Transcriptome changes in the phenylpropanoid pathway in senescing leaves of Toona sinensis

Juanjuan Sui1 · Changqing Qu1 · Jingxia Yang1 · Wenna Zhang2 · Yuntao Ji1

Received: 13 September 2018 / Revised: 27 May 2019 / Accepted: 3 June 2019 / Published online: 17 June 2019
© Franciszek Górski Institute of Plant Physiology, Polish Academy of Sciences, Kraków 2019

Abstract
Toona sinensis is a deciduous tree native to eastern and southeastern Asia that has important culinary and cultural values. To expand current knowledge of the transcriptome and functional genomics in this species, a de novo transcriptome sequence analysis of young and mature leaf tissues of T. sinensis was performed using the Illumina platform. Over 8.1 Gb of data were generated, assembled into 64,541 unigenes, and annotated with known biological functions. Proteins involved in primary metabolite biosynthesis were identified based on similarities to known proteins, including some related to biosynthesis of carbohydrates, amino acids, lipids, and energy. Analysis of unigenes differentially expressed between young and mature leaves (transcriptomic libraries ‘YL’ and ‘ML’, respectively) showed that the KEGG pathways of phenylpropanoid, naringenin, lignin, cutin, suberin, and wax biosynthesis were significantly enriched in mature leaves. These results not only expand knowledge of transcriptome characteristics for this valuable species, but also provide a useful transcriptomic dataset to accelerate the researches on its metabolic mechanisms and functional genomics. This study can also further the understanding of unique aromatic metabolism and Chinese medicinal properties of T. sinensis.

Keywords
Toona sinensis · Transcriptome analysis · Phenylpropanoid · Leaf senescence

Abbreviations
DEGs  Differentially expressed genes
FPKM  Fragments per kilobase of exon model
GC  Gallocatechin
COG  Clusters of orthologous groups of protein
GO  Gene ontology
KEGG  Kyoto encyclopedia of genes and genomes
PAL  Phenylalanine ammonia lyase
4CL  4-Coumarate-CoA ligase
CHS  Chalcone synthase
CCR  Cinnamoyl-CoA reductase
CAD  Cinnamyl-alcohol dehydrogenase
C4H  Cinnamate 4-hydroxylase
F3′5′H  Flavonoid 3′,5′-hydroxylase
FLS  Flavonol synthase
HCT  p-Hydroxycinnamoyl-CoA:chiquimato/quinato p-hydroxycinnamoyl transferase
CYP98A  Coumaroylquinate (coumaroylshikimate) 3′-monooxygenase
CoumCoA3H  Coumaric acid CoA hydrolase
REF  Coniferyl-aldehyde dehydrogenase
POD  Peroxidase

Introduction

Chinese mahogany (Taihe Toona sinensis Roem, syn. Cedrela sinensis, family Meliaceae) is a perennial woody tree that grows 25 m high, and is used as a source of food, timber, and medicine, particularly in the Anhui province of China. Regarded as nutritious food, the edible buds and young leaves are commonly used to make the condiment Toona Paste, which has a floral and onion-like flavor (Park...
et al. 1996; Edmonds and Staniforth 1998). The unique flavor results from various natural compounds including triterpenes, phenolics, flavonoids, and lysine amino acid (Mu et al. 2007; Zhou et al. 2011; Kakumu et al. 2014; Zhang et al. 2015). The mature, fibrous leaves of *T. sinensis* are used in Chinese traditional medicines to treat conditions ranging from diarrhea and other intestinal complaints to reproductive concerns and cancer. Recently, other biological properties of *T. sinensis* leaf extracts have been reported, including anti-inflammatory, analgesic, inhibition of boil growth inhibition, antioxidant, anti-diabetic, and anti-neoplastic, as well as anti-atherosclerotic, and inhibition of replication of the severe acute respiratory syndrome (SARS) coronavirus and of the pandemic influenza A (H1N1) virus (Hsu et al. 2003; Chia et al. 2010; Huang et al. 2012; Yang et al. 2013, 2014; You et al. 2013).

The nutritional value and potential health benefits of *T. sinensis* require further investigation. Currently, only very limited information is available about the compounds contributing to the flavor of young leaves and the medicinal content of mature leaves of *T. sinensis*. Only a few reports have addressed the effects of flavonoids on the taste of young leaves. Flavonoids, lysine, and polyphenols increase the antioxidant capacity of plant cells and associated tissues, and are responsible for the antioxidant properties of *T. sinensis* buds and young leaves (Wang et al. 2007; Vinodhini and Lokeswari 2014).

Recent rapid developments in bioinformatics have allowed the transcriptome approach to emerge as a powerful method for direct sequencing. RNA-Seq, or whole transcriptome shotgun sequencing, can now be used for transcriptome studies due to its high-throughput and high-resolution capabilities (Young et al. 2010; Torre et al. 2014). RNA-Seq allows analysis of complex transcriptional regulation and variable metabolic pathways of different flavonoids, including across different groups or tissues (Shi et al. 2014). Previous transcriptome studies in *T. sinensis* using other species allowed increased understanding of multiple aspects of the biochemistry, development, and metabolism of leaves and shoots, as well as new insights into the biosynthesis of metabolic compounds (Long et al. 2014; Wang et al. 2015).

In this study, we sequenced the transcriptomes of young and mature leaves of *T. sinensis*. RNA sequencing data was de novo assembled and annotated, and candidate gene expression changes were characterized. For the first time, molecular regulation of the phenylpropanoid and naringenin biosynthesis pathways was characterized in this species. Transcriptome differences between young and mature leaves described in the current study provide crucial resources for gene annotation and discovery, and gene function analysis. Moreover, our sequencing results enhance understanding of biosynthesis of phenylpropanoid and cutin, and provide insights into the potential molecular mechanisms of pharmacological action in *T. sinensis*, which can promote production and yield of phenylpropanoid for medicinal or culinary purposes of *T. sinensis*.

**Materials and methods**

**Plant materials and growth conditions**

Mature 5-year trees of the *T. sinensis* cultivar ‘Heiyouchun’ were sampled from a *T. sinensis* industry demonstration zone in Taihe County, Anhui, China. The first to third pinnate fronds with purple color were identified as young leaves (YL), and the sixth to eighth green pinnate fronds were considered as mature leaves (ML) (Fig. 1a). Young and mature leaves were harvested randomly from three *T. sinensis* clones, which were propagated by asexual reproduction and thus had the same genetics as the ‘Heiyouchun’ cultivar. At least 20 YL or ML were mixed in each sample pool for RNA-seq analysis. All samples were immediately immersed in liquid nitrogen and stored at −80 °C.

**RNA extraction and cDNA library construction**

Total RNA was extracted from leaf samples with TRIzol Reagent (Cat. #15596026, Invitrogen, Carlsbad, CA, USA) and then treated with DNase I (Invitrogen, Cat. #18047019) according to the established methods. To determine RNA quality and concentration, 1 µl of each RNA sample was electrophoresed (2%, agarose, 1x TBE) and quantified using a NanoDrop ND–1000 (Thermo Scientific). In addition, RNA integrity number (RIN) was determined with the Agilent 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA). At least 20 µg of total RNA was combined with oligo(dT) magnetic beads (concentration ≥ 250 ng/µl, OD260/280 = 1.8~2.2, OD260/230 ≥ 2.0, 28S:18S ≥ 1.0) and used to confirm that the RNA integrity number (RIN) value was greater than 8.0 before further library construction. RNA-Seq libraries were prepared using the TruSeq RNA Sample Prep Kit (RS-122-2001, Illumina Inc., San Diego, CA, USA). Buffer reagent was used to fragment the extracted mRNA, and the resulting fragmented mRNA was reverse transcribed into cDNA, with purified short fragments used for end repair and ligated with adaptors. The cDNA was enriched by PCR amplification and quality was confirmed with BioAnalyzer, after which RT-PCR was used to quantify the cDNA library, and it was sequenced (Illumina HiSeq™ 4000, BGI, Shenzhen, China), generating paired-end reads with 150 bp in length.
Data processing and de novo assembly

Raw reads were preprocessed using the filter-fq software (v1.2.0; https://github.com/bowentan/filterfq) to discard the adapters (> 5 bp), low-quality fragments (with a quality score $Q \leq 19$) or N (unknown nucleotide) content $> 5\%$, and those fragments shorter than 50 bp (including redundant sequences). This clean, high-quality data were used to...
calculate $Q_{20}$ and $Q_{30}$ values, levels of GC content and sequence duplication, and for all downstream analyses. The resulting paired-end reads were clustered using TGICL software (Pertea et al. 2003) to analyze the length and distribution of the transcriptional and unigene clusters. Paired-end sequences were separated into two files, “left” reads into the “left.fq” file and “right” reads into the “right.fq” file. Reads that uniquely mapped to the left contigs were considered to be derived from *T. sinensis*. Any reads matching to genus qualified them as right reads. Unmatched reads at this stage of the process were considered a set of singleton reads and also placed into the right.fq file. Potential transcripts and unigenes were assembled from the pooled clean reads of left.fq and right.fq files using Trinity software (v20140717) (Manfred 2011).

**Gene annotation and analysis**

Trinotate was used to perform the functional annotation of unigenes and ORFs (Bryant et al. 2017). We processed all unigene sequences for identification and functional annotation including homology search with known databases including NCBI’s Nt (nonredundant nucleotide sequences), GO (gene ontology), COG (Cluster of Orthologous Groups), and KEGG (Kyoto Encyclopedia of Genes and Genomes). The highest similarity of aligned proteins was used for functional annotation of unigene sequences. First, BLASTx and BlastN (both with parameters of match length $\geq 90$ bp, e value $< 1 \times 10^{-5}$, and the allowance of $\leq 1$ mismatch and 1 gap, and identity $\geq 90\%$) were used to align unigenes to protein databases and Nt, respectively. Subsequently, ESTScan software was used to determine sequence direction. Then Blast2GO was employed to determine GO annotation against the GO database for unigenes annotated by NCBI Nr (nonredundant protein sequences) (Conesa et al. 2005; Götz et al. 2008). InterProScan (v5) was used to give further protein annotation. Prediction of protein-coding regions was performed with OrfPredictor software (Min et al. 2005). Additionally, GO functional classification of all unigenes was performed using the Web Gene Ontology Annotation Plot (WEGO) software (Ye et al. 2006), which visualizes and characterizes gene functions and distributions across different pathways.

**Analysis of differentially expressed genes (DEGs)**

Gene expression levels were estimated by mapping clean reads to the Trinity transcript assembly using RNA-Seq by Expectation–Maximization (RSEM) (Li and Dewey 2011) for each sample. The abundance of each gene was normalized and calculated using the unigene expression via the Reads Per Kilo bases per Million reads (RPKM) method (Mortazavi et al. 2008) as follows:

$$\text{FPKM} = \frac{10^6 C}{NL/10^3},$$

where $C$ and $N$ represent the counts of mapped reads uniquely aligned to a unigene and the sum of reads sequenced that were uniquely aligned to total unigenes, respectively, and $L$ represents the sum of a unigene in base pairs.

The DEGSeq package in R was used to conduct differential expression analysis for the young and mature leaves by modeling count data with negative binomial distributions (Anders and Huber 2010). $P$ values were adjusted to reduce false positives due to multiple testing (Storey and Tibshirani 2003), with a $q$ value $< 0.05$ and llog$_2$ (ratio)$\geq 1$ set as the thresholds for significantly differential expression between the two samples. The identified differentially expressed genes (DEGs) were analyzed according to KEGG enrichment pathways and GO functional categories. GO enrichment analyses were conducted in GOseq with the Wallenius’ noncentral hypergeometric distribution used to search for and map all significantly enriched GO terms among the DEGS (Young et al. 2010). KEGG online tools were used for pathway enrichment analysis of the DEGs (http://www.kegg.jp/) (Mao et al. 2005).

**Results**

**Sequencing and de novo assembly of the transcriptome of *T. sinensis* young and mature leaves**

To eliminate genetic differences of individual cultivars, at least 20 young leaves (YL) or mature leaves (ML) from three individual *T. sinensis* ‘Heiyouchun’ cultivars were mixed in each sample pool for RNA extraction. To construct high-quality YL and ML RNA libraries, total RNA quality was determined by agarose gel electrophoresis, Nanodrop, and RNA integrity number (RIN) value is shown in Supplemental Fig. 1. The quality of YL RNA was 38 µg (concentration = 411 ng/µl, OD$_{260/280}$ = 2.1, OD$_{260/230}$ = 2.0, 28S/18S = 1.8, RIN = 10), and quality of ML RNA was 24 µg (concentration = 532 ng/µl, OD$_{260/280}$ = 2.1, OD$_{260/230}$ = 1.3, 28S/18S = 1.7, RIN = 8.9). The quality was satisfactory for use in constructing libraries.

To generate a complete *T. sinensis* leaf transcriptome, two cDNA libraries from YL and ML were constructed and sequenced using the Illumina HiSeq™ 4000 platform, generating 3.82 and 3.10 Gb of raw RNA-seq data, respectively. After deletion of adaptor-polluted, redundant, and
other low-quality sequences, 3.32 and 2.80 Gb clean reads of YL and ML, respectively, were retained and assembled. For these clean reads, the Q30 scores (sequencing error rate, 0.1%) were 97.32% and 95.44%, and GC contents were 40.06% and 40.26%, generated from the transcriptome libraries of ‘YL’ and ‘ML’, respectively (Fig. 1b, Table 1).

After filtration, the Trinity tool was used to assemble independent high-quality clean sequences from each library, which were further merged, generating 102,881 transcripts and 64,541 unigenes. These transcripts were 107,527,675 bp and 53,892,623 bp with unigene GC contents of 40.16% and 39.94% of YL and ML, respectively (Table 2). An overview of the sequence size distribution of transcripts and unigenes is shown in Supplemental Table 1.

The quality and quantity of raw sequence data were sufficient to perform further analysis. 63.16% (40,767) of the unigenes were between 200 and 600 bp in length, 12.11% (7814) were between 600 and 1000 bp, 14.19% (9160) were between 1 and 2 kb, 8.90% (5743) were between 2 and 4 kb, and unigenes of lengths more than 4 kb accounted for only 1.64% (1057) (Table 3).

### Functional annotation

To obtain functional annotations, we subjected all generated unigenes to BLASTx alignment using a serial blast with a cut-off e value 1e−5, in the NCBI databases and sequence homologies. In total, 75,779,461 raw reads (27.80% of the total reads) were annotated. Of these unigenes, 1746 were annotated with the Nr database and 726 with Nt (Fig. 1c); 33,791 with UniProt (including Swiss-Prot, TrEMBL, and PIR-PSD); 20,515 with GO (Supplemental Table 2), 8696 with COG; 5482 with KEGG; and 23,970 with PFAM.

The 20,515 unigenes annotated with GO were assigned to categories including molecular functions, cellular processes, and biological processes (Fig. 2). The two most abundant unigene sequences belonged to cellular processes (4644, 22.63%) and metabolic processes (4375, 21.33%) within biological processes. Unigenes involved in cellular processes were distributed in cell and cell parts (6044 unigenes, 33.83%), organelles (2052, 11.48%), and plasma membrane (3377, 18.90%). Unigenes involved in molecular functions played roles in binding (4744, 42.59%) and catalytic activity (4107, 36.87%), whereas 20.53% represented activity proteins, including transporters, structural molecules, molecular transducers, enzyme regulators, receptors, antioxidants, electron carriers, and transcription factors.

COG analysis aligned 8696 unigenes for functional classification (Fig. 3). For 14.26% (1240 unigenes), a general function was predicted while translation, ribosomal structure, and biogenesis accounted for 9.50% (826), posttranslational modification was related to 8.41% (732), 6.41% (556) were engaged in carbohydrate transport and metabolism, amino acid transport and metabolism accounted for involved 5.76% (501), replication functions were predicted for 4.37% (380), and 2.94% (256) were involved in transcription.
Fig. 2 GO annotation categories of assigned unigenes. The annotated 20,515 unigenes were assigned to GO annotation categories of molecular function, cellular process, and biological process categories.
Assembled unigenes were assigned to metabolic pathways in the KEGG database based on sequence similarity (Fig. 4). Of the 5482 unique mapped sequences, 14.61% (801) were assigned to amino acid metabolism pathways and 8.31% (456) to ribosome metabolism and translation pathways; 2.77% (152) were involved in the immune system; 2.04% (112) were classified under biosynthesis of secondary metabolites; 1.17% (64) under metabolism of terpenoids and polyketides; 0.97% (53) were assigned to phenylpropanoid biosynthesis; and 0.53% (14) to flavonoid biosynthesis.

Differentially expressed genes between young and mature leaves

To identify genes with different expression levels between YL and ML, the unigene expression levels were calculated with the RPKM method, which accounts for effects of both sequencing depth and gene length on the read count (Fig. 5a). A total of 15,172 unigenes had differential expression (with \( q \) value < 0.05 and \( \log_2 \) (ratio) ≥ 1) between the two samples and thus were identified as differentially expressed genes (DEGs). Among these DEGs, 9648 were up-regulated and 5524 were down-regulated in ML compared with YL (Fig. 5b). DEGs mapped within each GO term category were counted. The hypergeometric test revealed that a total of 67 functional groups, including molecular functions, cellular components, and biological processes, showed remarkable enrichment in DEGs compared with the transcriptomic background (Fig. 6).

Several enriched pathways, including amino acid biosynthesis, signal transduction, and metabolic pathways, were identified using KEGG enrichment analysis of DEGs. A total of 308 pathways by DEGs are shown in Table 4 and Supplement Table 4, with 22 metabolic pathways significantly over-represented. Significantly highly enriched pathways of YL samples were primarily related to plant biological and human pathogen resistance metabolism pathways, including general ribosome (ko03010), cell cycle (ko04110), RNA transport (ko03013), ribosome biogenesis in eukaryotes...
(ko03008), DNA replication (ko03030); HTLV-I infection (ko05166), Fanconi anemia pathway (cellular response to DNA interstrand crosslink) (ko03460), and systemic lupus erythematosus (ko05332). The most enriched pathways in samples of ML were related to secondary metabolism pathway, including ribosome (ko03010); phenylpropanoid biosynthesis (ko00940); pathogenic E. coli infection (ko05130); axon guidance (ko04360), hypertrophic cardiomyopathy (HCM) (ko05410); cutin, suberin, and wax biosynthesis (ko00037), dilated cardiomyopathy (ko05414); photosynthesis antenna protein (ko00196); malaria (similar to plant galactolipid metabolism pathway) (ko05144); flavonoid biosynthesis (ko00941); nitrogen metabolism (ko00910); carotenoid biosynthesis (ko00906); limonene and pinene degradation and stilbenoid, diarylheptanoid and gingerol biogenesis (ko00906). The phenylpropanoid biosynthesis exclusive ribosome in those metabolic pathways related specific medical traits of T.sinensis was a most significantly

Fig. 4  KEGG metabolic pathways of assembled unigenes. Assembled unigenes were assigned to metabolic pathways in the KEGG database based on sequence similarity

© Springer
enriched pathway, which included 33 up-regulated genes and 20 down-regulated genes.

Differentially expressed genes (DEGs) related to phenylpropanoid biosynthesis in mature leaves

Many unigenes related to phenylpropanoid biosynthesis were identified in ML transcriptome (Fig. 7). Transcriptome analysis revealed that 53 enzyme genes related to phenylpropanoid biosynthesis (Table 5) were up-regulated compared with YL, including genes related to the general phenylpropanoid pathway [4CL (6)], caffeic acid biosynthesis [Coum-CoA3H (2), HCT (4), CYP98A (2)], and the later steps of lignin biosynthesis [CCR (5), CAD (5), REF1 (2), POD (5), CAD (3)] (Supplemental Fig. 1). These results indicated that caffeoyl-CoA, flavonoids, and lignin were each metabolized in an enzyme-dependent manner and accumulated in ML extracts. In addition, almost all major enzyme genes involved in cutin, suberin, and wax biosynthesis were annotated in this pathway (Supplemental Fig. 2; Supplemental Table 3). Despite this increased information, the complexity of the molecular mechanism for the biosynthesis of cutin, suberin, and wax in mature leaves of *T. sinensis* remains uncertain and requires further study.

Discussion

Comparison of software packages for detecting gene differential expression of *T. sinensis* young and mature leaves

Transcriptome sequencing can be used to efficiently and effectively analyze the cellular transcriptome. Many computational software packages and pipelines have already been widely used during RNA-seq data analysis, including edgeR (Robinson et al. 2010), DESeq (Anders and Huber 2010), DEGSeq (Wang et al. 2010), and limma (Smyth 2004).

edgeR is normally used to determine differential expression with empirical Bayes estimation and exact tests based on a negative binomial model. edgeR can be used for small numbers of replicates with over-dispersed data to assess differential gene expression. TMM normalization and Benjamini–Hochberg procedures are used as default to control sequencing depths and FDR (Anders and Huber 2010).

edgeR is normally used to determine differential expression with empirical Bayes estimation and exact tests based on a negative binomial model. edgeR can be used for small numbers of replicates with over-dispersed data to assess differential gene expression. TMM normalization and Benjamini–Hochberg procedures are used as default to control sequencing depths and FDR, respectively (Robinson et al. 2010).

Similar to edgeR, DESeq also uses a negative binomial model, a scaling factor normalization procedure and the Benjamini–Hochberg procedure to control sequencing depths and FDR of different samples, but exhibits more general dispersion estimation and balanced selection of DEGs. DESeq is technically possible to use with experiments without any biological replicates but this is not recommended (Anders and Huber 2010).

Limma was originally used for microarray data analysis but was later extended to RNA-seq data. TMM normalization of the edgeR package and ‘voom’-conversed log₂ scale
are used to determine weight prior to linear modeling. The Benjamini–Hochberg procedure is used as default to estimate FDR (Smyth 2004).

DEGseq exports gene expression values in a table format, which are then directly processed by edgeR. It analyzes gene expression based on a random sampling model or raw counts in Poisson distribution model. DEGseq can
| New leaf vs. mature leaf | Map               | Count1 | Count2 | Count3 | Count4 | p  | q   | Up_Count | Down_Count | Significance |
|-------------------------|-------------------|--------|--------|--------|--------|----|-----|----------|------------|--------------|
| Ribosome                | map03010          | 265    | 66     | 1745   | 2923   | 4.45E−53 | 1.37E−50 | 48        | 217         | Yes          |
| Phenylpropanoid biosynthesis | map00940      | 53     | 41     | 1957   | 2948   | 0.001006 | 0.01822 | 33        | 20          | Yes          |
| Pathogenic *Escherichia coli* infection | map05130  | 34     | 14     | 1976   | 2975   | 1.57E−05 | 0.000805 | 17        | 17          | Yes          |
| Axon guidance           | map04360          | 27     | 14     | 1983   | 2975   | 0.000767 | 0.016878 | 14        | 13          | Yes          |
| Hypertrophic cardiomyopathy (HCM) | map05410   | 21     | 9      | 1989   | 2980   | 0.000899 | 0.017299 | 14        | 7           | Yes          |
| Cutin, suberin and wax biosynthesis | map00073  | 23     | 6      | 1987   | 2983   | 1.95E−05 | 0.000858 | 13        | 10          | Yes          |
| Dilated cardiomyopathy  | map05414          | 19     | 8      | 1991   | 2981   | 0.001438 | 0.023305 | 13        | 6           | Yes          |
| Arrhythmogenic right ventricular cardiomyopathy (ARVC) | map05412 | 19     | 7      | 1991   | 2982   | 0.000682 | 0.016161 | 12        | 7           | Yes          |
| Cell adhesion molecules (CAMs) | map04514     | 16     | 5      | 1994   | 2984   | 0.000873 | 0.017299 | 11        | 5           | Yes          |
| ECM–receptor interaction | map04512       | 13     | 4      | 1997   | 2985   | 0.002626 | 0.036758 | 10        | 3           | Yes          |
| Malaria                 | map05144          | 12     | 3      | 1998   | 2986   | 0.002005 | 0.029412 | 10        | 2           | Yes          |
| Hematopoietic cell lineage | map04640       | 11     | 2      | 1999   | 2987   | 0.001367 | 0.023305 | 9         | 2           | Yes          |
| RNA transport           | map03013          | 86     | 68     | 1924   | 2921   | 4.96E−05 | 0.001911 | 8         | 78          | Yes          |
| Photosynthesis-antenna proteins | map00196     | 8      | 0      | 2002   | 2989   | 0.000677 | 0.016161 | 8         | 0           | Yes          |
| HTLV-I infection        | map05166          | 61     | 24     | 1949   | 2965   | 2.83E−09 | 2.90E−07 | 6         | 55          | Yes          |
| Systemic lupus erythematosus | map05322      | 29     | 8      | 1981   | 2981   | 2.36E−06 | 0.000145 | 6         | 23          | Yes          |
| Gap junction            | map04540          | 15     | 5      | 1995   | 2984   | 0.001683 | 0.025917 | 6         | 9           | Yes          |
| Progesterone-mediated oocyte maturation | map04914   | 41     | 16     | 1969   | 2973   | 1.08E−06 | 8.28E−05 | 2         | 39          | Yes          |
| Cell cycle              | map04110          | 82     | 29     | 1928   | 2960   | 3.61E−13 | 5.55E−11 | 1         | 81          | Yes          |
| Ribosome biogenesis in eukaryotes | map03008     | 51     | 32     | 1959   | 2957   | 6.66E−05 | 0.002278 | 1         | 50          | Yes          |
| DNA replication         | map03030          | 32     | 18     | 1978   | 2971   | 0.000545 | 0.01527  | 0         | 32          | Yes          |
| Fanconi anemia pathway  | map03460          | 28     | 12     | 1982   | 2977   | 0.000125 | 0.003855 | 0         | 28          | Yes          |
also be applied to identify differential expression of exons or pieces of transcripts with or without a small number of replicates. In our study, to get higher sequencing depth and detect subtle gene expression changes, we directly pooled 20 individual biological replicates together into YL and ML sample groups. Due to lack of replicates, DEGseq was more suitable than the other programs to conduct differential gene expression analysis. When we use DEGseq package, it will first homogenize the sample when analyzing single replicate (this homogenization process will avoid the biasness to some extent) according to the internal arithmetic method, and then we analyze the difference based on the data after homogenization instead of directly analyzing the difference between the original data input. To sort off reliable DEGs, the software accounting calculates the corresponding p value and corrected q value. In addition, DESeq detected DEGs based on the level of gene expression according to the negative binomial distribution of statistical methods. The obtained p value will be corrected to control false-positive results according to Benjamini and Hochberg methods. The corrected q value $< 0.05$ and $\lvert \log_2 \text{ratio} \rvert \geq 1$ set as the thresholds is defined as DEGs.

Characterization, assembly, and gene annotation of leaves of \textit{T. sinensis}

In this study, using transcriptome sequencing analysis, we obtained 64,541 unigenes with an N50 value of 1563 bp and a mean length of 835 bp and used these for assembly evaluation by comparison with NCBI and sequence homologies. In total, 20,515 (43.06\%) of these unigenes were successfully annotated using BLAST searches of the public Nr, PFAM, Swiss-Prot, GO, COG, and KEGG databases. The resulting RNA-Seq data provided a high-quality annotated assembly for \textit{T. sinensis} generated by comprehensive analysis. Distribution patterns annotated similarly across several databases indicated that YL and ML of \textit{T. sinensis} undergo multiple unique developmental processes (Fig. 1; Supplemental Table 2). The large number of annotated enzymes suggests the presence of genes associated with different pathways of primary and secondary metabolite biosynthesis across life stages (Zhang et al. 2016; Zhao et al. 2017).

Differentially expressed unigenes in phenylpropanoid biosynthesis

Our findings demonstrate that the phenylpropanoid and lignin biosynthesis pathways were among the most enriched. Nine differentially expressed unigenes, including 4CL, CoumCoA3H, HCT, CYP98A, CCR, REF1, CAD, and POD, were up-regulated in YL and ML. ML were significantly enriched in phenylpropanoids, consistent with increased content of flavonoid, lignin, cutin, and wax. In plants, control of phenylpropanoid biosynthesis is complex and plays a significant role in pathogen resistance, anthocyanin biogenesis, and pharmacology (Jimenez and Riguera 1994). In this transcriptome study, we identified most of the
catabolic genes associated with phenylpropanoid synthesis, demonstrating an understanding of the precise pathway in plants (Shi et al. 2013). Genetic, molecular, and biochemical evidence suggests that synthesis and catabolism of phenylpropanoid amino acids are regulated by previously undescribed coordinated mechanisms (Burkhard et al. 2001; Grabherr et al. 2011). Information from the current study will advance understanding of the regulation of phenylpropanoid metabolism in T. sinensis, which will provide valuable information for the future production of high-phenylpropanoid crops with medical applications.

### Author contribution statement
Conceived and designed the experiments: JS, CQ, JY, and YJ. Conducted the experiments: JS, CQ, and JY. Performed data analysis: JS, CQ, and WZ. Wrote a draft of the manuscript: WZ and YJ.

### Table 5
Changes in transcript abundance of candidate genes related to phenylpropanoid biosynthesis in old leaves and new leaves

| Gene       | Mature leaf normalization | Young leaf normalization | log2 fold change | P value | Up/down | PFAM name      | PFAM description                                                                 |
|------------|---------------------------|--------------------------|------------------|---------|---------|---------------|----------------------------------------------------------------------------------|
| c35946_g1 | 1.21997                   | 0.013848952              | 6.46093          | 1.20E−11| Up      | p450          | Cytochrome P450                                                                   |
| c36184_g1 | 3.25325                   | 0.055395809              | 5.8796           | 7.72E−29| Up      | Methyltransf_3 | O-Methyltransferase                                                              |
| c44663_g2 | 0.31952                   | 0.013848952              | 4.52804          | 0.00067591| Up      | p450          | Cytochrome P450                                                                   |
| c34177_g1 | 30.7897                   | 2.215832377              | 3.79653          | 2.43E−221| Up      | Peroxidase    | Peroxidase                                                                       |
| c22474_g1 | 7.08745                   | 0.52662019               | 3.75142          | 5.25E−52| Up      | CYP98A3       | C3'H; Coumaroylquininate (coumaroyl-shikimate) 3'-monooxygenase                   |
| c16277_g1 | 1.33616                   | 0.110791619              | 3.59217          | 8.85E−11| Up      | p450          | Cytochrome P450                                                                   |
| c56414_g1 | 1.77186                   | 0.166187428              | 3.41438          | 2.47E−13| Up      | Beta-glucosidase. | Beta-glucosidase                                                                |
| c43864_g1 | 4.61846                   | 0.498562285              | 3.21157          | 1.11E−30| Up      | Peroxidase    | Peroxidase                                                                       |
| c6701_g1  | 211.026                   | 26.14682205              | 3.01271          | 0       | Up      | Epimerase     | NAD dependent epimerase/dehydratase family                                        |
| c35488_g1 | 0.58094                   | 0.083093714              | 2.80557          | 0.00012938| Up      | Aldedh       | Aldehyde dehydrogenase family                                                    |
| c39663_g1 | 49.4378                   | 7.173757321              | 2.78481          | 4.78E−271| Up      | Transferase   | Transferase family                                                               |
| c44868_g1 | 96.0581                   | 17.44967997              | 2.46071          | 0       | Up      | Glyco_hydro_1 | Glycosyl hydrolase family 1                                                      |
| c41387_g1 | 140.965                   | 29.11049785              | 2.27572          | 0       | Up      | Methyltransf_3 | O-Methyltransferase                                                              |
| c51066_g1 | 65.036                    | 15.31694131              | 2.08611          | 4.85E−255| Up      | Glyco_hydro_1 | Glycosyl hydrolase family 1                                                      |
| c51757_g1 | 31.3416                   | 222.6080602              | −2.8284          | 0       | Down    | Glyco_hydro_3 | Glycosyl hydrolase family 3N-terminal domain                                     |
| c36298_g2 | 5.02512                   | 44.45513707              | −3.1451          | 2.35E−274| Down    | ADH_N         | Alcohol dehydrogenase GroES-like domain                                          |
| c44584_g1 | 0.52284                   | 11.46693255              | −4.455           | 1.57E−91| Down    | Peroxidase    | Peroxidase                                                                       |
| c31990_g1 | 0.11619                   | 7.450736368              | −6.0029          | 9.94E−64| Down    | Peroxidase    | Peroxidase                                                                       |
| c42734_g1 | 0.23238                   | 2.188134472              | −3.2352          | 1.84E−15| Down    | Peroxidase    | Peroxidase                                                                       |
| c51571_g3 | 0.66808                   | 3.185259042              | −2.2533          | 2.35E−15| Down    | Peroxidase    | Peroxidase                                                                       |
| c37802_g1 | 0.05809                   | 0.99712457               | −4.1013          | 4.73E−09| Down    | Peroxidase    | Peroxidase                                                                       |
| c57658_g1 | 0.01452                   | 0.775541332              | −5.7387          | 5.28E−08| Down    | p450          | Cytochrome P450                                                                   |
| c15574_g1 | 0.05809                   | 0.609353904              | −3.3908          | 1.86E−05| Down    | Peroxidase    | Peroxidase                                                                       |
| c346_g1   | 28.0012                   | 214.4648762              | −2.9372          | 0       | Down    | adh_short    | Short-chain dehydrogenase                                                        |
| c29453_g1 | 77.5262                   | 193.9684267              | −1.3231          | 0       | Down    | ADH_N        | Alcohol dehydrogenase GroES-like domain                                          |
| c46059_g1 | 215.47                    | 94.72683142              | 1.18564          | 0       | Up      | p450          | Cytochrome P450                                                                   |
| c35840_g1 | 23.6732                   | 84.7001926              | −1.8391          | 5.20E−281| Down    | Glyco_hydro_3 | Glycosyl hydrolase family 3N-terminal domain                                     |
| c37180_g1 | 35.8148                   | 100.1279255             | −1.4832          | 1.03E−244| Down    | Glyco_hydro_1 | Glycosyl hydrolase family 1                                                      |
| c38152_g1 | 28.1174                   | 2.077342854              | 3.75865          | 6.25E−201| Up      | Peroxidase    | Peroxidase                                                                       |
| c53977_g1 | 0.63903                   | 0.166187428             | 1.94308          | 0.00124426| Up      | Peroxidase    | Peroxidase                                                                       |
| c4677_g1  | 0.29047                   | 0.01384952              | 4.39054          | 0.00128935| Up      | 4CL          | 4-Coumarate-CoA ligase                                                           |
| c58363_g1 | 0.01452                   | 0.249281142             | −4.1013          | 0.00340948| Down    | Peroxidase    | Peroxidase                                                                       |
Acknowledgements This work was supported by the earmarked fund for Key Projects of Natural Science Research of Education Department of Anhui Province (KJ2016A871, KJ2016A872), University Innovative Platform Project in Anhui Province (2016), Research and Development Team of Toona sinensis Deep Processing Technology (XDHXTD201702).

Compliance with ethical standards

Conflict of interest The authors have declared that no competing interests exist.

Data availability All relevant supporting data can be found within the additional files accompanying this article. RNA-Seq raw data reads have been deposited in NCBI SRA under the accession number PRJNA516485.

References

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106

Bryant DM, Johnson K, Di Tommaso T, Tickle T, Couger MB, Payzin-Terter S, Adiconis X, Fan L, Raychowdhury R, Zeng Q (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

Jiang L, Zhang J, Tian X, Wu S, Zhang Q et al (2014) De novo assembly of the desert tree Haloxylon ammodendron (C. A. Mey.) based on RNA-Seq data provides insight into drought response, gene discovery and marker identification. BMC Genom 15(1):1111

Manfred GG (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

Mao X, Cai T, Olyarchuk JG, Wei L (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics 21(19):3787–3793

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140

Shi Y, Yan X, Zhao P et al (2013) Transcriptomic analysis of a tertiary rilant plant, extreme xerophyte Reaumuria soongorica to identify genes related to drought adaptation. PLoS One 8(5):e53993

Shi SG, Yang M, Zhang M, Wang P, Kang YX, Liu JJ (2014) Genome-wide transcriptome analysis of genes involved in flavonoid biosynthesis between red and white strains of Magnolia sprengeri pamp. BMC Genom 15:706

Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:1

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100(16):9440–9445

Torre S, Tattini M, Brunetti C et al (2014) RNA-seq analysis of Quercus pubescens leaves: de novo transcriptome assembly, annotation and functional markers development. PLoS One 9(11):e112487

Vinodhini V, Lokeswari TS (2014) Antioxidant activity of the isolated compounds, methanolic and hexane extracts of Toona ciliata leaves. Int J Eng Technol 4(3):135–138

Wang KJ, Yang CR, Zhang YJ (2007) Phenolic antioxidants from Chinese toon (fresh young leaves and shoots of Toona ciliata leaves). Int J Eng Technol 4(3):135–138

Wang L, Feng Z, Wang X et al (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26(1):136–138

Wang X, Li ST, Li J, Li CF, Zhang YS (2015) De novo transcriptome sequencing in Pueraria lobata to identify putative genes involved in isoflavones biosynthesis. Plant Cell Rep 34(5):733–743

Yang SJ, Zhao Q, Xiang HM (2013) Antiproliferative activity and apoptosis-inducing mechanism of constituents from Toona sinensis on human cancer cells. Cancer Cell Int 13 (1):12

Yang HL, Huang PJ, Liu YR et al (2014) Toona sinensis Inhibits LPS-induced inflammation and migration in vascular smooth muscle.
cells via suppression of reactive oxygen species and NF-B signaling pathway [J]. Oxid Med Cell Longev 2014:1–16
Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34:293–297
You HL, Chen CJ, Eng HL, Liao PL, Huang ST (2013) The effectiveness and mechanism of Toona sinensis extract inhibit attachment of pandemic influenza A (H1N1) virus. Evid-Based Complement Altern Med 4:479718
Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11(2):R14
Zhang MF, Jiang LM, Zhang DM, Jia GX (2015) De novo transcriptome characterization of Lilium ‘Sorbonne’ and key enzymes related to the flavonoid biosynthesis. Mol Genet Genom 290:399–412
Zhang X, Song ZQ, Liu T, Guo LL, Li XF (2016) De novo assembly and comparative transcriptome analysis provide insight into lysine biosynthesis in Toona sinensis Roem. Int J Genom 3:6735209
Zhao H, Ren LP, Fan XY, Tang KJ, Li B (2017) Identification of putative flavonoid-biosynthetic genes through transcriptome analysis of Taihe Toona sinensis bud. Acta Physiol Plant 39:122
Zhou QW, Wang CO, Cai W, Gu YJ (2011) Nutrition quality analysis of different Toona sinensis cultivars in Dazhu county. J Sichuan For Sci Technol 32(4):117–119

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.