G3, 2021, 11(8), jkab182
DOI: 10.1093/g3journal/jkab182
Advance Access Publication Date: 29 May 2021
Investigation

Full-length transcriptome analysis of pecan (*Carya illinoinensis*) kernels

Chengcai Zhang, Huadong Ren,* Xiaohua Yao,* Kailiang Wang, and Jun Chang

Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Hangzhou 311400, China

*Corresponding author: No. 73, Da Qiao Road, Fuyang, Hangzhou 311400, Zhejiang, China. Email: renhd@163.com (H.R.); yaoxh168@163.com (X.Y.)

Abstract

Pecan is rich in bioactive components such as fatty acids (FAs) and flavonoids and is an important nut type worldwide. Therefore, the molecular mechanisms of phytochemical biosynthesis in pecan are a focus of research. Recently, a draft genome and several transcriptomes have been published. However, the full-length mRNA transcripts remain unclear, and the regulatory mechanisms behind the quality components biosynthesis and accumulation have not been fully investigated. In this study, single-molecule long-read sequencing technology was used to obtain full-length transcripts of pecan kernels. In total, 37,504 isoforms of 16,702 genes were mapped to the reference genome. The numbers of known isoforms, new isoforms, and novel isoforms were 9013 (24.03%), 26,080 (69.54%), and 2411 (6.51%), respectively. Over 80% of the transcripts (30,751, 81.99%) had functional annotations. A total of 15,465 alternative splicing (AS) events and 65,761 alternative polyadenylation events were detected; wherein, the retained intron was the predominant type (5652, 36.55%) of AS. Furthermore, 1894 long noncoding RNAs and 1643 transcription factors were predicted using bioinformatics methods. Finally, the structural genes associated with FA and flavonoid biosynthesis were characterized. A high frequency of AS accuracy (70.31%) was observed in FA synthesis-associated genes. This study provides a full-length transcriptome data set of pecan kernels, which will significantly enhance the understanding of the regulatory basis of phytochemical biosynthesis during pecan kernel maturation.

Keywords: alternative splicing; *Carya illinoinensis*; PacBio; lncRNA; fatty acid; flavonoid

Introduction

Pecan [*Carya illinoinensis* (Wangenh.) K. Koch], native to North America is an important tree nut crop in the world. Pecan ($2n = 32$) belongs to family Juglandaceae, *Carya*, with an estimated genome size of 650 Mb (Huang *et al.* 2019). Its kernels contain appreciable amounts of bioactive phytochemicals, such as fatty acids (FAs), polyphenols, flavonoids, and ellagic acid (Venkatachalam and Sathe 2006; Bolling *et al.* 2011; Zhang *et al.* 2019a). Oleic acid (C18:1) is the major fraction of monounsaturated FAs, which has beneficial effects on cardiovascular disease (Perdomo *et al.* 2015), total cholesterol, and low-density lipoprotein cholesterol (Fonolla-Joya *et al.* 2016). Flavonoids are a large class of natural bioactive compounds that exert diverse favorable bioeffects for human health, such as, anti-cancer, cardioprotective, and anti-diabetic (Wang *et al.* 2018). Thus, pecan is an excellent source of dietary bioactive food components and has remarkable protective effects against chronic human diseases. Therefore, the biosynthesis mechanisms of bioactive phytochemicals in pecan kernels are a research focus (Huang *et al.* 2019; Zhang *et al.* 2019a).

In recent years, RNA sequencing (RNA-Seq) has been used to determine the molecular basis of bioactive component biosynthesis in pecan kernels and has supplied a set of candidate genes associated with FA and flavonoid biosynthesis (Huang *et al.* 2017; Mattison *et al.* 2017; Jia *et al.* 2018; Zhang *et al.* 2019a). However, the transcriptomes in these studies were generated by *de novo* assembly using Illumina short-read sequencing. With this tool, it is difficult to provide full-length sequences for transcripts and cannot offer alternatively spliced (AS) and alternative polyadenylation (APA) events for each RNA (Cheng *et al.* 2017). In eukaryotes, AS can generate multiple mRNAs from the same gene and increase the diversity of the transcriptome and proteome (Stamm *et al.* 2005). It may influence the stability, the subcellular location, and the function of the protein. In *Arabidopsis thaliana*, over 80% of multiple-exon genes exist as AS events (Zhu *et al.* 2017); the occurrence rate of AS events increase with increasing exon number (Ruan *et al.* 2018). AS plays vital roles in development, signal transduction, and stress response in plants (Staiger and Brown 2013; Tang *et al.* 2016; Laloum *et al.* 2018). APA is a widespread mRNA-processing mechanism across all eukaryotic species that produce mRNAs with distinct 3′ termini, allowing them to interact with different regulators and perform gene regulation (Tian and Manley 2017; Vallejos Baier *et al.* 2017). However, little is known about AS and APA profiles in pecans.

Single-molecule long-read sequencing technology (SMRT) is a third-generation sequencing technology that can effectively provide full-length sequences of RNA without the need for short-read assembly, and offers more complete transcriptome data (Chao *et al.* 2018). Therefore, SMRT is broadly used in transcriptome and genome sequencing and is a superior strategy for novel

Received: February 25, 2021. **Accepted:** May 18, 2021

gene discovery, gene structural variation detection, and long noncoding RNA (lncRNA) prediction. Currently, it has been widely used in various plant species, such as strawberry (Li *et al.* 2017), arabica coffee (Cheng *et al.* 2017), rice (Zhang *et al.* 2019c), *Trifolium pratense* (Chao *et al.* 2018), olive (Rao *et al.* 2019), *Ginkgo biloba* (Ye *et al.* 2019), and *Pennisetum giganteum* (Li *et al.* 2020). For instance, in *G. biloba*, 12 290 AS events, 12 954 APA events, 2286 novel transcripts, and 1270 lncRNAs were observed (Ye *et al.* 2019). Moreover, extensive AS forms have been identified during bioactive phytochemical biosynthesis in plants (Zheng *et al.* 2017). For example, three FA biosynthesis-related *fatty acid desaturases* (*FADs*) exist in peanut (Ruan *et al.* 2018); the *4-coumarate-CoA ligase* (*4CL*) and *tyrosine aminotransferase* (*TAT*) genes show alternative splicing (AS) during salvianolic acid biosynthesis in *Salvia miltiorrhiza* (Xu *et al.* 2016), and eight flavonoid-related key structural genes have been observed to express different transcripts in tea (Qiao *et al.* 2019). Although many unigenes relevant to FA and flavonoid biosynthesis in pecan kernels have been reported (Huang *et al.* 2019; Zhang *et al.* 2019a), their AS characteristics have not been elucidated. Therefore, the information of AS events is important for understanding the genetic basis underlying bioactive metabolite production in pecan embryos. Recently, a structural genome of pecan assembled with 651.31 Mb in 3860 scaffolds was published (Huang *et al.* 2019), which offered a good opportunity to globally characterize AS, and APA events in pecan. LncRNAs are a type of RNA (>200 nt) without protein-coding capacity (Santosh *et al.* 2015). They play a crucial role in diverse biological processes in plants (Santosh *et al.* 2015; Wang *et al.* 2017a), such as photomorphogenesis, vernalization, male sterility, and abiotic stress tolerance (Jha *et al.* 2020; Sanchita *et al.* 2020). The prediction of lncRNAs in pecan will aid in the understanding of pecan kernel ripening regulation.

This study used SMRT technology to construct a full-length transcriptome of pecan kernels. The global AS and APA event identification, lncRNA prediction, and transcription factor (TF) prediction were performed. Finally, the structural genes associated with FA and flavonoid biosynthesis were characterized in-depth. We believe that the application of SMRT is useful to promote genetic studies and to uncover the mechanisms of bioactive component biosynthesis in pecan kernels.

Materials and methods

Plant material

Eight different development stages, *i.e.*, August 09, August 16, August 23, August 30, September 06, September 13, September 20, and September 28, of pecan kernels were sampled from two cultivars, "YLC28" and "Oconee," in 2018. The trees were 11-years-old and planted in Jiande (29°N, 119°W), China. After removing the shell and seed coat, the embryos were rapidly frozen in liquid nitrogen.

Library construction and SMRT sequencing

Total RNA was extracted by TRIzol (Life Technologies, Carlsbad, CA, USA). Equal amounts of total RNA from 16 samples (eight developmental stages of two cultivars) were combined to construct a representative sample for sequencing. The library construction and SMRT sequencing were performed as the description of Zhou *et al.* (2020) by Gene Denovo Biotechnology Co., Ltd. (Guangzhou, China).

Data processing and new isoforms annotation

The data processing and error correction were processed as described by Zhou *et al.* (2020). Circular consensus sequence (CCS) reads were extracted and classified into full-length nonchimeric (FLNC), nonfull-length (nFL), chimeras, and short reads. Subsequently, the high-quality consensus sequences were aligned to a draft genome of pecan (Huang *et al.* 2019) using GMAP (Wu and Watanabe 2005). The isoforms were classified into known isoforms (each being uniquely mapped to one known gene locus), novel isoforms (each showing the significant match to unannotated genomic locus), and new isoforms (each split mapped to distinct exons).

Each of the new isoforms was BLAST against four databases, including NR (http://www.ncbi.nlm.nih.gov), Swissprot (http://www.expasy.ch/sprot), KEGG (http://www.genome.jp/kegg), and GO (http://www.geneontology.org).

AS detection and validation

AS events were identified and classified into seven types, such as skipping exon (SE), retained intron (RI), alternative 5′ splice sites (A5), alternative 3′ splice sites (A3), mutually exclusive (MX) exons, alternative first (AF) exons, and alternative last (AL) exons (Chen *et al.* 2020). To validate the AS events, four genes were randomly selected, and the representative total RNA sample was employed as a template. Total RNA was reverse transcribed into cDNA using a PrimeScript 1st Strand cDNA Synthesis Kit (Takara, Dalian, China). PCR reactions were performed using KOD FX polymerase (Toyobo, Osaka, Japan). PCR products were separated on a 2% agarose gel.

Alternative polyadenylation, long noncoding RNA, and transcription factor analysis

APA detection, lncRNAs characterization, and TF analysis were performed following the procedures described by Chen *et al.* (2020).

Data availability

The PacBio SMRT sequencing data set have been submitted to the NCBI SRA database under BioProject accession number: PRJNA613367. Information of AS events (Supplementary Table S1). Primers for AS identification and validation (Supplementary Table S2). Information of APA events (Supplementary Table S3). Results of long noncoding RNA prediction (Supplementary Table S4). Results of TFs prediction (Supplementary Table S5). Isoforms in lipid metabolism of pecan (Supplementary Table S6). Isoforms in flavonoid biosynthesis of pecan (Supplementary Table S7). Supplementary material is available at figshare: https://doi.org/10.25387/g3.14608305.

Results and discussion

Full-length transcriptome sequencing and functional annotation

As gene expression and AS events have tissue and temporal-based characteristics (Qiao *et al.* 2019), to gain as many kernel development-related transcripts as possible, the transcriptome of a pooled sample (eight different developmental stages of pecan kernels from two cultivars) was sequenced using a PacBio Sequel platform. A total of 22 601 162 subreads (39.55 Gb) were generated, with an average length of 1750 bp and an N50 of 2420 bp (Table 1). Then, 485 150 CCS reads were extracted and classified into FLNC, nFL, full-length chimeras, and short reads. As a result,

Table 1 Summary of PacBio sequencing results

Terms	Amount
Total base (bp)	39,556,542,673
subreads number	22,601,162
subreads average length (bp)	1,750
subreads N50 (bp)	2,420
Number of CCS reads	485,150
Mean of CCS Read Length (bp)	2,312
Number of full-length reads	445,838
Number of FLNC reads	442,244
FLNC read average length (bp)	2,154
Number of unpolished consensus isoforms	236,820
Number of polished high-quality isoforms	194,991
Unpolished consensus isoforms average read length (bp)	2,123
Correct consensus number	194,992
Correct consensus average length (bp)	2,184
Correct consensus N50 length (bp)	2,650

442 244 FLNC reads (0.95 Gb) were obtained. Subsequently, 194 991 polished high-quality isoforms were generated after cluster analysis and correction of all FLNC reads. Then, all the polished high-quality isoforms were mapped to the reference genome. In total, 194 599 (99.80%) reads were successfully mapped, including 189 566 (97.22%) unique mapped reads and 5033 (2.58%) multiple mapped reads. In total, 37,504 isoforms of 16,702 gene loci were mapped onto the reference genome, including 9013 (24.03%) known isoforms, 2411 (6.51%) novel isoforms, and 26,080 (69.54%) new isoforms.

The functions of known isoforms which uniquely mapped to one known gene locus were annotated by the pecan genome annotation information (Huang *et al.* 2019). To predict the potential functions of each isoform, all the novel isoforms and new isoforms were aligned to four databases, including Nr, Swissprot, GO, and KEGG. Altogether, 30,751 (81.99%) isoforms exhibited homology with at least one database, including 3355 known isoforms, 25,520 new isoforms, and 1876 novel isoforms. Most of these (30,750, 99.99%) were matched to the Nr database, followed by the Swissprot database with 23,073 (75.03%). In total, 15,847 (51.53%) isoforms were annotated to the GO database and classified into 46 sub-categories of three key categories (Figure 1). Regarding "biological process," the terms related to "metabolic process" and "cellular process" were the main groups. Among the molecular function categories, "catalytic activity" was the most represented subcategory, followed by "binding" and "transporter activity." For the cellular component category, "cell" and "cell part" were the two largest subcategories. All transcripts were then matched to the KEGG database (Figure 1). Altogether, 7831 (25.47%) sequences were assigned to 137 pathways. Among these, metabolic pathways "biosynthesis of secondary metabolites," and "biosynthesis of antibiotics" were the most abundant. In addition, "starch and sucrose metabolism" and "fatty acid metabolism" were also significantly enriched (Figure 1).

AS identification and validation

AS plays an important role in various biological processes in plants (Staiger and Brown, 2013; Tang *et al.* 2016; Laloum *et al.* 2018). In total, 6749 genes produced two or more isoforms, and 15, 465 AS events were detected (Figure 2, Supplementary Table S1). Among the seven types of AS forms, RI predominated, accounting for 36.55% (5652) of the AS isoforms, followed by A3 (3960, 25.61%), A5 (2584, 16.71%), and SE (1816, 11.74%). Only 231 (1.49%) and 68 (0.44%) isoforms were AL and MX types of AS, respectively (Figure 2). Similarly, in the species of Moso bamboo (Wang *et al.* 2017b), strawberry (Li *et al.* 2017), *Populus* (Chao *et al.* 2018) and tea (Qiao *et al.* 2019), RI was also the most common type of AS event. To validate the authenticity of AS events, four genes (Gene004808, Gene007321, Gene005577, and Gene001662) presenting AS isoforms were randomly selected for RT-PCR (Supplementary Table S2). The results showed that the fragments of RT-PCR were consistent with the AS isoforms identified from SMRT data (Figure 3).

Alternative polyadenylation analysis

APA is a crucial post-transcriptional mechanism that generates mRNA isoforms with different 3′ ends (Tian and Manley 2017; Vallejos Baier *et al.* 2017). A total of 65, 761 APA events in 16, 702 genes were detected in the pecan SMRT data (Supplementary Table S3). Most genes (6048, 36.21%) were detected with one poly A site, followed by two poly A sites (3441 genes, 20.60%) and more than five poly A sites (2566 genes, 15.37%). APA enhances the diversity of transcripts, and the profile of APA events in different species is different. In *G. biloba*, most genes had more than five poly A sites (Ma *et al.* 2019). In *Manis javanica*, the main category was genes containing one poly A site; however, only 5.50% genes had more than five poly A sites (Ye *et al.* 2019).

Long noncoding RNA prediction

LncRNAs play crucial roles in diverse biological processes, such as photomorphogenesis, flowering regulation, and stress tolerance in plants (Santosh *et al.* 2015; Wang *et al.* 2017a; Jha *et al.* 2020; Sanchita *et al.* 2020). However, little is known about lncRNAs and their functions in pecan. In this study, putative lncRNAs were distinguished from unannotated transcripts using the CPC, CNCI, and Swissprot databases. A set of 1894 lncRNAs was identified using the three analytical methods (Figure 4, Supplementary Table S4). They were then classified into five groups based on their position relative to nearby protein-coding genes, including intergenic lncRNAs (421, 22.23%), bidirectional lncRNAs (87, 4.59%), intronic lncRNAs (155, 8.18%), antisense lncRNAs (184, 9.71%), and sense overlapping lncRNAs (916, 48.36%). Hence, full-length transcriptome sequencing is a powerful tool for the prediction of lncRNAs in plants (Chao *et al.* 2018; Qiao *et al.* 2019). The new lncRNAs might be related to embryo development and beneficial component synthesis in pecan and require further investigation.

Transcription factor prediction

TFs are involved in various biological processes in plants. However, little is known about the roles of TFs in the regulation of biological processes in pecan. In this study, 1643 isoforms belonging to 55 TF families were predicted (Supplementary Table S5). Therein, *bZIP* (125), *C3H* (119), *bHLH* (112), and *ARF* (98) were the predominant families, followed by *C2H2* (93), *MYB_related* (84), and *FAR1* (70) (Figure 5). Among these TFs, 18 novel genes were identified, including eight *FAR1s*, two *C2C2s* and *BBR-BPCs*, and one *bZIP*, *M-type*, *C3H*, *ERF*, *GATA*, and *NF-YC* (Supplementary Table S5). MYB has been reported to play crucial roles in secondary metabolism (Wei *et al.* 2019), abiotic/biotic stress tolerance (Shen *et al.* 2017), and reproduction (Meng *et al.* 2018). Meanwhile, several TFs have been shown to regulate FA biosynthesis in plants. For example, *WRI1* and *bZIP67* regulate FA synthesis in *Arabidopsis* (To *et al.* 2012; Mendes *et al.* 2013), and *GmWRI1a* positively regulates oil accumulation in soybean (Chen *et al.* 2018). Therefore, the TFs reported here provide a foundation for further functional characterization of TFs in pecan.

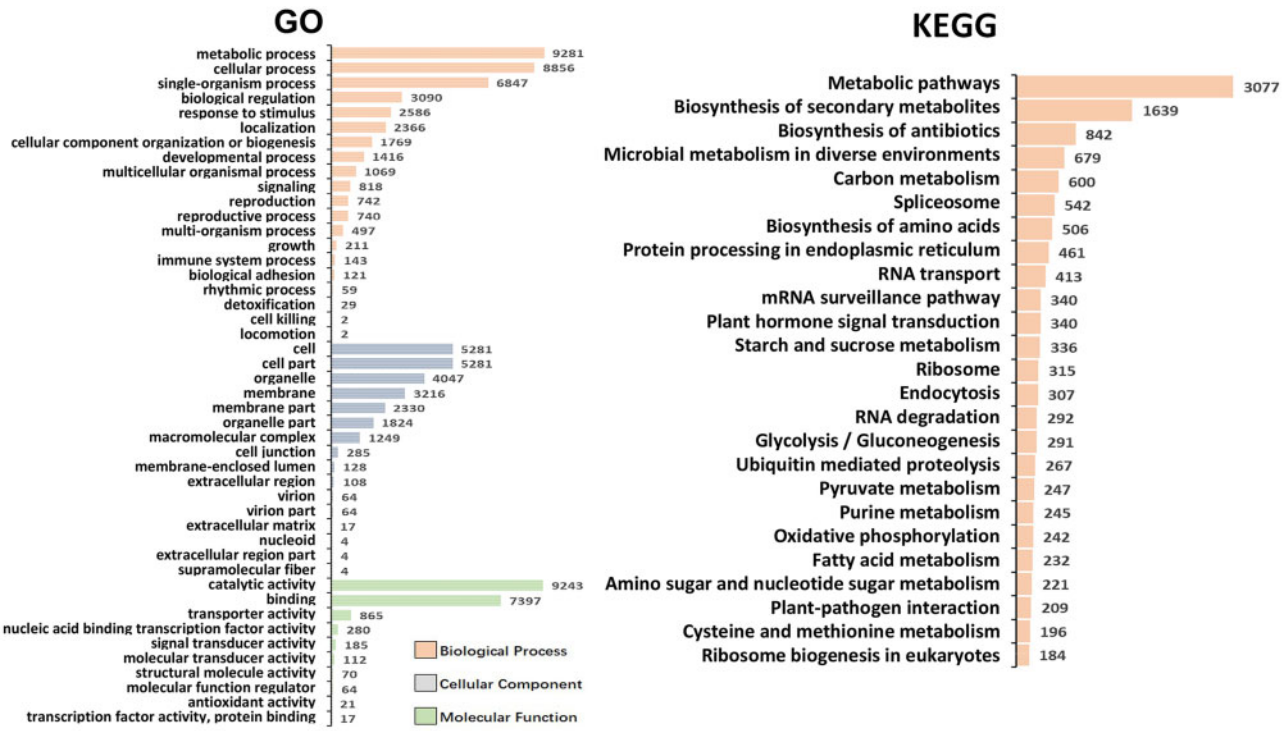

Figure 1 KEGG and GO functional classification of the pecan full-length transcriptome.

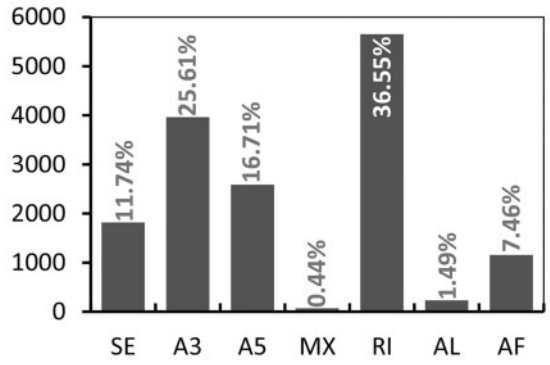

Figure 2 Summary of alternative splicing events.

Identification of lipid biosynthesis-related transcripts

Unsaturated FAs are abundant in pecan, which is one of the most important health components. To date, several studies have been performed to uncover the molecular mechanisms underlying oil accumulation during pecan nut maturation (Huang *et al.* 2017; Mattison *et al.* 2017; Jia *et al.* 2018). Many unigenes involved in *de novo* FA synthesis and triacylglyceride (TAG) synthesis pathways were isolated. However, the full-length of FAs-related transcripts has not been identified, and the post-transcriptional regulatory mechanisms of these genes have not been evaluated in pecan. According to the KEGG annotation results, 735 isoforms associated with lipid metabolism were identified, including 68 known isoforms, 610 new isoforms, and 57 novel isoforms (Supplementary Table S6). Then, with the emphasis on FA and TAG biosynthesis, 19 gene families including 64 genes were illustrated in the oil biosynthesis model (Figure 6).

In previous studies, a high AS ratio (61.6%) was observed in FA synthesis-associated genes in peanut (Ruan *et al.* 2018), and over 12 genes involved in α-linolenic acid (α-C18:3) metabolism showed AS events in *P. giganteum* (Li *et al.* 2020). Similarly, 70.31% (45 out of 64 genes) of the FA and TAG biosynthesis-related structural genes generated multiple protein isoforms (Figure 6). These high AS occurrences may be due to active transcription of FA-related genes and the rapid accumulation of FAs along with embryo maturation. Previous studies indicated that the AS occurrence rate is closely related to the phase of tissue development and location in plants (Stamm *et al.* 2005; Wang *et al.* 2017b; Zhu *et al.* 2017). Therefore, these observations imply that AS events might play crucial roles in FA metabolism, and the function of different isoforms needs to be studied in the future.

Characterization of transcripts associated with FAs synthesis

Acetyl-CoA carboxylase (ACCase) is a key enzyme for FA *de novo* synthesis (Sasaki and Nagano 2004). In plastids, ACCase comprises four subunits, including α-carboxyltransferase (α-CT), β-carboxyltransferase (β-CT), biotin carboxylase (BC), and biotin carboxyl carrier protein (BCCP). Broad AS events occurred in the *ACCase* subunits (Supplementary Table S6). The plastidial fatty acid synthase (FAS) system catalyzes *de novo* FA synthesis (Wei *et al.* 2012). This system consisted of β-ketoacyl-acyl carrier protein synthases (KASI, KASII, and KASIII), β-ketoacyl-ACP reductase (KAR), β-hydroxacyl-ACP dehydratase (HAD), and enoyl-ACP reductase (EAR). A set of *FASs* were isolated, including *KASI* (2), *KASII* (3), *KASIII* (2), *KAR* (2), *HAD* (1), and *EAR* (2). Except for *HAD*, all these genes expressed AS isoforms (Supplementary Table S6). FATA (acyl-ACP thioesterase A) and FATB (acyl-ACP thioesterase B) are the main determinants of FA chain length and the amount of saturated FA (Salas and Ohlrogge 2002). Two *FATAs* and two *FATBs* were obtained. Stearoyl-acyl desaturase (SAD) converts C18:0-ACP to C18:1-ACP, which is a key enzyme in determining the ratio between unsaturated and saturated FAs (Du *et al.* 2016). Four *SADs* including 12 isoforms were identified (Supplementary

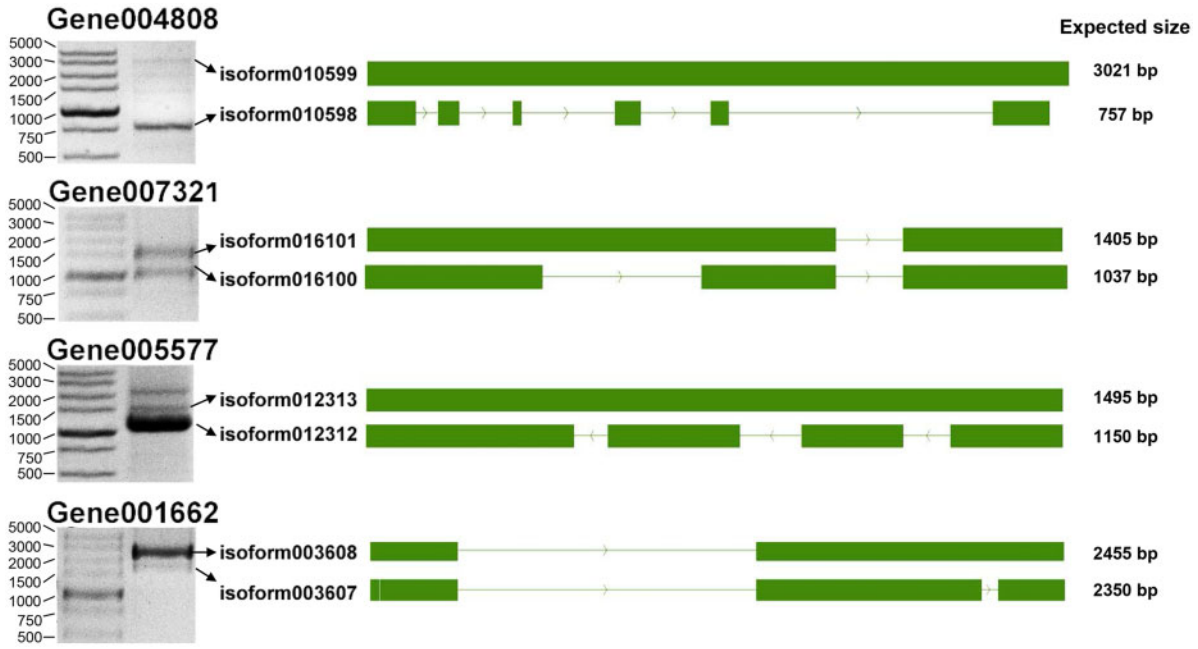

Figure 3 Validation of AS events using RT-PCR. The gel bands show DNA markers and the RT-PCR results. The arrows indicate the positions of different isoforms. The green boxes show the positions of exons, and lines show introns. The expected PCR sizes of each band was listed beside the structures.

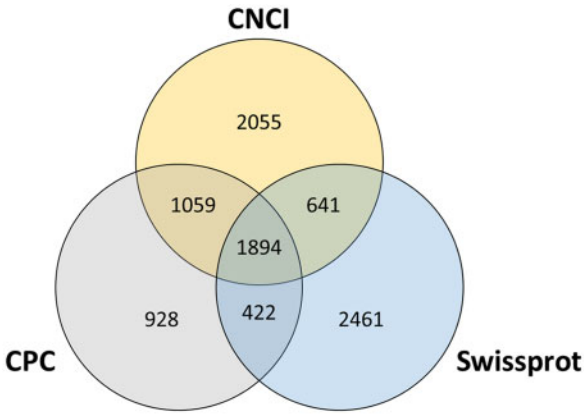

Figure 4 The Venn diagram of lncRNAs predicted by three databases.

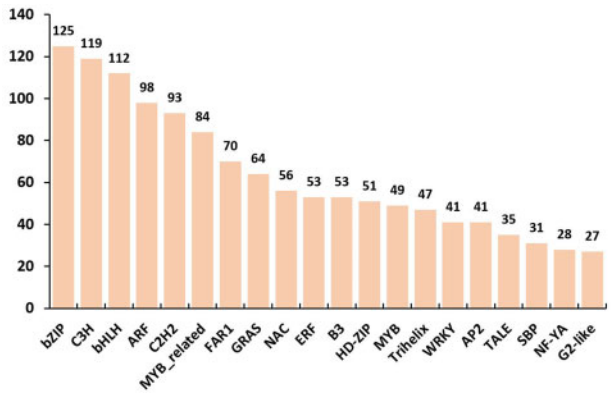

Figure 5 The number of top 20 transcription factors.

Table S6). Among them, different AS isoforms were generated for the two *SADs* of Gene006109 (Isoform013454 ~ Isoform013458) and Gene000740 (Isoform001516 ~ Isoform001520), respectively.

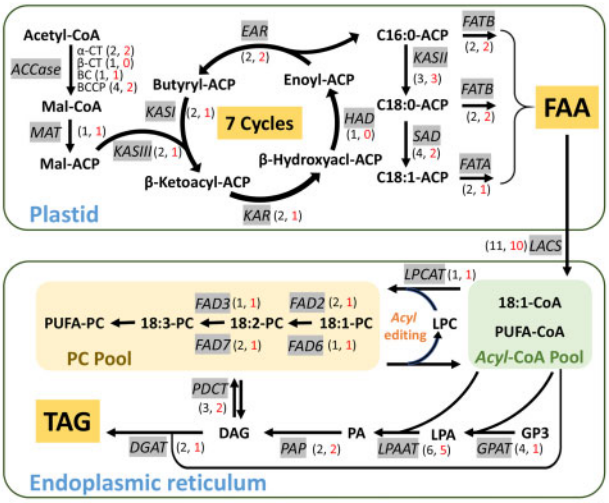

Figure 6 The proposed FA biosynthesis pathway of pecan. The numbers in brackets indicate the number of putative genes (in black font) and the number of AS genes (in red font).

Previously, two to six putative *SADs* were identified in pecan by using RNA-Seq (Huang *et al.* 2017; Jia *et al.* 2018; Xu *et al.* 2018), among them, two *SADs* may be responsible for the rapid accumulation of oleic acid in pecan kernels (Huang *et al.* 2017; Jia *et al.* 2018). Here, 10 AS isoforms of two *SADs* were identified which enhance our knowledge about FA desaturation in pecan seeds. Long-chain acyl-coenzyme A synthetases (LACS) catalyze free FAs to form long-chain acyl-CoA. Eleven *LACSs* were obtained, and 10 *LACSs* expressed multiple isoforms (Figure 6; Supplementary Table S6). Δ12-desaturase of FA desaturase 2 (FAD2) and FA desaturase 6 (FAD6) catalyze C18:1 to generate C18:2 (linoleic acid). Synthesis of α-C18:3 by C18:2 catalyzed by Δ15-desaturase, including FAD3, FAD7, and FAD8. Two, one, one and two *FAD2s*, *FAD3*, *FAD6*, and *FAD7s*, respectively, were obtained. Meanwhile, all presented splicing transcripts. Similarly,

Ruan (2018) also identified the presence of AS events in three *FADs* in peanuts. FAD significantly influences the FA composition, and the functions of different *FAD* isoforms in FA synthesis should be further studied in the future.

Characterization of transcripts associated with triacylglyceride biosynthesis

The assembly of TAG is catalyzed in-turn by glycerol-3-phosphate acyltransferase (GPAT), lysophosphatidic acid acyltransferase (LPAAT), phosphatidate phosphatase (PAP), and diacylglycerol O-acyltransferase (DGAT). Here, four, six, two, and two genes of *GPAT*, *LPAAT*, *PAP*, and *DGAT*, respectively, were obtained, and all generated AS transcripts (Supplementary Table S6). DGAT is a rate-limiting enzyme during TAG assembly and has been widely studied in plants (Zheng *et al.* 2017). In this study, a *DGAT1* with three AS transcripts, including isoform035438 (2110 bp), isoform035439 (2021 bp), and isoform035440 (2216 bp) were detected. A previous study found that AS is crucial for the regulation of gene expression and enzyme activity of *AhDGAT1* in peanut (Zheng *et al.* 2017). In addition, five *AhDGAT1* isoforms can generate high acyltransferase activity enzymes and complement the lethality phenotype of *Saccharomyces cerevisiae* strain H1246 (Zheng *et al.* 2017). In pecan, two to three putative *DGAT1s* and one *DGAT2* were obtained, and their expression patterns during embryo development have been reported (Huang *et al.* 2017; Jia *et al.* 2018). However, the post-transcriptional regulation of *DGAT1* in this plant has not been described previously. In this study, additional variants of pecan *DGAT1* were observed that will use to better understand the molecular mechanism of TAG biosynthesis in pecan. Therefore, the AS isoforms of *DGAT1s* should be characterized better in pecan studies.

Characterization of transcripts associated with flavonoid biosynthesis

Flavonoids are crucial beneficial components in pecan nuts; however, their biosynthesis mechanisms have not been fully elucidated (Zhang *et al.* 2019a,b). Two interconnected metabolic pathways underlie the biosynthesis of a wide range of flavonoids in plants, including the "phenylpropanoid biosynthesis pathway" and the "flavonoid biosynthesis pathway" (Figure 7). Phenylalanine ammonia-lyase (PAL) is the first enzyme in the "phenylpropanoid pathway," which converts phenylalanine into trans-cinnamic acid. One *PAL* with three transcript variants (isoform003607, isoform003608, and isoform003609) was obtained (Figure 7; Supplementary Table S7). These AS isoforms ranged from 1191 bp (isoform003609) to 2486 bp (isoform003608), and isoform003608 retained an intron near the 3′ end of the gene. Similar studies also found that *PAL* expressed multiple isoforms by AS in tea and olive (Xu *et al.* 2017; Rao *et al.* 2019). Cinnamate 4-hydroxylase (C4H) catalyzes trans-cinnamic acid into *p*-coumaric acid. The 4-coumarate-CoA ligase (4CL) converts *p*-coumaric acid into *p*-coumaroyl CoA. The *p*-coumaroyl CoA is precursor of flavonoids, lignins, and isoflavonoids. Two *C4Hs* and five *4CLs* were obtained; of which two *4CLs* were identified as having AS transcripts (Figure 7; Supplementary Table S7). Similar AS events in *4CLs* were also observed in *S. miltiorrhiza* (Xu *et al.* 2016).

CHS is a key enzyme during flavonoid biosynthesis. The expression of CHS can influence the accumulation of flavonoids and affect fruit development, floral color, stress tolerance, and other important physiological processes in plants. This enzyme catalyzes p-coumaroyl-CoA and malonyl-CoA to yield naringenin chalcone. Four *CHSs* were obtained, of which one gene expressed two transcript variants (isoform009353 and isoform009354). The longer transcript of isoform009354 (2103 bp) retained an intron (558 bp) near the 3′ end of the gene. Rao *et al.* (2019) also observed an IR-type AS event in a *CHS* gene in olive. Recently, three *CiCHSs* have been isolated and characterized in pecan kernels (Zhang *et al.* 2019b). Further studies should investigate how the AS of *CHS* regulates flavonoid synthesis in the pecan kernel. Chalcone-flavanone isomerase (CHI) is another rate-limiting enzyme in the flavonoid biosynthesis pathway, which catalyzes naringenin chalcone into flavanone. Two *CHIs* were obtained. After CHI, a flavanone 3-hydroxylase (F3H) catalyzes the conversion of flavanones into dihydroflavonols. One putative *F3H* was identified. Flavonol synthase (FLS) is able to convert dihydroflavonols to flavonols. Three *FLSs* were obtained, including two new isoforms (isoform012313 and isoform018838) and one novel isoform (isoform033121). Dihydroflavonol 4-reductase (DFR) converts dihydroflavonols into leucoanthocyanidins, which are subsequently converted to anthocyanins by anthocyanidin synthase (ANS). One *DFR* and one *ANS* were identified; the *DFR* expressed AS isoforms (isoform011396 and isoform011397). Leucoanthocyanidin reductase (LAR) and anthocyanidin reductase (ANR) catalyzed leucoanthocyanidins and anthocyanidin, respectively, to yield different types of flavan-3-ol monomers, respectively. The latter can link together to generate proanthocyanidins. One *LAR* and one *ANR* were obtained, in which the *ANR* generated two isoforms (isoform028612 and isoform028613) by AS. In summary, 22 genes associated with flavonoid biosynthesis were identified, including one *PAL*, two *4CL*, one *CHS*, one *DFR*, and one *ANR* transcribed AS isoform. These results were consistent with the reports in tea, olive, and kiwifruit, wherein many flavonoid biosynthesis-related genes expressed AS transcripts (Tang *et al.* 2016; Zhu *et al.* 2017; Qiao *et al.* 2019; Rao *et al.* 2019). Zhang *et al.* (2019a) and Huang *et al.* (2019) reported a set of flavonoid-related

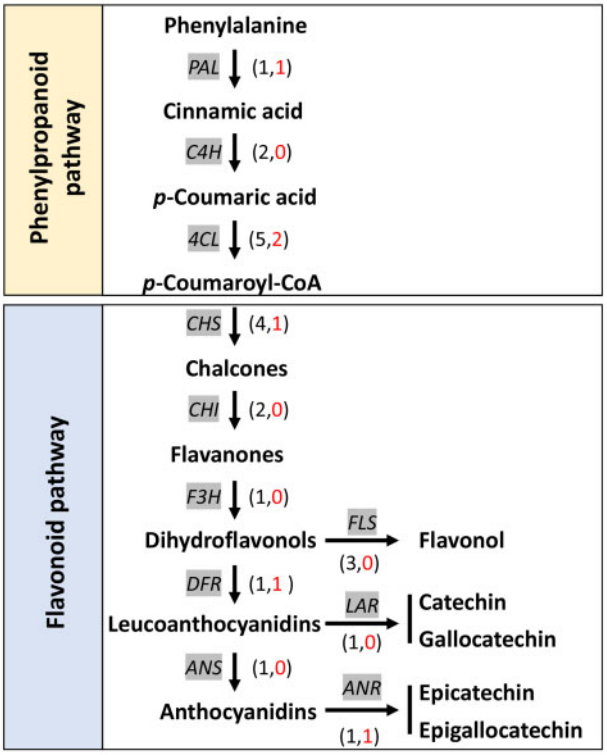

Figure 7 The proposed flavonoid biosynthesis pathway of pecan. The numbers in brackets indicate the number of putative genes (in black font) and the number of AS genes (in red font).

genes by using next-generation sequencing technology. However, the AS event of these genes has not been reported previously. AS transcripts may encode functional proteins instead of their conventional full-length transcripts (Zhu *et al.* 2017). The identification of AS of flavonoid-related genes in this study will aid in a more comprehensive understanding of flavonoid biosynthesis in pecan kernels.

Conclusions

The full-length transcriptome of pecan is reported for the first time in this study. A total of 37,504 isoforms were obtained, including 26,080 (69.54%) new isoforms and 2411 (6.51%) novel isoforms. A total of 15,465 AS events were observed, and the RI was the main type (5652, 36.55%). In addition, 65,761 APA events were detected, and 1894 lncRNAs and 1643 TFs were obtained. More importantly, 64 and 22 structural genes associated with FA and flavonoid biosynthesis were isolated, respectively. Meanwhile, a high AS ratio (70.31%) was observed in FA synthesis-associated genes. This study offers the full-length transcriptome data set of pecan kernels and presents a global view of AS events during pecan embryo development. Further studies are warranted to investigate the stage-specific gene AS patterns during kernel maturation as well as how these AS events influence the biosynthesis of bioactive components in pecan. We believe that our findings will promote the uncovering of the post-transcriptional regulation of pecan kernel development and can be used to understand the basis of FA and flavonoid biosynthesis in pecan kernels.

Acknowledgments

C.Z. carried out the experiments and written the draft articles. C.Z., H.R., K.W., and J.C. performed statistical analysis. X.Y., H.R., K.W., and J.C. revised the manuscript.
C.Z., H.R., and X.Y. designed the experiments.

Funding

This research was supported by the Fundamental Research Funds of CAF No. CAFYBB2018SY013, and the Fundamental Research Funds of CAF No. CAFYBB2017ZA004-8.

Conflicts of interest

The authors declare no conflicts of interest.

Literature cited

Bolling BW, Chen CYO, McKay DL, Blumberg JB. 2011. Tree nut phytochemicals: composition, antioxidant capacity, bioactivity, impact factors. A systematic review of almonds, brazils, cashews, hazelnuts, macadamias, pecans, pine nuts, pistachios and walnuts. Nutr Res Rev. 24:244–275.

Chao Y, Yuan J, Li S, Jia S, Han L, *et al.* 2018. Analysis of transcripts and splice isoforms in red clover (*Trifolium pratense* L.) by single-molecule long-read sequencing. BMC Plant Biol. 18:300.

Chen D, Du Y, Fan X, Zhu Z, Jiang H, *et al.* 2020. Reconstruction and functional annotation of *Ascosphaera apis* full-length transcriptome utilizing PacBio long reads combined with Illumina short reads. J Invertebr Pathol. 176:107475.

Chen L, Zheng Y, Dong Z, Meng F, Sun X, *et al.* 2018. Soybean (*Glycine max*) *WRINKLED1* transcription factor, *GmWRI1a*, positively regulates seed oil accumulation. Mol Genet Genomics. 293:401–415.

Cheng B, Furtado A, Henry RJ. 2017. Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. Gigascience. 6:1–13.

Du H, Huang M, Hu J, Li J. 2016. Modification of the fatty acid composition in Arabidopsis and maize seeds using a stearoyl-acyl carrier protein desaturase-1 (*ZmSAD1*) gene. BMC Plant Biol. 16:137.

Fonolla-Joya J, Reyes-García R, García-Martín A, López-Huertas E, Muñoz-Torres M. 2016. Daily intake of milk enriched with n-3 fatty acids, oleic acid, and calcium improves metabolic and bone biomarkers in postmenopausal women. J Am Coll Nutr. 35:529–536.

Huang R, Huang Y, Sun Z, Huang J, Wang Z. 2017. Transcriptome analysis of genes involved in lipid biosynthesis in the developing embryo of pecan (*Carya illinoinensis*). J Agric Food Chem. 65:4223–4236.

Huang Y, Xiao L, Zhang Z, Zhang R, Wang Z, *et al.* 2019. The genomes of pecan and Chinese hickory provide insights into *Carya* evolution and nut nutrition. Gigascience. 8:1–17.

Jia X, Li M, Luo H, Zhai M, Guo Z, *et al.* 2018. Transcriptome survey reveals candidate genes involved in lipid metabolism of *Carya illinoinensis*. Int J Agric Biol. 20:991–1004.

Jha UC, Nayyar H, Jha R, Khurshid M, Zhou M, *et al.* 2020. Long non-coding RNAs: emerging players regulating plant abiotic stress response and adaptation. BMC Plant Biol. 20:466.

Laloum T, Martín G, Duque P. 2018. Alternative splicing control of abiotic stress responses. Trends Plant Sci. 23:140–150.

Li Q, Xiang C, Xu L, Cui J, Fu S, *et al.* 2020. SMRT sequencing of a full-length transcriptome reveals transcript variants involved in C18 unsaturated fatty acid biosynthesis and metabolism pathways at chilling temperature in *Pennisetum giganteum*. BMC Genomics. 21:52.

Li Y, Dai C, Hu C, Liu Z, Kang C. 2017. Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-Seq in strawberry. Plant J. 90:164–176.

Ma JE, Jiang HY, Li LM, Zhang XJ, Li HM, *et al.* 2019. SMRT sequencing of the full-length transcriptome of the Sunda pangolin (*Manis javanica*). Gene. 692:208–216.

Mattison CP, Rai R, Settlage RE, Hinchliffe DJ, Madison C, *et al.* 2017. RNA-Seq analysis of developing recan (*Carya illinoinensis*) embryos reveals parallel expression patterns among allergen and lipid metabolism genes. J Agric Food Chem. 65:1443–1455.

Mendes A, Kelly AA, van Erp H, Shaw E, Powers SJ, *et al.* 2013. *bZIP67* regulates the omega-3 fatty acid content of *Arabidopsis* seed oil by activating *FATTY ACID DESATURASE3*. Plant Cell. 25:3104–3116.

Meng D, He M, Bai Y, Xu H, Dandekar AM, *et al.* 2018. Decreased sorbitol synthesis leads to abnormal stamen development and reduced pollen tube growth via an MYB transcription factor, *MdMYB39L*, in apple (*Malus domestica*). New Phytol. 217:641–656.

Perdomo L, Beneit N, Otero YF, Escribano Ó, Díaz-Castroverde S, *et al.* 2015. Protective role of oleic acid against cardiovascular insulin resistance and in the early and late cellular atherosclerotic process. Cardiovasc Diabetol. 14:75.

Qiao D, Yang C, Chen J, Guo Y, Li Y, *et al.* 2019. Comprehensive identification of the full-length transcripts and alternative splicing related to the secondary metabolism pathways in the tea plant (*Camellia sinensis*). Sci Rep. 9:2709.

Rao G, Zhang J, Liu X, Luo Y. 2019. Identification of putative genes for polyphenol biosynthesis in olive fruits and leaves using full-length transcriptome sequencing. Food Chem. 300:125246.

Ruan J, Guo F, Wang Y, Li X, Wan S, *et al.* 2018. Transcriptome analysis of alternative splicing in peanut (*Arachis hypogaea* L.). BMC Plant Biol. 18:139.

Salas JNJ, Ohlrogge JB. 2002. Characterization of substrate specificity of plant FatA and FatB acyl-ACP thioesterases. Arch Biochem Biophys. 403:25–34.

Sanchita PK, Trivedi MH. Asif 2020. Updates on plant long non-coding RNAs (lncRNAs): the regulatory components. Plant Cell Tiss Org. 140:259–269.

Santosh B, Varshney A, Yadava PK. 2015. Non-coding RNAs: biological functions and applications. Cell Biochem Funct. 33:14–22.

Sasaki Y, Nagano Y. 2004. Plant acetyl-CoA carboxylase: structure, biosynthesis, regulation, and gene manipulation for plant breeding. Biosci Biotechnol Biochem. 68:1175–1184.

Shen X, Guo X, Guo X, Zhao D, Zhao W, *et al.* 2017. *PacMYBA*, a sweet cherry R2R3-MYB transcription factor, is a positive regulator of salt stress tolerance and pathogen resistance. Plant Physiol Biochem. 112:302–311.

Staiger D, Brown JWS. 2013. Alternative splicing at the intersection of biological Timing, development, and stress responses. Plant Cell. 25:3640–3656.

Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, *et al.* 2005. Function of alternative splicing. Gene. 344:1–20.

Tang W, Zheng Y, Dong J, Yu J, Yue J, *et al.* 2016. Comprehensive transcriptome profiling reveals long noncoding RNA expression and alternative splicing regulation during fruit development and ripening in kiwifruit (*Actinidia chinensis*). Front Plant Sci. 7:335.

Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors. Nat Rev Mol Cell Biol. 18:18–30.

To A, Joubès J, Barthole G, Lécureuil A, Scagnelli A, *et al.* 2012. WRINKLED transcription factors orchestrate tissue-specific regulation of fatty acid biosynthesis in *Arabidopsis*. Plant Cell. 24:5007–5023.

Vallejos Baier R, Picao-Osorio J, Alonso CR. 2017. Molecular regulation of alternative polyadenylation (APA) within the drosophila nervous system. J Mol Biol. 429:3290–3300.

Venkatachalam M, Sathe SK. 2006. Chemical composition of selected edible nut seeds. J Agric Food Chem. 54:4705–4714.

Wang J, Meng X, Dobrovolskaya OB, Orlov YL, Chen M. 2017a. Non-coding RNAs and their roles in stress response in plants. Genom Proteom Bioinform. 15:301–312.

Wang T, Li Q, Bi K. 2018. Bioactive flavonoids in medicinal plants: structure, activity and biological fate. Asian J Pharmaceutical Sci. 13:12–23.

Wang T, Wang H, Cai D, Gao Y, Zhang H, *et al.* 2017b. Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (*Phyllostachys edulis*). Plant J. 91:684–699.,

Wei K, Wang L, Zhang Y, Ruan L, Li H, *et al.* 2019. A coupled role for *CsMYB75* and *CsGSTF1* in anthocyanin hyperaccumulation in purple tea. Plant J. 97:825–840.

Wei Q, Li J, Zhang L, Wu P, Chen Y, *et al.* 2012. Cloning and characterization of a β-ketoacyl-acyl carrier protein synthase II from *Jatropha curcas*. J Plant Physiol. 169:816–824.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 21: 1859–1875.

Xu Q, Zhu J, Zhao S, Hou Y, Li F, *et al.* 2017. Transcriptome profiling using single-molecule direct RNA sequencing approach for in-depth understanding of genes in secondary metabolism pathways of *Camellia sinensis*. Front Plant Sci. 8:1205.

Xu Z, Luo H, Ji A, Zhang X, Song J, *et al.* 2016. Global identification of the full-length transcripts and alternative splicing related to phenolic acid biosynthetic genes in *Salvia miltiorrhiza*. Front Plant Sci. 7:100.

Xu Z, Ni J, Shah FA, Wang Q, Wang Z, *et al.* 2018. Transcriptome analysis of pecan seeds at different developing stages and identification of key genes involved in lipid metabolism. PLoS One. 13: e0195913.

Ye J, Cheng S, Zhou X, Chen Z, Kim SU, *et al.* 2019. A global survey of full-length transcriptome of *Ginkgo biloba* reveals transcript variants involved in flavonoid biosynthesis. Ind Crops Prod. 139: 111547.

Zhang C, Yao X, Ren H, Chang J, Wang K. 2019a. RNA-Seq reveals flavonoid biosynthesis-related genes in pecan (*Carya illinoinensis*) kernels. J Agric Food Chem. 67:148–158.

Zhang C, Yao X, Ren H, Wang K, Chang J. 2019b. Isolation and characterization of three *chalcone synthase* genes in pecan (*Carya illinoinensis*). Biomolecules. 9:236.

Zhang G, Sun M, Wang J, Lei M, Li C, *et al.* 2019c. PacBio full-length cDNA sequencing integrated with RNA-seq reads drastically improves the discovery of splicing transcripts in rice. Plant J. 97: 296–305.

Zheng L, Shockey J, Guo F, Shi L, Li X, *et al.* 2017. Discovery of a new mechanism for regulation of plant triacylglycerol metabolism: the peanut diacylglycerol acyltransferase-1 gene family transcriptome is highly enriched in alternative splicing variants. J Plant Physiol. 219:62–70.

Zhou Y, Chen Z, He M, Gao H, Zhu H, *et al.* 2020. Unveiling the complexity of the litchi transcriptome and pericarp browning by single-molecule long-read sequencing. Postharvest Biol Tec. 168: 111252.

Zhu FY, Chen MX, Ye NH, Shi L, Ma KL, *et al.* 2017. Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in Arabidopsis seedlings. Plant J. 91: 518–533.

Communicating editor: B. Andrews