Acta Pharmaceutica Sinica B 2024;14(4):1878–1891

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

ORIGINAL ARTICLE

Crocus genome reveals the evolutionary origin of crocin biosynthesis

Zhichao Xu[a,b,c,†], Shanshan Chen[c,d,†], Yalin Wang[c,†], Ya Tian[c], Xiaotong Wang[c,d], Tianyi Xin[a,b], Zishan Li[c], Xin Hua[c], Shengnan Tan[e], Wei Sun[d], Xiangdong Pu[f], Hui Yao[a,b], Ranran Gao[d,*], Jingyuan Song[a,b,*]

[a]*Key Lab of Chinese Medicine Resources Conservation, State Administration of Traditional Chinese Medicine of the People's Republic of China, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100193, China*
[b]*State Key Laboratory of Basis and New Drug Development of Natural and Nuclear Drugs, Engineering Research Center of Chinese Medicine Resource, Ministry of Education, Beijing 100193, China*
[c]*College of Life Science, Northeast Forestry University, Harbin 150040, China*
[d]*Key Laboratory of Beijing for Identification and Safety Evaluation of Chinese Medicine, China Academy of Chinese Medical Sciences, Institute of Chinese Materia Medica, Beijing 100700, China*
[e]*Analysis and Testing Center of Northeast Forestry University, Harbin 150040, China*
[f]*School of Pharmacy, Anhui Medical University, Inflammation and Immune Mediated Diseases Laboratory of Anhui Province, Hefei 230032, China*

Received 26 August 2023; received in revised form 22 October 2023; accepted 9 November 2023

KEY WORDS

Crocus sativus;
WGT;
Crocin biosynthesis;
Carotenoids;
Apocarotenoids;
CCDs

Abstract *Crocus sativus* (saffron) is a globally autumn-flowering plant, and its stigmas are the most expensive spice and valuable herb medicine. *Crocus* specialized metabolites, crocins, are biosynthesized in distant species, *Gardenia* (eudicot) and *Crocus* (monocot), and the evolution of crocin biosynthesis remains poorly understood. With the chromosome-level *Crocus* genome assembly, we revealed that two rounds of lineage-specific whole genome triplication occurred, contributing important roles in the production of carotenoids and apocarotenoids. According to the kingdom-wide identification, phylogenetic analysis, and functional assays of carotenoid cleavage dioxygenases (CCDs), we deduced that the duplication, site positive selection, and neofunctionalization of *Crocus*-specific CCD2 from CCD1

*Corresponding authors.
E-mail addresses: jysong@implad.ac.cn (Jingyuan Song), rrgao1991@icmm.ac.cn (Ranran Gao).
[†]These authors made equal contributions to this work.
Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2023.12.013

members are responsible for the crocin biosynthesis. In addition, site mutation of CsCCD2 revealed the key amino acids, including I143, L146, R161, E181, T259, and S292 related to the catalytic activity of zeaxanthin cleavage. Our study provides important insights into the origin and evolution of plant specialized metabolites, which are derived by duplication events of biosynthetic genes.

1. Introduction

The stigmas of autumn-flowering *Crocus sativus*, named "red gold" saffron, are the source of the most expensive spice[1]. Owing to its prevalent uses as a spice, colorant, and medicine, saffron is an important cash crop distributed in Europe, the Mediterranean, and central Asia. Saffron has been introduced into China for nearly seven hundred years, and stigmas of saffron have been noted in the Chinese Pharmacopoeia for their medicinal usage in invigorating blood circulation and removing depression. The main active compounds of saffron stigmas are crocins, the glycosides of apocarotenoids, which have anticancer, anti-inflammatory, antioxidant, and antidiabetic activities and are of great pharmacological value, especially in the treatment of central nervous system and cardiovascular diseases[2−7].

The *Crocus* genus, belonging to Iridaceae, comprises approximately 100 perennial and seasonal flowering species that grow from corms[8]. Several classifications of the *Crocus* genus have been proposed based on morphological characteristics, flowering habits, and molecular evidence[8−11]. Phylogenetic analysis using the barcodes from chloroplast and nuclear loci determined that in the whole genus *Crocus*, the classification of section *Crocus* and *Nudiscapus*, which could be further separated into different series, is acceptable[8]. *C. sativus*, a sterile triploid species ($2n = 3x = 24$), is clustered in series *Crocus* of section *Crocus*, and the phylogenetic relationships among the *taxa* of *Crocus* series *Crocus via* chloroplast genome, genome-wide DNA polymorphism, genome survey, and multicolor fluorescent *in situ* hybridization revealed that triploid *C. sativus* originated from the autotriploidization of wild *Crocus cartwrightianus*[9,12,13]. Vegetative propagation and cultivation of *C. sativus* by corms prevented genetic segregation of the favorable traits of saffron, resulting in the worldwide cultivation of a unique clonal lineage.

Crocin biosynthesis and *in vitro* production are subjects of considerable interest[14−19]. Crocins are also enriched in the mature fruits of *Gardenia jasminoides*, and the flowers of *Buddleja davidii*. Previously, we reported the genome of *G. jasminoides* and fully elucidated the biosynthetic pathway of crocins (crocin I−V), including one carotenoid cleavage dioxygenase (GjCCD4a), one aldehyde dehydrogenase (GjALDH2C3), and two UDP-glucosyltransferases (GjUGT74F8 and GjUGT94E5)[14]. Crocin biosynthesis in saffron stigmas is initiated by carotenoid cleavage dioxygenase 2 (CsCCD2), which cleaves zeaxanthin at the 7,8/7′,8′ positions to produce crocetin dialdehyde[20,21]. The aldehyde dehydrogenase CsALDH3I1[22], the UDP-glucosyltransferase (UGT) CsUGT74AD1[22], and CsUGT91P3[23] perform, respectively, the dehydrogenation of crocetin dialdehyde to crocetin, and its glycosylation to crocins I−V. The homologous or more efficient genes involved in these crocin biosynthetic genes in *C. sativus* remain unclear. Owing to its large genome size and autotriploidization, high-quality whole-genome sequencing of *C. sativus* has not yet been achieved. Although short-read and long-read based transcriptomes have been reported and used to identify crocin biosynthetic genes in *C. sativus*[20,22,24−26], the gene structure, gene cluster, collinearity, and duplication events related to crocin biosynthesis, or the developmental mechanism of *Crocus*, are difficult to determine. Furthermore, the saffron genome will benefit mining and utilizing genes related to crocin biosynthesis and elucidating the genetic basis for crocin production in distantly related plants.

Here, we sequenced and assembled the haploid genome of triploid *C. sativus*, which is the first chromosome-level genome of Iridaceae species. The whole genome duplication event showed that *C. sativus* underwent two lineage-specific whole genome triplication (WGT) events, and the retained genes after these WGT events were mainly enriched in the metabolic pathway, revealing that the evolution of saffron and glucoside synthesis are closely linked. Comparative genomics, gene duplication, and functional verification of CsCCDs in saffron revealed that *Crocus*-specific *CsCCD2* genes related to crocin synthesis originated from the duplication of CsCCD1. Furthermore, the functional convergence of GjCCD4a and CsCCD2 drives the independent evolution of crocin biosynthesis in eudicots (*Gardenia*) and monocots (*Crocus*). Homologous modeling, molecular docking, and site mutations revealed the key amino acids related to the catalytic activity of CsCCD2. The *C. sativus* genome provides important insights into the convergent evolution of crocetin/crocin biosynthesis in distantly related plants.

2. Results

2.1. *Crocus sativus* genome assembly and annotation

A total of 451.10 Gb long reads with an N50 length of 11.45 kb using a third-generation Sequel sequencing platform, and 600 Gb short reads from Illumina sequencing were generated for the triploid *C. sativus* genome with a high level of heterozygosity (Supporting Information Tables S1 and S2). The filtered long-reads were error-corrected, trimmed, and initially assembled, and an approximately 7.59 Gb genome with contig N50 length of 299.03 kb for *C. sativus* was produced (Supporting Information Table S3). The assessment of genome assembly showed that 95.70% completed BUSCOs were identified, suggesting a high completeness of genome assembly. However, a high proportion of duplicated BUSCOs (66.40%) and a genome size that was much larger than predicted represented the occurrence of triploid-fused assembly. The redundancy assembly was further purged to produce a haplotype genome of 4.79 Gb with a contig N50 length of 353 kb (Table 1, Table S3). The assembled *Crocus* genome size

Table 1 The statistics of assembly and annotation of *C. sativus*.

Assembly and annotation	*C. sativus*
Total number of contigs	26,118
Assembly size (Mb)	4769.31
Contig N50 (bp)	361,768
Total number of pseudo-chromosomes	8
Total anchored genome size (Mb)	4637.29
Complete BUSCO for genome assembly	97.5%
Number of protein-coding genes	60,656
Complete BUSCO for annotation	92.2%
Repeat density	65.57%

is much larger than the genome size of *Asparagus* (*Asparagus officinalis*, 1.18 G)[27] from Asparagaceae and *Apostasia* (*Apostasia shenzhenica*, 349 Mb)[28] from Orchidaceae. The longest contig was 16.50 Mb. The purged genome presents high completeness with 93.20% BUSCO mapping, indicating that the genome assembly is of high quality. Furthermore, a library of chromosome conformation capture techniques (Hi-C) was constructed, and 677.66 Gb sequencing reads covering 144 × of the assembled genome size were produced to cluster and order, a total of 27,709 contigs into 8 pseudochromosome-level scaffolds with 97.5% BUSCO mapping (Table 1, Fig. 1, Supporting Information Table S4, Fig. S1). The completeness of *Crocus* genome assembly is also superior to the assembly of *A. officinalis* (BUSCO, 88.2%) and *A. shenzhenica* (BUSCO, 93.62%).

Approximately 65.57% (3,127,423,950 bp) of the draft *C. sativus* genome was composed of transposable elements (TEs) (Table 1, Supporting Information Table S5). LTR-RTs (long terminal repeat retrotransposons) are the major type of TEs in plants, and *Copia* and *Gypsy* are the two most prominent superfamilies of LTR-RTs. Among them, 60.78% of the total genome was annotated as LTR-RTs, of which the *Gypsy* and *Copia* superfamilies accounted for 95.82% of the total LTR elements (Supporting Information Fig. S2, Table S5). The proportion of LTR-RTs in *C. sativus* was similar to that in *A. officinalis*, and significantly higher than that in *A. shenzhenica*, which only occupied 14.45% of their genomes (Supporting Information Tables S6 and S7). In the *Crocus* genome, the proportions of *Copia* and *Gypsy* elements are similar; however, *Gypsy* elements are dominant in *A. shenzhenica*, and the *Copia* superfamily is primary in *A. officinalis*, suggesting the lineage specificity of LTR-RT duplication (Supporting Information Fig. S3). Next, we examined the insertion time of the intact LTR-RTs in *Crocus*, *A. officinalis*, and *A. shenzhenica.* The results showed that insertion times of the *Copia* and *Gypsy* superfamilies proliferated rapidly at ∼0.1 MYA in the *Crocus* genome (Supporting Information Fig. S4). However, the insertion and expansion of LTR-RTs in *A. shenzhenica* (Supporting Information Fig. S5) and *A. officinalis* (Supporting Information Figs. S6 and S7) occurred much earlier. The more recent burst of LTR-RT insertion in *Crocus* might be correlated with the large expansion of genome size.

Furthermore, we predicted 60,656 protein-coding genes from the repeat-masked genome of *C. sativus* through a combination of homology-based prediction, *de novo* prediction, and transcriptome-based prediction methods (Table 1, Supporting Information Table S8). Complete orthologs for 94.8% of the embryophyta BUSCOs were identified, indicating that the predicted protein-coding genes are largely complete (Supporting Information Table S9). We identified 57,964 duplicated genes, which were grouped into five different categories, including 31,373 whole-genome duplicates (WGD, 54.17%), 5075 dispersed repeats (DSD, 8,76%), 2764 tandem duplicates (TD, 4.78%), 8340 proximal duplicates (PD, 14.39%), and 10,408 transposed duplicates (TRD, 17.96%) (Supporting Information Fig. S8A). We compared the K_A, K_S, and K_A/K_S distributions for different modes of gene duplication and found higher K_A/K_S values for tandem duplication gene pairs (Fig. S8B−D), suggesting that the process of tandem duplication is ongoing, with faster sequence changes and positive selection than for genes generated through other modes of duplication. We further performed functional enrichment of these different modes of gene duplication, and comparative analysis showed that the functional enrichment of TD was significantly different from that of DSD, PD, and TRD (Supporting Information Figs. S9−S12, Tables S10−S14). The functional annotation of TDs showed that the functional terms related to secondary metabolic biosynthesis, including monoterpene synthesis and carotenoid synthesis, were enriched (Fig. S12).

2.2. Phylogenomic analysis and whole genome duplications of *C. sativus*

Here, 27,239 orthologous groups covering 409,744 genes were identified for 14 angiosperms, including 13 monocots and *Vitis vinifera*. A total of 1590 orthologous groups representing the low-copy gene families were chosen to construct the phylogenetic tree using *V. vinifera* as an outgroup with coalescent and concatenated models. The topologies of these two phylogenetic trees are consistent with the bootstrap value of 100% for all the nodes. The final phylogenetic relationships of *C. sativus* with other candidate species are also consistent with the Angiosperm Phylogeny Group IV botanical classification system, and the results showed that *C. sativus* from Iridaceae is sister to the branch of *Allium sativum*[29] from Amaryllidaceae and *A. officinalis*[27] from Asparagaceae. Molecular dating using the nucleotide sequences of the 1590 low-copy genes and six fossil age calibrations predicted that Iridaceae species diverged from Amaryllidaceae and Asparagaceae approximately 95.53 million years ago (MYA), with a 95% confidence interval (CI) of 75.53−111.17 MYA. The divergence time between Orchidaceae (*A. shenzhenica*) and Iridaceae species was inferred to be 111.86 MYA with a 95% CI of 97.11 and 123.30 MYA. By comparing 13 other plant species, we found 10,218 and 5303 gene families that appeared to expand and contract, respectively, in *C. sativus* (Fig. 2A). Functional analysis of 2573 rapidly expanded gene families in *C. sativus* revealed the marked enrichment of genes related to metabolic pathways and secondary metabolites, including carotenoid biosynthesis, which might be related to the biosynthesis of active ingredients of *Crocus* (Supporting Information Fig. S13, Tables S15 and S16).

During the evolution of plants, genome expansion is mainly driven by whole-genome duplication events or polyploidy and the proliferation of transposable elements. These events are a ubiquitous feature of plant genomes, not only increasing the diversity of genome sizes but also enriching the genetic information. Here, we analyzed WGD events in *C. sativus* genome. Syntenic analysis showed at least one WGD event in the *C. sativus* genome (Supporting Information Fig. S14). The distributions of synonymous substitutions per synonymous site (K_S) for *C. sativus* paralogs detected obvious peaks in the *C. sativus* genome (Supporting Information Figs. S15 and S16). Next, 'ksrates' was

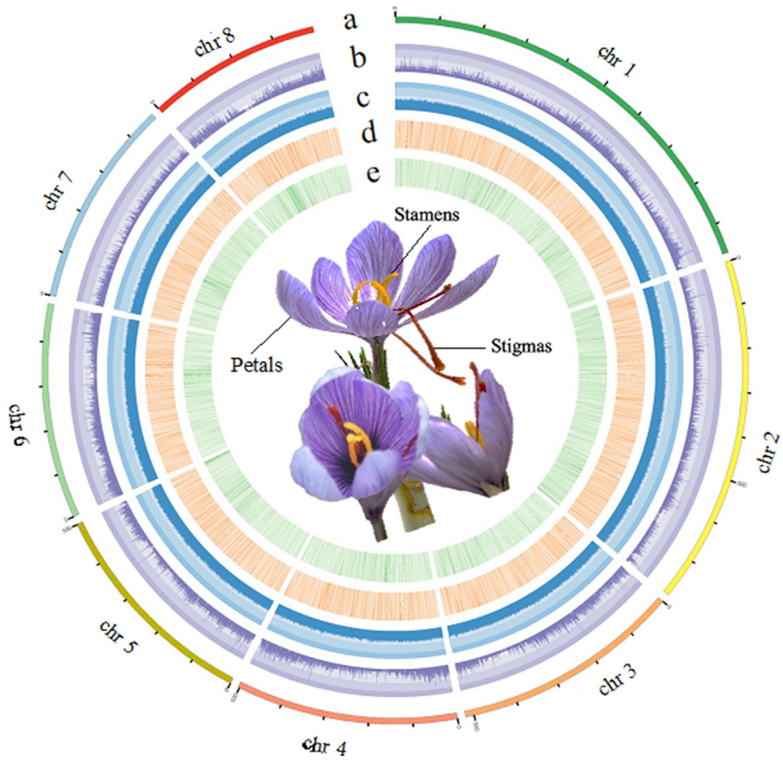

Figure 1 Genome features of the *C. sativus* genome. a, chromosome karyotypes in 100-kb windows; b, gene density; c, GC content; d, repeat sequence density of the *Gypsy* family; e, repeat sequence density of *Copia*. The middle part is the flower of *C. sativus*.

used to adjust the whole-genome duplication time according to the evolutionary rate. The distributions of K_S for all paralogous genes of *C. sativus* show three obvious peaks at 0.07 (a, Fig. 2B), 0.17 (b, Fig. 2B), and 0.39 (c, Fig. 2B), indicating that multiple WGD events have occurred (Fig. 2B). The paralogous genes localized at the K_S peak of 0.07 are interspersed repetitive sequences, which might be caused by explosive LTR insertion. The K_S distributions for homologous pairs between *C. sativus* and other species, *i.e.*, *A. sativum* and *A. officinalis* showed different values, 1.68 and 1.01, respectively, suggesting that these three species have evolved at different substitution rates. In addition, the species phylogeny indicated that *A. sativum* presented a larger branch length, that is, higher substitution rates, than *C. sativus* and *A. officinalis*. Then, the K_S value of divergence between *C. sativus* and *A. sativum*, and *A. officinalis* were further adjusted to 0.92 and 0.93, respectively (Fig. 2B and C). The two K_S peaks at 0.17 and 0.39 for paralogous genes of *C. sativus* are much less than the divergent K_S value with *A. sativum* and *A. officinalis*, suggesting that these two rounds of whole genome duplication events of *C. sativus* are species-specific. Furthermore, collinearity analysis combined with K_S values of the *C. sativus* genome revealed that two whole genome triplication (WGT) events occurred (Supporting Information Fig. S17). The K_S value of paralogous pairs for *A. sativum* and *A. officinalis*, respectively, also supported those two rounds of *A. officinalis*-specific WGD events and once *A. sativum*-specific WGD event occurred after their speciation (Supporting Information Figs. S18 and S19). Here, we estimated the mutation rate as 4.87×10^{-9} synonymous substitutions per synonymous site per year for *C. sativus*, *A. sativum*, and *A. officinalis* according to the divergent K_S value and time. Then, the *Crocus*-specific WGT events occurred at approximately 17.46 MYA (WGT-2) and 40.06 MYA (WGT-1), respectively (Fig. 2A). The slight K_S peak at 1.23 (d, Fig. 2B) for *Crocus* paralogous genes presented the ancestor whole genome duplication, which occurred before the spit with *A. sativum* and *A. officinalis*; in addition, the duplication time occurred at approximately 126.36 MYA, representing the tau (τ) duplication event. We further performed functional enrichment of WGD genes, and the results showed that metabolic and secondary metabolic biosynthesis were enriched, including terpenoid backbone and carotenoid biosynthesis, suggesting that WGD events play important roles in the biosynthesis of *Crocus* specialized metabolites (Supporting Information Fig. S20, Tables S17 and S18).

2.3. *Stigma-specific accumulation of apocarotenoids and coexpression network analysis*

The three stigmas of *C. sativus* exhibit a visible red color due to the presence of crocins, a kind of apocarotenoid. We collected samples from seven different organs of saffron: roots, peduncles, leaves, whole flowers, stamen, petals, and stigmas, to detect the accumulation of carotenoids and apocarotenoids. Based on targeted metabolome analysis of different organs and tissues, apocarotenoids (crocetin and five crocins) are largely enriched in stigmas and whole flowers; however, carotenoids such as β-carotene, zeaxanthin, and antheraxanthin are also detected in leaves with high accumulation (Fig. 3A).

We also mapped the RNA-seq reads from seven *C. sativus* organs to the assembled genome, and a total of 9430 genes (15.5%) were not expressed (FPKM <1) in any tested organ. Using *k*-means clustering, all the expressed genes were clustered into 48 clusters based on their expression profiles in different

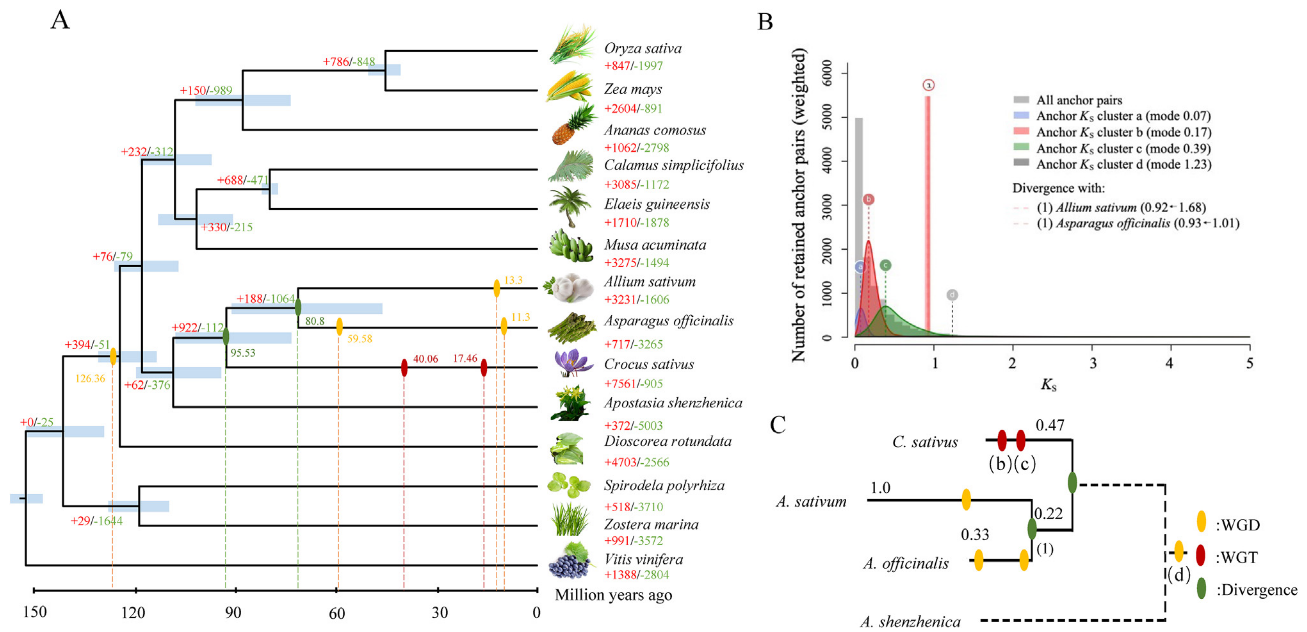

Figure 2 Evolution of the *C. sativus* genome. (A) Orthologous genes and phylogenetic analysis among *C. sativus* and 13 other angiosperms. Blue bars at nodes represent 95% credibility intervals of the estimated dates. The red and green numbers represent the expansion and contraction of gene families among 14 angiosperms. The ellipse in the branch represents the duplication events and divergence time. (B) K_S distributions of anchor pairs for the paralogs of *C. sativus*, and for the orthologs between *C. sativus* and *A. sativum*, and *C. sativus* and *A. officinalis*. The divergent K_S value of orthologs between *C. sativus* and *A. sativum* was adjusted from 1.68 to 0.92 (1), and the K_S value between *C. sativus* and *A. officinalis* was adjusted from 1.01 to 0.93 (1). Four K_S peaks for the paralogs of *C. sativus* named a, b, c, and d, were detected. (C) Phylogenetic tree for *C. sativus*, *A. sativum*, *A. officinalis*, and *A. shenzhenica* with the branch lengths of ortholog K_S distributions. The ellipse in the branch represents the duplication events. The red ellipse represents a WGT event, the yellow ellipse represents a WGD event, and the green ellipse represents the divergence. The dotted line represents the outgroup.

organs (Supporting Information Fig. S21). The functional enrichment for clusters 3, 5,13, 15, 17, 20, 34, and 43 that contained significantly or specifically expressed genes in stigmas showed that the terpenoid and carotenoid biosynthetic genes are significantly enriched (Supporting Information Fig. S22, Tables S19 and S20). In these clusters, 85 candidate transcription factors were further identified with FPKM >10. Furthermore, the co-expression network between structural genes in the carotenoid synthesis pathway and these transcription factors was performed, showed that MIKC_MADS, bHLH, and MYB might be related to the regulation of crocin biosynthesis and flower development (Supporting Information Fig. S23, Table S21).

2.4. Biosynthesis and evolution of carotenoids in Crocus

Carotenoids are lipid-soluble isoprenoids that act as the basis of plant pigments, such as flowers and fruits, and carotenogenesis genes have been completely elucidated in plants[30,31]. First, two molecules of geranylgeranyl diphosphate (GGPP) are condensed to form carotenoid skeleton phytoene by phytoene synthase (PSY), and phytoene is sequentially catalyzed to produce lycopene under a series of enzymes, including phytoene desaturase (PDS), 15-*cis*-ζ-carotene isomerase (Z-ISO), ζ-carotene desaturase (ZDS), and carotenoid isomerase (CRTISO)[32,33] (Fig. 3B). Here, we identified these coding genes related to upstream carotenoid biosynthesis in *Crocus* based on their homologs from *Arabidopsis thaliana*, and 7 *CsPSY*, 8 *CsPDS*, 5 *CsZISO*, 3 *CsZDS*, and 6 *CsCRTISO* genes were annotated (Supporting Information Table S22). PSY is the core rate-limiting enzyme in carotenoid biosynthesis[34], and this enzyme coding gene has evolutionarily expanded into a *PSY* gene family with gene numbers ranging from 1 to 13 in most monocot genomes, in contrast to one *AtPSY* from *A. thaliana*. Notably, *PSY* genes in *Crocus* expressed with tissue-specificity, such as *CsPSY5* and *CsPSY6*, are specifically and largely expressed in stigmas (Fig. 3C), and *CsPSY3* is uniquely transcribed in leaves. Additionally, the *CsPSY3*, *CsPSY5*, and *CsPSY6* genes showed high identities of 78.2%, 77.6%, and 75.5% with *AtPSY*, respectively, revealing their highly conserved phytoene synthase activity in plants (Table S22). In addition to the expansion of *CsPSY* genes, we also detected the expansion of *CsPDS*, *CsZISO*, *CsZDS*, and *CsCRTISO* genes in the *Crocus* genome, and these genes were distributed in different locations of the physical genome. In addition, we identified the significantly high expression of *CsZDS1*, *CsZDS2*, *CsCRTISO1-1*, and *CsCRTISO1-2* in stigmas and flowers in correlation with carotenoid accumulation in three stigmas of *Crocus*, suggesting their contribution to the coloration of stigmas (Table S22).

Beyond lycopene, carotenoid biosynthesis diverges into two branches with competing cyclization to δ-carotene and γ-carotene catalyzed by lycopene ε-cyclase (LCYE) and lycopene β-cyclase (LCYB), respectively[35]. Then, δ-carotene and γ-carotene are further transformed into α-carotene and β-carotene by LCYB, respectively. The hydroxylation of α-carotene and β-carotene produces different carotenoids, such as lutein, β-cryptoxanthin, zeaxanthin, etc, catalyzed by CYP97 family members (Fig. 3B). Here, we identified 3 *LCYE*, 2 *LCYB*, and 8 *CYP97* members in the *Crocus* genome (Table S22). Consistent with the observed coloration of stigmas, *CsLCYB2* presented

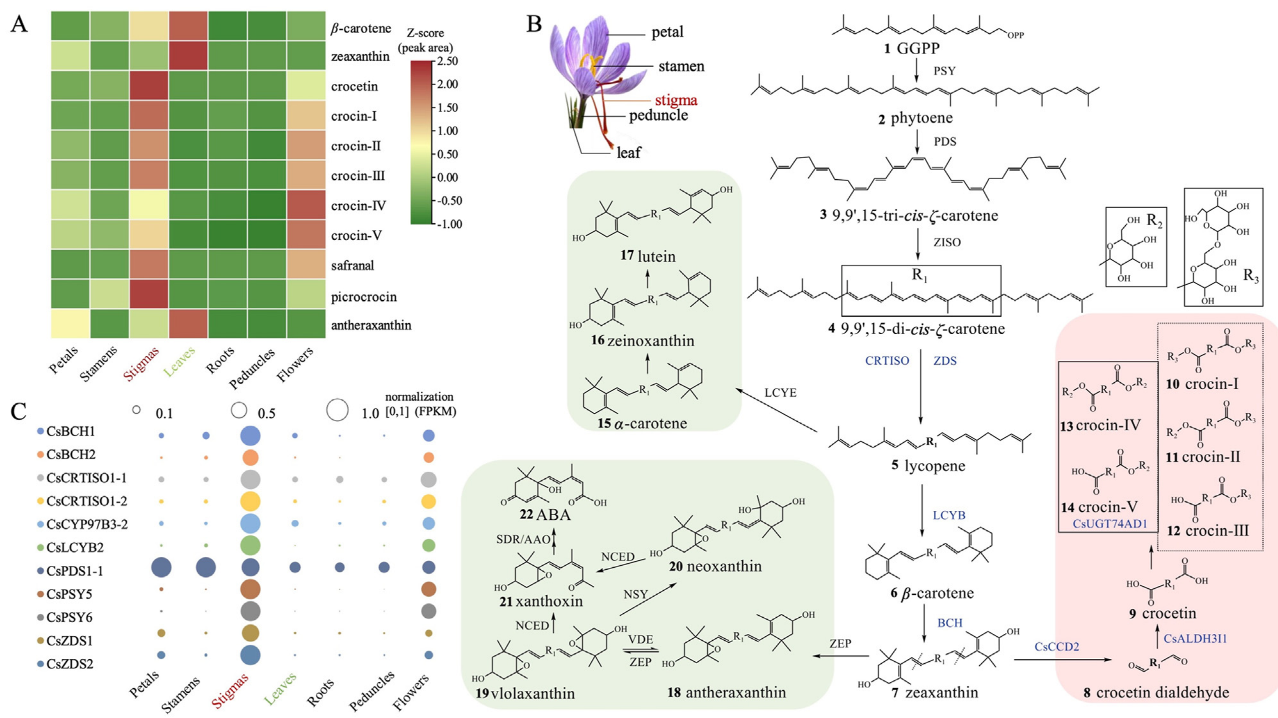

Figure 3 Proposed crocin biosynthesis and related gene expression in *C. sativus*. (A) Distribution of targeted carotenoids and apocarotenoids in different tissues of *C. sativus*. (B) The proposed biosynthetic pathways of carotenoids, abscisic acid (ABA), and crocins in *Crocus*. The pathway in the green box is the production of lutein and ABA biosynthesis, and the pink box is the biosynthesis pathway of crocins. (C) Gene expression of core genes involved in the biosynthetic pathways of carotenoids, ABA, and crocins in different tissues of *C. sativus*.

extremely high expression in *Crocus* flowers, particularly stigmas (Fig. 3C). However, silencing expression of *CsLCYE* genes in stigmas and specific expression of *CsLCYE1* and *CsLCYE2* in leaves of *Crocus* were detected, suggesting that the lutein pathway was inhibited in stigmas and activated in leaves. In addition, the *CYP97A* and *CYP97C* genes exhibited low expression in all tested tissues, and *CYP97B3-1* and *CYP97B3-2* showed specifically high expression in stigmas, indicating that *CYP97B* members might be responsible for the formation of β-cryptoxanthin and zeaxanthin in *Crocus*.

The 9-*cis*-epoxycarotenoid cleavage dioxygenase (NCED) family member NCED3 cleaves neoxanthin or violaxanthin to produce ABA (Fig. 3B). From the *Crocus* genome, 6 *NCED3* genes were identified; however, all *NCED3* genes showed silent expression in stigmas. Similarly, zeaxanthin epoxidase (ZEP) genes and neoxanthin synthase (NSY) genes in *Crocus* stigmas were expressed very slightly (FPKM <5), implying that ABA biosynthesis in *Crocus* stigmas might be significantly inhibited. This allows the metabolic flux towards the crocin pathway from carotenoids (*e.g.*, β-carotene, zeaxanthin, etc.) in stigmas.

The whole biosynthetic map of carotenoids from GGPP based on the *Crocus* genome was drawn (Fig. 3B), and compared to 12 other monocots, 2 eudicots (*V. vinifera* and *A. thaliana*) and *Amborella trichopoda*, we found significant expansion of carotenogenesis genes in *Crocus* (Table S23). Then, the carotenoid biosynthetic genes were mapped onto *Crocus* chromosomes (Table S22). Only *CsPDS2-2* and *CsPDS2-3* cluster with each other on chromosome 2, and there are no other tandem duplication genes for carotenoid biosynthesis in the *Crocus* genome. These results indicated that tandem duplication might not be the driving force of gene expansion related to carotenoid biosynthesis in *Crocus*. Distributions of K_S for all paralogous genes of carotenoid genes in *Crocus* showed that mean K_S values for three groups are consistent with the interspersed repetitive and two lineage-specific rounds of WGT events (Supporting Information Table S24), suggesting these gene duplication events contributed to the carotenoid biosynthesis. Additionally, the PD, TRD, and WGD events have played important roles in the expansion of carotenoid biosynthetic genes (Supporting Information Table S25).

2.5. *Catalytic activities and evolutionary origin of carotenoid cleavage dioxygenases in Crocus*

Crocins are the main valuable compounds in the stigmas of *Crocus* and the mature fruits of *Gardenia*, and crocin biosynthetic pathways in *Crocus* and *Gardenia* have been nearly completely elucidated. Generally, carotenoid cleavage dioxygenase (CCD), aldehyde dehydrogenase (ALDH), and UDP-glucosyltransferase (UGT) are responsible for crocin biosynthesis from carotenoids[35]. Here, we identified 20 *CCD* genes from the *Crocus* genome, including 6 *CCD1*, 1 *CCD1L*, 2 *CCD2*, 3 *CCD4*, 2 *CCD7*, and 6 *NCED* members (Supporting Information Table S26). Among them, two *CCD4* genes were highly expressed in flower tissues, such as petals, stamen, and stigmas, and silently expressed in roots, leaves, and peduncles (Fig. 4A). In *Crocus*, *CsCCD2* from *C. sativus* and *CaCCD2* from *Crocus ancyrensis* with 87% identity have been proven to cleave zeaxanthin into crocetin dialdehyde. Here, we found that two genome-annotated *CsCCD2* genes comprise the difference of three amino acids, site 151: I-L, site 230: K-E, and site 425: E-Q, and both genes are specifically expressed in stigmas of *C. sativus*, implying the same enzymatic activity (Supporting Information Fig. S24).

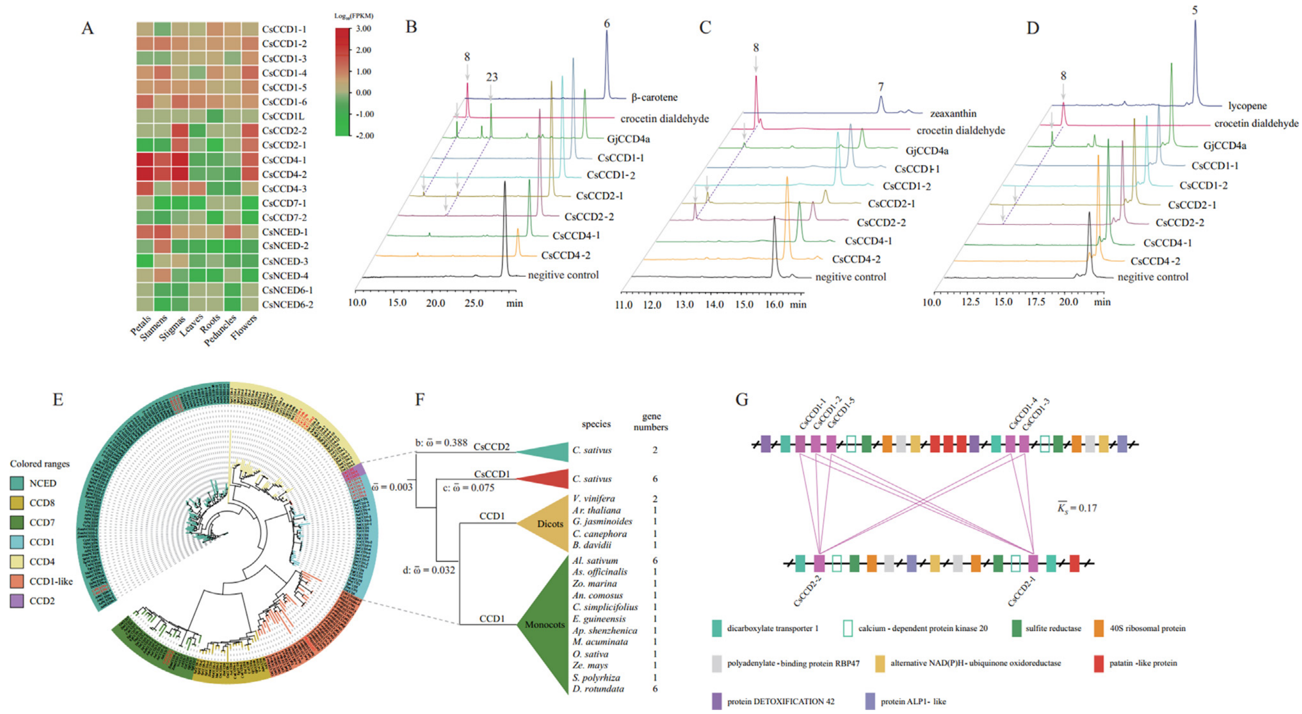

Figure 4 Functional identification and evolutionary analysis of CCD genes in *C. sativus*. (A) Gene expression patterns for *CsCCD* genes in different tissues of *C. sativus*. (B−D) The catalytic activities of candidate CsCCD genes, including CsCCD1-1, CsCCD1-2, CsCCD2-1, CsCCD2-2, CsCCD4-1, and CsCCD4-2, using *β*-carotene (B), zeaxanthin (C), and lycopene (D) as substrates, respectively. GjCCD4a was chosen as a positive control and has been indicated to cleave *β*-carotene (B), zeaxanthin (C), and lycopene (D) into crocetin dialdehyde, respectively. (E) Genome-wide identification and phylogenetic tree of the CCD gene family from the 14 angiosperms listed in Fig. 2A. (F) Phylogenetic topology of the CCD2 and CCD1 subfamilies. $\overline{\omega}$ represents the average ω ratio (K_A/K_S) of each branch. (G) Syntenic blocks for *CsCCD1* and *CsCCD2* genes in *C. sativus*. Different colored boxes represent the anchored gene pairs. The pink lines represent the synteny relationships between the *CsCCD1* and *CsCCD2* genes. The average K_S value of paralogous genes localized in the syntenic regions was calculated using CodeML of the PAML package.

We tested four highly expressed *CsCCD* genes, named *CsCCD4-1*, *CsCCD4-2*, *CsCCD2-1*, and *CsCCD2-2*, using carotenoid-producing *Escherichia coli* strains as host cells. GjCCD4a from *G. jasminoides* cleaved the 7,8 (7′,8′) positions of lycopene, *β*-carotene, and zeaxanthin into crocetin dialdehyde as a positive control (Fig. 4B−D). The results showed that *β*-carotene-producing strains with the plasmid containing *GjCCD4a*, *CsCCD2-1* and *CsCCD2-2*, respectively, produced two new peaks with high-performance liquid chromatography (HPLC) (Fig. 4B), and one new product with a retention time of 16.37 min and characteristic fragment ions ($[M+H]^+$: *m/z* 417.3152) that were the same as those of 8′-apo-*β*-carotenal. Another new peak at 11.88 min with characteristic fragment ions ($[M+H]^+$: *m/z* 297.1847) was identified as crocetin dialdehyde by comparison with the standard (Supporting Information Fig. S25). Zeaxanthin-producing strain containing *GjCCD4a*, *CsCCD2-1* and *CsCCD2-2*, respectively, also produced crocetin dialdehyde at 11.82 min (Fig. 4C). The lycopene-producing strain containing *GjCCD4a* could produce an obvious peak of crocetin dialdehyde at 11.82 min, and we could also observe slight product peaks for the strains containing *CsCCD2-1* and *CsCCD2-2*, respectively. The EIC extraction using LC-MS/MS confirmed that the peaks under the catalysis of *GjCCD4a*, *CsCCD2-1* and *CsCCD2-2* are identical to the fragment ions of crocetin dialdehyde (Fig. 4D, Fig. S25). These results suggest that CsCCD2-1 and CsCCD2-2 with 95.34% identity possess the same activity with GjCCD4a, which could cleave the 7−8/7′−8′ position of *β*-carotene, zeaxanthin, and lycopene. However, the cleavage activities for *CsCCD4-1* and *CsCCD4-2* were not detected at the different positions of *β*-carotene, zeaxanthin, and lycopene, although there is an unknown peak at 15.09 min for the catalysis of *CsCCD4-1* and *CsCCD4-2* to *β*-carotene. Additionally, the catalytic activities of two CCD1 enzymes (*CsCCD1-1* and *CsCCD1-2*) were detected, and the results showed that these two tested CCD1 enzymes could not perform any cleavage using lycopene, *β*-carotene, and zeaxanthin as substrates.

Compared with other monocots, *V. vinifera*, and *G. jasminoides*, CCD genes show a distinctive evolutionary profile. The phylogenetic tree showed that all *Crocus* CCD2 genes clustered into one branch, suggesting the lineage-specific evolution of CCD2 subfamily members (Fig. 4E). The CCD2 branch is sister to all CCD1 gene family members from the tested species, and the *Crocus* CCD1 genes are sister to the CCD1 genes of other monocots, *V. vinifera*, and *G. jasminoides*, revealing the rapid evolution of CsCCD1 genes (Fig. 4F). Generally, gene duplication has contributed to the evolution of novel gene functions related to adaptation[13]. Here, we found that the CsCCD2 genes presented significant genome synteny with CsCCD1 genes. Additionally, the average K_S value of paralogous genes localized in the syntenic regions was approximately 0.17, in accordance with the K_S value of WGT-2 (Fig. 4G). Our results supported that the CCD2 subfamily evolved from CsCCD1 members after the *Crocus*-specific

WGT event, and the rapid evolution of CsCCD2 gave rise to neofunctionalization to produce crocetin dialdehyde. The non-synonymous to synonymous rate ratios ($\omega = K_A/K_S$) for CCDs were analyzed using the two-ratio model of PAML, and the results showed that the ω values for the branches of CCD1 and CCD2 genes were less than 1 ($P < 0.01$), suggesting that CsCCD genes experienced purifying selection (Fig. 4F). Furthermore, we tested the site selection pressure of CsCCD2 using a branch-site model, and thirteen positive selection sites including sites, L174, T326, and N335 at the level $P > 99\%$ and sites S232, S257, E283, S292, H295, F341, S389, Q405, R450, and S508 at the level $P > 95\%$, were observed (Supporting Information Table S27). Our results showed that the positive selection of these amino acids might be related to the functional divergence of CsCCD2 genes.

2.6. *Identification of key amino acids responsible for the cleavage activity of CsCCD2-1*

Given the different catalytic activities of CsCCD1, CsCCD2, and CsCCD4, we aligned their amino acid sequences to identify the important residues related to CCD2 catalytic activity. Among the various residues, protein structure prediction with AlphaFold2 and molecular docking narrowed down the candidate residues, including I143, L146, R161, E181, T259, I260, S292, T326, M352, L353, and S364, which are distributed within 5 Å of the substrate, zeaxanthin (Fig. 5A and B). Among them, sites S292 and T326 have undergone positive selection. Molecular dynamics of the CCD2 and zeaxanthin complex revealed that eleven candidate residues are relatively rigid with low root-mean-square fluctuation (RMSF) values, suggesting their possible contributions to conserved substrate recognition or binding pocket stabilization (Fig. 5C).

To verify the importance of these amino acid residues, we constructed CsCCD2 mutants with a single substitution, and each mutant was transformed into zeaxanthin-producing strains for functional identification. Liquid chromatography showed that the I143A and E181A mutants led to an almost complete loss of the ability to cleave zeaxanthin, and only a slight peak for crocetin dialdehyde was detected (Fig. 5D and E). We observed that the I143 and E181 residues are distributed in the entrance of the CCD2 binding pocket, and the volumes of the protein binding pockets of the I143A and E181A mutants are obviously shrunken, suggesting that these two residues are crucial for the shape of the binding pocket (Supporting Information Fig. S26). In addition, the mutants L146A, R161A, T259A, and S292A significantly increased the catalytic activity of CCD2 toward zeaxanthin (Fig. 5D and E), and these residues are also distant from the active Fe^{II} center, suggesting their long-range effects on the catalytic center. Given the positive selection pressure of S292, we generated an additional mutant of S292F, and the result showed that the catalytic activity of S292F mutant was significantly reduced ($P < 0.05$, Fig. 5D and E).

3. Discussion

Asparagales, a monophyletic orchid order of flowering plants, is one of the largest angiosperm families, comprising 14 families and more than 36,200 species. This order contains crop plants (*e.g.*,

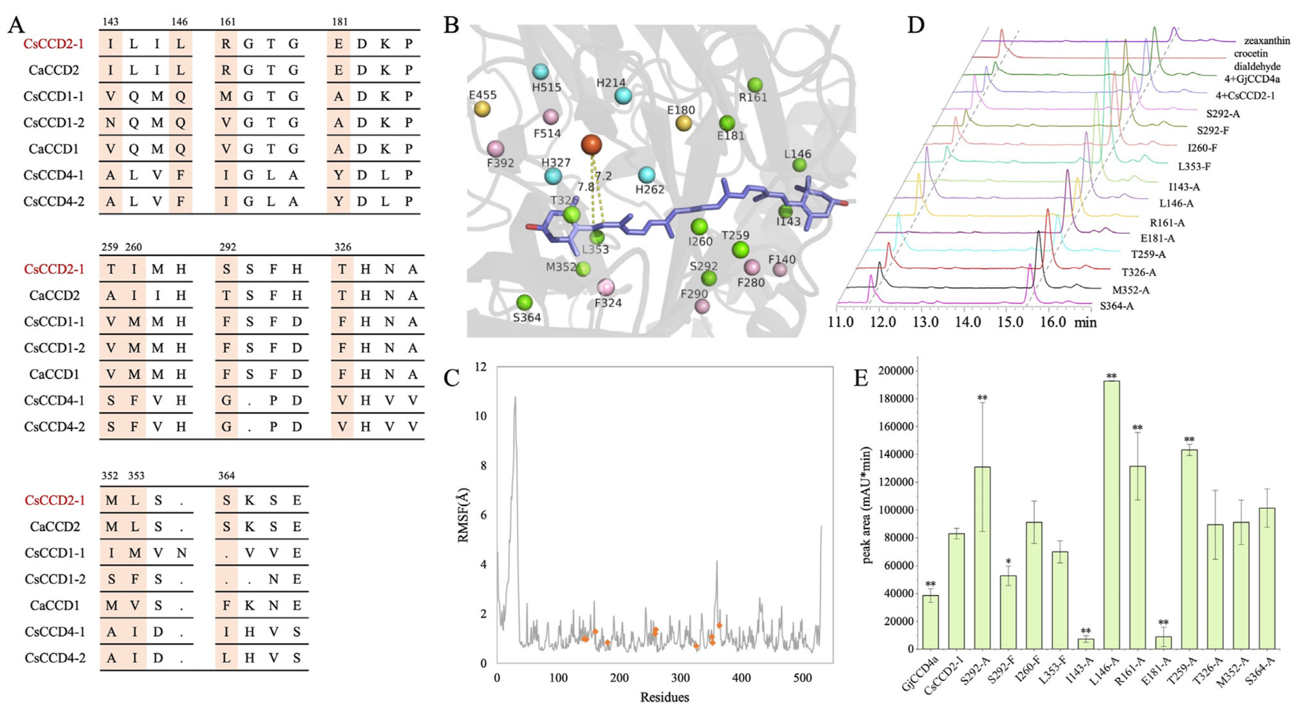

Figure 5 The identification of key amino acids responsible for the catalytic activity of CsCCD2-1. (A) Sequence alignment among CsCCD genes. (B) Molecular docking between CsCCD2-2 and zeaxanthin. The amino acids are selected around zeaxanthin within a 5 Å distance. (C) Molecular dynamics analysis of the CsCCD2-1 and zeaxanthin complex. The peaks represent the root-mean-square fluctuation (RMSF) value of all CsCCD2-1 amino acids. (D−E) The candite sites were mutated into alanine (Ala) to verify the catalytic activities of CsCCD2-1 mutants. (*) represents a significant difference ($P < 0.05$), and (**) represents an extremely significant difference ($P < 0.01$), compared with the catalytic activity of wild-type CsCCD2-1.

Allium and *Asparagus*), horticultural plants (*e.g.*, *Apostasia* and *Phalaenopsis*), and medicinal plants (*e.g.*, *Crocus* and *Dendrobium*). The genome size varies tremendously among Asparagales species, ranging from 0.38 pg (*Apostasia nuda*, Orchidaceae) to 75.90 pg (*Scilla mordakiae*, Asparagaceae) (https://cvalues.science.kew.org/). The genomes of representing species, such as *Allium*[29,36], *Asparagus*[27], *Apostasia*[28], *Phalaenopsis*[37], and *Dendrobium*[38], have been reported; importantly, our study sequenced the first chromosome-level genome of Iridaceae (*C. sativus*). The high BUSCO values for genome assembly and gene structure represented high completeness, which will be effectively applied to genome evolution and gene selection, although the continuity of contigs needs to be further improved using novel sequencing platforms, such as ultra-long reads and optical maps. The *C. sativus* genome and annotated coding genes will provide important genetic resources for evolutionary studies of Asparagales species. The phylogenetic tree supported the APGIV classification system, that Iridaceae species, *C. sativus*, is sister to Asparagaceae and Amaryllidaceae with divergence time of 95.53 MYA. The *Crocus*-specific LTR insertion after the split with *Allium* and *Asparagus* might have contributed to the large genome formation. Duplication events play important roles in the response to environmental stress and species diversity *via* gene dosage, subfunctionalization, neofunctionalization, and pseudogenization[39]. After the ancestral τ event shared by the majority of monocotyledons[40], at least two additional WGT events occurred in the *Crocus* genome. Through the adjusted K_S distributions of the *Allium* and *Asparagus* paralogs, two independent WGD events were identified in *A. officinalis*, and one WGD event was identified in *A. sativum*. These lineage-specific WGD events support their different biological processes, such as terpenoid and phenylpropanoid biosynthesis for *Crocus*. In addition, we determined that PD, TRD, and WGD gene duplication events were related to the duplication of carotenoid biosynthetic genes. This suggests their potential contribution to specific stigma development and crocin production.

Crocins are distributed in distant species, *Crocus* and *Gardenia*, and their evolutionary mechanism has attracted increasing attention[14]. CCD family members are responsible for the conversion of carotenoids into apocarotenoids, with CCD1 and CCD4 enzymes showing high diversity in substrate and cleavage specificities[41−43]. CCD1 family members have been reported to cleave a wide range of carotenoids at different positions 9,10; 9,10 (9′,10′); 5,6 (5′,6′), and CCD4 enzymes mainly cleave carotenoids at the unsymmetrical 9,10 or 7,8 positions[41]. In *C. sativus*, zeaxanthin is cleaved symmetrically at the 7,8 (7′,8′) positions by CsCCD2 to produce crocetin dialdehyde[20], while in *Gardenia*, the same reaction is carried out by GjCCD4a, which shares only 31% identity with CsCCD2, but different carotenoids, including β-carotene, lycopene, and zeaxanthin, can be accepted as substrates[14]. Importantly, *GjCCD4a* localized in the *Gardenia*-specific CCD4 tandem duplication, and the ancestral CCD4 conservatively cleaved the carotenoids at the unsymmetrical 9,10, however, duplicated *CCD4a* gene evolved the ability to produce crocetin dialdehyde. Additionally, our study indicated that CsCCD4 could not catalyze lycopene, β-carotene, and zeaxanthin to produce crocetin dialdehyde. These results suggested that the independent evolution of CsCCD2 and GjCCD4 contributed to crocin biosynthesis. Therefore, the convergent evolution of *CsCCD2* and *GjCCD4a* between *Crocus* and *Gardenia*, and the divergent evolution of *CCD4* genes between *Coffea* and *Gardenia* are the principal factors of crocin accumulation in a few distant and individual species. Notably, our results showed that CsCCD2-1 and CsCCD2-2 could also cleave the 7,8 (7′,8′) positions of all tested carotenoids, β-carotene, lycopene, and zeaxanthin, the same as the activity of GjCCD4a. However, the results presented different cleavage with previous findings, in which Frusciante et al., reported that CsCCD2 could not accept β-carotene and lycopene as substrates[20]. Recently, Wang et al., also supported the functional characteristics of CsCCD2, which can catalyze not only zeaxanthin cleavage but also β-carotene and lycopene cleavage[44].

Fang et al. also cloned the *FhCCD2* gene from *Freesia hybrida* of the Iridaceae family, and identified that *FhCCD2*, which showed 77.46% sequence identity with *CsCCD2*, could cleave zeaxanthin at the 7,8 (7′,8′) double bonds into the precursor substrate of crocin biosynthesis, the same activity as CsCCD2[45]. Our study indicated that the emergence of the ancestor CCD2 gene could be traced back to before the split between *Freesia* and *Crocus*. The divergence time of these two genera has been reported as 16.1 MYA with a 95% CI of 9.3 and 22.9 MYA (http://timetree.org), which is later than the WGT-2 event of *Crocus*, determining that the two rounds of WGT events are shared by *Freesia* and *Crocus*. Furthermore, our study indicated that the functional *CsCCD2* in *Crocus* evolved from the duplication of CsCCD1 genes, and the duplication occurred after the *Crocs* WGT-2 event. Therefore, we speculate herein that the functional CCD2 and crocin accumulation evolved after the WGT-2 event before the split between *Freesia* and *Crocus*. In addition, the replicon of CCD1 genes after the WGT-2 event experienced strongly positive selection to exhibit novel cleavage activity at the 7,8 (7′,8′) double bonds. Indeed, the catalytic mechanism of CCD1, CCD2, and CCD4, including the regioselective cleavage of substrates and enzyme promiscuity, needs to be deeply elucidated *via* the theories of quantum mechanics/molecular mechanics and enzyme-directed evolution.

4. Conclusions

In conclusion, we reported the first chromosome-level genome of the Iridaceae medicinal plant, *C. sativus*, providing important insights into the genome evolution and species speciation. Furthermore, using comparative genomic studies combined with *in vitro* assays, we reported the functional diversity and conservation of CCDs involved in crocin biosynthesis, and revealed the independent evolution of crocin biosynthesis between *Crocus* and *Gardenia*.

5. Experimental

5.1. *Plant materials*

The original *C. sativus* plants were collected from the planting base of Chongming Island, Shanghai. All independent tissues, including roots, peduncles, leaves, petals, stamens, stigmas, and flowers, were separated into three replicates for transcriptomes and metabolomes. High-quality DNA extracted from young leaves was used to construct different libraries for genome sequencing, including short insert fragments for Illumina sequencing and long fragments for SMRT sequencing.

5.2. *Sequel sequencing and genome assembly*

The high molecular weight genomic DNA of *C. sativus* was extracted as described for megabase-sized DNA preparation, and long DNA fragments (>20 kb) were selected using BluePippin. Long-read libraries were constructed following the protocols for the PacBio Sequel platforms (https://www.pacb.com/). The short-read libraries (300 and 500 kb) were constructed and sequenced using Illumina NovaSeq X Series. The raw reads from the Sequel platform were corrected, trimmed, and assembled by CANU (v2.0) with the default parameters[46]. Then, the redundant assembly of the FINAL contigs using the CANU pipeline was purged to improve the haploid assembly using purge_haplotigs with the default parameters[47]. The haplotigs were further polished by Illumina short reads three times using Pilon (v1.22) to improve the quality of genome assembly[48]. Finally, the completeness of the assembled genome was estimated by searching Benchmarking Universal Single-Copy Orthologs (BUSCO v4)[49]. Young leaves of *C. sativus* were fixed in 1% formaldehyde for Hi-C library construction. Cross-linked DNA was then lysed and digested using MboI restriction enzyme. DNA fragments were labeled with biotin and linked with blunt ends to construct an Illumina sequencing library. The clean sequencing data were mapped to the initial genome assembly, and valid Hi-C reads were used to correct the draft assembly. Then, the draft genome of *C. sativus* was assembled into chromosomes ($2n = 3x = 24$) using ALLHIC[50], and the mis-assembly was further manually corrected.

5.3. *Genome annotation and RNA-Seq analysis*

The RepeatModeler (v1.0.9) package was used to identify and classify the repeat elements of *C. sativus* genomes[51]. RepeatMasker (v4.0.6) was used to calculate and mask the repeat elements. The long terminal repeat retrotransposons (LTR-RTs) were identified by LTR_Finder (v1.0.6) and LTRharvest[52]. LTR_retriever was used to integrate the identification results. Protein-coding region identification and gene prediction were conducted through a combination of homology-based prediction, *de novo* prediction, and transcriptome-based prediction methods. Homologous proteins from six angiosperm genomes (*A. shenzhenica*[28], *A. officinalis*[27], *Dendrobium catenatum*[53], *Phalaenopsis equestris*[37], *Oryza sativa*[54] and *A. thaliana*[55]) were downloaded. Protein sequences were aligned to the assembly using GenBlastA (version 1.0.4)[56]. GeneWise (version 2.4.1)[57] was used to predict the exact gene structure of the corresponding genomic regions on each GenBlastA hit. Three *ab initio* gene prediction programs, Augustus (version 3.2.1)[58], GlimmerHMM (version 3.0.4)[59] and SNAP (version 2006-07-28)[60], were used to predict coding regions in the repeat-masked genome. Finally, RNA-seq data were mapped to the genome assembly using HiSAT2 (version 2.0.1)[61], StringTie (version 1.2.2)[62] and TransDecoder (version 3.0.1, https://github.com/TransDecoder/TransDecoder) were then used to assemble the transcripts and identify candidate coding regions into gene models. All gene models predicted from the above three approaches were combined by EvidenceModeler[63] into a nonredundant set of gene structures. Functional annotation of protein-coding genes was achieved using BLASTP (*E*-value 1e−05) against two integrated protein sequence databases: SwissProt and TrEMBL. Protein domains were annotated by using InterProScan (V5.30)[64]. The Gene Ontology (GO) terms for each gene were extracted with InterProScan. The pathways in which the genes might be involved were assigned by BLAST against the KEGG databases (release 84.0)[65], with an *E*-value cutoff of 1e−05.

5.4. *Ortholog detection and phylogenetic construction*

The amino acid sequences from *C. sativus* and 13 other angiosperms were clustered into orthologous groups using OrthoFinder (v.2.5.4)[66]. Low-copy genes were used to construct a phylogenetic tree using the RAxML package using the JTT + G + I substitution model for amino acid sequences with 1000 bootstrap replicates (v8.1.13)[67]. The eudicot *V. vinifera* was chosen as the outgroup. The divergence times of the tested species were calculated using the MCMCtree program based on the fossil-based age constraints[68]: 42−52 MYA for the divergence between *O. sativa*[54] and *Zea mays*[69], 80−85 MYA for the divergence between *Calamus simplicifolius*[70] and *Elaeis guineensis*[71], 103−134 MYA for the divergence between *Dioscorea rotundata*[72] and *O. sativa*, and 152−162 MYA for the divergence between *V. vinifera*[73] and *O. sativa*[54]. CAFÉ (v 3.1) was used to predict gene family expansion and contraction[74]. The expanded gene families were further annotated using KEGG and GO enrichment analyses.

5.5. *Whole genome duplication event analysis*

Syntenic blocks within *C. sativus* or between *C. sativus* and *A. officinalis*[27]/*A. sativum*[29] were identified based on paralogous or homologous gene pairs using MCscan (Python version)[75]. LAST was used to identify the homologs within the genome of *C. sativus*, and then to filter the gene pairs to remove tandem duplications and weak hits. A single linkage clustering is performed on the LAST output to cluster anchors into synteny blocks. For gene collinearity analyses, we also used BLASTP software to search the potential homologous gene pairs for each protein within and between genomes. Then, the BLAST results were selected as input for WGDI[75] to improve the synteny blocks. For inferences of WGD events, K_S values were estimated using the Nei-Gojobori method implemented in the YN00 program in the PAML (4.9 h) package[68]. Because the rate of evolution varies widely among species, we used ksrates[76] to correct the K_S values for the accurate identification of whole genome duplications and divergence times among tested species. First, ksrates was used to estimate the K_S values of one-to-one orthologs and paralogs. Then, rate adjustment was performed based on the phylogenetic tree with branch lengths equal to the K_S distances estimated from the ortholog K_S distributions.

5.6. *Co-expression network analysis between transcription factors and carotenoid biosynthetic genes*

The whole genes of *C. sativus* genome were clustered into various groups according to expression profiles in roots, peduncles, leaves, petals, stamens, stigmas, and flowers using the *k*-means clustering algorithm (R packages). The clusters that contained significantly or specifically expressed genes in stigmas were selected, and candidate transcription factors were further functionally annotated using PlantTFDB (v5.0) with default parameters[77]. The carotenoid biosynthetic genes of *C. sativus* were identified using the BLASTP (v2.11.0) search and the homologous evidence from *A. thaliana* as query sequences with an *E*-value cutoff of 1e−10. The co-expression networks between structural genes and transcription factors were constructed using Cytoscape (v3.7.1).

5.7. *Metabolome analysis of different crocus tissues using UPLC–MS/MS*

Fresh petals, stamens, stigmas, leaves, roots, peduncles, and flowers of *C. sativus* were respectively taken and powdered by adding liquid nitrogen, and 50% methanol was extracted by ultrasonic extraction for 30 min. The samples were analyzed by a 1290–6490 UPLC–QTOF (ESI) MS/MS system (Agilent) using an Acquity UPLC BEH C18 column (1.7 μm, 2.1 mm × 100 mm, Waters). The mobile phase was water with 0.1% formic acid (A) and acetonitrile (B) at a flow rate of 0.3 mL/min. The elution gradient was set as 0–5 min, 10%–50% B; 5–8 min, 50%–90% B; 10–11 min, 90%–100% B; 11–25 min, 100% B. Five microliters of 1 mg/mL filtered samples were injected into the systems. The following QTOF-MS parameters were used: scan range, 100–2000 Da; gas temperature, 350 °C; gas flow, 8 L/min; sheath gas temperature, 350 °C; sheath gas flow, 11 L/min; fragmentor, 120 V.

5.8. *Phylogenetic relationship and selection pressure analysis of CCD genes*

We downloaded the reported *A. thaliana* CCD genes, which were divided into the CCD1, CCD4, CCD7, CCD8, and NCED subfamilies[78]. The CCD gene family members in 13 angiosperm genomes were annotated using the BLASTP method with an *E*-value of 1e−5. The annotated CCD genes were manually filtered, and sequence alignment was performed using MAFFT (v7.487). The alignment gaps or poorly aligned regions were further removed using trimAl (v1.4. rev15). Then, IQ-TREE (v2.1.4-beta) was employed to construct the phylogenetic tree of CCD genes using JTT + R7 as the best-fit model with 1000 bootstraps. The selection pressure was estimated using the CodeML program of PAML packages using the phylogenetic tree as input. The tested CCDs were realigned *via* codon-based alignment using the ‘backtrans’ parameter of trimAl (v1.4. rev15). Here, branch models including the one-ratio model (M0) and two-ratio model (M2) of CodeML were used to calculate the ω value (K_A/K_S) of specific branches. ω values > 1, = 1, or <1 represent positive selection, neutral evolution, or purifying selection, respectively[68]. In addition, the branch-site model (Model 2, NSites = 2, fix_omega = 0, omega = 1.1) was employed to calculate the ω values for certain amino acid sites.

5.9. *Protein structure prediction, molecular docking, and molecular dynamics analysis*

We used AlphaFold2 to predict the protein structures of CsCCD2-1 and its mutants[79]. The resulting PDB files were visualized in PyMOL (v2.5). The structure of zeaxanthin was downloaded from the PubChem Data Bank in PDB format (PubChem CID: 5280899). AutoDock Vina software and AutoDock Tools were used for molecular docking analysis[80]. The AutoDockTools (ADT) graphical interface was applied to add the charges and polar hydrogens. The structures of the ligands were set to be routable in the ADT program, while the protein was kept as a rigid structure. AutoGrid was applied to the grid box, which was formed in the active site of the CsCCD2-1 protein with grid center coordinates $X = 1.599$, $Y = 1.502$, and $Z = -11.569$, and the sizes of x, y, and z are 33.0, 30.75, and 15.0, respectively. Out of the 9 different shapes obtained for each ligand–protein complex, and the best-ranked complex was examined by PyMOL (v2.5). GROMACS (v2018) was used to run all the simulations[81]. The ligand topologies were obtained with UCSF Chimera[82] and acpype[83]; and modeled with the all-atoms AMBER99SB force field[84]. The proteins were immersed in rectangular boxes filled with TIP3 water molecules, and then the necessary amount of counterions was added in electrostatically preferred positions, until the system was neutralized. The system underwent 50,000 steps of steepest descent energy minimization to remove steric overlap. Afterward, the systems were subjected to a two-step equilibration phase, namely NVT (number of particles, volume, and temperature) and NPT (number of particles, pressure, and temperature). The NVT equilibration was run for 100 ps (ps) to stabilize the temperature of the system, and the NPT was run for 100 ps to stabilize the pressure of the system by relaxing the system and keeping the protein restrained. All systems were subjected to a full 50 ns (ns) simulation under conditions of no restraints, an integration time step of 0.002 ps, and an xtc collection interval of 500 steps for 10 ps. The analyses of the trajectory files were performed using GROMACS utilities. The root mean square deviation (RMSD) was calculated using the parameter ‘gmx -rmsd’, and root mean square fluctuation (RMSF) analysis was performed using the parameter ‘gmx -rmsf’.

5.10. *Functional identification of CsCCDs and mutants*

The primers and plasmids used in this study are listed in Supporting Information Table S28. *CsCCD1-1*, *CsCCD1-2*, CsCCD2-1, CsCCD2-2, *CsCCD4-1*, and *CsCCD4-2* were ligated to the EcoR1/Kpn1 position of the pET32a series plasmids, respectively. The plasmid, which could produce carotenoids, including zeaxanthin, lycopene, and β-carotene, respectively, and the pET32a-*CCDs* plasmid were cotransferred into the BL21 (DE3) strain. The cells were induced with 0.3 mmol/L IPTG-induced and cultured overnight at 16 °C. Then, the cells were extracted with acetone, and the extracts were resuspended in ethyl acetate for HPLC and LC–MS/MS analyses. The primers for site-directed mutagenesis are listed in Supporting Information Table S29. The candidate sites of the CsCCD2-1 gene were specifically replaced with alanine (A) using the CloneExpress II one-step cloning kit. The recombinant plasmid pET32a-CCD2-mutants were further cotransferred with carotenoid-producing plasmids to verify the catalytic activities for six repetitions each.

Samples were analyzed using SHIMADZU LC-2050C instrument with a Thermo Scientific Hypersil GOLD C18 column (5 μm, 4.6 mm × 250 mm). The mobile phase was composed of water containing 0.1% formic acid (A) and acetonitrile (B). Different gradient elution procedures were performed. For zeaxanthin, the gradient was 10%–50% B at 0–5 min, 50%–90% B at 5–8 min, 90%–100% B at 8–10 min, and held at 100% B for 15 min and returned to 10% B within 1 min. For β-carotene and lycopene, the gradient was 10%–50% B at 0–5 min, 50%–90% B at 5–8 min, 90%–100% B in the next 2 min, held at 100% B for 30 min, and returned to 10% A in 1 min.

5.11. *Data availability*

The raw data of genome and transcriptome sequencing reported in this paper have been deposited in the Genome Sequence Archive in BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession number CRA007742 that are publicly accessible at http://bigd.big.ac.cn/gsa. The assembled genome and gene structures of *C. sativus* have been deposited in the Figshare (https://doi.org/10.6084/m9.figshare.21988667).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (81973424, 82073966, 82204346), the CAMS Innovation Fund for Medical Sciences (CIFMS) (Grant No. 2021-I2M-1-029, China), the National Key R&D Program of China (2023YFC3504800), the Fundamental Research Funds for the Central public welfare research institutes (ZZ16-YQ-047, ZZ16-ND-10-02, China), and the Key Scientific Research Foundation of the Higher Education Institutions of Anhui Province, China (KJ2021A0235).

Author contributions

Zhichao Xu and Jingyuan Song designed and coordinated the study. Zhichao Xu and Shanshan Chen assembled and annotated the genome. Xiaotong Wang and Hui Yao collected the samples. Shanshan Chen and Ya Tian conducted the bioinformatics analysis. Ranran Gao, Yalin Wang, Zishan Li, Xin Hua, Shengnan Tan, Tianyi Xin, Xiangdong Pu, and Wei Sun performed the experiments and analyzed the data. Zhichao Xu, Ranran Gao, and Jingyuan Song wrote and revised the manuscript.

Conflicts of interest

The authors declare no conflict of interest.

Appendix A. Supporting information

Supporting data to this article can be found online at https://doi.org/10.1016/j.apsb.2023.12.013.

References

1. Manzo A, Panseri S, Bertoni D, Giorgi A. Economic and qualitative traits of Italian Alps saffron. *J Mt Sci* 2015;**12**:1542−50.
2. Mykhailenko O, Kovalyov V, Goryacha O, Ivanauskas L, Georgiyants V. Biologically active compounds and pharmacological activities of species of the genus *Crocus*: a review. *Phytochemistry* 2019;**162**:56−89.
3. Ahmed S, Hasan M, Heydari M, Rauf A, Bawazeer S, Abu-Izneid T, et al. Therapeutic potentials of crocin in medication of neurological disorders. *Food Chem Toxicol* 2020;**145**:111739.
4. Bastani S, Vahedian V, Rashidi M, Mir A, Mirzaei S, Alipourfard I, et al. An evaluation on potential anti-oxidant and anti-inflammatory effects of Crocin. *Biomed Pharmacother* 2022;**153**:113297.
5. Boozari M, Hosseinzadeh H. Crocin molecular signaling pathways at a glance: a comprehensive review. *Phytother Res* 2022;**36**:3859−84.
6. Milani A, Basirnejad M, Shahbazi S, Bolhassani A. Carotenoids: biochemistry, pharmacology and treatment. *Br J Pharmacol* 2017;**174**:1290−324.
7. Kumar V, Kaur N, Wadhwa P. Clinical evidence on the effects of saffron (Crocus Sativus L.) in anxiety and depression. *World J Tradit Chin Med* 2022;**8**:181−7.
8. Harpke D, Meng S, Rutten T, Kerndorff H, Blattner FR. Phylogeny of *Crocus* (Iridaceae) based on one chloroplast and two nuclear loci: ancient hybridization and chromosome number evolution. *Mol Phylogenet Evol* 2013;**66**:617−27.
9. Nemati Z, Blattner FR, Kerndorff H, Erol O, Harpke D. Phylogeny of the saffron-crocus species group, *Crocus* series *Crocus* (Iridaceae). *Mol Phylogenet Evol* 2018;**127**:891−7.
10. Alsayied NF, Fernández JA, Schwarzacher T, Heslop-Harrison JS. Diversity and relationships of *Crocus sativus* and its relatives analysed by inter-retroelement amplified polymorphism (IRAP). *Ann Bot* 2015;**116**:359−68.
11. Frello S, Ørgaard M, Jacobsen N, Heslop-Harrison JS. The genomic organization and evolutionary distribution of a tandemly repeated DNA sequence family in the genus *Crocus* (Iridaceae). *Hereditas* 2004;**141**:81−8.
12. Nemati Z, Harpke D, Gemicioglu A, Kerndorff H, Blattner FR. Saffron (*Crocus sativus*) is an autotriploid that evolved in Attica (Greece) from wild *Crocus cartwrightianus*. *Mol Phylogenet Evol* 2019;**136**:14−20.
13. Schmidt T, Heitkam T, Liedtke S, Schubert V, Menzel G. Adding color to a century-old enigma: multi-color chromosome identification unravels the autotriploid nature of saffron (*Crocus sativus*) as a hybrid of wild *Crocus cartwrightianus* cytotypes. *New Phytol* 2019;**222**:1965−80.
14. Xu Z, Pu X, Gao R, Demurtas OC, Fleck SJ, Richter M, et al. Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. *BMC Biol* 2020;**18**:63.
15. Diretto G, López-Jiménez AJ, Ahrazem O, Frusciante S, Song JY, Rubio-Morage A, et al. Identification and characterization of apocarotenoid modifiers and carotenogenic enzymes for biosynthesis of crocins in *Buddleja davidii* flowers. *J Exp Bot* 2021;**72**:3200−18.
16. Zheng X, Mi J, Balakrishna A, Liew K, Ablazov A, Sougrat R, et al. *Gardenia* carotenoid cleavage dioxygenase 4a is an efficient tool for biotechnological production of crocins in green and non-green plant tissues. *Plant Biotechnol J* 2022;**20**:2202−16.
17. Liu TL, Yu S, Xu ZC, Tan JT, Wang B, Liu YG, et al. Prospects and progress on crocin biosynthetic pathway and metabolic engineering. *Comput Struct Biotechnol J* 2020;**18**:3278−86.
18. Ahrazem O, Diretto G, Argandoña J, Rubio-Moraga A, Julve JM, Orzáez D, et al. Evolutionarily distinct carotenoid cleavage dioxygenases are responsible for crocetin production in *Buddleja davidii*. *J Exp Bot* 2017;**68**:4663−77.
19. Pu X, He CN, Yang Y, Wang W, Hu KZ, Xu ZC, et al. *In vivo* production of five crocins in the engineered *Escherichia coli*. *ACS Synth Biol* 2020;**9**:1160−8.
20. Frusciante S, Diretto G, Bruno M, Ferrante P, Pietrella M, Prado-Cabrero A, et al. Novel carotenoid cleavage dioxygenase catalyzes the first dedicated step in saffron crocin biosynthesis. *Proc Natl Acad Sci U S A* 2014;**111**:12246−51.
21. Ahrazem O, Rubio-Moraga1 A, Berman J, Capell T, Christou P, Zhu CF, et al. The carotenoid cleavage dioxygenase CCD2 catalysing the synthesis of crocetin in spring crocuses and saffron is a plastidial enzyme. *New Phytol* 2016;**209**:650−63.
22. Demurtas1 OC, Frusciante S, Ferrante P, Diretto G, Azad NH, Pietrella M, et al. Candidate enzymes for saffron crocin biosynthesis are localized in multiple cellular compartments. *Plant Physiol* 2018;**177**:990−1006.
23. López-jimenez AJ, Frusciante S, Niza E, Ahrazem O, Rubio-Moraga Á, Diretto G, et al. A new glycosyltransferase enzyme from family 91, UGT91P3, is responsible for the final glucosylation step of crocins in saffron (*Crocus sativus L.*). *Int J Mol Sci* 2021;**22**:8815.
24. Yue J, Wang R, Ma XJ, Liu JY, Lu XH, Thakar SB, et al. Full-length transcriptome sequencing provides insights into the evolution of apocarotenoid biosynthesis in *Crocus sativus*. *Comput Struct Biotechnol J* 2020;**8**:774−83.
25. Ahrazem O, Argandoña1 J, Fiore A, Rujas A, Rubio-Moraga Á, Castillo R, et al. Multi-species transcriptome analyses for the regulation of crocins biosynthesis in *Crocus*. *BMC Genom* 2019;**20**:320.

26. Tan H, Chen XH, Liang N, Chen RB, Chen JF, Hu CY, et al. Transcriptome analysis reveals novel enzymes for apo-carotenoid biosynthesis in saffron and allows construction of a pathway for crocetin synthesis in yeast. *J Exp Bot* 2019;**70**:4819−34.
27. Harkess A, Zhou JS, Xu CY, Bowers JE, Hulst RV, Ayyampalayam S, et al. The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nat Commun* 2017;**8**:1279.
28. Zhang GQ, Liu KW, Li Z, Lohaus R, Hsiao YY, Niu SC, et al. The *Apostasia* genome and the evolution of orchids. *Nature* 2017;**549**: 379−83.
29. Sun X, Zhu S, Li N, Cheng Y, Zhao J, Qiao X, et al. A chromosome-level genome assembly of Garlic (*Allium sativum*) provides insights into genome evolution and allicin Biosynthesis. *Mol Plant* 2020;**13**: 1328−39.
30. Quian-Ulloa R, Stange C. Carotenoid biosynthesis and plastid development in plants: the role of light. *Int J Mol Sci* 2021;**22**:1184.
31. Badejo AA. Elevated carotenoids in staple crops: the biosynthesis, challenges and measures for target delivery. *J Genet Eng Biotechnol* 2018;**16**:553−62.
32. Cazzonelli CI, Pogson BJ. Source to sink: regulation of carotenoid biosynthesis in plants. *Trends Plant Sci* 2010;**15**:266−74.
33. Hirschberg J. Carotenoid biosynthesis in flowering plants. *Curr Opin Plant Biol* 2001;**4**:210−8.
34. Zhou XS, Rao S, Wrightstone E, Sun TH, Lui ACW, Welsch R, et al. Phytoene synthase: the key rate-limiting enzyme of carotenoid biosynthesis in plants. *Front Plant Sci* 2022;**13**:884720.
35. Li ZR, Ahn TK, Avenson TJ, Ballottari M, Cruz JA, Kramer DM, et al. Lutein accumulation in the absence of zeaxanthin restores non-photochemical quenching in the *Arabidopsis thaliana npq1* mutant. *Plant Cell* 2009;**21**:1798−812.
36. Liao NQ, Hu ZY, Miao JS, Hu XD, Lyu XL, Fang HT, et al. Chromosome-level genome assembly of bunching onion illuminates genome evolution and flavor formation in allium crops. *Nat Commun* 2022;**13**:6690.
37. Cai J, Liu X, Vanneste K, Proost S, Tsai WC, Liu KW, et al. The genome sequence of the orchid *Phalaenopsis equestris*. *Nat Genet* 2015;**47**:65−72.
38. Niu ZT, Zhu F, Fan YJ, Chao Li, Zhang BH, Zhu SY, et al. The chromosome-level reference genome assembly for *Dendrobium officinale* and its utility of functional genomics research and molecular breeding study. *Acta Pharm Sin B* 2021;**11**:2080−92.
39. Panchy N, Lehti-Shiu M, Shiu SH. Evolution of gene duplication in plants. *Plant Physiol* 2016;**171**:2294−316.
40. Jiao YN, Li JP, Tang HB, Paterson AH. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* 2014;**26**:2792−802.
41. Dhar MK, Mishra S, Bhat A, Chib S, Kaul S. Plant carotenoid cleavage oxygenases: structure−function relationships and role in development and metabolism. *Brief Funct Genomics* 2019;**19**:1−9.
42. Daruwalla A, Kiser PD. Structural and mechanistic aspects of carotenoid cleavage dioxygenases (CCDs). *BBA-MOL Cell Biol L* 2020; **1865**:158590.
43. Hou X, Rivers J, León P, McQuinn RP, Pogson BJ. Synthesis and function of apocarotenoid signals in plants. *Trends Plant Sci* 2016;**21**: 792−803.
44. Wang Y, Li SQ, Zhou Z, Sun LF, Sun J, Shen CP, et al. The functional characteristics and soluble expression of saffron CsCCD2. *Int J Mol Sci* 2023;**24**:15090.
45. Fang Q, Li YQ, Liu BF, Meng XY, YangZZ, Yang S, et al. Cloning and functional characterization of a carotenoid cleavage dioxygenase 2 gene in safranal and crocin biosynthesis from *Freesia hybrida*. *Plant Physiol Biochem* 2020;**154**:439−50.
46. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly *via* adaptive *k*-mer weighting and repeat separation. *Genome Res* 2017;**27**:722−36.
47. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinf* 2018;**19**:460.
48. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;**9**: e112963.
49. Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: assessing genomic data quality and beyond. *Curr Protoc* 2021;**1**:e323.
50. Zhang XT, Zhang SC, Zhao Q, Ming R, Tang HB. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat Plants* 2019;**5**:833−45.
51. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AJ, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* 2020;**117**:9451−7.
52. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 2007;**35**:265−8.
53. Zhang GQ, Xu Q, Bian C, Tsai WC, Yeh CM, Liu KW, et al. The *Dendrobium catenatum* Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. *Sci Rep* 2016;**6**:19029.
54. Du H, Yu Y, Ma YF, Gao Q, Cao YH, Chen Z, et al. Sequencing and *de novo* assembly of a near complete indica rice genome. *Nat Commun* 2017;**8**:15324.
55. Wang B, Yang XF, Jia YY, Xu Y, Jia P, Dang NX, et al. High-quality *Arabidopsis thaliana* genome assembly with nanopore and HiFi long reads. *Dev Reprod Biol* 2022;**20**:4−13.
56. She R, Chu JS, Wang K, Pei J, Chen N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res* 2009;**19**:143−9.
57. Birney E, Clamp M, Durbin R. GeneWise and genomewise. *Genome Res* 2004;**14**:988−95.
58. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res* 2006;**34**:435−9.
59. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* 2004;**20**:2878−9.
60. Korf I. Gene finding in novel genomes. *BMC Bioinf* 2004;**5**:59.
61. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**:357−60.
62. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* 2019;**20**:278.
63. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol* 2008;**9**:R7.
64. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. *Nucleic Acids Res* 2005;**33**:W116−20.
65. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M, et al. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:29−34.
66. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019;**20**:238.
67. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;**30**:1312−3.
68. Yang ZH. Paml 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007;**24**:1586−91.
69. Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science* 2009;**326**:1112−5.
70. Zhao HS, Wang SB, Wang JL, Chen CH, Hao SJ, Chen LF, et al. The chromosome-level genome assemblies of two rattans (*Calamus simplicifolius* and *Daemonorops jenkinsiana*). *GigaScience* 2018;**7**:1−11.

71. Singh R, Ong-Abdullah M, Low ET, Manaf MAA, Rosli R, Nookiah R, et al. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature* 2013;**500**:335−9.
72. Zhang YM, Chen M, Sun L, Wang Y, Yin J, Liu J, et al. Genome-wide identification and evolutionary analysis of NBS-LRR genes from *Dioscorea rotundata*. *Front Genet* 2020;**11**:484.
73. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007;**449**:463−7.
74. Bie TD, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 2006;**22**:1269−71.
75. Sun PC, Jiao BB, Yang YZ, Shan LX, Li T, Li XN, et al. WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol Plant* 2022;**15**:1841−51.
76. Sensalari C, Maere S, Lohaus R. *ksrates*: positioning whole-genome duplications relative to speciation events in *k*s distributions. *Bioinformatics* 2021;**38**:530−2.
77. Jin JP, Tian F, Yang DC, Meng YQ, Kong L, Luo JC, et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res* 2016;**45**:1040−5.
78. Tan BC, Joseph LM, Deng WT, Liu LJ, Li QB, Cline K, et al. Molecular characterization of the arabidopsis 9-*cis* epoxycarotenoid dioxygenase gene family. *Plant J* 2003;**35**:44−56.
79. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583−9.
80. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 2010;**31**:455−61.
81. Páll S, Zhmurov A, Bauer P, Abraham M, Lundborg M, Gray A, et al. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. *J Chem Phys* 2020;**153**:134110.
82. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 2004;**25**:1605−12.
83. Silva AWS, Vranken WF. ACPYPE-antechamber python parser interfacE. *BMC Res Notes* 2012;**5**:367.
84. Ponder JW, Case DA. Force fields for protein simulations. *Adv Protein Chem Struct Biol* 2003;**66**:27−85.