Author/s:
Zare, T;Paril, JF;Barnett, EM;Kaur, P;Appels, R;Ebert, B;Roessner, U;Fournier-Level, A

Title:
Comparative genomics points to tandem duplications of SAD gene clusters as drivers of increased α-linolenic (ω-3) content in S. hispanica seeds

Date:
2024-03-01

Citation:
Zare, T., Paril, J. F., Barnett, E. M., Kaur, P., Appels, R., Ebert, B., Roessner, U. & Fournier-Level, A. (2024). Comparative genomics points to tandem duplications of SAD gene clusters as drivers of increased α-linolenic (ω-3) content in S. hispanica seeds. Plant Genome, 17 (1), https://doi.org/10.1002/tpg2.20430.

Persistent Link:
https://hdl.handle.net/11343/351175

Received: 18 August 2023 | Accepted: 2 January 2024

DOI: 10.1002/tpg2.20430

The Plant Genome

ORIGINAL ARTICLE

Comparative genomics points to tandem duplications of *SAD* gene clusters as drivers of increased *α*-linolenic (*ω*-3) content in *S. hispanica* seeds

Tannaz Zare[1] | **Jeff F. Paril**[1] | **Emma M. Barnett**[1] | **Parwinder Kaur**[4] | **Rudi Appels**[5] | **Berit Ebert**[2] | **Ute Roessner**[3] | **Alexandre Fournier-Level**[1]

[1]School of BioSciences, The University of Melbourne, Parkville, Victoria, Australia

[2]School of Biology and Biotechnology, Ruhr-Universitat Bochum, Bochum, Germany

[3]Research School of Biology, The Australian National University, Canberra, Australian Capital Territory, Australia

[4]School of Agriculture and Environment, The University of Western Australia, Perth, Western Australia, Australia

[5]School of Agriculture, Food and Ecosystem Sciences, University of Melbourne, Parkville, Victoria, Australia

Correspondence
Alexandre Fournier-Level, School of BioSciences, The University of Melbourne, Parkville, 3010 VIC, Australia.
Email: afournier@unimelb.edu.au

Funding information
Research Training Program Scholarship; Alfred Nicholas Fellowship; Megan Klemm Postgraduate Research Scholarship; Norma Hilda Schuster (nee Swift) Scholarship

Abstract

Salvia hispanica L. (chia) is a source of abundant *ω*-3 polyunsaturated fatty acids (*ω*-3-PUFAs) that are highly beneficial to human health. The genomic basis for this accrued *ω*-3-PUFA content in this emerging crop was investigated through the assembly and comparative analysis of a chromosome-level reference genome for *S. hispanica*. The highly contiguous 321.5-Mbp genome assembly covering all six chromosomes enabled the identification of 32,922 protein-coding genes. Two whole-genome duplications (WGD) events were identified in the *S. hispanica* lineage. However, these WGD events could not be linked to the high *α*-linolenic acid (ALA, *ω*-3) accumulation in *S. hispanica* seeds based on phylogenomics. Instead, our analysis supports the hypothesis that evolutionary expansion through tandem duplications of specific lipid gene families, particularly the stearoyl-acyl carrier protein desaturase (*ShSAD*) gene family, is the main driver of the abundance of *ω*-3-PUFAs in *S. hispanica* seeds. The insights gained from the genomic analysis of *S. hispanica* will help establish a molecular breeding target that can be leveraged through genome editing techniques to increase *ω*-3 content in oil crops.

Plain Language Summary

Chia is an emerging crop that has been qualified as "superfood" because of it is a rich source of omega 3 fatty acids which have benefits for human nutrition and health. To

Abbreviations: 4DTv, four-fold degenerative (synonymous) sites; AA, amino acid sequences; ACP, acyl carrier protein; ALA, *α*-linolenic acid; BUSCO, Benchmarking Universal Single-Copy Orthologs; CDS, coding DNA sequence; CoA, acyl-coenzyme A; CRISPR, clustered regularly interspaced short palindromic repeats; DGAT, acyl-coenzyme A: diacylglycerol acyltransferases; ER, endoplasmic reticulum; FA, fatty acid; FAD, fatty acid desaturases; FAS, fatty acid synthase; gDNA, genomic DNA; GO, Gene Ontology; KAR, 3-ketoacyl-acyl carrier protein reductase; LA, linoleic acid; lncRNA, long non-coding RNA; LTP1, type 1 lipid transfer; LTR, long terminal repeat; misc_RNA, miscellaneous RNA; mRNA, messenger RNA; MYA, million years ago; NA, nervonic acid; NCBI, National Center for Biotechnology Information; ncRNA, small nuclear RNA; OA, oleic acid; PC, phosphatidylcholine; PDAT, phospholipid:diacylglycerol acyltransferase; PUFA, polyunsaturated fatty acid; SAD, stearoyl-ACP desaturase; SCO, single-copy orthogroup; TE, transposable element; TG, triglycerides; UniProt, Universal Protein Resource; WGD, whole-genome duplication.

understand the genetic make-up underlying this abundant production of fatty acids, we sequence the genome of a black commercial variety of chia and compare the gene content against other plant species. We found that the number of genes of a specific family producing stearoyl-ACP desaturase enzymes has been expanded in this particular plant species. This expansion in number of genes from this family was shown to have ocurred through tandem duplication, which is the doubling of a genome chunk occurring during the multiplication of the reproductive cells. We also found that these enzyme are being put under Natural Selection to remain efficient and this, with their larger that expected number, might explain why chia produces so much omega 3s.

1 | INTRODUCTION

Salvia hispanica L. (chia) is an oleaginous short-day flowering plant originating from Mexico and widely cultivated throughout Latin America, Australia, and Southeast Asia (Jamboonsri et al., 2012). Member of the Lamiaceae family, which comprises nearly 7100 species of flowering plants, *S. hispanica* belongs to the largest genus, *Salvia* spp. (sages) which regroups more than 1000 species. *Salvia* spp. are increasingly recognized as commercially important crops due to their nutraceutical and bioactive compounds among which are sterols, flavonoids, diterpenes, triterpenes, and polyphenols (Cahill, 2003; Georgiev & Pavlov, 2017; Harley et al., 2004; Walker et al., 2004).

Among the *Salvia* spp., *S. hispanica* is the most nutritionally valuable crop due to its health-promoting properties, including high levels of dietary fiber (35%), carbohydrates (5%), protein (18%–24%), lipids (31%–34%), antioxidants, and essential vitamins (da Silva et al., 2017; Timilsena et al., 2016; Zare et al., 2019). *Salvia hispanica* seeds are rich in essential fatty acids (81%), including α-linolenic acid (ALA, ω-3, 62%) and linoleic acid (LA, ω-6, 19%) with low ω-6:ω-3 ratio (0.3), which makes it one of the best sources of plant-based ω-3 (Oteri et al., 2023). Several studies examining the effect of ω-3 polyunsaturated fatty acids (ω-3-PUFAs) supplementation on human health suggest that they may help reduce several chronic diseases such as diabetes, cardiovascular and inflammatory disorders, hypertension, dyslipidemia, and kidney dysfunction (Arredondo-Mendoza et al., 2020; Creus et al., 2017; El-Feky et al., 2022; Y.-X. Liu et al., 2023; Meyer & De Groot, 2017; Ong et al., 2023; Onneken, 2018; Penson & Banach, 2020; Y. Xiao et al., 2022; X. Zhang et al., 2022).

The biosynthesis of PUFAs in plants involves a series of complex reactions in different subcellular compartments (Wallis et al., 2002). The de novo biosynthesis of 16- or 18-carbon fatty acids (FAs) takes place in plastids through the action of acetyl-acyl-coenzyme A (CoA) carboxylase and FA synthase (FAS) (Li-Beisson et al., 2013). After the conversion/elongation of C16:0 to C18:0, the C18:0-acyl carrier protein (ACP) (stearic acid) is desaturated to C18:1-ACP (oleic acid [OA], ω-9, oleic acid) in the chloroplast stroma by a soluble stearoyl-ACP desaturase (SAD) (Bates et al., 2013). The C18:1-ACP is further desaturated into C16:3 and C18:2/C18:3 by different plastidial membrane-bound FA desaturases (FAD5, FAD6, and FAD7/FAD8) (Browse & Somerville, 1991). FAs are next exported to the endoplasmic reticulum (ER) for conversion into acyl-CoAs before forming phosphatidylcholines (PCs) and triglycerides (TGs) (Block & Jouhet, 2015). Once synthesized, TGs are assembled into oil bodies and exported from the ER to be stored in the seed (Banaś et al., 2013).

High-quality genomes are providing valuable information on the evolution and functional divergence of key genes involved in oil biosynthetic pathways (Badouin et al., 2017; Lin et al., 2022; Shen et al., 2022; Unver et al., 2017; L. Wang et al., 2014). Expansion of the type 1 lipid transfer (*LTP1*) gene family and contraction of lipid degradation genes have been linked to high oil accumulation in sesame seeds (L. Wang et al., 2014). Neo-functionalization and expansion of the *SAD* gene family are thought to be responsible for the increased levels of OA in olives (Unver et al., 2017). However, the lack of sufficient genomic information for *S. hispanica* limited the exploration of the genetic basis of ω-3-PUFAs accumulation in this plant.

Early research determined chia's somatic chromosome number and DNA content ($2n = 2x = 12$, C-value $= 0.93 \pm 0.016$ pg, genome size $= \sim$460 Mb) (Estilai et al., 1990; Haque, 1980; Maynard & Ruter, 2022). In recent years, several studies have provided multi-tissue transcriptomes for *S. hispanica* in order to identify genes involved in secondary metabolite and oil biosynthesis (Gupta et al., 2021; Peláez et al., 2019; Sreedhar et al., 2015; Wimberley et al., 2020). In addition, a set of studies functionally characterized genes encoding fatty acid desaturases (FADs) against different biotic/abiotic stresses (Xue et al., 2018,

2023). These studies, together with a genome assembly for *S. hispanica* (L. Wang et al., 2022), provided new insights, but relatively little is known about the main drivers of high ω-3-PUFA accumulation in *S. hispanica* seeds.

Our study investigated the molecular mechanisms of oil biosynthesis in *S. hispanica* leveraging the assembly of a near-complete, high-quality chromosome-level reference genome (RefSeq: GCF_023119035.1). This enabled comparative genomic analysis to determine the occurrence of WGD events and gene family size and sequence evolution between *S. hispanica* and a subset of relevant species. We investigated if specific biological functions were overrepresented among significantly expanded gene families. In particular, our analysis seeks to probe the hypothesis that duplication and nucleotides substitutions in oil biosynthesis genes support the high production of ω-3 FAs in *S. hispanica*.

Core Ideas

- A high-quality chromosome-level reference genome of *S. hispanica* was assembled and analyzed.
- Ancestral whole-genome duplication events have not promoted high α-linolenic acid content in *S. hispanica* seeds.
- Tandem duplication of six stearoyl-ACP desaturase genes is a plausible cause for high ω-3 content in chia seeds.

2 | MATERIALS AND METHODS

2.1 | Plant material and genomic DNA extraction

A black-seed variety of *S. hispanica* L. (TCC Black 2014) was sourced from Chia Co. and Northern Australia Crop Research Alliance. Fresh young leaves were harvested from a 4-week-old individual *S. hispanica* plant, immediately frozen in liquid nitrogen, and stored at −80°C prior to genomic DNA (gDNA) isolation. High molecular weight gDNA was isolated using a cetyltrimethylammonium bromide method (Figure S1). The isolated gDNA was treated with RNase A following the method developed by Yoshinaga and Dalin (2016) and purified using NucleoMag NGS magnetic beads (Macherey-Nagel) prior to DNA libraries synthesis.

2.2 | Library construction, sequencing, and processing of the sequencing reads

For short-read sequencing, DNA libraries were synthesized from 3.9 μg of gDNA using the Illumina TruSeq DNA PCR-Free kit (Illumina) and sequenced on Illumina NovaSeq 6000 platform (Illumina) in 2 × 150 bp sequencing mode. Genewiz (Suzhou, China) conducted the Illumina library synthesis and sequencing. The reads quality was assessed using FastQC v0.11.9 (Andrews, 2010). Low-quality reads with an average quality per base below Q20 calculated over 4-bp sliding windows and leading bases with a quality score below Q20 were removed using Trimmomatic v0.39 (Table S1; Bolger et al., 2014). A total of 476 Gb of high-quality Illumina reads with an average length of 145 bp were retained for genome assembly.

DNA libraries were prepared for long-read sequencing using the SQK-LSK109 ligation sequencing kit (Oxford Nanopore Technologies) and sequenced on a MinION Mk1B portable device with FLO-MIN106D flowcell. The long-read sequencing was run for 48 h at 180 mV using the MinKNOW software v.2.0. Basecalling of long sequencing reads was performed with Guppy v5.0.11+2b6dbffa5 using the basecalling template_r9.4.1_450 bps_hac.jsn (Oxford Nanopore community; https://community.nanoporetech.com). Long reads were error-corrected with fmlrc2 v0.1.5 (J. R. Wang et al., 2018), resulting in 9 Gb of high-quality reads with an average length of 2825 bp.

For Hi-C sequencing, nuclei were isolated from young leaves of an individual *S. hispanica* plant, and in situ Hi-C library synthesis was performed by DNA Zoo at the University of Western Australia (Perth, Australia) as described in Rao et al. (2014; Figure S2). The sequencing of the Hi-C libraries (~300 bp insert size) was carried out on an Illumina NovaSeq 6000 platform (Illumina) in the 2 × 150 bp mode by Genewiz.

2.3 | Estimation of the genome size and genomic heterozygosity

The genome size of *S. hispanica* was estimated through k-mer frequency analysis of the sequencing reads. The k-mer distributions for sizes ranging from 15 to 21 mer were computed using Jellyfish v2.3.0 (Marçais & Kingsford, 2011). Default settings were applied for the computation (jellyfish count -C -m 15–21 -s 1000000000 -t 16 *.fastq -o kmer_reads.jf). The genome size, level of heterozygosity, and abundance of genomic repeats were estimated using GenomeScope v1.0.0 (Vurture et al., 2017). This estimation involved analyzing a k-mer frequency histogram generated by jellyfish (Marçais & Kingsford, 2011).

2.4 | De novo genome assembly and scaffolding

A meta-assembly approach was conducted using a hybrid approach combining long and short reads (Supplementary Figure S3). The hybrid assembly consisted of combining

the contigs assembled from short reads and error-corrected long reads using Platanus-allee with default parameters v2.2.2 (Kajitani et al., 2019). The error-corrected long reads were also used to generate a long-read-only assembly using Wtdbg2 v2.5 (Ruan & Li, 2020) with default parameters. Long-read-only assembly and the consensus scaffolds from the hybrid assembly were integrated into a non-redundant meta-assembly using QuickMerge v0.3 (Chakraborty et al., 2016) with the parameters "-hco 5.0 -c 1.5 -l 1000 -ml 8000 -t 16." Iterative polishing was performed using Racon v1.4.22 (Vaser et al., 2017) with Illumina short and corrected long reads sequencing data. Ambiguous regions (N's) and gaps within contigs were filled using Cobbler v0.6.1 (Warren, 2016). The gap-free contigs were then re-merged using RAILS v1.5.1 (Warren, 2016), and duplicated regions (haplotigs) were purged using purge_dups v1.2.5 (Guan et al., 2020) to remove misassembled or redundant contigs from the final set of haplotigs retained in the assembly. The final contig assembly was obtained after one round of Illumina short-read polishing and two rounds of corrected long-read polishing with Racon v1.4.22 (Vaser et al., 2017). The final contig assembly was subsequently scaffolded with Hi-C reads using the Juicer pipeline (Durand et al., 2016). The Hi-C-based contact map was constructed using 3D-DNA v180419 (Dudchenko et al., 2017) and manually curated using the JuiceBox v1.11.08 (Durand et al., 2016).

2.5 | Assessment of the assembly completeness

Short and long reads were mapped to the assembled genome using bwa-mem v0.7.17 (H. Li & Durbin, 2009). Genome completeness was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.2.2 (Simão et al., 2015) with different databases: "eukaryota_odb10," "eudicots_odb10," "viridiplantae_odb10," and "embryophyta_odb10." Published transcriptomes from different tissues of *S. hispanica* (Gupta et al., 2021; Klein et al., 2021; Peláez et al., 2019; Sreedhar et al., 2015; Wimberley et al., 2020) were mapped to the assembled genome using blastn v2.10.1 (Camacho et al., 2009) to further validate the completeness of the assembly.

2.6 | Genome annotation

The genome of *S. hispanica* was annotated using the National Center for Biotechnology Information (NCBI) Eukaryotic Genome Annotation Pipeline (Pruitt et al., 2007). Transposable elements (TEs) were identified by constructing a de novo library of repetitive sequences based on the assembled genome using the RepeatModeler v2.0.3 (Flynn et al., 2020). The generated library was then used to classify the TEs and tandem genomic repeats and to mask the low-complexity sequences within the genome using RepeatMasker v4.1.2 (Tarailo-Graovac & Chen, 2009). Annotation features over the genome assembly were visualized as a circos plot generated by pyCircos v0.3.0 (https://github.com/ponnhide/pyCircos) and Matplotlib package v3.5.1 (Hunter, 2007).

2.7 | Comparative genomics analysis

Comparative analysis of the *S. hispanica* genome was performed against that of *Salvia splendens* (scarlet sage) as a closely related species, *Sesamum indicum* (sesame) and *Erythranthe guttata* (monkey flower) as representatives of the Lamiales order, *Solanum lycopercicum* (tomato) as a relatively close species with a high-quality genome; and *Arabidopsis thaliana* (thale cress) and *Vitis vinifera* (wine grape) as outgroups. The reference genome sequences and annotations for these species were retrieved from NCBI (Sayers et al., 2021). The comparative genome analysis in this study was conducted using the compare_genomes analysis pipeline (Paril et al., 2022, 2023).

Protein sequences from *S. hispanica* and the species compared were used to define gene families or orthogroups as clusters of homologous genes using OrthoFinder v2.5.4 (Emms & Kelly, 2019). The hmmsearch function from HMMER v3.3.2 (Mistry et al., 2013) was used to search for gene families that corresponded to the orthogroups identified in the Protein Analysis Through Evolutionary Relationships (PANTHER, http://pantherdb.org) gene family database using PantherHMM v16.0 (Mi et al., 2021). Evolutionary relationships between gene families and contraction and expansion of gene families across species were tested using CAFE5 v5.0 (Mendes et al., 2020). Gene enrichment among the expanded or contracted gene families was analyzed using the Gene Ontology (GO) enrichment analysis tool (Ashburner et al., 2000) from the Universal Protein Resource (UniProt) database (Bateman et al., 2020).

The orthogroups containing only single-copy orthologs across every species (one-to-one orthologs) were aligned with MACSE v2.06 (Ranwez et al., 2011) and used to construct a phylogenetic tree through maximum likelihood as implemented in IQ-TREE v2.0.7 (Minh et al., 2020). The IQ-TREE software was also used to estimate site-specific evolutionary rates and divergence times between species through an empirical Bayes approach. Divergence times between *A. thaliana* and *V. vinifera* (115 million years, MYA) and *S. indicum* and *Solanum lycopersicum* (82 MYA) were inferred from the TimeTree of Life database (http://timetree.org; Kumar et al., 2017). The divergence time between *S. splendens* and *S. hispanica* was retrieved from L. Wang et al. (2022). The rates of nucleotide substitution among pairs of paralogs/orthologs were measured using the third codon transversion rates at fourfold degenerative (synonymous) sites (4DTv) to estimate

the likelihood of WGD events. The 4DTv values were calculated based on the alignment of each pair of coding DNA sequences (CDSs) within orthogroups across the selected species using MACSE v2.06 (Ranwez et al., 2011).

2.8 | Analysis of oil biosynthesis genes

The evolution of key lipid biosynthesis pathway enzymes was compared between *S. hispanica* and *S. splendens*, *S. indicum*, *E. guttata*, *S. lycopercicum*, *V. vinifera*, and *A. thaliana*. We focused on 35 well-characterized genes that play a role in various lipid biosynthesis pathways, including plastidial FA synthesis, FAS II, phospholipid and glycerolipid synthesis, TG biosynthesis, as well as genes involved in FA elongation, desaturation, and export. Gene and protein sequences were sourced from NCBI and UniProt and are listed in Table S2. The protein sequences encoded by these lipid pathway genes were queried against the protein sequences of genes annotated in our focal species using blastp (E-value $\leq 1e^{-4}$) to identify the orthogroups encoding for specific gene activity. Contraction and expansion of these gene families and rates of nucleotide substitution were tested as described in the previous section.

Gene duplication events, including WGD, tandem duplication, proximal duplication, transposed duplication, and dispersed duplication, were identified using the DupGen_finder pipeline (Qiao et al., 2019). Protein sequences were first aligned using blastp with E-value $< 1e^{-5}$, and the different modes of gene duplications between homologous gene pairs determined using the DupGen_finder.pl function from MCScanX (Qiao et al., 2019) with the following parameters were used: match_score: 50, match_size: 5, gap_penalty: −1, overlap_window: 5, e_value: $1e^{-5}$, and max gaps: 25. The chromosome ideogram plot and homologous synteny blocks were generated using the R\RIdeogram package (Hao et al., 2020).

The ratio between the number of nonsynonymous substitutions per nonsynonymous site (Ka) and the number of synonymous substitutions per synonymous site (Ks) in a pairwise alignment of two orthologous sequences was used to measure evolutionary differences between sequences. The KaKs_Calculator2.0 v2.0 (D. Wang et al., 2010) was used to calculate Ka/Ks ratio over 15-bp sliding windows of the CDSs of paralogs associated with lipid metabolism in the *S. hispanica* genome and homologous genes in other species.

2.9 | Gene expression analysis

The transcriptomic data of nine different tissue types were obtained from the NCBI database (accession number: PRJEB19614) (Gupta et al., 2021). The quality of the raw RNAseq reads was assessed using FastQC v0.11.9 (Andrews, 2010). The low-quality reads were removed from the raw data with Trimmomatic v0.39 (Bolger et al., 2014) using the following parameters: SLIDINGWINDOW:10:28 HEADCROP:50 MINLEN:80. The sequencing reads were then mapped to the *S. hispanica* reference genome with STAR v2.7.10a (Dobin et al., 2013). The expression levels of transcripts were quantified using Cuffquant v2.2.1 (Trapnell et al., 2012) by implementing the *S. hispanica* annotation file. The abundance of the transcripts was quantified using the FeatureCounts implemented in the R\Subread package v2.0.3 (Liao et al., 2019). Differential gene expression analysis was performed using edgeR v3.38.4 (Y. Chen et al., 2016; Robinson et al., 2010) and limma v3.52.3 (Law et al., 2014; Ritchie et al., 2015) packages. Dispersion estimates were fitted to the negative binomial generalized linear model, and the quasi-likelihood *F*-test was used to identify differentially expressed genes (Y. Chen et al., 2016; Lun et al., 2016; Robinson et al., 2010).

3 | RESULTS

3.1 | Chromosome-scale reference genome assembly and annotation of *S. hispanica*

The genome assembled for *S. hispanica* (RefSeq: GCF_023119035.1) consisted of 5304 contigs spanning 1556 scaffolds. The assembly covered ~321 Mb (N50 = 53 Mb; L50 = 3; largest scaffold = 58 Mb) with a GC content of 36% (Table S3). The statistics of the assembled genome in this study was compared with the concurrently published genome (L. Li et al., 2023) in Table S3. Hi-C reads analysis identified ~173 million contacts (Table S4), of which ~127 million and ~46 million were inter- and intra-chromosomal contacts used for genome super-scaffolding, respectively. The size of *S. hispanica* pseudo-chromosomes obtained through Hi-C scaffolding ranged from 40 to 58 Mb with spanned gaps of 491 to 741 bp (Table S5). The L90 = 6 (Figure S4) is consistent with the chromosome number reported by Maynard and Ruter (2022). The best k-mer distribution model was obtained for $k = 19$ and supported diploidy with 0.24% heterozygosity and 5.28% duplicated regions (Figure S5). Analysis of k-mer frequencies estimated a haploid genome size of 466 Mbp, consistent with the size of 460 Mbp reported by Maynard and Ruter (2022).

The completeness of the assembly assessed against a different set of lineage-specific core eukaryotic genes (Eukaryota $n = 255$, Eudicots $n = 2117$, Embryophyte $n = 1538$, and Viridiplantae $n = 410$) resulted in the retrieval of 98.4%, 93.6%, 95.3%, and 96.5% of complete single copy gene models, respectively (Figure S6). The average BUSCO score was relatively high (>95%) across all lineage sets. Around 94%

TABLE 1 Genomic features annotated in the *S. hispanica* genome (excluding pseudogenes).

Feature	Count	Mean length (bp)	Median length (bp)	Min length (bp)	Max length (bp)
Genes	36,993	2901	2291	62	163,900
All transcripts	54,009	1671	1452	62	16,722
mRNA	46,423	1753	1515	165	16,722
Misc_RNA	2381	2122	1859	167	13,008
tRNA	739	74	73	71	93
lncRNA	3758	979	729	78	5754
snoRNA	436	106	103	62	229
snRNA	223	138	120	98	197
rRNA	49	384	119	103	3191
Single exon transcripts	5396	1149	957	233	6688
CDSs	46,508	1379	1155	90	16,188
Exons	209,379	302	162	2	7672
Exons in coding transcripts	197,070	302	161	2	7062
Exons in non-coding transcripts	19,006	274	153	2	7672
Introns	166,729	355	142	30	99,611
Introns in coding transcripts	158,426	343	139	30	99,611
Introns in non-coding transcripts	14,710	437	196	32	63,506

Abbreviations: CDSs, coding DNA sequences; lncRNA, long non-coding RNA; Min, minimum; Max, maximum; mRNA, messenger RNA; rRNA, ribosomal RNAs; snoRNA, small nucleolar RNAs; snRNA, small nuclear RNAs; tRNA; transfer RNAs.

of the previously published *S. hispanica* transcripts (Gupta et al., 2021; Klein et al., 2021; Peláez et al., 2019; Sreedhar et al., 2015; Wimberley et al., 2020) mapped to the assembled genome (Figure S7). The high BUSCO score and the high mapping rate of transcripts indicated that the genome assembly contained nearly all the *S. hispanica* genes.

Additionally, 97.56% of the short reads re-mapped against the assembled genome indicating the high quality of the *S. hispanica* reference genome. However, 0.6% of the reads did not have their paired read mapped to the genome, and 5.9% of paired reads were mapped to a different chromosome. This potentially highlights repetitive sequences in the assembled genome and closely matches the estimated genome duplication rate of 5.28% determined by GenomeScope.

A total of 46,508 CDSs were annotated, encompassing 209,379 exons and 166,729 introns across all transcripts including messenger RNAs (mRNAs), miscellaneous RNAs (misc_RNAs), and small nuclear RNAs (ncRNAs) of class long non-coding RNAs (lncRNA). The repeat-masked assembly contained 39,616 genes, corresponding to 32,922 protein-coding genes, 4071 non-coding genes, and 2623 pseudogenes (Table S6). The total number of annotated transcripts (54,009) included 46,423 mRNAs, 3758 lncRNAs, 739 transfer RNAs, 436 small nucleolar RNAs, 223 ncRNAs, 49 ribosomal RNAs, and 2381 misc_RNAs (Table 1).

The genome of *S. hispanica* contained 44.14% of interspersed repeats, 1.81% and 0.43% of which were simple and low-complexity repeats, respectively (Table S7 and Figure S8). Long terminal repeats (LTRs) represented 13.54% of the genome, with Copia (41.35% of all LTRs) and Gypsy (40.26% of all LTRs) being the most abundant (Figure 1h), while DNA transposons represented only 3.69% of the genome (Table S7).

3.2 | Gene family evolution in the *S. hispanica* genome

The CDSs from the *S. hispanica* genome annotation were compared to those of *S. splendens*, *S. indicum*, *E. guttata*, *S. lycopercicum*, *V. vinifera*, and *A. thaliana* (Figure 2). The phylogeny inferred from 134 single-copy orthologs (SCOs) supported previously described evolutionary relationships among species. The SCOs alignment placed the *S. splendens* and *S. hispanica* with maximum nodal support, confirming their close ancestral relationship (Figure 2a). Their most recent common ancestor was dated to 9.6 MYA which supported previous report by L. Wang et al. (2022). The oilseed crops *S. indicum* and *S. hispanica* were estimated to have diverged ~58–59 MYA, similar to the estimated divergence time between *E. guttata* and *S. hispanica* (Figure 2a).

From the 320,180 genes found across all seven species, 305,234 genes (95.3%) were assigned to one or more of the 27,963 orthogroups. From the 46,608 CDSs annotated in the *S. hispanica genome*, 45,108 genes were assigned to an orthogroup, 134 being SCOs, and 2814 unique

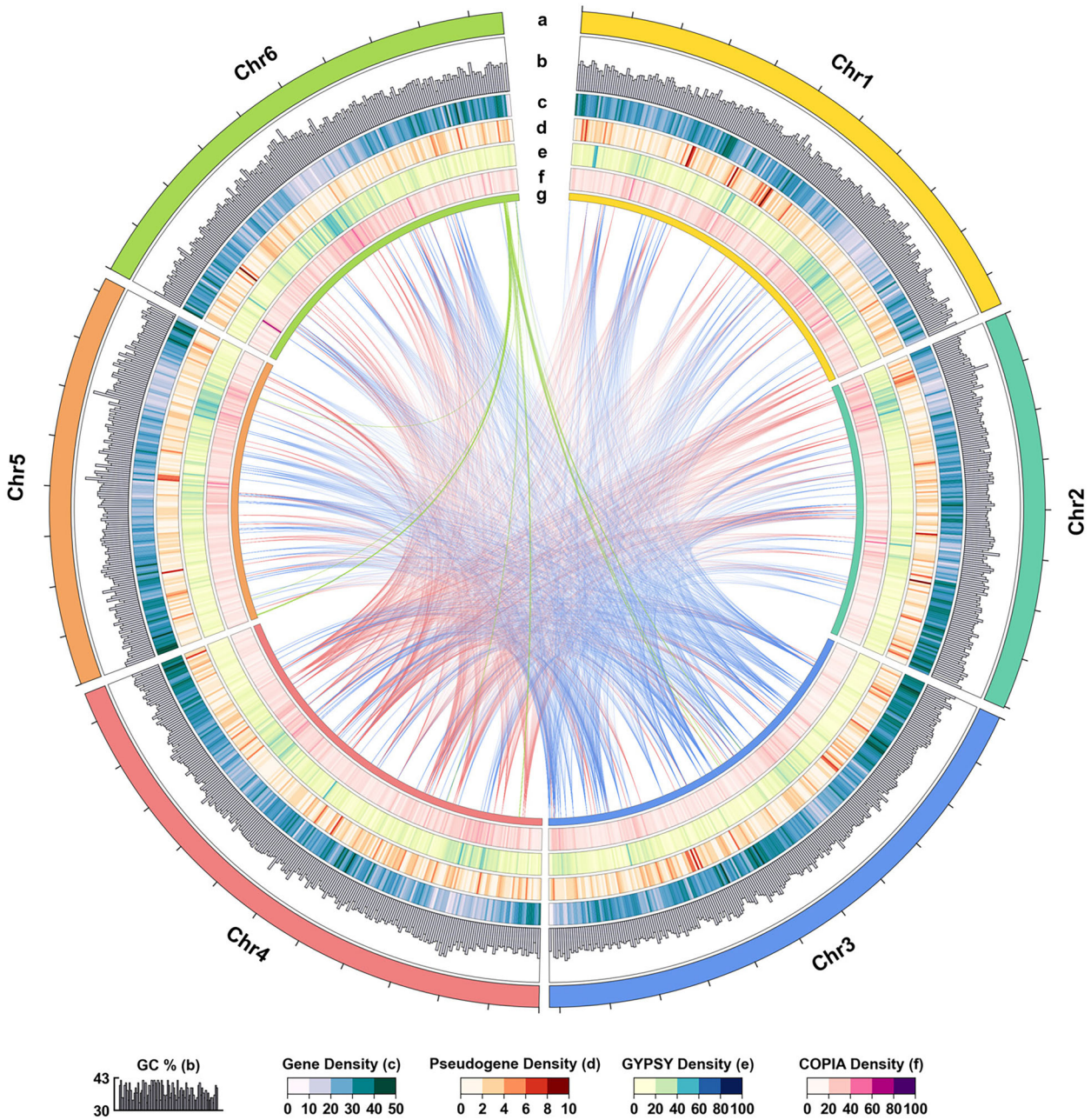

FIGURE 1 Genomic features of the *S. hispanica* reference genome assembly. (a) Chromosome layer showing the length of each chromosome with ticks indicating 5-Mbp intervals. (b) Distribution of the GC content at a window size of 250 kbp over the entire genome as a bar plot of percentage values with a lower bound of 30% and upper bound of 43%. (c) Distribution of protein-coding gene density over 250-kbp windows (values normalized between 0 and 50 across chromosomes). (d) Distribution of pseudogene density over 250-kbp windows (values normalized between 0 and 10 for all chromosomes). (e) and (f) Distribution of Gypsy and Copia long terminal repeat (LTR) density, respectively, over 250-kbp windows (values normalized between 0 and 100 across chromosomes). (g) The chord plot shows the synteny relationships across the genome's top five orthogroups (paralogs). The color of the internal chords is that of the chromosome containing the highest number of paralogs within each orthogroup.

paralogs form 682 orthogroups, leaving 1400 unassigned genes (Figure 2b). *Salvia splendens* contained the highest average number of paralogs within orthogroup (1.85), showing the highest genetic redundancy, followed by *A. thaliana* (1.12), *S. hispanica* (1.08), *V. vinifera* (0.96), *S. lycopersicum* (0.88), *S. indicum* (0.83), and *E. guttata* (0.74). The highest number of unique orthogroups was observed for *A. thaliana* (3047; 13,074 paralogs), followed by *S. splendens* (1471; 6236 paralogs), *V. vinifera* (1257; 6892 paralogs), *S. lycopersicum* (972; 4601 paralogs), *S. hispanica* (682; 2814 paralogs), *E. guttata* (617; 3738 paralogs), and *S. indicum* (402; 1790 paralogs; Figure 2b). The number of genes not assigned to any orthogroup varied across species, with the highest value for *S. splendens* (5081) followed by *A. thaliana*

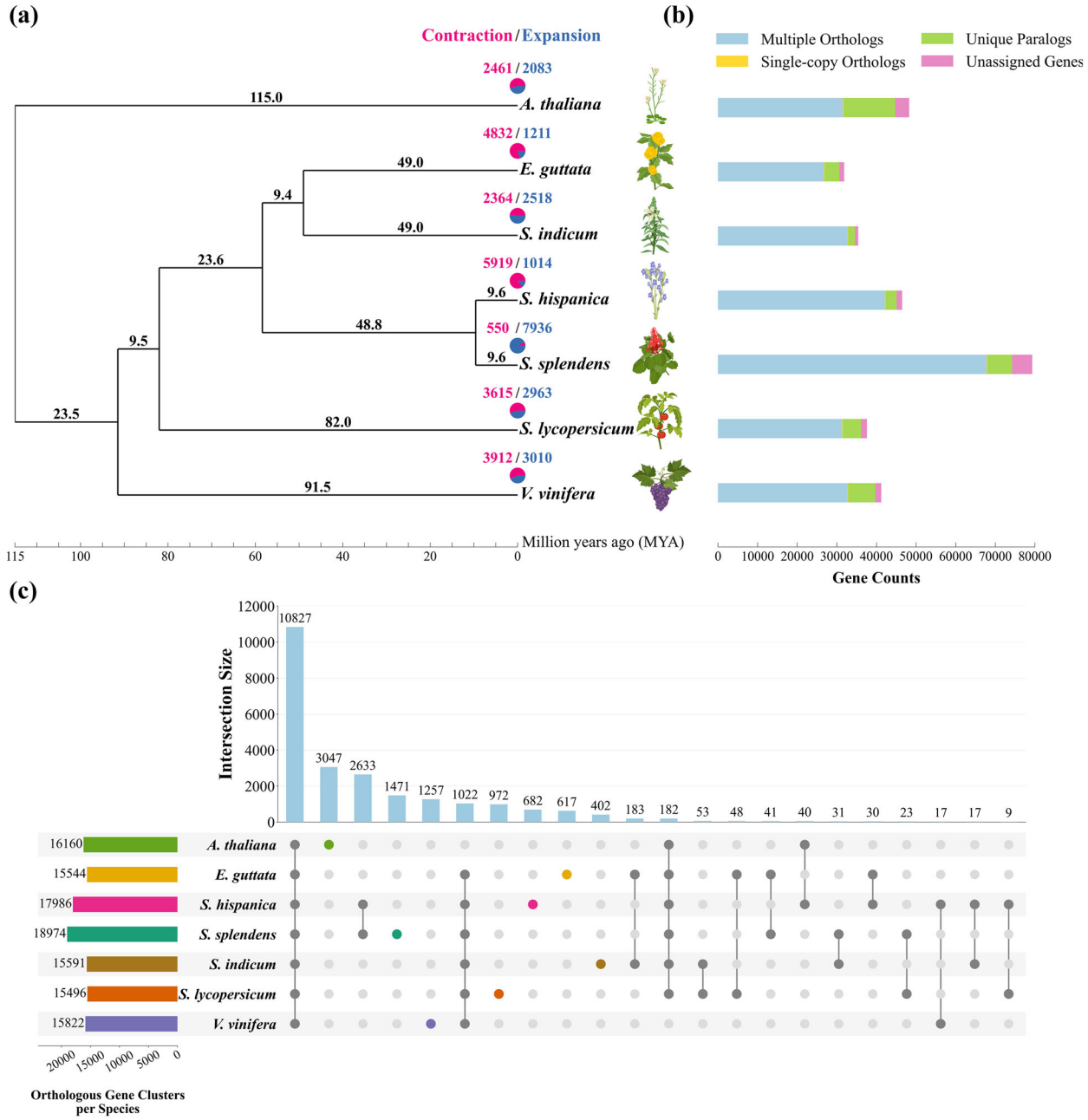

FIGURE 2 Evolution of *S. hispanica* and distribution of orthologous gene families across species. (a) Phylogenetic tree inferred from single-copy orthologs among selected species. Numbers on branches show divergence time in million years ago (MYA). The pie charts at the terminal branches show the contraction (pink) and expansion (dark blue) of gene families for each species. (b) Distribution of multiple orthologs, single copy orthologs, unique paralogs, and genes not associated with orthologs per species from orthogroup clustering by OrthoFinder. (c) The UpSet plot of the interactions between unique and shared orthologous gene clusters identified by OrthoFinder. The horizontal bar plot on the left shows the total number of orthogroups assigned to each species. The dark dots connected by solid lines show the species that include each cluster where the number of orthogroups within that cluster is indicated by the vertical bars on the top. Colored dots on the cluster map represent orthogroups unique to a species. Plant images are created with BioRender.com.

(3503), *S. lycopersicum* (1489), *V. vinifera* (1452), *S. hispanica* (1400), *E. guttata* (1216), and *S. indicum* (805; Figure 2b).

In total, 10,827 orthogroups were shared among all species (Figure 2c). The two closely related *Salvia* species (i.e., *S. splendens* and *S. hispanica*) contained the largest number of orthogroups (18,974 and 17,986, respectively), which is 15%–22% higher than that observed for other species. The high number of unique orthogroups in *A. thaliana* (3047) reflected the distant evolutionary relationships with the other species analyzed and its relevance as an outgroup (Figure 2c).

We next explored gene family expansion and contraction in *S. hispanica* compared to other selected species. *Arabidopsis thaliana*, *S. indicum*, *S. lycopersicum*, and *V. vinifera* showed a relatively even number of expanded versus contracted gene families (Figure 2a). *Erythranthe guttata* and *S. hispanica*, on the other hand, show a much higher number of contracted gene families, while *S. splendens* was the only species with a significant number of expanded gene families (7936). Interestingly, closely related *S. hispanica* exhibited the opposite pattern with a significant excess of contracted gene families (5919).

Among the gene families expanded in *S. hispanica*, significant enrichment was found for the maintenance of plant homeostasis, response to stress, and activation of defense mechanisms (Table S8). The top 10 gene families most unique to *S. hispanica* were highly enriched for specific biological processes: xenobiotic detoxification by transmembrane export in the plasma membrane (GO:1990961; $p < 5.70E^{-10}$), xenobiotic export from the cell (GO:0046618; $p < 5.70E^{-10}$), xenobiotic transport (GO:0042908; $p < 9.63E^{-11}$), plant-type primary cell wall biogenesis (GO:0009833; $p < 3.22E^{-04}$), galactose metabolic process (GO:0006012; $p < 9.83E^{-04}$), peptidyl-threonine dephosphorylation (GO:0035970; $p < 4.17E^{-08}$), toxin catabolic process (GO:0009407; $p < 8.70E^{-07}$), nucleotide-sugar transmembrane transport (GO:0015780; $p < 2.30E^{-02}$), S-glycoside catabolic process (GO:0016145; $p < 2.14E^{-04}$), and glucosinolate catabolic process (GO:0019762; $p < 2.14E^{-04}$; Figure S9). The 20 most enriched molecular functions and cellular component ontologies in the *S. hispanica* genome are presented in Table S9 and S10, respectively.

3.3 | Whole-genome duplications and speciation events

The occurrence of whole-genome duplication (WGD) events in the species studied was determined based on the distribution of the 4DTv among multi-copy paralogs (Figure S10). The 4DTv distribution for *S. hispanica* (Figure S10h) showed a high density at 0.1 (relative time to the most recent common ancestor) and at 0.3. The first peak at 0.1 corresponds to a relatively recent WGD event shared between *S. hispanica* and its most closely related species in our study, *S. splendens*. However, this event at 0.1 is not evident in the 4DTv distribution of *S. splendens* due to the masking effect of a very recent WGD event at 0.03 (Figure S10h). Both *S. indicum* and *S. lycopersicum* showed similar peaks at around 0.2 and 0.4, suggesting a more recent WGD; these peaks are less apparent for *E. guttata*, *A. thaliana*, and *V. vinifera* (Figure S10h).

The pairwise comparison of the 4DTv distribution in the two *Salvia* species indicated that the speciation event between them might have occurred quite recently (9.6 MYA as shown in Figure 2a) after a common WGD event shared across all *Salvia* species (Figure 3) and before the recent WGD event private to *S. splendens*. A comparison of *S. hispanica* with *S. indicum* and *E. guttata* (Figure 3) revealed that the *S. hispanica* genome has diverged from both species at the same time, supporting the estimated divergence time of 58.4 MYA (Figure 2a) and the absence of WGD private to *S. hispanica* in the *Salvia* lineage.

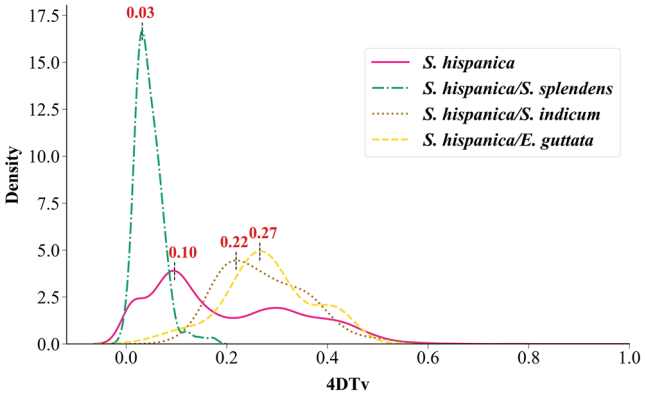

FIGURE 3 Distribution of transversion substitutions at fourfold degenerate sites (4DTv). Distribution of 4DTv for *S. hispanica* and pairwise 4DTv with *S. splendens*, *S. indicum*, and *E. guttata*. Peaks in pairwise 4DTv density indicate the relative time of divergence between species.

3.4 | Analysis of oil biosynthesis genes in *S. hispanica*

We investigated gene family expansion and particularly segmental duplication as a potential hypothesis for the high production of ω-3 FAs in *S. hispanica*. Key lipid synthesis genes including *SAD* and 3-ketoacyl-acyl carrier protein reductase (*KAR*) were significantly expanded in *S. hispanica* by one and four families, respectively ($p < 0.05$, Table S11). The increased number of *SAD* genes in *S. hispanica* cannot be explained by the WGD events: the *S. splendens* genome is twice larger than that of *S. hispanica*, having recently undergone a WGD event but does not contain twice the number of *SAD* genes. To understand the mechanisms underlying the expansion of specific gene families in *S. hispanica*, we investigated different modes of gene duplications (i.e., whole-genome, tandem, proximal, transposed, or dispersed duplications).

In *S. hispanica*, most of the *SAD* genes are located in the telomeric region of chromosome 1 (11 out of 13 genes), and the remaining two are located on chromosomes 3 and 4 (Figure 4). Gene duplication analysis (E-value $< 10^{-5}$) revealed that six of the *ShSAD* genes (*ShSAD2*, *ShSAD3*, *ShSAD4-a*, *ShSAD4-b*, *ShSAD5*, *ShSAD6*, and *ShSAD7*) form

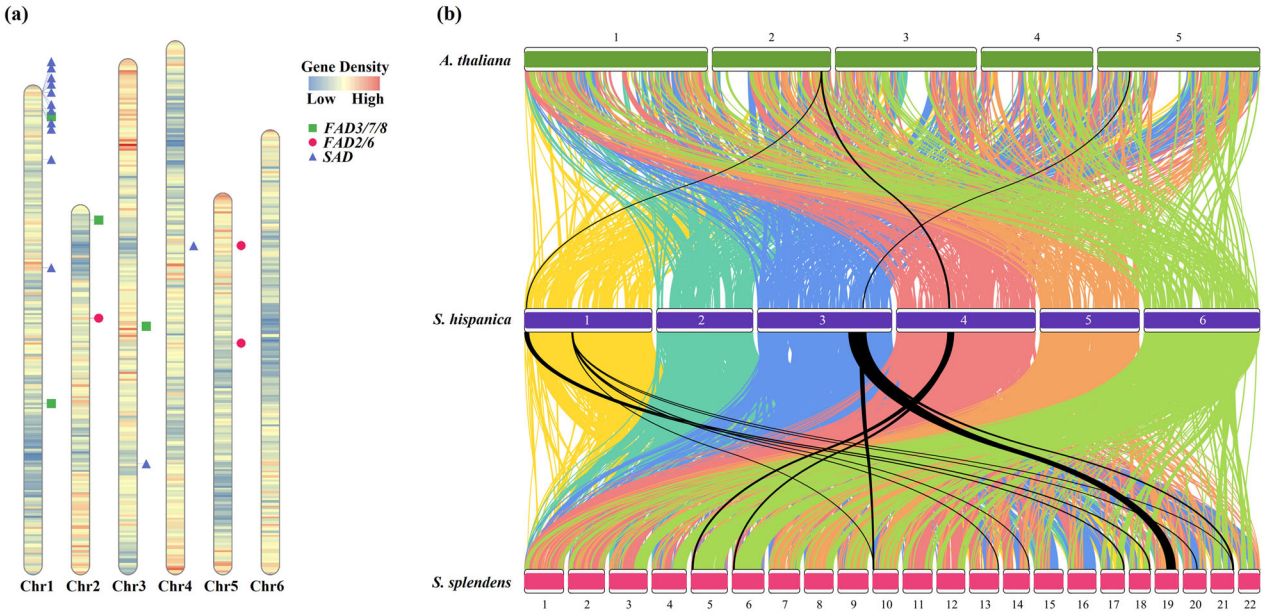

FIGURE 4 Chromosome ideogram and synteny analysis of *S. hispanica* genome. (a) Ideogram showing the gene density distribution and position of key fatty acid (FA) synthesis genes on *S. hispanica* chromosomes. The tandem array of *ShSAD* genes is located in the telomeric region of chromosome 1. (b) Synteny analysis of *S. hispanica* with *A. thaliana* and *S. splendens* using synteny blocks from DupGen_finder. Colors represent *S. hispanica* chromosomes. Black cords represent only synteny blocks containing *ShSAD* genes.

a tandem array located in the telomeric region of chromosome 1 (Figure 4a). Interestingly, the genes in this tandem array (excluding *ShSAD7*) were specific to *S. hispanica*, absent in other species studied. One gene upstream of this tandem array, the *ShSAD1* gene, was unique to *S. hispanica*, resulting from a duplication of *ShSAD13* located on chromosome 4.

ShSAD13 belongs to an orthogroup shared across species that includes four genes from *S. hispanica*. This orthogroup included *ShSAD13* and *ShSAD7* and *ShSAD11-a*, *ShSAD11-b*, and *ShSAD12*, which are dispersed duplicates located on chromosomes 1 and 3, respectively. The remaining three *ShSAD* genes formed an orthogroup unique to *S. hispanica* with *ShSAD8* and *ShSAD9* which are proximal duplicates (i.e., one gene apart), and *ShSAD10* which is a transposed duplicate of *ShSAD1*. This is different from the finding by Xue et al. (2023) who showed that all *ShSAD* genes are tandem duplicates. Instead, our analysis suggested that the *ShSAD* genes have been repeatedly duplicated in *S. hispanica* after its divergence from *S. splendens* (Figure 4b). The 11 *ShKAR* genes (Table S11) are spread across chromosomes 2–6, six of which directly resulted from WGD, including *ShKAR1*, *ShKAR2*, *ShKAR3*, *ShKAR5*, *ShKAR9*, and *ShKAR10*. *ShKAR5* is a tandem duplicate of *ShKAR6* which is one gene away from the pair of tandem duplicates formed by *ShKAR7* and *ShKAR8*. In addition, *ShKAR4* and *ShKAR11* are transposed duplicate and proximal duplicate pairs of *ShKAR10*, respectively.

We compared the expression of *ShSAD* and *ShKAR* genes, particularly those on the tandem array, with that of *ShFAD* genes in various *S. hispanica* tissues (Figure S11). This helped us determine their potential co-expression using transcriptomic data from a previous RNAseq study that included different developmental stages (Gupta et al., 2021). From the tandem array, *ShSAD2*, *ShSAD3*, and *ShSAD6* showed upregulation in reproductive tissues (i.e., the top half of the raceme, the bottom half of the raceme, and the flower). *ShSAD4* and *ShSAD7* from the tandem array were found to be co-expressed in various tissues, albeit at lower levels. These two genes showed more overactivity mainly in the vegetative tissues (i.e., cotyledon, shoot, day 12 leaf, and day 69 leaf), whereas *ShSAD2*, *ShSAD3*, and *ShSAD6* are co-expressed at a lower level. In reproductive tissues, *ShSAD11* and *ShSAD12* exhibited high expression levels, whereas *ShSAD1* and *ShSAD13* had higher expression levels in vegetative tissues despite notable co-expression in reproductive tissues. *ShSAD5* from the tandem array, as well as *ShSAD8*, *ShSAD9*, and *ShSAD10* did not show any expression in Gupta et al. (2021) dataset. Further investigation using other studies is required to confirm if these genes are, in fact, not expressed in *S. hispanica* tissues. We found that a total of nine *ShSAD* genes, including five from the tandem array on chromosome 1, were co-expressed in various tissues. Among these genes, five showed significantly high differential expression in reproductive tissues. Therefore, we speculate that *ShSAD* genes play a crucial role in FAs biosynthesis during the later stages of development, leading to the accumulation of high levels of ALA in *S. hispanica* seeds.

Of the 11 *ShKAR* genes, eight were found to be co-expressed in various tissues. Out of these eight genes,

ShKAR1, *ShKAR2*, *ShKAR3*, *ShKAR4*, *ShKAR5*, *ShKAR6*, and *ShKAR9* showed high expression levels in reproductive tissues. This suggests that similar to *ShSAD* genes, *ShKAR* genes also play a crucial role in high FAs biosynthesis in *S. hispanica*. *ShKAR11*, on the other hand, showed uniform expression across tissues, with higher expression in seeds. In contrast to *ShSAD* genes, most *ShKAR* genes also showed high expression in vegetative tissues. The *ShKAR* and *ShSAD* genes show different patterns of expression compared to the *ShFAD* genes. *ShFAD* genes are mainly overactive in vegetative tissues, but they are also co-expressed in reproductive tissues. However, their activity is reduced in reproductive tissues, especially for *ShFAD2a/b* and *ShFAD3a/b*, which suggests that they are not the primary factor driving high levels of ω-3 biosynthesis in *S. hispanica*. Instead, their contribution is augmented by the co-expression and overactivity of *ShSAD* and *ShKAR* genes, which provide necessary FA precursors for the biosynthesis of FAs.

Three characteristics were unique to the *SAD* gene family compared to the *KAR* gene family. First, the *S. hispanica* genome contained the highest number of *SAD* gene orthologs (13 copies) of all species studied here. However, the number of *KAR* family genes in *S. hispanica* was similar to that of other species. Second, *SAD* genes had the highest number of paralogs (10 genes including a private orthogroup of three genes) unique to *S. hispanica*, while *S. hispanica KAR* genes only showed one unique paralog. Third, *ShSAD* genes included a six-gene tandem array including five genes unique to *S. hispanica*, while Sh*KAR* genes included two tandem pairs with only one gene unique to *S. hispanica*.

We evaluated the evolutionary constraint on the *ShSAD* genes by analyzing the Ka/Ks ratio in orthogroups containing at least two genes. Comparison of pairwise Ka/Ks ratios between CDSs of *ShSAD*s unique to *S. hispanica* and orthologs from other species revealed a set of *ShSAD* genes with functional characteristics unique to *S. hispanica*. The orthogroup containing *ShSAD8*, *ShSAD9*, and *ShSAD10* showed an excess of non-synonymous substitution, with 14%–16% of the alignment length displaying a Ka/Ks ratio greater than 1 (Figure S12; Fisher's exact test: $p < 0.05$). In contrast, Ka/Ks ratios were less than 1 for 22%–27% of the alignments (Fisher's exact test: $p < 0.05$), indicative of purifying selection at other sites. However, most of the alignment lengths showed no substitutions, either being fully conserved (*ShSAD8-ShSAD9*) or not present across all species investigated (*ShSAD8-ShSAD10* and *ShSAD9-ShSAD10*). For the orthogroup containing *ShSAD7*, *ShSAD11-a/b*, *ShSAD12*, and *ShSAD13*, 68%–90% of the alignments length showed Ka/Ks ratios lesser than 1 (Fisher's exact test: p-value = 0.05) and only 0%–6% of Ka/Ks ratios greater than 1 (Fisher's exact test: $p < 0.05$), suggesting a strong purifying selection (Figures S13, S14 & S15).

4 | DISCUSSION

4.1 | Gene family evolution in *S. hispanica*

WGD is common in angiosperms, allowing the neofunctionalization of duplicated genes and the potential adaptation to novel conditions (Hahn et al., 2005; Hughes et al., 2014; S.-F. Li et al., 2020). In addition to the WGD event previously reported for *S. hispanica* (L. Li et al., 2023; L. Wang et al., 2022) shared with *S. splendens*, we identified an ancestral γ-WGD event shared with other species. Seeds of *S. splendens*, having undergone these two WDGs, contain 34.5% ALA (Joh et al., 1988) compared with 62% in *S. hispanica* seeds (Oteri et al., 2023). Therefore, the higher ω-3 production in *S. hispanica* compared to the other species studied here cannot be due to one of its WGD events.

WGD events in autopolyploids do not alter the dosage balance across molecular pathways, including protein modification and transcriptional regulation (Chang et al., 2022). The consistent gene expression after WGDs is due to dosage sharing and tight regulatory control mechanisms; in contrast, tandem duplications lead to the shuffling of regulatory elements (Rogers et al., 2017). Defoort et al. (2019) found that tandem duplications in *A. thaliana*, *S. lycopersicum*, and *Z. mays* can impact dosage balance in protein–protein interactions (Defoort et al., 2019). This may explain how tandem duplications of *ShSAD* genes increase FA synthesis in *S. hispanica* seeds. The effect of duplicated genes on dosage balance is more profound when genes encode for a limiting step of a metabolic pathway (Defoort et al., 2019), which is the case for *ShSAD* genes in FA biosynthesis. Further investigations are required to confirm that *S. hispanica* has only been subject to WGD leading to autopolyploidization.

The genome of *S. hispanica* showed the highest number of contracted gene families (5919), while *S. splendens* showed the highest number of expanded gene families (7936). The expansion or contraction of gene families has been associated with gene regulation (Baroncelli et al., 2016; Najafpour et al., 2020). Biological pathways are often regulated at the gene network level through regulatory hubs with key regulatory gene families expanded (Yu et al., 2017). In contrast, genes performing independent functions under purifying selection often show gene family contraction (Hess et al., 2018). Consequently, non-synonymous mutations cause immediate loss of function in single-copy genes but are neutral in expanded gene families due to functional redundancy (Force et al., 1999; Hahn et al., 2005).

The *S. hispanica* genome with predominantly contracted gene families is under purifying selection, and the selective removal of deleterious alleles potentially explains the genomic stability of key biological functions (Bray & West, 2005; Cvijović et al., 2018; dos Santos Maraschin et al.,

2019; Hough et al., 2013). Intense purifying selection has been observed across plant species. In *Zea mays* (maize), highly expressed genes experience stronger purifying selection and regulatory neofunctionalization leading to unique and independent functions that have increased photosynthetic efficiency and stress tolerance (Hughes et al., 2014). Similarly, paralogs with deleterious effects might have been removed from the *S. hispanica* genome by purifying selection and removed through the structural rearrangement of the gene families involved in oil biosynthesis for functional optimization. The adaptive gene arrangements and removal of deleterious paralogs might have reduced vulnerability to environmental pressures and increased reproductive success due to increased PUFA production. Oil seeds with a higher percentage of unsaturated FAs germinate earlier and have a higher growth rate in lower temperatures (Gao et al., 2020; Linder, 2000) due to the lower melting points of unsaturated FAs. Consequently, the active interaction between gene duplication, natural selection, and the functional optimization induced by cold environmental stress might have resulted in the *S. hispanica* genome coordinating high PUFA production in seeds.

4.2 | Tandem duplication as an essential evolutionary genomic mechanism

Tandem duplication, one of the main mechanisms of gene family expansion (Achaz et al., 2000; Lan & Pritchard, 2016), also supports phenotypic plasticity (Chang et al., 2022) by mediating the adaptive response of the secondary metabolism to environmental stress (Defoort et al., 2019). Tandem duplications are lineage-specific and often affect membrane proteins and biotic and abiotic response genes (Cai et al., 2023; Carretero-Paulet & Fares, 2012; Denoeud et al., 2014; Fischer et al., 2014; Hanada et al., 2008; Jiang et al., 2013; Kondrashov, 2012; Picart-Picolo et al., 2020; Rizzon et al., 2006). In the Lamiaceae family, species-specific tandem duplicates are responsible for the biosynthesis of flavonoids in *Scutellaria baicalensis* (Xu, Gao, et al., 2020), terpenoids in the *Lavandula angustifolia* (lavender) (J. Li et al., 2021), and diterpenoids in *Isodon rubescens* (Sun et al., 2023).

The tandem array of six *ShSAD* genes located in the telomeric region of chromosome 1 suggests a shared regulation. Tandem duplicates are known to be co-regulated with the sub-functionalization of expression at higher levels compared with segmental duplicates or WGD genes (Cannon et al., 2004; Casneuf et al., 2006). For example, the tissue-specific co-expression of unique tandem duplications involved in the biosynthesis of flavonoids was observed in *Carthamus tinctorius* (safflower; Wu et al., 2021). Similarly, seed-specific tandem duplicated gene pairs responsible for oil biosynthesis in *S. indicum* were shown to be co-expressed (Song et al., 2021).

Tandem duplications of lipid biosynthesis genes with effects consistent with those found here in *S. hispanica* have been evidenced across different species. In *Cajanus cajan* (pigeon pea), tandem duplicates control the biosynthesis of ALA (C. Liu et al., 2021). In *S. indicum*, a combination of lipid transfer gene family expansion due to tandem duplication and contraction of lipid degradation genes was identified as driving the high accumulation of FAs in seeds (L. Wang et al., 2014). The highly conserved domains in segmental and tandem duplicated wax ester synthase (*WSD1*) and *DGAT* genes in *Gossypium hirsutum* (cotton) were related to a rate-limiting process during high unsaturated FAs accumulation in seeds (Zhao et al., 2021). The expansion of *GmFAD2* genes in *Glycine max* (soybean) (Lakhssassi et al., 2021) and *OeB3* genes in *Olea europaea* (olive) (Qu et al., 2023) was associated with tandem duplications.

4.3 | *SAD* gene family expansion is an adaptive multi-stress response mechanism affecting FA biosynthesis

Contrary to most gene families involved in lipid synthesis in *S. hispanica*, the *ShSAD* and *ShKAR* gene families are substantially expanded. This extends previous findings from transcriptomic analysis which showed that the *SAD* gene family was expanded (L. Wang et al., 2022). In plants, the *SAD* and *KAR* gene families are critical for FA biosynthesis and the initiation of FA desaturation in the chloroplast (González-Thuillier et al., 2021; Li-Beisson et al., 2013). *SAD* genes encode the only known soluble desaturase in chloroplast stroma, which is essential for ALA biosynthesis. The SAD enzyme also controls the synthesis of ACP-bound oleic acid (18:1) from stearate, resulting in the first double bond at the α-end (You et al., 2014).

In *Camellia chekiangoleosa* seeds, the expansion of the *SAD* gene family leads to the high production of unsaturated FAs, which are thought to be adaptive (Shen et al., 2022). Similarly, in *S. hispanica* seeds with high FA content, a high number of unique *SAD* genes sit in a tandem array. *SAD* genes sitting in tandem have also been reported in *Linum usitatissimum* L. (flax seeds; You et al., 2014) and *Olea europaea* var. *sylvestris* (wild olive), where the expansion of duplicated *SAD* genes has allowed neofunctionalization to support the high production of OA (Unver et al., 2017).

The *S. hispanica* genome contains an increased number of tandem-duplicated *SAD* genes that are upregulated in reproductive tissues, leading to abundant production of PUFAs. This may represent an adaptive response to environmental stress, as seen in other plants (J. Chen et al., 2023; Feng et al., 2017; Zhao et al., 2021). The *ShSAD2* and *ShSAD7* genes in *S. hispanica* are overexpressed in response to cold stress (Xue et al., 2023). Differential substrate specificity of tandem

duplicates is believed to be the mechanism behind this adaptive response to stress, also leading to enhanced secondary metabolite synthesis (Chang et al., 2022; J. Li et al., 2021; Picart-Picolo et al., 2020; Tohge & Fernie, 2020; Y. Wang et al., 2015; H. Xiao et al., 2020; Xu, Pu, et al., 2020).

The function of *ShSAD11a* in seed oil formation was confirmed through heterologous expression studies in yeast and *A. thaliana* transgenic lines (Xue et al., 2023). The higher number of *SAD* genes in *S. hispanica* (13 genes) compared to *Perilla frutescens* (seven genes) is currently the main hypothesis for the higher accumulation of ALA in *S. hispanica* seeds (Xue et al., 2023). However, we also specifically hypothesize that the six-gene tandem array of *ShSAD* genes in the telomeric region of chromosome 1, including five species-specific genes (four of which co-expressed in reproductive tissues) might further explain the high accumulation of ω-3 FAs in *S. hispanica* seeds. For example, the regulation of the very long-chain monounsaturated nervonic acid (C24:1ω-9) in *Acer truncatum* (purple blow maple) is controlled by a 10-gene tandem array of 3-ketoacyl CoA synthetase (*KCS*) genes, encoding a rate-limiting enzyme that defines substrate and tissue specificity during FA elongation and is highly expressed in mature seeds (Ma et al., 2020).

The *ShSAD* tandem array includes five paralogs unique to *S. hispanica*, which were duplicated after the divergence from *S. splendens*. Our hypothesis is supported by the co-expression of four tandem genes (*ShSAD2*, *ShSAD3*, *ShSAD4*, and *ShSAD6*) with five other *ShSAD* genes (*ShSAD1*, *ShSAD7*, *ShSAD11*, *ShSAD12*, and *ShSAD13*) in reproductive tissues, as shown in Figure S11. Overactivity and co-expression of *ShSAD* genes in these tissues produce an abundance of C18:1-ACP, which serves as a substrate for the desaturation of FAs in both the plastid and ER. As a result, *S. hispanica* seeds accumulate relatively high levels of ω-3 FAs. L. Li et al. (2023) proposed that the high expression of ER-localized *ShFAD3* is the main driver of the high accumulation of ω-3 FA in *S. hispanica* seeds. However, our findings contradict this hypothesis. In fact, the expression of *ShFAD3* genes is decreased in reproductive tissues, requiring the extra boost from the *ShSAD* genes for further high production of ω-3 FA in *S. hispanica* seeds. The WGD *ShSAD* genes (*ShSAD7*, *ShSAD11-a/b*, *ShSAD12*, and *ShSAD13*) have been under evolutionary constraints to maintain their function. On the other hand, *ShSAD* orthologs unique to *S. hispanica* (*ShSAD8*, *ShSAD9*, and *ShSAD10*) show diverging non-synonymous mutations and increased rate of substitution at specific sites (Kryazhimskiy & Plotkin, 2008) while showing evidence of purifying selection elsewhere.

5 | CONCLUSIONS

This study generated a high-quality, chromosomal-level reference genome for *S. hispanica* to analyze the evolution of oil biosynthesis genes in this valuable oil-seed crop. Our analysis supports the expansion of the *ShSAD* gene family through tandem duplications as a potential driver of high ω-3 FAs accumulation in *S. hispanica* seeds. Comparative analysis of multiple chromosomal-level genomes is a powerful approach to assess the putative effect of gene copy number variation and other sources of structural variation across genomes. This work establishes valuable genomic resources for *S. hispanica* and prompts the need to further investigate structural variants at FA biosynthesis gene loci within and among species. This will enable the breeding of emerging crops or the horizontal transfer of genes across species, with the possibility of altering gene dosage through the introgression of arrays of paralogous genes to improve key traits.

AUTHOR CONTRIBUTIONS

Tannaz Zare: Data curation; formal analysis; investigation; methodology; visualization; writing—original draft. **Jefferson Paril**: Formal analysis; methodology; software; visualization. **Emma Barnett**: Project administration; resources. **Parwinder Kaur**: Resources; supervision. **Rudi Appels**: Funding acquisition; methodology; resources; writing—review and editing. **Berit Ebert**: Conceptualization; supervision; writing—review and editing. **Ute Roessner**: Funding acquisition; project administration; resources; supervision; writing—review and editing. **Alexandre Fournier-Level**: Conceptualization; formal analysis; methodology; project administration; supervision; writing—review and editing.

ACKNOWLEDGMENTS

This research was supported by a Research Training Program Scholarship, the Alfred Nicholas Fellowship, the Megan Klemm Postgraduate Research Scholarship, and the Norma Hilda Schuster (nee Swift) Scholarship from the University of Melbourne awarded to Tannaz Zare. Berit Ebert was supported by the Inaugural Botany Foundation Fellowship 2020 from the University of Melbourne Botany Foundation.

Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The Whole-Genome Shotgun (WGS) project of *S. hispanica* is available at DDBJ/ENA/GenBank under the accession JALPBU000000000. The genome assembly and annotation of *S. hispanica* (NCBI *Salvia hispanica* Annotation Release 100) are available from the NCBI database under GenBank GCA_023119035.1 and RefSeq GCF_023119035.1 accessions.

ORCID

Parwinder Kaur https://orcid.org/0000-0003-0201-0766
Alexandre Fournier-Level https://orcid.org/0000-0002-6047-7164

REFERENCES

Achaz, G., Coissac, E., Viari, A., & Netter, P. (2000). Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae*: A possible model for their origin. *Molecular Biology and Evolution*, *17*(8), 1268–1275.

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Arredondo-Mendoza, G. I., Jiménez-Salas, Z., Garza, F. J. G.-d. l., Solís-Pérez, E., López-Cabanillas-Lomelí, M., González-Martínez, B. E., & Campos-Góngora, E. (2020). Ethanolic extract of *Salvia hispanica* L. regulates blood pressure by modulating the expression of genes involved in BP-regulatory pathways. *Molecules*, *25*(17), 3875.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., & Eppig, J. T. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29.

Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., Lelandais-Brière, C., Owens, G. L., Carrère, S., & Mayjonade, B. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature*, *546*(7656), 148–152.

Banaś, W., Sanchez Garcia, A., Banaś, A., & Stymne, S. (2013). Activities of acyl-CoA: Diacylglycerol acyltransferase (DGAT) and phospholipid: Diacylglycerol acyltransferase (PDAT) in microsomal preparations of developing sunflower and safflower seeds. *Planta*, *237*, 1627–1636.

Baroncelli, R., Amby, D. B., Zapparata, A., Sarrocco, S., Vannacci, G., Le Floch, G., Harrison, R. J., Holub, E., Sukno, S. A., & Sreenivasaprasad, S. (2016). Gene family expansions and contractions are associated with host range in plant pathogens of the genus Colletotrichum. *BMC Genomics*, *17*(1), 1–17.

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L. G., & Garmiri, P. (2020). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489.

Bates, P. D., Fatihi, A., Snapp, A. R., Carlsson, A. S., Browse, J., & Lu, C. (2012). Acyl editing and headgroup exchange are the major mechanisms that direct polyunsaturated fatty acid flux into triacylglycerols. *Plant Physiology*, *160*(3), 1530–1539.

Bates, P. D., Stymne, S., & Ohlrogge, J. (2013). Biochemical pathways in seed oil synthesis. *Current Opinion in Plant Biology*, *16*(3), 358–364.

Block, M. A., & Jouhet, J. (2015). Lipid trafficking at endoplasmic reticulum–chloroplast membrane contact sites. *Current Opinion in Cell Biology*, *35*, 21–29.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bray, C. M., & West, C. E. (2005). DNA repair mechanisms in plants: Crucial sensors and effectors for the maintenance of genome integrity. *New Phytologist*, *168*(3), 511–528.

Browse, J., & Somerville, C. (1991). Glycerolipid synthesis: Biochemistry and regulation. *Annual Review of Plant Biology*, *42*(1), 467–506.

Cahill, J. P. (2003). Ethnobotany of chia, *Salvia hispanica* L. (Lamiaceae). *Economic Botany*, *57*(4), 604–618.

Cai, Z., Zhao, X., Zhou, C., Fang, T., Liu, G., & Luo, J. (2023). Genome-wide mining of the tandem duplicated type iii polyketide synthases and their expression, structure analysis of *Senna tora*. *International Journal of Molecular Sciences*, *24*(5), 4837.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 1–9.

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., & May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biology*, *4*(1), 1–21.

Carretero-Paulet, L., & Fares, M. A. (2012). Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. *Molecular Biology and Evolution*, *29*(11), 3541–3551.

Casneuf, T., De Bodt, S., Raes, J., Maere, S., & Van de Peer, Y. (2006). Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biology*, *7*, 1–11.

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, *44*(19), e147–e147.

Chang, J., Marczuk-Rojas, J. P., Waterman, C., Garcia-Llanos, A., Chen, S., Ma, X., Hulse-Kemp, A., Van Deynze, A., Van de Peer, Y., & Carretero-Paulet, L. (2022). Chromosome-scale assembly of the *Moringa oleifera* Lam. genome uncovers polyploid history and evolution of secondary metabolism pathways through tandem duplication. *The Plant Genome*, *15*(3), e20238.

Chen, J., Gao, J., Zhang, L., & Zhang, L. (2023). Tung tree stearoyl-acyl carrier protein Δ9 desaturase improves oil content and cold resistance of *Arabidopsis* and *Saccharomyces cerevisiae*. *Frontiers in Plant Science*, *14*, 1144853. https://doi.org/10.3389/fpls.2023.1144853

Chen, Y., Lun, A. T., & Smyth, G. K. (2016). From reads to genes to pathways: Differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, *5*, 1438.

Creus, A., Benmelej, A., Villafañe, N., & Lombardo, Y. B. (2017). Dietary Salba (*Salvia hispanica* L) improves the altered metabolic fate of glucose and reduces increased collagen deposition in the heart of insulin-resistant rats. *Prostaglandins, Leukotrienes and Essential Fatty Acids*, *121*, 30–39.

Cvijović, I., Good, B. H., & Desai, M. M. (2018). The effect of strong purifying selection on genetic diversity. *Genetics*, *209*(4), 1235–1278.

da Silva, B. P., Anunciação, P. C., da Silva Matyelka, J. C., Della Lucia, C. M., Martino, H. S. D., & Pinheiro-Sant'Ana, H. M. (2017). Chemical composition of Brazilian chia seeds grown in different places. *Food Chemistry*, *221*, 1709–1716.

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: A computational tool for the study of gene family evolution. *Bioinformatics*, *22*(10), 1269–1271.

Defoort, J., Van de Peer, Y., & Carretero-Paulet, L. (2019). The evolution of gene duplicates in angiosperms and the impact of protein–protein

interactions and the mechanism of duplication. *Genome Biology and Evolution*, *11*(8), 2292–2305.

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F., & Aprea, G. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, *345*(6201), 1181–1184.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21.

dos Santos Maraschin, F., Kulcheski, F. R., Segatto, A. L. A., Trenz, T. S., Barrientos-Diaz, O., Margis-Pinheiro, M., Margis, R., & Turchetto-Zolet, A. C. (2019). Enzymes of glycerol-3-phosphate pathway in triacylglycerol synthesis in plants: Function, biotechnological application and evolution. *Progress in Lipid Research*, *73*, 46–64.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., & Aiden, A. P. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95.

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems*, *3*(1), 99–101.

El-Feky, A. M., Elbatanony, M. M., Aboul Naser, A. F., Younis, E. A., & Hamed, M. A. (2022). *Salvia hispanica* L. seeds extract alleviate encephalopathy in streptozotocin-induced diabetes in rats: Role of oxidative stress, neurotransmitters, DNA and histological indices. *Biomarkers*, *27*(5), 427–440.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*, 1–14.

Estilai, A., Hashemi, A., & Truman, K. (1990). Chromosome number and meiotic behavior of cultivated chia, *Salvia hispanica* (Lamiaceae). *HortScience*, *25*(12), 1646–1647.

Feng, J., Dong, Y., Liu, W., He, Q., Daud, M., Chen, J., & Zhu, S. (2017). Genome-wide identification of membrane-bound fatty acid desaturase genes in *Gossypium hirsutum* and their expressions during abiotic stress. *Scientific Reports*, *7*(1), 45711.

Fischer, I., Dainat, J., Ranwez, V., Glémin, S., Dufayard, J.-F., & Chantret, N. (2014). Impact of recurrent gene duplication on adaptation of plant genomes. *BMC Plant Biology*, *14*(1), 1–15.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*(4), 1531–1545.

Gao, L., Chen, W., Xu, X., Zhang, J., Singh, T. K., Liu, S., Zhang, D., Tian, L., White, A., Shrestha, P., Zhou, X.-R., Llewellyn, D., Green, A., Singh, S. P., & Liu, Q. (2020). Engineering trienoic fatty acids into cottonseed oil improves low-temperature seed germination, plant photosynthesis and cotton fiber quality. *Plant and Cell Physiology*, *61*(7), 1335–1347.

Georgiev, V., & Pavlov, A. (2017). Genetic engineering and manipulation of metabolite pathways in *Salvia* spp. In V. Georgiev & A. Pavlov (Eds.), *Salvia biotechnology* (pp. 399–414). Springer International Publishing. https://doi.org/10.1007/978-3-319-73900-7_10

González-Thuillier, I., Venegas-Calerón, M., Moreno-Pérez, A. J., Salas, J. J., Garcés, R., von Wettstein-Knowles, P., & Martínez-Force, E. (2021). Sunflower (*Helianthus annuus*) fatty acid synthase complex: B-Ketoacyl-[acyl carrier protein] reductase genes. *Plant Physiology and Biochemistry*, *166*, 689–699.

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898.

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

Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C., & Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, *15*(8), 1153–1160.

Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K., & Shiu, S.-H. (2008). Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiology*, *148*(2), 993–1003.

Hao, Z., Lv, D., Ge, Y., Shi, J., Weijers, D., Yu, G., & Chen, J. (2020). RIdeogram: Drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Computer Science*, *6*, e251.

Haque, M. S. (1980). Karyotypes and chromosome morphology in the genus *Salvia* Linn. *Cytologia*, *45*(4), 627–640.

Harley, R. M., Atkins, S., Budantsev, A. L., Cantino, P. D., Conn, B. J., Grayer, R., Harley, M. M., De Kok, R. d., Krestovskaja, T. d., & Morales, R. (2004). *Labiatae. Flowering plants · Dicotyledons: Lamiales (except Acanthaceae including Avicenniaceae)* (pp. 167–275). Springer Science & Business Media.

Hess, K., Oliverio, R., Nguyen, P., Le, D., Ellis, J., Kdeiss, B., Ord, S., Chalkia, D., & Nikolaidis, N. (2018). Concurrent action of purifying selection and gene conversion results in extreme conservation of the major stress-inducible Hsp70 genes in mammals. *Scientific Reports*, *8*(1), 1–16.

Hough, J., Williamson, R. J., & Wright, S. I. (2013). Patterns of selection in plant genomes. *Annual Review of Ecology, Evolution, and Systematics*, *44*, 31–49.

Hughes, T. E., Langdale, J. A., & Kelly, S. (2014). The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Research*, *24*(8), 1348–1355.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(03), 90–95.

Jamboonsri, W., Phillips, T. D., Geneve, R. L., Cahill, J. P., & Hildebrand, D. F. (2012). Extending the range of an ancient crop, *Salvia hispanica* L.—A new ω3 source. *Genetic Resources and Crop Evolution*, *59*, 171–178.

Jiang, S.-Y., González, J. M., & Ramachandran, S. (2013). Comparative genomic and transcriptomic analysis of tandemly and segmentally duplicated genes in rice. *PLoS One*, *8*(5), e63551.

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*(1), 13–23.

Kajitani, R., Yoshimura, D., Okuno, M., Minakuchi, Y., Kagoshima, H., Fujiyama, A., Kubokawa, K., Kohara, Y., Toyoda, A., & Itoh, T. (2019). Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nature Communications*, *10*(1), 1702.

Klein, A., Husselmann, L. H., Williams, A., Bell, L., Cooper, B., Ragar, B., & Tabb, D. L. (2021). Proteomic identification and meta-analysis in *Salvia hispanica* RNA-Seq de novo assemblies. *Plants*, *10*(4), 765.

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1749), 5048–5057.

Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. *PLOS Genetics*, *4*(12), e1000304.

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, *34*(7), 1812–1819.

Lakhssassi, N., Zhou, Z., Cullen, M. A., Badad, O., El Baze, A., Chetto, O., Embaby, M. G., Knizia, D., Liu, S., & Neves, L. G. (2021). TILLING-by-sequencing+ to decipher oil biosynthesis pathway in soybeans: A new and effective platform for high-throughput gene functional analysis. *International Journal of Molecular Sciences*, *22*(8), 4219.

Lan, X., & Pritchard, J. K. (2016). Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science*, *352*(6288), 1009–1013.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), 1–17.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

Li, J., Wang, Y., Dong, Y., Zhang, W., Wang, D., Bai, H., Li, K., Li, H., & Shi, L. (2021). The chromosome-based lavender genome provides new insights into Lamiaceae evolution and terpenoid biosynthesis. *Horticulture Research*, *8*, 53. https://doi.org/10.1038/s41438-021-00490-6

Li, L., Song, J., Zhang, M., Iqbal, S., Li, Y., Zhang, H., & Zhang, H. (2023). A near complete genome assembly of chia assists in identification of key fatty acid desaturases in developing seeds. *Frontiers in Plant Science*, *14*, 1102715.

Li, S.-F., Wang, J., Dong, R., Zhu, H.-W., Lan, L.-N., Zhang, Y.-L., Li, N., Deng, C.-L., & Gao, W.-J. (2020). Chromosome-level genome assembly, annotation and evolutionary analysis of the ornamental plant Asparagus setaceus. *Horticulture Research*, *7*, 48.

Liao, Y., Smyth, G. K., & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, *47*(8), e47–e47.

Li-Beisson, Y., Shorrosh, B., Beisson, F., Andersson, M. X., Arondel, V., Bates, P. D., Baud, S., Bird, D., DeBono, A., & Durrett, T. P. (2013). Acyl-lipid metabolism. *The Arabidopsis Book*, *11*, e0161. https://doi.org/10.1199/tab.0161

Lin, P., Wang, K., Wang, Y., Hu, Z., Yan, C., Huang, H., Ma, X., Cao, Y., Long, W., & Liu, W. (2022). The genome of oil-Camellia and population genomics analysis provide insights into seed oil domestication. *Genome Biology*, *23*, 1–21.

Linder, C. R. (2000). Adaptive evolution of seed oils in plants: accounting for the biogeographic distribution of saturated and unsaturated fatty acids in seed oils. *The American Naturalist*, *156*(4), 442–458.

Liu, C., Wu, Y., Liu, Y., Yang, L., Dong, R., Jiang, L., Liu, P., Liu, G., Wang, Z., & Luo, L. (2021). Genome-wide analysis of tandem duplicated genes and their contribution to stress resistance in pigeonpea (*Cajanus cajan*). *Genomics*, *113*(1), 728–735.

Liu, Y.-X., Yu, J.-H., Sun, J.-H., Ma, W.-Q., Wang, J.-J., & Sun, G.-J. (2023). Effects of omega-3 fatty acids supplementation on serum lipid profile and blood pressure in patients with metabolic syndrome: A systematic review and meta-analysis of randomized controlled trials. *Foods*, *12*(4), 725.

Lun, A. T., Chen, Y., & Smyth, G. K. (2016). It's DE-licious: A recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. In E. Mathé & S. Davis (Eds.), *Statistical genomics: Methods in molecular biology* (Vol. 1418, pp. 391–416). Humana Press. https://doi.org/10.1007/978-1-4939-3578-9_19

Ma, Q., Sun, T., Li, S., Wen, J., Zhu, L., Yin, T., Yan, K., Xu, X., Li, S., & Mao, J. (2020). The *Acer truncatum* genome provides insights into nervonic acid biosynthesis. *The Plant Journal*, *104*(3), 662–678.

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770.

Maynard, R. C., & Ruter, J. M. (2022). DNA content estimation in the genus *Salvia*. *Journal of the American Society for Horticultural Science*, *147*(3), 123–134.

Mendes, F. K., Vanderpool, D., Fulton, B., & Hahn, M. W. (2020). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics*, *36*(22-23), 5516–5518.

Meyer, B. J., & De Groot, R. H. (2017). Effects of omega-3 long chain polyunsaturated fatty acid supplementation on cardiovascular mortality: The importance of the dose of DHA. *Nutrients*, *9*(12), 1305.

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T., & Thomas, P. D. (2021). PANTHER version 16: A revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research*, *49*(D1), D394–D403.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534.

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, *41*(12), e121–e121.

Najafpour, B., Cardoso, J. C., Canário, A. V., & Power, D. M. (2020). Specific evolution and gene family expansion of complement 3 and regulatory factor H in fish. *Frontiers in Immunology*, *11*, 568631.

Ong, K. L., Marklund, M., Huang, L., Rye, K.-A., Hui, N., Pan, X.-F., Rebholz, C. M., Kim, H., Steffen, L. M., & van Westing, A. C. (2023). Association of omega 3 polyunsaturated fatty acids with incident chronic kidney disease: Pooled analysis of 19 cohorts. *BMJ*, *380*, e072909.

Onneken, P. (2018). *Salvia hispanica* L (chia seeds) as brain superfood: How seeds increase intelligence. *Journal of Nutrition & Food Sciences*, *8*(684), 2.

Oteri, M., Bartolomeo, G., Rigano, F., Aspromonte, J., Trovato, E., Purcaro, G., Dugo, P., Mondello, L., & Beccaria, M. (2023). Comprehensive chemical characterization of chia (*Salvia hispanica* L.) seed oil with a focus on minor lipid components. *Foods*, *12*(1), 23.

Paril, J., Pandey, G., Barnett, E., Rane, R. V., Court, L., Walsh, T., & Fournier-Level, A. (2022). Rounding up the annual ryegrass genome: High-quality reference genome of *Lolium rigidum*. *Frontiers in Genetics*, *13*, 1012694.

Paril, J., Zare, T., & Fournier-Level, A. (2023). compare_genomes: A comparative genomics workflow to streamline the analysis of evolutionary divergence across genomes. *Current Protocols*, *3*(8), e876.

Peláez, P., Orona-Tamayo, D., Montes-Hernández, S., Valverde, M. E., Paredes-López, O., & Cibrián-Jaramillo, A. (2019). Comparative transcriptome analysis of cultivated and wild seeds of *Salvia hispanica* (chia). *Scientific Reports*, *9*(1), 9761.

Penson, P. E., & Banach, M. (2020). The role of nutraceuticals in the optimization of lipid-lowering therapy in high-risk patients with dyslipidaemia. *Current Atherosclerosis Reports*, *22*, 1–9.

Picart-Picolo, A., Grob, S., Picault, N., Franek, M., Llauro, C., Halter, T., Maier, T. R., Jobet, E., Descombin, J., & Zhang, P. (2020). Large tandem duplications affect gene expression, 3D organization, and plant–pathogen response. *Genome Research*, *30*(11), 1583–1592.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(1), D61–D65. https://doi.org.10.1093/nar/gkl842

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., & Paterson, A. H. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biology*, *20*(1), 1–23.

Qu, J., Wang, B., Xu, Z., Feng, S., Tong, Z., Chen, T., Zhou, L., Peng, Z., & Ding, C. (2023). Genome-wide analysis of the molecular functions of B3 superfamily in oil biosynthesis in olive (*Olea europaea* L.). *BioMed Research International*, *2023*, 6051511. https://doi.org/10.1155/2023/6051511

Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. (2011). MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One*, *6*(9), e22594.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., & Lander, E. S. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47.

Rizzon, C., Ponger, L., & Gaut, B. S. (2006). Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. *PLoS Computational Biology*, *2*(9), e115.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Rogers, R. L., Shao, L., & Thornton, K. R. (2017). Tandem duplications lead to novel expression patterns through exon shuffling in *Drosophila yakuba*. *PLOS Genetics*, *13*(5), e1006795.

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, *17*(2), 155–158.

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., & Klimke, W. (2021). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *49*(D1), D10.

Shen, T.-f., Huang, B., Xu, M., Zhou, P.-y., Ni, Z.-x., Gong, C., Wen, Q., Cao, F.-l., & Xu, L.-A. (2022). The reference genome of *Camellia chekiangoleosa* provides insights into *Camellia* evolution and tea oil biosynthesis. *Horticulture Research*, *9*, uhab083. https://doi.org/10.1093/hr/uhab083

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212.

Song, S., You, J., Shi, L., Sheng, C., Zhou, W., Dossou, S. S. K., Dossa, K., Wang, L., & Zhang, X. (2021). Genome-wide analysis of nsLTP gene family and identification of SiLTPs contributing to high oil accumulation in sesame (*Sesamum indicum* L.). *International Journal of Molecular Sciences*, *22*(10), 5291.

Sreedhar, R. V., Kumari, P., Rupwate, S. D., Rajasekharan, R., & Srinivasan, M. (2015). Exploring triacylglycerol biosynthetic pathway in developing seeds of Chia (*Salvia hispanica* L.): A transcriptomic approach. *PLoS One*, *10*(4), e0123580.

Sun, Y., Shao, J., Liu, H., Wang, H., Wang, G., Li, J., Mao, Y., Chen, Z., Ma, K., & Xu, L. (2023). A chromosome-level genome assembly reveals that tandem-duplicated CYP706V oxidase genes control oridonin biosynthesis in the shoot apex of Isodon rubescens. *Molecular Plant*, *16*(3), 517–532.

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, *25*(1), 1934–3396. https://doi.org/10.1002/0471250953.bi0410s25

Timilsena, Y. P., Adhikari, R., Barrow, C. J., & Adhikari, B. (2016). Physicochemical and functional properties of protein isolate produced from Australian chia seeds. *Food Chemistry*, *212*, 648–656.

Tohge, T., & Fernie, A. R. (2020). Co-regulation of clustered and neofunctionalized genes in plant-specialized metabolism. *Plants*, *9*(5), 622.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–578.

Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., Yang, M., He, L., Deng, T., & Escalante, F. J. (2017). Genome of wild olive and the evolution of oil biosynthesis. *Proceedings of the National Academy of Sciences*, *114*(44), E9413–E9422.

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204.

Walker, J. B., Sytsma, K. J., Treutlein, J., & Wink, M. (2004). Salvia (Lamiaceae) is not monophyletic: Implications for the systematics, radiation, and ecological specializations of Salvia and tribe Mentheae. *American Journal of Botany*, *91*(7), 1115–1125.

Wallis, J. G., Watts, J. L., & Browse, J. (2002). Polyunsaturated fatty acid synthesis: What will they think of next? *Trends in Biochemical Sciences*, *27*(9), 467.

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*(1), 77–80.

Wang, J. R., Holt, J., McMillan, L., & Jones, C. D. (2018). FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics*, *19*, 1–11.

Wang, L., Lee, M., Sun, F., Song, Z., Yang, Z., & 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 Communications*, *3*(4), 100326.

Wang, L., Yu, S., Tong, C., Zhao, Y., Liu, Y., Song, C., Zhang, Y., Zhang, X., Wang, Y., & Hua, W. (2014). Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biology*, *15*(2), 1–13.

Wang, Y., Wang, Q., Zhao, Y., Han, G., & Zhu, S. (2015). Systematic analysis of maize class III peroxidase gene family reveals a conserved subfamily involved in abiotic stress response. *Gene*, *566*(1), 95–108.

Warren, R. L. (2016). RAILS and Cobbler: Scaffolding and automated finishing of draft genomes using long DNA sequences. *Journal of Open Source Software*, *1*(7), 116.

Wimberley, J., Cahill, J., & Atamian, H. S. (2020). De novo sequencing and analysis of Salvia hispanica tissue-specific transcriptome and identification of genes involved in terpenoid biosynthesis. *Plants*, *9*(3), 405.

Wu, Z., Liu, H., Zhan, W., Yu, Z., Qin, E., Liu, S., Yang, T., Xiang, N., Kudrna, D., & Chen, Y. (2021). The chromosome-scale reference genome of safflower (*Carthamus tinctorius*) provides insights into linoleic acid and flavonoid biosynthesis. *Plant Biotechnology Journal*, *19*(9), 1725–1742.

Xiao, H., Wang, C., Khan, N., Chen, M., Fu, W., Guan, L., & Leng, X. (2020). Genome-wide identification of the class III POD gene family and their expression profiling in grapevine (*Vitis vinifera* L). *BMC Genomics*, *21*(1), 1–13.

Xiao, Y., Zhang, Q., Liao, X., Elbelt, U., & Weylandt, K. H. (2022). The effects of omega-3 fatty acids in type 2 diabetes: A systematic review and meta-analysis. *Prostaglandins, Leukotrienes and Essential Fatty Acids*, *182*, 102456.

Xu, Z., Gao, R., Pu, X., Xu, R., Wang, J., Zheng, S., Zeng, Y., Chen, J., He, C., & Song, J. (2020). Comparative genome analysis of *Scutellaria baicalensis* and *Scutellaria barbata* reveals the evolution of active flavonoid biosynthesis. *Genomics, Proteomics & Bioinformatics*, *18*(3), 230–240.

Xu, Z., Pu, X., Gao, R., Demurtas, O. C., Fleck, S. J., Richter, M., He, C., Ji, A., Sun, W., & Kong, J. (2020). Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. *BMC Biology*, *18*(1), 1–14.

Xue, Y., Chen, B., Win, A. N., Fu, C., Lian, J., Liu, X., Wang, R., Zhang, X., & Chai, Y. (2018). Omega-3 fatty acid desaturase gene family from two ω-3 sources, *Salvia hispanica* and *Perilla frutescens*: Cloning, characterization and expression. *PLoS One*, *13*(1), e0191432.

Xue, Y., Wu, F., Chen, R., Wang, X., Inkabanga, A. T., Huang, L., Qin, S., Zhang, M., & Chai, Y. (2023). Genome-wide analysis of fatty acid desaturase genes in chia (*Salvia hispanica*) reveals their crucial roles in cold response and seed oil formation. *Plant Physiology and Biochemistry*, *199*, 107737.

Yoshinaga, Y., & Dalin, E. (2016). *Standard operating procedures: RNase A cleanup of DNA samples*. The Genome Portal of the Department of Energy Joint Genome Institute.

You, F. M., Li, P., Kumar, S., Ragupathy, R., Li, Z., Fu, Y.-B., & Cloutier, S. (2014). Genome-wide identification and characterization of the gene families controlling fatty acid biosynthesis in flax (*Linum usitatissimum* L). *Journal of Proteomics & Bioinformatics*, *7*(10), 310–326.

Yu, D., Lim, J., Wang, X., Liang, F., & Xiao, G. (2017). Enhanced construction of gene regulatory networks using hub gene information. *BMC Bioinformatics*, *18*(1), 1–20.

Zare, T., Rupasinghe, T. W., Boughton, B. A., & Roessner, U. (2019). The changes in the release level of polyunsaturated fatty acids (ω-3 and ω-6) and lipids in the untreated and water-soaked chia seed. *Food Research International*, *126*, 108665.

Zhang, X., Ritonja, J. A., Zhou, N., Chen, B. E., & Li, X. (2022). Omega-3 polyunsaturated fatty acids intake and blood pressure: A dose-response meta-analysis of randomized controlled trials. *Journal of the American Heart Association*, *11*(11), e025071.

Zhao, Y.-P., Wu, N., Li, W.-J., Shen, J.-L., Chen, C., Li, F.-G., & Hou, Y.-X. (2021). Evolution and characterization of acetyl coenzyme A: Diacylglycerol acyltransferase genes in cotton identify the roles of GhDGAT3D in oil biosynthesis and fatty acid composition. *Genes*, *12*(7), 1045.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zare, T., Paril, J. F., Barnett, E. M., Kaur, P., Appels, R., Ebert, B., Roessner, U., & Fournier-Level, A. (2024). Comparative genomics points to tandem duplications of *SAD* gene clusters as drivers of increased α-linolenic (ω-3) content in *S. hispanica* seeds. *The Plant Genome*, *17*, e20430. https://doi.org/10.1002/tpg2.20430