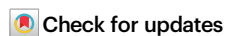


Genomic and epigenomic insight into giga-chromosome architecture and adaptive evolution of royal lily (*Lilium regale*)

Received: 5 November 2024

Accepted: 13 June 2025

Published online: 01 July 2025

Jie Sun^{1,3}, Xiangfeng Wang^{1,3}, Ke Wang^{1,2}, Dian Meng¹, Yu Mu¹, Lili Zhang¹, Jingxuan Wang¹, Gang Yao¹ & Li Guo¹✉

Lilies are popular ornamental and medicinal plants with gigantic genomes. Due to the challenge of assembling complex giga-genomes, our understanding of the genetic architecture, epigenetic regulation and evolution of large-genome plants such as lily remains limited. Here, we report a high-quality chromosome-level 35.6 Gb reference genome of royal lily (*Lilium regale*), a parent of many modern lily cultivars, using PacBio HiFi and Hi-C sequencing data. We show that genome expansion of *L. regale* is mainly caused by extensive proliferation of transposable elements resulting in long intergenic and intronic regions, along with whole-genome duplications and tandem repeats. *L. regale* genome is repeat-rich (80.06%) encoding abundant large genes (>10 Kb) with long introns that account for ~90% length of 67,862 genes encoded. Phylogenomics reveals significant gene family expansion related to defense response and biosynthesis of terpenoids, reflecting its adaptation strategy. Through multi-omic analysis, we reveal how transposable element activity and epigenetic regulations may impact transcription, alternative splicing, and three-dimensional organization, which contribute to its adaptive evolution. Collectively, this significantly improved lily genome assembly and annotation will serve as an essential resource for research on lily genetics, breeding, conservation biology, and angiosperm genome evolution.

Plants have the largest variation of genome sizes ranging from a few megabases to hundreds of gigabases. Genome sizes have profound implications on plant biology, influencing their growth, development, genome plasticity, and ecological adaptations^{1–3}. Despite its significance, the mechanisms underlying the emergence of large plant genomes, particularly how they maintain functional integrity while organizing gigabase-scale linear DNA within the nucleus, remain largely enigmatic. Proliferation of transposable elements (TEs), particularly long terminal repeat retrotransposons (LTR-RTs) that constitute a major portion of many plant genomes, can lead to substantial genome expansions and size variations⁴. On the other hand, host genomes often employ

epigenetic mechanisms to suppress the transcription and proliferation of retrotransposons⁴. Regions rich in retrotransposons, particularly heterochromatinized areas near centromeres and telomeres, are thought to inhibit crossing-over recombination^{5,6}. Two recombination mechanisms can remove LTR-RTs from host genomes: unequal recombination (UR), also known as intra-chromosomal homologous recombination, and illegitimate recombination. UR is a common process in angiosperms, effectively counteracting genome expansion by eliminating LTR-RTs. In contrast, in conifers, these removal mechanisms are largely inefficient, which may account for the immense genome sizes observed in many conifer species⁴.

¹Shandong Key Laboratory of Precision Molecular Crop Design and Breeding, Peking University Institute of Advanced Agricultural Sciences, Shandong Laboratory of Advanced Agricultural Sciences at Weifang, Weifang, Shandong, China. ²College of Life Sciences, Shandong Agricultural University, Tai'an, Shandong, China. ³These authors contributed equally: Jie Sun, Xiangfeng Wang. ✉e-mail: li.guo@pku-iaas.edu.cn

In addition, alternative splicing (AS) increases gene regulatory complexity by enabling the production of multiple protein isoforms from a single gene⁷. In TE-rich genomes, AS events may be influenced by TE insertions, contributing to transcriptomic diversity and facilitating adaptation to environmental stresses^{8,9}. Nonetheless, comprehensive genome-wide analyses of AS in the context of TE dynamics are scarce, particularly in plants with exceptionally large genomes. Three-dimensional (3D) genome organization, including chromosomal territories, compartments, and topologically associated domains (TADs), enables distal *cis*-regulatory elements to spatially interact with genes nearby, playing a crucial role in gene regulation^{10–12}. Current understanding of plant 3D chromatin structure is primarily derived from studies of smaller genomes, such as *Arabidopsis*¹³ and various crop species^{14–16}. The 3D genome architecture of non-model and gigagenome plants remains largely unexplored but will be essential to unraveling their structural and regulatory landscapes.

Lilium regale, renowned for its long-standing horticultural history and exceptional traits as the “royal lily”, has not yet had its genome sequenced and assembled at high quality. Historically, hybridization among *L. regale* and various lily germplasms has produced new varieties with enhanced stress tolerance, disease resistance, and fragrance¹⁷. The assembly of *L. regale* is imperative, given its value as a genetic resource for breeding and its remarkable resistance to biotic and abiotic stresses^{18,19}. Nevertheless, within the Liliales, numerous members, including lilies, tulips, and fritillaries, possess large genomes. Genome sizes of lilies were typically estimated between 30 and 48 Gb²⁰, which has posed significant challenges for genome assembly and has impeded studies on lily genomics and evolution. Only a few chloroplast and mitochondrial genomes^{21,22}, along with a reference genome assembly for David’s lily (*L. davidii*), have been reported²³. Advancements in long-read sequencing technologies, reduced costs, and improved genome assembly algorithms have recently enabled the sequencing of several organisms with giant genomes, such as African lungfish (45 Gb)²⁴, axolotl (32 Gb)²⁵, Antarctic krill (48 Gb)²⁶, Chinese pine (25.4 Gb)²⁷, and David’s lily (36.8 Gb)²³. However, assembling gigabase-scale genomes remains one of the most computationally intensive tasks in genomics due to technical complexities and the substantial computational and economic costs involved.

In this study, we de novo assembled a 35.6 Gb *L. regale* genome, one of the largest plant genomes sequenced to date, achieving high contiguity and quality in the Liliaceae. Through further comprehensive analyses, we determined that epigenetically regulated TE expansions are the primary drivers of the *L. regale* genome’s enlargement, revealing how epigenetic regulation may have orchestrated the maintenance and functionality of this extensive genomic landscape. Our findings provide insights into the evolutionary strategies employed by lilies to manage the dual pressure of TE proliferation and environmental adaptation.

Results

Assembly and annotation of the gigantic genome of *L. regale*

The *Lilium regale* accession MJ-PKU01, with a karyotype of $2n = 2 \times = 24$ (Supplementary Fig. 1a), was used for genome sequencing and assembly. Based on *k*-mer frequency analysis of 4.40 Tb (~124×) NGS reads (Supplementary Table 1), an estimated genome size of 35.8 Gb was revealed, respectively, with heterozygosity calculated as 3.32% (Supplementary Fig. 1b, c). In total, 806 Gb (~23×) PacBio HiFi reads were generated and assembled. With redundant sequences removed, we got the HiFi assembly of 35.6 Gb containing 3608 primary contigs with a N50 of 32.2 Mb (Table 1). As genome size increases, a general decline in contig N50 was observed across published large plant genomes (Fig. 1a and Supplementary Table 2). Despite the enormous size and high heterozygosity of the *L. regale* genome, our assembly achieved a high contig N50, improved over the recently published *L. davidii*²³ and *L. sargentiae*²⁸ assembly (Table 1). Further scaffolding

using 5.46 Tb (~154×) Hi-C reads (Fig. 1b and Supplementary Figs. 2 and 3) anchored 1818 contigs (34.4 Gb) onto 12 chromosomes ranging from 2.0 to 4.3 Gb (Supplementary Table 3), accounting for ~96.8% of the 35.6 Gb assembly. Furthermore, with the conserved repeat motif AAACCCT/AGGGTTT²⁹, 7 telomeres were detected (Supplementary Table 4). Finally, we obtained the 35.6 Gb high-quality chromosome-level assembly of *L. regale* genome. We extensively evaluated genome assembly using multiple methods. Hi-C interaction maps showed no apparent large-scale mis-assembly or any misplacement of contigs (Supplementary Figs. 2 and 3). HiFi and NGS read mapping rate of 99.55% and 95.44% (Table 1), a 63.8 value of QV calculated by Merqury³⁰ based on HiFi reads (Supplementary Table 5) and a BUSCO (Benchmarking Universal Single-Copy Orthologs) score of 96.16% (Table 1), together suggesting the high accuracy and completeness of our *L. regale* genome assembly. Genome annotation by integrating ab initio prediction, homologous proteins, and transcriptome-based evidence from different tissues using ISO-seq and RNA-seq (Supplementary Table 6) predicted 67,862 high-confidence gene models, of which 77.64% were annotated by eggNOG-mapper^{31,32}, Gene Ontology³³ (GO) and Kyoto Encyclopedia of Genes and Genome³⁴ (KEGG), and 52.27% were expressed in at least one tissue (TPM > 0).

Phylogenomics reveals the evolutionary history of the lily genome

To understand the evolutionary history of royal lily, we conducted a phylogenomic analysis of 16 species, including our *L. regale* genome and 15 other published genomes, including *L. davidii*²³, *L. sargentiae*, and *Gloriosa superba*²⁸. Ortholog identification and clustering using Orthofinder³⁵ classified 698,581 genes into 62,451 orthogroups with four categories: multi-copy, single-copy, unique-copy and others (Fig. 1d), 370 of which were composed of low-copy orthogroups (1 or 2 genes per orthologroup). In *Lilium*, 30,012 unique genes were identified, primarily associated with transcription regulation and post-transcriptional gene silencing (Supplementary Fig. 4). In contrast, *L. regale* had 15,691 unique genes, which were enriched in nucleotide excision repair and mRNA 3'-UTR binding. These genes may have evolved as adaptations to the exceptionally large genomes of *Lilium* (Supplementary Fig. 5). A phylogenetic tree was constructed using MCMCTREE³⁶ with the 370 low-copy orthogroups, with divergence times estimated and calibrated using fossil records from TimeTrees³⁷ (Supplementary Fig. 6a), showing that *G. superba* and (*L. davidii*, (*L. regale*, *L. sargentiae*)) diverged at ~95.6 million years ago (Mya). *L. regale* and *L. sargentiae* are more closely related than each with *L. davidii*, consistent with previous transcriptome-based analysis²⁸. Furthermore CAFE³⁸ analysis revealed significantly expanded orthogroups in *Lilium* primarily involved in terpenoid biosynthesis (Supplementary Fig. 5a). *L. regale* expanded orthogroups were also enriched in MAPK signaling pathway and defense responses mediated by cell wall thickening (Supplementary Fig. 5b). This reflected the genome evolution in *L. regale* to cope with environmental stress during its independent evolution^{39,40}.

Whole-genome duplications (WGDs) facilitate genome evolution and species adaptation of flowering plants⁴¹. A WGD (τ) event is shared by Asparagales and commelinid monocots and occurred after the split of Alismatales⁴². In addition, monocots such as rice, maize, and banana have undergone independent WGDs after the τ event⁴³. Interspecies comparison showed a 4:4 synteny ratio (Fig. 1e) between *L. regale* and *Dioscorea alata*, a species having two WGD events (the τ event and another independent WGD)⁴⁴, suggesting that *L. regale* also has undergone two WGD events. Furthermore, a 4:1 synteny ratio between *L. regale* and the ancestral monocot karyotype⁴⁵ (Supplementary Fig. 6b) supports the hypothesis that *L. regale* has undergone two additional WGDs since its divergence from the monocot ancestor.

Table 1 | Statistics for genome assembly and annotation of *Lilium regale*

Statistics	<i>Lilium davidii</i> ²³	<i>Lilium sargentiae</i> ²⁸	<i>Lilium regale</i> MJ-PKU01
Assembly size (bp)	36,680,000,000	47,092,191,210	35,569,194,365
Number of contigs	13,068	85,041	3608
Number of scaffolds	3853	35	1409
Number of chromosomes	12	–	12
Scaffold N50 (bp)	2,860,000,000	2,845,141,673	2,714,556,596
Contig N50 (bp)	7,720,000	929,352	32,225,586
Total length of pseudomolecules (bp)	–	–	34,433,330,367
GC content (%)	34.94	–	32.26
Repeat content (%)	88.31	82.32	80.06
Number of protein-coding genes	87,501	47,139	78,147
Number of high-confidence protein-coding genes	–	–	67,862
Average gene length (bp)	57,607	42,804	19,940
HiFi mapping rate (%)	–	–	99.55
ONT mapping rate (%)	99.10	–	–
NGS mapping rate (%)	97.80	–	95.44
BUSCO [†] of genome (%)	94.90	90.77	96.16
BUSCO [†] of protein (%)	79.50	88.30	88.20
QV (quality value)	30.18	–	63.80

[†]The lineage dataset is: embryophyta_odb10 (*L. davidii* and *L. sargentiae*) (Creation date: 2024-01-08, number of genomes: 50, number of BUSCOs: 1614); liliopsida_odb10 (*L. regale*) (Creation date: 2024-01-08, number of genomes: 15, number of BUSCOs: 3236).

We then examined the genome duplication history of *L. regale* by analyzing the distribution of synonymous substitution rates (*K*s) for syntenic gene pairs identified in synteny analysis (Fig. 1f). *K*s peaks correspond to genome duplication events (self-synteny) or species divergence (interspecies synteny). For *L. regale* syntenic paralogs, we observed three *K*s peaks (Fig. 1f). The first peak (*K*s ≈ 1.5) shared with *D. alata* represented the τ WGD (~125 Mya) widely present in monocots, including *L. regale* and *D. alata*. The second peak (*K*s ≈ 1.09) represented a WGD event in *L. regale* likely occurring at a similar time (~90.1 Mya) with the second WGD in *D. alata* judging by the two close peaks (Fig. 1f). Because the *K*s for the two peaks were both smaller than the *K*s for divergence peak of *L. regale* and *D. alata*, this second WGD should occur after the two species diverged, suggesting it's an independent WGD for both species (Supplementary Fig. 6a). Furthermore, synteny analysis between the two lilies *L. regale* and *L. davidii* revealed strong gene-level collinearity (Fig. 1g and Supplementary Fig. 6d). Previous studies^{23,28} have shown that *L. davidii* and *L. sargentiae* underwent independent WGD events, suggesting that the three *Lilium* species share a similar WGD history. Together, these analyses provide robust evidence that *Lilium* species has undergone an independent WGD. Finally, the third peak (*K*s ≈ 0.18) indicated a recent burst of gene duplications in *L. regale* likely through tandem or segmental duplications given their appearance as discontinuous fragments (Supplementary Fig. 6c). Tandem duplications are major genomic evolution forces for the microevolution of flowering plants, facilitating their adaptation to environmental changes⁴⁶. *L. regale* tandem-duplicated and WGD genes were enriched in biological processes such as stress tolerance, flavonoid biosynthesis, and monoterpenoid biosynthesis (Supplementary Fig. 7), indicating a specialization of terpenoid biosynthetic pathways (probably as sources of flower scent). Overall, we deciphered the genome evolution history in *L. regale* and how lilies may have evolved to cope with environmental stresses and herbivores through various mechanisms (Supplementary Figs. 4, 5, and 7).

Transposon-driven expansion and evolution of *L. regale* genome
Repeat annotation in *L. regale* identified 27.9 Gb repetitive sequences, occupying 80.06% of the genome (Supplementary Table 7), showing

their substantial contribution to *L. regale* genome expansion, similar to *L. davidii* (88.31%) and *L. sargentiae* (82.32%)^{23,28}. Most of the repetitive sequences were TEs (27.5 Gb), especially 19.8 Gb long terminal repeat retrotransposons (LTR-RTs) taking 72.3% of total TEs (Supplementary Fig. 8a and Supplementary Table 7). Distinct from humans whose genome contained a higher proportion of LINES (Long Interspersed Nuclear Elements) and SINES (Short Interspersed Nuclear Elements), the proportion of Ty3/Gypsy elements increased with genome size in plants (Fig. 2a). However, the fraction of different TE types in *L. regale* was comparable to smaller genomes such as maize and onion (Supplementary Table 8), suggesting that *L. regale* genome expansion was caused by amplification of all TE types. Despite a Gypsy-rich genome, the numbers of intact Gypsy (34,472) and Copia (35,771) elements were comparable (Fig. 2b and Supplementary Fig. 8b), with Tekay and Athila as the most abundant LTR/Gypsy subfamilies (Fig. 2b), which was consistent with the report in *L. davidii* and *L. sargentiae*^{23,28}.

Given the extensive insertions of LTR-RTs, we explored their genome-wide insertion dynamics and estimated the timing of insertions. A total of 79,821 intact LTRs (650.88 Mb), 4,249,484 truncated LTRs (7.13 Gb), and 19,230,536 solo LTRs (9.11 Gb) were identified in *L. regale*, comprising 3.28%, 35.90%, and 45.88% of the total LTR content (19,858.9 Mb), respectively (Supplementary Table 9), suggesting the rapid recombination of LTRs. All chromosomes exhibited high-density peaks, especially toward both ends, indicating LTR-RT hotspots (Fig. 2c), mainly composed of dense insertions from the same LTR lineage within a short timeframe. Among them, the hotspot bursts involving Gypsy-LTR elements were mainly derived from insertions of the Tekay lineage (Supplementary Fig. 9), consistent with whole-genome insertion of Tekay (Fig. 2d). By contrast, the hotspots involving Copia elements (Supplementary Fig. 8) showed no dominant subfamily types conserved among chromosomes (Supplementary Fig. 8f). In summary, *L. regale* genome expansion was accoupled with LTR insertion hotspots mostly towards the ends of chromosomes (Fig. 2c).

Unequal recombination between LTRs removes intermediate sequences and the formation of solo-LTRs⁴. In *L. regale*, the extensive formation of solo-LTRs implied a high recombination rate

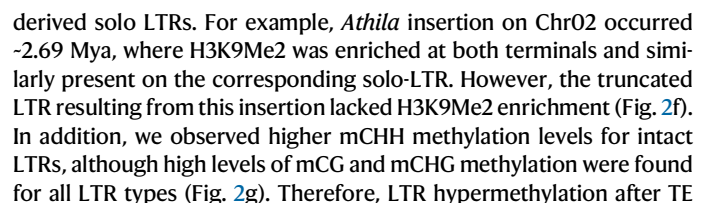


Fig. 1 | Genome assembly and phylogenomics reveal the evolutionary history of *Lilium regale* genome. **a** A comparison of large plant genomes (>10 G) published in recent years shown by a dot plot of genome size and contig N50. Different shapes represent genomes published in different years. The central line represents the linear regression fit ($R^2 = 0.14$), and the gray shaded area denotes the 95% confidence interval (CI) around the regression line ($n = 13$ species). Error bands were calculated using the standard error of the estimate. **b** Hi-C interaction heatmap of *L. regale* genome showing twelve pseudomolecules. Blue boxes represent scaffolds (pseudomolecules) Chr01–Chr12, arranged in decreasing size. **c** Circos plot showing the genomic features for the chromosome-level genome assembly of *L. regale*. Track A to H: chromosomes, GC content, gene density, TE density, Gypsy-LTR density, Copia-LTR density, exon density and intron density. The inner circle showing the flower of *L. regale* (source: author's own photograph). LTR long terminal repeat, TE transposable element. **d** Comparison of genes among 16 related species. Multi-copy genes: Genes belonging to orthogroups in which at least one species has more than one gene, while all

species have at least one gene. Single-copy genes: from the orthogroups in which the number of genes per species is 1. Unique-copy genes: from the orthogroups in which the number of genes is greater than 0 but only from one species. Other genes: from other orthogroups. **e** Dot plot showing the syntenic relationship of homeologous chromosomes between *L. regale* and *D. alata*. Representative synteny blocks signaling the whole-genome duplication events are color highlighted. The blue boxes represent homologous segments of *L. regale* in *D. alata*. Meanwhile, homologous segments from the same homologous segment sets are linked by blue lines. Similarly, the red boxes and lines denote homologous segments of *D. alata* in *L. regale*. **f** Synonymous substitution rate (Ks) density distributions of syntenic paralogs and orthologs. The WGD events and their associated time were labeled accordingly. **g** Syntenic comparison of *L. regale* with *L. davidii*. Numbers represent chromosome IDs, and the lines indicate syntenic blocks, which are conserved genomic regions with shared gene order. Source data are provided as a Source Data file.

insertion may help restrict the activity of TEs, perhaps as a genome defense mechanism.

Long genes and long introns in *L. regale*

L. regale genes were categorized into four groups: <1 kilobase (kb) ($n = 17,887$), 1–10 kb ($n = 29,597$), 10–100 kb ($n = 16,939$), and >100 kb ($n = 3,439$), accounting for 0.81%, 7.75%, 46.90%, and 44.54% of the total gene length, respectively (Fig. 3a). Long genes (>100 kb) were mostly enriched in responses to external and biotic stimuli (Supplementary Fig. 10). *L. regale* gene sizes varied substantially ranging from 75 bp to 1.51 Mb (average length: 19,940 bp) (Table 1 and Supplementary Table 10), longer than typical plant genes mainly attributed to the presence of long introns (average 4680 bp) much longer than usual^{23,28,47–49} (Supplementary Table 10) but shorter than *L. davidii* and *L. sargentiae*^{23,28}. As gene length increased, the number and length of exons and introns also increased (Supplementary Fig. 11a, b), with longer genes mainly occupied by intronic sequences (Fig. 3b). The lower GC content of introns compared to exons accounted for the negative correlation between gene length and GC content in *L. regale* (Fig. 3c). In addition, TE insertions (mainly LTRs) accounted for 72.52% of the total intronic length, indicating a major contribution of LTR insertion to ultra-long introns and genes. Interestingly, although *Copia* were about half as many as *Gypsy* genome-wide, their intronic abundance (~37% each) was comparable (Supplementary Fig. 11c), suggesting a more favorable contribution of *Copia* to intron inflation. Furthermore, regardless of tissues, ultra-long genes (>100 kb) and medium-long genes (10–100 kb) were expressed significantly higher than shorter genes (<10 kb) (Fig. 3d), suggesting that long genes were generally more transcriptionally active than shorter ones in *L. regale*. However, ultra-long genes had lower expression than medium-long genes, as was similarly found in *L. davidii*²³.

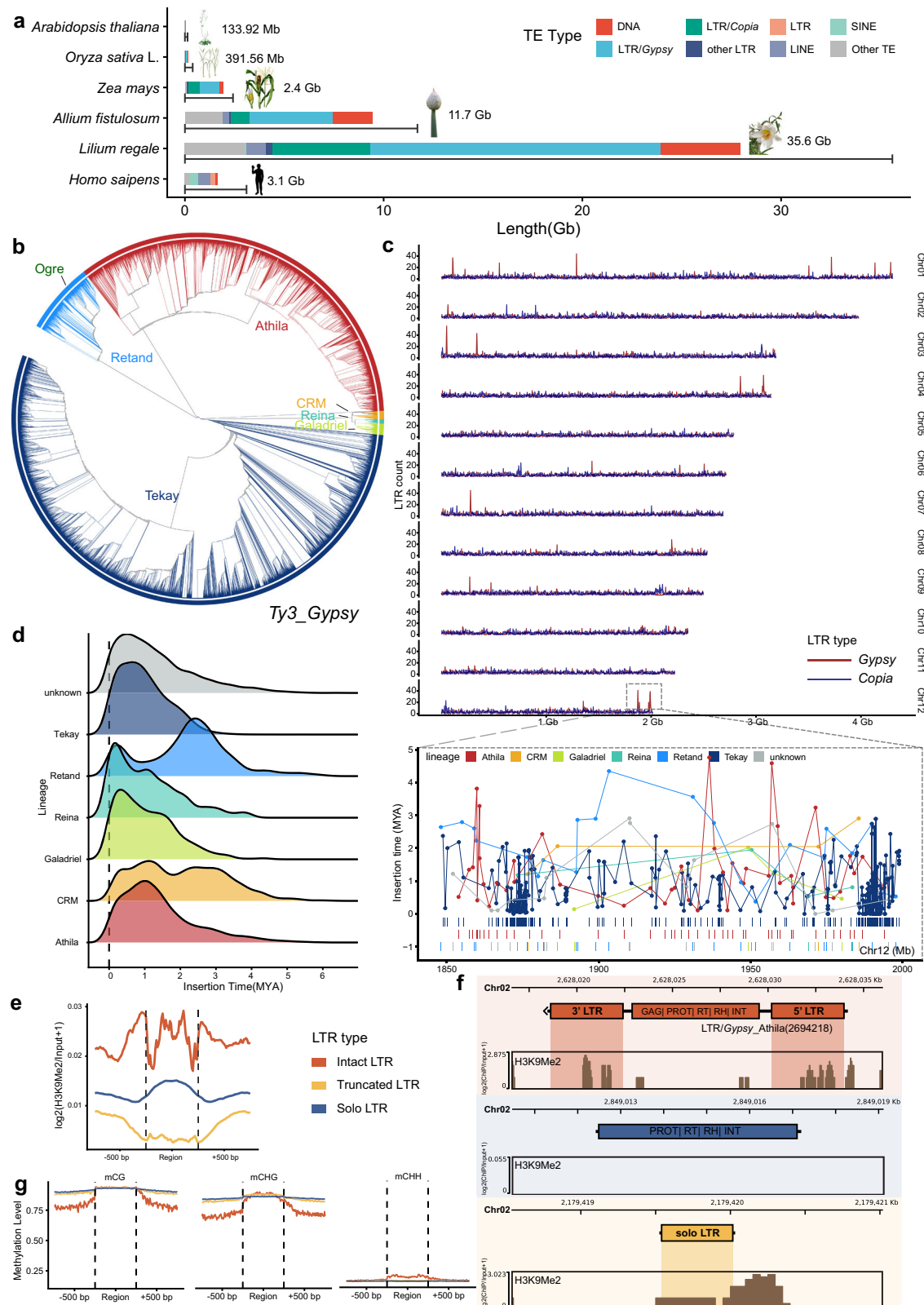
DNA methylation is a key factor regulating transcription^{26,27,50}. Whole-genome bisulfite sequencing profiling of leaves and flowers showed that *L. regale* genome had an overall high DNA methylation, with flowers lower than leaves (Supplementary Fig. 12a). Interestingly, ultra-long genes had the highest methylation levels, followed by medium-sized genes and smaller genes (Fig. 3e), suggesting a negative correlation between gene size and gene-body DNA methylation. For example, an ultra-long gene *LregChr03G072890* (200 kb) on Chr02 encoding a callose synthase was specifically expressed in flowers (Fig. 3f). Unlike its *Arabidopsis* homolog (*AT2G13680*)⁵¹, exons constituted a minor portion of *LregChr03G072890*, with the majority being introns with extensive TE insertions and higher methylation (Fig. 3f). Although its exons displayed substantial CG methylation, CHG and CHH methylation was much lower (Supplementary Fig. 12e, f). Therefore, ultra-long gene formation and transcriptional regulation in *L. regale* are likely associated with intronic TE insertions and DNA methylations.

Alternative splicing landscape of *L. regale* genome

Given that large genes with long introns are dominant in *L. regale*, we detected alternative splicing (AS) using PacBio ISO-seq from four tissues (Supplementary Table 6) to understand how splicing may regulate gene transcription. Seven major AS types were identified with A3 (alternative 3' splice sites) events (32.67%), RI (retention intron) events (27.40%) and A5 (alternative 5' splice sites) events (15.16%), constituting the three largest proportions (Fig. 4a and Supplementary Table 11). Compared to *Arabidopsis*, rice, and wheat, *L. regale* exhibited a much higher proportion of A3 and A5 events but fewer RI events (Fig. 4a and Supplementary Table 11). A3 events occurred more frequently in long genes (>10 kb), whereas RI and SE (skipping exons) events were less frequent in ultra-long genes (>100 kb) compared to shorter genes (Fig. 4b and Supplementary Fig. 13a), suggesting a length-dependent preference for AS events in *L. regale*.

To measure the exon splicing efficiency during different AS events, we calculated the percent spliced in index (PSI)⁵² of all AS events and found that long genes (>10 kb) exhibited higher PSI for A3 events (Supplementary Fig. 13b). Genes were categorized with PSI > 0.8 or <0.2 as high PSI genes (HPGs) and low PSI genes (LPGs), respectively. TE insertions impact AS of specific transcripts through various mechanisms, including intron retention and alternative donor or acceptor splice sites⁵³. We indeed observed a significantly higher TE density near HPGs than LPGs, particularly for LTRs and LINEs (Fig. 4c and Supplementary Fig. 13c), suggesting that TE density was likely associated with gene splicing in *L. regale* genome. Subsequently, we quantified various splicing types for HPGs and LPGs with LTR and LINE insertions and observed that HPG had much more A3 events than LPGs did (Fig. 4d). Therefore, LTR and LINE insertions in *L. regale* are likely associated with A3 occurrence and frequency.

DNA methylation modulates the activity of TEs as a defense mechanism for genomic stability⁵⁴. Considering the observed association of TE insertion with alternative splicing, we studied the effects of DNA methylation on AS events in *L. regale* by comparing the methylation levels near HPGs and LPGs with LTR and LINE insertions. CHG methylation levels in LPGs were significantly higher than in HPGs, primarily associated with intron regions (Fig. 4f and Supplementary Fig. 13e, f). Interestingly, although the TE density near LPGs was lower than in HPGs, the TE methylation of HPGs and LPGs was comparable (Fig. 4c and Supplementary Fig. 13g). This clearly indicated that the higher CHG methylation of LPG introns was independent of TE insertions but owing to non-TE methylation, perhaps collectively contributed to PSI differences among *L. regale* genes. In addition, H3K4Me3 was overall more enriched at the transcription start and end sites of LPGs than those of HPGs (Fig. 4f and Supplementary Fig. 13d). LPG undergoing major splicing events (A3, A5, RI) were expressed higher than HPGs, indicating that genes with lower splicing activity were more transcriptionally active (Fig. 4g). Interestingly, LPGs



enriched with H3K4Me3 were mainly involved in RNA biogenesis and nucleic acid endonuclease activity, suggesting putative involvement of these genes in RNA formation and splicing (Supplementary Fig. 14).

Three-dimensional genomic architecture of *L. regale* genome

A large genome not only requires special mechanisms to ensure proper transcription and splicing, but also presents a significant

challenge to DNA folding inside the nucleolus and faithful transmission of the genetic information during cell division. We analyzed *L. regale* 3D genome using leaf Hi-C data to understand how its gigachromosomes were organized within the nucleus for gene regulation, simulated whole-genome images obtained from 3D modeling demonstrated that its chromosomes were confined within a limited space inside nucleolus (Fig. 5a and Supplementary Fig. 15a).

Fig. 2 | Repeat element composition in *L. regale* genome and LTR insertion history. **a** The repeats content of *L. regale* compared to several representative organisms: *Arabidopsis thaliana*⁴⁷, *Oryza sativa*⁴⁸, *Allium fistulosum*¹²⁴, and *Homo sapiens*¹²⁵. DNA DNA transposons, LINE long interspersed nuclear elements, SINE short interspersed nuclear elements. The black line represents the whole genome sequence length, and the length of the colored bars indicates the total length of different repetitive sequences in the genome. **b** Phylogenetic analyses of intact LTR/Gypsy retrotransposons. **c** The counts of intact Gypsy and Copia on 2.5 Mb bins of 12 chromosomes, and a Tekay-dominant LTR burst on Chr12 1850–2000 Mb

region. Different colored dots represent individual intact LTR insertions of different lineages along the chromosomes, with the colored bars below showing the distribution of different lineages on the chromosomes. **d** The insertion time of intact LTR/Gypsy family in *L. regale*. **e** Enrichment of H3K9Me2 on different types of LTRs. **f**, The intact LTR/Gypsy in Chr02 and its truncated LTR and solo LTR as an example. H3K9Me2 is enriched on LTR repeat region. **g** The DNA methylation ratio (the proportion of 5mC at the given site) on different types of LTRs. **e–g** The figure legends are same. Source data are provided as a Source Data file.

Interestingly, unlike “chromosome territory”¹⁶ of *Arabidopsis*, maize, and wheat (Supplementary Fig. 15b–d), *L. regale* chromosomes appeared much more intertwined, consistent with its strong inter-chromosomal interactions present compared to smaller ones (Supplementary Fig. 15e), suggesting different spatial structural folding patterns between large and small genomes.

Next, we detected complex higher-order chromatin structures, including A/B and compartments, TADs, and chromatin loops in *L. regale* (Supplementary Data 1). Unsurprisingly, A compartments, characterized by higher intron density and more abundant H3K4Me3 enrichment associated with transcription activation, were distinct from B compartments, enriched with H3K27Me3 and H3K9Me2 marking repressive transcription and heterochromatin state, respectively (Supplementary Fig. 16a). Therefore, as for high-level 3D genome epigenetic signature, *L. regale* is similar to most published smaller genomes despite unique chromosome folding. Surprisingly, A compartments had higher GC content than B compartments in lily (Supplementary Fig. 16a), unlike soybean¹⁵ but similar to animals³⁵, reflecting potential differences in codon preference across species.

We also identified a total of 20,485 TADs (Supplementary Data 1) covering the majority of the lily genome, which substantially exceeded the reported for other plant genomes^{56,57}, raising the question of how TAD boundaries are formed and regulated in a large genome as that of the lily. Interestingly, a strong enrichment of H3K9Me2 and a modest enrichment of H3K4Me3 were observed within TAD boundaries (Fig. 5b), which were enriched with genes and LINE and SINE transposons (Fig. 5c and Supplementary Fig. 16b) despite showing an overall depletion of TEs (Supplementary Fig. 16b), suggesting specific types of TEs may be involved in TAD boundary dynamics. To understand the epigenetic regulation at TAD boundaries, we examined the histone modifications of genes flanking regions (GFRs), LINEs, and SINEs at TAD boundaries. GFRs were enriched by more H3K4Me3 than H3K9Me2, and genes with H3K4Me3 enrichment had housekeeping functions such as cytoskeleton and cell membrane (Supplementary Fig. 17a). As for LINEs and SINEs, H3K4Me3 was depleted while H3K9Me2 was significantly enriched at SINEs (Fig. 5d). Overall, H3K9Me2 enrichment at TAD boundaries mainly come from SINEs, while H3K4Me3 enrichment was likely associated with genes (Fig. 5c, d). As TEs are often hotspots for DNA methylation, we further examined the methylation levels of TEs near TAD boundaries and did not find a special pattern for the methylation of LINEs and SINEs enriched on TAD boundaries (Fig. 5e and Supplementary Fig. 16d), which was consistent with the overall depletion of TEs at TAD boundaries (Supplementary Fig. 16f). In summary, SINEs and H3K9Me2 might jointly contribute to the formation or maintenance of TAD boundaries in *L. regale*, as shown by the interval from 386–392 Mb on Chr02 (Fig. 5f), where two TAD boundaries were enriched with SINEs and H3K9Me2.

Genome evolution and epigenetic regulation for *L. regale*

Finally, we examined the consequence of TE invasion on gene functions by analyzing TE-rich (TRRs) and TE-poor regions (TPRs), defined as the top 100 regions with the highest and lowest TE densities in a 5 Mb window, respectively. TPR genes were enriched in housekeeping functions such as RNA splicing (RNA polymerase II, 3'-UTR-binding,

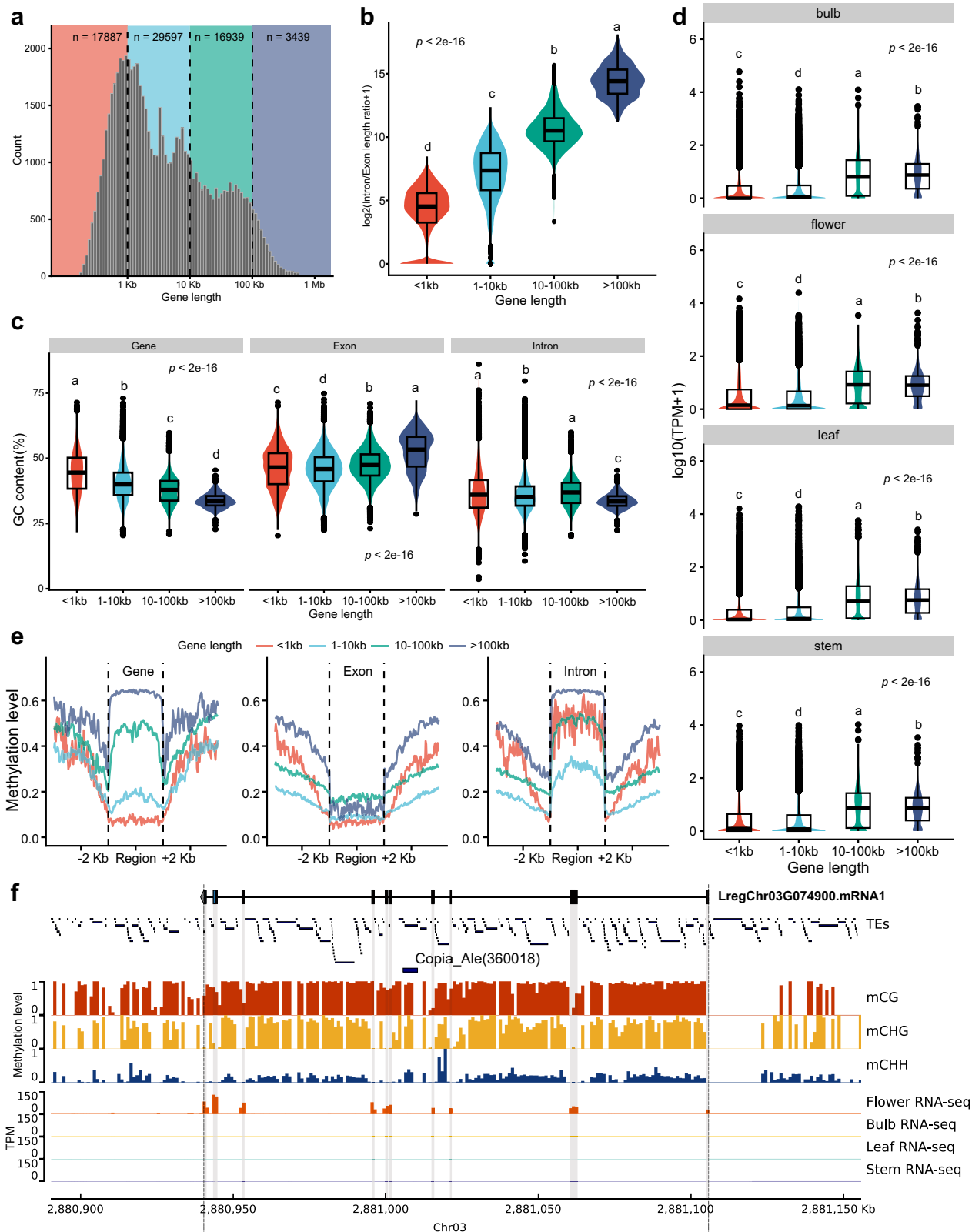
etc.) (Supplementary Fig. 18a), while TRR genes were enriched in stress response pathways (response to chitin, the synthesis and metabolism of trehalose, etc.) (Supplementary Fig. 18b). Some TPRs might act as TAD boundaries to ensure a tight regulation of these housekeeping genes, and experienced fewer TE insertions or were influenced by TE clearance mechanisms, resulting in lower splicing activity, which was associated with CHG hypermethylation. Interestingly, most TPR genes involved in RNA splicing were unique to *L. regale* (Fig. 1d and Supplementary Figs. 4 and 19). On the other hand, the TRRs with extensive TE insertions leading to the formation of ultra-long genes mainly encoded stress-related genes, and also exhibited a high number and activity of splicing events, generating the diverse transcripts likely contributed to environmental adaptation. In summary, *L. regale* has evolved an effective utilization of the dual roles endowed by TEs to cope with varied environmental challenges over its long evolutionary history.

Discussion

Genome size profoundly influences the growth, development, and ecological adaptation of organisms. Many members in Liliales, including lilies, tulips, and fritillaries, possess large genomes notoriously difficult to assemble. Combining PacBio HiFi and Hi-C data, we de novo assembled a 35.6 Gb genome of *L. regale* with the highest contiguity and quality of all lily genomes (Fig. 1a and Table 1). It raises the question of how lilies have evolved such a large genome. We show that *L. regale* genome is expanded primarily due to extensive inter-genic regions (96.01%), high content of TEs (80.06%, of which 72.3% are LTR retrotransposons), and large genes with long introns. Retrotransposons including LTRs, LINEs, and SINEs were the key drivers of genome expansion, and their precise regulation, along with an efficient transcription control system, enabled gigabase genomes to function properly.

We also revealed insights into how these LTRs were retained within the large genome. Unequal recombination between LTRs typically removes intermediate sequences and the formation of solo-LTRs⁴ in small genomes, while high-frequency gene conversion events often supplant the role of UR⁴ in larger genomes. However, the abundance of solo-LTRs in *L. regale* suggested rapid LTR recombination. Notably, epigenetic marker H3K9Me2 showed pronounced enrichment at terminal regions of LTRs (Fig. 2e), implying that it might act as a lock–stabilizing the duplicated terminal regions by restricting recombination due to heterochromatin, while the less-constrained central regions undergo frequent recombination generating numerous solo-LTRs. This broadly aligns with the findings reported before³⁸. Furthermore, given the observed TE expansion in *L. regale* (Fig. 2a), a similar mechanism might exist for other TE types. While the restriction of LTRs by H3K9Me2 appears independent of DNA methylation, RNA methylation specifically m6A does participate in the demethylation regulation of H3K9Me2³⁹. Our DNA methylation analysis indicated that prior to LTR insertion, the flanking regions likely exhibited relatively low methylation levels (Fig. 2g). After LTR insertion, high levels of methylation might suppress the activity of these LTRs as a defensive mechanism.

Lily genome encodes abundant large (>10 kb) genes (80% of all genes) with long introns largely due to 72.52% TE insertions. Genes with these extended introns could be correctly transcribed and



spliced, showing higher expression levels than smaller ones (Fig. 3d), consistent with findings in Chinese pine²⁷ and African lungfish²⁴. This suggested common strategies for maintaining the transcriptional efficiency of ultra-long genes by plants and animals. Alternative splicing (AS) contributes to key innovations over long evolutionary timescales⁶⁰, and plants have widespread AS events under stress conditions to better adapt to the environment^{8,9}. AS has been reported in

lilies but focused on individual genes so far, thus lacking a global perspective^{61,62}. By identifying genome-wide AS events, we observed a higher proportion of A3 events in lily compared to Arabidopsis and rice, but similar to those in wheat also having a large genome and high TE content (Fig. 4a and Supplementary Table 11). The significance of A3 events to large genome function, and whether this is common to giant genomes remains to be investigated in the future.

Fig. 3 | Ultra-long genes display unique transcriptional and epigenomic regulation. a The distribution of gene length in *L. regale*. **b** The intron/exon ratio of each gene between different length categories of full-length genes ($n = 15,667$ (< 1 kb), 34,295 (1–10 kb), 11,963 (10–100 kb), 952 (> 100 kb) genes). Genes without introns were regarded as “0”. **c** The GC content (%) of different length genes in full genes/exons/introns. The intron/exon ratio of each gene between different length genes ($n = 15,667$ (< 1 kb), 34,295 (1–10 kb), 11,963 (10–100 kb), 952 (> 100 kb) genes). Genes without introns were regarded as “0”. **d** The TPM of different length genes ($n = 5369$ (< 1 kb), 14,262 (1–10 kb), 12,668 (10–100 kb), 3171 (> 100 kb) genes) in four tissues. **e** The DNA Methylation ratio (The proportion of 5mC at the given site) of gene bodies, exons and introns between their upstream and downstream (± 2 Kb) in flower tissue. Colors represent different lengths of genes. **f** With

many TE insertions, *LregChr03G072890* expression under the control of methylation. The gene structure, TE insertions, intact LTR with its insertion time (year), methylation levels (CG, red; CHG, yellow; and CHH, blue; methylation level represent the proportion of 5mC at the given site), gene expression in four tissues (flower, orange; bulb, canary; leaf, cyan; and stem, purple) were plotted from the top to bottom successively. The shaded gray areas represent exon regions. **b–d** Groups labeled with different lowercase letters indicate significant differences ($P < 0.05$) based on LSD comparisons following one-way ANOVA, box plots denote the 25th percentile, the median and the 75th percentile, with minimum to maximum whiskers. Source data are provided as a Source Data file. In **b, c** gene outliers based on GC content are filtered out from total genes.

We proposed that alternative splicing in *L. regale* genome was likely regulated by two key factors: TE insertion and DNA methylations. Genes with intronic TE insertions, particularly those involving LTRs and LINEs, displayed high splicing activity, especially for A3 events (Fig. 4c, d). We also observed that DNA methylation in exons and introns of large genes was distinguishable (Fig. 3e and Supplementary Fig. 11e), and genes with low splicing activity were often associated with higher DNA methylation at splice sites. Previous studies showed that TE insertions can alter splicing patterns and even turn off gene expression⁶³, for example, through CHG methylation at splice sites in maize⁶⁴. We similarly observed that strong CHG methylation in non-TE regions of introns was associated with low splicing activity in *L. regale* (Supplementary Fig. 12g). This suggested that the mechanisms by which splicing was affected by DNA methylation and TE insertions were probably different. Notably, genes with low alternative splicing were primarily involved in RNA formation and splicing (Supplementary Fig. 14), and contained fewer TE insertions thus relatively stable, likely benefiting from rapid TE clearance. Combination with high CHG methylation at non-TE sites also might reduce splicing variants to stabilize gene transcription. Furthermore, lily-specific gene family analysis indicated that lilies had evolved numerous genes related to RNA splicing, such as poly(A)-specific ribonuclease activity. In conclusion, the extensive TE insertions contributed to regulating the alternative splicing of genes in *L. regale*, which likely evolved specific RNA splicing-related genes (Supplementary Fig. 14) to facilitate the proper splicing of the prevalent long genes as a mechanism for environmental adaptation.

Eukaryotic cells could manage packing their linear DNA of meters long inside the nucleus. The spatial organization of chromatin plays a key role in regulating gene expression^{15,65}, as demonstrated by many genomes of regular sizes, whereas the 3D spatial structure of giant genomes remains much of a mystery. In this study, we revealed the 3D chromatin organization of the lily and identified its tightly intertwined chromosomes (Fig. 5a), different from the compact and discrete chromosomal domains in species such as peanuts¹⁶, suggesting that the large chromosomes might lead to strong trans-interactions (Supplementary Fig. 13a). *L. regale* chromosomes could similarly be divided into A/B compartments and encode a large number of TAD-like structures. However, the mechanism of TAD boundary formation has remained unclear in plants due to the lack of CTCF-mediated looping⁶⁶. Our observations suggested that TAD boundaries in *L. regale*, due to infrequent TE insertions (Supplementary Fig. 15b), were less prone to mutations to ensure TAD integrity. As expected, TAD boundaries in *L. regale* were enriched with genes associated with core biological functions. For example, genes involved in endocytosis and cytoskeletal functions were enriched with H3K4Me3 (Supplementary Fig. 17a), while genes related to RNA splicing were enriched with H3K9Me2 (Supplementary Fig. 17b and Supplementary Data 2), indicating a defensive strategy in lily to suppress deleterious effects by extensive TE invasion.

In summary, we de novo assembled a gigantic genome of *L. regale* with the highest quality and contiguity within Liliaceae. We

demonstrate how transposon proliferation may have contributed to the expansion and evolution of *L. regale* genome, and how epigenetic regulation likely impacted transposon retention and orchestrated the proper genome function. This genome is an essential resource for lily genetics, breeding, conservation biology, and studies of angiosperm genome evolution, particularly for understanding evolutionary mechanisms of genome size expansion and organism adaptation in large-genome organisms.

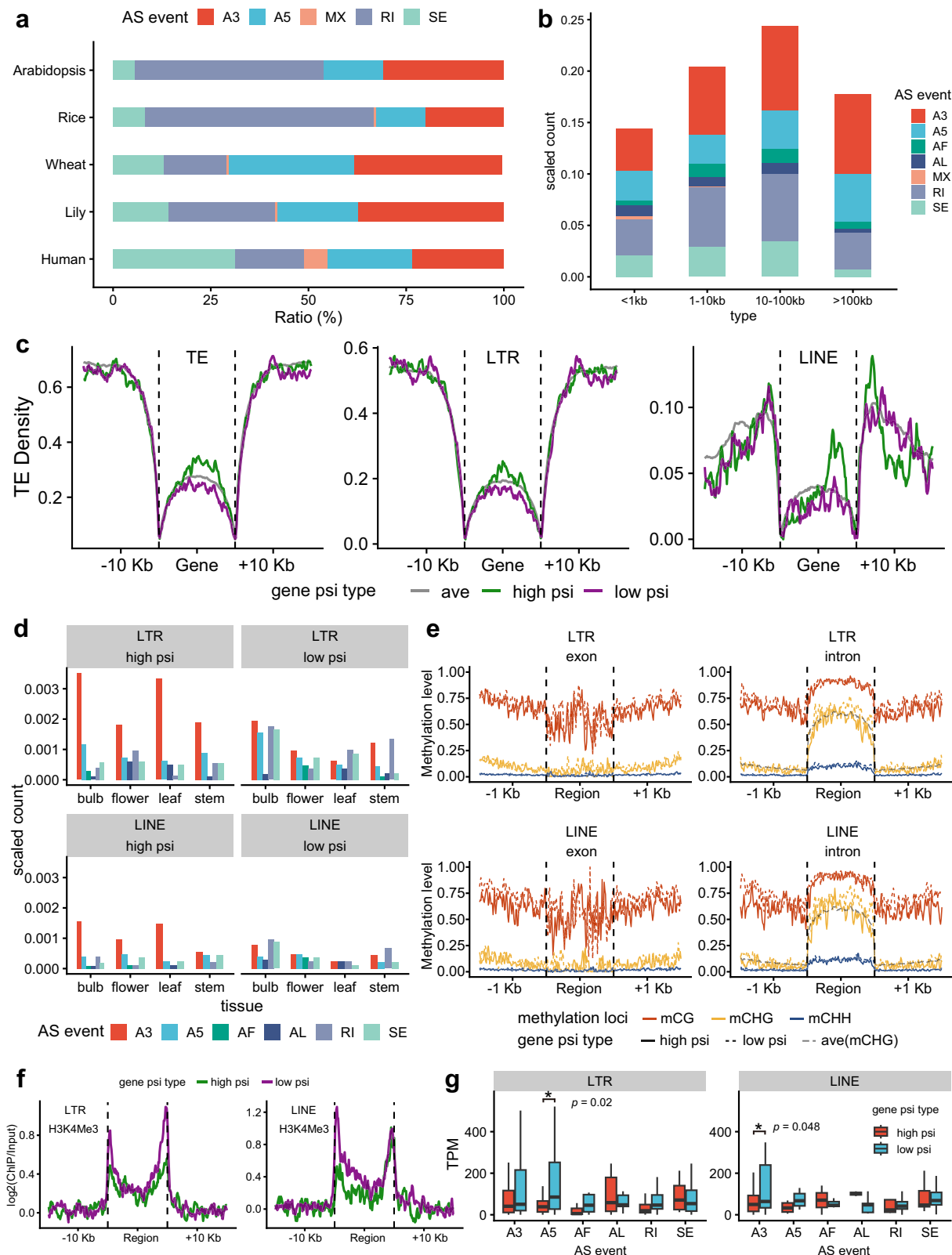
Methods

Plant materials and DNA sequencing

The royal lily (*Lilium regale*), sourcing from field collection, designated as “MJ-PKU01”, was used in this study due to its desirable agronomic characteristics, including broad-spectrum disease resistance, and abiotic stress tolerance. Bulbs of royal lily were planted at an approximate depth of 10 cm in soil and maintained in the greenhouse of Peking University Institute of Advanced Agricultural Sciences (PKU-IAAS) under natural conditions in Weifang, China (36.43° N, 119.6° E). Fresh leaves harvested from a single MJ-PKU01 plant were used for DNA extraction and sequencing. For Illumina short-read sequencing, DNA was extracted from about 2 g of leaves using a modified cetyltrimethylammonium bromide (CTAB) method. A sequencing library with an insert size of 350 bp was prepared using the VAHTS Universal DNA Library Prep Kit (Vazyme, Nanjing, China). Quality assessment of the library assessing DNA quantity, purity, and size range was conducted using Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA). The library was sequenced on the NovaSeq 6000 sequencing platform to produce pair-end sequence data (2×150 bp). For Pacific Biosciences (PacBio) sequencing, extraction of high-molecular-weight DNA was carried out as above. About 10 μ g of genomic DNA was used to prepare template libraries of 30–40 kb using the BluePippin Size Selection system (Sage Science, USA) following the manufacturer’s protocol (Pacific Biosciences, USA). The libraries were sequenced on the PacBio SEQUEL II platform with three SMRT flow cells.

Hi-C library construction and sequencing

Fresh leaves and flowers were cut into 2- to 3-mm strips and cross-linked with 1% formaldehyde at room temperature. The cross-linked tissues were treated with vacuum three times under a vacuum pressure of -60 to -80 kPa, maintained for 10 min each time, after which atmospheric pressure was restored. Subsequently, a final concentration of 0.125 M glycine was added for 5 min at room temperature to terminate fixation. The fixed tissues were homogenized with liquid nitrogen and resuspended in nuclei isolation buffer. The isolated nuclei were lysed with 0.1% SDS at 65 °C for 10 min and digested with HindIII (AAGCTAGCTT) at 37 °C overnight. The restriction fragments were blunt-ended and biotinylated with biotin-14-dCTP, diluted, and ligated with T4 ligase at 16 °C for 4 h. After the reversal of crosslinks at 65 °C overnight, the ligated DNA was purified and sheared to lengths of 300–500 bp with Covaris M220, pulled down with streptavidin beads, and prepared for Illumina sequencing



according to the standard Illumina library construction protocol. Libraries were quantified and sequenced using BGISEQ-2000 sequencing platform.

RNA-seq and Iso-seq sequencing

In this study, for substantiating transcripts to annotate the genome structure, we performed RNA-seq and Iso-seq of the total RNA isolated

from four different tissues (root, stem, flower, and leaf). Total RNA was extracted from ground tissues in liquid nitrogen using TRIzol reagent (Tiagen) following the protocol provided by the manufacturer. The integrity of the RNA was determined with the Agilent 2100 Bioanalyzer (Agilent Technologies) and agarose gel electrophoresis. The purity and concentration of the RNA were determined with the Nanodrop (Thermo Fisher Scientific) and Qubit (Thermo Fisher Scientific). The

Fig. 4 | Alternative splicing variations are associated with TE density and DNA methylation. **a** The alternative splicing (AS) events of *L. regale* compared to several representative organisms: *Arabidopsis thaliana*¹²⁶, *Oryza sativa*¹²⁷, *Triticum*¹²⁸ and *Homo sapiens*¹²⁹. **b** The events ratio of different length genes. scaled counts means “events count/perspective gene numbers”. The scaled count was calculated by events count/gene count. **c** The TE density (the proportion of TE length to the bases with non-zero coverage in the region) in gene bodies, upstream and downstream of HPG (high PSI gene, green), LPG (low PSI gene, purple) and average of all identified genes by PacBio Iso-seq (gray). **d** The counts of the level of different length genes between their upstream and downstream in HPGs and LPGs. **e** The DNA methylation level (CG, red; CHG, yellow; CHH, blue; methylation level represents the

proportion of 5mC at the given site) of the gene bodies, upstream and downstream of genes with LTR/LINE insertions (colored line, HPGs; colored dash, LPGs; gray dash, average of all genes at mCHG). **f** The enrichment of H3K4Me3 on HPGs and LPGs with LTR or LINE insertions. **g** The TPM (transcript per million mapped reads) of HPGs ($n = 255$ genes) and LPGs ($n = 230$ genes) with LTR or LINE insertions. **a, b, d** Colors indicate different AS events. **g** “*” denotes significant differences between groups under a two-tailed t test ($P = 0.02$ for LTR; $P = 0.048$ for LINE), with no significant differences observed among other groups; box plots denote the 25th percentile, the median and the 75th percentile, with minimum to maximum whiskers. Source data are provided as a Source Data file.

total RNA was used for RNA sequencing library construction following standard procedures of True-seq (Illumina) transcriptome sequencing and Iso-seq (PacBio) preparation kits. Finally, RNA-Seq and Iso-Seq were performed on the NovaSeq 6000 platform and the PacBio Sequel II platform, respectively. A total of 279 Gb RNA-seq data and 164 Gb clean Iso-Seq data were generated (Supplementary Table 6), which were then used for whole-genome protein-coding gene prediction.

De novo genome assembly and quality evaluation

To estimate the genome size and heterozygosity, KMC v3.2.1⁶⁷ was used to generate a 35 k -mer frequency distribution. Depth of k -mer = 1 is considered as an error. The formula for estimating the genome size is: Genome size = $(k\text{-mer num/main peak depth})$. The heterozygosity ratio was estimated by the Genomescope2 v2.0⁶⁸.

We used about 806 Gb PacBio HiFi reads comprising 51 M sub-reads with an average length of 15 kb for the primary assembly using Hifiasm v0.19.5⁶⁹ with the following parameter: `--hom-cov 22 -l`, which took around 2055.5 CPU hours to finish assembly. Then, a succession of searches was used to identify potential contaminants, plastids, and mitochondria in the generated assembly, using megaBLAST v2.12.0⁷⁰. To identify the haplotigs (haplotypes) and remove redundant sequences from a highly heterozygous genome, we used `purge_dups` v1.2.6⁷¹ with the cutoff parameters set to: `-l 2 -m 18 -u 72`. To acquire chromosome-level genome assembly, the assembled contigs were anchored using Hi-C data with the Juicer v1.5⁷² pipeline. YaHS v1.1⁷³ was used for the construction of chromosome-scale scaffolds from Hi-C data, followed by manual correction using Juicebox v1.11.08⁷⁴.

To evaluate the single-nucleotide quality of the genome, we mapped the clean Illumina reads sequenced from the same genomic DNA samples using bowtie2 v2.5.1⁷⁵. In addition, compleasm v0.2.6⁷⁶ with liliopsida_odb10 database of the Benchmarking Universal Single-Copy Orthologs (BUSCO)⁷⁷ was used to check the assembly quality with genome mode. BUSCO v5.8.0 was also used to check the annotation quality with protein mode with the liliopsida_odb10 database. Then, the Merquy v1.3³⁰ was used to assess the genome quality, with consensus quality value (QV) and completeness statistic values, indicating a level of accuracy and completeness in the assembled genome.

Identification of repetitive sequences and LTR analysis

We used two approaches to identify repeat elements in the genome: homolog-based prediction and de novo prediction. RepeatMasker v4.1.5⁷⁸ was used to find repetitive sequences based on the known database. RepeatModeler v1.0.11⁷⁹ was used to perform de novo prediction of repeat sequences and the results were combined as the library for RepeatMasker to identify and classify repeat elements, then the unclassified repeat elements (“unknown”) were further classified by DeepTE v2022-10-22⁸⁰. We identified 80.06% repetitive sequences in the assembly. Then the repetitive sequences of assembly were soft-masked by RepeatMasker for gene annotation. Intact LTRs were identified by LTR Retriever v2.9.6⁸¹ and further categorized in TESorter v1.4.6^{82,83}. The intact LTR-RTs were used to construct phylogenetic trees. The ends of the LTR retrotransposons were aligned using Mafft v7.505⁸⁴, and the insertion time (T) of the LTRs was calculated using the

formula $T = K/2r$, where K is the genetic distance calculated using the formula $K = -0.75 \ln(1 - 4\lambda/3)$. The nucleotide divergence rate (λ) between the two LTRs less than 0.75 was retained for further analysis. The nucleotide substitution rate “ r ” was set to 1.3×10^{-8} substitutions per site per year. Uncorrected pairwise distances were used to construct a neighbor-joining unrooted phylogenetic tree using fasttree v2.1.11⁸⁵ with the default parameters. *Gypsy* burst peaks were filtered by “count ≥ 35 ” in each 2.5 Mb bin, while *Copia* burst peaks were filtered by “count ≥ 20 ” in each 2.5 Mb bin. Solo-LTRs were identified using the script supplied by LTR Retriever, and then the potential solo-LTRs were classified by TESorter with truncated-LTRs and other TEs removed using bedtools v2.30.0⁸⁶. The homolog truncated and solo LTRs by using intact LTRs as queries in sequence alignment (BLAST), applying the criteria of identity $>80\%$ and e -value $< 1 \times 10^{-4}$.

Gene structure prediction

For genome annotation, ab initio prediction and similarity alignment were performed to annotate genome repeats via RepeatModeler and a soft-masked genome by RepeatMasker. Gene model prediction combining ab initio prediction, homologous proteins, RNA-seq, and Iso-seq evidence was constructed using the pipeline described as follows. For the ab initio prediction, we first trained the GeneMark-ET model using BRAKER2 v2.1.6⁸⁷ and then trained the SNAP⁸⁸ semi-HMM model using MAKER v2.31.11⁸⁹. Then, the protein data from *Allium fistulosum* v2.1⁹⁰, *Allium sativum*⁹¹, *Oryza. sativa* L⁴⁸, *Acanthochlamys bracteata*⁹², *Dioscorea alata*⁴⁴, and *Asparagus setaceus*⁹³ were downloaded through NCBI GenBank and used as homologous protein evidence. For transcriptome evidence, RNA-seq reads were first aligned to our genome assembly using HISAT2 v2.2.1⁹⁴. Reference-based assembly and de novo assembly of transcriptomes were performed subsequently using Trinity v2.8.4⁹⁵. Iso-seq data was analyzed using Iso-seq v4.0.0 pipeline⁹⁶. The above evidence was integrated by MAKER to predict the final gene model. The longest transcript of each predicted gene model was considered as its representative. Gene models were finally manually checked and corrected in IGV-GSaman v0.8.3 (<https://gitee.com/CJchen/IGV-SRNA>) with the support of mapped RNA-seq reads and previous annotations. RNA-seq data were quantified using Kallisto v0.48.0⁹⁷ as transcripts per million (TPM). The raw gene set was filtered to generate a core set of high-confidence protein-coding genes. Genes without protein (neither BLAST identity ≥ 40 nor e -value $< 1 \times 10^{-5}$) or EST (TPM in each tissue = 0) evidence were filtered out using a minimum cutoff of 100 amino acids.

Phylogenetic analysis

Orthologues and orthogroups in the 15 monocots^{23,28,44,90,91,98–105} and one dicotyledon¹⁰⁶ as an outgroup (Supplementary Table 12), were inferred using OrthoFinder v2.5.4³⁵ with default values setting and ‘-M msa’ activated. The longest predicted protein of each gene was used as the representative input for OrthoFinder analysis. 370 low-copy orthogroups were identified. TrimAl v1.4.12¹⁰⁷ was used to remove poorly aligned regions from protein multiple sequence alignments. RAXML v8.2.12¹⁰⁸ was used to build

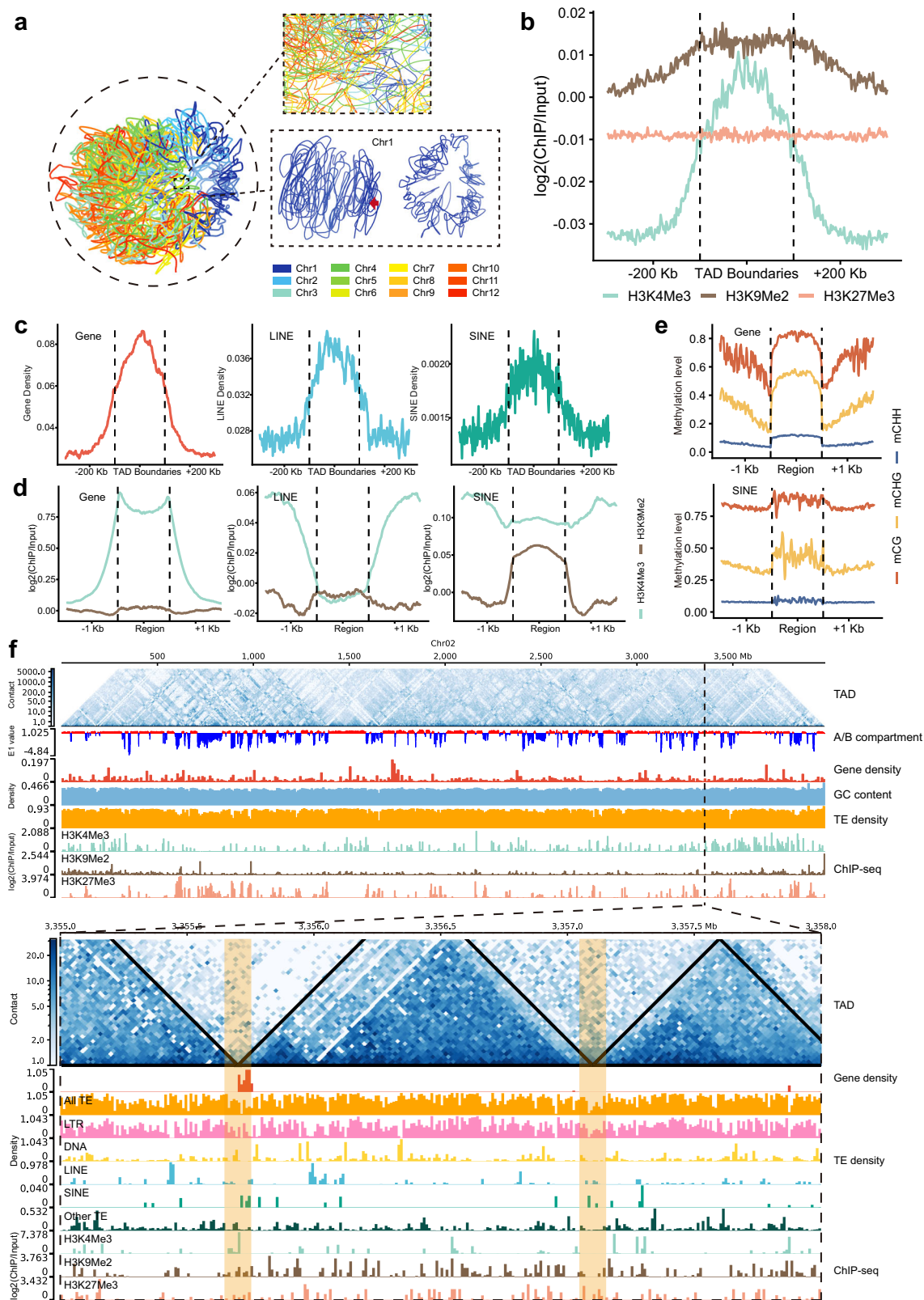


Fig. 5 | Epigenetic footprint and 3D genomic feature of *L. regale*. **a** The 3D model of each whole chromosome in the *L. regale* haplotype. Each color represents one chromosome. **b** The H3K4Me3, H3K9Me2 and H3K27Me3 enrichment on TAD boundaries, upstream and downstream. **c** Gene, LINE and SINE density between TAD boundaries upstream and downstream. **d** The H3K4Me3, H3K9Me2 and H3K27Me3 enrichment on genes, LINEs and SINEs on TAD boundaries. **e** The DNA methylation level (CG, red; CHG, yellow; CHH, blue; methylation level represents the proportion of 5mC at the given site) between upstream and downstream of

Gene and SINE on TAD boundaries. **f** Characteristics of the region on Chr02 as an example, which had SINEs and H3K9Me2 enrichment on TAD boundaries (High-lighted). The TAD domains (black triangles), A/B compartment (red, A compartment; blue, B compartment), gene density (red), GC content (%), TE density (orange, all TE; LTR, pink; DNA, gold; LINE, skyblue; SINE, green; other TE, darkgreen), ChIP-seq (cyan, H3K4Me3; brown, H3K9Me2; flesh, H3K27Me3) were plotted from the top to bottom successively.

maximum likelihood phylogenetic trees using the GAMMAJTT model, with *P. somniferum* as an outgroup. CAFE5³⁸ was used to infer gene gain and loss rates across genomes. The orthogroups generated by OrthoFinder were treated as distinct gene families and used as inputs for CAFE5 analysis. The identified genes underwent GO and KEGG enrichment analyses, with a significance threshold of $P < 0.05$. Divergence time estimates from TimeTree (www.timetree.org), a public database, were used for selecting the range of lower and upper uniform calibration priors. We used two pairwise divergence times, including (1) *Z. mays* and *A. comosus* (94.4–117 Mya), (2) “*A. shenzhenica*” and “*C. ensifolium*, *P. aphrodite*” (104.8–112.4 Mya), (3) *P. somniferum* and others (142.1–163.5 Mya). The CodeML and MCMCTREE programs in the PAML v4.948³⁶ were run to analyze amino acid substitution models and estimate species divergence times.

Whole-genome duplication analysis

The syntenic and WGD analysis was performed by WGD v0.6.5⁴⁵. We identified syntenic blocks by performing an all-against-all BLAST search and chaining the hits with a distance cutoff of five genes with WGD ‘icl’ module. Followed by calculating the K_s values for each anchor gene pair with the WGD ‘ks’ module. K_s dot plots for all anchor genes were created using the WGD ‘bk’ module. The Gaussian-fitted K_s peaks were generated using the WGD ‘pf’ module. *L. regale* and *AMK* exhibit extensive and well-maintained collinear blocks, with a syntenic block ratio of 4:1 (*L. regale*:*AMK*), indicating that *L. regale* has undergone WGD. Similarly, *L. regale* and *D. alata* also share extensive and well-maintained collinear blocks, with a ratio of 4:4, further supporting the occurrence of WGD events in *L. regale*. We calculated the K_s values for each gene pair in the aligned blocks obtained using WGD (v0.4.7) and plotted the distribution of all K_s values to infer WGD events in *L. regale*’s evolutionary history. Three K_s peaks were observed in *L. regale*, indicating multiple WGD events. Theoretically, each peak represents a WGD event. However, frequent recent gene duplications can also generate K_s peaks. To verify this, we mapped the gene pairs associated with the recent K_s peak onto chromosomes. These genes were not evenly distributed along the chromosomes, suggesting that the recent K_s peak resulted from random gene duplications rather than a true WGD event. The remaining two K_s peaks correspond to the τ event shared by monocots and a lineage-specific WGD event in *Lilium*, confirming that *L. regale* has experienced at least two WGD events in its evolutionary history.

To estimate the timing of WGD in *L. regale*, we used the “-ks” parameter in WGD to calculate the K_s values of syntenic block genes within *L. regale* and between *L. regale* and *D. alata*. We then applied the “-kp” parameter in WGD for Gaussian fitting of the K_s values, identifying the components associated with WGD peaks and calculating their mean K_s values and standard deviations. Using an estimated divergence time of 119.9 Mya and the average K_s value (1.45) of syntenic gene pairs between *L. regale* and *D. alata*, we estimated the evolutionary rate of *L. regale*. The synonymous substitution rate per site per million years (r) was calculated as $6.046e^{-4}$ ($r = K_s/2T$). Finally, we applied this r value and the mean WGD K_s value to estimate the WGD timing in *L. regale* using the formula $T = K_s/2r$.

Differential genes in this study were all performed enrichment analyzed of GO and KEGG pathways using the clusterProfiler v4.6.2¹⁰⁹ R package, in which gene length bias was corrected. Functional terms with corrected P value ≤ 0.05 were considered significantly enriched by differential genes.

Alternative splicing analysis

The identification of AS events was performed using the GFF file of transcripts obtained from the Iso-Seq pipeline, across different tissues, with SUPPA2 v2.3¹¹⁰. The parameters used were ‘-e SE SS MX RI FL -f ioe’. SUPPA2 generated seven types of events: A5, A3,

RI, SE, MX, AF, and AL. For each AS event in a gene, an ioe format file was generated to describe any form of the event. Specifically, the ioe file provided the numerators (one form of the event) and denominators (two forms of the event) of the transcripts, which assisted in PSI calculation. The quantification of transcripts was performed using IsoQuant v3.3.1¹¹¹ with FLNC sequences obtained from the Iso-Seq pipeline. The SUPPA2 “psiPerEvent” command was used to calculate the PSI values based on the TPM and ioe files obtained above.

Hi-C data processing

The Hi-C-Pro suite v3.1.0¹¹² was used for valid mapping of the Hi-C reads. Clean Hi-C data from individual accessions were mapped to the assembled genome (termed Self-mapping in this study) and a common reference genome using bowtie2. During mapping, the reads were first aligned using an end-to-end aligner, and the reads spanning ligation junctions were trimmed at their 3’ end and realigned to the genome. The aligned reads of both fragment mates were then paired in a single paired-end BAM file. Dangling-end reads, same-fragment reads, self-circled reads, self-ligation reads, and other invalid Hi-C reads were discarded from subsequent analyses. After removing duplications, valid pairs were used to generate raw Hi-C matrix at 10-kb, 20-kb, 50-kb, 100-kb, 200-kb, 500-kb, and 1-Mb resolutions. These matrices were normalized by the iterative correction and eigenvector decomposition (ICE) method. To investigate the contact domains, the valid pairs were converted into.hic format files with the juicer tool pre command; to calculate compartment signals and identify TAD-like or loops, the valid pairs were converted into.cool format files with hicConvertFormat command in HiCExplorer suite v3.7.2¹¹³; and to visualize the Hi-C contact map, the valid pairs were converted into.h5 format.

3D genome architecture identification

The miniMDS¹¹⁴ program was used to infer the 3D genome structures at the 1 mb ICE normalized contact matrix. Pymol v2.5.2¹¹⁵ was used to visualize 3D coordinates. Then, E1 values from the eigenvector decomposition on Hi-C contact maps were used to indicate the A/B compartment status. We used cooltools v0.5.4¹¹⁶ with the parameter “cooltools call-compartments” to obtain the E1, E2, and E3 values based on a 1 mb resolution Hi-C contact matrix. Because E2 or E3 values sometimes reflect A/B compartments, we manually checked the E1, E2, and E3 tracks with gene density and the plaid pattern in the Hi-C contact maps along each chromosome and obtained the final “E1” list. The direction of the eigenvalues is arbitrary; therefore, negative values were set to “A” and positive values were set to “B” based on their association with gene density. Compartment border was defined as the edge bin of the A/B compartments. TAD boundaries were identified based on a 100 kb resolution ICE normalized matrix with the parameters “--fdr 0.05.” by HiCExplorer. Loops were identified based on a 20 kb resolution ICE normalized matrix with the parameters “--fdr 0.05.” by HiCExplorer at the same.

Bisulfite sequencing and data processing

DNA was isolated from leaf and flower tissue to generate bisulfite sequencing (BS-seq) data. Bisulfite libraries were prepared¹¹⁷ and sequenced on (150 bp paired-end) the Illumina Novaseq 6000 system. To measure the genome-wide DNA methylation level, BSseq reads were first trimmed for quality and adapter sequences using fastp v0.38¹¹⁸, resulting in a total of 386.14 Gb (10.8× genomic coverage) clean reads. Bismark v0.24.0¹¹⁹, in conjunction with bowtie2 was then used to align the trimmed reads to the genome. The number of methylated and unmethylated reads per cytosine was determined using the BatMeth2¹²⁰. DNA methylation level of a genomic bin (e.g., 10 and 40 kb) was computed as the percentage of heavily methylated sites (which we defined as methylated reads contributing to at least

25% of all mapped reads) in that bin. This value was calculated for three contexts (i.e., CpG, CHG, and CHH), separately. Also, a value that we defined as the overall methylation level was calculated by summing across all three contexts.

ChIP library construction and sequencing

Chromatin Immunoprecipitation sequencing (ChIP-seq) data of three histone modifications (H3k4me3, H3K27me3, and H3K9me2) were generated for *L. regale* with two biological replicates¹⁷. Sequencing (150 bp paired-end) was performed on the Illumina NovaSeq 6000 system. ChIP-seq raw reads were trimmed for quality and adapter sequences using fastp and were then mapped to the *L. regale* genome using bowtie2. Enrichment peaks were called using MACS2 v2.2.7.1¹²¹ to verify the quality of ChIP-seq data. The ChIP-seq intensity of a genomic bin (i.e., 10 and 40 kb) was calculated using the bam-Coverage tool from DeepTools v3.5.1¹²² with the read coverage normalized in counts per million (CPM).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw sequencing data generated in this study have been deposited in **EBI** and **NCBI** under the accession PRJEB89667 and National Genomics Data Center (<https://ngdc.cncb.ac.cn>) under the accession CRA014739. The genome assembly has been deposited in the Genome Warehouse under the accession GWHEUVG000000000.1 and BioProject PRJCA022916. The genome assembly and annotation files are also available in figshare¹²³ (<https://doi.org/10.6084/m9.figshare.26345504.v1>). Source data are provided with this paper.

References

- Bhadra, S., Leitch, I. J. & Onstein, R. E. From genome size to trait evolution during angiosperm radiation. *Trends Genet.* **39**, 728–735 (2023).
- Peng, Y. et al. Plant genome size modulates grassland community responses to multi-nutrient additions. *N. Phytol.* **236**, 2091–2102 (2022).
- Šimová, I. & Herben, T. Geometrical constraints in the scaling relationships between genome size, cell size and cell cycle length in herbaceous plants. *Proc. Biol. Sci.* **279**, 867–875 (2012).
- Cossu, R. M. et al. LTR Retrotransposons Show Low Levels Of Unequal Recombination And High Rates Of Intralelement Gene Conversion In Large Plant Genomes. *Genome Biol. Evol.* **9**, 3449–3462 (2017).
- Lippman, Z. et al. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**, 471–476 (2004).
- Talbert, P. B. & Henikoff, S. Centromeres convert but don't cross. *PLoS Biol.* **8**, e1000326 (2010).
- Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18**, 437–451 (2017).
- Chaudhary, S., Jabre, I., Reddy, A. S. N., Staiger, D. & Syed, N. H. Perspective on alternative splicing and proteome complexity in plants. *Trends Plant Sci.* **24**, 496–506 (2019).
- Wang, L. et al. The pattern of alternative splicing and DNA methylation alteration and their interaction in linseed (*Linum usitatissimum* L.) response to repeated drought stresses. *Biol. Res.* **56**, 12 (2023).
- Bickmore, W. A. & van Steensel, B. Genome architecture: domain organization of interphase chromosomes. *Cell* **152**, 1270–1284 (2013).
- Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327–339 (2011).
- Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.* **20**, 437–455 (2019).
- Grob, S., Schmid, M. W. & Grossniklaus, U. Hi-C analysis in *Arabidopsis* identifies the KNOT, a structure with similarities to the *flamenco* locus of *Drosophila*. *Mol. Cell* **55**, 678–693 (2014).
- Dong, P. et al. 3D Chromatin architecture of large plant genomes determined by local A/B compartments. *Mol. Plant* **10**, 1497–1509 (2017).
- Ni, L. et al. Pan-3D genome analysis reveals structural and functional differentiation of soybean genomes. *Genome Biol.* **24**, 12 (2023).
- Zhang, X. et al. Chromatin spatial organization of wild type and mutant peanuts reveals high-resolution genomic architecture and interaction alterations. *Genome Biol.* **22**, 315 (2021).
- Vidalie, H. The lily as a cut flower: an ample range of cultivars. *PHM Revue Horticole*, 38–44 (2008).
- Gao, X. et al. The lre-miR159a-LrGAMYB pathway mediates resistance to grey mould infection in *Lilium regale*. *Mol. Plant Pathol.* **21**, 749–760 (2020).
- Sun, D. et al. LreEF1A4, a translation elongation factor from *Lilium regale*, is pivotal for cucumber mosaic virus and tobacco rattle virus infections and tolerance to salt and drought. *Int. J. Mol. Sci.* **21**, 2083 (2020).
- Pellicer, J. & Leitch, I. J. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *N. Phytol.* **226**, 301–305 (2020).
- Qu, K. et al. Comprehensive analysis of the complete mitochondrial genome of *Lilium tsingtauense* reveals a novel multi-chromosome structure. *Plant Cell Rep.* **43**, 150 (2024).
- Xu, L., Cao, Y., Yang, P. & Ming, J. The complete plastid genome of *Lilium regale* E.H.Wilson. *Mitochondrial DNA B Resour.* **6**, 806–807 (2021).
- Xu, S. et al. The evolutionary tale of lilies: giant genomes derived from transposon insertions and polyploidization. *Innovation* **5**, 100726 (2024).
- Wang, K. et al. African lungfish genome sheds light on the vertebrate water-to-land transition. *Cell* **184**, 1362–1376.e18 (2021).
- Nowoshilow, S. et al. The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554**, 50–55 (2018).
- Shao, C. et al. The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights. *Cell* **186**, 1279–1294.e19 (2023).
- Niu, S. et al. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **185**, 204–217.e14 (2022).
- Liang, Y. et al. The giant genome of lily provides insights into the hybridization of cultivated lilies. *Nat. Commun.* **16**, 45 (2025).
- Fajkus, J., Šýkorová, E. & Leitch, A. R. Telomeres in evolution and evolution of telomeres. *Chromosome Res* **13**, 469–479 (2005).
- Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
- Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
- Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–d314 (2019).
- Harris, M. A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
- Kanehisa, M. The KEGG database. *Novartis Found Symp.* **247**, 91–101; discussion 101–3, 119–28, 244–52 (2002).

35. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
36. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
37. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).
38. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2021).
39. Cao, Y. et al. The jasmonate-induced bHLH gene *SLUG* functions in terpene biosynthesis and resistance to insects and fungus. *J. Integr. Plant Biol.* **64**, 1102–1115 (2022).
40. Liu, G. et al. Diverse O-methyltransferases catalyze the biosynthesis of floral benzenoids that repel aphids from the flowers of waterlily *Nymphaea prolifera*. *Hortic. Res.* **10**, uhad237 (2023).
41. Panchy, N., Lehti-Shiu, M. & Shiu, S. H. Evolution of gene duplication in plants. *Plant Physiol.* **171**, 2294–2316 (2016).
42. McKain, M. R. et al. A Phylogenomic assessment of ancient polyploidy and genome evolution across the poales. *Genome Biol. Evol.* **8**, 1150–1164 (2016).
43. Zhang, L. et al. The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. *Plant Cell Environ.* **43**, 2847–2856 (2020).
44. Bredeson, J. V. et al. Chromosome evolution and the genetic basis of agronomically important traits in greater yam. *Nat. Commun.* **13**, 2001 (2022).
45. Sun, P. et al. WGDl: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* **15**, 1841–1851 (2022).
46. Wu, S., Han, B. & Jiao, Y. Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Mol. Plant* **13**, 59–71 (2020).
47. Hou, X., Wang, D., Cheng, Z., Wang, Y. & Jiao, Y. A near-complete assembly of an *Arabidopsis thaliana* genome. *Mol. Plant* **15**, 1247–1250 (2022).
48. Song, J. M. et al. Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol. Plant* **14**, 1757–1767 (2021).
49. Wang, X. et al. A near-complete genome sequence of einkorn wheat provides insight into the evolution of wheat A subgenomes. *Plant Commun.* **21**, 100768 (2023).
50. Li, J. et al. The methylation landscape of giga-genome and the epigenetic timer of age in Chinese pine. *Nat. Commun.* **14**, 1947 (2023).
51. Xie, B., Deng, Y., Kanaoka, M. M., Okada, K. & Hong, Z. Expression of *Arabidopsis* callose synthase 5 results in callose accumulation and cell wall permeability alteration. *Plant Sci.* **183**, 1–8 (2012).
52. Schafer, S. et al. Alternative splicing signatures in RNA-seq data: percent spliced in (PSI). *Curr. Protoc. Hum. Genet.* **87**, 11.16.1–11.16.14 (2015).
53. Ayarpadikannan, S., Lee, H. E., Han, K. & Kim, H. S. Transposable element-driven transcript diversification and its relevance to genetic disorders. *Gene* **558**, 187–194 (2015).
54. Bräutigam, K. & Cronk, Q. DNA methylation and the evolution of developmental complexity in plants. *Front Plant Sci.* **9**, 1447 (2018).
55. Chen, L. et al. Dynamic 3D genome reorganization during development and metabolic stress of the porcine liver. *Cell Discov.* **8**, 56 (2022).
56. Li, X. et al. Genomic rearrangements and evolutionary changes in 3D chromatin topologies in the cotton tribe (Gossypieae). *BMC Biol.* **21**, 56 (2023).
57. Liu, C., Cheng, Y. J., Wang, J. W. & Weigel, D. Prominent topologically associated domains differentiate global chromatin packing in rice from *Arabidopsis*. *Nat. Plants* **3**, 742–748 (2017).
58. Fedoroff, N. V. Presidential address. Transposable elements, epigenetics, and genome evolution. *Science* **338**, 758–767 (2012).
59. Li, Y. et al. N⁶-Methyladenosine co-transcriptionally directs the demethylation of histone H3K9me2. *Nat. Genet.* **52**, 870–877 (2020).
60. Wright, C. J., Smith, C. W. J. & Jiggins, C. D. Alternative splicing as a source of phenotypic diversity. *Nat. Rev. Genet.* **23**, 697–710 (2022).
61. Du, F. et al. Volatile composition and classification of *Lilium* flower aroma types and identification, polymorphisms, and alternative splicing of their monoterpene synthase genes. *Hortic. Res.* **6**, 110 (2019).
62. Wu, Z. et al. Alternative splicing provides a mechanism to regulate *LIHSA3* function in response to heat stress in lily. *Plant Physiol.* **181**, 1651–1667 (2019).
63. Galindo-González, L., Mhiri, C., Deyholos, M. K. & Grandbastien, M. A. LTR-retrotransposons in plants: engines of evolution. *Gene* **626**, 14–25 (2017).
64. Regulski, M. et al. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Res.* **23**, 1651–1662 (2013).
65. Delaneau, O. et al. Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* **364**, 6439 (2019).
66. Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* **15**, 234–246 (2014).
67. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
68. 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).
69. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
70. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinforma.* **10**, 421 (2009).
71. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
72. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
73. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, 808 (2023).
74. Dudchenko, O. et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under \$1000. Preprint at <https://www.biorxiv.org/content/10.1101/254797v1> (2018).
75. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
76. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics* **39**, btad595 (2023).
77. Manni, M., Berkeley, M. R., Seppely, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.* **1**, e323 (2021).
78. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. <http://www.repeatmasker.org/faq.html#faq3> (2013–2015).
79. Smit, A. & Hubley, R. RepeatModeler Open-1.0. <http://www.repeatmasker.org/faq.html#faq3> (2008–2015).
80. Yan, H., Bombarely, A. & Li, S. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics* **36**, 4269–4275 (2020).

81. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retro-transposons. *Plant Physiol.* **176**, 1410–1422 (2018).
82. Neumann, P., Novák, P., Hošťáková, N. & Macas, J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob. DNA* **10**, 1 (2019).
83. Zhang, R. G. et al. TESorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic. Res.* **9**, uhac017 (2022).
84. Rozewicki, J., Li, S., Amada, K. M., Standley, D. M. & Katoh, K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res.* **47**, W5–W10 (2019).
85. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
86. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
87. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* **3**, (2021).
88. Korf, I. Gene finding in novel genomes. *BMC Bioinforma.* **5**, 59 (2004).
89. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinforma.* **12**, 491 (2011).
90. Liao, N. et al. Chromosome-level genome assembly of bunching onion illuminates genome evolution and flavor formation in *Allium* crops. *Nat. Commun.* **13**, 6690 (2022).
91. Sun, X. et al. A chromosome-level genome assembly of garlic (*Allium sativum*) provides insights into genome evolution and alliin biosynthesis. *Mol. Plant* **13**, 1328–1339 (2020).
92. Xu, B. et al. Chromosome-level de novo genome assembly and whole-genome resequencing of the threatened species *Acanthochlamys bracteata* (Velloziaceae) provide insights into alpine plant divergence in a biodiversity hotspot. *Mol. Ecol. Resour.* **22**, 1582–1595 (2022).
93. Li, S. F. et al. Chromosome-level genome assembly, annotation and evolutionary analysis of the ornamental plant *Asparagus setaceus*. *Hortic. Res.* **7**, 48 (2020).
94. 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).
95. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
96. Guizard, S. et al. nf-core/isoseq: simple gene and isoform annotation with PacBio Iso-Seq long-read sequencing. *Bioinformatics* **39**, btad150 (2023).
97. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
98. Ai, Y. et al. The *Cymbidium* genome reveals the evolution of unique morphological traits. *Hortic. Res.* **8**, 255 (2021).
99. Chao, Y. T. et al. Chromosome-level assembly, genetic and physical mapping of *Phalaenopsis aphrodite* genome provides new insights into species adaptation and resources for orchid breeding. *Plant Biotechnol. J.* **16**, 2027–2041 (2018).
100. Chen, J. et al. A complete telomere-to-telomere assembly of the maize genome. *Nat. Genet.* **55**, 1221–1231 (2023).
101. Finkers, R. et al. Insights from the first genome assembly of onion (*Allium cepa*). *G3* **11**, jkab243 (2021).
102. Huang, H. R. et al. Telomere-to-telomere haplotype-resolved reference genome reveals subgenome divergence and disease resistance in triploid Cavendish banana. *Hortic. Res.* **10**, uhad153 (2023).
103. Ming, R. et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* **47**, 1435–1442 (2015).
104. Shirasawa, K. et al. Chromosome-scale haplotype-phased genome assemblies of the male and female lines of wild asparagus (*Asparagus kiusianus*), a dioecious plant species. *DNA Res.* **29**, dsac002 (2022).
105. Zhang, G. Q. et al. The *Apostasia* genome and the evolution of orchids. *Nature* **549**, 379–383 (2017).
106. Guo, L. et al. The opium poppy genome and morphinan production. *Science* **362**, 343–347 (2018).
107. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
108. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
109. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
110. Trincado, J. L. et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 40 (2018).
111. Pribelski, A. D. et al. Accurate isoform discovery with IsoQuant using long reads. *Nat. Biotechnol.* **41**, 915–918 (2023).
112. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
113. Ramírez, F. et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).
114. Rieber, L. & Mahony, S. miniMDS: 3D structural inference from high-resolution Hi-C data. *Bioinformatics* **33**, i261–i266 (2017).
115. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. (2015).
116. Abdennur, N. et al. Cooltools: Enabling high-resolution Hi-C analysis in Python. *PLoS Comput Biol* **20**, e1012067 (2024).
117. Liao, Y. et al. The 3D architecture of the pepper genome and its relationship to function and evolution. *Nat. Commun.* **13**, 3479 (2022).
118. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* **2**, e107 (2023).
119. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
120. Zhou, Q., Lim, J. Q., Sung, W. K. & Li, G. An integrated package for bisulfite DNA methylation data analysis with Indel-sensitive mapping. *BMC Bioinforma.* **20**, 47 (2019).
121. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
122. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
123. Sun, J. et al. Genomic and epigenomic insight into giga-chromosome architecture and adaptive evolution of royal lily (*Lilium regale*). *figshare* <https://doi.org/10.6084/m9.figshare.26345504.v1> (2025).
124. Hao, F. et al. Chromosome-level genomes of three key *Allium* crops and their trait evolution. *Nat. Genet.* **55**, 1976–1986 (2023).
125. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
126. Marquez, Y., Brown, J. W., Simpson, C., Barta, A. & Kalyna, M. Transcriptome survey reveals increased complexity of the

- alternative splicing landscape in *Arabidopsis*. *Genome Res.* **22**, 1184–1195 (2012).
127. He, W. et al. Full-length transcriptome reconstruction reveals genetic differences in hybrids of *Oryza sativa* and *Oryza punctata* with different ploidy and genome compositions. *BMC Plant Biol.* **22**, 131 (2022).
128. Gao, P. et al. Alternative splicing dynamics and evolutionary divergence during embryogenesis in wheat species. *Plant Biotechnol. J.* **19**, 1624–1643 (2021).
129. Glinos, D. A. et al. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**, 353–359 (2022).

Acknowledgements

This project was supported by the Natural Science Foundation for Distinguished Young Scholars of Shandong Province (ZR2023JQ010, L.G.), Taishan Scholars Program of Shandong Province (L.G.), and the Key R&D Program of Shandong Province (ZR202211070163, L.G.). The authors would like to thank the Bioinformatics Platform at Peking University Institute of Advanced Agricultural Sciences for providing the high-performance computing resources.

Author contributions

L.G. conceived and designed the project. L.Z. and G.Y. prepared the plant materials for sequencing. D.M. and Y.M. conducted experiments. J.S. performed genome assembly and bioinformatic analysis. X.W. conducted evolutionary genomic analysis. J.S. and K.W. conducted genomic alternative splicing analysis. J.S., X.W., and J.W. prepared the figures and tables. J.S., L.G., and X.W. wrote the manuscript. All authors read, revised and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-61289-w>.

Correspondence and requests for materials should be addressed to Li Guo.

Peer review information *Nature Communications* thanks Zhe Liang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

© The Author(s) 2025