



OPEN

DATA DESCRIPTOR

Haplotype-resolved genome assembly of *Coriaria nepalensis* a non-legume nitrogen-fixing shrub

Shi-Wei Zhao^{1,7}, Jing-Fang Guo^{1,7}, Lei Kong^{1,7}, Shuai Nie¹, Xue-Mei Yan¹, Tian-Le Shi¹, Xue-Chan Tian¹, Hai-Yao Ma¹, Yu-Tao Bao¹, Zhi-Chao Li¹, Zhao-Yang Chen¹, Ren-Gang Zhang², Yong-Peng Ma², Yousry A. El-Kassaby³, Ilga Porth⁴, Wei Zhao⁵✉ & Jian-Feng Mao^{1,6}✉

Coriaria nepalensis Wall. (Coriariaceae) is a nitrogen-fixing shrub which forms root nodules with the actinomycete *Frankia*. Oils and extracts of *C. nepalensis* have been reported to be bacteriostatic and insecticidal, and *C. nepalensis* bark provides a valuable tannin resource. Here, by combining PacBio HiFi sequencing and Hi-C scaffolding techniques, we generated a haplotype-resolved chromosome-scale genome assembly for *C. nepalensis*. This genome assembly is approximately 620 Mb in size with a contig N50 of 11 Mb, with 99.9% of the total assembled sequences anchored to 40 pseudochromosomes. We predicted 60,862 protein-coding genes of which 99.5% were annotated from databases. We further identified 939 tRNAs, 7,297 rRNAs, and 982 ncRNAs. The chromosome-scale genome of *C. nepalensis* is expected to be a significant resource for understanding the genetic basis of root nodulation with *Frankia*, toxicity, and tannin biosynthesis.

Background & Summary

Coriaria nepalensis Wall. ($2n = 40$)¹, also known as Masuri Berry, is a shrub belonging to the genus *Coriaria* of the unigeneric Coriariaceae family, and is mainly distributed in the Himalayan region. *C. nepalensis* is a non-legume nitrogen-fixing plant that forms root nodules with the actinomycete *Frankia*^{2,3}. The biological ability of nitrogen-fixation in this species contributes to its rehabilitation capacity of nutrient-poor degraded land^{4,5}; in combination with its osmotic adjustment function and drought tolerance^{6,7}, *C. nepalensis* improves the abiotic conditions and provides more suitable habitat for associated plant species^{8–10}. Furthermore, essential oils and extracts from *C. nepalensis* could be used as promising drugs due to their antimicrobial^{11,12} and anti-convulsant activities¹³. Traditionally, *C. nepalensis* has been used in folk medicine to treat ailments such as toothaches and traumatic injuries^{13,14}. The toxic and antibacterial properties of *C. nepalensis* provide an interesting opportunity for the development of a potent new and environmentally friendly pesticide for pest management¹⁵. Moreover, *C. nepalensis* bark offers an important source of hydrolysable tannin^{16,17}, an ideal treatment for tanning leather¹⁶.

The phylogenetic position of Coriariaceae is still debated¹⁸. Previous analyses based on plastid *rbcL* gene sequences^{19–21}, and the complete chloroplast genome¹⁴ placed Coriariaceae close to families in Cucurbitales. However, the nuclear genome has not yet been sequenced in Coriariaceae, although the genome assemblies of related taxa, such as in Datisceae²² and Begoniaceae²³, have been published.

Molecular genetic investigation of non-legume nitrogen-fixation and root nodulation from *Frankia* requires a high-quality genome assembly and functional annotation of the host plant. Additionally, such genomic

¹National Engineering Research Center of Tree Breeding and Ecological Restoration, Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, 100083, China. ²Yunnan Key Laboratory for Integrative Conservation of Plant Species with Extremely Small Populations, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China. ³Department of Forest and Conservation Sciences, Faculty of Forestry, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada. ⁴Département des Sciences du Bois et de la Forêt, Faculté de Foresterie, de Géographie et Géomatique, Université Laval, Québec, QC, G1V 0A6, Canada. ⁵Department of Ecology and Environmental Science, Umeå Plant Science Centre, Umeå University, Umeå, SE-901 87, Sweden. ⁶Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, Umeå, SE-901 87, Sweden. ⁷These authors contributed equally: Shi-Wei Zhao, Jing-Fang Guo, Lei Kong. ✉e-mail: zhao.wei@umu.se; jianfeng.mao@umu.se

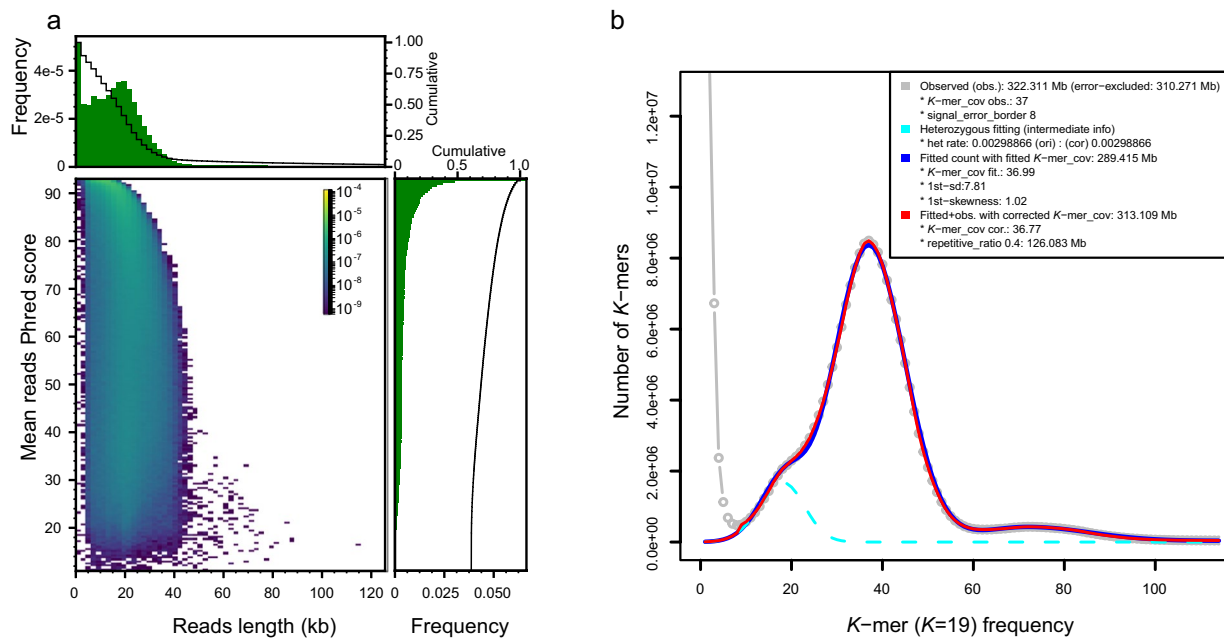


Fig. 1 Length and quality of PacBio HiFi reads and genome size survey. **(a)** Reads length and mean Phred score distribution of PacBio HiFi reads. **(b)** 19-mers frequency distribution estimated from PacBio HiFi sequences: observed K -mer (raw K -mer) frequencies (in grey), fitted K -mer frequencies (in blue) with skew normal distribution model, and overall fitting (in red) that concatenated observed and fitted K -mer frequencies.

resources may also be crucial to advance the phylogenetics of the unigeneric Coriariaceae family and the efficient exploration of *C. nepalensis*' valued natural products.

Here, we report a 620 Mb haplotype-resolved chromosome-scale assembly of *C. nepalensis* using a combination of high-quality PacBio HiFi (High Fidelity) long reads, Illumina reads, and Hi-C sequencing. The genome was assembled with contig N50 length of 11 Mb and 40 haplotype-resolved pseudochromosomes. We predicted 60,862 protein-coding genes, of which 99.5% were functionally annotated. Furthermore, 939 tRNAs, 7,297 rRNAs, and 982 ncRNAs were annotated. The provided genomic resources will be helpful for future functional studies in *C. nepalensis*.

Methods

Sample collection, library construction, and genome size estimation. Leave tissue samples for both genome and RNA sequencing were harvested in 2020 from a mature *C. nepalensis* individual growing in Kunming Botanical Garden which was transplanted from Songming county, Kunming, Yunnan province, China. Sampled leaves were immediately flash-frozen in liquid nitrogen and stored at -80°C until further use. High-quality genomic DNA was extracted from leaf tissue using the DNeasy Plant Mini Kit (QIAGEN, Inc.) and purified using the Mobio PowerClean Pro DNA Clean-Up Kit (MO BIO Laboratories, Inc.). DNA integrity was assessed using Agilent 4200 Bioanalyzer. Messenger RNA (mRNA), whose sequence information was later utilized in protein-coding gene structure prediction, was isolated from leaves using the NEBNext Poly(A) mRNA Magnetic Isolation Module, and RNA quality was determined with the Agilent 2100 BioAnalyzer.

We combined PacBio HiFi long reads sequencing, Illumina sequencing, and Hi-C scaffolding for *C. nepalensis* genome assembly. Genomic DNA fragments were prepared using g-Tubes and purified using AMPure PB beads for library construction and subsequent SMRT cell PacBio HiFi long reads sequencing. Fragment molecules were screened on BluePippin system. The library sequencing was performed on PacBio Sequel II platform, and ccs (<https://github.com/PacificBiosciences/ccs>) v6.2.0 was used to generate PacBio HiFi data. We obtained ~ 14.5 Gb ($\sim 40\times$) of HiFi sequencing data with an average length of 19 kb and N50 of 21 kb (Fig. 1a). As for Illumina sequencing, 150 bp paired-end PCR-free libraries were prepared and sequenced on Illumina HiSeq X Ten platform, and ~ 70 Gb ($\sim 200\times$) of Illumina raw data were obtained. We followed a standard procedure for Hi-C library preparation²⁴. In brief, leaf tissues were fixed with formaldehyde and the cross-linked DNA was digested with MboI restriction enzyme. Digested fragments were then biotinylated at 5' overhangs and joined to form chimeric junctions. After biotin-containing fragments were enriched and sheared, we constructed paired-end sequencing libraries. The Hi-C libraries were sequenced using the Illumina HiSeq X Ten platform and ~ 67 Gb of Hi-C raw data were obtained. RNA sequencing was performed on Illumina HiSeq X Ten platform after we constructed one sequencing library using the NEBNext Ultra RNA Library Prep Kit, and ~ 7.5 Gb (50 Mb reads) of raw data were acquired. Then, fastp²⁵ software was used for quality control to remove adapters and low-quality and too short Illumina reads (<60 bp). All clean reads were used for further genome assembly and gene predictions.

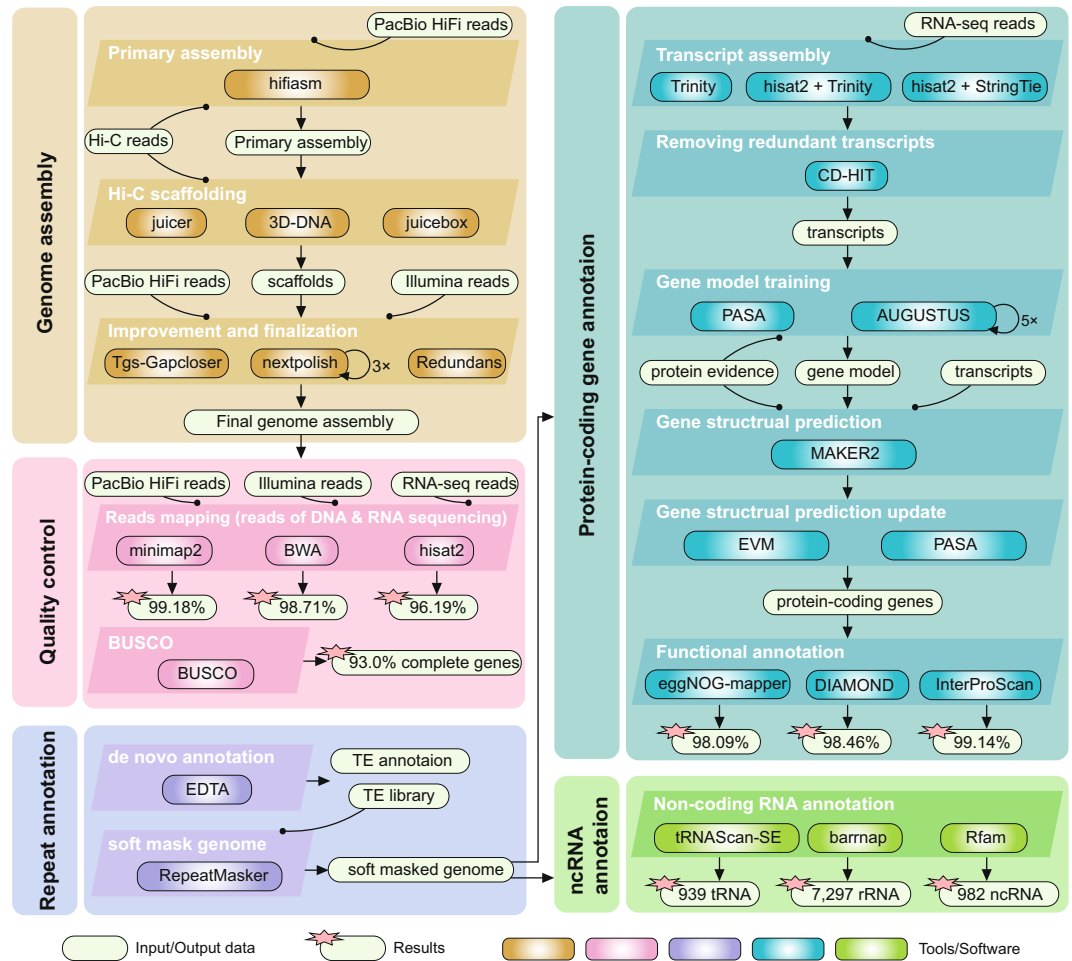


Fig. 2 Pipeline overview of genome assembly (yellow), quality control (pink), repeat annotation (purple), protein-coding gene annotation (blue), and ncRNA annotation. Boxes with different color shading represent the different software used in each analytical step.

Genomic characteristics including genome size, repeat content, and heterozygous rate were estimated based on K -mer frequencies. Through K -mer analysis ($K = 19$) of PacBio HiFi data with Jellyfish²⁶ v2.3.0, an overall *C. nepalensis* haplotype genome size of 313.1 Mb was estimated using findGSE v1.94.R²⁷ (Fig. 1b).

De novo genome assembly. *De novo* assembly involved three steps: primary assembly, Hi-C scaffolding, and polishing (Fig. 2). With PacBio HiFi reads and Hi-C reads as inputs, we used hifiasm²⁸ v0.16.1 to assemble the genome into contigs and obtained a haplotype-resolved assembly with two haplotypes for subsequent analysis. Further, the Hi-C reads that were mapped to the assembly using Juicer²⁹ v1.6. 3D-DNA³⁰ (-m haploid -i 150000 -r 0 -editor-repeat-coverage 5) were then used for preliminary Hi-C assisted chromosome assembly, and Juicebox³¹ (version 201008) was used to manually adjust the chromosome segmentation boundary and any wrong assembly, including switch error. Afterwards, we used 3D-DNA to re-scaffold each chromosome separately and used Juicebox to manually correct any visible error. We used TGS-GapCloser³² v1.0.1 (--min_match 1000 --minmap_arg '-x asm20') to fill the gaps (24 gaps were filled) with HiFi reads and performed three rounds of polishing using NextPolish³³ v1.4.0 based on Illumina reads, and removed redundant sequences identified by Redundans³⁴ v0.13c. Finally, a haplotype-resolved chromosomal level assembly with a total length of 620 Mb was obtained (Table 1). We obtained 40 pseudochromosomes, consistent with the chromosome number reported in a previous karyotype study¹. We named the chromosomes according to the descendent order of their lengths. Furthermore, as we were describing a haplotype-resolved genome assembly without parental information for subgenome phasing, we arbitrarily denoted the longer one from each pair of homologous chromosomes as haplotype genome “a” (with character “a” in the terminal of the chromosome IDs), while the other chromosome as haplotype genome “b” (with character “b”).

Chromosomes chr01–chr03 assemblies were significantly longer than the remaining chromosomes. The assembly of these three pairs of chromosomes was also difficult, showing Hi-C chromatin contact profiles distinct from others (Fig. 3a,b). These three pairs of chromosomes have a large number of gaps (in total 60) in the current assembly, while the other chromosomes had a total of only 2 gaps. Previous karyotype analysis¹ showed that *C. nepalensis* had three pairs of long chromosomes with extended heterochromatin regions, which

Features	Statistics
Sequencing	
Raw bases of WGS-PacBio HiFi (Gb)	~14.5
Raw bases of WGS-Illumina (Gb)	~70
Raw bases of Hi-C (Gb)	~67
Raw bases of RNA-seq (Gb)	~7.5
Assembly	
Genome size (Mb)	620.52
Number of pseudochromosomes	40
Chloroplast genome assembly (bp)	158,558
Mitochondria genome assembly (bp)	480,951
N50 of contigs (Mb)	10.97
L50 of contig	22
N50 of scaffolds (Mb)	12.9
L50 of scaffolds	11
Number of gaps	62
GC content (%)	34.78
Complete BUSCOs	1,338 (93.0%)
Annotation	
Number of protein-coding gene	60,862
Complete BUSCOs	1,440 (97.2%)
Average length of protein-coding gene (bp)	2,892.7
Average length of CDS (bp)	1,324
Average number of exons per transcript	6.3
Number of tRNA	939
Number of rRNA	7,297
Number of unclassified ncRNA	982

Table 1. Statistics of the haplotype-resolved genome assembly of *C. nepalensis*.

is concurrent with the three long chromosomes revealed in the present study. A high number (679,177) of tandem array repeats with the consensus sequence “ATCATTGCAAGTTATGCACAAAAGTTGTGTCTGTAGTGCAAACACTAGAAATTCGTTGACTTGCTTTGAAATAAGTTATTGACTTGAAATGACTCATTGAAATGATTTAAAGGTTAAACGAATGCACACTTTCCTTGCAATG” was identified on the three long chromosomes chr01–chr03 (Fig. 3c) using TRF³⁵ v4.09.1. We found the “TTTAGGG” characterized telomeric sequence in most chromosomes (Fig. 3c), indicating the high quality of our genome assembly.

In addition, a 158,558 bp chloroplast (Pt) genome and a 480,951 bp mitochondrial (Mt) genome were assembled based on short- and long- reads gained from genome sequencing using GetOrganelle³⁶ v1.7.5.0 (Table 1).

Repeat annotation. We performed *de novo* transposable element (TE) annotation using EDTA³⁷ v1.9.3 (--sensitive 1 --anno 1) which integrates homology-based and structure-based approaches for TE identification (Fig. 2). A TE library was generated and used for further repeat annotation with RepeatMasker (<http://www.repeatmasker.org/RepeatMasker/>) (-no_is -xsmall). The output repeat soft-masked genome sequence file was used for gene prediction. A total of 428 Mb (69.0%) of the assembly was annotated as TE (Table 2), of which 61 Mb (9.9%) were long terminal repeat (LTR) retrotransposons. Mutator transposons with 280 Mb (45.2%) in total length showed the highest genome occupation, and also a distribution similar to the high occupation tandem array mentioned above (Fig. 3c). Our further analysis revealed that the sequence motif of these tandem arrays is included inside the Mutator transposons.

Protein-coding genes prediction and other annotations. We collected 139,950 non-redundant protein sequences of the closely related species *Datisca glomerata*²², *Begonia fuchsoides*²², *Cucumis sativus*³⁸, *Vitis vinifera*³⁹, *Prunus persica*⁴⁰, and *Arabidopsis thaliana*⁴¹ as evidence for protein homology (Fig. 2). Three strategies were used to assemble RNA-seq reads into transcripts which were further used as transcriptional evidence for gene annotation. For transcripts assembly, (1) *de novo* assembly was performed using Trinity⁴² v2.13.2; (2) genome-guided assembly was performed using Trinity after reads were mapped to the genome assembly using HISAT2⁴³ v2.2.1; and (3) another genome-guided assembly was prepared using StringTie⁴⁴ v2.2.0 with reads mapping using HISAT2. We combined all these three sets of transcripts and obtained 77,555 transcript sequences after removing the redundant sequences with CD-HIT⁴⁵ v4.8.1. Gene structure was annotated using the PASA⁴⁶ v2.5.0 pipeline based on transcriptional evidence. Then, full-length gene sequences were identified by evidence of protein homologies. Based on the full-length gene set, a gene model used for *ab initio* gene structure prediction was trained and optimized using AUGUSTUS⁴⁷ 3.4.0.

Furthermore, the MAKER2⁴⁸ pipeline was used to predict the putative protein-coding gene structure. We performed *ab initio* predictions of gene structures using AUGUSTUS 3.4.0. The transcript evidence and

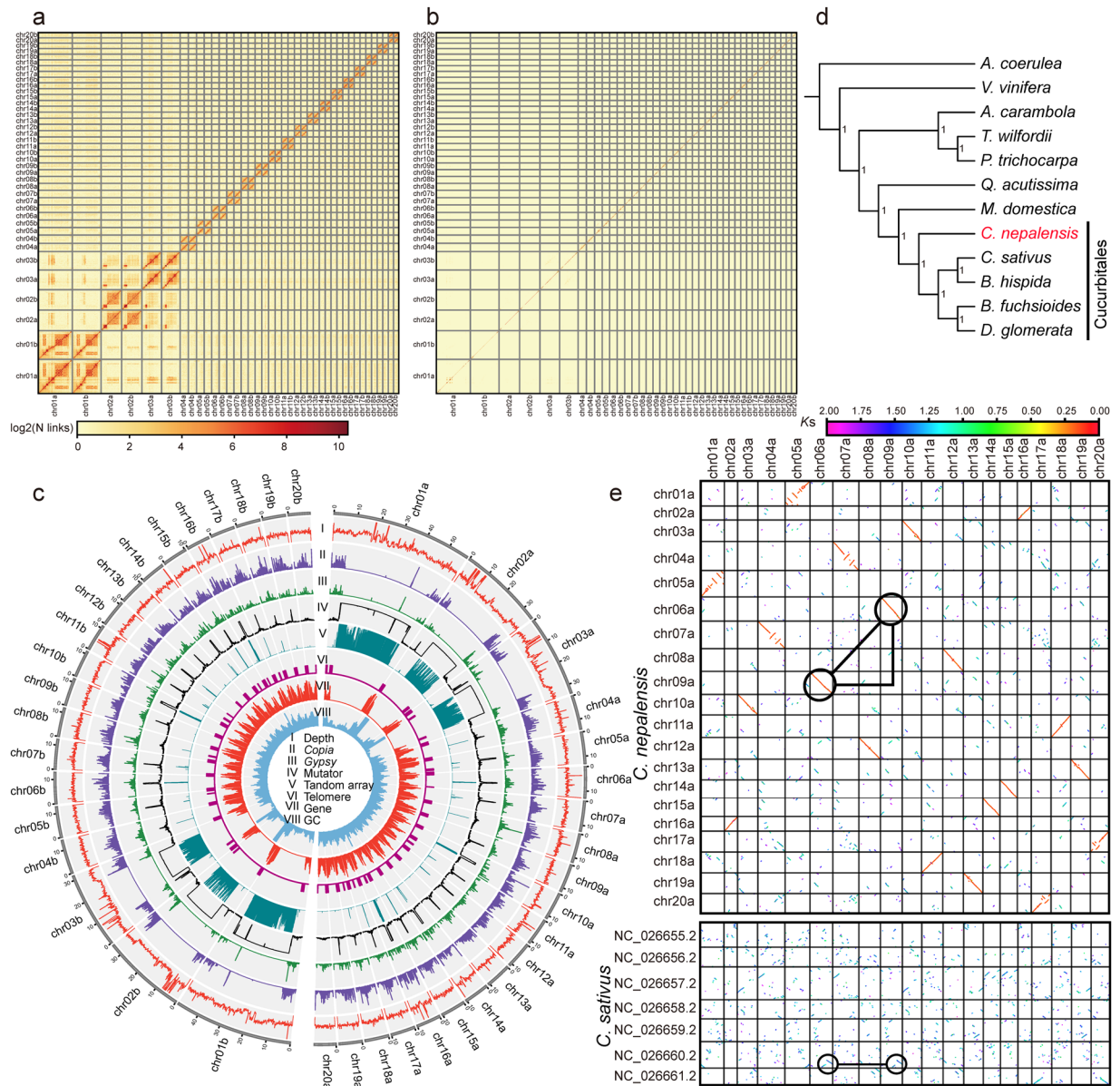


Fig. 3 Hi-C density heatmaps, genomic features and evolutionary history of *C. nepalensis*. **(a)** Hi-C chromatin contact density heatmap with a low threshold parameter (minimal mapping quality = 0). **(b)** Hi-C chromatin contact density heatmap with a high threshold parameter (minimal mapping quality = 1). **(c)** Distribution of genomic features of *C. nepalensis*. **I**: sequencing depth distribution of PacBio HiFi reads. **II–IV**: The density of *Copia* LTR-RTs, *Gypsy* LTR-TRs and Mutator TE. **V, VI**: Distribution of tandem array and telomere sequence. **VII, VIII**: Density of protein-coding gene and GC content. **(d)** Phylogenetic tree. **(e)** *Ks* dot plots of *C. nepalensis* haplotype genome “a” and *C. sativus*.

homologous protein evidence were aligned with the genome by BLAST+⁴⁹ v2.11.0 and optimized by exon-erate⁵⁰ 2.4.0. AUGUSTUS was used to integrate gene models from the above-mentioned gene prediction. To further improve the annotation accuracy, EvidenceModeler⁵¹ (EVM) v1.1.1 and PASA were used to integrate and update the gene prediction results. We annotated a final set of 60,862 protein-coding genes (Table 1), among which 30,622 genes were predicted for the haplotype subgenome with a longer set of chromosomes (haplotype genome “a”), and 30,240 genes for the haplotype subgenome “b”. We identified 26,489 putative gene families among *C. nepalensis* (haplotype genome “a”), *Aquilegia coerulea*⁵², *Vitis vinifera*³⁹, *Averrhoa carambola*⁵³, *Populus trichocarpa*⁵⁴, *Tripterygium wilfordii*⁵⁵, *Malus domestica*⁵⁶, *Datisca glomerata*²², *Begonia fuchsoides*²³, *Benincasa hispida*⁵⁷, *Cucumis sativus*³⁸, and *Quercus acutissima*⁵⁸, with OrthoFinder⁵⁹ v2.5.2. (Fig. 3d). Then, 1,199 orthogroups, with a minimum of 83.3% of the species having single-copy genes in any orthogroup, were used to infer the species tree with STAG⁶⁰, and the phylogenetic location of *C. nepalensis* was confirmed. *Ks* (synonymous substitutions) dot plots of haplotype genome “a” vs genome “a” and genome “a” vs *C. sativus* were

Superfamily	Number	Length (bp)	Percent (%)
Class I	102,847	63,608,643	10.25
LTR/Copia	46,213	29,513,519	4.76
LTR/Gypsy	12,994	7,928,273	1.28
LTR/unknown	39,113	24,243,803	3.91
nonLTR/pararetrovirus	730	347,201	0.06
nonLTR/LINE	3,797	1,575,847	0.25
Class II	675,085	314,548,320	50.69
TIR/hAT	17,702	5,852,661	0.94
TIR/CACTA	23,327	7,829,412	1.26
TIR/PIF-Harbinger	17,220	4,264,302	0.69
TIR/Mutator	563,911	280,243,005	45.2
TIR/Tc1_Mariner	9,638	3,344,642	0.54
Helitron	43,287	13,014,298	2.10
Other TEs	432,103	50,184,386	8.09
Total TEs	1,210,035	428,341,349	69.03

Table 2. Statistics of repeat annotation of the *C. nepalensis* genome.

Method	Database	Number	Percent (%)
eggNOG-mapper	eggNOG	59,758	98.09
	GO	33,948	55.73
	KEGG_KO	30,082	49.38
	KEGG_Pathway	18,595	30.52
	EC	12,864	21.12
	eggNOG	56,164	92.19
DIAMOND	COG	59,758	98.09
		59,981	98.46
	Swiss_Prot	47,641	78.20
	TrEMBL	59,933	98.38
	NR	59,896	98.32
InterProScan	<i>A.thaliana</i>	57,323	94.10
		60,396	99.14
	Pfam	51,101	83.88
	CDD	21,628	35.50
	SUPERFAMILY	40,074	65.78
	Interpro	53,945	88.55
	PANTHER	59,108	97.03
	Gene3D	42,846	70.33
	PIRSF	4,336	7.12
	PRINTS	8,886	14.59
	Coils	10,214	16.77
	TIGRFAM	6,982	11.46
MobiDBLite		26,460	43.43
	TMHMM	14,313	23.49
	Phobius	20,510	33.67
	SMART	19,932	32.72
Total		60,637	99.54

Table 3. Statistics of protein-coding gene functional annotation.

generated with WGD⁶¹ v0.62 (Fig. 3e), and one recent unique WGD (whole genome duplication) was revealed and was distinct from that found in *C. sativus*.

BUSCO⁶² was used for evaluating the completeness of the gene set. Out of 1,440 conserved genes, 1,400 (97.2%) were annotated, among which 1,365 (96.9%) were complete and duplicated BUSCO genes.

Three strategies were used for functional annotation of protein-coding genes (Fig. 2, Table 3): (1) we mapped gene sequences against eggNOG⁶³ 5.0 database using eggNOG-mapper⁶⁴ v2.1.6 (--target_taxa Viridiplantae) and annotated 98.1% of the genes, of which 55.7 and 49.4% were annotated with GO and KEGG items, respectively; (2) based on the principle of sequence similarity, we annotated 98.5% genes using DIAMOND⁶⁵ v2.0.12

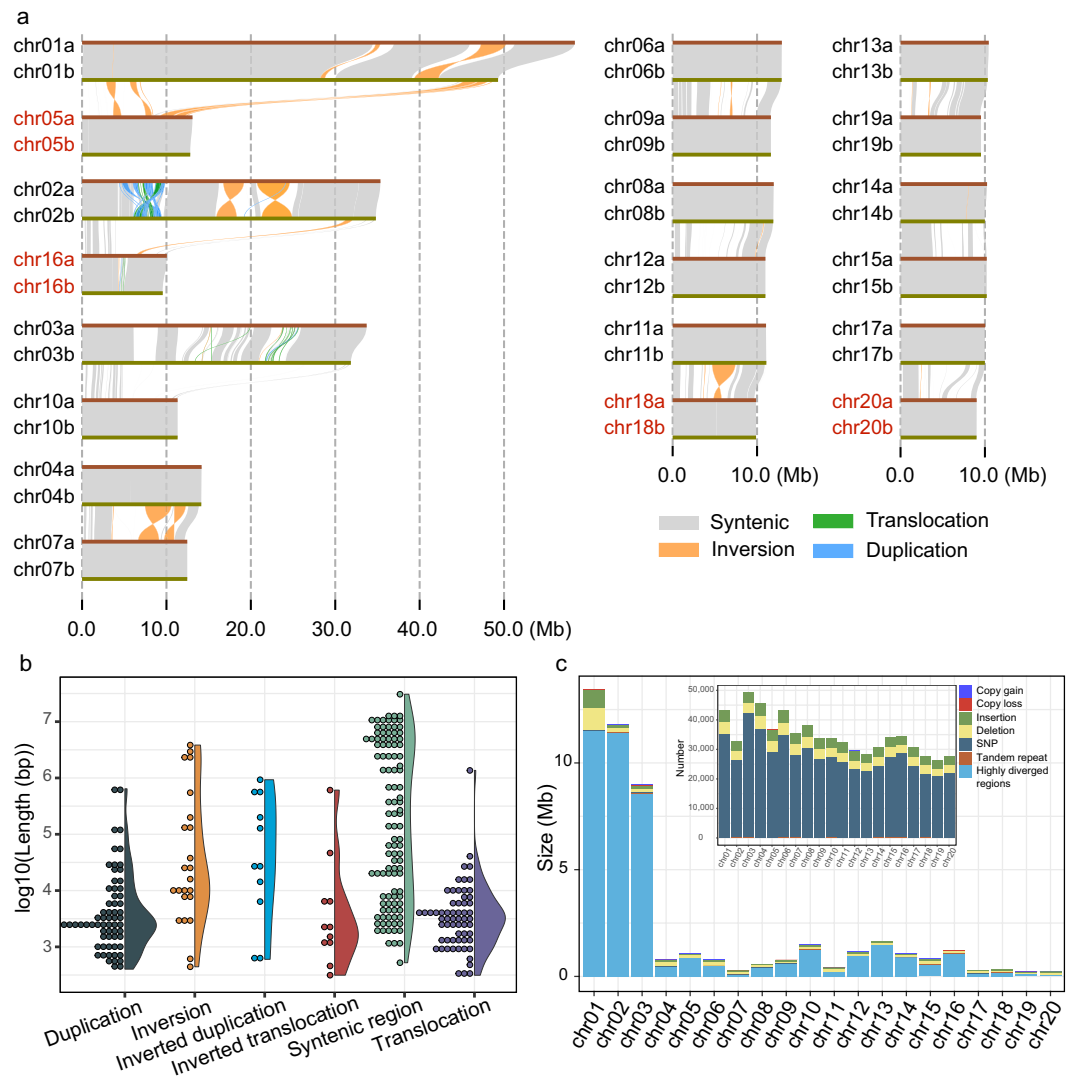


Fig. 4 Structural variation and statistics between two haplotype genome assemblies of *C. nepalensis*. **(a)** Structural variation between haplotype genomes. Subgenome “a” (chr01a–chr20a) is used as the reference sequence and subgenome “b” (chr01b–chr20b) is the query. **(b)** Size distributions of different types of structural variation between two haplotype assemblies. **(c)** Numbers and lengths of sequence differences on the syntenic region for each chromosome pair.

(--evalue $1e^{-5}$) against the following four protein databases: Swiss_Prot⁶⁶ (78.2%), TrEMBL⁶⁶ (98.4%), NR⁶⁷ (98.3%), and *Arabidopsis thaliana* genes⁴¹ (94.1%); (3) we annotated 99.1% of the genes against 14 databases using InterProScan⁶⁸ v5.52–86.0 (Table 3).

As for non-coding RNA (ncRNA) gene prediction (Fig. 2), we identified 939 tRNAs using tRNAScan-SE⁶⁹ v2.0.8, 7,297 rRNAs using Barrnap v0.9 (<https://github.com/tseemann/barrnap>) (--kingdom euk), and 982 other ncRNA using Rfam^{70,71} 16.6.

We predicted the genes in the two organelle genomes using OGAP (<https://github.com/zhangrengang/OGAP>). A total of 131 genes (89 protein-coding genes, 8 rRNAs, and 34 tRNAs) were annotated for the chloroplast genome, and 63 (42 protein-coding genes, 3 rRNAs, and 18 tRNAs) for the mitochondria genome.

Genome comparison between haplotype assemblies. The minimap2⁷² v2.24 was used to perform alignments between haplotype assemblies, and SyRI⁷³ v1.6 to identify syntenic regions and structural variations (e.g., duplications, inversions, and translocations). Plotsr⁷⁴ v0.5.4 was used for the visualization of the identified structural rearrangements (Fig. 4a). Chr01–chr03 pairs showed remarkable structural variation, while the syntenies of the other homologous chromosome pairs were mostly conserved in high collinearity with only few rearrangements. Syntenic regions were larger than the various types of structural variations (Fig. 4b). Sequence differences (local variation, e.g., SNPs, indels) on syntenic regions were identified (Fig. 4c). Highly diverged regions of long fragments were uneven among chromosome pairs, but the number of sequence differences were minor. Large fragments of collinearity between unpaired chromosomes were also detected (Fig. 4a).

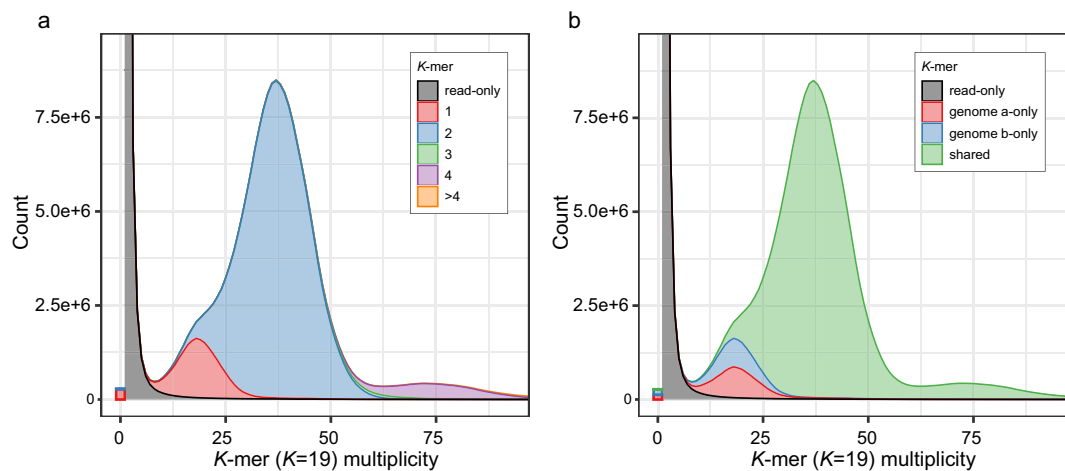


Fig. 5 Genome quality assessment with Merqury spectrum plot. **(a)** Copy number spectrum plot for haplotype assemblies of *C. nepalensis*. **(b)** Assembly spectrum plot for evaluating *K*-mer completeness.

Assembly	QV (quality value)	Error rate	Completeness (%)
Genome "a"	46.39	2.30e-05	94.12
Genome "b"	45.86	2.60e-05	93.68
Genome both "a" and "b"	46.12	2.44e-05	98.87

Table 4. Statistics of Merqury analysis for genome quality assessment.

Data Records

The raw data from PacBio HiFi, Illumina, and Hi-C sequencing were submitted to the SRA database (SRR22412655⁷⁵, SRR22026041⁷⁶, SRR22026042⁷⁷, SRR22026043⁷⁸). The haplotype-resolved genome assembly was deposited at Genbank with accession numbers GCA_027190085.1⁷⁹ and GCA_027186245.1⁸⁰. The genome assembly and gene annotation results of *C. nepalensis* were deposited in the figshare⁸¹ database.

Technical Validation

We mapped DNA and RNA sequencing reads to the final genome assembly for evaluation of the assembly quality (Fig. 2). A high read mapping rate of 99.2% was obtained when PacBio HiFi reads were mapped onto the genome using minimap2, and sequencing depth was counted and illustrated in the circos plot in Fig. 3c. We mapped the Illumina reads to the final assembly using BWA⁸² v0.7.17 and obtained a 98.7% reads mapping rate, and a low SNP heterozygosity level of ~0.0027% was obtained after SNPs were identified with SAMtools⁸³ v1.13. Furthermore, a single base error rate of ~0.0011% was acquired, and a read mapping rate of 96.2% was obtained when RNA-seq reads were mapped onto the final genome assembly using HISAT2. Since genome coverage by sequencing data was relatively high, our genome assembly has high completeness and continuity.

We performed further genome assembly quality control with Merqury⁸⁴ analysis (under $K = 19$) (Fig. 5, Table 4) based on PacBio HiFi reads. QVs (consensus quality values) for the individual haplotype genomes "a", "b", and shared for both "a" and "b" genomes are 46.39, 45.86, and 46.12, respectively. *K*-mer completeness scores for individual genomes "a", "b", and shared for both "a" and "b" genomes are 94.12, 93.68, and 98.87%, respectively. Again, our presented haplotype-resolved genome assembly was confirmed the good quality in completeness.

We further performed BUSCO assessments for the assembly (Table 1), where it was revealed that complete core genes (including single and multiple copies) accounted for 93.0%, while the missing gene rate accounted for only 4.9%, underscoring the good gene integrity of the assembly.

Code availability

All data processing commands and pipelines were carried out in accordance with the instructions and guidelines provided by the relevant bioinformatic software.

Received: 27 December 2022; Accepted: 20 April 2023;

Published online: 09 May 2023

References

- Oginuma, K., Nakata, M., Suzuki, M. & Tobe, H. Karyomorphology of *Coriaria* (Coriariaceae): Taxonomic implications. *The Botanical Magazine Tokyo* **104**, 297–308 (1991).
- Montserrat, P. Root nodules of *Coriaria*. *Nature* **182**, 475–475 (1958).
- Hu, C., Zhou, P., Zhou, Q., Chen, H. & Akkermans, A. D. L. Nodulation and molecular characterization of pure cultures isolated from root nodules of *Coriaria nepalensis*. *Chinese Science Bulletin* **43**, 695–698 (1998).
- Awasthi, P., Bargali, K., Bargali, S. S. & Jhariya, M. K. Structure and functioning of *Coriaria nepalensis* dominated shrublands in degraded hills of Kumaun Himalaya. I. Dry matter dynamics. *Land Degradation & Development* **33**, 1474–1494 (2022).

5. Mourya, N. R., Bargali, K. & Bargali, S. S. Impacts of *Coriaria nepalensis* colonization on vegetation structure and regeneration dynamics in a mixed conifer forest of Indian Central Himalaya. *Journal of Forestry Research* **30**, 305–317 (2019).
6. Bargali, K. & Tewari, A. Growth and water relation parameters in drought-stressed *Coriaria nepalensis* seedlings. *Journal of Arid Environments* **58**, 505–512 (2004).
7. Zeng, X. M., Xu, X. L., Yi, R. Z., Zhong, F. X. & Zhang, Y. H. Sap flow and plant water sources for typical vegetation in a subtropical humid karst area of southwest China. *Hydrological Processes* **35**, e14090 (2021).
8. Tiwari, M., Singh, S. P., Tiwari, A. & Sundriyal, R. C. Effect of symbiotic associations on growth of host *Coriaria nepalensis* and its facilitative impact on oak and pine seedlings in the Central Himalaya. *Forest Ecology and Management* **184**, 141–147 (2003).
9. Fang, S. Z., Li, H. Y. & Xie, B. D. Decomposition and nutrient release of four potential mulching materials for poplar plantations on upland sites. *Agroforestry Systems* **74**, 27–35 (2008).
10. Yan, K. *et al.* Current re-vegetation patterns and restoration issues in degraded geological phosphorus-rich mountain areas: A synthetic analysis of Central Yunnan, SW China. *Plant Divers* **39**, 140–148 (2017).
11. Ahmad, A., Khan, A., Kumar, P., Bhatt, R. P. & Manzoor, N. Antifungal activity of *Coriaria nepalensis* essential oil by disrupting ergosterol biosynthesis and membrane integrity against *Candida*. *Yeast* **28**, 611–617 (2011).
12. Kumar, P. *et al.* Antimicrobial activities of essential oil and methanol extract of *Coriaria nepalensis*. *Nat Prod Res* **25**, 1074–1081 (2011).
13. Zhao, F. *et al.* New sesquiterpenes from the roots of *Coriaria nepalensis*. *Tetrahedron* **68**, 6204–6210 (2012).
14. Fang, H. L., Shang, F. N., Qian, J. & Duan, B. Z. Phylogenetic relationship and characterization of the complete chloroplast genome of the *Coriaria nepalensis* Wall. in China, a least concern folk medicine. *Mitochondrial DNA Part B-Resources* **5**, 1718–1719 (2020).
15. Li, M. L. *et al.* Semisynthesis and antifedant activity of new acylated derivatives of tutin, a sesquiterpene lactone from *Coriaria sinica*. *Heterocycles* **71**, 1155–1162 (2007).
16. Guo, L. X., Qiang, T. T., Ma, Y. M., Wang, K. & Du, K. Optimisation of tannin extraction from *Coriaria nepalensis* bark as a renewable resource for use in tanning. *Industrial Crops and Products* **149**, 112360 (2020).
17. Guo, L. X., Qiang, T. T., Ma, Y. M., Ren, L. F. & Dai, T. T. Purification and characterization of hydrolysable tannins extracted from *Coriaria nepalensis* bark using macroporous resin and their application in gallic acid production. *Industrial Crops and Products* **162**, 113302 (2021).
18. Yokoyama, J., Suzuki, M., Iwatsuki, K. & Hasebe, M. Molecular phylogeny of *Coriaria*, with special emphasis on the disjunct distribution. *Mol Phylogenet Evol* **14**, 11–19 (2000).
19. Chase, M. W. *et al.* Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbcl*. *Annals of the Missouri Botanical Garden* **80**, 528–580 (1993).
20. Swensen, S. M., Mullin, B. C. & Chase, M. W. Phylogenetic affinities of Datisceae based on an analysis of nucleotide sequences from the plastid *rbcl* gene. *Systematic Botany* **19**, 157–168 (1994).
21. Swensen, S. M. The evolution of actinorhizal symbioses: Evidence for multiple origins of the symbiotic association. *American Journal of Botany* **83**, 1503–1512 (1996).
22. Griesmann, M. *et al.* Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* **361**, eaat1743 (2018).
23. Li, L. *et al.* Genomes shed light on the evolution of *Begonia*, a mega-diverse genus. *New Phytol* **234**, 295–310 (2022).
24. Xie, T. *et al.* De novo plant genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana*. *Mol Plant* **8**, 489–492 (2015).
25. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
26. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
27. Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* **34**, 550–557 (2017).
28. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
29. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**, 95–98 (2016).
30. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
31. Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst* **3**, 99–101 (2016).
32. Xu, M. *et al.* TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience* **9**, giaa094 (2020).
33. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
34. Prysacz, L. P. & Gabaldon, T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res* **44**, e113 (2016).
35. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
36. Jin, J. J. *et al.* GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol* **21**, 241 (2020).
37. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* **20**, 275 (2019).
38. Huang, S. *et al.* The genome of the cucumber, *Cucumis sativus* L. *Nat Genet* **41**, 1275–1281 (2009).
39. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
40. International Peach Genome, I. *et al.* The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat Genet* **45**, 487–494 (2013).
41. Arabidopsis Genome, I. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
42. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644–652 (2011).
43. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360 (2015).
44. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
45. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
46. Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654–5666 (2003).
47. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
48. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**, 188–196 (2008).
49. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
50. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

51. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments. *Genome Biol* **9**, R7 (2008).
52. Filiault, D. L. *et al.* The *Aquilegia* genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *Elife* **7**, e36426 (2018).
53. Wu, S. *et al.* The genome sequence of star fruit (*Averrhoa carambola*). *Hortic Res* **7**, 95 (2020).
54. Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa*. *Science* **313**, 1596–1604 (2006).
55. Tu, L. *et al.* Genome of *Tripterygium wilfordii* and identification of cytochrome P450 involved in triptolide biosynthesis. *Nat Commun* **11**, 971 (2020).
56. Duan, N. *et al.* Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. *Nat Commun* **8**, 249 (2017).
57. Xie, D. *et al.* The wax gourd genomes offer insights into the genetic diversity and ancestral cucurbit karyotype. *Nat Commun* **10**, 5158 (2019).
58. Fu, R. *et al.* Genome-wide analyses of introgression between two sympatric Asian oak species. *Nat Ecol Evol* **6**, 924–935 (2022).
59. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).
60. Emms, D. M. & Kelly, S. STAG: Species tree inference from all genes. *bioRxiv*, 267914 (2018).
61. Sun, P. *et al.* WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol Plant* **15**, 1841–1851 (2022).
62. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
63. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309–D314 (2018).
64. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol* **34**, 2115–2122 (2017).
65. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
66. Consortium, T. U. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480–D489 (2020).
67. Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **42**, D7–D17 (2013).
68. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
69. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955–964 (1997).
70. Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research* **49**, D192–D200 (2020).
71. Kalvari, I. *et al.* Non-coding RNA analysis using the Rfam database. *Curr Protoc Bioinformatics* **62**, e51 (2018).
72. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
73. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20**, 277 (2019).
74. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **38**, 2922–2926 (2022).
75. NCBI Sequence Read Archive (SRR22412655) <https://identifiers.org/insdc.sra:SRR22412655> (2022).
76. NCBI Sequence Read Archive (SRR22026041) <https://identifiers.org/insdc.sra:SRR22026041> (2022).
77. NCBI Sequence Read Archive (SRR22026042) <https://identifiers.org/insdc.sra:SRR22026042> (2022).
78. NCBI Sequence Read Archive (SRR22026043) <https://identifiers.org/insdc.sra:SRR22026043> (2022).
79. NCBI Assembly https://identifiers.org/insdc.gca:GCA_027190085.1 (2022).
80. NCBI Assembly https://identifiers.org/insdc.gca:GCA_027186245.1 (2022).
81. Zhao, SW. *et al.* Haplotype-resolved genome assembly of *Coriaria nepalensis*, a non-legume nitrogen-fixing shrub associated with *Frankia*, *figshare*, <https://doi.org/10.6084/m9.figshare.c.6292962.v1> (2023).
82. Li, H. J. A. P. A. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997v2> (2013).
83. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
84. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245 (2020).

Acknowledgements

This research was supported by the National Natural Science Foundation of China (32171816) and the National Key R&D Program of China (2022YFD2200103).

Author contributions

Jian-Feng Mao and Wei Zhao conceived and designed the study; Yong-Peng Ma collected the samples; Shi-Wei Zhao, Jing-Fang Guo, Lei Kong, Shuai Nie, Xue-Mei Yan, Tian-Le Shi, Xue-Chan Tian, Hai-Yao Ma, Yu-Tao Bao, Zhi-Chao Li, Zhao-Yang Chen, Ren-Gang Zhang performed bioinformatics; Shi-Wei Zhao drafted the manuscript; Jian-Feng Mao, Yousry A. El-Kassaby and Ilga Porth revised the manuscript. Shi-Wei Zhao, Jing-Fang Guo and Lei Kong contributed equally to this work.

Funding

Open access funding provided by Umea University.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.Z. or J.-F.M.

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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023