TYPE Original Research
PUBLISHED 20 March 2023
DOI 10.3389/fpls.2023.1102715

A near complete genome assembly of chia assists in identification of key fatty acid desaturases in developing seeds

Leiting Li[1], Jingjing Song[1], Meiling Zhang[2], Shahid Iqbal[3], Yuanyuan Li[4], Heng Zhang[1]* and Hui Zhang[5]*

[1]National Key Laboratory of Molecular Plant Genetics, Shanghai Center for Plant Stress Biology, Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai, China, [2]Center for Excellence in Brain Science and Intelligence Technology, Institute of Neuroscience, Chinese Academy of Sciences, Shanghai, China, [3]Institute of Plant Breeding and Biotechnology, Muhammad Nawaz Shareef University of Agriculture, Multan, Pakistan, [4]Centre for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai, China, [5]Shandong Provincial Key Laboratory of Plant Stress Research, College of Life Science, Shandong Normal University, Jinan, Shandong, China

OPEN ACCESS

EDITED BY
Mark Chapman,
University of Southampton,
United Kingdom

REVIEWED BY
Liangsheng Zhang,
Zhejiang University, China
Wei Fan,
Agricultural Genomics Institute at
Shenzhen (CAAS), China
Simone Scalabrin,
IGA Technology Services, Udine, Italy

*CORRESPONDENCE
Heng Zhang
✉ hengzhang@psc.ac.cn
Hui Zhang
✉ laohanzhang@hotmail.com

SPECIALTY SECTION
This article was submitted to Functional and Applied Plant Genomics, a section of the journal Frontiers in Plant Science

RECEIVED 19 November 2022
ACCEPTED 06 March 2023
PUBLISHED 20 March 2023

CITATION
Li L, Song J, Zhang M, Iqbal S, Li Y, Zhang H and Zhang H (2023) A near complete genome assembly of chia assists in identification of key fatty acid desaturases in developing seeds. *Front. Plant Sci.* 14:1102715. doi: 10.3389/fpls.2023.1102715

COPYRIGHT
© 2023 Li, Song, Zhang, Iqbal, Li, Zhang and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Chia is an annual crop whose seeds have the highest content of α-linolenic acid (ALA) of any plant known to date. We generated a high-quality assembly of the chia genome using circular consensus sequencing (CCS) of PacBio. The assembled six chromosomes are composed of 21 contigs and have a total length of 361.7 Mb. Genome annotation revealed a 53.5% repeat content and 35,850 protein-coding genes. Chia shared a common ancestor with *Salvia splendens* ~6.1 million years ago. Utilizing the reference genome and two transcriptome datasets, we identified candidate fatty acid desaturases responsible for ALA biosynthesis during chia seed development. Because the seed of *S. splendens* contains significantly lower proportion of ALA but similar total contents of unsaturated fatty acids, we suggest that strong expression of two *ShFAD3* genes are critical for the high ALA content of chia seeds. This genome assembly will serve as a valuable resource for breeding, comparative genomics, and functional genomics studies of chia.

KEYWORDS

chia, polyunsaturated fatty acids, transcriptome, FAD, HiFi

Introduction

Chia (*Salvia hispanica* L.) is an annual herbaceous crop belonging to the family of Lamiaceae, also commonly known as the mint family. Chia is native to central America and is believed to have served as a staple crop of the Aztec in pre-Columbian times (Valdivia-López and Tecante, 2015). Chia is currently cultivated for its seeds in Central and South America. Chia produces oily seeds with an oval shape and a diameter of ~2 mm. Thanks to its superior nutrient compositions, the chia seed is a trending functional food ingredient

(Muñoz et al., 2013; Cassiday, 2017). Chia seeds contain 30-40% total lipids, of which α-linolenic acid (ALA; C18:3, n-3), linoleic acid (LA; C18:2, n-6), and oleic acid (C18:1, n-9) account for ~60%, ~20%, and ~10% respectively (Ciftci et al., 2012; Kulczynski et al., 2019). ALA is an essential fatty acid (i.e., cannot be synthesized by human body) and up to 8-21% and 1-9% of ALA intake can be respectively converted to eicosapentaenoic acid (EPA; C20:5, n-3) and docosahexaenoic acid (DHA; C22:6, n-3) in the human body (Baker et al., 2016; Shahidi and Ambigaipalan, 2018). Studies indicate that these n-3 fatty acids are important for human development and growth (Li et al., 2019). The recommended Adequate Intake (AI) of ALA is 1.6 g/day for men and 1.1 g/day for women (Burns-Whitmore et al., 2019). In addition, a low n-6:n-3 ratio, as in the case of chia seeds, in the diet helps reduce inflammation (Simopoulos, 2002a; Simopoulos, 2002b; Lands, 2014). Chia seeds also have high contents of dietary fiber (up to 34.4%), proteins (16.5-24.2%), vitamin B3, multiple minerals (such as calcium, phosphorus, potassium, and ion), and antioxidants (Kulczynski et al., 2019). Because of these properties, chia seeds are increasingly used as an ingredient in food industry and restaurants.

In plants, fatty acid (FA) biosynthesis takes place within the plastid, where acetyl-coenzyme A (acetyl-CoA) is used as the main carbon donor for the initiation and elongation of acyl chains (Ohlrogge and Browse, 1995; Li-Beisson et al., 2013). During the elongation, fatty acids remain covalently attached to acyl carrier proteins (ACPs), which serve as a cofactor for FA biosynthesis. The fatty acids biosynthesis cycle is usually terminated when the acyl chain reaches 16 or 18 carbons in length, and two principal types of acyl-ACP thioesterases, FatA and FatB, hydrolyze acyl-ACP and release the corresponding FAs. Desaturation of common fatty acids (C16 and C18) begins at the C-9 position (Δ9) and progresses in the direction of the methyl carbon of the acyl chain. Thus, the conversion of stearic acid (C18:0) to α-linoleic acid ($C18:3^{\Delta9,12,15}$) involves the sequential action of three desaturases, including the stearoyl-ACP desaturase, the oleate desaturase, and the linoleate desaturase. In the model plant Arabidopsis, genetic analyses have identified the main enzymes with specific FA desaturase activities. While all the other FA desaturases are membrane-bound enzymes, the family of acyl-ACP desaturases (AADs) are stromal soluble enzymes that use stearoyl-ACP (C18:0) or palmitoyl-ACP (C16:0) as the substrate. The Arabidopsis genome encodes 7 AADs (Kachroo et al., 2007), named as FAB2 (FATTY ACID BIOSYNTHESIS 2) and AAD1-6. Genetic analyses indicate that FAB2, AAD1, ADD5, and AAD6 are redundant Δ9 stearoyl-ACP desaturases (SADs) (Kazaz et al., 2020), while AAD2 and AAD3 function as Δ9 palmitoyl-ACP desaturases (PADs) (Troncoso-Ponce et al., 2016). Further desaturation of oleic acids ($C18:1^{\Delta9}$) may take place within the plastid or the endoplasmic reticulum (ER). In the plastid, the oleic acids are incorporated into multiple types of glycerophospholipids and converted to C18:3 by FAD6 (FATTY ACID DESATURASE 6) and FAD7/8. Alternatively, the oleic acid may be exported and enters the acyl-CoA pool in the cytosol. The C18:1-CoA can be imported into ER, where it is incorporated into phosphatidylcholine (PC) and becomes sequentially desaturated by FAD2 and FAD3, which respectively prefer PC with C18:1 and C18:2 as the substrate. During seed development, the desaturated PCs are further converted to diacylglycerol (DAG) and triacylglycerol (TAG), the latter of which is the main form of storage lipids in the oil body of seeds.

In this study, we assembled a high-quality chia genome using accurate consensus long reads (PacBio HiFi reads) and genome-wide chromosome conformation capture (Hi-C). The chia genome is known to have 6 chromosomes ($2n = 12$) (Estilai et al., 1990), which in our study are composed of 21 main contigs, with telomere repeats at 8 ends of the chromosomes. Utilizing this highly accurate and complete genome, we annotated transposable elements and protein-coding genes in the chia genome. Compared to a recently published chromosome-level assembly of chia (Wang et al., 2022), our assembly has better contiguity and ~15% more gene models (35,850 vs. 31,069) thanks to the highly accurate CCS reads. Alignment analyses also revealed multiple Mb-size structural variations between two assemblies, demonstrating the importance of multiple high-quality genomes for the same species. Finally, making use of a published seed development transcriptome, we identified the main ER-localized linoleate desaturases that underlie the extremely high ALA content in chia seeds.

Results

Genome assembly

We selected a chia cultivar with a Mexico origin (Supplemental Figure 1) for the assembly of the genome. About 24.7 Gb of circular consensus sequencing reads with an average read length of 16.1 kbp were generated from a single sequencing cell (Supplemental Figure 2). K-mer-based analyses of the HiFi reads estimated the nuclear genome to be ~352.7 Mb in size (Supplemental Figure 3).

We performed genome assembly using the hifiasm assembler (Cheng et al., 2021). The initial assembly was 388.0 Mb, consisting of 666 contigs with a N50 length of 21.8 Mb and an L50 number of 7, indicating a high contiguity of the assembly. The longest 21 contigs have a total length of 361.7 Mb and a minimum length of 1.7 Mb, while other contigs are significantly shorter, 636 of which have lengths shorter than 150 kbp (Figure 1A). The average HiFi read depth on the 21 long contigs varies between 43 and 58, which are around the 54-fold coverage of the nuclear genome calculated from the k-mer distribution (Figure 1A; Supplemental Figure 3). In contrast, the rest 645 contigs have a read coverage varying from 0 to 557, suggesting that they originate either from fragments of highly repetitive regions or from the high-copy organellar genomes.

We next analyzed the plastid and mitochondrion genomes. From the initial assembly, we identified a circular contig (ptg000033c) with a length of 313,444 bp and an average read coverage of 557 folds. Genome annotation identified 151 mitochondrion-encoded genes, including 21 transfer RNAs, 6 ribosomal RNAs (rRNAs), and 124 protein-coding genes (Supplemental Figure 4), indicating that this contig is the complete mitochondrion genome. We also identified 4 other contigs that show 100% sequence identity but structural variations to the mitochondrion genome (Supplemental Figure 5).

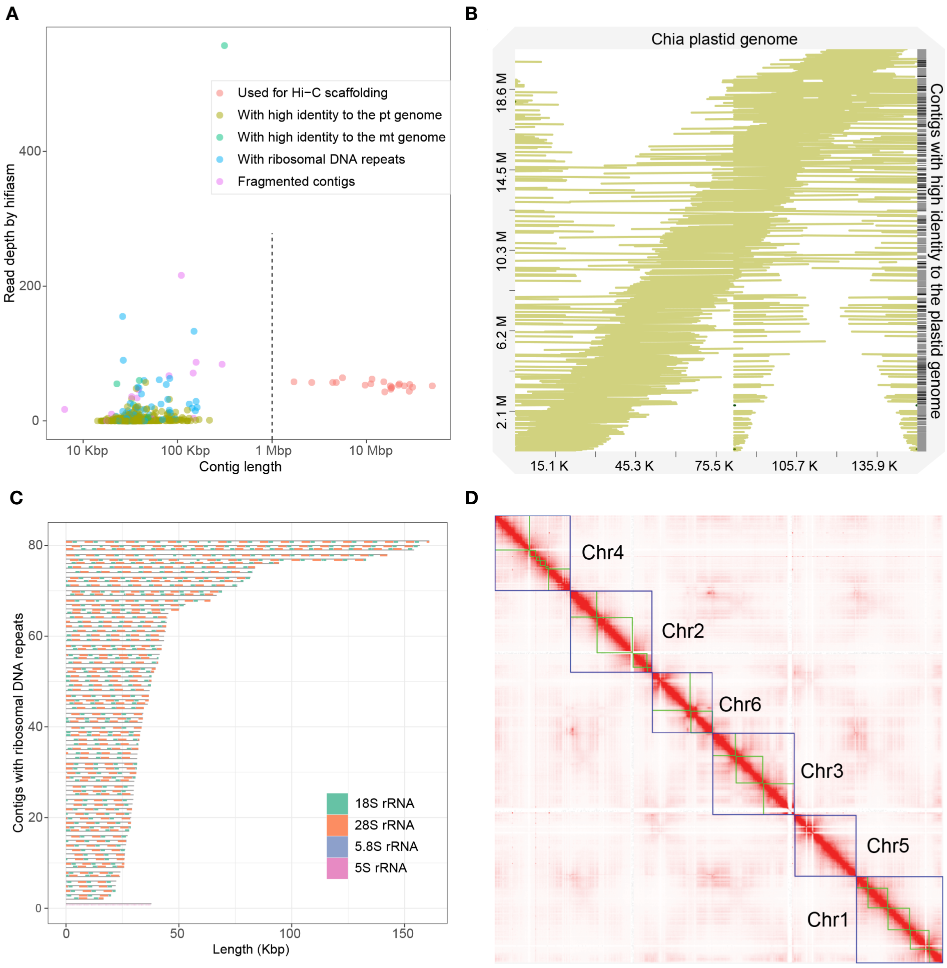

FIGURE 1
Assembly of the chia genome. **(A)** Dotplot showing the contig length and the read depth of the initial assembly. Contigs were classified into five categories based on the length, the read depth and their origins, as indicated in the legend. **(B)** Alignment of 538 initial contigs onto the chia plastid genome. The colored lines indicate the start and the end of the alignment relative to the plastid genome. Numbers in the X-axis indicate the length of the chia plastid genome while numbers in the Y-axis indicate the accumulated length of contigs that show high sequence identity to the plastid genome. **(C)** Structure of the 81 contigs containing ribosomal RNA repeats. **(D)** Hi-C contact map of the chia nuclear genome. Blue boxes indicate grouped pseudochromosomes, whereas green boxes indicate contigs.

Three of these contigs have a read depth similar to that of nuclear contigs (between 24 and 60) (Figure 1A). They might represent mitochondrial genome fragments recently transferred to the nuclear genome, or a minor population(s) of the heterozygous mitochondrial genome.

We could not identify a contig representing the complete plastid genome from the initial assembly. We thus assembled the plastid genome using Illumina short reads and the GetOrganelle software (Jin et al., 2020). The plastid genome has a length of 150,956 bp and 132 genes, including 87 protein-coding genes, 37 tRNA genes, and 8 rRNA genes (Supplemental Figure 6). Surprisingly, we found that 538 out of the 666 initial contigs could be mapped to the plastid genome with high coverage (>99%) and high identity rate (>99%) (Figure 1B). These contigs are short in length (14.2 to 217.6 kb) and most of them have low HiFi read coverage (with 530 contigs below 19-fold coverage) (Figure 1A). These plastid-originated contigs likely represent incompletely assembled plastid genome fragments and/or nuclear genome fragments with a plastid origin. The total length of these contigs was 20.7 Mb, accounting for most of the excessive part of the assembly compared to the predicted genome size.

Excluding the organellar-originated 543 contigs and the 21 high-confidence nuclear contigs, the remaining 102 contigs have a total length of 5.2 Mb. Ribosomal RNA (rRNA) repeats were identified in 81 of these contigs, indicating they were originated from genomic regions with high copy number of rRNA genes. Except for one contig mainly composed of 73 repeats of 5S rRNA, other contigs had a basic repeat unit of a “18S-5.8S-28S” structure with the copy number varied from 2 to 17 (Figure 1C). Considering the nuclear origin of most sequences, the 102 contigs were concatenated as Chr0.

We next used the 21 high-confidence nuclear contigs for Hi-C scaffolding. Based on ~180x (63.8 Gb) of Hi-C sequencing data, we clustered and ordered the 21 contigs into six pseudochromosomes,

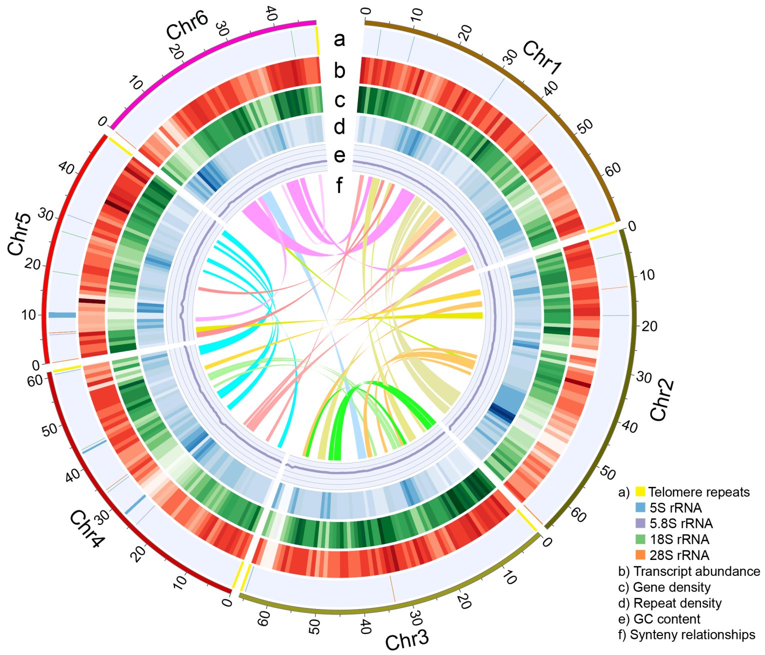

FIGURE 2
The nuclear genome of Shi_PSC_v1. Each ring indicates specific features of the nuclear genome. Data from non-overlapping 1-Mb windows were graphed: **(A)** Position of telomere repeats and ribosomal RNA genes; **(B)** Average transcript abundance; **(C)** Gene density; **(D)** LTR density; **(E)** GC content; **(F)** Synteny blocks >1 Mb in length.

whose sizes ranged from 47.8 Mb to 69.1 Mb (Figures 1D; 2, Table 1). The chromosome sequence names were decreasingly ordered based on sequence length. Chr5 was composed of a single contig while Chr4 contained the largest number (6) of contigs. The total length of the six pseudochromosomes was 361.7 Mb. The final v1 assembly (Shi_PSC_v1) of the chia genome composed of 9 sequences, seven of which (Chr0-Chr6) represent the nuclear genome, one for the mitochondrion genome, and one for the plastid genome.

Evaluation of genome assembly

We next evaluated the quality of the genome assembly using LTR Assembly Index (LAI) (Ou et al., 2018), Benchmarking Universal Single-Copy Orthologs (BUSCO) (Manni et al., 2021), Merqury (Rhie et al., 2020) and Illumina short reads. The whole genome had an LAI of 15.78, which was around the same level as the TAIR10 assembly of *Arabidopsis thaliana*, and could be considered as the reference level (Ou et al., 2018). The complete BUSCO of the chia genome assembly was 98.8%, indicating a high completeness of the gene space. Merqury compares k-mers from the assembly to those found in unassembled HiFi reads to estimate the completeness and accuracy. The completeness and quality value (QV) of Shi_PSC_v1 were 97.3 (out of 100) and 66.5 (>99.99% accuracy) respectively. Mapping of the Illumina short reads (Supplemental Table 1) against the chia genome assembly also revealed very high read mapping rate (99.9%) and a low apparent error rate (0.27%).

Genome annotation

For genome annotation, we first identified repetitive sequences in the Shi_PSC_v1 assembly. The analysis revealed that chia nuclear genome had a repeat content of 53.5% (Table 1). Similar to most plant genomes, retrotransposons accounts for the majority of the repetitive sequences of the genome. About half of the repeats were characterized as long terminal repeats (LTRs), with Gypsy (12.0% of the genome) and Copia (7.4% of the genome) being the main types. Besides, 65,851 simple repeats, 334 satellite sequences, 573 transfer RNAs (tRNAs) and 378 small nuclear RNAs (snRNAs) were also identified in the chia genome (Supplemental Table 2).

The repeat-masked assembly was then used for gene model prediction. Based on evidence from *ab initio* prediction, expressed sequence tags (ESTs) that assembled from the RNA-seq data by Gupta et al. (2021), and homologous protein sequences, a total of 35,850 protein-coding genes were annotated. Additionally, we also examined whether telomere signals were present at the end of each pseudochromosome. The results showed that all the six pseudochromosomes contain telomere repeats. Telomere repeats were detected at both ends of Chr3 and Chr4, and one end of Chr1, Chr2, Chr5, and Chr6 (Figure 2A). Comparing Shi_PSC_v1 to a recently published chia genome (Wang et al., 2022) revealed multiple Mb-size variations, including three inversions at the peri-telomeric region of Chr1 and the peri-centromeric regions of Chr2 and Chr3 (Supplemental Figure 7A). Further examination indicated that these regions are supported by raw reads in our assembly (Supplemental Figures 7B, C) but are composed of short

TABLE 1 Summary of chia genome assembly.

	Size	Number
Assembly features		
Estimated genome size	352,711,351 bp	
Total contigs	388,048,784 bp	666 contigs
Contig N50	21,830,104 bp	7 contigs
Longest contig	49,694,750 bp	
Chr1	69,924,378 bp	5 contigs/6997 genes
Chr2	66,361,501 bp	4 contigs/5756 genes
Chr3	66,031,358 bp	3 contigs/7894 genes
Chr4	61,126,009 bp	6 contigs/5365 genes
Chr5	49,694,750 bp	1 contig/4781 genes
Chr6	48,593,615 bp	2 contigs/4929 genes
Mitochondrial genome	313,444 bp	1 contig/151 genes
Plastid genome	150,956 bp	132 genes
GC content		37.00%
Annotation features		
Repetitive sequence		53.5%
Protein-coding genes		35,850

contigs concatenated together in the 2022 assembly (data not shown).

The complete BUSCO score of the protein sequences was 99.0%, close to the BUSCO score of the genome assembly (98.8%). Functional annotation showed that Gene Ontology (GO) terms (Gene Ontology, 2021), Pfam domains (Mistry et al., 2021), and InterPro families (Blum et al., 2021) were assigned to 58.9% (21,125), 72.0% (25,799), and 79.2% (28,405) of the protein-coding genes. In total, AHRD (Automated assignment of Human Readable Descriptions) function names were assigned to 89.5% (32,089) of the protein-coding genes (Boecker, 2021) (Supplemental Table 3). These metrics indicate high quality of the genome annotation.

Evolution of the chia genome

To understand the evolution of the chia genome, we selected five other species from the family of Lamiaceae, including three from the genus of *Salvia*, together with three species of Asterids and *Arabidopsis thaliana* for the orthology analysis (Figure 3A). A species tree constructed using orthologs shared in all analyzed species with STAG (Emms and Kelly, 2018) confirmed a close relationship between chia and *S. splendens*, as well as *S. bowleyana* and *S. miltiorrhiza* (Figure 3A). Using a reference divergence time of 115 million years ago (MYA) between Arabidopsis and other lineages (Hedges et al., 2015), chia was estimated to diverge with *S. splendens* ~6.2 million years ago (MYA) and the four *Salvia* species have a common ancestor ~21.8 MYA. The protein-coding genes of chia were assigned to 17,158 families. Relative to the common ancestor of chia and *S. splendens*, expansion in 528 families and reduction in 2,344 families were observed in chia (Figure 3A). In contrast, *S. splendens* had 8,777 expanded families and a large number of 2-copy gene families (Figure 3B). This is consistent with its recent tetraploidization event (Jia et al., 2021). Among the ten species analyzed, 8,812 families were shared while between 265 and 1,147 families were unique for each species (Figure 3C). Among the 720 gene families (2,529 genes) unique to chia, 72.6% of them were comprised of 2 or 3 members (Supplemental Figure 8) and the largest one contained 36 members. GO enrichment analysis was performed for genes in these chia-specific gene families. The results showed that the top enriched GO term in the category of biological process was “defense response” (GO:0006952) (Supplemental Figure 9), suggesting their potential roles in the environmental adaptation of chia. In addition, “acyl-[acyl-carrier-protein] desaturase activity” (GO:0045300) in the category of molecular function was enriched (Supplemental Figure 10). This expanded family mainly includes orthologous genes of *AtFAB2* (Supplemental Table 3; Supplemental Figure 11), the stearoyl-ACP (C18:0) or palmitoyl-ACP (C16:0) desaturases of Arabidopsis.

To investigate the whole-genome duplication events of chia, we performed intra-genome synteny analysis. In total, 323 synteny blocks with an average of 20.5 homologous gene pairs per block were identified (Figure 2F). The distribution of synonymous substitution rates (Ks) of these gene pairs revealed a single Ks peak at ~0.26 (Supplemental Figure 12), which was consistent with the whole genome duplication (WGD) event prior to the

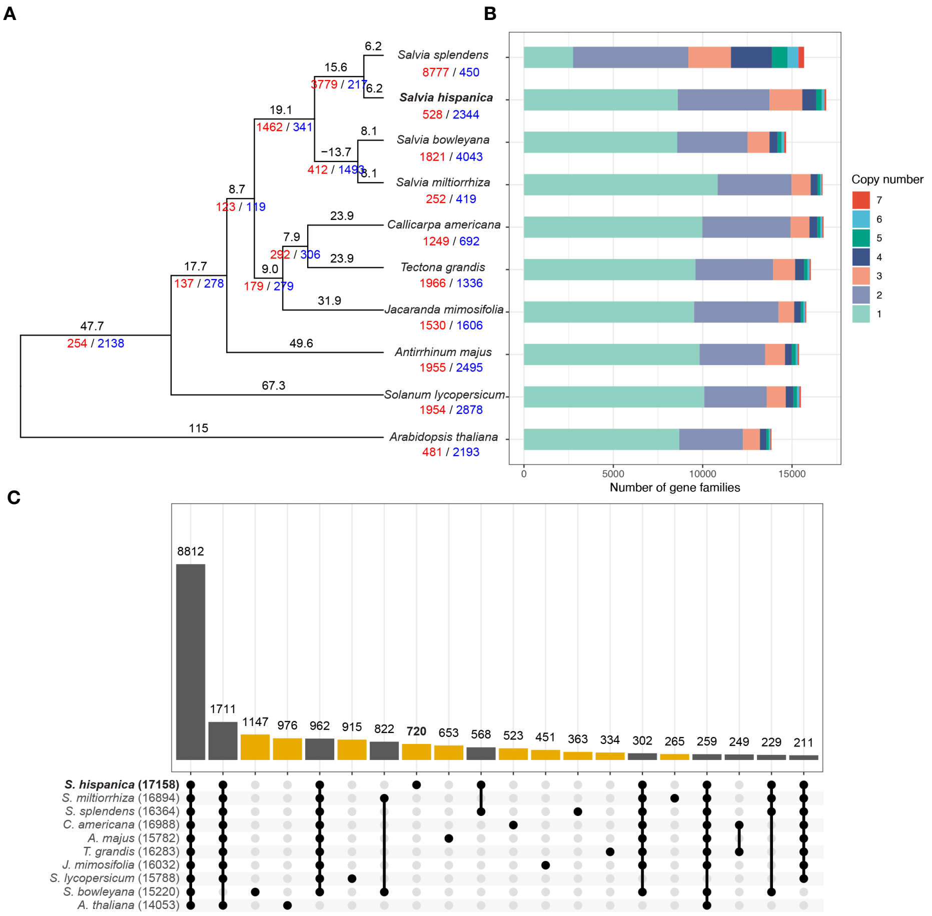

FIGURE 3
Evolution of the chia genome. **(A)** Phylogenetic tree for chia and 9 other plant species. Numbers of expanded and contracted gene families were indicated with red and blue numbers at each branch point. Branch length indicate the estimated divergence time in million years ago. **(B)** Numbers of gene families with different sizes (copy numbers) in each plant species. The X-axis indicate the number of gene families. **(C)** Upset plot indicating the number of gene families shared by different species. Yellow bars indicate numbers of species-specific gene families, whereas gene families that are commonly identified in two or more species are indicated by dark gray bars. Filled circles with a connected line below the x-axis indicate the species that share a specific number of gene families.

tetraploidization event of *S. splendens* (Jia et al., 2021). This indicates that this WGD event occurred before the divergence of chia and *S. splendens*.

Identification of genes involved in ALA biosynthesis

We next sought to identify genes underlying the high ALA content in chia seeds. We used kofamKOALA (Aramaki et al., 2020) to identify homologous genes of the lipid biosynthesis pathway (ko01004 of KEGG) in the chia genome (Supplemental Figure 13; Supplemental Table 4). We focused on genes encoding fatty acid desaturases. The analysis revealed 2 orthologs of *AtFatA* (K10782), 6 orthologs of *AtFatB* (K10781), 14 genes of the *AAD* family (K03921), 2 orthologs of *AtFAD2* (K10256), 2 orthologs of *AtFAD3* (K10257), and 2 orthologs of *AtFAD7/8* (K10257) among others (Figure 4A; Supplemental Figures 13, 14). Multiple sequence alignment (Supplemental Figure 15) indicated that *AtFAD7/8* and their orthologs in chia contain extra N-terminal sequences (plastid transit peptides) compared to the *AtFAD3* branch, consistent with their predicted localization in the plastid (Xue et al., 2018).

We utilized two published transcriptome dataset to help identify candidate ALA biosynthesis genes in the chia genome, one covering 13 different tissues or developmental stages of chia (Gupta et al., 2021) and one covering five different time points of chia seed development (3, 7, 14, 21, and 28 days after flower opening (DAF)) (Sreedhar et al., 2015). We reason that the ALA biosynthesis genes should be expressed at high levels during seed development. Indeed, we found that *Shi004382* (*ShFatA*), *Shi017381*, *Shi000260*, and *Shi006361* (*AtFAB2* orthologs), *Shi027338* and *Shi033531* (*AtFAD2* orthologs), and *Shi018884* and *Shi004328* (*AtFAD3* orthologs) are highly expressed in developing chia seeds, and their expression levels are decreased in the 28 DAF sample (Figure 4B). These genes are also expressed at significantly higher levels in developing seeds compared to other chia tissues/organs (Supplemental Figure 13). Although FAB2 homologs have either SAD or PAD activity, studies in Arabidopsis indicate that a single amino acid change (Tyr to Phe) is sufficient to confer PAD

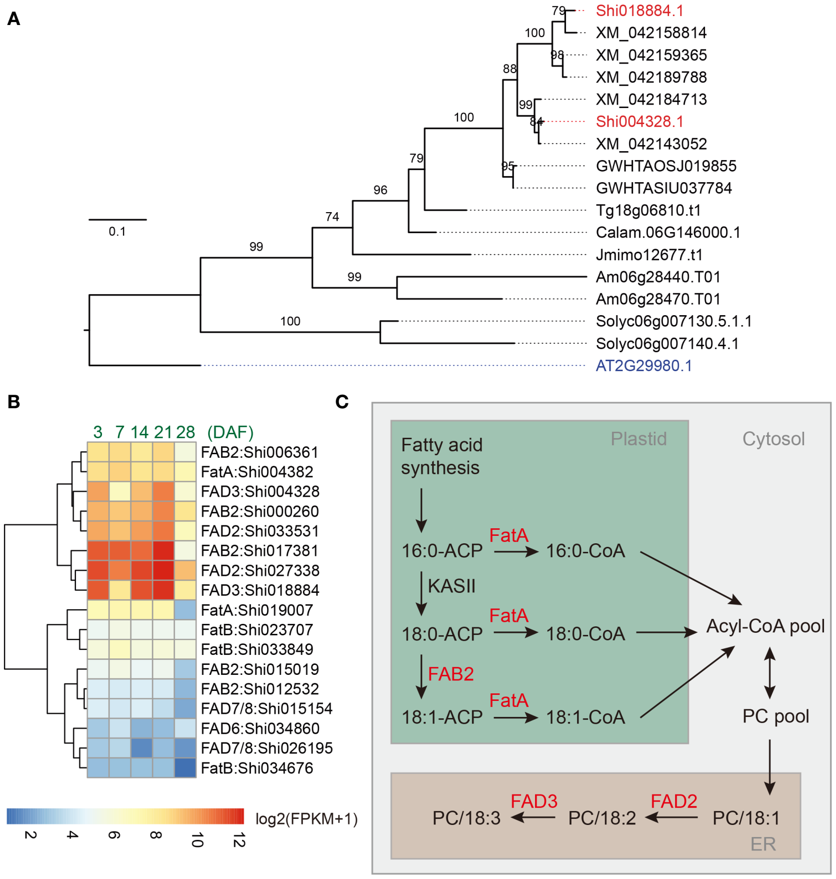

FIGURE 4
Identification of critical genes involved in fatty acid biosynthesis of chia seeds. **(A)** Phylogenetic tree of the *FAD3* genes. Shi: *Salvia hispanica*; AT: *Arabidopsis thaliana*; XM: *Salvia splendens*; GWHTAOSJ: *Salvia miltiorrhiza*; GWHTASIU: *Salvia bowleyana*; Tg: *Tectona grandis*; Jmimo: *Jacaranda mimosifolia*; Calam: *Callicarpa americana*; Am: *Antirrhinum majus*; Solyc: *Solanum lycopersicum*. **(B)** Expression pattern for *FatA*, *FatB*, *FAB2*, *FAD2*, *FAD3*, *FAD7/8*, and *FAD6* genes in developing chia seeds. DAF: Days after flower opening; FPKM: Fragments Per Kilobase of transcript per Million mapped reads. Only genes with maximum FPKM > 1 in seed samples were included in the plot. **(C)** A model for the biosynthesis of ALA in the chia genome. PC, phosphatidylcholine; ER, endoplasmic reticulum.

activity to AtFAB2 (SAD) (Troncoso-Ponce et al., 2016). The residue is predicted to locate at the bottom part of the substrate channel and the bulkier lateral chain of Phe may reduce the substrate binding pocket to better accommodate C16-ACP substrates. Multiple sequence alignment indicated that the highly expressed FAB2 homologs (*Shi017381*, *Shi000260*, and *Shi006361*) in chia seeds have a Tyr residue at the corresponding position, suggesting that they function as SADs (Supplemental Figure 16). In contrast, the two orthologs (*Shi015154* and *Shi026195*) of *AtFAD7/8*, the plastid localized omega-3 desaturase of Arabidopsis, were expressed at low to medium levels (FPKM values between 1.7 – 18.6) in developing seeds (Figure 4B; Supplemental Figure 13). In fact, multiple FA biosynthesis-related genes, such as genes encoding acyl carrier proteins (*Shi029800*, *Shi029801* and *Shi008432*), oil body-associated proteins (*Shi002948* and *Shi002148*), and lipid-transfer proteins (*Shi014949* and *Shi010250*), are also among the top 100 highly expressed genes in developing chia seeds (Supplemental Table 5). These results suggest a biosynthetic pathway involving plastid and ER localized enzymes, including *ShFAB2*, *ShFatA*, *ShFAD2* and *ShFAD3*, is responsible for the high ALA content in chia seeds (Figure 4C). Despite copy number variations were identified in some of these genes (Supplemental Table 4), we suggest that strong expression of fatty acid desaturase genes, particularly the ER localized FAD3s, are responsible for the high ALA content in chia seeds.

Discussion

De novo assembly of plant genomes has been greatly facilitated by the advancement of third-generation sequencing technologies that produce single-molecule long reads without the need of polymerase chain reactions. Commercially available 3rd-generation sequencing platforms suffer from high error rate of the raw reads (usually between 10-15%). The circular consensus sequencing (CCS) mode of PacBio significantly reduced consensus error rate by sequencing the same DNA insert multiple times. With carefully selected sizes of the DNA insert, a balance of sequencing length and accuracy can be achieved. In the current study, we performed CCS sequencing of the chia genomic DNA with a single SMRT cell, which produces 24.7 Gb of CCS data with median quality value of 31. The initial assembly included 666 contigs, while our analyses indicated that 623 of them originated from the organellar genomes or ribosome RNA repeats (Figure 1). The top 21 contigs have a total length of 361.7 Mb, which is slightly larger than the estimated genome size of 352.7 Mb based on k-mer analysis. Consistent with this high completeness of the nuclear genome, telomere repeats were identified at one or both ends of each of the six pseudochromosomes and rRNA repeats were identified in multiple chromosomes (Figure 2). Collapsing of repetitive regions was a common problem for *de novo* assembly of genomes with high repeat contents using longer but non-CCS PacBio reads. We did not observe similar phenomenon during the assembly of the chia genome. We reason that improved accuracy of the CCS mode helps resolving highly complex regions of the genome unless the repeat unit exceeds the read length, or the repeat sequences are highly similar.

Through phylogenetic and gene expression analyses, we identified candidate genes underlying high ALA contents of chia seeds. Two copies each of *ShFAB2*, *ShFAD2*, and *ShFAD3* exhibit very similar expression patterns (Figure 4B), suggesting these enzymes act together to promote the ALA content in chia seeds. This is consistent with the reported substrate channeling between FAD2 and FAD3 (Lou et al., 2014). Mature chia seeds have a lipid content of ~35%, of which up to 64% are ALA, the highest among all plant species (Muñoz et al., 2013; Kulczynski et al., 2019). Compared to its close relative, *S. splendens*, whose seeds were reported to have a ALA content of 34.5% and a LA content of 31.3% (Joh et al., 1988), the total content of ALA and LA of chia seeds are similar, suggesting that the elevated conversion rate from LA to ALA is the main event that drives high ALA content in chia seeds. In support of the idea that FAD3 is a rate limiting step in ALA biosynthesis, it was shown that overexpression of the rice *FAD3* gene is sufficient to increase the ALA content in seeds by ~28 fold (Liu et al., 2012). In addition to chia, seeds of flax (*Linum usitatissimum*) and perilla *(Perilla frutescens*) also have a relative ALA content around 60% (Ciftci et al., 2012). Although the genetic basis underlying their high ALA content remains to be determined, convergent high ALA contents in these species indicate that increasing omega-3 contents in seeds involve limited number of steps during evolution. This suggests a promising future for improving lipid composition in grains through transgenic or genome editing approaches.

Materials and methods

Library preparation and sequencing

Chia seeds were surface sterilized and grown in ½ MS medium supplemented with 0.7% agarose in a Percival growth chamber. Genomic DNA was extracted from two-week-old seedlings for genome survey sequencing and accurate consensus long-read sequencing (HiFi sequencing). The genome survey library was prepared and sequenced at the Genomics Core Facility of Shanghai Center for Plant Stress Biology following standard protocols. A 15-kb PacBio HiFi sequencing library were constructed and sequenced on a PacBio Sequel IIe platform at Berry Genomics (Beijing, China) following manufacturer's instructions. Etiolated 2-week-old seedlings were collected and used for crosslinking, proximity ligation, and library construction. The Hi-C library prepared by Biozeron (Shanghai, China) and sequenced at the Illumina NovaSeq platform with paired-end 150 bp sequencing mode.

Genome size estimation

To estimate the genome size of chia, 21 bp k-mer frequency of the PacBio HiFi reads was firstly counted with jellyfish (version 2.3.0) (Marcais and Kingsford, 2011). The k-mer frequency table was then used as input for GenomeScope2 (version 2.0) (Ranallo-Benavidez et al., 2020) to fit a diploid mathematical model to estimate the genome size, heterozygosity, and repetitiveness (Supplemental Figure 3).

Genome assembly

To assemble the nuclear genome using HiFi reads, three state-of-the-art genome assemblers were tested, including Flye (version 2.9) (Kolmogorov et al., 2019), HiCanu (version 2.2) (Nurk et al., 2020), and hifiasm (version 0.16.1) (Cheng et al., 2021). Flye applied a data structure of repeat graph (Kolmogorov et al., 2019). HiCanu was a modification of the Canu assembler (Koren et al., 2017) that was designed for HiFi reads with homopolymer compression, overlap-based error correction, and aggressive false overlap filtering (Nurk et al., 2020). Hifiasm is a genome assembler specifically designed for HiFi reads (Cheng et al., 2021). The previously estimated genome size was used as input parameter for Flye and HiCanu, while hifiasm does not require pre-estimated genome size. The results indicated that hifiasm with default parameters performed the best in terms of contiguity (Supplemental Table 6) and accuracy (Supplemental Figure 17).

To assemble the chia plastid genome, the GetOrganelle software (version 1.6.2) was used (Jin et al., 2020), which performs well in a comparison of chloroplast genome assembly tools (Freudenthal et al., 2020). GetOrganelle firstly extracted Illumina short reads that could be mapped to the embryophyte plastomes (a library composed of 101 plastid genomes) with bowtie2 (version 2.3.4.1) (Langmead and Salzberg, 2012) and then assembled them using SPAdes (version 3.13.0) (Bankevich et al., 2012). GetOrganelle produced three contigs representing the large single copy (LSC), small single copy (SSC) and inverted region (IR) of the chia plastid genome. Such three contigs were then aligned against the plastid genome of *Salvia miltiorrhiza* (accession number: NC_020431.1) (Qian et al., 2013), a close relative of chia. The alignment was performed with minimap2 (version 2.11) (Li, 2018) and visualized with D-Genies (version 1.3.1) (Cabanettes and Klopp, 2018). The three contigs were then ordered into a complete plastid genome using a customized Perl (version 5.34.0) script based on the BioPerl toolkit (version 1.7.4) (Stajich et al., 2002). Next, CHLOË (version

7c33699, https://chloe.plastid.org/) was used for the annotation of protein-coding genes, transfer RNAs, and ribosomal RNAs in the plastid genome.

To obtain the chia mitochondrial genome, we inspected contigs produced by hifiasm and found contig ptg000033c (length: 313,444 bp, read depth: 557) was circular and had the highest average read depth. Then we submitted this contig to the AGORA web tool (Jung et al., 2018) for genome annotation, with the protein-coding and rRNA genes of the *Salvia miltiorrhiza* mitochondrial genome (accession number: NC_023209.1) as a reference. The results of AGORA were then manually corrected by 1) removing protein-coding genes shorter than 30 amino acids, 2) removing protein-coding genes with pre-stop codons, 3) correcting mislabeled positions of ribosomal RNA genes. The chia mitochondrial genome was then visualized using OrganellarGenomeDRAW (OGDraw, version 1.3.1) (Greiner et al., 2019).

The "1-to-1" coverage and identity rate of contigs against the chia plastid and mitochondrial genomes were calculated using the dnadiff program of the MUMmer package (version 3.23) (Kurtz et al., 2004).

To obtain chia pseudochromosome sequences, the top 21 contigs in length and the Hi-C data was used for scaffolding. Illumina sequencing adapters and low-quality sequences of Hi-C data were trimmed by trim_galore (version 0.6.7, https://github.com/FelixKrueger/TrimGalore) with default parameters (quality score: 20; minimum length: 20 bp), which is a wrapper of cutadapt (version 3.4) (Martin, 2011). The clean Hi-C data were analyzed using Juicer (version 1.6) (Durand et al., 2016b), which produced high-quality DNA contact information. Then the 3D-DNA pipeline (version 180922) (Dudchenko et al., 2017) was used for ordering the contigs into pseudochromosomes. After visualizing the Hi-C contact map with Juicebox (version 1.9.1) (Durand et al., 2016a), we manually connect the contigs using "run-asm-pipeline-post-review.sh" of the 3D-DNA pipeline to avoid splitting the contigs.

Identification of rRNA repeats and telomere signatures

To predict the location of ribosomal RNA (rRNA) in the nuclear genome, Basic Rapid Ribosomal RNA Predictor (barrnap, version 0.9, https://github.com/tseemann/barrnap) was used, which using the nhmmer (version 3.1b1) (Wheeler and Eddy, 2013) to search the potential location of eukaryotes rRNA genes (5S, 5.8S, 28S, and 18S).

The telomere signature was examined using the program FindTelomeres (https://github.com/JanaSperschneider/FindTelomeres), which was a Python script for finding telomeric repeats (TTTAGGG/CCCTAAA). The results were further confirmed by TRF (version 4.09.1) (Benson, 1999) with parameters of "2 7 7 80 10 50 500 -m -d -h".

Genome circular plots were created in Circos (version 0.69.6) (Krzywinski et al., 2009). Dot plot of two genome assemblies was created using Assemblytics (Nattestad and Schatz, 2016). Visualization of the reads alignment file was performed using Integrative Genomics Viewer (IGV) (Thorvaldsdottir et al., 2013).

Genome quality evaluation

The quality of the genome assembly was evaluated using three methods, including Benchmarking Universal Single-Copy Orthologs (BUSCO) (version 5.0.0) (Manni et al., 2021), LTR Assembly Index (LAI) (version 2.9.0) (Ou et al., 2018) and Merqury (version 1.3) (Rhie et al., 2020). Merqury is a tool for reference-free assembly evaluation. Additionally, Illumina short reads were mapped to chia genome assembly using bwa-mem (version 0.7.17) (Li, 2013). The mapping rate and error rate of the Illumina short reads were estimated by SAMtools (version 1.15.1) (Li et al., 2009).

Genome annotation

A combined method was used for chia gene prediction, including *ab initio* prediction, EST discovery and protein homology search. To predict gene models, we firstly masked the repeats using RepeatMasker (version 3.1.2-p1) (Tarailo-Graovac and Chen, 2009). A species-specific repeat library was constructed for RepeatMasker using Repeatmodeler2 (version 2.0.2) (Flynn et al., 2020) and LTR_retriever (version 2.9.0) (Ou and Jiang, 2018). The LTR candidates for LTR_retriever was identified by LTR_FINDER_parallel (version 1.1) (Ou and Jiang, 2019) and LTRharvest (version 1.6.0) (Ellinghaus et al., 2008). LTR_FINDER_parallel is a parallel wrapper of LTR_FINDER (version 1.07) (Xu and Wang, 2007). The chia transcriptome of 13 tissue types (involved seeds, cotyledon, shoots, leaves, internodes, racemes, and flowers) (Gupta et al., 2021) were retrieved from the NCBI SRA database (accession number: PRJEB19614) and *de novo* assembled using Trinity (version 2.11.0) (Grabherr et al., 2011). The assembled transcripts were used as expressed sequence tags (EST) evidence for further gene model prediction. Seven sets of protein sequences downloaded from public databases were used as protein homology evidences, including *Arabidopsis thaliana* (version Araport11) (Cheng et al., 2017), *Antirrhinum majus* (version IGDBV1) (Li et al., 2019), *Callicarpa americana* (Hamilton et al., 2020), *Salvia miltiorrhiza* (version 1.0) (Song et al., 2020), *Salvia splendens* (Dong et al., 2018), *Tectona grandis* (Zhao et al., 2019) and the UniprotKB/Swiss-Prot dataset (version release-2020_04) (Poux et al., 2017).

Maker (version 3.01.03) (Campbell et al., 2014) was run three rounds to train AUGUSTUS (version 3.4.0) (Stanke and Waack, 2003) and SNAP (version 2006-07-28) (Korf, 2004) gene prediction parameters. GeMoMa (version 1.8) (Keilwagen et al., 2019) and MetaEuk (release 5) (Levy Karin et al., 2020) were used with the above mentioned protein homology datasets to discover gene models. Finally, EVidenceModeler (EVM, version 1.1.1) (Haas et al., 2008) was used to combine all the above gene prediction evidences. The est2geome and protein2genome features produced by Maker were used as transcript and protein evidence for EVM. The AUGUSTUS and SNAP gene models were used as *ab initio* prediction evidence for EVM. The GeMoMa and EetaEuk produced gene models were used as OTHER_PREDICTION evidence, which means they do not provide an indication of intergenic regions (Haas et al., 2008). As some of the gene models were overlapping with repetitive sequences, the final coding sequences and protein sequences were extracted from the unmasked genome assembly. Gene function annotation was performed by InterProScan (version 5.52-86.0) (Jones et al., 2014) and AHRD (version 3.3.3) (Boecker, 2021).

Genome evolution

Orthofinder (version 2.5.4) (Emms and Kelly, 2019) was used for the construction of orthologous groups. The STAG algorithm (Emms and Kelly, 2018) implemented in Orthofinder was used to estimate the species tree. Chia and other nine genomes were used for the construction of orthologous groups, including *Arabidopsis thaliana* (version Araport11) (Cheng et al., 2017), *Solanum lycopersicum* (version ITAG4.0) (Hosmani et al., 2019), *Antirrhinum majus* (version IGDBV1) (Li et al., 2019), *Tectona grandis* (Zhao et al., 2019), *Callicarpa americana* (Hamilton et al., 2020), *Jacaranda mimosifolia* (Wang et al., 2021), *Salvia bowleyana* (Zheng et al., 2021), *Salvia miltiorrhiza* (version 1.0) (Song et al., 2020), and *Salvia splendens* (version SspV2) (Jia et al., 2021). Gene family size expansion and contraction analysis was performed by CAFE5 (version 5.0.0) (Mendes et al., 2020). Synteny analysis was performed by the Python version of MCScan (version 1.1.17) (Tang et al., 2008). ParaAT (version 2.0) (Zhang et al., 2012) was used to prepare the alignment data for calculating Ks values, which was a wrapper of MUSCLE (version 3.8.1551) (Edgar, 2004) and PAL2NAL (version 13) (Suyama et al., 2006). KaKs_Calculator (version 2.0) (Wang et al., 2010) was used for calculating the Ks values using the YN model (Yang and Nielsen, 2000). The upset plot was created using the ggupset package (https://cran.r-project.org/package=ggupset) in R.

Gene expression analysis

Besides the chia transcriptome of 13 tissue types that retrieved from the NCBI SRA database (accession number: PRJEB19614) (Gupta et al., 2021), another set of transcriptome data for chia seed development was retrieved from the NCBI SRA database (accession number: PRJNA196477), which was sampled in 3, 7, 14, 21, and 28 DAF (Sreedhar et al., 2015). The raw RNA-seq data downloaded from the NCBI SRA database were firstly converted to FASTQ format using the fastq-dump command from the SRA Toolkit package (version 2.9.3, https://github.com/ncbi/sra-tools). Reads were then trimmed using trim_galore and then mapped to the chia reference genome by STAR (version 2.7.5c) (Dobin et al., 2013). Gene counts were summarized by featureCounts (version 2.0.1) (Liao et al., 2014). FPKM values were calculated using functions of the DESeq2 package (version 1.32.0) (Love et al., 2014) in the R platform (version 4.1.1) (R Core Team, 2021).

Multiple sequence alignment and phylogenetic tree construction

Visualization of multiple sequence alignment of the *FAD2*, *FAD3, FAD7, and FAD8* genes was performed using the Clustal Omega web tool (https://www.ebi.ac.uk/Tools/msa/clustalo/). Phylogenetic trees of the *FAB2/AAD*, *FAD2*, *FAD3*, *FAD7* and *FAD8* were constructed with the maximum likelihood method by IQ-TREE2 (Minh et al., 2020). The best-fitting amino acid substitution model was determined by ModelFinder (Kalyaanamoorthy et al., 2017).

Data availability statement

The datasets presented in this study can be found in online repositories. The genome assembly and corresponding sequencing data were deposited at NCBI (https://www.ncbi.nlm.nih.gov/) under accession number PRJNA864090 and at NGDC (https://ngdc.cncb.ac.cn/) under accession number PRJCA010915. The genome assembly and annotation data were deposited at CoGe (https://genomevolution.org/coge/) with genome ID 64745 for unmasked genome and genome ID 64746 for masked genome and figshare (https://doi.org/10.6084/m9.figshare.21976526).

Author contributions

LL performed data analyses; JS, MZ, and SI prepared plant materials; SI, YL, HeZ, and HuZ designed the project; LL and HeZ wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (31922008 to HeZ and 31900189 to LL), the Strategic Priority Research Program of CAS (XDB27040108 to HeZ), the Belt and Road Program of CAS (131965KYSB20190083-03 to HeZ), the Youth Innovation Promotion Association CAS (Y201844 to HeZ), and Central Guided Local Science and Technology Development Fund Project (YDZX2021079 to HuZ).

Acknowledgments

We thank the Core Facility for Genomics of Shanghai Center for Plant Stress Biology (PSC) for the construction of the Illumina sequencing library and the Core Facility for Bioinformatics of PSC for the maintenance of the high-performance computing (HPC) clusters.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1102715/full#supplementary-material

References

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., et al. (2020). KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 36, 2251–2252. doi: 10.1093/bioinformatics/btz859

Baker, E. J., Miles, E. A., Burdge, G. C., Yaqoob, P., and Calder, P. C. (2016). Metabolism and functional effects of plant-derived omega-3 fatty acids in humans. *Prog. Lipid Res.* 64, 30–56. doi: 10.1016/j.plipres.2016.07.002

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354. doi: 10.1093/nar/gkaa977

Boecker, F. (2021). *AHRD: Automatically annotate proteins with human readable descriptions and gene ontology terms* (Germany: University of Bonn, Bonn).

Burns-Whitmore, B., Froyen, E., Heskey, C., Parker, T., and San Pablo, G. (2019). Alpha-linolenic and linoleic fatty acids in the vegan diet: Do they require dietary reference Intake/Adequate intake special consideration? *Nutrients* 11, 2365. doi: 10.3390/nu11102365

Cabanettes, F., and Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* 6, e4958. doi: 10.7717/peerj.4958

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-p. *Curr. Protoc. Bioinf.* 48, 4 11 11–39. doi: 10.1002/0471250953.bi0411s48

Cassiday, L. (2017). Chia: superfood or superfad? *Inform* 28, 6–13. doi: 10.21748/inform.01.2017.06

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18, 170–175. doi: 10.1038/s41592-020-01056-5

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., and Town, C. D. (2017). Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* 89, 789–804. doi: 10.1111/tpj.13415

Ciftci, O. N., Przybylski, R., and Rudzińska, M. (2012). Lipid components of flax, perilla, and chia seeds. *Eur. J. Lipid Sci. Technol.* 114, 794–800. doi: 10.1002/ejlt.201100207

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dong, A. X., Xin, H. B., Li, Z. J., Liu, H., Sun, Y. Q., Nie, S., et al. (2018). High-quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant. *Gigascience* 7, giy068. doi: 10.1093/gigascience/giy068

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). *De novo* assembly of the *Aedes aegypti* genome using Hi-c yields chromosome-length scaffolds. *Science* 356, 92–95. doi: 10.1126/science.aal3327

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., et al. (2016a). Juicebox provides a visualization system for Hi-c contact maps with unlimited zoom. *Cell Syst.* 3, 99–101. doi: 10.1016/j.cels.2015.07.012

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., et al. (2016b). Juicer provides a one-click system for analyzing loop-resolution Hi-c experiments. *Cell Syst.* 3, 95–98. doi: 10.1016/j.cels.2016.07.002

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

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. doi: 10.1186/1471-2105-9-18

Emms, D., and Kelly, S. (2018). STAG: species tree inference from all genes. *BioRxiv*, 267914. doi: 10.1101/267914

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi: 10.1186/s13059-019-1832-y

Estilai, A., Hashemi, A., and Truman, K. (1990). Chromosome number and meiotic behavior of cultivated chia, salvia hispanica (Lamiaceae). *HortScience* 25, 1646–1647. doi: 10.21273/HORTSCI.25.12.1646

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* 117, 9451–9457. doi: 10.1073/pnas.1921046117

Freudenthal, J. A., Pfaff, S., Terhoeven, N., Korte, A., Ankenbrand, M. J., and Forster, F. (2020). A systematic comparison of chloroplast genome assembly tools. *Genome Biol.* 21, 254. doi: 10.1186/s13059-020-02153-6

Gene Ontology, C. (2021). The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.* 49, D325–D334. doi: 10.1093/nar/gkaa1113

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, W59–W64. doi: 10.1093/nar/gkz238

Gupta, P., Geniza, M., Naithani, S., Phillips, J. L., Haq, E., and Jaiswal, P. (2021). Chia (*Salvia hispanica*) gene expression atlas elucidates dynamic spatio-temporal changes associated with plant growth and development. *Front. Plant Sci.* 12, 667678. doi: 10.3389/fpls.2021.667678

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9, R7. doi: 10.1186/gb-2008-9-1-r7

Hamilton, J. P., Godden, G. T., Lanier, E., Bhat, W. W., Kinser, T. J., Vaillancourt, B., et al. (2020). Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing lamiaceae species, *Callicarpa americana*. *Gigascience* 9, giaa093. doi: 10.1093/gigascience/giaa093

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.* 32, 835–845. doi: 10.1093/molbev/msv037

Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E., et al. (2019). An improved *de novo* assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-c proximity ligation and optical maps. *BioRxiv*, 767764. doi: 10.1101/767764

Jia, K. H., Liu, H., Zhang, R. G., Xu, J., Zhou, S. S., Jiao, S. Q., et al. (2021). Chromosome-scale assembly and evolution of the tetraploid *Salvia splendens* (Lamiaceae) genome. *Hortic. Res.* 8, 177. doi: 10.1038/s41438-021-00614-y

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., dePamphilis, C. W., Yi, T. S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21, 241. doi: 10.1186/s13059-020-02154-5

Joh, Y.-G., Lee, O.-K., and Lim, Y.-J. (1988). Studies on the composition of fatty acid in the lipid classes of seed oils of the labiatae family. *J. Korean Appl. Sci. Technol.* 5, 13–23. doi: 10.12925/jkocs.1988.5.1.2

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Jung, J., Kim, J. I., Jeong, Y. S., and Yi, G. (2018). AGORA: organellar genome annotation from the amino acid and nucleotide references. *Bioinformatics* 34, 2661–2663. doi: 10.1093/bioinformatics/bty196

Kachroo, A., Shanklin, J., Whittle, E., Lapchyk, L., Hildebrand, D., and Kachroo, P. (2007). The *Arabidopsis* stearoyl-acyl carrier protein-desaturase family and the contribution of leaf isoforms to oleic acid synthesis. *Plant Mol. Biol.* 63, 257–271. doi: 10.1007/s11103-006-9086-y

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kazaz, S., Barthole, G., Domergue, F., Ettaki, H., To, A., Vasselon, D., et al. (2020). Differential activation of partially redundant Delta9 stearoyl-ACP desaturase genes is critical for omega-9 monounsaturated fatty acid biosynthesis during seed development in arabidopsis. *Plant Cell* 32, 3613–3637. doi: 10.1105/tpc.20.00554

Keilwagen, J., Hartung, F., and Grau, J. (2019). GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.* 1962, 161–177. doi: 10.1007/978-1-4939-9173-0_9

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37, 540–546. doi: 10.1038/s41587-019-0072-8

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly *via* adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinf.* 5, 59. doi: 10.1186/1471-2105-5-59

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

Kulczynski, B., Kobus-Cisowska, J., Taczanowski, M., Kmiecik, D., and Gramza-Michalowska, A. (2019). The chemical composition and nutritional value of chia seeds-current state of knowledge. *Nutrients* 11, 1242. doi: 10.3390/nu11061242

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5, R12. doi: 10.1186/gb-2004-5-2-r12

Lands, B. (2014). Historical perspectives on the impact of n-3 and n-6 nutrients on health. *Prog. Lipid Res.* 55, 17–29. doi: 10.1016/j.plipres.2014.04.002

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Levy Karin, E., Mirdita, M., and Soding, J. (2020). MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* 8, 48. doi: 10.1186/s40168-020-00808-x

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv.* 1303.3997. doi: 10.48550/arXiv.1303.3997

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, D., Wahlqvist, M. L., and Sinclair, A. J. (2019). Advances in n-3 polyunsaturated fatty acid nutrition. *Asia Pac J. Clin. Nutr.* 28, 1–5. doi: 10.6133/apjcn.201903_28(1).0001

Li, M., Zhang, D., Gao, Q., Luo, Y., Zhang, H., Ma, B., et al. (2019). Genome structure and evolution of *Antirrhinum majus* l. *Nat. Plants* 5, 174–183. doi: 10.1038/s41477-018-0349-9

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Li-Beisson, Y., Shorrosh, B., Beisson, F., Andersson, M. X., Arondel, V., Bates, P. D., et al. (2013). Acyl-lipid metabolism. *Arabidopsis Book* 11, e0161. doi: 10.1199/tab.0161

Liu, H. L., Yin, Z. J., Xiao, L., Xu, Y. N., and Qu le, Q. (2012). Identification and evaluation of omega-3 fatty acid desaturase genes for hyperfortifying alpha-linolenic acid in transgenic rice seed. *J. Exp. Bot.* 63, 3279–3287. doi: 10.1093/jxb/ers051

Lou, Y., Schwender, J., and Shanklin, J. (2014). FAD2 and FAD3 desaturases form heterodimers that facilitate metabolic channeling *in vivo*. *J. Biol. Chem.* 289, 17996–18007. doi: 10.1074/jbc.M114.572883

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8

Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A., and Zdobnov, E. M. (2021). BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* 38, 4647–4654. doi: 10.1093/molbev/msab199

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Mendes, F. K., Vanderpool, D., Fulton, B., and Hahn, M. W. (2020). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics*. 36, 5516–5518 doi: 10.1093/bioinformatics/btaa1022

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913

Muñoz, L. A., Cobos, A., Diaz, O., and Aguilera, J. M. (2013). Chia seed (*Salvia hispanica*): an ancient grain and a new functional food. *Food Rev. Int.* 29, 394–408. doi: 10.1080/87559129.2013.818014

Nattestad, M., and Schatz, M. C. (2016). Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* 32, 3021–3023. doi: 10.1093/bioinformatics/btw369

Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., et al. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* 30, 1291–1305. doi: 10.1101/gr.263566.120

Ohlrogge, J., and Browse, J. (1995). Lipid biosynthesis. *Plant Cell* 7, 957–970. doi: 10.1105/tpc.7.7.957

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Res.* 46, e126. doi: 10.1093/nar/gky730

Ou, S., and Jiang, N. (2018). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310

Ou, S., and Jiang, N. (2019). LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob DNA* 10, 48. doi: 10.1186/s13100-019-0193-0

Poux, S., Arighi, C. N., Magrane, M., Bateman, A., Wei, C. H., Lu, Z., et al. (2017). On expert curation and scalability: UniProtKB/Swiss-prot as a case study. *Bioinformatics* 33, 3454–3460. doi: 10.1093/bioinformatics/btx439

Qian, J., Song, J., Gao, H., Zhu, Y., Xu, J., Pang, X., et al. (2013). The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. *PloS One* 8, e57607. doi: 10.1371/journal.pone.0057607

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* 11, 1432. doi: 10.1038/s41467-020-14998-3

R Core Team (2021). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing). Available at: https://www.R-project.org/.

Rhie, A., Walenz, B. P., Koren, S., and Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* 21, 245. doi: 10.1186/s13059-020-02134-9

Shahidi, F., and Ambigaipalan, P. (2018). Omega-3 polyunsaturated fatty acids and their health benefits. *Annu. Rev. Food Sci. Technol.* 9, 345–381. doi: 10.1146/annurev-food-111317-095850

Simopoulos, A. P. (2002a). The importance of the ratio of omega-6/omega-3 essential fatty acids. *BioMed. Pharmacother.* 56, 365–379. doi: 10.1016/S0753-3322(02)00253-6

Simopoulos, A. P. (2002b). Omega-3 fatty acids in inflammation and autoimmune diseases. *J. Am. Coll. Nutr.* 21, 495–505. doi: 10.1080/07315724.2002.10719248

Song, Z., Lin, C., Xing, P., Fen, Y., Jin, H., Zhou, C., et al. (2020). A high-quality reference genome sequence of *Salvia miltiorrhiza* provides insights into tanshinone synthesis in its red rhizomes. *Plant Genome* 13, e20041. doi: 10.1002/tpg2.20041

Sreedhar, R. V., Kumari, P., Rupwate, S. D., Rajasekharan, R., and Srinivasan, M. (2015). Exploring triacylglycerol biosynthetic pathway in developing seeds of chia (*Salvia hispanica* l.): a transcriptomic approach. *PloS One* 10, e0123580. doi: 10.1371/journal.pone.0123580

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., et al. (2002). The bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12, 1611–1618. doi: 10.1101/gr.361602

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19 Suppl 2, ii215–ii225. doi: 10.1093/bioinformatics/btg1080

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi: 10.1093/nar/gkl315

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science* 320, 486–488. doi: 10.1126/science.1153917

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* 4, 10. doi: 10.1002/0471250953.bi0410s25

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Troncoso-Ponce, M. A., Barthole, G., Tremblais, G., To, A., Miquel, M., Lepiniec, L., et al. (2016). Transcriptional activation of two delta-9 palmitoyl-ACP desaturase genes by MYB115 and MYB118 is critical for biosynthesis of omega-7 monounsaturated fatty acids in the endosperm of arabidopsis seeds. *Plant Cell* 28, 2666–2682. doi: 10.1105/tpc.16.00612

Valdivia-López, MÁ, and Tecante, A. (2015). Chia (*Salvia hispanica*): A review of native Mexican seed and its nutritional and functional properties. *Adv. Food Nutr. Res.* 75, 53–75. doi: 10.1016/bs.afnr.2015.06.002

Wang, L., Lee, M., Sun, F., Song, Z., Yang, Z., and Yue, G. H. (2022). A chromosome-level genome assembly of chia provides insights into high omega-3 content and coat color variation of its seeds. *Plant Commun.* 3, 100326. doi: 10.1016/j.xplc.2022.100326

Wang, M., Zhang, L., and Wang, Z. (2021). Chromosomal-level reference genome of the Neotropical tree *Jacaranda mimosifolia* d. don. *Genome Biol. Evol.* 13, evab094. doi: 10.1093/gbe/evab094

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinf.* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Wheeler, T. J., and Eddy, S. R. (2013). Nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29, 2487–2489. doi: 10.1093/bioinformatics/btt403

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Xue, Y., Chen, B., Win, A. N., Fu, C., Lian, J., Liu, X., et al. (2018). Omega-3 fatty acid desaturase gene family from two omega-3 sources, salvia hispanica and perilla frutescens: Cloning, characterization and expression. *PloS One* 13, e0191432. doi: 10.1371/journal.pone.0191432

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17, 32–43. doi: 10.1093/oxfordjournals.molbev.a026236

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., et al. (2012). ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* 419, 779–781. doi: 10.1016/j.bbrc.2012.02.101

Zhao, D., Hamilton, J. P., Bhat, W. W., Johnson, S. R., Godden, G. T., Kinser, T. J., et al. (2019). A chromosomal-scale genome assembly of *Tectona grandis* reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways. *Gigascience* 8, giz005. doi: 10.1093/gigascience/giz005

Zheng, X., Chen, D., Chen, B., Liang, L., Huang, Z., Fan, W., et al. (2021). Insights into salvianolic acid b biosynthesis from chromosome-scale assembly of the salvia bowleyana genome. *J. Integr. Plant Biol.* 63, 1309–1323. doi: 10.1111/jipb.13085