Plant Communications
Research article

CellPress
Partner Journal

The genome of *Orychophragmus violaceus* provides genomic insights into the evolution of Brassicaceae polyploidization and its distinct traits

Kang Zhang[1,4], Yinqing Yang[1,4], Xin Zhang[1,4], Lingkui Zhang[1], Yu Fu[1], Zhongwei Guo[1], Shumin Chen[1], Jian Wu[1], James C. Schnable[2,*], Keke Yi[3,*], Xiaowu Wang[1,*] and Feng Cheng[1,*]

[1]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Key Laboratory of Biology and Genetic Improvement of Horticultural Crops of the Ministry of Agriculture, Sino-Dutch Joint Laboratory of Horticultural Genomics, Beijing 10008, China

[2]Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE 68588, USA

[3]Key Laboratory of Plant Nutrition and Fertilizer, Ministry of Agriculture, Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[4]These authors contributed equally to this article.

***Correspondence: James C. Schnable (schnable@unl.edu), Keke Yi (yikeke@caas.cn), Xiaowu Wang (wangxiaowu@caas.cn), Feng Cheng (chengfeng@caas.cn)**

https://doi.org/10.1016/j.xplc.2022.100431

ABSTRACT

Orychophragmus violaceus, referred to as "eryuelan" (February orchid) in China, is an early-flowering ornamental plant. The high oil content and abundance of unsaturated fatty acids in _O. violaceus_ seeds make it a potential high-quality oilseed crop. Here, we generated a whole-genome assembly for _O. violaceus_ using Nanopore and Hi-C sequencing technologies. The assembled genome of _O. violaceus_ was ~1.3 Gb in size, with 12 pairs of chromosomes. Through investigation of ancestral genome evolution, we determined that the genome of _O. violaceus_ experienced a tetraploidization event from a diploid progenitor with the translocated proto-Calepineae karyotype. Comparisons between the reconstructed subgenomes of _O. violaceus_ identified indicators of subgenome dominance, indicating that subgenomes likely originated via allotetraploidy. _O. violaceus_ was phylogenetically close to the _Brassica_ genus, and tetraploidy in _O. violaceus_ occurred approximately 8.57 million years ago, close in time to the whole-genome triplication of _Brassica_ that likely arose via an intermediate tetraploid lineage. However, the tetraploidization in _Orychophragmus_ was independent of the hexaploidization in _Brassica_, as evidenced by the results from detailed phylogenetic analyses and comparisons of the break and fusion points of ancestral genomic blocks. Moreover, identification of multi-copy genes regulating the production of high-quality oil highlighted the contributions of both tetraploidization and tandem duplication to functional innovation in _O. violaceus_. These findings provide novel insights into the polyploidization evolution of plant species and will promote both functional genomic studies and domestication/breeding efforts in _O. violaceus_.

Key words: *Orychophragmus violaceus*, eryuelan, genome assembly, polyploidization, subgenome differentiation, function innovation

Zhang K., Yang Y., Zhang X., Zhang L., Fu Y., Guo Z., Chen S., Wu J., Schnable J.C., Yi K., Wang X., and Cheng F. (2023). The genome of *Orychophragmus violaceus* provides genomic insights into the evolution of Brassicaceae polyploidization and its distinct traits. Plant Comm. **4**, 100431.

INTRODUCTION

Polyploidization has occurred frequently throughout the evolutionary history of the plant kingdom (Adams and Wendel, 2005; Soltis et al., 2009). It has been proposed that widespread and recurrent polyploidization contributes to both diversification and evolutionary innovation of plant species (Van de Peer et al., 2009; Schranz et al., 2012; Vekemans et al., 2012; Hughes et al., 2014; Zhang et al., 2019a). Polyploidization generates

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

multiple sets of subgenomes that coexist in one nucleus. These subgenomes are then subjected to sequence fractionation (gene loss) and genomic reshuffling, giving rise to the bulk of speciation.

The family Brassicaceae (crucifers) contains ∼340 genera and 4636 species, including the model plant species *Arabidopsis thaliana* (Arabidopsis) (Francis et al., 2021). In addition to the γ hexaploidization event shared by all pan-eudicot species, all Brassicaceae species descended from a common ancestor that experienced two subsequent tetraploidization events: β and α. Multiple additional polyploidization events have been identified in specific lineages within Brassicaceae, such as the hexaploidization event in the *Brassica* and *Raphanus* genera, the hexaploidization event in *Camelina sativa*, and the hexaploidization event in *Leavenworthia alabamica* (Wang et al., 2011; Haudry et al., 2013; Kagale et al., 2014; Mandáková et al., 2017a; Guo et al., 2021). A genomic block system has been constructed in Brassicaceae to enable comparisons of orthologous or paralogous regions (Schranz et al., 2006; Lysak et al., 2016). Using the Arabidopsis genome as a reference, 22 genomic blocks have been defined in the ancestral karyotype of Brassicaceae (Lysak et al., 2016). Based on the genomic block system, the three subgenomes in *Brassica rapa* have been reconstructed, and the karyotype of the ancestral diploid genomes (2*n* = 14) before the *Brassica* hexaploidization event has been deduced (Cheng et al., 2013). Comparisons among the three *Brassica* subgenomes support the idea that hexaploidization was realized through a two-step process. In the first step, an intermediate tetraploid genome formed through the merging of two diploid genomes, and, in the second step, the intermediate tetraploid genome was further merged with a third diploid genome and formed a hexaploid genome, which then re-diploidized into different *Brassica* species (Wang et al., 2011; Cheng et al., 2013). Among the three subgenomes in the present-day diploid *Brassica* genomes, the two subgenomes involved in the first step have been named MF1 and MF2 (more fractionated), and the third one involved in the second step has been named LF (least fractionated).

Orychophragmus violaceus belongs to the *Orychophragmus* Bunge genus within the Brassicaceae family (Figure 1A). *O. violaceus* is an evolutionarily close relative to *Brassica* (Warwick and Sauder, 2005) and is reportedly easily crossed with *Brassica* species (Liu and Li, 2007; Xu et al., 2019). Studies of the karyotype of *O. violaceus* indicate that this species likely evolved from a tetraploid ancestor (Lysak et al., 2007). The combination of the above factors suggests that *O. violaceus* could potentially be a descendent of the tetraploid intermediate lineage (MF1 + MF2) that ultimately led to the formation of the hexaploid ancestor of modern *Brassica*.

O. violaceus is an ornamental plant and an important oil crop. It blooms in February of the lunar calendar and is therefore called "eryuelan" (February orchid) in China. *O. violaceus* possesses a variety of flower colors, such as purple, pink, and white (Weng et al., 2000; Xinping et al., 2018). The seeds of *O. violaceus* consist of up to 50.29% oil by mass, which is higher than many widely grown oilseed crops, including soybean (∼17%), rapeseed (∼40%), and peanut (∼48%) (Zhongjin, 1992). The seed oil produced by *O. violaceus* is rich in health-promoting unsaturated fatty acids (UFAs) and does not include erucic acid—a harmful unsaturated acid produced by many plants, particularly brassicas (Zhongjin, 1992; Qiao et al., 2019). The abundance of very-long-chain fatty acids also makes the seed oil of *O. violaceus* a high-performance vegetable oil lubricant (Li et al., 2018). Therefore, *O. violaceus* is a potential source of high-quality oil with economic value. Furthermore, *O. violaceus* has a short life cycle, high nutrient-use efficiency, and the ability to thrive on a variety of land types, making it an ideal research model for green manure plants.

It remains unclear whether the aforementioned traits in *O. violaceus* benefited from the evolutionary innovation created by meso-tetraploidization. Although the genetic mechanisms that regulate these traits have not been investigated in *O. violaceus*, gene function studies in Arabidopsis, which is phylogenetically closely related to *O. violaceus*, provide valuable information. Using the high UFA content in *O. violaceus* seeds as an example, the overexpression of the diacylglycerol acyltransferase 1 (*DGAT1*) gene increases the seed oil content through an increase in triacylglycerol biosynthesis in Arabidopsis (Jako et al., 2001; Zhang et al., 2009) and in other crops, such as soybean (Lardizabal et al., 2008) and maize (Zheng et al., 2008). Fatty acid desaturation enzymes (fatty acid desaturases [FADs]), such as FAD6 and FAD2, are also key enzymes involved in the synthesis of UFAs (Okuley et al., 1994; Pham et al., 2012). This information provides a foundation for the identification of key genes that regulate oil production and quality in *O. violaceus*.

In this study, we investigated the ancestral evolution associated with the polyploidization event, subgenome differentiation, and genome divergence of *O. violaceus* from *Brassica*, as well as the putative regulatory genes involved in the formation of important traits, based on a high-quality genome assembly of *O. violaceus*. These findings provide insights into the genome and gene evolution of *O. violaceus* and offer an important foundation for future studies.

RESULTS

Genome assembly and annotation of *O. violaceus*

The 12 pairs of chromosomes in *O. violaceus* were confirmed by cytological observation (supplemental Figure 1). A total of ∼60 Gb of Illumina Solexa paired-end read data were produced, and K-mer counting estimated that the genome size of *O. violaceus* is ∼1.33 Gb (supplemental Figure 2), close to the size (1.44 Gb) measured previously by flow cytometry (Lysak et al., 2007). The *O. violaceus* genome was then assembled from Oxford Nanopore (ONT) sequencing data and scaffolded using high-throughput chromosome conformation capture (Hi-C) sequencing methods. Specifically, 218.00 Gb (∼164× coverage) of ONT read data (supplemental Table 1) were assembled into contigs, followed by sequence polishing and the removal of heterozygous contigs (materials and methods). Subsequently, 344 contigs remained, with a total size of 1.34 Gb (supplemental Table 2). These contigs were further linked into 198 scaffolds using ∼178× (237.23 Gb) coverage of Hi-C data, with a scaffold N50 of 100.34 Mb (supplemental Table 2). The 12 largest scaffolds comprised 158 contigs, accounting for

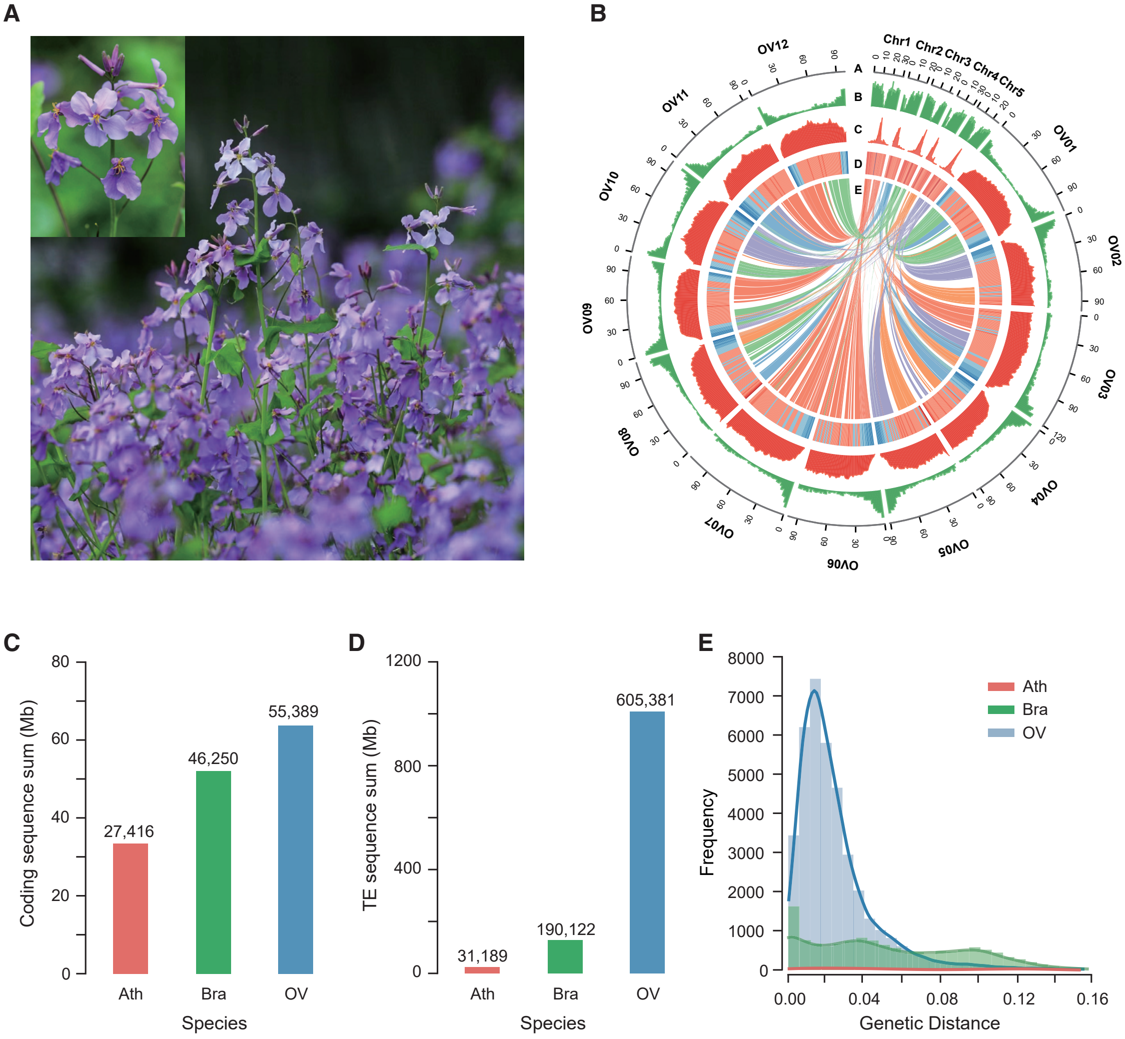

Figure 1. Genome assembly and characteristics of *Orychophragmus violaceus*.
(A) *O. violaceus* plant with a close-up view of its flowers.
(B) Chromosome size, gene density, TE density, GC content, and the genomic synteny relationship between *Arabidopsis thaliana* (upper right) and *O. violaceus* are indicated by **(A–E)**, respectively.
(C and D) Total length of coding genes **(C)** and TEs **(D)** in genomes of three species; the number of corresponding elements is placed on each bar.
(E) Distribution of the insertion time of full-length LTR retrotransposons in the genomes of the three species. *O. violaceus* experienced bulk LTR insertions relatively recently (genetic distance = 0.017). Ath, *Arabidopsis thaliana*; Bra, *Brassica rapa*; OV, *O. violaceus*.

90.95% (1.21 Gb) of the total assembly length and the total estimated genome size of *O. violaceus* (supplemental Figure 3 and supplemental Table 3). These 12 scaffolds were considered likely to correspond to the 12 chromosomes of *O. violaceus*. The completeness of the genome assembly was supported by a Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis (Seppey et al., 2019) in which 98.00% of the 1614 BUSCO genes were identified in the *O. violaceus* genome (supplemental Table 4).

A combination of *ab initio* and evidence-based methods was used to annotate gene models in the *O. violaceus* genome. Evidence-based methods included protein sequence alignments from related species and alignment of both short-read mRNA-seq and Nanopore-based full-length cDNA sequences to the *O. violaceus* reference genome. EVM was then used to integrate the evidence-based results with *ab initio* gene predictions (Haas et al., 2008). The final set of gene model annotations for *O. violaceus* contained 55 389 protein-coding genes (Figure 1B). More than 95% of these annotated gene models contained one or more conserved domains (supplemental Table 5). In addition, 136 microRNA genes, 1850 transfer RNAs, and 1329 ribosomal RNAs were identified and annotated in the *O. violaceus* genome.

Burst of TE insertions largely explains the expansion of the *O. violaceus* genome

Except for those of neo-polyploids, the genomes of sequenced Brassicaceae species are less than 0.7 Gb in size (Chen et al., 2022). Cytological observation confirmed that *O. violaceus* is a functionally diploid species (supplemental Figure 1). However,

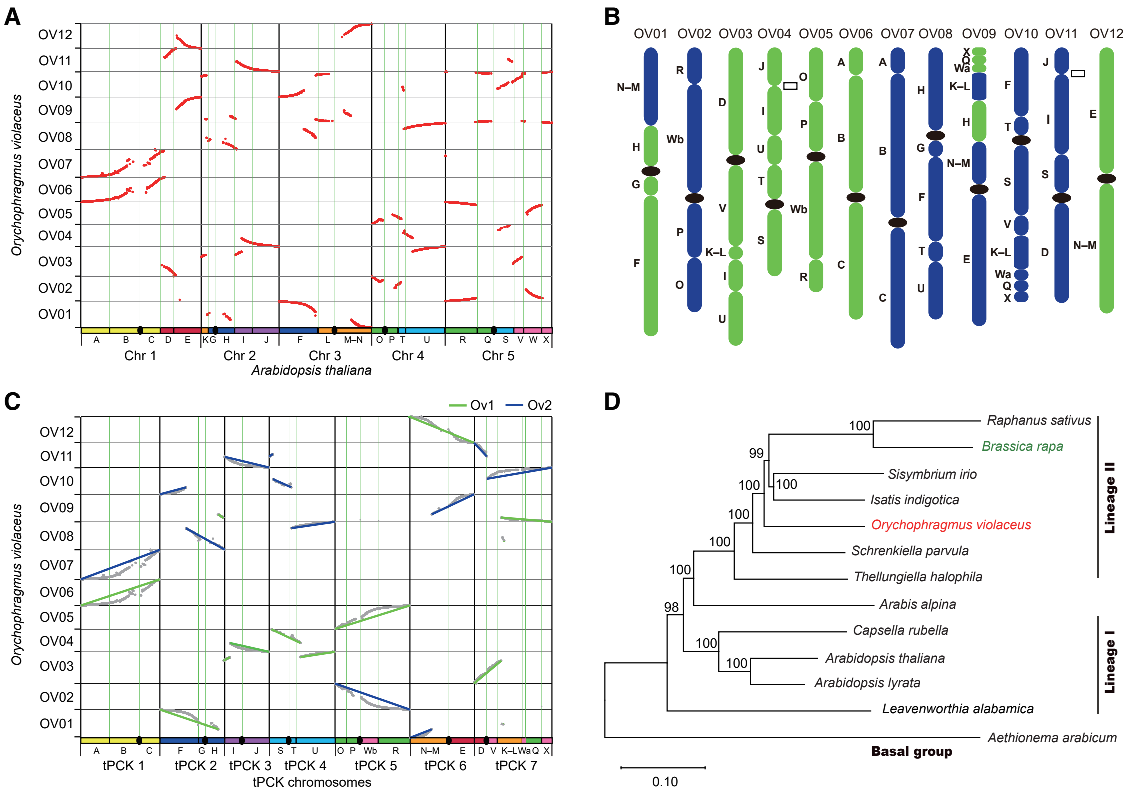

Figure 2. Diploid ancestor of *Orychophragmus violaceus* has a tPCK origin.

(A) Genomic synteny relationship between the genomes of Arabidopsis and *O. violaceus*.

(B) Distribution of the genomic blocks in the 12 chromosomes of *O. violaceus*. Filled rectangles represent 12 centromeres of *O. violaceus* that were inherited from the tPCK ancestor, whereas empty rectangles represent two ancestral centromeres that were inactivated in the *O. violaceus* genome.

(C) Syntenic fragments between *O. violaceus* and tPCK ancestral chromosomes.

(D) Phylogenetic tree, including 11 sequenced species from the Brassicaceae family, showing the close phylogenetic relationship between brassicas and *O. violaceus*.

the size of the *O. violaceus* genome is 1.33 Gb, which is larger than that of many other species in the Brassicaceae family (Lysak et al., 2009). We compared annotated genomic components among *O. violaceus* and two representative Brassicaceae diploid species, Arabidopsis (0.12 Gb genome, Brassicaceae lineage I) and *B. rapa* (0.49 Gb genome, Brassicaceae lineage II), to identify the factors that contribute to the increased genome size of *O. violaceus*. The sequences of the three genomes were separated into genic, intergenic, and centromeric regions. The genic region was further separated into exonic and intronic regions. We found that the intronic, centromeric, and intergenic regions were expanded in size in *O. violaceus* relative to Arabidopsis and *B. rapa*, whereas the quantity of exonic sequences in the genome showed little variation (Figure 1C and supplemental Figure 4). The expansion of intergenic regions was the largest contributor, explaining 77.33% of the size variation.

Annotation of repeat sequences across the *O. violaceus* genome (materials and methods and supplemental Table 2) made it possible to calculate the proportion of intronic, centromeric, and intergenic regions that consisted of different types of repeats, the ratios of which were compared against those of Arabidopsis and *B. rapa*. In all three categories, transposable elements (TEs) explained the largest increases in observed size (Figure 1D, supplemental Figure 4, and supplemental Table 6). Approximately 80% (79.43%) of the TEs in *O. violaceus* are long terminal repeat retrotransposons (LTR-RTs) from the Copia and Gypsy families. The age of individual LTR-RT insertions can be estimated by comparing the degree of sequence divergence between the two initially identical terminal repeat sequences on either side of the insertion. *O. violaceus* experienced a lineage-specific burst of LTR-RT insertions ~0.57 million years ago (MYA) (genetic distance = 0.016) (Figure 1E). This burst of new transposon insertions, which was estimated to have generated ~540 Mb of LTR-RTs, largely explained the substantial increase in genome size of *O. violaceus* relative to other Brassicaceae species.

Ancestral genome of *O. violaceus* has a tPCK origin

Genome synteny analysis performed between *O. violaceus* and Arabidopsis using SynOrths (Cheng et al., 2012a) identified 29 545 syntenic gene pairs between the two genomes. Using these syntenic gene pairs as pillars, we identified a set of 48 syntenic regions between the genomes of *O. violaceus* and Arabidopsis (materials and methods and Figure 2A). The previous genomic block information was mapped from the Arabidopsis genome to the *O. violaceus* genome based on these syntenic regions. In total, these syntenic relationships defined 44 genomic blocks in the 12 chromosomes of *O. violaceus*, 2 for each of the 22 genomic blocks in Arabidopsis (Figure 2B) (Lysak et al., 2016). This consistent 2:1 relationship indicates that the whole-genome duplication (WGD) event in the *O. violaceus* lineage occurred after the α event (tetraploidization) in Brassicaceae. To determine the origin of the diploid ancestor of *O. violaceus*, genomic block associations in the *O. violaceus* genome were further compared with the three major diploid karyotypes known in Brassicaceae: ACK (ancestral crucifer karyotype), PCK (proto-Calepineae karyotype), and tPCK (translocation PCK) (Cheng et al., 2013). The tPCK karyotype explained more of the genomic block associations observed in *O. violaceus* than either of the other two karyotypes (supplemental Table 7). In addition, both genomic block association groups (N–M/E and D/V/K–L/Wa/Q/X) that were specific to tPCK were identified in both duplicated copies within the *O. violaceus* genome (Figure 2C). These observations provide strong evidence that the diploid ancestor(s) of *O. violaceus* before the tetraploidization event had a tPCK origin.

The phylogenetic position of *O. violaceus* in the Brassicaceae family is also consistent with its evolution from the polyploidization of a tPCK ancestor. We performed genome comparisons among *O. violaceus* and 12 other Brassicaceae species whose genomes had been sequenced (supplemental Table 8). We identified 1773 syntenic gene families in all 13 Brassicaceae genomes and 212 499 synonymous nucleotide positions within these syntenic gene families (materials and methods). A phylogenetic tree constructed using these synonymous loci placed *O. violaceus* close to *Brassica* and *Schrenkiella parvula* in Brassicaceae lineage II but distant from Arabidopsis (lineage I) (Figure 2D). Considering that *Brassica*, *Sisymbrium irio*, *Isatis indigotica*, and *S. parvula* evolved from the tPCK diploid ancestor, the phylogenetic position of *O. violaceus* supports the idea that *O. violaceus* possesses a duplicated tPCK-like genome.

Gene fractionation in *O. violaceus* is not as strong as that in *Brassica*

The syntenic fragments between the genomes of *O. violaceus* and Arabidopsis were transferred to those between *O. violaceus* and the tPCK system, based on the genomic block associations in tPCK. A total of 10 break points were observed in the genome of *O. violaceus* compared with the ancestral chromosomes (Figures 2C and 3, supplemental Table 9). Because the break points occurred at different positions of the two copies of paralogous fragments generated by WGD, two copies were reconstructed for each of the seven ancestral chromosomes. By recognizing and relinking the 10 break points based on the tPCK chromosomes, we reconstructed seven pairs of ancestral chromosomes in the genome of *O. violaceus* (Figure 2C and supplemental Figure 5). Accordingly, we traced the rearrangements of tPCK chromosomes during the rediploidization of *O. violaceus*, i.e., to deduce how the two sets of tPCK chromosomes (7 × 2 = 14) evolved as the current 12 chromosomes of *O. violaceus*. We revealed that nine ancestral chromosomes were rearranged as seven *O. violaceus* chromosomes through large segmental translocation and the inactivation of two centromeres (Figure 3A and 3B, supplemental Table 10), whereas the other five ancestral chromosomes were inherited as the five current chromosomes (Figure 3C).

Subgenome dominance is one of the features that accompany allopolyploidization (Bird et al., 2018); in this phenomenon, genomic fragments from one subgenome retain more genes and are composed of more highly expressed paralogs than those of the other subgenome(s) (Senchina et al., 2003; Wang et al., 2006; Buggs et al., 2010; Woodhouse et al., 2010; Schnable et al., 2011; Tang et al., 2012; Murat et al., 2013; Pont et al., 2013; Akama et al., 2014; Li et al., 2014; Renny-Byfield et al., 2015). For each of the seven reconstructed pairs of tPCK chromosomes, one copy generally exhibited a higher gene content (retained more genes) than the other copy (Figure 4A), except for the short arm of tPCK4, where large fragment deletions were found. A difference in gene density was also observed when comparing the two copies of reconstructed chromosomes in units of genomic blocks (supplemental Table 11). Based on the consistent differences in gene content, we assigned the copy of each of the seven reconstructed chromosomes with the higher density of retained genes to subgenome 1 (Ov1) and the other copy of each of the seven chromosomes to subgenome 2 (Ov2) (Figure 4A). We compared the gene expression levels between syntenic paralogs in the two subgenomes and found that genes on Ov1 were significantly more likely to be expressed at higher levels than their paralogs in Ov2 than vice versa (Figure 4B). Dominance in gene expression was negatively associated with the density of TEs in the gene flanking regions between the two subgenomes (Figure 4C). These results indicate that the tetraploidization event in *O. violaceus* was accompanied by the subgenome dominance phenomenon.

Of the 24 401 genes in the Arabidopsis genome, 6296 and 6135 were completely absent from the genomes of *B. rapa* (as a representative of the *Brassica*) and *O. violaceus.* These genes likely represent either new genes gained in the lineage leading to Arabidopsis or genes lost in the common ancestor of *B. rapa* and *O. violaceus* prior to divergence and polyploidization. In *O. violaceus*, slightly more than half of these 18 105 presumptive ancestral genes were retained on both subgenomes (9579; 52.9%), and the remainder were largely reduced to single-copy status, with 30.1% of gene losses occurring on the Ov2 subgenome. By contrast, in *Brassica*, only approximately 14.5% of the presumed ancestral genes were retained across all three subgenomes. However, as an ancient hexaploid, *Brassica* has more potential gene copies to lose. To control for this, we calculated the proportion of all potential gene copies retained in each species. Of the 36 210 genes likely present in the initial tetraploid ancestor of *O. violaceus*, approximately 80.6% were still present in the modern *O. violaceus* genome. Of the 54 789 genes likely present in the initial hexaploid ancestor of *B. rapa*, only 55.8% were still present in the modern *B. rapa* genome. The level of gene loss in *O. violaceus* after the tetraploidization event was significantly lower ($p < 2.2 \times 10^{-16}$) than that in *B. rapa* after the *Brassica* hexaploidization event.

WGD event of *O. violaceus* is independent from that of *Brassica*

To estimate the age of the WGD event in *O. violaceus* and examine whether the tetraploidization of *O. violaceus* served as the first step in the hexaploidization of *Brassica*, we calculated the synonymous substitution rate (K_s) between pairs of paralogs in the subgenomes (materials and methods). It was estimated that the WGD of

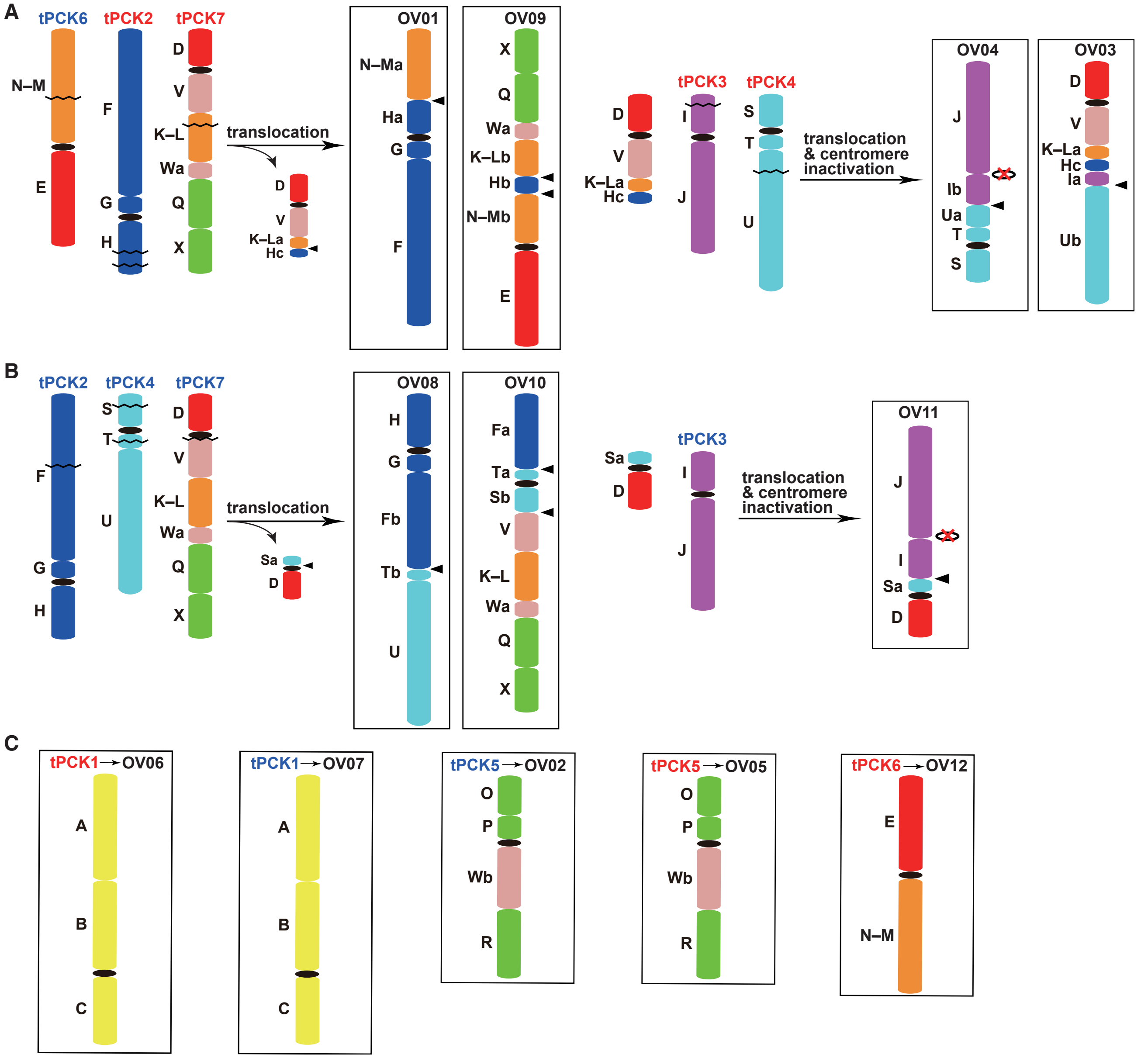

Figure 3. Deduced scenario by which the *Orychophragmus violaceus* chromosomes were derived from two sets of tPCK ancestral chromosomes.

(A) Ancestral chromosomes tPCK2, 3, 4, 6, and 7 were rearranged as the current chromosomes OV01, 03, 04, and 09 through translocations and centromere inactivation.

(B) Ancestral chromosomes tPCK2, 3, 4, and 7 were rearranged as the current chromosomes OV08, 10, and 11 through translocations and centromere inactivation.

(C) Ancestral chromosomes tPCK1 (two), 5 (two), and 6 were retained as OV02, 05, 06, 07, and 12, with their structures remaining unchanged. The color scheme of genomic blocks follows that of a previous study (Lysak et al., 2016). The black zigzag line denotes the break point. The black triangle denotes the position of genomic block fusion in the *O. violaceus* genome. No inversion was found during the rearrangements of tPCK chromosomes. The black ellipse denotes the centromere, whereas the hollow ellipse marked by the red cross denotes the inactivated/lost ancestral centromere. One set of tPCK chromosomes was categorized as Sub1 (red font), and the other set was categorized as Sub2 (blue font).

O. violaceus occurred less than 8.57 MYA (K_s = ~0.24). The meso-tetraploidization event was close in time but occurred later than the meso-hexaploidization event in *Brassica* (K_s = ~0.29) (Cheng et al., 2013) (Figure 5A). Furthermore, we investigated the evolutionary relationship between *O. violaceus* and *Brassica* at the subgenome level. The two subgenomes of *O. violaceus* were treated as two individual "genomes", as were the three subgenomes of *B. rapa* (as a representative of *Brassica*). Syntenic orthologous/paralogous genes were then identified among the two subgenomes of *O. violaceus*, the three subgenomes of *B. rapa*, and the three tPCK genomes of *S. parvula*, *I. indigotica*, and *S. irio*. In total, we identified 1544 syntenic gene families. A phylogenetic tree was then constructed with 143 959 synonymous loci from these syntenic gene families. As shown in

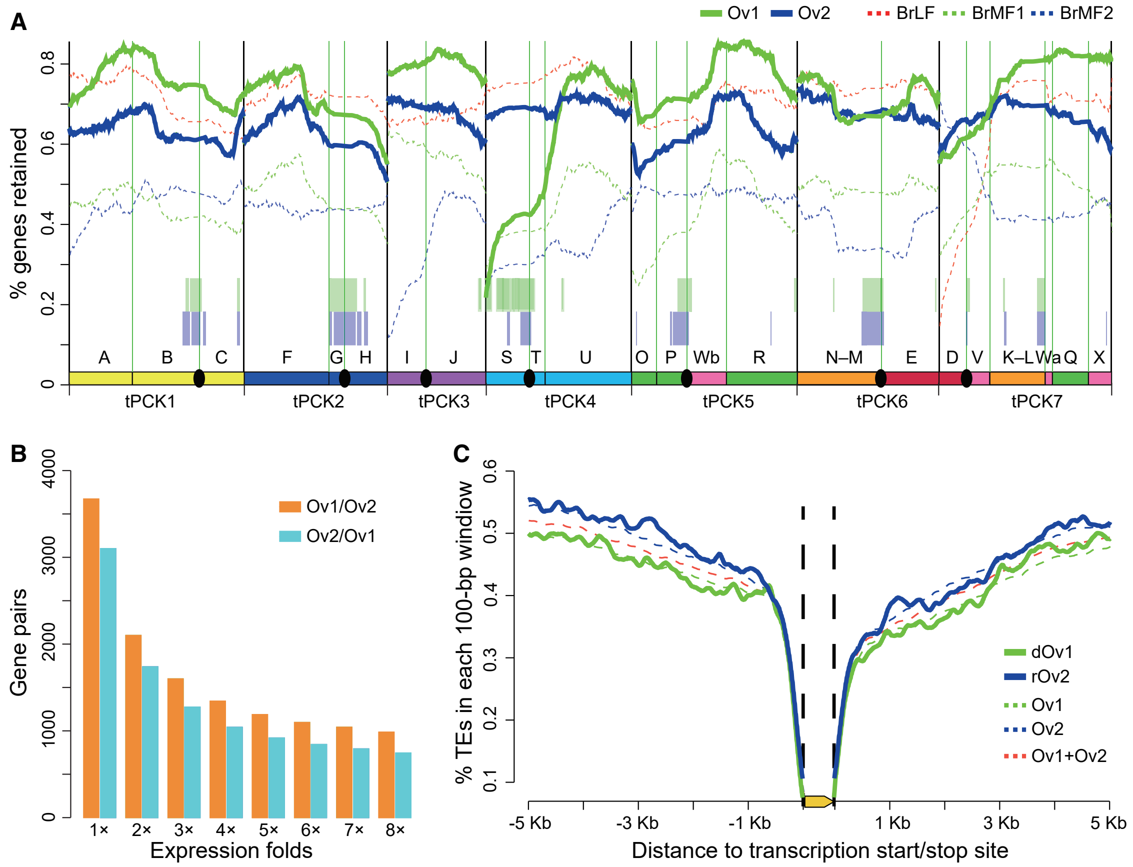

Figure 4. Subgenome dominance phenomenon accompanied the meso-tetraploidization event in *Orychophragmus violaceus*.
(A) Variation in gene density between the two subgenomes of *O. violaceus* (solid lines colored in green and blue) and among the three subgenomes of *B. rapa* (dashed lines colored in red, green, and blue), with subgenomes of *O. violaceus* retaining more genes than those of *B. rapa*, on average.
(B) Number of dominantly expressed genes between paralogous gene pairs from the two subgenomes of *O. violaceus*. Green denotes genes from Ov1 that are dominantly expressed over their paralogs from Ov2, whereas blue denotes genes from Ov1 that are dominantly expressed over their paralogs from Ov2.
(C) Distribution of TE sequences in the flanking regions of genes from different gene sets.

Figure 5B, all subgenomes of *B. rapa* (MF1, MF2, and LF) formed a single monophyletic clade that diverged from the two subgenomes of *O. violaceus*. This suggests that the two subgenomes of *O. violaceus* and the three subgenomes of *B. rapa* resulted from different WGD events. One subgenome of *O. violaceus* (Ov1) appeared to share a more recent common ancestor with *I. indigotica* and *S. irio*, whereas the other subgenome, Ov2, appeared to have diverged earlier (Figure 5B). This result supports a model in which the tetraploidization that led to the *O. violaceus* lineage was an allotetraploidization event resulting from a wide hybrid cross independent of the *Brassica* hexaploidization.

We compared the break and fusion points of the genomic blocks in the two subgenomes of *O. violaceus* with subgenomes in the *Brassica* genomes to further validate whether *O. violaceus* and *Brassica* had different WGD origins. As mentioned above, the MF1 and MF2 subgenomes in *Brassica* genomes, which were considered to be involved in the first step of the hexaploidization of brassicas, were used for comparisons. Ten break points—six break points occurring at Ov1 and four break points occurring at Ov2—occurred in the genome of *O. violaceus* during genomic rearrangements after tetraploidization. Four and five genomic fusions occurred within Ov1 and Ov2, respectively, whereas only two occurred between Ov1 and Ov2 of *O. violaceus* (supplemental Figure 6A and supplemental Table 12). Among the break points, only one (between blocks D/V) was also broken in subgenome MF1 of *Brassica* and *Raphanus sativus* (supplemental Figure 6). However, the break point between blocks D and V overlapped with a centromeric region, and centromeric regions are frequently the sites of genomic rearrangements. It is plausible that the break point between D and V may have occurred independently in *Brassica* and *O. violaceus*, rather than having been inherited from a common ancestor. None of the genomic block fusions in the genomes of *O. violaceus* were observed in the subgenomes of the *Brassica* species. Although *O. violaceus* shares a close phylogenetic relationship with *Brassica* species, there is no evidence to support the possibility that *O. violaceus* is descended from the tetraploid intermediate *Brassica* ancestor.

Hot spots of genomic rearrangement after WGD in *O. violaceus* and *Brassica*

Among the 10 aforementioned break points in the *O. violaceus* genome, except for the one between the D/V block, six were

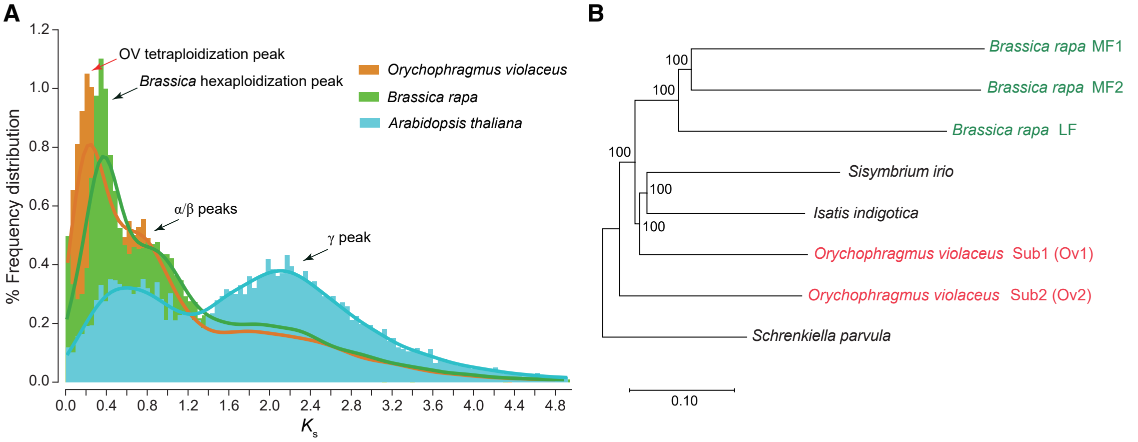

Figure 5. Close phylogenetic relationship with independent meso-polyploidization events between the genomes of *Orychophragmus violaceus* and *Brassica rapa*.

(A) Frequency distribution of K_s values between orthologous gene pairs in the genomes of Arabidopsis, *B. rapa*, and *O. violaceus*.

(B) Phylogenetic tree of two subgenomes of *O. violaceus* and three subgenomes of *B. rapa*, as well as *S. parvula* with a tPCK genome, which is ancestral to *O. violaceus* and *B. rapa*.

located in genomic regions close to but not overlapping with break points observed in the *Brassica* genomes (supplemental Figure 7). We defined these genomic regions as hot spots for genomic rearrangements. Six of the seven genomic regions corresponded to the common break points shared by *Brassica* and *R. sativus* genomes (referred to as the *Brassica*–*Raphanus* genome hereafter), indicating that these break points occurred before the divergence of the *Brassica* and *Raphanus* genera. For example, a break point occurred at block N–M in Ov2 of the *O. violaceus* genome, and a break point in the *Brassica*–*Raphanus* genome occurred 59 Arabidopsis genes away, as evaluated based on their genomic synteny relationships using the Arabidopsis genome as the reference. Other similar break points occurred at blocks H, I, S, and T of the *O. violaceus* genome, and the break points were 48–91 genes away in the *Brassica*–*Raphanus* genome from the locations of these break points in the *O. violaceus* genome (supplemental Table 13). In addition, we found one genomic region with extremely nearby break points between the genomes of *O. violaceus* and *Brassica*; that is, this break event occurred after the divergence of the *Brassica* and *Raphanus* genera. Specifically, the break point at the end of block H in Ov1 of the *O. violaceus* genome was only one gene away from the break point that occurred in the genome of *B. rapa* (supplemental Table 13). This high frequency (7/10) of closely neighboring break points was significantly biased from that of randomly distributed break points (permutation test, $p < 0.01$). These findings indicate that hot spots for genomic intervals occurred during reshuffling and rediploidization after the polyploidization of the Brassicaceae species. To explore whether the formation of hot spots was promoted by specific elements, we analyzed the sequence composition in most spot-involved genomic regions. The proportion of repetitive sequences in most hot spots (six of eight involved intervals)—or, more specifically, of the Gypsy-type LTR-RTs (five of eight involved intervals)—was slightly higher than that in the genome (supplemental Table 14), suggesting that TEs might facilitate the recurrent genomic rearrangements toward certain genomic regions.

Interval refinement of the G and K–L blocks in the ancestral karyotypes of Brassicaceae

The multispecies comparisons among *O. violaceus*, *B. rapa*, and *R. sativus*, ordered according to the tPCK, also revealed a potential transposition event. According to the tPCK, this potential transposed segment corresponded to *AT2G04039*–*AT2G05160* in the K–L block, as defined by Arabidopsis gene loci (Lysak et al., 2016). It was found to have been transposed from the K–L block and inserted between the F and G blocks (Figure 2C and supplemental Figure 6A). The potential transposition event was observed in both the subgenomes of *O. violaceus*, as well as in the two subgenomes (LF and MF2) of *B. rapa* and *R. sativus*, whereas the involved segment was not detected in the third subgenome (MF1), indicating that it had been lost during genome evolution (supplemental Figure 6). This suggests that this hypothetical event occurred in the ancestors of these three species. Further re-evaluation revealed that this segment was part of the G block (named G′) (Figure 6A) rather than the K–L block and should be located between the F and G blocks in the tPCK2 chromosome, given the connections of F/G′/G blocks present in all three species. The new positioning of this segment was confirmed by comparisons with crucifer genomes of the ACK karyotype, such as that of *Capsella rubella* (Figure 6A). In both genomes, this segment had the same location as in the tPCK, indicating that the modified connection was not introduced during the evolution of the species in Brassicaceae Lineage II from the ACK ancestor. The previous inaccurate assignment of this segment to the K–L block, causing an illusory transposition, probably occurred because this segment is located between a part of the K–L (referred to as [K–L]′) and G blocks in Arabidopsis. The [K–L]′/G connection was detected only in Arabidopsis but not in any other crucifer species (Mandáková and Lysak, 2016). Accordingly, we refined the intervals of the G and [K–L]′ blocks of the tPCK

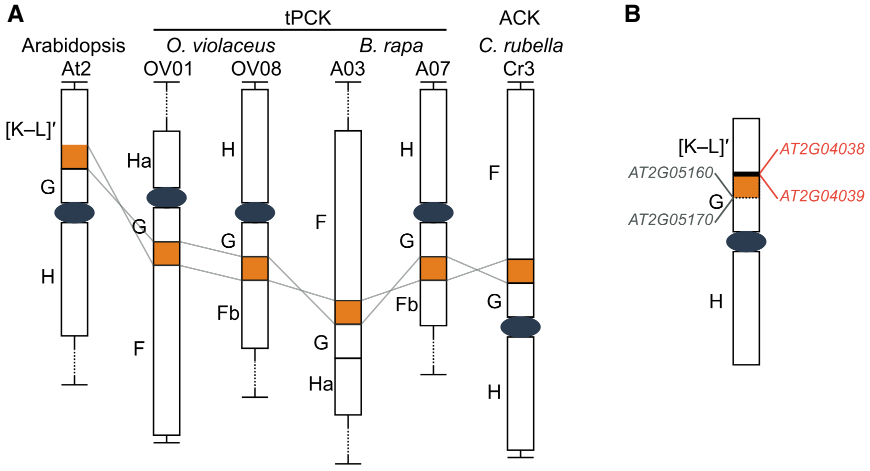

Figure 6. Refinement of intervals for genomic blocks K–L and G.

(A) Syntenic relationships regarding the K–L block among genomes of Arabidopsis, *O. violaceus* (tPCK), *B. rapa* (tPCK), and *C. rubella* (ACK). The [K–L]′ block refers to the part of K–L at the At2 chromosome of Arabidopsis. Except in Arabidopsis, a segment in the [K–L]′ block (highlighted in orange) is located between the G and F blocks in all other investigated species, indicating that it was inaccurately assigned to the [K–L]′ block.

(B) Refined intervals of genomic blocks [K–L]′ and G using the corresponding Arabidopsis genes as references. The interval defined by Lysak et al. (2016) is indicated by gray font and a dashed line between blocks, whereas the newly refined interval is indicated by orange font and a solid line between blocks [K–L]′ and G. The dark gray ellipse denotes the centromere.

by extending the G block from *AT2G05170–AT2G07690* to *AT2G04039–AT2G07690* and shortening the [K–L]′ block from *AT2G01060–AT2G05160* to *AT2G01060–AT2G04038* (Figure 6B).

Gene duplications associated with the evolution of distinctive traits in *O. violaceus*

We generated a set of fatty acid and lipid biosynthesis genes (73 genes in total) reported in Arabidopsis and identified 126, 125, and 70 orthologs in *O. violaceus*, *B. rapa*, and *S. parvula*, respectively (supplemental Table 15). *O. violaceus* retained/duplicated genes in oil synthesis regulation comparable with those of *B. rapa*, reflecting a much higher gene retention/duplication ratio in *O. violaceus*. Notably, three copies of *DGAT1*, a gene known to be closely associated with increased seed oil content (Jako et al., 2001), were identified in *O. violaceus* compared with two and one gene copy in *B. rapa* and *S. parvula*, respectively (Figure 7A). Two of the three *DGAT1*s in *O. violaceus* were paralogs duplicated through tetraploidization, whereas the third was duplicated from one of the paralogs through tandem duplication (Figure 7B). A similar scenario was also observed for the genes *FAD2* and *FAD6*, which encode enzymes critical to the synthesis of UFAs and, specifically, to the production of linoleic acid—the most abundant fatty acid in seeds of *O. violaceus*. *FAD2* and *FAD6* are responsible for transforming oleic acid (18:1) to linoleic acid (18:2) in the endoplasmic reticulum and plastid, respectively (Dar et al., 2017). We found five, two, and one copy of *FAD2* (Figure 7C), as well as three, two, and one copy of *FAD6* in *O. violaceus*, *B. rapa*, and *S. parvula*, respectively (Figure 7E). Both *FAD2* and *FAD6* retained the two paralogs generated by tetraploidization and were further augmented through tandem duplication or transposition-induced duplication (Figure 7D and 7F). These functional genes, exhibiting increased copy numbers in *O. violaceus*, may play a role in enabling high-quality and high-quantity oil production in *O. violaceus*.

Identification of key gene pathways regulating flowering in *O. violaceus*

We collected genes involved in flowering-time regulation in Arabidopsis and analyzed their orthologs and expression patterns in *O. violaceus*. We collected 174 genes in total from the four main pathways of flowering-time regulation in Arabidopsis (Pajoro et al., 2014), and we identified 263 orthologous genes in *O. violaceus* using comparative genomic analysis (supplemental Table 16).

We performed transcriptome profiling to investigate the pathways and genes important for flowering in *O. violaceus*. Leaf samples were collected at two developmental stages, vegetative growth and bolting, before and after cold treatment. Analysis of mRNA sequencing (mRNA-seq) data revealed that six genes related to flowering regulation showed significant expression fold changes ($p < 0.01$ and $q < 0.01$) (supplemental Figure 8 and supplemental Table 17). Among the six genes, three (two copies of *FLC* and one copy of *AGL19*) were from the vernalization pathway, two (*TSF* and *SMZ*) were from the photoperiod pathway, and one (*SOC1*) was from the meristem response and development pathway. Both copies of *FLC* and one *SMZ* were downregulated, whereas *AGL19*, *TSF*, and *SOC1* were upregulated. Combined with information from previous studies, these data indicated that low temperatures suppressed the expression of *FLC* and activated the expression of *AGL19*, whereas long day length suppressed the expression of *SMZ* and activated the expression of *TSF*. The expression change of genes from the vernalization and photoperiod pathways activated the expression of *SOC1*, which then promoted the flowering of *O. violaceus*. These results indicate that the vernalization and photoperiod gene pathways are important for flowering regulation in *O. violaceus*.

DISCUSSION

O. violaceus is a phylogenetically important species for understanding the evolutionary history of the Brassicaceae family and has significant economic potential as an ornamental plant and high-quality oil crop. We systematically investigated the genome evolution of *O. violaceus* through genome assembly and revealed that the larger genome size of *O. violaceus* compared with other Brassicaceae species is mostly due to TE insertions. We also confirmed and characterized the meso-tetraploidization event that occurred at 8.57 MYA in *O. violaceus*, thereby enriching our knowledge of the recurrent polyploidization events in Brassicaceae (Lysak et al., 2007; Franzke et al., 2011;

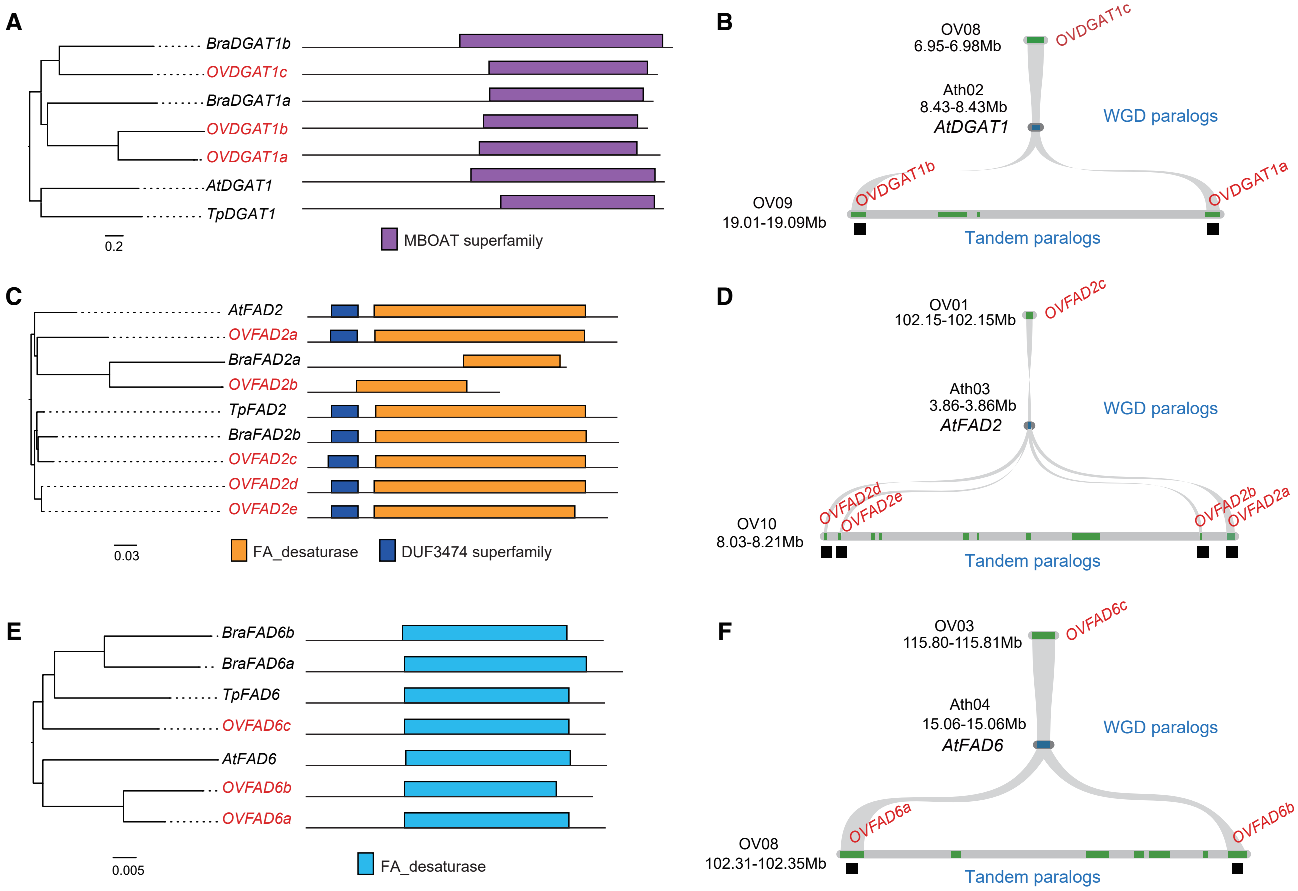

Figure 7. Genes associated with oil yield and content were duplicated through tetraploidization and tandem duplication in the *Orychophragmus violaceus* genome.
(A) Phylogenetic tree and functional domains in multiple copies of the *DGAT1* gene.
(B) WGD and tandem duplicated orthologous genes of *DGAT1* in the genome of *O. violaceus*.
(C) Phylogenetic tree and functional domains in the multiple copies of *FAD2*.
(D) WGD and tandem duplicated orthologs of *FAD2* in *O. violaceus*.
(E) Phylogenetic tree and functional domains in the multiple copies of *FAD6*.
(F) WGD and tandem duplicated orthologs of *FAD6* in *O. violaceus*.

Mandáková et al., 2017a). We further deduced that the diploid ancestor of *O. violaceus* before WGD had a tPCK origin. The subgenome dominance phenomenon was also observed in *O. violaceus*, but the level of gene fractionation was much lower than in *Brassica*. Comparative genomic analysis showed that *O. violaceus* had a close phylogenetic relationship with *Brassica*, although no evidence supported it as the intermediate tetraploid ancestor of *Brassica*. Moreover, comparative analysis of oil metabolism-related pathways supported the idea that multi-copy genes generated by both WGD and tandem duplication contributed to trait evolution in *O. violaceus*.

The two polyploidization events of *O. violaceus* and *Brassica* were distributed closely in the phylogenetic tree, which suggests that the two-step hexaploidization of *Brassica* was not an accidental event. We revealed that *O. violaceus* experienced a meso-tetraploidization event, which is consistent with a previous report based on a cytogenetics approach (Lysak et al., 2007). We further found that *O. violaceus* was tetraploidized from a tPCK ancestor whose karyotype was the same as that of the diploid ancestor of *Brassica*. In addition, K_s analysis of paralogs duplicated by polyploidizations showed that the divergence of the ancestor of the subgenomes in *O. violaceus* and *Brassica* occurred within a close historical period (Figure 5A). Previous studies have proposed a two-step evolution model for the *Brassica* hexaploidization event that involves a tetraploid as an intermediate ancestor of hexaploid *Brassica*. The close phylogenetic relationship between *O. violaceus* and *Brassica* suggested that *O. violaceus* was a candidate tetraploid ancestor of *Brassica*. However, we found only one break point in the genomic block, and no fusions were shared between the genomes of *O. violaceus* and the subgenomes MF1 and MF2 (the intermediate tetraploid) in *Brassica*. This suggests that the tetraploid *O. violaceus* ancestor may not have experienced the same re-diploidization process as the *Brassica* ancestor. Furthermore, the subgenome-based phylogenetic analysis positioned the three subgenomes of *B. rapa* into a single clade that diverged from the two subgenomes of *O. violaceus* (Figure 5B). It was then inferred that *O. violaceus* was not the intermediate tetraploid ancestor of *Brassica*, although it has a close phylogenetic position and comparable polyploidization time with that of *Brassica*. Nevertheless, the evolutionarily close relationship between the tetraploidization of *O. violaceus* and the hexaploidization of *Brassica* indicates that polyploidization events occurred frequently during the evolution of the *Brassica* ancestor, providing the necessary genome materials for the two-step hexaploidization of *Brassica*. In addition, comparison between *O. violaceus* and *Brassica* led to the refinement of two blocks, K–L and G, whose boundary was imprecisely defined in previous studies (Cheng et al., 2013; Lysak et al., 2016). We noticed that recent studies also revisited the boundaries of [K–L]′ and the G blocks according to cytogenetic results (Mandáková et al., 2019; Bayat et al., 2021). The end boundary of [K–L]′ proposed here (*AT2G04038*) is very close to that reported previously (*AT2G04032*). The high consistency between the comparative genomics and cytogenetic approaches strongly supports the accuracy of this refinement. By contrast, the G block was inferred to start at *AT2G04039* in our study, which is somewhat far from the boundary deduced recently (*AT2G05350*). This difference is likely to be attributed to the fact that the sequences in the G block are highly repetitive and genes in the G block have been heavily lost, increasing the difficulty of accurately determining its boundaries.

Genome size variation is a major feature of different species. In Brassicaceae, genome sizes vary from hundreds of millions of bases to eight gigabases (Lysak et al., 2009; Mandáková et al., 2017b). Two major factors, TE insertion and genome polyploidization, contribute to variations in genome size among plant species. A massive TE insertion can alter genome size and other features within a short period (Yang et al., 2019). Reports of different genomes show that TE insertion level is positively related to genome size. For example, the small genome of Arabidopsis has a TE proportion of only 16.69%, whereas in many species with huge genome sizes in lineage III of Brassicaceae, up to 67.19% of the genome is estimated to be occupied by TEs (Hloušková et al., 2019). *O. violaceus* has a relatively large genome compared with many other diploid Brassicaceae genomes reported (Lysak et al., 2009; Chen et al., 2022). The availability of the *O. violaceus* genome and its large genome size render it a representative for investigating the dominant factor causing genome expansion in Brassicaceae. TEs occupied 79.43% of the *O. violaceus* genome compared with 37.51% of the *B. rapa* genome, indicating that massive TE insertions have made a major contribution to the larger genome size of *O. violaceus*. Polyploidization doubles genome size until extensive genome fractionation/sequence loss occurs. Therefore, an expanded genome size fueled by polyploidization is commonly found in neo-polyploids, such as rapeseed (*Brassica napus*) and wheat (*Triticum aestivum*), in which duplicated genes and subgenomes have not been extensively fractionated through rediploidization. However, paleo-polyploidization events typically make a limited contribution to genome size, as the duplicated genes and subgenomes have been fractionated almost to the level of diploids. For example, Arabidopsis experienced three rounds of paleo-polyploidization—the γ, β, and α events—but it has a small genome (127 Mb). The ancestor diploid genome of *Brassica* experienced a further hexaploidization event but has an intermediate genome size of ~500 Mb. Similarly, our analysis found that the tetraploidization of *O. violaceus* contributed only slightly to the increased size of the *O. violaceus* genome. There were 55 389 and 46 250 genes annotated in *O. violaceus* and *B. rapa*, respectively. The sum of the coding sequences was 63.72 Mb in *O. violaceus*, slightly larger than that (52.07 Mb) in *B. rapa*. Genes duplicated by hexaploidization in *B. rapa* experienced stronger fractionation than genes duplicated by tetraploidization in *O. violaceus*. Using the Arabidopsis genome as a reference, we compared the gene duplication ratio between *B. rapa* and *O. violaceus* and showed that *O. violaceus* retained/duplicated more multi-copy genes after the tetraploidization event than did *B. rapa* after the hexaploidization event. As discussed above, these two polyploidization events occurred within a close historical period. Therefore, these results suggest that the more copies of genes generated by polyploidization, the stronger the fractionation of duplicated genes after polyploidization, which may be due to the more relaxed selection on more copies of genes duplicated by polyploidization. In addition to the higher retention of paralogs generated by polyploidization, many genes were duplicated through tandem duplication in the *O. violaceus* genome. These tandemly duplicated genes were important for the high oil content and quality of *O. violaceus* seeds. These findings and the genome/gene resources for *O. violaceus* will contribute to functional studies of this potential oil crop and green manure model plant, as well as providing valuable information for the improvement of traits, such as oil quality, in other crops.

Although *O. violaceus* is widely known as an ornamental plant, the genetic mechanism that underlies its flowering regulation is poorly understood. Here, a transcriptome assay provided a glimpse of the flowering-control pathways in *O. violaceus*. The importance of the transcriptional repressor *FLC* in regulating flowering time has been well documented in many Brassicaceae species (Tadege et al., 2001; Yuan et al., 2009; Hou et al., 2012; Xiao et al., 2013). There were two *FLC* copies in *O. violaceus*, and their expression was greatly reduced after cold treatment, consistent with the behaviors of their homologs in brassicas (Zhao et al., 2019; Akter et al., 2021). Their downregulation highlights the central role of *FLC* in the appropriate initiation of flowering processes. Our transcriptomic data revealed the activation of a key floral activator, *SOC1*, which belongs to the *TM3* clade of MADS-box genes, during the vernalization process (Lee and Lee, 2010). Intriguingly, two copies of *AGL19*, another floral activator in the same clade as *SOC1* (Becker and Theißen, 2003), were also upregulated after cold treatment in *O. violaceus*, and one copy was significantly differentially expressed (supplemental Table 17). Although previous studies have indicated that *AGL19* and *SOC1* may function partially independently (Schönrock et al., 2006), their activation suggests that this clade of MADS-box genes is a key component of the genetic network that controls flowering in *O. violaceus*. These results in *O. violaceus* extend our understanding of flowering regulation in Brassicaceae.

MATERIALS AND METHODS

Plant material

O. violaceus was planted in the greenhouse at the Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences (Beijing, China). Leaves of 6-week-old plants were collected and used for genome sequencing. The leaf, stem, flower, and root organs were also collected for mRNA-seq analysis.

Genome sequencing and assembly

DNA was isolated from the leaves by a standard genomic DNA extraction method using magnetic beads. We generated ~60 Gb of Illumina Solexa

150-bp paired-end reads. Approximately 200 Gb of ONT sequencing data were generated using the Oxford Nanopore PromethION sequencing platform, corresponding to approximately 147× coverage of the estimated genome size of *O. violaceus*. The average length of the ONT reads was ~15 kb, and the maximum length of the reads reached 833 kb. Minimap, followed by Miniasm (Li, 2016), was used to assemble the genome with ONT reads longer than 35 kb using default parameters. The resultant contigs (344 in total) were polished using Racon and Pilon with ~35× coverage of the longest ONT reads and ~44× coverage of Illumina paired-end reads. Purge Haplotigs (Roach et al., 2018) was used with default parameters to remove heterozygous sequences.

Pseudochromosome construction using Hi-C data

Fresh *O. violaceus* leaves were sampled for Hi-C sequencing. The Hi-C library was constructed using the Proximo Hi-C plant kit following standard protocols with the HindIII enzyme. About 288 million 150-bp paired-end reads were generated on the Illumina HiSeq 2000 platform. The Hi-C reads were mapped to the assembled contigs using Juicer (https://github.com/aidenlab/juicer). ALLHIC (Zhang et al., 2019b) was used to group, order, and orient the contigs (scaffolding). The linking results were manually curated to correct mis-joins and mis-assemblies by visualizing the interaction heatmap using Juicebox (https://github.com/aidenlab/Juicebox). The intra-chromosomal Hi-C heatmap was plotted using HiCPlotter (https://github.com/kcakdemir/HiCPlotter).

Repetitive element prediction

Homology-based and *de novo* approaches were used for repeat annotation. LTR_FINDER (Xu and Wang, 2007) and RepeatScout (Price et al., 2005) were used to generate a *de novo* repeat library for the *O. violaceus* genome. The *de novo* repeat library was classified using PASTEClassifier (Hoede et al., 2014) and merged with the Repbase (Jurka et al., 2005) database to produce the final repeat library. RepeatMasker (Tarailo-Graovac and Chen, 2009) was used to predict the repeat sequences in the genome of *O. violaceus* with the repeat library. The insertion time of the LTR-RTs was estimated using a previously described method (Cai et al., 2018). In brief, the terminal repeat sequences on both sides of an LTR-RT were extracted and compared using MUSCLE (Edgar, 2004), and the insertion time was then calculated based on the nucleotide mismatches between them.

Identification of the centromeric region

Centromere-associated repeats, such as CRB (Lim et al., 2007) and CL3 (Wang et al., 2019), were collected and aligned to the genome of *O. violaceus* using BLASTN (Altschul et al., 1990). The sequences of the best hits in the *O. violaceus* genome were extracted and aligned to the genome again. The distributions of the resultant hits were manually examined to locate the centromeric region in *O. violaceus*.

Protein-coding gene prediction and annotation

After pre-masking TE sequences, genes were predicted via *ab initio* prediction, homology-based searches, and mRNA-seq-assisted prediction. Genescan (Burger and Karlin, 1997), Augustus (Stanke and Waack, 2003), GlimmerHMM (Majoros et al., 2004), GeneID (Blanco et al., 2007), and SNAP (Korf, 2004) were used for *ab initio* gene prediction. For homology-based searches, we collected protein sequences of *A. thaliana*, *Arabidopsis lyrata*, *B. rapa*, *Brassicaoleracea*, *S. irio*, and *S. parvula* and used GeMoMa (Keilwagen et al., 2016) software for prediction. mRNA-seq datasets derived from roots, leaves, and stems of *O. violaceus* were used to assist in gene prediction. Specifically, HISAT2 (Kim et al., 2015), StringTie (Pertea et al., 2015), and TransDecoder (https://github.com/TransDecoder/TransDecoder) were used to assemble the mRNA-seq data into unigenes and to generate transcript-based gene models. All gene prediction datasets were combined by EVM (Haas et al., 2008) to generate the final gene set of *O. violaceus*, and the results were modified with PASA (Campbell et al., 2006). The parameters for HISAT2 were –max-intronlen 20000, –min-introlen 20; those for PASA were -align_tools gmap, -maxIntroLen 20000. Default parameters were used for the other tools.

The predicted genes of *O. violaceus* were further aligned to the Non-redundant (Nr) (Marchler-Bauer et al., 2011), EuKaryotic Orthologous Groups (KOG) (Koonin et al., 2004), Gene Ontology (GO) (Dimmer et al., 2012), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), and TrEMBL (Boeckmann et al., 2003) databases using BLASTP (Altschul et al., 1990) with an *E* value of 1×10^{-5}, and the most significant hits were retained. InterPro (Zdobnov and Apweiler, 2001) was used to annotate motifs and functional domains in the predicted genes. In addition, the *O. violaceus* genes were assigned to KEGG pathway maps based on the most probable Swiss-Prot hit for each gene.

Identification of syntenic genes

Syntenic orthologs were identified among the genomes of *O. violaceus* and other Brassicaceae species (Arabidopsis, *A. lyrata*, *Aethionema arabicum*, *Arabis alpina*, *B. rapa*, *C. rubella*, *I. indigotica*, *L. alabamica*, *R. sativus*, *S. irio*, *S. parvula*, and *Thellungiella halophila*) using SynOrths with default parameters (Cheng et al., 2012a). The Arabidopsis genome was used as the subject genome, and the others were used as query genomes.

K_s and phylogenetic analysis

Coding sequences of paralogous gene pairs or orthologous gene pairs were aligned using MUSCLE (Edgar, 2004). The synonymous nucleotide substitution rate per synonymous site (K_s) was calculated based on the sequence alignments following Nei and Gojobori's method as implemented in the KaKs_calculator (Zhang et al., 2006). For phylogenetic tree analysis, the genotypes of the K_s loci were concatenated, and a phylogenetic tree was constructed with the concatenated genotypes using the neighbor-joining method implemented in MEGA software (Kumar et al., 2016).

Determination of genomic fragments in synteny

Based on the syntenic gene pairs identified as described above, large-scale syntenic genomic fragments between the two genomes were identified by linking adjacent syntenic gene pairs. Considering local structural variations and the potential errors of genome assembly in one or both genomes, syntenic gene pairs may not be distributed immediately adjacent to the other syntenic gene pairs in one or both genomes. If two pairs of syntenic genes were interrupted by fewer than 50 genes or separated by fewer than 300 kb in the two genomes, then they were merged and considered to be one pair of syntenic fragments.

Genome blocks in the *O. violaceus* genome

The Arabidopsis genome contains one set of ancestral genomic blocks. Based on the identified syntenic fragments between the two genomes, block information from the Arabidopsis genome was mapped to the *O. violaceus* genome. For each block in Arabidopsis, there were two copies in the genome of *O. violaceus*.

Reconstruction of two subgenomes in the *O. violaceus* genome

Based on the syntenic fragments identified between the genomes of *O. violaceus* and Arabidopsis, as well as the chromosomal arrangements of the tPCK ancestor, we constructed two sets of ancestral chromosomes in the *O. violaceus* genome following methods reported elsewhere (Cheng et al., 2012b). The gene densities in the two sets of ancestral chromosomes were then compared. Because one copy of each of the seven ancestral chromosomes had more genes than the other, the set of chromosomes with a higher gene density was grouped together as subgenome 1 (Ov1), whereas the other set of chromosomes was grouped together as subgenome 2 (Ov2).

Comparison of expression between paralogs

Comparisons of expression levels between paralogs were conducted using the mRNA-seq leaf data for *O. violaceus*; the generation, processing, and quantification were performed as described above. The expression levels were compared between the paralogous gene pairs of the two subgenomes in *O. violaceus*. The number of paralogous pairs that showed two- to eight-fold expression differences was counted.

TE distribution variation around genes

A 100-bp sliding window with a 10-bp increment was used to scan the 5′ and 3′ flanking regions (5 kb) of each gene and to count the TE nucleotides. In each window, the ratio of TE nucleotides was calculated, and the average TE ratio was further calculated for a subset of genes in *O. violaceus*. These average values of the TE ratio were plotted to estimate the TE density in the flanking regions of the corresponding gene set in *O. violaceus*.

Agronomic trait-related gene analysis

Genes involved in the fatty acid biosynthesis and degradation pathway and the flowering time pathway in Arabidopsis were retrieved from the TAIR database (https://www.arabidopsis.org). Through syntenic gene analysis of the genomes of *O. violaceus*, *B. rapa*, and *S. parvula*, we obtained the syntenic orthologs of these Arabidopsis genes in the three species. We further identified the orthologous genes that showed no genomic synteny based on sequence homology (BLASTP with a cutoff of $E \le 1 \times 10^{-10}$ and coverage $\ge$80%). Syntenic genes and orthologous genes were integrated and further examined based on functional domain structures.

Flowering-related transcriptome data analysis

Fresh leaves were collected from mature plants in the vegetative phase and from bolting plants after 2 months of cold treatment. For transcriptome sequencing, mRNA was extracted using the Dynabeads mRNA DIRECT Kit (Illumina, San Diego, CA, USA). Sequencing libraries were generated using the VAHTS mRNA-seq v2 Library Prep Kit (Illumina, San Diego, CA, USA) following the manufacturer's recommendations. The libraries were sequenced at Biomarker Technologies Corporation (Beijing, China). The paired-end reads were aligned to the genome of *O. violaceus* using HISAT2 (version 2.1) (Kim et al., 2015), and expression was calculated using featureCounts (version 1.6.4) (Liao et al., 2014) with default parameters. Transcripts per million values were calculated using a local Perl script (available upon request). DESeq2 (version 1.20.0) (Love et al., 2014) was used to identify the differentially expressed genes with a cutoff of |fold change| $\ge$ 2 and *q*-value $\le$ 0.01.

ACCESSION NUMBERS

All genomic sequences and annotation datasets of the *O. violaceus* genome were deposited in the Genome Warehouse in the National Genomics Data Center (Chen et al., 2021) under accession number GWHBGBQ00000000 and are publicly accessible at https://ngdc.cncb.ac.cn/gwh. These genomic data are also available at http://www.bioinformaticslab.cn/pubs/OV_data/.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

FUNDING

This work was supported by the National Natural Science Foundation of China (NSFC grants 31722048 and 31972411), the Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences, and the Key Laboratory of Biology and Genetic Improvement of Horticultural Crops, Ministry of Agriculture and Rural Affairs, P.R. China. This work was also supported by the China Agricultural Research System—Green Manure (CARS-22). We thank Wei Zhang for kindly providing the photographs of eryuelan.

AUTHOR CONTRIBUTIONS

Conceptualization, F.C., X.W., and K.Y.; formal analysis, K.Z., X.Z., Y.Y., Z.G., L.Z., and Y.F.; visualization, K.Z., Y.Y., and X.Z.; resources, J.C.S.; writing – original draft, F.C., K.Z., Y.Y., and X.Z.; writing – review & editing, F.C., K.Z., J.C.S., X.W., and K.Y.; Supervision, F.C. and X.W.

ACKNOWLEDGMENTS

No conflict of interest declared.

Received: April 21, 2022
Revised: July 9, 2022
Accepted: August 24, 2022
Published: September 7, 2022

REFERENCES

Adams, K.L., and Wendel, J.F. (2005). Polyploidy and genome evolution in plants. Curr. Opin. Plant Biol. **8**:135–141.

Akama, S., Shimizu-Inatsugi, R., Shimizu, K.K., and Sese, J. (2014). Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. Nucleic Acids Res. **42**:e46.

Akter, A., Itabashi, E., Kakizaki, T., Okazaki, K., Dennis, E.S., and Fujimoto, R. (2021). Genome triplication leads to transcriptional divergence of *FLOWERING LOCUS C* genes during vernalization in the genus *Brassica*. Front. Plant Sci. **11**:619417.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Bayat, S., Lysak, M.A., and Mandáková, T. (2021). Genome structure and evolution in the cruciferous tribe Thlaspideae (Brassicaceae). Plant J. **108**:1768–1785.

Becker, A., and Theißen, G. (2003). The major clades of MADS-box genes and their role in the development and evolution of flowering plants. Mol. Phylogenet. Evol. **29**:464–489.

Bird, K.A., VanBuren, R., Puzey, J.R., and Edger, P.P. (2018). The causes and consequences of subgenome dominance in hybrids and recent polyploids. New Phytol. **220**:87–93.

Blanco, E., Parra, G., and Guigó, R. (2007). Using geneid to identify genes. Curr. Protoc. Bioinform. **64**:e56.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31**:365–370.

Buggs, R.J., Chamala, S., Wu, W., Gao, L., May, G.D., Schnable, P.S., Soltis, D.E., Soltis, P.S., and Barbazuk, W.B. (2010). Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. Mol. Ecol. **19** (*Suppl 1*):132–146.

Burger, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268**:78–94.

Cai, X., Cui, Y., Zhang, L., Wu, J., Liang, J., Cheng, L., Wang, X., and Cheng, F. (2018). Hotspots of independent and multiple rounds of LTR-retrotransposon bursts in *Brassica* species. Hortic. Plant J. **4**:165–174.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M., and Buell, C.R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genom. **7**:327.

Chen, H., Wang, T., He, X., Cai, X., Lin, R., Liang, J., Wu, J., King, G., and Wang, X. (2022). BRAD V3.0: an upgraded Brassicaceae database. Nucleic Acids Res. **50**:D1432–D1441.

Chen, M., Ma, Y., Wu, S., Zheng, X., Kang, H., Sang, J., Xu, X., Hao, L., Li, Z., Gong, Z., et al. (2021). Genome Warehouse: a public repository housing genome-scale data. Dev. Reprod. Biol. **19**:584–589.

Cheng, F., Wu, J., Fang, L., and Wang, X. (2012a). Syntenic gene analysis between *Brassica rapa* and other Brassicaceae species. Front. Plant Sci. **3**:198.

Cheng, F., Mandáková, T., Wu, J., Xie, Q., Lysak, M.A., and Wang, X. (2013). Deciphering the diploid ancestral genome of the mesohexaploid *Brassica rapa*. Plant Cell **25**:1541–1554.

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., and Wang, X. (2012b). Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. PLoS One **7**:e36442.

Dar, A.A., Choudhury, A.R., Kancharla, P.K., and Arumugam, N. (2017). The FAD2 gene in plants: occurrence, regulation, and role. Front. Plant Sci. **8**:1789.

Dimmer, E.C., Huntley, R.P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M.J., Bely, B., Browne, P., Mun Chan, W., Eberhardt, R., et al. (2012). The UniProt-GO annotation database in 2011. Nucleic Acids Res. **40**:D565–D570.

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

Francis, A., Lujan-Toro, B.E., Warwick, S.I., Macklin, J.A., and Martin, S.L. (2021). Update on the Brassicaceae species checklist. Biodivers. Data J. **9**:e58773.

Franzke, A., Lysak, M.A., Al-Shehbaz, I.A., Koch, M.A., and Mummenhoff, K. (2011). Cabbage family affairs: the evolutionary history of Brassicaceae. Trends Plant Sci. **16**:108–116.

Guo, X., Mandáková, T., Trachtová, K., Özüdoğru, B., Liu, J., and Lysak, M.A. (2021). Linked by ancestral bonds: multiple whole-genome duplications and reticulate evolution in a Brassicaceae tribe. Mol. Biol. Evol. **38**:1695–1714.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. **9**:R7.

Haudry, A., Platts, A.E., Vello, E., Hoen, D.R., Leclercq, M., Williamson, R.J., Forczek, E., Joly-Lopez, Z., Steffen, J.G., Hazzouri, K.M., et al. (2013). An atlas of over 90, 000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. **45**:891–898.

Hloušková, P., Mandáková, T., Pouch, M., Trávníček, P., and Lysak, M.A. (2019). The large genome size variation in the *Hesperis* clade was shaped by the prevalent proliferation of DNA repeats and rarer genome downsizing. Ann. Bot. **124**:103–120.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville, H. (2014). PASTEC: an automatic transposable element classification tool. PLoS One **9**:e91929.

Hou, J., Long, Y., Raman, H., Zou, X., Wang, J., Dai, S., Xiao, Q., Li, C., Fan, L., Liu, B., et al. (2012). A *Tourist*-like MITE insertion in the upstream region of the *BnFLC.A10* gene is associated with vernalization requirement in rapeseed (*Brassica napus* L.). BMC Plant Biol. **12**:238.

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

Korf, I. (2004). Gene finding in novel Genomes. BMC Bioinf. **5**:59.

Jako, C., Kumar, A., Wei, Y., Zou, J., Barton, D.L., Giblin, E.M., Covello, P.S., and Taylor, D.C. (2001). Seed-specific over-expression of an Arabidopsis cDNA encoding a diacylglycerol acyltransferase enhances seed oil content and seed weight. Plant Physiol. **126**:861–874.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110**:462–467.

Kagale, S., Koh, C., Nixon, J., Bollina, V., Clarke, W.E., Tuteja, R., Spillane, C., Robinson, S.J., Links, M.G., Clarke, C., et al. (2014). The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. Nat. Commun. **5**:3706.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28**:27–30.

Keilwagen, J., Wenk, M., Erickson, J.L., Schattat, M.H., Grau, J., and Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. **44**:e89.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods **12**:357–360.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., et al. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. **5**:R7.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. **33**:1870–1874.

Lardizabal, K., Effertz, R., Levering, C., Mai, J., Pedroso, M.C., Jury, T., Aasen, E., Gruys, K., and Bennett, K. (2008). Expression of *Umbelopsis ramanniana DGAT2A* in seed increases oil in soybean. Plant Physiol. **148**:89–96.

Lee, J., and Lee, I. (2010). Regulation and function of SOC1, a flowering pathway integrator. J. Exp. Bot. **61**:2247–2254.

Li, A., Liu, D., Wu, J., Zhao, X., Hao, M., Geng, S., Yan, J., Jiang, X., Zhang, L., Wu, J., et al. (2014). mRNA and small RNA transcriptomes reveal insights into dynamic homoeolog regulation of allopolyploid heterosis in nascent hexaploid wheat. Plant Cell **26**:1878–1900.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics **32**:2103–2110.

Li, X., Teitgen, A.M., Shirani, A., Ling, J., Busta, L., Cahoon, R.E., Zhang, W., Li, Z., Chapman, K.D., Berman, D., et al. (2018). Discontinuous fatty acid elongation yields hydroxylated seed oil with improved function. Nat. Plants **4**:711–720.

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

Lim, K.-B., Yang, T.-J., Hwang, Y.-J., Kim, J.S., Park, J.-Y., Kwon, S.-J., Kim, J., Choi, B.-S., Lim, M.-H., Jin, M., et al. (2007). Characterization of the centromere and peri-centromere retrotransposons in *Brassica rapa* and their distribution in related *Brassica* species. Plant J. **49**:173–183.

Liu, M., and Li, Z.Y. (2007). Genome doubling and chromosome elimination with fragment recombination leading to the formation of Brassica rapa-type plants with genomic alterations in crosses with Orychophragmus violaceus. Genome **50**:985–993.

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

Lysak, M.A., Mandáková, T., and Schranz, M.E. (2016). Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. Curr. Opin. Plant Biol. **30**:108–115.

Lysak, M.A., Cheung, K., Kitschke, M., and Bures, P. (2007). Ancestral chromosomal blocks are triplicated in brassiceae species with varying chromosome number and genome size. Plant Physiol. **145**:402–410.

Lysak, M.A., Koch, M.A., Beaulieu, J.M., Meister, A., and Leitch, I.J. (2009). The dynamic ups and downs of genome size evolution in Brassicaceae. Mol. Biol. Evol. **26**:85–98.

Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics **20**:2878–2879.

Mandáková, T., and Lysak, M.A. (2016). Chromosome preparation for cytogenetic analyses in *Arabidopsis*. Curr. Protoc. Plant Biol. **1**:43–51.

Mandáková, T., Li, Z., Barker, M.S., and Lysak, M.A. (2017a). Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. Plant J. **91**:3–21.

Mandáková, T., Hloušková, P., German, D.A., and Lysak, M.A. (2017b). Monophyletic origin and evolution of the largest crucifer genomes. Plant Physiol. **174**:2062–2071.

Mandáková, T., Pouch, M., Brock, J.R., Al-Shehbaz, I.A., and Lysak, M.A. (2019). Origin and evolution of diploid and allopolyploid Camelina genomes were accompanied by chromosome shattering. Plant Cell **31**:2596–2612.

Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. **39**:D225–D229.

Murat, F., Zhang, R., Guizard, S., Flores, R., Armero, A., Pont, C., Steinbach, D., Quesneville, H., Cooke, R., and Salse, J. (2013). Shared subgenome dominance following polyploidization explains grass genome evolutionary plasticity from a seven protochromosome ancestor with 16K protogenes. Genome Biol. Evol. **6**:12–33.

Okuley, J., Lightner, J., Feldmann, K., Yadav, N., Lark, E., and Browse, J. (1994). Arabidopsis *FAD2* gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. Plant Cell **6**:147–158.

Pajoro, A., Biewers, S., Dougali, E., Leal Valentim, F., Mendes, M.A., Porri, A., Coupland, G., Van de Peer, Y., van Dijk, A.D.J., Colombo, L., et al. (2014). The (r)evolution of gene regulatory networks controlling Arabidopsis plant reproduction: a two-decade history. J. Exp. Bot. **65**:4731–4745.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. **33**:290–295.

Pham, A.-T., Shannon, J.G., and Bilyeu, K.D. (2012). Combinations of mutant *FAD2* and *FAD3* genes to produce high oleic acid and low linolenic acid soybean oil. Theor. Appl. Genet. **125**:503–515.

Pont, C., Murat, F., Guizard, S., Flores, R., Foucrier, S., Bidet, Y., Quraishi, U.M., Alaux, M., Doležel, J., Fahima, T., et al. (2013). Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. Plant J. **76**:1030–1044.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics **21**:i351–i358.

Qiao, Q., Wang, X., Ren, H., An, K., Feng, Z., Cheng, T., and Sun, Z. (2019). Oil content and nervonic acid content of acer truncatum seeds from 14 regions in China. Hortic. Plant J. **5**:24–30.

Renny-Byfield, S., Gong, L., Gallagher, J.P., and Wendel, J.F. (2015). Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. Mol. Biol. Evol. **32**:1063–1071.

Roach, M.J., Schmidt, S.A., and Borneman, A.R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinf. **19**:460.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. USA **108**:4069–4074.

Schönrock, N., Bouveret, R., Leroy, O., Borghi, L., Köhler, C., Gruissem, W., and Hennig, L. (2006). Polycomb-group proteins repress the floral activator *AGL19* in the *FLC*-independent vernalization pathway. Genes Dev. **20**:1667–1678.

Schranz, M.E., Lysak, M.A., and Mitchell-Olds, T. (2006). The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci. **11**:535–542.

Schranz, M.E., Mohammadin, S., and Edger, P.P. (2012). Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Curr. Opin. Plant Biol. **15**:147–153.

Senchina, D.S., Alvarez, I., Cronn, R.C., Liu, B., Rong, J., Noyes, R.D., Paterson, A.H., Wing, R.A., Wilkins, T.A., and Wendel, J.F. (2003). Rate variation among nuclear genes and the age of polyploidy in Gossypium. Mol. Biol. Evol. **20**:633–643.

Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: assessing genome assembly and annotation completeness. Methods Mol. Biol. **1962**:227–245.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., Depamphilis, C.W., Wall, P.K., and Soltis, P.S. (2009). Polyploidy and angiosperm diversification. Am. J. Bot. **96**:336–348.

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics **19**:ii215–ii225.

Tadege, M., Sheldon, C.C., Helliwell, C.A., Stoutjesdijk, P., Dennis, E.S., and Peacock, W.J. (2001). Control of flowering time by *FLC* orthologues in *Brassica napus*. Plant J. **28**:545–553.

Tang, H., Woodhouse, M.R., Cheng, F., Schnable, J.C., Pedersen, B.S., Conant, G., Wang, X., Freeling, M., and Pires, J.C. (2012). Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. Genetics **190**:1563–1574.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinform. **Chapter 4**. Unit 4.10-14.

Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L., and Vandepoele, K. (2009). The flowering world: a tale of duplications. Trends Plant Sci. **14**:680–688.

Vekemans, D., Proost, S., Vanneste, K., Coenen, H., Viaene, T., Ruelens, P., Maere, S., Van de Peer, Y., and Geuten, K. (2012). Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. Mol. Biol. Evol. **29**:3793–3806.

Wang, G.-x., He, Q.-y., Zhao, H., Cai, Z.-x., Guo, N., Zong, M., Han, S., Liu, F., and Jin, W.-w. (2019). ChIP-cloning analysis uncovers centromere-specific retrotransposons in *Brassica nigra* and reveals their rapid diversification in *Brassica* allotetraploids. Chromosoma **128**:119–131.

Wang, J., Tian, L., Lee, H.S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C., Doerge, R.W., Comai, L., et al. (2006). Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics **172**:507–517.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.H., Bancroft, I., Cheng, F., et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet. **43**:1035–1039.

Warwick, S.I., and Sauder, C.A. (2005). Phylogeny of tribe Brassiceae (Brassicaceae) based on chloroplast restriction site polymorphisms and nuclear ribosomal internal transcribed spacer and chloroplast trnL intron sequences. Can. J. Bot. **83**:467–483.

Weng, D., Wang, H., and Weng, J. (2000). Studies on FlavonoidsL in leaves and stalks of *Orychophragmus violaceus* (L.). Chin. Wile Plant Resourc. **5**:13–15.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S., and Freeling, M. (2010). Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. PLoS Biol. **8**:e1000409.

Xiao, D., Zhao, J.J., Hou, X.L., Basnet, R.K., Carpio, D.P.D., Zhang, N.W., Bucher, J., Lin, K., Cheng, F., Wang, X.W., et al. (2013). The Brassica rapa *FLC* homologue *FLC2* is a key regulator of flowering time, identified through transcriptional co-expression networks. J. Exp. Bot. **64**:4503–4516.

Xinping, J., Yanming, D., Xiaobo, S., Lijian, L., Zheng, X., Xiaoqing, L., and Jiale, S. (2018). Analysis of anthocyanin composition in the flower of *Orychophragmus violaceus*. China Agric. Bull. **10**:60–64.

Xu, C., Huang, Q., Ge, X., and Li, Z. (2019). Phenotypic, cytogenetic, and molecular marker analysis of Brassica napus introgressants derived from an intergeneric hybridization with *Orychophragmus*. PLoS One **14**:e0210518.

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

Yang, S., Cheng, C., Qin, X., Yu, X., Lou, Q., Li, J., Qian, C., and Chen, J. (2019). Comparative cyto-molecular analysis of repetitive DNA provides insights into the differential genome structure and evolution of five *Cucumis* species. Hortic. Plant J. **5**:192–204.

Yuan, Y.-X., Wu, J., Sun, R.-F., Zhang, X.-W., Xu, D.-H., Bonnema, G., and Wang, X.-W. (2009). A naturally occurring splicing site mutation in the *Brassica rapa FLC1* gene is associated with variation in flowering time. J. Exp. Bot. **60**:1299–1308.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17**:847–848.

Zhang, K., Wang, X., and Cheng, F. (2019a). Plant polyploidy: origin, evolution, and its influence on crop domestication. Hortic. Plant J. **5**:231–239.

Zhang, M., Fan, J., Taylor, D.C., and Ohlrogge, J.B. (2009). *DGAT1* and *PDAT1* acyltransferases have overlapping functions in *Arabidopsis* triacylglycerol biosynthesis and are essential for normal pollen and seed development. Plant Cell **21**:3885–3901.

Zhang, X., Zhang, S., Zhao, Q., Ming, R., and Tang, H. (2019b). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat. Plants **5**:833–845.

Zhang, Z., Li, J., Zhao, X.Q., Wang, J., Wong, G.K.S., and Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Dev. Reprod. Biol. **4**:259–263.

Zhao, N., Liu, C., Meng, Y., Hu, Z., Zhang, M., and Yang, J. (2019). Identification of flowering regulatory genes in allopolyploid *Brassica juncea*. Hortic. Plant J. **5**:109–119.

Zheng, P., Allen, W.B., Roesler, K., Williams, M.E., Zhang, S., Li, J., Glassman, K., Ranch, J., Nubel, D., Solawetz, W., et al. (2008). A phenylalanine in DGAT is a key determinant of oil content and composition in maize. Nat. Genet. **40**:367–372.

Zhongjin, L. (1992). Oil analysis and utilization evaluation of *Orychophragmus violaceus* seeds. Chin. Wild Plants **3**:1–5.