BMC Plant Biology

RESEARCH

Open Access

Assembling the chromosome-level genome of *Arctium Lappa* L. provides new insights into carotenoid biosynthesis

Xingxing Feng[1], Kuanqiang Tang[2], Yanxing Wang[1], Na Zhang[1], Shuo Li[1], Liang Luo[1], Enqi Liu[1], Anhui Chen[1], Shuai Wang[1], Tingguo Kang[3], Deqiang Dou[3], Chuanyin Hu[4] and Yuwei Dong[1*]

Abstract

Background *Arctium lappa* L. (*A. lappa*), a plant of significant economic importance, is cultivated worldwide for its applications in food and medicine, presenting a compelling opportunity for genetic research within the *Asteraceae* family. Despite the importance of *A. lappa* genetic resources, the genome of the Peixian yellow *A. lappa* cultivar has not yet been sequenced.

Results This study presents a chromosome-level genome of the Peixian yellow *A. lappa*, the genome consists of a comprehensive 1.69 Gb reference and a scaffold N50 of 91.02 Mb. By utilizing Hi-C data, we anchored 99.12% of the assembled sequences onto 18 chromosomes and identified 40,665 protein-coding genes. The genome is predominantly composed of repetitive sequences, which constitute 68.32%, with long terminal repeats (LTRs) accounting for 65.66% of the total. A comparative genomic analysis with the Gansu *A. lappa* genome revealed significant genetic variations, including 2,972,809 single nucleotide polymorphisms (SNPs), 608,644 insertions/deletions (INDELs), and 11,800 structural variations (SVs). Furthermore, transcriptomic analyses comparing Peixian yellow *A. lappa* and Liu chuan li xiang *A. lappa* have shed light on the molecular mechanisms underlying carotenoid biosynthesis in Peixian yellow *A. lappa*. A total of 16 differentially expressed genes (DEGs) were specifically associated with the carotenoid biosynthesis pathway. Among these, several genes, including *PSY1*, *PDS*, *BCH2*, *VDE*, *NCED1*, and *NCED2*, were downregulated in Peixian yellow *A. lappa*. In contrast, other genes such as *DXS*, *IDI1*, *CCD8B*, *CCS*, and *CCD4* were upregulated. Additionally, several transcription factors linked to carotenoid biosynthesis were identified, including *MYC2*, *WRK35*, and *HY5*. Most of these transcription factors were downregulated in Peixian yellow *A. lappa*, with the exception of *Arcla.08g00830*, which encodes *MYC2*.

Conclusion This high-quality genomic resource is poised to enhance genome-informed breeding strategies and improve *A. lappa* quality, offering valuable insights for breeding programs aimed at enhancing desirable traits in *A. lappa*.

Keywords A. lappa, Genome assembly, HiFi, Carotenoid biosynthesis

*Correspondence:
Yuwei Dong
dongyuwei66@163.com
[1]School of Food and Bio-engineering, Xuzhou University of Technology, Xuzhou 221018, China
[2]College of Life Sciences and Agriculture and Forestry, Qiqihar University, Qiqihar 161006, China
[3]School of Pharmacy, Liaoning University of Traditional Chinese Medicine, Dalian 116600, China
[4]Xuzhou Tian Ma Jing An Food., Ltd, Xuzhou 221600, China

Introduction

Arctium lappa L. (*A. lappa*), known as "Niubang" in China, is an edible and medicinal plant belonging to the *Asteraceae* family [1]. In Europe and Japan, the fleshy root of *A. lappa* is consumed as a vegetable, while its fruit has long been utilized for medicinal purposes in China [2]. The fleshy root is rich in essential nutrients, including inulin, polysaccharides, lignans, carbohydrates, phenolic compounds, and carotene [3, 4]. Despite its nutritional and medicinal benefits, research into the molecular genetics of *A. lappa* lags behind that of other closely related *Asteraceae* plants, primarily due to the limited availability of genome assembly and annotation data [5–8]. Establishing a high-quality genome assembly is crucial for uncovering genetic variations that elucidate the genetic and molecular foundations of desirable plant traits [9]. Therefore, generating a high-quality genome assembly from a distinct *A. lappa* cultivar is essential for identifying genetic variations that contribute to trait diversity.

Genomic structural variations (SVs), including insertions/deletions (InDels), single-nucleotide polymorphisms (SNPs), duplications, and inversions, represent vital sources of genetic variation that influence key agronomic traits in crops [10, 11]. However, technological limitations and the lack of high-quality reference genomes have hindered the identification of genetic variants, particularly SVs, among *A. lappa* species. To address this issue, it is necessary to develop additional reference genomes from a broader range of accessions. This will facilitate the identification of SVs and their evolutionary dynamics, clarifying their roles in shaping essential horticultural traits relevant to *A. lappa* domestication and breeding.

Peixian yellow *A. lappa* is a cultivar specific to China, characterized by its golden yellow skin, which has gained popularity among consumers due to its high carotenoid content. Carotenoids not only serve as precursors to vitamin A but also play a crucial role in protecting vision and mitigating cardiovascular and cerebrovascular diseases [12]. The biosynthesis of carotenoids in plants has been extensively studied and documented [13, 14]. Recent studies have examined the regulation of carotenoid biosynthesis pathways in fruits and vegetables by various transcription factors (TFs), including the *bHLH*, *WRKY*, and *bZIP* families [15–17]. These transcription factors play critical roles in regulating the expression of carotenoid-related genes, influencing carotenoid accumulation during plant development and in response to environmental factors. For instance, the transcription factor *SlWRKY35* in tomato fruits activates the expression of the *SlDXS1* gene, promoting carotenoid accumulation [17]. In citrus fruits, the transcription factor *CsMPK6* interacts with *CsMYC2*, a member of the *bHLH* family, reducing its ability to bind promoter regions, which affects carotenoid biosynthesis. Additionally, WRKY proteins bind to W-box elements in promoter regions, regulating genes involved in carotenoid synthesis [17, 18]. *bZIP* transcription factors, characterized by their conserved *bZIP* domains, specifically recognize and bind to cis-acting elements such as the G-box (CACGTG) and C-box (GACGTC). Despite extensive research on carotenoid biosynthesis in various plant species, the regulatory genes involved in carotenoid biosynthesis in *A. lappa* remain largely unexplored.

This study focuses on the Chinese-specific cultivar Peixian yellow *A. lappa*, for which a high-quality chromosome-scale genome assembly was developed. Comparative genomic analysis revealed SVs between the Peixian yellow *A. lappa* and the Gansu *A. lappa* genomes. Establishing this high-quality reference genome not only illuminates the genetic diversity and evolutionary history of *A. lappa* but also provides a foundation for genomic selection of desirable traits in *A. lappa* and other species within the *Asteraceae* family.

Materials and methods

Plant materials and sequencing libraries preparations

Whole-genome sequencing was performed on individual plants of the Peixian yellow *A. lappa*, a Chinese-specific cultivar grown in Peixian, Jiangsu Province. A single plant of Peixian yellow *A. lappa* were used for sequencing and they were clonal. For genome assembly, genomic DNA was extracted from the young leaves of Peixian yellow *A. lappa* using a DNAsecure Plant Kit (Tiangen Biotech, Beijing, China), and DNA quality was detected by agarose gel electrophoresis. The A260/A280 ratio of the DNA was 1.81, as measured by the Qubit Fluorometer.

For single-molecule real-time (SMRT) sequencing, PacBio libraries were constructed following the standard protocols outlined by Pacific Biosciences [19]. Briefly, high-quality genome DNA was sheared to ~20 kb, the damaged ends were repaired and ligated with a blunt-end adaptor, and the resulting libraries were sequenced using the PacBio Sequel platform.

Hi-C libraries were constructed following a modified standard protocol [20]. Leaves of Peixian yellow *A. lappa* were crosslinked with 4% formaldehyde, and chromatin was extracted and digested with 400U of Dpn Ⅱ restriction enzyme at 37 °C. Cross-linked fragments were labeled with biotin-14-dCTP, followed by blunt-end repair and DNA ligation with T4 DNA ligase (NEB). After reverse crosslinking and DNA purification, the fragments were processed by fragmentation, end repair, and enrichment of biotin-labeled Hi-C samples using streptavidin C1 magnetic beads. Finally, Illumina paired-end sequencing adapters were ligated to the fragments.

Quality control of the Hi-C libraries was performed before sequencing on an Illumina PE150 platform.

Genome assembly and quality assessment

The genome size, heterozygosity ratio, and repeat sequence ratio of the Peixian yellow *A. lappa* genome were analyzed through *k-mer* distribution analysis (*k* = 17) using Illumina short reads which were generated through whole genome sequencing (WGS), following the method described by Marcais et al. [21].

Draft genome assembly was performed using Hifiasm (v0.16.1) (https://hifiasm.readthedocs.io/en/latest/index.html) [22] and subsequently polished with Racon and Nextpolish (v1.2.4) (https://github.com/Nextomics/NextPolish) based on PacBio HiFi reads and Illumina short-reads using Pilon (v1.2.2) (https://github.com/broadinstitute/pilon/releases/tag/v1.24) [23]. Hi-C data were used to assemble the chromosome-level genome using ALLHiC (v0.9.8) software (https://github.com/tangerzhang/ALLHiC).

For anchored contigs, clean read pairs were generated from the Hi-C library and mapped to the polished Peixian yellow *A. lappa* genome via BWA (v0.7.17-r1188) (https://github.com/lh3/bwa/releases) [24] with the default parameters. Paired reads that were mapped to a different contig were used for Hi-C associated scaffolding. The percentage of valid Hi-C read pairs successfully mapped to the genome was 16.34% and 25.51%, respectively.

To evaluate the quality of the genome assembly, several complementary methods were employed. First, genomic completeness was assessed using conserved plant genes from the Benchmarking Universal Single-Copy Ortholog (BUSCO v3.0.2) (https://busco.ezlab.org/) and Core Eukaryotic Genes Mapping Approach (CEGMA v2.5) databases (https://bioweb.pasteur.fr/packages/pack@CEGMA@v2.5) [25, 26]. Second, Illumina short reads were mapped back to the assembled genome using BWA (v0.7.8) [24] to calculate coverage rates and average sequencing depth. The quality of the genome assembly was further examined using the LTR Retriever package (v1.0.7) (https://github.com/oushujun/LTR_retriever) [27].

Construct circos plot

The Circos software was used to generate the circos plot [28]. First, prepare the input files for each track, ensuring that the display style (e.g., heatmap, line, etc.) and color scheme for each track are specified appropriately. Once the files are ready, configure the plot by following the software's guidelines, and then execute the command: circos -conf circosall. conf -outputfile niubang. circus.

Gene prediction and genome annotation

Protein-coding genes in the Peixian yellow *A. lappa* genome were identified through a combination of *de novo* gene prediction, homology-based prediction, and transcriptome data integration [29]. Repeat contents were determined using a hybrid approach of *de novo* and homology-based predictions. Repeat content identification: initially, a *de novo* repeat database was constructed using LTR (v1.06) (https://github.com/xzhub/LTR_Finder), RepeatScout (v1.0.5) (https://github.com/Dfam-consortium/RepeatScout), and RepeatModeler (v2.0.1) (http://www.repeatmasker.org/RepeatModeler/) [30]. Repeat elements were then identified using TRF (v4.09) (https://tandem.bu.edu/trf/trf.html) [31] based on this datab ase. For homology-based identification, RepeatMasker (v4.1.0) (https://www.repeatmasker.org/) [32] and Prot einMask (v4.1.0) [32] were utilized. The results from *de novo* and homology-based predictions were combined to comprehensively annotate repeat elements across the genome.

For RNA-seq-based prediction, transcriptome sequences from six tissues-root, leaf, petiole, seed, flower, and hypocotyledonary axis were aligned to the genome using TopHat (v2.1.1) (https://ccb.jhu.edu/software/tophat/index.shtml) [33]. Gene prediction was carried out by integrating evidence from homology-based, *de novo*, and transcript-assisted methods. Homology-based prediction: tools such as BlastAll (v2.2.26), Solar (v0.9.6), and GeneWise (v2.4.1) were employed for protein alignment. *De novo* prediction: gene models were predicted using five ab initio tools-Augustus (v3.2.3) (https://anaconda.org/bioconda/augustus) [34], Geneid (v1.4) (https://genome.crg.cat/software/geneid/index.html), Genescan (v1.0) (http://genes.mit.edu/GENSCAN.html), Glimmer (v3.0.4) (https://ccb.jhu.edu/software/glimmerhmm/), and SNAP (v2013.11.29) (https://snap.stanford.edu/) [35].

Transcript-assisted prediction: transcriptome data supported the refinement of gene models. The identified gene models from these approaches were integrated using EVM (v1.1.1) (http://evidencemodeler.github.io/) [36] to produce a comprehensive and high-confidence gene set. Gene annotation: gene functions were annotated by comparing predicted protein sequences against the SwissProt database (https://www.uniprot.org/) [37], TrEMBL, and NCBI non-redundant protein databases, selecting the best hits for functional annotation. Protein motifs and domains were identified using InterProScan (v5.35–74.0.0) (http://www.ebi.ac.uk/interpro/) [38] with reference to the InterPro (v29.0) database (https://www.ebi.ac.uk/interpro/) [39]. Functional classifications for genes were assigned based on Gene Ontology (GO) terms (http://www.geneontology.org/) [40] and Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.kegg.jp/) [41] pathways using the best-matched hits.

tRNAs were annotated using tRNAscan-SE (v1.4) [42]. rRNAs, miRNAs, and snRNAs: Predicted using INFERNAL (v1.1.2) (https://anaconda.org/bioconda/infernal) [43] with the Rfam database (release 9.1) (https://rfam.org/).

Comparative genomics and genome evolution analysis

To investigate the evolutionary history of the Peixian yellow *A. lappa*, we conducted a comparative genomic analysis using the genomes of 14 other plant species, including *Helianthus annuus* (GenBank: GCA_002127325.2) [44], *Lactuca sativa* (GenBank: GCA_002870075.4) [45], *Cynara cardunculus* (GenBank: GCA_001531365.2) [46], *Cichorium intybus* (GenBank: GCA_023525715.1) [47], *Artemisia annua Linn* (GenBank: GCA_003112345.1) [48], *Daucus carota* (GenBank: GCA_001625215.2) [49], *Solanum tuberosum* (GenBank: GCA_000226075.1) [50], *Solanum lycopersicum* (GenBank: GCA_036512215.2) [51], *Capsicum frutescens* L. (GenBank: GCA_011745845.1) [52], *Coffea canephora* (GenBank: GCA_049114775.1) [53], *Arabidopsis thaliana* (GenBank: GCA_000001735.2) [54], and *Oryza sativa* L.(GenBank: GCA_034140825.1) [55].

To analyze genome evolution, orthologous gene families, including single-copy and multicopy families, were identified across the Peixian yellow *A. lappa* and other genomes using OrthoMCL (http://orthomcl.org/orthomcl/) [56]. Single-copy gene families were aligned using MUSCLE (http://www.drive5.com/muscle/) [57], generating a superalignment matrix that served as the basis for phylogenetic tree construction.

Maximum likelihood (ML) phylogenetic trees were constructed using RaxML (http://sco.hits.org/exelixis/web/software/raxml/index.html) [58]. Divergence times among species were estimated using Mcmctree (http://abacus.gene.ucl.ac.uk/software/paml.html) [59], with calibration points derived from the TimeTree database (http://www.timetree.org/) and a previous report [60]. To identify gene family expansions and contractions, CAFÉ (v4.2) (https://hahnlab.github.io/CAFE/) [61] was employed.

Analysis of WGD events and genome synteny

WGD events and synteny in the Peixian yellow *A. lappa* genome were analyzed using MCScanX (http://chibba.pgml.uga.edu/mcscan2/). Paralogs were identified through pairwise analysis, and WGD events were inferred based on the distribution of fourfold synonymous third-codon transversion (4DTv) and synonymous substitution rate (*Ks*) values for each paralogous pair.

Collinearity between genomes was also analyzed using MCScanX [62] with default parameters. To detect syntenic regions, protein sequences were first aligned in an all-versus-all manner using BLASTP, and collinear paralogous pairs were identified through MCScanX [62].

Structural variant identification

To identify SVs, the Peixian yellow *A. lappa* genome was aligned to the Gansu *A. lappa* genome (GenBank number: GCA_023525745.1) [47] assembly using the MUMmer4 toolkit (https://mummer4.github.io/install/install.html) [63] with the parameters --mum -c 100 -l 100. Alignment blocks were filtered using delta-filter with parameters -i 80 -l 100 -1 to ensure high-quality alignments.

Subsequently, SVs such as SNPs, InDels, tandem expansions, repeat expansions, tandem contractions, and repeat contractions were identified using MUMmer4. Functional annotations of these SVs were performed using the ANNOVAR package.

(https://annovar.openbioinformatics.org/en/latest/) [64], enabling the association of SVs with potential functional impacts. The criteria for filtering SNPs, small InDels, and SVs are as follows: SNPs, small InDels (< 50 bp), and SVs (structural variations, ≥ 50 bp, including large insertions, deletions, and inversions).The SVs in carotenoid biosynthesis genes between Peixian yellow *A. lappa* and Gansu *A. lappa* were identified based on the carotenoid biosynthesis pathway from RNA-seq data.

Comparative transcriptome analysis of fleshy roots from Peixian yellow *A. lappa* and Liu chuan li xiang *A. lappa* species

Peixian yellow *A. lappa* is a cultivar specific to China, grown in Peixian, Jiangsu Province, and is characterized by its yellow epidermis. Liu chuan li xiang *A. lappa* is one of the main varieties cultivated in Japan, distinguished by its cylindrical root shape, smooth and straight bar-like appearance, and black epidermis. Fleshy roots from two *A. lappa* varieties were harvested and immediately frozen in liquid nitrogen. Total RNA was extracted using the RNeasy Plant Mini Kit (Qiagen). The RNA Integrity Numbers (RIN) for the RNA-seq samples were 9.60, 9.00, 10.00, 9.90, 10.00, and 9.80, respectively. Transcriptome profiling was carried out as previously described [65]. Briefly, three biological replicates per sample were sequenced on the Illumina NovaSeq 6000 (Illumina, USA). Raw RNA reads were filtered and trimmed to eliminate low-quality sequences, resulting in clean reads. These reads were then aligned to the draft reference genome using HISAT2 (https://github.com/DaehwanKimLab/hisat2) with default parameters. Gene-level read counts were obtained using FeatureCounts [66].

The Fragments Per Kilobase of transcript per Million mapped reads (FPKM) value for each gene was calculated based on the gene length and the number of reads mapped to it, providing a normalized measure of gene expression. To identify DEGs, DESeq2 was used

for normalization and statistical analysis. Multiple testing correction was performed using the Benjamini–Hochberg method. Genes with an adjusted *p*-value (P-adj) < 0.05 were considered differentially expressed. To explore the functional significance of DEGs, GO enrichment analysis was conducted using the GOseq R package, which accounts for gene length bias. GO terms with a corrected *p*-value < 0.05 were deemed significantly enriched. Additionally, KOBAS software.

(http://bioinfo.org/kobas) was employed to test for the enrichment of DEGs in KEGG pathways, with pathways showing a q-value < 0.05 considered significantly enriched.

The carotenoid biosynthesis genes in Peixian yellow *A. lappa* were identified from the carotenoid biosynthesis pathway. The bioinformatics tools used are listed in Table S1.

RNA isolation and real-time quantitative PCR analysis

Total RNA was isolated from the fleshy roots of two *A. lappa* varieties using the TransZol Plant Kit (TRANS, Beijing, China) following the manufacturer's instructions. cDNA was synthesized using the FastQuant RT Kit (with gDNase) (TIANGEN, Beijing, China) according to the manufacturer's protocol. Relative gene expression was quantified by RT-qPCR using the FastStart Universal SYBR Green Master (ROX) (Roche, Mannheim, Germany) on a Stratagene MX3005P Real-Time PCR System (Applied Biosystems, Waldbronn, Germany), following the manufacturer's guidelines. Three independent biological replicates were performed for each sample, and relative expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method, with normalization to β-actin. Primer sequences are listed in Table S22.

Results

Chromosome-level high-quality genome assembly of the Peixian yellow *A. lappa*

We assembled a chromosome-level genome of *A. lappa* through *de novo* sequencing of Peixian yellow *A. lappa*, a prominent local cultivar from Jiangsu Province, China. This cultivar is distinguished by its golden-yellow skin, which contrasts with the brown-skinned varieties, such as Liu chuan li xiang *A. lappa* (Fig. 1a and Fig. S1).

Initially, *k-mer* analysis (k = 17) estimated the genome size to be 1.69 Giga base-pairs (Gb), with a heterozygosity rate of 0.24% (Fig. S2). A total of 108 G Illumina reads were generated, providing a sequence coverage of 64.09×. Repetitive sequences, accounting for 75.78% of the genome, are derived from *k*-mer analysis (Table S2). We then generated 161 Gb of PacBio HiFi reads (95.54× genome coverage) (Table S3) and assembled the contigs using Hifiasm [22]. The resulting assembly had a total contig length of 1.71 Gb, with a contig N50 of 74.25 mega base-pairs (Mb). To improve the assembly's structural accuracy, we integrated Hi-C paired-end reads, which allowed us to anchor the contigs into 18 chromosomes, accounting for 99.12% of the assembled genome (Fig. 1b and c). This Hi-C-based assembly demonstrated exceptional quality, further supported by chromatin interaction data, achieving a scaffold N50 value of 91.02 Mb (Table S4; Table S5).

Multiple assessments confirmed the high quality of the assembled Peixian yellow *A. lappa* genome. BUSCO analysis indicated that 98.2% of the detected gene models were complete, underscoring the genome's exceptional completeness (Table S6 and Fig. S3). Additionally, CEGMA results revealed the presence of 243 core eukaryotic genes, further supporting the assembly's completeness (Table S7). A GC depth scatter plot confirmed the absence of contamination in the assembly (Fig. S4). The LTR Assembly Index (LAI) score of 20.14, derived from LTR annotation, the QV was 50.9691, met the standards for a reference-quality genome.

Notably, a comparison with the genomes of two other *A. lappa* species [47, 67] (Table S8) revealed that the Peixian yellow *A. lappa* exhibits the highest BUSCO score, indicating superior genome completeness. Furthermore, the contig N50 of the Peixian yellow *A. lappa* is 74 Mb, significantly higher than the 6.88 Mb observed in the Dongbei *A. lappa* [67]. This 10-fold increase in contig continuity suggests that the Peixian yellow *A. lappa* assembly is likely more accurate at both the contig and scaffold levels.

For genome annotation, complementary approaches were employed, including *de novo* gene prediction, protein-based homology searches, and RNA sequence-aided predictions. A total of 40,665 genes were predicted, with an average gene length of 4,355.36 bp and an average coding sequence length of 1,105.64 bp. Each gene contained an average of 4.50 exons, with an average exon length of 245.74 bp. The total gene count was comparable to that of *L. sativa* (Table S9). Furthermore, the Peixian yellow *A. lappa* genome encoded 2,162 microRNAs (miRNAs), 1,005 transfer RNAs (tRNAs), 7,684 ribosomal RNAs (rRNAs), and 2,860 small nuclear RNAs (snRNAs) (Table S10). Functional annotation assigned functions to 39,814 genes (97.91%), with 50.80% of the genes categorized into Gene Ontology (GO) terms and 74.88% mapped to KEGG pathways (Table S11; Fig. S5).

The Peixian yellow *A. lappa* genome provides insights into its evolution history. Polyploidization, driven by whole-genome duplication (WGD) events and the proliferation of transposable elements (TEs), was a key factor in genome size expansion, consistent with findings in other plants [68]. Repetitive sequences constituted 68.32% of the genome (Table S12). Among these, LTRs were the most prevalent, accounting for 65.66%, followed

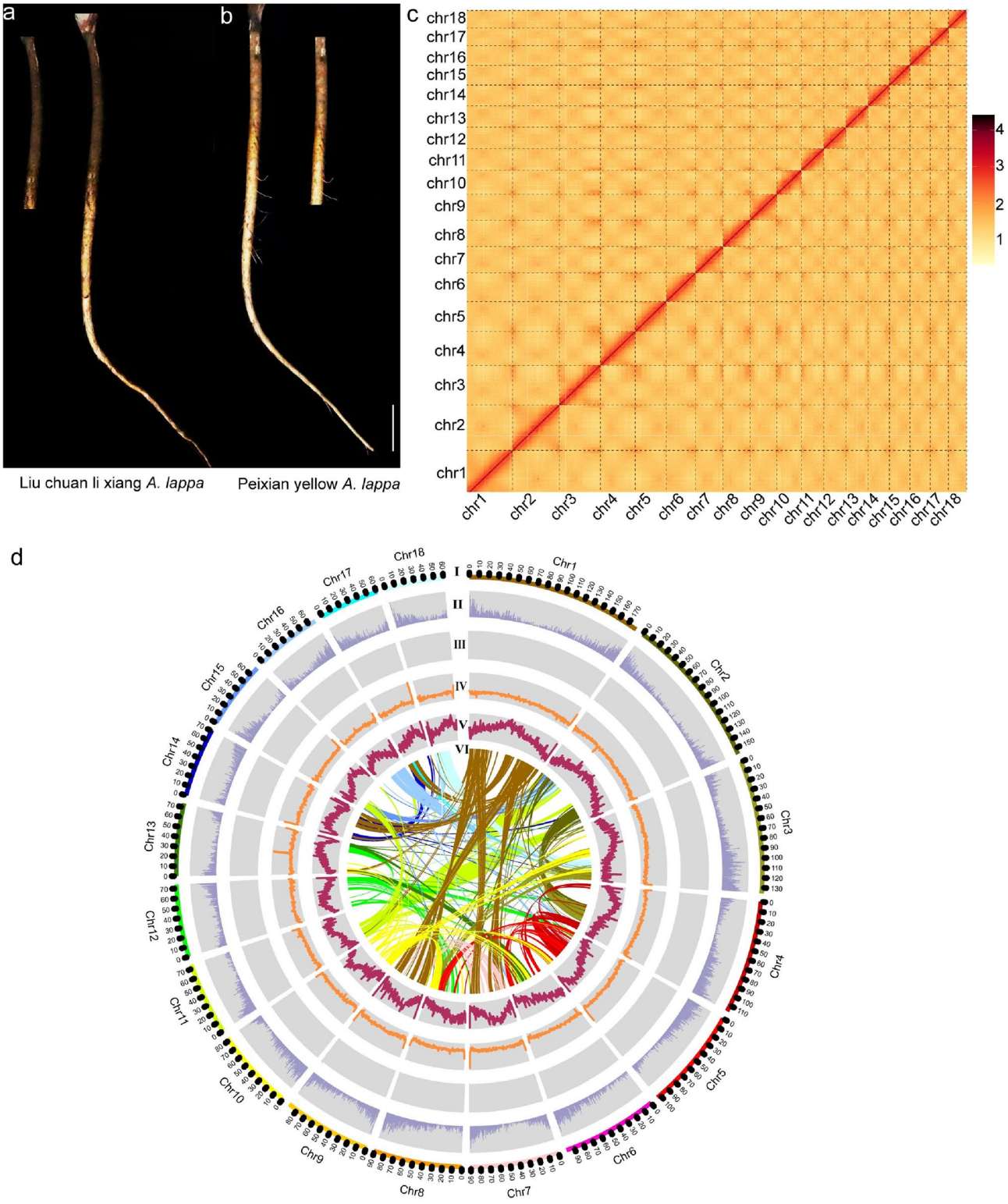

Fig. 1 Morphological and genomic characteristics of the *A. lappa* plant. **a** Phenotype of Liu chuan li xiang *A. lappa* displaying its characteristic brown skin. **b** Phenotype of the Peixian yellow *A. lappa*, notable for its golden yellow skin. Bar = 5 cm. **c** Genome-wide chromatin interaction map of the 18 chromosomes in the Peixian yellow *A. lappa* genome, illustrating the high-quality Hi-C data that supports chromosome-level assembly. **d** Circos plot depicting genomic features of the Peixian yellow *A. lappa* genome: (Ⅰ) The 18 chromosomes; (Ⅱ) Gene density; (Ⅲ) Transcriptional activity across the genome; (Ⅳ) GC content distribution; (Ⅴ) Density of repetitive sequences; (Ⅵ) Syntenic block relationships within the genome. 2n = 32

by DNA transposable elements (0.43%) and LINEs (0.06%) (Table S13).

Comparative genomics, genome evolution, and WGD events

To explore the evolutionary history of Peixian yellow *A. lappa*, we performed a comparative genomic analysis by comparing its genome with those of 14 other plant species. This analysis designated 44,294 gene families, of which 6,805 were common across all 15 species, and 660 were identified as single-copy gene families (Fig. 2a).

Focusing specifically on gene families in Peixian yellow *A. lappa*, comparisons with closely related species such as *H. annuus*, *L. sativa*, *C. cardunculus*, *C. intybus*, and *A. annua* revealed 6,253 genes unique to *A. lappa* (Fig. 2b). Functional enrichment analysis indicated that these unique genes were significantly associated with critical biological processes, including the nucleobase-containing compound biosynthetic process, cellular macromolecule biosynthetic process, and the regulation of cellular biosynthetic processes (Table S14).

Based on the phylogenetic tree analysis, the divergence time between Peixian yellow *A. lappa* and Gansu *A. lappa* was estimated to be approximately 3.5 million years ago (Mya) (Fig. 3a). Comparative genomic analysis revealed 1,251 significantly expanded gene families and 1,978 significantly contracted gene families in the Peixian yellow *A. lappa* genome (Fig. 3b).

GO enrichment analysis of the expanded gene families highlighted their significant involvement in key biological processes, including regulation of transcription, DNA-templated processes, regulation of cellular macromolecule biosynthetic processes, and regulation of nucleobase-containing compound metabolic processes (Table S15). These findings suggest that the expanded gene families may have played a role in shaping the unique genomic and functional characteristics of the Peixian yellow *A. lappa*.

Previous studies have identified a whole-genome triplication event (WGT-1) in *Asterids* II species such as *C. cardunculus* and *H. annuus* [44]. In this study, we performed a whole-genome duplication (WGD) analysis of Peixian yellow *A. lappa*. The analysis revealed a peak at approximately 0.28 in the 4DTv values, indicating a genome duplication event in the Peixian yellow *A. lappa* genome (Fig. 4a). To further explore this event, interspecific synteny analysis identified 9,275 syntenic genes (22.81%) within Peixian yellow *A. lappa*, supporting the hypothesis that the genome duplication involved WGT-1 and occurred prior to the diversification of *Asterids* II species (Fig. 4b).

Additionally, we conducted an analysis to identify positively selected genes by using Peixian yellow *A. lappa* as the foreground branch and *H. annuus*, *L. sativa*, *C. cardunculus*, *S. tuberosum*, and *S. melongena*. as background branches. This analysis, based on a likelihood ratio test (LRT), identified 496 positively selected genes (q-value < 0.05) (Table S16). GO enrichment of these genes highlighted significant associations with DNA metabolic processes, nucleotide-excision repair, and DNA repair in biological processes.

Comparison of SVs between the Peixian yellow *A. lappa* and Gansu *A. lappa* genomes

Using the Peixian yellow *A. lappa* genome as the reference, we identified a total of 2,972,809 SNPs and 608,644 small InDels. Additionally, 11,800 SVs ranging from 50 bp to 10,000 bp in size were detected, comprising 3,349 insertions, 3,248 deletions, 1,096 tandem expansions, 924 tandem contractions, 1,578 repeat expansions, and 1,605 repeat contractions (Fig. 5a).

The genomic distribution of these SVs is shown in Fig. 5b. The majority of SVs (73.11%) were located in intergenic regions, while 8.51% were found in exonic regions, 7.02% in intronic regions, and 5.84% in upstream regions. The relatively low proportion of SVs in exonic regions is likely attributed to their potential loss-of-function effects and the influence of strong purifying selection during plant genome evolution [69].

Analysis of key genes relevant to carotenoid biosynthesis

Carotenoids are known to contribute to the yellow, orange, and red pigmentation in flowers, fruits, and fleshy roots, respectively [13]. To investigate the molecular mechanisms responsible for the epidermal yellowing observed in the Peixian yellow *A. lappa*, we conducted a comparative transcriptome analysis of the annual fleshy roots from two *A. lappa* varieties: Liu chuan li xiang *A. lappa* (characterized by a black epidermis) and Peixian yellow *A. lappa* (characterized by a yellow epidermis).

The analysis focused on the expression patterns of protein-coding genes associated with the carotenoid biosynthesis pathway. Differential expression of these genes between the two cultivars provided insights into the genetic and regulatory mechanisms driving carotenoid accumulation and the resulting yellow pigmentation in the epidermis of the Peixian yellow *A. lappa*. A total of 6,918 DEGs were identified between the Peixian yellow *A. lappa* and the Liu chuan li xiang *A. lappa*. Among these, 3,796 genes were upregulated and 3,122 genes were downregulated in the Peixian yellow *A. lappa* (Fig. 6a, Table S17 and Table S18). The GO analysis of DEGs is presented in Table S19. KEGG pathway enrichment analysis revealed that these DEGs were significantly enriched in pathways such as plant-pathogen interaction, fatty acid elongation, glycolysis, oxidative phosphorylation, and carotenoid biosynthesis (Fig. 6b).

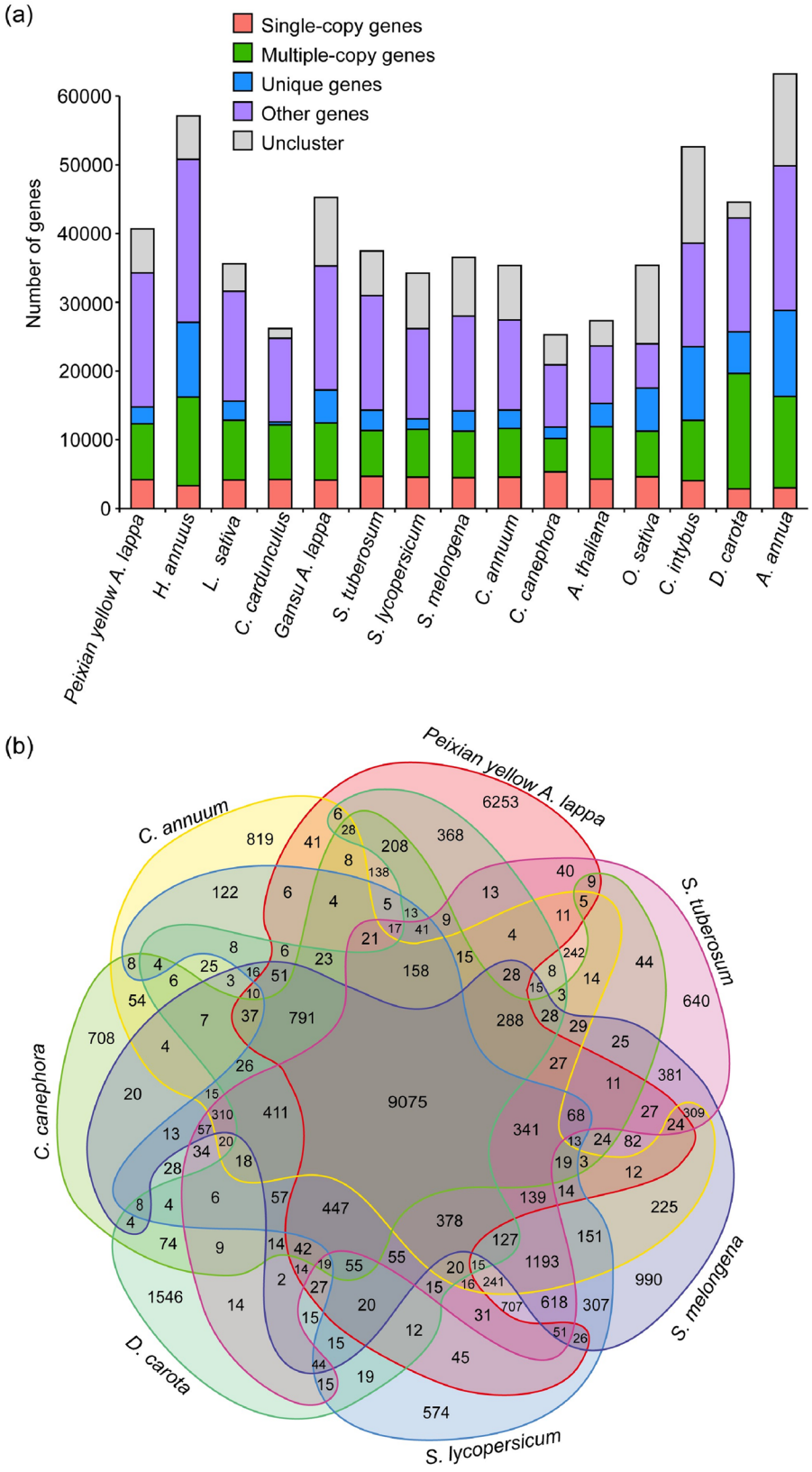

Fig. 2 Genome evolution of Peixian yellow *A. lappa*. **a** Distribution of gene families across 15 plant species, including *H. annuus*, *L. sativa*, *C. cardunculus*, Gansu *A. lappa*, *S. tuberosum*, *S. lycopersicum*, *S. melongena*, *C. annuum*, *C. canephora*, *A. thaliana*, *O. sativa*, *C. intybus*, *D. carota*, and *A. annua*. **b** Common and unique gene families identified among six species: *A. annua*, *A. lappa*, *C. cardunculus*, *C. intybus*, *H. annuus*, and *L. sativa*

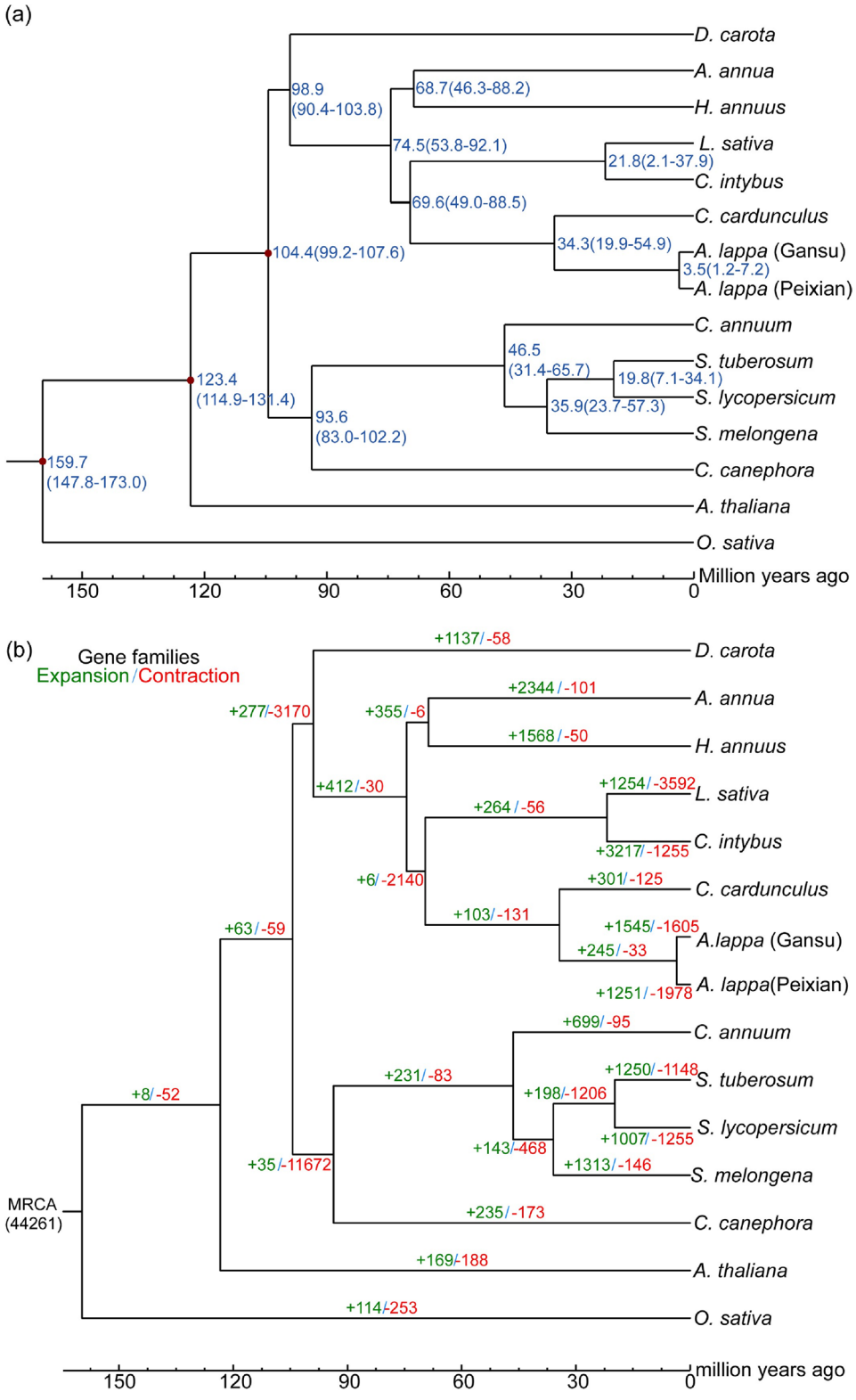

Fig. 3 Phylogenetic tree and gene family analysis of Peixian yellow *A. lappa* and related species. **a** Inferred phylogenetic tree constructed using single-copy orthologous gene families from Peixian yellow *A. lappa* and 14 other plant species. Divergence times (in million years ago, Mya) are indicated at each node with 95% credibility intervals. **b** Analysis of gene family expansions and contractions. Numbers of expanded and contracted gene families are displayed as plus (+) and minus (-) signs, respectively, followed by their corresponding counts

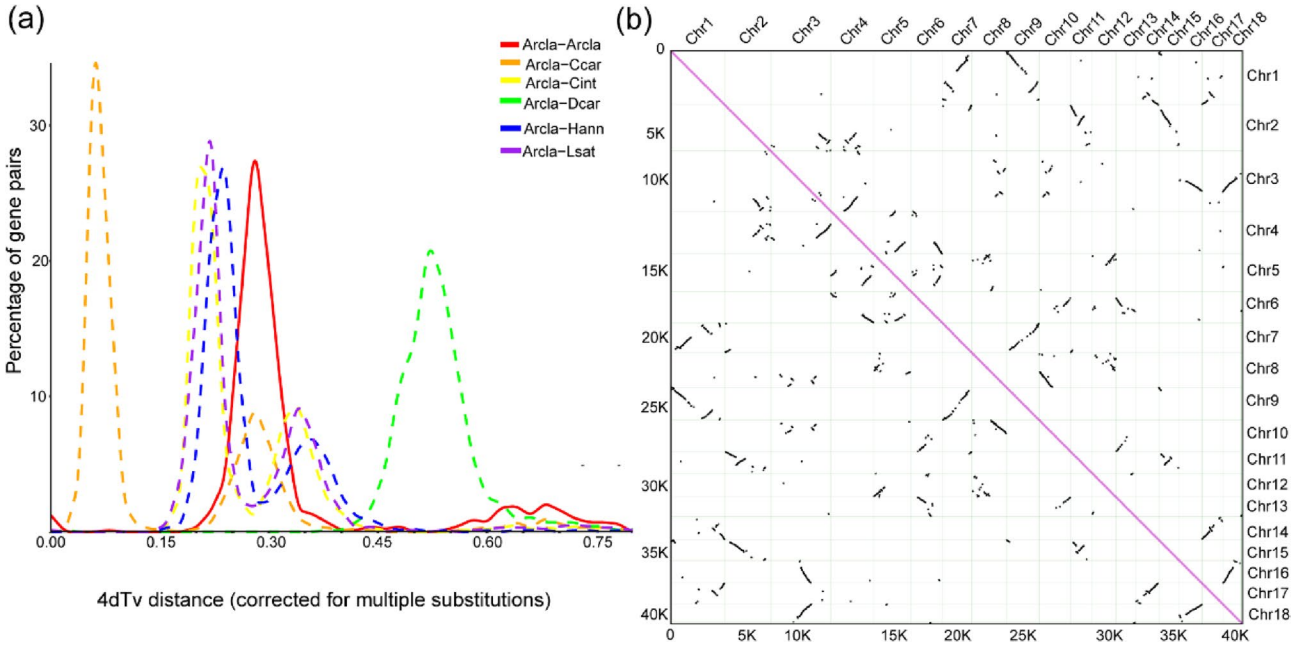

Fig. 4 Comparative genomic and evolutionary analysis of the Peixian yellow *A. lappa* genome. **a** Distribution of 4DTv values for duplicated gene pairs in syntenic blocks within the genomes of Peixian yellow *A. lappa*, *C. cardunculus*, *C. intybus*, *D. carota*, *H. annuus*, and *L. sativa*. **b** Visualization of syntenic blocks within the genome of Peixian yellow *A. lappa*.

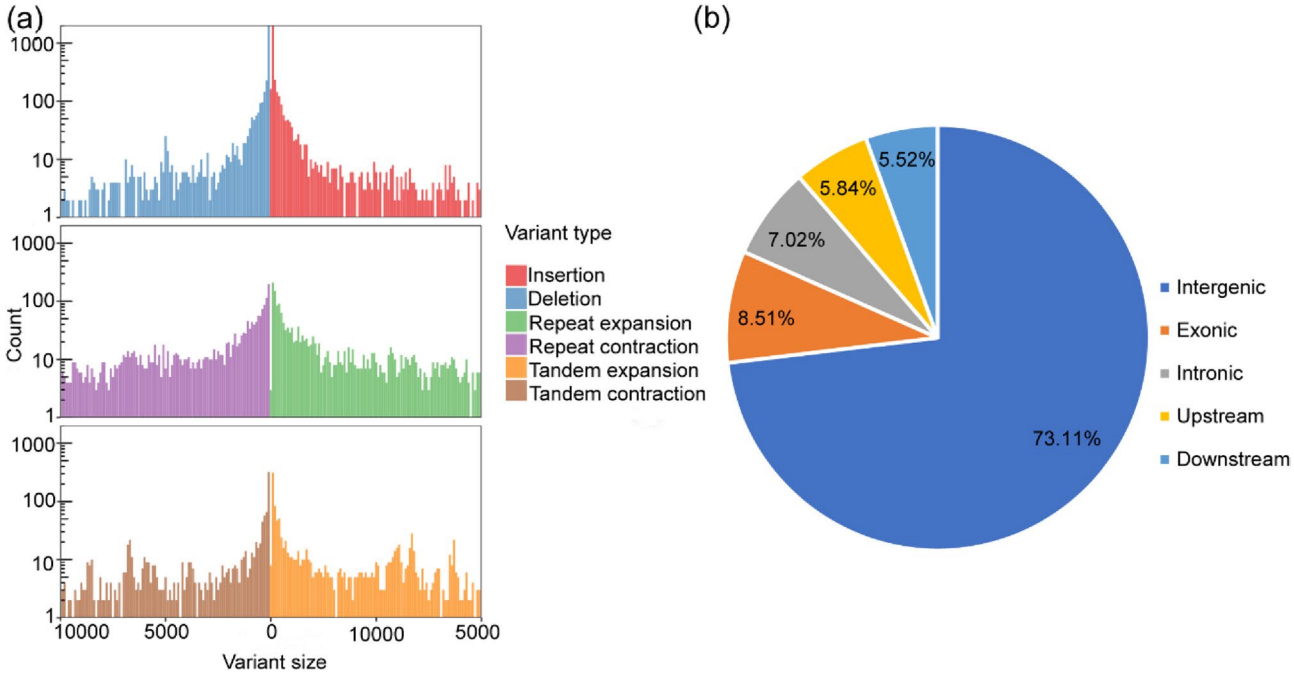

Fig. 5 Structural variants between the Peixian yellow *A. lappa* and Gansu *A. lappa* genomes. **a** Distribution of structural variants: size distributions of insertions, deletions, repeat expansions, repeat contractions, tandem expansions, and tandem contractions identified between the Peixian yellow *A. lappa* and Gansu *A. lappa* genomes. **b** Genomic localization of SVs: proportions of SVs located in different genomic regions, including exonic regions, intronic regions, upstream regions, and intergenic regions

A total of 36 genes were found to participate in carotenoid biosynthesis, with their FPKM values presented in Table S20. Among these, 20 genes showed no significant difference, including *Arcla.12g08340* (encoding *DXS*), *Arcla.18g09760*, *Arcla.09g18310*, and *Arcla.09g18300* (encoding *DXS2*), *Arcla.03g05740*, *Arcla.04g06880*, and *Arcla.15g13580* (encoding *IDI1*), *Arcla.08g12100* (encoding *IDI2*), *Arcla.15g07380*,

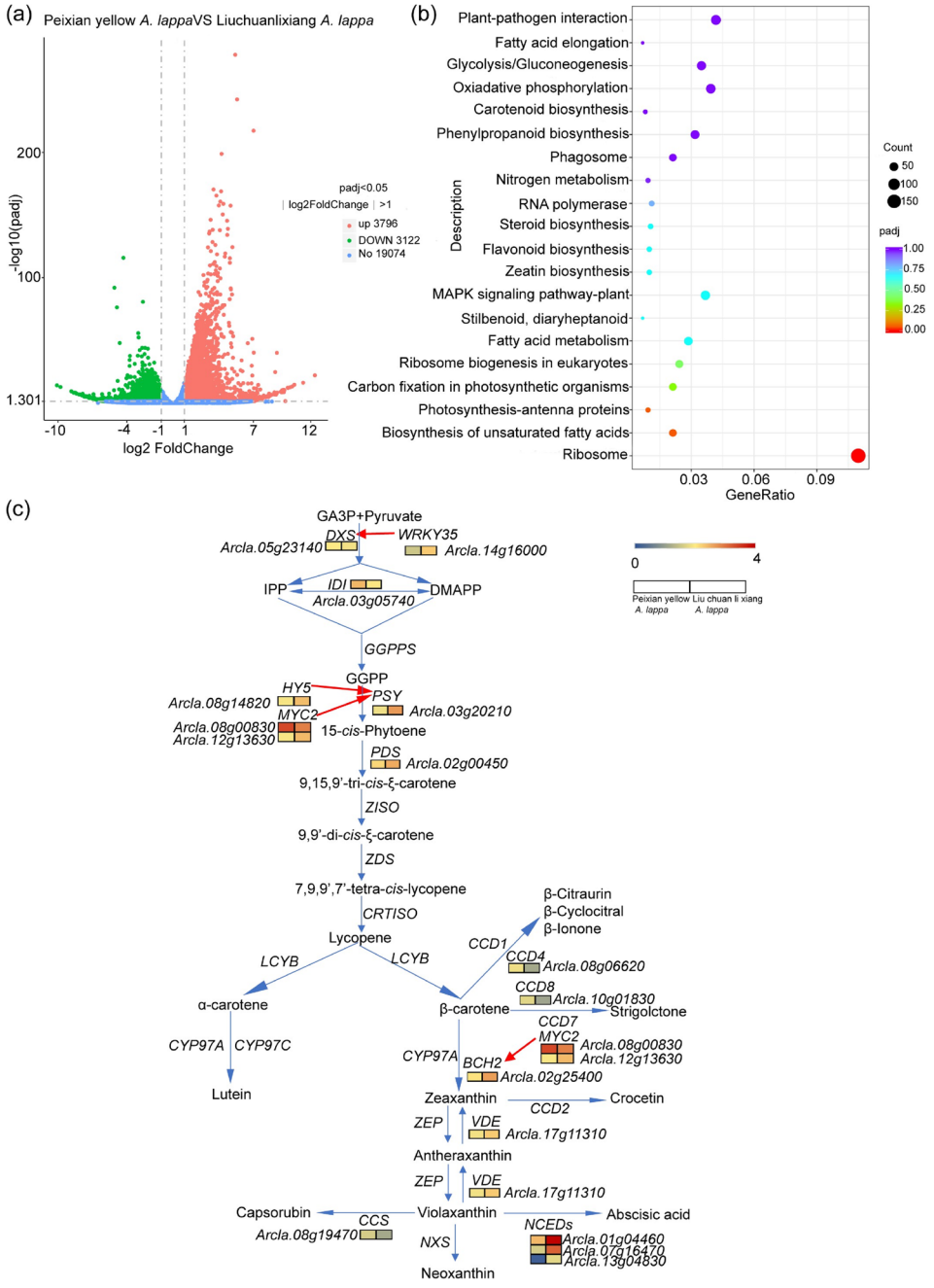

Fig. 6 Comparative transcriptomic and pathway analysis of carotenoid biosynthesis. **a** Volcano plot of DEGs between the Peixian yellow *A. lappa* and Liu chuan li xiang *A. lappa*. **b** Results of KEGG pathway enrichment for DEGs. **c** Key genes involved in the carotenoid biosynthesis pathway as identified through KEGG analysis, showcasing their differential expression between the two *A. lappa* varieties

Arcla.18g05610, and *Arcla.15g07390* (encoding *GGPPS*), *Arcla.03g18220* (encoding *ZDS*), *Arcla.14g11270*, and *Arcla.01g13440* (encoding *CRTSO*), *Arcla.17g09320* (encoding *LCYB*), *Arcla.03g05390* (encoding *CYP97A3*), *Arcla.10g02770* (encoding *CYP97C1*), *Arcla.06g14560*, *Arcla.08g15360*, and *Arcla.05g06970* (encoding *CCD1*), and *Arcla.12g01490* (encoding *CCD7*).

Notably, 16 DEGs were specifically associated with the carotenoid biosynthesis pathway. Several of these genes were downregulated in the Peixian yellow *A. lappa*, including *Arcla.03g20210* (encoding *PSY1*), *Arcla.02g00450* (encoding *PDS*), *Arcla.02g25400* (encoding *BCH2*), *Arcla.17g11310* (encoding *VDE*), *Arcla.01g04460* and *Arcla.07g16470* (both encoding *NCED1*), and *Arcla.13g04830* (encoding *NCED2*). In contrast, several genes were upregulated in the Peixian yellow *A. lappa*, including *Arcla.05g23140* (encoding *DXS*), *Arcla.03g05740* (encoding *IDI1*), *Arcla.10g01830*

(encoding *CCD8B*), *Arcla.08g19470* (encoding *CCS*), and *Arcla.08g06620* (encoding *CCD4*) (Fig. 6c; Table S20).

Additionally, several transcription factors associated with carotenoid biosynthesis were identified. These included *Arcla.08g00830* and *Arcla.12g13630* (both encoding *MYC2*), *Arcla.14g16000* (encoding *WRK35*), and *Arcla.08g14820* (encoding *HY5*). Most of these transcription factors were downregulated in the Peixian yellow *A. lappa*, with the exception of *Arcla.08g00830* (Fig. 6c; Table S20). The qRT-PCR results were consistent with the RNA-seq data (Fig. S6).

Comparative analysis of carotenoid-related genes between Peixian yellow *A. lappa* and Gansu *A. lappa*

We further analyzed SVs in carotenoid biosynthesis genes between the genomes of Peixian yellow *A. lappa* and Gansu *A. lappa*. The Gansu *A. lappa* genome contains the following SVs compared to the Peixian yellow *A. lappa*: *Arcla.05g23140* (encoding *DXS*) shows a deletion at positions 87,938,332 and 87,938,384. *Arcla.18g09760* (encoding *DXS2*) has a deletion in the intergenic region between positions 33,733,084 and 33,733,142. *Arcla.03g05740* (encoding *IDI1*) exhibits a tandem expansion in the intergenic region between positions 18,246,329 and 18,259,800. *Arcla.04g06880* (encoding *IDI1*) has a repeat contraction in the intergenic region between positions 28,244,917 and 28,250,577. *Arcla.08g12100* (encoding *IDI2*) shows an insertion at position 63,304,950 in the intergenic region. *Arcla.18g05610* (encoding *GGPPS*) contains a deletion at positions 14,252,244 and 14,252,296 in the intergenic region. *Arcla.10g02770* (encoding *CYP97C1*) has an insertion at positions 5,073,575 and 5,073,578. *Arcla.08g15360* (encoding *CCD1*) exhibits an insertion at position 73,433,584. *Arcla.01g04460* (encoding *NCED1*) contains an insertion at position 9,383,957 in the intergenic region. *Arcla.13g04830* (encoding *NCED2*) has an insertion between positions 14,786,428 and 14,786,432 in the intergenic region.

Additionally, *Arcla.08g00830* and *Arcla.12g13630* (both encoding *MYC2*) show the following differences: *Arcla.08g00830* has a repeat expansion between positions 4,041,474 and 4,069,315. *Arcla.12g13630* has an insertion between positions 66,832,819 and 66,832,828 in the intergenic region (Table S23).

Discussion

Comparison of the Peixian yellow *A. lappa* genome with other *A. lappa* genomes

High-quality, chromosome-level genome assemblies are essential for molecular breeding and for uncovering the molecular basis of economically important traits in plants [70]. In recent years, Hi-C sequencing technology, which utilizes long-range linkage information across tens of megabases, has become an efficient method for assembling chromosome-scale scaffolds in many large eukaryotic genomes [71]. *A. lappa*, an important fleshy root vegetable in the composite family, has been the subject of several genome assemblies at the chromosome level [47, 67]. However, research on the molecular genetics of *A. lappa* lags behind that of other closely related *Asteraceae* plants, hindering the identification of functional genomic variations and the dissection of the genetic determinants of complex traits in *A. lappa*. In this study, we generated a high-quality, chromosome-level genome for Peixian yellow *A. lappa*, which has a genome size of 1.69 Gb and a scaffold N50 value of 91.02 Mb (Table S8). This assembly is 10 times longer than the Dongbei *A. lappa* genome (1.79 Gb, scaffold N50 = 6.88 Mb) (Table S8).

Increasing evidence suggests that SVs can drive significant phenotypic and morphological variation by influencing gene dosage, function, and regulation in crop species [72]. In our study, we analyzed SVs associated with carotenoid synthesis in Peixian yellow *A. lappa* and Gansu *A. lappa* (Table S23). This is the first study to identify SVs between these two *A. lappa* genomes.

Given the rapid advances in sequencing technologies, it is crucial to further characterize genetic variation across different *A. lappa* species, subspecies, and varieties and to explore how these variations impact the genetic control of key horticultural traits in *Arctium*. Taken together, the availability of this complete genome assembly for Peixian yellow *A. lappa* provides a valuable resource for genetic breeding, as well as for evolutionary and comparative studies within the *Arctium* genus.

Carotenoid biosynthetic pathways in Peixian yellow *A. lappa*

Carotenoids are a group of isoprenoid metabolites that are vital for life. In plants, carotenoids are essential for photosynthesis and photoprotection, acting as light-harvesting pigments and structural components of photosystems. Their accumulation in fruits and fleshy roots contributes to the vivid orange, yellow, or red colors, which have significant ecological and agronomical value [13, 73]. Beyond their fundamental roles in plants, carotenoids are also crucial for human nutrition and health. As a result, substantial efforts have been made to increase carotenoid levels in food crops, enhancing their nutritional value and health benefits [74, 75]. Over the past few decades, significant progress has been made in understanding the carotenoid biosynthetic pathway and its regulation [76]. Several genes involved in carotenoid biosynthesis have been identified, including *PSY*, *PDS*, *VDE*, *NCED*, *DXS*, *IDI*, *CCD*, and *CCS* [77–80]. In this study, we performed a comparative transcriptomic analysis of two *A. lappa* varieties with distinct skin colors and identified DEGs associated with carotenoid biosynthesis.

Key genes, such as *PSY1*, *PDS*, *BCH2*, *VDE*, *NCED1*, and *NCED2*, were found to be downregulated in the Peixian yellow *A. lappa*, while *DXS*, *IDI1*, *CCD8B*, *CCS*, and *CCD4* were upregulated.

As the first enzyme in the methylerythritol phosphate (MEP) pathway, *DXS* plays a critical role in carotenoid biosynthesis [81]. It has been proposed that upregulation of *SlDXS1* in the MEP pathway can effectively increase carotenoid content in tomato fruit [82]. Thus, the upregulation of *DXS* in Peixian yellow *A. lappa* may contribute to the enhanced carotenoid content in this variety.

Transcription factors (TFs) play a central role in regulating the expression of carotenogenic genes. Recent advances have identified and validated various transcription factors that either activate or suppress the expression of carotenoid structural genes. Additionally, new carotenogenic transcription factors continue to be discovered and characterized [83]. Among these, *HY5* can bind to the G-box elements in the promoters of carotenoid biosynthesis genes like *PSY*, as well as chlorophyll synthesis genes, promoting both chlorophyll and carotenoid accumulation in response to light [84]. *MYC2*, a member of the bHLH family, has also been shown to regulate carotenoid biosynthesis. In *Citrus*, *CsMYC2* activates the expression of genes like *CsPSY* and *CsBCH*, which are responsible for fruit peel coloration [85]. Furthermore, *WRKY* TFs are regulated at multiple levels within various gene regulatory networks, including those of carotenoid biosynthesis. *WRKY* TFs can bind to W-box elements in the promoter regions of target genes [86]. In tomato fruit, *SlWRKY35* directly activates the expression of the *1-deoxy-D-xylulose 5-phosphate synthase* (*SlDXS1*) gene in the MEP pathway, leading to enhanced carotenoid accumulation [17].

Despite extensive research on carotenoid regulation in other species, the transcription factors involved in carotenoid biosynthesis in *A. lappa* have not been reported. In our study, we found that TFs such as *MYC2*, *WRKY35*, and *HY5*, known to be involved in carotenoid regulation were downregulated in Peixian yellow *A. lappa*. This suggests that *WRKY35* likely activates the expression of *DXS*, *HY5* activates *PSY1*, and *MYC2* regulates the expression of both *PSY* and *BCH2*. However, the mechanisms by which these transcription factors regulate carotenoid biosynthesis in *A. lappa* appear to be complex and warrant further investigation to fully elucidate their roles.

Our findings lay the groundwork for future studies on carotenoid biosynthesis and regulation in *A. lappa*. Additionally, these results contribute to the comprehensive identification of genetic variants associated with key horticultural traits, providing valuable genomic resources for developing gene-targeted strategies aimed at enhancing desirable traits in *A. lappa*.

Conclusions

In this study, we assembled a high-quality, chromosome-level genome of Peixian yellow *A. lappa* with a total length of 1.69 Gb using PacBio HiFi and Hi-C technologies. A total of 99.12% of the sequences were successfully anchored to 18 chromosomes. The scaffold N50 value was 91.02 Mb, and the BUSCO score was 98.2%, indicating a high-quality genome assembly. Gene annotation identified 40,665 genes. Among the sequences, 68.32% were repetitive, with LTRs representing the largest fraction of the repeats (65.66%). Evolutionary analysis suggested that Peixian yellow *A. lappa* have undergone a WGD event. Additionally, a combined transcriptomic analysis of Peixian yellow *A. lappa* and Liu chuan li xiang *A. lappa* identified 16 DEGs involved in carotenoid synthesis. Notably, the expression of *DXS*, *IDI1*, *CCD8B*, *CCS*, and *CCD4* genes was significantly upregulated, while the expression of *PSY1*, *PDS*, *BCH2*, *VDE*, *NCED1*, and *NCED2* genes was significantly downregulated. These expression patterns of DEGs related to carotenoid biosynthesis may play a crucial role in the coloration of the epidermis in Peixian yellow *A. lappa*.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-025-07539-x.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

We thank all our colleagues for providing useful discussions and technical assistance.

Authors' contributions

X.F. drafted the manuscript. K. T., Y. W., and N. Z. methodology. S. L. and L. L. conceptualization. E. L., A. C., and S. W. investigation. T. K., D. D and C. H. project administration. Y. D. writing-review, and editing. The authors read and approved the final manuscript.

Funding

This work was supported by the Jiangsu Province Industry-University-Research Cooperation Project (BY20231322).

Data availability

The genome assemblies and gene annotations have been deposited at GenBank database the accession number PRJNA1171327. The transcriptome data were deposited at the SRA database with accession number PRJNA1172597.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 12 May 2025 / Accepted: 9 October 2025
Published online: 05 November 2025

References

1. Chan YS, Cheng LN, Wu JH, Chan E, Kwan YW, Lee SM, et al. A review of the pharmacological effects of *Arctium lappa* (burdock). Inflammopharmacology. 2011;19(5):245–54. https://doi.org/10.1007/s10787-010-0062-4.
2. Wang D, Bădărau AS, Swamy MK, Shaw S, Maggi F, da Silva LE, López V, Yeung AWK, Mocan A. Atanasov AG.Arctium secies secondary metabolites chemodiversity and bioactivities. Front Plant Sci. 2019;10:834. https://doi.org/10.3389/fpls.2019.00834.
3. Shyam M, Sabina EP. Harnessing the power of *Arctium lappa* root: a review of its pharmacological properties and therapeutic applications. Nat Prod Bioprospect. 2024;14(1):49. https://doi.org/10.1007/s13659-024-00466-8.
4. Moro TMA. Burdock (*Arctium lappa* L) roots as a source of inulin-type fructans and other bioactive compounds: current knowledge and future perspectives for food and non-food applications. Food Res Int. 2021;141:109889. https://doi.org/10.1016/j.foodres.2020.109889.
5. Bicknell R, Gaillard M, Catanach A, McGee R, Erasmuson S, Fulton B, et al. Genetic mapping of the LOSS OF PARTHENOGENESIS locus in *Pilosella piloselloides* and the evolution of apomixis in the lactuceae. Front Plant Sci. 2023;14:1239191. https://doi.org/10.3389/fpls.2023.1239191.
6. Domont J, Thiblet M, Etienne A, Santos HAD, Cadalen T, Hance P, Gagneul D, Hilbert JL, Rambaud C. CRISPR /Cas9-targeted mutagenesis of CiGAS and CiGAO to reduce bitterness in Chicory (*Cichorium intybus* L). Front Biosci. 2023;28(9):201. https://doi.org/10.31083/j.fbl2809201.
7. Liu L, Xue Y, Luo J, Han M, Liu X, Jiang T, et al. Developing a UV-visible reporter-assisted CRISPR/Cas9 gene editing system to alter flowering time in *Chrysanthemum indicum*. Plant Biotechnol J. 2023;21(8):1519–521. https://doi.org/10.1111/pbi.14062.
8. Zhu Y, Liu Y, Wang W, Li H, Liu C, Dou L, et al. Identification and characterization of CYC2-like genes related to floral symmetric development in *Tagetes erecta* (*Asteraceae*). Gene. 2023;889:147804. https://doi.org/10.1016/j.gene.2023.147804.
9. Nunn A, Rodríguez-Arévalo I, Tandukar Z, Frels K, Contreras-Garrido A, Carbonell-Bejerano P, et al. Chromosome-level *Thlaspi arvense* genome provides new tools for translational research and for a newly domesticated cash cover crop of the cooler climates. Plant Biotechnol J. 2022;20(5):944–63. https://doi.org/10.1111/pbi.13775.
10. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell. 2020;182(1):145–61. https://doi.org/10.1016/j.cell.2020.05.021.
11. Li H, Wang S, Chai S, Yang Z, Zhang Q, Xin H, et al. Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. Nat Commun. 2022;13(1):682. https://doi.org/10.1038/s41467-022-28362-0.
12. Meléndez-Martínez AJ, Mandić AI, Bantis F, Böhm V, Borge GIA, Brnčić M, et al. A comprehensive review on carotenoids in foods and feeds: status quo, applications, patents, and research needs. Crit Rev Food Sci Nutr. 2022;62(8):1999–2049. https://doi.org/10.1080/10408398.2020.1867959.
13. Liang MH, Li XY. Involvement of transcription factors and regulatory proteins in the regulation of carotenoid accumulation in plants and algae. J Agric Food Chem. 2023;71(48):18660–73. https://doi.org/10.1021/acs.jafc.3c05662.
14. Liang MH, Zhu J, Jiang JG. Carotenoids biosynthesis and cleavage related genes from bacteria to plants. Crit Rev Food Sci Nutr. 2018;58(14):2314–33. https://doi.org/10.1080/10408398.2017.1322552.
15. Fu CC, Han YC, Kuang JF, Chen JY, Lu WJ. Papaya CpEIN3a and CpNAC2 co-operatively regulate carotenoid biosynthesis-related genes CpPDS2/4, CpLCY-e and CpCHY-b during fruit ripening. Plant Cell Physiol. 2017;58(12):2155–65. https://doi.org/10.1093/pcp/pcx149.
16. Stanley L, Yuan YW. Transcriptional regulation of carotenoid biosynthesis in plants: so many regulators, so little consensus. Front Plant Sci. 2019;10:1017. https://doi.org/10.3389/fpls.2019.01017.
17. Yuan Y, Ren S, Liu X, Su L, Wu Y, Zhang W, et al. SlWRKY35 positively regulates carotenoid biosynthesis by activating the MEP pathway in tomato fruit. New Phytol. 2022;234(1):164–78. https://doi.org/10.1111/nph.17977.
18. Alquezar B, Rodrigo MJ, Zacarías L. Regulation of carotenoid biosynthesis during fruit maturation in the red-fleshed orange mutant Cara Cara. Phytochemistry. 2008;69(10):1997–2007. https://doi.org/10.1016/j.phytochem.2008.04.020.
19. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW. Real-time DNA sequencing from single polymerase molecules. Methods Enzymol. 2010;472:431–55. https://doi.org/10.1016/s0076-6879(10)72001-2.
20. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012;58(3):268–76. https://doi.org/10.1016/j.ymeth.2012.05.001.
21. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de Novo assembler. GigaScience. 2012;1(1):18. https://doi.org/10.1186/2047-217x-1-18.
22. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–5. https://doi.org/10.1038/s41592-020-01056-5.
23. 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(11):e112963. https://doi.org/10.1371/journal.pone.0112963.
24. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.
25. Parra G, Bradnam K, Korf I.CEGMA : a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061–7. https://doi.org/10.1093/bioinformatics/btm071.
26. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2. https://doi.org/10.1093/bioinformatics/btv351.
27. Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR assembly index (LAI). Nucleic Acids Res. 2018;46(21):e126. https://doi.org/10.1093/nar/gky730.
28. Chen C, Wu Y, Xia R. A painless way to customize circos plot: from data preparation to visualization using TBtools. iMeta. 2022;1(3):e35. https://doi.org/10.1002/imt2.35.
29. Keilwagen J, Hartung F, Grau JGMM. Homology-based gene prediction utilizing intron position conservation and RNA-seq data. Methods Mol Biol. 2019;1962:161–77. https://doi.org/10.1007/978-1-4939-9173-0_9.
30. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. Repeatmodeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117(17):9451–7. https://doi.org/10.1073/pnas.1921046117.
31. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80. https://doi.org/10.1093/nar/27.2.573.
32. Tarailo-Graovac M, Chen N. Using repeatmasker to identify repetitive elements in genomic sequences. Curr Protocols Bioinf. 2009. https://doi.org/10.1002/0471250953.bi0410s25.
33. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with tophat and cufflinks. Nat Protoc. 2012;7(3):562–78. https://doi.org/10.1038/nprot.2012.016.
34. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 2004;32:W309–312. https://doi.org/10.1093/nar/gkh379.
35. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59. https://doi.org/10.1186/1471-2105-5-59.
36. 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(1):R7. https://doi.org/10.1186/gb-2008-9-1-r7.
37. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al. The SWISS-PROT protein knowledgebase and its supplement trembl in 2003. Nucleic Acids Res. 2003;31(1):365–70. https://doi.org/10.1093/nar/gkg095.
38. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37:D211–215. https://doi.org/10.1093/nar/gkn785.
39. Liao N, Hu Z, Miao J, Hu X, Lyu X, Fang H, et al. Chromosome-level genome assembly of bunching onion illuminates genome evolution and flavor

formation in allium crops. Nat Commun. 2022;13(1):6690. https://doi.org/10.1038/s41467-022-34491-3.
40. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al. The gene ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32:D258–261. https://doi.org/10.1093/nar/gkh036.
41. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014;42:D199–205. https://doi.org/10.1093/nar/gkt1076.
42. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955–64. https://doi.org/10.1093/nar/25.5.955.
43. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–5. https://doi.org/10.1093/bioinformatics/btt509.
44. Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, et al. The sunflower genome provides insights into oil metabolism, flowering and asterid evolution. Nature. 2017;546(7656):148–52. https://doi.org/10.1038/nature22380.
45. Reyes-Chin-Wo S, Wang Z, Yang X, Kozik A, Arikit S, Song C, et al. Genome assembly with *in vitro* proximity ligation data and whole-genome triplication in lettuce. Nat Commun. 2017;8:14953. https://doi.org/10.1038/ncomms14953.
46. Acquadro A, Barchi L, Portis E, Mangino G, Valentino D, Mauromicale G, et al. Genome reconstruction in *Cynara cardunculus* taxa gains access to chromosome-scale DNA variation. Sci Rep. 2017;7(1):5617. https://doi.org/10.1038/s41598-017-05085-7.
47. Fan W, Wang S, Wang H, Wang A, Jiang F, Liu H, et al. The genomes of chicory, endive, great burdock and yacon provide insights into Asteraceae palaeo-polyploidization history and plant inulin production. Mol Ecol Resour. 2022;22(8):3124–40. https://doi.org/10.1111/1755-0998.13675.
48. Shen Q, Zhang L, Liao Z, Wang S, Yan T, Shi P, Liu M, Fu X, Pan Q, Wang Y, et al. The genome of Artemisia annua provides insight into the evolution of Asteraceae family and Artemisinin biosynthesis. Mol Plant. 2018;11(6):776–88. https://doi.org/10.1016/j.molp.2018.03.015.
49. Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang J, et al. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat Genet. 2016;48(6):657–66. https://doi.org/10.1038/ng.3565.
50. Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, et al. Genome sequence and analysis of the tuber crop potato. Nature. 2011;475(7355):189–95. https://doi.org/10.1038/nature10158.
51. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012;485(7400):635–41. https://doi.org/10.1038/nature11119.
52. Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in capsicum species. Nat Genet. 2014;46(3):270–8. https://doi.org/10.1038/ng.2877.
53. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G, et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science. 2014;345(6201):1181–4. https://doi.org/10.1126/science.1255274.
54. Analysis of the genome sequence of the flowering plant *Arabidopsis Thaliana*. Nature. 2000;408(6814):796–815. https://doi.org/10.1038/35048692.
55. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science. 2002;296(5565):92–100. https://doi.org/10.1126/science.1068275.
56. Li L, Stoeckert CJ Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89. https://doi.org/10.1101/gr.1224503.
57. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7. https://doi.org/10.1093/nar/gkh340.
58. Stamatakis ARAML. Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033.
59. Yang Z. PAML 4: Phylogenetic analysis by aximum likelihood. Mol Biol Evol. 2007;24(8):1586–91. https://doi.org/10.1093/molbev/msm088.
60. The Amborella genome and the evolution of flowering plants. Science. 2013;342(6165):1241089. https://doi.org/10.1126/science.1241089.
61. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006;22(10):1269–71. https://doi.org/10.1093/bioinformatics/btl097.
62. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCscanx: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7):e49. https://doi.org/10.1093/nar/gkr1293.
63. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics. 2016;32(19):3021–3. https://doi.org/10.1093/bioinformatics/btw369.
64. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015;10(10):1556–66. https://doi.org/10.1038/nprot.2015.105.
65. Ying S, Su M, Wu Y, Zhou L, Fu R, Li Y, et al. Trichome regulator *SlMIXTA*-like directly manipulates primary metabolism in tomato fruit. Plant Biotechnol J. 2020;18(2):354–63. https://doi.org/10.1111/pbi.13202.
66. Liao Y, Smyth GK, Shi W. Featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30. https://doi.org/10.1093/bioinformatics/btt656.
67. Yang Y, Li S, Xing Y, Zhang Z, Liu T, Ao W, et al. The first high-quality chromosomal genome assembly of a medicinal and edible plant *Arctium lappa*. Mol Ecol Resour. 2022;22(4):1493–507. https://doi.org/10.1111/1755-0998.13547.
68. Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 2009;10(10):725–32. https://doi.org/10.1038/nrg2600.
69. Guan J, Xu Y, Yu Y, Fu J, Ren F, Guo J, et al. Genome structure variation analyses of Peach reveal population dynamics and a 1.67 mb causal inversion for fruit shape. Genome Biol. 2021;22(1):13. https://doi.org/10.1186/s13059-020-02239-1.
70. Zhang X, Lin S, Peng D, Wu Q, Liao X, Xiang K, et al. Integrated multi-omic data and analyses reveal the pathways underlying key ornamental traits in carnation flowers. Plant Biotechnol J. 2022;20(6):1182–96. https://doi.org/10.1111/pbi.13801.
71. Cai X, Chang L, Zhang T, Chen H, Zhang L, Lin R, et al. Impacts of allopolyploidization and structural variation on intraspecific diversification in brassica Rapa. Genome Biol. 2021;22(1):166. https://doi.org/10.1186/s13059-021-02383-2.
72. Renner SS, Wu S, Pérez-Escobar OA, Silber MV, Fei Z, Chomicki G. A chromosome-level genome of a Kordofan melon illuminates the origin of domesticated watermelons. Proc Natl Acad Sci USA. 2021. https://doi.org/10.1073/pnas.2101486118.
73. Sun T, Rao S, Zhou X, Li L. Plant carotenoids: recent advances and future perspectives. Mol Hortic. 2022;2(1):3. https://doi.org/10.1186/s43897-022-00023-2.
74. Liang MH, He YJ, Liu DM, Jiang JG. Regulation of carotenoid degradation and production of apocarotenoids in natural and engineered organisms. Crit Rev Biotechnol. 2021;41(4):513–34. https://doi.org/10.1080/07388551.2021.1873242.
75. Zheng X, Giuliano G, Al-Babili S. Carotenoid biofortification in crop plants: citius, altius, fortius. Biochim Et Biophys Acta Mol Cell Biol Lipids. 2020;1865(11):158664. https://doi.org/10.1016/j.bbalip.2020.158664.
76. Wang XJ, Luo Q, Li T, Meng PH, Pu YT, Liu JX, et al. Origin, evolution, breeding, and omics of apiaceae: a family of vegetables and medicinal plants. Hortic Res. 2022;9:uhac076. https://doi.org/10.1093/hr/uhac076.
77. Estévez JM, Cantero A, Reindl A, Reichler S, León. 1-Deoxy-D-xylulose-5-phosphate synthase, a limiting enzyme for plastidic isoprenoid biosynthesis in plants. J Biol Chem. 2001;276(25):22901–9. https://doi.org/10.1074/jbc.M100854200.
78. Rodríguez-Villalón A, Gas E, Rodríguez-Concepción M. Phytoene synthase activity controls the biosynthesis of carotenoids and the supply of their metabolic precursors in dark-grown *Arabidopsis* seedlings. Plant J. 2009;60(3):424–35. https://doi.org/10.1111/j.1365-313X.2009.03966.x.
79. Jahns P, Holzwarth AR. The role of the xanthophyll cycle and of lutein in photoprotection of photosystem II. Biochim Biophys Acta. 2012;1817(1):182–93. https://doi.org/10.1016/j.bbabio.2011.04.012.
80. Ablazov A, Mi J, Jamil M, Jia KP, Wang JY, Feng Q, et al. The apocarotenoid Zaxinone is a positive regulator of strigolactone and abscisic acid biosynthesis in *Arabidopsis* roots. Front Plant Sci. 2020;11:578. https://doi.org/10.3389/fpls.2020.00578.
81. Cheng Y, Xiang N, Chen H, Zhao Y, Wang L, Cheng X, et al. The modulation of light quality on carotenoid and tocochromanol biosynthesis in mung bean (*Vigna radiata*) sprouts. Food Chem. 2023;6:100170. https://doi.org/10.1016/j.fochms.2023.100170.
82. Lois LM, Rodríguez-Concepción M, Gallego F, Campos N, Boronat A. Carotenoid biosynthesis during tomato fruit development: regulatory role of

1-deoxy-D-xylulose 5-phosphate synthase. Plant J. 2000;22(6):503–13. https://doi.org/10.1046/j.1365-313x.2000.00764.x.
83. Gong J, Zeng Y, Meng Q, Guan Y, Li C, Yang H, et al. Red light-induced Kumquat fruit coloration is attributable to increased carotenoid metabolism regulated by FcrNAC22. J Exp Bot. 2021;72(18):6274–90. https://doi.org/10.1093/jxb/erab283.
84. Toledo-Ortiz G, Johansson H, Lee KP, Bou-Torrent J, Stewart K, Steel G, et al. The HY5-PIF regulatory module coordinates light and temperature control of photosynthetic gene transcription. PLoS Genet. 2014;10(6):e1004416. https://doi.org/10.1371/journal.pgen.1004416.
85. Yue P, Jiang Z, Sun Q, Wei R, Yin Y, Xie Z, et al. Jasmonate activates a *CsMPK6-CsMYC2* module that regulates the expression of β-citraurin biosynthetic genes and fruit coloration in orange (*Citrus sinensis*). Plant Cell. 2023;35(4):1167–85. https://doi.org/10.1093/plcell/koac363.
86. Mahiwal S, Pahuja S, Pandey GK. Review :tructural-functional relationship of WRKY transcription factors: unfolding the role of WRKY in plants. Int J Biol Macromol. 2024;257:128769. https://doi.org/10.1016/j.ijbiomac.2023.128769.

Publisher's Note