frontiers in Plant Science

ORIGINAL RESEARCH
published: 23 February 2016
doi: 10.3389/fpls.2016.00163

Transcriptome Analysis Identifies Candidate Genes Related to Triacylglycerol and Pigment Biosynthesis and Photoperiodic Flowering in the Ornamental and Oil-Producing Plant, *Camellia reticulata* (Theaceae)

Qiu-Yang Yao[1,2†], Hui Huang[1†], Yan Tong[1], En-Hua Xia[1,2] and Li-Zhi Gao[1]*

[1] Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species in Southwest China, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China, [2] University of Chinese Academy of Sciences, Beijing, China

OPEN ACCESS

Edited by:
Danièle Werck,
Centre National de la Recherche Scientifique, France

Reviewed by:
Hubert Schaller,
Centre National de la Recherche Scientifique, France
Mingshu Cao,
AgResearch Grasslands Research Center, New Zealand
Qing Liu,
Commonwealth Scientific and Industrial Research Organisation, Australia

***Correspondence:**
Li-Zhi Gao
lgao@mail.kib.ac.cn

[†] These authors have contributed equally to this work.

Specialty section:
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

Received: 16 July 2015
Accepted: 30 January 2016
Published: 23 February 2016

Citation:
Yao Q-Y, Huang H, Tong Y, Xia E-H and Gao L-Z (2016) Transcriptome Analysis Identifies Candidate Genes Related to Triacylglycerol and Pigment Biosynthesis and Photoperiodic Flowering in the Ornamental and Oil-Producing Plant, Camellia reticulata (Theaceae). Front. Plant Sci. 7:163. doi: 10.3389/fpls.2016.00163

Camellia reticulata, which is native to Southwest China, is famous for its ornamental flowers and high-quality seed oil. However, the lack of genomic information for this species has largely hampered our understanding of its key pathways related to oil production, photoperiodic flowering process and pigment biosynthesis. Here, we first sequenced and characterized the transcriptome of a diploid *C. reticulata* in an attempt to identify genes potentially involved in triacylglycerol biosynthesis (TAGBS), photoperiodic flowering, flavonoid biosynthesis (FlaBS), carotenoid biosynthesis (CrtBS) pathways. *De novo* assembly of the transcriptome provided a catalog of 141,460 unigenes with a total length of ~96.1 million nucleotides (Mnt) and an N50 of 1080 nt. Of them, 22,229 unigenes were defined as differentially expressed genes (DEGs) across five sequenced tissues. A large number of annotated genes in *C. reticulata* were found to have been duplicated, and differential expression patterns of these duplicated genes were commonly observed across tissues, such as the differential expression of *SOC1_a*, *SOC1_b,* and *SOC1_c* in the photoperiodic flowering pathway. Up-regulation of *SAD_a* and *FATA* genes and down-regulation of *FAD2_a* gene in the TAGBS pathway in seeds may be relevant to the ratio of monounsaturated fatty acid (MUFAs) to polyunsaturated fatty acid (PUFAs) in seed oil. *MYBF1*, a transcription regulator gene of the FlaBS pathway, was found with great sequence variation and alteration of expression patterns, probably resulting in functionally evolutionary differentiation in *C. reticulata*. *MYBA1_a* and some anthocyanin-specific biosynthetic genes in the FlaBS pathway were highly expressed in both flower buds and flowers, suggesting important roles of anthocyanin biosynthesis in flower development. Besides, a total of 40,823 expressed sequence tag simple sequence repeats (EST-SSRs) were identified in the *C. reticulata* transcriptome, providing valuable marker resources for further basic and applied researches on this economically important *Camellia* plant.

Keywords: *Camellia reticulata*, transcriptome, triacylglycerol biosynthesis, photoperiodic flowering pathway, pigment biosynthesis, EST-SSRs

INTRODUCTION

Camellia reticulata Lindley, one of the most famous plants in the family Theaceae (Huang et al., 2013), is an evergreen flowering tree or shrub naturally distributed in Southwest China (Ming et al., 2000). It primarily occurs in mountainous regions with poor soil conditions and a long dry season due to the subtropical plateau monsoon climates. Natural populations with different polyploidy levels (i.e., $2n = 2x, 4x, 6x$; $x = 15$) of *C. reticulata* resulted from natural hybridization and polyploidization were found across Southwest China (Liu and Gu, 2011), which form a greatly valuable gene pool to be explored for *Camellia* breeding programs in the future. Compared with the common camellia oil plant, *C. oleifera*, *C. reticulata* is very welcome in this region because of its higher economic returns especially in terms of seed size, seed yield and oil content (Liu and Ma, 2010). It is well recognized that the oil extracted from *Camellia* seeds is similar to that from well-known olive in FA composition, making it a valuable edible oil source for daily consumption. Indeed, the camellia oil contains a high content of unsaturated FAs (UFAs), comprising 60–80% oleic acid and 5–10% linoleic acid (Liu and Ma, 2010; Ma et al., 2011) and nearly undetectable content of very long-chain FAs (VLCFA, $>C_{20}$) (Ma et al., 2011). The ratio of monounsaturated FAs (MUFAs) to saturated FAs (SFAs) in the camellia oil is close to the optimal ratio following the Somopoulos' "Omega Diet" (Simopoulos and Robinson, 1999). These features are considered having balanced and healthy effects in reducing the risk of obesity, cancer, and heart disease. Researches in the rapeseed and *Arabidopsis* indicated that genes involved in triacylglycerol (TAG) biosynthesis (TAGBS) pathway are very much related to the oil composition and yield (Baud and Lepiniec, 2010). The TAGBS pathway mainly includes two conceptually simplified systems: the biosynthesis and modification of fatty acids (FAs), and TAG assembly (Baud and Lepiniec, 2010; Xu et al., 2011). FAs are *de novo* synthesized in the plastids with acetyl-CoA as a common precursor, and exported toward the cytosolic compartment as FA-CoA esters, which mainly contains FAs with double bonds of less than two and carbons of no longer than 18. FA modification generates a variety of FAs and involves enzymes contributing to FA elongation ($>C_{20}$) and polyunsaturated FA (PUFA) biosynthesis, such as BETA-KETOACYL -[ACYL-CARRIER PROTEIN] REDUCTASE (KAR) (Slabas et al., 1992), and BETA-KETOACYL-COA SYNTHASE (KCS) (Joubes et al., 2008), FATTY-ACID DESATURASE (FAD) (Okuley et al., 1994; Ma and Browse, 2006), and PHOSPHATIDYLCHOLINE:DIACYLGLYCEROL CHOLINEPHOSPHOTRANSFERASE (PDCT) (Lu et al., 2009). The TAG assembly occurs in the endoplasmic reticulum and involves four consecutive enzymatic reactions catalyzed by GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE (GPAT), LYSOPHOSPHOLIPID ACYLTRANSFERASE (LPAT), PHOSPHATIDIC ACID PHOSPHATASE (PAP), and DIACYLGLYCEROL ACYLTRANSFERASE (DGAT) (Baud and Lepiniec, 2010). Although the majority of candidate genes in TAGBS pathway were identified in the *C. oleifera* transcriptome using the 454 sequencing platform (Xia et al., 2014), their expression patterns are still largely unexplored in other *Camellia* species such as *C. reticulata*. The regulation of unique FA composition of camellia oil requires to be further studied, which will generate useful information to guide *Camellia* improvement programs.

C. reticulata is also a well-known camellia flower, which has been cultivated as a popular gardening plant in its indigenous region for at least 1300 years (Yu and Bruce, 1980). This ornamental woody plant is notable for large flowers, brilliant colors, various cultivars, and long florescence (Yu and Bruce, 1980; Liu and Gu, 2011). Extensive research efforts have advanced our knowledge regarding the genetic basis of photoperiodic flowering process in higher plants, particularly in *A. thaliana* for example. The photoperiodic flowering pathway overlaps with the circadian rhythm network, which acts as an endogenous timing system, enabling plants to promote flowering in response to photoperiod (Mas, 2005; Andres and Coupland, 2012). The photoperiod and irradiance are mainly perceived by mature leaves and mediated by many genes, mainly including *FLAVIN-BINDING KELCH REPEAT F BOX PROTEIN* (*FKF1*) (Imaizumi et al., 2005), *GIGANTEA* (*GI*) (Fowler et al., 1999), and *CONSTANS* (*CO*) (Putterill et al., 1995; Valverde et al., 2004). *CO* as the key regulator in the photoperiodic flowering pathway actively promotes the transcription of *FLOWERING LOCUS T* (*FT*) gene under long-day conditions (Putterill et al., 1995; Valverde et al., 2004). The FT protein, also known as florigen, is believed to move to the shoot apical meristem where the flowering process occurs (Amasino and Michaels, 2010). It has been proposed that FT interacts with FLOWERING LOCUS D (FD) to form a FT-FD complex in the meristem (Wigge et al., 2005). The FT-FD complex promotes the transcription activation of *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS* (*SOC1*) and a floral meristem-identity gene *AP1APETALA1* (*AP1*) (Putterill et al., 1995; Wigge et al., 2005). *SOC1* encodes a MADS box transcription factor that activates another floral meristem identity gene *LEAFY* (*LFY*) (Putterill et al., 1995; Samach et al., 2000). *LFY* and *APETALA1* (*AP1*) subsequently induce flower development at the anlagen of shoot apical meristem according to the ABC model (Jack, 2004; Andres and Coupland, 2012). A recent transcriptome sequencing and analysis of the summer-flowering *C. azalea* has provided insights into its floral bud development, which characterized some genes involved in the photoperiodic flowering pathway (Fan et al., 2015). More information about this pathway from other *Camellia* plants will

Abbreviations: ABA, abscisic acid; CDS, coding DNA sequence; CrtBS, carotenoid biosynthesis; CTAB, cetyl trimethylammonium bromide; DE, differentially expressed; EST, expressed sequence tag; FA, fatty acid; FDR, false discovery rate; flaBS, flavonoid biosynthesis; FPKM, expected number of fragments per kilobase of transcript sequence per millions base pairs sequenced; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; Knt, kilo-nucleotides; KO, KEGG orthology; Mnt, million nucleotides; MUFA, monounsaturated fatty acid; NCBI, National Center for Biotechnology Information; NR, the non-redundant protein database; PA, proanthocyanidin; PUFA, polyunsaturated fatty acid; qRT-PCR, quantitative reverse transcription polymerase chain reaction; RNA-Seq, RNA sequencing; SFA, saturated fatty acid; SSR, simple sequence repeat; Swiss-Prot, UniProtKB/Swiss-Prot; TAGBS, triacylglycerol biosynthesis; TAIR10, *A.*thaliana proteins database; TrEMBL, UniProtKB/TrEMBL; UFA, unsaturated fatty acid.

facilitate horticulturists to breed new cultivars with a wide-range blooming periods.

Flavonoids and carotenoids, two major groups of colorful pigments, are supposed to be involved in flower coloration in *Camellia* species (Nishimoto et al., 2004; Li et al., 2009). The pathways of flavonoid biosynthesis (FlaBS) and carotenoids biosynthesis (CrtBS) have been clearly uncovered in *A. thaliana* and some other plant species (Cunningham and Gantt, 1998; Winkel-Shirley, 2001; Cazzonelli and Pogson, 2010). Flavonoids are derived from phenylalanine. For the first steps, PHENYLALANINE AMMONIA-LYASE (PAL) catalyzes the phenylalanine into cinnamate, and subsequently, CINNAMATE 4-HYDROXYLASE (C4L) and 4-COUMARATE-COA LIGASE (4CL) catalyze the conversion of cinnamate to p-coumaroyl-CoA. Then, CHALCONE SYNTHASE (CHS) as a rate-limiting enzyme converts the p-coumaroyl-CoA into chalcone (Dao et al., 2011). Flavones, flavonols, anthocyanins and proanthocyanidins (PAs) are synthesized from the common chalcone precursor along the FlaBS pathway. Evidences indicated that FlaBS pathway genes are largely regulated at the transcription level by a complex of transcription factors (TFs) including R2R3 MYB TFs (Czemmel et al., 2012) and basic helix–loop–helix (bHLH) TFs (Nesi et al., 2000). For example, the MYBA1 in *Vitis vinifera* is a R2R3 MYB TF specifically controlling the expression of *UFGT*, which encodes an enzyme responsible for the anthocyanin biosynthesis and able to determine the grape berry colors (Walker et al., 2007). Carotenoids are a group of isoprenoid molecules with characteristic color in the yellow to red range (Cazzonelli and Pogson, 2010). Carotenoids are synthesized from the five-carbon building blocks isopentenyl diphosphate (IPP) and its double-bond isomer dimethylallyl diphosphate (DMAPP), both of which are produced by the plastid-localized methylerythritol phosphate pathway (Phillips et al., 2008). Three IPP molecules are added to DMAPP by GERANYLGERANYL DIPHOSPHATE SYNTHASE (GGPS) to generate geranylgeranyl diphosphate (GGPP), a common precursor for the biosynthesis of carotenoids and several other groups of plastidic isoprenoids (Hirschberg, 2001). Although genes involved in these pathways have been extensively studied in other plants, more studies on these networks are needed in *Camellia* species. Transcriptome sequencing uncovered the candidate genes in the FlaBS pathway in *Camellia*, such as *C. sinensis* (Shi et al., 2011), *C. taliensis* (Zhang et al., 2015), and *C. chekiangoleosa* (Wang et al., 2014). However, there is still a lack of comprehensive expression profiling of the FlaBS genes. In particular, there is so far no information for the CrtBS pathway in *Camellia* plants.

The economic importance of *Camellia* species is largely due to the demand for young leaves (e.g., tea leaves from *C. sinensis*), ornamental flowers, and seed oil. Genes that are involved in TAGBS, FlaBS and CrtBS pathways and photoperiodic flowering are supposed to be highly relevant to the above-mentioned agricultural and horticultural traits. While comprehensive analysis on gene networks related to photoperiodic flowering and CrtBS is so far absent in *Camellia* plants, genes involved in TAGBS and FlaBS were reported in *C. sinensis* (Shi et al., 2011), *C. chekiangoleosa* (Wang et al., 2014), and *C. oleifera* (Xia et al., 2014). But, they are still far away from being well characterized in *Camellia* because of scarce information of gene expression profiles. Indeed, although RNA sequencing (RNA-Seq) datasets have been reported in several *Camellia* species, large-scale gene expression profiles were only available in *C. sinensis* to identify genes activated during cold acclimation (Wang et al., 2013), because other studies are unable to provide a broad expression analysis because of the limitation of sequencing strategies, such as 454 sequencing (Wu et al., 2013; Wang et al., 2014; Xia et al., 2014) or Illumina sequencing with a single library (Shi et al., 2011). Compared with the other *Camellia* species, very few genomic resources are available for *C. reticulata*. Nevertheless, a full catalog of genes and their expression profiles in *C. reticulata* are needed to better understand the above mentioned biological pathways.

Here, Illumina RNA-Seq was employed to *de novo* sequence the transcriptome of the diploid *C. reticulata*. The transcriptomes from the five tissues under normal development conditions were separately sequenced and compared to present a global survey of expression profiles as well as differential expression analysis in *C. reticulata*. A large set of unigenes was obtained to identify the majority of genes related to TAGBS, FlaBS, CrtBS, and photoperiodic flowering pathways. The characterization of EST-SSRs (expressed sequence tag, simple sequence repeat) from the transcriptome of *C. reticulata* has expanded valuable marker resources in breeding programs of the camellia community.

MATERIALS AND METHODS

Plant Material, RNA Isolation and Illumina Sequencing

The wild *C. reticulata* plant YB1-2, which was previously characterized as a diploid (Xia et al., 1994; Liu and Gu, 2011; Huang et al., 2013), was collected from Sichuan Province, China (**Supplementary Table S1**). A total of five tissues, including leaf buds, mature leaves, flower buds, flowers and immature fruits, were used for transcriptome sequencing and quantitative reverse transcription polymerase chain reaction (qRT-PCR) validation; blackening seeds were only used for qRT-PCR (**Supplementary Figure S1**). Samples were immediately frozen in liquid nitrogen and stored at −80°C. Total RNA of each sample was isolated using a modified CTAB (cetyl trimethylammonium bromide) method (Li L. et al., 2008). RNA samples were treated with RNase-free DNase I (Takara) to avoid DNA contamination. RNA quality was assessed using an Agilent 2100 BioAnalyzer.

For each RNA sample, Poly-A mRNA was enriched and then used to prepare a 350-bp paired-end cDNA library (2 × 100 nt) according to the Illumina protocol. Paired-end sequencing was performed on the Illumina HiSeq™ 2000 platform.

Sequence Data Processing and *de novo* Assembly

FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to evaluate and visualize the sequence quality before and after trimming process. The trimming process was performed with Trimmomatic (Lohse et al., 2012), sequentially removing adapters, the first seven bases at the 5′ end of the reads

due to sequence bias (Hansen et al., 2010), low quality bases (phred score < 20), and short reads (<50 nt). The clean reads were *de novo* assembled using Trinity software (Grabherr et al., 2011), with parameter settings of "–group_pairs_distance 450 –min_contig_length 200 –CPU 4" and default options otherwise.

Post-Assembly Processing

Three rounds of filtering process were performed to construct a high-quality nuclear transcriptome assembly. First, putative protein-coding sequences most likely from plant epiphytes and pathogens were identified and discarded using a previously described taxonomy-based method (Krasileva et al., 2013). Next, a second round of filtering process was performed using DeconSeq (Schmieder and Edwards, 2011) against a local database constructed from genomic sequences of candidate contaminant species (i.e., human, bacteria, virus, fungi, and top 20 epiphyte or pathogen species inferred from the taxonomy-based filtering step). Finally, mitochondrial and plastidial transcripts were identified and removed by searching against publicly available mitochondrion and plastid genomes of dicotyledonous plants.

Then the filtered sequences were subjected to cluster analysis using CD-HIT-EST (Huang et al., 2010) with the following parameters: -r 1 and -c 0.9. The longest sequence of each cluster was used as a representative sequence (i.e., unigene).

Sequence Annotation

Homology searches were performed against the UniProtKB/Swiss-Prot (Swiss-Prot), UniProtKB/TrEMBL (TrEMBL), NCBI (National Center for Biotechnology Information) non-redundant protein database (NR), *A. thaliana* proteins database (TAIR10), and *V. vinifera* proteins database (IGGP 12X) using a BLASTx procedure with an *e*-value threshold of 10^{-3}. For each unigene, the best hit against these databases was used to predict sequence orientation, coding DNA sequence (CDS), and polypeptide sequence with GeneWise software (Birney and Durbin, 2000). The predicted peptide sequences were analyzed to assign corresponding protein domains and families by using the pfam_scan tool against the Pfam database (Punta et al., 2012). Gene ontology (GO) were obtained based on the NR database annotations using Blast2GO software (version 2.6.5) (Conesa et al., 2005), and then WEGO software (Ye et al., 2006) was used to obtain GO functional classifications. The unigenes were also submitted to the online KEGG (Kyoto Encyclopedia of Genes and Genomes) Automatic Annotation Server (KAAS) to obtain enzyme commission (EC) numbers and associated KEGG orthology (KO) identifiers that are directly linked to objects in the KEGG pathway map (Moriya et al., 2007). The KAAS annotation was performed with single-directional best hit method using angiosperm species data sets as reference. We focused on unigenes assigned to four pathways: photoperiodic flowering, TAGBS, FlaBS, and CrtBS. Finally, we performed a manual curation for these unigenes.

Transcript Abundance and Expression-Based Analysis

Paired end reads were aligned to the assembly by Bowtie (Langmead and Salzberg, 2012), and the resulting alignments were used to estimate expression abundances in FPKM (expected number of fragments per kilobase of transcript sequence per millions base pairs sequenced) by RSEM (Li and Dewey, 2011). The Bioconductor tool edgeR (Robinson et al., 2010) was employed to calculate the expression abundance fold change based on pairwise comparisons of normalized FPKM among five sequenced tissues using exact statistical method with default settings. Differentially expressed genes (DEGs) were defined with a threshold of fold change ≥ 4 and false discovery rate (FDR) ≤ 0.001. Enrichment analyses of DE gene sets in KEGG pathways or the GO database were performed using the online tool KOBAS (KEGG Orthology Based Annotation System, http://kobas.cbi.pku.edu.cn/) (Xie et al., 2011) with the expressed genes (FPKM ≥ 1) as background (corrected $p \leq 0.05$).

qRT-PCR Analysis

Primers used for qRT-PCR assays were designed with the Primer Express software (version 3.0, Applied Biosystems), and they amplified PCR products varied from 80 to 159 bp (**Supplementary Table S2**). The housekeeping gene *ELONGATION FACTOR 1 ALPHA* (*EF1A*) was used as internal reference control for normalization (Nicot et al., 2005). qRT-PCR assays were performed as described previously (Xu et al., 2011). The relative expression of the genes was calculated using the $2^{-\Delta\Delta Ct}$ method.

EST-SSR Identification and Polymorphism Survey

To increase confidence, only unigenes with length ≥500 nt were used for EST-SSR identification. They were searched by using the MISA program (Thiel et al., 2003), mining EST-SSRs with 2- to 6-nt motifs and a minimum length of 12 nt.

Twenty EST-SSRs were chosen for polymorphism survey in 24 individual *C. reticulata* plants (**Supplementary Tables S1**, **S3**). PCR primer pairs were designed using Invitrogen Vector NTI (version 10) with standard criteria. DNA extraction was performed as described previously (Doyle and Doyle, 1987). Standard PCR amplifications and electrophoresis were performed as described previously (Tong et al., 2013). The number of alleles per locus (NA), effective number of alleles (NE), Shannon's diversity index (I), observed heterozygosity (Ho), and expected heterozygosity (He) were estimated with POPGENE software (version 1.32) (Yeh et al., 1999). The polymorphic information content (PIC) value was calculated according to a previously described formula (Botstein et al., 1980).

RESULTS AND DISCUSSION

Sequencing, *de novo* Assembly and Assessment of the *C. reticulata* Transcriptome

Distinct cDNA libraries of leaf buds, mature leaves, flower buds, flowers, and immature fruits (**Supplementary Figure S1**) from a wild diploid *C. reticulata* plant were separately sequenced with Illumina HiSeq™ 2000, generating a total of 394.9 million 100 nt paired-end raw reads. After a stringent trimming process, about 311.3 million clean reads (78.8% of total raw data) with

TABLE 1 | Summary of the sequencing data and transcriptome assembly of *C. reticulata*.

	Raw reads	Clean reads	Retain rate (%)	Clean nucleotides (nt)
Leaf buds	75,791,012	57,655,612	76.1	4,700,656,880
Mature leaves	95,982,836	74,474,740	77.6	6,113,913,256
Flower buds	70,418,086	57,626,233	81.8	4,929,278,585
Flowers	71,987,410	59,351,076	82.4	5,114,541,561
Immature fruits	80,714,310	62,182,007	77.0	5,323,094,674
Total	394,893,654	311,289,668	78.8	26,181,484,956
	Contigs	**Unigenes**		
Total number	232,428	141,460		
Total length (nt)	186,434,209	96,117,212		
Mean length (nt)	802	679		
N50 (nt)	1,255	1,080		
GC content (%)	42.4	41.8		
Number of length ≥ 500 nt	119,933	57,162		
Number of length ≥ 1000 nt	63,845	27,880		
Reads mapping rate (%)	84.5	80.8		

26.2 billion nucleotides in total were retained (**Table 1**). The high-quality reads were *de novo* assembled into contigs with Trinity, which has been shown to be the best single k-mer assembler for *de novo* assembly from RNA-Seq short reads (Grabherr et al., 2011; Zhao et al., 2011). As direct output from the assembly procedure often includes contaminants and redundant sequences, assembled sequences were subjected to do a post-assembly processing to obtain a high-quality nuclear transcriptome (see Material and methods). As a result, an assembly of 232,428 contigs with a total size ∼186.4 million nucleotides (Mnt) was established for *C. reticulata*. After clustering analysis, 141,460 unigenes with a total size ∼96.1 Mnt were generated (**Table 1**). The unigenes ranged from 200 to 9,880 nt, with a mean length of 679 nt and an N50 length of 1080 nt (**Table 1** and **Figure 1A**).

Though a systematic evaluation standard is not yet available for transcriptome assembly of non-model organisms (Martin and Wang, 2011), we tried to overview the quality of our assembly with several broadly used parameters and by comparing it to independent databases from closely related species. Reads mapping showed that the proportion of reads assembled was 84.5% (**Table 1**), which is a comparable alignment rate to that of other *de novo* assemblies. 86.3% of the mapped paired-end reads aligned concordantly, showing good physical evidence of sequence contiguity. The unigenes was compared to available ESTs and protein sequences of *Camellia* plants from NCBI. Of 111,905 non-redundant *Camellia* ESTs, 96,092 (85.9%) ESTs were represented in our assembly (Megablast, $E = 10^{-9}$), among which 72,924 (75.9%) ESTs were matched with more than 80% identity and 80% coverage (**Supplementary Figure S2**). Of 256 non-redundant *Camellia* proteins (≥300 AA), 255 (99.6%) matched to at least one assembled unigene (BLASTx, $E = 10^{-9}$), among which 215 (84.3%) were represented with ≥80% coverage. These results demonstrated that the assembly successfully constructed a large number of homologous transcripts with desirable lengths. Furthermore, chimeric assembly level was inspected using the top longest assembled sequences (Van Belleghem et al., 2012). All of the 10 longest unigenes (≥7400 nt) in our assembly positively matched with long gene sequences in public databases without obvious chimeric evidence, and 9 of them were confirmed to have good alignment confidence (70% identity and 80% coverage) at both the cDNA and protein levels (**Supplementary Table S4**). These assessments indicate that the *C. reticulata* transcriptome constructed herein possessed desirable completeness, accuracy, and contiguity.

To facilitate the access and utilization of the *C. reticulata* transcriptome data, the sequences of raw reads and unigenes were deposited in the NCBI (BioProject ID PRJNA297756). The full list of transcript sequences is available upon request.

Functional Annotation and Gene Ontology Classification

To annotate the transcriptome with putative functions, similarity searches for each unigene were performed using BLASTx against five public databases: Swiss-Prot, TrEMBL, NR, TAIR10 and the *Vitis vinifera* protein database. With an *E*-value threshold of 10^{-3}, 69,922 (49.4% of the total) unigenes were significantly matched with known proteins in at least one of the searched databases (**Table 2** and **Supplementary Table S5**), and 42,042 (29.7% of the total) unigenes had significant hits with proteins presented in all of the five databases (**Table 2**). All of the 69,922 unigenes matched to known proteins were considered as putative protein-coding sequences. The length distribution of these putative protein-coding sequences is shown in **Figure 1A**, showing that the longer sequence would have a higher probability of being matched.

For the NR annotations, the top-hit species distribution is shown in **Figure 1B**. We found that 36.75% of the mapped unigenes had the highest matches to genes from *V. vinifera*, followed by *Theobroma cacao* (10.02%) and *Solanum lycopersicum* (6.85%) (**Figure 1B**). Overall, 54.29% of the matched sequences were mapped with similarity >80% (**Figure 1C**), and 76.01% of the matched sequences showed strong homology with an $E < 10^{-100}$ (**Figure 1D**). Notably,

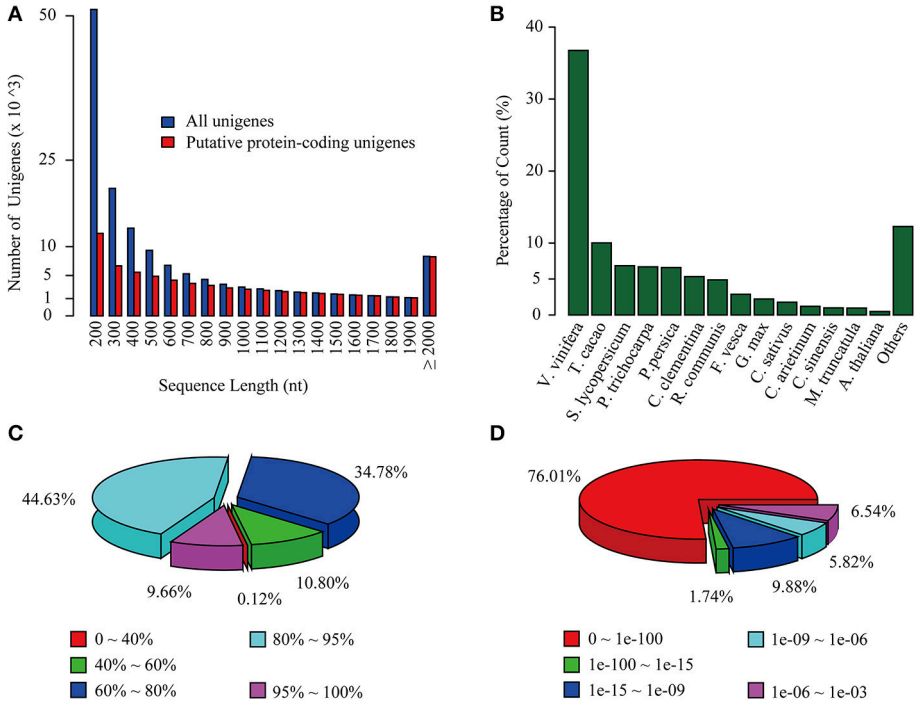

FIGURE 1 | Overview of *C. reticulata* transcriptome assembly and its homology search against the NR database. (A) Length distribution of total unigenes (blue) and putative protein-coding unigenes (red). **(B)** Species, **(C)** similarity, and **(D)** *E*-value distributions of the top BLAST hits for each unigene against the NR database.

TABLE 2 | Summary of functional annotation for *C. reticulata* unigenes.

Annotation	Number of unigenes	Percentage of matched unigenes (%)	Number of unique homologs
Total unigenes	141,460		
Annotated unigenes	69,922	49.4	
BLASTx AGAINST PUBLIC PROTEIN DATABASES			
NR	68,026	48.1	41,413
TrEMBL	62,618	44.3	37,374
Swiss-Prot	49,469	35.0	16,373
A. thaliana proteins (TAIR10)	56,664	40.1	17,368
V. vinifera proteins	59,968	42.4	15,763
Unigenes matched at least one database	69,922	49.4	
Unigenes matched all five databases	42,042	29.7	
DOMAIN/GO ANNOTATION			
Search against PFAM database	47,693	33.7	
Annotated with GO terms	39,301	27.8	

the annotated genes of *Aribidopsis* and *V. vinifera* on average have more than three counterparts in *C. reticulata* transcriptome (17,368 vs. 56,664 and 15,763 vs. 59,968, respectively, see **Table 2**), suggesting that the hypothesis of genome duplication events in *C. sinensis* probably might occur in *C. reticulata* (Shi et al., 2011).

Sequence orientation, CDS, and polypeptide sequence were predicted using GeneWise software (Birney and Durbin, 2000). For each putative polypeptide sequence, its functional domain and associated family was inferred by pfam_scan analysis against the Pfam database. The top 10 abundant families and domains are listed in **Supplementary Table S6**. The "pentatricopeptide repeat" (PF01535), which is thought to be involved in organelle biogenesis (Lurin et al., 2004), was the most abundant protein family. The "protein kinase domain" (PF00069), which functions in a process called phosphorylation, was the most abundant protein domain.

The GO classification system that describes gene function into three major categories (biological processes, molecular functions, and cellular components) and additional subcategories was also applied to putative gene functions. Overall, 39,301 unigenes were assigned to 131,466 GO terms (3736 unique GO terms; **Supplementary Table S7**) that were further classified into 47 subcategories (**Supplementary Figure S3**). The most abundant GO subcategories for biological processes, molecular functions, and cellular components were metabolic process (GO:0008152), binding (GO:0005488), and cell (GO:0005623), respectively.

KEGG Pathway Analysis

KEGG annotation and pathway assignment can help clarify the biological functions of genes in terms of networks (Moriya

et al., 2007). A total of 21,940 unigenes were matched to 3,278 unique KO groups (**Supplementary Table S7**), and 6,803 unigenes were assigned with 769 unique EC numbers. Pathway mapping assigned 13,316 unigenes (2,116 KO groups) into 333 functional pathways (**Supplementary Table S7**). "Metabolic pathways" (ko01100) had the largest number of KO identifiers (794, 37.5%), followed by "biosynthesis of secondary metabolites" (352, 16.6%, ko01110), "microbial metabolism in diverse environments" (129, 6.1%, ko01120), "ribosome" (119, 5.6%, ko03010), and "spliceosome" (103, 4.9%, ko03040). The KEGG pathway annotation results provided valuable information that allowed us to relate genes to specific physical processes such as TAGBS, FlaBS, CrtBS, and photoperiodic flowering, which will be demonstrated in the following paragraphs.

EST-SSR Marker Identification and Polymorphism Survey

57,162 unigenes with a length of at least 500 nt were used to identify EST-SSRs. As a result, a total of 40,823 EST-SSRs were detected in 25,188 unigenes (44.1% of the total searched unigenes; **Table 3** and **Supplementary Table S8**), equivalent to an average frequency of one SSR per 1.74 kilo-nucleotides (Knt) of the transcriptome sequences. This result was slightly more frequent than those reported in a corresponding study on tea (2.41 Knt) that used different parameters (Tan et al., 2013). The most abundant repeat type was trinucleotide (14,688, 36.0%), followed by dinucleotide (13,837, 33.9%), and tetranucleotide (5,759, 14.1%). Out of 411 repeat motifs identified, the most frequent was AG/CT (11,367, 27.8%), followed by AAG/CTT (3,369, 8.3%), ACC/GGT (2,711, 6.6%), ATC/ATG (2,176, 5.3%), and AGG/CCT (1,646, 4.0%) (**Supplementary Table S8**).

Out of 20 EST-SSR primer pairs chosen for polymorphism survey, 18 were successfully amplified in 24 *C. reticulata* accessions (**Supplementary Tables S1**, **S3** and **Supplementary Figure S4A**). Among the 18 working primer pairs, 17 successfully amplified PCR products with the expected sizes, and one pair (SSR11) generated a larger PCR product (∼270 bp) than expected size of 159 bp, suggesting that there may be a short intron within the amplicon. Our polymorphism survey showed that seven primers pairs could amplify polymorphism bands among the 24 *C. reticulata* representative samples (**Supplementary Figures S4B–H**). The PIC values varied from 0.33 to 0.69 with an average of 0.50, and the NA is 3.57 (**Table 4**). These results clearly demonstrated that EST sequences derived from RNA-Seq are effective and valuable resources to develop polymorphic SSR markers, which are essential in various applications such as population genetic structuring and linkage mapping.

Transcriptome Expression Profiling and Differential Expression Analysis

Expression levels of unigenes were determined by aligning the RNA-Seq reads from each library to the assembly. The default empirical abundance threshold of 1 FPKM was used to evaluate whether a gene was positively expressed (Vogel and Marcotte, 2012; Gonzalez-Porta et al., 2013). As a result, 94,450 unigenes were positively expressed in at least one of the libraries. Among them, 17,956 unigenes were expressed in all five tissues, and 8,837, 13,963, 3,265, 10,540, and 7,094 unigenes were specifically expressed in leaf buds, mature leaves, flower buds, flowers, and immature fruits, respectively (**Supplementary Figure S5A**). Based on expression levels, the sequence depths of expressed unigenes varied across a broad range. The top 1% of abundantly expressed unigenes accounted for 60% of the overall abundance value, while the top 10% abundantly expressed unigenes were 86% of the overall abundance value (**Supplementary Figure S5B**).

A total of 22,229 unigenes were defined as DEGs (fold change = 4 and FDR = 0.001); and 18,639 (83.85%) of these were putative protein-coding genes, indicating specific enrichment of functional genes in this dataset. All of the DEGs were assigned to five groups (DELB, DEML, DEFB, DEFL, and DEFR) according to the tissue identity where their highest expression occurred (**Supplementary Table S9**). As a result, Group DEML was assigned with the maximum number of DEGs (5,451), followed by Group DELB (5,073), Group DEFL (4,599), Group DEFR (3,756), and Group DEFB (3,350) (**Supplementary Table S9**). KEGG pathway or GO enrichment analysis of the DE unigenes in each group indicated that unigenes preferentially expressed in a tissue are highly related to the specific functions of that tissue (**Supplementary Table S10**). For example, Group DEML expressed most preferentially in mature leaves enriched genes involved in "photosynthesis," "carbon fixation," "fructose, mannose and nitrogen metabolism," "porphyrin and chlorophyll metabolism," as well as "carotenoid biosynthesis." Group DEFB and DEFL showed preferential expression of genes involved in different functional processes, which are related to flower development in different stages (flower buds and flowers), such

TABLE 3 | EST-SSRs present in the *C. reticulata* transcriptome.

Repeat type	Number of repeat units											
	3	4	5	6	7	8	9	10	11	12	≥13	Total
Dinucleotide	–	–	–	3650	2723	2807	2780	1511	353	13	4	13,841
Trinucleotide	–	8891	3246	1596	862	83	3	5	0	0	2	14,688
Tetranucleotide	4560	842	303	45	4	0	3	1	1	0	0	5759
Pentanucleotide	2215	594	57	5	0	0	0	0	0	0	0	2871
Hexanucleotide	2939	634	48	29	8	1	3	2	0	0	0	3664
Total												40,823

TABLE 4 | Allelic diversity attributes of seven polymorphic EST-SSRs.

	N	NA	NE	PIC	Ho	He	I
SSR1	24	4	2.40	0.41	0.71	0.58	1.01
SSR3	24	3	2.78	0.56	0.71	0.64	1.06
SSR11	23	3	2.38	0.54	0.87	0.58	0.94
SSR16	23	2	1.84	0.41	0.54	0.46	0.65
SSR17	24	4	2.19	0.33	1.00	0.54	0.89
SSR18	24	5	3.71	0.69	0.58	0.73	1.41
SSR19	24	4	2.59	0.55	0.92	0.61	1.06
Mean	24	3.57	2.56	0.50	0.76	0.59	1.00

N, The sample size.

as "plant hormone signal transduction," "pigment accumulation" (GO:0043476), "carotenoid biosynthesis" and "phenylpropanoid biosynthesis." Genes from group DEFR were highly expressed in immature fruits, and included genes functioned in "flavonoid biosynthesis," "phenylpropanoid biosynthesis," "flavone and flavonol biosynthesis," and "biosynthesis pathways of secondary metabolites." Collectively, the DEGs are able to mirror physiological differences among the five tissues. These results will facilitate the identification of genes that function in specific physiological programs.

qRT-PCR Validation

To evaluate the reliability of RNA-Seq analysis, 22 gene homologs related to TAGBS and FlaBS pathways were selected for qRT-PCR test in six tissues (leaf buds, mature leaves, flower buds, flowers, immature fruits, and blackening seeds) (**Supplementary Figure S1** and **Supplementary Table S1**). In general, our qRT-PCR results show a high degree of consistency with the RNA-Seq results (**Figures 2**, **3**). Expression patterns of nine (40.9%) cases (**Figures 2J,K**, **3A–E,H,J**) fit well with the RNA-Seq results across all five tissues. Ten (45.5%) cases (**Figures 2B,C,E–G,I**, **3F,G,I,K**) had almost similar expression patterns but with very small partial inconsistences compared to the RNA-Seq results. It is rational and acceptable that there are certain differences in direct comparisons between the RNA-Seq and qRT-PCR results due to different normalization methods, bias in library preparation in RNAseq, and other technical biases (Bustin, 2002; Li et al., 2010; Zheng et al., 2011).

Triacylglycerol Biosynthesis Pathway in *C. reticulata*

Based on KEGG pathway annotation of the *C. reticulata* transcriptome, we totally identified 93 unigenes for the TAGBS pathway (**Supplementary Table S11**; **Figure 4**). Most of the genes involved in TAGBS pathway were present in the assembly, except for two *KCS* genes, *FATTY ACID ELONGATION 1* (*FAE1*, namely *KCS18*) and *KCS9* (**Supplementary Table S11**). We failed to identified *FAE1* and *KCS9* in the traditional camellia oil plant *C. oleifera*, based on previously published transcriptome dataset (Xia et al., 2014). *FAE1* is involved in the elongation of C_{18} to C_{22} FAs (Kunst et al., 1992), while *KCS9* is involved in the elongation of C_{22} to C_{24} FAs (Kim et al., 2013). *Arabidopsis fae1* mutant had a deficiency in producing very long-chain FAs (VLCFA) in seed oil (Kunst et al., 1992). Thus, the absence of *FAE1* might be an important factor contributing to the camellia oil with very low level of VLCFA.

The genes of this pathway showed high similarities to *Arabidopsis* with an average identity of 70% at amino acid sequence level, suggesting their potentially functional conservation. Many of these genes had multiple copies in *C. reticulata* compared with *Arabidopsis* (63 duplicated unigenes matched to 24 *Arabidopsis* unique homolog genes), such as *STEAROYL-ACP DESATURASE* (*SAD*, 3 unigenes), *FAD2* (2 unigenes), and *DGAT2* (2 unigenes). This phenomenon was also found in other *Camellia* species including *C. sinensis*, *C. taliensis,* and *C. oleifera* (data not shown). The results seemingly support the hypothesis that *Camellia* species might share whole genome duplication events during the course of evolution before their speciation (Shi et al., 2011). Gene expression profiles of the TAGBS pathway were shown in **Figure 5A**. For all of the 93 TAGBS unigenes, 42 were defined as DEGs (fold change = 4, FDR = 0.001). There are 29 (69%) DEGs up-regulated in flowers or flower buds, and some of them are FA desaturase homologs, including *SAD_a*, *FAD2_a*, *FAD2_b*, *FAD3*, *FAD8_b*, and *FAD8_c*. These results indicate that a variety of fatty acids could be demanded for flower development. FA desaturase genes are often related to the modification of membrane fluidity in response of cold hardiness (Kodama et al., 1994; Matteucci et al., 2011; Wang et al., 2013). The flowers of *C. reticulata* often bloom in winter when the temperatures are quite low (**Supplementary Figures S6E,F, S7**). Thus, high expression of these desaturase genes may explain the freezing tolerance of *C. reticulata* flowers and flower buds in winter.

Camellia oil contains 60–80% oleic acid and 5–10% linoleic acid. It is well characterized in *Arabidopsis* that plastid-localized enzymes were responsible for oleic acid biosynthesis with 18:0-ACP as precursor: SAD catalyzes the conversion of stearoyl-ACP to oleoyol-ACP (Shanklin and Somerville, 1991), while FATTY ACYL-ACP THIOESTERASE A (FATA) plays a key role in the formation of free oleic acid (Hellyer et al., 1992; Moreno-Perez et al., 2012). *FAD2* encodes an oleate desaturase that required for PUFA biosynthesis in endoplasmic reticulum (Okuley et al., 1994). In this study, *FATA* gene was observed with only one copy, while *SAD* and *FAD2* were observed with 3 and 2 copies, respectively (**Supplementary Table S11**). Different expression patterns were observed among the copies of *SAD* and *FAD2*.

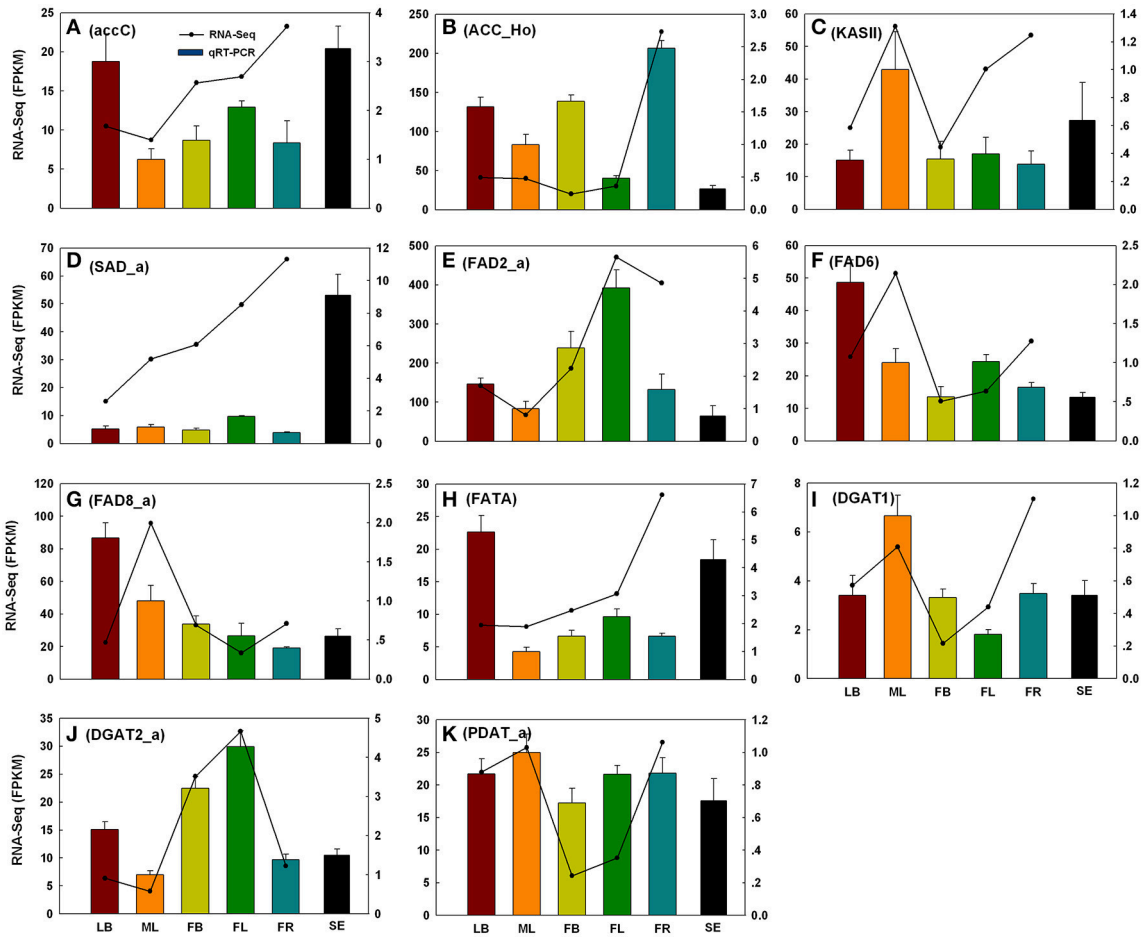

FIGURE 2 | qRT-PCR validations of 11 putative genes involved in TAG biosynthesis. The histograms show the qRT-PCR results of accC **(A)**, ACC_Ho **(B)**, KAS II **(C)**, SAD_a **(D)**, FAD2_a **(E)**, FAD6 **(F)**, FAD8_a **(G)**, FATA **(H)**, DGAT1 **(I)**, DGAT2_a **(J)**, and PDAT_a **(K)**; the line charts show the FPKM values of these genes. qRT-PCR results represent the mean (±SD) of three biological replicates. Gene abbreviations can be referenced in **Supplementary Table S12**. LB, leaf buds; ML, mature leaves; FB, flower buds; FL, flowers; FR, immature fruits; SE, seeds.

FAD2_a was up-regulated in young fruit, while *FAD2_b* was down-regulated in young fruit (**Figure 5A**). Clustering analysis of gene expression profiles shows that, *SAD_a* and *FATA* were co-expressed genes in the same subcluster, while *SAD_b* and *SAD_c* were in different subclusters (**Figure 5A**). Using qRT-PCR, we also investigated the expression levels of several TAGBS-related genes in ripening seeds. While *FAD2_a* had a lower expression level in seeds than in mature leaves (**Figure 2E**), *SAD_a* and *FATA* had several fold higher expression level in seeds than in mature leaves (**Figures 2D,H**). These observations suggest that *SAD_a* and *FATA* may be key genes responsible for MUFA production in *C. reticualta* seeds. Up-regulation of *SAD_a* gene and down-regulation of *FAD2_a* gene may be an efficient way to control the production ratio of MUFAs to PUFAs. Previously, Xia et al. (2014) found that parallel evolution of *FAD2* genes may occur between *Camellia* species and olive, which might result in the similar FA composition in their seed oil. Moreover, *FAD2* RNAi experiment in *Camelina sativa* showed evidence that *FAD2* is an efficient target for genetic manipulation toward lowering the PUFA content and increasing the oleic acid content (Nguyen et al., 2013).Our data here show that the transcription levels of *SAD*, *FAD2*, and *FATA* were highly regulated in seeds of *C. reticulata*. Their expression patterns may well explain why the FA composition of camellia oil was dominated by oleic acid but with relative low PUFA levels.

Photoperiodic Flowering Pathway in *C. reticulata*

To better understand the flowering transition process in *C. reticulata*, we conducted a field study to record the timing series of major vegetative and reproductive events over an entire year (**Supplementary Figures S6A–L**). Our observation showed that the flowering transition of *C. reticulata* often occurs at the apex of branch around mid-April when the day length is more than 12 h in Sichuan, China (**Supplementary Figures S6A**, **S7**). This result suggests that the photoperiodic flowering pathway may determine flowering time control of *C. reticulata*.

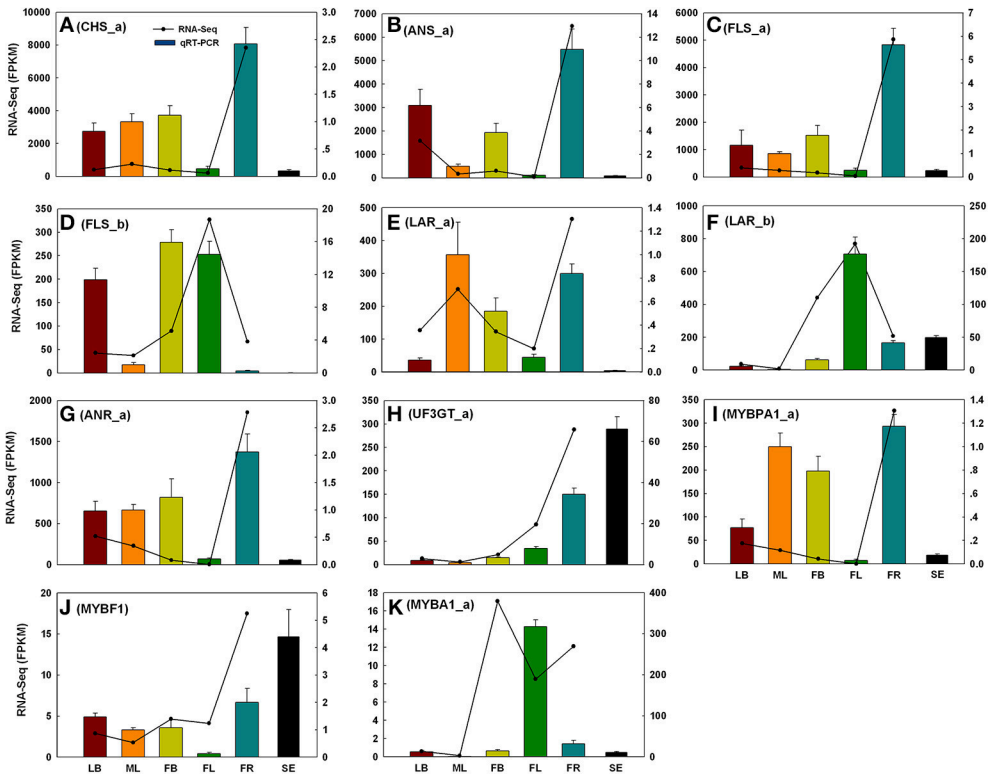

FIGURE 3 | qRT-PCR validations of 11 putative genes involved in flavonoid biosynthesis. The histograms show the qRT-PCR results of CHS_a **(A)**, ANS_a **(B)**, FLS_a **(C)**, FLS_b **(D)**, LAR_a **(E)**, LAR_b **(F)**, ANR_a **(G)**, UF3GT_a **(H)**, MYBPA1_a **(I)**, MYBF1 **(J)**, and MYBA1_a **(K)**; the line charts show the FPKM values of these genes. qRT-PCR results represent the mean (± SD) of three biological replicates. Gene abbreviations can be referenced in **Supplementary Table S13**. LB, leaf buds; ML, mature leaves; FB, flower buds; FL, flowers; FR, immature fruits; SE, seeds.

In the model plant, *A. thaliana*, genes involved in photoperiodic flowering pathway were well characterized (Andres and Coupland, 2012). Homology searches successfully identified 50 unigenes potentially involved in this pathway (**Supplementary Table S12**). Of them, 36 (72%) candidate genes were represented with full-length CDSs, such as *CRYPTOCHROMEs* (*CRYs*), *LATE ELONGATED HYPOCOTYL* (*LHY*), *TIMING OF CAB EXPRESSION 1* (*TOC1*), *FKF1*, *GI*, *SOC1*, and *FT*. However, *CO* and its transcription repressor *CYCLING-DOF-FACTOR-1 (CDF1)* were absent in the assembly. *CO* was expressed at an extremely low level in *Arabidopsis* and was only actively detected in leaves or shoots under long-day condition (Putterill et al., 1995). Thus, the possibility cannot be excluded that the unique expression pattern of *CO* led to the lack of itself in our results. Alternatively, we found several homologs of the *CO* gene family (e.g., *COL2, COL4, COL5,* and *COL9*; **Supplementary Table S12**). Similar to genes in TAGBS pathway, many genes involved in the photoperiodic flowering pathway appeared to have multiple copies in *C. reticulata*. For instances, *FD* was observed with 4 copies, and both *SOC1* and *AP1* were observed with three copies (**Supplementary Table S12**). Based on sequence comparisons, the candidate genes in this pathway showed moderate similarity to *Arabidopsis* with an average identity of 62% at amino acid sequence level. However, great sequence differences were found in *ELF3_a*, *ELF3_b,* and *LHY* with identities lower than 37%. The complex evolutionary process may bring about sequence variations and functional differentiation of these genes between *C. reticulata* and *Arabidopsis*.

The gene expression patterns of the photoperiodic flowering pathway genes across the tested tissues were shown in **Figure 5B**. A total of 20 putative genes were identified as DEGs, including *FT*, *SOC1* (*SOC1*_a and *SOC1*_c), *LFY* (*LFY_a* and *LFY_b*), and *AP1* (*AP1_a*, *AP1_b,* and *AP1_c*) (**Figure 5B**). The photoperiod and irradiance are mainly perceived by mature leaves. Nevertheless, many of these DEGs were not specially expressed in mature leaves. For examples, *SOC1_a* was up-regulated in mature leaves and leaf buds, and *FT* was up-regulated in mature leaves and young fruits (**Figure 5B**). These results suggest that these genes may play important roles in multiple developmental processes. The duplicated genes in this pathway demonstrate differential expression patterns across these tissues. For example, while *SOC1_a* was up-regulated in mature leaves and leaf buds, *SOC1_b* and *SOC1_c* were actively expressed in leaf buds and flowers (**Figure 5B**). Many studies suggested that the alteration of spatiotemporal expression is an important indicator of functional divergence in duplicated genes (Ganko et al., 2007). Thus, these duplicated

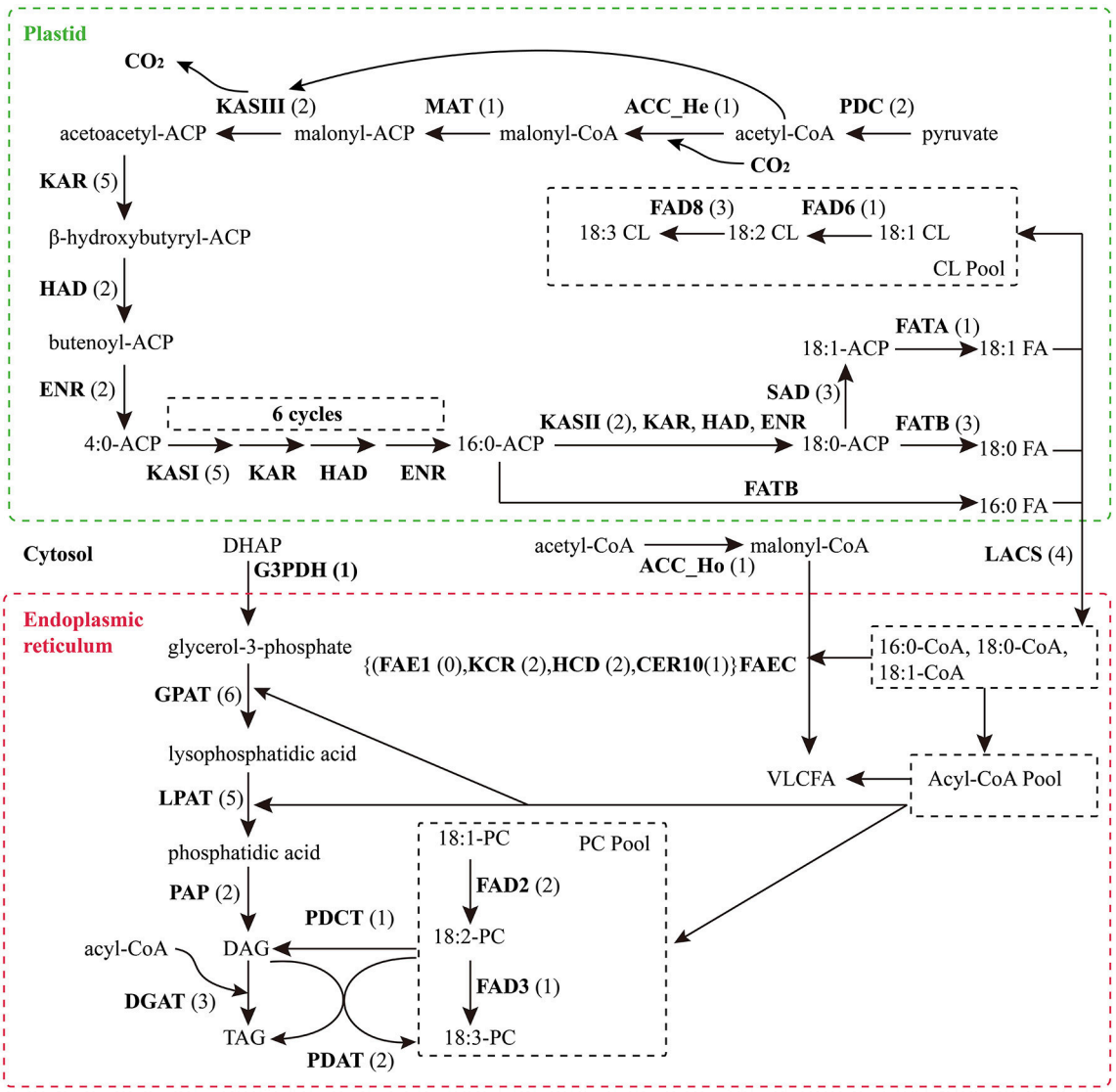

FIGURE 4 | Putative genes related to TAG biosynthesis in *C. reticulata*. The numbers in brackets following each gene name indicate the number of unigenes annotated to that gene. For each gene abbreviation, its full name and other information can be referenced in **Supplementary Table S11.**

genes may play diverse roles in multiple developmental processes.

Flavonoid Biosynthesis Pathway in *C. reticulata*

Flavonoids include the four major subgroups of flavones, flavonols, anthocyanins and PAs that play vital roles in flower coloration and many other plant physiological processes. Homology searches identified 56 unigenes as candidate genes that are related to the FlaBS pathway (**Supplementary Table S13**, **Figure 6A**). Most of these genes had multiple copies in *C. reticulata* compared to *Arabidopsis*, except for six genes (*ACC_Ho*, *FNSII*, *UGT75B2*, *AOMT*, *MYBF1,* and *F3′5′H*; **Supplementary Table S13**). Expression profiles for these genes were shown in **Figure 5C**. Overall, 89.3% (50/56) of these putative genes were defined as DEGs, indicating that most of these genes were extensively regulated at the level of transcription.

In grape, *VviMYB5* (*MYB5a* and *MYB5b*) (Deluc et al., 2006, 2008), *VviMYBF1* (Czemmel et al., 2009), *VviMYBPA1* (Bogs et al., 2007), and *VviMYBA1* (Kobayashi et al., 2004) were characterized as R2R3 MYB TF genes that play major roles in the production control of total flavonoids, flavonol, PAs, and anthocyanin pigments, respectively. The homologs of these TF genes were also identified in our analysis. Neither *MYB5a* nor *MYB5b* in *C. reticulata* is differentially expressed across tissues (**Figure 5C**), which is consistent to their putative roles as general regulators of flavonoid precursors (Deluc et al., 2006, 2008). *VviMYBF1* in grape has two functional MYB domains and serves as a specific regulator in expression control of *FLAVONOL*

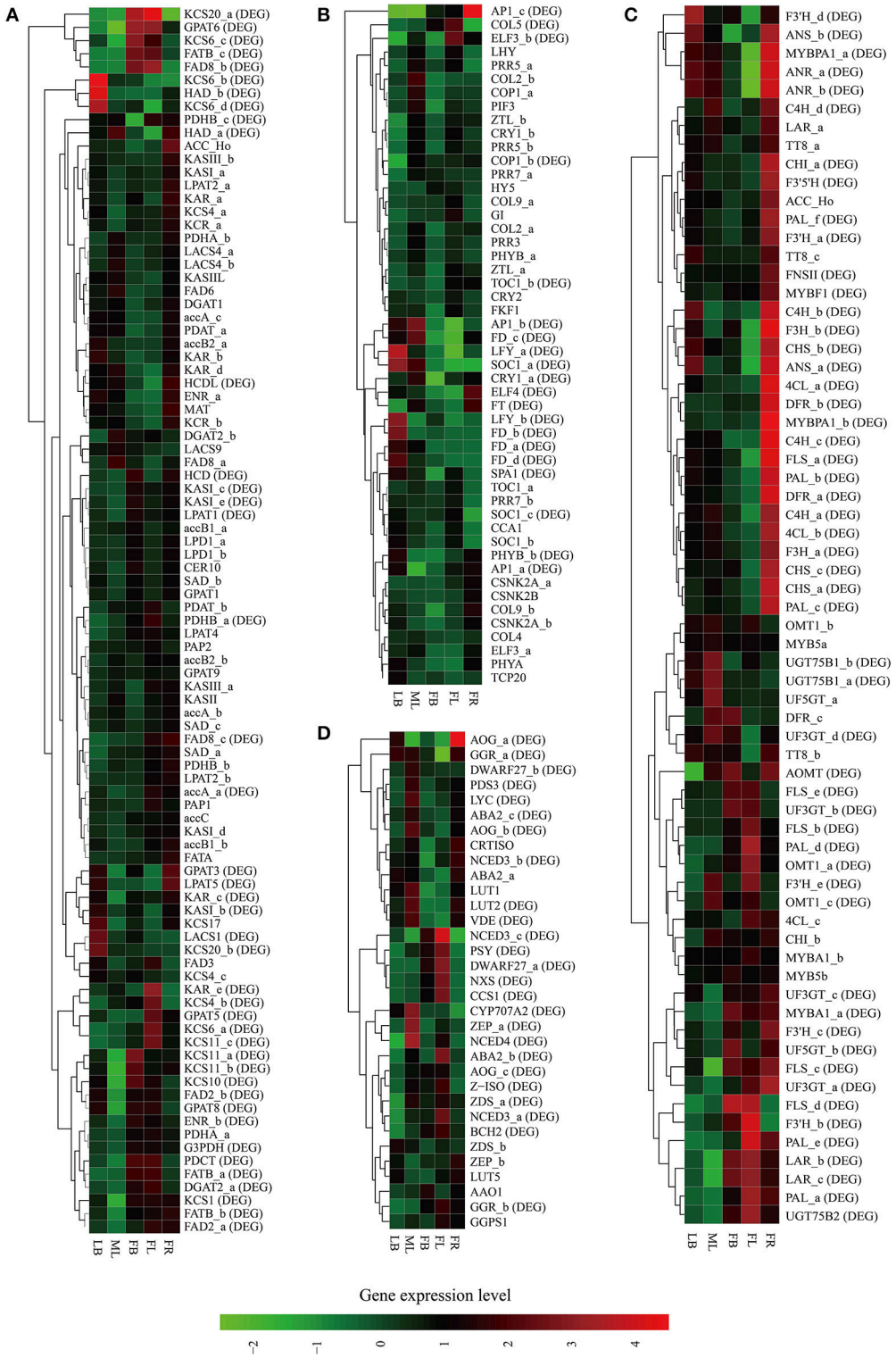

FIGURE 5 | Heat map representation and hierarchical clustering of putative genes involved in *C. reticulata* flowering time control **(A)**, TAG biosynthesis **(B)**, flavonoid biosynthesis **(C)**, and carotenoid biosynthesis **(D)**. For gene abbreviations in **Figures 4A–D**, their full name and other information can be referenced in **Supplementary Tables S11–S14**, respectively. Expression profiles of these genes across five tissues (LB, leaf buds; ML, mature leaves; FB, flower buds; FL, flowers; FR, immature fruits; SE, seeds) are shown, and DE genes (fold change ≥ 4, FDR ≤ 0.001) are indicated as "DE" in brackets. Green and red colors are used to represent low-to-high expression levels, and color scales correspond to the mean centered log2-transformed FPKM values. A hierarchical cluster dendrogram is shown on the left.

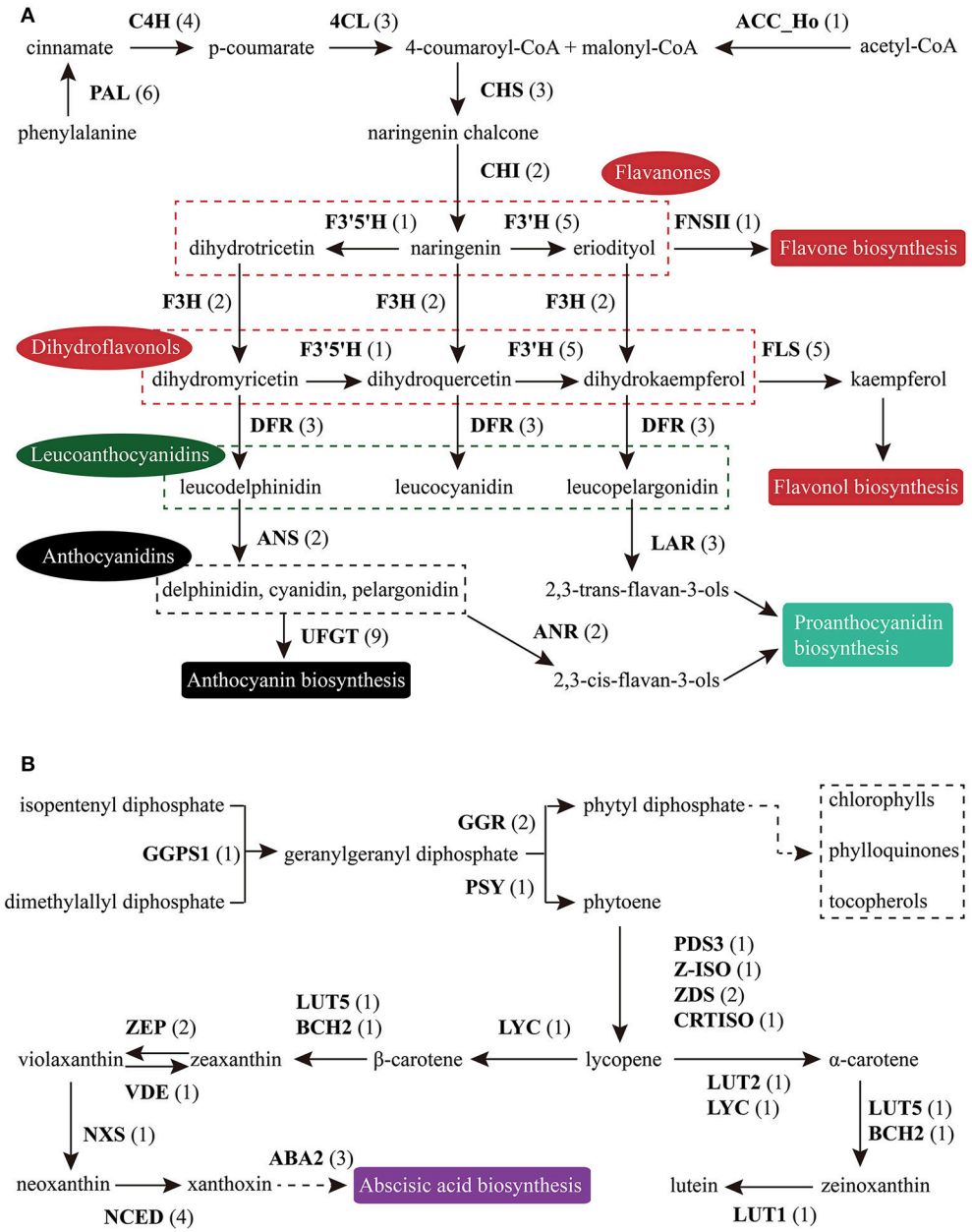

FIGURE 6 | Putative genes involved in FlaBS (A) and CrtBS (B). The numbers in brackets following each gene name indicate the number of unigenes annotated to that gene. For each gene abbreviation, its full name and other information can be checked in **Supplementary Table S13**, **14**.

SYNTHASE (*FLS*) gene to regulate the flavonol production (Czemmel et al., 2009). The *C. reticulata MYBF1* was found with only a single MYB domain (**Supplementary Figure S8**), making great difference with its homolog *VviMYBF1*. Furthermore, none of five copies of *FLS* seems to be correlated with *MYBF1* at expression level, which is inconsistent with previous observation that *MYBF1* is a positive regulator of *FLS* gene (Czemmel et al., 2009). Thus, *MYBF1* in *C. reticulata* might have evolved a novel function, which was preferentially expressed in seeds (**Figure 3J**).

Two copies of *MYBPA1* were observed in the *C. reticulata* transcriptome, *MYBPA1_a* was a bit more conserved than *MYBPA1_b* compared to *VviMYBPA1* (**Supplementary Table S13**). While *MYBPA1_b* showed specific expression in young fruits, *MYBPA1_a* were highly expressed in leaf buds, mature leaves and young fruits. ANTHOCYANIDIN REDUCTASE (ANR) and LEUCOANTHOCYANIDIN REDUCTASE (LAR) are known as two key enzymes in the biosynthesis of PAs. While ANR catalyzes the formation of cis-flavan-3-ol (an oligomeric form of PAs) from anthocyanidin

(Xie et al., 2003), LAR catalyzes the formation of tran-flavan-3-ol (another oligomeric form of PAs) from leucoanthocyanidin (Mauge et al., 2010). Previous report showed that *VviMYBPA1* positively activated the expression of *ANR* and *LAR* to specifically control PAs production (Bogs et al., 2007). In this study, the PA-specific biosynthetic gene *ANR* had two copies (*ANR_a* and *ANR_b*) in the assembly, and both of them had highly similar expression patterns with *MYBPA1_a* across tissues (**Figures 3G,I**, **4C**). Among three copies of *LAR*, only *LAR_a* appeared high similar expression patterns with *MYBPA1_a*, which was also validated by qRT-PCR (**Figures 3E,I**). The expression patterns of *LAR_a*, *ANR_a,* and *ANR_b* correlated with that of *MYBPA1_a* suggest that the regulation of *C. reticulata MYBPA1_a* may be similar to its homologs in grape.

In this study, two copies of *MYBA1* were identified in the *C. reticulata* transcriptome. *MYBA1_a* was more conserved than *MYBA1_b* compared to *VviMYBA1* (**Supplementary Table S13**). *MYBA1_a* was identified as a DEG, which was highly expressed in flower buds, flowers and young fruits but lowly expressed in leaf buds and mature leaves (**Figures 3K**, **5C**). Clustering analysis of gene expression profiles showed that *MYBA1_a* shared the same subcluster with several anthocyanin-specific biosynthetic genes, including *ANTHOCYANIDIN 3-O-GLUCOSYLTRANSFERASE* (*UF3GT_a* and *UF3GT_c*) and *ANTHOCYANIDIN 5-O-GLUCOSYLTRANSFERASE* (*UF5GT_b*). *MYBA1_a* and these anthocyanin-specific biosynthetic genes were highly expressed in both flower buds and flowers (**Figures 3H,K**, **5C**), indicating their important roles of anthocyanin biosynthesis in flower developments.

Carotenoid Biosynthesis Pathway in *C. reticulata*

Similar to flavonoids, carotenoids are a diverse group of colorful plant pigments. They play vital roles in many essential physiological processes such as the photosynthesis, flower coloration, and production of phytohormones (Cazzonelli and Pogson, 2010). In the assembly, we successfully identified 33 unigenes potentially involved in the CrtBS pathway (**Figure 6B** and **Supplementary Table S14**). Different from the FlaBS pathway, most (70%) of these genes in this pathway appeared a single copy. Seven genes had multiple copies, including *GERANYLGERANYL REDUCTASE* (*GGR*), *ZETA-CAROTENE DESATURASE* (*ZDS*), *ZEAXANTHIN EPOXIDASE* (*ZEP*), *NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 3* (*NCED 3*), *ABA-GLUCOSYLTRANSFERASE* (*AOG*), *BETA-CAROTENE ISOMERASE* (*DWARF27*), and *XANTHOXIN DEHYDROGENASE* (*ABA2*) (**Supplementary Table S14**). The expression profiles of these CrtBS-related genes are shown in **Figure 5D**. Most (25/33) of these genes were defined as DEGs, and the majority of these DEGs were up-regulated in mature leaves or flowers (**Figure 5D**). Leaf tissues typically accumulate lutein, beta-carotene, violaxanthin, and neoxanthin for the photosynthesis, antenna assembly, and photoprotection (Cazzonelli and Pogson, 2010). Thus, genes related to the synthesis of these compounds are reasonably up-regulated in mature leafs, such as *GGR_a*, *LUTEIN1* (*LUT1*), *LUT2*, and *LYCOPENE CYCLASE* (*LYC*) (**Figure 5D**). *PHYTOENE SYNTHASE* (*PSY*), which encodes a rate-limiting enzyme that functions in the first committed step of the CrtBS pathway (Cazzonelli and Pogson, 2010), was markedly up-regulated in both flower buds and flowers (**Figure 5D**). Clustering analysis of gene expression profiles showed that *NEOXANTHIN SYNTHASE* (*NXS*) and *NCED3_c* had similar expression patterns with *PSY* (**Figure 5D**). In abscisic acid (ABA) biosynthesis, the NXS enzyme catalyzes the conversion of violaxanthin into neoxanthin (Al-Babili et al., 2000), and the NCED enzyme subsequently catalyzes the cleavage of neoxanthin, which represents the first committed step of ABA biosynthesis (Qin and Zeevaart, 1999). Previously, *ZEP* and *ABA2* in *Arabidopsis* were also identified as ABA biosynthetic genes (Gonzalez-Guzman et al., 2002). **Figure 5D** show that *ABA2_b*, *NCED4*, *NCED3a*, *ZEP_a* were actively expressed in the camellia flowers. Collectively, ABA biosynthesis genes may play important roles in *C. reticulata* flowers for developmental processes.

Carotenoids provide flowers with distinct colors, ranging from yellow to orange or red. In yellow-flowered *Camellia* species, such as *C. chrysantha*, carotenoids were reported to contribute to yellow color of its petal (Hwang et al., 1992; Nishimoto et al., 2004). Red-flowered *Camellia* species, such as *C. reticulata*, are reported to have high amounts of anthocyanins but few carotenoids (Li J. B. et al., 2008). As illustrated in **Figure 5D**, several genes directly related to the formation of pigment compounds (e.g., lutein, zeaxanthin, carotene), such as *LUT1*, *LUT2*, *LUT5*, and *LYC*, were expressed at relatively low levels in flower buds and flowers of *C. reticulata*. Low expression level of these genes in flower may account for a low level of carotenoid-based color compounds in camellia flower. These data also supports that it is the flavonoid biosynthesis pathway rather than carotenoids biosynthesis pathway playing dominant roles in *C. reticulata* flower coloration. Activation of these genes that involved in the synthesis of carotenoid-based color compounds might be useful in genetic manipulations of flower color in *Camellia* species in future.

CONCLUSIONS

We first report the *de novo* assembly of the *C. reticulata* transcriptome. *De novo* assembly has provided a catalog of 141,460 unigenes with a total length of ∼96.1 million nucleotides and an N50 of 1080 nt. Systematic evaluation indicated a good quality of the transcriptome assembly, which is suitable for further studies such as EST-SSRs mining and pathway analysis. We identified the majority of candidate genes related to TAGBS, FlaBS, CrtBS, and photoperiodic flowering pathways. The results also showed that FA desaturase genes were highly regulated, which may be related to cold hardiness response and FA composition control of camellia seed oil. Both the flaBS and CrtBS pathways were extensively regulated in *C. reticulata*. The former mainly participated in the biosynthesis and metabolism of proanthocyanidins, flavonols and anthocyanin pigments for flower color, while the latter play essential role in ABA synthesis rather than color maintenance. These useful resources will further enhance basic and applied researches on this economically important *Camellia* plant.

AUTHOR CONTRIBUTIONS

LG designed the research. QY, HH, YT performed necessary experiments. QY, HH, YT, and EX performed bioinformatics analyses and interpreted the results. QY, HH, and LG wrote the paper. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Wen-Kai Jiang, Yuan Liu, Kui Li, Yun-Long Liu, Li-Ping Zhang, and Bang Liu for providing technical assistance with bioinformatics or experiments. This work was supported by the Project of Innovation Team of Yunnan Province, Top Talents Program of Yunnan Province (20080A009), Hundreds of Oversea Talents Program of Yunnan Province, and National Science Foundation of China (U0936603) to LG and the National Science Foundation of China (31200515) and Surface Project of Natural Science Foundation of Yunnan Province (2012FB179) to HH.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2016.00163

Supplementary Figure S1 | Tissues of wild diploid *C. reticulata* used for RNA-Seq (A–E) and qRT-PCR validation (A–F). (A) leaf buds with length of 30–50 mm collected in March, 2012; **(B)** mature leaves collected in July, 2012; **(C)** flower buds with a diameter of 15–25 mm collected in December, 2012; **(D)** fully opened flowers collected in December, 2012; **(E)** immature fruits with a diameter of 20–30 mm collected in March, 2012; **(F)** blackening seeds collected in July, 2013.

Supplementary Figure S2 | Identity and coverage distribution of the best BLAST hits for each unigene against the *Camellia* EST database. We used megaBLAST ($E \leq 10^{-9}$) to search against the local *Camellia* EST database and extracted the best hit of each sequence for analysis. The green line indicates identity distribution of the best match, while the red line shows the coverage distribution of the best match.

Supplementary Figure S3 | GO Classification of the *C. reticulata* transcriptome. GO terms assigned to each unigenes by BLAST2GO are summarized into three main GO categories (biological process, cellular component, molecular function) and 47 subcategories using the web-based tool WEGO.

Supplementary Figure S4 | Polymorphism survey of SSR markers with 24 individual *C. reticulata* plants. (A) Genomic DNA extracted from 24 individuals using the modified CTAB method. Genomic DNA resolved by electrophoresis on 1% agarose gel and visualized with ethidium bromide staining. Lanes 1–24: 24 individuals; M: 1-kb DNA ladders. **(B–H)** Polymorphism survey results showed that 7 SSR loci were polymorphic. PCR products were resolved by electrophoresis on 10% non-denaturing polyacrylamide gels and visualized by silver staining. Lanes 1–24: 24 individuals; M: 50-bp DNA ladders.

Supplementary Figure S5 | Expression characteristics of the expressed unigenes. (A) Venn diagram shows the presence of the expressed unigenes among five tissue types. **(B)** Cumulative distributions of average expression levels of the expressed unigenes in the five tested tissues. The average cumulative abundance of unigenes across the five tested tissues was calculated by sorting unigenes according to their descending expression levels. The cumulative values of expression levels and unigene number are displayed as a percentage of overall expression level values and total unigenes, respectively.

Supplementary Figure S6 | Major seasonal development (vegetative and reproductive) events of *C. reticulata*. (A) Flowering began after the formation of a new branch (mid-April); **(B)** the floral primordial began to form a new flower bud (late-April); **(C,D)** flower bud growth in June and October respectively; **(E)** the first camellia flower began to bloom in late November; **(F)** a flower would last ∼5 days and then faded away; **(G)** fruiting began when the flower faded away; **(H)** growth of young fruit and leaf bud (mid-March); **(I)** leaf bud breaking the sheaths (late-March); **(J)** continued young fruit growth while the leaf bud began to form a new branch (late-March to mid-April); **(K)** continued young fruit growth (June); **(L)** a fully ripened fruit splitting the capsule from the top (mid-August to late-September).

Supplementary Figure S7 | The relationship between the weather and *C. reticulata* reproductive phenology. The data shown here are based on the daily report from Weather China (http://www.weather.com.cn/).

Supplementary Figure S8 | Comparison of the conserved domains between *C. reticulata MYBF1* and *VviMYBF1*. Conserved domains were detected by BLASTp searches against the NR database on the NCBI web site. Domain hits with an E-value threshold of 0.001 are listed.

Supplementary Table S1 | Details of 24 *C. reticulata* individuals used in this study.

Supplementary Table S2 | The putative genes and corresponding primers used for qRT-PCR analysis.

Supplementary Table S3 | Details of 20 EST-SSRs used in the polymorphism survey of *C. reticulata*.

Supplementary Table S4 | The top 10 assembled unigenes examined for chimerical assembly errors.

Supplementary Table S5 | Sequence homology of the *C. reticulata* unigenes. Top BLAST hits ($E \leq 10^{-3}$) from five databases (Swiss-Prot, NR, TrEMBL, TAIR10, and grape protein database) for all the unigenes are shown.

Supplementary Table S6 | The 10 most abundant PFAM families/domains for *C. reticulata* unigenes.

Supplementary Table S7 | GO annotation, KO assignment and KEGG pathway annotation for *C. reticulata* unigenes. GO terms were obtained based on the NR database annotations using Blast2GO software.

Supplementary Table S8 | List of SSR motifs and their frequencies in the unigenes of *C. reticulata*.

Supplementary Table S9 | List of DE unigenes and their expression values. DE unigenes were assigned to 5 groups (DELB, DEML, DEFB, DEFL, and DEFR) based on their tissues where their highest expression occurred.

Supplementary Table S10 | KEGG pathway and GO enrichment analysis of DE unigenes.

Supplementary Table S11 | Detailed information of putative genes related to TAG biosynthesis in the *C. reticulata* transcriptome.

Supplementary Table S12 | Detailed information of putative genes related to photoperiodic flowering pathway in the *C. reticulata* transcriptome.

Supplementary Table S13 | Detailed information of putative genes related to flavonoid biosynthesis in the *C. reticulata* transcriptome.

Supplementary Table S14 | Detailed information of putative genes related to carotenoid biosynthesis in the *C. reticulata* transcriptome.

REFERENCES

Al-Babili, S., Hugueney, P., Schledz, M., Welsch, R., Frohnmeyer, H., Laule, O., et al. (2000). Identification of a novel gene coding for neoxanthin synthase from *Solanum tuberosum*. *FEBS Lett.* 485, 168–172. doi: 10.1016/S0014-5793(00)02193-1

Amasino, R. M., and Michaels, S. D. (2010). The timing of flowering. *Plant Physiol.* 154, 516–520. doi: 10.1104/pp.110.161653

Andres, F., and Coupland, G. (2012). The genetic basis of flowering responses to seasonal cues. *Nat. Rev. Genet.* 13, 627–639. doi: 10.1038/nrg3291

Baud, S., and Lepiniec, L. (2010). Physiological and developmental regulation of seed oil production. *Prog. Lipid Res.* 49, 235–249. doi: 10.1016/j.plipres.2010.01.001

Birney, E., and Durbin, R. (2000). Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* 10, 547–548. doi: 10.1101/gr.10.4.547

Bogs, J., Jaffe, F. W., Takos, A. M., Walker, A. R., and Robinson, S. P. (2007). The grapevine transcription factor VvMYBPA1 regulates proanthocyanidin synthesis during fruit development. *Plant Physiol.* 143, 1347–1361. doi: 10.1104/pp.106.093203

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic-linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32, 314–331.

Bustin, S. (2002). Quantifcation of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *J. Mol. Endocrinol.* 29, 23–39. doi: 10.1677/jme.0.0290023

Cazzonelli, C. I., and Pogson, B. J. (2010). Source to sink: regulation of carotenoid biosynthesis in plants. *Trends Plant Sci.* 15, 266–274. doi: 10.1016/j.tplants.2010.02.003

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Cunningham, F. X., and Gantt, E. (1998). Genes and enzymes of carotenoid biosynthesis in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 49, 557–583. doi: 10.1146/annurev.arplant.49.1.557

Czemmel, S., Heppel, S. C., and Bogs, J. (2012). R2R3 MYB transcription factors: key regulators of the flavonoid biosynthetic pathway in grapevine. *Protoplasma* 249(Suppl. 2), S109–S118. doi: 10.1007/s00709-012-0380-z

Czemmel, S., Stracke, R., Weisshaar, B., Cordon, N., Harris, N. N., Walker, A. R., et al. (2009). The grapevine R2R3-MYB transcription factor VvMYBF1 regulates flavonol synthesis in developing grape berries. *Plant Physiol.* 151, 1513–1530. doi: 10.1104/pp.109.142059

Dao, T. T. H., Linthorst, H. J. M., and Verpoorte, R. (2011). Chalcone synthase and its functions in plant resistance. *Phytochem. Rev.* 10, 397–412. doi: 10.1007/s11101-011-9211-7

Deluc, L., Barrieu, F., Marchive, C., Lauvergeat, V., Decendit, A., Richard, T., et al. (2006). Characterization of a grapevine R2R3-MYB transcription factor that regulates the phenylpropanoid pathway. *Plant Physiol.* 140, 499–511. doi: 10.1104/pp.105.067231

Deluc, L., Bogs, J., Walker, A. R., Ferrier, T., Decendit, A., Merillon, J. M., et al. (2008). The transcription factor *VvMYB5b* contributes to the regulation of anthocyanin and proanthocyanidin biosynthesis in developing grape berries. *Plant Physiol.* 147, 2041–2053. doi: 10.1104/pp.108.118919

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Fan, Z. Q., Li, J. Y., Li, X. L., Wu, B., Wang, J. Y., Liu, Z. C., et al. (2015). Genome-wide transcriptome profiling provides insights into floral bud development of summer-flowering *Camellia azalea*. *Sci. Rep.* 5:9729. doi: 10.1038/srep09729

Fowler, S., Lee, K., Onouchi, H., Samach, A., Richardson, K., Coupland, G., et al. (1999). *GIGANTEA*: a circadian clock-controlled gene that regulates photoperiodic flowering in *Arabidopsis* and encodes a protein with several possible membrane-spanning domains. *EMBO J.* 18, 4679–4688. doi: 10.1093/emboj/18.17.4679

Ganko, E. W., Meyers, B. C., and Vision, T. J. (2007). Divergence in expression between duplicated genes in *Arabidopsis*. *Mol. Biol. Evol.* 24, 2298–2309. doi: 10.1093/molbev/msm158

Gonzalez-Guzman, M., Apostolova, N., Belles, J. M., Barrero, J. M., Piqueras, P., Ponce, M. R., et al. (2002). The short-chain alcohol dehydrogenase ABA2 catalyzes the conversion of xanthoxin to abscisic aldehyde. *Plant Cell* 14, 1833–1846. doi: 10.1105/tpc.002477

Gonzalez-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.* 14:R70. doi: 10.1186/gb-2013-14-7-r70

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* 38:e131. doi: 10.1093/nar/gkq224

Hellyer, A., Leadlay, P. F., and Slabas, A. R. (1992). Induction, purification and characterization of acyl-ACP thioesterase from developing seeds of oil seed rape (*Brassica napus*). *Plant Mol. Biol.* 20, 763–780. doi: 10.1007/BF00027148

Hirschberg, J. (2001). Carotenoid biosynthesis in flowering plants. *Curr. Opin. Plant Biol.* 4, 210–218. doi: 10.1016/S1369-5266(00)00163-1

Huang, H., Tong, Y., Zhang, Q. J., and Gao, L. Z. (2013). Genome size variation among and within *Camellia* species by using flow cytometric analysis. *PLoS ONE* 8:e64981. doi: 10.1371/journal.pone.0064981

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Hwang, Y. J., Yoshikawa, K., Miyajima, I., and Okubo, H. (1992). Flower colors and pigments in hybrids with *Camellia chrysantha*. *Sci. Hortic.* 51, 251–259. doi: 10.1016/0304-4238(92)90123-T

Imaizumi, T., Schultz, T. F., Harmon, F. G., Ho, L. A., and Kay, S. A. (2005). FKF1 F-box protein mediates cyclic degradation of a repressor of CONSTANS in *Arabidopsis*. *Science* 309, 293–297. doi: 10.1126/science.1110586

Jack, T. (2004). Molecular and genetic mechanisms of floral control. *Plant Cell* 16, S1–S17. doi: 10.1105/tpc.017038

Joubes, J., Raffaele, S., Bourdenx, B., Garcia, C., Laroche-Traineau, J., Moreau, P., et al. (2008). The VLCFA elongase gene family in *Arabidopsis thaliana*: phylogenetic analysis, 3D modelling and expression profiling. *Plant Mol. Biol.* 67, 547–566. doi: 10.1007/s11103-008-9339-z

Kim, J., Jung, J. H., Lee, S. B., Go, Y. S., Kim, H. J., Cahoon, R., et al. (2013). *Arabidopsis 3-ketoacyl-CoA synthase 9* is involved in the synthesis of tetracosanoic acids as precursors of cuticular waxes, suberins, sphingolipids, and phospholipids. *Plant Physiol.* 162, 567–580. doi: 10.1104/pp.112.210450

Kobayashi, S., Goto-Yamamoto, N., and Hirochika, H. (2004). Retrotransposon-induced mutations in grape skin color. *Science* 304, 982. doi: 10.1126/science.1095011

Kodama, H., Hamada, T., Horiguchi, G., Nishimura, M., and Iba, K. (1994). Genetic enhancement of cold tolerance by expression of a gene for chloroplast omega-3 fatty acid desaturase in transgenic tobacco. *Plant Physiol.* 105, 601–605.

Krasileva, K. V., Buffalo, V., Bailey, P., Pearce, S., Ayling, S., Tabbita, F., et al. (2013). Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol.* 14:R66. doi: 10.1186/gb-2013-14-6-r66

Kunst, L., Taylor, D. C., and Underhill, E. W. (1992). Fatty acid elongation in developing seeds of *Arabidopsis thaliana*. *Plant Physiol. Biochem.* 30, 425–434.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26, 493–500. doi: 10.1093/bioinformatics/btp692

Li, J. B., Hashimoto, F., Shimizu, K., and Sakata, Y. (2008). Anthocyanins from red flowers of *Camellia* cultivar 'Dalicha'. *Phytochemistry* 69, 3166–3171. doi: 10.1016/j.phytochem.2008.03.014

Li, J. B., Hashimoto, F., Shimizu, K., and Sakata, Y. (2009). A new acylated anthocyanin from the red flowers of *Camellia hongkongensis* and characterization of anthocyanins in the section *Camellia* species. *J. Integr. Plant Biol.* 51, 545–552. doi: 10.1111/j.1744-7909.2009.00828.x

Li, L., Fu, Q. T., and Yu, D. Q. (2008). An effective protocol for the isolation of RNA from cycad leaves. *Acta Bot. Yunnanica* 30, 593–596.

Liu, L. Q., and Gu, Z. J. (2011). Genomic *in situ* hybridization identifies genome donors of *Camellia reticulata* (Theaceae). *Plant Sci.* 180, 554–559. doi: 10.1016/j.plantsci.2010.12.006

Liu, X. K., and Ma, Y. H. (2010). Camparison of fatty acid between *Camellia reticulata* f. simpex seed and *Camellia oleifera* seed in Yunnan. *J. Kunming Univ.* 32, 56–58.

Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., et al. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res.* 40, W622–W627. doi: 10.1093/nar/gks540

Lu, C. F., Xin, Z. G., Ren, Z. H., Miquel, M., and Browse, J. (2009). An enzyme regulating triacylglycerol composition is encoded by the *ROD1* gene of *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 106, 18837–18842. doi: 10.1073/pnas.0908848106

Lurin, C., Andres, C., Aubourg, S., Bellaoui, M., Bitton, F., Bruyere, C., et al. (2004). Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* 16, 2089–2103. doi: 10.1105/tpc.104.022236

Ma, J. L., Ye, H., Rui, Y. K., Chen, G. C., and Zhang, N. Y. (2011). Fatty acid composition of *Camellia oleifera* oil. *J. Consum. Protect. Food Safety* 6, 9–12. doi: 10.1007/s00003-010-0581-3

Ma, X., and Browse, J. (2006). Altered rates of protein transport in *Arabidopsis* mutants deficient in chloroplast membrane unsaturation. *Phytochemistry* 67, 1629–1636. doi: 10.1016/j.phytochem.2006.04.008

Martin, J. A., and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat. Rev. Genet.* 12, 671–682. doi: 10.1038/nrg3068

Mas, P. (2005). Circadian clock signaling in *Arabidopsis thaliana*: from gene expression to physiology and development. *Int. J. Dev. Biol.* 49, 491–500. doi: 10.1387/ijdb.041968pm

Matteucci, M., D'Angeli, S., Errico, S., Lamanna, R., Perrotta, G., and Altamura, M. M. (2011). Cold affects the transcription of fatty acid desaturases and oil quality in the fruit of *Olea europaea* L. genotypes with different cold hardiness. *J. Exp. Bot.* 62, 3403–3420. doi: 10.1093/jxb/err013

Mauge, C., Granier, T., D'Estaintot, B. L., Gargouri, M., Manigand, C., Schmitter, J. M., et al. (2010). Crystal structure and catalytic mechanism of leucoanthocyanidin reductase from *Vitis vinifera*. *J. Mol. Biol.* 397, 1079–1091. doi: 10.1016/j.jmb.2010.02.002

Ming, T. L., Gu, Z. J., Zhang, W. J., and Xie, L. S. (2000). *Monograph of the Genus Camellia*. Kunming: Yunnan Science and Technology Press.

Moreno-Perez, A. J., Venegas-Caleron, M., Vaistij, F. E., Salas, J. J., Larson, T. R., Garces, R., et al. (2012). Reduced expression of *FATA* thioesterases in *Arabidopsis* affects the oil content and fatty acid composition of the seeds. *Planta* 235, 629–639. doi: 10.1007/s00425-011-1534-5

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35, W182–W185. doi: 10.1093/nar/gkm321

Nesi, N., Debeaujon, I., Jond, C., Pelletier, G., Caboche, M., and Lepiniec, L. (2000). The *TT8* gene encodes a basic helix-loop-helix domain protein required for expression of *DFR* and *BAN* genes in *Arabidopsis* siliques. *Plant Cell* 12, 1863–1878. doi: 10.1105/tpc.12.10.1863

Nguyen, H. T., Silva, J. E., Podicheti, R., Macrander, J., Yang, W. Y., Nazarenus, T. J., et al. (2013). *Camelina* seed transcriptome: a tool for meal and oil improvement and translational research. *Plant Biotechnol. J.* 11, 759–769. doi: 10.1111/pbi.12068

Nicot, N., Hausman, J. F., Hoffmann, L., and Evers, D. (2005). Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. *J. Exp. Bot.* 56, 2907–2914. doi: 10.1093/jxb/eri285

Nishimoto, S., Hashimoto, F., Shimizu, K., and Sakata, Y. (2004). Petal coloration of interspecific hybrids between *Camellia chrysantha* × *C. japonica*. *J. Jpn. Soc. Horticul. Sci.* 73, 189–191. doi: 10.2503/jjshs.73.189

Okuley, J., Lightner, J., Feldmann, K., Yadav, N., Lark, E., and Browse, J. (1994). *Arabidopsis FAD2* gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. *Plant Cell* 6, 147–158. doi: 10.1105/tpc.6.1.147

Phillips, M. A., Leon, P., Boronat, A., and Rodriguez-Concepcion, M. (2008). The plastidial MEP pathway: unified nomenclature and resources. *Trends Plant Sci.* 13, 619–623. doi: 10.1016/j.tplants.2008.09.003

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* 40, D290–D301. doi: 10.1093/nar/gkr1065

Putterill, J., Robson, F., Lee, K., Simon, R., and Coupland, G. (1995). The *CONSTANS* gene of *Arabidopsis* promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. *Cell* 80, 847–857. doi: 10.1016/0092-8674(95)90288-0

Qin, X. Q., and Zeevaart, J. A. D. (1999). The 9-cis-epoxycarotenoid cleavage reaction is the key regulatory step of abscisic acid biosynthesis in water-stressed bean. *Proc. Natl. Acad. Sci. U.S.A.* 96, 15354–15361. doi: 10.1073/pnas.96.26.15354

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Samach, A., Onouchi, H., Gold, S. E., Ditta, G. S., Schwarz-Sommer, Z., Yanofsky, M. F., et al. (2000). Distinct roles of CONSTANS target genes in reproductive development of *Arabidopsis*. *Science* 288, 1613–1616. doi: 10.1126/science.288.5471.1613

Schmieder, R., and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* 6:e17288. doi: 10.1371/journal.pone.0017288

Shanklin, J., and Somerville, C. (1991). Stearoyl-acyl-carrier-protein desaturase from higher-plants is structurally unrelated to the animal and fungal homologs. *Proc. Natl. Acad. Sci. U.S.A.* 88, 2510–2514. doi: 10.1073/pnas.88.6.2510

Shi, C. Y., Yang, H., Wei, C. L., Yu, O., Zhang, Z. Z., Jiang, C. J., et al. (2011). Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics* 12:131. doi: 10.1186/1471-2164-12-131

Simopoulos, A. P., and Robinson, J. (1999). *The Omega Diet: The Lifesaving Nutritional Program Based on the Diet of the Island of Crete*. New York, NY: HarperCollins.

Slabas, A. R., Chase, D., Nishida, I., Murata, N., Sidebottom, C., Safford, R., et al. (1992). Molecular cloning of higher-plant 3-oxoacyl-(acyl carrier protein) reductase. Sequence identities with the nodG-gene product of the nitrogen-fixing soil bacterium *Rhizobium meliloti*. *Biochem. J.* 283(Pt 2), 321–326. doi: 10.1042/bj2830321

Tan, L. Q., Wang, L. Y., Wei, K., Zhang, C. C., Wu, L. Y., Qi, G. N., et al. (2013). Floral transcriptome sequencing for SSR marker development and linkage map construction in the tea plant (*Camellia sinensis*). *PLoS ONE* 8:e81611. doi: 10.1371/journal.pone.0081611

Thiel, T., Michalek, W., Varshney, R. K., and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley *(Hordeum vulgare* L.). *Theor. Appl. Genet.* 106, 411–422. doi: 10.1007/s00122-002-1031-0

Tong, Y., Wu, C. Y., and Gao, L. Z. (2013). Characterization of chloroplast microsatellite loci from whole chloroplast genome of *Camellia taliensis* and their utilization for evaluating genetic diversity of *Camellia reticulata* (Theaceae). *Biochem. Syst. Ecol.* 50, 207–211. doi: 10.1016/j.bse.2013.04.003

Valverde, F., Mouradov, A., Soppe, W., Ravenscroft, D., Samach, A., and Coupland, G. (2004). Photoreceptor regulation of CONSTANS protein in photoperiodic flowering. *Science* 303, 1003–1006. doi: 10.1126/science.1091761

Van Belleghem, S. M., Roelofs, D., Van Houdt, J., and Hendrickx, F. (2012). *De novo* transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle *Pogonus chalceus* (Coleoptera, Carabidae). *PLoS ONE* 7:e42605. doi: 10.1371/journal.pone.0042605

Vogel, C., and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* 13, 227–232. doi: 10.1038/nrg3185

Walker, A. R., Lee, E., Bogs, J., McDavid, D. A. J., Thomas, M. R., and Robinson, S. P. (2007). White grapes arose through the mutation of two similar and adjacent regulatory genes. *Plant J.* 49, 772–785. doi: 10.1111/j.1365-313X.2006.02997.x

Wang, X. C., Zhao, Q. Y., Ma, C. L., Zhang, Z. H., Cao, H. L., Kong, Y. M., et al. (2013). Global transcriptome profiles of *Camellia sinensis* during cold acclimation. *BMC Genomics* 14:415. doi: 10.1186/1471-2164-14-415

Wang, Z. W., Jiang, C., Wen, Q., Wang, N., Tao, Y. Y., and Xu, L. A. (2014). Deep sequencing of the *Camellia chekiangoleosa* transcriptome revealed candidate genes for anthocyanin biosynthesis. *Gene* 538, 1–7. doi: 10.1016/j.gene.2014.01.035

Wigge, P. A., Kim, M. C., Jaeger, K. E., Busch, W., Schmid, M., Lohmann, J. U., et al. (2005). Integration of spatial and temporal information during floral induction in *Arabidopsis*. *Science* 309, 1056–1059. doi: 10.1126/science.1114358

Winkel-Shirley, B. (2001). Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol.* 126, 485–493. doi: 10.1104/pp.126.2.485

Wu, H. L., Chen, D., Li, J. X., Yu, B., Qiao, X. Y., Huang, H. L., et al. (2013). *De novo* characterization of leaf transcriptome using 454 sequencing and development of EST-SSR markers in tea (*Camellia sinensis*). *Plant Mol. Biol. Rep.* 31, 524–538. doi: 10.1007/s11105-012-0519-2

Xia, E. H., Jiang, J. J., Huang, H., Zhang, L. P., Zhang, H. B., and Gao, L. Z. (2014). Transcriptome analysis of the oil-rich tea plant, *Camellia oleifera*, reveals candidate genes related to lipid metabolism. *PLoS ONE* 9:e104150. doi: 10.1371/journal.pone.0104150

Xia, L. F., Gu, Z. J., Wang, Z. L., Xiao, T. J., Wang, L., and Kondo, K. (1994). Dawn on the origin of *Camellia reticulata*–the new discovery of its wild diploid in Jinshajiang Valley. *Acta Bot. Yunnannica* 16, 255–262.

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., et al. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 39, W316–W322. doi: 10.1093/nar/gkr483

Xie, D. Y., Sharma, S. B., Paiva, N. L., Ferreira, D., and Dixon, R. A. (2003). Role of anthocyanidin reductase, encoded by *BANYULS* in plant flavonoid biosynthesis. *Science* 299, 396–399. doi: 10.1126/science.1078540

Xu, R., Wang, R., and Liu, A. (2011). Expression profiles of genes involved in fatty acid and triacylglycerol synthesis in developing seeds of Jatropha (*Jatropha curcas* L.). *Biomass Bioen.* 35, 1683–1692. doi: 10.1016/j.biombioe.2011.01.001

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., et al. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* 34, W293–W297. doi: 10.1093/nar/gkl031

Yeh, F. C., Yang, R. C., and Boyle, T. (1999). *Popgene Version 1. 31 Quick User Guide*. Alberta: University of Alberta and Centre for International Forestry Research.

Yu, T. T., and Bruce, B. (1980). The origin and classification of the garden varieties of *Camellia reticulata*. *Am. Camellia Yearb.* 1980, 1–29.

Zhang, H. B., Xia, E. H., Huang, H., Jiang, J. J., Liu, B. Y., and Gao, L. Z. (2015). *De novo* transcriptome assembly of the wild relative of tea tree (*Camellia taliensis*) and comparative analysis with tea transcriptome identified putative genes associated with tea quality and stress response. *BMC Genomics* 16:4. doi: 10.1186/s12864-015-1494-4

Zhao, Q. Y., Wang, Y., Kong, Y. M., Luo, D., Li, X., and Hao, P. (2011). Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* 12(Suppl. 14):S2. doi: 10.1186/1471-2105-12-S14-S2

Zheng, W., Chung, L. M., and Zhao, H. (2011). Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics* 12:290. doi: 10.1186/1471-2105-12-290

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.