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,*], 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, 300 Fenglin Road, Shanghai, China

[2] Center for Excellence in Brain Science and Intelligence Technology, Institute of Neuroscience, Chinese Academy of Sciences, 320 Yueyang Road, 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, 300 Fenglin Road, Shanghai, China

[5] Shandong Provincial Key Laboratory of Plant Stress Research, College of Life Science, Shandong Normal University, Ji'nan, Shandong, China

[*] Correspondence:

hengzhang@psc.ac.cn

laohanzhang@hotmail.com

Abstract

Chia is an annual crop whose seeds have the highest content of α-linolenic acid (ALA) of any plant species. We generated a high-quality assembly of the chia genome using circular consensus sequencing 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 serve 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, 2002, 2002; 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

63 acids remain covalently attached to acyl carrier proteins (ACPs), which serve as a cofactor
64 for FA biosynthesis. The fatty acids biosynthesis cycle is usually terminated when the acyl
65 chain reaches 16 or 18 carbons in length, and two principal types of acyl-ACP thioesterases,
66 FatA and FatB, hydrolyze acyl-ACP and release the corresponding FAs. Desaturation of
67 common fatty acids (C16 and C18) begins at the C-9 position (Δ9) and progresses in the
68 direction of the methyl carbon of the acyl chain. Thus, the conversion of stearic acid (C18:0)
69 to α-linoleic acid (C18:$3^{\Delta 9,12,15}$) involves the sequential action of three desaturases, including
70 the stearoyl-ACP desaturase, the oleate desaturase, and the linoleate desaturase. In the model
71 plant Arabidopsis, genetic analyses have identified the main enzymes with specific FA
72 desaturase activities. While all the other FA desaturases are membrane-bound enzymes, the
73 family of acyl-ACP desaturases (AADs) are stromal soluble enzymes that use stearoyl-ACP
74 (C18:0) or palmitoyl-ACP (C16:0) as the substrate. The Arabidopsis genome encodes 7
75 AADs (Kachroo et al., 2007), named as FAB2 (FATTY ACID BIOSYNTHESIS 2) and
76 AAD1-6. Genetic analyses indicate that FAB2, AAD1, ADD5, and AAD6 are redundant Δ9
77 stearoyl-ACP desaturases (SADs) (Kazaz et al., 2020), while AAD2 and AAD3 function as
78 Δ9 palmitoyl-ACP desaturases (PADs) (Troncoso-Ponce et al., 2016). Further desaturation of
79 oleic acids (C18:$1^{\Delta 9}$) may take place within the plastid or the endoplasmic reticulum (ER). In
80 the plastid, the oleic acids are incorporated into multiple types of glycerophospholipids and
81 converted to C18:3 by FAD6 (FATTY ACID DESATURASE 6) and FAD7/8. Alternatively,
82 the oleic acid may be exported and enters the acyl-CoA pool in the cytosol. The C18:1-CoA
83 can be imported into ER, where it is incorporated into phosphatidylcholine (PC) and becomes
84 sequentially desaturated by FAD2 and FAD3, which respectively prefer PC with C18:1 and
85 C18:2 as the substrate. During seed development, the desaturated PCs are further converted
86 to diacylglycerol (DAG) and triacylglycerol (TAG), the latter of which is the main form of
87 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). The six chia chromosomes are composed of 21 main contigs, with telomere repeats at 8 ends of the chromosomes. Utilizing this highly accurate and complete genome and 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) that has 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 high sequence identity (100%) to the mitochondrion genome (**Supplemental Figure 5**). We reason that these contigs may represent mitochondrial genome fragments recently transferred to the nuclear genome.

We could not identify a single contig representing the 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 misassembled 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 rest 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, whose sizes ranged from 47.8 Mb to 69.1 Mb (**Figure 1D; Figure 2; Table 1**). 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 and Copia 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), 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**).

The complete BUSCO score of the protein sequences was 97.6%, 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 by STAG (Emms and Kelly, 2018) confirmed a close relationship between chia and *S. splendens*, as well as *S. bowleyana* and *S. miltiorrhiza* (**Figure 3A**). After calibrating divergence time using data retrieved from the TimeTree database (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 7**) 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 8**), 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 9**). This expanded family mainly includes orthologs genes of *AtFAB2*

(**Supplemental Table 3; Supplemental Figure 10**), 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 11**), which was consistent with the whole genome duplication (WGD) event prior to the 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 12; Supplemental Table 4**). We focused on genes encoding fatty acid desaturases. The analysis revealed 2 homologs of *AtFatA* (K10782), 6 homologs of *AtFatB* (K10781), 14 genes of the *AAD* family (K03921), 2 homologs of *AtFAD2* (K10256), and 4 homologs of *AtFAD3/7/8* (K10257), among others (**Supplemental Figure 12**). Further phylogenetic analyses separated *AtFAD3/7/8* (K10257) into 2 branches, each containing 2 orthologs of *AtFAD3*, and *AtFAD7/8* (**Figure 4A**; **Supplemental Figure 13**). Multiple sequence alignment (**Supplemental Figure 14**) 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 candidate genes should be expressed at significantly higher levels compared to their counterparts 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 12**). 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 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 15**). 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 12**). In fact, the most highly expressed genes in developing chia seeds also contain 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*) (**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 bases 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 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 previous estimated genome size by GenomeScope2 was used as input parameter for Flye and HiCanu. While hifiasm do 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 16**).

To assemble the chia plastid genome, the GetOrganelle software (version 1.6.2) (Jin et al., 2020), which performs well in a systematic comparison of chloroplast genome assembly tools, (Freudenthal et al., 2020) was used. GetOrganelle firstly extracted Illumina short reads that could be mapped to the embryophyte plastomes (a library composed of 101 plastid genomes) by 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 reference. The results of AGORA were then manually corrected by 1) removing protein-coding genes less 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 by Juicer (version 1.6) (Durand et al., 2016), 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 by Juicebox (version 1.9.1) (Durand et al., 2016), 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).

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). 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 for 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).

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 that 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). The data 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

The genome assembly and corresponding sequencing data were deposited at NCBI under BioProject number PRJNA864090 and at NGDC under accession number PRJCA010915.

Funding

This work was supported by the National Natural Science Foundation of China (31922008 to Heng.Z. and 31900189 to L.L.), the Strategic Priority Research Program of CAS (XDB27040108 to Heng.Z.), the Belt and Road Program of CAS (131965KYSB20190083-03 to Heng.Z.), and the Youth Innovation Promotion Association CAS (Y201844 to Heng.Z.).

Author contributions

L.L. performed data analyses; J.S., M.Z., and S.I. prepared plant materials; S.I., Y.L., He.Z., Hu.Z. designed the project; L.L. and He.Z. wrote the manuscript.

Acknowledgements

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. The authors declare no conflict of interest.

References

Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H (2020) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics **36:** 2251-2252

Baker EJ, Miles EA, Burdge GC, Yaqoob P, Calder PC (2016) Metabolism and functional effects of plant-derived omega-3 fatty acids in humans. Prog Lipid Res **64:** 30-56

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19:** 455-477

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res **27:** 573-580

Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, Richardson L, Salazar GA, Williams L, Bork P, Bridge A, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA, Necci M, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A, Finn RD (2021) The InterPro protein families and domains database: 20 years on. Nucleic Acids Res **49:** D344-D354

Boecker F (2021) AHRD: Automatically Annotate Proteins with Human Readable Descriptions and Gene Ontology Terms. Universitäts-und Landesbibliothek Bonn

Burns-Whitmore B, Froyen E, Heskey C, Parker T, 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**

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

494 **Campbell MS, Holt C, Moore B, Yandell M** (2014) Genome Annotation and Curation
495 Using MAKER and MAKER-P. Curr Protoc Bioinformatics **48:** 4 11 11-39

496 **Cassiday L** (2017) Chia: superfood or superfad? Inform **28:** 6-13

497 **Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD** (2017)
498 Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome.
499 Plant J **89:** 789-804

500 **Cheng H, Concepcion GT, Feng X, Zhang H, Li H** (2021) Haplotype-resolved de novo
501 assembly using phased assembly graphs with hifiasm. Nat Methods **18:** 170-175

502 **Ciftci ON, Przybylski R, Rudzińska M** (2012) Lipid components of flax, perilla, and chia
503 seeds. European Journal of Lipid Science and Technology **114:** 794-800

504 **Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M,**
505 **Gingeras TR** (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29:**
506 15-21

507 **Dong AX, Xin HB, Li ZJ, Liu H, Sun YQ, Nie S, Zhao ZN, Cui RF, Zhang RG, Yun QZ,**
508 **Wang XN, Maghuly F, Porth I, Cong RC, Mao JF** (2018) High-quality assembly of
509 the reference genome for scarlet sage, *Salvia splendens*, an economically important
510 ornamental plant. Gigascience **7**

511 **Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS,**
512 **Machol I, Lander ES, Aiden AP, Aiden EL** (2017) De novo assembly of the *Aedes*
513 *aegypti* genome using Hi-C yields chromosome-length scaffolds. Science **356:** 92-95

514 **Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL**
515 (2016) Juicebox Provides a Visualization System for Hi-C Contact Maps with
516 Unlimited Zoom. Cell Syst **3:** 99-101

517 **Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL** (2016)
518 Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments.
519 Cell Syst **3:** 95-98

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:** 1792-1797

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics **9:** 18

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

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol **20:** 238

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A **117:** 9451-9457

Freudenthal JA, Pfaff S, Terhoeven N, Korte A, Ankenbrand MJ, Forster F (2020) A systematic comparison of chloroplast genome assembly tools. Genome Biol **21:** 254

Gene Ontology C (2021) The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res **49:** D325-D334

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol **29:** 644-652

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

Gupta P, Geniza M, Naithani S, Phillips JL, Haq E, 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

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol **9:** R7

Hamilton JP, Godden GT, Lanier E, Bhat WW, Kinser TJ, Vaillancourt B, Wang H, Wood JC, Jiang J, Soltis PS, Soltis DE, Hamberger B, Buell CR (2020) Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, *Callicarpa americana*. Gigascience **9**

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S (2015) Tree of life reveals clock-like speciation and diversification. Mol Biol Evol **32:** 835-845

Hosmani PS, Flores-Gonzalez M, van de Geest H, Maumus F, Bakker LV, Schijlen E, van Haarst J, Cordewener J, Sanchez-Perez G, Peters S (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

Jia KH, Liu H, Zhang RG, Xu J, Zhou SS, Jiao SQ, Yan XM, Tian XC, Shi TL, Luo H, Li ZC, Bao YT, Nie S, Guo JF, Porth I, El-Kassaby YA, Wang XR, Chen C, Van de Peer Y, Zhao W, Mao JF (2021) Chromosome-scale assembly and evolution of the tetraploid *Salvia splendens* (Lamiaceae) genome. Hortic Res **8:** 177

Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, Li DZ (2020) GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol **21:** 241

Joh Y-G, Lee O-K, Lim Y-J (1988) Studies on the Composition of Fatty Acid in the Lipid Classes of Seed Oils of the Labiatae Family. Journal of the Korean Applied Science and Technology **5:** 13-23

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M,

Yong SY, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics **30:** 1236-1240

Jung J, Kim JI, Jeong YS, Yi G (2018) AGORA: organellar genome annotation from the amino acid and nucleotide references. Bioinformatics **34:** 2661-2663

Kachroo A, Shanklin J, Whittle E, Lapchyk L, Hildebrand D, 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

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods **14:** 587-589

Kazaz S, Barthole G, Domergue F, Ettaki H, To A, Vasselon D, De Vos D, Belcram K, Lepiniec L, Baud S (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

Keilwagen J, Hartung F, Grau J (2019) GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. Methods Mol Biol **1962:** 161-177

Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol **37:** 540-546

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res **27:** 722-736

Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics **5:** 59

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res **19:** 1639-1645

Kulczynski B, Kobus-Cisowska J, Taczanowski M, Kmiecik D, Gramza-Michalowska A (2019) The Chemical Composition and Nutritional Value of Chia Seeds-Current State of Knowledge. Nutrients **11**

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol **5:** R12

Lands B (2014) Historical perspectives on the impact of n-3 and n-6 nutrients on health. Prog Lipid Res **55:** 17-29

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods **9:** 357-359

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

Li D, Wahlqvist ML, Sinclair AJ (2019) Advances in n-3 polyunsaturated fatty acid nutrition. Asia Pac J Clin Nutr **28:** 1-5

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

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34:** 3094-3100

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:** 2078-2079

Li M, Zhang D, Gao Q, Luo Y, Zhang H, Ma B, Chen C, Whibley A, Zhang Y, Cao Y, Li Q, Guo H, Li J, Song Y, Zhang Y, Copsey L, Li Y, Li X, Qi M, Wang J, Chen Y, Wang D, Zhao J, Liu G, Wu B, Yu L, Xu C, Li J, Zhao S, Zhang Y, Hu S, Liang C, Yin Y, Coen E, Xue Y (2019) Genome structure and evolution of *Antirrhinum majus* L. Nat Plants **5:** 174-183

Li-Beisson Y, Shorrosh B, Beisson F, Andersson MX, Arondel V, Bates PD, Baud S, Bird D, Debono A, Durrett TP, Franke RB, Graham IA, Katayama K, Kelly AA, Larson T, Markham JE, Miquel M, Molina I, Nishida I, Rowland O, Samuels L, Schmid KM, Wada H, Welti R, Xu C, Zallot R, Ohlrogge J (2013) Acyl-lipid metabolism. Arabidopsis Book **11:** e0161

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30:** 923-930

Liu HL, Yin ZJ, Xiao L, Xu YN, 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

Lou Y, Schwender J, Shanklin J (2014) FAD2 and FAD3 desaturases form heterodimers that facilitate metabolic channeling in vivo. J Biol Chem **289:** 17996-18007

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **15:** 550

Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM (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

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

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal **17:** 10-12

Mendes FK, Vanderpool D, Fulton B, Hahn MW (2020) CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol **37:** 1530-1534

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A (2021) Pfam: The protein families database in 2021. Nucleic Acids Res **49:** D412-D419

Muñoz LA, Cobos A, Diaz O, Aguilera JM (2013) Chia seed (*Salvia hispanica*): an ancient grain and a new functional food. Food reviews international **29:** 394-408

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S (2020) HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res **30:** 1291-1305

Ohlrogge J, Browse J (1995) Lipid biosynthesis. Plant Cell **7:** 957-970

Ou S, Chen J, Jiang N (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res **46:** e126

Ou S, Jiang N (2018) LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol **176:** 1410-1422

Ou S, Jiang N (2019) LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA **10:** 48

Poux S, Arighi CN, Magrane M, Bateman A, Wei CH, Lu Z, Boutet E, Bye AJH, Famiglietti ML, Roechert B, UniProt Consortium T (2017) On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. Bioinformatics **33:** 3454-3460

Qian J, Song J, Gao H, Zhu Y, Xu J, Pang X, Yao H, Sun C, Li X, Li C, Liu J, Xu H, Chen S (2013) The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. PLoS One **8:** e57607

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

Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun **11:** 1432

Rhie A, Walenz BP, Koren S, Phillippy AM (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol **21:** 245

Shahidi F, Ambigaipalan P (2018) Omega-3 Polyunsaturated Fatty Acids and Their Health Benefits. Annu Rev Food Sci Technol **9:** 345-381

Simopoulos AP (2002) The importance of the ratio of omega-6/omega-3 essential fatty acids. Biomed Pharmacother **56:** 365-379

Simopoulos AP (2002) Omega-3 fatty acids in inflammation and autoimmune diseases. J Am Coll Nutr **21:** 495-505

Song Z, Lin C, Xing P, Fen Y, Jin H, Zhou C, Gu YQ, Wang J, Li X (2020) A high-quality reference genome sequence of *Salvia miltiorrhiza* provides insights into tanshinone synthesis in its red rhizomes. Plant Genome **13:** e20041

Sreedhar RV, Kumari P, Rupwate SD, Rajasekharan R, Srinivasan M (2015) Exploring triacylglycerol biosynthetic pathway in developing seeds of Chia (*Salvia hispanica* L.): a transcriptomic approach. PLoS One **10:** e0123580

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res **12:** 1611-1618

Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics **19 Suppl 2:** ii215-225

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res **34:** W609-612

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008) Synteny and collinearity in plant genomes. Science **320:** 486-488

Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics **Chapter 4:** Unit 4 10

Troncoso-Ponce MA, Barthole G, Tremblais G, To A, Miquel M, Lepiniec L, Baud S (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

Valdivia-López MÁ, Tecante A (2015) Chia (*Salvia hispanica*): A review of native Mexican seed and its nutritional and functional properties. Advances in food and nutrition research **75:** 53-75

Wang D, Zhang Y, Zhang Z, Zhu J, Yu J (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics **8:** 77-80

Wang M, Zhang L, Wang Z (2021) Chromosomal-Level Reference Genome of the Neotropical Tree *Jacaranda mimosifolia* D. Don. Genome Biol Evol **13**

Wheeler TJ, Eddy SR (2013) nhmmer: DNA homology search with profile HMMs. Bioinformatics **29:** 2487-2489

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res **35:** W265-268

Xue Y, Chen B, Win AN, Fu C, Lian J, Liu X, Wang R, Zhang X, Chai Y (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

722 **Yang Z, Nielsen R** (2000) Estimating synonymous and nonsynonymous substitution rates
723 under realistic evolutionary models. Mol Biol Evol **17:** 32-43

724 **Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, Dai L** (2012) ParaAT: a parallel tool
725 for constructing multiple protein-coding DNA alignments. Biochem Biophys Res
726 Commun **419:** 779-781

727 **Zhao D, Hamilton JP, Bhat WW, Johnson SR, Godden GT, Kinser TJ, Boachon B,**
728 **Dudareva N, Soltis DE, Soltis PS, Hamberger B, Buell CR** (2019) A chromosomal-
729 scale genome assembly of *Tectona grandis* reveals the importance of tandem gene
730 duplication and enables discovery of genes in natural product biosynthetic pathways.
731 Gigascience **8**

732 **Zheng X, Chen D, Chen B, Liang L, Huang Z, Fan W, Chen J, He W, Chen H, Huang L,**
733 **Chen Y, Zhu J, Xue T** (2021) Insights into salvianolic acid B biosynthesis from
734 chromosome-scale assembly of the Salvia bowleyana genome. J Integr Plant Biol **63:**
735 1309-1323
736

737 **Table 1.** Summary of chia genome assembly and annotation

	Salvia hispanica	**Number of contigs**
Assembly		
Contigs		
Estimated genome size	352,711,351 bp	
Total length (number) of contigs	388,048,784 bp	666
N50 (L50) of contigs	21,830,104 bp	7
N90 (L90) of contigs	4,780,990 bp	18
Longest contig	49,694,750 bp	
Contigs used for Hi-C scaffolding	361,724,111 bp	21
Contigs with high identity to the plastid genome	20,671,896 bp	538
Contigs with high identity to the mitochondrial genome	458,295 bp	5
Contigs with ribosomal DNA repeats	3,787,617 bp	81
Pseudomolecules		
Chr1	69,924,378 bp	5
Chr2	66,361,501 bp	4
Chr3	66,031,358 bp	3
Chr4	61,126,009 bp	6
Chr5	49,694,750 bp	1
Chr6	48,593,615 bp	2
Mitochondrial genome	313,444 bp	
Plastid genome	150,956 bp	
Annotation		
GC content	37.00%	
Repeat sequences	53.5%	
LTR/Copia	7.4%	
LTR/Gypsy	12.0%	
Number of protein-coding genes	35,850	
Complete BUSCOs Genome assembly Proteome	 98.8% 97.6%	
Simple repeats	65,875	
Satellite	334	
tRNA	586	
snRNA	378	

738

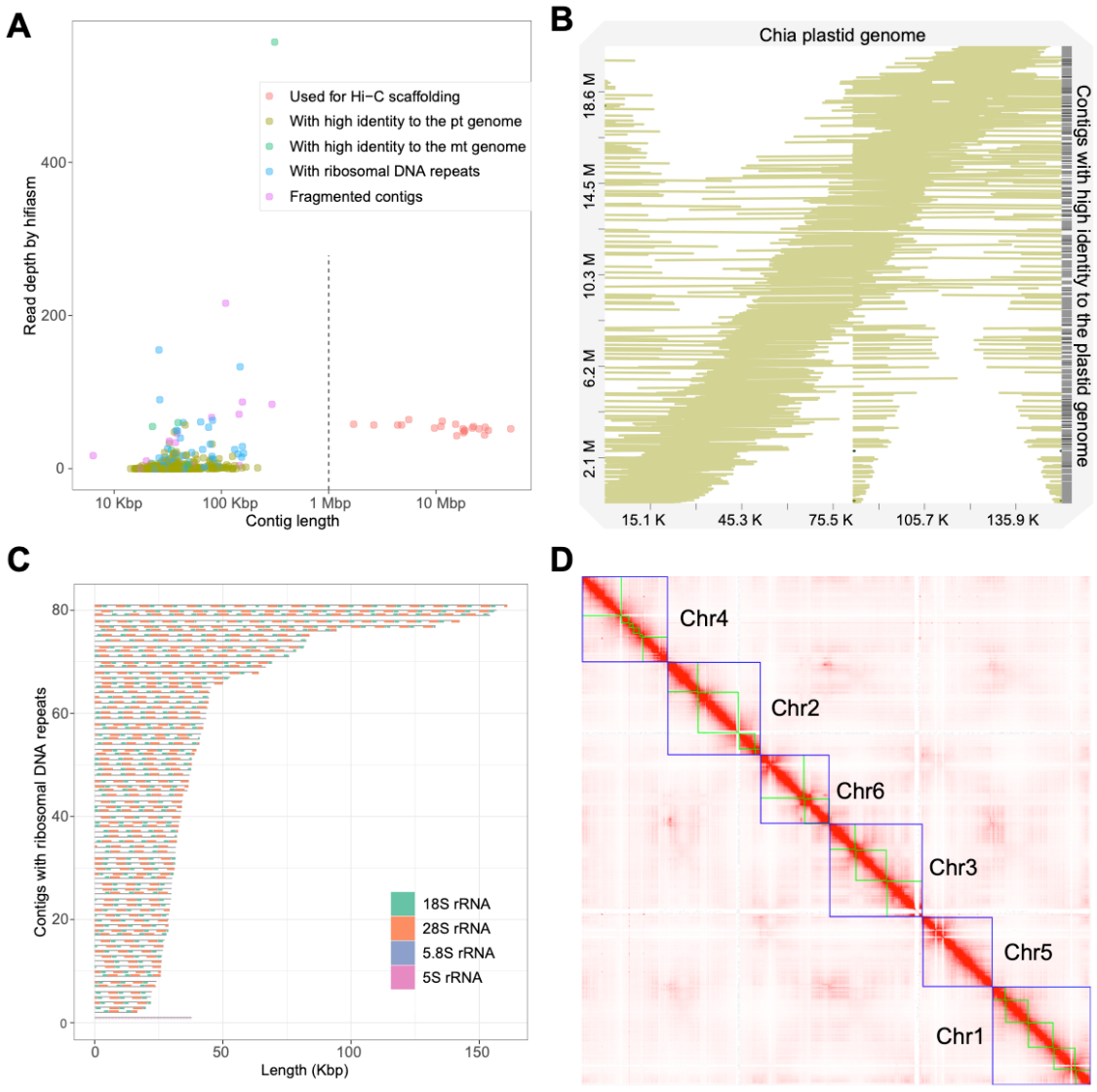
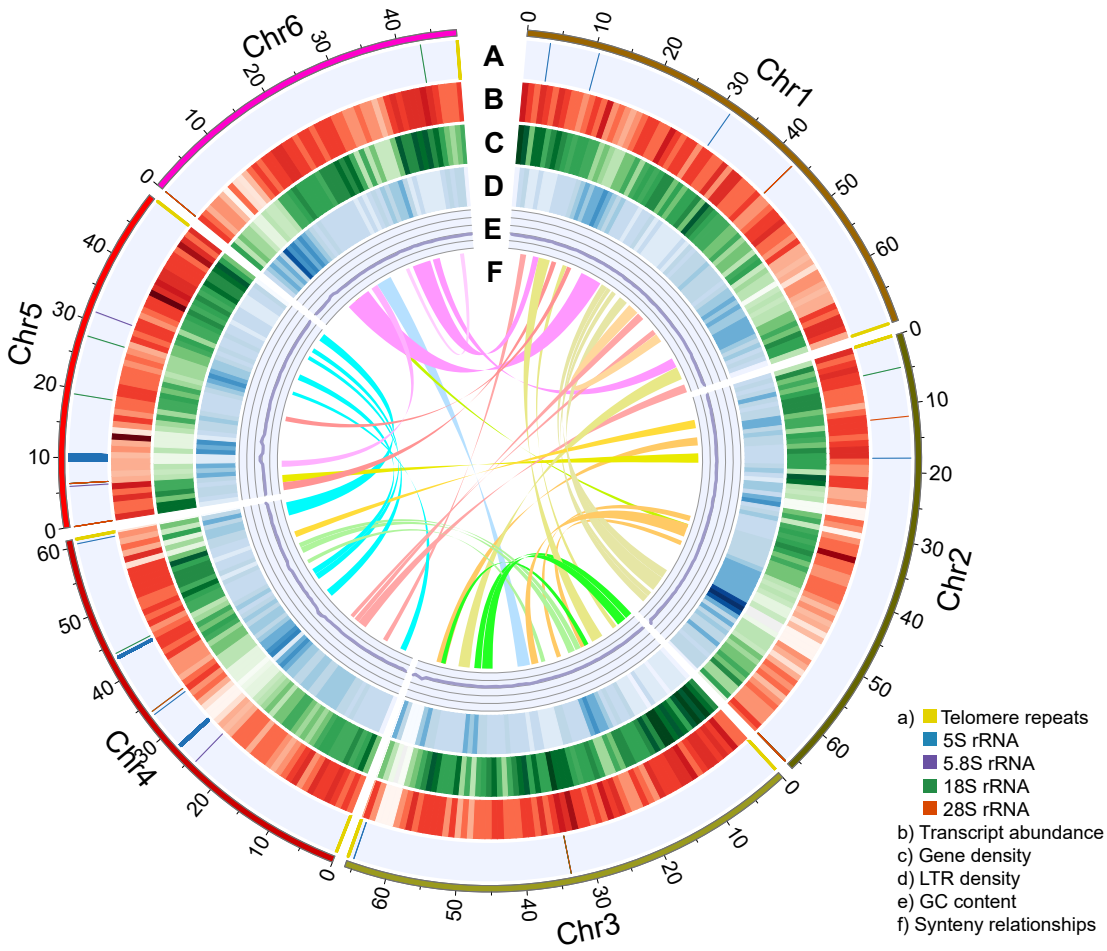


Figure 1 Assembly of the chia genome. **A)** Dotplot showing the contig length and the read depth of the initial assembly. Contigs were classified into five categories based on the length, the read depth and their origins, as indicated in the legend. **B)** Alignment of 538 initial contigs onto the chia plastid genome. **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.



747

748 **Figure 2** The nuclear genome of Shi_PSC_v1. Each ring indicates specific features of the

749 nuclear genome. Data from non-overlapping 1-Mb windows were graphed: **A)** Position of

750 telomere repeats and ribosomal RNA genes; **B)** Average transcript abundance; **C)** Gene

751 density; **D)** LTR density; **E)** GC content; **F)** Synteny blocks >1 Mb in length.

752

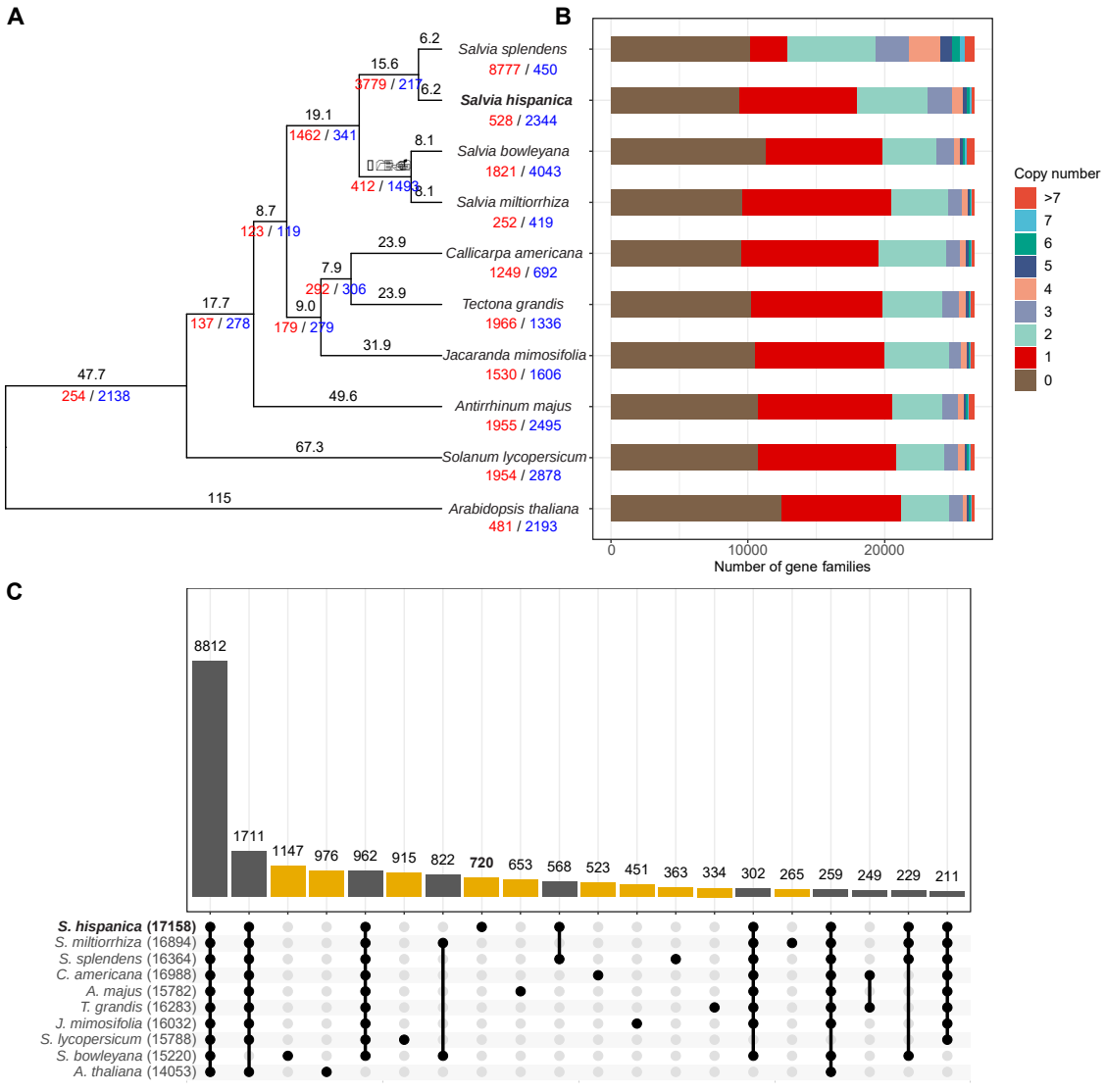


Figure 3 Evolution of the chia genome. **A)** Phylogenetic tree for chia and 9 other plant species. Numbers of expanded and contracted gene families were indicated by red and blue numbers at each branch point. Branch length indicate the estimated divergence time in million years ago. **B)** Numbers of gene families with different copy numbers in each plant species. **C)** Upset plot indicating the number of gene families shared by different species. Yellow bars indicate species-specific gene families.

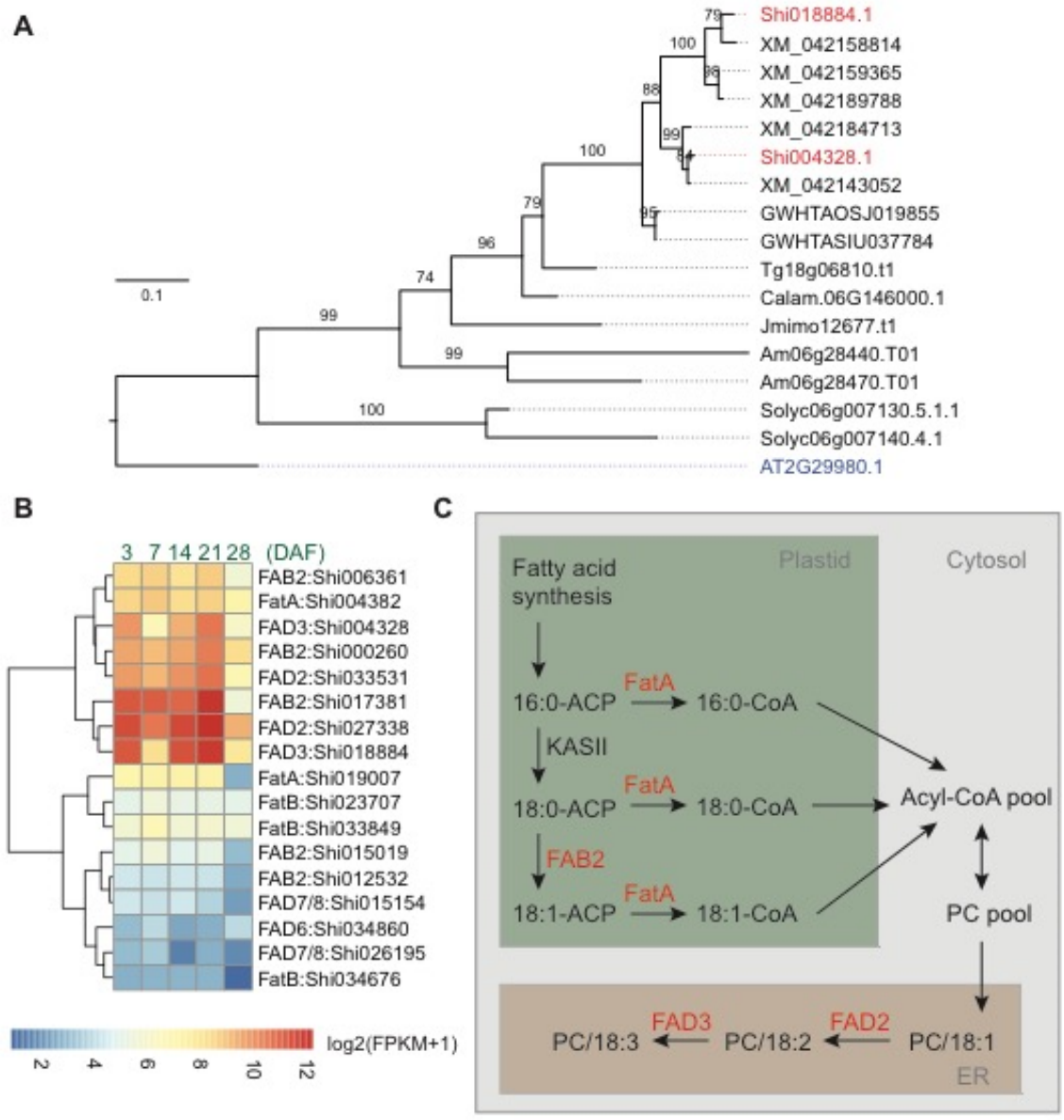


Figure 4 Identification of critical genes involved in fatty acid biosynthesis of chia seeds. **A**) Phylogenetic tree of the *FAD3* genes. Shi: *Salvia hispanica*; AT: *Arabidopsis thaliana*; XM: *Salvia splendens*; GWHTAOSJ: *Salvia miltiorrhiza*; GWHTASIU: *Salvia bowleyana*; Tg: *Tectona grandis*; 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

770 model for the biosynthesis of ALA in the chia genome. PC: phosphatidylcholine; ER:

771 endoplasmic reticulum.

772

Parsed Citations

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. The authors declare no conflict of interest.

References

Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H (2020) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 36: 2251-2252
Google Scholar: Author Only Title Only Author and Title

Baker EJ, Miles EA, Burdge GC, Yaqoob P, Calder PC (2016) Metabolism and functional effects of plant-derived omega-3 fatty acids in humans. Prog Lipid Res 64: 30-56
Google Scholar: Author Only Title Only Author and Title

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19: 455-477
Google Scholar: Author Only Title Only Author and Title

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580
Google Scholar: Author Only Title Only Author and Title

Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, Richardson L, Salazar GA, Williams L, Bork P, Bridge A, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA, Necci M, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A, Finn RD (2021) The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 49: D344-D354
Google Scholar: Author Only Title Only Author and Title

Boecker F (2021) AHRD: Automatically Annotate Proteins with Human Readable Descriptions and Gene Ontology Terms. Universitäts-und Landesbibliothek Bonn
Google Scholar: Author Only Title Only Author and Title

Burns-Whitmore B, Froyen E, Heskey C, Parker T, 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

Cabanettes F, Klopp C (2018) D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6: e4958
Google Scholar: Author Only Title Only Author and Title

Campbell MS, Holt C, Moore B, Yandell M (2014) Genome Annotation and Curation Using MAKER and MAKER-P. Curr Protoc Bioinformatics 48: 4 11 11-39
Google Scholar: Author Only Title Only Author and Title

Cassiday L (2017) Chia: superfood or superfad? Inform 28: 6-13
Google Scholar: Author Only Title Only Author and Title

Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD (2017) Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J 89: 789-804
Google Scholar: Author Only Title Only Author and Title

Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18: 170-175
Google Scholar: Author Only Title Only Author and Title

Ciftci ON, Przybylski R, Rudzińska M (2012) Lipid components of flax, perilla, and chia seeds. European Journal of Lipid Science and Technology 114: 794-800
Google Scholar: Author Only Title Only Author and Title

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21
Google Scholar: Author Only Title Only Author and Title

Dong AX, Xin HB, Li ZJ, Liu H, Sun YQ, Nie S, Zhao ZN, Cui RF, Zhang RG, Yun QZ, Wang XN, Maghuly F, Porth I, Cong RC, Mao JF (2018) High-quality assembly of the reference genome for scarlet sage, Salvia splendens, an economically important ornamental plant. Gigascience 7
Google Scholar: Author Only Title Only Author and Title

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL (2017) De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 356: 92-95

Google Scholar: Author Only Title Only Author and Title

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL (2016) Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst 3: 99-101
Google Scholar: Author Only Title Only Author and Title

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL (2016) Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst 3: 95-98
Google Scholar: Author Only Title Only Author and Title

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797
Google Scholar: Author Only Title Only Author and Title

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9: 18
Google Scholar: Author Only Title Only Author and Title

Emms D, Kelly S (2018) STAG: species tree inference from all genes. BioRxiv: 267914
Google Scholar: Author Only Title Only Author and Title

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20: 238
Google Scholar: Author Only Title Only Author and Title

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A 117: 9451-9457
Google Scholar: Author Only Title Only Author and Title

Freudenthal JA, Pfaff S, Terhoeven N, Korte A, Ankenbrand MJ, Forster F (2020) A systematic comparison of chloroplast genome assembly tools. Genome Biol 21: 254
Google Scholar: Author Only Title Only Author and Title

Gene Ontology C (2021) The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 49: D325-D334
Google Scholar: Author Only Title Only Author and Title

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644-652
Google Scholar: Author Only Title Only Author and Title

Greiner S, Lehwark P, Bock R (2019) OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res 47: W59-W64
Google Scholar: Author Only Title Only Author and Title

Gupta P, Geniza M, Naithani S, Phillips JL, Haq E, 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
Google Scholar: Author Only Title Only Author and Title

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9: R7
Google Scholar: Author Only Title Only Author and Title

Hamilton JP, Godden GT, Lanier E, Bhat WW, Kinser TJ, Vaillancourt B, Wang H, Wood JC, Jiang J, Soltis PS, Soltis DE, Hamberger B, Buell CR (2020) Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, Callicarpa americana. Gigascience 9
Google Scholar: Author Only Title Only Author and Title

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S (2015) Tree of life reveals clock-like speciation and diversification. Mol Biol Evol 32: 835-845
Google Scholar: Author Only Title Only Author and Title

Hosmani PS, Flores-Gonzalez M, van de Geest H, Maumus F, Bakker LV, Schijlen E, van Haarst J, Cordewener J, Sanchez-Perez G, Peters S (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
Google Scholar: Author Only Title Only Author and Title

Jia KH, Liu H, Zhang RG, Xu J, Zhou SS, Jiao SQ, Yan XM, Tian XC, Shi TL, Luo H, Li ZC, Bao YT, Nie S, Guo JF, Porth I, El-Kassaby YA, Wang XR, Chen C, Van de Peer Y, Zhao W, Mao JF (2021) Chromosome-scale assembly and evolution of the tetraploid Salvia splendens (Lamiaceae) genome. Hortic Res 8: 177
Google Scholar: Author Only Title Only Author and Title

Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, Li DZ (2020) GetOrganelle: a fast and versatile toolkit for accurate de

novo assembly of organelle genomes. Genome Biol 21: 241
Google Scholar: Author Only Title Only Author and Title

Joh Y-G, Lee O-K, Lim Y-J (1988) Studies on the Composition of Fatty Acid in the Lipid Classes of Seed Oils of the Labiatae Family. Journal of the Korean Applied Science and Technology 5: 13-23
Google Scholar: Author Only Title Only Author and Title

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30: 1236-1240
Google Scholar: Author Only Title Only Author and Title

Jung J, Kim JI, Jeong YS, Yi G (2018) AGORA: organellar genome annotation from the amino acid and nucleotide references. Bioinformatics 34: 2661-2663
Google Scholar: Author Only Title Only Author and Title

Kachroo A, Shanklin J, Whittle E, Lapchyk L, Hildebrand D, 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
Google Scholar: Author Only Title Only Author and Title

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14: 587-589
Google Scholar: Author Only Title Only Author and Title

Kazaz S, Barthole G, Domergue F, Ettaki H, To A, Vasselon D, De Vos D, Belcram K, Lepiniec L, Baud S (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
Google Scholar: Author Only Title Only Author and Title

Keilwagen J, Hartung F, Grau J (2019) GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. Methods Mol Biol 1962: 161-177
Google Scholar: Author Only Title Only Author and Title

Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37: 540-546
Google Scholar: Author Only Title Only Author and Title

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27: 722-736
Google Scholar: Author Only Title Only Author and Title

Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5: 59
Google Scholar: Author Only Title Only Author and Title

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19: 1639-1645
Google Scholar: Author Only Title Only Author and Title

Kulczynski B, Kobus-Cisowska J, Taczanowski M, Kmiecik D, Gramza-Michalowska A (2019) The Chemical Composition and Nutritional Value of Chia Seeds-Current State of Knowledge. Nutrients 11
Google Scholar: Author Only Title Only Author and Title

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12
Google Scholar: Author Only Title Only Author and Title

Lands B (2014) Historical perspectives on the impact of n-3 and n-6 nutrients on health. Prog Lipid Res 55: 17-29
Google Scholar: Author Only Title Only Author and Title

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359
Google Scholar: Author Only Title Only Author and Title

Levy Karin E, Mirdita M, Soding J (2020) MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome 8: 48
Google Scholar: Author Only Title Only Author and Title

Li D, Wahlqvist ML, Sinclair AJ (2019) Advances in n-3 polyunsaturated fatty acid nutrition. Asia Pac J Clin Nutr 28: 1-5
Google Scholar: Author Only Title Only Author and Title

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997
Google Scholar: Author Only Title Only Author and Title

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094-3100
Google Scholar: Author Only Title Only Author and Title

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079
Google Scholar: Author Only Title Only Author and Title

Li M, Zhang D, Gao Q, Luo Y, Zhang H, Ma B, Chen C, Whibley A, Zhang Y, Cao Y, Li Q, Guo H, Li J, Song Y, Zhang Y, Copsey L, Li Y, Li X, Qi M, Wang J, Chen Y, Wang D, Zhao J, Liu G, Wu B, Yu L, Xu C, Li J, Zhao S, Zhang Y, Hu S, Liang C, Yin Y, Coen E, Xue Y (2019) Genome structure and evolution of Antirrhinum majus L. Nat Plants 5: 174-183
Google Scholar: Author Only Title Only Author and Title

Li-Beisson Y, Shorrosh B, Beisson F, Andersson MX, Arondel V, Bates PD, Baud S, Bird D, Debono A, Durrett TP, Franke RB, Graham IA, Katayama K, Kelly AA, Larson T, Markham JE, Miquel M, Molina I, Nishida I, Rowland O, Samuels L, Schmid KM, Wada H, Welti R, Xu C, Zallot R, Ohlrogge J (2013) Acyl-lipid metabolism. Arabidopsis Book 11: e0161
Google Scholar: Author Only Title Only Author and Title

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30: 923-930
Google Scholar: Author Only Title Only Author and Title

Liu HL, Yin ZJ, Xiao L, Xu YN, 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
Google Scholar: Author Only Title Only Author and Title

Lou Y, Schwender J, Shanklin J (2014) FAD2 and FAD3 desaturases form heterodimers that facilitate metabolic channeling in vivo. J Biol Chem 289: 17996-18007
Google Scholar: Author Only Title Only Author and Title

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15: 550
Google Scholar: Author Only Title Only Author and Title

Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM (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
Google Scholar: Author Only Title Only Author and Title

Marcais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764-770
Google Scholar: Author Only Title Only Author and Title

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal 17: 10-12
Google Scholar: Author Only Title Only Author and Title

Mendes FK, Vanderpool D, Fulton B, Hahn MW (2020) CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics
Google Scholar: Author Only Title Only Author and Title

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 37: 1530-1534
Google Scholar: Author Only Title Only Author and Title

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A (2021) Pfam: The protein families database in 2021. Nucleic Acids Res 49: D412-D419
Google Scholar: Author Only Title Only Author and Title

Muñoz LA, Cobos A, Diaz O, Aguilera JM (2013) Chia seed (Salvia hispanica): an ancient grain and a new functional food. Food reviews international 29: 394-408
Google Scholar: Author Only Title Only Author and Title

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S (2020) HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res 30: 1291-1305
Google Scholar: Author Only Title Only Author and Title

Ohlrogge J, Browse J (1995) Lipid biosynthesis. Plant Cell 7: 957-970
Google Scholar: Author Only Title Only Author and Title

Ou S, Chen J, Jiang N (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res 46: e126

Google Scholar: Author Only Title Only Author and Title

Ou S, Jiang N (2018) LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol 176: 1410-1422
Google Scholar: Author Only Title Only Author and Title

Ou S, Jiang N (2019) LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA 10: 48
Google Scholar: Author Only Title Only Author and Title

Poux S, Arighi CN, Magrane M, Bateman A, Wei CH, Lu Z, Boutet E, Bye AJH, Famiglietti ML, Roechert B, UniProt Consortium T (2017) On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. Bioinformatics 33: 3454-3460
Google Scholar: Author Only Title Only Author and Title

Qian J, Song J, Gao H, Zhu Y, Xu J, Pang X, Yao H, Sun C, Li X, Li C, Liu J, Xu H, Chen S (2013) The complete chloroplast genome sequence of the medicinal plant Salvia miltiorrhiza. PLoS One 8: e57607
Google Scholar: Author Only Title Only Author and Title

R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
Google Scholar: Author Only Title Only Author and Title

Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun 11: 1432
Google Scholar: Author Only Title Only Author and Title

Rhie A, Walenz BP, Koren S, Phillippy AM (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21: 245
Google Scholar: Author Only Title Only Author and Title

Shahidi F, Ambigaipalan P (2018) Omega-3 Polyunsaturated Fatty Acids and Their Health Benefits. Annu Rev Food Sci Technol 9: 345-381
Google Scholar: Author Only Title Only Author and Title

Simopoulos AP (2002) The importance of the ratio of omega-6/omega-3 essential fatty acids. Biomed Pharmacother 56: 365-379
Google Scholar: Author Only Title Only Author and Title

Simopoulos AP (2002) Omega-3 fatty acids in inflammation and autoimmune diseases. J Am Coll Nutr 21: 495-505
Google Scholar: Author Only Title Only Author and Title

Song Z, Lin C, Xing P, Fen Y, Jin H, Zhou C, Gu YQ, Wang J, Li X (2020) A high-quality reference genome sequence of Salvia miltiorrhiza provides insights into tanshinone synthesis in its red rhizomes. Plant Genome 13: e20041
Google Scholar: Author Only Title Only Author and Title

Sreedhar RV, Kumari P, Rupwate SD, Rajasekharan R, Srinivasan M (2015) Exploring triacylglycerol biosynthetic pathway in developing seeds of Chia (Salvia hispanica L.): a transcriptomic approach. PLoS One 10: e0123580
Google Scholar: Author Only Title Only Author and Title

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12: 1611-1618
Google Scholar: Author Only Title Only Author and Title

Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 Suppl 2: ii215-225
Google Scholar: Author Only Title Only Author and Title

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34: W609-612
Google Scholar: Author Only Title Only Author and Title

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008) Synteny and collinearity in plant genomes. Science 320: 486-488
Google Scholar: Author Only Title Only Author and Title

Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics Chapter 4: Unit 4 10
Google Scholar: Author Only Title Only Author and Title

Troncoso-Ponce MA, Barthole G, Tremblais G, To A, Miquel M, Lepiniec L, Baud S (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

Google Scholar: Author Only Title Only Author and Title

Valdivia-López MÁ, Tecante A (2015) Chia (Salvia hispanica): A review of native Mexican seed and its nutritional and functional properties. Advances in food and nutrition research 75: 53-75

Google Scholar: Author Only Title Only Author and Title

Wang D, Zhang Y, Zhang Z, Zhu J, Yu J (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics 8: 77-80

Google Scholar: Author Only Title Only Author and Title

Wang M, Zhang L, Wang Z (2021) Chromosomal-Level Reference Genome of the Neotropical Tree Jacaranda mimosifolia D. Don. Genome Biol Evol 13

Google Scholar: Author Only Title Only Author and Title

Wheeler TJ, Eddy SR (2013) nhmmer: DNA homology search with profile HMMs. Bioinformatics 29: 2487-2489

Google Scholar: Author Only Title Only Author and Title

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35: W265-268

Google Scholar: Author Only Title Only Author and Title

Xue Y, Chen B, Win AN, Fu C, Lian J, Liu X, Wang R, Zhang X, Chai Y (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

Google Scholar: Author Only Title Only Author and Title

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32-43

Google Scholar: Author Only Title Only Author and Title

Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, Dai L (2012) ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. Biochem Biophys Res Commun 419: 779-781

Google Scholar: Author Only Title Only Author and Title

Zhao D, Hamilton JP, Bhat WW, Johnson SR, Godden GT, Kinser TJ, Boachon B, Dudareva N, Soltis DE, Soltis PS, Hamberger B, Buell CR (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

Google Scholar: Author Only Title Only Author and Title

Zheng X, Chen D, Chen B, Liang L, Huang Z, Fan W, Chen J, He W, Chen H, Huang L, Chen Y, Zhu J, Xue T (2021) Insights into salvianolic acid B biosynthesis from chromosome-scale assembly of the Salvia bowleyana genome. J Integr Plant Biol 63: 1309-1323

Google Scholar: Author Only Title Only Author and Title