Chromosome-level reference genome assembly for the protected resource plant, *Zenia insignis*

Authors

Si-Yun Chen[1,2,4], Zhi-Yun Yang[1,4], Rong Zhang[1,2,3], Hui Liu[1,2,3], Ting-Shuang Yi[1,2]

Affiliations

1. Germplasm Bank of Wild Species, Yunnan Key Laboratory of Crop Wild Relatives Omics, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, Yunnan, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
3. Key Laboratory of Plant Diversity and Specialty Crops, Chinese Academy of Sciences, Beijing 100093, China
4. these authors contributed equally: Si-Yun Chen and Zhi-Yun Yang.

corresponding author(s): Hui Liu (liuhui1@mail.kib.ac.cn), Ting-Shuang Yi (tingshuangyi@mail.kib.ac.cn)

Abstract

Zenia insignis Chun, from the subfamily Dialioideae of Fabaceae, is a tree species of significant economic and ecological values. It is a threatened species as per the IUCN Red List. The absence of a reference genome has impeded research on *Z. insignis* despite its importance. In this study, we generated a reference genome for *Z. insignis*, with a contig N50 of 6.02Mb, a total length of 361.11 Mb, and 97.71% of the sequences assembled into 14 pseudo-chromosomes. The BUSCO assessment score for completeness is 97.30%, and the LAI index assessment score for continuity is 14.57. To our knowledge, this is the first reported nuclear genome of a species belonging to subfamily Dialioideae. The reference genome will provide a valuable resource for the phylogenomic studies of the Fabaceae family and facilitate further research on *Z. insignis*.

Background & Summary

Zenia insignis Chun, a member of the family Fabaceae, is a tree species recognized as one of the Grade-II Wild Plants of National Priority Protection in China[1] and is listed as Near Threatened (NT) species on the International Union for Conservation of Nature (IUCN) Red List[2]. This species was named by the renowned botanist Woon-Young Chun in honor of Hongjun Ren (H.C. Zen)[3]. The species is primarily distributed in southern and southwestern China at an altitude range of 200–1000 meters and northern Vietnam[4]. *Zenia insignis* boasts a robust root system and strongly resists drought and infertile conditions[5]. In the karst regions of southern and southwestern China, it has been utilized as an outstanding tree species for afforestation on rocky terrain and as a pioneering species for the control of rocky desertification[5-7]. Additionally, it has significant economic value. The wood of *Z. insignis* is

versatile and suitable for producing doors, boards, agricultural implements, and furniture[8,9]. The plant's tender leaves serve as livestock feed and green manure in rice fields[7-9].

Research on *Z. insignis* has predominantly concentrated on seedling[10,11], reforestation technologies[5,6], and phylogenetic studies[12-19]. Though partial and complete plastome sequences[12-14,16] and nuclear sequences[15,17-19] have been utilized in constructing the phylogenetic tree of the Fabaceae family, the deep phylogenetic relationships within Fabaceae remain to be fully elucidated. A chromosome-level genome sequence of *Z. insignis* would be instrumental in clarifying these phylogenetic relationships. In this study, we have successfully obtained the reference genome of *Z. insignis* and performed comparative analyses with other representative plant genomes to further our understanding of the evolutionary history of the Fabaceae.

In this study, we integrated multiple sequencing data types to assemble and annotate the genome of *Z. insignis*. Specifically, we used PacBio continuous long reads (CLR) (174.14 Gb, 363.43× coverage), next-generation sequencing (NGS) reads (44.14 Gb, 91.96× coverage), Hi-C data (59.63 Gb, 124.45× coverage), and full-length transcriptome reads. The resulting reference genome consists of 14 pseudo-chromosomes and spans a total length of 352.84 Mb with a contig N50 of 6.02 Mb. The genome's LAI index score is 14.57. The GC content of the genome is 34.38%, and the proportion of repeat sequences is 35.47%, with long terminal repeat (LTR) elements being the most prevalent, accounting for 9.03%. We also obtained the chloroplast genome sequence of *Z. insignis* with a length of 159,390 bp.

We predicted 33,322 gene models, of which 90.46% (30,143) were functionally annotated. Analysis of orthologous genes between *Z. insignis* and ten other representative plant species revealed 608 orthogroups, with at least 72.72% of species having single-copy genes in any given orthogroup. Additionally, 1,625 gene families underwent significant expansion, and 741 gene families underwent significant contraction in *Z. insignis*. In the whole genome duplication analysis, two collinear block *Ks* peaks were observed at 0.245±0.001 and 1.693±0.001, suggesting that *Z. insignis* has undergone an additional whole genome duplication (WGD) event after the γ WGT event shared by all core eudicots.

Methods

Genome size estimation and sequencing

Fresh young leaves of *Z. insignis* were collected from the Kunming Botanical Garden (KBG), Kunming Institute of Botany, Chinese Academy of Sciences (KIB), Yunnan Province, China. The voucher specimen (Collection Number: Yi20005) had been deposited at the Herbaria (KUN) of Kunming Institute of Botany, Chinese Academy of Sciences, Yunnan Province, China. The leaves were immediately flash-frozen in liquid nitrogen to preserve their cellular structure and molecular content. Subsequently, the samples were transferred to an ultra-low temperature freezer maintained at –80°C for storage.

DNA/RNA extraction, library construction, and sequencing processes were provided by Wuhan Frasergen Bioinformatics Co., Ltd. The major steps are summarized below:

1. Genomic DNA was extracted from the tender leaves of *Z. insignis* using the modified cetyltrimethylammonium bromide (CTAB) method[20,21]. Total RNA was isolated using a Trizol reagent (Invitrogen, CA, USA). The quantity and quality of the DNA/RNA were determined using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA).

2. The quality-compliant DNA was further tested for purity and integrity using 0.8% agarose gel electrophoresis and a Qubit dsDNA HS Assay Kit on a Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). RNA contamination was identified using a 1.5% agarose gel, and purity and integrity were evaluated using a NanoDrop 2000 and a Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

3. According to the manufacturer's instructions, a short-read library was constructed using the VAHTS Universal DNA Library Prep Kit (Vazyme, Nanjing, China) for MGI. The library's quantification and sizing were measured using a Qubit 3.0 Fluorometer and a Bioanalyzer 2100 system (Agilent Technologies, CA, USA). Paired-end sequencing of the short-read library was conducted on an MGI-SEQ 2000 platform (BGI, China).

4. To perform long-read sequencing, the library was prepared in accordance with the PacBio single-molecule real-time (SMRT) protocol and then sequenced on a Pacific Biosciences Sequel II System in CLR mode.

5. The construction of the Hi-C library commenced with the cross-linking of fresh leaves of *Z. insignis* under vacuum infiltration with 3% formaldehyde. Following lysis, digestion, labelling, ligation, purification, adaptor addition, and PCR amplification, the Hi-C libraries were quantified and sequenced on the MGI-SEQ 2000 platform.

6. For short-read transcriptome sequencing, RNA was isolated from Z. insignis young leaves using the QIAGEN kit according to the manufacturer's instructions. The short-read library was subsequently constructed and sequenced on the MGI-SEQ 2000 platform using paired-end sequencing.

7. The full-length transcriptome sequencing was conducted using the Pacific Biosciences DNA Template Prep Kit 2.0 for library construction (Pacific Biosciences, CA, USA). Subsequently, the SMRT sequencing was performed on a Pacific Biosciences Sequel II platform.

Our study generated 44.14 Gb of raw data for the whole genome, 174.14 Gb of PacBio CLR data, 59.63 Gb of Hi-C data, and 81.92 Mb of full-length transcriptome data (Table 1).

The genome size of *Z. insignis* was assessed using flow cytometry analysis conducted on a BD FACScalibur flow cytometer (BD Biosciences). The test sample preparation was conducted in accordance with the manufacturer's guidelines and aligned with established

methodologies from previous literature[22,23]. As a standard reference, we utilized *Zea mays* (maize)[24], a species with a well-documented genome size of 2.3 Gb. The analysis yielded a result indicating that the genome size of *Z. insignis* is approximately 480 ± 4.8 Mb (Figure 1A).

A *K-mer* analysis was conducted on *Z. insignis* to estimate its genome size. Next-generation sequencing (NGS) data was obtained from Frasergen and processed Fastp v0.19.3[25]. This involved the removal of adapter sequences, the filtration of reads deemed to be of insufficient length, and the elimination of those with low quality. Subsequently, Jellyfish v2.2.10[26] was employed to calculate the frequency distribution of the cleaned data using 21-*mers*. Subsequently, GenomeScope v2.0[27] was employed to estimate the fundamental characteristics of the genome. The genome size of *Z. insignis* was estimated to be approximately 364.94 Mb, with a heterozygosity rate of 0.842% and a repeat content of 1.17% (Figure 1B).

De novo genome assembly

Figure 2 illustrates the methodology employed for the chromosome-level genome assembly of *Z. insignis*. The objective was to utilize three distinct assemblers, compare their resulting assemblies, and achieve a high-quality assembly. The assemblers employed were Falcon v0.3.0[28], Canu v2.1.1[29], and Flye v2.9[30]. The following section provides a summary of the assembly process.

1. The Falcon software was employed to assemble the CLR long reads. The resulting contigs were then polished using long reads with GCpp v2.0.0[31]. Subsequently, the processes above yielded the assembly contig, designated as Assembly_Contig1. Subsequently, Assembly_Contig1 was subjected to further refinement with short reads, employed NextPolish v1.31[32]. The output from the process was designated as the 'Falcon contig'. To eliminate redundancy, the Falcon contig was processed with Purge_Dups v1.2.5[33], and the resulting non-redundant contigs were designated as the 'Falcon_purge contigs'.

2. The Canu software was employed to assemble the CLR long reads, and the result was subsequently polished with long reads by GCpp. Through these processes, we obtained Assembly_Contig2. Assembly_Contig2 was then polished with short reads using NextPolish, with the output from this step designated as the 'Canu_contig'. The Canu_contig was subsequently subjected to further processing to eliminate redundancy, with the result that the non-redundant contigs were designated as the 'Canu_purged contigs'. Purge_Dups also employed the process of eliminating redundancy.

3. The third assembler employed was Flye, which accepted the error-corrected reads generated by Canu as input. The output from Flye was polished with long reads by GCpp, resulting in the generation of Assembly_Contig3. Subsequently, the contig was polished with short reads using NextPolish, resulting in the 'Flye_contig'. The Flye_contig was subjected to

149 a process of redundancy elimination using Purge_Dups, resulting in the generation of a set
150 of non-redundant contigs designated as the 'Flye_purge contigs'.

151 4. Six distinct results were obtained after completing the three steps above. The
152 following contigs were identified: Falcon contig, Canu contig, Flye_contig, Falcon_purge
153 contig, Canu_purge contig, and Flye_purge contig.

154 BUSCO v5.4.5[34] with 1614 embryophyta single-copy orthologs was employed to assess
155 the completeness of the assembly, the N50[35] metric provided insights into the distributions
156 of contigs, and the LTR Assembly Index (LAI) v2.9.0[36] was employed to evaluate the
157 contiguity of the assembled genome. The results are presented in Tables 2–4. Notably, the
158 Canu_purged assembly exhibited the highest N50 value, the Flye_purged assembly achieved
159 the optimal BUSCO score, and the Falcon_purged contig assembly exhibited the highest LAI
160 score. In other words, the result generated by Purge_Dups exhibited a better contiguity.
161 Manual inspection using Hi-C reads revealed a low redundancy in Falcon_purge assembly
162 (Supplementary Figure 1). Consequently, the results generated by Purge_Dups were
163 selected for further comparison. A comparison of the Falcon_purge assembly with other
164 results generated by Purge_Dups revealed that it exhibited high contiguity with N50 of 6.02
165 Mb, BUSCO completeness of 97.30%, and LAI score of 14.57.

166 During the chromosome assembly phase, we integrated Hi-C data with the previously
167 assembled contigs to achieve chromosome-level scaffolding. Specifically, Juicer was
168 employed to align the Hi-C reads to the contig genome. Subsequently, the 3D-DNA pipeline[37]
169 was used to map the contigs onto chromosome-level scaffolds. For a visual representation of
170 the Hi-C data, Juicebox v1.22[37] was employed. The results of the chromosome assembly also
171 demonstrated that the Purge_Dups results exhibited lower redundancy, compared to the
172 rest two assemblies (Supplementary Figure 1). The Purge_Dups result was chosen as the
173 optimal choice, and subsequently, the heatmap of the result was manually adjusted (Figure
174 3A). The final outcome of this process was the generation of 14 pseudo-chromosomes with a
175 total length of 352.84 Mb, as illustrated in Figures 3A and 3B.

176 The GetOrganelle v1.7.3.3[38] software was employed to assemble the chloroplast genome,
177 successfully acquiring the complete chloroplast genome sequence depicted in Figure 3C. The
178 newly assembled chloroplast genome spans 159,390 base pairs (bp) and is annotated using
179 the PGA tool[39], as shown in Figure 3C, drawn by OGDRAW v1.3.1[40]. For comparative analysis
180 of the chloroplast genome, we employed the MAFFT plugin within the Geneious Prime
181 software suite[41] to align our sequence with a published chloroplast genome of *Z. insignis*[16]
182 (NC_045299.1) obtained from the National Center for Biotechnology Information(NCBI). As
183 illustrated in Figure 3D, the alignment result revealed that apart from some single nucleotide
184 polymorphisms (SNPs) in non-coding regions, there were no significant differences in large-
185 scale sequences.

186 **Repeat annotation**

187 The RepeatModeler v2.0.1[42] software was employed to predict the occurrence of repeat
188 sequences within the genome of *Z. insignis*. RepeatModeler integrates three *de novo* repeat-
189 finding programs: RECON v1.08[43], RepeatScout v1.0.6[44], and LTRHarvest v5.9[45]/LTR_retriever
190 v2.9.0[46]. For our analysis, we employed rmblast v2.13.0[47] as the search engine.
191 Subsequently, the output from RepeatModeler was processed using RepeatMasker v4.1.4[48]
192 to identify and mask the repetitive elements within the *Z. insignis* genome.

193 Table 5 presents the results of the repeat sequence analysis for *Z. insignis*. A total of
194 128.08 Mb was identified as repetitive sequences, constituting 35.47% of the pseudo-
195 chromosome sequences. The analysis revealed four primary classes of transposable
196 elements within the genome: 9.03% for long terminal repeats (LTRs), 3.08% for DNA
197 elements, 1.81% for long interspersed nuclear elements (LINEs), and 0.08% for short
198 interspersed nuclear elements (SINEs).

199 **Gene prediction and functional annotation**

200 A hybrid approach was implemented to achieve high-quality gene prediction, incorporating
201 *de novo* prediction, homology-based prediction, and RNA-Seq-assisted prediction. We
202 employed a similar pipeline to the AnnoSmk v1.0[49] pipeline for gene prediction. Firstly, the
203 RepeatMasker v4.1.4[48] software was employed to generate repeat evidence. For homology-
204 based prediction, protein data for *Arabidopsis thaliana*[50] was sourced from The Arabidopsis
205 Information Resource (TAIR) and employed to construct homologous gene models using
206 genBlastG v1.38[51] as protein evidence. For transcriptome evidence, HISAT2 v2.2.1[52] was
207 employed to map paired-end short transcriptome reads to the *Z. insignis* genome. Then, the
208 mapped reads were assembled using Trinity v2.11.0[53]. Concurrently, we processed full-
209 length transcriptome data with the IsoSeq3 v3.4.0[54] pipeline to produce high-quality
210 consensus transcript sequences. Minimap2 v2.22[55] was employed to align both short and
211 long-read data, from which the transcriptome evidence was derived. The integrated repeat,
212 protein, and transcript evidence were then subjected to gene prediction using MAKER2
213 v3.01.03[56]. Three rounds of MAKER2 analysis were executed. In the second and third rounds,
214 the output generated through *de novo* gene prediction using AUGUSTUS v3.4.0[57] was also
215 included. BUSCO was employed to evaluate the gene prediction results after each round of
216 MAKER2, yielding scores of 46.60%, 96.60%, and 96.80%, respectively (Table 6). Following
217 the third round of MAKER2, the results were input into the PASA pipeline v2.5.0[58], resulting
218 in the final output. The final BUSCO score for the AnnoSmk pipeline was 97.20% (Table 6). In
219 conclusion, a total of 33,322 gene models were predicted.

220 The functional annotation of protein-coding genes was conducted using the eggNOG-
221 mapper website[59], which employs the eggNOG-mapper v2.1.12[60] and the eggNOG 5
222 database. This tool utilizes several comprehensive databases, including Pfam[61], InterPro[62],
223 Gene Ontology (GO)[63], and the Kyoto Encyclopedia of Genes and Genomes (KEGG)[64], to
224 annotate protein sequences. This analysis resulted in the successful functional annotation of
225 30,143 protein-coding genes, representing 90.46% of the total gene models.

226 **Comparative genomics and phylogenetic analysis**

227 The identification of orthologous genes represents a central objective in the fields of
228 comparative genomics and phylogenetics. In this study, we compared the protein sequences
229 of *Z. insignis* with those of ten other representative plant species, including *Arabidopsis*
230 *thaliana*[50](GCA_000001735.2), *Cercis chinensis*[65](PRJCA007288), *Eriobotrya*
231 *japonica*[66](GWHAAZU00000000.1), *Glycine max*[67](GCA_000004515.5), *Lysidice*
232 *rhodostegia*[68], *Medicago truncatula*[69](GCA_003473485.2), *Senna tora*[70](GCA_014851425.1),
233 *Quercus mongolica*[71](GCA_011696235.1), *Vitis vinifera*[72](GCA_030704535.1), and *Polygala*
234 *tenuifolia*[73](TCMPG20196). These were employed to discern orthologs and paralogs. The
235 OrthoFinder v2.5.4[74] software was used to predict orthologous and paralogous gene
236 families. The protein sequences from the eleven species were input into OrthoFinder,
237 resulting in the clustering of 39,196 genes into 30,977 gene families. The analysis revealed
238 8,698 orthogroups that are shared by all species. Of these, 40 orthogroups comprise single-
239 copy genes, while 608 orthogroups have at least 72.72% of species exhibiting single-copy
240 genes within each orthogroup.

241 We used the 608 orthogroups obtained from our analysis to construct phylogenetic trees.
242 Initially, the seqkit tool v0.16.1[75] was employed to extract the corresponding nucleic acid
243 sequences. Subsequently, the DNA sequences were aligned using MUSCLE v3.81511[76], and
244 the sequences were concatenated using seqkit to generate a supermatrix of genes. The
245 supermatrix was subsequently employed as the input for constructing the phylogenetic tree,
246 which was generated using RAxML v8.2.12[77] with the maximum likelihood method. The most
247 appropriate model for RAxML, GTR+F+R3, was determined using IQTree v2.1.4-beta[78].
248 Subsequently, the divergence times at the nodes of the phylogenetic tree using R8s v1.81,
249 with calibration time points derived from the TimeTree database[79]. The calibration points
250 included the divergence times between *M. truncatula* and *G. max* (55 million years ago,
251 MYA), *M. truncatula* and *L. rhodostegia* (68 MYA), and *G. max* and *L. rhodostegia* (68 MYA).
252 Finally, the CAFE v5.0[80] program was employed to understand the expansion and contraction
253 of gene families across the phylogenetic tree. The results indicate that *Z. insignis* (subfamily:
254 Dialioideae) diverged from *S. tora* (subfamily: Caesalpinioideae) and *M. truncatula*
255 (subfamily: Papilionoideae) around 51 MYA. The analysis of gene family dynamics identified
256 1,625 gene families with significant expansion (LRT test, $p \leq 0.05$) and 741 gene families with
257 significant contraction (LRT test, $p \leq 0.05$) in the genome of *Z. insignis* (Figure 4A).

258 We conducted GO and KEGG enrichment analyses to understand the biological
259 significance of the gene families that exhibited significant expansion and contraction with *a*
260 *P- value* cutoff of 0.05. The GO and KEGG analyses were performed using the eggNOG-
261 mapper web platform[59]. We employed TBtools v2.086[81] to carry out the enrichment analysis
262 of the identified gene families. The GO enrichment analysis revealed that the families that
263 exhibited significant expansion were predominantly involved in molecular function and
264 biological process (Supplementary Figure 2A). The terms of the molecular function are

mainly 'metal ion transmembrane transporter activity', 'inorganic cation transmembrane transporter activity', and 'cellulose synthase activity' (Supplementary Figure 2A). And the terms of the biological process are mainly 'hydrocarbon metabolic process' and 'olefin metabolic process' (Supplementary Figure 2A). Furthermore, the KEGG enrichment analysis revealed that these families are significantly enriched in pathways such as 'alpha-Linolenic acid metabolism', 'MAPK signaling pathway-plant', 'metabolism of terpenoids and polyketides', 'biosynthesis of other secondary metabolites', and 'ion channels' (Supplementary Figure 2C). Conversely, for the families that exhibited a significant contraction, the GO enrichment analysis was predominantly involved in the biological process of 'embryo development ending in seed dormancy', 'embryo development', 'seed development', 'fruit development', and 'proteolysis involved in protein catabolic process' (Supplementary Figure 2B). The corresponding KEGG enrichment analysis identified pathways including 'protein families: metabolism', 'ubiquitin mediated proteolysis', 'chromosome and associated proteins', 'protein processing in endoplasmic reticulum', and 'brite hierarchies' (Supplementary Figure 2D).

Identification of whole genome duplication events

To elucidate the whole genome duplication (WGD) events in *Z. insignis*, a comprehensive analysis was conducted using the WGDI toolkit v0.6.5[82] with various parameters. We used *V. vinifera*[72] as a reference genome. First, we used BLAST+ v2.15.0[83] to identify homologous genes (e-value 10^{-5}). We then drew a dot plot of collinear genes with the parameter -d (Figure 5A and B). Next, we conducted a collinear analysis with the parameter -icl, calculated *Ks* values with the parameter -ks, combined the results of the collinear analysis and *Ks* values with the parameter -bi, and fitted *Ks* peaks with the parameter -pf. Finally, we plotted a *Ks* dot plot with the parameter -bk (Figure 5C and D) and a *Ks* peak plot with the parameter -kf (Figure 4B). The dot-plot and *Ks* dot-plot results for *Z. insignis* showed a high degree of similarity to those shown in Figure 4A and C. They were cross-validated with the results from *V. vinifera* in Figure 5B and D. The *Ks* peak analysis revealed two peaks at 0.245±0.001 and 1.693±0.001, suggesting that *Z. insignis* likely experienced a lineage-specific WGD event, as illustrated in Figure 4B.

Technical Validation

As part of our sequencing pipeline, we conducted a rigorous sample quality assessment for both DNA and RNA samples used in the study. The specific procedures and criteria for this assessment are detailed in our paper's 'Methods' section, ensuring that only high-quality samples were selected for subsequent sequencing and analysis.

In our study, we employed two distinct methodologies to estimate the genome size of *Z. insignis*. The first method involved Flow cytometry analysis, which provided a quick and accurate measurement of the relative DNA content within the cells of *Z. insignis*. The second method was *K-mer* analysis, which leverages the frequency distribution of DNA sequences to predict genome size. The details have been mentioned in the 'Method' section. Both

approaches were crucial in validating the genome size and ensuring the reliability of our genomic data.

In our assessment of the *de novo* genome assembly for *Z. insignis*, we utilized a suite of metrics, including N50, BUSCO, and LAI, to evaluate the performance of various assembly software tools. The results of these evaluations are presented in Tables 2, 3, and 4, which compare the assembly quality, completeness, and continuity across different assemblers. Based on these assessments, Falcon-purged output demonstrated the most favorable balance of completeness and continuity, making it the choice for subsequent downstream analyses. In summary, we have successfully generated a high-quality reference genome for *Z. insignis*, which may promote further biological and comparative genomic studies of this species.

Code Availability

We used all software or pipelines, depending on the tools' manuals or protocols. The software's version is described in the methods section, and the parameters are given if not default.

Acknowledgements

This research was supported by the Major Program for Basic Research Project of Yunnan Province (202401BC070001), the Youth Talents Special Project of Yunnan Province 'Xingdian Talents Support Program' (XDYC-QNRC-2022-0423), the National Natural Science Foundation of China, key international (regional) cooperative research project (No. 31720103903), the Yunnan Provincial Basic Research Program Youth Project (2019FD058), and the National Natural Science Foundation of China (No. 32270247). We are grateful to the Germplasm Bank of Wild Species, particularly the Molecular Biology Experiment Center and the iFlora High Performance Computing Center, the Kunming Botanical Garden, Ms. Ding-Jie Wang, and Mr. Zuo-Ying Xiahou.

Author contributions

T.-S. Y. and H. L. Designed the experiments. S.-Y. C. and Z.-Y. Y. performed the experiments. S.-Y. C., H. L., and R. Z. analyzed the data. S.-Y. C., T.-S. Y. and H. L. wrote the paper. All Authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

References

1 The State Council. *List of Wild Plants of National Priority Protection(First batch)*, <https://www.gov.cn/gongbao/content/2000/content_60072.htm> (2000).

2 IUCN. *The IUCN Red List of Threatened Species*, <http://www.iucnredlist.org/> (2024).

3 Chun, W.-Y. A new genus in the Chinese flora. *Sunyatsenia* **6**, 196–198 (1946).

4 Xu, L.-R. *et al. Fabaceae(Leguminosae)* In *Flora of China* Vol. 10. Wu, Z.-Y., Raven, P. H., & Hong, D.-Y. Eds 27 Science Press & Missouri Botanical Garden Press (2010).

5 Ao, H., He, D., Zhang, Z., Xia, H. & Yang, W. *Zenia insignis* communities in limestone areas of Guangdong. *Tropical Geography* **17**, 275–276+278–282, http://dx.chinadoi.cn/10.13284/j.cnki.rddl.000227 (1997).

6 Zhang, Z.-P., He, D.-Q., Ao, H.-X., Yang, W.-J. & Wu, H.-M. The characteristic of physiological ecology of the main tree species used for afforestation at limestone regions in North Guangdong. *Acta Phytoecol. Sin.* **17**, 133–142 (1993).

7 Zhang, Y. Growth performance on acid soil and cultivation technique of *Zenia insignis*. *Guangdong Forestry Science and Technology* **23**, 76–79, http://dx.chinadoi.cn/10.3969/j.issn.1006-4427.2007.01.016 (2007).

8 Hou, L., Li, Y., Li, P., Yang, Y. & Xie, Y. Study of *Zenia insignis* Chun. comprehensive utilization. *Sci. Silvae Sin.* **37**, 139–143, http://dx.chinadoi.cn/10.11707/j.1001-7488.20010322 (2001).

9 Yu, G. & Zhang, Y. A preliminary study on the wood fibre form of *Zenia insignis* introduced in Fujian Province. *Journal of Fujian Forestry Science and Technology* **25**, 31–33, http://dx.chinadoi.cn/10.3969/j.issn.1002-7351.1998.01.006 (1998).

10 Qin, Y.-R., Jiang, G.-M., Cen, Z.-Y., Chen, Y.-S. & Lan, C.-Y. Research on the seed's germination characteristic of *Zenia insignis* as the vanguard tree for afforestation in karst areas. *Seed* **27**, 15–21, http://dx.chinadoi.cn/10.3969/j.issn.1001-4705.2008.12.005 (2008).

11 Tang, W., Pan, B., Wei, H., Huang, Z. & Lu, Q. Comparative study on the characteristics of seed germination and seedling growth of different *Zenia insignis* provenances in Guangxi. *Seed* **38**, 93–98, http://dx.chinadoi.cn/10.16590/j.cnki.1001-4705.2019.12.93 (2019).

12 Bai, H.-R., Oyebanji, O., Zhang, R. & Yi, T.-S. Plastid phylogenomic insights into the evolution of subfamily Dialioideae (Leguminosae). *Plant Diversity* **43**, 27–34, https://doi.org/10.1016/j.pld.2020.06.008 (2021).

13 Herendeen, P. S. *Phylogenetic relationships in Caesalpinioid Legumes: a preliminary analysis based on morphological and molecular data* In *Advances in legume systematics, part 10.* Klitgaard, B. B. & Bruneau, A. Eds 37–62 Royal Botanic Gardens (2003).

14 Kajita, T., Ohashi, H., Tateishi, Y., Bailey, C. D. & Doyle, J. J. rbcL and legume phylogeny, with particular reference to Phaseoleae, Millettieae, and allies. *Syst. Bot.* **26**, 515–536, http://dx.doi.org/10.1043/0363-6445-26.3.515 (2001).

15 Koenen, E. J. M. *et al.* Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. *New Phytol.* **225**, 1355–1369, https://doi.org/10.1016/j.molp.2021.02.006 (2020).

16 Lai, Q., Tu, T. & Zhang, D. The complete plastid genome of *Zenia insignis* Chun (Leguminosae). *Mitochondrial DNA Part B-Resour.* **4**, 2926–2927, https://doi.org/10.1080/23802359.2019.1660266 (2019).

17 The Legume Phylogeny Working Group. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon* **66**, 44–77, https://doi.org/10.12705/661.3 (2017).

18 Zhang, R. *et al.* Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of Leguminosae. *Syst. Biol.* **69**, 613–622, https://doi.org/10.1093/sysbio/syaa013 (2020).

19 Zhao, Y. *et al.* Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. *Mol. Plant.* **14**, 748–773, https://doi.org/10.1016/j.molp.2021.02.006 (2021).

20 Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11–15 (1987).

21 Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S. & Thompson, W. F. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* **1**, 2320–2325, https://doi.org/10.1038/nprot.2006.384 (2006).

22 Doležel, J. & Bartoš, J. Plant DNA flow cytometry and estimation of nuclear genome size. *Ann. Bot.* **95**, 99–110, https://doi.org/10.1093/aob/mci005 (2005).

23 Doležel, J., Greilhuber, J. & Suda, J. Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* **2**, 2233–2244, https://doi.org/10.1038/nprot.2007.310 (2007).

24 Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115, https://doi.org/10.1126/science.1178534 (2009).

25 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).

26 Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k-mers*. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).

27 Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).

28 Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054, https://doi.org/10.1038/nmeth.4035 (2016).

29 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736, https://doi.org/10.1101/gr.215087.116 (2017).

30 Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546, https://doi.org/10.1038/s41587-019-0072-8 (2019).

31 Seifert, D. & Töpfer, A. *GitHub*, <https://github.com/PacificBiosciences/gcpp> (2021).

32 Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255, https://doi.org/10.1093/bioinformatics/btz891 (2019).

33 Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).

34 Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).

35 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921, https://doi.org/10.1038/35057062 (2001).

36 Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126–e126, https://doi.org/10.1093/nar/gky730 (2018).

37 Dudchenko, O. *et al. De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2016).

38 Jin, J.-J. *et al.* GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* **21**, 241, https://doi.org/10.1186/s13059-020-02154-5 (2020).

39 Qu, X.-J., Moore, M. J., Li, D.-Z. & Yi, T.-S. PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* **15**, 50, https://doi.org/10.1186/s13007-019-0435-7 (2019).

40 Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **47**, W59–W64, https://doi.org/10.1093/nar/gkz238 (2019).

41 Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649, https://doi.org/10.1093/bioinformatics/bts199 (2012).

42 Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).

43 Bao, Z. & Eddy, S. R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276, https://doi.org/10.1101/gr.88502 (2002).

44 Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).

45 Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18, https://doi.org/10.1186/1471-2105-9-18 (2008).

469 46 Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for
470 identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–
471 1422, https://doi.org/10.1104/pp.17.01310 (2017).
472 47 Smit, A., Hubley, R. & Rosen, J. *RMBlast*, <https://www.repeatmasker.org/rmblast/>
473 (2024).
474 48 Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences.
475 *Curr. Protoc. Bioinf.* **5**, 4.10.11–14.10.14,
476 https://doi.org/10.1002/0471250953.bi0410s05 (2004).
477 49 Yan, H. & Liu, H. *AnnoSmk*, <https://github.com/yanhui-k/AnnoSmk> (2024).
478 50 The Arabidopsis Genome, I. Analysis of the genome sequence of the flowering plant
479 *Arabidopsis thaliana*. *Nature* **408**, 796–815, https://doi.org/10.1038/35048692
480 (2000).
481 51 She, R. *et al.* genBlastG: using BLAST searches to build homologous gene models.
482 *Bioinformatics* **27**, 2141–2143, https://doi.org/10.1093/bioinformatics/btr342
483 (2011).
484 52 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome
485 alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**,
486 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
487 53 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data
488 without a reference genome. *Nat. Biotechnol.* **29**, 644–652,
489 https://doi.org/10.1038/nbt.1883 (2011).
490 54 Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-
491 molecule long-read sequencing. *Nat. Commun.* **7**, 11708,
492 https://doi.org/10.1038/ncomms11708 (2016).
493 55 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**,
494 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
495 56 Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database
496 management tool for second-generation genome projects. *BMC Bioinformatics* **12**,
497 491, https://doi.org/10.1186/1471-2105-12-491 (2011).
498 57 Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically
499 mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**, 637–
500 644, https://doi.org/10.1093/bioinformatics/btn013 (2008).
501 58 Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal
502 transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666,
503 https://doi.org/10.1093/nar/gkg770 (2003).
504 59 *eggNOG-mapper*, <http://eggnog-mapper.embl.de/> (2024).
505 60 Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J.
506 eggNOG-mapper v2: functional annotation, orthology assignments, and domain
507 prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829,
508 https://doi.org/10.1093/molbev/msab293 (2021).
509 61 Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**,
510 D412–D419, https://doi.org/10.1093/nar/gkaa913 (2020).
511 62 Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427,
512 https://doi.org/10.1093/nar/gkac993 (2022).

513 63 Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature*
514 *Genet.* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
515 64 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a
516 reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–
517 D462, https://doi.org/10.1093/nar/gkv1070 (2015).
518 65 Li, J. *et al.* The nearly complete assembly of the *Cercis chinensis* genome and
519 Fabaceae phylogenomic studies provide insights into new gene evolution. *Plant*
520 *Commun.*, 100422, https://doi.org/10.1016/j.xplc.2022.100422 (2022).
521 66 Jiang, S., An, H., Xu, F. & Zhang, X. Chromosome-level genome assembly and
522 annotation of the loquat (*Eriobotrya japonica*) genome. *GigaScience* **9**, giaa015,
523 https://doi.org/10.1093/gigascience/giaa015 (2020).
524 67 Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**,
525 178–183, https://doi.org/10.1038/nature08670 (2010).
526 68 Liu, H. & Yi, T.-S. *unplished Lysidice rhodostegia data* (Kunming Institute of Botany,
527 Chinese Academy of Sciences, 2024).
528 69 Young, N. D. *et al.* The *Medicago* genome provides insight into the evolution of
529 rhizobial symbioses. *Nature* **480**, 520–524, https://doi.org/10.1038/nature10625
530 (2011).
531 70 Kang, S.-H. *et al.* Genome-enabled discovery of anthraquinone biosynthesis in *Senna*
532 *tora*. *Nat. Commun.* **11**, 5875, https://doi.org/10.1038/s41467-020-19681-1 (2020).
533 71 Ai, W. *et al.* A chromosome-scale genome assembly of the Mongolian oak (*Quercus*
534 *mongolica*). *Mol. Ecol. Resour.* **22**, 2396–2410, https://doi.org/10.1111/1755-
535 0998.13616 (2022).
536 72 Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization
537 in major angiosperm phyla. *Nature* **449**, 463–467,
538 https://doi.org/10.1038/nature06148 (2007).
539 73 Meng, F. *et al.* Genome assembly of *Polygala tenuifolia* provides insights into its
540 karyotype evolution and triterpenoid saponin biosynthesis. *Hortic. Res.-England* **10**,
541 uhad139, https://doi.org/10.1093/hr/uhad139 (2023).
542 74 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for
543 comparative genomics. *Genome Biol.* **20**, 238, https://doi.org/10.1186/s13059-019-
544 1832-y (2019).
545 75 Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for
546 FASTA/Q File Manipulation. *PLoS One* **11**, e0163962,
547 https://doi.org/10.1371/journal.pone.0163962 (2016).
548 76 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high
549 throughput. *Nucleic Acids Res.* **32**, 1792–1797, https://doi.org/10.1093/nar/gkh340
550 (2004).
551 77 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
552 large phylogenies. *Bioinformatics* **30**, 1312–1313,
553 https://doi.org/10.1093/bioinformatics/btu033 (2014).
554 78 Minh, B. Q. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic
555 inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534,
556 https://doi.org/10.1093/molbev/msaa015 (2020).

79 Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819, https://doi.org/10.1093/molbev/msx116 (2017).

80 Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518, https://doi.org/10.1093/bioinformatics/btaa1022 (2020).

81 Chen, C. *et al.* TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. *Mol. Plant.* **16**, 1733–1742, https://doi.org/10.1016/j.molp.2023.09.010 (2023).

82 Sun, P. *et al.* WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant.* **15**, 1841–1851, https://doi.org/10.1016/j.molp.2022.10.018 (2022).

83 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, https://doi.org/10.1186/1471-2105-10-421 (2009).

84 NCBI Sequence Read Archive. *SRR29423588*, <https://www.ncbi.nlm.nih.gov/sra/SRR29423588> (2024).

85 NCBI Sequence Read Archive. *SRR29423585*, <https://www.ncbi.nlm.nih.gov/sra/SRR29423585> (2024).

86 NCBI Sequence Read Archive. *SRR29423587*, <https://www.ncbi.nlm.nih.gov/sra/SRR29423587> (2024).

87 NCBI Sequence Read Archive. *SRR29423586*, <https://www.ncbi.nlm.nih.gov/sra/SRR29423586> (2024).

88 NCBI Sequence Read Archive. *SRR29423581*, <https://www.ncbi.nlm.nih.gov/sra/SRR29423581> (2024).

89 NCBI Sequence Read Archive. *SRR29423580*, <https://www.ncbi.nlm.nih.gov/sra/SRR29423580> (2024).

90 NCBI Sequence Read Archive. *SRR29423579*, <https://www.ncbi.nlm.nih.gov/sra/SRR29423579> (2024).

91 NCBI Sequence Read Archive. *SRR29423578*, <https://www.ncbi.nlm.nih.gov/sra/SRR29423578> (2024).

92 NCBI Sequence Read Archive. *SRR29423589*, <https://www.ncbi.nlm.nih.gov/sra/SRR29423589> (2024).

93 NCBI Sequence Read Archive. *SRR29423583*, <https://www.ncbi.nlm.nih.gov/sra/SRR29423583> (2024).

94 NCBI Sequence Read Archive. *SRR29423582*, <https://www.ncbi.nlm.nih.gov/sra/SRR29423582> (2024).

95 NCBI Sequence Read Archive. *SRR29423584*, <https://www.ncbi.nlm.nih.gov/sra/SRR29423584> (2024).

96 NCBI GenBank. *JBEVYO000000000*, <https://www.ncbi.nlm.nih.gov/nuccore/JBEVYO000000000> (2024).

97 Science Data Bank. *Genomic squences of Zenia insignis*, <https://doi.org/10.57760/sciencedb.10931> (2024).

98 Science Data Bank. *Annotation of Zenia insignis*, <https://doi.org/10.57760/sciencedb.07828> (2024).

99 NCBI GenBank. *PP882779*, <https://www.ncbi.nlm.nih.gov/nuccore/PP882779> (2024).

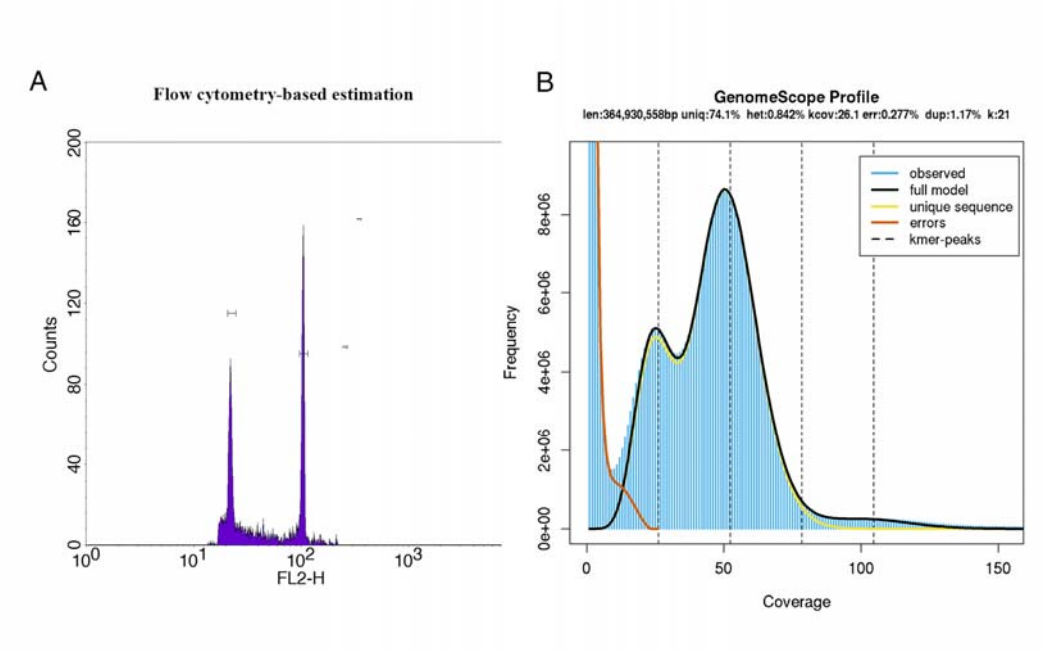


Figure 1 The result of genome size estimation for *Z. insignis*. (A) Flow cytometry-based estimation result of *Z. insignis* (left peak, the sample) and *Zea mays* (right peak, the reference). (B) The 21-*kmer* distribution estimation result was obtained using GenomeScope2. The distribution of *21-kmer* exhibited two major peaks, indicating certain heterozygosity within the genome, and high-frequency *k-mers* with coverage near 1 that was possibly generated by sequencing errors.

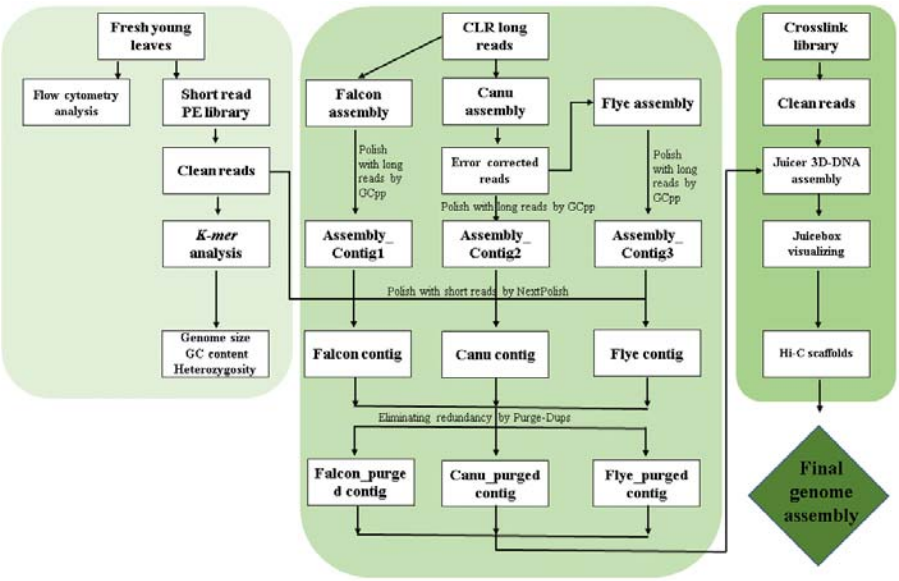


Figure 2 The pipeline for the chromosome-level genome assembly of *Z. insignis*.

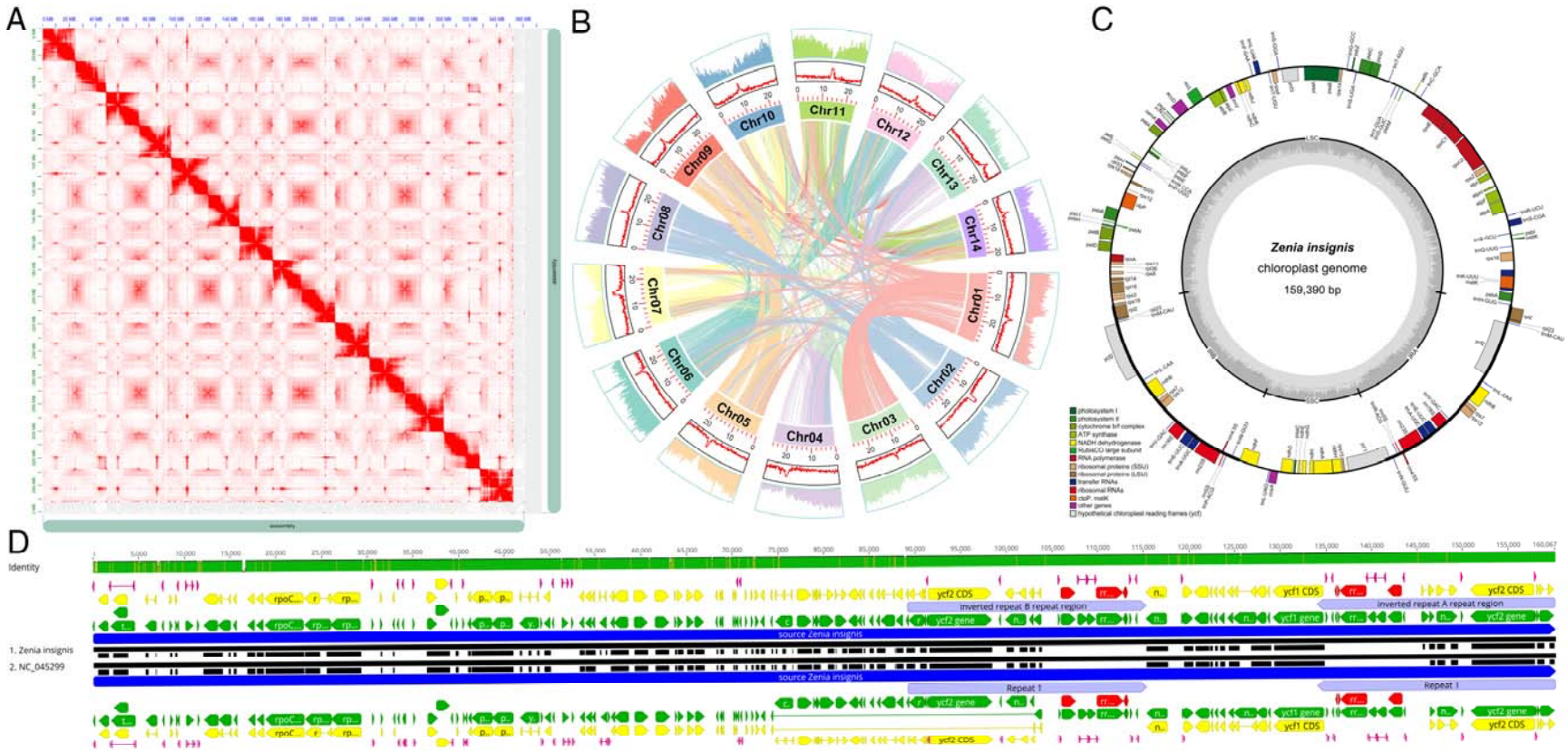


Figure 3 The chromosome and plastome information of *Z. insignis*. (A) The interaction heat map of Hi-C among chromosomes for the *Z. insignis* genome. The Hi-C interaction matrix shows the pairwise correlations among the 14 pseudomolecules. (B) The high-quality genomic landscape of *Z. insignis*. The genome diagram has four circles from the outside: gene density, GC content, chromosomal length, and intraspecies collinearity. (C)The visualization of the plastome genome of *Z. insignis* using OGDRAW. The outside circle displays genes color-coded according to their functional groups. The second circle

617 illustrates the large single-copy (LSC), inverted repeat (IRA, IRB), and small single-copy (SSC) regions. The third circle delineates the percentage of GC
618 content and genomic positions in kilobase pairs. (D) Compared analysis of plastome for *Z. insignis*. Gene(green), CDS (yellow), RNA (red), and IR regions
619 (grey) have been color-coded according to their functional groups. The identity of the two plastome sequences exhibited high overall similarity, except for
620 some regions in non-coding sequences.

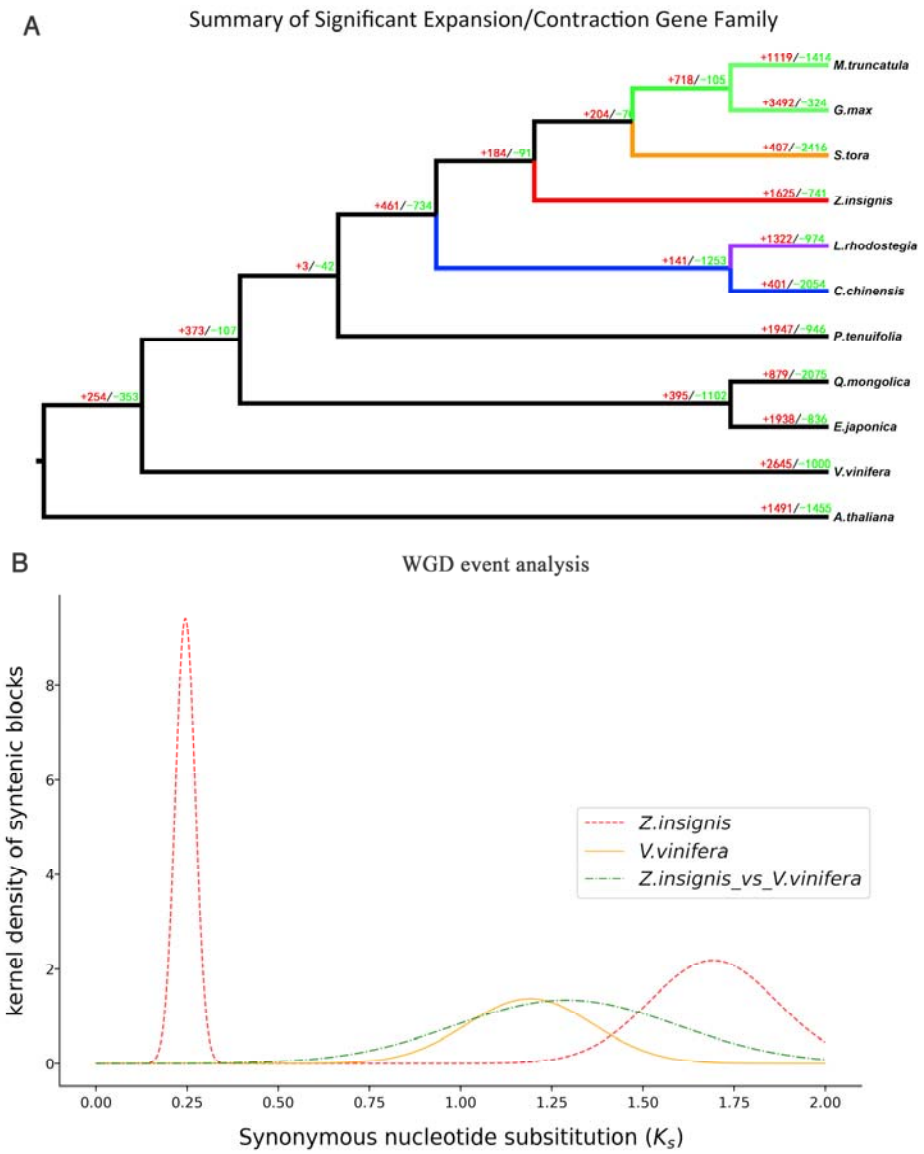


Figure 4 The result of gene family evolution and distribution of *Ks* values. (A) The result of gene family evolution. The plus/minus sign corresponding numbers reflect the expansion or contraction of gene families. The color of the branch indicates different subfamilies of Fabaceae. (B) The distribution of *Ks* values between *Z. insignis* and *V. Vinifera*. The *Ks* distribution curve of *Z. insignis, V. Vinifera, and Z. insignis vs. V. Vinifera* exhibited a significant overlap in the right *Ks* peak.

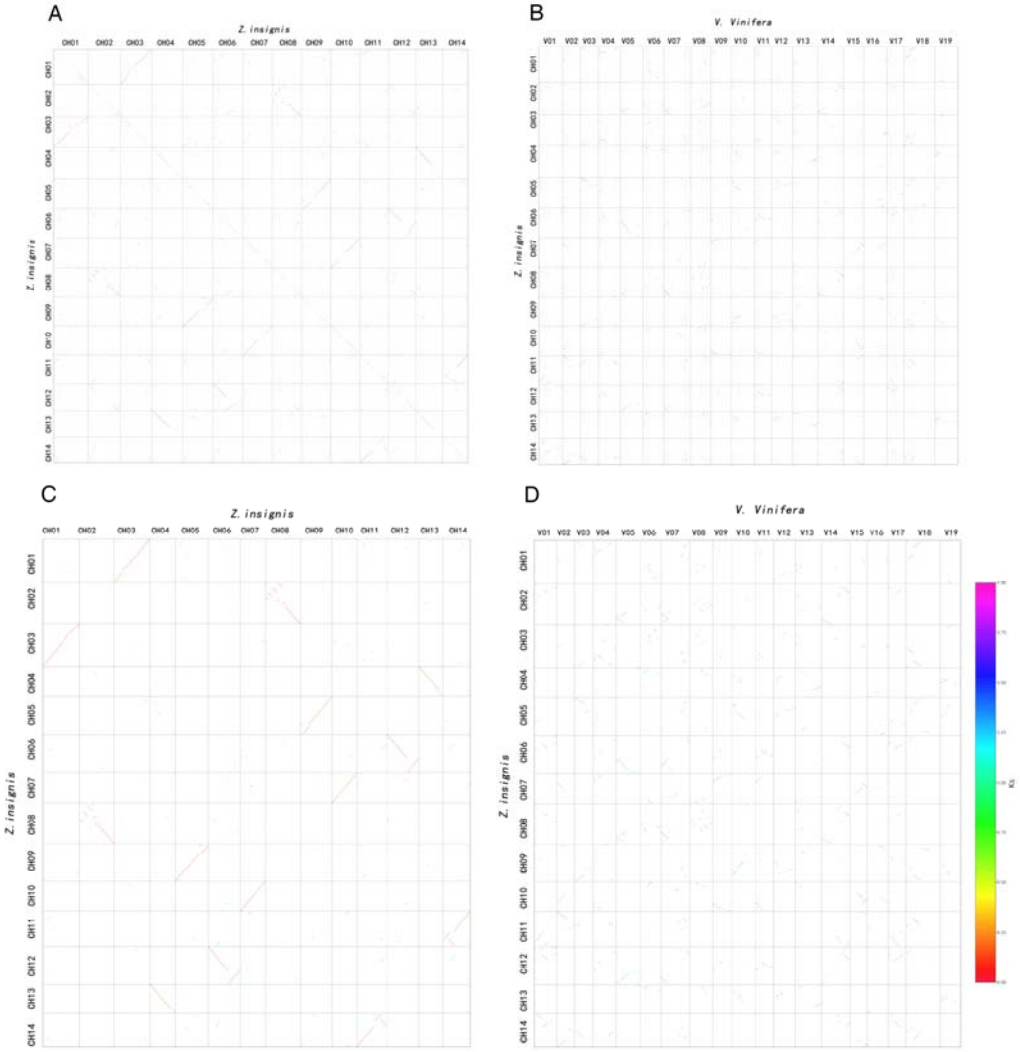
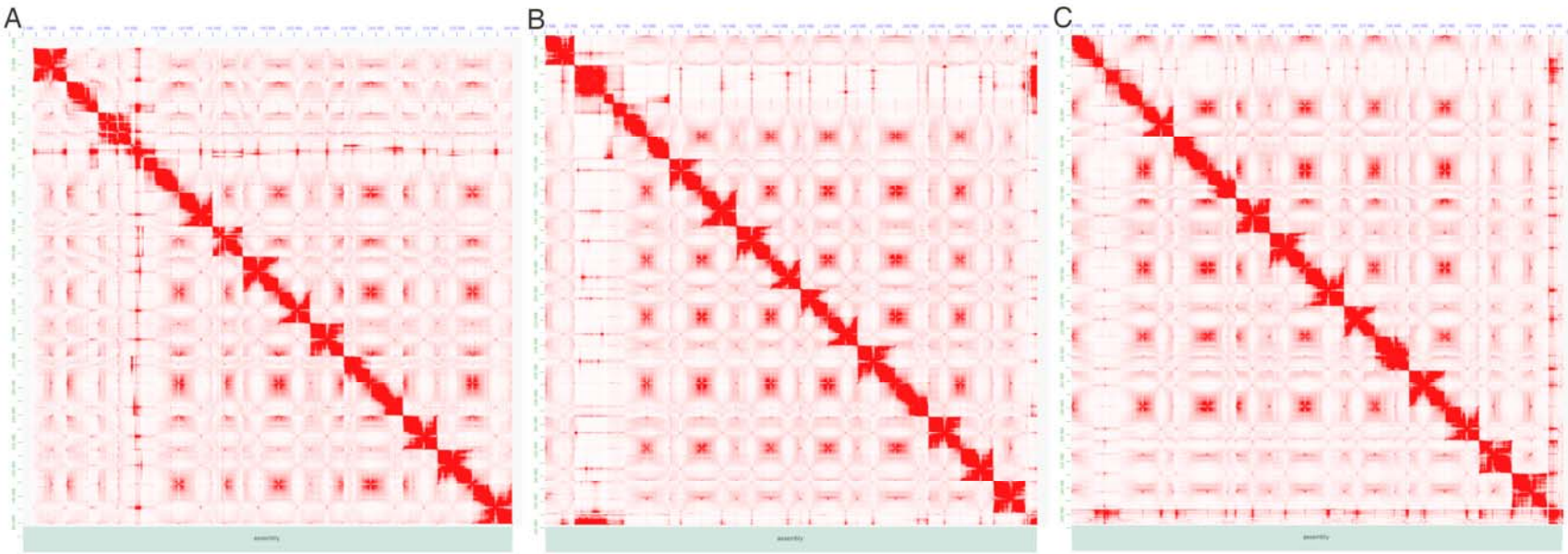
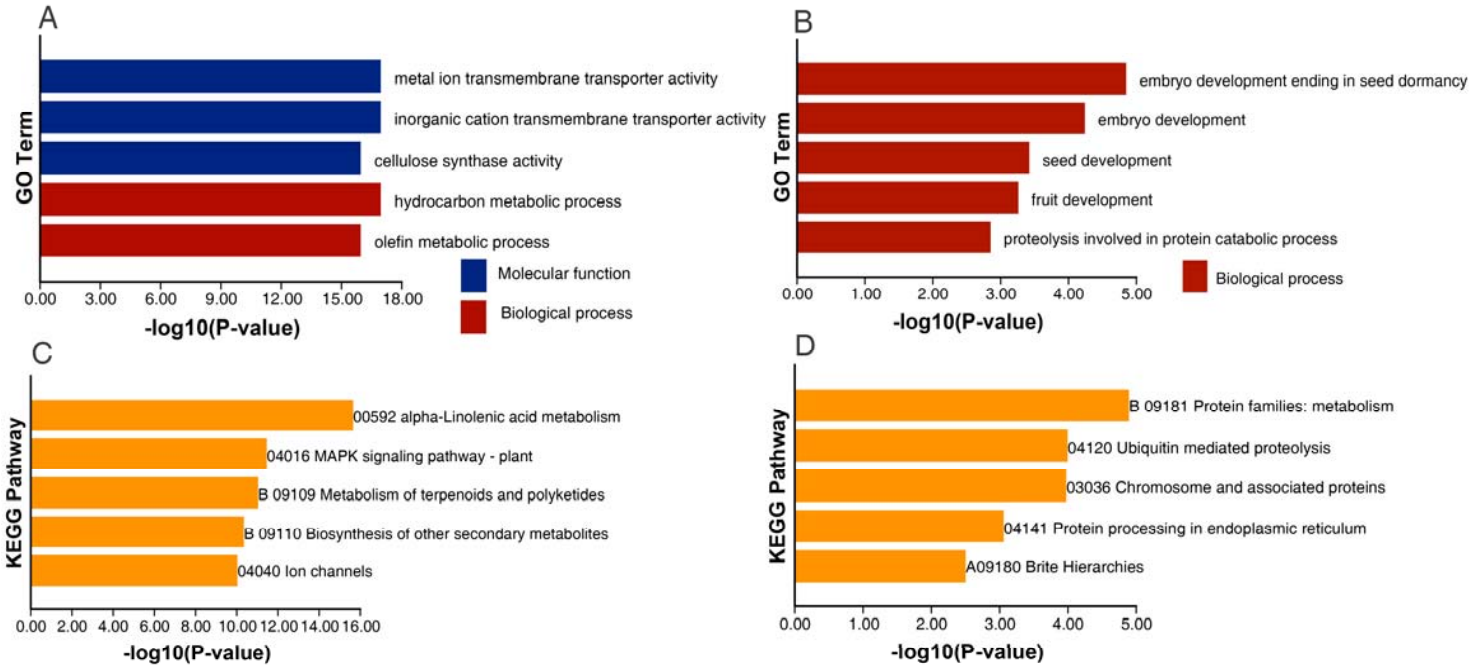


Figure 5 The result of WGDI analysis for *Z. insignis*. (A) The homologous genes dot plot of genomic synteny between chromosomes of *Z. insignis*. The best-hit genes are indicated in red, the second best-hit genes are in blue, and the other hit genes are in gray. (B) Dot plot of genomic synteny between *Z. insignis* and *V. Vinifera*. The best-hit genes are indicated in red, the second best-hit genes in blue, and the other hit genes are in gray (C) *Ks* dot plot of genomic synteny between chromosomes of *Z. insignis*. The color of the dot indicates the *Ks* of the gene pair. (B) *Ks* dot plot of genomic synteny between *Z. insignis* and *V. Vinifera*. The color of the dot indicates the *Ks* of the gene pair.



Supplementary Figure 1 The heat map of the Hi-C interaction matrix generated by Purge_Dups. (A) Heat map of Hi-C for Falcon_purge. (B) Heat map of Hi-C for Canu_purge. (C) The Heat map of Hi-C for Flye_purge.



Supplementary Figure 2 The result of enrichment analyses for GO and KEGG. (A)The result of GO enrichment analyses for significant expansion gene families. (B) The result of GO enrichment analyses for significant contraction gene families. (C) The result of KEGG enrichment analyses for significant expansion gene families. (B) The result of KEGG enrichment analyses for significant contraction gene families.

644 Table 1 Statistics of sequencing data

Survey data	PacBio CCS data	Hi-C data	Full-length Transcriptome data
44140.85Mb	174,139.06Mb	59,630.13Mb	81.92Mb
91.96×	363.43×	124.45×	—

Table 2 The results of the assessment for N50

Nx	Canu_contig		Canu_purged_contig		Falcon_contig		Falcon_purged_contig		Flye_contig		Flye_purged_contig	
	Size	Number	Size	Number	Size	Number	Size	Number	Size	Number	Size	Number
N90	92,660	1,400	1,431,948	43	363,551	152	1,503,027	68	172,335	253	309,802	192
N80	113,319	822	4,428,622	31	1,150,475	87	2,685,811	49	529,601	137	653,152	114
N70	156,297	367	7,093,928	23	2,280,795	59	3,819,453	38	952,759	84	1,139,158	72
N60	320,232	71	8,669,316	18	3,626,040	43	4,916,748	29	1,478,745	53	1,600,360	46
N50	4,693,084	29	10,026,892	14	4,819,753	31	6,024,120	22	2,345,742	32	2,516,555	29
N40	7,292,602	20	12,366,826	11	5,750,257	23	7,011,975	17	3,977,660	20	4,128,350	18
N30	10,196,961	13	14,068,457	8	7,098,875	16	7,310,681	12	6,920,119	12	7,140,693	11
N20	14,068,457	8	14,855,825	5	7,527,208	9	8,509,734	7	9,355,035	7	9,355,035	7
N10	15,123,523	4	15,261,855	3	9,877,742	4	11,240,351	3	9,936,040	4	10,281,987	3
TotalLength	592,944,931	2,337	377,600,313	448	457,757,591	598	360,998,718	266	371,463,918	1,303	354,656,638	744

Table 3 The results of the assessment for BUSCO

	Canu_contig	Canu_purged_contig	Falcon_contig	Falcon_purged_contig	Flye_contig	Flye_purged_contig
Complete BUSCOs (C)	98.50%	98.50%	97.50%	97.30%	98.60%	98.60%
Complete and single-copy BUSCOs (S)	44.10%	77.30%	65.80%	79.80%	76.70%	76.90%
Complete and duplicated BUSCOs (D)	54.40%	21.20%	31.70%	18.40%	21.90%	21.70%
Fragmented BUSCOs (F)	0.30%	0.60%	1.00%	1.20%	0.40%	0.60%
Missing BUSCOs (M)	1.20%	0.90%	1.50%	1.50%	1.00%	0.80%
Total BUSCO groups	1614	1614	1614	1614	1614	1614

647 Table 4 The results of the assessment for LAI

	Chromosome From	To	Intact	Total	raw_LAI	LAI
Canu_contig	1	592,944,931	0.0088	0.2003	4.38	0
Canu_purged_contig	1	377,600,313	0.0126	0.2099	5.98	11.61
Falcon_contig	1	457,757,591	0.0077	0.1684	4.57	4.62
Falcon_purged_contig	1	360,998,718	0.0119	0.1329	8.94	14.57
Flye_contig	1	371,463,918	0.0109	0.1868	5.83	11.46
Flye_purged_contig	1	354,656,638	0.0108	0.1689	6.4	12.03

648 Table 5 The result of repeat sequences annotation

Type	Length(bp)	Rate (%)
DNAs	11,129,683	3.08
LINEs	6,520,335	1.81
SINEs	298,378	0.08
LTRs	32,617,056	9.03
Rolling-circles	1,280,052	0.35
Small RNA	6,975,086	1.93
Satellites	3,559,151	0.99
Simple repeats	1,063,047	0.29
Low complexity	64,922,128	17.98
Unknown	128,083,818	35.47
Total	11,129,683	3.08

649 Table 6 The result of BUSCO for gene prediction

	R1	R2	R3	R3+PASA
Complete BUSCOs (C)	46.60%	96.60%	96.80%	97.20%
Complete and single-copy BUSCOs (S)	39.90%	44.80%	43.90%	44.20%
Complete and duplicated BUSCOs (D)	6.70%	51.80%	52.90%	53.00%
Fragmented BUSCOs (F)	3.50%	2.30%	2.30%	1.90%
Missing BUSCOs (M)	49.90%	1.10%	0.90%	0.90%
Total BUSCO groups	1614	1614	1614	1614

650