scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly and annotation of *Cercis chuniana* (Fabaceae)

Wei Gong[1,3], Wanzhen Liu[2,3], Mengxue Feng[1,3], Chuxing Liang[1], Zhilong Huang[1], Meixiu Lin[1], Qiongqiong Lin[1], Rongjing Zhang[1] ✉ & Hanghui Kong[2] ✉

The genus *Cercis* L., at the base of the subfamily Cercidoideae of Fabaceae, is known for its ecological adaptability and significant medicinal, ornamental, and economic value. However, the lack of a high-quality genome hinders the understanding of the evolution of *Cercis* and Fabaceae. In this study, we present a chromosome-level genome of *Cercis chuniana* by combining Illumina short reads, PacBio HiFi long reads, and Hi-C data. The final genome size is 355.53 Mb, consisting of 12 contigs with a N50 of 42.34 Mb. Notably, 344.24 Mb, corresponding to 96.82% of the genome, was anchored to seven chromosomes. The assembly comprises 24.83% repetitive sequences, including 19.32% long terminal repeats. Additionally, a total of 33,837 protein-coding genes were predicted in the genome, with 32,709 (96.67%) genes successfully annotated. The high-quality genome assembly of *C. chuniana* not only bridges the existing gap in genomic data and offers important resources for molecular studies of this species, but also provides essential insights for future studies on speciation, functional and comparative genomics within the Fabaceae family.

Background & Summary

The Fabaceae family is the third-largest family of angiosperms, comprising approximately 770 genera and 19,500 species, which are globally distributed and adapted to diverse ecological environments[1,2]. Fabaceae plants possess considerable economic importance, being widely used in food, medicine, fodder, timber, industry, and soil stabilization[3–5]. In recent years, advances in sequencing technology have significantly accelerated phylogenetic and genomic research in Fabaceae[2,6–10].

Cercis L., an important genus within the Fabaceae family, comprises nine species known for their ornamental and medicinal value[11,12]. *Cercis* species are widely distributed across temperate and subtropical regions, adapting to various ecological environments and exhibiting unique morphological characteristics[13]. Consequently, *Cercis* has become one of the most popular choices for ornamental landscaping. However, despite their medicinal and horticultural significance, genomic research on *Cercis* is relatively limited, with only a few species having their genomes sequenced to date[6,10]. Among these species, *C. chuniana*, a small tree or shrub occurring in subtropical evergreen broadleaved forest, shows an asymmetrical leaf blade[14,15], which makes it easily identified morphologically. Based on phylogenetic reconstruction, all sampled populations of *C. chuniana* form a monophyletic clade and is distinguished as one of the earliest diverging species in the *Cercis* genus[16,17]. Therefore, both morphological and phylogenetic evidence showed that *C. chuniana* is a well-resolved species[17]. Given its morphological distinctiveness and phylogenetic position, a comprehensive genomic analysis of *C. chuniana* is essential for understanding its evolutionary history and speciation processes.

To facilitate phylogenetic and speciation studies of *C. chuniana*, a chromosome-scale genome was produced using Illumina short-reads, PacBio long-reads, and Hi-C sequencing data. A contig-level genome assembly consists of 12 contigs with an N50 of 42.34 Mb. After Hi-C scaffolding, seven scaffolds were obtained with an N50 of 49.72 Mb, representing a substantial improvement over the published genome[6,10], which had N50 of 421.03 kb and 47.43 Mb, respectively. These seven scaffolds were anchored to seven chromosomes, spanning 344.24 Mb and corresponding to 96.82% of the assembled sequences. The genome comprises 24.83% repetitive sequences,

[1]College of Life Sciences, South China Agricultural University, Guangzhou, 510614, China. [2]Key Laboratory of National Forestry and Grassland Administration on Plant Conservation and Utilization in Southern China, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. [3]These authors contributed equally: Wei Gong, Wanzhen Liu, Mengxue Feng. ✉e-mail: zhangrongj@scau.edu.cn; konghh@scbg.ac.cn

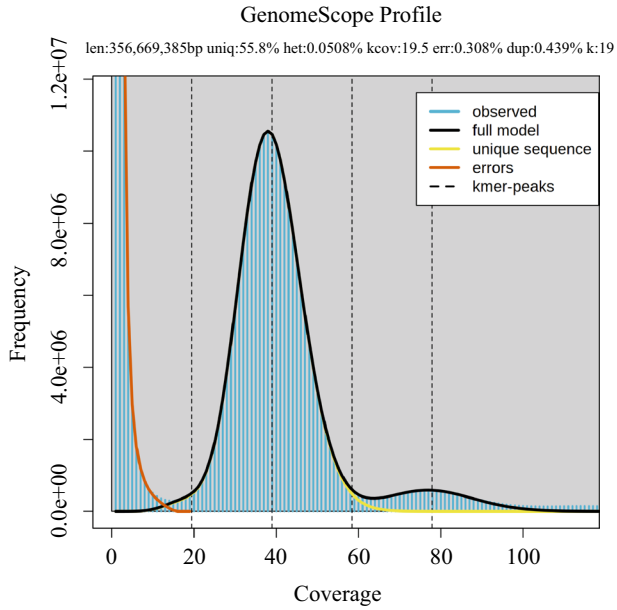

Fig. 1 Genome survey of *Cercis chuniana* based on the Kmer distribution analysis using a Kmer size of 19. The horizontal axis represents the Kmer depth, while the vertical axis represents the corresponding number of Kmers at each depth.

and 33,837 protein-coding genes, of which 32,709 (96.67%) were functionally annotated. This high-quality genome assembly offers a valuable resource for molecular studies of *C. chuniana*, providing foundational data for future studies in speciation, functional genomics, and comparative genomics within the Fabaceae family.

Methods

Sample collection and DNA sequencing. Fresh, young leaves of *Cercis chuniana* were collected from Dadongshan Nature Reserve, Qingyuan, Guangdong, China (24°55′42″ N, 112°43′41″ E), and quickly flash-frozen in liquid nitrogen for several hours. Then, the samples were stored in a −80 °C freezer before being sent to Novogene Corporation (Tianjin, China) for sequencing. A voucher specimen of *C. chuniana* was stored at the herbarium of College of Life Sciences, South China Agriculture University (SCAUB), with a voucher number of MX221002. Total genomic DNA was extracted from the frozen leaves using CTAB method[18] and was purified using phenol/chloroform/ isoamyl alcohol mixture to remove protein contaminants. The DNA was then precipitated, washed with ethanol, and resuspended in TE buffer for genome sequencing.

Library construction and genome sequencing. For Illumina sequencing, a library with a short-fragment size of 350 bp was prepared using the Covaris ultrasonicator. The DNA fragments were further treated with end repair, A-tailing, adapter ligation, purification, and PCR amplification. The prepared library was sequenced on an Illumina NovaSeq 6000 platform (Illumina, USA) to generate 150-bp paired-end (PE150) reads.

For PacBio HiFi sequencing, the purified DNA samples were randomly fragmented with a Covaris ultrasonicator into 15–18 kb sequences. These DNA fragments were enriched and purified using magnetic beads. After damage and end repair of the DNA fragments, stem-loop sequencing adapters were ligated to both ends of fragments. After quality control checks, the constructed libraries were used for PacBio HiFi sequencing.

Hi-C library preparation and sequencing followed the protocol of Belton *et al.*[19] with minor modifications. Cells were treated with paraformaldehyde to fix the DNA conformation, and the DNA was then digested with the restriction enzyme DPN II (GATC) to produce sticky ends. After end repair, purification, and adaptor ligation, the Hi-C library was sequenced on the Illumina NovaSeq 6000 platform (Illumina, USA). In total, 105.68 Gb of sequencing data was generated, including 20.92 Gb (58.84×) of Illumina reads, 14.78 Gb (41.57×) of PacBio HiFi reads, 48.99 Gb (137.79×) of Hi-C data, and 20.99 Gb of transcriptome data. The PacBio HiFi long reads had a scaffold N50 of 17.83 kb, with an average length of 17.91 kb.

Genome size estimation. The genome size of *C. chuniana* was estimated using a k-mer (k = 19) analysis-based approach with Illumina PE short reads. Jellyfish v2.1.4[20] was employed to count k-mer in the DNA sample. The GenomeScope (http://qb.cshl.edu/genomescope/) was then used to estimate the genome size. A 19-mer analysis of the sequenced genome revealed that *C. chuniana* had a total genome size of 357 Mb with 0.05% heterozygosity (Fig. 1).

Genome assembly. Quality control of the raw Illumina data was performed using fasp[21]. The quality control process involved filtering out reads containing adapter sequences, removing reads with more than 10% of undetermined nucleotide bases, and excluding paired-end reads where either read contains more than 20% of low-quality bases in the single-end sequencing read.

	Statistics
Number of contigs	12
Number of scaffolds	7
Genome size (bp)	355,533,734
Contig N50 (bp)	42,339,124
Scaffold N50 (bp)	49,715,743
Chromosome/total	7
GC content (%)	36.02%
Protein-coding gene	33,837

Table 1. Summary of the *Cercis chuniana* genome assembly.

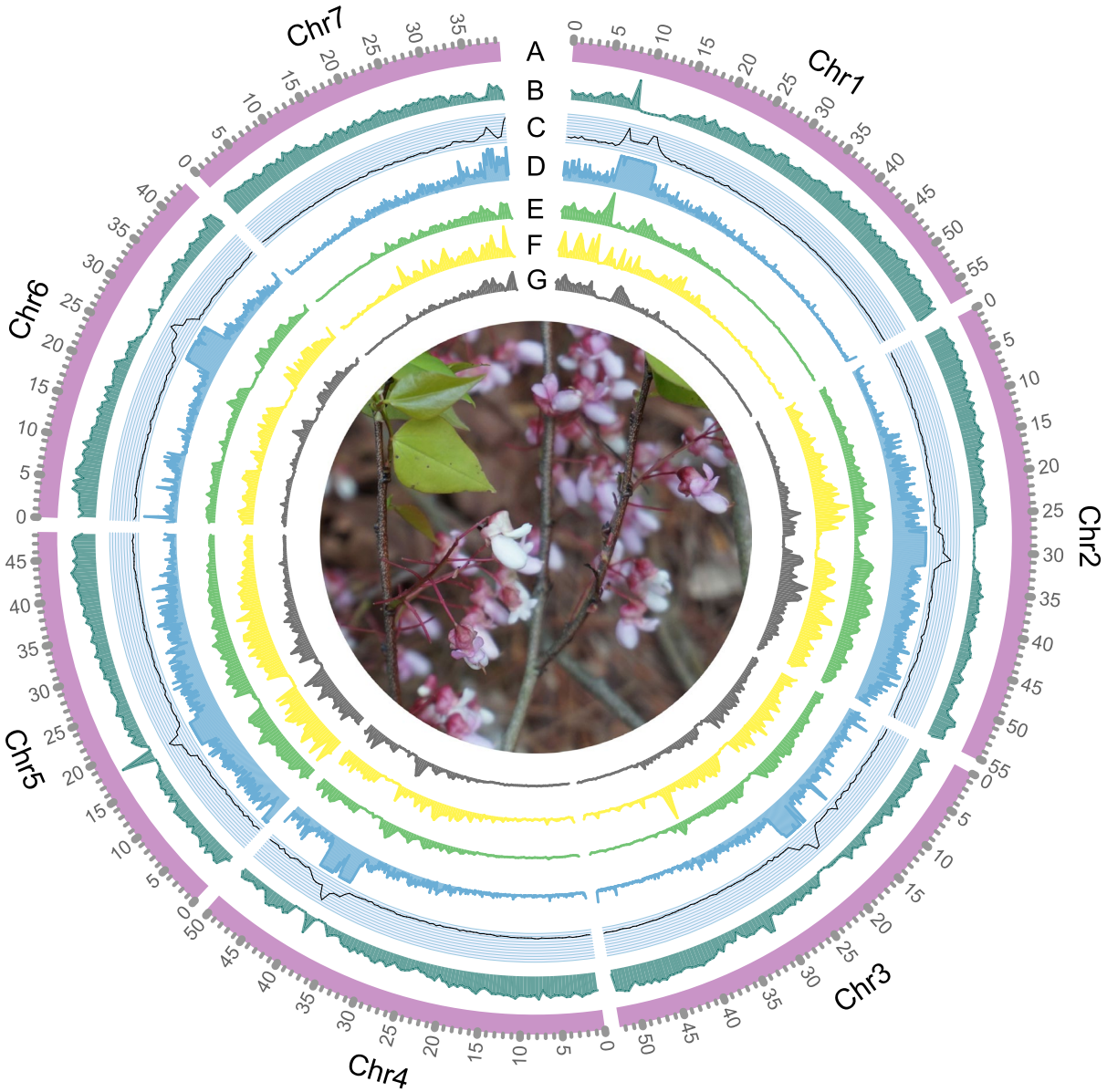

Fig. 2 The genomic landscape of seven pseudochromosomes of *Cercis chuniana*. Circles from outside to inside: A, seven super-scaffolds. Lengths are shown in Mb. B, gene number; C, GC content; D, repeat density; E, total LTR density; F, LTR-*Copia* density; G, LTR-Gypsy density.

The primary assembly of PacBio HiFi long reads was generated using Hifiasm v0.19.8[22] with the default parameters. To eliminate heterozygous and redundant regions, BWA[23] and Purge_Dups v1.2.5[24] were used with the parameters ("-s 70"). After conducting quality control, Hi-C reads were aligned to the assembly

Type	Length (bp)	Percent of the genome (%)
DNAs	11,873,486	3.34
LINEs	3,682,585	1.04
SINEs	353,725	0.10
LTRs	68,678,540	19.32
Other	806	0.00023
Unknown	6,950,182	1.95
Total repeats	88,290,758	24.83

Table 2. Annotation of repeat elements in the *Cercis chuniana* genome. Abbreviations: DNAs, DNA elements; LINEs, long interspersed nuclear elements; SINEs, short interspersed nuclear elements; LTRs, long terminal repeats.

	Total	Nr	Swissprot	KEGG	COG	TrEMBL	GO	Overall
Number	33,837	31,704	23,187	24,095	23,715	31,374	17,569	32,709
Percentage	100%	93.70%	68.53%	71.21%	70.09%	92.72%	51.92%	96.67%

Table 3. The genome functional annotation results of *Cercis chuniana*.

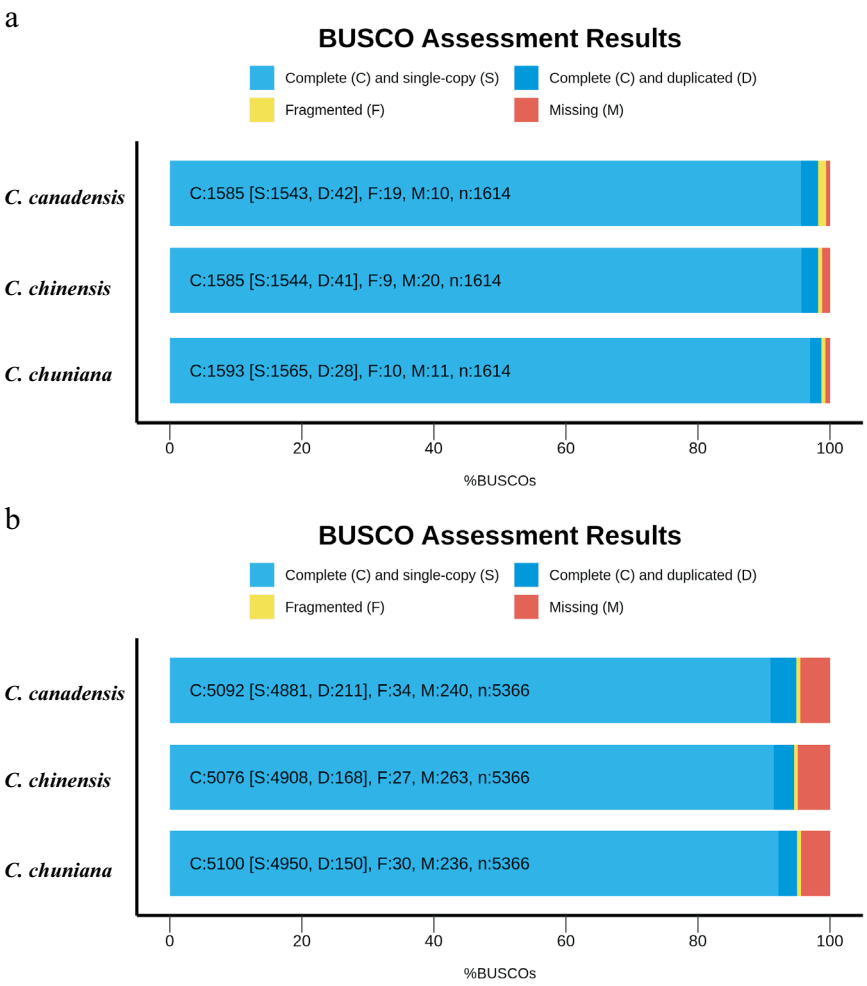

Fig. 3 BUSCO assessment results of the whole genome assembly in *Cercis chuniana* based on eudicots_odb10 database (**a**) and fabales_odb10 database (**b**). The number of complete (C) orthologs (including single-copy (S) and duplicated (D) orthologs) fragmented (F) orthologs, and missing (M) orthologs for *C. canadensis*, *C. chinensis*, and *C. chuniana* are shown in each bar.

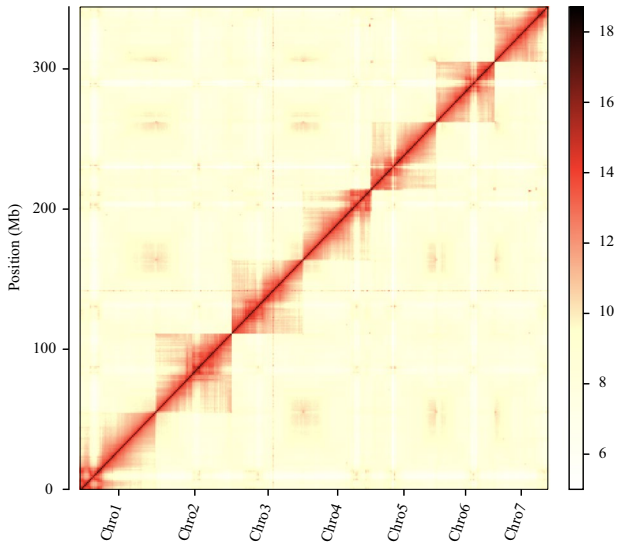

Fig. 4 Heat map of Hi-C assembly of *Cercis chuniana*. The scale bar represents the interaction frequency of Hi-C links.

using Juicer v1.6[25], and the primary contigs were anchored into chromosomes using 3D-DNA v180922[26]. The contig assembly result was carefully reviewed, and any assembly errors were manually corrected using Juicebox v2.20.00[27]. The completeness and contiguity of genome assemblies were evaluated using the BUSCO v5.4.7[28] and the LTR Assembly Index (LAI)[29]. The final genome assembly of the *C. chuniana* achieved the chromosome-level assembly resolution, with a size of 355.53 Mb, consisting of 12 contigs and seven scaffolds. The GC content of the assembly was 36.02%, with N50 sizes of 42.34 Mb for contigs and 49.72 Mb for scaffolds (Table 1; Fig. 2).

Genome annotation. For the annotation of repetitive elements in the *C. chuniana* genome, a *de novo* annotated repetitive sequence database was created using RepeatScout v1.0.6[30], LTR_FINDER[31], and RepeatModeler v2.0.5[32]. This database was then aligned with the genome using RepeatMasker v4.1.6[33] (http://www.repeatmasker.org/RepeatMasker) to identify and mask repeat elements. Additionally, tandem repeats were identified using Tandem Repeats Finder v4.0.9[34], and RepeatMasker and RepeatProteinMask v4.1.6[35] were used to identify above tandem repeats in the genome through homology-based alignment with the RepBase database (http://www.girinst.org/repbase). The results from *de novo* prediction and homology-based alignment were integrated, and the repetitive sequences were classified and statistically analyzed. The analysis revealed that approximately 24.83% of *C. chuniana* genome consists of repetitive elements, including LTRs (19.32%), DNA transposons (3.34%), unclassified elements (1.95%), LINEs (1.04%), SINEs (0.10%), and other elements (Table 2).

For protein-coding gene annotation in *C. chuniana*, gene regions were identified using Maker v2.31.8[36] based on transcribed RNA, *ab initio* gene predictions, and homologous proteins. For the transcriptome-based gene prediction, HISAT2 v2.2.1[37] was used to align RNA transcripts, and StringTie v2.1.6[38] was employed to generate a genome-guided annotation files from RNA-seq alignment production. TransDecoder v5.7.1 (https://github.com/TransDecoder/TransDecoder) was then employed to predict coding regions. Diamond v0.8.36[39] was used to align sequences against the NR (Non-Redundant Protein Sequence), UniProt (Universal Protein), and OrthoDB database, filtering for sequences with alignment consistency percentages above 0.8 and e-values less than 1e−5.

For homology-based annotation, protein sequences from *Arabidopsis thaliana, Bauhinia variegate*, *Cercis chinensis*, *Cercis canadensis*, *Glycine max*, *Oryza sativa*, and *Trifolium pretense* were aligned with the genome of *C. chuniana* using Miniprot[40]. The homology-based alignment files, transcriptome alignment files, and repetitive sequence annotations were integrated as evidence for *ab initio* gene prediction using the MAKER v2.31.8 pipeline[36], SNAP v1.0[41], Augustus v3.3.1[42], and GeneMark[43] were employed to perform *ab initio* gene prediction based on Hidden Markov models. Subsequently, MAKER was run again to integrate the three sources of evidence and generate a non-redundant set of gene models. In total, 33,837 protein-coding genes were annotated in the *C. chuniana* genome (Table 1).

Functional annotation of the coding genes was conducted through homology-based searches against the NCBI non-redundant (NR) protein database using BLASTP v2.2.28[44]. Furthermore, InterProScan 5.48–83.0[45] and DIAMOND v2.0.8[39] (--evalue 1e-5--max-target-seqs. 5) were employed to assign functional annotations based on Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), clusters of orthologous groups (COG), Swiss-Prot, and TrEMBL databases. As a result, a total of 32,709 genes (96.67% of the predicted protein-coding genes) were functionally annotated, including 17,569 with GO terms, 24,095 with KEGG annotations, 31,704 with NR protein hits, and 23,715 with COG classifications (Table 3).

Data Records

All high-throughput sequencing data generated in this study have been deposited in the National Genomics Data Center (NGDC) database under the project accession number PRJCA029422. The clean sequencing data from the genome libraries are available in the Genome Sequence Archive (GSA) under the accession numbers CRA018567 (HiFi), CRA018586 (HiC), and CRA018585 (RNA-seq)[46–48]. The genome assembly has been deposited in the NCBI under accession number JBPDIX000000000[49]. The annotation file is available from the figshare archive[50].

Technical Validation

The assembled genome of *Cercis chuniana* was evaluated using the BUSCO v5.4.7[28] with the eudicots_odb10 and fabales_odb10 databases. Among the 1,614 core eudicot genes and 5,366 core Fabales genes, 98.70% (including 96.96% single-copy genes and 1.74% duplicated genes) and 95.04% (including 92.25% single-copy genes and 2.79% duplicated genes), respectively, were successfully identified in the *C. chuniana* genome. Compared with *C. canadensis* and *C. chinensis*, *C. chuniana* exhibited a higher proportion of complete BUSCOs in both databases (Fig. 3), indicating an improvement in assembly completeness. Furthermore, 96.82% of the Hi-C data were successfully anchored to the seven pseudo-chromosomes, further validating the accuracy of the chromosome-level assembly (Fig. 4). The organization of interaction contacts within and between chromosome regions was clearly visualized in the Hi-C heatmap, providing additional support for the structural integrity of the assembly. In addition, the final LTR Assembly Index (LAI) score was 18.57, suggesting that the assembly quality of *C. chuniana* meets the standard for reference-grade genomes.

Code availability

No specific script was used in this study. The codes and pipelines used in data processing were all executed according to the manuals and protocols of the published bioinformatics tools. The version and parameters of software are described in the Methods section.

Received: 11 September 2024; Accepted: 1 July 2025;
Published online: 08 July 2025

References

1. Lewis, G. P., Wood, J. R. I. & Lavin, M. *Steinbachiella* (Leguminosae: Papilionoideae: Dalbergieae), endemic to Bolivia, is reinstated as an accepted genus. *Kew Bull.* **67**, 789–796 (2012).
2. Legume Phylogeny Working Group. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon.* **66**, 44–77 (2017).
3. Li, J. S., Zhao, S., Ren, C. S. & Cercis, L. Germplasm resources and research progress. *J. Northwest Forestry Uni.* **6**, 145–152 (2021).
4. Yang, Y. *et al*. Study on embryo rescue of interspecific hybridization in *Cercis* Linn. *Plant Physiol. J.* **11**, 2173–2180 (2022).
5. Meng, Y. *et al*. Phytochemistry and pharmacology of genus *Cercis*: *A Review*. *Rec. Nat. Prod*. (2024).
6. Griesmann, M. *et al*. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science.* **361**, 1743 (2018).
7. Egan, A. N. & Vatanparast, M. Advances in legume research in the genomics era. *Aust. Syst. Bot.* **32**, 459–483 (2019).
8. Koenen, E. J. *et al*. Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. *New Phytol.* **225**, 1355–1369 (2020).
9. Zhao, Y. *et al*. Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. *Mol. Plant.* **14**, 748–773 (2021).
10. Li, J. *et al*. The nearly complete assembly of the *Cercis chinensis* genome and Fabaceae phylogenomic studies provide insights into new gene evolution. *Plant Commun.* **4**, 100422 (2023).
11. Hao, G. *et al*. A phylogenetic and biogeographic study of *Cercis* (Leguminosae). *Acta Bot. Sinica.* **43**, 1275–1278 (2001).
12. Fritsch, P. W. & Cruz, B. C. Phylogeny of *Cercis* based on DNA sequences of nuclear ITS and four plastid regions: Implications for transatlantic historical biogeography. *Mol. Phylogenet. Evol.* **62**, 816–825 (2012).
13. Wunderlin, R.P., Larsen, K., Larsen, S.S. Cercideae. In: Polhill, R. M., Raven, P. H. (Eds.), *Advances in Legume Systematics*. Part 1. Royal Botanic Gardens, Kew, pp. 107–116 (1981).
14. Chen, D. Z. *et al*. *Cercis* Linnaeus, Sp. Pl. 1:374. 1753. *Flora China.* **10**, 5–6 (2010).
15. Metcalf, F. P. *Cercis chuniana* F.P.Metcalf. *Lingnan Sci. J.* **19**, 551 (1940).
16. Fritsch, P. W. *et al*. Leaf adaptations and species boundaries in North American *Cercis*: Implications for the evolution of dry floras. *Am. J. Bot.* **105**, 1577–1594 (2018).
17. Liu, W. Z. *et al*. Population dynamics linked to glacial cycles in *Cercis chuniana* F. P. Metcalf (Fabaceae) endemic to the montane regions of subtropical China. *Evol. Appl.* **14**, 2647–266 (2021).
18. Doyle, J. J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).
19. Belton, J. M. *et al*. Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods.* **58**, 268–276 (2012).
20. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* **27**, 764–770 (2011).
21. Chen, S. *et al*. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* **34**, 884–890 (2018).
22. Cheng, H. *et al*. Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnol.* **40**, 1332–1335 (2022).
23. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/pdf/1303.3997v2.pdf (2013).
24. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics.* **36**, 2896–2898 (2020).
25. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016a).
26. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science.* **356**, 92–95 (2017).
27. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016b).
28. Waterhouse, R. M. *et al*. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
29. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, 126–126 (2018).

30. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics.* **21**(suppl 1), 351–358 (2005).
31. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**(suppl_2), 265–268 (2007).
32. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA.* **117**, 9451–9457 (2020).
33. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. http://www.repeatmasker.org (2015).
34. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
35. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protoc. Bioinformatics* **25**, 4.10. 11–14.10. 14 (2009).
36. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics.* **12**, 1–14 (2011).
37. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods.* **12**, 357–360 (2015).
38. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
39. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods.* **12**, 59–60 (2015).
40. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics.* **39**, 14 (2023).
41. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics.* **5**, 1–9 (2004).
42. Stanke, M. *et al*. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics.* **7**, 1–11 (2006).
43. Lomsadze, A. *et al*. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).
44. Trapnell, C. *et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
45. Finn, R. D. *et al*. InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
46. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA018567 (2025).
47. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA018586 (2025).
48. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA018585 (2025).
49. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBPDIX000000000.1 (2025).
50. Kong, H. H. Genome annotation of *Cercis chuniana* (Fabaceae). *figshare.* https://doi.org/10.6084/m9.figshare.29262773.v2 (2025).

Acknowledgements

This study was supported by grants from the National Natural Science Foundation of China (32270218, 31470312) and South China Botanical Garden, Chinese Academy of Sciences (E5R02710). We thank Dr. Xing Guo for her great support in the sequencing platform and bioinformatics analysis.

Author contributions

H.K. and W.G. conceived the project. W.L. and M.F. conducted the experiments and coordinated the research activities. M.F., C.L., Z.H., M.L. and R.Z. collected the samples. W.L., M.F. and Q.L. performed data analyses. H.K. and W.G. drafted the manuscript. H.K., W.G., and R.Z. revised and finalized the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.Z. or H.K.