Global Transcriptome Analysis and Identification of the Flowering Regulatory Genes Expressed in Leaves of *Lagerstroemia indica*

Zhenyu Zhang,* Peng Wang,* Ya Li, Lingling Ma, Linfang Li, Rutong Yang, Yuzhu Ma, Shu'an Wang, and Qing Wang

Flowering time is an important trait for ornamental plants, and flowering regulation has thus been both a focus of and challenge to researchers. *Lagerstroemia indica* is an important summer flowering tree in China and has been introduced abroad as a key parent of new cultivars; no previous reports have addressed the regulation of flowering time in this species. In this study, 28,567,778 × 2 reads were obtained from leaves of *L. indica*. A total of 37,325 unigenes were assembled with an average length of 849.56 bp, and 17,506 (46.90%) unigenes were significantly matched to known genes in the nr database of GenBank. The annotated sequences were clustered into putative functional categories using the Gene Ontology framework. Potential genes and their functions were predicted by the Cluster of Orthologous Groups analysis and Kyoto Encyclopedia of Genes and Genomes pathway mapping. A total of 115 unigenes related to flowering time control were discovered. Ten homologous genes of the *CONSTANS-like* (*COL*) gene family were identified based on transcript data. Phylogenetic analysis of the *CONSTANS* and *COL* genes from *L. indica* and other species grouped them into three clades. The transcriptome dataset and outcome of the analysis provide a valuable new resource for research on the functional genomics and molecular mechanisms of flowering control in *L. indica*.

Introduction

ALL HIGHER PLANTS ARE unable to initiate reproductive development immediately after germination and must undergo a vegetative period (Duan and Guo, 2004). The vegetative periods of most woody plants range from 4 to 20 years and are dependent mainly on species characteristics (Weigel and Nilsson, 1995; Pillitter *et al.*, 2003). However, the vegetative period of *Lagerstroemia indica* (crape myrtle) is only 1 or 2 years, significantly shorter than those of most other species. *L. indica* originated in China and historical records of its cultivation in this country date back to 1700 years, well before its cultivation in other areas of the world (Liu *et al.*, 2008). Known for its long-lasting flowering in midsummer, wide range of flower colors (white, red, purple, and their combined variants), extensive diversity of growth habits (miniature shrub to large tree), and appealing exfoliating bark (cinnamon colored to gray colored), *L. indica* is a key parent for crape myrtle breeding (Cabrera, 2002). Previous research on *L. indica* has focused on surveying its genetic resources and its genetic diversity in China over the past several decades (Pei and Rong, 1983; Min *et al.*, 2008). To date, no research has been conducted on the regulation of flowering time in this species.

Flowering is a crucial developmental transition from the vegetative to the reproductive phase and is properly timed by a number of intrinsic and environmental cues. The regulation of flowering time has been studied in *Arabidopsis thaliana*, *Oryza sativa*, *Vitis vinifera*, and *Populus trichocarpa* (Yuceer *et al.*, 2002; Griffiths *et al.*, 2003; Boss *et al.*, 2004; Almada *et al.*, 2009). Four regulatory pathways, including the photoperiod, vernalization, autonomous, and gibberellin (GA) pathways, have been found in *A. thaliana* (Boss *et al.*, 2004). In the photoperiod pathway, *CONSTANS* (*CO*) and *GIGANTEA* (*GI*), which receive the light and circadian clock signals that activate *FLOWERING LOCUS T* and other genes, play very important roles (Ding *et al.*, 2009). In the vernalization pathway, *FLOWERING LOCUS C* (*FLC*) plays a central role in the regulation of flowering time (Boss *et al.*, 2004). Three key genes with similar sequences and functions, *GIBBERELLIC ACID INSENSITIVE* (*GAI*), *REPRESSOR OF GAI-3* (*RGA*), and *RGA-like 1*, are involved in the GA pathway (Cao *et al.*, 2006). The autonomous pathway is not influenced by the environment, and the genes in this pathway have no cascade amplification effect with each other (Lee *et al.*, 1994; Chou and Yang, 1998; Schomburg *et al.*, 2001; Quesada *et al.*, 2003;

Institute of Botany, Jiangsu Province & Chinese Academy of Sciences, Nanjing, China.

*These two authors contributed equally to this work.

Simpson *et al.*, 2003; Ausín *et al.*, 2004; Lim *et al.*, 2004; Schmitz *et al.*, 2007).

The *CO* gene, which has multiple copies in most higher plants, was first cloned from *A. thaliana* using map-based cloning in 1995 (Putterill *et al.*, 1995). Taken together, the *CO* and *CONSTANS-like* (*COL*) genes form the *COL* gene family (Griffiths *et al.*, 2003), which has 17 and 16 homologous genes in *A. thaliana* and *O. sativa*, respectively (Griffiths *et al.*, 2003). The amino termini of the proteins encoded by the *COL* genes possess one or two B-box domains and one CCT domain, which function in protein–protein interactions (Putterill *et al.*, 1995). Based on the variation in B-box domains, the *COL* gene family in *A. thaliana* may be separated into three groups (Robson *et al.*, 2001). Wu *et al.* (2014) identified 26 *CO* homologs (*GmCOLs*) in the soybean genome and classified them into three clades, which are conserved among flowering plants. Chou *et al.* (2013) identified 12 full-length *E. pusilla* *CO*-like (*EpCOL*) genes and divided them into four groups.

In recent years, next-generation sequencing (NGS) technologies have emerged as powerful tools for high-throughput sequence determination and have increased the rate and efficiency of gene discovery (Clark *et al.*, 2013; Wang *et al.*, 2013). Compared with the sequencing of expressed sequence tags, NGS techniques possess higher throughput, lower cost, and more rapid RNA sequencing. Zhang *et al.* (2012) identified numerous putative sequences related to flowering time control and flower development in *Dendrocalamus latiflorus* by comparing its floral transcriptome to those of known plants, and 290 unigenes were discovered that were homologous to known flowering-related genes from other species. Zhang *et al.* (2011) conducted a transcriptome profile analysis of the flowering molecular processes of a *Poncirus trifoliata* mutant and its wild-type counterpart using massively parallel signature sequencing, thereby identifying a large number of genes involved in early flowering development. These studies and others have shown that transcriptome sequencing provides a rapid and efficient method for identifying genes related to flowering time control.

L. indica is one of the most well-known ornamental plants worldwide. Flowering time is an important trait for ornamental plants, and no previous studies have examined the regulation of flowering time in *L. indica*. The genes that regulate flowering time are expressed in the leaves of the plant. Thus, the goals of this study were to characterize the transcriptome of the leaves of *L. indica*, identify the genes related to flowering time control, and highlight the *COL* gene and determine its evolutionary characteristics. The transcriptome dataset and outcome of the analysis provide a valuable new resource for research on the functional genomics of and molecular mechanisms of flowering control in *L. indica*.

Materials and Methods

Plant material

The *L. indica* plants used in this study were collected from the Institute of Botany, Jiangsu Province & Chinese Academy of Sciences, Nanjing, China. The plants were cultured under the appropriate conditions.

RNA extraction and library preparation for transcriptome analysis

Leaves were collected from *L. indica* plants on May 12, June 14, and July 18, 2012. All samples were frozen immediately in liquid nitrogen and stored at -80°C until use. The total RNA from the mutant leaves was extracted using the Trizol reagent according to the manual instructions (Invitrogen). The total RNA concentration was quantified using UV spectrophotometry, and the quality was checked using electrophoresis in a 1% agarose gel. Equal volumes of RNA from each of the three samples were pooled. Paired-end libraries with approximate average insert lengths of 200 bp were synthesized using the Genomic Sample Prep Kit (Illumina) according to the manufacturer's instructions. Before cluster generation, the library concentration and size were assayed using the Agilent DNA1000 kit (Agilent) on a 2100 Bioanalyzer (Agilent). Libraries were sequenced as 100-mer62 on a Hi-Seq 2000 equipped with a paired-end module at the Shanghai Hanyu Bio-tech Co. Ltd.

Illumina reads processing and assembly

A Perl script was written to remove low-quality sequences (reads with a base quality of less than 20). The high-quality reads were then assembled with Trinity (Grabherr *et al.*, 2011), using default settings except for the K-mer value, to construct unique consensus sequences.

Gene annotation

To find protein coding sequences from different contigs, gene prediction was conducted using the GetORF function of EMBOSS (Rice *et al.*, 2000). The predicted proteins from the contigs were matched to the nonredundant protein sequences (nr) database of GenBank using blastp with an *E*-value threshold of 10^{-5} . A gene name was assigned to each protein sequence based on its best BLAST hit.

The GoPipe program (Chen *et al.*, 2005) was then used to perform gene ontology (GO) functional classification for these predicted proteins. To evaluate the completeness of our transcriptome library and the effectiveness of our annotation process, we searched the annotated unigene sequences for the possible functions involved in COG classification (www.ncbi.nlm.nih.gov/COG/). The predicted proteins of the contigs were matched to the KEGG database to obtain KO numbers using blastx with an *E*-value threshold of 10^{-3} (Kanehisa *et al.*, 2010).

Phylogenetic analysis and domain identification

A BLAST search was conducted against the *L. indica* transcriptome using the full-length amino acid sequences of the *CO* and *COL* homologs from *A. thaliana* and *V. vinifera* obtained from the NCBI (www.ncbi.nlm.nih.gov/) database. Amino acid sequences with B-box and CCT domains were chosen and used to BLAST search for additional *CO* and *COL* homologs in the *L. indica* transcriptome. The BLAST search continued until no further new homologs appeared. A total of 10 *L. indica* *CO* homologs (*LiCOLs*) were obtained. The full-length protein sequences of the *CO* homologs from *L. indica*, *A. thaliana*, and *V. vinifera* were aligned using ClustalW in MEGA5.1 (www.megasoftware.net/) with the default parameters. The phylogenetic tree was generated in

TABLE 1. STATISTICS OF cDNA SEQUENCES OF *LAGERSTROEMIA INDICA* GENERATED BY THE ILLUMINA HI-SEQ 2000 SEQUENCER

Dataset name	Lagerstroemia indica
Number of reads	28,567,778 × 2
High-quality reads	27,395,509 × 2 (95.8%)
Average read length (bp)	100
Number of contigs	57,593
Average length of contigs (bp)	914.06
Number of unigenes	37,325
Average length of unigenes (bp)	849.56
Total length of unigenes (bp)	31,709,704

MEGA5.1 using the neighbor-joining method with 1000 bootstrap replicates (Saitou and Nei, 1987; Tamura *et al.*, 2011). The MEME program was used to predict conserved motifs with the following parameters: any number of repetitions, maximum of 15 motifs, and optimum motif width ≥ 22 and ≤ 100 (Bailey *et al.*, 2009).

Results

Transcriptome sequencing and assembly

Using an Illumina Hi-Seq 2000 sequencer, 28,567,778 × 2 reads were obtained, with an average read length of 100 bp (Table 1). After the removal of low-quality sequences from the raw data, 27,395,509 × 2 high-quality reads (95.8%) were obtained (Table 1). The clean sequence data were deposited in the Short Read Archive database of the NCBI under the accession number SRP (034678). Based on the high-quality reads, a total of 57,593 contigs were assembled, with an average length of 914.06 bp. All contigs then were merged into 37,325 unigenes, with an average length of 849.56 bp, using the Trinity method (Table 1). The lengths of these unigenes ranged from 201 to 27,285 bp (Fig. 1).

Functional annotation

A total of 17,506 (46.90%) unigenes were significantly matched to known genes in the GenBank nr database. For unigenes longer than 2000 bp, 3458 (94.51%) of 3659 uni-

FIG. 1. Length distribution of unigenes in the leaves of *Lagerstroemia indica*.

FIG. 2. Matching percentage of *L. indica* unigenes with different lengths to entries in the GenBank nr databases. *E*-value cutoff of $1.0E-5$.

genes were significantly matched to known genes in the nr database (Fig. 2). The match efficiencies were 19.09% and 55.57% for unigenes in the 201–500 bp and 501–1001 bp ranges, respectively (Fig. 2). The *E*-value distribution of the top hits in the databases showed that 65.41% of the matched

FIG. 3. *E*-value distribution of BLAST hits for matched *L. indica* unigene sequences against nr databases. *E*-value cutoff of $1.0E-5$.

sequences possessed strong homology ($<1.0e^{-50}$) and that 34.59% of the matched sequences possessed moderate homology (between $1.0e^{-5}$ and $1.0e^{-50}$) (Fig. 3).

GO annotation and COG annotation

A total of 11,915 unigenes could be assigned to GO classes with 91,094 GO terms based on their similarity to sequences with previously known functions. A total of 91,094 GO terms were retrieved and 33.80%, 33.20%, and 32.99% of the terms were assigned to the molecular function, cellular component, and biological process categories, respectively (Fig. 4).

Using the Cluster of COG database, 6703 unigenes were classified into 25 COG categories (Fig. 5), of which “Signal transduction mechanisms” was the most common (1750, 14.06%), followed by “General function prediction only” (1342, 10.78%) and “Posttranslational modification, protein turnover, chaperones” (1126, 9.04%) (Fig. 5). “Nuclear structure” (42, 0.34%), “Extracellular structures” (31, 0.25%), and “Cell motility” (3, 0.02%) were the three least prevalent COG categories (Fig. 5).

KEGG pathway mapping

To identify the biological pathways in *L. indica*, we mapped the annotated sequences to the reference canonical pathways contained in the KEGG database. Out of 37,325 unigenes, 16,993 (45.53%) were annotated and 5593

(32.91%) significant matches were assigned to 312 KEGG pathways. The KEGG pathways were classified into six categories: human diseases (1582), organismal systems (1075), cellular processes (762), environmental information processing (622), genetic information processing (1570), and metabolism (3384) (Fig. 6). The most prevalent category was metabolism (3384, 60.50%), which included carbohydrate metabolism (809), energy metabolism (379), amino acid metabolism (455), lipid metabolism (362), nucleotide metabolism (279), metabolism of co-factors and vitamins (256), and other subcategories (Fig. 6). The least common category was environmental information processing (622, 11.12%), which included signal transduction (580), membrane transport (39), and signaling molecules and interaction (3) (Fig. 6).

Identification of putative genes related to flowering time

As shown in Supplementary Table S1 (Supplementary Data are available online at www.liebertpub.com/dna), 115 unigenes showed homology to known flowering time genes. Of these unigenes, 75 were divided into four regulatory pathways of flowering time: the photoperiod, vernalization, autonomous, and GA pathways.

A total of 34 unigenes were identified as related to the photoperiod pathway (Supplementary Table S1). The *CO* and *COL* genes, the most important genes in the photoperiod pathway, had 10 homologous unigenes in the leaves of

FIG. 4. Gene ontology (GO) terms for the transcriptomic sequences of *L. indica*.

FIG. 5. COG classifications of unigenes in *L. indica*. Color images available online at www.liebertpub.com/dna

L. indica. A *GI* unigene was identified among the regulators of *CO* transcription and protein stability. Five unigenes encoding putative photoreceptor apoproteins, including *PHYTOCHROME A (PHYA)*, *PHYB*, *CRYPTOCHROME 1 (CRY1)*, and *CRY2*, were also identified. The central circadian clock genes, homologs of *LATE ELONGATED HYPOCOTYL*, *EARLY FLOWERING 4*, and *CIRCADIAN CLOCK ASSOCIATED 1*, were present in the leaves of *L. indica*. Furthermore, the unigenes of *ZTL*, *SPA*, *COP1*, *PSEUDO-RESPONSE REGULATOR 1 (PRR1)*, *PRR5*, and *PRR7* were all expressed in the leaves of *L. indica*.

A total of 17 unigenes were related to the vernalization pathway (Supplementary Table S1). The *FLC* gene modulates downstream signaling, and all other genes in the vernalization pathway affect the flowering period by altering the expression of *FLC*. A homologous gene of *FRIGIDA*, the gene that regulates the expression of *FLC*, was also identified. Homologous genes of *VERNALIZATIONINDEPENDENT 3*, *FERTILIZATION INDEPENDENT ENDOSPERM*, *MULTICOPY SUPPRESSOR OF IRA1*, *SILENT INFORMATION REGULATOR 2*, and *CURLY LEAF*, all of which are putative repressors of *FLC*, were present in the leaves of *L. indica*.

For the autonomous pathway, 12 unigenes were discovered (Supplementary Table S1). Six unigenes showing homology to *LUMINDEPENDENS*, *FY*, *FPA*, *FLOWERING LOCUS D*, and *FCA* were discovered. Additionally, six homologous unigenes of *DICER-like (DCL)*, a gene related

to the autonomous pathway, including *DCL1*, *DCL2*, and *DCL3A*, were discovered in this study.

Twelve putative homologous unigenes of GA-signaling pathway genes were also found (Supplementary Table S1). Homologs of the three key genes in this pathway, *GAI*, *RGA-like*, and *RGA*, were discovered in this study. A homologous unigene of *SPY*, located upstream of *GAI/RGA*, was identified in *L. indica*, as was one for *GAMYB*, a newly discovered gene of the GA pathway.

Besides the genes of the four known pathways, an additional 40 unigenes related to flowering time control were discovered (Supplementary Table S1). The homologous unigenes for *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* and *LEAFY*, which are the ultimate receptors of the four pathways, were identified. The homologous unigenes of the MADS gene family, *APETALA*, *AGAMOUS-like*, *DEFICIENS*, and *SHORT VEGETATIVE PHASE*, were all found in *L. indica*. Furthermore, the homologous unigenes of *SQUAMOSA PROMOTER BINDING PROTEIN-like*, *HEME ACTIVATOR PROTEIN*, *FD*, and *EMBRYONIC FLOWER 2* were also expressed in this species.

Identification and phylogenetic analysis of CO homologs

Ten *CO* and *COL* homologs were identified (*LiCOLs*) through BLAST searching with the full-length amino acid

FIG. 6. KEGG metabolism pathway categories assigned with *L. indica* unigenes.

sequences of the *CO* and *COL* homologs from *A. thaliana* and *V. vinifera*. Phylogenetic analysis clarified the evolutionary relationships of these *LiCOLs* (Fig. 7). Each *LiCOL* was then assigned a name based on its level of homology to the *A. thaliana CO* and *COLs*. All nucleotide sequences of the *LiCOLs* are available in GenBank (KF986327-KF986333, KJ717945-KJ717947). The 10 *LiCOLs* were classified into three clades: four genes in Clade I, two genes in Clade II, and four genes in Clade III (Fig. 7). Clade I contained two subclades that nested two genes, as did Clade III (Fig. 7).

By analyzing the motifs of the *COL* proteins, a total of 15 conserved motifs were identified (Fig. 7 and Supplementary Table S2). Motif 1 was represented by the typical CCT domain of 43 amino acids (aa) and was present in all *CO* and *COL* genes. Motif 2 was represented by the typical B-box domain of 37 aa, and motif 3 was represented by the divergent B-box domain of 32 aa (Fig. 7 and Supplementary Table S2). Excepting the B-box domain and CCT domain,

Clades I to III contained three to five different conserved motifs. Additionally, the subclades of Clades I and III contained different conserved motifs. For example, one subclade of Clade III contained motifs 6 and 12, but the other subclade only contained motif 10.

A reciprocal BLAST search then identified potential orthologous relationships among the *CO* homologs of *A. thaliana*, *L. indica* and *V. vinifera* (Fig. 8). Supporting the conservation of the observed clades among species in the phylogenetic analysis, the BLAST best hits for all of the *A. thaliana CO* homologs were identified as *L. indica* homologs within the same clade and vice versa, excepting *AtCOL15*, for which the best hit was *LiCOL8*.

Discussion

Previous research has indicated that the same pathways of flowering control exist in different plants (Sun *et al.*, 2007).

FIG. 7. Phylogenetic analysis and conserved motif analysis of *COL* family protein from *L. indica*, *Arabidopsis thaliana*, and *Vitis vinifera*. A total of 10 *COL* family proteins in *L. indica* (KF986327-KF986333, KJ717945-KJ717947), 17 in *A. thaliana* (Q39057.1, O50055.1, Q96502.1, Q9SK53.1, Q940T9.2, Q9FHH8.2, Q8LG76.2, Q9C9A9.1, Q9SSE5.1, Q9LUA9.1, O23379.2, Q9LJ44.2, O82256.1, O22800.2, Q9C7E8.1, Q8RWD0.2), and 9 in *V. vinifera* (XP_002282509.1, XP_002263458.1, XP_002277953.1, XP_002264506.2, XP_002265377.2, XP_002274384.2, XP_002268490.1, XP_002275375.2, XP_002276181.1) were used to construct the NJ tree. Numbers at nodes indicate the value of 1000 bootstrap analyses. Each motif is represented by a number in a colored box. Specific lengths, locations, and *p*-values of each motif can be found in Supplementary Table S2. Color images available online at www.liebertpub.com/dna

FIG. 8. BLAST best hits of *CO* homologs in *L. indica*, *A. thaliana*, and *V. vinifera*. The BLAST best hits are shown by dotted lines with an arrowhead, and the protein pairs of the reciprocal best hits are shown by lines with double-head arrows.

By comparing the unigenes in the leaves of *L. indica* with the NCBI database, at least 115 unigenes were discovered that showed homology to known flowering related genes from other plants (Supplementary Table S1). The number of unigenes related to flowering time control in *L. indica* was slightly greater than the number in *Poncirus trifoliata* (Zhang *et al.*, 2011), but was significantly lower than that in *D. latiflorus* (Zhang *et al.*, 2012), which included a large amount of floral meristem identity genes. In this study, homologous genes in the four pathways related to flowering time control were more prevalent than in *P. trifoliata* and *D. latiflorus* (Zhang *et al.*, 2011, 2012). This result showed that the leaf transcriptome may be superior to the floral and floral bud transcriptomes on flowering time control.

The work of previous researchers has shown that *CO* and *COL* genes possess one or two B-box domains at the N terminus and one CCT domain at the C terminus (Robson *et al.*, 2001; Griffiths *et al.*, 2003; Almada *et al.*, 2009). In this study, all *LiCOL* genes possessed one or two B-box domains and one CCT domain. The present number of *COL* homologous genes identified in *L. indica* was less than the 17 genes found in *A. thaliana*, 16 in *O. sativa*, and 14 in *V. vinifera*, the genomes of which have been completely sequenced (Robson *et al.*, 2001; Griffiths *et al.*, 2003; Almada *et al.*, 2009). A lack of genome sequence information may have led to the detection of fewer *COL* homologous genes in *L. indica*.

All genes of the *COL* family are divided into three classes, designated as clades I to III, based on their B-box domains (Robson *et al.*, 2001; Griffiths *et al.*, 2003). In this study, the *CO* and *COL* genes were divided into three classes in an NJ phylogenetic tree using their full-length

protein sequences. The results of the conserved motif analysis and reciprocal BLAST search supported the conservation of the clades among species in the phylogenetic analysis.

Acknowledgments

The authors thank Jian-liang Wang for management of cultivated land, agricultural independent innovation funds (CX[11]1039), public service platform of science and technology of Jiangsu province (BM2012058), natural science funds of Jiang Su province (BK2012377), and the modern agriculture project of Nanjing (201201021).

Disclosure Statement

The authors declare that there are no conflicts of interest regarding the publication of this article.

References

- Almada, R., Cabrera, N., Casaretto, J.A., Ruiz-Lara, S., and Villanueva, E.G. (2009). *VvCO* and *VvCOL1*, two *CONSTANS* homologous genes, are regulated during flower induction and dormancy in grapevine buds. *Plant Cell Rep* **28**, 1193–1203.
- Ausín, I., Alonso-Blanco, C., Jarillo, J.A., Ruiz-García, L., and Martínez-Zapater, J.M. (2004). Regulation of flowering time by *FVE*, a retinoblastoma-associated protein. *Nat Genet* **36**, 162–166.
- Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J.Y., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202–W208.
- Boss, P.K., Bastow, R.M., Mylne, J.S., and Dean, C. (2004). Multiple pathways in the decision to flower, enabling, promoting, and resetting. *Plant Cell* **16**, S18–S31.
- Cabrera, R.I. (2002). Evaluating and promoting the cosmopolitan and multipurpose *Lagerstroemia*. In XXVI International Horticultural Congress, Toronto, Canada, 177–184.
- Cao, D., Cheng, H., Wu, W., Soo, H.M., and Peng, J. (2006). Gibberellin mobilizes distinct DELLA-dependent transcriptomes to regulate seed germination and floral development in *Arabidopsis*. *Plant Physiol* **142**, 509–525.
- Chen, Z.Z., Xue, C.H., Zhu, S., Zhou, F.F., Ling, X.B., et al. (2005). GoPipe, streamlined gene ontology annotation for batch anonymous sequences with statistics. *Prog Biochem Biophys* **32**, 187–191.
- Chou, M.L., and Yang, C.H. (1998). *FLD* interacts with genes that affect different developmental phase transitions to regulate *Arabidopsis* shoot development. *Plant J* **15**, 231–242.
- Chou, M.L., Shih, M.C., Chan, M.T., Liao, S.Y., Hsu, C.T., Huang, Y.T., Chen, J.J., Liao, D.C., Wu, F.H., and Lin, C.S. (2013). Global transcriptome analysis and identification of a *CONSTANS*-like gene family in the orchid *Erycina pusilla*. *Planta* **237**, 1425–1441.
- Clark, S.M., Vaitheeswaran, V., Ambrose, S.J., Purves, R.W., and Page, J.E. (2013). Transcriptome analysis of bitter acid biosynthesis and precursor pathways in hop (*Humulus lupulus*). *BMC Plant Biol* **13**, 12.
- Ding, Y., Dai, S.L., and Ma, Y.P. (2009). Photoperiodic control of flowering in higher plant. *North Horticulture* **9**, 106–110.
- Duan, Y.X., and Guo, W.W. (2004). Flowering regulating genes and their relations with juvenility in woody plants. *J Chin Biotechnol* **24**, 22–26.
- Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z.H., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644–652.
- Griffiths, S., Dunford, R.P., Coupland, G., and Laurie, D.A. (2003). The evolution of *CONSTANS*-like gene families in barley, rice, and *Arabidopsis*. *Plant Physiol* **131**, 1855–1867.
- Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38**, D355–D360.
- Lee, I., Aukerman, M.J., Gore, S.L., Lohman, K.N., Michaels, S.D., Weaver, L.M., John, M.C., Feldmann, K.A., and Amasino, R.M. (1994). Isolation of *LUMINIDEPENDENS*, a gene involved in the control of flowering time in *Arabidopsis*. *Plant Cell* **6**, 75–83.
- Lim, M.H., Kim, J., Kim, Y.S., Chung, K.S., Seo, Y.H., Lee, I., Kim, J., Hong, C.B., Kim, H.J., and Parka, C.M. (2004). A new *Arabidopsis* gene, *FLK*, encodes an RNA binding protein with K homology motifs and regulates flowering time via *FLOWERING LOCUS C*. *Plant Cell* **16**, 731–740.
- Liu, Y.S., Zetter, R., Ferguson, D.K., and Zou, C. (2008). *Lagerstroemia* (*Lythraceae*) pollen from the Miocene of eastern China. *Grana* **47**, 262–271.
- Min, W., Ping, S., Xiang, R.X., and Xiang, Z.Q. (2008). Recent advances in *Lagerstroemia indica* resources and breeding. *Shandong Forestry Sci Technol* **2**, 66–68.
- Pei, F.W., and Rong, Z.Z. (1983). *The Flora of China* (52-2). (Sciences Press, Beijing). pp. 67–92.
- Pillitter, L.J., Walling, L.L., and Lovatt, C.A. (2003). Regulation of flowering in the Washington navel orange, floral genes. *Proc Int Soc Citricult* **1**, 201–204.
- 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.
- Quesada, V., Macknight, R., Dean, C., and Simpson, G.G. (2003). Autoregulation of *FCA* pre-mRNA processing controls *Arabidopsis* flowering time. *Sci Signal* **22**, 3142.
- Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS, the European molecular biology open software suite. *Trends Genet* **16**, 276–277.
- Robson, F., Costa, M.M.R., Hepworth, S.R., Vizir, I., Reeves, P.H., Putterill, J., and Coupland, G. (2001). Functional importance of conserved domains in the flowering-time gene *CONSTANS* demonstrated by analysis of mutant alleles and transgenic plants. *Plant J* **28**, 619–631.
- Saitou, N., and Nei, M. (1987). The neighbor-joining method, a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.
- Schmitz, R.J., Hong, L., Fitzpatrick, K.E., and Amasino, R.M. (2007). *DICER-LIKE 1* and *DICER-LIKE 3* redundantly act to promote flowering via repression of *FLOWERING LOCUS C* in *Arabidopsis thaliana*. *Genetics* **176**, 1359–1362.
- Schomburg, F.M., Patton, D.A., Meinke, D.W., and Amasino, R.M. (2001). *FPA*, a gene involved in floral induction in *Arabidopsis*, encodes a protein containing RNA-recognition motifs. *Plant Cell* **13**, 1427–1436.
- Simpson, G.G., Dijkwel, P.P., Quesada, V., Henderson, I., and Dean, C. (2003). *FY* is an RNA 3' end-processing factor that

- interacts with *FCA* to control the *Arabidopsis* floral transition. *Cell* **113**, 777–787.
- Sun, C.H., Deng, X.J., Fang, J., and Chu, C.C. (2007). An overview of flowering transition in higher plants. *Hereditas* **29**, 1182.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., *et al.* (2011). MEGA5, molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731–2739.
- Wang, R., Xu, S., Jiang, Y., Jiang, J., Li, X., Liang, L., He, J., Peng, F., and Xia, B. (2013). *De novo* sequence assembly and characterization of *Lycoris aurea* transcriptome using GS FLX titanium platform of 454 pyrosequencing. *PLoS One* **8**, e60449.
- Weigel, D., and Nilsson, O. (1995). A developmental switch sufficient for flower initiation in diverse plants. *Nature* **377**, 495–500.
- Wu, F., Price, B.W., Haider, W., Seufferheld, G., Nelson, R., and Hanzawa, Y. (2014). Functional and evolutionary characterization of the *CONSTANS* gene family in short-day photoperiodic flowering in soybean. *PLoS One* **9**, e85754.
- Yuceer, C., Harkess, R.L., Land, S.B., and Luthe, D.S. (2002). Structure and developmental regulation of *CONSTANS-LIKE* genes isolated from *Populus deltoides*. *Plant Sci* **163**, 615–625.
- Zhang, J.Z., Ai, X.Y., Sun, L.M., Zhang, D.L., Guo, W.W., *et al.* (2011). Transcriptome profile analysis of flowering molecular processes of early flowering trifoliate orange mutant and the wild-type [*Poncirus trifoliata* (L.) Raf.] by massively parallel signature sequencing. *BMC Genomics* **12**, 63.
- Zhang, X.M., Zhao, L., Larson-Rabin, Z., Li, D.Z., and Guo, Z.H. (2012). *De novo* sequencing and characterization of the floral transcriptome of *Dendrocalamus latiflorus* (Poaceae, Bambusoideae). *PloS One* **7**, e42082.

Address correspondence to:

Ya Li, PhD

Institute of Botany

Jiangsu Province & Chinese Academy of Sciences

Nanjing 210014

China

E-mail: yalicnbg@aliyun.com

Received for publication April 18, 2014; received in revised form May 18, 2014; accepted May 19, 2014.