scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A high-quality Chromosome-level genome assembly of *Gynostemma guangxiense* (Cucurbitaceae)

Xiao Zhang[1], Hao Zhang[2], Chen Chen[1] & Yuemei Zhao[3] ✉

Gynostemma guangxiense **X. X. Chen & D. H. Qin, belonging to the family Cucurbitaceae, is a perennial creeping herbaceous plant endemic to China with potential medicinal and health value. Here, we report the high-quality chromosome-level genome of *G. guangxiense* obtained by integrating Illumina short read, PacBio high-fidelity (HiFi) long read, Hi-C, and RNA-Seq technologies. The genome is anchored to 11 pseudochromosomes with a total size of 565.18 Mb, with a scaffold N50 of 52.63 Mb, achieving a complete BUSCO of 98.00%. Furthermore, we identified 27,527 protein-coding genes, of which 97.75% were functionally annotated. This genome provides an important molecular foundation for adaptive evolution, genetic conservation, and effective development of valuable medicinal plant resources within the *Gynostemma* genus.**

Background & Summary

Gynostemma guangxiense X. X. Chen & D. H. Qin (Cucurbitaceae) is a climbing leathery vine endemic to the limestone mountain forests in Guangxi Province, China. Its pedate leaves comprise (3-)5-7 ovate-elliptic or obovate leaflets that are nearly glabrous. Although similar in growth habit and leaf morphology to *G. pentaphyllum*, it differs in its female flowers, which form 1-2-flowered cymes, and in its trigonous-spherical fruits[1] (Fig. 1). Unlike *G. pentaphyllum*, *G. guangxiense* has a sweet taste, and its whole herb is traditionally used to treat hepatitis and chronic bronchitis[2]. *Gynostemma* is the only known genus besides *Panax* that produces dammarane-type saponins[3]. Gypenosides II, IV, VIII, and XII are structural homologs of ginsenosides Rb1, Rb2, Rd1, and Rf2, respectively[4].

As an important sister species of *G. guangxiense*, *G. pentaphyllum* is a traditional Chinese herb having potential in treating hypertension, hyperlipidemia, and inflammation[5,6], with additional health benefits including anticancer and immunomodulatory effects[7,8]. These medicinal value generate significant interest in the plants of the *Gynostemma* genus. However, the exclusive reliance on the widely distributed *G. pentaphyllum* has caused overexploitation and depletion of wild resources, leading to its classification as a national second-class protected species[9]. Comprehensive physiological, biochemical, and molecular studies across *Gynostemma* species are crucial for enhancing conservation efforts and broadening therapeutic applications. These researches will also resolve taxonomic uncertainties within this genus.

Previous metabolic studies on *G. guangxiense* have identified high leaf saponin and flavonoid content[10,11], but reports on polysaccharides are scarce. Among the essential trace elements in the human body, Fe, Mn, Zn and Cu are the most important. These four elements are more abundant in *G. guangxiense* than in *G. pentaphyllum*, which has great development and utilization value[12,13]. Current genomic studies on *G. guangxiense* primarily cover chloroplast genome characterization[14] and ISSR-PCR-based phylogenetics[15]. With the development of sequencing technologies, an increasing number of medicinal plant genomes have been sequenced and assembled[16]. These studies have laid a good molecular foundation for the identification and utilization, quality screening, biosynthesis of important components, and genetic improvements of medicinal plant species.

Here, we present a high-quality chromosome-level genome assembly of *G. guangxiense* obtained by integrating Illumina short read, PacBio high-fidelity (HiFi) long read, Hi-C, and RNA-Seq technologies. The assembled genome size was 565.18 Mb with a scaffold N50 of 52.63 Mb, achieving a complete BUSCO score of 98.00%. A

[1]Xi'an Botanical Garden of Shaanxi Province, Institute of Botany of Shaanxi Province, Xi'an, Shaanxi, 710061, China. [2]School of Pharmacy, Shaanxi University of International Trade & Commerce, Xi'an, Shaanxi, 712046, China. [3]School of Biological Sciences, Guizhou Education University, Guiyang, Guizhou, 550018, China. ✉e-mail: zhaoyuemei@gznc.edu.cn

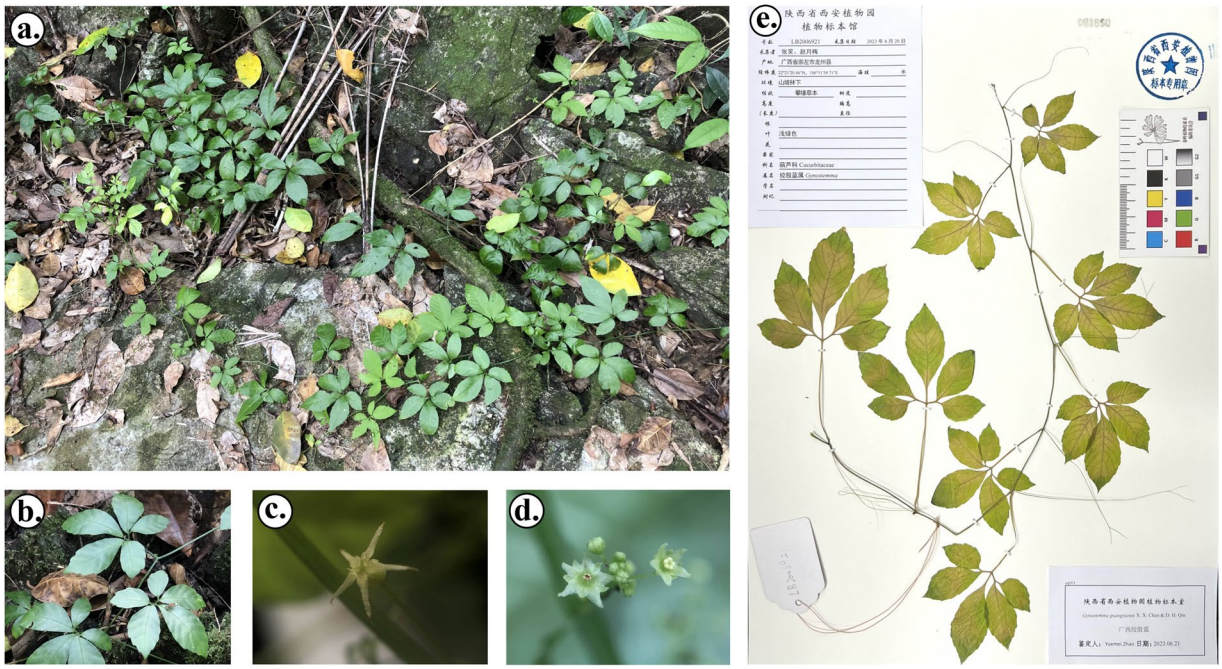

Fig. 1 Morphology and genomic features of *G. guangxiense*. (**a**) habitat; (**b**) leaves; (**c**) male flowers (panicles); (**d**) female flowers (cymose); (**e**) A scan of the herbarium specimen of *G. guangxiense* (Collector: Xiao Zhang and Yuemei Zhao; Collection site: Longzhou, Chongzuo, Guangxi, China; Environmental conditions: Under the forests on the hillside).

Read types	No. of clean bases	No. of clean reads	Length of N50 (bp)	Q30 rate (%)	GC(%)	Coverage (×)	Mapping rate (%)
Illumina short reads	30,289,621,088	220,451,660	—	96.04	34.35	59.43	99.68
HiFi reads	24,576,460,284	1,250,810	19,981	—	—	43.42	99.94
Hi-C reads	58,564,559,478	396,427,114	47,452,197	96.68	35.86	110.10	98.74
RNA reads	10,512,350,572	71,009,204	—	97.66	42.09	—	99.39

Table 1. The DNA and RNA sequencing statistics of *G. guangxiense*.

total of 562.11 Mb (99.45%) of the assembled sequences was anchored to 11 pseudochromosomes. Genome annotation predicted 27,527 protein-coding genes, and the ratio of TE repetitive sequences was 67.38%. In conclusion, this reference genome of *G. guangxiense* provides valuable information that lays an important molecular foundation for adaptive evolution, genetic conservation, and the effective development of important medicinal plant resources within the *Gynostemma* genus in the future.

Methods

Sample collection and sequencing. Healthy plant material was obtained from a female individual of *G. guangxiense* growing in Chongzuo, Guangxi, China (22.35°N, 106.86°E), and a voucher herbarium specimen was identified and deposited at the Institute of Botany of Shaanxi Province under the voucher number LB2006921 (Fig. 1e). Genomic DNA was extracted from fresh young leaves following the protocol of the DNeasy Plant Mini Kit (QIAGEN, Hilden, Germany), and randomly fragmented. A paired-end short read (150 bp) library was constructed and sequenced on the Illumina NovaSeq platform under 250 bp - 700 bp insert. Approximately, 33.64 Gb (59.43 × coverage) of short read data were obtained. For HiFi sequencing (15 - 20 kb insert size), a 20 kb library was constructed following the protocol for the PacBio Revio platform, and circular consensus sequencing (CCS) was performed (Table 1). A total of 24.58 Gb of HiFi long clean reads with an N50 of 19,981 bp were obtained for de novo assembly (Table 1). A Hi-C library was also sequenced on an Illumina NovaSeq platform with paired-end reads of 150 bp.The experiment processes were including (1) live samples were treated with 1–3% formaldehyde at room temperature for 10–30 minutes; (2) cutting the genome with Thermo Scientific HindIII restriction enzyme, and then repair the end and add biotin; (3) the interacting fragments were linked using T4 DNA ligase to form a ring; (4) using ultrasound to break the fragment again and the biotin-containing fragments were captured using magnetic beads to create libraries and sequenced. Finally, 62.32 Gb reads (110.10 × coverage) (Table 1) were generated. Total RNA was extracted from five tissues (terminal buds, mature leaves, stems, flowers and fruits) using TRNzol Universal Reagent (TIANGEN BIOTECH DP424), and uniformly mixed after leveling the concentration. The concentration of the extracted nucleic acids was determined with NanoDrop spectrophotometer (Qubit 4.0)

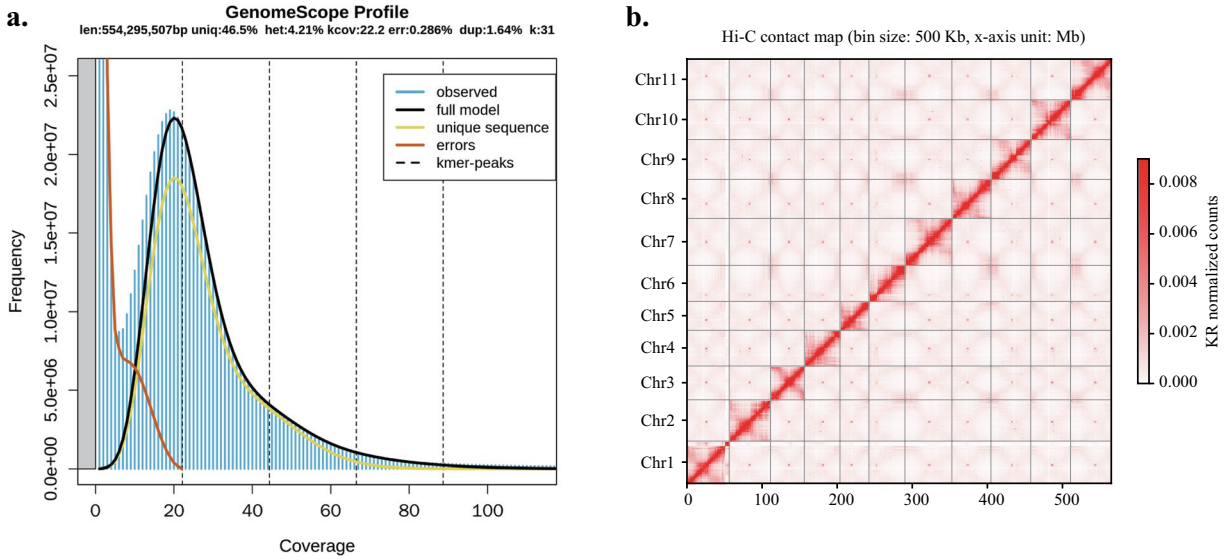

Fig. 2 Genome size estimation and Hi-C heatmap of *G. guangxiense*. (**a**) Genome size estimation by 17-K-mer analysis. (**b**) Heatmap of pseudochromosomes after Hi-C assisted assembly.

and integrity was detected with Qsep 400. The RNA-seq library was constructed and sequenced on the Illumina NovaSeq platform with 150 bp paired-end reads, generating a total of 10.68 Gb of clean data for genome annotation (Table 1).

Genome size estimation. To determine the chromosomal ploidy of the sample, flow cytometry analysis was first conducted using a Partec CyFlow Space system (Jindi Future Biotechnology, Beijing, China), and a smudge plot was employed. Illumina short reads were filtered using fastp v0.23.4[17] software to remove adapters and low-quality reads. Thereafter, a total of 30.29 GB clean reads were used for genome survey analysis. A 31-bp k-mer with quality-filtered reads was counted and satisfied using KMC v3.2.1[18] to calculate genome size, repeat content, and heterozygosity. Subsequently, the K-mer data were fitted and analyzed using the skewed normal distribution model in GenomeScope v2.0[19] software. The final genome evaluation results showed that the genome size of *G. guangxiense* was approximately 554 Mb, with a heterozygosity of 4.21% and a repeat content of 53.5%, indicating that it was a complex genome (Fig. 2a).

Genome assembly. After quality control, the PacBio HiFi reads were assembled into contigs using the Hifiasm v0.19.8[20] software with default parameters to obtain a preliminary assembly version of the genome, which was 803.27 Mb in size with a contig N50 of 44.97 Mb. The Hi-C clean reads were then aligned to the reference genome using Juicer v1.6 software[21] with default parameters. To obtain valid pair reads, noisy reads, including low-quality reads, duplicated reads, single-ended reads, and reads of three or more positions aligned on the reference genome, were filtered, automatically clustered, sorted, and oriented using the YaHS software[22] to generate Hi-C and assembly files. We then used Juicebox v1.11.08[23] to manually inspect and adjust the draft assembly, and visualized the Hi-C contact maps of the genome assembly. Finally, approximately 562.11 Mb of scaffold was anchored to 11 longest scaffolds that were identified as pseudochromosomes, featuring a contig and scaffold N50 value of 47.45 Mb and 52.63 Mb, respectively (Tables 2, 3, Fig. 2b). The complete BUSCO score for the assembly was 98.00% (Table 4). Although this complete genomic sequences was high-quality, there were still some gaps, such as at the end of Chr1 and in the middle of Chr2 and Chr4 (Table 3, Fig. 2b). These regions have high GC content and were repeat-rich regions which may represent potential centromeric and telomeric regions (Fig. 3).

Annotation of repeat sequences. Repeat sequences constitute a substantial portion of the *G. guangxiense* genome (70.59%) and can be broadly classified into tandem repeats (TRs) and transposable elements (TEs) according to their distribution in the genome. First, Tandem Repeats Finder v. 4.09[24] was used to search for TRs, and LTR elements were investigated using LTR FINDER v1.07[25]. And then, a non-redundant TE library of the *G. guangxiense* genome was constructed by extensive de novo TE annotator (EDTA) v1.0[26]. TE repeat sequences were predicted using RepeatMasker[27] with the default parameters. Finally, the results showed that the predicted proportion of TE repeat sequences was 67.38%, with long terminal repeats (LTRs) and long interspersed nuclear elements (LINEs) accounting for 44.55% and 2.05% of the genome, respectively (Table 5).

Annotation of non-coding RNAs. Based on the assembled genome sequence, software INFERNAL v1.1.4[28] trained on Rfam v14.8[29] wes used to predict the non coding RNA (ncRNA) of the genome. Meanwhile, tRNAscan-SE v2.00[30] with parameter "--thread 4 -E -I" was used to predict tRNA, and RNAmmer v1.20[31] with parameter "-S euk -m lsu,ssu,tsu -gff" was used to build models to predict rRNA and its various subunits. The final results were further integrated to obtain the prediction results of ncRNA in the *G. guangxiense* genome. In total, 696 tRNA, 4,282 rRNA and 692 ncRNA were indentified, respectively (Table 6).

Genomic feature	Value
Total genome size (bp)	565,181,624
pseudochromosomes number	11
pseudochromosomes length (bp)	562,106,698
Contig number	80
Contig N50 (bp)	47,452,197
Contig N90 (bp)	18,294,638
Scaffold number	68
Scaffold N50 (bp)	52,626,209
Scaffold N90 (bp)	44,970,829
Largest scaffold length (bp)	62,093,050
HIC Anchor ratio (%)	99.45
Protein-coding gene number	27,527

Table 2. Statistics of the *G. guangxiense* genome assembly and annotation.

ID	Length (bp)	No. of gaps
Chr1	56,234,087	4
Chr2	54,742,393	2
Chr3	44,970,829	0
Chr4	47,118,517	4
Chr5	38,281,868	1
Chr6	47,829,778	1
Chr7	62,093,050	0
Chr8	51,889,994	0
Chr9	52,699,468	0
Chr10	52,626,209	0
Chr11	53,620,505	1

Table 3. Summary of the eleven pseudochromosomes.

Type	Assembly		Annotation	
	Number	Percentage (%)	Number	Percentage (%)
Complete BUSCO (C)	1,582	98.00	1,559	96.59
Complete and single-copy BUSCO (S)	1,292	80.00	1,523	94.36
Complete and duplicated BUSCO (D)	290	18.00	36	2.23
Fragmented BUSCO (F)	23	1.40	11	0.68
Missing BUSCO (M)	9	0.60	44	2.73
Total BUSCO groups searched	1,614	100.00	1,614	100.00

Table 4. The BUSCO results of assembly and annotation.

Gene prediction and functional annotation. The presence of repeat sequences often makes it difficult to predict and annotate encoding genes. In this study, we shielded the TE sequences before predicting the encoding genes. Protein-coding gene annotation was performed using a combination of ab initio-, homology-, and transcriptome-based prediction methods. For ab initio prediction, Augustus v3.5.0[32] and GeneMark-ES v5.1[33] were employed to predict protein-coding genes. Second, the protein files of five species (*G. pentaphyllum*, *Momordica charantia*, *Cucumis sativus*, *Nicotiana tabacum* and *Arabidopsis thaliana*) were downloaded from the National Center for Biotechnology Information (NCBI) and aligned to the query genome as homologous proteins using GeMoMa v1.90[34]. Third, quality-controlled RNA-Seq data from the five tissues were mapped onto assembled genomes using HISAT2 v.2.2.1[35]. Based on the alignment results, the mapped reads were assembled using StringTie v1.3.3b[36], and PASA v2.5.2[37] was used to predict the UTR and variable splicing regions of the initially obtained gene set. To obtain the final non-redundant gene set, Evidence Modeler v1.1.1[38] was used to combine the three gene datasets with weights of 1, 1, 5, and 10 for GeneMark, Augustus, GeMoMa, and PASA, respectively. We also assessed the completeness of the gene set based on comparisons with the plant single-copy ortholog gene database (Embryophyta_odb10) using BUSCO v4.0.6[39]. The results showed that 96.59% of the Embryophyta odb10 gene set was completely covered by genome annotations (Table 4). The gene sets obtained from gene structure annotations were searched against six known protein databases including the NCBI non-redundant protein (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), Eukaryotic Orthologous

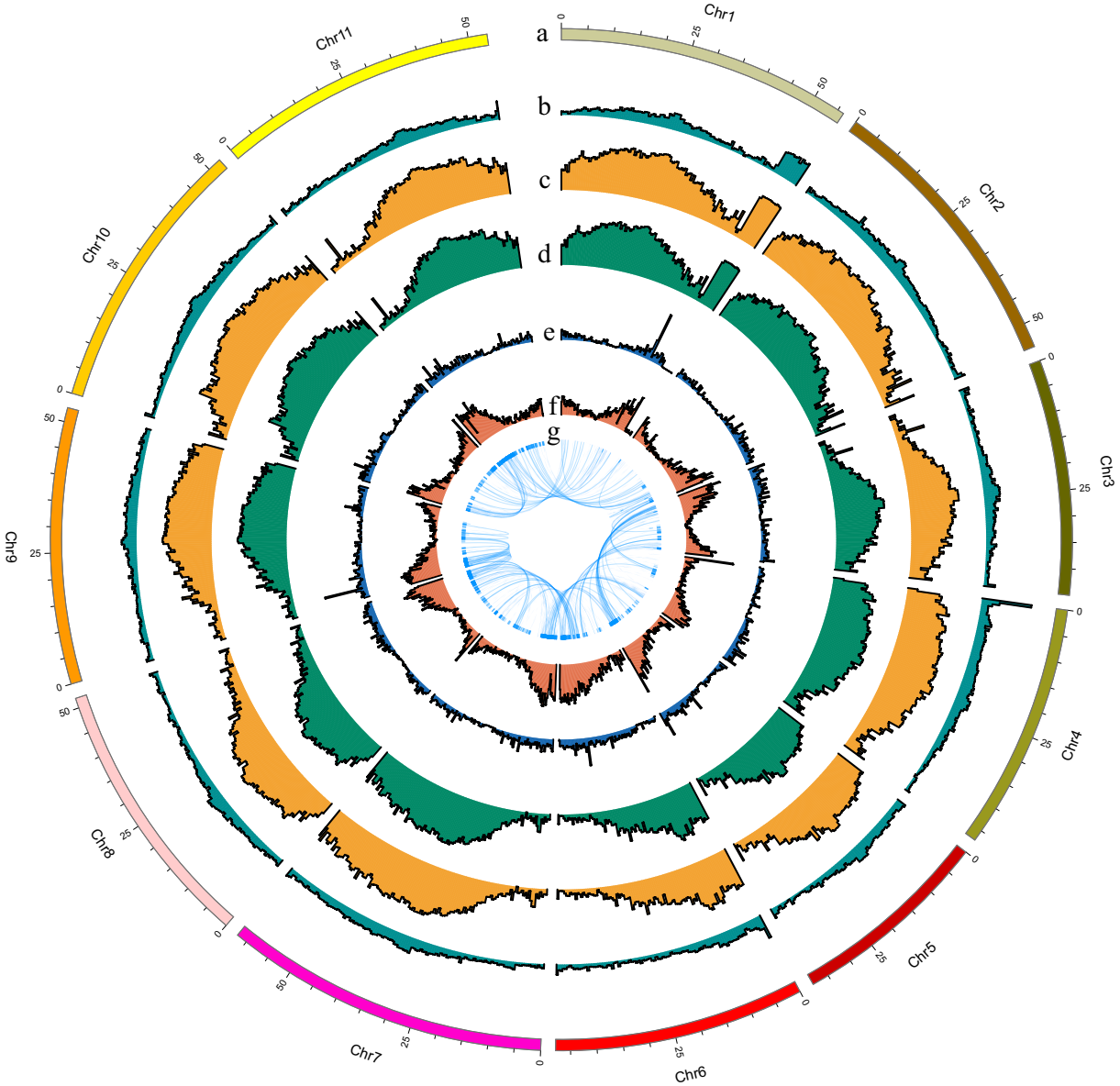

Fig. 3 Genome map of *G. guangxiense*. (a) The 11 pseudochromosomes. (b) GC content. (c) Total repeat sequences density. (d) TE density. (e) TR density. (f) Gene density. (g) Intragenomic collinearity.

Groups (KOG), InterPro, and Swiss-Prot using BLAST v2.2.31[40] with an e-value less than 1e-5. In total, 27,527 protein-coding genes were predicted, of which 97.75% were functionally annotated (Tables 2, 7).

Data Records

The clean Illumina reads, PacBio HiFi reads, Hi-C reads and RNA-seq reads for the *G. guangxiense* genome has been deposited in the NCBI Sequence Read Archive (SRA) database under accession number SRR32030229[41], SRR32030228[42], SRR32030227[43], and SRR32030226[44], respectively, under BioProject accession number PRJNA1209686[45]. The final chromosome assembly is available in the National Center for Biotechnology Information (NCBI) GenBank under the accession No. JBLAST000000000[46]. The genome and its repeat annotation for *G. guangxiense* have been uploaded to figshare under accession number 28283174[47] and 30086362[48]. Details and summary of functional annotation of *G. guangxiense* has been uploaded to figshare under accession number 30204301[49].

Technical Validation

The completeness of the assembly and annotation was assessed based on the Embryophyta odb10 database using BUSCO v4.0.6, with default parameters. The assembly completeness of BUSCO was 98.00% (n = 1,614), and the annotated proteins completeness of BUSCO was 96.59% (n = 1,614) (Table 4). Clean Illumina, HiFi and RNA-Seq reads were mapped onto the assembled genome using BWA v0.7.15[50], minimap2 v2.28[51], and

Class	Order	Super family	Number of elements	Length of sequence (bp)	Percentage of sequence (%)
Class I			400,323	263,342,926	46.59
	LTR		363,672	251,773,205	44.55
		Gypsy	175,983	163,724,046	28.97
		Copia	93,454	47,383,137	8.38
		Unknown	93,985	40,641,224	7.19
		Other	250	24,798	0
	LINE		36,650	11,569,681	2.05
		L1	36,576	11,563,975	2.05
		Other	74	5,706	0
	SINE		1	40	0
		Other	1	40	0
Class II			457,921	117,502,812	20.79
	DNA		390,126	85,056,003	15.05
		Helitron	228,045	39,968,025	7.07
		DTM	69,072	16,012,087	2.83
		MULE-MuDR	6,612	2,381,050	0.42
		DTC	35,463	13,690,314	2.42
		DTA	28,128	7,708,415	1.36
		DTH	13,785	2,991,356	0.53
		CMC-EnSpm	5,517	1,623,425	0.29
		Other	3,504	681,331	0.12
	MITE		66,482	32,140,153	5.69
		DTA	13,435	2,408,814	0.43
		DTM	47,664	29,206,667	5.17
		Other	5,383	524,672	0.09
	RC		1,313	306,656	0.05
		Other	1,313	306,656	0.05
Total TEs			858,244	380,845,738	67.38
Tandem Repeats			16,071	1,517,360	0.27
Low complexity			29,337	1,444,478	0.26
Simple repeats			131,760	5,211,869	0.92
Unknown			58,993	9,958,107	1.76
Other			124	9,431	0
Total Repeats			1,094,529	398,986,983	70.59

Table 5. Statistics of repeat sequences in the *G. guangxiense* genome.

Type		Number	Average_length (bp)	Total_length (bp)	Percentage (%)
tRNA		696	75.01	52,207	0.0092
regulatory		8	56.25	450	0.0001
rRNA		4,282	510.58	2,186,320	0.39
	18S	15	952.93	14,294	0.0025
	28S	150	3,363.76	504,564	0.089
	5.8S	196	153.38	30,062	0.0053
	5S	3,409	118.12	402,678	0.071
ncRNA		692	115.18	79,703	0.0141
	snRNA	494	105.32	52,030	0.0092
	miRNA	83	127.43	10,577	0.0019
	spliceosomal	87	142.23	12,374	0.0022
	other	28	168.64	4,722	0.0008

Table 6. Statistics of non coding RNA in the *G. guangxiense* genome.

HISAT2[35] to assess the genomic integrity and accuracy. The mapping reads were 99.68%, 99.94%, and 99.39%, respectively. These results indicated that the *G. guangxiense* genome assembly was of high quality.

Database	Gene Number	Annotation Ratio (%)
Total genes	27,527	—
KOG (EuKaryotic Orthologous Groups)	13,345	48.48
KEGG (Kyoto Encyclopedia of Genes and Genomes)	8,856	32.17
NR (non-redundant)	26,061	94.67
SwissProt (Swiss Institute of Bioinformatics and Protein Information Resource)	19,004	69.04
Interpro	26,194	95.16
GO (Gene ontology)	11,652	42.33
Annotated genes	26,907	97.75

Table 7. Statistical summary of the annotation of the *G. guangxiense* genome using six databases (KOG, KEGG, Nr, Swissprot, InterPro, and GO).

Data availability

The final chromosome assembly of *G. guangxiense* is available in the GenBank under the accession No. JBLAST000000000. The clean sequencing reads generated from three platform-specific sequencing runs, along with the final genome assembly, have been deposited in the NCBI Sequence Read Archive (SRA, accession numbers: SRR32030226 - SRR32030229).

Code availability

The software utilized in this study were executed in strict adherence to the official guidelines of published bioinformatics programs. Anything not mentioned in Methods was run with default settings. No custom codes were used.

Received: 19 June 2025; Accepted: 12 February 2026;
Published online: 23 February 2026

References

1. Chen S., Lu A., & Charles J. Flora of China. Vol. 19. Beijing: Missouri Botanical Garden Press. (2011).
2. Chen, X. X. & Qin, D. H. A new species of the genus *Gynostemma* from Guangxi. *Acta Botanica Yunnanica.* **10**(4), 495–496 (1988).
3. Chen, Q. *et al*. Transcriptome sequencing of *Gynostemma pentaphyllum* to identify genes and enzymes involved in triterpenoid biosynthesis. *International Journal of Genomics.* **2016**, 1–10 (2016).
4. Kao, T., Huang, S., Inbaraj, B. & Chen, B. Determination of flavonoids and saponins in *Gynostemma pentaphyllum* (Thunb.) Makino by liquid chromatography-mass spectrometry. *Analytica Chimica Acta.* **626**, 200–211 (2008).
5. Gou, S. H. *et al*. Anti-atherosclerotic effect of *Fermentum Rubrum* and *Gynostemma pentaphyllum* mixture in high-fat emulsion-and vitamin D3-induced atherosclerotic rats. *Journal of the Chinese Medical Association.* **81**, 398–408 (2018).
6. Babich, O. *et al*. Medicinal plants to strengthen immunity during a pandemic. *Pharmaceuticals.* **13**, 313–314 (2020).
7. Li, Y. *et al*. Anti-cancer effects of *Gynostemma pentaphyllum* (thunb.) makino (jiaogulan). *Chinese Medicine.* **11**, 43–45 (2016).
8. Choi, K. T. Botanical characteristics, pharmacological effects and medicinal components of korean *Panax ginseng* C. a Meyer. *Acta Pharmacologica Sinica.* **29**(11), 09–18 (2008).
9. Li, Z. H. *et al*. A review on studies of systematic evolution of *Gynostemma* Bl. *Acta Botanica Boreali-Occiden-talia Sinica.* **32**, 2133–2138 (2012).
10. Liu, S., Lin, R. & Hu, Z. Comparison of stem and leaf structures and total gypenosides among 5 species of *Gynostemma*. *Journal of Fujian Agriculture and Forestry University (Natural Science Edition).* **35**(5), 495–499 (2006).
11. Jiang, W., Zhou, Y. & Li, J. Assaying of total flavonoids in 6 kinds of *Gynostemma* made in Guangxi. *Chinese Pharmacy.* **7**(1), 74–75 (2006).
12. Li, X., Liu, S., Yi, L. & Li, C. Seasonal variations in the contents of total saponins, total flavonoids and mineral elements of three species of the genus *Gynostemma*. *Journal of Chinese Medicinal Materials.* **35**(1), 26–30 (2012).
13. Peng, X., Wang, T., Luo, Z. & Liu, S. Nutritional components of three species of *Gynostemma* (Cucurbitales: Cucurbitaceae). *Journal of Mountain Agriculture and Biology.* **36**(1), 89–91 (2017).
14. Zhao, Y., Zhang, X., Zhou, T., Chen, X. & Ding, B. Complete chloroplast genome sequence of *Gynostemma guangxiense*: genome structure, codon usage bias, and phylogenetic relationships in Gynostemma (Cucurbitaceae). *Brazilian Journal of Botany.* **46**, 351–365 (2023).
15. Wang, C. *et al*. Identification of seven plants of *Gynostemma* BL. by ISSR-pcr. *Chinese Traditional and Herbal Drugs.* **39**(4), 588–591 (2008).
16. Zhang, X. *et al*. Diploid chromosome-level reference genome and population genomic analyses provide insights into Gypenoside biosynthesis and demographic evolution of *Gynostemma pentaphyllum* (Cucurbitaceae). *Horticulture Research.* **10**(1), uhac231 (2023).
17. Chen, S. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* **34**, i884–i890 (2018).
18. Deorowicz, S., Kokot, M., Grabowski, S. & Debudaj-Grabysz, A. KMC 2: fast and resource-frugal k-mer counting. *Bioinformatics.* **31**(10), 15 (2015).
19. Vurture, G. W. *et al*. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics.* **33**(14), 2202–2204 (2017).
20. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods.* **18**, 170–175 (2021).
21. Durand, N. C. *et al*. Juicer provides a one-click system for Analyzing loop-resolution hi-C experiments. *Cell Systems.* **3**, 95–98 (2016).
22. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics.* **39**, btac808 (2023).
23. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems.* **3**, 99–101 (2016).
24. Zhao, X. & Hao, W. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research.* **35**, 265–268 (2007).
25. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research.* **27**, 573–580 (1999).

26. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome biology.* **20**, 275–259 (2019).
27. Jurka, J. *et al*. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research.* **110**, 462–467 (2005).
28. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* **29**(22), 2933–5 (2013).
29. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research.* **33**(Database issue), D121–4 (2005).
30. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research.* **25**(5), 955–64 (1997).
31. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research.* **35**(9), 3100–8 (2007).
32. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research.* **34**, 435–439 (2006).
33. Alexandre, L., Burns, P. D. & Mark, B. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research.* **42**, e119 (2014).
34. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods in Molecular Biology.* **1962**, 161–177 (2019).
35. Kim, D., Langmead, B. & Salzberg, S. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods.* **12**, 357–360 (2015).
36. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA - seq reads. *Nature Biotechnology.* **33**, 290–295 (2015).
37. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research.* **31**, 5654–5666 (2003).
38. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology.* **9**, R7 (2008).
39. Simão, F. A. *et al*. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* **31**, 3210–3212 (2015).
40. Altschul, S. *et al*. Basic local alignment search tool. *Journal of Molecular Biology.* **215**, 403–410 (1990).
41. *NCBI Sequence Read Archive*,https://identifiers.org/ncbi/insdc.sra:SRR32030229 (2025).
42. *NCBI Sequence Read Archive*,https://identifiers.org/ncbi/insdc.sra:SRR32030228 (2025).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32030227 (2025).
44. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32030226 (2025).
45. *NCBI National Genomics Data Center (NGDC) database* https://identifiers.org/ncbi/bioproject:PRJNA1209686 (2025).
46. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBLAST000000000 (2025).
47. Zhang, X. The genome annotation of *Gynostemma guangxiense*. *figshare.* https://doi.org/10.6084/m9.figshare.28283174.v1 (2025).
48. Zhang, X. The repeat annotation of *Gynostemma guangxiense*. *figshare.* https://doi.org/10.6084/m9.figshare.30086362.v1 (2025).
49. Zhang, X. Details and summary of functional annotation of *Gynostemma guangxiense*. *figshare.* https://doi.org/10.6084/m9.figshare.30204301 (2025).
50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* **25**, 1754–1760 (2009).
51. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* **34**, 3094–3100 (2018).

Acknowledgements

We would like to express our gratitude to Kindstar Sequenon Co., Ltd for its support of the sequencing technology. This study was financially supported by the Project of the Science and Technology Program of Shaanxi Academy of Science (No. 2024k-32), National Natural Science Foundation of China (No. 32000256, 31900273), and the Innovation Capability Support Program of Shaanxi (No. 2025ZC-KJXX-119).

Author contributions

X.Z. and Y.Z. conceived and designed the study. X.Z. and Y.Z. prepared the samples. X.Z., H.Z., and C.C. analyzed the data. X.Z. and H.Z. wrote the manuscript. C.C. and Y.Z. revised 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 Y.Z.