nature plants

Article

https://doi.org/10.1038/s41477-025-02060-3

Two Melanthiaceae genomes with dramatic size difference provide insights into giant genome evolution and maintenance

Received: 27 September 2021

Accepted: 26 June 2025

Published online: 01 August 2025

Check for updates

Peng Zeng[1,5], Hang Zong[1,5], Yuwei Han[1,5], Weixiong Zhang[1,5], Zunzhe Tian[1,5], Botong Zhou[1,5], Juan He[1,5], Yongting Zhang[1,5], Xiaonan Liu[2,4,5], Lin Liu[1], Tinggan Zhou[1], Qiong Li[1], Yang Yu[1], Yingmei Peng[1], Wenbo Zhu[1], Simei He[3], Guanghui Zhang[3], Huifeng Jiang[2,6] ✉, Shengchao Yang[3,6] ✉ & Jing Cai[1,6] ✉

To characterize genome size evolution in the Melanthiaceae family, we sequenced the genomes of *Paris polyphylla* var. *yunnanensis* (54.58 Gb/1C) and *Veratrum dahuricum* (Turcz.) O. Loes. (3.93 Gb/1C). Using a hierarchical bottom-up chromosome assembly strategy, we successfully assembled the five giant chromosomes of *P. polyphylla*, with the largest chromosome reaching 14.14 Gb. We observed widespread secondary diagonal signals in a Hi-C interaction heat map of *P. polyphylla*, which suggests a higher-order helical structure (with ~250 Mb per turn) of interphase chromatin. Our genome assemblies reveal that *P. polyphylla* has not undergone recent whole-genome duplications since its divergence from *V. dahuricum*. Gene family analysis showed that the five DNA repair pathways are all significantly enriched in the expanded gene families in *P. polyphylla*. Our results not only report a giant, high-quality plant genome, but also reveal how giant chromosomes evolved and were retained.

Genome size (GS) is a crucial aspect in the evolution of land plants as it influences numerous biological processes. The amount of DNA in a haploid genome (1C-value) also exhibits astonishing diversity, varying over 24,000-fold[1,2]. So far, researchers have sequenced species from most families on the plant phylogenetic tree, resulting in over 1,720 assemblies (Supplementary Table 1). However, the majority of plant genomic studies have focused on species with small genomes. Only ~10% of the assemblies are greater than the median GS for angiosperms, which is 2.45 Gb/1C[1]. In the Plant C-value Database (https://cvalues.science.kew.org, Supplementary Table 2), over 600 species (top 5%) have extremely large GS (>20 Gb/1C). However, <1% (10 of 1,720) of the assemblies have GS >20 Gb/1C, which remain severely under-represented (Supplementary Table 1). Knowledge derived from small genomes may not be applicable to all genomes. As ref. 3 demonstrated, in small and medium-sized genomes, the repeat proportion is positively correlated with GS, but for genomes >10 Gb/1C, the repeat proportion does not

[1]Shaanxi Key Laboratory of Qinling Ecological Intelligent Monitoring and Protection, School of Ecology and Environment, Northwestern Polytechnical University, Xi'an, China. [2]Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, National Technology Innovation Center of Synthetic Biology, Tianjin, China. [3]State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, National and Local Joint Engineering Research Center on Germplasm Innovation and Utilization of Chinese Medicinal Materials in Southwest China, Yunnan Agricultural University, Kunming, China. [4]Present address: Cooperative Innovation Center of Industrial Fermentation (Ministry of Education and Hubei Province), Key Laboratory of Fermentation Engineering (Ministry of Education), Hubei Key Laboratory of Industrial Microbiology, National "111" Center for Cellular Regulation and Molecular Pharmaceutics, Hubei University of Technology, Wuhan, China. [5]These authors contributed equally: Peng Zeng, Hang Zong, Yuwei Han, Weixiong Zhang, Zunzhe Tian, Botong Zhou, Juan He, Yongting Zhang, Xiaonan Liu. [6]These authors jointly supervised this work: Huifeng Jiang, Shengchao Yang, Jing Cai. ✉e-mail: jiang_hf@tib.cas.cn; shengchaoyang@163.com; jingcai@nwpu.edu.cn

https://doi.org/10.1038/s41477-025-02060-3

Table 1 | Summary of assembly metrics and completeness of both genomes and annotated genes

Species	Genome assembly	Number of sequences	Total length (bp)	N50 (bp)	N90 (bp)	Longest (bp)
P. polyphylla	Contigs	78,381	52,745,548,966	1,517,703	334,444	257,336,657
	Chromosomes	5	47,765,251,326	9,900,240,954	7,583,609,810	14,138,182,717
	Un-anchored	34,838	5,178,980,651	674,000	58,386	92,110,000
V. dahuricum	Contigs	7,626	3,553,867,786	2,215,577	264,688	18,611,606
	Chromosomes	16	3,262,103,997	219,022,602	161,792,680	262,217,294
	Un-anchored	4,557	293,313,289	106,880	24,764	6,462,241
Species		**Complete**	**Complete and single copy**	**Complete and duplicated**	**Fragmented**	**Missing**
P. polyphylla	Genome	1,578 (97.77%)	1,256 (77.82%)	322 (19.95%)	8 (0.5%)	28 (1.73%)
	Gene	1,523 (94.4%)	1,211 (75.03%)	312 (19.3%)	12 (0.7%)	79 (4.9%)
V. dahuricum	Genome	1,586 (98.27%)	628 (38.91%)	758 (59.36%)	4 (0.25%)	24 (1.49%)
	Gene	1,491 (92.4%)	734 (45.5%)	757 (46.9%)	39 (2.4%)	84 (5.2%)

increase with GS. Recently, ref. 4 published the genome sequence of Lanzhou lily (*Lilium davidii* var. *unicolor*, 35.6 Gb/1C) and proposed that the giant GS was mainly due to recent transposon insertions (TEs, with a peak at ~1 million years ago (Ma)) and polyploidization. Nevertheless, the timing of TE bursts in giant genomes varies greatly, even within the same genus, such as in *Pinus*: *Pinus tabuliformis* (6 Ma)[5], *Pinus lambertiana* (16 Ma)[6] and *Pinus taeda* (17 Ma)[6,7]. Therefore, more studies on species with giant genomes are needed to clarify the mechanism of genome expansion.

As the most diverse family in terms of GS, Melanthiaceae species from 16 genera are classified into 5 tribes, Chionographideae, Heloniadeae, Melanthieae, Parideae and Xerophylleae[8] (Supplementary Fig. 1). The mean GS of the Parideae tribe is ~50 Gb/1C (the largest being 150 Gb/1C), while the mean GSs of the other four tribes range from 1.23 to 3.69 Gb/1C[8]. This indicates that Parideae underwent substantial genome expansion, probably through transposon insertion or genome duplication after diverging from the Xerophylleae tribe 52.4 Ma[9]. In the Parideae tribe, *Paris polyphylla* var. *yunnanensis* (2n = 2× = 10) has a giant flow cytometry-based GS of 54.58 ± 0.54 Gb/1C (Supplementary Table 3) and consists of only 5 chromosomes. The average chromosome length is estimated at 10.9 Gb, ranking among the largest chromosome classes of organisms recorded in the Plant C-value Database (Supplementary Fig. 2a and Table 2). Although chromosomes are fundamental organizational units in eukaryotic genomes, how extremely long chromosomes (at 10-Gb level) fold into higher-order structures and perform their functions remains unclear. The giant chromosomes of *P. polyphylla* provide a valuable opportunity to reveal how such extremely long chromosomes have evolved and been maintained.

To investigate the mechanism underlying the dramatic genome expansion in the lineage leading to *P. polyphylla*, we sequenced the genomes of *P. polyphylla* and *Veratrum dahuricum* (Turcz.), representatives of Parideae and Melanthieae tribes, respectively. Here we present a chromosome-level genome assembly of *P. polyphylla*, spanning 52.75 Gb/1C. This assembly is compacted into 5 chromosomes[10,11], with sizes ranging from 7.58 to 14.14 Gb and an average length of 9.55 Gb. The longest chromosome exceeds the size of the Lanzhou lily's largest chromosome by more than 3-fold and also surpasses the length of the South American lungfish's largest chromosome, which is 12.03 Gb[12] (Supplementary Fig. 2b). The Darwin Tree of Life Project (https://www.darwintreeoflife.org/) has released the largest chromosome assembly so far (92 Gb/1C) for *Viscum album*, which consists of 10 chromosomes, with the longest chromosome being 10.8 Gb, still smaller than that of *P. polyphylla* (Supplementary Fig. 2b). Another species in the Melanthiaceae family, *V. dahuricum*, possesses a significantly smaller assembly, with a GS of 3.55 Gb/1C and an average chromosome length of 0.20 Gb. The stark contrast in GS among closely related species provides a valuable opportunity to study the mechanism underpinning large genome and chromosome evolution.

Results and Discussion

Chromosomal-level assembly of two Melanthiaceae genomes

For *P. polyphylla*, we obtained a genome assembly of 52.75 Gb/1C in length with a contig N50 of 1.52 Mb (Table 1), using a HiFi assembly and additional round of hybrid assembly protocol with HiFi reads, ONT reads and Hi-C reads from the Pacbio Revio, PromethION and MGISEQ-2000 platforms, respectively (Supplementary Table 4). The assembly is equivalent to 96.64% of the whole genome (54.58 Gb/1C) based on flow cytometry (Supplementary Fig. 3 and Table 3) or 77.75% of the whole genome (67.85 Gb/1C) based on *K*-mer analysis with short reads (Supplementary Fig. 4). We assessed the completeness and accuracy of the assemblies using three different methods: BUSCO gene set alignment by compleasm[13], *K*-mer-based quality assessment by Merqury[14], and mapping of genomic short reads and full-length transcripts. The results of compleasm showed 97.77% of the 1,614 BUSCO genes in complete form in the genome, while Merqury showed a base-level quality value (QV) of 46.01 (indicating accuracy higher than 99.99%). In addition, 99.95% of the genomic short reads could be mapped to the assembly and covered 98.32% of the assembly, and 97.06% of the 177,874 full-length transcripts could be mapped to the assembly using BLAT with the maximum intron size set to 500 kb. All the above results indicated high accuracy and integrity of the assembly. Repeat sequence annotation showed that the assembly was mainly composed of repetitive sequences (96.8%), including 36.21% of *Ty3/Gypsy* and 21.61% of *Helitron* (Supplementary Table 5). Gene annotation resulted in a set of 82,505 protein-coding genes, 93.47% of which had functional annotations (Supplementary Table 6).

With a hierarchical bottom-up chromosome assembly strategy comprising temporary scaffolding, chromosome grouping and iterative chromosome anchoring (Extended Data Fig. 1), we successfully anchored 90.5% of contigs (47.7 out of 52.75 Gb) to the 5 chromosomes, with the largest chromosome being 14.14 Gb in length (Fig. 1a, Table 1 and Supplementary Table 7; please refer to Supplementary Notes 1–3 for more details on chromosome assembly). To validate the quality of chromosome assembly, we carried out fluorescence in situ hybridization (FISH) experiments using 9 probes designed for putative centromeres, telomeres, 5.8S\18S\28S\5S ribosomal DNA (rDNA) and 3 chromosome-specific repeat sequences, respectively (Extended Data Fig. 2, Supplementary Figs. 5 and 6 and Table 8). The chromosomal locations of the probe signals aligned well with the chromosome assembly. Consistent with the hypothesis that centromeres have high mutation rates, we have found many different tandem repeats near the constriction region of chromosomes and chose the most

https://doi.org/10.1038/s41477-025-02060-3

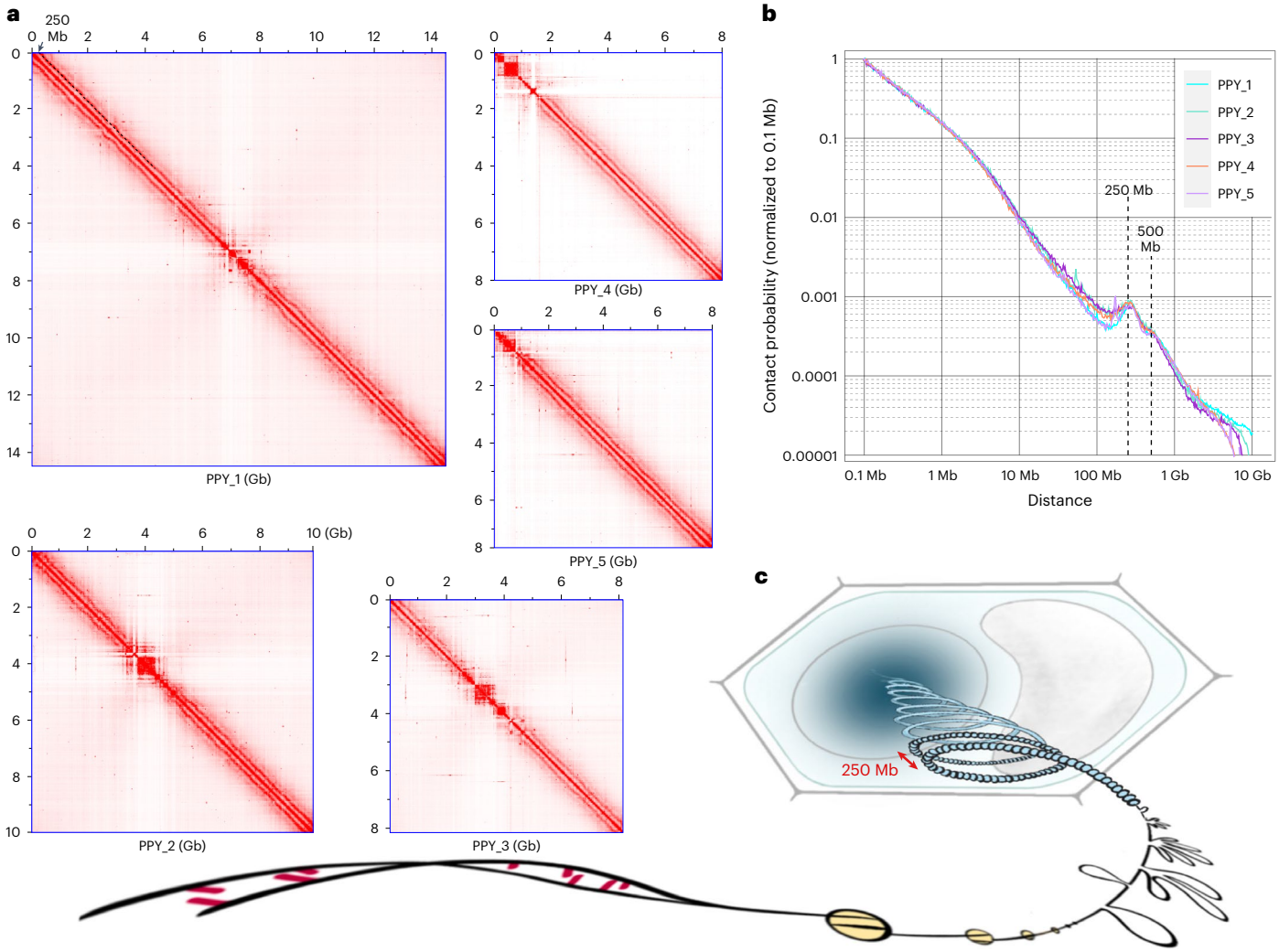


Fig. 1 | Overview of *P. polyphylla* chromosomes and higher-order structure. **a**, Hi-C interaction heat map of the 5 chromosomes in *P. polyphylla*. The *P. polyphylla* chromosomes are labelled PPY_1 to PPY_5 following descending chromosome size with parallel secondary diagonal lines (~250 Mb away from main diagonal), suggesting a helical chromatin structure. **b**, Normalized interaction intensity plots for each of the 5 *P. polyphylla* chromosomes. Genome-wide interaction intensity was plotted over genomic distance from 0.1 Mb to 10 Gb after normalization to the interaction intensity at 0.1 Mb by setting interaction intensity at 0.1 Mb as 1. The dashed lines mark a peak at 250 Mb, representing the interaction over one helical loop. **c**, Model of *P. polyphylla* chromatin structure. Third-order hierarchical folding structure of *P. polyphylla* chromatin is shown from the nucleosome (at periods of 200 bp) to helices (at periods of ~250 Mb).

prevalent 79-mer repeat as the target to design oligo probes for FISH. The results showed that the centromere probe signals were found on all 5 pairs of chromosomes near the chromosomal contractions except the longest pair of chromosomes (PPY_1) and one of the third pair of chromosomes (PPY_3). These results showed not only high diversity in centromere sequences among chromosomes[15,16], but also high polymorphism between homologous chromosomes, which has been recently identified by comparison of centromeres from two complete human genomes[17], suggesting that dynamic evolution of centromeres is going on in *P. polyphylla*.

For *V. dahuricum*, we obtained a 3.55 Gb/1C genome assembly with a contig N50 of 2.21 Mb and QV of 42.85 using HiFi reads (Supplementary Table 4), and further anchored 3.26 Gb (91.83%) of contigs to 16 chromosomes with Hi-C data (Supplementary Fig. 7 and Table 4). Repeat and gene annotation of the assembly identified 85.78% of the genome as repetitive sequences (including a higher 39.74% of *Ty3/Gypsy* and a lower 1.71% of *Helitron* component than in *P. polyphylla*) and 104,721 protein-coding genes, 99.3% of which had functional annotations (Supplementary Tables 5 and 6), more than the numbers of total and functional protein-coding genes in *P. polyphylla*.

Putative helical chromatin structure in giant chromosomes of *P. polyphylla*

The Hi-C data not only assisted our chromosome-level assembly but also provided novel insights into the three-dimensional (3D) structural information of the *P. polyphylla* genome. In the Hi-C interaction heat map (Fig. 1a), we observed a pair of secondary diagonals ~250 Mb away from the main diagonal. This suggests that these genomic parts interact intensely with neighbouring regions on the genome, contributing to the signals on the main diagonal, and also with other genomic regions at a constant distance away, which contribute to the parallel secondary diagonals. In addition, outside of the secondary diagonals, a pair of weak diagonals was just discernible. A plot of the Hi-C interaction intensities over genomic distance revealed a peak at ~250 Mb and a shoulder at ~500 Mb (Fig. 1b). A helical model of a higher-order chromatin structure (Fig. 1c) offers a compelling explanation for the pattern of multiple parallel diagonals in the Hi-C interaction map in *P. polyphylla*. In this model, sequences distant from each other on the same chromosome are brought into close proximity in 3D space due to their location on the same side of neighbouring loops on a helix. When the helix is highly compressed, similar to a compressed spring, points

on the same side of neighbouring three or more loops can be brought even closer together, accounting for the existence of diagonals beyond the secondary ones. We also observed similar secondary diagonals in the published Hi-C data from leaf tissues of ginkgo[18], suggesting that the chromatin helical structure in interphase cells is not a special case limited to *P. polyphylla*. It is worth noting that secondary diagonals along the main diagonals in Hi-C interaction maps have previously been observed only in purified metaphase cells (not in interphase cells) in barley[19], humans[20], chickens[21] and axolotls[22]. These findings are interpreted as evidence of a higher-order helical structure of metaphase chromosomes. In these studies, the distance between the primary and secondary diagonals varied from 12 Mb to 35 Mb, representing the genomic length of each helical turn.

In our research, the distance between these diagonals in *P. polyphylla* was ~250 Mb (Fig. 1b,c), significantly larger than that observed in purified metaphase cells (<35 Mb). We determined the Hi-C interaction intensity between pairs of 10-Mb bins separated by a distance of 240–260 Mb. When we plotted the intensity distribution across chromosomes, it was consistently around 10,000 pairs of Hi-C reads/bin pair in most chromosome regions, with a noticeable decrease at some centromeres and telomeres (Supplementary Fig. 8). This suggests that parallel diagonals are prevalent in most areas of the *P. polyphylla* chromosomes and only absent in regions of high repetitive heterochromatin. Further sequence analysis of the regions delineated by secondary diagonals revealed no specific repeat type enrichment or significant difference in gene density compared with the whole genome (Supplementary Fig. 9). This helical structure might represent a unique 3D conformation of large chromosomes in interphase nuclei, potentially preventing DNA entanglement and promoting efficient gene transcriptional regulation. To validate this hypothesis and investigate its functional implications, further experiments, including high-resolution Hi-C and microscopy analyses of selected cells in both mitosis and interphase, are essential. Such studies will contribute to our understanding of the high-order chromatin structure of large chromosomes.

Contrasting histories of polyploidization between *P. polyphylla* and *V. dahuricum*

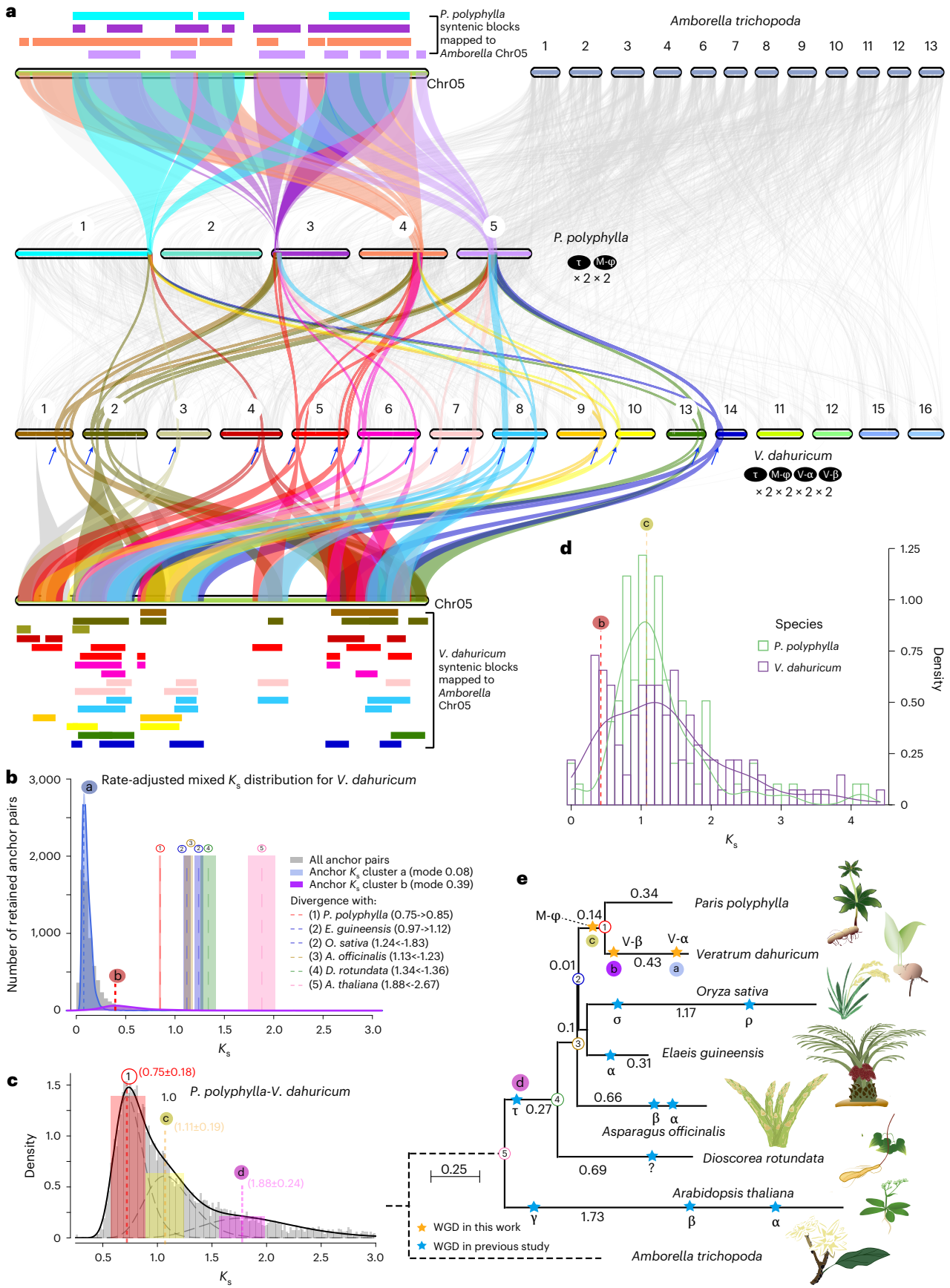
Polyploidization, which instantly and significantly increase genome size, has been a key factor in the evolution of numerous plant lineages[23–25]. However, its long-term contribution to genome size remains unclear[26,27]. We reconstructed the polyploidization history in the lineages leading to *P. polyphylla* and *V. dahuricum* by using *Amborella trichopoda* as outgroup. This was achieved by combining evidence of genome synteny within and between species with the distribution of synonymous substitutions per synonymous site (K_s) between paralogous genes retained in syntenic blocks (anchor pairs) within each species. Surprisingly, our synteny analysis within each species did not reveal any recent polyploidization events post divergences in the giant genome of *P. polyphylla* (Supplementary Fig. 10). However, it did uncover two polyploidization events (V-α and V-β) in the smaller genome of *V. dahuricum* (Supplementary Fig. 11), although the synteny blocks corresponding to the older V-β event accounted for only 14% of the genome (Supplementary Table 9). Interspecies collinearity revealed a syntenic depth ratio of 4:1 between *P. polyphylla* and *A. trichopoda* (Supplementary Fig. 12) and 10:1 between *V. dahuricum* and *A. trichopoda* (Supplementary Fig. 13), suggesting two ancient whole-genome duplication (WGDs) for *P. polyphylla* and at least three WGDs for *V. dahuricum*. Notably, the syntenic depth ratio between *P. polyphylla* and *V. dahuricum* stands at 1:5 (Supplementary Fig. 14) rather than 4:X, indicating extensive recombination and gene loss after their divergence. We used Chr05 of *A. trichopoda* as the query to demonstrate a syntenic depth of 1:4:16 for *A. trichopoda*:*P. polyphylla*:*V. dahuricum* (Fig. 2a). Hence, we infer that *P. polyphylla* underwent two ancient WGD events (τ and M-φ shared with *V. dahuricum*) and *V. dahuricum* experienced two additional WGD events (V-α and V-β) post divergence from *P. polyphylla*.

The polyploidization history was further cross-validated using the K_s distribution of intraspecies anchor pairs. Significant K_s distribution peaks indicative of polyploidization events were detected and aligned with the K_s peaks of orthologous gene pairs between species, which enabled localization of the polyploidization events on the branches of the phylogenetic tree (Fig. 2b–e). The phylogenetic relationships among Liliales (*P. polyphylla* and *V. dahuricum*), Commelinids (*Oryza sativa* and *Elaeis guineensis*) and Asparagales (*Asparagus officinalis*) in Fig. 2e were determined by phylogenomic analysis, and are different from those inferred by previous chloroplast genome- and transcriptome-based phylogenies (Supplementary Note 4). K_s analysis of intraspecies comparison in *V. dahuricum* identified two peaks (K_s = 0.08 and 0.39) after its divergence from *P. polyphylla* (K_s = 0.75 and adjusted to 0.85 after rate adjustment), which represent two polyploidization events (V-α and V-β) (Fig. 2b). Two peaks were also identified in the interspecies K_s analysis between *P. polyphylla* and *V. dahuricum*, indicating two shared ancestral polyploidization events (τ and M-φ) (Fig. 2c and Supplementary Fig. 15). Furthermore, we calculated the K_s of paralogue gene pairs in *P. polyphylla* and *V. dahuricum* identified by collinear orthologues of chr05 in *A. trichopoda* (Fig. 2a). The K_s distribution of these paralogue in *V. dahuricum* also showed two peaks. The peak at K_s = 0.4 in *V. dahuricum* is highly consistent with the V-β event and the peak around K_s = 1.1 in both *V. dahuricum* and *P. polyphylla* represents the M-φ event in the common ancestor of the two Melanthiaceae species (Fig. 2d).

Our results confirmed a common WGD (M-φ) before the divergence of *P. polyphylla* and *V. dahuricum*, followed by two polyploidizations in *V. dahuricum* but no recent polyploidization in *P. polyphylla* since their divergence (Fig. 2e). This finding offered a convincing explanation for the large differences in gene number between the two species

Fig. 2 | Genome synteny between *P. polyphylla* and *V. dahuricum* and their contrasting WGD histories. a, Interspecies synteny plotted with JCVI[76]. Links represent collinearity among *P. polyphylla*, *V. dahuricum* and *A. trichopoda*. The coloured ribbon represents the 1:4:16 collinearity between chromosome 5 of *A. trichopoda*, 4 chromosomes of *P. polyphylla* and 12 chromosomes of *V. dahuricum*. Coloured rectangles represent the corresponding syntenic blocks mapped to *Amborella* chromosome 5. Blue arrows point to the 16 homologous copies of *V. dahuricum*. **b**, WGD signatures in paralogue K_s distributions within the *V. dahuricum* genome using the 'ksrates' software. Block-wise median K_s values for each syntenic block (Supplementary Fig. 11) were calculated and plotted in grey histogram. Two putative WGD events could be inferred (two peaks in blue and purple, with dashed lines to mark the modes of these components as age estimates of WGD events). The mode of rate-adjusted K_s of the nodes labelled with numbers on the phylograms in **e** are marked with coloured boxes, indicating a range of ±1s.d. **c**, Mixed modelling of K_s distribution between *P. polyphylla* and *V. dahuricum*. Based on 11,453 anchor pairs between *P. polyphylla* and *V. dahuricum*, mixed modelling of K_s distribution using the 'wgd' software showed three components, labelled in **e**, representing divergence between *P. polyphylla* and *V. dahuricum*, an M-φ event in the common ancestor of the two Melanthiaceae species and an older τ event. Coloured rectangles mark the range of mean ± s.d. for each component. **d**, K_s distribution of paralogue gene pairs indirectly inferred from collinear orthologues of *A. trichopoda*. The homologous chromosome pairs from the most recent polyploidization in *V. dahuricum* were excluded from the analysis. The two K_s distributions have two peaks,with the peak at K_s = 0.4 of *V. dahuricum* coinciding well with the 'b' event in **b**, and the peak at K_s = 1.1 shared by the two species coinciding with the 'c' event in **c**. **e**, Inferred WGD histories of *P. polyphylla* and *V. dahuricum* are marked on the phylogeny of six representative monocot species and *Arabidopsis*, with outgroup *A. trichopoda*. Branch lengths are in proportion to K_s values. WGD events from other studies are marked on relevant branches.

https://doi.org/10.1038/s41477-025-02060-3



https://doi.org/10.1038/s41477-025-02060-3

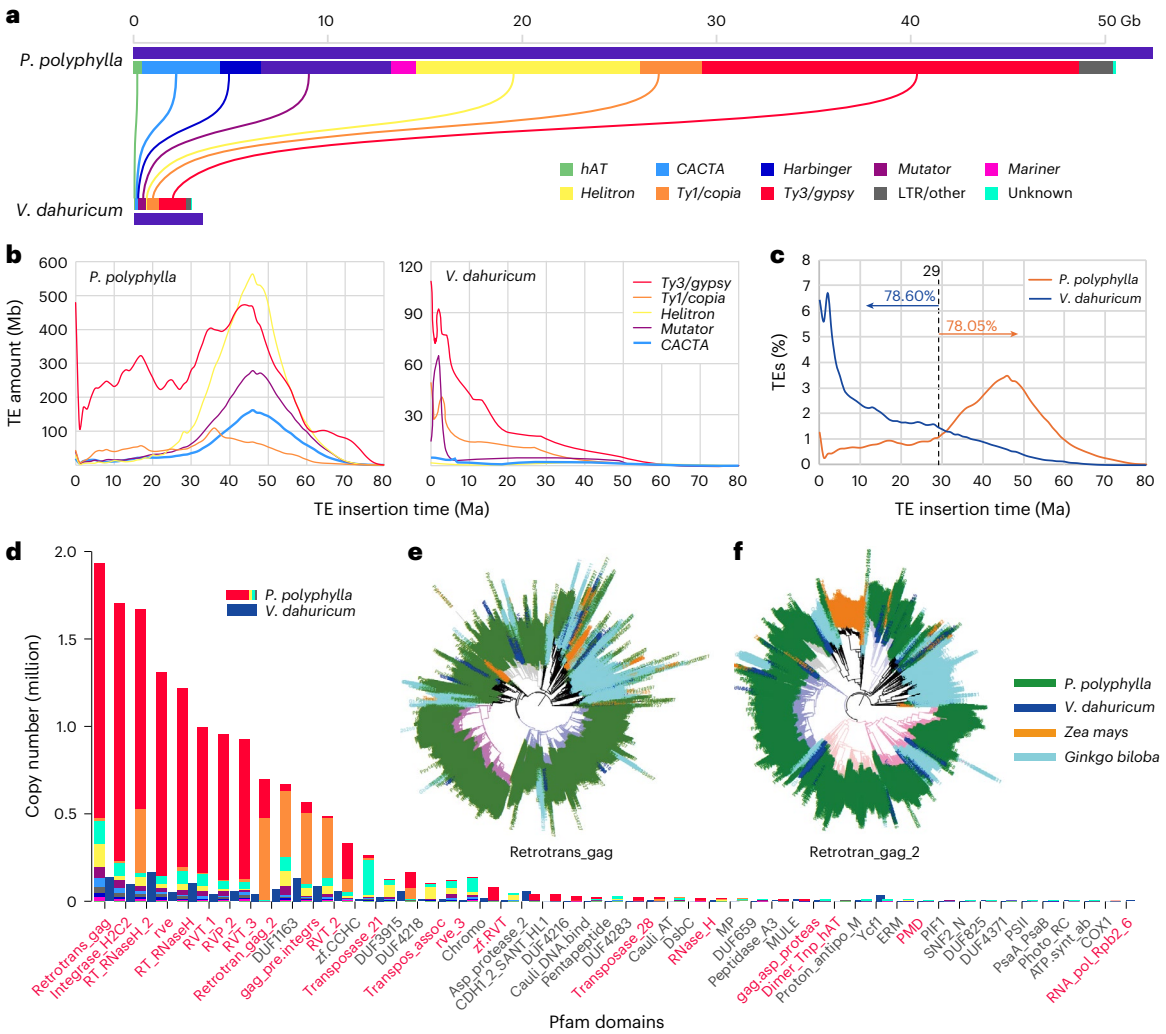


Fig. 3 | Different transposable element (TE) insertion histories in two Melanthiaceae species. a, TE family composition in two genomes. **b**, TE insertion time of the five most abundant TE superfamilies in *P. polyphylla* and *V. dahuricum*. **c**, Proportional distribution of TEs in the two species. The two curves intersect at ~29 Ma, with 78.05% of the TEs in *P. polyphylla* older than this time and 78.60% of the TEs in *V. dahuricum* younger than it. **d**, Bar plot showing copy numbers of the 50 most abundant protein domains in the two species. Bars representing domains from the *P. polyphylla* genome are further subdivided and filled with different colours representing copy numbers of the different TE superfamilies. Colour code as in **a**. Domains with names in red are TE-associated protein domains. **e**,**f**, Phylogenetic relationships among copies of two representative gag polypeptide of LTR, Retrotrans_gag (**e**) and Retrotran_gag_2 (**f**), respectively representing retrotransposons of the *Ty3/gypsy* and *Ty1/copia*.

and revealed the reason why the species with a smaller genome has more genes. Therefore, it could be inferred that mechanisms other than WGD were responsible for the genome expansion in *P. polyphylla*.

Contribution of repetitive sequences to genome expansion

To explore the possible contribution of repetitive sequences to the huge genome of *P. polyphylla*, we identified TEs in both genomes on the basis of de novo annotation and searching of entries in the Repbase TE library. The total repetitive sequences content amounted to 96.80% and 85.78% (including 21.61% and 1.71% *Helitron* components) for the *P. polyphylla* and *V. dahuricum* genomes, respectively (Fig. 3a and Supplementary Table 5). We further corroborated these findings using a whole-genome protein domain search, revealing that TE-associated domains were the most prevalent within the genome of *P. polyphylla* (Fig. 3d). An extended protein domain search across 78 plant species with sequenced genomes demonstrated a significant positive correlation between the quantity of TE domains and genome size (Supplementary Table 10 and Extended Data Fig. 3). This underscores the pivotal role of TE insertions in determining genome size diversity among plants. Notably, certain *P. polyphylla* domains, such as RVT_1 and RVT_2, were below the regression line, suggesting a lower copy number than expected. This discrepancy might arise from the pronounced divergence of ancient TEs, rendering them elusive for homologue identification. In a previous study focusing on repeat content using unassembled reads, it was observed that the proportion of repeat content does not continuously increase with an increase in genome size, particularly in plant species with large genomes[3]. Our assessment of repeat sequences in the *P. polyphylla* genome surpasses previous estimations, which were based on short-reads clustering with Repeatexplorer2 in two other *Paris* species, ranging from 76% to 83%[3]. This difference could be attributed to both methodological distinctions and differences in repeat definition (see Supplementary Note 5 for further details).

In the two genomes, the most abundant category of TEs was the long terminal repeat (LTR) *Ty3/gypsy*, constituting 36.21% (19.1 Gb) and 39.74% (1.41 Gb) of the *P. polyphylla* and *V. dahuricum* genomes, respectively (Fig. 3a). A divergence analysis of the top 5 TE categories (*Ty3/gypsy*, *Helitron*, *Mutator*, *CACTA* and *Ty1/copia*) that compose 83.69% of the *P. polyphylla* genome exposed a major broad peak at ~45 Ma in the temporal distribution of TE insertions. These data imply that ancient TE bursts facilitated the predominance of TE content in

the *P. polyphylla* genome (Fig. 3b). In contrast, the TEs in the *V. dahuricum* genome displayed only narrow peaks below 5 Ma, indicating that the majority of TEs in this genome are relatively recent (Fig. 3c). The proportional distribution of TEs in the two species revealed an intersection point at 29 Ma, indicating that 78.05% of the TEs in *P. polyphylla* were inserted earlier than 29 Ma, whereas 78.6% of TEs in *V. dahuricum* were inserted later than this time (Fig. 3c). These results suggest that ancient TE bursts contributed to the dominance of TE content in the *P. polyphylla* genome, while most TEs in the *V. dahuricum* genome are very young. Given that *Ty3/gypsy* and *Ty1/copia* are the two most abundant LTR categories in *P. polyphylla* (Supplementary Fig. 6 and Table 5), we constructed the phylogeny of these LTRs on the basis of the alignments of 4 LTR functional domains from the genomes of *P. polyphylla*, *V. dahuricum* and two outgroups (*Zea mays* and *Ginkgo biloba*). The results showed that the majority of *Ty3/gypsy* and *Ty1/copia* were clustered into species-specific clades rather than into shared clades across species. This suggests that the majority of LTRs emerged after the split of the two species. Moreover, numerous *P. polyphylla*-specific clades exhibited a brush-like tree topology (Fig. 3e,f and Supplementary Fig. 16), implying that many LTR duplication events occurred within a short period of time, accounting for the TE insertion peak (Fig. 3b). Previous research on the large genome of *Fritillaria* species within Liliales found that most TE families exhibited high divergence and low copy number. This suggests that the vast genome in *Fritillaria* plants arose from TE accretion rather than from recent extensive TE amplification, indicating inefficient DNA removal mechanisms in these plants[28]. We noted that distinct TE categories accumulated differently in *P. polyphylla*; for example, some TEs such as *CACTA* accumulated slowly with a flat broad peak, whereas other TEs such as *Helitron* accumulated rapidly with a sharp peak. In particular, *Helitron*, unlike all other types of TE entry in the de novo TE library, is a Class II TE, which generally lacks typical transposon features[29]. *Helitron* has significantly higher repeat copies than other types of TE (Supplementary Fig. 17), accounting for >21% of the *P. polyphylla* genome. Therefore, the genome expansion in *P. polyphylla* seems to be the consequence of both ancient and recent amplification, as well as both rapid and slow TE accumulation, suggesting that inefficient DNA removal might also contribute to its huge genome.

TE insertions have not only played an important role in *P. polyphylla* genome expansion but have also had an impact on the structure of protein-coding genes. We found that the TE insertions in introns increased the length of many genes, resulting in a bimodal gene length distribution (Supplementary Fig. 18 and Note 6).

Genetic and epigenetic changes accompanying genome expansion

Greater copy number and unique sequence changes in *P. polyphylla* genes related to nucleotide repair-associated pathways. Maintaining chromosome stability is particularly challenging for the 10-Gb chromosomes of *P. polyphylla*. Several DNA repair pathways, including base excision repair, nucleotide excision repair (NER), mismatch repair, homologous recombination and non-homologous end joining, play a crucial role in safeguarding the genome against damage from various mutagens and preserving stability[30]. Our study consistently revealed that all these pathways were significantly enriched (Supplementary Fig. 19) in both expanded gene families in *P. polyphylla* in comparisons among 11 angiosperm species (Supplementary Tables 11 and 12), and those genes with non-synonymous substitutions correlated with genome size in the analysis on 49 plant species (Supplementary Tables 13–15).

Our analysis, based on Café[31] in 10 monocots with *Arabidopsis* as the outgroup (Supplementary Table 11), detected 1,793 expanded gene families in *P. polyphylla* (Supplementary Fig. 20). We conducted Kyoto Encyclopedia of Genes and Genomes (KEGG)[32] pathway enrichment analysis on these 1,793 gene families and found significant enrichment in all 5 DNA repair pathways (Supplementary Fig. 19 and Table 12). In addition to examining gene family size, we also conducted protein sequence variation analysis on orthologous genes across 49 plant genomes of varying sizes (Supplementary Table 13). These genomes are categorized into two groups: one group consists of 37 small genomes with GS < 2.45 Gb/1C, while the other group comprises 12 large genomes with GS > 2.45 Gb/1C. Through this analysis, we identified 731 orthologous gene families that exhibit significant variations across different genome size groups (Supplementary Table 14 and Note 7). Among these 731 gene families, we again observed significant enrichment in the 5 DNA repair pathways (Supplementary Fig. 19 and Table 15), suggesting convergent evolution of key genes in numerous plant lineages with large genomes (GS > 2.45 Gb/1C).

In the NER pathway, *P. polyphylla* exhibited the highest gene count (216) across all DNA repair pathways and plant species under consideration (Supplementary Table 16). This prompted a deeper investigation into gene family expansion and sequence evolution within this pathway for *P. polyphylla*. A total of 61 *P. polyphylla* genes (compared with 36 *V. dahuricum* genes) from 15 expanded families were mapped to 10 KEGG entries in the NER pathway (Fig. 4a,b, and Supplementary Tables 16 and 17). Furthermore, 2 genes in the NER pathway displayed mutations unique to *P. polyphylla*, while 9 others demonstrated rapid evolution. These findings suggest that both new gene copies and novel mutations might have influenced the evolutionary trajectory of the NER pathway in *P. polyphylla*, potentially as a response to genome expansion and the challenges of maintaining genome stability.

Higher DNA 5mC methylation in *P. polyphylla* than in *V. dahuricum*. The process of DNA methylation is essential in silencing transposons and maintaining genome stability[33]. In line with the higher TE content in *P. polyphylla* compared with *V. dahuricum*, *P. polyphylla* demonstrated a greater proportion of 5-methylcytosine (5mC) at CpG dinucleotides (94.24%) than *V. dahuricum* (90.72%) based on nanopore signals. Both proportions surpass that of maize, which is 86%[34]. To delve deeper into the methylation levels, we categorized all methylated CpG sites according to the level of methylation at those sites, defined as the ratio of methylated CpG reads to total reads at the same site. More than half of the sites (61.53% in *P. polyphylla* and 63.59% in *V. dahuricum*) displayed full methylation (100%, that is, all reads at the site were methylated). Meanwhile, sites exhibiting zero to intermediate methylation levels (0%–70%) constituted 4.77% and 9.20% of all CpG sites in *P. polyphylla* and *V. dahuricum*, respectively. In addition, sites with high methylation levels (71%–99%) made up 33.71% and 27.21% of all CpG sites in *P. polyphylla* and *V. dahuricum*, respectively, and these formed a peak around the 94% methylation level (Fig. 4c and Supplementary Table 18). Notably, compared with *P. polyphylla*, the *V. dahuricum* genome had a higher percentage of unmethylated sites at 0% (2.67% vs 0.51% in *P. polyphylla*) and fully methylated sites at 100% (63.59% vs 61.53% in *P. polyphylla*) (Fig. 4c). For high methylation levels between 90% and 99%, the proportion of sites at each methylation level was generally higher in *P. polyphylla* compared with *V. dahuricum*.

Our genome-wide 5mC analysis revealed that while *P. polyphylla* exhibited higher overall DNA methylation, *V. dahuricum* demonstrated a greater percentage of sites with 100% methylation. This is probably attributable to the latter's higher ratio of recently inserted TEs with low divergence, which are rigorously methylated via the RNA-directed DNA methylation pathway. When compared with maize, the *P. polyphylla* genome possesses a greater TE proportion, whereas the *V. dahuricum* genome's TE proportion is similar to that of maize at 85%[35], but it has an elevated ratio of TE at low sequence (~1%) (Supplementary Fig. 21). The increased TE proportion in *P. polyphylla* and the more recent TE duplications in *V. dahuricum* may account for their heightened methylation levels compared with maize. The spontaneous molecular process of 5mC deamination in methylated repeat regions can lead to C-to-T

https://doi.org/10.1038/s41477-025-02060-3

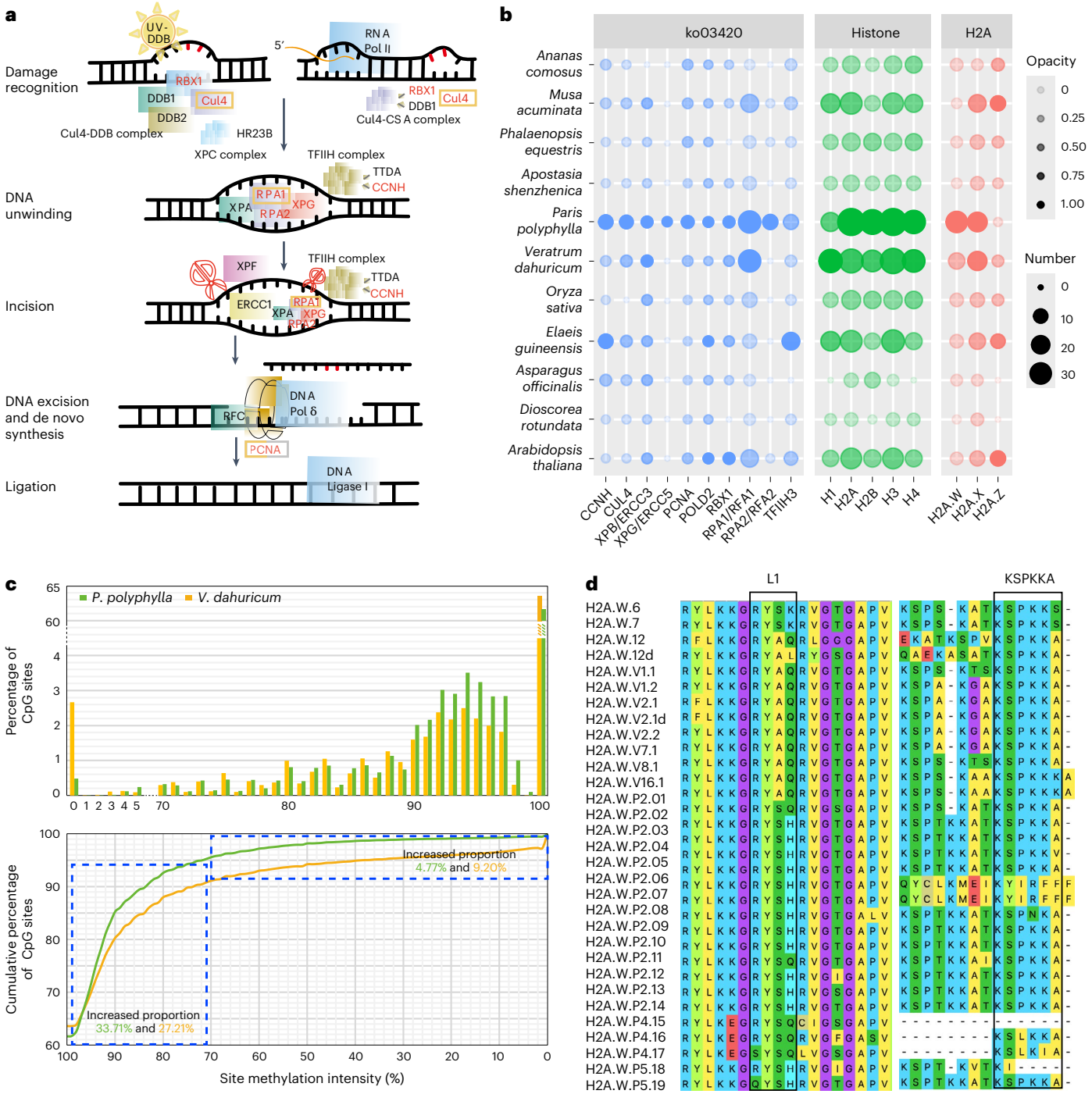


Fig. 4 | Genomic features counteracting TE activity in the *P. polyphylla* genome. a, Genes of the nucleotide excision repair pathway. Genes in red are from gene families that evolved more copies in *P. polyphylla*. Genes in grey and yellow line boxes represent positively selected genes and rapidly evolving genes, respectively. CCNH, cyclin H; CSA, DNA damage-binding protein; Cul4, cullin 4; DDB1/2, DNA damage-binding protein 1/2; ERCC1, DNA excision repair protein ERCC-1; HR23B, UV excision repair protein RAD23; PCNA, proliferating cell nuclear antigen; Pol δ, DNA polymerase delta subunit 3/4; RBX1, RING-box protein 1; RFC, replication factor C subunit 1; RPA1/2, replication factor A1/2; TFIIH, transcription initiation factor TFIIH subunit; TTDA, TFIIH basal transcription factor complex TTD-A subunit; XPA, DNA-repair protein complementing XP-A cells; XPC, xeroderma pigmentosum group C-complementing protein; XPF, DNA excision repair protein ERCC-4; XPG, DNA excision repair protein ERCC-5. **b**, Gene number of different gene families. Eleven angiosperm genomes were used for gene family clustering. Blue dots represent 10 expanded gene families in the ko03420 pathway, green dots represent the 5 families of histone genes and red dots represent the 3 subfamilies of histone H2A. The opacity of a dot represents the relative abundance calculated by dividing the number of genes at that dot by the maximum number of this gene type. **c**, CpG site methylation level in *P. polyphylla* and *V. dahuricum*. Top: distribution of CpG sites plotted over different methylation levels for *P. polyphylla* and *V. dahuricum* genomes. Bottom: methylated proportion of all CpG sites plotted over different cut-off percentages (*x* axis) over which a CpG site could be categorized as methylated. The two blue dashed boxes indicate the increased proportion of methylated sites with methylation intensity dropping from 99% to 71% and from 70% to 0%. **d**, Multiple sequence alignments of H2A.W proteins. The L1 loop region is marked by a black box. KSPKKA is the typical motif at the C terminal of H2A.W. Gene names with 'P' and 'V' indicate copies from *P. polyphylla* and *V. dahuricum*, respectively. Unique amino acid substitutions exist in many *P. polyphylla* H2A.W proteins in the L1 loop.

mutations and a reduction in GC content[36,37], potentially explaining the reduced GC content in the *P. polyphylla* genome.

Increasing copies of histone gene family along the *P. polyphylla* lineage. In addition to exploring the NER pathway and TE silencing via DNA methylation, we also investigated the histone-coding genes within *P. polyphylla*. These genes encode the primary protein components of chromatin and play a crucial role in epigenetic regulation of chromatin structure[38]. Notably, our whole-genome Pfam[39] search results revealed a significant correlation between the copy number of the 'histone' protein domain and genome size (Supplementary Fig. 22a). Moreover, utilizing HistoneDB 2.0 (ref. 40), we identified 136 histones in *P. polyphylla*. This number exceeded that of the other 10 species included in our gene family clustering analysis (Supplementary Fig. 23 and Table 19).

Among all the histone gene families, the most substantial expansion of a histone subfamily in the *P. polyphylla* lineage was discovered in genes encoding H2A.W, a variant of the core nucleosome histone H2A. The evolutionary history of the *H2A.W* subfamily was further elucidated through phylogenetic analysis. On the basis of the maximum-likelihood and neighbour-joining trees of 19 *P. polyphylla* and 9 *V. dahuricum* *H2A.W* genes, along with collinearity patterns among the regions where these genes are located, we hypothesized the presence of 2 ancestral copies of *H2A.W* in each of the ancestral genomes of the *P. polyphylla* and *V. dahuricum* lineages. Nevertheless, these 4 copies probably coexisted in the common ancestor of *P. polyphylla* and *V. dahuricum*, given that their divergence predates the split between the two species. During speciation, incomplete lineage sorting led the *P. polyphylla* and *V. dahuricum* lineages to inherit 2 of the 4 copies each. Subsequently, from these 2 ancestral genes, 8 copies in *V. dahuricum* emerged through two rounds of WGDs, while the ninth copy arose from a recent gene duplication. In contrast, the 19 *H2A.W* genes in *P. polyphylla* can be traced back to a series of gene duplications, notably including highly uneven tandem gene duplication, resulting in 4 copies stemming from one ancestral gene and 15 copies from the other (Supplementary Fig. 24).

The H2A.W molecule possesses a distinctive C-terminal motif, KSPKKA, which participates in the formation of higher-order interfibre interactions contributing to heterochromatin structure. This motif can maintain genomic TE silencing independently of the genome and the histone methylation H3K9me pathway[41,42]. Our analysis revealed that the *H2A.W* genes in *P. polyphylla* exhibit species-specific variations in the loop region of the core histone domain (Fig. 4d and Supplementary Fig. 22b). This particular loop region has been implicated in the dimerization of H2As[43]. Consequently, the unique modification at this motif might enhance the stability of H2A dimers in the nucleosomes, resulting in more stable heterochromatin within the genomic TE regions.

In summary, the above genetic (genes in DNA repair pathways and genes coding for *H2A.W*) and epigenetic (CpG methylation) changes accompanying genome expansion in *P. polyphylla* could be an evolutionary response to maintain the larger genome.

Conclusions

In this study, we report a high-quality assembly of *P. polyphylla* with a total length of 52.75 Gb, boasting the largest chromosome ever reported. Our innovative hierarchical bottom-up chromosome assembly strategy offers a promising method for tackling the assembly of other giant plant genomes. The data from this research augment our comprehension of the composition and chromosomal architecture of exceptionally large genomes (~50 Gb/1C), which have been relatively underexplored. The genome assemblies described herein provide a unique avenue to investigate the mechanisms underlying genome size expansion and stability. Notably, despite diverging only ~73.4 Ma, the two species exhibited a staggering 13-fold disparity in their genome size. Our analyses suggest that TE insertions play a pivotal role in this expansion, particularly influencing the growth of intergenic regions and introns. While WGD may have catalysed the formation of additional protein-coding genes, its impact on overall genome size appears limited in this context, as WGD mainly facilitated an increase in genes in the smaller genome. Moreover, we observed a significant expansion of the *H2A.W* histone gene subfamily in *P. polyphylla*, which is implicated in heterochromatin formation and TE silencing. In addition, the *P. polyphylla* genome displays pronounced methylation, a key component of the TE silencing mechanism.

Methods

Sample collection and usage

The *P. polyphylla* samples were collected from a nursery in Kunming, Yunnan Province, while the *V. dahuricum* samples were collected from native habitats in Jilin Province, China. Fresh *P. polyphylla* leaves from three healthy individual plants were harvested for flow cytometry. A plant labelled 'Leaf-1' was used for whole-genome DNA sequencing using Oxford Nanopore Technology (ONT), Pacbio Revio sequencing and next generation sequencing. Seedling individuals were used for Hi-C and Pore-C libraries and three flowering individuals were used for RNA extraction. For *V. dahuricum*, three mature individuals were collected for flow cytometry, with the plant labelled 'LL-1' used for DNA sequencing, and 'LL-2' and 'LL-3' used for Hi-C and RNA sequencing (RNA-seq), respectively.

Flow cytometry experiment

The genome sizes of *P. polyphylla* and *V. dahuricum* were estimated using flow cytometry. Ginkgo (with a genome size of 10.61 Gb/1C) and tomato (with a genome size of 0.88 Gb/1C) were used as reference plants for *P. polyphylla* and *V. dahuricum*, respectively. Young leaves were used for estimation. Samples were placed in 0.8 ml of pre-cooled mGb dissociation buffer (45 mM $MgCl_2 \cdot 6H_2O$, 20 mM MOPS, 30 mM sodium citrate, 1% (w/v) PVP 40, 0.2% (v/v) TritonX-100, 10 mM Na_2EDTA, 20 µl ml^{-1} β-mercaptoethanol, pH7.5) and finely chopped with a sharp blade. After incubation on ice for 10 min, the sample was filtered through a 40-µm filter to obtain a nuclear suspension. Appropriate volumes of propidium iodide (PI) and RNAase solution were then added to the suspension to achieve a working concentration of 50 µg ml^{-1} for both, followed by staining in the dark for 0.5–1 h. The stained sample suspension was mixed with the reference sample suspension in an appropriate ratio and analysed using a BD FACScalibur flow cytometer at 488 nm excitation. A total of 10,000 particles were collected for each sample, with a coefficient of variation (c.v.%) controlled to within 5%. Data analysis was conducted using Modifit 3.0 software. PI dye intercalates into the DNA double strand, and under 488 nm excitation, it emits fluorescence proportional to the DNA content. By comparing the fluorescence intensity peaks of the sample and the reference plant, and using the C-value of the reference, the genome size of the sample can be calculated using the formula: DNA content of sample = DNA content of reference × (fluorescence intensity of sample/fluorescence intensity of reference). We measured the genome sizes of *P. polyphylla* and *V. dahuricum*, with resulting average of three replicates of 54.58 Gb/1C and 3.93 Gb/1C, respectively (Supplementary Table 3 and Fig. 3).

Genome sequencing

High-quality genomic DNA was extracted from *P. polyphylla* leaves and *V. dahuricum* stems, respectively. The extracted DNA was used for genomic library construction following the protocols of the different sequencing platforms. For *P. polyphylla* long-read sequencing, genomic DNA libraries were sequenced on the Revio and PromethION platforms, obtaining 1.83 Tb HiFi reads and 2.95 Tb of QC-passed long-read sequencing data (including 57.28 Gb ultra-long reads and 60.64 Gb Pore-C reads with restriction enzyme DpnII), respectively. For *V. dahuricum*, two HiFi libraries were sequenced using the circular consensus sequencing (CCS) mode of the PacBio Sequel System, generating 55.03 Gb of self-corrected long-read data. We also sequenced 192.49 Gb

of ONT reads. For short-read sequencing, paired-end libraries with an insert size of 250 bp were constructed and sequenced using the 150-bp paired-read mode, resulting in 1.81 Tb of clean data for *P. polyphylla* on the DNBSEQ-T7 platform and 135 Gb of data for *V. dahuricum* on the Illumina NextSeq 500 platform. For Hi-C, the sequencing methods were the same as for short-read sequencing, resulting in 5.90 Tb and 436.84 Gb of Hi-C (DpnII) data for *P. polyphylla* and *V. dahuricum*, respectively (Supplementary Table 4).

Estimation of genome size based on *k*-mer frequency

We estimated genome size on the basis of *k*-mer frequency of short-read sequencing data using Jellyfish (v.2.2.10)[44] with commands 'jellyfish count -m K -s 10000 M -t 20 -c 7 -C' and 'jellyfish histo', where *K* is the length of oligomers (*K* = 21 for *V. dahuricum* and *K* = 39 for *P. polyphylla*). The frequency results were then used to calculate genome size with GenomeScope2, resulting in 67.85 Gb/1C for *P. polyphylla* and 4.07 Gb/1C for *V. dahuricum* (Supplementary Fig. 4).

Genome contig building, scaffolding and chromosome construction

Hifiasm (0.19.8-r603 with default parameters), a de novo assembler for PacBio HiFi reads[45], was used to perform de novo genome assembly for both *P. polyphylla* and *V. dahuricum*. On the basis of 33.28× and 14× depth of HiFi reads, we generated contigs of 56.34 Gb/1C (N50 1.18 Mb) and 3.55 Gb/1C (N50 2.21 Mb) for *P. polyphylla* and *V. dahuricum*, respectively. For *V. dahuricum*, we aligned 111× depth of Hi-C data (Supplementary Table 4) to the contigs using Juicer (v.1.6)[46] and used 3D-DNA (v.180922)[47] to anchor the contigs with the interaction information. We visualized the interaction data among contigs using Juicebox (v.1.11.08)[48,49] to manually correct and re-order the inappropriate 3D-DNA contig placements into the final 16 chromosomes.

For *P. polyphylla*, we added ONT and Hi-C data to further extend the contigs by local hybrid assembly. Briefly, we first mapped the 107× Hi-C data, 33.28× HiFi reads and 53.71× ONT reads to the contigs of the HiFi assembly. On the basis of the Hi-C alignment results, the contigs were divided into 9 groups using the cluster module of HapHiC[50] (v.1.0.3, parameters: –RE AAGCTT –threads 10 –aln_format bam –correct_nrounds 2 –bin_size 1000). For each group of contigs, all the HiFi and ONT reads mapped were subjected to local hybrid assembly using Hifiasm, scaffolding by ONT reads using Longstitch[51] and polishing by HiFi reads with Racon[52]. Finally, we mapped the polished scaffolds in the 9 groups with the initial HiFi-based assembly using minimap2 (ref. 53), retained the longer sequences of the two assemblies, combined these sequences with the contigs that were not clustered by HapHiC to purge haplotigs using purge_dups[54], and obtained the final assembly (52.75 Gb/1C, contig N50 1.52 Mb).

For the ~10-Gb-long chromosomes of *P. polyphylla*, we could not concatenate the contigs into chromosomes accurately using JuiceBox due to frequent programme freezing, even when using a 128-CPU and 256-Gb memory computer (Extended Data Fig. 1a–d). Thus, we developed a hierarchical bottom-up chromosome assembly strategy to solve this issue. First, all contigs were split at assembly errors using YaHS[55] and ordered using 3D-DNA on the basis of the Hi-C interaction maps. Then, adjacent contigs with strong Hi-C interaction signals were linked as temporary scaffolds (Extended Data Fig. 1e,f). Finally, ordering was performed for all 19,902 linked temporary scaffolds (size >100 Kb and total length 52.09 Gb). Using manual adjustment, temporary scaffolds were grouped into 5 giant chromosomes (Extended Data Fig. 1g). For each chromosome, we broke down the temporary scaffolds of each chromosome into contigs (Extended Data Fig. 1h) and then ran 3D-DNA to obtain the anchored contigs of the chromosome, as well as the un-anchored contigs. We next calculated the interaction strengths between all un-anchored and anchored contigs of the 5 chromosomes to determine the corresponding strongest interacting anchored contigs, and then reassigned the un-anchored contigs to the chromosome where the strongest interacting anchored contigs were located. Finally, we repeated the anchoring of contigs for each chromosome (Extended Data Fig. 1i). Lastly, on the basis of the assembly files of the 5 chromosomes and all un-anchored contigs, we ran the 3D-DNA pipeline shell 'run-asm-pipeline-post-review.sh' to generate chromosomal sequences and the whole-genome Hi-C interaction map.

Completeness and accuracy of the assembly

The software compleasm (v.0.2.6) was used to perform completeness evaluation, with parameters '-t 50 -l embryophyta_odb10'. The software Merqury (v.1.3) was used to perform accuracy evaluation. First, we counted the *k*-mer using the command 'meryl *k* = 21 memory=150 threads=30' for short reads and merged *k*-mer with the command 'meryl *k* = 21 memory=180 threads=60 union-sum'. Then, we used 'merqury.sh' to calculate the quality value (QV) of the assembly.

Repetitive sequence annotation

To identify the repetitive sequences in the genomes, we first searched simple tandem repeats in the genome using Tandem Repeats Finder[56] (v.4.09, with parameters '-d -h Match=2 Mismatch=7 Delta=7 PM=80 PI=10 Minscore=50 MaxPeriod=20'). We then annotated the TEs on the basis of two strategies: homology-based and de novo prediction. For homology-based prediction, we applied RepeatMasker (v.4.1.0, http://www.repeatmasker.org/) and its RepeatProteinMask package to search the genome with the nucleotide and protein libraries of TEs curated by Repbase (https://www.girinst.org/repbase). For de novo prediction, we first built a non-redundant de novo TE library with the Extensive de novo TE Annotator (EDTA v.1.9.4) pipeline[57] following the classification system based on a curated library of rice TEs (https://github.com/oushujun/EDTA/tree/master/database). Among all the entries in the de novo library, *Helitrons* were predicted by the HelitronScanner software[58] integrated in the EDTA pipeline and were classified to two types using TEsorter software[59]: (1) autonomous *Helitron* with *Helitron* encoding Rep/Helicase and RPA proteins, and (2) non-autonomous *Helitron* without Rep/Helicase and RPA proteins. Next, we searched the genome with this de novo TE library using RepeatMasker (v.4.1.0, rmblastn parameters: -num_alignments 9999999 -gapopen 24 -gapextend 6 -mask_level 101 -complexity_adjust -word_size 9 -xdrop_ungap 450 -xdrop_gap_final 225 -xdrop_gap 112 -min_raw_gapped_score 225 -dust no -num_threads 4). For the generated TEs from the de novo prediction, a minimum similarity of 60% was required to assign a particular sequence to a TE repeat family. Finally, we merged and calculated the total length of all detected TEs.

We used a Perl script (parseRM.pl, available at https://github.com/4ureliek/Parsing-RepeatMasker-Outputs) to parse the raw de novo alignment outputs from RepeatMasker and estimate TE insertion time with the parameters '-l 100,1 -m 0.0025' for *P. polyphylla* and '-l 100,1 -m 0.0026' for *V. dahuricum*. The '-m' parameter was to set substitution rate, which was calculated using r8s. This process uses the corrected percentage of divergence (from these alignment files) of each copy to its matching repeat in the de novo library, and the TE insertion time (in Ma) is equal to divergence / 100 / (substitution rate × 2).

Construction of LTR trees

We selected *Z. mays*[35] and *G. biloba*[60] as outgroups at different phylogenetic distances. Pfam annotation of the four genomes was carried out to identify the functional domains of the LTRs. Amino acid sequences mapped to protein domains with no less than 50% coverage were extracted for the Retrotrans_gag, RVT_1, Retrotran_gag_2 and RVT_2 domains of *P. polyphylla*, *V. dahuricum*, *Z. mays* and *G. biloba*, respectively. The extracted amino acid sequences were aligned using MAFFT (v.7.471, –thread -1 –auto –phylipout –reorder), and the global alignment results were then used to build phylogenetic trees with FastTree (v.2.1.11)[61], with visualization using FigTree (v.1.4.4).

Centromeric repeats

Centromeric regions are usually located in the middle of chromosomes and typically consist of satellite DNA and short stationary repeats arranged in tandem arrays[62,63]. We scanned the simple tandem repeats of *V. dahuricum* with copy numbers >100 and performed all-versus-all BLASTn for their consensus sequences (each tandem repeat has a consensus sequence). A 77-bp repeat unit ('GCATAGCGAGGAGTCACCGAAGTGGTCCAAGCGTAGTGTGTGGTG TGGCAACTTGTCGTTACCTGCAACTTGTCGTT') was found, which was homologous to most consensus sequences. Most consensus sequences were located in the middle of the chromosomes, that is, a presumed centromere region. We therefore inferred that this 77-bp repeat unit in *V. dahuricum* was centromeric satellite DNA. For *P. polyphylla*, a 79-bp repeat unit was detected ('TAAAGAAACCTAGTCTCGTAATAATAGAGAGAAAACGAAAGGGTC GCCGAGCTATAAATGCCACGGGTATACCTATTCG') and identified as potential centromeric satellite DNA. However, no sequence similarity was found between the centromeric repeat units of the two species. The density distribution of centromere units shows that they are mainly located in the putative centromere region (Supplementary Table 20).

Telomere repeats

Telomeres are repeat sequences located at the ends of linear chromosomes and range in length from 300 bp to many kilobases, usually consisting of guanine-rich repeats 6–8 bp in length. The *V. dahuricum* simple tandem repeats predicted by Tandem Repeats Finder were filtered, and those with a PeriodSize of 2–20 bp and copy number >100 were selected. By mapping the chromosomal locations of various consensuses, we found that only 'TTTAGGG' conformed to the telomere sequence signature. 'TTTAGGG' is a typical telomeric sequence in plants, also known as the *Arabidopsis*-type as it was first discovered in *Arabidopsis thaliana*[64]. We also discovered the same telomere sequence in *P. polyphylla*. The density distribution of telomere units shows that they are mainly located at the ends of chromosomes (Supplementary Table 21).

Fluorescent in situ hybridization (FISH)

Whole plants of *P. polyphylla* were obtained from the Key Laboratory of Medicinal Plant Biology of Yunnan Province, Yunnan Agricultural University, Kunming, China. Chromosome spread preparation was performed as previously described[65,66]. Briefly, root tips were collected into cold 0.05% colchicine solution, incubated at 4 °C for 6 h in the dark, fixed in Carnoy's fixative (ethanol:acetic acid at 3:1) at 4 °C for 20–24 h and washed in 70% ethanol. The root tips were then softened in an enzyme mixture (2.5% pectolyase and 2.5% cellulase) at 37 °C for 90–120 min, macerated on slides, steamed at 55 °C for 30–60 s and air dried for at least 20 min at 37 °C.

To verify the accuracy of the assembly, a FISH experiment targeting centromeres, telomeres and tandem repeats was performed following the protocols in refs. 66–68 with slight modifications. For probe selection (Supplementary Table 8), the 7-bp telomeric repeat (TTTAGGG) and the 79-bp centromeric repeat were repeated 30 times and 3 times, respectively, to generate ~200 bp probe sequences. The probe sequences for the 28S, 18S and 5.8S rDNA regions were selected as the 300-bp regions with the highest specificity across the genome from the full-length of rDNA sequences. Additional tandem repeats included a 323-bp repeat on chromosome PPY_2 (2–323), a 200-bp repeat on PPY_2 (2–200) and a 197-bp repeat on PPY_3 (3–197). All probe sequences were synthesized by Tsingke Biotechnology and then labelled by PCR using PCR Dig Labelling kit (Roche, 11636090910).

Slides were pretreated as previously described[65] in 2× SSC, 45% acetic acid and 4% formaldehyde successively. For hybridization, probes were mixed with hybridization buffer (50% formamide, 10% dextran sulfate, 0.25% SDS, 125 ng μl^{-1} salmon sperm DNA, 2× SSC), denatured at 95 °C for 6 min, applied to slides and covered with coverslips. The whole slides were denatured on a hot plate (80 °C) for 10 min and hybridized at 37 °C overnight. Post-hybridization washes included 10% formamide (in 2× SSC) and 0.2% Tween-20 (in 4× SSC). Anti-digoxigenin fluorescein (Roche, 11207741910) and the anti-digoxigenin rhodamine (Roche, 11207750910) were applied for probe detection. Slides were counterstained with 8 µl of 1 µg ml^{-1} DAPI (4′,6-diamidino-2-phenylindole) solution for 10 min in the dark, washed in 2× SSC for 2 min and sealed with antifade mounting medium (Beyotime).

For the 5S rDNA probe, a hybridization–elution–hybridization approach was used, involving repeated ethanol dehydration and denaturation in 70% formamide at 73 °C. Fluorescence signals were captured using a Nikon AX/AX R laser confocal microscope system and processed with ImageJ (1.52i) software.

WGD inference

WGD events were inferred from genomic synteny and peaks in K_s (synonymous substitutions per synonymous site) distribution. Intragenomic syntenic blocks were detected using i-ADHoRe[69] with the parameters 'alignment_method=gg2, number_of_threads=4, gap_size=35, cluster_gap=40, cloud_gap_size=40, cloud_cluster_gap=40, max_gaps_in_alignment=40, q_value = 0.75, prob_cutoff=0.01, anchor_points=3, level_2_only=false, multiple_hypothesis_correction=FDR'.

Using the 'ksd' module in wgd (v.1.2)[70], we calculated the K_s of both the paralogous gene pairs from the intraspecies syntenic blocks and the orthologous gene pairs between species constructed according to the reciprocal best hit protocol. The ksrates (v.1.1.1)[71] programme was used to detect WGD signatures in paralogue K_s distribution and to perform orthologue K_s adjustment according to branch length differences from the common ancestral nodes to terminal nodes representing current species. A phylogenetic tree in newick format of eight species was used as the input tree of ksrates programme: *P. polyphylla*, *V. dahuricum*, (*Oryza sativa*, *Elaeis guineensis*, *Asparagus officinalis*, *Dioscorea rotundata*, *Arabidopsis thaliana* and *Amborella trichopoda*); where we placed *P. polyphylla* and *V. dahuricum* sister to *O. sativa* and *E. guineensis* to match the phylogenies generated in this study (Fig. 2e, Supplementary Fig. 20). Both *P. polyphylla* and *V. dahuricum* were regarded as focal species for WGD detection and K_s adjustment, respectively.

5mC DNA methylation of CpG sites

As nanopore sequencing is sensitive enough to distinguish methylated (5mC) and unmethylated cytosine bases on CpG sites, we identified 5mC in the CpG sites with Nanopolish (v.0.13.2) and compared the methylation frequency in the *P. polyphylla* and *V. dahuricum* genomes.

Using an in-house script, the counted numbers of CpG sites in the *P. polyphylla* and *V. dahuricum* genomes were found to be 1,285,466,703 and 145,637,881, respectively. For each species, we first mapped the ONT reads to the genome and sorted the alignments according to the coordinates on the contigs/scaffolds. Next, the genome was split into 500 (*P. polyphylla*) or 50 parts (*V. dahuricum*). For each read, we extracted the corresponding alignment from 'sorted.bam' using the command 'samtools view -F0x900 sorted.bam contig1 contig2 contig3 … | cat header - | samtools view -@ 5 -Sb - > part.bam'. The 'part.bam' file was converted to a fasta-format file of ONT reads (reads.fa). We used the read IDs to extract the corresponding parts from the 'readdb' files in fast5 format as the input file for Nanopolish. The reads.fa was compressed using 'bgzip reads.fa && mv read.fa.gz read.index' and indexed using 'samtools faidx read.index'. Finally, for each part of the genome, we used 'nanopolish call-methylation -t 8 –reads=read –bam=part.bam –genome=part.fa –methylation=cpg > part.cpg.csv' to call the methylation of CpG sites, with the methylation counts then converted to methylation frequency using the Python script 'calculate_methylation_frequency.py', which is available in the Nanopolish (v.0.13.2) package.

After filtering sites with ONT depths <10, the number of sites for which the methylation frequency could be calculated was found to be

1,030,353,645 (80.15% of *P. polyphylla* CpGs) and 118,935,593 (81.67% of *V. dahuricum* CpGs). After methylation calling, we checked the methylation distribution of called sites and found that *P. polyphylla* had a higher percentage of methylated sites at high methylation levels (80%–99%), although the ratio of fully methylated sites in *P. polyphylla* (61.53%) was lower than that of *V. dahuricum* (63.59%). The average methylation level for the *P. polyphylla* genome was 94.24%, higher than that of *V. dahuricum* (90.72%).

Gene family expansion and contraction analysis and functional enrichment analysis

Using the protein sets of 11 species (Supplementary Table 11), 28,970 gene families were clustered by OrthoMCL; we reconstructed the evolutionary history of gene family expansion and contraction using CAFE (v.5)[31]. Compared with the common ancestor of *P. polyphylla* and *V. dahuricum*, we identified 1,793 and 2,577 expanded and contracted gene families, respectively, in *P. polyphylla*. We carried out KEGG functional enrichment analysis of the expanded gene families in *P. polyphylla* against all *P. polyphylla* genes (background gene set) using TBtools (v.2.121)[72] with the plant BackEnd files. We identified 131 enriched KEGG pathways (Supplementary Table 12) in the expanded gene families ($p < 0.05$).

Pfam domain annotation

We searched the whole genomes of *P. polyphylla* and *V. dahuricum* for protein domains using the Pfam (release 33.1) database. The genomes were first split into 248-kb fragments with a 1-kb overlap between two adjacent windows. The fragments were then translated into amino acid sequences with the 'transeq' command in EMBOSS (v.6.6.0)[73] in six phases. We next used HMMER (v.3.3.1)[74] to search for these amino acid sequences, then converted the search results into genome coordinates, removed redundancies and counted the number of different domains on the genomes.

Histones prediction

The curated alignment results of five types of histone (H1, H2A, H2B, H3 and H4) were downloaded from https://www.ncbi.nlm.nih.gov/research/histonedb/ and profile HMMs were contructed using hmmbuild (HMMER 3.3.1). Core and linker histone genes in *Arabidopsis*[75] were searched by the HMMs, and the main alignment regions were considered as the core regions of the five histones. Proteins of 11 species (Supplementary Table 11) were searched by the HMMs and filtered out hits covered >50% of core regions. The proteins corresponding to these matches were regarded as histones. We also downloaded histone sequences of *Oryza sativa* and *Arabidopsis thaliana* with subfamily information from HistoneDB 2.0 and used them as targets. We compared the predicted histones with them and searched for the best hit as the subfamily of the predicted histones.

Statistical analysis

Bootstrapping was used to test the phylogenetic tree nodes. Spearman's rank correlation was performed to calculate the correlation between the genome sizes and domain counts. Two-tailed Student's *t*-test was used to compare two samples. *P* values were used for KEGG enrichment to avoid high false rates across multiple tests.

Reporting summary

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

Data availability

The sequencing data for *Paris polyphylla* var. *yunnanensis* are available from NCBI Broproject PRJNA1130132 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1130132) and the China National GeneBank Database, under accession number CNP0002103, https://db.cngb.org/search/project/CNP0002103/ The *Veratrum dahuricum* data were deposited at the National Genomic Data Center, under accession number PRJCA005207, https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA005207. The de novo *Helitron* TEs were uploaded to the general repository Zenodo at https://doi.org/10.5281/zenodo.15401237 (ref. 77).

Code availability

The code and shell scripts used for chromosome assembly are described in GitHub at https://github.com/zengpeng2012/link_contigs_to_temporary_scaffolds (ref. 78).

References

1. Leitch, I. J. & Leitch, A. R. in *Plant Genome Diversity* Vol. 2 (eds Greilhuber, J. et al.) 307–322 (Springer, 2013).
2. Pellicer, J., Hidalgo, O., Dodsworth, S. & Leitch, I. J. Genome size diversity and its impact on the evolution of land plants. *Genes* **9**, 88 (2018).
3. Novák, P. et al. Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nat. Plants* **6**, 1325–1329 (2020).
4. Xu, S. et al. The evolutionary tale of lilies: giant genomes derived from transposon insertions and polyploidization. *Innovation* **5**, 100726 (2024).
5. Niu, S. et al. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **185**, 204–217 (2022).
6. Stevens, K. A. et al. Sequence of the sugar pine megagenome. *Genetics* **204**, 1613–1626 (2016).
7. Neale, D. B. et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* **15**, R59 (2014).
8. Pellicer, J., Kelly, L. J., Leitch, I. J., Zomlefer, W. B. & Fay, M. F. A universe of dwarfs and giants: genome size and chromosome evolution in the monocot family Melanthiaceae. *New Phytol.* **201**, 1484–1497 (2014).
9. Ji, Y. et al. Plastome phylogenomics, biogeography, and clade diversification of Paris (Melanthiaceae). *BMC Plant Biol.* **19**, 543 (2019).
10. Filion, W. G. & Vosa, C. G. Quinacrine fluorescence studies in *Paris polyphylla*. *Can. J. Genet. Cytol.* **22**, 417–420 (1980).
11. Wang, L.Li, Y.-F.Tang, R.-H. & Gu, Z.-J. Mapping of 8–26S rDNA loci in four species of the genus *Paris* by fluorescence in situ hybridization (FISH). *J. Syst. Evol.* **42**, 419–426 (2004).
12. Schartl, M. et al. The genomes of all lungfish inform on genome expansion and tetrapod evolution. *Nature* **634**, 96–103 (2024).
13. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics* **39**, btad595 (2023).
14. 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).
15. Tsukahara, S. et al. Centrophilic retrotransposon integration via CENH3 chromatin in *Arabidopsis*. *Nature* **637**, 744–748 (2025).
16. Gong, Z. et al. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell* **24**, 3559–3574 (2012).
17. Logsdon, G. A. et al. The variation and evolution of complete human centromeres. *Nature* **629**, 136–145 (2024).
18. Liu, H. et al. The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution. *Nat. Plants* **7**, 748–756 (2021).
19. Kubalová, I. et al. Helical coiling of metaphase chromatids. *Nucleic Acids Research* **51**, 2641–2654 (2023).
20. Hayashida, M. et al. Higher-order structure of human chromosomes observed by electron diffraction and electron tomography. *Microsc. Microanal.* **27**, 149–155 (2021).
21. Gibcus, J. H. et al. A pathway for mitotic chromosome formation. *Science* **359**, eaao6135 (2018).

22. Schloissnig, S. et al. The giant axolotl genome uncovers the evolution, scaling, and transcriptional control of complex gene loci. *Proc. Natl Acad. Sci. USA* **118**, e2017176118 (2021).
23. Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
24. Wendel, J. F., Jackson, S. A., Meyers, B. C. & Wing, R. A. Evolution of plant genome architecture. *Genome Biol.* **17**, 37 (2016).
25. Cai, L. et al. Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. *New Phytol.* **221**, 565–576 (2019).
26. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
27. Landis, J. B. et al. Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* **105**, 348–363 (2018).
28. Thomas, J. & Pritham, E. J. Helitrons, the eukaryotic rolling-circle transposable elements. *Microbiol. Spectr.* **3**, 893–926 (2015).
29. Kelly, L. J. et al. Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytol.* **208**, 596–607 (2015).
30. Hardy, P.-O. & Chaconas, G. The nucleotide excision repair system of *Borrelia burgdorferi* is the sole pathway involved in repair of DNA damage by UV light. *J. Bacteriol.* **195**, 2220–2231 (2013).
31. 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).
32. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res.* **53**, D672–D677 (2025).
33. He, X.-J., Chen, T. & Zhu, J.-K. Regulation and function of DNA methylation in plants and animals. *Cell Res.* **21**, 442–465 (2011).
34. Gent, J. I. et al. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res.* **23**, 628–637 (2013).
35. Schnable, P. S. et al. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
36. Hershberg, R. & Petrov, D. A. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* **6**, e1001115 (2010).
37. Long, H. et al. Evolutionary determinants of genome-wide nucleotide composition. *Nat. Ecol. Evol.* **2**, 237–240 (2018).
38. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).
39. Mistry, J. et al. Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
40. Draizen, E. J. et al. HistoneDB 2.0: a histone database with variants—an integrated resource to explore histones and their variants. *Database* **2016**, baw014 (2016).
41. Lei, B. & Berger, F. H2A variants in *Arabidopsis*: versatile regulators of genome activity. *Plant Commun.* **1**, 100015 (2020).
42. Kawashima, T. et al. Diversification of histone H2A variants during plant evolution. *Trends Plant Sci.* **20**, 419–425 (2015).
43. Bönisch, C. & Hake, S. B. Histone H2A variants in nucleosomes and chromatin: more or less stable? *Nucleic Acids Res.* **40**, 10719–10741 (2012).
44. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
45. 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).
46. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
47. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
48. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
49. 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 *bioRxiv* https://doi.org/10.1101/254797 (2018).
50. Zeng, X. et al. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat. Plants* **10**, 1184–1200 (2024).
51. Coombe, L. et al. LongStitch: high-quality genome assembly correction and scaffolding using long reads. *BMC Bioinformatics* **22**, 534 (2021).
52. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
53. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
54. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
55. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
56. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
57. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
58. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of *Helitron* transposons in many plant genomes. *Proc. Natl Acad. Sci. USA* **111**, 10263–10268 (2014).
59. Zhang, R.-G. et al. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic. Res.* **9**, uhac017 (2022).
60. Guan, R. et al. Draft genome of the living fossil *Ginkgo biloba*. *GigaScience* **5**, 49 (2016).
61. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
62. Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001).
63. Hartley, G. & O'Neill, R. J. Centromere repeats: hidden gems of the genome. *Genes* **10**, 223 (2019).
64. Richards, E. J. & Ausubel, F. M. Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. *Cell* **53**, 127–136 (1988).
65. Aliyeva-Schnorr, L., Ma, L. & Houben, A. A fast air-dry dropping chromosome preparation method suitable for FISH in plants. *J. Vis. Exp.* **16**, e53470 (2015).
66. Bolaños-Villegas, P., Yang, X., Makaroff, C. A. & Jauh, G.-Y. Protocol for the preparation of *Arabidopsis* meiotic chromosome spreads and fluorescent in situ hybridization. *Bio Protoc.* **4**, e1102 (2014).
67. Naish, M. et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* **374**, eabi7489 (2021).
68. Kubalová, I. et al. Helical coiling of metaphase chromatids. *Nucleic Acids Res.* **51**, 2641–2654 (2023).
69. Proost, S. et al. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, e11 (2012).
70. Zwaenepoel, A. & Van de Peer, Y. wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**, 2153–2155 (2019).

71. Sensalari, C., Maere, S. & Lohaus, R. ksrates: positioning whole-genome duplications relative to speciation events in KS distributions. *Bioinformatics* **38**, 530–532 (2022).
72. Chen, C. et al. TBtools-II: a "one for all, all for one" bioinformatics platform for biological big-data mining. *Mol. Plant* **16**, 1733–1742 (2023).
73. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).
74. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
75. Probst, A. V., Desvoyes, B. & Gutierrez, C. Similar yet critically different: the distribution, dynamics and function of histone variants. *J. Exp. Bot.* **71**, 5191–5204 (2020).
76. Tang, H. et al. JCVI: a versatile toolkit for comparative genomics analysis. *iMeta* **3**, e211 (2024).
77. Zeng, P. Two Melanthiaceae genomes with dramatic size difference provide insights into giant genome evolution and maintenance (V1.0) [data set]. *Zenodo* https://doi.org/10.5281/zenodo.15401237 (2025).
78. GitHub Link short contigs to longer temporary scaffolds. *Zenodo* https://doi.org/10.5281/zenodo.15872340 (2025).

Acknowledgements

We thank W. Wang and K. Wang of Northwestern Polytechnical University (NWPU) and G. Zhang of Zhejiang University for constructive suggestions during paper preparation; X. Yang and Y. Jia of Xi'an Jiaotong University for help in genome assembly and chromosome compaction; L. Zhang of the Chinese Institute of Brain Science in Beijing for support in computation resources; and Z. Li of VIB-UGent Center for Plant Systems Biology for suggestions on WGD analysis. This work was supported by the National Key R&D Program of China (2022YFC3400300), the Thousand Talents Plan (5113190037), Talents Team Construction Fund of NWPU, Open Research Project of the Cross-Cooperative Team of the Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, and Fundamental Research Funds for the Central Universities (3102019JC007) to J.C., and the Major Science and Technique Programs in Yunnan Province (2019ZF011-1) and Science and Technology Innovation Team of Yunnan (202105AE160011) to S.Y.

Author contributions

J.C., H.J. and S.Y. conceived and led the study; P.Z. managed the project; P.Z. and J.C. wrote the paper; P.Z., G.Z. and Z.T. collected and sequenced the plant material; P.Z. assembled and annotated the genomes; Y.Z. and L.L. performed FISH experiments; P.Z. conducted gene family clustering and comparative genomics. P.Z., H.Z., B.Z., Z.T. and T.Z. performed chromosome construction; P.Z. and Z.T. conducted WGD analysis; P.Z. and T.Z. conducted methylation analysis; P.Z., Y.H., Z.T., W. Zhang, T.Z. and Y.Y. performed gene family analysis; W. Zhang, X.L., J.H., Q.L., Y.P., S.H. and W. Zhu, conducted protein function verification. All authors read and approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

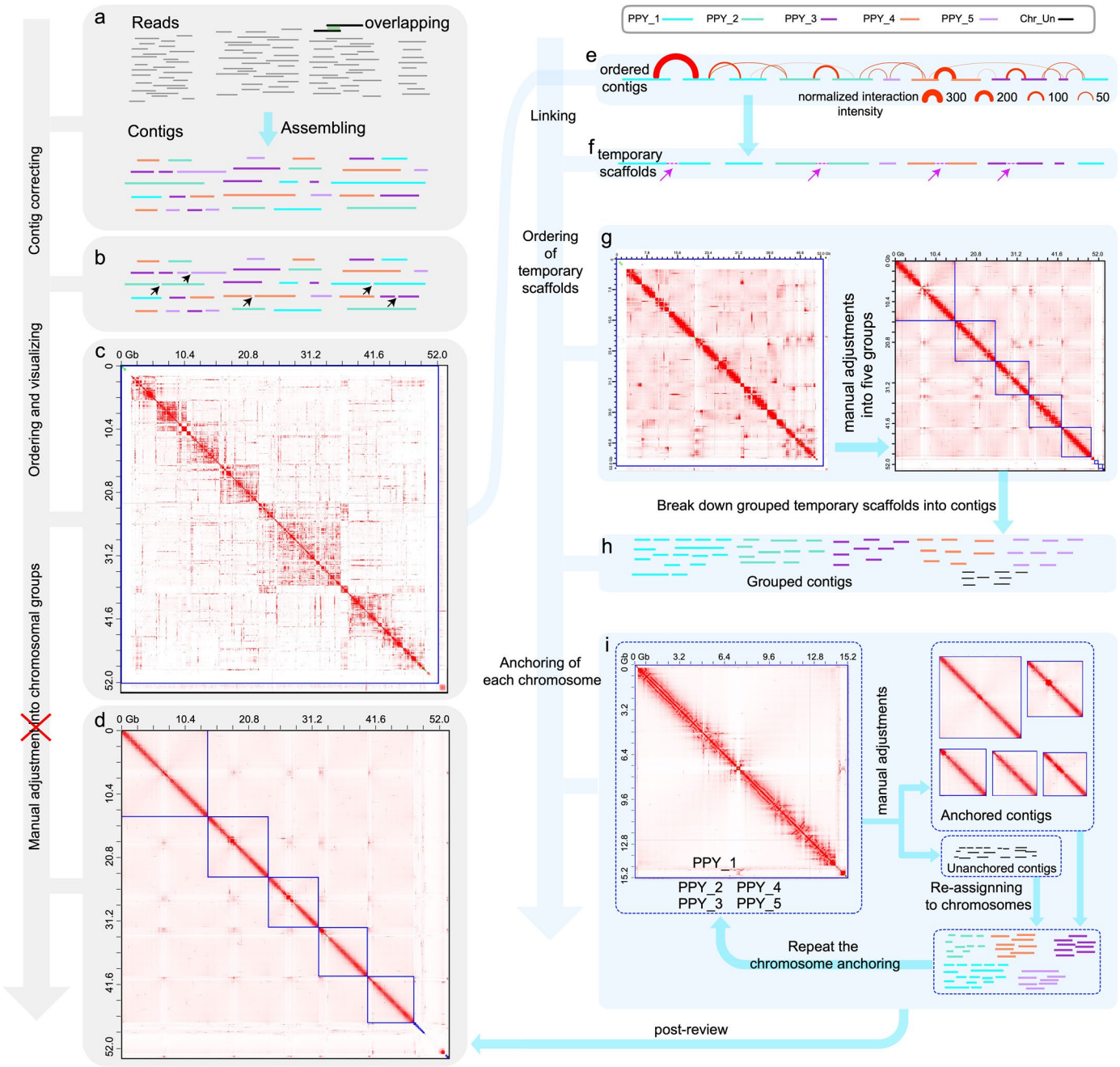
Extended data is available for this paper at https://doi.org/10.1038/s41477-025-02060-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-025-02060-3.

Correspondence and requests for materials should be addressed to Huifeng Jiang, Shengchao Yang or Jing Cai.

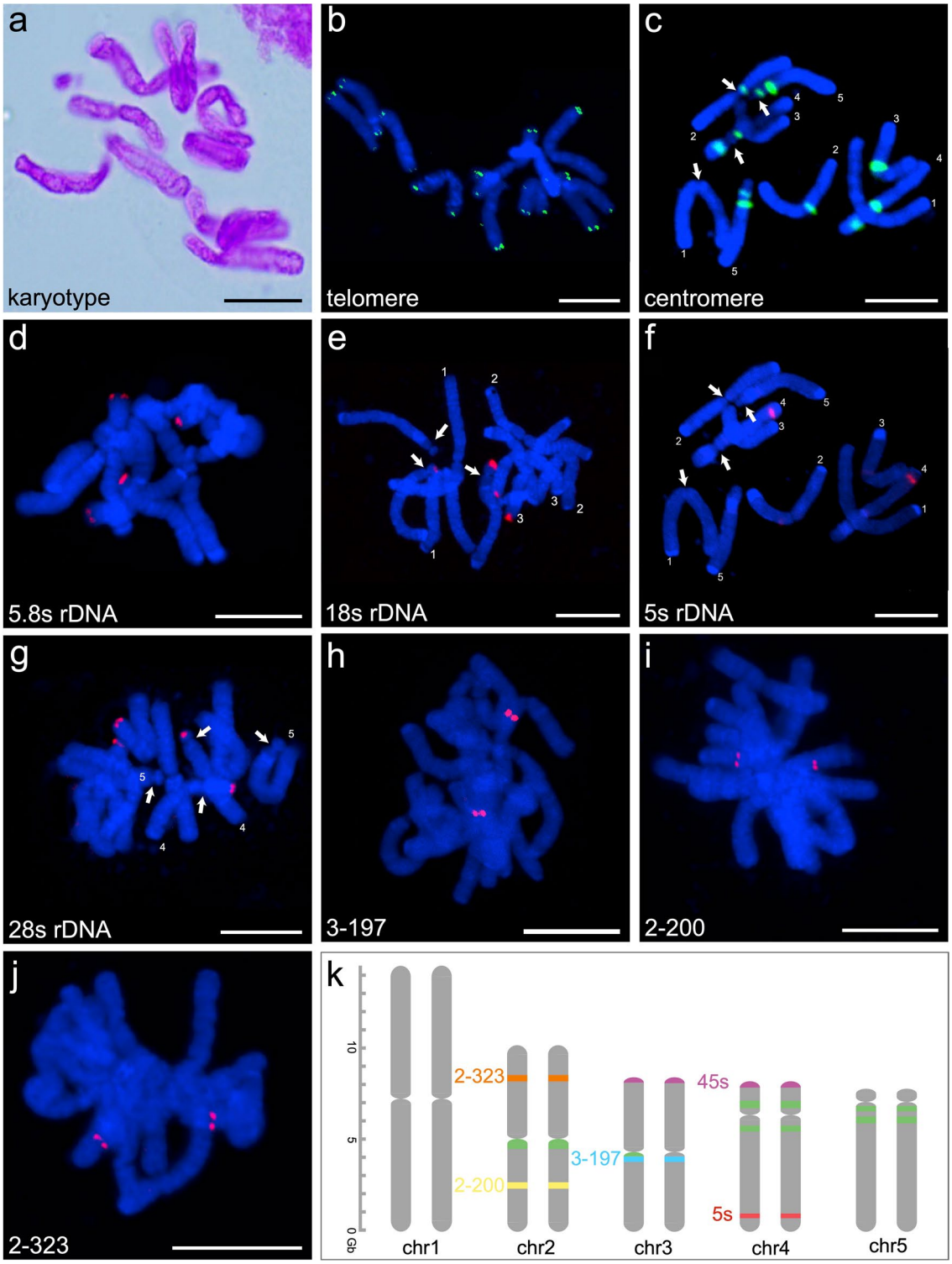
Peer review information *Nature Plants* thanks Robert Van Buren, D. Blaine Marchant, Jean-Marc Aury and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41477-025-02060-3



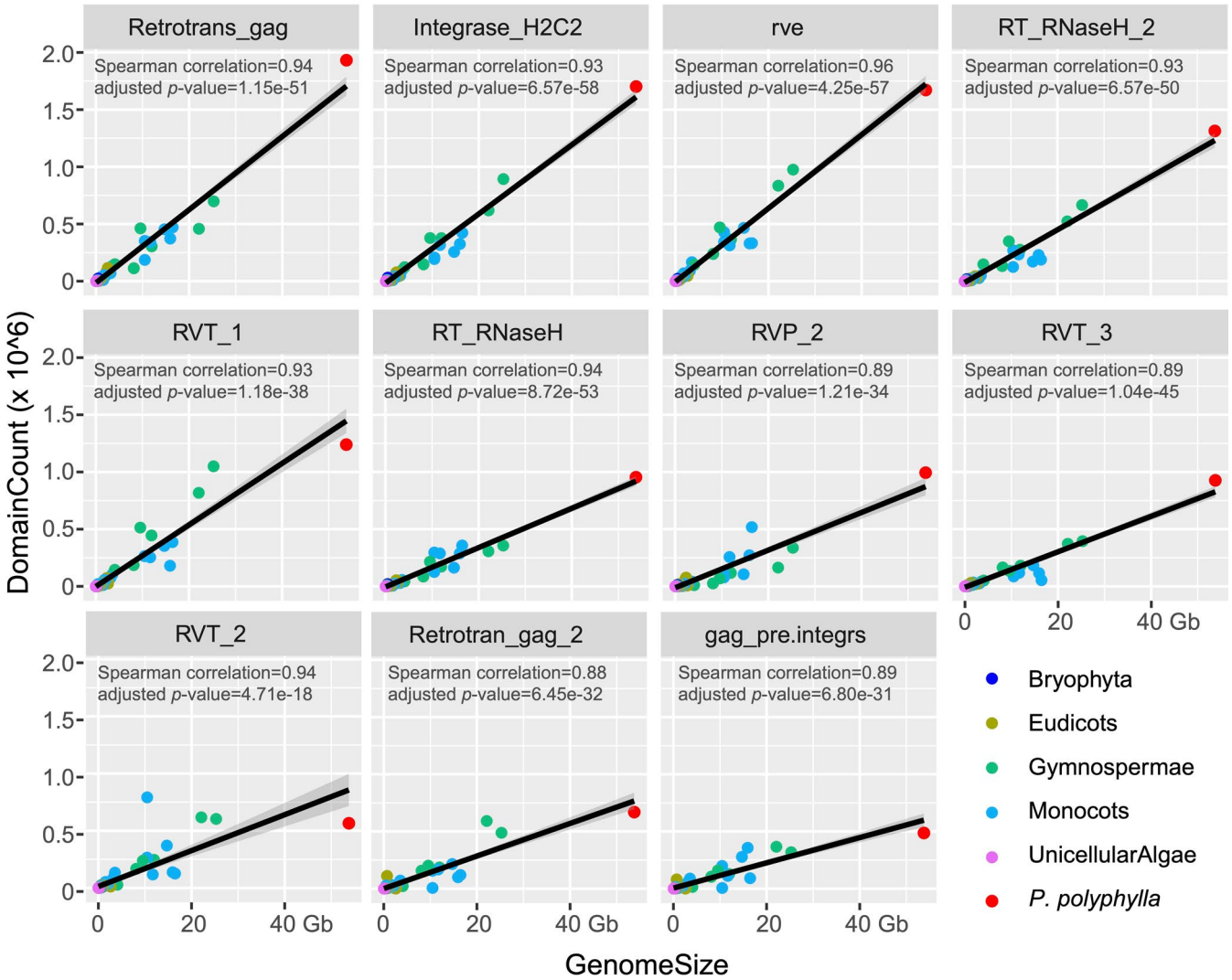
Extended Data Fig. 1 | Hierarchical bottom-up chromosome assembly strategy for huge genome of *P. polyphylla*. (**a**) *De novo* assembly of HiFi reads and coloured solid lines represent contigs from different chromosomes. (**b**) Contigs correcting by YaHS, a Hi-C scaffolding tool. Black arrows point to the contigs cut by YaHS. (**c**) Initial Hi-C interaction heatmap of contigs generated by 3D-DNA. For most species, a general pipeline of a->b->c->d will generate acceptable chromosome-level assemblies. The giant 52.75 Gb/1 C genome comprising 78,381 contigs made it almost impossible to manually manipulate these contigs into final chromosomes (**d**). (**e**) Ordered contigs in **c** were linked into temporary scaffolds (**f**) where adjacent contigs have a normalized interaction intensity >=100. Purple arrows point to the linked contigs. (**g**) Ordering and visualizing for all temporary scaffolds. Manual adjustments were conducted to group temporary scaffolds into five ultra-long chromosomes according to the interaction data between scaffolds. (**h**) Foreach chromosome group, all temporary scaffolds were broken into contigs. (**i**) Two iterative cycles of anchoring of contigs for each chromosome. Anchoring and visualizing was performed on grouped contigs, with manual adjustments for each chromosome. Un-anchored contigs were reassigned to the chromosomes where the strongest interacting anchored contigs were located. Anchoring of contigs was repeated using anchored contigs and reassigned un-anchored contigs. (**d**) Final whole-genome Hi-C interaction heatmap were generated by pooling the five chromosomes and un-anchored contigs and run post-review function of 3D-DNA.

https://doi.org/10.1038/s41477-025-02060-3



Extended Data Fig. 2 | Karyotype and FISH in metaphase cells of *P. polyphylla*. (**a**) Chromosomes of metaphase root tip cells stained with modified carbol fuchsin dye. (**b**–**j**) FISH images with different types of probes, with the probe name indicated in the lower left corner of each image. All individual numbers in the figure represent the chromosome identifiers. The arrows indicate the chromosome primary constrictions. The scale bar is 20 μm. FISH with 45 s rDNA probes including three probes for 5.8 s rDNA (**d**), 18 s rDNA(**e**) and 28 s rDNA(**g**). All the three probes showed four signals at one end of four chromosomes, among which two are at the short arm of a pair of acrocentric chromosomes. (**f**) FISH with 5 s rDNA probe was performed on the same slide after the washing away the centromere probes in **c**. (**h**) FISH with a chromosome3-specific repeat probe 3-197. (**i**, **j**) FISH with two chromosome2-specific repeat probes: 2-200 and 2-323. (**k**) The schematic diagram of *P. polyphylla* chromosomes with all probe signals except the telomere probes. Green bands represent the centromere probes while other bands were labeled with probe name in same color. All FISH experiments were independently repeated at least three times with consistent results.

https://doi.org/10.1038/s41477-025-02060-3



Extended Data Fig. 3 | Correlations between genome size and TE domains. Dot plot of genome size and abundance of 11 most relevant protein domains in genomes of 78 species (including four algae, two bryophytes, nine gymnosperms, 32 eudicots, and 31 monocots). Each dot represents a species. All 11 domains were from transposable elements (TEs). Associations between genome size and domain count were tested using cor.test(GenomeSize, DomainCount, conf.level = 0.95, methods = "spearman"), all analysed correlations are significantly positive (p-values < 0.001). A linear model estimated trend line and calculated 95% confidence interval around the trend (grey fill) are plotted (two-sided). Retrotrans_gag: Retrotransposon gag protein of LTR gypsy-type, Integrase_H2C2: Integrase zinc binding domain, rve: Integrase core domain, RT_RNaseH_2: RNase H-like domain found in reverse transcriptase, RVT_1: Reverse transcriptase (RNA-dependent DNA polymerase) 1, RT_RNaseH: Rnase H-like domain found in reverse transcriptase, RVP_2: Retroviral aspartyl protease, RVT_3: Reverse transcriptase-like, RVT_2: Reverse transcriptase (RNA-dependent DNA polymerase) 2, Retrotran_gag_2: gag-polypeptide of LTR copia-type, gag_pre.integrs: GAG-pre-integrase domain.

nature portfolio

Corresponding author(s): Jing Cai

Last updated by author(s): May 16, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software had been used for data collection.
Data analysis	The following software/tools/algorithms/packages were used: 3D-DNA (v180922), ASTRAL (v5.7.3), Augustus (v3.3.3), BLAST (2.7.1), blat (v35), BUSCO (v3.0.2), compleasm (v0.2.6), CAFE5, EDTA(v1.9.4), EVidenceModeler (v1.1.1), FastTree (2.1.11), GeneWise (2.4.1), GlimmerHMM (v3.0.4), GMAP (version 2018-07-04), Hifiasm (v0.19.8), HMMER (3.3.1), i-ADHoRe (3.0), ImageJ (1.52i), InterProScan (v5.45-80.0), Jellyfish (v2.2.10), bwa (0.7.17), JuiceBox (v1.11.08), Juicer (v1.6), ksrates (v1.1.1), mafft (v7.471), MCScanX, minimap2 (2.17-r974), MUSCLE-v3.8.31, Nanopolish-v0.13.2, NextDenovo (v2.2-beta.0), NextPolish (v1.1.0), OrthoFinder(2.4.0), OrthoMCL(1.4), PAML package (v.4.9i), RAxML-NG (v.1.0.0-master), RepeatMasker (version 4.1.0), Solar (0.9.6), Sonicparanoid (1.3.5), TBtools (v2.121), Trinity (v.2.8.4), stringtie (2.2.1), hisat2(2.1.0), wgd (v1.2), Pfam (v33.1), R, Python, Perl. Specific parameters used during run-time are provided in the methods. All software or scripts are available from official websites or GitHub as indicated in the methods.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The sequencing data for Paris polyphylla var. yunnanensis are available from NCBI Broproject PRJNA1130132 and the China National GeneBank Database (CNGBdb, https://db.cngb.org/) under accession number CNP0002103. The Veratrum dahuricum data were deposited at the National Genomic Data Center (https://bigd.big.ac.cn/bioproject/) under accession number PRJCA005207. The de novo Helitrons TEs were uploaded to the general repository Zenodo (https://doi.org/10.5281/zenodo.15401237).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No samples size is calculated. The sample size of both species in our study was sufficient for de novo genome sequencing.
Data exclusions	Low quality data are excluded for genome assembly.
Replication	No sample was re-sequenced, because it is a de novo genome sequencing project.
Randomization	Randomization is not required for genome sequencing by the state of the art of genomics.
Blinding	Blinding was not necessary for this study as we used all data for the genome assembly and analyses.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Nuclei were isolated from young leaves
Instrument	BD FACScalibur
Software	CXP v2.2
Cell population abundance	abundance >8000 cells
Gating strategy	Total nuclei populations were gated using PI intensity (Fig. S3).

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.