*DNA Research*, 2021, 28(6), 1–8
https://doi.org/10.1093/dnares/dsab022
Advance Access Publication Date: 19 October 2021
Resource Article: Genomes Explored

Resource Article: Genomes Explored

# High-quality genome assembly of an important biodiesel plant, *Euphorbia lathyris* L.

**Mingcheng Wang [1†], Zhijia Gu[2†], Zhixi Fu[3], and Dechun Jiang [4*]**

[1]Institute for Advanced Study, Chengdu University, Chengdu 610106, China, [2]Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China, [3]College of Life Sciences, Sichuan Normal University, Chengdu 610101, China, and [4]CAS Key Laboratory of Mountain Ecological Restoration and Bioresource Utilization & Ecological Restoration and Biodiversity Conservation Key Laboratory of Sichuan Province, Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu 610041, China

*To whom correspondence should be addressed. Tel. +86 13228121601, Fax: +028-82890288. Email: jiangdc@cib.ac.cn

[†]These authors contributed equally.

Received 14 July 2021; Editorial decision 7 October 2021; Accepted 14 October 2021

## Abstract

Caper spurge, *Euphorbia lathyris* L., is an important energy crop and medicinal crop. Here, we generated a high-quality, chromosome-level genome assembly of caper spurge using Oxford Nanopore sequencing, Illumina sequencing, and Hi-C technology. The final genome assembly was ~988.9 Mb in size, 99.8% of which could be grouped into 10 pseudochromosomes, with contig and scaffold N50 values of 32.6 and 95.7 Mb, respectively. A total of 651.4 Mb repetitive sequences and 36,342 protein-coding genes were predicted in the genome assembly. Comparative genomic analysis showed that caper spurge and castor bean clustered together. We found that no independent whole-genome duplication event had occurred in caper spurge after its split from the castor bean, and recent substantial amplification of long terminal repeat retrotransposons has contributed significantly to its genome expansion. Furthermore, based on gene homology searching, we identified a number of candidate genes involved in the biosynthesis of fatty acids and triacylglycerols. The reference genome presented here will be highly useful for the further study of the genetics, genomics, and breeding of this high-value crop, as well as for evolutionary studies of spurge family and angiosperms.

**Key words:** caper spurge, biodiesel plant, genome assembly, nanopore sequencing, oil metabolism

## 1. Introduction

Energy plays a vital role in today's social and economic development. With the global population explosion and rapid increase in human activities, the world energy demand is rising. By 2040, the world will need twice as much energy as it produces today without continuous improvements in energy efficiency, as predicted by the International Energy Agency (IEA). At present, more than 85% of all the energy we consume is provided by fossil fuels,[1] which mainly includes oil, gas, and coal. Since fossil fuels originate from dead life forms and take millions of years to form, they are finite and non-renewable. The continuous depletion of fossil fuels will most likely lead to a global energy crisis.[2] In addition, fossil fuel combustion is the key contributor to greenhouse gas emissions[3] and is also the major source of air pollution, which poses a serious threat to human and ecosystem health.[4–6] Renewable and alternative energy sources are key to solving the energy crisis, environmental pollution, and climate change.[1]

Biodiesel has attracted considerable attention as a renewable fuel that is produced from vegetable oils and animal fats.[7] It has much lower emissions of greenhouse gases and pollutants than petroleum-based diesel fuel.[8,9] Biodiesel comprised mono-alkyl esters of long-chain fatty acids produced by a catalyzed reaction of triglycerides with an alcohol, such as methanol.[9] Raw materials constitute more than 80% of overall biodiesel production cost.[10] Thus, it is crucial to select a suitable feedstock to achieve sustainable production.[9] The spurge family (Euphorbiaceae) contains a considerable amount of well-known biodiesel plants, such as rubber [*Hevea brasiliensis* (Willd. ex A.Juss.) Müll.Arg.], tung tree [*Vernicia fordii* (Hemsl.) Airy Shaw], castor bean (*Ricinus communis* L.), cassava (*Manihot esculenta* Crantz), physic nut (*Jatropha curcas* L.), and caper spurge (*Euphorbia lathyris* L.). Among these species, caper spurge, a biennial herb native to the Mediterranean area, has been considered as an ideal source of biodiesel because of its high seed productivity, high seed oil content (48 wt%), and low-cost non-edible property.[11,12] It has been confirmed that caper spurge seed can produce high-quality biodiesel at the laboratory scale.[12–14] In addition, caper spurge latex is an important industrial hydrocarbon source due to its high content of triterpenoids (50% of its dry weight).[15] The seeds of caper spurge are utilized as traditional Chinese medicine, and also have great potential for use in modern medicine due to the high content of lathyrane diterpenes.[16–18] However, a better understanding of oil biosynthesis, active ingredients biosynthesis, and other aspects in caper spurge has been hampered by the lack of genome information.

Previous studies have demonstrated that caper spurge is diploid ($2n = 2x = 20$).[19,20] In this study, we assembled and reported a chromosomal-level genome assembly of caper spurge by integrating Oxford Nanopore Technologies (ONT), high-through chromosome conformation capture (Hi-C), and Illumina sequencing. Using this high-quality genome, we further performed genome annotation and comparative genomic analysis with other plant species. We also identified oil metabolism genes within this genome. Our caper spurge genome assembly will provide a valuable genomic resource for genetics, genomics, and the breeding of caper spurge. This genome is also beneficial for investigating the evolutionary history of the spurge family as well as angiosperms.

## 2. Materials and methods

### 2.1 Plant materials and DNA sequencing

Fresh leaves were collected from an adult plant of caper spurge (Fig. 1) growing in Kunming, Yunnan Province, Southwestern China. Total genomic DNA was extracted using the CTAB method[21] for genome sequencing. The library for ONT sequencing was constructed using the 1D ligation sequencing kit (SQK-LSK108, ONT, UK) and sequenced on the ONT PromethION platform. For Illumina sequencing, paired-end libraries with an insertion size of 350 bp were constructed following the manufacturer's instructions and sequenced on an Illumina NovaSeq 6000 platform. For Hi-C analysis, ∼2 g of young leaves were collected and immediately frozen in liquid nitrogen. The Hi-C libraries were then constructed as described previously,[22] including chromatin extraction and digestion, DNA ligation, and purification. The DNA was sheared to a mean fragment size of 350 bp and sequenced on an DNBSEQ-T7 platform (MGI, Shenzhen, China).

### 2.2 RNA sequencing

To aid in genome annotation, fresh tissues of the leaf, stem, and flower were collected for RNA sequencing (RNA-seq). Total RNA was extracted from these tissues, and residual DNA was removed using a DNA-free DNA Removal kit (Thermo Fisher Scientific, USA). The RNA-seq libraries were then constructed and sequenced on an Illumina HiSeq 2500 platform. Low-quality reads were excluded from downstream analysis with Trimmomatic v.0.36.[23]

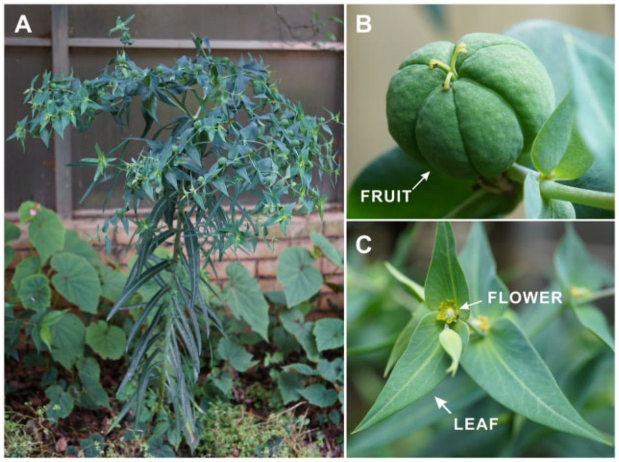

**Figure 1.** Morphological features of caper spurge. (A) Whole plant, (B) fruit, and (C) flower and leaf.

### 2.3 Genome assembly and assessment

Before assembly, the genome size of caper spurge was estimated using a 17-mer frequency distribution analysis of Illumina short reads with Jellyfish v2.2.9.[24] The FAST5 files generated by Nanopore sequencers were converted to FASTQ format using Guppy v3.2.2,[25] and low quality reads with mean_qscore_template < 7 were filtered out. The high-quality ONT reads were then *de novo* assembled using NextDenovo v2.3.1 (https://github.com/Nextomics/NextDenovo) with the main parameters of reads_cutoff = 1k and seed_cutoff = 16444. The assembly process (summarized in Fig. 2) can be divided into four major steps: (i) the ONT subreads were self-corrected and consistent sequences were obtained with NextCorrect; (ii) correlations of consistent sequences were captured and the preliminary genome was assembled with NextGraph; (iii) the assembled contigs were refined using ONT long reads and Illumina short reads with Racon and Nextpolish; and (iv) potential allelic haplotigs were identified and removed by similarity searches with the main parameters of identity = 0.8 and overlap = 0.8. The contigs were anchored into chromosomes using LACHESIS v1.0[26] based on the Hi-C data. Finally, placement and orientation errors exhibiting obvious discrete chromatin interaction patterns were manually adjusted.

The genome assembly quality was assessed by a combination of reads mapping, transcript alignment, Benchmarking Universal Single-Copy Orthologs (BUSCO)[27] analysis, and long terminal repeat (LTR) assembly index (LAI) scores.[28] First, the Illumina short reads were mapped to the assembly using BWA v0.7.17.[29] Second, RNA-seq reads from the leaf, stem, and flower were *de novo* assembled into unigenes using Trinity v2.5.1,[30] and all unigenes were mapped to the genome using BLAT v36.[31] Third, the coverage of complete BUSCO genes was examined by BUSCO v3.0.2 with the Embryophyta odb10 dataset. Fourth, LAI scores were calculated in a sliding window of 3-Mb across the entire genome using LTR_retriever v2.8.[32]

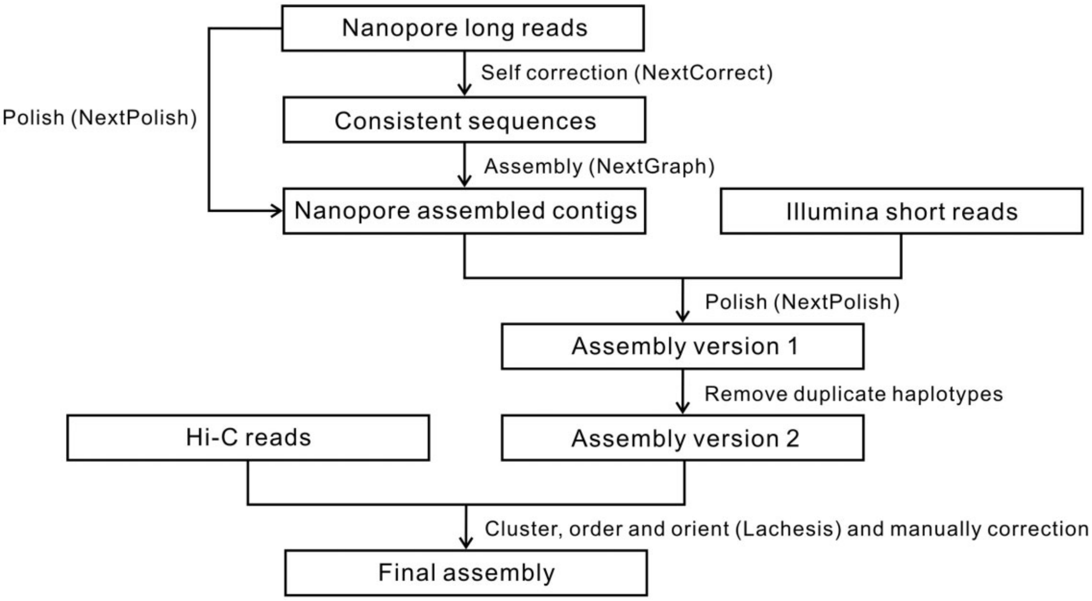

**Figure 2.** Overview of the pipeline used for the assembly of the caper spurge genome.

## 2.4 Genome annotation

RepeatMasker v4.0.7[33] was used to perform homology searches of repeats against the assembly based on a comprehensive database including a *de novo* repeat library generated by RepeatModeler v1.0.11[34] and the 'Viridiplantae' repeat library from the Repbase database v22.11.[35] To investigate the dynamics of LTR-retrotransposons (RTs), we performed *de novo* detection of intact LTR-RTs which are characterized by specific structural signals, including LTR pairs, palindromic motifs, target site duplications (TSDs), primer binding sites (PBSs), polypurine tract (PPT), as well as sites of reverse transcriptase (RT), Integrase (IN) and RNaseH (RH). The locations of LTR pairs, LTR motifs, and TSDs were initially predicted by LTR_Finder v1.06[36] and LTRharvest v1.5.10[37] with the main parameters of min LTR length = 100 bp, max LTR length = 7000 bp, and min LTR similarity = 90%. LTR_Finder was then used to detect PBSs, PPT, and RT domains by build-in aligning and counting modules. RT identification includes a dynamic programming to process frame shift. Finally, LTR_retriever was used to integrate these predictions, remove false positives, and estimate insertion times of intact LTR-RTs.

The repeat-masked genome was then subjected to gene prediction. First, a *de novo* prediction was performed using Augustus v3.2.3[38] with default *Arabidopsis thaliana* parameters. Second, a homology-based prediction was conducted by aligning the protein sequences of *R. communis*,[39] *M. esculenta*,[40] *H. brasiliensis*,[41] *Populus trichocarpa*,[42] and *Arabidopsis thaliana*[43] to the caper spurge genome with TBLASTN v2.2.31+[44] and GeneWise v2.4.1.[45] Third, a comprehensive transcriptome database was built using *de novo* and genome-guided RNA-seq assembly, and gene structures were further predicted using PASA v2.2.0.[46] Finally, all these gene predictions were integrated into a consensus gene set using EVidenceModeler v1.1.1.[47] FPKM (Fragments per Kilobase per Million) was calculated to evaluate the expression level of each gene using Cufflinks v2.2.1.[48]

Gene functions were assigned by aligning the protein sequences to the SwissProt and TrEMBL databases[49] using DIAMOND v0.9.22.[50] Protein motifs and domains were annotated using InterProScan v5.31.[51] Gene ontology (GO) IDs were assigned using Blast2GO v2.5.[52] Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway mapping was achieved using the KEGG Automatic Annotation Server (https://www.genome.jp/kaas-bin/kaas_main).

## 2.5 Comparative genomic analysis

Orthologous and paralogous gene groups were identified using BLASTP[44] and OrthoMCL v2.0.9[53] based on the annotated genes of caper spurge and eight other sequenced plant species, namely five Euphorbiaceae species (*R. communis*, *M. esculenta*, *H. brasiliensis*, *J. curcas*,[54] and *V. fordii*[55]) and three outgroup species (*P. trichocarpa*, *A. thaliana*, and *Vitis vinifera*). Subsequently, single-copy orthogroups among the nine species were selected for species tree construction. For each orthogroup, protein sequences were aligned by MAFFT v7.313,[56] and conserved sites were extracted from the alignments using Gblocks v0.91b.[57] Gene trees were then constructed using RAxML v8.2.11[58] under the PROTGAMMAILGX model with 100 bootstrap replicates. Finally, a coalescent-based species tree was constructed using ASTRAL v4.11.1.[59]

Divergence times among the nine species were estimated using MCMCTREE implemented in PAML v4.9e.[60] The calibration time for the divergence between *V. vinifera* and *P. trichocarpa* (105–115 million years ago, Mya) was obtained from the TimeTree database (http://www.timetree.org). Gene family expansions and contractions were further determined using CAFÉ v3.1.[61]

Polyploidization analysis was performed based on an all-against-all BLASTP search. Syntenic blocks among genomes were identified using MCScanX[62] with default parameters. For each syntenic gene pair, the synonymous substitution rate ($K_s$) was calculated using the 'add_ka_and_ks_to_collinearity.pl' script from MCScanX. To perform comparisons of genome structures between species, positional relationships among inter-chromosomal syntenic gene pairs from the MCScanX outputs were illustrated using the 'Advanced Circos' and 'Dual Systeny Plot' from TBtools v1.098653.[63]

## 2.6 Identification of oil metabolism genes

We identified oil metabolism genes in the caper spurge genome by performing homology searches against the *Arabidopsis* genes that are known to be involved in the biosynthesis of fatty acids (FA) and triacylglycerols (TAGs).[64,65] The candidate genes were further filtered by checking their Enzyme Commission (EC) number.

# 3. Results and discussion

## 3.1 A High-quality genome assembly

Genome sequencing of caper spurge generated a total of 97.2 Gb ONT long reads data, 152.1 Gb Hi-C data, and 48.9 Gb Illumina data (Supplementary Table S1). The ONT long reads were preliminary *de novo* assembled into 130 contigs with a total length of 988.9 Mb, which was close to the genome size estimation by *k*-mer analysis (1.07 Gb; Supplementary Fig. S1). These contigs were then clustered and oriented onto 10 pseudochromosomes based on the Hi-C data (Fig. 3). Approximately 99.8% (986.8 Mb) of the final assembly was assigned to pseudochromosomes, with contig and scaffold N50 values of 32.6 and 95.7 Mb, respectively (Fig. 3B; Table 1; Supplementary Tables S2 and S3).

The high quality and completeness of our caper spurge genome were indicated by the following evidence: (i) 99.4% of the Illumina reads could be successfully aligned to the genome (Supplementary Table S4); (ii) the RNA-seq reads (Supplementary Table S5) were *de novo* assembled into 299,174 unigenes, 88.7% of which had a length coverage of >90% within a single scaffold (Supplementary Table S6); (iii) 1,580 (97.9%) complete BUSCO genes were identified in the genome (Supplementary Table S7); (iv) high LAI scores (17.9 on average) were observed in all pseudochromosomes (Supplementary Fig. S2); and (v) the Hi–C interaction matrices displayed a diagonal pattern for the intra-chromosomal interactions in all pseudochromosomes (Fig. 3A; Supplementary Fig. S3).

## 3.2 Genome annotation

A total of 651.4 Mb (65.9% of the genome) repetitive sequences were identified within the caper spurge genome (Supplementary Table S8), 80.8% (526.1 Mb) of which were transposable elements (TEs). Nearly 72% of the TEs had a divergence rate of < 20% (Supplementary Fig. S4), indicating a recent burst of TEs. As expected, LTR-RTs were the most abundant repeat class, occupying 43.6% of the genome.

Based on the repeat-masked genome, we predicted 36,342 protein-coding genes with mean transcript size of 2,863 bp. Compared to the other five Euphorbiaceae species, caper spurge has the shortest average transcript length and the lowest average exon

**Table 1.** Statistics of the assembly and annotation of the caper spurge genome

| Genome feature | Value |
|---|---|
| Genome assembly | |
| Estimated genome size (by *k*-mer analysis) (Mb) | 1,071.07 |
| Total length (Mb) | 988.90 |
| GC content (%) | 36.82 |
| Contig N50 (Mb) | 32.59 |
| Longest contig (Mb) | 70.11 |
| Contig number | 137 |
| Pseudochromosome number | 10 |
| Scaffold N50 (Mb) | 95.72 |
| Longest scaffold (Mb) | 140.42 |
| Genome annotation | |
| Total length of repeats (Mb) | 651.44 |
| Number of protein-coding genes | 36,342 |
| Mean transcript length (bp) | 2,863.12 |
| Mean coding sequence length (bp) | 1,105.56 |
| Mean exon length (bp) | 245.53 |
| Mean intron length (bp) | 501.76 |
| Functionally annotated genes | 34,391 |

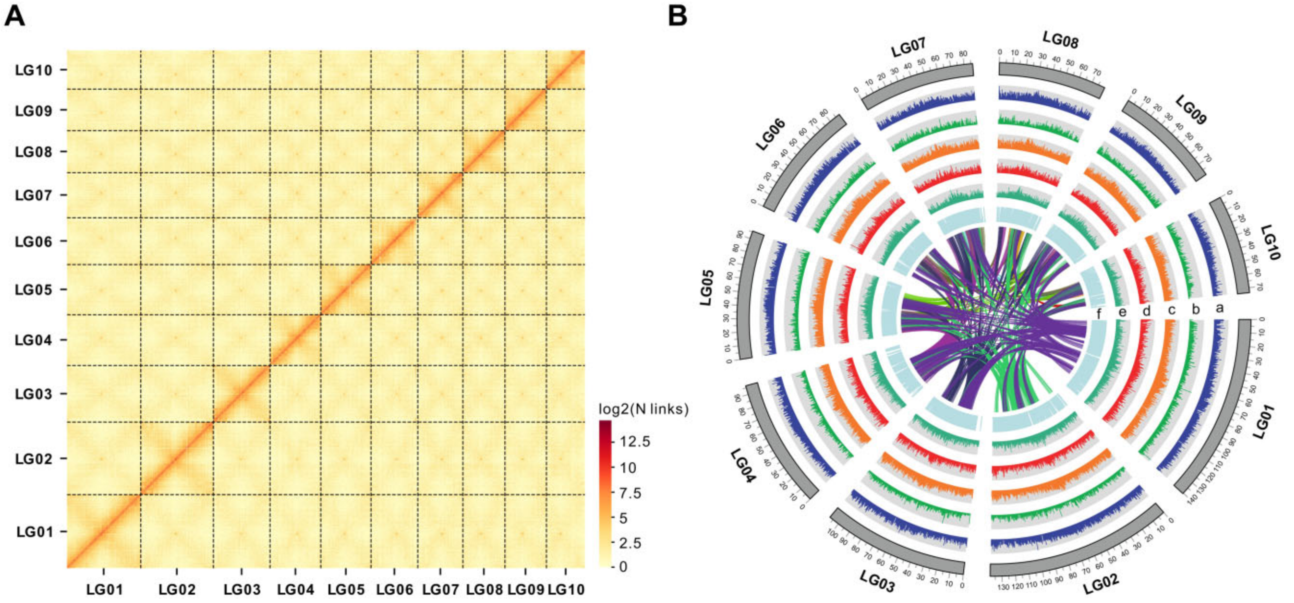

**Figure 3.** Hi–C assisted assembly of caper spurge pseudomolecules. (A) Heatmap showing Hi–C interactions at a resolution of 100 kb. (B) Genome features in non-overlapping windows of 500 kb across the caper spurge genome. Tracks from outside to inside are as follows: (a) GC content, (b) gene density, (c) repeat density, (d) LTR/*Gypsy* density, (e) LTR/*Copia* density, and (f) density of collinear blocks between caper spurge and castor bean.

number per gene (Supplementary Table S9). The predicted caper spurge genes obtained high completeness scores (90.7% complete BUSCOs; Supplementary Table S7). Overall, 34,391 (94.6%) of the protein-coding genes were functionally annotated by at least one public database, and 21,368 (59.5%) genes could be assigned to GO terms (Supplementary Table S10).

## 3.3 Evolution history of caper spurge

OrthoMCL analysis identified 1,247 single-copy orthogroups among caper spurge and eight other plant species. Based on these single-copy genes, we constructed a fully supported species tree (Fig. 4A). Caper spurge was clustered with castor bean, and the divergence between these two species was estimated to have occurred ~50.2 Mya. However, the genome size of caper spurge (1.07 Gb) is approximately three times larger than that of the castor bean (367 Mb). The $K_s$ distribution of orthologs between caper spurge and castor bean showed a major peak ~0.59, which is obviously younger than the two peaks identified from the analysis of paralogs in caper spurge (~0.79) and castor bean (~1.25), indicating that no independent whole-genome duplication (WGD) event had occurred in the caper spurge genome after its split from the castor bean (Fig. 4C). We identified 12,281 intact LTR-RTs with a total length of 109.7 Mb (11.1% of the genome) in the caper spurge genome. In contrast, only 1,565 intact LTR-RTs with a total length of 10.8 Mb (3.2% of the genome) were identified in the castor bean genome. We then estimated the insertion time of intact LTR-RTs in caper spurge. Approximately 48.8% of the intact LTR-RTs were younger than 2 million years, with median insertion times of 1.92 and 1.67 Mya for *Copia* and *Gypsy* elements, respectively (Fig. 4D). These results indicated that recent substantial amplification of LTR-RTs in the caper spurge genome contributed significantly to its genome expansion. Comparative analysis of the genome structures between caper spurge and castor bean revealed that 'LG04' and 'LG08' in the caper spurge genome are highly homologous with 'Chr8' and 'Chr10' in the castor bean genome, respectively (Supplementary Fig. S5). However, highly diverged chromosomal structures were observed among other chromosomes of these two genomes, which might be caused by frequent inter-chromosomal recombinations.

Gene family clustering analysis also identified a total of 7,114 genes that were unique to caper spurge (Fig. 4B), of which 6,346 (89.2%) were supported by gene expression data (FPKM > 0.5)

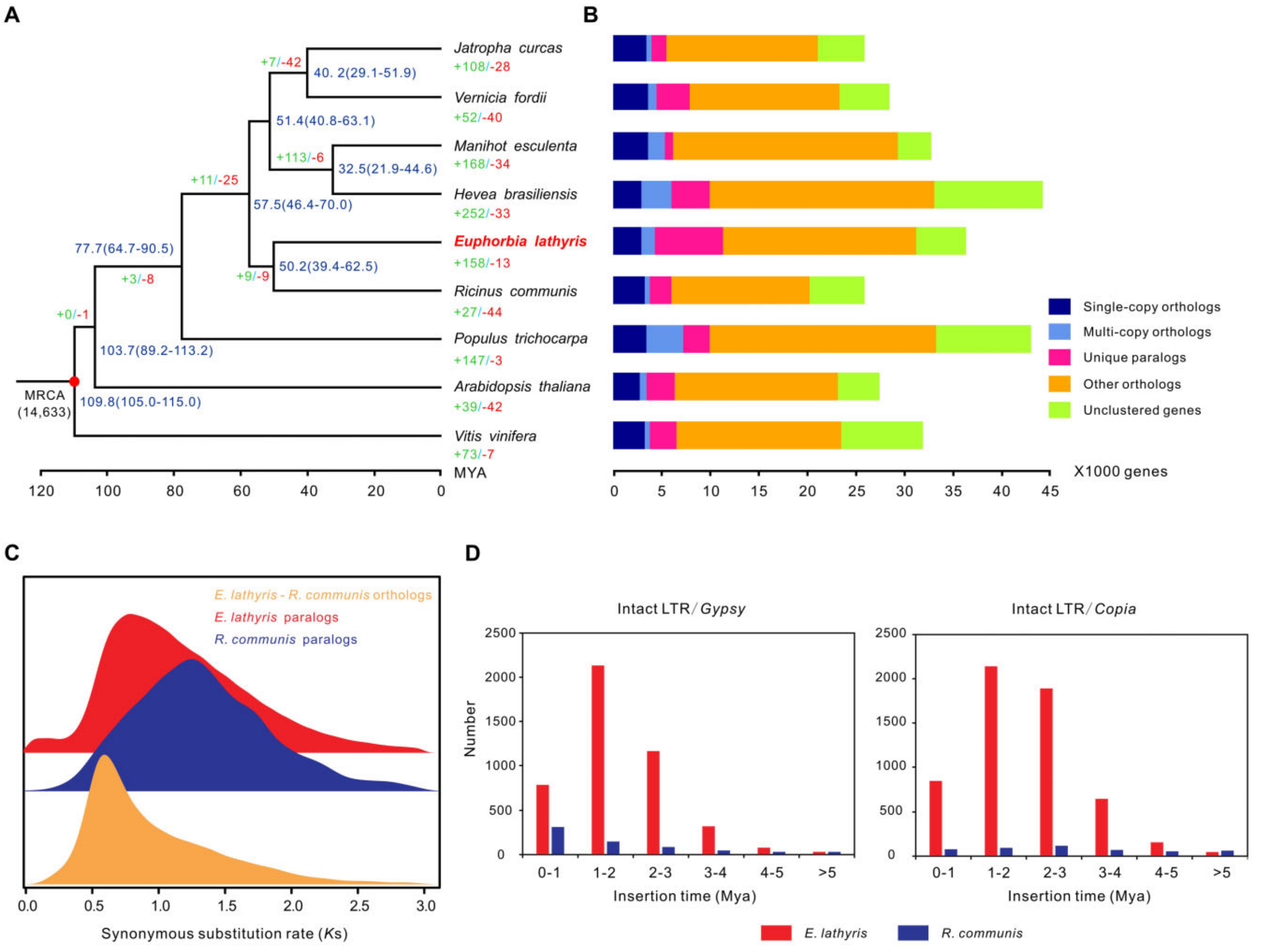

**Figure 4.** Phylogenetic and evolutionary analyses of the caper spurge genome. (A) A species tree based on 1,247 single-copy orthogroups from nine plant species. Gene family expansion and contraction are denoted in green and red numbers, respectively. Numbers on nodes represent the inferred divergence times with 95% confidence intervals. Red dots represent the calibration time of *A. thaliana*–*V. vinifera* divergence. (B) Clusters of orthologous and paralogous gene families in the nine plant genomes. (C) $K_s$ distributions for the whole paranome identified from the whole genome of caper spurge and castor bean. (D) Insertion age distribution of caper spurge intact LTR-RTs in comparison to castor bean.

**Table 2.** Information of genes involved in fatty acid biosynthesis in caper spurge

| Gene | Full name | EC number | Candidate genes |
|---|---|---|---|
| *α-PDH* | Pyruvate dehydrogenase alpha subunit, E1a component of pyruvate dehydrogenase complex | 1.2.4.1 | *Casp00914*, *Casp01950* |
| *β-PDH* | Pyruvate dehydrogenase beta subunit, E1b component of pyruvate dehydrogenase complex | 1.2.4.1 | *Casp18671*, *Casp22878*, *Casp24468*, *Casp30494* |
| *LPD* | Dihydrolipoamide dehydrogenase, E3 component of pyruvate dehydrogenase complex | 1.8.1.4 | *Casp05139*, *Casp05320* |
| *α-CT* | Carboxyltransferase alpha subunit of heteromeric ACCase | 6.4.1.2 | *Casp28361* |
| *BC* | Biotin carboxylase; subunit of heteromeric ACCase | 6.3.4.14 | *Casp30777*, *Casp33262* |
| *BCCP* | Biotin carboxyl carrier protein; subunit of heteromeric ACCase | 6.4.1.2 | *Casp18174* |
| *MCMT* | Malonyl-CoA: ACP malonyltransferase | 2.3.1.39 | *Casp30938* |
| *KASI* | Ketoacyl-ACP synthase I | 2.3.1.41 | *Casp13790*, *Casp14708*, *Casp18048*, *Casp22802*, *Casp25869*, *Casp27208* |
| *KASIII* | Ketoacyl-ACP synthase III | 2.3.1.180 | *Casp06545* |
| *ER* | Enoyl-ACP reductase | 1.3.1.9 | *Casp34837* |
| *KAR* | Ketoacyl-ACP reductase | 1.1.1.100 | *Casp00027*, *Casp01493*, *Casp02324*, *Casp04669*, *Casp05925*, *Casp07320*, *Casp08713*, *Casp09230*, *Casp09232*, *Casp14336*, *Casp16791*, *Casp18406*, *Casp20097*, *Casp20488*, *Casp28998*, *Casp29498*, *Casp30017*, *Casp35223* |
| *HAD* | Hydroxyacyl-ACP dehydrase | 4.2.1.59 | *Casp27479* |
| *LS* | Lipoate synthase | 2.8.1.8 | *Casp14434*, *Casp21423* |
| *LT* | Lipoyltransferase | 2.3.1.181 | *Casp10186*, *Casp32501* |
| *HACPS* | Holo-ACP synthase | 2.7.8.7 | *Casp16258*, *Casp35859* |

and/or functional annotation. Among all the species analysed, caper spurge harboured the highest number of unique paralogs (Fig. 4B). These caper spurge-specific genes have significantly ($P < 2.2e-16$; Wilcoxon rank-sum test) shorter transcript length and lower exon number than other caper spurge genes (Supplementary Fig. S6). The GO enrichment analysis of these unique genes revealed that they were mainly related to 'negative regulation of endosperm development', 'positive regulation of unsaturated fatty acid biosynthetic process', 'fatty acid homeostasis', and 'regulation of embryonic development' (Supplementary Fig. S7). CAFÉ analysis identified 158 significantly ($P < 0.01$) expanded gene families in caper spurge (Fig. 4A), which were highly enriched in 'lateral root morphogenesis', 'flavonol biosynthetic process', 'brassinosteroid metabolic process', and 'response to UV-B' (Supplementary Fig. S8).

### 3.4 Genes involved in oil metabolism

In the caper spurge genome, we identified a total of 46 and 50 genes that were possibly involved in the biosynthesis of FA and TAGs, respectively (Tables 2 and 3). Of these genes, caper spurge has only one copy of the *α-CT*, *BCCP*, *MCMT*, *KASIII*, *ER*, *HAD*, *LPCAT*, *PDCT*, *DAG-CPT*, and *PDAT* genes. It is worth noting that one *DGAT* gene (*Casp18910*) was found to be unique to caper spurge. All these 96 candidate genes were located on pseudochromosomes. These candidate genes provide essential information for future studies of genome editing and molecular breeding of caper spurge, which might improve its seed oil content and biodiesel quality.

## 4. Conclusion

In this study, we present a chromosome-level genome assembly of caper spurge with high accuracy and completeness. This high-quality genome provides valuable insights into the evolutionary history of caper spurge with respect to other Euphorbiaceae species. It is evidenced in the present study that no independent whole-genome duplication event occurred in caper spurge after its split from the castor bean, and recent substantial amplification of LTR-RTs contributed significantly to its genome expansion. This study also identified a number of candidate genes involved in the biosynthesis of fatty acids and triacylglycerols. These candidate genes provide essential information for future studies of genome editing and molecular breeding of caper spurge, which might improve its seed oil content and biodiesel quality. The genome information is also a valuable genetic resource for investigating the active ingredients biosynthesis, gene family evolution, karyotype evolution, and phylogenomics of the spurge family and angiosperms.

## Supplementary data

Supplementary data are available at DNARES online.

## Acknowledgements

We thank Dr Zhiqiang Wang from Chengdu University for his useful suggestions on plant material collection and design of our experiment.

## Funding

This work was funded by the National Natural Science Foundation of China (31900322; 32070410), the Sichuan Science and Technology Bureau (2020YFH0005) and start-up funds provided by Chengdu University.

## Accession numbers

The raw sequence data have been deposited at the NCBI under the BioProject PRJNA745156. RNA-seq data of three tissues are available under the

**Table 3.** Information of genes involved in triacylglycerol biosynthesis in caper spurge

| Gene | Full name | EC number | Candidate genes |
|---|---|---|---|
| *GDPH* | Pyruvate dehydrogenase alpha subunit, E1a component of pyruvate dehydrogenase complex | 1.1.1.8 | *Casp14382*, *Casp14733*, *Casp23079*, *Casp34246* |
| *GPAT* | Glycerol-3-phosphate acyltransferase | 2.3.1.15<br>2.3.1.198 | *Casp00601*, *Casp01161*, *Casp01918*, *Casp12493*, *Casp15506*, *Casp21713*, *Casp23256*, *Casp28812* |
| *LPAAT* | 1-Acylglycerol-3-phosphate acyltransferase | 2.3.1.51 | *Casp10065*, *Casp32526*, *Casp34377*, *Casp34939*, *Casp35048* |
| *LPCAT* | 1-Acylglycerol-3-phosphocholine acyltransferase, lysophospholipid acyltransferase | 2.3.1.23 | *Casp21483* |
| *PDCT* | Phosphatidylcholine: diacylglycerol cholinephosphotransferase | 2.7.8.2 | *Casp13015* |
| *DAG-CPT* | Diacylglycerol choline phosphotransferase | 2.7.8.2 | *Casp29028* |
| *PP* | Phosphatidate phosphatase | 3.1.3.4 | *Casp07550*, *Casp07551*, *Casp07870*, *Casp08150*, *Casp08151*, *Casp35333* |
| *PLA2* | Phospholipase A2 | 3.1.1.4 | *Casp01365*, *Casp24368*, *Casp25591*, *Casp28164* |
| *FAD2* | Oleate desaturase | 1.14.99.33 | *Casp15112*, *Casp23296*, *Casp26601*, *Casp30072*, *Casp30075*, *Casp30077* |
| *FAD3* | Linoleate desaturase | 1.14.99.33 | *Casp15112*, *Casp23296*, *Casp26601*, *Casp30072*, *Casp30075*, *Casp30077* |
| *CK* | Choline kinase | 2.7.1.32 | *Casp14050*, *Casp31645* |
| *CCT* | Choline-phosphate cytidylyltransferase | 2.7.7.15 | *Casp27667*, *Casp27672*, *Casp29053*, *Casp30787* |
| *PDAT* | Phospholipid: diacylglycerol acyltransferase | 2.3.1.43 | *Casp18970* |
| *DGAT* | Acyl-CoA: diacylglycerol acyltransferase | 2.3.1.20 | *Casp00790*, *Casp18910*, *Casp25700*, *Casp33665* |
| *MAGAT* | Monoacylglycerol acyltransferase | 2.3.1.22 | *Casp07919*, *Casp13179*, *Casp15810* |

Sequence Read Archive (SRA) accession numbers SRR15084678, SRR15084679, and SRR15084680. ONT long reads data and Illumina data are available under the SRA accession numbers SRR15533036 and SRR15531403. Hi-C data are available under the SRA accession numbers SRR15533037 and SRR15534417. The genome assembly, annotations, and predicted peptides are available on FigShare at the link: https://doi.org/10.6084/m9.figshare.14909913.v1.

## Conflict of interest

None declared.

## References

1. Abas, N., Kalair, A. and Khan, N. 2015, Review of fossil fuels and future energy technologies, *Futures*, **69**, 31–49.
2. Pimentel, D., Hurd, L.E., Bellotti, A.C., et al. 1973, Food production and the energy crisis, *Science*, **182**, 443–9.
3. Höök, M. and Tang, X. 2013, Depletion of fossil fuels and anthropogenic climate change—a review, *Energy Policy*, **52**, 797–809.
4. Kampa, M. and Castanas, E. 2008, Human health effects of air pollution, *Environ. Pollut.*, **151**, 362–7.
5. Perera, F. 2017, Pollution from fossil-fuel combustion is the leading environmental threat to global pediatric health and equity: solutions exist, *Int. J. Environ. Res. Public Health*, **15**, 16.
6. Ma, X., Zhang, T., Ji, C., et al. 2021, Threats to human health and ecosystem: looking for air-pollution related damage since 1990, *Renew. Sustain. Energy Rev.*, **145**, 111146.
7. Ma, F. and Hanna, M.A. 1999, Biodiesel production: a review, *Bioresour. Technol.*, **70**, 1–15.
8. Van Gerpen, J. 2005, Biodiesel processing and production, *Fuel Process. Technol.*, **86**, 1097–107.
9. Mahlia, T.M.I., Syazmi, Z.A.H.S., Mofijur, M., et al. 2020, Patent landscape review on biodiesel production: technology updates, *Renew. Sustain. Energy Rev.*, **118**, 109526.
10. Haas, M.J., McAloon, A.J., Yee, W.C. and Foglia, T.A. 2006, A process model to estimate biodiesel production costs, *Bioresour. Technol.*, **97**, 671–8.
11. Ayerbe, L., Tenorio, J.L., Ventas, P., Funes, E. and Mellado, L. 1984, *Euphorbia lathyris* as an energy crop—part 1. Vegetative matter and seed productivity, *Biomass*, **4**, 283–93.
12. Wang, R., Hanna, M.A., Zhou, W.W., et al. 2011, Production and selected fuel properties of biodiesel from promising non-edible oils: *Euphorbia lathyris* L., *Sapium sebiferum* L. and *Jatropha curcas* L. *Bioresour. Technol.*, **102**, 1194–9.
13. Zapata, N., Vargas, M., Reyes, J.F. and Belmar, G. 2012, Quality of biodiesel and press cake obtained from *Euphorbia lathyris*, *Brassica napus* and *Ricinus communis*, *Ind. Crops Prod.*, **38**, 1–5.
14. Adeniyi, A.G., Ighalo, J.O., Adeoye, A.S. and Onifade, D.V. 2019, Modelling and optimisation of biodiesel production from *Euphorbia lathyris* using ASPEN Hysys, *SN Appl. Sci.*, **1**, 1–9.
15. Nemethy, E.K., Otvos, J.W. and Calvin, M. 1981, Hydrocarbons from *Euphorbia lathyris*, *Pure Appl. Chem.*, **53**, 1101–8.
16. Lu, J., Li, G., Huang, J., et al. 2014, Lathyrane-type diterpenoids from the seeds of *Euphorbia lathyris*, *Phytochemistry*, **104**, 79–88.
17. Luo, D., Callari, R., Hamberger, B., et al. 2016, Oxidation and cyclization of casbene in the biosynthesis of Euphorbia factors from mature seeds of *Euphorbia lathyris* L, *Proc. Natl. Acad. Sci. USA*, **113**, E5082–9.
18. Zhang, C.Y., Wu, Y.L., Zhang, P., Chen, Z.Z., Li, H. and Chen, L.X. 2019, Anti-inflammatory lathyrane diterpenoids from *Euphorbia lathyris*, *J. Nat. Prod.*, **82**, 756–64.
19. Bowden, W.M. 1940, Diploidy, polyploidy, and winter hardiness relationships in the flowering plants, *Am. J. Bot.*, **27**, 357–71.
20. Jin, M.Y., Ma, C., Wei, W.L. and Feng, S.S. 2007, Karyotype studies of new energy plant *Euphorbia lathyris* L, *Chin. J. Oil Crop Sci.*, **29**, 213–5.
21. Doyle, J.J. and Doyle, J.L. 1987, A rapid DNA isolation procedure for small quantities of fresh leaf tissue, *Phytoch. Bull.*, **19**, 11–5.
22. Louwers, M., Splinter, E., Van Driel, R., De Laat, W. and Stam, M. 2009, Studying physical chromatin interactions in plants using Chromosome Conformation Capture (3C), *Nat. Protoc.*, **4**, 1216–29.

23. Bolger, A.M., Lohse, M. and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, **30**, 2114–20.
24. Marçais, G. and Kingsford, C. 2011, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, *Bioinformatics*, **27**, 764–70.
25. Wick, R.R., Judd, L.M. and Holt, K.E. 2019, Performance of neural network basecalling tools for Oxford Nanopore sequencing, *Genome Biol*., **20**, 1–10.
26. Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. and Shendure, J. 2013, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions, *Nat. Biotechnol*., **31**, 1119–25.
27. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–2.
28. Ou, S., Chen, J. and Jiang, N. 2018, Assessing genome assembly quality using the LTR Assembly Index (LAI), *Nucleic Acids Res*., **46**, e126.
29. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, **25**, 1754–60.
30. Haas, B.J., Papanicolaou, A., Yassour, M., et al. 2013, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, *Nat. Protoc*., **8**, 1494–512.
31. Kent, W.J. 2002, BLAT—the BLAST-like alignment tool, *Genome Res*., **12**, 656–64.
32. Ou, S. and Jiang, N. 2018, LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons, *Plant Physiol*., **176**, 1410–22.
33. Tarailo-Graovac, M. and Chen, N. 2009, Using RepeatMasker to identify repetitive elements in genomic sequences, *Curr. Protoc. Bioinformatics*, **25**, 4–10.
34. Price, A.L., Jones, N.C. and Pevzner, P.A. 2005, De novo identification of repeat families in large genomes, *Bioinformatics*, **21**, i351–8.
35. Jurka, J., Kapitonov, V.V., Pavlicek, A., et al. 2005, Repbase update, a database of eukaryotic repetitive elements, *Cytogenet. Genome Res*., **110**, 462–7.
36. Xu, Z. and Wang, H. 2007, LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons, *Nucleic Acids Res*., **35**, W265–8.
37. Ellinghaus, D., Kurtz, S. and Willhoeft, U. 2008, LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons, *BMC Bioinf*., **9**, 18.
38. Stanke, M., Keller, O., Gunduz, I., et al. 2006, AUGUSTUS: ab initio prediction of alternative transcripts, *Nucleic Acids Res*., **34**, W435–9.
39. Xu, W., Wu, D., Yang, T., et al. 2021, Genomic insights into the origin, domestication and genetic basis of agronomic traits of castor bean, *Genome Biol*., **22**, 113.
40. Bredeson, J.V., Lyons, J.B., Prochnik, S.E., et al. 2016, Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity, *Nat. Biotechnol*., **34**, 562–70.
41. Liu, J., Shi, C., Shi, C.C., et al. 2020, The chromosome-based rubber tree genome provides new insights into spurge genome evolution and rubber biosynthesis, *Mol. Plant*., **13**, 336–50.
42. Tuskan, G.A., Difazio, S., Jansson, S., et al. 2006, The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), *Science*, **313**, 1596–604.
43. Arabidopsis Genome Initiative. 2000, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, *Nature*, **408**, 796–815.
44. Camacho, C., Coulouris, G., Avagyan, V., et al. 2009, BLAST+: architecture and applications, *BMC Bioinf*., **10**, 421.
45. Birney, E., Clamp, M. and Durbin, R. 2004, GeneWise and genomewise, *Genome Res*., **14**, 988–95.
46. Haas, B.J., Delcher, A.L., Mount, S.M., et al. 2003, Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies, *Nucleic Acids Res*., **31**, 5654–66.
47. Haas, B.J., Salzberg, S.L., Zhu, W., et al. 2008, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments, *Genome Biol*., **9**, R7.
48. Trapnell, C., Roberts, A., Goff, L., et al. 2012, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nat. Protoc*., **7**, 562–78.
49. Bairoch, A. and Apweiler, R. 2000, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res*., **28**, 45–8.
50. Buchfink, B., Xie, C. and Huson, D.H. 2015, Fast and sensitive protein alignment using DIAMOND, *Nat. Methods*., **12**, 59–60.
51. Hunter, S., Apweiler, R., Attwood, T.K., et al. 2009, InterPro: the integrative protein signature database, *Nucleic Acids Res*., **37**, D211–5.
52. Conesa, A., Götz, S., García-Gómez, J.M., et al. 2005, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics*, **21**, 3674–6.
53. Li, L., Stoeckert, C.J. and Roos, D.S. 2003, OrthoMCL: identification of ortholog groups for eukaryotic genomes, *Genome Res*., **13**, 2178–89.
54. Chen, M.S., Niu, L., Zhao, M.L., et al. 2020, De novo genome assembly and Hi-C analysis reveal an association between chromatin architecture alterations and sex differentiation in the woody plant *Jatropha curcas*, *GigaScience*, **9**, giaa009.
55. Zhang, L., Liu, M., Long, H., et al. 2019, Tung tree (*Vernicia fordii*) genome provides a resource for understanding genome evolution and improved oil production, *Genomics, Proteomics Bioinf*., **17**, 558–75.
56. Katoh, K. and Standley, D.M. 2013, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Mol. Biol. Evol*., **30**, 772–80.
57. Castresana, J. 2000, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, *Mol. Biol. Evol*., **17**, 540–52.
58. Stamatakis, A. 2014, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics*, **30**, 1312–3.
59. Mirarab, S., Reaz, R., Bayzid, M.S., et al. 2014, ASTRAL: genome-scale coalescent-based species tree estimation, *Bioinformatics*, **30**, i541–8.
60. Yang, Z. 2007, PAML 4: phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol*., **24**, 1586–91.
61. De Bie, T., Cristianini, N., Demuth, J.P. and Hahn, M.W. 2006, CAFE: a computational tool for the study of gene family evolution, *Bioinformatics*, **22**, 1269–71.
62. Wang, Y., Tang, H., DeBarry, J.D., et al. 2012, MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity, *Nucleic Acids Res*., **40**, e49.
63. Chen, C., Chen, H., Zhang, Y., et al. 2020, TBtools: an integrative toolkit developed for interactive analyses of big biological data, *Mol. Plant*., **13**, 1194–202.
64. Beisson, F., Koo, A.J., Ruuska, S., et al. 2003, Arabidopsis genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database, *Plant Physiol*., **132**, 681–97.
65. Li-Beisson, Y., Shorrosh, B., Beisson, F., et al. 2013, Acyl-Lipid Metabolism, *Arabidopsis Book*, **11**, e0161.