DNA Research, 2022, 29(3), 1–10
https://doi.org/10.1093/dnares/dsac009
Advance Access Publication Date: 25 April 2022
Resource Article: Genomes Explored

OXFORD

Resource Article: Genomes Explored

Chromosome-level genome assembly and characterization of *Sophora Japonica*

Weixiao Lei[1†], Zefu Wang[2†], Man Cao[1†], Hui Zhu[1], Min Wang[1], Yi Zou[1], Yunchun Han[1], Dandan Wang[1], Zeyu Zheng[1], Ying Li[1], Bingbing Liu[3]*, and Dafu Ru[1]*

[1]State Key Laboratory of Grassland Agro-Ecosystems, and College of Ecology, Lanzhou University, Lanzhou 730000, China, [2]Key Laboratory of Bio-resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610000, China, and [3]Institute of Loess Plateau, Shanxi University, Taiyuan 030006, China

*To whom correspondence should be addressed. Tel. 13880788291. Email: rudf@lzu.edu.cn (D.R.); Tel. 13880788291. lbb2015@sxu.edu.cn (B.L.)

[†]The first three authors contributed equally to this work.

Received 14 December 2021; Editorial decision 3 April 2022; Accepted 7 April 2022

Abstract

Sophora japonica is a medium-size deciduous tree belonging to Leguminosae family and famous for its high ecological, economic and medicinal value. Here, we reveal a draft genome of *S. japonica*, which was ~511.49 Mb long (contig N50 size of 17.34 Mb) based on Illumina, Nanopore and Hi-C data. We reliably assembled 110 contigs into 14 chromosomes, representing 91.62% of the total genome, with an improved N50 size of 31.32 Mb based on Hi-C data. Further investigation identified 271.76 Mb (53.13%) of repetitive sequences and 31,000 protein-coding genes, of which 30,721 (99.1%) were functionally annotated. Phylogenetic analysis indicates that *S. japonica* separated from *Arabidopsis thaliana* and *Glycine max* ~107.53 and 61.24 million years ago, respectively. We detected evidence of species-specific and common-legume whole-genome duplication events in *S. japonica*. We further found that multiple TF families (e.g. *BBX* and *PAL*) have expanded in *S. japonica*, which might have led to its enhanced tolerance to abiotic stress. In addition, *S. japonica* harbours more genes involved in the lignin and cellulose biosynthesis pathways than the other two species. Finally, population genomic analyses revealed no obvious differentiation among geographical groups and the effective population size continuously declined since 2 Ma. Our genomic data provide a powerful comparative framework to study the adaptation, evolution and active ingredients biosynthesis in *S. japonica*. More importantly, our high-quality *S. japonica* genome is important for elucidating the biosynthesis of its main bioactive components, and improving its production and/or processing.

Key words: *Sophora japonica*, genome, nanopore, Hi-C, WGD

1. Introduction

Sophora japonica (2n = 28) is a popular perennial ornamental and medicinal plant of the subfamily Papilionoideae commonly known as Chinese scholar tree or Japanese pagoda tree.[1] Widely distributed in Asia and Europe, it has been cultivated for more than 3,000 years in China. As a medicinal plant, almost every part of *S. japonica*, including the leaves, buds, flowers, seeds and bark, is used medicinally in Asia. For example, the buds and fruits are commonly used as

ingredients in traditional Chinese medicine,[2–4] and flavonoid components of the buds and fruits reportedly have haemostatic properties.[5] Moreover, the dry flowers (*Huaihua* or Flos Sophorae) and flower buds (*Huaimi* or Flos Sophorae Immaturus) are included in both Chinese and European Pharmacopeia and contain diverse kinds of components such as tetraglycosides, flavones, isoflavones, isoflavone tetraglycosides and flavonol glycosides (rutin, for instance).[6] According to Chinese pharmacopeia records, clinical trials and modern pharmacological studies, rutin is one of the main effective components and reportedly has more than 40 therapeutic properties.[7,8]

Sophora japonica is also widely used in urban greening programmes and planted by highways as it has high resistance to pests and diseases, pollution tolerance and adaptability to diverse environmental factors. In urban areas of Beijing, China, where *S. japonica* accounts for 81% of the street trees,[9] and elsewhere it plays important roles in microclimate regulation, maintenance of healthy urban ecosystems and improvement of air quality. However, despite its medicinal and ecological value, very little genomic information is available for *S. japonica*.

Therefore, we have assembled and annotated the *S. japonica* genome using long reads obtained from Oxford Nanopore (ONT) sequencing and short reads from Illumina sequencing. This yielded a genome assembly of 511.49 Mb with a ~17.34 Mb contig N50 size. Using Hi-C data, we associated 91.62% of the assembled bases with 14 pseudo-chromosomes, with an improved N50 of 31.32 Mb. We identified a species-special whole-genome duplication (WGD) event and found many of the duplicated genes caused by WGD have contributed to *S. japonica*'s adaptation and biosynthesis. We further performed whole-genome resequencing for 35 individuals to examine the population structure. The high-quality chromosome-level *S. japonica* genome provides robust foundations for investigating pests and diseases resistance, pollution tolerance and environmental adaptability and biosynthesis of its main active ingredients of the species.

2. Materials and methods

2.1. Sample collection and sequencing

Leaves were collected from a *S. japonica* individual in Lanzhou, Gansu Province, China (36°2′57″N, 103°51′34″E) (Fig. 1). Genomic DNA was extracted with a QIAGEN® Genomic Kit following the manufacturer's recommendations. An Illumina paired-end library (with insert size of 350 bp) was constructed using an Illumina genomic DNA sample preparation kit. The Illumina NovaSeq 6000 platform was then used for sequencing and yielded ~103.70 Gb of raw sequence data (~193.55× coverage of the genome) (Supplementary Table S1). A Nanopore library was constructed and sequenced with a PromethION sequencer (Oxford Nanopore Technologies, UK). A total of 60.45 Gb (~112.83×) of long reads were generated (Supplementary Table S1). The Hi-C library was constructed using young leaves from the same *S. japonica* individual. Illumina NovaSeq 6000 platform was used to generate 60.73 Gb Hi-C data (Supplementary Table S1). We also collected four tissue samples (including leaves, stems, flowers and fruits) from the same individual for RNA-Seq. For each sample, the fresh tissue was used to construct cDNA library. Finally, a total of 27.47 Gb paired-end reads were obtained for all RNA-Seq samples (Supplementary Table S1).

2.2. Genome size and heterozygosity estimation

To estimate the size and heterozygosity level of the *S. japonica* genome, 17-mer spectrum analysis was performed following a procedure previously applied in *Oplegnathus fasciatus* genome research.[10] Based on the total number of k-mers (74,577,100,335), the estimated genome size is ~535.77 Mb (Supplementary Fig. S1 and Table S2). The estimated heterozygosity level and repetition frequency of the *Styphnolobium japonicum* genome are 1.19% and 52.77%, respectively (Supplementary Fig. S1 and Table S2).

2.3. Assembly of the *S. japonicum* genome

We used the previously generated 60.45 Gb ONT single molecule long reads for genome assembling. NextDenovo (version 2.3) (https://github.com/Nextomics/NextDenovo), including sequencing error correction, preliminary assembly and genome polishing, was performed under the recommended pipeline. The NextCorrect module was used for raw read correction and consensus sequence (CNS) extraction. The NextGraph module was used for preliminary assembly, and the NextPolish (version 1.2.2)[11] module for genome polishing. According to previously described procedures,[12] the *S. japonica* assembly was further refined based on 60.73 Gb Hi-C data (Supplementary Table S1) to improve the primary genome assembly to the chromosome level. BUSCO (version 4)[13] was further used to assess the completeness of the genome assembly with embryophyta_odb10 database.

2.4. Annotation of repeat and non-coding RNAs

We used two strategies to predict repeats in *S. japonica* genome. For homology-based strategy, RepeatMasker (version 4.1.0)[14] was performed for homology searching of repetitive elements against Repbase database (version 23).[15] For *ab initio* identification, LTR Finder (version 1.07)[16] and RepeatModeler (version 2.0) were performed, respectively.

To identify non-coding RNA (ncRNA) sequences, two strategies were used: database searching and model-based prediction. rfam_scan.pl (version 1.0.4) was used to search against Rfam database (version 11.0)[17] to detect small nuclear RNA (snRNA) and microRNA (miRNA) sequences, and tRNAscan-SE (version 2.0)[18] was used (with eukaryotic parameters) to predict transfer RNA (tRNA) sequences. In addition, RNAmmer (version 1.2)[19] was used to predict ribosomal RNA (rRNA) and its subunits.

2.5. Prediction and functional annotation of protein-coding genes

To maximize the reliability of the gene annotation process, repetitive regions in the assembled genome were masked before gene annotation. The prediction was performed by homology-based, transcriptome-based and *ab initio* strategies to identify high-quality protein-coding genes. In homology-based prediction process, protein-coding sequences of seven species (*Arabidopsis thaliana*, *Cicer arietinum*, *Glycine soja*, *Lupinus albus*, *Populus trichocarpa*, *Trifolium pratense* and *Vigna unguiculata*) were downloaded from NCBI or Phytozome (v12.1; https://phytozome.jgi.doe.gov/) (Supplementary Table S3) and align with our *S. japonica* genome using TBLASTN (*E*-value $< 1 \times 10^{-5}$). We used GeneWise (version 2.4.1)[20] to align homologous genome sequences with matched proteins to define gene models. To generate transcriptome-based predictions based on RNA-Seq data, we sequenced RNA-Seq for leaves, stems, flowers and fruits of *S. japonica*, respectively. All RNA reads were initially aligned by HISAT2 (version 2.1.0)[21] based on the reference genome and further assembled into transcripts using Trinity (version 2.6.6).[22] These assembled sequences were compared with

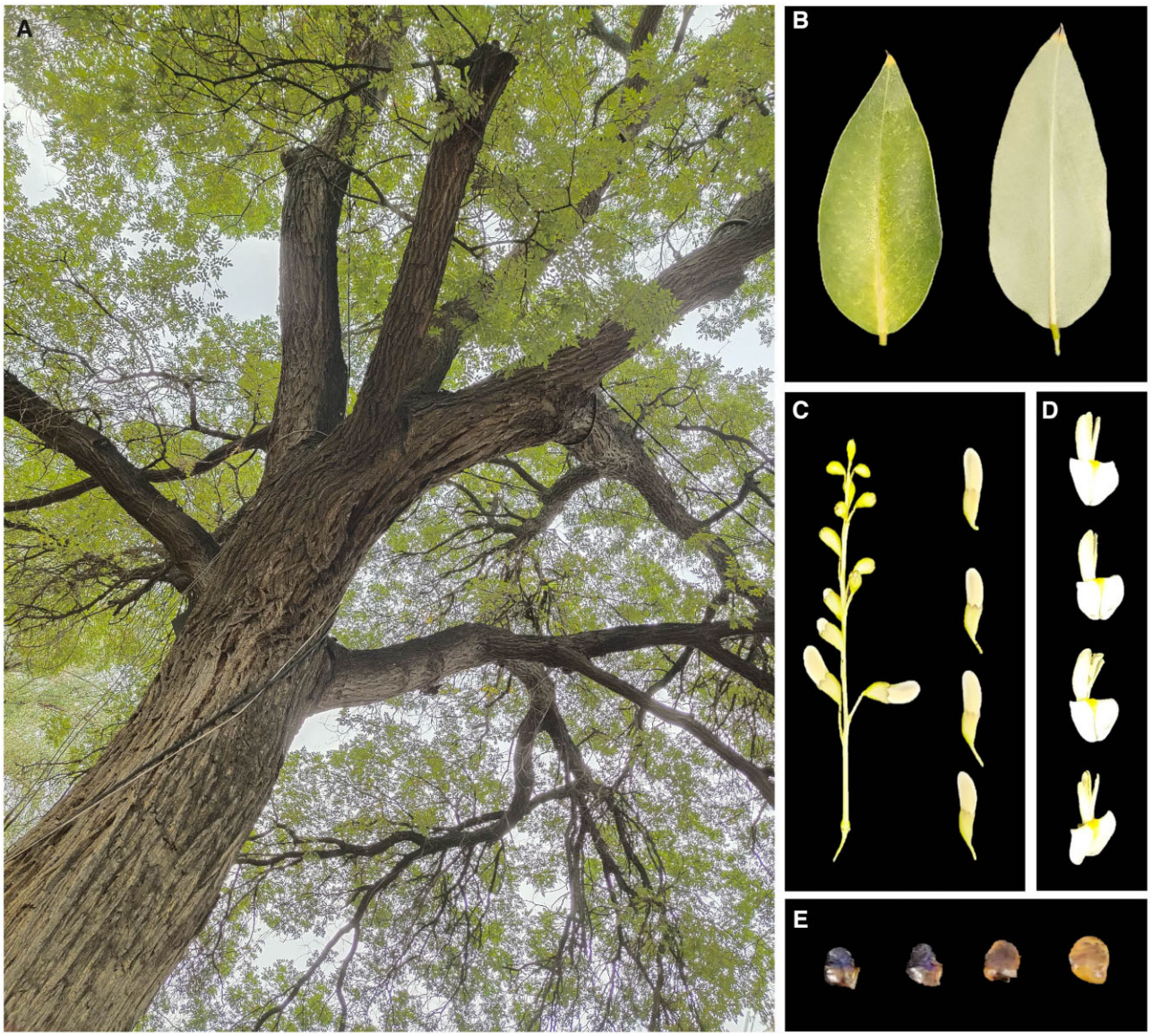

Figure 1. (A) Photo of *S. japonica* tree; (B) photo of an *S. japonica* leaf; (C) flower buds (*Huaimi*) of *S. japonica*; (D) flowers (*Huaihua*) of *S. japonica*; (E) seeds of *S. japonica*.

the *S. japonica* genome using the Program to Assemble Spliced Alignments (PASA) (version 2.3.3).[23] Then we clustered the valid transcript alignments based on the positions obtained from the genome mapping, and finally assembled them into gene structures with the scripts from PASA packages (version 2.3.3) ('build_comprehensive_transcriptome.dbi' and 'pasa_asmbls_to_training_set.dbi') following default parameter settings. For *ab initio* prediction, the genome was subjected to analysis by Augustus[24] with parameters trained using PASA self-trained gene models. EVM (version 1.1.1)[25] was then used to merge the gene models obtained by the three methods into a final consensus set. The gene model was finally updated by PASA to generate untranslated regions and alternative splicing variants.

We used BLASTP (*E*-value $< 1 \times 10^{-5}$)[26] to annotate functions of protein-coding genes based on entries in the Swiss-Prot (version 2020_04),[27] TrEMBL (version 2020_04)[27] and NR databases (release 20200502). The protein domains were predicted by searching against InterPro database (version v84)[28] using InterProScan (version 5.32-71.0).[29] The Gene Ontology (GO) terms were obtained from the corresponding InterPro entries. The pathways for each gene were predicted by BBH method on KAAS website (https://www.genome.jp/tools/kaas/) based on Kyoto Encyclopedia of Genes and Genomes (KEGG)[30] databases.

2.6. Phylogeny and evolution of gene families

To explore *S. japonica*'s evolutionary relationships, we used OrthoFinder (version 2.3.8)[31] to cluster its genes with genes from 16 other dicot species: *Vigna radiata*, *Vigna unguiculate*, *Phaseolus vulgaris*, *Glycine max*, *G. soja*, *Cajanus cajan*, *Medicago truncatula*, *T. pratense*, *C. arietinum*, *Lotus japonicus*, *L. albus*, *Lupinus angustifolius*, *Arachis duranensis*, *Arachis ipaensis*, *Senna tora and A. thaliana* (Supplementary Table S3). Before clustering, the genes for each species were filtered. Only the longest transcript of each gene was retained and short protein-coding sequences (fewer than 40 amino

acids) or error sequences (stop codons appearing prematurely) were removed. We identified 204 groups of strictly single-copy genes and used them to construct maximum likelihood trees to assess their evolutionary relationships using GTRGAMMA models in RAxML (version 8.2.12).[32] We also used the 'mcmctree' programme in the PAML (version 4.9j) package to estimate divergence times,[33] based on known approximate divergence times of *A. duranensis* and *M. truncatula* [49–58 million years ago (Ma)], as well as *G. max* and *A. thaliana* (98–117 Ma) (http://www.timetree.org/).

The expansion and contraction of gene families are important drivers of metabolite variation and species adaptation during plant evolution.[34] Thus, we explored the expansion and contraction of orthologous gene families in the *S. japonica* genome using CAFE (version 2.2) with default parameters.[35] The expansive and contractive gene families in *S. japonica* were extracted using home-made perl script and further applied to GO enrichment analyses using TBtools (version 1.075).[36]

2.7. Whole-genome duplications

We used the MCScanX (version 1.1.11) package to search for collinear blocks (defined as regions containing more than five collinear genes) between pseudochromosomes of pairs of the 16 genomes listed above and our *S. japonica* genome. The number of synonymous substitutions per synonymous site values of collinearity of orthologous gene pairs were determined using the Perl script 'add_ka_and_ks_to_collinearity.pl' implemented in MCScanX.[37] We further determined the expanded genes which caused by WGD waves and also applied to GO enrichment analyses.

2.8. Population genetic analyses

An Illumina NovaSeq 6000 sequencing platform and PE150 sequencing strategy were used for whole-genome resequencing of 35 samples with high coverage (~18×) (Supplementary Table S4). BWA (version 0.7.17)[38] was used to map the reads to the chromosome-level genome of *S. japonicum*. Picard software (version 2.23.6) was applied to remove duplicates (broadinstitute.github.io/picard). SAMtools (version 1.10)[39] was applied to identify single-nucleotide polymorphisms (SNPs) and InDels. To obtain high quality SNPs, we filtered the SNPs with the following criteria: (i) no InDel within 5 bps window; (ii) root mean square quality no less than 20; (iii) missing rate <20%; (iv) minimum allele frequency no <0.05 and *P*-value of Hardy–Weinberg equilibrium no <0.001. In addition, the base quality and mapping quality were set to 20 and 30, respectively.

To analyse phylogenetic relationships, we used TreeBeST (version 1.9.2)[40] to construct a neighbour-joining tree, ADMIXTURE (version 1.23)[41] to infer the population structure, and PLINK (version 1.07) for principal component analysis (PCA)-based population clustering,[42] respectively.

2.9. Demographic history

We used pairwise sequentially Markovian coalescence (PSMC) (version 0.6.5-r67) modelling[43] to infer demographic history. It is capable of reconstructing the history of changes in population size over time using the distribution of the most recent common ancestor between two alleles within an individual. SAMtools (version 1.3)[39] was used to obtain consensus sequences and divide them into non-overlapping 100 base pair bins, with the following parameters: -N25 -t15 -r5 -p '$4+25\times2+4+6$'. A generation time of 7 years and mutation rate of 3.65×10^{-9} per nucleotide per year were used to convert the scaling time and population size to actual time and size.

3. Results

3.1. Assembly of the *S. japonicum* genome

To perform a *de novo* assembly of the *S. japonica* genome (estimated genome size ~535.77 Mb; Supplementary Table S2 and Fig. S1), we combined several sequencing technologies and assembly strategies (see Materials and methods). A total of 60.45 Gb ONT reads, corresponding to ~112.83-fold coverage of the estimated genome size, were generated and used for *de novo* assembly (Supplementary Table S1). The primary genome assembly of *S. japonica* includes 110 contigs with N50 = 17.34 Mb and longest contig of 32.91 Mb (Table 1). The genome is 511,488,806 bp long with an average GC content of 33.77% (Table 1). The genome size is similar to that of assembled *C. arietinum* and *L. japonicus* (532.29 Mb, 544.14 Mb, respectively).[44,45]

The assembly was further refined using 60.73 Gb Hi-C data (Supplementary Table S1) and previously described procedures[12] for chromosome-level anchorage. *Sophora japonica* has 2n = 28 chromosomes, according to karyotype analyses, and in total 468.63 Mb (91.62%) contig sequences were anchored onto 14 chromosomes (Fig. 2A and B). The final scaffold genome was 511,485,306 bp long in size, a bit larger than the contig genome. In addition, the scaffold N50 was increased to 31.32 Mb, with a longest scaffold of 60.53 Mb and 98.1% coverage according to BUSCO (version 4, dataset: embryophyta_odb10) analysis (single, duplicated, fragmented and

Table 1. Summary of the genome assembly and annotation tables

Genome assembly		
Estimated genome size	535.77 Mb	
N50 length (contig)	17.34 Mb	
Longest contig	32.92 Mb	
Number of contigs	110	
Total length of contigs	511.49 Mb	
N50 length (scaffold)	31.32 Mb	
Longest scaffold	60.53 Mb	
Number of scaffolds	99	
Total length of scaffolds	511.49 Mb	
Average GC content (%)	33.77	
BUSCO score of assembly (%)	98.1 (S: 88.7, D: 9.4), F: 1.0, M: 0.9	
Transposable elements		
Annotation	Percent (%)	Total length (bp)
DNA	5.59	27,990,056
LINE	1.17	5,880,495
LTR	37.98	190,194,740
SINE	0.15	762,931
Satellite	0.18	904,219
Simple repeat	9.37	46,924,249
Small RNA	0.09	438,768
Unknown	20.31	101,699,257
Total	53.13	271,762,340
Protein-coding genes		
Predicted genes	31,000	
Average genes length(bp)	4,864.39	
Average CDS length (bp)	1,304.15	
Average exons per gene	5.63	
Average exon length (bp)	231.49	
Average intron length (bp)	620.11	
BUSCO score of annotation (%)	97.4 (S: 87.4, D: 10.0), F: 1.0, M: 1.6	

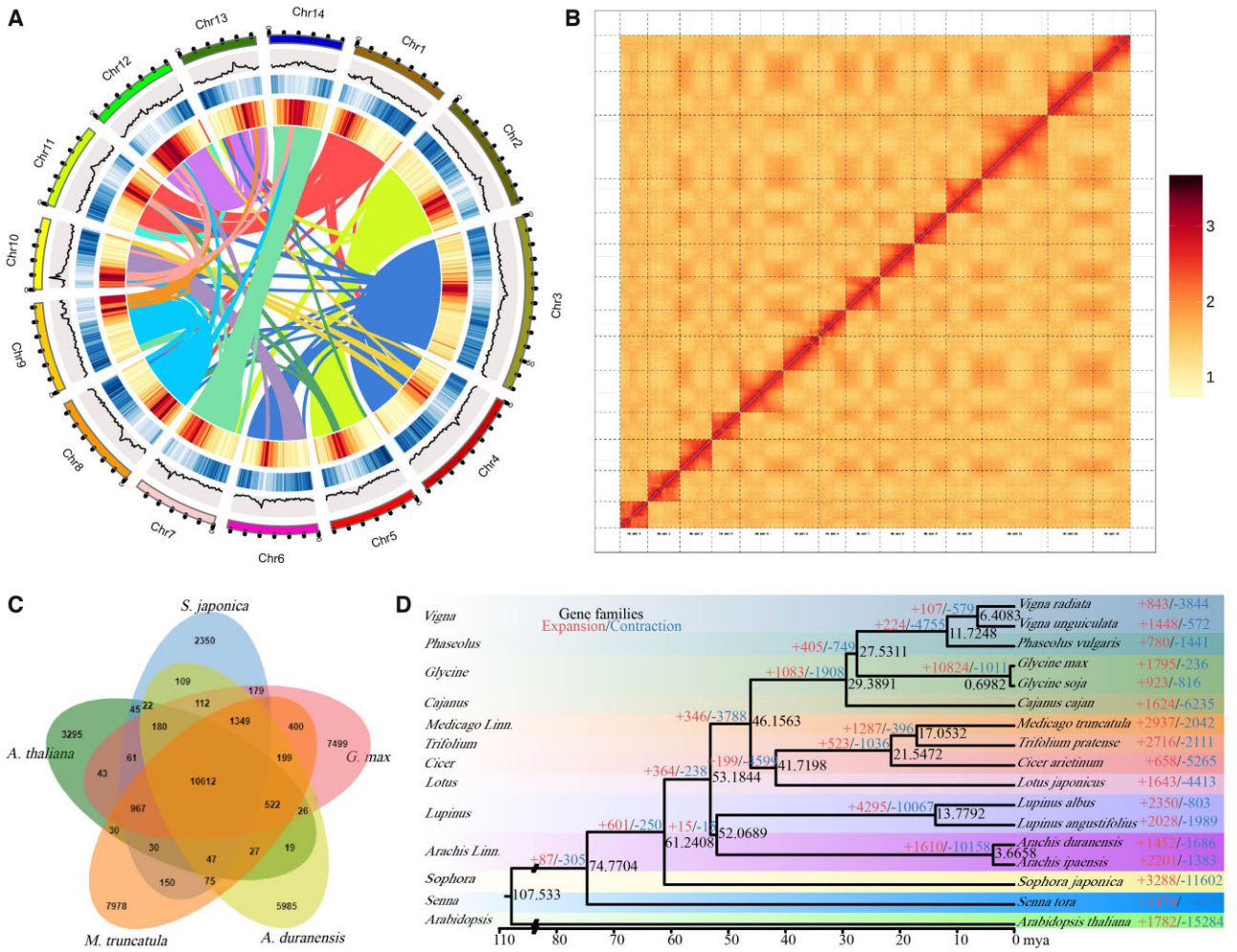

Figure 2.. (A) Circos plot showing genomic features of *S. japonica*. Concentric circles, outer to inner, show GC density, gene density, repetitive sequence density and collinearity, respectively. (B) Heat map of chromatin contact matrices generated by aligning the Hi-C dataset to the *S. japonica* genome. (C) Phylogenetic relationships and divergence times of commelinid plants obtained using the maximum likelihood (ML) method with *A. thaliana* as a distant outgroup. Divergence times were estimated using the 'mcmctree' program incorporated in the PAML package.

missed percentages: 88.7%, 9.4%, 1.0% and 0.9%, respectively) (Table 1). Furthermore, 99.45% of the Illumina data could be mapped to the chromosome-level genome (coverage of 97.81%), and 97.47%, 97.12% and 96.64% of the assembled genome sequence could be covered by at least 4-fold, 10-fold and 20-fold Illumina data, respectively (Supplementary Table S5). These results clearly suggest that our assembly has high quality and is relatively complete.

3.2. Repeat and ncRNAs annotation

We identified ~271.76 Mb repetitive elements in total, accounting for 53.13% of the genome (Table 1). Of these, DNA transposons, long terminal repeat, long interspersed nuclear elements and short interspersed nuclear elements accounted for 5.59%, 37.98%, 1.17% and 0.15% of the genome, respectively (Table 1). The fraction of repetitive sequences in the genome is comparable to other genomes, like those of grapevine (41%),[46] castor bean (50%)[47] and soybean (59%),[48] but less than that seen in the genomes of sorghum (62%)[49] and maize (85%).[50]

The final ncRNA annotation results included 456 miRNA, 75 tRNA, 1,488 rRNA and 160 snRNA sequences with average lengths of 165.23, 73.81, 398.98 and 141.29 bp, respectively, accounts for a small part of the genome (Supplementary Table S6).

3.3. Prediction and functional annotation of protein-coding genes

We generated 31,000 gene models in total. Coding sequences of the predicted genes are 4,864.39 bp long and have 5.63 exons, on average (Table 1). In addition, among 31,000 predicted protein-coding genes, 97.07% (30,092), 92.17% (28,574), 97.29% (30,160), 82.07% (25,442), 96.38% (29,879) and 46.23% (14,330) were obtained from searches against the InterPro, GO, TrEMBL, Swiss-PROT, NR and KEGG database, respectively. In total, functional annotations resulted in the assignment of putative functional annotations for 30,721 (99.1%) genes (Supplementary Table S7). BUSCO assessment indicated that 97.4% of the highly conserved plant genes (1,614 in total) are completely present in the genome (single, duplicated, fragmented and missed percentages: 87.4%, 10.0%, 1.0% and 1.6%, respectively) (Table 1). These results clearly show that the annotated gene set of *S. japonica* is relatively complete.

3.4. Phylogeny and evolution of gene families

From comparison with published genomes of the 16 species listed in Materials and methods, we found *S. japonica* shared 10,612 gene families with four other species (*A. thaliana*, *M. truncatula*, *A. duranensis* and *G. max*) and contained 2,350 unique gene families (Fig. 2C). Meanwhile, we identified 204 single-copy orthologous genes. Phylogenetic inferences based on 204 single-copy orthologous genes indicates that *S. japonica* separated from *A. thaliana* and *G. max* about 107.53 Ma and 61.24 Ma, respectively (Fig. 2D). After comparing gene families of all 17 species, we inferred that the *S. japonica* genome includes 3,288 extended families and 11,602 contracted families (Fig. 2D). Among them, 568 expanded and 325 contracted gene families were significant ($P < 0.01$, Supplementary Tables S8 and S9). GO enrichment analysis indicated that gene families related to 'response to fungus (GO: 0009620)', 'regulation of root development (GO: 2000280)', 'regulation of response to stress (GO: 0080134)', 'flavonoid metabolic process (GO: 0009812)', 'defense response to oomycetes (GO: 0002229)' have all expanded (Supplementary Table S8 and Fig. S2A), which may explain the high general environmental resistance of *S. japonica*. While gene families related to 'terpene biosynthetic process (GO: 0051762)', 'sesquiterpene metabolic process (GO: 0051761)', 'ribosomal subunit (GO: 0044391)', 'alkaloid biosynthetic process (GO: 0035825)' have all contracted (Supplementary Table S9 and Fig. S2B). Furthermore, the *PAL* TF family, which plays an important role in the process of plant resistance through regulating the synthesis of plant antibiotics,[51] expanded from four in *A. thaliana* and five in *M. truncatula* to six in *S. japonica* genome (Supplementary Fig. S2C). In addition, we found the *BBX* TF family, which is widely involved in plant growth and development[52,53] and plays important role in abiotic stress response,[54,55] expanded from 32 members in *A. thaliana* and 24 in *M. truncatula* to 38 in *S. japonica* genome (Supplementary Fig. S2D).

3.5. Whole-genome duplications

WGDs (polyploidization) have played a major role in the angiosperms' evolutionary history. The Ks value for collinear gene pairs indicated that a WGD has recently occurred in the evolution of *S. japonica*, and evidence of the pan-legume WGD event was also observed in its genome (Fig. 3A).

Intragenomic syntenic analysis of *S. japonica* revealed strong collinearity between large segments of different chromosomes (Fig. 3B). The widespread occurrence of one-to-one syntenic blocks confirms that WGD events have occurred in the evolutionary history of the *S. japonica* genome. Furthermore, we identified 2:2 *S. japonica*–*G. max* and 2:1 *S. japonica*–*M. truncatula* syntenic depth ratios (Fig. 3C). This clearly confirms that two WGD events have occurred in *S. japonica*'s evolution as there is documented evidence that *G. max* has undergone two WGD events[48] and *M. truncatula* has undergone one.[56] We further found that the most expanded genes which caused by WGD waves were enriched in 'response to water (GO: 0009415)', 'response to salt stress (GO: 0009651)', 'response to abiotic stimulus (GO: 0009628)' as well as 'regulation of biosynthetic process (GO: 0009889)' (Supplementary Tables S10 and S11).

3.6. Population genetic analyses

We generated short Illumina reads for 35 *S. japonica* individuals from 10 populations sampled in various parts of its range, with an average depth of ~18× (Fig. 4A and Supplementary Tables S4 and S12). After mapping to the reference *S. japonica* genome and quality control, we obtained 6,250,321 high-quality SNPs with the mapping ratio and coverage of 97.92% and 92.35%, respectively, on average (Supplementary Table S12). The genetic diversity (π) of *S. japonica* was calculated to be 0.0034 based on the high-quality SNPs (Supplementary Table S13). Using *S. tora* as an outgroup, we explored the phylogenetic relationships among the 35 accessions by examining whole-genome genetic variations. The neighbour-joining phylogenetic tree we obtained showed that the 35 accessions cluster in a single group (Fig. 4D). According to bar plots of assignment probabilities from ADMIXTURE analysis and cross-validation analysis of the 35 individuals, the best number of clusters (K) is 1 (lowest CV error = 0.65). PCA-based analysis of *S. japonica* populations also clustered the individuals into a single group (Fig. 4B). In summary, there is no obvious differentiation among the geographical groups of *S. japonica*.

3.7. Demographic history

The strong selection pressure likely exerted on *S. japonica* through its long domestication process is expected to have substantially affected the effective population size (N_e) of the existing genetic clusters. To address this issue, we estimated Ne using the PSMC method (Li and Durbin, 2011) (Fig. 4C). The results indicate that the ancestral N_e of *S. japonica* peaked ~2 Ma then continuously declined. It is obvious that during Naynayxungla glaciation period, *S. japonica* underwent a sharp decline.

4. Discussion

In this study, we report a high-quality, chromosome-level genome for *S. japonica* obtained by a combination of Nanopore sequencing and Hi-C technology. The genome survey results indicated that the heterozygosity of *S. japonica* genome was relatively high (1.19%) and similar to that of *Pyrus bretschneideri* (1.2%), *Quercus lobata* (1.25%), *Jasminum sambac* (1.5%) and *Castanopsis tibetana* (1.32%).[57–62] The final assembled genome of *S. japonica* is 511.49 Mb in length, consisting of 110 contigs with a contig N50 length of 17.34 Mb, which is much higher than that of *P. bretschneideri* (35.7 kb), *Q. lobata* (~18 kb) and *C. tibetana* (3.3 Mb), while comparable to that of *J. sambac* (17.5 Mb). Besides, it is also relatively high compared with other legumes, such as *V. unguiculata* (10.9 Mb),[63] *L. albus* (1.76 Mb)[64] and *V. radiata* (2.8 Mb).[65] Further Hi-C scaffolding placed 91.62% of the assembled genome on 14 pseudochromosomes (Lachesis Groups) ranging in size from 24.80 to 60.53 Mb with an improved N50 length of 31.32 Mb. The BUSCO assessment indicated that 98.1% of the conserved genes could be detected in the assembled genome. With the highly complete genome assembly, we obtain a high-quality gene set with better contiguity and annotation, which contains 31,000 protein coding genes, 456 miRNAs, 75 tRNAs, 1,488 rRNAs and 160 snRNAs. Phylogenomic analysis based on 204 single-copy orthologous genes revealed that *S. tora* and *S. japonica* were placed successively sister to all other Fabaceae species, and the two diverged at 74.77 Ma.

Our comparative genomics analyses revealed that *S. japonica* and other legume species shared the common legume WGD event[66,67] and experienced an independent WGD event which was first identified in our study (Fig. 2). WGD is recognized as a powerful enabler of plant genome expansion and evolution.[68] WGD (or paleopolyploidy) has been shown to contribute significantly to plant adaptation during global environmental changes in the evolutionary history of angiosperms.[69] For example, WTD events in Asteraceae, Lamiaceae and Apiaceae have contributed to the environmental adaptation during the Cretaceous period.[70–72] Here, using the chromosome-scale genome of

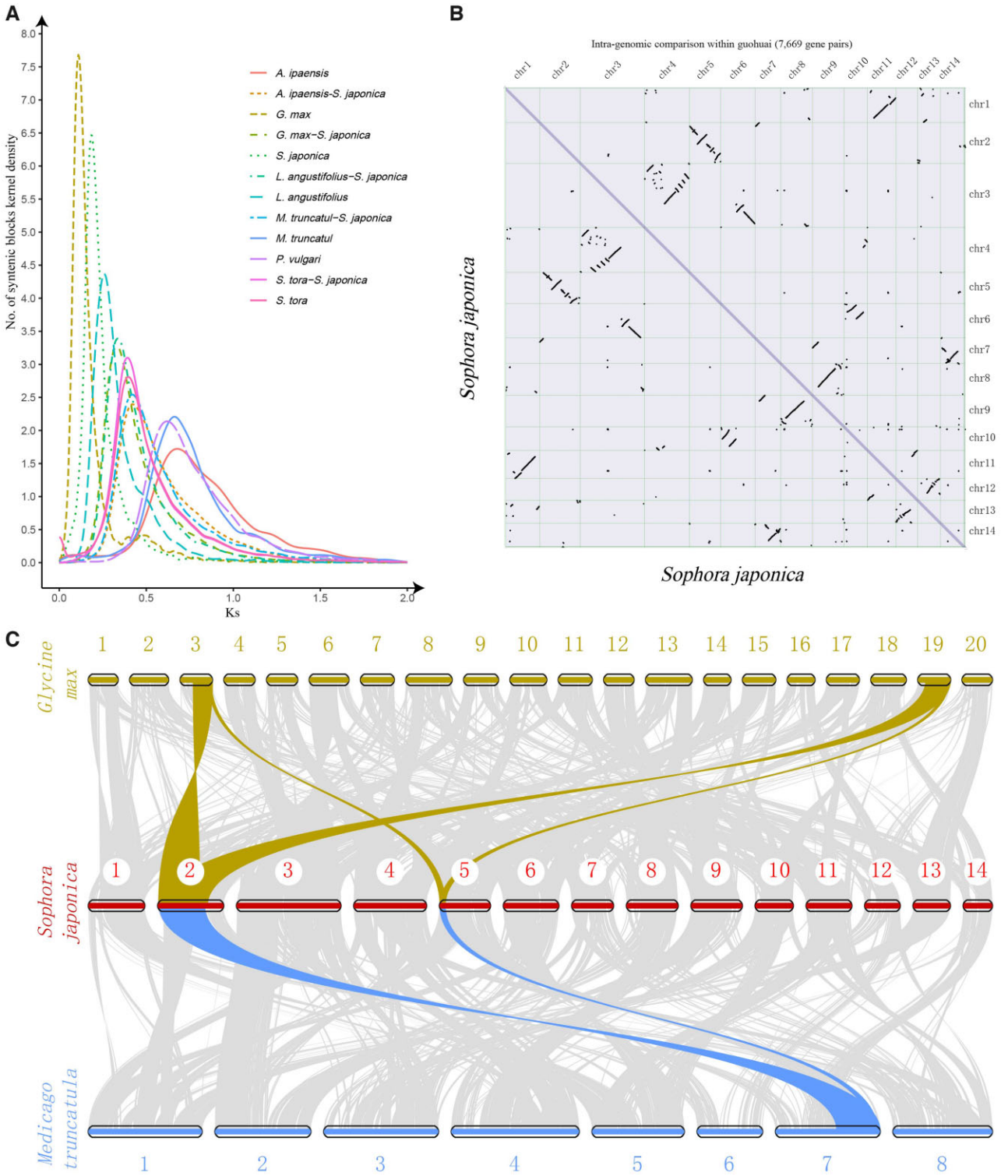

Figure 3. Results of comparative genomic analyses. (A) Ks distribution of syntenic blocks. (B) Dot plot of syntenic blocks identified by MCscanX in the *S. japonica* genome. (C) MCscanX identified synteny blocks (involving ≥5 collinear genes) between *S. japonica, G. max* and *M. truncatula*.

S. japonica, we first identified a species-specific WGD event, which was confirmed in *G. max* and *M. truncatula* genomes by Ks value profiles and collinearity analysis (Fig. 3). As *Glycine*-specific duplication is estimated to have occurred ~13 Ma,[48,66] and legume-common tetraploidy event at ~59 Ma,[66] we calculated that a species-specific WGD of *S. japonica* occurred ~19.42–29.32 Ma (Ks ~0.21) (Fig. 2D). Besides, we found that gene families related to oxidoreductase activity, cellular response, defence response, response, regulation, positive regulation, negative regulation and immune responses have all expanded, most of which are caused by WGD waves (Supplementary Tables S10 and S11

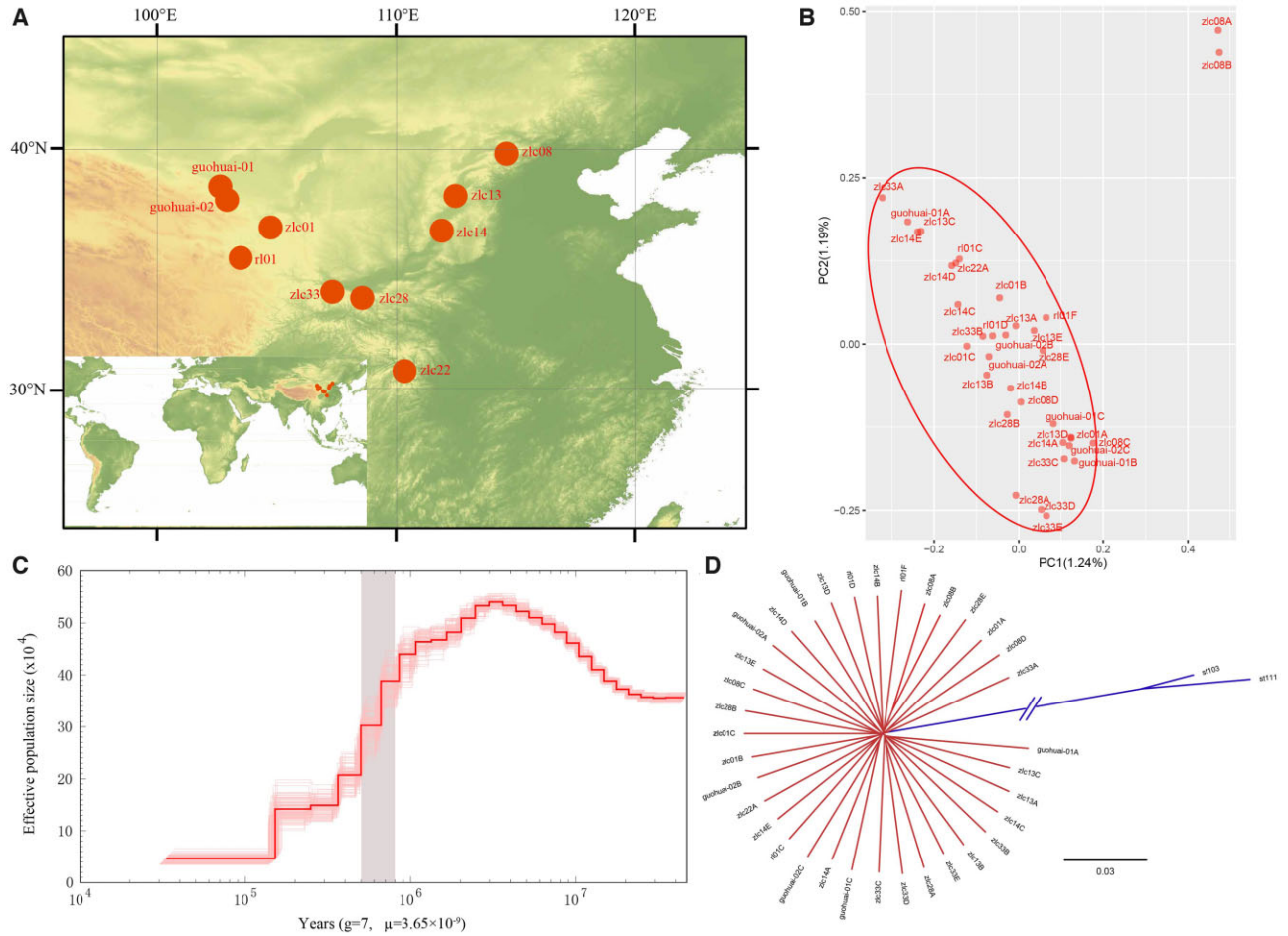

Figure 4. (A) Map showing current distribution of *S. japonica*. (B) Principal component analysis (PCA) plots showing scores of the first two principal components. (C) Demographic history of *S. japonica*, showing PSMC estimates of the species' effective population size (Ne). The time scale on the *x*-axis was calculated assuming a mutation rate per generation (μ) of $3.65\times e^{-9}$ and generation time (g) of 7 years. The time of the Naynayxungla glaciation is highlighted in grey vertical bars. (D) Neighbour joining tree obtained with *Senna tora* as the outgroup. Red and blue indicate species of *S. japonica* and *S. tora*, respectively.

and Fig. S2). We further examined members of TF families, such as *BBX* and *PAL*, which have been determined to be involved in abiotic stress responses, were increased in *S. japonica* (Supplementary Fig. S2C and D). We thus propose that WGD events may have increased *S. japonica*'s adaptability to diverse environmental factors and high resistance to pests and diseases.

Re-sequencing of 35 *S. japonica* individuals from surrounding areas of Gansu detected no obvious differentiation in samples from different geographic regions (Fig. 4B and D), so we speculate that the sampled *S. japonica* trees may have originated from the same nursery base. It is reasonable to infer that the *S. japonica* planted throughout the country may stem from very limited sources. To further study the species' genetic diversity, we calculated its π value and found that it is moderate compared with those of other species (Supplementary Table S13). However, this does not mean that the π value of *S. japonica* will not decrease in the near future. In particular, we notice that, the effective population size of *S. japonica* was continuously declined since 2 Ma (Fig. 4C). Thus, *S. japonica* from various sources or wild varieties should be planted to ensure that its genetic diversity does not decrease in the future.

In conclusion, we have constructed a chromosome-level reference genome for *S. japonica*. The genome is 511.49 Mb long, with a contig N50 size of 17.34 Mb and 98.1% coverage according to BUSCO analysis, indicating that the assembly has high completeness. We also detected evidence of two WGD events in the species' evolutionary history. The high-quality genome and whole genome resequencing data of *S. japonica* provide further foundations for comparative genomics and analyses of both the adaptation and evolution of dicotyledons.

Supplementary data

Supplementary data are available at *DNARES* online.

Funding

This study was supported by the National Natural Science Foundation of China (grant no. 32001085) and Fundamental Research Funds for Central Universities (grant no. lzujbky-2020-34, lzujbky-2020-ct02).

Conflict of interest

The authors have no conflict of interest to declare.

Data availability

Data acquired in this Whole Genome Shotgun project has been deposited in NCBI under project number PRJNA814452 and the BIG

Data Center (http://bigd.big.ac.cn) under project number PRJCA005733. The genome assembly file is available at NCBI. All the annotation tables containing results of the draft genome analysis are available at figshare doi.org/10.6084/m9.figshare.14790132.

References

1. Orwa, C., Mutua, A., Kindt, R., et al. *Agroforestree Database: A Tree Reference and Selection Guide, Version 4.0*. Nairobi, Kenya: World Agroforestry Centre, 2009.
2. Miao, M.S., Cheng, B.L. and Jiang, N. 2014, Effect of *Sophora japonica* total flavonoids on mouse models of hyperglycemia and diabetes model, *Appl. Mech. Mater.*, **664**, 397–401.
3. Chinese Pharmacopoeia Committee. *Chinese Pharmacopoeia, Part 1*. Beijing, China: China Medical Science Press, 2015.
4. He, X., Bai, Y., Zhao, Z., et al. 2016, Local and traditional uses, phytochemistry, and pharmacology of *Sophora japonica* L.: a review, *J. Ethnopharmacol.*, **187**, 160–82.
5. Ishida, H., Umino, T., Tsuji, K., et al. 1989, Studies on the antihemorrhagic substances in herbs classified as hemostatics in Chinese medicine. X. on hemostatic activities of the parched herbs for hemostatics, *Yakugaku Zasshi.*, **109**, 179–83.
6. Kim, J.M. and Yun-Choi, H.S. 2008, Anti-platelet effects of flavonoids and flavonoid-glycosides from *Sophora japonica*, *Arch. Pharm. Res.*, **31**, 886–90.
7. Kamal, K. 2010, Rutin natural bioflavonoid: traditional and medicinal uses, *Pharmacologyonline.*, **1**, 931–7.
8. Ganeshpurkar, A. and Saluja, A.K. 2017, The pharmacological potential of rutin, *Saudi Pharm. J.*, **25**, 149–64.
9. Zheng, X.P. and Zhang, Q.X. 2011, Status and prospects of urban landscape plants' application in Beijing, *Chinese Landsc. Architect.*, **5**, 81–5.
10. Xiao, Y., Xiao, Z., Ma, D., Liu, J., Li, J. 2019, Genome sequence of the barred knifejaw *Oplegnathus fasciatus* (Temminck & Schlegel, 1844): the first chromosome-level draft genome in the family Oplegnathidae, *GigaScience.*, **8**, giz013.
11. Hu, J., Fan, J., Sun, Z., Liu, S. 2020, NextPolish: a fast and efficient genome polishing tool for long-read assembly, *Bioinformatics*, **36**, 2253–5.
12. Zhang, D.C., Guo, L., Guo, H., et al. 2019, Chromosome-level genome assembly of golden pompano (*Trachinotus ovatus*) in the family Carangidae, *Sci. Data.*, **6**, 216.
13. Simão, F.A., Waterhouse, R.M., Ioannidis, P., et al. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–2.
14. Tarailo-Graovac, M. and Chen, N. 2009, Using RepeatMasker to identify repetitive elements in genomic sequences, *Curr. Protoc. Bioinformatics.*, **25**, 4.10.
15. Bao, W., Kojima, K.K. and Kohany, O. 2015, Repbase update, a database of repetitive elements in eukaryotic genomes, *Mob. DNA.*, **6**, 11.
16. Xu, Z. and Wang, H. 2007, LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons, *Nucleic Acids Res.*, **35**, W265–8.
17. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A. 2005, Rfam: annotating non-coding RNAs in complete genomes, *Nucleic Acids Res.*, **33**, D121–4.
18. Lowe, T.M. and Eddy, S.R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res.*, **25**, 955–64.
19. Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.-H., Rognes, T., Ussery, D.W. 2007, RNAmmer: consistent and rapid annotation of ribosomal RNA genes, *Nucleic Acids Res.*, **35**, 3100–8.
20. Birney, E., Clamp, M. and Durbin, R. 2004, GeneWise and genomewise, *Genome Res.*, **14**, 988–95.
21. Kim, D., Langmead, B. and Salzberg, S.L. 2015, HISAT: a fast spliced aligner with low memory requirements, *Nat. Methods.*, **12**, 357–60.
22. Haas, B.J., Papanicolaou, A., Yassour, M., et al. 2013, *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, *Nat. Protoc.*, **8**, 1494–512.
23. Haas, B.J., Delcher, A.L., Mount, S.M., et al. 2003, Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies, *Nucleic Acids Res.*, **31**, 5654–66.
24. Stanke, M., Steinkamp, R., Waack, S., Morgenstern, B. 2004, AUGUSTUS: a web server for gene finding in eukaryotes, *Nucleic Acids Res.*, **32**, W309–12.
25. Haas, B.J., Salzberg, S.L., Zhu, W., et al. 2008, Automated eukaryotic gene structure annotation using EVi-denceModeler and the program to assemble spliced alignments, *Genome Biol.*, **9**, R7.
26. Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
27. Bairoch, A. and Apweiler, R. 2000, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res.*, **28**, 45–8.
28. Blum, M., Chang, H., Chuguransky, S., et al. 2021, The InterPro protein families and domains database: 20 years on, *Nucleic Acids Res.*, **49**, D344–54.
29. Quevillon, E., Silventoinen, V., Pillai, S., et al. 2005, Interproscan: protein domains identifier, *Nucleic Acids Res.*, **33**, W116–20.
30. Ogata, H., Goto, S., Sato, K., et al. 1999, KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res.*, **27**, 29–34.
31. Emms, D.M. and Kelly, S. 2019, OrthoFinder: phylogenetic orthology inference for comparative genomics, *Genome Biol.*, **20**, 238.
32. Stamatakis, A. 2006, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, *Bioinformatics.*, **22**, 2688–90.
33. Yang, Z. 2007, PAML 4: phylogenetic analysis by maximum likelihood, *Mol Biol Evol.*, **24**, 1586–91.
34. Denoeud, F., Carretero-Paulet, L., Dereeper, A., et al. 2014, The coffee genome provides insight into the convergent evolution of caffeine biosynthesis, *Science.*, **345**, 1181–4.
35. De Bie, T., Cristianini, N., Demuth, J.P., Hahn, M.W. 2006, CAFE: a computational tool for the study of gene family evolution, *Bioinformatics.*, **22**, 1269–71.
36. Chen, C., Chen, H., Zhang, Y., et al. 2020, TBtools: an intergrative toolkit developed for interactive analyses of big biological data, *Molecular Plant.*, **13**, 1194–202.
37. Wang, Y.P., Tang, H., DeBarry, J.D., et al. 2012, MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity, *Nucleic Acids Res.*, **40**, e49–14.
38. Li, H. and Durbin, R. 2010, Fast and accurate long-read alignment with Burrows-Wheeler transform, *Bioinformatics*, **26**, 589–95.
39. Li, H., Handsaker, B., Wysoker, A., et al.; 1000 Genome Project Data Processing Subgroup. 2009, The sequence alignment/map format and SAMtools, *Bioinformatics*, **25**, 2078–9.
40. Vilella, A.J., Severin, J., Ureta-Vidal, A., et al. 2009, EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates, *Genome Res.*, **19**, 327–35.
41. Alexander, D.H., Novembre, J. and Lange, K. 2009, Fast model-based estimation of ancestry in unrelated individuals, *Genome Res.*, **19**, 1655–64.
42. Purcell, S., Neale, B., Todd-Brown, K., et al. 2007, PLINK: a tool set for whole-genome association and population-based linkage analyses, *Am. J. Hum. Genet.*, **81**, 559–75.
43. Li, H. and Durbin, R. 2011, Inference of human population history from individual whole-genome sequences, *Nature*, **475**, 493–6.
44. Varshney, R.K., Song, C., Saxena, R.K., et al. 2013, Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement, *Nat. Biotechnol.*, **31**, 240–6.
45. Li, H., Jiang, F., Wu, P., Wang, K., Cao, Y. 2020, A high-quality genome sequence of model legume *Lotus japonicus* (MG-20) provides insights into the evolution of root nodule symbiosis, *Gene.*, **11**, 483.

46. Jaillon, O., Aury, J., Noel, B., et al. 2007, The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, *Nature*, **449**, 463–7.
47. Chan, A.P., Crabtree, J., Zhao, Q., et al. 2010, Draft genome sequence of the oilseed species *Ricinus communis*, *Nat. Biotechnol.*, **28**, 951–6.
48. Schmutz, J., Cannon, S.B., Schlueter, J., et al. 2010, Genome sequence of the palaeopolyploid soybean, *Nature*, **463**, 178–83.
49. Paterson, A.H., Bowers, J.E., Bruggmann, R., et al. 2009, The Sorghum bicolor genome and the diversification of grasses, *Nature*, **457**, 551–6.
50. Schnable, P.S., Ware, D., Fulton, R.S., et al. 2009, The B73 maize genome: complexity, diversity, and dynamics, *Science*, **326**, 1112–5.
51. Olsen, K.M., Lea, U.S., Slimestad, R., et al. 2008, Differential expression of four Arabidopsis PAL genes; PAL1 and PAL2 have functional specialization in abiotic environmental-triggered flavonoid synthesis, *J. Plant Physiol.*, **165**, 1491–9.
52. Crocco, C.D., Ocampo, G.G., Ploschuk, E.L., Mantese, A., Botto, J.F. 2018, Heterologous expression of AtBBX21 enhances the rate of photosynthesis and alleviates photoinhibition in *Solanum tuberosum*, *Plant Physiol.*, **177**, 369–80.
53. Laubinger, S., Marchal, V., Gentilhomme, J., et al. 2006, Arabidopsis SPA proteins regulate photoperiodic flowering and interact with the floral inducer CONSTANS to regulate its stability, *Development*, **133**, 3213–22.
54. Wang, Q., Tu, X., Zhang, J., Chen, X., Rao, L. 2013, Heat stress-induced BBX18 negatively regulates the thermotolerance in Arabidopsis, *Mol. Biol. Rep.*, **40**, 2679–88.
55. Liu, X., Li, R., Dai, Y., Chen, X., Wang, X. 2018, Genome-wide identification and expression analysis of the B-box gene family in the Apple (*Malus domestica* Borkh.) genome. *Mol. Genet. Genomics*, **293**, 303–15.
56. Young, N.D., Debellé, F., Oldroyd, G.E.D., et al. 2011, The Medicago genome provides insight into the evolution of rhizobial symbioses, *Nature*, **480**, 520–4.
57. Vurture, G.W., Sedlazeck, F.J., Nattestad, M., et al. 2017, GenomeScope: fast reference-free genome profiling from short reads, *Bioinformatics*, **33**, 2202–4.
58. Wu, J., Wang, Z., Shi, Z., et al. 2013, The genome of the pear *Pyrus bretschneideri* Rehd, *Genome Res.*, **23**, 396–408.
59. Ramos, A.M., Usié, A., Barbosa, P., et al. 2018, Data descriptor: the draft genome sequence of cork oak, *Sci. Data*, **5**, 180069.
60. Xu, S., Ding, Y., Sun, J., et al. 2022, A high-quality genome assembly of *Jasminum sambac* provides insight into floral trait formation and Oleaceae genome evolution, *Mol. Ecol. Resour.*, **22**, 724–39.
61. Sork, V.L., Fitz-Gibbon, S.T., Puin, D., et al. 2016, First draft assembly and annotation of the genome of a California endemic oak *Quercus lobate* Nee (Fagaceae), *G3 (Bethesda)*, **6**, 3485–95.
62. Sun, Y., Guo, J., Zeng, X., et al. 2021, Chromosome-scale genome assembly of *Castanopsis tibetana* provides a powerful comparative framework to study the evolution and adaptation of Fagaceae trees, *Mol. Ecol. Resour.*, **22**, 1178–89.
63. Lonardi, S., Munoz-Amatriain, M., Liang, Q., et al. 2019, The genome of cowpea (*Vigna unguiculata* [L.] Walp.), *Plant J.*, **98**, 767–82.
64. Xu, W., Zhang, Q., Yuan, W., et al. 2020, The genome evolution and low-phosphorus adaptation in white lupin, *Nat. Commun.*, **11**, 1069.
65. Ha, J., Satyawan, D., Jeong, H., et al. 2021, A near-complete genome sequence of mungbean (*Vigna radiata* L.) provides key insights into the modern breeding program, *Plant Genome*, **14**, e20121.
66. Wang, J., Sun, P., Li, Y., et al. 2017, Hierarchically aligning 10 legume genomes establishes a family-level genomics platform, *Plant Physiol.*, **174**, 284–300.
67. Zhao, Y., Zhang, R., Jiang, K., et al. 2020, Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogenfixing symbiosis in Fabaceae, *Mol. Plant.*, **14**, 1–26.
68. Jackson, S. and Chen, Z.J. 2010, Genomic and expression plasticity of polyploidy, *Curr. Opin. Plant Biol.*, **13**, 153–9.
69. Wu, S., Han, B. and Jiao, Y. 2020, Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms, *Mol. Plant*, **13**, 59–71.
70. Badouin, H., Gouzy, J., Grassa, C.J., et al. 2017, The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution, *Nature*, **546**, 148–52.
71. D'Hont, A., Denoeud, F., Aury, J.-M., et al. 2012, The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants, *Nature*, **488**, 213–7.
72. Iorizzo, M., Ellison, S., Senalik, D., et al. 2016, A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution, *Nat. Genet.*, **48**, 657–66.