LETTER

The evolutionary tale of lilies: Giant genomes derived from transposon insertions and polyploidization

Sujuan Xu,[1,2,20] Runzhou Chen,[3,20] Xinqi Zhang,[1,2,20] Yufeng Wu,[4,20] Liuyan Yang,[5,20] Zongyi Sun,[6] Zhitao Zhu,[4] Aiping Song,[1] Ze Wu,[1,2] Ting Li,[1,2] Biao Jin,[7] Shihui Niu,[8] Xin-Cheng Huang,[1,4] Si-Jie Liu,[1,4] Cheng-Ao Yang,[1,4] Guixia Jia,[9] Yanhong He,[3] Fang Du,[10] Minmin Chen,[5] Fei Chen,[11] Wenhe Wang,[12] Hongmei Sun,[13] Yongyao Fu,[14] Weibiao Liao,[15] Huaidi Pei,[16] Xuewei Wu,[17] Sixiang Zheng,[18] Jia-Yu Xue,[1,4,*] Guogui Ning,[3,*] Ray Ming,[19,*] and Nianjun Teng[1,2,*]

[1]Key Laboratory of Landscaping, Ministry of Agriculture and Rural Affairs, College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China
[2]Lily Science and Technology Backyard Qixia of Jiangsu, Nanjing 210043, China
[3]National Key Laboratory for Germplasm Innovation & Utilization of Horticultural Crops, Huazhong Agricultural University, Wuhan 430070, China
[4]State Key Laboratory for Crop Genetics and Germplasm Enhancement, Bioinformatics Center, Nanjing Agricultural University, Nanjing 210000, China
[5]Forestry and Pomology Research Institute, Shanghai Academy of Agricultural Sciences, Shanghai 201403, China
[6]Grandomics Biosciences, Wuhan 430070, China
[7]College of Horticulture and Landscape, Yangzhou University, Yangzhou 225009, China
[8]College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, China
[9]School of Landscape Architecture, Beijing Forestry University, Beijing 100083, China
[10]Shanxi Agricultural University, Jinzhong 030801, China
[11]National Key Laboratory for Tropical Crop Breeding, College of Breeding and Multiplication, Sanya 572025, China
[12]College of Landscape Architecture, Beijing University of Agriculture, Beijing 102206, China
[13]College of Horticulture, Shenyang Agricultural University, Shenyang 110866, China
[14]School of Advanced Agriculture and Bioengineering, Yangtze Normal University, Chongqing 408100, China
[15]College of Horticulture, Gansu Agricultural University, Lanzhou 730070, China
[16]Institute of Biotechnology, Gansu Academy of Agricultural Sciences, Lanzhou 730070, China
[17]School of Agriculture, Yunnan University, Kunming 650091, China
[18]Hunan Institute of Agricultural Environment and Ecology, Hunan Academy of Agricultural Sciences, Changsha 410125, China
[19]Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou 350002, China
[20]These authors contributed equally
*Correspondence: xuejy@njau.edu.cn (J.-Y.X.); ggning@mail.hzau.edu.cn (G.N.); rayming@illinois.edu (R.M.); njteng@njau.edu.cn (N.T.)

Received: July 8, 2024; Accepted: October 20, 2024; Published Online: October 24, 2024; https://doi.org/10.1016/j.xinn.2024.100726

Citation: Xu S., Chen R., Zhang X., et al., (2024). The evolutionary tale of lilies: Giant genomes derived from transposon insertions and polyploidization. The Innovation **5(6)**, 100726.

Dear Editor,

Lily (*Lilium* spp.), known as the "king of bulbous flowers," has high ornamental and medicinal value due to its attractive, fragrant flowers and nutritious bulbs. The *Lilium* genus comprises approximately 115 perennial bulbous herbal species distributed in the Northern hemisphere, including 55 species and 18 varieties that were discovered in China. The rich diversity of lilies in China likely contributes to their extensive utilization and provides valuable resources for scientific studies. However, their giant genomes pose a challenge to high-quality genome assembly, and there is no reference genome for this lineage. To date, only a few giant genomes have been published due to technical difficulties and high costs.[1] The substantial diversity in genome size among organisms is of fundamental biological significance. However, the correlation between organismal complexity and genome size remains tenuous.[2] Sequencing and assembling large/complex genomes remain challenging due to issues such as polyploidy, high heterozygosity, and high repeat ratios. Because they have significantly larger genomes than other eukaryotes, almost all *Lilium* species could serve as ideal models in which to study the relationship between organismal complexity and genome size. *L. davidii* var. *unicolor*, the only edible sweet lily variety, has been cultivated for ~150 years in Lanzhou, China, providing an important source of income for local farmers. We selected this variety for genome sequencing to acquire valuable data for studying the giant genomes of lilies and, thereby, facilitate their genetic improvement and enhance breeding efforts.

CHROMOSOME-SCALE ASSEMBLY OF THE LILY MEGA-GENOME

We estimated the genome size of *L. davidii* var. *unicolor* to be approximately 38.01 Gb by flow cytometry analysis and 37.62 Gb with 2.18% heterozygosity by *K-mer* distribution analysis. Karyotype analysis showed that *L. davidii* var. *unicolor* is diploid, with 12 pairs of giant chromosomes. To assemble this extremely large genome, we generated 3.32 Tb of Illumina short-read and 2.25 Tb of Nanopore long-read data. We employed the optimized NextDenovo pipeline, which is suitable for giant genomes, to produce a preliminary non-redundant contig-level assembly of 36.68 Gb (13,068 contigs, N50 = 7.72 Mb). Subsequently, 4.45 Tb of High-throughput Chromosome Conformation Capture (Hi-C) data were produced for physical mapping, and after four rounds of manual adjustment, 96.99% of the contigs were anchored onto the 12 pseudochromosomes, corresponding to 12 chromosomes (Figure 1A).

We assessed the completeness and accuracy of the assembly using various methods. First, we measured the relative physical lengths of 12 sets of chromosomes from three somatic cells. The lengths of all chromosomes in the assembly were proportional to the observed physical lengths. We then realigned the Illumina and Nanopore reads to the genome, with high mapping rates of 97.80% and 99.10%, respectively. The long terminal repeat (LTR) assembly index (LAI) value was estimated to be >10, and base-level accuracy analysis yielded a quality value (QV) of 30.18, far exceeding the standard for a reference-level assembly. We also identified telomeric sequences in two pseudochromosomes at both ends and five pseudochromosomes at one end, confirming the high completeness of the assembly. Genome annotation yielded 87,501 protein-coding genes, and 78,348 (89.54%) genes could be functionally annotated. Finally, benchmarking universal single-copy orthologs (BUSCO) evaluation captured 94.90% completeness in the assembly. These findings highlight the high completeness, accuracy, and contiguity of this lily genome assembly.

RECENT TE EXPANSIONS AND POLYPLOIDIZATION CREATED THE SUPER-LARGE GENOME OF *L. DAVIDII* VAR. *UNICOLOR*

The main factors affecting genome size are the accumulation of repetitive DNA sequences and polyploidization. We identified an extremely high percentage of repetitive sequences (88.31%) in the lily genome through annotation, with transposable elements (TEs) accounting for 84.19% of the genome. Among these TEs, LTR retrotransposons (LTR-RTs) are the major components (64.40%), with Copia and Gypsy accounting for 16.62% and 31.53%, respectively. An estimation of the insertion time of the LTR-RTs revealed the sharply accelerated accumulation of LTR-RTs over the past 5 million years. The accumulation rate of Copia reached a peak ~1.65 million years ago (mya), showing a burst of Copia insertions, whereas the burst of Gypsy insertions occurred more recently, at approximately 0.89 mya (Figure 1B). The explosive insertions of Copia and Gypsy elements near the peak accounted for 29.6% and 22.1% of total TE

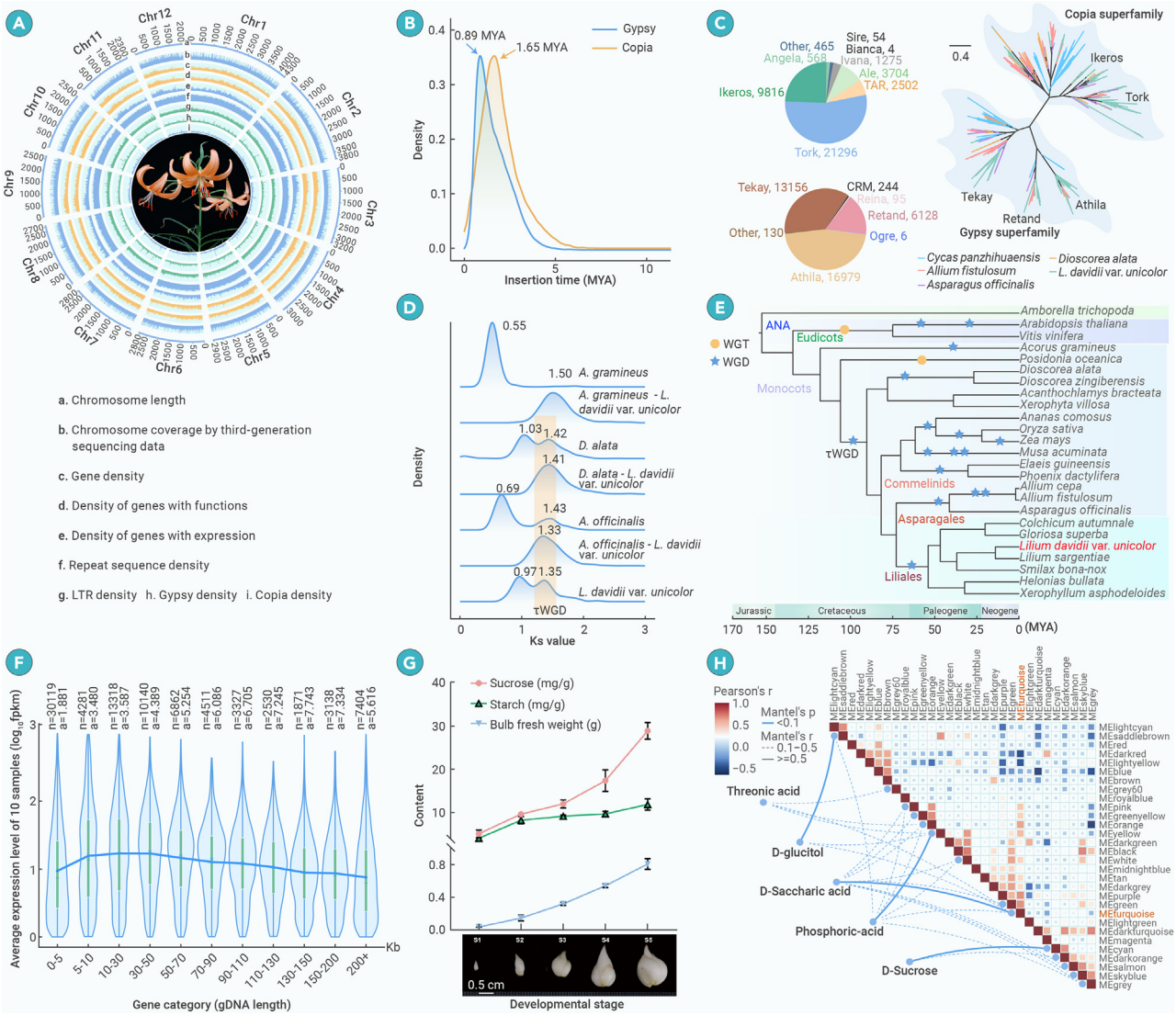

Figure 1. The lily genome and multi-omic analysis (A) Genomic features of *L. davidii* var. *unicolor*. (B) Estimated insertion times of Gypsy and Copia retrotransposons. (C) Phylogeny and classification of Gypsy and Copia retrotransposon subclasses. (D) Identification of WGD events. (E) Phylogenetic position and divergence time of the lily. "τ" refers to the WGD shared by all monocots but Acorales and Alismatales. (F) Expression patterns of genes with different lengths and intron numbers in the lily genome. (G) Fresh weight and starch and sucrose contents during lily bulblet development. (H) Correlation analysis between glycolytic metabolites and co-expression modules.

insertions, respectively. Therefore, these recent intensive TE insertion events largely explain the giant genome of this lily.

Phylogenetic analysis indicated that both the Gypsy and Copia LTR-RTs in lily comprise multiple subclasses (Figure 1C). Of these, two Gypsy subclasses (Athila and Tekay) and two Copia subclasses (Tork and Ikeros) outnumbered the others, suggesting they contributed significantly to the expanded genome size. Different subclasses of TEs have expanded in different plants during their evolution, with Athila, Retand, Tekay, and Tork having undergone explosive expansion in lily (Figure 1C). Some subclasses (such as Athila) are reported to show target-site bias in pericentromeric heterochromatin and thus suppress recombination, which in theory should have resulted in lower LTR-RT removal rates and facilitated genome expansion.[3]

Apart from TE insertions, whole-genome duplications (WGDs) might also directly contribute to the expansion of genome size. We therefore analyzed WGDs that have occurred in the lily genome. Analysis of the age distributions of synonymous substitution sites (Ks) for paralogs retained in collinear regions showed two signature peaks at 1.35 and 0.97, indicating that two rounds of WGDs occurred in the lily genome (Figure 1D). Intergenomic collinearity analysis revealed ratios of 4:2 between lily and *Acorus gramineus*, 4:4 between lily and *Asparagus officinalis*, and 4:4 between lily and *Dioscorea alata*. Considering that *A. gramineus* experienced only one WGD and *A. officinalis* and *D. alata* each experienced two WGDs,[4,5] the ratios of the intergenomic collinearity blocks support the notion that two WGDs occurred in the lily genome. Based on the positions of the Ks peaks representing speciation events (Figure 1D), we determined that both WGDs in lily occurred after its divergence from *A. gramineus* (1.50), with one ancient WGD (τ) being shared with Asparagales and Dioscoreales (as evidenced by a larger Ks peak value than the speciation peaks) and one more recent WGD having occurred independently in the lily lineage (showing a smaller Ks peak value compared to the speciation peaks).

Using genes from 563 low-copy orthogroups, our reconstructed monocot phylogeny resolved lily as the sister group of Asparagales, with a divergence time dating back to 72.7 mya (Figure 1E), which is consistent with the findings of most studies employing nuclear genes.[6] By contrast, the hypothesis that Liliales and Asparagales are successive sisters to commelinids is mainly supported by studies using plastid genomes,[7] possibly reflecting the cyto-nuclear conflict arising from different modes of inheritance (biparental vs. uniparental). Although the onion and garlic genomes from Asparagales both underwent two more rounds of WGDs,[8] their genomes are less than half the size of the lily genome. This finding suggests that the lily genome has been better preserved after the TE insertion events and WGDs, reflecting distinct evolutionary modes for lily and the two Asparagales species. Unequal recombination is considered to be a major LTR-RT removal mechanism in plants[9] and has been used to explain the formation of giant genomes: giant genomes have a significantly lower rate of unequal recombination, in theory leading to a slower LTR removal rate, the

continuous accumulation of LTRs, and a giant genome.[10] Perhaps the lily genome has an even lower rate of unequal recombination than other plant genomes.

LILY'S ULTRA-LONG GENES RESULT FROM ULTRA-LONG INTRONS AND SHOW A NEGATIVE CORRELATION BETWEEN EXPRESSION LEVEL AND GENE LENGTH

Lily contains a large proportion of long genes (57.61 kb on average), with genes longer than 50 kb (defined as ultra-long genes) accounting for 33.88% of all annotated genes. However, the coding sequences of lily genes have an average length of only 847.17 bp, comprising an average of 3.97 exons and 213.72 bp per exon, which are comparable to those of other plants with significantly smaller genomes. These findings suggest that extremely long introns are the primary cause for the ultra-long genes in lily. Indeed, the average intron length of lily genes is ~19.13 kb, the second longest among all published plant genomes (second only to that of *Cycas panzhihuaensis*).

In the gymnosperm *Pinus tabuliformis*, large genes with long introns exhibit higher expression levels than the others, which was attributed to greater chromatin accessibility.[10] To investigate the expression patterns and factors influencing the long genes in the lily genome, we conducted a comprehensive analysis of several gene characteristics. Among all investigated factors, gene length showed significant but variable correlations with expression levels (Figure 1F): a positive correlation between gene length and expression levels was observed for genes shorter than 50 kb, but a negative correlation was observed for genes that exceeded 50 kb. This contrasting trend in gene expression was uniquely observed in *L. davidii* var. *unicolor*, distinguishing it from other plants possessing long genes. Additionally, the *L. davidii* var. *unicolor* genes with the highest expression levels typically contained 3 or 4 introns (Figure 1F).

Overall, lily genes exhibit an expression pattern distinct from that of *P. tabuliformis* genes: a trend of increasing expression for genes shorter than 50 kb and decreasing expression for genes longer than 50 kb. Perhaps 50 kb represents a transition point that limits the efficiency of transcription or intron splicing, resulting in lower efficiency for longer genes.

LILY BULBLET DEVELOPMENT AND CARBOHYDRATE METABOLISM

Lily bulbs are nutrient-storing organs that serve as important resources for the pharmaceutical and food industries in East Asia. To examine nutrient accumulation and the underlying mechanisms during lily bulblet development, we collected bulblets at five developmental stages and subjected them to comprehensive cytological, transcriptomic, and metabolomics analyses. Starch and sucrose accumulation in bulblets was detected throughout bulblet development (Figure 1G), and a significant number of genes in the glycolytic metabolic pathway were highly expressed exclusively in bulblets, demonstrating clear organ-specific expression patterns. Furthermore, wide-targeted metabolomics detected a total of 870 metabolites in lily bulblets across the five developmental stages, including carbohydrates, lipids, phenolic acids, and so on, demonstrating rich metabolic diversity. Metabolome-transcriptome correlation analysis revealed a significant association between carbohydrate metabolites and one gene expression module (turquoise), which likely includes genes encoding enzymes in carbohydrate and starch metabolic pathways (Figure 1H).

In conclusion, we successfully assembled a chromosome-level lily genome, the largest plant genome published to date. Our analysis of this genome provided an in-depth understanding of the features and mechanisms of giant genomes, as well as other characteristics, including prevalent ultra-long genes and their expression profiles. The high-quality lily reference genome will lay the foundation for genetic and molecular research on lilies, including population genetics and evolutionary genomics to trace the origin and evolutionary history of different lily cultivars and genome-wide association studies to identify key genes associated with important agronomic traits. The lily reference genome will also facilitate the mining of the abundant genetic resources within lilies and the identification of candidate genes associated with key traits and provide essential genetic resources for molecular breeding.

DATA AND CODE AVAILABILITY

All genomic and transcriptomic data for *Lilium davidii* var. *unicolor* have been deposited in the China National GeneBank Database (CNP0005511). More detailed results of this paper are provided at https://doi.org/10.6084/m9.figshare.27222642.

REFERENCES

1. Sun, Y., Shang, L., Zhu, Q.H., et al. (2022). Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci. ***27***: 391–401.
2. Hidalgo, O., Pellicer, J., Christenhusz, M., et al. (2017). Is there an upper limit to genome size? Trends Plant Sci. ***22***: 567–573.
3. Peterson-Burch, B.D., Nettleton, D., and Voytas, D.F. (2004). Genomic neighborhoods for Arabidopsis retrotransposons: a role for targeted integration in the distribution of the Metaviridae. Genome Biol. ***5***: R78.
4. Ren, R., Wang, H., Guo, C., et al. (2018). Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. Mol. Plant ***11***: 414–428.
5. Harkess, A., Zhou, J., Xu, C., et al. (2017). The asparagus genome sheds light on the origin and evolution of a young Y chromosome. Nat. Commun. ***8***: 1279.
6. Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., et al. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. Nature ***574***: 679–685.
7. Li, H.T., Luo, Y., Gan, L., et al. (2021). Plastid phylogenomic insights into relationships of all flowering plant families. BMC Biol. ***19***: 232.
8. Hao, F., Liu, X., Zhou, B., et al. (2023). Chromosome-level genomes of three key *Allium* crops and their trait evolution. Nat. Genet. ***55***: 1976–1986.
9. Cossu, R.M., Casola, C., Giacomello, S., et al. (2017). LTR retrotransposons show low levels of unequal recombination and high rates of intraelement gene conversion in large plant genomes. Genome Biol. Evol. ***9***: 3449–3462.
10. Niu, S., Li, J., Bo, W., et al. (2022). The Chinese pine genome and methylome unveil key features of conifer evolution. Cell ***185***: 204–217.e14.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (grant no. 2019YFD1000400), the "JBGS" Project of Seed Industry Revitalization in Jiangsu Province (JBGS[2021]093), the Project for Crop Germplasm Resources Conservation of Jiangsu Province (grant no. 2021-SJ-011), the Seed Industry Innovation Project (2021NK1005) from the Department of Science and Technology of Hunan Province, the High-level Key Discipline Construction Project for Traditional Chinese Medicine—Resources Science of Chinese Medicinal Materials from National Administration of Traditional Chinese Medicine, the Fundamental Research Funds for the Central Universities from Nanjing Agricultural University, and the Pilot Incubation Project of Modern Flower and Horticultural Industry Technology from Nanjing Oriole Island Modern Agricultural Development Co., Ltd. This work was supported by the high-performance computing platform of Bioinformatics Center, Nanjing Agricultural University.

DECLARATION OF INTERESTS

The authors declare no competing interests.