生物学过程密切相关. 冷、热、光等逆境胁迫也会导致转录组发生异常的可变剪接[3]. 基本的可变剪接方式主要包括内含子保留（retained intron，RI）、外显子跳跃（skipping exon，SE）、可变5'端（alternative 5' splice site，A5SS）、可变3'端（alternative 3' splice site，A3SS）、互斥外显子（mutually exclusive exon，MXE）、可变第一外显子（alternative first exon，AFE）、可变最后外显子（alternative last exon，ALE）等7种方式[4]. 内含子保留是拟南芥等植物基因组中最常见的可变剪接类型，约占拟南芥可变剪接事件的40%[5]. 本工作以模式生物拟南芥为研究对象，通过分析拟南芥的种子（seed）、根（root）、叶（leaf）、花（flower）、花梗（pedicel）、节间（internode）、长角果（pod）等7个组织的RNA-seq测序数据，系统研究了拟南芥基因表达和可变剪接发生的组织特异性.

1 材料与方法

1.1 数据集构建

2016年，Klepikova等[6]通过Illumina HiSeq 2000平台完成了拟南芥79个组织的转录组（RNA-seq）数据测定，数据的读段长度等于50 bp，读段类型为单端测序，测序结果提交到了NCBI数据库（PRJNA314076）. 我们下载了种子（长角果长到1.5 cm时采集）、根（7 d龄的根顶端组织）、叶（发芽后第12 d）、花（始花期）、花梗（始花期）、节间（始花期土壤到第一片叶组织之间）、长角果（长度为1.5 cm）共7种组织的RNA-seq原始测序

* 国家自然科学基金（61662055，61671256，31660322）、内蒙古自然科学基金（2018MS03024）和内蒙古自治区高等学校青年科技英才支持计划资助项目.
** 通讯联系人. Tel：86-472-5951944
蔡禄. E-mail：nmcailu@163.com
邢永强. E-mail：xingyongqiang1984@163.com
收稿日期：2019-06-24，接受日期：2019-09-10

数据（raw data），每个组织有2个生物学重复，建立了包含14个RNA-seq测序样本的数据集（表1）.

Table 1　RNA−seq Dataset

Sra number	Tissue	Raw reads	Clean reads	Dropped rate/%	Alignment rate/%
SRR3581878	Seed	34050503	31487845	7.50	94.43
SRR3581712	Seed	35342896	32136187	9.07	91.91
SRR3581836	Root	25220183	23927918	5.12	98.49
SRR3581356	Root	24251789	23244475	4.15	98.73
SRR3581676	Leaf	25075794	23895262	4.71	98.71
SRR3581843	Leaf	24332850	23218893	4.58	98.72
SRR3581865	Flower	26980183	26314071	2.47	98.93
SRR3581699	Flower	27036642	26363924	2.49	98.85
SRR3581703	Pedicel	23954061	22998146	3.99	98.34
SRR3581869	Pedicel	32365293	31051798	4.06	98.40
SRR3581705	Internode	24020644	23243999	3.23	98.80
SRR3581871	Internode	23678087	22761633	3.87	98.70
SRR3581876	Pod	34255917	32274538	5.78	98.71
SRR3581710	Pod	32012167	30634493	4.30	98.68

使用NCBI提供的SRA Toolkit工具包的fast-dump（2.9.2），将sra格式的原始测序数据转化为可读的fastq格式. 表1显示14个样本的平均测序深度约28.0 million，满足基于RNA-seq测序数据进行基因表达分析的数据深度要求.

1.2　质控分析

首先，使用高通量测序数据质量评估的常用工具FastQC（v0.11.8）查看了数据集的数据质量[7]，结果显示raw data中存在一定的冗余读段. 为剔除这些低质量读段，使用Trimmomatic（0.38）[8]处理raw data（参数设置：leading=3；Trailing=3；SLIDINGWINDOW=4：15；MINLEN=50 bp，其他参数使用默认值），得到可用于下游分析的干净数据（clean data）（表1），平均深度约26.8 million. 图1以种子的SRR3581878样本为例展示了clean data的读段质量分析结果. 可以发现，读段每个位点的碱基质量分数Q值均大于30，即每个位点的测序准确率均大于99.9%.

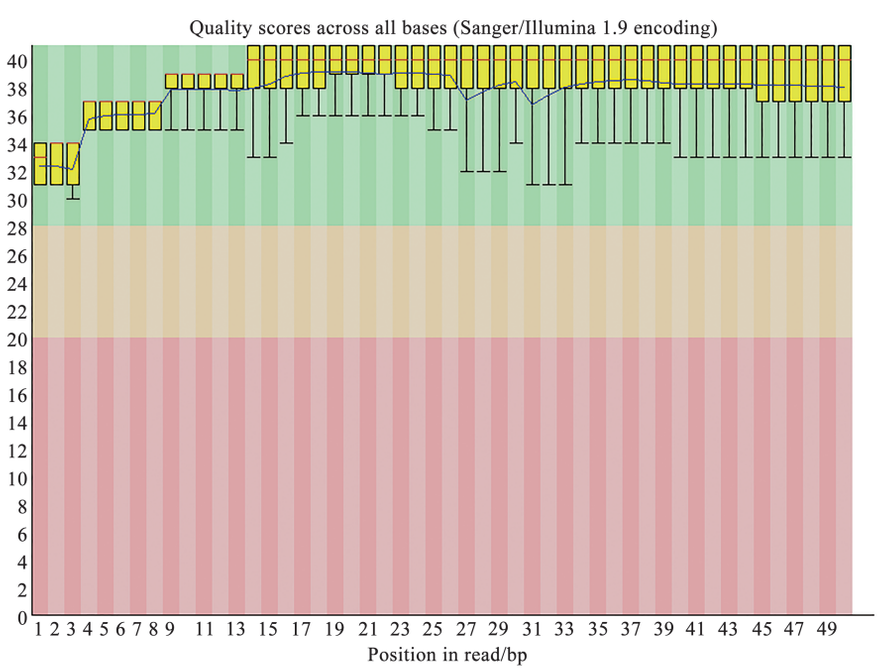

Fig. 1　Average sequencing quality per nucleotide site of the sample SRR3581878

The *x*-axis denotes the nucleotide site of read and the *y*-axis denotes the base quality. The top and lower edges of yellow rectangle represent upper and lower quartiles of base quality at specific site. The blue line represents mean of base quality.

1.3　转录本及基因的表达水平定量化

使用Patro等开发的Salmon[9]工具进行转录本的定量化. Salmon（v0.12.0）是一款不需要序列比对就可以快速完成转录本定量的RNA-seq数据分析工具，主要以TPM值（transcripts per million）描述转录本的表达量. 它的使用流程包括基于转录组fasta序列文件建立索引、基于fastq格式的测序文件进行转录本定量. 2017年，Zhang等[10]构建了拟

南芥的注释文件（AtRTDv2_QUASI_19April2016.gtf）及相应的转录组fasta格式序列文件（AtRTDv2_QUASI_19April2016.fa）[10]. 该注释文件共包含34 212个基因及对应的82 190个转录本，远大于Araport数据库提供的转录本数量（58 699）[11]. 因此，我们使用AtRTDv2_QUASI_19April2016.fa作为参考转录组建立索引. 因为读段长度小于75 bp，建立索引时的-k参数设置为25；同时打开keepDuplicates参数. 转录本定量分析时文库类型选择参数设置为A；打开seqBias和gcBias参数进行序列特异性和GC含量偏性校准. 以每个样本的转录本定量分析输出文件作为R包tximport（1.10.0）的输入文件，得到由14个样本的基因读段数（count）构成的基因表达矩阵[12].

1.4 差异表达基因的鉴定

使用基于负二项分布模型的DESeq2工具（R包，v1.22.1）进行拟南芥不同组织差异表达基因的鉴定[13]. DESeq2筛选差异表达基因的步骤主要包括：a. 输入已经获得的基因表达矩阵（行名为样本名，列名为基因名）；b. 设置分组信息以及构建dds对象；c. 使用DESeq函数估计离散度，并进行标准化，得到res对象结果；d. 设定阈值，本工作选取$padj < 0.01$且$\log_2 FC > 1$（FC表示基因表达水平差异倍数）的基因为上调基因，选取$padj < 0.01$且$\log_2 FC < -1$的基因为下调基因.

1.5 差异可变剪接事件的鉴定

目前，已有SUPPA2等多款识别可变剪接事件的工具[4]. SUPPA2擅长定量化注释文件中已有的可变剪接事件，本工作只关注拟南芥注释文件中已存在的可变剪接事件，所以采用了流行的SUPPA2进行可变剪接的定量化分析. SUPPA2是一款鉴定多种条件下差异可变剪接事件的工具[14]. 我们使用该工具鉴定了拟南芥7个组织中的5类可变剪接事件，并进一步识别了不同组织间的差异可变剪接事件. SUPPA2（2.3）识别差异可变剪接事件时主要包括以下几个步骤：a. generateEvents. 基于GTF注释文件的外显子信息识别可变剪接事件. b. psiPerEvent. 基于第一步获得的记录可变剪接事件的文件和Salmon提供的转录本丰度值（TPM）文件，计算每个样本中每个可变剪接事件的包含率（PSI值）. PSI的取值范围介于0和1之间. 统计分析$0 < PSI < 1$的可变剪接事件，可鉴定出各个组织中存在的可变剪接事件数. c. diffSplic. 以转录本丰度值文件以及第一步和第二步生成的结果文件作为输入信息，产生描述不同组织间差异可变剪接事件的文件. 其中ΔPSI表示不同组织间可变剪接事件包含率的差异（$\Delta PSI = PSI_2 - PSI_1$）；$P$-value描述差异显著性程度，本工作选取$\Delta PSI > 0.1$且$P$-value < 0.05的差异可变剪接事件为有效差异事件.

1.6 基因富集分析

使用Yu等[15]开发的功能较为强大的R包clusterProfiler（v3.10.1）对鉴定的差异表达基因及发生差异可变剪接事件的基因进行GO和KEGG富集分析. 在进行富集分析时需在R环境下上载Bioconductor上提供的拟南芥注释信息（org.At.tair.db），该工具提供了友好的可视化分析工具.

2 结果与讨论

2.1 样本数据质量分析

使用FastQC和Trimmomatic工具对拟南芥不同组织的RNA-seq raw data进行质控分析，得到平均深度约26.8 million的clean data用于下游分析（表1）. 基于fastq格式的clean data和拟南芥转录组序列文件，使用Salmon和tximport工具得到样本的转录本（82190×14维）和基因表达矩阵（34212×14维）. 基于基因表达矩阵，对7种拟南芥组织（14个样本）进行聚类分析（图2a）. 可以看出所有样本的组间差异大于组内差异，每个组织的2个重复样本均可以聚在一起，7个不同的组织可以明显分开，说明样本收集和测序数据质量的可靠性都很高，可用于不同组织的差异表达基因和差异可变剪接分析. 图2a也显示生殖器官种子和花可以与营养器官叶、节间、根等显著区分. 作为连接营养器官和生殖器官的花梗组织基因表达图谱与叶组织更接近. 另外，通过观察图2a对行（基因）的聚类结果，不难发现不同的组织存在特异高表达基因集（图2a中的红色条带），也存在一些在某几种组织中表达量较高，而在其他组织低表达的基因集. 这些基因集的识别和功能分析将极大地促进拟南芥的组织特异性分化和发育机理的解读，这也是我们日后开展植物组织分化研究的工作内容之一.

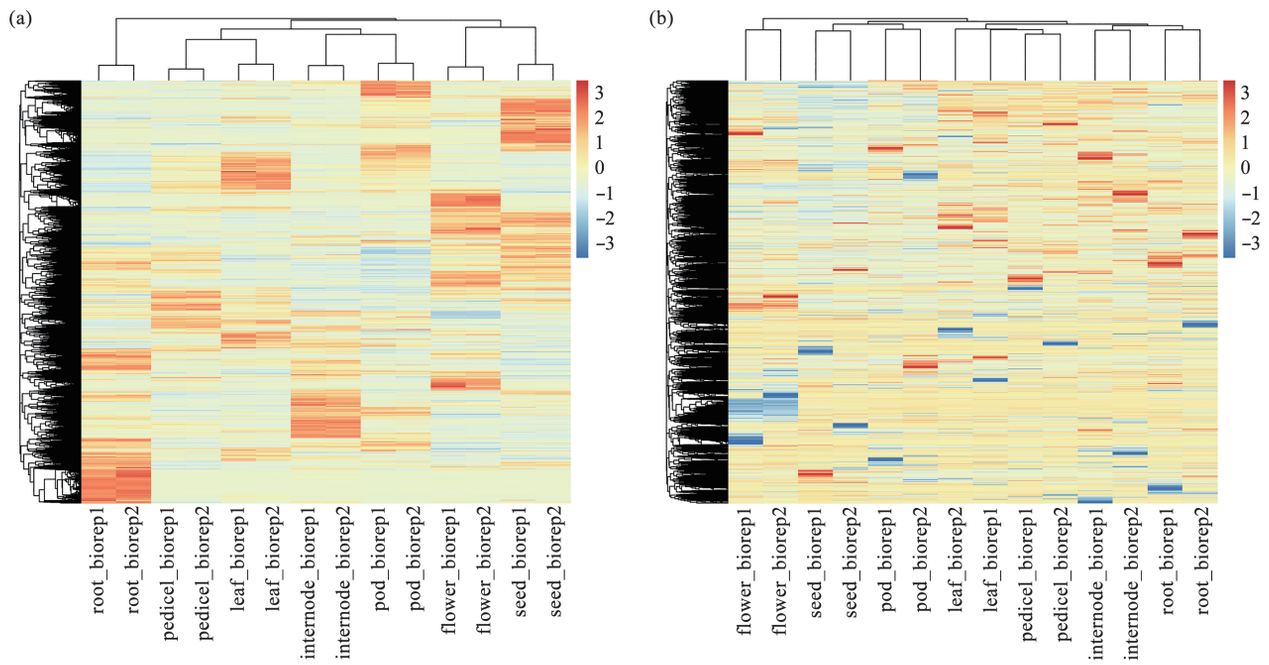

Fig. 2　Clustering of samples in *Arabidopsis* based on gene expression level（a）and inclusion level of AS（b）

2.2　组织间差异表达基因的鉴定

拟南芥 AtRTDv2_QUASI 注释文件中包含 33 536 个重叠于 Araport 的 gtf 注释文件的基因，其中编码基因共 27 566 个. 以至少在一个样本中的基因表达值大于 0 为筛选条件，共识别了 29 640 个基因，其中编码基因 26 581 个. 共 18 920 个基因在所有样本中均表达，其中编码基因 18 325 个. 叶组织的基因表达数最小（22 885），花器官的基因表达数最大（25 954）.

以基因表达矩阵作为 DESeq2 的输入文件，鉴定了拟南芥 7 种组织间的差异表达基因（表 2）. 结果显示，在 **1.4** 节规定的阈值下鉴定的组织间上调基因和下调基因数量均大于 2 000，说明不同的组织间存在大量的差异表达基因，这些基因对组织的特异性发育发挥着重要作用. 叶组织和花器官间共识别了 9 280 个显著差异表达的基因，以此为例通过火山图直观地展示了二者之间的差异表达基因分布特征（图 3）. 可以看出，大多数差异表达基因的表达值变化在 1 000 倍之内. 表 2 显示：叶和花梗组织间的差异表达基因数最少，表明这两个组织的转录组差异较小；根和长角果组织间的差异表达基因数最多，表明两个组织间的转录组差异较大，客观地反映了营养器官和生殖器官的表达谱差异性. 该结论与下文 **3.1** 节的聚类结果相同，也与组织间形态学差异一致.

Table 2　Differentially expressed genes between different tissues of *Arabidopsis*

Tissue		Seed	Root	Leaf	Flower	Pedicel	Internode	Pod
Seed	up	0						
	down	0						
Root	up	5566	0					
	down	5347	0					
Leaf	up	5130	5300	0				
	down	5239	4564	0				
Flower	up	3636	5130	4649	0			
	down	4455	5239	4631	0			
Pedicel	up	4735	5205	2930	3834	0		
	down	4635	4354	2372	3247	0		
internode	up	4921	4052	3745	4336	3228	0	
	down	5358	4623	4013	4568	4179	0	
Pod	up	4565	5519	3089	4633	3439	3127	0
	down	5073	5904	4362	5350	4907	4230	0

“up” indicates that the expression level is up-regulated（column/row name）；“down” indicates that the expression level is down-regulated（column/row name）.

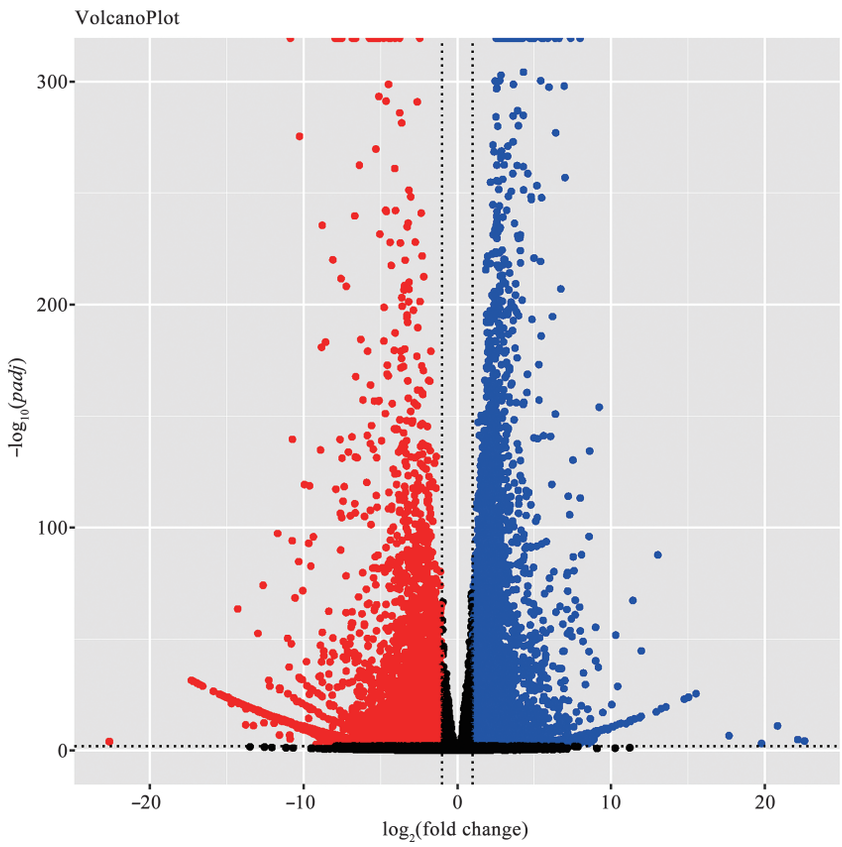

Fig. 3　Differentially expressed gene distribution between leaf and flower

The cutoff of up-regulated genes is *padj* < 0.01 & $\log_2 FC$ > 1 and the cutoff of down-regulated genes is *padj* < 0.01 & $\log_2 FC$ < −1. The *x*-axis represents $\log_2 FC$ and the *y*-axis represents $-\log_{10} padj$. The red dot denotes the down-regulated gene， the blue dot denotes the up-regulated gene and the black dot denotes the non-differentially expressed gene.

2.3　差异表达基因的功能分析

为进一步阐明差异表达基因的生物学功能，使用clusterprofiler进行差异表达基因的GO和KEGG功能富集分析.

2.3.1　差异表达基因GO富集分析

GO可以从GO-BP（生物学过程）、GO-MF（分子功能）、GO-CC（细胞组分）三个方面对输入的基因列表分别进行注释. 我们对7个组织间（共21种组合）的差异表达基因分别进行了GO富集分析. 考虑到文章的篇幅限制，这里仅以营养器官叶组织和生殖器官花组织间差异表达基因为例，介绍GO-BP富集分析结果. 表2显示在阈值为$|\log_2 FC|>1$且*padj*<0.01时，相比花器官，在叶组织中表达水平显著上调的基因共4 649个，显著下调的基因共4 631个. 为保证差异表达基因功能分析的可靠性，这里以阈值为$|\log_2 FC|>4$（即*FC*相差16倍）且*padj*<0.01，筛选得到极显著差异上调的基因551个、下调的基因1 666个. 将此基因列表作为Clusterprofiler包进行GO富集分析（*P*<0.01）的输入文件，得到的富集分析结果见图4和图5. 结果显示，551个极显著上调的基因共富集到系统性获得抗性（systemic acquired resistance）、水杨酸响应（response to salicylic acid）、过氧化氢代谢调控（regulation of hydrogen peroxide metabolic process）、免疫调控（regulation of immune system process）等共135个GO-BP词条. 系统获得性抗性是指植物的某个局部受到逆境因子侵染的时候，会产生逆境信号物质，在其他非侵染部位甚至植物整体和个体之间诱导产生对这种逆境的抗性机制. 该词条的显著富集将有助于植物叶组织抵抗外部侵染等逆境胁迫[16]. 研究表明，水杨酸与植物的光合作用、呼吸代谢、气孔关闭、抗逆性等生物学过程密切相关，该词条的显著富集进一步证实了水杨酸在双子叶植物叶组织中有重要的生物学功能[17]. 过氧化

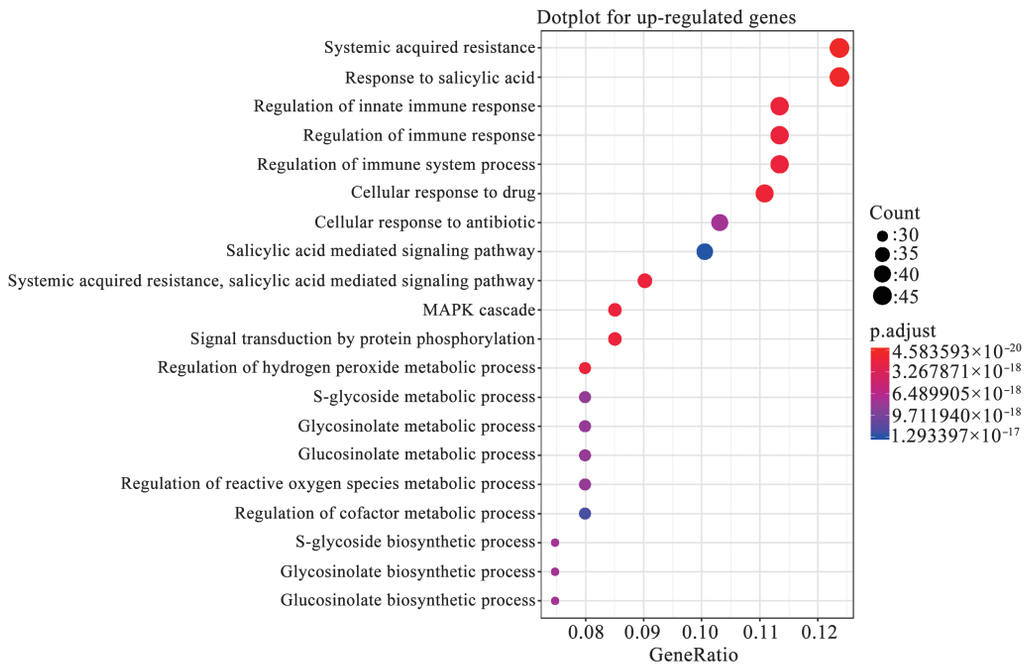

Fig. 4　The GO−BP functional enrichment of significantly up−regulated genes between leaf and flower

The *x*-axis is GeneRatio， which denotes the percentage of total DEGs in the given GO term. The *y*-axis is the description information of enriched GO-BP terms. The color of the point denotes the value of *p.adjust* and the size denotes the enriched number of differential genes under GO terms.

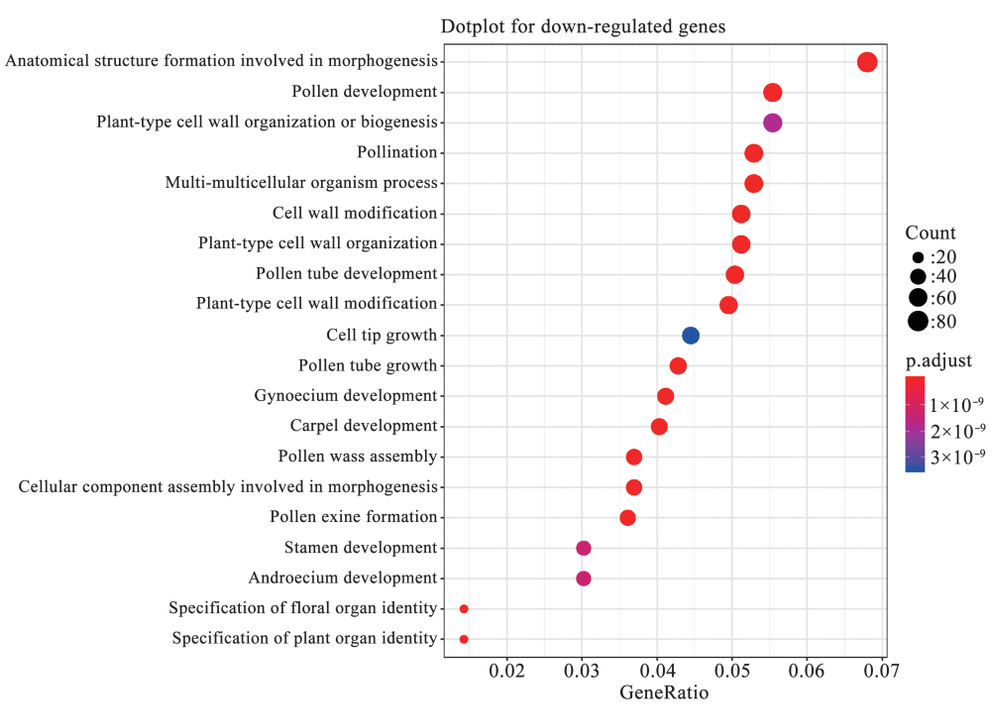

Fig. 5　The GO−BP functional enrichment of significantly down−regulated genes between leaf and flower

氢代谢是植物中清除活性氧的关键酶，也是植物耐受胁迫所必需的，其主要发生在叶绿体细胞器中，因此在叶组织中过氧化氢代谢相关显著富集[18]. 叶组织中存在大量的叶绿体，研究发现叶绿体能够

感知病菌侵染，并迅速地将信号传递给细胞核，促使植物建立免疫防御体系[19]. 免疫调控词条的显著富集也印证了叶绿体在植物免疫调控中的重要作用.

花器官中极显著上调（叶组织中下调）的1 666个基因富集到了花粉外壁形成（pollen exine formation）、花粉壁组装（pollen wall assembly）、花粉管发育（pollen tube development）、涉及形态发生的解剖结构形成（anatomical structure formation involved in morphogenesis）、花器官的形成（floral organ formation）等共63个GO-BP词条. 显然，富集的大多数词条均与花粉外壁形成、花粉壁形成、花粉管发育等花器官发育相关. 证实我们筛选的差异表达基因显著地反映了组织发育特异性特征. 为更明确地展示极显著差异表达基因的功能，分别构建了上述上调基因和下调基因前30个GO-BP词条的网络重叠图（附件图S1和S2）. 结果显示，在花器官中显著上调基因的GO-BP词条形成了与花粉管发育、花器官形态发生等相关的两个模块，清楚地展示了上调基因的生物学功能.

2.3.2　差异表达基因的KEGG通路富集分析

KEGG是一个整合了基因组、化学和系统功能信息的综合数据库，该数据库有助于把基因及表达信息作为一个整体的网络进行研究. 我们对7个组织间（共21种组合）的差异表达基因分别进行了KEGG富集分析. 这里仍以叶组织和花器官间的差异表达基因为例，介绍KEGG富集分析结果. 以筛选到的551个极显著上调基因和1 666个极显著下调基因分别作为Clusterprofiler包进行KEGG富集分析（P<0.01）时的输入文件，得到的富集结果见图6. 极显著上调的基因共富集到硫代葡萄糖苷生物合成（glucosinolate biosynthesis）、2-氧代羧酸代谢（2-oxocarboxylic acid metabolism）两条KEGG通路. 硫代葡萄糖苷是十字花科植物中重要的次生代谢物，而且在幼嫩的叶片、枝芽、种子中硫代葡萄糖苷生物合成活性较高，伴随植物组织的成熟合成能力变弱[20]. 我们选取的组织为幼嫩的叶片，所以该通路的显著富集进一步证实了硫代葡萄糖苷合成在植物叶组织中有重要的生物学功能.

极显著下调的基因富集到了角质、小檗碱和蜡生物合成（cutin， suberine and wax biosynthesis）、苯丙烷类生物合成（phenylpropanoid biosynthesis）、戊糖和葡萄糖醛酸的相互转化（pentose and glucuronate interconversions）3条KEGG通路. 植物苯丙烷类物质在植物中普遍存在，有数千种不同的化学结构形式，包括总黄酮、黄酮醇、香豆素、木质素、花青素等苯类化合物，这些化合物对植物发育以及植物逆境胁迫中的应答发挥重要作用[21].

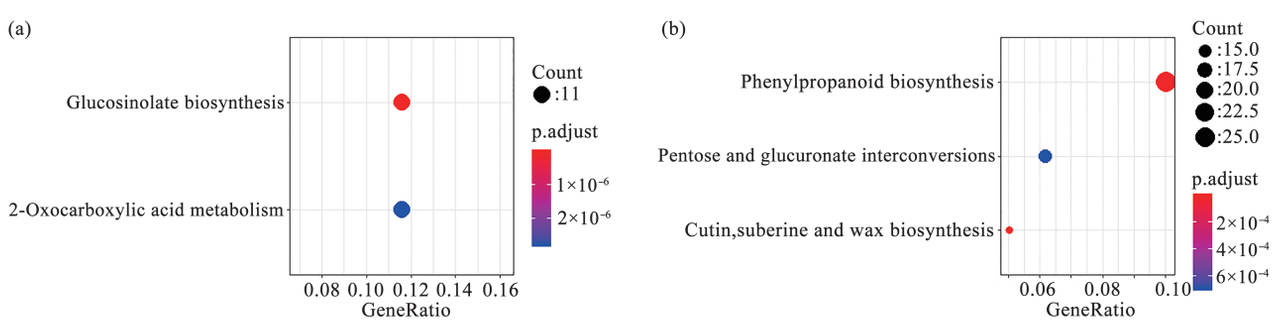

Fig. 6　KEGG enrichment analysis of significantly up-regulated（a）and significantly down-regulated genes（b）between leaf and flower

2.4　可变剪接事件鉴定及差异可变剪接分析

植物pre-mRNA可变剪接与组织分化、发育等生物学过程密切相关，并且冷、热、光等逆境胁迫也会导致转录组发生异常的可变剪接. 我们使用SUPPA2工具鉴定了拟南芥花、叶等7种组织中的内含子保留、外显子跳跃、可变5'、可变3'、互斥外显子共5类可变剪接事件. 进一步，使用SUPPA2的diffSplice模块识别了7种组织间的差异可变剪接事件，并针对发生差异可变剪接事件的基因进行了生物学功能分析.

2.4.1　不同组织的可变剪接事件鉴定

使用SUPPA2的generateEvents模块识别了AtRTDv2_QUASI_19April2016.gtf注释文件中存在的内含子保留等5类可变剪接事件；结合转录本丰

度值，使用SUPPA2的psiPerEvent模块计算每个样本中每个可变剪接事件的包含率（*PSI*值）. 通过统计0<*PSI*<1的可变剪接事件，鉴定了各个组织中发生的各类可变剪接事件数（详见**1.5**节）. 图7列举了识别的拟南芥7种组织中存在的各类可变剪接事件频数. 可以看出：内含子保留在5种可变剪接类型中占比最高（7种组织中的平均占比为49.7%）；而外显子跳跃占所有可变剪接事件的比率较低（7种组织中的平均占比为4.6%）；可变5'、可变3'在7种组织中占所有可变剪接事件的平均比率分别为16.0%和29.5%；互斥外显子的发生频率很低，可能与互斥外显子真实的发生频率较低以及互斥外显子的结构较复杂，造成SUPPA2的识别率不准确相关. 该结论与文献报道的植物基因组可变剪接频率分布特征基本一致. 由图7也可以看出各个组织中发生的各类可变剪接事件数中，节间组织中发生了数量最多的可变剪接事件（32 919），而在种子组织中发生了数量最少的可变剪接事件（27 384），从可变剪接的频数角度反映了不同组织的转录图谱复杂性. 值得一提的是，本工作也基于不同组织可变剪接事件的包含率（*PSI*值）对样本和基因进行了聚类分析（图2b），结果显示基于*PSI*值可以实现对组织内样本的正确识别，对组织间的聚类结果也与图2a基本一致. 该结果表明除基因表达水平外，pre-mRNA的可变剪接也是拟南芥组织特异性的重要表征. 图2b的行（可变剪接事件）聚类结果显示，不同的组织明显存在特异的高包含率和低包含率可变剪接事件集（图2b中的红色和蓝色条带），表明可以通过调节可变剪接异构体的包含率实现对组织特异性的调控. 今后开展植物可变剪接的研究将重点关注此类可变剪接事件集.

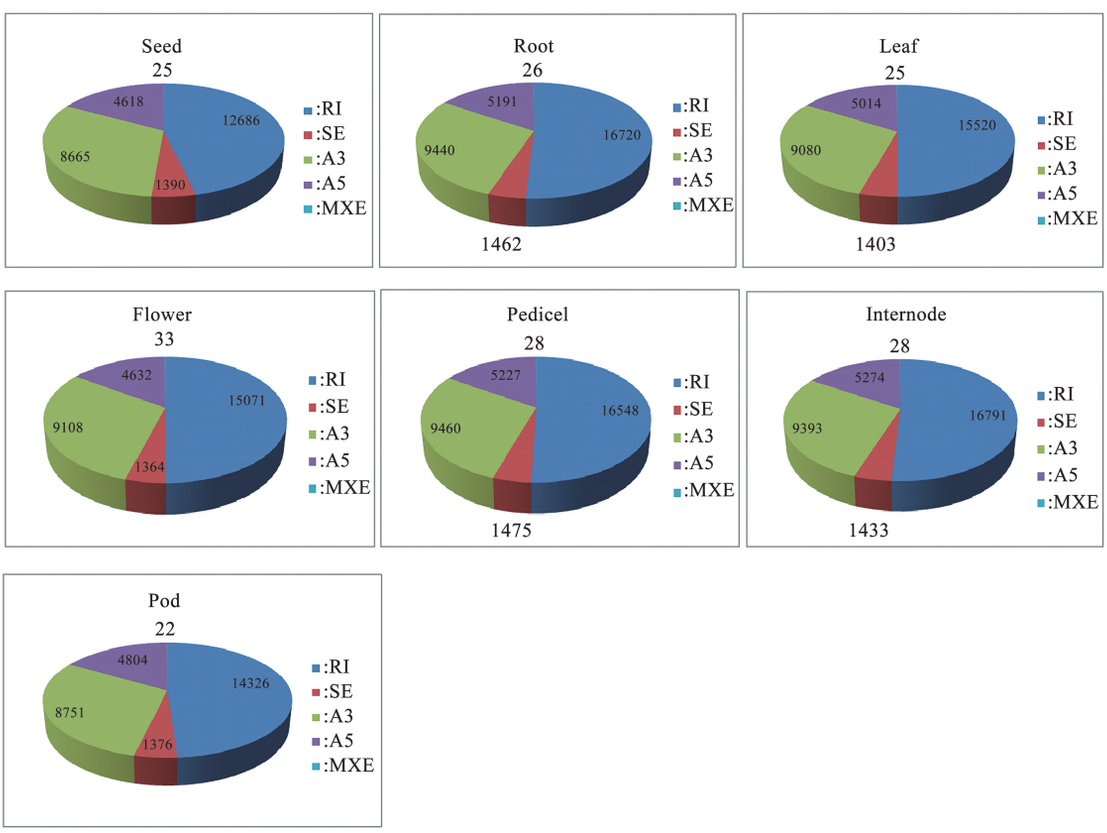

Fig. 7　Distribution of alternative splicing events in seven tissues

RI：Retained intron；SE：Skipping exon；A5：Alternative 5' splice site；A3：Alternative 3' splice site；MXE：Mutually exclusive exon.

2.4.2　不同组织间差异可变剪接事件鉴定

使用 SUPPA2 的 diffSplice 模块，以 Δ*PSI*>0.1 且 *P*-value<0.05 为阈值鉴定不同组间的差异可变剪接事件. 共得到 7 种组织之间（21 种组合）的差异可变剪接事件（表 3）. 其中花和花梗之间鉴定出的差异可变剪接事件最多（2 510），而节间和长角果之间鉴定出数量最少的可变剪接事件（468）. 不难发现营养器官与生殖器官间的差异可变剪接事件丰度显著高于营养器官或生殖器官内部不同组织间的差异可变剪接事件丰度. 另外，也统计了发生差异可变剪接事件的基因数量，约 18% 的基因中存在两个或两个以上的差异可变剪接事件，基因数量变化规律与差异可变剪接事件数变化趋势一致.

为进一步阐明发生差异可变剪接事件基因的生物学功能，我们使用 Clusterprofiler 对不同组织间发生差异可变剪接事件的基因进行了 GO 功能富集分析. 以叶组织和花器官间的差异可变剪接事件为例介绍富集分析结果，将叶和花组织间发生差异可变剪接事件的 736 个基因作为 Clusterprofiler 包进行 GO 富集分析的输入文件，以 *P*<0.05 为阈值，得到的富集分析结果见图 8. 分别富集到了四萜类生物合成过程（tetraterpenoid biosynthetic process）、类胡萝卜素生物合成过程（carotenoid biosynthetic process）、四萜代谢过程（tetraterpenoid metabolic process）、3-磷酸甘油醛代谢过程（glyceraldehyde-

Table 3　Frequency of differential alternative splicing events between seven tissues

	RI	SE	A3	A5	MXE	Sum*	Gene Num
Seed-Root	990	12	193	151	0	1346	1106
Seed-Leaf	1246	26	188	150	0	1610	1290
Seed-Flower	1065	20	155	126	0	1366	1129
Seed-Pedicel	1374	25	194	148	0	1741	1370
Seed-Internode	1264	30	202	159	1	1656	1318
Seed-Pod	649	25	153	110	0	937	782
Root-Leaf	690	24	158	140	1	1013	829
Root-Flower	667	22	152	115	1	957	811
Root-Pedicel	728	31	168	145	2	1074	848
Root-Internode	656	28	138	147	1	970	770
Root-Pod	391	30	156	117	1	695	589
Leaf-Flower	566	24	162	91	0	843	736
Leaf-Pedicel	395	17	125	81	1	619	523
Leaf-Internode	496	17	142	87	1	743	610
Leaf-Pod	281	17	114	73	0	485	421
Flower-Pedicel	1488	158	474	388	2	2510	1895
Flower-Internode	1399	137	477	354	2	2369	1841
Flower-pod	1006	135	398	321	1	1861	1471
Pedicel-Internode	430	15	122	79	1	646	560
Pedicel-Pod	277	22	110	96	0	505	434
Internode-Pod	278	24	101	65	0	468	397

* Sum represents the sum of five types of differentially alternative splicing events；Gene Num represents the number of genes occurring differential alternative splicing events.

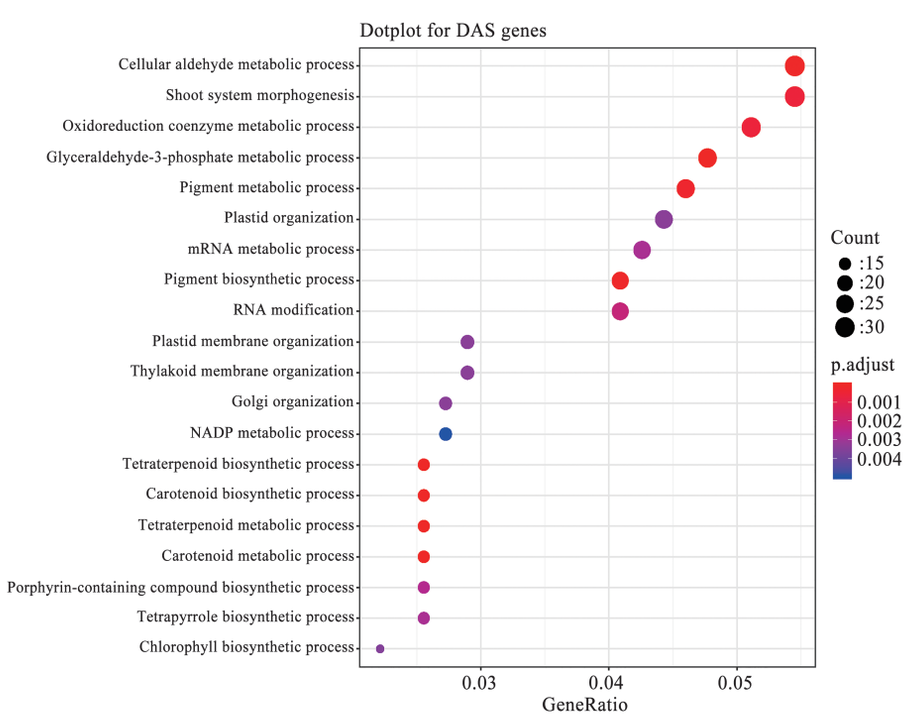

Fig. 8　GO−BP functional enrichment of genes undergoing differentially alternative splicing between leaf and flower

3-phosphate metabolic process）等92个GO-BP词条. 在植物中四萜类的胡萝卜素是植物光合作用中重要的色素，有吸收与传递光能以及抗氧化的作用[22]，四萜类生物合成途径对植物的生长发育有重要的调控作用. 该词条的富集说明可以通过差异可变剪接的方式调控该词条相关基因的表达，进而实现对叶组织或花器官组织特异性分化的调控. 同时也观察到92个GO-BP词条中还涉及花形态发生（flower morphogenesis）、叶形态发生（leaf morphogenesis）以及叶发育（leaf development）、心皮发育（carpel development）等生物学过程，证实可以通过差异可变剪接的调控方式实现基因组对拟南芥花器官发育的调控，进而证实了可变剪接在植物组织器官发育和分化过程中的重要性.

3　结　　论

随着组学测序技术的迅猛发展，拟南芥、水稻、玉米等植物的转录组测序已完成. 真核生物的转录和pre-mRNA的剪接相互耦合，但针对植物基因组的可变剪接分析的工作相对较少. 本工作基于GEO数据库的RNA-seq数据，识别了拟南芥的种子、根、叶、花、花梗、节间、长角果共7种组织的表达基因和可变剪接事件，进而鉴定了不同组织之间的差异表达基因和差异可变剪接事件，并进行了相应的生物学功能分析. 该研究工作的开展有助于阐明植物基因表达及可变剪接的组织特异性调控机制，将来拟系统分析多种植物的基因表达和可变剪接事件变化规律，并重点研究不同植物间的保守基因的基因表达和可变剪接事件特征.

附件　图S1，S2见本文网络版（http://www.CNKI.net或http://www.pibb.ac.cn）.

参　考　文　献

[1] 周新成, 王海燕, 卢 诚, 等. 植物功能基因选择性剪接研究进展. 热带农业科学, 2012, **32**(2): 36-41
Zhou X C Wang H Y, Lu C, *et al*. Chinese Journal of Tropical Agriculture, 2012, **32**(2): 36-41

[2] Jang Y H, Lee J H, Park H Y, *et al*. OsFCA transcripts show more complex alternative processing patterns than its *Arabidopsis* counterparts. J Plant Biol, 2009, **52**(2): 161-166

[3] Barbazuk W B, Fu Y, McGinnis K M, *et al*. Genome-wide analyses of alternative splicing in plants: opportunities and challenges. Genome Res, 2008, **18**(9): 1381-1392

[4] Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, *et al*. Intron retention is a major phenomenon in alternative splicing in *Arabidopsis*. Plant J, 2004, **39**(6): 877-885

[5] Marquez Y, Brown J W S, Simpson C, *et al*. Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. Genome Res, 2012, **22**(6): 1184-1195

[6] Klepikova A V, Kasianov A S, Gerasimov E S, *et al*. A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. Plant J, 2016, **88**(6): 1058-1070

[7] Kroll K W, Mokaram N E, Pelletier A R, *et al*. Quality control for RNA-Seq (QuaCRS): an integrated quality control pipeline. Cancer Inform, 2014, **13**(Suppl 3):7-14

[8] Bolger A M, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014, **30**(15): 2114-2120

[9] Patro R, Duggal G, Love M I, *et al*. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods, 2017, **14**(4): 417-419

[10] Zhang R, Calixto C , Marquez Y , *et al*. A high quality *Arabidopsis* transcriptome for accurate transcript-level analysis of alternative splicing. Nucleic Acids Res, 2017, **45**(9): 5061-5073

[11] Krishnakumar V, Contrino S, Cheng C Y, *et al*. ThaleMine: a warehouse for *Arabidopsis* data integration and discovery. Plant Cell Physiol, 2017, **58**(1):e4

[12] Soneson C, Love M I, Robinson M D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res, 2015, **4**:1521

[13] Love M I, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. 2014, Genome Biol, **15**(12):550

[14] Trincado J L, Entizne J C, Hysenaj G, *et al*. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol, 2018, **19**(1):40

[15] Yu G, Wang L G, Han Y, *et al*. ClusterProfiler: an R package for comparing biological themes among gene clusters. OMICS, 2012, **16**(5):284-287

[16] 张艳秋, 崔崇士. 植物系统获得性抗性研究进展. 东北农业大学学报, 2008, **39**(12): 113-117
Zhang Y Q, Cui C S. Journal of Northeast Agricultural University, 2008, **39**(12): 113-117

[17] Shen C, Yang Y, Liu K, *et al*. Involvement of endogenous salicylic acid in iron-deficiency responses in *Arabidopsis*. J Exp Bot, 2016,

67(14): 4179-4193

[18] Slesak I, Libik M, Karpinska B, *et al*. The role of hydrogen peroxide in regulation of plant metabolism and cellular signalling in response to environmental stresses. Acta Biochim Pol, 2007, **54**(1):39-50

[19] Lv R, Li Z, Li M, *et al*. Uncoupled expression of nuclear and plastid photosynthesis-associated genes contributes to cell death in a lesion mimic mutant. Plant Cell, 2019, **31**(1): 210-230

[20] Porter A J R, Morton A M, Kiddle G, *et al*. Variation in the glucosinolate content of oilseed rape (*Brassica napus L*.) leaves. Ann Appl Biol, 1991, **118**(2): 461-467

[21] Boudet A M. Evolution and current status of research in phenolic compounds. Phytochemistry, 2007, **68**(22-24): 2722-2735

[22] 王凌健, 方欣, 杨长青, 等. 植物萜类次生代谢及其调控. 中国科学:生命科学, 2013, **43**(12): 1030-1046
Wang L J, Fang X, Yang C Q, *et al*. SCIENTIA SINICA Vitae, 2013, **43**(12):1030-1046

Differential Analysis of Gene Expression and Alternative Splicing in Different Tissues of *Arabidopsis thaliana*[*]

XING Yong-Qiang[1,2)**], HE Ze-Xue[1)], LIU Guo-Qing[1,2)], CAI Lu[1,2)**]

([1)] *School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou* 014010, China;

[2)] *The Inner Mongolia Key Laboratory of Functional Genome Bioinformatics, Inner Mongolia University of Science and Technology, Baotou* 014010, China)

Abstract　Alternative splicing is crucial for post-transcriptional regulation and is responsible for transcriptome and proteome diversity. In recent years, with the completion of transcriptome sequencing of plants such as *Arabidopsis* thaliana, *Oryza sativa,* and *Maize*, researchers found that pre-mRNA alternative splicing in plant is involved with tissue differentiation and development, *etc*. In this work, RNA-seq data was downloaded from GEO database. Trimmomatic, Salmon, DESeq2, SUPPA2 and other tools were employed to detect expression genes and alternative splicing events in seed, root, leaf, flower, pedicel, internode and pod across the *Arabidopsis*. Then, differentially expressed genes and differential alternative splicing events were identified throughout 7 tissues. Furthermore, the comparison between leaves and flowers was taken as an example to display the corresponding biological functions. In this study, the tissue specificity of gene expression and alternative splicing in *Arabidopsis* was systematically investigated. It's helpful for elucidating gene expression mechanism in plant genome.

Key words　*Arabidopsis thaliana*, differentially expressed gene, differential alternative splicing, tissue specificity

DOI：10.16476/j.pibb.2019.0139

---

* This work was supported by grants from The National Natural Science Foundation of China（61662055，61671256，31660322），the Natural Science Foundation of Inner Mongolia（2018MS03024）and The Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region.

** Corresponding author. Tel：86-0472-5951944

CAI Lu. E-mail：nmcailu@163.com

XING Yong-Qiang. E-mail：xingyongqiang1984@163.com

Received：June 24，2019　Accepted：September 10，2019