Horticulture Research

Horticulture Research, 2022, **9**: uhab083
https://doi.org/10.1093/hr/uhab083

Article

The reference genome of *Camellia chekiangoleosa* provides insights into *Camellia* evolution and tea oil biosynthesis

Teng-fei Shen[1,†], Bin Huang[2,†], Meng Xu[1,†], Peng-yan Zhou[1], Zhou-xian Ni[1], Chun Gong[2], Qiang Wen[2,*], Fu-liang Cao[1,*] and Li-An Xu[1,*]

[1]Co-Innovation Center for Sustainable Forestry in Southern China, Key Laboratory of Forest Genetics and Biotechnology Ministry of Education, Nanjing Forestry University, Nanjing 210037, China
[2]Jiangxi Provincial Key Laboratory of Camellia Germplasm Conservation and Utilization, Jiangxi Academy of Forestry, Nanchang, Jiangxi 330047, China
*Corresponding authors. E-mail: laxu@njfu.edu.cn, fuliangcaonjfu@163.com, jxwenqiang@aliyun.com
†These authors contributed equally.

Abstract

Camellia oil extracted from *Camellia* seeds is rich in unsaturated fatty acids and secondary metabolites beneficial to human health. However, no oil-tea tree genome has yet been published, which is a major obstacle to investigating the heredity improvement of oil-tea trees. Here, using both Illumina and PicBio sequencing technologies, we present the first chromosome-level genome sequence of the oil-tea tree species *Camellia chekiangoleosa* Hu. (CCH). The assembled genome consists of 15 pseudochromosomes with a genome size of 2.73 Gb and a scaffold N50 of 185.30 Mb. At least 2.16 Gb of the genome assembly consists of repetitive sequences, and the rest involves a high-confidence set of 64 608 protein-coding gene models. Comparative genomic analysis revealed that the CCH genome underwent a whole-genome duplication event shared across the *Camellia* genus at ~57.48 MYA and a γ-WGT event shared across all core eudicot plants at ~120 MYA. Gene family clustering revealed that the genes involved in terpenoid biosynthesis have undergone rapid expansion. Furthermore, we determined the expression patterns of oleic acid accumulation- and terpenoid biosynthesis-associated genes in six tissues. We found that these genes tend to be highly expressed in leaves, pericarp tissues, roots, and seeds. The first chromosome-level genome of oil-tea trees will provide valuable resources for determining *Camellia* evolution and utilizing the germplasm of this taxon.

Introduction

World-famous *Camellia* (Theaceae) plants are valued not only for their aesthetic contributions to landscaping but also for the nutritional and health benefits of beverages containing their compounds and of their edible oils. Constituting one of the four major oil-bearing groups of trees worldwide, oil-tea trees refer to the general name of several *Camellia* species whose seeds have a high oil content and are cultivated for their edible value. Camellia oil or tea-oil extracted from *Camellia* seeds is rich in unsaturated fatty acids (UFAs) and a variety of secondary metabolites beneficial to human health and is known as "oriental olive oil" due to its high oleic oil content and antioxidant activity [1]. Camellia oil and its by-products are also widely used in the medicinal and cosmetic industries. Oil-tea tree species are documented as traditional woody edible oil crop species in East and Southeast Asia, with their cultivation and use for edible oil in China dating back more than 2000 years [2]. In 2020, the planting area of oil-tea trees was approximately 4.3 million hectares in China, accounting for more than 95% of the global camellia tea-oil resources, and its annual output value exceeded 116 billion yuan. China is currently the only country with a substantial production of tea oil, and the main cultivated species are *Camellia oleifera*, *Camellia meiocarpa*, and *Camellia chekiangoleosa* [3].

Camellia oil has an ideal fatty acid profile and is a natural competitor of olive oil. UFAs are an important component of biomembranes and play a critical role in regulating the biological processes of cells. UFAs are also essential for human survival and play a vital role in regulating physiological processes, such as maintenance of the nervous system and regulation of glucose and lipid metabolism [4]. The biosynthesis of UFAs is a very complex process. Acetyl coenzyme A is polymerized in the chloroplast stroma to form saturated fatty acids (SFAs) with 16–18 carbons via the catalysis of fatty acid synthases, and then SFAs are converted to palm acid and oleic acid via Δ9 fatty acid desaturase [stearoyl-ACP desaturase (SAD)] [5]. Oleic acid is desaturated by Δ12 fatty acid desaturase (FAD2 and FAD6) to form linoleic acid, which is further desaturated by Δ15 desaturase (FAD3, FAD7, and FAD8) to synthesize α-linolenic acid

Received: 12 November 2021; **Accepted:** 18 December 2021; **Published:** 18 January 2022; **Corrected and Typeset:** 24 January 2022

(ALA) or by Δ6 desaturase to synthesize γ-linolenic acid (GLA). ALA and GLA are further processed into docosahexaenoic acid and arachidonic acid, respectively, both of which are essential for humans [5].

As important bioactive components of camellia oil, secondary metabolites, including phenols, proanthocyanidins, tocopherols, and carotenoids, are gaining increased amounts of attention because of their health benefits [6]. Various triterpenoids have been isolated from *Camellia* and investigated for their bioactivity. Cycloartanol, β-amyrin, and squalene are the three main triterpenoids in camellia oil, with gross contents of 1043.30 mg/kg, 878.24 mg/kg, and 133.26 mg/kg, respectively. A triterpenoid saponin from camellia oil exerts both antioxidant and antimutagenic properties in humans and animals. In camellia oil, squalene, a common precursor of triterpenoids, has a variety of physiological activities, such as antiaging, antitumor, and antioxidant activities [7]. In the biosynthesis of triterpenoids, two molecules of FPP are catalyzed by squalene synthase (SQS) to synthesize squalene, which is further catalyzed by squalene epoxidase (SQE) to form 2,3-oxido-squalene. Finally, 2,3-oxido-squalene undergoes a series of protonation, cyclization, rearrangement, and deprotonation through the cyclization of different types of oxosqualene cyclases (OSCs) to form a triterpenoid backbone [8].

Oil-tea trees are monoecious and allogamous plants that can be vegetatively propagated and live for thousands of years. Although the naturally long generation time of oil-tea trees has traditionally hindered the breeding of these species, considerable efforts have been made to study the flowering and fruiting and biosynthesis of camellia oils from biochemical, physiological, or molecular genetic perspectives over the last several decades. The oil-tea tree species *Camellia chekiangoleosa* Hu. (CCH), a diploid in the genus *Camellia*, is naturally distributed in the mountainous areas in Fujian, Jiangxi, Hunan, Zhejiang, and Anhui provinces and has been introduced to Europe, America, and Australia as an ornamental plant species [9]. Extensive trials on CCH have been conducted in various provinces of China to improve its oil quality and yield [10]. Dried CCH kernels consisted of 60.3% crude fats, 8.8% crude protein, and 10.3% camellia saponins, and the oil content exceeded 60%. CCH oil mostly consists of oleic acid, linoleic acid, stearic acid, and palm acid, of which UFAs (mainly oleic acid and linoleic acid) account for approximately 90%, resulting in its very highly valuable nutritional and health-promoting functions [11]. However, the genetic bases for the growth and development of these *Camellia* species, especially the biosynthesis of bioactive compounds in camellia oil, are not yet understood due to lack of a reference genome. A high-quality reference genome could provide researchers with great convenience. Several tea (*Camellia sinensis*) genome sequences have recently been released and there are increasing multiomics studies based on their genomes, transcriptomes, proteomes, non-coding RNA, and genome-wide association studies that have fully revealed the biological properties of tea and provided researchers with new insights to better study the medicinal and economic value of tea [12–14]. The lack of a reference genome sequence is a major obstacle for basic and applied biology on oil-tea trees. We herein present a high-quality genome sequence of CCH, and reveal the biosynthesis of UFAs and terpenoids in camellia oil. This genome sequence will facilitate the understanding of *Camellia* genome evolution and tea oil biosynthesis and will promote germplasm utilization for breeding improved oil-tea tree varieties.

Results

Chromosome-level assembly of the CCH genome

The genome size of CCH was evaluated by two methods before formal assembly (Supplementary Fig. S1). To obtain a chromosome-level reference genome of CCH, we generated PacBio HiFi reads (51.09 Gb, ~19-fold genome coverage) and Illumina Hi-C reads (283.40 Gb, ~102-fold genome coverage). By using the hifiasm software as an assembly tool, we ultimately yielded a 2.73 Gb genome of CCH that covers 97.40% of the scaffolds and consists of 15 pseudochromosomes (scaffold N50 = 185.30 Mb) (Fig. 1). At least 2.16 Gb of repetitive sequences accounted for 79.09% of the CCH genome assembly. Long terminal repeat (LTR) retrotransposons are the most dominant class of transposable elements in the CCH genome, accounting for approximately 64.56% of the genome, among which *Copia* and *Gypsy* elements are the two most dominant classes of LTR retrotransposons, constituting 6.55% and 34.47% of the genome, respectively (Supplementary Table S1). Through *ab initio* modeling, protein-based searches, and transcript analysis of long-read isoform sequencing and short-read RNA sequencing data, a high-confidence set of 64 608 protein-coding gene models encoding a total of 66 579 proteins, 64 130 of which were annotated in at least one database, was established (Table 1, Supplementary Fig. S2). Two methods were used to assess the completeness of the CCH genome. First, the statistical results of BUSCO showed that the CCH genome covered 2177 (93.59%) of 2326 complete gene models, of which 1940 (83.40%) genes were present as single copies and 237 genes were present as multiple copies. Second, an LTR assembly index score (LAI) of 11.53 indicated that the CCH genome has high sequence continuity. Taken together, these results indicate that the assembly of the CCH genome is of high quality and meets the reference genome standards (Supplementary Fig. S3).

Phylogenetic status of CCH

A total of 21 747 gene families were identified via comparisons of protein sequences homologous to CCH and 15 other species, of which the number of single-copy orthologous genes was 154. To confirm the relationship between CCH and other species, we constructed a

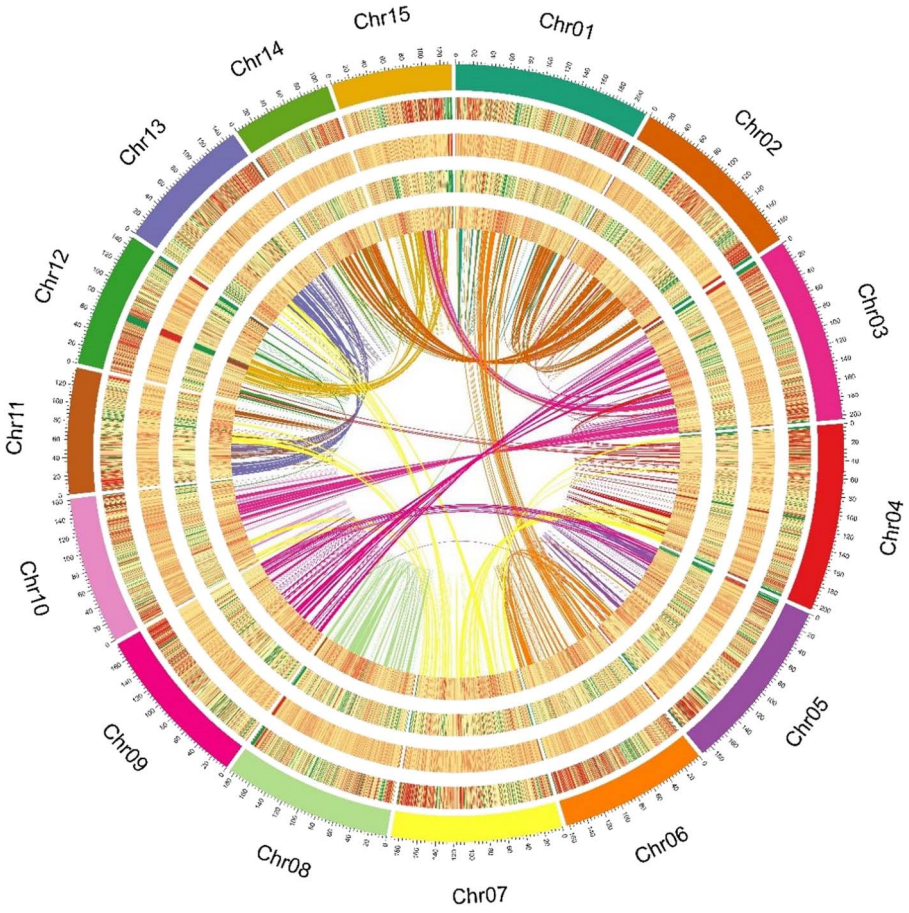

Figure 1. Features of the CCH genome. The outermost ring represents the 15 pseudochromosomes of CCH, and the second to fifth circles represent the genes, total TEs, *Copia* element distribution and *Gypsy* element distribution, with the green color representing a lower density and the red color representing a higher density. The innermost region indicates the colinear regions within the genome of CCH.

Table 1. Summary of genome assembly and annotation information for CCH, CSS-BY, CSS-SCZ, and DASZ

	CCH	CSS-BY	CSS-SCZ	DASZ
Genome size (Gb)	2.73	3.25	2.94	3.11
N50 of contigs (Kb)	1921.46	625.11	600.46	2589.77
N50 of scaffolds (Mb)	185.25	195.68	/	201.21
GC content (%)	39.23	38.24	38.25	38.98
Number of genes	64 608	40 812	50 525	33 021
Average CDS length (bp)	1125	/	1086	1139.09
Average exon length (bp)	280	5.2	245	211.7
Average exon number per gene	5.02	5.2	5.1	5.38
Repeat sequence length (Gb)	2.16	2.41	2.55	2.72
Percentage of repeat sequences (%)	79.09	74.13	87	87.41
LAI score	11.53	/	12.45	/
Sequences anchored to chromosomes (%)	97.4	97.87	86.73	99.55
BUSCO	93.8	88.13	94.4	93.2

Note: CCH (*Camellia chekiangoleosa*), CSS-BY (*Camellia sinensis* var. *sinensis* "Biyun"), CSS-SCZ (*C. sinensis* var. *sinensis* "Shuchazao"), and DASZ (an ancient tea tree of species *Camellia sinensis*) all belong to the family Theaceae, genus *Camellia*.

high-confidence phylogenetic tree with *Zea mays* and *Elaeis guineensis* as the outgroup species by using 154 single-copy orthologous genes shared between CCH and 15 other species. According to the results, CCH and *C. sinensis* were clustered on the same branch, which belongs to the family Theaceae, and the result is consistent with research based on chloroplast genomes (Fig. 2a) [15]. Furthermore, we also constructed a phylogenetic tree of divergence time, and the results showed that the family Theaceae (CCH and *C. sinensis*) separated from the family Actinidiaceae (*Actinidia chinensis*) ~71.22 (49.22–93.81) MYA (Supplementary Fig. S4).

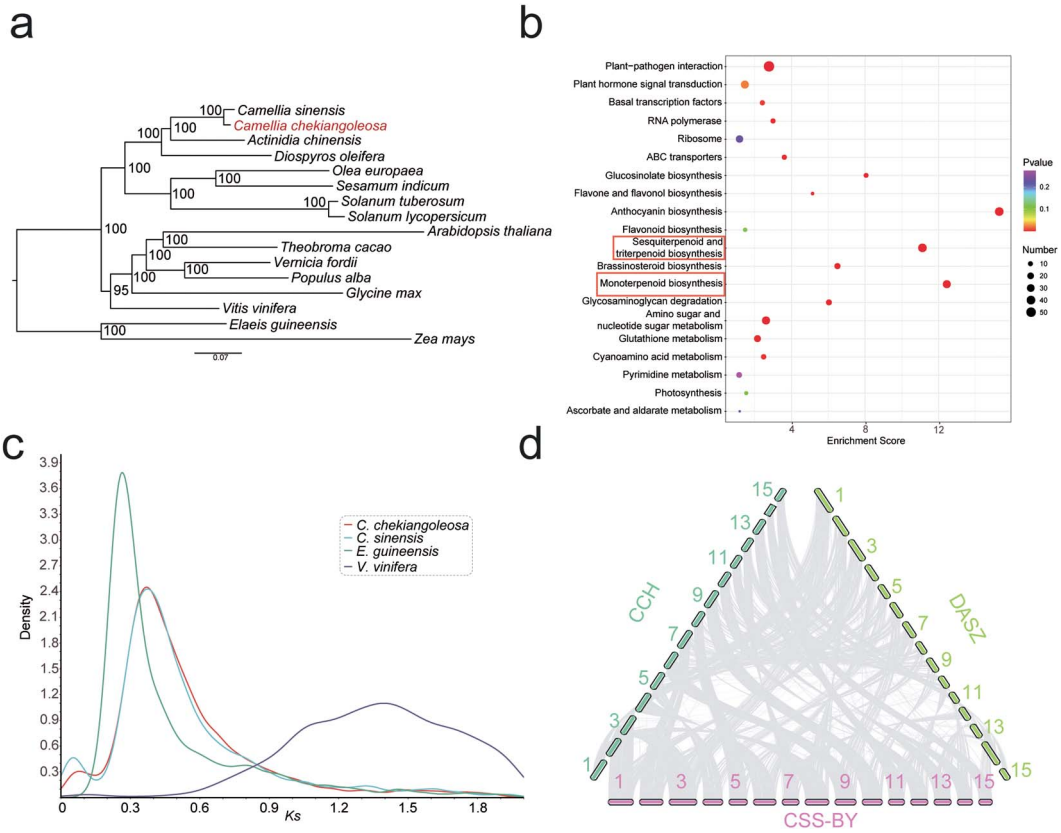

Figure 2. Comparative genome analysis. (a) Dated phylogeny for 16 plant species with monocots constituting an outgroup. The bootstrap value is given in black. (b) KEGG enrichment analysis of 414 gene families undergoing rapid expansion in CCH. (c) Density distribution of *Ks* for paralogous gene pairs of the four plant species. (d) Interspecific collinearity at the chromosome level across CCH, DASZ, and CSS-BY. The gray lines indicate connected matched gene pairs.

Expansion and contraction of gene families

Comparing the gene families of CCH, *Elaeis guineensis*, *Vernicia fordii*, *Diospyros oleifera*, and *Olea europaea*. A total of 18 577 gene families were found in five species, and we found that the KEGG enrichment analysis of CCH-specific gene families (2264 gene families) revealed that "fatty acid biosynthesis" (ko00061, enrichment score = 2.83) was one of the significantly enriched terms (Supplementary Fig. S5, Supplementary Fig. S6, Supplementary Table S2). We identified 3017 and 2941 gene families in the CCH genome that underwent expansion and contraction, respectively, with 414 gene families (8168 genes) undergoing rapid expansion and 131 gene families (265 genes) undergoing rapid contraction (Fig. 2a, Supplementary Fig. S7, Supplementary Table S3). KEGG enrichment analysis of the 414 rapidly expanding gene families revealed that 15 KEGG pathways were significantly enriched (*P*-value <0.05), with "sesquiterpenoid and triterpenoid biosynthesis" (ko00909, enrichment score = 11.06) and "monoterpenoid biosynthesis" (ko00902, enrichment score = 12.39) among the significantly enriched pathways, suggesting that the rapid expansion of genes associated with monoterpene, sesquiterpene, and triterpene biosynthesis may be related to the unique properties of CCH (Fig. 2b, Supplementary Table S4).

Whole-genome duplication and collinearity

To identify the whole-genome duplication (WGD) events experienced by CCH, we analyzed the *Ks* distribution of CCH, *C. sinensis*, *E. guineensis*, and *Vitis vinifera*. The distribution of *Ks* showed one peak at ~1.3–1.5 in the genome of CCH, *C. sinensis*, *E. guineensis*, and *V. vinifera*, indicating that these species shared an ancient γ-WGT (whole-genome triplication) event that was also shared by all core eudicot plants (Fig. 2c). We also noticed one peak at ~0.3–0.4 in the genome of CCH and *C. sinensis*, indicating that they shared a recent WGD event that was shared across members of the genus *Camellia* (Fig. 2c). Studies have shown that tea plants experienced only one WGD event after the γ-WGT event, and some genes involved in catechin and caffeine biosynthesis expanded and were retained following the WGD event, contributing to the flavor compounds of tea plants [16]. A number of studies have found that the ancestors of eudicots had a γ-WGT event 120 million years ago, and the genes preserved from this ancient WGT event were mostly associated with water acquisition and salt stress, which occurred during the arid Cretaceous period, so it is assumed that this genome-wide replication event provided the genetic basis for plant species to adapt to the harsh survival environment during this period [17, 18]. As a member of eudicots, several tea genome sequences supported the genus

Camellia, which also underwent the γ-WGT event [19, 20]. According to Fig. 2c, the grape underwent only one WGD event (γ-WGT), and the genus *Camellia* had a recent WGD event (~57.48 million years ago). We think that this event may have weakened the relationship between paralogous homologous gene pairs in WGT events, resulting in their peaks being less easily observed in the region of ~1.3–1.5. The peak value of orthologous gene pairs between CCH and *C. sinensis* (*Ks* = 0.5) was lower than both the value between CCH and *V. vinifera* (*Ks* = 0.8) and the peak value between CCH and *E. guineensis* (*Ks* = 1.6), implying that the divergence time between CCH and *C. sinensis* occurred later; these results correspond to the phylogenetic relationships (Fig. 2a, Supplementary Fig. S8).

To better understand the evolution of CCH, we determined the collinearity relationship between CCH and CSS-BY. The results showed that there was high collinearity between CCH and CSS-BY, indicating that there was no large-scale structural variation after the divergence of CCH and CSS-BY. For most collinear regions, one chromosome of CCH corresponded to one chromosome of CSS-BY; for example, CchChr1, CchChr2, CchChr3, CchChr4, CchChr5, CchChr6, CchChr7, CchChr8, CchChr9, CchChr10, CchChr11, CchChr12, CchChr13, CchChr14, and CchChr15 of CCH corresponded to CsChr3, CsChr8, CsChr1, CsChr6, CsChr4, CsChr9, CsChr2, CsChr5, CsChr10, CsChr12, CsChr14, CsChr7, CsChr13, CsChr15, and CsChr11 of CSS-BY, respectively (Fig. 2d, Supplementary Fig. S9a). In addition, we further analyzed the intergenomic collinearity between CCH and DASZ. Although one chromosome of CCH usually corresponds to one chromosome of DASZ, the collinear regions between the two are not as strong as those between CCH and CSS-BY, indicating that they have a distant relationship (Fig. 2d, Supplementary Fig. S9b).

Tandem duplication and positive selection within the CCH genome

KEGG enrichment analysis revealed that 9200 tandemly duplicated genes (14.24% of all genes on the chromosomes) were enriched in "sesquiterpenoid and triterpenoid biosynthesis" (ko00909, enrichment score = 3.51), "monoterpenoid biosynthesis" (ko00902, enrichment score = 3.59), "anthocyanin biosynthesis" (ko00942, enrichment score = 6.74), "phenylpropanoid biosynthesis" (ko00940, enrichment score = 2.60), "flavonoid biosynthesis" (ko00941, enrichment score = 2.78), "plant hormone signal transduction" (ko04075, enrichment score = 1.61), and "plant-pathogen interaction" (ko04626, enrichment score = 2.12), indicating that the tandemly duplicated genes are involved in secondary metabolite biosynthesis and plant–environment interactions (Supplementary Fig. S10a).

A total of 119 genes were subjected to positive selection in the CCH genome. KEGG enrichment analysis showed that these genes were enriched in "base excision repair" (ko03410, enrichment score = 17.52), "DNA replication" (ko03030, enrichment score = 5.61), "vitamin B6 metabolism" (ko00750, enrichment score = 16.19), "terpenoid backbone biosynthesis" (ko00900, enrichment score = 6.92), and "fatty acid degradation" (ko00071, enrichment score = 4.29), indicating that these genes may be involved in DNA repair and plant secondary metabolism (Supplementary Fig. S10b).

UFA biosynthesis-associated genes

Studies have shown that *FAD* and *SAD* genes are essential for the biosynthesis of UFAs in a variety of oilseed plant species [21, 22]. A total of 32 genes involved in the biosynthesis of acyl-lipid desaturase and acyl-ACP desaturase were identified in the CCH genome in this study, of which there were 11 *CchSAD* and 17 *CchFAD* genes, respectively. We found that all *SAD* genes clustered onto one branch according to our phylogenetic tree based on the homologous sequences of CCH, *Glycine max*, *Oryza sativa*, *Sesamum indicum*, *Olea europaea*, *Arabidopsis thaliana*, and *Arachis hypogaea*, indicating that the expansion of *SAD* genes occurred after the divergence of these species (Fig. 3). *FAD1/2/3/6/7/8* of *FAD* genes clustered onto one branch, while *FAD4/5* clustered onto another branch. Among *FAD1/2/3/6/7/8*, the products of *FAD6/7/8* are localized in plastids and have similar structures; the products of *FAD2* and *FAD3* are localized in the endoplasmic reticulum and use phosphatidylcholine as the preferred substrate; *FAD2* and *FAD3* are key genes involved in SFA desaturation and encode key enzymes for oleic and linoleic acid desaturation, respectively [23]. In fact, some of the genes involved in fatty acid biosynthesis are expressed specifically in seeds and are important organs for the rapid accumulation of lipids. For example, one type of the *CchFAD2* gene, *CchFAD2A* (Cch15G000175), is expressed over one hundred times more in the seeds than in other tissues, while another type, *CchFAD2B* (Cch10G003830), is highly expressed only in the shoots (Fig. 4). Members of the *CchSAD* (*CchSAD2*, Cch05G001837) gene family, which are closely involved in fatty acid synthesis and highly expressed in seeds, may be closely related to the accumulation of UFAs (Fig. 4).

Terpenoid biosynthesis pathway in CCH

Camellia oil is rich in a variety of secondary metabolites that are beneficial to human health, such as terpenoids, and elucidating the potential terpenoid biosynthesis pathway of CCH will help us better understand the medicinal value of camellia oil. We found that genes involved in terpenoid biosynthesis underwent rapid expansion in the CCH genome, and a total of 86 genes involved in terpenoid biosynthesis were identified in the whole-genome sequence of CCH in this study, including 61 *CchTPS*, five *CchSQS*, five *CchSQE*, and 15 *CchOSC* genes, which were named in order of their position on the chromosome (Fig. 5). Chromosomal localization of genes involved in terpenoid biosynthesis revealed that most of them were distributed on the chromosomes in accordance with tandem duplication (Fig. 5a, Supplementary Fig. S10a). To investigate the

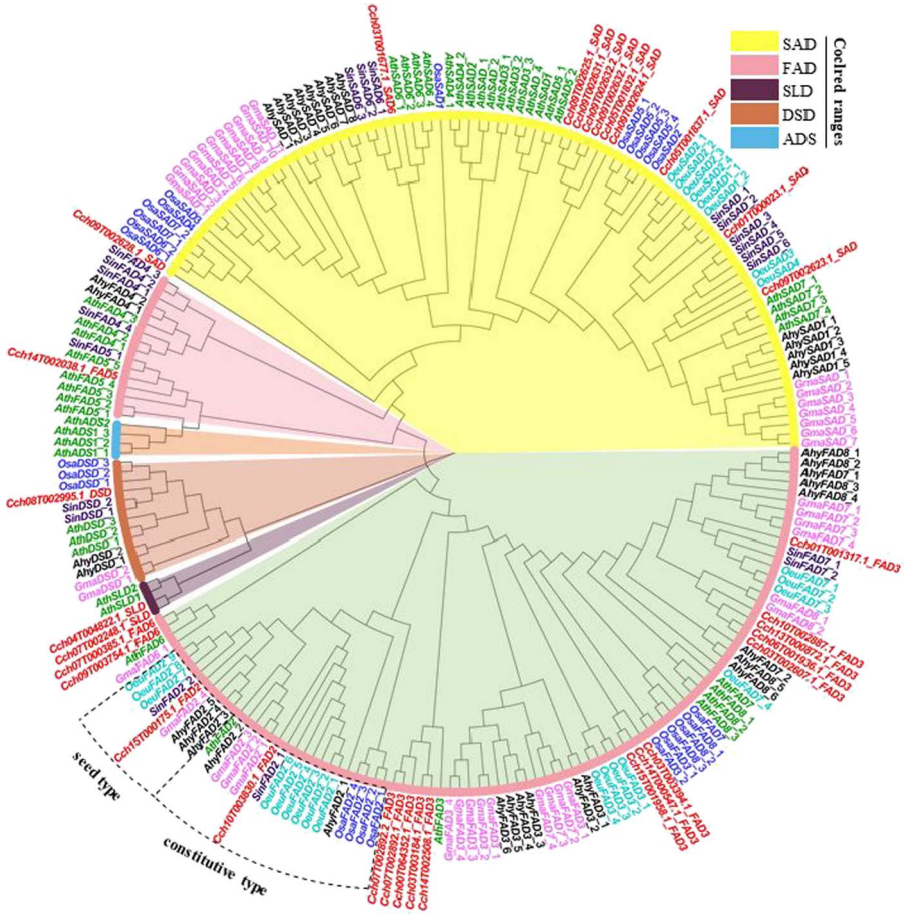

Figure 3. Phylogenetic tree of *FAD* and *SAD* candidate genes from CCH, *Oryza sativa*, *Olea europaea*, *Camellia sinensis*, *Arabidopsis thaliana*, *Glycine max*, and *Arachis hypogaea*. The CCH *FAD* and *SAD* candidate genes show high similarity to known *FAD* and *SAD* genes, respectively.

potential functions of these genes in CCH, we analyzed the expression of the 86 genes in 6 tissues, which were found to be highly expressed in the shoots and stems but expressed at low levels in the leaves, pericarp tissues, roots, and seeds (Fig. 5c, Supplementary Fig. S11).

In addition, *CchTPSs* could be divided into five subfamilies according to our phylogenetic tree, TPS-a, TPS-b, TPS-c, TPS-e/f, and TPS-g subfamilies, with 35, 16, 1, 5, and 5 members, respectively. To determine the potential function of *CchTPSs*, the NR database annotation revealed that 22 of 61 *CchTPS* genes encoded monoterpene synthases, 34 encoded sesquiterpene synthases, and five encoded diterpene synthases, and the *CchTPS* genes were found mainly to encode sesquiterpene synthases (Supplementary Table S5). We also found that different types of terpene synthase-encoding genes were distributed on chromosomes unevenly, with those encoding monoterpene synthases tending to be distributed on Chr5, Chr8, Chr10, and Chr12; those encoding sesquiterpene synthases tending to be distributed on Chr2, Chr4, Chr6, and Chr8; and those encoding diterpene synthases tending to be distributed on Chr8. According to Fig. 5a and Supplementary Table S5, we found that the *TPS* genes are located on Chr1 (one gene), Chr2 (four genes), Chr4 (six genes), Chr5 (four genes), Chr6 (16 genes), Chr8 (18 genes), Chr10 (four genes), Chr11 (one gene), Chr12 (four genes), Chr13 (two genes), and Chr15 (one gene), and 12 of the 61 *TPS* genes are tandem. Many studies supported the *TPS* genes are tandem. For example, the *TPS* genes in *Oryza sativa*, *Ricinus communis*, *Solanum lycopersicum*, *Medicago truncatula*, and *Lotus japonicus* were often arranged to generate gene clusters and distributed in tandem duplication [24–28].

Discussion

In this study, by combining PacBio, Hi-C, and Illumina sequencing technologies, we yielded the first chromosome-level reference genome of an oil-tea tree species. The assembled genome consisted of 15 pseudochromosomes with a size of approximately 2.73 Gb. Both BUSCO and LAI scores indicated that the assembly quality was high and met the reference genome level. The percentage of repetitive sequences in the genome of CCH was lower than that in DASZ (3.11 Gb) and CSS-SCZ (2.94 Gb) and higher than that in CSS-BY (3.25 Gb) (Table 1) [20, 29, 30].

Combining the gene families that expanded and contracted during the evolution of CCH and the genes subjected to positive selection, we found that genes related

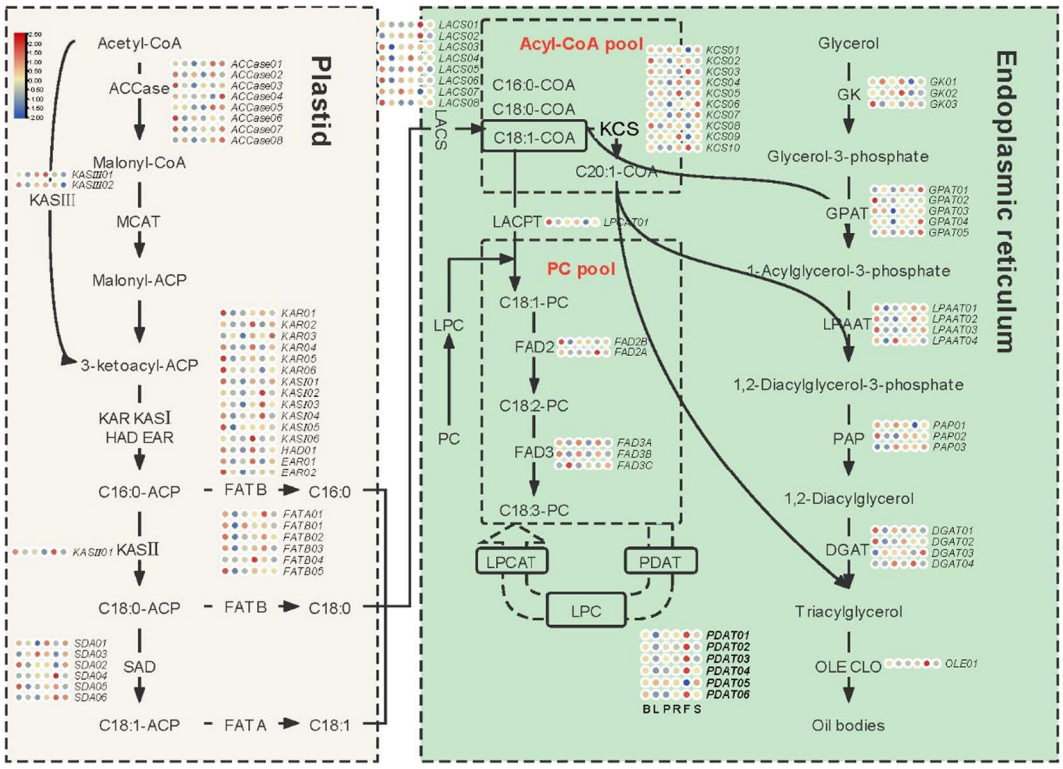

Figure 4. Tissue-specific relative expression profiles (red–blue scale) of genes implicated in oleic acid biosynthesis (heatmap). B: buds; L: leaves; P: pericarp tissues; R: roots; F: fruits; S: stems.

to fatty acid synthesis expanded and those related to linoleic acid synthesis contracted throughout evolution. The genes subjected to positive selection include fatty acid-degrading- and FAD-binding-related functions, all of which are inextricably linked to lipid metabolism. We speculate that during the evolutionary process of CCH, to continuously adapt to the environment, genes related to lipid synthesis evolved adaptively, eventually leading to the gradual development of high-lipid characteristics. Two types of *FAD2* and *SAD2* genes were found in several species, those expressed specifically in the seeds and those constitutively expressed, with the former highly expressed in the seeds and the latter expressed evenly across different tissues. In this study, *CchFAD2A* and *CchSAD2* were expressed specifically in the seeds, while *CchFAD2B* was constitutively expressed. Our previous study showed that the expression of *CchFAD2A* was much higher than that of *CchFAD2B* in the seeds, while the expression of *CchFAD2B* was more similar across various tissues.

Terpenoids are among the most diverse plant secondary metabolites and have significant research value. Terpenoid biosynthesis begins with acetyl coenzyme A or pyruvate and glyceraldehyde-3-phosphate; the former undergoes a six-step condensation reaction to produce isopentenyl diphosphate (IPP), while the latter undergoes a seven-step condensation reaction to produce IPP [31]. IPP and its isomer, DMIPP, generate monoterpenes, sesquiterpenes, diterpenes, and triterpenes under the action of different enzymes, including those encoded by *TPS* genes associated with monoterpene, sesquiterpene, and diterpene biosynthesis [32], and *SQS*, *SOE*, and *OSC* genes associated with triterpene biosynthesis [33–35]. The number of *TPS* genes in angiosperms ranged from 40 to 152 [36], and a total of 61 *CchTPS* genes containing both C- and N-terminal structural domains were identified in the CCH genome in this study, the number of which was much greater than the 23 *CsTPS* genes identified in CSS-SCZ and the 45 *CsTPS* genes identified in Tieguanyin [37], indicating that the *CchTPS* gene family in the CCH genome expanded (Fig. 5a). The phylogenetic tree showed that *CchTPSs* could be divided into five subfamilies, of which the TPS-a subfamily was the largest, which was consistent with the findings in CSS-SCZ [37], *Arabidopsis thaliana* [38], and *S. lycopersicum*, in contrast to the results in Tie-guanyin, in which TPS-b was found to be the largest subfamily. Functional annotation of 61 *CchTPSs* revealed that 22 *CchTPSs* were associated with monoterpene synthesis, 34 *CchTPSs* were associated with sesquiterpene synthesis, and five *CchTPSs* were associated with diterpene synthesis; in addition, a significant expansion of monoterpene- and sesquiterpene-encoding genes occurred in the CCH genome compared with the Tie-guanyin and CSS-SCZ genomes (Fig. 2b, Fig. 4) [37]. We also found that *CchTPS* genes were unevenly distributed on chromosomes and were mostly tandemly repeated, suggesting that CCH underwent genetic expansion during evolution, which is consistent with the findings of studies in *S. lycopersicum*, Tie-guanyin, *Ricinus communis.*, *Avena sativa*, *M. truncatula*, *A. thaliana*, and *Lotus*

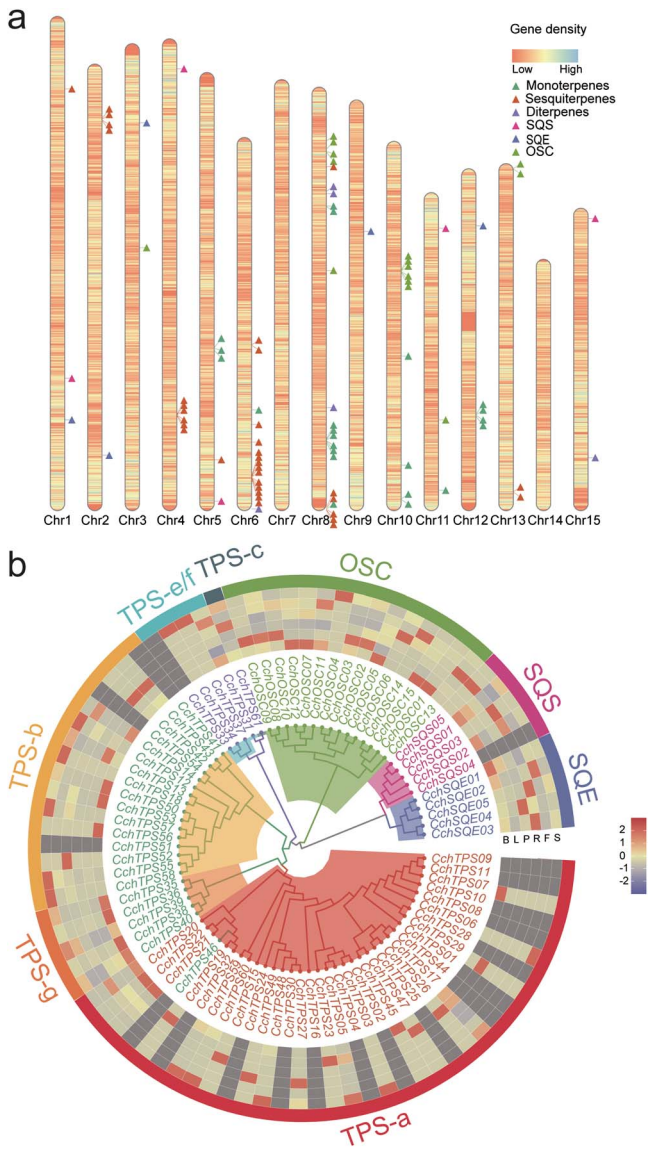

Figure 5. Genes involved in the proposed terpenoid biosynthesis pathway. (a) Chromosomal distribution pattern of genes involved in terpenoid biosynthesis. (b) Heatmap of genes involved in terpenoid biosynthesis, where B, L, P, R, F, and S represent buds, leaves, pericarp tissues, roots, seeds, and stems, respectively.

corniculatus [25, 27, 28]. The expression of *CchTPSs* was significantly different in the six different studied tissues, with most *CchTPS* genes tending to be highly expressed in the shoots and stems but expressed at low levels in the leaves, pericarp tissues, roots, and seeds; however, in CSS-SCZ, most *CsTPSs* tended to be highly expressed in the flowers and leaves. *CchTPS* genes are specifically expressed in different tissues and organs, suggesting that these genes play different functions (Fig. 5a).

SQS is a key enzyme in triterpene biosynthesis and has six relatively conserved structural domains. Two *AthSQS* genes have been identified in *A. thaliana*, but the product of only one has catalytic activity [39]. In this study, five *CchSQS* genes were identified in CCH, with *CchSQS03* and *CchSQS05* being more highly expressed in the roots. SQE is widely found in plants, animals, and humans, and it usually catalyzes squalene to form 2,3-oxidosqualene. In this study, five *CchSQE* genes were identified, *CchSQE01* was expressed at the highest level in the buds, *CchSQE02* was expressed at the highest level in the seeds, *CchSQE04* was expressed at the highest level in the pericarp tissues, and *CchSQE03* and *CchSQE05* were expressed at the highest level in the roots; moreover, seven *TwSQE* genes were identified in *Tripterygium wilfordii* and had the highest expression in the roots and flowers [40]. Two *SgSQE* genes were identified in *Siraitia grosvenorii*, and their expression was greatest on day 15 of fruit development. OSC catalyzes the generation of sterols and triterpenoid precursors from 2,3-oxidosqualene, which is a key step in the product diversity of triterpenoids. In this study, 15 *CchOSC* genes were identified in the CCH genome, the number of which was much greater than the four *CoOSC* genes found in *C. oleifera*, suggesting that *CchOSC* genes also underwent expansion. These findings imply that the high number of *CchOSC* gene family members may be closely related to their triterpene products (Fig. 5b).

Conclusions

In this study, the first high-quality chromosome-level reference genome of oil-tea trees was obtained by various techniques, and gene structure and comparative genomic analyses were performed. We identified the expression pattern of UFA- and terpenoid biosynthesis-associated genes in six tissues. It was found that genes involved in monoterpenes, sesquiterpenes, and triterpenes in the CCH genome underwent rapid expansion. Our study provides a useful reference genome for revealing *Camellia* evolution and investigating its medicinal value.

Materials and methods

Library construction and genome sequencing

High-quality genomic DNA was extracted from fresh leaves of CCH using the Plant Genomic DNA Kit (Tiangen, China) according to the manufacturer's instructions. The library was constructed using the TruSeq DNA LT Sample Prep Kit according to the manufacturer's instructions, and sequencing was performed on the Illumina HiSeq X platform (Illumina, San Diego, CA, USA) in 150 bp paired-end mode. Hi-C libraries were prepared according to a previous report [41], and paired-end 150 bp sequencing was performed on the Illumina HiSeq platform (Illumina, San Diego, CA, USA) following quality control measures. The raw sequences were filtered using fastp with the default parameters [42]. A SMRT library was then generated using the PacBio Sequel II platform (Pacific Biosciences, Menlo Park, CA, USA) according to the manufacturer's instructions.

Genome assembly and annotation

Genome assembly was performed using hifiasm in "no purge" mode, and heterozygosity was removed using

purge_dups to obtain the draft genome of CCH [43]. The chromosomes were clustered, sorted, and corrected based on Hi-C interaction information using 3D-DNA [44]. Finally, the Hi-C interaction matrix was imported into Juicebox for manual inspection. The integrity of the assembled genomes was subsequently assessed using BUSCO [45].

EDTA was used to predict LTR, terminal inverted repeat, and Helitron elements [46]. Two methods were used to predict the protein-coding genes of CCH. First, we mapped the transcriptome of CCH by using hisat2 [47]. Second, MAKER was used to align the genes of homologous protein sequences of *Actinidia chinensis*, *A. thaliana*, DASZ, CCS-SCZ, *O. europaea*, *Sesamum indicum*, *Vaccinium corymbosum*, *V. vinifera*, and *D. oleifera* to the CCH genome [48]. tRNAs were identified using tRNAscan-SE [49]; rRNAs were identified using Barrnap (https://github.com/tseemann/barrnap); and miRNAs, snRNAs, and snoRNAs were identified using Rfam [50].

Gene family clustering and phylogenetic analysis

After filtering the protein sequences to those less than 30 amino acids in length, we clustered the filtered protein sequences of CCH, *Z. mays*, *E. guineensis*, *S. indicum*, *Theobroma cacao*, *V. vinifera*, *G. max*, *A. thaliana*, *V. fordii*, *A. chinensis*, *C. sinensis*, *D. oleifera*, *O. europaea*, *S. lycopersicum*, *Solanum tuberosum*, and *Populus alba* based on similarity by OrthoFinder [51].

MAFFT was used to perform the multiple sequence alignment [52]. Afterward, RAxML was used to construct a phylogenetic tree by concatenating the 154 single-copy orthologous genes [53] after exacting the conserved region by TrimAL [54]. The correction time points were obtained using TimeTree [55], and the divergence time was estimated using PAML [56]. Gene family expansion and contraction were detected using CAFE [57] and visualized using ggtree [58].

WGD and collinearity

We analyzed the *Ks* distribution to discover WGD events within CCH, *C. sinensis*, *E. guineensis*, and *V. vinifera*. KaKs_calculator was used to calculate *Ks* values [59]. The collinearity relationships between CCH, CSS-BY, and DASZ were determined using JCVI after exacting the longest protein sequences of each gene [60].

Tandem duplication and positive selection

To identify tandemly duplicated genes in the CCH genome, we first extracted the longest protein sequences of each gene and then used blastp to identify homologous genes.

The protein sequences of single-copy gene family members shared across CCH, *C. sinensis*, *D. oleifera*, *O. europaea*, *S. indicum*, *C. sinensis*, *S. lycopersicum*, and *Solanum tuberosum* were aligned by MAFFT [52]. Then, we detected the positively selected genes using the branch-site mode with CCH serving as the foreground branch and the other seven species constituting the background branch [56].

Terpenoid-related gene identification

Genes involved in terpenoid biosynthesis were identified using HMMER against the proteome of CCH ($e < 10–5$). PF03936 and PF01397 were used to identify the *CchTPS* gene; PF00494, the *CchSQS* gene; PF08491, the *CchSQE* gene; and PF13243 and PF13249, the *CchOSC* gene. Sequences shorter than 200 amino acids were excluded from further analysis.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31860179 and 31260184), Key Research and Development Program of Jiangxi Province (20201BBF61003 and 20161BBF60122), the Natural Science Foundation of Jiangxi Province, China (20151BAB204030), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), and the Doctor Initial Project of Jiangxi Academy of Forestry (2021521001). We are grateful to Mr Yin-Cong Gu and Ms Dong An (Shanghai OE Biotechnology Co., Ltd.) for their technical support in genome data analysis.

Author contributions

L.X., F.C., Q.W., and M.X. conceived the project. T.S., B.H., M.X., P.Z., Z.N., and C.G. participated in the data analysis. M.X., T.S., and B.H. wrote the manuscript. All authors have read and approved the final version of the paper.

Data availability

The whole genome sequence data reported in this paper have been deposited in the Genome Warehouse in the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation, under accession number GWHBGBN00000000, which is publicly accessible at https://ngdc.cncb.ac.cn/gwh.

Conflict of interest

The authors declare no competing interests.

Supplementary data

Supplementary data is available at *Horticulture Research Journal* online.

References

1. Zhu M, Shi T, Chen Y *et al.* Prediction of fatty acid composition in camellia oil by 1H NMR combined with PLS regression. *Food Chem*. 2019;**279**:339–46.
2. Zhou J, Ai Z, Wang H *et al.* Phosphorus alleviates aluminum toxicity in Camellia oleifera seedlings. *Int J Agric Biol*. 2019;**21**:237–43.
3. Zhang D, Yu J, Zhang R *et al.* Teaoil Camellia—Eastern "olive" for the world. *Acta Hortic*. 2006;**769**:43–8.

4. Sokoła-Wysoczańska E, Wysoczanski T, Wagner J *et al.* Polyunsaturated fatty acids and their potential therapeutic role in cardiovascular system disorders—a review. *Nutrients*. 2018;**10**:1561.
5. He M, Qin C-X, Wang X *et al.* Plant unsaturated fatty acids: biosynthesis and regulation. *Front Plant Sci*. 2020;**11**:390.
6. Wang Y, Sun D, Chen H *et al.* Fatty acid composition and antioxidant activity of tea (Camellia sinensis L.) seed oil extracted by optimized supercritical carbon dioxide. *Int J Mol Sci*. 2011;**12**: 7708–19.
7. Qian J, Liu Y, Ma C *et al.* Positive selection of squalene synthase in Cucurbitaceae plants. *Int J Genomics*. 2019;**2019**:1.
8. Shang Y, Huang S. Multi-omics data-driven investigations of metabolic diversity of plant triterpenoids. *Plant J*. 2019;**97**: 101–11.
9. Xie Y, Wang X. Comparative transcriptomic analysis identifies genes responsible for fruit count and oil yield in the oil tea plant Camellia chekiangoleosa. *Sci Rep*. 2018;**8**:6637.
10. Wang X, Zeng Q, del Mar Contreras M *et al.* Profiling and quantification of phenolic compounds in Camellia seed oils: natural tea polyphenols in vegetable oil. *Food Res Int*. 2017;**102**:184–94.
11. Guo H, Tan H, Zhou J. Proximate composition of Camellia chekiangoleosa Hu fruit and fatty acid constituents of its seed oil. *Journal of Zhejiang University (Agriculture & Life Sciences)*. 2010;**36**:662–9.
12. Liu Z-W, Li H, Liu J-X *et al.* Integrative transcriptome, proteome, and microRNA analysis reveals the effects of nitrogen sufficiency and deficiency conditions on theanine metabolism in the tea plant (Camellia sinensis). *Hortic Res.* 2020;**7**:65.
13. Lu L, Chen H, Wang X *et al.* Genome-level diversification of eight ancient tea populations in the Guizhou and Yunnan regions identifies candidate genes for core agronomic traits. *Hortic Res*. 2021;**8**:190.
14. Xia E-H, Tong W, Wu Q *et al.* Tea plant genomics: achievements, challenges and perspectives. *Hortic Res.* 2020;**7**:7.
15. Yin X, Li T, Huang B *et al.* Complete chloroplast genome of Camellia chekiangoleosa (Theaceae), a shrub with gorgeous flowers and rich seed oil. *Mitochondrial DNA Part B*. 2021;**6**: 840–1.
16. Wang Y, Chen F, Ma Y *et al.* An ancient whole-genome duplication event and its contribution to flavor compounds in the tea plant (Camellia sinensis). *Hortic Res.* 2021;**8**:176.
17. Chen J, Hao Z, Guang X *et al.* Liriodendron genome sheds light on angiosperm phylogeny and species–pair differentiation. *Nature Plants*. 2019;**5**:18–25.
18. Alix K, Gérard PR, Schwarzacher T *et al.* Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Ann Bot*. 2017;**120**:183–94.
19. Xia E-H, Zhang HB, Sheng J *et al.* The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Mol Plant*. 2017;**10**:866–77.
20. Chen J-D, Chao Z, Ma JQ *et al.* The chromosome-scale genome reveals the evolution and diversification after the recent tetraploidization event in tea plant. *Hortic Res*. 2020;**7**:63.
21. Yukawa Y, Takaiwa F, Shoji K *et al.* Structure and expression of two seed-specific cDNA clones encoding stearoyl-acyl carrier protein desaturase from sesame, *Sesamum indicum* L. *Plant Cell Physiol*. 1996;**37**:201–5.
22. Combs R, Bilyeu K. Novel alleles of FAD2-1A induce high levels of oleic acid in soybean oil. *Mol Breed*. 2019;**39**:79.
23. Zheng Y, Chen C, Liang Y *et al.* Genome-wide association analysis of the lipid and fatty acid metabolism regulatory network in the mesocarp of oil palm (Elaeis guineensis Jacq.) based on small noncoding RNA sequencing. *Tree Physiol*. 2019;**39**:356–71.
24. Shimura K, Okada A, Okada K *et al.* Identification of a biosynthetic gene cluster in rice for momilactones. *J Biol Chem*. 2007;**282**:34013–8.
25. King AJ, Brown GD, Gilday AD *et al.* Production of bioactive diterpenoids in the Euphorbiaceae depends on evolutionarily conserved gene clusters. *Plant Cell*. 2014;**26**:3286–98.
26. Matsuba Y, Zi J, Jones AD *et al.* Biosynthesis of the diterpenoid lycosantalonol via nerylneryl diphosphate in *Solanum lycopersicum*. *PLoS One*. 2015;**10**:e0119302.
27. Naoumkina MA, Modolo LV, Huhman DV *et al.* Genomic and coexpression analyses predict multiple genes involved in triterpene saponin biosynthesis in *Medicago truncatula*. *Plant Cell*. 2010;**22**:850–66.
28. Krokida A, Delis C, Geisler K *et al.* A metabolic gene cluster in Lotus japonicus discloses novel enzyme functions and products in triterpene biosynthesis. *New Phytol*. 2013;**200**:675–90.
29. Zhang W, Zhang Y, Qiu H *et al.* Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nat Commun*. 2020;**11**:3719.
30. Wei C, Yang H, Wang S *et al.* Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. *Proc Natl Acad Sci U S A*. 2018;**115**: E4151–8.
31. Pu X, Dong X, Li Q *et al.* An update on the function and regulation of methylerythritol phosphate and mevalonate pathways and their evolutionary dynamics. *J Integr Plant Biol*. 2021;**63**:1211–26.
32. Alicandri, E, Paolacci, A. R, Osadolor, S. *et al.* On the evolution and functional diversity of terpene synthases in the Pinus species: a review. *J Mol Evol*. 2020;**88**:253–83.
33. Wang J-R, Lin JF, Guo LQ *et al.* Cloning and characterization of squalene synthase gene from Poria cocos and its up-regulation by methyl jasmonate. *World J Microbiol Biotechnol*. 2014;**30**:613–20.
34. Laranjeira S, Amorim-Silva V, Esteban A *et al.* Arabidopsis Squalene Epoxidase 3 (SQE3) complements SQE1 and is important for embryo development and bulk squalene epoxidase activity. *Mol Plant*. 2015;**8**:1090–102.
35. Xue Z, Duan L, Liu D *et al.* Divergent evolution of oxidosqualene cyclases in plants. *New Phytol*. 2012;**193**:1022–38.
36. Chen F, Tholl D, Bohlmann J *et al.* The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J*. 2011;**66**:212–29.
37. Zhou H-C, Shamala LF, Yi X-K *et al.* Analysis of terpene synthase family genes in Camellia sinensis with an emphasis on abiotic stress conditions. *Sci Rep*. 2020;**10**:933.
38. Aubourg S, Lecharny A, Bohlmann J. Genomic analysis of the terpenoid synthase (AtTPS) gene family of Arabidopsis thaliana. *Mol Gen Genomics*. 2002;**267**:730–45.
39. Busquets A, Keim V, Closa M *et al.* Arabidopsis thaliana contains a single gene encoding squalene synthase. *Plant Mol Biol*. 2008;**67**: 25–36.
40. Liu Y, Zhou J, Hu T *et al.* Identification and functional characterization of squalene epoxidases and oxidosqualene cyclases from Tripterygium wilfordii. *Plant Cell Rep*. 2020;**39**:409–18.
41. Rao SSP, Huntley MH, Durand NC *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014;**159**:1665–80.
42. Chen S, Zhou Y, Chen Y *et al.* Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England)*. 2018;**34**:i884–90.
43. Cheng H, Concepcion GT, Feng X *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;**18**:170–5.

44. Dudchenko O, Batra SS, Omer AD *et al.* De novo assembly of the genome using Hi-C yields chromosome-length scaffolds. *Science*. 2017;**356**:92–5.
45. Simão FA, Waterhouse RM, Ioannidis P *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*. 2015;**31**:3210–2.
46. Ou S, Su W, Liao Y *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*. 2019;**20**:275.
47. Kim D, Paggi JM, Park C *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;**37**:907–15.
48. Cantarel BL, Korf I, Robb SMC *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*. 2008;**18**:188–96.
49. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 1997;**25**:955–64.
50. Kalvari I, Argasinkska J, Quinones-Olvera N *et al.* Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res*. 2018;**46**:D335–42.
51. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;**16**:157.
52. Katoh K, Misawa K, Kuma K-I *et al.* MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;**30**:3059–66.
53. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*. 2014;**30**:1312–3.
54. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;**25**:1972–3.
55. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics (Oxford, England)*. 2006;**22**:2971–2.
56. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. 1997;**13**:555–6.
57. De Bie T, Cristianini N, Demuth JP *et al.* CAFE: a computational tool for the study of gene family evolution. *Bioinformatics (Oxford, England)*. 2006;**22**:1269–71.
58. Yu G, Smith DK, Zhu H *et al.* Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol*. 2017;**8**:28–36.
59. Zhang Z, Li J, Zhao XQ *et al.* KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics*. 2006;**4**:259–63.
60. Tang H, Bowers JE, Wang X *et al.* Synteny and collinearity in plant genomes. *Science*. 2008;**320**:486–8.