Research paper

Transcriptome and Microsatellite Analysis of *Camellia brevistyla*

Tse-Yen Liu,[1] Chia-Chen Wu[2,3]

【Summary】

Camellia brevistyla is a native tree species in Taiwan with economic importance for cooking oil and tea-seed oil. In this study we built the first transcriptome database of *C. brevistyla*, establishing a valuable resource for future investigations. We also developed and validated microsatellite markers for analyzing genetic relationships within the species. We obtained 65,722 unigenes and identified 25,996 simple sequence repeats (SSRs) within 47,595 unigenes and designed 8,113 SSR primers. Among these, 2,158 SSRs were longer than 18 base pairs (bps), primarily tri-nucleotide repeats, and the largest percentage (37%) were CT/TC repeats. Of the randomly selected 90 primer pairs, 49 (54.44%) were successfully amplified using a polymerase chain reaction (PCR) and then analyzed with a 1.5% agarose gel. In a gene ontology (GO) analysis, 34,009 unigenes were assigned 1 or more GO terms at level 2. In a Kyoto Encyclopedia of Genes and Genomes (KEGG) database analysis, 7,842 unigenes were annotated against the KEGG database, locating 16,109 enzyme-catalyzed sites on 141 KEGG pathways. Fatty-acid related metabolism involved 23 sites catalyzed by enzymes. We expect that this database and the SSR markers identified will contribute to ongoing molecular breeding, variety identification, gene cloning, and genetic analyses.

Key words: *Camellia brevistyla,* simple sequence repeat (SSR), transcriptome, gene annotation.

Liu TY, Wu CC. 2023. Transcriptome and microsatellite analysis of *Camellia brevistyla*. Taiwan J For Sci 38(3):233-48.

[1] Forest Protection Division, Taiwan Forestry Research Institute, 53 Nanhai Rd., Zhongzheng District, Taipei 100051, Taiwan. 林業試驗所森林保護組，100051臺北市中正區南海路53號。

[2] Silviculture Division, Taiwan Forestry Research Institute, 53 Nanhai Rd., Zhongzheng District, Taipei 100051, Taiwan. 林業試驗所育林組，100051臺北市中正區南海路53號。

[3] Corresponding author, E-mail: chiachen@tfri.gov.tw 通訊作者

Received May 2023, Accepted July 2023. 2023年5月送審2023年7月通過。

研究報告

細葉山茶轉錄體與微衛星體序列分析

劉則言[1)] 吳家禎[2,3)]

摘 要

細葉山茶(*Camellia brevistyla*)是臺灣原生樹種，種籽可以生產食用油，具有重要的經濟價值。本研究我們建立了細葉山茶的轉錄組(transcriptome)分析資料庫，並且初步測試可用的簡單重複序列(simple sequence repeat, SSR)引子。轉錄組定序，組裝後得到65,722個重組序列(unigene)，並且在47,595個重組序列中找到25,996個簡單重複序列，並且設計出8,113組聚合酶鏈鎖反應的引子對，其中簡單重複序列長度大於18 bp的有2,158組，以三核甘酸(tri-nucleotide)的重複形式，以及CT/TC重複(37%)出現最多次。實際測試90組簡單重複序列中，有49組可以成功得到擴增產物(amplicon)，成功率為54.44%。轉錄組資料庫分析中，有34,009個重組序列被註解並取得到基因本體(gene ontology, GO)分析資料，在KEGG分析資料中，總共有7,842個重組序列分析取得16,109關於代謝途徑的酵素反應位置，其中和脂肪酸代謝途徑相關的有23個酵素反應位置。本研究建立了臺灣原生細葉山茶的轉錄體資料庫，並發現多組簡單重複序列引子，預期將有助於日後細葉山茶的分子育種、品種鑑定、基因選殖與遺傳分析。

關鍵詞：細葉山茶、簡單序列重複、轉錄組、基因註解。

劉則言、吳家禎。2023。細葉山茶轉錄體與微衛星體序列分析。台灣林業科學38(3):233-48。

INTRODUCTION

Tea-seed oil (*Camellia* oil) is an edible vegetable oil extracted from the seeds of *C. oleifera*, a member of the Theaceae family. Another species, *C. brevistyla,* native to Taiwan, is used for tea-seed oil production. *C. oleifera* is grown in southern Taiwan, while *C. brevistyla* is widely cultivated in northern regions of Taiwan.

C. brevistyla is an evergreen tree that reaches up to 15 m in height. It features dark-green leaves and produces fragrant white flowers. The fruit is a small, round, green capsule containing 2~6 seeds from which oil is extracted. Tea-seed oil, recommended as a high-quality edible oil by the Food and Agricultural Organization (FAO) (Cheng et al. 2018) is renowned for its nutritional value, including over 85% unsaturated fatty acids (omega-9) (Sahari and Amooi 2013, Cheng et al. 2018) and other compounds including triterpenoid saponins, phenolic compounds, flavonoids, catechins, α-tocopherol, and squalene (Sahari and Amooi 2013, Cheng et al. 2018, Wang et al. 2019, Luan et al. 2020, Wu et al. 2020). Tea-seed oil from *C. oleifera* and *C. brevistyla* is commonly used in cooking and also has traditional applications in medicine and cosmetics, such as for relieving stomachaches, treating burn wounds, and for hair and skin care (Cheng et al. 2018). Its functional values include antioxidant and anti-inflammatory effects, modulation of the

microbiota, and alleviation of the progression of Alzheimer's disease in rats (Sahari and Amooi 2013, Cheng et al. 2015, Luan et al. 2020, Wu et al. 2020, L.Wang et al. 2021). Long-term intake of tea-seed oil may also assist in treating cardiovascular and cerebrovascular diseases, reducing cholesterol levels, and protecting the liver (Cheng et al. 2015).

Simple sequence repeats (SSRs), also known as microsatellite DNA, are a type of genetic marker that consists of short tandem repeats of nucleotide sequences, usually 2 to 6 base pairs in length, which are dispersed throughout the genome. SSRs are highly polymorphic and can be used in a wide range of applications, including population genetics, plant breeding, and plant identification (Merritt et al. 2015, Vieira et al. 2016, Wu et al. 2018). The advantage of SSR markers over other DNA markers include high polymorphism, co-dominant inheritance, and ease of use, making them a valuable tool for plant breeders and geneticists (Merritt et al. 2015). There are 2 classes of SSRs, genomic SSRs (located in non-coding genomic regions) and genic SSRs (found in expressed sequences) such as EST-tags or transcriptomes. Genic SSRs are generally more evolutionarily conserved within and across related species. Thus, genic-SSRs located in transcriptional regions that contribute to agronomic traits are easily used for gene cloning, map construction, and marker-assisted selection (MAS). However, the polymorphism of genic-SSRs is much lower than that of genomic-SSRs (Dutta et al. 2011, Zhang et al. 2014).

The transcriptome refers to the complete set of RNA transcripts produced by a cell or tissue at a particular time. Transcriptome analysis is an important tool for understanding gene expressions and regulation in plants. In recent years, RNA sequencing (RNA-Seq) has emerged as a powerful tool for transcriptome analysis in plant research for quantification of gene expression levels, novel gene discovery and molecular marker development (Dutta et al. 2011, Shi et al. 2011, Xia et al. 2014, Zhang et al. 2015, Cheng et al. 2023).

In this study, we obtained the *C. brevistyla* transcriptome, then did the analysis of SSR mining, primer validation, gene annotation, and metabolism analysis. These transcriptomic data will be informative and useful for future investigations of tea-seed oil, including genetic analysis, fatty acid metabolism, and DNA marker identification.

MATERIALS AND METHODS

Plant materials and RNA extraction

We collected roots, leaves, stems, and fruits from a 30-yr-old *C. brevistyla* tree (SMS1. decimal degrees: 24.966902, 121.590566) in Muzha, Taipei City, for RNA extraction (Chang et al. 1993). The fresh plant material was flash-frozen using liquid nitrogen and promptly stored in a refrigerator at -80 °C.

Sequencing, SSR mining, and preliminary validation

We selected the Illumina Hi-Seq 2000 (100 bp pair-end) as a transcriptome sequencing system and for library construction. Sequencing was performed by Yourgene Company (New Taipei City, Taiwan). After sequencing, we pooled all trimmed reads and adopted Velvet (Oases) which uses de Bruijn graphs for the *de novo* assembly algorithm (Zerbino and Birney 2008). We used Sputnik to search for SSRs in the obtained unigene sequences to get SSR information and SSR primer data (Cardle et al., 2000). SSR counts for mining were set to 4 for di-nucleotides, 3 for tri-nucleotides, 2 for tetra-nucleotides, and 1 for penta-nucleotides. For instance, AT”AT

AT AT AT" is an example of 4 repeat counts of a di-nucleotide. Therefore, the minimum SSR length was 10 bp for 4 repeats of a di-nucleotide and 1 repeat of a penta-nucleotide. After detecting SSRs, we used Primer 3 to design polymerase chain reaction (PCR) primers with an optimum melting temperature (60 °C) and other default settings for the identified SSRs (Untergasser et al. 2012). We preliminarily validated the designed primer (SSR length longer than 38 bp) by electrophoresis with a 1.5% (w/v) agarose gel for 4 *C. brevistyla* and 4 *C. oleifera* samples.

Unigene annotation, gene ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathway analyses

The unigenes were annotated with BLAST (blastx) against the NCBI NR database of green plants with an E-value cutoff of 10-5. Results were read into an MS- Excel file. Functional categorization using GO (http://www.geneontology.org) was carried out based on results of BLASTX matches from the plant protein dataset of NR databases using Blast2GO (Conesa et al. 2005) with an E-value cutoff of 10^{-5}. GO term distributions at level 2 for each GO type were calculated and saved as Excel tables. Pathway assignments were determined with the KEGG pathway database using the results of blastx with an E-value cutoff of 10^{-5}.

RESULTS

Sequencing data and *de novo* assembly

In total, 59,328,094 paired-reads were obtained with 5,992,137,494 base pairs. Trimming reduced the data to 58,019,588 clean reads with a total of 5,606,915,588 bp. The average length of reads was 96.5 bp (Table 1).

Clean reads were used for *de novo* assembly. In total, 65,722 unigenes with a total of 43,078,194 bp were obtained with lengths from 200 to 6,483 bp. The average unigene length was 655.46 bp with an N50 value of 852 bp (Table 2). The length distribution of unigenes is shown in Fig. 1. Numbers of unigenes with lengths exceeding 300 and 2000 bp were 50,927 and 1494, respectively, accounting for 92% and 8.7% of the total unigene bases (Table 2). Detailed sequences and names of the 65,722 unigenes are recorded in Supplementary File 1.

SSR mining, SSR primer design, and preliminary validation

In total, 25,996 SSRs were found in 16,891 (25.7%) of the 65,722 unigenes (Supplementary Table S1), for a density

Table 1. Sequencing statistics for the *Camellia brevistyla* transcriptome

Sequencing statistic	Sequencing results
Number of initial reads (paired)	59,328,094
Read length (bp)	101
Total sequencing base (bp)	5,992,137,494
Number of reads after trimming (paired)	58,748,083
Average read length after trimming (bp)	96.5
Total clean sequencing base (bp)	5,606,915,558
Number of reads after trimming and removing control reads (paired)	58,019,588

of 1 SSR/1.65 kb (25,996/43,078,194 bp). Among all mined SSRs, di-nucleotides were the most numerous with 11,772, followed by tri-nucleotides (10,098), tetra-nucleotides (1,866), and penta-nucleotides (2,260) (Fig. 2). We then designed 8,113 primer pairs using computational tools (Supplementary Table S2) with the SSR data. The number of primer pairs containing SSRs longer than 18 bp was 2158, with tri-nucleotide repeats being the predominant type (1,019) (Fig. 2). The most common repeat motif was CT/TC, accounting for the largest percentage (23%), followed by AG/GA repeats (13%) (Fig. 3). The number of SSRs is shown in Fig. 4. The highest repeat count was 3 times for lengths of 12, 16 and 20 bp in tri-, tetra- and penta-nucleotides, respectively. The highest frequency of repeat counts was for di-nucleotide (TC) with 39 repeat counts for 80 bp. These high frequency

Table 2. Statistics for *Camellia brevistyla de novo* assembly

Statistics of *de novo* assembly	Value
Number of unigenes	65,722
Total base of all unigenes (bp)	43,078,194
Longest length among unigenes (bp)	6,483
Shortest length among unigenes (bp)	200
Average length of unigenes (bp)	655
N50 (bp)	852
Total bases with unigene length over 300 bp (bp)	39,419,917
Number of unigenes with length over 300 bp	50,927
The percentage of unigenes with length over 300 bp (%)	0.920
Total bases with unigene length over 2000 bp (bp)	3,768,084
Number of unigenes with length over 2000 bp	1,494
The percentage of unigenes with length over 2000 bp (%)	0.087

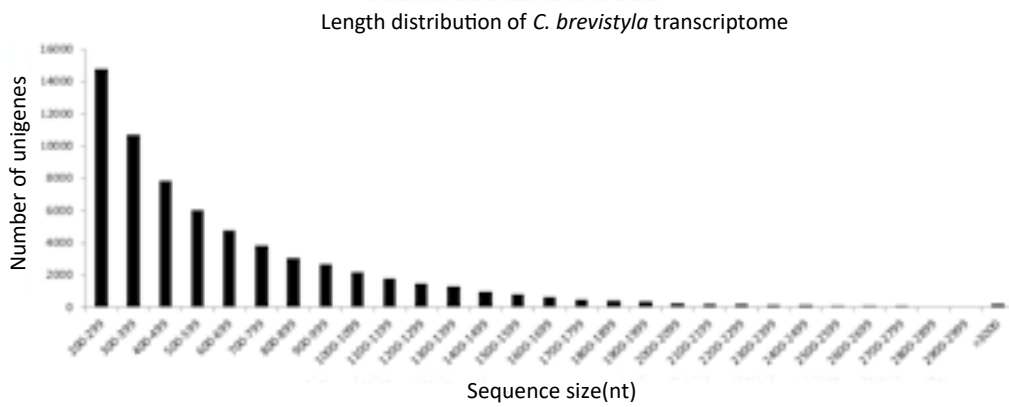

Fig. 1. Unigene length distribution of the *Camellia brevistyla* transcriptome.

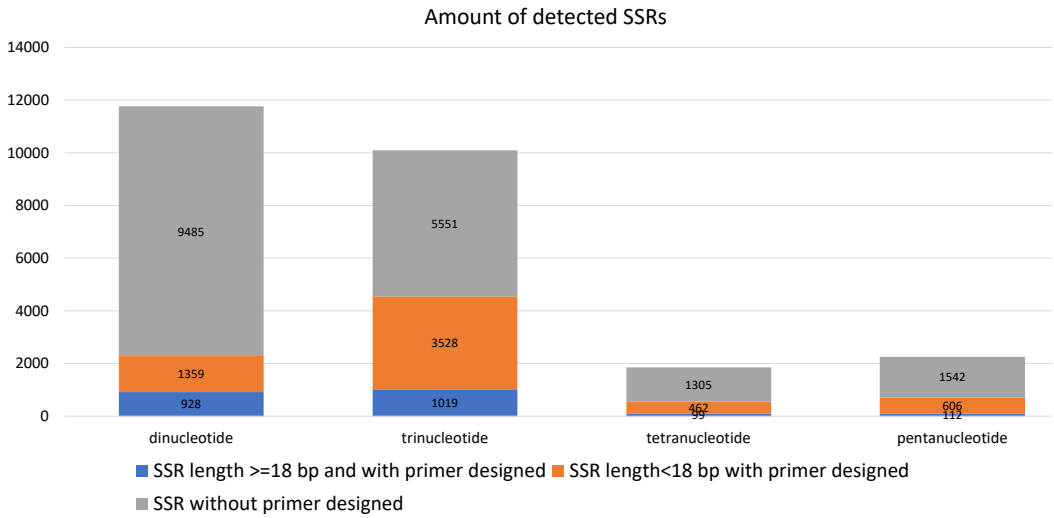

Fig. 2. Bar plot showing the amount of detected simple sequence repeats (SSRs) in *the Camellia brevistyla* transcriptome unigene.

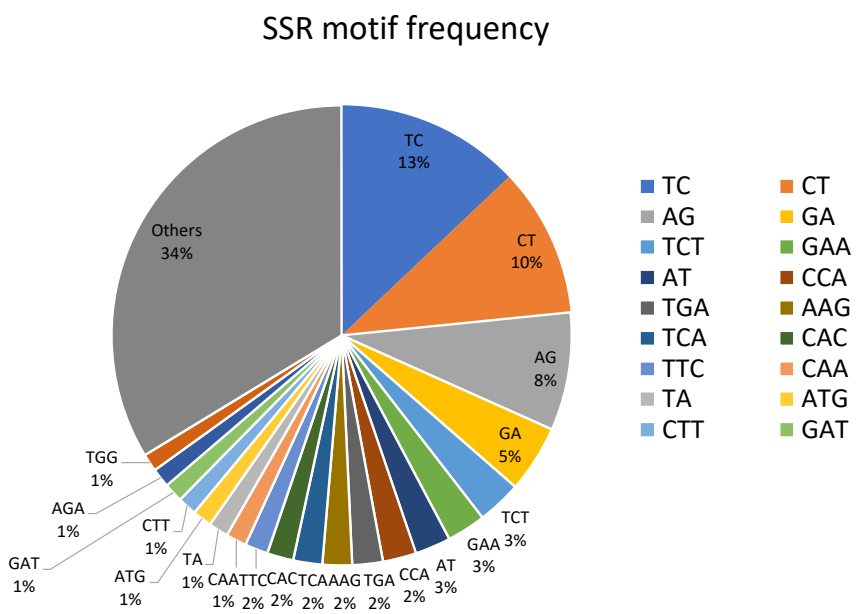

Fig. 3. Frequency of detected simple sequence repeats (SSRs) in the *Camellia brevistyla* transcriptome unigene. This pie chart shows SSR frequencies with primer designed and with SSR lengths of ≥ 18 bp.

data (66.11%) were distributed over 2 to 6 repeat counts.

We conducted a preliminary random test of 90 primer pairs with SSR lengths longer than 38 bp. There were 49 primer pairs which could be successfully (54.44%) amplified by the PCR, and 31 of these 49 primer pairs showed single bands with size-matches on electrophoresis using a 1.5% agarose gel. The results of electrophoresis showed 11 primer pairs with PCR amplicons with longer than the expected size. Five primer pairs had multiple amplicons and 2 primer pairs displayed PCR amplicons shorter than the expected size (Fig. 5). The 90 primer pairs tested in this article are listed in Supplementary Table S3.

Functional, KEGG map annotation, and GO classification

In total, 65,722 unigenes were annotated using the BLASTX protein database. Homologous genes came from several species, with 36% of the unigenes having the highest homology with genes from *Vitis vinifera*, followed by *Populus trichocarpa* (10%), *Ricinus communis* (9%), *Glycine max* (4%), *Arabidopsis* spp. (2%), and *Camellia* spp. (1%). Additionally, 27% of the unigenes did not find a match in the NR plant protein database (no hit found) (Fig. 6).

In the course of the GO analysis, 34,009 unigenes were assigned one or more GO terms at level 2. The results yielded 177,720 GO terms, including 87,807 for biological processes, 46,271 for molecular functions, and 43,642 for cellular components. Among the assigned GO terms, the majority were distributed across 20 subcategories of biological processes, followed by 11 subcategories of molecular functions and 9 subcategories of cellular components. The GO analysis showed that for biological processes, unigenes involved in metabolic processes (GO:0008152) (25%) and cellular processes (GO:0009987) (24%) were highly represented. For molecular functions, binding (GO:0005488) (41%) was most highly represented followed by catalytic activity (GO:0003824) (41%). For cellular components, cell (GO:0005623) (38%) was the most highly represented (Fig. 7).

The KEGG database analysis yielded annotations for 7842 unigenes, encompassing 16,109 enzyme-catalyzed sites across 141 KEGG pathways. Most of the pathways rep-

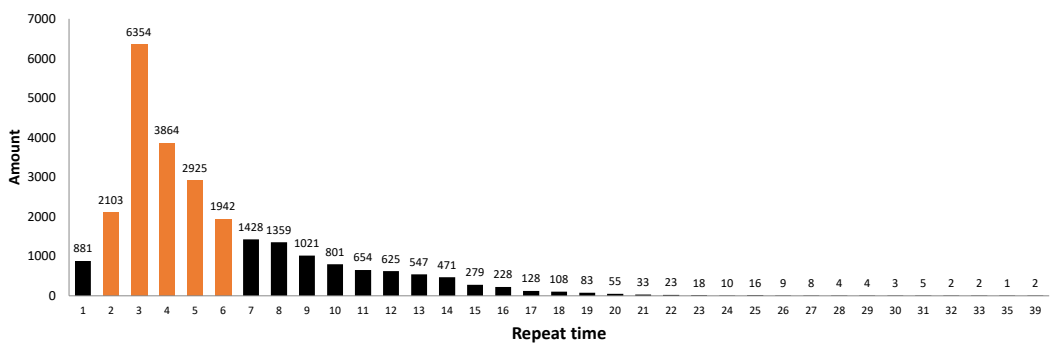

Fig. 4. Distribution of simple sequence repeats (SSRs) in the *Camellia brevistyla* transcriptome. This chart shows all detected SSRs in unigenes. Orange bars represent the majority of SSRs occurring among detected SSRs.

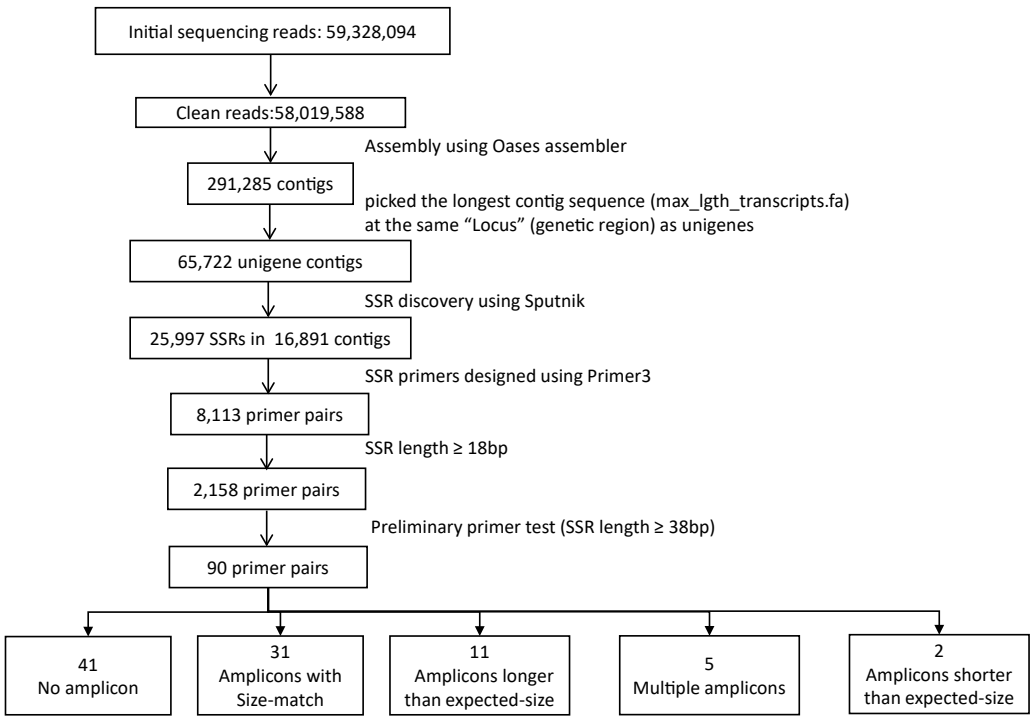

Fig. 5. Work flow of the study and results of preliminary simple sequence repeat (SSR) primer validation.

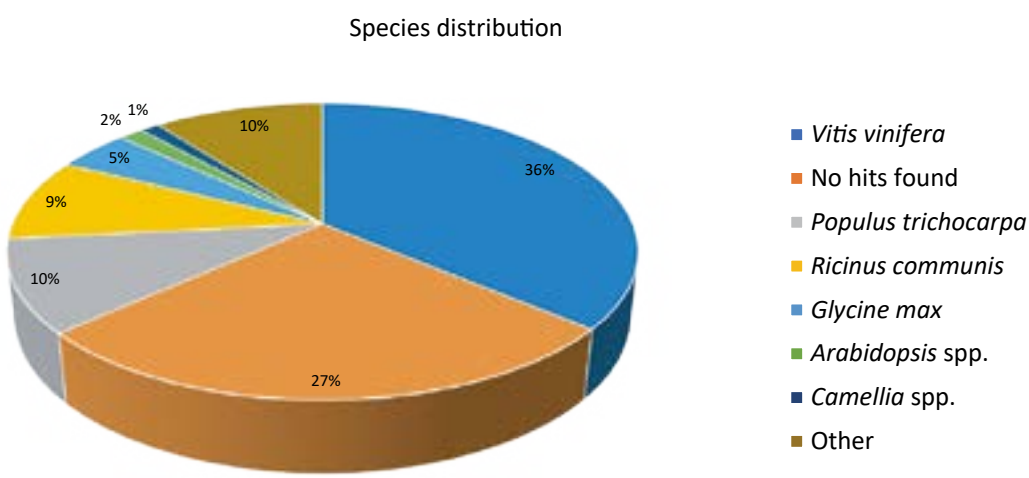

Fig. 6. Species distribution from BLASTx annotation of the *Camellia brevistyla* transcriptome.

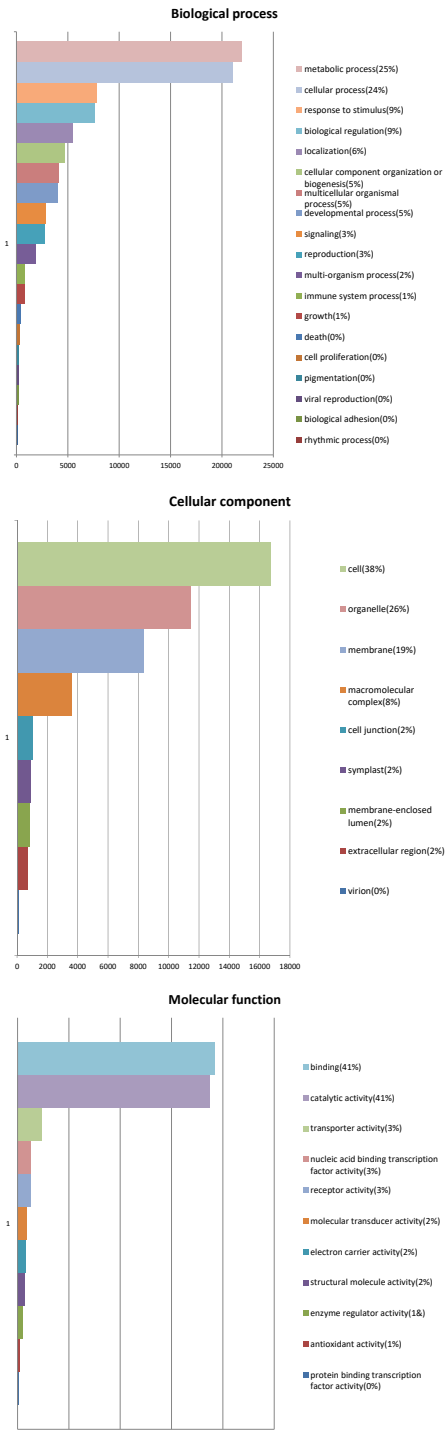

Fig. 7. Annotation results for the *Camellia brevistyla* transcriptome at level 2 gene ontology (GO) terms.

resented by unique sequences were related to purine metabolism (880 unigenes). The starch and sucrose metabolism T cell receptor signaling pathway and pyrimidine metabolism were also significantly represented. Notably 180 unigenes were related to fatty acid metabolism which may be strongly related to the synthesis of tea-seed oil in *C. brevistyla* seeds (Fig. 8). Totals of 106, 99, and 48 unigenes were respectively related to fatty acid biosynthesis, biosynthesis of unsaturated fatty acids, and fatty acid elongation. In our results, 23 enzymes were annotated and involved in fatty acid-related metabolism (Table 3). The number of unigenes assigned to KEGG pathways is listed in Supplementary Table S4, and unigenes annotated to fatty acid metabolism are listed in Supplementary Tables S5-S8. Results of BLASTx analysis are in Supplementary Table S9.

DISCUSSION

Characterization of *C. brevistyla* transcriptome and the unigene discovery

Next-generation sequencing brings advantages of time-saving, lower costs, and higher informatic throughput for gene discovery of non-modeled organisms (Dai et al., 2015, Wu et al. 2017). In this study, 65,722 unigenes were obtained with a total length of 5,606,915,558 bp and N50 of 852 bp (Table 1). The number of assembly unigenes and the length of N50 were greater than from a previous report on *C. oleifera* transcriptomes (Xia et al. 2014), while they were less than results of Tai et al.'s and Wu et al.'s reports (Tai et al. 2015, Wu et al. 2022), due to sequencing reads and sample collection. In our study, we mixed leaves, stems, seeds, and roots in equal-molar ratios for sequencing library construction; however, this approach may introduce bias in the sequencing and read

assembly for specific expressing genes. The sequencing depth and throughput are relative to the quality of assembly contigs. In general, the higher the accuracy of assembly contigs the more sequencing throughput is necessary. However, this also increases the sequencing cost. The total number of bases of unigenes we obtained was similar to *C. sinensis* transcriptomes (Shi et al. 2011). Thus, to the best of our knowledge, an initial transcriptome database construction for a non-model plant based on 5~8 giga-bp of sequencing throughput for each test sample is widely accepted by researchers with budget considerations (Tai et al. 2015, Zhang et al. 2015, Wu et al. 2022). In the annotation results, the most matched species was *V. vinifera*, coinciding with *C. sinensis* and *C. oleifera*. Matching species annotation with higher percentages for *P. trichocarpa, R. communis,* and G. max were also similar to other *Camellia* spp. (Shi et al. 2011, Xia et al. 2014). The annotation results for *C. brevistyla* demonstrated similarities with other *Camellia* species in previous reports before genome data were reported. Today, decreasing costs and continuing development of sequencing technology have resulted in the generation of numerous transcriptomes of *Camellia* species with long-length sequencing, and a draft genome of *C. sinensis* was reported (Wang et al. 2021, Ma et al. 2022, Liu et al. 2023). To explore molecular aspects of *Camellia* species in the future, more genomic data would be useful as references for further research. However, the *C. brevistyla* transcriptome is insufficient, as there is only 1

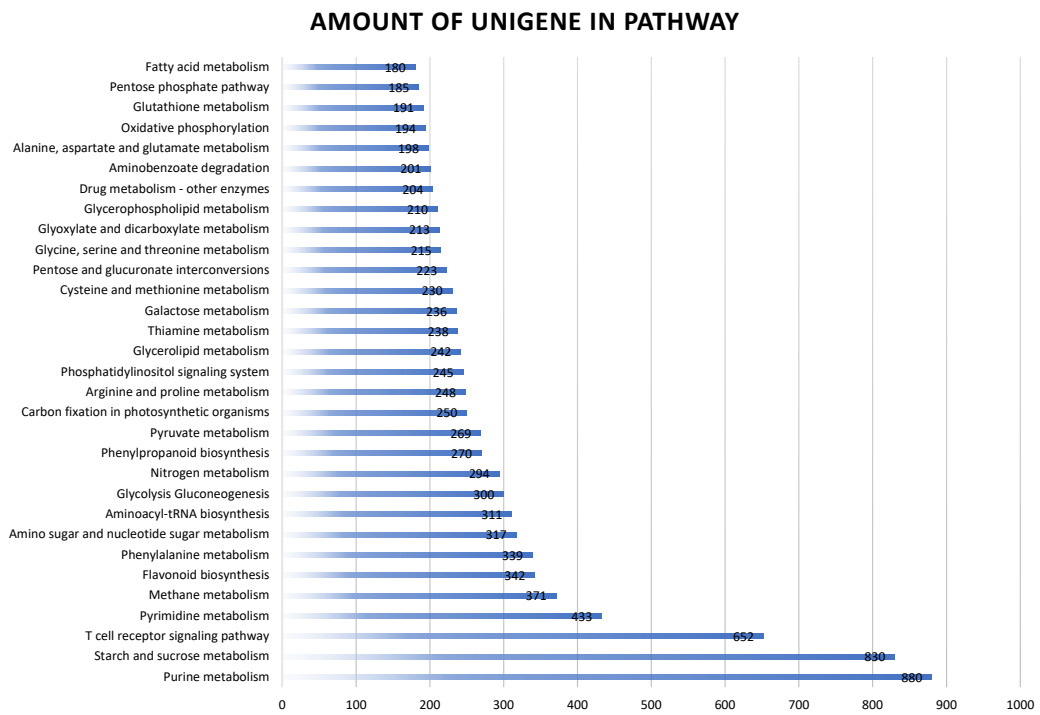

Fig. 8. Amount of unigenes of *Camellia brevistyla* transcriptome in KEGG pathways.

transcriptome available with similar sequencing throughput as our present report (Wu et al. 2022). Therefore, our sequencing data are important and informative for further genetic analyses of *C. brevistyla*. In addition, a population genetic analysis was conducted using inter-simple sequence repeats (ISSRs), revealing that gene flow of *C. brevistyla* among regions was limited (Su et al. 2017). In this study, SSR markers were co-dominant and different from ISSR markers. ISSR markers are dominant primers and can only be used for studying genetic diversity. On the other hand, SSR markers are co-dominant and highly variable, capable of displaying genetic information from both parents and are suitable for breeding and genetic research.

SSR mining and preliminary validation

With the progress of sequencing technol-

Table 3. 23 enzyme-catalyzed sites associated with fatty acid biosynthesis predicted in the *Camellia brevistyla* transcriptome

Pathway in KEGG (amount of unigene)	KEGG EC number	Enzyme
Fatty acid metabolism (180)	6.2.1.3	Long-chain acyl-CoA synthetase
	1.3.3.6	acyl-CoA oxidase
	4.2.1.17	enoyl-CoA hydratase
	1.1.1.35	3s-hydroxyacyl-CoA dehydrogenase
	2.3.1.16	ketoacyl-CoA thiolase
	2.3.1.9	acetyl-CoA C-acetyltransferase
	6.2.2.20	long chain fatty acid ligase
	5.3.3.8	enoyl-CoA hydratase
	5.1.2.3	3-hydroxybutyryl-CoA epimerase
	1.1.1.1	alcohol dehydrogenase
	1.14.15.3	alkane 1-monooxygenase
Fatty acid biosynthesis (106)	6.4.1.2	acetyl-CoA carboxylase
	2.3.1.41	3-oxoacyl-[acyl-carrier-protein] synthase I
	2.3.1.180	acetoacetyl-[acyl-carrier protein] synthase
	2.3.1.100	3-oxoacyl-[acyl-carrier protein] reductase
	2.3.1.39	malonyl-CoA-ACP transacylase
Biosynthesis of unsaturated fatty acids (99)	1.14.19.2	stearoyl-CoA desaturase
	2.3.1.16	ketoacyl-CoA thiolase
	1.3.3.6	acyl-CoA oxidase
	3.1.2.2	acyl-coenzyme A thioesterase
Fatty acid elongation (48)	2.3.1.16	ketoacyl-CoA thiolase
	4.2.1.17	enoyl-CoA hydratase
	3.1.2.22	palmitoyl-protein thioesterase

ogy, development of molecular markers has become easier than 2 decades ago. Especially for SSR markers, in the past, researchers had to use time-consuming methods, such as enrichment methods for SSR mining (Hamarsheh and Amro 2011, Vieira et al. 2016). SSR markers are highly informative and widely applied in diverse biological research areas such as taxonomical, phylogenetic, evolutionary and breeding studies (Hamarsheh and Amro 2011, Liu et al. 2012, Cochard et al. 2015, Merritt et al. 2015). For the purposes of SSR mining and primer design, time-saving and efficient high-throughput sequencing technology was used in this study. In total, 25,996 SSRs were found in 16,891 (25.7%) of the 65,722 unigenes (Supplementary Table S1), with a density of 1 SSR/1.65 kb. The SSR density we reported with a density of 1 SSR/1.65 kb. The SSR density coincides with that for *C. sinensis* EST-SSR (1 SSR/1.61 kb) (Sahu et al. 2012). Except for mono-nucleotide SSR, we found that di-nucleotide SSRs were dominant in *C. brevistyla*, this coincides with most *Camellia* species (Sahu et al. 2012, Xia et al. 2014, Wu et al. 2022). Among detected SSR sites, the most frequently occurring motif in the di-nucleotide was TC/AG (36%) with the same results for *C. oleifera* and *C. sinensis* (Xia et al. 2014, Ma et al. 2022).

In this study, 90 SSR markers were preliminarily validated, and 49 SSR primer pairs (54.4%) successfully yielded PCR amplicons. This percentage was lower than previously reported ratios of 60%-92.2% amplifications (Zhang et al. 2012). Because the 41 SSR primer sets have no amplicons, the location of the primers may be across splice sites, large introns, or poor-quality sequences (Zhang et al. 2014). These 49 SSR primer sets with successful amplification from this study are recommended for further genetic experiments. For further selection of SSR primer pairs, avoidance of di-nucleotide SSRs could decrease genotyping stuttering; longer SSR lengths may produce more polymorphism because of DNA replication slippage. However, choosing the suitable SSR markers is important not randomly choose, for a coarser scale with more distantly diverged species or taxa identification, lower SSR repeat times or interrupted repeats are suitable. (Merritt et al. 2015, Wu et al. 2018).

Fatty acid biosynthesis in *C. brevistyla*

C. brevistyla is famous for its edible seed-oil and high nutritional value. The primary composition of fatty acids in *C. brevistyla* and *C. oleifera* seeds is omega-9, which differs from the seed-oil of *C. sinensis* (Zhang et al. 2007, Liang et al. 2017). Overall fatty acid biosynthesis pathways are well studied in eukaryotes. Two enzyme systems of acetyl-CoA carboxylase (ACC), and fatty acid synthase (FAS) complex are involved in the biosynthesis of fatty acids, from the beginning of acetyl-CoA up to a chain length of C16 or C18 (Xia et al. 2014). Four fatty acid-relative pathways were annotated in our KEGG pathway analysis: fatty acid metabolism, fatty acid biosynthesis, biosynthesis of unsaturated fatty acids, and fatty acid elongation. In these four pathways, 23 enzymes were annotated with unigenes (Table 3). These enzymes play important roles in fatty acid synthesis. Fatty acid biosynthesis begins with acetyl-CoA, which is initially catalyzed by ACC (EC: 6.4.1.2) to form malonyl-CoA. Then, malonyl-ACP is produced by malonyl-CoA ACP transacylase (MCMT, EC: 2.3.1.39). Acetoacetyl-acyl-carrier protein synthase (KAS III, EC: 2.3.1.180), 3-oxoacyl-acyl-carrier protein reductase (KAR, EC: 1.1.1.100), 3R-hydroxyacyl-ACP dehydrase (HAD, EC: 4.2.1.17), and enoyl-ACP reductase (EAR, EC: 1.3.1.9) are

involved in the elongation process. Beta-ketoactyl-ACP synthase I (KAS I, EC: 2.3.1.41) is involved in further elongation. Stearoyl-CoA desaturase (AAD, EC: 1.14.19.2) was identified for the synthesis of unsaturated fatty acids from the acyl group esterified to ACP. When the acyl group was removed from the ACP by acyl-ACP thioesterase (OAH, EC: 3.1.2.14), or when acyl-ACP released the free fatty acid, the elongation of fatty acids stopped. Interestingly, FAD2 (delta(12)-fatty-acid desaturase, EC: 1.4.19.6) which desaturates oleic acid (C18:1) to generate linoleic acid (C18:2) was not annotated in our results. This seems to infer lower expression of FAD2 in *C. brevistyla*, which contains oleic acid rather than linoleic acid (Zhang et al., 2007, Xia et al. 2014, Gong et al. 2020). Thus, *C. brevistyla* transcriptome data cover most of the enzymes relative to fatty acid biosynthesis, including elongation and metabolism. This infers that this transcriptome data are useful and accurate for further research of fatty acid biosynthesis in *C. brevistyla*, such as gene cloning and differential gene expression studies.

CONCLUSIONS

From this research, we provide analyzed transcriptome data of *C. brevistyla*, including 65,722 unigenes obtained using *de novo* assembly and 8113 SSR primer pairs detected. Among these, 49 SSR primer pairs were initially validated and recommended, and 23 enzyme-catalyzed sites associated with fatty acid biosynthesis were predicted with relative unigene sequences in our results. As far as we know, there are fewer data on tea-seed oil species, especially the *C. brevistyla* transcriptome than for *C. oleifera*. So these data can support functional gene cloning, population genetics, molecular taxonomical identification, and breeding research in the future.

Data access statement

All data are deposited in the open access DataOne database, named Transcriptomes of *Camellia brevistyla*. Supplementary files may be downloaded from the link: https://metacat.tfri.gov.tw/tfri/view/urn%3Auuid%3Ac746c6e7-2291-44e3-890b-62b8fc6015d6 (or search keyword "Camellia brevistyla" at the web site of https://metacat.tfri.gov.tw/tfri)

The 2 files below are available:

1. Sequence of the transcriptome unigenes
2. Supplementary table S1-S8

LITERATURE CITED

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R. 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156(2):847-54.

Chang S, Puryear J, Cairney J. 1993. A simple and efficient method for isolating RNA from pine trees. Plant Mol. Biol. Report. 11: 113-6. https://doi.org/10.1007/BF02670468.

Cheng L, Li M, Wang Y, Han Q, Hao Y, Qiao Z, et al. 2023. Transcriptome-based variations effectively untangling the intraspecific relationships and selection signals in Xinyang Maojian tea population. Front. Plant Sci 14:1114284. https://doi.org/10.3389/fpls.2023.1114284.

Cheng X, Yang T, Wang Y, Zhou B, Yan L, Teng L, et al. 2018. New method for effective identification of adulterated *Camellia* oil basing on *Camellia oleifera*-specific DNA. Arab. J. Chem. 11(6):815-26. https://doi.org/10.1016/j.arabjc.2017.12.025.

Cheng YT, Lu CC, Yen GC. 2015. Beneficial effects of *Camellia* oil (*Camellia oleifera* Abel.) on hepatoprotective and gastropro-

tective activities. J. Nutr. Sci. Vitaminol. (Tokyo). 61:S1002. https://doi.org/10.3177/jnsv.61.S100.

Cochard B, Carrasco-Lacombe C, Pomiès V, Dufayard JF, Suryana E, Omorè A, Durand-Gasselin T, Tisnè S. 2015. Pedigree-based linkage map in two genetic groups of oil palm. Tree Genet. Genomes 11:68. https://doi.org/10.1007/s11295-015-0893-7.

Dai F, Tang C, Wang Z, Luo G, He L, Yao L. 2015. De novo assembly, gene annotation, and marker development of mulberry (*Morus atropurpurea*) transcriptome. Tree Genet. Genomes 11(2):26. https://doi.org/10.1007/s11295-015-0851-4.

Dutta S, Kumawat G, Singh BP, Gupta DK, Singh S, Dogra V, Gaikwad K, et al. 2011. Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* (L.) Millspaugh]. BMC Plant Biol. 11:17. https://doi.org/10.1186/1471-2229-11-17.

Gong W, Song Q, Ji K, Gong S, Wang L, Chen L, et al. 2020. Full-length transcriptome from *Camellia oleifera* Seed provides insight into the transcript variants involved in oil biosynthesis. J. Agric. Food Chem. 68(49):14670-83. https://doi.org/10.1021/acs.jafc.0c05381.

Hamarsheh O, Amro A. 2011. Characterization of simple sequence repeats (SSRs) from *Phlebotomus papatasi* (Diptera: Psychodidae) expressed sequence tags (ESTs). Parasites and Vectors 4:189. https://doi.org/10.1186/1756-3305-4-189.

Liang H, Hao BQ, Chen GC, Ye H, Ma J. 2017. *Camellia* as an oilseed crop. HortScience 52:488-97. https://doi.org/10.21273/HORTSCI11570-16.

Liu H, Liu Q, Chen Y, Zhu Y, Zhou X, Li B. 2023. Full-length transcriptome sequencing provides insights into flavonoid biosynthesis in *Camellia nitidissima* Petals. Gene 850:146924. https://doi.org/10.1016/j.gene.2022.146924.

Liu M, Qiao G, Jiang J, Yang H, Xie L, Xie J, Zhuo R. 2012. Transcriptome sequencing and de novo analysis for Ma bamboo (*Dendrocalamus latiflorus* Munro) using the illumina platform. PLoS One 7(10):e46766. https://doi.org/10.1371/journal.pone.0046766.

Luan F, Zeng J, Yang Y, He X, Wang B, Gao Y, Zeng N. 2020. Recent advances in *Camellia oleifera* Abel: A review of nutritional constituents, biofunctional properties, and potential industrial applications. J. Funct. Foods 75:104242. https://doi.org/10.1016/j.jff.2020.104242.

Ma D, Fang J, Ding Q, Wei L, Li Y, Zhang L, Zhang X. 2022. A survey of transcriptome complexity using full-length isoform sequencing in the tea plant *Camellia sinensis.* Mol. Genet. Genomics 297:1243-55. https://doi.org/10.1007/s00438-022-01913-2.

Merritt BJ, Culley TM, Avanesyan A, Stokes R, Brzyski J. 2015a. An empirical review: Characteristics of plant microsatellite markers that confer higher levels of genetic variation. Appl. Plant Sci 3(8):1500025. https://doi.org/10.3732/apps.1500025.

Sahari M, Amooi M. 2013. Tea seed oil: Extraction, compositions, applications, functional and antioxidant properties. Acad. J. Med. Plants 1:68-79.

Sahu J, Sarmah R, Dehury B, Sarma K, Sahoo S, Sahu M, Barooah M, Modi MK, Sen P. 2012. Mining for SSRs and FDMs from expressed sequence tags of *Camellia sinensis*. Bioinformation 8(6):260-6. https://doi.org/10.6026/97320630008260.

Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, Sun J, et al. 2011. Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-

specific compounds. BMC Genomics 12:131. https://doi.org/10.1186/1471-2164-12-131.

Su MH, Hsu TH, Wang CN, Lin KH, Chiang MC, Kang RD, et al. 2017. Genetic diversity of a novel oil crop, *Camellia brevistyla*, revealed by ISSR DNA markers. Korean J. Hortic. Sci. Technol. 35(5):588-98. https://doi.org/10.12972/kjhst.20170063.

Tai Y, Wei C, Yang H, Zhang L, Chen Q, Deng W, et al. 2015b. Transcriptomic and phytochemical analysis of the biosynthesis of characteristic constituents in tea (*Camellia sinensis*) compared with oil tea (*Camellia oleifera*). BMC Plant Biol. 15:190. https://doi.org/10.1186/s12870-015-0574-6.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3 - new capabilities and interfaces. Nucleic Acids Res. 40(15):e115. https://doi.org/10.1093/nar/gks596.

Vieira MLC, Santini L, Diniz AL, Munhoz C de F. 2016. Microsatellite markers: What they mean and why they are so useful. Genet. Mol. Biol. 39(3):312-28. https://doi.org/10.1590/1678-4685-GMB-2016-0027.

Wang F, Chen Z, Pei H, Guo Z, Wen D, Liu R, Song B. 2021. Transcriptome profiling analysis of tea plant (*Camellia sinensis*) using Oxford Nanopore long-read RNA-Seq technology. Gene 769:145247. https://doi.org/10.1016/j.gene.2020.145247.

Wang L, Ahmad S, Wang X, Li H, Luo Y. 2021. Comparison of Antioxidant and Antibacterial Activities of *Camellia* Oil From Hainan With *Camellia* Oil From Guangxi, Olive Oil, and Peanut Oil. Front. Nutr. 8:667744. https://doi.org/10.3389/fnut.2021.667744.

Wang RY, Tung YT, Chen SY, Lee YL, Yen GC. 2019. Protective effects of camellia oil (*Camellia brevistyla*) against indomethacin-induced gastrointestinal mucosal damage *in vitro* and *in vivo*. J. Funct. Foods 62:103539. https://doi.org/10.1016/j.jff.2019.103539.

Wu CC, Chu FH, Ho CK, Sung CH, Chang SH. 2017. Comparative analysis of the complete chloroplast genomic sequence and chemical components of *Cinnamomum micranthum* and *Cinnamomum kanehirae*. Holzforschung 71(3):189-97. https://doi.org/10.1515/hf-2016-0133.

Wu CC, Chu FH, Ho CK, Chang JM, Chang SH. 2018. Development of simple sequence repeat markers in *cinnamomum kanehirae hayata* using illumina-based sequencing. Taiwan J. For. Sci. 33(3):197-211.

Wu CC, Tung YT, Chen SY, Lee WT, Lin HT, Yen GC. 2020. Anti-inflammatory, antioxidant, and microbiota-modulating effects of camellia oil from *Camellia brevistyla* on acetic acid-induced colitis in rats. Antioxidants 9(1):58. https://doi.org/10.3390/antiox9010058.

Wu Q, Tong W, Zhao H, Ge R, Li R, Huang J, et al. 2022. Comparative transcriptomic analysis unveils the deep phylogeny and secondary metabolite evolution of 116 *Camellia* plants. Plant J. 111(2):406-21. https://doi.org/10.1111/tpj.15799.

Xia EH, Jiang JJ, Huang H, Zhang LP, Zhang HBin, Gao LZ. 2014. Transcriptome analysis of the oil-rich tea plant, *Camellia oleifera,* reveals candidate genes related to lipid metabolism. PLoS One 9(8):e104150. https://doi.org/10.1371/journal.pone.0104150.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821-9. https://doi.org/10.1101/gr.074492.107.

Zhang HBin, Xia EH, Huang H, Jiang JJ, Liu BY, Gao LZ. 2015. De novo transcriptome assembly of the wild relative of tea tree (*Camellia taliensis*) and comparative analysis with tea transcriptome identified putative genes associated with tea quality and stress response. BMC Genomics 16:298. https://doi.

org/10.1186/s12864-015-1494-4.

Zhang D, Tan X, Chen H. 2007. Characteristics and molecular genetics of lipid biosynthesis in tea-oil tree seed, Seed Science and Biotechnology.

Zhang H, Wei L, Miao H, Zhang T, Wang C. 2012. Development and validation of genic-SSR markers in sesame by RNA-seq. BMC Genomics 13:316. https://doi.org/10.1186/1471-2164-13-316.

Zhang W, Tian D, Huang X, Xu Y, Mo H, Liu Y, et al. 2014. Characterization of flower-bud transcriptome and development of genic SSR markers in Asian lotus (*Nelumbo nucifera* Gaertn.). PLoS One 9(11):e112223. https://doi.org/10.1371/journal.pone.0112223.