Genetics and Molecular Research

Transcriptomic analysis of *Camellia ptilophylla* and identification of genes associated with flavonoid and caffeine biosynthesis

M.M. Li[1], J.Y. Xue[1], Y.L. Wen[2], H.S. Guo[1], X.Q. Sun[1], Y.M. Zhang[1] and Y.Y. Hang[1]

[1]Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing, China
[2]Jiangsu Kaiji Biotechnology Co., Ltd., Yixing, China

Corresponding author: Y.Y. Hang
E-mail: hangyueyu@cnbg.net

Genet. Mol. Res. 14 (4): 18731-18742 (2015)
Received July 27, 2015
Accepted October 26, 2015
Published December 28, 2015
DOI http://dx.doi.org/10.4238/2015.December.28.22

ABSTRACT. *Camellia ptilophylla*, or cocoa tea, is naturally decaffeinated and its predominant catechins and purine alkaloids are *trans*-catechins and theobromine Regular tea [*Camellia sinensis* (L.) O. Ktze.] is evolutionarily close to cocoa tea and produces *cis*-catechins and caffeine. Here, the transcriptome of *C. ptilophylla* was sequenced using the 101-bp paired-end technique. The quality of the raw data was assessed to yield 70,227,953 cleaned reads totaling 7.09 Gbp, which were assembled *de novo* into 56,695 unique transcripts and then clustered into 44,749 unigenes. In catechin biosynthesis, leucoanthocyanidin reductase (LAR) catalyzes the transition of leucoanthocyanidin to *trans*-catechins, while anthocyanidin synthase (ANS) and anthocyanidin reductase (ANR) catalyze *cis*-catechin production. Our data demonstrate that two *LAR* genes (*CpLAR1* and *CpLAR2*) by *C. ptilophylla* may be advantageous due to the combined effects of this quantitative trait, permitting increased leucoanthocyanidin consumption for the synthesis of *trans*-catechins. In contrast, the only *ANS* gene observed in *C. sinensis* (*CsANS*) shared high identity (99.2%) to one homolog from *C. ptilophylla* (*CpANS1*), but lower identity (~80%)

to another (*CpANS2*). We hypothesized that the diverged *CpANS2* might have lost its ability to synthesize *cis*-catechins. *C. ptilophylla* and *C. sinensis* each contain two copies of ANR, which share high identity and may share the same function. Transcriptomic sequencing captured two *N*-methyl nucleosidase genes named *NMT1* and *NMT2*. *NMT2* was highly identical to three orthologous genes *TCS2*, *PCS2,* and *ICS2*, which did not undergo methylation *in vitro*; in contrast, *NMT1* was less identical to *TCS*, *PCS* and *ICS*, indicating that *NMT1* may undergo neofunctionalization.

Key words: Cocoa tea; Transcriptome; Trans-catechins biosynthesis; Theobromine biosynthesis

INTRODUCTION

Camellia ptilophylla, commonly referred as cocoa tea, is a naturally decaffeinated tea breed. This tea belongs to the *Camellia* Sect. *Thea* (L.) Dyer in the tea family (Yang et al., 2007; Peng et al., 2008) and was discovered in Guangdong Province by Prof. Zhang Hongda. Cocoa tea has an orchid-like aroma, a pure and rich taste, and its decaffeinated nature makes it a suitable drink for people of all ages and for those advised not to consume caffeine, including pregnant women, insomniacs, and individuals with elevated blood pressure. In addition, cocoa tea is considered to be a healthcare drink, and is an effective antipyretic, has roles in detoxification and blood lipid clearance, and can lower blood sugar. Cocoa tea is therefore referred to as 'Bai-sui' tea by local citizens, meaning that this tea could make its drinkers live as long as 100 years.

Regular tea, *C. sinensis* (L.) O. Ktze., is evolutionarily close to cocoa tea, and although they are both assigned to the genus *Camellia* systematically, the main metabolic compounds produced by the two species are quite different. The predominant catechins and purine alkaloids in regular tea are *cis*-catechins and caffeine (1,3,7-methylxanthine); in contrast, cocoa tea mainly contains *trans*-catechins and theobromine (3,7-methylxanthine). The *trans*-catechins include (+)-gallocatechin gallate (GCG), (+)-gallocatechin gallate (GCG), (+)-catechin gallate (CG), (+)-gallocatechin (GC), (+)-catechin (C), and the *cis*-catechins include (-)-epigallocatechin gallate (EGCG), (-)-epigallocatechin (EGC), (-)-epicatechin gallate (ECG), and (-)-epicatechin (EC). On one hand, the different catechin isoforms may account for differences in the specific health-care effects of cocoa tea and regular tea; whereas, the replacement of caffeine by theobromine creates a wider consumer population for cocoa tea.

The molecular mechanisms that lead to differences in the predominant catechins and purine alkaloids in the two tea species remains poorly understood. Stafford (1990) hypothesized that an epimerase converts C and GC to EC and EGC, respectively. However, Xie et al. (2003) were unable to find this epimerase in plants but discovered two independent pathways involved in the synthesis of *trans*- and *cis*-catechins. One pathway (the *cis*-catechins synthesis pathway) converts leucoanthocyanidin into anthocyanidin via anthocyanidin synthase (ANS) activity and then to EC and EGC by anthocyanidin reductase (ANR) activity (Xie et al., 2004). The second pathway (the *trans*-catechins synthesis pathway) converts leucoanthocyanidin into C and GC by the action of the monomeric enzyme leucoanthocyanidin reductase (LAR) (Furukawa et al., 2002; Tanner et al., 2003). To date, the gene encoding LAR has been characterized in tea and in many other plants (Bogs et al., 2005; Henry-Kirk et al., 2012; Thill et al., 2012), and the gene encoding

ANS was first characterized in mutant *Zea mays* L. (Menssen et al., 1990). However, investigations involving ANS have mostly focused on flower color (Nakatsuka et al., 2005). The gene encoding ANR was termed *BANYLUS* and isolated from *Arabidopsis thaliana* (L.) Heynh. (Xie et al, 2003) and *Medicago truncatula* Gaertn (Xie et al, 2004). Additionally, the conversion of anthocyanidin into EC and EGC by ANR was confirmed in *C. sinensis* (Punyasiri et al., 2004; Singh et al., 2009).

Caffeine biosynthesis involves three *S*-adenosyl-L-methionine (SAM)-dependent methylation steps (Suzuki, 1972) and a methyl nucleoside step (Negishi et al., 1988). Theobromine and caffeine are products of the second-to-last and last methylation steps catalyzed by *N*-methyltransferase, respectively. The protein sequences of the *N*-methyltransferases involved in caffeine biosynthesis are highly identical (>80%) in three *Camellia* species (Yoneyama et al., 2006). Only a few studies have investigated the possible substrate specificities of these enzymes (Kato and Mizuno, 2004). For example, Kato et al. (1999) isolated an *N*-methyltransferase from *C. sinensis* that catalyzed the last two steps of caffeine biosynthesis. Nevertheless, the specificities of these enzymes remain unclear.

The transcriptome of *C. sinensis* was sequenced by Shi et al. (2011), and a number of genes involved in the synthesis of catechins and purine alkaloids were identified from the data, providing much information for further studies. However, the corresponding information for *C. ptilophylla* is still lacking. Therefore, in this study we sequenced the transcriptome of *C. ptilophylla* using a high-throughput RNA-seq technique and generated 7.09 Gb of high-quality data for comprehensive analysis. A complete analysis of the dataset followed by comparison with data from *C. sinensis*, enabled us to identify genes related to the synthesis of catechins and caffeine/theobromine, and provided clues for further functional characterization to better understand the differences between cocoa tea and regular tea.

MATERIAL AND METHODS

Material collection, preparation, and sequencing

Fresh buds and tender leaves of *C. ptilophylla* were collected from the field of the Nanjing botanical garden of Mem. Sun Yat. Sen, Jiangsu, China, and were immediately frozen in liquid nitrogen, transferred to the laboratory, and stored at -80°C prior to RNA extraction.

The buds and tender leaves of *C. ptilophylla* were carefully cleaned and total RNA was extracted using TRIzol (Life Technologies, USA) and subsequently treated with DNase I (Takara, Dalian, China) following the manufacturer instructions. The RNA quality was determined on 1% agarose gel electrophoresis, and the concentration was determined using an Agilent 2100 Bioanalyzer (Agilent Technologies, USA). The mRNA was then enriched using an Oligo d(T) primer and fragmented, and cDNA synthesis was performed using random hexamer primers with a size selection of 200 bp to prepare the cDNA library. The cDNA library was then sequenced on an Illumina HiSeq 2000 platform (Illumina, San Diego, CA, USA).

De novo assembly

The raw sequence data were stored in a FASTQ format with each read being 101 bp. FASTQ reads were quality checked using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). The adaptor sequences, homopolymers, and low quality bases were trimmed/

filtered from the raw FASTQ data using SeqPrep (https://github.com/jstjohn/SeqPrep) and Sickle (https://github.com/najoshi/sickle), with over 5% unknown nucleotides and low-quality reads, and more than 20% of bases possessing a quality value ≤ 20 (an error possibility of 0.01). Reads with lengths less than 50 bp (after trimming the low quality bases/adapter sequences) were removed. To facilitate access to and use of the *C. ptilophylla* transcriptome sequencing data, the clean reads were deposited in the NCBI Sequence Read Archive (SRA) under the accession No. SRX547138. *De novo* transcriptome assembly was carried out using the assembly program Trinity (Grabherr et al., 2011) with an optimized k-mer length of 25. To reduce redundancy, clustering was performed with CD-HIT (version 4.0) (Li and Godzik, 2006), and sequences with a minimum 95% identity were merged into a single representative unigene.

Annotation of unigenes and classification

The unigenes were used as query sequences to search against the non-redundant (nr) protein database at NCBI (http://www.ncbi.nlm.nih.gov) with an E-value cut-off of $1e^{-5}$. The annotations of the best hits were recorded. A gene ontology (GO) analysis of *C. ptilophylla* (http://www.geneontology.org/) was further performed to categorize the functions of the unigenes by Blast2GO, and the unigenes were assigned to the "biological process", "cellular component," and "molecular function" subontologies Clusters of orthologous groups of proteins (COGs) and the Kyoto encyclopedia of genes and genomes (KEGG) were used to predict possible functional classifications and molecular pathways.

Sequences analysis

The LAR, ANS, ANR, and caffeine synthase (CS) sequences from *Camellia* were extracted from NCBI (http://www.ncbi.nlm.nih.gov/). The identities of the gene sequences were calculated using Geneious 4.8 (http://www.geneious.com/).

RESULTS

Transcriptome sequencing and *de novo* assembly

In order to obtain high quality transcriptome sequencing data, fresh-collected buds and tender leaves of *C. ptilophylla* were carefully washed to remove any microbes adhered to the surface, in order to reduce avoidable contamination. After total RNA was extracted, Poly A^+ RNA was isolated and reverse transcription, and cDNA library construction, and sequencing on the Illumina HiSeq 2000 platform (Illumina) were performed. The paired-end sequencing yielded 2x101-bp reads from either end of the cDNA fragment. In our study, 92,885,314 reads were generated, totaling 9.38 Gb of raw data. After eliminating the adapter sequences, ambiguous and low-quality reads, and removing short sequences (<50 bp), 70,227,953 (7.09 Gb, 75.60%) high-quality reads remained. These reads were assembled into 56,695 unique transcripts (UTs) by the Trinity program (Grabherr et al., 2011) The sequence length of the 56,695 UTs ranged from 301 to 12,021 bp, out of which, 43,063 UTs (75.96%) were ≥500 bp, 28,062 UTs (49.50%) were ≥ 1000 bp, and 11,286 UTs (19.91%) were ≥ 2000 bp, with an average length of 1317 bp and a N50 of 1877 bp (Table 1; Figure 1).

To reduce redundancy caused by alternative splicing, PCR or sequencing errors, all 56,695 UTs were further clustered by 95% identity into 44,749 unigenes using CD-HIT (Li and Godzik, 2006).

Table 1. Summary of *Camellia ptilophylla* transcriptome sequences.

Terms	Number
Total reads	92,885,314
Total bases (bp)	9,381,416,714
After cleaned	70,227,953
No. of unique transcripts	56,695
No. of unigenes	44,749
Average length (bp)	1,317
N50 (bp)	1,877

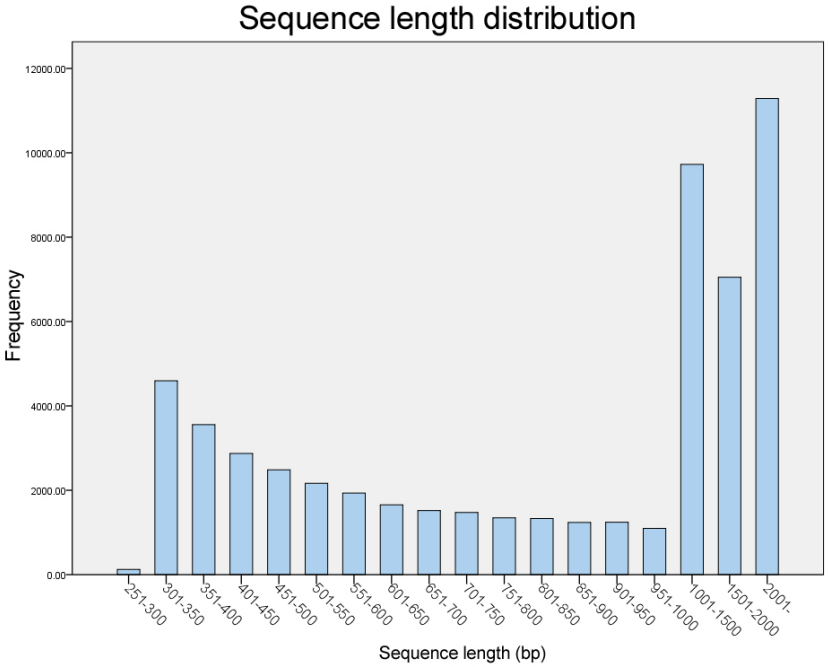

Figure 1. Length distribution of *Camellia ptilophylla* transcriptomic sequences.

Functional annotation and pathway analysis

The unigenes were subjected to BLASTX analysis at an e-value cut-off of $1.0e^{-5}$ against the NCBI Nr database. The best hit for each unigene with the highest sequence similarity from the database was chosen and the annotations were retrieved. In total, 36,345 unigenes (81.22%) had at least one match to known protein sequences in the database (**Table S1**), indicating that most of the assembled unigenes were functionally identifiable. The E-value distribution of the best hits in the Nr database revealed that 57.09% of the mapped sequences have significant homology

(less than $1.0e^{-60}$), while the other 42.91% of the homologous sequences range between $1.0e^{-5}$ and $1.0e^{-60}$ (Figure 2a). The distribution of similarity showed that 27.39% of the query sequences have a similarity higher than 80%, while 72.61% of the hit have a similarity ranging from 20 to 80% (Figure 2b). The top blast hits for the unigenes were widely distributed across many species. Among the species with the best hits, *Vitis vinifera* accounted for 48.56 (17,649), *Amygdalus persica* accounted for 9.93% (3609), *Lycopersicon esculentum* accounted for 9.09% (3303), *Ricimus communis* accounted for 7.81% (2840), and *Populus balsamifera* ssp *trichocarpa* accounted for 6.88% (2500) (Figure 2c).

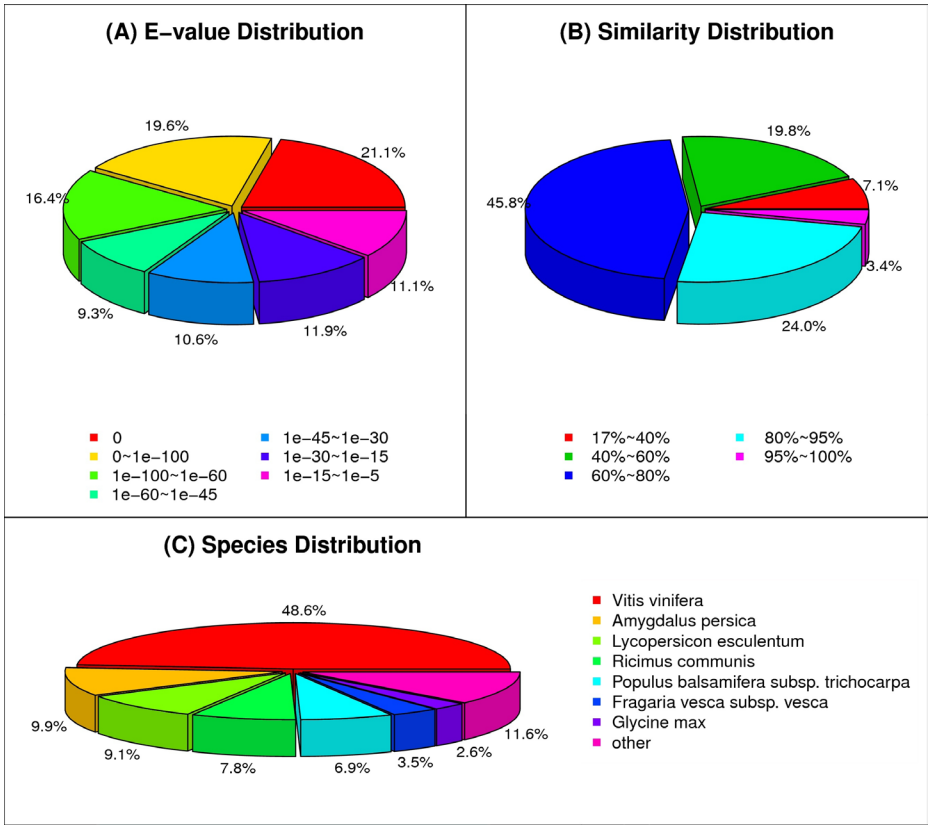

Figure 2. Characteristics of the homology search of unigenes against the non-redundant (nr) database.

GO assignments were performed to functionally categorize the annotated unigenes, which resulted in 227,924 unigenes that mapped to at least one GO term, and were classified into 55 functional subcategories (Figure 3). Of these, assignments to the “biological process” made up the majority (110,817, 48.62%) followed by “cellular component” (84,745, 37.18%) and “molecular function” (32,362, 14.20%). Under the biological process category, cellular process (17,606 unigenes, 15.89%) and metabolic process (16,948 unigenes, 15.29%) were highly represented, indicating that some important activities occur in *C. ptilophylla*.

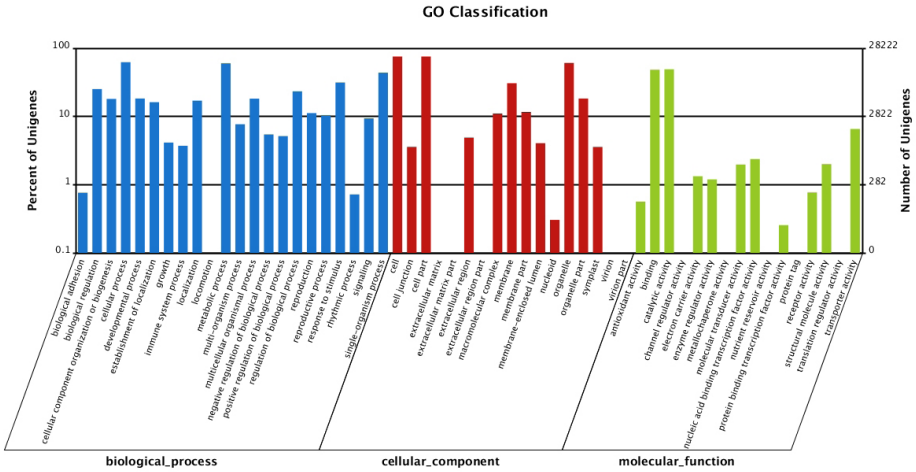

Figure 3. Gene ontology (GO) classification of *Camellia ptilophylla* transcriptome sequences.

We further compared the assembled unigenes to the COGs database for an in-depth analysis. Of the assembled sequences, 14,995 significantly matched and were subsequently grouped into 24 functional categories (Figure 4). The five largest categories included “general function prediction only” (2408 unigenes, 16.06%), “posttranslational modification, protein turnover, and chaperones” (1359 unigenes, 9.06%), “replication, recombination, and repair” (1197 unigenes, 7.98%) “transcription” (1184 unigenes, 7.90%), and “signal transduction mechanisms” (1016 unigenes, 6.78%). The “secondary metabolite biosynthesis, transport, and catabolism” category included approximately 2.52% (379 unigenes) of COG matched unigenes.

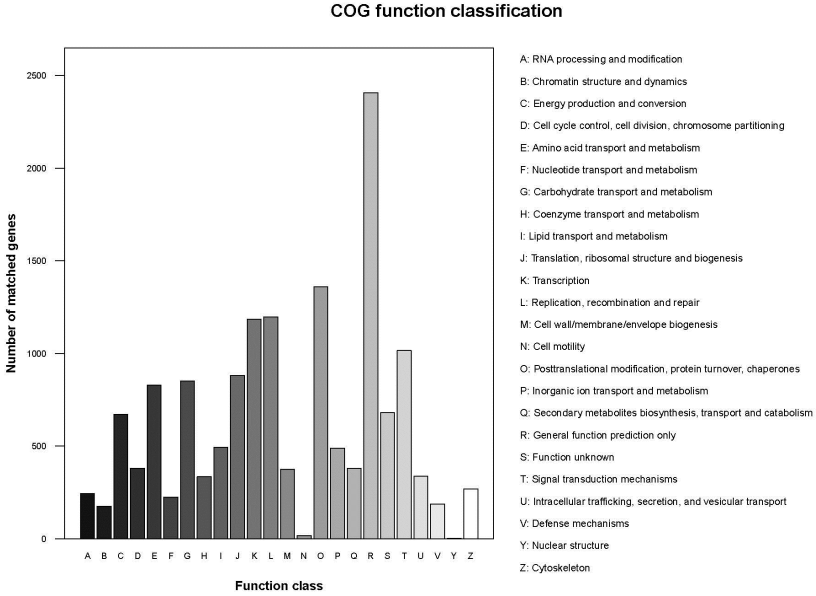

Figure 4. Clusters of orthologous groups (COG) function classification of *Camellia ptilophylla* transcriptomic sequences.

KEGG pathway assignments were then performed to provide alternative functional annotations for the enzymes of the different biochemical pathways and their enzyme commission (EC) numbers. A total of 12,408 unigenes were assigned to 324 KEGG pathways (**Table S2**). The "ribosome" category was the most prominently represented (220 unigenes, 1.61%), followed by "biosynthesis of amino acids" (200 unigenes, 1.61%), "carbon metabolism" (198 unigenes, 1.60%), and "spliceosome" (196 unigenes, 1.58%).

Sequences analyses of catechin biosynthesis-associated genes in *C. ptilophylla* and *C. sinensis*

Expression of the *LAR*, *ANS,* and *ANR* genes, which are involved in the synthesis of catechins, were detected in the transcriptomes of both *C. ptilophylla* and *C. sinensis*.

Two copies of the *LAR* genes (*CpLAR1* and *CpLAR2*) were identified form the transcriptome of *C. ptilophylla*, while only one copy of the *LAR* gene (*CsLAR*) was found in *C. sinensis* (GU992401). The nucleotide sequences of *CpLAR1* and *CpLAR2* shared only 79.9% identity because *CpLAR2* was an incomplete ORF due to sequencing issues. However, *CpLAR1* shared 99.2% sequence identity with *CsLAR*, possessing eight nucleotide substitutions.

Two intact genes (*CpANS1* and *CpANS2* sharing 82.2% sequence identity with each other) and one partial gene (*CpANS3*) were identified from our dataset. Compared with the only copy of *CsANS* from *C. sinensis* (AY830416), *CpANS1* and *CpANS2* shared 99.2% and 81.9% sequence identities with the single *CsANS* gene, respectively. Similar to other members of the 2OG-Fe(II) oxygenase superfamily, *CpANSs* in addition to those from *C. sinensis,* contain active sites that are believed to coordinate iron binding at the catalytic centers of the iron-containing soluble oxygenases and 2-oxoglutarate-dependent enzymes (Lukacin and Britsch, 1997).

Two genes encode ANR, *CpANR1* and *CpANR2*, which, in this dataset, shared only 79.5% identity. These two copies of *ANRs* were also identified in *C. sinensis*. Interestingly, *CpANR1* showed a much higher homology to *CsANR1* (GU992402) (97.3%), and *CpANR2* showed a higher homology to *CsANR2* (GU992400) (99%), suggesting that these two gene pairs have an orthologous relationship.

Sequence analysis of genes related to alkaloid biosynthesis in *C. ptilophylla* and *C. sinensis*

In our study, two genes encoding *N*-methyl transferase were identified and designated *NMT1* and *NMT2*, and were found to share 93.5% sequence identity. Compared with the caffeine synthase genes (*TCSs*) in *C. sinensis*, *TCS2* (AB031281) shared 98.6 and 92.3% identity with *NMT1* and *NMT2*, respectively, and *TCS1* (AB031280) shared 92.2 and 86.4% identity, respectively. The sequences of *NMT1* and *NMT2* differed by the presence of indels at the beginning of the coding regions, demonstrating the uniqueness of *NMT1* in *C. ptilophylla*.

DISCUSSION

Why sequence the transcriptome of *C. ptilophylla*?

Tea is considered to be one of the top three most popular drinks worldwide, because

of its fragrant, thirst-quenching, and healthcare properties. Nearly 2 billion people from over 160 countries consume about 250 thousand tons of tea every year. As an important economic crop, tea also attracts the attention of scientists who aim to study its healthcare components and mechanisms of biosynthesis. Nowadays, studies on efficient chemicals have developed far beyond studies on biosynthesis pathways. However, this is not unexpected, because knowledge of the genetic background of tea is lacking. Very large genome sizes of *Camellia* species have been reported because of frequent polyploidy events. For example, *C. sinensis* was predicted to possess a genome as large as ~3 Gb and it is possible that the genome of *C. ptilophylla* is even larger (Huang et al., 2013); therefore, the cost and time needed to sequence the whole genomes of these plants would be very high. Thus, a more economical and targeted research plan is needed to sequence the transcriptome, because this would detect the genes expressed in the leaves and young buds, which are used to make tea.

The transcriptome of *C. sinensis* was previously sequenced on an Illumina GA IIx platform using the 75-bp paired-end technique, leading to the generation of 2.32 Gb clean data and identification of genes related to flavonoid and caffeine biosynthesis pathways (Shi et al, 2011). Comparatively, here, the transcriptome of *C. ptilophylla* was sequenced on an Illumina HiSeq 2000 platform using the modified 101-bp paired-end technique. The raw data went through quality control measures, in which adapters were trimmed and short sequences were removed, to yield 70,227,953 cleaned reads. The processed data were then assembled *de novo* into 56,695 UTs with an average length of 1317 bp. The upgraded sequencing platform and assembly software helped us to obtain more abundant and reliable data and laid a firm foundation for further genetic and evolutionary analyses of *C. ptilophylla*. Moreover, when comparing our data with that obtained for *C. sinensis*, we discovered similarities and differences in critical genes that regulate the synthesis of the main chemical components, catechins and alkaloids. Additionally, regulatory mechanisms were explored and discussed, providing a basis for future study.

Genetic discrepancies underlying the different catechins components of *C. ptilophylla* and *C. sinensis*

The biosynthesis of different catechins involves complex pathways and enzymes, and most of these pathways have not been elucidated to date. Fortunately, the synthesis pathways of simple catechins, including those of C, EC, GC, and EGC, have been well studied (Rani et al., 2012). These simple catechins are synthesized from leucoanthocyanidin, which is derived from flavonoids. All of the genes encoding enzymes related to catechins biosynthesis were identified in the transcriptome of *C. ptilophylla*, suggesting the presence of a complete catechins biosynthesis pathway in this species.

There are two potential outcomes of the catechins biosynthesis pathway. The LAR protein has been shown to catalyze the conversion of leucoanthocyanidin to *trans*-catechin, while ANS and ANR catalyze the transition of *cis*-catechin (Tanner et al, 2003; Xie et al, 2004). *In vitro* expression analysis of LAR verified its activity in catalyzing leucoanthocyanidin to *trans*-catechin C (Tanner et al, 2003); additionally, the overexpression of *PtrLAR3* in *Populus tomentosa* was shown to significantly increase the amount of proanthocyanidins (Yuan et al., 2012). Therefore, compared with *C. sinensis*, which contains only one *LAR* gene, the two *LAR* genes of *C. ptilophylla* may confer an advantage due to the combined effects of this quantitative trait, which allows for the utilization of more leucoanthocyanidin to synthesize more *cis*-catechin. However, the more likely

factor responsible for the increased *cis*-catechin levels is the divergence of *CpLAR2*. Because *CpLAR1* is highly identical to *CsLAR*, these genes are probably functionally identical. As the more diverged gene, *CpLAR2*, is more likely to have evolved specific and enhanced catalytic synthesis of *trans*-catechin, accounting for the accumulation of *cis*-catechins in *C. ptilophylla*.

Three *ANS* genes, which are involved in the synthesis of *cis*-catechin, were found in the *C. ptilophylla* transcriptome, including two intact genes (*CpANS1* and *CpANS2*) and one partial gene (*CpANS3*). Compared with the two *ANS* genes observed in *C. sinensis*, one gene pair (*CpANS1* and *CsANS1*) was highly identical, while the other (*CpANS2* and *CsANS2*) was much less similar. Therefore, it is hypothesized that the diverged *CpANS2* may have lost the ability to synthesize *cis*-catechin, causing a partial, if not total, failure of this pathway. The other gene family involved in the synthesis of *cis*-catechins is *ANR*. Both *C. ptilophylla* and *C. sinensis* were shown to possess two *ANR* genes. The *ANR* genes of the two species were highly identical at the sequence level and are therefore likely to share the same function.

Therefore, the discrepancy between the *trans*- and *cis*-catechins produced by the two *Camellia* species likely involves two gene families that are associated with the synthesis pathways of LAR and ANS. Evidence suggests that the diverged genes *CpLAR2* and *CpANS2* underlie dominant activity of *trans*-catechin and decreased *cis*-catechin in *C. ptilophylla*. Moreover, the quantitative nature of the traits and enzyme kinetics may also be involved in the process.

Differences in the alkaloid components of *C. ptilophylla* and *C. sinensis*

The caffeine synthesis pathway starts at xanthosine, which is converted to 7-methylxanthosine, 7-methylxanthine, 3,7-methylxanthine (theobromine), and finally, caffeine. The entire pathway includes three methylation reactions that are catalyzed by *N*-methyltransferases (Suzuki, 1972), and one degradation reaction, which is catalyzed by *N*-methyl nucleosidase (Negishi et al, 1988). Other studies have also reported a separate process for converting 7-methylxanthine to caffeine involving the alternative intermediate 1,7-methylxanthine instead of theobromine (Suzuki, 1972; Kato et al., 1996; Ashihara et al., 1998).

Yoneyama et al. (2006) successfully cloned the *N*-methyltransferase-encoding genes from *C. sinensis*, *C. ptilophylla*, and *C. irrawadiensis*, and subsequently introduced them into *Escherichia coli* for heterologous expression. As a result, the *N*-methyltransferase encoded by *TSC1* in *C. sinensis* synthesized caffeine; however, the exact step at which this enzyme functioned could not be determined. This is because the last two steps of both pathways involve this gene product, while the gene products in *C. ptilophylla* (*PCS1*) and *C. irrawadiensis* (*ICS1*) specifically catalyze 7-methylxanthine to theobromine at the third step of one pathway. Therefore, the functional differences between TSC1 in *C. sinensis* and PCS1 and ICS1 in *C. ptilophylla* and *C. irrawadiensis* were confirmed; the former participates in the synthesis of theobromine, 1,7-methylxanthine and caffeine, while the latter participates in the synthesis of theobromine only.

Data obtained from transcriptome sequencing revealed two genes that were homologous to *PCS1* and *PCS2*, which were *NMT1* and *NMT*2. In our comparison, *NMT1* shared 86.4, 92.5, 92.8, 92.3, 98.7, and 94% sequence identity with the previously identified *TCS1*, *PCS1*, *ICS1*, *TCS2*, *PCS2,* and *ICS2* genes, respectively, from *C. sinensis*, *C. ptilophylla*, and *C. irrawadiensis* (Yoneyama et al, 2006). Similarly, *NMT2* shared 92.2, 95.4, 95.7, 98.6, 99.8, and 99% identity, respectively. *NMT2* shared 99.8% sequence identity with *PCS2*. The 0.2% sequence discrepancy was most likely due to differences between populations or individuals. Interestingly, *NMT1* also

shared the highest identity with *PCS2* (98.7%) but shared less identity with *PCS1* (92.5%), suggesting that *NMT1* is slightly more diverged than the *PCS2* duplicate in the sampled population. Our data did not capture the *PCS1* orthologous gene, possibly because of its low expression level or due to insufficient sampling. *TCS2*, *PCS2,* and *ICS2* were reported to be inactive for methylation *in vitro* (Yoneyama et al, 2006), and our sequence comparison suggests similar results with *NMT2* and *NMT1*. However, unidentified factors (i.e., protein modification, protein folding, etc.) are thought to underlie the precise activities of proteins *in vitro*. In fact, the slightly more diverged *NMT1* may undergo neofunctionalization, and more in-depth experiments should be performed to fully elucidate the deeper functional characteristics.

The possession of the artificial chimeric genes *PCS1* and *TCS2* was demonstrated to restore *E. coli*'s ability to synthesize caffeine (Yoneyama et al, 2006), indicating the possible caffeine synthesis activities of *PCS2*, *TCS2,* and *ICS2* encoded products *in vivo*, although different environments influence protein folding. These genes, in addition to *NMT1,* also require further functional characterization.

The conversion of 7-methylxanthosine to 7-methylxanthine is catalyzed by *N*-methyl nucleosidase, which was isolated from the leaves of *C. sinensis* (Negishi et al, 1988). However, the genes encoding this enzyme have not been cloned and therefore do not exist in public databases and were not analysed in the present study.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by grants from the Prospective Industry-Academia-Research project of Jiangsu Province (#BY2012211) and Institute of Botany, Jiangsu Province and Chinese Academy of Sciences (#SQ201302). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

REFERENCES

Ashihara H, Kato M and Ye CX (1998). Biosynthesis and metabolism of purine alkaloids in leaves of cocoa tea (*Camellia ptilophylla*). *J. Plant Res.* 111: 599-604.

Bogs J, Downey MO, Harvey JS, Ashton AR, et al. (2005). Proanthocyanidin synthesis and expression of genes encoding leucoanthocyanidin reductase and anthocyanidin reductase in developing grape berries and grapevine leaves. *Plant Physiol*. 139: 652-663.

Furukawa T, Eshima A, Kouya M, Takio S, et al. (2002). Coordinate expression of genes involved in catechin biosynthesis in *Polygonum hydropiper* cells. *Plant Cell Rep*. 21: 385.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol*. 29: 644-652.

Henry-Kirk RA, McGhie TK, Andre CM, Hellens RP, et al. (2012). Transcriptional analysis of apple fruit proanthocyanidin biosynthesis. *J. Exp. Bot*. 63: 5437-5450.

Huang H, Tong Y, Zhang Q and Gao L (2013). Genome size variation among and within *Camellia* species by using flow cytometric analysis. *PLoS One* e64981.

Kato M and Mizuno K (2004). Caffeine synthase and related methyltransferases in plants. *Front. Biosci*. 9: 1833-1842.

Kato M, Kanehara T, Shimizu H, Suzuki T, et al. (1996). Caffeine biosynthesis in young leaves of *Camellia sinensis*: *In vitro* studies on N-methyltransferase activity involved in the conversion of xanthosine to caffeine. *Physiol. Plant.* 98: 629-636.

Kato M, Mizuno K, Fujimura T, Iwama M, et al. (1999). Purification and characterization of caffeine synthase from tea leaves. *Plant Physiol.* 120: 579-586.

Li W and Godzik A (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658-1659.

Lukacin R and Britsch L (1997). Identification of strictly conserved histidine and arginine residues as part of the active site in Petunia hybrida flavanone 3beta-hydroxylase. *Eur. J. Biochem*. 249: 748-757.

Menssen A, Höhmann S, Martin W, Schnable PS, et al. (1990). The En/Spm transposable element of *Zea mays* contains splice sites at the termini generating a novel intron from a dSpm element in the A2 gene. *EMBO J*. 9: 3051-3057.

Nakatsuka T, Nishihara M, Mishiba K and Yamamura S (2005). Two different mutations are involved in the formation of white-flowered gentian plants. *Plant Sci.* 169: 949-958.

Negishi O, Ozawa T and Imagawa H (1988). N-Methyl Nucleosidase from Tea Leaves. *Agric. Biol. Chem.* 52: 169-175.

Peng L, Song X, Shi X, Li J, et al (2008). An improved HPLC method for simultaneous determination of phenolic compounds, purine alkaloids and theanine in *Camellia* species. *J. Food Compos. Anal.* 21: 559-563.

Punyasiri PA, Abeysinghe IS, Kumar V, Treutter D, et al. (2004). Flavonoid biosynthesis in the tea plant *Camellia sinensis*: properties of enzymes of the prominent epicatechin and catechin pathways. *Arch. Biochem. Biophys*. 431: 22-30.

Rani A, Singh K, Ahuja PS and Kumar S (2012). Molecular regulation of catechins biosynthesis in tea [*Camellia sinensis* (L.) O. Kuntze]. *Gene* 495: 205-210.

Shi CY, Yang H, Wei CL, Yu O, 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.

Singh K, Rani A, Paul A, Dutt S, et al (2009). Differential display mediated cloning of anthocyanidin reductase gene from tea (*Camellia sinensis*) and its relationship with the concentration of epicatechins. *Tree Physiol*. 29: 837-846.

Stafford HA (1990). Pathway to proanthocyanindins (condensed tannins), flavan-3-ols, and unsubstituted flavans. CRC Press, New York.

Suzuki T (1972). The participation of S-adenosylmethionine in the biosynthesis of caffeine in the tea plant. *FEBS Lett*. 24: 18-20.

Tanner GJ, Francki KT, Abrahams S, Watson JM, et al. (2003). Proanthocyanidin biosynthesis in plants. Purification of legume leucoanthocyanidin reductase and molecular cloning of its cDNA. *J. Biol. Chem*. 278: 31647-31656.

Thill J, Regos I, Farag MA, Ahmad AF, et al. (2012). Polyphenol metabolism provides a screening tool for beneficial effects of *Onobrychis viciifolia* (sainfoin). *Phytochemistry* 82: 67-80.

Xie DY, Sharma SB, Paiva NL, Ferreira D, et al. (2003). Role of anthocyanidin reductase, encoded by BANYULS in plant flavonoid biosynthesis. *Science* 299: 396-399.

Xie DY, Sharma SB and Dixon RA (2004). Anthocyanidin reductases from *Medicago truncatula* and *Arabidopsis thaliana*. *Arch. Biochem. Biophys*. 422: 91-102.

Yang XR, Ye CX, Xu JK and Jiang YM (2007). Simultaneous analysis of purine alkaloids and catechins in *Camellia sinensis*, *Camellia ptilophylla* and *Camellia assamica* var. Kucha by HPLC. *Food Chem*. 100: 1132-1136.

Yoneyama N, Morimoto H, Ye CX, Ashihara H, et al. (2006). Substrate specificity of N-methyltransferase involved in purine alkaloids synthesis is dependent upon one amino acid residue of the enzyme. *Mol. Genet. Genomics* 275: 125-135.

Yuan L, Wang L, Han Z, Jiang Y, et al. (2012). Molecular cloning and characterization of *PtrLAR3*, a gene encoding leucoanthocyanidin reductase from *Populus trichocarpa*, and its constitutive expression enhances fungal resistance in transgenic plants. *J. Exp. Bot*. 63: 2513-2524.