

The chromosome-level genome assembly of *Dryopteris fragrans* reveals transposon-mediated genome evolution and adaptation

ying chang

changying@neau.edu.cn

College of Life Sciences, Northeast Agricultural University

weicong DAI

College of Life Sciences, Northeast Agricultural University

Qia Wang

Kunming Institute of Botany, Chinese Academy of Sciences

Yuhan Fang

South China Botanical Garden

xiaojie qiu

College of Life Sciences, Northeast Agricultural University

chunhua song

College of Life Sciences and Technology, Harbin Normal University

dongrui zhang

College of Life Sciences, Northeast Agricultural University

shucai wang

Laboratory of Plant Molecular Genetics & Crop Gene Editing, School of Life Sciences, Linyi University

Article

Keywords:

Posted Date: March 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-7268223/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Abstract

Ferns are ancient vascular plants pivotal to plant evolution research. Although sequencing technologies have advanced fern genomic studies, scarce genomic resources for specialist-habitat ferns limit insights into their genome evolution. *Dryopteris fragrans* (L.) Schott is a fern endemic to sun-exposed volcanic-lava habitats; here, we generated its high-quality chromosome-level genome assembly, and explored the drivers of its genomic evolution and habitat adaptation via whole-genome duplication (WGD) detection, gene family evolution analysis and other approaches. No recent WGD event was detected in *D. fragrans*, while transposable elements (TEs)—the major genomic component, associated with the expansion of environment-adaptive gene families—were identified as the primary evolutionary driver. Specifically, TEs shape gene structure by forming clade-specific long-intron genes, regulate gene expression through promoter insertion, and increase alternative splicing events in host genes. This study reports the first high-quality genome of a volcanic-lava-adapted fern, revealing TEs as potential key drivers of *D. fragrans*' genomic evolution and habitat adaptation. Our findings advance understanding of TE functions in non-seed plant evolution, and provide valuable genomic resources for researching early land plant adaptation and regulatory innovation.

Introduction

Ferns represent one of the oldest surviving lineages of vascular plants and are valuable model groups for investigating plant evolution^{1–3}. With approximately 10,600 extant species, ferns constitute the second largest lineage (after angiosperms) among vascular plants⁴. Ferns have notoriously immense genomes (average 1C = 12.3 billion bases (Gb); maximum 1C = 160.45 Gb) and very high chromosome numbers (average, 40.5; maximum, 720)^{5,6}. For a long time, this feature of ferns was thought to be associated with whole-genome duplication (WGD) events^{7,8}. However, in published fern genomes, WGD events are less common in ferns than in seed plants^{9–13}. On the other hand, in all published fern genomes, repetitive sequences were found to be the predominant component of the genome, with transposable element (TE) being the major contributors^{9–13}. Consistent with recent comprehensive analyses of WGD in leptosporangiate ferns, these observations suggest that in the absence of frequent WGD events, fern genome evolution may have relied on small-scale duplications¹⁴, wherein TE—the dominant genomic component—likely play a pivotal role.

TE act as “jumping genes” in the genome, shaping plant genomic diversity and playing dual roles in genomic structural variation and functional innovation^{15–17}. TEs are divided into two major classes: Class I retrotransposons that mobilize via an RNA intermediate and Class II DNA transposons that move through a cut-and-paste mechanism¹⁸. In seed plants, the insertion of TEs has been shown to be closely associated with genome expansion^{19,20}, gene family evolution^{21,22} and environmental adaptation²³. Moreover, environmental stresses such as ultraviolet stress, drought, or pathogen attack can trigger bursts of TE activity, leading to adaptive genetic variation²⁴. TE insertions can regulate gene expression by altering gene structure or via insertional mutagenesis, supplying raw material for evolutionary

innovation²⁵. While in seed plants, the ability of TE to promote genome evolution and improve plant environmental adaptability has been supported by numerous studies, research on TEs in ferns remains relatively scarce.

The current availability of high-quality fern genomes remains limited to only a few lineages: three species from Pteridaceae^{9,10,13}, two from Salviniaceae¹¹, one from Cyatheaceae¹², four from Cibotiaceae^{26,27}, and one from Lygodiaceae²⁸. Although these fern genomic resources have enhanced our understanding of this group to a certain extent, existing studies have mainly focused on ferns inhabiting humid and/or shaded habitats. Genomic research on ferns from specific habitats is relatively scarce, which may contain new perspectives for understanding the genomic evolution of this group. Take *D. fragrans* as an example, it is a fern species growing in direct sunlight on volcanic lava habitats in the northern temperate zone, and belongs to the basal lineage of Dryopteridaceae, the most species-rich family of ferns^{29,30}. Currently, research on *D. fragrans* has mainly focused on terpenoid metabolism and the medicinal value of its active components, with few studies on its genomic evolution and environmental adaptability³¹.

In this study, we present a chromosome-level genome assembly of *D. fragrans*. Annotation analyses revealed that its genome harbors significantly higher TE content than other reported fern genomes, but with lower TE integrity. We explored the evolution of gene families potentially associated with adaptation to specialized habitats and the potential relationship between gene expansion and TEs. Additionally, we investigated the potential impacts of TEs on gene structure and expression regulation. Similar TE-mediated genome evolution has been reported in seed plants^{15,32–34}; our study extends this concept to ferns, laying a foundation for future comparative studies in non-seed plants. These findings suggest that TE-driven genomic dynamics may be a universal feature of land plant evolution, rather than being unique to seed plants.

Results

Chromosome-level genome assembly and annotation

Ferns exhibit an alternation of generations between diploid sporophytes and haploid gametophytes. To obtain a high-quality haplotype-resolved genome of *D. fragrans*, spores were cultured into gametophytes, which were then subjected to propagation and culture. The resulting callus was used for DNA extraction, library construction, and sequencing. K-mer analysis estimated the genome size to be approximately 4.48 Gb (Fig. S1a). The new genome was generated using a combination of Oxford Nanopore Technologies (ONT) long reads (average sequencing depth of 58×; Table S1), Illumina short reads (coverage depth of 52×; Table S2), and high-throughput chromosome conformation capture data (Hi-C; average coverage depth of 111×). The final assembly comprised 41 chromosomes with a total size of 4.45 Gb and a contig N50 of 5.70 Mb (Fig. 1a; Table 1; Fig. S1b; Table S3). Furthermore, we identified

telomeric sequences at both ends of 16 chromosomes and at one end of 14 chromosomes (Fig. S1c; Table S4).

We annotated the *D. fragrans* genome using a combined strategy (ab initio, homology-based and transcriptome-based predictions; Tables S5–S8). In total, 33,489 high-confidence gene models were annotated, with a mean gene length of 36,768 bp (Table 1); 75.11% of genes were supported by transcript evidence (Table S8). Benchmarking Universal Single Copy Orthologs (BUSCO) analysis recovered 1,388 complete BUSCOs (86.0%), including 1,296 single-copy (80.3%) and 92 duplicated (5.7%) complete BUSCOs (Fig. 1b). Additional Core Eukaryotic Genes Mapping Approach (CEGMA) assessment also supported the high reliability of our gene models (Fig. S2). A total of 22,879 protein-coding genes (68.32% of all genes) were assigned functional annotations in public databases (Fig. S3; Table. S9). Together, these results indicate that the gene-structure annotation of the *D. fragrans* genome is of sufficient quality to support downstream analyses.

In the *D. fragrans* genome, we annotated 3.77 Gb of repetitive sequences, accounting for 85.39% of the assembly (Table S10). Long terminal repeat (LTR) retrotransposons are the dominant component: LTR/Gypsy and LTR/Copia elements constitute 54.35% and 17.75% of the genome, respectively. We also detected asymmetric repeat accumulation at chromosome ends for 20 chromosomes (Fig. 1c; Fig. S4). Notably, 10 of these chromosomes contain telomeric sequences at both ends. This pattern suggests that the two chromosomal arms within the same chromosome may differ in terms of repeat content. The underlying mechanisms, including potential differences in insertion or removal dynamics, recombination or DNA repair activity, and selective constraints, require further investigation.

a. Circos plot of the *D. fragrans* genome, divided into 10 Mb windows, showing (1) chromosome length and number; (2) content of repetitive sequences; (3) content of Solo-LTRs; (4) content of Gypsy; (5) content of Copia; (6) gene density; and (7) GC content, with the innermost circle representing synteny. **b.** BUSCO assessment of genome completeness across seven fern species. **c.** Distribution of repeat content along chromosomes visualized via bar plots and heatmaps. Each chromosome is divided into 50 windows (x-axis), and the repeat density within each window is shown on the y-axis and by color intensity.

Recent WGD has not been detected in *D. fragrans*.

WGD is a widespread and important process in plant genome evolution, providing raw gene material for functional diversification and thereby accelerating evolutionary innovation. To investigate the WGD history of *D. fragrans*, we first surveyed genome-wide synteny. We identified 75 syntenic blocks containing 399 paralogous gene pairs, representing only 2.38% of the gene set, a proportion too low to support a recent WGD (Fig. 2a). The synonymous substitution rate (K_s) distributions of both syntenic pairs and all homologous gene pairs showed a single peak at $K_s \approx 1.2$, which was consistent with one ancient duplication event rather than recent polyploidy (Fig. 2b; Fig. S5). Using the multi-taxon polyploidy search (MAPS) framework with genome and transcriptome data from *D. fragrans* and other ferns, we detected only a single WGD shared broadly among leptosporangiate ferns, likely dating to the Permian–

Triassic interval (Fig. 2c). In summary, we found no evidence of recent WGD in *D. fragrans*, and its genomic evolution may be driven by other factors.

a. Intragenomic synteny analysis of *D. fragrans*, illustrating the distribution and retention of homologous gene blocks. **b.** Ks distribution for non-tandem duplicated homolog pairs. **c.** MAPS-based detection of WGD signals: the phylogenetic tree of selected fern species is shown below, with nodes N2–N6 indicated; above, the proportion of duplicated genes in each subtree is plotted, where the green, red, and blue curves represent the null simulation, positive simulation, and observed data, respectively. The shaded regions denote confidence intervals.

TE-related dispersed duplicates (DSD) are the main drivers of gene family expansion in *D. fragrans*

To explore whether the evolution of specific gene families contributes to the genomic evolution of *D. fragrans*, protein-coding genes from 12 species were clustered into 26,136 orthogroups, of which 31,115 *D. fragrans* genes (93.8% of the gene set) were assigned. A species tree inferred from 253 single-copy orthogroups and molecular dating data indicates that *D. fragrans* diverged from the lineage shared with water ferns and maidenhair ferns at ~ 168.6 Ma (Middle Jurassic; Fig. 2a). To investigate gene-family dynamics potentially linked to adaptation, we conducted gene ontology (GO) enrichment on 761 *D. fragrans*-specific orthogroups (4,615 genes) and 2,262 significantly expanded orthogroups (8,774 genes). Both lineage-specific and expanded families were enriched for functions related to redox metabolism, terpene biosynthesis, stress perception/defense, and protein homeostasis (Fig. 3b, c). Compared with the other genes, the genes within the *D. fragrans*-specific and expanded orthogroups presented stronger transcriptional responses to ultraviolet (UV) treatment (Fig. S6)³⁵. Taken together, these results suggest that the evolution and expansion of these gene families may have contributed to the adaptation of *D. fragrans* to its specific habitat.

To further explore the main modes of gene family expansion in *D. fragrans*, we identified 28,193 duplicated genes and classified them into five categories: 486 whole-genome duplicates (WGD; 1.7%), 2,456 tandem duplicates (TD; 8.7%), 2,165 proximal duplicates (PD; 7.7%), 980 transposed duplicates (TRD; 3.5%), and 21,747 dispersed duplicates (DSD; 77.1%). Compared with the other classes, the PD and TD pairs presented lower Ks values and relatively higher nonsynonymous/synonymous substitution rate (Ka/Ks) values (Fig. 3d; Fig. S7), indicating a more recent origin and faster sequence divergence, which is consistent with ongoing tandem and proximal duplications. We then compared the overlap between expanded orthogroups and genes derived from the five duplication modes (WGD, TD, PD, TRD, DSD). The majority of genes within expanded orthogroups are derived from DSD (Fig. 3e). In addition, by assessing genome-wide correlations between TE abundance and gene counts of different duplication classes, we observed that DSD gene counts were positively correlated with TEs (Pearson's R = 0.79), a correlation stronger than that between TRD gene counts and TEs (Pearson's R = 0.75; Fig. S8). Taken together, the expanded gene families potentially associated with the adaptation of *D. fragrans* to its specific habitat are mainly derived from DSD, and DSD is closely related to TEs, suggesting that TEs may be the main driver of genomic evolution in *D. fragrans*.

a. Species tree with divergence time estimates and numbers of expanded/contracted gene families. The tree was inferred from 253 single-copy orthogroups; divergence times are indicated in Mya. **b.** GO enrichment of *D. fragrans*-specific orthogroups. **c.** GO enrichment of orthogroups significantly enriched in *D. fragrans*. **d.** Distributions of Ka/Ks ratios for genes originating from five duplication modes (WGD, TD, PD, TRD, DSD). **e.** Overlap analysis between genes in expanded orthogroups and genes derived from each duplication mode (WGD, TD, PD, TRD, DSD). Among them, WGDp, TDp, PDp, TRDp, and DSDp refer to positively selected genes. **f.** GO enrichment for genes originating from WGD pairs.

Low Integrity and continuous insertion of TE in the *D. fragrans* genome.

A comparison with published fern genomes revealed that both the relative proportion and the absolute base count of TEs are elevated in *D. fragrans* (Fig. 4a). However, the TE complement in this genome is characterized not by the preservation of numerous intact elements but by pervasive fragmentation and rapid turnover. Several lines of evidence support this interpretation. First, the LTR assembly index (LAI) was low (LAI = 9.73), indicating reduced overall integrity of LTR sequences relative to those of other ferns (Fig. 4b). Second, intact LTR retrotransposons represent only ~ 0.58% of the assembly, which is consistent with widespread truncation or degradation. Third, analyses of Kimura substitution spectra and pairwise sequence identity revealed a pattern of continuous, frequent LTR integration rather than a single, recent burst (Fig. 4c-f); in contrast, several non-LTR TE classes presented signatures of episodic amplification. Notably, we observed peaks at a Kimura substitution level of ~ 35 and at a sequence identity of ~ 0.75, which may correspond to the ancient leptosporangiate WGD signal.

Subfamily level classification of ~ 46,130 intact LTRs further reveals heterogeneous dynamics: Copia elements are dominated by the Ale, Tork and Ivana subfamilies, whereas Gypsy elements are primarily Athila, Reina and Tekay (Fig. 4d; Fig. S9a). When a neutral substitution rate ($\mu = 1.3 \times 10^{-8}$ substitution site⁻¹ year⁻¹) was used for insertion-time estimation, the Ivana subfamily presented evidence of relatively recent expansion, whereas most other LTR lineages accumulated more gradually (Fig. S9b,c). Taken together, these observations are most consistent with a high-turnover regime in which ongoing insertion of LTRs is rapidly followed by processes that fragment or remove copies.

a. Comparison of LTR content and genome size between *D. fragrans* and several other plant species. **b.** LAI index of the *D. fragrans* genome, with regions > 10 shown in blue and regions < 10 shown in red. **c.** Kimura substitution levels of *D. fragrans* TEs. **d.** Phylogenetic trees of the six major LTR/Copia (left) and LTR/Gypsy (right) families. **e.** Identity distribution of TEs within the genome compared with a custom-built TE database. **f.** Identity distribution of LTRs within the genome compared with a custom-built TE database.

TE insertion shapes gene structure in *D. fragrans*.

TE act as a "double-edged sword" in genomes: on the one hand, their insertion into genic regions can disrupt gene structure and impair function; on the other hand, they provide genetic variation resources for evolution. Since the most direct effect of TE insertion into genic regions is altering gene length, we

first investigated the associations between TEs and gene structure in *D. fragrans* through cross-species comparisons. The results revealed that among the fern lineages, with the exception of two aquatic ferns with small genomes, the other four ferns presented a broader range of gene lengths (Fig. 5a). Further analysis of coding sequence (CDS) and intron regions revealed that CDS length was relatively conserved across all plant groups, whereas intron length varied significantly—with a subset of genes harboring exceptionally long introns (> 10 kb) in the four ferns. Notably, a strong positive correlation was observed between gene length and intron length in *D. fragrans* ($R = 0.99$; Fig. S10), indicating that gene length is determined primarily by intron length.

Further investigations revealed that TE insertions are prevalent within the introns of *D. fragrans* genes. Moreover, compared with TE in other genomic regions, these TE insertions located in gene regions exhibit higher stability (Fig. S11). All long-intron genes contained TE insertions (Fig. 5b), and a positive correlation was detected between the proportion of TE-derived sequences in introns and intron length ($R = 0.71$). To further investigate the impacts of varying proportions of TE-derived sequences relative to intron length on genes, we classified the genes into four categories according to TE insertion proportions relative to intron length and intron length itself: short-intron no TE-insertion genes (SNG), short-intron low TE-ratio genes (SLG, TE ratio < 0.5), short-intron high TE-ratio genes (SHG), and long-intron TE-insertion genes (LG) (Fig. 5c; Fig. S12). In addition, genes with TE insertions possessed more introns than other genes, with LG genes harboring the largest number of introns (Fig. S13).

Further analysis of TE types inserted into introns revealed that various TE families were present in the SLG, SHG, and LG genes. Among these, LTR retrotransposons dominated in both insertion number and total length (Fig. 5d), indicating that LTRs are the primary TE type that are inserted into genic regions. A comparison of TE sequence identity distributions in introns of the three TE-containing gene categories revealed that SLG genes had significantly lower TE identity than did SHG genes and LG genes (Fig. 5e; Fig. S14), suggesting that more ancient TE insertions occur in SLG genes.

a. Comparison of the full gene length, CDS length, and intron length distributions between *D. fragrans* and other representative plant species. **b.** Density plot showing the relationship between intron length and the proportion of TE-derived sequences in introns. Red and blue represent long-intron genes and short-intron genes, respectively; solid lines indicate linear regression fits. **c.** Classification of genes into five categories on the basis of intron length and TE content proportion, with the number of genes in each category displayed. **d.** Insertion number and total length of different TE families in introns of TE-containing genes. **e.** Sequence identity distribution of TEs in introns of TE-containing genes.

TE insertions in promoter regions may affect gene expression.

To investigate the potential impacts of TE insertions on gene expression, we compared the expression levels of the four gene categories across different tissues and distinct developmental stages of sporangia in *D. fragrans*³⁶. The results revealed that LG genes generally presented the highest expression levels, whereas SHG genes presented very low expression (Fig. 6a). Since the promoter region plays a key role in gene expression regulation, we examined TE insertions in the promoters of

these four gene categories. We found that, compared with the other categories, the SHG genes presented significantly more TE insertions in their promoters (Fig. 6b), which may account for the relatively low expression levels of SHG genes.

To further explore the relationship between the length of TE insertions in promoter regions and gene expression levels, we divided the TE insertion length in the promoter regions into four 500-bp intervals and compared their effects on gene expression. Across the four gene categories, gene expression levels were consistently the lowest when the TE insertion length ranged from 1500 to 2000 bp (Fig. 6c). Notably, in SHG genes, TE insertions with lengths ranging from 0 to 1500 bp may promote gene expression. These findings suggest that TE can influence gene expression by being inserted into promoter regions.

We used four tandemly duplicated UDP-glycosyltransferases (UDPGTs) as a representative example to illustrate the impacts of TE insertions in the promoter regions of SHG genes on gene expression levels and the underlying potential mechanisms. The promoter region of *UDPGT1* contained a 1912-bp TE insertion, while *UDPGT2* and *UDPGT3* had 518 bp and 344 bp TE insertions, respectively, and no TE insertions were detected in the promoter of *UDPGT4* (Fig. 6d). The transcriptome data revealed no expression of *UDPGT1* or *UDPGT4*, whereas *UDPGT2* and *UDPGT3* were expressed in glandular trichomes and spore-bearing leaves. Further investigation revealed that the loss of the TATA-box core element in the promoter of *UDPGT1* due to extensive TE insertions likely accounts for the lack of expression. In contrast, the promoter regions of *UDPGT2* and *UDPGT3* harbored TE insertions that contained additional functional cis-acting elements, such as CAAT-boxes and MYB-binding sites, suggesting that TEs may modulate gene expression by introducing regulatory elements.

a. Expression levels of four gene categories (LG, SHG, SLG, SNG) across different tissues and stages of sporangium development ($p < 0.001$). **b.** Length distribution of TE insertions in the promoter regions of the four gene categories. **c.** Comparison of gene expression levels across different insertion size intervals of TEs in the promoter regions of the four gene categories. The x-axis represents the insertion length of TEs, with intervals of 500 bp. **d.** Expression statistics for four *UDPGT* genes. The structure of each gene is shown at the bottom, where wide blocks represent exons, narrow blocks represent introns, and white arrows indicate the promoter regions. The red bars above the gene structures represent TEs. The histograms above show the distribution of transcriptome reads, with the numbers on the left indicating the normalized read peak values.

TE insertions can enhance alternative splicing (AS) of genes

To investigate whether TE insertion into genic regions regulates AS, we performed a genome-wide systematic analysis of alternative splicing in *D. fragrans* and identified a total of 105,203 alternative splicing events. The results showed that approximately 66% of the genes exhibited alternative splicing characteristics, with an average of 6 alternative splicing isoforms per gene. Among all AS types, alternative first exon (AF) events accounted for the highest proportion, reaching 23.54%. Notably, as

many as 80.42% of mutually exclusive exon (MX) events occurred in TE-inserted genes (Fig. S15a), suggesting that TE insertion may be one of the factors driving MX-type AS events. Further comparison of AS characteristics across different gene categories revealed that, compared with SNG genes without TE insertions, TE-inserted genes (LG, SHG and SLG) had significantly higher proportions of alternative splicing genes (ASGs) and greater numbers of AS events. Among these, LG genes not only had the highest proportion of AS occurrence but also the largest number of AS events (Fig. S15b, c).

Discussion

We generated a high-quality chromosome-level genome assembly of *D. fragrans*—a fern species inhabiting sun-exposed volcanic-lava habitats (Fig. 1; Fig. S1). As the first genome of a specialized-habitat fern within Dryopteridaceae, this assembly not only fills a critical gap in fern genome research but also provides a foundational resource for comparative genomics studies of early land plants.

In this study, we found that the core factor driving genome evolution in *D. fragrans* is not frequent WGD (Fig. 2; Fig. S5), but rather TEs associated with the expansion of gene families related to adaptation to specialized environments (Fig. 3; Fig. S6-8). This TE-mediated gene expansion pattern dependent on DSDs is consistent with the mechanistic models in angiosperms, where mobile elements regulate gene transposition and the acquisition of new genes^{37,38}. Our results extend this evolutionary paradigm to leptosporangiate ferns, suggesting that distantly diverged land plant lineages share convergent evolutionary paths in gene family expansion. This convergence implies that, in the absence of large-scale WGD, TE-mediated small-scale gene duplication may represent a universal strategy for adaptive evolution—a pattern particularly relevant for early land plant lineages with conserved diploidy¹⁴.

Combined with analyses of TE insertion frequency, this study suggests that TEs in the *D. fragrans* genome exhibit a more distinct dynamic evolutionary pattern of continuous insertion and rapid removal (Fig. 4; Fig. S9). The habitat of *D. fragrans* is characterized by intense ultraviolet radiation and large temperature fluctuations, all of which have been confirmed to induce TE activation in plants^{39–42}. Rapid removal of intact TEs can minimize genomic instability caused by TE hyperactivity, while continuous TE insertion events can reserve abundant genetic variation for the adaptive evolution of the species. This dynamic TE pattern may represent a previously unrecognized genomic adaptation strategy of ferns in specialized habitats, a conclusion that merits further validation in other fern species adapted to similar habitats.

Further studies revealed that TEs can also insert into the genic regions of *D. fragrans* to affect gene structure, leading to the presence of a class of genes with longer introns in the *D. fragrans* genome (Fig. 5; Fig. S10). Furthermore, genes with TE insertions have a greater number of introns (Fig. S13) and more alternative splicing events (Fig. S15), and TEs within genic regions exhibit higher stability (Fig. S11). Meanwhile, TE insertions in gene promoter regions may regulate gene expression by introducing additional cis-acting elements (Fig. 6). This regulatory effect has been previously reported in seed plants

^{16,22}, and the present study extends this finding to ferns, suggesting that the TE-mediated regulatory innovation mechanism is conserved across clades.

This study still has certain methodological and interpretive limitations. First, this study only used the genome of a single species, *D. fragrans* from a specialized habitat, to explore the driving role of TEs in fern genome evolution, which has a certain limitation in representativeness. Second, all analyses in this study were based on experimental materials from a single individual, so it was impossible to assess the variation characteristics of TE insertion patterns and copy numbers within natural populations; future population-level sequencing analyses are needed to clarify whether the TE characteristics observed in this study are fixed or variable across different habitats. Third, this study did not analyze the epigenetic landscape regulating TE stability and promoter activity; to further explore the mechanism by which TEs affect gene expression, future studies could investigate relevant epigenetic modification characteristics or use gene editing technology to knockout TE insertions within genes, thereby providing direct experimental evidence for the regulatory functions of TEs.

Materials and Methods

Plant materials and genome sequencing

Plant materials used in this study were collected from the Wudalianchi region, Heilongjiang Province, China. Spores were isolated from sporophytes of *D. fragrans* and subsequently cultured to the gametophyte stage. Thereafter, gametophyte tissues were propagated through successive generations on sterile solid media containing (per liter) 4.43 g Murashige and Skoog (MS) basal salts, 10 g sucrose, and 7.5 g agar. Cultures were maintained under controlled environmental conditions (22°C, 16-hr light/8-hr dark photoperiod, 60% humidity) for 6–8 weeks to ensure robust biomass accumulation.

DNA was extracted from the gametophytes of *D. fragrans* via the SDS method, followed by purification with the QIAGEN® Genomic Kit (Cat# 13343). DNA quality was assessed on 1% agarose gels to check for degradation and contamination. Purity was evaluated via a NanoDrop™ One UV–Vis spectrophotometer (Thermo Fisher Scientific), with OD260/280 ratios between 1.8 and 2.0 and OD260/230 ratios between 2.0 and 2.2. The DNA concentration was measured via a Qubit® 3.0 fluorometer (Invitrogen, USA).

For long-read sequencing, 2 µg of qualified DNA per sample was used as input for Oxford Nanopore Technologies (ONT) library preparation. DNA was size-selected via the BluePippin system (Sage Science, USA) before end-repair and A-tailing with the NEBNext Ultra II End Repair/dA-tailing Kit (Cat# E7546). Adapter ligation was performed via an LSK109 kit (Oxford Nanopore Technologies). Library quality was quantified with a Qubit® 3.0 fluorometer. Sequencing was performed on the GridION X5/PromethION platform (Oxford Nanopore Technologies).

Genome Assembly

The genome assembly was carried out via ONT reads via a hybrid approach. First, raw Nanopore reads were basecalled from FAST5 to FASTQ via Guppy (version 3.2.2 + 9fe0a78)⁴³. Low-quality reads (mean_qscore_template < 7) were filtered out. De novo genome assembly was conducted via NextDenovo (v2.3.1)⁴⁴, which employs an overlap-layout-consensus (OLC) strategy. Given the high error rate of ONT reads, subreads were self-corrected via NextCorrect to generate consistent sequences (CNS reads). After CNS correlation analysis with NextGraph, a preliminary genome assembly was constructed.

To refine the assembly, contigs were corrected via Racon (v1.3.1)⁴⁵ (with ONT long reads) and polished via Nextpolish (v1.3.0)^{46,47} (using Illumina short reads). Redundant contigs were removed via similarity searches (identity ≥ 80%, overlap ≥ 80%). The completeness of the genome was assessed via BUSCO (v4.0.5)⁴⁸ and CEGMA (v2)⁴⁹. Assembly accuracy was evaluated by mapping Illumina paired-end reads to the genome with BWA (0.7.12-r1039)⁵⁰ and SAMtools (v1.4)⁵¹ to assess the mapping rate and genome coverage. Base accuracy was further calculated with BCFtools (v1.8.0)⁵². Additionally, RNA-seq reads were aligned to the genome to assess gene coverage, and mitochondrial sequences were excluded by submitting the draft genome to the NT library for sequence filtering⁵³. For telomere prediction, the quartet_teloexplorer.py script from quarTeT was utilized, in which the species was specified as "plant", and all other parameters were set to default values⁵⁴.

Hi-C scaffolding

To anchor the scaffolds to chromosomes, Hi-C libraries were prepared from the genomic DNA of the reference cultivar. Freshly harvested leaves were vacuum infiltrated in nuclei isolation buffer, fixed with formaldehyde, and ground into powder. Nuclei were isolated and digested with DpnII. Biotin-14-dCTP was incorporated into the DNA, and unligated DNA ends were removed via T4 DNA polymerase exonuclease activity. The ligated DNA was sheared into 300 – 600 bp fragments, repaired, and A-tailed before being purified via biotin–streptavidin pull-down. The Hi-C library was quantified and sequenced on the Illumina NovaSeq/MGI-2000 platform.

Hi-C raw data quality was controlled via Hi-C-Pro (v3.1.0)⁵⁵, with filtering for low-quality sequences (quality score < 20), adaptor sequences, and sequences shorter than 30 bp. Clean paired-end reads were aligned to the draft genome via Bowtie2 (v2.3.2)⁵⁶. Valid interaction pairs were identified and retained via Hi-C-Pro (v3.1.0) for further analysis. The scaffolds were clustered, ordered, and oriented via LACHESIS, with parameters set for minimum resites, link density, and the noninformative ratio. Manual adjustment was performed to correct any orientation errors and improve chromosome-level scaffold placement.

Gene annotation and functional annotation

Gene prediction was carried out via three independent approaches: ab initio prediction, homology-based prediction, and RNA-seq-based prediction. For homology-based gene prediction, the GeMoMa (v1.6.1)⁵⁷

tool was used to align homologous peptides from closely related species to the repeat-masked genome assembly, allowing us to obtain gene structure information. For RNA-seq-based prediction, filtered mRNA-seq reads were aligned to the reference genome via STAR (v2.7.3a)⁵⁸. Transcripts were assembled with StringTie (v1.3.4d)⁵⁹, and open reading frames (ORFs) were predicted via PASA (v2.3.3)⁶⁰. Additionally, RNA-seq reads were assembled de novo via StringTie, and the resulting transcripts were analyzed via PASA to generate a training set. Augustus (v3.3.1)⁶¹ was then employed for ab initio gene prediction via this training set. The final gene set was generated by integrating the predictions from the three approaches via EVidenceModeler (v1.1.1)⁶⁰.

For functional annotation, gene models were annotated by comparing the predicted proteins against several public databases: SwissProt, Non-Redundant Protein Database (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG), Clusters of Orthologous Groups of proteins (KOG), and GO. Protein domains and GO terms were identified via InterProScan (5.32–71.0)⁶². For the other four databases, BLASTP (v2.7.1)⁶³ was used to compare the predicted protein sequences against the public protein databases, and the best hits (with the lowest E value) were retained. The results from all five databases were concatenated to provide comprehensive functional annotations for the gene models.

Phylogenetic reconstruction and gene-family evolution

We retrieved publicly available genome assemblies and annotations for the species included in this study. Orthogroups were inferred with OrthoFinder (v2.5.5)⁶⁴ via default settings. Single-copy orthologs identified by OrthoFinder were aligned with MAFFT (v7.525)⁶⁵, and poorly covered alignment positions (coverage < 60%) were removed via trimAl (v1.4)⁶⁶. Model selection and maximum-likelihood tree inference were performed with IQ-TREE (v2.3.3)⁶⁷, and node support was assessed with 1,000 ultrafast bootstrap replicates. Divergence times were estimated with MCMCtree in the PAML (v4.10.7) package⁶⁸; analyses used a burn-in of 200,000 iterations followed by 10,000 sampled iterations with a sampling frequency of 5. Gene family expansion and contraction were inferred via CAFE (v5.0)⁶⁹. Species trees were visualized and annotated with FigTree (v1.4.4).

TE Annotation and LTR Analysis

For TE annotation, a *D. fragrans* TE library was constructed via EDTA (v2.2)⁷⁰. EDTA incorporates both structure- and homology-based detection programs to annotate the predominant TE classes found in plant genomes. Unknown TEs were isolated separately and reclassified via DeepTE⁷¹. The complete TEs were further categorized via Tesorter (v1.4.6)⁷². Ultimately, a comprehensive TE database for *D. fragrans* was established. Subsequent summary statistics and visualizations were performed via the R packages tidyverse and ggplot2.

Gene Structure Analysis

First, the GFF3 files of the species under analysis were filtered to retain only protein-coding genes. The positions of exons and genes were extracted, and intron positions were identified by calculating the complement via BEDTools (v2.31.1)⁷³. Gene length, exon length, and intron length were then quantified, with visualizations created via the R packages ggplot2 and ggridges. Genes with log10-transformed intron lengths greater than four were defined as having long introns. TE insertions within these long-intron genes and other genes were assessed via bedtools. The types and lengths of TEs inserted into introns were then summarized.

Identification of Solo/Intact LTRs

To evaluate the dynamics of LTR retrotransposons, intact and solo LTR elements in the *D. fragrans* genome were identified and classified. An intact LTR was defined as a retrotransposon containing two recognizable LTR sequences (both ends) within 1,000 bp of the flanking regions, along with an internal coding region, typically including domains such as gag, pol, and reverse transcriptase.

In contrast, a single LTR was defined as a partial element retaining only a single LTR sequence along with a truncated internal region or lacking it entirely. These genes were identified via EDTA and further validated through sequence homology and structure-based annotation.

WGD analysis

To assess the WGD history of *D. fragrans*, first, conserved collinear blocks within the assembled genome were identified via intragenomic synteny analysis via JCVI⁷⁴. Only gene pairs located within syntenic blocks containing at least five collinear genes were retained for further analysis. The number and size of these blocks were used to evaluate large-scale duplication events.

To estimate the timing of gene duplications, Ks was calculated for paralogous gene pairs via ParaAT (v2.0)⁷⁵ and KaKs_Calculator (v3.0)⁷⁶. The Ks distribution was fitted with a Gaussian model to detect putative peaks indicative of ancient or recent WGD events.

In addition, the Multitaxon Paleopolyploidy Search (MAPS)⁷⁷ approach was applied to identify phylogenetically congruent duplication events across ferns. Gene family trees were reconstructed via genome and transcriptome data from multiple species via OrthoFinder (v2.5.5). To reduce transcript redundancy, transcriptome datasets were clustered via cd-hit (v4.8.1)⁷⁸ at a 90% identity threshold prior to analysis. Duplicated signals were then mapped onto internal nodes of the species phylogeny. To assess the significance of the observed duplication patterns, we performed simulation-based comparisons against both null and positive models.

Full-Length Transcriptome Sequencing and Alternative Splicing Analysis

To characterize the full-length transcriptome and identify alternative splicing (AS) events in *D. fragrans*, third-generation sequencing was performed via the PacBio CCS platform on samples collected from

various tissues and developmental stages. HiFi reads were processed with the Iso-Seq3 (v4.2.0) pipeline to generate full-length nonconcatemer (FLNC) reads. These FLNC reads were then aligned to the reference genome via Minimap2-2.27 (r1193)⁷⁹, and the resulting alignments were integrated to produce a preliminary GTF annotation file. To refine the transcript models, we used SQANTI3 (v5.3.6)⁸⁰ to compare the full-length transcriptome-based annotations with the existing genome annotations, resulting in a final, high-confidence gene annotation set. For AS analysis, SUPPA2 (v2.4)⁸¹ was used to identify AS events on the basis of the full-length transcriptome annotation. The inclusion levels (percent spliced-in, PSI) of splicing events were calculated by integrating second-generation (Illumina) transcriptome data.

Declarations

Acknowledgements

This work was supported by the National Natural Science Foundation of China (no. 32270394 to Y Chang and no.32370243 to YH Fang), a startup fund from Linyi University (LYDX2019BS039). We gratefully acknowledge financial support from the “Double First-Class” initiative of Heilongjiang Province for the advantageous and characteristic discipline of Chinese Materia Medica Biogenetics. NGS, ONT, and Hi-C sequencing of the genome were conducted via grandomic methods. Full-length transcriptome sequencing was performed by BerryGenomics Co., Ltd.

Data availability

The *D. fragrans* genome assembly and all of the raw sequencing data have been deposited at China National Center for Bioinformation (CNCB)^{82,83}, under the BioProject accession number PRJCA042196.

Authors' contributions

Y.C. and S.W. conceived the study. W.D., Q.W. and Y.F. designed and managed the major scientific objectives. D.Z. and C.S. managed the plant materials. X.Q. assembled the genome and estimated the genome size. W.D. annotated the genome and transposons and performed the data analysis. D.Z. contributed to the full-length transcriptome sequencing. W.D., Q.W., Y.F. and S.W. led the manuscript preparation. All the authors read and approved the final manuscript.

References

1. Testo, W., Sundue, M.: A 4000-species dataset provides new insight into the evolution of ferns. *Mol. Phylogenet. Evol.* **105**, 200–211 (2016)
2. Kenrick, P., Crane, P.R.: The origin and early evolution of plants on land. *Nature*. **389**, 33–39 (1997)
3. Nildas, K.J., Tiftneyt, B.: H. Patterns in vascular land plant diversification. (1983)

4. I, P.: A community-derived classification for extant lycophytes and ferns. *J. Syst. Evol.* **54**, 563–603 (2016)
5. Sessa, E.B., Der, J.P.: Chapter Seven - Evolutionary Genomics of Ferns and Lycophytes. In: Rensing, S.A. (ed.) *Advances in Botanical Research*, vol. 78, pp. 215–254. Academic (2016)
6. Fernández, P., et al.: A 160 Gbp fern genome shatters size record for eukaryotes. *iScience* **27**, (2024)
7. Huang, C.-H., Qi, X., Chen, D., Qi, J., Ma, H.: Recurrent genome duplication events likely contributed to both the ancient and recent rise of ferns. *J. Integr. Plant Biol.* **62**, 433–455 (2020)
8. Pelosi, J.A., Kim, E.H., Barbazuk, W.B., Sessa, E.B.: Phylotranscriptomics Illuminates the Placement of Whole Genome Duplications and Gene Retention in Ferns. *Front. Plant. Sci.* **13**, (2022)
9. Fang, Y., et al.: The genome of homosporous maidenhair fern sheds light on the euphyllophyte evolution and defences. *Nat. Plants.* **8**, 1024–1037 (2022)
10. Marchant, D.B., et al.: Dynamic genome evolution in a model fern. *Nat. Plants.* **8**, 1038–1051 (2022)
11. Li, F.-W., et al.: Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants.* **4**, 460–472 (2018)
12. Huang, X., et al.: The flying spider-monkey tree fern genome provides insights into fern evolution and arborescence. *Nat. Plants.* **8**, 500–512 (2022)
13. Zhong, Y., et al.: Genomic Insights into Genetic Diploidization in the Homosporous Fern *Adiantum nelumboides*. *Genome Biol. Evol.* **14**, evac127 (2022)
14. Chen, H., et al.: Revisiting ancient polyploidy in leptosporangiate ferns. *New Phytol.* **237**, 1405–1417 (2023)
15. Alseekh, S., Scossa, F., Fernie, A.R.: Mobile Transposable Elements Shape Plant Genome Diversity. *Trends Plant Sci.* **25**, 1062–1064 (2020)
16. Hirsch, C.D., Springer, N.M.: Transposable element influences on gene expression in plants. *Biochim. et Biophys. Acta (BBA) - Gene Regul. Mech.* **1860**, 157–165 (2017)
17. Hassan, A.H., Mokhtar, M.M., El Allali, A.: Transposable elements: multifunctional players in the plant genome. *Front. Plant. Sci.* **14**, (2024)
18. Wicker, T., et al.: A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007)
19. Zhang, X., Wessler, S.R.: Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. *Proceedings of the National Academy of Sciences* **101**, 5589–5594 (2004)
20. Staton, S.E., Burke, J.M.: Evolutionary transitions in the Asteraceae coincide with marked shifts in transposable element abundance. *BMC Genom.* **16**, 623 (2015)
21. Wicker, T., et al.: Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.* **19**, 103 (2018)
22. LTR-retrotransposons in plants: Engines of evolution. *Gene.* **626**, 14–25 (2017)

23. Li, Z.-W., et al.: Transposable Elements Contribute to the Adaptation of *Arabidopsis thaliana*. *Genome Biol. Evol.* **10**, 2140–2150 (2018)
24. Makarevitch, I., et al.: Transposable elements contribute to activation of maize genes in response to abiotic stress. *PLoS Genet.* **11**, e1004915 (2015)
25. Lisch, D.: How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61 (2013)
26. Qin, G., et al.: Chromosome-Scale Genome of the Fern *Cibotium barometz* Unveils a Genetic Resource of Medicinal Value. *Horticulturae*. **10**, 1191 (2024)
27. Wei, Z.: Resolving the Stasis-Dynamism Paradox: Genome Evolution in Tree Ferns
28. Pelosi, J., et al.: The genome of the vining fern *Lygodium microphyllum* highlights genomic and functional differences between life phases of an invasive plant. Preprint at. (2025). <https://doi.org/10.1101/2025.03.06.640867>
29. Sessa, E.B., Zimmer, E.A., Givnish, T.J.: Phylogeny, divergence times, and historical biogeography of New World *Dryopteris* (Dryopteridaceae). *Am. J. Bot.* **99**, 730–750 (2012)
30. Zuo, Z.-Y., et al.: A revised classification of Dryopteridaceae based on plastome phylogenomics and morphological evidence, with the description of a new genus, *Pseudarachniodes*. *Plant. Divers.* **47**, 34–52 (2025)
31. Chen, L., et al.: Microbial-type terpene synthases significantly contribute to the terpene profile of glandular trichomes of the fern *Dryopteris fragrans* (L.). *Plant J.* **121**, e70079 (2025)
32. Hassan, A.H., Mokhtar, M.M.: El Allali, A. Transposable elements: multifunctional players in the plant genome. *Front. Plant. Sci.* **14**, 1330127 (2024)
33. Zhang, Y., et al.: Evolutionary rewiring of the wheat transcriptional regulatory network by lineage-specific transposable elements. *Genome Res.* **31**, 2276–2289 (2021)
34. Choulet, F., et al.: Megabase Level Sequencing Reveals Contrasted Organization and Evolution Patterns of the Wheat Gene and Transposable Element Spaces. *Plant. Cell.* **22**, 1686–1701 (2010)
35. Song, C., Guan, Y., Zhang, D., Tang, X., Chang, Y.: Integrated mRNA and miRNA Transcriptome Analysis Suggests a Regulatory Network for UV–B-Controlled Terpenoid Synthesis in Fragrant Woodfern (*Dryopteris fragrans*). *IJMS* **23**, 5708 (2022)
36. Lu, Z., Huang, Q., Zhang, T., Hu, B., Chang, Y.: Global transcriptome analysis and characterization of *Dryopteris fragrans* (L.) Schott sporangium in different developmental stages. *BMC Genom.* **19**, 471 (2018)
37. Xiao, Y., Wang, J.: Understanding the Regulation Activities of Transposons in Driving the Variation and Evolution of Polyploid Plant Genome. *Plants.* **14**, 1160 (2025)
38. Oliver, K.R., McComb, J.A., Greene, W.K.: Transposable Elements: Powerful Contributors to Angiosperm Evolution and Diversity. *Genome Biol. Evol.* **5**, 1886–1901 (2013)
39. Mhiri, C., Borges, F., Grandbastien, M.-A.: Specificities and Dynamics of Transposable Elements in Land Plants. *Biology.* **11**, 488 (2022)

40. Negi, P., Rai, A.N., Suprasanna, P.: Moving through the Stressed Genome: Emerging Regulatory Roles for Transposons in Plant Stress Response. *Front. Plant. Sci.* **7**, (2016)
41. Roquis, D., et al.: Genomic impact of stress-induced transposable element mobility in *Arabidopsis*. *Nucleic Acids Res.* **49**, 10431–10447 (2021)
42. Thieme, M., et al.: Experimentally heat-induced transposition increases drought tolerance in *Arabidopsis thaliana*. *New Phytol.* **236**, 182–194 (2022)
43. Wick, R.R., Judd, L.M., Holt, K.E.: Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019)
44. Hu, J., et al.: An efficient error correction and accurate assembly tool for noisy long reads. Preprint at. (2023). <https://doi.org/10.1101/2023.03.09.531669>
45. Vaser, R., Sović, I., Nagarajan, N., Šikić, M.: Fast and accurate de novo genome assembly from long uncorrected reads. 068122 Preprint at (2016). <https://doi.org/10.1101/068122>
46. Hu, J., Fan, J., Sun, Z., Liu, S.: NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics.* **36**, 2253–2255 (2020)
47. Chen, S., Zhou, Y., Chen, Y., Gu, J.: fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* **34**, i884–i890 (2018)
48. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M.: BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* **31**, 3210–3212 (2015)
49. Parra, G., Bradnam, K., Korf, I.: CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics.* **23**, 1061–1067 (2007)
50. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics.* **25**, 1754–1760 (2009)
51. Li, H., et al.: The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* **25**, 2078–2079 (2009)
52. Danecek, P., McCarthy, S.A.: BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics.* **33**, 2037–2039 (2017)
53. Kim, D., Langmead, B., Salzberg, S.L.: HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods.* **12**, 357–360 (2015)
54. Lin, Y., et al.: quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* **10**, uhad127 (2023)
55. Servant, N., et al.: HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015)
56. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* **9**, 357–359 (2012)
57. Keilwagen, J., et al.: Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89–e89 (2016)

58. Dobin, A., et al.: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013)
59. Kovaka, S., et al.: Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019)
60. Haas, B.J., et al.: Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008)
61. Stanke, M., Diekhans, M., Baertsch, R., Haussler, D.: Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics*. **24**, 637–644 (2008)
62. Zdobnov, E.M., Apweiler, R.: InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. **17**, 847–848 (2001)
63. McGinnis, S., Madden, T.L.: BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20–W25 (2004)
64. Emms, D.M., Kelly, S.: OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019)
65. Katoh, K.: MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002)
66. Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T.: trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. **25**, 1972–1973 (2009)
67. Minh, B.Q., et al.: IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020)
68. Yang, Z.: PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007)
69. 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)
70. Ou, S., et al.: Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019)
71. Yan, H., Bombarely, A., Li, S.: DeepTE: a computational method for *de novo* classification of transposons with convolutional neural network
72. Zhang, R.-G., Wang, Z.-X., Ou, S., Li, G.-Y.: TESorter: lineage-level classification of transposable elements using conserved protein domains. Preprint at. (2019). <https://doi.org/10.1101/800177>
73. Quinlan, A.R., BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *CP Bioinf.* **47**, (2014)
74. Tang, H., et al.: JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* **3**, e211 (2024)
75. Zhang, Z., et al.: ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* **419**, 779–781 (2012)
76. Zhang, Z.: KaKs_Calculator 3.0: Calculating Selective Pressure on Coding and Non-Coding Sequences. *Genom. Proteom. Bioinform.* **20**, 536–540 (2022)
77. Li, Z., et al.: Early genome duplications in conifers and other seed plants. *Sci. Adv.* (2015). <https://doi.org/10.1126/sciadv.1501084>

78. Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W.: CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. **28**, 3150–3152 (2012)
79. Li, H.: Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**, 3094–3100 (2018)
80. Pardo-Palacios, F.J., et al.: SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat. Methods*. **21**, 793–797 (2024)
81. Trincado, J.L., et al.: SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 40 (2018)
82. Chen, T., et al.: The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. *Genom. Proteom. Bioinform.* **19**, 578–583 (2021)
83. Members, C.N.C.B.-N.G.D.C., Partners, et al.: Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2025. *Nucleic Acids Res.* **53**, D30–D44 (2025)

Tables

Table 1. Genome assembly statistics of *D. fragrans*.

Genome		
Size (Gb)	4.45	
Contig N50 (Mb)	5.70	
Scaffold N50 (Mb)	108.12	
Contig number	1,740	
Sequence anchored on chromosome (%)	99.33	
GC content (%)	40.53	
High-copy repeat content (%)	85.39	
Protein-coding gene		
Mean gene length (bp)	36,768	
Mean CDS length (bp)	1,301	
Mean exon length (bp)	251	
Mean intron length (bp)	8,497	
Mean exon number per gene	5.17	
Gene number	33,489	
Noncoding	Number	Total (bp)
miRNA	116	13,673
tRNA	3,811	287,962
rRNA	514	721,680

Figures

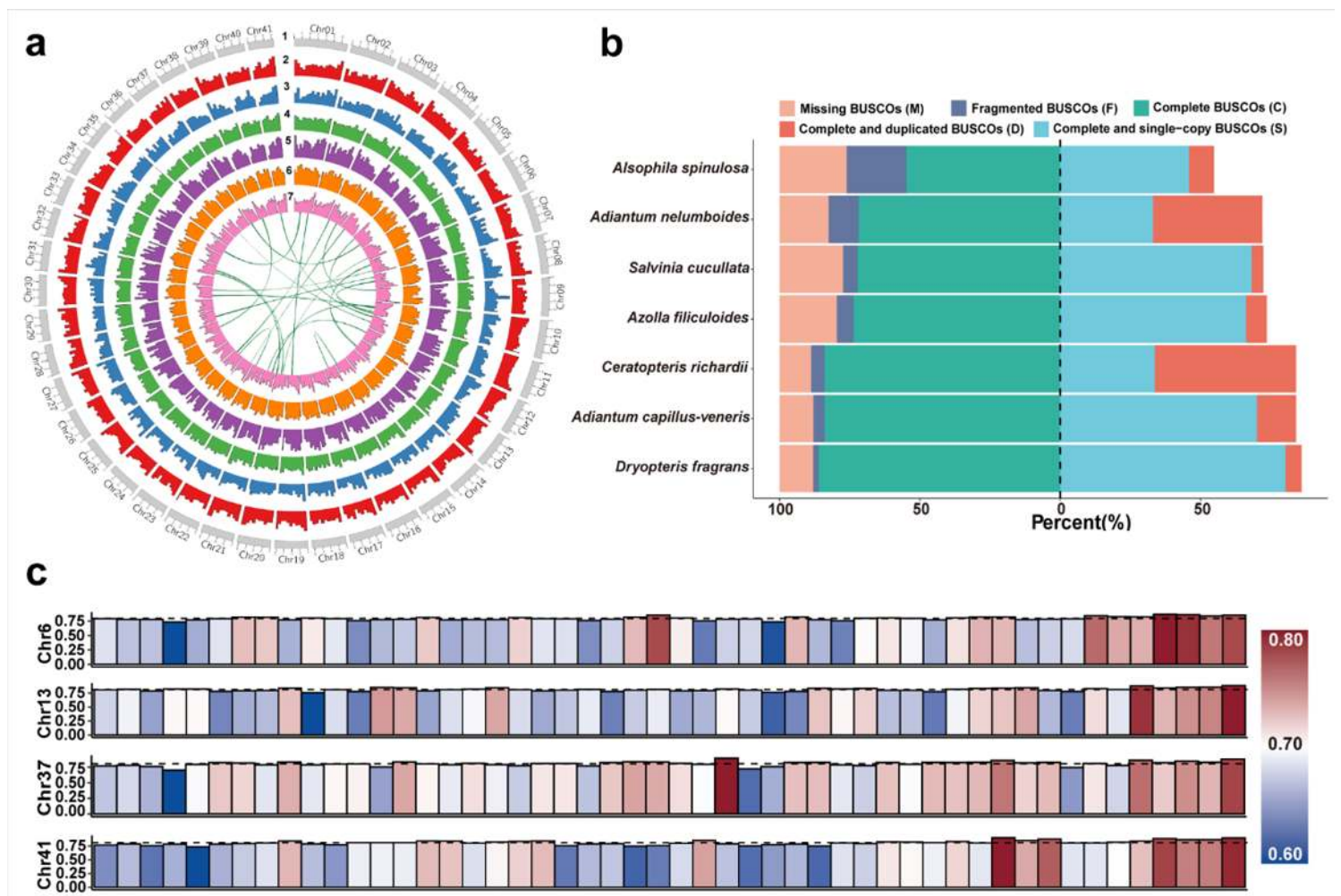


Figure 1

Chromosome-level genome assembly of *D. fragrans*.

a. Circos plot of the *D. fragrans* genome, divided into 10 Mb windows, showing (1) chromosome length and number; (2) content of repetitive sequences; (3) content of Solo-LTRs; (4) content of Gypsy; (5) content of Copia; (6) gene density; and (7) GC content, with the innermost circle representing synteny. **b.** BUSCO assessment of genome completeness across seven fern species. **c.** Distribution of repeat content along chromosomes visualized via bar plots and heatmaps. Each chromosome is divided into 50 windows (x-axis), and the repeat density within each window is shown on the y-axis and by color intensity.

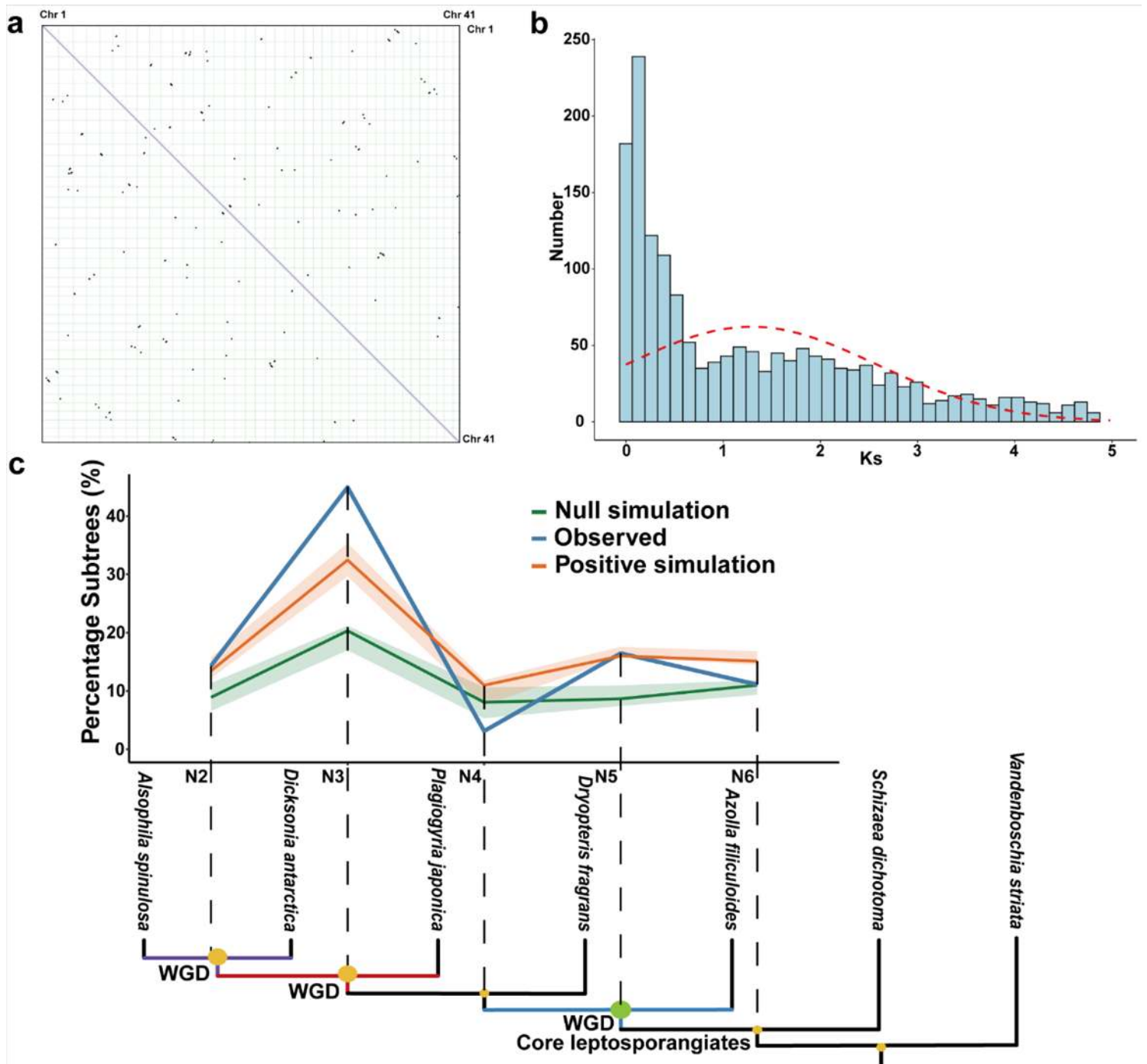


Figure 2

WGD analysis.

a. Intragenomic synteny analysis of *D. fragrans*, illustrating the distribution and retention of homologous gene blocks. **b.** Ks distribution for non-tandem duplicated homolog pairs. **c.** MAPS-based detection of WGD signals: the phylogenetic tree of selected fern species is shown below, with nodes N2–N6 indicated; above, the proportion of duplicated genes in each subtree is plotted, where the green, red, and blue curves represent the nullsimulation, positivesimulation, and observeddata, respectively. The shaded regions denote confidence intervals.

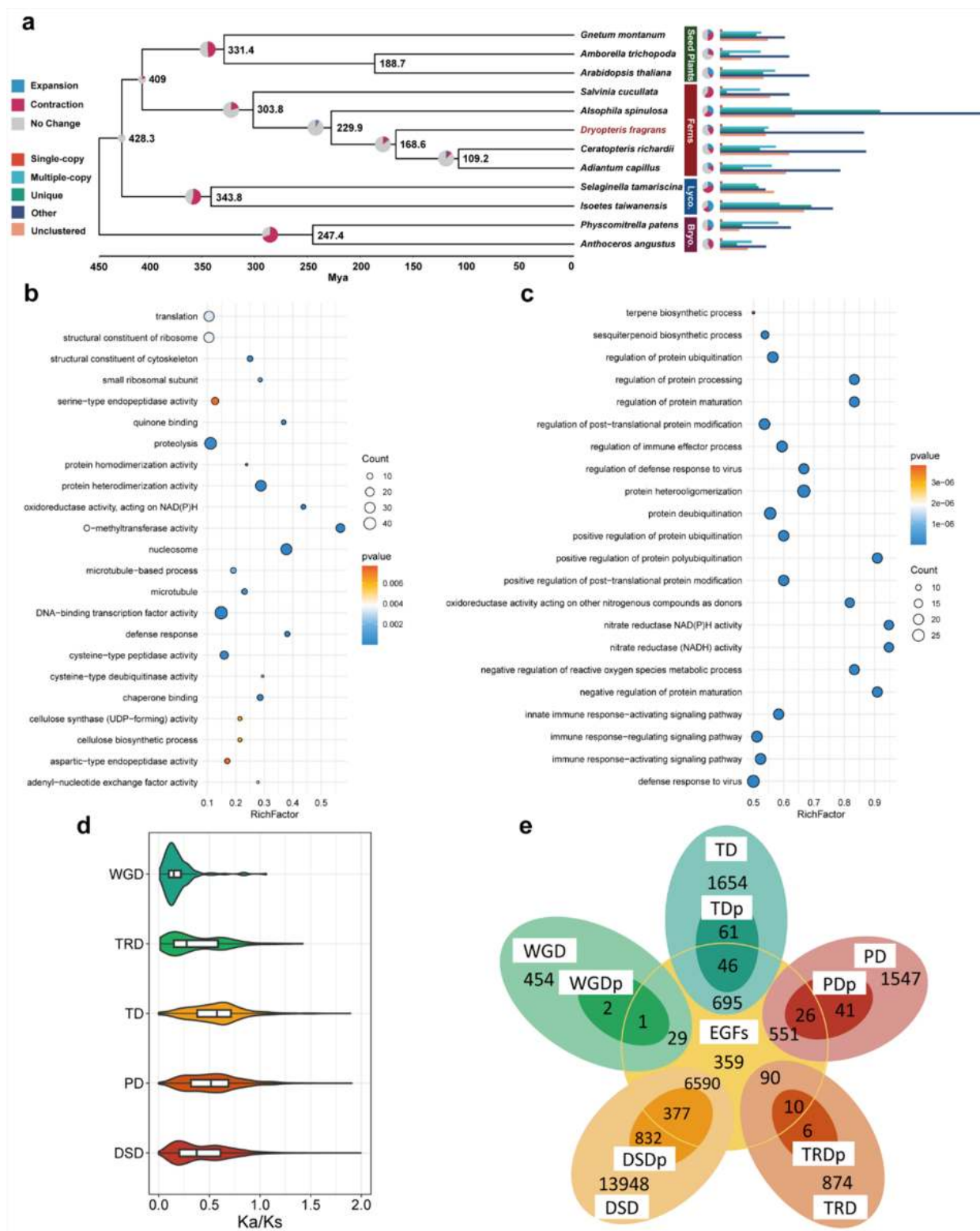


Figure 3

Phylogeny and gene-family evolution.

a. Species tree with divergence time estimates and numbers of expanded/contracted gene families. The tree was inferred from 253 single-copy orthogroups; divergence times are indicated in Mya. **b.** GO enrichment of *D. fragrans*-specific orthogroups. **c.** GO enrichment of orthogroups significantly enriched in

D. fragrans. **d.** Distributions of Ka/Ks ratios for genes originating from five duplication modes (WGD, TD, PD, TRD, DSD). **e.** Overlap analysis between genes in expanded orthogroups and genes derived from each duplication mode (WGD, TD, PD, TRD, DSD). Among them, WGDp, TDp, PDp, TRDp, and DSDp refer to positively selected genes. **f.** GO enrichment for genes originating from WGD pairs.

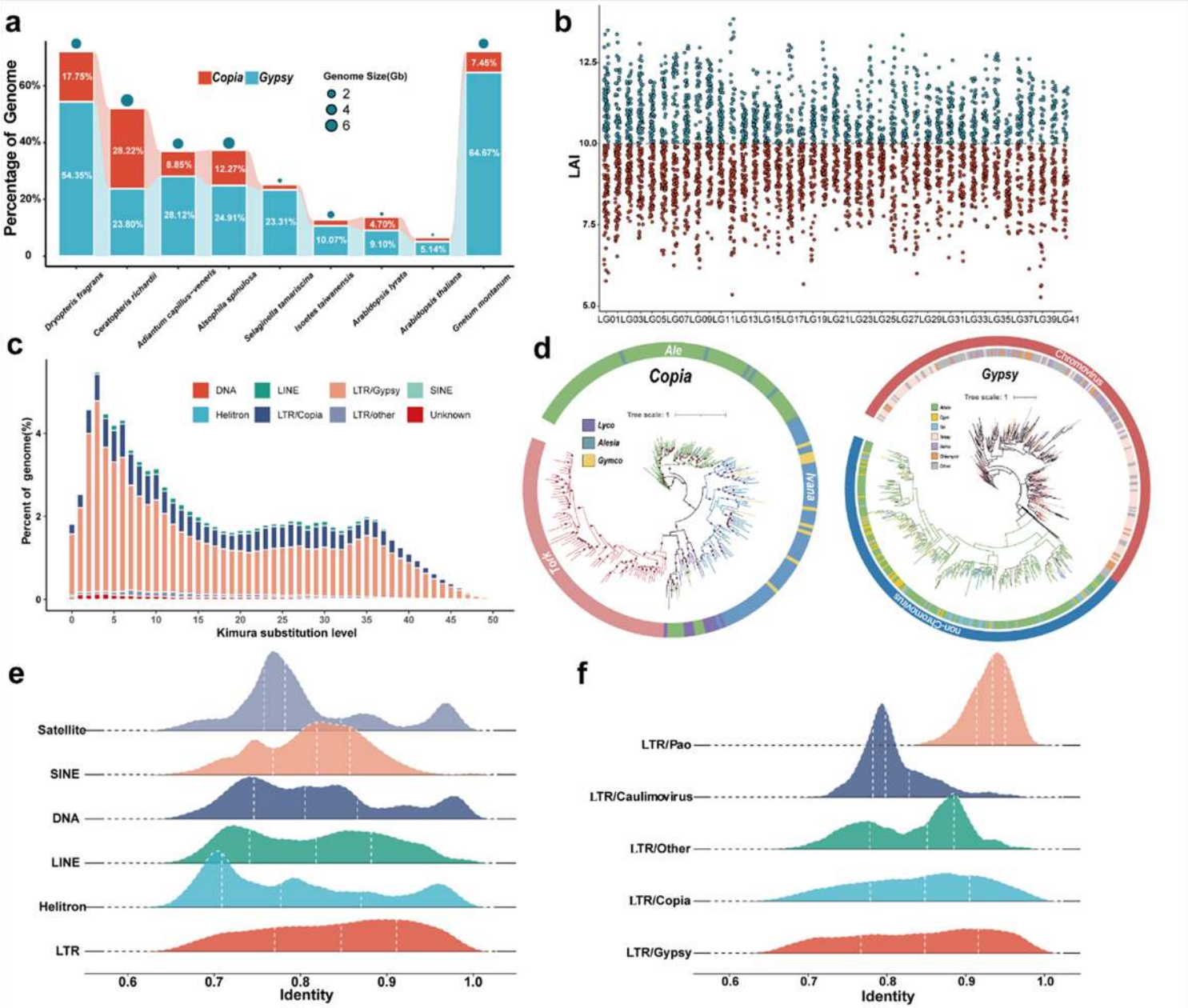


Figure 4

Annotation and characteristics of TEs in *D. fragrans*.

a. Comparison of LTR content and genome size between *D. fragrans* and several other plant species. **b.** LAI index of the *D. fragrans* genome, with regions >10 shown in blue and regions <10 shown in red. **c.** Kimura substitution levels of *D. fragrans* TEs. **d.** Phylogenetic trees of the six major LTR/Copia (left) and LTR/Gypsy (right) families. **e.** Identity distribution of TEs within the genome compared with a custom-

built TE database. **f.** Identity distribution of LTRs within the genome compared with a custom-built TE database.

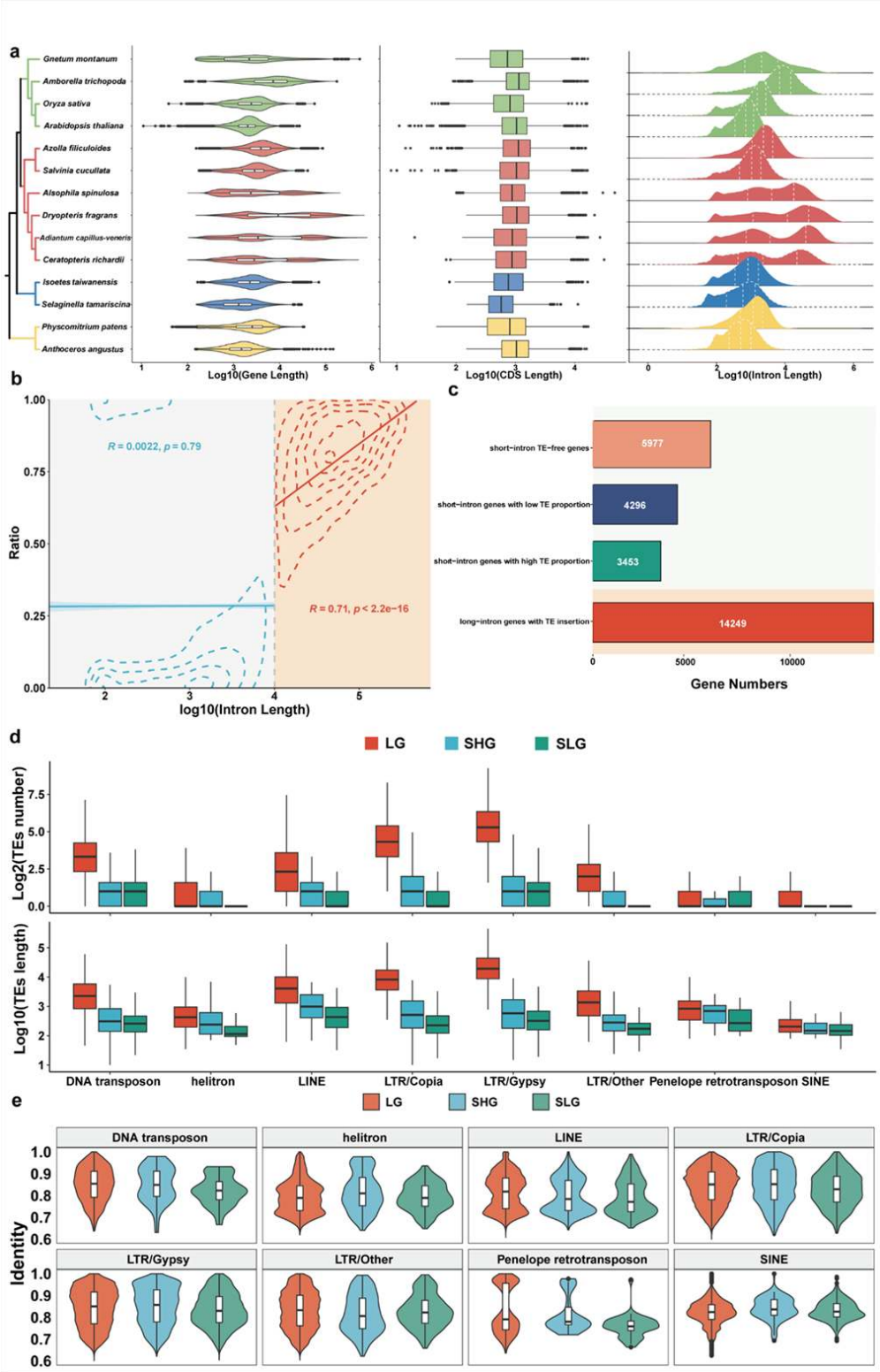


Figure 5

Effects of TEs on gene structure in *D. fragrans*.

a. Comparison of the full gene length, CDS length, and intron length distributions between *D. fragrans* and other representative plant species. **b.** Density plot showing the relationship between intron length and the proportion of TE-derived sequences in introns. Red and blue represent long-intron genes and short-intron genes, respectively; solid lines indicate linear regression fits. **c.** Classification of genes into five categories on the basis of intron length and TE content proportion, with the number of genes in each category displayed. **d.** Insertion number and total length of different TE families in introns of TE-containing genes. **e.** Sequence identity distribution of TEs in introns of TE-containing genes.

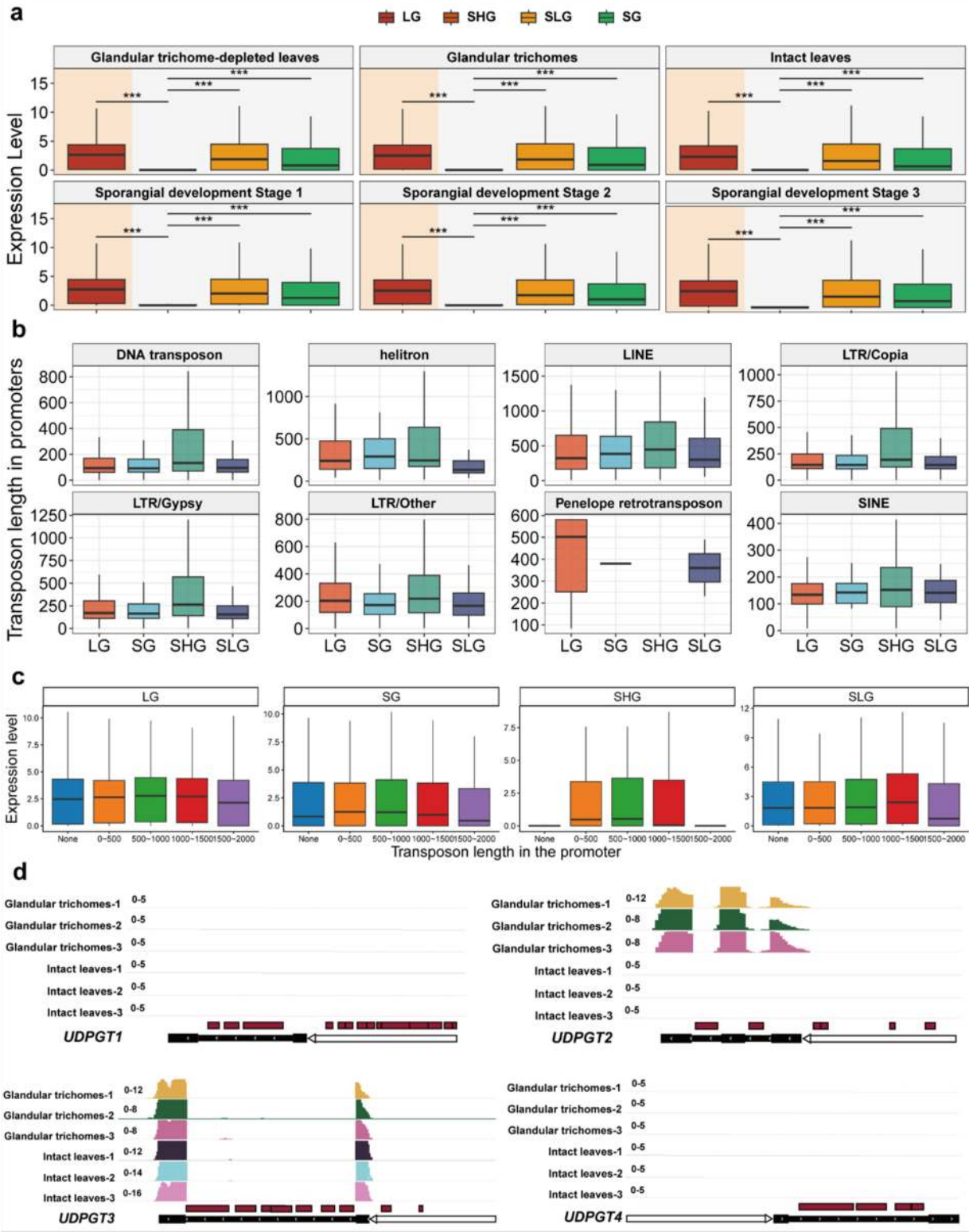


Figure 6

Impact of TE insertion in promoter regions on gene expression.

a. Expression levels of four gene categories (LG, SHG, SLG, SNG) across different tissues and stages of sporangium development ($p < 0.001$). **b.** Length distribution of TE insertions in the promoter regions of the four gene categories. **c.** Comparison of gene expression levels across different insertion size intervals of TEs in the promoter regions of the four gene categories. The x-axis represents the insertion length of TEs, with intervals of 500 bp. **d.** Expression statistics for four *UDPGT* genes. The structure of each gene is shown at the bottom, where wide blocks represent exons, narrow blocks represent introns, and white arrows indicate the promoter regions. The red bars above the gene structures represent TEs. The histograms above show the distribution of transcriptome reads, with the numbers on the left indicating the normalized read peak values.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTable.xlsx](#)
- [SupplementaryFigure.docx](#)