



OPEN

DATA DESCRIPTOR

Chromosome level genome assembly of colored calla lily (*Zantedeschia elliottiana*)

Yi Wang^{1,6}, Tuo Yang^{2,6}, Di Wang^{1,3}, Rongxin Gou^{1,3}, Yin Jiang^{1,3}, Guojun Zhang³, Yuhong Zheng⁴, Dan Gao⁵, Liyang Chen⁵, Xiuhai Zhang¹ & Zunzheng Wei¹

The colored calla lily is an ornamental floral plant native to southern Africa, belonging to the *Zantedeschia* genus of the Araceae family. We generated a high-quality chromosome-level genome of the colored calla lily, with a size of 1,154 Mb and a contig N50 of 42 Mb. We anchored 98.5% of the contigs (1,137 Mb) into 16 pseudo-chromosomes, and identified 60.18% of the sequences (694 Mb) as repetitive sequences. Functional annotations were assigned to 95.1% of the predicted protein-coding genes (36,165). Additionally, we annotated 469 miRNAs, 1,652 tRNAs, 10,033 rRNAs, and 1,677 snRNAs. Furthermore, Gypsy-type LTR retrotransposons insertions in the genome are the primary factor causing significant genome size variation in Araceae species. This high-quality genome assembly provides valuable resources for understanding genome size differences within the Araceae family and advancing genomic research on colored calla lily.

Background & Summary

Zantedeschia spp, commonly known as calla lily, is a perennial herbaceous flowering plant belonging to genus *Zantedeschia* of the family Araceae. It is typically found in swamps and hills regions of South Africa^{1,2}. Through its unique spathes and decorative foliage, calla lily has become popular tubers flowering plants worldwide. It is usually divided into two groups: white calla lily and colored calla lily³. Colored calla lily is a significant economic horticultural crop that have been among the top cut flower and tuber exports in New Zealand for the past three decades, while also contributing substantially to the horticultural export revenues of the Netherlands and the United States. Furthermore, the tubers of colored calla lilies have medicinal value and are effective in treating certain gastrointestinal and trauma-related illnesses.

Through k-mer and flow cytometry analysis, the genome size of *Zantedeschia elliottiana* cv. 'Jingcai Yangguang' was ~1.2 Gb, with a genome heterozygosity of 1.9% and a repeat sequence proportion of 67.84% (Figs. 1, 2). The *de-novo* assembly of the genome used 84.30X Illumina paired-end short reads (100.31 Gb), 36.92X HiFi reads (43.93 Gb) and 141.45X Hi-C reads (168.18 Gb). We first assembled the genome by HiFi reads and generated a 1,154 Mb contig sequence with 42 Mb contig N50 size (Table 1). Using Hi-C reads, 98.50% of the contigs were anchored into 16 pseudo-chromosomes (Fig. 3, Table 1). The transposable elements content of the total genome in the final annotation is 60.18%, of which LTR retroelement accounted for the largest proportion (51.54%). On the contrary, the proportion of DNA transposons was only 3.73% (Table 2). A total of 36,165 protein-coding genes were predicted, of which 95.1% could be functionally annotated through the InterPro⁴, Pfam⁵, Swiss-Prot⁶, NCBI Non-redundant protein (NR)⁷ and Kyoto Encyclopedia of Genes and Genomes (KEGG)⁸ databases (Table 3). In addition, 10,033 rRNA, 1,677 snRNA, 469 miRNA and 1,652 tRNA in *Zantedeschia elliottiana* cv. 'Jingcai Yangguang' genome were obtained by non-coding RNA annotation (Table 4). Using BUSCO evaluation, 98% of the core genes can be identified, including 95.7% of complete single-copy

¹Institute of Grassland, Flowers and Ecology, Beijing Academy of Agriculture and Forestry Sciences, Beijing, 100097, China. ²College of Horticulture, China Agricultural University, Beijing, 100193, China. ³College of Horticultural Science & Technology, Hebei Key Laboratory of Horticultural Germplasm Excavation and Innovative Utilization/Hebei Higher Institute Application Technology Research and Development Center of Horticultural Plant Biological Breeding, Hebei Normal University of Science & Technology, Qinhuangdao, 66004, China. ⁴Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing Botanical Garden, Mem. Sun Yat-Sen, Nanjing, 210014, China. ⁵Smartgenomics Technology Institute, Tianjin, 301700, China. ⁶These authors contributed equally: Yi Wang, Tuo Yang. ✉e-mail: zhangxiuhai@baafs.net.cn; weizunzheng@163.com

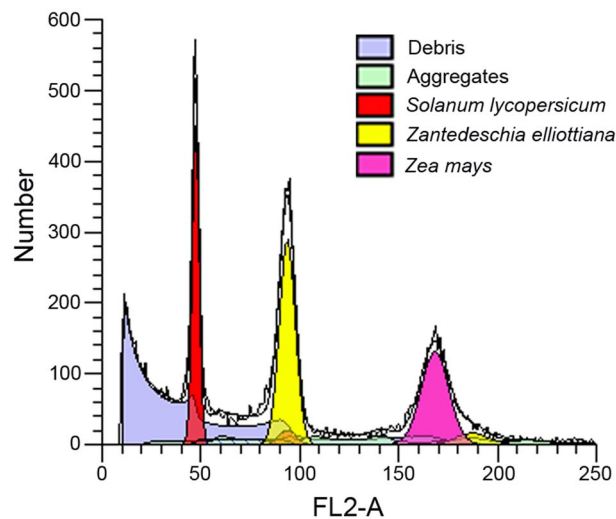


Fig. 1 Genome size estimation of *Zantedeschia elliptiana* cv. 'Jingcai Yangguang' by flow cytometry. Tomato and maize were used as internal references to genome size estimation.

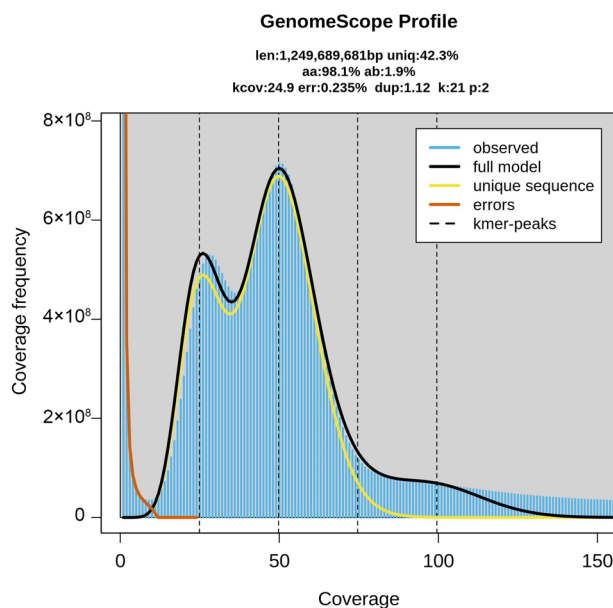


Fig. 2 Genome size estimation of *Zantedeschia elliptiana* cv. 'Jingcai Yangguang' using Illumina reads.

genes and 2.3% of duplicated genes (Table 1). 93.83–95.23% of RNA-seq reads from eight *Zantedeschia elliptiana* cv. 'Jingcai Yangguang' tissues (tuber, leaf, pistil, root, spathe, stamen, stem and style) could be mapped to the genome. 99.02% of Illumina reads and 98.42% of HiFi reads were correctly mapped to the genome. The LTR Assembly Index (LAI) of the genome was 18.43, which directly proved that the genome has high continuity (Table 1). LTR insertion time analysis showed that Araceae plants had different LTR bursts during genome evolution, and different types of LTR have different burst states. For Copia-type LTR retrotransposons, *Pistia stratiotes* and *Zantedeschia elliptiana* cv. 'Jingcai Yangguang' had the same insertion time. Interestingly, *Amorphophallus konjac* and *Colocasia esculenta* experienced two outbreaks of Copia and Gypsy. The time interval between the two outbreaks of *Colocasia esculenta* were obvious, while *Amorphophallus konjac* were close. Analysis also showed that Gypsy of *Pistiastratiotes* had recently experienced an outbreak (Fig. 4a). As a branch of Araceae family, Lemnaceae plants have a smaller genome size and number of genes than True-Araceae plants. However, the genome size of True-Araceae plants is not related to the number of genes. Correlation analysis further explained the high correlation between genome size and transposable elements. Gypsy-type LTR retrotransposons had the highest correlation with genome size (Fig. 4b).

Here, a high-quality chromosome-level assembly of *Zantedeschia elliptiana* cv. 'Jingcai Yangguang' was assembled, revealing the fundamental cause of genome size variation in the Araceae family.

Assembly characteristics	<i>Z. elliotiana</i> cv. 'Jingcai Yangguang'
Total length of contigs (bp)	1,154,500,755
N50 length of contigs (bp)	42,376,536
Total number of contigs	375
Longest contigs	80,375,493
Total gap size (bp)	6,700
Total sequences anchored to the pseudo-chromosomes (bp)	1,137,238,020
place rate (%)	98.50
Number of annotated genes	36,165
Percentage of transposon element sequences (%)	60.18
Complete BUSCOs (%)	98.00
Fragmented BUSCOs (%)	0.80
Missed BUSCOs (%)	1.20
LAI	18.43

Table 1. Summary of the *Z. elliotiana* genome.

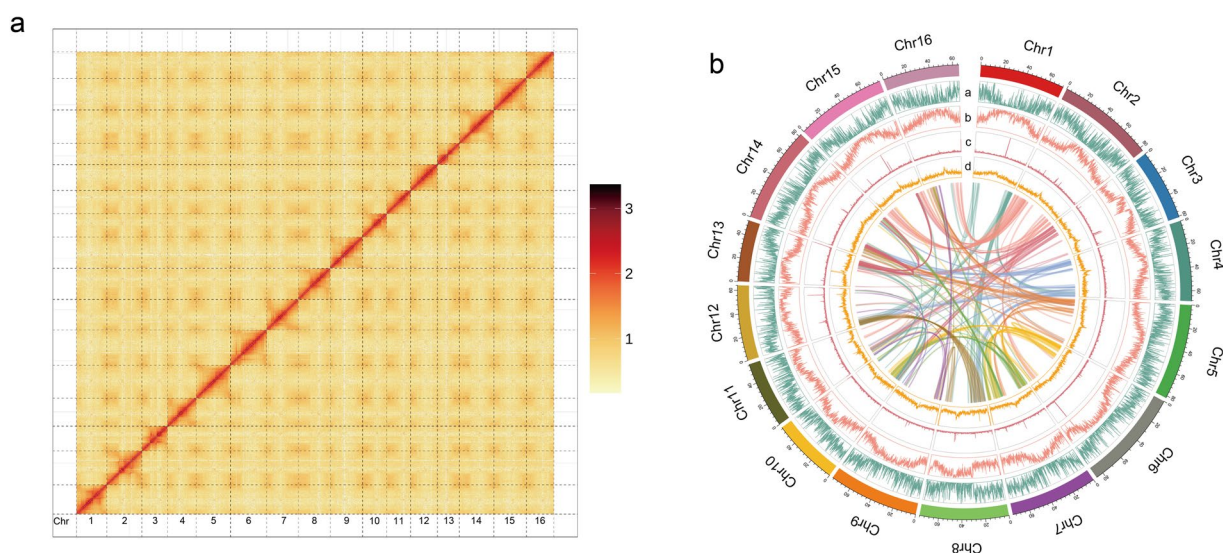


Fig. 3 Characteristics of *Zantedeschia elliotiana* cv. 'Jingcai Yangguang' genome. **(a)** Hi-C heatmap of the *Zantedeschia elliotiana* cv. 'Jingcai Yangguang' genome. **(b)** Circos plot of *Zantedeschia elliotiana* cv. 'Jingcai Yangguang' genome. **(a)** Gene density, **(b)** TE density, **(c)** Tandem repeats density, **(d)** GC content and syntenic blocks.

Methods

Sample collection and sequencing. 'Jingcai Yangguang' is a variant of *Zantedeschia elliotiana* cv. 'Black Magic' with a chromosome number of $2n = 2x = 32$. It was initially cultivated in 2015 by Di Zhou, a former associate researcher in our team. Its young leaves were collected for genome sequencing, and the sequencing material was sourced from the same plant to ensure accuracy of the sequencing. Eight tissues (tuber, leaf, pistil, root, spathe, stamen, stem and style) were sampled for transcriptome sequencing, and the sequencing results were used for gene structure annotation.

The FastPure Plant DNA Isolation Mini Kit (Vazyme, CHN) was employed for DNA extraction from leaf tissue. In liquid nitrogen, fresh leaves were pulverized into a fine powder, and genomic DNA was isolated according to the manufacturer's guidelines. NanoDrop 2000 (Thermo Scientific, USA) and gel electrophoresis were utilized to evaluate the concentration and purity of the isolated DNA.

The high-quality DNA was used to construct a genomic library, and the library construction and sequencing work were completed at Novogene Co., Ltd. in Beijing. The library is then size-selected using BluePippin (Sage Science, USA) to obtain fragments of the desired size range, which is typically ~15 kb for HiFi sequencing. The purified and size-selected library is then sequenced on the PacBio Sequel II system (Pacific Biosciences, USA). For Illumina sequencing, a short-read sequencing library was constructed with an insert size of ~250 bp and sequenced on an Illumina NovaSeq. 6,000 platform (Illumina, USA). The Hi-C library was constructed using the same leaf sample as previously described. Briefly, nuclear DNA was fixed with formaldehyde and digested with the restriction enzyme DpnII (NEB, UK). Biotinylated nucleotides were added to the termini

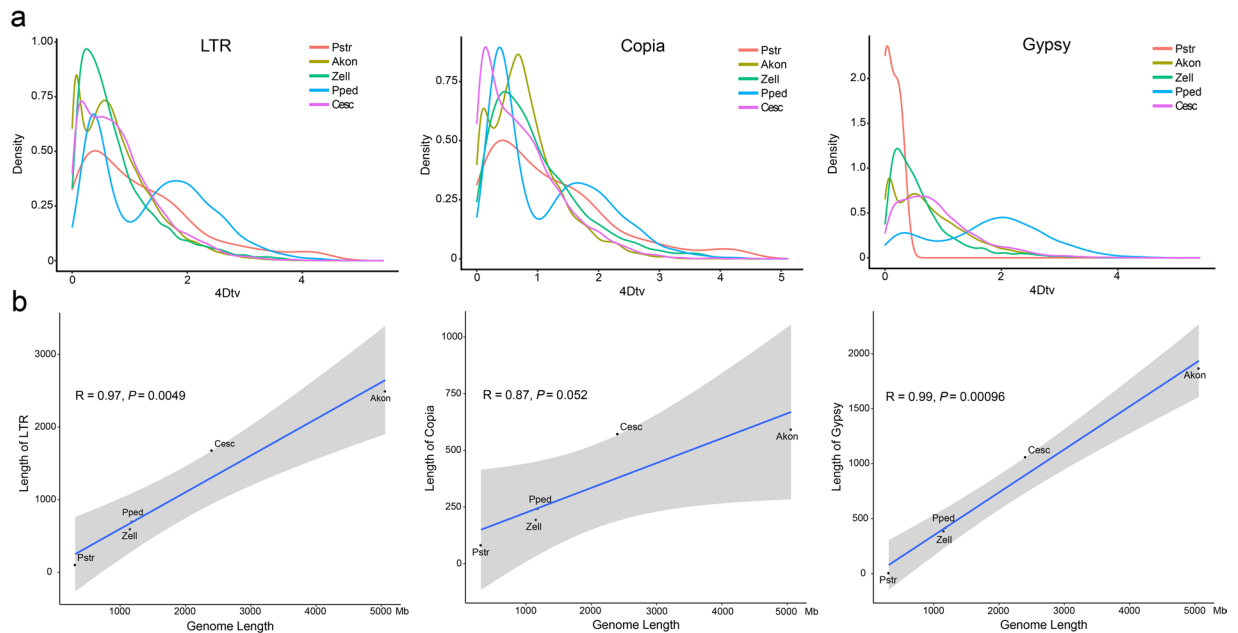


Fig. 4 The influence of LTRs on genome size. **(a)** The insertion time of LTRs (Copia and Gypsy) was predicted by 4Dtv. Pstr, *Pistia stratiotes*; Akon, *Amorphophallus konjac*; Zell, *Zantedeschia elliottiana* cv. ‘Jingcai Yangguang’; Pped, *Pinellia pedatisecta*; Cesc, *Colocasia esculenta*. **(b)** Analysis of the correlation between the total length of LTRs and the genome size.

of the fragmented DNA, followed by enrichment and size selection to obtain fragments approximately 500 bp. The library was sequenced on the Illumina NovaSeq. 6,000 platform (Illumina, USA).

The RNAprep Pure Plant Kit (TIANGEN, CHN) was used to extract RNA from 8 different tissues (tuber, leaf, pistil, root, spathe, stamen, stem and style). The tissue samples were ground with liquid nitrogen and lysis buffer was added to extract RNA. The RNA was isolated according to the manufacturer’s guidelines. RNA-seq libraries were generated and sequenced on an NovaSeq. 6,000 platform (Illumina, USA).

Genome size estimation. Two methods, k-mer and flow cytometry analysis, were employed to estimate the genome size of *Zantedeschia elliottiana* cv. ‘Jingcai Yangguang’. For flow cytometry analysis, the DNA content of *Zantedeschia elliottiana* cv. ‘Jingcai Yangguang’ was assessed using the BD Accuri C6 flow cytometer (BD Biosciences, USA), with tomato and maize as reference standards (Fig. 1). The frequency distribution of k-mer was assessed using Jellyfish (v1.0.0) (-C -m 21 -G 2)⁹. Using GenomeScope (v2.0) (-p 2 -k 21)¹⁰ to calculate the genome size and heterozygosity level with k-mer size = 21 (Fig. 2).

De-novo genome assembly. Firstly, contigs were assembled from HiFi reads using hifiasm (v0.19.5) (<https://github.com/chhylp123/hifiasm>) with default parameters. Subsequently, Hi-C reads were aligned to contigs using HICUP (v0.7.3)¹¹ to evaluate the efficiency of data. Following that, contigs were anchored into 16 pseudo-chromosomes using YaHS (v1.1) with default parameters (Fig. 3). Finally, the assembled genome was manually corrected with Juicebox (v1.11.08) (Table 1)¹².

Completeness evaluation of the assembled genome. Benchmarking Universal Single-Copy Orthologs (BUSCO v5.4.5, embryophyta_odb10)¹³, and LTR Assembly Index (LAI, LTR_retriever v2.9.0)¹⁴ were used to determine the completeness of the genome, respectively (Table 1).

Genome prediction and annotation. The annotation pipeline employed for predicting repeat elements consisted of both homology-based and *de-novo* approaches. In the homology-based approach, alignment searches were conducted against the Repbase database (<http://www.girinst.org/repbase>)¹⁵ to identify homologous evidence, which was subsequently predicted using RepeatProteinMask (v4.1.0) (<http://www.repeatmasker.org/>). For *de-novo* annotation, a *de-novo* library was constructed using LTR_FINDER (v1.07)¹⁶, RepeatScout (v1.0.6) (<http://www.repeatmasker.org/>)¹⁷, and RepeatModeler (v2.0.4) (<http://www.repeatmasker.org/RepeatModeler.html>)¹⁸. The annotation process was then performed using Repeatmasker (v4.1.0) (<http://repeatmasker.org/>)¹⁹.

To annotate the gene structure, a strategy incorporating *de-novo* prediction, protein-based homology, and transcriptome were employed. Protein sequences from *Amorphophallus konjac*, *Colocasia esculenta*, *Lemna minuta*, *Spirodela polyrhiza*, *Pistia stratiotes* and *Pinellia pedatisecta* were mapped to their respective genome using WUblast (v2.0)²⁰. GeneWise (v2.4.1)²¹ was utilized to predict the gene structures in the genomic regions identified by WUblast (v2.0). The gene structures generated by GeneWise (v2.4.1) were referred to as the Homo-set. Additionally, gene models produced by PASA (v2.5.2)²², which served as training data for *de-novo* gene prediction programs. Five *de-novo* gene prediction programs, namely AUGUSTUS (v2.5.5)²³, Genscan

<i>Z. Elliottiana</i> cv. 'Jingcai Yangguang'		
Without N gaps: 1,154,500,755		
Repetitive sequences	Length (bp)	Ratio (%) in genome
LTR Retroelement	595,082,160	51.54
Gypsy (LTR)	387,224,277	33.54
Copia (LTR)	194,442,787	16.84
LINE	47,578,872	4.12
SINE	293,029	0.03
DNA transposons	43,021,358	3.73
Other/Unspecified/Unknown	29,613,899	2.57

Table 2. Classification of repetitive sequences in *Z. Elliottiana* cv. 'Jingcai Yangguang' genome.

Database	<i>Z. Elliottiana</i> cv. 'Jingcai Yangguang'	
	Gene numbers	Ratio (%)
NR	23,081	63.80
Swiss-Prot	16,854	46.60
KEGG	16,690	46.10
Pfam	19,198	53.10
GO	15,276	42.20
Annotated	34,406	95.10
Total	36,165	—

Table 3. Statistics of gene functional annotation.

Type	Number	Average length (bp)	Total length (bp)	% of genome	
miRNA	469	117.95	55,320	0.004792	
tRNA	1,652	75.21	124,254	0.010763	
rRNA	18S	1,571	1759.02	2,763,428	0.239360
	28S	6,091	143.58	874,541	0.075750
	5.8S	1,532	159.30	244,048	0.021139
	5S	839	115.46	96,874	0.008391
snRNA	CD-box	1,408	106.49	149,933	0.012987
	HACA-box	69	147.61	10,185	0.000882
	splicing	200	134.86	26,971	0.002336

Table 4. Classification of non-coding RNAs in *Z. Elliottiana* cv. 'Jingcai Yangguang' genome.

(v1.0)²⁴, Geneid (v1.4)²⁵, GlimmerHMM (v3.0.1)²⁶ and SNAP (v2013.11.29)²⁷, were employed to predict coding regions within the repeat-masked genome. To perform transcript-based annotations, the clean data were aligned to the genome assembly using TopHat (v2.0)²⁸, and Cufflinks (v2.1.1)²⁹. These results were combined by EvidenceModeler (v1.1.1)²², which generated a non-redundant set of gene annotations.

The predicted protein sequences were functionally annotated through searches in five databases: NR⁷, InterPro⁴, KEGG⁸, Pfam⁵ and Swiss-Prot⁶. Gene Ontology (GO)³⁰ annotation was performed using InterProScan (v5.52–86.0)³¹ (Table 3). Blast (v2.2.26) (E-value threshold of 1E-5) were used to align the protein sequences of *Zantedeschia Elliottiana* to these databases for gene function annotation.

Noncoding RNA (ncRNA) annotation was conducted using tRNAScan (v1.4)³² and blast (v2.2.26)³³ for predicting tRNA and rRNA, respectively. Furthermore, miRNA and snRNA were identified through alignment with the Rfam database³⁴ using INFERNAL (v1.0)³⁵.

Estimation of LTR retrotransposons insertion timing. The full-length LTR retrotransposons were aligned to the ClariTeRep³⁶ datasets using blastn (blast, v2.2.26). The insertion time of each LTR retrotransposon was calculated. The alignment of the 5' and 3' LTRs was performed using MUSCLE (v5.1)³⁷, and the EMBOSS software package (v6.6.0)³⁸ was used to calculate the accumulated divergence³⁹.

Data Records

The raw data (PacBio HiFi reads, Illumina reads, and Hi-C sequencing reads) used for genome assembly were deposited in the SRA at NCBI SRR24273711-SRR24273714^{40–43}.

The RNA-seq data were deposited in the SRA at NCBI SRR24273483-SRR24273490^{44–51}. The genome assembly and annotation files are available in Figshare (<https://doi.org/10.6084/m9.figshare.22656112>)⁵² and GenBank under the accession JARZZO000000000⁵³.

Technical Validation

Firstly, the Hi-C heatmap exhibits the accuracy of genome assembly, with relatively independent Hi-C signals observed between the 16 pseudo-chromosomes (Fig. 2a). Moreover, we aligned RNA and DNA reads to the final determined genome to assess the accuracy of genome assembly. For the alignment of DNA reads, Illumina reads were aligned using BWA (v0.7.17)⁵⁴ with default parameters, while HiFi reads were aligned using minimap2 (v2.24-r1122)⁵⁵ with default parameters. The mapping rate for Illumina reads was 99.02%, while the mapping rate for HiFi reads was 98.42%. For the alignment of RNA reads, transcriptomic data from different tissues were individually mapped to the final determined genome using HISAT2 (v2.2.1)⁵⁶ with default parameters. The mapping rates for the respective tissue-specific transcriptomic data ranged from 93.83% to 95.23%. Furthermore, we evaluated the completeness of the genome using BUSCO (v5.4.5, embryophyta_odb10)¹³, and LAI (LTR_retriever, v2.9.0)¹⁴ (Table 1). Overall, these assessments individually confirmed the accuracy and completeness of the genome assembly.

Code availability

All data processing commands and pipelines were carried out in accordance with the instructions and guidelines provided by the relevant bioinformatic software. There were no custom scripts or code utilized in this study.

Received: 7 June 2023; Accepted: 22 August 2023;

Published online: 09 September 2023

References

- Letty, C. The Genus *Zantedeschia*. (1973).
- Yao, J.-L., Rowland, R. E. & Cohen, D. Karyotype studies in the genus *Zantedeschia* (Araceae). *S. Afr. J. Bot.* **60**, 4–7 (1994).
- De Hertogh, A. & Le Nard, M. *The physiology of flower bulbs*. (1993).
- Finn, R. D. *et al.* InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res* **45**, D190–D199 (2016).
- Finn, R. D. *et al.* Pfam: the protein families database. *Nucl. Acids Res* **42**, D222–D230 (2013).
- Bairoch, A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res* **28**, 45–48 (2000).
- O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucl. Acids Res* **44**, D733–D745 (2016).
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457–D462 (2015).
- Marçais, G. & Kingsford, C. Jellyfish: A fast k-mer counter. *Tutorialis e Manuais* **1**, 1–8 (2012).
- 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 (2020).
- Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4** (2015).
- Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**, 1119–1125 (2013).
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
- Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucl. Acids Res* **46**, e126–e126 (2018).
- Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462–467 (2005).
- Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265–W268 (2007).
- Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
- Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457 (2020).
- Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinformatics* **25** (2009).
- She, R., Chu, J. S.-C., Wang, K., Pei, J. & Chen, N. genBlastA: Enabling BLAST to identify homologous gene sequences. *Genome Res* **19**, 143–149 (2008).
- Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
- Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
- Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–W439 (2006).
- Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol* **268**, 78–94 (1997).
- Guigó, R. Assembling Genes from Predicted Exons in Linear Time with Dynamic Programming. *J. Comput. Biol* **5**, 681–702 (1998).
- Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
- Korf, I. Gene finding in novel genomes. *BMC Bioinform* **5**, 1–9 (2004).
- Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
- Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562–578 (2012).
- Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).
- Mulder, N. & Apweiler, R. InterPro and InterProScan. *Humana Press*, 59–70 (2007).
- Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucl. Acids Res* **25**, 955–964 (1997).
- Mount, D. W. Using the Basic Local Alignment Search Tool (BLAST). *Cold Spring Harb Protoc.* **17** (2007).
- Griffiths-Jones, S. Rfam: annotating non-coding RNAs in complete genomes. *Nucl. Acids Res* **33**, D121–D124 (2004).
- Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).

36. Daron, J. *et al.* Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome biology* **15**, 1–15 (2014).
37. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res* **32**, 1792–1797 (2004).
38. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).
39. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. USA* **101**, 12404–12410 (2004).
40. NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRR24273711> (2023).
41. NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRR24273712> (2023).
42. NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRR24273713> (2023).
43. NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRR24273714> (2023).
44. NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRR24273483> (2023).
45. NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRR24273484> (2023).
46. NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRR24273485> (2023).
47. NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRR24273486> (2023).
48. NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRR24273487> (2023).
49. NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRR24273488> (2023).
50. NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRR24273489> (2023).
51. NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRR24273490> (2023).
52. Yang, T. Genome annotation files of *Zantedeschia elliottiana* 'Jingcai Yangguang', *figshare*, <https://doi.org/10.6084/m9.figshare.22656112> (2023).
53. Wang, Y. *Zantedeschia* hybrid cultivar cultivar Jingcaiyanguang, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:JARZZO000000000> (2023).
54. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
55. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
56. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (32071812), Beijing Academy of Agriculture and Forestry Sciences Specific Projects for Building Technology Innovation Capacity (KJCX20230108; KJCX20230801; KJCX20230811).

Author contributions

Z.W. and X.Z. designed the study and led the research. Y.W. and T.Y. wrote the draft manuscript. Y.W., T.Y., D.G. and L.C. contribute to the genome assembly and annotation. Y.W., T.Y., D.W., R.G., Y.J., D.G. and L.C. participated in genome evolution analysis. Z.W., X.Z., G.Z. and Y.Z. contributed substantially to the revisions. The final manuscript has been read and approved by all authors.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.Z. or Z.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023