Chromosome-scale *Salvia hispanica* L. (Chia) genome assembly reveals rampant *Salvia* interspecies introgression

Julia Brose[1], John P. Hamilton[2,3], Nicholas Schlecht[4], Dongyan Zhao[1], Paulina M. Mejía-Ponce[5], Arely Cruz Pérez[5], Brieanne Vaillancourt[2], Joshua C. Wood[2], Patrick P. Edger[6], Salvador Montes-Hernandez[7], Guillermo Orozco de Rosas[8], Björn Hamberger[4], Angélica Cibrian Jaramillo[5], and C. Robin Buell[2,3,9]

[1]Department of Plant Biology, Michigan State University, East Lansing, MI, USA

[2]Center for Applied Genetic Technologies, University of Georgia, Athens, GA, USA

[3]Department of Crop and Soil Sciences, University of Georgia, Athens, GA, USA

[4]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA

[5]National Laboratory for Genomics of Biodiversity (UGA-Langebio), CINVESTAV, Irapuato, Guanajuato, Mexico

[6]Department of Horticulture, Michigan State University, East Lansing, MI, USA

[7]Campo Experimental Bajío, Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Km 6.5 carretera Celaya-San Miguel de Allende, C.P. 38110, Celaya, Guanajuato, México.

[8]CHIABLANCA SC DE RL, Acatic, Jalisco. Mexico

[9]Institute of Plant Breeding, Genetics, and Genomics, University of Georgia, Athens, GA, USA

ABSTRACT

Salvia hispanica L. (Chia), a member of the Lamiaceae, is an economically important crop in Mesoamerica, with health benefits associated with its seed fatty acid composition. Chia varieties are distinguished based on seed color including mixed white and black (Chia pinta) and black (Chia negra). To facilitate research on Chia and expand on comparative analyses within the Lamiaceae, we generated a chromosome-scale assembly of a Chia pinta accession and performed comparative genome analyses with a previously published Chia negra genome assembly. The Chia pinta and negra genome sequences were highly similar as shown by a limited number of single nucleotide polymorphisms and extensive shared orthologous gene membership. There is an enrichment of terpene synthases in the Chia pinta genome relative to the Chia negra genome. We sequenced and analyzed the genomes of 20 Chia accessions with differing seed color and geographic origin revealing population structure within *S. hispanica* and interspecific introgressions of *Salvia* species. As the genus *Salvia* is polyphyletic, its evolutionary history remains unclear. Using large-scale synteny analysis within the Lamiaceae and orthologous group membership, we resolved the phylogeny of *Salvia* species. This study and its collective resources further our understanding of genomic diversity in this food crop and the extent of inter-species hybridizations in *Salvia*.

PLAIN LANGUAGE SUMMARY

Chia pinta is an economically important crop due to the high fatty acid present in the seeds. There are multiple types of Chia based on the seeds color including mixed which and black (Chia pinta), black (Chia negra), and white (Chia blanca). We generated a genome assembly of Chia pinta and compared it to existing genome assemblies. While the assemblies are highly similar there are key differences in terpene synthase composition between Chia pinta and Chia negra. We also sequenced 20 other Chia accessions with different seed color and geographic origin to determine a population structure within Chia. We generated genomic resources to further our understanding of this food crop.

ABBREVIATIONS

BGC Biosynthetic gene cluster

BUSCO Benchmarking Universal Single Copy Orthologs

GO Gene ontology

SNP Single nucleotide polymorphism

TIR Terminal inverted repeat

TPS Terpene synthase

WGS Whole genome shotgun

INTRODUCTION

Chia (*Salvia hispanica* L.) belongs to the largest genus within the Lamiaceae containing approximately 980 species (Hu et al., 2018). Chia is a notable and economically important species within the *Salvia* genus attributable to the high nutritional value of its seeds which contain 16-26% protein, 23-41% fiber, and 20-34% polyunsaturated fatty acids, of which, 60% is α-lineolic acid (Muñoz et al., 2013). Historically, Chia was the third most economically important crop in Mesoamerica, only behind maize and amaranth, due to its use in religious practices and as a medicine (Valdivia-López & Tecante, 2015). The medicinal properties of Chia include treatments for gastrointestinal, respiratory, urinary, obstetrics, skin, central nervous, and ophthalmologic issues (Cahill, 2003). The traditional uses of Chia revolve around religious practices which contributed to the decrease of Chia prominence and cultivation in the 15th century following the invasion by conquistadors (Cahill, 2003). Chia was introduced to Spain where it was named by Linnaeus as *Salvia hispanica* referencing the presumed origin of Spain (Baldivia, 2018). While Chia originated in present day Mexico and Guatemala, it has since been distributed throughout the world resulting in the emergence of diverse varieties (Cahill, 2004).

Chia varieties are characterized by their seed color and origin. The widely cultivated Chia blanca has a white seed coat while Chia negra has a black seed coat that can occur in wild and cultivated populations. Other seed coat colors include mixes of black and white seeds. Morphological characteristics distinguishing cultivated from wild accessions mirror traits observed in other domesticated species, such as decreased apical dominance, increased branching, increased seed size, decreased pubescence, increased florescence length determinism, increased anthocyanin pigmentation, variation in seed coat color and patterns, increased plant height, and closed calyxes (Cahill, 2004). While phenotypically distinct, dietary proteins are similar in wild and cultivated Chia accessions although wild accessions with higher levels of polyunsaturated fatty acids have been reported (Peláez et al., 2019).

Robust genomic resources for the Lamiaceae facilitate comparative genomic analysis. Within the Lamiaceae there are seven subfamilies with chromosome-scale genomes [Ajugoideae, Callicarpoideae, Nepetoideae, Lamiodeae, Scutellariodeae, and Tectonoideae] (Dong et al., 2018; Zhao et al., 2019a; b; Hamilton et al., 2020; He et al., 2022; Li et al., 2022; Shen et al., 2022; Sun et al., 2022; Pan et al., 2023). Current genomic resources for Chia include a genome assembly derived from an Australian black seeded variety (Chia negra; Wang et al., 2022), a white seeded variety (Chia blanca; Li et al., 2023), and a Mexican Chia (Alejo-Jacuinde et al., 2023) as well as transcriptomes constructed from wild and cultivated seeds (Peláez et al., 2019). Expanding the number and diversity of chia accessions with genome assemblies and sequence will facilitate our understanding of genetic diversity of this important crop as well as provide resources for more informed breeding programs. In addition to diversity within Chia, three other *Salvia* species occur in the same region in Mesoamerica (*Salvia uruapan* Fern., *Salvia tiliifolia* Vahl., and *Salvia polystachya* Ort.) that have similar uses as *S. hispanica* (Cahill, 2003). These species are challenging to distinguish from each other, but no reports indicate hybridization with *S. hispanica*. A phylogeny of *Salvia,* based on 91 nuclear genes, places Chia within *Salvia* sect. *Potiles* in a monophyletic clade (Lara-Cabrera et al., 2021). However, the *Salvia* genus has yet to be fully resolved and remains polyphyletic with *S. tiliifolia* being placed within two separate clades: the Angulatae and Polystachyae (Lara-Cabrera et al., 2021). Therefore, additional phylogenetic analyses are necessary to achieve a comprehensive resolution of the *Salvia* genus.

In this study, we report on the genome sequence of a Chia pinta accession, comparative analyses with published Chia genomes, and analysis of genetic diversity in a set of 20 Chia accessions revealing population structure between domesticated and wild Chia species and evidence of interspecies hybridization of *S. tiliifolia* with Chia.

RESULTS AND DISCUSSION

Chia Genome

We selected a Chia pinta accession from Acatic**,** Jalisco, Mexico that produces mixed color seeds and is grown as a superfood source throughout Mexico. Using 5.7 million PacBio long reads (36.5 Gb) representing ~100x coverage of the predicted ~355 Mbp Chia genome (Wang et al., 2022), we assembled the Chia pinta (2n=2x=12) genome using Canu (Koren et al., 2017). Whole genome shotgun (WGS) reads were used to generate a k-mer (k=21) distribution profile using GenomeScope indicating an estimated genome size of 338 Mbp with 62.6% unique kmers and 0.5% heterozygosity. The initial Canu assembly was error corrected using the raw PacBio reads using Arrow (Pacific Biosciences) followed by three rounds of error correction with the Illumina WGS reads using Pilon (Walker et al., 2014). The error-corrected assembly consisted of 2,094 contigs with a total length of 425.14 Mbp, which is substantially larger than the previously estimated genome size. Haplotigs were removed from the assembly using purgeHaplotigs (Roach et al., 2018) (-a = 50%) with an output consisting of "primary contigs" representing the putative haploid genome sequence, "haplotigs" containing diverged haplotypes, and "artefacts" representing contigs with very low or extremely high read coverage. Following removal of haplotigs, the "primary contigs" size decreased from 425 Mbp to 343 Mbp (Table 1). Manual examination of Chia vs. Chia self-alignments of contigs in the 'purged assembly' revealed five pairs of contigs that were putative residual haplotigs. Removal of these contigs resulted in a 'purged assembly' containing 407 contigs with an N50 contig length of 1.5 Mbp and a total size of 343.2 Mbp. The distribution of k-mers from WGS reads in the final assembly was examined using KAT (Mapleson et al., 2017) revealing a single peak indicating a haploid assembly with few retained haplotigs.

Using Hi-C sequence data, the contigs were assembled into six pseudomolecules, consistent with the known chromosome number of Chia and the Chia negra Australian Black (hereafter Chia negra) genome assembly (Wang et al., 2022). The final Chia pinta genome assembly was 342 Mb final with an N50 of 62Mb, of which, 99.64% of the assembly was anchored to one of the six pseudochromosomes (Table 1). Metrics for the final chromosome assembly were calculated using only the six chromosomes. The GC content of the final assembly was 36.6% consistent with the previously published Chia negra genome (Wang et al., 2022). Alignment of Illumina WGS reads to the final assembly revealed 98.4% of the reads aligned to the genome, of which, 99.5% were properly paired. Alignment of RNA-seq reads from a diverse set of tissue types (leaf, inflorescence, stem, and root) showed an overall alignment rate between 93.7% and 96.0%. To confirm the quality of the Chia pinta assembly, we used Benchmarking Universal Single Copy Orthologs (BUSCO) (Simão et al., 2015) to determine the representation of conserved orthologs in the final assembly. In total, 97.4% of the BUSCO orthologs were complete with 86.6% as single copy, 10.8% duplicated, 0.7% fragmented, and 1.9% missing. Overall, these results indicate a high-quality Chia pinta genome assembly.

Repetitive Sequences and Transposable Element Annotation in the Chia pinta genome

Using *de novo* repetitive sequence identification with RepeatModeler coupled with sequences from the Viridiplantae RepBase, RepeatMasker masked 46.8% of the Chia pinta genome. With respect to transposable elements, retroelements were the dominant sequence with 40,151 retroelements occupying 15.15% of the Chia pinta genome while DNA transposons (36,807 elements) accounted for 4.86%. Unclassified interspersed repeats represented the largest number of elements with 378,795 or 26.11% of the genome. The remaining repetitive elements included rolling circle, small RNA, satellites, simple repeats, and low complexity sequences make up less than 1% of the genome.

The Extensive *de-novo* TE Annotator (EDTA) was used to annotate the Chia pinta genome for transposable elements revealing 314,306 elements spanning 149,780,410 bp (43.64%) of the Chia pinta genome. Long terminal repeats comprise 21.33% of the genome, of which, 5.7% were *Copia* elements and 11.45% were *Gypsy* elements; unknown long terminal repeats comprise 4.13% of the genome. Terminal inverted repeat (TIR) sequences represent 20.01% of the genome with the largest portion (12.09%) belonging to Tc1_Mariner family. The remaining TIRs are PIF_Harbinger (3.26%), hAT (2.32%), Mutator (1.80%), and CACTA (0.54%). Helitrons are non-terminal inverted repetitive elements and comprise 2.3% of the genome.

Annotation of the Chia Pinta Genome

We annotated the Chia pinta genome for protein-coding genes resulting in 59,062 working gene models corresponding to 41,279 loci (Table 2). Working gene models had an average transcript length of 3.1 kbp, coding sequence (CDS) length of 1,217 bp, exon length of 279 bp and intron length of 240 bp. Working gene models exhibited an average of 5.8 exons, with 13.6% of transcripts being single-exon genes. The high confidence model set, a subset of the working set which have expression and/or protein evidence, contains 53,053 gene models representing 35,480 loci (Table 2). The high confidence set has an average transcript length of 3.3 kbp, exon length of 226 bp, intron length of 244 bp, and 6.1 exons per model; 6,105 gene models are single exon models. We selected the longest model as a representative for each gene locus from the working and high confidence model sets. With respect to BUSCO representation, the high confidence representative models are 94.8% complete, of which, 84.8% are complete and single copy while 10% are complete and duplicated; 1.9% are fragmented and 3.3% are missing. For the working representative models, 95.7% are complete with 85.5% complete and single copy and 10.2% complete and duplicated; 1.7% fragmented and 2.6% missing. Overall, the BUSCO results indicate a robust annotation of the Chia pinta genome.

Comparative Analyses of Chia Genome Assemblies

There are currently three published long-read, chromosome-scale Chia genome assemblies: Chia blanca (Li et al., 2023), Chia negra (Wang et al., 2022), and Mexican Chia (Alejo-Jacuinde et al., 2023). BUSCO analysis of all three published Chia genomes revealed that all of these assemblies were high quality and with robust gene annotation datasets. Syntenic orthologs (syntelogs) were identified between all four assemblies revealing a high degree of synteny between these genome assemblies (Figure 1) with limited disruptions that may be due to assembly artifacts in the various genome assemblies. Due to the high degree of similarity between the four Chia genomes, we performed detailed comparisons of our Chia pinta genome to the chromosome-scale black seeded Chia negra in which 73.62% of the genes were colinear within 1,178 syntenic blocks (Figure 1). Chia negra is a 344Mb genome assembly with 99.05% anchored on to chromosomes and 3.3Mb unanchored (Wang et al., 2022) with 428 gaps, amounting to a total of 191.2 kbp Ns. A total of 1,278,367 Single Nucleotide Polymorphisms (SNPs) were identified between the Chia negra and Chia pinta genomes that were distributed throughout the genome with 10.0% (127,210) residing in genic regions, 75.6% (967,385) in intergenic regions, and 14.4% (184,772) within intronic regions of the Chia pinta genome.

Using Orthofinder with the predicted proteomes of both Chia pinta and Chia negra, we identified 20,580 orthogroups, of which, 358 orthogroups (2,738 genes) were unique to Chia pinta while 462 orthogroups (1,458 genes) were unique to Chia negra. Gene ontology (GO) enrichment of the genes unique to Chia pinta revealed differences in certain biological process, cellular components, and molecular function ontologies. Of particular interest was the enrichment of the GO terms “defense response”, and “diterpenoid biosynthetic process” with 45 terpene synthases identified in the GO terms “diterpenoid biosynthetic process” and “terpene synthase activity”.

BLASTP was used to search all representative proteins in Chia pinta and Chia negra against a collection of known terpene synthases (TPSs). TPSs greater than 350 amino acids were used to create a phylogeny to determine the relationships among the TPSs. After filtering, a total of 111 TPSs in Chia pinta and 53 in Chia negra were identified. To confirm that this is not due to annotation errors, Chia pinta TPS transcript sequences were used in a BLASTN search against the Chia negra genome; no additional terpene synthases were identified in Chia negra indicating these sequences are absent in the Chia negra genome assembly. A phylogeny was constructed with putative TPS protein sequences from Chia pinta, Chia negra, and functionally characterized TPSs to assign Chia TPSs to closest known functionally characterized TPSs. Despite GO enrichment annotation of ’diterpenoid biosynthetic process’, most enriched TPSs are within the TPS-a and to a lesser degree TPS-b subfamilies which produce sesqui- and monoterpenes, indicating an expansion of volatile terpenes. The discrepancy on the GO terms claiming diterpenoid processes yet finding sesqui- and monoterpene synthases can be explained by GO enrichment often misannotated TPSs as diTPSs.

The TPS-a subfamily contains 56 putative TPSs in Chia pinta and only four in Chia negra. Of the 56 putative Chia pinta TPSs, 38 were found to enriched relative to Chia negra. The enriched TPSs reside in clades that do not contain a Chia negra TPS. To further understand the genomic context of the enriched TPSs, biosynthetic gene clusters (BGCs) membership and synteny were used. There are 16 BGCs containing TPSs in Chia pinta present on chromosomes 1, 2, 3, 4, and 6. Notably, six of these BGCs contain 23 out of the 56 Chia pinta specific TPS-a subfamily genes (Figure 2). This coincides with the expansion of the TPS-a subfamily in Chia pinta. All Chia pinta enriched TPS-a BGCs contain syntenic genes between Chia pinta, Chia negra, and *S. miltiorrhiza* (Figure 2). However, Chia pinta only shares one syntenic TPS with Chia negra and three syntenic TPSs with *S. miltiorrhiza.* Many of the TPSs present in Chia pinta’s BGCs appear to be tandem duplications, most notably in the teal and green BGCs (Figure 2)**.** However, some of the TPSs present in the green BGC are less than 350 amino acids indicating they may be truncated.

The origin and expansions of TPS-a genes were examined through synteny with *S. miltiorrhiza*. Two separate BGCs, purple and orange, contain paralogous TPSs yet are in distinct syntenic blocks (Figure 2). Work in *S. miltiorrhiza* characterized orthologs of these genes (89% identity) as (-)-5-epi-eremophilene synthases in which three TPSs (*SmSTPS1*, *SmSTPS2*, and

SmSTPS3) had differential gene expression yet identical biochemical activity (Fang et al., 2017). The purple BGC contains one TPS that is a syntelog of *SmSTPS1*, but there are no syntelogs of *SmSTPS2* or *SmSTPS3* (Figure 2) suggesting that a single gene was maintained and was tandemly duplicated or that structural rearrangements occurred disrupting synteny with *SmSTPS2* or *SmSTPS3*. The orange BGC contains TPSs that are equally related to *SmSTPS1* but are not syntenic with the *S. miltiorrhiza* SmSTPS cluster. Instead, the homologs have moved into a different syntenic block entirely. Additionally, there is a notable difference in gene expression profiles of the purple and orange BGCs with the orange BGC largely expressed in the leaf and stem whereas the purple clade has its highest expression in roots amongst the different paralogs (Figure 2). This may exemplify how a BGC can evolve by duplication and subfunctionalization resulting in distinct spatial gene expression patterns. The teal and yellow BGCs indicate that there are no syntenic TPSs in *S. miltiorrhiza*. The minor enrichment in TPS-b genes present in Chia pinta is largely due to expansion of a single clade. The closest functionally characterized enzyme to this expanded clade was and (–)-exo-α-bergamotene synthase, having between 62-67% identity for this clade.

Finding such a large difference in TPS-a abundance and identifying many of them within BGCs between Chia pinta and Chia negra further supports the diversity that exists not just within the *Salvia* genus, but even within Chia accessions. One potential source of the TPS expansion could be due to sequencing gaps in the Chia negra genome assembly. Specially, there are gaps in the purple BGC region of the Chia negra genome sequence. Therefore, these TPSs could be present within the species, but were not captured by the genome assembly. However, for the remaining five BGCs there are no assembly gaps in the Chia negra genome assembly and when the predicted transcripts for the TPSs were searched against the Chia negra genome, there were no hits for these regions. To determine if the TPSs are unique to Chia pinta, we examined the BGCs for syntelogs in the two other long-read Chia genome assemblies. The teal, orange, pink, green, and yellow BGCs contain syntelogs in Chia pinta, Chia blanca, and Mexican Chia whereas the purple BGC contains only syntelogs between Chia pinta and Mexican Chia. Thus, diversity in TPSs is present between Chia accessions suggesting variation in terpenoid profiles that may be associated with local adaptation.

Lamiaceae Phylogeny and Gene Family Expansions

To determine the evolutionary relationships of Lamiaceae species with Chia pinta, a species phylogeny was constructed using high-quality available genome sequences from 23 species from seven tribes in the Lamiaceae (Figure 3). Using the multiple sequence alignment option in Orthofinder, 923,746 genes were assigned to orthogroups. As shown in Figure 3, the Nepetoideae tribe is sister to Ajugoideae, Lamiodeae, and Scutellariodeae, the Callicarpoideae and Tectonoideae are sister to all other species, and the Premnoideae is sister to all other subfamilies. The relationships between the tribes in this genome-derived tree differs from a published phylogeny derived from 520 single copy transcripts (Godden et al., 2019) in which the Nepetoideae is sister to Ajugoideae, Lamiodeae, Scutellariodeae, Premnoideae, and Tectonoideae. The topology difference between these two phylogenetic estimates could be due to a combination of species sampling and data quality differences.

Gene expansions and contractions of single copy orthologs throughout the Lamiaceae were identified using CAFE (Figure 3A) and placed on the species tree phylogeny revealing large expansions and contractions throughout the Lamiaceae. The node branching of the Nepetoideae indicates a gene family expansion of 1,506 genes and contraction of 1,688 genes. The branch point from *S. hispanica* and *Salvia splendens* reveals 2,901 gene expansions and 12,295 gene contractions indicating substantial differences within the *Salvia* genus.

Synteny between genomes serve as a tool for examining evolution reflecting ancestral conservation of gene order. Using Chia pinta as the reference genome, we examined synteny within 11 chromosome-scale assemblies, spanning six tribes of the Lamiaceae family, revealing extensive conservation among the genomes (Figure 3B). In total, 182 Chia pinta genes were found to have a one-to-one syntenic relationship across all 11 species.

The polyphyletic nature of *Salvia* is highlighted by orthogroup membership. Of the 39,379 orthogroups containing 211,888 genes there were 12,987 orthogroups, containing 165,520 genes, in common among all *Salvia* (Figure 4A). The next highest number of orthogroups are unique to *S. rosmarinus* closely followed by *S. officinalis* and then *S. splendens* (Figure 4A). We also performed syntenic analyses between the genomes of four *Salvia* species to further our understanding of the species relationship in this polyphyletic genus. As expected, Chia pinta shares extensive synteny with other *Salvia* species (Figure 4b). *S. splendens* is reported to be a tetraploid (Jia et al., 2021). Based on orthogroup membership, 25% (4,684) of orthogroups shared by *S. splendens* and Chia pinta contain two *S. splendens* genes for each Chia pinta gene. This pattern reflects that *S. splendens* is a tetraploid and Chia pinta is a diploid. There are also two syntenic blocks in *S. splendens* for each block within Chia pinta, the syntenic blocks exist across four chromosomes in *S. splendens* (Figure 4b and 4c). It has been reported that there is a single shared whole genome duplication between Chia pinta and *S. splendens* and an additional duplication just in *S. splendens* (Jia et al., 2021; Wang et al., 2022). Therefore, the four unique chromosomes syntenic to a single chromosome in Chia pinta could be due to chromosomal fusions in Chia pinta or chromosomal fissions in *S. splendens.* Within the *Salvia* genus there are large regions of fragmented synteny between Chia pinta and *S. officinalis* as well as between *S. splendens* and *S. officinalis.* The fragmentation could be present due to different ancestry of Chia pinta and *S. officinalis*. As *Salvia* is a polyphyletic genus (Lara-Cabrera et al., 2021), this could be indicative of how distantly related these two species are. An alternative hypothesis is that they share a common ancestor, but the divergence time between species is so long that conserved genetic regions have been differentially fractionated (i.e. unique gene loss patterns). This is consistent with the large gene family expansions and contractions in the node that splits the *Salvia* species.

Population Structure of Chia

Seed coat color is a frequent descriptor for Chia accessions with Chia white seeded blanca varieties while Chia negra, Chia cualac, and Chia xonostli are predominately black-seeded (Figure 5a). Chia pinta seeds are a mix of both black and white seeds (Figure 5a). A diversity panel of 19 Chia accessions including wild and cultivated accessions along with two *S. tiliifolia* accessions with origins throughout Mexico was constructed and sequenced to reveal genetic diversity among accessions and provide insight into population structure of cultivated and wild Chia varieties. The percentage of reads aligned to the Chia pinta genome ranged from 95.5% to 97.7% for the *S. tiliifolia* samples and 96.3%-98.5% for the Chia varieties suggesting that the two species share substantial sequence similarity. Population structures were inferred with admixture with k=2 to k =13. Population structure admixture results suggested through the cross-validation error plot that there are two possible number of populations: four and nine as the local minima being at four and the global minima at nine in the cross-validation error plot.

Using a k=4, broad population groups are present that can be assigned to known categories of Chia: Chia pinta (yellow), *S. tiliifolia* (purple), Chia negra and Chia Xonostli (blue), and Chia Cualac (pink). The population structure indicates that the phenotypic and origin grouping reflects the genetic structure of the population. Chia pinta accessions are domesticated Chia varieties whereas Chia negra and Chia Xonostli are classified as wild due to their more open calyx and other wild traits. Chia negra is in the same population group with the less widely known Chia Xonostli which is similar to Chia negra yet categorized differently due to its domesticated traits. Historically, Chia Xonostli was found in the states of Jalisco, Guanajuato, Veracruz, and Hidalgo. Chia Cualac was reported to be semi-domesticated and forms their own group with some admixture from Chia Xonostli (Peláez et al., 2019). This follows the hypothesis that wild introgressions are present throughout the populations. One *S. tiliifolia* accession is admixed with Chia pinta. *S. tiliifolia* is nearly indistinguishable from Chia and is known to grow in the same areas as Chia pinta; thus, it is possible that these species hybridize and form a population of *S. tiliifolia* that is highly admixed with Chia pinta. Feral hybrid accessions could continue to evolve through hybridization with domesticated Chia yielding the admixture present within one accession of *S. tiliifolia* (Figure 5).

CONCLUSIONS

In this study, a high-quality chromosome-scale genome assembly of Chia pinta was generated that allowed for additional genomic comparisons within the economically important crop including three other long-read, chromosome-scale Chia assemblies that showed extensive synteny among the genome sequences. Comparative genomic tools were used to determine differences within Chia accessions and throughout the Lamiaceae. Interestingly, Chia pinta was enriched in TPSs and contains novel TPSs compared to the Chia negra with some TPSs located within BGCs and syntenic with *S. miltiorrhiza*. Further examination of TPSs within BGCs among the four Chia genome assemblies revealed further diversification suggestive of variation in terpenoid biosynthesis among varieties. Through sequencing of a diversity panel, the population structure of Chia revealed introgression with other *Salvia* species.

MATERIALS AND METHODS

Plant materials

Different Chia varieties were collected throughout Mexico. Plants were grown in an experimental field in Celaya, Guanajuato, Mexico (20.578°, −100.822°) at the Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (INIFAP).

Nucleic acid isolation, library construction, and sequencing

For construction of a reference genome, DNA was isolated from medium-sized leaves from a mature plant (13.5 weeks old) of accession SM_ACJ2017 using a modified protocol from Doyle and Doyle (1987) and Healey et al. (2014). Large insert (>15kb, >20kb) PacBio libraries were made with the SMRTbell™ Template Prep Kit and sequenced on the PacBio Sequel platform at the University of Georgia, Georgia Genomics and Bioinformatics Core (GGBC, UG Athens, GA, RRID:SCR_010994). A whole genome shotgun library for reference error correction was prepared using the Illumina TruSeq Nano DNA Library Preparation Kit and sequenced in paired-end mode, 150 nt in length on a HiSeq 4000 at the Michigan State University Research Technology Support Facility (RTSF). Whole genome shotgun libraries for use in error correction and diversity panel variant analyses were constructed as described previously in Hardigan *et al.* 2016 (Hardigan et al., 2016) and sequenced at the Michigan State University RTSF in paired-end mode on a HiSeq4000 generating 150 nt reads. RNA was isolated from three biological replicates from a core set of tissues (leaf, inflorescence, lateral stem, secondary root) from the reference accession SM_ACJ2017 as described previously in Peláez *et al.* 2019 (Peláez et al., 2019). RNA-seq libraries were prepared using the Illumina TruSeq Stranded mRNA Library Preparation Kit and sequenced on an Illumina HiSeq 4000 generating 150 nt paired end reads for one replicate and 50 nt single end reads for the other two replicates; library preparation and sequencing were performed at the Michigan State University Research Technology Support

Facility (RTSF). A Phase Genomics Proximo Hi-C library was prepared from Chia pinta leaf tissue and sequenced by Phase Genomics (Seattle, WA) on the NextSeq 500 generating paired end 150 nt reads.

Chia pinta genome assembly

PacBio reads greater than 10 kbp (1.2 million reads, 21.6 Gb) were used to generate the initial assembly using Canu (v1.7; Koren et al., 2017) with a corrected ErrorRate of 0.15%. The initial assembly was polished with the raw PacBio reads using Arrow in the SMRT Analysis package (v5.0.1.9585; Pacfici Biosciences), followed by three rounds of error correction with 56 million Illumina WGS reads (150 nt paired-end WGS reads, 45X coverage) using Pilon (v1.22; Walker et al., 2014). Potential haplotigs were purged using purgeHaplotigs (v1.0.4; Roach et al., 2018) with the "maximum match score (-m)" of 500% and "-a = 50% ". Contigs were scaffolded to a chromosome scale assembly using Hi-C reads and Proximo pipeline with an input chromosome number of six by Phase Genomics (Bickhart et al., 2017). Scaffolded contigs were visualized with Juicebox (v1.9.8; Durand et al., 2016).

Genome annotation

A custom repeat library (CRL) was generated using RepeatModeler (v2.0.1;Flynn et al., 2020) and protein coding genes were removed from the CRL using ProteinExcluder (v1.2; Campbell et al., 2014). The Viridiplantae RepBase repeats (v20150807) were then added to create the final CRL. The genome assembly was hard and soft masked using RepeatMasker (v4.1.0; Smit et al.) with the CRL with the parameters: -s -nolow -no_is. RNA-seq libraries were cleaned using Cutadapt (v2.9; Martin, 2011) (--times 2 --minimum-length 100 --quality-cutoff 10) and then aligned to the genome assembly with HISAT2 (v2.2.0;Kim et al., 2019) (--max-intronlen 5000 --rna-strandness RF –dta –no-unal). The RNA-seq alignments were then assembled into transcript assemblies using Stringtie (v2.1.1; Kovaka et al., 2019).

Ab initio gene models were predicted on the soft-masked genome assembly using the BRAKER2 pipeline (v2.1.5; Brůna et al., 2021) using the leaf RNA-seq library CHI_AA as a source for hints. The *ab initio* gene models were then refined using PASA2 (v2.4.1; Campbell et al., 2006) with the RNA-seq transcript assemblies as a source of transcript evidence to produce the working gene model set. High confidence gene models were selected from the working gene model set by first calculating working gene model abundances of the RNA-seq libraries for the working gene models with Kallisto (v0.46.0; Bray et al., 2016), then searching the working gene models against PFAM (v32.0; Mistry et al., 2021) with HMMER (v3.2.1; Mistry et al., 2013). Working gene models with a TPM >1 in at least one RNA-seq library or a non-transposable element related PFAM domain match and no partial or containing an internal stop codon were identified as high confidence gene models. Functional annotation was assigned to the working gene model by searching the protein sequences against the Arabidopsis proteome (TAIR10), PFAM (v32.0; Mistry et al., 2021) and the Swiss-Prot plant proteins (release 2015_08). Search results were processed in the same order and the function of the first hit encountered was assigned to the gene model. Repetitive elements were identified using EDTA (v2.1.0; Ou et al., 2019) with the parameters species set to “others” and step set to “all”.

Genome quality assessment

Quality assessment of the genome assembly was performed by aligning WGS reads cleaned for low quality bases and adaptors using Cutadapt (v3.4; Martin, 2011) to the final assembly using BWA-mem (v0.7.16a; Li, 2013). Assemblathon.pl (https://github.com/KorfLab/Assemblathon/blob/master/assemblathon_stats.pl) was used to generate genome metrics. BUSCO (v3.1.0.Py3; Simão et al., 2015) embryophyta_odb10 was used to determine genic representation in the final assembly. Jellyfish (v.2.3.0; Marçais & Kingsford, 2011) with the option -m 21 was used to count kmers that were then visualized in GenomeScope (v2.0; Ranallo-Benavidez et al., 2020) with kmer length 21 was used to verify genome size and heterozygosity from the WGS reads from Chia pinta (CHI_AN). The Kmer Analysis Toolkit (v2.4.1; Mapleson et al., 2017) was used to examine the assembly for retained haplotigs. Synteny between the chia genome assemblies (Wang et al., 2022; Alejo-Jacuinde et al., 2023; Li et al., 2023) was analyzed using GENESPACE (v.1.1.10;Lovell et al., 2022). Syntenic comparison between Chia pinta and Chia negra was also performed using MCScanX (Wang et al., 2012).

Lamiaceae phylogeny and comparative analysis

Publicly available genomes of *Callicarpa armericana* (Hamilton et al., 2020), *Cleorodendrum inerme* (He et al., 2022) , *Hyssopus officinalis* (Lichman et al., 2020), *Nepeta cataria* (Lichman et al., 2020), *Nepeta mussinii* (Lichman et al., 2020), *Ocimum basilicum* (Bornowski et al., 2020), *Origanum majorana* (Bornowski et al., 2020), *Origanum vulgare* (Bornowski et al., 2020), *Perilla frustescens*(Zhang et al., 2021; Tamura et al., 2022), *Pogostemon cablin* (Shen et al., 2022), *Salvia miltiorrhiza* (Pan et al., 2023), *Salvia officinalis* (Li et al., 2022), *Salvia rosmarinus* (Bornowski et al., 2020), *Salvia splendens* (Jia et al., 2021), *Scutellaria baicalensis* (Zhao et al., 2019b), *Scutellaria barbata* (Xu et al., 2020), *Tectona grandis* (Zhao et al., 2019a), *Thymus quinquecostatus* (Sun et al., 2022)*, Lavandula angustifolia* (Hamilton et al., 2023) and *Premna obstusifolia* (He et al., 2022) were obtained and quality assessed using BUSCO (v5.5.0; Simão et al., 2015) embryophyta_odb10. Species with genome assembly complete BUSCO scores greater than 90% and annotation complete BUSCO scores greater than 80% were used in further comparative analysis. Orthogonal genes and species tree phylogeny were built using OrthoFinder (v.2.5.4; Emms & Kelly, 2019) with options -M msa -T raxml. The species tree output was covered into an ultrametric tree using the make_ultrametric command in

OrthoFinder (v.2.5.4; Emms & Kelly, 2019). Branch lengths were rescaled using the *Premna obstusifolia* divergence date of 16.06 MYA retrieved from the TimeTree of Life resource (Kumar et al., 2022). Gene family expansions and contractions were identified using CAFE (v.4.2.1; Han et al., 2013) with the following scripts with default parameters: cafetutorial_report_analysis.py and cafetutorial_draw_tree.py. Syntelogs through the Lamiaceae were obtained for the chromosome scale assemblies within the Lamiaceae and visualized using GENESPACE (v.1.1.10; Lovell et al., 2022).

Gene ontology term enrichment

Gene ontology (GO) terms were assigned to high confidence Chia pinta genes using InterProScan (v5.63-95.0; Jones et al., 2014). GO descriptions were added using the ontologyIndex package (Greene et al., 2017) and enrichment was calculated using the topGO R package (Alexa & Rahnenfuhrer, 2010). GO terms with an FDR adjusted p-value < 0.05 were considered significant.

Terpene synthase identification

BGCs were identified in Chia pinta, Chia negra, and *S. miltiorrhiza* with PlantiSMASH (Kautsar et al., 2017). Enriched TPSs identified in the various BGCs were searched with NCBI BLAST the nonredundant protein database to identify the closest functionally characterized TPSs. To extract all TPSs from the genome, the high confidence representative protein models were blasted against a reference set of known TPSs enzymes representing TPSs across all subfamilies. The BLAST hits with an E-value 1E-5 or better were selected. These gene models were filtered to remove any sequences smaller than 350 amino acids to ensure a quality phylogeny and minimize pseudogenes. The final set of putative and reference TPS sequences were aligned using clustal omega (v1.2.4; Sievers et al., 2011). A phylogenetic tree of the alignment was built via RAXML (v8.2.12; Stamatakis, 2014) with the PROTGAMMA AUTO model, algorithm a, and 1000 bootstraps. Gene expression of terpene synthases was calculated using the single end RNA-seq libraries and Cufflinks (v.2.2.1; Roberts et al., 2011) with the options -b and -u to generate FPKM values for all Chia pinta genes. Orthologous genes from Chia pinta, Chia negra (Wang et al., 2022) , Chia blanca (Li et al., 2023)and the Mexican Chia variety (Alejo-Jacuinde et al., 2023) were identified using OrthoFinder (v.2.5.4; Emms & Kelly, 2019) with options -M msa -T raxml.

Population structure analysis

Whole genome shotgun reads from the diversity panel were cleaned using Cutadapt (v3.4; Martin, 2011) and aligned to the Chia genome using BWA-mem (v0.7.16a; Li, 2013). PicardTools (v2.20.8; Picard toolkit, 2019) commands SortSam, MarkDuplicates, BuildBamIndex, and CollectAlignmentSummaryMetrics were used to sort, convert files, and generate alignment metrics. The GATK (v4.1.2.0; Van der Auwera & O'Connor, 2020) HaplotypeCaller with default parameters was used to call variants. GenomicsDBImport with default parameters was used to merge the varieties into a single VCF file and genotyped using GenotypeGVCFs.Separated. SNPs were selected using the SelectVariants command. Hard filtering of the SNPs was performed using the parameters QD < 2.0, QUAL < 30.0, SOR > 3.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, MQRankSum-12.5, ReadPosRankSum < -8.0. Additional filtering was performed using VCFTools (v0.1.16; Danecek et al., 2011) with filtering –freq2 and –max-alleles 2 to retain only bialleleic sites, minor allele frequency of 0.071, --max-missing 0.9, --minQ 30, --min-meanDP 15, --max-meanDP 39.

SNPs were called relative to the Chia negra reference genome (Wang et al., 2022) using nucmer from MUMmer (v4.0; Marçais et al., 2018) with the options –maxgap=2500, --minmatch=11, and --mincluster=25. SNPs were quality filtered using the delta-filter command in MUMmer with the -r flag. (v4.0; Marçais et al., 2018). The SNP set from the diversity panel and from the alignment of the two genome assemblies were combined and converted into bed format using PLINK 2.0 (v.alpha2.3; Purcell & Chang; Chang et al., 2015) resulting in 156,829 total SNPs. Population structure was inferred with Admixture (v.1.3.0; Alexander et al., 2009) and a SNP phylogenetic tree built with SNPhylo (v.20160204; Lee et al., 2014) using default parameters.

ACKNOWLEDGMENTS

Funds for this study were provided by a grant to C.R.B. from the National Science Foundation Plant Genome Research Program (IOS-1444499), the Georgia Research Alliance, Georgia Seed Development, and the University of Georgia. JB was supported by Michigan State University.

CONTRIBUTIONS

ACJ and CRB conceived of the study. SM-HGOR, PMM-P and ACP collected samples. PMM-P and ACP prepared materials. JB, JPH, NS, DZ, JCW and BV performed data analyses and drafted the manuscript. PPE, BH, ACJ, and CRB supervised and performed project administration. All authors approved of the manuscript.

DATA AVAILABILITY STATEMENT

The raw sequence reads are available in the National Center for Biotechnology Information Sequence Read Archive under BioProject PRJNA744892. The genome assembly, annotation, and large data sets (genome assembly, genome annotation) reported in this study are available in Figshare via 10.6084/m9.figshare.24546049.

CONFLICT OF INTERESTS

The authors declare no conflict of interests.

REFERENCES

Alejo-Jacuinde, G., Nájera-González, H.R., Chávez Montes, R.A., Gutierrez Reyes, C.D., Barragán-Rosillo, A.C., Perez Sanchez, B., Mechref, Y., López-Arredondo, D., Yong-Villalobos, L., & Herrera-Estrella, L. (2023). Multi-omic analyses reveal the unique properties of chia (Salvia hispanica) seed metabolism. *Communications Biology*, *6*. https://doi.org/10.1038/s42003-023-05192-4

Alexa, A., & Rahnenfuhrer, J. (2010). topGO: Enrichment Analysis for Gene Ontology

Alexander, D.H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109

Van der Auwera, G.A., & O'Connor, B. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. 1st Editio. O'Reilly Media.

Baldivia, A.S. (2018). A Historical Review of the Scientific and Common Nomenclature Associated with Chia: From *Salvia hispanica* to *Salvia mexicana* and Chian to Salba. *Agricultural Research & Technology: Open Access Journal*, *18*. https://doi.org/10.19080/artoaj.2018.18.556047

Bickhart, D.M., Rosen, B.D., Koren, S., Sayre, B.L., Hastie, A.R., Chan, S., Lee, J., Lam, E.T., Liachko, I., Sullivan, S.T., Burton, J.N., Huson, H.J., Nystrom, J.C., Kelley, C.M., Hutchison, J.L., Zhou, Y., Sun, J., Crisà, A., Ponce De León, F.A., Schwartz, J.C., Hammond, J.A., Waldbieser, G.C., Schroeder, S.G., Liu, G.E., Dunham, M.J., Shendure, J., Sonstegard, T.S., Phillippy, A.M., Van Tassell, C.P., & Smith, T.P.L. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, *49*, 643–650. https://doi.org/10.1038/ng.3802

Bornowski, N., Hamilton, J.P., Liao, P., Wood, J.C., Dudareva, N., & Buell, C.R. (2020). Genome sequencing of four culinary herbs reveals terpenoid genes underlying chemodiversity in the Nepetoideae. *DNA Research*, *27*, 1–12. https://doi.org/10.1093/dnares/dsaa016

Bray, N.L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*, 525–527. https://doi.org/10.1038/nbt.3519

Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, *3*, 1–11. https://doi.org/10.1093/nargab/lqaa108

Cahill, J.P. (2003). Ethnobotany of Chia, *Salvia hispanica* L. (Lamiaceae). *Economic Botany*, *57*, 604–618. https://doi.org/10.1663/0013-0001(2003)057[0604:EOCSHL]2.0.CO;2

Cahill, J.P. (2004). Genetic diversity among varieties of Chia (*Salvia hispanica* L.). *Genetic Resources and Crop Evolution*, *51*, 773–781. https://doi.org/10.1023/B:GRES.0000034583.20407.80

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M., & Robin, C.R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics*, *7*, 1–17. https://doi.org/10.1186/1471-2164-7-327

Campbell, M.S., Law, M.Y., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., Ware, D., Shiu, S.H., Childs, K.L., Sun, Y., Jiang, N., & Yandell, M. (2014). MAKER-P: A Tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiology*, *164*, 513–524. https://doi.org/10.1104/pp.113.230144

Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., & Lee, J.J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 1–16. https://doi.org/10.1186/s13742-015-0047-8

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dong, A.X., Xin, H.B., Li, Z.J., Liu, H., Sun, Y.Q., Nie, S., Zhao, Z.N., Cui, R.F., Zhang, R.G., Yun, Q.Z., Wang, X.N., Maghuly, F., Porth, I., Cong, R.C., & Mao, J.F. (2018). High-quality assembly of the reference genome for scarlet sage, *Salvia splendens,* an economically important ornamental plant. *GigaScience*, *7*, 1–10. https://doi.org/10.1093/gigascience/giy068

Doyle, J.J., & Doyle, J.L. (1987). A rapid DNA isolation procedure from small quantities of fresh leaf tissues.. *Pytochemical Bulletin*, *19*, 11–15

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*, 99–101. https://doi.org/10.1016/j.cels.2015.07.012

Emms, D.M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*, 238. https://doi.org/10.1186/s13059-019-1832-y

Fang, X., Li, C.Y., Yang, Y., Cui, M.Y., Chen, X.Y., & Yang, L. (2017). Identification of a novel (-)-5-epieremophilene synthase from salvia miltiorrhiza via transcriptome mining. *Frontiers in Plant Science*, *8*, 1–11. https://doi.org/10.3389/fpls.2017.00627

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 of the United States of America*, *117*, 9451–9457. https://doi.org/10.1073/pnas.1921046117

Godden, G.T., Kinser, T.J., Soltis, P.S., Soltis, D.E., & Chaw, S.M. (2019). Phylotranscriptomic Analyses Reveal Asymmetrical Gene Duplication Dynamics and Signatures of Ancient Polyploidy in Mints. *Genome Biology and Evolution*, *11*, 3393–3408. https://doi.org/10.1093/GBE/EVZ239

Greene, D., Richardson, S., & Turro, E. (2017). OntologyX: A suite of R packages for working with ontological data. *Bioinformatics*, *33*, 1104–1106. https://doi.org/10.1093/bioinformatics/btw763

Hamilton, J.P., Godden, G.T., Lanier, E., Bhat, W.W., Kinser, T.J., Vaillancourt, B., Wang, H., Wood, J.C., Jiang, J., Soltis, P.S., Soltis, D.E., Hamberger, B., & Robin Buell, C. (2020). Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, Callicarpa americana. *GigaScience*, *9*, 1–11. https://doi.org/10.1093/gigascience/giaa093

Hamilton, J.P., Vaillancourt, B., Wood, J.C., Wang, H., Jiang, J., Soltis, D.E., Buell, C.R., & Soltis, P.S (2023). Chromosome-scale genome assembly of the 'Munstead' cultivar of Lavandula angustifolia. *BMC Genomic Data*, *24*, 75

Han, M. V., Thomas, G.W.C., Lugo-Martinez, J., & Hahn, M.W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution*, *30*, 1987–1997. https://doi.org/10.1093/molbev/mst100

Hardigan, M.A., Crisovan, E., Hamilton, J.P., Kim, J., Laimbeer, P., Leisner, C.P., Manrique-Carpintero, N.C., Newton, L., Pham, G.M., Vaillancourt, B., Yang, X., Zeng, Z., Douches, D.S., Jiang, J., Veilleux, R.E., & Buella, C.R. (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. *Plant Cell*, *28*, 388–405. https://doi.org/10.1105/tpc.15.00538

He, Z., Feng, X., Chen, Q., Li, L., Li, S., Han, K., Guo, Z., Wang, J., Liu, M., Shi, C., Xu, S., Shao, S., Liu, X., Mao, X., Xie, W., Wang, X., Zhang, R., Li, G., Wu, W., Zheng, Z., Zhong, C., Duke, N.C., Boufford, D.E., Fan, G., Wu, C.I., Ricklefs, R.E., & Shi, S. (2022). Evolution of coastal forests based on a full set of mangrove genomes. *Nature Ecology and Evolution*, *6*, 738–749. https://doi.org/10.1038/s41559-022-01744-9

Healey, A., Furtado, A., Cooper, T., & Henry, R.J. (2014). Protocol: A simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods*, *10*, 1–8. https://doi.org/10.1186/1746-4811-10-21/COMMENTS

Hu, G.X., Takano, A., Drew, B.T., Liu, E. De, Soltis, D.E., Soltis, P.S., Peng, H., & Xiang, C.L. (2018). Phylogeny and staminal evolution of *Salvia* (Lamiaceae, Nepetoideae) in East Asia. *Annals of Botany*, *122*, 649–668. https://doi.org/10.1093/AOB/MCY104

Jia, K.H., Liu, H., Zhang, R.G., Xu, J., Zhou, S.S., Jiao, S.Q., Yan, X.M., Tian, X.C., Shi, T. Le, Luo, H., Li, Z.C., Bao, Y.T., Nie, S., Guo, J.F., Porth, I., El-Kassaby, Y.A., Wang, X.R., Chen, C., Van de Peer, Y., Zhao, W., & Mao, J.F. (2021). Chromosome-scale assembly and evolution of the tetraploid Salvia splendens (Lamiaceae) genome. *Horticulture Research*, *8*. https://doi.org/10.1038/s41438-021-00614-y

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kautsar, S.A., Suarez Duran, H.G., Blin, K., Osbourn, A., & Medema, M.H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research*, *45*, W55–W63

Kim, D., Paggi, J.M., Park, C., Bennett, C., & Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*, 907–915. https://doi.org/10.1038/s41587-019-0201-4

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

Kovaka, S., Zimin, A. V., Pertea, G.M., Razaghi, R., Salzberg, S.L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*, 1–13. https://doi.org/10.1186/s13059-019-1910-1

Kumar, S., Suleski, M., Craig, J.M., Kasprowicz, A.E., Sanderford, M., Li, M., Stecher, G., & Hedges, S.B. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, *39*, 1–6. https://doi.org/10.1093/molbev/msac174

Lara-Cabrera, S.I., Perez-Garcia, M. de la L., Maya-Lastra, C.A., Montero-Castro, J.C., Godden, G.T., Cibrian-Jaramillo, A., Fisher, A.E., & Porter, J.M. (2021). Phylogenomics of Salvia L. subgenus Calosphace (Lamiaceae). *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.725900

Lee, T.H., Guo, H., Wang, X., Kim, C., & Paterson, A.H. (2014). SNPhylo: A pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics*, *15*, 1–6. https://doi.org/10.1186/1471-2164-15-162

Li, C.Y., Yang, L., Liu, Y., Xu, Z.G., Gao, J., Huang, Y.B., Xu, J.J., Fan, H., Kong, Y., Wei, Y.K., Hu, W.L., Wang, L.J., Zhao, Q., Hu, Y.H., Zhang, Y.J., Martin, C., & Chen, X.Y. (2022). The sage genome provides insight into the evolutionary dynamics of diterpene biosynthesis gene cluster in plants. *Cell Reports*, *40*, 111236. https://doi.org/10.1016/j.celrep.2022.111236

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

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*. https://doi.org/10.3389/fpls.2023.1102715

Lichman, B.R., Godden, G.T., Hamilton, J.P., Palmer, L., Kamileen, M.O., Zhao, D., Vaillancourt, B., Wood, J.C., Sun, M., Kinser, T.J., Henry, L.K., Rodriguez-Lopez, C., Dudareva, N., Soltis, D.E., Soltis, P.S., Robin Buell, C., & O'Connor, S.E. (2020). The evolutionary origins of the cat attractant nepetalactone in catnip. *Science Advances*, *6*, 1–14. https://doi.org/10.1126/sciadv.aba0721

Lovell, J.T., Sreedasyam, A., Schranz, M.E., Wilson, M., Carlson, J.W., Harkess, A., Emms, D., Goodstein, D.M., & Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife*, *11*, 1–20. https://doi.org/10.7554/ELIFE.78526

Mapleson, D., Accinelli, G.G., Kettleborough, G., Wright, J., & Clavijo, B.J. (2017). KAT: A K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, *33*, 574–576. https://doi.org/10.1093/bioinformatics/btw663

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*, 1–14. https://doi.org/10.1371/journal.pcbi.1005944

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

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

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913

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 Res*, *41*, e121. https://doi.org/10.1093/nar/gkt263

Muñoz, L.A., Cobos, A., Diaz, O., & Aguilera, J.M. (2013). Chia Seed (*Salvia hispanica*): An Ancient Grain and a New Functional Food. *Food Reviews International*, *29*, 394–408. https://doi.org/10.1080/87559129.2013.818014

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., Jiang, N., Hirsch, C.N., & Hufford, M.B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, *20*, 1–18. https://doi.org/10.1186/s13059-019-1905-y

Pacfici Biosciences. SMRT tools.

Pan, X., Chang, Y., Li, C., Qiu, X., Cui, X., Meng, F., Zhang, S., Li, X., & Lu, S. (2023). Chromosome-level genome assembly of Salvia miltiorrhiza with orange roots uncovers the role of Sm2OGD3 in catalyzing 15,16-dehydrogenation of tanshinones. *Horticulture Research*, *10*. https://doi.org/10.1093/hr/uhad069

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*, 1–11. https://doi.org/10.1038/s41598-019-45895-5

Picard toolkit. (2019). . *Broad Institute, GitHub repository*, Purcell, S., & Chang, C. PLINK 2.0 alpha 2. 3

Ranallo-Benavidez, T.R., Jaron, K.S., & Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*. https://doi.org/10.1038/s41467-020-14998-3

Roach, M.J., Schmidt, S.A., & Borneman, A.R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies.. *BMC bioinformatics*, *19*, 460. https://doi.org/10.1186/s12859-018-2485-7

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, *12*. https://doi.org/10.1186/gb-2011-12-3-r22

Shen, Y., Li, W., Zeng, Y., Li, Z., Chen, Y., Zhang, J., Zhao, H., Feng, L., Ma, D., Mo, X., Ouyang, P., Huang, L., Wang, Z., Jiao, Y., & Wang, H. bin. (2022). Chromosome-level and haplotype-resolved genome provides insight into the tetraploid hybrid origin of patchouli. *Nature Communications*, *13*, 1–15. https://doi.org/10.1038/s41467-022-31121-w

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., & Higgins, D.G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*. https://doi.org/10.1038/msb.2011.75

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*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smit, A., Hubley, R., & Green, P. *RepeatMasker Open-4.0*.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Sun, M., Zhang, Y., Zhu, L., Liu, N., Bai, H., Sun, G., Zhang, J., & Shi, L. (2022). Chromosome-level assembly and analysis of the *Thymus* genome provide insights into glandular secretory trichome formation and monoterpenoid biosynthesis in thyme. *Plant Communications*, *3*, 100413. https://doi.org/10.1016/j.xplc.2022.100413

Tamura, K., Sakamoto, M., Tanizawa, Y., Mochizuki, T., & Bono, H. (2022). Resource Article : Genomes Explored A highly contiguous genome assembly of red perilla (Perilla frutescens) domesticated in Japan 1–8

Valdivia-López, M.Á., & Tecante, A. (2015). Chia (*Salvia hispanica*): A Review of Native Mexican Seed and its Nutritional and Functional Properties. *Advances in Food and Nutrition Research* (pp. 53–75). Academic Press Inc.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., & Earl, A.M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE*, *9*, e112963. https://doi.org/10.1371/journal.pone.0112963

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*, 100326. https://doi.org/10.1016/j.xplc.2022.100326

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., Kissinger, J.C., & Paterson, A.H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*, *40*, e49. https://doi.org/10.1093/nar/gkr1293

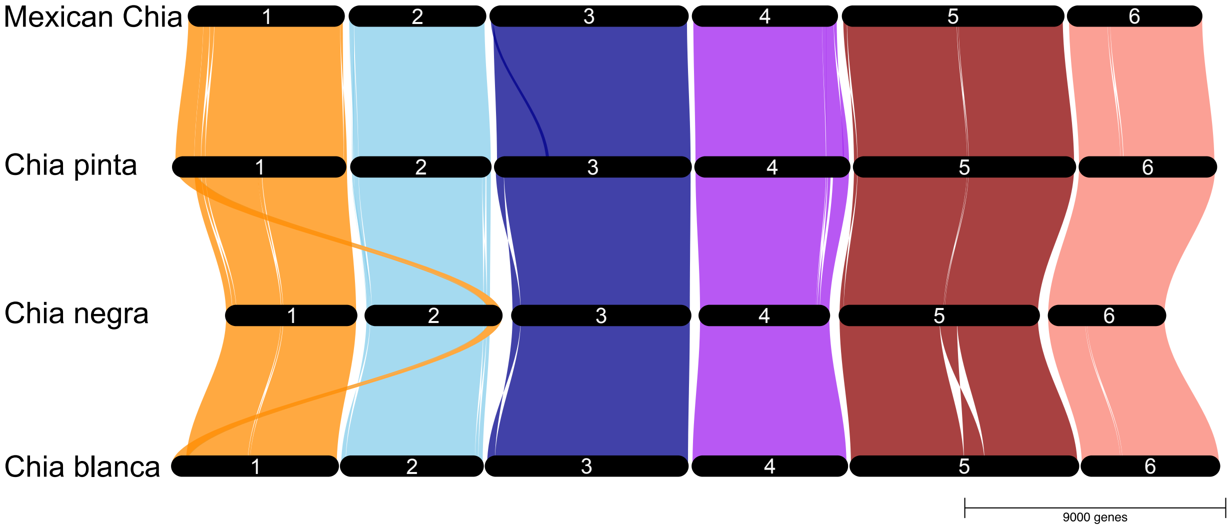
Xu, Z., Gao, R., Pu, X., Xu, R., Wang, J., Zheng, S., Zeng, Y., Chen, J., He, C., & Song, J. (2020). Comparative Genome Analysis of *Scutellaria baicalensis* and *Scutellaria barbata* Reveals the Evolution of Active Flavonoid Biosynthesis. *Genomics, Proteomics and Bioinformatics*, *18*, 230–240. https://doi.org/10.1016/j.gpb.2020.06.002

Zhang, Y., Shen, Q., Leng, L., Zhang, D., Chen, S., Shi, Y., Ning, Z., & Chen, S. (2021). Incipient diploidization of the medicinal plant Perilla within 10,000 years. *Nature Communications*, *12*, 1–13. https://doi.org/10.1038/s41467-021-25681-6

Zhao, D., Hamilton, J., Bhat, W., Johnson, S., Godden, G., Kinser, T., Boachon, B., Dudareva, D., Soltis, P., & Soltis, D. (2019)(a). A chromosomal-scale genome assembly of *Tectona grandis* reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways. *GigaScience*, *8*, giz005

Zhao, Q., Yang, J., Cui, M.Y., Liu, J., Fang, Y., Yan, M., Qiu, W., Shang, H., Xu, Z., Yidiresi, R., Weng, J.K., Pluskal, T., Vigouroux, M., Steuernagel, B., Wei, Y., Yang, L., Hu, Y., Chen, X.Y., & Martin, C. (2019)(b). The Reference Genome Sequence of *Scutellaria baicalensis* Provides Insights into the Evolution of Wogonin Biosynthesis. *Molecular Plant*, *12*, 935–950. https://doi.org/10.1016/j.molp.2019.04.002

774 **Figures and Tables**



775

776 **Figure 1. Synteny of the Chia genomes.** The top track is the Mexican Chia genome (Alejo-
777 Jacuinde et al., 2023), the second track is the Chia pinta genome reported in this study, the
778 third genome is Chia negra (Wang et al., 2022), and the bottom track represents the Chia
779 blanca genome (Li et al., 2023). The ribbons indicate syntenic blocks between the genomes
780 identified using GENESPACE (v.1.1.10;Lovell et al., 2022).

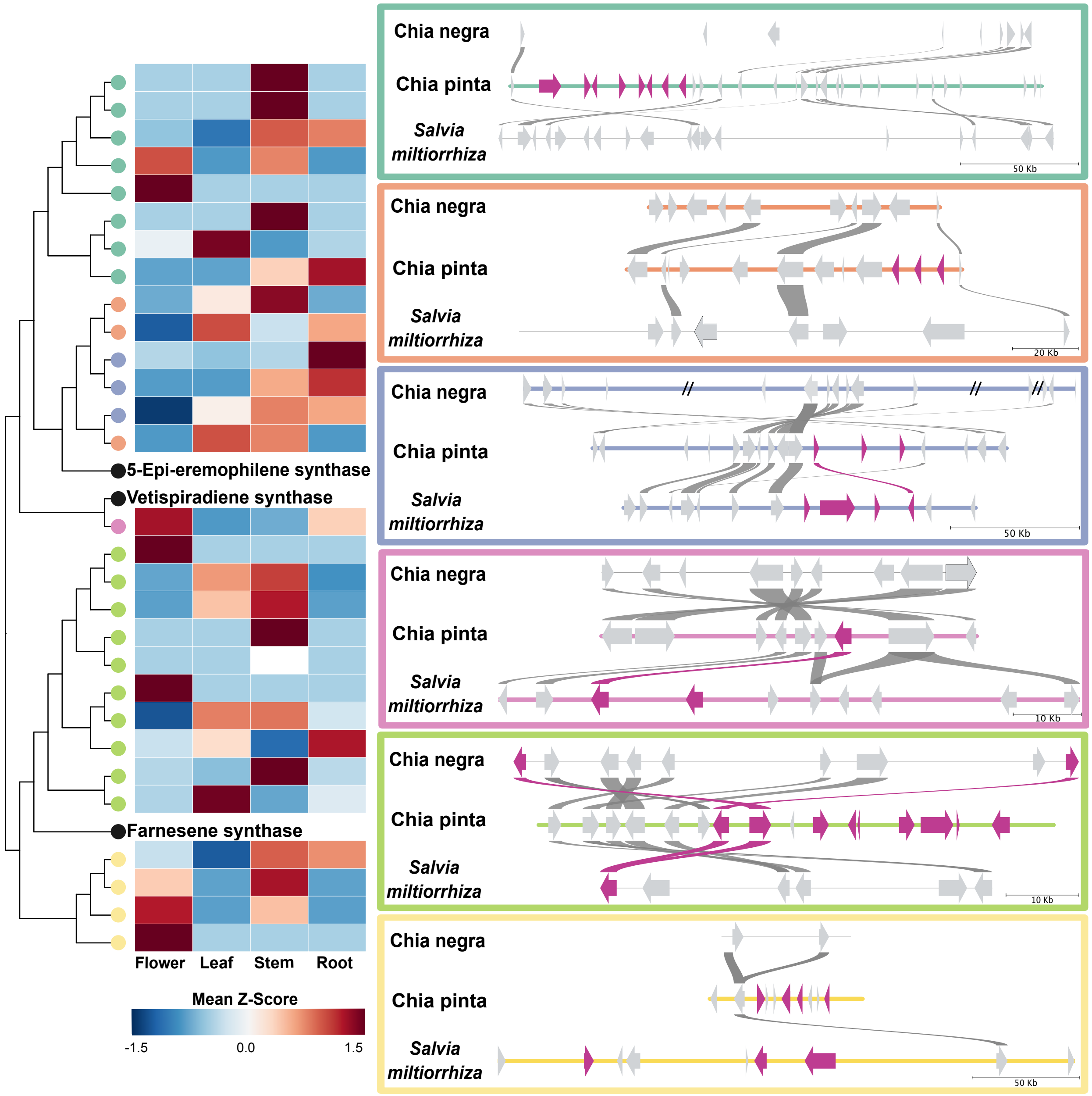


Figure 2. Chia pinta TPS-a Biosynthetic Gene Cluster Expression and Synteny. A phylogeny of the Chia pinta terpene synthase (TPS-a) genes present in biosynthetic gene clusters (BGCs) with representative functionally characterized reference TPSs is shown. The Chia pinta phylogeny was generated using RAxML (v8.2.12; Stamatakis, 2014). The heatmap of gene expression was constructed from flower, leaf, stem, and root tissue using expression values generated by Cufflinks (v.2.2.1; Roberts et al., 2011) with z-scores range from -1.5 to 1.5. Chia pinta genes (circles on the phylogeny) are colored by their respective BGC and correspond to the outlined syntenic BGCs; genes in black are known TPS. Biosynthetic gene clusters (BGCs) were identified by PlantiSmash (Kautsar et al., 2017) with boxes colored to match the clades in the phylogeny. Syntenic regions were determined using MCScanX (Wang et al., 2012) between Chia pinta, Chia

negra, and *S. miltorrhiza*. Synteny is indicated as lines between the genes (arrows). The color of the gene and syntenic line is determined by the presumed identity assigned by PlantiSmash where hot pink indicate TPSs; slashes through the line indicate gaps in the assembly. Grey genome lines indicate that it is not a TPS BGC.

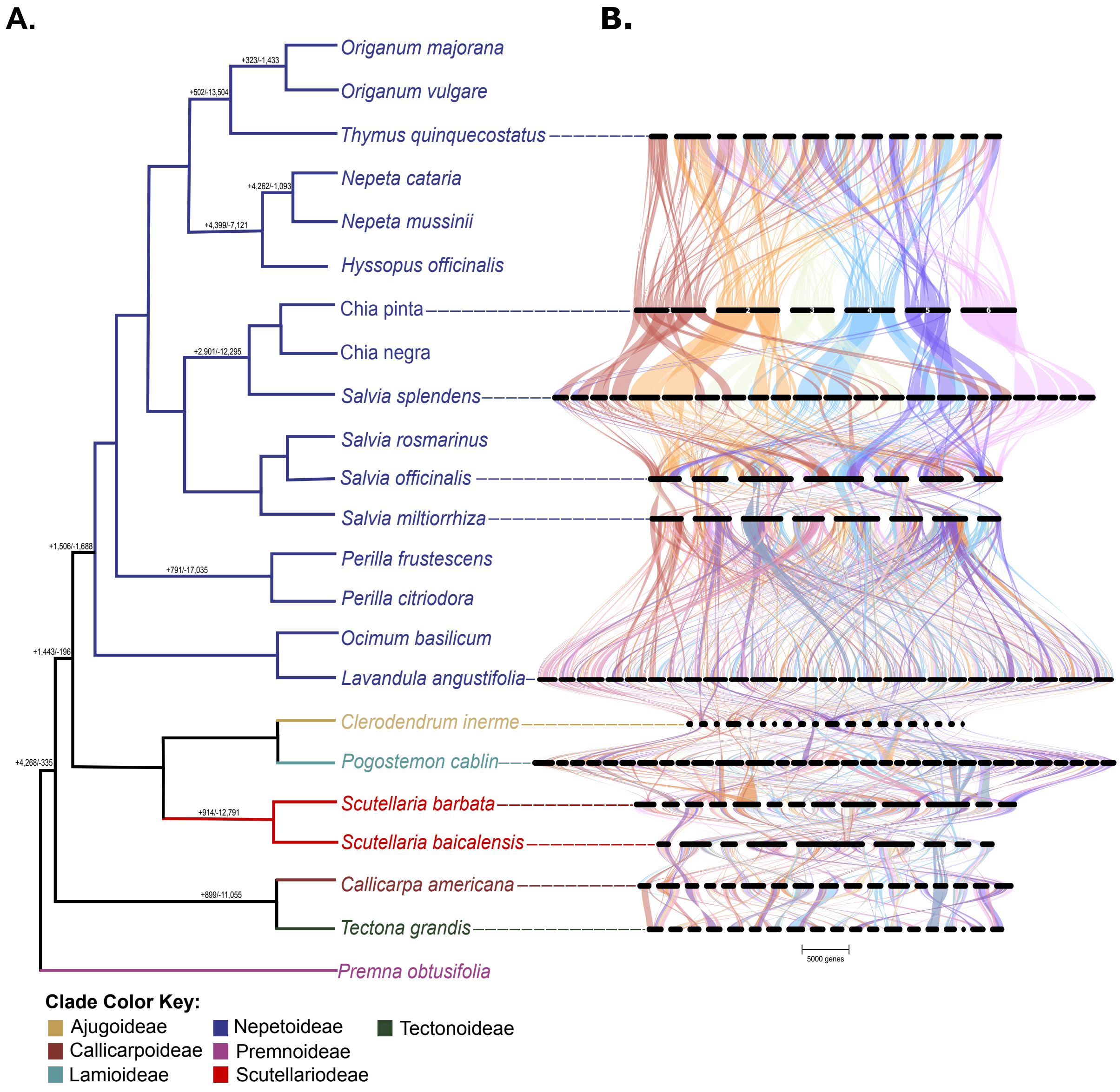


Figure 3. Lamiaceae phylogeny and synteny. A. A species phylogeny was generated using OrthoFinder (v.2.5.4;Emms & Kelly, 2019) using publicly available chromosome-scale Lamiaceae genomes. Numbers on branches indicated with (+) are gene family expansions and (-) are gene family contractions using CAFE (v.4.2.1; Han et al., 2013). **B.** The GENESPACE (v.1.1.10; Lovell et al., 2022) syntenic map of orthologous regions within chromosome-scale Lamiaceae genome assemblies are shown using the Chia pinta as the reference genome. Chromosomes are scaled by their physical length.

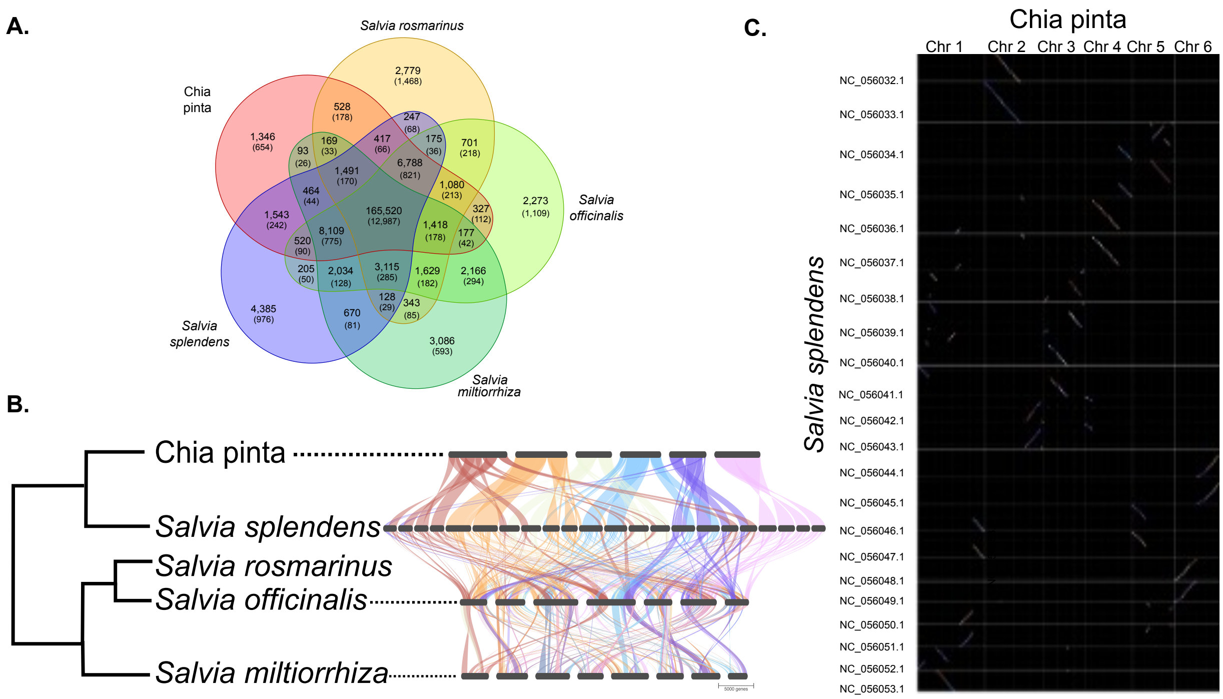


Figure 4. *Salvia* gene orthology and synteny. A. *Salvia* orthogroup intersections between Chia pinta, *Salvia rosmarinus*, *Salvia officinalis*, *Salvia splendens,* and *Salvia miltiorrhiza* as determined by OrthoFinder (v.2.5.4; Emms & Kelly, 2019). Numbers of orthologous groups and genes in parentheses are reported. **B.** GENESPACE (v.1.1.10; Lovell et al., 2022) syntenic map of orthologous regions within chromosome-scale *Salvia* genome assemblies using Chia pinta as the reference genome. **C.** Synteny dotplot for the anchor genes between Chia pinta and *Salvia splendens* generated in GENESPACE (v.1.1.10; Lovell et al., 2022). Chia pinta includes 21,720 genes with BLAST hits. *Salvia splendens* includes 25,958 genes with blast hits.

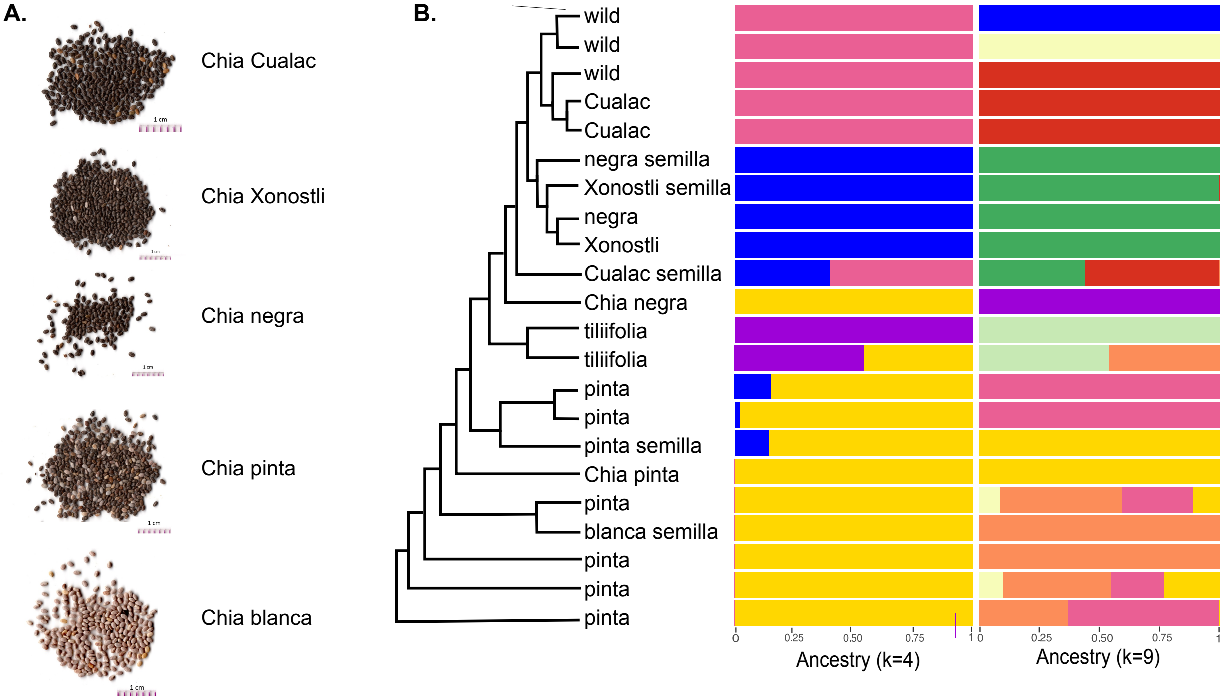


Figure 5. Population structure of Chia. **A.** Representative seed images of Chia varieties. **B.** SNP phylogeny was built using SNPhylo (v.20160204; Lee et al., 2014). Admixture (v.1.3.0; Alexander et al., 2009) population structure of 20 Chia accessions and 2 *Salvia tiliifolia* accessions was generated from 156,829 SNPs. Populations from the minima on the cross-validation plot was determined using k=4 and k=9.

Table 1. Chia pinta Genome Assembly Metrics

	Input assembly	Purged Assembly	Final Chromosome-scale Assembly
Number of Contigs/ Chromosomes	2,094	407	6
Total length (bp)	425,143,449	343,219,856	341,980,016
Maximum Contig Length (bp)	9,374,111	9,374,111	67,233,260
Minimum Contig Length (bp)	1,684	2,780	57,181,130
N50 Contig Length (bp)	1,150,825	1,506,829	62,351,092
Average Contig Length (bp)	203,029	858,050	56,996,669

Table 2. Chia pinta Genome Annotation Metrics

	High Confidence Model Set	High Confidence Representative Model Set	Working Model Set	Working Model Representative Set
Number of Gene Models	53,053	35,480	59,062	41,279
Number of Loci	35,480	35,480	41,279	41,279
Average Transcript Length (bp)	3,300.5	2,889.0	3,104.3	2,661.1
Average CDS Length (bp)	1,283.4	1,196.6	1,216.8	1,109.9
Average Exon Length (bp)	280.2	283.7	279.1	280.8
Average Intron Length (bp)	244.2	229.8	239.8	225.2
Average No. Exons per Model	6.1	5.3	5.8	4.9
Single Exon Transcripts	6,105	6,043	8,062	7,999