- A high-quality chromosome-level reference genome of *S. hispanica* was assembled and analysed.
- Ancestral whole-genome duplication events have not promoted the 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.

Comparative Genomics Points to Tandem Duplications of *SAD* Gene Clusters as Drivers of Increased ω-3 Content in *S. hispanica* Seeds

Tannaz Zare, Jeff F. Paril, Emma M. Barnett, Parwinder Kaur, Rudi Appels, Berit Ebert, Ute Roessner, and Alexandre Fournier-Level

Affiliations: Tannaz Zare, Jeff F. Paril, Emma M. Barnett and Alexandre Fournier-Level: School of BioSciences, The University of Melbourne, Parkville, 3010 VIC, Australia

Berit Ebert: School of Biology and Biotechnology, Ruhr-Universitat Bochum, 44801 Bochum, Germany

Ute Roessner: Research School of Biology, The Australian National University, Canberra, 2600 ACT, Australia

Parwinder Kaur: School of Agriculture and Environment, The University of Western Australia, Perth, 6009 WA, Australia

Rudi Appels: School of Agriculture, Food and Ecosystem Sciences, University of Melbourne, Parkville, 3010 VIC, Australia

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

Abbreviations: 4DTv, four-fold degenerative (synonymous) sites; AA, amino acid sequences; ACC, acyl-coenzyme A carboxylase; ACP, acyl carrier protein; ALA, α-linolenic acid; BUSCO, Benchmarking Universal Single-Copy Orthologs; CDSs, coding DNA sequences; CoA, acyl-coenzyme A; CRISPR, clustered regularly interspaced short palindromic repeats; DG, diacylglycerol; DGAT, acyl-coenzyme A: diacylglycerol acyltransferases; EFA, essential fatty acid; ER, endoplasmic reticulum; EST, expressed sequence tag; FA, fatty acid; FAD, fatty acid desaturases; FAS, fatty acid synthase; gDNA, genomic DNA; GO, Gene Ontology; KAR, 3-ketoacyl-acyl carrier protein reductase; KCS, 3-ketoacyl CoA synthetase; LA, linoleic acid; lncRNA, long non-coding RNA; LPCAT, acyl-coenzyme A:lysophosphatidylcholine acyltransferases; LTP1, type 1 lipid transfer; LTRs, Long terminal repeats; misc_RNA, miscellaneous RNA; MN, meganucleases; MRCA, most recent common ancestor; mRNA, messenger RNA; MYA, million years; NA, nervonic acid; NACRA, Northern Australia Crop Research Alliance; NCBI, National Center for Biotechnology Information; ncRNA, small nuclear RNA; OA, oleic acid; PC, phosphatidylcholine; PDAT, phospholipid:diacylglycerol acyltransferase; PUFA, polyunsaturated fatty acid; rRNA, ribosomal RNA; SA, stearic acid; SAD, stearoyl-ACP desaturase; SCOs, single-copy orthogroups; snoRNA, small nucleolar RNA; TALEN, transcription activator-like effector nucleases; TE, transposable element; TG, triglycerides; tRNA, transfer RNA; UniProt, Universal Protein Resource; WGD, whole genome duplication; WSD1, wax ester synthase 1; ZFN, zinc-finger nucleases

ABSTRACT

Salvia hispanica L. (chia) is an abundant source of ω-3 polyunsaturated fatty acids (PUFAs) that are highly beneficial to human health. The genomic basis for this accrued PUFA content in this emerging crop was investigated through the assembly and comparative analysis of a chromosome-level reference genome for *S. hispanica* (321.5 Mbp). The highly contiguous 321.5Mbp genome assembly, which covers 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 (ACP) 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 leveraging advanced genome editing techniques and will greatly support breeding efforts for improving ω-3 content in other oil crops.

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 7,100 species of flowering plants, *S. hispanica* belongs to the largest genus, *Salvia* spp. (sages) which regroups more than 1,000 species. *Salvia* spp. are increasingly recognised as commercially important crops due to their nutraceutical and bioactive compounds among which sterols, flavonoids, diterpenes, triterpenes and polyphenols (Cahill, 2003; Harley et al., 2004; Walker et al., 2004; Georgiev & Pavlov, 2017).

Amongst the *Salvia* spp., *S. hispanica* is the most nutritionally valuable crop due to its health-promoting properties, including high levels of dietary fibre (35%), carbohydrates (5%), protein (18-24%), lipids (31-34%), antioxidants and essential vitamins (Timilsena et al., 2016; da Silva et al., 2017; Zare et al., 2019). *S. hispanica* seeds are rich in essential fatty acids (EFAs) (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 (PUFAs) supplementation on human health suggest they may help reduce several chronic diseases such as diabetes, cardiovascular and inflammatory disorders, hypertension, dyslipidemia, and kidney dysfunction (Creus et al., 2017; Meyer & De Groot, 2017; Onneken, 2018; Arredondo-Mendoza et al., 2020; Penson & Banach, 2020; El-Feky et al., 2022; Xiao et al., 2022; Zhang et al., 2022; Liu et al., 2023; Ong et al., 2023).

The biosynthesis of PUFAs in plants involves a series of complex reactions in different subcellular compartments. The *de novo* biosynthesis of 16- or 18-carbon fatty acids (FAs) takes place in plastids through the action of acetyl-CoA carboxylase (ACC) and FA synthase (FAS) (Li-Beisson

et al., 2013). After the conversion/elongation of C16:0 to C18:0, the C18:0-ACP (SA, stearic acid) is desaturated to C18:1-ACP (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, 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 synthesised, 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 (Wang et al., 2014; Badouin et al., 2017; Unver et al., 2017; Lin et al., 2022; Shen et al., 2022). Expansion of the type 1 lipid transfer (*LTP1*) gene family and contraction of lipid degradation genes have been linked to the high oil accumulation in sesame seeds (Wang et al., 2014). Neo-functionalization and expansion of the *SAD* gene family is 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* had 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 ± 0.016 pg, genome size = ~ 460 Mb) (Haque, 1980; Estilai et al., 1990; 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 (Sreedhar et al., 2015; Peláez et al., 2019; Wimberley et al., 2020; Gupta et al., 2021). In addition, a set of studies functionally characterised genes encoding fatty acid desaturases (FADs) against different biotic/abiotic stresses (Xue et al., 2018) (Xue et al., 2023). These studies, together with

a genome assembly for *S. hispanica* (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 species. We investigated if specific biological functions overrepresented among significantly expanded gene families. In particular, our analysis seek to probe the hypothesis that duplication and nucleotides substitutions in oil biosynthesis genes support the high production of ω-3 FAs in *S. hispanica*.

2 MATERIALS AND METHODS

2.1 Plant material and genomic DNA extraction

A black-seed variety of *S. hispanica* L. was sourced from Chia Co. and Northern Australia Crop Research Alliance (NACRA; Kununurra, Western Australia). Fresh young leaves were harvested from a four-week-old individual *S. hispanica* plant, immediately frozen in liquid nitrogen and stored at -80 °C prior to the isolation of genomic DNA (gDNA). High molecular weight gDNA was isolated using a cetyltrimethylammonium bromide (CTAB) method (Murray & Thompson, 1980; Supplemental Figure S1). The isolated gDNA was treated with RNase A following the method developed by Yoshinaga & Dalin (Yoshinaga & Dalin, 2016); and purified using NucleoMag™ NGS magnetic beads (Macherey-Nagel, Düren, Germany) prior to DNA libraries synthesis.

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

For short-read sequencing, DNA libraries were synthesised from 3.9 µg of gDNA using the Illumina TruSeq DNA PCR-Free kit (Illumina, San Diego, CA, USA) and sequenced on Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA) in 2×150bp 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 4bp sliding windows and leading bases with a quality score below Q20 were removed using Trimmomatic v0.39 (Supplemental Table S1; Bolger et al., 2014). A total of 476Gb of high-quality Illumina reads with an average length of 145bp was retained for genome assembly.

For long-read sequencing, DNA libraries were prepared using the SQK-LSK109 ligation sequencing kit (Oxford Nanopore Technologies, Oxford, UK) and sequenced on a MinION Mk1B portable device with FLO-MIN106D flowcell. The long-read sequencing was run for 48hrs at 180mV 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_450bps_hac.jsn (Oxford Nanopore community, https://community.nanoporetech.com). Long reads were error-corrected with fmlrc2 v0.1.5 (Wang et al., 2018) resulting in 9Gb of high-quality resds with average length 2,825bp.

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; Supplemental Figure S2). The sequencing of the Hi-C libraries (~300bp insert size) was carried out on an Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA) in the 2×150bp mode by Genewiz (Suzhou, China).

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 size ranging from 15-mer to 21-mer were computed using Jellyfish v2.3.0 (Marçais & Kingsford, 2011) and the genome size, level of the heterozygosity, and abundance of genomic repeats were estimated using GenomeScope v1.0.0 (Vurture et al., 2017).

2.4 *De novo* genome assembly and scaffolding

A meta-assembly approach was conducted using a hybrid combination of long and short reads. The hybrid assembly consisted in 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 (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* (Sreedhar et al., 2015; Peláez et al., 2019; Wimberley et al., 2020; Gupta et al., 2021; Klein et al., 2021) 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, 2014; O'Leary et al., 2016). 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 visualised 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; Paril et al., 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 analysed 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). 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 *S. 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* (9.6 MYA) was retrieved from Wang et al. (2022). The rates of nucleotide substitution amongst pairs of paralogs/orthologs were measured using the third codon transversion rates at four-fold degenerative (synonymous) sites (4DTv) to estimate the likelihood of WGD events. The 4DTv values were calculated based on the alignment of each pair of 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 involved in lipid and FAs biosynthesis pathways retrieved from NCBI and UniProt (Supplemental 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 a 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: $1e^{-5}$, 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 (Wang et al., 2010) was used to calculate Ka/Ks ratio over 15bp sliding windows of the CDSs of paralogs associated with lipid metabolism in the *S. hispanica* genome and homologous genes in other species.

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 in 5,304 contigs spanning 1,556 scaffolds. The assembly covered ~321Mb (N50=53Mb; L50=3; largest scaffold=57Mb) with a GC content of 36% (Table 1). Hi-C reads analysis identified ~173 million contacts (Supplemental Table S3), 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 40Mb to 58Mb with spanned gaps of 491bp to 741bp (Supplemental Table S4); The L90=6 matched the chromosome number determined through flow cytometry (Figure 1 and Supplemental Figure S3; Maynard & 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 (Supplemental Figure S4). Analysis of k-mer frequencies estimated a haploid genome size of 466Mbp, consistent with the size of 460Mbp 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 (Supplemental Figure S5). The average BUSCO score was relatively high (> 95%) across all lineage sets. Around 94% of the previously published *S. hispanica* transcripts (Sreedhar et al., 2015; Peláez et al., 2019; Wimberley et al., 2020; Gupta et al., 2021; Klein et al., 2021) mapped to the assembled genome (Supplemental Figure S6). 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 mRNAs, misc_RNAs, and ncRNAs of class lncRNA. The repeat-masked assembly contained 39,616 genes, corresponding to 32,922 protein-coding genes, 4,071 non-coding genes and 2,623 pseudogenes (Supplemental Table S5). The total number of annotated transcripts (54,009) included 46,423 messenger RNAs (mRNAs), 3,758 long non-coding RNAs (lncRNAs), 739 transfer RNAs (tRNAs), 436 small nucleolar RNAs (snoRNAs), 233 small nuclear

281 RNAs (ncRNAs), 49 ribosomal RNAs (rRNAs), and 2,381 miscellaneous RNAs (misc_RNAs;
282 Table 2).

283 The genome of *S. hispanica* contained 44.14% of interspersed repeats, 0.80% and 0.19% of
284 which were simple and low complexity repeats, respectively (Supplemental Table S6 and
285 Supplemental Figure S7). Long terminal repeats (LTRs) represented 11.05% of the genome, with
286 Copia (5.60%) and Gypsy (5.45%) being the most abundant (Figure 1h), when DNA transposons
287 only represented 3.69% of the genome (Supplemental Table S6 and Supplemental Figure S7).

288 **3.2 Gene family evolution in the *S. hispanica* genome**

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

298 From the 320,180 genes found across all seven species, 305,234 genes (95.3%) were assigned
299 to one or more of the 27,963 orthogroups. In *S. hispanica*, a total of 45,108 genes were assigned
300 to orthogroups, 134 being SCOs, and 2,814 unique paralogs form 682 orthogroups, leaving 1,400
301 unassigned genes (Figure 2b). *S. splendens* contained the highest average number of paralogs
302 within orthogroup (1.85), showing the highest genetic redundancy, followed by *A. thaliana* (1.12),

303 *S. hispanica* (1.08), *V. vinifera* (0.96), *S. lycopersicum* (0.88), *S. indicum* (0.83), and *E. guttata*
304 (0.74). The highest number of unique orthogroups was observed for *A. thaliana* (3047; 13074
305 paralogs), followed by *S. splendens* (1471; 6,236 paralogs), *V. vinifera* (1257; 6,892 paralogs), *S.*
306 *lycopersicum* (972; 4,601 paralogs), *S. hispanica* (682; 2814 paralogs), *E. guttata* (617; 3,738
307 paralogs), and *S. indicum* (402; 1,790 paralogs; Figure 2b). The number of genes not assigned to
308 any orthogroup varied across species, with the highest value for *S. splendens* (5,081) followed by
309 *A. thaliana* (3,503), *S. lycopersicum* (1,489), *V. vinifera* (1,452), *S. hispanica* (1400), *E. guttata*
310 (1,216), and *S. indicum* (805; Figure 2b).

311 In total, 10,827 orthogroups were shared among all species, with 682 orthogroups being unique
312 to *S. hispanica* (Figure 2c). The two closely related *Salvia* species (i.e., *S. splendens* and *S.*
313 *hispanica*) contained the largest number of orthogroups (18,974 and 17,986, respectively), which
314 is 15-22% higher than that observed for other species. The high number of unique orthogroups in
315 *A. thaliana* (3,047) reflected the distant evolutionary relationships with the other species analysed
316 and its relevance as an outgroup (Figure 2c).

317 We next explored gene family expansion and contraction in *S. hispanica* compared to other
318 selected species. *A. thaliana*, *S. indicum*, *S. lycopersicum* and *V. vinifera* showed a relatively even
319 number of expanded *vs*. contracted gene families (Figure 2a). *E. guttata* and *S. hispanica*, on the
320 other hand, show a much higher number of contracted gene families, while *S. splendens* was the
321 only species with significant number of expanded gene families (7936). Interestingly, closely
322 related *S. hispanica* exhibited the opposite pattern with a significant excess of contracted gene
323 families (5,919).

324 Among the gene families expanded in *S. hispanica.,* a significant enrichment was found for
325 maintenance of plant homeostasis, response to stress and activation of defence mechanisms

(Supplemental Table S7). The top 10 gene families most unique to *S. hispanica* were highly enriched for specific biological processes: xenobiotic detoxification by transmembrane export in plasma membrane (GO:1990961; $P<5.70E^{-10}$), xenobiotic export from 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; $P<2.14E^{-04}$; Supplemental Figure S8). The 20 most enriched molecular function and cellular component ontologies in the *S. hispanica* genome are presented in Supplemental Table S8 and S9, respectively.

3.3 Whole genome duplications and speciation events

The occurrence of WGD events in species studied was determined based on the distribution of the 4DTv among multi-copy paralogs (Supplemental Figure S9). The 4DTv distribution for *S. hispanica* (Supplemental Figure S9h) showed a high density at 0.1 (relative time to the most recent common ancestor) and at 0.3. The first peak corresponded to a relatively recent WGD event in *S. hispanica* that is only shared with closely related species *S. splendens*. However, this peak in the 4DTv distribution of *S. splendens* is masked by a very recent WGD event at 0.03 (Supplemental Figure S9h). 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* (Supplemental Figure S9h).

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.

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*. 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 6 of the *ShSAD* genes (*ShSAD2*: XP_047955596.1; *ShSAD3*: XP_047970931.1; *ShSAD4-a* isoform X1: XP_047970902.1; *ShSAD4-b* isoform X2: XP_047970909.1; *ShSAD5*: XP_047970920.1; *ShSAD6*:

XP_047955488.1; and *ShSAD7*: XP_047954777.1) form 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 (XP_047969270.1) was unique to *S. hispanica,* resulting from a duplication of *ShSAD13* (XP_047982897.1) located on chromosome 4.

ShSAD13 belongs to an orthogroup shared across species that included 4 genes from *S. hispanica*. This orthogroup included *ShSAD13*, *ShSAD7*, and *ShSAD11-a* isoform X1 (XP_047955195.1), *ShSAD11-b* isoform X2 (XP_047955196.1) and *ShSAD12* (XP_047971758.1), which are dispersed duplicates located on chromosome 1 and 3, respectively. The remaining three *ShSAD* genes formed an orthogroup unique to *S. hispanica* with *ShSAD8* (XP_047966762.1) and *ShSAD9* (XP_047955361.1) which are proximal duplicates (ie. one gene apart) and *ShSAD10* (XP_047938734.1) which is a transposed duplicate of *ShSAD1*. This is different to the finding by Xue and colleagues who showed that all *ShSAD* genes are tandem duplicates (Xue et al., 2023). Instead, our analysis suggested that the *ShSAD* genes have been repeatedly duplicated in *S. hispanica,* after its divergence from *S. splendens* (Figure 4b).

The eleven *ShKAR* genes are spread across chromosomes 2 to 6, six of which directly resulted from WGD, including *ShKAR1* (XP_047958713.1), *ShKAR2* (XP_047960775.1), *ShKAR3* (XP_047964327.1), *ShKAR5* (XP_047940213.1), *ShKAR9* (XP_047939697.1), and *ShKAR10* (XP_047952057.1). *ShKAR5* is also a tandem duplicate of *ShKAR6* (XP_047944250.1) which is one gene away from the pair of tandem duplicates formed by *ShKAR7* (XP_047940355.1) and *ShKAR8* (XP_047944629.1). In addition, *ShKAR4* (XP_047937601.1) and *ShKAR11* (XP_047945899.1) are transposed duplicate and proximal duplicate pairs of *ShKAR10*, respectively.

392 Three characteristics were unique to the *SAD* gene family compared to the *KAR* gene family.
393 First, the *S. hispanica* genome contained the highest number of *SAD* gene orthologs (13 copies) of
394 all species studied here. However, the number of *KAR* family genes in *S. hispanica* was similar to
395 that of other species. Second, *SAD* genes had the highest number of paralogs (10 genes including
396 a private orthogroup of 3 genes) unique to *S. hispanica*, while *S. hispanica KAR* genes only showed
397 one unique paralog. Third, *ShSAD* genes included a six-gene tandem array including 5 genes
398 unique to *S. hispanica*, while Sh*KAR* genes included two tandem pairs with only one gene unique
399 to *S. hispanica*.

400 We evaluated the evolutionary constraint on the *ShSAD* genes by analysing the Ka/Ks ratio in
401 orthogroups containing at least two genes. Comparison of pairwise Ka/Ks ratios between CDSs of
402 *ShSAD*s unique to *S. hispanica* and orthologs from other species revealed a set of *ShSAD* genes
403 with functional characteristics unique to *S. hispanica*. The orthogroup containing *ShSAD8*,
404 *ShSAD9*, and *ShSAD10* showed an excess of non-synonymous substitution, with 14 to 16% of the
405 alignment length displaying a Ka/Ks ratio greater than 1 (Supplemental Figure 10; Fisher's exact
406 test: $P<0.05$). In contrast, Ka/Ks ratios were less than 1 for 22 to 27% of the alignments (Fisher's
407 exact test: $P<0.05$), indicative of purifying selection at other sites. However, most of the
408 alignments length showed no substitutions, either being fully conserved (*ShSAD8-ShSAD9*) or not
409 present across all species investigated (*ShSAD8-ShSAD10* and *ShSAD9-ShSAD10*). For the
410 orthogroup containing *ShSAD7*, *ShSAD11-a/b*, *ShSAD12*, and *ShSAD13*, 68 to 90% of the
411 alignments length showed Ka/Ks ratios lesser than 1 (Fisher's exact test: p-value=0.05) and only
412 0-6% of Ka/Ks ratios greater than 1 (Fisher's exact test: $P<0.05$), suggesting a strong purifying
413 selection (Supplemental Figures 11, 12 & 13).

414

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; Li et al., 2020). In addition to the WGD event previously reported for *S. hispanica* (Wang et al., 2022; Li et al., 2023) shared with *S. splendens*, we identified an ancestral γ-WGD event shared with other species. *Salvia* species having undergone these two WDGs (e.g., *S. splendens*) neither show an increased expression of lipid genes nor a high oil accumulation in seeds. 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.

Whole genome duplication events 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). Tandem duplications in *A. thaliana*, *S. lycopersicum*, and *Z. mays* were shown to impact dosage balance in protein-protein interactions and could explain that *ShSAD* genes tandem duplications increased 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.

The genome of *S. hispanica* showed the highest number of contracted gene families (5,919), when *S. splendens* showed the highest number of expanded gene families (7,936). The expansion or contraction of gene families has been associated with gene regulation (Baroncelli et al., 2016;

437 Najafpour et al., 2020). Biological pathways are often regulated at the gene network level through
438 regulatory hubs with key regulatory gene families expanded (Yu et al., 2017). In contrast, genes
439 performing independent functions under purifying selection often show gene family contraction
440 (Hess et al., 2018). Consequently, non-synonymous mutations cause immediate loss of function in
441 single-copy genes but are neutral in expanded gene families due to functional redundancy (Force
442 et al., 1999; Hahn et al., 2005).

443 The *S. hispanica* genome with predominantly contracted gene families is under purifying
444 selection, and the selective removal of deleterious alleles potentially explains the genomic stability
445 of key biological functions (Bray & West, 2005; Hough et al., 2013; Cvijović et al., 2018; dos
446 Santos Maraschin et al., 2019). Intense purifying selection has been observed across plant species.
447 In *Zea mays* (maize), highly expressed genes experience stronger purifying selection and
448 regulatory neofunctionalization leading to unique and independent functions that have increased
449 photosynthetic efficiency and stress tolerance (Hughes et al., 2014). Similarly, paralogs with
450 deleterious effects might have been removed from the *S. hispanica* genome, and a structural
451 reorganisation could have occurred within gene families involved in oil biosynthesis.

452 **4.2 Tandem duplication as an essential evolutionary genomic mechanism**

453 Tandem duplication, one of the main mechanisms of the gene family expansion (Achaz et al.,
454 2000; Lan & Pritchard, 2016), is also supporting phenotypic plasticity (Chang et al., 2022) by
455 mediating the adaptive response of the secondary metabolism to environmental stress (Defoort et
456 al., 2019). Tandem duplications are lineage-specific and often affect membrane proteins and biotic
457 and abiotic response genes (Rizzon et al., 2006; Hanada et al., 2008; Carretero-Paulet & Fares,
458 2012; Kondrashov, 2012; Jiang et al., 2013; Denoeud et al., 2014; Fischer et al., 2014; Picart-

Picolo et al., 2020; Cai et al., 2023). In the Lamiaceae family, species-specific tandem duplicates are responsible for the biosynthesis of flavonoids in *Scutellaria baicalensis* (Xu et al., 2020a), terpenoids in the *Lavandula angustifolia* (lavender) (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 sub-functionalization of expression at higher levels compared with segmental duplicates or WGD genes (Cannon et al., 2004; Casneuf et al., 2006). For example, 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 (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 (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) were 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 only showed that the *SAD* gene family was expanded (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 (Li-Beisson et al., 2013; González-Thuillier et al., 2021). *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 is 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 increased number of tandem-duplicated *SAD* genes in the *S. hispanica* genome, leading to the abundant production of PUFAs, might represent an adaptive response to environmental stress as shown in other plants (Feng et al., 2017; Zhao et al., 2021; Chen et al., 2023). 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 (Wang et al.,

2015; Picart-Picolo et al., 2020; Tohge & Fernie, 2020; Xiao et al., 2020; Xu et al., 2020b; Li et al., 2021; Chang et al., 2022).

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* (7 genes) is currently the main hypothesis for the higher accumulation of ALA in *S. hispanica* seeds (Xue et al., 2023). However, we also specifically hypothesise that the six-gene tandem array of *ShSAD* genes in the telomeric region of chromosome 1, including five species-specific, co-expressed genes might further explain the high accumulation of ω-3 FAs in *S. hispanica* seeds. For example, regulation of the very long-chain monounsaturated nervonic acid (NA, 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 defining substrate and tissue specificity during FA elongation, and 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*. The overactivity of this *SAD* gene tandem array produces an abundance of C18:1-ACP as a substrate for the desaturation of FAs in both the plastid and ER, and consequently the relatively high accumulation of ω-3 FAs in *S. hispanica* seeds. This contradicts the hypothesis that the high expression of ER localised *ShFAD3* drives the high ω-3 FA accumulation in *S. hispanica* seeds (Li et al., 2023). 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

524 specific sites (Kryazhimskiy & Plotkin, 2008) while showing evidence of purifying selection
525 elsewhere.

526

527 **5 CONCLUSIONS**

528 This study generated a high-quality, chromosomal-level reference genome for *S. hispanica* to
529 analyse the evolution of oil biosynthesis genes in this valuable oil-seed crop. Our analysis
530 suggested that the expansion of the *ShSAD* gene family through tandem duplications is a driver of
531 high ω-3 FAs accumulation in *S. hispanica* seeds. Comparative analysis of multiple chromosomal-
532 level genomes is able to assess the putative effect of gene copy number variation and other source
533 of structural genome variation. This work establishes valuable genomic resources in chia and
534 prompts the need to investigate further structural variants at FA biosynthesis gene loci within and
535 among species. This will enable the breeding of emerging crop or the horizontal transfer of genes
536 across species, with the possibility of altering gene dosage through the introgression of arrays of
537 paralogous genes to improve key traits.

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 TZ. BE was supported by the Inaugural Botany Foundation Fellowship 2020 from the University of Melbourne Botany Foundation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Tannaz Zare: 0000-0002-7194-6800

Jeff F. Paril: 0000-0002-5693-4123

Emma Barnett: 0009-0002-6786-0972

Parwinder Kaur: 0000-0003-0201-0766

Rudi Appels: 0000-0002-5369-5227

Berit Ebert: 0000-0002-6914-5473

Ute Roessner: 0000-0002-6482-2615

Alexandre Fournier-Level: 0000-0002-6047-7164

SUPPLEMENTAL MATERIAL

Supplemental Figures

Supplemental Figure S1: Workflow of genomic DNA extraction from *S. hispanica*'s leaf tissue.

Supplemental Figure S2: An overview of the in situ Hi-C library preparation and sequencing.

Supplemental Figure S3: The Hi-C contact map of the *S. hispanica* genome assembly.

Supplemental Figure S4: Genome size estimation with GenomeScope using different k-mer lengths.

Supplemental Figure S5: Evaluation the completeness of the *S. hispanica* genome.

Supplemental Figure S6: Distribution of mapped tissue-specific transcripts from published assembled transcriptomes against the *S. hispanica* assembled genome.

Supplemental Figure S7: Distribution of repeat content and selected genomic features in each chromosome of *S. hispanica*.

Supplemental Figure S8: Gene Ontology (GO) enrichment analysis highlighting the top 10 enriched biological processes in the comparative genomics study of *S. hispanica* (p-value <0.05).

Supplemental Figure S9: Distribution of the Kernel density estimate (KDE) for transversion substitutions at fourfold degenerate sites (4dTV) for selected taxa studied here.

Supplemental Figure S10: Ka/Ks plots for three *ShSAD* genes of orthogroup OG0021142.

Supplemental Figure S11: Ka/Ks plots for three *ShSAD* genes of orthogroup OG0001529.

Supplemental Figure S12: Ka/Ks plots for a *ShSAD* gene (XP_047982897.1) and orthologous genes from *S. Splendens* belonging to orthogroup OG0001529.

Supplemental Figure S13: Ka/Ks plots for a ShSAD gene (XP_047982897.1) and orthologous genes from *A. thaliana* belonging to orthogroup OG0001529.

Supplemental Tables

Supplemental Table S1. Percentage of trimmed paired-end Illumina reads with Trimmomatic

Supplemental Table S2. List of lipid genes and their functions used in whole genome comparative analysis of *S. hispanica*.

Supplemental Table S3. Summary statistics of inferred Hi-C contacts and mapped Hi-C reads to the *S. hispanica* draft assembly.

Supplemental Table S4. Summary statistics of the six pseudo-chromosomes of *S. hispanica* obtained by Hi-C scaffolding.

Supplemental Table S5. Identified proteins, RNA molecules, genes, and pseudogenes in *S. hispanica* chromosomes

Supplemental Table S6. Summary of repeat elements in the genome assembly of *S. hispanica*.

Supplemental Table S7. Gene Ontology (GO) enrichment analysis highlighting the top 20 enriched biological processes in the comparative genomics study of *S. hispanica* (p-value <0.05).

Supplemental Table S8. Gene Ontology (GO) enrichment analysis highlighting the top 20 enriched molecular functions in the comparative genomics study of *S. hispanica* (p-value <0.05).

595 **Supplemental Table S9.** Gene Ontology (GO) enrichment analysis highlighting the enriched
596 cellular components in the comparative genomics study of *S. hispanica* (p-value <0.05).

597

598 **DATA AVAILABILITY**

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

603 **REFERENCES**

604 Achaz, G., Coissac, E., Viari, A., & Netter, P. (2000). Analysis of intrachromosomal
605 duplications in yeast Saccharomyces cerevisiae: a possible model for their origin.
606 *Molecular biology and evolution, 17*(8), 1268-1275.

607 Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. *Babraham*
608 *Bioinformatics*.

609 Arredondo-Mendoza, G. I., Jiménez-Salas, Z., Garza, F. J. G.-d. l., Solís-Pérez, E., López-
610 Cabanillas-Lomelí, M., González-Martínez, B. E., & Campos-Góngora, E. (2020).
611 Ethanolic Extract of Salvia hispanica L. Regulates Blood Pressure by Modulating the
612 Expression of Genes Involved in BP-Regulatory Pathways. *Molecules, 25*(17), 3875.

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

616 Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., Lelandais-Brière, C.,
617 Owens, G. L., Carrère, S., & Mayjonade, B. (2017). The sunflower genome provides
618 insights into oil metabolism, flowering and Asterid evolution. *nature, 546*(7656), 148-
619 152.

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

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

628 Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E.,
629 Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A.,
630 Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L. G., & Garmiri, P.
631 (2020). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research,*
632 *49*(D1), D480-D489.

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

636 Bates, P. D., Stymne, S., & Ohlrogge, J. (2013). Biochemical pathways in seed oil synthesis.
637 *Current opinion in plant biology, 16*(3), 358-364.

638 Block, M. A., & Jouhet, J. (2015). Lipid trafficking at endoplasmic reticulum–chloroplast
639 membrane contact sites. *Current opinion in cell biology, 35*, 21-29.

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

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

644 Browse, J., & Somerville, C. (1991). Glycerolipid synthesis: biochemistry and regulation.
645 *Annual review of plant biology, 42*(1), 467-506.

646 Cahill, J. P. (2003). Ethnobotany of chia, Salvia hispanica L.(Lamiaceae). *Economic Botany,*
647 *57*(4), 604-618.

648 Cai, Z., Zhao, X., Zhou, C., Fang, T., Liu, G., & Luo, J. (2023). Genome-Wide Mining of the
649 Tandem Duplicated Type III Polyketide Synthases and Their Expression, Structure
650 Analysis of Senna tora. *International Journal of Molecular Sciences, 24*(5), 4837.

651 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T.
652 L. (2009). BLAST+: architecture and applications. *BMC bioinformatics, 10*, 1-9.

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

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

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

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

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

670 Chen, J., Gao, J., Zhang, L., & Zhang, L. (2023). Tung tree stearoyl-acyl carrier protein Δ9
671 desaturase improves oil content and cold resistance of Arabidopsis and Saccharomyces
672 cerevisiae. *Frontiers in Plant Science, 14*.

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

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

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

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

684 Defoort, J., Van de Peer, Y., & Carretero-Paulet, L. (2019). The evolution of gene duplicates in
685 angiosperms and the impact of protein–protein interactions and the mechanism of
686 duplication. *Genome biology and evolution, 11*(8), 2292-2305.

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

690 dos Santos Maraschin, F., Kulcheski, F. R., Segatto, A. L. A., Trenz, T. S., Barrientos-Diaz, O.,
691 Margis-Pinheiro, M., Margis, R., & Turchetto-Zolet, A. C. (2019). Enzymes of glycerol-
692 3-phosphate pathway in triacylglycerol synthesis in plants: Function, biotechnological
693 application and evolution. *Progress in lipid research, 73*, 46-64.

694 Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim,
695 M. S., Machol, I., Lander, E. S., & Aiden, A. P. (2017). De novo assembly of the Aedes
696 aegypti genome using Hi-C yields chromosome-length scaffolds. *science, 356*(6333), 92-
697 95.

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

701 El-Feky, A. M., Elbatanony, M. M., Aboul Naser, A. F., Younis, E. A., & Hamed, M. A. (2022).
702 Salvia hispanica L. seeds extract alleviate encephalopathy in streptozotocin-induced
703 diabetes in rats: Role of oxidative stress, neurotransmitters, DNA and histological
704 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.

Georgiev, V., & Pavlov, A. (2017). Genetic Engineering and Manipulation of Metabolite Pathways in Salvia Spp. *Salvia Biotechnology*, 399-414.

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: β-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.

730 Gupta, P., Geniza, M., Naithani, S., Phillips, J. L., Haq, E., & Jaiswal, P. (2021). Chia (Salvia
731 hispanica) Gene Expression Atlas Elucidates Dynamic Spatio-Temporal Changes
732 Associated With Plant Growth and Development. *Frontiers in Plant Science, 12*, 667678.

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

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

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

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

744 Harley, R. M., Atkins, S., Budantsev, A. L., Cantino, P. D., Conn, B. J., Grayer, R., Harley, M.
745 M., De Kok, R. d., Krestovskaja, T. d., & Morales, R. (2004). Labiatae. *Flowering*
746 *Plants· Dicotyledons: Lamiales (except Acanthaceae including Avicenniaceae), 7*, 167-
747 275.

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

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

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

757 Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science &*
758 *engineering, 9*(03), 90-95.

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

762 Jiang, S.-Y., González, J. M., & Ramachandran, S. (2013). Comparative genomic and
763 transcriptomic analysis of tandemly and segmentally duplicated genes in rice. *PLoS One,*
764 *8*(5), e63551.

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

769 Klein, A., Husselmann, L. H., Williams, A., Bell, L., Cooper, B., Ragar, B., & Tabb, D. L.
770 (2021). Proteomic Identification and Meta-Analysis in Salvia hispanica RNA-Seq de
771 novo Assemblies. *Plants, 10*(4), 765.

772 Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a
773 changing environment. *Proceedings of the Royal Society B: Biological Sciences,*
774 *279*(1749), 5048-5057.

775 Kryazhimskiy, S., & Plotkin, J. B. (2008). The Population Genetics of dN/dS. *PLOS Genetics,*
776 *4*(12), e1000304.

777 Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: a resource for timelines,
778 timetrees, and divergence times. *Molecular biology and evolution, 34*(7), 1812-1819.

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

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

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

787 Li, J., Wang, Y., Dong, Y., Zhang, W., Wang, D., Bai, H., Li, K., Li, H., & Shi, L. (2021). The
788 chromosome-based lavender genome provides new insights into Lamiaceae evolution and
789 terpenoid biosynthesis. *Horticulture research, 8*.

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

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

796 Li-Beisson, Y., Shorrosh, B., Beisson, F., Andersson, M. X., Arondel, V., Bates, P. D., Baud, S.,
797 Bird, D., DeBono, A., & Durrett, T. P. (2013). Acyl-lipid metabolism. *The Arabidopsis*
798 *book/American Society of Plant Biologists, 11*.

799 Lin, P., Wang, K., Wang, Y., Hu, Z., Yan, C., Huang, H., Ma, X., Cao, Y., Long, W., & Liu, W.
800 (2022). The genome of oil-Camellia and population genomics analysis provide insights
801 into seed oil domestication. *Genome biology, 23*, 1-21.

802 Liu, C., Wu, Y., Liu, Y., Yang, L., Dong, R., Jiang, L., Liu, P., Liu, G., Wang, Z., & Luo, L.
803 (2021). Genome-wide analysis of tandem duplicated genes and their contribution to stress
804 resistance in pigeonpea (Cajanus cajan). *Genomics, 113*(1), 728-735.

805 Liu, Y.-X., Yu, J.-H., Sun, J.-H., Ma, W.-Q., Wang, J.-J., & Sun, G.-J. (2023). Effects of Omega-
806 3 Fatty Acids Supplementation on Serum Lipid Profile and Blood Pressure in Patients
807 with Metabolic Syndrome: A Systematic Review and Meta-Analysis of Randomized
808 Controlled Trials. *Foods, 12*(4), 725.

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

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

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

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

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

821 Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T., & Thomas, P. D.
822 (2021). PANTHER version 16: a revised family classification, tree-based classification
823 tool, enhancer regions and extensive API. *Nucleic Acids Research, 49*(D1), D394-D403.

824 Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler,
825 A., & Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for
826 phylogenetic inference in the genomic era. *Molecular biology and evolution, 37*(5), 1530-
827 1534.

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

Motyka, S., Koc, K., Ekiert, H., Blicharska, E., Czarnek, K., & Szopa, A. (2022). The current state of knowledge on Salvia hispanica and Salviae hispanicae semen (Chia Seeds). *Molecules, 27*(4), 1207.

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*.

Onneken, P. (2018). Salvia hispanica L (Chia Seeds) as brain superfood: how seeds increase intelligence. *J Nutr Food Sci, 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. *bioRxiv*, 2022.2007. 2018.499821.

Paril, J., Zare, T., & Fournier-Level, A. (2023). compare_genomes: a comparative genomics workflow to streamline the analysis of evolutionary divergence across genomes. *bioRxiv*, 2023.2003. 2016.533049.

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.

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

858 Picart-Picolo, A., Grob, S., Picault, N., Franek, M., Llauro, C., Halter, T., Maier, T. R., Jobet, E.,
859 Descombin, J., & Zhang, P. (2020). Large tandem duplications affect gene expression,
860 3D organization, and plant–pathogen response. *Genome Research, 30*(11), 1583-1592.

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

864 Qu, J., Wang, B., Xu, Z., Feng, S., Tong, Z., Chen, T., Zhou, L., Peng, Z., & Ding, C. (2023).
865 Genome-Wide Analysis of the Molecular Functions of B3 Superfamily in Oil
866 Biosynthesis in Olive (Olea europaea L.). *BioMed Research International, 2023*.

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

869 Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T.,
870 Sanborn, A. L., Machol, I., Omer, A. D., & Lander, E. S. (2014). A 3D map of the human
871 genome at kilobase resolution reveals principles of chromatin looping. *Cell, 159*(7),
872 1665-1680.

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

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

877 Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature methods,*
878 *17*(2), 155-158.

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

882 Shen, T.-f., Huang, B., Xu, M., Zhou, P.-y., Ni, Z.-x., Gong, C., Wen, Q., Cao, F.-l., & Xu, L.-A.
883 (2022). The reference genome of Camellia chekiangoleosa provides insights into
884 Camellia evolution and tea oil biosynthesis. *Horticulture research, 9*.

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

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

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

895 Sun, Y., Shao, J., Liu, H., Wang, H., Wang, G., Li, J., Mao, Y., Chen, Z., Ma, K., & Xu, L.
896 (2023). A chromosome-level genome assembly reveals that tandem-duplicated CYP706V
897 oxidase genes control oridonin biosynthesis in the shoot apex of Isodon rubescens.
898 *Molecular Plant, 16*(3), 517-532.

899 Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements in
900 Genomic Sequences. *Current Protocols in Bioinformatics, 25*(1), 4.10.11-14.10.14.

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

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

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

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

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

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

917 Wang, D., Zhang, Y., Zhang, Z., Zhu, J., & Yu, J. (2010). KaKs_Calculator 2.0: a toolkit
918 incorporating gamma-series methods and sliding window strategies. *Genomics,*
919 *proteomics & bioinformatics, 8*(1), 77-80.

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

922 Wang, L., Lee, M., Sun, F., Song, Z., Yang, Z., & Yue, G. H. (2022). A chromosome-level
923 genome assembly of chia provides insights into high omega-3 content and coat color
924 variation of its seeds. *Plant Communications, 3*(4), 100326.

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

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

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

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

936 Wu, Z., Liu, H., Zhan, W., Yu, Z., Qin, E., Liu, S., Yang, T., Xiang, N., Kudrna, D., & Chen, Y.
937 (2021). The chromosome-scale reference genome of safflower (Carthamus tinctorius)
938 provides insights into linoleic acid and flavonoid biosynthesis. *Plant Biotechnology*
939 *Journal, 19*(9), 1725-1742.

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

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

946 Xu, Z., Gao, R., Pu, X., Xu, R., Wang, J., Zheng, S., Zeng, Y., Chen, J., He, C., & Song, J.
947 (2020a). Comparative genome analysis of Scutellaria baicalensis and Scutellaria barbata
948 reveals the evolution of active flavonoid biosynthesis. *Genomics, proteomics &*
949 *bioinformatics, 18*(3), 230-240.

950 Xu, Z., Pu, X., Gao, R., Demurtas, O. C., Fleck, S. J., Richter, M., He, C., Ji, A., Sun, W., &
951 Kong, J. (2020b). Tandem gene duplications drive divergent evolution of caffeine and
952 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., Li, L., Liu, X., Jiang, H., Zhao, Y., Wei, S., Lin, N., & Chai, Y. (2021). Molecular Cloning and Characterization of FAD6 Gene from Chia (Salvia hispanica L.). *Biochemical Genetics, 59*, 1295-1310.

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.

Xue, Y., Yin, N., Chen, B., Liao, F., Win, A. N., Jiang, J., Wang, R., Jin, X., Lin, N., & Chai, Y. (2017). Molecular cloning and expression analysis of two FAD2 genes from chia (Salvia hispanica). *Acta physiologiae plantarum, 39*, 1-12.

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). *J Proteomics Bioinform, 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, M., Fan, J., Taylor, D. C., & Ohlrogge, J. B. (2009). DGAT1 and PDAT1 acyltransferases have overlapping functions in Arabidopsis triacylglycerol biosynthesis

and are essential for normal pollen and seed development. *The Plant Cell, 21*(12), 3885-3901.

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.

FIGURES AND TABLES

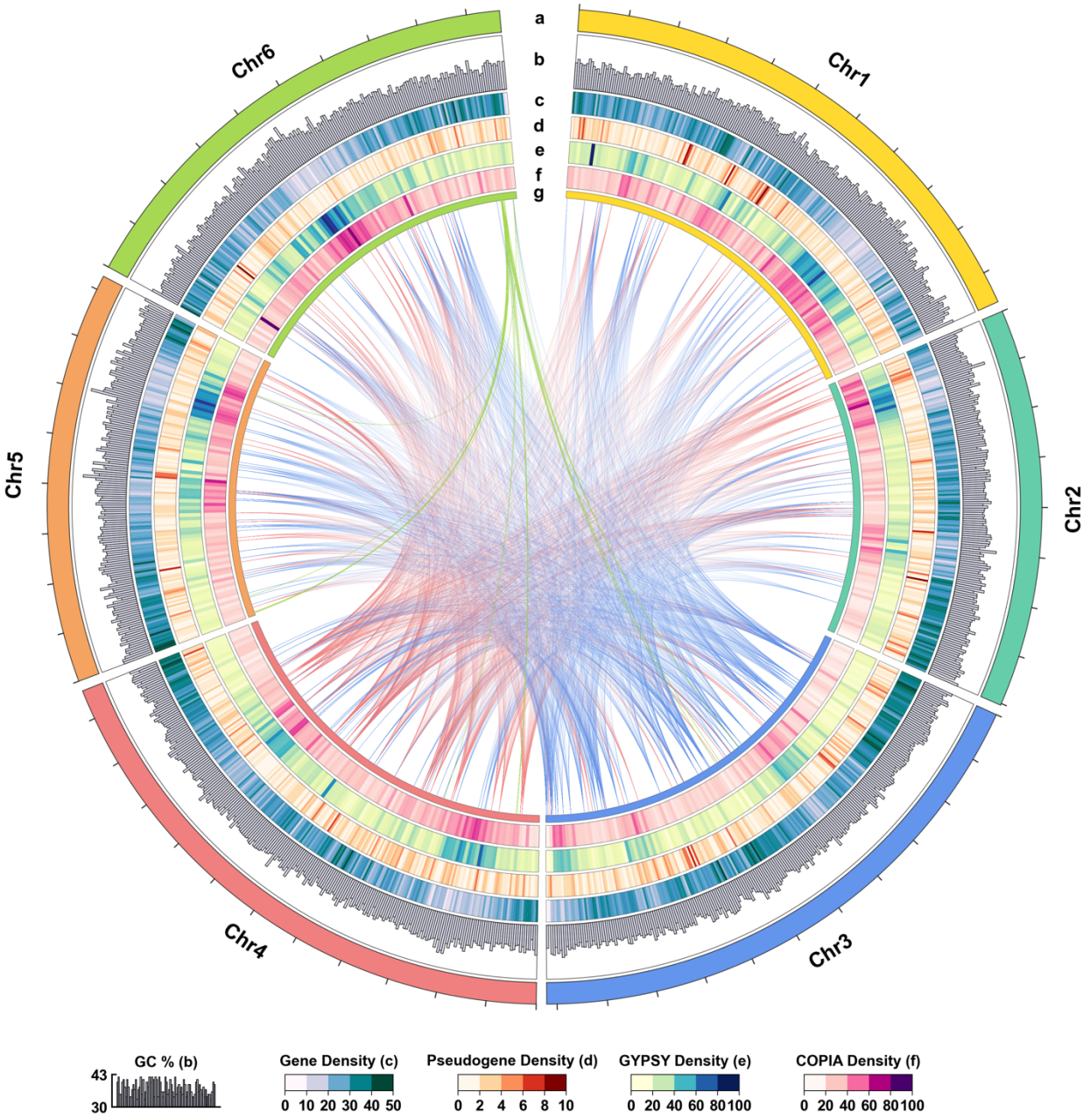


Figure 1. Genomic features of the *S. hispanica* reference genome assembly. (a) Chromosome layer showing the length of each chromosome with ticks indicating 5Mbp intervals. (b) Distribution of the GC content at a window size of 250kbp over the entire genome as a bar plot of percentage

992 values with a lower bound of 30% and upper bound of 43%. (c) Distribution of protein coding
993 gene density over 250kbp windows (values normalized between 0 and 50 across chromosomes).
994 (d) Distribution of pseudogene density over 250kbp windows (values normalized between 0 and
995 10 for all chromosomes). (e) and (f) Distribution of Gypsy and Copia LTR density, respectively,
996 over 500kbp windows (values normalized between 0 and 100 across chromosomes). (g) The chord
997 plot shows the synteny relationships for the top five orthogroups (paralogs) across the genome.
998 The color of the internal chords is that of the chromosome containing the highest number of
999 paralogs within each orthogroup.

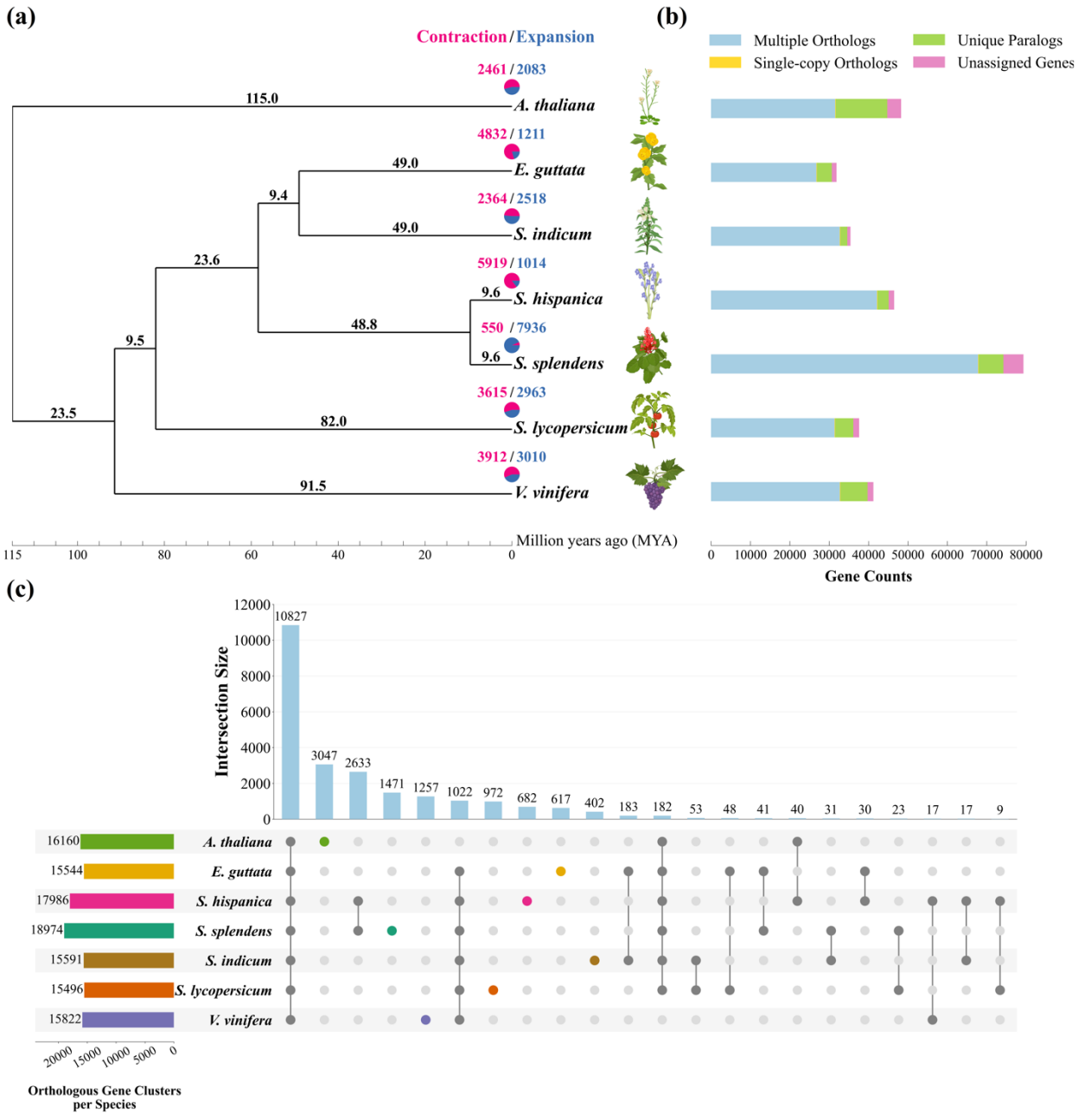


Figure 2. Evolution of *S. hispanica* and distribution of orthologous gene families across species. (a) Phylogenetic tree inferred from single-copy orthologs among selected species. Numbers on branches show divergence time in MYA. The pie charts at the terminal branches show the contraction (pink) and 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 in 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.

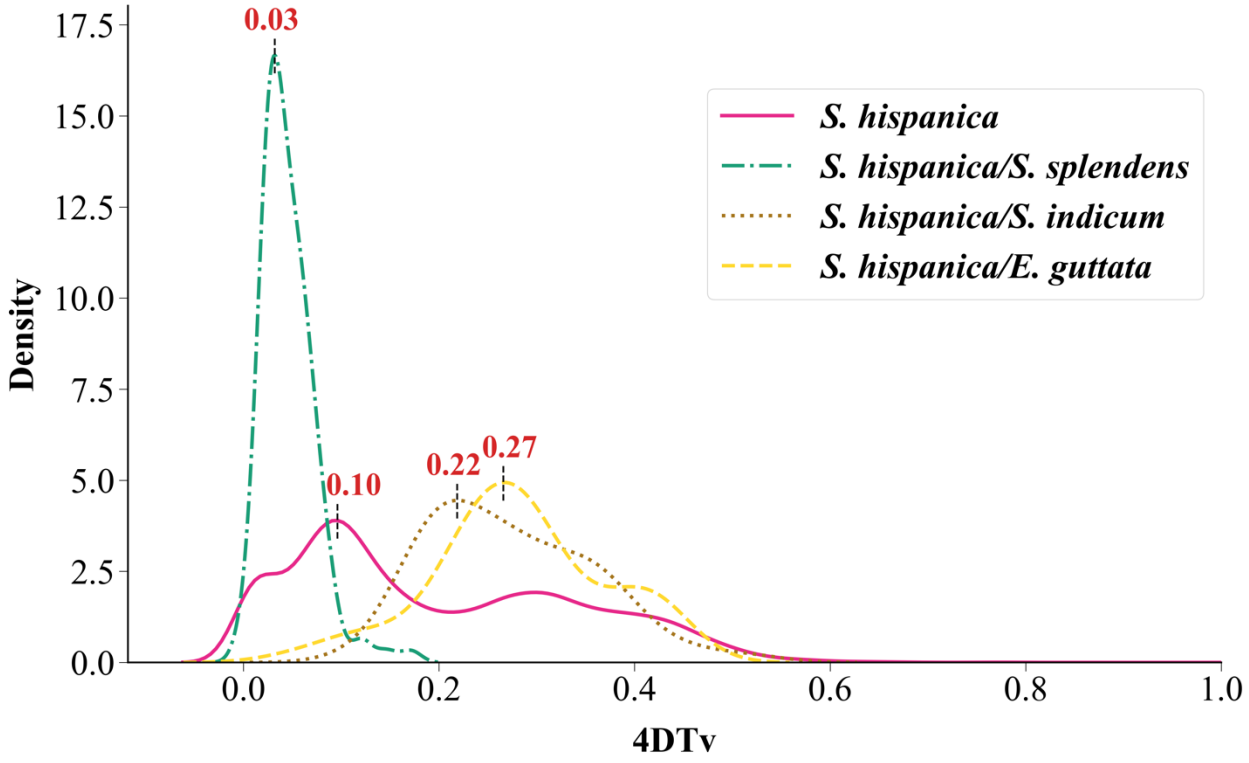


Figure 3. Distribution of transversion substitutions at fourfold degenerate sites (4DTv). Distribution of 4DTv for *S. hispanica* and pairwise 4DTv with *S. splendens*, *S. indicum*, and *E. guttata*. Peaks in pairwise 4DTv density indicate the relative time of divergence between species.

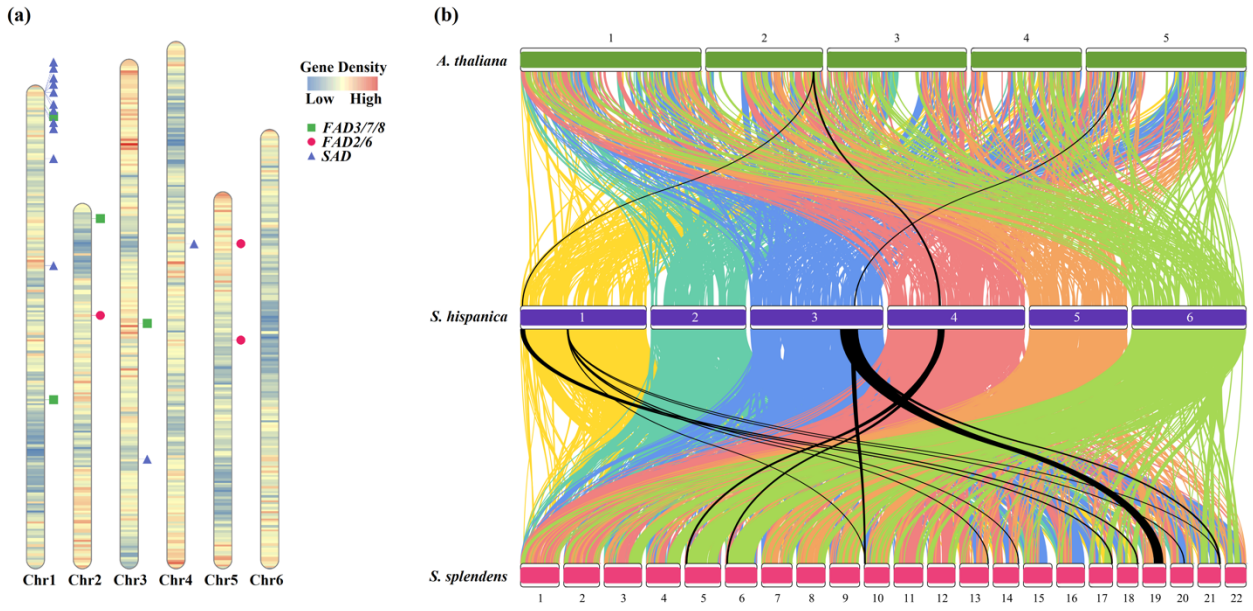


Figure 4. Chromosome ideogram and synteny analysis of *S. hispanica* genome. (a) Ideogram showing the gene density distribution and position of key FA synthesis genes on *S. hispanica* chromosomes. The tandem array of *ShSAD* genes is located in the telomeric region of chromosome 1. (b) Synteny analysis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

1024 **Table 1.** Statistics for the chromosome-level reference genome assembly of *S. hispanica*.

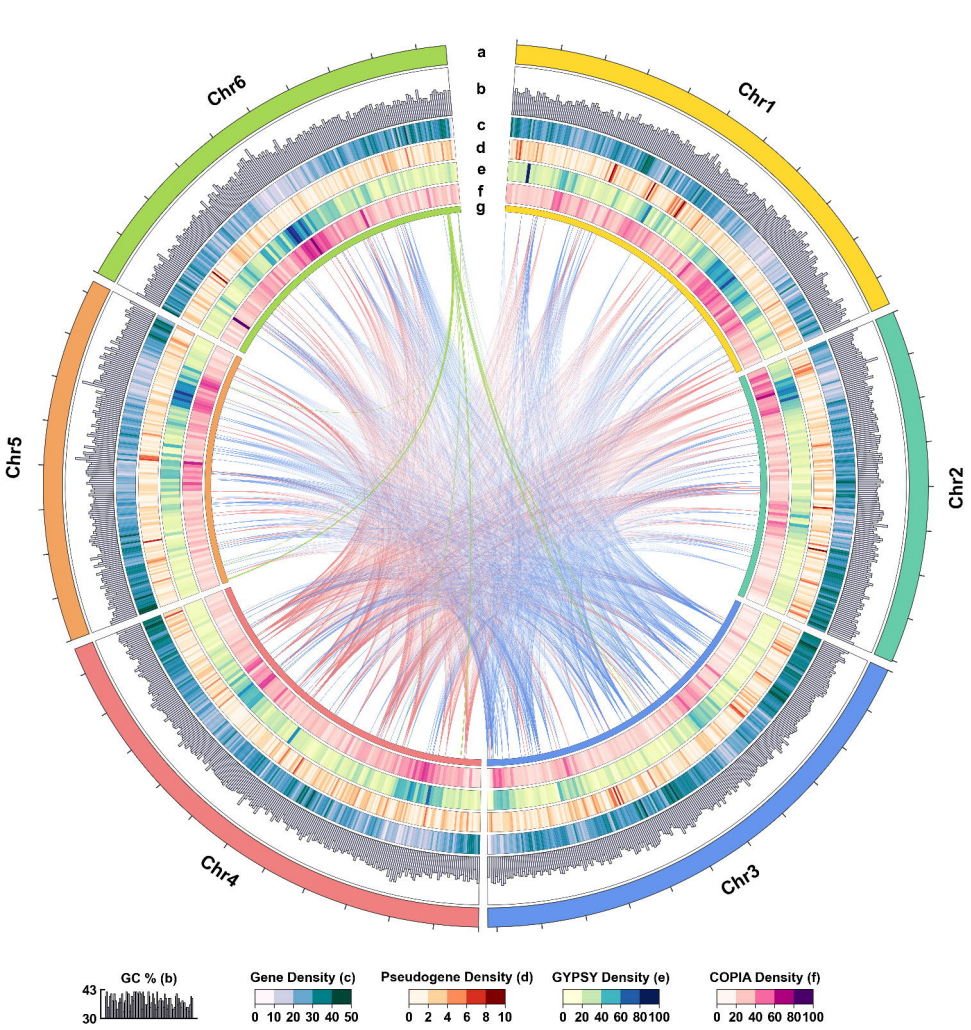
Assembly Features	***S. hispanica* Genome Assembly**
Total assembly length (bp)	321,469,233
Contigs ≥ 1k bp	5,301
Contigs ≥ 10 kbp	4,307
Contigs ≥ 25 kbp	4,016
Contigs ≥ 50 kbp	3,858
Largest contig length (kbp)	894
Scaffolds ≥ 1 kbp	1,553
Scaffolds ≥ 10 kbp	572
Scaffolds ≥ 25 kbp	288
Scaffolds ≥ 50 kbp	130
Largest Scaffold length (kbp)	57,985
% Main genome in scaffolds > 50 kbp	95.66
GC (%)	36
N50	53,190,533
L50	3
N90	40,175,719
L90	6
Mismatches (N's)	613,150
Gap (%)	0.19

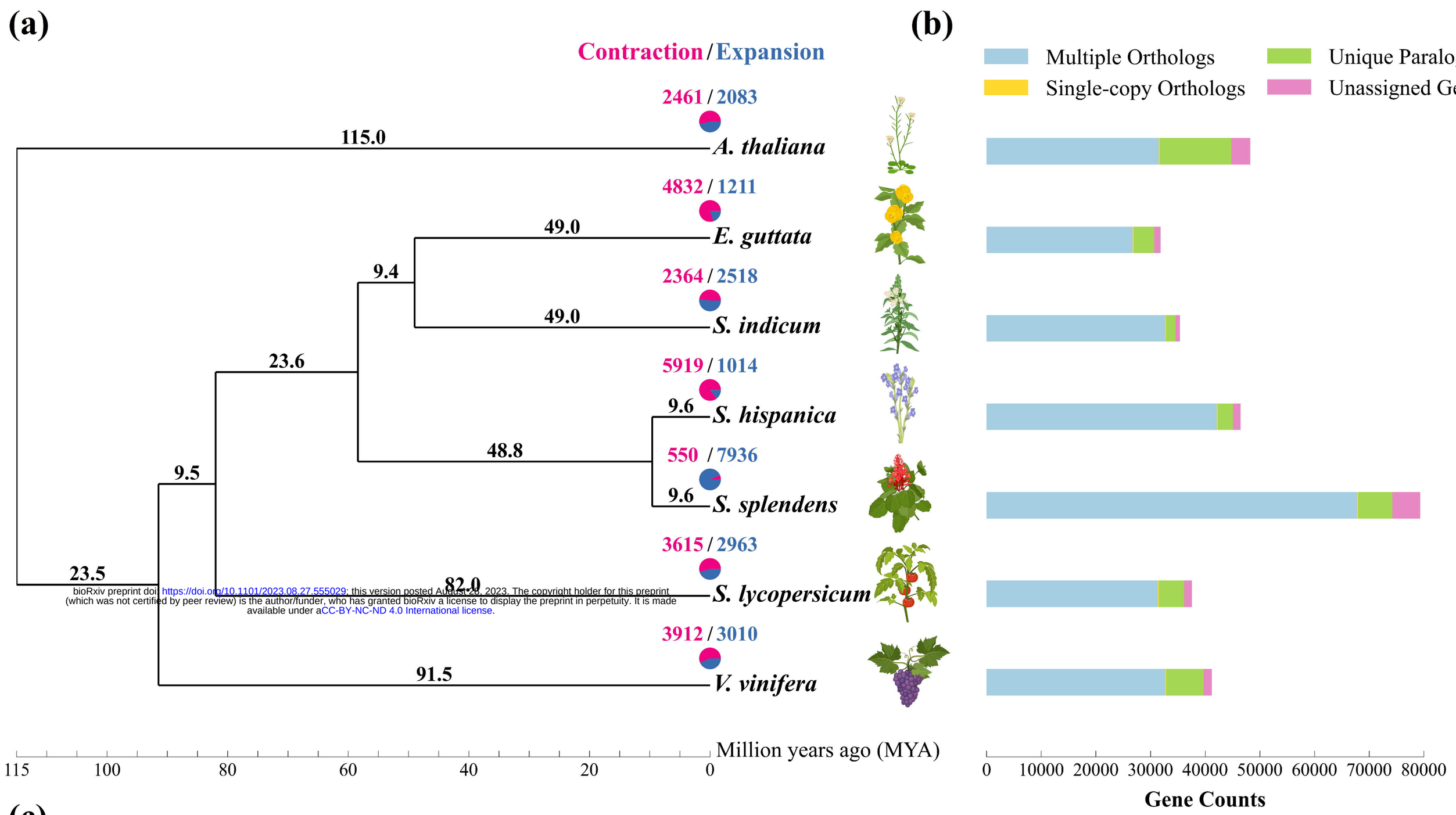
1025

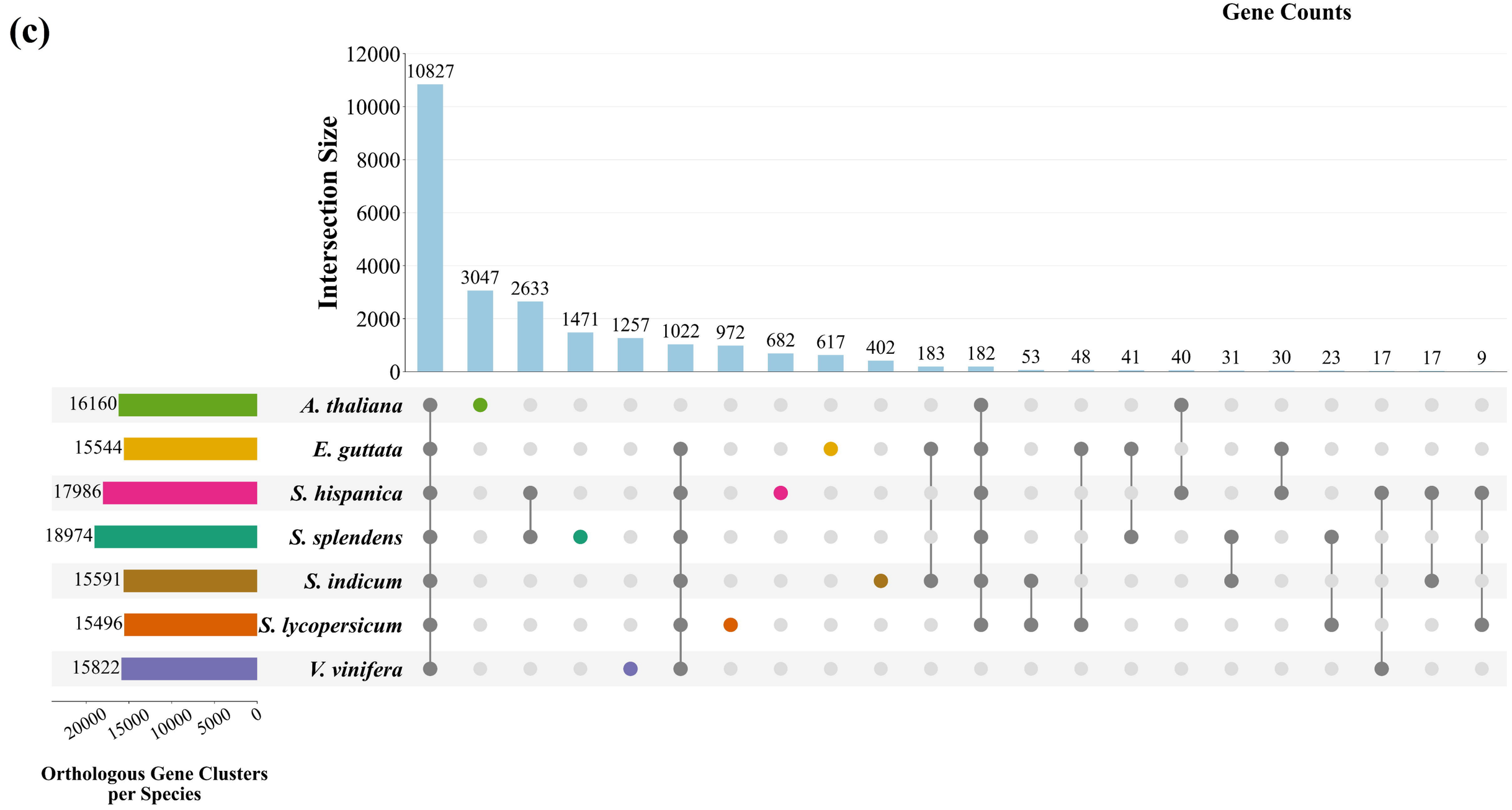
1026 **Table 2.** Count and length of the annotated genomic features in the *S. hispanica* genome
1027 (excluding pseudogenes).

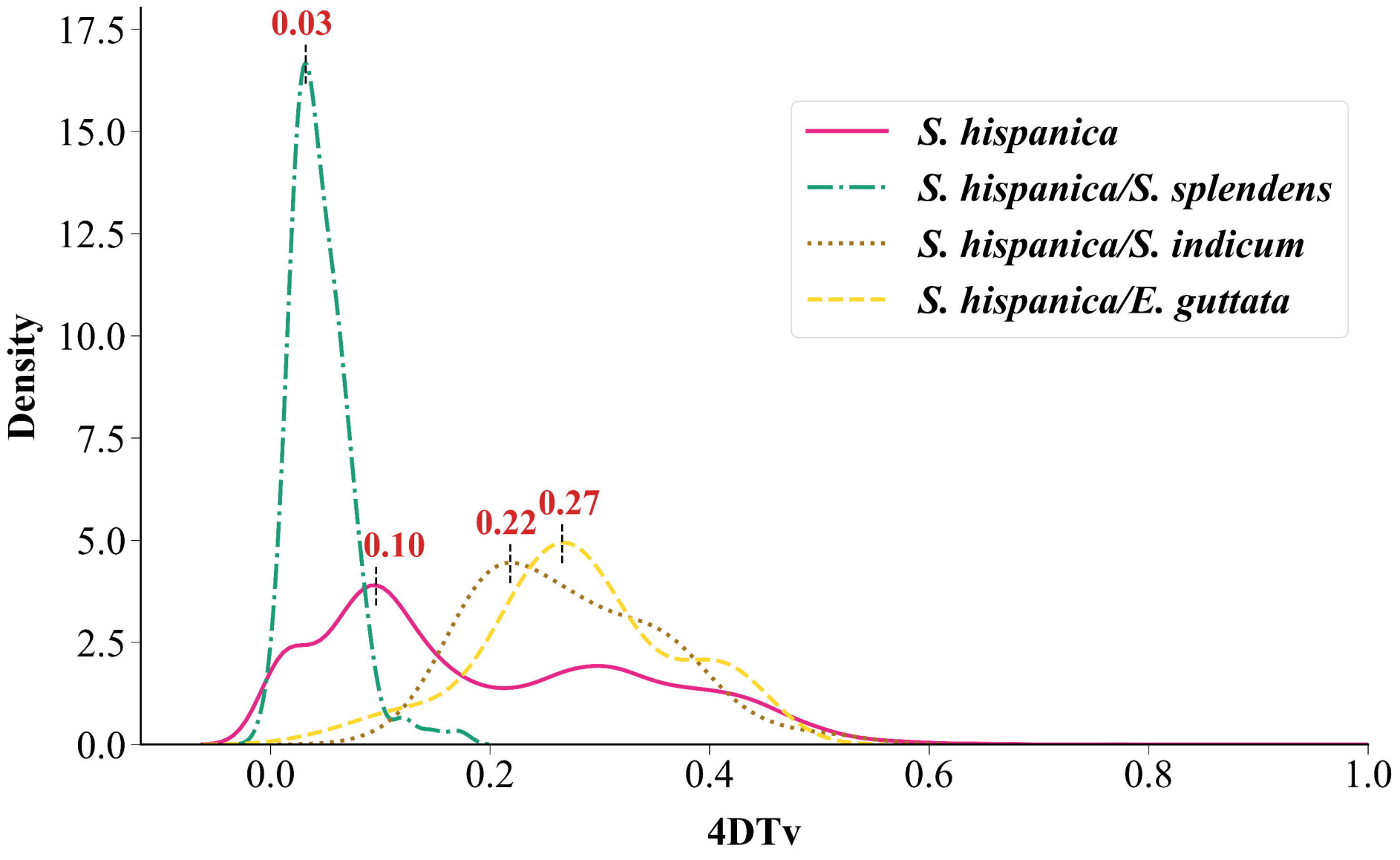
Feature	Count	Mean length (bp)	Median length (bp)	Min length (bp)	Max length (bp)
Genes	36,993	2,901	2,291	62	163,900
All transcripts	54,009	1,671	1,452	62	16,722
mRNA	46,423	1,753	1,515	165	16,722
Misc_RNA	2,381	2,122	1,859	167	13,008
tRNA	739	74	73	71	93
lncRNA	3,758	979	729	78	5,754
snoRNA	436	106	103	62	229
snRNA	223	138	120	98	197
rRNA	49	384	119	103	3,191
Single exon transcripts	5,396	1,149	957	233	6,688
CDSs	46,508	1,379	1,155	90	16,188
Exons	209,379	302	162	2	7,672
Exons in coding transcripts	197,070	302	161	2	7,062
Exons in non-coding transcripts	19,006	274	153	2	7,672
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

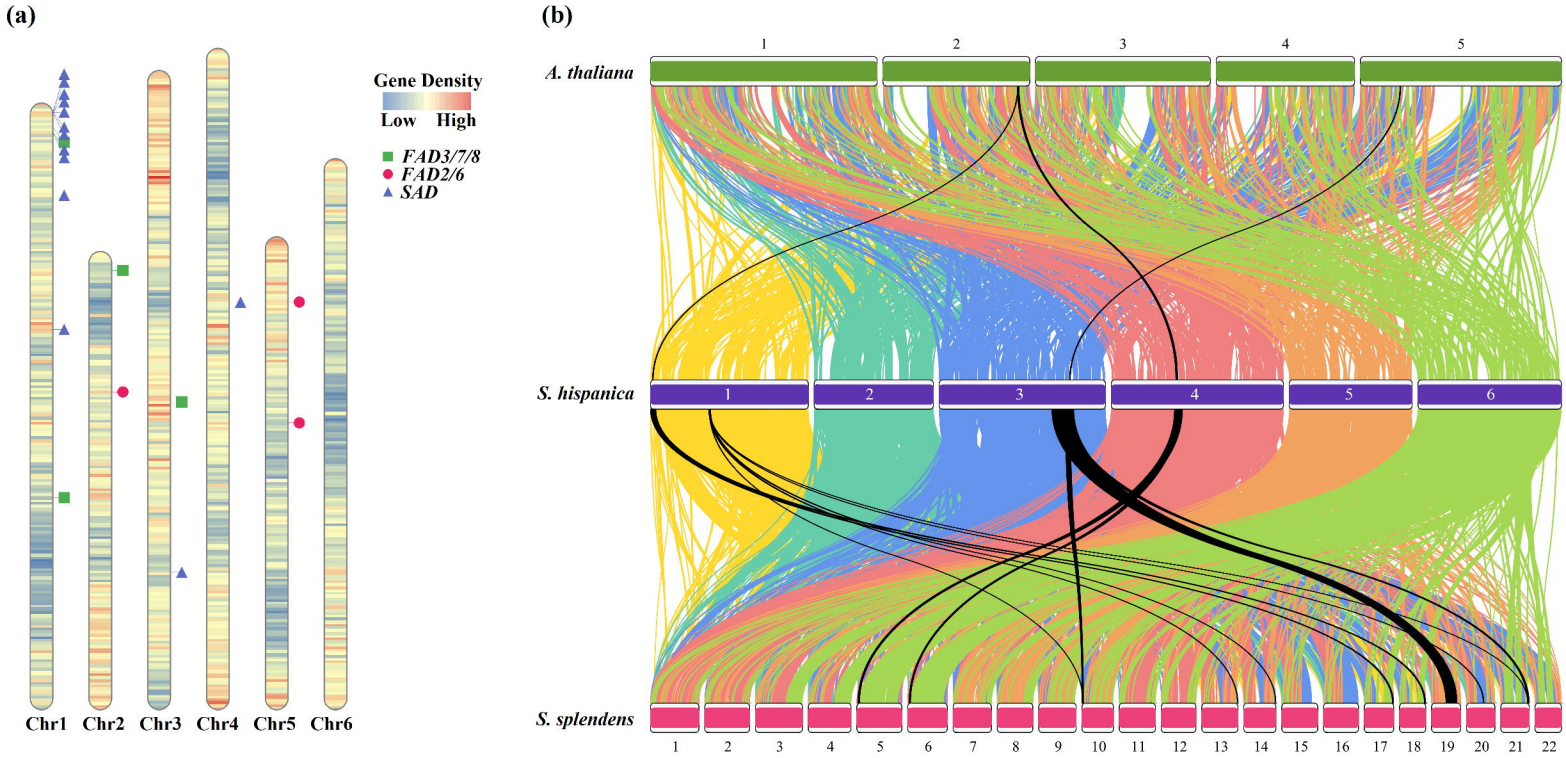
1028


a
b
c
d
e
f
g
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
GC % (b)
43
30
Gene Density (c)
0 10 20 30 40 50
Pseudogene Density (d)
0 2 4 6 8 10
GYPSY Density (e)
0 20 40 60 80 100
COPIA Density (f)
0 20 40 60 80 100






0.03
0.10
0.22
0.27
S. hispanica
S. hispanica/S. splendens
S. hispanica/S. indicum
S. hispanica/E. guttata
Density
4DTv
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0


(a)
Gene Density
Low High
FAD3/7/8
FAD2/6
SAD
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
(b)
A. thaliana
S. hispanica
S. splendens
1
2
3
4
5
6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22