



OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the ornamental plant *Alcea rosea*

Xi Chen, Xiu Li, Shengwen Tang, Jiao Ma, Zhangshun Zhu, Fangwen Li & Xiaoqing Shi

Alcea rosea, a member of the Malvaceae family, is celebrated for its rich floral palette and global horticultural significance. Here, we present a high-quality reference genome for *A. rosea*, achieving a genome assembly size of 1.01 Gbp, with a Contig N50 length of 36.61 Mbp. The genome sequence was successfully mapped to 21 chromosomes, and the scaffold N50 length reached 52.57 Mbp, with a scaffold genome completeness of 99.6%. A total of 565.84 Mbp (comprising 56% of the genome) of repetitive sequences were identified, with transposable elements being predominant, particularly long terminal repeat (LTR) elements, which accounted for 48.44% of the genome. 51,436 genes were annotated. Among these predicted genes, the average gene length and coding sequence (CDS) length were 2739.92 bp and 1242.54 bp, respectively.

Background & Summary

The Malvaceae family encompasses both economically vital crops and ornamental species. Economically, it includes globally significant crops such as cotton (*Gossypium* spp.) for natural fiber production and cacao (*Theobroma cacao*) for chocolate manufacturing^{1–3}. Ornamentally, the genus *Hibiscus* (e.g., *H. syriacus*, *H. mutabilis*) has been extensively studied for its horticultural value. As a member of the Malvaceae, hollyhock (*A. rosea*) holds unique dual significance. Historically cultivated in China for over two millennia, it was documented in ancient pharmacopeias like Compendium of Materia Medica for its edible and medicinal properties. Modern studies have identified bioactive compounds in hollyhock, including flavonoids and polysaccharides with demonstrated anti-inflammatory and antimicrobial activities^{4–7}. Beyond its traditional uses, hollyhock exhibits exceptional floral diversity, with petal colors ranging from white to black and intricate variegation patterns, making it an ideal system for studying pigmentation genetics (Fig. 1).

Prior genomic efforts within the Malvaceae family have predominantly focused on economically pivotal species. The genus *Gossypium* (cotton) boasts multiple high-quality genomes, including the diploid *G. raimondii* and tetraploid *G. hirsutum*, which revolutionized fiber development studies^{8,9}. Similarly, the cacao genome and its re-sequencing efforts have accelerated research on flavonoid biosynthesis^{10,11}. For ornamental taxa, chromosome-level assemblies of *H. syriacus*¹² and *H. mutabilis*¹³ revealed genetic bases for floral color transition and woody growth.

Notably, *A. rosea* genomic exploration remains limited. Previous studies have conducted transcriptome sequencing on single-petaled and double-petaled red-flowered *A. rosea*, identifying several genes potentially associated with double-petaled flower formation^{14,15}. Additionally, Wang *et al.*¹⁶ performed genome sequencing of *A. rosea* (genome data yet to be published) and resequenced 32 germplasm resources to investigate its genetic diversity. This contrasts with related genera like *Ceiba pentandra* (kapok tree) and *Bombax ceiba*, whose genomes enabled comparative analyses of fiber evolution¹⁷. The lack of a hollyhock nuclear genome hinders exploration of its hallmark traits: petal pigmentation diversity and mucilage biosynthesis pathways.

Despite these advances in related species, *A. rosea* genomic exploration remains limited. To address this, we present the first chromosome-level genome assembly of *A. rosea*. Combining PacBio HiFi (52x) sequencing, Hi-C scaffolding, and RNA-seq annotation, we generated a 1.01 Gbp genome with scaffold N50 of 52.57 Mbp, anchoring 92.6% (1.01 Gbp) onto 21 chromosomes. We annotated 51,436 protein-coding genes, providing a critical resource for elucidating the genetic basis of hollyhock's ornamental and medicinal traits.

Chengdu Botanical Garden, Chengdu, Sichuan, China. ✉e-mail: shixiaoqing87@126.com



Fig. 1 The diverse color and petal types of *A. rosea*.

Methods

Sampling. The plant seed material for genome sequencing originated from wild Hollyhock seeds preserved in the Germplasm Bank of Wild Species in Southwest China, collected from Songpan, Aba, Sichuan, with the collection number SCU-10-427. Seeds were sown in plug trays during March and maintained in the nursery of Chengdu Botanical Garden (30°46'N, 104°07'E) under controlled conditions: temperature 15–20 °C, natural photoperiod, and peat-based substrate (pH 5.5–6.0). At one month post-sowing, seedlings developed approximately 3–4 true leaves, from which leaf samples were collected for preliminary genomic analysis.

Seedlings were then transplanted into 14-cm pots under adjusted conditions: temperature 20–25 °C (other parameters unchanged). By four months post-sowing, plants entered a rapid vegetative growth phase (25–35 °C), and young leaves were harvested for genome sequencing.

At 14 months post-sowing (mature flowering stage, 20–30 °C), multiple tissues (leaves, stems, fruits, flowers, and roots) from the same individual were collected for transcriptome sequencing.

Chromosome fluorescence by *in situ* hybridization. Take the root of *A. rosea*, which has active meristematic tissue, and use nitrous oxide to induce cell mitosis. After obtaining mid-phase cells, prepare chromosome specimens and perform fluorescence *in situ* hybridization (FISH) using fluorescent probes based on the conserved repetitive sequences of telomeres, 5SrDNA and 18SrDNA. After DAPI staining, observe chromosomes under a fluorescence microscope to determine the karyotypic features of the species. Through chromosome squashing of root tips and FISH, it was revealed that *A. rosea* has a total of 42 chromosomes, with lengths ranging from 0.8 to 2.0 μm. The majority of these chromosomes were found to be metacentric or submetacentric. Using 5S rDNA and 18S rDNA repeated sequence probes for FISH analysis, we identified 8 chromosomes with strong 5S rDNA (red) hybridization signals and 6 chromosomes with strong 18S rDNA (green) hybridization signals, confirming the diploid nature of the sample with a karyotype of $2n = 2x = 42$ (Fig. 2).

Genome size estimation. To estimate the genome size of *A. rosea*, 100X short read data (PE150; DNBSEQ) performed K-mer analysis, and SOAPnuke (v 1.5.6)¹⁸ was used to filter the original sequencing data, and the reads with low quality, adaptor contamination and PCR duplication were filtered out. The remaining clean reads were reused for subsequent analysis. The 17,21 kmer frequency is quickly counted using the Jellyfish (v2.1.4)¹⁹, then GenomeScope (v1.0)²⁰ is used to fit the 17,21 kmer spectrum and estimate the genome's characteristics. The genomic survey, employing a negative binomial distribution model, estimated the species' genome size to be around 1.2 Gbp with a heterozygosity rate of 0.41% and 76.3% repeat content, indicative of a low heterozygous, high repeat genome (Fig. 3).

Genome sequencing. DNA was extracted using the long - fragment magnetic bead method. The procedure was as follows: Take about 0.3–3 g of tissue sample, grind it into a flour - like consistency in liquid nitrogen, and transfer it into a 5 mL centrifuge tube. Add 3 mL of Tissue Lysis Buffer 1, mix by shaking, and incubate at 65 °C for 30 minutes. After incubation, centrifuge at low temperature for 10 minutes, transfer the supernatant to a new corresponding centrifuge tube, add 1 mL of Precipitation Buffer 2, incubate at low temperature for 5 minutes, centrifuge at low temperature for 10 minutes, and transfer the supernatant to a 2 mL centrifuge tube. Add 900 μL of Lysis Supernatant to each tube, then add 900 μL of DNA Binding Buffer and 50 μL of Magnetic Bead Buffer. Gently invert and mix several times, centrifuge briefly for 2 seconds, let stand at room temperature for 5 minutes to allow binding. Place the 2 mL centrifuge tubes on a magnetic rack until the magnetic beads are adsorbed and the liquid becomes clear. Discard the supernatant, wash the magnetic bead precipitate twice with 75% ethanol, air - dry at room temperature, and add 80–300 μL of pre - warmed Elution Buffer at 50 °C. Gently tap the centrifuge tube to resuspend the magnetic beads, let stand at room temperature for 5 minutes, place on the magnetic rack until the liquid is clear, transfer the supernatant to a new 1.5 mL centrifuge tube, and store it in a refrigerator at 4 °C for nucleic acid quality control. After being disrupted by Megaruptor (approximately 15–20 K), the DNA

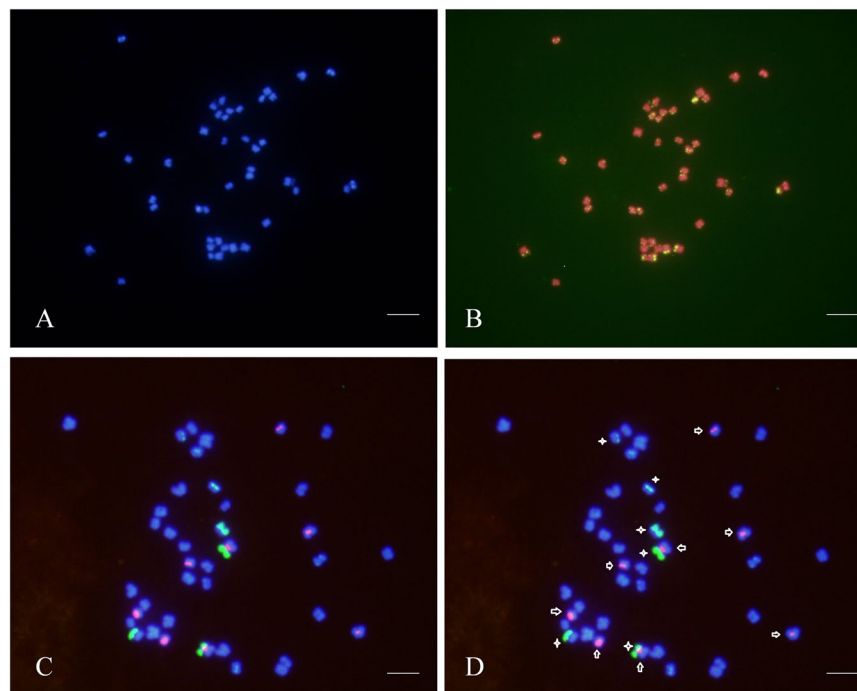


Fig. 2 Chromosomal Fluorescence *In Situ* Hybridization (FISH) Images of *A. rosea*, AB: Telomere Fluorescence *In Situ* Hybridization (FISH) results, showing *in situ* hybridization of telomeric repeat sequences (green). CD: Chromosomal rDNA Fluorescence *In Situ* Hybridization (FISH) results, displaying 5S rDNA in red (indicated by arrows) and 18S rDNA in green (indicated by asterisks). Scale bar: 5 μ m.

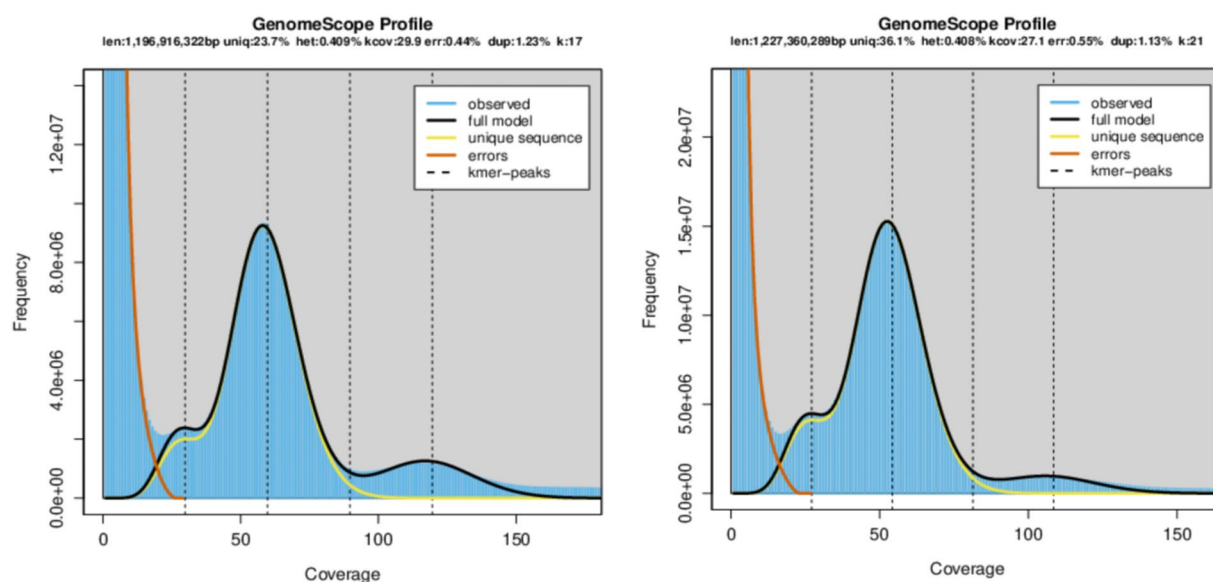


Fig. 3 17,21-kmer Distribution Curves. Note: The x-axis represents the k-mer depth (coverage), while the y-axis denotes the frequency of k-mers at that particular depth.

underwent damage repair and end repair before being ligated with known adaptors. Following enzymatic digestion, selection, and final transformation into a dumbbell-shaped HiFi library, it was subjected to quality checks. Post-qualification, the sample underwent PacBio Sequel II sequencing, subsequent data analysis, and interpretation of results.

Transcriptome sequencing. Transcriptome sequencing is used to assist in genome annotation. Total RNA was extracted from flowers, leaves, roots, stems, and fruits of the same hollyhock plant using the CTAB method²¹.

Type	Number	Rate (%)
Contig		
Complete BUSCOs (C)	2315	99.5
Complete and single-copy BUSCOs (S)	1145	49.2
Complete and duplicated BUSCOs (D)	1170	50.3
Fragmented BUSCOs (F)	5	0.2
Missing BUSCOs (M)	6	0.3
Total BUSCO groups searched	2326	100
Scaffold		
Complete BUSCOs (C)	2326	99.6
Complete and single-copy BUSCOs (S)	1146	49.3
Complete and duplicated BUSCOs (D)	1170	50.3
Fragmented BUSCOs (F)	5	0.2
Missing BUSCOs (M)	5	0.2
Total BUSCO groups searched	2326	100
Gene-set		
Complete BUSCO	1588	98.4
Complete and single-copy BUSCOs	853	52.9
Complete and duplicated BUSCOs	735	45.5
Fragmented BUSCOs	3	0.2
Missing BUSCOs	23	1.4
Total BUSCO groups searched	1614	—

Table 1. Statistics of the evaluation results of genome BUSCO.

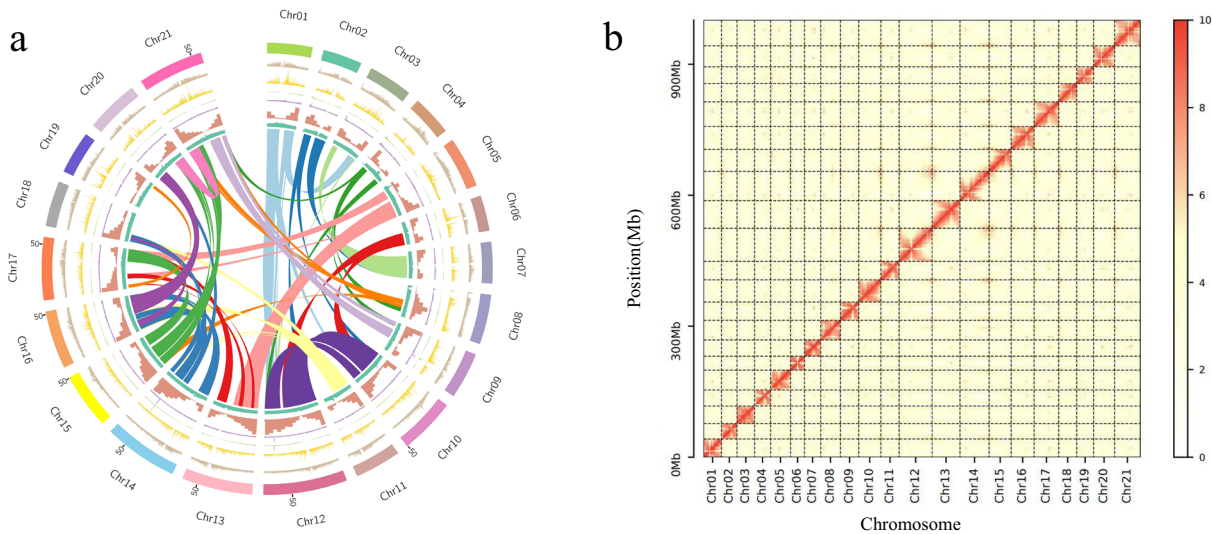


Fig. 4 The genomic features and Hi-C map of *A. rosea*. **(a)** genomic features of *A. rosea* (From outside to inside are genomic collinearity, GC content, gene number, repeat sequence features (transposon, DNA_class.stat, Copia, gypsy) and chromosomes), **(b)** Hi-C map of the *A. rosea*, Color intensity indicates the frequency of Hi-C interaction links from low (yellow) to high (red).

mRNA was then enriched from the total RNA to construct strand-specific transcriptome libraries. These libraries were sequenced using the high-throughput DNBSEQ platform, followed by bioinformatics analysis.

Genome assembly and Hi-C analysis. The PacBio Sequel II platform generated 62.48 Gbp of HiFi long-read data (52 × coverage; 33.117 Gbp and 29.370 Gbp from two cells, respectively), which were assembled with Hifiasm (v0.19.6)²². The resulting contigs spanned 1.09 Gbp with an N50 of 36.61 Mbp. The genome assembly quality was assessed using BUSCO (v5.1.2)²³, with the completeness evaluation library being eudicots_odb10. BUSCO assessment showed that the contig genome completeness was 99.5% (Table 1, Fig. 4).

Hi-C libraries were prepared using DpnII restriction enzyme and sequenced on the DNBSEQ platform, producing 291.85 Gbp of 150 bp paired-end reads. To anchor contigs onto chromosomes, the Hi-C clean data were mapped to the assembled contigs using BWA (v0.7.12)²⁴, and then erroneous mappings (MAPQ = 0) and

Type	Repeat Length(bp)	% of genome
Trf	118191505	11.7
Repeatmasker	150729880	14.92
Proteinmask	115731537	11.45
De novo	516766961	51.14
Total	565840497	56

Table 2. Statistical results of repeat repeats.

Type	Length (bp)	% in genome
DNA	46813547	4.63
LINE	18109441	1.79
SINE	1129454	0.11
LTR	489490650	48.44
Other	2900	0.000287
Unknown	4892388	0.48
Total	539481100	53.39

Table 3. Classification of TEs.

Type	Number
Number of genes	51436
Gene length(bp)	2739.92
CDS length (bp)	1242.54
Exons per gene	5.49
Exons length (bp)	226.23
Intron length (bp)	333.31

Table 4. Gene structure prediction statistics.

duplicates were filtered by the Juicer²⁵ pipeline (v1.5.6) to obtain the interaction matrix. Following, approximately 1.01 Gbp reads were used to anchor the contigs into chromosomes with 3D-DNA²⁶ pipeline (v180,922). And 3D-DNA pipeline was used to remove select short contigs using default parameters. The Hi-C contact maps were then reviewed with Juicebox²⁷ Assembly Tools (v2.15.07).

The chromosome-level genome assembly of *A. rosea* was generated by integrating PacBio HiFi long reads (62.48 Gbp, 52 × coverage) and Hi-C data (291.85 Gbp). The final assembly spans 1.01 Gbp across 21 scaffolds (scaffold N50 = 52.57 Mbp), with a contig N50 of 36.61 Mbp and a BUSCO completeness of 99.6% (eudicots_odb10) (Table 1).

Repeat sequence annotation. The methods for repeat sequence annotation are categorized into two types: homology-based alignment and de novo prediction. Homology-based alignment relies on the RepBase²⁸, RepeatMasker (v4.1.2)²⁹ is used to identify and classify sequences similar to known repeat sequences. For de novo prediction, a novel repeat sequence library is first constructed using software such as RepeatModeler (v2.0.3)³⁰ and LTRharvest (v1.6.2)³¹, and then predictions are made using RepeatMasker. In addition, TRF (4.10.0)³² is utilized to locate tandem repeat sequences within the genome. A total of 565.84 Mbp (56% of the genome) of repetitive sequences were identified (Table 2), with transposable elements (TEs) predominating, among which LTR elements accounted for 48.44% of the genome (Table 3).

Gene structure prediction. Gene structure prediction integrates three approaches: homology-based prediction, de novo prediction, and transcriptome-assisted prediction. Homology-based prediction utilizes annotation information from six closely related species: *H. cannabinus*, *G. raimondii*, *G. hirsutum*, *Corchorus. capsularis*, *C. olitorius*, and *T. cacao*. Firstly, GeMoMa (v1.9)³³ is employed for homology prediction. Based on the results of homology prediction, genes with complete structures are selected for training the de novo prediction software, Augustus (v3.0.3)³⁴ and SNAP (v11/29/2013)³⁵. RNA-seq data is aligned using HISAT (v2.0.4)³⁶ and assembled into transcripts with StringTie (v1.2.2)³⁷. Finally, EvidenceModeler (v1.1.1)³⁸ is used to consolidate all data, resulting in the final gene set. The completeness and proportion of fully predicted genes are assessed by comparing the gene set against the embryophyta_odb10 database using BUSCO. Ultimately, 51,436 genes were annotated. Among these predicted genes, the average gene length and coding sequence (CDS) length were 2739.92 bp and 1242.54 bp, respectively (Table 4). The average number of exons per gene was 5.49, with average exon and intron lengths of 226.23 bp and 333.31 bp, respectively. The BUSCO completeness assessment of the gene set revealed

	Number	Percentage(%)
Total	51,436	100
Nr-Annotated	47,889	93.1
Swissprot-Annotated	39,623	77.03
KEGG-Annotated	38,407	74.67
KOG-Annotated	38,218	74.3
TrEMBL-Annotated	47,899	93.12
Interpro-Annotated	46,386	90.18
GO-Annotated	31,171	60.6
Overall	48,045	93.41

Table 5. Genome function annotation result.

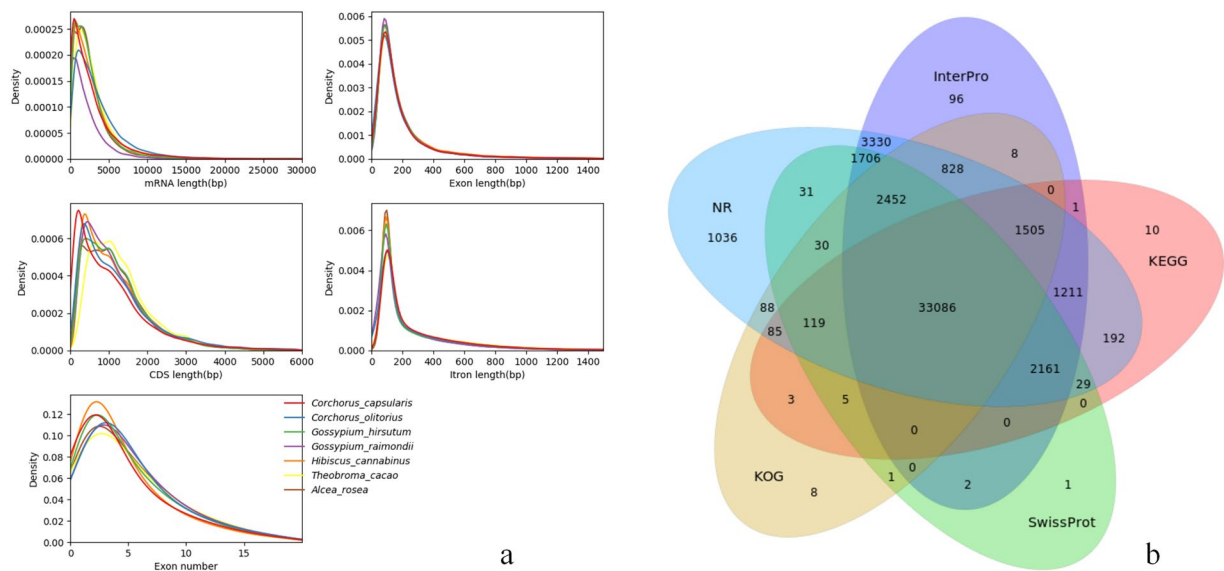


Fig. 5 Gene structure prediction and gene function annotation. (a) Statistical plots of the prediction results of the gene structure, (b) Gene functional annotation results for number statistics in five databases: NR, InterPro, KEGG, SwissProt and KOG.

coverage of 1,588 conserved proteins, with a completeness of 98.4%, including 853 (52.9%) single-copy genes and 735 (45.5%) multi-copy genes (Table 1).

Functional annotation of genes. Functional annotation of genes is performed using InterProScan (v5.28-67.0)³⁹ to search secondary structural domain databases and obtain gene function information. The comparison databases used include SwissProt⁴⁰, TrEMBL, KEGG⁴¹, InterPro, NR, KOG, and GO⁴². The functional annotation statistics of the predicted genes in seven databases (Nr, Swissprot, KEGG, KOG, TrEMBL, Interpro, GO) are shown in the Table 5. The Venn diagram illustrating the number of gene annotations across five databases (Nr, Interpro, KEGG, Swissprot, and KOG) is shown in the Fig. 5b, with 33,086 genes annotated in all five databases. The statistical results of KOG functional annotation categorization indicated that the five most represented gene functions were ‘General function prediction only’, ‘Signal transduction mechanisms’, ‘Transcription’, ‘Function unknown’, and ‘Posttranslational modification, protein turnover, chaperones’ (Fig. S1). The KEGG pathway annotation results showed that the five most involved metabolic pathways were global and overview maps, carbohydrate metabolism, translation, folding, sorting, and degradation, and signal transduction - membrane transport (Fig. S2). The GO secondary node annotation classification statistics indicated that the most represented categories in terms of gene number were binding, cellular processes, catalytic activity, metabolic processes, and cellular component (Fig. S3).

Data Records

The genome sequence data used for genome assembly and annotation have been deposited in the Genome Sequence Archive (GSA) of the National Genomics Data Center (NGDC) under accession number CRA024326⁴³. Both the raw reads and chromosome assembly have been submitted to NCBI with the accession number of SRP590903⁴⁴ and GCA_050717195.1⁴⁵. Annotation files can be accessed on Figshare⁴⁶.

Technical Validation

We assessed the genome integrity using BUSCO, using the comparison library eudicots_odb10, the assembly integrity of contig genome was 99.5%, and the assembly integrity of scaffold genome was 99.6%, and the number and proportion of fully predicted genes were evaluated. The results showed that the gene set fully covered 1,588 conserved proteins (98.4% integrity), of which 863 (52.9%) were single copies and 735 (45.5%) were multiple copies. The three evaluation results showed that the genome assembly quality of hollyhock was high.

Code availability

Software programs and pipelines were conducted as specified in the instruction manuals and published protocols of bioinformatic tools. Detailed information on software versions, code, and parameters can be found in the Methods section. No specific or custom code was used in this study.

Received: 4 March 2025; Accepted: 26 June 2025;

Published online: 04 July 2025

References

- Wang, D. H. *et al.* MaGenDB: a functional genomics hub for Malvaceae plants. *Nucleic Acids Research* **48**, D1076–D1084, <https://doi.org/10.1093/nar/gkz953> (2019).
- Wang, M. J. *et al.* Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nature Genetics* **51**, 224–229, <https://doi.org/10.1038/s41588-018-0282-x> (2019).
- Sun, M. Y., Gu, Y. Y., Glisan, S. L. & Lambert, J. D. Dietary Cocoa Ameliorates Non-Alcoholic Fatty Liver Disease and Increases Markers of Antioxidant Response and Mitochondrial Biogenesis in High Fat-Fed Mice. *The Journal of Nutritional Biochemistry* **92**, 108618, <https://doi.org/10.1016/j.jnutbio.2021.108618> (2021).
- Seyyednejad, S. M., Koochak, H. & Darabpour, E. A survey on *Hibiscus rosa-sinensis*, *Alcea rosea* L. and *Malva neglecta* Wallr as antibacterial agents. *Asian Pacific Journal of Tropical Medicine*, 351–355 [https://doi.org/10.1016/S1995-7645\(10\)60085-5](https://doi.org/10.1016/S1995-7645(10)60085-5) (2010).
- Mert, T., Fafal, T., Kivak, B. & Yalin, H. T. Antimicrobial and Cytotoxic Activities of the Extracts Obtained from the Flowers of *Alcea Rosea* L. *Hacettepe University Journal of the Faculty of Pharmacy* **30**, 17–24, https://www.researchgate.net/publication/265098890_Antimicrobial_and_Cytotoxic_Activities_of_the_Extracts_Obtained_from_the_Flowers_of_Alcea_Rosea_L (2010).
- Ammar, N., El-Hakeem, R. A., El-Kashoury, E. S. & El-Kassem, L. A. Evaluation of the phenolic content and antioxidant potential of *Althaea rosea* cultivated in Egypt. *Journal of The Arab Society for Medical Research* **8**, 48, <https://doi.org/10.4103/1687-4293.123786> (2013).
- Hosaka, H., Mizuno, T. & Iwashina, T. Flavonoid Pigments and Color Expression in the Flowers of Black Hollyhock (*Alcea rosea* ‘Nigra’). *Bulletin of the National Museum of Nature and Science* **38**, 69–75 (2012).
- Ma, Z. *et al.* High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement. *Nat Genet* **53**, 1385–1391, <https://doi.org/10.1038/s41588-021-00910-2> (2021).
- Zhang, T. *et al.* Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat Biotechnol* **33**, 531–537, <https://doi.org/10.1038/nbt.3207> (2015).
- Argout, X. *et al.* The cacao Criollo genome v2.0: an improved version of the genome for genetic and functional genomic studies. *BMC Genomics* **18**, 730, <https://doi.org/10.1186/s12864-017-4120-9> (2017).
- Cornejo, O. E. *et al.* Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. *Commun Biol* **1**, 167, <https://doi.org/10.1038/s42003-018-0168-6> (2018).
- Kim, Y. M. *et al.* Genome analysis of *Hibiscus syriacus* provides insights of polyploidization and indeterminate flowering in woody plants. *DNA Res* **24**, 71–80, <https://doi.org/10.1093/dnares/dsw049> (2017).
- Yang, Y. Z. *et al.* A High-Quality, Chromosome-Level Genome Provides Insights Into Determinate Flowering Time and Color of Cotton Rose. *Frontiers in plant science* **13**, 818206, <https://doi.org/10.3389/fpls.2022.818206> (2022).
- Gao, W. *et al.* Transcriptome analysis in *Alcea rosea* L. and identification of critical genes involved in stamen petaloid. *Scientia Horticulturae* **293**, <https://doi.org/10.1016/j.scienta.2021.110732> (2022).
- Luo, Y. *et al.* Transcriptomics analyses reveal the key genes involved in stamen petaloid formation in *Alcea rosea* L. *BMC Plant Biol* **24**, 551, <https://doi.org/10.1186/s12870-024-05263-6> (2024).
- Wang, Y. *et al.* Genetic Diversity and Population Structure Analysis of Hollyhock (*Alcea rosea* Cavan) Using High-Throughput Sequencing. *Horticulturae* **9**, <https://doi.org/10.3390/horticulturae9060662> (2023).
- Shao, L. *et al.* High-quality genomes of *Bombax ceiba* and *Ceiba pentandra* provide insights into the evolution of Malvaceae species and differences in their natural fiber development. *Plant Communications* **5**, 100832, <https://doi.org/10.1016/j.xplc.2024.100832> (2024).
- Luo, R., Liu, B., Xie, Y., Li, Z. & Liu, Y. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, <https://doi.org/10.1186/2047-217X-1-18> (2012).
- Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764, <https://doi.org/10.1093/bioinformatics/btr011> (2011).
- Vurtture *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btx153> (2017).
- Gambino, G., Perrone, I. & Griboudo, I. A Rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants. *Phytochemical Analysis* **19**, 520–525, <https://doi.org/10.1002/pca.1078> (2008).
- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175, <https://doi.org/10.1038/s41592-020-01056-5> (2021).
- Manni, M., Berkeley, M. R., Seppely, M., Simao, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* **38**, 4647–4654, <https://doi.org/10.1093/molbev/msab199> (2021).
- Heng, L. & Richard, D. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **14**, <https://doi.org/10.1093/bioinformatics/btp324> (2010).
- Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98, <https://doi.org/10.1016/j.cels.2016.07.002> (2016).
- Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92, <https://doi.org/10.1126/science.aal332> (2017).
- Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**, 99–101, <https://doi.org/10.1016/j.cels.2015.07.012> (2016).
- Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic & Genome Research* **110**, 462–467, <https://doi.org/10.1159/000084979> (2005).

29. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics* **25**, <https://doi.org/10.1002/0471250953.bi0410s25> (2009).
30. Flynn, J. M., Hubley, R., Rosen, J., Clark, A. G. & Smit, A. F. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 201921046, <https://doi.org/10.1073/PNAS.1921046117> (2020).
31. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18, <https://doi.org/10.1186/1471-2105-9-18> (2008).
32. Gary, B. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580, <https://doi.org/10.1093/nar/27.2.573> (1999).
33. Jens, K. *et al.* Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research* **44**, 89, <https://doi.org/10.1093/nar/gkw092> (2016).
34. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, 435–439, <https://doi.org/10.1093/nar/gkl200> (2006).
35. Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research* **35**, 3823–3835, <https://doi.org/10.1093/nar/gkm238> (2007).
36. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357–360, <https://doi.org/10.1038/nmeth.3317> (2015).
37. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**, 290–295, <https://doi.org/10.1038/nbt.3122> (2015).
38. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7, <https://doi.org/10.1186/gb-2008-9-1-r7> (2008).
39. Philip, J. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 1236–1240 <https://doi.org/10.1093/bioinformatics/btu031> (2014).
40. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* **31**(1), 365–370, <https://doi.org/10.1093/nar/gkg095> (2003).
41. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30, <https://doi.org/10.1093/nar/28.1.27> (2000).
42. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29, <https://doi.org/10.1038/75556> (2000).
43. NGDC Genome Sequence Archive <https://ngdc.cncb.ac.cn/gsub/submit/gsa/subCRA039584> (2025).
44. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRP590903> (2025).
45. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_050717195.1 (2025).
46. Chen, X. *et al.* Chromosome-level genome assembly of the ornamental plant *Alcea rosea*. Figshare <https://doi.org/10.6084/m9.figshare.28801490> (2025).

Acknowledgements

This work was supported by the Chengdu Park City Development Administration. We thank BGI for their technical services.

Author contributions

Xi Chen and Xiaoqing Shi conceived and designed the experiments. Xi Chen, Xiaoqing Shi, Xiu Li performed the research. Shengwen Tang, Jiao Ma, Zhangshun Zhu and Fangwen Li analysed the data. Xi Chen and Xiaoqing Shi wrote the paper. All authors approved this manuscript.

Competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-025-05473-z>.

Correspondence and requests for materials should be addressed to X.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025