The chromosome-scale genome sequence of *Triadica sebifera* provides insight into fatty acids and anthocyanin biosynthesis

Jie Luo¹, Wenyu Ren¹, Guanghua Cai¹, Liyu Huang¹, Xin Shen², Na Li³, Chaoren Nie³, Yingang Li ² & Nian Wang ¹

The Chinese tallow tree (*Triadica sebifera*) can produce oil with high content of unsaturated fatty acids in seeds and shows attractive leaf color in autumn and winter. Here, the 739 Mb chromosome-scale genome sequence of the Chinese tallow tree was assembled and it reveals the Chinese tallow tree is a tetraploid. Numerous genes related to nutrition assimilation, energy utilization, biosynthesis of secondary metabolites and resistance significantly expanded or are specific to the Chinese tallow tree. These genes would enable the Chinese tallow tree to obtain high adaptability. More genes in fatty acids biosynthesis in its genome, especially for unsaturated fatty acids biosynthesis, and higher expression of these genes in seeds would be attributed to its high content of unsaturated fatty acids. Cyanidin 3-O-glucoside was identified as the major component of anthocyanin in red leaves. All structural genes in anthocyanin biosynthesis show significantly higher expression in red leaves than in green leaves. Transcription factors, seven MYB and one bHLH, were predicted to regulate these anthocyanin biosynthesis genes. Collectively, we provided insight into the polyploidization, high adaptability and biosynthesis of the high content of unsaturated fatty acids in seeds and anthocyanin in leaves for the Chinese tallow tree.
In the spurge family (Euphorbiaceae), there are ca. 300 genera and 8000 species. Numerous plants in this family can be used for generating bioenergy and bioproducts. For example, the rubber tree (*Hevea brasiliensis*) produces a commercially viable amount of natural rubber, and castor bean (*Ricinus communis*) is used to make castor oil, which is a strong laxative. Although plants in the spurge family are important to us, there is a lack of understanding of these plants, especially regarding their genomic and genetic information. Recently, the genome sequences of some plants in the spurge family, including the rubber tree, castor bean, physic nut (*Jatropha curcas*), tung tree (*Vernicia fordii*), and cassava (*Manihot esculenta*), were released. The unigenes of leafy spurge (*Euphorbia esula*) is a perennial weed in the spurge family, was also released with transcriptome assembly. The Chinese tallow tree (CTT) is also an important species in the spurge family and is native to China. The previous Latin name of this plant was *Sapium sebiferum*, and this name was recently changed to *Triadica sebifera* (http://loranorthamerica.org/Triadica_sebifera). The CTT has been cultivated in China for at least 1500 years. Seeds of the CTT can produce up to ca. 40% oil; thus, the CTT is also an important energy oil plant in the spurge family. However, even as knowledge has been obtained, there is no genome resource for this important tree species.

Seeds of the CTT can produce two types of oil, stillingia oil and tallowy fat. Both types of oil are important resources for the food industry and biofuel production. The tallow fat is produced in the white coat of the seeds, while the stillingia oil accumulates in the seed kernels. Moreover, there is a relatively high content of unsaturated fatty acids (FAs), of which ca. 30% is linoleic acid (C18:2) and 50% is linolenic acid (C18:3). In addition to being used for bioenergy production, the CTT is also considered an ornamental tree in China because of its attractive leaf color in autumn and winter. In the central regions of China, the CTT is widely grown, especially in rural areas that comprise mainly mountains. The red, purple, and yellow colors of the CTT leaves attract much attention in autumn and winter. Now there are ca. 20 cultivars in China used as ornamental trees, such as 'Hongzi Jiaren', 'Pudazi', and 'Huixiang' (Personal communication). Therefore, the CTT has become an important tourism resource in some places in the central regions of China, e.g., Luotian and Dawu Counties of Hubei province. In recent years, some elite CTT lines with attractive leaf colors have been introduced to some large cities in China for landscaping. Additionally, the CTT also produces some secondary metabolites that can be used as medicine. For example, nine compounds, shikimic acid, kaempferol, quercetin, isoquercetin, hyperin, kaempferol-3-O-beta-D-glucopyranoside, kaempferol-3-O-beta-D-glucopyranoside, gallic acid, and rutin, were extracted from leaves of the CTT and these chemical compounds are the compositions of some herbal medicines. Though the CTT is an important tree with high economic value, the understanding of this plant is very limited, possibly due to the lack of genomic resources.

The CTT is also an important plant for ecological study. This tree has high adaptability and grows fast and well under various conditions, including canal and stream banks, steep mountain slopes, sandy beaches, and alkaline, saline, or acid soils. Some lines were introduced into the United States of America (USA) in the 18th century and then spread rapidly. Now the CTT is considered a serious invader in the southeastern USA that is overgrowing native plants. Some studies have uncovered the underlying mechanisms for its invasive characteristics. For example, a higher flavonoid concentration in the roots and higher concentration of flavonoids, and lower concentration of tannins were found in introduced CTTs than in native plant populations. However, there are no reports on genomic levels to reveal its invasive mechanisms.

To meet the requirement of high-quality genomic resources of CTT, we assembled a chromosome-scale genome sequence of the CTT. The genome evolution of CTT and some species in the spurge family were analyzed. The cause of the high adaptability of CTT was predicted by a comprehensive investigation of the features of its genome. We also examined the compositions of anthocyanin in red CTT leaves. The biosynthesis pathways for the formation of FAs in seeds and anthocyanin in red leaves were also dissected with the CTT genome sequence. In summary, we obtained a valuable genome resource for the CTT and provided insight into polyploidization, high adaptability, and biosynthesis of the high content of unsaturated FAs in seeds and anthocyanin in leaves for this tree plant.

**Results**

Chromosome-scale genome of the Chinese tallow tree (CTT) revealed that it is a tetraploid. The CTT line used for genome assembly showed red leaves in the field in Luotian County, Hubei Province, China (N 31.05, E 115.66, H 387.7 m) (Fig. 1a, c). Its female flowers grow in the lower part of the inflorescence, while the male flowers grow in the upper part (Fig. 1b). Seeds of the CTT are coated with white wax (Fig. 1d). To estimate the genome size of the CTT, 104.86 Gb PE reads were used to perform 17-bp K-mer analysis (Supplementary Data 1). Interestingly, a total of 3 peaks can be observed for the K-mer distribution (Supplementary Fig. 1). The distribution of K-mer frequency suggests that the genome of the CTT may be tetraploid and that the tetraploid genome size of the CTT is 2.95 Gb. Additionally, there was also an increased frequency after 100-fold coverage, suggesting a high percentage of repeated sequences in the CTT genome.

The predicted percent of repeats in the genome was 64.80%. Subsequently, the genome size of another two CTT lines was also estimated by flow cytometry. By setting the diploid genome size of poplar as 912 Mb, the genome sizes of the two CTT lines were estimated as 2934 and 3010 Mb (Supplementary Data 2). Therefore, the two estimations were almost the same.

To assemble the CTT genome, a total of 42.81 Gb clean HiFi long reads were generated (Supplementary Data 1). The genome was then assembled by using two different software packages, Canu and HiFiAsm. Both software programs produced a total of ca. 2.9 Gb contigs (Supplementary Data 3), while haplotype assembly was also produced by HiFiAsm. Both haplotypes produced by HiFiAsm were ca. 1.4 Gb (Supplementary Data 3). Moreover, the 2.9 Gb contigs produced by Canu were also purged, and the resulting size was ca. 787 Mb (Supplementary Data 3). These results also suggested that the genome of the CTT is tetraploid. Therefore, the 2.9 Gb contigs assembled by Canu (assigned as tetraploid), one of the 1.4 Gb haplotype genome produced by HiFiAsm (assigned as diploid), and 787 Mb purged contigs (assigned as monoploid) were selected for further analyses. The size of N50 for the diploid and monoploid contigs were 8.4 and 29.4 Mb (Supplementary Data 3), respectively. Detailed information on the three types of data is listed in Supplementary Data 3. The three types of assemblies were first assessed for their genome completeness by BUSCO. Clearly, all three datasets showed high completeness (Fig. 1e); however, both tetraploid and diploid genomes showed >85% complete and duplicated BUSCOs, while the monoploid genome showed 77.2% and 20.5% complete and single-copy and complete and duplicated BUSCOs, respectively. These results also indicated that the genome of the CTT is tetraploid.

Second, all three types of assemblies were then anchored by using 127.09 Gb of Hi-C data. Unfortunately, only the monoploids were successfully clustered into large scaffolds. The resulting scaffolds were also subjected to a one-by-one alignment,
and some small scaffolds that were totally covered by larger scaffolds were removed. We finally obtained 25 scaffolds with a total size of 739 Mb. A total of 22 scaffolds were >10 Mb in size, and the Hi-C matrix plot for these scaffolds is shown in Fig. 1f. A previous report suggested that there might be 88 chromosomes in the genome of the CTT (http://floranorthamerica.org/Triadica_sebifera); thus, our result supported this prediction.

Because there were no previous nomenclature of chromosomes for the CTT, we assigned these scaffolds according to their sizes, and this information is listed in Supplementary Data 4. To evaluate the monoploid genome sequence, a total of 18 RNA-Seq data were mapped onto this genome. The results showed that most RNA-Seq data had an average mapping rate of 94.28% and that most mapping rates were >90% (Supplementary Data 5).
Genome annotation. When comprehensively investigating the 739 Mb monoploid genome sequence, the GC level was 32.21%. The repeat sequence of the monoploid genome was then identified; in total, 64.81% of the sequences were annotated as repeats (Table 1). This result highly agrees with the 17-bp K-mer estimation (Supplementary Fig. 1). Of these 64.81% repeated sequences, LTRs accounted for a large proportion (49.4%) (Table 1). In the Euphorbiaceae family, LTRs account for 14.4, 11, and 65.9% of the whole genomic sequence in R. communis, H. brasiliensis, and J. curcas, respectively (Supplementary Data 10). Therefore, the proportion of the LTR sequences varies greatly among different species in the Euphorbiaceae family.

The repeat-masked monoploid genome sequence was then subjected to gene prediction with de novo, homology, and transcript-based approaches. In total, 32,579 genes encoding 43,536 proteins were predicted with average lengths of 6240 and 2097 bp for genes and coding DNA sequences (CDS) (Table 1), respectively. The genome features, including the distribution of GC content, repeat sequences, and genes, are illustrated in Table 1.

To investigate the expanded and contracted genes in the genome of the CTT, the genomes of another 11 plants, namely, A. thaliana, Eucalyptus grandis, H. brasiliensis, J. curcas, L. usitatissimum, M. esculenta, Oryza sativa, P. trichocarpa, R. communis, S. purpurea, and V. vinifera were used for gene cluster analysis. In total, 63,750 orthogroups were identified for all 12 plants. Of these 63,750 orthogroups, 8580 were common for all 12 plants. Of these 63,750 orthogroups, 8580 were common for all 12 plants (Fig. 2a). A total of 398 orthogroups were only found in the CTT genome. These 398 orthogroups including 1781 genes, is considered to be specific for the CTT (Supplementary Data 9). Functional analysis of these 1781 genes showed enrichment in a number of biological processes, such as organonitrogen compound metabolic process, phosphorus metabolic process, carbohydrate metabolic process, organic acid metabolic process, and hydroxy compound metabolic process (Fig. 2b). Clearly, all these biological processes enable plants to obtain more energy and nutrition.

Of the 8580 orthogroups, we identified 151 orthogroups with single-copy genes. A phylogenetic tree was constructed by using these 151 single-copy genes in the 12 plants. Clearly, the five plants in the Euphorbiaceae family are clustered together (Fig. 2c). Moreover, the CTT shows the closest distance to R. communis. The divergent analysis revealed that the CTT and R. communis shared a common ancestor that diverged from the common ancestor of H. brasiliensis, J. curcas, and M. esculenta 36.7 million years ago (MYA) (Fig. 2c). The data also suggested that the CTT and R. communis diverged at 32.8 MYA.

A further investigation of gene cluster analysis revealed numerous genes showing expansion or contraction in the 12 genomes. For the CTT, there were 5287 and 4938 expanded and contracted orthogroups (Fig. 2c), respectively. Of the 5287 expanded orthogroups, a total of 53 orthogroups, including 762 genes, reached the significantly expanded level (P < 0.05) (Supplementary Data 10). Functional analysis of these 762 genes revealed that the top 5 enriched biological processes were macromolecule metabolic process, organic cyclic compound metabolic process, cellular aromatic compound metabolic process, and regulation of cellular process (Fig. 2d). Most of these biological processes are related to the biosynthesis of complex secondary metabolites that would enable plants to obtain higher resistance to various stresses. Additionally, there were numerous genes annotated as resistance (R) genes in 53 orthogroups (Supplementary Data 10). Therefore, the functions of these expanded genes in the CTT genome suggest that these genes may enable the CTT to obtain high stress or disease resistance. To investigate what types of duplications contribute to the formation of these significantly expanded genes, gene duplication types for the whole CTT monoploid genome were analyzed: in total, 3527, 4356, 1005, 1802, and 21,889 for a singleton, dispersed, proximal, tandem, and WGD (Supplementary Data 11), respectively; for the 762 significantly expanded genes, there were 2, 100, 217, 141, and 302, respectively.

Obviously, the proximal and tandem duplications in the 762 genes were significantly increased (P < 0.05, Chi-square test). Usually, proximal and tandem duplications are considered the consequence of tandem duplication of large fragments or single genes. Thus, these analyses suggest that the major driver of expanded genes is tandem duplications in the CTT genome.

---

**Table 1 Statistics of the Chinese tallow tree (CTT) genome assembly and annotation.**

| Features                        | Number | Size/mean size (bp) |
|---------------------------------|--------|---------------------|
| Genomes assembly                |        |                     |
| Raw assembly                    | 897    | 786,950,343         |
| Contig NSD Length               | 29,398,129 |
| Shortest sequence length        | 12,575 |
| Longest sequence length         | 40,396,491 |
| Pseudochromosomes               | 22     | 739,402,270         |
| Gene                            | 32,579 | 6240                |
| Transcript                      | 43,536 | 2097                |
| Coding DNA sequence (CDS)       | 43,536 | 1317                |
| Exon                            | 273,019| 334                 |
| Intron                          | 229,483| 808                 |
| Repeat sequence                 |        |                     |
| Type                            | Number | Length (bp)         | Percent (%) |
| LINEs                           | 7373   | 3,497,379           | 0.47        |
| LTR elements                    | 277,801| 365,708,966         | 49.46       |
| DNA transposons                 | 18,037 | 15,691,900          | 2.12        |
| Rolling-circles                 | 1692   | 1,376,237           | 0.19        |
| Unclassified                    | 225,792| 78,602,547          | 10.63       |
| Low complexity                  | 36,541 | 1,790,042           | 0.24        |
| Simple repeats                  | 193,233| 8,623,656           | 1.17        |
| Total                           | 760,433| 475,290,727         | 64.28       |

This result suggested that the quality of the monoploid genome sequence was sufficient for further analysis. In conclusion, we generated a chromosome-scale monoploid genome sequence for the tetraploid CTT genome.
In the above analyses, the CTT and *R. communis* were predicted to share a common ancestor. To compare the two genomes, the monoploid genome of the CTT was first aligned with the genome of *R. communis*. Clearly, most collinear blocks in the *R. communis* genome matched 2 counterparts (Fig. 3a), suggesting that a genome duplication occurred after the divergence between the CTT and *R. communis*. Moreover, a self-comparison of the monoploid genome of the CTT also suggested that there is another collinear block for *R. communis* duplication occurred after the divergence between the CTT and other four plants in Euphorbiaceae, namely, *R. communis*, *J. curcas*, *H. brasiliensis*, *M. esculenta*, and *O. sativa*. The 4DTv values of the genomes of *R. communis* and *J. curcas* showed one 4DTv peak, while the other four genomes showed at least two peaks. The diploid and monoploid genomes of the CTT showed a similar value (Fig. 3b). Moreover, the self-comparison of the monoploid genome of the CTT also showed a very recent duplication occurring in the common ancestor of these plants and that this duplication occurred outside the Euphorbiaceae family. The unusual peak of the monoploid genome of the CTT at 0.468 may be attributed to confusion of paralogs within its genome during the analysis. Only the genomes of *R. communis* and *J. curcas* showed one 4DTv peak, while the other four genomes showed at least two peaks. The diploid and monoploid genomes of the CTT showed a similar value of Peak 2 (ca. 0.05). Additionally, the other two plants, *H. brasiliensis* and *M. esculenta*, showed values of 0.081 and 0.123 at Peak 2, respectively. These values of Peak 2 suggest that the genomes of all 3 plants were duplicated after they diverged from their common ancestor of *R. communis* or *J. curcas*. Moreover, the diploid genome of the CTT also showed the third unique peak (Peak 1, 0.002), and the value is much less than the value of the other two peaks. This result suggests that there was a very recent duplication of the CTT genome. According to the above analyses, we predicted an evolutionary model for the Euphorbiaceae family (Fig. 3d). In this model, there was a common duplication outside this family, one duplication occurring in *H. brasiliensis* and *M. esculenta* and two additional duplications in the CTT genome. The two duplications in the CTT genome highly support our prediction of the tetraploidy of this plant.

**Dissection of biosynthesis pathways of fatty acids (FAs).** The seed of CTT produces stellinica oil inside the kernel and tallowy fat in the white coat outside the kernel (Fig. 4a). After pollination,
ca. 4 months is required for the development of a mature seed for the CTT (Fig. 4a). The FA compositions of stillingia oil and tallowy fat are totally different. There is ca. 60% palmitinic acid (C16:0) and 30% oleic acid (C18:1) in tallowy fat and 30% linoleic acid (C18:2) in stillingia oil. Thus, there is a high accumulation of saturated and unsaturated FA in the coat and kernel, respectively.

To dissect the biosynthesis pathways of FA in the seeds of the CTT, a homology search of genes involved in this pathway was conducted by using the information released in the Acyl Lipids database (http://aralip.plantbiology.msu.edu/pathways/pathways). Seed samples of four different developmental stages, namely, 1, 3, 5, and 16 weeks after pollination (WAP), were also collected (Fig. 4a). To compare the FA-related genes with other plants, a homology search of genes involved in the FA pathway was conducted for Jatropha curcas, Manihot esculenta, Ricinus communis, CCT diploid, and CCT monoploid. An evolutionary model predicted for some plants in the Euphorbiaceae family. A black dot indicates that a whole genome duplication occurred. The line length roughly indicates divergent time.

Identification of candidate genes for the formation of red leaves. The CTT is an important landscaping tree species and can show charming red leaves in autumn and winter (Fig. 1a). However, the leaves of the CTT before autumn are green (Fig. 5a). Additionally, the young leaves are sometimes light red. To investigate the mechanisms underlying the formation of red leaves, three types of plant materials, namely, green, light red, and red leaves, were collected (Fig. 5a). Anthocyanins in these samples were extracted and then examined by UHPLC-MS/MS. The products were compared with numerous known metabolites of anthocyanins. In positive ionization mode, the product showed parent and daughter ion peaks at m/z 449.11 and 287.06, respectively (Fig. 5b). This pattern is identical to the spectrum of Cyanidin 3-O-glucoside (Cy3-g) in a previous report. Therefore, the major composition of anthocyanin in leaves of the CTT is Cy3-g. The content of this metabolite in green, light red, and red leaves was 0.12 ± 0.0, 9.30 ± 0.2, and 139.26 ± 1.35 ng/g fresh weight (FW), respectively (Fig. 5c).

Because the anthocyanin biosynthesis pathway has been well studied in V. vinifera and A. thaliana, structural genes encoding enzymes in this pathway were used as queries to search the...
genome of the CTT. A total of 23 genes encoding the seven enzymes, chalcone synthase (CHS), chalcone isomerase (CHI), dihydroflavonol 4-reductase (DFR) and leucoanthocyanidin dioxygenase (LDOX) and uridine diphosphate-glucose:flavonoid 3-O-glucosyltransferase (UFGT) and O-methyl transferase (OMT), shown in Fig. 5d were identified (Supplementary Data 13). The expression of these 23 genes in green and red leaves was also calculated and compared by using RNA-Seq data. Interestingly, all nine differentially expressed genes (DEGs) showed higher expression in red leaves than in green leaves (Fig. 5d). Moreover, at least one DEG was identified for each of the seven enzymes, CHS, CHI, F3H, ANS, DFR, LDOX, and UFGT, suggesting that the higher expression of these genes in the red leaves would be attributed to the formation of red color for the leaves of the CTT. Genes of the MYB-bHLH-WD40 (MBW) complex, ELONGATED HYOCOTYL5 (HY5), and CONSTITUTIVE PHOTOMORPHOGENIC1 (COP1) were reported to regulate structural genes in anthocyanin biosynthesis in a number of studies (summarized in Fig. 5d). Thus, DEGs of these genes in the CTT were also identified, except WD40 genes, due to a large number of genes in this gene family in the genome. In total, seven MYB genes and one bHLH gene showed significantly higher expression in red leaves than in green leaves (False discovery rate <0.01, fold change >2.0). Therefore, these 8 transcription factors (TFs) could be the determinant genes for the formation of red leaves for the CTT in autumn and winter.

Discussion

The CTT can grow well in a number of stress conditions in China and is now regarded as an invasive plant in southeastern USA. In this study, we were able to provide the chromosome-scale genome sequence for the CTT. According to the analysis of its genome, the fact that the CTT is a tetraploid would be the most interesting finding. Usually, polyploid plants have more growth vigor than diploid plants. Therefore, the tetraploid CTT genome would be the major diver of its high adaptability character. Additionally, we also found that some genes related to disease resistance, nutrition, energy utilization, and secondary metabolite biosynthesis are specific to the CTT or expand significantly in its genome. These genes...
would also be another driver for the tree’s high adaptability character.

In addition, the tetraploid genome enables CTT to get high adaptability, however, this character increases the difficulty in genome assembly. In this study, we were not able to obtain the chromosome-scale sequence for the CTT diploid genome, and only chromosome-scale sequence for the CTT diploid genome, and only genome assembly. In this study, we were not able to obtain the monoploid genome and the tetraploid genome. Meanwhile, both H. brasiliensis and M. esculenta show the second peak (Peak 2, Fig. 3c); however, their peak values are not the same. Therefore, these data suggest that WGD occurs in these two plants at different times. The initial assembly for monoploid and tetraploid genomes CTT are 2.88 Gb and 787 Mb, respectively, while the genome

![Fig. 5 Analysis of anthocyanin biosynthesis genes in the Chinese tallow (CTT) genome.](image)

(a) Three types of leaf samples with different colors were used for anthocyanin measurement and RNA-Seq analysis. (b) Determination of the major composition of anthocyanin in CTT leaves. Anthocyanins in red leaves were extracted and examined by using a Q-exactive plus ultra-high-performance liquid chromatography instrument with a mass spectrometer (UHPLC/MS) (Thermo Fisher, USA) (upper panel), and the content of Cyanidin 3-O-glucoside (Cy3-g) was measured with high-performance liquid chromatography (HPLC) (Waters Alliance-e2695, Waters Corporation, USA) instrument (lower panel). (c) Contents of Cy3-g in three types of leaf samples. (d) Anthocyanin biosynthesis genes in the CTT genome. The upper panel indicates the regulation and structure genes for anthocyanin synthesis. The bottom panel indicates differentially expressed genes (DEGs) between the red and green leaves for the regulation and structure genes of anthocyanin in the CTT. Columns in this table from left to right are enzyme name, gene ID, Log2FC, red to green leaves reads per kilobase per million mapped reads (RPKM) values for red leaves, RPKM for green leaves, false discovery rate (FDR) for the comparison of red to green RPKM. The full names of these enzymes are CHS chalcone synthase, CHI chalcone isomerase, F3H flavanone 3-hydroxylase, ANS anthocyanidin synthase, DFR dihydroflavonol 4-reductase, LDOX leucoanthocyanin dioxygenase, UFGT uridine diphosphate-glucose:flavonoid 3-O-glucosyltransferase, OMT O-methyl transferase, and bHLH basic helix-loop-helix.

| Name           | Gene ID   | Log2FC | Red   | Green  | RPKM | FDR  |
|----------------|-----------|--------|-------|--------|------|------|
| CHS            | evm.model.Scaffold_16.1044 | 5.37   | 6942.84 | 215.45 | 3.35E-05 |
| CHI            | evm.model.Scaffold_11.1149 | 5.13   | 597.21  | 23.34  | 8.39E-10 |
| F3H            | evm.model.Scaffold_15.1503  | 3.40   | 22.18   | 2.55   | 6.36E-06 |
| DFR            | evm.model.Scaffold_15.498   | 3.84   | 10.77   | 0.82   | 1.49E-04 |
| LDOX           | evm.model.Scaffold_10.259   | 5.02   | 226.32  | 9.66   | 8.41E-10 |
| UFGT           | evm.model.Scaffold_02.775   | 7.05   | 285.35  | 2.68   | 2.72E-20 |
| UFGT           | evm.model.Scaffold_14.299   | 5.46   | 267.93  | 7.84   | 8.49E-10 |
| OMT            | evm.model.Scaffold_01.1171  | 2.61   | 1701.09 | 356.71 | 4.53E-03 |
| OMT            | evm.model.Scaffold_02.379   | 4.51   | 9.06    | 0.43   | 3.77E-04 |
| OMT            | evm.model.Scaffold_19.804   | 4.09   | 5.31    | 0.27   | 1.62E-03 |
| MYB            | evm.model.Scaffold_04.157   | 3.91   | 74.72   | 6.92   | 6.62E-06 |
| MYB            | evm.model.Scaffold_10.1161  | 2.49   | 15.81   | 3.70   | 6.96E-04 |
| MYB            | evm.model.Scaffold_06.1011  | 1.85   | 26.95   | 10.13  | 8.70E-03 |
| MYB            | evm.model.Scaffold_02.77    | 1.63   | 17.04   | 7.18   | 5.69E-03 |
| MYB            | evm.model.Scaffold_22.444   | 2.49   | 15.91   | 3.74   | 8.85E-04 |

suggesting that a common duplication occurs for all five plants. This peak at 0.37 was also detected in V. vinifera in a previous report. Because an ancient duplication is considered to occur in the V. vinifera genome24, plants in Euphorbiaceae also harbor this ancient WGD. Within the Euphorbiaceae family, there are no additional WGD occurs in these two plants. Meanwhile, both H. brasiliensis and M. esculenta show the second peak (Peak 2, Fig. 3c); however, their peak values are not the same. Therefore, these data suggest that WGD occurs in these two plants at different times. The CTT showing small values at Peak 2 and Peak 3 may suggest that WGD events occurred recently (Fig. 3d).
sizes for *J. curcas*, *R. communis*, *H. brasiliensis*, and *M. esculenta* are 264, 336, 1590, and 742 Mb respectively. The genome sizes of these five plants in the Euphorbiaceae family seem not to match the WGD and polyploidization levels in their genome. However, after gaining insight into the size of LTRs in these genomes, it is easy to explain these inconsistencies. LTR takes up ca. 50 and 65% of the whole genomes of the CTT and *H. brasiliensis*, respectively (Table 1), while the other three genomes show lower proportions of LTRs. Thus, these data suggest that a burst of LTRs is the major contributor to genome size.

In seeds of the CTT, two types of oil, stillingia oil and tallowy fat, are accumulated. There is a very high content of unsaturated FAs (C18:2 and C18:3) in stillingia oil. FAD2 and FAD3 are considered the two major enzymes for the biosynthesis of C18:2 and C18:3. In this study, we found two and three genes encoding FAD2 and FAD3 in the CTT monoploid genome (Fig. 4b), respectively. In the Euphorbiaceae family, there was no gene encoding FAD3 in the *J. curcas* and *M. esculenta* genome, while one and two genes in *R. communis* and *H. brasiliensis* genomes, respectively. This data further indicates that there are many more genes encoding FAD3 in tetraploid of the CTT (the number would be 12) than all other species in the Euphorbiaceae family. In further analysis, we found that the 3 FAD3 genes were produced by WGD in the CTT monoploid genome (Supplementary Data 12). In our above analysis, two rounds of WGD occurred in the CTT monoploid, *M. esculenta* and *H. brasiliensis* genomes, and they shared the first round of WGD (Fig. 3c, d). However, there are only 0 and 2 FAD3 genes in *M. esculenta* and *H. brasiliensis* genomes, respectively. The less FAD3 genes in *M. esculenta* and *H. brasiliensis* suggested that FAD3 was not duplicated or deleted in the second round of WGD. In other words, more FAD3 genes in the CTT monoploid genome are mainly attributed to the second round of WGD.

In the dissection of the formation of attractive leaf color in the CTT, we first uncovered Cy3-g was the major composition of anthocyanins in red leaves (Fig. 5a-c). The identification of structural genes for anthocyanin biosynthesis revealed that genes encoding six enzymes were differently expressed in red and green CTT leaves. Especially one gene *evm.model.Scaffold_01.1171* encoding an anthocyanin 3-O-glucosyltransferase showed ca. fivefold higher expression in red than green CTT leaves. The anthocyanin 3-O-glucosyltransferase in Arabidopsis specifically glycosylates the 3-position of the flavonoid C-ring. Because Cy3-g is the major component of anthocyanins in red leaves of CTT, our results are consistent with studies reported in Arabidopsis. This indicates the expression of these genes encoding the 8 TFs in Fig. 5d are activated in the autumn and winter, which suggests these genes might play vital roles in forming red leaves in autumn and winter. However, the mechanisms for the increased expression of these genes are not clear. In Arabidopsis, the gene encoding anthocyanidin 3-O-glucosyltransferase showed increased expression due to a change in light (increased light/dark ratio) under cold conditions. This might also suggest that *evm.model.Scaffold_01.1171* would also respond to light change and cold conditions. The red color of CTT leaves is formed in the late autumn and early winter, thus, photoperiod, light intensity, and the environmental temperature are also changed in this growth condition. Therefore, the activated expression of genes related to anthocyanin biosynthesis in the CTT would also be attributed to the changed conditions of light and temperature.

In conclusion, a chromosome-scale genome was assembled for the CTT in this study, and it provided a valuable genomic resource for this plant. According to the comprehensive analyses of this valuable genome resource, we uncovered novel knowledge to understand the unique characteristics of this plant, such as high adaptability, high content of unsaturated FAs in seeds, and attractive leaf color in autumn and winter. Moreover, we also found two recent WGDs of the CTT genome, and these duplications made it a tetraploid. These results will help us to better utilize this plant in the future.

Methods

**Plant materials.** A CTT line was found in the field of Lutou County, Hubei Province, China (N 31.05°, E 115.66°, H 387.7 m) (Fig. 1a). Seeds of this CTT line were harvested at 1, 3, 5, and 16 WAP and used for RNA isolation in 2020. A total of 5–10 seeds were mixed as one biological sample and three biological samples were collected at each time point. Meanwhile, the red and green leaves were also collected from the same CTT line. The red and green leaves were collected at different time points. Similarly, three biological samples were collected for green and red leaves, respectively. These samples were then used for anthocyanin determination and RNA isolation. This CTT line was then propagated in a woody plant medium (WPM) by using young shoots as the initial explants. Tissue culture plants were then grown in a greenhouse for further experiments.

**Genomic sequencing.** DNA was isolated from the young CTT plants that propagated from the CTT line found in Lutou County by using the modified cetyltrimethylammonium bromide (CTAB) method. RNA was isolated from seeds and red and green leaves of the CTT by using an RNeasy Plant Mini Kit according to the manufacturer’s instructions (DP432, TIANGEN Biotech (Beijing) Co., Ltd., Beijing, China). DNA and RNA that met the required quality were sent for sequencing. For pair-end (PE) read genomic sequencing and RNA sequencing (RNA-Seq), 150-PE libraries were constructed, and these libraries were sequenced by using the MGISEQ-2000 platform (BGI, Shenzhen, China).

**Isoform sequencing (Iso-Seq) of mRNA and long high-fidelity (HiFi) sequencing of genomic DNA were conducted by the Single-molecule real-time (SMRT) Pacific Biosciences (PacBio) platform. The Hi-C library was also constructed using fresh leaves of the young tissue culture plants of the CTT, and it was sequenced on the Illumina HiSeq platform NovaSeq 6000 (Illumina, San Diego, CA). All library construction, sequencing, and raw read filtering was conducted according to the manufacturer’s instructions.

**K-mer estimation and flow cytometry detection.** The 150-bp PE DNA reads of the CTT were filtered by using Trimomatic software with default parameters. All PE reads were also filtered by using this software in further analyses. The clean PE reads were then used for K-mer analysis. The distribution of the 17-bp K-mer was calculated with “kmerFreq” implemented in GCE software with parameter setting as “-k 17 31”. Genome size and repeat content were estimated with “gce” program with parameter setting as “-H 1”. Young leaves of the CTT and the poplar line “NL895” were harvested for flow cytometry detection using a cell analyzer (BD LSRFortessa, BD Biosciences, San Jose, CA, USA). The cell cycle analysis was conducted in our recent studies. The preparation of samples and fluorescence detection were performed according to the manufacturer’s instructions. The monoploid genome size of poplar “NL895” was set as 456 Mb, and it was used as the control sample to estimate the genome size of the CTT.

**Genome assembly.** Two software programs, Canu (Canu v2.1.1) and HiFiasm (HiFiasm v.0.16.1) were used to conduct genome assembly by using HiFi reads. The produced contigs of Canu were purged by using purge_dups (v1.0.1) [https://github.com/dfguan/purge_dups]. The raw contigs produced by Canu and finally purged contigs produced by purge_dups were used as the tetraploid and monoploid genomes of the CTT for further studies, respectively. For the HiFiasm assembly, different parameters were used, and the final parameter was “-l 3” to obtain the diploid genome of the CTT for further studies. All three types of assemblies were assessed by using Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis (The lineage dataset: embryophyta_odb10, Creation date: 2020-10-10, number of species: 50, number of BUSCOs: 16143). The monoploid genome size of Canu was also calculated with “kmerFreq” implemented in GCE software with parameter setting as “-k 17 31”. Genome size and repeat content were estimated with “gce” program with parameter setting as “-H 1”.

**Repeat sequence identification, gene prediction, and functional annotation.** Transposable elements (TEs) were identified by using a combination of homology-based and de novo approaches. The software programs Repeat-ProteinMask, RepeatModeler 2.0, LTR Finder v. 1.0.6, and RepeatMasker v.4.0.5 were used to perform this analysis.

Three strategies were employed for gene prediction, namely, ab initio prediction, homology-based prediction, and transcriptome-based prediction. The repeat-masked chromosome-scale genome of the CTT monoploid and the diploid contigs were used for gene prediction. First, CTT genome sequences were aligned
with protein sequences of five plants, namely, A. thaliana, V. vinifera, O. sativa, P. trichocarpa, and J. regia, and gene structures were predicted with the Exonerate pipelines version exonerate-1.0.0 with default parameters. The whole genome sequence data, including PE short reads, HiFi reads, Hi-C interaction data, and Jacobson reference genome, were used for the integrated analysis with PASA pipelines. The protein sequences of all gene models were aligned with the Swiss-Prot, TrEMBL, TAIR, and Nr databases with the diamond blastp program. All these databases were downloaded on 2021-09-01. The descriptions or Gene Ontology (GO) terms were assigned with an automated assignment of human-readable descriptions (AHRD) pipelines. Functional enrichment analyses were performed with TBtools pipelines. All above procedures and parameters were set according to the manufacturer’s instructions.

Divergence time estimation, gene family expansions and contractions, synteny analysis. Gene families among 12 plants were identified with OrthoFinder (version 2.5.4). Briefly, the longest protein sequence of each gene model of the 12 plants was extracted, and an all-by-all blast was conducted. The orthologous groups were then identified with default parameters implemented in OrthoFinder software. The phylogenetic trees of these 12 plants were constructed with single-copy genes by using the software RAxML (the model parameter was set as PROTGAMMAAUTO). Expansion and extraction of gene families were analyzed using the Computational Analysis of Gene Family Evolution (CAFE) version 3.1. The divergence times for these 12 plants were estimated using the MCMCTree program. The time tree was calibrated by using the known divergence time for some pairs in the 12 plants (http://www.timetree.org/). The synteny between the CTT and castor bean or within the CTT monoploid genome was identified with Augustus (https://github.com/Gaius-Augustus/Augustus). All statistical analyses were performed with the statistical computing programming language R (version 4.0.5). All statistical analyses were conducted with a significance level of $p = 0.05$ ($p < 0.05$). Sample sizes and replicates are demonstrated in the corresponding descriptions.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. The whole genome sequence data, including PE short reads, HiFi reads, Hi-C interaction reads, transcriptome data, and genome files, have been deposited in the NCBI under accession number PRJA813698.

Received: 20 March 2022; Accepted: 20 July 2022; Published online: 04 August 2022

References
1. Zhang, L. et al. Tung tree (Vernicia fordii) genome provides a resource for understanding genome evolution and improved oil production. Genom. Proteom. Bioinf. 17, 558–575 (2019).
2. Liu, J. et al. The chromosome-based rubber tree genome provides new insights into spurge genome evolution and rubber biosynthesis. Mol. Plant 13, 336–350 (2020).
3. Wang, W. Q. et al. Cassava genome from a wild ancestor to cultivated varieties. Nat. Commun. 5, 5110 (2014).
4. Xu, W. et al. Genomic insights into the origin, domestication and genetic basis of agronomic traits of castor bean. Genome Biol. 22, 113 (2021).
5. Chan, A. P. et al. Brz gene family sequence of the oilseed species Ricinus communis. Nat. Biotechnol. 28, 951–953 (2010).
6. Horvath, D. P. et al. Gene space and transcriptome assemblies of leafy spurge (Euphorbia esula) identify promoter sequences, repetitive elements, high-quality markers, and a full-length chloroplast genome. Weed Sci. 66, 355–367 (2018).
7. Eberhardt, T. L., Li, X. B., Shupe, T. F. & Hse, C. Y. Chinese tallow tree (Sapium sebiferum) utilization: characterization of extractives and cell-wall chemistry. Wood Fiber Sci. 39, 319–324 (2007).
8. Zhi, Y. et al. Comparative lipidomics and proteomics of lipid droplets in the mesocarp and seed tissues of Chinese tallow (Triadica sebifera). Front. Plant Sci. 8, 1339 (2017).
9. Bai, M. Y., Zaka, S., Iqbal, S., Sabir, A. W. & Khan, S. A. Sapium sebiferum oil and fat analysis by chromatography. Fett Wiss. Technol. 85, 359–362 (1983).
10. Aitzenmüller, K., Xin, Y. N., Werner, G. & Gronheim, M. High-performance liquid chromatographic investigations of stillingia oil. J. Chromatogr. 603, 165–173 (1992).
11. Wang, H. Q., Zhao, C. Y. & Chen, R. Y. Studies on chemical constituents from leaves of Sapium sebiferum. Zhongguo Zhong Yao Za Zhi 32, 1179–1181 (2007).
12. Pile, L. S. et al. Mechanisms of Chinese tallow (Triadica sebifera) invasion and their management implications - A review. For. Ecol. Manag. 404, 1–13 (2017).
13. Siemann, E. & Rogers, W. E. Genetic differences in growth of an invasive tree species. Ecol. Lett. 4, 514–518 (2001).
14. Huang, W. et al. Resource allocation to defence and growth are driven by different responses to generalist and specialist herbivory in an invasive plant. J. Ecol. 98, 1157–1167 (2010).
15. Wang, Y. et al. Genetic variation in anti-herbivore chemical defences in an invasive plant. J. Ecol. 100, 894–904 (2012).
16. Tian, B. L., Pei, Y. C., Huang, W., Ding, J. Q. & Siemann, E. Increasing flavonoid concentrations in root exudates enhance associations between arbuscular mycorrhizal fungi and an invasive plant. ISME J. 15, 1919–1930 (2021).
17. Tuskal, G. A. et al. The genome of black cottonwood. Populus trichocarpa (Torr. & Gray). Science 313, 1596–1604 (2006).
18. Wang, Y., Li, J. & Paterson, A. H. MScScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. Bioinformatics 29, 1458–1460 (2013).
19. Wang, Y. et al. MScanX: a toolkit for detection and evolutionary analysis of gene synteny and colinearity. Nucleic Acids Res. 40, e49 (2012).
20. Lambert, M. et al. A high-throughput UHPLC-QqQ-MS method for polyphenol profiling in rose wines. Molecules 20, 7890–7914 (2015).
21. Jaakola, L. New insights into the regulation of anthocyanin biosynthesis in fruits. Trends Plant Sci. 18, 477–483 (2013).
22. Chen, Z. J. Molecular mechanisms of polypholid and hybrid vigor. Trends Plant Sci. 15, 57–71 (2010).
23. Washburn, J. D. & Bircher, J. A. Polyploids as a “model system” for the study of heterosis. Plant Reprod. 27, 1–5 (2014).
24. Jallion, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463–466 (2007).
25. Wu, P. Z. et al. Integrated genome sequence and linkage map of physic nut (Jatropha curcas L.), a biodiesel plant. *Plant J.* 81, 810–821 (2021).
26. Kubo, H., Nawa, N. & Lupsea, S. A. Anthocyaninless1 gene of Arabidopsis thaliana encodes a UDP-glucose:flavonoid-3-O-glucosyltransferase. *J. Plant Res.* 120, 445–449 (2007).
27. Jones, P., Messner, B., Nakajima, J. I., Schaffner, A. R. & Saito, K. UGT73C6 and UGT78D1, glycosyltransferases involved in flavonol glycoside biosynthesis in Arabidopsis thaliana. *J. Biol. Chem.* 278, 43910–43918 (2003).
28. Sostamo, A. J., Piippo, M., Allahverdiyeva, Y., Batichkova, N., Aro, E. M. Light has a specific role in modulating Arabidopsis gene expression at low temperature. *BMC Plant Biol.* 8, 13 (2008).
29. Porebski, S., Bailey, L. G. & Baum, R. B. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol content. *Plant Mol. Biol. Rep.* 15, 3–15 (1997).
30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120 (2014).
31. Liu, B. et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quant. Biol.* 35, 62–67 (2015).
32. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736 (2017).
33. Cheng, H. Y., Conception, G. T., Feng, X. W., Zhang, H. W. & Li, H. Halotlotype-de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18, 170–176 (2021).
34. Simar, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212 (2015).
35. Wingett, S. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* 4, 1310 (2015).
36. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 3, 91–98 (2016).
37. Dudchandon, O. et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* 356, 92 (2017).
38. Dudchandon, O. et al. The juicebox assembly tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Preprint at bioRxiv 254779 (2018).
39. Jurca, R. R. RepeatMasker: a database and electronic journal of repetitive elements. *Trends Genet.* 16, 418–420 (2000).
40. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc Bioinformatics*. Chapter 4: Unit 4.10 (2009).
41. Xu, Z. & Wang, H.LTR-FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268 (2007).
42. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580 (1999).
43. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 8, 31 (2005).
44. Butler, A. J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 39, 5654–5666 (2003).
45. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439 (2006).
46. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two efficient tools for the prediction of full-length genes. *Nucleic Acids Res.* 32, W688–W691 (2004).
47. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9, 4 (2008).
48. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* 18, 366–366 (2021).
49. Chen, C. J. et al. TBtools: an integrative toolkit developed for interactive analysis of large phylogenies. *J. Mol. Plant* 12, 1194–1200 (2020).
50. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238 (2019).
51. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313 (2014).
52. De Bié, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271 (2006).
53. Puttick, M. N. MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics* 35, 5321–5322 (2019).
54. Wu, L., D.-P. Wan, H.-L., Zhang, S. & Yu, J. γ-MYN: a new algorithm for estimating Ka and Ks with consideration of variable substitution rates. * Biol. Direct* 4, 20 (2009).
55. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915 (2019).
56. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930 (2014).
57. Varet, H., Brillet-Gueguen, L., Coppee, J. Y. & Dillies, M. A. SARTools: a DESeq2- and EdgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data. *PLoS ONE* 11, e0157022 (2016).
58. Adjé, F. et al. Optimization of anthocyanin, flavonol and phenolic acid extractions from Delonix regia tree flowers using ultrasound-assisted water extraction. *Ind. Crops Products* 32, 439–444 (2010).
59. Wang, H., Asher, K., Zhan, C. & Wang, N. A. Transcriptomic and metabolic analysis of fruit development and identification of genes involved in raffinose and hydrolysable tannin biosynthesis in walnuts. *J. Agric. Food Chem.* 69, 8050–8062 (2021).
60. Jaakola, L. New insights into the regulation of anthocyanin biosynthesis in fruits. *Trends Plant Sci.* 18, 477–483 (2013).
61. Zhao, Q. et al. Expression of structural genes related to anthocyanin biosynthesis of *Vitis amurensis*. *J. For. Res.* 27 (2016).
62. Cantalapiedra, C. P., Ana, H. P., Ivica, L., Peer, B. & Jaime, H. C. eggNOG mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* 38, 5825–5829 (2021).