

Carthamus tinctorius L. genome sequence provides insights into synthesis of unsaturated fatty acids

Yuanyuan Dong

Ministry of Education, Jilin Agricultural University

Xiaojie Wang

Wenzhou Medical University

Xiuming Liu

Ministry of Education, Jilin Agricultural University

Na Yao

Ministry of Education, Jilin Agricultural University

Yang Jing

Ministry of Education, Jilin Agricultural University

Linna Du

Ministry of Education, Jilin Agricultural University

Xiaowei Li

Ministry of Education, Jilin Agricultural University

Nan Wang

Ministry of Education, Jilin Agricultural University

Weican Liu

Ministry of Education, Jilin Agricultural University

Fawei Wang

Ministry of Education, Jilin Agricultural University

Xiaokun Li

Wenzhou Medical University

Haiyan Li (✉ hyli99@163.com)

Hainan University

Yuanxin Wang

Ministry of Education, Jilin Agricultural University

Research Article

Keywords: Safflower, Genome, Whole genome duplication, Evolutionary history, Fatty acid biosynthesis

Posted Date: October 18th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3367700/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Domesticated safflower (*Carthamus tinctorius* L.) is a widely cultivated edible oil crop. Here, we present the genome assembly for *C. tinctorius* variety *Jihong01*, which was obtained by integrating Oxford Nanopore Technologies (ONT) and BGI-SEQ500 sequencing results. The assembled genome was 1.04 Gb, and consisted of 32,379 protein-coding genes, 97.71% of which were functionally annotated. Safflower had a recent whole genome duplication (WGD) event in evolution history and diverged from sunflower approximately 37.3 million years ago. Comparative genomic and differential gene expression analyses at five stages of seed development revealed that genes encoding fatty acid desaturase 2, fatty acid desaturase 6 play key roles in determining linoleic acid composition. Analyses of fatty acids contents obtained from seeds at different stages revealed FAD2 and FAD6 are crucial for the biosynthesis of LA, the major fatty acids component of seed lipids. Our data can be used in breeding programs to improve quality traits and provide a foundation for further research on the natural properties of safflower.

Introduction

Safflower (*Carthamus tinctorius* L.) is a diploid ($2n = 24$) dicot in the family Compositae (also known as Asteraceae) [1]. It is an annual plant that is predominantly self-pollinated. This herbaceous crop is adapted to hot and dry environments due to its deep root system and xerophytic spines. Therefore, it is widely cultivated in arid and semiarid regions [2]. Safflower is thought to have been domesticated in the fertile region over 4,000 years ago, and it has a long history of cultivation in Asia, the Mediterranean region, Europe, and the Americas [3, 4].

Safflower is mainly an oil crop present, It has been grown as birdseed and as a source of oil for the paint industry [5, 6]. In some areas such as Western Europe, safflower is cultivated as a source of Safflor Yellow (SY) that produced in the floret, used as a natural dyestuff. It is valuable as an edible oil crop because it produces a large amount of oil (approx. 25% oil content in seeds), which is rich in octadecadienoic acids and contained more than 70% linoleic acid (LA) [7]. LA is a compound with a high degree of unsaturation, as one kind of essential PUFA (poly unsaturated fatty acid), which is necessary in diet composition for humans and animals. As one of the oldest sources of oil for humans worldwide, and the main economic traits of cultivated safflower varieties are related to its composition proportion in LA [8, 9]. Although not as a mainstream oilseed crop in current world, it has been cultivated with wide geographic distribution, the species diversity could be an important resource for genetic breeding.

The genetic diversity and natural variations in safflower have been studied using several molecular and analytical methods in recent decades [3, 10]. More recent studies have provided molecular information for safflower including its complete chloroplast genome [11], full-length transcriptome [12], and the locations of 2,008,196 single nucleotide polymorphisms, which were identified from recombinant inbred safflower lines [13]. The results of those studies and others indicated that the genetic architecture and evolution of safflower domestication are complex [14], and this has limited safflower breeding. Genetic breeding programs have used hybridization to breed new cultivars [15] and have characterized the safflower

germplasm using various molecular markers, including expressed sequence tag, inter simple sequence repeat, single nucleotide polymorphism, and simple sequence repeat markers [12, 16–19]. High-quality safflower reference genome can reveal genetic structure and phylogenetic details, as well as biosynthesis processes of bioactive compounds.

To gain insights into the genetic landscape of safflower, we chose the safflower cultivar *Jihong01* for genome assembly, which is a landrace widely cultivated and collected from western China. This economically important landrace is also used extensively for breeding novel safflower strains with improved medicinal properties. In this research, we provide a genome overview that includes details of genome evolution, gene family expansion, and putative genes for unsaturated fatty acids biosynthesis and composition. This reference genome will serve as a platform for investigating the genome background, and for identifying important genes to exploit in genetic breeding programs.

Results

Genome sequencing and assembly

Genomic DNA was extracted from leaves of the “*Jihong01*” safflower variety and sequenced on BGI-SEQ500 and Oxford Nanopore Technologies (ONT) platforms. We obtained 101 Gb short reads data and 130 Gb long reads data in total. By a 17-mer frequency statistics, the genome size of safflower genome was estimated to be 1,066.2 Mb (Figure S1). The primary contigs was assembled by NECAT software (<https://github.com/xiaochuanle/NECAT>) with an N50 of 8.6 Mb. The initial assembly was error-corrected by Pilon [20], and the redundant sequences were removed by HaploMerger2 [21]. After several steps of polishing, total length of the final assembly was 1,061.1 Mb which is close to the estimated genome size (Table S1). To generate a chromosomal-level assembly of the safflower genome, a high-throughput chromosome conformation capture (Hi-C) library was constructed and produced 51 Gb valid reads (47×). 98.82% of the final assembly was anchored to 12 pseudochromosomes with length ranging from 68.06 Mb to 106.66 Mb. GC content of the final assembly was 38.37%. The completeness of our assembly was evaluated by BUSCO (Benchmarking Universal Single Copy Orthologs) [22]. In *embryophyta_odb9* dataset, 90.3% BUSCO genes was complete, 2.4% fragmented and 7.3% missing in the assembly. Also, we used LTR Assembly Index (LAI) [23] values to evaluate the quality of non-coding region in the assembly (Fig. 1). The LAI score of our assembly was 21.94, suggesting a high-quality safflower genome assembly.

Genome annotation

We identified 63.4% of the assembly as repetitive sequences. The proportion is comparatively close to artichoke (58.4%) but lower than sunflower (74.7%) and lettuce (74.2%), which may be the reason why the genome size of lettuce or sunflower is two to three times larger than that of artichoke and safflower. The most abundant transposable elements (TEs) were long terminal repeat (LTR) retrotransposons, accounting for 54.2% of the assembly. Like most plants, *Gypsy* (26.9%) and *Copia* (25.2%) were two dominant LTR super families. Insertion time of LTR was estimated at 1.5 Mya based on the sequence divergence of all LTRs, later than artichoke (Figure S2). DNA transposons covered 7.1% of the assembly.

A total of 32,379 protein-coding genes were predicted using a combination of homology prediction and transcripts supporting. Distributions of gene set parameters had consistent trend with other plants (Figure S3). The gene set covered 93.9% complete BUSCOs of embryophyte BUSCO groups. 97.71% of predicted proteins were functionally annotated against public protein databases (InterPro, UniProt and KEGG). Besides protein-coding genes, we also annotated 131 miRNAs, 998 tRNAs, 3,017 rRNAs and 1,408 snRNAs.

Gene family and phylogeny analysis

Phylogenetic tree was reconstructed based on CDS sequences of 212 single-copy gene families. Plants of Monocotyledons, Rosidae and Asteridae were separated into respective branches and each species were at the reported evolutionary positions. Safflower and sunflower were clustered into one branch in the phylogenetic tree (Fig. 2a). Divergence time of safflower and sunflower was estimated at approximately 37.3 Mya, after the whole genome duplication event at the basal of Compositae family [24]. For comparative analysis, we chose high-quality proteins of 10 oil plant species (oil palm: *Elaeis guineensis*, soybean: *Glycine max*, sunflower: *Helianthus annuus*, Jatropha: *Jatropha curcas*, walnut: *Juglans regia*, flax; *Linum usitatissimum*, olive tree: *Olea europaea*, castor: *Ricinus communis*, sesame: *Sesamum indicum* and maize: *Zea mays*) together with our predicted proteins. All proteins were clustered into 27,600 gene families by OrthoMCL [25] pipeline, within which 495 gene families were safflower-specific. Compared to olive tree and sesame, safflower shared more gene families with sunflower (Fig. 2b), indicating a closer relationship between safflower and sunflower.

Gene family size changes was evaluated by CAFE software [26]. As for safflower, 516 gene families were under size expansion and 2,126 gene families were under size contraction. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment was implemented to the expanded gene families. 225 of 2,751 genes were markedly enriched in fatty acids biosynthesis and metabolism pathways, including linoleic acid metabolism (map00591), alpha-linolenic acid metabolism (map00592) and biosynthesis of unsaturated fatty acids (map01040) (Figure S4). Expansion of gene families involved in fatty acid metabolism especially in unsaturated fatty acids biosynthesis may result in high oil production in safflower.

WGD in safflower genome

As a member of Compositae family, safflower had a recent whole genome duplication (WGD) event in evolution history. We used WGD pipeline [27] to calculate Ks distribution of paralogs in safflower, sunflower, lettuce, artichoke and coffee tree respectively (Fig. 3a). After the γ duplication event in eudicots (peak of coffee tree), safflower had experienced another WGD event which was also found in artichoke and lettuce ($K_s \sim 0.75-1$). Besides, this round of duplication was a triplication event, illustrated by the residues of triplication (Fig. 1) and triplicate synteny blocks between coffee tree and safflower. In Compositae, two rounds of WGD events existed in Heliantheae species, so there were still several 1:2 synteny blocks between safflower and sunflower (Fig. 3b).

Unsaturated fatty acid biosynthesis pathway

Cultivated safflower is an important oil crop rich in unsaturated fatty acids. Safflower oil has the relatively higher polyunsaturated/saturated ratios than other edible oil. We annotated a total of 1,586 genes involved in lipid metabolism in safflower genome. Proteins of 96 genes were functionally enriched in biosynthesis of unsaturated fatty acids pathway (Table S2). In plant, fatty acid desaturases (FAD) catalyse the desaturation reactions of fatty acids. Stearoyl-ACP desaturase (SAD) is a soluble FAD in plastids, transforming stearic acid (C18:0) to oleic acid (C18:1). FAD2 and FAD6 catalyse further desaturation from oleic acid to linoleic acid (C18:2). FAD2 is localised in the endoplasmic reticulum (ER) while FAD6 in the plastids. In a previous study, researchers had isolated 11 members of FAD2 family in safflower [5]. In our assembly, 29 copies of *FAD2* genes were annotated, as well as 4 *SAD* genes and 1 *FAD6* gene (Table S3). *FAD2* genes were amplified by tandem duplication and formed 2 gene clusters located on chromosome 9 and chromosome 11 respectively (Fig. 4a). Phylogenetic analysis of *FAD2* gene family in safflower and other oil crops indicated that *FAD2* gene family were significantly expanded in safflower and sunflower, besides, multi-copy of *FAD2* genes were found in flax genome (Fig. 4b).

We also sequenced transcriptome of seed in five development stages after flowering (DAF6, DAF12, DAF18, DAF24, DAF30) using RNA-seq, each stage with three duplications. Expression patterns of 96 genes in biosynthesis of unsaturated fatty acids pathway were analysed using Mfuzz package. The expression of *FAD2* genes reached peak at DAF18 to DAF24 (Figure S5, cluster 2 and cluster 7), suggesting that the biosynthesis of linoleic acid may be the most active on the five or six day after flowering. Another dominant expression pattern was peak at DAF6 and decreasing with development (Figure S5, cluster 9). Genes in this pattern were mainly acyl-CoA oxidase and very-long-chain enoyl-CoA reductase.

Changes in the fatty acids composition and levels during seed developing stage

Fatty acids are essential for plant growth and development. FAs are synthesized in plastids and to a large extent transported to the endoplasmic reticulum for modification and lipid assembly. Fatty acid composition and contents are the most important indicators to measure the lipid quality. We examined the fatty acid composition of seed storage lipids in developing seeds, the content of fatty acids of seeds in five development stages was measured by LC-MS. Compositional analyses of seed oil revealed that, palmitic acid (C16:0), stearic acid (C18:0), oleic acid (C18:1) and linoleic acid (C18:2) accounted for a predominant proportion of the lipid content (Fig. 5). The contents of oleic acid (C18:1) and linoleic acid (C18:2) were the highest at the maturity stage, C18:1n7, C18:1n9 and C18:2n6 reached 1,776, 1,666, 1,510 µg/g DW, increased by 29.85, 5.42, and 2.56 times when compared to the early stages of grain formation, respectively. According to the data results, C18:3n3 and C18:3n6 were not belong to the high content of PUFA, and their contents decreased first and then increased in the seed development stage. In addition to this, most of the fatty acids were increased in the seeds except that C14:1 was decreased. The composition and content of fatty acids in safflower seeds during seed development indicated that the synthesis and accumulation of polyunsaturated fatty acids C18:2 was the main factor determining the oil quality of safflower seeds.

To gain a better understanding of the relationship between genes and fatty acids species, the Pearson correlation test was performed for the intensity of fatty acids and the expression pattern of genes during the safflower seeds development stage. The correlation between essential genes and fatty acids associated with lipids synthesis was identified.

Here, molecular varieties and composition of these fatty acids (C18:1, C18:2, C18:3) have important effects on the composition and quality of lipids. Our results showed that a total of 28 genes were significantly correlated with C18:1, C18:2 and C18:3 molecular species metabolites that exhibit a Pearson correlation coefficient > 0.7 and $p\text{-value} < 0.05$. Among them, expression pattern of 15 FAD2 genes and 4 FAD6 genes were positively correlated with C18:1n7, C18:1n9 and C18:2n6 composition patterns during seed developing stages (Fig. 6). Among them, two FAD6 genes Cti_chr11_01893 and Cti_chr11_01895 expression patterns showed significant correlation with C18:2 contents during seed development stage. These results indicated that FAD2 and FAD6 genes is responsible for the high proportion of C18:2 in developing safflower seeds.

Discussion

Previous studies on the Compositae have reported genome sequences for *H. annuus* [28], lettuce (*Lactuca sativa*) [29], and globe artichoke (*Cynara cardunculus* var. *scolymus*) [30]. Those studies have provided scientific resources for detailed analyses of genome evolution, functional gene exploration, metabolic pathway construction, and molecular breeding programs. Here, we report the genome sequence of this economically important crop species and describe the key functional genes involved in fatty acid synthesis, so that these new genomic resources are fully available to other relevant researchers.

This research describes a high-quality safflower genome, with a size of 1.04 Gb and 12 pseudochromosomes. There are many karyotypes in the ancestor species of safflower with 10, 11, 12, 22, and 32 pairs of chromosomes, many of which are self-incompatibility species[31]. The current karyotype of cultivated safflower could have originated from wild ancestor *C. tinctorius*, that with $2n = 24$ chromosomes karyotype. This high-quality genome information will be useful for analysing sequences of homologous species, and provides genetic evidence for the nutritional compounds encoded in the *C. tinctorius* genome.

First safflower (*Carthamus tinctorius* L.) cultivar 'Anhui-1' genome was sequencing using PacBio Sequel (Pacific Biosciences) combined with Illumina Hiseq 2500 sequencing platform and focused on biosynthetic pathways of LA and HSYA[32]. Here, in this work, we sequenced and denovo assembled genome of safflower cultivar 'Jihong01' used BGI-SEQ500 and Oxford Nanopore Technologies (ONT) platform and gained another high quality genome assembly results. Meanwhile, we paid more attention to UFA contents during seed development stages. Seed oil fatty acids composition continues to be important traits for safflower breeding [33]. Recent studies on the molecular mechanisms of lipid metabolism have identified a number of genes that form the genetic basis of this trait. The factors determining seed oil composition were found to be complex. In cultivated safflower, lipid metabolic

pathways underlie the natural trait of seed oil content, and candidate genes in the genome were identified to be involved in lipid metabolism and reported before. The unsaturated fatty acid synthesis in ER are important during safflower seed development. Our analyses indicated that the genes involved in fatty acids composition in safflower seeds have undergone expansion during evolution. These analyses may provide clues about the biochemical relevance of lipid composition in seeds.

In terms of the fatty acid composition of oil, there is a lower proportion of LA in sesame (C18:2, 32.95–52.94%) [34] than in safflower (C18:2, 63.9–76.1%) [33, 35, 36]. However, the number of genes involved in fatty acid elongation, biosynthesis, and degradation are similar among the genomes of safflower, sunflower, sesame [34], grape [37], capsicum [38], and *Arabidopsis* [39]. A few key enzymes in the desaturation metabolic pathway regulate unsaturated fatty acid biosynthesis in the ER. Genetic evolution, including genome duplication or gene family expansion, is crucial for generating new gene functions and/or for intensifying pathways [40, 41]. Divergence from an ancestral genome can result in an evolutionary bias towards the production of specific natural products. The creation of duplicates can result in gene expansion, contraction, or loss [42]. *FAD2* gene family encoding enzymes that catalyse linolenic acid biosynthesis have enlarged via whole-genome duplication, which resulted in large-scale gene family expansion. In particular, the gene related to LA synthesis, *FAD2*, has undergone expansion in safflower. The *FAD2* and *FAD6* homologs involved in unsaturated fatty acid biosynthesis showed the highest transcript levels at the DAF18 and DAF24 stage of seed development. Previous studies found that their expression patterns were related to the LA content, with high transcript levels during seed development [33, 43]. These findings provide substantial novel insights into the reasons for the high proportion of LA in safflower oil.

Here, the assembled genome consisted mostly of repeats, coding and non-coding RNAs, and other related sequences. This information allowed us to reconstruct the evolutionary history of safflower, which includes a large-scale safflower-specific whole-genome duplication events. Candidate genes, including *FAD2* and *FAD6*, which encode key functional enzymes related to LA composition, were duplicated in the whole-genome duplication event, and correlation analysis of gene expression and contents of fatty acids also indicated that the biosynthesis of LA was preferred by specific *FAD2* and *FAD6* genes. The sequencing of the safflower genome represents a cornerstone for studying the genetic properties of important metabolic pathways and for advancing molecular breeding programs aimed at producing new safflower cultivars.

Conclusions

The safflower genome assembly represents a cornerstone for future research programs aimed at exploiting the economic properties of safflower, while also considering agricultural constraints and human nutritional needs. The candidate *FAD2* and *FAD6* genes revealed by our integrated approach provide a genetic resources of unsaturated fatty acid biosynthesis and provide a genetic landscape for safflower germplasm utilization.

EXPERIMENTAL PROCEDURES

DNA extracting and sequencing

Safflower variety seedlings (*Jihong01*) collected and planted in an experimental field of Jilin Agriculture University, Changchun City, China were used in this research and stored in Engineering Research Center of Bioreactor and Pharmaceutical Development, Ministry of Education. High-quality genomic DNA was extracted using the Qiagen kit. Subsequently, the extraction process focused on selecting large-size fragments, which were accomplished through automated gel electrophoresis. Following this, the DNA underwent treatment involving the end-repair/dA tailing module, and subsequently, it was ligated to an adaptor using the ONT 1D ligation sequencing kit. The prepared library was loaded onto flow cells and subjected to sequencing using the Nanopore PromethION platform.

Genome assembly

K-mer frequency was calculated by Jellyfish v2.26 [44] and the genome size was estimated using GenomeScope [45]. The initial contigs were assembled by NECAT with default parameters using Nanopore reads longer than 5 kb. The initial assembly was error-corrected by Pilon with short reads. Size of the initial assembly was a little larger than estimated, so we used HaploMerger2 software to remove redundant contigs in the initial assembly. Then the assembly was error-corrected again. Reads generated by Hi-C library was filtered strictly by HiC-Pro pipeline [46] to discard invalid reads pairs. We used Juicer [47] and 3D *de novo* assembly (3D-DNA) pipeline [48] to anchor contigs to pseudochromosomes.

Repeat annotation and LTR insertion time

Repetitive sequences were annotated using a combination of *de novo* and homology strategy. We used RepeatModeler [49], LTR_FINDER [50] and TRF [51] software for *de novo* repeats identification based on repetitive sequences features. Then RepeatMasker and RepeatProteinMask were used to annotate transposon elements based on RepBase. LTR insertion time was estimated based on the divergence of LTR pairs. Intact LTRs were identified using LTR_FINDER software in the four Compositae genomes. Then we used MUSCLE [52] to align LTR pairs and distmat to calculate K values under the Kimura two-parameter model. With the K values of LTR pairs, the insertion time was calculated using formula $T = K/2r$, where r was the rate of nucleotide substitution and set as 7×10^{-9} per site per generation here [53, 54].

Genes prediction and function annotation

Protein-coding genes were predicted based on both homolog proteins and transcripts. Proteins of seven plants (*Arachis hypogaea*, *Brassica napus*, *Glycine max*, *Helianthus annuus*, *Lactuca sativa*, *Ricinus communis* and *Sesamum indicum*) was downloaded from NCBI database. We first aligned these proteins with the assembly using BLAT [55], then the alignment was input to Genewise [56] to get homology annotations. RNA-seq reads were mapped to the assembly using HISAT2 [57] and transcripts were assembled using StringTie [58]. All of the evidences were integrated to the final protein-coding gene set

by GLEAN [59]. Protein-coding genes were assessed for conserved protein domains in the ProDom, ProSiteProfiles, SMART, PANTHER, Pfam, PIRSF and ProSitePatterns databases using InterProScan [60]. Also, amino acid sequences were aligned to the following protein databases: Swiss-Prot, TrEMBL, Kyoto Encyclopaedia of Genes and Genomes (KEGG) using BLASTP (e-value < 1e-5) for function annotation.

Comparative analysis

All protein-coding gene sequences of the 10 oil plant species (*Elaeis guineensis*, *Glycine max*, *Helianthus annuus*, *Jatropha curcas*, *Juglans regia*, *Linum usitatissimum*, *Olea europaea*, *Ricinus communis*, *Sesamum indicum* and *Zea mays*) was downloaded from the NCBI database. The longest transcript of each gene without frame shift or internal termination was selected and translated into amino acid sequences for subsequent analyses. First, we used BLASTP for an all-to-all proteins alignment under the e-value of 1e-5. Then the ortholog genes were clustered into groups using OrthoMCL with a Markov inflation index of 1.5 and a maximum e-value of 1e-5. One-to-one single-copy ortholog groups were joined to a super-gene and aligned using MUSCLE [52]. Phylogenetic tree was constructed using RAXML [61] with *Z. mays* and *E. guineensis* as outgroups, under GTR + Optimization of substitution rates and GAMMA model of rate heterogeneity. The divergence time was estimated using MCMCTREE in PAML packages [62] based on HKY85 model and correlated rates molecular clock model. The size changes of each gene families were calculated by CAFE [26] with the random birth and death model.

WGD events and synteny

We used WGD software [27] to identify WGD events in the evolutionary process of safflower, sunflower, lettuce, artichoke and coffee tree. Synteny blocks between safflower and sunflower and between safflower and coffee tree were identified and displayed using jcv packages [63].

Identification of unsaturated fatty acids biosynthesis genes

We used BLASTP to identified safflower FAD2 based on amino acids sequences of *Arabidopsis thaliana* FAD2 (NP_187819.1) with e-value of 1e-10.

Fatty acids content analysis

Dry seed samples were collected at five different stages of seed development for the purpose of analyzing fatty acid content and composition. Methyl esters (FAME) were prepared from each seed sample. Subsequently, quantitative analysis of fatty acids within these FAME samples was conducted utilizing Gas Chromatography-Mass Spectrometry (GC-MS) through the Agilent Technologies 6890N/5975B system. The methods employed for this analysis were in accordance with those described by Ecker et al.[64]. The fatty acids present, including saturated fatty acids (SFAs), monounsaturated fatty acids (MUFAs), and polyunsaturated fatty acids (PUFAs), were subjected to quantitative calculations. These calculations were carried out based on a standard fatty acid methyl ester mix.

Declarations

Funding

This research was funded by The Science and Technology Development Project of Jilin province (20210203179SF), Science and Technology Research Project of the Education Department of Jilin Province (JJKH20210617KJ).

Author contributions

Y.D., X.W., X.L., and H.L. performed some data analyses. J.Y., N.Y. and L.D. managed samples and tissues. Y.W. prepared materials and uploaded data. Y.D., X.W., X.L., and J.Y. performed some data analyses and prepared graphics. N.Y., X.L., N.W., X.L., and W.L., prepared the libraries. H.L., and X.L. assisted in data analysis and in the overall design of the project. F.W., H.L., and X.L. developed the figure of the study and assisted with manuscript preparation. Y.D., X.W., X.L., J.Y., F.W., and H.L. wrote the manuscript. All of the authors reviewed and approved the final manuscript.

Availability of data and materials

The raw sequence reads were deposited in China National GeneBank DataBase (CNGB db) under Project No. CNP0004859 and CNP0004861.

Ethics approval and consent to participate

No specific were required for voucher specimens for this study. Voucher specimens were prepared and deposited at the Jilin agricultural University. Yuanyuan Dong and Xiaojie Wang undertook the identification work of the plant material. The authors have complied with all relevant institutional and national guidelines and legislation in experimental research and field studies on plants, including the collection of plant materials for this study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Chapman MA, Burke JM: **DNA sequence diversity and the origin of cultivated safflower (*Carthamus tinctorius* L.; Asteraceae).** *BMC Plant Biol* 2007, **7**:60.
2. Dajue L, Mündel H-H: **Safflower, *Carthamus Tinctorius* L**, vol. 7: Bioversity International; 1996.
3. Panahi B, Ghorbanzadeh Neghab M: **Genetic characterization of Iranian safflower (*Carthamus tinctorius*) using inter simple sequence repeats (ISSR) markers.** *Physiol Mol Biol Plants* 2013,

4. McPherson MA, Good AG, Topinka AK, Yang RC, McKenzie RH, Cathcart RJ, Christianson JA, Strobeck C, Hall LM: **Pollen-mediated gene flow from transgenic safflower (*Carthamus tinctorius* L.) intended for plant molecular farming to conventional safflower.** *Environ Biosafety Res* 2009, **8**(1):19-32.
5. Cao S, Zhou XR, Wood CC, Green AG, Singh SP, Liu L, Liu Q: **A large and functionally diverse family of Fad2 genes in safflower (*Carthamus tinctorius* L.).** *BMC Plant Biol* 2013, **13**:5.
6. Matthaus B, Ozcan MM, Al Juhaime FY: **Fatty acid composition and tocopherol profiles of safflower (*Carthamus tinctorius* L.) seed oils.** *Nat Prod Res* 2015, **29**(2):193-196.
7. Rapson S, Wu M, Okada S, Das A, Shrestha P, Zhou XR, Wood C, Green A, Singh S, Liu Q: **A case study on the genetic origin of the high oleic acid trait through FAD2-1 DNA sequence variation in safflower (*Carthamus tinctorius* L.).** *Front Plant Sci* 2015, **6**:691.
8. Li R, Beaudoin F, Ammah AA, Bissonnette N, Benchaar C, Zhao X, Lei C, Ibeagha-Awemu EM: **Deep sequencing shows microRNA involvement in bovine mammary gland adaptation to diets supplemented with linseed oil or safflower oil.** *BMC Genomics* 2015, **16**:884.
9. Kazuma K, Takahashi T, Sato K, Takeuchi H, Matsumoto T, Okuno T: **Quinochalcones and flavonoids from fresh florets in different cultivars of *Carthamus tinctorius* L.** *Biosci Biotechnol Biochem* 2000, **64**(8):1588-1599.
10. Sehgal D, Rajpal VR, Raina SN, Sasanuma T, Sasakuma T: **Assaying polymorphism at DNA level for genetic diversity diagnostics of the safflower (*Carthamus tinctorius* L.) world germplasm resources.** *Genetica* 2009, **135**(3):457-470.
11. Lu C, Shen Q, Yang J, Wang B, Song C: **The complete chloroplast genome sequence of Safflower (*Carthamus tinctorius* L.).** *Mitochondrial DNA A DNA Mapp Seq Anal* 2016, **27**(5):3351-3353.
12. Chen J, Tang X, Ren C, Wei B, Wu Y, Wu Q, Pei J: **Full-length transcriptome sequences and the identification of putative genes for flavonoid biosynthesis in safflower.** *BMC Genomics* 2018, **19**(1):548.
13. Bowers JE, Pearl SA, Burke JM: **Genetic Mapping of Millions of SNPs in Safflower (*Carthamus tinctorius* L.) via Whole-Genome Resequencing.** *G3 (Bethesda)* 2016, **6**(7):2203-2211.
14. Pearl SA, Bowers JE, Reyes-Chin-Wo S, Micheltore RW, Burke JM: **Genetic analysis of safflower domestication.** *BMC Plant Biol* 2014, **14**:43.
15. Mayerhofer M, Mayerhofer R, Topinka D, Christianson J, Good AG: **Introgression potential between safflower (*Carthamus tinctorius*) and wild relatives of the genus *Carthamus*.** *BMC Plant Biol* 2011, **11**:47.
16. Pearl SA, Burke JM: **Genetic diversity in *Carthamus tinctorius* (Asteraceae; safflower), an underutilized oilseed crop.** *Am J Bot* 2014, **101**(10):1640-1650.
17. Lee GA, Sung JS, Lee SY, Chung JW, Yi JY, Kim YG, Lee MC: **Genetic assessment of safflower (*Carthamus tinctorius* L.) collection with microsatellite markers acquired via pyrosequencing method.** *Mol Ecol Resour* 2014, **14**(1):69-78.

18. Chapman MA, Hvala J, Strever J, Matvienko M, Kozik A, Michelmore RW, Tang S, Knapp SJ, Burke JM: **Development, polymorphism, and cross-taxon utility of EST–SSR markers from safflower (*Carthamus tinctorius* L.).** *Theoretical and Applied Genetics* 2009, **120**(1):85-91.
19. Yang Y-X, Wu W, Zheng Y-L, Chen L, Liu R-J, Huang C-Y: **Genetic diversity and relationships among safflower (*Carthamus tinctorius* L.) analyzed by inter-simple sequence repeats (ISSRs).** *Genetic Resources and Crop Evolution* 2007, **54**(5):1043-1051.
20. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK *et al*: **Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.** *PLoS One* 2014, **9**(11):e112963.
21. Huang S, Kang M, Xu A: **HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly.** *Bioinformatics* 2017, **33**(16):2577-2579.
22. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics* 2015, **31**(19):3210-3212.
23. Ou S, Chen J, Jiang N: **Assessing genome assembly quality using the LTR Assembly Index (LAI).** *Nucleic Acids Res* 2018, **46**(21):e126.
24. Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH: **Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years.** *Mol Biol Evol* 2008, **25**(11):2445-2455.
25. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**(9):2178-2189.
26. De Bie T, Cristianini N, Demuth JP, Hahn MW: **CAFE: a computational tool for the study of gene family evolution.** *Bioinformatics* 2006, **22**(10):1269-1271.
27. Zwaenepoel A, Van de Peer Y: **wgd-simple command line tools for the analysis of ancient whole-genome duplications.** *Bioinformatics* 2019, **35**(12):2153-2155.
28. Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, Lelandais-Briere C, Owens GL, Carrere S, Mayjonade B *et al*: **The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution.** *Nature* 2017, **546**(7656):148-152.
29. Reyes-Chin-Wo S, Wang Z, Yang X, Kozik A, Arikait S, Song C, Xia L, Froenicke L, Lavelle DO, Truco MJ *et al*: **Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce.** *Nat Commun* 2017, **8**:14953.
30. Scaglione D, Reyes-Chin-Wo S, Acquadro A, Froenicke L, Portis E, Beitel C, Tirone M, Mauro R, Lo Monaco A, Mauromicale G *et al*: **The genome sequence of the outbreeding globe artichoke constructed de novo incorporating a phase-aware low-pass sequencing strategy of F1 progeny.** *Sci Rep* 2016, **6**:19427.
31. Garnatje T, Garcia S, Vilatersana R, Vallès J: **Genome size variation in the genus *Carthamus* (Asteraceae, Cardueae): systematic implications and additive changes during allopolyploidization.** *Ann Bot* 2006, **97**(3):461-467.

32. Wu Z, Liu H, Zhan W, Yu Z, Qin E, Liu S, Yang T, Xiang N, Kudrna D, Chen Y *et al*: **The chromosome-scale reference genome of safflower (*Carthamus tinctorius*) provides insights into linoleic acid and flavonoid biosynthesis.** *Plant Biotechnol J* 2021, **19**(9):1725-1742.
33. Sabzalain MR, Saeidi G, Mirlohi A: **Oil content and fatty acid composition in seeds of three safflower species.** *Journal of the American Oil Chemists' Society* 2008, **85**(8):717-721.
34. Wei X, Liu K, Zhang Y, Feng Q, Wang L, Zhao Y, Li D, Zhao Q, Zhu X, Zhu X *et al*: **Genetic discovery for oil production and quality in sesame.** *Nat Commun* 2015, **6**:8609.
35. Bozan B, Temelli F: **Chemical composition and oxidative stability of flax, safflower and poppy seed and seed oils.** *Bioresource technology* 2008, **99**(14):6354-6359.
36. Ma DW, Wierzbicki AA, Field CJ, Clandinin MT: **Preparation of conjugated linoleic acid from safflower oil.** *Journal of the American Oil Chemists' Society* 1999, **76**(6):729-730.
37. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C *et al*: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449**(7161):463-467.
38. Kim S, Park M, Yeom S-I, Kim Y-M, Lee JM, Lee H-A, Seo E, Choi J, Cheong K, Kim K-T: **Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species.** *Nature genetics* 2014, **46**(3):270-278.
39. Arabidopsis Genome I: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**(6814):796-815.
40. Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, Liberles DA: **Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms.** *J Exp Zool B Mol Dev Evol* 2007, **308**(1):58-73.
41. Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, Seo E, Choi J, Cheong K, Kim KT *et al*: **Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species.** *Nat Genet* 2014, **46**(3):270-278.
42. Ibarra-Laclette E, Lyons E, Hernandez-Guzman G, Perez-Torres CA, Carretero-Paulet L, Chang TH, Lan T, Welch AJ, Juarez MJ, Simpson J *et al*: **Architecture and evolution of a minute plant genome.** *Nature* 2013, **498**(7452):94-98.
43. Li Y, Beisson F, Pollard M, Ohlrogge J: **Oil content of *Arabidopsis* seeds: the influence of seed anatomy, light and plant-to-plant variation.** *Phytochemistry* 2006, **67**(9):904-915.
44. Marcais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics* 2011, **27**(6):764-770.
45. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC: **GenomeScope: fast reference-free genome profiling from short reads.** *Bioinformatics* 2017, **33**(14):2202-2204.
46. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E: **HiC-Pro: an optimized and flexible pipeline for Hi-C data processing.** *Genome Biol* 2015, **16**:259.

47. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL: **Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments.** *Cell Syst* 2016, **3**(1):95-98.
48. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP *et al*: **De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds.** *Science* 2017, **356**:92-95.
49. **RepeatModeler Open-1.0** [<http://www.repeatmasker.org>]
50. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons.** *Nucleic Acids Res* 2007, **35**(Web Server issue):W265-268.
51. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**(2):573-580.
52. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**(5):1792-1797.
53. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H *et al*: **The Arabidopsis lyrata genome sequence and the basis of rapid genome size change.** *Nat Genet* 2011, **43**(5):476-481.
54. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M: **The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana.** *Science* 2010, **327**(5961):92-94.
55. Kent WJ: **BLAT-the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656-664.
56. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res* 2004, **14**(5):988-995.
57. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements.** *Nat Methods* 2015, **12**(4):357-360.
58. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** *Nat Biotechnol* 2015, **33**(3):290-295.
59. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM: **Creating a honey bee consensus gene set.** *Genome Biol* 2007, **8**(1):R13.
60. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G *et al*: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30**(9):1236-1240.
61. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30**(9):1312-1313.
62. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24**(8):1586-1591.
63. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH: **Synteny and Collinearity in Plant Genomes.** *Science* 2008, **320**:486-488.
64. Ecker J, Scherer M, Schmitz G, Liebisch G: **A rapid GC-MS method for quantification of positional and geometric isomers of fatty acid methyl esters.** *J Chromatogr B Analyt Technol Biomed Life Sci* 2012,

Figures

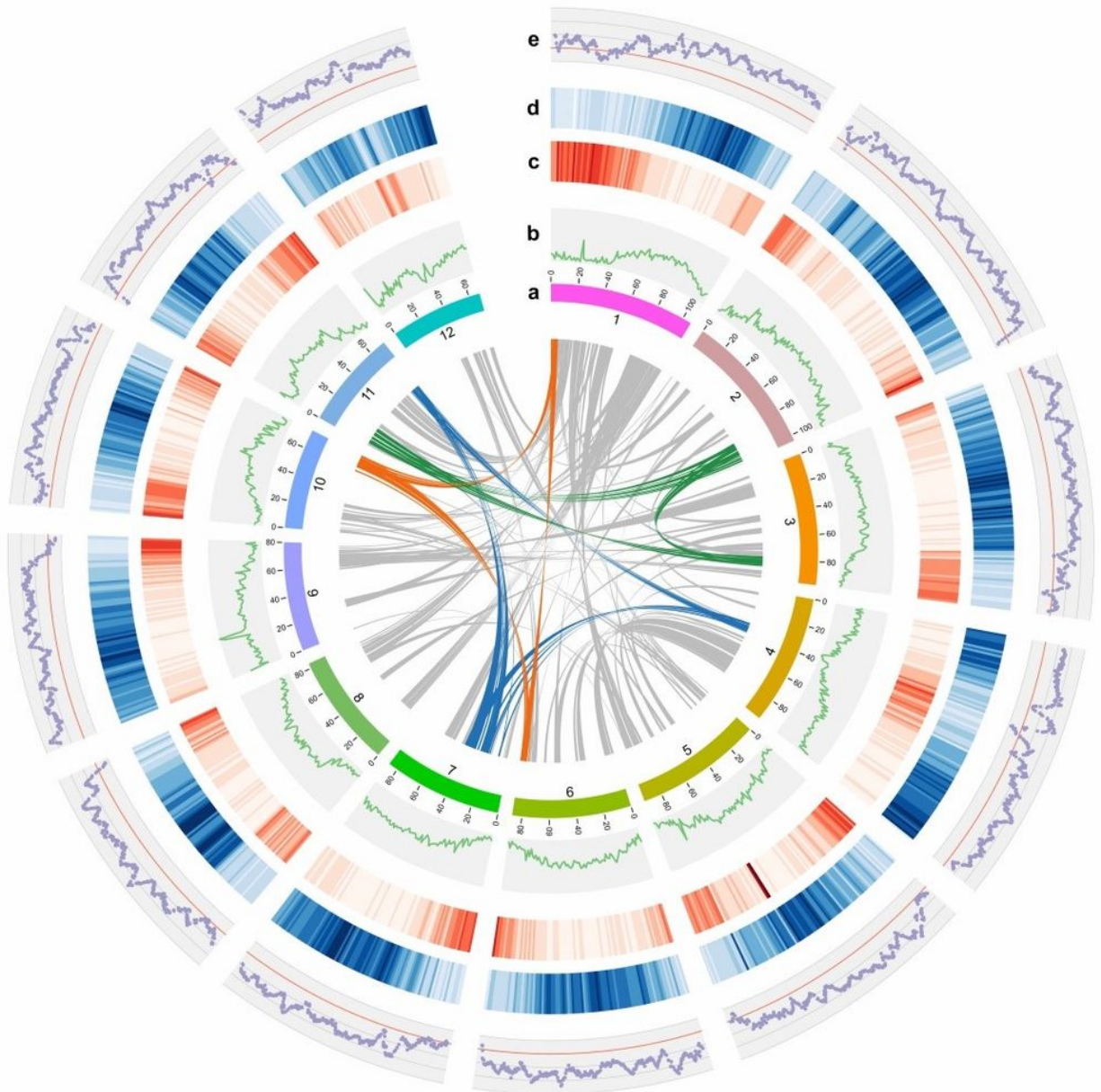


Figure 1

Overview of the *Carthamus tinctorius* genome. a. Chromosomal pseudomolecules. b. GC content (1 Mb windows). c. Gene density (1 Mb windows). d. TE density (1 Mb windows). e. LAI score (3 Mb windows)

with 300 Kb sliding step). Inner grey ribbons indicate links of syntenic blocks, while colored ribbons highlight the residues of whole genome triplication.

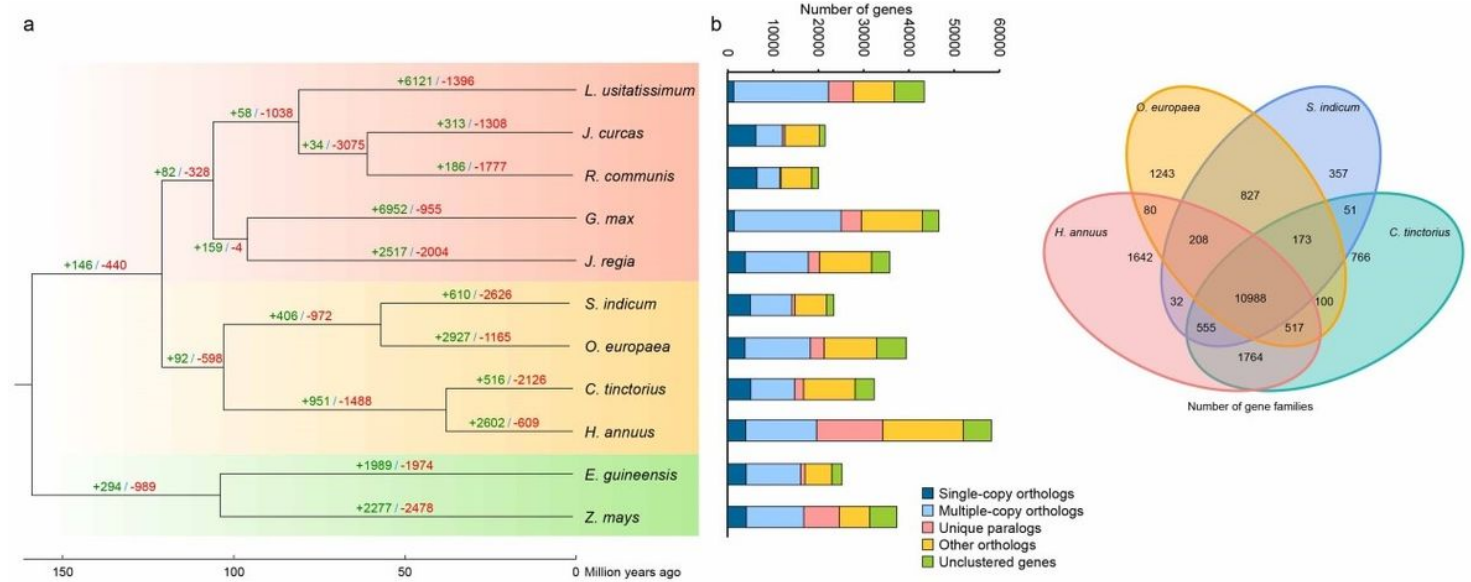


Figure 2

Comparative analysis of *Carthamus tinctorius* with other oil crops. (a) Phylogeny, divergence time and gene family expansion/contraction of 11 species. The green numbers are families under size expansion while the red numbers are families under size contraction. The vertical stacked column right is the orthologous genes in 11 species. (b) Venn diagram of safflower, sesame, olive tree and sunflower.

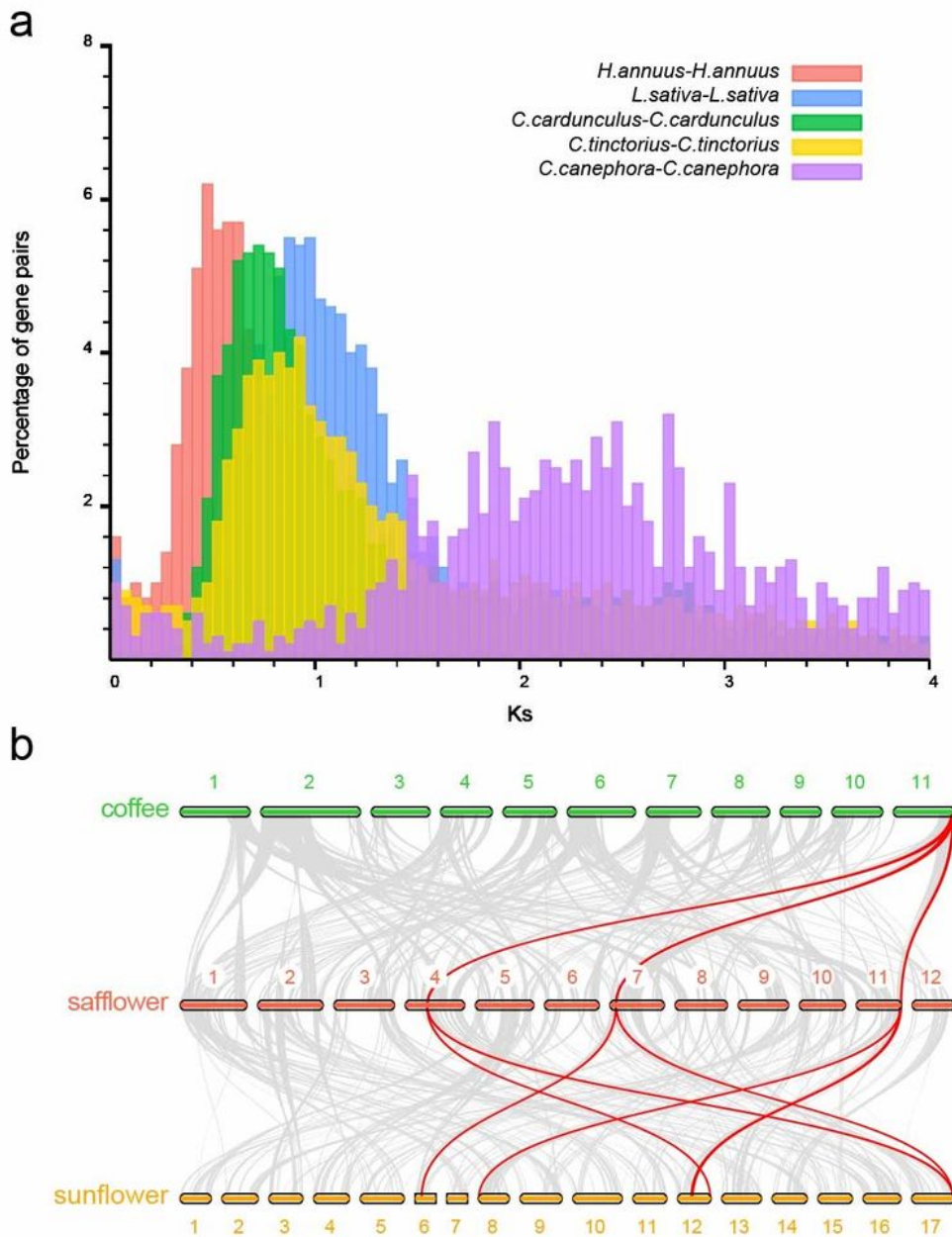


Figure 3

WGD events in *Carthamus tinctorius* evolutionary history. (a) Histogram of five species paralog Ks distributions. (b) Macrosynteny of gene regions among coffee, safflower and sunflower. Grey lines indicate the synteny blocks between each two species. Red lines highlight the 1:3 between coffee and safflower and 1:2 between safflower and sunflower synteny block corresponding relations.

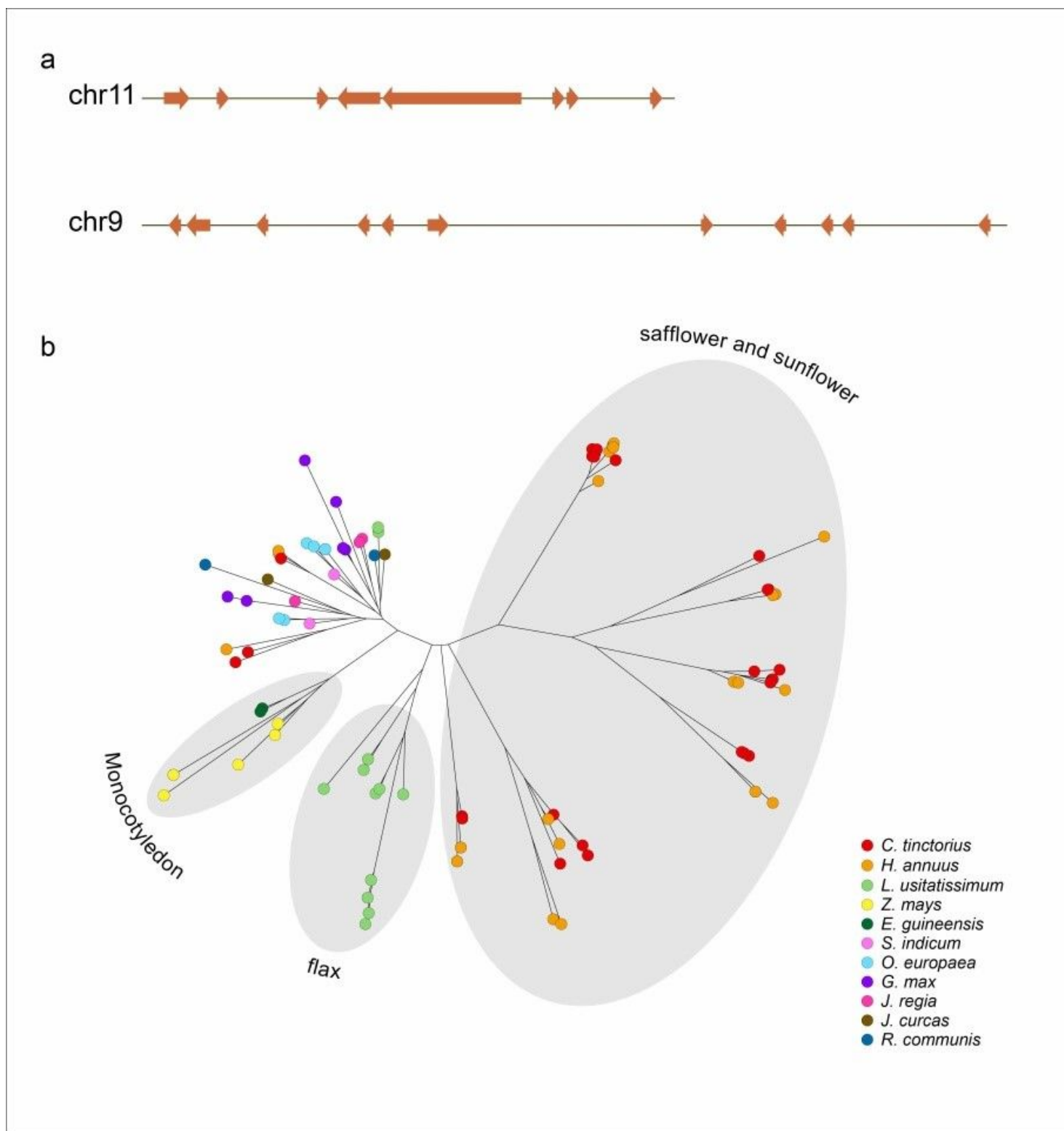
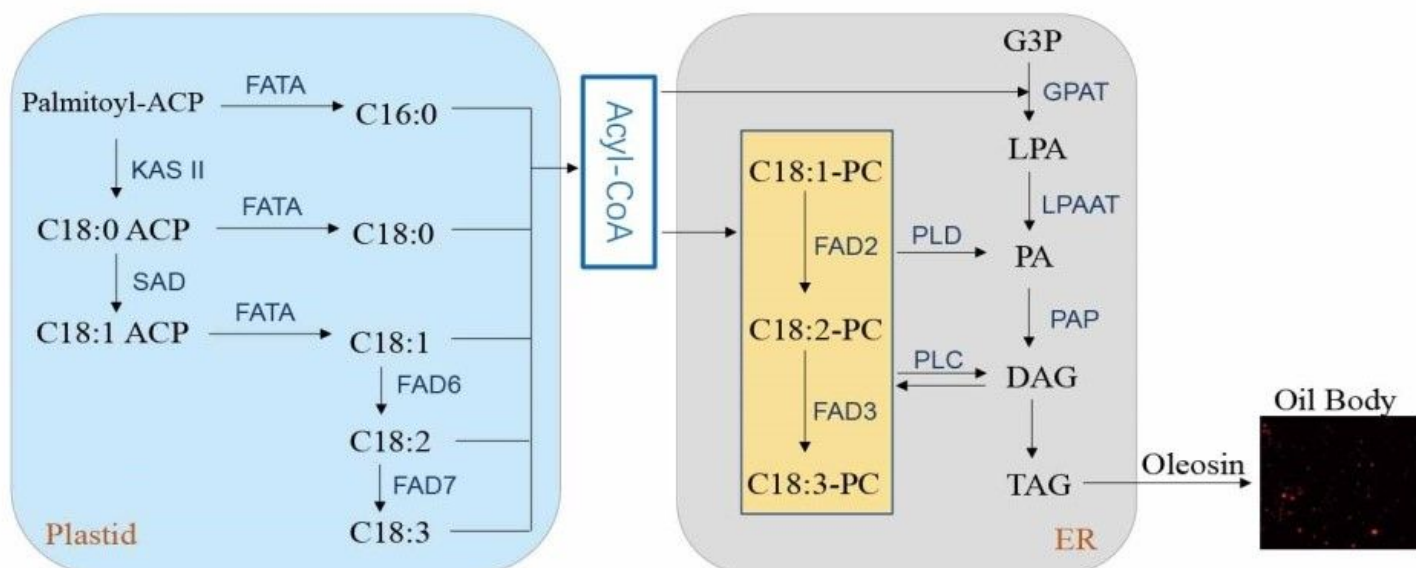


Figure 4

FAD2 gene clusters. (a) Two *FAD2* gene clusters on chromosome 9 and chromosome 11. The line indicates chromosome segment. Brown arrows indicate *FAD2* genes. (b) Phylogenetic analysis of the *FAD2* gene family among 11 species. Each circle indicates a *FAD2* gene. Different colors represent different species.

a



b

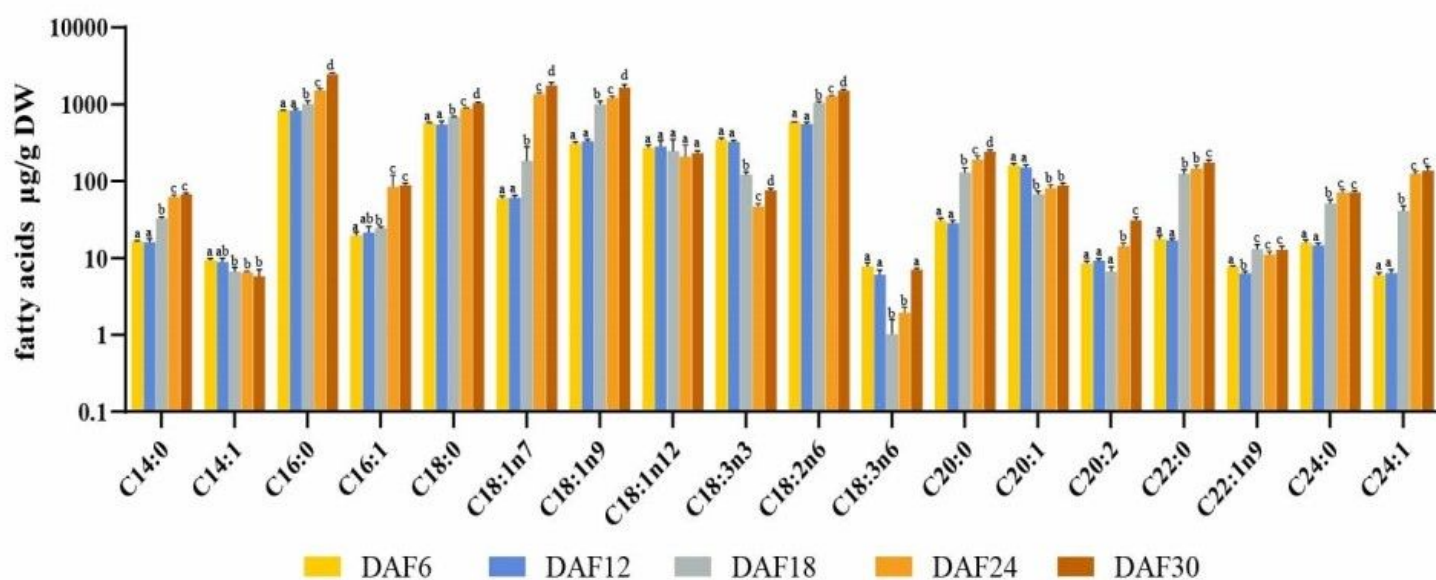


Figure 5

Fatty acids biosynthesis and contents of seed lipids. (a) Fatty acid and oil biosynthesis in safflower. (b) Fatty acids contents of seed storage lipids during developing stages

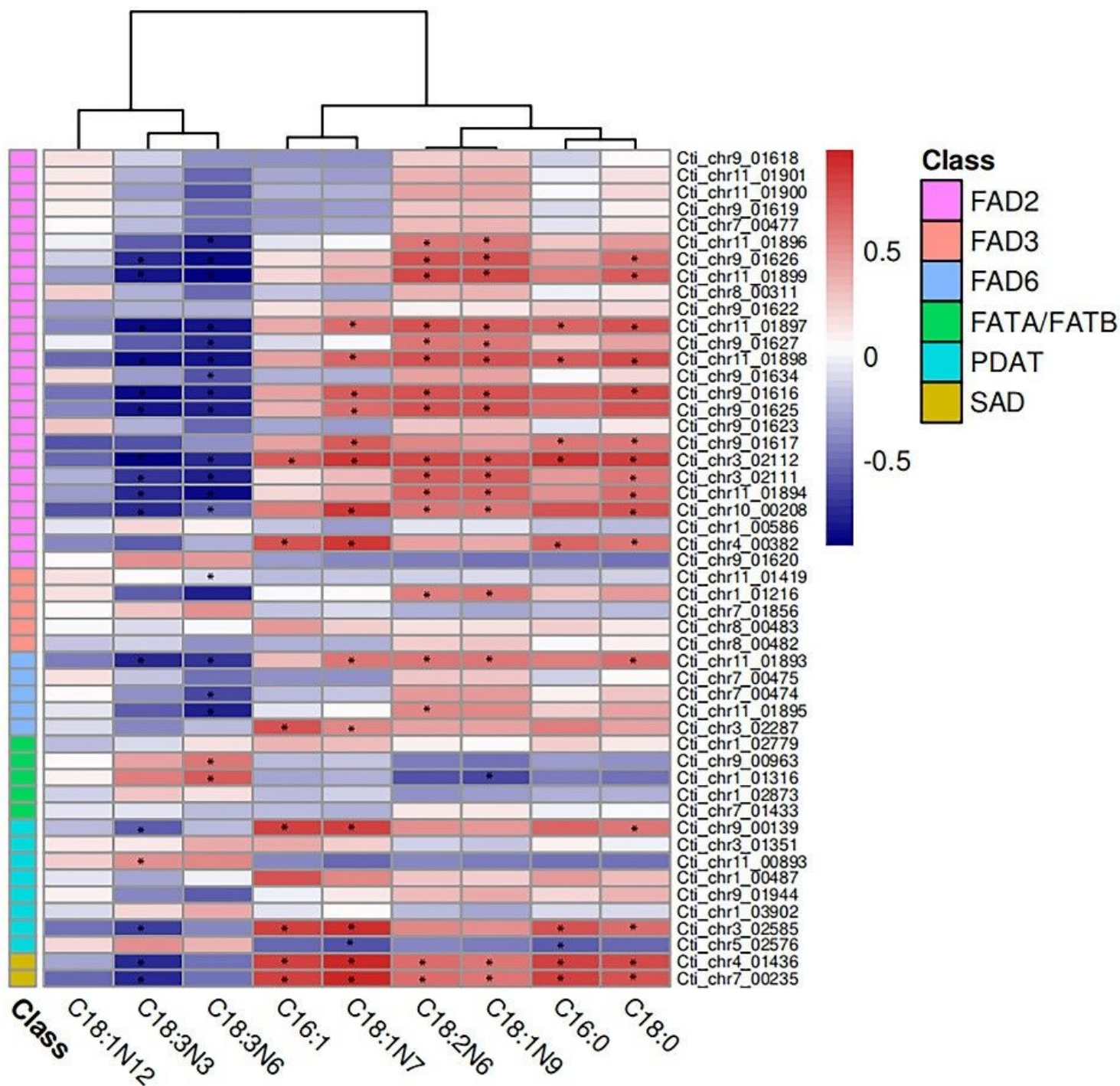


Figure 6

Correlation coefficient between gene expression level and contents of fatty acids. * $p < 0.05$

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementaryfile.docx](#)